Title

Altered Subgenomic RNA Expression in SARS-CoV-2 B.1.1.7 Infections

Authors

Matthew D Parker[1,2,3], Benjamin B. Lindsey[4,5], Dhruv R Shah[5], Sharon Hsu[2,5], Alexander J Keeley[4,5], David G Partridge[4], Shay Leary[6], Alison Cope[4], Amy State[4], Katie Johnson[4], Nasar Ali[4], Rasha Raghei[4], Joe Heffer[7], Nikki Smith[5], Peijun Zhang[5], Marta Gallis[5], Stavroula F Louka[5], Max Whiteley[5], Benjamin H Foulkes[5], Stella Christou[5], Paige Wolverson[5], Manoj Pohare[5], Samantha E Hansford[5], Luke R Green[5], Cariad Evans[4], Mohammad Raza[4], Dennis Wang[1,2,3,8], Silvana Gaudieri[6,9,10], Simon Mallal[9,10], The COVID-19 Genomics UK (COG-UK) consortium[11+], Thushan I. de Silva[4,5*]

[+]Full list of consortium names and affiliations located in the supplementary material

*Corresponding Author

Email: t.desilva@sheffield.ac.uk

Affiliations

[1] Sheffield Biomedical Research Centre, The University of Sheffield

[2] Sheffield Bioinformatics Core, The University of Sheffield

[3] The Department of Neuroscience/Neuroscience Institute, The University of Sheffield

[4] Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK.

[5] The Florey Institute for Host-Pathogen Interactions & Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK.

[6] Institute for Immunology and Infectious Diseases, Murdoch University, Murdoch, Western Australia, Australia

[7] IT Services, The University of Sheffield, Sheffield, UK

[8] Department of Computer Science, The University of Sheffield, Sheffield, UK

[9] Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[10] School of Human Sciences, University of Western Australia, Crawley, Western Australia, Australia

[11] https://www.cogconsortium.uk

One Sentence Summary

The recently emerged and more transmissible SARS-CoV-2 lineage B.1.1.7 shows greater subgenomic RNA expression in clinical infections and enhanced expression of a noncanonical subgenomic RNA near ORF9b.

Abstract

SARS-CoV-2 lineage B.1.1.7 viruses are more transmissible, may lead to greater clinical severity, and result in modest reductions in antibody neutralization. subgenomic RNA (sgRNA) is produced by discontinuous transcription of the SARS-CoV-2 genome and is a crucial step in the SARS-CoV-2 life cycle. Applying our tool (periscope) to ARTIC Network Oxford Nanopore genomic sequencing data from 4400 SARS-CoV-2 positive clinical samples, we show that normalised sgRNA expression profiles are significantly increased in B.1.1.7 infections (n=879). This increase is seen over the previous dominant circulating lineage in the UK, B.1.177 (n=943), which is independent of genomic reads, E gene cycle threshold and days since symptom onset at sampling. A noncanonical sgRNA which could represent ORF9b is found in 98.4% of B.1.1.7 SARS-CoV-2 infections compared with only 13.8% of other lineages, with a 16-fold increase in median expression. We hypothesise that this is a direct consequence of a triple nucleotide mutation in nucleocapsid (28280:GAT>CAT, D3L) creating a transcription regulatory-like sequence complementary to a region 3' of the genomic leader. These findings provide a unique insight into the biology of B.1.1.7 and support monitoring of sgRNA profiles in sequence data to evaluate emerging potential variants of concern.

Main Text

The recently emerged SARS-CoV-2 lineages B.1.1.7 (20I/501Y.V1 or VOC-202012/01), B.1.351 (20H/501Y.V2 or VOC-202012/02) and P.1 (20J/501Y.V3 or VOC-202101/02) (*1*) have been classified as variants of concern by public health agencies. An increasing body of evidence suggests B.1.1.7 is more transmissible (*2*, *3*) and rapidly became the dominant circulating virus in the United Kingdom (UK) during October 2020 to February 2021 (Figure 1A). To date, B.1.1.7 has been reported in 93 countries (https://cov-lineages.org/, 22nd February 2021), with increasing prevalence. Recent data suggest a similar pattern for P.1 in Manaus (*4*), Brazil (*5*), and now 19 other countries (https://cov-lineages.org/, 22nd February 2021), where it appears to be more transmissible than it's ancestral lineage B.1.1.28 (*4*). Along with B.1.351, these variants of concern harbour functionally important mutations in the SARS-CoV-2 Spike protein, some of which demonstrate evidence of convergent evolution across all three lineages (Supplementary Table S1). Preliminary data (*6*, *7*) shows that B.1.1.7 positive diagnostic respiratory samples may have lower cycle threshold (Ct) values, therefore higher viral loads, compared to other lineages. These findings suggest a potential reason for enhanced transmissibility, though they did not account for potential confounders such as days since symptom onset at sampling. Many of these studies also use S gene target failure (SGTF) as a surrogate for the presence of B.1.1.7 (*6*), which might misclassify samples, depending on the prevalence of B.1.1.7 (*8*). Several analyses from community tested cases also suggest increased mortality associated with B.1.1.7 (*8*). Reasons for the potential viral load increase and enhanced mortality are currently unclear (*9*).

Genomic surveillance has been critical in rapidly identifying these variants and Nanopore sequencing of ARTIC Network (*10*) prepared SARS-CoV-2 amplicons is used by many laboratories to generate this data. We have recently reported an approach to quantify subgenomic RNA (sgRNA) expression profiles from genomic sequence data, which is produced as a result of a critical step in the SARS-CoV-2 replication cycle (*11*). sgRNA is produced from the genomically encoded SARS-CoV-2 RNA Dependent RNA polymerase (RdRp) using discontinuous transcription of the positive, single-stranded SARS-CoV-2

genome from the 3' end. Negatively stranded RNAs are produced, which are shorter than the genome, owing to a template switch from the ORF to the leader sequence at the 5' end of the genome when RdRp encounters a transcription regulatory sequence in the genome body (TRS-B) to a complementary TRS 3' of the leader sequence (TRS-L). All sgRNAs therefore contain a leader sequence at their 3' end which can be used computationally for their identification. There are thought to be nine such canonical sgRNAs; Spike:S, E: Envelope, M: Membrane, N: Nucleocapsid, ORF3a, ORF6, ORF7a, ORF8 and ORF10, although multiple studies have found negligible ORF10 expression (*12*, *13*).

As part of COVID-19 Genomics Consortium UK (COG-UK) (*14*) we are sequencing SARS-CoV-2 positive nose-throat swabs from healthcare workers and patients at Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK ('Pillar 1' testing). Additionally, to relieve pressure on centralised sequencing services, we also sequence a selection from 'Pillar 2' testing, which represent SARS-CoV-2 positive samples from the community, tested at the UK's Lighthouse Laboratories (Figure S1). We hypothesised that we would see differences in sgRNA expression profiles in distinct lineages of SARS-CoV-2, in particular, increased sgRNA in B.1.1.7 that may relate to its altered phenotype.

We stratified sgRNA expression by lineage in 4400 SARS-CoV-2 sequences that reached our previously defined quality control thresholds (>90% genome coverage, >50K mapped reads, Table S2), normalized for the genomic coverage from the corresponding amplicon for the five most abundantly expressed open reading frames (S, E, M, N and ORF6). We focused our analyses primarily on comparing B.1.1.7 with B.1.177, the dominant lineage in the UK, which was displaced by B.1.1.7. These two lineages are also the most represented sequences from both sampling pillars in our dataset (Figure 1A&B, B.1.1.7; Pillar 1: 729, Pillar 2: 150, B.1.177; Pillar 1: 764 Pillar 2: 179). A significant increase in normalized sgRNA expression of S, N, and to a lesser extent E, ORF6 and M, was apparent in B.1.1.7 SARS-CoV-2 infections (Figure 1C&D, Wilcoxon effect sizes; S: 0.475, E: 0.191, M 0.0700, N: 0.469, ORF6: 0.105). Negligible differences were seen in other ORF sgRNA (Figure S2). Consistent with previous findings (*6*, *15*), we also found significantly decreased E gene Ct

(ECT; in house diagnostic assay (*16*); B.1.177 median=25, B.1.1.7 median=23) and significantly greater Relative Light Units (RLU; Hologic Panther platform; B.1.177 median=1121, B.1.1.7 median = 1177) in B.1.1.7 infections compared to B.1.177 infections (Figure 1E&F). Accordingly, total mapped reads and total genomic RNA reads (i.e. non-sgRNA) were also significantly higher for B.1.1.7 sequences compared to B.1.177 (Figure S3).

