MSc Project Report 2023-2024

Genome-wide Association Study (GWAS) of SARS-CoV-2 Variants in a Gambian Cohort

Candidate number: 2401997

Page limit: 60

Project length: 48 pages

Standard Submitted in part fulfilment of the requirements for the degree of MSc Health Data Science September 2024

Abstract

Background: The COVID-19 pandemic, driven by SARS-CoV-2, has had heterogeneous impacts globally and within Africa. In The Gambia, limited genomic surveillance and resource-constrained healthcare settings posed unique challenges in tracking viral lineage emergence and their potential associations with viral load. This study investigated relationships between viral genetic variation and viral load, measured by RT-PCR cycle threshold (Ct) values, in a cohort of over 1,600 Gambian SARS-CoV-2 genomes.

Methods: Viral genome sequences with matched Ct data were compiled from routine surveillance. Standardized lineage assignment used Pangolin, followed by phylogenetic reconstruction and identification of ten most prevalent lineages. Quality control included linkage disequilibrium pruning and minor allele frequency filtering. Principal component analysis assessed population structure, revealing significant stratification with lineage A.29 as a genetic outlier. Three complementary genome-wide association study models tested SNP-Ct associations: a primary model with all samples, a sensitivity model excluding genetic outliers, and a stratified model removing A.29 samples to control for population structure artifacts.

Results: Phylogenetic analysis revealed distinct clusters reflecting local transmission dynamics and multiple viral introductions aligned with epidemic waves. Ten dominant lineages accounted for most genomes, with Delta variants (AY.34.1, B.1.617.2) being most prevalent. Population structure analysis identified significant genetic stratification, particularly A.29 lineage separation. Simple linear regression identified one significant SNP at position 15222, but this association disappeared when A.29 samples were excluded (p-value increased from 2.71×10^{-5} to 5.62×10^{-3}), indicating population stratification artifact rather than genuine association. Lineage-specific analyses revealed no genome-wide significant associations. One synonymous mutation in B.1.416 lineage showed significant association with Ct values but represents a lineage-specific effect rather than population-wide signal.

Conclusion: This study demonstrates critical methodological considerations for viral GWAS, highlighting how population structure can create spurious associations. The loss of significance when controlling for genetic stratification suggests that apparent SNP-Ct associations primarily reflect population structure rather than functional viral genetic effects. These findings emphasize the predominance of host and technical factors over viral genetic variation in determining Ct values, and underscore the necessity for rigorous population structure control in viral genomic association studies.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Stephan for his exceptional guidance throughout the GWAS workflow. His patient support whenever I encountered difficulties was invaluable, and his willingness to review all of my drafts late near the submission deadline significantly enhanced the quality of this work. Stephan's clear explanations and thoughtful feedback helped me overcome challenges and maintain momentum throughout the project.

1. Background & Purpose

The coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has created an unprecedented global public health crisis since its emergence in late 2019. SARS-CoV-2 primarily transmits via respiratory droplets, leading to clinical presentations ranging from asymptomatic infections to severe respiratory illness and death *(Morawska & Milton, 2020)*.

Globally, as of July 2025, there have been approximately 778 million confirmed COVID-19 cases and 7.1 million reported deaths *(World Health Organization, 2025)*. The pandemic's impact has been highly heterogeneous across regions. Europe, for instance, has reported about 281 million cases (36.1% of global total) and 2.3 million deaths (32.4% of global fatalities), whereas Africa—despite comprising roughly 17% of the world's population—has recorded only 9.6 million cases (1.2%) and 176,000 deaths (2.5%) *(World Health Organization, 2025)*.

The pandemic's course has been shaped by evolving viral variants with differing transmissibility and immune escape capabilities, resulting in multiple waves of infections worldwide *(Lewis, 2022)*.

1.2 Epidemiology of SARS-CoV-2 and Its Impact in The Gambia

In The Gambia, the COVID-19 pandemic posed significant public health challenges from its onset. The first confirmed case was reported on 17 March 2020, involving an imported infection from the United Kingdom *(World Health Organization, 2020)*.

By late 2022, official figures documented approximately 12,586 cases and 372 deaths *(World Health Organization, 2022)*, yet modelling and excess mortality analyses suggest that true infection and fatality rates were considerably higher *(COVID-19 Excess Mortality Collaborators, 2022)*. Limited healthcare infrastructure, a concentration of testing services in the capital, and widespread underdetection impeded the timely and accurate assessment of pandemic dynamics, especially in rural communities. These health system constraints hindered evidence-based allocation of resources and tailored public health policymaking *(Gilbert et al., 2020)*.

The spread of SARS-CoV-2 in The Gambia was marked by multiple epidemic waves, each associated with the emergence of distinct viral lineages and repeated introductions from abroad *(Sallah et al., 2023)*. Seasonal influences, though less pronounced than in temperate countries, were relevant: surges tended to coincide with the rainy season (June–October), when crowded indoor conditions and overlapping respiratory infections likely contributed to higher transmission *(Jarju et al., 2021)*. Similar seasonal and introduction-linked patterns were reported elsewhere in West Africa *(Salyer et al., 2021)*.

Clinically, Gambian patients displayed symptoms consistent with global case descriptions, including fever, cough, fatigue, sore throat, and loss of smell or taste *(Jarju et al., 2021)*. The burden of severe disease and mortality was disproportionately borne by the elderly and individuals with comorbidities such as hypertension, diabetes, and chronic respiratory disease, aligning with wider African and international evidence *(Salyer et al., 2021)*. However, weak vital registration systems and limited diagnostic capacity meant that true morbidity and mortality were likely undercounted *(COVID-19 Excess Mortality Collaborators, 2022)*.

A major limitation in pandemic management was the scarcity of genomic surveillance. Between 2020 and early 2022, more than 11,900 COVID-19 cases were reported, but only about 1,638 genomes were sequenced *(Kanteh et al., 2023)*. This paucity of data reflects a broader challenge, as only ~1% of global SARS-CoV-2 sequences originate from Africa *(Kanteh et al., 2023)*. As a result, real-time monitoring of emerging variants was hindered, limiting understanding of transmission dynamics. In The Gambia, each epidemic wave was associated with the introduction of new lineages first identified abroad, underscoring the importance of strengthening genomic surveillance to support epidemic preparedness and control *(Sallah et al., 2023; Kanteh et al., 2023)*.

1.2.1 Waves of Introduction and Transmission Dynamics in the Gambia

SARS-CoV-2 is primarily transmitted through respiratory droplets and aerosols, especially in poorly ventilated indoor environments (Morawska & Milton, 2020). Evidence from both global and African studies has shown that asymptomatic and presymptomatic carriers contributed substantially to viral spread, complicating containment efforts (Buitrago-Garcia et al., 2020). In The Gambia, where large extended households and communal social activities are common, household-based transmission played a particularly important role, consistent with findings from West Africa and other resource-limited settings (Sallah et al., 2023).

The Gambian epidemic was also characterised by multiple distinct waves, each associated with particular viral lineages and heavily influenced by repeated introductions from abroad combined with local seasonal patterns:

First Wave (March–July 2020): The earliest cases were dominated by lineage B.1.416, most likely introduced by international travellers from Europe. Epidemiological and phylogenetic evidence reveal several importation events, including multiple introductions directly from Senegal, which shares porous borders and strong cross-border travel with The Gambia. This initial wave coincided with the start of the rainy season, leading to increased indoor gatherings and a pattern of household-driven transmission similar to that observed with other respiratory viruses (Sallah et al., 2023).

Second Wave (November 2020–February 2021): The Alpha (B.1.1.7), Eta, and B.1.1.420 lineages became predominant. The Gambia's epidemic curve closely tracked the appearance and spread of these variants in West Africa, with evidence for continued introductions from Europe and Asia. The period between the first and second waves saw a relative lull, likely reflecting both behaviour and partial effects of public health measures (Sallah et al., 2023).

Third Wave (July–October 2021): The Delta variant (AY.34.1 sub-lineage) surged, again aligning with the rainy season. Increased indoor activity during rains facilitated higher local transmission (Sallah et al., 2023).

Fourth Wave (December 2021–January 2022): Omicron BA.1.1 rapidly replaced circulating variants, leading to the steepest epidemic wave. This reflected the Omicron-driven surge seen globally (Sallah et al., 2023).

Seasonality and Transmission: Notably, the first and third waves fell during the rainy season (July–October), when indoor crowding likely amplified transmission risks. The convergence of environmental and behavioural factors with viral introductions played a pivotal role in shaping local outbreaks (Sallah et al., 2023).

1.2.2 Challenges in Surveillance and Healthcare Infrastructure

The progression of the pandemic in The Gambia was shaped by several critical challenges that directly impacted the availability, representativeness, and analytical validity of SARS-CoV-2 genomic and diagnostic data *(Kanteh et al., 2023)*:

Limited Testing and Genomic Capacity
Due to significant resource constraints, PCR testing in The Gambia was predominantly carried out in a limited number of accredited laboratories located mainly in urban centres, particularly within the capital region. Testing prioritised symptomatic individuals based on strict eligibility criteria, which resulted in two key biases within the collected data. First, there was a pronounced urban sampling bias, as most testing—and consequently most SARS-CoV-2 genomic sequencing and CT count data—originated from urban populations. This left rural and peri-urban communities substantially underrepresented, despite differences in population density, exposure risk, and healthcare access that may influence viral transmission dynamics and viral load measures. Second, there was a disease severity bias since testing predominantly targeted symptomatic or clinically significant cases, thus excluding many asymptomatic or mild infections from surveillance. This focus limited the ability to capture the full spectrum of viral genetic diversity and Ct distributions across different clinical presentations *(Kanteh et al., 2023)*.

Beyond these primary biases, the study also identified additional constraints affecting data collection and representativeness. Limited laboratory capacity and logistic challenges restricted sample throughput and the timeliness of testing, resulting in potential delays and sample degradation. These operational issues may have further influenced the quality and comparability of CT counts, a critical measure in associating viral genetics with viral load. Furthermore, incomplete contact tracing limited longitudinal follow-up of cases, thereby reducing opportunities to study within-host viral evolution or transmission-linked SNP patterns over time. Finally, socio-economic and behavioural factors including differences in healthcare-seeking behaviour by age, gender, and geographic location may have introduced further complexities in the dataset, as populations less likely to engage with formal healthcare services were inadequately sampled *(Kanteh et al., 2023)*.

Together, these biases and limitations underscore the challenges of generalising SNP–Ct association findings across The Gambia's diverse population and highlight the importance of accounting for sampling frameworks and underlying demographic factors in any genomic epidemiology analysis *(Kanteh et al., 2023)*.

Socio-Cultural Barriers to Accurate Surveillance

Beyond technical and infrastructural challenges, socio-cultural factors significantly hindered the accurate reporting and documentation of COVID-19 cases and deaths in The Gambia. Widespread stigma associated with COVID-19 infection led to fears of blame, ostracism within local communities, and, in some cases, loss of employment or social status. Individuals experiencing symptoms or exposed to the virus were sometimes reluctant to seek testing or report their condition, concerned that a confirmed diagnosis might result in isolation from family or exclusion from communal activities, which are central aspects of Gambian social life. Additionally, many Gambians rely on daily wage work, and periods of illness or mandatory quarantine could result in direct loss of income, further discouraging individuals from reporting symptoms or positive test results. In certain settings, misinformation about the virus fostered denial or minimized perceived risk, reducing public willingness to engage with health protocols. These socio-cultural and economic repercussions compounded existing health system limitations, contributing to underdocumentation and the incomplete surveillance of infection rates and outcomes within the country *(Kanteh et al., 2023; Salyer et al., 2021)*.

These efforts, although resource-constrained, provided important insights into the introduction, evolution, and spread of predominant lineages including Alpha, Delta, and Omicron in Gambia, and helped address a critical shortage of African genomic data needed for global variant surveillance and SNP-based association studies *(Kanteh et al., 2023)*.

1.3 Viral Load, Cycle Threshold (CT), and the Relevance of Genome-Wide Association in COVID-19

Viral load refers to the quantity of viral RNA present in a clinical specimen collected from an infected individual, most commonly from respiratory tract samples such as nasal or throat swabs. This viral RNA load provides an indirect measure of how much virus is present within the host at the time of sampling. Quantitative reverse transcriptase polymerase chain reaction (RT-PCR) assays are the gold standard diagnostic tools for detecting SARS-CoV-2 RNA. These assays amplify viral genetic material through successive cycles, producing a cycle threshold (Ct) value, which represents the number of amplification cycles required to reach a detectable level of viral RNA (Puhach et al., 2023; Shenoy et al., 2021).

Importantly, CT counts are inversely proportional to viral load—the lower the CT count, the higher the concentration of viral RNA in the sample. Because RT-PCR quantifies viral RNA rather than infectious virus particles directly, CT counts serve as a proxy measure of viral load rather than a direct count of viable virus. While viral load measured by Ct correlates broadly with infectious virus presence and potential transmissibility, it does not distinguish between live, replication-competent virus and non-infectious RNA fragments (Puhach et al., 2023). Despite this limitation, CT counts remain a widely used and practical surrogate biomarker for estimating viral burden within patients.

Viral load plays a critical role in several aspects of SARS-CoV-2 infection dynamics and patient outcomes. Higher viral loads, indicated by low cycle threshold (Ct) values, generally correlate with increased infectiousness. Individuals with greater amounts of virus shed more viral particles, raising the likelihood of transmission to others in close contact. Goyal et al. (2021) demonstrated this relationship through a within-host mathematical model simulating viral load kinetics based on patient data. Their study linked viral RNA levels, approximated by CT counts, to transmission probability, showing that viral loads above a certain threshold markedly increase infectiousness. This model underscores the importance of viral load variation in driving transmission dynamics and superspreading events within communities.

Elevated viral loads are also associated with more severe disease manifestations and poorer clinical outcomes. Strong viral replication can amplify immune responses and cause tissue damage, resulting in severe respiratory symptoms and complications. Shenoy et al. (2021) conducted a systematic review analysing quantitative viral load data from multiple studies, finding that lower CT counts correlate with increased disease severity and mortality. Conversely, lower viral loads typically correspond to milder or asymptomatic infections. Monitoring viral load over time, often through serial Ct measurements, can further inform disease prognosis and progression, aiding clinical management decisions.

Genome-wide association studies (GWAS) identify statistical links between genetic variations, particularly single nucleotide polymorphisms (SNPs), and phenotypic traits by systematically scanning the genome across many samples. In viral genomics, GWAS analyses genotype data encoded as SNP presence or absence and correlates these with quantitative traits such as cycle threshold (Ct) values through linear regression models that adjust for confounders like viral lineage or sampling date. Each SNP is tested independently, producing effect sizes and p-values, with stringent multiple testing corrections applied to minimize false positives. This approach enables discovery of genetic variants or mutational patterns that influence viral load, transmissibility, or disease severity. For example, a GWAS on *Mycobacterium tuberculosis* (Naz et al., 2023) identified SNPs linked to drug resistance, while another on *Plasmodium falciparum* (Manske et al, 2012) revealed mutations associated with immune evasion. These studies demonstrate GWAS's power to elucidate complex genotype-phenotype relationships in pathogens, making it a valuable tool for investigating how SARS-CoV-2 genetic diversity affects CT counts and viral dynamics.

2. Aim and Objectives

Primary Objective

The primary objective of this study is to investigate the relationship between viral genetic variation and viral load in SARS-CoV-2 infections from The Gambia. Specifically, the study aims to evaluate whether variation in viral load, using RT-PCR cycle threshold (Ct) values as a proxy, is influenced predominantly by single nucleotide polymorphisms (SNPs) within the viral genome.

Specific Objectives

1. To compile and analyse a dataset of over 1,600 whole-genome SARS-CoV-2 sequences from The Gambia, each with matched CT counts.
2. To identify genetic variants (SNPs) that are significantly associated with differences in CT counts across infected individuals.
3. To identify changes in Amino acids and if that impacts protein formation

Hypothesis

Null Hypothesis (H0):
There is no association between viral genetic variation (SNPs) within SARS-CoV-2 genomes and viral load as measured by RT-PCR cycle threshold (Ct) values in the Gambian cohort.
This means that none of the SNP loci genotyped in this dataset have a statistically significant effect on CT counts, once appropriate quality control and confounder adjustments have been made

Alternative Hypothesis (H1):
At least one viral genetic variant (SNP) within SARS-CoV-2 genomes is significantly associated with variation in viral load as measured by RT-PCR cycle threshold (Ct) values in the Gambian cohort. This asserts that there exists at least one SNP for which the genotype has a statistically significant effect on the CT count (after quality control and adjustment for confounders).

3. Materials and Methods

The flow chart (Figure 1) describes the workflow for the project. We began by assigning standardized lineages to all sequences in the dataset using Pangolin, a tool for phylogenetic lineage assignment. These sequences were then split into two groups: All Lineages and Top 10 Lineages, which contains the 10 most prevalent lineages within the dataset. Both groups of genomes were aligned, and GWAS was conducted to identify associations between genetic variants (SNPs) and the phenotype (CT count).

For All Lineages, a phylogenetic tree was first constructed, followed by simple linear regression. Using the PCA results for all lineages, three models were developed to assess the impact of subpopulations identified in the PCA: a primary model with all samples, a sensitivity model excluding outliers, and a stratified model removing the most divergent lineage. For Top 10 Lineages, we conducted lineage-specific simple linear regression analyses. Results were visualised using Manhattan plots to show SNP-phenotype associations, QQ plots to compare expected versus observed results under the null hypothesis, and PCA plots to identify clustering related to population structure.

.

