Analyzing SARS-CoV-2 in Pennsylvania Wastewater Samples: SNP Distribution, Temporal Trends, and Correlation with Population

Alyssa Ramirez

13 December, 2024

Background and Overview

The monitoring and detection of SARS-CoV-2 RNA in wastewater has emerged as another tool in aiding the surveillance of SARS-CoV-2 in populations. Wastewater surveillance provides non-invasive, cost-effective methods for insights and trends in the spread in SARS-CoV-2 RNA within a region. This approach has been useful in understanding hot spots, strain development, drift in other regions and overall helpful insights to the public health. Wastewater surveillance has been proven to be efficient in identifying patterns of viral transmission, monitoring genetic drift, and supporting epidemiological studies.(Peccia *et al.*, 2022) This study utilizes Pennsylvania wastewater surveillance systems (PAWSS) data, a statewide surveillance system that collects waste samples in diverse communities. Data from this study were collected from multiple collection sites: wastewater treatment facilities (WWTFs), individual, buildings or congregate settings, correctional facilities, dormitories, long-term care facilities and schools.(Pennsylvania Department of Health, 2024) The data focuses on population sizes, sequencing information, temporal trends and waste water sample collecting methods. Wastewater surveillance is specifically effective for identifying SARS-CoV-2 viral RNA, including non-traditional cases that could potentially go unnoticed by clinical testing. These also include even asymptomatic and oligosymptomatic samples. Wastewater surveillance offers a specific advantage over traditional clinical testing by using large-scale monitoring of SARS-CoV-2 RNA with minimal resources. As stated, this approach "only needs to test one sample for mass screening of SARS-CoV-2. Clinical testing, on the other hand, may require 10,000 individual tests and bears the associated costs for sample collection and investments in high-throughput infrastructure" (Peccia *et al.*, 2022) This highlights the cost-effectiveness and efficiency of wastewater surveillance as a whole. This study focuses on the essential factors that influence the reliability and effectiveness of wastewater surveillance, including sampling methods, population size, and temporal trends. Sampling methods, such as composite and grab sampling. These methods differ in their ability to provide viral RNA signals. Inspecting population size impacts the concentration of viral RNA and the potential risk of diluting detectable viral RNA in larger populations. On the other hand, temporal trends reveal how SARS-CoV-2 RNA levels can vary overtime and how population dynamics influence viral community infection rates. Bash and Rstudio technologies were used to create visual representations of data.

Research Question

The primary aim of this study is to investigate how sampling methods (composite vs. grab), collection periods, and population size influence the detection and reliability of SARS-CoV-2 RNA in wastewater. By analyzing sequencing metadata, this research seeks to uncover factors that impact viral RNA detection, with implications for improving public health surveillance systems.

Methods

Wastewater data collection Wastewater samples were collected from various locations across Pennsylvania from the Pennsylvania Wastewater Surveillance System (PAWSS). Samples were gathered used 2 types of filtering mechanisms, composite and grab sampling methods. Composite samples were collected over a duration of 24 hours and considered a "average" of waste. Compared to grab sampling, which is a one-time sample.(Kmush *et al.*, 2022)

RNA Purification and Sequencing Following extraction from composite and grab samplings, reverse transcription to synthesize complementary DNA (cDNA) is used on the purified RNA. The purified RNA is synthesized using Reverse Transcription Polymerase Chain Reaction (RT-PCR). The cDNA is amplified and prepared for sequencing using the Illumina COVIDSeq™ Assay, which targets the SARS-CoV-2 genome, including "regions prone to single nucleotide polymorphisms (SNPs)."Sequencing was performed on an Illumina platform, generating high-throughput data with robust coverage of genomic regions of interest."(Surveillance of infectious disease through wastewater sequencing: Detect sars-cov-2 variants and other respiratory viruses in the community, 2024) To eliminate contaminants and ensure accurate data, quality control metrics were applied to sequencing reads. The sequencing metadata were obtained from the National Library of Medicine BioProject database (Accession: PRJNA1039783), submitted by the Pennsylvania Department of Health Bureau of Laboratories.(Health Bureau of Laboratories Submission Group, 2023)

Data Processing and Analysis Sequencing data were processed using a bioinformatics makefile pipeline, designed to analyze the genomic metadata. GFF annotation files and VCF datasets were utilized to identify SNPs and functional data. The pipeline consisted of the following key steps:

1.GFF Annotation Parsing: The genomic feature file (GFF) was read and parsed to extract gene annotations, enabling identification of specific SARS-CoV-2 genomic regions.[2023Djaffardjy]

