Faculty of Computer and Information Sciences
Biomedical Department
Mansoura University

High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing

Alaa Mohamed Al-Adl

Abstract

The rapid development of the 2019 Coronavirus Disease (COVID-19), a global pandemic which affects millions of people globally, requires sensitive and high-performance approaches for diagnosing, monitoring and determining the SARS-genetic CoV-2's epidemiology. In this study we used the protocol COVIDSeq involving multiple-plex-PCRs, bar-coding, and the sequencing of high-throughput samples and decided to support SARS-genetic CoV-2's epidemiology. We used 752 clinical samples out of a total of 1536 samples. A thorough analysis revealed a number of six highly confidential COVIDSeq samples that detected SARS-CoV-2 in RT-PCR in 21 and 16 samples classified as inconclusive and pan-sarbon positive respectively suggesting that COVIDSeq could be used as a confirmatory test. Thus, we have analysed COVIDSeq as an additional advanced tool for the development of SARS-CoV-2 gene epidemiology as a potential high sensitivity test.

Introduction

Coronavirus 2019 (COVID-19) is the world pandemic that affects millions of people around the globe, imposing enormous burdens on national health and socio-economic welfare systems. Methods used for testing are mainly subdivided into serological antigen-antbody based tests, nuclear acid-based amplification testing, and sequencing based assays. Some of these approaches have also been adapted to enable higher performances. While serological tests are high-sensitivity and low-specific tests.The Gold Standard for detection and diagnosis has been Nucleic acid-based amplifica refractological treatment, such as quantitative real-time PCR (QRT-PCR), but the negative RT-PCR does not prevent infection in clinical cases which are suspected. The rapid advanced sequence sequencing and analysis techniques of the next generation have enabled us to understand and interpret the genetic makeup of SARS-CoV-2 and its epidemiological evolution. The full SARS-CoV-2 genome sequence was decidibated by viral RNA sequencing from the initial cluster of cases. This study describes application of the recently approved COVIDSeq Protocol for clinical use by the US Food and Drug Authority (US FDA). The protocol provides for the high performance detection and genetic epidemiology of SARS-CoV-2 iso to be used in one sequence with a multiplex amplicon-based PCR enrichment and with barcodes with a performance of 1536 samples using the NovaSeq S4 flow cell. Our analytical assessment suggests that the protocol COVIDSeq may be a sensitive detection approach, with genetic genetic epidemiological insight available. This is the first COVID realistic assessment to our best knowledge.

Related Work

- The COVIDSeq approach also offered insight into SARS-CoV-2 isolates' genetic epidemiology and evolution. A sigto-a large number of genomes that had an insight into the prevalent lineage/clades of the virus could be carried out for this phylogenetic examination [23–25]. The analysis also reports for the first time in two lines B.1.112 and B.1.99.India. Indian. In previous reports, B.1.112 and B.1.99 were reported respectively from the USA and UK. Implies their origin, distribution and possible introductions to India beyond India. Travellers from those countries would need more information to confirm this the as sumption.
- Our analysis shows that the technical duplicates and the high concorrence of SARS-CoV-2 detection between COVIDSeq approaches and RT-PCR approaches are highly consistent. Our comparative analysis of the RT-PCR and COVIDSeq SARS-CoV-2 Detecting test has shown that the COVIDseq test is comparable to that of RT-PCR in terms of sensitivity, precision and accuracy. COVIDSeq protocol, deleted from RT-PCR assays, suggest a comparable sensitivity of a sequence-based test compared to RT-pCR in samples precluded as non-conclusive (21/35) and pansardine (16/ 43) and negative (6/19).. This was spilled into 43/97 additional samples and a potential 5.71 percent gain in all the samples.
- Finally, our analysis suggests that COVIDSeq is a high performance, high-performance, SARS-Co V-2, based approach. COVIDSeq also has an additional benefit to enable SARS-CoV-2 genetic epidemiology.

❑ Shakya M, Ahmed SA, Davenport KW, Flynn MC, Lo C-C, Chain PSG. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Sci Rep. 2020; 10: 1723. https://doi.org/10.1038/s41598-020-58356-1 PMID: 32015354.

