SARS-COV-2 TARGETED SEQUENCE ANALYSIS WITH SRSLY

INTRODUCTION

The sequencing of pathogen genomes is an integral part of epidemiology, especially during a global pandemic[1]. A pathogen's genome can dictate the severity of disease it causes, mode of infection, and can help researchers develop therapeutic strategies[1]. As genomic datasets grow, phylogenic analyses of genetic variation can help elucidate the geographic origins and transmission of the pathogen, estimate mutation rates, and help to predict the emergence of more virulent strains[2]. Such studies warrant efficient, fast, and affordable sequencing techniques that allow researchers to obtain the genomic sequences of viruses isolated from infected patients in real-time. An additional challenge is isolating and identifying pathogenic genetic material from biological samples, which contain large amounts of host genetic material.

Analyses of viral sequences is of significance today due to the global Coronavirus Disease 2019 (COVID-19) pandemic with about 2.5 million case in US alone as of June 2020[3]. This acute and highly morbid respiratory disease is caused by the novel SARS-CoV-2, a betacoronavirus which contains a ~30kb single-stranded, positive sense RNA genome[4]. To date, many variants of the SARS-CoV-2 genome have been identified globally[5-8]. Characterization of these mutations and their effect on virulence and host immunity is vital for disease management.

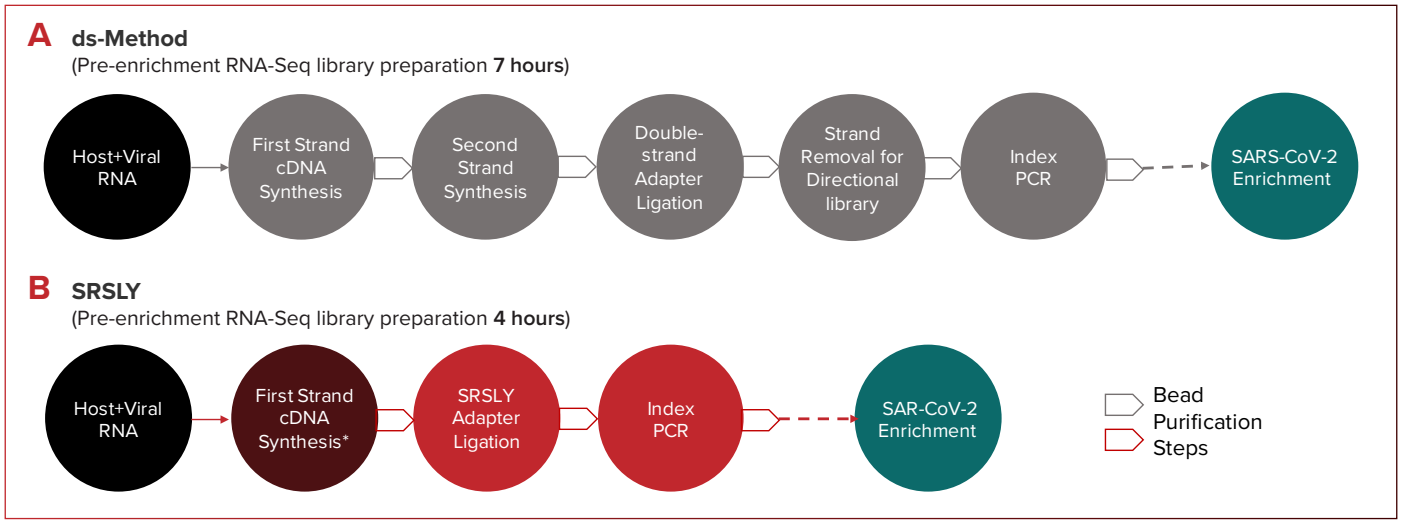


Figure 1. Schematic for RNA-Seq and viral enrichment approach. A. ds-Method - Traditional RNA-Seq library preparation workflow that involves more than 5 steps B. SRSLY - Claret Bioscience's novel approach for RNA-Seq that requires only three steps to generate directional libraries prior to target enrichment of viral sequences. Black – Input RNA. Grey – Traditional double-stranded RNA-Seq approach. Dark-Red - Any user-preferred first-strand synthesis method may be used. Red - SRSLY library preparation from first-strand cDNA. Teal – Any downstream target enrichment approach. Note: Some ds-Methods combine steps to reduce bead purification steps.

