Article

Intrahost SARS-CoV-2 k-mer identification method (iSKIM) for rapid detection of mutations of concern reveals emergence of global mutation patterns

Ashley Thommana [1,2], Migun Shakya [3], Jaykumar Gandhi [1], Christian K. Fung [1], Patrick S.G. Chain [3] Irina Maljkovic Berry [1,4], and Matthew A. Conte[1,*]

[1] Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, MD, USA
[2] Montgomery Blair High School, Silver Spring, MD, USA
[3] Los Alamos National Laboratory, Biosecurity and Public Health Group, Bioscience Division, Los Alamos, NM, USA
[4] Integrated Research Facility, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Frederick, MD, USA
* Corresponding author

Abstract: Despite unprecedented global sequencing and surveillance of SARS-CoV-2, timely identification of the emergence and spread of novel variants of concern (VoCs) remains a challenge. Several million raw genome sequencing runs are now publicly available. We sought to survey these datasets for intrahost variation to study emerging mutations of concern. We developed iSKIM ("intrahost SARS-CoV-2 k-mer identification method") to relatively quickly and efficiently screen the many SARS-CoV-2 datasets to identify intrahost mutations belonging to lineages of concern. Certain mutations surged in frequency as intrahost minor variants just prior to, or while lineages of concern arose. The Spike N501Y change common to several VoCs was found as a minor variant in 834 samples as early as October 2020. This coincides with the timing of the first detected samples with this mutation in the Alpha/B.1.1.7 and Beta/B.1.351 lineages. Using iSKIM, we also found that Spike L452R was detected as an intrahost minor variant as early as September 2020, prior to the observed rise of the Epsilon/B.1.429/B.1.427 lineages in late 2020. iSKIM rapidly screens for mutations of interest in raw data, prior to genome assembly, and can be used to detect increases in intrahost variants, potentially providing an early indication of novel variant spread.

Keywords: SARS-CoV-2, COVID-19, variants of concern, intrahost variation, mutation, genomic sequencing, bioinformatics, computational genomics

1. Introduction

The unprecedented biomedical research focus on the COVID-19 pandemic has provided an unparalleled amount of genomic data for studying virus evolutionary processes in novel and more detailed ways. Researchers have submitted and made public full-length SARS-CoV-2 genomes together with, and to a lesser extent, the accompanying raw sequencing data in efforts to monitor and surveil how the virus is changing in near real-time [1]. For example, a mutation causing an amino acid change in the Spike protein, D614G, which likely increased the fitness of SARS-CoV-2, spread globally from early to mid-2020 and has since become effectively fixed [2,3]. The amount of change in SARS-CoV-2 genomes remained low until late 2020, when SARS-CoV-2 lineage B.1.1.7 [4], subsequently designated 'Alpha' by the World Health Organization (WHO) [5], was identified in the United Kingdom [6]. Alpha exhibited a fitness advantage allowing it to outcompete other circulating lineages [7,8]. The fitness advantage of the Alpha lineage was likely driven by the presence of a number of novel mutations, particularly within the

Spike gene where the N501Y change in the receptor binding domain has been shown to increase binding affinity to the ACE-2 receptor [9]. Alpha has also been shown to have a modest ability to evade neutralizing antibodies from prior infection or vaccination [10]. Additional lineages with genetic changes predicted or known to impact spread, disease severity, diagnostic or therapeutic escape, and identified to cause significant community transmission in multiple countries, have been deemed by the WHO as "variants of interest" (VoI) [5]. Such lineages can then be deemed "variants of concern" (VoC) if they also show the ability to cause a detrimental change in COVID-19 epidemiology, increase in virulence, and/or decrease in public health measures [5]. Several additional VoIs and VoCs have been identified: B.1.351/Beta first identified in South Africa [11], P.1/Gamma and P.2/Zeta first identified in Brazil [12,13], B.1.617.2/Delta and AY/Delta first identified in India [14,15], and B.1.1.529/Omicron and BA/Omicron first identified in South Africa and Botswana [16,17].

Most of the genetic sequence analysis of SARS-CoV-2 has focused on consensus genome sequences. However, viruses often exhibit variation within an individual host and exist (and transmit) as a population of variants [18,19]. High throughput genome sequencing methods have been developed to analyze intrahost variation present within genome sequencing data [20,21] and many of the SARS-CoV-2 sequencing experiments are performed using amplicon-based sequencing [22,23]. Intrahost variation in SARS-CoV-2 has now been characterized from several different perspectives including mutation profile differences between intrahost and consensus SNPs [24], specifically within the context of specific geographical regional dynamics [25], across time within the same patients [26,27], and within patients with cancer [28].

Studying intrahost dynamics across hundreds of thousands or millions of samples remains a computationally challenging endeavor both in terms of disk storage of input data and output files, as well as raw compute power. Due to these limitations, previous studies of SARS-CoV-2 intrahost variation have focused on up to ~15,000 datasets [29]. Improving the speed of existing software for analyzing intrahost variation has shown promise [30]. However, alternative approaches for analyzing this large amount of data remain appealing. Counting of relatively short sequences of length k, or 'k-mers', has proven to be a very fast and efficient bioinformatics approach for many different types of high throughput sequencing datasets due to the ability to avoid more traditional and time-consuming alignment and post-processing steps [31–33]. For instance, a k-mer based tool, fastv, has been developed for detecting SARS-CoV-2 and other pathogens in high throughput sequencing data by providing a set of pathogen-specific k-mers [34]. In addition to providing a SARS-CoV-2 specific k-mer sets, fastv can take as input arbitrary user provided k-mers to allow for flexibility in what a user can screen for. Here we present iSKIM ("intrahost SARS-CoV-2 k-mer identification method") as a novel approach with lineage-specific k-mers for the SARS-CoV-2 VoCs/VoIs. These VoC/VoI specific k-mers can then be used for quick k-mer screening of SARS-CoV-2 sequencing datasets to identify samples containing VoC/VoI mutations as intrahost variants and/or consensus variants. iSKIM provides post-processing tools to summarize the screening results and can enable researchers to prioritize particular samples for more complex analyses such as reference-based genome assembly, curation and downstream analysis when sequencing or analyzing many samples at once.

We also apply iSKIM by scanning for VoC/VoI mutations across publicly available SARS-CoV-2 data in the NCBI Sequence Read Archive (SRA) database. Our analysis identified patterns and trends regarding the frequency of VoC mutations among the datasets and the intrahost diversity of samples. Further application of this technique to newly deposited SARS-CoV-2 raw data may provide an earlier way to forecast potential increases of novel mutations that may become fixed in current or emerging variants.

2. Materials and Methods

2.1. *Variant of Concern lineage-specific k-mer generation*

K-mer sequences of 21bp in length were generated for each of the PANGO lineages [4] investigated in this study (B.1.1.7/Alpha, B.1.351/Beta, P.1/Gamma, P.2/Zeta, B.1.429/Epsilon, B.1.526/Iota and B.1.617.2/Delta). The lineage defining and most common mutations for each lineage were obtained from several sources for validation and completeness [35–37]. Typically, lineage defining mutations are listed as as amino acid changes in proteins (e.g. Spike N501Y) and not typically has genome reference coordinates. However, to generate k-mers, the nucleotide changes are required. Representative sets of genomes for each lineage were obtained from NCBI and the corresponding lineage defining mutations were matched for those listed in amino acid coordinates to reference coordinates (e.g. Spike N501Y is A23063T). Mutations were defined based on the coordinates of the SARS-CoV-2 reference genome (NCBI accession number NC_045512.2) [38]. Bgzip (version 1.9) and tabix (version 1.9) [39,40] were used to create a compressed and indexed VCF file containing lineage specific mutations, separately for each mutation. These compressed and indexed VCF files were then used to create a consensus reference containing each mutation using the bcftools (version 1.9) 'consensus' command by supplying the NC_045512.2 reference and each mutation specific VCF file. A BED file containing the reference coordinate positions 10bp upstream to 11bp downstream of each mutation were created. The bedtools (v.2.30.0) [41] 'getfasta' command was then used by supplying the mutation fasta file previously generated and the appropriate BED file, to generate a 21bp FASTA file for each mutation. Each 21bp k-mer FASTA file was then combined for each lineage to represent the set of mutations for each lineage. Additionally, a set of comparison reference k-mers for each lineage was generated in a similar fashion using simply the SARS-CoV-2 reference sequence (NC_045512.2) at these same posistions, but without any mutations.

