The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape

Darren P Martin[1]*, Steven Weaver[2], Houryiah Tegally[3], Emmanuel James San[3], Stephen D Shank[2], Eduan Wilkinson[3], Jennifer Giandhari[3], Sureshnee Naidoo[3], Yeshnee Pillay[3], Lavanya Singh[3], Richard J Lessells[3], NGS-SA[4&], COVID-19 Genomics UK (COG-UK)[5$], Ravindra K Gupta[6,7], Joel O Wertheim[8], Anton Nekturenko[9], Ben Murrell[10], Gordon W Harkins[11], Philippe Lemey[12], Oscar A MacLean[13], David L Robertson[13],Tulio de Oliveira[3,14]*, Sergei L Kosakovsky Pond[2]*

Affiliations

[1]Institute of Infectious Diseases and Molecular Medicine, Division Of Computational Biology, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa.
[2]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Pennsylvania, USA
[3]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine & Medical Sciences, University of KwaZulu- Natal, Durban, South Africa
[4]http://www.krisp.org.za/ngs-sa/ngs-sa_network_for_genomic_surveillance_south_africa/
[5]https://www.cogconsortium.uk
[6]Clinical Microbiology, University of Cambridge, Cambridge, UK
[7]Africa Health Research Institute, KwaZulu-Natal, South Africa
[8]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[9]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, Pennsylvania, USA.
[10]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden.
[11]South African Medical Research Council Capacity Development Unit, South African National Bioinformatics Institute, University of the Western cape, Bellville, South Africa.
[12]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium.
[13]MRC-University of Glasgow Centre for Virus Research, Scotland, UK.
[14]Department of Global Health, University of Washington, Seattle, US.

*Corresponding authors.

[&]Full list of consortium names and affiliations are in the appendix

[$]Full list of consortium names and affiliations are in the appendix

Abstract

The emergence and rapid rise in prevalence of three independent SARS-CoV-2 “501Y lineages", B.1.1.7, B.1.351 and P.1, in the last three months of 2020 has prompted renewed concerns about the evolutionary capacity of SARS-CoV-2 to adapt to both rising population immunity, and public health interventions such as vaccines and social distancing. Viruses giving rise to the different 501Y lineages have, presumably under intense natural selection following a shift in host environment, independently acquired multiple unique and convergent mutations. As a consequence all have gained epidemiological and immunological properties that will likely complicate the control of COVID-19. Here, by examining patterns of mutations that arose in SARS-CoV-2 genomes during the pandemic we find evidence of a major change in the selective forces acting on immunologically important SARS-CoV-2 genes (such as N and S) that likely coincided with the emergence of the 501Y lineages. In addition to involving continuing sequence diversification, we find evidence that a significant portion of the ongoing adaptive evolution of the 501Y lineages also involves further convergence between the lineages. Our findings highlight the importance of monitoring how members of these known 501Y lineages, and others still undiscovered, are convergently evolving similar strategies to ensure their persistence in the face of mounting infection and vaccine induced host immune recognition.

Introduction

In the first eleven months of the SARS-CoV-2 pandemic (December 2019 - October 2020), the evolution of the virus worldwide was in the context of a highly susceptible new host population [1,2]. Other than the early identification of the D614G substitution in the viral spike protein [3–5] that probably increased its transmissibility without impacting its pathogenesis, few mutations were epidemiologically significant and the evolutionary dynamics of the virus was predominantly characterized by a mutational pattern of slow and selectively-neutral random genetic drift. This behavior is consistent with exponential growth in a population of naive susceptible hosts that do not exert significant selective pressures on the pathogen prior to transmission events [2]. Past pandemics and long-term evolutionary dynamics of RNA viruses attest to the fact that such an evolutionary “lull” does not necessarily last. Indeed, in late 2020, three relatively divergent SARS-CoV-2 lineages emerged in rapid succession: (i) B.1.1.7 or 501Y.V1 which will hereafter be referred to as V1 [6] (ii) B.1.351 or 501Y.V2 which will hereafter be referred to as V2 [7] and (iii) P.1 or 501Y.V3 which will hereafter be referred to as V3 [8].

Viruses in each of the three lineages (which will hereafter be collectively referred to as 501Y lineages) have multiple signature (or lineage defining) deletions and amino acid

changing substitutions (Figure 1), many of which impact key domains of the spike protein: the primary target of both infection and vaccine induced immune responses. Prior to this, while many distinct spike mutations had been observed, all circulating SARS-CoV-2 lineages were defined by small numbers of mutations. All of the 501Y lineages also have significantly altered phenotypes: increased human ACE2 receptor affinity (V1, V2 and V3) [9–11], increased transmissibility (V1 and V2) [12–14], increased capacity to overcome prior infection and/or vaccination-induced immunity (V2 and V3) [15–19] and associations with increased virulence (V1) [20]. Why did the heavily mutated 501Y lineages all arise on different continents at almost the same time? Was this due to an intrinsic property of SARS-CoV-2 and its mutability changing, or was it a shift in the host selective environment extrinsic to the virus?

Evidence that natural selection has played a pivotal role in the emergence of V1, V2 and V3 can be found in the remarkable patterns of independently evolved convergent mutations that have arisen within the members of these lineages (Figure 1). One of the most striking of these parallel changes is a nine nucleotide deletion between genome coordinates 11288 and 11296 (here and hereafter all nucleotide and amino acid coordinates refer to the GenBank reference genome NC_045512). This deletion is within the portion of ORF1ab that encodes non-structural protein 6 (nsp6): a component of the SARS-CoV-2 membrane-tethered replication complex that likely influences the formation and maturation of autophagosomes [21], and decreases the effectiveness of host innate antiviral defences by reducing the responsiveness of infected cells to type I interferons [22]. Independently evolved deletions at this site have been repeatedly found prior to the emergence of V1, V2 and V3 and it is also found with other V1 signature deletion mutations in the newly reported SARS-CoV-2 lineage, B.1.525 (https://github.com/cov-lineages/pango-designation/issues/4). Although the biological consequences of the 11288 to 11296 deletion remain unknown, its convergent evolution implies that, in the context of the B.1.525 and 501Y lineages at least, it is likely highly adaptive.

Additionally, there are four convergent spike gene mutations that are each shared between members of different 501Y lineages. Almost all the spike genes of sequences in these lineages carry the N501Y mutation at a key receptor binding domain (RBD) site that increases the affinity of the spike protein for human ACE2 receptors by ~3.5 fold [9,10]. The vast majority of V2 and V3 variants and some more recent V1 variants also have a spike E484K mutation. Whereas in the presence of 501N, 484K has a modest positive impact on ACE2 binding [9], when present with 501Y, these mutations together synergistically increase ACE2-RBD binding affinity ~12.7 fold [10,11]. Crucially, E484K and other mutations at S/484 also frequently confer protection from neutralization by both convalescent sera [23], vaccine elicited antibodies [17,24–26], and some monoclonal antibodies [23,26]. There is therefore increasing evidence that viruses carrying the E484K

mutation (with or without 501Y) will be able to more frequently infect both previously infected [27] and vaccinated individuals [17,19,24].

