Data in Brief 55 (2024) 110607

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Optimization data for an ARTIC-/Illumina-based whole-genome sequencing protocol and pipeline for SARS-CoV-2 analysis

Christian Bundschuh[a,*], Niklas Weidner[a,b], Julian Klein[a], Tobias Rausch[c], Nayara Azevedo[c], Anja Telzerow[a,c], Katharina Laurence Jost[a], Paul Schnitzler[a], Hans-Georg Kräusslich[a,d], Vladimir Benes[c]

[a] *Medical Faculty Heidelberg, Department of Infectious Diseases Virology, Heidelberg University, Heidelberg, Germany*
[b] *Medical Faculty Heidelberg, Department of Infectious Diseases, Microbiology and Hygiene, Heidelberg University, Heidelberg, Germany*
[c] *Genomics Core Facility, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
[d] *Deutsches Zentrum für Infektionsforschung, partner site Heidelberg, Germany*

ARTICLE INFO

Article history:
Received 10 April 2024
Revised 29 May 2024
Accepted 4 June 2024
Available online 12 June 2024

Dataset link: SARS-CoV-2 analysis pipeline for short-read, paired-end illumina sequencing (Original data)

Keywords:
Virology
Epidemiology
Whole genome sequencing
Bioinformatics
SARS-CoV-2

ABSTRACT

In January 2021, Germany commenced surveillance of SARS-CoV-2 variants under the Corona Surveillance Act, which ceased in July 2023. The objective was to bolster pandemic control, as specific alterations in amino acids, particularly within the spike protein, were linked to heightened transmission and decreased vaccine effectiveness.

Consequently, our team conducted whole genome sequencing using the commercially accessible ARTIC protocol on Illumina's NextSeq500 platform and MiSeq for SARS-CoV-2 positive samples obtained from patients at Heidelberg University Hospital, affiliated hospitals, and the public health office in the Rhine-Neckar/Heidelberg region. Throughout the pandemic, we refined the existing ARTIC V4 protocol as well as our bioinformatics pipeline, the details of which are outlined

[*] Corresponding author.
E-mail address: christian.bundschuh@med.uni-heidelberg.de (C. Bundschuh).

https://doi.org/10.1016/j.dib.2024.110607

in this report. This report reflects the protocol for the MiSeq analysis, the protocol for the NextSeq500 can be found in our previous publication.

Specifications Table

Subject	*Genomics; Virology*
Specific subject area	*Establishment of an optimized ARTIC- and Illumina-based Whole-genome sequencing protocol and pipeline for SARS-CoV-2*
Type of data	*Raw data* *Protocol, Figure* *Processed data/optimized protocol*
Data collection	*Inclusion criteria was a PCR cycle threshold of <32.* *RNA isolation was performed by the utilization of automated magnetic bead-based nucleic acid extraction protocols, including QIASymphony, DSP Virus/Pathogen mini-Kit (Qiagen) or the Chemagic Viral DNA/RNA 300 Kit H96 (PerkinElmer).* *Library preparation adhered mainly to the original ARTIC protocol but with the described optimizations.* *Sequencing data was generated on an Illumina NextSeq500 and MiSeq and the data evaluation was performed with the described bioinformatic tools, such as trim galore, kraken2, Burrows-Wheeler Alignment tool, SAMtools, Alfred, FreeBayes, bcftools, Ensembl Variant Effect Predictor, Pangolin and Nextclade.*
Data source location	*Heidelberg, Germany* *Heidelberg University and European Molecular Biology Laboratory Heidelberg (EMBL)*
Data accessibility	Repository name: GitHub/Zenodo Data identification number: tobiasrausch/covid19 [1] Direct URL to data: https://doi.org/10.5281/zenodo.10847332 Instructions for accessing these data: Python-based script
Related research article	Bundschuh, C., Weidner, N., Klein, J., Rausch, T., Azevedo, N., Telzerow, A., Mallm, J. P., Kim, H., Steiger, S., Seufert, I., Börner, K., Bauer, K., Hübschmann, D., Jost, K. L., Parthé, S., Schnitzler, P., Boutros, M., Rippe, K., Müller, B., Bartenschlager, R., Kräusslich, H. G. & Benes, V. 2024. Evolution of SARS-CoV-2 in the Rhine-Neckar/Heidelberg Region 01/2021 - 07/2023. *Infect Genet Evol*, 119**,** 105,577 [2].