2.2. *Obtaining and formatting NCBI SRA data*

The NCBI SRA database was queried using the phrase "SARS-CoV-2" on May 12, 2021. The BioProject accession identifiers associated with the reads were generated by navigating to the related database section of the page or by querying the BioProject database with "(SARS-CoV-2) AND bioproject_sra[filter] NOT bioproject_gap[filter]." Only SARS-CoV-2 samples sequenced with the Illumina platform were used for this study. Any samples without the collection month and year, and without a geographic location (country) were not used in the dataset. Any samples with a collection date prior to Nov 2019 were removed. The resulting dataset consisted of 411,805 SRA samples. These NCBI SRA files were downloaded via NCBI FTP [42] . SRA files were converted to gzip compressed fastq files using the 'fastq-dump' program from NCBI SRA [43,44] toolkit version 2.10.9 with the following parameters: '*--split-3 --gzip*'.

2.3. *Screening NCBI SRA data for Variant of Concern k-mers*

Fastv (version 0.8.1) [34] was ran on each NCBI SRA fastq.gz file in paired mode (--in1 and --in2) for SRA accessions with paired-end reads, and simply (--in1) for SRA accessions with single-end reads. The custom k-mer sets for each lineage (B.1.1.7, P.1, B.1.351, B.1.429, B.1.617.1, B.1.617.2, B.1.526.) were supplied separately to fastv with the '-k' option and both html ('-h') and JSON ('-j') output files were generated.

The Fastv JSON output for both the lineage k-mers and corresponding reference k-mer sets was parsed and the proportion of lineage to reference counts were used to determine if mutations belonging to each lineage were present at a minor variant level (> 1% to 50%) or as fixed mutations (>99%). This threshold of 1% or greater was chosen to capture a large amount of minor variants without approaching the error rate of various

Illumina instruments [45]. A minimum coverage of 5 reads of the reference allele and a minimum of coverage of 5 reads of the mutation allele were required for candidate minor variants.

2.4. *Inspecting for primer induced mutations using ARTIC primer schemes*

The popular ARTIC primer schemes for versions 1 through version 4.1 were downloaded from https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019. The BED files were visualized in IGV [46] alongside specific VoC mutations corresponding to the N501Y and L452R Spike changes to verify that these were not primer induced mutations.

2.5. *Comparison of iSKIM to LoFreq and ngs_mapper on select NCBI SRA data*

834 samples containing the N501Y Spike change in samples from October 2020 and the 68 samples containing the L452R Spike change in samples from September 2020 as identified by iSKIM were run through ngs_mapper (version v1.5.4) [47] and LoFreq (version 2.1.4) to compare the k-mer generated call frequencies to frequencies generated by reference-based read assembly. The Wuhan-Hu-1 genome (NCBI accession: NC_045512.2) was used as the reference in both cases.

2.6. *Phylogenetic analysis of select SARS-CoV-2 genomes*

NCBI blastn [48] version 2.11.0+ was used to query the consensus genomes of the 834 NCBI SRA samples identified by iSKIM as having the N501Y Spike change as a minor variant against the GISAID EpiCov database obtained April 17, 2022. The consensus genomes of these 834 samples were downloaded from GISAID. A blastn e-value cut-off ('-evalue') of 1e-250 and a percent identity cut-off ('-perc_identity') of 99.9 were used. The resulting top 5 blast hits for each query sequence were taken. An additional custom set of 3,243 background reference samples were obtained from the NextStrain SARS-CoV-2 global build [49] and added. These were combined with the 834 query sequences and a multiple sequence alignment to the Wuhan-Hu-1 reference (NCBI accession: NC_045512.2) was generated using MAFFT [50] version v7.475 with the following settings: '--auto --keeplength –addfragments'. This alignment was used as input to generate a maximum likelihood phylogeny using FastTree version 2.1.11 [51]. The same process was used to generate a tree for the 68 NCBI SRA samples identified by iSKIM as having the L452R Spike change present as a minor variant, except that the top 50 blastn hits were used instead, and all other settings remained the same as described above. Trees were visualized and formatted using Figtree version 1.4.4 .

3. Results

3.1. *iSKIM analysis of SARS-CoV-2 NCBI SRA data by month*

411,805 samples obtained from the NCBI SRA with collection dates spanning 14 months (February 2020 – April 2021) were screened using iSKIM for mutations belonging to the following lineages of concern/interest: B.1.1.7/Alpha, B.1.351/Beta, P.1/Gamma, P.2/Zeta, B.1.429/Epsilon, B.1.526/Iota and B.1.617.2/Delta. VoC mutations that were either fixed or present as a minor variant in each sample (>1%, see methods for details) were then tabulated across all samples. The spike N501Y change, which is present in most samples of B.1.1.7/Alpha, B.1.351/Beta, and P.1/Gamma, was found in a number of samples either as a fixed variant or as a minor variant (Table 1). Several patterns emerged from examining the N501Y change across each month in this period. 834 samples were detected that had the N501Y change present as a minor variant in October of 2020. The number of samples with N501Y present as a minor variant then decreased in November 2020, as the B.1.1.7/Alpha lineage became more prevalent and as the number of samples

fixed for N501Y increased. There were 34 samples collected from Australia that had the fixed N501Y change prior to the emergence of B.1.1.7/Alpha or the other VoCs in the June/July of 2020. These samples from Australia have been previously identified in other studies [52,53]. 32 of these Australian samples are assigned to Pango [54] lineage B.1.1.136 and two were assigned to Pango lineage B.1.1 which suggests additional convergence of the N501Y change.

Table 1. Number of samples containing the Spike N501Y change present as a fixed variant or minor variant in NCBI SRA samples across each month as identified by iSKIM. Numbers in bold represent months where a high frequency of samples with the N501Y change was identified prior to emergence first the Alpha VoC.

Month and Year	# of NCBI SRA samples screened	# NCBI SRA samples fixed for N501Y	Fraction of NCBI SRA samples fixed for N501Y	# samples with N501Y present as a minor variant	Fraction of samples with N501Y present as a minor variant
February 2020	298	0	0.0000	0	0.0000
March 2020	14279	0	0.0000	3	0.0002
April 2020	16396	0	0.0000	2	0.0001
May 2020	8085	0	0.0000	1	0.0001
June 2020	10381	**31**	0.0030	4	0.0004
July 2020	10344	**3**	0.0003	5	0.0005
August 2020	9646	0	0.0000	0	0.0000
September 2020	11000	19	0.0017	5	0.0005
October 2020	22710	240	0.0106	**834**	**0.0367**
November 2020	22671	1618	0.0714	56	0.0025
December 2020	26274	10405	0.3960	80	0.0030
January 2021	69019	49666	0.7196	442	0.0064
February 2021	61025	51801	0.8488	216	0.0035
March 2021	81301	73298	0.9016	220	0.0027
April 2021	28507	24882	0.8728	53	0.0019

Another key spike change, L452R, has been shown to be associated with increased transmission (*in vivo*), infectivity (*in vivo*), and causes reduced antibody neutralization from infected patients and vaccinated individuals [55], as well as escaping HLA-A24-restricted cellular immunity [56]. Spike L452R also shows a similar pattern as N501Y (Table 2). 68 samples were detected that had the L452R change present as a minor variant in September of 2020. The number of samples with L452R present as a minor variant then decreased in October 2020 and then in December 2020, three months later, B.1.429/Epsilon became more prevalent and the number of samples fixed for L452R increased.

Table 2. Number of samples containing the Spike L452R change present as a fixed variant or minor variant in NCBI SRA samples across each month as identified by iSKIM. Numbers in bold represent the month where a high frequency of samples with the L452R change was identified prior to emergence first in the Epsilon VoI.

Month and Year	# of NCBI SRA samples screened	# NCBI SRA samples fixed for L452R	Fraction of NCBI SRA samples fixed for L452R	# samples with L452R present as a minor variant	Fraction of samples with L452R present as a minor variant
February 2020	298	0	0.0000	0	0.0000
March 2020	14279	0	0.0000	2	0.0001
April 2020	16396	0	0.0000	1	0.0001
May 2020	8085	0	0.0000	0	0.0000
June 2020	10381	0	0.0000	7	0.0007
July 2020	10344	0	0.0000	0	0.0000
August 2020	9646	0	0.0000	11	0.0011
September 2020	11000	0	0.0000	**68**	**0.0062**
October 2020	22710	8	0.0004	15	0.0007
November 2020	22671	17	0.0007	2	0.0001
December 2020	26274	257	0.0098	11	0.0004
January 2021	69019	1525	0.0221	201	0.0029
February 2021	61025	1293	0.0212	172	0.0028
March 2021	81301	1381	0.0170	113	0.0014
April 2021	28507	825	0.0289	23	0.0008

3.2. *Phylogenetic analysis of early N501Y and L452R minor variant samples*

To confirm that the samples containing the N501Y and L452R spike changes identified as minor variants were not all or primarily found in the same outbreaks or in close transmission chains, global phylogenetic analyses including the consensus genomes of these samples from October 2020 (N501Y) and September 2020 (L452R) were performed. The resulting global trees indicate that the samples containing these minor variant changes emerged independently multiple times (Figure 1 and Figure 2, respectively) and were not part of close transmission chains or related outbreaks. These findings indicate a pattern where a mutation presents itself as a minor variant a few months prior to gaining prevalance as a fixed mutation. The majority of the 834 samples identified as having the N501Y change as a minor variant surge in October 2020 (Table 1) were assigned to Pango lineage B.1.177 and its sublineages (n=477, Table S1). Similarly, of the 68 samples identified as having L452R as a minor variant surge in September 2020 (Table 2), B.1.177 and its sublineages were the most common although not the majority (n=31, Table S2).

