ANALYZING THE SARS-CoV-2 GENOME WITH TARGET ENRICHED RNA-Seq

Sarah Trusiak[1], Jonathan Nowacki[1], Ranjit Kumar[1], Spencer Debenport[1], and Rachel Kasinskas[1]
[1]Roche Sequencing Solutions. Wilmington, MA

INTRODUCTION

Since the emergence of the SARS-CoV-2 and its resulting disease, COVID-19, in late 2019, over 100 million people have been infected worldwide and over 2 million people have died. RNA sequencing (RNA-seq) enables research into the host's transcriptional response to viral infection and the RNA genome of the virus itself. Samples from infected hosts typically contain a single SARS-CoV-2 strain, while environmental samples like soil and sewage often contain a mixture of strains. When studying the RNA genome of viruses in mixed-RNA samples, such as total RNA from a human host or environmental sample, it is necessary to enrich for the viral RNA molecules above the much-more-abundant background RNA. RNA-seq with hybridization-based target enrichment (TE) is a powerful technology that enables high-throughput genome studies of SARS-CoV-2, which are critical for discovering new viral mutations as well as tracking viral transmission and the emergence of new strains.

To enable targeted sequencing of the SARS-CoV-2 genome from complex samples, we have developed a KAPA SARS-CoV-2 TE panel and workflow. Our probe panel covers 100% of the reference genome and >99.7% of 184 publicly available SARS-CoV-2 sequences (NCBI). Using this panel, a new target-enriched RNA-seq workflow was developed that incorporates the KAPA SARS-CoV-2 TE panel into a modified version of the HyperCap Workflow v3 that includes the KAPA RNA HyperPrep Kit. Performance of this panel and workflow was then tested using samples prepared with varying viral copy numbers of different SARS-CoV-2 strains and human RNA background to mimic the diversity found in real clinical samples.

EXPERIMENTAL DESIGN

Samples:

RNA samples were made by mixing 20 ng or 100 ng of Universal Human Reference RNA (UHR) (Agilent) with either 0, 10, 1000, 10,000, or 1,000,000 copies of Twist Synthetic SARS-CoV-2 RNA controls #1-6 (Twist Biosciences). The six SARS-CoV-2 RNA control strains were mixed in equal ratios prior to addition into UHR.

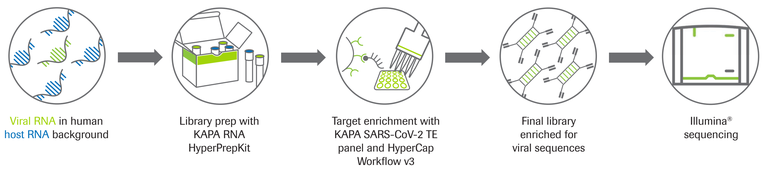

Figure 1: The KAPA SARS-CoV-2 TE panel with the KAPA RNA HyperPrep Kit and HyperCap Workflow v3 enables high-throughput targeted sequencing of the SARS-CoV-2 genome.

Workflow:

All RNA samples were then processed into sequencing libraries using the KAPA RNA HyperPrep kit. The following modifications to the KAPA RNA HyperPrep Kit protocol (KR1350 v2.17) were necessary to prepare the RNA libraries for downstream TE capture:

- In the adapter ligation reaction, 5 uL of 15 uM KAPA Universal Adapter was used in place of the KAPA Unique Dual-Indexed Adapters recommended in the original protocol.
- Post ligation, the 0.63x KAPA Pure Bead purification was omitted and only the second purification of 0.7x was performed.
- Libraries were amplified with KAPA HiFi HotStart ReadyMix and KAPA UDI Primer Mixes as per Chapter 4 of the KAPA HyperCap v3 workflow.
- To ensure that sufficient amplified library was available for target capture, PCR cycling conditions were adjusted as follows: for libraries with 20 ng of UHR input RNA,15 cycles; for libraries with 100 ng of input UHR RNA, 13 cycles.

