

Sergey Tkachev, Leila Shigapova, Nurislam Shaikhutdinov, Elena Shagimardanova

**OPTIMIZATION OF HIGH-THROUGHPUT SEQUENCING
METHODS OF TICK-BORNE ENCEPHALITIS VIRUS
GENOMES**

Guidelines

KAZAN FEDERAL UNIVERSITY
2023

INTRODUCTION

Tick-borne encephalitis virus (TBEV), a member of the *Flavivirus* genus of the *Flaviviridae* family, is the causative agent of an infectious disease of the human central nervous system transmitted by vector ticks. TBEV has a single-stranded (+) RNA genome approximately 10.5–11 kb in size, which includes a type I 5' cap, 5' and 3' non-coding regions, and one long open reading frame (ORF) encoding three structural (C, prM, and E) and seven non-structural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) proteins (Knipe and Howley, 2013). Of the structural proteins, E protein is the main envelope glycoprotein that is exposed on the surface of the virion and plays a key role in the TBEV life cycle, including cell receptor binding, membrane fusion, and virion assembly. This protein is responsible for the formation of neutralizing antibodies against TBEV and induces a protective response in the host.

Currently, TBEV is divided into three main subtypes - European (TBEV-Eur), Far Eastern (TBEV-FE) and Siberian (TBEV-Sib) (Ecker et al., 1999; Gritsun et al., 1993; King et al., 2012). In addition, two putative TBEV subtypes have been described: Baikalian (Demina et al., 2010) and Himalayan (Dai et al., 2018). TBEV-Sib is the most common subtype and has been found in all regions where TBEV has been identified, with the exception of Western and Central Europe.

Although attempts to describe the variability of TBEV, and especially TBEV-Sib, have been made earlier (Demina et al., 2010; Kovalev et al., 2009; Pogodina et al., 2004; Tkachev et al., 2011; Zlobin et al., 2001a, 2001b), a detailed analysis of the genetic diversity of different TBEV subtypes, as well as the geographical distribution of genetic variants of different TBEV subtypes in Eurasia, has not been carried out.

In case of virus collection studies, the use of “classical” approaches based on Sanger sequencing is inefficient and time consuming. The solution to this problem may be the use of new generation sequencing (NGS) approaches to analyze large sets of samples of TBEV strains. Based on recent publications, methods for mass genomic sequencing of TBEV using NGS for viruses and TBEV in particular have not yet been clearly developed and optimized, so the goal of our project was to optimize NGS methods for complete genome sequencing of TBEV.

RNA isolation methods for further sequencing based on NGS approaches are different and can lead to ambiguous results, so it was necessary to develop an RNA isolation algorithm followed by complete genome sequencing, which would fill the gap in data on TBEV genetics by sequencing the genomes of various TBEV subtypes and their genetic lines from different collections of viruses. This approach was proposed, firstly, due to the fact that currently only about 250 complete sequences of TBEV genomes are presented in the GenBank databases (of which only about 50 are for TBEV-Sib), and secondly, only complete genome sequences are the most accurate and informative, allow us to describe the genetic diversity and evolutionary processes that occur with the virus.

MATERIALS AND METHODS

TBEV strains samples.

Within the framework of the ongoing project, 300 samples of tick-borne encephalitis virus (TBEV) strains isolated from various sources (ticks, small mammals, clinical specimens, etc.) from various geographical regions of Eurasia were taken. Brain fragments of white laboratory mice infected with TBEV strains were immersed in the DNA/RNA Shield reagent (Zymo Research, USA), which ensures the stability of nucleic acids during storage/transportation of samples at ambient temperature (4-25°C). The reagent effectively inactivates viruses and lyses cells, as described in the manufacturer's instructions.

Optimization of the viral RNA isolation method.

To identify the optimal method of RNA isolation, we used several methods, followed by a quantitative assessment of the isolated RNA (Table 1):

1. Isolation of RNA from the brain tissues of infected mice using the NucleoZOL kit (Takara Bio, Japan). To do this, tissue (25 mg) was blast frozen with liquid nitrogen and mechanically homogenized using a mortar and pestle according to the manufacturer's protocol.

