The unique evolutionary dynamics of the SARS-CoV-2 Delta variant

Adi Stern*†[1,2], Shay Fleishon*[3], Talia Kustin[1,2], Edo Dotan[1,2], Michal Mandelboim[3,4], Oran Erster[3], Israel Consortium of SARS-CoV-2 sequencing[Ψ], Ella Mendelson[3,4], Orna Mor[3,4], Neta S. Zuckerman†[3]

1 The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel.

2 Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel

3 Central Virology Laboratory, Public Health Services, Ministry of Health and Sheba Medical Center, Tel-Hashomer, Israel.

4 School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel.

* Co-equal authorship

† Corresponding authors

[Ψ] Israel Consortium of SARS-CoV-2 sequencing: Neta S. Zuckerman, Orna Mor, Efrat Dahan Bucris, Michal Mandelboim, Danit Sofer, Dana Bar-Ilan, Miranda Geva, Omer Asraf, Oran Erster, Gideon Rechavi, Efrat Glick-Saar, Nir Rainy, Chen Weiner, Reut Sorek-Abramovich, Yevgeni Yegorov, Anna Vishnevsky, Patricia Benveniste-Lekovitz, Abu Hamad Ramzia, Adina Bar Chaim, Ella Mendelson.

Abstract

The SARS-Coronavirus-2 (SARS-CoV-2) driven pandemic was first recognized in late 2019, and the first few months of its evolution were relatively clock-like, dominated mostly by neutral substitutions. In contrast, the second year of the pandemic was punctuated by the emergence of several variants that bore evidence of dramatic evolution. Here, we compare and contrast evolutionary patterns of various variants, with a focus on the recent Delta variant. Most variants are characterized by long branches leading to their emergence, with an excess of non-synonymous substitutions occurring particularly in the Spike and Nucleocapsid proteins. In contrast, the Delta variant that is now becoming globally dominant, lacks the signature long branch, and is characterized by a step-wise evolutionary process that is ongoing. Contrary to the "star-like" topologies of other variants, we note the formation of several distinct clades within Delta that we denote as clades A-E. We find that sequences from the Delta D clade are dramatically increasing in frequency across different regions of the globe. Delta D is characterized by an excess of non-synonymous mutations, mostly occurring in ORF1a/b, some of which occurred in parallel in other notable variants. We conclude that the Delta surge these days is composed almost exclusively of Delta D, and discuss whether selection or random genetic drift has driven the emergence of Delta D.

Introduction

The COVID-19 pandemic was first recognized in Wuhan, China in December 2019 and the etiological agent of the disease SARS-Coronavirus-2 was sequenced around ten days after the disease was formally discovered [1]. During the few months of the pandemic the evolution of the virus was relatively predictable, with substitutions accumulating at a fixed pace of about one substitution every second week [2][3][4][5], at a rate compatible with a molecular clock based on neutral evolution, and no evidence of dramatic positive selection was observed [6]. One notable exception was the D614G substitution in the Spike (S) protein, which quickly rose to fixation and indeed later evidence showed that this mutation is associated with increased transmissibility [7][8][9][10]. However, since the fall of 2020, several SARS-CoV-2 variants have emerged with particular genomic and epidemiological features, all suggestive of some form of selection operating on them. The first such variant was B.1.1.7 (Alpha), first detected in the U.K. Its most prominent characteristic was the overall wealth of substitutions found in it, especially in the region encoding for S [11]. Subsequently, it was found that B.1.1.7 spreads exceptionally rapidly [12][13] and its high transmissibility led to it displacing the originally circulating strains in many different countries around the world.

Overall, variants are characterized based on well-defined phylogenetic clades, and classified based on available evidence for increased transmissibility, virulence, or escape from the immune system / therapeutics. Variants with these characteristics are classified as variants of concern (VOC), variants of interest (VOI) or variants under monitoring (VUM) by health agencies such as the world health organization (WHO, www.who.int). Thus far, the WHO has defined four VOCs: B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta), and four VOIs: B.1.525 (Eta), B.1.526 (Iota), B.1.617.1 (Kappa) and C.37 (Lambda). An additional twelve variants previously classified as VOIs, such as B.1.427/9 originating in California [14], have now been reclassified as VUMs. For the purpose of our investigation herein, we collectively denote a VOC/VOI/VUM as a VO.

Following the detection of B.1.1.7, there has been a surge of different variants reported globally, which were characterized by punctual rises in frequencies of a particular variant, often accompanied by its demise a few months later. In the past few months, focus has turned to the Delta variant, which was first detected in India and has recently increased in

prevalence globally. The Delta variant currently seems to be displacing all other variants, including the highly dominant and contagious Alpha variant, in numerous countries across the globe [15]. In this study, we utilize evolutionary genomics to explain the rise and fall of different variants across time, with an emphasis on exploring the patterns of evolution in the globally-ascending Delta variant in comparison to other VOs.

Results

Genomic features common and unique across VOs

We begin with a global analysis of a representative sample of SARS-CoV-2 sequences. Within the phylogeny of these sequences, most tips (Fig. 1; white dots) are spread across the entire time-line of the tree, with short branches separating the various clades. This reflects the intense sampling of isolates, coupled with the molecular clock of the virus, which dictates, on average, approximately one substitution every second week [2]. However, the branches leading to the VOs represented in the tree are exceptionally long, representing the accumulation of many substitutions. We find that these branches share the following common features: (1) most have a higher proportion of non-synonymous (NS) substitutions as compared to non-VOs (Fig. 2A) ($p=0.01$, t-test (Methods)), (2) the entire 3' portion of the genome is enriched for NS substitutions, with a particular emphasis on the S and N genes, but also on ORF3a and ORF8 (Fig. 2B), (3) NS substitutions in S are mostly located in the N terminal domain (NTD) and receptor binding domain (RBD), (Fig. 2C), and (4) often, parallel substitutions are observed among the various VOs, such as at positions 452, 484, or 501 of the Spike protein [16].

When focusing on the Spike protein, we found that all VOs are characterized by substitutions prevalent particularly in NTD and (RBD), both critical targets of neutralizing antibodies [17] (Fig. 2C). Interestingly, the one exception is the Alpha VOC, which bears only one mutation associated with immune evasion (a deletion at position 144).