Amplified libraries were purified following the KAPA HyperCap Workflow v3 with KAPA HyperPure beads. 1000 ng of each library underwent singleplex target capture, washing, and recovery using the KAPA HyperCapture Reagent Kit and KAPA HyperCapture Bead Kit as per the "<40 Mbp Capture Target Size" protocol in Chapter 5 of the KAPA HyperCap Workflow v3. Both 20 ng and 100 ng UHR libraries were hybridized to the KAPA SARS-CoV-2 TE panel (KAPA HyperExplore MAX 0.5Mb T2 #1000004753) overnight for at least 16 hours. An additional 1000 ng of each 20 ng UHR viral titration library was hybridized to the KAPA SARS-CoV-2 TE panel for only 1 hour to before washing and recovery to compare performance of the shortened time to overnight hybridization. Paired-end sequencing (2 x 75bp) was performed on the Illumina NextSeq 550.

TARGET-ENRICHED RNA-SEQ YIELDS 1X VIRAL GENOME COVERAGE IN SAMPLES CONTAINING 1,000 VIRAL COPIES IN 20 ng OR 100 ng OF HOST RNA

A

Human background	Viral copies	% Aligned reads	Mean target coverage	% Bases covered at 1X	% Bases covered at 10X	% Bases covered at 20X	Fold enrichment
20 ng RNA	10	0.05%	2.2	7.5%	5.7%	4.2%	50,600x
	1,000	4.75%	178.6	99.5%	98.8%	97.0%	47,529x
	10,000	46.56%	1795.4	100.0%	99.9%	99.9%	4,656x
	1,000,000	94.35%	3653.5	100.0%	99.9%	99.9%	943x
100 ng RNA	10	0.02%	1.1	7.9%	7.7%	4.2%	106,000x
	1,000	1.59%	58.6	97.6%	87.1%	71.3%	79,333x
	10,000	5.61%	207.2	100.0%	99.4%	95.6%	2,805x
	1,000,000	84.94%	3314.3	100.0%	99.9%	99.9%	4,247x

B

		% Bases covered at 1X				
Human background	Viral copies	25K reads	100K reads	250K reads	1M reads	8M reads
20 ng RNA	10	2.7%	5.4%	7.0%	7.4%	8.1%
	1,000	90.6%	98.1%	99.4%	99.5%	99.5%
	10,000	99.9%	99.9%	100.0%	100.0%	100.0%
	1,000,000	99.9%	99.9%	99.9%	100.0%	100.0%
100 ng RNA	10	1.2%	4.0%	6.4%	7.9%	9.1%
	1,000	57.8%	86.3%	95.1%	97.6%	98.1%
	10,000	85.0%	98.2%	99.8%	100.0%	100.0%
	1,000,000	99.9%	99.9%	100.0%	100.0%	100.0%

Table 1. SARS-CoV-2 genome alignment & coverage metrics from viral copy titration samples. Viral copy titration samples were processed through the KAPA SARS-CoV-2 TE and HyperCap Workflow v3 in triplicate. As expected, the "0 viral copy" samples had no valid aligned reads and are thus not included in the tables. **A.** Average alignment & coverage metrics (n=3) from viral copy titration datasets when downsampled to 1 million read pairs. **B.** Additional downsampling analysis of the datasets shows the fewest total sequencing reads required to achieve full 1X genome coverage (n=3) for each viral copy level.

SAVE VALUABLE TIME WITH HYBRIDIZATION AS SHORT AS 1 HOUR

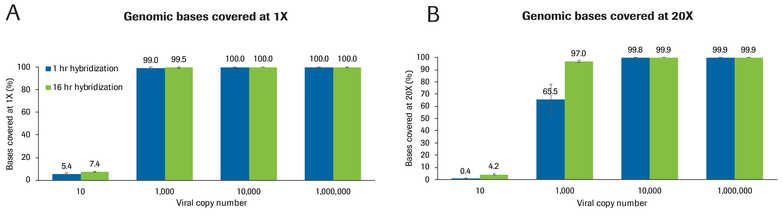

Figure 2. Shortening the hybridization time to 1 hour yields equivalent 1X SARS-CoV-2 genome coverage compared to 16-hour hybridization. Libraries generated from samples containing the indicated number of viral copies in a background of 20 ng of human RNA were hybridized to the SARS-CoV-2 TE panel for 1 hour or 16 hours and processed through the HyperCap v3 workflow in triplicate. **A.** The 1-hour hybridization time and 16-hour hybridization time yield similar SARS-CoV-2 genome coverage at 1X for all viral copy levels. **B.** Samples with higher viral loads also yield similar genome coverage at 20X for the 1-hour and 16-hour hybridization times, with somewhat reduced coverage for lower viral loads. 1 million Illumina NextSeq read pairs (2x75 bp).