2. Isolation of RNA from brain suspension using the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol.

3. Isolation of RNA from the brain tissues of infected mice, previously mechanically homogenized with a mortar and pestle, using the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol.

4. Isolation of RNA from brain suspension using the Quick-RNA™ Viral Kit (Zymo Research, USA) according to the manufacturer's protocol.

Table 1. Quantification of RNA samples No. 1,2,3 with NanoPhotometer® NP80 spectrophotometer (Implen, Germany) and a Qbit 3.0 Fluorometer BR Assay Kit (Invitrogen, USA) using different isolation methods.

1. NucleoZOL (from 25 mg of tissue)		2. QiA Viral RNA (from 140 µl of suspension)		3. QiA Viral RNA (from tissue)		4. Zymo (from 200 µl of suspension)		
Sampl es	Nanondrop, ng/µl	Qubit (BR), ng/µl	Nanondrop, ng/µl	Qubit (BR), ng/µl	Nanondrop, ng/µl	Qubit (BR), ng/µl	Nanondrop, ng/µl	Qubit (BR), ng/µl
1	237.9	151.4	28.8	26.4	97.3	74.5	57.0	41.8
2	221.6	146.5	29.4	22.4	124.4	87.0	162.8	77.2
3			14.0	9.8	43.4	35.6		

A comparison of the above methods for RNA isolation showed that the most appropriate way to isolate total RNAs in terms of price/quality/extraction time was to use the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

Viral RNA isolation.

The process of viral RNA isolation was carried out using the QIAamp Viral RNA Kit, according to the manufacturer's protocol. The quality and quantity of the isolated RNA was assessed using a Nanondrop spectrophotometer (ThermoFisher Scientific). The concentration of nucleic acids varied from 10 to 250 ng/µl (Table 2).

Table 2. An example of RNA concentration measuring with Nanodrop spectrophotometer (ThermoFisher Scientific).

Sample	Concentration, ng/μl	Sample	Concentration, ng/μl	Sample	Concentration, ng/μl
5977	105.0	13075	102.5	2033-85	31.2
8740	234.7	13080	75.4	2018-85	29.5
8788	22.4	13169	80.2	2045-85	18.6
11573	29.2	13191	41.7	2021-85	19.1

Primary screening of RNA samples for the presence of specific TBEV RNA was carried out using PCR of target RNA fragments of the TBEV genome using a set of oligonucleotide primers complementary to the E-NS1 gene region of TBEV of various subtypes and genetic lineages (Table 3). The use of PCR with these primers made it possible to reliably determine the presence of TBEV RNA in the sample.

Table 3. PCR primer sequences.

Primer's name	Sequence
TBEV_E7	5' – ggcatagaaaggctgacagtg -3'
TBEV_E10	5'- gatacctctctccacacaaccag -3'

Reverse transcription followed by PCR was performed using the BioLabMix kit (Novosibirsk, Russia). According to the following scheme (Table 4), a reaction mixture containing the necessary components was prepared and added to the RNA samples.

Table 4. Components for PCR amplification and the volume of the reaction mixture per 1 sample.

Component	Volume, μl
RT-PCR-RV	12.5
BioMaster mix	1.0
TBEV_E7 primer (1 a.u/ml)	1.0
TBEV_E10 primer (1 a.u/ml)	1.0
RNA	2.0
Nuclease free water	7.5

PCR was performed using the temperature conditions recommended below (Table 5).

Table 5. Amplification program for the detection of tick-borne encephalitis virus.

Step	Temperature, °C	Incubation time	Cycles number
Reverse transcription	45	30 min	1
Pre-denaturation	95	5 min	1
Denaturation	95	30 sec	35
Renaturation	48	30 sec	
Elongation	72	1 min	
Storage	4	~	

Amplification products were registered by electrophoresis in agarose gel (1%) in the presence of ethidium bromide. To determine the product fragments lengths, Gene Ruler (Thermo Scientific) length markers were used, which allow identification of DNA fragments from 100 to 10,000 base pairs in length (Fig. 1).

