Distinct Patterns of Emergence of SARS-CoV-2 Spike Variants including N501Y in Clinical Samples in Columbus Ohio

Huolin Tu,[1] Matthew R Avenarius,[1,2,] Laura Kubatko[3,], Matthew Hunt,[1] Xiaokang Pan,[1] Peng Ru,[4] Jason Garee,[1] Keelie Thomas,[2] Peter Mohler,[5] Preeti Pancholi,[2] Dan Jones[1,2,4]

[1]James Molecular Laboratory at Polaris, The Ohio State University James Cancer Center, Columbus, OH, USA.
[2]Department of Pathology, The Ohio State University Wexner Medical Center, Columbus, OH, USA.
[3]Department of Evolution, Ecology and Organismal Biology, The Ohio State University James Cancer Center, Columbus, OH, USA.
[4]The Ohio State University Comprehensive Cancer Center, The Ohio State University James Cancer Center, Columbus, OH, USA.
[5]Departments of Physiology and Internal Medicine and Davis Heart and Lung Research Institute, The College of Medicine and Ohio State University Wexner Medical Center

Corresponding Author:
Dan Jones, MD, PhD
The Ohio State James Molecular Laboratory
Innovation Center at Polaris
2001 Polaris Parkway
Columbus, OH 43240
Phone: 614-293-4993
E-mail: daniel.jones@osumc.edu

Text Page count: 15
References: 24 references
Tables: 2
Figures: 2
Supplement: 1 Table, 2 FASTA files

Abstract

Following the worldwide emergence of the p.Asp614Gly shift in the Spike (S) gene of SARS-CoV-2, there have been few recurring pathogenic shifts occurring during 2020, as assessed by genomic sequencing. This situation has evolved in the last several months with the emergence of several distinct variants (first identified in the United Kingdom and South Africa, respectively) that illustrate multiple changes in the S gene, particularly p.Asn501Tyr (N501Y), that likely have clinical impact. We report here the emergence in Columbus, Ohio in December 2020 of two novel SARS-CoV-2 clade 20G variants. One variant, that has become the predominant virus found in nasopharyngeal swabs in the December 2020-January 2021 period, harbors S p.Gln677His (Q677H) membrane glycoprotein (M) p.Ala85Ser (A85S) and nucleocapsid (N) p.Asp377Tyr (D377Y) mutations, with additional S mutations in subsets. The other variant present in two samples, contains S N501Y, which is a marker of the UK-B.1.1.7 (clade 20I/501Y.V1) strain, but lacks all other mutations from that virus. It is from a different clade and shares multiple mutations with the clade 20G viruses circulating in Ohio prior to December 2020. These two SARS-CoV-2 viruses emerging now in several parts of the United States add to the diversity of S gene shifts occurring worldwide and support multiple independent acquisition of S N501Y (in likely contrast to the unitary S D614G shift) occurring first during this period of the pandemic.

Introduction

SARS-CoV-2 genomic sequencing has facilitated surveillance efforts to track shifts in viral isolates worldwide (Brufsky, 2020). The emergence in March-April of 2020 of the D614G mutation defining the more transmissible G-strain has been the primary shift during the first nine months of the pandemic (Korber et al, 2020). This variant has been shown to have increased cell binding and viral spread in *in vitro* models (Mok et al, 2020; Hu et al 2020). Within the last several months, however, emergence of several distinct SARS-CoV-2 strains with additional likely pathogenic changes have occurred. These include the rapid spread of novel variants in the United Kingdom (UK, Technical Advisory Group, 2020; European Centre for Disease Prevention and Control, 2020) and South Africa (Tegally et al, 2020) containing several likely pathogenic but distinct mutations in the Spike (S) gene, particularly N501Y. The rapid transmissibility of these variants (Davies, 2020) and the sudden occurrence of multiple changes in the S gene has raised concerned about shifts in the pattern of COVID-19 disease and possible variability in response to antibody therapies or vaccines.

Here, we report the results of SARS-CoV-2 genomic surveillance from April 2020 through January 2021 in Columbus, Ohio. These data reveals a parallel recent shift in the predominant 20C>20G clade that contains 3 new variants and the emergence of a new virus that harbors the S N501Y variant but is different from the UK-B.1.1.7 (20I/501Y.V1) and the South African variant B.1.351 (20H/501Y.V2).