1. Value of the Data

- Allows the easy establishment of a free, public and open-source SARS-CoV-2 whole genome sequencing bioinformatics pipeline.
- Full protocol on the library preparation allows for an easy establishment in any laboratory with the necessary tools.
- Protocol optimizations reduce hands-on time and costs per analysis.

2. Background

Towards the end of 2019, a novel coronavirus (2019-nCoV), later named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the World Health Organization (WHO), emerged as the causative agent for a surge in pneumonia cases known as coronavirus disease (COVID-19) in Wuhan, China [3]. The rapid spread of the virus led to the escalation of the epidemic in China into a global pandemic, with over 676 million confirmed SARS-CoV-2 cases by March 2023 (data not updated beyond that point) [4,5].

The Spike (S) protein, comprising 1273 amino acids, was early identified as crucial for SARS-CoV-2 host cell entry (Xia et al., 2020). Consequently, various mutations in the S protein have been associated with increased viral transmissibility, disease severity, reinfection despite natural immunity, and vaccine efficacy. As a result, monitoring circulating variants and their respective mutations has become a vital epidemiological strategy for pandemic control [6], making viral genome sequencing essential for epidemiological surveillance.

This report outlines our group's adapted and refined whole genome sequencing protocol, ARTIC V4, initially developed by New England Biolabs [7], as well as the optimized bioinoformatics pipeline.

3. Data Description

For the easy establishment of a free, public and open-source SARS-CoV-2 whole genome sequencing bioinformatics pipeline we provide the necessary tools in this publication.

The method section is separated in two distinctive parts:

First part is the wet-lab part of the protocol. Here, we provide the applied protocols for RNA purification, library preparation (ARTIC protocol) with the respective substeps, such as cDNA synthesis, cDNA amplification, fragmentation/end prep, adaptor ligation, PCR Enrichment of Adaptor-ligated DNA/Dual Index PCR, library pooling, pool cleanup and library quality control.

These steps do not generate any data, but are the necessary wet-lab preparation of the samples to generate the bioinformatic/sequencing data for the specimen analysis.

Second part contains the required tools for the bioinformatic pipeline setup for the sequence analysis.

The necessary steps for the data analysis include:

Sequence adapter trimming.
Host-read contamination removal.
Alignment to SARS-CoV-2 reference genome.
Variant calling.
Consensus computation.
Lineage classification.
Summary report generation.

Therefore we mainly utilized the following scripts/programs (a complete overview of all utilized programs/scripts is provided in chapter 5 Data analysis):

For bioinformatics processing, we first filtered adapters and contaminating host reads [8].

We then mapped the reads to the SARS-CoV-2 reference genome, masked the priming regions and called variants using FreeBayes [9].

Variant annotation was performed using the Ensembl Variant Effect Predictor [10].

For lineage classification, we first generated a viral consensus sequence using iVar [11], which was then classified using Nextclade [12] and Pangolin [13].

For process optimization, we utilized a custom Python script, available in the GitHub source code repository (https://doi.org/10.5281/zenodo.10847332), to consolidate all quality control metrics, lineage labels, and variants resulting in amino acid changes. This script facilitated the generation of comprehensive reports, including metadata tracking sheets and gzipped FASTA files for all viral assemblies. These files were prepared for immediate upload to the German electronic sequencing data hub (DESH) operated by the RKI.

