periscope: sub-genomic RNA identification in SARS-CoV-2 ARTIC Network Nanopore Sequencing Data

*Matthew D Parker[1,9], *Benjamin B Lindsey[4,5], Shay Leary[2], Silvana Gaudieri[2,3,6], Abha Chopra[2], Matthew Wyles[9], Adrienn Angyal[5], Luke R Green[5], Paul Parsons[7], Rachel M Tucker[7], Rebecca Brown[5], Danielle Groves[5], Katie Johnson[4], Laura Carrilero [7], Joe Heffer[10], David Partridge[4,5], Cariad Evans[4], Mohammad Raza [4], Alexander J Keeley[4,5], Nikki Smith[5], Dennis Wang[1,8], Simon Mallal[2,3], Thushan I de Silva[4,5]

[1] Sheffield Bioinformatics Core, The University of Sheffield, Sheffield, UK

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

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

[4] Sheffield Teaching Hospitals NHS Foundation Trust, Department of Virology/Microbiology, Sheffield, UK.

[5] The Florey Institute, Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK.

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

[7] Department of Animal and Plant Sciences, Alfred Denny Building, The University of Sheffield, S10 2TN

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

[9] Neuroscience Institute, The University of Sheffield, Sheffield, UK

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

https://github.com/sheffield-bioinformatics-core/periscope

* These authors contributed equally to this work.

[+] Corresponding Author

Keywords

SARS-CoV-2, sub-genomic RNA, Nanopore, Expression

Abstract (250 words)

We have developed periscope, a tool for the detection and quantification of sub-genomic RNA in ARTIC network protocol generated Nanopore SARS-CoV-2 sequence data. We applied periscope to 1155 SARS-CoV-2 sequences from Sheffield, UK. Using a simple local alignment to detect reads which contain the leader sequence we were able to identify and quantify reads arising from canonical and non-canonical sub-genomic RNA. We were able to detect all canonical sub-genomic RNAs at expected abundances, with the exception of ORF10, suggesting that this is not a functional ORF. A number of recurrent non-canonical sub-genomic RNAs are detected. We show that the results are reproducible using technical replicates and determine the optimum number of reads for sub-genomic RNA analysis. Finally variants found in genomic RNA are transmitted to sub-genomic RNAs with high fidelity in most cases. This tool can be applied to tens of thousands of sequences worldwide to provide the most comprehensive analysis of SARS-CoV-2 sub-genomic RNA to date.

Introduction

SARS-CoV-2 is responsible for the most significant pandemic since 1918. Understanding variation within sub-genomic RNA synthesis within the human host may have important implications for the study of SARS-CoV-2 biology and evolution. Thanks to advances in sequencing technology and collaborative science, over 40,000 SARS-CoV-2 genomes have been sequenced worldwide in only the first six months of the outbreak.

The genome of SARS-CoV-2 comprises a single positive-sense RNA molecule of approximately 29kb in length. While the 1a and 1b polyproteins are translated directly from this genomic RNA (gRNA), all other proteins are translated from sub-genomic RNA intermediates(Stern and Kennedy 1980; Sola et al. 2015). Sub-genomic RNAs are produced through discontinuous transcription during negative strand synthesis followed by positive-strand synthesis to form mRNA. The resulting sub-genomic RNA contains a leader sequence derived from the 5’ untranslated region of the genome and a transcription regulating sequence (TRS) 5’ of the open reading frame (ORF). The template switch occurs during sub-genomic RNA synthesis due to a conserved core sequence within the TRS 5’ of each ORF (TRS-B) and the TRS within the leader sequence (TRS-L)(Zúñiga et al. 2004). The conserved core sequence leads to base pairing between the TRS-L and the nascent RNA molecule transcribed from the TRS-B resulting in a long-range template switch and incorporation of the 5' leader sequence(Sola et al. 2015). Understanding variation within sub-genomic RNA synthesis within the human host may have important implications for the study of SARS-CoV-2 biology and evolution.

Beyond the regulation of transcription, sub-genomic RNA may also play a role in the evolution of coronaviruses and the template switching required for sub-genomic RNA synthesis may explain the high rate of recombination seen in coronaviruses(Simon-Loriere and Holmes 2011; Wu and Brian 2010). While the majority of sub-genomic RNA relate to known ORFs, novel, non-canonical sub-genomic RNA are also produced(Kim et al.), although the biological function of this is unclear. Sub-genomic RNAs have also been shown to modulate host cell translational processes(Patel et al. 2013).

It has previously been shown that sub-genomic RNA transcript abundances can be quantified from full RNASeq data by calculation of Reads Per Kilobase of transcript, per Million mapped reads (RPKM) or by using so-called “chimeric” fragments containing the leader and TRS(Irigoyen et al. 2016). From two independent repeats, the correlation between these two measurement methods was 0.99 (R^2).

The ARTIC Network([CSL STYLE ERROR: reference with no printed form.]) protocol for the sequencing of the SARS-CoV-2 has been widely employed worldwide to characterise the genetic diversity of this novel coronavirus. The COVID-19 Genomics Consortium (COG-UK)(2020) in the UK, alone, has 10165 ARTIC nanopore sequences (Correct 12th June 2020), while internationally GISAID contains thousands more similar datasets (8775 with “nanopore” in the metadata and 3660 list “artic”, 14th June 2020). This protocol involves the amplification of 98 overlapping regions of the SARS-CoV-2 genome in two pools of 49 amplicons to provide full sequence coverage when sequenced with, in our case, Oxford

Nanopore sequencing devices. All known SARS-CoV-2 ORF TRS sites are contained within one or more amplicons in this panel (Figure 1A).

We therefore hypothesised that we could detect and quantify the levels of normalised sub-genomic RNA (sgRNA) to both identify novel non-canonical sgRNA and provide an estimate of ORF sub-genomic RNA expression in SARS-CoV-2 sequence data generated by the ARTIC Network protocol. Here we present a tool for these purposes from this rich dataset and apply it to 1155 SARS-CoV-2 sequences derived from clinical isolates in Sheffield, UK.

Results

Evidence for Sub-Genomic RNA

We designed a tool, periscope (https://github.com/sheffield-bioinformatics-core/periscope), to re-analyse raw ARTIC Network Nanopore sequencing data from SARS-CoV-2 isolates to identify sub-genomic RNA based on the detection of the leader sequence at the 5' end of reads as described previously(Leary et al. 2020).

The recommended bioinformatics standard operating procedure to process ARTIC network sequencing data to produce a consensus sequence involves selecting reads between 400 and 700 base pairs and the trimming of primer and adapter sequence. In most cases this removes reads which might provide evidence for sub-genomic RNA. Mapping raw data from this protocol reveals the presence of reads at ORF TRS sites which are sometimes shorter than the full ARTIC Network amplicon and contain leader sequence at their 5' end. These reads, we believe, are the result in the case of Pool 1, priming from primer 1 of the pool (which is homologous to most of the leader sequence) and also, in both pools uni-directional amplification from the 3’ primer which results in a truncated amplicon when the template is a sub-genomic RNA (Figure 1B). Interestingly we also see longer reads which are the result of priming from the 3’ end of the adjacent amplicon (Figure 1B).

To separate genomic from subgenomic reads, we employ the following workflow using snakemake(Köster and Rahmann 2012); Raw ARTIC Network Nanopore sequencing reads that pass QC are collected and aligned to the SARS-CoV-2 reference, reads are filtered out if they are unmapped or supplementary alignments. We do not perform any length filtering. Each read is assigned an amplicon. We search the read for the presence of the leader sequence (5’-AACCAACTTTCGATCTCTTGTAGATCTGTTCT-3’) using a local alignment. If we find the leader with a strong match it is likely that that read is from amplification of sub-genomic RNA. We assign reads to an ORF. Using all of this information we then classify each read into genomic, sub-genomic or non-canonical sub-genomic (Figure 1D) and produce summaries for each amplicon and ORF including normalised pseudo-expression values.

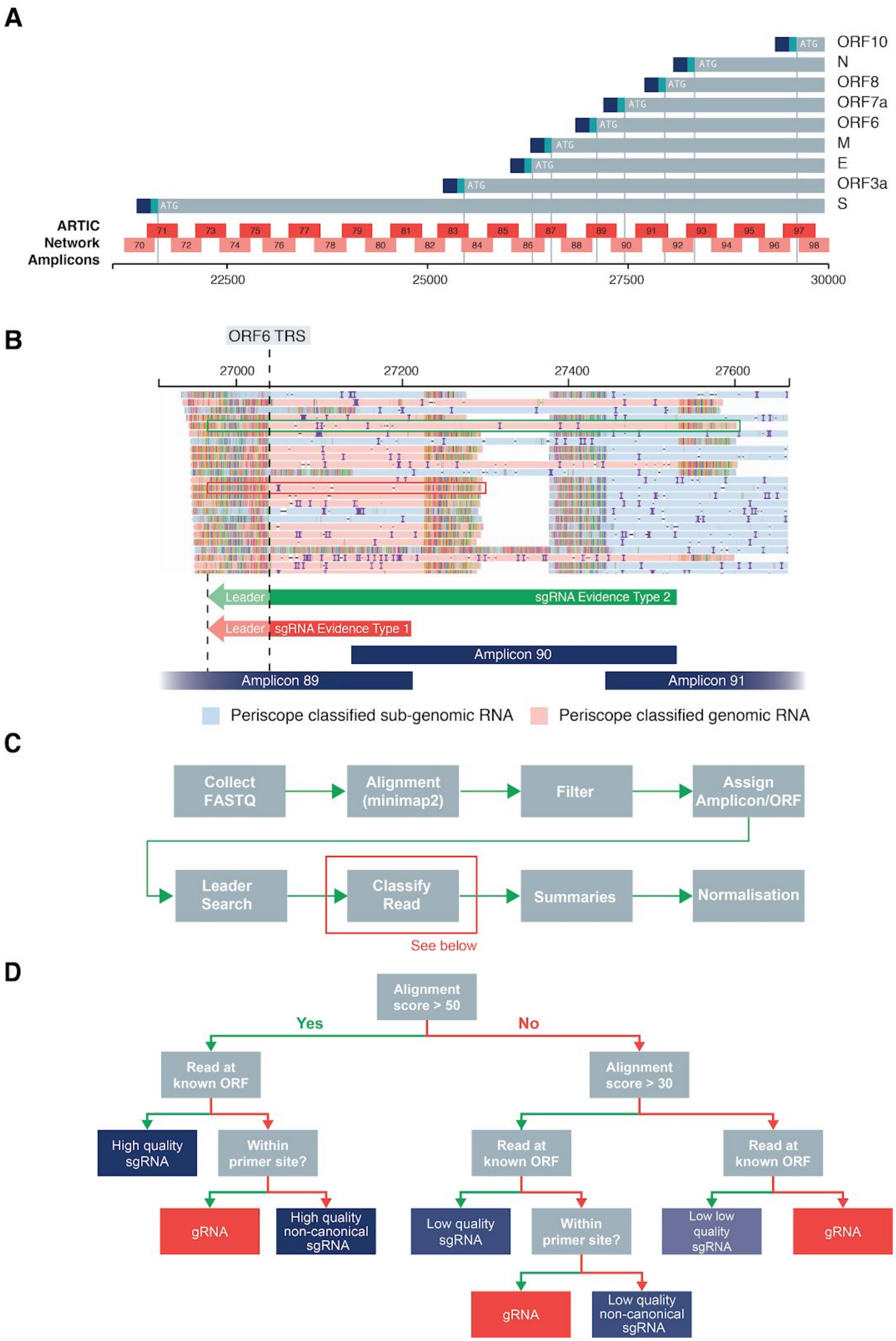

A
ORF10
N
ORF8
ORF7a
ORF6
M
E
ORF3a
S
ATG
ARTIC Network Amplicons
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
22500
25000
27500
30000
B
ORF6 TRS
27000
27200
27400
27600
Leader
sgRNA Evidence Type 2
Leader
sgRNA Evidence Type 1
Amplicon 90
Amplicon 89
Amplicon 91
Periscope classified sub-genomic RNA
Periscope classified genomic RNA
C
Collect FASTQ
Alignment (minimap2)
Filter
Assign Amplicon/ORF
Leader Search
Classify Read
Summaries
Normalisation
See below
D
Alignment score > 50
Yes
No
Read at known ORF
Alignment score > 30
High quality sgRNA
Within primer site?
Read at known ORF
Read at known ORF
gRNA
High quality non-canonical sgRNA
Low quality sgRNA
Within primer site?
Low low quality sgRNA
gRNA
gRNA
Low quality non-canonical sgRNA

