EVOLUTION OF SARS-CoV-2 VARIANTS IN GEOGRAPHIC LOCATIONS WITH VARYING CASE INCIDENCE

.

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI'I AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

MICROBIOLOGY

AUGUST 2022

By

Claire Jisu Fraser

Thesis Committee:
Marguerite Butler, Chairperson
Sladjana Prišić
Vivek Nerurkar

Keywords: selection, phylogenetics, evolution, diversity

TABLE OF CONTENTS

Acknowledgements…… 3

Abstract…….. 4

List of Figures and Tables…….. 5

Introduction…… 6

Materials and Methods…..……10

Results……16

Discussion……..27

References……..35

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Committee Chair, Dr. Marguerite Butler for assisting me in developing this project through providing the essential framework for this study. She has spent countless hours teaching me proper methods for conducting lab work, as well as how to run various analyses. I am where I am today because of skills I have learned under Dr. Butler's instruction and am eternally grateful for everything she has done for me. I would also like to thank my other committee members, Dr. Sladjana Prisic and Dr. Vivek Nerurkar, for working with me and providing feedback for my thesis.

This endeavor would not have been possible without the generous funding from the University of Hawai'i at Manoa Graduate Student Organization (Award # 20-12-03) and the Hawai'i State Department of Health, State Laboratories Division, who also provided the samples to sequence. The technical support and advanced computing resources from University of Hawaii Information Technology Services – Cyberinfrastructure, funded in part by the National Science Foundation MRI award # 1920304, are also gratefully acknowledged.

A huge thank you to my fellow grad student, Ethan Hill, for collaborating with me through providing assistance running various analyses, interpreting results, and being an essential support system for me. Special thanks to the Hawai'i Pandemic Applied Modeling Work Group (HiPAM) for assistance in developing interesting questions to address in my study. Many thanks to the sequencing volunteer team at UH who helped with sequencing efforts and provided feedback during the development of my defense. Finally, I'd like to express my extreme gratitude for my family and friends for their encouragement and support through my studies.

ABSTRACT

The evolutionary dynamics of SARS-CoV-2 over the course of the pandemic are of great concern. Prior studies have identified accelerated protein evolution early in the pandemic, followed by a period of increased selective constraints. While there are expected changes in rates of mutation influenced by viral spread, the translation to selection is not well studied. In addition, new variants are emerging as SARS-CoV-2 continues to evolve and some are spreading more rapidly due to mutations that provide a viral advantage. A comparative analysis between Los Angeles County and Hawai'i, localities with rigorous surveillance sequencing programs and varying case incidence, was conducted to test whether there are evolutionary differences of B.1.1.7, the first identified variant of concern, and B.1.243, a rapidly spreading variant in Hawai'i, across localities. Of these four locality-variant scenarios, only B.1.243 in LA County occurred at low case incidence and the remaining three scenarios were at high case incidence. With the rapid spread of virus, I generally found elevated rates of evolution - both synonymous and nonsynonymous substitution. The two variants had very different histories, with B.1.243 evolving locally in Hawaii before being introduced into LA County in a single migration event, whereas B.1.1.7 travelled between localities repeatedly. Yet, there were great differences in diversity of genome sequences overall and elevation of number of sites under positive and negative selection in Los Angeles County relative to the same variant in Hawaii. Within each local outbreak of each variant, I found little evidence for an early phase of protein adaptation followed by an increase in constraints as found in a previous study across the US. Instead, I find that in general more sites are under negative selection than positive selection at this point of the pandemic. Overall, these results show support for variability in evolutionary dynamics across localities and suggest differences in selective pressures across different populations.

LIST OF TABLES AND FIGURES

Table 1. Genomes from GISAID included in this study……...12

Table 2. Substitution Rates per Gene…….……...17

Figure 1. Case incidence and Effective Reproduction number through time……19

Figure 2. Kernel Density Estimates through time……..20

Table 3. Genomes in each bin based on KDP plots……...21

Figure 3. IQTREE Phylogenies.....22

Table 4. Pairwise Nucleotide Distance for B.1.1.7……....23

Table 5. Pairwise Nucleotide Distance for B.1.243……...23

Figure 4. Total sites under selection between phases……24

Table 6. Common sites under selection between variants in ORF1a and S……..25

Figure 5. Heatmap of sites under selection……26

Table 7. Sites under selection in B.1.1.7 S gene……27

Table 8. Sites under selection in B.1.243 S gene……...27

INTRODUCTION

December 2019 marked the start of the coronavirus disease 2019 (COVID-19) pandemic with unprecedented spread of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), however the experience within different localities has been highly variable. In theory, differences in magnitude of viral replication provide the potential for differences in viral evolutionary dynamics, such as mutation rates, selection, and genome diversity among circulating variants, but it is unclear whether differences in evolution are actually observed. Thanks to collaborative efforts of public health and research labs across the globe, over 12M genomes have been sequenced and shared on the Global Initiative on Sharing Avian Influenza Data (GISAID) online repository by 15 April 2022 (Khare et al., 2021), providing an unprecedented supply of data to identify differences in evolutionary dynamics and epidemiological parameters across localities in real time (Bi et al., 2020; Nie et al, 2020; Miller et al., 2021).

Theoretically, viral evolution can be affected by increased viral spread, as more transmission events provide a greater opportunity for mutation and ultimately increase the level of genetic diversity (Scholle et al., 2013; Castellano et al., 2019). Interestingly, for some viruses, when variation in mutation rates have been observed, they have tended to be beneficial mutations (Jiang et al., 2010), which presents an interesting possibility that SARS-CoV-2 may be gaining in fitness as the pandemic has accelerated. Indeed, accelerated protein evolution was detected within the first few months post-origination of SARS-CoV-2, followed by an increase in selective constraints through time (Rochman et al., 2021; Lin et al., 2021; Chaw et al., 2020). Specifically, an elevated number of sites under positive selection was interpreted as accelerated protein evolution, and an increase in sites under negative selection was interpreted as an increase

in selective constraints (Lin et al., 2021; Lynch and Conery, 2000). Various methods have been developed to test for signals of selection by comparing the rates of nonsynonymous (dN) and synonymous (dS) substitutions, as varying dN rates are signs of positive selection when elevated, and negative selection when low (Li et al., 1985; Yang and Nielsen, 2000; Goldman and Yang 1994; Felsenstein, 2001).

As the virus continues to spread, certain localities across the globe experience quick rises in case counts and upon investigation, these surges are typically signals of emerging variants with a fitness advantage (Davies et al., 2020; Lindstrøm et al., 2021; Ramanathan et al., 2021). The PANGOLIN lineage B.1.1.7 (O'Toole et al., 2021; Alpha) variant was one of the first identified variants of concern, originating in Europe, and is 40-90% more transmissible than the ancestral lineages (Davies et al., 2020; Lindstrøm et al., 2021; Ramanathan et al., 2021; Scripps Research, 2021). Other variants, which are not classified as variants of concern globally, may yet dominate locally, spreading rapidly within a small population. In late 2021, the B.1.243 variant represented 40% of all cases in the state of Hawai'i, while representing a smaller proportion of cases in other US states (Maison et al., 2021). Therefore, I can test whether individual variants will show a common evolutionary dynamic with a pattern of accelerated protein evolution in the early exponential growth phase of viral proliferation, followed by an increase of selective constraints in the waning phase.

In addition to the possibility of elevated rates of evolution across the genome, variation through rates of mutation and selection across gene regions is well known in SARS-CoV-2 (Kumar Das et al., 2021; Rahman et al., 2021). Variability in the evolutionary dynamics across the genome highlights regions under selective pressures, typically dependent on the function of the coded protein (Liu et al., 2008). High mutation rates have been observed in ORF1ab, S,

ORF3a, and N gene regions, with low rates of mutation occurring in E, ORF6, ORF7a, ORF7b, and ORF10 (Kumar Das et al., 2021; Rahman et al., 2021). ORF1ab makes up ~70% of the viral genome and is involved in viral transcription, replication, and immune evasion; ORF1a is the larger of the two glycoproteins coded by the ORF1ab gene and includes non-structural protein (nsp) 1 to 11 (Brant et al., 2021; Wu et al., 2020). The S gene region codes for the spike protein, which is a critical functional protein for the virus as it is the antigenic region that binds to the host cell receptor and is involved in host cell entry (Brant et al., 2021; Wu et al., 2020). The spike protein is subsequently cleaved into 2 subunits post translation and prior studies have detected a correlation between clade growth and mutation accumulation in the S1 subunit (Kistler et al., 2022) and adaptive evolution concentrated in the S1 subunit in other coronaviruses (Kistler and Bedford, 2021). Therefore, variants that are successful are expected to have elevated levels of positive selection in the S1 region prior to or shortly post-origination, due to the accumulation of advantageous mutations that provide a viral advantage.

In contrast, ORF7a has a low mutation rate and is therefore a good region of comparison. ORF7a plays a role in interaction with the host immune response (Yashvardhini et al., 2021) and may be highly conserved. If so, it may display elevated rates of negative selection (Monteiro et al., 2010; Choudhuri, 2014). In addition to variability in the rates of mutation across the genome, selection is also variable across the genome and positive selection has been detected within the ORF1ab, S, ORF3a, E, and N gene regions, while negative selection was found within ORF1ab and S (Lin et al., 2021; Emam et al., 2021; Cao et al., 2021, Liu et al., 2020; Kochan et al., 2021; Miller et al., 2020).

