Chan Lab SARS-CoV-2 Probe Analysis

Analysis date: April 20, 2020
Report date: April 24, 2020

Summary

We aligned Chan lab probes to SARS-CoV-2 genomes from two complementary databases: NCBI (fewer genome sequences, higher quality) with ~800 genomes and GISAID (more sequences, poorer quality) with >7000 genomes. Our analysis suggests that the RdRp forward and reverse primers should be redesigned because it doesn't match the viral genome at multiple positions. All other probes and primers appear to correctly target stable genome regions. ORF1ab probes lie in a region that may be difficult to sequence, as 3.2% of GISAID genomes have systematic blocks of N's in this region. The uncertainty in this sequence is likely caused by technical DNA sequencing factors and represents poor data quality from the GISAID database. However, it does represent some uncertainty, which should be considered when interpreting ORF1ab probe results.

Background

The Chan lab aims to diagnose human SARS-CoV-2 infection using probes that simultaneously target multiple regions of the viral genome. Several DNA probes were designed to target each of three genome regions (ORF1ab, RdRp and E genes) of the human SARS-CoV-2 genome sequence (Fig 1). This analysis aims to measure the match fidelity of each probe to its target sequence and identify any mutations in the targeted regions found in available human SARS-CoV-2 genome sequences.

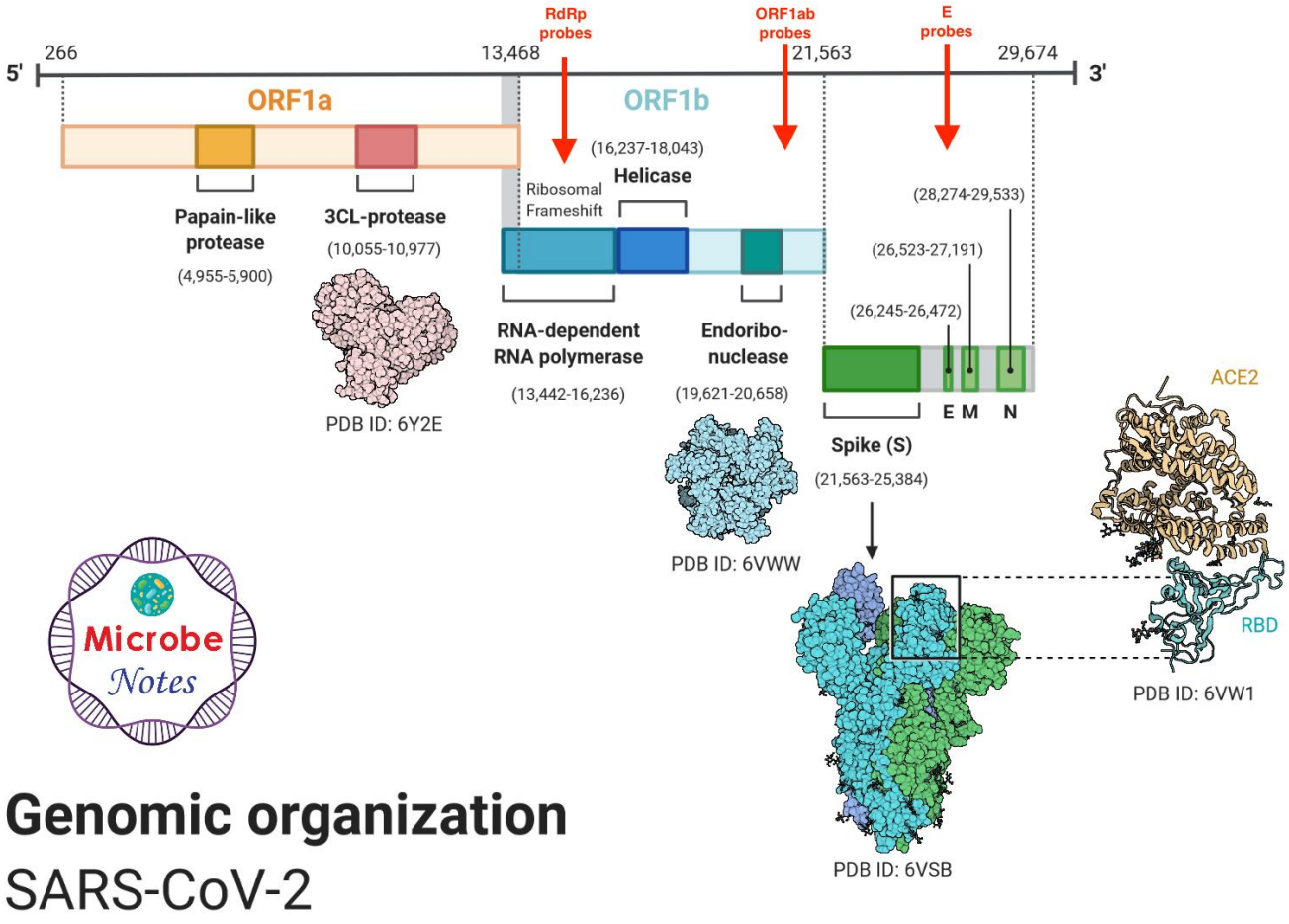


Figure 1: Schema of human SARS-CoV-2 genome (~30kb). Probes target three regions, ORF1ab, RdRp and E genes. Red arrows indicate approximate target positions.

Methods

Probe sequences

Several probes targeting each of ORF1ab, RdRp and E genes of the SARS-CoV-2 genome were provided by the Chan lab (Table 1).

Region	Forward Primer	Capture Probe
E gene	CGTTAATAGTTAATAGCGTACTTCTTTTTC	TACTTCTTTTTCTTGCTTTCGTGGTATTCT
RdRp	ATTAAGTGAGATGGTCATGTGTGGCGGCTCA	GTGGCGGTTCACTATATGTTAAACCAGGTG
ORF1ab	AGGTTTCAAACTTTACTTGCTTTACATAGA	TACTTGCTTTACATAGAAGTTATTTGACTC
	Reporter Probe	**Reverse Primer**
E gene	CTAGTTACACTAGCCATCCTTACTGCGCTT	ATATTGCAGCAGTACGCACACAATCGAAGC
RdRp	GAACCTCATCAGGAGATGCCACAACTGCTT	CAAATGTTAAAGACACTATTAGCATAAGCAG
ORF1ab	GGACAGCTGGTGCTGCAGCTTATTATGTGG	TCCTAGGTTGAAGATAACCCACATAATAAG
	Reporter Probe #2	
E gene	AGCCATCCTTACTGCGCTTCGATTGTGTGC	

Region	Reverse Primer (reverse complemented)
E gene	GCTTCGATTGTGTGCGTACTGCTGCAATAT
RdRp	CTGCTTATGCTAATAGTGTCTTTAACATTTG
ORF1ab	CTTATTATGTGGGTTATCTTCAACCTAGGA

Table 1: Probe sequences provided by the Chan lab. Reverse primers were reverse complemented to align along with other probes.

SARS-CoV-2 Genome Processing

Genome sequences were obtained from two global databases: GISAID and NCBI.

GISAID (https://www.gisaid.org/) is a database for influenza virus sequences and now receives public submissions of SARS-CoV-2 genomes. 7,716 sequences were downloaded from GISAID's EpiCov gateway on April 20 2020 in FASTA format. Sequences were selected using the following criteria: complete, high coverage, excluding low coverage and human. The database indicates that some samples are derived from cell culture, however the metadata was not publicly available, GISAID did not reply to emails and thus it was not possible to filter out these sequences, Fig 2.

NCBI is an established nucleotide repository and serves as a second source of global SARS-CoV-2 genomes. 825 sequences (selection criteria, nucleotide completeness: 'complete') were downloaded from the NCBI Virus Hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi) on 20 April 2020. The sequences include isolates from multiple sources (Fig 3).

