Benefits of Hybrid Capture for Viral Surveillance of SARS-CoV-2

Danny Antaki, Tong Liu, Kristin Butcher, Bryan Höglund, Jean Challacombe, Esteban Toro
Twist Bioscience

1. Abstract

Diagnostics and vaccines against novel SARS-CoV-2 (SCV-2) strains rely on viral genome sequencing. Researchers have gravitated towards the cost-effective and highly sensitive amplicon-based (e.g. ARTIC) and hybrid capture sequencing (e.g. SARS-CoV-2 NGS Assay) to selectively target the SCV-2 genome. To demonstrate the advantages of hybrid capture for viral surveillance, we implemented predictive modeling to assess areas in the SCV-2 genome that may be more prone to mutations that could impact primer efficiency in the multiplexing PCR step during amplicon sequencing.

We analyzed 383,656 genome sequences of variant of concern (VOC) and variant of interest (VOI) isolates from GISAID and found 101,432 viruses (27%) had ≥1 mismatch in the last 6 base pairs from the 3' end of ARTIC primers. In contrast, only 38 viruses (0.01%) had enough mutations (≥10) predicted to have a similar effect on hybrid capture sequencing. Our approach identified 4 isolates with excess mutations in ARTIC primers, which we produced synthetic genomes of and compared the performance of ARTIC amplicon sequencing with Twist Bioscience's SARS-CoV-2 NGS Assay. For each strain, we observed dropouts in sequencing coverage for amplicon libraries only but not for hybrid capture. Incidentally, we observed an additional dropout of ARTIC amplicon 72 in two samples caused by a recurrent 1bp mismatch in the reverse primer. We also compared dropout using invariant reference controls (Twist Synthetic SARS-CoV-2 RNA Control 2) and found hybrid capture sequenced 99.5% of bases at ≥50X coverage compared to 92.1% for amplicon sequencing.

Taken altogether, our results demonstrate that hybrid capture is more robust to genomic variation and leads to fewer dropout events. Although amplicon sequencing provides a low-cost platform for viral surveillance, it comes at the added cost of sequencing dropout which can cause incorrect classification of viral lineages. This is further compounded by the certainty that mutations will accumulate in SCV-2, making it a matter of time until more primers fail. Therefore we suggest that Hybrid Capture is a robust and comprehensive solution for viral surveillance.

2. Methods

1,067,579 SCV-2 genomic sequences were acquired from GISAID (2021-04-21) in a multiple sequence alignment format, and variants were called using the *faToVcf* tool (**Figure 2.1**). Restricting to VOC lineages, a total of 6,777 unique alleles were observed (588,922 mutations total) in ARTIC primers. We then selected viruses with mutations in the last 6bp from the 3-prime end of ARTIC primers, resulting in 101,432 isolates. We then constructed synthetic genomes of 4 candidate viral genomes (**Table 2**) to test if the mutations would lead to dropouts after amplicon sequencing.

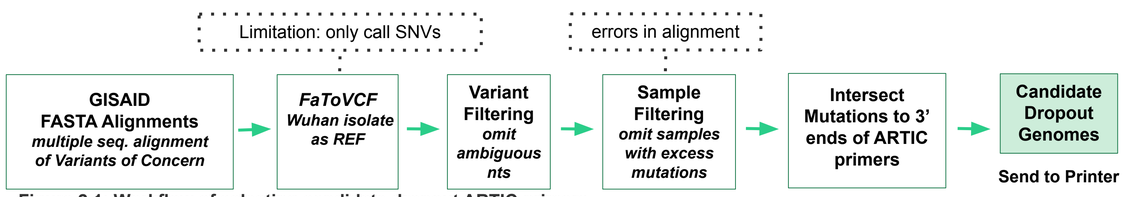

Figure 2.1: Workflow of selecting candidate dropout ARTIC primers

GISAID Accession Numbers	WHO Label	Pango Lineage	Collection date	Location	Mutated ARTIC Primer	Mutations	Gene
EPI_ISL_837547	Epsilon	B.1.429	2020-12-23	Washington, United States	nCoV-2019_73_LEFT	c.21984_21985delGG, c.21986_21988GTG>CAT	S
EPI_ISL_1366445	Epsilon	B.1.429	2021-02-16	California, United States	nCoV-2019_24_LEFT	c.7056_7059CTGG>AAAA	ORF1a
EPI_ISL_1540525	Alpha	B.1.1.7	2021-02-25	Prague, Czech Republic	nCoV-2019_87_LEFT	c.26214_26215delTT, c.26216_26217TG>AC	ORF3a
EPI_ISL_1108224	Alpha	B.1.1.7	2021-02-08	England, United Kingdom	nCoV-2019_41_RIGHT	c.12466_12468AGC>GAA	ORF1a

