INTERNATIONAL JOURNAL OF MEDICAL AND APPLIED SCIENCES
eISSN:2320-3137
Earthjournals Publisher
www.earthjournals.in

REVIEW ARTICLE

Next-generation sequencing and its clinical application on SARS-CoV-2

Shilpa K[1], Bernaitis L[2], Benita Mary L[3]

1. Department of Molecular Microbiology, Omix Research and Diagnostics Laboratories, Bangalore, Karnataka, India.
2. Department of Microbiology, Nandha Siddha Medical College and Hospital, Erode, Tamilnadu, India.
3. Department of Physiology,Rajas Dental College & Hospital, Kavalkinaru junction, Tirunelveli, Tamilnadu ,India.

Corresponding author: Dr.Bernaitis L, Associate Professor, Department of Microbiology, Nandha Siddha Medical College and Hospital, Erode - Tamilnadu, India. Orcid: 0000-0003-0192-3467

Publication history: Received on 24 July 2021, Accepted on 30 August 2021, Published online 6 September 2021

ABSTRACT:

Next-generation sequencing (NGS) is a new technologyused for DNA and RNA sequencing and variant/mutationdetection. NGS can sequence hundreds and thousands ofgenes or whole genome in a short period of time. Thesequence variants/mutations detected by NGS have beenwidely used for disease diagnosis, prognosis, therapeuticdecision, and follow up of patients. The capacity of itsmassive parallel sequencing offers new opportunities forpersonalized precision medicine. We also present an outline of current repositories and databases that provide access to SARS-CoV-2 genomic data and associated metadata. Finally, we offer general advice and guidelines for the appropriate sharing and deposition of SARS-CoV-2 data and metadata, and suggest that more efficient and standardized integration of current and futureSARS-CoV-2-related data would greatly facilitate the struggle against this new pathogen.

KeyWords: COVID-19, SARS-CoV-2, NSG (Next Generation Sequencing)

INTRODUCTION

NGSis a new technology for DNA and RNA sequencingand variant/mutation detection. This technology combinesthe advantages of unique sequencing chemistries, differentsequencing matrices, and bioinformatics technology. Such acombination allows a massive parallelsequencing of variouslengths of DNA or RNA sequences or even whole genomewithin a relatively short period of time. It is a revolutionarysequencing technology after Sangersequencing.NGS involves several major steps in sequencing.
Forexample, DNA NGS involves DNA fragmentation, librarypreparation, massive parallel sequencing, bioinformaticsanalysis, and variant/mutation annotation and interpretation1.

The COVID-19 epidemic was declared a public health emergency of international concern. Since the outbreak of COVID-19, there has been considerable discussion on the origin and human-to-human transmission of the causative virus. Interestingly, ACE2 receptors are also reported to be expressed in the kidney and gastrointestinal tract, tissues known to harbor SARS-CoV-2.

Notably, next-generation sequencing (NGS) has been applied to the study of COVID-19 and has greatly promoted SARS-CoV- 2 origin tracing. It is critical to explore the potential origins and mechanism of SARS-CoV-2 to control the spread of COVID-19 and further improve the therapeutic regimen2.

NGS PLATFORMS

Currently available NGS platforms apply different approaches to achieve high-throughput sequencing. The differences in sequencing approach in turn influences the sequencing quality, quantity and choice of application. The general approach for a typical NGS run begins with genomic DNA extraction from test samples, library preparation which involves DNA fragmentation, ligation of adaptors, adaptor sequencing, and sample enrichment and finally sequencing.

Illumina

Illumina , is perhaps the most popular among currently available NGS platforms offering various scalable options that complement requirements of different study designs, cost of sequencing and intended use of the sequencing data. Illumina offers a method for selecting
an optimum sequencing platform via its sequencing platform comparison tool3.

MATERIALS AND METHODS:

Samples from nasal, nasopharyngeal, and oropharyngeal swabs were obtained according to the standard protocol and collected in 3 ml sterile viral transport medium (VTM) tube or 1ml of TRIzol reagent (Invitrogen). All the samples were transported to the laboratory at a cold temperature (2–8˚C) within 72 hours post collection, and stored at -80˚C till further used.

