IRSTI 34.15.31 https://doi.org/10.26577/eb.2022.v92.i3.08

D.K. Kamalova[1,4]*, M.Zh. Zhurinov[2], G.A. Tasanova[3], A.O. Amirgazin[4], K.K. Mukanov[4], A.B. Shevtsov[4]

[1]«Eurasian National University named after L.N. Gumilyov», Kazakhstan, Nur-Sultan
[2]«D.V. Sokolsky Institute of Fuel,Catalysis and Electrochemistry» JSC, Kazakhstan, Almaty
[3]Non-profit limited company «A.Baitursynov kostanay regional university», Kazakhstan, Kostanay
[4]«National Center for Biotechnology», Kazakhstan, Nur-Sultan
*e-mail: kamalova@biocenter.kz

DEVELOPMENT OF A PROTOCOL FOR WHOLE GENOME SEQUENCING OF THE SARS-COV-2 VIRUS

Whole-genome sequencing of the SARS-CoV-2 virus during the pandemic has become an essential part of the epidemiological control of the spread of coronavirus infection. This made it possible to get actual data of circulating genetic variants and mutational changes of SARS-CoV-2 during the pandemic. The obtained results provided an opportunity for researchers to collect additional data to evaluate the virulence, infectivity, and the likelihood of evading an immune response when using vaccines and therapeutic monoclonal antibodies. There are different technological approaches for whole-genome sequencing of SARS-CoV-2, but multiplex PCR amplification is used most often due to ease of implementation and cost-effectiveness. When performing research on viral genome sequencing, researchers need to optimize existing sequencing protocols with available reagents or develop new approaches. This study is devoted to the development of a protocol for whole-genome sequencing of the SARS-CoV-2 virus based on the use of RT-PCR amplification using 39 primer pairs in 3 reaction mixtures. The protocol made it possible to obtain 15 complete genome data of the SARS-CoV-2 and can probably be used for large-scale studies on viral genome sequencing.

Key words: COVID-2019, SARS-CoV-2, RNA, whole genome sequencing.

Д.К. Камалова[1,4]*, М.Ж. Жұрынов[2], Г.А. Тасанова[3], А.О. Амиргазин[4], Қ.Қ. Мұқанов[4], А.Б. Шевцов[4]

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан, Нұр-Сұлтан қ.
[2]«Д.В. Сокольский атындағы Жанармай, Катализ және Электрохимия Институты» АҚ, Қазақстан, Алматы қ.
[3]«А. Байтұрсынов атындағы Қостанай өңірлік университеті» КеАҚ, Қазақстан, Алматы қ.
[4]Ұлттық биотехнология орталығы, Қазақстан, Нұр-Сұлтан қ.
*e-mail:kamalova@biocenter.kz

SARS-COV-2 вирусының толықгеномды секвенерлеу хаттамасын әзірлеу

Пандемия кезеңінде SARS-CoV-2 вирусын толықгеномды секвенерлеу коронавирус инфекциясының таралуын бақылаудағы эпидемиологияның ажырамас бөлігі болып қалыптасты. Бұл пандемия кезеңінде SARS-CoV-2 вирусының мутациялық өзгергіштігі мен айналымда жүрген генетикалық түрлері туралы жаңартылған деректер алуға мүмкіндік берді. Алынған нәтижелер зерттеушілерге екпелер мен терапевтік моноклоналды антиденелерді пайдалану кезінде вируленттілік, инфекциялық және иммундық жауаптан жалтару ықтималдығы туралы қосымша деректерді жинақтауға мүмкіндік берді. Қазіргі таңда SARS-CoV-2 вирусын толықгеномды секвенерлеудің әр түрлі технологиялық тәсілдері бар, бірақ мультиплексті ПТР амплификация өзінің жеңіл қолданылуы мен үнемшілдігі тұрғысынан жиі қолданылады. Вирустар геномын секвенерлеу алдында зерттеушілер секвенерлеу хаттамасын өздеріне қолжетімді реагенттерді қолдана отырып, оңтайландырып алуы қажет немесе жаңа тәсілдер әзірлеулері керек. Мақалада жүргізілген зерттеу жұмыстары, 3 реакциялық қоспада 39 жұп праймерді қолдану арқылы, КТ-ПТР амплификациясын қолдану негізінде SARS-CoV-2 вирусын толықгеномды секвенерлеу хаттамасын әзірлеуге бағытталған. Хаттама SARS-CoV-2 вирусының 15 толық геномдық мәліметін алуға мүмкіндік береді және кең ауқымды вирустық геномды секвенерлеуге арналған зерттеулерде пайдаланылуы мүмкін.

Түйін сөздер: COVID-2019, SARS-CoV-2, РНҚ, толықгеномды секвенерлеу.