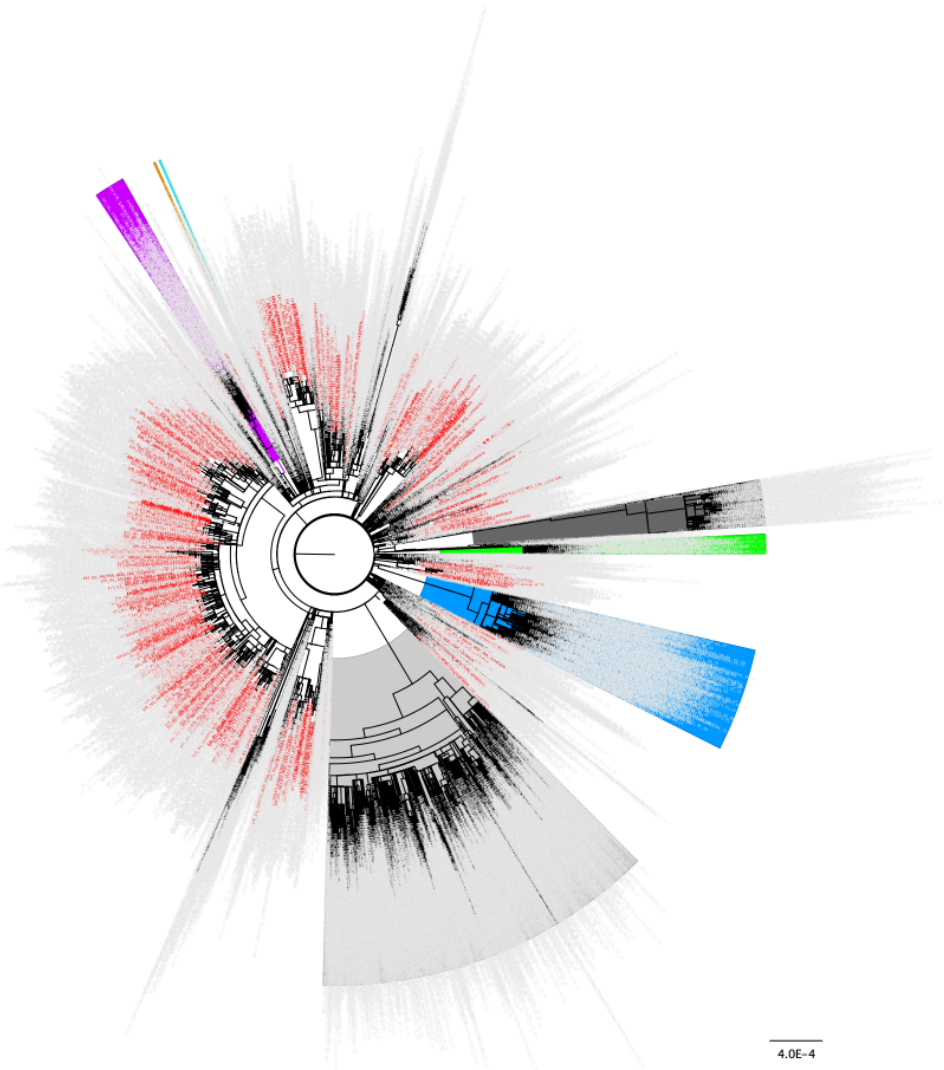

Figure 1. Distribution of the 843 samples containg Spike N501Y as a minor variant from October 2020 across the global SARS-CoV-2 phylogeny indicating independent emergence. Background lineages include Alpha/B.1.17 samples highlighted in blue, Gamma/P.1 highlighted in green, Beta/B.1.351 highlighted in purple, Epsilon/B.1.429 highlighted in orange, Iota/B.1.526 highlighted in turquoise, Delta/B.1.617.2 highlighted in grey, and Omicron/BA.1/BA.2 highlighted in dark grey. None of the 834 samples containing Spike N501Y as a minor variant in Octoboer 2020 were present in these highlighted lineages. Non VoC/VoI lineages are not highlighted. 834 samples identified as having the N501Y change present as a minor variant in October 2020 (Table 1) are colored in red. 3,243 total background genomes were included in this analysis.

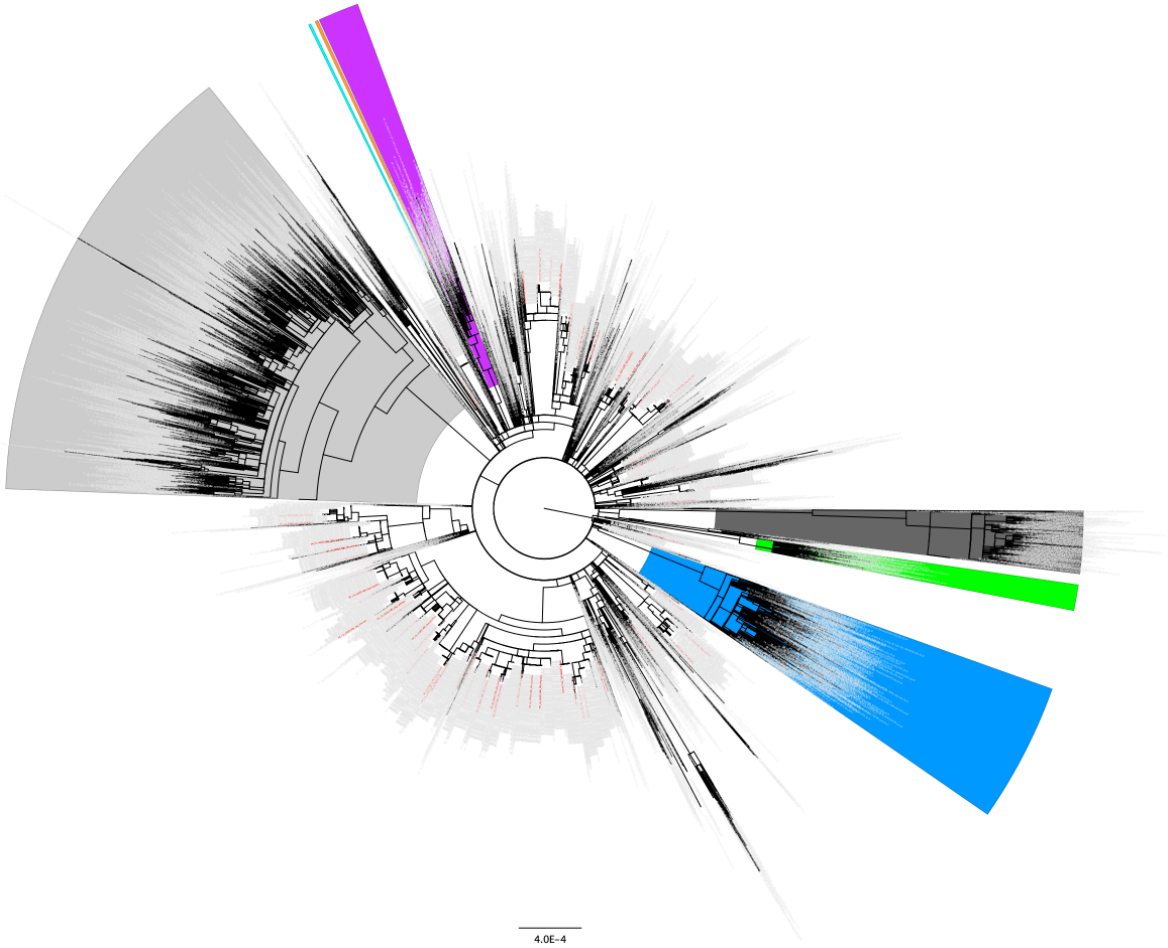

Figure 2. Distribution of the 68 samples containing Spike L452R as a minor variant from September 2020 across the global SARS-CoV-2 phylogeny indicating independent emergence. Background lineages include Alpha/B.1.17 samples highlighted in blue, Gamma/P.1 highlighted in green, Beta/B.1.351 highlighted in purple, Epsilon/B.1.429 highlighted in orange, Iota/B.1.526 highlighted in turquoise, Delta/B.1.617.2 highlighted in grey, and Omicron/BA.1/BA.2 highlighted in dark grey. None of the 68 samples containing Spike L452R as a minor variant in September 2020 were present in these highlighted lineages. Non VoC/VoI lineages are not highlighted. 68 samples identified as having the L452R change present as a minor variant in September 2020 (Table 2) are colored in red. 3,243 background genomes were included in this analysis.

