TECHNICAL NOTE

Sequencing guidelines for COVID-19 surveillance using the Illumina COVIDSeq™ Test (RUO Version)

Understand the impact of genome coverage on SARS-CoV-2 variant analysis

illumina®

The need for COVID-19 surveillance

The COVID-19 pandemic has raged across the world for more than a year.[1] The emergence of new SARS-CoV-2 variants that are potentially more contagious or deadly have raised concerns about public health efforts, certain diagnostic tests, and vaccines developed to combat the pandemic.[2] This highlights the need for genomic surveillance to identify and monitor new SARS-CoV-2 variants. To address this need, we offer the Illumina COVIDSeq Test (RUO Version). This amplicon- based next-generation sequencing (NGS) assay includes primers designed to detect mutations and characterize RNA from the SARS-CoV-2 virus to help clinical research labs identify and track the emergence and prevalence of new SARS-CoV-2 variants and lineages.

When using the Illumina COVIDSeq Test for viral surveillance, it is important to understand the impact genome coverage has on variant analysis. In NGS, genome coverage is determined by the read length and depth. When setting up a sequencing run, laboratories need to balance the read lengths and read depths sufficiently to achieve comprehensive viral genome coverage with investments in sequencing costs and analysis time. This technical note evaluates the impact of various read lengths and depths on SARS-CoV-2 genome coverage, and provides guidelines for laboratories adopting the Illumina COVIDSeq Test (RUO Version) for surveillance.

Experimental design

Sequencing parameters

Sixteen COVID-positive samples (Ct values ≤ 30) were sequenced at a read length of 2 × 151 bp on the NextSeq™ 550Dx instrument (RUO Mode). To evaluate the effect of read length on genome coverage, reads were trimmed to various lengths:

- 1 × 36 bp
- 2 × 51 bp
- 2 × 101 bp
- 2 × 36 bp
- 2 × 76 bp
- 2 × 151 bp

To evaluate the effect of read depth on genome coverage, each read length was downsampled to various depths:

- 10M reads
- 2.5M reads
- 1M reads
- 0.25M reads
- 5M reads
- 2M reads
- 0.5M reads
- 0.1M reads

Analysis with DRAGEN™ COVID Lineage App

FASTQ sequencing files from the NextSeq 550Dx instrument were input to the Illumina DRAGEN COVID Lineage App for alignment to a SARS-CoV-2 reference genome. Starting from FASTQ files, the app performs mapping/alignment, variant calling, and consensus sequence generation. Access the software in BaseSpace™ Sequence Hub.

Results

Impact of read length and depth on SARS-CoV-2 genome coverage

Sequencing data from the 16 COVID-positive samples across a range of read lengths and depths were evaluated for performance. For this technical note, performance is defined as the portion of a consensus genome that exceeds 10× coverage. Plotting performance as a function of read length and depth shows significant improvement with paired-end reads over single reads. Paired-end reads substantially increase the number of unique reads that pass duplicate removal, resulting in better performance at low read length (Figure 1). When considering paired-end sequencing, comprehensive coverage is achieved at read lengths of 2 × 51 bp to 2 × 76 bp, with read depths of 1-2M reads per sample (Figure 1 and Figure 2).

Shorter read lengths and shallower read depths are more cost-effective and efficient in terms of bases sequenced and analysis time (Figure 3). However, it is important to note that the sample set used in this technical note is small and includes ideal test samples (Ct values ≤ 30). For analysis of real-world samples that potentially have low viral titers, laboratories may choose to increase read length or depth to ensure comprehensive genome coverage. Deletions of increasing size have been observed in newly identified SARS-CoV-2 variants.[3-5] Sequencing reads lengths of 2 × 76 bp or shorter may not detect larger deletions, so longer read lengths may be preferred to ensure full coverage and characterization of SARS-CoV-2 variants for COVID-19 surveillance.