Figure 1. Periscope Algorithm Design Details

A. *ARTIC network amplicon layout with respect to ORF TRS positions of SARS-CoV-2. Blue and aqua at the end of each orf signifies leader and TRS respectively.* ***B.*** *Read pileup at ORF6 TRS showing two types of reads which support the existence of sub-genomic RNAs. Type 1 (Red): Result from 3'->5' amplification from the closest primer to the 3' of the TRS site, and Type 2 (Green): These result from 3'->5' amplification from the adjacent amplicons 3' primer (i.e. the second closest 3' primer).* ***C.*** *Overview of periscope workflow.* ***D.*** *Decision tree for read classification.*

Detection of Sub-Genomic RNA

We were able to detect sub-genomic RNAs with a high leader alignment score from all canonical ORFs in multiple samples with the exception of ORF10 (Figure 2, Supplementary Table S1, Supplementary File S1). As shown in Figure 2, sub-genomic RNA from the N ORF was the most abundant sub-genomic RNA, found in 97.3% of samples, consistent with a published report using direct RNA Nanopore sequencing *in vitro(Kim et al.; Alexandersen et al. 2020)*. Genomic RNA raw reads also differed at ORF start sites (Figure 2A), presumably due to differences in amplification efficiency and number of mapped reads.

Like previously published reports(Kim et al.; Taiaroa et al. 2020; Alexandersen et al. 2020) we were unable to find strong evidence of sgRNA supporting the presence of ORF10 (Supplementary Table 2, Figure 2C) with only 0.95% of samples containing high or low quality sub-genomic RNA calls at this ORF. We aligned the 12 reads from these samples to a reference composing or ORF10 and leader (Supplementary Figure 2B). On manual review of these results ten (4 HQ) of these reads are falsely classified as sub-genomic RNA. Two reads remain, one read from each of samples SHEF-C0840 and SHEF-C58A5. These reads seem to represent ORF10 sub-genomic RNA as they have an almost complete match to the leader and the remainder of the reads is a strong match to ORF10.

Normalisation of Sub-Genomic Read Abundance

Beyond defining the presence of reads which are a result of amplification of sub-genomic RNA, we hypothesized that we could quantify the level of sub-genomic RNA present in a sample using either total mapped reads or genomic RNA reads from the same amplicon as denominators for normalisation. Normalisation of this kind would be analogous to traditional RNAseq analysis where reads per million (RPM) are calculated to allow comparisons between datasets where the number of reads affect the amount of each transcript detected. In the case of ARTIC Network Nanopore sequencing data, which involves PCR of small (~400bp) overlapping regions of the SARS-CoV-2 genome, amplification efficiency of each amplicon should also be taken into account. Because we have a median of 258,210 mapped reads for our samples (Supplementary Figure S1), we normalised both genomic RNA or sub-genomic RNA per 100,000 mapped reads (genomic reads per hundred thousand, gRPHT, or sub-genomic reads per hundred thousand, sgRPHT, respectively, Figures 2B and D).

For genomic RNA, normalising per 100,000 mapped reads across all samples reduces the standard deviation from 2966 to 784 across all ORFs, with the standard deviation for N reduced from 3159 to 591.

In our second approach, we determine the amplicon from which the sub-genomic RNA has originated, using methods from the ARTIC Network Field Bioinformatics package([CSL STYLE ERROR: reference with no printed form.]). We then normalise per 1000 gRNA reads from the same amplicon. If a sub-genomic RNA has resulted from more than one amplicon

(Figure 1B) the resulting normalised counts from each amplicon are summed giving us sub-genomic reads per 1000 genomic RNA reads (sgRPTg) for every ORF (Figure 2E).

Periscope outputs both methods of normalisation so that the user can decide which is more appropriate in their case and determine whether the conclusions of their analysis are consistent across both approaches.

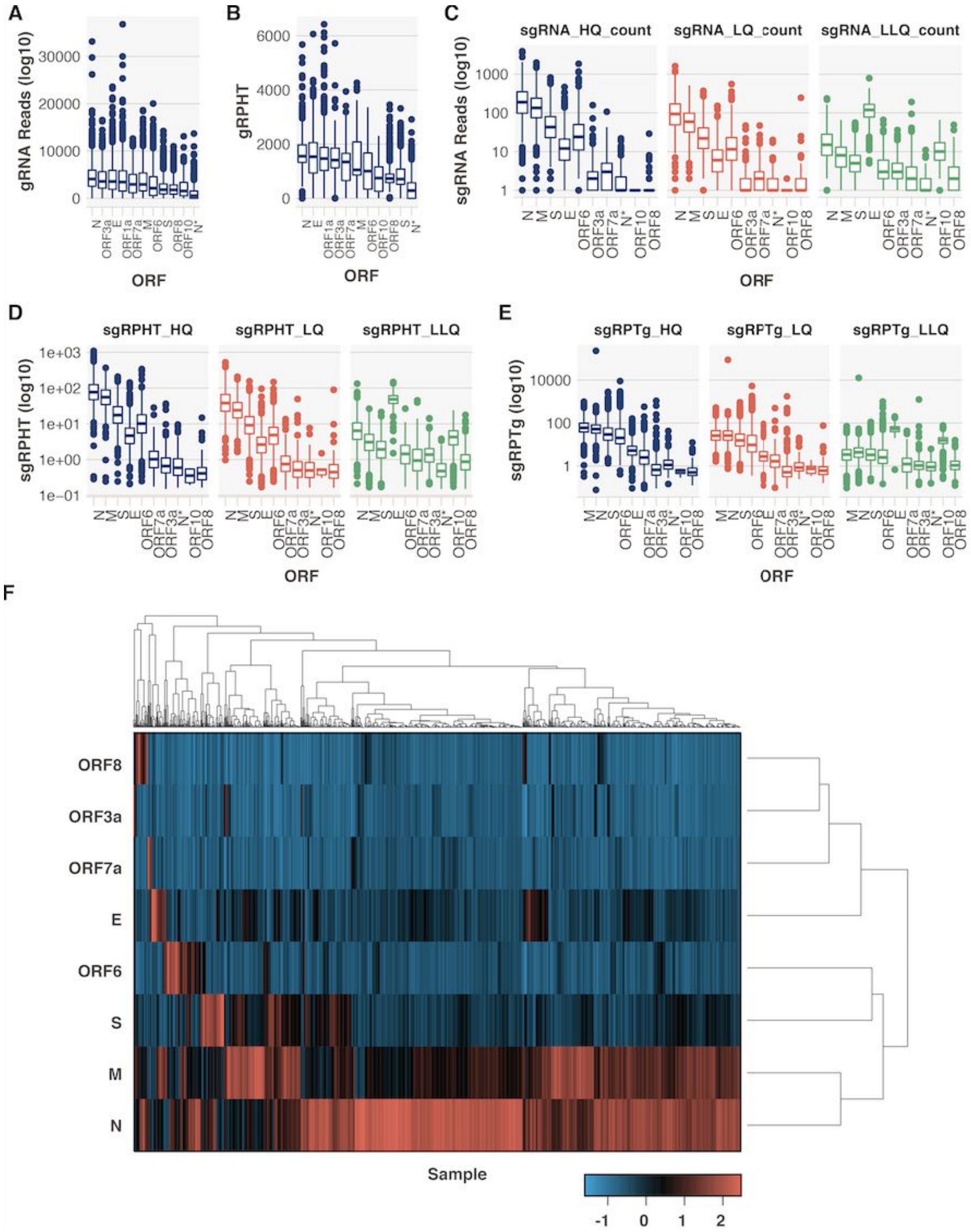


Figure 2. Detection and Quantification of Sub-Genomic RNA in SARS-CoV-2 Isolates in 1155 SARS-CoV-2 sequences

A. *Number of reads supporting genomic RNA at each of the ORFs. If multiple amplicons cover the ORF then this represents the sum of reads for those amplicons.* ***B.*** *Number of reads supporting sub-genomic RNA at each ORF.* ***C.*** *sub-genomic RNA levels separated by quality.* ***D****. sgRNA normalised to mapped reads.* ***E.*** *sgRNA normalised per 1000 genomic RNAs.* ***F.*** *Heatmap of subgenomic RNA expression across 1,155 SARS-CoV-2 sequences. Data is normalised per ORF and row means per sample subtracted and therefore negative heatmap values do not constitute down-regulation.*

Technical Replicates & Batch Effects

To assess the reproducibility of sub-genomic RNA analysis using ARTIC Network Nanopore sequencing data and periscope, we analysed two samples that were subject to four technical replicates each (Supplementary File S4); cDNA was independently prepared from the same swab extracted RNA, and subject to independent amplification using the recommended ARTIC Network polymerase chain reaction (PCR) and sequenced (Figures 3A & 3B). Pearson's correlation coefficient is >0.87 for all normalised sub-genomic abundances between replicates from the same sample.

Next, we treated our sub-genomic RNA abundance values like an RNASeq dataset and asked whether other factors could be influencing expression. To do this we used an unsupervised principal component analysis (Figure 3D, E, F and G) and coloured samples by the different categorical variables that could affect expression, like sequencing run ("batch effect"), the ARTIC primer version, the number of mapped reads and E gene cycle threshold (CT, Diagnostic test, normalised to RNAseP CT value, see materials and methods) as this is an indicator of the amount of virus present in an isolate and a proxy for quality (Figure 3G).

There did not appear to be a significant correlation between any of the above variables and the expression values in the PCA analysis.