❑ Langat P, Raghwani J, Dudas G, Bowden TA, Edwards S, Gall A, et al. Genome-wide evolutionary dynamics of influenza B viruses on a global scale. PLoS Pathog. 2017; 13: e1006749. https://doi.org/10.1371/journal.ppat.1006749 PMID: 29284042

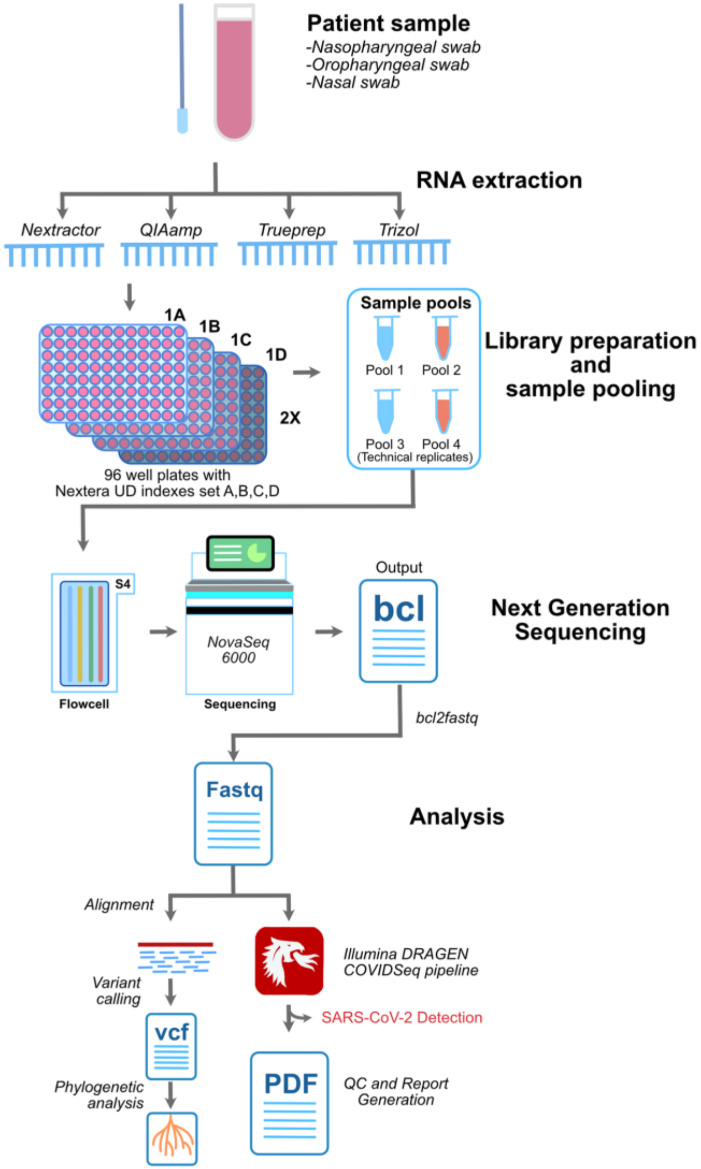
Methods

Patients and samples: The study was agreed on by the Committee on Institutional Human Ethics (IHEC no. Dated CSIR IGIB/IHEC/2020-21/2001) of the CSIR Institute for Genomics and Integrative Biology and the Ethics Committee revoked patient consent. Samples were obtained in the standard Protocol from nasal, nasopharyngeal and oropharyngeal swabs and collected in 3 ml of Viral Transportation Tube Sterile (VTM) or 1 ml of TRIzol reagent (Invitrogen). All samples were transported to the laboratory at cold temperatures (2-8 µC) and stored at -80 µC for further use within 72 hours following collection.

RNA extraction was conducted with a Biosafety Level 2 (BSL-2) facility in a pre-amplification environment. Four different methods have been used to isolate the RNA. A total of 140 µl of VTM was used for manual RNA extraction. Before insulation, thermal inactivation of the VTM samples at 50 µl was carried out for 30 minutes. Using the QIAamp1 Viral RNA Mini Kit (QIAGEN) as manufacturer's instructions, RNA was extracted from 140 µL of VTM samples following heat inactivation. 200 µl VTM has been transferred to a 96-well deep cartridge board (VN143), which has been equipped with the kit, for the automated beaded magnet extraction method and extracted by Nextractor1 NX48S.

