Research Article

iMedPub Journals
www.imedpub.com

Journal of Molecular Microbiology

2022
Vol.6 No.6:001

Rapid Lineage Assignment and Phylogenetic Tracking of SARS-CoV-2 Cases through Automated Library Preparation, Sequencing and Bioinformatics Analysis

Mark Pandori[1*], Andrew J Gorzalski[1], Irina St Louis[1], Danielle Siao[1], Lauren Siao[1], Diego Bunuel[1], Stephanie Van Hooser[1], David C Hess[1], Heather Kerwin[2], Joel Sevinsky[2], Kevin Libuit[2] and Suhash Verma[2]

[1]Department of Pathology and Laboratory Medicine, University of Nevada, Reno, Nevada, US

[2]Department of Microbiology and Immunology, University of Nevada, Reno, Nevada, US

[*]**Corresponding author:** Mark Pandori, Department of Pathology and Laboratory Medicine, University of Nevada, Reno, NV, US, Tel: 4156329183; E-mail: MPANDORI@MED.UNR.EDU

Received date: July 28, 2022, Manuscript No. IPMM-22-14176; **Editor assigned date:** August 01, 2022, PreQC No. IPMM-22-14176 (PQ); **Reviewed date:** August 16, 2022, QC No. IPMM-22-14176**; Revised date:** September 29, 2022, Manuscript No. IPMM-22-14176 (R); **Published date:** October 07, 2022, DOI: 10.36648/IPMM.6.6.001

Citation: Pandori M, Gorzalski AJ, Louis IS, Siao D, Siao L, et al. (2022) Rapid Lineage Assignment and Phylogenetic Tracking of SARS-COV-2 Cases through Automated Library Preparation, Sequencing and Bioinformatic Analysis. J Mol Microbiol. Vol:6 No:6

Abstract

The COVID-19 pandemic has provided a stage to illustrate that there is considerable value in obtaining rapid, whole genomic information about pathogens. Herein we describe the utility of automated SARS-CoV-2 library preparation, genomic sequencing and a bioinformatic analysis pipeline to provide rapid, near "real-time" SARS-CoV-2 variant description. We evaluated the turnaround time, accuracy and other quality parameters obtained from clear labs Dx automated sequencing instrumentation from analysis of continuous clinical samples from January 1, 2021 to October 6, 2021. Additionally, we illustrate instances where near real-time analysis provided intelligence relevant to concurrent disease control investigations.

Keywords: SARS-CoV-2; Sequencing; Bioinformatics; Phylogenetic; Disease control

Introduction

Genomic sequencing of infectious agents provides data that describes organisms at the Single Nucleotide Polymorphic (SNP) matrix, the highest level of discriminatory capability. For epidemiologic investigations, SNP level discrepancies can provide powerful intelligence in the determination of the relatedness (phylogenetic) of cases. Barriers have existed to having sequence and phylogenetic information concurrent to the investigational process: Sequencing itself is a laboratory process of significant complexity that can require extensive hands on time, followed by lengthy sequencing processes. When raw sequence data is obtained it is initially unusable and requires computational processes prior to providing utility for epidemiologists and disease control investigators. These processes can be arcane and time consuming, lengthening further the time between specimen collection and the moment of utility in an investigation [1].

Contact tracing efforts can stem transmission networks and this work is well-supported by phylogenetic information [2,3]. Early detection of pathogen variants can allow a more rational and effective channeling of public health resources [4]. It can be reasoned that significantly reducing the time from case detection to case description through sequencing analysis would impact COVID-19 disease control efforts. We sought to construct a system of SARS-CoV-2 sequencing and data collection and analysis that provides rapid, near "real time" phylogenetic intelligence. This system was used to augment surveillance for the state of Nevada and was found to offer significant impact to disease control investigation and to variant detection.

Materials and Methods

SARS-CoV-2 RNA isolation and detection through RT-PCR analysis

RT-PCR was performed using the CDC influenza SARS-CoV-2 multiplex assay. Nucleic acid extractions were performed by apostle mag touch nucleic acid extraction automation systems. Specimens found positive with a Ct value less than 30 were subjected to targeted sequencing. Additionally, experiments were performed to evaluate the ability to sequence specimens of higher Ct value and this cutoff was elevated to 33 based upon those results.

