Long-amplicon variant-robust genome capture of SARS-CoV-2 using molecular inversion probes

Eric D Boyden, Arjun D Patel, Joseph M Vieira,
Jack M Amaral, Gregory J Porreca, Patrick C Saunders

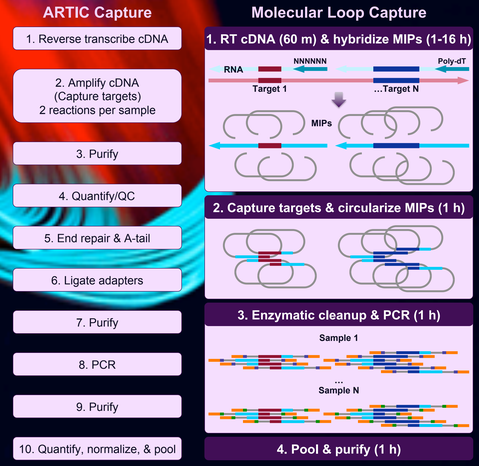
Molecular Loop Biosciences, Inc.
info@molecularloop.com

Background

Combating pandemics such as Covid-19 requires routine and systematic complete viral sequencing of positive samples to identify and track emergent variant strains. The PCR-based ARTIC assay is effective but prone to amplification dropouts when new mutations arise, which has required multiple primer revisions. We previously presented an alternative variant-robust short-read sequencing assay that uses tiled molecular inversion probes (MIPs) for high redundancy (7.5X mean tiling depth), which improves coverage uniformity and insures against amplification and mutation dropouts, thereby facilitating superior genome completion rates across a broad range of Ct values. Our highly scalable and easily automatable chemistry requires less than 2 hours of manual labor across 4 addition-only steps including a combined RT + hybridization step, and may be performed using either a 1-day or overnight workflow (Figure 1).

However, short-read NGS assays may be unable to phase distant variants and show reduced sensitivity to modestly sized indels. Sequencers by Pacific Biosciences and Oxford Nanopore Technologies can generate reads that are several hundred to thousands of bases long, with more power to phase variants and identify larger indels (e.g. ≤0.5X read length), but capitalizing on these advantages requires sufficiently long library molecules, which is not a common feature of published MIP assays.

Figure 1. Capture workflow comparison



Methods & Results

To establish that our technology can produce templates that are optimal for long read sequencing, we designed panels of MIPs to capture 99.5% of the SARS-CoV-2 genome (Wuhan-Hu-1) in overlapping elements of either 675 or 1215 bp, with average tiling depths of 22.5X or 40.5X respectively (Figure 2). We then evaluated the performance of the SC2.225 and SC2.675 probesets on synthetic RNA from Wuhan and BA.1 strains using either a 1-day or overnight workflow (Figure 3). The SC2.1215 amplicons were not compatible with Illumina sequencing.

All probesets produced sharp amplicons of the predicted sizes (Figure 4). Nearly all reads aligned (Figure 5) and were on target (Figure 6). The insert sizes of aligned read pairs were consistent with the expected sizes (Figure 7). Mean coverage was similar for both samples (Figure 8), with ≥96% of the genome and ≥99% of capturable BA.1 mutations covered to a depth of ≥5X (Figures 9 & 10). *Since only ~1/3 of each 675 bp amplicon was sequenced, 3X higher coverage would be expected with long reads.*

This work represents a significant achievement in the expansion of MIP capture to new applications. We have shown that MIPs can capture much larger elements than those for which they are commonly designed, which will improve their utility for long-read sequencing platforms.

This study was funded by Molecular Loop Biosciences.

Figure 3. Experimental design & analysis

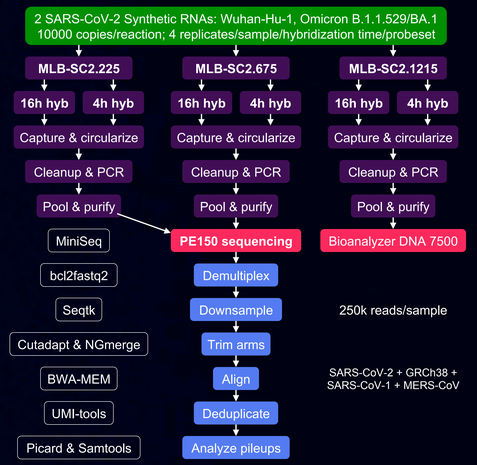
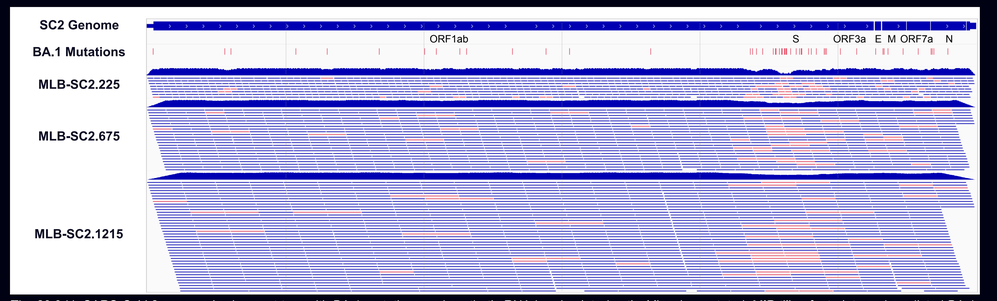
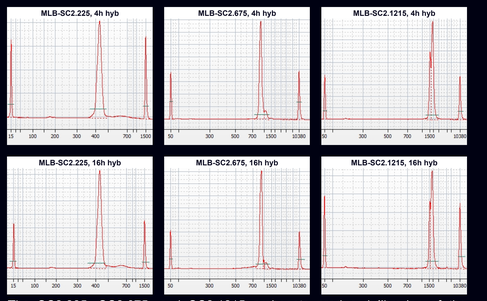


