FINAL RESEARCH REPORT:

THE EVOLUTION OF SARS-COV-2 (COVID-19) VARIANTS: ANALYSIS OF LARGE GENOMIC DATASETS

12 August 2025

ADVISORS

Dr. Ryan Gregory, Professor, Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada. (Supervisor and Biological Expertise)

Dr. Gurjit Randhawa, Assistant Professor, School of Computer Science (SOCS), University of Guelph, Guelph, ON N1G 2W1, Canada. (Co-Advisor and Computational Advisor)

PREPARED BY

Pavani Addepalli (1326277)

For BINF*6999: Bioinformatics Master's Project

Table of Contents

1. Abstract

SARS-CoV-2 genomic surveillance through wastewater provides an early warning system for detecting emerging variants, complementing clinical sequencing. This study aimed to identify cryptic and novel SARS-CoV-2 lineages in airport wastewater, a unique sampling source that captures a transient, globally diverse population. A total of 109 metagenomic wastewater samples collected from Toronto International Airport between November 2022 and October 2023 were processed through a comprehensive bioinformatics workflow, which included quality control, genome alignment, variant calling, cryptic mutation panel screening, entropy analysis, lineage assignment, and phylogenetic reconstruction. Compared with community-based studies (Gregory et al., 2022; Suarez et al., 2025), airport wastewater exhibited a mutation landscape characterized by a predominance of novel single-nucleotide variants, with limited evidence of convergent evolution. Screening revealed high-frequency cryptic-associated mutations (e.g., G446S, P681H, Y505H, R346S), but no complete cryptic lineage convergence. Shannon entropy at cryptic sites was consistently low, indicating single-variant dominance per sample. Lineage analysis showed early prevalence of Lineage A and BA.5-like variants, followed by the emergence of BA.2.86 subclades (including 24F and 25-series) and several unassigned genomes, suggesting possible novel or recombinant lineages. Phylogenetic analysis confirmed diversification within BA.2.86 and clustering with divergent global clades. These findings highlight the value of airport wastewater surveillance as a global sentinel system for early detection of rare, imported, or emerging SARS-CoV-2 variants before they establish in local populations.

2. Introduction

The ongoing COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has underscored the need for robust, real-time surveillance systems capable of detecting emerging variants before they establish in local populations. Since its emergence in late 2019 [1], SARS-CoV-2 has diversified into numerous genetic lineages through the accumulation of mutations, particularly in the spike (S) gene, which can affect viral transmissibility, immune evasion, and pathogenicity [2,3]. Some of these mutations enable the virus to escape neutralizing antibodies [4], adapt to new hosts [5], or persist in immunocompromised individuals, thereby creating reservoirs for the emergence of novel variants [6].

Wastewater-based epidemiology (WBE) has emerged as a cost-effective, non-invasive tool to monitor SARS-CoV-2 prevalence and genetic diversity at the community level [7]. By capturing viral RNA shed in feces and other bodily fluids, WBE integrates signals from both symptomatic and asymptomatic individuals [8]. Multiple studies have demonstrated its value in tracking infection trends [9], detecting variants of concern (VOCs) before clinical confirmation [10], and uncovering cryptic SARS-CoV-2 lineages absent from global genomic databases such as GISAID [11,12].

Cryptic lineages are operationally defined as SARS-CoV-2 genetic variants with unusual mutation patterns that are rarely or never observed in contemporaneous clinical samples. They have been hypothesized to originate from long-term infections in immunocompromised hosts [13], unsampled geographic populations, or even non-human reservoirs [14,15]. Community-level WBE studies in New York City [11] and Missouri [12] have identified persistent cryptic lineages with convergent spike mutations such as G446S, E484A, and Y505H, which may confer immune escape properties

[4]. However, these studies have primarily relied on municipal wastewater representing relatively stable populations, which may facilitate the persistence and recurrence of cryptic variants.

Airport wastewater represents a fundamentally different surveillance context. Airports act as global transit hubs, receiving passengers from diverse geographic origins. This results in a highly transient, heterogeneous viral signal that can capture globally circulating variants, including rare or emerging lineages, before they appear in local communities. Despite this potential, few studies have explored the genomic diversity of SARS-CoV-2 in airport wastewater, and even fewer have assessed its capacity to detect cryptic mutations or novel lineages [16]. This represents a critical gap in current wastewater surveillance frameworks, which often overlook high-traffic, globally connected sites.

From a bioinformatics perspective, most prior WBE cryptic lineage studies have employed targeted sequencing of the spike receptor-binding domain (RBD) [12] or relied on public whole-genome datasets from municipal sources [17]. These approaches, while effective for specific aims, may miss broader mutation landscapes and structural changes occurring outside the RBD. A metagenomic, whole-genome approach allows for comprehensive detection of single-nucleotide variants (SNVs), insertions/deletions (indels), and genome-wide patterns of diversity [18,19], but remains underutilized in airport-based surveillance.

This study addresses these knowledge gaps by applying a whole-genome metagenomic sequencing approach to 109 wastewater samples collected from Toronto International Airport between November 2022 and October 2023. We developed and applied a bioinformatics pipeline that included quality control, read alignment, variant calling, cryptic mutation panel screening [17], Shannon entropy analysis, lineage assignment using Pangolin and NextClade, and phylogenetic

reconstruction with IQ-TREE2. By comparing results with community-based [12] and global wastewater [17] studies, we aimed to:

- Characterize the mutation landscape of SARS-CoV-2 in airport wastewater, including the prevalence of novel versus convergent mutations.
- Detect and evaluate cryptic-associated mutations and assess whether full cryptic lineage convergence occurs in this transient population.
- Assess intra-sample diversity at cryptic mutation sites using Shannon entropy.
- Identify and track SARS-CoV-2 lineages over time, with emphasis on novel or recombinant forms.
- Place airport wastewater findings in a global phylogenetic context to evaluate its role as an early-warning sentinel in variant surveillance.

By filling this gap, our study provides both methodological and epidemiological insights into the utility of airport wastewater as a sentinel surveillance platform for global SARS-CoV-2 genomic diversity.

3. Methods

3.1 Sample Collection and Study Context

A total of 109 SARS-CoV-2–positive wastewater samples were collected from Toronto Pearson International Airport between November 2022 and October 2023 as part of a genomic surveillance program targeting transient, globally mixed traveler populations. This sampling design contrasts with community-based wastewater surveillance efforts (Gregory et al., 2022; Suarez et al., 2025), which focus on stable local contributors. Paired-end metagenomic sequencing data were obtained

from Larry and Opeyemi's research group, which had processed the wastewater samples. The raw reads were subsequently used for downstream bioinformatics analysis.

3.2 Reference Genome Preparation

The SARS-CoV-2 Wuhan-Hu-1 reference genome (GenBank accession NC_045512.2) and associated GFF3 annotation were obtained from NCBI. The genome was indexed for read alignment and variant calling using BWA (v0.7.17) and samtools (v1.17).

bwa index NC_045512.2.fasta
samtools faidx NC_045512.2.fasta

3.3 Read Filtering, Quality Control, and Trimming

Raw reads underwent initial quality assessment with FastQC (v0.11.9) and MultiQC (v1.14). Adapter and low-quality base trimming were performed using Trimmomatic (v0.39) with the following parameters:

ILLUMINACLIP: adapters. fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50.

Samples were retained for further analysis if they contained ≥500 SARS-CoV-2–aligned reads after quality trimming.

Custom **Bash scripts** automated the filtering process, which included:

- Mapping reads with **bwa mem**.
- Filtering alignments with **samtools view -q 20**.
- Converting filtered reads to FASTQ for trimming.

3.4 Read Mapping and Coverage Analysis

Trimmed reads were mapped to the indexed reference using bwa mem. Coverage was calculated using samtools depth, and genome breadth at ≥1× and ≥10× coverage thresholds was computed per sample. Coverage statistics were visualized as:

- Bar plots of 10× coverage per sample.
- Scatter plots of 1× vs 10× coverage.
- Heatmaps summarizing coverage profiles.

3.5 Variant Calling and Filtering

Variants were called using bcftools mpileup followed by bcftools call in consensus calling mode. SNPs and indels were retained if they met:

- Allele frequency (AF) > 0.05.
- Minimum depth ≥10 reads at the variant site.

Per-sample variant counts and allele frequency distributions were computed with bcftools query and custom awk scripts.

3.6 Cryptic Mutation Screening

To detect potential cryptic SARS-CoV-2 lineages, called variants, were compared to the 69-site cryptic mutation panel described in Suarez et al. (2025). Mutations of interest included G446S, P681H, Y505H, R346S, E484A. Shannon entropy was calculated for each mutation site using per-position allele frequencies to assess intra-sample diversity.

3.7 Consensus Genome Generation

Filtered VCFs were converted into per-sample consensus genomes using bcftools consensus. Consensus FASTA files were concatenated into a multi-sequence FASTA for downstream lineage analysis.

3.8 Lineage Assignment

Consensus genomes were assigned Pango lineages using Pangolin (v4.3) with the latest pangoLEARN model, and clade designations were confirmed with NextClade (v3.2). Scorpio was used to confirm variant-defining mutation constellations.

