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

Rahul C. Bhoyar[1], Abhinav Jain[1,2], Paras Sehgal[1,2], Mohit Kumar Divakar[1,2], Disha Sharma[1], Mohamed Imran[1,2], Bani Jolly[1,2], Gyan Ranjan[1,2], Mercy Rophina[1,2], Sumit Sharma[1], Sanjay Siwach[1], Kavita Pandhare[1], Swayamprabha Sahoo[3], Maheswata Sahoo[3], Ananya Nayak[3], Jatindra Nath Mohanty[3], Jayashankar Das[3], Sudhir Bhandari[4], Sandeep K Mathur[4], Anshul Kumar[4], Rahul Sahlot[4], Pallavali Rojarani[5], Juturu Vijaya Lakshmi[5], Araveti Surekha[5], Pulala Chandra Sekhar[5], Shelly Mahajan[6], Shet Masih[6], Pawan Singh[6], Vipin Kumar[6], Blessy Jose[6], Vidur Mahajan[6], Vivek Gupta[7], Rakesh Gupta[7], Prabhakar Arumugam[1,2], Anjali Singh[1,2], Ananya Nandy[1,2], P.V. Raghavendran[1,2], Rakesh Mohan Jha[1,2], Anupama Kumari[1,2], Sheetal Gandotra[1,2], Vivek Rao[1,2], Mohammed Faruq[1,2], Sanjeev Kumar[1,2], Betsy Reshma G[1,2], Narendra Varma G[1], Shuvra Shekhar Roy[1,2], Antara Sengupta[1,2], Sabyasachi Chattopadhyay[1,2], Khushboo Singhal[1,2], Shalini Pradhan[1], Diksha Jha[1,2], Salwa Naushin[1,2], Saruchi Wadhwa[1,2], Nishu Tyagi[1,2], Mukta Poojary[1,2], Vinod Scaria[1,2]*, Sridhar Sivasubbu[1,2]*

[1]CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110025, India

[2]Academy for Scientific and Innovative Research, Human Resource Development Centre Campus, Ghaziabad, Uttar Pradesh, India

[3]IMS and SUM Hospital, Siksha "O" Anusandhan (Deemed to be University), Kalinga Nagar, Bhubaneswar, Odisha 751003, India

[4]Sawai Man Singh Medical College, Jaipur, Rajasthan 302004, India

[5]Kurnool Medical College, Kurnool, Andhra Pradesh 518002, India

[6]Center for Advanced Research in Imaging, Neuroscience & Genomics, Ring Road, Defence Colony, Delhi, 110024, India

[7]Government Institute of Medical Sciences, NOIDA, India

Corresponding authors

Sridhar Sivasubbu- s.sivasubbu@igib.res.in

Vinod Scaria- vinods@igib.res.in

Abstract

The rapid emergence of coronavirus disease 2019 (COVID-19) as a global pandemic affecting millions of individuals globally has necessitated sensitive and high-throughput approaches for the diagnosis, surveillance and for determining the genetic epidemiology of SARS-CoV-2. In the present study, we used the COVIDSeq protocol, which involves multiplex-PCR, barcoding and sequencing of samples for high-throughput detection and deciphering the genetic epidemiology of SARS-CoV-2. We used the approach on 752 clinical samples in duplicates, amounting to a total of 1536 samples which could be sequenced on a single S4 sequencing flow cell on NovaSeq 6000. Our analysis suggests a high concordance between technical duplicates and a high concordance of detection of SARS-CoV-2 between the COVIDSeq as well as RT-PCR approaches. An in-depth analysis revealed a total of six samples in which COVIDSeq detected SARS-CoV-2 in high confidence which were negative in RT-PCR. Additionally, the assay could detect SARS-CoV-2 in 21 samples and 16 samples which were classified inconclusive and pan-sarbeco positive respectively suggesting that COVIDSeq could be used as a confirmatory test. The sequencing approach also enabled insights into the evolution and genetic epidemiology of the SARS-CoV-2 samples. The samples were classified into a total of 3 clades. This study reports two lineages B.1.112 and B.1.99 for the first time in India. This study also revealed 1,143 unique single nucleotide variants and added a total of 73 novel variants identified for the first time. To the best of our knowledge, this is the first report of the COVIDSeq approach for detection and genetic epidemiology of SARS-CoV-2. Our analysis suggests that COVIDSeq could be a potential high sensitivity assay for detection of SARS-CoV-2, with an additional advantage of enabling genetic epidemiology of SARS-CoV-2.

Keywords:

COVID-19, COVIDSeq, Next Generation Sequencing, Amplicon, SARS-CoV-2, Genetic Epidemiology, Diagnosis

Introduction

Coronavirus Disease 2019 (COVID-19) has emerged as a global epidemic affecting millions of individuals globally and imposes a huge burden on the socio-economic welfare and healthcare systems of nations. At present the need for assays for rapid detection for diagnosis and surveillance, understanding the genetic epidemiology and evolution of the virus would be central for managing the spread of the epidemic (J. Lu *et al.*, 2020; Meredith *et al.*, 2020). The advantage of quick sequencing of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome led to the development of polymerase chain reaction (PCR) based diagnostic assays that leveraged rapid identification of infected individuals to get fast medical support or quantization essential to both patient management and incidence tracking (Wu *et al.*, 2020). Identification of early imported cases in France helped to prevent immediate secondary transmission (Bernard Stoecklin *et al.*, 2020)**.** Singapore's enhanced surveillance and containment strategy also led to the suppressed expansion of SARS-CoV-2 (de Lusignan *et al.*, 2020)**.** On similar grounds, The Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) have rapidly expanded their national surveillance system to combat SARS-CoV-2 (Ng *et al.*, 2020). Coupled with a highly accurate and high-throughput method of detection, this approach will become more effective in dealing with COVID-19.

A number of approaches have been widely used for the detection of SARS-CoV-2 from clinical samples. Some of these approaches have also been adapted to enable higher throughputs. These methods are majorly subdivided into antigen-antibody based serological assays, nucleic acid based amplification assays and sequencing based assays. While serological assays are rapid detection tests, they have low sensitivity and specificity (Döhla *et al.*, 2020). Nucleic acid based amplification such as Real-Time PCR (RT-PCR) has been the gold standard in detection and diagnosis, but a negative RT-PCR does not eliminate the possibility of infection in clinically suspected cases (Wang *et al.*, 2020). Such results should be carefully interpreted to avoid false negative reporting (Kucirka *et al.*, 2020). Moreover, these tests have been developed for diagnostic purposes and do not provide much information on the nature of the virus, its genetic information and evolutionary pattern. In this regard, recently developed Next generation sequencing based methods are potentially a good alternative for detection of SARS-CoV-2 ('First NGS-based COVID-19 diagnostic', 2020).