Sequencing

Extracted RNA samples, tested by way of the CDC influenza SARS-CoV-2 (Flu SC2) multiplex assay are diluted as follows: If Ct<15.0, then 1:1000; If Ct>=15.0 and Ct<=18.0, then 1:100; If Ct>=18.0 and Ct<23.0, then 1:10; If Ct>=23.0, no dilution required. Specimens with Ct values of 30 or above were appraised for suitability in sequencing and for part of the project, specimens with Ct values equal to or less than 33 were included. The clear lab Dx sequencing methodology is complex and lengthy and is provided as supplemental material.

Bioinformatics analysis

Analysis of SARS-CoV-2 genomic data was performed using titan clear labs and titan augur run, two Workflow Description Language (WDL) workflows within the Theiagen Public Health Viral Genomics (PHVG) Dock store collection. PHVG WDL workflows consist of freely available containerized bioinformatics algorithms, such as those hosted on the StaPH-B DockerHub repository, and were designed to run on general-purpose containerized workflow execution infrastructures including light-weight compute resources running miniwdl, local or cloud based High Performance Compute (HPC) systems with access to the cromwell engine, or various web applications that provide a graphical user interface to non-technical users, such as DNA nexus or terra. For this study, Titan clear labs and titan augur run were accessed and run using the terra platform. Source code for the PHVG WDL workflows has been made publicly available with the AGPL-3.0 license on GitHub. The titan clear labs work flow was utilized to generate consensus assemblies from raw clear labs read data and to perform SARS-CoV-2 lineage and clade designations for each sample. Briefly, human reads were removed from clear labs read data using the NCBI SRA human scrubber. De-hosted reads were then assembled as per the Artic nCoV-2019 novel coronavirus bioinformatics protocol with a modification whereby the artic minion normalize flag was adjusted to 20000 to account for clear labs sequencing depths. The resulting consensus assemblies were then analyzed with pangolin and next clade to perform lineage and clade designations, respectively [5,6]. NCBI'S VADR tool was also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

The sequence alignment file (BAM) generated by the titan clear labs workflow was visualized and manually assessed using the CLC Genomic Workbench software.

The titan augur run workflow was utilized to perform phylogenetic and cluster analysis of the SARS-CoV-2 datasets. Titan augur run executes subcommands from the Next strain Augur Toolkit through the use of a modified version of the broad institute's SARS-CoV-2_nextstrain WDL workflow to construct SARS-CoV-2 Maximum Likelihood (ML) and time trees as well as an auspice compatible JSON file for interactive visualization. Phylogenetic tree visualizations were constructed by uploading the auspice compatible JSON files generated by titan augur run to the auspice web application.

Results

From January 1, 2021 through October 6, 2021, 10,102 specimens found positive for SARS-CoV-2 were retained or obtained by the Nevada State Public Health Laboratory (NSPHL) for genomic sequencing. Specimens with the cycle threshold (Ct value) of 30 or lower (as per CDC SARS-CoV-2/Influenza A/B Multiplex Real-Time PCR) were subject to automated library preparation and sequencing on clear lab DX instrumentation as described in materials and methods (section sequencing). Ninety-one (91) exceptions to the Ct cutoff occurred whereby RNA samples with Ct between 30 and 40 were analyzed. Preparation of extracted RNA for initiation of sequencing included 60 to 90 minutes of laboratory work time per 64 specimens. This time included dilution of RNA samples and loading of specimens onto the sequencing platform. Upon completion of the procedure, sequence data (both FASTQ and FASTA formats) was available through a cloud-based system within 20 hours (after 10/15/21, 12 hours) of extracted RNA being loaded to the system. For lineage assessment and for public health utilization of the data, generated FASTQ files were manually or automatically uploaded to a specific bioinformatics computational pipeline.

Assessment of data quality and coverage

From January 1, 2021 through October 6, 2021, 10,102 extracted RNA samples from positive cases were analyzed by sequencing. We assessed the performance of the automated sequencing platform with regard to genomic coverage and depth. For input specimens with a Ct value of 30 or lower, sequencing on the clear labs Dx platform resulted in genomic assemblies with greater than or equal to 90% coverage of the reference SARS-CoV-2 genome in 70% of instances (7026/10102). Coverage of 75% of the reference genome was accomplished in 80% of instances (8091/10102). Coverage (%) of SARS-CoV-2 genome was assessed on the basis of Ct value (through Ct 30) as determined by diagnostic RT-PCR for cases overall and no significant correlation was found (R=0.0362). However, for specimens with Ct value greater than 30 (91 instances), the correlation between Ct value and coverage was more substantial (R=0.82) [7]. Specimens with Ct values greater than 30 were found to have a mean genomic coverage of 61.8% while specimens with Ct values below 30 were found to have a mean coverage of 82.1%. For specimens with Ct values greater than 30, coverage of at least 90% of the genome occurred at a frequency of 30.5%.