The programs/scripts available in the aforementionedGitHub repository have the subsequent folder structure [1]:

example: This folder contains an exemplary data analysis with a SARS-CoV-2 COG-UK sample
kraken2: contains the files for the respective application as well as a readme
ref: includes all files for pangolin and a readme on how to update and use pangolin
scripts: contains the scripts required for the process of adapter-trimming, host-read contamination removal, sequence alignment, variant calling, consensus computation and lineage classification
src: contains the updated pangolin files
README.md contains an explanation of the repository and the subfolders/scripts etc.
Makefile is the script for pangolin updating
License contains the license clause
The generated data sets are exported as fastq-Files (for the sequence, after the assembly) and Excel-Files with the relevant specimen data (including the mutation patterns, the clade and Pangolin classification).

4. Experimental Design, Materials and Methods

4.1. RNA purification

For the analysis of SARS-CoV-2 sequencing, RNA was extracted from specimens collected from the upper respiratory tract, including nasopharyngeal and oropharyngeal swabs, as well as pharyngeal washes. The extraction process involved automated magnetic bead-based nucleic acid extraction protocols.

Our group employed either the QIASymphony DSP Virus/Pathogen mini-Kit from Qiagen or the Chemagic Viral DNA/RNA 300 Kit H96 from PerkinElmer for magnetic bead RNA extraction, following the protocols provided by the manufacturers.

4.2. ARTIC protocol

The NEBNext ARTIC SARS-CoV-2 FS Library Prep Kit for Illumina (E7650) contains the enzymes, buffers and oligonucleotides required to convert a broad range of total RNA into high quality, targeted, libraries for next-generation sequencing on the Illumina platform. Primers targeting the human EDF1 (NEBNext ARTIC Human Primer Mix 1) and NEDD8 (NEBNext ARTIC Human Primer Mix 2) genes are supplied as optional internal controls. The fast, user-friendly workflow also has minimal hands-on time. Each kit component must pass rigorous quality control standards, and for each new lot the entire set of reagents is functionally validated together by construction and sequencing of indexed libraries on an Illumina sequencing platform. For larger volume requirements, customized and bulk packaging is available by purchasing through the OEM/Bulks department at NEB.

4.3. cDNA synthesis

The presence of carry-over products can interfere with sequencing accuracy, particularly for low copy targets. Therefore, it is important to carry out the appropriate no template control (NTC) reactions to demonstrate that positive reactions are meaningful.

Gently mix and spin down the LunaScript RT SuperMix reagent. Prepare the cDNA synthesis reaction as described in Table 1:

- For no template controls, add water instead of RNA.
- Each RNA plate should have a no template (H_2O) and Twist RNA (12,500 copies) as controls.
- The wells to add the controls should be constant, as presented in Fig. 1.

Table 1
LunaScript RT SuperMix pipetting scheme.

Component	1 rxn Volume (µL)	96-well plate Volume (µL)
RNA sample	4	–
LunaScript RT SuperMix	1	100
MM/strip tube:	25 µL	

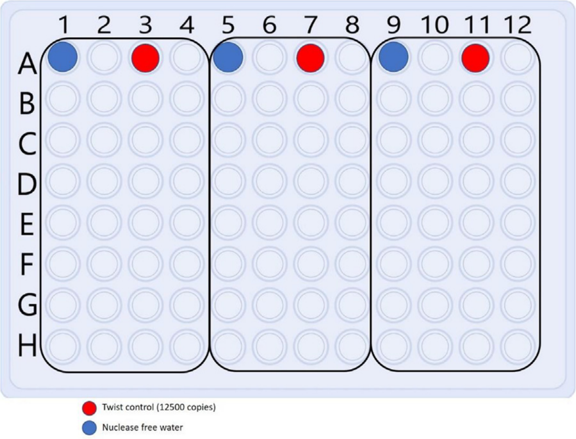

Fig. 1. Plate positions for controls.

Table 2
Thermocycler settings for cDNA synthesis reaction.

Cycle Step	Temperature	Time	Cycles
Primer Annealing	25 °C	2 min	1
cDNA Synthesis	55 °C	20 min	
Heat Inactivation	95 °C	1 min	
Hold	4 °C	∞	

*Set heated lid to 105 °C.