The rapid advancement of Next generation sequencing technology and analysis methods has enabled understanding the genetic makeup of SARS-CoV-2 and interpreting its evolutionary epidemiology. Viral RNA sequencing from the initial cluster of cases deciphered the full genome sequence of SARS-CoV-2 (Zhu *et al.*, 2020). This led to other sequencing based studies for detailed genomic characterization of the virus (R. Lu *et al.*, 2020). Combined genetic and epidemiological studies have been suggested to provide insights into spread of the infection, evolutionary patterns and genetic diversity of the virus (J. Lu *et al.*, 2020; Meredith *et al.*, 2020), for further assisting in effective management and preventive measures. Genomic surveillance coupled with agent-based modelling in Australia has been observed as an excellent approach to investigate and regulate COVID-19 transmission (Rockett *et al.*, 2020). Towards these efforts, several openly available databases have also been developed such as the Global Initiative on Sharing All Influenza Data (GISAID) that facilitates rapid and open sharing of SARS-CoV-2 genome sequences (Shu and McCauley, 2017). Thus, along with detection, sequencing based methods may also provide an added advantage of understanding the genetic epidemiology of the outbreak.

In the present study, we describe the application of the COVIDSeq protocol recently approved by the US FDA for clinical use. This protocol envisages high-throughput detection and genetic epidemiology of SARS-CoV-2 isolates using a multiplex PCR amplicon based enrichment followed by barcoding with a throughput of 1536 samples in a single sequencing run using NovaSeq S4 flow cell. Our analysis suggests that COVIDSeq protocol could be a sensitive approach for detection with additional insights offered through genetic epidemiology with respect to the genetic lineages. To the best of our knowledge this is the first real-life evaluation of COVIDSeq protocol.

Materials and Methods

Patients and Samples

The study was approved by the Institutional Human Ethics Committee (IHEC No. Dated CSIR-IGIB/IHEC/2020-21/01). 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 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 QIAamp® Viral RNA Mini kit (QIAGEN) as per the manufacturer's instructions. For 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 Nextractor® NX-48S instrument (Genolution Inc.) as instructed by the manufacturer. After 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 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 method.

Real Time PCR for SARS-CoV-2

To detect SARS-CoV-2 viral infection, one-step Real-Time PCR assay was performed using STANDARD M nCoV Real-Time detection kit (SD Biosensor, Korea), targeting the nCoV2 specific ORF1ab (RdRp) and pan-sarbeco specific E genes on LightCycler® 480 System (Roche) and ABI 7500 Fast DX (Applied Biosystems) as per the manufacturer's instructions.

Library preparation and sequencing

The libraries were prepared using Illumina COVIDSeq protocol (Illumina Inc, USA). The first strand synthesis was carried out in Biosafety level 2 (BSL-2) plus environment following standard protocols. The synthesized cDNA was amplified using a multiplex polymerase chain reaction (PCR) protocol, producing 98 amplicons across the SARS-CoV-2 genome (https://artic.network/). The primer pool additionally had primers targeting human RNA, producing an additional 11 amplicons. The PCR amplified product was later processed for tagmentation and adapter ligation using IDT for Illumina Nextera UD Indexes Set A, B, C, D (384 indexes, 384 samples). Further enrichment and cleanup was performed as per protocols provided by the manufacturer (Illumina Inc). All samples were processed as batches in a 96-well plate that consisted of one of COVIDSeq positive control HT (CPC HT) and one no template control (NTC); these 96 libraries were pooled together in a tube. Pooled samples were quantified using Qubit 2.0 fluorometer (Invitrogen Inc.) and fragment sizes were analyzed in Agilent Fragment analyzer 5200 (Agilent Inc). The pooled library was further normalized to 4nM concentration and 25 μl of each normalized pool containing index adapter set A, B, C, and D were combined in a new microcentrifuge tube to a final concentration of 100pM and 120pM. For sequencing, pooled libraries were denatured and neutralized with 0.2N

NaOH and 400mM Tris-HCL (pH-8). Replicas of each 384 sample pools were loaded onto the S4-flow cell following NovaSeq-XP workflow as per the manufacturer's instructions (Illumina Inc). Dual indexed single end sequencing with 36bp read length was carried out on NovaSeq 6000 platform.

Data Processing

The raw data generated in binary base call (BCL) format from NovaSeq 6000 was processed using DRAGEN COVIDSeq Test Pipeline (Illumina Inc.) on the Illumina DRAGEN v3.6 Bio-IT platform as per standard protocol. The analysis involves sample sheet validation, data quality check, FASTQ generation, and SARS-CoV-2 detection when at least 5 SARS-CoV-2 probes are detected. Further samples with SARS-CoV-2 and at least 90 targets detected were processed for alignment, variant calling and consensus sequence generation.

For in-depth analysis, we additionally analysed the data using a custom pipeline. This included demultiplexing the raw data to FASTQ files using bcl2fastq (v2.20) followed by quality assessment of the FASTQ files using Trimmomatic (v0.39) (Bolger, Lohse and Usadel, 2014). An average base quality of Q30 and read length cut-off of 30 bps were used for trimming, apart from the adapter sequences. We followed a recently published protocol to perform reference-based assembly (Poojary *et al.*, 2019). As per protocol, the trimmed reads were aligned to the human reference genome (GRCh38 / hg38) and SARS-CoV-2 Wuhan-Hu-1 reference genome (NC_045512.2) using HISAT2-2.1 (Kim, Langmead and Salzberg, 2015)**.** The reads mapped to hg38 were further discarded and the unaligned reads were extracted using samtools (v 1.10) (Li *et al.*, 2009). The unaligned reads were further mapped to the Wuhan Hu-1 genome and the alignment statistics were evaluated (Wu *et al.*, 2020). The data was merged for duplicates for the variant calling and consensus sequence generation. Variant calling was performed using VarScan (v2.4.4) for samples with genome coverage greater than 99% (Koboldt *et al.*, 2009). Samtools (v 1.10) (Li *et al.*, 2009), bcftools (v 1.10.2), and seqtk (version 1.3-r114) (Shen *et al.*, 2016) were used to generate the consensus sequence. We have also evaluated the correlation coefficient with p-value < 0.01 between the duplicates total reads and genome coverage.

Annotation of Genetic Variants and Comparison with existing datasets

The variants were systematically annotated using ANNOVAR (Wang, Li and Hakonarson, 2010). Annotations on genomic loci and functional consequences of the protein were retrieved from RefSeq. Custom databases were created for annotations on functional consequences, potential immune epitopes, protein domains and evolutionary conservation scores. Genomic loci associated with common error prone sites and diagnostic primer/probe regions were manually curated and were systematically converted to datatables compatible with ANNOVAR for added annotation options. All the filtered variants were checked with other viral genomes submitted from India and worldwide. Genomes with alignment percentage of at least 99 and gap percentage < 1 were filtered as high quality. A total of 1372 high quality genomes from India out of 1888 submitted till July 28, 2020 were included in the analysis. Similarly global genomes submitted till August 07, 2020 were included, accounting to 29177 high quality genomes out of a total of 79764 genomes submitted. Details of the samples, originating and submitting laboratories are listed in **Supplementary Table 1a.** Mutation information provided by Nextstrain (Hadfield *et al.*, 2018) till August 08, 2020 was also used for comparison.