3.3. *Comparison of VoC/VoI mutations*

The VoCs/VoIs that were analyzed (B.1.1.7/Alpha, B.1.351/Beta, P.1/Gamma, P.2/Zeta, B.1.429/Epsilon, B.1.526/Iota and B.1.617.2/Delta) constituted a total of 108 lineage specific mutations, some of which are present in two or more linages (for example, Spike N501Y in Alpha, Beta and Gamma). Of these 108 mutations evaluated with iSKIM between February 2020 and April 2021, n=15 mutations had a substantial rise (n>30 samples) as minor variants prior to fixation (Figure 3 and Figure S1), including the N501Y and L452R Spike changes. n=11 (73.3%) of these mutations were located on either the Spike (n=10) or Nucleocapsid (n=1) structural proteins and the remaining n=4 mutations were located on non-structural proteins (Figure 4). n=42 of the screened VoC mutations had candidate minor variants and fixed variants follow the same growth patterns, where both increase as VoIs/VoCs emerged. Interestingly, n=17 of the mutations had no

substantial growth of minor variants despite a rise in the number of fixed variants (Figure S1 and summarized in Table S3). Of those fixed mutations, n=11 (64.7%) were located on non-structural proteins and the others were located on Spike (n=3) and Nucleocapsid (n=3).

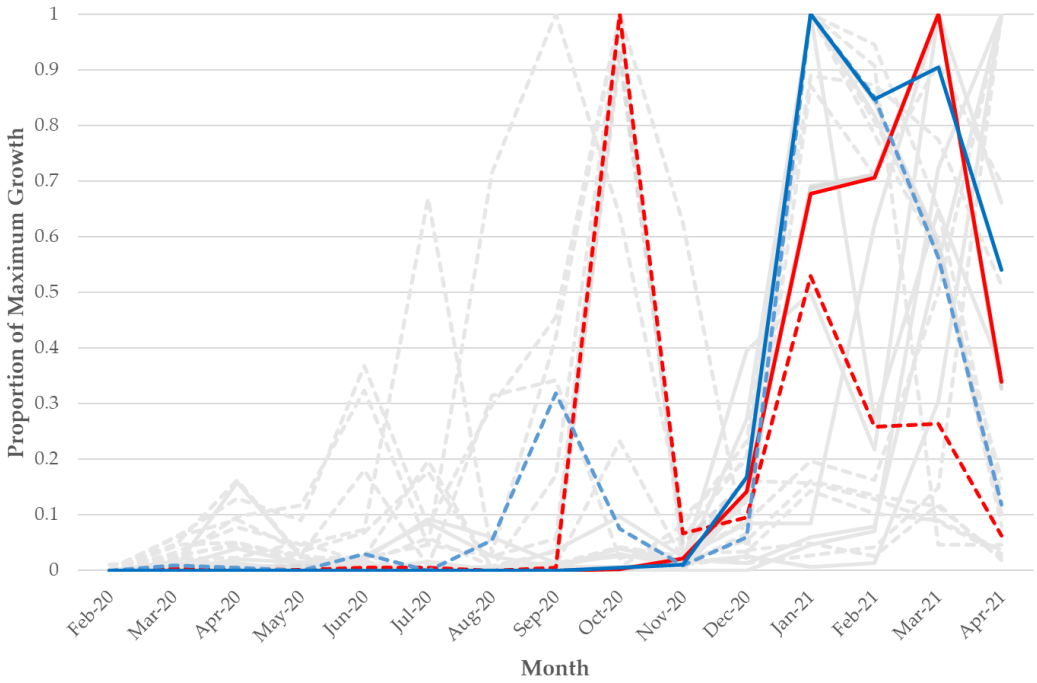

Figure 3. Frequency over time of the n=15 VoC/VoI mutations that had a substantial increase as a minor variant prior to a rise as a fixed variant across 411,805 NCBI SRA SARS-CoV-2 samples. The Y-axis is scaled by the maximum count for each particular mutation either as a minor variant or fixed mutation (whichever was higher for each). Dotted lines represent minor variant mutations and solid lines represent fixed mutations. The red solid and dotted lines represent the A23063T/N501Y mutation/change and the blue solid and dotted lines represent the T22917G/L452R mutation/change. The grey lines represent the other 13 VoC/VoI mutations that had a substantial increase as a minor variant prior to a rise as a fixed variant (each is also found in Figure S1).

Of the mutations that were screened for by iSKIM, 50.5% were located on the Spike protein. 36.4% were not located on any of the four structural proteins. However, 73.3% of the mutations that peaked as candidate minor variants prior to their fixed variants' peaks were located on the Spike protein (Figure 4) which was significant (one-proportion z-test, $p = 0.0387$). Of the mutations that had no substantial number of samples containing the mutation as minor variants despite a substantial rise in the number samples containing the mutation as fixed variants, 64.7% of the mutations were not located on structural proteins, which was also significant (one-proportion z-test, $p = 0.0765$). Mutations that are located on non-structural proteins are more likely to fall into the category of having no substantial rise in minor variant presence paired with a substantial rise in the number of fixed variants.

Mutations of Concern (Amino Acid Notation)	Protein Type	Protein Segment	Lineages					
			P.1/ Gamma	B.1.1.7/ Alpha	B.1.351/ Beta	B.1.429/ Epsilon	B.1.617.1/ Iota	B.1.617.2/ Delta
G5230T (K1655N)					X			
G17014T (D260Y)						X		
G17523T (M1352I)							X	
G21600T (S13I)						X		
G28881T (R203M)							X	X
C23271A (A570D)				X				
G22132T (R190S)			X					
T22917G (L452R)						X	X	X
A23063T (N501Y)			X	X	X			
G21974T (D138Y)			X					
C23604A (P681H)				X				
G23012C (E484Q)							X	
T24506G (S982A)				X				
G28048T (R52I)				X				
G24914C (D1118H)				X				

PROTEIN SEGMENT - LEGEND
- NTD
- RBD
- SARS-CoV-like_Spike_SD1-2_S1-S2_S2
- ORF 8
- NSP3
- Beta domain of Nsp13
- Stalk domain of Nsp13
- Nucleocapsid

PROTEIN TYPE LEGEND
- Spike
- Nucleocapsid
- Not a structural protein

Figure 4: n=15 VoC/VoI mutations that appeared as candidate minor variants prior to becoming fixed variants were mostly associated with the spike protein including on the NTD and RBD protein domains. 'X' denotes which lineage(s) each mutation is predominantly found in.

3.4. *Comparison of iSKIM to established minor variant detection software*

To evaluate whether our k-mer based method, iSKIM, produces results that can be repeated by a reference-based assembly method, the samples that iSKIM identified as having the N501Y and L452R changes surge as minor variants (Table 1 and Table 2) were separately run through the variant calling tool LoFreq and ngs_mapper's built in variant caller (basecaller.py). Analyzing the 68 samples identified by iSKIM as having L452R as a minor variant in the month of September 2020 (Table 2 and listed in Table S2), revealed that LoFreq did not identify any of these samples as having sufficient alternate nucleotides to be considered candidate minor variants nor fixed variants for the L452R change, while ngs_mapper identified all 68 samples as having a call frequency between 0 and 0.01. To standardize the comparison between iSKIM and ngs_mapper, the ratio of the T22917G mutation (L452R) to reference was used. The alternate nucleotide (G) to the reference nucleotide (T) [calculated as (# of G)/ (# of T)] was compared. The iSKIM ratio tended to be slightly higher than the ngs_mapper ratio (Figure S2). All ratios from both methods were close to 0.01 indicating a low frequency of the mutation in the samples.