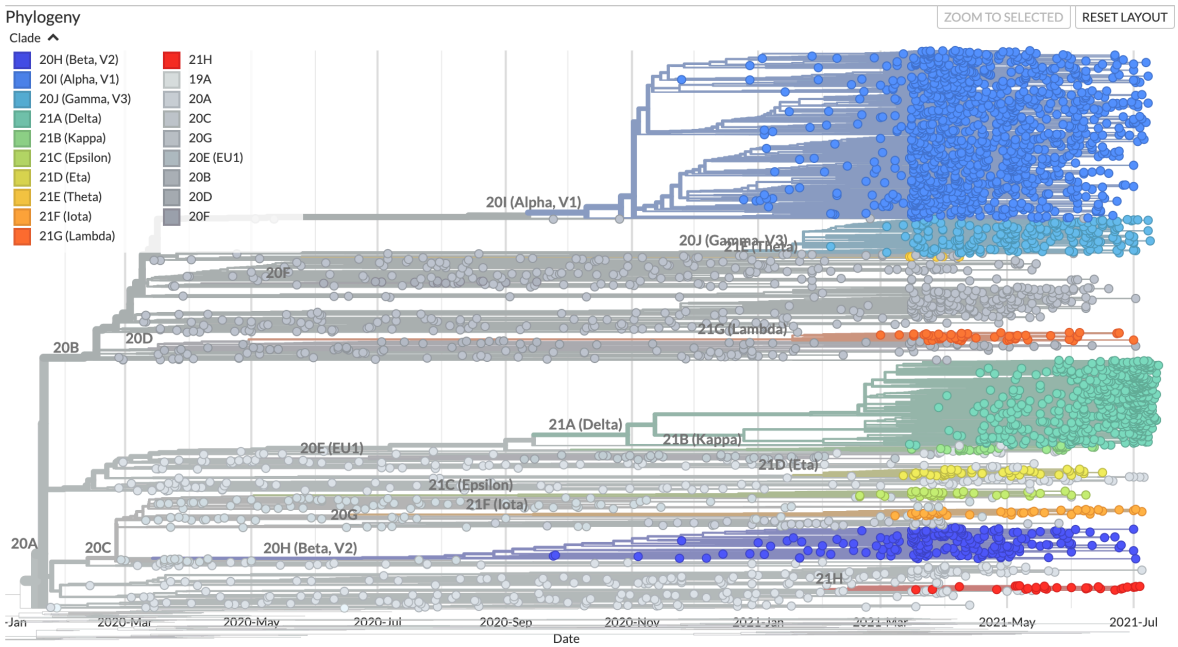

Figure 1. Time-aligned phylogeny of a representative sample of SARS-CoV-2 isolates. Different clades corresponding to VOCs and VOIs are colored. Notably, long branches lead to most VOCs/VOIs, suggesting an increased rate of evolution leading to their emergence. The phylogeny was generated using Nextstrain [18] on July 24 2021.

All VOs bear an amino-acid replacement at either position 203, 204, or 205 of the nucleocapsid protein (N), often combined. While it has previously been suggested that the K203/R204 polymorphisms create a non-canonical sgRNA [19], this is not expected to be the case for either the Delta or the Beta variants (Fig. S1). We thus suggest that the amino-acid replacements themselves may be adaptive. Accumulating evidence suggest that N has a crucial role in evasion from the cell autonomous innate immune response [20], [21], in viral assembly [22], and in interactions with cellular co-factors [23], and we suggest that amino-acid replacements at this region may increase replication capacity of the virus.