Methods:

This study was approved by the Institutional Review Board for the utilization of residual RNA samples from routine clinical SARS-CoV-2 PCR testing for viral sequencing. Briefly, standard PCR-based detection of SARS-CoV-2 was initiated by extraction of viral RNA from nasopharyngeal (NP) swabs (KingFisher™ Flex Magnetic Particle Processor, ThermoFisher). The viral RNA was analyzed, in most cases, using the TaqPath COVID-19 Combo Kit with an Applied Biosystems 7500 Fast Dx Real-Time PCR instrument (ThermoFisher) for SARS-CoV2 detection. SARS-CoV-2 virus sequence was then detected by next-generation sequencing (NGS) using a validated clinical assay in the James Molecular Laboratory at The Ohio State University. Residual RNA from PCR-based testing was reverse-transcribed using SuperScript™ VILO™ cDNA Synthesis Kit (ThermoFisher). NGS was performed using primer sets that tiled the entire SARS-CoV-2 genome (Ion AmpliSeq SARS-CoV-2 Research Panel, ThermoFisher), with library preparation and sequencing performed on Ion Chef and S5, respectively (Ion Torrent, Life Technologies). This panel included primers for the co-amplification of human housekeeping genes to assess RNA quality.

Analysis was performed in the Ion Browser with COVID-19 annotation plugins that produced consensus FASTA files using the IRMA method (reference strain: NC_045512.2). For tree-building, individual COVID-19 sequence FASTA files were combined into a single multifasta file with a custom shell script. The multifasta files were aligned using MAFFT (Katoh et al, 2002) (version 7.453) using default settings. MAFFT alignment files were analyzed for maximum likelihood using RAxML (Stamatakis, 2006) (version 8.2.12) using the GTRGAMMA model with 1000 bootstraps. The tree was produced using Dendroscope (Huson and Scornavacca 2012)

(version 3.7.2) with default settings. Numbers at the tree branches represent percent of bootstraps supporting a branch (i.e. 30 = 300/1000 runs supporting this branch). Strain typing and clades were designated using the most recent NextStrain nomenclature, with clade designation as 20G throughout if the clade-defining mutations listed in Supplementary Table 1 are present (Bedford et al, 2021).

The sequence of the COH.20G/501Y variant in sample D32 was confirmed by an independent SARS-CoV-2 genomic sequencing and analysis method. Briefly, RNA was reverse-transcribed using SuperScript™ VILO™ cDNA Synthesis Kit (ThermoFisher). Libraries were produced with KAPA HyperPrep and DI Adapter Kit (Roche), SARS-CoV-2 viral sequences were captured with COVID-19 Capture Panel covering the entire genome (IDT) and the products were sequenced on the NextSeq 550 (Illumina). The analysis pipeline including BaseSpace, a custom pipeline using GATK tools and DRAGEN RNA Pathogen Detection software (Illumina). Sequences of the two COH.20G/501Y viruses were deposited to GISAID.org as EPI_ISL_832378 and EPI_ISL_826521; an example of COH.20G/677H was deposited as EPI_ISL_826463.

Results

Summary of sequencing results in the early and mid-pandemic period

RNA extracted from PCR-positive nasopharyngeal samples from the Columbus OH area was sequenced in April (n= 56), May (n = 71), June (n=21), July (n=16), September (n = 11) and December 2020 (n =36) and January 2021 (n = 24) for surveillance purposes. A total of 235 NP samples were sequenced. In April 2020, two samples were positive for the S-strain with the remainder representing the G strain.

Aside from the G strain-defining changes (Supplementary Table 1), there were very few recurrent non-conservative/non-synonymous changes observed in April and May 2020 samples. In that period, most of the G-strain positive cases represented G strain alone or an unspecified G branch (17.3%, commonly with ORF8 p.Ala51Val) or the 20C clade bearing ORF3 p.Gln57His (80.3%), with few representing clade 20B (2.4%). In June and July 2020, as apparent infection rates in Columbus decreased, there was a proportional increase in clade 20B (40.5% of samples), with fewer clade G/unspecified (21.7%) and clade C viruses (37.8%). In September, coinciding with an increase in PCR positivity rates in the area, 20C clade samples again predominated (72.7% of samples), with some showing additional variants closely matching the newly designated NextStrain 20G clade (Bedford, 2021), with the remaining being 20A or 20B clade viruses. Samples were not obtained during the months of October and November 2020.