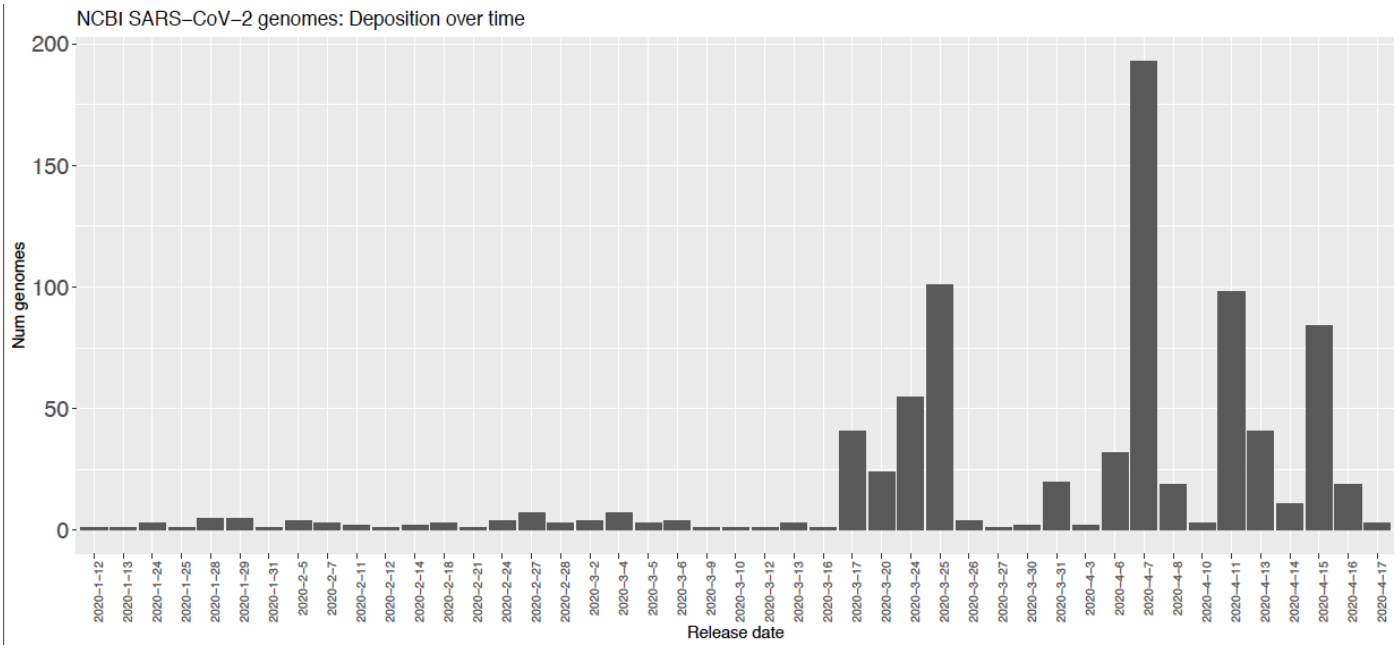


Figure 3: Number of SARS-CoV-2 genomes deposited in NCBI, by release date. The database shows unsteady sequence growth and results will likely need to be frequently updated (weekly). Release date is not easily available for the GISAID database.

Bioinformatic Analysis

Software created for this project is available on GitHub at https://github.com/BaderLab/Chan-Covid19. It is currently private to the Bader lab but will be made publicly available upon project release or publication.

Analysis was performed using shell-scripting, R and Mathematica. Sequences of lower quality were excluded from analysis. These include sequences with >550 N's, >10 gaps and a length >30kb and length <29kb.

Multiple sequence alignment (MSA) was performed using MAFFT version 7 (https://mafft.cbrc.jp/alignment/software/). We aligned all genomes and as well as the probe sequences using default parameters. MAFFT has an iterative alignment algorithm and is useful for sequences containing large gaps.

BLASTN (version 2.10.0+) was used to count the number of mismatches of each probe against target genomes, using the blastn_short tool which is optimized for query sequences < 50bp. This was run with -outfmt 6 (tab-delimited output) and num_alignments 8000, which returns all results.

MSA results were visualized using sequence logos generated by custom R code. For each position in the multiple sequence alignment, the frequency of each nucleotide A, C , G, T, H (gaps, represented by a hyphen '-' in the MSA) or N (undefined nucleotide) was counted at each nucleotide position and compared to the given probe sequence to count mismatches. The frequency of each nucleotide was visualized as the

height of corresponding letters in sequence logos. Matching logos were calculated to represent nucleotide frequency in all aligned sequences and mismatch logos were calculated to represent the frequency of nucleotides being a mismatch to the original probe sequences.

MSAs were visually evaluated using UniProt UGENE (http://ugene.net/). This visualization permitted an assessment of the probe-genome alignment region to identify systematic gaps in the reference genome MSA.

Results

Quality Control of GISAID and NCBI SARS-CoV-2 Genomes

There is a source and quality difference between GISAID and NCBI sequences, Fig 4. In general, GISAID had ten-fold more sequences than NCBI, but these sequences were of poorer quality. Individual GISAID genomes (N=7,716 sequences) could have up to 5,813 total indeterminate nucleotides or "N"s (median: 14 N; mean: 87 N per genome). In comparison, NCBI (825 sequences) genomes had up to 632 N's (median: 0; mean: 7 N). GISAID genomes also had up to 202 gaps per genome (denoted as hyphens in the sequence), whereas NCBI had no gaps in any genome. 824 NCBI and 7,512 sequences from GISAID passed quality control criteria and were included in this analysis.

GISAID has a more diverse geographic distribution of genome sequences compared to NCBI (Fig 4B-C). Overall, 97 countries are represented, with 55% making up the top three countries (27% USA, 16% England and 11% Australia). On the other hand, 87% of sequences in NCBI come from the USA, with almost all sequences from the top three countries (96%; 87% USA, 7.6% China, 1.3% Spain).

Interestingly, some countries in the GISAID collection have higher than average number of Ns. e.g. 94% sequences from Germany (total 63 sequences), 79% from England, (994 / 1256), 67% from Hong Kong (25 / 37) and 66% from Scotland (189 / 287).

There are genomic "hotspots" of N's in both the GISAID and NCBI genomes (Fig5). These occur at the 5' and 3' ends of the genome, but also at other locations. In GISAID sequences, there is a hotspot of N's precisely where the ORF1ab probes align.

