<u>Original Paper</u>

Mutational Patterns Observed in SARS-CoV-2 Genomes Sampled From Successive Epochs Delimited by Major Public Health Events in Ontario, Canada: Genomic Surveillance Study

David Chen[1], BMedSci; Gurjit S Randhawa[2], MSc, PhD; Maximillian PM Soltysiak[1], BSc; Camila PE de Souza[3], PhD; Lila Kari[4], PhD; Shiva M Singh[1], PhD; Kathleen A Hill[1], MSc, PhD

[1]Department of Biology, Western University, London, ON, Canada

[2]School of Mathematical and Computational Sciences, University of Prince Edward Island, Charlottetown, PE, Canada

[3]Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada

[4]School of Computer Science, University of Waterloo, Waterloo, ON, Canada

Corresponding Author:
Kathleen A Hill, MSc, PhD
Department of Biology
Western University
1151 Richmond Street
London, ON, N6A 5B7
Canada
Phone: 1 519 661 2111 ext 81337
Email: khill22@uwo.ca

Abstract

Background: The emergence of SARS-CoV-2 variants with mutations associated with increased transmissibility and virulence is a public health concern in Ontario, Canada. Characterizing how the mutational patterns of the SARS-CoV-2 genome have changed over time can shed light on the driving factors, including selection for increased fitness and host immune response, that may contribute to the emergence of novel variants. Moreover, the study of SARS-CoV-2 in the microcosm of Ontario, Canada can reveal how different province-specific public health policies over time may be associated with observed mutational patterns as a model system.

Objective: This study aimed to perform a comprehensive analysis of single base substitution (SBS) types, counts, and genomic locations observed in SARS-CoV-2 genomic sequences sampled in Ontario, Canada. Comparisons of mutational patterns were conducted between sequences sampled during 4 different epochs delimited by major public health events to track the evolution of the SARS-CoV-2 mutational landscape over 2 years.

Methods: In total, 24,244 SARS-CoV-2 genomic sequences and associated metadata sampled in Ontario, Canada from January 1, 2020, to December 31, 2021, were retrieved from the Global Initiative on Sharing All Influenza Data database. Sequences were assigned to 4 epochs delimited by major public health events based on the sampling date. SBSs from each SARS-CoV-2 sequence were identified relative to the MN996528.1 reference genome. Catalogues of SBS types and counts were generated to estimate the impact of selection in each open reading frame, and identify mutation clusters. The estimation of mutational fitness over time was performed using the Augur pipeline.

Results: The biases in SBS types and proportions observed support previous reports of host antiviral defense activity involving the SARS-CoV-2 genome. There was an increase in U>C substitutions associated with adenosine deaminase acting on RNA (ADAR) activity uniquely observed during Epoch 4. The burden of novel SBSs observed in SARS-CoV-2 genomic sequences was the greatest in Epoch 2 (median 5), followed by Epoch 3 (median 4). Clusters of SBSs were observed in the spike protein open reading frame, ORF1a, and ORF3a. The high proportion of nonsynonymous SBSs and increasing dN/dS metric (ratio of nonsynonymous to synonymous mutations in a given open reading frame) to above 1 in Epoch 4 indicate positive selection of the spike protein open reading frame.

Conclusions: Quantitative analysis of the mutational patterns of the SARS-CoV-2 genome in the microcosm of Ontario, Canada within early consecutive epochs of the pandemic tracked the mutational dynamics in the context of public health events that

instigate significant shifts in selection and mutagenesis. Continued genomic surveillance of emergent variants will be useful for the design of public health policies in response to the evolving COVID-19 pandemic.

Introduction

SARS-CoV-2 is responsible for the global COVID-19 pandemic, and there have been 4,109,931 total confirmed COVID-19 cases in Canada as of August 12, 2022 [1]. As the most populated province in Canada, Ontario reported 1,394,524 confirmed COVID-19 cases and 52,998 hospitalizations as of August 13, 2022 [2], among a population estimated to be 15,007,816 during the second quarter of 2022 [3]. The adaptive evolution of more transmissible and virulent COVID-19 variants associated with different acquired mutations over time may lead to increased case counts, increased mortality rates, and reduced effectiveness of general COVID-19 vaccines [4]. The emergence of novel SARS-CoV-2 variants of concern over time may in part be attributed to both the innate error rate of SARS-CoV-2 replication and the different sources of host somatic mutagenesis that cause nonrandom patterns of mutation types and counts in viral genomes. Some known mechanisms that drive SARS-CoV-2 genomic evolution are commonly associated with host antiviral defenses, including the antiviral activity of (1) apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) family causing C>U nucleotide substitutions, (2) reactive oxygen species (ROS) causing G>U substitutions, and (3) adenosine deaminase acting on RNA (ADAR) causing A>G and U>C substitutions [5,6]. Each host-specific antiviral defense mechanism may generate a unique set of mutation types and abundances over time, known as a mutational signature, which can be used to identify which and to what extent specific mutational processes contribute to all of the mutations observed in each genome [7]. Tracking the abundance of different substitution types over time can provide insights into the contribution of each mechanism of host antiviral defense to nucleotide changes in the SARS-CoV-2 genome [8].

Acquired mutations at specific sites in open reading frames [9] or near N6-methyladenosine (m6A) methylation sites [10] of the SARS-CoV-2 genome may confer advantages in viral transmissibility, host invasion, and reproduction, and modulate the severity of the clinical symptoms of COVID-19. Previous research has identified 8 m6A methylation sites as potential sites of negative regulation of viral infection [10]. Quantifying how the landscape of SARS-CoV-2 mutations in the SARS-CoV-2 genome changes over time can reveal how viral evolution may be associated with specific patterns of mutation burden, mutation types, and genomic locations of mutations, as well as different selection pressures. Moreover, the ratio of nonsynonymous to synonymous mutations in a given open reading frame, known as the dN/dS ratio, can be used to estimate the extent of and change in positive or negative selection of protein-coding regions across the SARS-CoV-2 genome over time.

The initial spread of COVID-19 in China was limited in part due to movement restrictions set in Wuhan in January 2020, followed by the rapid implementation of nonpharmaceutical interventions, including case isolation, physical distancing, wearing face masks, and contact tracing. These interventions were effective at reducing the seropositivity rate below the threshold for an epidemic [11]. However, international travel [12] and a reduction in nonpharmaceutical interventions due to lifting of regional mandates [12,13] have been associated with the spread of novel variants and waves of infections around the world since the initial COVID-19 outbreak in Wuhan. Similarly, changes in public health policy over the course of the pandemic may influence the transmission and emergence of new SARS-CoV-2 variants in Ontario, resulting in fluctuations in the prevalence of variants and genomic diversity reflected in the landscape of novel mutations observed in each epoch.

Previous SARS-CoV-2 genomic studies have characterized the landscape of mutation types and allele frequencies around the world, including but not limited to the United States [14], Qatar [15], the United Kingdom [16], Uruguay [17], and Canada [18]. Public health policies implemented in different countries to reduce COVID-19 transmission may in part influence the emergence of novel SARS-CoV-2 genetic variants [19] and mutation rates that can lead to the evolution of resistance to vaccines [20]. For instance, genomic epidemiology of the first 2 waves of SARS-CoV-2 in Canada revealed that the number of sublineages imported to Canada reduced by over 10-fold when restrictions of foreign nationals were implemented in March 2020. Thus, public health policies and travel restrictions can affect the number of opportunities for novel SARS-CoV-2 lineages to seed new outbreaks or challenge existing lineages. The change in SARS-CoV-2 genomic diversity in the viral population over time is attributed to both the impact of different public health policies and the introduction of novel international variants.

Similar to Canada, Spain employed a multistage nonuniform lockdown, while China employed more immediate and widespread lockdown procedures to reduce COVID-19 transmission compared with Canada [21,22]. Moreover, inconsistent provincial and territorial public health responses to the COVID-19 pandemic in Canada have been reported to be less effective at reducing COVID-19 transmission and SARS-CoV-2 genetic diversity in the Canadian population [23]. Effective epidemiological surveillance requires rigorous and reliable COVID-19 testing of case counts as well as genome sequencing to track the genomic mutations that are associated with increased transmission. A database of SARS-CoV-2 genomic sequences is necessary to track the genomic evolution of SARS-CoV-2 across the world. The Global Initiative on Sharing All Influenza Data (GISAID) is a database of influenza

genomic sequences, and associated clinical and epidemiological data, which has facilitated the rapid public sharing of SARS-CoV-2–related data necessary for genomic surveillance during the COVID-19 pandemic [24].

Thus, characterizing the evolution of the SARS-CoV-2 mutational landscape in the Canadian province of Ontario as a microcosm can reveal in part how Ontario-specific public health decisions and the introduction of novel variants during different time-based epochs may in part be associated with changes in the mutational landscape. This study is the first to characterize the change in the mutational landscape of 24,244 SARS-CoV-2 genomes sampled in Ontario from January 1, 2020, to December 31, 2021, at a large scale and over time, expanding on previous studies by increasing the sample size [25] and reporting the results of SARS-CoV-2 genomic surveillance over successive epochs of time [4]. In line with these studies, we also confirm previous reports of nonneutral selection in open reading frames, such as the spike and nucleocapsid open reading frames, as well as ORF3A in a larger curated data set of SARS-CoV-2 genomes sampled from the Ontario microcosm. Moreover, genomic surveillance is useful for quantifying the diversity of different variants over time as a measure of the effectiveness of public health policies to reduce transmission [26] and make rapid policy changes as well as predict conserved genomic regions undergoing negative selection as promising targets for vaccine design [27].

The surface-exposed SARS-CoV-2 spike glycoprotein, similar to other class I viral fusion proteins, such as influenza virus haemagglutinin and HIV envelope glycoprotein, regulates viral entry into host cells by changing conformation from a metastable unliganded state to a liganded stable state [28,29]. Previous studies have suggested that the spike protein mainly binds to the angiotensin II receptor to enter host cells [30]. In studies of viral challenge by SARS-CoV with similar class I viral fusion proteins as in SARS-CoV-2, polyclonal antibody responses targeting the spike protein were effective in inhibiting viral entry and decreasing viral load [31]. The design of antiviral agents that target the spike protein, a known region of positive selection that influences viral transmission, has been proposed to target the spike protein-angiotensin converting enzyme II binding interface, the auxiliary receptors involved in viral-cell fusion, or the specific epitopes of receptor binding motifs [32].

Recently, conserved pan-variant epitopes across the Alpha, Beta, Gamma, and Epsilon variant spike proteins have been successfully targeted using a neutralizing antibody fragment [33]. The changes in the patterns of mutation types, mutation genomic locations, and selection pressures of open reading frames associated with viral fitness between successive epochs suggest that the prediction of highly confident conserved epitopes and the resulting design of therapeutics with sustained efficacy may remain a significant challenge. The design of specific therapeutic approaches and vaccines for SARS-CoV-2 requires ongoing genomic surveillance to monitor their efficacy in combatting novel as well as increasingly resistant and transmissible SARS-CoV-2 variants [34]. As the COVID-19 pandemic continues, the genomic diversity of SARS-CoV-2 in Ontario may increase owing to the emergence of novel mutations acquired in part due to the host immune response and the increasing pool of infected hosts. Moreover, selection for emergent variants with mutations that confer a fitness advantage, characterized by increased immune escape and transmissibility, may further drive the genomic evolution of SARS-CoV-2 [35,36].