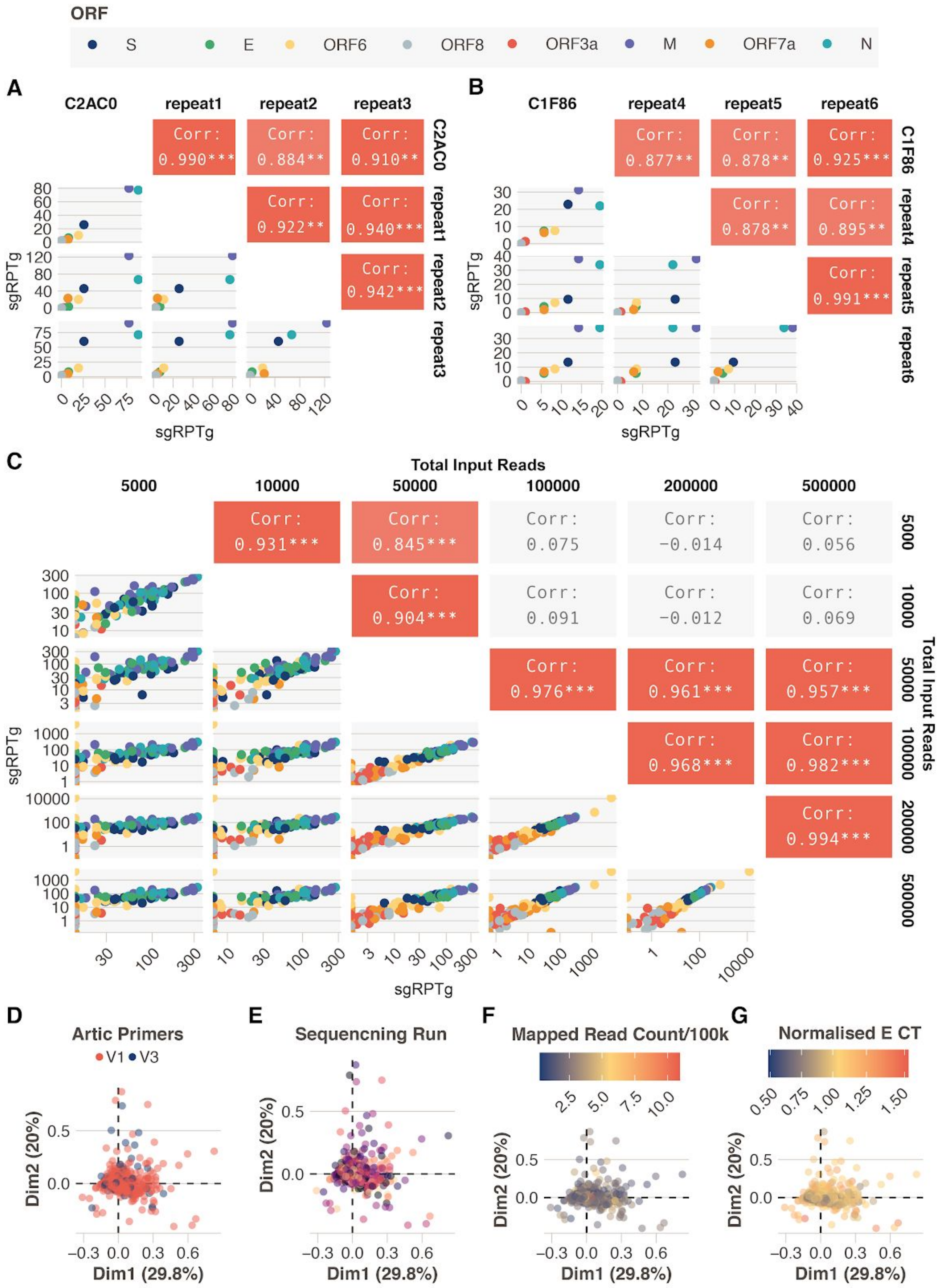


Figure 3. Technical Replicates, Detection Limit, and Batch Effects

***A & B.** Four technical replicates of two samples additional to the Sheffield cohort. Correlation coefficients between sub-genomic RNA normalised per 1000 genomic RNA reads (sgRPTg) from Pearson. **C.** Downsampling of reads from 23 high coverage (>1million mapped reads) samples. The number of reads as input to periscope was downsampled with seqtk(Li 2012) to 5,10,50,100,200, and 500 thousand reads. **D, E, F, G.** Unsupervised PCA analysis coloured by, primer version (**D**), sequencing run (**E**), total mapped read count (**F**), or normalised E CT (**G**).*

Lower Limit of Detection

The number of reads generated from any sequencing experiment is likely to vary between samples and between runs. The median mapped read count in our dataset is 258,210, but varies between 9105 and 3,260,686 (Supplementary Figure S1). In our experience we generally see low total amounts of sub-genomic RNA compared to the genomic counterparts, therefore its detection is likely to suffer when a sample has lower amounts of reads. To determine the effect of lower coverage on the detection of sub-genomic RNA we downsampled 23 samples which had > 1million mapped reads to lower read counts with seqtk(Li 2012). We chose high (500k reads) medium (200k and 100k), and low read counts (50k,10k and 5k) and ran periscope on this downsampled data (Supplementary File S3). In the absence of a ground truth, we correlated the abundance of sub-genomic RNA, pairwise, between downsampled datasets (Figure 3C). If coverage did not affect the abundance estimates then all coverage levels would show a high correlation coefficient when compared to each other. As expected, lower counts of 5,000 and 10,000 reads do not correlate with those generated from 100,000, 200,000 and 500,000 reads (R < 0.1). Samples with 50,000 reads seem to perform well when compared to 100,000 and 200,000 reads with a correlation coefficient of 0.976 and 0.961 respectively.

Non-Canonical Sub-Genomic RNA

In addition to sub-genomic RNA for known ORFs we can use periscope to detect novel, non-canonical sub-genomic RNA. We previously applied periscope to detect one such novel sub-genomic RNA (N*) which is a result of the creation of a new TRS site by a triplet variation at position 28881 to 28883 which results in production of a truncated N ORF(Leary et al. 2020).

To classify sub-genomic RNAs as non-canonical supporting reads must fulfil two criteria; 1. The start position does not fall in a known TRS region (+/- 10bp from the leader junction), and 2. The start position must not fall +/-5bp from a primer sequence. We chose to implement the second criteria because we noticed a pattern of novel sub-genomic RNAs being detected at amplicon edges due to erroneous leader matches to the primers sites.

We found evidence of non-canonical sub-genomic RNAs supported by 2 or more reads in 913 samples (Supplementary Figure S3), some of these could likely be reclassified as known, as we notice these predictions fall close to known ORF start sites (Supplementary File S2).

SHEF-C0118 contains 177 reads (HQ & LQ) which support a non-canonical sub-genomic RNA at position 25744 between ORFs 3a and E (Figure 4B). Total sgRNA count is high compared to most other ORFs in this isolate (Figure 4C, Supplementary Tables S3 & S4). Four other samples (26 with 1 or more) each contain >1 read supporting this sub-genomic RNA.

377 samples have evidence (1 or more, 155 have 2 or more, +/-5 from 10639) for a non-canonical sub-genomic RNA at position 10639 (HQ or LQ, Figure 4D). In this case there is a TRS-*like* sequence close to the leader in this non-canonical sub-genomic RNA; `ACGAAC` -> `ACG`**G**`AC`. Two samples have significant support with 103 HQ reads each (SHEF-CE04A, SHEF-CA0D5). It is possible that this represents an independent ORF1b sub-genomic RNA.

226 (1 or more, 62 have 2 or more, +/-5 from 5785) samples have evidence for a non-canonical sub-genomic RNA at position 5785 (Figure 4E), there is no core TRS sequence present and there doesn’t appear to be a productive start codon.

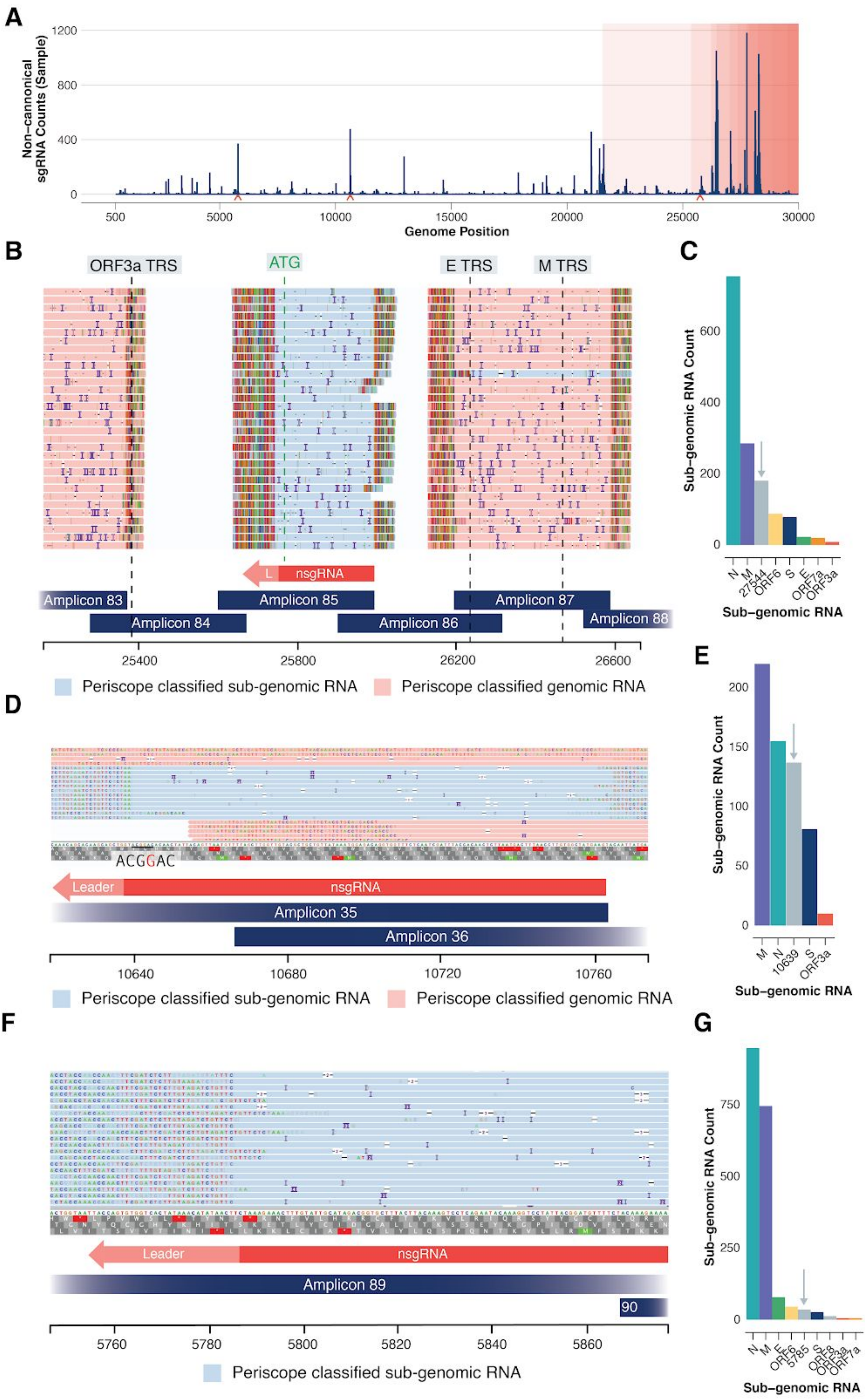

A
Non−cannonical sgRNA Counts (Sample)
1200
800
400
0
500
5000
10000
15000
20000
25000
30000
Genome Position
B
ORF3a TRS
ATG
E TRS
M TRS
L
nsgRNA
Amplicon 83
Amplicon 84
Amplicon 85
Amplicon 86
Amplicon 87
Amplicon 88
25400
25800
26200
26600
Periscope classified sub-genomic RNA
Periscope classified genomic RNA
C
Sub−genomic RNA Count
600
400
200
0
N
M
27544
ORF6
S
E
ORF7a
ORF3a
Sub−genomic RNA
D
ACGGAC
Leader
nsgRNA
Amplicon 35
Amplicon 36
10640
10680
10720
10760
Periscope classified sub-genomic RNA
Periscope classified genomic RNA
E
Sub−genomic RNA Count
200
150
100
50
0
M
N
10639
S
ORF3a
Sub−genomic RNA
F
Leader
nsgRNA
Amplicon 89
90
5760
5780
5800
5820
5840
5860
Periscope classified sub-genomic RNA
G
Sub−genomic RNA Count
750
500
250
0
N
M
E
ORF6
5785
S
ORF8
ORF3a
ORF7a
Sub−genomic RNA

Figure 4. Non-Canonical Sub-Genomic RNA

We classified reads as supporting non-canonical sub-genomic RNA as described in Figure 1D. ***A.*** *Histogram (bin width=100) of genome positions (>500) showing the number of samples a novel sub-genomic RNA has been detected. Red arrows indicate evidence highlighted in B and C.* ***B.*** *Non-canonical sub-genomic RNA with strong support in SHEF-C0118 at position 25744.* ***C.*** *Raw sub-genomic RNA levels (HQ&LQ) in SHEF-C0118 show high relative amounts of this non-canonical sub-genomic RNA at position 25744.* ***D.*** *Non-canonical sub-genomic RNA with strong support in SHEF-CE04A at 10369 is also supported in additional* 377 *samples. In close proximity to the leader is a sequence which could be considered a "weak" TRS;* *ACGAAC* -> *ACG**G**AC***.** ***E.*** *Raw sub-genomic RNA levels (HQ&LQ) in SHEF-CE04A show high relative amounts of this non-canonical sub-genomic RNA at 10639. (ORFs with sub-genomic RNA evidence shown)* ***F.*** *Non-canonical sub-genomic RNA at 5785 (SHEF-BFD90 shown for illustration) found in* 226 *samples.* ***G.*** *Count of HQ&LQ raw sub-genomic RNAs in SHEF-D02E5.*

Variants in Sub-Genomic RNA

An advantage of both having reads from both genomic and subgenomic RNA is the ability to examine how genomic variants are represented in sub-genomic RNAs. For variants found in our isolates by the ARTIC Network nanopolish(Simpson 2018) pipeline we interrogated the bases called (pysam(Gilman et al. 2019) pileup) at the variant position in genomic and sub-genomic RNA to determine if there were detectable differences (this tool is integrated into periscope). As we can only discern genomic and sub-genomic from a small subset of amplicons, the chance that a variant falls within these amplicons is low, but where the two do coincide, variants called in genomic RNA were supported in reads from sub-genomic RNA.