2.Gene Extraction: Relevant genomic features were filtered to produce a table of gene names and coordinates for downstream analyses.[2023Djaffardjy]

3.VCF File Integration: Variant call format (VCF) files were parsed, cleaned, and merged with metadata to create a unified dataset containing SNP-level information for all samples.[2023Djaffardjy]

4.SNP Annotation: SNPs were annotated with gene information by merging the tidy VCF data with the gene table and metadata.[2023Djaffardjy] The pipeline worked effortlessly due to the `Bash` (Project, 2024) and `Rstudio`(Team, 2023) function scripts previously provided by Professor Zimmerman, University of San Francisco.

Data Visualization and Statistical Analysis Visualizations and analyses were conducted using R programming. R packages that were used are data manipulation are `ggplot2` (Wickham, 2016), `dplyr` (Wickham *et al.*, 2023) and `tidyverse` (Wickham and others, 2023b). `Rmarkdown` was used for report generation and rendering Markdown files into HTML and pdf. (Xie and others, 2023) The R package `ggthemes`(Arnold, 2022) was used to create figures illustrating SNP distribution, temporal trends, and relationships between key variables, such as population size and RNA detection levels. To create color on the figures `RColorBrewer`(Neuwirth, 2023) was used hand in hand with `ggthemes`. The R package `magrittr` (Bache and Wickham, 2023) was used to utilize the pipe operator in code. To manipulate and visualize variant call format (VCF) files, `vcfR`(Knaus and others, 2023) was used. To enable report generation and integration of R code `knitr` was used.(Xie, 2023) To make sure all of the pipeline worked correctly `testthat` was used.(Wickham and others, 2023a) Lastly, figures and tables were generated to summarize and visualize the data. OpenAI's ChatGPT (OpenAI, 2024) was used to create all figures and table.

Results

Impact of Sampling Methods on Viral RNA Detection The analysis of composite and grab samples showed the distinct differences in RNA detection consistency and data accuracy. Composite samples, which were collected over the course of 24 hours, provided a higher average in sequencing coverage compared to the grab samples.(Kmush *et al.*, 2022) As showed in Figure 5, the distribution of sampling methods showed a drastic difference in count. Composite sampling minimizes temporal variability which results in a overall higher value in representative number of SARS-CoV-2 RNA levels. On the other hand, grab samples collect more at a single time but show a major deduction in sequencing coverage and high variability. Which is likely due to the temporal fluctuations of viral RNA levels in wastewater.(Kmush *et al.*, 2022)

Summary of Sequencing Coverage by Sampling Method Table 1 summarizes the sequencing coverage for different sampling matrices and methods. Composite samples consistently outperformed grab samples in sequencing coverage, reinforcing their reliability for wastewater surveillance. This table is used to support figure 5, the table provides a detailed statistical metric on sequencing coverage for composite and grab sampling methods. While Figure 5 visible demonstrates the distribution coverage, Table 1 highlights quantitative differences.

Temporal Trends in Sequencing Coverage Temporal fluctuations in sequencing coverage were observed across Pennsylvania, as shown in Figure 2 peaks in the figure show sequencing coverage corresponding to periods of increases viral prevalence in the Pennsylvania region. This suggests that the alignment with SARS-CoV-2 case surges are representative. This also highlights the utilities of wastewater surveillance in real-time epidemiological trends. Even though, some moths exhibited a lower coverage, which could potentially explain reduced viral shedding or sample processing errors.

Population Size and RNA Detection Levels The relationship between population size and overall RNA detection levels is represented in Figure 3. A positive correlation between the two variables can suggest that larger population sizes contribute to more consistent viral RNA in wastewater. However, the relationship is complex and influenced by multiple factors. As stated, "Although the association between population size and viral RNA concentrations may be driven by absolute number of COVID-19 cases, it is important to state that RNA concentrations may also be affected independently. As population size increases, sewage flow increases due to increased water usage causing dilution of viral RNA." (Carrat *et al.*, 2022) This further supports what is seen in Figure.4, where there is a small negative correlation between population size and average sequencing read lengths. These figures can support the idea that RNA fragmentation leads to an increase in sample diversity. Which can be insufficdent in producing accurate and reliable data while in a larger population size.