Here, we compared the novel changes in the proportions of single base substitution (SBS) types and the total burden of novel SBSs from SARS-CoV-2 genomes across 4 successive epochs. We tested if observed SBSs in open reading frames cluster near genes associated with increased viral fitness or known m6A sites across 4 successive epochs. Selection using the dN/dS ratios of different coding regions and the diversity of substitution types across the whole genome was compared between the 4 epochs to identify which coding regions were likely conserved. Finally, estimations of the rates of novel SBSs and the change in mutational fitness over time were used to estimate the evolution of the SARS-CoV-2 mutational landscape in Ontario.

This comprehensive study investigating the SARS-CoV-2 mutational landscape in the Ontario microcosm provides a foundation for research into simulating genomic epidemiology parameterized using both genomic and public health factors to inform public health decision-making as well as predict the activity of mutational processes and selection pressures that drive SARS-CoV-2 genomic evolution.

Methods

Data Collection

Data for 24,244 complete (>29,000 base pairs), high coverage (<1% N's), and nonduplicate SARS-CoV-2 genomic sequences in FASTA format sampled from January 1, 2020, to December 31, 2021, were retrieved from the GISAID database on January 1, 2022 [20]. Metadata associated with each genomic sequence, including sampling date and GISAID clade assignment [37], were also retrieved from the GISAID database. Duplicate sequences were removed if the FASTA header and associated metadata were the same. Whole-genome alignment of SARS-CoV-2 genomic sequences to the 29,903 base-pair reference genome (GenBank accession number MN908947.3) was conducted using MAFFT (Multiple Alignment with Fast Fourier Transform) version 7 with default parameters and keeping the alignment length between the sequence and the reference genome. Custom code was used to identify SBSs observed in aligned SARS-CoV-2 genomic sequences compared to the reference genome by iteratively checking if the base type at each position of the genomic sequence and reference genome was different, and if so, updating the data table with the observed SBS position, the reference base, and the alternate base for downstream analysis of mutational patterns. Custom code used in this study was implemented using Python version 3.8.0. The final position of each observed SBS is referenced based on the reference genome position. Minor allele frequency, defined as the proportion of the SARS-CoV-2 population sampled during a given epoch with the allele, was calculated for each minority allele at each base position.

Sequences were assigned to 4 different epochs based on time as follows: Epoch 1 from April 1, 2020, to August 31, 2020 (n=2256); Epoch 2 from September 1, 2020, to February 28, 2021 (n=4443); Epoch 3 from March 1, 2021, to August 31, 2021 (n=12,864); and Epoch 4 from September 1, 2021, to December 31, 2021 (n=4102) (Figure 1). The start of each epoch closely coincided with major public health events that may be associated with reduced opportunities for viral transmission (eg, province-wide lockdowns with closure of nonessential businesses, introduction of international travel restrictions, and warmer weather) or increased viral transmission (eg, reopening of schools, lifting of international travel restrictions, and cooler weather), including the Ontario State of Emergency declaration in Epoch 1, the 2020 return to school in Epoch 2, expanded COVID-19 vaccine eligibility in Epoch 3, and the 2021 return to school in Epoch 4. Epoch-specific genomic mutations, defined as mutations in 1 epoch not observed in any previous epoch, were used for the downstream mutational pattern, clustering, selection, and diversity analyses. Mutations specific to Epoch 1 were determined by identifying mutations not previously observed in SARS-CoV-2 genomic sequences sampled from January 1, 2020, to March 31, 2020, in the collected data set (n=579).

Figure 1. Timeline of the 4 successive epoch time periods annotated with major public health events from April 1, 2020, to December 31, 2021. The public health events are relevant in the timeline given their impact on virus spread. The number of new COVID-19 case counts (blue) and number of vaccine doses administered (red) are plotted over time. The Ontario State of Emergency Declaration (March 17, 2020) instituted a province-wide lockdown imposing severe restrictions on human behavior with the intent to greatly limit the spread of infection. Re-entry from lockdown was phased in steps through the provincial Roadmap to Reopen. Significant public health policies associated with changes in viral transmission in the timeline include the Roadmap to Reopen and return to school.
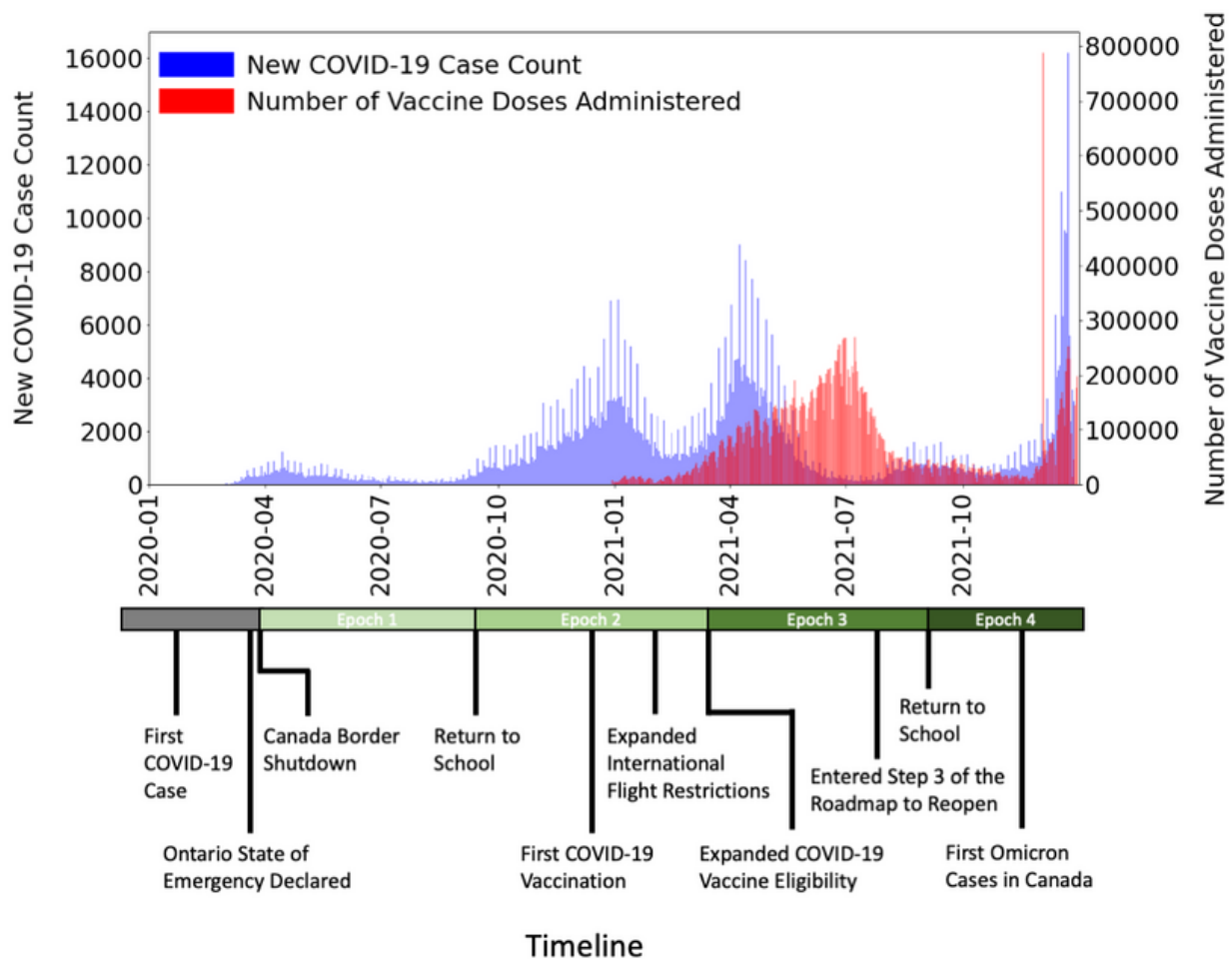
Ethical Considerations

We confirm that all secondary analyses of research data from the GISAID database were performed in accordance with relevant usage guidelines and regulations. The genomic data retrieved from the GISAID database were deidentified and could not be linked to patients' identities [24].

Generation of SBS-96 Mutational Catalogues

Each SBS for each SARS-CoV-2 genomic sequence was assigned to 1 of 96 possible classes, where each class was defined by 6 base substitutions represented by the pyrimidine of the Watson-Crick base pair (C>A, C>G, C>U, U>A, U>C, and U>G) and the flanking 5′ and 3′ bases of the SBS that forms the local trinucleotide context [7]. For example, both the C>A pyrimidine substitution and G>U purine substitution were referred to by their pyrimidine base pair, C>A. For each

SARS-CoV-2 genomic sequence, a mutational catalogue of the count of each of the 96 SBSs observed was generated.

To compare the mutational catalogues between multiple sequences in each epoch, the mean proportion of each of the 96 SBS types for each epoch was generated. First, the proportion of each of the 96 SBS types was calculated by dividing the count of each of the 96 SBS types by the total count of all SBS mutations observed in each SARS-CoV-2 genomic sequence. Second, the mean proportion of each of the 96 SBS types was calculated by summing the proportion of each of the 96 SBS types across all SARS-CoV-2 genomic sequences assigned to the same epoch and dividing the sum by the number of SARS-CoV-2 genomic sequences in the epoch.

Somatic Mutation Clustering Near Open Reading Frames and m6A Methylation Sites

The distribution of sites where novel SBSs were observed along the whole genome was analyzed to detect the presence of SBS clusters in specific open reading frames or near known m6A methylation sites [10]. Open reading frames with observed clusters of epoch-specific mutations were identified if the positions of SBSs in the open reading frame were different from a random sample of 100 positions in the open reading frame using a 2-sample Kolmogorov-Smirnov test. Eight known m6A methylation sites of the SARS-CoV-2 genome were previously identified [10]. Observed clusters of epoch-specific mutations near m6A methylation sites were identified if the positions of SBSs within a window of ±1500 base pairs of one of the m6A methylation sites were different from a random sample of 100 positions in the same window using a 2-sample Kolmogorov-Smirnov test. The ±1500 base-pair window flanking each m6A methylation site was chosen based on a previous study that first identified the m6A methylation sites and reported clustering statistics near these sites using the same ±1500 base-pair window [10].

dN/dS Ratio of Open Reading Frames

To quantify the change in the selection pressures of each open reading frame over time, the mean dN/dS ratio of each open reading frame was compared between different epochs. First, we calculated the proportion of synonymous (pS) and nonsynonymous SBSs (pN) by dividing the count of pS and pN by the count of synonymous and nonsynonymous sites, respectively, for each open reading frame. Next, we estimated the count of synonymous (dS) and nonsynonymous (dN) substitutions per site using the Nei and Gojobori method shown

in Equations 1 and 2 [38,39]. Null values of the dN/dS ratio, such as when zero nonsynonymous substitutions in an open reading frame were observed in a given SARS-CoV-2 genomic sequence, were excluded from the analyses.

