viruses

MDPI

Article

A Pre-Vaccination Baseline of SARS-CoV-2 Genetic Surveillance and Diversity in the United States

Adam A. Capoferri [1,2,*], Wei Shao [3], Jon Spindler [1], John M. Coffin [4], Jason W. Rausch [1] and Mary F. Kearney [1]

[1] HIV Dynamics and Replication Program, Center for Cancer Research, NCI-Frederick, Frederick, MD 21702, USA; jspindler@mail.nih.gov (J.S.); rauschj@mail.nih.gov (J.W.R.); kearneym@mail.nih.gov (M.F.K.)
[2] Department of Microbiology and Immunology, Georgetown University, Washington, DC 20007, USA
[3] Advanced Biomedical Computing Science, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA; shaow@mail.nih.gov
[4] Department of Molecular Biology and Microbiology, Tufts University, Boston, MA 02129, USA; john.coffin@tufts.edu
* Correspondence: adam.capoferri@nih.gov or ajl160@georgetown.edu

Abstract: COVID-19 vaccines were first administered on 15 December 2020, marking an important transition point for the spread of SARS-CoV-2 in the United States (U.S.). Prior to this point in time, the virus spread to an almost completely immunologically naïve population, whereas subsequently, vaccine-induced immune pressure and prior infections might be expected to influence viral evolution. Accordingly, we conducted a study to characterize the spread of SARS-CoV-2 in the U.S. pre-vaccination, investigate the depth and uniformity of genetic surveillance during this period, and measure and otherwise characterize changing viral genetic diversity, including by comparison with more recently emergent variants of concern (VOCs). In 2020, SARS-CoV-2 spread across the U.S. in three phases distinguishable by peaks in the numbers of infections and shifting geographical distributions. Virus was genetically sampled during this period at an overall rate of ~1.2%, though there was a substantial mismatch between case rates and genetic sampling nationwide. Viral genetic diversity tripled over this period but remained low in comparison to other widespread RNA virus pathogens, and although 54 amino acid changes were detected at frequencies exceeding 5%, linkage among them was not observed. Based on our collective observations, our analysis supports a targeted strategy for worldwide genetic surveillance as perhaps the most sensitive and efficient means of detecting new VOCs.

Keywords: SARS-CoV-2; COVID-19; SARS-CoV-2 evolution; SARS-CoV-2 in United States; variants of concern; viral evolution

Citation: Capoferri, A.A.; Shao, W.; Spindler, J.; Coffin, J.M.; Rausch, J.W.; Kearney, M.F. A Pre-Vaccination Baseline of SARS-CoV-2 Genetic Surveillance and Diversity in the United States. *Viruses* **2022**, *14*, 104. https://doi.org/10.3390/v14010104

Academic Editor: Corinne Ronfort

Received: 17 December 2021
Accepted: 4 January 2022
Published: 7 January 2022

1. Introduction

The first reported case of COVID-19 in the U.S. was on January 20th, 2020 [1]. Since that time, through to the beginning of January 2022, there have been more than 57 million U.S. cases (19.3% of global) and 828,000 deaths (15.2% of global) [2]. By late 2020, several vaccines had been developed, tested, and approved, and since Dec 15, 2020, have been widely administered in the U.S. and the world. The end of 2020 was also marked by the emergence of SARS-CoV-2 variants with greater transmissibility, virulence, and partial resistance to current preventives and treatments [3,4]. Though they share some common genetic features, such 'variants of concern' (VOCs) appear to have emerged independently in different regions throughout the world [5,6]. Among these, the Delta VOC, which began to spread rapidly in the U.S. in early 2021, became the dominant form in the U.S until the emergence and spread of the Omicron VOC. In this work, we provide a detailed characterization of the early spread of SARS-CoV-2 in three phases across the different geographic regions of the U.S. and assess the varying levels of genetic surveillance and

genetic diversity in each region and phase. Assessing the spread and genetics of SARS-CoV-2 in the U.S. in 2020, largely in the absence of immune selection pressure from prior infection or vaccine administration, will allow us and others to contrast these patterns with the surveillance and genetics of the virus in the future, with emphasis on improving the early detection of new variants.

2. Materials and Methods

2.1. Sources for COVID-19 Cases and Deaths in the U.S. in 2020

To characterize the early spread of SARS-CoV-2 across different regions in the U.S., we extracted data for U.S. COVID-19 cases and deaths between 1 January 2020 and 15 December 2020 from https://COVIDtracking.com/data (accessed 17 December 2020) [7] or https://COVID.cdc.gov (accessed 5 January 2022) [2]. U.S. regions were assigned based on the four Census Regions of the United States (U.S. Census Bureau) [8]. Viral spread in the U.S. was divided into three phases based on COVID-19 case peaks where the derivative of the trough was approximately zero: Phase 1 of winter–spring (1 January 2020–31 May 2020), Phase 2 of summer (1 June 2020–31 August 2020), and Phase 3 of fall (1 September 2020–15 December 2020). The estimated 2019 population for each U.S. state was accessed by the U.S. Census Bureau (Supplementary Materials Table S1) to normalize the incidence of COVID-19 cases and deaths in the sub-regional areas. GraphPad Prism V.8.4.3 was used to visualize the data.

2.2. Sources and Numbers of SARS-CoV-2 Sequences Used in Surveillance and Diversity Analyses

A total of 36,299 full-length (29,782 bp), high-coverage SARS-CoV-2 genomes from humans in the U.S. with infections between 20 January 2020 and 15 December 2020 were obtained from gisaid.org (accessed on 18 December 2020) [9,10] for surveillance and sequence analysis. The numbers of sequences obtained were: Phase 1 (22,434), Phase 2 (11,893), and Phase 3 (2072). All sequences used in the analyses can be found in the Supplemental Dataset. To minimize miscounting of mutations in our sequence alignments, we replaced rare instances of the standard ambiguous base designation 'N', which would be considered a match with any nucleotide (i.e., A, T, C, G), and thus artificially reduce the apparent mutation rate, with gaps, so that no comparison with the reference at these positions would be made. SARS-CoV-2 Clade O, Clade GV, Cruise-ship, and other U.S. territory sequences were excluded from the analysis due to the limited number of sequences. Sequences from VOCs: Alpha (Pango lineage B.1.1.7, Nextstrain 20I, GISAID clade GR, originally isolated in the U.K.), Beta (B.1.351, 20H, GH, South Africa), Gamma (P.1, 20J, GR, Brazil), and Epsilon (B.1.427+B.1.429, 21C, GH, United States) were accessed from gisaid.org on 1 April 2021; and Delta (B.1.617.2, 21A, GK, India) were accessed from gisaid.org on 27 July 2021 (see also Section 2.5) [9–11]. The VOCs (Alpha, Beta, Gamma, and Epsilon) were sampled in the U.S. between 1 November 2020 and 31 March 2021 because of their earlier introduction or emergence in the U.S compared to Delta. All VOC sequences available at the time were accessed resulting in a dataset with 37 Alpha sequences, 38 Beta, 26 Gamma, and 31 Epsilon VOC. An additional 15,733 Delta sequences were added to the final dataset corresponding to the later detection of the Delta VOC. Gap-stripped alignments were generated using the FFT-NS-1 200PAM/k = 2 algorithm of MAFFT v7.450 [12,13]. Additional analyses and data handling of SARS-CoV-2 sequences were conducted using Geneious Prime® 2020.2.4 and several in-house-generated Perl scripts available at https://github.com/Wei-Shao/COV2-Analysis(accessed 17 December 2020).

2.3. Analysis of SARS-CoV-2 Genetic Surveillance in the U.S. in 2020

To assess the relationship between the case distribution and genetic surveillance of SARS-CoV-2 in the U.S., genomic sequences from each phase were separated by region and clade for each month in 2020. To estimate the number of COVID-19 cases within each GISAID clade, the number of sequences was multiplied by the total monthly new

COVID-19 cases. RStudio v1.3 [14] and GraphPad Prism V.8.4.3 were used to tabulate data and for data visualization.