Table 2: Sequencing Dropout of Select SCV-2 Variants

For the four synthetic genomes and an additional synthetic control genome of the Wuhan isolate (Twist Synthetic SARS-CoV-2 RNA Control 2), we carried out **amplicon sequencing** and **hybrid capture sequencing** using the v3 protocol from the ARTIC network and the SARS-CoV-2 NGS Assay respectively (**Figure 2.2**). Sequencing was carried out on an Illumina NextSeq 500 and reads were trimmed to 74bp and downsampled to 100,000 reads. Alignment was performed using BWA-MEM while marking duplicates. Since the viral synthetic genomes were constructed in 5kb fragments, we only considered primer pairs that fully spanned an individual fragment, resulting in the omission of 5-6 ARTIC primer pairs depending on the sample.

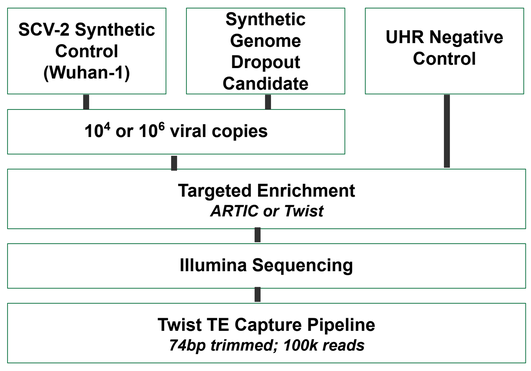

Figure 2.2: Workflow of performance comparison between ARTIC amplicon sequencing and hybrid capture sequencing

3. Mismatches in Amplicon Primers and Hybrid Capture Probes

The total number of mismatches that intersected either ARTIC amplicon primers or Twist hybrid capture probes was counted for each virus (**Figure 3.1**). Since Twist hybrid capture probes cover the entirety of the SCV-2 genome, the number of mismatches is equal to the total number of mismatches called for a given virus (mean = 33). In contrast, the 218 ARTIC primers, ranging 22-30bp in length, overlap 4,908 base pairs (16.4%) of the SCV-2 genome, and were found to have an average of 1.9 mismatches per virus. Twist hybrid capture probes are robust to mismatches exhibiting 50% efficiency at ≥10 mismatches.

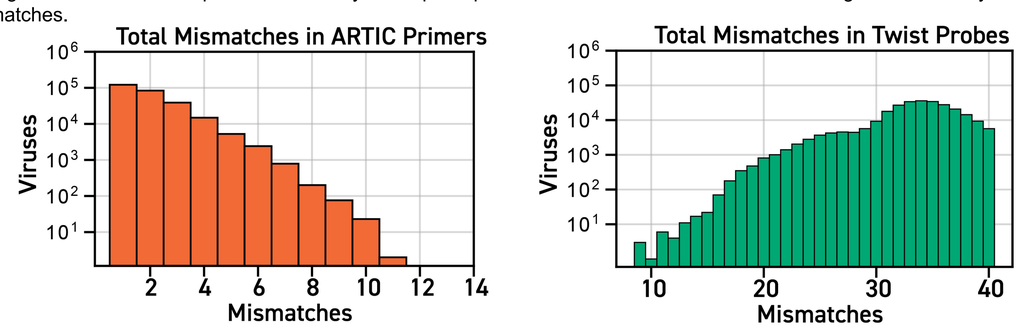

Figure 3.1 VOC Mismatches in ARTIC Amplicon Primers and Twist Hybrid Capture Probes.

For each virus, the total number of mismatches within a sole primer or probe was quantified (**Figure 3.2**). For ARTIC primers, only mismatches within the last 6bp from the 3-prime end were considered. 101,432 distinct VOC viruses (27%) had at least 1 for ARTIC primers. In contrast, the number of VOC viruses with 10 or more mismatches (50% efficiency; orange line) in Twist probes was 38 isolates (0.01%).