Rapid emergence of a clade 20G virus with shared S, N and M mutations:

When sequencing resumed in late December, we noted the emergence of a distinct 20G clade that had acquired the following variants: S p.Gln677His (Q677H), M p.Ala85Ser (A85S)

and N p.Asp377Tyr (D377Y, Table 1A) and is designated COH.20G/677H (Figure 1, brackets). During the week of Dec 21st 2020, these 3 variants was co-detected in 1 of 10 samples (10%), but were detected in 6/20 (30%), 6/10 (60%) and 8/13 (61.5%) of samples in the following weeks. Three virus samples also showed N D377Y without the other two changes, indicating it was an earlier change. Other changes seen in a subsets of COH.20G/677H cases included ORF1AB p.Ile529Val and S p.Thr95Ile followed by either S p.Leu5Phe or S p.Asn914Ser.

In all cases, these co-occurring variants arose in a 20G clade variant branch that had been present in Columbus since at least September 2020. The backbone was defined by ORF1AB: p.Met2606Ile (c.7818G>A), p.Leu3352Phe (c.10054C>T), p.Thr4847Ile (c.14540C>T), p.Leu6053Leu (c.18159A>G), p.His7013His (c.21039C>T); ORF3A: p.Gly172Val (c.515G>T); ORF8: p.Ser24Leu (c.71C>T); N: p.Pro67Ser (c.199C>T) and p.Pro199Leu (c.596C>T).

Emergence of a distinctive clade 20G virus harboring S N501Y

In late December (12/30/20) and January (1/6/21), we detected two samples with a 20G strain backbone that had acquired S N501Y, as well as ORF8 R52I in one case (designated COH.20G/501Y, Figure 1, arrow) which are both changes present in the UK-B.1.1.7 strain. In contrast to the B.1.1.7 strain, which has a 20B origin, the S N501Y and ORF8 R52I variants identified in this case were on a 20G background common to our area, as defined by ORF1AB: p.Leu3352Phe (c.10054C>T), p.Leu6053Leu (c.18159A>G), p.His7013His (c.21039C>T), ORF3A: p.Gly172Val (c.515G>T), ORF8: p.Ser24Leu (c.71C>T), N: p.Pro67Ser (c.199C>T), and p.Pro199Leu (c.596C>T). Both samples also shared mutations in ORF1AB (p.Thr999Ile, p.Ala1074Val and p.Leu1313Leu, Table 1B) indicating they likely arose from the same precursor.

In addition to lacking the characteristic N p.ArgGly203LysArg (c.608_610delGGGinsAAC) marking the 20B clade, COH.20G/501Y viruses also lack the other mutations seen in B.1.1.7, as summarized in Supplementary Table 1.

Discussion

We report the presence of two samples in which SARS-Co2 had acquired the S N501Y variant in late December 2020/January 2021 in Columbus Ohio. This particular amino acid change was first highlighted in a clinical sample in the United Kingdom in association with other novel S variants and a clade 20B backbone (ECDC, 2020; Davies et al, 2020), with the combination named as the B.1.1.7 strain and a Next Strain designation as 20I/501Y.V1 (Bedford et al, 2021). The same N501Y mutation was subsequently found in a clade 20C strain in South Africa, where it was associated with a different set of additional S variants (Tegally et al, 2020), with Next Strain designation as 20H/501Y.V2 or B.1.351. In late December 2020, the incidence of detection of both of these variants began markedly increasing, implicating S 501Y (with or without other S mutations) in increased transmissibility. The virus with S N501Y identified in Columbus (COH.20G/501Y) has a 20G backbone but lacks nearly all of the reported consensus changes in 20I/501Y.V1 as well as those in the 20H/501Y.V2. This favors an independent acquisition of this variant in a 20G clade branch that has been consistently present in Ohio since at least September 2020.