834 samples from October 2020 were identified by iSKIM as having N501Y present as a minor variant (Table 1 and listed in Table S1). Lofreq identified 338 samples as possessing the mutation as a minor variant and 5 as a fixed variant. All 834 samples were run through ngs_mapper, and again, the ngs_mapper output had a lower ratio of mutation to reference nucleotide (A23063T for N501Y) for each sample compared to iSKIM (Figure S3, Figure S4, Figure S5). For the 343 samples that were identified as N501Y variants by all three methods (LoFreq, ngs_mapper, and iSKIM) and registered as either minor or fixed variants, two trends emerged. At higher ratios of mutation to reference nucleotides, when the mutation was fixed or biallelic, iSKIM's calculated ratio was greater than that of ngs_mapper and LoFreq (Figure S6). However, at lower ratios, when the mutation was a candidate minor variant, iSKIM's ratio was in between the calculated ratios for ngs_mapper and LoFreq (Figure S7 and Figure S8).

4. Discussion

The COVID-19 pandemic and response has led to an unprecedented amount of genomic data generation and sharing worldwide across publicly available databases. The amount of SARS-CoV-2 genomes and genomic datasets now represents over an order of magnitude greater data than any other previously studied virus [57–59]. This includes data across space and time to encompass various waves of the pandemic. In this study we sought to leverage the whole genome sequencing data that is publicly available in the NCBI SRA database to discover sample datasets that contain VoC defining mutations present as intrahost minor variants. Previous studies of SARS-CoV-2 intrahost variation have been performed at smaller scales due to the computational limitations of intrahost analysis [24–27]. To perform intrahost analysis at a much larger scale we took a different approach by generating short k-mer sequences encompassing VoC mutations that could be used to quickly scan the raw SARS-CoV-2 sequencing reads in the SRA database. Our k-mer based tool, iSKIM, allowed for the scan of over 400,000 raw genomic sequencing datasets totaling dozens of terabytes of raw data.

The analysis of these publicly available data at this scale revealed several patterns. We scanned for SARS-CoV-2 VoC/VoI mutations from the beginning of the pandemic through the emergence of Delta (February 2020 – April 2021). 108 total lineage specific mutations were screened and 15 of these mutations had a substantial increase as minor variants in samples detected one to five months prior to fixation. Based on our results, certain mutations appear in the population as minor variants a few months prior these mutations being seen as fixed mutations in larger numbers of samples. Of the 15 mutations identified with this pattern, 10 were located on the Spike protein, which was statistically significant. Conversely, 17 mutations had no substantial increase in the presence of minor variants despite a rise in the number of samples posessing these mutations as fixed variants. 11 (64.7%) of these mutations were located on non-structural proteins of SARS-CoV-2, which was also statistically significant. One possible explanation of this finding is that many of these latter mutations do not confer a fitness advantage to the virus, and are neutral mutations that emerged in lineages alongside advantagous mutations that then hitchhike to fixation.

A comparison of iSKIM to LoFreq and ngs_mapper was performed to confirm the accuracy of the iSKIM results. iSKIM consistently called the Spike L452R change at a slighlty higher frequency than ngs_mapper, while LoFreq did not call this as a minor variant intrahost mutation in the 68 samples from September 2020 detectected by iSKIM. This finding may be explained by the fact that LoFreq employs additional filtering steps that include accounting for strand-bias and high alignment error probability that are not taken into account by the reference-free approach of iSKIM. The minor variant intrahost results of the Spike N501Y change in the 834 samples from October 2020 were comparable across all three methods. In this instance, iSKIM called this intrahost mutation at a slightly higher frequency than ngs_mapper, but at a lower frequency than LoFreq. Therefore, if all 400,000+ NCBI SRA samples analyzed with iSKIM had been analyzed with Lofreq as well, it is possible additional samples containing these VoC mutations may have been identified. However, this is not currently computational feasible. Nonetheless, the results indicate that iSKIM results correspond well with results from established reference-assembly-based methods, ngs_mapper and LoFreq.

Many of the 834 samples from October 2020 that contained the N501Y Spike change as a minor intrahost variant belonged to the B.1.177 Pango lineage, as well as several from the B.1.36.28, B.1.36.17, B.1.221.1/B.1.221.2 lineages. Each of these lineages have been shown to have been involved in multiple recombination events during the emergence of the Alpha/B.1.1.7 VoC lineage, and none of the recombinant viruses contained a full complement of the Alpha/B.1.1.7 mutations [60]. This pattern is also observed in our analysis of the 834 samples from October 2020, where a subset of the Alpha/B.1.1.7 defining mutations (specifically N501Y), but not all lineage defining mutations, are

present as intrahost variants. This may be an important consideration when studying recombination in SARS-CoV-2 [61–63].

The large majority of the data analyzed in this study were generated using amplicon sequencing approaches which have been shown to be susceptible to producing varying levels of false primer induced mutations [64]. Primer trimming is a common bioinformatics step employed to remove these artifacts [65]. One shortcoming in the vast amount of SARS-CoV-2 data present in the NCBI SRA is the lack of sufficient metadata and details of the specific primer sets that were used for each run. While many NCBI SRA entries do include the sequencing strategy details that were used, for example, typical ARTIC protocols [22], these primer sets are updated regularly and primer sequences are not included with the sequencing submission. Therefore, it was not feasible in this study to primer trim each of the 400,000+ samples that were analzed. However, for the main findings of the L452R and N501Y changes, the popular ARTIC primer schemes were taken into account during our analyses as the 834 and 68 samples identified were generated with the ARTIC protocol. Neither of the mutations (T22917G/L452R or A23063T/N501Y) overlapped with ARTIC primers (see Methods). Therefore, these two intrahost mutations that we have highlighted were not impacted by primer induced mutations in the samples identified.

In this study we focused on SARS-CoV-2 Illumina sequencing data that was available in the NCBI SRA as this represented a very large amount of data. There is also a large amount of Oxford Nanopore SARS-CoV-2 sequencing data as well as other platforms such as PacBio. The error profiles of these longer read technologies differ from that of Illumina. Incorporating iSKIM support for these other data with varying error profiles should entale adjusting sevearl settings within iSKIM to account for differences between reads and k-mers while also providing sufficient sensitivity for detection. Similarly, this may also provide a way to account for k-mer erosion if additional SARS-CoV-2 mutations accumulate within the chosen k-mer sequences.

In addition to screening the large number of raw samples publicly available in the NCBI SRA, iSKIM can also be used to rapidly screen for newly sequenced samples that contain VoC or other mutations of interest prior to the more time-consuming and computationally expensive steps of reference-based genome assembly and curation. This can allow researchers to prioritize particular samples for reference-based genome assembly, other downstream analyses, or early reporting when turnaround time is critical. This has been particularly useful during periods of the pandemic when new VoCs are emerging, but not yet taken over as the dominant variant circulating. The iSKIM results also provide a complementary view of the data alongside typical consensus genome results.

Important putative and known mutations in the SARS-CoV-2 genome have been identified that may allow the virus to escape immune defenses [66–69]. Studies of patients with various forms of immunosuppression have revealed divergent SARS-CoV-2 virus sequences [70–72]. Additional mutations rarely observed in genome sequences sampled from clinical settings have been found in abunadnce in certain wastewater surveillance [73]. Animal reservoirs also pose a potential source of additional SARS-CoV-2 variation with capability for spillback into humans [74,75]. One additional way in which iSKIM could be applied would be to manually gather and curate this growing list of SARS-CoV-2 mutations not seen in previous or currently circulating VoC lineages. These mutations would then be added as sets of k-mers to iSKIM and could be used to screen all newly submitted raw sequencing datasets as a way to provide an early warning that known mutations may be emerging first as minor variant mutations. It still may be difficult to descern which of these mutations may be more important to pursue experimentally and which are less critical. However, this approach could provide a slightly earlier detection to when many of these mutations are soon then seen at the consensus level as is the current paradigm for early detection and warning.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Figure S1 – Plots of the number of minor and fixed variants for each mutation within each VoI/VoC over time identified by iSKIM across the 411,805 NCBI SRA runs screened.

Figure S2. The ratio of the alternate nucleotide to the reference nucleotide for each of the 68 samples from September 2020 containing Spike L452R as a minor variant identified by the iSKIM method (blue) is consistently slightly greater than that of the ngs_mapper method (orange).

Figure S3. The ratio of alternate nucleotide to the reference nucleotide for each SRA sample identified as having N501Y present as a minor variant by iSKIM. iSKIM has a higher ratio than ngs_mapper overall. The black line indicates where the one-to-one relationship falls.

Figure S4. Zoom in of the lower-left portion of Figure S3. The ratio of the alternative nucleotide to the reference nucleotide for N501Y candidate minor variant samples with iSKIM ratio less than 0.1. These lower ratios also indicate that the iSKIM method and ngs_mapper method give different results with iSKIM reporting a higher ratio than ngs_mapper overall. The black line indicates where the one-one-relationship falls.