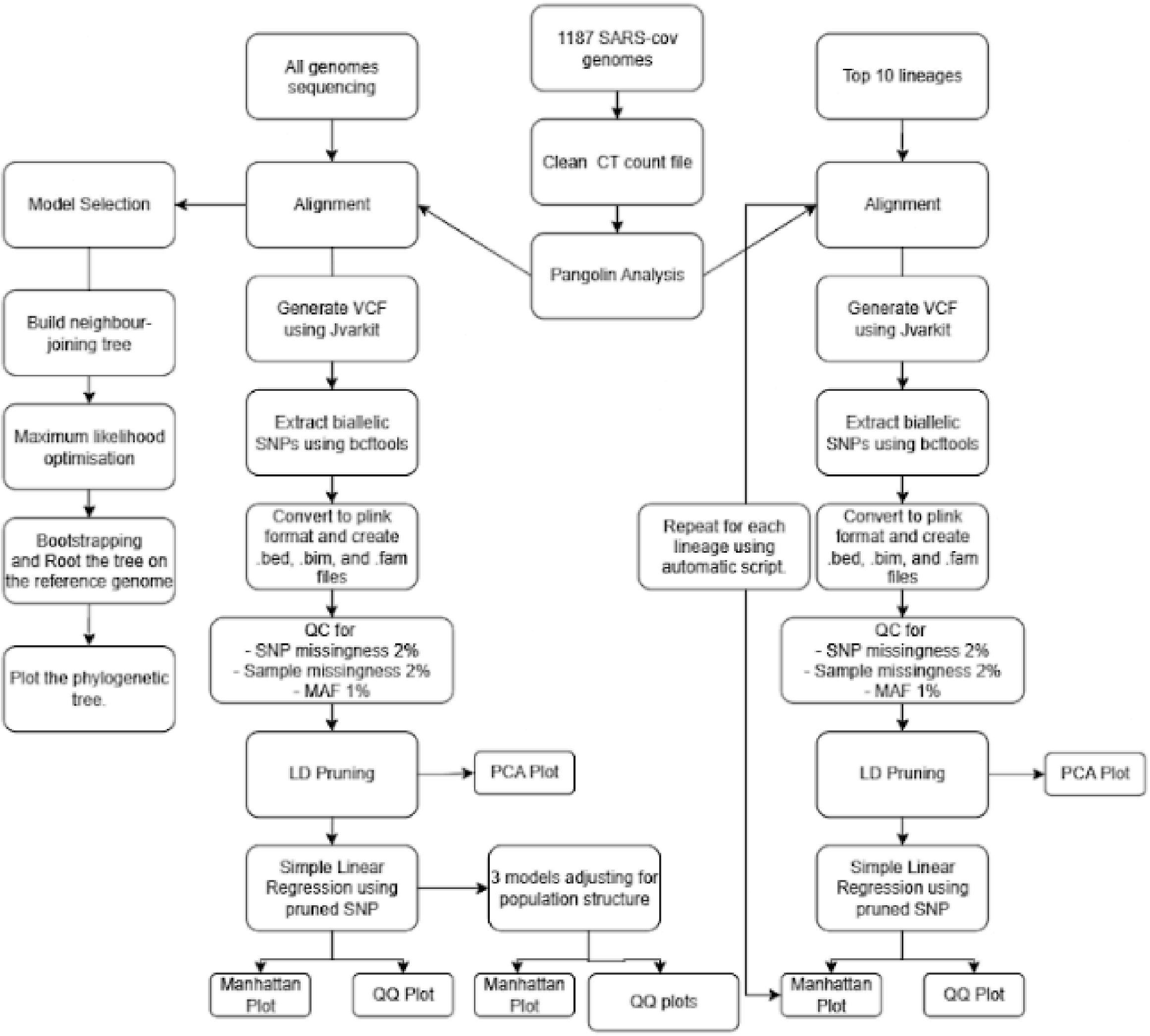

1187 SARS-cov genomes
Clean CT count file
Pangolin Analysis
All genomes sequencing
Alignment
Model Selection
Build neighbour-joining tree
Maximum likelihood optimisation
Bootstrapping and Root the tree on the reference genome
Plot the phylogenetic tree.
Generate VCF using Jvarkit
Extract biallelic SNPs using bcftools
Convert to plink format and create .bed, .bim, and .fam files
QC for
- SNP missingness 2%
- Sample missingness 2%
- MAF 1%
LD Pruning
PCA Plot
Simple Linear Regression using pruned SNP
3 models adjusting for population structure
Manhattan Plot
QQ Plot
Manhattan Plot
QQ plots
Top 10 lineages
Alignment
Generate VCF using Jvarkit
Extract biallelic SNPs using bcftools
Repeat for each lineage using automatic script.
Convert to plink format and create .bed, .bim, and .fam files
QC for
- SNP missingness 2%
- Sample missingness 2%
- MAF 1%
LD Pruning
PCA Plot
Simple Linear Regression using pruned SNP
Manhattan Plot
QQ Plot

Figure 1. Analytical workflow for genome-wide association study (GWAS) of SARSCoV-2 viral load using cycle threshold (Ct) values.
The flowchart outlines the process for lineage assignment, genome alignment, SNP extraction, quality control, phylogenetic analysis, and association testing to identify relationships between viral genetic variants and Ct-based viral load measures in Gambian samples. Results are visualized with principal component analysis, Manhattan, and QQ plots for both all lineages and the ten most prevalent lineages in the dataset.

3.1 Data Processing

This section outlines the initial pre-processing of the CT count dataset, including cleaning and formatting, followed by the assignment of SARS-CoV-2 lineages. The most prevalent top ten lineages were extracted for further study.

3.1.1 Data Cleaning CT count Excel sheet

Both the 'screening' and 'confirmatory' columns in the CT (cycle threshold) dataset were converted to numeric data types. Missing values in the screening column were removed, resulting in the exclusion of 424 rows. The screening CT count was selected for further analysis, as it most closely represents the viral load at the time of initial sampling. The dataset was then examined and cleaned for extraneous white spaces to ensure accurate matching with genome names in subsequent FASTA files.

For conversion to PLINK format, both Family ID (FID) and Individual ID (IID) columns were required. As FID was absent, a new FID column containing all zeros was created. The 'FID', 'IID', and 'ct_value_screening' columns were then extracted to generate the 'plink.txt' file.

Merge checks were performed for each lineage, and in all cases, fewer than 20 samples per lineage were excluded due to missing CT counts encountered during data cleaning. For example, for the AY.34.1 lineage, 24 samples were identified in the .FAM file but were absent from 'plink.txt' because they had previously been excluded for missing CT counts.

3.1.2 Assigning Lineages and Extraction of Top 10 Genomes

To assign SARS-CoV-2 lineages, the initial large dataset was divided into ten batches of approximately equal size. Each batch was processed individually using the Pangolin software to generate lineage designations, and output files from all batches were merged to create a combined lineage assignment dataset. The distribution of lineages was plotted using a bar plot in R to identify the most prevalent lineages. The ten most frequently occurring lineages were selected for further analysis. Genome identifiers corresponding to each of these top ten lineages were extracted from the combined dataset and saved into separate text files, which were subsequently used to filter and generate a distinct FASTA file for each lineage.

3.1.3 Multiple Sequence Alignment

The reference genome (NC_045512.2) was added to each lineage-specific FASTA file to serve as a standard for alignment and comparison. Unidentified nucleotides represented by 'N' were replaced with '-' to indicate missing nucleotides. Multiple sequence alignment was then performed for each lineage using MAFFT with the --auto setting, which automatically selects optimal alignment parameters based on the characteristics of the sequences.

3.1.4 Phylogenetic Tree

Phylogenetic analysis was conducted using aligned SARS-CoV-2 genome sequences from the Gambian dataset alongside lineage assignments derived from Pangolin.

Data Preparation and Labelling:

Aligned genome sequences in FASTA format (gambia_only_aligned.fasta) were imported using the readDNAStringSet function from the Biostrings R package. The Pangolin lineage assignments were read from a CSV file (lineage_report.csv) with the read_csv function. Sequence headers in the FASTA file were replaced with combined lineage and taxon identifiers, embedding lineage information within the sequence names. Sequences lacking lineage data retained their original headers. The relabelled sequences were saved as a new FASTA file (gambia_only_aligned_pangolin.fasta) for subsequent analysis.

Model Selection:
The relabeled FASTA file was imported using the read.dna function from the ape package with the format specified as FASTA. The sequences were then converted into a phyDat object, which is suitable for downstream phylogenetic analyses. Model selection was performed with the modelTest function from the phangorn package, comparing the Jukes-Cantor (JC69) and Felsenstein 1981 (F81) substitution models. The model with the lowest corrected Akaike Information Criterion (AICc) value was selected. Although additional substitution models were evaluated, they were not supported by the dist.ml function; therefore, only the models compatible with this function were tested.

Tree Construction:
A distance matrix was computed under maximum likelihood estimation according to the chosen substitution model (dist.ml). This matrix was used to infer an initial neighbor-joining (NJ) tree with the NJ function. Any negative branch lengths were adjusted to zero to maintain biological plausibility. The NJ tree was further optimised using maximum likelihood via the pml and optim.pml functions, applying stochastic subtree pruning and regrafting (SPR) for topology refinement. Optimisation was performed with iterative likelihood improvement and tracing enabled to monitor convergence.

Bootstrap Analysis:
Node support was assessed by bootstrap resampling with 30 replicates (bootstrap.pml), using nearest neighbour interchange (NNI) optimisation during resampling for improved accuracy. A fixed random seed (set.seed(123)) was applied to ensure reproducibility.
The ML-optimised tree was rooted using the Wuhan-Hu-1 reference genome (accession B|hCoV-19/Wuhan/WH01/2019|EPI_ISL_406798|2019-12-26) as an outgroup via the root function, with multifurcations resolved for clarity. The rooted tree was saved in Newick format (gambia_ml_tree_pangolin_rooted_tree.nwk).

Visualisation:
Phylogenetic trees were visualised using the ggtree package, integrating a metadata dataframe linking tip labels to Pangolin lineages by parsing lineage from sequence headers. Tip labels were colour-coded by lineage and internal nodes annotated with bootstrap values. The tree image was exported as a high-resolution PNG file (gambia_ml_tree_pangolin_rooted_tree.png).

3.2 GWAS analysis

3.2.1 Variant Calling

Following multiple sequence alignment, variant calling was performed for both the top 10 lineages and all lineages using Jvarkit, a utility suite for genomic data processing. This generated a Variant Call Format (VCF) file that included key columns describing each detected polymorphism:

CHROM: Designates the chromosome (or contig) where the variant is located.

POS: Specifies the genomic position of the variant relative to the reference genome.

ALT: Lists the alternative nucleotide(s) observed amongst the sequences at that position.

QUAL: Provides a Phred-scaled confidence score for each variant, reflecting the probability of an incorrect call −10×log10(P) (P is the probability that the variant call is incorrect).

FILTER: Indicates any variant-level quality filters that were not passed.

INFO: Offers additional annotation, such as allele frequency, sequencing depth, or other variant metrics.

The resulting VCF file contained multiallelic variants, meaning multiple alternative alleles were observed at some loci. In the context of viral genomes—particularly for haploid viruses like SARS-CoV-2—the term “alleles” refers to variant frequencies observed in the population rather than individual virions. While viral populations commonly acquire multiple mutations through error-prone replication, it remains rare for distinct changes to become highly prevalent at the *exact* same nucleotide position.

To ensure compatibility with downstream analyses in PLINK, which only accepts biallelic variants, all non-biallelic (multiallelic) sites were removed using the bcftools . This step extracted only those sites featuring a single alternative allele, improving compatibility without substantially impacting interpretation, as single-position multiallelic variation is uncommon in viral datasets.

3.2.3 Conversion to PLINK Format

The filtered VCF file containing only biallelic variants (.vcf.gz) was converted to PLINK’s binary format, producing three files: .bed , .bim , and .fam . This conversion enables genotype data to be efficiently stored and analysed for genome-wide association studies (GWAS).

The .bed file stores the genotype calls for each sample at every variant site in a compact binary format.

The .bim file serves as a variant metadata index. In viral datasets, the chromosome field simply contains the reference genome name since the SARS-CoV-2 genome is single-chromosome (~30,000 nucleotides). The unique SNP identifier is generated by combining the reference genome and the SNP position. The genetic distance column, which in human data reflects recombination probability, is set to zero as recombination mapping is not relevant in haploid viruses. The .bim file also includes the position of each variant (in base pairs along the genome) and enumerates the reference and alternate nucleotides. Al represents the effect Allele; it is the allele on which association statistics are reported.

Chromosome (CHR)	SNP ID (SNP)	Genetic distance (CM)	Base-pair position (BP)	Allele 1 (A1)	Allele 2 (A2)
NC_045512.2	NC_045512.2:1006	0	1006	T	G
NC_045512.2	NC_045512.2:4012	0	4012	T	C
NC_045512.2	NC_045512.2:4399	0	4399	T	G

Table 1 - First 3 rows form the AY.34.1_aligned_auto.QC.unique.bim file

The .fam file records sample metadata. While its six columns typically describe family relationships, sex, and phenotype in human GWAS, for viral data only the individual ID is relevant—this ID is a combination of the Pangolin lineage and the original genome name. All other metadata fields are either set to defaults or marked as missing.

Family ID (FID)	Individual ID (IID)	Paternal ID	Maternal ID	Sex	Phenotype
0	134101	0	0	0	-9
0	134115	0	0	0	-9
0	135100	0	0	0	-9

Table 2 - First 3 rows form the AY.34.1_aligned_auto.QC.unique.fam file

For the all lineages work flow The VCF and the relevant Plink files were generated using the same workflow mentioned for the top 10 lineages. However, To conduct the GWAS for associations between SARS-CoV-2 genetic variants and CT counts, I first prepared genotype data compatible with phenotype sample identifiers. Because Pangolin lineage assignments appended lineage labels to sample IDs in the genotype .fam file (formatted as lineage|sampleID , e.g. AY.34|60685), while phenotype Ct count data contained sample IDs without lineage prefixes (e.g. 60685), a new set of PLINK input files was generated to ensure matching sample identifiers.

A new .fam file was created by extracting the sample ID portion after the pipe symbol from the original .fam file, and replacing each entry with the appropriate PLINK .fam format fields. The corresponding .bed and .bim files were duplicated to maintain consistency.

```
awk 'BEGIN{OFS="\t"} {split($2,a,"|"); print 0, a[2], 0, 0, 0, -9}'
all_lineages.QC.unique.fam > all_lineages.QC.unique.GWAS_prep.fam cp
all_lineages.QC.unique.bed all_lineages.QC.unique.GWAS_prep.bed cp
all_lineages.QC.unique.bim all_lineages.QC.unique.GWAS_prep.bim
```

3.2.4 Quality Control (QC) and Confounder Adjustment

Quality Control

Quality control procedures were applied to ensure that the resulting genetic association analyses were not confused by technical artifacts or poor-quality data. Specific filters included:

SNP missingness (<2%): Single nucleotide polymorphisms (SNPs) for which genotypes could not be reliably determined in more than 2% of samples were excluded. This threshold helps minimize false associations due to technical errors during sequencing or genotyping.

Sample missingness (<2%): Samples (genomes) missing more than 2% of genotype calls across all sites were removed, helping to avoid bias from incomplete or low-quality data.

Minimum Allele Frequency (MAF >0.01): The minimum allele frequency threshold was set at 0.01 (i.e., at least 1% of the genomes contained the minor allele). Although a MAF threshold of 0.05 is standard, a lower value was adopted here due to limited sample size as a cutoff of 0.05 would exclude many variants present in the dataset.

MAF filtering helps exclude very rare SNPs, which may introduce false positives or reduce the ability to detect true genetic associations because there are too few carriers in the dataset.

LD Pruning

Linkage disequilibrium (LD) pruning was performed to reduce redundancy among genetic variants and ensure independence among SNPs used in the association analysis. In viral genomes, high LD can occur because mutations are inherited together during replication. LD pruning removes SNPs that are highly correlated ($r^2 > 0.2$) within overlapping genomic windows (window size: 50 SNPs; step size: 5 SNPs). This reduces the chance that multiple correlated SNPs, representing the same mutation event or haplotype, are included in downstream analyses, which could otherwise confound statistical tests.

Window size: 50 SNPs per window.

Step size: Advances the window by 5 SNPs each time, creating overlap.

LD threshold (r^2): For each window, pairs of SNPs with $r^2 > 0.2$ were identified; one of each highly correlated pair was removed.

3.2.5 Linear regression models

CT Count Data Quality Assessment and Normality Testing

CT count data underwent systematic quality control and normality assessment before genetic association analysis. Only samples that passed linkage disequilibrium pruning and minor allele frequency filtering were included in the normality assessment to ensure consistency with the final analytical dataset. Lineage covariate data and CT screening values were merged by sample identifier, retaining only samples present in both datasets that met genetic quality control criteria. Biologically implausible CT counts of zero were identified and removed, as these represent technical failures rather than true viral load measurements.

Descriptive statistics, including mean, median, standard deviation, skewness, and kurtosis, were calculated for the cleaned CT counts from quality-controlled samples. Normality was assessed through multiple approaches: visual inspection using histograms with normal distribution overlays, quantile-quantile plots comparing observed values to theoretical normal quantiles, and formal statistical tests including Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests. CT count distributions were examined both overall and stratified by viral lineage using boxplots to identify lineage-specific distributional differences. Skewness values between -0.5 and 0.5 with kurtosis values between 2 and 4 were considered indicative of approximate normality, supporting the use of linear regression models. This comprehensive assessment ensured that phenotype distributions from the final analytical sample met the underlying assumptions required for valid genome-wide association testing and accurate interpretation of genomic inflation factors.

3.2.5 Principal Component Analysis and Population Structure Assessment

Principal component analysis was conducted using eigenvector output from standard genomic PCA workflows, with sample lineage information systematically extracted from sample identifiers using standardized parsing to ensure consistent classification across all samples. A critical analytical decision was made to define population structure relative to an empirically-derived center point, calculated as the overall mean of PC1 and PC2 coordinates across all samples, providing an unbiased reference point independent of any specific lineage characteristics. From this center, we computed Euclidean distances for each sample using the formula $\sqrt{[(PC1 - mean\_PC1)^2 + (PC2 - mean\_PC2)^2]}$, creating the distance_from_center variable as a continuous metric of genetic deviation from the population mean. Outliers were defined using a conservative statistical threshold approach where samples were classified as outliers if their distance from the population center exceeded the mean distance plus two standard deviations (mean + 2SD), a criterion designed to capture approximately 5% of samples under normal distribution assumptions while maintaining statistical rigor. To comprehensively assess lineage-specific population structure, we calculated multiple statistics including lineage centroid coordinates as the mean PC1 and PC2 values for each lineage, standard deviations of PC1 and PC2 within each lineage to measure internal genetic diversity, and sample counts per lineage to assess representation. The distance_centroid_from_origin variable was computed to quantify how far each lineage centroid lies from the overall population center, providing a measure of lineage-specific genetic deviation. We systematically calculated pairwise Euclidean distances between all lineage centroids to construct a complete distance matrix quantifying genetic separation between different lineages, identifying both the most genetically distant and most similar lineage pairs. Additional population structure metrics included the mean distance from center for samples within each lineage, maximum distance from center within each lineage, the number and percentage of outlier samples per lineage, and the overall percentage of outliers in the dataset. These comprehensive statistics enabled systematic identification of which lineages deviate most significantly from the population mean, assessment of within-lineage genetic diversity, and quantification of between-lineage genetic distances.