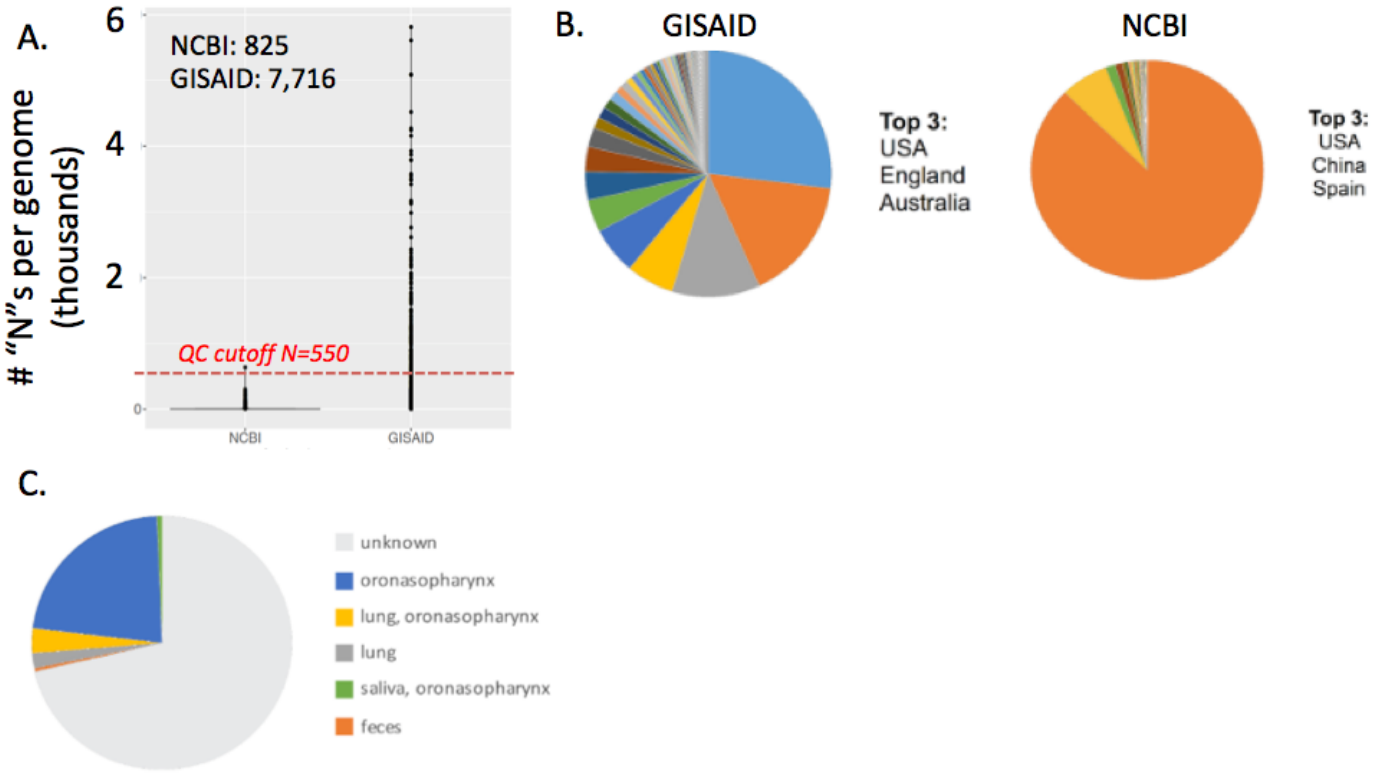


Figure 4: Overall sequence quality, geographic distribution and isolate source of GISAID and NCBI. (A) Number of indeterminate nucleotides per genome in GISAID (7,716 sequences) and NCBI (825 sequences). (B) Proportion of sequences from different countries. (C) Distribution of isolation sources of sequences in NCBI.

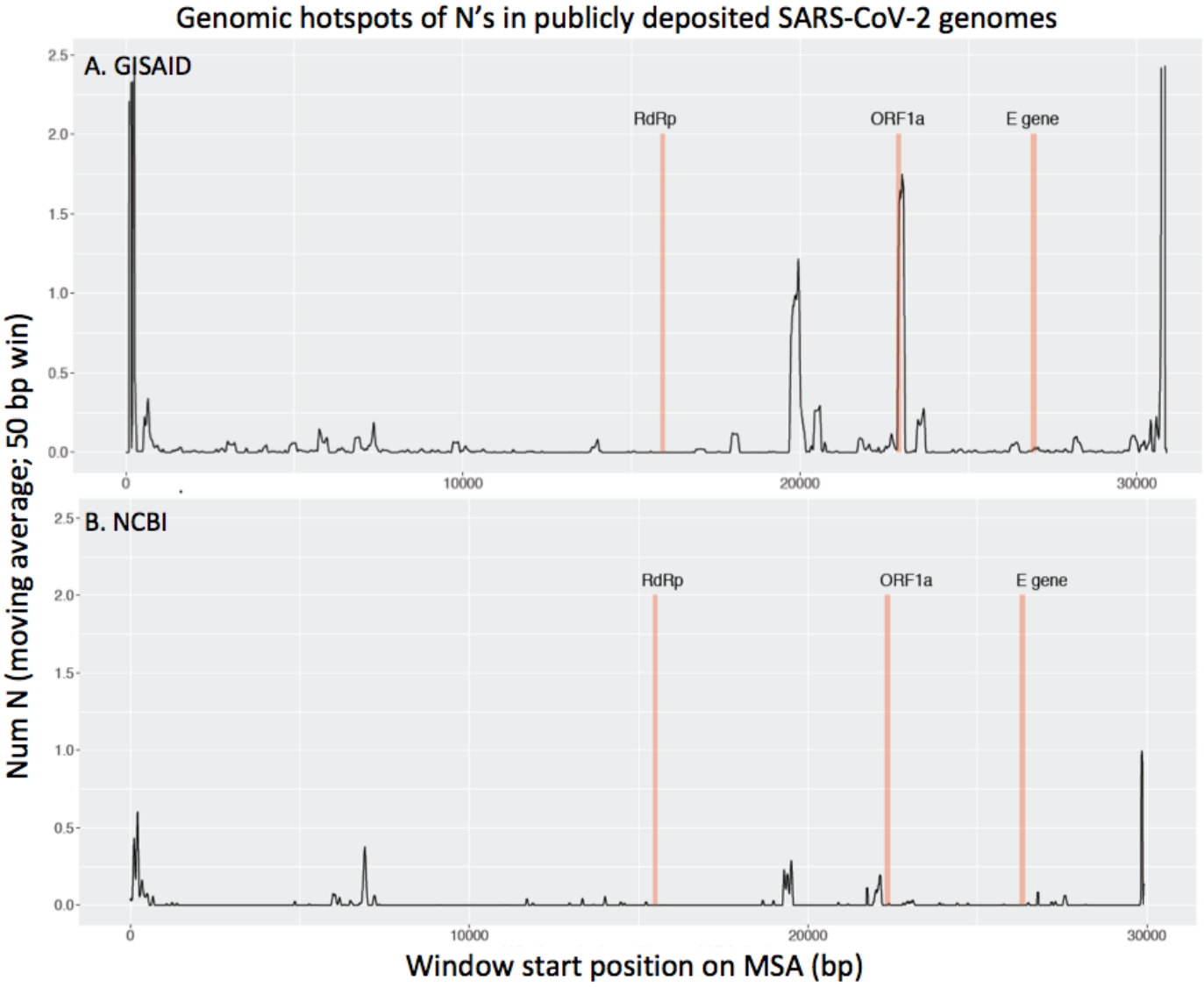


Figure 5: Genomic distribution of N's in publicly deposited SARS-CoV-2 genomes.
The graphs show a moving average of the number of N's across the genomes for (A) GISAID genomes (N=7,512 sequences) and (B) NCBI (N=824 sequences) (50 bp sliding window). Rose coloured bars indicate the position at which probes align for RdRp, ORF1ab and E genes.

Mismatch Count for All Probes

Our main objective is to explore how well each probe matched to the target region. This was accomplished by analyzing the results of primer/probe to viral genome alignment using BLASTN (version 2.10.0+). Our results showed a low number of mismatches between the target sequence and probe/primer regions see Fig 6 for details. Length analysis using BLASTN identified an 8 nucleotide length difference between the RdRp forward primer sequence and all genomes - there were 4 nucleotides from both ends of the primer (at positions 1-4 and 28-31) that were not aligned to any genomes. The length of the probes or primers also showed a discrepancy between the ORF1ab alignment region, however we could not identify a trend (see further analysis).