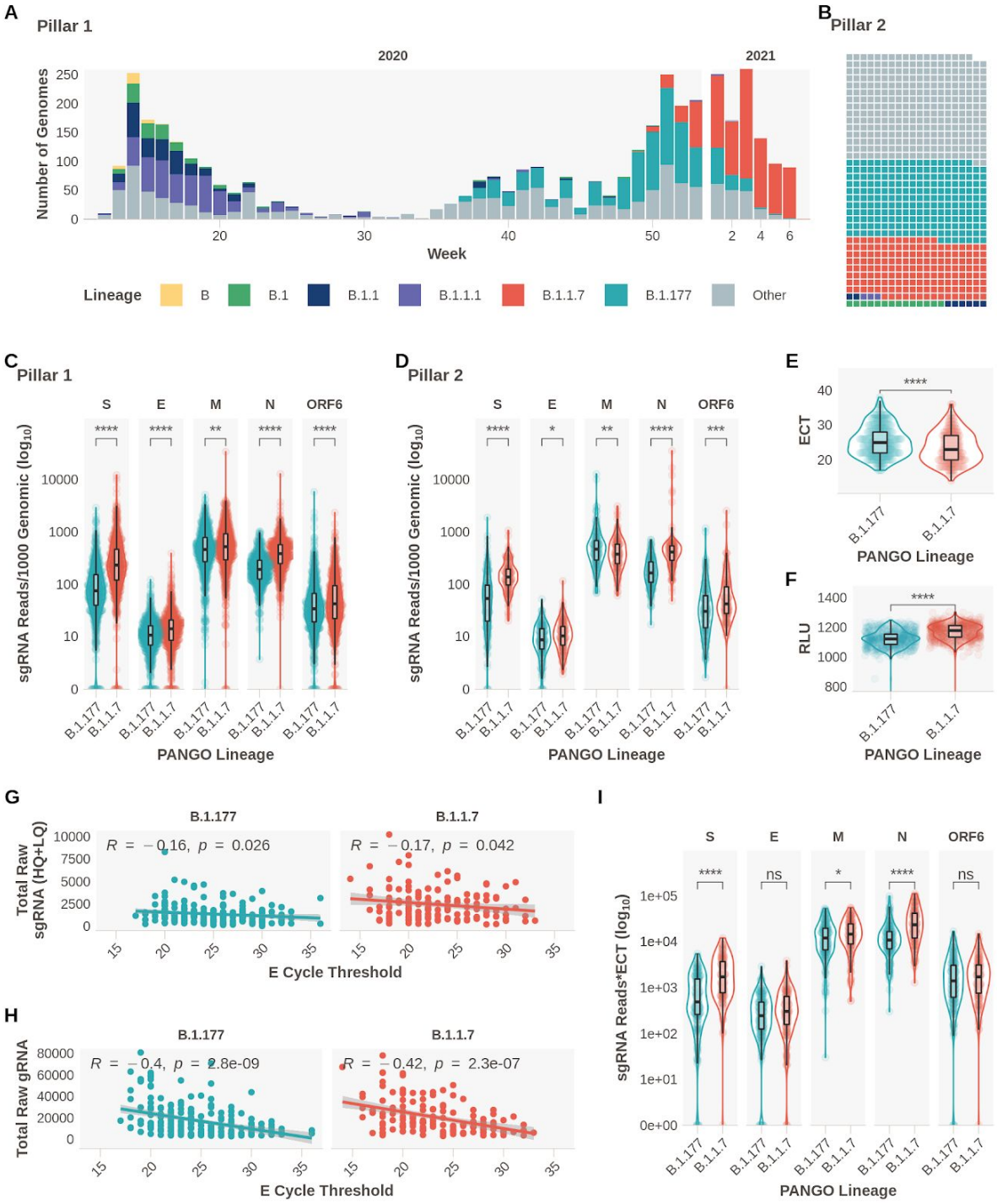


Figure 1. Subgenomic RNA Expression is Increased in B.1.1.7 infections

A. *Pillar 1 SARS-CoV-2 lineages over time (B.1.1.7$_n$ = 729, B.1.177$_n$ =764).* ***B.*** *Representation of lineage composition of pillar 2 data (B.1.1.7$_n$ = 150, B.1.177$_n$ = 179).* ***C.*** *sgRNA expression in samples of lineages B.1.177 and B.1.1.7 in pillar 1 samples from the most highly expressed ORFs.* ***D.*** *sgRNA expression in samples of lineages B.1.177 and B.1.1.7 in pillar 2 samples from the most highly expressed ORFs.* ***E.*** *E gene cycle threshold (ECT) for B.1.177 (n=257) and B.1.1.7 (n=185) lineages.* ***F.*** *Relative light units (RLU) for B.1.177 (n=626) and B.1.1.7 (n=626) lineages.* ***G****. E gene cycle threshold compared to total raw sgRNA counts (High and Low Quality) for B.1.177 and B.1.1.7. Correlation coefficient and p value using Pearson.* ***H.*** *E gene cycle threshold compared to total raw genomic RNA counts for B.1.177 and B.1.1.7. Correlation coefficient and p value using Pearson.* ***I.*** *Raw sgRNA counts normalised to E gene cycle threshold (sgRNA*ECT). All p-values (except G & H) calculated using an unpaired Wilcoxon signed rank test, and adjusted for multiple testing with the Holm method (**** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05)*

We observed a weak but significant negative correlation between raw sgRNA reads and ECT (R = -0.16 and p=0.026 for B.1.177, R = -0.17 and p=0.042 for B.1.1.7), with a greater negative correlation between genomic reads and ECT (R = -0.4 and p<0.0001 for B.1.177, R = -0.42 and p<0.0001 for B.1.1.7; Fig 1G&H). To ensure the observed sgRNA expression increase in B.1.1.7 was not due to the effect of ECT alone (as a surrogate of viral load), we normalized sgRNA to ECT and repeated our comparisons. The differences between B.1.1.7 and B.1.177 were still apparent for S, N and M sgRNA (Figure 1I), suggesting that the increase in B.1.1.7 sgRNA expression is seen independently of any difference in ECT between lineages. To ensure the changes observed were not due to parental lineage changes in sgRNA expression such as those observed with the R203K/G204R mutation (*17*) we analysed normalized sgRNA comparing B.1.1.7 and B.1.177 with their respective ancestral lineages (Figure 2A&B). An increase in sgRNA is still observed for B.1.1.7 compared to B.1.1 in all major ORFs and interestingly, for B.1.177 compared to B.1 for ORFs S, E, M and N (Figure 2A&B). This demonstrates the importance of stratifying sequences by lineage when studying sgRNA.