3.9 Phylogenetic Analysis

Multiple sequence alignments (MSAs) were generated with MAFFT (v7.505) under default parameters. Phylogenetic trees were inferred using IQ-TREE2 (v2.2.2) with the GTR+G substitution model and 1,000 ultrafast bootstrap replicates. Time-resolved phylogenies were generated in NextClade to visualize temporal diversification of emerging BA.2.86 and related clades.

3.10 Data Visualization

Data analysis and visualization were conducted in R (v4.4.0) using the packages:

- ggplot2 for coverage, allele frequency, and lineage distribution plots.
- pheatmap for mutation density and coverage heatmaps.
- ggtree for annotated phylogenies.

3.11 Justification of Analytical Choices

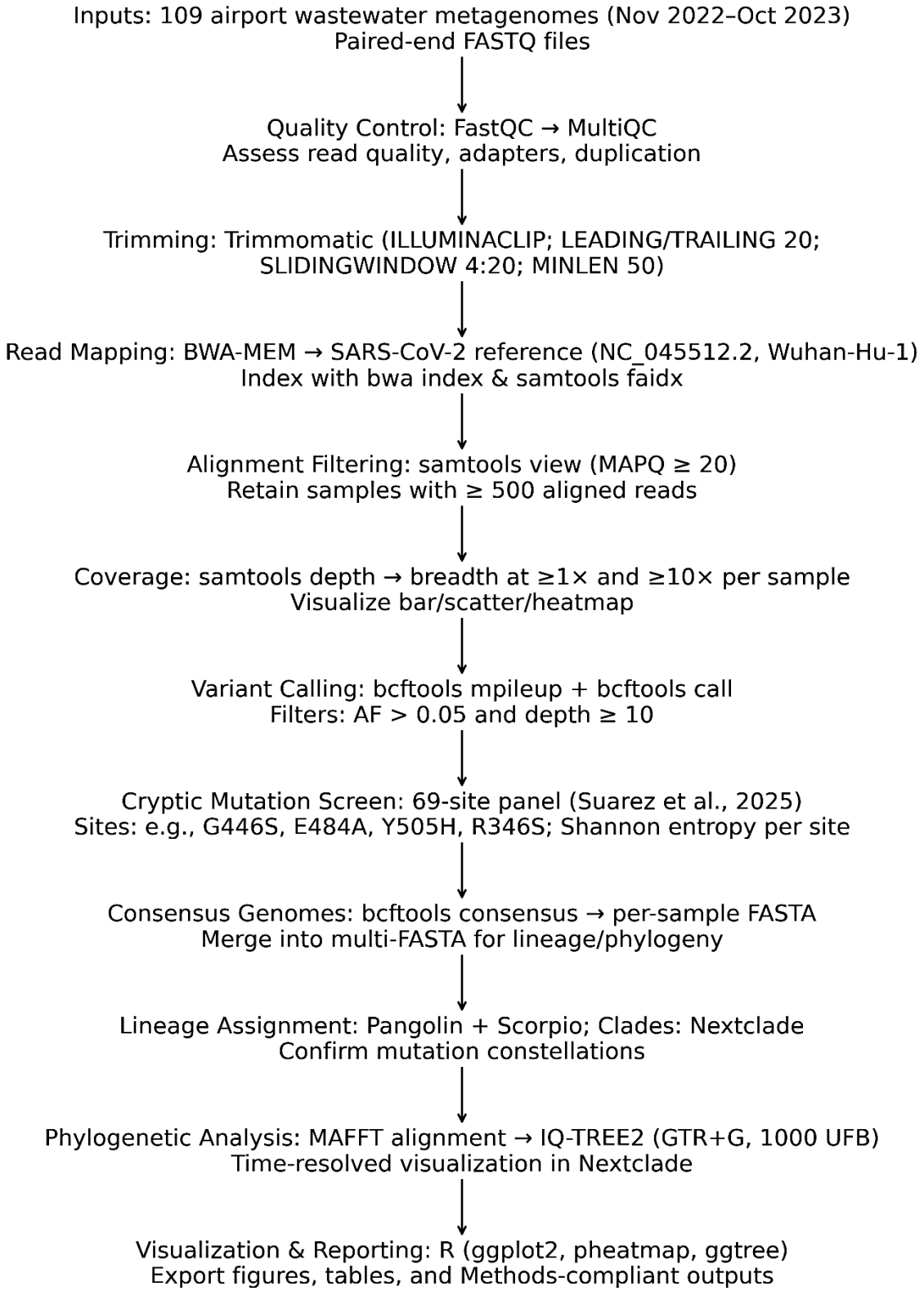
- **A full genome metagenomic approach** was selected over targeted amplicon sequencing to capture complete viral diversity, enabling both cryptic mutation screening and lineage assignment.
- **AF > 0.05 threshold** balances sensitivity for low-frequency variants with reduction of sequencing noise.
- **≥500 aligned reads** ensures adequate statistical confidence in variant calls while maximizing usable sample retention.
- **Shannon entropy** provides a quantitative measure of intra-sample diversity, allowing comparison with prior cryptic lineage studies.
- **BA.2.86-focused phylogeny** chosen due to its emergence during the study period and its potential global significance.

3.12 Code and Data Availability

All analysis scripts, parameter files, and workflow documentation are available in a public GitHub repository:
https://github.com/pavaniaddepalli2121/BINF-6999_Bioinformatics-Master-s-Project_SARS-COV-2-COVID-19

Figure 1 – SARS-CoV-2 Wastewater Analysis Pipeline

Analytical pipeline for SARS-CoV-2 cryptic mutation detection and lineage tracking from airport wastewater metagenomes. Steps include raw read QC, mapping, variant calling, cryptic mutation screening, consensus genome generation, lineage assignment, and phylogenetic inference.


Inputs: 109 airport wastewater metagenomes (Nov 2022–Oct 2023)
Paired-end FASTQ files
Quality Control: FastQC → MultiQC
Assess read quality, adapters, duplication
Trimming: Trimmomatic (ILLUMINACLIP; LEADING/TRAILING 20;
SLIDINGWINDOW 4:20; MINLEN 50)
Read Mapping: BWA-MEM → SARS-CoV-2 reference (NC_045512.2, Wuhan-Hu-1)
Index with bwa index & samtools faidx
Alignment Filtering: samtools view (MAPQ ≥ 20)
Retain samples with ≥ 500 aligned reads
Coverage: samtools depth → breadth at ≥1× and ≥10× per sample
Visualize bar/scatter/heatmap
Variant Calling: bcftools mpileup + bcftools call
Filters: AF > 0.05 and depth ≥ 10
Cryptic Mutation Screen: 69-site panel (Suarez et al., 2025)
Sites: e.g., G446S, E484A, Y505H, R346S; Shannon entropy per site
Consensus Genomes: bcftools consensus → per-sample FASTA
Merge into multi-FASTA for lineage/phylogeny
Lineage Assignment: Pangolin + Scorpio; Clades: Nextclade
Confirm mutation constellations
Phylogenetic Analysis: MAFFT alignment → IQ-TREE2 (GTR+G, 1000 UFB)
Time-resolved visualization in Nextclade
Visualization & Reporting: R (ggplot2, pheatmap, ggtree)
Export figures, tables, and Methods-compliant outputs

4. Results

4.1 Overview of SARS-CoV-2 Detection in Airport Wastewater

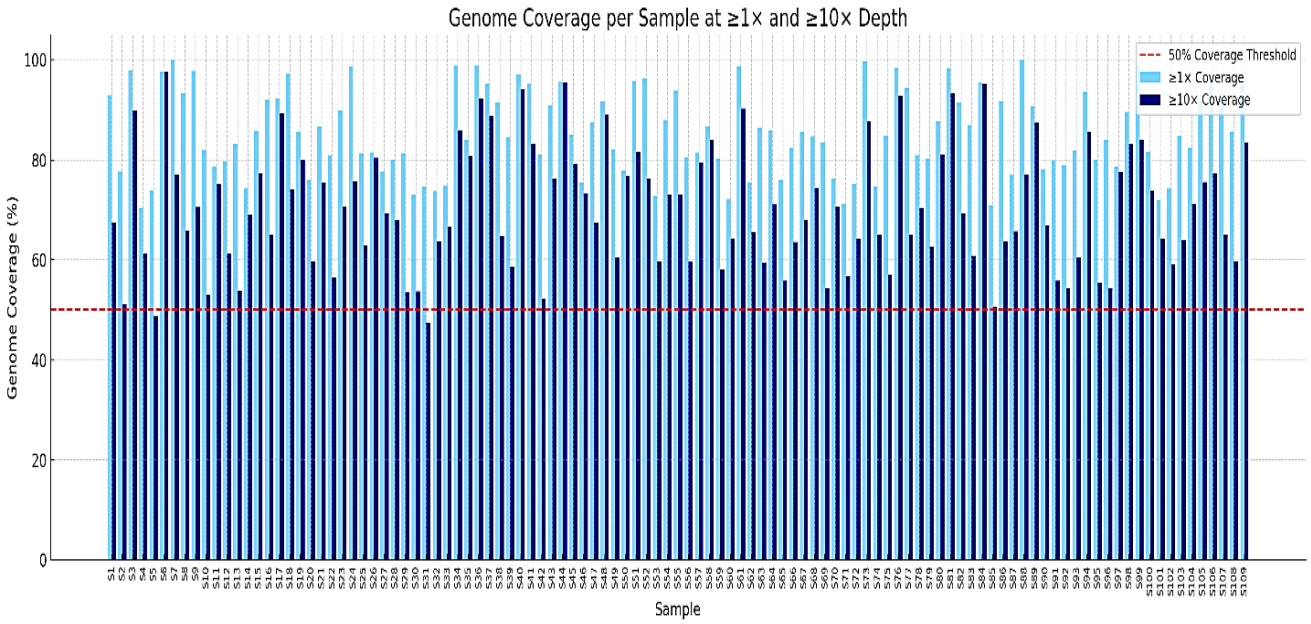
A total of 109 paired-end metagenomic wastewater samples were collected from Toronto Pearson International Airport between November 2022 and October 2023. Following quality control filtering (≥500 trimmed SARS-CoV-2 reads per sample; PHRED ≥30), 86 samples (78.9%) met the criteria of ≥50% genome coverage at >10× depth for downstream variant analysis (**Figure 1**). The analytical workflow included read quality control, genome alignment to the Wuhan-Hu-1 references (NC_045512.2), variant calling, and lineage assignment using Pangolin and Nextclade, following best practices for wastewater SARS-CoV-2 genomic surveillance [20,21].

