Development of a High-Throughput NGS Workflow for SARS-CoV-2 Whole-Genome Sequencing

00821

Sun Hee Rosenthal,[a] Anna Gerasimova,[a] Rolando Ruiz-Vega,[a] Kayla Livingston,[a] **Ron M Kagan**, Yan Liu, Ben Anderson, Renius Owen, Laurence Bernstein, Alla Smolgovsky, Dong Xu, Rebecca Chen, Andrew Grupe, Pranoot Tanpaiboon, Felicitas Lacbawan

Quest Diagnostics, San Juan Capistrano, CA, USA; [a]These authors have contributed equally to this work.

Introduction

- Monitoring for new mutations in SARS-CoV-2 is crucial for identifying diagnostic and therapeutic targets and provides important insights to achieve a more effective COVID-19 control strategy.
- Next-generation sequencing (NGS) has been widely used for whole-genome sequencing of SARS-CoV-2.
- However, the scalability of NGS methods may be limited by the complexity of the workflow.
- Here, we address this limitation by designing a workflow optimized for high-throughput performance.

Methods

- Our NGS workflow is similar to a previously reported tailed PCR amplicon approach.[1]
- Libraries were prepared by a 2-step PCR method, but optimized to improve amplicon balance, integrate robotic liquid handlers, and minimize amplicon dropout for viral genomes harboring primer-binding site mutation(s).
  - Modified ARTIC network v3 primers were used for SARS-CoV-2 whole-genome amplification.
  - A modified touchdown PCR method was used, which gradually reduced the annealing temperature from 65℃ to 55℃ (0.7℃/second) within each cycle.
- Sequencing was performed on the Illumina NovaSeq 6000 and the Illumina MiSeq, and bioinformatic analysis was performed using an in-house pipeline.[2]

Results

- Compared to conventional PCR using a constant annealing temperature, the use of touchdown PCR in the optimized workflow reduced the proportion of specimens with amplicon dropout for most SARS-CoV-2 lineages amplified with the ARTIC v3 primers (**Table 1**).
- Compared to the original ARTIC v3 workflow, coverage of the SARS-CoV-2 genome was reduced by about 30% in the high-throughput workflow (**Figure 1**) but was still adequate to generate complete SARS-CoV-2 genome consensus sequences (**Table 2**).
- Modifying primer balance and pool assignments optimized coverage for relatively low-performing amplicons (26, 29, 67, 68, 74, 75, and 79) (**Figure 1**).
- Among the 186 specimens sequenced to determine the limit of detection of the optimized workflow, 91% to 100% with cycle threshold (Ct) values <28 generated ≥97% consensus sequence (**Figure 2A**), and 100% with Ct=28 and 88% with Ct=29 generated >90% consensus sequence (**Figure 2B**).
- Complete consensus sequences were generated from 96% of specimens using the optimized high-throughput workflow compared to only 66% using the original ARTIC v3 workflow (**Table 2**).

Results (continued)

Table 1. Amplicon Dropout by PCR Method and Clade

Clade[a]	Specimens tested, n	Amplicon dropout, %, by annealing temperature	
		Conventional (65℃)	Touchdown (65℃-55℃)
20A	12	0.34	0
20B	10	0.51	0
20C	3	1.7	0.34
20G	24	0.21	0
20H (Beta)	1	2.04	0
20I (Alpha)	9	0	0
21C (Epsilon)	17	1.08	0
21F (Iota)	3	0	0
Total	79	0.5	0.01

[a] SARS-CoV-2 clades were assigned with Nextclade[3] v1.3.0.

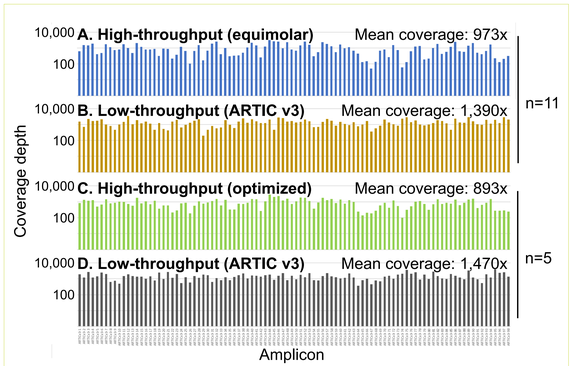

Figure 1. SARS-CoV-2 genome coverage achieved by the high-throughput workflow (A, equimolar primer pools; C, optimized for low-performing amplicons) and low-throughput (B, D) workflows on matched clinical samples with Ct=24. Coverage was computed at a normalized depth of 200,000 mapped reads.

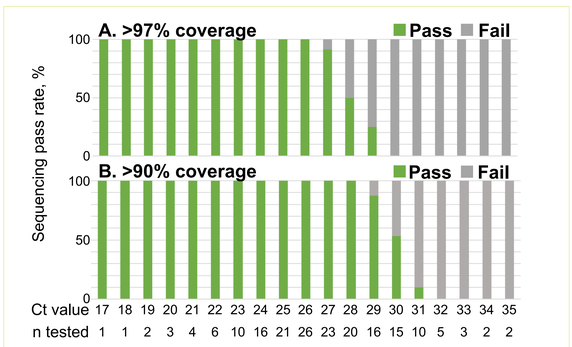

Figure 2. Sequencing pass rate (A, >97% consensus sequence coverage; B, >90%) by cycle threshold (Ct) value.