Phylogenetic Analysis

A total of 495 samples that had at least 99% genome coverage were considered for this analysis, along with the dataset of SARS-CoV-2 genomes from India deposited in GISAID. The sample names and the name of the originating and submitting institutions are listed in **Supplementary Table 1b.** We followed a previously described protocol for phylogenetic clustering (Jolly and Scaria, 2020 under review). A total of 26 COVIDSeq genomes having Ns > 5% were removed from the analysis. Genomes from GISAID having Ns > 5% and ambiguous dates of sample collection were also excluded from the analysis. The phylogenetic network was built using the analysis protocol for SARS-CoV-2 genomes provided by Nextstrain (Hadfield *et al.*, 2018). The genome sequences were aligned using MAFFT to the reference genome and problematic variant positions were masked (Katoh and Toh, 2008). A raw phylogenetic tree was constructed using IQTREE and the raw tree was refined to construct a molecular-clock phylogeny, infer mutations, and identify clades (Nguyen *et al*., 2015). The resulting phylogenetic tree was viewed using Auspice, an interactive visualization web-application provided by Nextstrain. Lineages were also assigned to the genomes using the Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) package (Rambaut *et al*., 2020). The phylogenetic distribution of the lineages was visualized and annotated using iToL (Letunic and Bork, 2016)**.**

Comparison of RT-PCR test with the sequencing based COVIDSeq test

Initially, all the samples underwent RT-PCR based screening for the presence of SARS-CoV-2 RNA. Out of these 752 samples, 655 (87.1%) samples were RT-PCR positive, 43 (5.7%) were pan-sarbeco, 35 (4.6%) were inconclusive and 19 (2.5%) were negative. We compared the sample type (e.g. positive, pan-sarbeco, inconclusive and negative) WGS output and calculated percent of genome covered, sensitivity, specificity, accuracy, precision and gain of detection rate. The methodology adopted in this study has been represented in **Figure 1.**

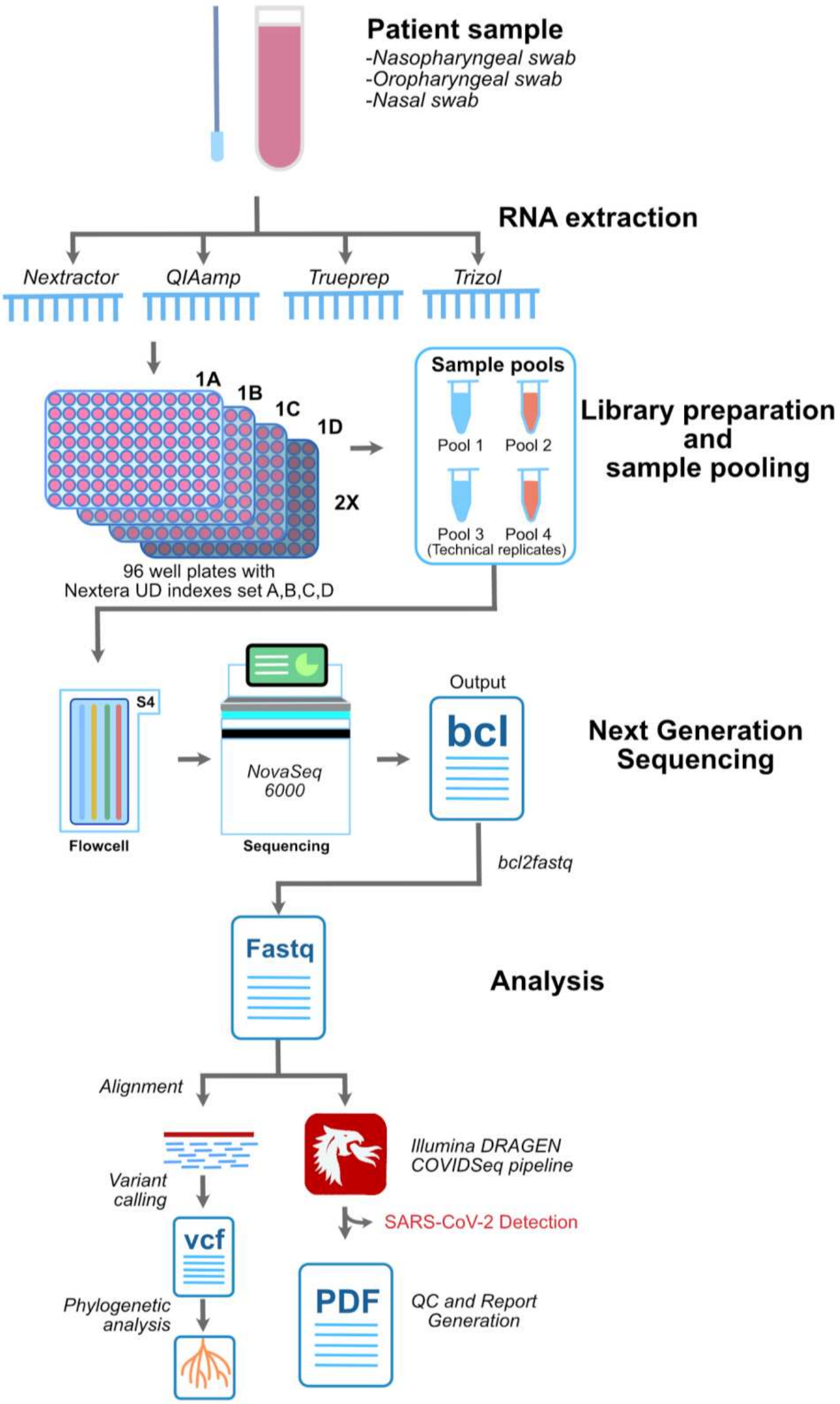

Figure 1. Schematic summarising the sampling, library preparation, sequencing and analysis employed in this study.

Results

The sample panel consisted of a total of 752 samples. Among these, 655 (87.1%) were SARS-CoV-2 positive on RT-PCR as per the diagnostic guidelines laid out by the Indian Council for Medical Research (ICMR). We included 19 samples that were RT-PCR negative (2.5%) and 43 samples (5.7%) were categorised as pan-sarbeco, since they were positive for the E gene primers only. A total of 35 samples (4.6%) were considered inconclusive as the samples had one of the two genes (i.e. ORF1ab gene) tested positive. Apart from this, the sequencing panel consisted of 8 CPC HT and 8 NTC as internal process controls making total samples to 768. The quality of the pooled library was checked by agarose gel electrophoresis and fragment analyzer which showed the fragment size to be around 300bp. The panel was sequenced in technical duplicates making it a total of 1536 samples in total. The sequencing was performed for 36 cycles. The runtime of the sequencer was 11 hours. Sequencing generated a total of 705.64 Gb of data with 86.90% cluster passing filter and 95.62% above the quality cutoff of QC30. Sequencing generated on an average of approximately 8.4 million reads for the 1,536 samples.