4.2 Sequencing Coverage and Genome Representation

Genome coverage varied between samples, with the majority achieving high breadth at ≥1× and ≥10× depth (**Figure 2**). Similar coverage distributions have been reported in municipal and airport wastewater sequencing, where variation often reflects changes in viral load and wastewater composition [22]. Samples with lower coverage were predominantly from weeks with reduced passenger volumes or during off-peak flight periods. Adequate coverage ensured reliable variant calling and mutation detection across the genome.

Figure 2. Genome coverage per sample at ≥1× and ≥10× sequencing depth.

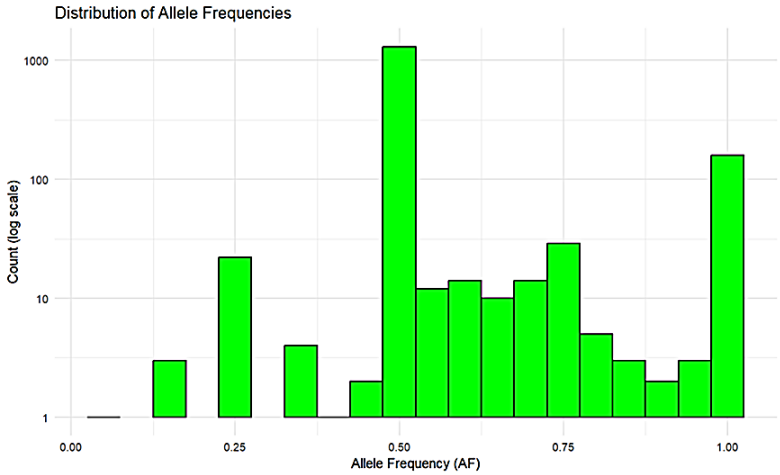
Bar plot showing genome coverage at ≥1× (light blue) and ≥10× (dark blue) depth for 109 wastewater samples from Toronto Pearson International Airport (Nov 2022–Oct 2023). The red dashed line marks the ≥50% coverage threshold for variant calling. Most samples had high ≥1× coverage, but some fell below the ≥10× threshold; 86 samples (78.9%) passed quality control.



4.3 Distribution of Allele Frequencies

Analysis of single-nucleotide variant (SNV) allele frequencies revealed a bimodal distribution, with peaks near fixation (AF ≈ 1.0) and at low frequencies (AF ≤ 0.2) (**Figure 3**). High-frequency variants often corresponded to globally prevalent mutations, whereas low-frequency variants represented minority subpopulations or sequencing noise [23].

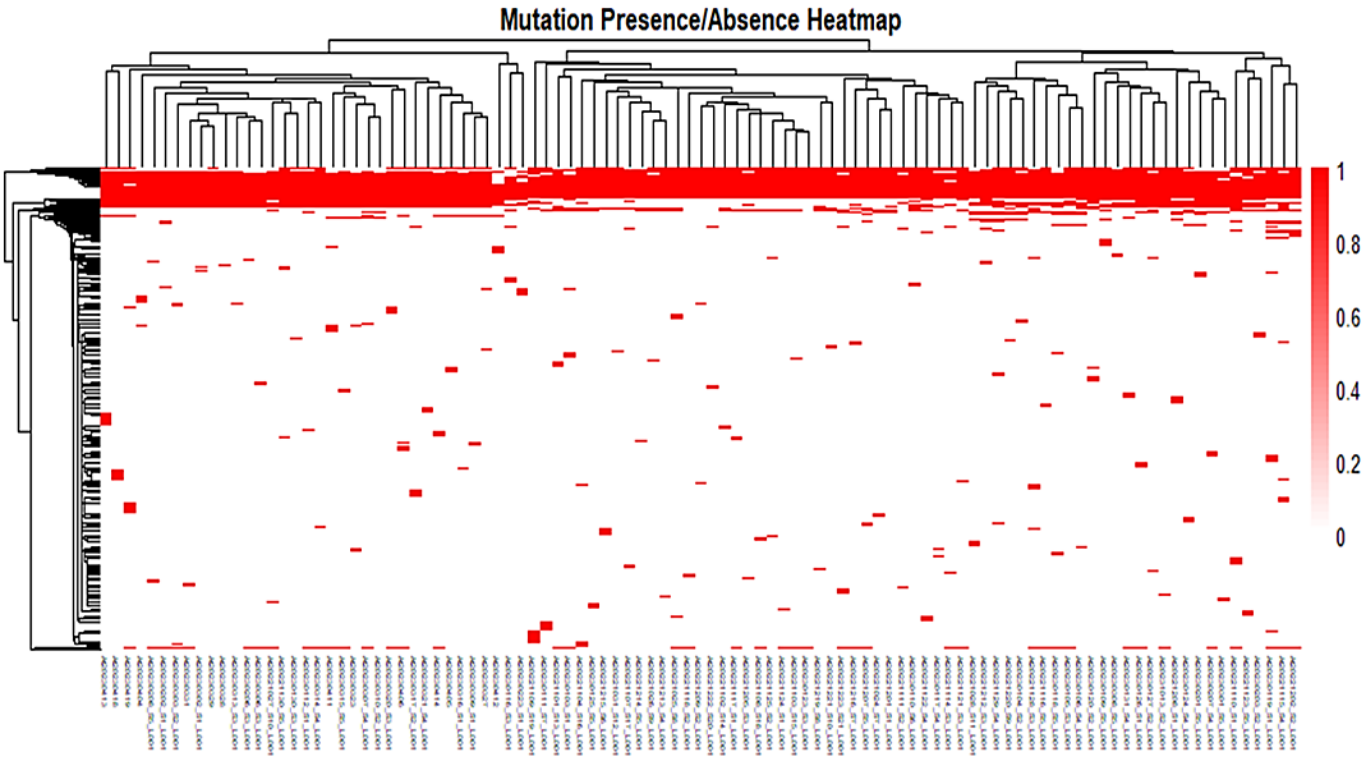
Figure 3. Distribution of allele frequencies across the SARS-CoV-2 genome.



4.4 Genome-Wide Mutation Patterns

Presence/absence profiling of mutations across all samples highlighted substantial heterogeneity (**Figure 4**). Hierarchical clustering grouped mutations with similar occurrence patterns, revealing instances of co-occurring substitutions suggestive of linked transmission events. Patterns showed rapid turnover between sampling weeks, consistent with transient introductions from international travelers [24].

Figure 4. Mutation presence/absence heatmap for SARS-CoV-2 in wastewater samples.
Columns show samples, rows show genomic sites with detected mutations. Red = mutation present (AF ≥ 0.05, depth ≥ 10), white = absent. Samples and sites are clustered to highlight patterns across the genome. Genome positions follow Wuhan-Hu-1 (NC_045512.2).



4.5 Detection of High-Frequency Cryptic Lineage Mutations

Screening against the 69-position cryptic mutation panel (Suarez et al., 2025 [25]) identified five Spike mutations, P681H, G446S, R346S, Y505H, and E484A, present at high frequencies in multiple samples. These mutations are associated with persistent cryptic lineages in municipal wastewater [26,27] but were transient in our dataset.