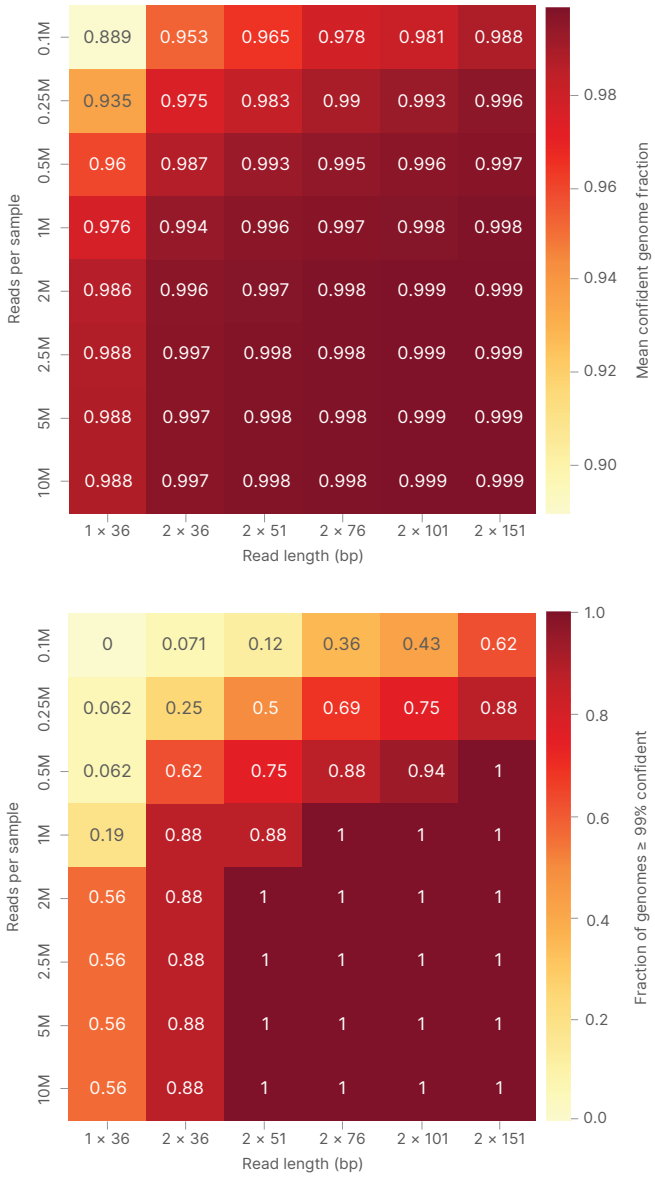

Figure 1: Genomic coverage performance—Mean confident genome fraction (top) and fraction of genomes > 0.99 confident (bottom), two metrics of genomic coverage performance, are plotted in heat maps as a function of read length and depth. Values close to 1 represent more complete coverage.

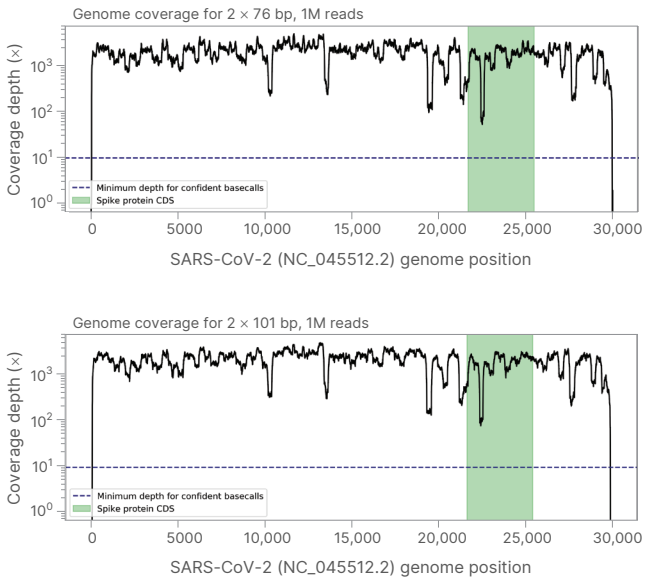

Figure 2: SARS-CoV-2 genome coverage—Demonstrated genome coverage for SARS-CoV-2 generated at read lengths of 2 × 76 bp (top) and 2 × 101 bp (bottom) at a read depth of 1M reads. Coverage is shown with all reads above the minimum depth for confident base calls (dashed blue line). The region corresponding to the spike (S) protein is shaded in green.