It is possible that the days since symptom onset at sampling may vary between lineages in our dataset, either due to changes in sampling practice over time or presentation of individuals to healthcare services, which in turn could confound our sgRNA findings. We obtained information on symptom duration at sampling for 2327 samples in our pillar 1 genome sequenced cohort. A modest difference in the distribution of the days since symptom onset between B.1.177 and B.1.1.7 was apparent, where B.1.1.7 infections appeared to be sampled earlier (Figure 2C). We therefore applied a generalized linear regression model to adjust for any effect this change in distribution might have on sgRNA expression levels in lineages B.1, B.1.1, B.1.177 and B.1.1.7. Subgenomic RNA was still significantly elevated in B.1.1.7 in S, N and M and to a lesser extent in E (Figure 2D, Fig S4, Tables S3&S4). A smaller increase in sgRNA in B.1.177 (compared to B.1) was still seen, though only for M and N. Furthermore, as sampling practice and sequencing in healthcare workers in our hospital remained consistent throughout the period when B.1.177 and B.1.1.7 were circulating, we hypothesized that these data were less likely to be affected by unknown confounders in sampling. Subgenomic RNA levels remained significantly higher in B.1.1.7 compared to B.1.177 infections when the comparison was restricted to samples from healthcare workers (Figure S5). No difference in sgRNA levels for B.1.1.7 was seen by sex or age (Figure S6 and S7). Relative light units in the diagnostic assay also varied over time in B.1.1.7 compared to B.1.177 infections, with significantly higher RLU at 0, 1, 2, 3 days since symptom onset (Figure 2E). Data since symptom onset for samples with Ct values were insufficient to replicate this analysis.

From these cross-sectional data we created a pseudo infection time course using normalized sgRNA expression at each day following symptom onset between B.1.1.7 and B.1.177 infections. Significantly higher sgRNA for B.1.1.7 compared to B.1.177 was observed early in infection for S and N sgRNA, with lower magnitude differences for E sgRNA (Figure 2F&S8). This increase appeared to be greatest on day one following symptom onset. Using the same methods for sgRNA estimation, we have previously shown that *ACE2* and *TMPRSS2* expression alter the kinetics of sgRNA expression *in vitro*, leading to a peak in expression at an earlier stage of infection (*11*). It is possible that the increased

affinity for *ACE2* conferred by N501Y in the B.1.1.7 spike protein (*18*, *19*) enhances sgRNA profiles early in clinical infection. Experimental validation of these findings in cell and animal models would be extremely valuable.

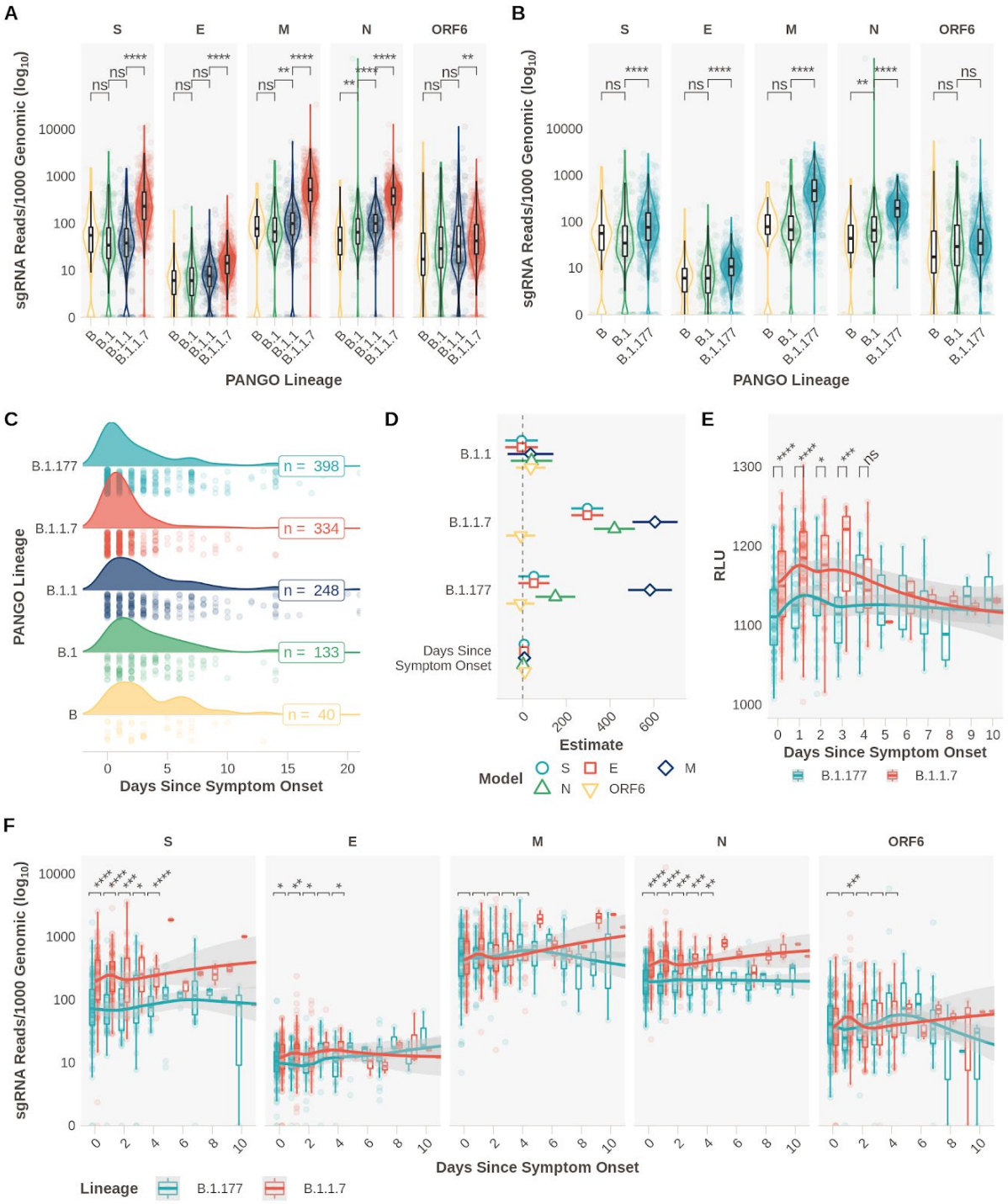


Figure 2. Relationship between day of symptom onset at sampling and subgenomic RNA levels in B.1.1.7 and B.1.177 infections

A. *Subgenomic RNA data from most highly expressed ORFs for B.1.1.7 and respective ancestral lineages in Pillar 1 data (B=$_n$51, B.1=$_n$164, B.1.1$_n$=302, ,B.1.1.7$_n$=729).* ***B.*** *Subgenomic RNA data from most highly expressed ORFs for B.1.177 and respective*

ancestral lineages in Pillar 1 data (B_n=51, $B.1_n$=164, $B.1.177_n$=764). ***C.*** *Distribution of the reported days since symptom at the time of sampling for B (n=40), B.1 (n=133), B.1.1 (n=248), B.1.177 (n=398) and B.1.1.7 (n=334) lineage infections.* ***D.*** *Model estimates and 95% confidence intervals from generalized linear model comparing effect of lineage and day of symptom onset on sgRNA expression for mostly highly abundant ORFs. Reference lineage B.1. No interaction was observed between any lineage and day of symptom onset. One B.1 sample with outlier N expression (SHEF-CC370) was excluded.* ***E.*** *Pseudo time course of RLU values for B.1.1.7 and B.1.177, plot restricted to 10 days. (0 Days; $B.1.177_n$ = 92, $B.1.1.7_n$= 106. 1 Day; $B.1.177_n$ = 52, $B.1.1.7_n$= 102. 2 Days; $B.1.177_n$ = 28, $B.1.1.7_n$= 43. 3 Days; $B.1.177_n$ = 31, $B.1.1.7_n$= 14. 4 Days; $B.1.177_n$ = 17, $B.1.1.7_n$= 19.).* ***F.*** *Normalized subgenomic RNA expression for each of the highest expressed ORFs stratified by the reported days since symptom onset, loess model. Note: differing y-axis scales, plot restricted to 10 days. (0 Days; $B.1.177_n$ = 115, $B.1.1.7_n$= 93. 1 Day; $B.1.177_n$ = 53, $B.1.1.7_n$= 93. 2 Days; $B.1.177_n$ = 31, $B.1.1.7_n$= 47. 3 Days; $B.1.177_n$ = 31, $B.1.1.7_n$= 12. 4 Days; $B.1.177_n$ = 19, $B.1.1.7_n$= 18.). All p-values calculated using an unpaired Wilcoxon signed rank test, and adjusted for multiple testing with Holm (**** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05, ns or blank = not significant).*