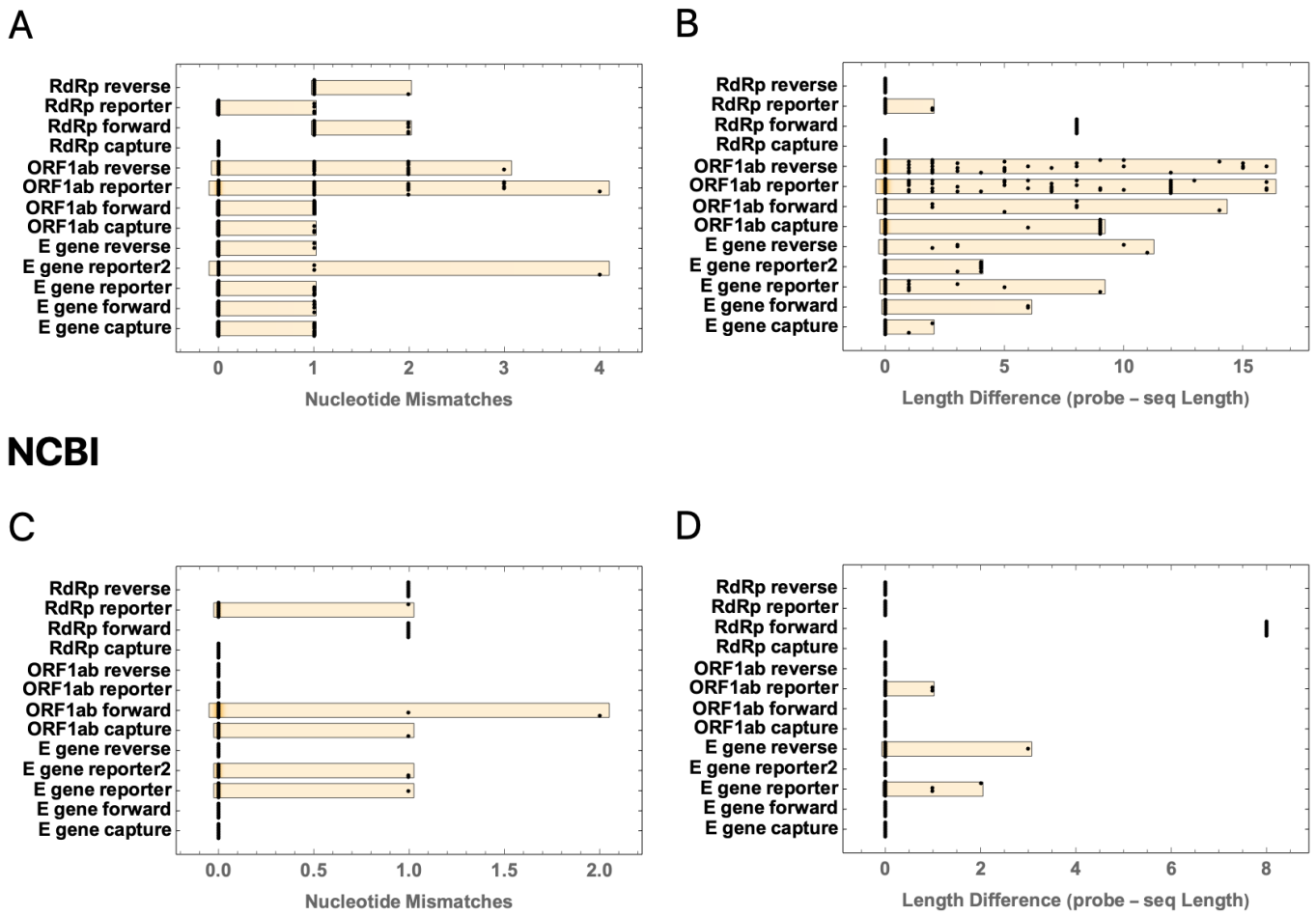
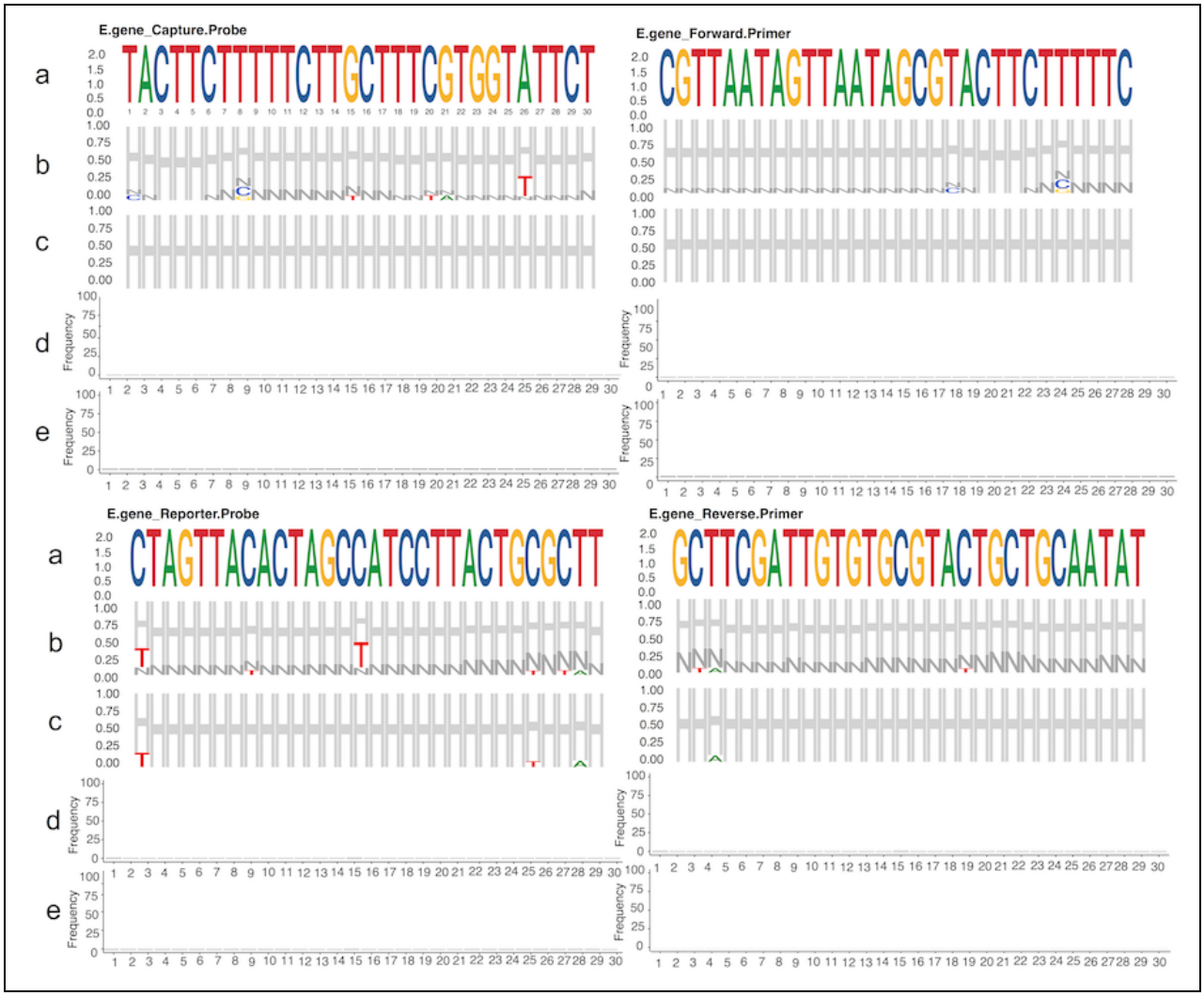


Figure 6: Using BLASTN we aligned the probe or primer sequences to the set of whole viral genomes retrieved from (A-B) GISAID (n>7100) and (C-D) NCBI (n=824) databases. A,C) The number of nucleotide mismatches between the query (probe or primer) sequence and the viral genome sequences. We observe a low frequency (<0.002) of mismatches in all E gene probes and primers; a low frequency (<0.011) of mismatches in ORF1ab probes and primers; RdRp reverse primer has mismatches with all genomes; a low frequency (<0.001) of the genomes had 1 mismatch with the RdRp reporter probe; a high frequency (~1) of the genomes had 1 mismatch with the RdRp forward primer; and zero mismatches were seen with the RdRp capture probe. B, D) The difference between the length of the probe sequence and the aligned genome sequence identifies query sequences that did not completely align with the target region.