Data were prepared for three complementary regression models through systematic covariate construction and exclusion list generation, each addressing different aspects of population stratification and potential confounding. The primary analysis model includes all samples with both A.29 lineage indicator and principal components as covariates, designed to capture both discrete lineage effects and continuous population structure through comprehensive covariate inclusion that maintains maximum statistical power. The sensitivity analysis model excludes individual genetic outliers identified through the statistical threshold approach while retaining all lineages and covariates, specifically testing whether results are driven by extreme genetic variants that might not represent typical lineage characteristics. The stratified analysis model excludes all A.29 samples and uses only principal components as covariates, testing whether population structure effects persist in the absence of the potentially most divergent lineage and providing evidence for lineage-specific versus general population structure effects. The A29 indicator variable was created as a binary numeric variable coded as 1 for A.29 lineage samples and 0 for all other lineages, enabling direct coefficient interpretation in regression models and allowing quantification of A.29-specific effects relative to all other lineages combined. Principal components were retained in their original continuous form to preserve maximum information about population structure, while sample identifiers were systematically parsed to ensure compatibility with standard genetic analysis software, with family IDs set to 0 to indicate unrelated samples and individual IDs extracted from composite identifiers. Rather than modifying the primary dataset, formal exclusion lists were generated containing sample identifiers in standard format for the sensitivity and stratified models, preserving complete data integrity and allowing exact replication of analytical decisions. An alternative covariate file containing only the A29 indicator variable was also prepared to avoid potential multicollinearity issues between the lineage indicator and principal components in certain analytical contexts. This three-model analytical framework was specifically designed to maximize statistical power while systematically testing key assumptions about population structure, with sample sizes carefully preserved across models: the primary model retaining the full sample size for maximum power, the sensitivity model removing only statistical outliers (typically less than 5% of samples based on the 2SD threshold), and the stratified model removing only A.29 samples while preserving substantial power for testing population structure effects in the remaining lineages, thereby enabling robust assessment of result consistency across different analytical conditions and assumptions about population stratification.

Potential Limitations and Mitigation Strategies

The analysis framework acknowledges that principal components may not capture all relevant population structure, particularly rare or recent admixture events. However, PC1 and PC2 typically explain the largest components of genetic variation and provide the most stable basis for population stratification control. The decision to focus on A.29 as a specific lineage of interest was made based on preliminary analysis suggesting it represents the most genetically divergent lineage, but the analytical framework is designed to be robust to this choice through the multi-model comparison strategy.

Top 10 Lineage PCA

Principal Component Analysis was performed individually for each of the top 10 most prevalent SARS-CoV-2 lineages using eigenvector files generated through standard genomic PCA workflows using an automatic script. For each lineage, the corresponding eigenvector file (.PCA.eigenvec) was processed to extract the first two principal components (PC1 and PC2) from columns 3 and 4 respectively. PCA plots were generated using R scripting with systematic visualization parameters including blue circular points (pch=19) at

70% size (cex=0.7) to represent individual samples within each lineage's principal component space.

3.2.6 Simple Linear Regression

Using the prepared genotype files, linkage disequilibrium (LD)-pruned variants were selected by extracting the samples listed in the all_lineages.QC.unique.LDpruned.prune.in file. The phenotype file specifying cycle threshold (Ct) values (ct_phenotype_plink.txt) was provided in PLINK format.
Genome-wide association studies (GWAS) were performed with PLINK2 utilizing a simple linear regression model without covariates. Specifically, for both datasets (all lineages combined and the top 10 lineages subset), a linear regression model was fitted to evaluate the association between each single nucleotide polymorphism (SNP) and the CT count phenotype.
This analysis produced a results file with the extension .linear containing several key columns, including SNP identifiers, effect sizes (beta coefficients), standard errors, test statistics, and p-values, which together describe the strength and significance of the association between viral genetic variants and viral load as estimated by Ct.

CHROM: The chromosome or reference genome name, which in this study is NC_045512.2 .

POS: The genomic position of the SNP on the reference genome.

REF: The nucleotide present in the reference genome at the given position.

ALT: The alternative nucleotide (mutation) observed in the sample set.

A1_FREQ: The frequency of the allele coded as ‘A1’ within the lineage, representing its prevalence in the dataset.

BETA: The estimated effect size of the SNP on CT count. A positive BETA value indicates that the SNP is associated with an increase in CT count, while a negative value indicates a decrease.

P-value: The statistical significance of the association between the SNP and CT count, representing the probability that the observed association is due to random chance.

To account for multiple testing, a significance threshold was calculated using a Bonferroni correction: 0.05 divided by the number of SNPs tested (n). SNP associations with p-values below this threshold were considered statistically significant.

QQ Plot

Simple linear regression analysis results were processed using systematic quality control and statistical assessment procedures designed to evaluate the distribution of association p-values and assess potential systematic bias or population stratification effects. The GWAS output file containing linear regression results was processed with stringent filtering criteria, retaining only variants with valid p-values between 0 and 1 while excluding missing values and boundary cases that could introduce artifacts in downstream statistical calculations. This conservative filtering approach ensures that the QQ plot analysis reflects only legitimate statistical tests and removes potential computational errors or extreme values that could distort the assessment of p-value distribution patterns.

Quantile-quantile plot construction followed standard genomic analysis protocols with expected $-\log_{10}(p)$ values calculated using the rank-based formula $-\log_{10}(\text{rank}/(n+1))$, where rank represents the ordered position of each p-value and n represents the total number of valid variants, providing theoretical quantiles under the null hypothesis of no association. Observed $-\log_{10}(p)$ values were computed directly from the association p-values after sorting in ascending order to enable direct comparison against expected values. The genomic inflation factor lambda was calculated as the ratio of the median observed chi-square statistic to the theoretical median chi-square value under the null hypothesis, where chi-square statistics were derived from p-values using the inverse chi-square transformation qchisq(1-P, df=1). This lambda calculation provides a quantitative measure of systematic inflation in test statistics, with values near 1.0 indicating appropriate control of population structure and systematic effects, while values substantially greater than 1.0 suggest the presence of confounding factors such as population stratification, relatedness, or technical artifacts that inflate association signals across the genome. Summary statistics included total variant counts after quality control filtering, median expected and observed $-\log_{10}(p)$ values for central tendency assessment, and identification of variants meeting genome-wide significance thresholds ($p < 5\times10^{-8}$) to assess the presence of strong association signals that exceed conventional statistical significance criteria for genome-wide association studies.

Manhattan Plot

Manhattan plot visualisation for the primary simple linear regression analysis was implemented using systematic data processing and statistical threshold determination to enable comprehensive assessment of association patterns across the SARS-CoV-2 genome. GWAS output files were processed with stringent quality control filtering, retaining only variants with valid p-values between 0 and 1 while excluding missing values and ensuring valid genomic positions for accurate spatial representation of associations. The $-\log_{10}(P)$ transformation was applied to all valid p-values to enable standard genomic visualization approaches, with genomic positions extracted from the POS column and converted to numeric format for proper spatial ordering along the viral genome. A critical analytical decision was made to implement Bonferroni correction thresholds calculated specifically for the number of variants tested, using the formula $-\log_{10}(0.05/\text{N\_SNPs})$ to establish conservative significance criteria that account for multiple testing burden. This adaptive threshold approach ensures appropriate statistical rigor while avoiding overly conservative corrections that might obscure genuine associations in smaller genomic datasets. Manhattan plot construction included comprehensive visual elements with all variants plotted as position versus $-\log_{10}(P)$ values, Bonferroni significance thresholds displayed as horizontal reference lines, and variants exceeding the threshold highlighted with distinct coloring to enable immediate identification of statistically significant associations. Summary statistics were systematically compiled including total variant counts after quality control, Bonferroni threshold values in both $-\log_{10}(P)$ and p-value formats, and comprehensive identification of significant variants with positional and statistical details for the top associations.

Lineage-Specific Analysis Framework

Individual lineage analyses were conducted for eight major SARS-CoV-2 lineages using standardized processing workflows designed to enable systematic comparison of association patterns across different viral genetic backgrounds while controlling for lineage-specific characteristics that could influence statistical power and association detection. Each lineage dataset underwent identical quality control procedures with filtering for valid p-values, removal of variants with zero allele frequency to eliminate uninformative markers that could introduce statistical artifacts, and extraction of valid genomic positions for proper spatial analysis. This zero allele frequency filtering represents a critical analytical decision that ensures all analyzed variants contribute meaningful genetic information while removing monomorphic sites that provide no association signal but could dilute statistical power. For each lineage, both quantile-quantile plots and Manhattan plots were constructed using lineage-specific parameter calculations, including individual Bonferroni correction thresholds based on the number of variants tested within each lineage dataset, lineage-specific genomic inflation factor calculations to assess population structure control within each genetic background, and identification of significant associations meeting lineage-appropriate statistical criteria. The genomic inflation factor lambda was calculated individually for each lineage using the standard median chi-square ratio approach, enabling assessment of whether population structure control remains effective within more homogeneous genetic backgrounds compared to the overall analysis. QQ plots were constructed with expected quantiles calculated using lineage-specific sample sizes and variant numbers, while Manhattan plots included lineage-specific Bonferroni thresholds and highlighted significant associations according to each lineage's multiple testing burden. This systematic approach enables direct comparison of association strength, population structure control effectiveness, and statistical power across different viral lineages while maintaining appropriate statistical rigor for each individual analysis. Visualization outputs were organized using grid layouts with four plots per panel to enable efficient comparison across lineages, with each plot labeled for clear identification and summary statistics including lambda values and significant variant counts displayed for rapid assessment of analysis quality and association patterns within each lineage context.

3.2.7 A Three-Model Framework for GWAS Analysis with Adaptive Significance Thresholds and Population Structure Control

Genome-wide association study results from all three regression models were systematically processed and visualized using comprehensive statistical assessment approaches designed to evaluate both association significance and population structure control effectiveness. GWAS output files were processed with standardized filtering criteria, retaining only additive genetic effects (ADD test) to focus on actual SNP associations while removing rows with missing or zero p-values that could introduce artifacts in downstream visualization and statistical calculations. A critical analytical decision was made to implement adaptive Bonferroni correction thresholds calculated individually for each model using the formula 0.05/N_SNPs, where N_SNPs represents the total number of valid SNPs tested in each specific model, ensuring appropriate multiple testing correction that accounts for the actual number of statistical tests performed rather than applying arbitrary universal thresholds. This model-specific approach addresses the reality that different sample exclusion criteria across the three models may result in varying numbers of analyzable SNPs, requiring correspondingly adjusted significance thresholds to maintain statistical rigor. For each model, we calculated $-\log_{10}(P)$ values to enable standard GWAS visualization approaches, computed minimum and maximum p-values to assess the dynamic range of association signals, identified the number of SNPs meeting Bonferroni significance criteria, and systematically tracked both the adaptive threshold values and the corresponding $-\log_{10}(P)$ equivalents for visualization purposes.

Manhattan plots were constructed for individual models and in combined format using position-specific plotting against -$\log_{10}$(P) values, with model-specific adaptive Bonferroni thresholds displayed as horizontal reference lines color-coded to match each model's data points, enabling direct visual assessment of which associations exceed the appropriate significance threshold for each analytical approach. Quantile-quantile plots were generated to assess deviation from the null hypothesis of no association, with expected -$\log_{10}$(P) values calculated using the points function to generate theoretical quantiles for comparison against observed values sorted in ascending order. The genomic inflation factor lambda was computed for each model using the median chi-square statistic divided by the theoretical median under the null hypothesis, providing a quantitative measure of population structure control effectiveness where values near 1.0 indicate appropriate control and values substantially greater than 1.0 suggest residual population stratification that could inflate association signals. QQ plots included both individual model assessments and combined visualization with model-specific expected value calculations to account for different sample sizes and SNP numbers across the three analytical approaches. Comprehensive summary statistics were systematically compiled including the number of SNPs tested per model, model-specific Bonferroni thresholds with scientific notation formatting, minimum and maximum p-values, median p-values for central tendency assessment, lambda values for population structure evaluation, and counts of SNPs meeting Bonferroni significance criteria. Top association hits were identified for each model through systematic ranking by p-value significance, with explicit classification of each association as either meeting or failing to meet the model-specific Bonferroni threshold, enabling direct comparison of association strength and significance consistency across the three analytical approaches and providing evidence for the robustness of identified signals across different population structure control strategies.

3.3 Connect SNP to genes

3.3.2 SNP-to-Gene Mapping and Functional Annotation

The annotated SARS-CoV-2 reference genome file (R_Genes_NC_045512.2.GFF3) was downloaded directly from the GenBank database, containing comprehensive gene feature annotations including genomic coordinates of coding sequences (CDS) used to map single nucleotide polymorphisms identified in association analyses to specific viral genes. Significant SNPs were systematically filtered from both pan-lineage and lineage-specific linear regression results based on statistical significance thresholds, with the pan-lineage analysis using Bonferroni-corrected criteria ($P < 0.0007246$) and lineage-specific analyses using more permissive thresholds ($P < 0.01$) combined with allele frequency filtering (A1_FREQ > 0) to eliminate uninformative monomorphic variants that could introduce analytical artifacts. A critical analytical decision was made to convert SNP positions into BED format intervals with mutation start positions defined as one base pair upstream of the SNP position (POS - 1) and end positions equal to the SNP position, enabling precise spatial overlap detection with gene boundaries while accounting for standard genomic coordinate conventions. CDS features were systematically extracted from the GFF3 annotation file using awk processing and converted into BED-formatted coordinates specifying chromosome, start, end, and comprehensive gene annotations including gene names, protein products, and protein identifiers. Location-based mapping of SNPs to genes was performed using bedtools to identify overlaps between SNP mutation intervals and CDS regions, producing comprehensive linkage files connecting variants to their corresponding genes and enabling systematic functional annotation across all lineages analyzed.

3.3.3 Amino Acid Change Analysis and Functional Impact Assessment

For SNPs located within coding sequences, comprehensive amino acid change analysis was conducted through systematic sequence extraction, mutation simulation, and protein translation workflows designed to assess the functional consequences of identified associations. The SARS-CoV-2 reference genome sequence (NC_045512.2) was downloaded from NCBI to provide the template DNA sequence for functional analysis, with gene-specific sequences extracted based on CDS coordinates identified through the mapping procedure. A systematic mutation analysis pipeline was implemented where reference DNA sequences were modified to incorporate the alternate allele at the identified SNP position, followed by translation of both reference and mutated sequences using the standard genetic code to identify amino acid changes. Critical analytical parameters included calculation of codon position within genes, determination of codon numbers affected by each mutation, and assessment of whether nucleotide changes resulted in synonymous or non-synonymous amino acid substitutions. For non-synonymous changes, comprehensive functional impact assessment was performed through systematic classification of amino acid properties including hydrophobicity, polarity, charge, and aromaticity characteristics of both reference and alternate amino acids. Conservative changes were defined as amino acid substitutions maintaining similar physicochemical properties, while non-conservative changes involved substitutions between amino acids with different property profiles, providing a framework for predicting potential functional consequences. The analysis workflow included generation of protein context sequences surrounding mutation sites to assess local structural implications, creation of reference and alternate protein sequence files for downstream analysis, and systematic compilation of mutation summaries including SNP positions, nucleotide changes, codon numbers, amino acid changes, and functional impact classifications. This comprehensive approach enables systematic assessment of whether identified genetic associations correspond to functionally relevant protein-altering mutations that could contribute to phenotypic differences across viral lineages or represent neutral variants that may tag causal mutations through linkage disequilibrium.

Results

4.1 Descriptive Analysis

Understanding the evolutionary relationships among SARS-CoV-2 genomes is critical for elucidating local transmission dynamics and viral diversification patterns. Phylogenetic analysis provides a framework to identify dominant lineages, track the introduction and spread of variants, and infer evolutionary pressures shaping viral populations. Here, we present a maximum likelihood phylogenetic tree constructed from Gambian SARS-CoV-2 genomes, which reveals the major circulating lineages and their genetic relatedness, contextualising the virus’s evolutionary history within The Gambia’s epidemic landscape.

The frequency bar plot depicts the cumulative prevalence of each SARS-CoV-2 lineage among approximately 1,600 genomes collected throughout the pandemic in The Gambia. This analysis was compared to the timing of epidemic wave peaks observed in the epidemiological study conducted by Kanteh et al. (2023). The comparison helps clarify the key distinction between overall lineage frequency and epidemic wave peaks, highlighting the difference between a rapid, sharp transmission surge followed by a decline, versus a more sustained, prolonged transmission pattern.

4.1.1 Polygenetic Tree

The phylogenetic tree (Figure 2) constructed from SARS-CoV-2 genomes sampled in the Gambian cohort reveals a detailed picture of the evolutionary relationships shaped by local transmission dynamics and viral diversification. Consistent with established analytical frameworks (Chen et al., 2023), this approach identifies the ten most prevalent lineages, including AY.34.1, B.1.617.2, BA.1.1, and B.1.1.7, which form prominent clusters in the tree. These dense clusters are indicative of major contributors to local outbreaks, as similarly observed by Sun et al. (2022), who reported strong clustering of dominant variants responsible for epidemic waves. The presence of multiple sub-lineages within these clusters, such as diverse AY. *and B.1.1.* variants, aligns with phylodynamic findings from Attwood et al. (2022), corroborating ongoing viral evolution and adaptation in response to local selective pressures.

The tree further distinguishes separate clusters for lineages like AY.34.1 and B.1.416, which lack intermediate genomes bridging them. This pattern strongly suggests multiple independent introductions rather than continuous evolution within a single local transmission chain, a conclusion supported by phylogeographic interpretations in Jaimes et al. (2020). Additionally, isolated tips and short branches likely reflect either sporadic introductions or less-transmissible transmission chains, in agreement with structural patterns documented by Chen et al. (2023).

Branch length analysis reveals that the scale bar represents 0.003 substitutions per site, corresponding to approximately 0.3% genetic divergence or roughly 90 nucleotide differences across the ~30,000 bp SARS-CoV-2 genome. This relatively small evolutionary distance reflects the recent timeframe of pandemic spread and the slow evolutionary rate of SARS-CoV-2 compared to other RNA viruses. The short branch lengths observed throughout the tree are consistent with rapid epidemic expansion and limited time for substantial genetic divergence, particularly evident in the dense clustering of closely related sequences within major lineages.

Bootstrap analysis reveals robust support (0.8–0.9) for major lineage-defining branches, providing confidence in their distinct evolutionary origins. However, within-lineage branch support drops to around 0.3, reflecting shallow genetic divergence due to rapid epidemic expansions and the resulting high genomic similarity among descendant viruses.