Д.К. Камалова*[1,4], М.Ж. Журинов[2], Г.А. Тасанова[3],
А.О. Амиргазин[4], К.К. Муканов[4], А.Б. Шевцов[4]

[1]Евразийский национальный университет имени Л.Н. Гумилева, Казахстан, г. Нур-Султан
[2]АО «Институт топлива, катализа и электрохимии им. Д.В. Сокольского, Казахстан, г. Алматы
[3]НАО «Костанайский региональный университет имени А. Байтурсынова», Казахстан, г. Костанай
[4]Национальный центр биотехнологии, Казахстан, г. Нур-Султан
*e-mail:kamalova@biocenter.kz

Разработка протокола полногеномного секвенирования вируса SARS-CoV-2

Полногеномное секвенирование вируса SARS-CoV-2 за период пандемии стало неотъемлемой частью эпидемиологического контроля за распространением коронавирусной инфекции. Это позволило обеспечить получение актуальных данных о циркулирующих генетических вариантах и мутационных изменениях вируса SARS-CoV-2 в период пандемии. Полученные результаты предоставили возможность исследователям накопить дополнительные данные о вирулентности, инфекционности, о вероятности уклонения от иммунного ответа при использовании вакцин и терапевтических моноклональных антител. Существуют разные технологические подходы для полногеномного секвенирования SARS-CoV-2, но мультиплексная ПЦР амплификация используется наиболее часто из-за простоты выполнения и экономичности. При выполнении исследований по секвенированию геномов вирусов исследователям необходимо оптимизировать существующие протокола секвенирования под доступную реагентику или разрабатывать новые подходы. Данные исследования посвящены разработке протокола полногеномного секвенирования вируса SARS-CoV-2 на основе использования ОТ-ПЦР амплификации с использованием 39 пар праймеров в 3 реакционных смесях. Протокол позволил получить данные 15 полных геномов вируса SARS-CoV-2 и может быть использован для крупномасштабных исследований по секвенированию геномов вирусов.

Ключевые слова: COVID-2019, SARS-CoV-2, РНК, полногеномное секвенирование.

Introduction

Mankind throughout its history has repeatedly faced various kinds of pandemics. One of the latest pandemics that has affected all of humanity is the COVID-2019 (Corona Virus Disease 2019) pandemic caused by the SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus 2). SARS-CoV-2 belonging to the Coronaviridae family, genus Betacoronavirus [1], is a single-stranded positive-polarity RNA-genome virus.

The first information about unknown pneumonia was made public on 31st of the December in 2019 in Wuhan, China. On the 3rd of January, 44 patients were confirmed infected and the whole genome sequence of the SARS-CoV-2 (nCoV2019) virus was obtained. Due to the exponential growth in numbers of infected people, the Chinese government introduced unprecedented quarantine measures on the 23rd January, with the closure of cities and the isolation of people. The number of new COVID-19 cases has decreased after a two-month surge. By the end of March, China had 81,554 confirmed cases of COVID-19, of which 50,007 were in Wuhan [2]. Nevertheless, the infection quickly spread beyond China and as of April 4, 2020, a total of 1,051,635 people had a confirmed diagnosis of COVID-19 in 208 countries, in which more than 75,000 deaths have been reported [3]. On March 11th, 2020, the World Health Organization (WHO) announced that COVID-19 should be characterized as a pandemic [4]. At great economic cost, many countries have taken unprecedented measures to restrict the spread of the virus, including quarantines, closing borders, imposing restrictions on crowds, and imposing nationwide lockdowns [5]. As of February 20, 2022, there have been over 422 million confirmed cases and over 5.8 million deaths worldwide [6].

The control and fight against any virus outbreaks directly depend on various activities, including effective diagnostic procedures. Knowledge of mutational changes in the pathogen genome is crucial for assessing the adaptive changes of the virus, which can accumulate and affect its diagnosis, transmissibility, and pathogenicity [7]. Sequencing plays an important role in controlling the efficiency of antiviral drugs and vaccines, identifying determinants of drugs resistance and variants capable of evading vaccine-immunity, and detecting possible recombination.

Advances in genome sequencing technology have resulted in an exceptionally large number of SARS-CoV-2 genomes being sequenced during the COVID-19 pandemic [8]. Genome sequences have been made publicly available through several repositories, including a data exchange service

hosted by the Global Initiative to Share All Influenza Data (GISAID) (https://www.gisaid.org/) [9,10]. In addition to public genome repositories, open source platforms for real-time data visualization and genomic data analysis also work, including NextStrain (https://nextstrain.org) and CoV-GLUE (http://cov-glue.cvr.gla.ac.uk).

Analysis of the genome-wide data from the first months made it possible to differentiate genetic lines and mutations that increase transmissibility which led to rapid spread to all continents. For example, the D614G mutation in the S protein of SARS-CoV-2 was critical for S1 cleavage, facilitating virus fusion with the host cell membrane [11,12]. The G614 mutant, which originated in Europe in February 2020, quickly leaked to other parts of the world and accounted for more than 74% of all published sequences by June 2020 [13].