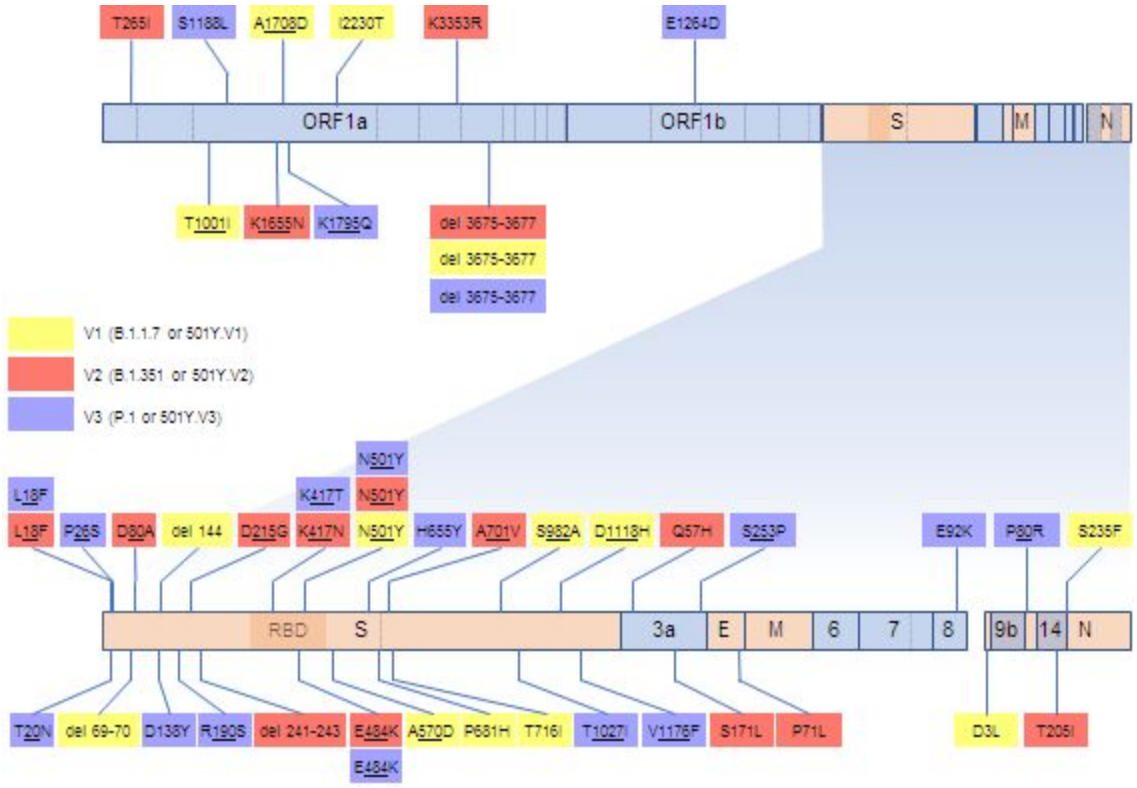


Figure 1. SARS-CoV-2 genome map indicating the locations and encoded amino acid changes of what we considered here to be signature mutations of V1, V2 and V3 sequences. Genes represented with light blue blocks encode non-structural proteins and genes in orange encode structural proteins:S encodes the spike protein, E the envelope protein, M the matrix protein, and N the nucleocapsid protein. Within the S-gene, the receptor binding domain (RBD) is indicated by a darker shade and the site where the S protein is cleaved into two subunits during priming for receptor binding and cell entry is indicated by a dotted vertical line.

A third RBD site that is mutated in both V2 and V3 is S/417. Whereas V2 sequences generally carry a K417N mutation, V3 sequences carry a K417T mutation. Both the K417N and K417T mutations can reduce the affinity of spike for ACE2, particularly in conjunction with the 501Y and E484K mutations [24], but both also have a moderately positive impact on spike expression [9] and these and other mutations at S/417 provide modest protection from neutralization by some convalescent sera [23,26], vaccine induced antibodies [26] and some neutralizing monoclonal antibodies [16,23,26].

A fourth spike gene mutation that is shared by ~42% of V2 sequences and by all V3 sequences is L18F. This amino acid change is predicted to have a modest impact on the structure of spike [28] and also protects from some neutralizing monoclonal antibodies [29]. Viruses carrying the L18F mutation have been detectably increasing in prevalence since the start of the pandemic such that they now account for ~10% of sampled SARS-CoV-2 sequences.

These five convergent mutations in different rapidly spreading SARS-CoV-2 lineages is compelling evidence that they each, either alone or in combination, provide some significant fitness advantage. The individual and collective fitness impacts of the other signature mutations in V1, V2 and V3 remain unclear. A key way to infer the fitness impacts of these mutations is to examine patterns of synonymous and non-synonymous substitutions at the codon sites where the mutations occurred [30]. Specifically, it is expected that the most biologically important of these mutations will have occurred at codon sites that display substitution patterns that are dominated by non-synonymous mutations (i.e. mutations that alter encoded amino acid sequences); patterns that are indicative of positive selection.

Here, using a suite of phylogenetics-based natural selection analysis techniques, we examine patterns of positive selection within the protein coding sequences of viruses in the V1, V2 and V3 lineages to identify the mutations that have most likely contributed to the increased adaptation of these lineages. We find that the emergence of the 501Y lineages coincided with a marked global change in positive selection signals, indicative of a general shift in the selective environment within which SARS-CoV-2 is evolving. Against this backdrop the 501Y lineages all display evidence of substantial ongoing adaptation that in many cases involves mutations that converge on the signature mutations of other 501Y lineages, but also involve multiple other convergent mutations at non-signature sites: a pattern which suggests both that viruses in all three lineages may be climbing very similar adaptive peaks, and, therefore, that viruses in all three lineages are likely in the process of converging on a similar adaptive endpoint.

Results and Discussion

There has been a recent detectable shift in selective pressures acting on circulating SARS-CoV-2 variants

Analyses of positive selection on SARS-CoV-2 genomes undertaken prior to the emergence of 501Y lineages revealed mutational patterns dominated by neutral evolution [2]. There were, however, indications that some sites in the genome had experienced episodes of positive selection [3–5]. Through regular analyses of global GISAID data (30), starting in March 2020, we tracked the extent and location of positive

and negative selective pressures on SARS-CoV-2 genomes (Figure 2). The power of these analyses to detect evidence of selection acting on individual codon sites progressively increased throughout 2020 with increasing numbers of sampled genome sequences and sequence diversification.

Even accounting for this expected increased power of detection, it is evident that a significant shift in selective pressures occurred ~11 months after SARS-CoV-2 cases were first reported in Wuhan City in December 2019. Specifically, during November 2020 this change in selection pressures manifested in substantial increases in the numbers of SARS-CoV-2 codon sites that were detectably evolving under both positive and negative selection. This increase accelerated through January 2021, with sites found to be evolving under diversifying positive selection in S, N, and ORF3a (MEME[31] or FEL [30] selection detection method $p \leq 0.0001$) rapidly increasing in density from a baseline of between zero and four codons per Kb in November 2020, to 10 or more codons per Kb at the beginning of 2021 (Figure 2A). This sudden increase in the numbers of sites that were detectably evolving under positive selection coincided with epidemic surges in multiple parts of the world in both hemispheres, many of which were driven by the emerging V1 and V2 lineages.

Among the 37 signature mutation sites in V1, V2 and V3 (Figure 1), 14 were detectably evolving under positive selection prior to December 2020 whereas this number increased to 31 by January 2021 (Figure 2C). The only signature mutation shared between any of the three 501Y lineages that was detectably evolving under positive selection before December 2020 was S/18 which was weakly detected for the first time in October.

Our regular tracking of positively selected SARS-CoV-2 codon sites prior to November 2020 therefore yielded no clear indications that non-synonymous substitutions at the crucial RBD sites S/417, S/484 or S/501 (the other key convergent signature mutation sites in the 501Y lineages), provided SARS-CoV-2 with any substantial fitness advantages in the first 11 months of the pandemic. Instead, the sporadic weak selection signals that these analyses yielded between July and November were of adaptive amino acid substitutions in the Spike N-terminal domain (S/18, and the V3 signature site, S/26), near the furin cleavage site (the V3 signature site, S/655), and in the C-terminal domain (the V3 signature site, S/1027, and the V1 signature site, S/1118). Conversely, for much of the latter half of 2020 the relatively strong and consistently detected selection signals at the V1 signature site, N/3, and the V2 signature sites, ORF1ab/265 (nsp2 codon 85) and ORF3a/57, clearly indicated that some substitutions at these sites were likely adaptations.