The level of SARS-CoV-2 genetic surveillance in the U.S. in 2020 was also compared to levels in other developed nations in the same time period (U.K. and in Australia). Sequence data from the U.K. and Australia were obtained under the same selection criteria as for the U.S. The number of cases was accessed on the same days using https://coronavirus.data.gov.uk/ (accessed 17 December 2020) [15] and the National Notifiable Diseases Surveillance System (http://www9.health.gov.au/cda/source/rpt_3.cfm, accessed 17 December 2020) [16]. The level of sequencing was determined from the number of sequences obtained monthly and the number of monthly cases of COVID-19 using an in-house generated bioinformatic script to visualize data (script available at https://github.com/aacapoferri/COV2, accessed 18 December 2020).

2.4. Mutation Detection and Measurements of Genetic Diversity and Divergence

Sequences obtained as described above were used to detect the emergence of new SARS-CoV-2 mutations and to assess the clade genetic diversity and divergence across the 3 phases of spread in the U.S. in 2020. All sequences for each clade and phase were included in each analysis with the exception of clade GH, where the number of sequences was too high for measurements of genetic diversity and divergence and, therefore, 2500 sequences were randomly subsampled.

SARS-CoV-2 mutation frequencies were determined for each clade/phase compared to either the majority-rule of Phase 1 consensus sequence, the Wuhan-Hu-1 reference genome (GenBank accession, NC_045512.2), or the VOCs [17]. To exclude clade-defining mutations and amplification/sequencing errors, several steps were taken. First, majority-rule consensus sequences for each clade were generated from all genomes in Phase 1 and used as a reference for detecting new mutations that emerged in Phases 2 and 3. This approach allowed majority clade-associated mutations to be omitted for the detection of new mutations only. Second, a threshold $\geq$5% frequency was used to eliminate mutations that were rarely detected and, therefore, could be PCR errors, sequencing errors, or real but not determined to be sustained in the population with at least 95% confidence (see Section 3.4) [18]. Our approach to setting a threshold was as follows: first, the 99th percentile for all clades (G/GH/GR/L/S/V) across Phases 1, 2, and 3 for all non-zero mutation frequencies at each nucleotide position was calculated (average of 3.84%); second, to account for error mutations, the upper–outer fence [Q3 + 1.5IQR] of the 99th percentiles was calculated (6.09%); finally, to resolve this range, the median was rounded to the nearest whole percent (5%). Mutation frequencies were plotted and annotated using the "Mutation frequency for SARS.R" script (an example provided on https://github.com/aacapoferri/COV2, accessed 17 December 2020). Mutations that were present in any phase at $\geq$5% of the population were noted for each G-based clade and in each phase. Heatmaps were generated using GraphPad Prism V.8.4.3 to visualize the persistence and emergence of mutations that were present in at least one phase in greater than $\geq$5% of the surveyed populations.

Mutation distributions were determined by assessing the number of mutations per sequence for each clade during each phase by Hamming distance. To examine the number of mutations per sequence in the G-based clades during each phase, the distribution of the number of mutations relative to the Wuhan-Hu-1 reference genome was generated using an in-house script available at https://github.com/Wei-Shao/COV2-Analysis (accessed 17 December 2020).

Statistical shifts in population structure (divergence) were determined using a test for panmixia with a statistical cut-off at $p < 10^{-3}$ [19]. Population genetic diversity was calculated as average pair-wise distance (APD) in MEGAX for each clade/phase [20]. APD was determined using p-distance and included transitions/transversions with rates among sites as uniform where gaps/missing data were treated as a complete deletion. All sequences were used for each clade/phase for calculating APD, except in clade GH during

Phases 1 and 2, where random subsampling of 2500 sequences was performed. When calculating the APD for each clade per month, 50 sequences were randomly subsampled in triplicate. Where there were fewer than 50 sequences available for a given clade/month, all sequence were utilized unless there were fewer than 10, in which case that clade/month was excluded from the analysis.

Identical SARS-CoV-2 genomes were collapsed to determine the number of different variants in the dataset. A simple linear regression was determined for clades G, GH, and GR in GraphPad Prism V.8.4.3 with the linear equation and goodness-of-fit (R^2) reported. The slope was understood as the rate of change in %APD/month. The length of the SARS-CoV-2 genome is ~30,000 base pair, which when multiplied by the slope, gave an approximate number of nucleotide changes/month for a given G-based clade.

To examine mutations in mapped epitope sites, majority-rule consensus sequences were generated for the G-based clades in each phase and were aligned to the Wuhan-Hu-1 reference genome. For comparison and fine-mapping, previously described T-cell and B-cell epitopes from Spike and Nucleocapsid are listed and referenced in Table S2. Nonsynonymous mutations observed in the G-based clades that differed from the reference in either T-cell or B-cell epitopes were noted. To distinguish between purifying and positive selection, the dN/dS ratio was calculated for each codon position in Spike and Nucleocapsid using an in-house Perl script at https://github.com/Wei-Shao/COV2-Analysis (accessed 17 December 2020).

2.5. Analyses of Variants of Concern (VOCs)

The five VOCs used in this study included the Alpha, Beta, Gamma, Delta, and Epsilon sampled within the U.S. The Delta VOC was assigned as GISAID clade GK (defined after 2020). The numbers of mutations per sequence were determined for each VOC and compared to the corresponding values calculated for the GISAID clades during Phase 3. The APD for each VOC dataset was calculated and compared to the APDs of each G-based clade in the U.S. during 2020. Because the SARS-CoV-2 Spike protein is particularly important for vaccine and therapeutic strategies, we searched for and characterized VOC-defining S gene mutations in the U.S. viral population even when their frequencies did not meet our 5% threshold.

2.6. Phylogenetic Analysis

We reconstructed a phylogenetic tree using a random subsample of 300 sequences for each clade (G/GH/GR) and sequences from the circulating VOCs, as described in the dataset. Sequencing depth was an important consideration due to the random subsampling and the relatively small sequence dataset for several VOCs. From our analysis in Figure S5, we found that we could detect mutations present at frequencies $\geq$1% with 95% probability using datasets of 300 sequences. An alignment of 1,084 sequences was generated with MAFFT v7.450 (FFT-NS-1 200PAM/k = 2 algorithm) [12,13]. The final alignment was 28,744 bp in length after ends were trimmed. A maximum-likelihood phylogeny was estimated with RAxML-NG v.1.0.0 [21,22] using the GTR+I+G4 substitution model after optimizing with ModelTest-NG v0.1.7 [23] in raxmlGUI 2.0 [24]. Optimized model parameters were as follows: $-$lnL = 101,634.84, AIC_C = 215,879.42, freq[A, C, G, T] = [0.30, 0.17, 0.18, 0.35], R[A-C, A-G, A-T, C-G, C-T, G-T] = [0.16, 0.73, 0.13, 0.11, 2.47, 1.00], Among-site rate variation by proportion of invariable sites (I) = 0.63, and Variable sites (G) with Gamma distribution shape parameter = 1.01. The outgroup was set to the Wuhan-Hu-1 reference genome. Tree visualization was performed in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/, accessed 17 December 2020).

2.7. Viral Genetic Surveillance Resources

Several databases were used to compare and contrast global trends and our U.S.-based specific analysis. These included: PANGO lineages (https://cov-lineages.org/, accessed 17 December 2020) [25], NextStrain (https://nextstrain.org/sars-cov-2/, accessed 17 Decem-

ber 2020) [26], Global Initiative on Sharing Avian Influenza Data (https://www.gisaid.org/, accessed 17 December 2020) [9,10], Outbreak.info (https://outbreak.info/situation-reports, accessed 17 December 2020) [27], Observable (https://observablehq.com/@spond/linkage-disequilibirum-in-sars-cov-2, accessed 17 December 2020), Virological forum (https://virological.org/, accessed 17 December 2020), Los Alamos National Laboratory (https://cov.lanl.gov/content/index, accessed 17 December 2020), and the U.S. Centers for Disease Control and Prevention (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html, accessed 17 December 2020).