The COVID-19 detection was performed using the DRAGEN COVIDSeq Test pipeline that implements SARS-CoV-2 detection criteria of at least 5 SARS-CoV-2 targets to be considered as positive. Out of the 1,504 samples, DRAGEN COVIDSeq Test pipeline successfully annotated 1,352 samples. Further 136 samples were classified as undetected, and 16 failed the internal quality check. This corresponds to 676 unique samples in which SARS-CoV-2 was detected, 68 unique samples in which SARS-CoV2 was undetected and 8 unique samples which failed the assay. There was no discordance in the annotations between any of the 752 sample duplicates considered, suggesting a cent percent concordance in the detection. The total runtime for the DRAGEN COVIDSeq pipeline was 374 minutes. The stepwise runtime is summarised in **Supplementary Table 2.**

All samples were also further considered for in-depth alignment and on average 8.4 million raw reads were generated for 1,536 samples, which were trimmed at base quality Q30 and read length of 30 bps that lead to an average of 7.9 million reads. The trimmed reads were further aligned to the human reference genome (GRCh38/hg38) and SARS-CoV-2 genome (NC_045512.2). On an average we found 2.4 million human reads with mapping percentage of 30.73% and 5.04 million SARS-CoV-2 reads with mapping percentage of 63.89% respectively. The unmapped reads from the human aligned files were extracted and mapped to the SARS-CoV-2 reference genome (NC_045512.2) to increase its specificity and 4.4 million such reads (79.34%) mapped to it with 6322X coverage. **Figure 2** summarises the concordance of aligning reads as well as genome coverage across the duplicates. The data has been summarized in the **Supplementary Table 3.**

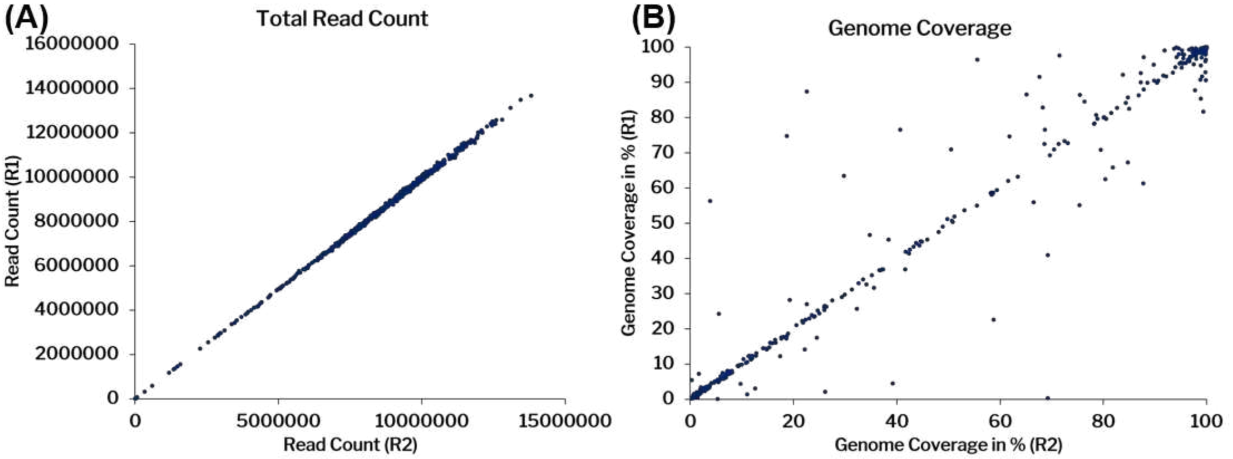

Figure 2. Concordance in the aligning reads (A) and coverage (B) across the replicate samples considered in the analysis.

The mean coverage was also plotted for all the samples across 98 PCR amplicons covering the whole SARS-CoV-2 genome represented in **Figure 3.** The mean coverage across the amplicons was ~14256x for the positive samples considered (706 samples with genome coverage > 5%). We have found 20 amplicons had coverage ±2 standard deviations (SD) of this value, out of which 16 amplicons had coverage < 2 SD and 4 amplicons had > 2 SD.

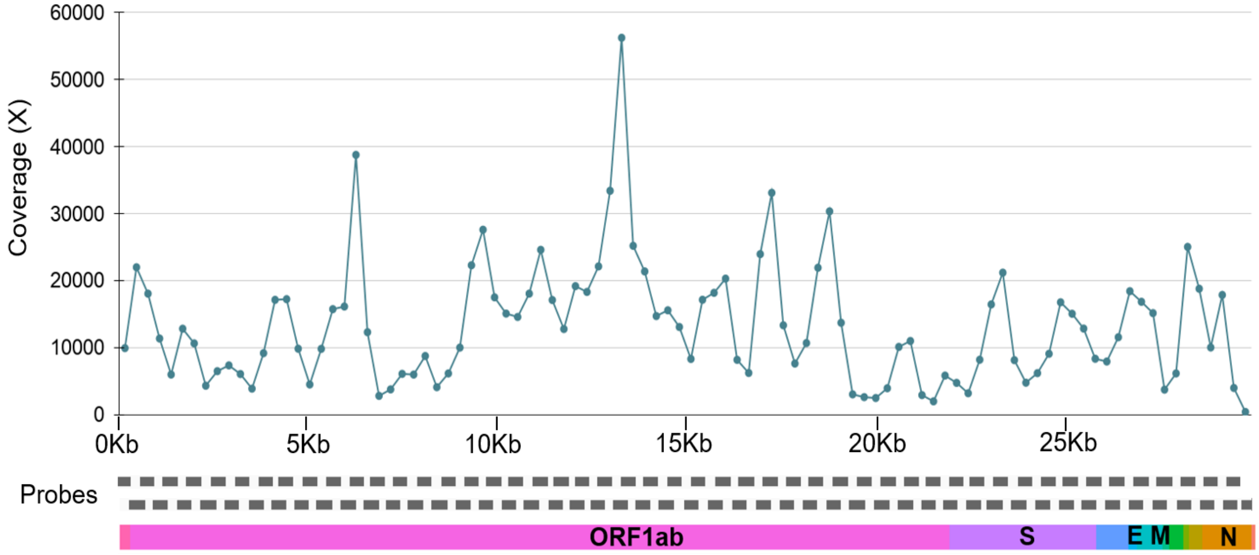

Figure 3. Mean coverage for the amplicons across the SARS-CoV-2 genome.

