RESEARCH ARTICLE

Variation at Spike position 142 in SARS-CoV-2 Delta genomes is a technical artifact caused by dropout of a sequencing amplicon [version 1; peer review: 2 approved]

Theo Sanderson[1,2], Jeffrey C. Barrett[1]

[1]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
[2]Francis Crick Institute, London, NW1 1AT, UK

v1 **First published:** 10 Nov 2021, **6**:305
https://doi.org/10.12688/wellcomeopenres.17295.1
Latest published: 10 Nov 2021, **6**:305
https://doi.org/10.12688/wellcomeopenres.17295.1

Abstract

Public SARS-CoV-2 genomes from the Delta lineage show complex and confusing patterns of mutations at Spike codon 142, and at another nearby position, Spike codon 95. It has been hypothesised that these represent recurrent mutations with interesting evolutionary dynamics, and that these mutations may affect viral load. Here we show that these patterns, and the relationship with viral load, are artifacts of sequencing difficulties in this region of the Delta genome caused be a deletion in the binding site for the 72_RIGHT primer of the ARTIC V3 schema. Spike G142D should be considered a lineage-defining mutation of Delta.

Keywords

delta, genome, artic

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 10 Nov 2021	✓ view	✓ view

1. **James Davis**, University of Chicago, Chicago, USA

2. **Joseph Fauver**, University of Nebraska Medical Center, Omaha, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Theo Sanderson (theo.sanderson@crick.ac.uk), Jeffrey C. Barrett (jb26@sanger.ac.uk)

Author roles: Sanderson T: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Barrett JC**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust [210918]; COG-UK; the Medical Research Council part of UK Research and Innovation; the National Institute of Health Research; and Genome Research Limited, operating as the Wellcome Sanger Institute.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Sanderson T and Barrett JC. **Variation at Spike position 142 in SARS-CoV-2 Delta genomes is a technical artifact caused by dropout of a sequencing amplicon [version 1; peer review: 2 approved]** Wellcome Open Research 2021, **6**:305 https://doi.org/10.12688/wellcomeopenres.17295.1

First published: 10 Nov 2021, **6**:305 https://doi.org/10.12688/wellcomeopenres.17295.1

Introduction

The ARTIC Network amplicon protocol is one of the most widely used approaches to sequence severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes. It consists of approximately 100 pairs of PCR primers which each amplify a ~340 base-pair section of cDNA from the SARS-CoV-2 genome for subsequent sequencing. Version 3 of the ARTIC schema was designed using the Wuhan-Hu-1 reference sequence (MN908947) and released in March of 2020[1,2]. Mutations that arise in SARS-CoV-2 in these primer sites might affect how well each section of the genome is amplified, and thus the quality and completeness of the resulting sequence. If such mutations increase in frequency, for example by occurring on widespread lineages, they can cause substantial issues of interpretation when using public sequence datasets.

Here we investigate the effect of a deletion found in the genome of all Delta lineage SARS-CoV-2 that profoundly reduces the amplification of ARTIC V3 amplicon 72. In particular, we investigate its effects on calls for spike mutations G142D and T95I, which have recently been argued to be recurring within the Delta lineage[3] and associated with higher viral load. We show that these observations are actually a direct consequence of amplicon 72 failure, and that G142D is a lineage-defining mutation for Delta.

Methods

Except where otherwise specified, the following analyses were performed on a set of COG-UK genomes[4] generated by the Wellcome Sanger Institute[5] from 1 March to 30 June 2021 for which C_t value (the PCR cycle threshold for detecting the virus in the diagnostic test: higher numbers mean less virus in the sample) data were available.

We analysed multiple sequence alignments, and raw reads in the Sequence Read Archive, to observe the true underlying dynamics at the sites corresponding to G142 and T95.

We used a simple Python script to extract residues at certain locations from COG-UK and GISAID multiple sequence alignments, created from consensus sequences aligned to the Wuhan Hu-1 reference sequence. We also calculated coverage from 10 randomly selected genomes corresponding to various categories, using `samtools depth`.

To examine underlying reads, we mapped GISAID genomes to their accessions in the SRA, by firstly connecting each GISAID to its GenBank entry (where available). We then queried GenBank for corresponding SRA accessions using a script that we make available in our code repository. We downloaded 7 random genomes, and used a Python script to create a pile-up with the `pysam` library from which we extracted the residue at each position of interest, as well as the start and end positions of the read.