For specimens with 90% or greater coverage of reference genome, in all instances were equipment batch runs were maximized (32 specimens tested/run) mean depth of coverage per base was 1124x with a standard deviation of 1157x (min: 12x, max: 15,111x).

Lineage assessment

While 90% or greater coverage was accomplished in 70% of cases, lineage assessments by PANGOLIN were accomplished in 80% (6,779/8,468) of cases, omitting cases where PANGOLIN made a call of lineage "B.1" when coverage of the corresponding genome was less than 90% [8]. Such instances are complicated by an inability of PANGOLIN to make a higher granularity call rather than an accurate call of the lineage B.1 simply due to an absence of calls at particular genomic positions.

Error rate assessment

We sought to determine empirically the accuracy of long read sequencing (ONT) on the clear lab instrument by examining the performance of sequencing a synthetically constructed genome (WuHan-1, Twist Biosciences). Sequencing of the chemically synthesized genome was performed twice and the error rate observed at every sequenced base in the genome was assessed. The error rate was determined by the number of miscalls vs. the number of accurate calls. Two independent sequence tests of the synthetically constructed genome on the ONT platform included one run with an average depth of coverage of 6821x ("Run 1") and another with average read depth of 341x ("Run 2"). For Run 1, the mean error rate for bases with 1000x or greater coverage was 0.86%. For RUN 2, the mean error rate for bases with 45x or greater coverage was 0.99%. The same calculation an Illumina-based sequencing run showed an average error rate of 0.17%.

Certain base sites showed notably higher rates than average on both of the sequencing platforms evaluated (Tables 1 and 2). We generated a list of high sequencing error sites by applying the following criteria: For each ONT sequencing run we calculated the average sequencing error rate for the entire genome. We then determined the specific error rate for each position in the genome, and compared it to the overall sequencing error rate. For each sequencing run we identified the positions in the genome that had error rates 10-fold or higher than the average error rate. We then filtered these sites based on two additional criteria: 1) The error rate had to be at least 5-fold higher than average in the other ONT sequencing run and 2) We filtered out sites that were in the bottom 10% of coverage for either run. This yielded a set of 93 base positions that demonstrated reproducible elevated error rate. The error rate across these 93 sites was 11.9%. Of these 93 base sites 4 were a, 41 were C, 41 were G and 7 were T. Thus, in our data set GC base pairs (88% of high error sites) were far more likely to generate a high error site than AT base pairs. A summary of the highest error-rates sites is shown in Tables 1 and 2.

Table 1: ONT sequenching: sites with 20X over mean*error rates.

Base location	ONT Error Rate (ER)	ONT ER fold over mean	ONT Error Rate (ER)	ONT ER fold over mean	Illumina Error Rate (ER)	Illumina ER fold over mean	Gene location
	Run 1	**Run 1**	**Run 2**	**Run 2**			
5736	36.79%	43	31.13%	31	0.22%	1.3	NSP3
227	29.20%	34	28.73%	29	0.18%	1.09	UTR (untranslated)
3903	21.91%	25	13.76%	14	0.41%	2.5	NSP3
6078	21.52%	25	25.19%	25	0.28%	1.7	NSP3
16507	21.16%	25	23.16%	23	0.23%	1.4	HEU CASE
17708	20.18%	23	18.50%	19	0.10%	0.61	HEU CASE
20931	19.01%	22	24.23%	24	0.23%	1.4	2'-0-ribose methyltransferase
28568	18.81%	22	17.88%	18	0.27%	1.6	nudeocapsid phosphoprotein
7392	17.46%	20	22.40%	23	0.26%	1.5	NSP3
Mean Base	0.86%	1	0.99%	1	0.17%	1	
*20X as determined by average of two runs							

Table 2: Illumina sequencing: Sites with 10X over mean error rates.