Genomic Variability Across SARS-CoV-2 Genes Genomic variability is seen in Figure 1, this figure represents each gene seen in the data associated with its SNP count. The figure has revealed that there is variability across the SARS-CoV-2 genes but the S gene prevailed over all other genes. This finding is consistent with the critical role of the spike protein in viral infectivity and host interaction, making it a hotspot for mutations. This observation aligns with global studies on SARS-CoV-2 evolution, emphasizing the need for continued surveillance of the S gene to monitor emerging variants.(Larsen *et al.*, 2022)

Variant Distributions in Pennsylvania The distribution of reference and alternate genome variants in Figure 6 show a predominance of thymine (T) as the most frequent alternate genome nucleotide. This finding may reflect regional patterns in viral evolution, further underscoring the importance of localized genomic surveillance.

Discussionn

In conclusion, data suggest that sampling methods, population size and temporal trends significantly influence the detection of SARS-CoV-2 in wastewater. When it came down to which sampling method was deemed reliable and most effective, composite samples won over grab samples. Population size and RNA detection did not correlate and had a small role against RNA detection showing trends that were not predicted before going through the data. Although the population size data was not helpful to the overall study, it was helpful in the visualization of the S genome, seen as the most common in the population. Temporal trends indicated variation over time in sequencing coverage that could possibly reflect changes in the SARS-CoV-2 genome.