Human populations are connected by travel, whether they will still experience the same evolutionary dynamics is not clear. The genetic diversity of SARS-CoV-2 genomes found within

a population is determined by both de-novo evolution that arises from mutational accumulation associated with community transmission but can also be exchanged via importation from other communities (Otto et al., 2021). It is therefore interesting to examine communities connected by human travel such as Los Angeles (LA) County and the state of Hawai'i, but experienced different pandemic histories. In addition, sequencing efforts in both localities were similar, with ~5-10% of all cases sequenced (CDC, 2022; California Department of Public Health, 2022b). Despite strict interventions set in place in both localities following surges in case counts, early in the pandemic Hawaii had very low case counts while that of Los Angeles soared (Governor of the State of Hawai'i, 2020; State of Hawai'i, Department of Health, 2020; California Department of Public Health, 2022a). LA County was the county with the highest case incidence in the United States (US) as of 15 April 2022 (Dong et al., 2020) and comprises 88 cities with a total population of 10M. Alternatively, the state of Hawai'i (HI) is among the five states in the US with the lowest case incidence as of 15 April 2022 and comprises 5 counties with a total population of 1.45M (U.S. Census Bureau, 2022). Many evolutionary studies of SARS-CoV-2 have been conducted on a large-scale, however variation in the evolutionary dynamics of SARS-CoV-2 within a local population is not as clear, particularly between connected populations.

In this study, I will use genomic SARS-CoV-2 data across LA County and Hawai'i to test for the impact that varying levels of viral spread has on mutation rates and identify whether these differences also have an impact on selection across the genome. I will also explore whether the B.1.1.7 and B.1.243 variants follow an expected pattern of evolution and if they are under different selective constraints within different host localities.

MATERIALS AND METHODS

Models for Selection at Nucleotide Sites

Inferring selection at sites along the genome relies on testing for mutation rates. In particular, whether or not synonymous or non-synonymous mutations occur at exceptionally higher or lower rates as compared to the random fixation of neutral mutations (Kimura, 1991). Regions of the genome under negative and positive selection can be identified by detection of elevated synonymous and nonsynonymous substitution rates (Lynch and Conery, 2000). Selection analyses used in this study tested for selection across the entire phylogeny, as well as within smaller regions of the tree.

Experimental Design

Sequences from the state of Hawai'i were used as a representative of a geographic location with relatively low case incidence until mid-December 2021. The entire state was considered a single system for the purpose of this analysis. For comparison, LA County is the US County with the highest case incidence from June 2020 to September 2021 (Dong et al., 2020). The B.1.1.7 variant was selected as it underwent rapid spread in both localities, with a greater total proportion of SARS-CoV-2 sequences in California than in Hawai'i. The B.1.243 variant rapidly spread in Hawai'i representing a greater proportion of all sequences than in other US states and was selected as an additional variant to analyze in this study. The two variants had sufficient sample size and similar timing across the locales for this study: B.1.1.7 and B.1.243, with similar timeframes across the two localities (Table 1). Selection analyses were conducted on gene regions with high rates of mutation (ORF1a and S) and low rates of mutation (ORF7a; Kumar Das et al., 2021; Rahman et al., 2021).

Next-generation sequencing and GISAID

Samples of SARS-CoV-2 collected from Hawai'i were obtained from the Hawai'i State Department of Health and provided to us in the form of purified RNA. Two hundred eighty-five samples were processed using the NEBNext ARTIC SARS-CoV-2 Companion Kit (E7660S, NEB, USA) compatible with sequencing using the Oxford Nanopore Technologies (ONT) MinION device (R9.4.1, ONT, UK) following recommended protocols developed by the ARTIC Network (Quick, 2020). VarSkip primers (NEB) were used for amplification. Subsequent processing of the samples was completed using the nCoV-2019 novel coronavirus bioinformatics protocol (ARTIC, https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). Medaka was used to generate consensus sequences and variant call rather than nanopolish. Tablet v1.21.02.08 (Milne, 2012) was used to visualize coverage across the genome. Final lineage assignment was made using the Pangolin COVID-19 Lineage Assigner version 3.1.16 (O'Toole et al., 2021). Thirty-two genomes were identified as B.1.1.7 and included in subsequent analyses. 226 samples had >95% genome coverage and average depth ranging from 220-483x. Amplicon dropouts were observed in regions 3471-3778 and 8284-8709 bp.

To supplement the Hawai'i dataset, sequences were also downloaded from online databases. Two thousand eighty SARS-CoV-2 genomes were collected from GISAID for a low incidence locality (the State of Hawai'i) and a high incidence locality (Los Angeles County) on 4 May 2022 (Khare et al., 2021; Table 1) to explore differences in viral evolutionary dynamics with respect to incidence.

Table 1. Number of genomes (*n)* included in this study. These are a subset of the data available on GISAID that passed Nextstrain filters (% of All genomes).
*An additional 32 genomes sequenced in house are included in the Hawai'i B.1.1.7 dataset count below.

Location/Variant	***n***	**% of All Genomes**	**Collection Dates**
HI / B.1.1.7	687*	98.6%	1/21/21 - 7/14/21
HI / B.1.243	737	100%	7/8/20 - 2/8/21
LA County / B.1.1.7	801	94.7%	1/4/21 - 8/12/21
LA County / B.1.243	287	98.3%	6/20/20 - 4/6/21

Case Counts through Time and Division of Sequences

Total case counts were downloaded on 4 May 2022 for the State of Hawai'i (Centers for Disease Control and Prevention, 2022) and for LA County (California Department of Public Health, 2022a). Variant counts were determined by the total number of sequences available on GISAID (Khare et al., 2021) and 7-day averages were calculated using the R package zoo v1.8-10 (Zeileis and Grothendieck, 2005). For each locality, case count was plotted through time using kernel density plots to determine the point in time where case counts peaked within each locality using the R package ggplot2 v3.3.6 (Wickham, 2016) in the R statistical computing environment v4.0.4 (R Core Team, 2021). This date was used to bin genomes into exponential and waning viral growth phases at each locality for each strain.

Nextstrain Pipeline: Filtering and Phylogenetic Reconstruction

The NextStrain pipeline was used to filter the data for quality, align sequences, build the datasets and generate starting trees for downstream Bayesian inference phylogenetic reconstruction (Hadfield et al., 2018). I used quality filters on the genomes, excluding those with <27,000 bp and no collection date (Table 1). The sequences are aligned to the reference sequence (hCoV-19/Wuhan/Hu-1/2019) using MAFFT v7.475, masking regions that are prone to sequencing errors.

Maximum-likelihood (ML) phylogenies were reconstructed using IQTREE v2.1.4-beta (Nguyen et al., 2015) allowing the software to select the best fitting evolutionary model from the data, timescale inferred by TreeTime 0.8.0 (Sagulenko et al., 2018) with 1000 bootstrap replicates. ML phylogenies were reconstructed for each variant at each locality for starting trees for Bayesian Inference (see below), as well as for each variant within the localities pooled to visualize migration of the virus between localities.

Estimation of Epidemiological Parameters

For each variant at each locality, phylogenies were estimated by Bayesian inference using the Markov chain Monte Carlo (MCMC) method implemented by BEAST v1.10.4 (Suchard et al., 2018) using the starting ML trees generated above on the Mana University of Hawai'i High Performance Computing Cluster. Bayesian inference was used to estimate effective reproduction number thorough time (Rt). Epidemiological parameters also have expected changes with increased viral spread, where transmission from one infected individual to a greater number of healthy individuals will result in an increase in case counts and can be quantified by the effective reproduction number (Irons and Raftery, 2021; Achaiah et al., 2020; Billah et al., 2020; Rt). The

MCMC analysis was run for 300,000,000 generations in duplicate under a Hasegawa-Kishino-Yano (HKY) model, with evolutionary rate mean of 0.00135 s/s/y, under a strict clock assumption and Birth-Death Skyline Serial model (Stadler et al., 2012) with sampling every 20,000 generations. All tree element operators were set to zero to fix the starting tree, Re was estimated under a lognormal prior with a mean of 0.0 and a variance of 1.0. Convergence was determined by effective sampling size values (ESS >200) after 15% burn-in visualized through Tracer v1.7.1 (Rambaut et al., 2018). The Birth-Death Skyline Serial model is used to estimate the Rt, allowing for serially sampled data and rate changes over time. Substitution rates were also estimated assuming a log normal distribution with a range of 0.0007 to 0.002 s/s/y (Miller et al., 2020; Chaw et al., 2020). Rt plots were generated using the bdskytools package in R (Plessis, 2016).

Pairwise Nucleotide Distance Calculation

To determine the proportion of nucleotide differences within and between each group of sequences, the genetic distance was calculated using MEGA v11 (Tamura et al., 2021) estimated by 1000 bootstrap replicates under the p-distance Model/Method, assuming uniform substitution rates among sites.

Selection Analyses

Bayesian Inference (BI) and Maximum Likelihood (ML) methods were used to infer sites with detectable selection. Detected sites were considered significant under ML methods if posterior values >0.9 and under Bayesian methods if the p-value <0.1. Signals of selection were tested for in ORF1ab, S, and ORF7a gene regions using Fast, Unconstrained Bayesian AppRoximation (FUBAR; Murrell et al., 2013), Fixed Effects Likelihood (FEL; Kosakovsky

Pond and Frost, 2005), and the Mixed Effects Model of Evolution (MEME; Murrell et a., 2012) methods using the HyPhy Interactive command-line v2.2.4 on the Mana University of Hawai'i high performance computing cluster and visualized through the HyPhy Vision web server (Weaver et al., 2018). FUBAR and FEL methods were used to detect both positive and negative pervasive selection. MEME, in contrast, tests specifically for pervasive and episodic positive selection. Sequence data were trimmed to single-gene coding regions excluding the stop codon and used to conduct each selection analysis along their respective phylogeny. Sites detected with significance under at least one method were considered in this study. The ggplot2 and ggpubr packages in R were used to generate lollipop plots to compare the number of sites under positive or negative selection in either the exponential or waning phases (Wickham, 2016; Kassambara, 2020). The tidyverse package was used in addition to generate heatmaps for comparison of the position of sites under selection across phases, localities, and variants (Wickham et al., 2019).