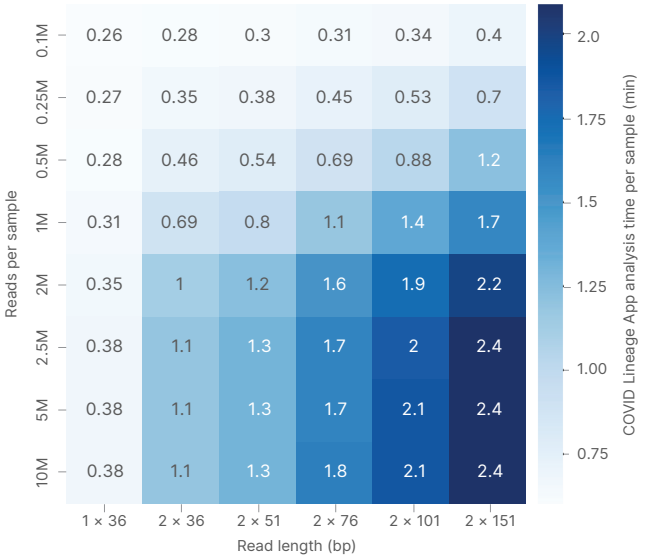

Figure 3: Sequencing considerations—Analysis time is plotted in a heat map as a function of read length and depth. Times represent total analysis time (detection, mapping/alignment, and variant calling) per sample using the DRAGEN COVID Lineage App 3.5.1.

Summary

The emergence and spread of new SARS-CoV-2 variants during the COVID-19 pandemic highlights the need for sequencing-based viral surveillance. The Illumina COVIDSeq Test (RUO Version) is an ideal sequencing solution for confirming SARS-CoV-2 variants and lineages. The data presented in this technical note shows that ≥ 99% genome coverage was achieved for all samples analyzed using read lengths of 2 × 76 bp and longer and read depths of 1M reads per sample and higher. Minor coverage improvements are observed with increases in read length and depth. However, viral titer and quality of extracted RNA may impact the required read configuration and depth to obtain optimal consensus genome calling. Laboratories should balance analysis time, cost, and coverage requirements.

Learn more

Illumina COVIDSeq Test, illumina.com/products/by-type/clinical-research-products/covidseq.html

References

1. World Health Organization. WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV). 30 January 2020.
2. Baric, RS. Emergence of a highly fit SARS-CoV-2 variant. *N Engl J Med*. 2020;383:2684–2686.
3. McCarthy KR, Rennick LJ, Nambulli S, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science*. 2021; doi:10.1126/science.abf6950.
4. Addetia A, Xie H, Roychoudhury P, et al. Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates. *J Clin Virol*. 2020; 129:104523.
5. Rosenthal SH, Kagan RM, Gerasimova A, et al. Identification of eight SARS-CoV-2 ORF7a deletion variants in 2,726 clinical specimens. *bioRxiv*. 2020; doi.org/10.1101/2020.12.10.418855.

Glossary

Genome coverage–The percentage of target bases for a particular genome that are sequenced a given number of times.

Read length–The number of base pairs sequenced from a DNA fragment. Read length can vary depending on sample type, application, and coverage requirements.

Read depth–The number of times a base is read in a sequencing run (ie, the number of reads per base).

FASTQ–A text file that contains the sequence data and corresponding quality scores. Typically, this file type is input for various secondary analysis software applications.

Consensus sequence–A DNA sequence resulting from alignment that represents the most abundant nucleotide at each position.

Paired-end sequencing–The process of sequencing a DNA fragment from both ends in the same run, enabling more accurate alignment and detection of small structural variants.

Mean confident genome fraction–The mean fraction of the genome which has a coverage level greater than 10× coverage.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-00088 v1.0