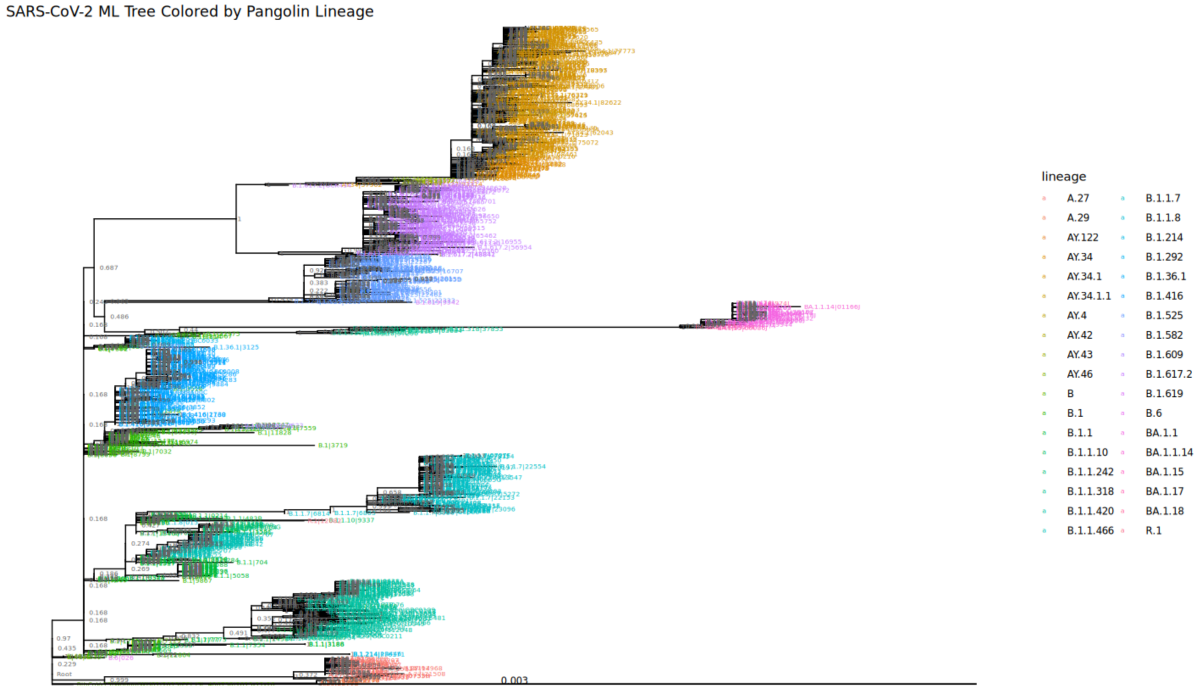


Figure 2. Maximum likelihood phylogenetic tree of SARS-CoV-2 genomes from the Gambian cohort**.** This tree illustrates the evolutionary relationships among SARS-CoV-2 samples, with branches coloured by Pangolin lineage assignments. Major lineages included are AY.34.1, B.1.1.7, B.1.1.420, B.1.1.466, B.1.1, B.1.416, B.1.525, B.1.617.2, B.1, and BA.1.1. The tree was constructed using a neighbour-joining method based on a distance matrix calculated with the best-fit evolutionary model. Negative branch lengths were corrected, followed by maximum likelihood optimization with stochastic rearrangement. Bootstrap values were estimated from 30 replicates to assess branch support. The tree was rooted on the Wuhan-Hu-1 reference genome, and visualisation was performed with ggtree in R, with tip labels coloured by lineage. The scale bar represents 0.003 substitutions per site, indicating the evolutionary distance corresponding to approximately 0.3% genetic divergence.

4.1.2 Top 10 Lineages

The 10 most prevalent lineages identified were AY.34.1 (262 genomes), B.1.416 (184 genomes), B.1.617.2 (160 genomes), B.1.1.7 (140 genomes), B.1.525 (129 genomes), B.1 (133 genomes), B.1.1.420 (96 genomes), B.1.1.466 (31 genomes), B.1.1 (45 genomes), and BA.1.1 (28 genomes) (Figure 3). Delta variant sub-lineage AY.34.1 was the most prevalent, followed by B.1.416 and the parental Delta lineage B.1.617.2, highlighting Delta variants as dominant contributors during the study period.

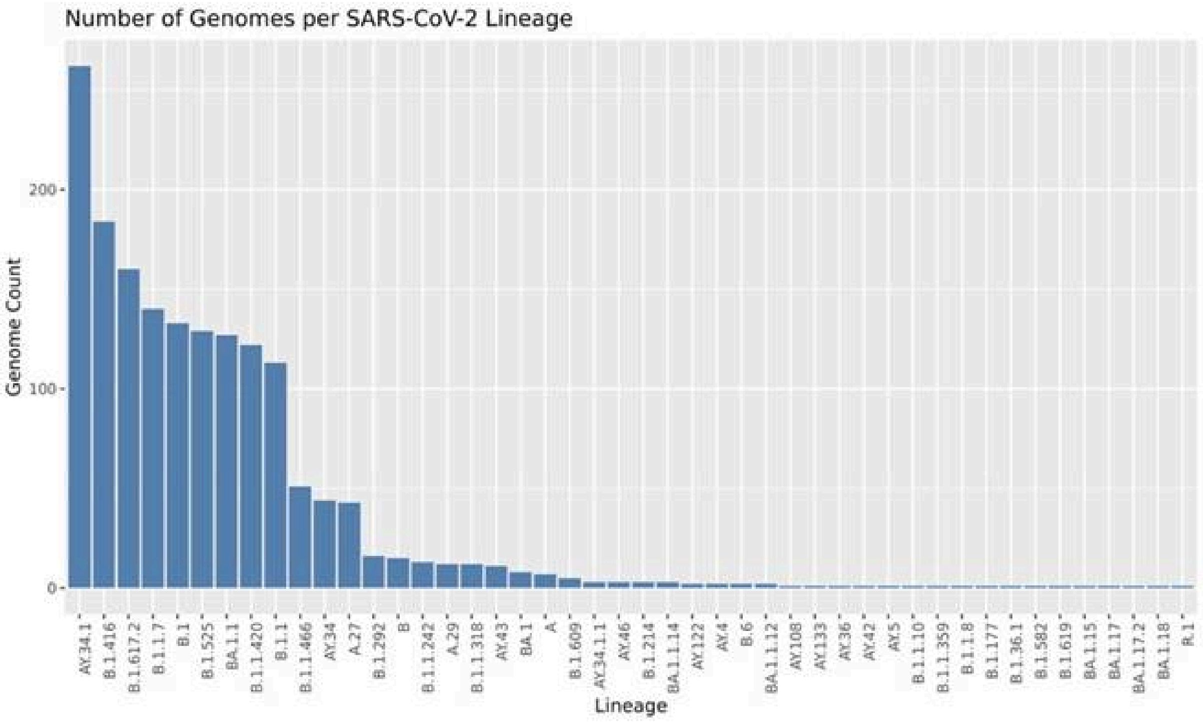


Figure 3: Displays the frequency of each lineage within the dataset before any cleaning of Screening count which excluded 400 genomes.

4.2 Exploratory analysis

CT Count Distribution Assessment

Visual assessment of the cleaned CT count distribution revealed an approximately bellshaped distribution centered around 22-24 cycles, with some notable departures from perfect normality (Figure 7). The histogram showed the distribution was moderately rightskewed with a longer tail extending toward higher CT counts, and the theoretical normal overlay indicated imperfect alignment with the observed data, particularly in the distributional tails. The quantile-quantile plot provided more detailed evidence of deviations from normality, with points following the diagonal reference line reasonably well in the central range but exhibiting systematic departures at the extremes. Specifically, the lower quantiles fell below the theoretical line while upper quantiles rose above it, creating a characteristic S-shaped pattern indicative of heavier tails than expected under normality. Several extreme outliers were observed at high CT counts (>60), contributing to the distributional asymmetry. While these deviations from perfect normality were noted, the overall distribution remained sufficiently bell-shaped to support the use of linear regression methods for subsequent genetic association analyses.

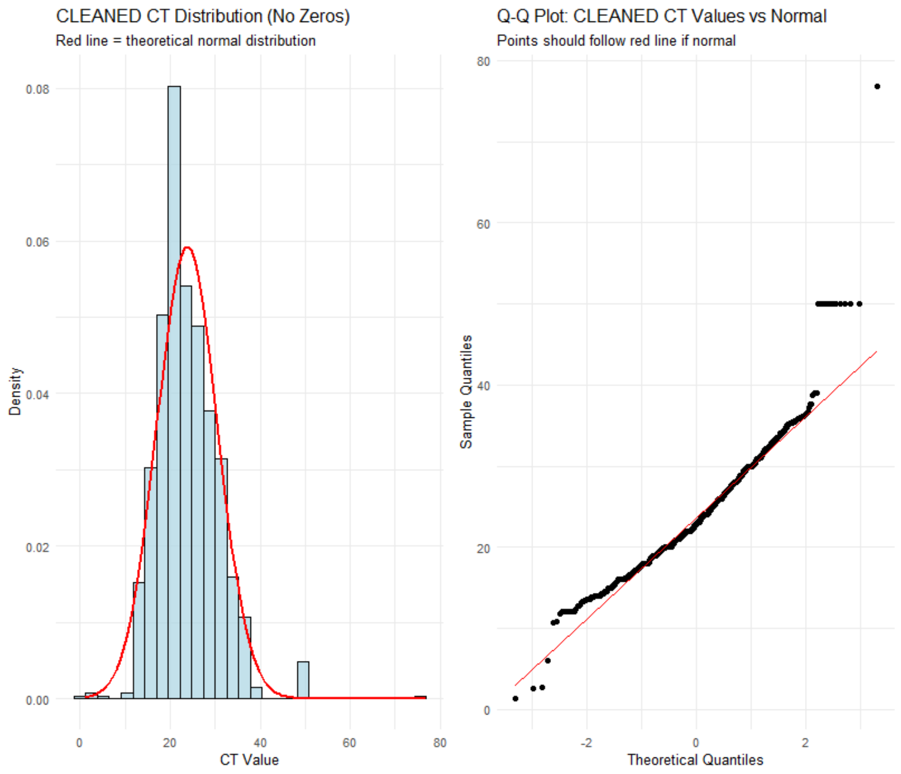


Figure 4: Two-panel visualization of the cleaned cycle threshold (CT) value distribution following removal of biologically implausible zero values from the SARS-CoV-2 dataset (n = 1,037 samples). The left panel displays a histogram showing the empirical distribution of CT counts with 30 bins, overlaid with a theoretical normal distribution curve (red line) using the sample mean (23.81) and standard deviation (6.73). The right panel presents a quantile-quantile (Q-Q) plot comparing the observed CT count distribution against theoretical normal quantiles, with the red diagonal line representing perfect normality. Points that deviate from this line indicate departures from normality, particularly visible in the tails where extreme values show greater deviation than expected under a normal distribution. This analysis demonstrates that while the CT counts approximate a normal distribution in the central range, significant skewness (1.197) and excess kurtosis (8.22) indicate non-normal characteristics, particularly in the upper tail representing samples with higher CT counts.

4.2.1 Population Structure Analysis via Principal Component Analysis

Principal component analysis was conducted on 1,187 viral genome samples representing 36 distinct SARS-CoV-2 lineages. Population structure assessment focused on the first two principal components (PC1 and PC2) to identify genetic clustering patterns and potential population stratification that could confound genome-wide association analyses shown in figure 5.

4.2.2 Outlier Detection and Population Deviations

Individual samples were assessed for deviation from the population center using Euclidean distance calculations in PC1-PC2 space. The population center was established at the mean coordinates (PC1 = 0.000, PC2 = 0.000), consistent with PCA standardisation. A distance threshold of 0.091293 (population mean + 2 standard deviations) was applied to identify outlier samples.

Twenty-four samples (2.02% of the total dataset) exceeded the distance threshold and were classified as population outliers. The most extreme outliers were concentrated in two lineages: AY.34.1 (maximum distance = 0.555) and A.29 (distance = 0.330). Notably, all nine samples from lineage A.29 were classified as outliers, indicating this lineage represents a distinct genetic cluster.

Lineage-Specific Population Structure

Lineage-specific centroids were calculated to assess population clustering patterns. The analysis revealed substantial heterogeneity in lineage positioning within the PC space, with A.29 showing the greatest deviation from the population center (centroid distance = 0.330).

Statistical evaluation identified one lineage (A.29) as significantly deviating from the population center, exceeding the threshold of mean centroid distance plus one standard deviation. This lineage demonstrated both the highest individual sample distances and complete separation from the main population cluster.

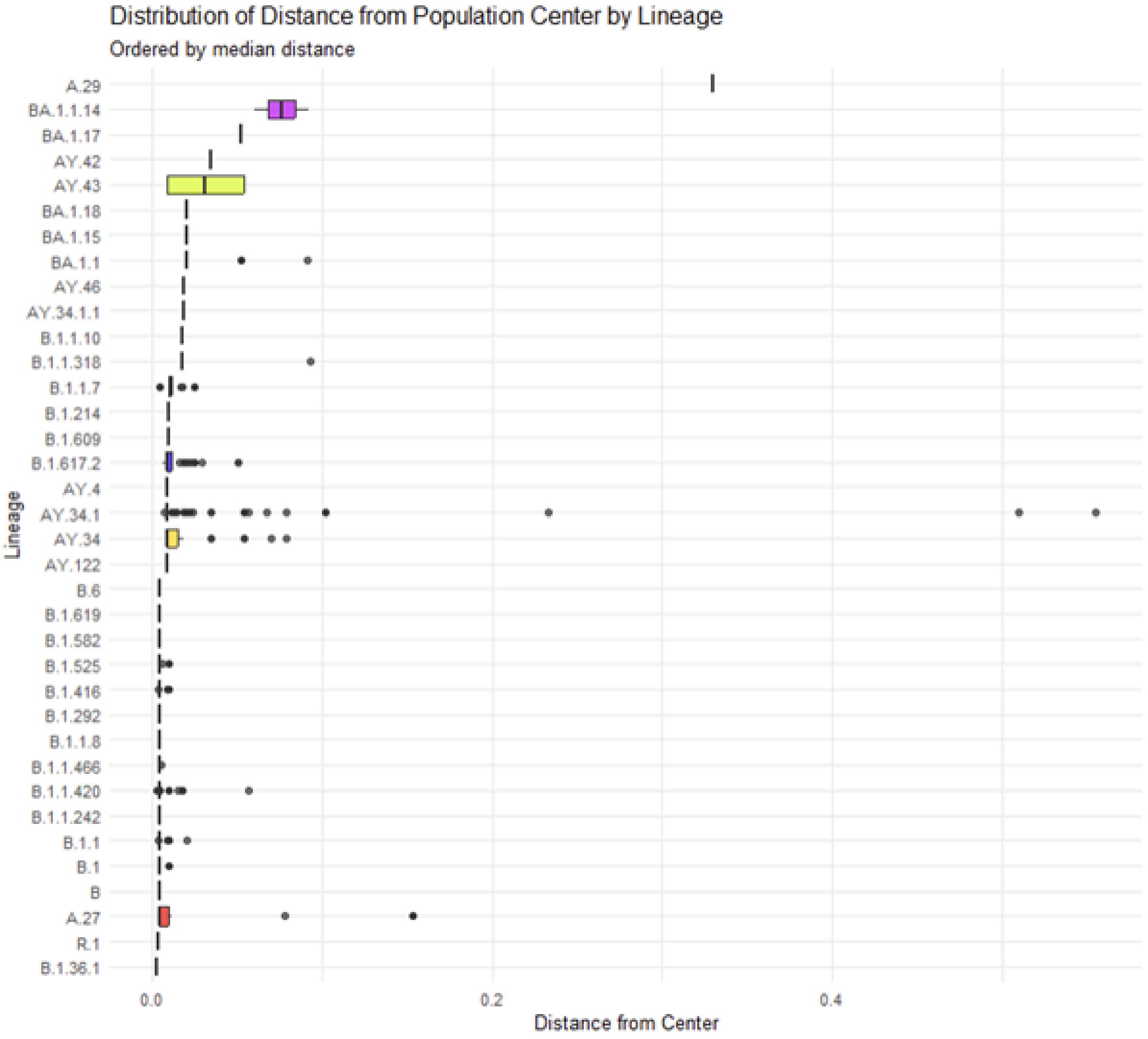


Figure 5: Box plot displaying the distribution of genetic distances from the population center for each SARS-CoV-2 lineage. Lineages are ordered by median distance from the population center (ascending from bottom to top). Each box represents the interquartile range (25th-75th percentile) of distances within a lineage, with the horizontal line indicating the median. Whiskers extend to 1.5 times the interquartile range, and outliers are shown as individual points beyond the whiskers. Lineages positioned higher on the y-axis show greater genetic divergence from the overall population average, indicating more unique genetic characteristics.

4.2.3 Pairwise Lineage Relationships

To quantify genetic differentiation between lineages and identify the most divergent lineage pairs, we calculated pairwise Euclidean distances between all lineage centroids in principal component space. This analysis helps identify which lineages are most genetically distinct and provides quantitative support for visual clustering patterns observed in PCA plots.

Pairwise distances between lineage centroids revealed pronounced separation patterns. The most distant lineage pairs consistently involved A.29, with distances ranging from 0.335 to

0.349 relative to other major lineages (Table 3). This finding reinforces A.29's position as a genetic outlier identified in individual sample analysis.

Lineage 1	Lineage 2	Distance
BA.1.1.14	A.29	0.3488875
BA.1.17	A.29	0.3410708
AY.42	A.29	0.3405024
AY.34.1.1	A.29	0.3383099
AY.46	A.29	0.3383099
AY.43	A.29	0.3376017
AY.34	A.29	0.3358480
A.27	A.29	0.3357208
B.1.617.2	A.29	0.3356440
AY.34.1	A.29	0.3353186

Table 3: Pairwise Euclidean distances between SARS-CoV-2 lineage centroids in principal component space, ranked by distance in descending order. Each row represents a unique lineage pair, with distances calculated between the mean PC1 and PC2 coordinates of each lineage's centroid. Higher distance values indicate greater genetic differentiation between lineage pairs in the two-dimensional PCA space. Only the top 10 most distant pairs are shown to highlight the lineages with the greatest genetic divergence.

In contrast, several B.1 sublineages (B.1.1.8, B.1.292, B.1.582, B.1.619, B.6) showed identical centroid positions (pairwise distance = 0.000), indicating complete genetic clustering. This suggests these represent closely related viral variants with minimal population structure differences, likely reflecting recent common ancestry or limited evolutionary divergence..

4.2.4 Population Structure Visualisation and Implications

Visual Confirmation of Genetic Clustering
Visual inspection of PC1 versus PC2 scatter plots confirmed the numerical analyses, revealing A.29 as a clearly separated cluster in the lower-left quadrant of the PCA space. The remaining 35 lineages formed a relatively tight cluster around the population center, with minimal separation between most lineages (Figure 6).

Box plot analysis of distance distributions by lineage confirmed A.29 as the most divergent population, with consistently higher distances from the population center compared to all other lineages.

Quantitative Distance Analysis
Distance matrix analysis of SARS-CoV-2 lineages revealed A.29 as a consistently divergent outlier with pairwise distances of 0.335-0.349 from all other major lineages (BA.1.1.14, BA.1.17, AY.42, etc.). In contrast, several B.1 sublineages showed complete genetic clustering (distance = 0.000), indicating minimal population structure differences between closely related variants.