Results

In UK data, all Delta sequences with a known residue at spike position 142 are G142D

A convenient, human-readable way to represent a SARS-CoV-2 genome is to simply list all positions that differ from the Wuhan reference, either in nucleotide or amino acid co-ordinates. This summary is available, for example, from GISAID in the `metadata.tsv` files. When represented this way, just 65% of our set of UK genomes are annotated as having the G142D mutation, and there is an unspoken assumption that the remaining 35% of sequences *do not* harbour the mutation. However, there is a third possibility: that we lack information on the residue at spike position 142.

To distinguish these possibilities, it is necessary to look at the nucleotides corresponding to position 142 in a multiple-sequence alignment. We have built a simple tool to make converting from amino acid to nucleotide coordinates more straightforward (`codon2nucleotide.theo.io`, see also Table 1). Spike position 142 is encoded by a codon at nucleotide positions 21986-8. In the case of Delta, the relevant nucleotide mutation is a mutation at position 21987 from G in the reference sequence to A in Delta, which creates the spike G142D substitution. Plotting the distribution of residues in our dataset shows that 65% of Delta sequences indeed have an A at the 21987 position – however nearly all of the remaining 35% of sequences do not have the G seen in the reference (and 100% of Alpha sequences), but instead have an N indicating that the nucleotide at this position is unknown (Figure 1).

Table 1. Summary of the true nucleotides at each position discussed in this work in the Wuhan reference, Delta lineage, and AY.4 sublineage, and the amino acid mutations they result in.

Genomic position	Reference nucleotide	Delta nucleotide (AA mutation)	AY.4 nucleotide (AA mutation)
21987	G	A (G142D)	A (G142D)
21846	C	C	T (T95I)

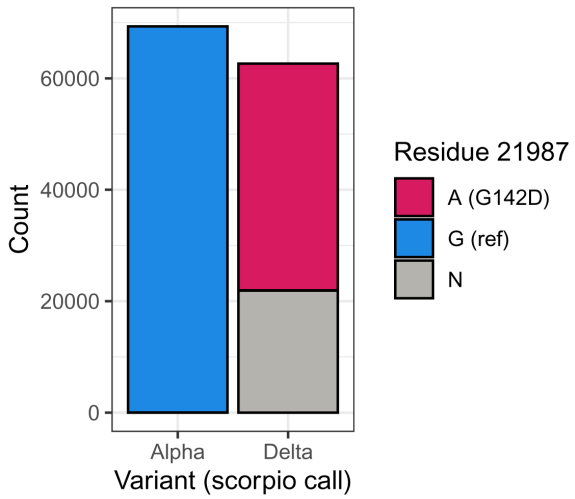

Figure 1. G142D is fixed in Delta, with almost all Delta sequences where the nucleotide at position 21987 has an A at this position. In contrast, Alpha contains the reference G at this position. 35% of Delta sequences have N at this position, indicating that the position does not have sequencing coverage.

Missing data at spike position 142 results from reduced coverage of ARTIC V3 amplicon 72, caused by a deletion in the binding region for its right-hand primer

To better understand how sequencing coverage of amplicon 72 affects these mutations, we examined the depth of coverage in this area for four sets of 10 representative genomes (Figure 2). The top panel shows coverage of Delta sequences with position 21987 A (red) and N (grey). All these sequences show a >50-fold drop in coverage for amplicon 72, compared with 71 and 73, and for about one-third of sequences, the coverage is so low that no consensus sequence is called, resulting in the N at this position. This phenomenon is caused by a Delta-lineage defining deletion at positions 22029–22034, which is in the binding site for the right hand primer for amplicon 72 (found at coordinates 21904–21933, Figure 3). The deletion removes the region to which the 5' end of the primer would bind, reducing the binding site by five nucleotides, and T_m from 66°C to 60°C. The deletion itself is covered by amplicon 73, so it is not affected by the amplicon 72 drop-out, and is called consistently in Delta viruses.

