Performance Characteristics of Sequencing Assays for Identification of the *SARS-CoV-2* Viral Genome

Danny Antaki, Mara Couto-Rodriguez, Tong Liu, Kristin Butcher, Esteban Toro, Bryan Höglund, Xavier Jirau Serrano, Joseph E. Barrows, Bradley A. Connor, Christopher E. Mason, Niamh B. O'Hara, Dorottya Nagy-Szakal

Biotia

TWIST BIOSCIENCE

1. Abstract

As the SARS-CoV-2 (SCV-2) virus evolves, diagnostics and vaccines against novel 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. We provide an in silico predictive model to compare these 2 technologies and present additional data on the high scalability of the SARS-CoV-2 NGS Assay.

In silico work implemented alignments of 383,656 genome sequences of variant of concern (VOC) and variant of interest (VOI) isolates (GISAID). We compared the clinical performance of ARTIC V3 amplicon sequencing and SARS-CoV-2 NGS Assay (Twist Bioscience) using viral isolates with mutations within amplicon primers and hybrid capture probes. The miniaturized hybrid capture workflow was optimized and clinically evaluated to support high-throughput (384-plex), and sequencing data was processed by COVID-DX software.

We 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. We then created synthetic genomes of 4 isolates with excess mutations in ARTIC primers and observed dropouts in sequencing coverage for amplicon libraries but not hybrid capture. In fact, we observed an incidental dropout of amplicon 72 in 2 samples caused by a recurrent 1bp mismatch. Both assays detected a wide range of variants (~99.9% coverage at 5X depth) in clinical samples (CT value < 30) collected in NY (Spring 2020-Spring 2021). The distribution of the number of reads and on target rates were more uniform among specimens within amplicon-based sequencing. However, uneven genome coverage and primer dropouts, some in the spike protein, were observed on VOC/VOI and other isolates highlighting some limitations of an amplicon-based approach.

The SARS-CoV-2 NGS Assay is a comprehensive and scalable sequencing tool for variant profiling, yields more consistent coverage and smaller dropout rate compared to ARTIC (0.05% vs. 7.7%).

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.

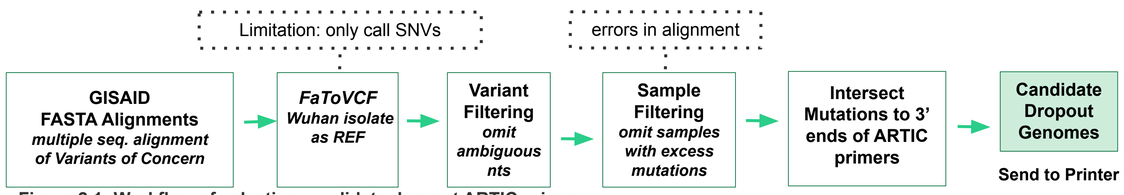


Figure 2.1: Workflow of selecting candidate dropout ARTIC primers

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.

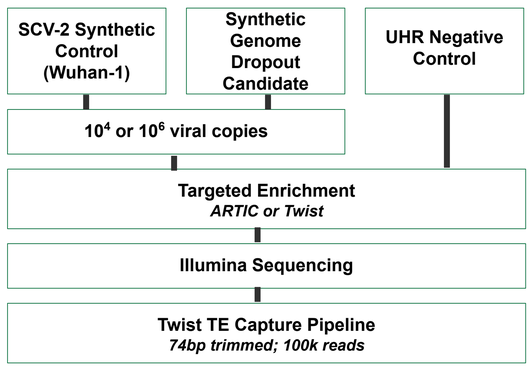


Figure 2.2: Workflow of performance comparison between ARTIC amplicon sequencing and hybrid capture sequencing

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.

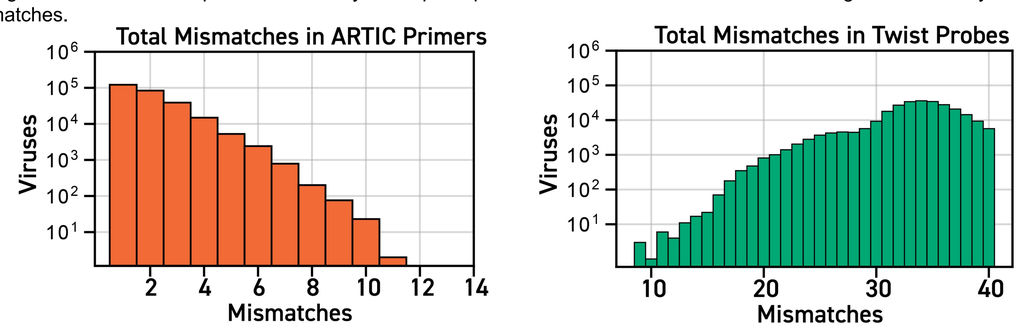


Figure 3.1 VOC Mismatches in ARTIC Amplicon Primers and Twist Hybrid Capture Probes.