RNA isolation

RNA extraction was carried out in a pre-amplification environment with a Biosafety level 2 (BSL-2) facility. RNA isolation was done using four different methods. For manual RNA extraction, a total of 140 µl of the VTM medium was used; prior to isolation, the VTM samples were subjected to heat inactivation at 50˚C for 30 minutes. After heat inactivation, the RNA was extracted from 140 µl of VTM samples using QIAamp1 Viral RNA Mini kit (QIAGEN) as per the manufacturer's instructions. For the automated magnetic bead-based extraction method, 200 µl of VTM was transferred to a 96-well deep well cartridge plate supplied with the kit (VN143), and extraction was performed on Nextractor1 NX-48S instrument (Genolution Inc.) as instructed by the manufacturer. After a bead-based capture and washing process, the RNA sample was eluted in 40 µl of the elution buffer. For RNA isolation using Trueprep AUTO v2 universal cartridge-based sample prep device, (Molbio Diagnostics Pvt. Ltd.) 500 µl of the VTM was added to the 2.5 ml of lysis buffer provided with the kit. After pipette mixing, 3 ml of the mixture was dispensed in the provided cartridge; the final RNA was eluted in 50 µl of elution buffer.

For RNA from TRIzol reagent, the tubes containing swabs were vortexed briefly. The overall content of the TRIzol tube was transferred into a 1.5 ml tube, followed by the addition of 200 µl of chloroform and mixed by inverting the tubes several times. After 5 minutes of incubation, the 1.5 ml tubes were centrifuged for 15 minutes at 12,000 RPM at 4˚C. The upper clear aqueous layer which contains the RNA was transferred to new tubes. An equal amount of isopropanol was added to the tubes containing the RNA. Contents of the tubes were mixed by inverting the tubes several times and tubes were incubated for 10 minutes on ice followed by centrifugation for 10 minutes at 12,000 RPM at 4˚C. The supernatant was discarded and the RNA pellet was dissolved in 30 µl of RNase-free water after 2 ethanol washes. TURBO DNase (Ambion, Applied Biosystems) treatment was given to the isolated RNA to remove genomic DNA contamination in the samples followed by RNA purification using the phenol/chloroform method4.

Library preparation and sequencing

A step-by-step protocol describing the library preparation and sequencing protocol can be found at Protocol Exchange. For all libraries, each sample was pooled (7 µL/sample) and library PCR products were purifified with SPRIselect beads (A66514, Beckman Coulter). The PoC, test, and pilot cohorts were purifified as follows: ratio 0.8:1 (beads:library), and the extended cohort with 1:1 (beads:library) (Beckman Coulter). Due to NSA products in the fragment analyzer profifile (Supplementary Fig. 3c) in the test cohort and pilot cohort, we performed size selection purifification (220–350 bp) using the Pippin Prep system (Pippin HT, Sage Science). Library quality was assessed with the 5200 Agilent Fragment Analyzer (ThermoFisher) and Qubit 2.0 Fluorometermina) using 75 bp paired-end sequencing5.

RESULTS:

Detection of SARS-CoV-2 by mNGS

First, we demonstrated that SARS-CoV-2 could be detected by direct long-read third-generation metatranscriptomic sequencing from nasopharyngeal (NP) swab specimens of COVID-19 patients. SARS-CoV-2 was identifified in 31/40 (77.5%) samples that were positive for SARS-CoV-2 by the diagnostic RT-PCR using the online CosmosIDbioinformatics program.Time to detection of SARS-CoV-2 reads ranged from 1 min (cycle threshold [Ct], 16.0) to 15 h (Ct, 33.4) after the start of the sequencing run, which correlated with the RT-PCR Ct values . In the 8 samples where SARS-CoV-2 was unable to beidentifified, the Ct values ranged from 21.0 to 36.6, with mean and median values of 29.1 and 29.0, respectively. We considered that perhaps an abundance of host reads could have masked the SARS-CoV-2 reads in these samples, but there was no relationship between the number of SARS-CoV-2 reads that were detected and human reads detected (Fig. 1a). Lower Ct values were associated with increased sequencing coverage of the SARS-CoV-2 reference genome, a decreased number of total sequencing reads compared to the fifirst SARS-CoV-2 read detection, a greater proportion of SARS-CoV-2 total matches by CosmosID, and a decreasing number of days from the onset of symptoms. We observed that the most severe cases were spread out along the range of Ct values . No SARS-CoV-2 reads were identified aligning in any of the 10 samples obtained from patients that were suspected of having SARS-CoV-2 infection but were

negative by RT-PCR. A summary of sequencing reads and taxonomic classification of sequencing reads are provided in the supplemental material, respectively.

Identifying coinfections

Next, we examined the samples for possible coinfections. Of the 40 COVID-19-positive samples, 5 (12.5%) revealed organisms of clinical relevance detected in high abundance (50% relative to normal levels in microbiota). These included Haemophilus influenzae (n 2; 5%), Moraxella catarrhalis(n 1; 2.5%), human metapneumovirus (hMPV) (n 1; 2.5%), and human alphaherpesvirus 1 (n 1; 2.5%) . In our COVID-19-negative samples, Moraxella catarrhalis(n 1; 10%) was identified. Unfortunately, standard-of-care testing was not performed to detect these pathogens. No fungal or protist coinfections were detected.