Inline graphic 1

Inline graphic 2

Site-Specific Shannon Diversity Index

To compare the differences in genetic diversity at each genomic site over time, the Shannon diversity index was calculated at each genomic site for SARS-CoV-2 genomic sequences in each epoch. The mean diversity index value across the whole genome and site-specific diversity index values were compared between SARS-CoV-2 genomic sequences from different epochs [40].

Nextstrain: Annual Rate of Novel SBSs and Estimation of Mutational Fitness

Out of 24,244 SARS-CoV-2 genomic sequences included in this study, 7398 SARS-CoV-2 genomic sequences with complete sampling dates, including day, month, and year, were used in the Nextstrain Augur 18.0.0 and Auspice 2.32.1 pipelines [41] for phylogenetic analysis and visualization, respectively [42]. The Nextstrain set of pipelines was used to estimate the rate of novel substitutions per year and the change in mutation fitness over time. Mutation fitness is defined as a metric that predicts viral reproduction and transmissibility based on the contribution of multiple somatic mutations that have been annotated to affect lineage growth, such as mutations that confer a fitness advantage by evading host immunity or decreasing generation time [43].

Results

Host Antiviral Defense Mechanisms are Associated With Nonrandom Biases in the Prevalence of Different SBS Types

Patterns of mutation types and their proportions commonly associated with mutagenesis due to host antiviral defense mechanisms were quantified by classifying the novel SBSs observed in each epoch into the SBS-96 classification scheme. Next, the SBS-96 mutation types and mean proportions were compared between different epochs (Figure 2). There were 1767 unique epoch-specific SBSs in Epoch 1, while there were 3573, 4822, and 2076 such SBSs in Epochs 2, 3, and 4, respectively (Multimedia Appendix 1).

Figure 2. The mean proportion of single base substitution (SBS)-96 mutation types of each of the 4 different epochs. The mean proportion shows, on average across all SARS-CoV-2 genomic sequences sampled during the epoch, what proportion each SBS type makes up out of all 96 possible SBS types. At least 80% of novel SBSs in each epoch were C>U, C>A, or U>C substitutions that have previously been attributed to the activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide-like, reactive oxygen species, and adenosine deaminase acting on RNA, respectively.
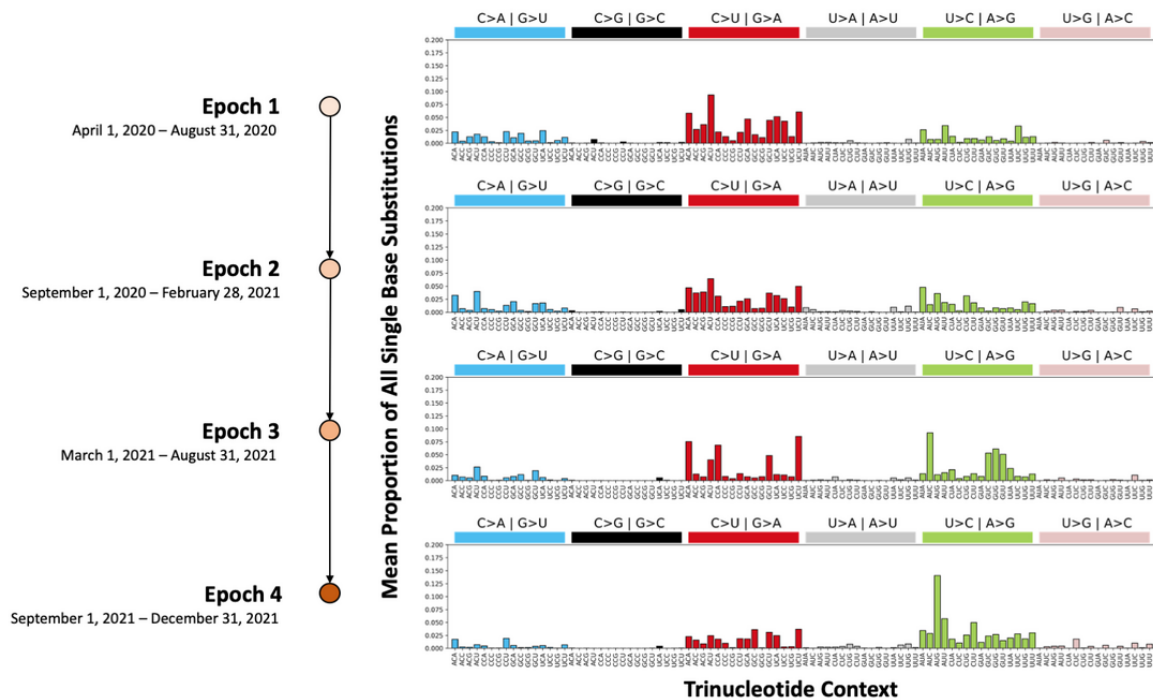
Across each of the 4 epochs, at least 80% of the novel SBSs were C>U, C>A, or U>C substitutions that have previously been attributed to the activity of APOBEC, ROS, and ADAR, respectively. Moreover, we observed that pyrimidine substitutions were more prevalent between each pair of purine and pyrimidine substitutions, constituting between 55.4% and 59.2% of all substitutions. The mean proportion of the C>A substitution type increased from 13.5% of all substitutions in Epoch 1 to 22.4% in Epoch 4. Conversely, the mean proportion of the C>T substitution type decreased from 67.9% in Epoch 1 to 48.4% in Epoch 4. The mean proportion of the T>C substitution type was relatively consistent among Epochs 1 to 3, ranging from 16.8% to 17.1% of all substitutions observed in each respective epoch. During Epoch 4, there was an increase in the mean proportion of the U>C substitution type at the AUG trinucleotide context compared to previous epochs. An increase in the mean proportion of the U>C substitution type to 19.2% of all substitutions was observed in Epoch 4 compared to Epochs 1 to 3. Visualizing each SARS-CoV-2 genomic sequence using a 3D uniform manifold approximation and projection plot of the SBS-96 mutation types and counts showed that genomic sequences sampled from the same epoch tended to uniquely cluster together (Multimedia Appendix 2).

Average Burden of Novel SBSs Differs Between Successive Epochs

To analyze the variation in the number of observed mutations over time, the average cumulative SBS mutational burden and average burden of novel SBSs first observed in each epoch were compared between different epochs. Comparing the clade composition of different epochs, we observed that clade G (Delta) and GR (Gamma) sequences made up the majority of sequences sampled in Epochs 1 and 2, clade GRY (Alpha) and GK sequences made up the majority of sequences sampled in Epoch 3, and clade GK (Delta) sequences were the sole majority of sequences sampled in Epoch 4 (Figure 3A). We observed that across successive epochs, the cumulative number of observed SBSs increased as expected, with a median cumulative mutational burden of 10 SBSs in Epoch 1, 21 SBSs in Epoch 2, 37 SBSs in Epoch 3, and 41 SBSs in Epoch 4 (Figure 3B).

Figure 3. Mutational burden of single base substitutions (SBSs). Global Initiative on Sharing All Influenza Data (GISAID) clade assignments for each SARS-CoV-2 genomic sequence are based on known marker mutations associated with 8 high-level phylogenetic groupings. (A) The number of SARS-CoV-2 genomic sequences colored by GISAID clade assignment and sampled from each of the 4 epochs. (B) The distribution of the total number of SBSs observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. (C) The count distribution of the number of novel epoch-specific SBSs first observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. (D) The density distribution of the number of novel epoch-specific SBSs first observed in SARS-CoV-2 genomic sequences colored by GISAID clade assignment and sampled from each of the 4 epochs.
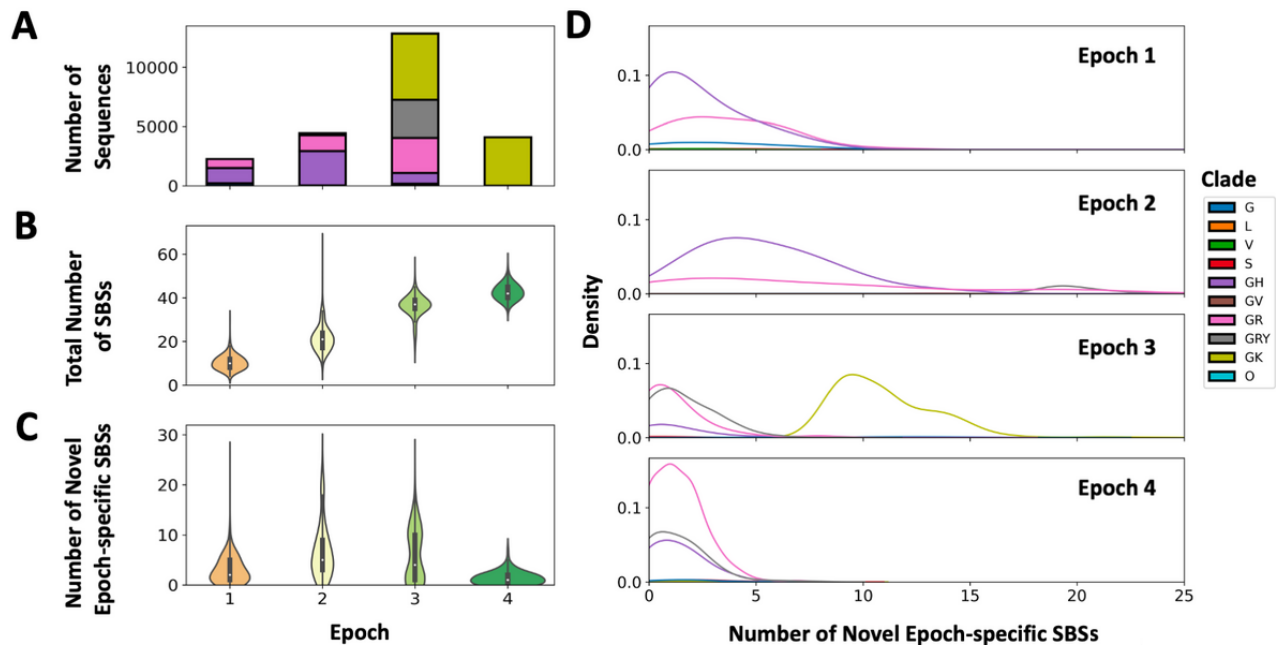
In addition, we observed the greatest maximum number of novel SBSs in Epoch 1 (maximum 26; median 2), followed by Epoch 2 (maximum 25; median 5), Epoch 3 (maximum 24; median 4), and Epoch 4 (maximum 8; median 1) (Figure 3C). The distribution of the mutational burden of novel SBSs in Epochs 2 and 3 was bimodal, with 2 different clade-specific populations of SARS-CoV-2 genomes with different mean counts of novel SBSs observed. In contrast, Epoch 1 and Epoch 4 showed a unimodal distribution of novel SBSs across all genomes sampled during the time period.