Probe-genome Alignment: E gene

Analysis of the E gene identified that regions covered by the probes are 100% stable or extremely close to it for all positions, as indicated by the MSA consensus sequences (Fig 7). Similar results were found for both NCBI and GISAID. At a frequency of 0.36%, a C>T mismatch was detected in the reporter probe in NCBI. Fig 7. Other mismatches (H, N and T) were present at a very low frequency less than 0.25%. The region covered is very stable within and around the probe region of interest (Fig 8).


E.gene_Capture.Probe
E.gene_Forward.Primer
E.gene_Reporter.Probe
E.gene_Reverse.Primer
a
b
c
d
e
Frequency

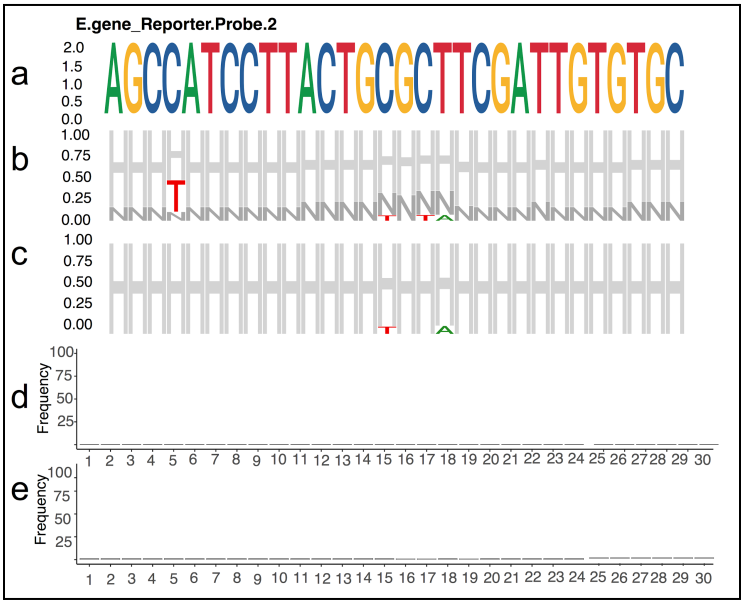

E.gene_Reporter.Probe.2
a
b
c
d
e
Frequency

Figure 7: Consensus sequence and mismatches for NCBI and GISAID results using the multiple sequence alignment. We generated an MSA consensus sequence for all sequences to identify the probability (height) of nucleotide occurrence at each position. The position can indicate the nucleotides (A, C, G, T) or a gap (H) or an undefined nucleotide (N). **Mismatch:** shows the nucleotides that were found as a mismatch when compared to the original probe sequence (a). The height of each letter indicates nucleotide mismatch frequency compared to all mismatched nucleotides (b: GISAID, c: NCBI). The mismatch frequency in the genome sequences is indicated in histograms (d: GISAID, e: NCBI).

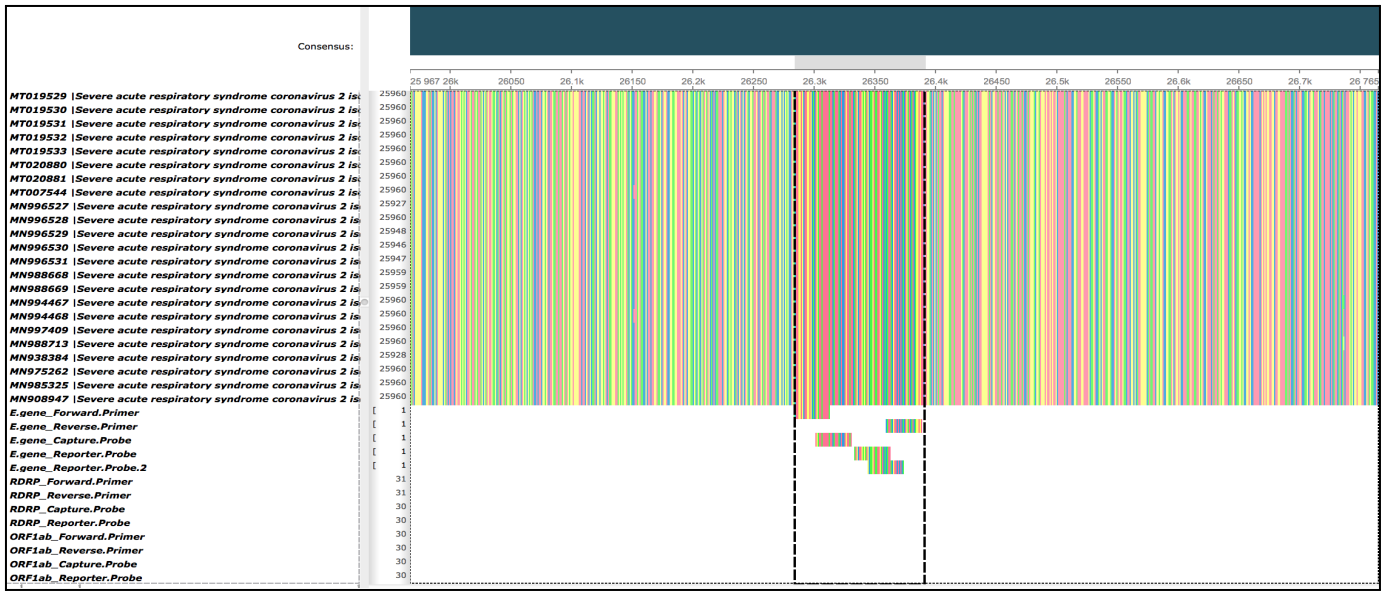


Figure 8: View of subset of E gene probe MSA: Alignment of the probe and primer sequences to the E gene region, as viewed in UGENE The black dotted rectangle indicates the probe regions. The region covered is very stable within and around the probe region of interest.

Probe-genome Alignment: RdRp gene

Analysis of the RdRp gene identified 3 mutations in the forward primer region (G>A, A>G , T>C) and one in the reverse primer (T>C) present at a frequency of 100%. No genome contains the base present in the primers at these mismatch positions (Fig 9). Results are identical in both NCBI and GISAID. Thus, the RdRp forward and reverse primers contain non-matching positions and should be redesigned. Additionally:

- The T>C mutation in the forward primer is not present in the capture probe, which contains a T at that position (pos 7) as in the viral sequences.
- The region covered is very stable within and around the probe region of interest Fig 10.
- The mutations that were identified are visible on a close zoomed view of the multiple sequence alignment (see Fig 10 bottom panel).
- A few other mutations occuring are observable in each probes/primers but are present at a very low frequency (<0.25%).