The controls should be switched each plate!
Seal the plate, vortex and spin prior to placing it in the thermocycler.
Incubate the reactions in a thermocycler with the settings shown in Table 2.
Samples can be stored at −20 °C for up to a week.

4.4. cDNA amplification

Note: 2.5 µl cDNA input is recommended. If using less than 2.5 µl of cDNA, add nuclease-free water to a final volume of 2.5 µl. We recommend setting up the cDNA synthesis and cDNA amplification reactions in different rooms to minimize cross-contamination of future reactions.

Gently mix and spin down reagents. Prepare the split pool cDNA amplification reactions as described in Tables 3 and 4.

Seal the plate, vortex and spin prior to placing it in the thermocycler.
Incubate reactions in a thermocycler with the settings shown in Table 5.
Combine the Pool A and Pool B PCR reactions for each sample.

Table 3
Pipetting scheme for Pool Set A.

Component	1 rxn Volume (µL)	96-well plate Volume (µL)
cDNA	2.5	–
Q5 Hot Start High-Fidelity 2X MM	2.5	250
ARTIC SARS-CoV-2 Primer Mix 1	0.7	70

Table 4
Pipetting scheme for Pool Set B.

Component	1 rxn Volume (µL)	96-well plate Volume (µL)
cDNA	2.5	–
Q5 Hot Start High-Fidelity 2X MM	2.5	250
ARTIC SARS-CoV-2 Primer Mix 2	0.7	70

Table 5
Thermocycler settings for pool cDNA amplification reaction.

Cycle Step	Temperature	Time	Cycles
Initial Denaturation	98 °C	30 s	1
Denaturation	95 °C	15 s	35
Annealing/Extension	63 °C	5 min	
Hold	4 °C	∞	1

*Set heated lid to 105 °C.

Table 6
Expected Qubit BR values for controls and samples.

ID	Expected Qubit BR value
H_2O negative control	4–20
Positive control	100–200
Sample	100–200

4.5. Quality control

We used to perform Qubit as well as Tapestation 4150, but we found that Qubit alone was predictive of whether libraries would succeed. We measured the controls and 8 random samples (2 samples from each source plate) with Qubit. The expected Qubit BR values for the respective substances is shown in Table 6.

Samples can be stored at −20 °C for up to a week.

4.6. Fragmentation/End prep

Mix 11 µl of the pooled amplicons (from step 2.4) with 99 µl of nuclease free water in a new 96well plate (Eppendorf TwinPlate).

Ensure that the Ultra II FS Reaction Buffer is completely thawed. If a precipitate is seen in the buffer, pipette up and down several times to break it up, and quickly vortex to mix. Place on ice until use.

Vortex the Ultra II FS Enzyme Mix 5–8 s prior to use and place on ice.

Note: It is important to vortex the enzyme mix prior to use for optimal performance.

Table 7
Pipetting scheme for fragmentation/end prep reaction.

Component	1 rxn Volume (µL)	96-well plate Volume (µL)
cDNA amplified	6.5	–
NEBNext Ultra II FS reaction Buffer	1.5	150
NEBNext Ultra II FS Enzyme Mix	0.5	50

Table 8
Thermocycler settings for fragmentation/end prep reaction.

Temperature	Time
37 °C	30 min
65 °C	30 min
4 °C	∞

*Set heated lid to 75 °C.

Table 9
Pipetting scheme for adapter ligation reaction.

Component	1 rxn Volume (µL)	96-well plate Volume (µL)
FS reaction Mixture	8.5	–
NEBNext Adaptor for Illumina	0.5	50
NEBNext Ultra II Ligation MM	6	600

*Mix the Ultra II Ligation Master Mix by pipetting up and down several times prior to adding to the reaction.
**The NEBNext adaptor is provided in NEBNext Oligo kits.

Add the following components to a 0.2 ml thin wall PCR tube on ice. The respective pipetting scheme is shown in Table 7.

Seal the plate, vortex and spin prior to placing it in the thermocycler.

In a thermocycler, run a program with the settings shown in Table 8.