In Epoch 2, SARS-CoV-2 genomes from the GRY clade reported a higher burden of novel SBSs (median 19), while GH and GR clade genomes reported a lower burden of novel SBSs (GH median 5; GR median 5) (Figure 3D). In Epoch 3, SARS-CoV-2 genomes from the GR, GRY, and GH clades

reported a lower burden of novel SBSs (GR median 1; GRY median 1; GH median 1), while GK clade genomes reported a higher burden of novel SBSs (median 10) (Figure 3D).

Clusters of Novel SBSs and Selection of Open Reading Frames may be Associated With Viral Fitness

To identify open reading frames on SARS-CoV-2 genomes sampled in Ontario that may be associated with viral fitness, we observed the presence of clusters of novel SBSs in specific open reading frames compared between different epochs (Figure 4A). We consistently observed clusters of epoch-specific SBSs in the spike protein during Epochs 1, 2, and 3, but not during Epoch 4. Clusters of epoch-specific SBSs were observed in ORF1a during Epochs 2, 3, and 4. Likewise, clusters of epoch-specific SBSs were observed in ORF3a during Epochs 2 and 3, but not Epoch 4.

Figure 4. Minor allele frequency and clustering of novel epoch-specific single base substitutions (SBSs). (A) Scatterplot of the base position and the log-transformed minor allele frequency of novel epoch-specific SBSs first observed in each of the 4 epochs. The annotation of the SARS-CoV-2 genomic elements is shown above each plot. (B) Heatmap of the Kolmogorov-Smirnov test *P* value comparing the base positions of novel epoch-specific SBSs observed in each open reading frame of the SARS-CoV-2 genome and 100 randomly sampled base positions in the open reading frame, colored by *P* value (*P*<.05 is shown in red). Clusters of epoch-specific SBSs were observed in the spike protein, ORF1a, and ORF3a open reading frames during multiple epochs.
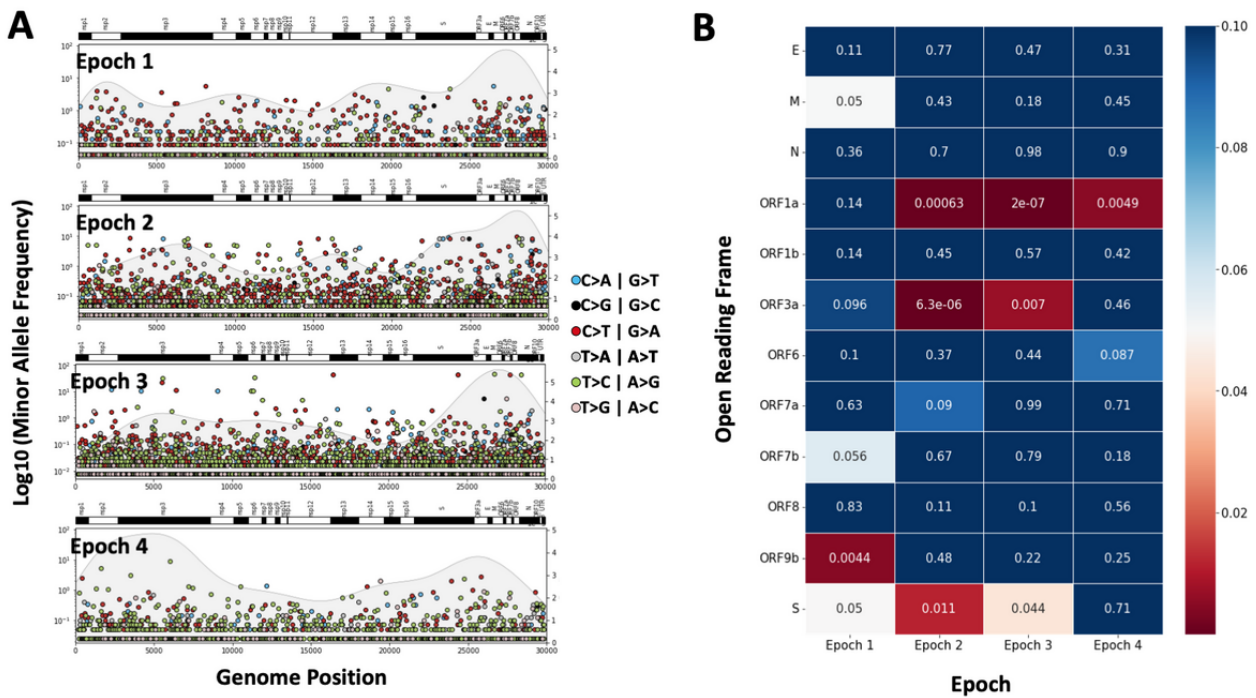
In addition, we tested for the existence of epoch-specific clusters of novel SBSs compared to a randomly sampled distribution of 100 substitution positions within a window of ±1500 base pairs of each of the 8 known m6A methylation sites using the 2-sample Kolmogorov-Smirnov test. In Epoch 2, there was 1 m6A methylation site at position 27525 in ORF6 of the SARS-CoV-2 reference genome with a significant cluster of novel SBSs (Figure 5A). In Epoch 3, there were 4 m6A methylation sites at position 27525 in ORF6 as well as positions 29428, 29450, and 29522 in ORF10 with a cluster of novel SBSs (Figure 5A). To confirm these findings at each of the 5 m6A methylation sites with potential clusters of novel SBSs within the window, we observed that the density of the positions of mutations did show nonrandom clustering with peaks of density that differed from the randomly sampled distribution (Figure 5B).

Figure 5. Clustering of novel epoch-specific single base substitutions (SBSs) within a ±1500 base-pair window around 8 N6-methyladenosine (m6a) methylation sites identified in the SARS-CoV-2 genome. (A) Heatmap of the Kolmogorov-Smirnov test *P* value comparing the base positions of the novel epoch-specific SBSs observed in each ±1500 base-pair window of 8 m6a methylation sites and 100 randomly sampled base positions in the same window, colored by *P* value (*P*<.05 is shown in red). (B) Density distribution of the novel epoch-specific SBSs (blue) and 100 randomly sampled base positions (orange) in the ±1500 base-pair window of 1 m6a methylation site in Epoch 2, and 4 m6a methylation sites in Epoch 3. The density distribution of SBSs is shown for 4 unique m6a methylation sites across 2 epochs since potential clusters of SBSs were detected within the ±1500 base-pair window around the m6a methylation sites.
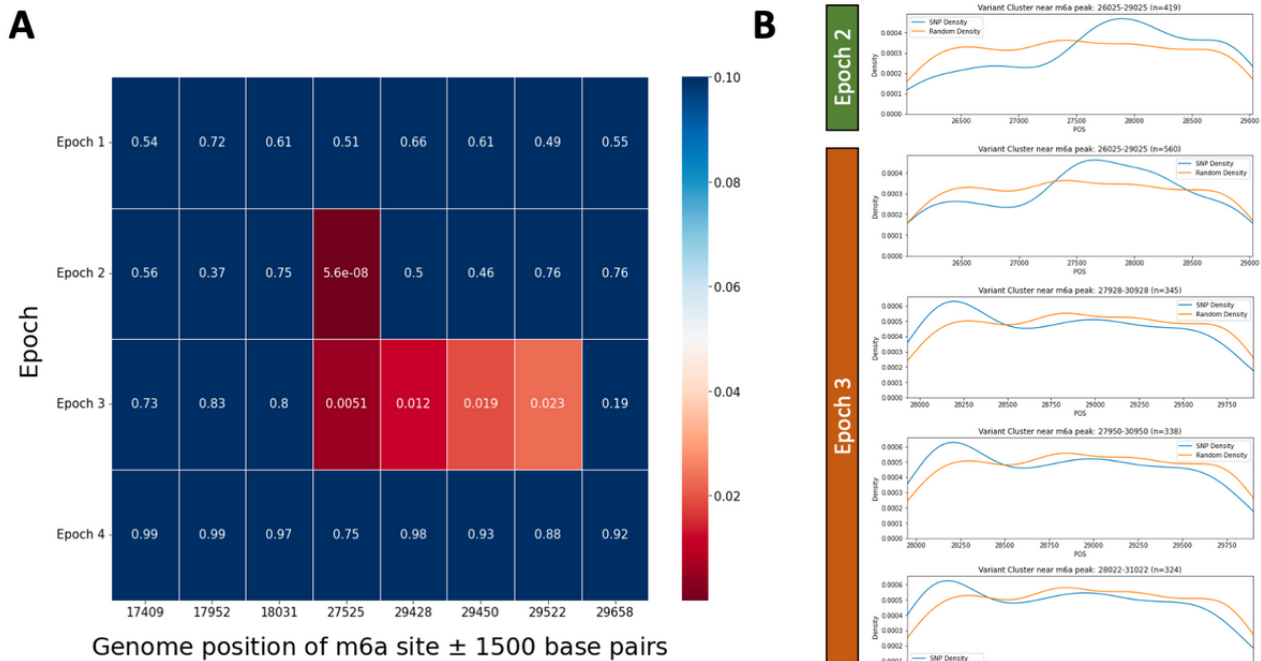
To quantify the selection pressures on different open reading frames of the SARS-CoV-2 genome over time, we compared the dN/dS ratio for each open reading frame between different epochs. We observed that across successive epochs, there was an increase in the median dN/dS ratio from −0.5 in Epoch 1 to 0.5 in Epoch 4 for the spike protein open reading frame (Figure 6). This suggests an increase in positive selection for nonsynonymous mutations in the spike protein. Conversely, the median dN/dS metric of ORF1b was consistently below 0 across all epochs, indicating negative selection for nonsynonymous mutations in ORF1b (Figure 6).