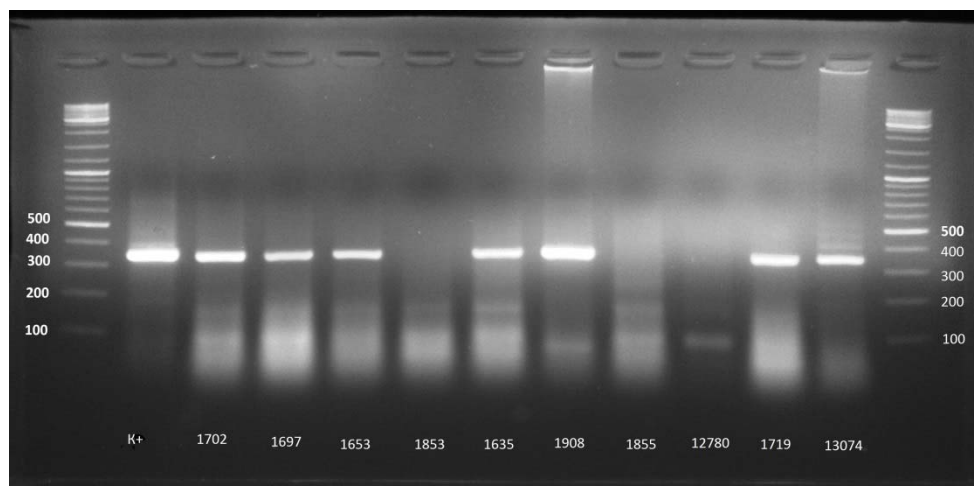


Figure 1. Example of electrophoresis of PCR fragments obtained with oligonucleotides TBEV_E7 and TBEV_E10. The outermost lanes are the Gene Ruler length markers.

Based on the results of electrophoresis, samples were taken with a PCR product of 341 bp in length for further work.

Preparation of RNA libraries.

An important criterion for the preparation of RNA libraries is the assessment of the integrity index of the isolated RNA (RNA Integrity Number (RIN)), and this indicator should be at least 7. Checking the RNA integrity of the most of our samples showed that this indicator is too low to use this RNA as a matrices (Fig.2). But, despite the low RIN, trial libraries were prepared and the results of their subsequent sequencing did not differ from samples with a high integrity index. Therefore, it was decided not to take this indicator into account when preparing libraries.

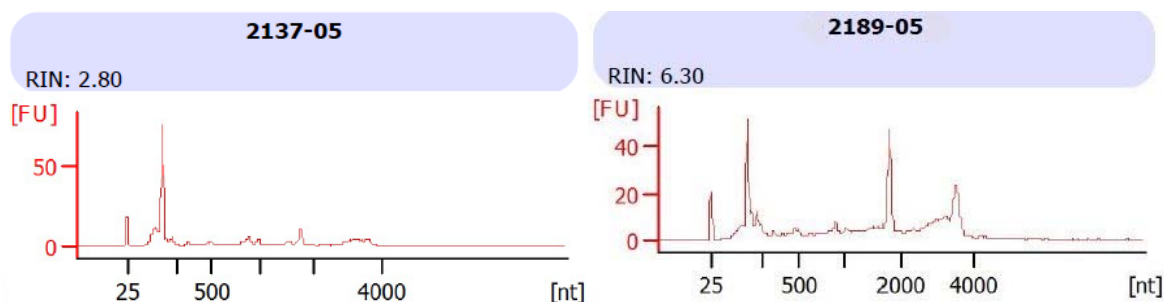


Figure 2. Evaluation of the integrity index of the isolated RNA using the Agilent 2100 Bioanalyzer instrument with the Eukaryote Total RNA Pico chip.