Figures

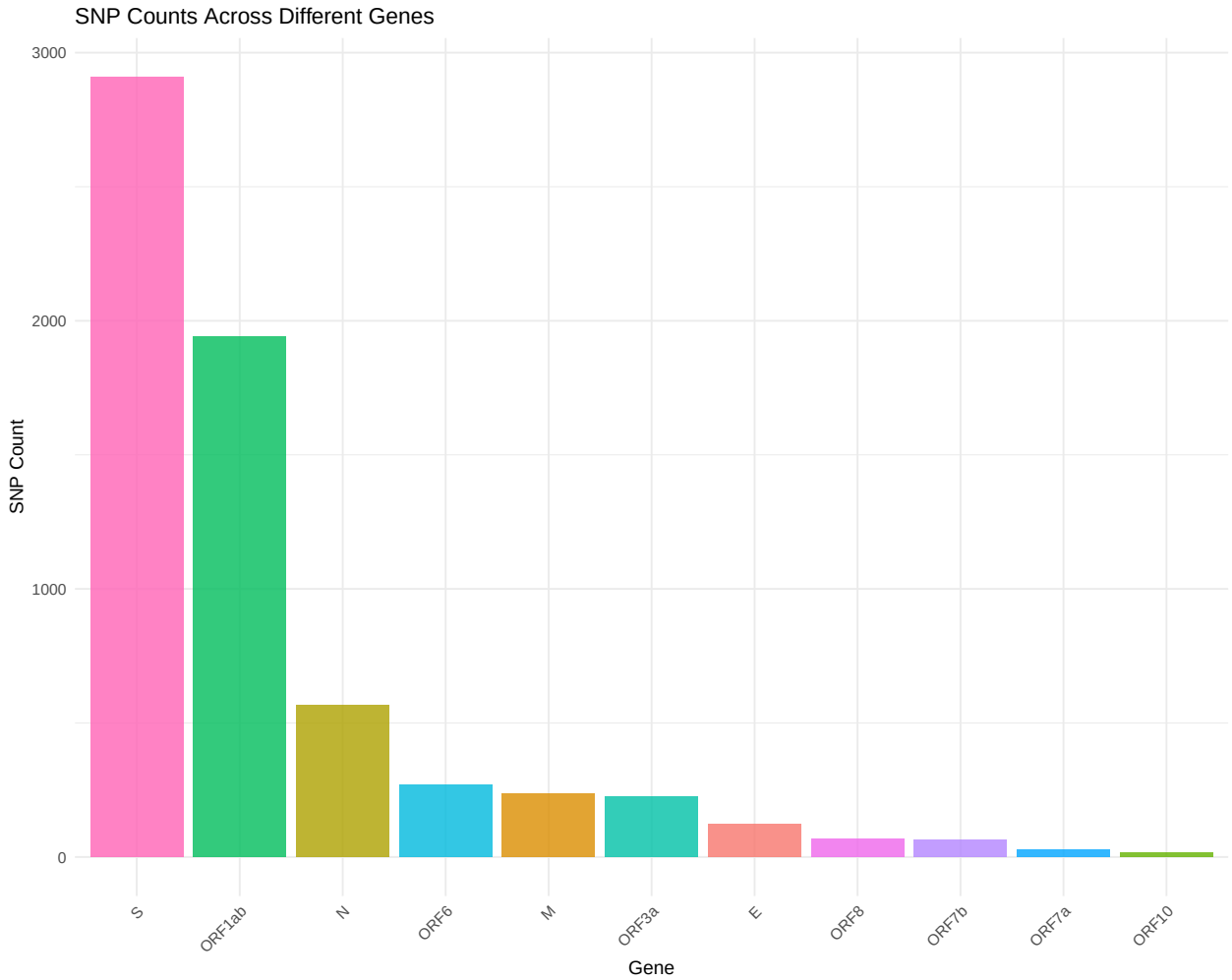

Figure 1:Figure: Bar plot showing the number of single nucleotide polymorphisms (SNPs) identified across different genes of SARS-CoV-2. The genes are labeled on the x-axis, and the y-axis indicates the count of SNPs. This figure highlights which genes exhibit the highest mutation frequency, with the S gene having the largest number of SNPs, suggesting greater genetic variability in this region. Such information is crucial for understanding the genetic diversity and evolution of the virus in wastewater samples

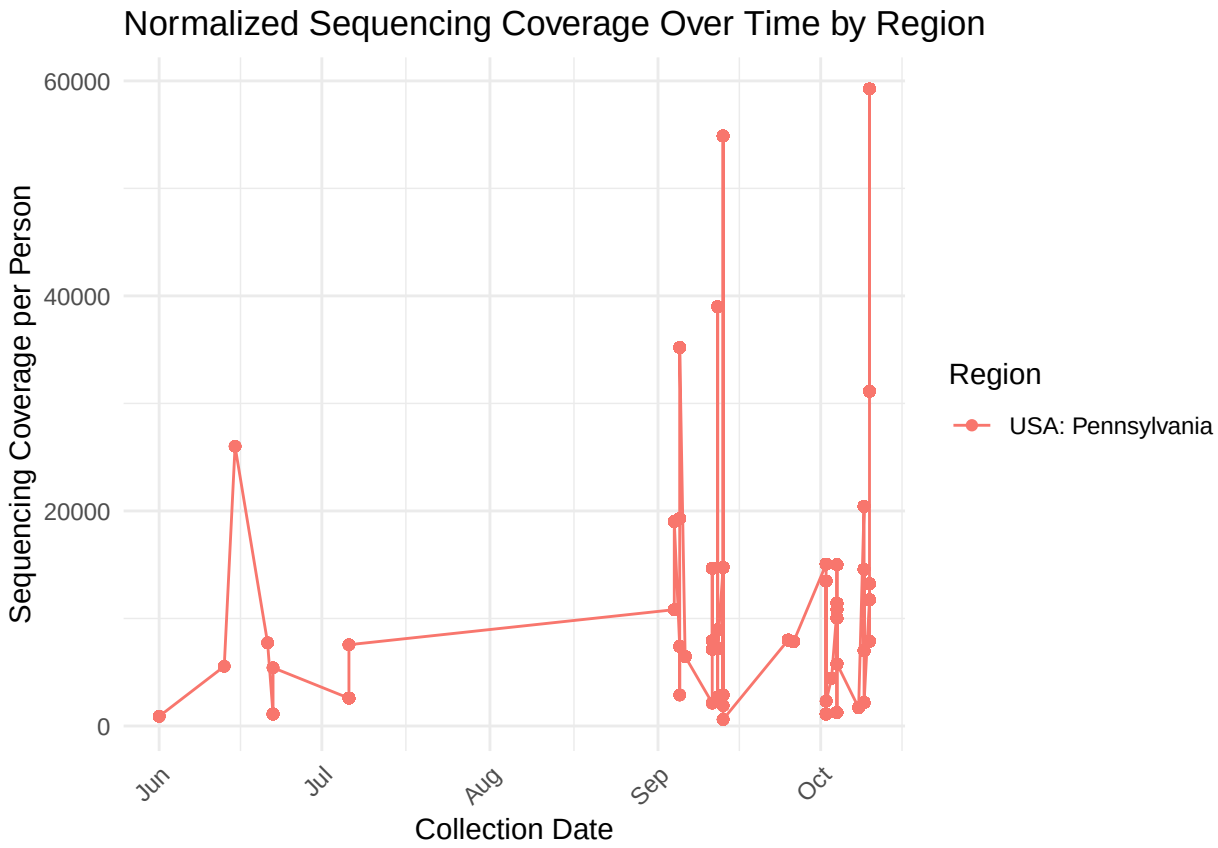

Figure 2:Temporal Trends in Normalized Sequencing Coverage Over Time in Pennsylvania The plot shows normalized sequencing coverage (SARS-CoV-2 RNA levels per person) over time for wastewater samples collected in Pennsylvania. Peaks in sequencing coverage may reflect variations in viral prevalence, sampling methods, or population size adjustments. The data are normalized to account for differences in population size represented by each sample.