Base location	Illumina error rate (ER)	Illumina ER fold over mean	ONT error rate	ONT ER fold over mean	Gene location
6669	0.0748	45	0.0597	6.9	NSP3

295 17	4.71	29	0.0045	0.52	Nucleocapsid phosphoprotein
3350	4.07	25	0.0399	4.6	NSP3
26715	3.15	19	0.0019	0.21	Membrane glycoprotein
6628	3.09	19	0.024	1.45	NSP3
9009	2.46	15	0.0062	0.72	NSP4
17385	2.17	13	0.0176	2.04	Heucase
26720	2.12	13	0.014	1.63	Membrane glycoprotein
25325	2.11	13	0.0032	0.37	Spike
28469	1.71	10	0.0182	2.12	Nucleocapsid phosphoprotein
Mean Base	0.0017	1	0.0086	1	

Next, we examined the type of sequencing error that occurred. Off all sequencing errors made in both ONT runs, 93.1% were transition errors (C->T, T->C, A->G, G->A). This bias has been reported before for ONT sequencing and is likely due to the chemical similarities of purines (A and G) and pyrimidines (C and T) [9]. Thus, it is more likely an ONT sequencing run would mistake one purine for another purine or one pyrimidine for another pyrimidine than mistake a purine for a pyrimidine.

We examined these errors with regard to their position in the genome by binning the genome into 5 kb sections (e.g. 1-5000 bp, 5001-10,000, etc.) and counting the number of high error rate sites in each bin. This analysis revealed no bias for genome position (data not shown) [10].

Lastly, we examined the sequence context for each of these 93 sites. To do this we took the 3 base pairs upstream and downstream of each error site to create a 7 base pair sequence. We then counted how many times that 7 bp sequences were found in our sequence of the synthetic genome. 7 base pair sequences appeared on average 3.2 times in the genomes (range of 1 to 9 times). 25 of the sequences in our set were unique in the genome, so no conclusions could be drawn from those. Another 59 sequences were found multiple times in our genome but only one of those instances demonstrated an elevated sequencing error rate. This suggests that the extent of sequence context examined here did not play a role in the sequencing error rate. There were 3 instances of sequences found twice in our set of 137 sites (GGCCACA, TCAGCAC, AGAGCAA) but in all cases there were additional instances of these 7 bp sequences in the genome that did not have an elevated sequencing error. The only 7 bp sequence that had evidence of sequence dependence was the sequence GACGGTT. This sequence appears three times in the synthetic genome and all three instances had an elevated error rate. This sequence had an average error rate of 13.1% across all three sites and both ONT runs. It is possible that there are more complex sequence contexts at work that our analysis did not reveal. None of these sites were observed to have an elevated error rate, using the same criteria on the synthetic genome sequenced with the illumina platform.

Turnaround time

We sought to determine the average turnaround time for generation of sequence/lineage data for specimens submitted to the NSPHL for diagnostic SARS-CoV-2 testing. Samples were drawn randomly from April through May of 2021 (98 samples) and the time from receipt at the Laboratory to the time when the lineage data was provided to a state public health database was determined. The overall average was 2.77 days (standard deviation, 1.2 days). The majority of samples were assessed diagnostically ("positive" or "negative" for SARS-CoV-2) by RT-PCR, sequenced and analyzed (lineage assigned, phylogenetic relations assessed) within 2 days (within 2 days: 57/98, 58%; within 3 days: 82/98, 84 %).

Control investigation impact of rapid sequencing and bioinformatics analysis

The ability to discern genomic differences rapidly at the single nucleotide level provides a capability to discriminate cases within the same lineage. Four illustrative examples are provided:

In February 2021, a cluster of 23 B.1.1.7 cases were detected in washoe county within a 10-day time frame. Genomic sequencing as described herein indicated that such cases were not genetically uniform and that two distinct clusters were visible, along with other genetically B.1.1.7 cases (Figure 1).

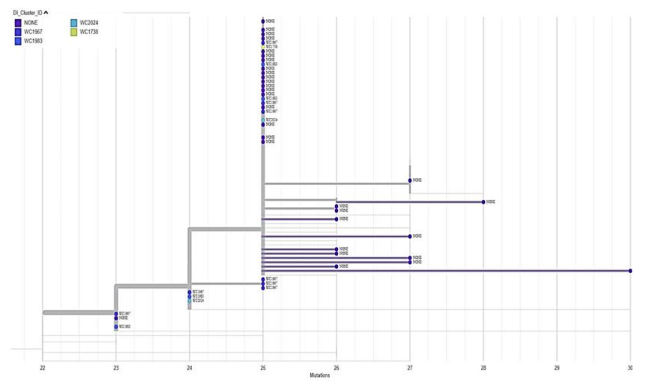

Figure 1: Cases of B.1.1.7 in Washoe County, February, 2021. This illustrates the four different B.1.1.7 clusters identified through genomic sequencing.