Posterior estimates of synonymous and nonsynonymous substitution rates for each phase of a variant within a locality were calculated by taking the average of the mean posterior substitution rates (either synonymous or nonsynonymous) estimated by the FUBAR method, across all sites within each gene region.

RESULTS

Variant Prevalence and Division into 2 Phases

The peaks in case incidence for both Hawai'i and LA County coincide with peaks in B.1.243 case counts (Figure 1). In both localities, B.1.1.7 replaced B.1.243 as it waned. B.1.1.7 dominated, while B.1.243 composed of a much smaller proportion of cases in LA County. On the other hand, B.1.1.7 and B.1.243 were both dominating variants across the state of Hawai'i. Case counts for B.1.1.7 in LA County were higher than case counts for B.1.243, however the total number of case counts including other circulating variants at the time was higher while B.1.243 was prevalent, compared to B.1.1.7 (Figure 1). Date ranges and genome counts for each bin are listed in Table 3 based on case count densities (Figure 2).

Table 2. Posterior Estimates of Synonymous (S) and Nonsynonymous (NS) rate means calculated across rates estimated at each site by the FUBAR selection analyses.
N+ = number of sites under positive selection, N- = number of sites under negative selection.
The Sc column indicates whether variants within each locality are considered relatively "high" or "low" case incidence based on sequence counts as a proxy for variant outbreak.

Gene	Variant	Length (bp)	Locality	Phase	S	NS	NS/S	N+	N-	Sc
ORF1a	B.1.1.7	13,218	LA County	Exponential	1.85	1.34	0.72	13	18	HIGH
				Waning	1.72	0.97	0.56	11	70	
			Hawai'i	Exponential	1.99	1.32	0.66	3	12	HIGH
				Waning	1.77	1.21	0.68	6	14	
	B.1.243		LA County	Exponential	1.8	1.09	0.61	10	35	LOW
				Waning	1.86	1.13	0.61	6	27	
			Hawai'i	Exponential	2.18	1.47	0.67	2	11	HIGH
				Waning	2.71	2.07	0.76	1	2	
S	B.1.1.7	3,822	LA County	Exponential	2.49	1.81	0.73	2	2	HIGH
				Waning	1.82	1.20	0.66	16	20	
			Hawai'i	Exponential	2.62	1.88	0.72	2	3	HIGH
				Waning	1.97	1.49	0.76	6	8	
	B.1.243		LA County	Exponential	2.73	1.51	0.55	6	19	LOW
				Waning	2.78	1.51	0.54	4	14	
			Hawai'i	Exponential	2.88	2.06	0.72	0	3	HIGH
				Waning	4.08	2.79	0.68	0	6	
ORF7a	B.1.1.7	366	LA County	Exponential	3.24	2.05	0.63	0	0	HIGH
				Waning	2.35	1.64	0.70	0	1	
			Hawai'i	Exponential	4.25	2.71	0.64	0	3	HIGH
				Waning	2.91	2.13	0.73	1	0	
	B.1.243		LA County	Exponential	2.78	2.05	0.74	1	0	LOW
				Waning	4.23	3.74	0.88	0	0	
			Hawai'i	Exponential	3.9	2.39	0.61	0	1	HIGH
				Waning	4.35	3.29	0.76	0	0	

Higher Substitution Rates in Hawai'i

For both variants, the rates of synonymous and nonsynonymous substitutions were higher in Hawai'i than LA County (Table 2). This was especially apparent for B.1.243.

Aggressive Government Response Following Surge in Cases in Both Localities

Case counts in Hawai'i began to increase in August 2020, with a large number of B.1.243 cases circulating within the state (Figure 1A, 1B). In response to the surge in cases, a stay-at-home order and limitations on gatherings to groups of five or fewer was implemented across the most populated island of Oahu on August 27, 2020 (Governor of the State of Hawai'i, 2020). As case counts remained at a 7-day-average of ~100, the Safe Travels Program was implemented in Hawai'i and extended beyond the final collection date included in this study, limiting travel to the islands on October 15, 2020 (State of Hawai'i, Department of Health, 2020). B.1.243 continued to spread rapidly for the next few months and began to decline with the introduction of the B.1.1.7 variant on January 21, 2021.

In November 2020, there was a large outbreak in LA County followed by a rapid implementation of safety measures including a stay-at-home order, mask mandate, and gatherings limited to household members (Figure 1A, 1B; County of Los Angeles Public Health, 2022). Nevertheless, LA County was hit with an increase in the B.1.243 variant. Following the decline of B.1.243 was an increase in the B.1.1.7 variant, which was coincident with a drop in overall case counts.

Initial Rt Value Estimates are Similar Between Variants

Upon introduction, the Rt values for both variants across localities were estimated to be ~1.25. Following estimates for B.1.1.7 dropped to ~0.75 and were unchanging, however in Hawai'i there was an estimated increase to ~1.5 around December 2020 (Figure 1). Rt values for B.1.243 in Hawai'i remained at ~1.25 over the entire time the variant circulated within the state while estimates for LA County declined to ~0.75 in April 2021, followed by another drop to ~0.6 by July 2021 (Figure 1).

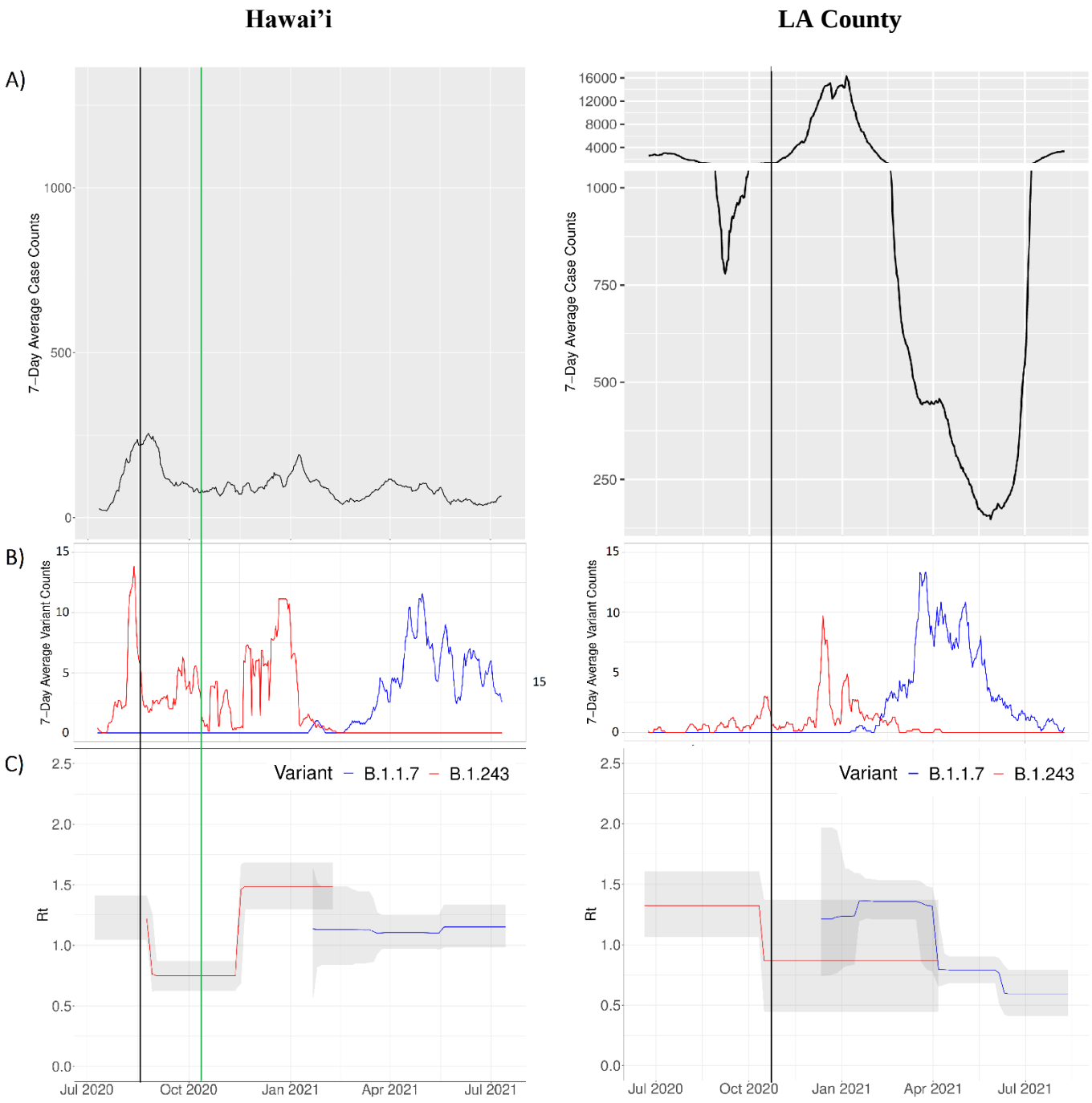

Figure 1.

A) 7-day average case incidence within Hawai'i (left) and LA County (right).

B) 7-day average variant counts within Hawai'i and LA County.

C) Effective reproduction number (Rt) for B.1.243 (red) and B.1.1.7 (blue) in each locality. Black vertical line indicates timing of safety measure implementations, green vertical line indicates start of Safe Travels program in Hawai'i which continued beyond July 2021.