Data processing: The data generated by DRAGEN COVIDSeq test pipeline (Illumina Inc.) on the Illumina DRAgen v3.6 Bio ATM platform in binary basis (BCL) format in NovaSeq 6000 were processed according to standard protocol. The analysis consists of sample sheet validation, quality data inspection, generation of FASTQ and detection of SARS-CoV-2 when five SARS-CoV-2 samples are detected. For alignment, variant and consensus-sequence-generation, further samples were processed with SARS-CoV-2 and at least 90 targets.

Annotation of genetic variants: ANNOVAR was used to annotate the variants systematically[34]. RefSeq has been taken up with annotations of genomic loci and protein functional consequences. For functional impact annotations, potential immune epitopes, protein areas and evolutionary conservation values, customised data bases were created.



Schematic summary of the analysis in this study. The methodology adopted in the sampling, library preparation, sequencing and analysis involving custom based pipeline and COVIDSeq pipeline employed in this study

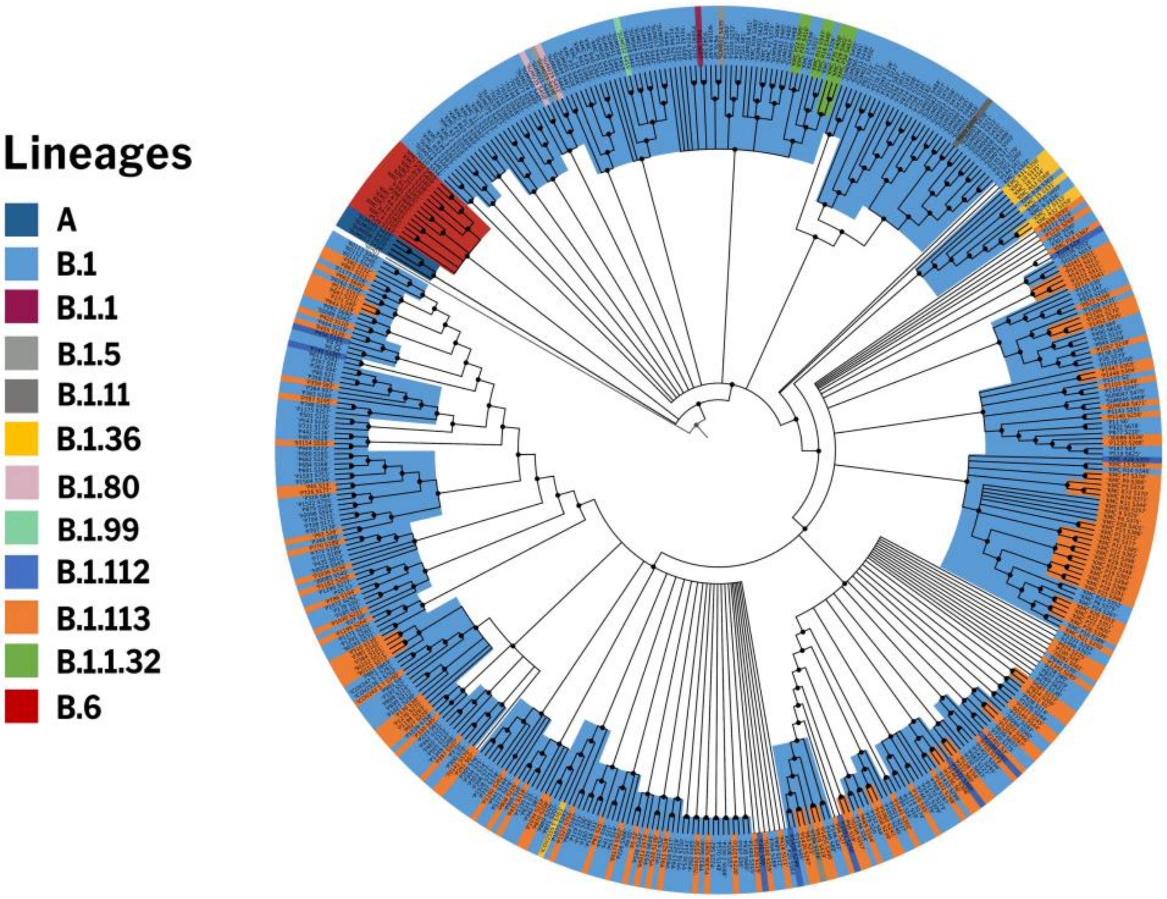
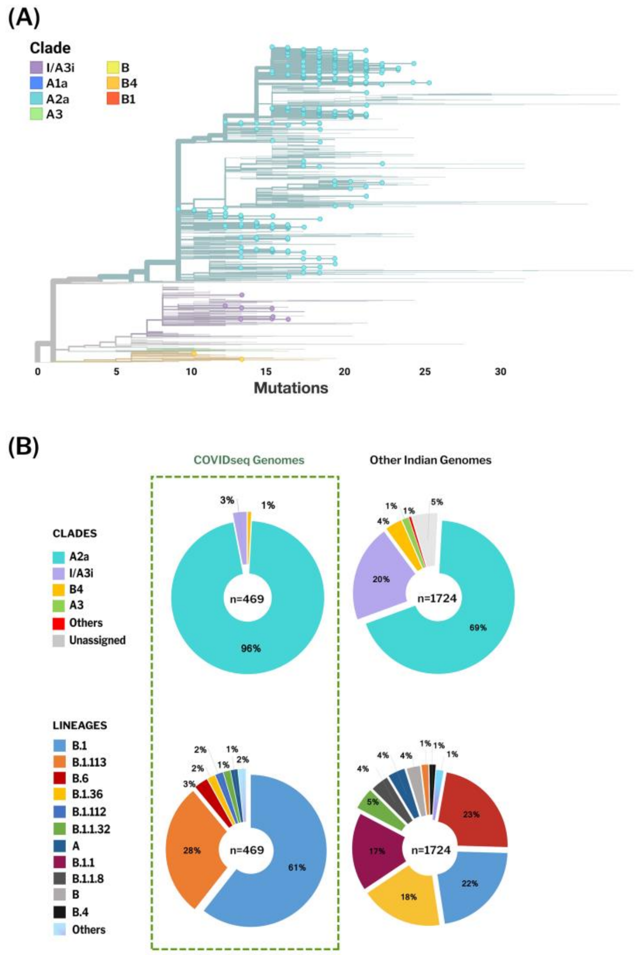