These data comported with information gathered through the disease control investigation process. The first cluster involved a private gathering which occurred on February 22, 2021, with 81 confirmed attendees, where 26 of them attended a secondary gathering held the following day. The contacts to the probable index case were found to possess identical genetics. The probable index case was a resident of another state, and was present at both gatherings.

42 cases were associated with an out of state volleyball tournament which occurred on February 27-28, 2021. The NSPHL notified WCHD of cases with B.1.1.7 and upon further investigation it was determined these initial cases were linked through athletics, the youth club and later to the high school volleyball teams. Sequence data provided the information that connected the high school cases to the volleyball club and consequently, the out of state tournament. This indicated that this tournament was the likely source of introduction for this particular cluster.

Figure 1 illustrates the four different B.1.1.7 clusters identified through genomic sequencing. The genomic sequencing data guided contact tracing efforts had linked cases on the 23 SNP mutation line had been associated to two unique clusters (WC1967 and WC1983). The 24 SNP mutation line shows three cases contact tracing had linked to three unique clusters were genetically identical. The 25 SNP mutation line calls attention to three cases that were separate from the largest cluster (WC1967), and the largest cluster on the 25 SNP mutation line shows cases from all four clusters were genetically related and several cases not associated to a cluster ("NONE") also were related.

By July 21, 2021, there were 14 unique B.1.617.2 (Delta variant) clusters detected with at least one case in each cluster sequenced. Two clusters are described further to describe the utility of SNP-level data.

A delta (B.1.617.2) cluster was identified through contact tracing and case investigation efforts for cases working in a manufacturing plant. Some cases were linked to the workplace but had other possible exposures. Through phylogenetic analysis, among the initial cases grouped into the cluster (WC2217), three cases were genetically distinct from the cluster and were residing in the same household. The household was originally associated to cluster WC2217 since the suspect index case for this household worked at the manufacturing warehouse and was believed to have been exposed in the workplace. However, there were other possible exposures, including travel history to a neighboring state. Sequence data demonstrated an introduction of the delta variant into the household independent from the workplace cluster WC2217. As a result, the household initially tied to WC2217 was disassociated from the manufacturing cluster. This example was used to remind disease investigators to consider all possible routes of exposure prior to associating a case with an infection chain.

One of the earliest identified B.1.617.2 clusters, WC2216, resulted in 19 cases, six hospitalizations, and three fatalities due to COVID-19 infection (Figure 2).

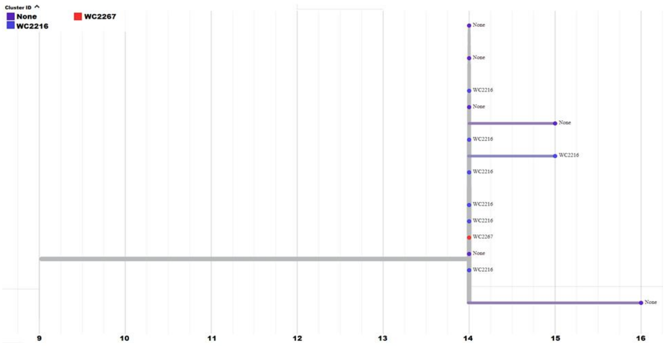

Figure 2: Phylogenetic relationships of B.1.617.2 (Delta) cases in July, 2020, in Washoe County, NV. The legend is stratified on epidemiologically identified clusters (WC_ID) and includes cases not linked to any cluster at all, identified as "None".

Virus detected and sequenced and located in this cluster has a spike mutation protein A222V. Phylogenetic analysis revealed a case associated with a different cluster, WC2267, which had zero SNP differences from cases linked to WC2216. Upon communicating with the disease investigators who investigated cases associated with WC2267, this separate social cluster resulted in the hospitalization of a couple in their 30s with no known underlying health conditions. Traditional case interviews had not yielded connections with these two clusters through named contacts or exposures, therefore genetic sequencing data alone demonstrated a linkage in these two notably virulent clusters.