Remaining apparent reversions in UK data result mostly from mixed infections or contamination

Almost all sequences (>99.9%) in our dataset had an A or N at position 21987. Nevertheless, in this dataset there were 31 sequences typed as Delta[6] which had G at position 21987, and would thus be candidate "revertants", where a second mutation, back to the reference allele, occurred at that position. We examined these in more detail. A possible explanation for this would be if the sample contained a mixture of lineages,

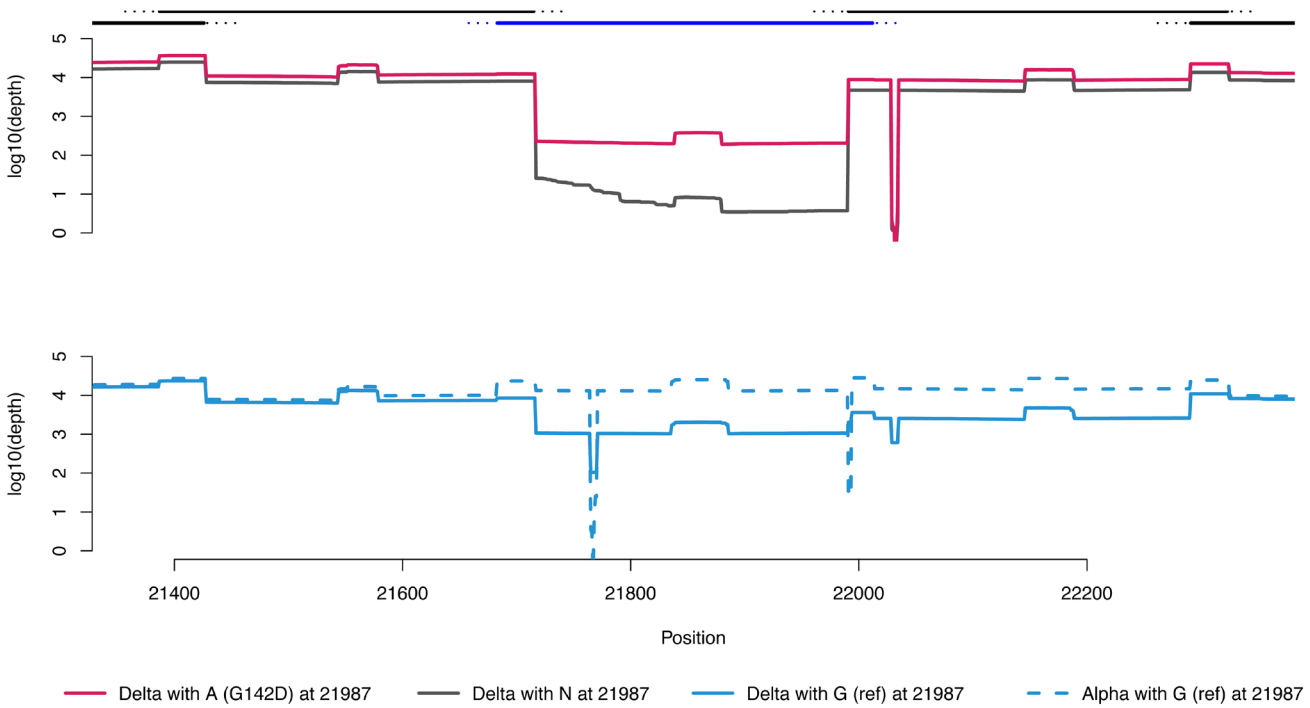

Figure 2. Local sequencing coverage (note log Y axis) near amplicon 72. Lines along the top show amplicons 71, 72 (blue) and 73, as well as parts of amplicons 70 and 74. Dashed portions are primers (so ought not to be present in the final sequence). In the top panel, the red line is average coverage for 10 random B.1.617.2 sequences with D at spike position 142; grey line average coverage for 10 random B.1.617.2 sequences with N (missing data) at that position. Bumps in coverage are caused by overlap of adjacent amplicons and overlap of paired-end illumina sequence reads in the middle of each amplicon. The B.1.617.2 lineage defining deletion at 22029–22034 causes the zero coverage dip in the right-hand primer for amplicon 72. In the bottom panel, Alpha genomes (blue-dashed) have good coverage throughout, and show dips at two characteristic deletions. The blue-solid line shows coverage for the small number of sequences typed as Delta with G at spike 142, which appear to be mixtures of Delta and Alpha.