Genome sequence analysis of SARS-CoV-2 (and other RNA viruses) extracted from biological samples requires conversion of RNA to sequence-ready libraries, followed by targeted enrichment of the complete viral genome (outlined in Figure 1). RNA-Seq library preparation involves conversion of RNA to first-strand complementary DNA (fs-cDNA). In traditional RNA-Seq methods (Figure 1A, called as ds-Methods henceforth) this step is usually followed by second-strand synthesis, which is necessary only because in these methods ligation of sequencing adapters require a double-stranded template. This additional 2nd strand synthesis step compromises directionality information: additional modifications or labeling of one or both strands are required to identify the original sequence. Retaining directionality data is particularly critical

in SARS-CoV-2 sequence-based studies because the virus uses both genome-derived and antisense transcripts for protein translation[9]. Moreover, the 2nd strand synthesis step may result in template loss (during subsequent purification steps), introduce sequence errors, and, increase overall protocol time. While dealing with biological samples, which typically contain picogram scale of target sequences, these shortcomings are exacerbated. A better solution would eliminate the need for 2nd-strand synthesis altogether.

One such solution is SRSLY, a single-stranded library preparation approach developed by Claret Bioscience LLC., for use with a variety of DNA inputs, including single-stranded molecules[10]. In RNA-seq applications, SRSLY involves a simple workflow that generates sequence-ready molecules within 3 hours starting directly from first-strand cDNA thereby eliminating the prerequisite 2nd strand synthesis step (Figure 1B). When coupled with robust enrichment strategies, the target sequences of interest can be enriched and further amplified enabling downstream sequence-specific analyses. While other single-stranded methods exist, they either involve multiple time-consuming steps or may introduce additional nucleotides to facilitate adapter ligation. Even ds-Methods require the addition of a terminal dAMP i.e. dA-tailing prior to adapter ligation. These steps result in inaccurate viral sequences at read ends, unlike SRSLY, which does not require terminal base modifications and therefore retains native sequences - a critical aspect in the discovery of novel mutations and in phylogenetic studies.

Apart from an efficienct library preparation method, a robust downstream target-enrichment approach is critical for accurate capture and analysis of viral molecules. Arbor Biosciences' myBaits system is a hybridization-based approach that is simultaneously highly specific and highly sensitive. This system is appropriate for enriching viral sequences from RNA-Seq libraries prepared from fresh viral isolates, highly degraded trace RNA sources, or even samples with high levels of genetic diversity. The myBaits SARS-CoV-2 Expert panel used in this application note comprises over 2,000 biotinylated ssRNA probes designed using all full and partial SARS-CoV-2 genome sequences available in NCBI as of January 31, 2020.

Here, starting with cDNA generated from a complex pool of human and SAR-CoV2 RNA using the NEBNext® Ultra II First-strand synthesis module upstream and targeted viral-enrichment using the Arbor Biosciences myBaits® Expert SARS- CoV2 panel downstream, we demonstrate the seamless integration of the SRSLY library preparation method into a target-enrichment workflow. This modular workflow detects down to 10 viral copies and enables accurate identification of specific variants within the SARS-CoV-2 genome.

EXPERIMENTAL DESIGN AND RESULTS

Directional RNA-Seq library preparation with SRSLY

To mimic total RNA inputs expected from biological samples, 50ng of human total RNA (Agilent) was spiked with SARS-CoV-2 viral copies in a dilution series from 1 million to 10 copies. For each contrived RNA sample, RNA fragmentation and First-strand synthesis were performed using the NEBNext® Ultra II first-strand synthesis module. Following cDNA purification, the SRSLY™ NanoPlus kit was used to convert the single-stranded template to sequence-ready libraries. Unlike standard RNA-Seq protocols, rRNA depletion was not performed to avoid loss of complexity during the additional purification step following depletion and because we did not expect the rRNA-derived fragments to interfere with subsequent viral enrichment. Pre-enrichment library quality was evaluated by sequencing the libraries to a low depth (~1 million reads) on an Illumina® MiSeq. The sequences were equally down-sampled to a depth of 1 million reads and RNA-Seq metrics were obtained. As expected, > 95% of the pre-enrichment reads were from the correct strand, with a majority of the reads mapping to rRNA genes (Figure 2). Prior to enrichment, viral reads were detectable only in the library with 1 million viral copies at the available sequencing depth (data not shown).