Discussion

To date, most methods for sequencing the genomes of infectious agents have required large amounts of laboratory and computational time, in addition to expertise. All of this served as a barrier to implementation of a rapid, near real-time utilization of genomic data. Herein, we sought to remove that barrier through the use of novel, available tools. The first of those is a commercially available device capable of fully automating the genomic sequencing process of SARS-CoV-2. The second was the use of an open source bioinformatics tool which rapidly converts raw sequence data to lineage and provides facile pathways to phylogenetic relatedness analysis.

Utilizing this novel system, we found that epidemiological data integrity is greatly improved when a multidisciplinary approach is used to consider both epidemiological and genetic data. Genomic sequencing provides the data that disease investigators and contact tracers can use to verify epidemiologic hypotheses. This is particularly useful in large outbreaks where a correct identification of clusters is necessary for proper contact tracing to stop further spread. Moreover, sequence data was found to serve as a good reference to confirm clusters identified so that discordances in information gathered through traditional investigation can be revisited. In multiple investigations, SNP level phylogenetic information was able to reveal individuals associated with the same cluster but who were not forthcoming otherwise during regular investigations.

It is notable, that a centralized laboratory system for sequencing and sequence analysis would likely not produce the benefits described. Turnaround times for shipping, batching and analyses in a centralized system would not foster the use of genomic data in near-real time to impact disease control. Certainly, centralized systems of capability and expertise could play other roles in genomics, medicine, and public health, particularly at a strategic level.

Part of the system used in this work included a novel software solution to the conversion of raw sequence data to that which is useful to public health professionals. This solution used herein was found to possess multiple advantages.

The first advantage is that it utilizes Google Cloud Platform (GCP) as a compute and storage resource. This provides a nearly inexhaustible and scalable amount of resources for the fastest turnaround times achievable. GCP parallelizes all analyses and is capable of handling large numbers of compute requests that are often required by genomic analysis. The object model for cloud storage offered by GCP ensures that storage space is never an issue.

The second advantage of this solution is the adoption of the Terra platform. Although most cloud platforms are capable of the scalability and cost efficiency mentioned above, most of these require extensive maintenance of cloud architecture using IT professionals. This can drive up cost, inhibit adoption, and stifle innovation. Terra removes the time consuming maintenance of a cloud-computing infrastructure, allowing scientists to focus on the content, not on becoming bioinformaticists or Linux system administrators. Training on Terra is simple, and Terra provides for multiple levels of permissions and security, making sure scientists are permitted access only what they need, and not anything else. Furthermore, as data sharing becomes necessary, these same permission and security features can allow effective data sharing between entities on the county, state, and federal levels.

The third advantage involves the use of Dockstore. The Terra platform utilizes workflows that are freely available on Dockstore, an online repository of genomic workflows for biomedical research and public health. Lastly, the entire system, from Google to Terra to Dockstore, is reproducible, portable, and can be implemented in a matter of days in just about any location globally [8-10].

Conclusion

The quality of sequence data generated by the clear labs, ONT driven platform was found to be higher than expected. We found average error rates similar to those previously observed and lower than seen previously. Certain sites in the SARS-CoV-2 genome were found to be more prone to sequencing error, with only one possible sequence theme consistently associated with error. Further investigation will be required.

References

1. Rossen JWA, Friedrich AW, Moran-Gilad J (2018) Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect 24:355-360
2. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, et al. (2020) Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis 20:1263-1271
3. Zuckerman NS, Pando R, Bucris E, Drori Y, Lustig Y, et al. (2020) Comprehensive Analyses of SARS-CoV-2 Transmission in a Public Health Virology Laboratory. Viruses 12:854
4. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, et al. (2021) Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature 595:707-712
5. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, et al. (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403-1407
6. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, et al. (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123
7. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, et al. (2020) Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. Nat Commun 11:6272
8. Sarkozy P, Jobbagy A, Antal P (2018) Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times. In: EMBEC and NBC 2017. Springer, Singapore 241–244
9. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H (2020) Benchmarking of long-read correction methods. NAR Genom Bioinform 2:lqaa037
10. Delahaye C, Nicolas J (2021) Sequencing DNA with nanopores: Troubles and biases. PLOS ONE 16:e0257521