3. Results and Discussion

3.1. Spread of SARS-CoV-2 in the U.S. Pre-Vaccination Did Not Coorelate with Geographic or Levels of Sequence Surveillance

To characterize the early spread of SARS-CoV-2 in the U.S., we compared the cases and deaths across each geographic region (Figure 1). The spread of SARS-CoV-2 in the U.S. in 2020 occurred in three phases marked by peaks in case numbers, hospitalizations, and deaths (Figure 1A,B). Phase 1, in the winter and spring of 2020, began with the introduction of SARS-CoV-2 from Europe and Asia [28] and was followed by a surge in cases resulting from community spread and mobility [29], especially in the northeast (Figure 1C–E) [30–33]. Phase 2 began in early June with accelerated community spread primarily in the south and west after mitigation policies were relaxed (Figure 1D,E). The start of Phase 3 in the fall of 2020 was marked by a surge in transmission in the Midwest (Figure 1D,E), followed by a nationwide increase, at or near the end of which time public vaccination was initiated (Figure S1 and Table S1).

The regional distribution of COVID-19 cases varied by phase and was not always correlated with the level of viral sequencing in the different regions (Figure 1G,H). For example, although the south had the greatest overall number of cases (Figure 1F), the majority of SARS-CoV-2 sequences were obtained from samples collected in the west (Figure 1G,H), an observation made by a previous group as well [34]. Multiple factors contributed to this disproportionate sequencing including the failure to initiate a national genetic surveillance plan early in the pandemic, limited funding for sequencing, and limited access to donor samples in some regions of the country. In total, viral sequences were obtained from 1.2% of reported U.S. cases in 2020, a low level compared to other developed nations such as the U.K. (8.1%) and Australia (6.2%) (Figure 1I,J). The aggregate rate of sequencing in the U.S. reflects a decrease from 8.4% in Phase 1 to 0.3% in Phase 3, a difference that can be partly explained by the long intervals between sample collection and sequence deposition in GISAID (median: ~100 days; Figure 1H,I). Since vaccines were first administered in late 2020, the rates of SARS-CoV-2 sequencing in the U.S. have increased significantly. Though this development is certainly encouraging with respect to early detection of emerging variants and increased efforts to better match case and sequencing geographical distributions, more targeted approaches are still needed to detect the evolution of new VOCs.

3.2. SARS-CoV-2 Genetic Diversity and Divergence Increased in the U.S. in 2020

To characterize the increasing genetic diversity and divergence of SARS-CoV-2 in the U.S. prior to both the introduction of vaccines and detection of the first VOC, high-coverage full-length genome sequences with GISAID submission dates on or before 15 December 2020 were analyzed. All early GISAID-assigned clades of SARS-CoV-2 (G/GH/GR/S/L/V) were identified in the U.S. in Phase 1 (Figure 2A). However, the G-based clades (G/GH/GR), defined by the D614G mutation in the Spike (S) gene [35,36], accounted for >99% of sequences by Phase 2 (Figure S2). SARS-CoV-2 variants with the D614G mutation have been shown to be more infectious and exhibit some degree of resistance to certain monoclonal antibodies [37], yet they maintain convalescent serum neutralization sensitivity [38] and do not appear to worsen clinical outcomes (more on VOCs below) [39].

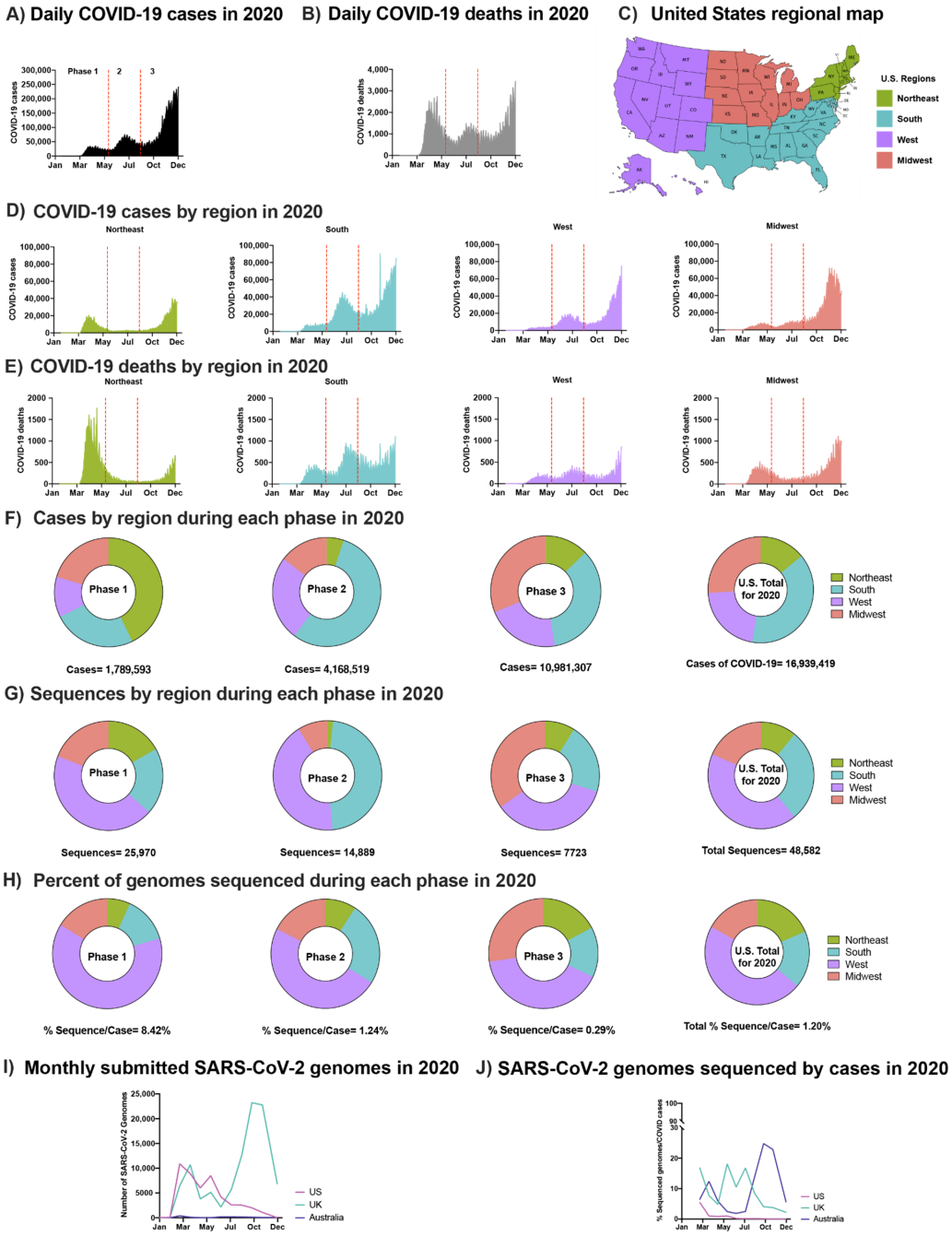

Figure 1. SARS-CoV-2 epidemic in the U.S. in 2020: (**A**) Daily COVID-19 cases in the U.S. in 2020. (**B**) Daily COVID-19 deaths in the U.S. in 2020. (**C**) U.S. regional map colored by region. (**D**) Number of COVID-19 cases in the U.S. in 2020 by region: Northeast, South, West, Midwest, respectively. (**E**) Number of COVID-19 deaths in the U.S. in 2020 by region. (**A**,**B**) and (**D**,**E**) Separation of phases is denoted by vertical dotted red lines. Data were smoothed by a moving 3-day average. (**F**) Proportion of COVID-19 cases by region during each phase and the overall contribution to the U.S. total in 2020. (**G**) Proportion of SARS-CoV-2 sequences accessed (submission as of 15 December 2020) by region during each phase and the overall contribution to the U.S. total in 2020 (**H**) The number of sequences per case were obtained by each region during each phase and the U.S. total in 2020. (**F**–**H**) Highlights Phases 1, 2, and 3, followed with U.S. total of 2020. (**I**) Total number of sequences submitted to GISAID from the U.K., Australia, and the U.S. by 15 December 2020. (**J**) Submitted SARS-CoV-2 genomes normalized to the number of COVID-19 cases from the U.K., Australia, and the U.S. (see Section 2.3).