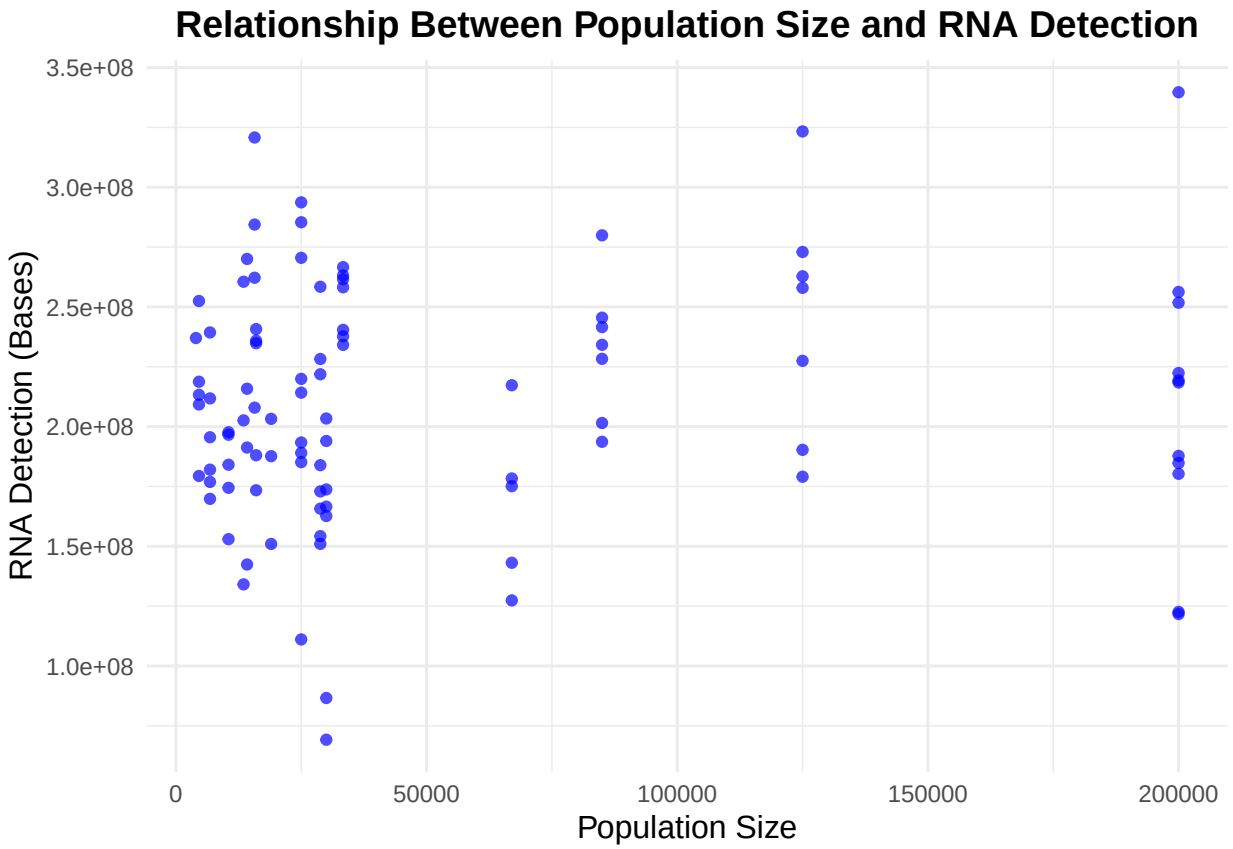

Figure 3:Scatter plot showing the relationship between population size and RNA detection levels (measured as bases) in wastewater samples. Each point represents a data sample, and the distribution illustrates whether population size affects RNA detection variability.