In a small subset of samples we have identified variants in the TRS sequence of some ORFs. Notably 6 samples have a variant at 27046C>T (ACGAAC to ACGAA**T**) in the TRS of ORF6, this variant is present in both the gRNA and sgRNA reads. 4 of these samples have low expression of ORF6 compared to the rest of the cohort (Figure 5E), although numbers are too low to compute statistical significance.

In addition we have one sample with a variant in the N ORF TRS (SHEF-C0F96, 28256C>T. CTAAACGAAC to **T**TAAACGAAC) which is found only in gRNA but not sgRNA reads. Interestingly, this sample has low expression of most ORFs (ORF N shown in Figure 5C) and has 233,127 reads which is around the median of the cohort. However, this mutation falls outside the core TRS and read counts for sub-genomic RNA are low so this result should be treated with caution. It is possible that this represents a sequencing error in the genomic RNA which is not found in the sub-genomic RNA due to the context around that position changing due to the inclusion of the leader sequence.

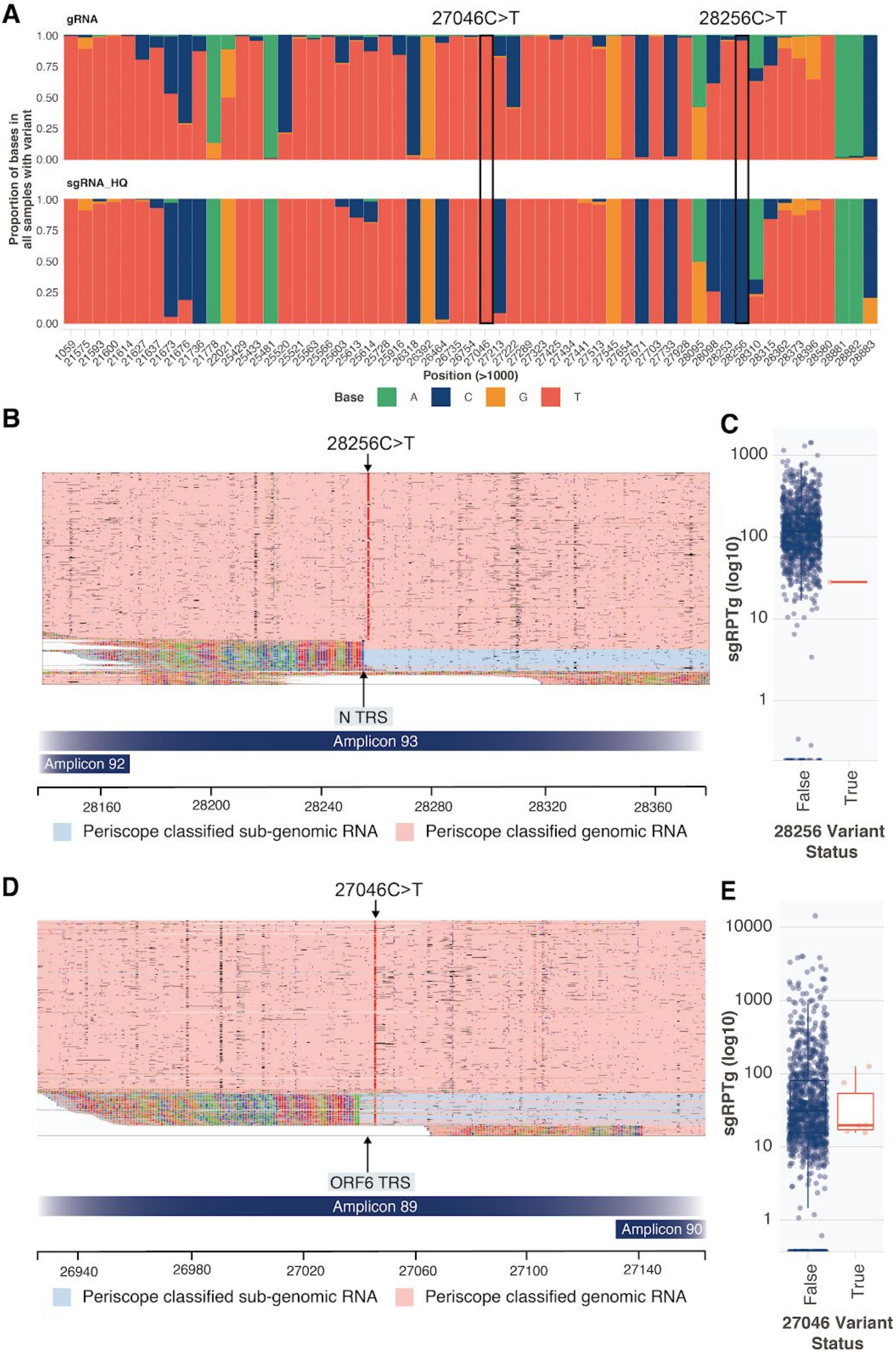

A
gRNA
27046C>T
28256C>T
sgRNA_HQ
Proportion of bases in all samples with variant
1.00
0.75
0.50
0.25
0.00
1059
21575
21593
21600
21614
21627
21637
21673
21676
21736
21778
22021
25429
25433
25481
25520
25521
25563
25566
25603
25613
25614
25728
25916
26318
26392
26464
26735
26754
27046
27213
27222
27289
27323
27425
27434
27441
27513
27545
27654
27671
27703
27733
27928
28095
28098
28253
28256
28310
28315
28362
28373
28396
28580
28881
28882
28883
Position (>1000)
Base
A
C
G
T
B
28256C>T
N TRS
Amplicon 93
Amplicon 92
28160
28200
28240
28280
28320
28360
Periscope classified sub-genomic RNA
Periscope classified genomic RNA
C
1000
100
10
1
sgRPTg (log10)
False
True
28256 Variant Status
D
27046C>T
ORF6 TRS
Amplicon 89
Amplicon 90
26940
26980
27020
27060
27100
27140
Periscope classified sub-genomic RNA
Periscope classified genomic RNA
E
10000
1000
100
10
1
sgRPTg (log10)
False
True
27046 Variant Status

Figure 5. Variants in Sub-Genomic RNA

A. *Base frequencies at each of the variant positions called by ARTIC in each sample, (multiple samples can be represented at one position), split by read class.* ***B.*** *SHEF-C0F96 has a 28256C>T variant, of high quality which sits in the ORF N TRS sequence. This variant is not present in sub-genomic reads and N expression is one of the lowest in the cohort (**C**).* ***C.*** *SHEF-C0C35 has 27046C>T variant of high quality which sits in the TRS sequence. This variant is present in both genomic and sub-genomic gRNA.* ***D.*** *ORF6 expression levels in these samples.*

Discussion

We have developed periscope, a tool that can be used on the tens of thousands of publically available sequencing datasets worldwide to detect sub-genomic RNA. Here we applied periscope to 1,155 SARS-CoV-2 sequences from Sheffield, UK. This was motivated by the detection of a novel sub-genomic RNA generated as a result of a triplet mutation found in a large number of worldwide SARS-CoV-2 isolates(Leary et al. 2020). Here, we expand that analysis showing that sub-genomic RNA analysis using periscope is reproducible, with strong correlations between sub-genomic RNA abundance levels between technical replicates, and has a limit of detection of at least 50,000 total reads.

We were able to detect reads supporting all annotated ORFs with the exception of ORF10 (Supplementary Table S1). We failed to find significant support for this ORF, mirroring previous findings(Kim et al.; Alexandersen et al. 2020), with only two reads in total for all 1155 samples in the dataset. Taken together this strongly suggests that ORF10 is not functional in SARS-CoV-2. The abundance of sub-genomic RNAs is in line with previously published reports of protein levels in SARS-CoV-2, with M, N and S showing the highest levels(Bouhaddou et al. 2020). In our data the median proportion of total sub-genomic RNA is 1.2% (Supplementary Figure S8), which is in agreement with reports, based on E gene, that subgenomic RNA represented 0.4% of total viral RNA(Wölfel et al. 2020).

Non-canonical sub-genomic RNAs are readily detected by periscope and we present examples where periscope was able to detect high abundances of specific non-canonical sub-genomic RNAs in a number of isolates which could indicate some functional significance. The non-canonical sub-genomic RNA at position 25744 in SHEF-C0118 is particularly interesting due to its high relative abundance compared to the canonical sub-genomic RNAs in the same sample. There does not appear to be a canonical TRS sequence in close proximity to the leader junction, but there exists a motif which has two mismatches to the canonical TRS; A**A**GAA**T**. An ATG downstream of the leader in these reads has an adequate Kozak sequence (AXXATGa) and would result in a N-terminal truncated 3a protein. The position of this novel sub-genomic RNA, alone, suggests that this could be a sub-genomic RNA that represents ORF 3b. Interestingly two forms of 3a protein in SARS-CoV have been noted in the literature(Huang et al. 2006). Alternatively, SARS-CoV contains a nested ORF within the 3a sub-genomic RNA, 3b, but the homolog of this protein is truncated early in SARS-CoV2, however, others note that a protein from this truncated 3b, of only 22 amino acids in length, could have an immune regulatory function(Konno et al. 2020). This non-canonical sub-genomic RNA could indicate production of this novel 3b protein in SARS-CoV-2 independent from the 3a sub-genomic RNA, a phenomenon that has been shown to occur in SARS-CoV(Hussain et al. 2005). We cannot explain, why, in this sample this non-canonical sub-genomic RNA is present in such high abundance. There are no genomic variants that contribute to a TRS sequence, for example. We also find evidence of highly recurrent non-canonical sub-genomic RNAs which have weaker evidence like those at 10639, which could represent an independent sub-genomic RNA for ORF 1b. Others, like those at 5785 have no apparent related ORF. These findings illustrate that, although much is unknown about the expression of non-canonical sub-genomic RNA, periscope could help define these novel transcripts in larger datasets.

We were able to integrate the sub-genomic RNAs for the presence of variants which were called in genomic RNA. In most cases, as expected, sub-genomic RNAs contain the same variants present in the genomic RNA. We found one case where, in the N sub-genomic RNA, a variant found in genomic RNA was not present. It is possible that this is due to sequencing errors. The surrounding bases for the variant in genomic vs sub-genomic differ, therefore base calling could be affected by this change in context.