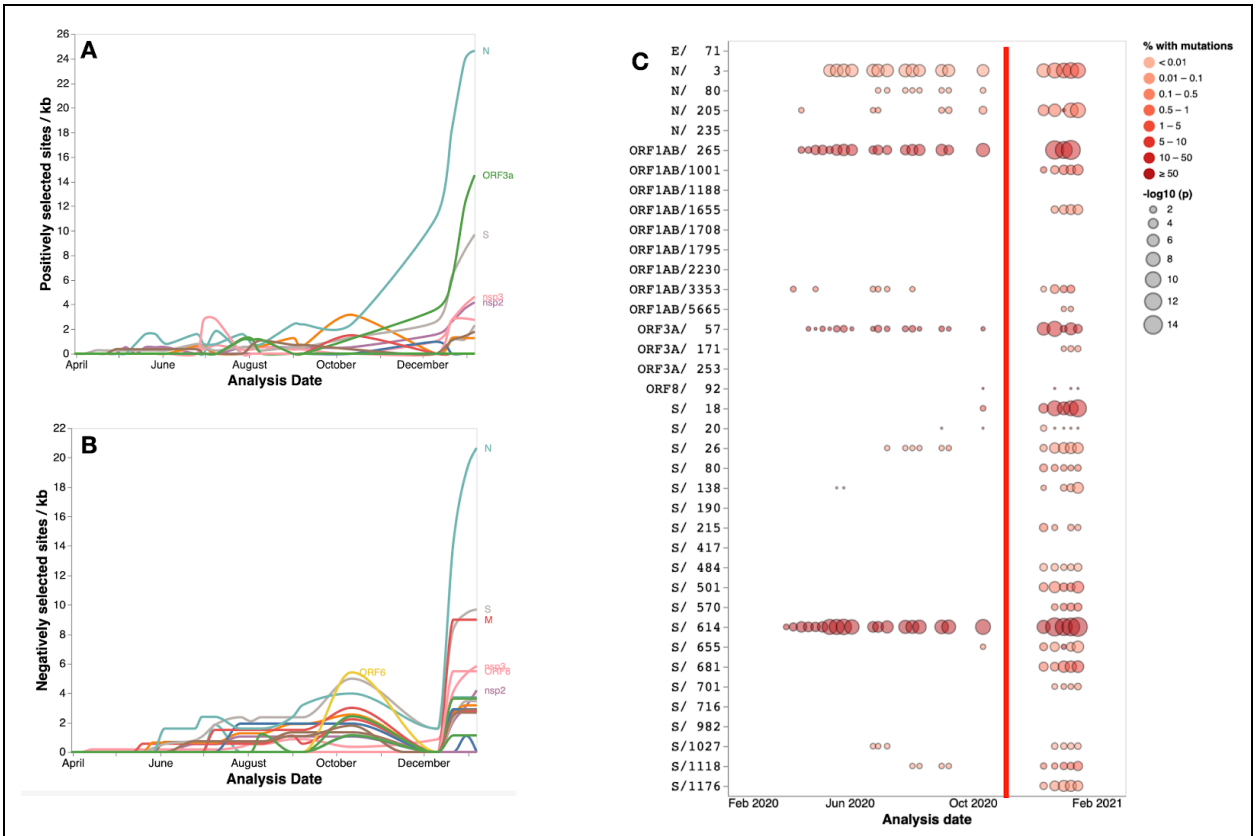


Figure 2. Signals of positive and negative selection at individual codon sites that were detectable with either the MEME or FEL methods at different times between March 2020 and January 2021. **A/B**. The gene-by-gene per Kb density of codons detectably evolving under positive/negative selection between March 2020 and January 2021. **C** Signals of positive selection detected at 37 V1, V2 and V3 signature mutation sites between March 2020 and January 2021. Also included for reference is S/614, the site of the D614G mutation that is present in all three of the 501Y lineages. For the circles the intensity of red indicates the proportion of analysed sequences that have mutations at a given site and the size of the circle indicates the statistical significance of the MEME and IFEL positive selection tests. The vertical red line indicates the 1st November 2020.

Taken together, these patterns of detectable selection suggest that the adaptive value of signature 501Y lineage RBD mutations may have only manifested after a selective shift that occurred shortly before November 2020.

Signals of ongoing selection within the V1, V2 and V3 lineages

We collected all sequences assigned to B.1.1.7 (V1), B.1.351 (V2), and P1 (V3) PANGO lineages [32] in GISAID [33] as of Feb 02, 2021, and subjected them to a comprehensive battery of gene- and codon-level selection tests (see Methods).

We first tested the 501Y lineage sequences to determine whether, at the level of individual genes, they were evolving under measurably different selective pressures than other SARS-CoV-2 lineages. To each of the individual V1, V2, and V3 gene datasets we added algorithmically-selected (see Methods) non-501Y lineage reference sequences that were representative of historical SARS-CoV-2 diversity (sampled prior to Oct 15th, 2020). We then analysed each of these datasets to compare the selective processes operating on the reference sequences relative to the V1, V2 and V3 lineage sequences since their emergence. The relative lack of sequence divergence in the analysed datasets meant that the power of these gene-wide analyses was expected to be low, especially for V3 for which few sequences were available. However, among the genes with sufficient clade-level phylogenetic resolution, using the BUSTED[S] selection detection method [31] we found evidence of episodic diversifying selection in the N-gene (all clades), nsp3 (V1/V2), S (V3), and RdRp, endornase, exonuclease, ORF7 and ORF8 in V1. It should be noted though that the gene-wide signal of positive selection detected in the V1 ORF8 is likely a consequence of a mutation that introduced a stop codon into this ORF at position 27.

We next examined the different lineage-specific datasets with background references for evidence of positive selection at individual codon sites using MEME [34] and IFEL [35]. These analyses revealed evidence of positive selection at 111 individual codon sites across all lineages including 43 in V1, 64 in V2, and 14 in V3 (Table S1; https://observablehq.com/@spond/combined-view-of-sites-under-selection-in-n501y-clades). This is indicative of substantial ongoing adaptation of V1 and V2 sequences since the emergence of these lineages. V3 sequences might also be adaptively evolving but there is too little sequence data currently available to convincingly detect selection at individual codon sites in this lineage .

Differentiating between convergent mutations and recombination

Given that five of the signature mutations found in the V1, V2 and V3 lineages are convergent, it is worth considering how, mechanistically, these mutations were brought together in the progenitors of these lineages. While independent mutations in two viral genomes at the same codon site could both yield the same amino acid change, this same result could also be achieved if only one genome acquired the mutation, and then embedded in a stretch of homologous sequence it was swapped into the other genome

by genetic recombination in a co-infected individual. Two coronavirus genomes replicating within the same cell will frequently exchange fragments of their genetic code [36] and there are tentative reports that recombination may be detectable both within the global SARS-CoV-2 genome dataset [37], and specifically within the V1 and V2 lineages [38]. Although it is certain that SARS-CoV-2 is recombining, it is extremely difficult to differentiate between convergent mutations and actual recombination events when recombination is occurring between highly similar sequences [39]. Regardless, it is very unlikely that any of the presently detectable convergent 501Y lineage mutations were derived through recombination between viruses in the different 501Y lineages because until January 2021 the lineages were geographically separated. It is nevertheless possible that independent convergent mutations seen in different sub-lineages of V1, V2 and V3 sequences might in reality be the consequence of either intra-lineage recombinational transfers of the mutations between genetically distinct sequences within the lineages, or transfers from co-circulating non-501Y lineage viruses. In either case, just as convergent mutation events are effectively indistinguishable from recombination events, so too are the practical implications of convergent mutations and recombination. Since we detected no convincing evidence of recombination in any of the datasets that we analysed here, we will hereafter refer exclusively to convergent mutations even though in reality some of the mutations in question could have conceivably been acquired through recombination. With the emergence of more divergent lineages such as V1, V2 and V3, it is likely that recombination will be both more readily detectable as these begin to co-circulate, and more likely to generate increasingly diverse variants.

Signals of ongoing mutational convergence at signature mutation sites

Notable among the lineage-specific positive selection signals detected at 111 individual codon sites of 501Y lineage sequences, were those detected at 21 of their signature mutation sites: 6/11 of the V1 signature sites, 8/14 of the V2 ones and 12/17 of the V3 ones (see underlined codon site numbers in Figure 1). Given that (i) each lineage was defined by the signature mutations along the phylogenetic tree branch basal to its clade, and (ii) that these basal branches were included in the lineage-specific selection analyses, these selection analysis results were biased in favour of detecting the signature mutations as evolving under positive selection. We therefore used the selection results for signature mutation sites only to identify the signature mutations that, relative to the background reference sequences, were evolving under the strongest degrees of positive selection.

Most noteworthy of the signature mutation sites that displayed the strongest evidence of lineage-specific positive selection are codons S/18, S/80, S/417, S/484, S/501, and

S/701 in that all are either suspected or known to harbour mutations with potentially significant fitness impacts [9,10,23,25,26,29,40].

There are particularly interesting ongoing mutational dynamics at codons S/18 and S/484 in the V1 lineage. Although neither of these codons are signature mutation sites in the V1 lineage, in multiple independent instances mutations have occurred at these sites that converge on signature mutations seen in the V2 and V3 lineages.