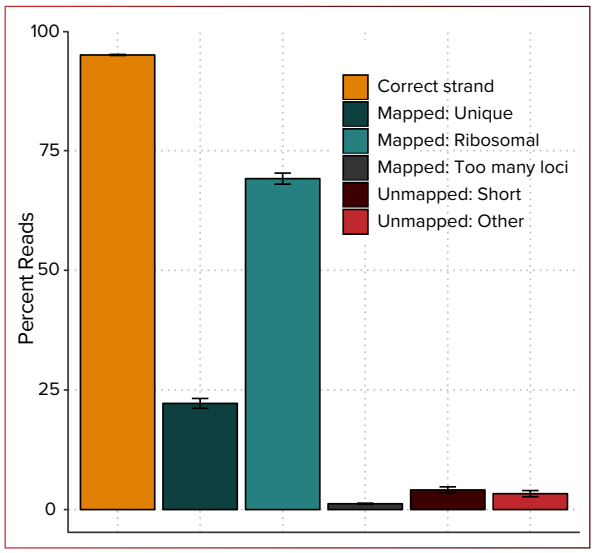
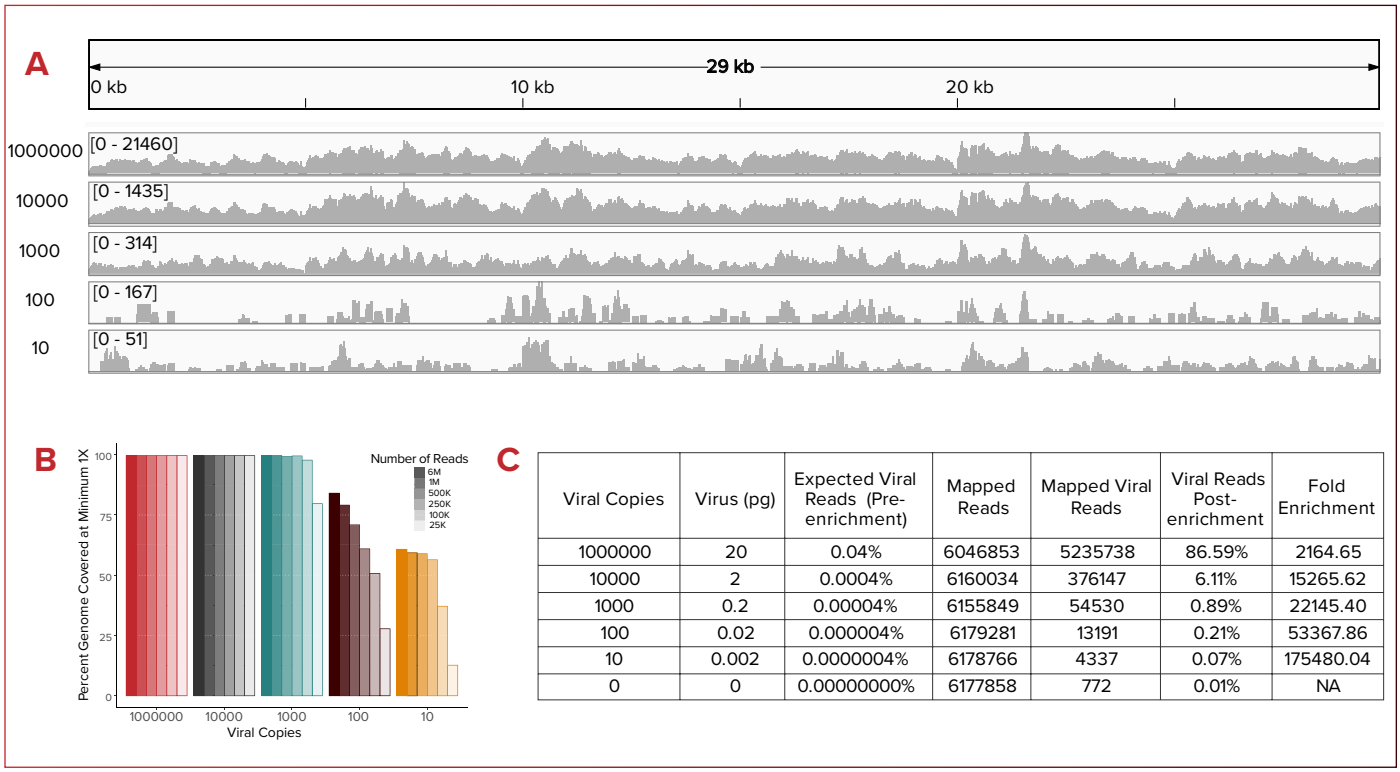


Figure 2. Average Mapping Metrics of Pre-enrichment Libraries. Reads were filtered and mapped to a combined genome of human and SARS-CoV-2 reference sequences using STAR aligner v2.7.3a.