If necessary, samples can be stored at –20 °C; however, a slight loss in yield (~20 %) may be observed. We recommend continuing with adaptor ligation before stopping.

4.7. Adaptor ligation

Add the following components directly to the FS Reaction Mixture. The respective pipetting scheme is shown in Table 9.

Seal the plate, vortex and spin prior to placing it in the thermocycler.

Incubate at 20 °C for 15 min in a thermocycler with the heated lid off.

Add 0.75 µl of • (red) USER® Enzyme to the ligation mixture from Step 4.3.

Mix well and incubate at 37 °C for 15 min with the heated lid set to ≥ 47 °C.

Samples can be stored overnight at –20 °C.

4.8. PCR enrichment of adaptor-ligated DNA/Dual index PCR

Add the following components to a sterile strip tube. The respective pipetting scheme is shown in Table 10.

Seal the plate, vortex and spin before starting the program in the thermocycler.

Place the tube on a thermocycler and perform PCR amplification using the PCR cycling conditions shown in Table 11.

Table 10
Pipetting scheme for enrichment of adaptor-ligated DNA/Dual Index PCR reaction.

Component	1 rxn Volume (μL)	96-well plate Volume (μL)
Adaptor Ligated DNA Fragments	3	–
NEBNext Library PCR MM	5	500
Dual Index Mix	3	–

*Oligos from NEB dual index 96-well plate. 4 different sets.
**The leftover of the Adaptor Ligated DNA Fragments can be stored overnight @ –20 °C.

Table 11
Thermocycler settings for enrichment of adaptor-ligated DNA reaction.

Cycle Step	Temperature	Time	Cycles
Initial Denaturation	98 °C	30 s	1
Denaturation	98 °C	10 s	6*
Annealing/Extension	65 °C	75 s	
Final Extension	65 °C	5 min	1
Hold	4 °C	∞	

* Set heated lid to 105 °C. The number of PCR cycles recommended should be viewed as a starting point and may need to be optimized for particular sample types.

4.9. Pooling of libraries

Without any QC step, all libraries from a plate should be pooled together.

Pipette 4 μL of each library into a strip with a multichannel pipette and combine all volume into a 1.5 mL Eppendorf tube (∼380 μL).

4.10. Cleanup of the pool

Vortex NEBNext Sample Purification Beads to resuspend.

Measure the pool volume and add 0.9X of resuspended beads to the PCR reaction. Mix well by pipetting up and down at least 10 times. Be careful to expel all of the liquid out of the tip during the last mix. Vortexing for 3–5 s on high can also be used. If centrifuging samples after mixing, be sure to stop the centrifugation before the beads start to settle out.

Incubate samples on bench top for at least 5 min at room temperature.

Place the tube/plate on an appropriate magnetic stand to separate the beads from the supernatant. If necessary, quickly spin the sample to collect the liquid from the sides of the tube or plate wells before placing on the magnetic stand.

After 5 min (or when the solution is clear), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets.

Note: do not discard the beads.

Add 200 μl of 80 % freshly prepared ethanol to the tube/plate while in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets.

Repeat the precious washing step for a total of two washes. Be sure to remove all visible liquid after the second wash. If necessary, briefly spin the tube/plate, place back on the magnet and remove traces of ethanol with a p10 pipette tip.

Air dry the beads for up to 5 min while the tube/plate is on the magnetic stand with the lid open.

Note: Do not over-dry the beads. This may result in lower recovery of DNA.

Remove the tube/plate from the magnetic stand. Elute the DNA target from the beads by adding 105 μl of 0.1× TE. Mix well by pipetting up and down 10 times, or on a vortex mixer. Incubate for at least 2 min at room temperature.

If necessary, quickly spin the sample to collect the liquid from the sides of the tube or platewells before placing back on the magnetic stand.

Place the tube/plate on the magnetic stand. After 5 min (or when the solution is clear), transfer 100 µl to a new eppendorf tube.