Whereas the L18F mutation is fixed in all currently sampled V3 lineage sequences, it has occurred (and persisted in descendent variants) at least twice in the V1 lineage and at least four times in the V2 lineage. S/18 falls within multiple different predicted CTL epitopes [41] and the L18F mutation is known to reduce viral sensitivity to some neutralizing monoclonal antibodies [29]. An F at residue S/18 is also observed in 10% of other known Sarbecoviruses and the L18F mutation is the tenth most common in currently sampled SARS-CoV-2 genomes. Having occurred independently numerous times since the start of the pandemic, S/18 is also detectably evolving under positive selection in the global SARS-CoV-2 genome dataset (MEME p-value = 6.9×10^{-13}).

More concerning than these convergent L18F mutations are the convergent E484K mutations that have independently arisen at least twice within the V1 lineage and are arising independently elsewhere and persisting in descendent variants. These mutations converge on a significant signature mutation found in the V2 and V3 lineages that decreases viral sensitivity to both infection- and vaccine-induced neutralizing antibodies [17,24–26,40] and, in conjunction with the N501Y mutation, increases the affinity of Spike for human ACE2 receptor sites [10,11].

Similar convergence patterns to those observed at S/18 and S/484 can be seen at seven other signature mutation sites: S/26, S/681, S/701, S/716, S/1176, and N/80. Of these S/681, S/701, and S/716 fall within 30 residues of the biologically important furin cleavage site (S/680 to S/689). Whereas some V2 sequences have independently acquired the signature V1 mutations, P681H and T716I, some V1 sequences have independently acquired the V2 signature mutation, A701V. Any of these three mutations might directly impact the priming of spike for ACE2 binding and, consequently, the efficiency of viral entry into host cells [42]. SARS-CoV-2 variants with deletions of the furin cleavage site have reduced pathogenicity [43,44] and, although presently untested, it is plausible that the P681H mutation - which falls within this site - might increase the efficiency of furin cleavage by replacing a less favourable uncharged amino acid with a more favourable positively charged basic one [42]. Whereas S/716 is not detectably under selection in either the lineage-specific or global SARS-CoV-2 datasets, sites S/681 and S/701 are both detectably evolving under positive selection in the global SARS-CoV-2

dataset (MEME p-values = 1.6 x 10^{-6} and 7.9 x10^{-3} respectively); important indicators that are consistent with the P681H and A701V mutations being adaptive.

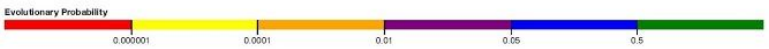
Non-convergent mutations at signature mutation sites might still be evolutionarily convergent

In addition to the ten positively selected signature mutation sites displaying evidence of convergent mutations between the V1, V2 and V3 lineages, five positively selected signature mutation sites (S/20, S/80, S/138, S/190 and S/215) display evidence of only divergent mutations (at the amino acid replacement level), where the same site is mutated as in another lineage, but to a different amino acid. All five of these codon sites are detectably evolving under positive selection in both individual lineage specific datasets and, with the exception of S/190, the global SARS-CoV-2 dataset (Figure 3).

The diverging mutations at these five sites might in fact also be contributing to the overall patterns of evolutionary convergence between the lineages; just via different routes. The N-terminal domain of spike where all five of these sites fall, (and S/80 in particular) is targeted by some monoclonal and infection-induced neutralizing antibodies [29] and it is therefore possible that these sites could be evolving under immunity-driven diversifying selective pressures. Therefore while mutations at S/20, S/80, S/138, S/190 and S/215 in different lineages do not converge on the same encoded amino acid states, they could nevertheless still be convergent on similar fitness objectives (immune escape or compensation for the fitness costs of other mutations): such as is likely the case with the also not-strictly-convergent V2 K417N and V3 K417T signature mutations.

Positive selection may be driving further convergence at non-signature mutation sites

In addition to lineage-specific signals of positive selection being detected at 21 of the signature mutation sites that characterize each of V1, V2 and V3 (Figure 1), such signals were also detected at 90 non-signature mutation sites Table S1). To test whether positive selection acting at these codon sites might be favouring convergent amino acid changes across the three lineages we examined each of the 90 non-signature mutation sites that were detectably evolving under positive selection in any one of the lineages for evidence of convergent mutations having occurred in either of the other two lineages. This revealed the occurrence of convergent mutations between sequences in different lineages at 26/90 (28.9%) of these sites including eleven in ORF1a and twelve in the S-gene (Figure 3).



Nucleotide	Site	V1	V2	V3	V1 subs to	V2 subs to	V3 subs to
1057	ORF1A/265		~		T I		
1312	ORF1A/350	✕	✕		I A	I	
1819	ORF1A/519	✕	✕		S	S	
2995	ORF1A/911	✕	✕		F	F	
3094	ORF1A/944	✕	✕		L	L	
3265	ORF1A/1001	~ +			I		
3826	ORF1A/1188			~	L		
3922	ORF1A/1220	✕	✕		L H	L S	
5098	ORF1A/1612	✕	✕		L	L	
5227	ORF1A/1655		~ +		K	N	
5386	ORF1A/1708	~ +			D	A	
5647	ORF1A/1795			~ +	K		Q
6523	ORF1A/2087	✕	✕		I	I	
6952	ORF1A/2230	~			T		
7726	ORF1A/2488	✕	✕		F	F	
8206	ORF1A/2648	✕	✕		I	I	
8659	ORF1A/2799	✕	✕		Y	Y H	
10321	ORF1A/3353		~		K		
11416	ORF1A/3718	✕	✕		F	F	
14827	ORF1B/454	✕	✕		I	I	
21108	ORF1B/2548	✕	✕		I	I	
21574	S/5	✕	✕		F L	F	
21613	S/18	✕	~✕	~✕	F	F	F
21619	S/20	✕	✕	~ +	T I	T	N
21622	S/21	✕	✕		K T	T	
21637	S/26	✕	+	~✕	S	L R	S
21640	S/27	✕	✕		S T	S	
21799	S/80	+	~ +		Y	A D	
21973	S/138	+		~	H		
22129	S/190			~ +	R		T
22204	S/215		~ +		Y	G H	
22225	S/222	✕	✕		V	V	
22810	S/417		~ +	~ +		N K	T
22984	S/475	✕	✕		V A	A	
23011	S/484	✕	~✕	~✕	E K	E K	K
23062	S/501	~✕	~✕	~✕	Y	Y N	Y
23269	S/570	~ +			D Y		
23308	S/583	✕	✕		D	D	
23425	S/622	✕	✕		F L	F	
23518	S/653	✕	✕		V	V	
23524	S/655			~	Y		
23602	S/681	~✕	✕		H,R	H	
23662	S/701	✕	~✕		V	V	
23707	S/716	~✕	✕		I V	I	
23995	S/812	✕	✕		S	S	
24235	S/892	✕	✕		V	V	
24505	S/982	~ +			A S		
24619	S/1020	✕	✕		S V	S	
24640	S/1027			~ +			T
24913	S/1118	~ +			P D		
25087	S/1176		✕	~✕		E	E
25216	S/1219	✕	✕		V	V	
25560	ORF3A/57		~		H	Y	
25782	ORF3A/131	✕	✕		C	L C	
25902	ORF3A/171		~			L	
26148	ORF3A/253			~ +	S		P
28279	N/3	~			L		
28342	N/24	✕	✕		I	I	
28510	N/80	✕	✕	~ +	L	L	R
28705	N/145	✕	✕		H Y	H	
28876	N/202	✕		✕	S		S
28885	N/205		~				
29365	N/365	✕	✕		L S	S L	

Selection legend

1. ~ : Signtature mutation
2. + : MEME support for at least one branch
3. ✕ : convergent subsitutions to another lineage (no MEME support)
4. ✕ : convergent subsitutions to another lineage AND MEME support

Figure 3. Genome sites where signature and convergent mutations occur within the V1, V2 and V3 lineage sequences. Labels within the coloured blocks indicate amino acid substitutions occurring at individual sites whereas the colours of the blocks indicate model-based predictions of the probable evolutionary viability of the observed amino acid substitutions based on the numbers of times these substitutions have been observed in related coronaviruses that infect other host species. The absence of colour indicates unprecedented substitutions, red indicates highly unusual substitutions and green indicates common substitutions seen at homologous sites in non-SARS-CoV-2 coronaviruses.