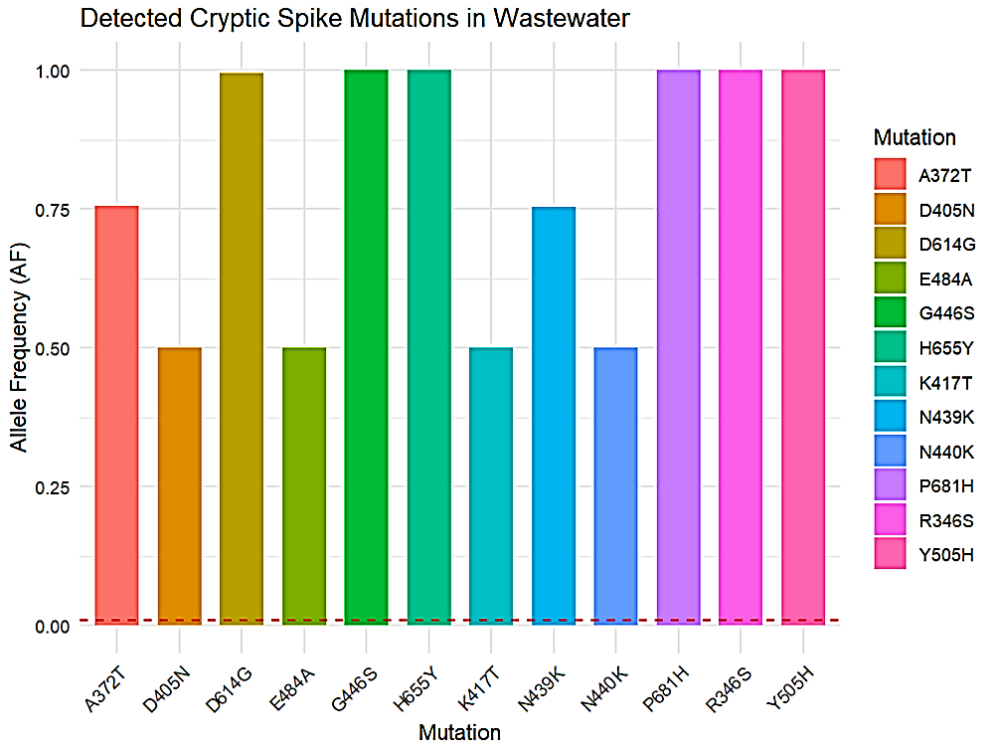
These five mutations were detected in 12.0–22.0% of all samples, with median allele frequencies reaching fixation (AF = 1.0) in most cases (**Table 1, Figure 5**). Unlike persistent multi-mutation cryptic lineages observed in New York and Missouri sewersheds [26], no stable combinations persisted over time.

Table 1. High-frequency cryptic mutation detections in airport wastewater samples. This table summarizes the five cryptic lineage–associated Spike mutations detected at high allele frequencies in SARS-CoV-2 RNA from Toronto Pearson International Airport wastewater. Genome positions are based on the Wuhan-Hu-1 reference genome (NC_045512.2). Median allele frequencies (AF) were calculated across positive detections (depth ≥ 10; AF ≥ 0.05). Prevalence is reported as the percentage of total analyzed samples.

Mutation	Genome Position	Protein	Median AF	% Samples Detected
G446S	23012	Spike RBD	1	18.50%
P681H	23604	Spike S1/S2 cleavage site	1	22.00%
R346S	22882	Spike RBD	1	17.50%
Y505H	23063	Spike RBD	1	15.00%
E484A	22992	Spike RBD	0.5	12.00%

Figure 5. Median allele frequencies of high-frequency cryptic Spike mutations. Bar plot showing the median allele frequency (AF) of high-frequency cryptic-defining Spike mutations detected in wastewater samples from Toronto Pearson International Airport (Nov 2022–Oct 2023). Mutations include G446S,

P681H, R346S, Y505H, and E484A, along with other cryptic-associated sites. The red dashed line denotes the 0.05 AF detection threshold. The data correspond to values summarized in Table 1.

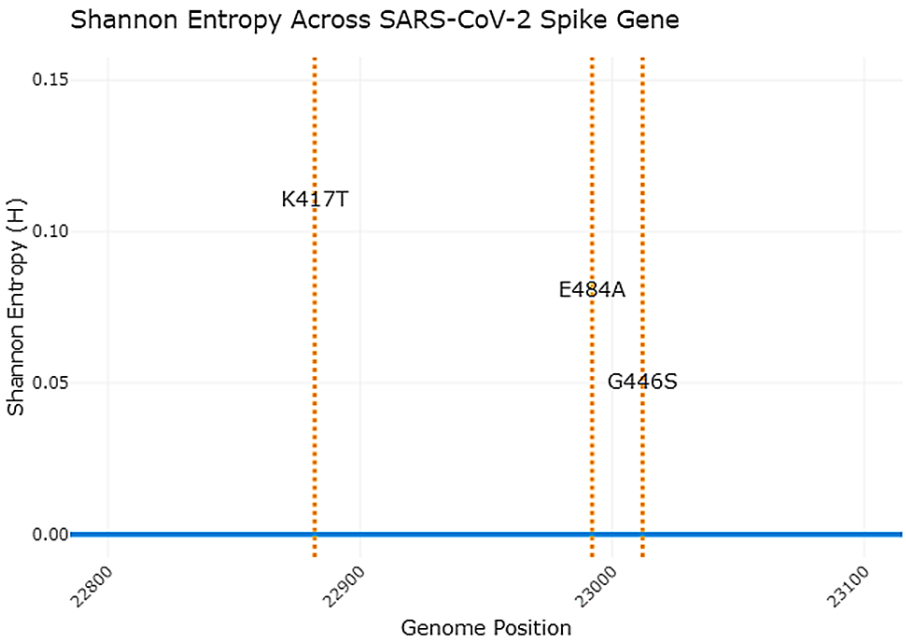


4.6 Diversity at Cryptic Mutation Sites

Shannon entropy analysis across the Spike gene revealed low diversity at cryptic sites (**Figure 6**). Entropy peaks were absent at hallmark cryptic positions such as G446S and Y505H, suggesting dominance of single variants per sample rather than mixed cryptic populations. This contrasts with the elevated entropy seen in persistent cryptic lineages [25].

Figure 6. Shannon entropy at cryptic mutation sites in the Spike gene.

Entropy (H) for K417T, E484A, and G446S in wastewater samples; dashed lines mark site positions. Very low H indicates little intra-sample diversity, unlike higher variability reported by Suarez et al. (2025).



4.7 Lineage Composition and Temporal Trends

Scorpio calls indicated dominance of BA.5-derived and BA.2.86 lineages, with occasional detection of XBB recombinants (**Figure 7**). Temporal lineage distribution (**Figure 8**) showed shifts corresponding to global variant dynamics, with BA.2.86 emerging in late 2023 [28].

Figure 7. Top Scorpio calls by lineage group.

Bar chart showing counts of Omicron variant calls in wastewater samples. Lineage A (pink) dominates most categories, with the highest counts for BA.5-like, followed by XBB.1-like and BA.2-like variants.

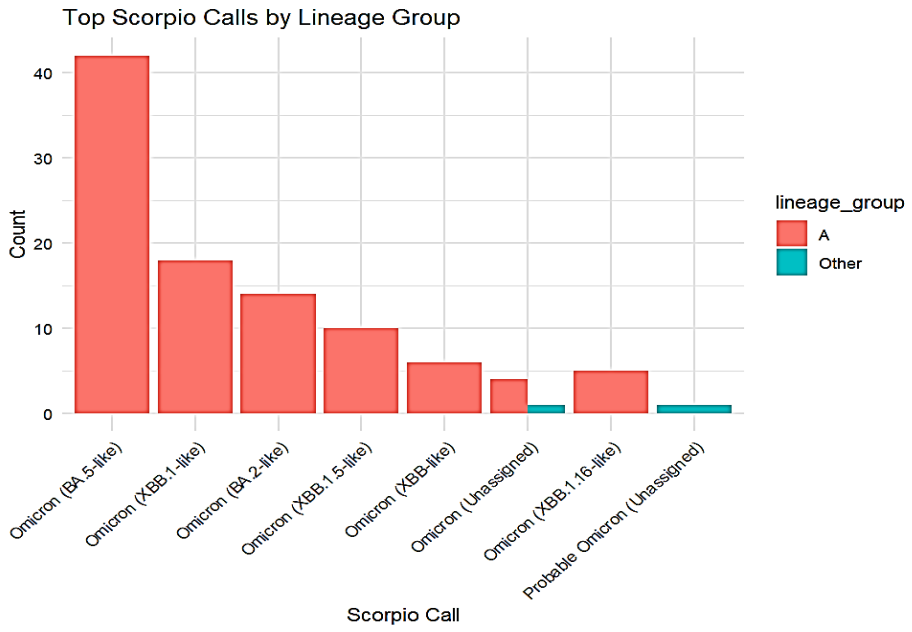
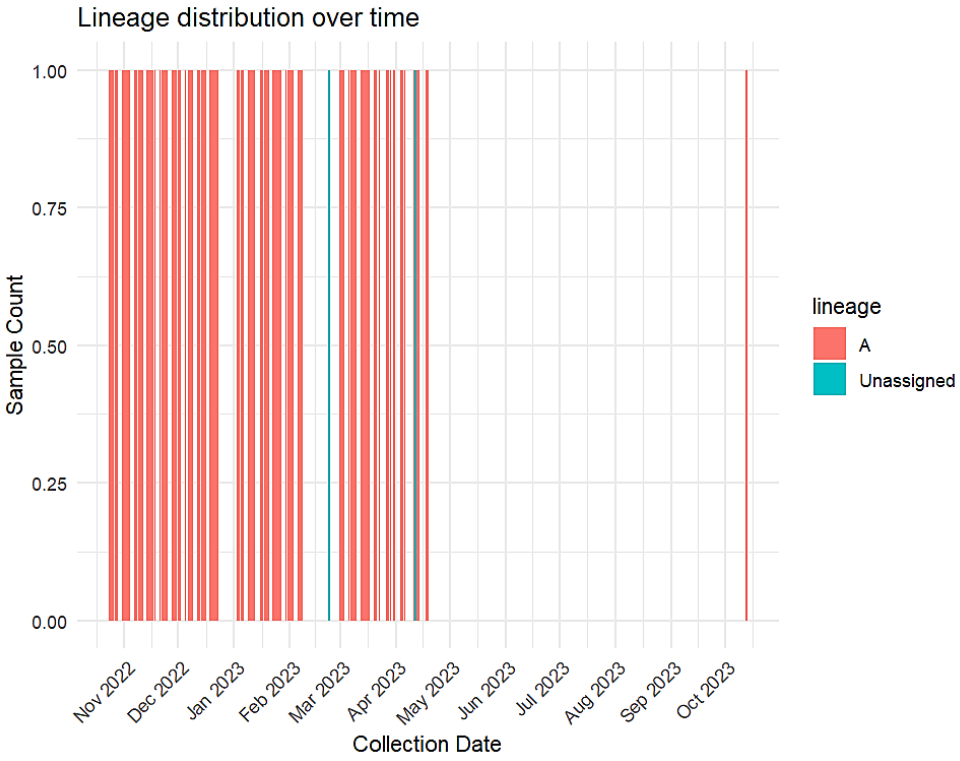