Repeat the cleanup steps with the 0.9× of resuspended beads

Elute the DNA target from the beads by adding 55 µl of 0.1× TE and mix well

Place the tube/plate on the magnetic stand. After 5 min (or when the solution is clear), transfer 50 µl to a new eppendorf tube and store at –20 °C.

4.11. Library QC

Measure the concentration of the final pool with Qubit DNA BR (HS DNA reagents) kit.

Assess the library size distribution with Agilent Tapestation 4150 with D1000 tape and D1000 reagents. The sample may need to be highly diluted (in rare cases).

The expected concentration is between 100 and 200 ng/ul by BR Qubit.

A peak sized at 250 bp is expected on a Tapestation 4150, based on a 30-minute fragmentation time.

4.12. Sequencing

Sequencing was conducted using an Illumina MiSeq System. This system is a desktop sequencer supported by automated software, offering various analytical capabilities including whole genome sequencing, exome sequencing, whole transcriptome sequencing, mRNA-Seq, and methylation sequencing.

The MiSeq sequencing method is based on 2-channel sequencing by synthesis, employing fluorescence-labeled nucleotide bases. These bases are incorporated into the (c)DNA template strands, and detection is achieved accordingly. Each sequencing cycle involves reversible terminator-bound deoxyribonucleotide triphosphates (dNTPs), minimizing incorporation bias and reducing error rates.

Users can adjust settings and configurations to meet specific requirements, such as choosing between high or mid output flow cells and opting for paired-end reads of either 75 or 150 bp read length. For our experiments, we utilized paired-end reads with a 75 bp length (Fig. 2).

5. Data Analysis

To analyze the ARTIC sequencing data, samples were consistently sequenced in paired-end mode with a read length of 75 bp, utilizing the NEBNext ARTIC library preparation kit. Our data analysis involved a multi-stage pipeline comprising several tools for tasks such as sequencing adapter trimming, alignment, quality control, mutation calling, viral genome assembly, and lineage classification. The specific tools used in each stage are detailed in the following sections.

5.1. Trim galore (https://zenodo.org/badge/latestdoi/62039322)

To mitigate the potential presence of sequencing adapters in the final sequencing read caused by short input fragments, we utilized trim galore with default parameters to effectively trim off these adapters.

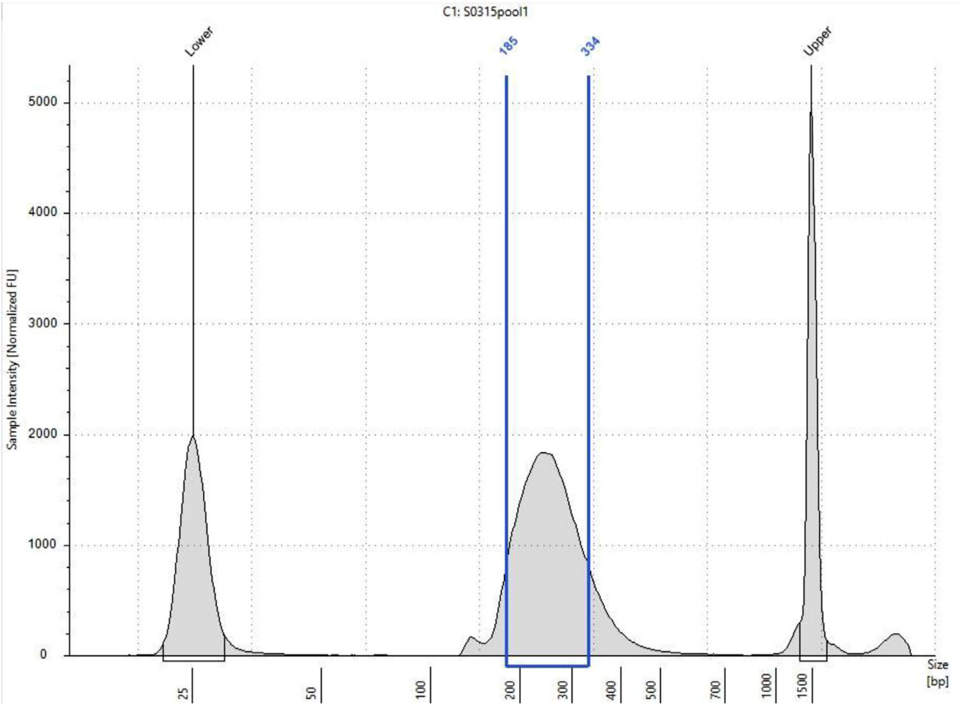

Fig. 2. Example of final library size distributions on a Tapestation 4150.