Noncanonical sgRNA are the result of RdRp template switching from regions of the genome with no canonical TRS-B site. These noncanonical sgRNA tend to be enriched around canonical sites, presumably due to the frequency of RdRp template switching that occurs in close proximity. We quantified the expression of these noncanonical sgRNA in pillar 1 B.1.1.7 infections. We observed low levels of these transcripts throughout the genome across all samples (*11*) and compared the presence of noncanonical sgRNA transcripts between B.1.1.7 and B.1.177 infections (Figure 3A). In addition to the noncanonical sgRNA resulting from the nucleocapsid R203K/G204R mutation (N*, Figure S2) (*17*), B.1.1.7 samples were also significantly enriched for reads which support the production of a noncanonical sgRNA from genomic position 28282 (Figure 3B,C&D, Fig S9, Wilcoxon unpaired p = <2e-16, Wilcoxon effect size = 0.797, $B.1.1.7_n$=717, $other_n$=430). The B.1.1.7 genomic sequence contains a triple nucleotide mutation at 28280, 28281, and 28282

GAT>CTA, resulting in an amino acid substitution, D3L, in the nucleocapsid protein. This triplet results in enhanced complementarity between the 28282 genomic region and the sequence 3' of the leader at the 5' end of the genome (Figure 3E), which may have resulted in a novel TRS-B-like site that drives higher sgRNA production from this locus. Of note, this site is upstream of the ORF9b ATG and this noncanonical sgRNA retains the full coding region of the ORF9b, but lacks the canonical start codon of N. We propose that this represents the ORF9b sgRNA, which was detected at low levels in 13.8% (430/3127) non-B.1.1.7 samples (median normalised expression=0.32), but is found in almost all pillar 1 B.1.1.7 infections (98.4%, 717/729), with a 16-fold increase in expression (median normalised expression=6.02 , Figure 3A&C).

Interestingly, we noticed non-B.1.1.7 samples in our dataset with low levels of putative ORF9b sgRNA did not contain the nucleocapsid D3L mutation. We propose the following hypothesis to explain our findings: in non-B.1.1.7 SARS-CoV-2, the weak complementarity around the ORF9b ATG causes low levels of noncanonical sgRNA to be produced which has high complementarity to the genome, but crucially contains the CTA triplet (Figure S9A). This could have led to a transcriptionally driven recombination event resulting in the GAT > CTA mutation in the genomic sequence (Figure 3E panel 2&3). This new TRS-B-like site with increased complementarity in B.1.1.7 viruses, drives higher sgRNA production from this locus, akin to sites of canonical sgRNA production (Figure 3E Panel 4). A similar event has been described previously more 3' in the N ORF (*17*), and mutations which result in increased complementarity and noncanonical sgRNA formation have been shown to occur in coronaviruses (*20*). This raises the possibility that an ORF9b protein could be produced independently of other upstream sgRNAs with greater efficiency in B.1.1.7. ORF9b has been shown to regulate interferon responses (*21*–*23*) and has been shown to be present in SARS-CoV (*24*) and SARS-CoV-2 infections (*13*, *24*, *25*). This finding could therefore be important for investigating the greater transmissibility of B.1.1.7 and requires urgent experimental validation. We explored whether the nucleocapsid D3L mutation has occurred in non-B.1.1.7 lineages in the global COG-UK sequence data set (14th of February 2021). 77 sequences with D3L were noted in a variety of SARS-CoV-2 lineages and appear as

homoplasies in a global phylogeny, suggesting that this event may have occurred independently on several occasions (Figure S10).

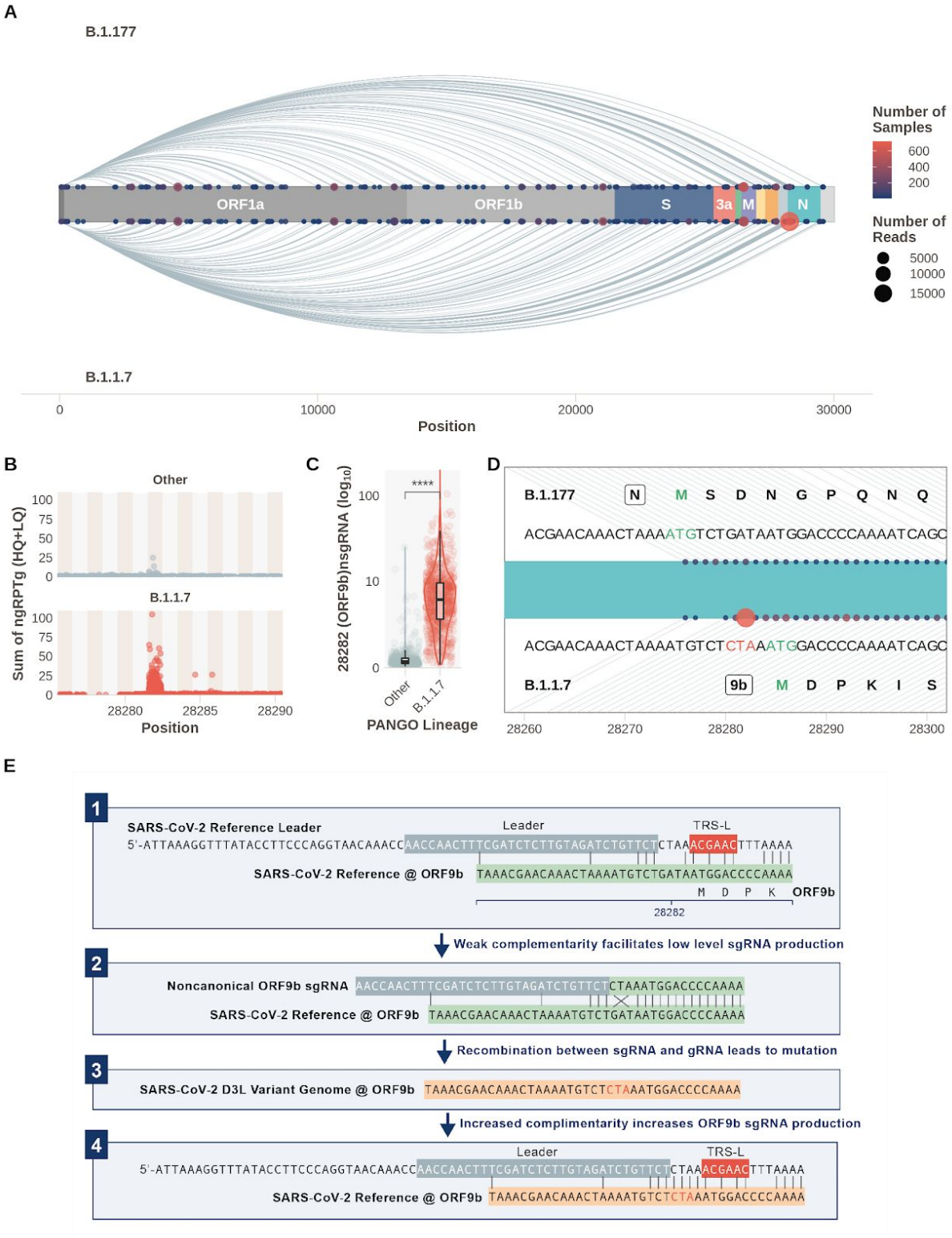


Figure 3. A Noncanonical sgRNA representing ORF9b is highly expressed in B.1.1.7 due to a triple nucleotide mutation in nucleocapsid leading to the D3L substitution.