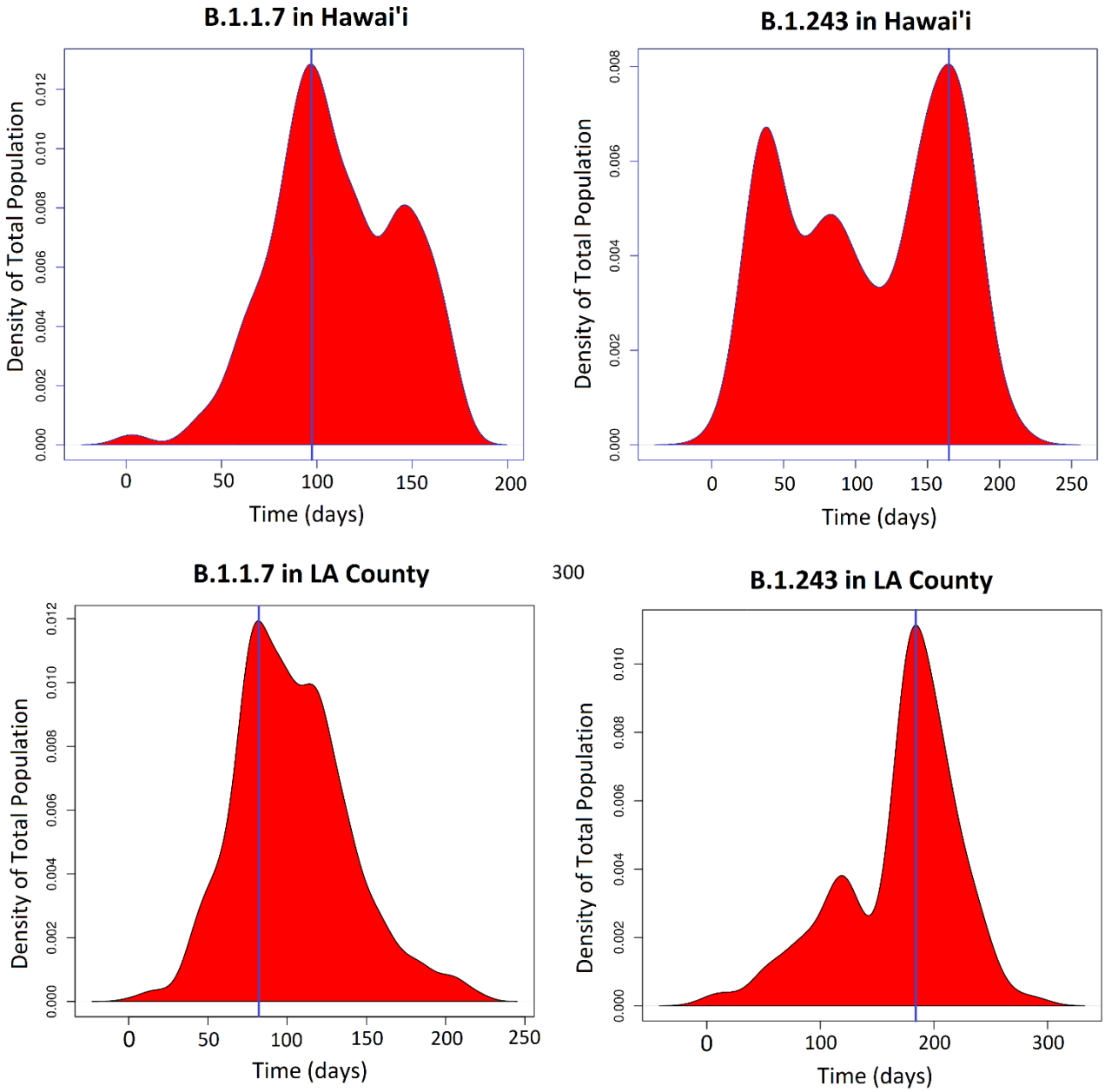

Figure 2. Kernel Density Plots for B.1.1.7 and B.1.243 in LA County and Hawai'i case counts through time. Blue lines indicate break in time to split genomes between exponential and waning phases.

Table 3. Number of genomes in each bin based on case counts through time.

Location/Variant	Phase	# of Genomes	Collection Dates
HI / B.1.243	Exponential	590	7/8/20 - 12/18/20
HI / B.1.243	Waning	147	12/19/20 - 2/8/21
LA County / B.1.243	Exponential	112	6/20/20 - 12/20/20
LA County / B.1.243	Waning	175	12/21/20 - 4/6/21
HI / B.1.1.7	Exponential	245	1/21/21 - 4/27/21
HI / B.1.1.7	Waning	442	4/28/21 - 7/14/21
LA County / B.1.1.7	Exponential	249	1/4/21 - 3/26/21
LA County / B.1.1.7	Waning	552	3/27/21 - 8/12/21

Higher Pairwise Nucleotide Distance in LA County

The within-population pairwise nucleotide distance shows greater diversity within the LA County community than in Hawai'i for both variants (Table 4). This difference is greater for B.1.243, as well as a very large between-population distance with ~5x greater difference between localities (Table 4). B.1.1.7, in contrast, is much more similar between LA County and Hawai'i populations.

Phylogenetic Reconstruction and Highly Supported Estimation of Epidemiological Parameters

Bayesian inference phylogenies used for mutational analyses and estimation of Rt (phylogenies not shown) resulted in an estimated sample size (ESS) values >2000 for each variant within each locality, providing high support for the parameter values estimated based on the phylogeny.

Maximum-likelihood trees for B.1.1.7 shows a mixture of LA County and Hawai'i genomes interspersed throughout the tree, while phylogenetic reconstruction of B.1.234 resulted in 2 distinct clades with only one introductory event into LA County from Hawai'i (Figure 3).

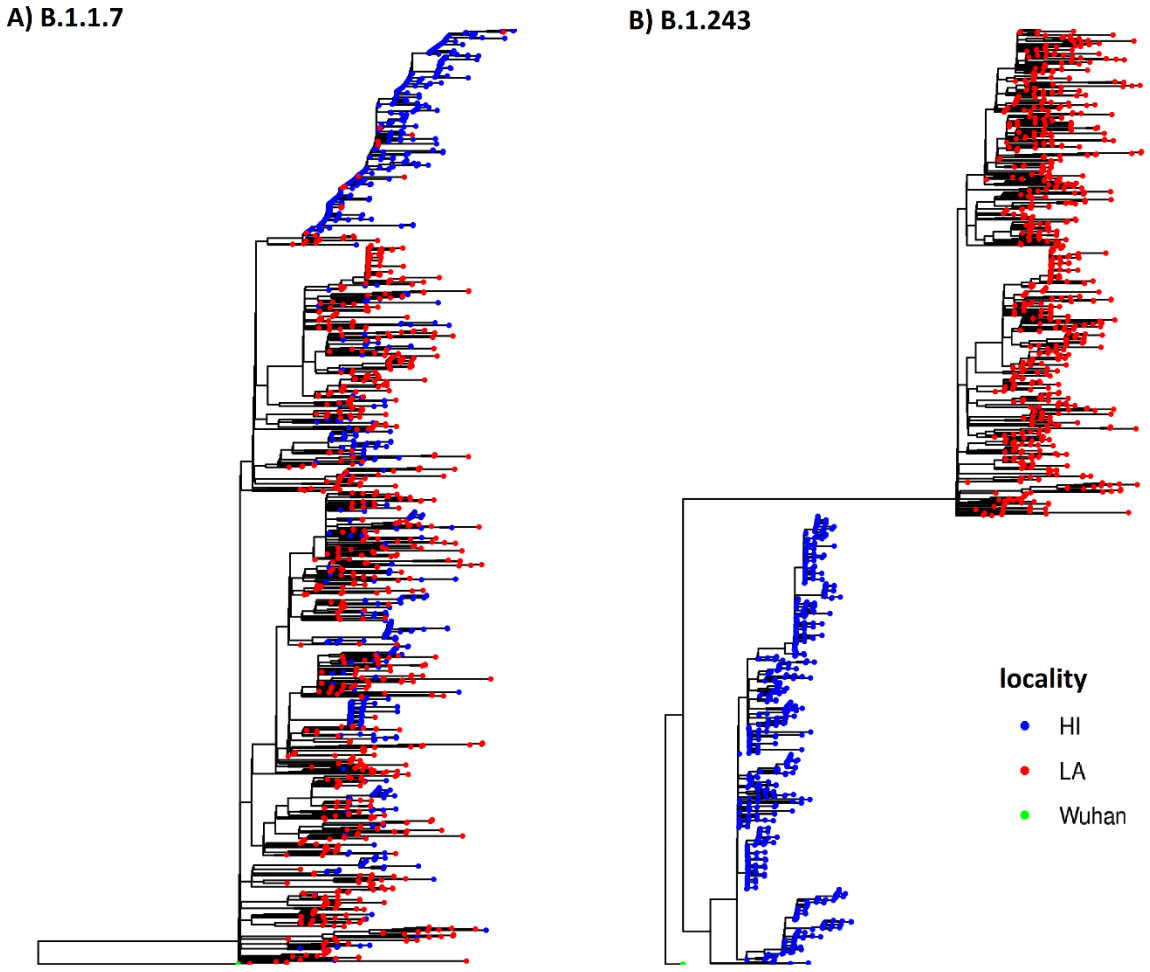

Figure 3. B.1.1.7 (A) and B.1.243 (B) ML phylogenies reconstructed from IQTREE. Genomes from Hawai'i are colored in blue, LA County in red, and the reference genome from Wuhan colored in green.

Table 4. B.1.1.7 Pairwise Nucleotide Distance (bp/length x 10^{-4})

B.1.1.7	Hawai'i	LA County
Hawai'i	4.50 ± 0.41	5.40 ± 0.43
LA County		5.57 ± 0.63

Table 5. B.1.243 Pairwise Nucleotide Distance (bp/length x 10^{-4})

B.1.243	Hawai'i	LA County
Hawai'i	2.67 ± 0.42	14.2 ± 1.5
LA County		7.50 ± 0.52

Opposing Trends Between Variants

There were more sites under selection for B.1.1.7 than B.1.243 within both localities (Figure 5). With high observed rates of mutation, a greater number of sites under selection were detected in ORF1a and S gene regions compared to ORF7a. Negative selection was more frequent than positive selection. The waning phase of B.1.1.7 has more sites under selection than the exponential phase in both localities, however the opposite is true for B.1.243 (Table 2; Figure 4).