The rapid accumulation of genomic data required the classification of the genotypes. Several attempts have been made to classify circulating SARS-CoV-2 strains into lineages or genotypes, with potential differences in transmissibility and disease severity. A study on 103 genomes of SARS-CoV-2 provided by Lu R. et al. showed the existence of two lineages, named L and S [14]. A comparative study of the 160 SARS-CoV-2 virus genomes from different countries identified three subtypes A, B and C, with differences in geographical spread and distribution [15]. However, such studies have been criticized for possible sampling bias and misinterpretation of the results [16,17,18,19]. Limited or incorrectly drawn sampling would skew the inferences, potentially hiding a part of circulation and resulting in inaccurate estimates of mutation rates [20]. The World Health Organization has decided to designate coronavirus mutations with letters of the Greek alphabet. The first four letters went to strains from Britain (Alpha), South Africa (Beta), Brazil (Gamma) and India (Delta), Botswana and the Republic of South Africa (Omicron). During the pandemic, multiple variants of the alpha-, beta-, gamma-, delta-, and omicron-SARS-CoV-2 viruses were identified.

During molecular epidemiology studies and implementing whole genome sequencing of pathogens researchers face the need to apply a suitable sequencing protocol which may need adjusting. Several approaches have been proposed for sequencing the SARS-CoV-2 virus. The metagenomics approach is obtaining cDNA and its total sequencing, its disadvantage is the need to obtain a large number of high-quality reads which is not always possible given low virus titers or admixing human RNA in samples [21]. Another approach is random amplification with one primer for whole genome sequencing, it allows to identify mixed variants, however much like metagenomics, requires a deep sequencing [22]. Yet another approach is multiplex-PCR-based enrichment before whole genome sequencing on 2nd and 3rd generation NGS sequencers [23,24].

The virus genome enrichment protocol based on specific amplification is easy to conduct and it is less costly. However, specific PCR-related points, variety of instruments in different laboratories, and different reagents, all require the optimization of the published protocols. The aim of this work was to develop a protocol for SARS-CoV-2 sequencing by using the PCR-enrichment followed by sequencing on the MiSeq platform (Illumina).

Materials and Methods

Clinical Samples and RNA Isolation

We used 30 RNA samples obtained from patients diagnosed with COVID-19. The samples were collected from the four regions of Kazakhstan.

The work with clinical samples was carried out BSL-3 facilities at the local branch of the National Center for Biotechnology (NCB) in Almaty (Kazakhstan). RNA was isolated from the clinical samples using the QIAamp Viral RNA Mini Kit (Cat. 52904, QIAGEN) according to the manufacturer's instructions. RNA was transferred to the NCB central laboratory in Nur-Sultan, Kazakhstan.

Primer design

The selection of primers was carried out using a consensus sequence created from the alignment of 500 full-genome sequences of the SARS-CoV-2 virus. The primers targeted conservative regions in the genome, with distance between primer-binding sites 800-1200 bp. Planned PCR products overlapped by 100-200 bp. Software FastPCR and NCBI PrimerBlast (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) were used to compute the specific primers. The primers were chemically synthesized at NCB.

Primers check

cDNA was obtained using the Reverta-L kit (Russia) in a total volume of 120 µl according to the manufacturer's instructions. With each matched primers-pair, a PCR reaction was performed, including: forward and reverse primers in a final concentration 300 nM, 1U. SynTaq DNA polymerase (Synthol, Russia), 0.2 mM of each dNTP, 1x PCR buffer (Sintol, Russia), 2.5 mM MgCl2, 3% DMSO, and 3 µl cDNA. The PCR amplification program is: first denaturation at 95°C for 3 minutes; 35 cycles:

95°C – 25 seconds, 59°C – 40 seconds, 72°C – 90 seconds, final elongation 72°C – 10 minutes.

Amplification products

The analysis of the amplified DNA fragments was performed by electrophoresis in 1.5% agarose gels stained with ethidium bromide. 1×TAE buffer was used as electrophoresis buffer. The gels were documented using the GelDoc (Bio-Rad) and analysed in QuantityOne (Bio-Rad). PCR-products sizes were compared to DNA Ladder 10kb (Thermo Scientific, #SM1293, 100–10000 bp).

Multiplex PCR Enrichment in Two-Step RT-PCR

For reverse transcription, RNAscribe RT kit (Biolabmix, Russia) was used according to the manufacturer's instructions. After the reverse transcription, 6 µl of cDNA was used for PCR. The reaction mixture included 2 µl of primers, 2U SynTaq DNA polymerase (Sintol, Russia), 0.2 mM each dNTP, 1-x PCR buffer (Sintol, Russia), 2.5 mM MgCl2, 3% DMSO. The PCR amplification program had denaturation at 95°C for 3 minutes; 42 cycles: 95°C – 25 seconds, 57°C – 40 seconds, 72°C – 3 minutes, final elongation 72°C – 10 minutes.