5.2. kraken2 [8]

kraken2 is a k-mer based method, which can be used for the filtering of contaminating human-derived reads in the sequencing data due to unspecific amplicon primer binding (host-read contamination removal). This process was optimized by conducting an alignment of the generated sequencing data against the host reference (build GRCh38) and a screening for potential laboratory contaminants.

5.3. Burrows-Wheeler alignment tool [14]

Burrows-Wheeler Alignment tool was applied for the alignment of the sequencing data against the SARS-CoV-2 reference genome (NC_045512.2).

5.4. SAMtools [15]

The sorting and indexing of the alignments were achieved by using SAMtools.

Secondary usage for the generation of the viral consensus sequence and its subsequent quality control (consensus computation).

5.5. Alfred [16]

Alfred was used for the quality control of the SAMtools data.

Secondary usage for the generation of the viral consensus sequence and its subsequent quality control (consensus computation).

5.6. iVar [11]

iVar was applied to mask priming regions and thereby prevent a possible variant calling bias in these regions. Due to the overlapping amplicon design of the applied protocol this did not result in further dropout (unobserved) regions.

Secondary usage for the generation of the viral consensus sequence and its subsequent quality control (consensus computation).

The parameters for iVar consensus and quality control were established in accordance with the Robert Koch Institute (RKI) SARS-CoV-2 sequencing submission criteria. These criteria include ensuring a consensus sequence with at least 90 % informative sequence, less than 5 % unobserved bases (Ns), a minimum coverage of 20×, and a minimum of 90 % read support for informative positions [17].

5.7. FreeBayes [9]

FreeBayes was utilized for variant calling.

5.8. bcftools [18]

bcftools was used for quality filtering and normalization of variants.

Secondary usage for the generation of the viral consensus sequence and its subsequent quality control (consensus computation).

5.9. Ensembl variant effect predictor [10]

Subsequently, all identified variants were annotated with the Ensembl Variant Effect Predictor for their functional consequences.

5.10. Pangolin [13]

Pangolin (with default parameters) was used for the conducting of lineage classification.

5.11. Nextclade [12]

Nextclade (with default parameters) was utilized for the clade classification process.

5.12. Custom python script (https://doi.org/10.5281/zenodo.10847332)

We utilized a custom Python script, available in the GitHub source code repository (https://doi.org/10.5281/zenodo.10847332), to consolidate all quality control metrics, lineage labels, and variants resulting in amino acid changes. This script facilitated the generation of comprehensive reports, including metadata tracking sheets and gzipped FASTA files for all viral assemblies. These files were prepared for immediate upload to the German electronic sequencing data hub (DESH) operated by the RKI.

5.13. GEAR genomics [19]

In rare instances, leveraging methods of GEAR genomics were utilized for the validation process of involved primer design and analysis of Sanger sequencing chromatograms, which was a prerequisite for the confirmation of noteworthy sequencing findings.

Limitations

Not applicable.

Ethics Statement

This study was approved by the ethics committee of the Medical Faculty at the University of Heidelberg for the analysis of proband samples by whole genome sequencing of the viral RNA (S-316/2021).

Furthermore, the sequencing of SARS-CoV-2 positive samples adhered to the governmental regulations outlined in the Coronavirus-Surveillanceverordnung (CorSurV) by Germany and Baden-Württemberg.

