Computational and Structural Biotechnology Journal 20 (2022) 2292–2296

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Reconstruction of the origin of the first major SARS-CoV-2 outbreak in Germany

Marek Korencak [a], Sugirthan Sivalingam [b,c], Anshupa Sahu [b,c], Dietmar Dressen [d], Axel Schmidt [b], Fabian Brand [b], Peter Krawitz [b], Libor Hart [e], Anna Maria Eis-Hübinger [a], Andreas Buness [b,c], Hendrik Streeck [a,*]

[a] Institute of Virology, University Hospital Bonn, Venusberg-Campus 1, Bonn 53127, Germany
[b] Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University of Bonn, Venusberg-Campus 1, Bonn 53127, Germany
[c] Core Unit for Bioinformatics Data Analysis, Medical Faculty, University of Bonn, Venusberg-Campus 1, Bonn 53127, Germany
[d] Labor Mönchengladbach MVZ Dr. Stein & Kollegen GbR, Tomphecke 45, Mönchengladbach 41169, Germany
[e] Department of Oral and Maxillofacial Surgery, University of Duisburg-Essen, Henricistr. 92, Essen 45136, Germany

ARTICLE INFO

Article history:
Received 11 March 2022
Received in revised form 6 May 2022
Accepted 6 May 2022
Available online 10 May 2022

Keywords:
SARS-CoV-2
Outbreak
Phylogenetic analysis
Sequencing

ABSTRACT

The first major COVID-19 outbreak in Germany occurred in Heinsberg in February 2020 with 388 officially reported cases. Unexpectedly, the first outbreak happened in a small town with little to no travelers. We used phylogenetic analyses to investigate the origin and spread of the virus in this outbreak. We sequenced 90 (23%) SARS-CoV-2 genomes from the 388 reported cases including the samples from the first documented cases. Phylogenetic analyses of these sequences revealed mainly two circulating strains with 74 samples assigned to lineage B.3 and 6 samples assigned to lineage B.1. Lineage B.3 was introduced first and probably caused the initial spread. Using phylogenetic analysis tools, we were able to identify closely related strains in France and hypothesized the possible introduction from France.

1. Introduction

In December 2019 China reported several fatal pneumonia cases. Shortly afterward, Zhou et. al. identified the cause of those deaths: a novel coronavirus, which was closely related to SARS-CoV and was later named SARS-CoV-2 [1]. Since then, the virus has spread through all continents, and the World Health Organization (WHO) has declared a pandemic. While in most contries the first SARS-CoV-2 outbreaks occurred in major cities including Milan [2], Manchester [3] or Chicago [4] or high-density traffic hubs [5–7], the first outbreak in Germany happened in Heinsberg, a small relatively unknown town with little to no tourism [8]. After a carnival session where super-spreading occurred, it was reported that about 3.1% of the local population was PCR-positive [8]. However, until today it is uncertain how the virus was introduced in the first place to this town and how it was able to spread from thereafter.

The virus strains circulating today evolved from the original Wuhan strain by accumulating different types of mutations. In general, RNA viruses have very high mutation rates, which can be up to a million times higher compared to their hosts, which may correlate with enhanced virulence and other traits considered beneficial for virus replication [9]. Sequencing data suggest that coronaviruses change slower than most other RNA viruses. This is likely due to a proofreading enzyme that corrects copying mistakes [10]. At the root of the phylogeny of SARS-CoV-2 are two lineages that were denoted as lineages A and B. The earliest lineage A virus (GISAID EPI_ISL_406801) was sampled on January 5, 2020. There are two nucleotide positions, which help us to distinguish between these two lineages. While the early lineage A shares those two nucleotides with the closest known bat virus, lineage B viruses have different nucleotides on these sites. An early representative of B lineage is Wuhan-Hu-1 (GenBank accession MN908947) sampled on December 26, 2019 [11]. Rambaut et. al. identified six lineages derived from lineage A (denoted A.1-A.6) and two descendant sublineages of A.1 (A.1.1 and A.3). They also described 16 lineages, which were directly derived from lineage B. Lineage B.1 is the predominant lineage globally and it has been divided into more than 70 sub-lineages. Creating common and generally agreed upon nomenclature of viruses circulating in different places will help to provide links between outbreaks that share similar viral

* Corresponding author.
E-mail address: hendrik.streeck@ukbonn.de (H. Streeck).

https://doi.org/10.1016/j.csbj.2022.05.011

genomes. For this purpose, an algorithm named Phylogenetic Assignment of Named Global Outbreak LINeages (pangolin) was implemented [11]. To date, millions of genomes of SARS-CoV-2 have been sequenced worldwide providing us with a detailed picture of the molecular evolution of the virus.