Multiplex PCR Enrichment in One-Step RT-PCR

One-step RT-PCR was performed using the BioMaster RT-PCR-Extra reagent kit (2×). The reaction mixture was made in a volume 25 µl including 1 µl of primer-mixture, 12.5 µl of 2× manufacturer-supplied mixture RT-PCR-Extra, 1 µl of BioMaster Extra-mix, 7 µl of RNA, and 2.5 µl of DMSO. The PCR amplification program included reverse transcription for 30 minutes at 50°C, denaturation at 93°C for 3 minutes; 42 cycles: 93°C – 15 seconds, 57°C – 40 seconds, 68°C – 2 minutes, and final elongation 68°C – 10 minutes.

Preparation of DNA libraries and high throughput whole genome sequencing

The preparation of DNA libraries from PCR fragments was performed using the Nextera DNA Flex Library Prep Kit (24 samples) (Cat. No. 20018704, Illumina, USA) according to the manufacturer's instructions. Sequencing was performed on a MiSeq sequencer platform (Illumina, USA) using MiSeq Reagent Kit v3, 600 Cycles (Catalog #MS-102-3003).

Bioinformatics

Bioinformatics analysis was done on Ubuntu 19.04 LTS operating system. Raw data quality control was performed using FastQC v0.11.7 [25] and MultiQC v1.8 [26] programs. Raw data trimming was performed using the Seqtk v1.3-r106 program [27].

The BWA v0.7.17-r1188 program [28] was used to map the reads to the reference sequence imported by accession number NC_045512.2 from the GenBank (NCBI) database. Variant identification and consensus sequencing were performed using FreeBayes [Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012] and BCFtools [Danecek, Petr, et al. "Twelve years of SAMtools and BCFtools." Gigascience 10.2 (2021): giab008.]

Results

For use in the PCR-enrichment, thirty-nine 39 pairs of primers were designed covering the entire virus genome (Table 1).

The use of all selected primers-pairs in a monoplex reaction resulted in expected fragments varying in length by 800-1200 bp. Figure 1 shows the electrophorerogram of 28 primer pairs.

The selected primers were divided into 3 sets (Table 1). The primers were mixed in equal molarity to 4 pmol/µl of each primer. The resulting mixtures were used in one-step RT-PCR and for the 2nd step of the two-step PCR. For reverse transcription in two-step RT-PCR, a mixture of R primers was used, at 2 pmol/µl of each primer. The PCR products from amplifications with three primer-sets were mixed and purified with AMPure XP magnetic beads (Beckman Coulter) in a 1:1 ratio. The quality of the obtained PCR products was checked by electrophoresis (Figure 2) and measuring concentration using the fluorimetric method.

No differences were found in the concentrations of PCR-products produced in multiplexed reactions on cDNA probes from different samples.

The results from whole-genome sequencing using the MiSeq platform (Illumina) included from 300 thousand to 600 thousand reads per sample, with average lengths from 172 bp up to 277 bp, and duplicated sequences totaling 71.9% to 81.5%. A large number of duplicated sequences can be explained by the sequencing of PCR products as a DNA template. The assembly of genomes by the mapping method resulted in complete genomes of 30 isolates of the SARS-CoV-2 virus. To evaluate coverage uniformity, we calculated the average coverage of each amplicon at a normalized depth of 500x reads per position. The whole genome sequencing protocol using one-step RT PCR resulted in a more uniform library (Figure 3A). However, two regions on the genome, i.e. the nt.8500-10000 and nt.22000-23000 regions have the coverage of less than 250x. The two-step RT PCR protocol had 8 to 12 genome regions with less than 250 coverage regions, as well as an unsequenced region (Figure 3B).