Fig 6. Phylogenetic distribution of PANGOLIN lineages in COVIDseq genomes. The distribution of lineages assigned by PANGOLIN in 469 CCOVIDseq genomes with the Wuhan/WH01 (EPI_ISL_406798) as reference.

Phylogenetic analysis

A total of 495 samples, including a dataset of SARS-CoV-2 genomes from India deposited in GISAID, had at least 99% genome coverage. Table S8 lists the sample names and names of the institution(s) originating and submitting. A protocol on phylogenetic clustering was previously described[36]. The analytic results removed a total of 26 COV to IIDSeq genomes with Ns > 5%. Also excluded from analysis were GISAID genomes having Ns > 5% and ambiguous data of the collection of the samples. Nextstrain [15] developed the phylogenetic network using the SARS-CoV-2 Analysis protocol.

The genome sequences were aligned with MAFFT and the reference genome were masked [37]. Using IQTREE, a raw phylogenetic tree was built to construct a phylogeny of the molecular-clock, to deduce mutations and to identify clades [38]. Apply, an interactive visualisation web application from Nextstrain, viewed the resulting phylogenetic tree.



Phylogenetic distribution of Indian SARS-CoV-2 genomes.

A) Phylogenetic trees generated by Nextstrain. 469 COVIDseq genomes reported from this study are highlighted. The 469 genomes cluster under clade A2a, I/A3i and B4,with A2a being the dominant clade.