CRediT Author Statement

Christian Bundschuh: Writing – review & editing, Writing – original draft, Investigation, Data curation. **Niklas Weidner:** Writing – review & editing, Investigation, Data curation. **Tobias Rausch:** Writing – review & editing, Formal analysis, Data curation. **Nayara Azevedo:** Investigation. **Anja Telzerow:** Investigation. **Katharina Laurence Jost:** Writing – review & editing, Investigation. **Paul Schnitzler:** Writing – review & editing, Supervision. **Hans-Georg Kräusslich:** Writing – review & editing, Supervision, Conceptualization. **Vladimir Benes:** Writing – review & editing, Investigation, Formal analysis.

Data Availability

SARS-CoV-2 analysis pipeline for short-read, paired-end illumina sequencing (Original data) (Github/Zenodo).

Acknowledgments

We thank all diagnostics employees of the Virology department of the University Hospital Heidelberg and EMBL GeneCore's NGS protocol-automation team for their support. This work was supported by the program for surveillance and control of SARS-CoV-2 mutations of the state of Baden-Württemberg and the German Federal Research Network Applied Surveillance (CorSurV).

For the publication fee we acknowledge financial support by Heidelberg University.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Rausch T. 2024. SARS-CoV-2 data analysis. https://doi.org/10.5281/zenodo.10847332.

[2] C. Bundschuh, N. Weidner, J. Klein, T. Rausch, N. Azevedo, A. Telzerow, J.P. Mallm, H. Kim, S. Steiger, I. Seufert, K. Börner, K. Bauer, D. Hübschmann, K.L. Jost, S. Parthé, P. Schnitzler, M. Boutros, K. Rippe, B. Müller, R. Bartenschlager, H.G. Kräusslich, V. Benes, Evolution of SARS-CoV-2 in the Rhine-Neckar/Heidelberg Region 01/2021 - 07/2023, Infect. Genet. Evol. 119 (2024) 105577.

[3] World Health Organization (WHO) 2020. Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020.

[4] Center for Systems Science and Engineering (Johns Hopkins University). 2023. *COVID-19 Dashboard* [Online]. Available: https://coronavirus.jhu.edu/map.html [Accessed 19/01/2023].

[5] World Health Organization (WHO) 2023. Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update.

[6] S.S. Abdool Karim, T. de Oliveira, New SARS-CoV-2 variants - clinical, public health, and vaccine implications, N. Engl. J. Med. (2021) published online 24 March 2021.

[7] Tyson, J.R., James, P., Stoddart, D., Sparks, N., Wickenhagen, A., Hall, G., Choi, J.H., Lapointe, H., Kamelian, K., Smith, A.D., Prystajecky, N., Goodfellow, I., Wilson, S.J., Harrigan, R., Snutch, T.P., Loman, N.J. & Quick, J. 2020. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*.

[8] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257.

[9] Garrison, E. & Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv,* 1207.

[10] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The ensembl variant effect predictor, Genome Biol. 17 (2016) 122.

[11] N.D. Grubaugh, K. Gangavarapu, J. Quick, N.L. Matteson, J.G. De Jesus, B.J. Main, A.L. Tan, L.M. Paul, D.E. Brackney, S. Grewal, N. Gurfield, K.K.A. Van Rompay, S. Isern, S.F. Michael, L.L. Coffey, N.J. Loman, K.G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar, Genome Biol. 20 (2019) 8.

[12] J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics 34 (2018) 4121–4123.

[13] A. Rambaut, E.C. Holmes, Á. O'Toole, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, Nat. Microbiol. 5 (2020) 1403–1407.

[14] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.

[15] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[16] T. Rausch, M. Hsi-Yang Fritz, J.O. Korbel, V Benes, Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing, Bioinformatics 35 (2019) 2489–2491.

[17] Robert Koch Institute (RKI). 2021. *Qualitätsvorgaben für die Sequenzdaten* [Online]. Available: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Qualitaetskriterien.pdf (Accessed 06/12/2021).

[18] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993.

[19] T. Rausch, M.H. Fritz, A. Untergasser, V. Benes, Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files, BMC Genomics 21 (2020) 230.