SARS-CoV-2 enrichment by probe hybridization

To perform target-enrichment, we coupled the SRSLY libraries (500 ng each) with the the myBaits Expert Virus SARS-CoV-2 Panel (Cat. No. 308716). Probe-hybridization was performed for 24 hours to improve enrichment of viral sequences at lower titers. The enriched libraries were amplified using the KAPA Hifi HotStart Ready Mix PCR kit. The resultant libraries were deeply sequenced on an Illumina® HiSeq 4000. All reads were mapped to a combined human (hg19) and SARS-CoV-2 (MN908947.3) genome.

After enrichment, we observed high coverage of the entire viral genome in samples with as low as 1,000 viral copies, and partial genome coverage at 10 copies (Figure 3A). We also estimated the breadth of genomic coverage with reads down-sampled from 6 million down to 25,000 reads. We observed at least 1X full coverage of the complete genome in samples with as low as 1,000 viral copies at only 100,000 reads. This breadth of coverage reduces only at the lowest titers and sequencing depths. For example, at the lowest titer - 10 viral copies - we still observe >50% of the genome with 250,000 reads (Figure 3B). Overall, viral genome enrichment ranged from 2,165-fold to 175,480-fold (Figure 3C).



Viral Copies	Virus (pg)	Expected Viral Reads (Pre-enrichment)	Mapped Reads	Mapped Viral Reads	Viral Reads Post-enrichment	Fold Enrichment
1000000	20	0.04%	6046853	5235738	86.59%	2164.65
10000	2	0.0004%	6160034	376147	6.11%	15265.62
1000	0.2	0.00004%	6155849	54530	0.89%	22145.40
100	0.02	0.000004%	6179281	13191	0.21%	53367.86
10	0.002	0.0000004%	6178766	4337	0.07%	175480.04
0	0	0.00000000%	6177858	772	0.01%	NA

Figure 3. Post-enrichment viral genome coverage and mapping metrics. A. Viral genome coverage – Reads mapping to MN908947.3 Reference were visualized with IGVtools2.4.13. Coverage across the genome at a sequencing depth of 1 Million reads is shown. Numbers in parentheses indicate the range of coverage. Lower coverage was seen at lower titers B. Breadth of coverage of viral genome for each titration of viral copies at different sequencing depth was calculated. Legend – for each viral titer the darker shades indicate a higher sequencing depth going from 6 million read to 25 thousand reads. C. Mapping metrics and fold-enrichment of viral sequences for each titration of viral copies – metrics for reads down-sampled to ~6 Million are shown. Spurious mapping to SARS-CoV-2 genome was observed at low levels in the control, which contained only human RNA and no viral copies

Detection of viral variants from enriched SRSLY libraries

The synthetic viral genome used in this experiment was generated using a unique SARS-CoV-2 isolate (Reference genome: MT007544.1)[11]. When compared to the RefSeq SARS-CoV-2 genome MN908947.3, this genome harbors specific mutations - 3 single nucleotide substitutions and a 10 nt deletion. At a sequencing depth of 1 million reads, we detected all three SNPs even at viral loads of 1,000 copies. At lower titers, we detected the first and the third SNPs, however the second SNP was not detected due to uneven coverage in that region. The deletion is detected even at titers of 100 viral copies by this approach (Figure 4). Recent studies have shown that saliva and nasal swabs from patients at the onset of symptoms have about 1,000 to 10,000 copies of the virus[12], suggesting that the combination of SRSLY and a downstream target-enrichment protocol like the myBaits SARS-CoV-2 system may facilitate the identification of novel viral variants even at the time of diagnosis.