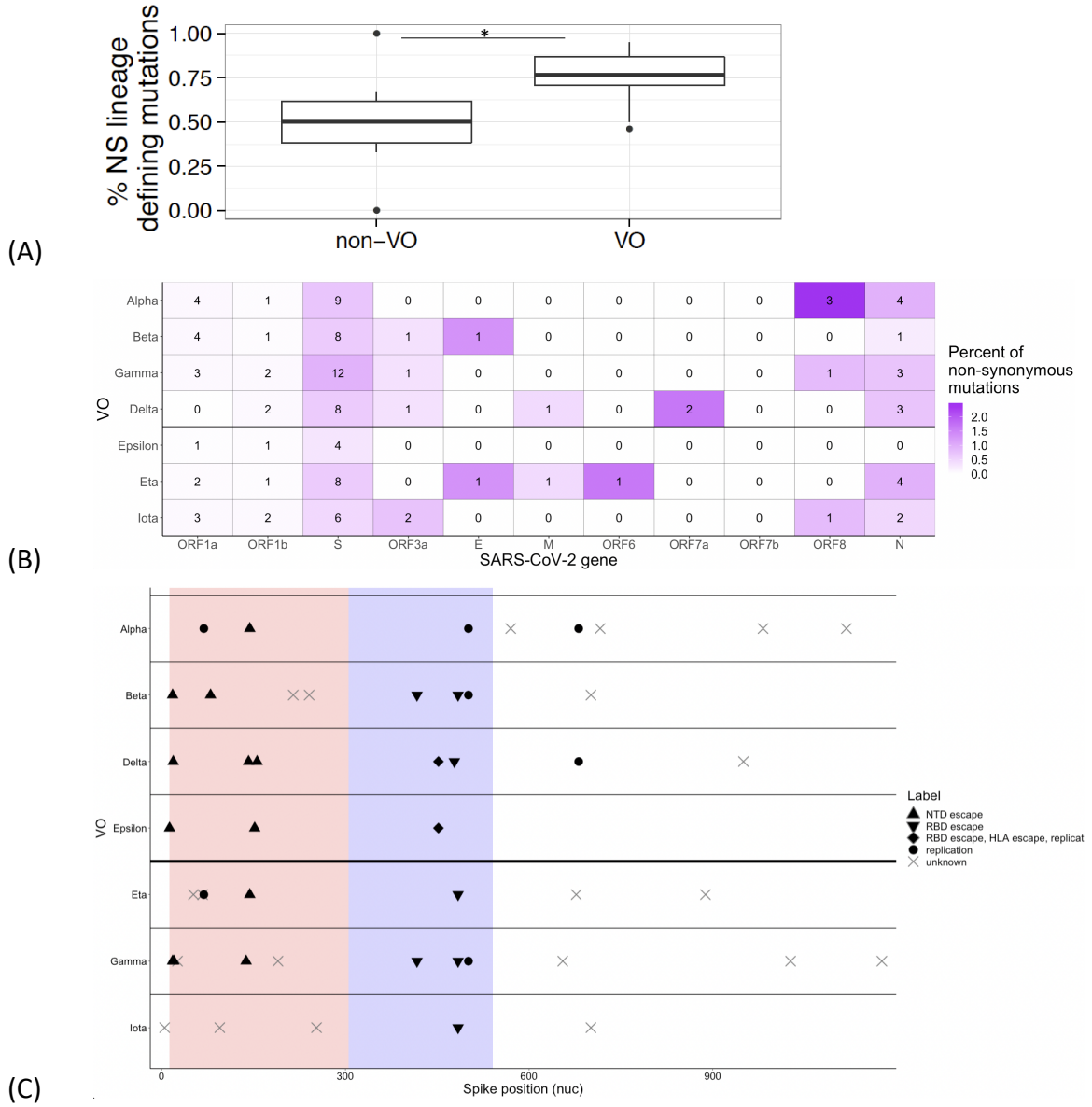

Figure 2. Comparison of lineage defining non-synonymous (NS) substitutions across VOs. (A) The proportion of NS (out of synonymous and non-synonymous substitutions) in VOs compared to non-VO B.1 lineages (Methods) (B) Summary of the number of non-synonymous substitutions per gene. Each cell is colored by the relative proportion of non-synonymous substitutions normalized by protein length. Deletions affecting coding regions are counted as non-synonymous in (A) and (B). (C) Non-synonymous substitutions with a focus on the spike protein. Pink and violet shading correspond to the NTD and RBD, respectively. Each substitution is shaped based on current evidence for the functional change it entails (Table S1), focusing on association with two global categories: escape from antibodies (squares and circles) and enhanced replication (regardless of the mechanisms). The D614G amino-acid substitution shared by all VOs was omitted for clarity.

Focus on the Delta variant and contrast with Alpha

Both the Delta and Alpha variants (corresponding to NextStrain lineages 21A and 20I, respectively), which spread worldwide, were inferred to have arisen first around the fall of 2020 (with confidence intervals spanning spring 2020 through winter 2020/2021) [18]. However, when analyzing the phylogenies of each variant, the tree topologies looks vastly different (Fig. 3A). VO phylogenies are characterized by multiple polytomies ("star"-like phylogenies) (Fig. 3A, Fig. S2), whereas the Delta phylogeny is more structured. The signature long branch leading to VOs, as described above, is absent in Delta, and it appears that some of the key signature substitutions associated with Delta (e.g., S: L452R, S: P681R) were created in a series of independent steps, as evidenced by multiple shared substitutions with the Kappa variant and other sequences. In fact, when performing a thorough analysis of all globally available Delta sequences (Methods), we were able to separate the Delta phylogeny into five distinct clades, which we label Delta A through E, each characterized by a specific set of substitutions (Methods) (Fig. 3B, Table 1). These clades encompass the three recently noted VOIs AY.1, AY.2 and AY.3 (Fig. 3A) and all other AY lineages defined by the Pango nomenclature system [24].

We went on to examine the lineage defining mutations of each clade. We first noted a high proportion of non-synonymous substitutions in some of the clades, in particular clades D and E, which are both characterized by a large number of overall lineage-defining mutations (Table 1). Interestingly, concurrent clade E sequences appear to be a result of recombination between an ancestral predicted E sequence and an ancestor of the B+C clade, since all clade E sequences share the substitutions unique to the B+C clade in the first ~5200 bases of their genomes (Table 1).

Next, we observed that when focusing on the *basal* Delta lineage (i.e., the branch leading to the emergence of the Delta variant), many lineage defining mutations also occurred independently in other VO lineages, in line with this being one of the four signatures we described above for all VOs. When zooming in to the five Delta clades and their associated lineage defining mutations, we found that theses mutations were also found in other VOs. This was particularly true for Delta D clade which is defined by four mutations common to other VOCs/VOIs (Table 1). This interpretation requires caution, since it does not consider the probability of independent acquisition of mutations due to the high mutation rate of the virus, or other technical artefacts of sequencing.

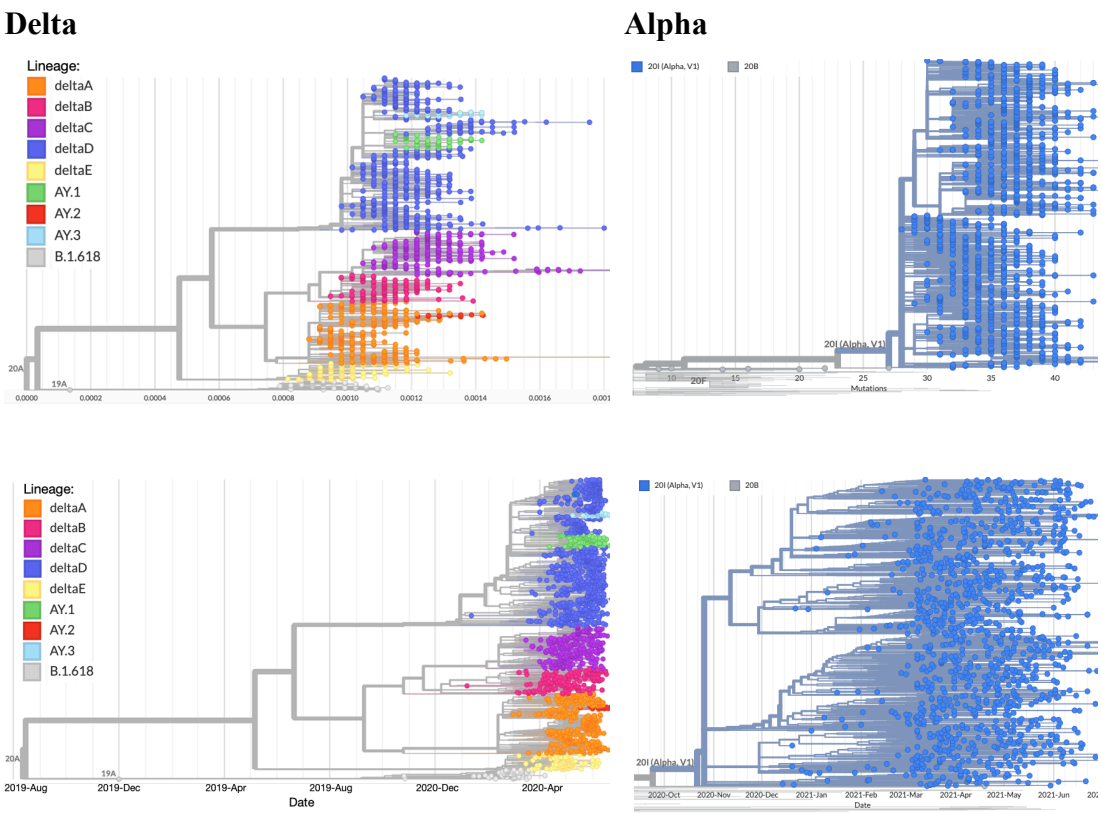

Figure 3. Phylogenies of the Delta and Alpha variants. The two upper panels are divergence-based phylogenies and the lower panels are time-aligned phylogenies. Clades in Delta are separated based on long branches or based on sub-variant designations (see main text) and color-coded accordingly. The B.1.618 was used as an outgroup for the Delta phylogeny, whereas 20B sequences were used as an outgroup for the Delta phylogeny. Substitutions defining the five clades of Delta A-E are specified in Table 1.