Evaluating the respiratory microbiome

Beyond determining coinfections, we analyzed the metagenomic profiles of these patients in order to uncover potential shifts in the microbiome that could impact patient outcomes. Specimens obtained from SARS-CoV-2-positive patients did have a significant reduction in the diversity of their bacterial communities at the species level as measured by the Shannon diversity index (P 0.0082), Chao richness estimate (P 0.0097) , and Simpson diversity index (P 0.018). We did not see significant differences at the genus and family levels. Given that we did see a decrease in diversity in the positive samples, we were interested in determining if there was decreasing diversity at lower Ct values. However, among these samples, we did not observe any relationship between Ct values and diversity (data not shown using the same analyses described above6.

Our Systematic Parallel Analysis of Endogenous RNA Regulation Coupled to Barcode Sequencing (SPAR-Seq) systemwas modified to simultaneously monitor COVID-19 viral targets and additional controls by multiplex PCR assays. For barcode sequencing, unique, dual-index C19-SPAR-Seq barcodes were used. Unique reverse 8-nucleotide barcodes were used for each sample, while forward 8-based barcodes were used to mark each half of the samples in 96- well plate to provide additional redundancy. These two sets of barcodes were incorporated into forward and reverse primers, respectively, after the universal adaptor sequences and were added to the amplicons in the second PCR reaction. The C19-SPAR-Seq analysis pipeline with the algorithms used is explained in detail in Supplementary with additional analytical tools described in Supplementary and below in the “Methods” sections. Computational requirements for the demultiplexing step is 32 GB RAM and minimum 1 GB network infrastructure, with a Linux-operating system. Demultiplexing and mapping. Illumina MiSeq sequencing data was demultiplexed based on perfect matches to unique combinations of the forward and reverse 8 nucleotide barcodes. Full-length forward and reverse reads were separately aligned to dedicated libraries of expected amplicon sequences using bowtiewith parameters –best -v 3 -k 1 -m 1. Read counts per amplicon were represented as reads per million or absolute read counts.

Filtering of low-input samples.

To remove samples with low amplifified product, likely reflecting low input due to ineffificient sample collection or degradation, before attempting to classify, we computed precision-recall curves for classifying control samples into ‘low amplifification’ and ‘high amplifification’ based on reads mapped to RNA amplicons but ignoring mapping to

genomic sequence, if applicable. The former group comprised all controls in which individual steps were omitted (H2O controls) and the latter comprised HEK293T as well as synthetic SARS-CoV-2 RNA controls. For each PoC, test, pilot, extended runs, we obtained the total mapped read threshold (including reads mapping to both human and viral amplicons) associated with the highest F1 score, representing the point with optimal balance of precision and recall. Samples with reads lower than this threshold were removed from subsequent steps. To assign positive and negative samples, we used negative (H2O and HEK293T) and positive (synthetic SARSCoV-2 RNA dilutions) internal controls for each run and calculated optimum cutoffs for viral reads (total reads mapping to all three viral amplicons) by PROC which defines the threshold for optimum positive predictive value (PPV) andnegative predictive value (NPV) for diagnostic tests. Thus, a sample was labeled positive if it had viral reads above the viral read threshold; negative if it had viral reads below the viral read threshold and human reads above the mapped read threshold; and inconclusive if it had both viral and human reads below the respective thresholds. Sample classification by heatmap clustering. Heatmap and hierarchical clustering of viral and control amplicons, log10(mapped reads + 1), was used to analyze and classify all samples. Samples with a total mapped read count lower than the RNA QC threshold were labeled as inconclusive and removed before the analysis. Known positive (high, medium, and low) and negative control samples were used as references to distinguish different clusters. In addition, dilutions of synthetic SARS-CoV-2 RNA were also included as controls and analyzed across different PCR cycles and primer pool conditions. Viral mutation assessment. To remove PCR and sequencing errors for the assessment of viral sequence variations, we determined the top enriched amplicon sequence. For this, firstly, paired end reads were stitched together to evaluate full length amplicons. The last 12 nucleotides of read1 sequence are used to join the reverse complement of read2 sequences. No mismatches were allowed for stitching criteria. The number of full length reads per unique sequence variation were counted for each amplicon per sample by matching the 10 nucleotides from the 3′ and 5′ end of the sequence with gene-specifific primers.