The RNA libraries sample preparation was performed with the KAPA RNA HyperPrep Kit (Roche, USA) according to the manufacturer's protocol. Optimal conditions were selected for some points of the protocol: fragmentation was carried out at 94°C for 4 minutes, adapters diluted 2 times were used for ligation.

When optimizing fragmentation, two options were tested. In the first case, fragmentation is performed for 5 minutes at 85°C, and this mode is recommended when using partially degraded RNA. Using this mode, we obtained libraries with longer fragments than recommended in the protocol (more than 440 bp) (Fig. 4).

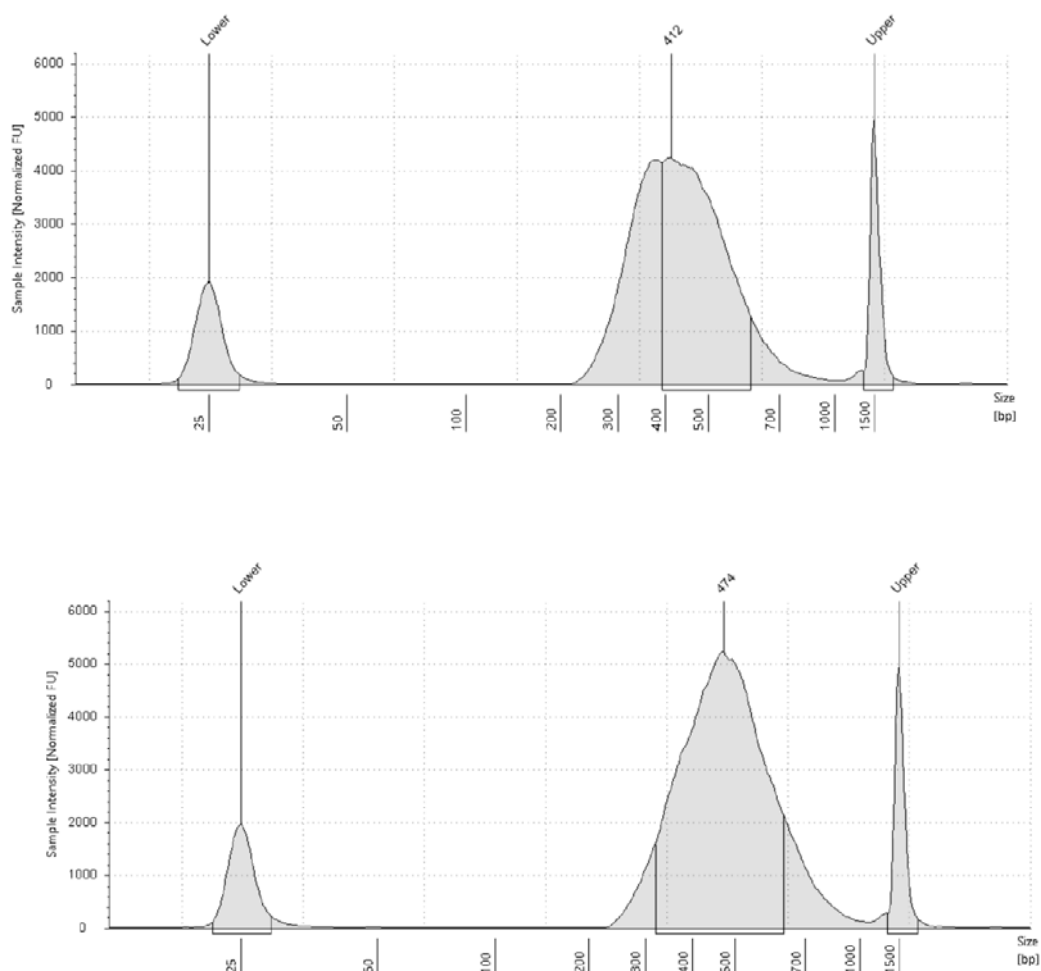


Figure 4. Length distribution of library fragments during fragmentation for 5 minutes at 85°C on Agilent Technologies 4150 TapeStation system using ScreenTape D1000 cartridges (Agilent Technologies, USA).