A)

Clade	Number of Sequences			Panmixia		%APD			%APD Fold-increase			%Different SARS-CoV-2 Variants		
	Phase 1	Phase 2	Phase 3	Phase 1→2	Phase 2→3	Phase 1	Phase 2	Phase 3	Phase 1→2	Phase 2→3	Phase 1→3	Phase 1	Phase 2	Phase 3
G	2274	2834	518	$p<10^{-6}$	$p<10^{-6}$	0.02	0.04	0.06	1.72	1.32	2.27	56	63	70
GH	14,707*	6299*	1314	$p<10^{-6}$	$p<10^{-6}$	0.02	0.05	0.06	2.50	1.20	3.00	67	75	56
GR	1648	2571	240	$p<10^{-6}$	$p<10^{-6}$	0.02	0.04	0.06	1.89	1.48	2.79	58	66	75
S	2900	85	#ND	$p<10^{-6}$	#ND	0.03	0.03	#ND	1.01	#ND	#ND	55	51	#ND
L	577	4	#ND	0	#ND	0.01	0.00	#ND	0.27	#ND	#ND	58	75	#ND
V	328	#ND	#ND	#ND	#ND	0.02	#ND	#ND	#ND	#ND	#ND	67	#ND	#ND

(#ND) Not enough sequences for proper analysis; (*) 2,500 sequences were randomly selected for Panmixia, %APD, and %Different SARS-CoV-2 Variants.

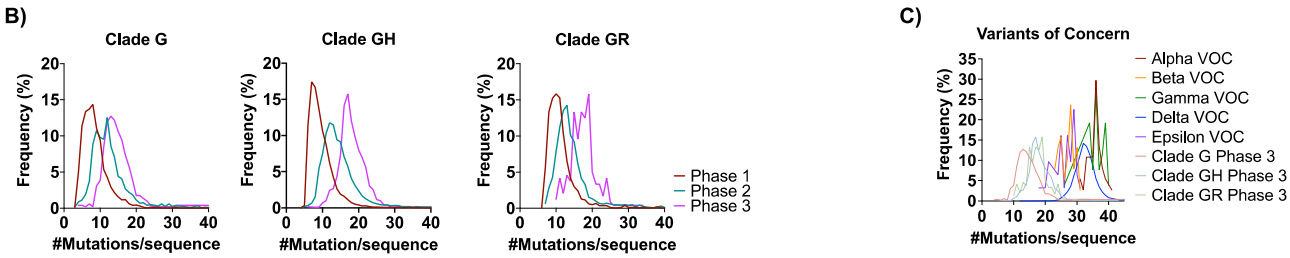

Figure 2. SARS-CoV-2 genetic diversity increased over time: (**A**) The number of sequences obtained from GISAID (https://www.gisaid.org/, accessed 17 December 2020) in the analysis for each clade by phase are reported. Genetic divergence was measured by panmixia probability [40] (significance cutoff, $p < 10^{-3}$) for clade/Phase with >11 sequences for Phase 1 → 2 and Phase 2 → 3. Genetic diversity was measured by average pair-wise distance (%APD) and the percent of different SARS-CoV-2 variants. (**B**) The distribution of the frequency for the number of mutations per sequence relative to the Wuhan-Hu-1 isolate was determined for the G-based clades in each Phase. (**C**) Distribution of the frequency for the number of mutations per sequence for the VOCs: Alpha, Beta, Gamma, Delta, and Epsilon. The number of mutations/sequence (mut/seq) at the maximum peak frequency were: Clade G (13 mut/seq), Clade GH (17 mut/seq), Clade GR (19 mut/seq), Alpha VOC (36 mut/seq), Beta (28 mut/seq), Gamma (36 mut/seq), Delta (32 mut/seq), and Epsilon (29 mut/seq). (# ND) Not enough sequences for proper analysis; (*) 2500 sequences were randomly selected for Panmixia, %APD, and %Different SARS-CoV-2 Variants.

The aggregate average pair-wise distance (APD) among the G-based clades increased from 0.02% in Phase 1 to 0.06% in Phase 3, reflective of 2.3-, 3.0-, and 2.8-fold increases for clades G, GH, and GR, respectively (Figure 2A and Figure S3). These rates translate to 1.95-nt/month (clade G), 2.85-nt/month (clade GH), and 2.22-nt/month (clade GR) when expressed as the average numbers of changes observed in the viral genome each month (Figure S3A). For comparison, the overall APD measured for VOC lineages that emerged in 2021 (i.e., the Alpha, Beta, Gamma, Delta, and Epsilon) was between 0.02% and 0.09% (Figure S3B). As these genetic distances were comparable to those measured in the G-based clades in the U.S. in 2020, there is no reason to suspect that the VOCs will naturally accrue mutations more rapidly than non-variant lineages as the pandemic continues.

The observed increase in genetic diversity of SARS-CoV-2 in the U.S. in 2020 was due to an increase in the number of unique variants comprising clade G (+14%) and clade GR (+17%) (Figure 2A). However, despite the 3-fold increase in APD, clade GH had an overall decrease in the number of different variants from Phase 1 to Phase 3 (−11%) (Figure 2A). This finding suggests that the increase in genetic diversity in the GH clade resulted not from an increase in the number of variants but rather from the spread of fewer variants that were more divergent. This finding is further supported by observations of the numbers of mutations observed per sequence over time for each clade (Figure 2B). Specifically, whereas in Phase 1, clades G, GH, and GR averaged 7, 7, and 10 mutations/sequence, respectively, these frequencies increased by 1.7-, 2.4-, and 1.8-fold in Phase 3, another indication of a disproportionate increase in GH clade divergence. Similarly, panmixia, a metric indicating the degree to which random populations remain unstructured (i.e., non-divergent) over time [40], was calculated for the respective clades. We found there was

significant divergence in both the G-based and S clades from Phase 1 to Phase 2 ($p < 10^{-6}$), demonstrating that there was not random unstructured population mixing and that viral evolution was directional (suggesting selective pressures for some mutations) (Figure 2A).

Parallel analysis of the Alpha, Beta, Gamma, Delta, and Epsilon VOCs showed that these variants contained 26–36 mutations/sequences relative to their clade consensus, nearly twice the average observed among the G clades in 2020 (Figure 2B,C). These data suggest that the VOCs emerged via an atypical evolutionary pathway in which key mutations were acquired together within individuals (humans and/or other hosts) rather than only sequentially over the course of multiple person-to-person transmissions, further emphasizing the need for more targeted surveillance approaches.

3.3. Phylogenetic Analysis Reveals the Evolutionary Relationships of G Clades in the U.S. in 2020 to the VOCs

To determine the phylogenetic relationships of the G-based clades and VOCs in the U.S., we reconstructed a maximum-likelihood phylogenetic tree with a total of 1084 sequences from the U.S. (Figure S4). Though each of the G based clades formed well-defined groups, there was some intermingling observed, which may have been a result of mis-categorization at the time the clades were defined. The tree structure also reveals the evolutionary relationships between the VOCs and the clades from which they were derived. Specifically, the Alpha and Gamma VOCs branch within Clade GR, while the Beta and Epsilon are within Clade GH. For our analysis, Delta falls within Clade G since, at the time of collection, Clade GK had not yet been defined. This representation further highlights the greater genetic distance between VOCs, especially Alpha, Gamma, and Delta, relative to the G clades, again suggesting that the VOCs emerged through distinct evolutionary pathways.

3.4. SARS-CoV-2 Mutations Present in ≥5% of the Population Detected in the U.S. in 2020

Levels of SARS-CoV-2 sequencing in the U.S. determine both the sensitivity with which we are able to detect emerging mutations and our degree of statistical confidence that we have done or can do so. From U.S. viral sequences submitted to GISAID prior to 15 December 2020, we were able to detect mutations present in the population at frequencies ≥5% in Phases 1 and 2 and ≥14.5% in Phase 3 with 95% confidence (Figure S5). Increasing the sensitivity of and confidence in this type of analysis would require greater sampling, as is now being pursued. For instance, to detect mutations present at 1% or 0.1% frequency with 99% confidence, the viral sequencing capacity in the U.S. needs to increase by ~7.8-fold relative to 2020 levels (to 460 sequences/day) or ~78-fold (4600 sequences/day), respectively (Figure S5). Though these goals are now within reach in the U.S. due to increased funding for SARS-CoV-2 sequencing, achieving this level of surveillance worldwide, especially in regions with less abundant resources and higher population densities, seems unlikely, and more practical alternatives must be considered. This approach is in agreement with a recommendation by the European Centre for Disease Prevention and Control advocating for targeted sequencing of SARS-CoV-2 cases in select populations (e.g., vaccine breakthrough and outbreak clusters) as a complement to broader surveillance [41].