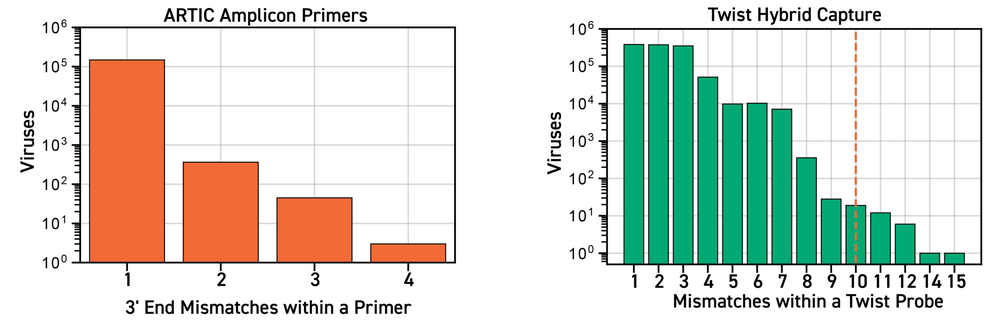

Figure 3.2: VOC Mismatches within a Single Amplicon Primer or Hybrid Capture Probe.

4. Candidate Dropouts and Results

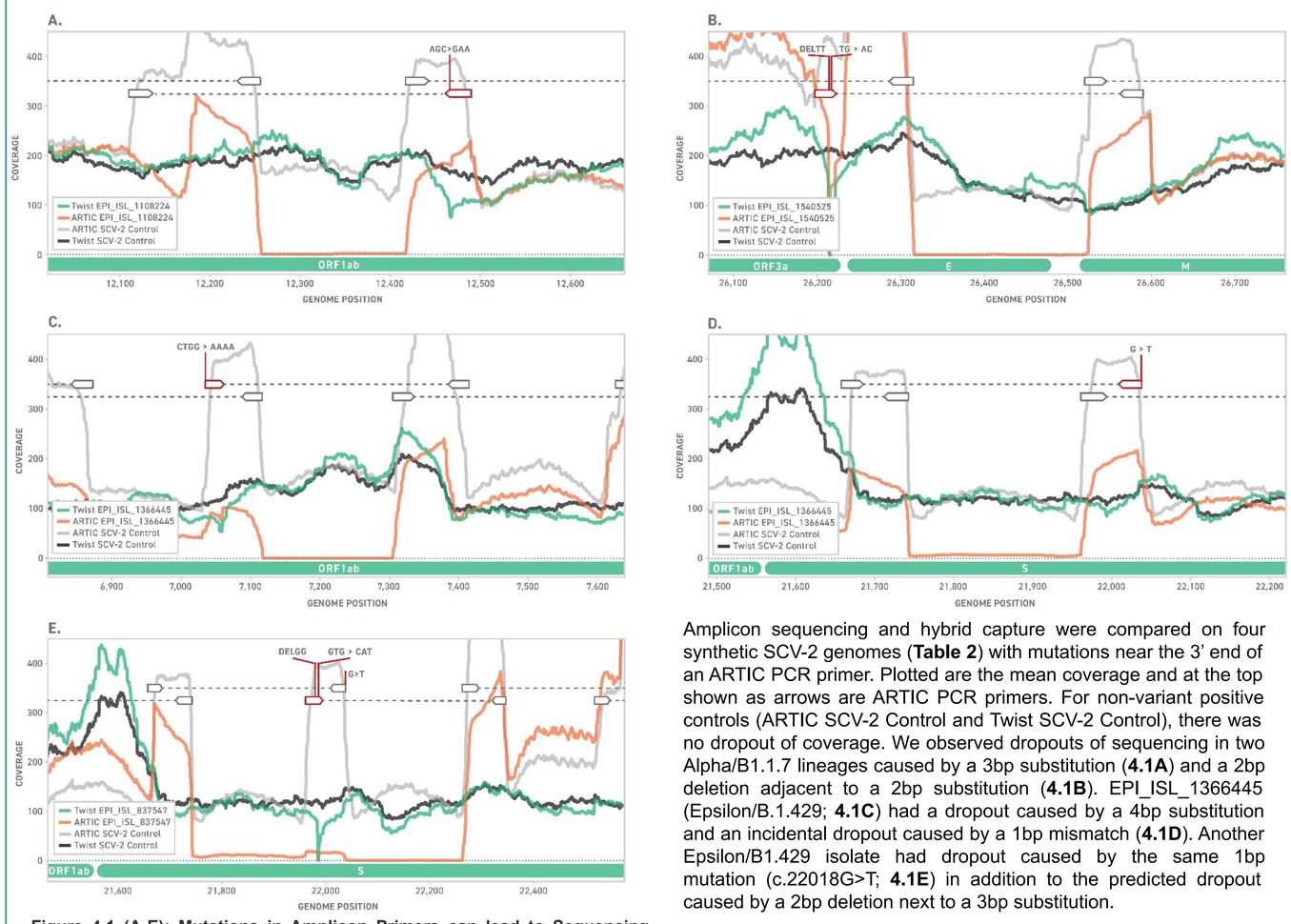

Figure 4.1 (A-E): Mutations in Amplicon Primers can lead to Sequencing Dropout

Amplicon sequencing and hybrid capture were compared on four synthetic SCV-2 genomes (**Table 2**) with mutations near the 3' end of an ARTIC PCR primer. Plotted are the mean coverage and at the top shown as arrows are ARTIC PCR primers. For non-variant positive controls (ARTIC SCV-2 Control and Twist SCV-2 Control), there was no dropout of coverage. We observed dropouts of sequencing in two Alpha/B1.1.7 lineages caused by a 3bp substitution (**4.1A**) and a 2bp deletion adjacent to a 2bp substitution (**4.1B**). EPI_ISL_1366445 (Epsilon/B.1.429; **4.1C**) had a dropout caused by a 