Since the first version of this pre-print, nearly identical viruses (currently classified in 20G clade as B.1.2 by NextStrain) have now been reported in Michigan (EPI_ISL_826428), Utah (EPI_ISL_812388) and Texas (multiple submissions, including EPI_ISL_784046), as of January 17, 2021. Related viruses with additional mutations were also reported from Wyoming (e.g., EPI_ISL_765985) and Louisiana (EPI_ISL_778960). These viruses have collection dates from early November 2020 to mid-January 2021 compatible with spread of this variant throughout the middle of the US during that period.

The S N501Y mutation, located within the receptor binding domain, is of particular concern for two reasons. First, the S protein with 501Y mutation displays increased affinity for ACE2 (Luan et al, 2020; Starr et al, 2020). Second, S 501Y mutation may impact association of receptor binding neutralizing antibodies including those in the Regeneron cocktail (Weisblum et al, 2020; Starr et al, 2020). The S N501Y mutation has also been shown to emerge spontaneously with viral passaging in a mouse model of SARS-CoV-2 infection (Gu et al, 2020), supporting its role in promoting viral spread and/or transmissibility. The only other shared mutation of COH.20G/501Y with B.1.1.7, in one case only, is the ORF8 R52I mutation. B.1.1.7 also has a deletion involving ORF8 that would likely inactivate its functions; such deletions have emerged in multiple strains of SARS-CoV-2 (Su et al, 2020) but were not present in COH.20G/501Y.

We also report the emergence of a predominant SARS-CoV-2 virus population with a 20G clade backbone that has single mutations in the S, M and N genes (Q677H, A85S and D377Y, respectively) in Columbus, Ohio in December. Since the first publication of this pre-print, there have been reports of closely related 20G variants also harboring S Q677H in several states in the upper Midwest, including Michigan, Wisconsin and Minnesota (Pater, 2021).

The S Q677H mutation disrupts a QTQTN consensus sequence (Figure 2) adjacent to the polybasic furin cleavage site spanning the S1 and S2 junction (Jacob et al, 2020). Prior to late 2020, that mutation had only been rarely reported in NextStrain where it was mostly seen sporadically outside of the United States (Kim et al, 2020). Deletions spanning the QTQTN motif have also been reported and may influence viral properties (Liu et al, 2020). The N D377Y

mutation has been uncommonly reported previously (Gupta et al, 2020). The rapid emergence of this variant across the Midwest merits close attention.

References

Bedford T**,** Hodcroft EB**,** Neher RA, Artic Network. Updated Nexstrain SARS-CoV-2 clade naming strategy. Available at https://virological.org/t/updated-nextstain-sars-cov-2-clade-naming-strategy/581, accessed 2021 Jan 11.

Brufsky A. Distinct Viral Clades of SARS-CoV-2: Implications for Modeling of Viral Spread. Journal of medical virology. 2020 Apr 20.

Davies NG, Barnard RC, Jarvis CI, Kucharski AJ, Munday J, Pearson CA, Russell TW, Tully DC, Abbott S, Gimma A, Waites W. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. medRxiv. 2020 Dec 26.

ECDC: European Centre for Disease Prevention and Control. Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom – 20 December 2020. ECDC: Stockholm; 2020.

Greaney AJ, Loes AN, Crawford KH, Starr TN, Malone KD, Chu HY, Bloom JD. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. bioRxiv. 2021 Jan 4.

Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, Hilton SK, Huddleston J, Eguia R, Crowe Jr JE, Bloom JD. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. bioRxiv. 2020 Sept 10.

Gu H, Chen Q, Yang G, He L, Fan H, Deng YQ, Wang Y, Teng Y, Zhao Z, Cui Y, Li Y. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science. 2020 Sep 25;369(6511):1603-7.

Gupta A, Sabarinathan R, Bala P, Donipadi V, Vashisht D, Katika MR, Kandakatla M, Mitra D, Dalal A, Bashyam MD. Mutational landscape and dominant lineages in the SARS-CoV-2 infections in the state of Telangana, India. medRxiv. 2020 Aug 26.