Table 1. Lineage defining mutations of Delta and its clades defined in this study.

	Mutation (genome)	Mutation type	Variant	% NS **	Independent emergence in other VOs ***
Delta	G210T	extragenic	5'UTR:210		
	C21618G	non-syn	S:T19R		none
	AGTTCA22029-	deletion	S:E156		none
	T22917G	non-syn	S:L452R		B.1.427, B.1.429, C36.3, C.37 (Lambda) (L452Q)
	C22995A	non-syn	S:T478K		none
	C23604G	non-syn	S:P681R		B.1.1.7, B.1.1.318 (P681H)
	G24410A	non-syn	S:D950N		B.1.621
	C25469T	non-syn	ORF3a:S26L		none
	T26767C	non-syn	M:I82T		B.1.1.318, C36.3, B.1.617.1 (Kappa) (I82C)
	T27638C	non-syn	ORF7a:V82A		none
	C27752T	non-syn	ORF7a:T120I		none
	A28271.	extragenic	3'UTR:28271		B.1.1.318, C.37 (Lambda), B.1.620
	A28461G	non-syn	N:D63G		none
	G28881T	non-syn	N:R203M		B.1. lineage (R203K)
	G29402T	non-syn	N:D377Y		none
	G29742T	extragenic	3'UTR:29742	1	none
A-D	G15451A	non-syn	NSP12b:G662S, ORF1b: G662S		none
	C16466T	non-syn	NSP13:P77L, ORF1b, P1000L		none
	GATTTC28248-	deletion	ORF8:D119-	1	none
A-C	C5184T *	non-syn	NSP3:P822L, ORF1a: P1640L		B.1.466.2
	C9891T	non-syn	NSP4:A446V, ORF1a: A23901V		B.1.1.318
	T11418C	non-syn	NSP6:V149A, ORF1a: V3718A	1	none
A	A5584G	syn	NSP3:T955T		none
	C11514T	non-syn	NSP6:T181I, ORF1a, T3750I		none
	C13019T	syn	NSP9:R111R		none
	C22227T	non-syn	S:A222V	0.5	none
B-C	C1191T *	non-syn	NSP2:P129L, ORF1A: P309L		none
	C1267T *	syn	NSP2:G154G		none
	T12946C	syn	NSP9:Y87Y		none
	A20262G	syn	NSP15:L214L		B.1.526 (Iota)
	C27739T	non-syn	ORF7a:L116F	0.4	none
C	C6539T	non-syn	NSP3:H1274Y, ORF1a: H2092Y		none
	C20320T	non-syn	NSP15:H234Y, ORF1b: H2285Y		none
	C24745T	syn	S:V1061V		none
	G29427A	non-syn	N:R385K	0.75	none
D	G4181T	non-syn	NSP3:A488S, ORF1a: A1306S		none
	C6402T	non-syn	NSP3:P1228L, ORF1a, P2046L		B.1.351.3 (Beta)
	C7124T	non-syn	NSP3:P1469S, ORF1a: P2287S		C.37 (Lambda)
	C8986T	syn	NSP4:D144D		none
	G9053T	non-syn	NSP4:V167L, ORF1a: V2390L		none
	C10029T	non-syn	NSP4:T492I, ORF1a: T3255I		C.37 (Lambda)
	A11201G	non-syn	NSP6:T77A, ORF1a: T3646A		B.1.617.1 (Kappa)
	A11332G	syn	NSP6:V120V		none
	C19220T	non-syn	NSP14:A394V, ORF1b: A1918V		none
	C27874T	non-syn	ORF7b:T40I		none
	G28916T	non-syn	N:G215C	0.82	none
E	C1191T *	non-syn	NSP2:P129L, ORF1a: P309L		none
	C1267T *	syn	NSP2:G154G		none
	C5184T *	non-syn	NSP3:P822L, ORF1a: P1640L		B.1.466.2*
	G9203A	non-syn	NSP4:D217N, ORF1a: D2980N		C.36.3
	T9678C	non-syn	NSP4:F375S, ORF1a: F3138S		none
	C11005A	non-syn	NSP6:H11Q, ORF1a: H3580Q		none
	A17496G	syn	NSP13:E420E		none
	A20396G	non-syn	NSP15:K259R, ORF1b: K2310R		B.1.617.1 (Kappa)
	A21792C	non-syn	S:K77T		none
	C28253T	syn	ORF8:F120F	0.7	Beta, Iota, Lambda (more)****

* Mutations shared by E clade and other clades
** % non-synonymous mutations, in coding regions only
*** VOs as defined by WHO (VOC/VOI – red, VUM - pink).
**** Site often masked from analyses (https://github.com/W-L/ProblematicSites_SARS-CoV2)

The prevalence of the five newly characterized Delta clades A-E was analyzed in several countries where the frequency of the Delta variant has been rapidly increasing over time [15] (Fig. S3). All countries exhibited a similar pattern, in which the Delta D clade has gained dominance over the other Delta clades (A,B,C,E) over time (Fig. 4A). In India, where this variant was first discovered, all the Delta clades initially co-occur, with the Delta D clade gaining dominance only in May. We note that the first sequences bearing the Delta D lineage defining mutations are from February-March 2021, and are predominantly from India, suggesting this clade was first created there.