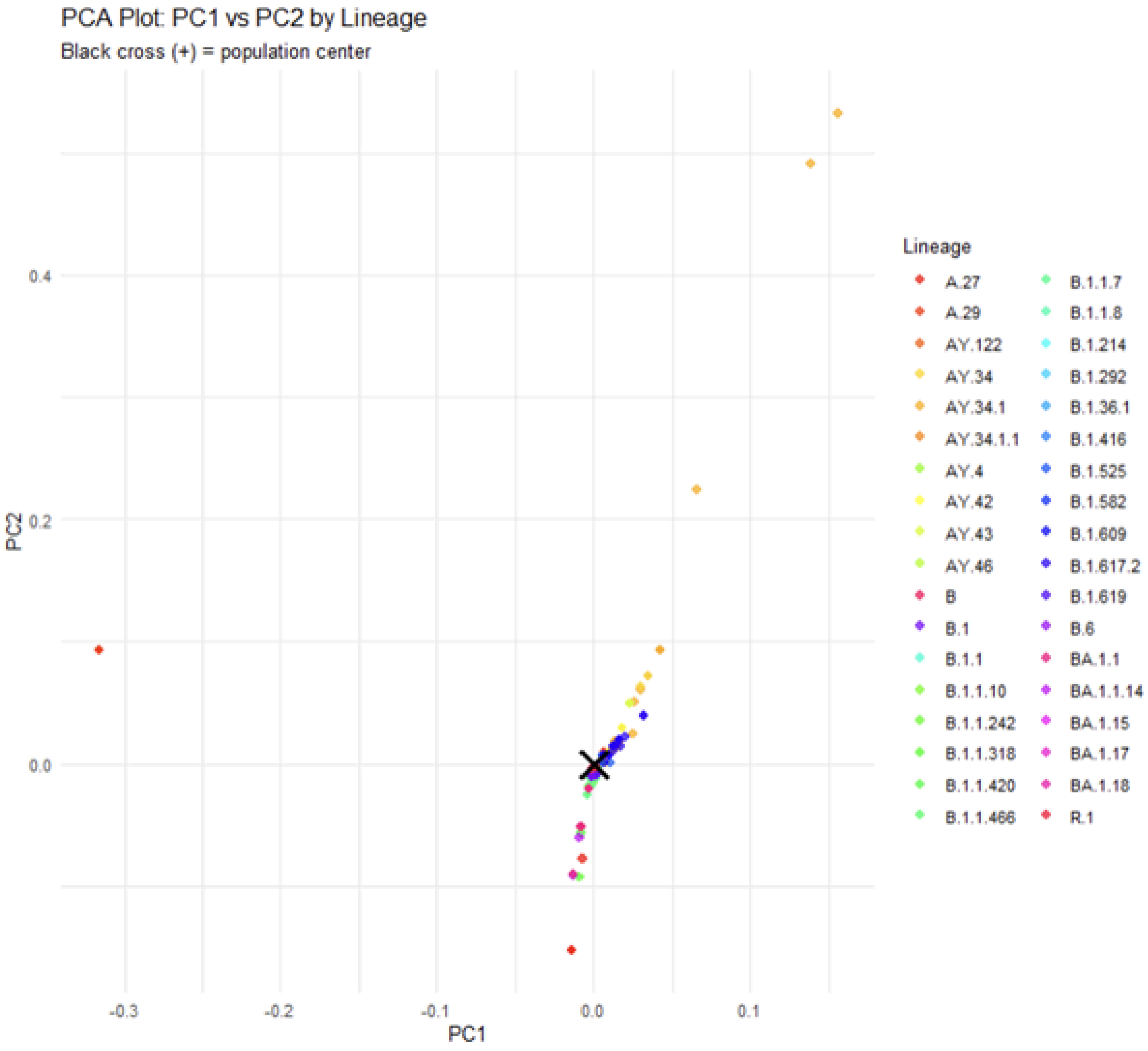
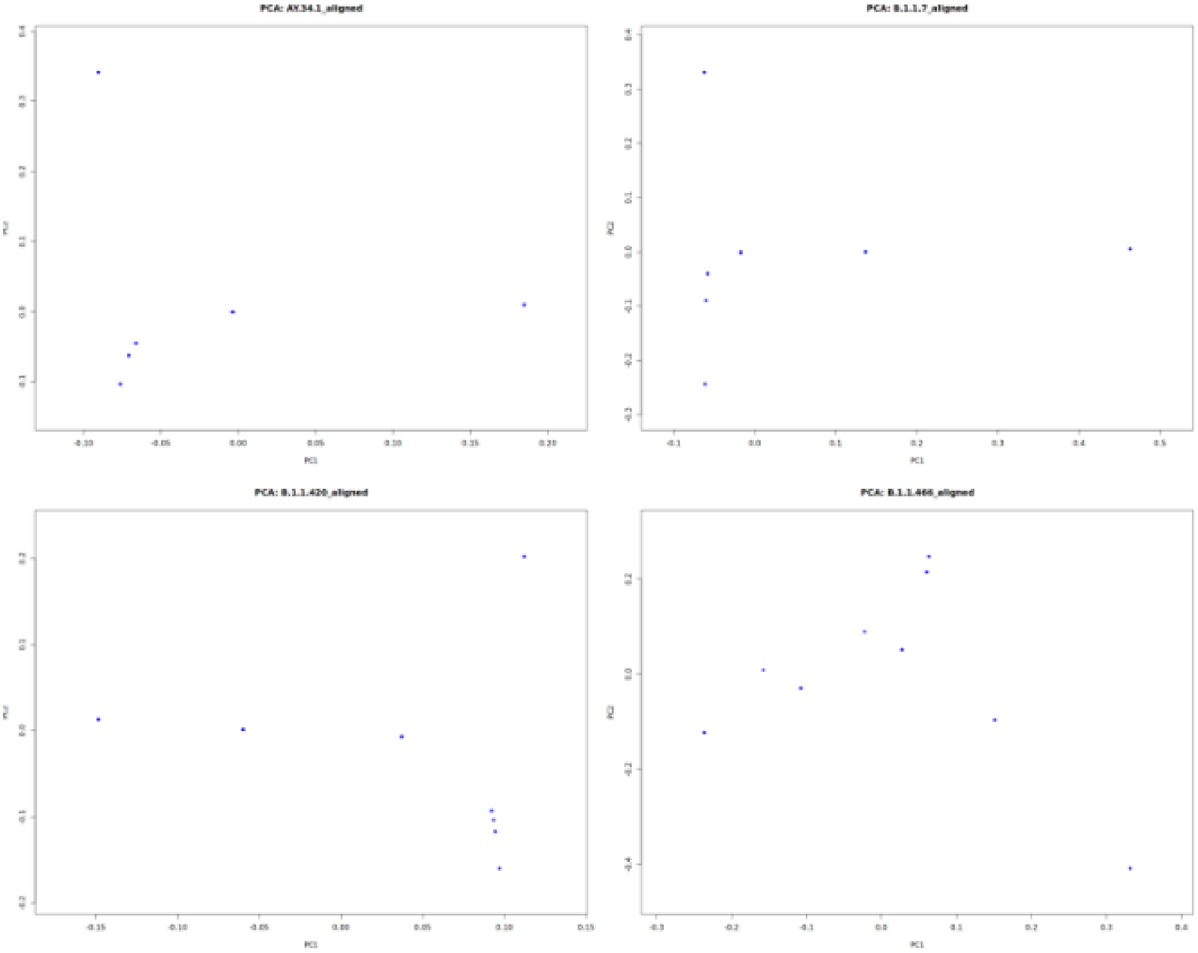


Figure 6: Principal Component Analysis (PCA) plot showing the distribution of SARS-CoV-2 samples across the first two principal components (PC1 and PC2). Each point represents an individual sample, colored by viral lineage assignment. The black cross (+) indicates the population center (mean of PC1 and PC2 across all samples). Points further from the center represent samples with greater genetic divergence from the population average. Clustering of points by color indicates genetic similarity within lineages, while separation between color groups demonstrates genetic differentiation between lineages.

Top 10 Lineages PCA

Principal Component Analysis (PCA) was performed separately for each of the top 10 viral lineages to assess underlying genetic variation and population structure. Across all PCA plots, genomes appeared widely distributed without forming distinct clusters, with individual points occasionally grouping into very small local clusters of up to 3–5 genomes (Figure 7). The number of single nucleotide polymorphisms retained after quality control ranged from 9 to 20 SNPs per lineage. The absence of distinct clustering indicates a largely homogeneous genetic background within lineages at the resolution examined.



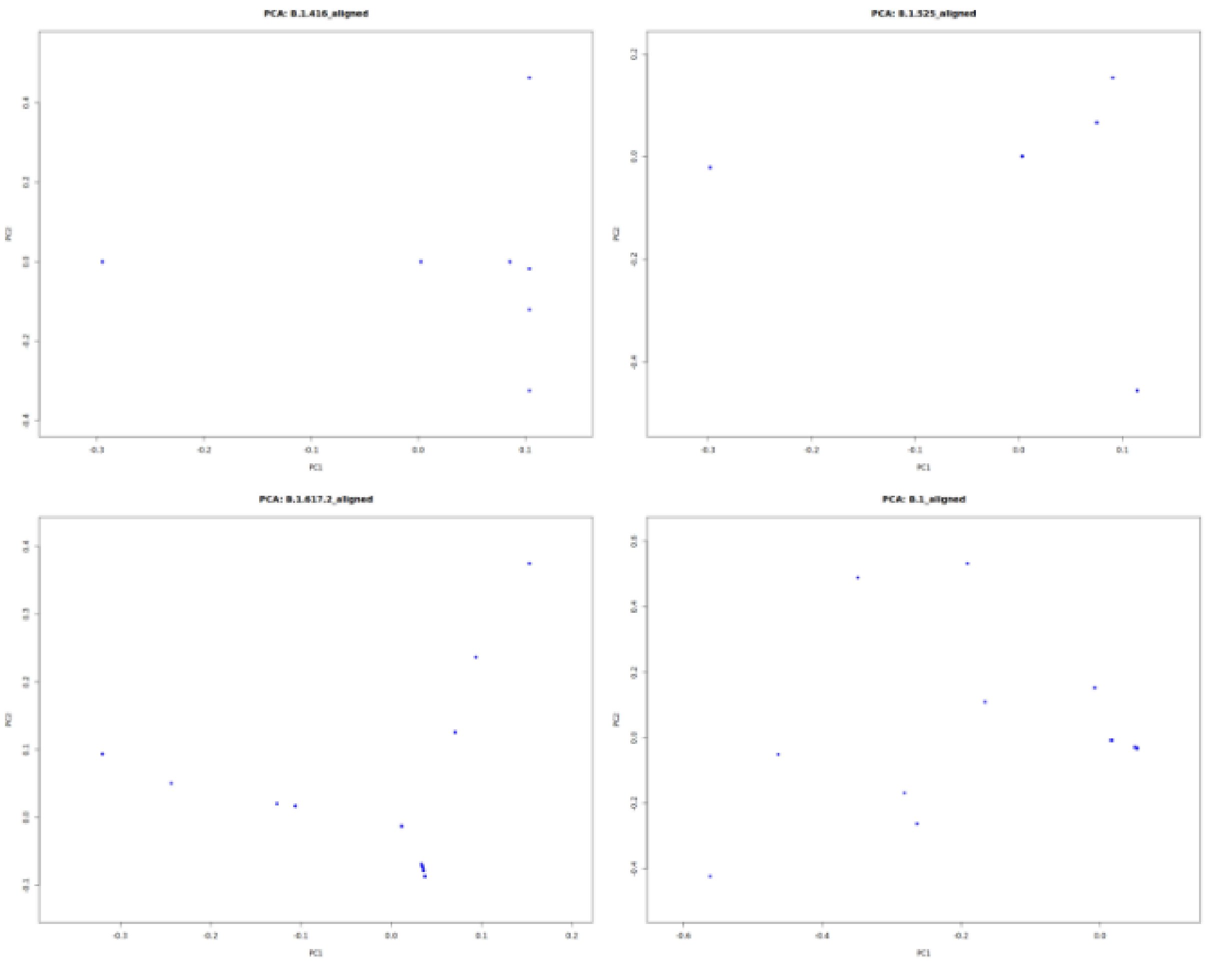


Figure 7: Multi-panel display showing Principal Component Analysis (PCA) results for top 10 individual SARS-CoV-2 lineages. Each panel represents a separate lineage (indicated by panel titles), displaying the distribution of samples within that lineage across the first two principal components (PC1 on x-axis, PC2 on y-axis). Individual points represent viral samples belonging to the specified lineage. The first panel shows AY.34.1, B.1.1.7, B.1.1.420 and B.1.1.466 respectively. And the second panel shows B.1.416, B.1.525, B.1.617.2 and B.1 respectively.

4.3 Simple Linear Regression

4.3.1 All Lineages

In genome-wide association studies (GWAS), QQ plots are essential quality control tools that compare the distribution of observed test statistics (p-values) against their expected distribution under the null hypothesis of no genetic association. Under ideal conditions with no true associations and no systematic bias, points should follow the diagonal line closely. Deviations from this line can indicate: true genetic associations (rightward tail deviation), population stratification (systematic upward shift), or technical artifacts.

This QQ plot (Figure 8) shows the results from a simple linear regression analysis examining genetic associations across SARS-CoV-2 lineages. The key findings are:

1. **Good Model Fit (λ = 0.902):** The genomic inflation factor below 1.0 indicates no evidence of population stratification or systematic inflation of test statistics, suggesting the PCA correction has effectively controlled for population structure.
2. **No Strong Associations:** The observed p-values closely follow the expected null distribution along the diagonal, with no clear deviation in the tail region. This suggests there are no variants with genome-wide significant associations with the phenotype being tested.
3. **Appropriate Sample Size:** With 69 variants analysed, the plot demonstrates that the statistical analysis is well-calibrated and free from major confounding factors that could lead to false positive associations.

The slight deflation (λ < 1.0) is not concerning and may reflect conservative analysis parameters or effective population structure correction through PCA.

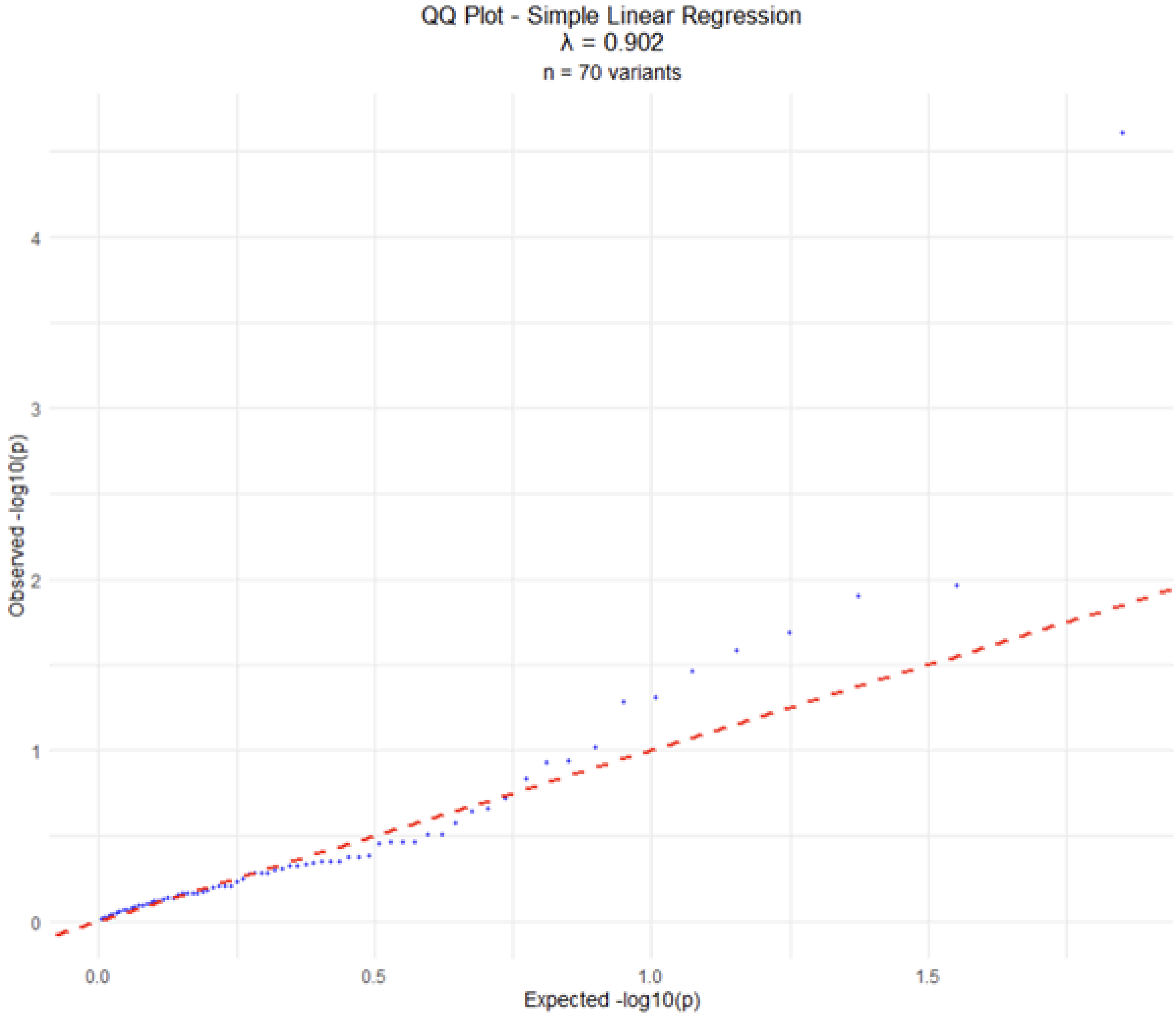


Figure 8: Quantile-quantile (QQ) plot displaying the distribution of observed versus expected -log$_{10}$(p-values) from genome-wide association analysis of SARS-CoV-2 lineage data. The x-axis shows expected -log$_{10}$(p-values) under the null hypothesis of no association, while the y-axis shows observed -log$_{10}$(p-values) from the actual analysis. Each point represents a genetic variant (n = 69 variants total). The red dashed diagonal line represents the expected distribution under the null hypothesis (λ

= 1.0). The genomic inflation factor (λ) of 0.902 indicates minimal population stratification or systematic bias in the analysis.