Figure S5. The ratio of the alternate nucleotide to the reference nucleotide for each of the 834 samples from Octoboer 2020 containing Spike N501Y as a minor variant identified by the iSKIM method (blue) is consistently slightly greater than that of the ngs_mapper method (orange).

Figure S6. The ratio of the alternate nucleotide to the reference nucleotide for each N501Y sample identified by the iSKIM method compared with LoFreq (orange) and ngs_mapper (blue) calls. The black line indicates where the one-to-one relationship with iSKIM falls. iSKIM has a higher ratio than both LoFreq and ngs_mapper.

Figure S7. Zoom-in of the lower left portion of Figure S6. At low ratios, LoFreq has slightly higher ratios than iSKIM, while ngs_mapper has slightly lower ratios than iSKIM. The black line indicates where the one-to-one relationship with iSKIM falls.

Figure S8. The ratio of the alternate nucleotide to the reference nucleotide for selected N501Y candidate minor variant samples with iSKIM ratios less than 0.1. The ratios for the iSKIM method (blue) are consistently slightly greater than that of the ngs_mapper method (orange), but consistently less than that of the LoFreq method (gray).

Table S1 – Pango lineages of the 834 genomes identified as having the N501Y minor variant.

Table S2 – Pango lineages of the 68 genomes identified as having the L452R minor variant.

Table S3 – Categorization based on the pattern of minor and fixed variant frequencies over time for the 108 mutations screened via iSKIM across the 411,805 NCBI SRA SARS-CoV-2 samples.

Table S4 – Acknowledgments of GISAID Originating Laboratories, Submitting Laboratories, and Authors for the 834 genomes identified as having the N501Y minor variant.

Table S5 – Acknowledgments of GISAID Originating Laboratories, Submitting Laboratories, and Authors for the 68 genomes identified as having the L452R minor variant.

Table S6 – Acknowledgments of GISAID Originating Laboratories, Submitting Laboratories, and Authors for the 3,243 genomes used as background reference genomes for phylogenetic analysis.

File S1 – FASTA file of k-mer sequences for each of the following variant lineages: B.1.1.7, P.1, B.1.351, B.1.429, B.1.617.1, B.1.617.2, B.1.526, B.1.621, BA.1, BA.2.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, M.A.C, I.M.B. A.T..; methodology, M.A.C., A.T..; software, M.A.C., J.G., A.T.; validation, M.A.C., A.T., M.S.; formal analysis, M.A.C., A.T., M.S., C.K.F., I.M.B.; investigation, M.A.C., A.T., I.M.B., M.S., P.S.G.C.; resources, M.A.C., I.M.B, P.S.G.C.; data curation, M.A.C., A.T.; writing—original draft preparation, M.A.C.; writing—review and editing, A.T., M.A.C., M.S., J.G., C.K.F., P.S.G.C., I.M.B., M.A.C.; visualization, A.T., M.A.C., I.M.B.; supervision, M.A.C., P.S.G.C., I.M.B; project administration, M.A.C., I.M.B..; funding acquisition, M.A.C., I.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Department of Defense Global Emerging Infections Surveillance (GEIS) section of the Armed Forces Health Surveillance Division (AFHSD) ProMIS IDs P0169_21_WR and P0130_22_WR.

Disclaimer: Material has been reviewed by the authors' respective institutions. There is no objection to its presentation and/or publication. The View(s) expressed are those of the authors and do not necessarily reflect the official policy of the Departments of the Army, the Department of Defense, or the U.S. Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable.

Acknowledgments: We thank the entire community of researchers around the world working on SARS-CoV-2 for their tireless work producing and sharing the immense amount of data that was used in the analysis here. It is a testament to what can be accomplished when the scientific community works with free and open data sharing. We would like to acknowledge all the authors and originating and submitting laboratories of the sequences from GISAID's EpiCov that were used in our analyses. The full list of each of these is provide in Table S4, Table S5, Table S6.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Chiara, M.; D'Erchia, A.M.; Gissi, C.; Manzari, C.; Parisi, A.; Resta, N.; Zambelli, F.; Picardi, E.; Pavesi, G.; Horner, D.S.; et al. Next Generation Sequencing of SARS-CoV-2 Genomes: Challenges, Applications and Opportunities. *Brief. Bioinform.* **2021**, *22*, 616–630, doi:10.1093/bib/bbaa297.
2. 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.; et al. Spike Mutation D614G Alters SARS-CoV-2 Fitness. *Nature* **2020**, doi:10.1038/s41586-020-2895-3.
3. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812-827.e19, doi:10.1016/j.cell.2020.06.043.
4. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407, doi:10.1038/s41564-020-0770-5.
5. Tracking SARS-CoV-2 Variants Available online: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed on 22 June 2021).
6. du Plessis, L.; McCrone, J.T.; Zarebski, A.E.; Hill, V.; Ruis, C.; Gutierrez, B.; Raghwani, J.; Ashworth, J.; Colquhoun, R.; Connor, T.R.; et al. Establishment and Lineage Dynamics of the SARS-CoV-2 Epidemic in the UK. *Science (80-.).* **2021**, *371*, 708–712, doi:10.1126/science.abf2946.
7. Volz, E.; Mishra, S.; Chand, M.; Barrett, J.C.; Johnson, R.; Geidelberg, L.; Hinsley, W.R.; Laydon, D.J.; Dabrera, G.; O'Toole, Á.; et al. Assessing Transmissibility of SARS-CoV-2 Lineage B.1.1.7 in England. *Nature* **2021**, *593*, 266–269, doi:10.1038/s41586-021-03470-x.

8. Washington, N.L.; Gangavarapu, K.; Zeller, M.; Bolze, A.; Cirulli, E.T.; Schiabor Barrett, K.M.; Larsen, B.B.; Anderson, C.; White, S.; Cassens, T.; et al. Emergence and Rapid Transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **2021**, *184*, 2587-2594.e7, doi:10.1016/j.cell.2021.03.052.
9. Bayarri-Olmos, R.; Johnsen, L.B.; Idorn, M.; Reinert, L.S.; Rosbjerg, A.; Vang, S.; Hansen, C.B.; Helgstrand, C.; Bjelke, J.R.; Bak-Thomsen, T.; et al. The Alpha/b.1.1.7 Sars-Cov-2 Variant Exhibits Significantly Higher Affinity for Ace-2 and Requires Lower Inoculation Doses to Cause Disease in K18-Hace2 Mice. *Elife* **2021**, *10*, 1–14, doi:10.7554/eLife.70002.
10. Planas, D.; Bruel, T.; Grzelak, L.; Guivel-Benhassine, F.; Staropoli, I.; Porrot, F.; Planchais, C.; Buchrieser, J.; Rajah, M.M.; Bishop, E.; et al. Sensitivity of Infectious SARS-CoV-2 B.1.1.7 and B.1.351 Variants to Neutralizing Antibodies. *Nat. Med.* **2021**, *27*, 917–924, doi:10.1038/s41591-021-01318-5.
11. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Detection of a SARS-CoV-2 Variant of Concern in South Africa. *Nature* **2021**, *592*, 438–443, doi:10.1038/s41586-021-03402-9.
12. Voloch, C.M.; da Silva Francisco, R.J.; de Almeida, L.G.P.; Cardoso, C.C.; Brustolini, O.J.; Gerber, A.L.; de C. Guimarães, A.P.; Mariani, D.; Mirella da Costa, R.; Ferreira, O.C.J.; et al. Genomic Characterization of a Novel SARS-CoV-2. *J. Virol.* **2021**, *95*, doi:doi.org/10.1128/ JVI.00119-21.
13. Sabino, E.C.; Buss, L.F.; Carvalho, M.P.S.; Prete, C.A.; Crispim, M.A.E.; Fraiji, N.A.; Pereira, R.H.M.; Parag, K. V.; da Silva Peixoto, P.; Kraemer, M.U.G.; et al. Resurgence of COVID-19 in Manaus, Brazil, despite High Seroprevalence. *Lancet* **2021**, *397*, 452–455, doi:10.1016/S0140-6736(21)00183-5.
14. Cherian, S.; Potdar, V.; Jadhav, S.; Yadav, P.; Gupta, N.; Das, M.; Rakshit, P.; Singh, S.; Abraham, P.; Panda, S.; et al. SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganism* **2021**, *2*, 1–11, doi:https://doi.org/10.3390/microorganisms9071542.
15. Ferreira, I.; Datir, R.; Papa, G.; Kemp, S.; Meng, B.; Singh, S.; Pandey, R.; Ponnusamy, K.; Radhakrishnan, V.; Sato, K.; et al. SARS-CoV-2 B.1.617 Emergence and Sensitivity to Vaccine-Elicited Antibodies. *bioRxiv* **2021**, 2021.05.08.443253.
16. Viana, R.; Moyo, S.; Amoako, D.G.; Tegally, H.; Scheepers, C.; Althaus, C.L.; Anyaneji, U.J.; Bester, P.A.; Boni, M.F.; Chand, M.; et al. Rapid Epidemic Expansion of the SARS-CoV-2 Omicron Variant in Southern Africa. *Nature* **2022**, doi:10.1038/s41586-022-04411-y.
17. Mendelson, M.; Venter, F.; Moshabela, M.; Gray, G.; Blumberg, L.; de Oliveira, T.; Madhi, S.A. The Political Theatre of the UK's Travel Ban on South Africa. *Lancet* **2021**, *398*, 2211–2213, doi:10.1016/s0140-6736(21)02752-5.
18. Domingo, E.; Sheldon, J.; Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 159–216, doi:10.1128/mmbr.05023-11.
19. Domingo, E.; Perales, C. Viral Quasispecies. *PLoS Genet.* **2019**, *15*, 1–20, doi:10.1371/journal.pgen.1008271.
20. Grubaugh, N.; Gangavarapu, K.; Quick, J.; Matteson, N.; De Jesus, J.G.; Main, B.; Tan, A.; Paul, L.; Brackney, D.; Grewal, S.; et al. An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and IVar. *Genome Biol.* **2019**, *20*, doi:10.1186/s13059-018-1618-7.
21. Wilm, A.; Aw, P.P.K.; Bertrand, D.; Yeo, G.H.T.; Ong, S.H.; Wong, C.H.; Khor, C.C.; Petric, R.; Hibberd, M.L.; Nagarajan, N. LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets. *Nucleic Acids Res.* **2012**, *40*, 11189–11201, doi:10.1093/nar/gks918.
22. Tyson, J.R.; James, P.; Stoddart, D.; Sparks, N.; Wickenhagen, A.; Hall, G.; Choi, J.H.; Lapointe, H.; Kamelian, K.; Smith, A.D.; et al. Improvements to the ARTIC Multiplex PCR Method for SARS-CoV-2 Genome Sequencing Using Nanopore. *bioRxiv Prepr. Serv. Biol.* **2020**, doi:10.1101/2020.09.04.283077.
23. Li, T.; Chung, H.K.; Pireku, P.K.; Beitzel, B.F.; Sanborn, M.A.; Tang, C.Y.; Hammer, R.D.; Ritter, D.; Wan, X.; Berry, I.M.; et al. Rapid High-Throughput Whole-Genome Sequencing of SARS- CoV-2 by Using One-Step Reverse Transcription-PCR Ampli