MULTIPLE SARS-COV-2 STRAIN VARIANTS CAN BE DETECTED IN A SINGLE REACTION

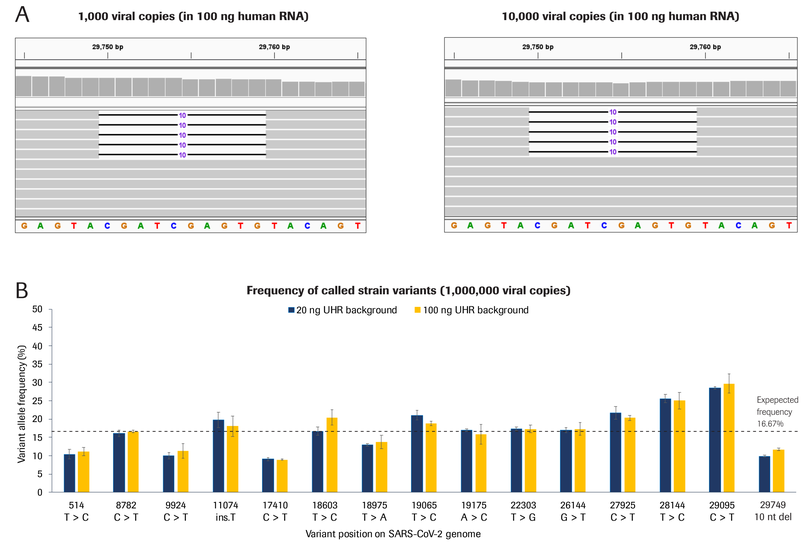

Figure 3. Detection of variants from six strains of SARS-CoV-2 within the same sample. Variant calling results for 100 ng human RNA samples with a total of one thousand (1,000), ten thousand (10,000), or one million (1,000,000) SARS-CoV-2 copies. **A.** IGV screenshots from representative 1,000 and 10,000 copy replicates show that even the most complex variant-a 10nt deletion at the 3' end of the SARS-CoV-2 genome-is detected in down to 1000 total viral copies (166 copies of strain with deletion). **B.** Alternate allele frequency analysis for the 1,000,000 viral copy datasets in 20 and 100ng human RNA background (n=3 each). All of the expected variants from the six mixed SARS-CoV-2 strain controls were detected at frequencies close to expected (dotted line 16.67%). Twist control strains (Accession ID) to variants are as follows: Strain 2 (MN908947.3) = reference strain; Strain 1 (MT007544.1) = variants at positions 19065, 22303, 26144, 29750; Strain 3 (LC528232) = variants at position 11074; Strain 4 (MT106054) = variants at positions 8782, 18603, 18975, 19175, 27925, 28144, 29095; Strain 5 (MT188340.1) = variants at positions 514, 17410; Strain 6 (MT118835.1) = variant at positions 9924.

CONCLUSION

The KAPA SARS-CoV-2 TE panel and target-enriched RNA-seq workflow enable high-throughput sequencing of the SARS-CoV-2 genome.

The results presented here demonstrate that:

- The SARS-CoV-2 genome can be covered at 1X (97%) when as few as 1,000 viral copies are present in 20 ng or 100 ng of human RNA background, from 1M Illumina NextSeq sequencer read pairs (2 x 75 bp).
- Some genomic sequence (7-9%) can be obtained from sample with as few as 10 total viral copies.
- One-hour hybridization offers a good balance between workflow speed and performance at all viral concentrations tested.
- Variants from six different SARS-CoV-2 strains were detected in the same sample near their expected frequency. A complex 10nt deletion in one strain was detected down to 1,000 total viral copies.

Disclaimer: Although the results of this study are promising, this KAPA RNA HyperPrep with HyperCap Workflow v3 protocol is still in development and has not been fully validated by Roche.