Hu J, He CL, Gao Q, Zhang GJ, Cao XX, Long QX, Deng HJ, Huang LY, Chen J, Wang K, Tang N. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity. bioRxiv. 2020 Jan 1.

Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic biology. 2012 Dec 1;61(6):1061-7.

Jacob JJ, Vasudevan K, Pragasam AK, Gunasekaran K, Kang G, Veeraraghavan B, Mutreja A. Evolutionary tracking of SARS-CoV-2 genetic variants highlights intricate balance of stabilizing and destabilizing mutations. bioRxiv. 2020 Dec 29.

Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002 Jul 15;30(14):3059-66.

Kim JS, Jang JH, Kim JM, Chung YS, Yoo CK, Han MG. Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. Osong Public Health and Research Perspectives. 2020 Jun;11(3):101.

Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020 Aug 20;182(4):812-27.

Liu Z, Yuan R, Li M, Lin H, Peng J, Xiong Q, Sun J, Li B, Wu J, Hulswit RJ, Bowden TA. Identification of a common deletion in the spike protein of SARS-CoV-2. J Virol. 22 June 2020.

Luan, B., Wang, H. and Huynh, T., Molecular Mechanism of the N501Y Mutation for Enhanced Binding between SARS-CoV-2's Spike Protein and Human ACE2 Receptor. bioRxiv. 2021 Jan 5.

Mok BW, Cremin CJ, Lau SY, Deng S, Chen P, Zhang AJ, Lee AC, Liu H, Liu S, Ng TT, Lao HY. SARS-CoV-2 spike D614G variant exhibits highly efficient replication and transmission in hamsters. bioRxiv. 2020 Aug 28.

Pater AA, Bosmeny MS, Barkau CL, Ovington KN, Chilamkurthy R, Parasrampuria, Eddington SB, Yinusa AB, White AA, Metz PE, Sylvain RJ, Hebert MM, Benzinger SW, Sinha K, Gagnon KT. Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. bioRxiv. 2021 Jan 13.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Nov 1;22(21):2688-90.

Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KH, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, King NP. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020 Sep 3;182(5):1295-310.

Su Y, Anderson D, Young B, Zhu F, Linster M, Kalimuddin S, Low J, Yan Z, Jayakumar J, Sun L, Yan G. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. bioRxiv. 2020 Mar 12.

Technical Advisory Group: Brief on the viral variant VOC-202012/01, 23 December 2020.

Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K. 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 Dec 22.

Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, Muecksch F, Rutkowska M, Hoffmann HH, Michailidis E, Gaebler C. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife. 2020 Oct 28;9:e61312.

Figure Legends.

Figure 1. Phylogenic relationship between viruses seen in late December in Columbus Ohio. Samples labeled "COH" are nasopharyngeal swabs from patients tested in Columbus, Ohio from 12/21-12/31/20. Most are clade 20G, with one each being 20A, 20B, and a 20G variant (var) that does not show every strain-defining mutation. One of the 20G/501Y viruses is included and marked with an arrow. Three examples of the emerging 20G/677H variant are bracketed, with the adjacent 20G/377Y viruses containing N D377Y but not S Q677H or M A85S. Reference sequences (FASTA files downloaded from GISAID.org) show recent examples of South Africa B.1.351/20H/501Y.V2 (EPI_ISL_745160, 12/4/2020), a B.1.351 20H/501Y.V2 strain collected in Australia (EPI_ISL_775245, 1/4/2021), a B.1.1.7/20I/501Y.V1 virus collected in the United States (EPI_ISL_779154, 1/4/2021) and a 20C-derived virus from Nevada with several distinct S variants (EPI_ISL_751557, 12/4/2020). The 2019-nCoV/USA-WA1/2020 SARS-CoV-2 reference stain (ATCC) was used as a sequencing control. See Methods for details on tree-building and interpretation.

Figure 2. Site of the Q677H mutation in Spike gene in the QTQTN motif conservation/furin cleavage site. Signal peptide (SP), N-terminal domain (NTD), receptor-binding domain (RBD), fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2), and transmembrane domain (TM).

Table 1A. Mutations present in the emergent clade 20G/677H SARS-CoV-2 virus.