The technical duplicates had a correlation coefficient of 0.99 (p-value < 0.00001) for reads and 0.984 (p-value < 0.00001) for the coverage.

For further genome assembly and variant calling, the alignment files were merged and variants were called using VarScan. Only 495 samples that had at least 99% of the genome covered were considered for variant call.

The analysis identified a total of 1,143 unique variants. 73 genetic variants were found to be novel in comparison with other Indian and global genome data and were reported for the first time. The median for the number of variants called were 12. The distribution of the variants per genome is summarised in **Figure 4A**. Of the 1,143 unique variants, a total of 1,104 variants were in the exonic region and 39 were in the downstream or upstream region. Of the 1,104 exonic variants, 639 variants were non-synonymous while 452 were synonymous. A total of 13 were found to be stopgain. The variant annotation data is summarized in **Figure 4B.**

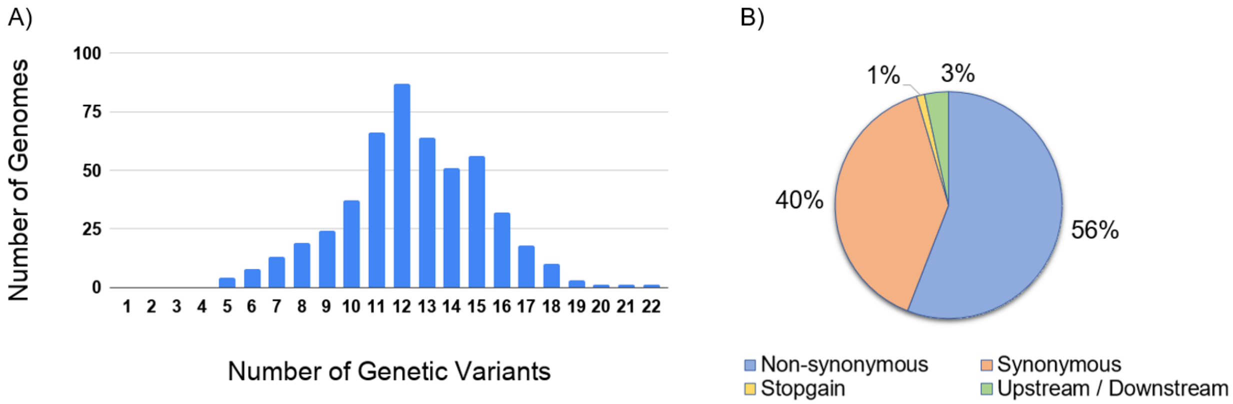

Figure 4. Variant number per genome and their annotation (A) Distribution of variants in the genomes with ≥ 99% coverage (B) Summary of the variant annotations.

Analysis of frequency of the variants across the genomes revealed a total of 89 variants that had a frequency ≥ 1% and were polymorphic. The variants were also mapped across the genes. The ORF1ab gene had the largest number of variants. Normalised for the length of the genes, ORF3a gene had the highest number. Similarly for non-synonymous variants ORF1ab gene had the highest number of variants and ORF3a had the highest normalized for the length of the gene.

To get an insight into the genetic epidemiology, the genomes were analyzed for their phylogenetic distribution. Phylogenetic reconstruction was done for 2193 genomes, including 469 genomes from this study and samples previously sequenced from Indian laboratories. The genome Wuhan/WH01 (EPI_ISL_406798) was used as the reference for constructing the tree. The resulting phylogenetic tree suggests that out of 469 COVIDSeq genomes, 451 genomes (96%) fell into the A2a clade while 14 genomes (3%) mapped to the I/A3i clade. A total of 4 genomes mapped to the B4 clade. The phylogenetic clusters for the genomes are summarised in **Figure 5.** The distribution of lineages assigned by PANGOLIN suggests a dominant occurrence of the lineages B.1 (n=286) and B.1.113 (n=134) as compared to other Indian genomes which show a dominance of B.6 and B.1 lineages. We also found 2 lineages in our dataset, B.1.112 (n=8) and B.1.99 (n=1), which have not been previously reported from India. **Figure 6** summarises the phylogenetic distribution of the lineages.