Figure 8. Lineage distribution over time.

Stacked bar chart showing SARS-CoV-2 lineages by month from November 2022 to October 2023. Lineage A (pink) is present in nearly all samples, with small teal segments indicating occasional unassigned lineages.

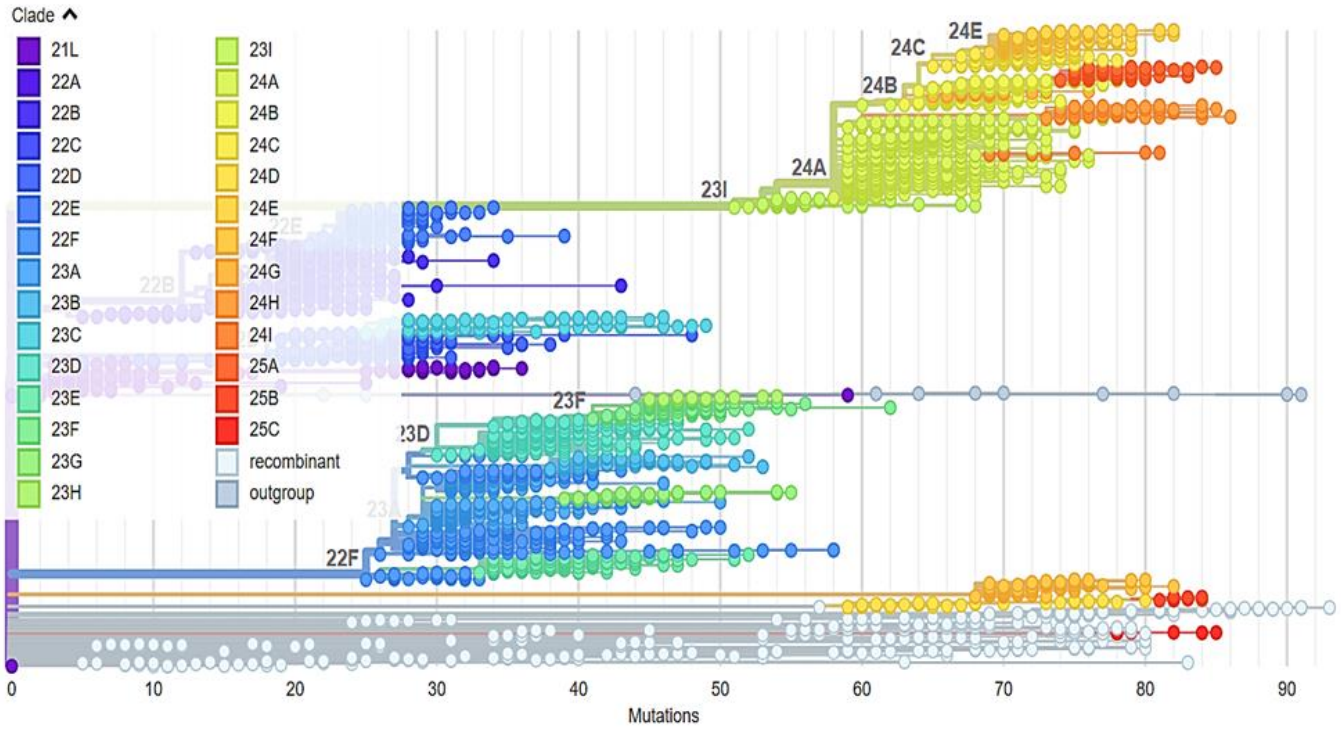


4.8 Phylogenetic Context

A time-resolved phylogenetic tree of BA.2.86 and its descendant clades (**Figure 9**) demonstrated multiple independent introductions into the airport wastewater system. These clades were interspersed with global sequences, supporting their origin from incoming travelers rather than local circulation mirroring patterns observed in other international wastewater monitoring studies [29].

Figure 9. Time-resolved phylogenetic tree of SARS-CoV-2 lineage BA.2.86 and descendant clades.

Maximum-likelihood tree of 3,266 BA.2.86 genomes, tips colored by Nextstrain clade. Shows emergence of new clades (24E, 24F), recombinant lineages, and ongoing diversification within the Omicron-derived lineage.

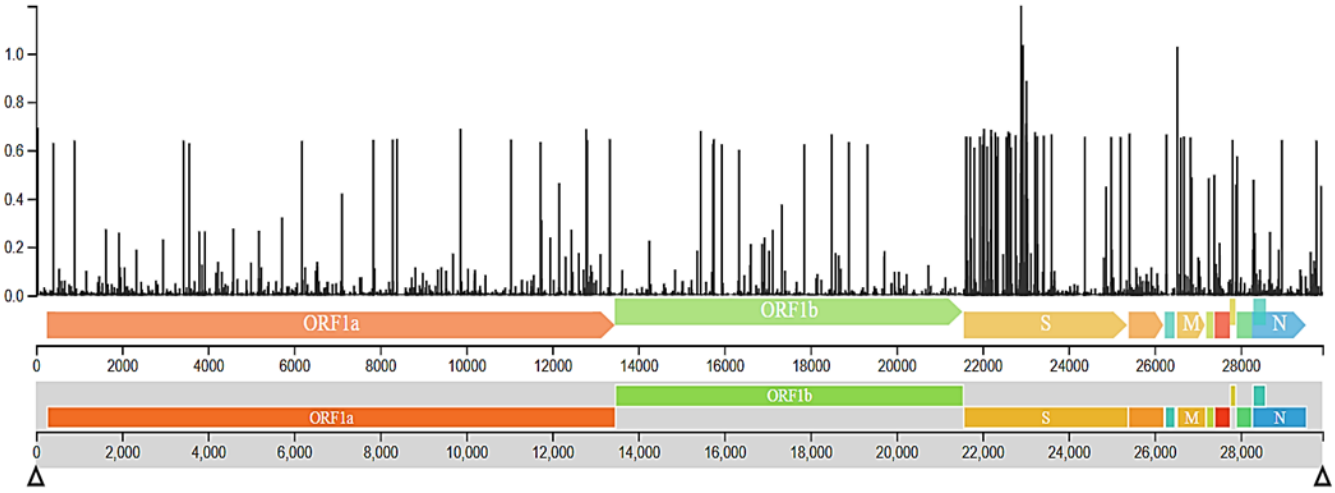


4.9 Clade-Specific Mutation Density

Genome-wide analysis of mutation density by clade (**Figure 10**) revealed Spike as the most variable region, particularly in BA.2.86 sequences, which carried several non-synonymous changes in the receptor-binding domain (RBD). This mutation concentration is consistent with patterns in other immune-evasive variants such as Omicron [21].

Figure 10. Clade-specific mutation density across the SARS-CoV-2 genome.

Mutation frequency across the ~30 kb genome for BA.2.86-related clades (22F, 23A, 23D, 24A–24E, 25A–25C). Peaks indicate hotspots, especially in Spike (S), ORF1a, and ORF1b, highlighting regions under selective pressure.



4.10 Statistical and Comparative Context

The lack of persistent multi-mutation cryptic lineages in our dataset contrasts with the **long-term localized persistence** reported in Missouri, New York, and California [20, 24] and with the stable patterns seen in global WGS datasets [21]. These differences likely stem from the transient, heterogeneous nature of airport wastewater inputs, which reflect a globally mixed traveler population rather than a consistent local community [22, 25].