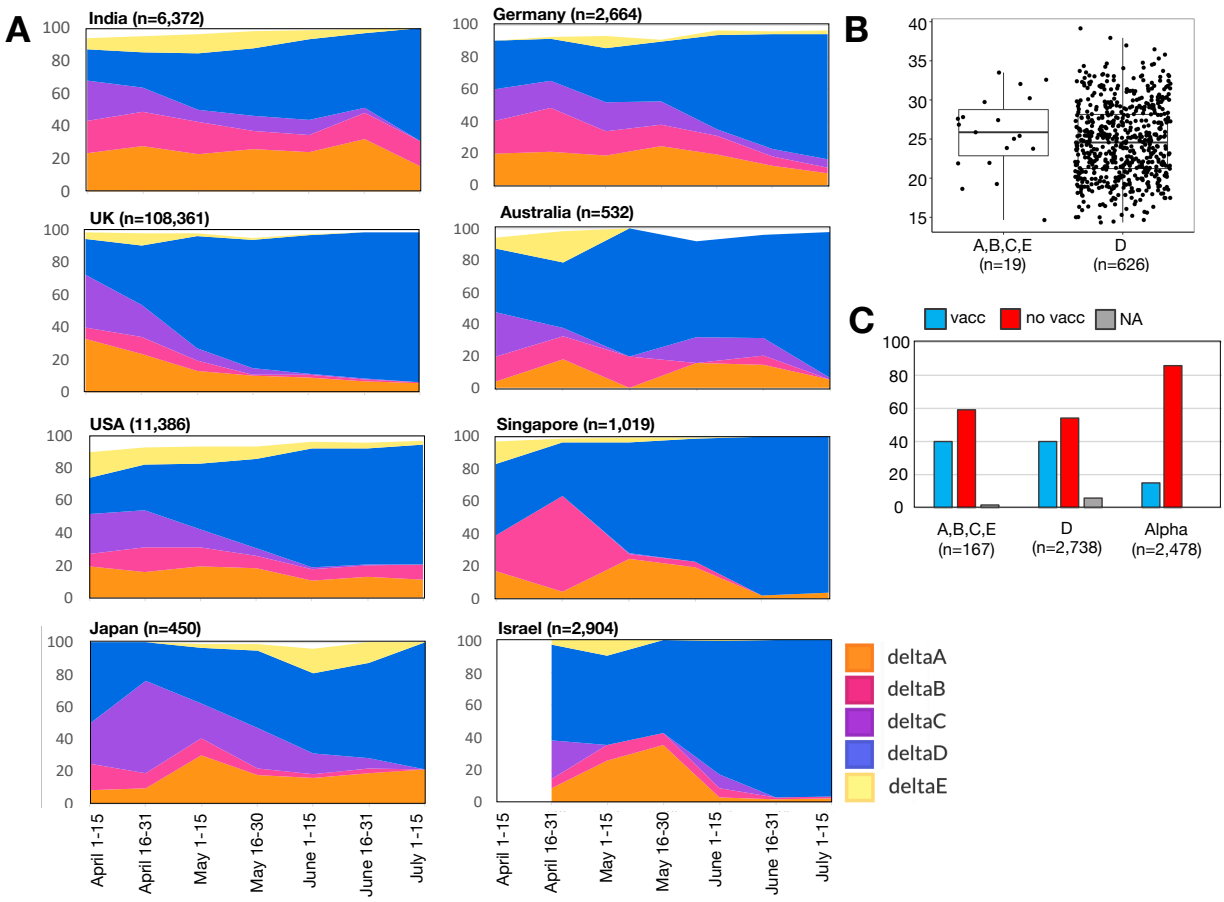

Figure 4. Prevalence and characteristics of Delta clades.
(A) Frequency of the five Delta clades A-E between April and July 2020 in countries where the Delta variant has rapidly increased. Frequencies were calculated per time-point, where non-complete/corrupted sequences that could not be classified as any of the Delta clades were removed from the analysis but not from the number of total sequences, hence not all frequencies sum up to 100%. The number of the total sequences evaluated per country is indicated in parentheses. (B) A comparison of cycle threshold (Ct) values from individuals infected by Delta D versus other clades of Delta, (C) breakdown of the proportion of vaccinated versus non- vaccinated individuals infected by Delta D, other clades of Delta and the Alpha variant.

Interrogation of Delta D infection characteristics

Next, we explored the characteristics of Delta infections in Israel, comparing Delta clade D sequences to other Delta clades. No significant differences were found in viral load, measured by the cycle threshold (Ct), between Delta D infections and infections from the other Delta clades (Fig. 4C), with the caveat that only a small number of samples had available Ct values for the non-clade D infections. Similarly, no striking differences were observed between Delta D and Delta A-E in age or gender (Table S2). We next focused on the proportion of vaccinated (two vaccine doses) and non-vaccinated individuals infected with the Delta variant. When comparing overall Delta clades with Alpha, it was notable that a higher proportion of vaccinated individuals become infected with Delta as compared to Alpha, whereas no notable differences were observed between Delta D and other Delta clades (Fig. 4C, Table S2). These results are line with reports on lower vaccine effectiveness against infection with regards to the Delta variant [25][26][27], but do not add any information why Delta D is more prevalent.

Discussion

The identification of new SARS-CoV-2 variants is justifiably causing constant trepidation around the world. However, our understanding of what drives the emergence and dynamics of these variants is somewhat lacking. Most VOs have displayed minor "blips", i.e., they increased in frequency rapidly in one location, yet this increase in frequency was also followed by a rapid decay. Only two VOCs - Alpha and Delta, displayed a more dramatic global pattern, increasing dramatically in frequency over most of the globe.

What makes these two variants unique? We first focus on non-synonymous substitutions associated with evasion from antibodies (Abs), which we denote as Abs-evasion substitutions. The Alpha variant stands out for its absence of such substitutions (with the exception of one amino-acid deletion at the NTD), which are abundant in all other VOs (Fig. 2C). This is line with the timeline during which Alpha began to spread, around the fall of 2020: the global population was not vaccinated, and it is not likely that recovered individuals constituted a large fraction of the global population. Thus, Abs-evasion substitutions bore no selective advantage in infections of naive individuals. On the other hand, we suggest that in other VOs spreading in 2020, Abs-evasion substitutions may have even incurred a fitness cost during infection of naive individuals, and this led to the limited

spread of most VOs, as for example occurred with the Beta variant in Israel [28]. In contrast to Alpha, the spread of the Delta VOC is currently occurring at a very different landscape, with an increase in immunized and recovered individuals, and thus Abs-evasion substitutions may confer a significant advantage. Accordingly, results herein and by others show more infections occurring in immunized individuals as compared to what was previously observed for the Alpha variant. At this point it should be noted that current reports from Israel and elsewhere suggest that the recent increase in observed breakthrough cases with the Delta variant are likely a combination of waning immunity [29] in addition to the ability of this variant to overcome the immune response to some extent [26]. This waning may have led to somewhat reduced vaccine effectiveness against infection [30], yet effectiveness against hospitalization and severe disease remain high [31].

We now discuss the enigmatic process of how VOs are created. When the unique long branch (representing a large accumulation of substitutions) of the Alpha VOC was first noted, it was suggested that Alpha arose in an immunocompromised individual chronically infected with SARS-CoV-2 [32][33]. It has been previously observed that during treatment of such individuals with convalescent plasma or with monoclonal antibodies, rapid evolution is observed, including the accumulation of various Spike amino-acid replacements that are prevalent in VOs [32]. Accordingly, one hypothesis that emerges is that all VOs were first created in immunocompromised individual infected with SARS-CoV-2, albeit this hypothesis is very difficult to confirm, since it is impossible to detect a "patient" zero from which VOs emerged.