Fi Cation with an Integrated Micro Fl Uidic System and Next-. *J. Clin. Microbiol.* **2021**, *59*.

24. Sapoval, N.; Mahmoud, M.; Jochum, M.D.; Liu, Y.; Leo Elworth, R.A.; Wang, Q.; Albin, D.; Ogilvie, H.A.; Lee, M.D.; Villapol, S.; et al. SARS-CoV-2 Genomic Diversity and the Implications for QRT-PCR Diagnostics and Transmission. *Genome Res.* **2021**, *31*, 635–644, doi:10.1101/GR.268961.120.
25. Armero, A.; Berthet, N.; Avarre, J.C. Intra-Host Diversity of Sars-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* **2021**, *13*, 1–15, doi:10.3390/v13010133.
26. Ko, S.H.; Mokhtari, E.B.; Mudvari, P.; Stein, S.; Stringham, C.D.; Wagner, D.; Ramelli, S.; Ramos-Benitez, M.J.; Strich, J.R.; Davey, R.T.; et al. High-Throughput, Single-Copy Sequencing Reveals SARS-CoV-2 Spike Variants Coincident with Mounting Humoral Immunity during Acute COVID-19. *PLoS Pathog.* **2021**, *17*, 1–20, doi:10.1371/journal.ppat.1009431.
27. Valesano, A.L.; Rumfelt, K.E.; Dimcheff, D.E.; Blair, C.N.; Fitzsimmons, W.J.; Petrie, J.G.; Martin, E.T.; Lauring, A.S. Temporal Dynamics of SARS-CoV-2 Mutation Accumulation within and across Infected Hosts. *PLoS Pathog.* **2021**, *17*, 1–15, doi:10.1371/journal.ppat.1009499.
28. Siqueira, J.D.; Goes, L.R.; Alves, B.M.; Carvalho, P.S. de; Cicala, C.; Arthos, J.; Viola, J.P.B.; de Melo, A.C.; Soares, M.A. SARS-CoV-2 Genomic Analyses in Cancer Patients Reveal Elevated Intrahost Genetic Diversity. *Virus Evol.* **2021**, *7*, 1–11, doi:10.1093/ve/veab013.
29. Rocheleau, L.; Laroche, G.; Fu, K.; Stewart, C.M.; Mohamud, A.O. Identification of a High-Frequency Intrahost SARS-CoV-2 Spike Variant with Enhanced Cytopathic and Fusogenic Effects. *MBio* **2021**, *13*, e00788-21, doi:doi.org/10.1128/mBio.00788-21.
30. Kille, B.; Liu, Y.; Sapoval, N.; Nute, M.; Rauchwerger, L.; Amato, N.; Treangen, T.J. Accelerating SARS-CoV-2 Low Frequency Variant Calling on Ultra Deep Sequencing Datasets. *2021 IEEE Int. Parallel Distrib. Process. Symp. Work. IPDPSW 2021 - conjunction with IEEE IPDPS 2021* **2021**, 204–208, doi:10.1109/IPDPSW52791.2021.00038.
31. Marçais, G.; Kingsford, C. A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers. *Bioinformatics* **2011**, *27*, 764–770, doi:10.1093/bioinformatics/btr011.
32. Melsted, P.; Pritchard, J.K. Efficient Counting of K-Mers in DNA Sequences Using a Bloom Filter. *BMC Bioinformatics* **2011**, *12*, doi:10.1186/1471-2105-12-333.
33. Marchet, C.; Boucher, C.; Puglisi, S.J.; Medvedev, P.; Salson, M.; Chikhi, R. Data Structures Based on K-Mers for Querying Large Collections of Sequencing Data Sets. *Genome Res.* **2021**, *31*, 1–12, doi:10.1101/gr.260604.119.
34. Chen, S.; He, C.; Li, Y.; Li, Z.; Iii, C.E.M. A Computational Toolset for Rapid Identification of SARS-CoV-2 , Other Viruses and Microorganisms from Sequencing Data. *Brief. Bioinform.* **2021**, *22*, 924–935, doi:10.1093/bib/bbaa231.
35. Tsueng, G.; Mullen, J.; Alkuzweny, M.; Cano, M.; Rush, B.; Haag, E.; Latif, A.A.; Zhou, X.; Qian, Z.; Andersen, K.G.; et al. Outbreak . Info Research Library : A Standardized , Searchable Platform to Discover and Explore COVID- 19 Resources and Data. *bioRxiv* **2022**, *2*, 1–19, doi:doi.org/10.1101/2022.01.20.477133.
36. Hodcroft, E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. Available online: https://covariants.org/.
37. Pickett, B.E.; Greer, D.S.; Zhang, Y.; Stewart, L.; Zhou, L.; Sun, G.; Gu, Z.; Kumar, S.; Zaremba, S.; Larsen, C.N.; et al. Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community. *Viruses* **2012**, *4*, 3209–3226, doi:10.3390/v4113209.
38. 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.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265–269, doi:10.1038/s41586-020-2008-3.
39. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, 1–4, doi:10.1093/gigascience/giab008.
40. Bonfield, J.K.; Marshall, J.; Danecek, P.; Li, H.; Ohan, V.; Whitwham, A.; Keane, T.; Davies, R.M. HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data. *Gigascience* **2021**, *10*, 1–6, doi:10.1093/gigascience/giab007.