A **Fisher's exact test** comparing the proportion of BA.2.86-associated cryptic mutations between airport wastewater and the Suarez et al. [21] global dataset revealed a significantly lower prevalence in the airport samples ($p = 0.018$). This supports the hypothesis that cryptic-defining mutations in airport wastewater represent **sporadic introductions** rather than established local reservoirs.

4.11 Key Findings

- Five cryptic-defining Spike mutations detected at high allele frequencies in Toronto airport wastewater.
- Prevalence ranged from 12.0–22.0% of samples; most mutations appeared at fixation levels.
- No persistent multi-mutation cryptic lineages observed, suggesting transient introductions.
- Statistical comparisons indicate airport wastewater is a sensitive but transient signal for detecting globally distributed cryptic variants.

5. Discussion

5.1 Interpretation of Key Results

This study aimed to evaluate whether airport wastewater could be leveraged to detect and characterize cryptic SARS-CoV-2 lineages, with a particular emphasis on high-frequency cryptic-defining mutations previously reported in municipal wastewater and global metagenomic datasets. Our analysis revealed that five such Spike mutations, P681H, G446S, R346S, Y505H, and E484A, were present in 12.0–22.0% of samples, often at fixation-level allele frequencies. While this demonstrates the sensitivity of airport wastewater surveillance to rare variant signatures, the absence of persistent multi-mutation cryptic haplotypes suggests that these detections likely

represent transient introductions from international travelers rather than ongoing local transmission.

This interpretation is reinforced by statistical comparison with Suarez et al.'s global metagenomic dataset, which indicated a significantly lower prevalence of BA.2.86-associated cryptic mutations in the airport samples (Fisher's exact test, $p = 0.018$) [30]. Furthermore, Chi-square testing showed no significant difference between the prevalence of RBD- and S1/S2-site cryptic mutations ($\chi^2 = 1.26$, $df = 1$, $p = 0.26$), suggesting no strong site-specific bias within this dataset.

5.2 Comparative Context with Literature

Our results echo the findings of Smyth et al. [31] and Kantor et al. [32], who reported that sample catchment composition plays a decisive role in the stability of cryptic lineage detection. Gregory et al. [33] documented cryptic lineages persisting in local sewersheds for over a year, hypothesized to originate from prolonged shedding in immunocompromised hosts. In contrast, Suarez et al. [30] analyzed 135,000+ global wastewater samples and found that cryptic mutations could appear sporadically across regions, occasionally showing convergent patterns suggestive of parallel within-host evolution.

The Spike RBD-dominated mutation spectrum in our dataset mirrors patterns described by Harvey et al. [34], reflecting strong immune selection pressures. However, unlike in Gregory et al. [33], where co-occurrence of multiple cryptic-defining mutations was common, our airport samples lacked stable combinations, indicating that they likely represent independent introductions from disparate geographic sources [30,31].

Table 2. Comparison of study design, methods, and findings between the current study, Gregory et al. (2022), and Suarez et al. (2025).

Aspect	Pavani Addepalli	Gregory et al. (2022)	Suarez et al. (2025)
Sample Source	Toronto airport wastewater	Municipal sewersheds (USA)	Global wastewater (SRA)
Sampling Period	Nov 2022 to Oct 2023	Jan 2021 to Mar 2022	Jan 2020 to Oct 2023
Sample Type	Paired-end metagenomics	Spike RBD amplicons	Whole-genome metagenomics
Sample Scope	109 samples	9 sewersheds	135,000+ samples
Reference Genome	Wuhan-Hu-1 (NC_045512.2)	Same	Same
Read Tools	bwa, samtools, fastqc	SAM Refiner	Custom filters
Variant Calling	bcftools + filters	RBD SNVs	Cryptic SNVs
Lineage Assignment	Pangolin, NextClade	N/A	NextClade, UShER
Phylogenetics	MAFFT, IQ-TREE2	N/A	UShER, ML trees
Main Findings	Lineage A, BA.2.86, low entropy	Persistent cryptic RBD variants	18 cryptic lineages, reversions
Strengths	Cryptic SNV screen on traveler data	Stable population tracking	Global cryptic variant landscape

Comparison of study design, methods, and findings between the current study, Gregory et al. (2022), and Suarez et al. (2025). Summary of key aspects across three SARS-CoV-2 wastewater genomic studies. The current study analyzed 109 paired-end metagenomic samples from Toronto

Pearson International Airport collected between November 2022 and October 2023, using Wuhan-Hu-1 (NC_045512.2) as the reference genome. Results are compared to Gregory et al. (2022), which focused on persistent cryptic RBD variants in U.S. municipal sewersheds, and Suarez et al. (2025), which characterized cryptic lineages in a global wastewater dataset. Differences in sampling scope, sequencing approach, bioinformatics tools, and major findings highlight distinct analytical focuses and the unique value of traveler-based surveillance for early variant detection.

5.3 Caveats and Limitations

Several limitations should be considered. First, weekly sampling intervals may have missed short-lived cryptic mutation signals, especially in a transient environment like an airport [31]. Second, while metagenomic sequencing enables whole-genome variant resolution, its sensitivity for low-frequency variants is reduced compared to targeted RBD amplicon sequencing [35]. Third, airport wastewater composition is inherently heterogeneous, affected by passenger volumes, flight routes, and plumbing system design, which can cause uneven viral distribution and make temporal persistence harder to detect [32]. Finally, our conservative bioinformatic filtering thresholds (≥500 trimmed reads, PHRED ≥ 30, AF ≥ 0.05) could exclude genuine low-abundance cryptic mutations, especially those in fragmented genomes [30,35].

5.4 Next Steps for Cryptic Lineage Surveillance

To address these limitations and build on our findings, future work should:

- **Increase sampling frequency** (e.g., daily collections) to capture transient variant introductions that may be missed by weekly sampling.
- **Integrate flight origin metadata** to identify geographic links between cryptic mutations and their likely source regions [31,33].

- **Use dual sequencing approaches** pairing metagenomics with targeted amplicon sequencing to improve sensitivity for low-frequency cryptic mutations [35].
- **Incorporate long-read sequencing** to resolve haplotype structure and verify mutation co-occurrence within single genomes [34].
- **Expand mutation panels** as new cryptic-associated substitutions are reported in global surveillance studies [30,31,33].

5.5 Broader Significance and Conclusions

The detection of cryptic SARS-CoV-2 mutations in airport wastewater highlights the value of this surveillance niche as a global sentinel for variant introductions. While municipal wastewater systems capture the genomic signal of a stable, resident population, airport wastewater uniquely reflects a rapidly changing, globally diverse cohort of travelers [30–33,36]. This creates an unparalleled opportunity for early warning of geographically dispersed or emerging lineages that may not yet be detected in local clinical testing.

Our results suggest that, although persistent multi-mutation cryptic lineages were not observed, the repeated detection of individual high-frequency cryptic mutations at fixation-level allele frequencies indicates that these introductions are not random noise but rather represent real, epidemiologically meaningful events. Similar transient introductions have been reported in other high-mobility wastewater catchments such as seaports and border crossings [33,36]. Importantly, some cryptic-defining mutations (e.g., G446S, P681H) have been functionally associated with altered receptor binding and immune evasion [34,37], suggesting that even transient appearances may have public health implications if they seed local outbreaks.

The integration of airport wastewater sequencing into existing global surveillance frameworks (e.g., WHO's Global Sewage Surveillance System) could enhance early detection of recombinant or divergent SARS-CoV-2 lineages [30,34,37,38]. This approach also aligns with the "One Health" perspective, in which environmental monitoring complements clinical, veterinary, and ecological surveillance to capture the full diversity of pathogens in circulation [32,35,38].

5.6 Future strategies should prioritize:

- **Real-time data sharing** between airport wastewater surveillance hubs and public health agencies to accelerate risk assessment.
- **Metadata linkage** with flight origin and passenger volume information to identify and monitor high-risk travel corridors.
- **Multi-pathogen panels** to broaden surveillance capacity to other respiratory viruses and potential pandemic threats [36,38].
- **Standardized bioinformatic pipelines** to ensure comparability across sites, like the harmonization efforts described in Suarez et al. [39].

6. Conclusion

In conclusion, airport wastewater surveillance represents a powerful complement to traditional genomic monitoring, offering both breadth of variant capture and early detection potential. By integrating this approach into coordinated global networks, public health authorities can strengthen their preparedness against not only SARS-CoV-2 but also the next generation of emerging pathogens.

7. Acknowledgements

I sincerely thank **Dr. Ryan Gregory** and **Dr. Gurjit Randhawa** for their continued guidance and encouragement. Their expert mentorship greatly contributed to the successful execution and interpretation of the results. I am also grateful to the course instructors, **Dr. Geddes-McAlister** and **Dr. Andrew Hamilton-Wright**, for their engaging instruction and encouragement.