It has been suggested that sub-genomic RNA abundance estimates from amplicon based sequencing data are largely a function of the quality of the RNA in the initial sample, as a marker for this, the authors used average read length(Alexandersen et al. 2020). The advantage of the ARTIC Network protocol over the AmpliSeq IonTorrent protocol used by the aforementioned study for SARS-CoV-2 sequencing is that the ARTIC Network protocol has a short, consistent amplicon length (mean is 389 and standard deviation is 11.2). The sub-genomic RNA amplicons described here have a range of between 200 and 400bp (Supplementary Figure S4), with some, as shown, having support from longer reads. The assay is designed inherently, to deal with samples with degraded RNA. Since sub-genomic RNA reads are a product of these amplicons we do not believe degradation plays a significant role in the determination of abundance levels in our dataset. Furthermore, using E gene ct as a surrogate for RNA quality we find no correlation with the total amount of sub-genomic RNA detected (Spearman Rank Correlation, rho=0.06977053 , Supplementary

Figure S6). Furthermore, coverage is also not correlated with sgRNA amount (Spearman Rank Correlation, rho=-0.01188594, Supplementary Figure S7).

The COVID-19 Genomics Consortium (COG-UK)(2020) in the UK, alone, has 10165 ARTIC nanopore sequences (Correct 12th June 2020) while internationally GISAID contains thousands more similar datasets (8775 with “nanopore” in the metadata and 3660 list “ARTIC”, 14th June 2020). The application of periscope could therefore provide significant insights into the sub-genomic RNA architecture of SARS-CoV-2 at an unprecedented scale. Furthermore, periscope can be provided new primer/amplicon locations and therefore would be suitable for any PCR based genomic analysis protocol and could be applied to other viral sequencing data where discontinuous transcription was the method of gene expression.

Periscope offers an opportunity to further understand regulation of the SARS-Cov-2 genome by identifying and quantifying sub-genomic RNA in large sequencing datasets. By taking advantage of the vast amount of ARTIC Network Nanopore sequencing datasets that have been generated during the unprecedented public health crisis, we can not only add value, but could potentially uncover biological consequences to variants which are present in the SARS-CoV-2 genome.

Methods

SARS-CoV-2 Isolate Collection and Processing

1155 samples from 1155 SARS-CoV-2 positive individuals were obtained from either throat or combined nose/throat swabs. Nucleic acids were extracted from 200µl of each sample using MagnaPure96 extraction platform (Roche Diagnostics Ltd, Burgess Hill, UK). SARS-CoV-2 RNA was detected using primers and probes targeting the E gene and the RdRp genes of SARS-CoV-2 and the human gene RNASEP, to allow normalisation, for routine clinical diagnostic purposes, with thermocycling and fluorescence detection on ABI Thermal Cycler (Applied Biosystems, Foster City, United States) using previously described primer and probe sets(Corman et al. 2020).

SARS-CoV-2 Isolate Amplification and Sequencing

Nucleic acids from positive cases underwent long-read whole genome sequencing (Oxford Nanopore Technologies (ONT), Oxford, UK) using the ARTIC Network protocol (accessed the 19th of April, https://artic.network/ncov-2019, https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w). In most cases 23 isolates and one negative control were barcoded per 9.4.1D nanopore flow cell. Following basecalling, data were demultiplexed using ONT Guppy (--require-both-ends). Reads were filtered based on quality and length (400 to 700bp), then mapped to the Wuhan reference genome (MN908947.3) and primer sites trimmed. Reads were then downsampled to 200x coverage in each direction. Variants were called using nanopolish(Simpson 2018).

Sub-genomic RNA Detection

Periscope consists of a Python based snakemake(Köster and Rahmann 2012) workflow which runs a python package that processes and classifies reads based on their configuration (Figure 1C).

Pre-Processing

Pass reads for single isolates are concatenated and aligned to MN908947.3 with minimap2(Li 2018) (-ax map-ont -k 15). It should be noted that adapters or primers are not trimmed. Bam files are sorted and indexed with samtools(Li et al. 2009).

Periscope - Read Filtering

Reads from this bam file are then processed with pysam(Gilman et al. 2019). If a read is unmapped or represents a supplementary alignment then it is discarded. Each read is then assigned an amplicon using the “find_primer” method of the ARTIC field bioinformatics package. We search for the leader sequence (5’-`AACCAACTTTCGATCTCTTGTAGATCTGTTCT`-3’) with biopython(Cock et al. 2009) local pairwise alignment (localms) with the following settings, match +2, mismatch -2, gap -10 and extension -0.1 with score_only set to true to speed up computation. The read is then assigned an ORF using a pybedtools(Dale et al. 2011) and a bed file consisting of all known ORFs +/-10 of the predicted leader/genome transition.

Periscope - Read Classification

We then classify reads (Figure 1D); if the alignment score is > 50 and the read is at a known ORF then it is classified as a “High Quality” sub-genomic RNA. If the read starts at a primer site then it is classified as genomic RNA, if not then it is classified as a “High Quality” non-canonical sub-genomic RNA supporting read. If the alignment score is > 30 but <= 50 and the read is at a known ORF then it is classified as a “Low Quality” sub-genomic RNA. If the read is within a primer site it is labelled as a genomic RNA, if not then it is a “Low Quality” sub-genomic RNA. Finally any reads with a score of <= 30 are dealt with. If they are at a known ORF then they are classified as a “Low Low Quality” sub-genomic RNA otherwise they are labelled as genomic RNA. The following tags are added to the reads for manual review of the periscope calls; XS: Alignment score, XA: Amplicon, AC: Read class, and XO: The read ORF. Reads are binned into qualities (HQ, LQ, LLQ etc) because we noticed that some sub-genomic RNAs were not classified as such due to a lower match to the leader. On manual review they are bona fide sub-genomic RNA. This quality rating negates the need to alter alignment score cut-offs continually to find the best balance between sensitivity and specificity. Restricting to HQ data means that sensitivity is reduced but specificity is increased, including LQ calls will decrease specificity but increase sensitivity.

Periscope - Sub-genomic RNA Normalisation

Once reads have been classified, the counts are summarised and normalised. Two normalisation schemes are employed:

1. **Normalisation to Total Mapped Reads**

   Total mapped reads per sample are calculated using pysam idxstats and used to normalise both genomic, sub-genomic and non-canonical sub-genomic RNA reads. Reads per hundred thousand total mapped reads are calculated per quality group.

2. **Normalisation to Genomic Reads from the Corresponding Amplicon**

   Counts of sub-genomic, non-canonical sub-genomic, and genomic are recorded on a per amplicon basis and normalisation occurs within the same amplicon to the total genomic read count/1000. If multiple amplicons contribute to the count of sub-genomic or non-canonical subgenomic then the normalised values are summed.

Periscope outputs several useful files which are described in more detail in the Supplementary Material, briefly these are; periscope’s processed bam file with associated tags, a per amplicons counts file, and a summarised counts file for both canonical and non-canonical ORFs.

Sub-genomic Variant Analysis

A python script is provided “variant_expression.py” that takes the periscope bam file and a VCF file of variants (usually from the ARTIC analysis pipeline) and for each position in the VCF (pyvcf(Casbon)) file extracts the counts of each base in each class of read (i.e. genomic, sub-genomic and non-canonical sub-genomic) and outputs these counts as a table. This tool also provides a useful plot (Supplementary Figure S5) of the base counts at each position for each class.

Analysis and Figure Generation

Further analysis was completed in R 3.5.2(Schulte et al. 2012) using Rstudio 1.1.442(Racine 2012), in general data was processed using dplyr (v0.8.3), figures were generated using ggplot2 (v3.3.1), both part of the tidyverse(Wickham et al. 2019) family of packages (v1.2.1). Plots themed with the ftplottools package (v0.1.5). GGally (v2.0.0) ggpairs was used for the matrix plots for downsampling and repeats. Where multiple hypothesis tests were carried out, multiple testing correction was carried out using Bonferroni-Holm. Reads were visualised in IGV(Robinson et al. 2011) and annotated with Adobe Illustrator.

Principal Component Analysis (PCA)

PCA was carried out to determine if any of the experimental variables were responsible for the differences in expression values between samples. Reads from ORF1a and ORF10 amplicons were removed from the analysis and expression values were normalised within each ORF:

$$\frac{x-min(x)}{max(x)-min(x)}$$

and then row means subtracted. The "PCA" function of the R package FactoMineR(Pagès 2014), (v1.41) was then used to carry out the PCA without further scaling and all other settings as default. The resulting PCA was plotted using the "fviz_pca_ind" function. Plots were coloured according to the variable in question.

Heatmap

Heatmap was constructed using the R package "heatmap3" (v1.1.7) and uses the same expression values used for the PCA analysis described above.

Periscope Requirements

Periscope is a snakemake(Köster and Rahmann 2012) workflow with a package written in python to implement read filtering and classification and is provided with a conda environment definition. It has been tested on a Dell XPS, core i9, 32gb ram, 1Tb SSD running ubuntu 18.04 and was able to process 10,000 reads per minute. To install periscope requires conda. To run periscope you will need the path to your raw fastq files from your ARTIC Network Nanopore sequencing run (unfiltered) and other variables defined in the Supplementary Material and on the github README.

Data Access

All raw and periscope processed sequencing data will be provided in due course on the European Nucleotide Archive under study ID: EGAS00001004520.

Periscope is freely available under GNU General Public License v3.0 from the Sheffield Bioinformatics Core github account (https://github.com/sheffield-bioinformatics-core/periscope)

Results of our full analysis can also be found as supplementary material and also on github (https://github.com/sheffield-bioinformatics-core/periscope-publication)

Abbreviations

TRS	Transcription regulation sequence (ACGAAC)
ORF	Open reading frame
sgRNA	sub-genomic RNA
gRNA	genomic RNA
sgRPHT	sub-genomic RNA reads per 100,000 mapped reads
sgRPTg	sub-genomic RNA reads per 1,000 genomic RNA reads
gRPHT	genomic RNA reads per 100,000 mapped reads

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).

Acknowledgements

We would like to thank the Sheffield Bioinformatics Core for useful discussions around the normalisation techniques. We thank all partners of and contributors to the COG-UK consortium, who are listed at https://www.cogconsortium.uk/about/.

Author Contributions

MDP developed periscope, analysed and interpreted data and was a major contributor to manuscript preparation. BL analysed and interpreted data and contributed to manuscript preparation. DW interpreted data and contributed to manuscript preparation. SL,SG, SM, AC originally conceived the method for sub-genomic RNA detection and contributed to manuscript preparation. MW, LRG, PP, DG, RT, RB, LC, AA, AK, KJ and NS collated, processed and sequenced samples, and reviewed the manuscript. JH is a developer of our LIMS system to ensure efficient sample processing, and reviewed the manuscript. MR, CE and DP collected and organised clinical samples and metadata and reviewed the manuscript. TdS conceived the study and contributed to manuscript preparation and oversight. All authors read and approved the final manuscript.

Disclosure Declaration

All authors declare there are no conflicts of interest.

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 were 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).

References

Alexandersen S, Chamings A, Bhatta TR. 2020. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.06.01.20119750v1.abstract.

Bouhaddou M, Memon D, Meyer B, White KM, Rezelj VV, Marrero MC, Polacco BJ, Melnyk JE, Ulferts S, Kaake RM, et al. 2020. The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell*. http://www.sciencedirect.com/science/article/pii/S0092867420308114.

Casbon J. PyVCF—A Variant Call Format Parser for Python 2012.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, Bleicker T, Brünink S, Schneider J, Schmidt ML, et al. 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* **25**. http://dx.doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**: 3423–3424.