4bp substitution and an incidental dropout caused by a 1bp mismatch (**4.1D**). Another Epsilon/B1.429 isolate had dropout caused by the same 1bp mutation (c.22018G>T; **4.1E**) in addition to the predicted dropout caused by a 2bp deletion next to a 3bp substitution.

We calculated the total number of mutations at each position from the 3-prime end for every ARTIC primer (**Figure 4.2**). We found a mutation rate of ≥5% of viruses analyzed (293,130) for the first 10bp from the 3-end. Four positions were mutated in ≥10% of the dataset. As SCV-2 continues to spread, more mutations will accumulate in these regions, complicating variant calling and potentially assigning incorrect variant lineages to samples.

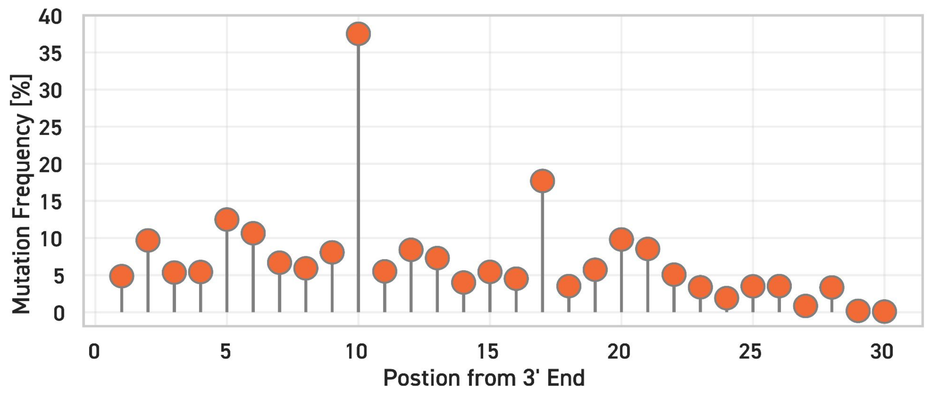

Figure 4.2: Accumulation of Mismatches in ARTIC Amplicon Primers Across 293,130 Isolates plotted from the 3-prime end