A. *Subgenomic RNA not attributed to a canonical TRS-B site in B.1.177 (top) and B.1.1.7 (bottom) SARS-CoV-2 infections (Pillar 1 data). Size of point is number of reads, colour is*

number of samples (N excluded for clarity).* ***B.*** *Total normalized noncanonical sgRNA (ngRPTg) in B.1.1.7 vs other lineages.* ***C.*** *B.1.1.7 has significantly increased expression of a noncanonical sgRNA at 28282. (Wilcoxon effect size, unpaired = 0.797, $Other_n$=430, $B.1.1.7_n$=717).* ***D.*** *Schematic of the noncanonical sgRNA (points) in the context of the SARS-CoV-2 genome around position 28282. Top shows the sequence present in B.1.177 and the amino acid sequence of the N protein. Bottom shows the sequence present in B.1.1.7 with the triplet CTA mutation and the closest ATG which represents the ORF9b methionine.* ***E.*** *Proposed model for sgRNA driven mutation of N D3L which leads to high ORF9b expression in B.1.1.7. All p-values calculated using an unpaired Wilcoxon signed rank test, and adjusted for multiple testing with holm (**** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05, ns = not significant)*

Taken together, our data suggests that sgRNA expression from existing ARTIC Nanopore sequencing data can be used in real time to examine the effect of SARS-CoV-2 variation on its ability to express its genome. We cannot say if the increased sgRNA expression is the cause or consequence of an increase in viral replication or a more efficacious entry, but our study provides further insight to guide exploration with mechanistic studies. A major advantage of this approach is that we can deconvolute the contribution of genomic and subgenomic RNA, which is impossible with current diagnostic PCR assays and we can, additionally, examine all ORFs simultaneously and discover noncanonical subgenomic RNA which could be of biological relevance. Finally we believe that sgRNA expression analysis should be carried out on all compatible genomic surveillance platforms to give an instant readout of altered expression profiles in emerging SARS-CoV-2 variants. This would use existing data to complement epidemiological and phylodynamic methods, and provide an early warning of variants that might be of concern with regards to greater transmissibility and/or disease severity.

Methods

Diagnostic SARS-CoV-2 Testing

SARS-CoV-2 positivity was determined from nose/throat swabs diagnostically by Sheffield Teaching Hospitals NHS Foundation trust either using the Hologic Panther to generate Relative Light Units (RLU) (*26*) or an in house dual E/RdRp real time PCR assay to generate a cycle threshold (ECT or RCT respectively) (*16*).

Sample Preparation, ARTIC Network PCR and Nanopore Sequencing

RNA was extracted from viral transport medium (VTM) with Qiagen QIAamp MinElute Virus Spin Kit (50). Resultant RNA was subject to the ARTIC network (*10*) tiled amplicon protocol and subsequently sequenced on an Oxford Nanopore GridION X5. Bases were called with the default basecaller in MinKNOW (currently guppy v4) with *--require-both-ends* set for de-multiplexing. Raw sequencing fastq files were subject to no processing after sequencing to ensure sgRNA leader sequences are retained as described previously (*11*).

Subgenomic Read Classification and Normalisation

Subgenomic RNAs were classified using periscope (*11*) as described previously. Briefly, reads containing the leader sequence at their start are identified by a local alignment, the quality of this alignment determines which quality “bin” sgRNA reads are placed (HQ, LQ or LLQ). The amplicon from which the sgRNA evidence was generated is determined and a count of genomic reads for this amplicon used to normalise the raw sgRNA read counts. Samples were excluded from the subsequent analysis if:

- Their consensus coverage was < 0.9
- They had less than 50,000 mapped reads (we have previously shown that fewer reads produce a less robust analysis)

Lineage Assignment

Lineages (*1*) were assigned using Pangolin (https://github.com/cov-lineages/pangolin v2.1.7).

Figure Generation, Statistical Analysis

Statistical analysis was performed in R (*27*). Figures were generated in R using tidyverse (*28*), apart from those that depict sequencing reads, which were generated in IGV (*29*). Tests between groups were performed using an unpaired Wilcoxon test using the rstatix package (https://github.com/kassambara/rstatix) , adjusting p-values for any multiple comparisons using the Holm method.

Phylogenetic Tree Generation

The grapevine pipeline (https://github.com/COG-UK/grapevine) was used for generating the phylogeny based on all data available on GISAID and COG-UK up until 16th February 2021. A representative sample of global sequences was obtained in 2 steps. First by randomly selecting one sequence per country per epi week followed by random sampling of the remaining sequences to generate a sample of 4000 sequences. The global tree was then pruned using code adapted from the tree-manip package (https://github.com/josephhughes/tree-manip). We then identified samples with D3L mutations and colour coded these tips according to their lineages. The visualisation was produced using R/ape, R/ggplot2, R/ggtree, R/treeio, R/phangorn, R/stringr, R/dplyr, R/aplot.

Ethics Approval and Consent

Individuals presenting with active COVID-19 disease were sampled for SARS CoV-2 sequencing at Sheffield Teaching Hospitals NHS Foundation Trust, UK using samples

collected for routine clinical diagnostic use. This work was performed under approval by the Public Health England Research Ethics and Governance Group for the COVID-19 Genomics UK consortium (R&D NR0195).

Data Availability

All SARS-CoV-2 consensus sequences that are of high enough quality are available on GISAID and ENA and from https://www.cogconsortium.uk/data/. All sgRNA expression data is provided as supplementary files and at https://github.com/sheffield-bioinformatics-core/periscope-variants-publication. All raw sequencing data will be available on ENA as soon as possible.

Code Availability

Periscope is available at https://github.com/sheffield-bioinformatics-core/periscope and the code used to generate the figures contained within this manuscript can be found as a supplementary files or at https://github.com/sheffield-bioinformatics-core/periscope-variants-publication

References

1. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. **5**, 1403–1407 (2020).

2. K. Leung, M. H. Shum, G. M. Leung, T. T. Lam, J. T. Wu, Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveill.* **26** (2021), doi:10.2807/1560-7917.ES.2020.26.1.2002106.

3. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, W. J. Edmunds, CMMID COVID-19 Working Group, Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *bioRxiv* (2020), , doi:10.1101/2020.12.24.20248822.

4. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings (2021), (available at https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586).

5. C. M. Voloch, R. da Silva F, L. G. P. de Almeida, C. C. Cardoso, O. J. Brustolini, A. L. Gerber, A. P. de C Guimarães, D. Mariani, R. M. da Costa, O. C. Ferreira, C.-U. Workgroup, LNCC-Workgroup, A. C. Cavalcanti, T. S. Frauches, C. M. B. de Mello, R. M. Galliez, D. S. Faffe, Terezinha M P, A. Tanuri, A. T. R. de Vasconcelos, Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *medRxiv*, 2020.12.23.20248598 (2020).

6. M. Kidd, A. Richter, A. Best, N. Cumley, J. Mirza, B. Percival, M. Mayhew, O. Megram, F. Ashford, T. White, E. Moles-Garcia, L. Crawford, A. Bosworth, S. F. Atabani, T. Plant, A. McNally, S-variant SARS-CoV-2 lineage B1.1.7 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-qPCR. *J. Infect. Dis.* (2021), doi:10.1093/infdis/jiab082.

7. T. Golubchik, K. A. Lythgoe, M. Hall, L. Ferretti, H. R. Fryer, G. MacIntyre-Cockett, M. de Cesare, A. Trebes, P. Piazza, D. Buck, J. A. Todd, C. Fraser, D. Bonsall, The COVID-19 Genomics UK (COG-UK) consortium, Early analysis of a potential link between viral load and the N501Y mutation in the SARS-COV-2 spike protein. *bioRxiv* (2021), , doi:10.1101/2021.01.12.20249080.