3.5. At Least 54 New Amino Acid Changes Emerged in 2020 and Persisted in ≥5% of the U.S. Population in at Least one Phase of the Spread

To detect the emergence of new SARS-CoV-2 mutations in the U.S. in 2020, majority-rule consensus sequences of each G clade from Phase 1 were compared to all respective sequences in the subsequent two phases (Tables S3–S5). Only mutations present in ≥5% of the population were included in the analyses for reasons described in Materials and Methods, although select mutations of specific interest present at lower frequencies were also noted and are discussed below. About half of the non-clade-defining mutations that arose in the U.S. in 2020 and persisted at frequencies ≥5% were nonsynonymous, i.e., clade G: ORF1a (9 nonsynonymous mutations), ORF1b (4), Spike (1), Nucleocapsid (3); clade GH: ORF1a (12), ORF1b (6), Spike (1), Nucleocapsid (6); and clade GR: ORF1a (6), ORF1b (3), Spike (4), Matrix (1). While some mutations arose and then declined during 2020 (or shifted

geographically), most persisted and increased in frequency (Figure 3A–C). In particular, clade G mutation N^{S194L} and clade GH mutations $ORF1a^{L3352F}$ and $ORF1b^{N1653D;R2613C}$ increased more than 40% from Phase 1 to Phase 3. Many new mutations were detected in Phase 3 despite the limitations of extremely shallowing sampling.

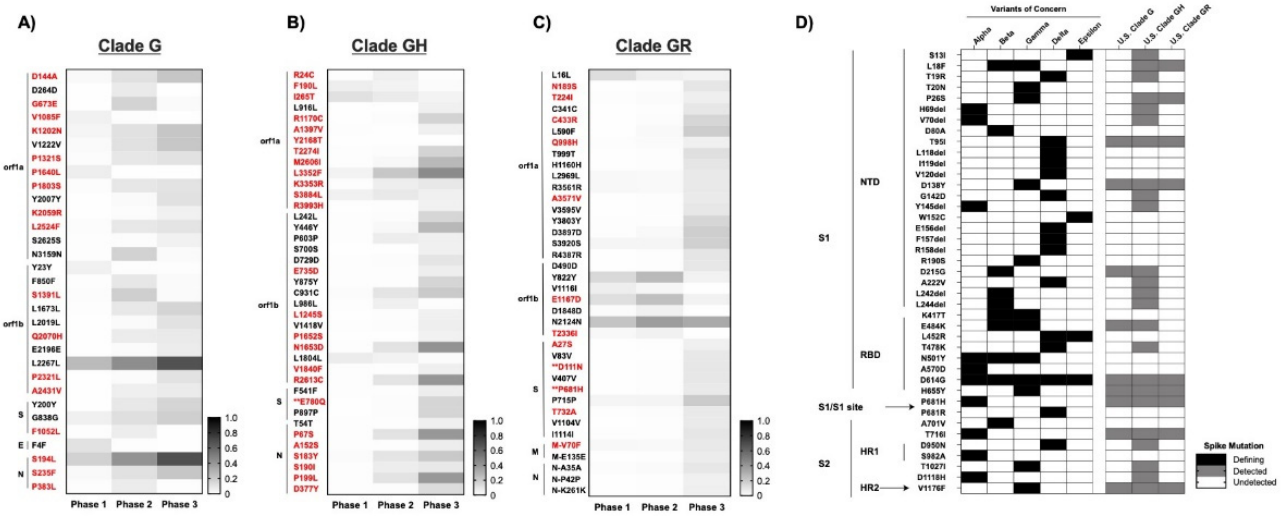

Figure 3. Emerged mutations in SARS-CoV-2 have increased in the U.S: (**A**–**C**) Non-clade-defining mutations of Clades G, GH, and GR. Emerging mutations are shown in which, during at least one Phase in 2020, the frequency exceeded 5%. Sequences were compared to the majority-consensus sequence for each respective clade in Phase 1. Mutation designations reflect the relative amino acid positions in the gene regions. There were no common mutations among the G-based clades. Red text denotes non-synonymous mutations. (**) Mutations that occur in T- or B-cell epitope regions. (**D**) Comparison of non-synonymous mutations and deletions in Spike between VOCs and the U.S. G-based clades across all three phases. The presence or absence of mutations is indicated by shading: VOC-defining mutations (black), mutations detected during at least one phase in the respective G-clade (gray), and undetected in either VOC or U.S. G-clades (white).

The ratio of nonsynonymous-to-synonymous fixation rates (dN/dS) serves to indicate the nature of selective pressure at individual coding positions in a gene. More specifically, this ratio can suggest neutral (dN/dS~1), negative/purifying (dN/dS < 1), or positive (dN/dS > 1) selection at a given site. The overall dN/dS for SARS-CoV-2 has been shown to be under negative selection [36]; however, we did observe some mutations in Spike that were under positive selection (dN/dS > 1). Several of these mapped to T- and B-cell epitopes, including clade G $Nucleocapsid^{S235F}$, clade GH S^{E780Q}, and clade GR $S^{D111N;P681H}$ (Figure S6). Some positions in predicted antibody epitopes in Spike, e.g., $S^{94;562;816}$, were found to be under purifying selection (dN/dS < 1). The same is true of the alanine residue in the furin cleavage site of Spike (RR**A**R), suggesting that substitution at this position confers a replicative disadvantage. This conclusion is supported by in vitro experiments wherein virus containing an RR**K**R polybasic cleavage site produces larger syncytia but exhibits decreased infectivity [42]. Conversely, the proline adjacent to the S1/S2 cleavage site (S^{681}) required for infection in human lungs [42–44] was under positive selection, indicative of room for evolutionary improvement. S^{5}, located 5′ to the start of the NTD/S1, was also under positive selection in all G-based clades. Outside of Spike, we identified several positively selective nonsynonymous changes in both B-cell and MHC-Class I/II T-cell epitopes in Nucleocapsid (Figure S7).

In accordance with an elevated level of importance, VOCs are defined primarily, though not exclusively, by their associated Spike mutations in functionally and immunogenically important regions within the N-terminal and Receptor Binding Domains [45]. Critical investigations have been and are being conducted to determine how emergent

Spike mutations affect vaccine efficacy [46,47], immunotherapeutics (i.e., monoclonal antibodies and convalescent plasma) [3,47], and viral transmission and pathogenesis [35,39,48]. Selection of Spike mutations among recipients of convalescent plasma is particularly concerning [48–50], since the partial protection conferred by such treatments is likely conducive to immune escape, and may have contributed to the emergence of the Alpha, Beta, Gamma, Delta, and Epsilon VOCs, and, most recently, the Omicron VOC [51].