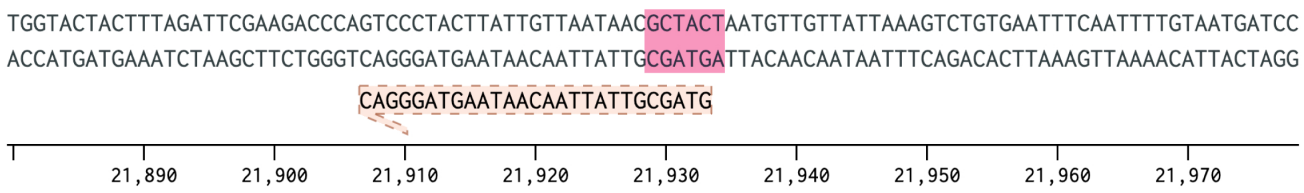

Figure 3. 72_RIGHT primer of the ARTIC V3 scheme shown against the SARS-CoV-2 reference genome, with the lineage defining deletion at 22029–22034 in Delta highlighted.

either in the patient during a mixed infection, or from laboratory contamination. If the other lineage is present at a level too low to affect the majority of the genome but amplifies successfully for amplicon 72, then the resultant genome would appear as a mosaic, with Delta sequence everywhere apart from at amplicon 72, which would be the other lineage. Given the time period of the study the most likely candidate would be B.1.1.7 (Alpha), which has a six base pair deletion at position 21764 within amplicon 72, resulting in H69/70del. We looked to see whether we could see this deletion ("-") in the apparent revertant sequences (those with G at 21987). In the majority of cases, we could (Figure 4), suggesting that this explanation explains most of the apparent revertants in our dataset. The bottom panel of Figure 2 shows concurrent Alpha sequences (dashed blue) have even coverage throughout this region, as well as the H69/70 deletion (and another at spike amino acid 144). The bottom solid blue line shows Delta sequences with a G at 21987, which show both intermediate coverage and mixed evidence for all three deletions (two from Alpha, one from Delta).

Apparent revertants in global sequencing data result mostly from untrimmed primer sequences

In a global dataset available on GISAID[7], 16% of Delta sequences have a G at position 21987, which is a far higher rate than the tiny number of apparent Alpha contaminants in the UK data described above. Indeed, testing for the presence of H69/70del suggests only a small proportion of these can be explained by contamination or mixed infections with B.1.1.7.

Residue 21987 lies within the 73_LEFT primer (21961–90) used for amplifying the amplicon immediately following amplicon 72. According to protocols, these primer sequences

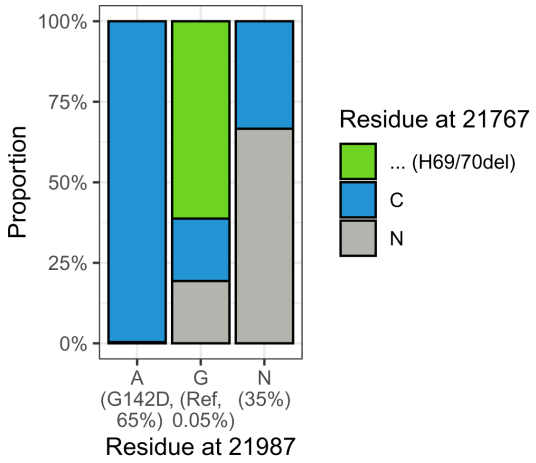

Figure 4. Relationship of the residue at 21987 in Delta lineage samples (representing spike 142) with the residue at 21767 representing spike 69/70del for the set of UK samples. Most of the small number of "revertants" with G at position 21987 also have a gap at 21767, corresponding to the spike 79/70 deletion found in B.1.1.7. This suggests that these sequences represent contamination from B.1.1.7 (either as a mixed infection or in the laboratory) that is specific to amplicon 72 because of the reduced efficiency of this amplicon in Delta samples.

should be trimmed from sequencing reads prior to down-stream analysis, but if this is not performed correctly, it could lead to miscalling position 21987, since the 73_LEFT primer contains the reference sequence for this region, with a G at position 21987, and because the reduced coverage of amplicon 72 would leave fewer actual viral reads to compete with those derived from the primer.

To investigate this effect we took a random sample of GISAID genomes for which the following conditions were satisfied:

- The genome is classified as Delta
- The genome has G at position 21987
- The mapped reads that led to the consensus sequence are deposited[1] in the Sequence Read Archive[8].

We sampled seven such genomes, all of which were from the USA (in part because many countries do not submit to the SRA). We downloaded the raw reads that had created these genomes and examined them to look for what evidence there was for the base at position 21987 (Figure 5).