8. N. G. Davies, C. I. Jarvis, W. J. Edmunds, N. P. Jewell, K. Diaz-Ordaz, R. H. Keogh, Increased hazard of death in community-tested cases of SARS-CoV-2 Variant of Concern 202012/01. *medRxiv* (2021), doi:10.1101/2021.02.01.21250959.

9. J. C. Brown, D. H. Goldhill, J. Zhou, T. P. Peacock, R. Frise, N. Goonawardane, L. Baillon, R. Kugathasan, A. L. Pinto, P. F. McKay, J. Hassard, M. Moshe, A. Singanayagam, T. Burgoyne, the ATACCC Investigators, PHE Virology Consortium, W. S. Barclay, 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. *Cold Spring Harbor Laboratory* (2021), p. 2021.02.24.432576.

10. Artic Network (2020), (available at https://artic.network/ncov-2019).

11. M. D. Parker, B. B. Lindsey, S. Leary, S. Gaudieri, A. Chopra, M. Wyles, A. Angyal, L. R. Green, P. Parsons, R. M. Tucker, R. Brown, D. Groves, K. Johnson, L. Carrilero, J. Heffer, D. Partridge, C. Evans, M. Raza, A. J. Keeley, N. Smith, D. Wang, S. Mallal, T. I.

de Silva, periscope: sub-genomic RNA identification in SARS-CoV-2 ARTIC Network Nanopore Sequencing Data. *bioRxiv* (2020), p. 2020.07.01.181867.

12. D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, H. Chang, The Architecture of SARS-CoV-2 Transcriptome. *Cell*. **181**, 914–921.e10 (2020).

13. Y. Finkel, O. Mizrahi, A. Nachshon, S. Weingarten-Gabbay, D. Morgenstern, Y. Yahalom-Ronen, H. Tamir, H. Achdout, D. Stein, O. Israeli, A. Beth-Din, S. Melamed, S. Weiss, T. Israely, N. Paran, M. Schwartz, N. Stern-Ginossar, The coding capacity of SARS-CoV-2. *Nature* (2020), doi:10.1038/s41586-020-2739-1.

14. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk, An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe*. **1**, e99–e100 (2020).

15. Tracking SARS-CoV-2 VOC 202012/01 (lineage B.1.1.7) dissemination in Portugal: insights from nationwide RT-PCR Spike gene drop out data (2021), (available at https://virological.org/t/tracking-sars-cov-2-voc-202012-01-lineage-b-1-1-7-dissemination-in-portugal-insights-from-nationwide-rt-pcr-spike-gene-drop-out-data/600).

16. H. Colton, M. Ankcorn, M. Yavuz, L. Tovey, A. Cope, M. Raza, A. J. Keeley, A. State, B. Poller, M. Parker, T. I. de Silva, C. Evans, Improved sensitivity using a dual target, E and RdRp assay for the diagnosis of SARS-CoV-2 infection: Experience at a large NHS Foundation Trust in the UK. *J. Infect.* (2020), doi:10.1016/j.jinf.2020.05.061.

17. S. Leary, S. Gaudieri, A. Chopra, S. Pakala, E. Alves, M. John, S. Das, S. Mallal, E. Phillips, Three adjacent nucleotide changes spanning two residues in SARS-CoV-2 nucleoprotein: possible homologous recombination from the transcription-regulating sequence. *bioRxiv* (2020), p. 2020.04.10.029454.

18. F. Ali, A. Kasry, M. Amin, The New SARS-CoV-2 Strain Shows a Stronger Binding Affinity to ACE2 Due to N501Y Mutation. *arXiv [physics.bio-ph]* (2021), (available at http://arxiv.org/abs/2101.01791).

19. J. C. Santos, G. A. Passos, The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. *Cold Spring Harbor Laboratory* (2021), p. 2020.12.29.424708.

20. S. Zúñiga, I. Sola, S. Alonso, L. Enjuanes, Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. *J. Virol.* **78**, 980–994 (2004).

21. H.-W. Jiang, H.-N. Zhang, Q.-F. Meng, J. Xie, Y. Li, H. Chen, Y.-X. Zheng, X.-N. Wang, H. Qi, J. Zhang, P.-H. Wang, Z.-G. Han, S.-C. Tao, SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. *Cell. Mol. Immunol.* **17**, 998–1000 (2020).

22. L. Han, M.-W. Zhuang, Y. Zheng, J. Zhang, M.-L. Nan, P.-H. Wang, C. Gao, SARS-CoV-2 ORF9b Antagonizes Type I and III Interferons by Targeting Multiple Components of RIG-I/MDA-5-MAVS, TLR3-TRIF, and cGAS-STING Signaling Pathways. *Cold Spring Harbor Laboratory* (2020), p. 2020.08.16.252973.

23. J. Wu, Y. Shi, X. Pan, S. Wu, R. Hou, Y. Zhang, T. Zhong, H. Tang, W. Du, L. Wang, J. Wo, J. Mu, Y. Qiu, K. Yang, L.-K. Zhang, B.-C. Ye, N. Qi, SARS-CoV-2 ORF9b inhibits RIG-I-MAVS antiviral signaling by interrupting K63-linked ubiquitination of NEMO. *Cell Rep.* **34**, 108761 (2021).

24. J.-P. Guo, M. Petric, W. Campbell, P. L. McGeer, SARS corona virus peptides recognized by antibodies in the sera of convalescent cases. *Virology*. **324**, 251–256 (2004).

25. H.-W. Jiang, Y. Li, H.-N. Zhang, W. Wang, X. Yang, H. Qi, H. Li, D. Men, J. Zhou, S.-C. Tao, SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG

and IgM responses. *Nat. Commun.* **11**, 3581 (2020).

26. A. J. Gorzalski, H. Tian, C. Laverdure, S. Morzunov, S. C. Verma, S. VanHooser, M. W. Pandori, High-Throughput Transcription-mediated amplification on the Hologic Panther is a highly sensitive method of detection for SARS-CoV-2. *J. Clin. Virol.* **129**, 104501 (2020).

27. R Core Team, R: A Language and Environment for Statistical Computing, (available at https://www.R-project.org).

28. H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, Others, Welcome to the Tidyverse. *Journal of Open Source Software*. **4**, 1686 (2019).

29. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

Acknowledgements

We thank the Sheffield Bioinformatics Core for their useful thoughts and discussions. We would like to thank the members of the Sheffield Biomedical Research centre for their continued support of the SARS-CoV-2 sequencing work in Sheffield.