These differences are also found in the estimated substitution rates. Each group has the highest rates of both synonymous and nonsynonymous substitutions in ORF7a, followed by S, then ORF1a (Table 2). However, ORF7a had very few sites identified as significantly experiencing either positive or negative selection.

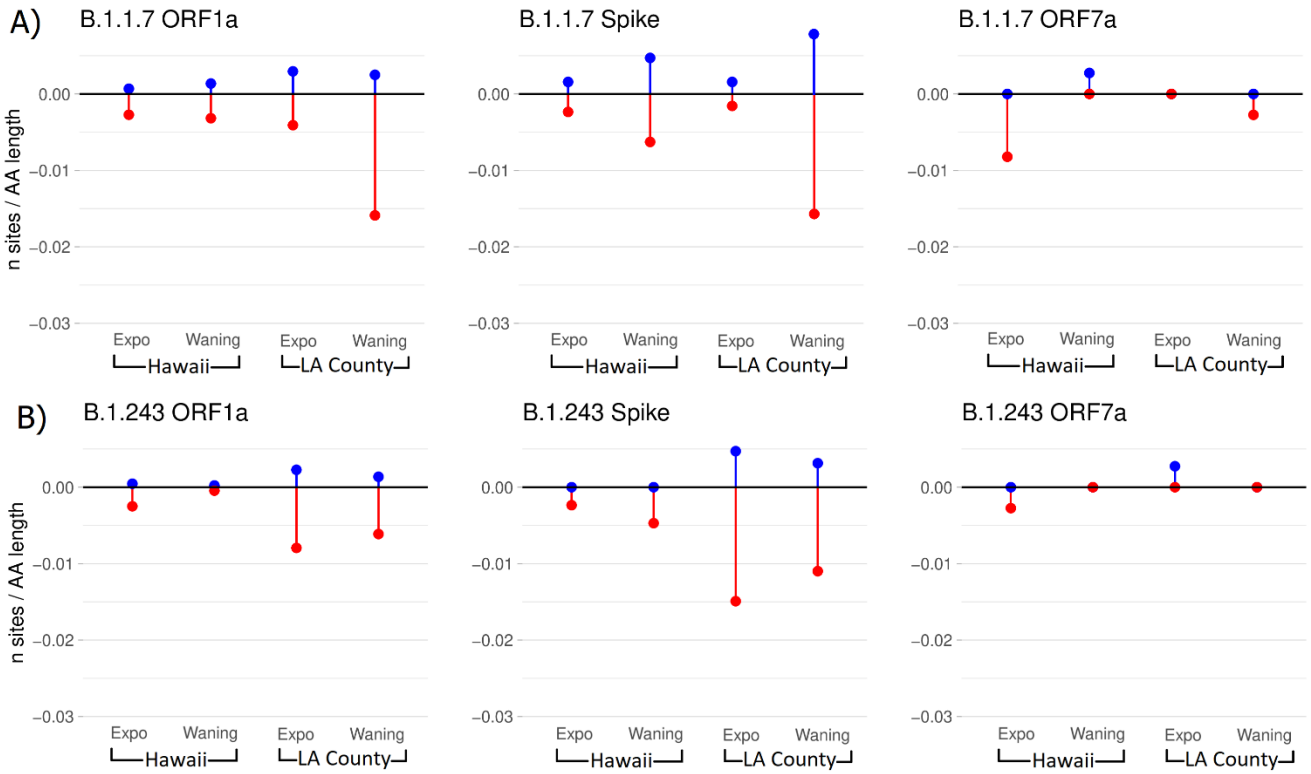

Figure 4. Number of sites in ORF1a, S, and ORF7a gene regions under selection in the exponential and waning phases for B.1.1.7 (A) and B.1.243 (B) in Hawai'i and LA County. Sites under positive selection (blue) are plotted above the line, and sites under negative selection (red) are plotted below the line. The number of sites are normalized to the length of the corresponding gene regions.

Selection is variable across gene regions, Expected Pattern of Evolution found in LA County B.1.1.7

The expected pattern of evolution is identified by elevated positive selection in the exponential phase and elevated negative selection in the waning phase. This pattern is observed in B.1.1.7 in LA County, but this combination does not occur in any of the other cases. Most gene regions follow either elevated positive selection in the exponential phase or elevated negative selection in the waning phase, but not both. B.1.243 tends to have the opposite pattern (Figure 4).

More sites with selection in LA County than Hawai'i, Most Sites are Unique

The majority of sites with detected selection for each variant were unique to a locality (Figure 5). Of the 174 sites in B.1.1.7, only 14 were shared across localities; of the 127 sites in B.1.243, only 3 were shared across localities. 25 of the total 181 sites were detected across both variants with majority of sites in common within with ORF1a region and no sites shared within ORF7a (Table 6). Site 924 in ORF1a was the only site detected across both localities and variants. The number of sites with detectable selection in both the exponential and waning phases are much higher in LA County than in Hawai'i.

Table 6. Common sites under selection across B.1.1.7 and B.1.243 variants. There were no common sites between variants within ORF7a.

Gene region	**Shared sites**
ORF1a	186, 549, 647, 918, 924, 1438, 1919, 2033, 2092, 2134, 2861, 3055, 3368, 3479, 3494, 3523, 3591, 3606, 3829, 4398
S	5, 138, 412, 424, 508

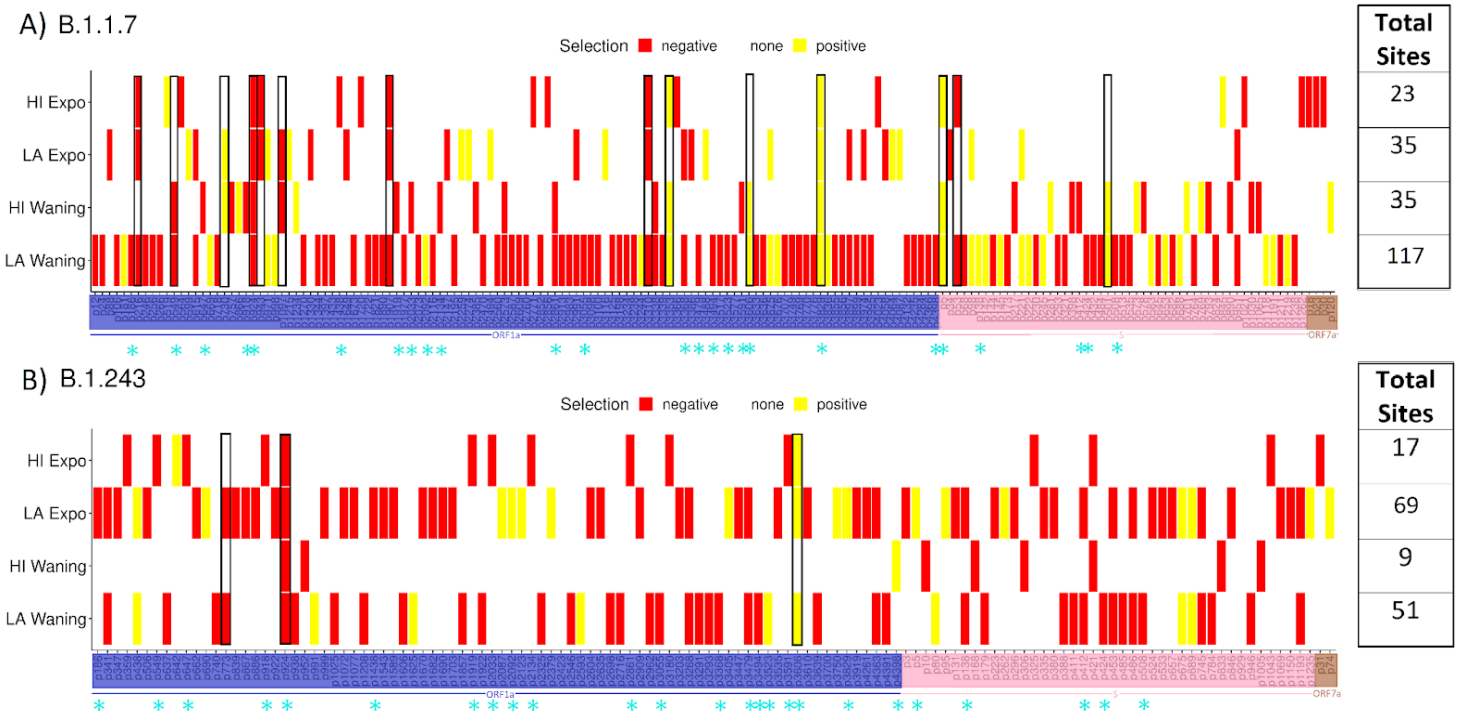

Figure 5. Heatmap of sites under positive (yellow) or negative (red) selection and corresponding amino acid position in ORF1a (blue) S (pink), and ORF7a (brown) gene regions for B.1.1.7 and B.1.243. Black boxes indicate sites with overlap across Hawai'i and LA County, blue asterisks indicate sites with overlap across variants, and table on the right lists the total number of sites under either positive or negative selection for the exponential and waning phases in Hawai'i and LA County.

More sites under selection in S1 than S2 in LA County

The S1 subunit has more sites under selection than S2, by 2x or more. This is true for both variants at both localities, except for B.1.1.7 in Hawaii. Comparing populations, Hawai'i has fewer sites under selection than LA County (Table 7 and 8).

Table 7. Number of sites under selection in B.1.1.7 in the S gene.

B.1.1.7	S1	S2
Hawai'i	9	9
LA County	28	9

Table 8. Number of sites under selection in B.1.243 in the S gene.