Most notable among the mutations occurring at these non-signature sites are spike L5F mutations seen in V1 and V2 lineages. S/5 has been detectably evolving under positive selection in the global SARS-CoV-2 dataset since May 2020 (MEME p-value = 1.6 x 10^{-16}) [3] and in both the V1 (p=0.08, marginal evidence) and V2 lineages (p=0.001). L5F mutations have arisen independently at least eight times in the V1 lineage and at least twice in the V2 lineage. The precise biological significance of the L5F mutation remains unknown but it falls within CTL epitopes that are predicted to be recognizable by a broad array of different HLA types (Campbell et al., 2020) and is therefore potentially a CTL escape mutation.

Collectively, convergent amino acid changes in individual clades of the V1 or V2 lineages respectively account for 17/236 (7.2%) and 16/103 (15.5%) of all the non-synonymous mutations mapping to internal branches of these clades. Conditioned on observed variability in the sequences, there are 50 polymorphic amino-acid sites shared by the two clades; the probability that 16 of those 50 will be convergent between two clades by chance alone is ≤ 10^{-6} (permutation test).

All 26 genome sites where convergent mutations occur are detectably evolving under positive selection in at least one of the V1, V2 or V3 lineages. Further, 20/26 of the codons where these convergent non-signature site mutations occur are also inferred to be evolving under either pervasive or episodic positive selection in the global SARS-CoV-2 dataset (MEME p-values < 0.05). Collectively this is strong evidence that an appreciable proportion of the convergent non-signature site mutations are adaptive.

Coevolution of spike codons hint at the enabling role of the N501Y mutation

In an effort to better understand the selective processes that might have driven, and might still be driving, the adaptive convergence of mutations in V1, V2 and V3 we separately examined patterns of amino acid substitution within each of the three lineages for evidence of different amino acid sites coevolving with one another. It is plausible that, as has been observed in other rapidly evolving viruses [45,46], adaptive mutations in particular genes of V1, V2 and V3 lineage viruses might have had an

associated fitness cost that was subsequently recouped by a secondary compensatory mutation elsewhere in those genes (a process called antagonistic epistasis). Alternatively, two or more individual mutations that in previous phases of the epidemic may have appeared to have little or no adaptive value when found alone, might synergise when occurring together within the V1, V2 and/or V3 viruses to yield substantial fitness benefits (called synergistic epistasis). Codon pairs with these antagonistic (dampening) and synergistic (enhancing) epistatic interactions might be expected to co-evolve to some degree. We attempted to test for such coevolving codon pairs using a Bayesian graphical model [45] approach which infers networks of conditional dependence relationships among codon sites.

Although the power of this test was expected to be severely limited by the low degrees of genetic diversity in the V1, V2 and V3 datasets, it nevertheless succeeded in identifying sets of somewhat concordant co-varying residues in the V1 and V2 datasets.

In V2, we found three-way coevolutionary interactions between positions S/417, S/501 and S/701. If this apparent epistasis is antagonistic then, given the centrality of S/501 in the emergence of the V1, V2 and V3 lineages, it is likely that N501Y was a primary adaptive mutation with the other two being compensatory for whatever fitness costs were associated with a Y residue at S/501. S/417 sits ~20 Å away from S/501 in the spike 3D structure (Figure 4) and might conceivably compensate for some less-than-optimal structural impact of the N501Y mutation. Given that both sites are within the ACE2 receptor binding motif, mutations at S/501 might also interact either antagonistically or synergistically with S/417 to optimize ACE2 binding affinity (as is suggested in [11] for this particular site pair) and/or maximize neutralization escape. Similarly, S/701 which sits ~102 Å from S/501 near the furin cleavage site and therefore potentially impacts furin-mediated priming of the spike protein for ACE2 binding, might synergistically interact with S/501 to expedite cellular entry.

Three two-way coevolutionary interactions were also detected in the S-gene of V1 lineage viruses. As with the V2 coevolutionary interactions these include pairs of sites in/near the RBD (S/366 and S/575 which are ~29 Å apart, and S/479 with S/490 which are ~14 Å apart) that are close enough in the 3D structure that one site might epistatically interact with the other with respect to ACE2 binding and/or neutralization escape. The detected coevolutionary interactions in the V1 spike also include a pair of sites where one, S/560, is near the RBD in the 3D structure and the other, S/716, is ~81 Å away near the furin cleavage site such that they too may be coevolving to optimize cellular entry.

While the clustering of coevolving sites in the RBDs of V1 and V2 lineage viruses suggests that mutations in this region might commonly involve fitness trade-offs

between ACE2 binding affinity and neutralization susceptibility, we must caution against over-interpreting these results. Given the small numbers of mutations informing these coevolution analyses, the inferred epistatic interactions are best viewed as speculative.

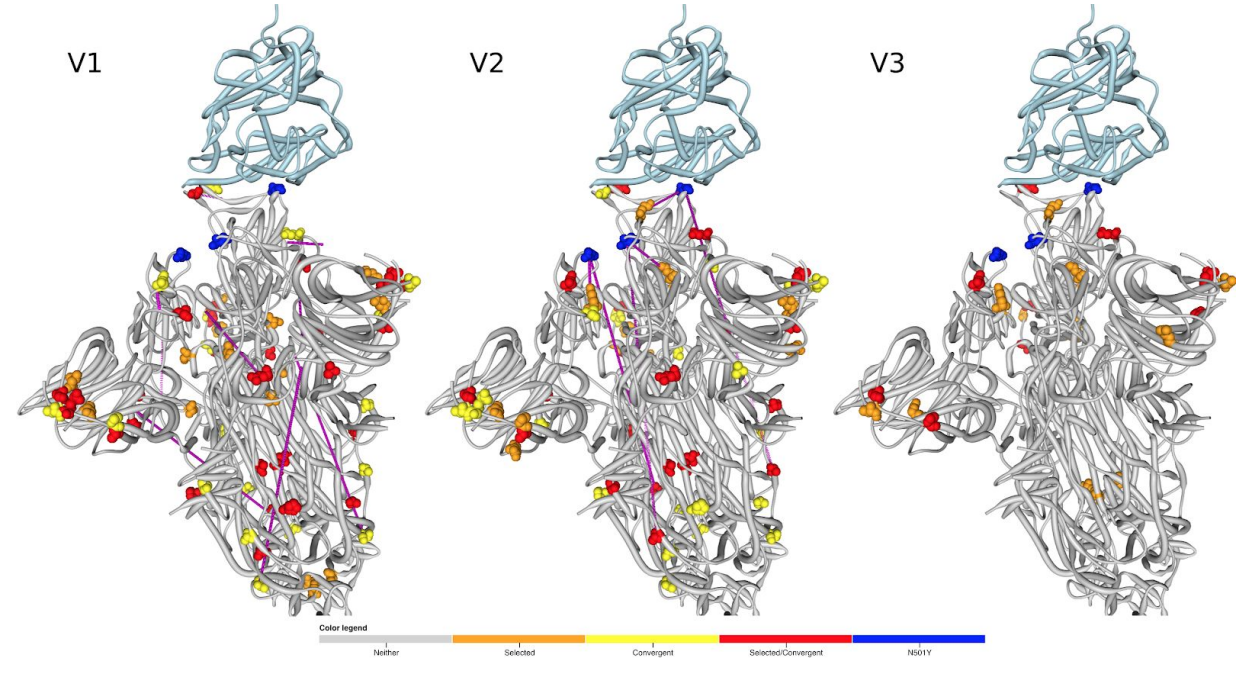


Figure 4. Locations of amino acids encoded by codons that are evolving under positive selection and/or encode convergent amino acid changes between the lineages mapped to the 3D structure of Spike (PDB 7DF4 structure; [47]). The human ACE2 receptor is shown in light blue. Signature mutations are **not** represented unless they are inferred to be under positive selection. Pairs of sites that are detectably coevolving within the different lineages are linked by purple lines.

What might have caused the sudden emergence of the 501Y lineages?

The sedentary pace of SARS-CoV-2 evolution in the first eleven months of the pandemic implies that, with the probable exception of the spike D614G mutation [3], almost none of the arising mutations during that time conferred any substantial fitness advantage. What this means is that from the start of the pandemic SARS-CoV-2 was likely near a fitness peak with respect to its capacity to infect and transmit between humans [2,48]. This could explain why, with the exception of the spike L18F mutation, neither the key convergent signature mutations in V1, V2 and V3, nor the additional convergent mutations that we report here rose substantially in frequency in the global SARS-CoV-2 population before October 2020. Given the numbers of infections that had occurred by October, all of these individual mutations, and even all of the potentially epistatically interacting pairs of these mutations, would have arisen independently

multiple times in different infected individuals without providing the viruses within which they occurred a detectable fitness advantage.