41. Quinlan, A.R.; Hall, I.M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26*, 841–842, doi:10.1093/bioinformatics/btq033.
42. NCBI SRA FTP Available online: ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/byrun (accessed on 23 August 2018).
43. Leinonen, R.; Sugawara, H.; Shumway, M. The Sequence Read Archive. *Nucleic Acids Res.* **2011**, *39*, 2010–2012, doi:10.1093/nar/gkq1019.
44. Kodama, Y.; Shumway, M.; Leinonen, R. The Sequence Read Archive: Explosive Growth of Sequencing Data. *Nucleic Acids Res.* **2012**, *40*, 2011–2013, doi:10.1093/nar/gkr854.
45. Stoler, N.; Nekrutenko, A. Sequencing Error Profiles of Illumina Sequencing Instruments. *NAR Genomics Bioinforma.* **2021**, *3*, 1–9, doi:10.1093/nargab/lqab019.
46. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Brief. Bioinform.* **2013**, *14*, 178–192, doi:10.1093/bib/bbs017.
47. Ngs_mapper Available online: https://ngs-mapper.readthedocs.io/en/latest/ (accessed on 28 July 2022).
48. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410, doi:10.1016/S0022-2836(05)80360-2.
49. Hadfield, J.; Megill, C.; Bell, S.M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R.A. NextStrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* **2018**, *34*, 4121–4123, doi:10.1093/bioinformatics/bty407.
50. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.
51. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0009490.
52. Tang, J.W.; Tambyah, P.A.; Hui, D.S. Emergence of a New SARS-CoV-2 Variant in the UK. *J. Infect.* **2020**, *82*, E27–E28, doi:doi.org/10.1016/j.jinf.2020.12.024.
53. Leung, K.; Shum, M.H.H.; Leung, G.M.; Lam, T.T.Y.; Wu, J.T. Early Transmissibility Assessment of the N501Y Mutant Strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Eurosurveillance* **2020**, *26*, doi:10.2807/1560-7917.ES.2020.26.1.2002106.
54. O'Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J.T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evol.* **2021**, *7*, 1–9, doi:10.1093/ve/veab064.
55. Deng, X.; Garcia-Knight, M.A.; Khalid, M.M.; Servellita, V.; Wang, C.; Morris, M.K.; Sotomayor-González, A.; Glasner, D.R.; Reyes, K.R.; Gliwa, A.S.; et al. Transmission, Infectivity, and Neutralization of a Spike L452R SARS-CoV-2 Variant. *Cell* **2021**, *184*, 3426-3437.e8, doi:10.1016/j.cell.2021.04.025.
56. Motozono, C.; Toyoda, M.; Zahradnik, J.; Saito, A.; Nasser, H.; Tan, T.S.; Ngare, I.; Kimura, I.; Uriu, K.; Kosugi, Y.; et al. SARS-CoV-2 Spike L452R Variant Evades Cellular Immunity and Increases Infectivity. *Cell Host Microbe* **2021**, *29*, 1124-1136.e11, doi:10.1016/j.chom.2021.06.006.
57. Khare, S.; Gurry, C.; Freitas, L.; B Schultz, M.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; TC Lee, R.; Yeo, W.; et al. GISAID's Role in Pandemic Response. *China CDC Wkly.* **2021**, *3*, 1049–1051, doi:10.46234/ccdcw2021.255.
58. Elbe, S.; Buckland-Merrett, G. Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health. *Glob. Challenges* **2017**, *1*, 33–46, doi:10.1002/gch2.1018.
59. Shu, Y.; McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality. *Eurosurveillance* **2017**, *22*, 2–4, doi:10.2807/1560-7917.ES.2017.22.13.30494.
60. 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.; et al. Generation and Transmission of Interlineage Recombinants in the SARS-CoV-2 Pandemic. *Cell* **2021**, *184*, 5179-

5188.e8, doi:10.1016/j.cell.2021.08.014.

61. Ignatieva, A.; Hein, J.; Jenkins, P.A. Ongoing Recombination in SARS-CoV-2 Revealed through Genealogical Reconstruction. *Mol. Biol. Evol.* **2022**, *39*, 1–11, doi:10.1093/molbev/msac028.
62. Pollett, S.; Conte, M.A.; Sanborn, M.; Jarman, R.G.; Lidl, G.M.; Modjarrad, K.; Maljkovic Berry, I. A Comparative Recombination Analysis of Human Coronaviruses and Implications for the SARS-CoV-2 Pandemic. *Sci. Rep.* **2021**, *11*, 1–11, doi:10.1038/s41598-021-96626-8.
63. Bolze, A.; White, S.; Basler, T.; Rossi, A.D.; Greninger, A.L.; Hayashibara, K.; Wyman, D.; Dai, H.; Cassens, T.; Tsan, K.; et al. Evidence for SARS-CoV-2 Delta and Omicron Co-Infections and Recombination. *medRxiv* **2022**, 1–24, doi:https://doi.org/10.1101/2022.03.09.22272113.
64. Maio, N. De; Walker, C.; Borges, R.; Weilguny, L.; Slodkowicz, G.; Goldman, N. Issues with SARS-CoV-2 Sequencing Data Available online: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
65. Liu, T.; Chen, Z.; Chen, W.; Chen, X.; Hosseini, M.; Yang, Z.; Li, J.; Ho, D.; Turay, D.; Gheorghe, C.P.; et al. A Benchmarking Study of SARS-CoV-2 Whole-Genome Sequencing Protocols Using COVID-19 Patient Samples. *iScience* **2021**, *24*, 102892, doi:10.1016/j.isci.2021.102892.
66. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; COVID-19 Genomics UK (COG-UK) Consortium; et al. SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* **2021**, *614*, doi:10.1038/s41579-021-00573-0.
67. Greaney, A.J.; Loes, A.N.; Crawford, K.H.D.; Starr, T.N.; Malone, K.D.; Chu, H.Y.; Bloom, J.D. Comprehensive Mapping of Mutations in the SARS-CoV-2 Receptor-Binding Domain That Affect Recognition by Polyclonal Human Plasma Antibodies. *Cell Host Microbe* **2021**, *29*, 463-476.e6, doi:10.1016/j.chom.2021.02.003.
68. Cao, Y.; Yisimayi, A.; Jian, F.; Song, W.; Xiao, T.; Wang, L.; Du, S.; Wang, J.; Li, Q.; Chen, X.; et al. BA.2.12.1, BA.4 and BA.5 Escape Antibodies Elicited by Omicron Infection. *Nature* **2022**, doi:10.1038/s41586-022-04980-y.
69. Greaney, A.J.; Starr, T.N.; Bloom, J.D. An Antibody-Escape Estimator for Mutations to the SARS-CoV-2 Receptor-Binding Domain. *Virus Evol.* **2022**, *8*, 1–8, doi:10.1093/ve/veac021.
70. Corey, L.; Beyrer, C.; Cohen, M.S.; Michael, N.L.; Bedford, T.; Rolland, M. SARS-CoV-2 Variants in Patients with Immunosuppression. *N. Engl. J. Med.* **2021**, *385*, 562–566.
71. Clark, S.A.; Clark, L.E.; Pan, J.; Coscia, A.; McKay, L.G.A.; Shankar, S.; Johnson, R.I.; Brusic, V.; Choudhary, M.C.; Regan, J.; et al. SARS-CoV-2 Evolution in an Immunocompromised Host Reveals Shared Neutralization Escape Mechanisms. *Cell* **2021**, *184*, 2605-2617.e18, doi:10.1016/j.cell.2021.03.027.
72. Nussenblatt, V.; Roder, A.E.; Das, S.; de Wit, E.; Youn, J.-H.; Banakis, S.; Mushegian, A.; Mederos, C.; Wang, W.; Chung, M.; et al. Yearlong COVID-19 Infection Reveals Within-Host Evolution of SARS-CoV-2 in a Patient With B-Cell Depletion. *J. Infect. Dis.* **2022**, *225*, 1118–1123, doi:10.1093/infdis/jiab622.
73. Smyth, D.S.; Trujillo, M.; Gregory, D.A.; Cheung, K.; Gao, A.; Graham, M.; Guan, Y.; Guldenpfennig, C.; Hoxie, I.; Kannoly, S.; et al. Tracking Cryptic SARS-CoV-2 Lineages Detected in NYC Wastewater. *Nat. Commun.* **2022**, *13*, 1–9, doi:10.1038/s41467-022-28246-3.
74. Hale, V.L.; Dennis, P.M.; McBride, D.S.; Nolting, J.M.; Madden, C.; Huey, D.; Ehrlich, M.; Grieser, J.; Winston, J.; Lombardi, D.; et al. SARS-CoV-2 Infection in Free-Ranging White-Tailed Deer. *Nature* **2022**, *602*, 481–486, doi:10.1038/s41586-021-04353-x.
75. Pickering, B.; Lung, O.; Maguire, F.; Kruczkiewicz, P.; Marchand-austin, A.; Massé, A.; Mcclinchey, H.; Aftanas, P.; Blais-savoie, J.; Chee, H.; et al. Highly Divergent White-Tailed Deer SARS-CoV-2 with Potential Deer-to-Human Transmission Abstract Wildlife Reservoirs of SARS-CoV-2 May Enable Viral Adaptation and Spillback from Animals to Humans . In North America , There Is Evidence of Unsustained Spill. *bioRxiv* **2022**.

719