In this study, we used phylogenetic analysis on samples collected from the first outbreak in Germany to retrospectively investigate the route of introduction and onward transmission of SARS-CoV-2. Our data demonstrate that there were two circulating lineages, B.3 and B.1, introduced at different time points, with lineage B.3 being introduced first. Using phylogenetic analysis, we observe that the majority of our samples could be assigned to one part of the European SARS-CoV-2 phylogenetic tree. This branch of the tree contains samples dating earlier, and of those the majority were from France. This data suggests France as possible source of this outbreak and also illustrating how phylogenetic analysis can retrospectively add insights regarding the spread of the virus.

2. Materials and methods

2.1. Sample collection

Throat swabs were taken by family doctors in their office from individuals showing signs of SARS-CoV-2 infection. The swabs were stored in Viral Transport Media (VTM) and sent to diagnostic laboratories for SARS-CoV-2 analyses by RT-qPCR. In total, we sequenced 90 selected samples from individuals diagnosed for SARS-CoV-2 infection during the first major outbreak in Germany in February and March 2020. Samples were provided by the clinical laboratory MVZ Labor Mönchengladbach and Institute of Virology, University Hospital Bonn. From both laboratories we obtained the original swab in VTM. RNA was isolated using QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's protocol. Extracted RNA was then stored at −80 °C until further experiments.

2.2. Whole genome sequencing

Viral RNA was used to prepare cDNA, which was target-enriched using QIAseq SARS-CoV-2 Panel (Qiagen). Libraries were prepared using FX DNA Library Preparation Kit (Qiagen) according to the manufacturer's protocol. Briefly, cDNA was fragmented, adapters were ligated, and samples were purified and quantified. Quality control of all samples was assessed using the TapeStation 4200 (Agilent) and then the samples were sequenced using the Illumina MiSeq Next Generation Sequencing (NGS) platform.

2.3. Genome assembly

Raw sequencing data were trimmed using cutadapt v3.2 [12]. The resulting reads were aligned to the SARS-CoV-2 reference genome (GenBank ID: MN908947.3) using minimap2 v2.17 [13]. The depth of the coverage was assessed using samtools depth v1.12 [14]. Primer sequences (ARTIC protocol) were soft-clipped from the alignment using the trim function in iVar v1.3 [15]. Consensus genome assemblies were built using samtools mpileup and the consensus function in iVar with default settings. Finally, QUAST v5.0.2 [16] was applied to evaluate the quality of the consensus genome assemblies. Coverage and consensus genome quality were confirmed by FastQC v0.11.9 [17] and MultiQC v1.10.1 [18].

2.4. Data quality and availability

The quality of the SARS-CoV-2 reference-based genome assemblies was checked by assessing the fraction of the covered genome, number of misassembles, number of mismatches, and indels per 100 kbp. A total of 89 of 90 samples showed a genome coverage of >90% (97.7% Mean; ±4.0% SD; 90.0% IQR) with the median depth of coverage 2950.25-fold. The data produced in this study were deposited in the GISAID portal with the submission date of February 12, 2021, and the location Heinsberg.

2.5. Phylogenetic analysis

Multiple sequence alignment (MSA) was performed using MAFFT v7.475 [19]. To ensure high-quality, known sequencing errors were masked using a custom python script [20]. Before the downstream analyses, sequences were kept if they were longer than 28.000 bp and had less than 0.05% missing bases. Columns that contained more than 50% gaps were also removed. After stringent quality control, the maximum likelihood (ML) based phylogenetic tree reconstruction was performed using FastTree v2.1.10 [21]. Pangolin v2.4.2 [11] was applied to determine the most likely SARS-CoV-2 lineage. The phylogenetic tree was visualized using FigTree v1.4.4 [22] and annotated with Pango lineage. In addition, we performed a Nextstrain [23] phylogenetic analysis [24] and lineage annotation for integrative analysis purposes. For that, we extracted SARS-CoV-2 sequence data and metadata from European samples from the GISAID database [25] from December 5, 2019 to April 4, 2020. Using default parameters for subsampling and analysis, we ran the Nextstrain workflow by setting the geographic areas to Europe, Germany, and North-Rhine Westphalia (NRW). respectively. The resulting JSON files were visualized using the web-based application Auspice [23].