The fitness advantages of viruses that led to the V1, V2 and V3 clades may have only manifested after October 2020 because the fitness landscape upon which SARS-CoV-2 had been evolving up until then underwent a topographical change that opened up new mutational pathways to increased fitness. Our discovery here of a major selective shift in the evolution of SARS-CoV-2 variants sampled before and after November 2020 appears to support the possibility of a sudden global SARS-CoV-2 fitness landscape change in approximately October 2020. The precise cause of this selective shift is unknown but increases in seropositivity and/or the relaxation of transmission prevention measures are obvious candidates.

However, the apparent sudden emergence of the 501Y lineages might not have required selective shifts and a changed or changing global fitness landscape. The emergence of these lineages could have been enabled by mutational jumps across a static global fitness landscape if such a landscape included the possibility of viruses being transmitted late during the course of long-term SARS-CoV-2 infections [6]. Such long-term infections are relatively common but mostly involve dead-end viral shedding into the digestive tract [49]. Occasionally, particularly in the context of immune-compromised patients [50,51], these long-term infections also involve chronic upper respiratory tract infections with transmissible virus shedding. The local SARS-CoV-2 fitness landscapes within chronically-infected patients may be either softened, or shifted: perhaps because the main selective pressures within these patients are immunity evasion instead of transmission. By allowing the accumulation of larger combinations of epistatically interacting mutations and facilitating better exploration of local fitness landscapes within individual patients, chronic infections could have enabled the serendipitous discovery of new peaks on the global fitness landscape; i.e., adaptations to a single individual enhancing characteristics such as transmissibility and population-scale immune evasion. The credibility of this "chronic-illness-emergence-hypothesis" is strengthened by the fact that SARS-CoV-2 variants with notable similarities to viruses in the 501Y lineages have been independently observed in chronically infected patients. Besides carrying N501Y and E484K mutations, some of these viruses have also carried the spike N-terminal domain deletions seen in V1 sequences [50,51] .

The chronic-illness-emergence-hypothesis hypothesis is also appealing because it explains both the high degree of divergence of the 501Y lineage viruses relative to their nearest ancestors, and the timing of their emergence. The apparent absence of substantial SARS-CoV-2 adaptations during the early pandemic would simply be a consequence of the durations of the chronic infections within which the progenitors of

these lineages evolved. Chronic infections lasting between three and five months have been well documented [50–53]. Chronic infections that occurred in April 2020 when the first global peak of infections occurred might be expected to only yield substantially adapted variants by July or August 2020 which is approximately concordant with estimates of the times when the most recent common ancestors of viruses in the V2 lineage arose [7]. Factoring in the time it would have taken for prototypes of these 501Y lineages to spread sufficiently from the index cases, their detection and characterization in the last three months of 2020 may have in fact been the expected outcome of chronic infections that started in April. Such outcomes could perhaps have even been predictable given the mounting selection pressures exerted by increasing population immunity in the latter half of 2020.

What patterns of selection, convergence and coevolution tell us about the adaptive value of 501Y lineage mutations

Although it is clear that in the regions of the world where the 501Y lineage viruses are expanding in prevalence these viruses have substantial fitness advantages over the SARS-CoV-2 variants that preceded them, it remains unclear what the precise biological underpinnings of these advantages are.

One possibility is that viruses in the 501Y lineages might be better at infecting people who have been previously infected. The convergent spike mutations in codons S/18, S/80, S/417, S/484, and S/501 all provide some degree of protection from known anti-SARS-CoV-2 neutralizing antibodies [25,29,40]. The places in the world where V2 and V3 seem to have first emerged have high to medium density population centres (Nelson Mandela Bay and Manaus respectively) that had large numbers of people who had anti-SARS-CoV-2 antibodies in October 2020; the time when it is assumed that these lineages began exponentially expanding [14,54]. Further, viruses in both lineages seem highly resistant to neutralization by the sera of previously infected individuals [15–17].

While this “evasion-of-preexisting-immunity” hypothesis might explain fitness advantages of the V2 and V3 lineage viruses, it does not convincingly explain the fitness advantage of V1 lineage viruses. Modelling of the spread of V1 in the UK suggested a much better fit under transmissibility advantages, than immunity evasion [20]. Additionally, very few of the signature V1 S-gene mutations are known to have any strong impact on antibody binding [17,40]. The deletion of codon S/144 in the N-terminal domain of Spike in these sequences appears to be strongly protective against some monoclonal antibodies [26,29] and the S982A mutation that they carry can be slightly protective against some Pfizer and Moderna vaccine induced antibodies [49]. Nevertheless, only very modest reductions have been observed in the capacity of polyclonal sera from vaccinated individuals to neutralize V1 pseudotyped viruses,

suggesting that V1 likely has limited immune escape adaptations [24,25,55,56]. Furthermore, V1, V2 and V3 signature mutations (including N501Y) were overrepresented in *in vitro* experimental evolution lines not exposed to any active immunity, suggesting that many of the apparent immune-evasion signature mutations in all of these lineages might contribute to increased R0 irrespective of their immune evasion properties [57].

Even if immune selection was not a primary contributor to the emergence of V1, recent reports on the discovery of V1 sequences carrying the signature V2 and V3 spike mutation, E484K [58], together with our discovery here in V1 sequences of spike N-terminal domain mutations (such as L5F, L18F, P26S, A27S, and A222V) at codon sites that are both detectably evolving under positive selection and also converge on mutations seen in V2 and V3 sequences, suggests that immune pressure is at present likely impacting the ongoing evolution of the V1 lineage.

Besides immune evasion, other plausible contributors to the fitness advantage of viruses in the 501Y lineages are: (1) that they produce more particles at anatomical sites suitable for optimal droplet or aerosol transmission, (2) that they produce particles that individually have physical properties that increase their survivability during transmission, or (3) that they have molecular features that make them better at entering human cells. V1 infections have been associated with increased viral loads at the nasopharynx [32] which suggests that increased transmissibility may simply be a consequence of increased numbers of viral particles per exhaled infectious droplet. There is, however, conflicting data on this point with some indications that V1 infections might not in fact be associated with higher viral loads [59,60]. Currently there is no available data on how well the V1, V2 and V3 lineage viruses survive in the environment relative to those of other SARS-CoV-2 lineages.

What is known with a high degree of confidence, however, is that the N501Y mutation that is shared by V1, V2 and V3 viruses increases the affinity of spike for human ACE2 receptors [9], that this affinity is synergistically enhanced by the E484K mutations found in V2 and V3 viruses [10,11], and that these mutations therefore likely increase the probability that contacts between virus particles and human cells will result in successful infections. Crucially, increased ACE2 binding affinity could by itself also contribute to the evasion of pre-existing immunity. Besides favouring ACE2 in competitions with neutralizing antibodies for spike protein binding sites, increased ACE2 binding affinities might shorten the time needed for viruses to enter cells and, therefore, shrink the window of their vulnerability to antibodies.

The time required for bound viruses to enter cells could also be reduced in V1 and V2 viruses by optimized furin-mediated priming of spike for ACE2 binding. Although V1, V2 and V3 viruses all carry spike mutations that might impact this process (P681H and

T716I in V1, S701V in V2, and H655Y in V3 [42], it is presently unknown whether any of them actually do. Here we detect both convergence in some V2 sequences on the V1 P681H and T716I signature mutations, and convergence in V1 sequences on the V2 signature mutation, A701V. The convergence of mutations at these sites, two of which are also detectably evolving under positive selection in the global SARS-CoV-2 dataset, suggests that selection in the V1 and V2 lineages may be favoruring further optimization of spike priming for ACE2 receptor binding.

The interplay between selection favouring optimization of furin-mediated spike priming, ACE2 binding, and escape from RBD targeting antibodies is perhaps evident in the weak signals of coevolution that we detect in V1 and V2 sequences between pairs of S gene codons within the RBD (such as S/479 and S/490 in V1 and S/501 and S/417 in V2) and sites in the RBD with sites close to the furin cleavage site (such as S/570 and S/716 in V1 and S/501 and S/701 in V2). These coevolution patterns and experimentally demonstrated synergism between N501Y and E484K mutations [10] collectively suggest that constellations of mutations in the 501Y lineages may have initially coalesced around a set of synergistic epistatic interactions between N501Y, an ACE2 binding optimization mutation, furin-mediate spike priming optimization mutations (potentially P681H, A701V, and T716I), and immune evasion mutations in the RBD (such as E484K, K417N and K417T).