3.6. Mutations in Spike and Implications for COVID-19 Vaccines

Since all current vaccines were designed to potentiate cellular and humoral immune memory responses against the SARS-CoV-2 Spike protein [52], the emergence of mutations in the viral gene encoding Spike warrant special attention, including those found at overall frequencies of less than 5% in all three phases. For instance, three mutations in known B-cell epitopes in Spike became prominent in Clades GH and GR starting in Phase 2 (Supplementary Materials Figure S6). The frequency of these mutations increased with each new phase of infections, perhaps constituting a rare example of enrichment due to immune resistance outside of VOCs or due to founder effects. It has been shown in studies using the Ad26.COV2.S (Janssen/Johnson & Johnson) vaccine that several VOCs (Alpha, Beta, and Gamma) produced similar CD4 and CD8 responses; however, there were reduced neutralizing antibody responses against the Beta and Gamma VOCs among vaccine recipients [53]. In contrast, studies in nonhuman primates vaccinated with mRNA-1273 (Moderna) showed measurable dose–responses of circulating and mucosal antibodies against the VOCs [54]. Administration of either BNT162b2 (Pfizer/BioNTech) or ChAdOx1 nCOV-19 (Oxford/AstraZeneca) generated neutralizing antibodies against the Delta VOC but with 3-to-5-fold lower titers compared to the Alpha VOC, as well as a modestly decreased vaccine effectiveness in Delta vs. Alpha [55,56]. Although the VOCs were not detected in the U.S. until late 2020 or early 2021, our analysis shows that many of their individual defining Spike mutations were present in the U.S. as early as Phase 1 (Figure 3D). However, with the exception of S^{P681H}, the average frequencies of these individual mutations were all <1%. S^{E484K} and S^{N501Y} mutations are present in multiple VOCs and are particularly concerning, having been reported to decrease susceptibility to antibody neutralization [3]. Moreover, neutralizing antibodies from COVID-19 patients or SARS-CoV-2-infected humanized mice were shown to be less effective against Alpha, Beta, and Gamma VOCs, each of which harbor the S^{E484K} and/or S^{N501Y} mutations [57], although immunity induced by mRNA-based vaccines appears to remain at least partially effective against E484K-containing VOCs [58–61]. With respect to the durability of vaccine immunity, antibody functionality persists at low levels for at least 6 months after primary mRNA-1273 vaccination when compared to the VOCs. However, a greater reduction in antibody recognition was observed in cases of the Gamma VOC, of which the efficacy for the BNT162b2, Ad26.COV2.S, and mRNA-1273 vaccines has been studied against the VOCs as well as support booster vaccinations to prolong protection [11,56,62–67]. Comparably, sera from individuals previously infected with the Beta or Gamma VOCs were less effective in neutralizing the Delta VOC than from a strain isolated from Australia early in the pandemic [68].

Fortunately, the evolutionary distance between these SARS-CoV-2 VOCs and complete resistance to adaptive immunity conferred by current vaccines remains substantial. Not only does vaccine-induced immunity remain highly effective against hospitalization and death caused by any of the current VOCs [69], but the sudden worldwide dominance of the Delta VOC in 2021 was due almost completely to its increased transmission and not vaccine resistance. Whether this will also prove to be true for more recently discovered VOC, Omicron, remains to be seen, since the relative transmissibility, vaccine resistance, and pathogenicity of this variant have yet to be fully characterized. However, the more than 30 Spike gene mutations in the Omicron VOC relative to the Wuhan-Hu-1 reference are again suggestive of an evolutionary environment distinct from the one most prevalent in the U.S. in 2020.

Although multiple variants harboring individual mutations thought to incrementally contribute to immune resistance were found in the U.S. at low frequencies early in 2020, our analysis indicates that, after nearly a year, any selective advantage conferred by these mutations was insufficient to significantly expand their representation in the viral population. It is therefore perhaps not surprising that, after analyzing tens of thousands of sequences representing millions of cases in the U.S. in 2020, we found no evidence of variants harboring multiple immune resistance mutations serially acquired over the course of several viral transmissions among the sequences analyzed. Of course, our analysis is not necessarily predictive of what will occur in the future with the now-dominant Omicron VOC, or newly emergent VOCs in the face of rising immune pressure due to prior infection and/or vaccination. Our work will, however, serve as an important baseline for comparison in the event of changes in the rate or mode of evolution of SARS-CoV-2.

4. Conclusions

Although the replication fidelity of SARS-CoV-2 is quite high relative to other RNA viruses [5,70], its genetic stability is countered by the expansive spread of the virus both in the U.S. and globally. On balance, evolution of this virus might be best characterized as slow but inexorable, driven largely by genetic drift but also influenced by selective pressures such as relative infectivity, relative transmissibility, and to a lesser extent, immune evasion. In contrast, the relatively recent and suddenly emergent VOCs are characterized by a significantly higher degree of genetic divergence that is rapidly acquired and clearly confers a replicative advantage. These variants were most likely the product of isolated cases in which the evolutionary environments differ substantially from the norm; e.g., from chronically infected immunosuppressed individuals, including those who received treatment with monoclonal antibodies or convalescent serum [49–51], and/or animal reservoirs.

Regardless of the evolutionary pathways taken, the emergence of more rapidly transmissible and partially immune resistant variants in both the U.S. and global viral populations threatens the continued efficacy of current treatments, and vaccines and will do so increasingly as the virus continues to spread and evolve. The sudden worldwide dominance of the Delta VOC in 2021, and the more recent emergence and spread of the Omicron VOC, exemplifies this concern. In response, scientists worldwide are coming together to increase timely SARS-CoV-2 sequencing and analysis to help guide decisions on current and future health policies. In addition, because it is becoming increasingly apparent that the current VOCs emerged from atypical evolutionary environments, targeted worldwide surveillance of high-risk human infections and animal reservoirs may be the best means of detecting future VOCs as they emerge, particularly in resource-limited regions.

In 2020, efforts to slow the spread of the virus within the U.S. were hampered by incomplete adherence to recommended preventative measures (e.g., mask wearing, hand washing, social distancing) as well as inconsistent and even conflicting messaging regarding these measures. These failures indirectly accelerated viral divergence, increased genetic diversity, and resulted in the accumulation of nonsynonymous mutations, many of which are within epitopes now associated with resistance to neutralizing antibodies and that may ultimately contribute to immune evasion. Now, in 2022, the Delta VOC rapidly spread across the U.S. as well as Omicron VOC, primarily among the unvaccinated population but also, to a lesser degree, among the vaccinated, raising the bar for herd immunity even higher.

As more people become infected and/or are vaccinated, selective pressure for immune-resistant variants will increase concomitantly. It is therefore essential that we monitor individuals who become infected post-vaccination, as well as those with prolonged infection (e.g., immunocompromised individuals), both to better understand the capacity of SARS-CoV-2 for escape from vaccine-induced immunity and to identify and isolate resistant variants early when they emerge. Finally, although our study focused on the U.S., we recognize that the emergence of potential new VOCs and variants of interest is an ongoing and global concern. We must all therefore diligently and intelligently use our resources to

detect and analyze new variants and, just as importantly, continue to encourage measures to prevent their spread.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/v14010104/s1, Figure S1: SARS-CoV-2 epidemic at the divisional level in the U.S., Figure S2: Proportion of SARS-CoV-2 genomes in each GISAID clade, Figure S3: Genetic distance over time in 2020 for GISAID clades and Variants of Concern, Figure S4 Phylogenetic analysis of SARS-CoV-2 in the U.S., Figure S5: Determining number of sequences of sampling at a given probability to detect variant frequency, Figure S6: Mutation in G clades that are located in MHC-II HLA-DR T-cell and B-cell epitopes, Figure S7: Selection pressure and epitopes detected in Nucleocapsid, Table S1: Demographics and Regional Division based on the U.S. Census Bureau, Table S2: SARS-CoV-2 T-cell and B-cell epitopes Table S3: Non-associated Clade G mutations that either persisted or emerged during 2020, Table S4: Non-associated Clade GH mutations that either persisted or emerged during 2020, Table S5: Non-associated Clade GR mutations that either persisted or emerged during 2020, Supplemental Dataset: Sequences Analyzed.

Author Contributions: Conception and design: A.A.C., J.M.C., J.W.R. and M.F.K. collection and assembly of data: A.A.C., W.S., J.S. analysis and interpretation of the data: A.A.C., W.S., J.M.C., J.W.R. and M.F.K. drafting of the article: A.A.C., J.M.C., J.W.R. and M.F.K. All authors wrote and approved the final article. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this research was provided to Mary F. Kearney by the National Cancer Institute's Intramural Center for Cancer Research which supports the HIV Dynamics and Replication Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available in the main text or the Supplementary Materials.

Acknowledgments: We gratefully acknowledge the authors, the originating and submitting laboratories for their sequence and metadata shared through GISAID, on which this research is based. We would also like to thank the COVID Tracking Project.

Conflicts of Interest: The authors declare no competing interest.