In all cases, there were a substantial number of reads with an A at position 21987. Plotting reads according their starting point in the genome, to simulate the effect of primer clipping, suggested that in all the cases examined clipped reads provided reasonable evidence for the possibility of A at at position 21987, with three cases where A was the significant majority residue in clipped reads.

However not all reads beginning prior to 29160 had an A at position 21987. An explanation has very recently been provided by others[9], who show that the pools containing the primers for ARTIC amplicons 71 and 73 can result in non-specific amplification of a hybrid amplicon that also incorporates amplicon 72 sequence (as well as the 73_LEFT primer sequence). This would also mean that even when clipping was carried out (which our analysis suggests it sometimes is not), the primer sequence could be presented in an unexpected context in which it would not be removed. This effect is particularly difficult to diagnose and correct in the case of short read sequencing.

T95I is subject to the same dropout effect, but is also genuinely limited to certain Delta sublineages

We will briefly discuss T95I, which was suggested to exhibit similar dynamics to G142D[3]. T95I corresponds to a mutation at nucleotide 21846 from C to T. This position is also found within amplicon 72 and so is expected to exhibit similar artifacts to those described above.

[1]We used https://github.com/nextstrain/ncov-ingest/blob/05b3b36d8264f017b1a931a5427903793cbef802/source-data/accessions.tsv to convert GISAID IDs to GenBank IDs, which we then scanned for SRA accessions