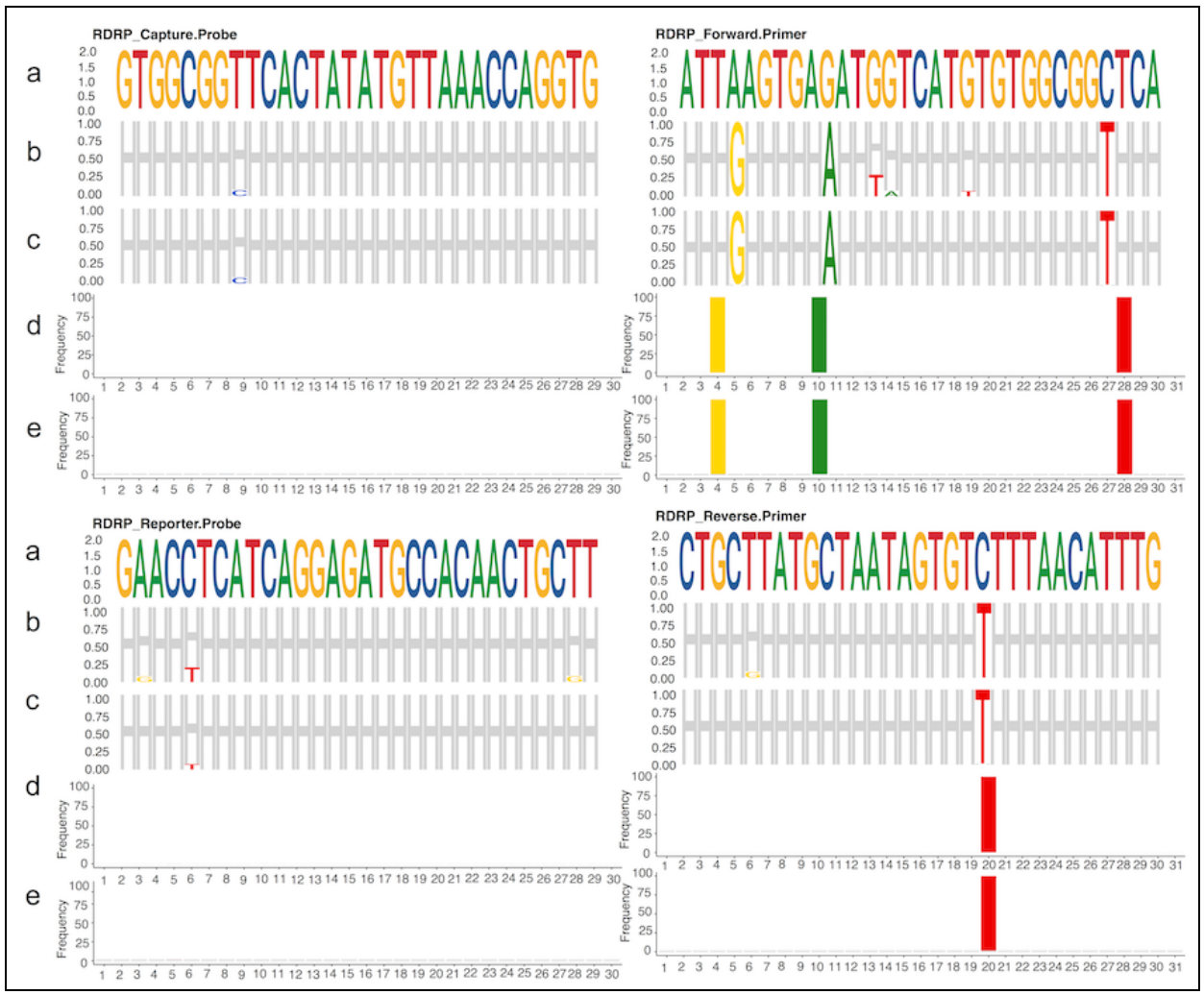
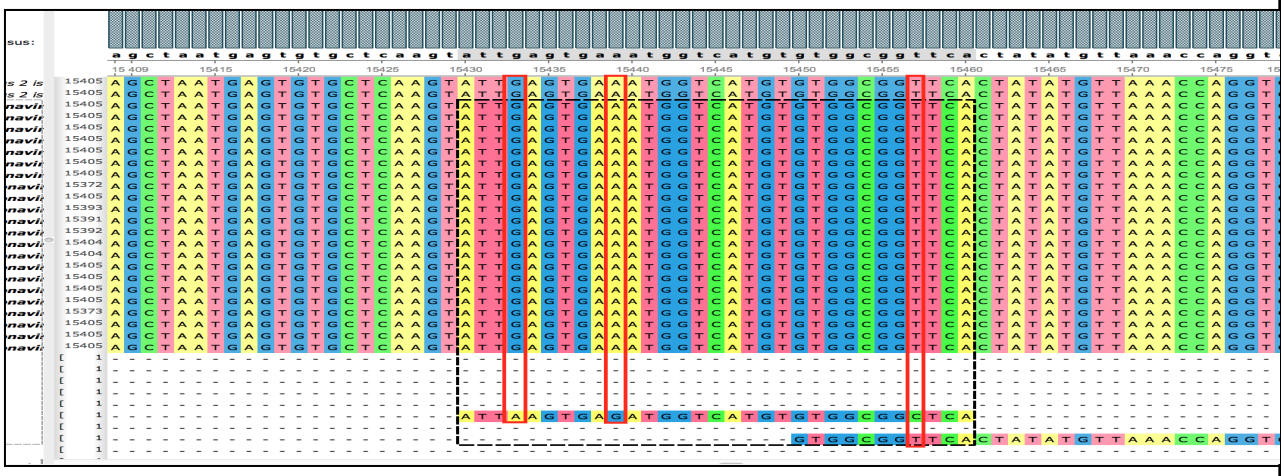


Figure 9: Consensus sequence and mismatches NCBI results for RdRp (legend as for Fig 7).

A. RdRp Forward Primer Mismatch View (UGene)



B. RdRp Reverse Primer Mismatch View (UGene)

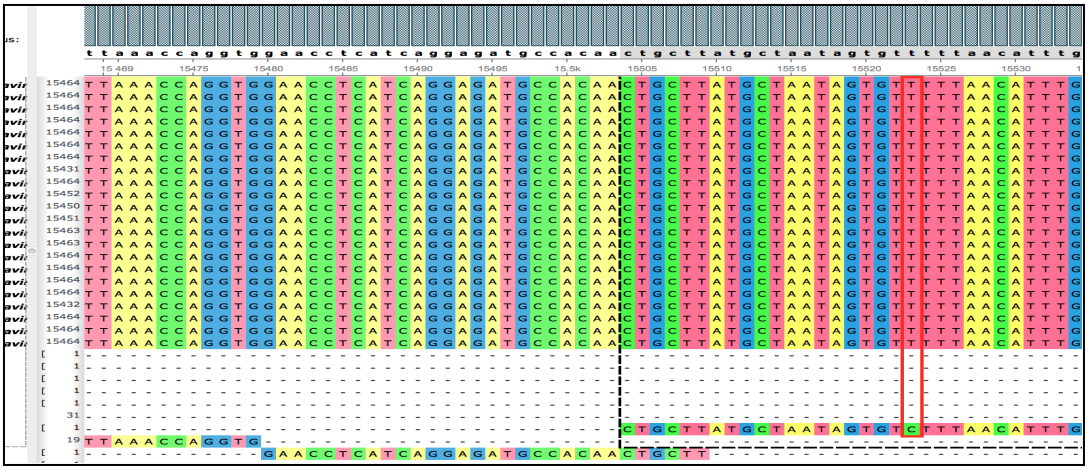


Figure 10: Multiple sequence alignment view using UGENE of the probe target regions for the RdRp gene (NCBI database) for a subset of sequences. Top: The black dotted rectangle indicates the forward primer and capture probe regions. The red rectangles indicate the 3 forward primer mutations (G>A, A>G, T>C) that were identified in the forward primer but not present in the capture probe. Bottom: The black dotted rectangle indicates the reverse primer region. The red rectangle indicates the mutation (T>C) that was identified in the sequence. The region within and around primers is stable.

Probe-Genome Alignment: ORF1ab gene

Analysis of the ORF1ab gene using available sequences identified no frequent mutations (A, C, G ,T). The only mutations that were detected, like T, G, A, were present at a very low frequency less than 0.25%, see Fig 11.