The other samples were fragmented for 4 minutes at 94°C and we obtained fragments that corresponded to the conditions of the protocol (fragments from 330 bp in length) (Fig. 5).

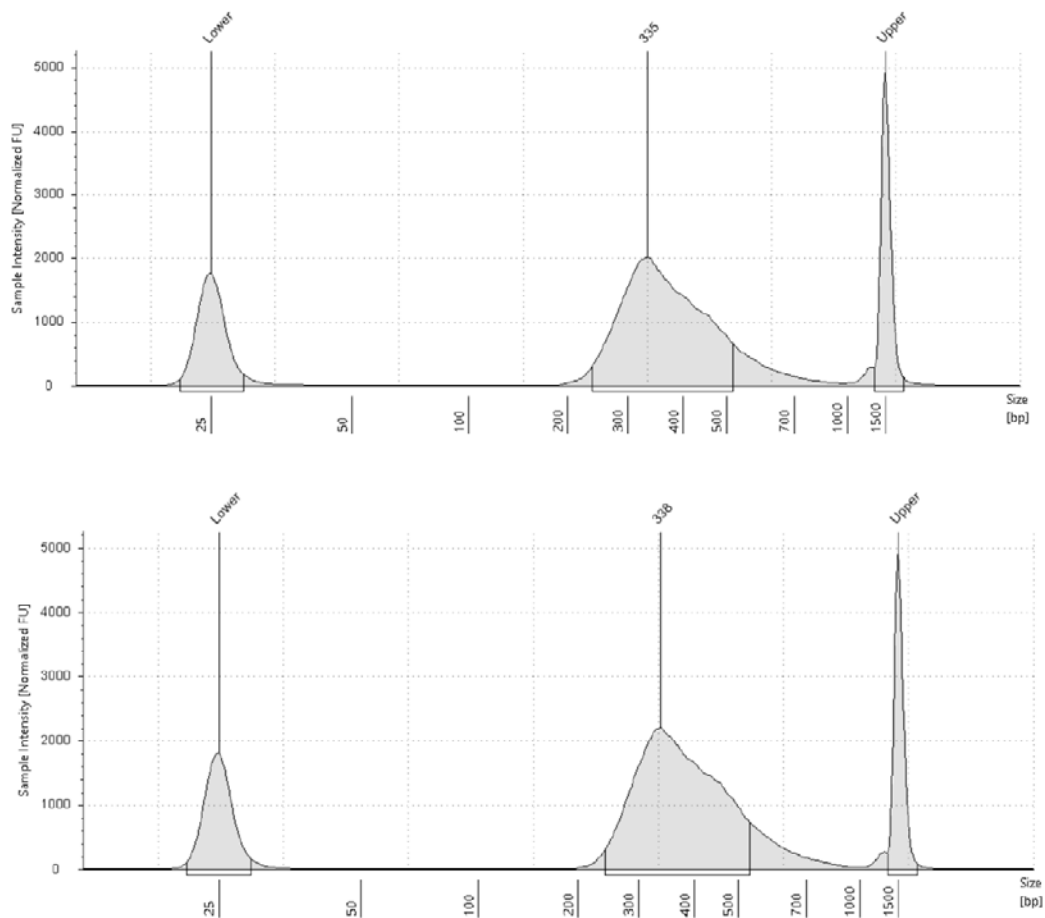


Figure 5. Length distribution of library fragments during fragmentation for 4 minutes at 94°C on an Agilent Technologies 4150 TapeStation system using ScreenTape D1000 cartridge (Agilent Technologies, USA).

The resulting libraries were quantified by measuring the concentration of the library with Qbit 3.0 Fluorometer HS Assay Kit (Invitrogen, USA), which varied from 10 to 50 ng/μl. Further, the quality of the final libraries was checked on 4150 TapeStation Agilent Technologies automated electrophoresis system using ScreenTape D1000 cartridge (Agilent Technologies, USA). The library was considered qualitative if the fragments were in the range of 200-700 bp, with a predominance in the range of 350-500 bp (Fig. 6).