All Lineages Simple Regression

In the simple linear regression analysis, a single SNP at position 15222 shown in figure 9 on the SARSCoV-2 reference genome (NC_045512.2) was identified as significantly associated with CT counts after applying a stringent Bonferroni-corrected significance threshold. This SNP is located within the coding sequence spanning positions 13467 to 21555, corresponding to the gene product YP_009724389.1, known as the polyprotein pp1ab. The pp1ab polyprotein plays a critical role in viral replication and transcription, and variants within this region may influence viral fitness or replication efficiency, potentially impacting viral load measured by CT counts.

The identification of only one significant SNP passing the conservative Bonferroni correction suggests a robust and unlikely false-positive association at this locus. However, the overall limited number of significant SNPs indicates that the simple linear regression model captures only a modest genetic influence on CT counts. This limitation likely arises from the model's inability to account for confounding factors such as viral lineage and host characteristics.

Therefore, while the SNP at position 15222 represents a promising candidate impacting viral load, more comprehensive multivariable analyses are necessary. Such models can adjust for confounders, enhance statistical power, and provide a clearer understanding of the complex genetic architecture underlying CT count variation.

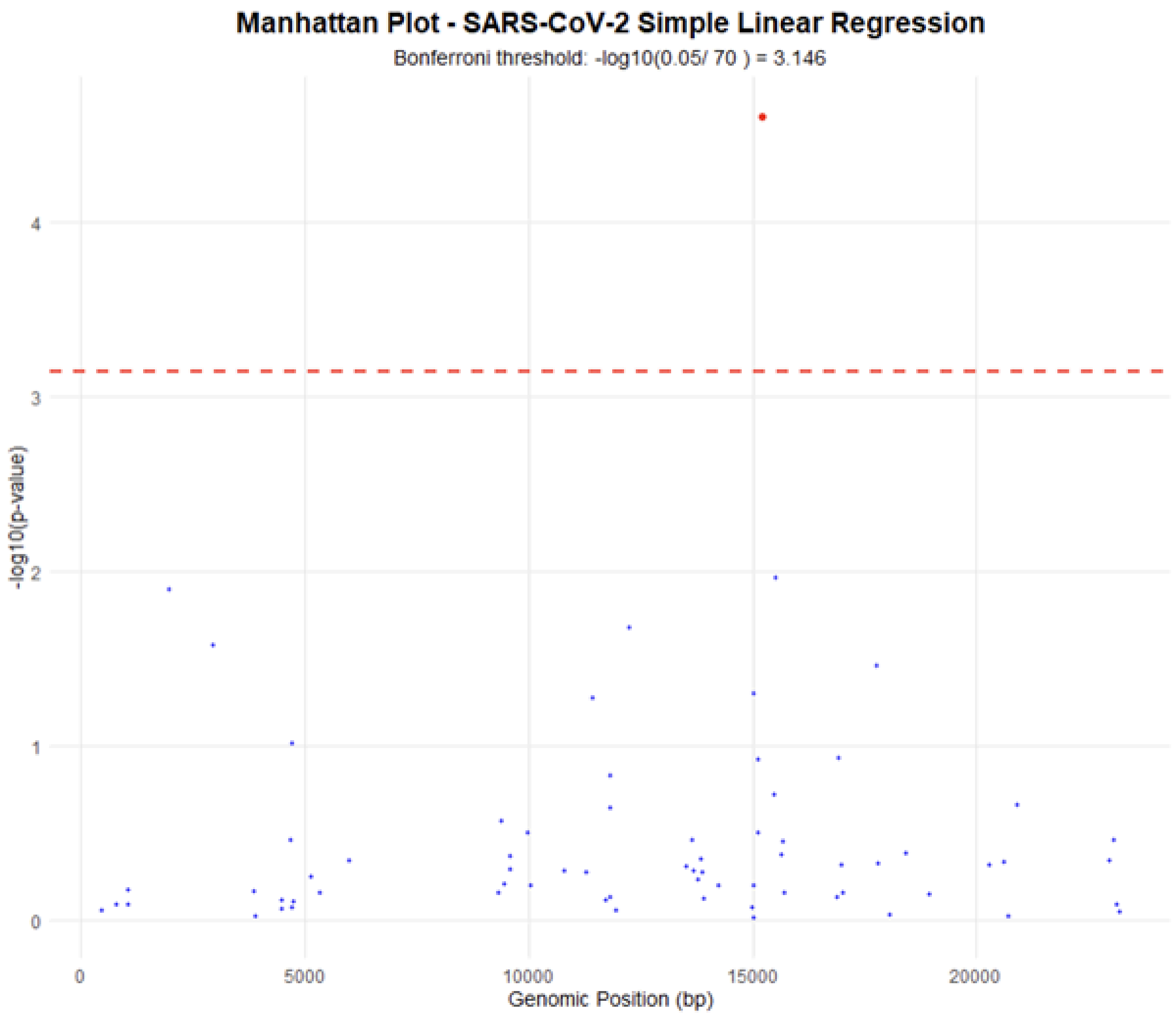


Figure 9 – Manhattan plot showing the association results of 69 SARS-CoV-2 SNPs with CT counts using simple linear regression across the viral genome. The x-axis indicates genomic position, and the y-axis shows –$\log_{10}$(p-values). The horizontal line represents the Bonferroni-corrected significance threshold (3.14). One SNP exceeds this threshold, indicating a significant association.

4.3.2 Top 10 Lineages

We conducted genome-wide association studies (GWAS) for cycle threshold (CT) values across ten SARS-CoV-2 lineages. Following quality control, the number of single nucleotide polymorphisms (SNPs) retained after minor allele frequency (MAF) filtering (MAF ≥ 0.01) and linkage disequilibrium (LD) pruning ranged from 0 to 20 (Table 4). The SNP counts reflect the number of variants used in the regression analysis after all QC filtering.

Table 4. Summary of GWAS results by lineage

Lineage	Sample Size (SNPs (n))	Threshold -log10(P)	No. significant SNPs	MAF	No pruned out (LD pruning)
AY.34.1	13	2.41497	0	0.01	3
B.1.1.7	9	2.25527	0	0.01	7

B.1.1.420	20	2.60206	0	0.01	16
B.1.1.466	19	2.57978	0	0.01	14
B.1.1	0	0	0	0.01	0
B.1.416	11	2.34242	0	0.01	5
B.1.525	14	2.44716	0	0.01	7
B.1.617.2	20	2.60206	0	0.01	19
B.1	12	2.41497	0	0.01	4
BA.1.1	0	1.60206	0	0.01	0

No SNPs reached genome-wide significance in any lineage after Bonferroni correction. Threshold values for $-\log_{10}(P)$ ranged from ~2.26 to 2.60 depending on the number of SNPs tested (0.05/n).

QQ Plots

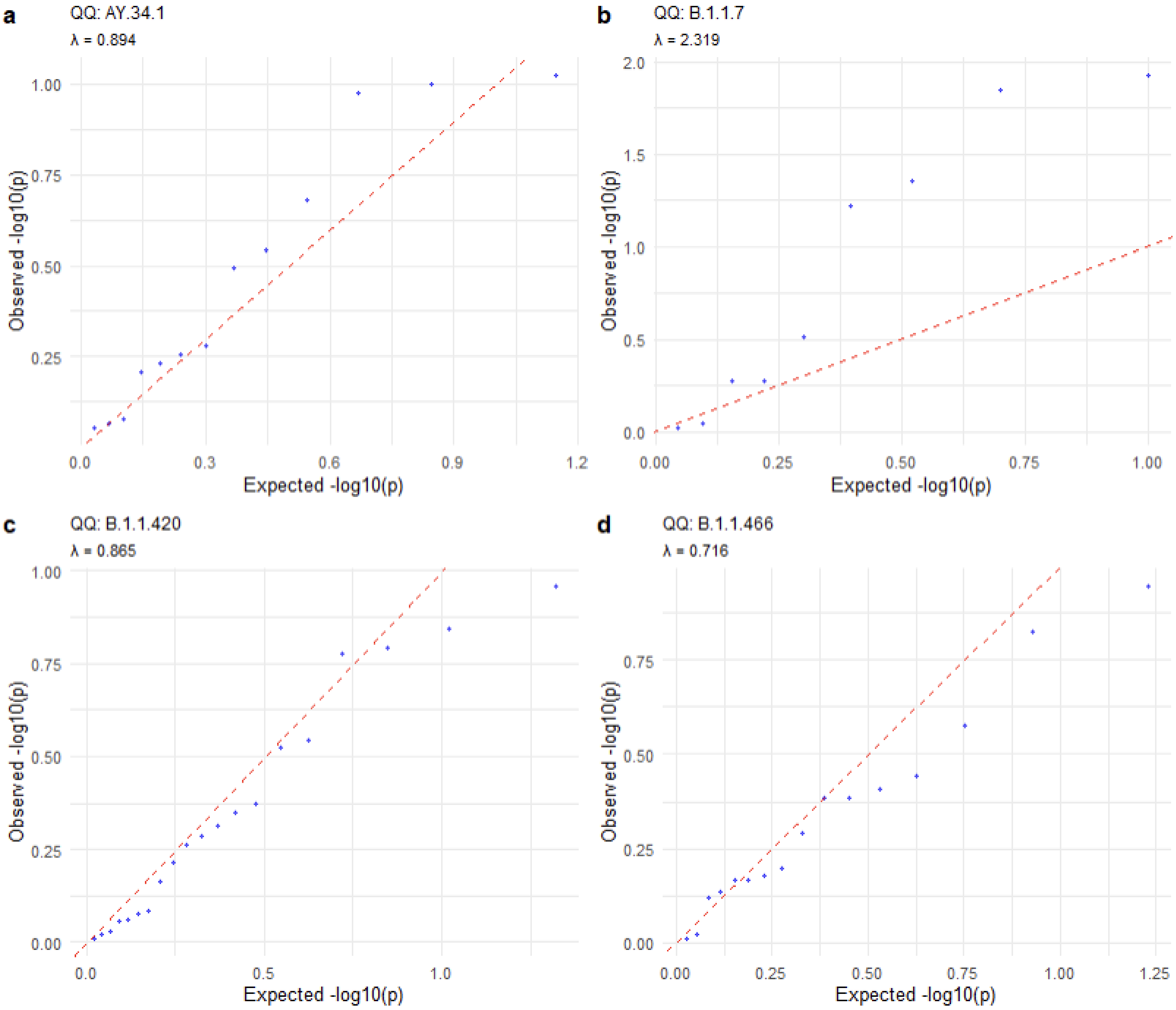
These QQ plots (Figure 10) represent individual genome-wide association analyses conducted separately for each major SARS-CoV-2 lineage, allowing assessment of genetic associations within specific viral lineages rather than across the entire population. This approach can reveal lineage-specific genetic effects that might be masked in population-wide analyses.

The analyses reveal distinct patterns of statistical behavior across different SARS-CoV-2 lineages. Three lineages demonstrate excellent calibration with genomic inflation factors close to the ideal value of 1.0. AY.34.1 ($\lambda = 0.894$), B.1.1.420 ($\lambda = 0.865$), and B.1.1.466 ($\lambda = 0.716$) all show observed p-values closely following the expected distribution along the diagonal, indicating no systematic bias and no strong genetic associations detected within these lineages.

Four lineages display mild to moderate genomic inflation above the expected threshold. B.1.416 ($\lambda = 1.195$), B.1.525 ($\lambda = 1.302$), B.1.617.2 ($\lambda = 1.163$), and B.1 ($\lambda = 1.327$) show inflation that could reflect residual population stratification within these specific lineages, small sample sizes leading to statistical instability, or potentially genuine genetic associations that are specific to these viral variants.

The most concerning pattern appears in B.1.1.7 ($\lambda = 2.319$), which shows substantial systematic inflation that may indicate inadequate population structure correction or technical issues specific to this lineage's analysis. This level of inflation warrants careful examination of the underlying data and analytical approach for this particular lineage.

Overall the majority of lineages show well-controlled analyses with appropriate statistical behavior. The variation in λ values across lineages suggests that genetic architecture and population structure may differ between SARS-CoV-2 lineages, highlighting the value of lineage-specific analyses. Lineages with $\lambda > 1.2$ may warrant additional investigation for population substructure or represent genuine biological differences in genetic association patterns.


a
QQ: AY.34.1
λ = 0.894
Observed -log10(p)
1.00
0.75
0.50
0.25
0.0
0.3
0.6
0.9
1.2
Expected -log10(p)
b
QQ: B.1.1.7
λ = 2.319
Observed -log10(p)
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Expected -log10(p)
c
QQ: B.1.1.420
λ = 0.865
Observed -log10(p)
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
Expected -log10(p)
d
QQ: B.1.1.466
λ = 0.716
Observed -log10(p)
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
Expected -log10(p)

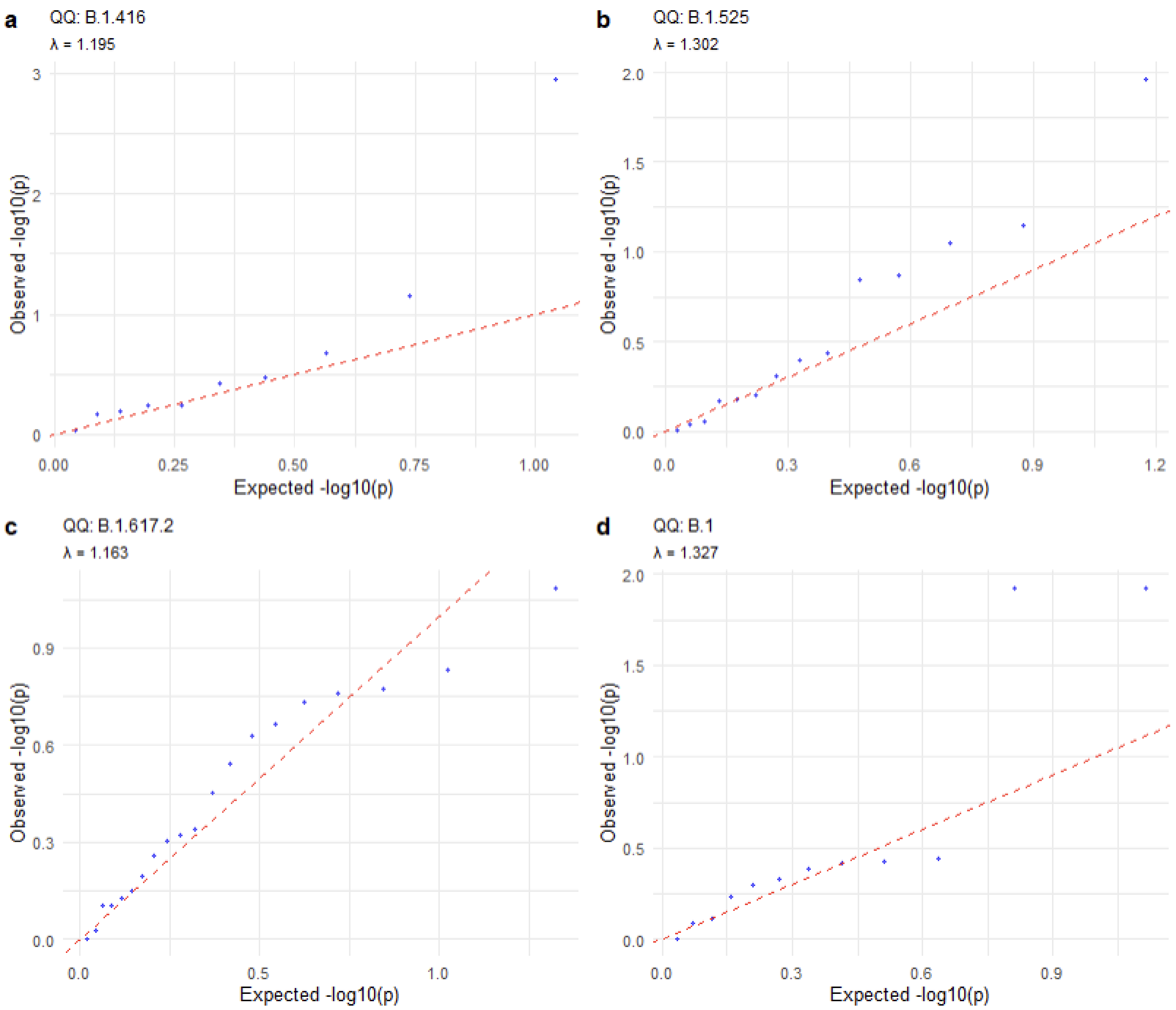


Figure 10: Multi-panel quantile-quantile (QQ) plots showing the distribution of observed versus expected -log$_{10}$(p-values) from simple linear regression analyses performed separately for individual SARS-CoV-2 lineages. Each panel (a-h) represents a different lineage as indicated by the panel titles (e.g., QQ_AY.34.1, QQ_B.1.1.7, etc.). The x-axis shows expected -log$_{10}$(p-values) under the null hypothesis, while the y-axis shows observed -log$_{10}$(p-values) from the lineage-specific analysis. Each point represents a genetic variant tested within that lineage. The red dashed diagonal line represents perfect adherence to the null distribution (λ = 1.0). The genomic inflation factor (λ) for each lineage is displayed in the upper left corner of each panel.

Manhattan Plots

The Manhattan plots (Figure 11) for all analyzed lineages reveal distinct patterns of genetic association across the SARS-CoV-2 genome. After quality control filtering, the number of SNPs available for analysis varied considerably between lineages, with some retaining as few as 13 variants while others maintained up to 20 variants for testing.