Figure 6. Distribution of the dN/dS metric (ratio of nonsynonymous to synonymous mutations in a given open reading frame) calculated from the novel epoch-specific single base substitutions observed in SARS-CoV-2 genomic sequences sampled during the 4 different epochs, grouped by open reading frame. The numbers of SARS-CoV-2 genomic sequences where a nonnull dN/dS metric could be calculated for the open reading frame are shown.
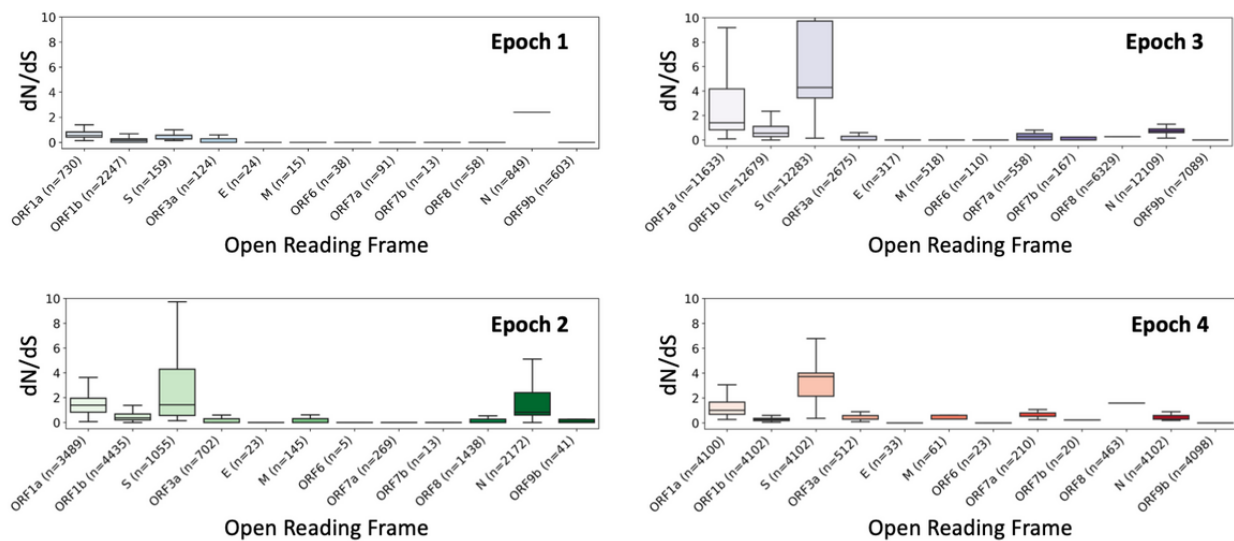
To characterize the diversity of different sites in the SARS-CoV-2 genome based on the different mutation types observed at the same site, we calculated the Shannon diversity index at each base position along the SARS-CoV-2 whole

genome (Multimedia Appendix 3). Then, we qualitatively compared the overall profile of Shannon diversity indices between different epochs to identify if diverse mutation types tended to be observed at specific open reading frames. We found that the most recent Epoch 4 showed a relatively broad diversity across the entire genome instead of peaks of high diversity near ORF6a, ORF7a, ORF7b, ORF8, and nucleocapsid protein observed in the previous 3 epochs. Moreover, Epoch 4 most closely resembled Epoch 1 in the shared peak of site-specific diversity in nonstructural protein (NSP) 2 of ORF1a.

Increases in SARS-CoV-2 Mutational Fitness Across Divergent Lineages Occur in Spurts

To estimate the rate of novel substitutions per year and potential divergence events indicated by changes in mutational fitness,

we built a rooted maximum likelihood phylogenetic tree using the Nextstrain Augur pipeline with 7398 SARS-CoV-2 genomic sequences having complete sampling dates sampled in Ontario (Figure 7). Eighteen unique clades were represented in the tree, including several variants of concern, such as Delta (21A, 21I, 21J), Gamma (20J), Alpha (20I), and Beta (20H) (Figure 6). The tree shows that the COVID-19 pandemic in Ontario is caused by multiple different lineages of SARS-CoV-2 viruses, many of which are also concurrently observed in the same epoch. The positive linear rate of acquiring novel SBSs was observed to be around 23 mutations per year.

Figure 7. Mutational fitness of 7398 SARS-CoV-2 genomic sequences over time, colored by variant type identified using the Augur pipeline. Mutation fitness is a relative unitless metric that compares the estimated fitness of a given SARS-CoV-2 genomic sequence compared to the Wuhan reference genome, based on mutations annotated to be associated with a change in fitness. Higher mutational fitness suggests increased viral reproduction and transmissibility. Four high-level clusters of SARS-CoV-2 genomes were observed based on similarities in mutational fitness and SARS-CoV-2 variant composition. Clusters are denoted by a red marker at the earliest cluster-specific phylogenetic branch point.
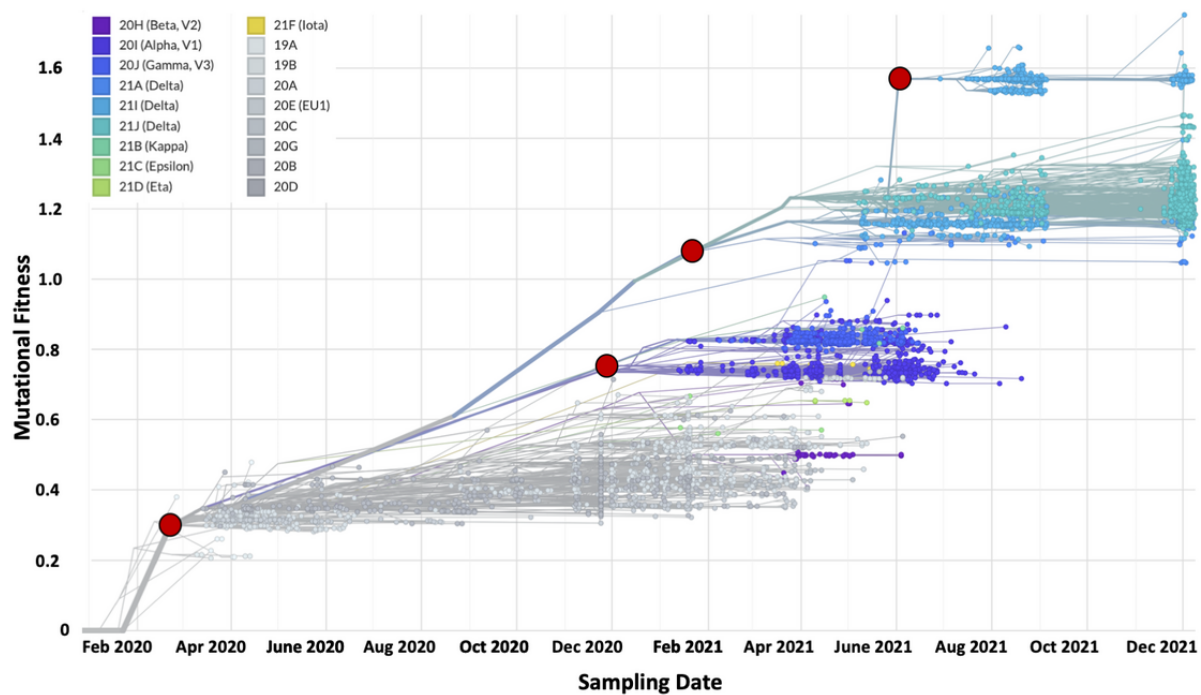
The phylogenetic analysis revealed the timepoints of divergence of SARS-CoV-2 clades based on mutational fitness into 4 clusters. Cluster 1 was marked by an early divergence in mutational fitness in March 2020 of Epoch 1 based on branch length compared to the reference genome. Cluster 2 was predominantly composed of Alpha and Gamma variants of concern that diverged from Cluster 1 in December 2020 of Epoch 2. Following this, Cluster 3 was predominantly composed of Delta variants of concern that diverged from Cluster 2 in April 2021 of Epoch 3. Lastly, Cluster 4 was the most divergent cluster based on branch length and showed the highest mutational fitness, consisting of Delta variants of concern associated with the 20I clade observed from July 2021 in Epochs 3 and 4.

Discussion

In this study, we quantified the changes in the landscape of SBSs observed in SARS-CoV-2 genomic sequences sampled from 4 successive epochs in Ontario from March 2020 to December 2021 during the COVID-19 pandemic. Among sequences sampled during Epoch 1, we observed positive selection and a high median dN/dS ratio above 1 in the nucleocapsid protein–coding region. Moreover, we observed clustering of novel epoch-specific SBSs, positive selection, and higher base-specific Shannon diversity in the spike protein– and nucleocapsid protein–coding regions near the 3′ end of the SARS-CoV-2 genome. During Epoch 3, we observed that there was a bimodal distribution of the number of novel epoch-specific SBSs in the GK clade and a lower number of novel

epoch-specific SBSs in the GH, GV, and GRY clades. Similar to Epoch 2, there was a higher prevalence of novel SBSs in the spike protein and ORF3a-coding regions as well as positive selection of ORF1a and the spike protein–coding region. Finally, we uniquely noted that there was an increase in the mean proportion of the U>C substitution type up to 19.2% of all substitutions, especially at the AUG trinucleotide context, in Epoch 4 compared to previous epochs.

We observed that the open reading frames of SARS-CoV-2 viruses, including ORF1a and the spike protein, were increasingly impacted by positive selection over time, and this was in part consistent with mutations associated with host antiviral defenses. Future therapies can target these regions of positive selection since they affect the fitness of the SARS-CoV-2 virus [44]. Previous studies have similarly observed positive selection in the SARS-CoV-2 spike protein open reading frame [44,45] and ORF1a [46,47]. SARS-CoV-2 ORF1a and ORF1b genomic RNAs are translated and cleaved into 16 NSPs [48] associated with viral genomic replication as well as suppression of host immune response and gene expression [49]. For example, the NSP 7, 8, and 12 complex forms viral replicase machinery [50,51] and the NSP 10-16 complex is involved with capping viral mRNA transcripts and immune response evasion [52]. Due to the broad function of multiple NSPs translated from ORF1a for viral replication and the modulation of host immune response, future research to screen or design novel pharmaceutical drugs that target specific NSPs may show promise [53]. However, novel nonsynonymous mutations associated with positive selection affecting ORF1a and spike coding regions may not be conserved over time, so new iterations of pharmaceutical drugs, vaccines, or antigen tests targeting these regions may be required to maintain specificity and efficacy.

In Ontario, Epoch 1 marked the start of the pandemic from when the Canadian borders were shut down in late March 2020 to the return to school in October 2020. Similar to other regions around the world, such as the Baltimore-Washington area in the United States [54], the early measures designed to reduce local transmission of COVID-19 may have been compromised in part by the introduction of COVID-19 from national and international sources as well as the interconnectedness of the Ontario region. The high median dN/dS ratio above 1 in the nucleocapsid protein–coding region suggested that positive selection for nonsynonymous substitutions drove the early genomic evolution of the nucleocapsid protein and its function in viral genome packaging, despite being generally understood as a conserved region of the coronavirus genome [55]. Likewise, the high Shannon diversity index peaks near NSP1 and NSP2, which are involved in viral gene expression [56] and RNA binding [55], respectively, as well as the 3′ end of the SARS-CoV-2 genome, indicate a high prevalence of multiple different substitution types observed at a single genomic site. Since Epoch 1 is the first time period of the COVID-19 pandemic, the negative selection observed at all other open reading frames aside from the nucleocapsid protein–coding region may be attributed to the lag between the observation of a deleterious mutation and its subsequent selective removal from the gene pool [57]. Thus, SARS-CoV-2 genomic evolution associated with open reading

frames involved in viral genome packaging and the resulting modulation of immune response were likely early events during Epoch 1 in the Ontario microcosm.