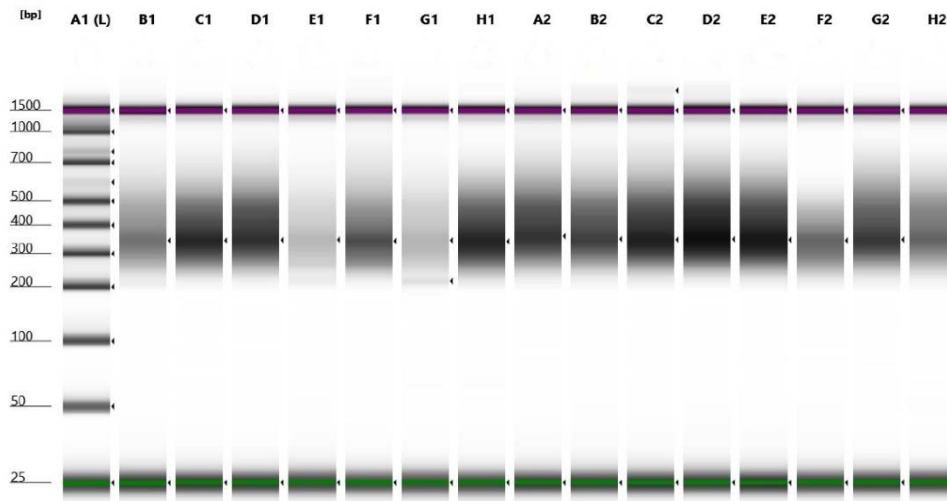


Figure 6. Electrophoregram of resulting libraries before hybridization on Agilent Technologies 4150 TapeStation system using ScreenTape D1000 cartridge (Agilent Technologies, USA). Lane A1(L) is an automatic marker with known fragments, lanes B1-H2 are resulting libraries.

Targeted enrichment of libraries.

Targeted enrichment of libraries was performed with SeqCap EZ technology (Roche, Switzerland). For this, a panel of oligonucleotides was designed and ordered, corresponding to various fragments of the TBEV genomes of all known sequences of TBEV various subtypes and genovariants currently present in the GenBank database (<https://www.ncbi.nlm.nih.gov/nucleotide/>). The enriched DNA library was qualitatively assessed with Agilent Technologies 4150 TapeStation automated electrophoresis system using ScreenTape D1000 cartridge (Agilent Technologies, USA). The average length range of the library should be in the length range of 200-700 bp, with a predominance in the length range of 300-400 bp (Fig. 7).