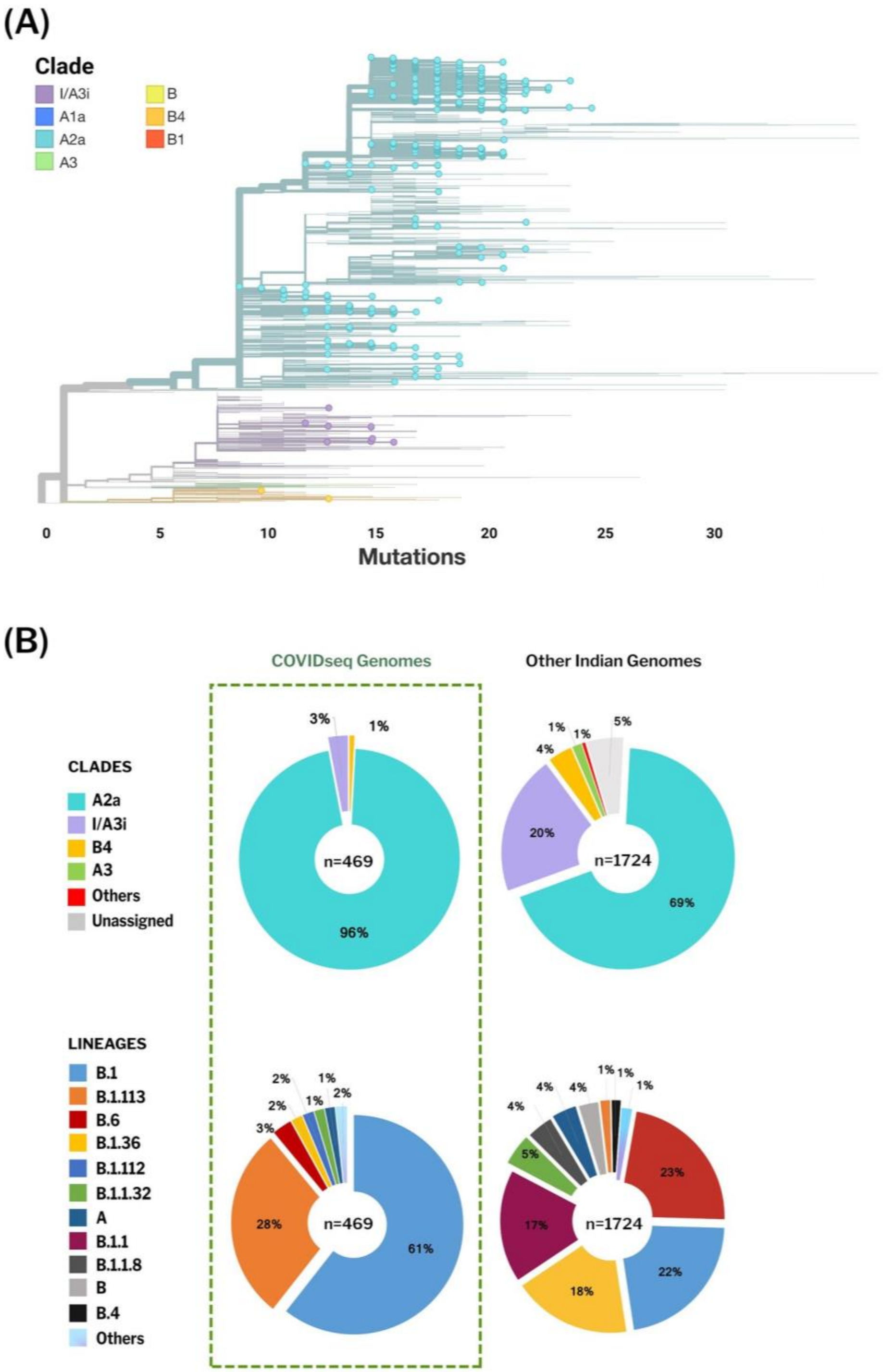

Figure 5. Phylogenetic map of Indian SARS-CoV-2 genomes. 469 genomes reported from this study are highlighted (A) and the proportion of the clades and lineages representing the genomes (B)

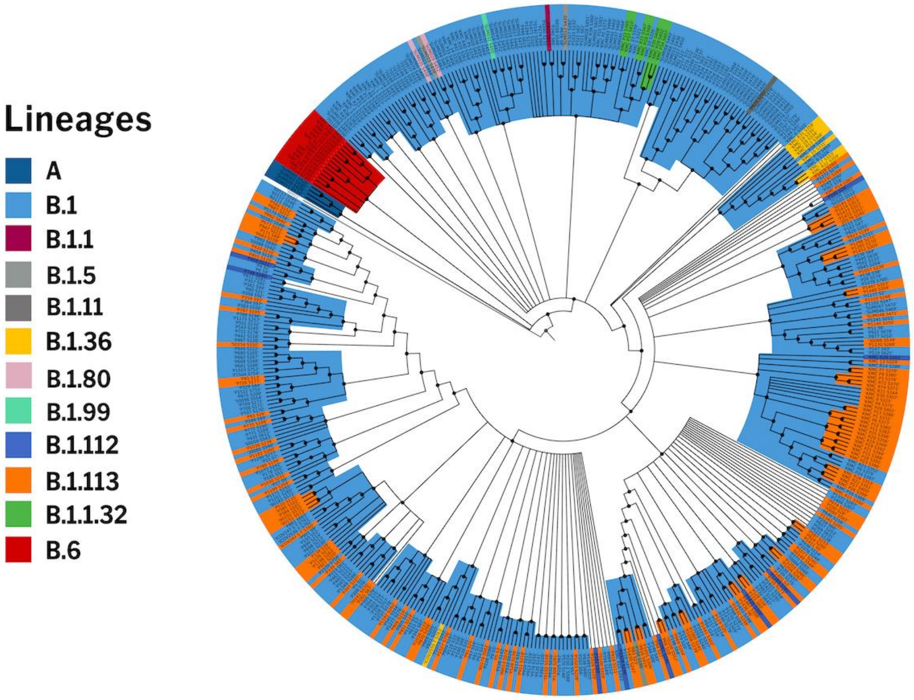

Figure 6. Phylogenetic distribution of lineages as annotated by PANGOLIN

The sensitivity of the assay was benchmarked across 649 RT-PCR as well as COVIDSeq confirmed dataset. Analysis revealed the assay had a sensitivity of 97.53% compared to RT-PCR. Since we had only 19 RT-PCR negative samples, we did not assess the specificity of the assay. The comparison of RT-PCR with COVIDSeq assay has been summarized in **Supplementary 4a**. Notwithstanding, the DRAGEN COVIDSeq protocol identified SARS-CoV-2 in the samples which were negative for RT-PCR for SARS-CoV-2. Additionally, SARS-CoV-2 was detected by the protocol in 21 samples which were inconclusive and 16 samples which were annotated pan-sarbeco. We further analysed these samples in great detail to check whether multiple genomic regions were covered in the sequencing experiments. **Figure 7** summarises the coverage plots across the genome for the 6 samples which were negative in RT-PCR and detected by COVIDSeq pipeline. The coverage plots for the samples which were inconclusive and pan-sarbeco in RT-PCR were detected by COVIDSeq assay represented in **Supplementary Figure 1**.

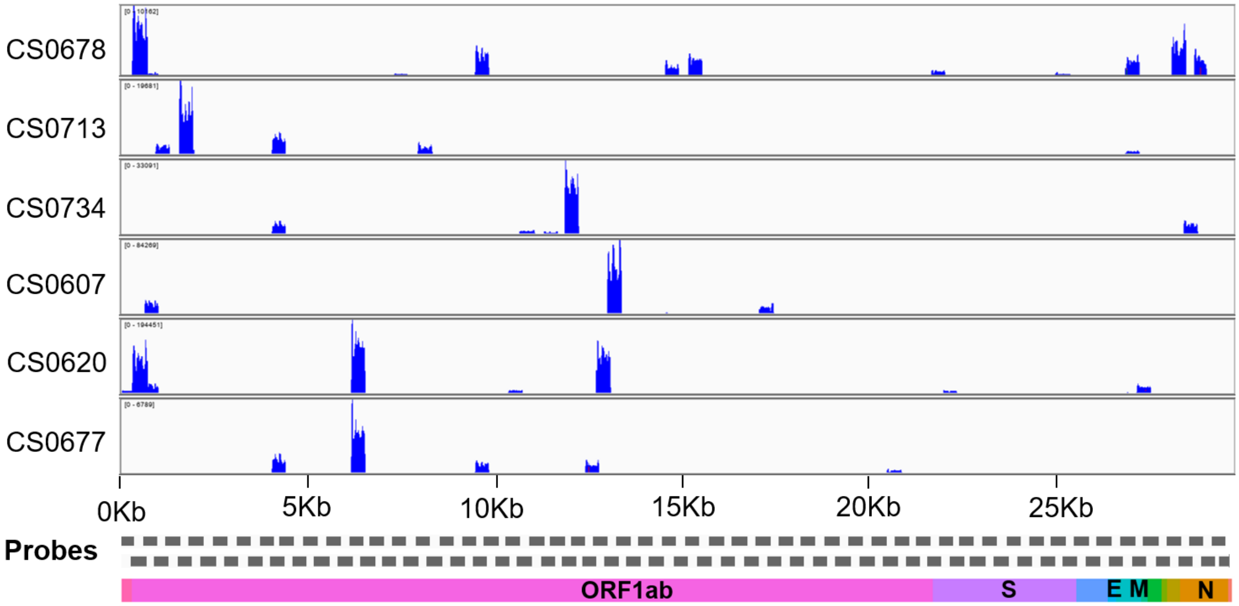

Figure 7. Coverage plot across the genome for samples which were negative on RT-PCR assays in which the DRAGEN COVIDSeq Pipeline detected SARS-CoV-2.