Results (continued)

Table 2. Generation of Complete Consensus Sequences

Clade[a]	Specimens with complete[b] consensus sequences generated, % (n), by workflow	
	High-throughput[c]	ARTIC v3[d]
19B	100 (3/3)	0 (0/3)
20A	83.0 (44/53)	62.2 (33/53)
20B	96.2 (102/106)	59.4 (63/106)
20C	88.7 (87/98)	47.9 (47/98)
20G	96 (168/175)	81.1 (142/175)
20H (Beta)	100 (7/7)	0 (0/7)
20I (Alpha)	96.9 (691/713)	78.2 (558/713)
20J (Gamma)	100 (44/44)	45.4 (20/44)
21A (Delta)	100 (1/1)	0 (0/1)
21C (Epsilon)	98.0 (99/101)	1.0 (1/101)
21D (Eta)	100 (10/10)	100 (10/10)
21F (Iota)	97.7 (128/131)	61.8 (81/131)
21G (Lambda)	100 (1/1)	0 (0/1)
21H	100 (3/3)	66.6 (2/3)
Total	96.0 (1,388/1,446)	66.1 (957/1,446)

[a] SARS-CoV-2 clades were assigned with Nextclade[3] v1.3.0.
[b] Defined by obtaining ≥97% SARS-CoV-2 genome coverage (100% coverage within the region of interest).
[c] Illumina NovaSeq 6000 with the SP reagent kit using 2 x 251 cycles.
[d] Illumina MiSeq with the 600 cycle v3 kit using 2 x 251 cycles.

- As part of the CDC National SARS-CoV-2 surveillance project (NS3), the high-throughput workflow was used to track the proportions of variants being monitored (Alpha, Gamma, Iota, Epsilon, Lambda, and others) and variants of concern (Delta) on a weekly basis from January through September 2021 (**Figure 3**).
- The Alpha and Iota variants became more prevalent through May, when the Delta variant emerged and rapidly increased in prevalence to >99% of all specimens sequenced in mid-September (**Figure 3**), consistent with national variant trends reported by the CDC.[4]

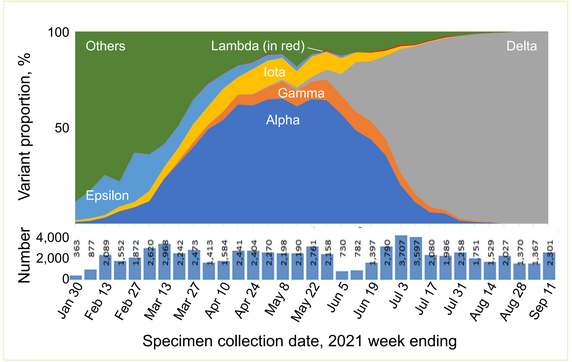

Figure 3. Weekly proportions of SARS-CoV-2 variants calculated against the total number of specimens sequenced each week.

Results (continued)

- Intra-assay precision was assessed for 3 replicates using 188 clinical specimens; 175/188 (93.1%) specimens met quality control requirements for 2 or more replicates and all lineage and clade assignments were concordant.
- Inter-assay precision was assessed for 168 clinical specimens over 3 runs; 160/168 (95.2%) specimens met quality control requirements for 2 or more replicates and all lineage and clade assignments were concordant.

Summary of Key Findings

- We present an optimized WGS workflow for SARS-CoV-2 that can process up to 2,688 samples in a single NovaSeq 6000 run without compromising sensitivity or robustness, and with fewer amplicon dropout events compared to the standard ARTIC protocol.
- Because a 2-step PCR protocol may increase primer-dimer formation resulting in lower sequencing quality and/or lower coverage,[1] we implemented 3 cleanup steps, including bead cleanups and automated agarose gel size selection, to obtain high-quality sequences with only a 30% reduction in coverage. Nevertheless, the coverage still far exceeded the minimum necessary to generate high-quality SARS-CoV-2 genome consensus sequences.
- Implementing touchdown PCR for library amplification reduced amplicon dropouts compared to the original ARTIC v3 protocol. These techniques are also applicable to the newly released ARTIC v4.1 primer sets with improved coverage of Delta and Omicron lineages (data not shown).
- We also demonstrated the utility of the methods described here for high-throughput viral genome surveillance encompassing over 65,000 clinical samples in 2021. Changes in prevalence of SARS-CoV-2 lineages corresponded to published data.[4]

References

1. Gohl DM, Garbe J, Grady P, et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genomics*. 2020;21(1):863. doi:10.1186/s12864-020-07283-6
2. Rosenthal SH, Gerasimova A, Ruiz-Vega R, et al. Development and validation of a high-throughput next-generation sequencing workflow for SARS-CoV-2 whole genome sequencing: results from over 65,000 clinical cases. *In Review*. Preprint posted online November 12, 2021. doi:10.21203/rs.3.rs-997210/v1
3. Nextclade. https://clades.nextstrain.org/
4. Variant Proportions. Centers for Disease Control and Prevention COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker/#variant-proportions