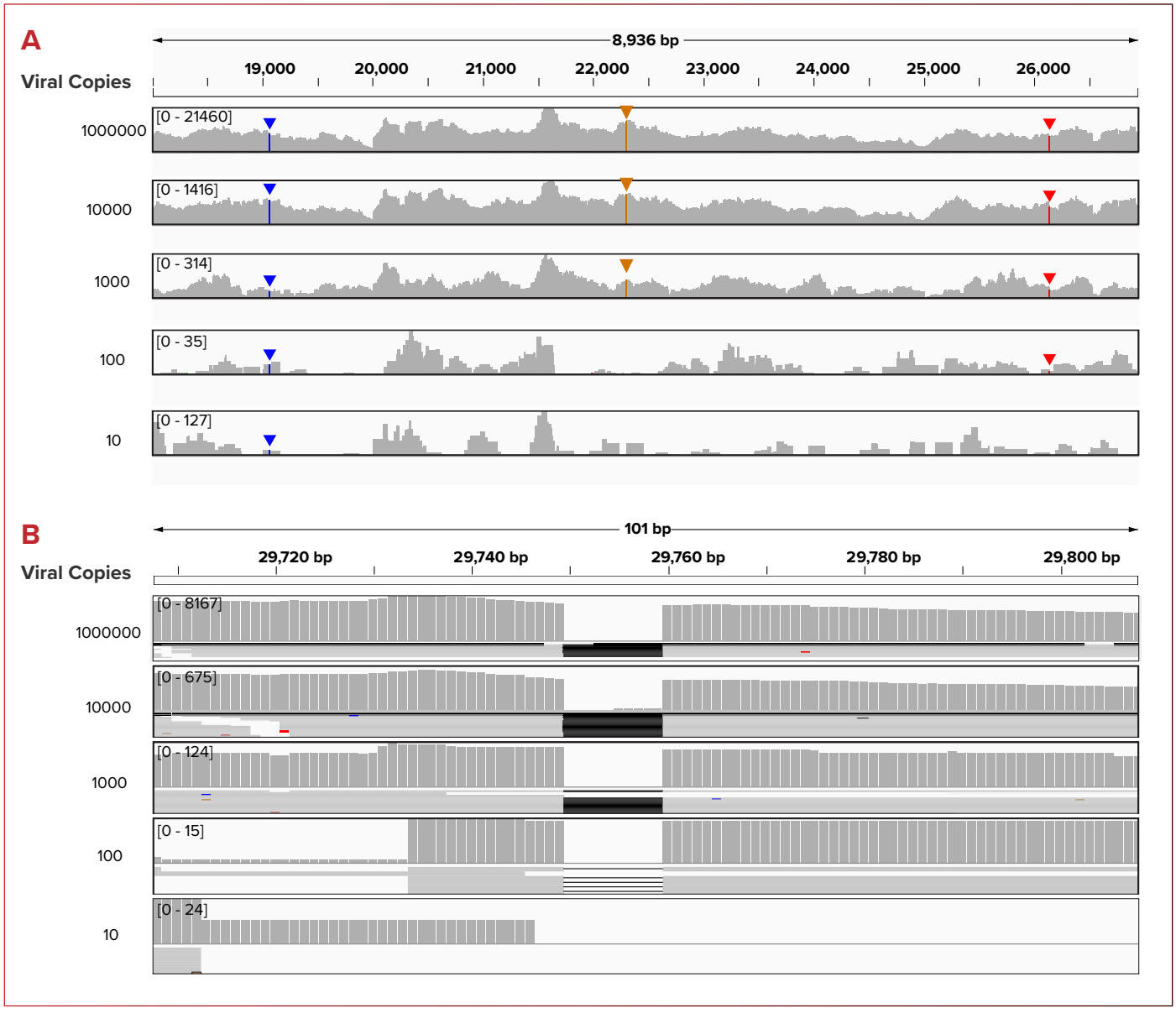


Figure 4. Detection of variants in MT007544.1 from 1 million reads of post-enrichment libraries. Each track shows IGVtools coverage of MN908947.3 Reference genome at a depth of 1 million reads for viral titers provided outside each box. Numbers in parentheses indicate the range of coverage. A. Three SNPs present between position 19000 to 26500 are detected even at a viral load of 1000 copies, indicated by colored triangles blue – 19065 (T>C), yellow – 22303 (T>G), red- 26144 (G>T) B. A 10 nt deletion between 29750-29759 indicated by lack of coverage is detected even at a viral load of 100 copies.

CONCLUSION

Here we demonstrate the utility and high sensitivity of SRSLY in a SARS-CoV-2 target enrichment workflow for downstream NGS-based analyses. We detect the virus down to 10 viral genome copies in a background of 50 ng of human gRNA and observe high breadth of coverage, particularly at viral loads that are thought to be clinically relevant[12]. We were also able to detect known genetic variations within the viral genome at titers as low as 100 copies, which indicate that SRSLY retains these genetic variations and that the downstream is tolerant to these mutations. While we have showcased the utility of the NEBNext First-strand sythesis module and the myBaits Expert SARS-CoV-2 Panel in this application note, the SRSLY protocol integrates with other modules such as the Zymo-Seq RiboFree Universal cDNA Kit and the Twist Bioscience® Twist SARS-CoV-2 Research Panel (write to technicalsupport@claretbio.com for more information).