Author Contributions

Contributor Role*	Contributor
Conceptualization	MDP, TIdS
Data curation	MDP, TIdS, BL, DS, AK, DP, SH, MR, CE
Formal Analysis	MDP, TIdS, BL, SH
Funding acquisition	DW, TIdS
Investigation	PZ, MG, SL, MW, BF, SC, PW, MP, LG
Methodology	MDP, TIdS
Project administration	NS, SH
Resources	CE, MR, DP, AC, AS, KJ, NA, RR
Software	MDP, JH
Supervision	MDP, DW, TIdS
Validation	Not Applicable
Visualization	MDP, BL, SH
Writing – original draft	MDP, TIdS
Writing – review & editing	MDP, DW, BL, SH, TIdS

*Contributions as described at: https://casrai.org/credit/

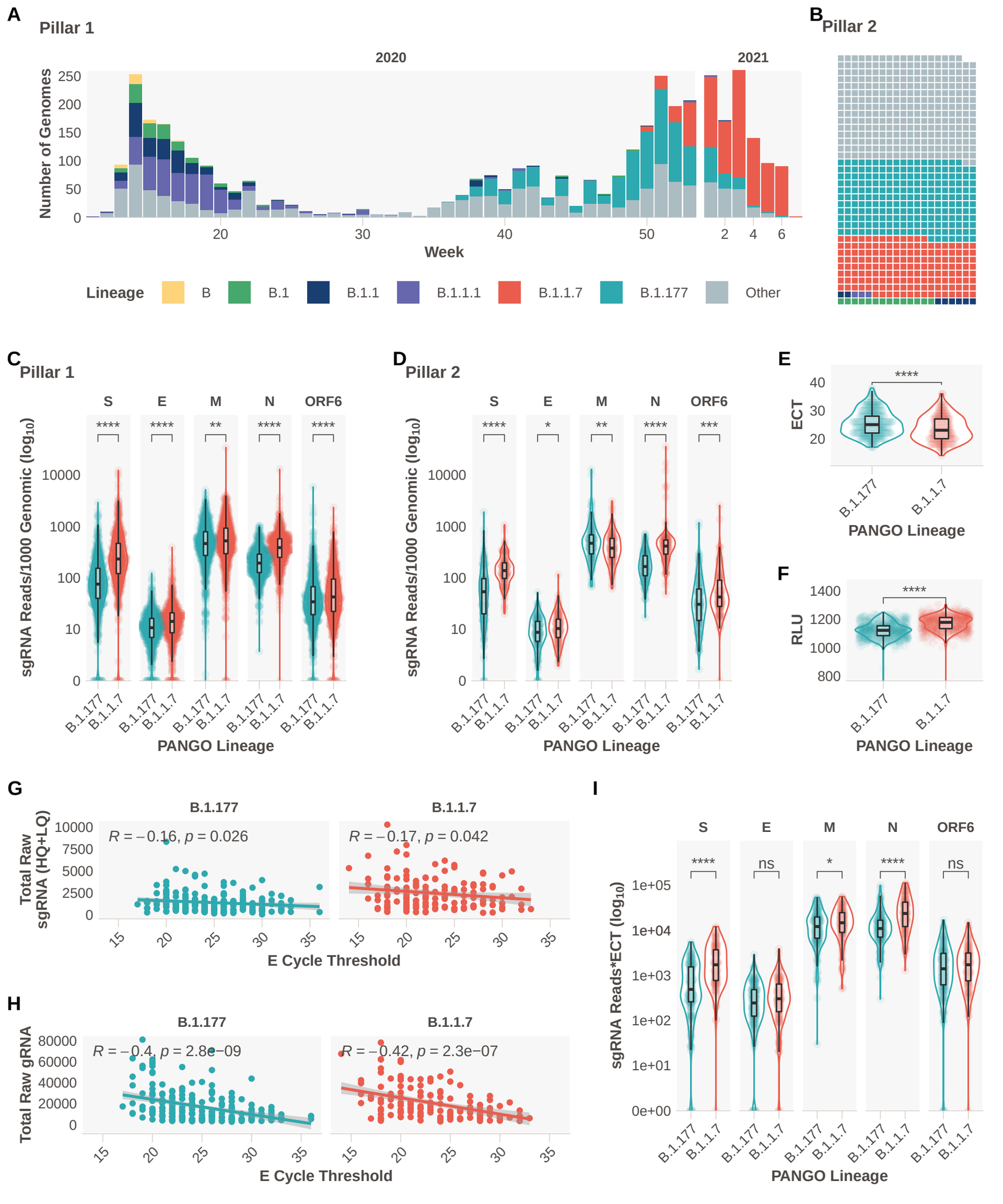
Funding

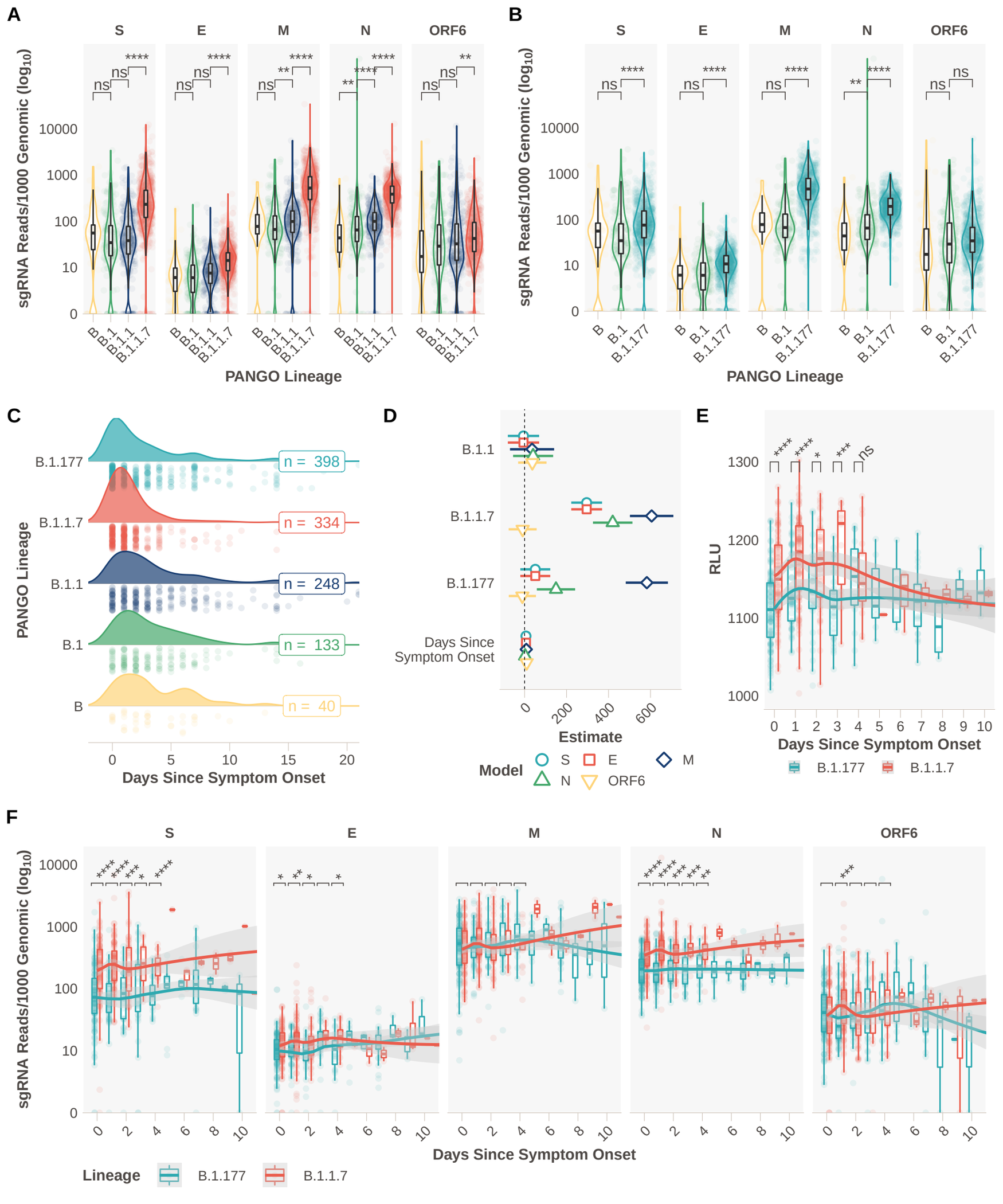
Sequencing of SARS-CoV-2 samples was undertaken by the Sheffield COVID-19 Genomics Group as part of the COG-UK CONSORTIUM and supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of