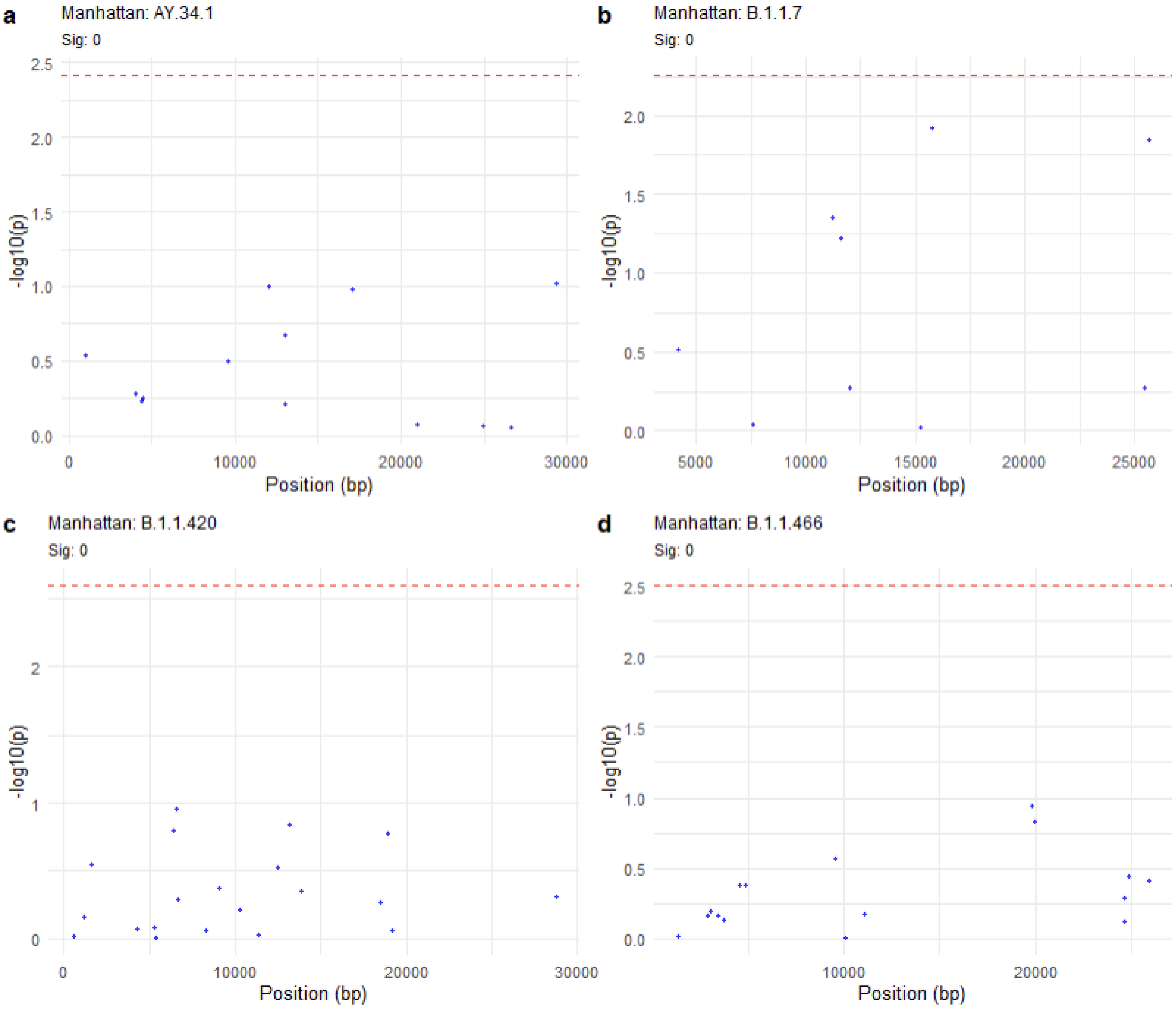
Association peaks across lineages were generally of low to moderate magnitude, with most maximum –log$_{10}$(P) values ranging between 1.0 and 2.5. The B.1.415 lineage displayed one of the strongest individual association signals, with a peak reaching approximately –log$_{10}$(P)

= 3.0 in the early genome region. Similarly, several other lineages showed notable peaks that approached or occasionally exceeded their respective significance thresholds, indicating potential genetic associations worthy of further investigation.

Lineages B.1.617.2 and B.1.1.420 contained the highest number of SNPs after quality control filtering (20 each), resulting in the most stringent Bonferroni thresholds ($-\log_{10}(P) \approx 2.60$). The B.1.617.2 and B.1 plots displayed some of their strongest association peaks towards the latter part of the genome (20,000–30,000 bp).

Across all Manhattan plots, several global patterns emerged. Most lineages displayed polygenic association patterns rather than single dominant loci, with signals distributed across multiple genomic positions. The most prominent peaks across lineages occurred early (3,000–5,000 bp) or mid-genome (10,000–15,000 bp), rather than late in the genome. Notable exceptions were B.1.617.2 and B.1, which had prominent peaks in the 20,000–30,000 bp range. Some lineages exhibited relatively uniform low-signal profiles (e.g., B.1.1.420), while others showed more heterogeneous association landscapes with several distinct peaks of varying magnitude.

These findings suggest that genetic variation may influence the tested phenotype through multiple loci with modest to moderate individual effects. The variation in association patterns between lineages indicates that different viral variants may have distinct genetic architectures, supporting the analytical approach of conducting lineage-stratified association studies in viral genomics research.


a
Manhattan: AY.34.1
Sig: 0
-log10(p)
0.0
0.5
1.0
1.5
2.0
2.5
0
10000
20000
30000
Position (bp)
b
Manhattan: B.1.1.7
Sig: 0
-log10(p)
0.0
0.5
1.0
1.5
2.0
5000
10000
15000
20000
25000
Position (bp)
c
Manhattan: B.1.1.420
Sig: 0
-log10(p)
0
1
2
0
10000
20000
30000
Position (bp)
d
Manhattan: B.1.1.466
Sig: 0
-log10(p)
0.0
0.5
1.0
1.5
2.0
2.5
10000
20000
Position (bp)

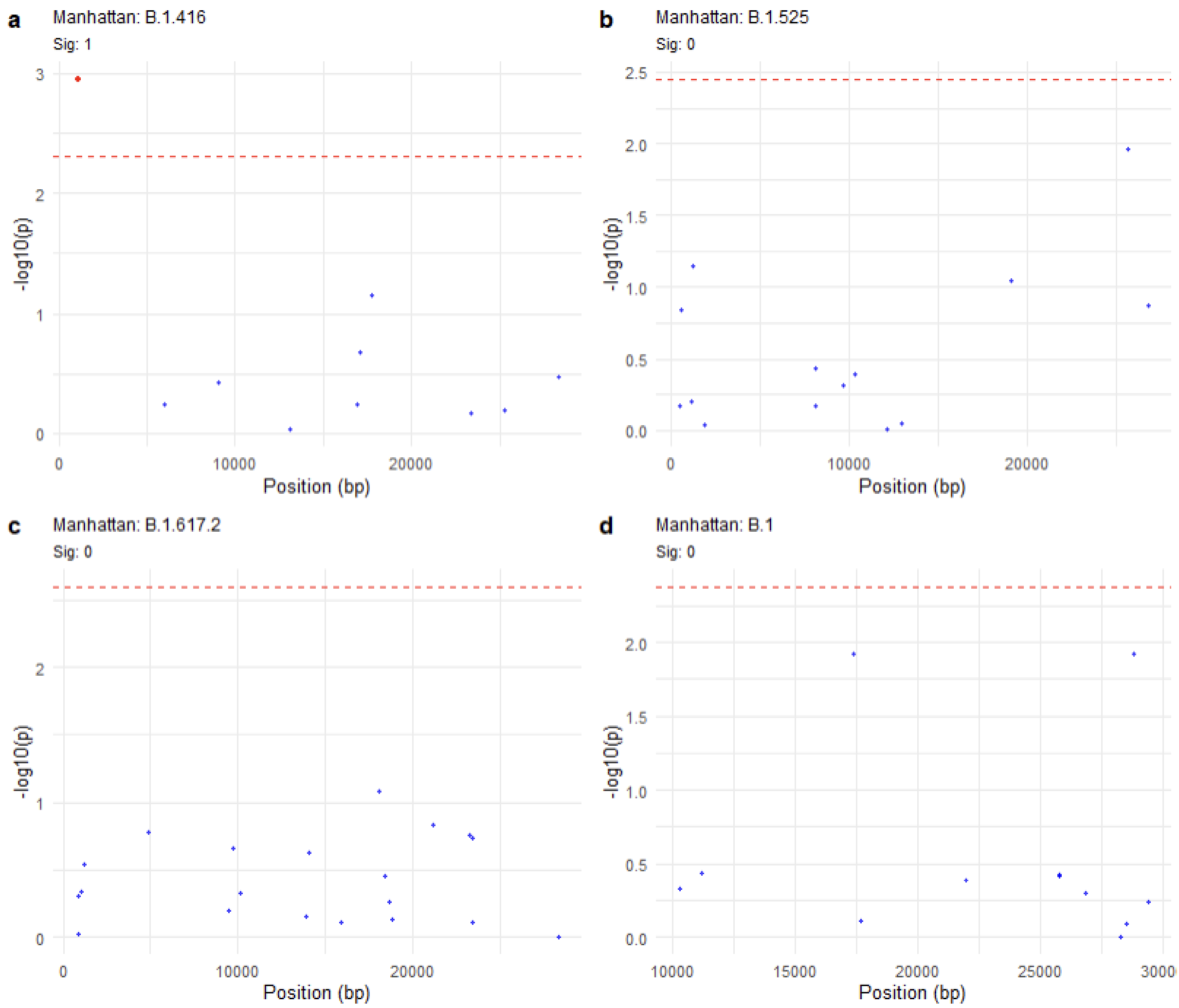


Figure 11: Multi-panel Manhattan plots displaying genome-wide association results from simple linear regression analyses performed separately for individual SARS-CoV-2 lineages. Each panel (a-d) represents a different lineage as indicated by panel titles (e.g., Manhattan AY.34.1, Manhattan B.1.1.7, etc.). The x-axis shows genomic position in base pairs (bp) across the SARS-CoV-2 genome, while the y-axis displays the negative logarithm of association p-values ($-\log_{10}(P)$) for each tested genetic variant. Each point represents a single nucleotide polymorphism (SNP) that passed quality control filters within that lineage. The horizontal red dashed line indicates the genome-wide significance threshold adjusted for multiple testing using Bonferroni correction. The significance threshold varies by lineage depending on the number of SNPs tested after quality control filtering.

4.3.3 Three-Model Approach for Population Structure Control

Model 1 (Primary Analysis): Includes all samples (n=total) with A.29 binary covariate plus PC1 and PC2 as covariates. This model directly controls for the major population stratification (A.29 vs. all others) while accounting for residual population structure through principal components, providing the most comprehensive approach to maintain statistical power while controlling confounding.

Model 2 (Sensitivity Analysis): Excludes individual outlier samples while retaining the A.29 covariate and PC controls on remaining samples. This tests whether results are robust after removing the most extreme genetic outliers, ensuring that findings are not driven by a few highly divergent samples that could disproportionately influence associations.

Model 3 (Stratified Analysis): Completely removes all A.29 samples and analyses only the remaining population using PC1 and PC2 controls. This approach tests associations within the genetically homogeneous population, eliminating the major source of stratification entirely to determine if signals persist in the absence of the most divergent lineage, thereby distinguishing true host genetic effects from population structure artifacts.

Manhattan Plot Pattern Analysis

Model 1 (Primary Analysis): Controls for A.29 using a binary covariate while retaining all samples. Lambda = 0.941 indicates slight test statistic deflation, suggesting potentially over-conservative correction. One SNP (position 15222) reaches Bonferroni significance (P = 2.71e-05).

Model 2 (Sensitivity Analysis): Removes individual outliers but retains A.29 covariate correction. Shows similar results to Model 1 with the same significant SNP (P = 2.81e-05) and more pronounced deflation (Lambda = 0.909), indicating that outlier removal may have been overly conservative.

Model 3 (Stratified Analysis): Completely excludes A.29 samples, eliminating the major source of population structure. The previously significant SNP loses significance entirely (P = 5.62e-03), with Lambda = 1.053 showing slight inflation but closer to ideal.

The Significance Drop in Model 3 (Figure 12) Reveals Population Stratification Artifact. The dramatic loss of significance for the top SNP when A.29 samples are removed (from P = 2.7e-05 to P = 5.6e-03) strongly suggests this association was driven by population structure rather than genuine host genetic effects. This is a classic example of population stratification creating spurious associations.

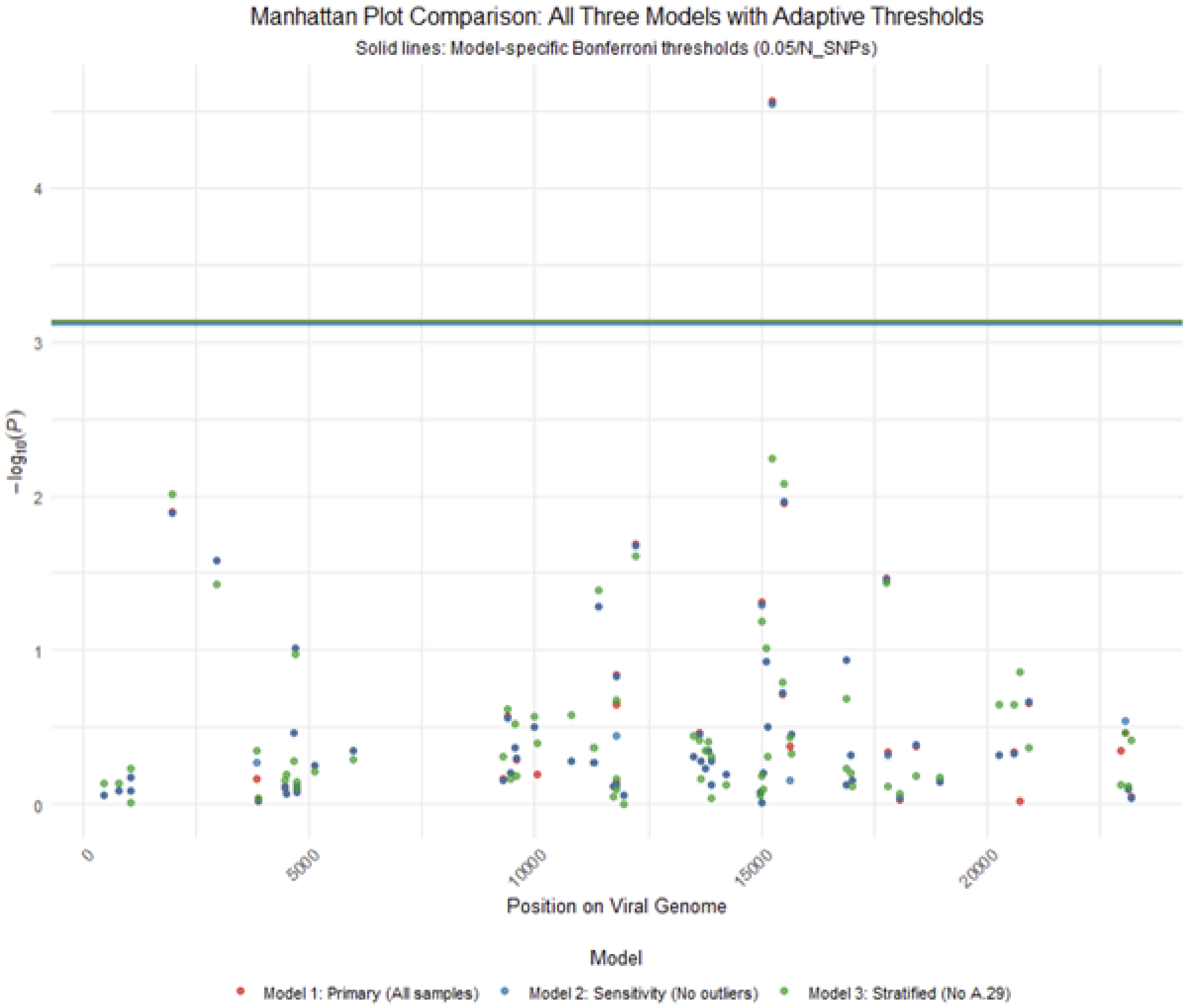


Figure 12: Comparative Manhattan plot displaying genome-wide association results from three different analytical models applied to the complete SARS-CoV-2 dataset. The x-axis shows genomic position across the viral genome (0-30,000 bp), while the y-axis displays the negative logarithm of association p-values ($-\log_{10}(P)$) for each tested genetic variant. Three models are represented by different colored points: Model 1 (Primary - all samples) in red, Model 2 (Sensitivity - no outliers) in blue, and Model 3 (Stratified - no A.29) in green. The horizontal gray line indicates the genome-wide significance threshold using model-specific Bonferroni corrections that account for the number of variants tested in each analysis. This adaptive threshold approach ensures appropriate multiple testing correction while allowing for direct comparison of association strengths across the three analytical approaches.

QQ Plot Pattern Analysis

All three analytical models demonstrate a consistent deviation from the expected null distribution. The observed pattern in figure 13 is characterised by an initial departure from the diagonal line in the lower p-value range, followed by convergence toward the null expectation in the intermediate range, and subsequent flattening in the tail region. This pattern indicates several key findings: first, the presence of a limited number of genuine genetic associations or residual population structure effects manifesting in the tail distribution; second, the majority of tested variants conform to the null hypothesis of no association, as evidenced by the convergence in the middle range; and third, potential overcorrection or deflation in the intermediate p-value range, possibly resulting from conservative covariate adjustment strategies. The remarkable consistency of this pattern across all three models suggests that underlying population structure effects persist despite covariate inclusion, indicating that the observed deviations reflect genuine biological signal rather than model-specific artifacts. This concordance across analytical approaches strengthens confidence in the robustness of the identified associations while highlighting the limitations of standard population structure correction methods in viral genomics applications.