Epoch 2 was the time period from the September 2020 return to school and the introduction of the first COVID-19 vaccines in December 2020 to the expansion of COVID-19 vaccine eligibility to the general public in February 2021. Previous studies have suggested that COVID-19 vaccines likely play a role in initially reducing the genomic diversity of SARS-CoV-2 [58] and that early vaccine candidates targeting the spike protein involved in viral entry would likely be therapeutically effective against SARS-CoV-2 variants in Epoch 2 [59]. SARS-CoV-2 variants associated with clades GH and GR were the majority populations observed in both Epochs 1 and 2, but the introduction of variants associated with clade GRY (UK B.1.1.7 strain) was unique to Epoch 2 [60]. Similar to Epoch 1, clustering of novel epoch-specific SBSs and high base-specific Shannon diversity near the 3′ end of the SARS-CoV-2 genome was observed. Coupled with the positive selection observed in the spike protein– and nucleocapsid protein–coding regions near the 3′ end of the SARS-CoV-2 genome, variants observed in Epoch 2 showed increased genomic diversity in regions associated with viral attachment and entry as well as RNA genome packaging [61]. As time progressed in Epoch 2, selection pressure against SARS-CoV-2 variants with relatively decreased transmission, rate of replication, or immune defense evasion may have driven genomic evolution in favor of variants with novel mutations associated with increased fitness. Taken together, the introduction of vaccines targeting the spike protein is consistent with selection for novel mutations in the spike protein open reading frame introducing less virus susceptibility to vaccine-induced immune responses during Epoch 2 in the Ontario microcosm.

Epoch 3 spanned from March 2021 following the expansion of COVID-19 vaccine eligibility in Ontario to the September 2021 return to school. By August 5, 2021, 72% of Canadians had received 1 or more doses of a COVID-19 vaccine and 61% of Canadians were fully vaccinated with 2 doses, which were comparable to vaccination statistics observed in Ontario by this time [62]. Moreover, there was increasing demand on intensive care unit resources, and implementation of stay-at-home orders in Ontario as well as federal-mandated COVID-19 testing and 14-day quarantine of international air travelers to Canada were new measures to reduce COVID-19 case counts [62]. The emergence of GK clade SARS-CoV-2 variants with relatively high counts of novel epoch-specific SBSs and GH, GV, and GRY clade SARS-CoV-2 variants with relatively lower counts of novel SBSs comprised the bimodal distribution in the number of novel epoch-specific SBSs across all variants sampled in Epoch 3. Moreover, Epoch 3 variants were characterized by a relatively high prevalence of novel SBSs observed in the spike protein– and ORF3a-coding regions similar to Epoch 2. Likewise, both ORF1a and the spike protein–coding region showed a dN/dS ratio above 1, suggesting continued positive selection associated with nonsynonymous mutations in genomic regions involved with viral transmission and immune evasion. Moreover, the peak in the site-specific Shannon diversity index near the 3′ end of the SARS-CoV-2 genome indicates that the

nonsynonymous mutations in the spike protein–coding region are diverse in the observed alternate base at each mutation site. Population mixing among vaccinated and unvaccinated populations is consistent as a contributor to the increased infection rates among vaccinated individuals than expected in a fully vaccinated population [63]. Therefore, the observed positive selection associated with novel mutations in genomic regions with impact on SARS-CoV-2 transmission and immune evasion may have been driven in part by viral evolution in the human host population with variable immune responses due to the heterogeneity of individual vaccination statuses.

Epoch 4 was from the September 2021 return to school to the end of December 2021. This epoch followed the start of Step 3 of the Roadmap to Reopen, allowing for increased numbers of people at indoor and outdoor gatherings and increased capacity at nonessential venues with the requirement of face coverings in indoor settings [64]. The first case of the emergent SARS-CoV-2 Omicron variant was identified on November 22, 2021, during Epoch 4 [65]. Interestingly, we noted that there was an increase in the proportion of U>C substitutions in SARS-CoV-2 genomes sampled from Epoch 1 to Epoch 4. Our finding is consistent with a previous report of ADAR-induced editing of A>G and complementary T>C substitutions as mutations observed more commonly in genomes sampled from late 2020 onwards [66]. Moreover, the degree of RNA deamination has been reported as a potential determinant of viral immunogenicity and infectivity in emergent minor viral populations, warranting further investigation into RNA deamination as the main driver of SARS-CoV-2 genomic evolution [66]. Compared with Epoch 3, the lower median number of novel epoch-specific SBSs observed in Epoch 4 variants and the lower dN/dS ratio of the spike protein–coding region may be due to a combination of the shorter time period, a decrease in the mutation rate, and a reduction in positive selection pressure for nonsynonymous mutations. SBSs unique to Epoch 4 were clustered in ORF1a, namely NSP2- and NSP3-coding regions associated with viral replication, and were predominantly nonsynonymous mutations as evidenced by a dN/dS ratio above 1. These findings confirm a previous report of positive selection driving the genomic evolution of NSP2 and NSP3, and the high transmissibility of COVID-19 [67]. Compared to previous epochs, the marked increase in nonsynonymous mutations and relatively higher dN/dS ratio above 2 in the NSP8 region of variants sampled in Epoch 4 may be potential mechanisms for increasing stability of the SARS-CoV-2 viral replication and transcription complex [68]. Further research is required to determine the set of genomic mutations unique to Omicron variant genomes that may provide further insights into its mechanisms of increased transmissibility, immune evasion, and decreased pathogenicity [69]. The mutation fitness of SARS-CoV-2 genomic sequences sampled in Ontario was observed to increase in spurts over short time periods, likely coinciding with the introduction of novel SARS-CoV-2 variants with acquired genomic mutations that confer a fitness advantage [70]. Thus, future research predicting the functional impact of different sets of mutations on fitness could improve the surveillance of emergent SARS-CoV-2 variants for public health and inform the design of specific antiviral therapies [71].

This study used specific dates associated with the enactment of government public health policies to examine subsequent epoch-specific mutational patterns in SARS-CoV-2 genomic sequences. The nature of this study does not permit assessment of causation between government public health policies and mutational patterns due to the potential for other contributing and potentially confounding factors, including regional weather patterns, time lag between instantiation of public health policy and practical implementation, and the development of natural and vaccine immunity in the population. Future studies may identify specific time periods when these additional factors are impactful and assess their association with mutational patterns.

As the COVID-19 pandemic continues, there is a possibility of co-infection with other respiratory pathogens [72], as well as reinfection or co-infection with multiple different variants of SARS-CoV-2 [73]. Moreover, successive selective sweeps caused in part by both mutations that confer increased fitness [74,75] and homologous recombination may give rise to novel variants [76], such as the BA.2 Omicron variant. Thus, further research into the clinical impacts of co-infection and reinfection with different SARS-CoV-2 strains may highlight the interplay between genomic variation and COVID-19 symptoms and severity.

Another consideration is the impact of COVID-19 seasonality due to regional differences in environmental factors, including temperature and humidity, that can influence viral transmission, the diversity of SARS-CoV-2 variants selected based on tolerance to different environmental conditions, and the resulting case counts [77]. Interestingly, the annual winter influenza peaks in case counts reduced during 2020 and 2021, suggesting that COVID-19–related public health measures may impact the seasonal transmission of other respiratory viruses [78]. Thus, the development of public health policies should take into account the variation in the seasonality of COVID-19 and other respiratory viruses so that health care systems are prepared for fluctuations in case counts. Further surveillance of SARS-CoV-2 genomic variation and transmission patterns across Ontario can inform effective public health decision-making and serve as a microcosm of the COVID-19 pandemic as the case count of the novel Omicron variant increases. Future design of specific antiviral therapeutics should consider ongoing genomic surveillance as a tool to identify candidate targets [79,80].

In summary, we uniquely noted a bimodal distribution in epoch-specific counts of SBSs in sequences sampled during Epoch 3, where there was a high count observed in GK clade sequences and a lower count observed in GH, GV, and GRY clade sequences. Moreover, we uniquely observed an increase in the mean proportion of the U>C substitution type up to 19.2% of all substitutions, especially at the AUG trinucleotide context, in Epoch 4 compared to previous epochs. We confirmed previous reports of positive selection and clustering of SBSs near or within the ORF1a-, nucleocapsid protein–, and spike protein–coding regions.

We characterized the mutational profile of 24,244 SARS-CoV-2 genomic sequences sampled from January 1, 2020, to December 31, 2021, in Ontario, Canada. Our findings highlight how SARS-CoV-2 genomic sequences sampled from different epochs

harbor different patterns in mutational types, counts, and clusters that may be associated with differences in the transmissibility and virulence of SARS-CoV-2. Nonrandom biases in the abundance of different SBS types are consistent with the activity of host antiviral defense mechanisms and are in agreement with previous reports of the impact of host antiviral defense activity on the SARS-CoV-2 genome. Clusters of epoch-specific SBSs were observed in the spike protein, envelope protein, membrane protein, ORF3a, ORF6, and ORF7a open reading frames across all epochs, as well as near 4 unique m6A methylation sites during Epochs 2 and 3. Positive selection of the spike protein open reading frame, responsible for encoding the spike protein involved in viral entry, was observed. The estimated mutational fitness of SARS-CoV-2 genomic sequences increased in short-term spurts over time, suggesting that only a subset of somatic mutations confers a fitness advantage.

The microcosm of Ontario uniquely focuses on the evolution of the SARS-CoV-2 mutational profile associated with Ontario-specific public health events and policies. The mutational analysis of SARS-CoV-2 genomic sequences can in part reflect the impact of different public health policies during different epochs, such as the limiting of travel across Canadian borders in Epoch 1, and the impact on the genetic diversity of the SARS-CoV-2 viral population. This study of the mutational profile of SARS-CoV-2 in Ontario may serve as a model of the evolution of the SARS-CoV-2 mutational profile for comparison with other regions around the world that have implemented similar or different public health policies.

Further research of therapeutic agents designed to target conserved epitopes under negative selection, such as ORF1b, may shed light on how genomic surveillance can be a useful tool to inform the development of more effective antiviral therapies. Simulation tools used to project the evolution of SARS-CoV-2 genetic diversity due to somatic mutations or prediction models of COVID-19 waves may be parameterized using the mutational profiles and time points observed from SARS-CoV-2 sequences included in this study. To track the emergence of novel SARS-CoV-2 variants with reduced vaccine efficacy in the future, increased genomic surveillance is required, and it will inform public health policies associated with vaccine boosters as well as the implementation of nonpharmaceutical interventions such as wearing face masks and physical distancing.