SRSLY is a robust library preparation method that eliminates 2nd strand synthesis - a step that is prerequisite in tradition RNA-Seq workflows. (For more information about comparative performance metrics between SRSLY and other double-stranded approaches for RNA-Seq please refer to our technical note - https://www.claretbio.com/s/ClaretBio-Technical-Note-SRSLY-for-RNASeq.pdf). Omitting this potentially error-prone step makes SRSLY better-suited for studies of viral RNA genomes, including phylogenetics (or phylogenetic analyses). More recent studies have highlighted the importance of studying the viral transcriptome for understanding SARS-CoV-2 virulence. The inherently directional libraries synthesized with SRSLY can also facilitate improved understanding of the viral transcriptome apart from the genome; by retaining sense and antisense provenance of the transcripts when derived from samples such as nasal swabs, whole blood and from in-vivo cell-culture experiments. Together, this study demonstrates that SRSLY can be easily incorporated into existing RNA-Seq library preparation and target-enrichment workflows reducing total protocol time; while retaining sequence information and directionality of input molecules.

REFERENCES

1 Helmy, Y. A. *et al*. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. *J Clin Med* **9**, doi:10.3390/jcm9041225 (2020).

2 Vankadari, N. Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. *Gene* **752**, 144792, doi:10.1016/j.gene.2020.144792 (2020).

3 WHO. *Coronavirus disease 2019* https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 2020).

4 Zhu, N. *et al*. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733, doi:10.1056/NEJMoa2001017 (2020).

5 Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**, 9241-9243, doi:10.1073/pnas.2004999117 (2020).

6 Coronaviridae Study Group of the International Committee on Taxonomy of, V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5, 536-544, doi:10.1038/s41564-020-0695-z (2020).

7 Pachetti, M. *et al*. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* **18**, 179, doi:10.1186/s12967-020-02344-6 (2020).

8 Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep*, 100682, doi:10.1016/j.genrep.2020.100682 (2020).

9 Kim, D. *et al*. The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921 e910, doi:10.1016/j.cell.2020.04.011 (2020).

10 Troll, C. J. *et al*. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics* **20**, 1023, doi:10.1186/s12864-019-6355-0 (2019).

11 Caly, L. *et al*. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. *Med J Aust* **212**, 459-462, doi:10.5694/mja2.50569 (2020).

12 Pan, Y., Zhang, D., Yang, P., Poon, L. L. M. & Wang, Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis* **20**, 411-412, doi:10.1016/S1473-3099(20)30113-4 (2020).

SRSLY™

Claret Bioscience's SRSLY™ NGS library preparation kits for Illumina® provide reagents and proprietary adapters that convert a variety of fragmented DNA inputs (as low as 10 nt in length) to sequence-ready molecules. The kits are provided in a modular format that allow maximum flexibility in the choice of Index PCR primers and DNA purification strategies.

CLARETBIO PRODUCT INFORMATION:

To order the SRSLY kits email at orders@claretbio.com

Kit	Reactions	Catalog Number
SRSLY PicoPlus (Input < 10 ng)	24 or 96 reactions	CBS-K250B-24, CBS-K250B-96
SRSLY NanoPlus (Input > 10 ng)	24 or 96 reactions	CBS-K150B-24, CBS-K150B-96
Unique Dual Index Primers	24 or 96 reactions	CBS-UD-24, CBS-UD-96
Clarefy Beads	24 or 96 reactions	CBS-BD-24 or CBS-BD-96

myBaits®

myBaits® target capture kits from Arbor Biosciences provide pools of in-solution biotinylated RNA probes plus reagents for highly-efficient, scalable targeted sequencing. Arbor Biosciences' proprietary oligo synthesis technology delivers high-quality, complex baits at extremely competitive pricing. In addition, they offer complimentary bait design and project development assistance from a team of expert scientists.

ARBOR BIOSCIENCES PRODUCT INFORMATION:

To order the myBaits kits email sales@arborbiosci.com

Kit	Reactions	Catalog Number
myBaits Expert SARS-CoV-2 panel	16, 48 or 96 reactions	308716, 308748, 308796