Table 1 – Primers for PCR amplification of the SARS-CoV-2 genome

Primers set 1		Primers set 2		Primers set 3	
F_4-29	aaaggtttataccttcccaggtaaca	F_590	gaaataccagtggcttaccgca	F_1498	ctcttatgttggttgccataacaagt
F_2249	ggacaaattgtcacctgtgcaa	F_2996	tctggtgagtttaaattggcttcac	F_3730	ttggtgctgaccctatacattctt
F_4591	tgaaactcttgttacaatgccactt	F_5339	ccacctgctctacaagatgcttat	F_6091	ccagttaactggttataagaaacctgc
F_6839	gcatctatgccgactactatagcaa	F_7619	tgtgttaattgtgatacattctgtgct	F_8385	acaacattgctttgatatggaacgtta
F_9230	acacacgttatgtgctcatggat	F_9918	gtggagcaatggatacaactagc	F_10560	ggagttcatgctggcacagactta
F_11438	aatgctttagatcaagccatttcca	F_12221	gtggctaaatctgaatttgaccgt	F_12991	acttggtagtttagctgccaca
F_13757	tagacggtgacatggtaccaca	F_14630	cagtagctgcacttactaacaatgtt	F_15376	tgtagcttgtcacaccgtttctat
F_16172	acacttcaaggtattgggaacct	F_16915	ctgacatcacatacagtaatgccatta	F_17667	gggtgttatcacgcatgatgtt
F_18439	tccagagttagtgctaaaccacc	F_19196	attgcaatgtcgatagatatcctgc	F_19845	gggtgtggacattgctgctaat
F_20752	gatagtgcaacattacctaaaggcat	F_22480	ccctgcatacactaattctttcacac	F_22898	ctccctcagggtttttcgcytt
F_23108	gaacttctacatgcaccagcaa	F_23853	accgtgctttaactggaatagct	F_24608	gcagaaatcagagcttctgctaa
F_25350	cagtgctcaaaggagtcaaattaca	F_26097	aattgttgatgagcctgaagaacat	F_26844	tccatgtggtcattcaatccagaa
F_27612	cgtctatcagttacgtgccagat	F_28402	aggtttacccaataatactgcgtct	F_29860	gattgaaccagcttgagagcaa
R_886	agtgcatgaagctttaccagca	R_1733	tgctttataatccaaacctttcacagt	R_2376	tgcgtgacaaatgtttcacctaaat
R_3132	gcaccaaattccaaaggtttacctt	R_3872	cttcctctttaggaatctcagcgat	R_4632	gagcagcttcttccaaatttaagc
R_5472	gcaagaatctaaattggcatgttgaa	R_6228	cattgttaacatgccaaacaataggtt	R_6980	ggttgagtagattaaagaacctaggca
R_7715	aacgatgtaagaagactggtcagt	R_8530	accacccttaagtgctatctttgtt	R_9650	aacataaccatccactgaatatgtgctaa
R_10300	gcatagaatgtccaataaccctgagtt	R_10792	cagaaagaggtcctagtatgtcaaca	R_11656	gtctaaagtagcggttgagtaaacaa
R_12336	cattgtctgcatagcactagtaactt	R_13100	cccactagctagataatctttgtaagc	R_13962	acgttcacctaagttggcgtat
R_14659	aattaccgggtttgacagtttgaaa	R_15521	ccgtgacagcttgacaaatgttaa	R_16285	atggtctacgtatgcaagcacc
R_17063	tggtccctggagtgtagaatact	R_17796	agtcccaaaatctttgaggctaca	R_18575	gcccataagacaaatacgactctgt
R_20430	gccacattttctaaactctgaagtctt	R_20072	gtttgggacctacagatggttgta	R_20874	ggtgcaactcctttatcagaacca
R_21640	cacgtgtgaaagaattagtgtatgca	R_22512	agattctgttggttggactctaaagt	R_23233	caaattgttggaaaggcagaaacttt
R_24029	atctgcaagtgtcactttgttgaa	R_24730	gcaagaagactacaccatgaggt	R_25495	ggagtgaggcttgtatcggtat
R_26227	gagtacataagttcgtactcatcagc	R_26983	atgtcacagcgtcctagatggt	R_27838	cttgcagttcaagtgagaaccaaa
R_28496	tcatctggactgctattggtgtt	R_29327	tgcgtcaatatgcttattcagcaa	R_30760	cctaagaagctattaaaatcacatgg

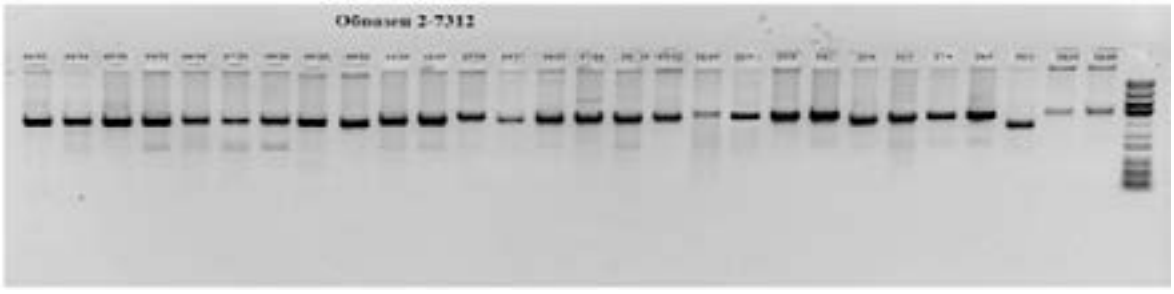

Figure 1 – Analysis of SARS-CoV-2 amplification results. The produced PCR products cover the entire SARS-CoV-2 genome (M), molecular weight marker (Thermo Scientific, #SM1293) (100-10000 bp, from 100-10000).