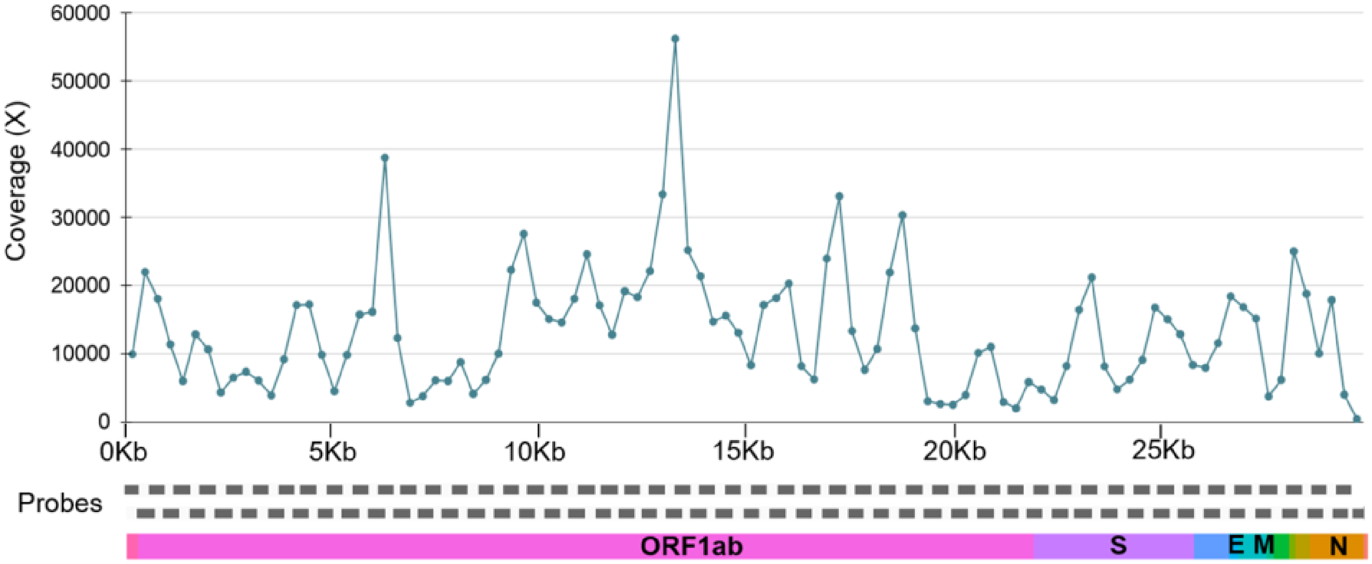
B) The proportion of clades and PANGOLIN lineages representing the Indian genomes. B.1 and B.1.113 are the dominant lineages in COVIDseq genomes whereas other Indian genomes show a dominance of B.6 and B.1 lineages.

Result

There were a total of 752 samples in the sample panel. Of those, the diagnostic guide set by the Indian Council for Medical Research (ICR) was 655 (87.1 percent) positive RT-PCR SARS daCoV-2 (ICMR). We included 19 RT-PCR negative (2.5%) and 43 (5.7%) samples, because they were positive only for primary E genes. Samples were classified as pan-Sarbotic samples. E gene is conserved in all the beta-CoVs. The presence of either SARS-CoV or SARS-CoV-2 or other bat-CoVs is detected in the individual sample.

The COVID-19 detection was conducted using the DRAGEN COVIDSeq test pipeline, which has at least 5 SARS-CoV-2 detection criteria to be seen as positive. Of the 1,504 samples, 1,352 were successfully recorded in the DRAGEN COVIDSeq Test pipeline. An additional 136 samples were undetected and the internal quality control failed 16. This is the same as sixty-six unique samples detected with SARS-CoV-2, sixty-eight unusual samples that did not detected SARS-CoV2 and eight unique test samples.