References

1. Holshue, M.L.; DeBolt, C.; Lindquist, S.; Lofy, K.H.; Wiesman, J.; Bruce, H.; Spitters, C.; Ericson, K.; Wilkerson, S.; Tural, A.; et al. First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* **2020**, *382*, 929–936. [CrossRef] [PubMed]
2. Centers for Disease Control and Prevention. COVID-19 Response. COVID-19 Case Surveillance Public Data Access, Summary, and Limitations. Available online: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36 (accessed on 11 June 2020).
3. Wang, P.; Casner, R.G.; Nair, M.S.; Wang, M.; Yu, J.; Cerutti, G.; Liu, L.; Kwong, P.D.; Huang, Y.; Shapiro, L.; et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe* **2021**, *29*, 747–751.e4. [CrossRef] [PubMed]
4. Zhang, W.; Davis, B.D.; Chen, S.S.; Sincuir Martinez, J.M.; Plummer, J.T.; Vail, E. Emergence of a Novel SARS-CoV-2 Variant in Southern California. *JAMA* **2021**, *325*, 1324–1326. [CrossRef] [PubMed]
5. Wu, A.; Wang, L.; Zhou, H.-Y.; Ji, C.-Y.; Xia, S.Z.; Cao, Y.; Meng, J.; Ding, X.; Gold, S.; Jiang, T.; et al. One year of SARS-CoV-2 evolution. *Cell Host Microbe* **2021**, *29*, 503–507. [CrossRef]
6. Plante, J.A.; Mitchell, B.M.; Plante, K.S.; Debbink, K.; Weaver, S.C.; Menachery, V.D. The variant gambit: COVID-19's next move. *Cell Host Microbe* **2021**, *29*, 508–515. [CrossRef]
7. The COVID Tracking Project. Available online: https://covidtracking.com/about-data/data-summary (accessed on 17 December 2020).
8. Available online: https://www.census.gov/geographies/reference-maps/2010/geo/2010-census-regions-and-divisions-of-the-united-states.html (accessed on 17 December 2020).
9. Elbe, S.; Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **2017**, *1*, 33–46. [CrossRef]
10. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data—From vision to reality. *Euro Surveill. Bull. Eur. Sur Les Mal. Transm. Eur. Commun. Dis. Bull.* **2017**, *22*, 30494. [CrossRef]
11. Tao, K.; Tzou, P.L.; Nouhin, J.; Gupta, R.K.; de Oliveira, T.; Kosakovsky Pond, S.L.; Fera, D.; Shafer, R.W. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **2021**, *22*, 757–773. [CrossRef]

12. 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. [CrossRef]
13. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
14. R Team. *RStudio: Integrated Development for R*; R CoreTeam: Vienna, Austria, 2020.
15. Coronavirus (COVID-19) in the UK. Available online: https://coronavirus.data.gov.uk/ (accessed on 17 December 2020).
16. Coronavirus (COVID-19) Case Numbers and Statistics. Available online: https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-case-numbers-and-statistics (accessed on 17 December 2020).
17. 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. [CrossRef]
18. Kumar, S.; Tao, Q.; Weaver, S.; Sanderford, M.; Caraballo-Ortiz, M.A.; Sharma, S.; Pond, S.L.K.; Miura, S. An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. *Mol. Biol. Evol.* **2021**, *38*, 3046–3059. [CrossRef] [PubMed]
19. Jordan, M.R.; Kearney, M.; Palmer, S.; Shao, W.; Maldarelli, F.; Coakley, E.P.; Chappey, C.; Wanke, C.; Coffin, J.M. Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. *J. Virol Methods* **2010**, *168*, 114–120. [CrossRef] [PubMed]
20. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]
21. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **2019**, *35*, 4453–4455. [CrossRef] [PubMed]
22. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
23. Darriba, D.; Posada, D.; Kozlov, A.M.; Stamatakis, A.; Morel, B.; Flouri, T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* **2019**, *37*, 291–294. [CrossRef] [PubMed]
24. Edler, D.; Klein, J.; Antonelli, A.; Silvestro, D. raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. *Methods Ecol. Evol.* **2021**, *12*, 373–377. [CrossRef]
25. 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*, veab064. [CrossRef]
26. 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. [CrossRef]
27. Julia, L.; Mullen, G.T.; Latif, A.A.; Alkuzweny, M.; Cano, M.; Haag, E.; Zhou, J.; Zeller, M.; Hufbauer, E.; Matteson, N.; et al. Outbreak.info. Available online: https://outbreak.info/ (accessed on 17 December 2020).
28. Worobey, M.; Pekar, J.; Larsen, B.B.; Nelson, M.I.; Hill, V.; Joy, J.B.; Rambaut, A.; Suchard, M.A.; Wertheim, J.O.; Lemey, P. The emergence of SARS-CoV-2 in Europe and North America. *Science* **2020**, *370*, 564–570. [CrossRef]
29. Chang, S.; Pierson, E.; Koh, P.W.; Gerardin, J.; Redbird, B.; Grusky, D.; Leskovec, J. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **2021**, *589*, 82–87. [CrossRef]
30. Gonzalez-Reiche, A.S.; Hernandez, M.M.; Sullivan, M.J.; Ciferri, B.; Alshammary, H.; Obla, A.; Fabre, S.; Kleiner, G.; Polanco, J.; Khan, Z.; et al. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **2020**, *369*, 297–301. [CrossRef]
31. Ladner, J.T.; Larsen, B.B.; Bowers, J.R.; Hepp, C.M.; Bolyen, E.; Folkerts, M.; Sheridan, K.; Pfeiffer, A.; Yaglom, H.; Lemmer, D.; et al. An Early Pandemic Analysis of SARS-CoV-2 Population Structure and Dynamics in Arizona. *mBio* **2020**, *11*, e02107-20. [CrossRef]
32. Bedford, T.; Greninger, A.L.; Roychoudhury, P.; Starita, L.M.; Famulare, M.; Huang, M.L.; Nalla, A.; Pepper, G.; Reinhardt, A.; Xie, H.; et al. Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **2020**, *370*, 571–575. [CrossRef]
33. Lemieux, J.E.; Siddle, K.J.; Shaw, B.M.; Loreth, C.; Schaffner, S.F.; Gladden-Young, A.; Adams, G.; Fink, T.; Tomkins-Tinch, C.H.; Krasilnikova, L.A.; et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **2021**, *371*, eabe3261. [CrossRef]
34. Wang, R.; Chen, J.; Gao, K.; Hozumi, Y.; Yin, C.; Wei, G.W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun. Biol.* **2021**, *4*, 228. [CrossRef] [PubMed]
35. 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. [CrossRef] [PubMed]
36. Dearlove, B.; Lewitus, E.; Bai, H.; Li, Y.; Reeves, D.B.; Joyce, M.G.; Scott, P.T.; Amare, M.F.; Vasan, S.; Michael, N.L.; et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 23652–23662. [CrossRef]
37. Li, Q.; Wu, J.; Nie, J.; Zhang, L.; Hao, H.; Liu, S.; Zhao, C.; Zhang, Q.; Liu, H.; Nie, L.; et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **2020**, *182*, 1284–1294.e9. [CrossRef] [PubMed]