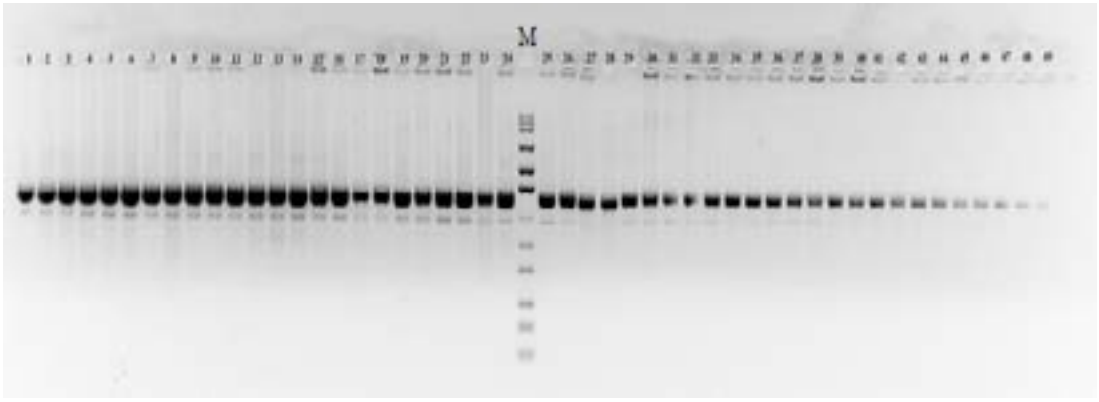

Figure 2 – Electrophoresis of pooled samples. These products were produced in multiplexed PCR with three primer-sets shown in Table 1.
(M), molecular weight marker; 1-15 PCR products obtained in one-step RT-PCR; 16-30 PCR products obtained in one-step RT-PCR.

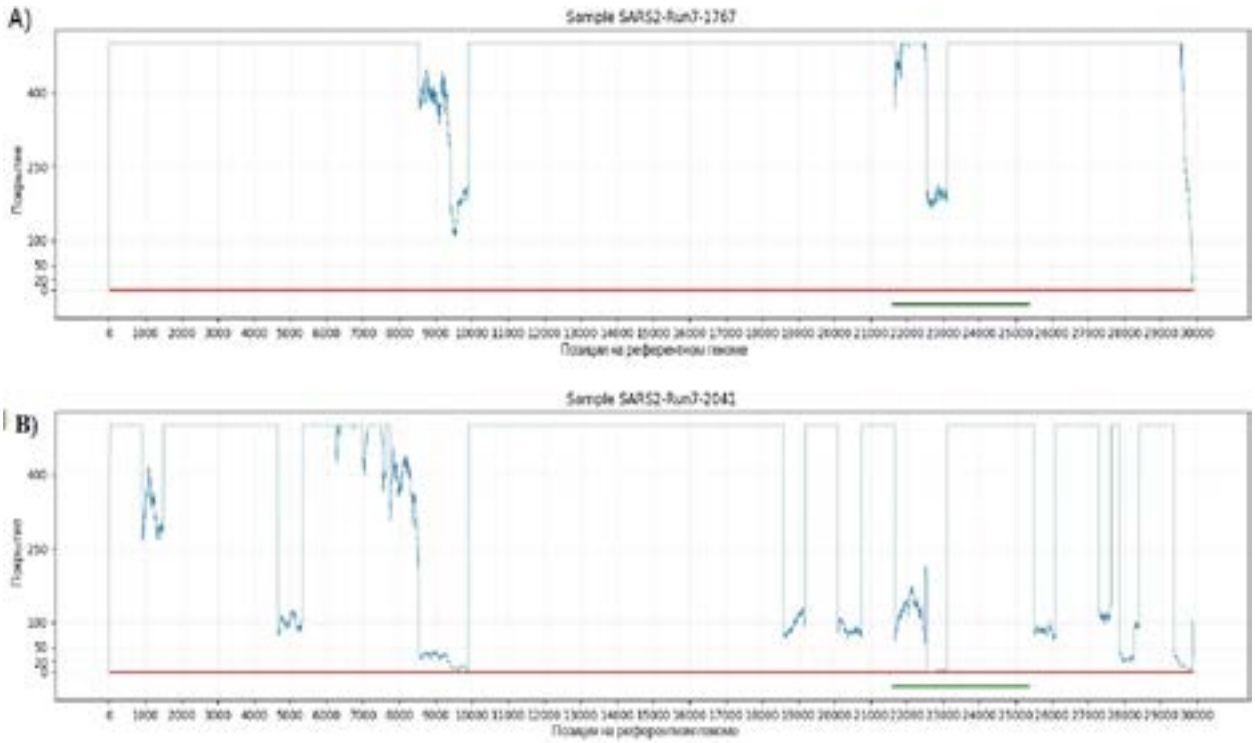

Figure 3 – Histogram of the average depth of amplicon sequencing.
(A) using one-step RT PCR, (B) using two-step RT PCR

Discussion

The reduction in the costs of whole genome sequencing (WGS) facilitated the introduction of WGS into genetic epidemiology. In the majority of studies, the use of WGS data is retrospective. With such use, epidemiologists obtain the data on phylogenetic relationships, evolutionary changes, drug resistance factors and virulence after the studied outbreak has been eliminated [29]. This approach does not apply to pandemics or outbreaks that pose a global threat or have the potential to develop into a pandemic. The first real-time genomic analysis of the behavior of the pathogen was applied during the outbreak of the Ebola virus in 2014-2015, by tracking changes in the structures

of circulating lineages and distribution pathways [30].