DISCUSSION:

Amplicon-based sequencing

Amplicon sequencing enables researchers to restrict the scope of their analysis only to alimited number/type of sequences of choice. This approach is highly specific, but requires significant a priori knowledge of the sequence that is to be ‘targeted.’ Diagnostic RT-PCR tests for the detection of SARS-CoV-2 nucleic acids from clinical specimens, which are based on very specific primers for the amplification of discrete regions of the genome of the virus, could be considered a specialized form of amplicon sequencing. Amplicon-based approaches for the sequencing of SARS-CoV-2 adopt an enrichment workf low consisting of firststrand cDNA synthesis followed by genome amplification with multiplex PCRs. The objective is to produce pools of amplicons that cover either the entire length or the discrete portions of the viral genome. Several different multiplex PCR designs, differing in the number and size of amplicons, have been proposed for SARS-CoV-2. Amplicon sequencing is highly specific and robust to low amounts of RNA and degraded samples, and less sequencing is required with respect to the metatranscriptomic approach since non-viral reads are rare. While amplicon sequencing is

theoretically convenient and cheap, it presents some limitations which should be considered. Firstly, because of differences in primer efficiency, or possible variants in the primer annealing regions, amplification across the genome can be biased, with decreased coverage in specific genomic regionsand/or 3and 5UTRs regions missed altogether leading to an incomplete assembly. Moreover, since the primers are designed on the reference SARS-CoV-2 genome sequence, this approach may not identify large structural variants and can present systematic limitations in the presence of high levels of genomic divergence. While the amplicon-based approach is highly dependable for the reconstruction of the most prevalent genome variant in a viral population, a recent study suggests that it provides highly biased representation of minor allele frequencies with respect tothat derived from metatranscriptomics experiments performed on the same samples.

Several commercial kits and non-commercial protocols are available for SARS-CoV-2 amplicon preparation, some of which are tailored to particular NGS platforms. Since sequencing depth is a marginal consideration, libraries can be sequenced on benchtop platforms with a mid-throughput (i.e. Illumina NextSeq and Miseq; Ion torrent platforms,etc.). Additionally, when combined with the short turn-around times of Single MoleculeSequencing (SMS) technologies such as ONT and PacBio, amplicon sequencing ofSARS-CoV-2 can be used for rapid surveillance of transmission chains, as exemplified by the approach adopted by the ARTIC network for real-time monitoring of the COVID-19 outbreak in the United Kingdom, where a fast, ampliconbased protocol successfully applied to previous viral outbreaksestablished a rapid in house tiling multiplex PCR protocol for the simultaneous detection and sequencing of several respiratory viruses which includes a large part of the SARSCoV-2 genome. The Wang protocol has also been suggested for diagnostic usage as it shows higher sensitivity than approved RT-qPCR tests. While several SARS-Cov-2 genome sequencing protocols using tiled amplicons are available for the PacBio platform (see https://www.pacb.com/research-focus/microbiology/COVID-19- sequencing-tools-and-resources/), to our knowledge they have been scarcely used until now, although a major study of the introduction and spread of SARS-CoV-2 in the New York City area used both the PacBio and the Illumina technologies. The robustness of amplicon-sequencing to degraded and low concentrations of RNA is evident from studies of environmental specimens, where this approach is followed by sequencing with Ion torrent or ONT for wastewater samples, and by Sanger sequencing for a patient breathing air sample and for a door handle swab.

Hybrid capture-enrichment sequencing

Similar to amplicon-based sequencing, hybrid capture is a sequencing strategy that enables researchers to target only predefined sequences or regions of a genome that are relevant to their specific interests. Target-enrichment strategies using hybrid capture were originally developed for human genomic studies, to enable the rapid and cost-effective sequencing of the exons of protein coding genes (exome sequencing). Exome sequencing is still considered the method of choice for the study of genetic variation in protein coding loci in humans, as it achieves a good trade-off between the specificity of amplicon based enrichment, and the sensitivity (to different types of genetic variants) of shotgun sequencing at significantly lower costs. Hybrid capture enriches targeted genetic material through hybridization to specific biotinylated probes, allowing a considerably reduced sequencing depth compared with shotgun Metatranscriptomics8.