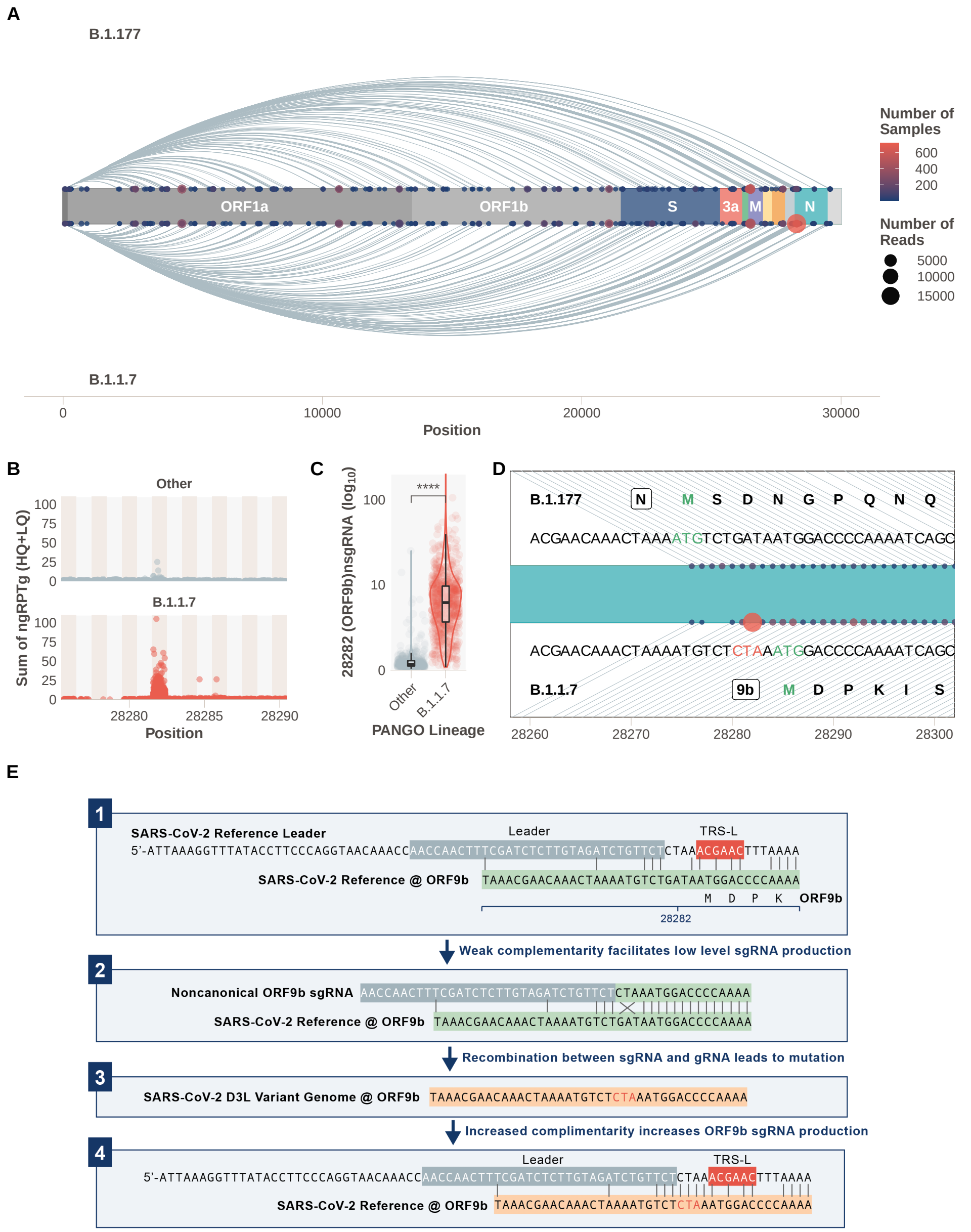
Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. MDP and DW are funded by the NIHR Sheffield Biomedical Research Centre (BRC - IS-BRC-1215-20017). TIdS is supported by a Wellcome Trust Intermediate Clinical Fellowship (110058/Z/15/Z). We thank all partners of and contributors to the COG-UK consortium, who are listed at https://www.cogconsortium.uk/about/.

Competing Interests

The authors note no competing interests.


A
Pillar 1
2020
2021
Number of Genomes
250
200
150
100
50
0
20
30
40
50
2
4
6
Week
Lineage
B
B.1
B.1.1
B.1.1.1
B.1.1.7
B.1.177
Other
B
Pillar 2
C
Pillar 1
S
E
M
N
ORF6
****
**
sgRNA Reads/1000 Genomic (log10)
10000
1000
100
10
0
B.1.177
B.1.1.7
PANGO Lineage
D
Pillar 2
*
***
E
ECT
40
30
20
F
RLU
1400
1200
1000
800
G
B.1.177
B.1.1.7
Total Raw sgRNA (HQ+LQ)
10000
7500
5000
2500
0
R = −0.16, p = 0.026
R = −0.17, p = 0.042
15
20
25
30
35
E Cycle Threshold
H
Total Raw gRNA
80000
60000
40000
20000
R = −0.4, p = 2.8e−09
R = −0.42, p = 2.3e−07
I
ns
sgRNA Reads*ECT (log10)
1e+05
1e+04
1e+03
1e+02
1e+01
0e+00


A
S
E
M
N
ORF6
sgRNA Reads/1000 Genomic (log_{10})
10000
1000
100
10
0
ns
****
**
B
B.1
B.1.1
B.1.1.7
PANGO Lineage
B
B.1
B.1.177
C
PANGO Lineage
B.1.177
B.1.1.7
B.1.1
B.1
B
n = 398
n = 334
n = 248
n = 133
n = 40
0
5
10
15
20
Days Since Symptom Onset
D
B.1.1
B.1.1.7
B.1.177
Days Since Symptom Onset
0
200
400
600
Estimate
Model
S
E
M
N
ORF6
E
RLU
1300
1200
1100
1000
****
*
***
ns
0 1 2 3 4 5 6 7 8 9 10
Days Since Symptom Onset
B.1.177
B.1.1.7
F
S
E
M
N
ORF6
sgRNA Reads/1000 Genomic (log_{10})
10000
1000
100
10
0
****
***
**
*
0 2 4 6 8 10
Days Since Symptom Onset
Lineage
B.1.177
B.1.1.7


A
B.1.177
ORF1a
ORF1b
S
3a
M
N
Number of Samples
600
400
200
Number of Reads
5000
10000
15000
B.1.1.7
0
10000
20000
30000
Position
B
Other
B.1.1.7
Sum of ngRPTg (HQ+LQ)
100
75
50
25
0
28280
28285
28290
Position
C
28282 (ORF9b)nsgRNA (log10)
****
100
10
0
Other
B.1.1.7
PANGO Lineage
D
B.1.177
N
M S D N G P Q N Q
ACGAACAAACTAAAATGTCTGATAATGGACCCCAAAATCAGC
ACGAACAAACTAAAATGTCTCTAAATGGACCCCAAAATCAGC
B.1.1.7
9b
M D P K I S
28260
28270
28280
28290
28300
E
1
SARS-CoV-2 Reference Leader
Leader
TRS-L
5'-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAA
SARS-CoV-2 Reference @ ORF9b
TAAACGAACAAACTAAAATGTCTGATAATGGACCCCAAAA
M D P K
ORF9b
28282
Weak complementarity facilitates low level sgRNA production
2
Noncanonical ORF9b sgRNA
AACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAATGGACCCCAAAA
SARS-CoV-2 Reference @ ORF9b
TAAACGAACAAACTAAAATGTCTGATAATGGACCCCAAAA
Recombination between sgRNA and gRNA leads to mutation
3
SARS-CoV-2 D3L Variant Genome @ ORF9b
TAAACGAACAAACTAAAATGTCTCTAAATGGACCCCAAAA
Increased complimentarity increases ORF9b sgRNA production
4
Leader
TRS-L
5'-ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAA
SARS-CoV-2 Reference @ ORF9b
TAAACGAACAAACTAAAATGTCTCTAAATGGACCCCAAAA