The COVID-19 pandemic mobilized the healthcare system of all countries and highlighted the need to introduce WGS into the pandemic control and monitoring system. In Kazakhstan, controlling genomic variability of the SARS-CoV-2 virus was introduced fairly soon. However, existing deficiencies in the infrastructure for genomic research of dangerous pathogens resulted in slowing down the research, because sequencing protocols needed to be optimized. Actually, the first sequencing-confirmation of SARS-CoV-2 in Kazakhstan and the first local data submitted to GISAID have been obtained by Sanger sequencing. Subsequent genomes were sequenced on MiSeq, but enrichment was required. The monoplex enrichment needs a high-titer virus sample which or a large amount of clinical sample, because each genome required 28 to 30 OT-PCR reactions. The use of prefabricated panels incurs high financial costs. In this work, we have developed a protocol for NGS sequencing of the SARS-CoV-2 virus with PCR enrichment in three reaction mixtures.

The WGS monitoring of the SARS-CoV-2 virus allowed describing multiple mutations and the discrimination of genetic lines, which when juxtaposed to epidemiological and functional data enabled predicting epidemic waves. The first functional missense mutation D614G in the gene encoding the spike protein (S-protein) was identified in March 2020 in Europe and the mutation rapidly propagated across the world. The D614G change was shown to enhance the virus replication in human lung epithelial cells and primary human respiratory tissues by increasing the infectivity and stability of virions, which explains the evolutionary success of the mutant strains [31]. The emergence of the "delta" variant has raised concerns about the immune response evasion in vaccinated and recovering individuals [32].

Active mutational dynamics in SARS-COV-2 highlights the importance of global genomic surveillance of the virus and could lead to early detecting of new emerging variants ahead of a possible epidemic wave. The whole genome sequencing method is currently being used to study the genetic diversity of the SARS-CoV-2 virus to supplement phylogenetic analysis, haplotype network analysis and genetic diversity studies [32].

Conclusion

Methods for complete-genome sequencing of SARS-CoV-2 were improved in this work. Protocols were optimized to include the PCR-enrichment with 39 designed primer-pairs. The enrichment of genomic sequences allowed using for WGS low-copy-virus samples or limited quantities of clinical samples. The results enable using 2nd and 3rd generation sequencing platforms for WGS with SARS-CoV-2 samples. This article presents the improved protocols and the utilization to conduct monitoring of SARS-CoV-2 genetic diversity in Kazakhstan.

Acknowledgment

This study was supported by the program BR10965271 "Development of highly effective medicinal substances from plant materials with antiviral activity against COVID-19 and similar viral infections" of the Ministry of Education and Science of the Republic of Kazakhstan.

References

1 Yang X, Yu Y, Xu J, Shu H, Xia JA, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study // Lancet Resp Med. – 2020. – Vol. 8. – P. 475–81. doi: 10.1016/S2213-2600(20)30079-5

2 Li, Z., Guan, X., Mao, N., Luo, H., Qin, Y., He, N., & Gao, G. F. Antibody seroprevalence in the epicenter Wuhan, Hubei, and six selected provinces after containment of the first epidemic wave of COVID-19 in China // The Lancet Regional Health-Western Pacific. -2021. – Vol. 8. 100094.

3 of the International, C. S. G. The species Severe acute respiratory syndrome- related coronavirus: classifying 2019-nCoV and naming it SARS- CoV-2 // Nat. Microbiol. –2020. –Vol. 5. – P. 536-544.

4 Corman V. M., Muth D. Niemeyer D. Hosts and sources of endemic human coronaviruses // Adv. Virus Res. – 2018. – Vol. 100. – P. 163-188.

5 Annan A., Baldwin H.J., Corman V.M., Klose S.M., Owusu M., Nkrumah E.E., et al. Human betacoronavirus 2c EMC/2012-related viruses in bats, Ghana and Europe // Emerg. Infect. Dis. – 2013. – Vol. 19. – P. 456-459.

6 https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---22-february-2022

7 Taubenberger, J.K., Kash, J.C. Influenza virus evolution, host adaptation, and pandemic formation // Cell Host Microbe. – 2010, – Vol. 7. – P. 440–451.

8 World Health Organization. Summary table of SARS cases by country, 1 November 2002-7 August 2003. Weekly Epidemiological Record // Relevé épidémiologique hebdomadaire. -2003. – Vol. 78(35). – P. 310-311.