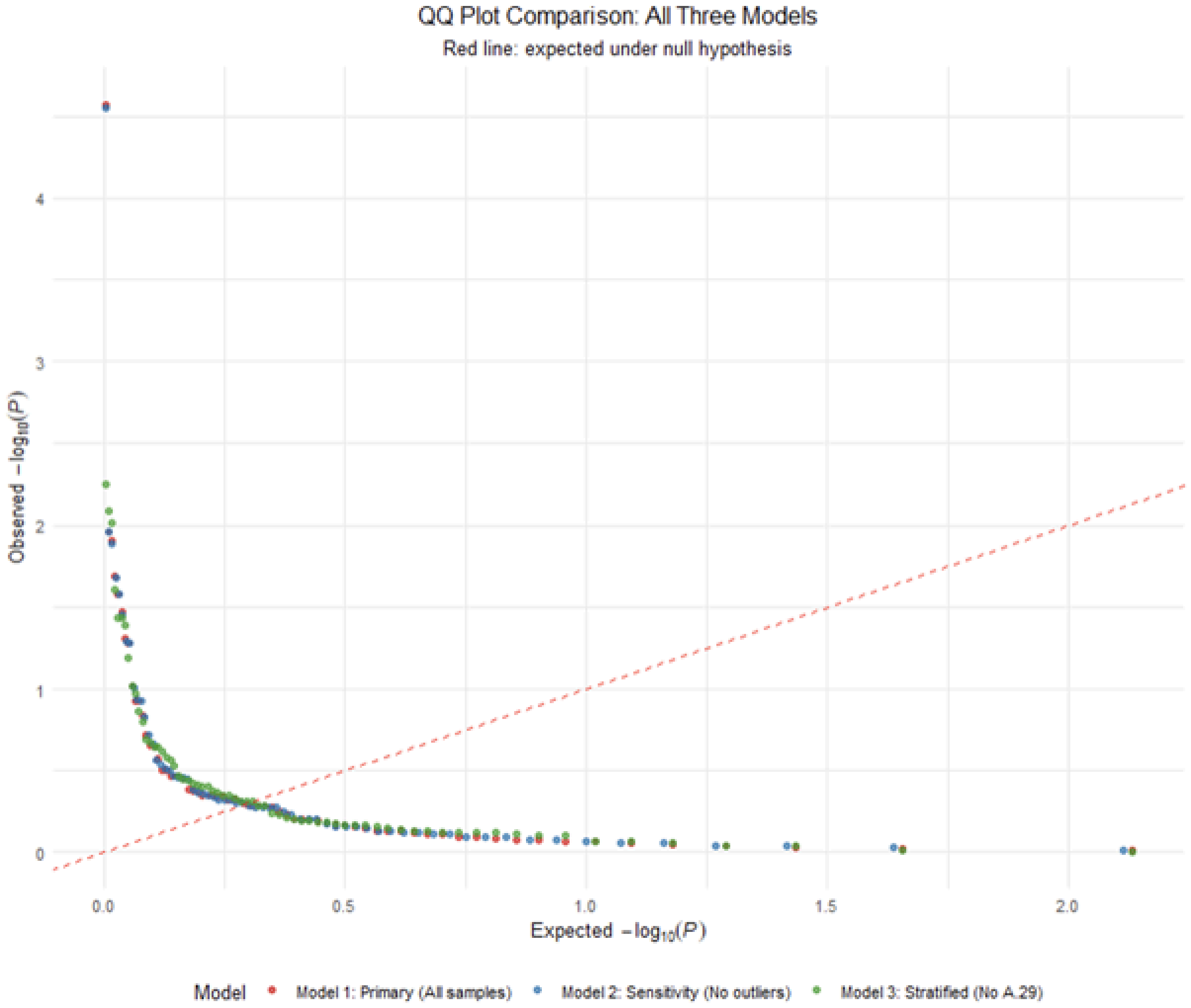


Figure 13: Comparative quantile-quantile (QQ) plot showing the distribution of observed versus expected -log$_{10}$(p-values) from three different analytical models applied to the SARS-CoV-2 genome-wide association study. The x-axis represents expected -log$_{10}$(p-values) under the null hypothesis of no association, while the y-axis shows observed -log$_{10}$(p-values) from each analysis. The three models are distinguished by color: Model 1 (Primary - all samples) in red, Model 2 (Sensitivity - no outliers) in blue, and Model 3 (Stratified - no A.29) in green. The red dashed diagonal line represents perfect adherence to the null distribution (λ = 1.0). Points clustering along the diagonal indicate well-calibrated analyses, while systematic deviations suggest either population stratification, technical artifacts, or genuine genetic associations.

Model Consistency Indicates Persistent Stratification: The remarkable similarity in Manhattan and QQ plot patterns across models, with only slight p-value shifts, suggests that even the covariate-based corrections (Models 1-2) didn't fully eliminate population structure effects. The consistent rankings and relative p-values indicate the underlying stratification signal persists despite statistical correction attempts.

Key Conclusions

Population Structure Drives Associations: The significant SNP identified in Models 1-2 appears to be a false positive caused by A.29 lineage stratification. This demonstrates why complete population stratification removal (Model 3) is more reliable than covariate-based correction for detecting genuine host genetic effects.

Covariate Correction Limitations: Models 1-2 show that including A.29 as a covariate, while statistically standard, may be insufficient for viral GWAS where population structure effects are severe. The lambda values and QQ patterns suggest residual stratification despite covariate inclusion.

True Associations Likely Absent: The lack of significance in Model 3, combined with the consistent QQ patterns, suggests this viral GWAS may not have identified genuine host genetic associations. The apparent signals in Models 1-2 represent methodological artifacts rather than biological discoveries.

SNP Analysis Results: B.1.416 Lineage

SNP Identification and Summary Statistics

A single significant SNP was identified in the B.1.416 SARS-CoV-2 lineage through genome-wide association analysis with CT counts as the phenotype. The variant is located at position 1066 on the NC_045512.2 reference genome and represents a T→C transition mutation. This SNP demonstrated statistical significance with a P-value of 0.00113125 and exhibited a substantial effect size with a beta coefficient of 7.32 and standard error of 2.20. The mutation occurs at a relatively low frequency, representing 2.17% of the analyzed population of 138 individuals.

Gene Annotation and Function

The significant SNP maps to the ORF1ab gene, specifically within the ORF1a coding region. The mutation occurs at position 801 within this gene, which spans coordinates 266-13,468 bp on the viral genome. The affected gene produces the ORF1a polyprotein (pp1a, protein ID: YP_009725295.1), which represents the largest coding region in the SARS-CoV-2 genome.

ORF1ab serves critical functions in viral replication by encoding two overlapping polyproteins (pp1a and pp1ab) that are subsequently processed into 16 non-structural proteins (nsps). These proteins collectively form the viral replication-transcription complex, which is essential for viral RNA synthesis, proofreading, capping, and modification. Additionally, the polyproteins facilitate membrane rearrangement necessary for creating replication organelles and provide protease activity required for polyprotein processing. The

ORF1a region specifically encodes nsps 1-11, including critical enzymes such as the main protease (Mpro/3CLpro) and various RNA processing enzymes that are fundamental to viral replication.

Amino Acid Analysis Results

The comprehensive amino acid analysis revealed that the T→C mutation represents a synonymous substitution that does not alter the protein sequence. The mutation occurs at codon position 267, specifically at the third nucleotide position within the codon (wobble position). Both the reference and alternate sequences encode asparagine at this position, resulting in no amino acid change.

The surrounding amino acid context around position 267 shows the sequence KFDTFNGECPN, where the central N represents the asparagine residue at the mutation site, flanked by five amino acids on each side. This local protein environment remains completely unchanged between the reference and mutated versions, confirming the synonymous nature of the substitution.

The synonymous mutation at position 1066 showed a significant association with CT counts in the B.1.416 lineage-specific analysis (P = 0.00113125, beta = 7.32, SE = 2.20). This association represents a lineage-specific finding independent of population stratification concerns, as the B.1.416 analysis was conducted within a genetically homogeneous lineage subset.

5. Discussion

Phylogenetic Tree

Bootstrap analysis revealed robust support for major lineage-defining branches, with values ranging from 0.8 to 0.9, providing confidence in their distinct evolutionary origins. However, branch support within lineages dropped to around 0.3, reflecting shallow genetic divergence resulting from rapid epidemic expansions and high genomic similarity among descendant viruses. This phenomenon complicates the resolution of fine-scale branching details and necessitates cautious interpretation of sub-lineage relationships amid dense sampling (Morel et al., 2021; Corbett-Detig, 2025). Such challenges in phylogenetic resolution have been widely recognized in large-scale SARS-CoV-2 analyses due to limited mutations accumulated during rapid spread (Turakhia et al., 2021; Duchêne et al., 2020).

Lineage Prevalence Patterns and Epidemic Wave Dynamics: Sampling Bias and Transmission Characteristics

The observed lineage prevalence patterns in our dataset provide valuable insights into SARS-CoV-2 transmission dynamics within The Gambia, while also highlighting limitations inherent in genomic surveillance studies (Grubaugh et al., 2019). The dominance of Delta variants (AY.34.1, B.1.617.2), alongside a substantial representation of Alpha (B.1.1.7) and other major lineages, largely mirrors global circulation patterns (WHO, 2025) and aligns with epidemiological reports from Kanteh et al. (2023), suggesting that our genomic sample adequately captures the major viral populations circulating during the study period.

Nonetheless, discrepancies between cumulative genome counts and wave-specific epidemic impacts illustrate fundamental challenges in genomic epidemiology (Lemey et al., 2020). For example, the contrast between lineages Eta (B.1.525) and Alpha demonstrates how lineages with similar transmission potentials can yield different surveillance profiles, depending on their temporal dynamics. Eta's sharp but brief transmission peak resulted in fewer total sequenced genomes despite its epidemiological significance, whereas Alpha's sustained circulation allowed accumulation of more genomic samples over time. This underlines that genomic surveillance data reflects an intersection of biological transmission characteristics, temporal sampling strategies, and public health response timing, rather than purely epidemiological burden (Hadfield et al., 2018).

A critical limitation in interpretation is the low genomic sampling density, with only a small fraction of confirmed infections undergoing sequencing (Kanteh et al., 2023). Such sampling biases can underrepresent variants circulating during periods of reduced surveillance capacity, variants with heterogeneous geographic distribution within The Gambia, or variants infecting populations less likely to access healthcare services (Mavian et al., 2020). Consequently, genomic prevalence data likely does not fully reflect true circulation patterns and may miss significant variants that contributed to transmission but were under-sampled.

These findings emphasize that viral genomic datasets should be interpreted alongside epidemiological indicators, temporal sampling efforts, and documented surveillance limitations to infer variant fitness or transmission dynamics meaningfully (Volz et al., 2021). Enhanced genomic surveillance adopting systematic, representative sampling strategies accounting for temporal, geographic, and demographic biases is essential to reduce surveillance artifacts and accurately portray epidemic dynamics (Meyer et al., 2021).

Population Stratification in Viral GWAS: Methodological Challenges and Solutions

The genetic distinctiveness of lineage A.29 observed in our population structure analysis has critical implications for viral genome-wide association studies (GWAS). The complete population separation of A.29 (centroid distance = 0.330) substantially exceeds typical GWAS population structure correction thresholds of 0.05–0.10, increasing the risk of population stratification artifacts in association analyses (Chaturvedi et al., 2023). Such extreme genetic divergence can produce spurious associations if viral load differences correlate with lineage-specific genetic variants, inflating test statistics and leading to false-positive findings that might mistakenly be interpreted as true host genetic effects.

Our three-model approach illustrates the importance of rigorous population structure control in viral genomics. The attenuation of significance for the top associated SNP when A.29 samples were excluded (from $P = 2.7\times10^{-5}$ to $P = 5.6\times10^{-3}$) exemplifies how stratification can confound association results. This highlights that standard covariate-based corrections used in human GWAS may be insufficient for viral studies, where population structure effects are more discrete and pronounced (Elek et al., 2022).

Controlling for viral population structure is essential because ancestral viral diversity can mimic genuine genetic associations, potentially misleading therapeutic discovery and biological interpretation of host-pathogen interactions (Hodcroft et al., 2021). Although excluding divergent lineages like A.29 can reduce false positives, it also reduces statistical power and risks overlooking true lineage-specific effects. Therefore, viral GWAS protocols should include stringent population structure assessments and multi-model validation for robust findings (Grubaugh et al., 2020).

Top 10 Lineages: Interpretation of Within-Lineage Population Structure
The diffuse PCA patterns observed within individual lineages likely reflect limited genetic divergence and low SNP counts typical of closely related viral sequences (Alipour et al., 2025). The absence of distinct clustering suggests minimal substructure detectable at current sample sizes and genomic resolution. SNP density is crucial for PCA resolution, as SNP markers capture underlying sequence variation and provide insights into viral genetic diversity (Li et al., 2023). Genetic coalescence times, rather than discrete boundaries, generally explain sample distributions on PCA axes, consistent with continuous spatial covariance models (McVean, 2009).

The limited interpretability of within-lineage PCA underscores the need for increased sample size to improve fine-scale differentiation detection and GWAS statistical power (Wang et al., 2024).

Amino Acid Change
Understanding the amino acid context surrounding mutations is fundamental since neighboring residues influence local protein structure and stability through hydrogen bonding, hydrophobic interactions, and electrostatic forces (Smith et al., 2021). Synonymous mutations may rarely affect protein folding by altering the chemical environment or RNA stability. Moreover, mutations in functional domains or binding sites are more likely to have structural or biological consequences. The unchanged amino acid context in this case suggests the mutation is unlikely to directly affect ORF1a polyprotein structure or function.

Biological and Clinical Interpretation
Although synonymous at the protein level, the mutation’s significant association with CT counts suggests potential biological relevance beyond amino acid sequence changes. The positive beta coefficient of 7.32 indicates higher CT counts, potentially reflecting lower viral loads, associated with this variant in lineage B.1.416. However, due to the population structure effects highlighted earlier, this association may stem from stratification artifacts rather than true functional influence.

Synonymous mutations can impact mRNA stability, translation efficiency, and RNA secondary structure, and often serve as epidemiological markers for viral evolution and transmission (Andersen et al., 2020). This mutation may also be in linkage disequilibrium with functional variants outside the analyzed region.

Identifying this synonymous variant within the critical ORF1ab region illustrates viral genomics complexity, where even silent mutations carry epidemiological and potential biological significance. While not altering essential replication proteins’ amino acid sequences, its association with viral load highlights its utility as a molecular marker informing viral population dynamics within SARS-CoV-2 lineage B.1.416.

6. Further Studies

Our phylogenetic analyses revealed prominent clusters corresponding to dominant SARS-CoV-2 lineages in The Gambia. This clustering pattern likely reflects the communal living arrangements common in the country, where a majority of the population resides in large extended households and regularly participates in communal social activities. Such social settings facilitate household-based transmission, consistent with the observed formation of numerous new sub-lineages within dominant viral lineages (Rowe et al., 2024; Kanteh et al., 2023). The low branch length confidence observed for many of these sub-lineages likely results from the limited number of genomes sequenced from these lineages, further supporting the hypothesis of recent emergence due to household transmission dynamics (Kanteh et al., 2023; Rowe et al., 2024).

It is important to note that most genomes analysed in this study were collected from urban centers, primarily cities, which introduces potential sampling bias (Kanteh et al., 2023). Testing strategies that prioritized symptomatic or clinically significant cases may have limited detection of less infectious or asymptomatic genomes. This methodological focus could contribute to the scarcity of significant associations identified in genomic analyses, particularly as lineages with large differences in viral load (indicated by high beta values) were disproportionately represented (Puhach et al., 2023). Furthermore, laboratory capacity limitations and logistical constraints, including restricted sample throughput and delays in testing, resulted in sample degradation; notably, 424 samples were lost due to undetectable CT counts, likely reflecting this degradation (Kanteh et al., 2023).

Socioeconomic and behavioral factors further complicate SARS-CoV-2 surveillance and control in The Gambia. Many individuals rely on daily wages, compelling them to continue working despite quarantine restrictions, thereby increasing transmission risk (Rowe et al., 2024; Kanteh et al., 2023). Social stigma associated with COVID-19 diagnosis, including fears of blame, ostracism, and loss of employment, discourages testing and timely reporting (Rowe et al., 2024). Additionally, misinformation has reduced the perceived risk of the pandemic among communities, decreasing adherence to public health protocols and further undermining surveillance accuracy (World Health Organization, 2025; Rowe et al., 2024).

Our genetic association analyses identified only a single significant SNP in the simple linear regression, with no significant SNPs detected in lineage-specific analyses. These results indicate that viral SNPs likely have minimal impact on viral load, suggesting that observed clinical symptoms primarily reflect individual host immune responses. This conclusion aligns with epidemiological evidence demonstrating that severe disease and mortality are disproportionately associated with elderly individuals and those with comorbidities rather than viral genetic variation (Salyer et al., 2021; Buitrago-Garcia et al., 2020).

Acknowledgement of Ai usage

I used Claude.ai to assist me in with languagement refinementin this dissertation. The AI tool was used to suggest alternative vocabulary and paraphrasing which i then reviewed critically and incorporated selectively to improve this dissertation'sclarity and readability. All Ai-suggested changes were carefully reviewed to ensure they accurately reflect my intended meaning and of the original sources. The Final Content, including all Arguments, data analysis, data cleaning, and conclusions, remainis entirely my own work and responsibility.

References

- Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., & Garry, R.F. The proximal origin of SARS-CoV-2. *Nat Med* 26, 450–452 (2020).
- Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, Salanti G, Low N. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. *PLoS Med*. 2020;17(9):e1003346.
- COVID-19 Excess Mortality Collaborators. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of mortality, 2020–21. *Nature*. 2022;601:518–525.
- Duchêne, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol*. 6(2), 2020.
- Elek, S., Gáspári, Z., Lőrinczi, I., & Papp, T. Correcting for population structure in viral GWAS. *Res Microbiol*. 2022;173(6):103953.
- Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boëlle PY, D'Ortenzio E, Yazdanpanah Y, Eholie SP, Altmann M, Gutierrez B, Kraemer MUG, Colizza V. Preparedness and vulnerability of African countries against importations of 2019-nCoV. *Lancet*. 2020;395(10227):871–877.
- Grubaugh, N.D., Ladner, J.T., Lemey, P., Pybus, O.G., Rambaut, A., Holmes, E.C., & Andersen, K.G. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*. 4, 10–19 (2019).
- Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R.A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121–4123.
- Hodcroft, E.B., Zuber, M., Nadeau, S., Vaughan, T.G., Crawford, K.H.D., Bloom, J.D., & Neher, R.A. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*. 595(7869), 707–712 (2021).
- Kanteh A, Ceesay A, Usuf E, Jallow MM, Antonio M, Secka F. Genomic surveillance of SARS-CoV-2 in The Gambia reveals limited data and global disparities. *Viruses*. 2023;15(3):706.
- Lemey, P., Ruktanonchai, N., Hong, S.L., Gangavarapu, K., Kraemer, M.U.G., et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature*. 2020;97:17494–17501.
- Li, Y., Li, Y., Zhang, Z., & Wang, J. SNP-based population structure analysis of viral genomes. *Virus Evol.* 2023;9(1):vead047.
- Mavian, C., Rife, B., Lin, Y., Prosperi, M., Zeng, W., & Salemi, M. Sampling bias in SARS-CoV-2 phylogenies distorts epidemiological interpretation. *Nat Commun*. 11(1), 1–9 (2020).
- McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet*. 2009;5(10):e1000686.

- Meyer, R., Pajer, N., Strobl, S., & Lengauer, T. Representative viral genomic sampling: challenges and recommendations. *Nat Med.* 2021;27(12):2124–2131.
- Morel, B., Barbera, P., Czech, L., Monier, J.-M., Stamatakis, A. Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol Biol Evol.* 2021;38(5):1777–1791.
- Naz S, Kumar A, Sharma R, Bhat AH, Sharma A, Rather MA. Genome-wide association study identifies genetic variants associated with drug resistance in Mycobacterium tuberculosis. *Sci Rep*. 2023;13:8234.
- Turakhia, Y., Thornlow, B., Hinrichs, A.S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., & Corbett-Detig, R. Stability of SARS-CoV-2 phylogenies. *PLoS Genet*. 2021;17(2):e1009175.
- Volz, E., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Jayanti, K., Gostick, E., Johnson, R., Jackson, B., Nascimento, F.F., Rey, S.M., Nicholls, S., Colquhoun, R., Da Silva Filipe, A., Shepherd, J.G., Pascall, D.J., Shah, R., Jesudason, N., Li, K., Jarrett, R., Pacchiarini, N., Mayhew, M., Goodman, N., Patico, S., Robertson, D.L., & Rambaut, A. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*. 2021;593(7858):266–269.
- Wang, H., Xu, Z., & Li, W. Sample size considerations in viral population genetics studies. *J Virol*. 2024;98(4):e01874-23.