Data Availability

The open-source analysis code, sequence data sets, and output data sets are available at http://github.com/HillLab/SARS_CoV_2_Ontario_Genomic_Surveillance under the terms of the Creative Commons Attribution 4.0 International License.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The number of unique and shared single base substitutions observed in SARS-CoV-2 genomic sequences sampled from each of the 4 epochs.
[PNG File , 106 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Three-dimensional uniform manifold approximation and projection plot of the single base substitution-96 mutation types and counts observed in 24,244 SARS-CoV-2 genomic sequences sampled in Ontario, colored by the epoch when they were sampled. Each SARS-CoV-2 genomic sequence is represented as 1 circle marker. Lighter shades of each marker color indicate that the SARS-CoV-2 genomic sequence was sampled at an earlier timepoint during the same epoch compared to sequences colored in darker shades of the same color. For example, a SARS-CoV-2 genomic sequence sampled during April 2020 in Epoch 1 would be shown in lighter red, while a SARS-CoV-2 genomic sequence sampled during August 2020 in Epoch 1 would be shown in darker red.
[PNG File , 183 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Mean site-specific Shannon diversity index at each base position of the SARS-CoV-2 genome for SARS-CoV-2 genomic sequences sampled from each of the 4 epochs. Greater Shannon diversity index values at 1 base position suggest that the types of nucleotide bases are more diverse and evenly distributed between the different types compared to lower Shannon diversity values. . [PNG File , 60 KB-Multimedia Appendix 3]

References

1. COVID-19 epidemiology update: Key updates. Government of Canada. URL: https://health-infobase.canada.ca/covid-19/ [accessed 2022-07-21]
2. Ontario COVID-19 Data Tool. Public Health Ontario. URL: https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=summary [accessed 2022-08-19]
3. Population estimates, quarterly. Statistics Canada. URL: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901 [accessed 2022-08-19]
4. Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc Natl Acad Sci U S A 2021 Jul 20;118(29):e2104241118 [FREE Full text] [doi: 10.1073/pnas.2104241118] [Medline: 34292871]
5. Azgari C, Kilinc Z, Turhan B, Circi D, Adebali O. The Mutation Profile of SARS-CoV-2 Is Primarily Shaped by the Host Antiviral Defense. Viruses 2021 Mar 02;13(3):394 [FREE Full text] [doi: 10.3390/v13030394] [Medline: 33801257]
6. Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience 2021 Feb 19;24(2):102116 [FREE Full text] [doi: 10.1016/j.isci.2021.102116] [Medline: 33532709]
7. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, PCAWG Mutational Signatures Working Group, PCAWG Consortium. The repertoire of mutational signatures in human cancer. Nature 2020 Feb 05;578(7793):94-101 [FREE Full text] [doi: 10.1038/s41586-020-1943-3] [Medline: 32025018]
8. Mourier T, Sadykov M, Carr MJ, Gonzalez G, Hall WW, Pain A. Host-directed editing of the SARS-CoV-2 genome. Biochem Biophys Res Commun 2021 Jan 29;538:35-39 [FREE Full text] [doi: 10.1016/j.bbrc.2020.10.092] [Medline: 33234239]
9. Khateeb J, Li Y, Zhang H. Emerging SARS-CoV-2 variants of concern and potential intervention approaches. Crit Care 2021 Jul 12;25(1):244 [FREE Full text] [doi: 10.1186/s13054-021-03662-x] [Medline: 34253247]
10. Liu J, Xu Y, Li K, Ye Q, Zhou H, Sun H, et al. The mA methylome of SARS-CoV-2 in host cells. Cell Res 2021 Apr 28;31(4):404-414 [FREE Full text] [doi: 10.1038/s41422-020-00465-7] [Medline: 33510385]
11. Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, et al. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. Lancet Infect Dis 2020 Jul;20(7):793-802 [FREE Full text] [doi: 10.1016/S1473-3099(20)30230-9] [Medline: 32247326]
12. Yang J, Li J, Lai S, Ruktanonchai C, Xing W, Carioli A, et al. Uncovering two phases of early intercontinental COVID-19 transmission dynamics. J Travel Med 2020 Dec 23;27(8):taaa200 [FREE Full text] [doi: 10.1093/jtm/taaa200] [Medline: 33094347]
13. Han E, Tan MMJ, Turk E, Sridhar D, Leung GM, Shibuya K, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. The Lancet 2020 Nov;396(10261):1525-1534. [doi: 10.1016/s0140-6736(20)32007-9]
14. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun Biol 2021 Feb 15;4(1):228 [FREE Full text] [doi: 10.1038/s42003-021-01754-6] [Medline: 33589648]
15. Benslimane FM, Al Khatib HA, Al-Jamal O, Albatesh D, Boughattas S, Ahmed AA, et al. One Year of SARS-CoV-2: Genomic Characterization of COVID-19 Outbreak in Qatar. Front Cell Infect Microbiol 2021 Nov 17;11:768883 [FREE Full text] [doi: 10.3389/fcimb.2021.768883] [Medline: 34869069]
16. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Oxford Virus Sequencing Analysis Group (OVSG), COVID-19 Genomics UK (COG-UK) Consortium, et al. SARS-CoV-2 within-host diversity and transmission. Science 2021 Apr 16;372(6539):eabg0821 [FREE Full text] [doi: 10.1126/science.abg0821] [Medline: 33688063]
17. Elizondo V, Harkins GW, Mabvakure B, Smidt S, Zappile P, Marier C, et al. SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay. Emerg Microbes Infect 2021 Dec 15;10(1):51-65 [FREE Full text] [doi: 10.1080/22221751.2020.1863747] [Medline: 33306459]
18. Paré B, Rozendaal M, Morin S, Kaufmann L, Simpson SM, Poujol R, et al. Patient health records and whole viral genomes from an early SARS-CoV-2 outbreak in a Quebec hospital reveal features associated with favorable outcomes. PLoS One 2021 Dec 2;16(12):e0260714 [FREE Full text] [doi: 10.1371/journal.pone.0260714] [Medline: 34855869]
19. Grubaugh ND, Hodcroft EB, Fauver JR, Phelan AL, Cevik M. Public health actions to control new SARS-CoV-2 variants. Cell 2021 Mar 04;184(5):1127-1132 [FREE Full text] [doi: 10.1016/j.cell.2021.01.044] [Medline: 33581746]

20. Sjaarda CP, Guthrie JL, Mubareka S, Simpson JT, Hamelin B, Wong H, et al. Temporal Dynamics and Evolution of SARS-CoV-2 Demonstrate the Necessity of Ongoing Viral Genome Sequencing in Ontario, Canada. mSphere 2021 Jun 30;6(3):00011-21. [doi: 10.1128/msphere.00011-21]

21. Gao J, Zhang P. China's Public Health Policies in Response to COVID-19: From an "Authoritarian" Perspective. Front Public Health 2021 Dec 15;9:756677 [FREE Full text] [doi: 10.3389/fpubh.2021.756677] [Medline: 34976920]

22. David G, Rafael H, Ayelén RB, Inmaculada L, Amparo L, Marina P, et al. Perimeter confinements of basic health zones and COVID-19 incidence in Madrid, Spain. BMC Public Health 2022 Feb 03;22(1):216 [FREE Full text] [doi: 10.1186/s12889-022-12626-x] [Medline: 35109838]

23. Cyr A, Mondal P, Hansen G. An Inconsistent Canadian Provincial and Territorial Response During the Early COVID-19 Pandemic. Front Public Health 2021 Sep 27;9:708903 [FREE Full text] [doi: 10.3389/fpubh.2021.708903] [Medline: 34646800]

24. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017 Mar 30;22(13):30494 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.13.30494] [Medline: 28382917]

25. Zhang M, Li L, Luo M, Liang B. Genomic characterization and evolution of SARS-CoV-2 of a Canadian population. PLoS One 2021 Mar 4;16(3):e0247799 [FREE Full text] [doi: 10.1371/journal.pone.0247799] [Medline: 33662015]

26. Romano CM, Melo FL. Genomic surveillance of SARS-CoV-2: A race against time. Lancet Reg Health Am 2021 Sep;1:100029 [FREE Full text] [doi: 10.1016/j.lana.2021.100029] [Medline: 34386792]

27. Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, et al. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. Sci Rep 2020 Sep 24;10(1):15643 [FREE Full text] [doi: 10.1038/s41598-020-72533-2] [Medline: 32973171]

28. Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol 2003 Aug 15;77(16):8801-8811 [FREE Full text] [doi: 10.1128/jvi.77.16.8801-8811.2003] [Medline: 12885899]

29. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y, Wang H. The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. Front Immunol 2020 Oct 7;11:576622 [FREE Full text] [doi: 10.3389/fimmu.2020.576622] [Medline: 33117378]

30. Davidson AM, Wysocki J, Batlle D. Interaction of SARS-CoV-2 and Other Coronavirus With ACE (Angiotensin-Converting Enzyme)-2 as Their Main Receptor. Hypertension 2020 Nov;76(5):1339-1349. [doi: 10.1161/hypertensionaha.120.15256]

31. Traggiai E, Becker S, Subbarao K, Kolesnikova L, Uematsu Y, Gismondo MR, et al. An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. Nat Med 2004 Aug 11;10(8):871-875 [FREE Full text] [doi: 10.1038/nm1080] [Medline: 15247913]

32. Yan W, Zheng Y, Zeng X, He B, Cheng W. Structural biology of SARS-CoV-2: open the door for novel therapies. Signal Transduct Target Ther 2022 Jan 27;7(1):26 [FREE Full text] [doi: 10.1038/s41392-022-00884-5] [Medline: 35087058]

33. Walls AC, Park Y, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020 Apr 16;181(2):281-292.e6 [FREE Full text] [doi: 10.1016/j.cell.2020.02.058] [Medline: 32155444]

34. Mannar D, Saville J, Sun Z, Zhu X, Marti M, Srivastava S, et al. SARS-CoV-2 variants of concern: spike protein mutational analysis and epitope for broad neutralization. Nat Commun 2022 Aug 18;13(1):4696 [FREE Full text] [doi: 10.1038/s41467-022-32262-8] [Medline: 35982054]

35. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. The Lancet Microbe 2021 Sep;2(9):e481-e484. [doi: 10.1016/s2666-5247(21)00121-x]

36. Martin D, Weaver S, Tegally H, San E, Shank S, Wilkinson E, NGS-SA, COVID-19 Genomics UK (COG-UK), et al. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv 2021 Jul 25:2021 [FREE Full text] [doi: 10.1101/2021.02.23.21252268] [Medline: 33688681]

37. Clade and lineage nomenclature, March 2, 2021. GISAID. URL: https://www.gisaid.org/resources/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ [accessed 2022-06-01]

38. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986 Sep;3(5):418-426. [doi: 10.1093/oxfordjournals.molbev.a040410] [Medline: 3444411]

39. Maiti AK. Evolutionary shift from purifying selection towards divergent selection of SARS-CoV2 favors its invasion into multiple human organs. Virus Res 2022 May;313:198712 [FREE Full text] [doi: 10.1016/j.virusres.2022.198712] [Medline: 35176330]

40. Ghanchi NK, Nasir A, Masood KI, Abidi SH, Mahmood SF, Kanji A, et al. Higher entropy observed in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. PLoS One 2021 Aug 31;16(8):e0256451 [FREE Full text] [doi: 10.1371/journal.pone.0256451] [Medline: 34464419]

41. Nextstrain. GitHub. URL: https://github.com/nextstrain [accessed 2022-12-10]

42. Huddleston J, Hadfield J, Sibley T, Lee J, Fay K, Ilcisin M, et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw 2021 Jan;6(57):2906 [FREE Full text] [doi: 10.21105/joss.02906] [Medline: 34189396]

XSL•FO
RenderX

43. Obermeyer F, Jankowiak M, Barkas N, Schaffner S, Pyle J, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. medRxiv 2022 Feb 16:1208 [FREE Full text] [doi: 10.1101/2021.09.07.21263228] [Medline: 35194619]

44. Emam M, Oweda M, Antunes A, El-Hadidi M. Positive selection as a key player for SARS-CoV-2 pathogenicity: Insights into ORF1ab, S and E genes. Virus Res 2021 Sep;302:198472 [FREE Full text] [doi: 10.1016/j.virusres.2021.198472] [Medline: 34118359]