The Delta variant, however, appears to be quite different in this context from other VOs, especially in comparison to Alpha. The tree topology of Delta is highly structured, suggesting that its spread was a slower and more step-wise process. The Delta clade includes five newly characterized clades, suggesting two possibilities: either Delta arose early on during the pandemic, and rounds of random genetic drift led to its separation into several clades, or - selection led to the formation of these clades. The high proportion of non-synonymous substitutions during the emergence of some, but not all of these clades supports selection, yet this is inconclusive. We further note that of the five clades A-E, Delta D seems to be repeatedly gaining dominance in various countries across the globe. We go on to discuss two hypotheses: the first is that the rise in frequency of Delta D is due to founder effects, and the second is that Delta D arose due to positive selection.

One possibility that arises, is that the increase in Delta D is a combination of founder effects together with a shift in the landscape of immunized and recovered individuals across the globe. Accordingly, Delta D in itself has no selective advantage over other Delta clades; yet all Delta clades bear a selective advantage over the Alpha variant and other variants that predated Delta. In particular, it is possible that the advantage Delta bears is in its ability to overcome some of the defenses of immunized individuals as compared to Alpha, as evident from the data herein and elsewhere [34][35][36]. Thus, Delta D first increased in frequency over the period of March to May 2021 in India merely due to genetic drift, and later as the proportion of immunized/recovered individuals increased, the already prevalent Delta D took over. The caveat in this hypothesis is that infections from all Delta clades are evident already in April across the globe. It is possible that these are a biased sample from incoming travelers who did not go on to create transmission chains.

A second hypothesis is that Delta D may be under positive selection. In line with this, 82% of the lineage specific mutations of this clade are non-synonymous, similar to the proportion that characterizes VOs (Table 1, Fig. 2A). However, this clade lacks additional substitutions in S, and is characterized by seven amino-acid replacements in the ORF1a/b polyprotein. This is particularly perplexing as the lineage defining mutations of the main Delta lineage are completely devoid of mutations in ORF1a/b (Table 1). An additional non-synonymous substitution is evident in the ORF7b gene (T40I) and in the N gene (G215C), quite proximal to the 203-205 region discussed above. Notably, four of the eleven substitutions unique to the Delta D clade are observed in other VOs (Table 1), suggesting that selection may have driven the emergence of this clade. If so, we suggest that the first mutations that were fixed in the basal Delta lineage, which are mostly in the S and N genes, bore the highest selective advantage. We suggest that the additional mutations fixed in Delta D (and possibly in other clades), bore a smaller selective advantage, and were hence contingent on a larger viral population size, which was enabled due to the S and N gene mutations. This mode of step-wise evolution, from large effect to small effect mutations, is in line with evolutionary theory and has been previously observed in other RNA viruses [37], and it remains to be confirmed whether and how this pattern will be recapitulated in the future.

To summarize, we have used a comparative approach to detect a unique mode of evolution present in Delta. This step-wise mode of evolution characterized both the formation of the Delta D clade, and its subsequent spread, and is in stark contrast to the evolution observed in other VOs. In particular, the global increase in Delta frequency has occurred concurrently with the increase in Delta D, suggesting that what is now labelled as "Delta" worldwide is actually specifically the Delta D clade.

Methods

Ethics statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Sheba Medical Center institutional review board (7045-20-SMC). For the Israel cohort, patient consent was waived as the study used remains of clinical samples and the analysis used anonymous clinical data.

Study cohorts

Global sample sequences used in this study were downloaded from GISAID (global initiative on sharing all influenza data) [38]. Sequence and patient data for the Israel Delta cohort was obtained via the Israel Consortium of SARS-CoV-2 sequencing, a national surveillance system to identify circulating and imported variants via sequencing, established in December 2020 by Israel Ministry of Health and Public Health Services. Details on sequencing and bioinformatics are given below.

Construction of phylogenetic trees

Phylogenetic trees were constructed using NextStrain's Augur pipeline [18]. Sequences were aligned to SARS-CoV-2 reference genome (NC_045512.2) using MAFFT [39], and a time-resolved phylogenetic tree was constructed with IQ-Tree [40] and TreeTime [41] under the generalized time reversible (GTR) substitution model and visualized with auspice [18]. Lineage nomenclature was attained from Pangolin Lineages [24].
Dating of internal nodes are reported based on NextStrain, which in turns relies on IQ-Tree [40] ancestral sequence reconstruction, dating, and assignment of confidence intervals for these dates.

Comparison of lineage defining mutations for VOs and non-VOs

Lineage defining mutations of VOs were extracted from CoVariants [15] and outbreak.info [42]. non-VO lineages were selected from Pango lineages [43][44] based on the following criteria: at least 1,000 sequences were available, and earliest data of detection varied between March and November 2020. Of these, twelve lineages were randomly chosen with an emphasis on lineages prevalent across different continents. All lineages were derived from the B.1 lineage. For each lineage, we randomly sampled at least 250 sequences and focused on substitutions present in more than 90% of the sequences, yet absent from the lineage defining mutations of the ancestral lineages. Substitutions were then classified as extragenic, synonymous, NS, or deletions based on data from ViruSurf [45]. When estimating the fraction of NS substitutions, we grouped deletions and NS, and calculated their fraction out of both NS and synonymous. A one-sided t-test with unequal variances was used to assess the statistical significance of the higher fraction of NS in VOs.

Characterization of Delta sub-clades

When observing the phylogenetic tree of the Delta variant in NextStrain's global analysis (www.nextstrain.org, July 24th 2021), we observed a strong separation into five distinct clades, which was recapitulated in both global and continent based builds. The five clades were all based on internal branches with at least three substitutions. Since NextStrain is based on a sample of sequences, we next verified the veracity of the clades by downloading ~3000 sequences identified as B.1.617.2 between March 1 - July 15 2021, which were collected from Virusurf [45], without "N"s (undetermined nucleotide), to facilitate identification of *bona fide* substitutions. A representative sample from around the globe was further analyzed to avoid biases towards countries who sequence more intensely. Substitutions, as compared to the SARS-CoV-2 reference NC_045512.2, were identified in each sequence using Coronapp [46], and clustered based on similar unique substitutions that are not part of the B.1.617.2 lineage defining mutations. A similar search using ViruSurf [45] was conducted for additional clade signature substitutions, to validate each clade.