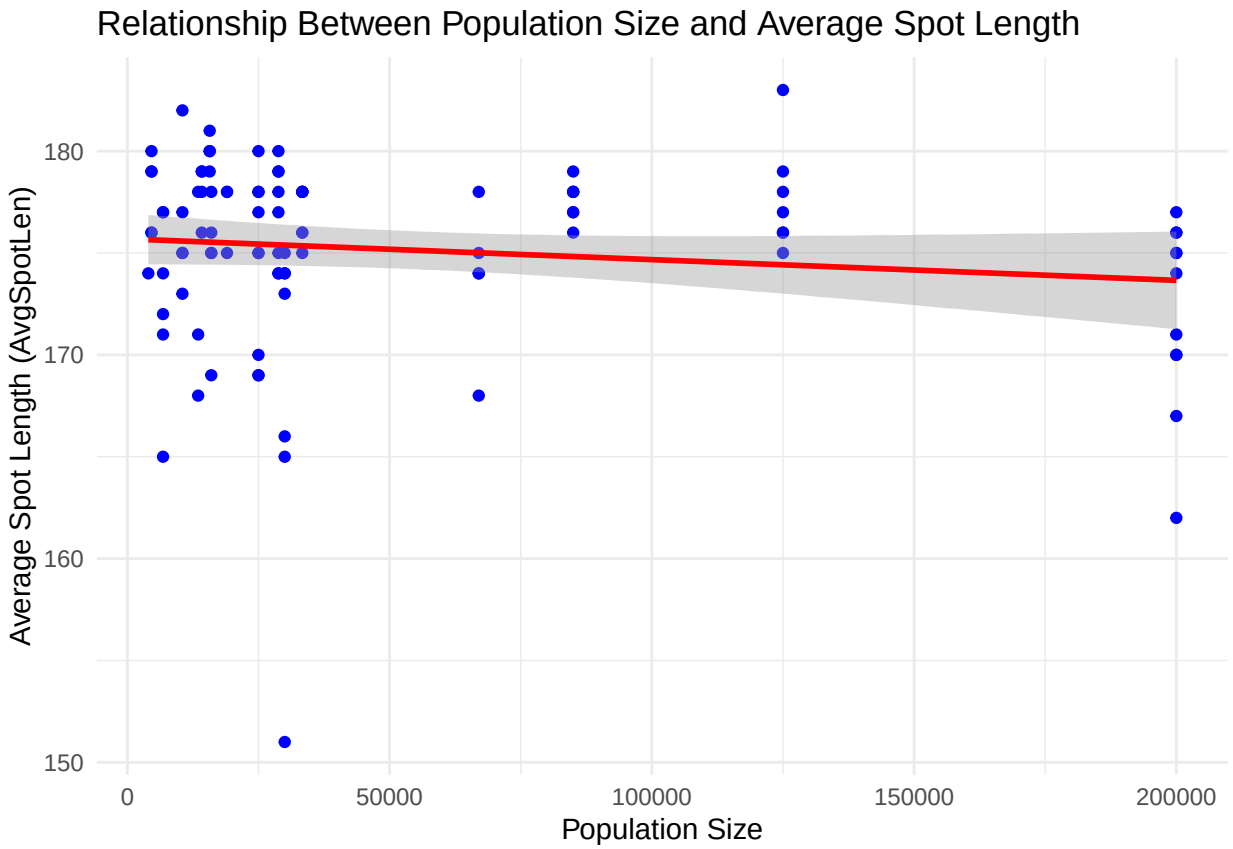

Figure 4: Scatter plot showing the relationship between population size and average spot length of sequencing reads. Each blue point represents a sample, and the red regression line indicates the overall trend. The figure suggests a slight negative correlation, implying that as population size increases, the average spot length may decrease slightly.