45. Yang H, Chen C, Wang J, Liao H, Yang C, Chen C, et al. Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. Proc Natl Acad Sci U S A 2020 Dec 01;117(48):30679-30686 [FREE Full text] [doi: 10.1073/pnas.2007840117] [Medline: 33184173]

46. Velazquez-Salinas L, Zarate S, Eberl S, Gladue DP, Novella I, Borca MV. Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. Front Microbiol 2020 Oct 23;11:550674 [FREE Full text] [doi: 10.3389/fmicb.2020.550674] [Medline: 33193132]

47. Yépez Y, Marcano-Ruiz M, Bezerra RS, Fam B, Ximenez JP, Silva Jr WA, et al. Evolutionary history of the SARS-CoV-2 Gamma variant of concern (P.1): a perfect storm. Genet. Mol. Biol 2022;45(1):e20210309. [doi: 10.1590/1678-4685-gmb-2021-0309]

48. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature 2021 Jan 09;589(7840):125-130. [doi: 10.1038/s41586-020-2739-1] [Medline: 32906143]

49. Tsai P, Wang M, Yang D, Liang K, Chou S, Chiou S, et al. Genomic variance of Open Reading Frames (ORFs) and Spike protein in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). J Chin Med Assoc 2020 Aug;83(8):725-732 [FREE Full text] [doi: 10.1097/JCMA.0000000000000387] [Medline: 32773643]

50. te Velthuis AJW, van den Worm SHE, Snijder E. The SARS-coronavirus nsp7+nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. Nucleic Acids Res 2012 Feb;40(4):1737-1747 [FREE Full text] [doi: 10.1093/nar/gkr893] [Medline: 22039154]

51. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat Commun 2019 May 28;10(1):2342 [FREE Full text] [doi: 10.1038/s41467-019-10280-3] [Medline: 31138817]

52. Rosas-Lemus M, Minasov G, Shuvalova L, Inniss N, Kiryukhina O, Wiersum G, et al. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. bioRxiv 2020 Apr 26:2020 [FREE Full text] [doi: 10.1101/2020.04.17.047498] [Medline: 32511376]

53. Raj R. Analysis of non-structural proteins, NSPs of SARS-CoV-2 as targets for computational drug designing. Biochem Biophys Rep 2021 Mar;25:100847 [FREE Full text] [doi: 10.1016/j.bbrep.2020.100847] [Medline: 33364445]

54. Thielen P, Wohl S, Mehoke T, Ramakrishnan S, Kirsche M, Falade-Nwulia O, et al. Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area. JCI Insight 2021 Mar 22;6(6):e144350 [FREE Full text] [doi: 10.1172/jci.insight.144350] [Medline: 33749660]

55. Dutta NK, Mazumdar K, Gordy JT. The Nucleocapsid Protein of SARS–CoV-2: a Target for Vaccine Development. J Virol 2020 Jun 16;94(13):e00647-20. [doi: 10.1128/jvi.00647-20]

56. Mendez AS, Ly M, González-Sánchez AM, Hartenian E, Ingolia NT, Cate JH, et al. The N-terminal domain of SARS-CoV-2 nsp1 plays key roles in suppression of cellular gene expression and preservation of viral gene expression. Cell Rep 2021 Oct 19;37(3):109841 [FREE Full text] [doi: 10.1016/j.celrep.2021.109841] [Medline: 34624207]

57. Morales A, Rice A, Ho A, Mordstein C, Mühlhausen S, Watson S, et al. Causes and Consequences of Purifying Selection on SARS-CoV-2. Genome Biol Evol 2021 Oct 01;13(10):196 [FREE Full text] [doi: 10.1093/gbe/evab196] [Medline: 34427640]

58. Neisen M, Anand P, Silvert E, Suratekar R, Pawlowski C, Ghosh P, et al. COVID-19 vaccines dampen genomic diversity of SARS-CoV-2: Unvaccinated patients exhibit more antigenic mutational variance. MedRxiv. URL: https://www.medrxiv.org/content/10.1101/2021.07.01.21259833v1 [accessed 2022-12-09]

59. Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc Natl Acad Sci U S A 2020 Sep 22;117(38):23652-23662 [FREE Full text] [doi: 10.1073/pnas.2008281117] [Medline: 32868447]

60. Bano I, Sharif M, Alam S. Genetic drift in the genome of SARS COV-2 and its global health concern. J Med Virol 2022 Jan 23;94(1):88-98 [FREE Full text] [doi: 10.1002/jmv.27337] [Medline: 34524697]

61. Jungreis I, Sealfon R, Kellis M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. Nat Commun 2021 May 11;12(1):2642 [FREE Full text] [doi: 10.1038/s41467-021-22905-7] [Medline: 33976134]

62. Detsky AS, Bogoch II. COVID-19 in Canada: Experience and Response to Waves 2 and 3. JAMA 2021 Sep 28;326(12):1145-1146. [doi: 10.1001/jama.2021.14797] [Medline: 34424275]

63. Fisman DN, Amoako A, Tuite AR. Impact of population mixing between vaccinated and unvaccinated subpopulations on infectious disease dynamics: implications for SARS-CoV-2 transmission. CMAJ 2022 Apr 25;194(16):E573-E580 [FREE Full text] [doi: 10.1503/cmaj.212105] [Medline: 35470204]

64. Ontario Moving to Step Three of Roadmap to Reopen on July 16. Government of Ontario. 2021. URL: https://news.ontario.ca/en/release/1000501/ontario-moving-to-step-three-of-roadmap-to-reopen-on-july-16 [accessed 2022-08-19]

XSL•FO
RenderX

65. Ulloa AC, Buchan SA, Daneman N, Brown KA. Estimates of SARS-CoV-2 Omicron Variant Severity in Ontario, Canada. JAMA 2022 Apr 05;327(13):1286-1288 [FREE Full text] [doi: 10.1001/jama.2022.2274] [Medline: 35175280]

66. Ringlander J, Fingal J, Kann H, Prakash K, Rydell G, Andersson M, et al. Impact of ADAR-induced editing of minor viral RNA populations on replication and transmission of SARS-CoV-2. Proc Natl Acad Sci U S A 2022 Feb 08;119(6):e2112663119 [FREE Full text] [doi: 10.1073/pnas.2112663119] [Medline: 35064076]

67. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J Med Virol 2020 Jun 28;92(6):584-588 [FREE Full text] [doi: 10.1002/jmv.25719] [Medline: 32083328]

68. Reshamwala SMS, Likhite V, Degani MS, Deb SS, Noronha SB. Mutations in SARS-CoV-2 nsp7 and nsp8 proteins and their predicted impact on replication/transcription complex structure. J Med Virol 2021 Jul 14;93(7):4616-4619 [FREE Full text] [doi: 10.1002/jmv.26791] [Medline: 33433004]

69. Jung C, Kmiec D, Koepke L, Zech F, Jacob T, Sparrer KMJ, et al. Omicron: What Makes the Latest SARS-CoV-2 Variant of Concern So Concerning? J Virol 2022 Mar 23;96(6):e0207721 [FREE Full text] [doi: 10.1128/jvi.02077-21] [Medline: 35225672]

70. Wei C, Shan K, Wang W, Zhang S, Huan Q, Qian W. Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genomics 2021 Dec;48(12):1111-1121 [FREE Full text] [doi: 10.1016/j.jgg.2021.12.003] [Medline: 34954396]

71. Kumar A, Asghar A, Dwivedi P, Kumar G, Narayan RK, Jha RK, et al. A Bioinformatics Tool for Predicting Future COVID-19 Waves Based on a Retrospective Analysis of the Second Wave in India: Model Development Study. JMIR Bioinform Biotech 2022 Sep 22;3(1):e36860 [FREE Full text] [doi: 10.2196/36860] [Medline: 36193192]

72. Lin D, Liu L, Zhang M, Hu Y, Yang Q, Guo J, et al. Co-infections of SARS-CoV-2 with multiple common respiratory pathogens in infected patients. Sci China Life Sci 2020 Apr 05;63(4):606-609 [FREE Full text] [doi: 10.1007/s11427-020-1668-5] [Medline: 32170625]

73. Tillett RL, Sevinsky JR, Hartley PD, Kerwin H, Crawford N, Gorzalski A, et al. Genomic evidence for reinfection with SARS-CoV-2: a case study. The Lancet Infectious Diseases 2021 Jan;21(1):52-58. [doi: 10.1016/s1473-3099(20)30764-7]

74. Wang X, Hu M, Jin Y, Wang B, Zhao Y, Liang L, et al. Global Mutational Sweep of SARS-CoV-2: From Chaos to Order. Front Microbiol 2022 Feb 8;13:820919 [FREE Full text] [doi: 10.3389/fmicb.2022.820919] [Medline: 35211106]

75. Mastriani E, Rakov AV, Liu S. Isolating SARS-CoV-2 Strains From Countries in the Same Meridian: Genome Evolutionary Analysis. JMIR Bioinform Biotech 2021 Jan 22;2(1):e25995 [FREE Full text] [doi: 10.2196/25995] [Medline: 33497425]

76. Kozlakidis Z. Evidence for Recombination as an Evolutionary Mechanism in Coronaviruses: Is SARS-CoV-2 an Exception? Front Public Health 2022 Mar 17;10:859900 [FREE Full text] [doi: 10.3389/fpubh.2022.859900] [Medline: 35372203]

77. Liu X, Huang J, Li C, Zhao Y, Wang D, Huang Z, et al. The role of seasonality in the spread of COVID-19 pandemic. Environ Res 2021 Apr;195:110874 [FREE Full text] [doi: 10.1016/j.envres.2021.110874] [Medline: 33610582]

78. Maharaj AS, Parker J, Hopkins JP, Gournis E, Bogoch II, Rader B, et al. The effect of seasonal respiratory virus transmission on syndromic surveillance for COVID-19 in Ontario, Canada. The Lancet Infectious Diseases 2021 May;21(5):593-594. [doi: 10.1016/s1473-3099(21)00151-1]

79. Gupta E, Mishra RK, Kumar Niraj RR. Identification of Potential Vaccine Candidates Against SARS-CoV-2 to Fight COVID-19: Reverse Vaccinology Approach. JMIR Bioinform Biotech 2022 Apr 26;3(1):e32401 [FREE Full text] [doi: 10.2196/32401] [Medline: 35506029]

80. Math RK, Mudennavar N, Javaregowda PK, Savanur A. In Silico Comparative Analysis of the Functional, Structural, and Evolutionary Properties of SARS-CoV-2 Variant Spike Proteins. JMIR Bioinform Biotech 2022 May 30;3(1):e37391 [FREE Full text] [doi: 10.2196/37391] [Medline: 35669291]

Abbreviations

ADAR: adenosine deaminase acting on RNA
APOBEC: apolipoprotein B mRNA editing enzyme catalytic polypeptide-like
GISAID: Global Initiative on Sharing All Influenza Data
m6A: N6-methyladenosine
NSP: nonstructural protein
ROS: reactive oxygen species
SBS: single base substitution