Consistently, the samples have over 5% of the genomic region covered in the COVIDSeq protocol suggesting that the protocol could provide for a potentially more sensitive detection assay compared to RT-PCR.

Since the RNA samples were derived from multiple protocols for RNA extraction, we could also get an insight into the compatibility of the protocols with the COVIDSeq test. Of the samples which were SARS-CoV-2 positive on RT-PCR, 182 samples were processed using QIAamp® Viral RNA Mini kit, 264 samples using Nextractor® NX-48S (Genolution, Korea), 201 samples on Trueprep® AUTO v2 (Molbio Diagnostics Pvt. Ltd.) and 8 samples using TRIzol based extraction method. Of these, COVIDSeq detected SARS-CoV-2 in 168 of 182 samples (92.3%) from the QIAamp® extracted samples, 263 of 264 samples (99.6%) from the bead based automated method using Nextractor® NX-48S, 194 of 201 samples (96.5%) from the Truprep extraction method and 7 of 8 samples (87.5%) from TRIzol based method suggesting compatibility with different extraction methods. The details of the RT-PCR and COVIDSeq samples are summarized in **Supplementary Table 4b**.

Discussion and Conclusions

A number of high-throughput approaches have recently been employed for the detection as well as sequencing of SARS-CoV-2, while RT-PCR based approaches are widely considered as the gold-standard for detection. These include shotgun approaches (Harilal *et al.*, 2020), capture based (Wen *et al.*, 2020; Xiao *et al.*, 2020) as well as amplicon based (Baker *et al.*, 2020) approaches followed by Next Generation Sequencing. Typically the multiplex barcoded library sequencing has been implemented for sample numbers less than 96. A number of approaches have been suggested to increase the throughput of sequencing using barcoded libraries (Palmieri *et al.*, 2020; Schmid-Burgk *et al.*, 2020). There is a paucity of data on higher order multiplex barcoding and sequencing approaches in clinical samples.

In the present report, we evaluated the COVIDSeq approach for high-throughput detection of SARS-CoV-2 which uses multiplex PCR followed by barcoded libraries and sequencing on a next-generation sequencing platform which envisages sequencing 1536 samples per flow cell. We analysed 752 clinical samples in technical duplicates.

Our analysis suggests a high concordance between technical duplicates and a high concordance of detection of SARS-CoV-2 between the COVIDSeq as well as RT-PCR approaches. Our comparative analysis of SARS-CoV-2 detection with RT-PCR and COVIDSeq test showed that COVIDSeq test outperformed with increased sensitivity, precision and accuracy. COVIDSeq protocol detected SARS-CoV-2 in samples previously categorised as inconclusive (21/35), pan-sarbeco (16/43) and negative (6/19) using RT-PCR assays suggesting a higher sensitivity of the sequencing based assay compared to RT-PCR. This corresponded to an additional 43/97 samples and a potential gain of 5.71% of samples of the whole dataset and 44.33% of the samples which were considered inconclusive (N=97), suggesting that the sequencing approach could be used as a potential orthogonal approach to confirm cases which are doubtful or inconclusive in RT-PCR. Notwithstanding the advantage, 16 samples which were annotated positive in RT-PCR were missed in the COVIDSeq approach. Our analysis also suggests the protocol is compatible with different approaches for RNA isolation suggesting a wider applicability in clinical settings where pooling from different labs becomes inevitable.

The COVIDSeq approach additionally provided insights into the genetic epidemiology and evolution of the SARS-CoV-2 isolates. Phylogenetic analysis could be performed for a significantly large number of genomes which gave insights into the prevalent lineage/clades of the virus (Langat *et al.*, 2017; Michie *et al.*, 2020; Shakya *et al.*, 2020). This analysis also reports two lineages B.1.112 and B.1.99 for the first time in India.

Furthermore, a total of 1,143 unique variants were contributed by this analysis to the global repertoire of genetic variants. As expected, a significant number of variants were non-synonymous in nature (Kryazhimskiy, Bazykin and Dushoff, 2008; Tang *et al.*, 2020). The present analysis adds a total of 73 novel variants identified for the first time in genomes. Apart from the throughput of sample analysis, the COVIDSeq approach is also remarkable in terms of speed, with a sequencing time of 11 hours and analysis timeline of 6 hours. Given that the NovaSeq 6000 sequencer used in the present study can handle two S4 flow cells in parallel, this could be potentially scaled to a throughput of 1536x2 samples that can be handled in parallel.

In conclusion, our analysis suggests that COVIDSeq is a high-throughput sequencing based approach which is sensitive for detection of SARS-CoV-2. In addition, COVIDSeq has an additional advantage of enabling genetic epidemiology of SARS-CoV-2.

Acknowledgements

Authors acknowledge Samatha Mathew for reviewing the manuscript and providing constructive criticism. Authors acknowledge Vigneshwar Senthilvel for creating the graphical abstract of the COVIDSeq pipeline. AJ, MD, BJ, PS acknowledges research fellowships from CSIR and MP acknowledges a research fellowship from BINC-DBT. DS acknowledges a research fellowship from Intel. The authors SS, MS, AN acknowledge the SOAU for their PhD fellowship.

Funding

Authors acknowledge funding for the work from the Council of Scientific and Industrial Research (CSIR), India through grants CODEST and MLP2005. The funders had no role in the design of experiment, analysis or decision to publish.

Conflicts of Interest

Authors declare no conflicts of interest.

Data availability

Raw datasets are available at NCBI short Read Archive with Project ID PRJNA655577. The replicate datasets are accessible with IDs SUB7891477 and SUB7864921.

References

Baker, D. J. *et al.* (2020) ‘CoronaHiT: large scale multiplexing of SARS-CoV-2 genomes using Nanopore sequencing’. doi: 10.1101/2020.06.24.162156.

Bernard Stoecklin, S. *et al.* (2020) ‘First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020’, *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 25(6). doi: 10.2807/1560-7917.ES.2020.25.6.2000094.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics* , 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bani, J, and Scaria V (2020) 'Computational Analysis and Phylogenetic clustering of SARS-nCov-2 genomes'. Available at: https://docs.google.com/document/d/1B5NxWFwsRz_vD5Y6EwjKxkRamsPLVfs1MjVoziU1Zq0/edit?usp=embed_facebook (Accessed: 4 August 2020).

Döhla, M. *et al.* (2020) ‘Rapid point-of-care testing for SARS-CoV-2 in a community screening setting shows low sensitivity’, *Public health*, 182, pp. 170–172. doi:

10.1016/j.puhe.2020.04.009.

'First NGS-based COVID-19 diagnostic' (2020) *Nature biotechnology*, 38(7), p. 777. doi: 10.1038/s41587-020-0608-y.