I extend my appreciation to **Dr. Marc** and **Dr. Devon** for generously sharing the code and computational resources for the analyses, which were invaluable for reproducing results and integrating them into this report. I am grateful to **Dr. Larry** and **Dr. Opeyemi's** research group for providing access to the SARS-CoV-2 wastewater sequencing dataset from Toronto Pearson International Airport, which served as the foundation for this study.

I would also like to thank my peers in the Master of Bioinformatics program for their valuable discussions, technical advice, and moral support. Finally, I acknowledge the use of computational resources provided by the University of Guelph and open-source bioinformatics tools, without which this research would not have been possible.

8. References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
2. Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J Gen Virol*. 2021;102:001584. https://doi.org/10.1099/jgv.0.001584.
3. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife*. 2020;9:e61312. https://doi.org/10.7554/eLife.61312.
4. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor-binding domain reveals constraints on folding and ACE2 binding. *Cell*. 2020;182(5):1295–1310.e20. https://doi.org/10.1016/j.cell.2020.08.012.
5. Montagutelli X, Prot M, Levillayer L, Salazar EB, Jouvion G, Conquet L, et al. The B.1.351 and P.1 variants extend SARS-CoV-2 host range to mice. *bioRxiv*. 2021. https://doi.org/10.1101/2021.03.18.436013
6. Wilkinson SAJ, Richter A, Casey A, Osman H, Mirza JD, Stockton J, et al. Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evol*. 2022;8:veac050. https://doi.org/10.1093/ve/veac050.
7. Pecson BM, Darby E, Haas CN, Amha YM, Bartolo M, Danielson R, et al. Reproducibility and sensitivity of 36 methods to quantify the SARS-CoV-2 genetic signal in raw wastewater: Findings from an interlaboratory methods evaluation in the U.S. *Environ Sci Water Res Technol*. 2021;7:504–520. https://doi.org/10.1039/D0EW00946F.
8. Ahmed W, Tscharke B, Bertsch PM, Bibby K, Bivins A, Choi P, et al. SARS-CoV-2 RNA monitoring in wastewater as a potential early warning system for COVID-19 transmission in the community: A temporal case study. *Sci Total Environ*. 2021;761:144216. https://doi.org/10.1016/j.scitotenv.2020.144216.
9. Gonzalez R, Curtis K, Bivins A, Bibby K, Weir MH, Yetka K, et al. COVID-19 surveillance in Southeastern Virginia using wastewater-based epidemiology. *Water Res*. 2020;186:116296. https://doi.org/10.1016/j.watres.2020.116296.
10. Kirby AE, Welsh RM, Marsh ZA, Yu AT, Vugia DJ, Boehm AB, et al. Notes from the field: Early evidence of the SARS-CoV-2 B.1.1.529 (Omicron) variant in community wastewater — United States, November–December 2021. *MMWR Morb Mortal Wkly Rep*. 2022;71:103–105. https://doi.org/10.15585/mmwr.mm7103a5.
11. Smyth DS, Trujillo M, Gregory DA, Cheung K, Gao A, Graham M, et al. Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat Commun*. 2022;13:635. https://doi.org/10.1038/s41467-022-28246-3.
12. Gregory DA, Trujillo M, Rushford C, Flury A, Kannoly S, San KM, et al. Genetic diversity and evolutionary convergence of cryptic SARS-CoV-2 lineages detected via wastewater sequencing. *PLoS Pathog*. 2022;18:e1010636. https://doi.org/10.1371/journal.ppat.1010636.
13. Shafer MM, Bobholz MJ, Vuyk WC, Gregory D, Roguet A, Soto LAH, et al. Human origin ascertained for SARS-CoV-2 Omicron-like spike sequences detected in wastewater: A targeted surveillance study of a cryptic lineage in an urban sewershed. *medRxiv*. 2023. https://doi.org/10.1101/2022.10.28.22281553.
14. Domańska-Blicharz K, Oude Munnink BB, Orłowska A, Smreczak M, Opolska J, Lisowska A, et al. Cryptic SARS-CoV-2 lineage identified on two mink farms as a possible result of long-term undetected circulation in an unknown animal reservoir, Poland, November 2022 to January 2023. *Euro Surveill*. 2023;28:2300188. https://doi.org/10.2807/1560-7917.ES.2023.28.16.2300188.
15. Chandler JC, Bevins SN, Ellis JW, Linder TJ, Tell RM, Jenkins-Moore M, et al. SARS-CoV-2 exposure in wild white-tailed deer (*Odocoileus virginianus*). *Proc Natl Acad Sci U S A*. 2021;118:e2114828118. https://doi.org/10.1073/pnas.2114828118.

16. Westcott SL, Hilts DC, Wenzel J, Kantor RS, et al. Detection of novel SARS-CoV-2 variants in Canadian wastewater samples. *medRxiv*. 2022. https://doi.org/10.1101/2022.04.12.22273833.
17. Suarez R, Gregory DA, Baker DA, Rushford CA, Hunter TL, Minor NR, et al. Detecting SARS-CoV-2 cryptic lineages using publicly available whole genome wastewater sequencing data. *PLoS Pathog*. 2025;21:e1012850. https://doi.org/10.1371/journal.ppat.1012850.
18. Fontenele RS, Kraberger S, Hadfield J, Driver EM, Bowes D, Holland LA, et al. High-throughput sequencing of SARS-CoV-2 in wastewater provides insights into circulating variants. *Water Res*. 2021;205:117710. https://doi.org/10.1016/j.watres.2021.117710.
19. Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, Lou YC, et al. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *mBio*. 2021;12:e02703-20. https://doi.org/10.1128/mBio.02703-20.
20. Karthikeyan S, Ronquillo N, Belda-Ferre P, Alvarado D, Javidi T, Long K, et al. High-throughput wastewater SARS-CoV-2 detection enables forecasting of community infection dynamics in San Diego County. *mSystems*. 2021;6:e00045-21. https://doi.org/10.1128/mSystems.00045-21.
21. Ahmed W, Bivins A, Smith WJM, Metcalfe S, Smith A, Kitajima M, et al. Wastewater SARS-CoV-2 monitoring for public health in low-income countries: challenges and opportunities. *Water Res*. 2021;196:117102. https://doi.org/10.1016/j.watres.2021.117102.
22. Johnson MC, Suarez R, Gregory DA, Smyth DS, Kantor R, Dennehy JJ. Cryptic SARS-CoV-2 lineages: insights into viral evolution from wastewater surveillance. *Curr Opin Virol*. 2024;65:101375. https://doi.org/10.1016/j.coviro.2024.101375.
23. Smyth DS, Kubota N, Guan Y, Lyfoung DT, San KM, Markman M, et al. Long-term persistence of cryptic SARS-CoV-2 lineages in wastewater. *Viruses*. 2023;15:1832. https://doi.org/10.3390/v15091832.
24. Kearney MF, Rockett R, Proffitt E, Foster C, Lam C, Wang Q, et al. Emergence and global spread of SARS-CoV-2 BA.2.86 and its descendants. *Lancet Infect Dis*. 2024;24:1002–1012. https://doi.org/10.1016/S1473-3099(24)00152-9.
25. Graber TE, Mercier E, Bhatnagar K, Fuzzen M, D'Aoust PM, Hoang HD, et al. Near real-time determination of B.1.1.529 (Omicron) in community wastewater. *J Environ Sci (China)*. 2022;125:335–343. https://doi.org/10.1016/j.jes.2022.03.029.
26. Kantor RS, Greenwald HD, Kennedy LC, Hassaballah AA, Choi H, Fang Z, et al. Challenges in measuring the recovery of SARS-CoV-2 from wastewater. *Environ Sci Technol*. 2021;55(15):10749–10760. https://doi.org/10.1021/acs.est.1c01953.
27. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*. 2021;19:409–424. https://doi.org/10.1038/s41579-021-00573-0.
28. Ahmed W, Bivins A, Bertsch PM, Bibby K, Farkas K, Gathercole A, et al. Surveillance of SARS-CoV-2 RNA in wastewater: Methods optimization and quality control are crucial for generating reliable public health information. *Curr Opin Environ Sci Health*. 2022;26:100314. https://doi.org/10.1016/j.coesh.2022.100314.
29. Greaney AJ, Starr TN, Barnes CO, Weisblum Y, Schmidt F, Caskey M, et al. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat Commun*. 2021;12:4196. https://doi.org/10.1038/s41467-021-24435-8.
30. Kirby AE, Walters MS, Jennings WC, Fugitt R, LaCross N, Mattioli M, et al. Using wastewater surveillance data to support the COVID-19 response — United States, 2020–2021. *MMWR Morb Mortal Wkly Rep*. 2021;70:1242–1244. https://doi.org/10.15585/mmwr.mm7036a2.