B.1.243	S1	S2
Hawai'i	5	3
LA County	25	9

DISCUSSION

The results from this study support the idea that surges in case incidence have had an impact on the evolution of SARS-CoV-2. Overall, SARS-CoV-2 case incidence in Hawai'i was low in the early part of the COVID-19 pandemic, yet certain variants reached high incidence locally. Alternatively, overall case incidence in LA County was high, but B.1.243 had low incidence. Looking at the data, Hawai'i both B.1.243 and B.1.1.7 reached high incidence, and in LA County B.1.1.7 reached high incidence (Table 1). The data reported here therefore represent three high incidence variant-locality combinations which I will call "scenarios" and one low incidence scenario. I do find that with high case incidence, there is more viral evolution. The rates of neutral evolution were higher in the three high incidence scenarios than in the low, however higher evolution did not translate directly in viral genetic diversity and higher levels of selection. There was little evidence for the hypothesis of accelerated protein adaptation followed by increased constraint, which I discuss below, however, I did find that there are more sites under negative selection than positive selection, and that there are more sites under selection (both positive and negative) in Los Angeles than in Hawaii. Taken together, there is significant

evolution occurring across sites and lineages of SARS-CoV-2, with very different dynamics of selection and viral adaptation occurring across localities which is discussed below.

There are several important differences to consider between the variants. B.1.243 was temporally earlier in the pandemic and at a time when both localities instituted strict lockdowns and was one of many variants that was never classified as a variant of concern. B.1.1.7 eventually replaced B.1.243 in both localities, coincident with a dip in case counts in each place. B.1.1.7 transmission seemed to be much higher than B.1.243 in LA County suggesting greater difficulty in control of the B.1.1.7 variant (Figure 1). The spread of both variants in Hawai'i was similar. In addition, B.1.243 originated in Hawaii and was later introduced to Los Angeles by a single introduction, and therefore the two localities represent two evolutionarily independent clades (Figure 3). B.1.1.7 on the other hand, has experienced more mixing between the two localities, yet as I discuss below, there are still large differences between these localities regarding the sites under selection.

Low Support for Expected Pattern of Evolution

There was little evidence for the hypothesis of accelerated protein adaptation followed by increased constraint, when comparing exponential versus waning phases with regard to the number of sites under selection. In some cases, there were more sites under positive selection in the exponential phase, and in other cases more sites under negative selection in the waning phase, but both predictions were only held in only the ORF1a gene region of B.1.1.7 in LA County (Table 2, Figure 4). B.1.243 generally opposed the expected pattern.

There was little to no selection occurring in the ORF7a gene region, so the focus here is on ORF1a and S. The increase in selective constraints through time that is observed in the ORF1a gene region in LA County B.1.1.7 genomes, corresponds to the pattern of evolution

detected in other studies (Rochman et al., 2021; Lin et al., 2021; Chaw et al., 2020). While there was an expected increase in negative selection in the ORF1a and S gene regions for B.1.1.7 in both localities, there was also typically an increase in positive selection indicating the continuation of adaptive evolution and lower selective constraints (Lin et al., 2021; Lynch and Conery, 2000). The expectation for this pattern is likely due to the opportunity for exploration of the fitness landscape being highest as the virus is introduced, with mutations occurring randomly across the genome. Although most mutations are neutral and have no impact on the virus, rare mutations may provide some sort of viral advantage or fitness impact (Korber et al., 2020; Kimura, 1991). As the virus continues to mutate, advantageous mutations have the potential to accumulate through fixation and selective constraints begin to increase as these mutations are causing a detrimental effect at an elevated rate. The ORF1a and S gene regions are likely under greater selective pressures, as they code for proteins involved in viral proliferation and transmission, and have the potential to acquire mutations that have a major functional impact (Brant et al., 2021; Wu et al., 2020). However, other gene regions may reveal additional evolutionary variation (Lin et al., 2021; Emam et al., 2021; Cao et al., 2021).

I note that these results conflict with what was found in the Lin et al. (2020) study, that reported early protein adaptation followed by increased constraint. However, this study includes many more sequences. The Lin et al., (2020) study set an unnecessarily high data filter (100% complete genomes), which limited the dataset to 265 to represent the entire US. Small sample size in the number of sequences included in a mutational analysis may result in the underrepresentation of mutational frequencies across the genome and ultimately skew the results.

Higher Levels of Neutral Evolution in Hawai'i, More sites under selection in LA County

There is a clear signal of greater synonymous substitution reflecting higher levels of neutral evolution in Hawai'i overall, which was unexpected in comparison to LA County, as Hawai'i is expected to have lower case incidence. The number of sites under selection, however, do not reflect these differences observed in substitution rates, despite the incorporation of these rates to detect selection (Li et al., 1985; Yang and Nielsen, 2000; Goldman and Yang 1994; Felsenstein, 2001). There were more sites under selection in LA County than in Hawai'i for both variants (Figure 6), which is particularly surprising for B.1.243 as incidence in LA County was much lower. This may be explained by random mutations across the genome providing no advantage nor disadvantage for the virus, resulting in relatively low selective pressures overall in Hawai'i, and may indicate differences in the environmental pressures across Hawai'i and LA County.

Sites Under Selection are Unique to Variants, Localities, and Time

The variants differed in the dynamics of sites under selection. Across both localities, B.1.1.7 contained more sites under selection than B.1.243, with the exception of the period of exponential increase in LA County. Furthermore, the majority of sites under selection were unique to not only the locality, which would also be explained by differences in environmental pressures across localities. Most sites under selection were also unique to either the exponential or waning phase, which may indicate variable environmental pressures across the genome through time (Figure 5).

Site 924 (bp2772) in ORF1a was the only site under selection detected across both localities and both variants. This site falls within non-structural protein 4 (nsp4), which plays a

role in the formation of double-membrane vesicles required for replication in coronaviruses (Oostra et al., 2007; Gadlage et al., 2010). Co-mutations were found between ORF1ab-C2772T and S-A1841G, however only the A1841G mutation in the spike protein was found across all genomes in this study, indicating active selection on ORF1ab-C2772T. C2772T mutations at this site were found across the globe in ~67% of all variants by April 9th, 2020 (Zhang et al., 2021). Co-mutations detected with this mutation are located in ORF1ab-C14144T and in S-A1841G (Zhou et al., 2020). The S-A1841G mutation results in the D614G substitution in the spike protein, which has been linked to increased transmissibility (Korber et al., 2020; Plante et al., 2021).

B.1.243 Genomes Between LA County and Hawai'i are Very Different

For B.1.243, the pairwise nucleotide distances indicate that LA County and Hawaii represent two distinct populations separated by a large distance (Tables 4 and 5), which is confirmed by the phylogeny (Figure 3). One transmission event from Hawai'i to LA County resulted in the subsequent spread within LA County. B.1.1.7 on the other hand, seemed to jump back and forth between localities, indicated by the mixing of sequences from each population interspersed throughout the tree (Figure 3) and lower pairwise nucleotide distances (Tables 4 and 5).

Within each individual locality, the pairwise nucleotide distances in LA County were higher than in Hawai'i, which is expected due to higher spread resulting in an increased opportunity for mutation and variation between genomes within a community (Scholle et al., 2013; Castellano et al., 2019). Implementation of strict travel restrictions in Hawaii occurred

while B.1.243 was in circulation, which may have limited the opportunity for spread within an introduction to other populations and may explain lower nucleotide distance values.

High diversity despite low case counts of B.1.243 in LA County may be explained by mixing with other populations contributing to the diversity within LA County. An alternative explanation might be an underrepresentation of B.1.243, but this is not likely given the strong sequencing effort in LA County to sequence ~10% of all positive cases (California Department of Public Health, 2022). It is clear, however, that B.1.243 evolved independently within each locality (Figure 4, Tables 4 and 5).

Evolution is variable within different gene regions, More sites under selection in S1 than S2 for B.1.1.7

Separation of the sites with detected selection in the spike protein into the two subunits, S1 and S2, sheds some light onto a potential relationship between variants of concern and sites under selection. LA County had more sites under selection in the S gene region than Hawai'i for both variants, particularly B.1.1.7, with a large proportion of these falling within the S1 subunit.

A recent study conducted by Kistler et al. (2022) identified a correlation between clade growth and mutation accumulation in the S1 region resulting in a high ratio of nonsynonymous to synonymous substitutions, supporting this hypothesis. The S1 subunit is the antigenic region that binds to the host cell receptor, whereas the S2 subunit is involved in viral-host cell membrane fusion and is more conserved than S1 in the Sarbecovirus genus (Walls et al., 2020). Therefore, it makes sense that higher rates of mutation or greater selection within the S1 region would be found in variants such as B.1.1.7, that have acquired mutations that confer an advantage. The increase in nonsynonymous substitutions may provide greater antigenic variation

within the binding region and ultimately result in a structural protein that has a higher affinity for the host cell receptor (Korber et al., 2020). Alternatively, if an advantageous mutation has already been acquired, fixation and conservation of these changes is likely as it results in higher viral fitness.

Rt Change Precedes Dynamic Changes in Case Counts

Rt is typically estimated using epidemiological models including data on data such as population size and which proportions of the population are susceptible, infected, and recovered (Brinks et al., 2020; Bicher et al., 2022; Ramos et al., 2021), however it can also be estimated using genetic data (Stadler et al., 2012; Lai et al., 2020; Farah et al., 2020). With the exception of the sharp peak in B.1.243 cases in LA County at the end of 2020, the average Rt values for each variant preceded dynamic changes in case counts through time, following the same trend. The changes in Rt followed those in case counts well. The fidelity of the genetic data to the epidemiological estimates are impressive considering that the genomic sampling represents a comparatively small subset of cases, which highlights the incredible amount of information found in sequence data and its potential usefulness for public health, especially as the pipeline from sequence generation to analysis can now be conducted in near real time.