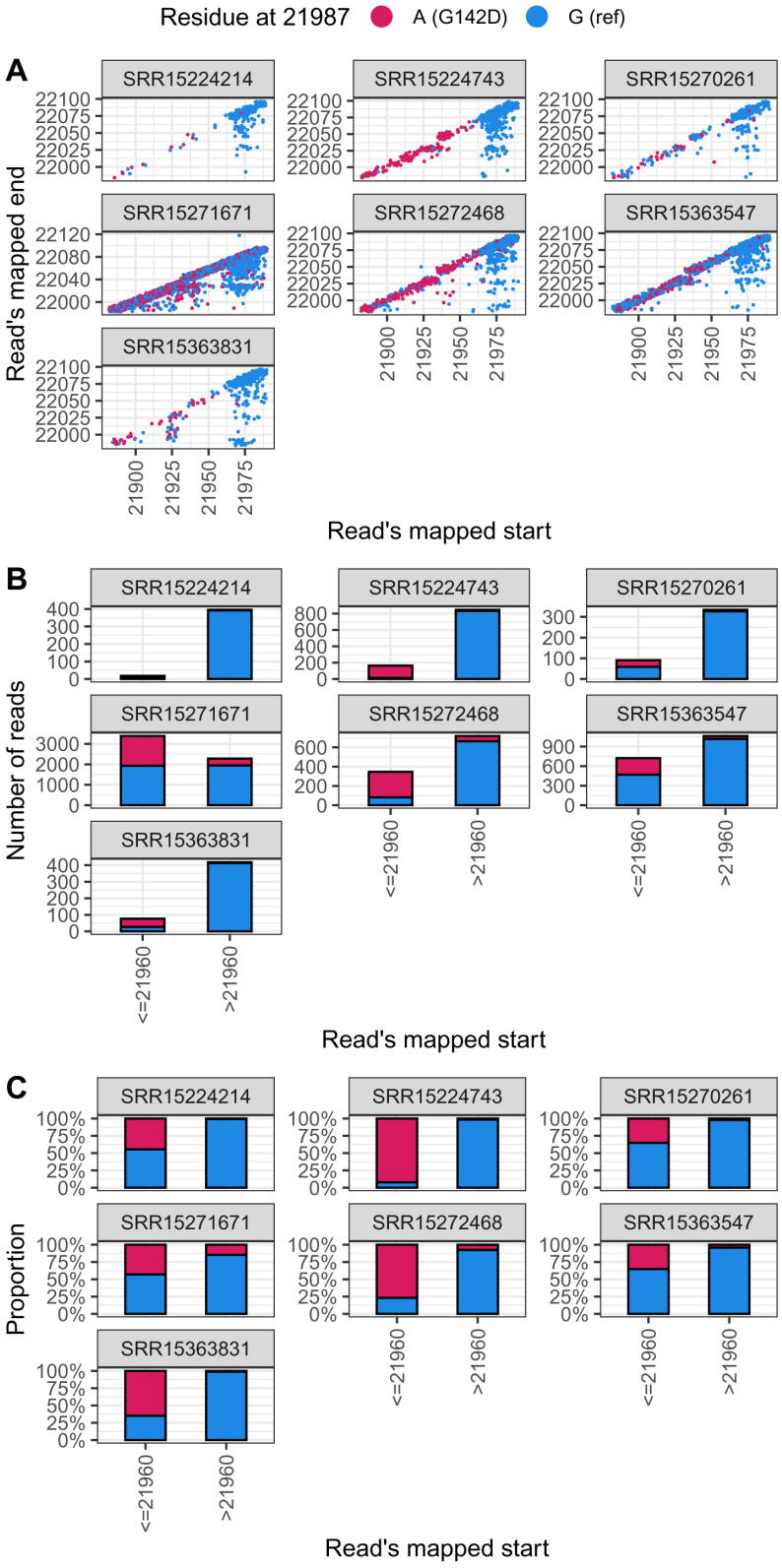

Figure 5. G residues at position 21987 in Delta result in large part from reads which appear to have come from the 73_LEFT primer. This plot shows seven samples randomly selected from the SRA with a G at position 21987 of the consensus sequence. In all cases there is substantial evidence for A at this position. Specifically reads which must have come from amplicon 72 (indicated as <= 29160) show substantial evidence for A, whereas reads likely to have come from the 73_LEFT primer sequence (indicated as > 29160) are overwhelmingly G. (Panel A features some jitter to reduce overplotting.)

T95I differs from G142D in that this mutation is *not* fixed within Delta. Rather there is genuine underlying biological variation of genotypes, over which the technical effects described above are layered. This can be seen in Figure 6, which shows that T95I is fixed in the AY.4 sublineage of Delta, but absent from any of the other designated sublineages. As anticipated, a substantial amount of sequences from all Delta sublineages have N at this position, reflecting the reduced efficiency of amplification at amplicon 72. As previously, this technical effect can explain the relationship between detection of T95I and higher viral loads reported in Shen *et al.*[3].

An increased likelihood of correctly calling position 142 for high viral load samples creates a spurious correlation between C_t values and genotype

These observations provide a clear hypothesis that could explain the observation of a correlation between viral load and the presence of the G142D mutation reported in Shen *et al.*[3]. For samples with low viral load (high C_t values), one might expect that the reduced amplification efficiency for amplicon 72 would be more likely to lead to an N at position 21987 than for samples with higher viral loads. To formally test this we plotted the residue at 21987 by C_t value, and indeed found that though at average C_t values the residue was able to be called correctly as A, as C_t value approached 30 almost all calls became Ns[2] (Figure 7). It is therefore likely that the observed relationship between genotype and C_t value is caused not by the genotype affecting the viral load, but instead the reverse, with the amount of viral material in the sample affecting the ability to detect the genotype.

[2]A trend towards more N calls at extremely low C_t values (high viral loads, $C_t \approx 10$) was also seen, which may reflect very large amounts of DNA template titrating out primers, leading to reduced coverage at this amplicon.

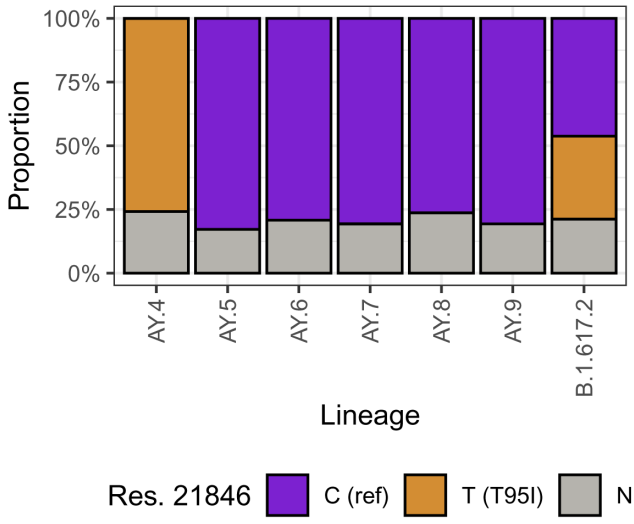

Figure 6. Distribution of nucleotides at 21846 (nucleotide T encodes T95I) for different sublineages within Delta. T95I is fixed in AY.4, but absent from other currently designated Delta sublineages.