- In NCBI, the region covered by and around the probes is stable with no mutations. There are only some Hs at a frequency of 1.3% which likely represent background noise. However, the GISAID results for the ORF1ab regions are different (Fig 11). We observe an unusual pattern with a frequency of N's in the reporter probe and reverse primer that is absent in the forward primer and capture probes at a frequency of 3.2% (Fig 12) .
- In Fig 13, the view has been adjusted to show some genomes with N's in the reporter probe and reverse primer regions. This region of N's is about 220 bp long and is associated in this example with samples from Australia. The N region starts within the reporter probe and reverse primer region and extends in the 5' direction.

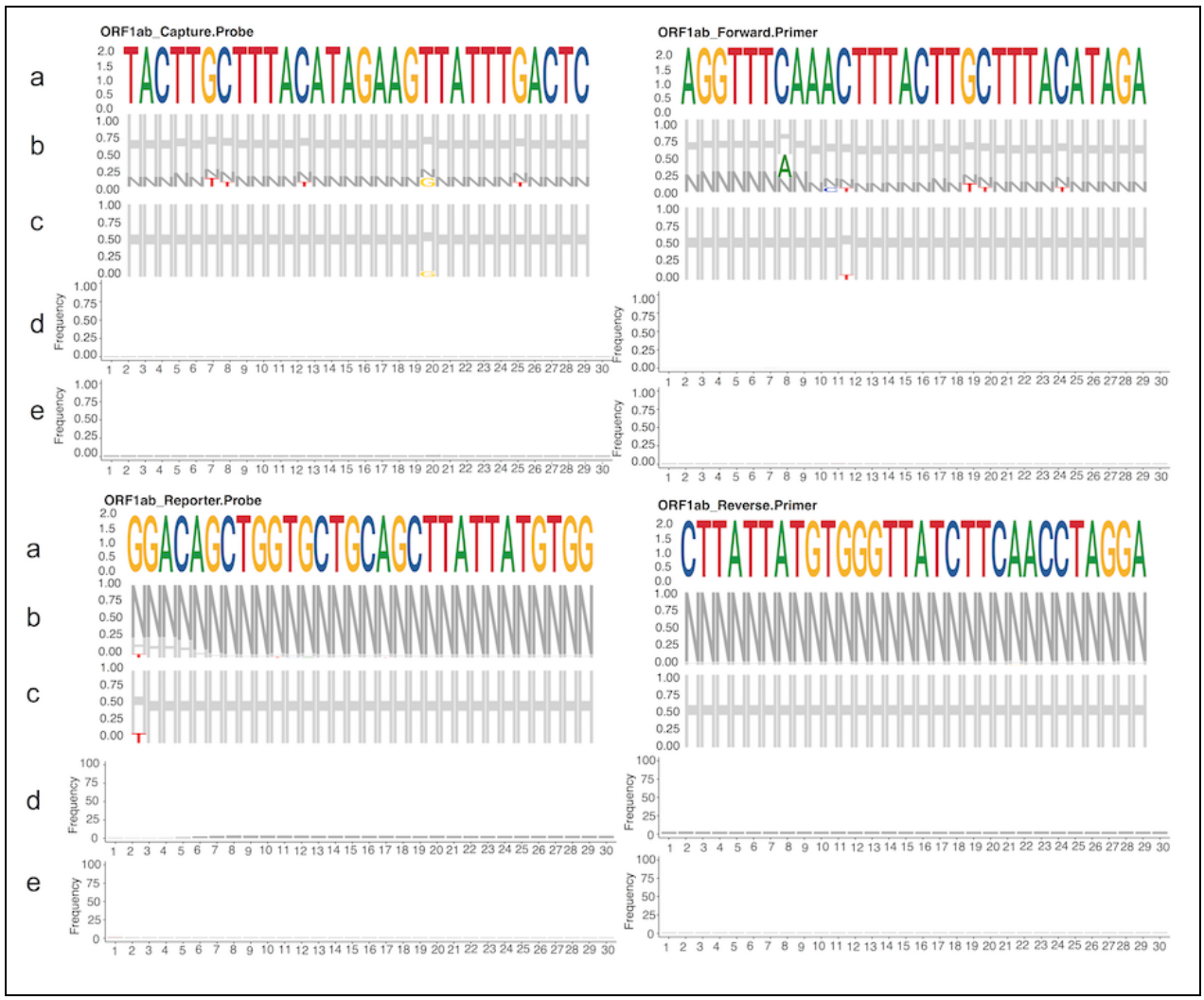


Figure 11:Consensus sequence and mismatches NCBI results for ORF1ab (legend as for Fig 7).