Hadfield, J. *et al.* (2018) 'Nextstrain: real-time tracking of pathogen evolution', *Bioinformatics* , 34(23), pp. 4121–4123. doi: 10.1093/bioinformatics/bty407.

Harilal, D. *et al.* (2020) 'SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission', *Clinical chemistry*. doi: 10.1093/clinchem/hvaa187.

Katoh, K. and Toh, H. (2008) 'Recent developments in the MAFFT multiple sequence alignment program', *Briefings in bioinformatics*, 9(4), pp. 286–298. doi: 10.1093/bib/bbn013.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: a fast spliced aligner with low memory requirements', *Nature methods*, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

Koboldt, D. C. *et al.* (2009) 'VarScan: variant detection in massively parallel sequencing of individual and pooled samples', *Bioinformatics* , 25(17), pp. 2283–2285. doi: 10.1093/bioinformatics/btp373.

Kryazhimskiy, S., Bazykin, G. A. and Dushoff, J. (2008) 'Natural selection for nucleotide usage at synonymous and nonsynonymous sites in influenza A virus genes', *Journal of virology*, 82(10), pp. 4938–4945. doi: 10.1128/JVI.02415-07.

Kucirka, L. M. *et al.* (2020) 'Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure', *Annals of internal medicine*. doi: 10.7326/M20-1495.

Langat, P. *et al.* (2017) 'Genome-wide evolutionary dynamics of influenza B viruses on a global scale', *PLoS pathogens*, 13(12), p. e1006749. doi: 10.1371/journal.ppat.1006749.

Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic acids research*, 44(W1), pp. W242–5. doi: 10.1093/nar/gkw290.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* , 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Lu, J. *et al.* (2020) 'Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China', *Cell*, 181(5), pp. 997–1003.e9. doi: 10.1016/j.cell.2020.04.023.

Lu, R. *et al.* (2020) 'Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding', *The Lancet*, 395(10224), pp. 565–574. doi: 10.1016/S0140-6736(20)30251-8.

de Lusignan, S. *et al.* (2020) 'Emergence of a Novel Coronavirus (COVID-19): Protocol for Extending Surveillance Used by the Royal College of General Practitioners Research and Surveillance Centre and Public Health England', *JMIR public health and surveillance*, 6(2), p. e18606. doi: 10.2196/18606.

Meredith, L. W. *et al.* (2020) 'Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study', *The Lancet infectious diseases*. doi: 10.1016/S1473-3099(20)30562-4.

Michie, A. *et al.* (2020) 'Genome-Scale Phylogeny and Evolutionary Analysis of Ross River Virus Reveals Periodic Sweeps of Lineage Dominance in Western Australia, 1977-2014', *Journal of virology*, 94(2). doi: 10.1128/JVI.01234-19.

Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular biology and evolution*, 32(1), pp. 268–274. doi: 10.1093/molbev/msu300.

Ng, Y. *et al.* (2020) 'Evaluation of the Effectiveness of Surveillance and Containment Measures for the First 100 Patients with COVID-19 in Singapore - January 2-February 29, 2020', *MMWR. Morbidity and mortality weekly report*, 69(11), pp. 307–311. doi: 10.15585/mmwr.mm6911e1.

Palmieri, D. *et al.* (2020) ‘REMBRANDT: A high-throughput barcoded sequencing approach for COVID-19 screening’, *Molecular Biology*. bioRxiv.

Poojary, M. *et al.* (2019) ‘Computational Protocol for Assembly and Analysis of SARS-nCoV-2 Genomes’, *Research Reports*. Available at: http://www.companyofscientists.com/index.php/rr/article/view/165 (Accessed: 2 August 2020).

Rambaut, A. *et al.* (2020) ‘A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology’, *Nature microbiology*. doi: 10.1038/s41564-020-0770-5.

Rockett, R. J. *et al.* (2020) ‘Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling’, *Nature medicine*. doi: 10.1038/s41591-020-1000-7.

Schmid-Burgk, J. L. *et al.* (2020) ‘LAMP-Seq: Population-Scale COVID-19 Diagnostics Using Combinatorial Barcoding’, *Molecular Biology*. bioRxiv.

Shakya, M. *et al.* (2020) ‘Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life’, *Scientific reports*, 10(1), p. 1723. doi: 10.1038/s41598-020-58356-1.

Shen, W. *et al.* (2016) ‘SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation’, *PloS one*, 11(10), p. e0163962. doi: 10.1371/journal.pone.0163962.

Shu, Y. and McCauley, J. (2017) ‘GISAID: Global initiative on sharing all influenza data - from vision to reality’, *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 22(13). doi: 10.2807/1560-7917.ES.2017.22.13.30494.

Tang, X. *et al.* (2020) ‘On the origin and continuing evolution of SARS-CoV-2’, *National Science Review*, 7(6), pp. 1012–1023. doi: 10.1093/nsr/nwaa036.

Wang, K., Li, M. and Hakonarson, H. (2010) ‘ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data’, *Nucleic acids research*, 38(16),

p. e164. doi: 10.1093/nar/gkq603.

Wang, Y. *et al.* (2020) ‘Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak’, *Journal of medical virology*, 92(6), pp. 538–539. doi: 10.1002/jmv.25721.

Wen, S. *et al.* (2020) ‘High-coverage SARS-CoV-2 genome sequences acquired by target capture sequencing’, *Journal of medical virology*. doi: 10.1002/jmv.26116.

Wu, F. *et al.* (2020) ‘A new coronavirus associated with human respiratory disease in China’, *Nature*, 579(7798), pp. 265–269. doi: 10.1038/s41586-020-2008-3.

Xiao, M. *et al.* (2020) ‘Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples’, *Genome medicine*, 12(1), p. 57. doi: 10.1186/s13073-020-00751-4.

Zhu, N. *et al.* (2020) ‘A Novel Coronavirus from Patients with Pneumonia in China, 2019’, *The New England journal of medicine*, 382(8), pp. 727–733. doi: 10.1056/NEJMoa2001017.

Supplementary Data

Supplementary Table 1a: GISAID acknowledgement table for global genomes used in the study.

Supplementary Table 1b: GISAID acknowledgement table for Indian genomes used in the study.

Supplementary Table 2: DRAGEN COVIDSeq Test Pipeline time summary for each task.

Supplementary Table 3: Data summary of the COVIDSeq, RT-PCR and custom pipeline analysis. NA- Not Applicable.

Supplementary Table 4a: Summary of the COVIDSeq assay comparison with RT-PCR.

Supplementary Table 4b: Comparison of different RNA extraction methods and detection of the SARS-CoV-2 with RT-PCR and COVIDSeq test.

Supplementary Figure 1: Coverage plots for 37 samples that were inconclusive and pan-sarbeco by RT-PCR and detected positive for SARS-CoV-2 by sequencing**.**