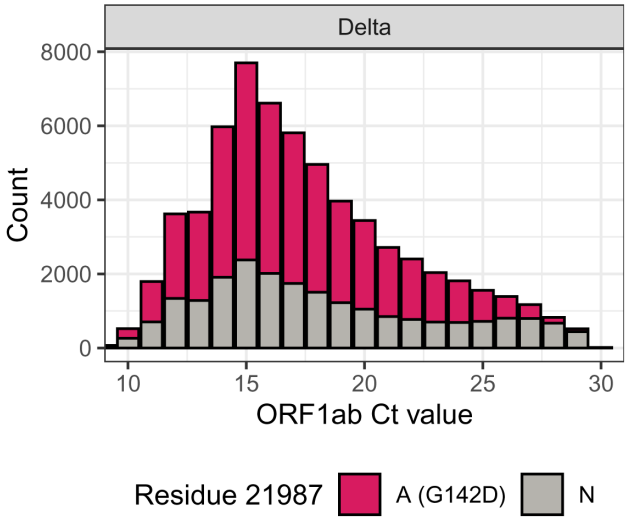

Figure 7. Relationship of C_t value and residue at position 21987 for COG-UK Delta samples until 30 June 2021.

Discussion

We have shown the apparent diversity within Delta at spike 142 and immediately adjacent regions reflects reduced amplification of a sequencing region rather than underlying biology. G142D is fixed in Delta, with essentially all apparent back mutations being artifacts. Similarly, T95I is fixed in AY.4, and any apparent reversions in that clade are artifactual. This suggests there is no current basis to expect a biological causative relationship between the presence of G142D or T95I and viral load.

For most analyses of Delta sequences, it is likely to be advantageous to mask out the entire amplicon 72 sequence to avoid being misled by the various technical effects at these sites. In general when analysing consensus sequences processed with unknown pipelines, masking out ARTIC V3 primers is often likely to be desirable; for example, D950N is another Delta mutation found within an ARTIC V3 primer site which exhibits spurious apparent reversions in global Delta sequences. Assumptions that the failure to detect a given mutation implies the absence of that mutation are prone to mistakes – it is important to consider the alternative that we lack information about a particular site.

Regular monitoring of amplicon coverage may assist in early detection of mutations at primer binding sites, and the artifacts that these may cause. Trimming primer sequences is an essential part of any pipeline for generation of consensus sequences, but may not be being performed robustly for a significant proportion of current global sequences. Importantly, open deposition of raw reads allowed us to examine these effects, and could even allow large-scale correction of erroneous sequences if suitable infrastructure were developed. The non-specific

amplification of amplicon 72 sequence noted by Cerutti *et al.*[9] further complicates these effects.

The ARTIC Network has released a V4 of the amplicon scheme[10], which resolves these issues[11], and we would encourage researchers to move to it to allow detection across the genome given that almost all sequenced genomes are currently Delta.

Data availability

Underlying data

COG-UK genomes (including the subset sequenced at the Wellcome Sanger Institute) that we analysed here, are available from https://www.cogconsortium.uk/tools-analysis/public-data-analysis-2/ as well as through GenBank.

International SARS-CoV-2 sequence data is available from GISAID at EpiCov.org.

Raw sequence reads are available from the Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra.

Code that we used for our analysis is available at https://doi.org/10.5281/zenodo.5608686 and includes a version of the Ct dataset.

Software availability

Source code for Codon2Nucleotide: https://github.com/theosanderson/codon2nucleotide

Archived source code as at time of publication:

https://doi.org/10.5281/zenodo.5608674

License: MIT

Acknowledgements

We thank the thousands of labs around the world who have deposited SARS-CoV-2 sequence data to public databases like GISAID and the SRA, as well as the Wellcome Sanger Institute Covid-19 Surveillance Team and COVID-19 Genomics UK (COG-UK) consortium. We also acknowledge @babarlelephant's early insights into these issues.