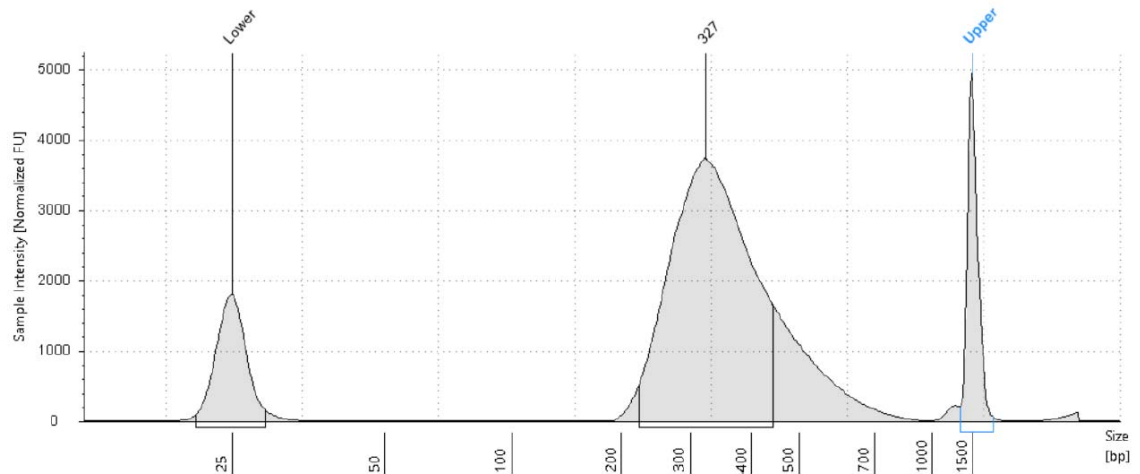


Figure 7. Fragments length distribution in the final pool of libraries on Agilent Technologies 4150 TapeStation system using ScreenTape D1000 cartridge (Agilent Technologies, USA).

The sequencing of the resulting library was performed with new generation high-performance sequencer MiSeq (Illumina). The variant of sequencing of paired end fragments (2x150) was used, the total number of cycles was 300.

Bioinformatics processing of NGS data and assembly of tick-borne encephalitis virus genomes

The process of bioinformatics processing of tick-borne encephalitis virus genomic data consisted of the following steps:

1. Quality control of raw reads with the FastQC program, which allows to determine the overall quality profile along the reads by the Phred score metric, determine the number of adapters, the range of read lengths, the presence of contamination using the GC composition and a number of other important parameters that determine the overall quality of libraries ;
2. Trimming (cleaning) data with the fastp program (Chen S. et al., 2018) from low-quality reads and trimming from technical sequences (adapters);

Further steps depend on the number of reads per sample. Thus, at sufficiently high coverage of the sample, i.e. in the presence of large number of reads per sample, a scenario is possible in which the virus genome is assembled *de novo*. With a small sample coverage, i.e. a small number of reads per sample, it is preferable to use the data mapping method with published reference TBEV genomes.

1. In the first case, with a high coverage of the library, we used the mnaviralSPAdes software, a *de novo* assembler for RNA virus datasets that takes cleaned pair-ended reads as input to the algorithmic chain and outputs sequences in fasta format.
2. The QUAST software (Gurevich A. et al., 2013) was used to evaluate the quality of the resulting assembly. At this stage, the total length of the assembly and the number of long,

- continuous DNA regions (contigs) are estimated. If the length of the assembly was from 11107 p. (total length of the GenBank reference genome: MH645618) or more, we considered that we had assembled the genome of the specific TBEV being studied.
3. Since the rnaviralSPAdes assembler software can assemble several contigs, we use the longest contig, which is stated as the TBEV genome.
 4. In case of insufficient sample coverage for genome assembly, the obtained data were mapped using the bwa mem software (Li H., 2013) with the reference TBEV genome (GenBank: MH645618).
 5. The resulting .sam files (a text format for storing biological sequences) are converted into .bam files (the binary equivalent of SAM) using the samtools utility suite with sorting and further deduplication with the picard software.
 6. Deduplicated .bam files are used to determine variants (alleles) with the GATK (Genome Analysis Toolkit) utility suite, which allows to determine variants and write them into vcf format convenient for further manipulations. After getting .vcf files, the obtained variants are filtered in each individual file according to such criteria as the depth of coverage of the variant (at least 10 reads per variant) and the quality of the variant definition (at least 20).
 7. Using the FastaAlternateReferenceMaker utility included in the GATK software package, an alternative reference genome sequence is generated in the specified interval in the fasta format. Given a set of variants, FastaAlternateReferenceMaker replaces the reference bases in the variation locations with the bases specified in the corresponding entries in the set.

Thus, during data processing, two approaches were used to determine the TBEV genome sequence, which eliminates the possibility of low-quality genome sequences.