38. Weissman, D.; Alameh, M.G.; de Silva, T.; Collini, P.; Hornsby, H.; Brown, R.; LaBranche, C.C.; Edwards, R.J.; Sutherland, L.; Santra, S.; et al. D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. *Cell Host Microbe* **2021**, *29*, 23–31.e4. [CrossRef] [PubMed]
39. Volz, E.; Hill, V.; McCrone, J.T.; Price, A.; Jorgensen, D.; O'Toole, Á.; Southgate, J.; Johnson, R.; Jackson, B.; Nascimento, F.F.; et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **2021**, *184*, 64–75. [CrossRef] [PubMed]
40. Achaz, G.; Palmer, S.; Kearney, M.; Maldarelli, F.; Mellors, J.W.; Coffin, J.M.; Wakeley, J. A robust measure of HIV-1 population turnover within chronically infected individuals. *Mol. Biol. Evol.* **2004**, *21*, 1902–1912. [CrossRef] [PubMed]
41. European Centre for Disease Prevention and Control. *Guidance for Representative and Targeted Genomic SARS-Cov-2 Monitoring*; European Centre for Disease Prevention and Control: Solna Municipality, Sweden, 2021.
42. Hoffmann, M.; Kleine-Weber, H.; Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* **2020**, *78*, 779–784.e5. [CrossRef] [PubMed]
43. Zhou, H.; Chen, X.; Hu, T.; Li, J.; Song, H.; Liu, Y.; Wang, P.; Liu, D.; Yang, J.; Holmes, E.C.; et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* **2020**, *30*, 2196–2203.e3. [CrossRef]
44. Andersen, K.G.; Rambaut, A.; Lipkin, W.I.; Holmes, E.C.; Garry, R.F. The proximal origin of SARS-CoV-2. *Nat. Med.* **2020**, *26*, 450–452. [CrossRef]
45. Konings, F.; Perkins, M.D.; Kuhn, J.H.; Pallen, M.J.; Alm, E.J.; Archer, B.N.; Barakat, A.; Bedford, T.; Bhiman, J.N.; Caly, L.; et al. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat. Microbiol.* **2021**, *6*, 821–823. [CrossRef]
46. Greaney, A.J.; Starr, T.N.; Gilchuk, P.; Zost, S.J.; Binshtein, E.; Loes, A.N.; Hilton, S.K.; Huddleston, J.; Eguia, R.; Crawford, K.H.D.; et al. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **2021**, *29*, 44–57.e9. [CrossRef]
47. Starr, T.N.; Greaney, A.J.; Addetia, A.; Hannon, W.W.; Choudhary, M.C.; Dingens, A.S.; Li, J.Z.; Bloom, J.D. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **2021**, *371*, 850–854. [CrossRef]
48. Choi, B.; Choudhary, M.C.; Regan, J.; Sparks, J.A.; Padera, R.F.; Qiu, X.; Solomon, I.H.; Kuo, H.H.; Boucau, J.; Bowman, K.; et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **2020**, *383*, 2291–2293. [CrossRef]
49. Kemp, S.A.; Collier, D.A.; Datir, R.P.; Ferreira, I.A.T.M.; Gayed, S.; Jahun, A.; Hosmillo, M.; Rees-Spear, C.; Mlcochova, P.; Lumb, I.U.; et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **2021**, *592*, 277–282. [CrossRef] [PubMed]
50. Avanzato, V.A.; Matson, M.J.; Seifert, S.N.; Pryce, R.; Williamson, B.N.; Anzick, S.L.; Barbian, K.; Judson, S.D.; Fischer, E.R.; Martens, C.; et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **2020**, *183*, 1901–1912. [CrossRef] [PubMed]
51. 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 antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *medRxiv* **2021**. [CrossRef]
52. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **2020**, *586*, 516–527. [CrossRef]
53. Alter, G.; Yu, J.; Liu, J.; Chandrashekar, A.; Borducchi, E.N.; Tostanoski, L.H.; McMahan, K.; Jacob-Dolan, C.; Martinez, D.R.; Chang, A.; et al. Immunogenicity of Ad26.COV2.S vaccine against SARS-CoV-2 variants in humans. *Nature* **2021**, *596*, 268–272. [CrossRef]
54. Corbett, K.S.; Nason, M.C.; Flach, B.; Gagne, M.; O'Connell, S.; Johnston, T.S.; Shah, S.N.; Edara, V.V.; Floyd, K.; Lai, L.; et al. Immune correlates of protection by mRNA-1273 vaccine against SARS-CoV-2 in nonhuman primates. *Science* **2021**, *373*, eabj0299. [CrossRef]
55. Planas, D.; Veyer, D.; Baidaliuk, A.; Staropoli, I.; Guivel-Benhassine, F.; Rajah, M.M.; Planchais, C.; Porrot, F.; Robillard, N.; Puech, J.; et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **2021**, *596*, 276–280. [CrossRef] [PubMed]
56. Lopez Bernal, J.; Andrews, N.; Gower, C.; Gallagher, E.; Simmons, R.; Thelwall, S.; Stowe, J.; Tessier, E.; Groves, N.; Dabrera, G.; et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* **2021**, *385*, 585–594. [CrossRef]
57. Yuan, M.; Huang, D.; Lee, C.D.; Wu, N.C.; Jackson, A.M.; Zhu, X.; Liu, H.; Peng, L.; van Gils, M.J.; Sanders, R.W.; et al. Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. *Science* **2021**, *373*, 818–823. [CrossRef]
58. Garcia-Beltran, W.F.; Lam, E.C.; St. Denis, K.; Nitido, A.D.; Garcia, Z.H.; Hauser, B.M.; Feldman, J.; Pavlovic, M.N.; Gregory, D.J.; Poznansky, M.C.; et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **2021**, *184*, 2372–2383.e9. [CrossRef] [PubMed]
59. Edara, V.V.; Norwood, C.; Floyd, K.; Lai, L.; Davis-Gardner, M.E.; Hudson, W.H.; Mantus, G.; Nyhoff, L.E.; Adelman, M.W.; Fineman, R.; et al. Infection- and vaccine-induced antibody binding and neutralization of the B.1.351 SARS-CoV-2 variant. *Cell Host Microbe* **2021**, *29*, 516–521.e3. [CrossRef]
60. Kuzmina, A.; Khalaila, Y.; Voloshin, O.; Keren-Naus, A.; Boehm-Cohen, L.; Raviv, Y.; Shemer-Avni, Y.; Rosenberg, E.; Taube, R. SARS-CoV-2 spike variants exhibit differential infectivity and neutralization resistance to convalescent or post-vaccination sera. *Cell Host Microbe* **2021**, *29*, 522–528.e2. [CrossRef] [PubMed]

61. Shen, X.; Tang, H.; McDanal, C.; Wagh, K.; Fischer, W.; Theiler, J.; Yoon, H.; Li, D.; Haynes, B.F.; Sanders, K.O.; et al. SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral spike vaccines. *Cell Host Microbe* **2021**, *29*, 529–539.e3. [CrossRef]
62. Pegu, A.; O'Connell, S.; Schmidt, S.D.; O'Dell, S.; Talana, C.A.; Lai, L.; Albert, J.; Anderson, E.; Bennett, H.; Corbett, K.S.; et al. Durability of mRNA-1273 vaccine-induced antibodies against SARS-CoV-2 variants. *Science* **2021**, *373*, 1372–1377. [CrossRef] [PubMed]
63. Mbaeyi, S. The Advisory Committee on Immunization Practices' Interim Recommendations for Additional Primary and Booster Doses of COVID-19 Vaccines—United States, 2021. *MMWR Morb. Mortal Wkly. Rep.* **2021**, *70*, 1545–1552. [CrossRef]
64. Abu-Raddad, L.J.; Chemaitelly, H.; Butt, A.A. Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. *N. Engl. J. Med.* **2021**, *385*, 187–189. [CrossRef]
65. Haas, E.J.; Angulo, F.J.; McLaughlin, J.M.; Anis, E.; Singer, S.R.; Khan, F.; Brooks, N.; Smaja, M.; Mircus, G.; Pan, K.; et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: An observational study using national surveillance data. *Lancet* **2021**, *397*, 1819–1829. [CrossRef]
66. Sadoff, J.; Gray, G.; Vandebosch, A.; Cárdenas, V.; Shukarev, G.; Grinsztejn, B.; Goepfert, P.A.; Truyers, C.; Fennema, H.; Spiessens, B.; et al. Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N. Engl. J. Med.* **2021**, *384*, 2187–2201. [CrossRef] [PubMed]
67. Sheikh, A.; McMenamin, J.; Taylor, B.; Robertson, C. SARS-CoV-2 Delta VOC in Scotland: Demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **2021**, *397*, 2461–2462. [CrossRef]
68. Liu, C.; Ginn, H.M.; Dejnirattisai, W.; Supasa, P.; Wang, B.; Tuekprakhon, A.; Nutalai, R.; Zhou, D.; Mentzer, A.J.; Zhao, Y.; et al. Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **2021**, *184*, 4220–4236.e13. [CrossRef]
69. Thompson, M.G.; Burgess, J.L.; Naleway, A.L.; Tyner, H.; Yoon, S.K.; Meece, J.; Olsho, L.E.W.; Caban-Martinez, A.J.; Fowlkes, A.L.; Lutrick, K.; et al. Prevention and Attenuation of Covid-19 with the BNT162b2 and mRNA-1273 Vaccines. *N. Engl. J. Med.* **2021**, *385*, 320–329. [CrossRef] [PubMed]
70. Rausch, J.W.; Capoferri, A.A.; Katusiime, M.G.; Patro, S.C.; Kearney, M.F. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 24614–24616. [CrossRef] [PubMed]