Gilman P, Janzou S, Guittet D, Freeman J, DiOrio N, Blair N, Boyd M, Neises T, Wagner M, Others. 2019. *PySAM (Python Wrapper for System Advisor Model" SAM")*. National Renewable Energy Lab.(NREL), Golden, CO (United States) https://www.osti.gov/biblio/1559931.

Huang C, Narayanan K, Ito N, Peters CJ, Makino S. 2006. Severe acute respiratory syndrome coronavirus 3a protein is released in membranous structures from 3a protein-expressing cells and infected cells. *J Virol* **80**: 210–217.

Hussain S, Pan J 'an, Chen Y, Yang Y, Xu J, Peng Y, Wu Y, Li Z, Zhu Y, Tien P, et al. 2005. Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. *J Virol* **79**: 5288–5295.

Irigoyen N, Firth AE, Jones JD, Chung BY-W, Siddell SG, Brierley I. 2016. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. *PLoS Pathog* **12**: e1005473.

Kim D, Lee J-Y, Yang J-S, Kim JW, Narry Kim V, Chang H. The architecture of SARS-CoV-2 transcriptome. http://dx.doi.org/10.1101/2020.03.12.988865.

Konno Y, Kimura I, Uriu K, Fukushi M, Irie T. 2020. SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is further increased by a naturally occurring elongation variant. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.05.11.088179v1.abstract.

Köster J, Rahmann S. 2012. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**: 2520–2522.

Leary S, Gaudieri S, Chopra A, Pakala S, Alves E, John M, Das S, Mallal S, Phillips E. 2020. Three adjacent nucleotide changes spanning two residues in SARS-CoV-2 nucleoprotein: possible homologous recombination from the transcription-regulating sequence. *bioRxiv* 2020.04.10.029454. https://www.biorxiv.org/content/10.1101/2020.04.10.029454v1.abstract (Accessed June 11, 2020).

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H. 2012. seqtk Toolkit for processing sequences in FASTA/Q formats.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Pagès J. 2014. *Multiple Factor Analysis by Example Using R*. CRC Press.

Patel RK, Burnham AJ, Gebhart NN, Sokoloski KJ, Hardy RW. 2013. Role for subgenomic mRNA in host translation inhibition during Sindbis virus infection of mammalian cells. *Virology* **441**: 171–181.

Racine JS. 2012. RStudio: a platform-independent IDE for R and Sweave. *J Appl Econometrics* **27**: 167–172.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Schulte E, Davison D, Dye T, Dominik C. 2012. A Multi-Language Computing Environment for Literate Programming and Reproducible Research. *Journal of Statistical Software* **46**. http://dx.doi.org/10.18637/jss.v046.i03.

Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nat Rev Microbiol* **9**: 617–626.

Simpson J. 2018. Nanopolish: Signal-level algorithms for MinION data. *Github Available at: https://github com/jts/nanopolish [Accessed January 10, 2019]*.

Sola I, Almazán F, Zúñiga S, Enjuanes L. 2015. Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu Rev Virol* **2**: 265–288.

Stern DF, Kennedy SI. 1980. Coronavirus multiplication strategy. I. Identification and characterization of virus-specified RNA. *J Virol* **34**: 665–674.

Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L. 2020. Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.03.05.976167v2.abstract.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* **4**: 1686.

Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, Niemeyer D, Jones TC, Vollmar P, Rothe C, et al. 2020. Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**: 465–469.

Wu H-Y, Brian DA. 2010. Subgenomic messenger RNA amplification in coronaviruses. *Proc Natl Acad Sci U S A* **107**: 12257–12262.

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

2020. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe*. http://www.sciencedirect.com/science/article/pii/S2666524720300549.

Artic Network. https://artic.network/ncov-2019 (Accessed June 11, 2020a).

fieldbioinformatics. Github https://github.com/artic-network/fieldbioinformatics (Accessed June 15, 2020b).

Supplementary Material

Supplementary Files

supplementary_file_1_cannonical_orf_counts.csv

All canonical ORF counts for all 1155 samples.

supplementary_file_2_novel_orf_counts.csv

All non-canonical ORF counts for all 1155 samples.

supplementary_file_3_downsampling.csv

All canonical ORF counts for all samples used in the downsampling experiment.

supplementary_file_4_repeats.csv

All canonical ORF counts for all samples used in the replicates experiment.

Periscope README

Below we have provided instructions for the installation, execution and interpretation of periscope.

Requirements

periscope runs on MacOS, unix and unix subsystem for windows 10.

You will need:

- conda
- Your raw fastq files from the ARTIC protocol

- Periscope installation

In our hands periscope takes around 1 minute per 10,000 reads on a single core on a Dell XPS core i9 with 32Gb ram and 1Tb SSD.

Installation

```
git clone https://github.com/sheffield-bioinformatics-core/periscope.git
&& cd periscope
conda env create -f environment.yml
conda activate periscope
python setup.py install
```

Execution

```
conda activate periscope
periscope \
    --fastq-dir <PATH_TO_DEMUXED_FASTQ> \
    --output-prefix <PATH_TO_OUTPUT> \
    --sample <SAMPLE_NAME> \
    --resources <PATH_TO_PERISCOPE_RESOURCES_FOLDER> \
    --score_cutoff <ALIGNMENT_CUTOFF_FOR_sgRNA> \
    --threads <THREADS_FOR_MAPPING>
```

Output Files

Filename	**Description**
`<OUTPUT_PREFIX>.fastq`	A merge of all files in the fastq directory specified as input.
`<OUTPUT_PREFIX>_periscope_counts.csv`	The counts of genomic, sub-genomic and normalisation values for *known* ORFs
`<OUTPUT_PREFIX>_periscope_amplicons.csv`	The amplicon by amplicon counts, this file is useful to see where the counts come from. Multiple amplicons may be represented more than once where they may have

	contributed to more than one ORF.
`<OUTPUT_PREFIX>_periscope_novel_counts.csv`	The counts of genomic, sub-genomic and normalisation values for *non-canonical* ORFs

`<OUTPUT_PREFIX>.bam &`
`<OUTPUT_PREFIX>.bam.bai`
	minmap2 mapped reads and index with no adjustments made.

`<OUTPUT_PREFIX>_periscope.bam &`
`<OUTPUT_PREFIX>_periscope.bam.bai`
	This is the original input bam file and index created by periscope with the reads specified in the fastq-dir. This file, however, has tags which represent the results of periscope:
	● XS is the alignment score
	● XA is the amplicon number
	● XC is the assigned class (gDNA or sgDNA)
	These are useful for manual review in IGV or similar genome viewer. You can sort or colour reads by these tags to aid in manual review and figure creation.

Examining Base Frequencies of Called Variants in periscope

This script will take the pass vcf from the ARTIC Network pipeline and examine the periscope bam file for the bases present at that position. It will split the counts by read class and output a plot showing contribution at each base at each site in the VCF.

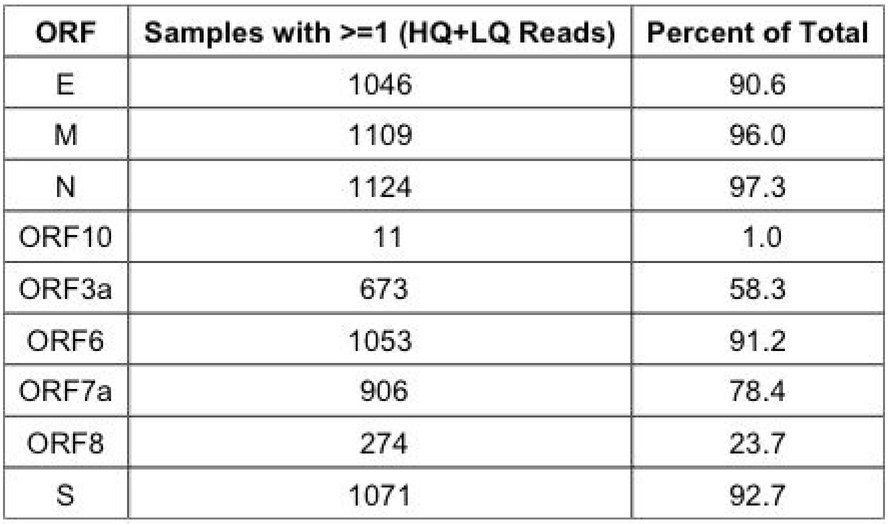
```
conda activate periscope

gunzip <ARTIC_NETWORK_VCF>.pass.vcf.gz

<PATH_TO_PERISCOPE>/periscope/periscope/scripts/variant_expression.py \
    --periscope-bam <PATH_TO_PERISCOPE_OUTPUT_BAM> \
    --vcf <ARTIC_NETWORK_VCF>.pass.vcf \
    --sample <SAMPLE_NAME> \
    --output-prefix <OUPUT_PREFIX>
```

Filename	**Description**
`<OUTPUT_PREFIX>_base_counts.csv`	Counts of each base at each position
`<OUTPUT_PREFIX>_base_counts.png`	Plot of each position and base composition

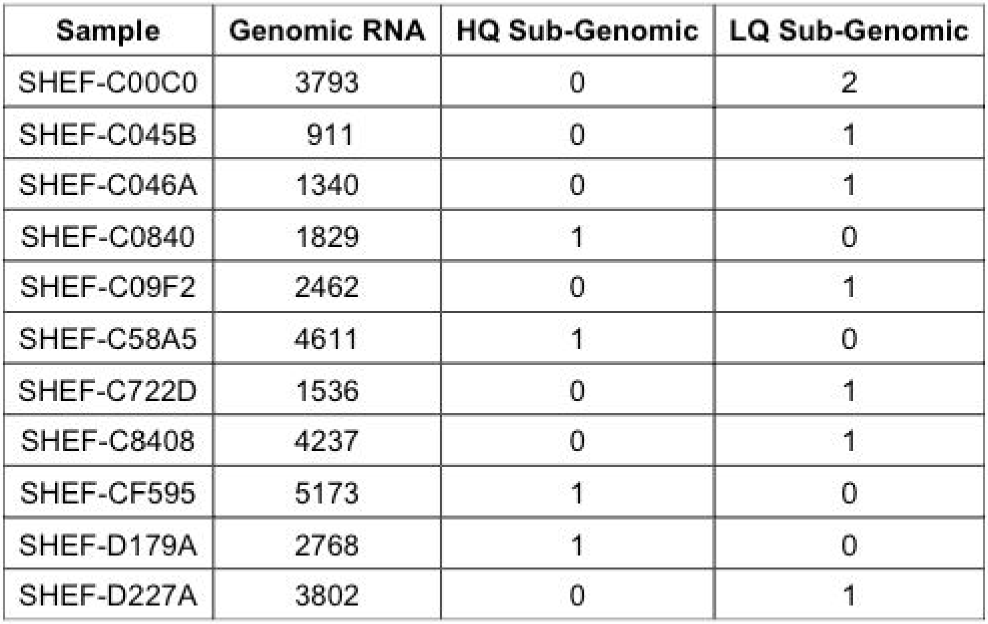
Supplementary Tables



ORF	Samples with >=1 (HQ+LQ Reads)	Percent of Total
E	1046	90.6
M	1109	96.0
N	1124	97.3
ORF10	11	1.0
ORF3a	673	58.3
ORF6	1053	91.2
ORF7a	906	78.4
ORF8	274	23.7
S	1071	92.7

Supplementary Table S1 - sub-genomic RNAs detected for each canonical ORF

Raw counts of the number of sub-genomic RNAs found across all 1155 samples of the cohort with 1 or more HQ or LQ read.