Contamination

Contamination is a persistent concern in clinical microbiology, and the wide-ranging detection ability of the NGS technology can exacerbate this issue. Contamination can be traced to laboratory environments, reagents, or personnel, as well as non-traditional sources of contamination stemming from the nature of NGS technology. These challenges can be addressed with standard sampling and laboratory sanitation protocols, as well as the implementation of standards in the bioinformaticworkflflow, such as those that require minimum pathogen abundance thresholds to reduce false positives.Studies indicate that when an mNGS sample is a true negative, false-positive readings from background contaminants may increase . The NGS limits of detection for causative pathogens can be monitored with positive controls that are spiked into samples, and spiked controls may also help to provide some level of background suppression . Further investigation into the balance between positive controls, background contaminants, and their effects on performance is warranted to eliminate the existing challenges. The unique nature of next-generation sequencing also yields contamination in downstream processes of the clinical workflflow, including data analysis. When samples are multiplexed, some sequences can be misclassified during the de-multiplexing stage to an incorrect index (index hopping), leading to inaccurate test results if not properly addressed during bioinformaticanalysis . Additionally, with the possibility of over 99% of reads being from the human host , massive amounts of host genetic material, as well as that of the healthy, host microbiome, can complicate the interpretation of results and need to be removed fromthe dataset to prevent “contamination” of bioinformatic analyses. Finally, errors in reference databases can cause problems for specimen identification, and these databases need to beroutinely monitored by regulatory authorities to maintain reliability and up-to-dateinformation9.

CONCLUSION:

To conclude, the data presented herein are useful for clinical and research teams who want to implement SARS-CoV-2 sequencing and chose the most suitable protocol according to the application. To summarise, mNGS remains the gold standard for samples with high viral load to obtain a maximum of information without any bias. For low- and mid-Ct values, RVOP leads to very high coverage as well, enabling genome end sequencing, contrary to amplicon methods. For higher Ct values, amplicon-based enrichment is a very interesting alternative, in particular ARTIC-ONT protocol that did not show any major dropout issues in this present evaluation. However, as a reminder, loss of coverage in any given region of the genome should alert to a potential rearrangement or an SNP in primerannealing or probe-hybridizing regions and would require further validation using unbiased metagenomic sequencing10.

REFERENCES:

1.Dahui Qin, 2019.Next-generation sequencing and its clinical application,CancerBiol MedCancer Biol Med,16(1):5-9.

2. XiaominChen,Yutong Kang, Jing Luo, Kun Pang, Xin Xu,et.al.2021.Next-Generation Sequencing Reveals the Progression of COVID-19,Frontiers in Cellular and Infection Microbiology,11(1):1-14.

3.Aquillah M. Kanzi, James Emmanuel San, Benjamin Chimukangara, Eduan Wilkinson, Maryam Fish,et.al.2020.Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance,Frontiers in Genetics,11(1):1-18.
4.Rahul C. Bhoyar, Abhinav Jain, Paras Sehgal, Mohit Kumar Divakar, DishaSharma,et.al.2021.High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing,PLOS ONE,1-16.
5.MassabUmair, AamerIkram, Muhammad Salman, Adnan Khurshid, Masroor Alam, et.al.2021. Whole-genome sequencing of SARS-CoV-2 reveals the detection of G614 variant in Pakistan,PLOS ONE,1-11.
6.Heba H. Mostafa,John A. Fissel,BrianFanelli,Yehudit Bergman,VictoriaGniazdowski, et.al.2020. Metagenomic Next-Generation Sequencing of NasopharyngealSpecimens Collected from Confirmed and Suspect COVID-19Patients,mbio.asm.org,11(6):1-12.
7.Marie-Ming Aynaud,J. JavierHernandez,SedaBarutcu,UlrichBraunschweig, Kin Chan,et.al.2021.A multiplexed, next generation sequencing platform for high-throughput detection of SARS-CoV-2.NATURE COMMUNICATIONS,12(1405):1-10.
8.Matteo Chiara, Anna Maria D'Erchia, Carmela Gissi, Caterina Manzari, Antonio Parisi,et.al.2020.Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities,Briefings in Bioinformatics,00(00):1-15.
9. Goldin John, Nikhil Shri Sahajpal, Ashis K. Mondal, SudhaAnanth, ColinWilliams,et.al.2021.Next-Generation Sequencing (NGS) in COVID-19: A Tool forSARS-CoV-2 Diagnosis, Monitoring New Strains and Phylodynamic Modeling in Molecular Epidemiology,Curr. Issues Mol. Biol.43:845–867.
10.CarolineCharre, ChristopheGinevra,MarinaSabatier,HadrienRegue, Gregory Destras,et.al.2020.Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation,Virus Evolution,6(2):1-8.

Paper cited as: Shilpa K, Bernaitis L, Benita Mary LNext-generation sequencing and its clinical application on SARS-CoV-2. International Journal of Medical and applied Sciences.2021;10(2): 61-68.