Epistasis could also explain why SARS-CoV-2 variants carrying N501Y mutations are only now starting to dominate the pandemic. The increase in ACE2 binding affinity and modest protection from some neutralizing antibodies that are provided by N501Y, even in the absence of changes at other sites [9,40], likely fostered the continual low-but-stable prevalence of this mutation globally. If S/501Y does indeed have a propensity to epistatically interact with, and realize the adaptive benefits of, multiple different mutations impacting ACE2 binding, spike priming, cell entry and/or immune evasion, then this could explain its centrality during the coalescence of the V1, V2 and V3 signature mutation constellations.

Where is the evolution of the 501Y lineages headed?

Regardless of specific details of how or why the V1, V2 and V3 lineages evolved, it is apparent that the evolution of viruses within these lineages is continuing and that much of this evolution is adaptive, involving both sequence diversification and further mutational convergence between viruses in the different lineages. We can speculate based on our selection analyses, the convergence patterns that we detect, and model based predictions of codons expected to be found is SARS-CoV-2 based on the evolution of related coronaviruses in other host species (http://hyphy.org/w/index.php/PRIME), that the culmination of this current stage of

convergence might be a virus containing a subset of codons expressing 350I, 519S, 911F, 944L, 1220L, 1612L, 2087I, 2488F, 2648I, 2799F and 3718F in ORF1a; 454I and 3548F in ORF1b; 5F, 18F, 21T, 26S, 27S, 215Y/G/H, 222V, 417N/T, 484K, 501Y, 583D, 622F, 653V, 681H, 701V, 716I, 812S, 892V, 1020S, 1176F, and 1219V in the S-gene; 131C in ORF3a; and 24I, 80L/R, 145Y, and 365L/S in the N-gene.

The most important issue here is not whether this particular “super variant” ever arises, but rather that the convergent mutations that have already arisen in members of the different 501Y lineages implies that these viruses are presently on, and are actively scaling, a broad new peak in the fitness landscape. Whatever SARS-CoV-2 variants eventually summit that peak could be a considerably bigger problem for us than any variants that we currently know in that they might have any combinations of increased transmissibility, altered virulence and/or increased capacity to escape population immunity.

Although only time can test the accuracy of this prediction, it should be possible using *in vitro* evolution to infer some amino acid sequence features at the adaptive summit of this fitness peak. An obvious approach would be to use replicated, laboratory infections of either synthesised or sampled live viruses carrying complements of mutations that are representative of the current standing diversity within the V1, V2 and V3 lineages. In the presence of mixed sera from multiple previously infected and/or vaccinated individuals these infections would create the appropriate conditions both for genetic recombination to occur, and for selection to rapidly sort multiple recombination-generated combinations of input immune evasion, cell entry and replication impacting mutations. Although the chimaeras that ultimately dominate these *in vitro* infections will doubtlessly be cell-culture optimized (as opposed to transmission between, and replication within, humans optimized) they should nevertheless carry many combinations of mutations that will be relevant to the continuing pandemic and which should include some of the most concerning mutation combinations that might arise during 2021: perhaps particularly so when/if some of these 501Y lineage viruses start naturally recombining with each other. Even if the concerning combinations that are discovered are only triplets or quartets of mutations, these would still be invaluable hints at what we should start looking for when it comes to trawling the rapidly growing pool of SARS-CoV-2 genomic surveillance data for potential vaccine escape mutants and other potentially problematic variants of concern.

Methods

Unless specified otherwise, all analyses were performed on a single gene (e.g. S) or peptide product (e.g. nsp3); since genes/peptides are the targets of selection.

We aligned all sequences from an individual lineage (i.e. V1, V2 or V3) and reference (GISAID unique haplotypes in the corresponding gene/peptide) sequences to the GenBank reference genome protein sequence for the corresponding segment using the codon-aware alignment tool, ***bealign*** which is part of the BioExt Python package (https://github.com/veg/BioExt), also used by HIV-TRACE [61] with the HIV-BETWEEN-F scoring matrix (similar viral sequences). This approach did not consider insertions relative to the reference genome in subsequent analyses. Deletions were treated as missing data and were not specifically tested for with codon models. No convincing signals of recombination were detected in any of the alignments using RDP5 [62].

Figure 5. An example of how phylogenetic trees were partitioned into two non-overlapping sets of branches during selection analyses. A foreground clade (here illustrated in orange) is nested within a background tree (illustrated in blue). In our study the foreground clade comprised the subtree relating sequences in either the V1, V2 or V3 lineages to one another and the background clade the tree relating the 501Y lineage sequences to a set of algorithmically selected SARS-CoV-2 reference sequences that were representative of SARS-CoV-2 genetic diversity sampled before October 15th 2020..

Because the codon-based selection analyses that we performed gain no power from including identical sequences, and minimal power from including sequences that are essentially identical, we filtered V1, V2, V3, and reference (GISAID) sequences using pairwise genetic distances complete linkage clustering with the tn93-cluster tool (https://github.com/veg/tn93). All groups of sequences that are within **D** genetic distance (Tamura-Nei 93) of *every* other sequence in the group and represented by a single (randomly chosen) sequence in the group. We set **D** at *0.0001* for *lineage-specific* sequence sets, and at *0.0015* for GISAID reference (or "background") sequence sets. We restricted the reference set of sequences to those sampled before Oct 15th, 2020.

We inferred a maximum likelihood tree from the combined sequence dataset using *raxml-ng* using default settings (GTR+G model, 20 starting trees). We partitioned internal branches in the resulting tree into two non-overlapping sets used for testing (e.g., orange and blue branches in Figure 5) and annotated the Newick tree. Because of lack of phylogenetic resolution in some of the segments/genes, not all analyses were possible for all segments/genes. In particular this is true when lineage V1, V2 or V3 sequences were not monophyletic in a specific region, and no internal branches could be labeled as belonging to the focal lineage.

We used HyPhy v2.5.27 (http://www.hyphy.org/) [63] to perform a series of selection analyses. Analyses in this setting need to account for a well-known feature of viral evolution [45] where terminal branches include "dead-end" (maladaptive or deleterious on the population level) [64] mutation events within individual hosts which have not been "seen" by natural selection, whereas internal branches must include at least one transmission event. However, because our tree is reduced to only include unique haplotypes, even leaf nodes could represent "transmission" events, if the same haplotype was sampled more than once (and the vast majority were). We performed:

1. Gene-level tests for selection on the internal branches of the V1,V2 or V3 clades using BUSTED [31] with synonymous rate variation enabled.
2. Codon site-level tests for episodic diversifying (MEME) [31] and pervasive positive or negative selection (FEL) [30] on the internal branches of the V1,V2 or V3 clades.
3. Epistasis/co-evolution inference on substitutions along internal branches of the V1,V2 or V3 clades using Bayesian Graphical models [45].
4. Model based predictions of codons expected to arise is SARS-CoV-2 based on the evolution of related coronaviruses in other host species was determined using the PRIME method http://hyphy.org/w/index.php/PRIME.

We combined all the results using a Python script and visualized results using several open source libraries in ObservableHQ (*https://observablehq.com/@spond/n501y-clades*).

Acknowledgements

We gratefully acknowledge all of the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based (Table S2).

SLKP was supported by the following grants from the U.S. National Institutes of Health R01 AI134384 (NIH/NIAID), R01 AI140970 (NIH/NIAID), and a RAPID award from the US National Science Foundation 2027196 (NSF/DBI,BIO).

DLR is funded by the Medical Research Council (MC_UU_1201412) and Wellcome Trust (220977/Z/20/Z). OAM is funded by the Wellcome Trust (206369/Z/17/Z).

COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute.

PL acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS), the EU grant 874850 MOOD and the Wellcome Trust through project 206298/Z/17/Z.

JOW was supported by an NIH-NIAID R01 AI135992.

SEJ and HT are supported by H3ABioNet, an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) funded by the National Human Genome Research Institute of the National Institutes of Health under Award Number U24HG006941.