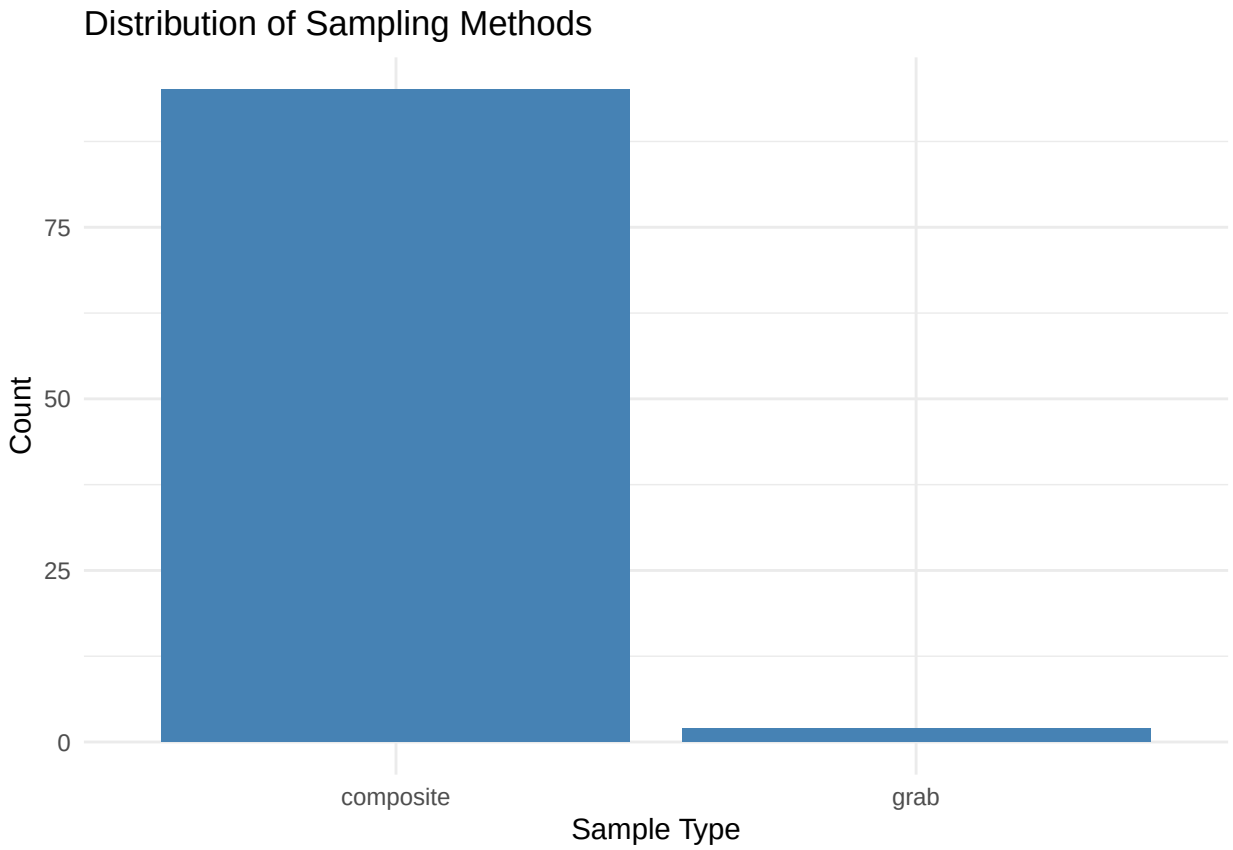

Figure 5: Bar plot showing the distribution of sampling methods. The majority of samples were collected using composite sampling, while grab sampling was used much less frequently.