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%).

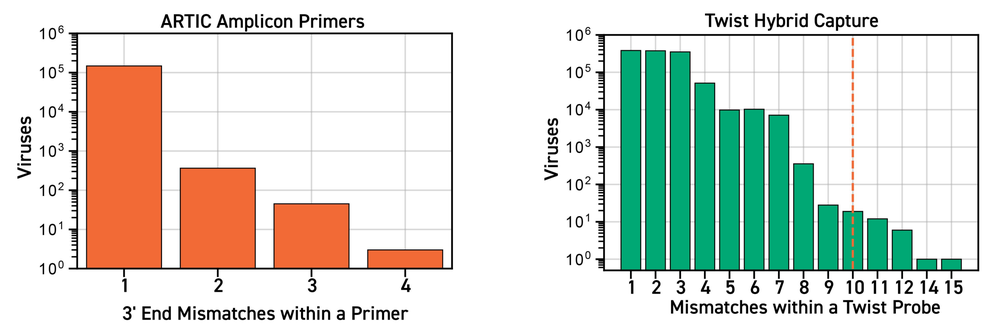


Figure 3.2: VOC Mismatches within a Single Amplicon Primer or Hybrid Capture Probe.

4. Candidate Dropouts and Results

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

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 4.2**). 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.

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.

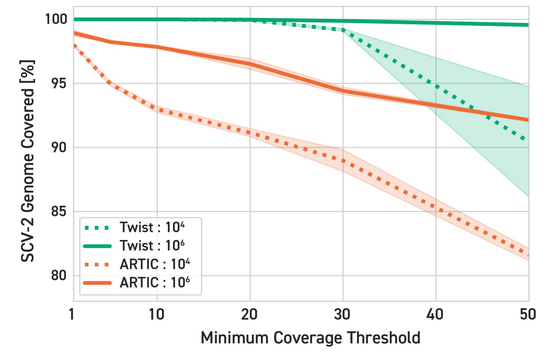


Figure 4.2: Twist's SARS-CoV-2 NGS Assay covers more of the SCV-2 genome than amplicon sequencing.

5. Clinical Data and Conclusions

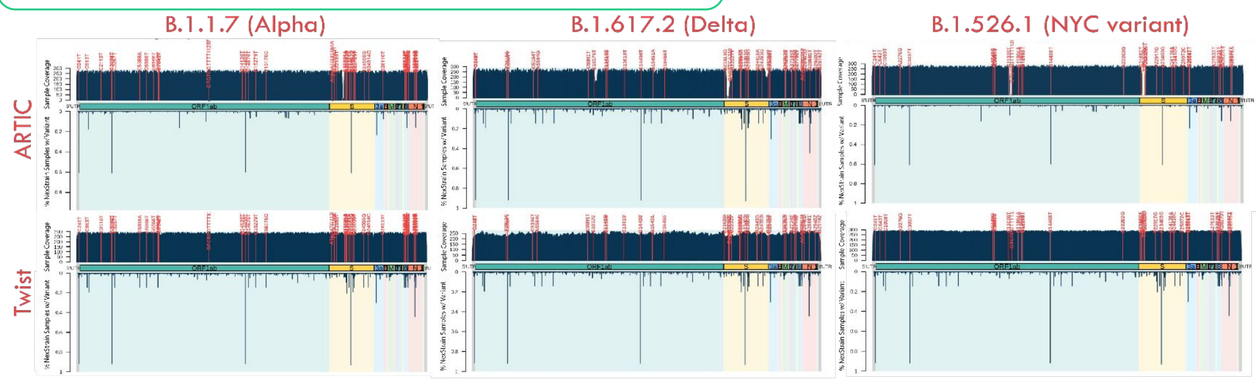


Figure 5.1: Comparison of percent genome coverage at 5x depth between Twist hybrid capture and ARTIC workflow on clinical specimens with different variant profiles. Twist's SARS-CoV-2 NGS Assay showed uniformity of coverage (>99%) across the viral genome, while the ARTIC workflow showed dropouts on clinical samples with detected SCV-2 from the Alpha (B.1.1.7), Delta (B.1.617.2) and NYC (B.1.526.1) lineages. The percent of the SCV-2 genome recovered and genetic variants identified compared to the reference genome are indicated (top) with high callable coverage (blue) shown. The proportion of known genetic variants of SCV-2 strains as reported in NextStrain as of 03-07-2021 from across the world are shown (bottom). Variants are only shown in genomic regions that have sufficient coverage.