Figure 2. SARS-CoV-2 Omicron BA.1 genome schematic and MIP tiling footprints



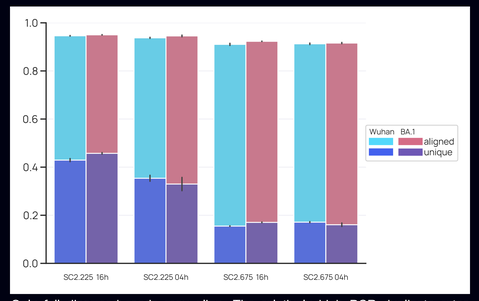
The 29.9 kb SARS-CoV-2 genome is shown at top, with BA.1 mutations and synthetic RNA breakpoints (vertical lines) annotated. MIP tiling footprints and predicted BA.1 coverage profiles are shown below; red MIPs are potentially affected by BA.1 mutations, blue MIPs are not. MIPs that overlap breakpoints will not capture synthetic RNA.

Figure 4. Library electropherograms



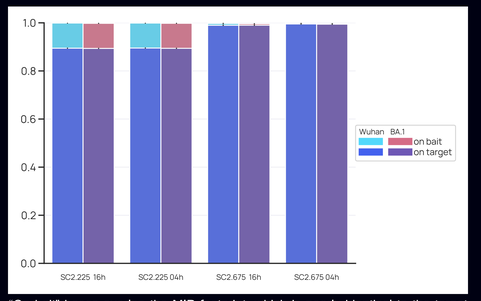
The SC2.225, SC2.675, and SC2.1215 probesets produced libraries of the expected size (413 bp, 864 bp, and 1404 bp respectively) on Bioanalyzer traces.

Figure 5. Fraction of reads aligned/unique



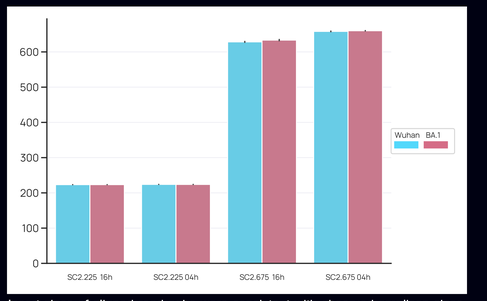
Only full disarmed reads may align. The relatively high PCR duplicate rates suggest that <250k reads/sample is required to maximize sequencing efficiency.

Figure 6. Fraction of bases on bait/target



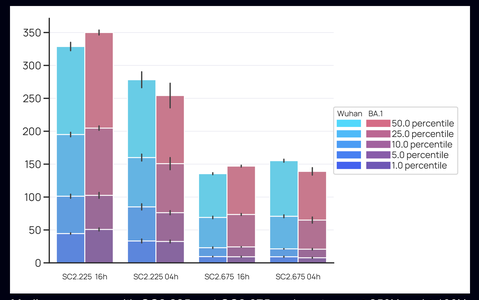
"On bait" bases overlap the MIP footprint, which is nearly identical to the target footprint. However overlapping portions of read pairs are considered "off target".

Figure 7. Mean read pair insert size



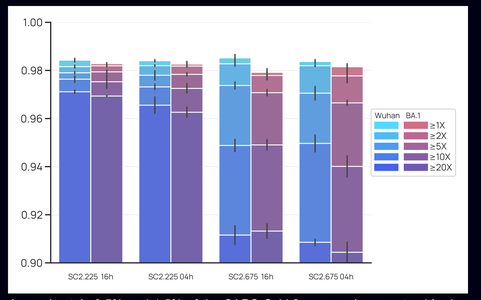
Insert sizes of aligned read pairs were consistent with observed amplicon sizes. A small fraction of shorter amplicons reduced the SC2.675-16h mean insert size.

Figure 8. Percentile target coverage



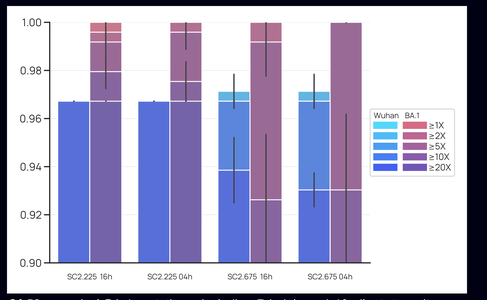
Median coverage with SC2.225 and SC2.675 probesets was ≥250X and ≥100X, respectively. SC2.675 coverage would be ~3X higher with long read sequencing.

Figure 9. Fraction of genome ≥X coverage



Approximately 0.5% and 1.5% of the SARS-CoV-2 genome is uncapturable due to proximity to genome termini and synthetic RNA breakpoints, respectively.

Figure 10. Fraction of BA.1 mutations ≥X



Of 59 canonical BA.1 mutations, including 7 indels and 10 silent, two sites are uncapturable in the synthetic Wuhan RNA due to their proximity to breakpoints.