This study highlights the rapid evolution of SARS-CoV-2 and potential for different evolutionary dynamics related to viral fitness that can occur between communities, even when connected by travel. These findings have several implications: larger outbreaks do lead to more evolution. However, not all evolution is equal with regard to viral fitness. We see that environmental pressures can be variable across localities, variants within the same locality can

evolve differently, and the dynamics of evolution can vary on many different levels. While interventions to control viral spread within a community are essential to prevent de novo evolution, it is also important to identify methods of control between communities. Human travel patterns serve to move diverse variants around the globe, providing an opportunity for the variants to experience a greater diversity of selective environments. Imported variants may thrive in a new environment and accumulate mutations that ultimately result in variants that are more fit. In addition, recombination in SARS-CoV-2 has been detected and is of greater concern when diverse variants are mixed within a community (Haddad et al., 2021; Ignatieva et al., 2022; Lacek et al., 2022; Jackson et al., 2021). Application of the methods used in this study across localities that differ in epidemiological factors may provide insight into evolution and variability of sequences within a community and impacts of safety measure implementation. Further studies should be conducted to test for support of a correlation between evolution in the spike protein subunits and variants of concern and explore the variability in genome diversity between populations that vary in levels of mixing with other populations.

REFERENCES

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S. A., Zhang, T., Gao, W., Cheng, C., Tang, X., Wu, X., Wu, Y., Sun, B., Huang, S., Sun, Y., Zhang, J., … Feng, T. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of

their close contacts in Shenzhen, China: A retrospective cohort study. *The Lancet Infectious Diseases*, *20*(8), 911–919. https://doi.org/10.1016/s1473-3099(20)30287-5

Bicher, M., Rippinger, C., Schneckenreither, G., Weibrecht, N., Urach, C., Zechmeister, M., Brunmeir, D., Huf, W., & Popper, N. (2022). Model based estimation of the SARS-COV-2 immunization level in Austria and consequences for herd immunity effects. *Scientific Reports*, *12*(1), 2872. https://doi.org/10.1038/s41598-022-06771-x

Billah, M. A., Miah, M. M., & Khan, M. N. (2020). Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PLOS ONE*, *15*(11). https://doi.org/10.1371/journal.pone.0242128

Brant, A. C., Tian, W., Majerciak, V., Yang, W., & Zheng, Z.-M. (2021). SARS-COV-2: From its discovery to genome structure, transcription, and replication. *Cell & Bioscience*, *11*(1). https://doi.org/10.1186/s13578-021-00643-z

Brinks, R., Küchenhoff, H., Timm, J., Kurth, T., & Hoyer, A. (2020). Epidemiological measures for informing the general public during the SARS-COV-2-outbreak: Simulation study about bias by incomplete case-detection. *MedRxiv*. https://doi.org/10.1101/2020.09.23.20200089

U.S. Census Bureau. (2022, April 15). *Honolulu County, Hawai'i*. Census.gov. Retrieved April 15, 2022, from https://www.census.gov/

California Department of Public Health. (2022a). *Covid-19 time-series metrics by county and state - statewide COVID-19 cases deaths tests*. California Health and Human Services Open Data Portal. Retrieved May 4, 2022, from https://data.chhs.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state/resource/046cdd2b-31e5-4d34-9ed3-b48cdbc4be7a

California Department of Public Health. (2022b). *Tracking Variants*. California Health and Human Services Open Data Portal. Retrieved July 30, 2022, from https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID-Variants.aspx

Castellano, D., Eyre-Walker, A., & Munch, K. (2019). Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other Homininae. *Genome Biology and Evolution*, *12*(1), 3550–3561. https://doi.org/10.1093/gbe/evz215

Centers for Disease Control and Prevention (CDC). (2022, May 4). *United States covid-19 cases and deaths by State over time*. Centers for Disease Control and Prevention. Retrieved May 4, 2022, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

Chaw, S.-M., Tai, J.-H., Chen, S.-L., Hsieh, C.-H., Chang, S.-Y., Yeh, S.-H., Yang, W.-S., Chen, P.-J., & Wang, H.-Y. (2020). The origin and underlying driving forces of the SARS-COV-2 outbreak. *Journal of Biomedical Science*, *27*(1). https://doi.org/10.1186/s12929-020-00665-8

Choudhuri, S. (2014). Fundamentals of Molecular Evolution. *Bioinformatics for Beginners*, 27–53. https://doi.org/10.1016/b978-0-12-410471-6.00002-5

County of Los Angeles Public Health. (2020, November 27). *Public health to add additional safety modifications to health officer order - targeted safer at home order comes after 5-day average of new cases is 4,751*. LISTING OF DEPARTMENT OF PUBLIC HEALTH PRESS RELEASES. Retrieved April 15, 2022, from http://publichealth.lacounty.gov/phcommon/public/media/mediapubdetail.cfm?unit=media&ou=ph&prog=media&prid=2830

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waites, W., Wong, K. L., van Zandvoort, K., Silverman, J. D., Diaz-Ordaz, K., Keogh, R., Eggo, R. M., … Edmunds, W. J. (2020). Estimated transmissibility and impact of SARS-COV-2 lineage B.1.1.7 in England. *Science*, *372*, eabg3055. https://doi.org/10.1101/2020.12.24.20248822

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19

in Real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/s1473-3099(20)30120-1

Farah, S., Atkulwar, A., Praharaj, M. R., Khan, R., Gandham, R., & Baig, M. (2020). Phylogenomics and phylodynamics of SARS-COV-2 retrieved genomes from India. *Future Virology*, *15*(11). https://doi.org/10.1101/2020.06.23.20138222

Felsenstein J. (2001). Taking variation of evolutionary rates between sites into account in inferring phylogenies, J Mol Evol., 2001, vol. 53 (447-455).

Gadlage, M. J., Sparks, J. S., Beachboard, D. C., Cox, R. G., Doyle, J. D., Stobart, C. C., & Denison, M. R. (2010). Murine hepatitis virus nonstructural protein 4 regulates virus induced membrane modifications and replication complex function. *Journal of Virology*, *84*(1), 280–290. https://doi.org/10.1128/jvi.01772-09

Goldman, N., and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA-sequences. Mol. Biol. Evol. 11:725–736.

Governor of the State of Hawai'i. (2020, August 26). *Gov. Ige Approves Mayor Caldwell's Stay at Home Order*. Retrieved April 15, 2022, from https://governor.Hawai'i.gov/newsroom/latest-news/governors-office-news-release-gov-ige-approves-mayor-caldwells-stay-at-home-order/.

Haddad, D., John, S. E., Mohammad, A., Hammad, M. M., Hebbar, P., Channanath, A., Nizam, R., Al-Qabandi, S., Al Madhoun, A., Alshukry, A., Ali, H., Thanaraj, T. A., & Al-Mulla, F. (2021). SARS-COV-2: Possible recombination and emergence of potentially more virulent strains. *PLOS ONE*, *16*(5), e0251368. https://doi.org/10.1371/journal.pone.0251368

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of Pathogen Evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Ignatieva, A., Hein, J., & Jenkins, P. A. (2022). Ongoing recombination in SARS-COV-2 revealed through genealogical reconstruction. *Molecular Biology and Evolution*, *39*(2), msac028. https://doi.org/10.1093/molbev/msac028

Irons, N. J., & Raftery, A. E. (2021). Estimating SARS-COV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proceedings of the National Academy of Sciences*, *118*(31), e2103272118. https://doi.org/10.1073/pnas.2103272118

Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., Nicholls, S. M., O'Toole, Á., Pacchiarini, N., Poplawski, R., Scher, E., Todd, F., Webster, H. J., Whitehead, M., Wierzbicki, C., … Rambaut, A. (2021). Generation and transmission of interlineage recombinants in the SARS-COV-2 pandemic. *Cell*, *184*(20), 5179–5188. https://doi.org/10.1016/j.cell.2021.08.014

Jiang, X., Mu, B., Huang, Z., Zhang, M., Wang, X., & Tao, S. (2010). Impacts of mutation effects and population size on mutation rate in asexual populations: A simulation study. *BMC Evolutionary Biology*, *10*(1). https://doi.org/10.1186/1471-2148-10-298

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Khare, S., Gurry, C., Freitas, L., B Schultz, M., Bach, G., Diallo, A., Akite, N., Ho, J., TC Lee, R., Yeo, W., Core Curation Team, G. I. S. A. I. D., & Maurer-Stroh, S. (2021). GISAID's role in pandemic response. *China CDC Weekly*, *3*(49), 1049–1051. https://doi.org/10.46234/ccdcw2021.255

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics*, *66*(4), 367–386. https://doi.org/10.1266/jjg.66.367

Kistler, K., & Bedford, T. (2021). *eLife*; 10:e64509. https://doi.org/10.7554/eLife.64509

Kistler, K. E., Huddleston, J., & Bedford, T. (2022). Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-COV-2. *Cell Host & Microbe*, *30*(4). https://doi.org/10.1016/j.chom.2022.03.018

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., McDanal, C., Perez, L. G., Tang, H., … Wyles, M. D. (2020). Tracking changes in SARS-COV-2 spike: Evidence that D614g increases infectivity of the COVID-19 virus. *Cell*, *182*(4), 812–827. https://doi.org/10.1016/j.cell.2020.06.043

Kosakovsky Pond, S. L., & Frost, S. D. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, *22*(5), 1208–1222. https://doi.org/10.1093/molbev/msi105

Koçhan, N., Eskier, D., Suner, A., Karakülah, G., & Oktay, Y. (2021). Different selection dynamics of S and rdrp between SARS-COV-2 genomes with and without the dominant mutations. *Infection, Genetics and Evolution, 91*, 104796. https://doi.org/10.1016/j.meegid.2021.104796

Lacek, K. A., Rambo-Martin, B. L., Batra, D., Zheng, X.-yu, Sakaguchi, H., Peacock, T., Keller, M., Wilson, M. M., Sheth, M., Davis, M. L., Borroughs, M., Gerhart, J., Hassell, N., Shepard, S. S., Cook, P. W., Lee, J., Wentworth, D. E., Barnes, J. R., Kondor, R., & Paden, C. R. (2022). Identification of a novel SARS-COV-2 delta-omicron recombinant virus in the United States. *BioRxiv*. https://doi.org/10.1101/2022.03.19.484981

Lai, A., Bergna, A., Acciarri, C., Galli, M., & Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-COV-2. *Journal of Medical Virology*, *92*(6), 675–679. https://doi.org/10.1002/jmv.25723

Li, W. H., C. I. Wu, and C. C. Luo. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, 2(1), 150–174.