Gene	Nucleotide	cDNA	Amino acid	Cases	% in last month
N	G29402T	c.1129G>T	D377Y	24	38.1
S	G23593T	c.2031G>T	Q677H	21	33.3
M	G26775T	c.253G>T	A85S	19	30.2
Subgroup					
ORF1AB	A1850G	c.1585A>G	I529V	8	12.7
S	C21846T	c.284C>T	T95I	10	15.9
Minor (1)					
S	C21575T	c.13C>T	L5F	3	4.8
Minor (2)					
S	A25563G	c.2741A>G	N914S	3	4.8

Table 1B. Novel variants in the 20G/501Y sample in Columbus and other sites in the United States.

Gene	nucleotide	cDNA	amino acid
ORF1AB	C3261T	c.2996C>T	T999I
	C3486T	c.3221C>T	A1074V
	C4202T	c.3937C>T	L1313L
S	A23063T	c.1501A>T	N501Y
Present in 1 of 2 viruses			
ORF8	G28048T	c.155G>T	R52I

Includes COH.20G/501Y (EPI_ISL_832378 and EPI_ISL_826521) and similar viruses reported by other groups, including EPI_ISL_826428, EPI_ISL_812388, EPI_ISL_784046 (See Discussion for details).

Figure 1

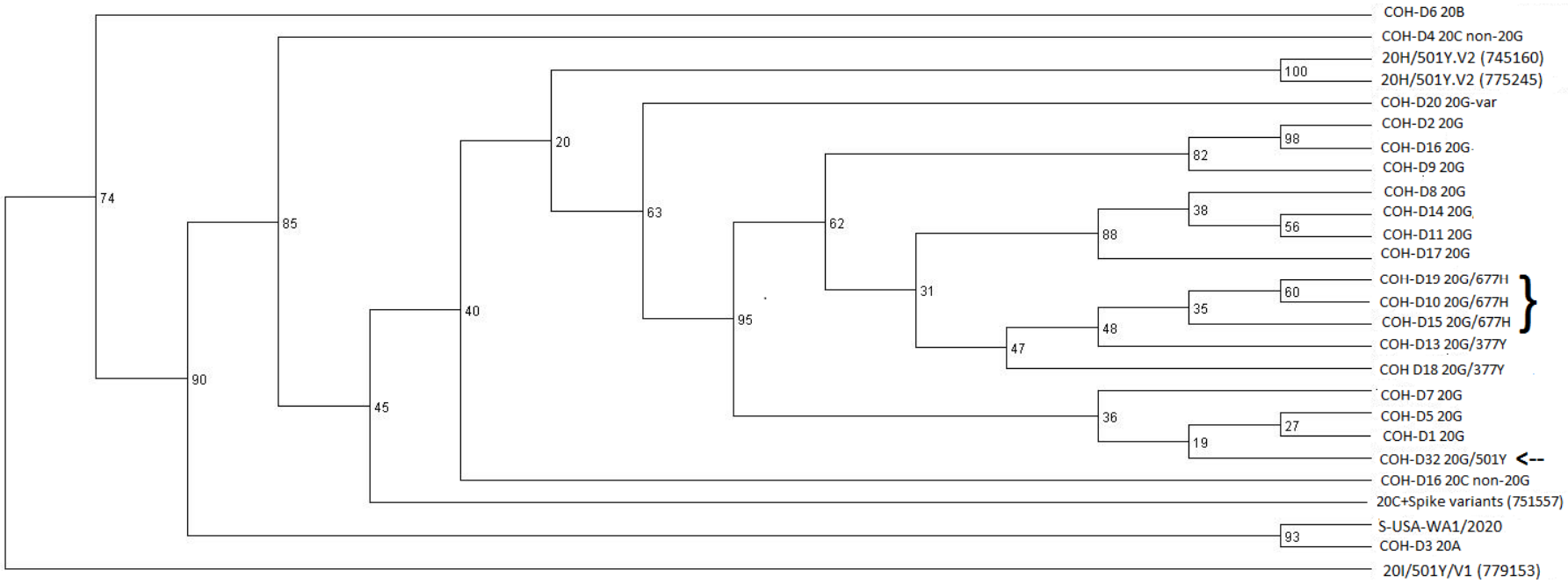

Figure 2