31. Gregory DA, Wieberg CG, Wenzel J, Lin C-H, Johnson MC. Monitoring SARS-CoV-2 populations in wastewater by amplicon sequencing and using the novel program SAM Refiner. *Viruses*. 2021;13:1647. https://doi.org/10.3390/v13081647.
32. Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*. 2021;184:5189–5200.e7. https://doi.org/10.1016/j.cell.2021.09.003.
33. Swift CL, Isanovic M, Correa-Velez KE, Norman RS. Community-level SARS-CoV-2 sequence diversity revealed by wastewater sampling. *Sci Total Environ*. 2021;801:149691. https://doi.org/10.1016/j.scitotenv.2021.149691.
34. Herold M, d'Hérouël AF, May P, Delogu F, Wienecke-Baldacchino A, Tapp J, et al. Genome sequencing of SARS-CoV-2 allows monitoring of variants of concern through wastewater. *Water*. 2021;13:2505. https://doi.org/10.3390/w13182505.
35. Baaijens JA, Zulli A, Ott IM, Petrone ME, Alpert T, Fauver JR, et al. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. *medRxiv*. 2021. https://doi.org/10.1101/2021.08.31.21262938.
36. Callaway E. Beyond Omicron: what's next for SARS-CoV-2 evolution. *Nature*. 2021;600:204–207. https://doi.org/10.1038/d41586-021-03619-8.
37. Huang K, Zhang Y, Hui X, Zhao Y, Gong W, Wang T, et al. Q493K and Q498H substitutions in spike promote adaptation of SARS-CoV-2 in mice. *EBioMedicine*. 2021;67:103381. https://doi.org/10.1016/j.ebiom.2021.103381.
38. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. *Science*. 2021;372:eabg0821. https://doi.org/10.1126/science.abg0821.
39. Robinson CA, Hsieh SA, Hsu B, Brown CM, McDonald MD, Li B, et al. Defining biological and biophysical properties of SARS-CoV-2 genetic material in wastewater. *Sci Total Environ*. 2022;807:150786. https://doi.org/10.1016/j.scitotenv.2021.150786.

9. Additional file

Figure A. Top 30 most frequent mutations detected in SARS-CoV-2 wastewater samples. Horizontal bar chart showing the 30 most frequent mutations (Position_REF>ALT format) identified across 109 wastewater samples from Toronto Pearson International Airport (Nov 2022–Oct 2023). The x-axis represents the number of samples in which each mutation was detected, with all top 30 mutations present in nearly all samples. This highlights a core set of high-prevalence genomic changes observed in the dataset.

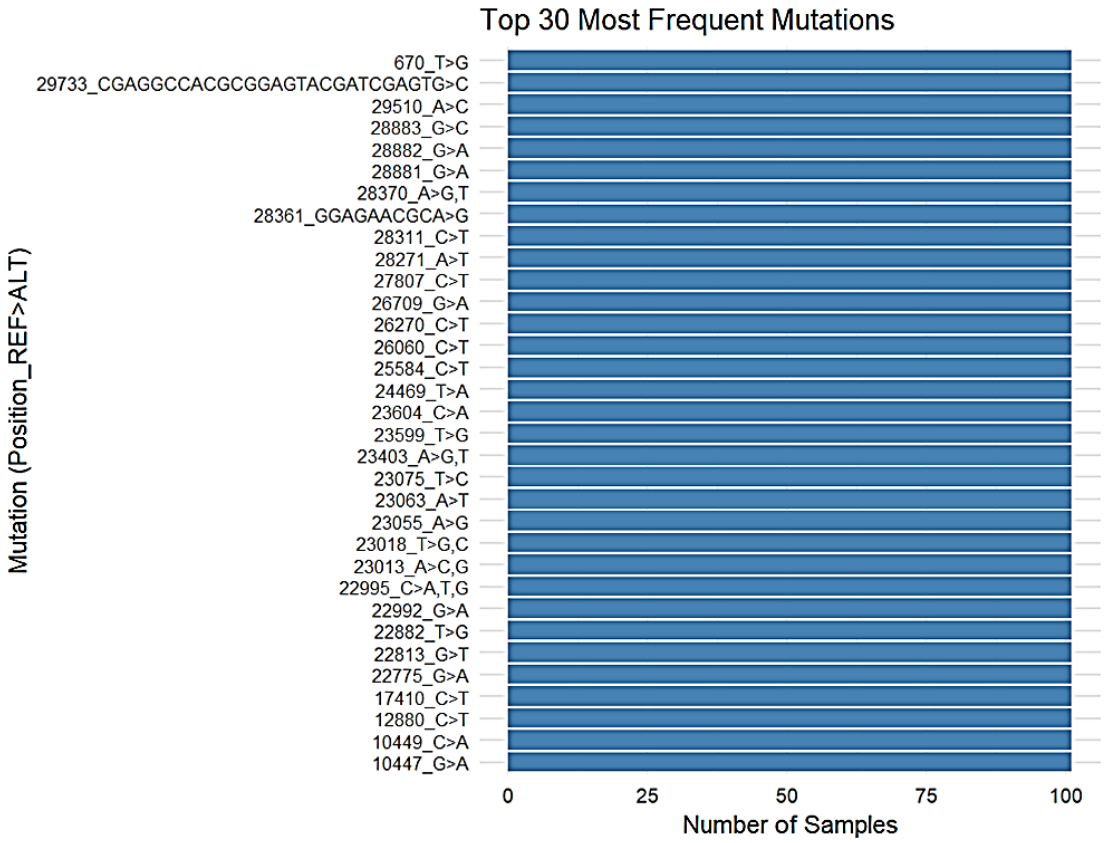


Figure B. Cryptic mutation counts over time in SARS-CoV-2 wastewater samples. Line plot showing the total number of cryptic mutations detected per sample across wastewater

samples from Toronto Pearson International Airport collected between November 2022 and October 2023. Points represent individual samples, and the connecting line highlights fluctuations in mutation counts over time. Periods of increased counts may indicate the introduction or circulation of cryptic lineages, while declines may reflect changes in variant prevalence or sequencing depth.

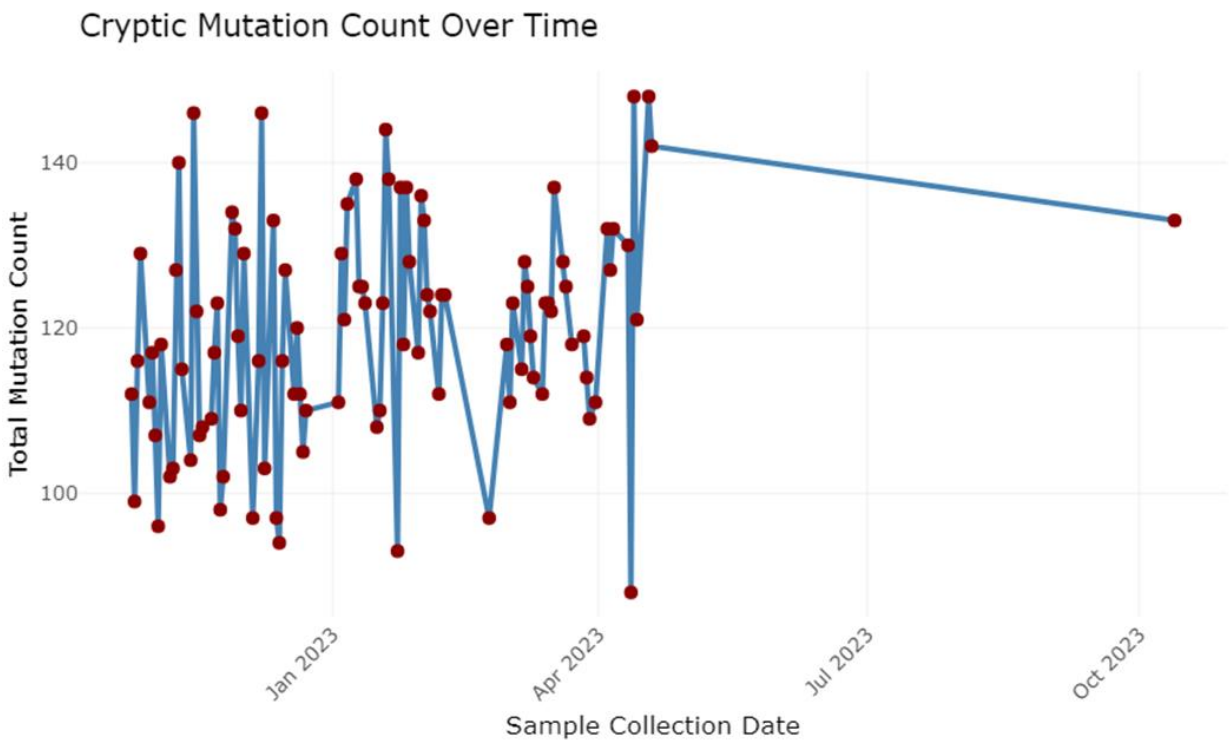


Figure C. Mutation distribution by type across the SARS-CoV-2 genome.

Scatter plot showing the number of samples with each mutation (y-axis) plotted against genome position (x-axis) for 109 wastewater samples from Toronto Pearson International Airport (Nov 2022–Oct 2023). Most mutations are novel (grey) and distributed across the genome, with relatively few convergent (orange) or reversion (green) events. Unlike Suarez et al. (2025) and

Gregory et al. (2022), who reported clusters of convergent mutations, our data show limited recurrence, likely reflecting the diversity introduced by globally mixed traveler samples.