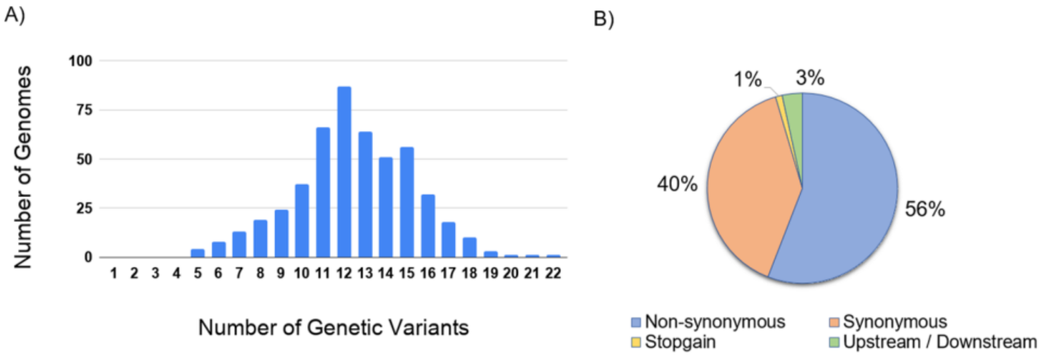
Since the RNA samples were taken from several RNA extraction protocols, we can also learn more about the COVIDSeq test compatibility of the protocols. 182 samples have been taken using the QIAamp1 Viral RNA Mini set, 264 with the use of the NX-48S Nextractor1 (Geneva, Korea), the TrueP1 AUTO v2 Samples (MolbioDiagnostics Pvt., Ltd.) and the TRizol-based extraction method, eight samples have been treated. Of those, in 168 (92.3 percent) of 182 samples (92.3%) extracted of samples from QIAamp1, COVIDSeq detected SARS-CoV-2 in 264 (99.6%) percentage samples from Nextractor1 NX-48S, of which 194 of 201 samples have been extracted (96.5 percent)



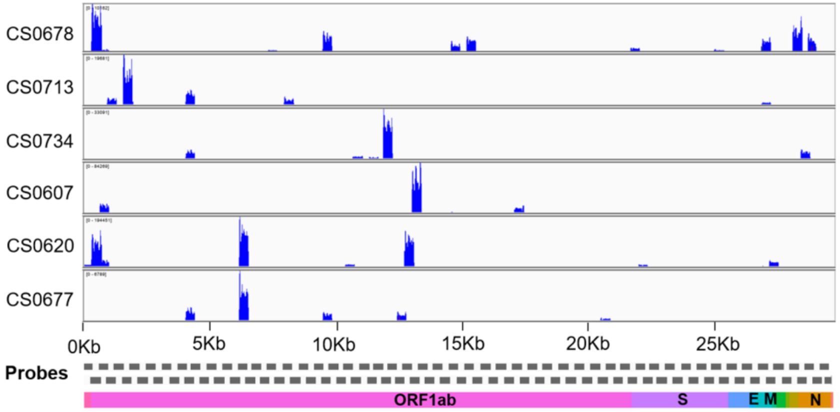
The line plot for the mean coverage of SARS-CoV-2 genome. The mean coverage for the 98 amplicons across the SARS-CoV-2 genome

The mean coverage for all samples of 98 PCR amplicons covering the entire SARS-CoV-2 genome as shown in Figure 3 was also reported. The medium amplicon coverage for the positive samples was ~14256x (706 genome-covered samples >5%). We discovered 20 amplicons with a coverage of ±2 standard deviations (SD), 16 amplicons with a coverage of less than 2 SD and 4 amplicons with a coverage of less than 2 SD.For reads and 0.984 (p-value < 0.00001) for coverage there was a correlation coefficient of 0.99 for the technical duplicates.

The alignment files were merged and the varying variants called by VarScan for further genome assembly and variant calling. The variant call was taken into account only 495 samples, which had a minimum of 99 percent of the genome covered. Of these 495 samples 91 were Andhra Pradesh samples, 63 were Odisha samples, and 341 were Delhi samples. A total of 1,143 unique variants were identified in the analysis. Compared to other Indian and global genome data, 73 genetic variants have been found to be novel and reported for the first time. The average of the so-called variants was 12.



Variant number per genome and their annotation.
(A) Distribution of variants in the genomes with 99% coverage
(B) Summary of the variant annotations



The coverage plot across the SARS-CoV-2 genome. The coverage plot constructed using Integrative Genome Viewer (IGV) for samples that were negative on RT-PCR assays but detected by DRAGEN COVIDSeq Pipeline.