Nasopharyngeal specimens that tested positive with FDA EUA authorized RT-PCR tests in reference clinical laboratories were collected in New York (Spring 2020 - Spring 2021, under Advarra IRB, Pro00042824). Specimens were processed using the SARS-CoV-2 NGS Assay (Twist), ARTIC SARS-CoV-2 FS Library Prep Kit (New England Biolabs), sequenced with Illumina NextSeq 550 and profiled the viral genome and identified mutations with the COVID-DX software (Biotia).

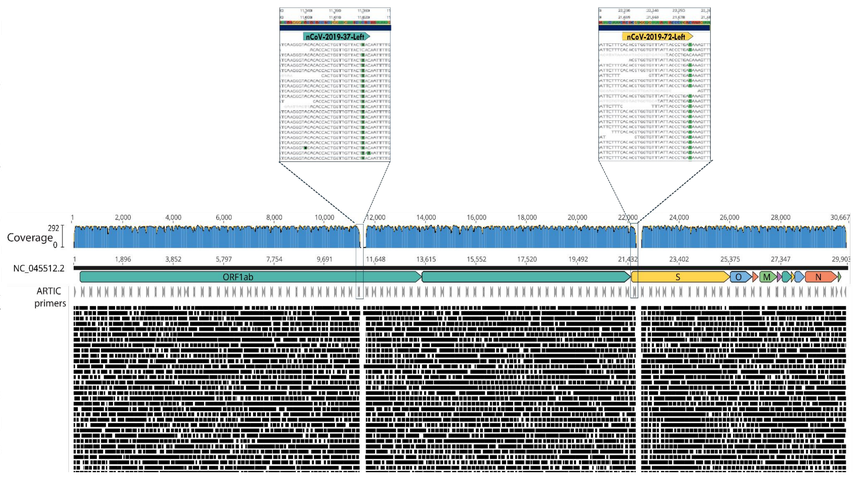
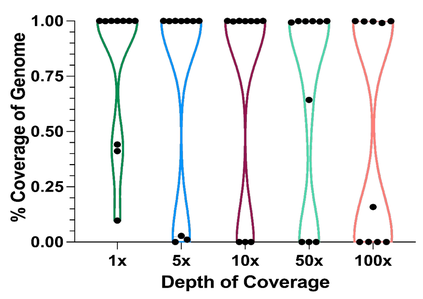


Figure 5.2: A clinical specimen (B.1.526.1 lineage) displaying ARTIC primer dropouts at the Orf1ab and S gene locations leading to insufficient genome coverage thus causing missing a SNP.



CT Value	Presence/Absence		% Percent coverage of genome at 10x depth	
	1x	0.2x	1x	0.2x
13.189	Positive	Positive	1	1
14.03	Positive	Positive	1	1
21.769	Positive	Positive	1	1
21.983	Positive	Positive	0.99	0.99
26.916	Positive	Positive	0.99	0.99
27.599	Positive	Positive	0.99	0.99
28.657	Positive	Positive	0.99	0.99
31.109	Negative	Negative	0	0
33.475	Positive	Negative	0.43	0
33.474	Positive	Negative	0.82	0

Figure 5.3: The miniaturized SARS-CoV-2 NGS assay workflow was optimized on clinical samples (n=10). The SARS-CoV-2 NGS Assay with decreased library preparation volume (0.2x from 2.5x) provided sufficient coverage across the viral genome at different depths on clinical specimens. The table shows that reduced volumes of 0.2x and 1x yield ample coverage to call presence/absence and genetic variants on samples with a CT value below 30.

The miniaturized hybrid capture workflow can be adapted to support high-throughput sequencing, SARS-CoV-2 detection and surveillance with decreased cost and automation.

Financial Disclosures: All authors are employees and/or shareholders of Twist Bioscience or Biotia.