Sample	Genomic RNA	HQ Sub-Genomic	LQ Sub-Genomic
SHEF-C00C0	3793	0	2
SHEF-C045B	911	0	1
SHEF-C046A	1340	0	1
SHEF-C0840	1829	1	0
SHEF-C09F2	2462	0	1
SHEF-C58A5	4611	1	0
SHEF-C722D	1536	0	1
SHEF-C8408	4237	0	1
SHEF-CF595	5173	1	0
SHEF-D179A	2768	1	0
SHEF-D227A	3802	0	1

Supplementary Table S2 - Samples with predicted sub-genomic RNA for ORF10

Samples with any HQ or LQ evidence of ORF10 sub-genomic RNA (amplicons 97 and 98).

Twelve reads in total across the whole cohort putatively support a sub-genomic RNA for ORF10 .

Read Start Position	Contributing Amplicon	Total Genomic Read Count	Total HQ Sub-Genomic Read Count	Total LQ Sub-Genomic Read Count	Total Sub-Genomic Read Count
25744	84,85	6019	120	33	153
25745	85	1389	4	5	9
25746	85	1389	1	2	3
25748	85	1389	2	1	3
25749	85	1389	1	0	1
25754	85	1389	2	0	2
25755	85	1389	1	0	1
25732	85	1389	0	1	1
25735	85	1389	0	1	1
25742	85	1389	0	1	1
25753	85	1389	0	1	1
25766	85	1389	0	1	1

Supplementary Table S3 - Non-canonical sub-genomic RNA at position 25744 in sample SHEF-C0118 has strong support

ORF	Contributing Amplicon	Total Genomic Read Count	Total HQ Sub-Genomic Read Count	Total LQ Sub-Genomic Read Count	Total Sub-Genomic Read Count
S	71,72	4939	44	31	75
ORF3a	83,84	8147	3	2	5
E	86,87,88	11183	12	7	19
M	87,88	6590	192	90	282
ORF6	89	564	60	24	84
ORF7a	90,91	5580	0	16	16
N	93,94	8928	500	251	751
Non-Canonical sgRNA	**84,85**	**6019**	**131**	**46**	**177**

Supplementary Table S4 - Canonical sub-genomic RNA in SHEF-C0118

Showing only those ORF sub-genomic RNAs with supporting Reads

Read Start Position	Contributing Amplicon	Total Genomic Read Count	Total HQ Sub-Genomic Read Count	Total LQ Sub-Genomic Read Count	Total Sub-Genomic Read Count
10639	35,36	4698	103	15	118
10640	35	3583	0	1	1
10641	35	3583	2	3	5
10642	35,36	4698	1	4	5
10643	36	1115	1	0	1
10644	35	3583	1	0	1
10645	35	3583	2	2	4
10647	35	3583	0	1	1

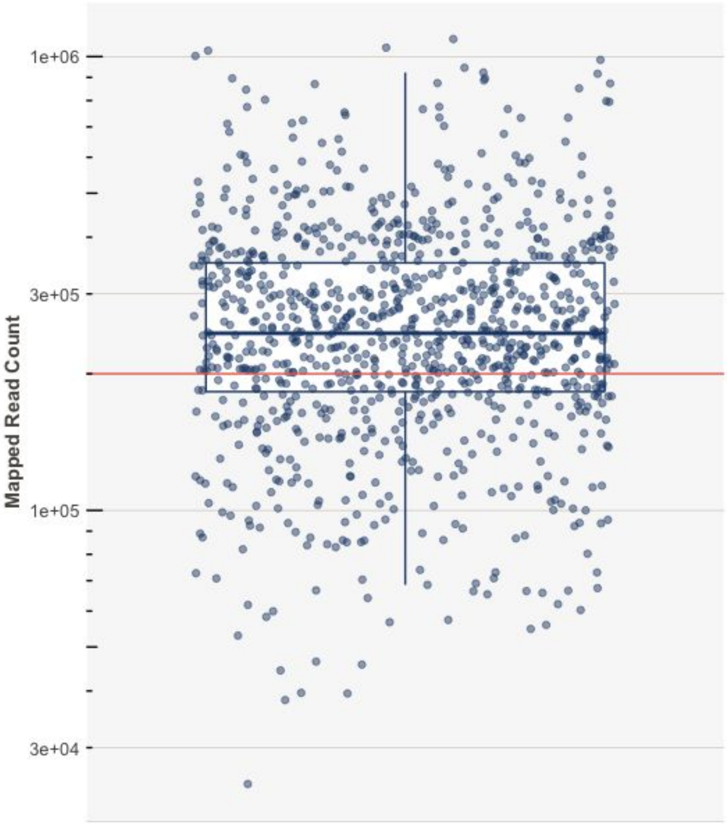
Supplementary Table S5 - Non-canonical sub-genomic RNA at 10639 in SHEF-CE04A has strong support

ORF	Contributing Amplicon	Total Genomic Read Count	Total HQ Sub-Genomic Read Count	Total LQ Sub-Genomic Read Count	Total Sub-Genomic Read Count
S	71,72	4945	64	16	80
ORF3a	84,85	6990	8	1	9
M	87	4456	176	43	219
ORF6	89	1527	0	0	0
N	93	8083	116	38	154
Non-Canonical sgRNA	**35,36**	**4698**	**110**	**26**	**136**

Supplementary Table S6 - Canonical sub-genomic RNA in SHEF-CE04A

Showing only those ORF sub-genomic RNAs with supporting Reads

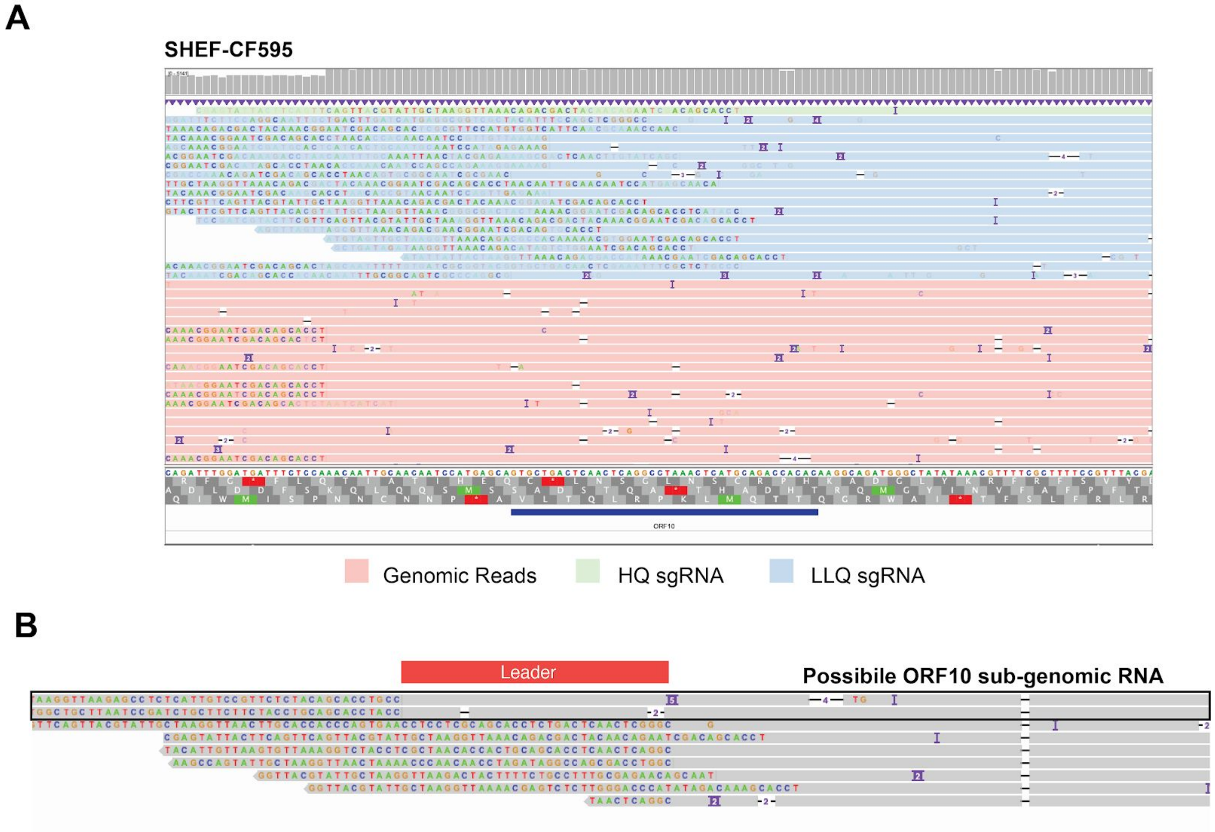
Supplementary Figures



minimum	q1	median	mean	q3	maximum
9105	183117	258210	312151.580170411	378579	3260686

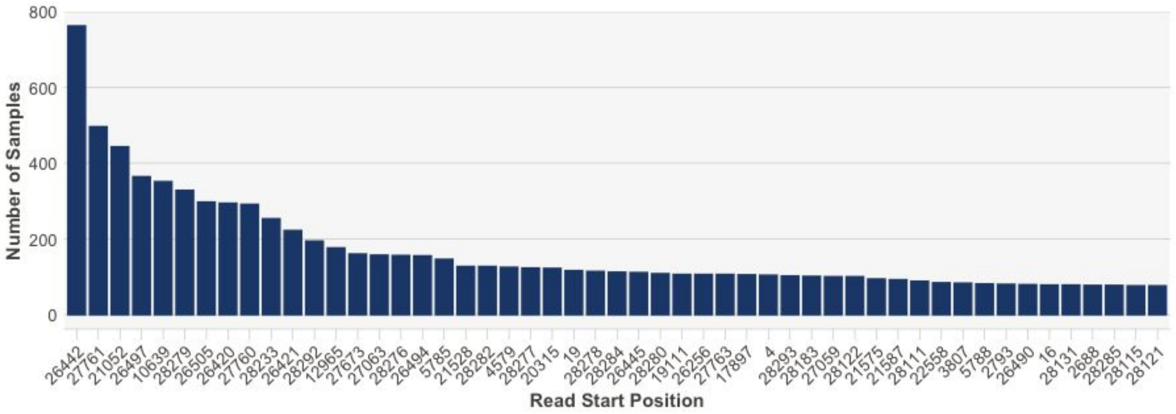
Supplementary Figure S1 - Mapped read counts in 1155 SARS-CoV-2 isolates from Sheffield

Reads were mapped with minimap2 to MN908947.3, sorted with samtools and mapped reads counted with samtools flagstat. Coral horizontal line represents 200,000 mapped reads.