Lin, Z., Qing, H., Li, R., Zheng, L., & Yao, H. (2021). Evolution trace of SARS-COV-2 from January 19 to March 12, 2020, in the United States. *Journal of Medical Virology*, *93*(12), 6595–6604. https://doi.org/10.1002/jmv.27225

Lindstrøm, J. C., Engebretsen, S., Kristoffersen, A. B., Rø, G. Ø., Palomares, A. D.-L., Engø-Monsen, K., Madslien, E. H., Forland, F., Nygård, K. M., Hagen, F., Gantzel, G., Wiklund, O., Frigessi, A., & de Blasio, B. F. (2021). Increased transmissibility of the alpha SARS-COV-2 variant: Evidence from contact tracing data in Oslo, January to February 2021. *Infectious Diseases*, *54*(1), 72–77. https://doi.org/10.1080/23744235.2021.1977382

Liu, Q., Zhao, S., Hou, Y., Ye, S., Sha, T., Su, Y., Zhao, W., Bao, Y., Xue, Y., & Chen, H. (2020). Ongoing natural selection drives the evolution of SARS-COV-2 genomes. *MedRxiv*. https://doi.org/10.1101/2020.09.07.20189860

Liu, J., Zhang, Y., Lei, X. *et al.* (2008). Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol* 9, R69. https://doi.org/10.1186/gb-2008-9-4-r69

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Maison, D. P., Ching, L. L., Cleveland, S. B., Tseng, A. C., Nakano, E., Shikuma, C. M., & Nerurkar, V. R. (2021). Algorithm for the quantitation of variants of concern for rationally designed vaccines based on the isolation of SARS-COV-2 hawai‘i lineage B.1.243. *BioRxiv*. https://doi.org/10.1101/2021.08.18.455536

Miller, D., Martin, M. A., Harel, N., Tirosh, O., Kustin, T., Meir, M., Sorek, N., Gefen-Halevi, S., Amit, S., Vorontsov, O., Shaag, A., Wolf, D., Peretz, A., Shemer-Avni, Y., Roif-Kaminsky, D., Kopelman, N. M., Huppert, A., Koelle, K., & Stern, A. (2020). Full genome viral sequences inform patterns of SARS-COV-2 spread into and within Israel. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19248-0

Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., Shaw, P. D., & Marshall, D. (2012). Using tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. https://doi.org/10.1093/bib/bbs012

Monteiro, F., Marcet, P., & Dorn, P. (2010). Population genetics of triatomines. *American Trypanosomiasis*, 169–208. https://doi.org/10.1016/b978-0-12-384876-5.00008-3

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). Fubar: A fast, unconstrained bayesian approximation for inferring selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205. https://doi.org/10.1093/molbev/mst030

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, *8*(7). https://doi.org/10.1371/journal.pgen.1002764

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nie, Q., Li, X., Chen, W., Liu, D., Chen, Y., Li, H., Li, D., Tian, M., Tan, W., & Zai, J. (2020). Phylogenetic and phylodynamic analyses of SARS-COV-2. *Virus Research*, *287*, 198098. https://doi.org/10.1016/j.virusres.2020.198098

Oostra, M., te Lintelo, E. G., Deijs, M., Verheije, M. H., Rottier, P. J., & de Haan, C. A. (2007). Localization and membrane topology of coronavirus nonstructural protein 4: Involvement of the early secretory pathway in replication. *Journal of Virology*, *81*(22), 12323–12336. https://doi.org/10.1128/jvi.01506-07

Otto, S. P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, D. J. D., & Ogden, N. H. (2021). The origins and potential future of SARS-COV-2 variants of concern in the evolving COVID-19 pandemic. *Current Biology*, *31*(14). https://doi.org/10.1016/j.cub.2021.06.049

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., Yeats, C., Du Plessis, L., Maloney, D., Medd, N., Attwood, S. W., Aanensen, D. M., Holmes, E. C., Pybus, O. G., & Rambaut, A. (2021). Assignment of epidemiological lineages in an emerging pandemic using The pangolin tool. *Virus Evolution*, *7*(2), veab064. https://doi.org/10.1093/ve/veab064

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Zhang, X., Muruato, A. E., Zou, J., Fontes-Garfias, C. R., Mirchandani, D., Scharton, D., Bilello, J. P., Ku, Z., An, Z., Kalveram, B., Freiberg, A. N., Menachery, V. D., Xie, X., … Shi, P.-Y. (2020). Spike mutation D614G alters SARS-COV-2 fitness. *Nature*, *592*(7852), 116–121. https://doi.org/10.1038/s41586-020-2895-3

Plessis, L. (2016). bdskytools: BDSKY Tools. R package version 0.0.1.0.

Quick, J. (2020, April 9). *NCoV-2019 Sequencing protocol V2 (GunIt)*. protocols.io. Retrieved April 26, 2022, from https://dx.doi.org/10.17504/protocols.io.bdp7i5rn

Ramanathan, M., Ferguson, I. D., Miao, W., & Khavari, P. A. (2021). SARS-COV-2 B.1.1.7 and B.1.351 spike variants bind human ACE2 with increased affinity. *The Lancet Infectious Diseases*, *21*(8), 1070. https://doi.org/10.1016/s1473-3099(21)00262-0

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Ramos, A. M., Vela-Pérez, M., Ferrández, M. R., Kubik, A. B., & Ivorra, B. (2021). Modeling the impact of SARS-COV-2 variants and vaccines on the spread of covid-19. *Communications in Nonlinear Science and Numerical Simulation*, *102*, 105937. https://doi.org/10.1016/j.cnsns.2021.105937

Rochman, N. D., Wolf, Y. I., Faure, G., Mutz, P., Zhang, F., & Koonin, E. V. (2021). Ongoing global and regional adaptive evolution of SARS-COV-2. *Proceedings of the National Academy of Sciences*, *118*(29). https://doi.org/10.1073/pnas.2104241118

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vex042

Scholle, S. O., Ypma, R. J., Lloyd, A. L., & Koelle, K. (2013). Viral substitution rate variation can arise from the interplay between within-host and Epidemiological Dynamics. *The American Naturalist*, *182*(4), 494–513. https://doi.org/10.1086/672000

Scripps Research. (2021, March 30). *B.1.1.7 Variant of COVID-19 Virus Spreading Rapidly in United States*. Retrieved April 15, 2022, from https://www.scripps.edu/news-and-events/press-room/2021/20210330-andersen-b117-variant-covid19.html.

Shetty, R. M., Achaiah, N. C., & Subbarajasetty, S. B. (2020). R0 and RE of COVID-19: Can we predict when the pandemic outbreak will be contained? *Indian Journal of Critical Care Medicine*, *24*(11), 1125–1127. https://doi.org/10.5005/jp-journals-10071-23649

Stadler, T., Kühnert, D., Bonhoeffer, S., & Drummond, A. J. (2012). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, *110*(1), 228–233. https://doi.org/10.1073/pnas.1207965110

State of Hawai'i, Department of Health. (2020, October 15). *Hawai'i COVID-19 Daily News Digest October 15, 2020*. Retrieved April 15, 2022, from https://health.Hawai'i.gov/news/covid-19/Hawai'i-covid-19-daily-news-digest-october-15-2020/.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and Phylodynamic Data Integration using beast 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Tamura, K., Stecher, G., & Kumar, S. (2021). Mega11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, function, and antigenicity of the SARS-COV-2 spike glycoprotein. *Cell*, *181*(2). https://doi.org/10.1016/j.cell.2020.02.058

Weaver, S., Shank, S. D., Spielman, S. J., Li, M., Muse, S. V., & Kosakovsky Pond, S. L. (2018). Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. *Molecular Biology and Evolution*, *35*(3), 773–777. https://doi.org/10.1093/molbev/msx335

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, LD., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Yang, Z. H., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17:32–43.

Yashvardhini, N., Jha, D. K., Kumar, A., Sayrav, K., & Gaurav, M. (2021). Genetic variations in the ORF7A protein of SARS-COV-2 and its possible role in vaccine development. *Biomedical Research and Therapy*, *8*(8), 4497–4504. https://doi.org/10.15419/bmrat.v8i8.688

Zeileis A, Grothendieck G (2005). "zoo: S3 Infrastructure for Regular and Irregular Time Series." Journal of Statistical Software, 14(6), 1–27. doi: 10.18637/jss.v014.i06.

Zhang, M., Li, L., Luo, M., & Liang, B. (2021). Genomic characterization and evolution of SARS-COV-2 of a Canadian population. *PLOS ONE*, *16*(3), e0247799. https://doi.org/10.1371/journal.pone.0247799

Zhou, Z.-J., Qiu, Y., Pu, Y., Huang, X., & Ge, X.-Y. (2020). BioAider: An efficient tool for viral genome analysis and its application in tracing SARS-COV-2 transmission. *Sustainable Cities and Society*, *63*, 102466. https://doi.org/10.1016/j.scs.2020.102466