Library preparation, sequencing and processing

Israel cohort samples were processed as follows: RNA was extracted from 200µl respiratory samples with the MagNA PURE 96 (Roche, Germany), according to the manufacturer instructions. Libraries were prepared using COVID-seq library preparation kit, as per manufacturer's instructions (Illumina). Library validation and mean fragment

size was determined by TapeStation 4200 via DNA HS D1000 kit (Agilent). Libraries were pooled, denatured and diluted to 10pM and sequenced on NovaSeq (Illumina). Fastq files were subjected to quality control using FastQC (www.bioinformatics.babraham.ac.uk/ projects/fastqc/) and MultiQC [47] and low-quality sequences were filtered using trimmomatic [48]. Sequences were mapped to the SARS-CoV-2 reference genome (NC_045512.2) with Burrows-Wheeler aligner (BWA) mem [49]. Resulting BAM files were sorted and indexed using SAMtools suite [50]. Breadth and depth of sequencing was calculated from sorted BAM files using a custom python script. Consensus Fasta sequences were assembled for each sample using iVar (https://andersen-lab.github.io/ivar/html/index.html), with positions <5 nucleotides determined as Ns. Multiple alignment of sample sequences with the reference Wuhan sequence (NC_045512.2) was performed with MAFFT using default parameters [39]. All data generated via the Israel Consortium of SARS-CoV-2 sequencing, including the Israel cohort data in this manuscript, is regularly deposited and available in GISAID.

References

[1] A. Wu *et al.*, "Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China," *Cell Host Microbe*, vol. 27, no. 3, pp. 325–328, Mar. 2020, doi: 10.1016/j.chom.2020.02.001.

[2] X. Li *et al.*, "Transmission dynamics and evolutionary history of 2019-nCoV," *J. Med. Virol.*, vol. 92, no. 5, pp. 501–511, May 2020, doi: 10.1002/jmv.25701.

[3] S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, and G. Baele, "Temporal signal and the phylodynamic threshold of SARS-CoV-2," *Virus Evol.*, vol. 6, no. 2, Jul. 2020, doi: 10.1093/ve/veaa061.

[4] Y. M. Bar-On, A. Flamholz, R. Phillips, and R. Milo, "SARS-CoV-2 (COVID-19) by the numbers," *Elife*, vol. 9, Apr. 2020, doi: 10.7554/eLife.57309.

[5] M. A. Martin, D. VanInsberghe, and K. Koelle, "Insights from SARS-CoV-2 sequences," *Science (80-.).*, vol. 371, no. 6528, pp. 466–467, Jan. 2021, doi: 10.1126/science.abf3995.

[6] L. van Dorp, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, and F. Balloux, "No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2," *Nat. Commun.*, vol. 11, no. 1, p. 5986, Dec. 2020, doi: 10.1038/s41467-020-19818-2.

[7] P. Arora, S. Pöhlmann, and M. Hoffmann, "Mutation D614G increases SARS-CoV-2 transmission," *Signal Transduct. Target. Ther.*, vol. 6, no. 1, p. 101, Dec. 2021, doi: 10.1038/s41392-021-00502-w.

[8] J. A. Plante *et al.*, "Spike mutation D614G alters SARS-CoV-2 fitness," *Nature*, vol. 592, no. 7852, pp. 116–121, Apr. 2021, doi: 10.1038/s41586-020-2895-3.

[9] B. Korber *et al.*, "Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus," *Cell*, Jul. 2020, doi: 10.1016/j.cell.2020.06.043.

[10] B. Zhou *et al.*, "SARS-CoV-2 spike D614G change enhances replication and transmission," *Nature*, vol. 592, no. 7852, pp. 122–127, Apr. 2021, doi: 10.1038/s41586-021-03361-1.

[11] Meera Chand; Susan Hopkins; Gavin Dabrera; Christina Achison; Wendy Barclay; Neil Ferguson; Erik Volz; Nick Loman; Andrew Rambaut; Jeff Barrett, "Investigation of novel SARS-COV-2 variant Variant of Concern 202012/01," 2020, [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf.

[12] N. G. Davies *et al.*, "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England," *Science (80-.).*, vol. 372, no. 6538, p. eabg3055, Apr. 2021, doi: 10.1126/science.abg3055.

[13] E. Volz *et al.*, "Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England," *Nature*, vol. 593, no. 7858, pp. 266–269, May 2021, doi: 10.1038/s41586-021-03470-x.

[14] X. Deng *et al.*, "Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant.," *Cell*, Apr. 2021, doi: 10.1016/j.cell.2021.04.025.

[15] E. B. Hodcroft, "CoVariants: SARS-CoV-2 Mutations and Variants of Interest," 2021.

[16] A. Wu *et al.*, "One year of SARS-CoV-2 evolution," *Cell Host Microbe*, vol. 29, no. 4, pp. 503–507, Apr. 2021, doi: 10.1016/j.chom.2021.02.017.

[17] F. Amanat *et al.*, "SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD, and S2," *Cell*, vol. 184, no. 15, pp. 3936-3948.e10, Jul.