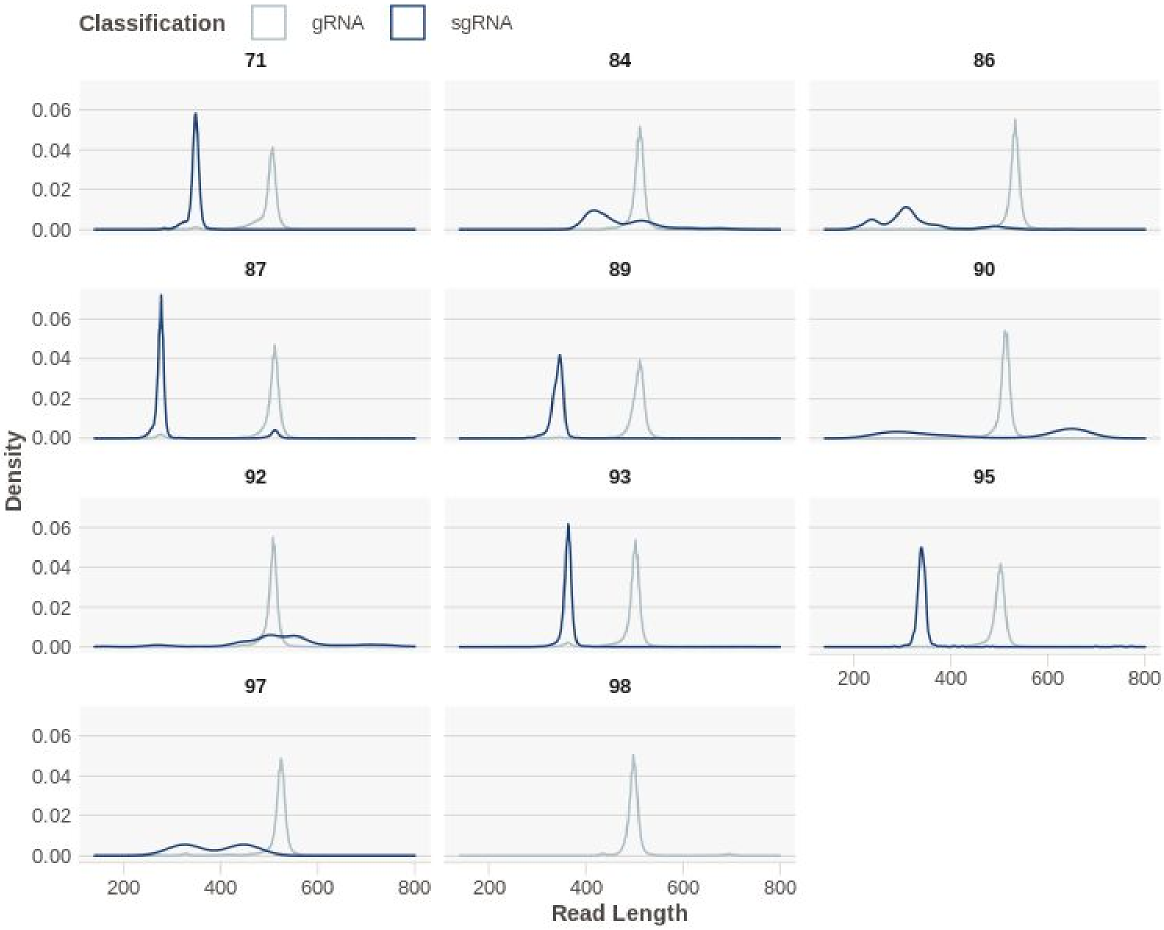
Supplementary Figure S2 - Manual review of periscope bam files for ORF10

A. SHEF-CF595 has 1 read classified as HQ sub-genomic RNA at the predicted ORF10 leader junction (light green). It is clear from this IGV screenshot that this read does not contain a valid leader sequence. **B.** All HQ and LQ sub-genomic reads, 12 in total, aligned with minimap2 to a reference consisting of ORF10 and leader sequence, 3 reads failed to map. Two reads could be bona fide ORF10 sub-genomic RNAs (Highlighted in the black box with a good match to the leader) from samples SHEF-C0840 and SHEF-C58A5.



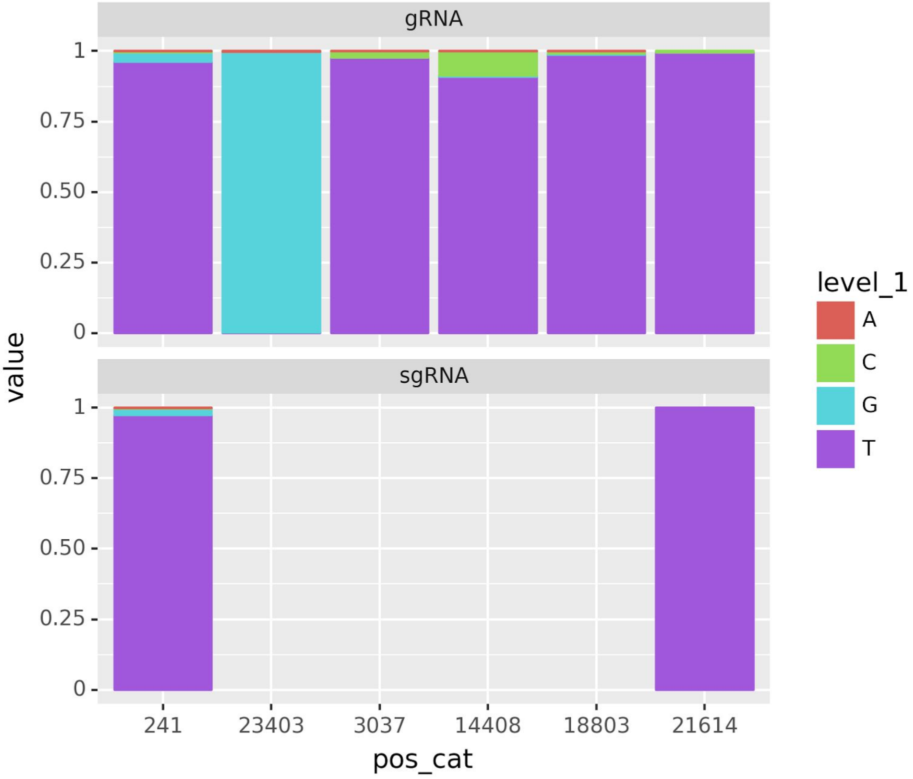
Supplementary Figure S3 - Number of samples with at the most frequently represented non-canonical sub-genomic RNAs

The number of samples each non-canonical sub-genomic RNA was found in. This is an exact position match, and includes sgRNAs that could be just outside the +/-10 of the leader junction. Sites with support in > 100 samples shown.



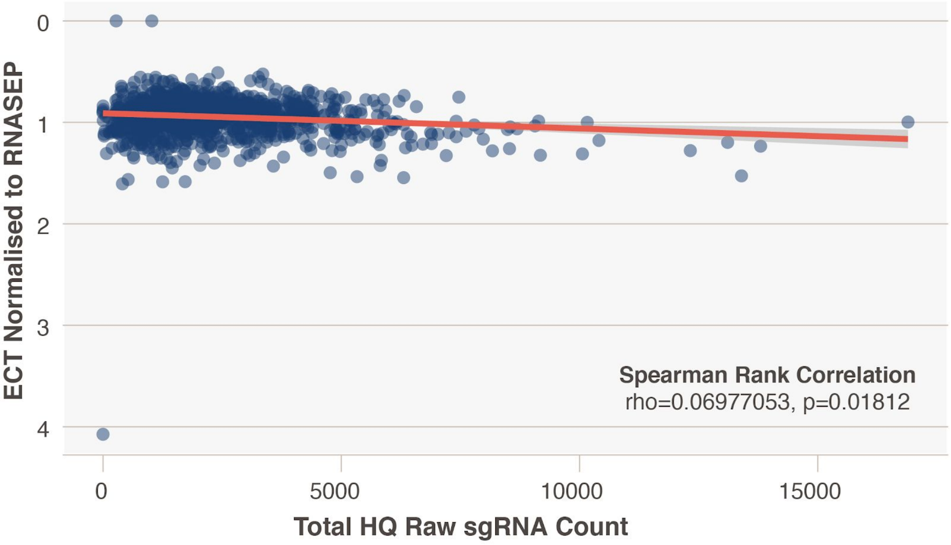
Supplementary Figure S4 - Size of reads classed as genomic and sub-genomic

At each of the amplicons responsible for the production of sub-genomic RNA supporting reads we examined the size of the two classes of treads. Genomic RNA is between 400 and 600bp, and in most cases sub-genomic RNA is shorter at around 200-300bp. For some amplicons, i.e. amplicon 90, reads which span two amplicons contribute to sgRNA calls and therefore we see a larger read length in this case.



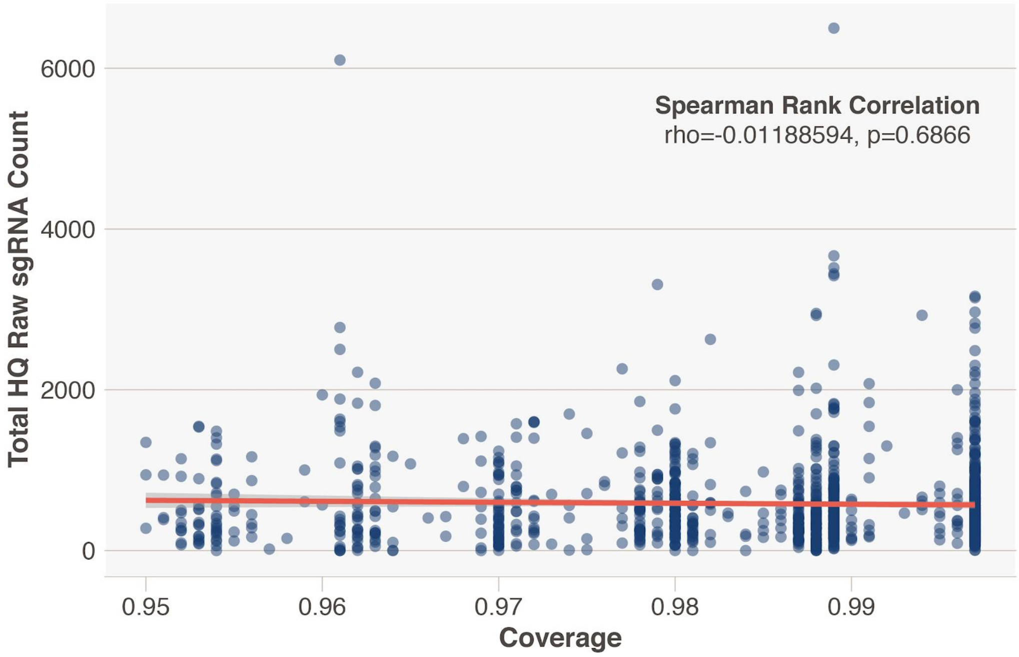
Supplementary Figure S5 - Example of periscope output for variant analysis

problematic variants as it has now been shown that this variant has a functional effect on sub-genomic RNA production.



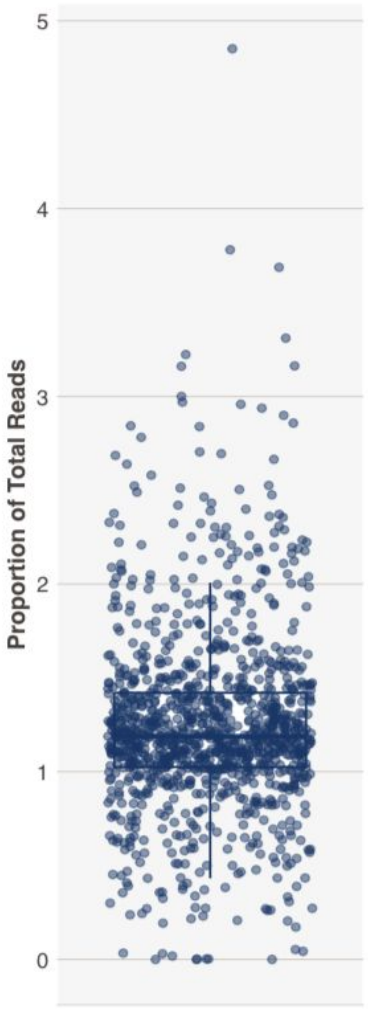
Supplementary Figure S6 - Normalised Ect is not correlated with the raw amount of sub-genomic RNA detected

Total raw HQ sub-genomic RNA counts are not correlated with normalised Ect (Ect/RNASEP). Y-axis reversed for ease of understanding.



Supplementary Figure S7 - Consensus coverage not correlated with the raw amount of sub-genomic RNA detected

Total raw HQ sub-genomic RNA counts are not correlated with consensus genome coverage.



Supplementary Figure S8 - sgRNA as a Proportion of the Total Reads

Total sub-genomic (HQ+LQ) reads as a proportion of the total reads. Median of 1.188%.