References

1. Quick J, Loman N, ARTIC Network: **Artic.network.** 2020.
Reference Source
2. Tyson JR, James P, Stoddart D, *et al.*: **Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore.** *bioRxiv.* 2020; 2020.09.04.283077.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Shen L, Triche TJ, Bard JD, *et al.*: **Spike protein ntd mutation g142d in sars-cov-2 delta voc lineages is associated with frequent back mutations, increased viral loads, and immune evasion.** *medRxiv.* 2021.
Publisher Full Text
4. Nicholls SM, Poplawski R, Bull MJ, *et al.*: **CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance.** *Genome Biol.* 2021; **22**(1): 196.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Vöhringer HS, Sanderson T, Sinnott M, *et al.*: **Genomic reconstruction of the sars-cov-2 epidemic in england.** *medRxiv.* 2021.
Publisher Full Text
6. O'Toole Á, Scher E, Underwood A, *et al.*: **Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool.** *Virus Evol.* 2021; **7**(2): veab064.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** *Glob Chall.* 2017; **1**(1): 33–46.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Leinonen R, Sugawara H, Shumway M, *et al.*: **The sequence read archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): 19–21.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Cerutti L: **Missing g21987a mutation in sars-cov-2 delta variants due to non-specific amplification by artic v3 primers.** *Virological.org.* 2021.
Reference Source
10. Quick J, ARTIC Network: **Artic.network.** 2021.
Reference Source
11. Davis JJ, Wesley Long S, Christensen PA, *et al.*: **Analysis of the artic version 3 and version 4 sars-cov-2 primers and their impact on the detection of the g142d amino acid substitution in the spike protein.** *bioRxiv.* 2021.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 18 May 2022

https://doi.org/10.21956/wellcomeopenres.19122.r50103

✔ **Joseph Fauver**
University of Nebraska Medical Center, Omaha, NE, USA

Sanderson and Barrett take a deep dive into a lineage defining mutation associated with SARS-CoV-2 genomes of the Delta lineage. This was a timely and important piece of detective work that determined there was indeed an amino acid mutation (G142D) in the Spike gene for Delta lineage samples. Their investigation demonstrated that the G->A SNP at nucleotide position 21987 is real and is a clade defining mutation, but that it is often not sequenced due to amplicon dropouts. The ARTIC primer sets have been the workhorse of many SARS-CoV-2 sequencing labs and have been incredibly reliable, however amplicon based approaches are subject to reductions in sensitivities due to mutations in primer binding sites, and in this case a deletion seen in all Delta genomes caused amplicon 72, which covers the site in question, to drop out. The authors also demonstrate that reversions at this site are likely the result of either i) contamination of sequencing libraries with Alpha sequences, or ii) untrimmed primer sequences. Finally, and one of the most useful and interesting points in this manuscript, the authors show that association of genomes with the G142D mutation and a lower CT value/higher viral load is the result of amplicon 72 not dropping out when there is more virus in the sample. There is a correlation between CT value and genome completeness, and in general, more SARS-CoV-2 RNA in a sample will result in more complete genomes, even if there are mutations in primer binding sites.

Taken together, this manuscript represents a comprehensive investigation into a pertinent mutation in Delta lineage genomes. My biggest take away from this manuscript is that work like this should be done routinely for all SARS-CoV-2 sequencing projects. Although Delta lineage viruses are no longer circulating, the issues identified here will likely occur again in other lineages at different sites in the genome. All Next Generation Sequencing strategies have associated biases and understanding how those biases and influence consensus sequence genomes and interpretations for clinical or evolutionary outcomes is crucial.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Next Generation Sequencing, genomic epidemiology, virology, parasitology, vector biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 29 November 2021

https://doi.org/10.21956/wellcomeopenres.19122.r47040

James Davis
Consortium for Advanced Science and Engineering, University of Chicago, Chicago, IL, USA

The article by Sanderson and Barrett offers a timely warning about the over the interpretation of SARS-CoV-2 sequencing results. They clearly lay out the cause of the ambiguity for the G142D mutation in Delta variants. I think this needs to be published so that as many people as possible 1) can see the importance of using an up to date primer scheme and 2) can see the importance of submitting sequencing reads to public archives.

I found no obvious or major issues with the text. My only criticism is of the last section of the results, which tries to demonstrate a trend between higher Ct values and ambiguity at position 21987. The two plots look similar to me, especially if they had been plotted as a fraction of the total samples. I think I would need to see statistics to be convinced that these two sets are different (especially since the title of the section is "Increased likelihood... ... ").

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, computational biology, microbiology, sars-cov-2, evolution, machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.