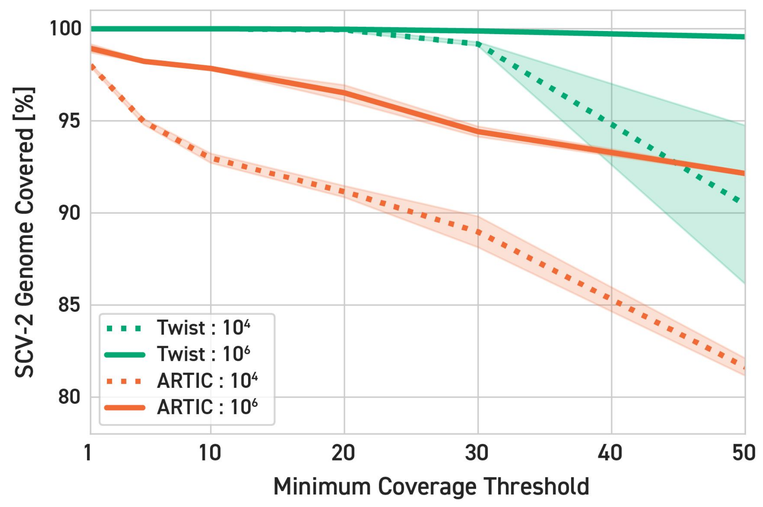

Figure 4.3: Twist's SARS-CoV-2 NGS Assay covers more of the SCV-2 genome than amplicon sequencing.

The total percent of the SCV-2 genome covered by hybrid capture and amplicon sequencing was compared at 10^4 and 10^6 viral copies (**Figure 4.3**). Shaded regions represent one standard deviation from the mean for two replicates. At a viral titer of 10^6 hybrid capture covered 100% of bases at 1X in comparison to 98.9% of bases at 1X for amplicon sequencing. At a lower titer of 10^4, hybrid capture was still able to cover 100% of sequenceable bases in contrast to 98% of bases at 1X. Both methods are able to effectively cover most of the SCV-2 genome, but surveillance of extant and emerging SCV-2 strains is contingent on variant calling which requires high coverage for accurate interpretation. At 50X coverage hybrid capture covered 99.5% and 90.4% of the genome at 10^6 and 10^4 viral copies in comparison to 92.1% genome covered at 10^6 copies and 81.6% genome covered at 10^4 copies.

5. Conclusions

Since the onset of the SCV-2 outbreak in 2019, researchers have been studying the evolution of the virus's genome. By monitoring how variations in the viral genome emerge and propagate throughout the population through genome sequencing, health providers around the globe can make more informed decisions about public health. Likewise, pharmaceutical research in treatments, cures, and vaccines for SCV-2 is contingent on ongoing genome surveillance. When this work was carried out the Alpha lineage was dominant in the United States, which has been subsumed by the Delta lineage, demonstrating the pressing need to implement a sequencing method that is robust to genomic variation. Here in this study we have demonstrated the benefits of hybrid capture for genome sequencing of SCV-2 over amplicon sequencing.

The results presented in this technical note are corroborated by a recent study published by Klempt et al. (2020). Their study compared three methods for sequencing SARS-CoV-2 samples: Target capture from Twist and Illumina, and amplicon sequencing from Paragon. Target Capture methods presented greater uniformity in coverage, and lower false positive rates compared to amplicon sequencing. Additionally, a recent study by Doddapaneni et al. (2021) showed that Twist Bioscience's target capture based assay has further benefits in SARS-CoV-2 surveillance as it allows for the simultaneous target capture and quantitation of subgenomic coronavirus RNAs, which is not possible with amplicon sequencing.

Although amplicon sequencing provides a low-cost platform for viral surveillance, it comes at the added cost of sequencing dropout which can cause incorrect classification of viral lineages. This is further compounded by the certainty that mutations will accumulate in SCV-2 (**Figure 4.2**), making it a matter of time until more primers fail. In conclusion, effective viral surveillance requires comprehensive genome sequencing as a reliable platform for the continuous monitoring of variants as they occur.

Klempt, P. et al. (2020). Performance of Targeted Library Preparation Solutions for SARS-CoV-2 Whole Genome Analysis. Diagnostics. 2020; 10(10):769. https://doi.org/10.3390/diagnostics10100769

Doddapaneni, H. et al. (2021). Oligonucleotide capture sequencing of the SARS-CoV-2 genome and subgenomic fragments from COVID-19 individuals. PLoS ONE 16(8): e0244468. https://doi.org/10.1371/journal.pone.0244468

Financial Disclosures: All authors are employees and shareholders of Twist Bioscience