2021, doi: 10.1016/j.cell.2021.06.005.
[18] J. Hadfield *et al.*, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, Dec. 2018, doi: 10.1093/bioinformatics/bty407.
[19] C. E. M. R. D. W. S. G. S. M. T. C.-19 G. U. (COG-U. consortium; T. I. de S. Matthew D Parker; Benjamin B. Lindsey; Dhruv R Shah; Sharon Hsu; Alexander J Keeley; David G Partridge; Shay Leary; Alison Cope; Amy State; Katie Johnson; Nasar Ali; Rasha Raghei; Joe Heffer; Nikki Smith; Peijun Zhang; Marta Gallis; Stavroula F Louka; Hai, "Altered Subgenomic RNA Expression in SARS-CoV-2 B.1.1.7 Infections," *bioRxiv*, 2021, doi: https://doi.org/10.1101/2021.03.02.433156.
[20] L. G. T. B.-K. R. Z.-A. P. P. B. V. X. W. U. R. T. O. B. S. R.-H. C. B. Krogan, "Evolution of enhanced innate immune evasion by the SARS-CoV-2 B.1.1.7 UK variant," *bioRxiv*, 2021, doi: https://doi.org/10.1101/2021.06.06.446826.
[21] S. Wang *et al.*, "Targeting liquid–liquid phase separation of SARS-CoV-2 nucleocapsid protein promotes innate antiviral immunity by elevating MAVS activity," *Nat. Cell Biol.*, vol. 23, no. 7, pp. 718–732, Jul. 2021, doi: 10.1038/s41556-021-00710-0.
[22] E. Nikolakaki and T. Giannakouros, "SR/RS Motifs as Critical Determinants of Coronavirus Life Cycle," *Front. Mol. Biosci.*, vol. 7, Aug. 2020, doi: 10.3389/fmolb.2020.00219.
[23] Y. Cong *et al.*, "Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle," *J. Virol.*, vol. 94, no. 4, Jan. 2020, doi: 10.1128/JVI.01925-19.
[24] A. R. Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, JT McCrone, Chris Ruis, Khali Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, David Aanensen, Eddie Holmes, Oliver Pybus, "pangolin: lineage assignment in an emerging pandemic as an epidemiological tool."
[25] Y. Lustig *et al.*, "Neutralising capacity against Delta (B.1.617.2) and other variants of concern following Comirnaty (BNT162b2, BioNTech/Pfizer) vaccination in health care workers, Israel," *Eurosurveillance*, vol. 26, no. 26, Jul. 2021, doi: 10.2807/1560-7917.ES.2021.26.26.2100557.
[26] D. Planas *et al.*, "Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization," *Nature*, Jul. 2021, doi: 10.1038/s41586-021-03777-9.
[27] P. Mlcochova *et al.*, "SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion," *Nature*, Sep. 2021, doi: 10.1038/s41586-021-03944-y.
[28] T. Kustin *et al.*, "Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2-mRNA-vaccinated individuals," *Nat. Med.*, Jun. 2021, doi: 10.1038/s41591-021-01413-7.
[29] A. H. Yair Goldberg, Micha Mandel, Yinon M. Bar-On, Omri Bodenheimer, Laurence Freedman, Eric J. Haas, Ron Milo, Sharon Alroy-Preis, Nachman Ash, "Waning immunity of the BNT162b2 vaccine: A nationwide study from Israel," *medRxiv*, 2021, doi: https://doi.org/10.1101/2021.08.24.21262423.
[30] A. Fowlkes, M. Gaglani, K. Groover, M. S. Thiese, H. Tyner, and K. Ellingson, "Effectiveness of COVID-19 Vaccines in Preventing SARS-CoV-2 Infection Among Frontline Workers Before and During B.1.617.2 (Delta) Variant Predominance — Eight U.S. Locations, December 2020–August 2021," *MMWR. Morb. Mortal. Wkly. Rep.*, vol. 70, no. 34, pp. 1167–1169, Aug. 2021, doi: 10.15585/mmwr.mm7034e4.
[31] A. Sheikh, J. McMenamin, B. Taylor, and C. Robertson, "SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness,"

Lancet, vol. 397, no. 10293, pp. 2461–2462, Jun. 2021, doi: 10.1016/S0140-6736(21)01358-1.

[32] S. A. Kemp *et al.*, "SARS-CoV-2 evolution during treatment of chronic infection," *Nature*, vol. 592, no. 7853, pp. 277–282, Apr. 2021, doi: 10.1038/s41586-021-03291-y.

[33] D. Frampton *et al.*, "Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study," *Lancet Infect. Dis.*, Apr. 2021, doi: 10.1016/S1473-3099(21)00170-5.

[34] B. E. Y. Po Ying Chia, Sean Wei Xiang Ong, Calvin J Chiew, Li Wei Ang, Jean-Marc Chavatte, Tze-Minn Mak, Lin Cui, Shirin Kalimuddin, Wan Ni Chia, Chee Wah Tan, Louis Yi Ann Chai, Seow Yen Tan, Shuwei Zheng, Raymond Tzer Pin Lin, Linfa Wang, Yee-Sin Leo, Vernon J L, "Virological and serological kinetics of SARS-CoV-2 Delta variant vaccine-breakthrough infections: a multi-center cohort study," *medRxiv*, 2021, doi: https://doi.org/10.1101/2021.07.28.21261295.

[35] R. K. G. Petra Mlcochova, Steven Kemp, Mahesh Shanker Dhar, Guido Papa, Bo Meng, Swapnil Mishra, Charlie Whittaker, Thomas Mellan, Isabella Ferreira, Rawlings Datir, Dami A. Collier, Sujeet Singh, Rajesh Pandey, Robin Marwal, Meena Datta, Shantanu Sengupta, Kalaia, "SARS-CoV-2 B.1.617.2 Delta variant emergence and vaccine breakthrough," *medRxiv*, 2021, doi: 10.21203/rs.3.rs-637724/v1.

[36] J. Lopez Bernal *et al.*, "Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant," *N. Engl. J. Med.*, p. NEJMoa2108891, Jul. 2021, doi: 10.1056/NEJMoa2108891.

[37] A. Stern *et al.*, "The Evolutionary Pathway to Virulence of an RNA Virus," *Cell*, vol. 169, no. 1, pp. 35-46.e19, Mar. 2017, doi: 10.1016/j.cell.2017.03.013.

[38] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data - from vision to reality.," *Euro Surveill.*, vol. 22, no. 13, 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[39] K. Katoh, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.

[40] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–74, Jan. 2015, doi: 10.1093/molbev/msu300.

[41] P. Sagulenko, V. Puller, and R. A. Neher, "TreeTime: Maximum-likelihood phylodynamic analysis," *Virus Evol.*, vol. 4, no. 1, Jan. 2018, doi: 10.1093/ve/vex042.

[42] M. A. Haendel *et al.*, "The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment," *J. Am. Med. Informatics Assoc.*, vol. 28, no. 3, pp. 427–443, Mar. 2021, doi: 10.1093/jamia/ocaa196.

[43] A. Rambaut *et al.*, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nat. Microbiol.*, vol. 5, no. 11, pp. 1403–1407, Nov. 2020, doi: 10.1038/s41564-020-0770-5.

[44] Á. O'Toole *et al.*, "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2," *Wellcome Open Res.*, vol. 6, p. 121, May 2021, doi: 10.12688/wellcomeopenres.16661.1.

[45] A. Canakoglu, P. Pinoli, A. Bernasconi, T. Alfonsi, D. P. Melidis, and S. Ceri, "ViruSurf: an integrated database to investigate viral sequences," *Nucleic Acids*

Res., vol. 49, no. D1, pp. D817–D824, Jan. 2021, doi: 10.1093/nar/gkaa846.
[46] D. Mercatelli, L. Triboli, E. Fornasari, F. Ray, and F. M. Giorgi, "Coronapp: A web application to annotate and monitor SARS-CoV-2 mutations.," *J. Med. Virol.*, vol. 93, no. 5, pp. 3238–3245, May 2021, doi: 10.1002/jmv.26678.
[47] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016, doi: 10.1093/bioinformatics/btw354.
[48] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data.," *Bioinformatics*, vol. 30, no. 15, pp. 2114–20, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
[49] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics*, vol. 25, no. 14, pp. 1754–60, Jul. 2009, doi: 10.1093/bioinformatics/btp324.
[50] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009, doi: 10.1093/bioinformatics/btp352.