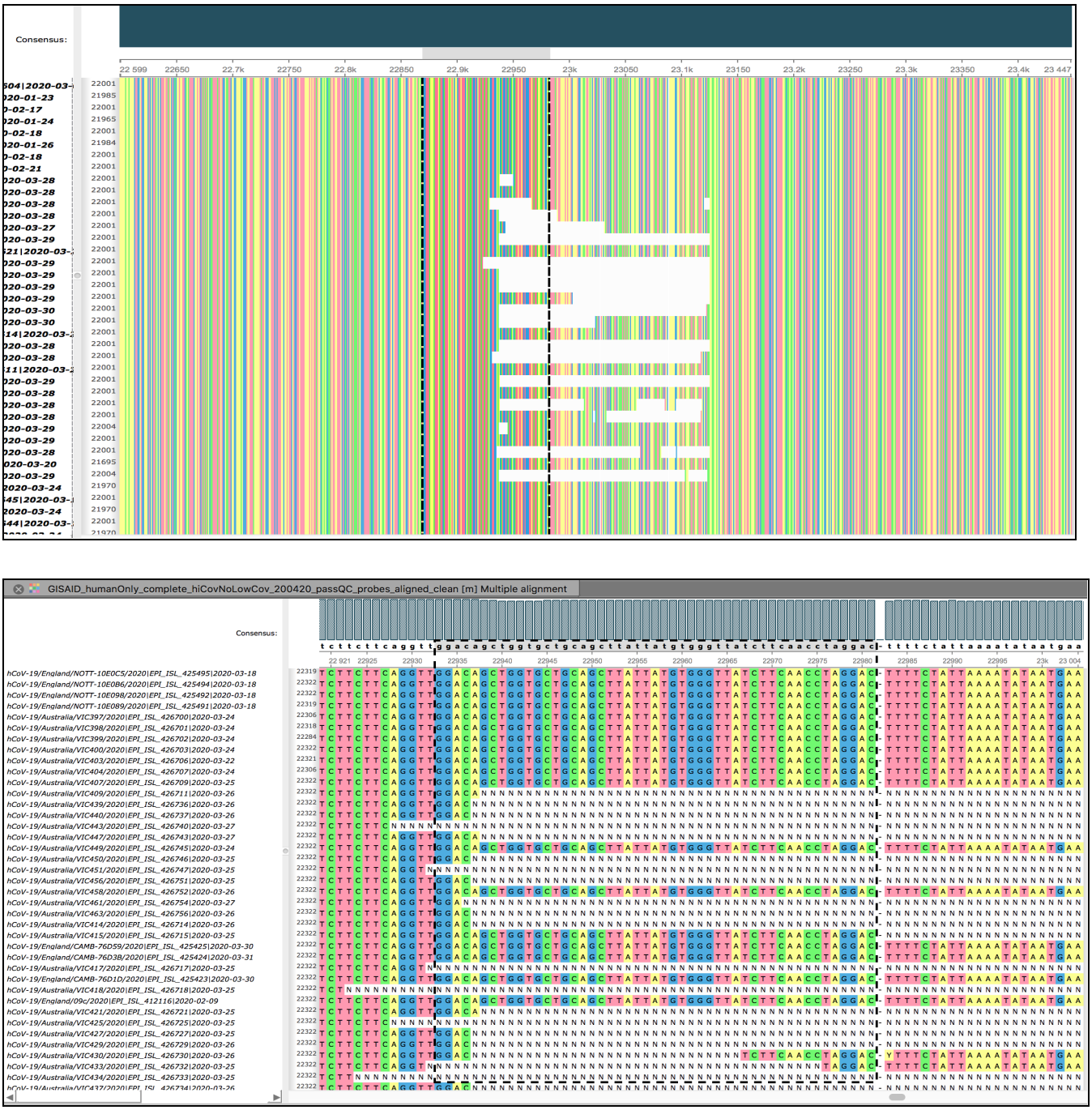


Figure 12**:** Multiple sequence alignment view using UGENE of the ORF1ab probe region for GISAID sequences. The top panel shows a broader view, and the bottom panel shows a zoomed in view. Unlike the NCBI sequences, GISAID sequences presented a region of N's covering the reporter probe and reverse primer regions. The black dotted rectangle indicates the probe regions. The samples with N's come from Australia in this example. Bottom: Closer look at the N's region. The black dotted rectangle indicates reporter primer and reverse probe regions. On the left side, the names of the samples indicate that the samples with N's are coming from Australia. The N region starts at the reporter probe and reverse primer region and continues in the 3'direction.

Genome Quality around ORF1ab Probe Region

As shown above, GISAID and NCBI genomes show hotspots for "N" nucleotides. In the GISAID database, one of these starts exactly where ORF1ab probes align (see Fig 5), but in the NCBI database this is not the case. In both collections, there is a strong significant correlation between the total number of N's in the genome and the number of N's in the 500bp region around alignment of ORF1ab probes; for GISAID, correlation is 0.64 ($p < 2e\text{-}16$) and for NCBI, the correlation is 0.62 ($p < 2e\text{-}16$) as shown in Fig 13.

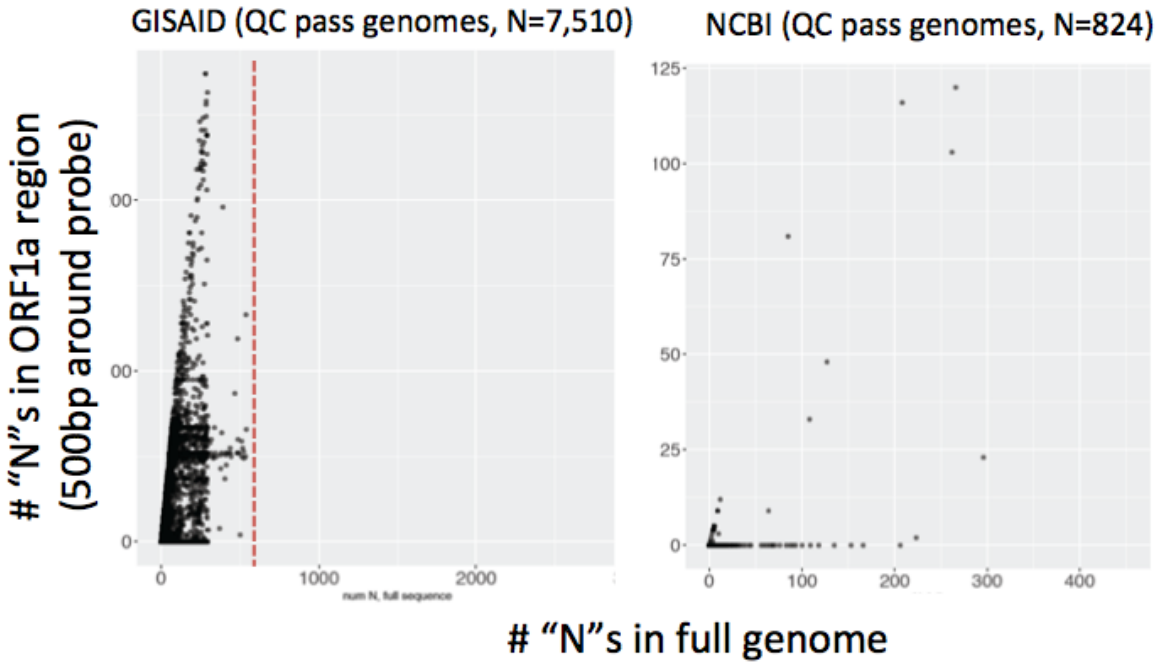


Figure 13. Ns in the genome (x-axis) as compared to those in the 500bp around the region of alignment of ORF1ab probes (y-axis). Data shown for GISAID sequences (left: 7,510 sequences passing QC) and NCBI (right: 824 sequences passing QC).

Extended Data

Link to Multiple sequence alignment files and result files: CURRENT REPORT

Note: Multiple sequence alignment files can be viewed using the freely downloadable UGENE (http://ugene.net/) visualization tool. Probes are located at the bottom of the alignment. RdRp probes are aligned at about 15.5kb, ORF1ab at about 22.3kb and Egene at about 26.3Kb (based on the NCBI alignment).

Extended data also contain the mismatch frequencies and counts at each nucleotide position for each probe.

Questions for Chan Lab

- N's in the ORF1ab region seems to be somehow correlated with overall N's in the genome sequence. What could cause N's to be so systematic in the ORF1ab region? Mutation (deletions)? 3D RNA structure that hinders high-quality sequencing?
- For RdRp, the C mutation in the forward primer is not present in the capture probe which contains a T at that same position. How is this internal data inconsistency possible?

Next steps

The virus is evolving (https://academic.oup.com/nsr/article/doi/10.1093/nsr/nwaa036/5775463) and the viral genome databases are growing (e.g. NCBI: estimate of 60 new sequences per day) , which suggests that we should repeat this analysis regularly. We suggest to:

- Continue to monitor new genome sequences:
    - As the number of sequences available in the databases is growing daily, a new alignment would be created with new sequences only: it could be done, weekly, biweekly, monthly or as requested.
    - It is possible to automate sequence download from NCBI but not GISAID which has to be manually downloaded.
    - If the number of sequences is too large, they might be separately analyzed by geographic region or phylogenetic clade. Results would be provided for each group. This may help us improve signal to noise ratios in the analysis.
    - GISAID is working to make additional genome sequence metadata available that could improve our genome selection criteria.

Extras

Gene		Forward primer	Reverse primer	Capture	Reporter	Reporter 2
E gene	**NCBI**	None	None	None	Pos1 (C>T) 0.36%	None
	GISAID	None	None	None	None	None
RdRp	**NCBI**	Pos 4 (A>G) 98% Pos 10 (G>T) 98% Pos 28 (C>T) 98% All Pos (H) 1.4%	Pos 20 (C>T) 98% All Pos (H) 1.4%	All Pos H 1.3%	All Pos H 1.4%	-
	GISAID	Pos 4 (A>G) 100% Pos 10 (G>T) 100% Pos 28 (C>T) 100%	Pos 20 (C>T) 100%	None	None	-
ORF1ab	**NCBI**	All pos H 1.4%	All pos H 1.3%	All pos H 1.3%	All pos H 1.3%	-
	GISAID	None	All pos N 3.2%	None	All pos N 3.2%	-

Table 2. Percent mismatch of probe-genome alignment with CoV-2 genome sequences from NCBI (824 sequences) and GISAID (7,512). Only genome sequences with <500 N's, <10 gaps and <30kb considered. * Mismatches are reported in the table if they are present in equal or more than 0.25% of the sequences (all results are available in extended data).

Contributions:
Maria Abou Chakra
Zoe Clarke
Shirley Hui
Ruth Isserlin
Shraddha Pai
Delaram Pouyabahar
Veronique Voisin
Gary Bader