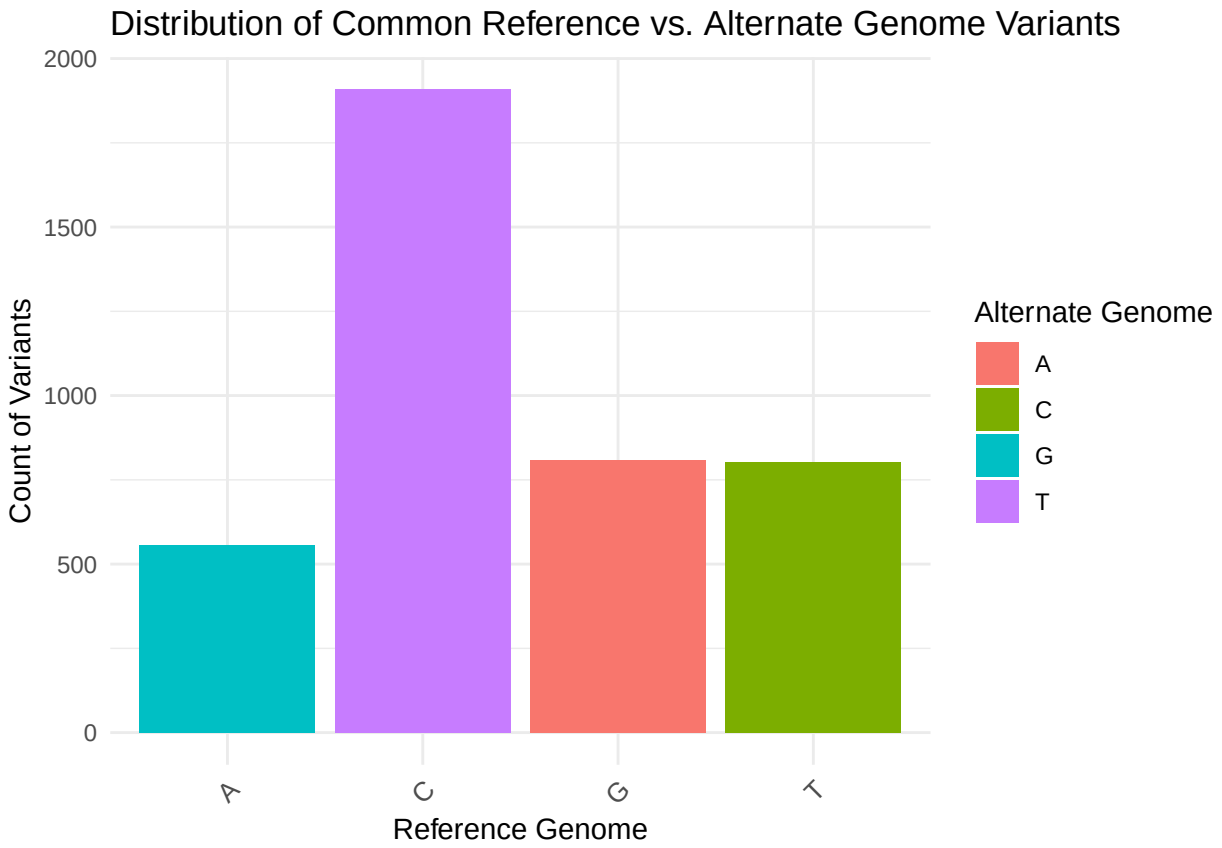

Figure 6: Distribution of Common Reference vs. Alternate Genome Variants. This bar plot displays the counts of variants classified by their reference genome nucleotide and corresponding alternate genome variants. Variants are categorized as A, C, G, and T based on their alternate genome.

Tables

Table 1: Summary of Sequencing Coverage by Sampling Method

Sample Matrix	Sample Type	Mean Coverage (Bases)	Median Coverage (Bases)	Minimum Coverage (Bases)	Maximum Coverage (Bases)	Number of Samples
post grit removal	composite	231564431	245510106	142390614	323318001	2212
raw wastewater	composite	231422246	237004344	69190509	339673573	4351
raw wastewater	grab	222379868	222379868	222379868	222379868	66

Table 1:Summary of Sequencing Coverage Metrics by Sampling Method.This table presents key statistics such as mean, median, minimum, and maximum coverage—stratified by sample type (composite vs. grab) and sample matrix (e.g., raw wastewater, post-grit removal).

Sources Cited

Arnold,J.B. (2022) Ggthemes: Extra themes, scales and geoms for 'ggplot2'.

Bache,S.M. and Wickham,H. (2023) Magrittr: A forward-pipe operator for r.

Carrat,F. *et al.* (2022) Seroprevalence of sars-cov-2 in france by july 2021: Results from the second nationwide epicov survey. *PLOS Medicine*, **19**, e1003896.

Health Bureau of Laboratories Submission Group,P.D. of (2023) Sequencing of sars-cov-2 rna from wastewater influent samples collected in pennsylvania for genomic surveillance.

Kmush,B.L. *et al.* (2022) Comparability of 24-hour composite and grab samples for detection of sars-cov-2 rna in wastewater. *FEMS Microbes*, **3**, 1–5.

Knaus,B.J. and others (2023) VcfR: Manipulate and visualize vcf data.

Larsen,D.A. *et al.* (2022) Wastewater surveillance for sars-cov-2 rna in denmark, 2020–2021. *Environmental Health Perspectives*, **130**, 065001.

Neuwirth,E. (2023) RColorBrewer: ColorBrewer palettes.

OpenAI (2024) ChatGPT: Advanced ai language model.

Peccia,J. *et al.* (2022) Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nature Biotechnology*, **40**, 181–186.

Pennsylvania Department of Health (2024) Pennsylvania wastewater surveillance system (pawss).

Project,G. (2024) Bash (bourne again shell).

Surveillance of infectious disease through wastewater sequencing: Detect sars-cov-2 variants and other respiratory viruses in the community (2024) Illumina.

Team,R. (2023) RStudio: Integrated development for r.

Wickham,H. (2016) Ggplot2: Elegant graphics for data analysis Springer-Verlag New York.

Wickham,H. *et al.* (2023) Dplyr: A grammar of data manipulation.

Wickham,H. and others (2023a) Testthat: Unit testing for r.

Wickham,H. and others (2023b) Tidyverse: Easily install and load the 'tidyverse'.

Xie,Y. (2023) Knitr: A general-purpose package for dynamic report generation in r.

Xie,Y. and others (2023) Rmarkdown: Dynamic documents for r.