The Network for Genomic Surveillance South Africa (NGS-SA) is supported by the Strategic Health Innovation Partnerships Unit of the South African Medical Research Council, with funds received from the South African Department of Science and Innovation.

GWH is supported by a grant from the US National Institutes of Health (1U01AI152151-01)

Conflict of Interest

JOW has received funding from Gilead Sciences, LLC (completed) and the CDC (ongoing) via grants and contracts to his institution unrelated to this research.

Bibliography

1. Dearlove, B. *et al.* A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci USA* **117,** 23652–23662 (2020).
2. MacLean, O. A. *et al.* Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans. *BioRxiv* (2020). doi:10.1101/2020.05.28.122366
3. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182,** 812-827.e19 (2020).
4. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11,** 6013 (2020).
5. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* (2020). doi:10.1038/s41586-020-2895-3
6. Rambaut, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* (2020).
7. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* (2020). doi:10.1101/2020.12.21.20248640
8. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *virological.org* (2021). at <https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586>

9. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182,** 1295-1310.e20 (2020).
10. Zahradnik, J. *et al.* SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. *BioRxiv* (2021). doi:10.1101/2021.01.06.425392
11. Nelson, G. *et al.* Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *BioRxiv* (2021). doi:10.1101/2021.01.13.426558
12. Volz, E. *et al.* Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv* (2021). doi:10.1101/2020.12.30.20249034
13. Public Health England. Investigation of novel SARS-CoV-2 variant. Variant of Concern. (2020). at <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957504/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5_England.pdf>
14. CAB, P., Russell, T. W., Davies, N. & Kucharski, A. J. Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in South Africa | CMMID Repository. (2021). at <https://cmmid.github.io/topics/covid19/sa-novel-variant.html>

15. Cele, S. *et al.* Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. *medRxiv* (2021). doi:10.1101/2021.01.26.21250224
16. Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *BioRxiv* (2021). doi:10.1101/2021.01.18.427166
17. Wu, K. *et al.* mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. *BioRxiv* (2021). doi:10.1101/2021.01.25.427948
18. Hoffmann, M. *et al.* SARS-CoV-2 variants B.1.351 and B.1.1.248: Escape from therapeutic antibodies and antibodies induced by infection and vaccination. *BioRxiv* (2021). doi:10.1101/2021.02.11.430787
19. Garcia-Beltran, W. F. *et al.* Circulating SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *medRxiv* (2021). doi:10.1101/2021.02.14.21251704
20. Horby, P. *et al.* Non-parametric analysis of fatal outcomes associated with B1.1.7. *Imperial College London* (2021).
21. Cottam, E. M. *et al.* Coronavirus nsp6 proteins generate autophagosomes from the endoplasmic reticulum via an omegasome intermediate. *Autophagy* **7,** 1335–1347 (2011).
22. Xia, H. *et al.* Evasion of Type I Interferon by SARS-CoV-2. *Cell Rep.* **33,** 108234 (2020).
23. Greaney, A. J. *et al.* Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *BioRxiv* (2021). doi:10.1101/2020.12.31.425021

24. Collier, D. A. *et al.* SARS-CoV-2 B.1.1.7 sensitivity to mRNA vaccine-elicited, convalescent and monoclonal antibodies. *medRxiv* (2021). doi:10.1101/2021.01.19.21249840
25. Wang, P. *et al.* Increased Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7 to Antibody Neutralization. *BioRxiv* (2021). doi:10.1101/2021.01.25.428137
26. Wang, Z. *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *BioRxiv* (2021). doi:10.1101/2021.01.15.426911
27. Vasques Nonaka, C. K. *et al.* Genomic Evidence of a Sars-Cov-2 Reinfection Case With E484K Spike Mutation in Brazil. (2021). doi:10.20944/preprints202101.0132.v1
28. Nguyen, T. T. *et al.* Genomic Mutations and Changes in Protein Secondary Structure and Solvent Accessibility of SARS-CoV-2 (COVID-19 Virus). *BioRxiv* (2020). doi:10.1101/2020.07.10.171769
29. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *BioRxiv* (2021). doi:10.1101/2021.01.14.426475
30. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22,** 1208–1222 (2005).
31. Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32,** 1365–1371 (2015).
32. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5,** 1403–1407 (2020).
33. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative

contribution to global health. *Global Challenges* **1,** 33–46 (2017).

34. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8,** e1002764 (2012).
35. Pond, S. L. K. *et al.* Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput. Biol.* **2,** e62 (2006).
36. Su, S. *et al.* Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.* **24,** 490–502 (2016).
37. Varabyou, A., Pockrandt, C., Salzberg, S. L. & Pertea, M. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *BioRxiv* (2020). doi:10.1101/2020.09.21.300913
38. Ignatieva, A., Hein, J. & Jenkins, P. A. Evidence of ongoing recombination in SARS-CoV-2 through genealogical reconstruction. *BioRxiv* (2021). doi:10.1101/2021.01.21.427579
39. Posada, D. & Crandall, K. A. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* **98,** 13757–13762 (2001).
40. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29,** 44-57.e9 (2021).
41. Campbell, K. M., Steiner, G., Wells, D. K., Ribas, A. & Kalbasi, A. Prediction of SARS-CoV-2 epitopes across 9360 HLA class I alleles. *BioRxiv* (2020). doi:10.1101/2020.03.30.016931
42. Garry, R. F. *et al.* Spike protein mutations in novel SARS-CoV-2 ‘variants of

concern' commonly occur in or near indels. *virological.org* (2021).

43. Lau, S.-Y. *et al.* Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. *Emerg. Microbes Infect.* **9,** 837–842 (2020).
44. Johnson, B. A. *et al.* Furin Cleavage Site Is Key to SARS-CoV-2 Pathogenesis. *BioRxiv* (2020). doi:10.1101/2020.08.26.268854
45. Poon, A. F. Y., Lewis, F. I., Pond, S. L. K. & Frost, S. D. W. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput. Biol.* **3,** e231 (2007).
46. Du, W. *et al.* Mutation of the second sialic acid-binding site of influenza A virus neuraminidase drives compensatory mutations in hemagglutinin. *PLoS Pathog.* **16,** e1008816 (2020).
47. Xu, C. *et al.* Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Sci. Adv.* **7,**
48. Zhan, S. H., Deverman, B. E. & Chan, Y. A. SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? *BioRxiv* (2020). doi:10.1101/2020.05.01.073262
49. Gaebler, C. *et al.* Evolution of antibody immunity to SARS-CoV-2. *Nature* (2021). doi:10.1038/s41586-021-03207-w
50. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383,** 2291–2293 (2020).
51. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* (2021). doi:10.1038/s41586-021-03291-y
52. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding

from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183,** 1901-1912.e9 (2020).

53. Baang, J. H. *et al.* Prolonged severe acute respiratory syndrome coronavirus 2 replication in an immunocompromised patient. *J. Infect. Dis.* **223,** 23–27 (2021).
54. Buss, L. F. *et al.* Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. *Science* **371,** 288–292 (2021).
55. Shen, X. *et al.* SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral Spike vaccines. *BioRxiv* (2021). doi:10.1101/2021.01.27.428516
56. Muik, A. *et al.* Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. *BioRxiv* (2021). doi:10.1101/2021.01.18.426984
57. Szemiel, A. M. *et al.* In vitro evolution of Remdesivir resistance reveals genome plasticity of SARS-CoV-2. *BioRxiv* (2021). doi:10.1101/2021.02.01.429199
58. Public Health England. Investigation of novel SARS-CoV-2 Variant of Concern Technical briefing 5. (2021). at <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf>
59. Walker, A. S. *et al.* Increased infections, but not viral burden, with a new SARS-CoV-2 variant. *medRxiv* (2021). doi:10.1101/2021.01.13.21249721
60. Brown, J. C. *et al.* Increased transmission of SARS-CoV-2 lineage B.1.1.7 (VOC 2020212/01) is not accounted for by a replicative advantage in primary airway cells

or antibody escape. *BioRxiv* (2021). doi:10.1101/2021.02.24.432576

61. Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J. & Wertheim, J. O. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Mol. Biol. Evol.* **35,** 1812–1819 (2018).
62. Martin, D. P. *et al.* RDP5: A computer program for analysing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* (2020). doi:10.1093/ve/veaa087
63. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37,** 295–299 (2020).
64. Pybus, O. G. *et al.* Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol. Biol. Evol.* **24,** 845–852 (2007).