9 Zaki A.M., van Boheemen S., Bestebroer T.M., Osterhaus A.D., Fouchier R.A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia // N. Engl. J. Med. – 2012. – Vol. 367. – P. 1814-1820.

10 Reusken C.B., Haagmans B.L., Muller M.A., Gutierrez C., Godeke G.J., Meyer B., et al. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study // Lancet Infect. Dis. – 2013. – Vol.13. –P. 859-866.

11 Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182: 812–827. e819. pmid:32697968

12 Zhang L, Jackson CB, Mou H, Ojha A, Peng H, et al. (2020) SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity // Nature communications. – 2020. – Vol.11. – P. 1–9.

13 Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant // Cell . -2020. – Vol.183. –p. 739–751.

14 Lu R., Zhao X., Li J., et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding // Lancet. – 2020. – Vol. 395. – P. 565-574.

15 Wan Y., Shang J., Graham R., et al. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus // J. Virol. – 2020. – Vol. 94. doi: 10.1128/JVI.00127-20

16 de Groot R.J., Baker S.C., Baric R., Enjuanes L., Gorbalenya A.E., Holmes K.V., et al. Family Coronaviridae. Virus Taxonomy: Classification and Nomenclature of Viruses // Ninth Report of the International Committee on Taxonomy of Viruses. London. – 2012. – P. 806-828.

17 Wan Y., Shang J., Graham R., et al. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus // J. Virol. – 2020. – Vol. 94(7). e00127-20.

18 Sun J., He W.T., Wang L., et al., COVID-19: epidemiology, evolution, and cross-disciplinary perspectives // Trends. Mol. Med. – 2020. – Vol. 26. – P. 483-495.

19 Neurath M.F. Covid-19 and immunomodulation in IBD // Gut. – 2020. – Vol. 69. – P. 1335–1342.

20 Masters P.S. The molecular biology of coronaviruses. advances in virus research // Academic Press. – 2006. – Vol. 66. – P. 193-292.

21 Manning, J. E., Bohl, J. A., Lay, S., Chea, S., Sovann, L., Sengdoeurn, Y., & Karlsson, E. A. Rapid metagenomic characterization of a case of imported COVID-19 in Cambodia // Biorxiv. -2020. doi: 10.1101/2020.03.02.968818

22 Chrzastek, K., Tennakoon, C., Bialy, D., Freimanis, G., Flannery, J., & Shelton, H. A random priming amplification method for whole genome sequencing of SARS-CoV-2 and H1N1 influenza A virus // bioRxiv – 2021. doi: https://doi.org/10.1101/2021.06.25.449750

23 Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R., & Kuroda, M. A proposal of alternative primers for the ARTIC Network's multiplex PCR to improve coverage of SARS-CoV-2 genome sequencing // BioRxiv. – 2020. doi: https://doi.org/10.1101/2020.03.10.985150

24 Resende, P. C. et. al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms // BioRxiv. – 2020. doi: https://doi.org/10.1101/2020.04.30.069039

25 Mukanov K. K., Shevtsov A. B. Berdimuratova K. T, Amirgazin A. O, Kuibagarov M. A, Lutsay V. B. Optimization of PCR purification using silica-coated magnetic beads // Eurasian Journal of Applied Biotechnology. – 2020. – №. 1. – P. 81-89.

26 Andrews S. et al. FastQC: a quality control tool for high throughput sequence data. – 2010.

27 Ewels P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report //Bioinformatics. – 2016. – Т. 32. – №. 19. – С. 3047-3048.

28 Li H. Seqtk—Toolkit for Processing Sequences in FASTA/Q Formats; 2008.

29 Octavia, S., Ang, M. L., Van La, M., Zulaina, S., Saat, Z. A. A. S., Tien, W. S., Lin, R. T. Retrospective genome-wide comparisons of Salmonella enterica serovar Enteritidis from suspected outbreaks in Singapore // Infection, Genetics and Evolution. – 2018. – P. 229-233.

30 Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Carroll, M. W. Real-time, portable genome sequencing for Ebola surveillance // Nature. – 2016. – Vol. 530(7589). – P. 228-232.

31 Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., Shi, P. Y. Spike mutation D614G alters SARS-CoV-2 fitness // Nature. – 2021. – Vol. 592(7852). – P. 116-121.

32 Shastri, J., Parikh, S., Aggarwal, V., Agrawal, S., Chatterjee, N., Shah, R. & Pandey, R. Severe SARS-CoV-2 breakthrough reinfection with delta variant after recovery from breakthrough infection by alpha variant in a fully vaccinated health worker // Frontiers in Medicine. – 2021. – P. 1379.

33 Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, Nao N, Tanaka R, Hashino M, Takahashi T, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak // Proc Natl Acad Sci USA. -2020. – Vol. 117. – P. 20198–20201.