LITERATURE

- 1) Chen S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor // *Bioinformatics*. 2018. Vol. 34, № 17. P. i884–i890.
- 2) Dai X., Shang G., Lu S., Yang J., Xu J. A new subtype of eastern tick-borne encephalitis virus discovered in Qinghai-Tibet Plateau, China // *Emerg. Microbes Infect.* 2018. 7 (1), 74. <https://doi.org/10.1038/s41426-018-0081-6>.
- 3) Demina T.V., Dzhioev Y.P., Verkhozina M.M., Kozlova I.V., Tkachev S.E., Plyusnin A., Doroshchenko E.K., Lisak O.V., Zlobin V.I. Genotyping and characterization of the geographical distribution of tick-borne encephalitis virus variants with a set of molecular probes // *J. Med. Virol.* 2010. 82 (6), 965–976. <https://doi.org/10.1002/jmv>.
- 4) Ecker M., Allison S.L., Meixner T., Heinz F.X. Sequence analysis and genetic classification of TBEV from Europe and Asia // *J. Gen. Virol.* 1999. 80, 179–185. <https://doi.org/10.1099/0022-1317-80-1-179>.
- 5) Gritsun T.S., Frolova T.V., Pogodina V.V., Lashkevich V.A., Venugopal K., Gould E.A. Nucleotide and deduced amino acid sequence of the envelope gene of the Vasilchenko strain of TBE virus; comparison with other flaviviruses // *Virus Res.* 1993. 27, 201–209. [https://doi.org/10.1016/0168-1702\(93\)90082-X](https://doi.org/10.1016/0168-1702(93)90082-X).
- 6) Gurevich A. et al. QUAST: quality assessment tool for genome assemblies // *Bioinformatics*. 2013. Vol. 29, № 8. P. 1072–1075.
- 7) Virus taxonomy: classification and nomenclature of viruses. In: King A.M.Q., Adams M.J., Carstens E.B., Lefkowitz E.J. (Eds.), Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego. 2012.
- 8) Knipe D.M., Howley P.M. (Eds.) *Fields Virology*, 6th edition. Lippincott Williams & Wilkins, a Wolters Kluwer business, Philadelphia, USA. 2013.
- 9) Kovalev S.Y., Chernykh D.N., Kokorev V.S., Snitkovskaya T.E., Romanenko V.V. Origin and distribution of tick-borne encephalitis virus strains of the Siberian subtype in the Middle Urals, the north-west of Russia and the Baltic countries // *J. Gen. Virol.* 2009. 90, 2884–2892. <https://doi.org/10.1099/vir.0.012419-0>.
- 10) Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM // arXiv [q-bio.GN]. 2013.
- 11) Pogodina V.V., Bochkova N.G., Karan L.S., Trukhina A.G., Levina L.S., Malenko G.V., Druzhinina T.A., Lukashenko Z.S., Dul'keit O.F., Platonov A.E. The Siberian and Far-Eastern subtypes of tick-borne encephalitis virus registered in Russia's Asian regions: genetic and antigen characteristics of the strains // *Vopr. Virusol.* 2004. 49, 20–25. (In Russian).
- 12) Tkachev S.E., Demina T.V., Dzhioev Y., Kozlova I.V., Verkhozina M.M., Doroshchenko E.K., Lisak O.V., Bakhvalova V.N., Paramonov A.I., Zlobin V.I. Genetic studies of tick-borne encephalitis virus strains from Western and Eastern Siberia. In: Růžek, D. (Ed.), *Flavivirus Encephalitis*. InTech, Croatia, 2011. pp. 235–254.
- 13) Zlobin V.I., Demina T.V., Mamaev L.V., Butina T.V., Belikov S.I., Gorin O.Z., Dzhioev Iu.P., Verkhozina M.M., Kozlova I.V., Voronko I.V., Adel'shin R.V., Grachev M.A., Analysis of genetic variability of strains of tick-borne encephalitis virus by primary structure of a fragment of the membrane protein E gene // *Vopr. Virusol.* 2001a. 46 (1), 12–16 (In Russian).
- 14) Zlobin V.I., Demina T.V., Belikov S.I., Butina T.V., Gorin O.Z., Adel'shin R.V., Grachev M.A. Genetic typing of tick-borne encephalitis virus based on an analysis of the levels of homology of a membrane protein gene fragment // *Vopr. Virusol.* 2001b. 46 (1), 17–22 (In Russian).