2.6. Variant calling, annotation and clustering

Variant calling was performed using the BAM file created in the aforementioned consensus genome assembly step using ivar. Minimum alignment quality and depth were set to 20 and 10 for an alternative allele to be called. Gene-based annotation of the variants to identify any consequence on the protein-coding level was assessed by Annovar [26]. SNP-based identity-by-state (IBS) clustering was performed using a hierarchical clustering approach from the R-package SNPRelate [27].

2.7. Ethics approval

The study was approved by the Ethics Committee of the Medical Faculty of the University of Bonn (approval number 085/20) and has been registered at the German Clinical Trials Register (https://www.drks.de, identification number DRKS00021306, study arm 1).

3. Results and discussion

We sequenced the viral genomes of 90 (23%) of the 388 SARS-CoV-2 cases that were reported in the Heinsberg district in February and March 2020. After quality control, we retained 89 samples for phylogenetic analysis. Phylogenetic tree annotated by Pangolin annotation system revealed that the samples clustered into groups 1 and 2 (Fig. 1A). The majority of samples belonging to group 1 were assigned to pangolin lineage B.3 (74 samples), whereas, in group 2, the majority of samples were assigned to lineage B.1 (6 samples). Samples belonging to lineage B.3 were collected early in the outbreak (before March 13, 2020, Fig. 1B), indicating that this lineage caused the initial outbreak. Lineage B.1 was introduced at a later time point (after March 13, 2020, Fig. 1B). Interestingly, as the pandemic was progressing, B.1 lineage became the predominant strain worldwide [28]. Similar observation came from the first

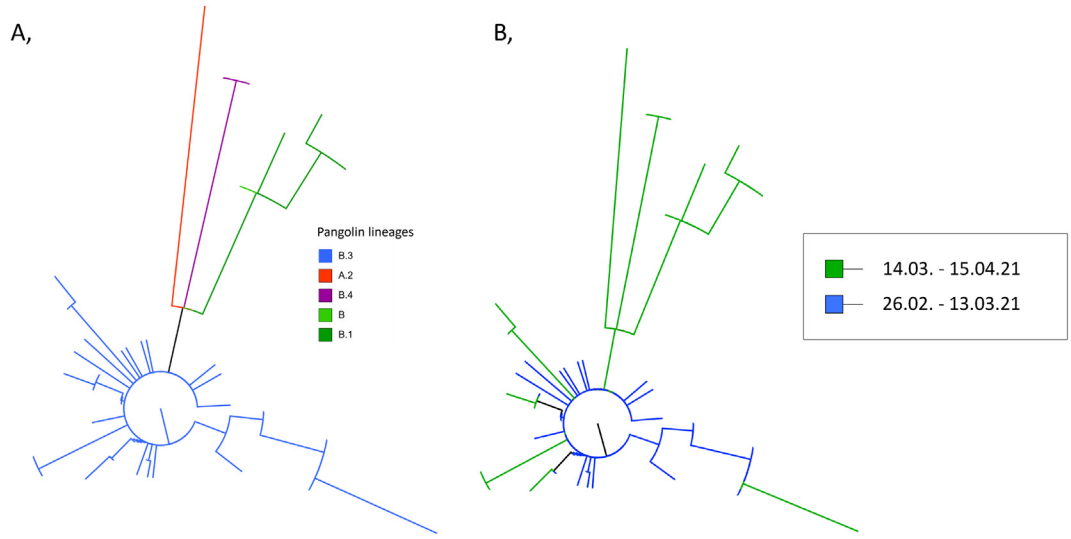

Fig. 1. Clustering and phylogenetic tree reconstruction. A) Phylogenetic tree of the study samples generated using FastTree, branches were colored by Pangolin lineage assignment. B) Same as A) but branches were colored by swab collection date.

and most affected region in Italy, where 344 out of 346 SARS-CoV-2 genomes where interspersed within B-sub lineages. Lineage B.1 was identified here in the second half of February 2020. Later it was identified in the Netherlands, the UK, and Central Europe, supporting our data [29].

To further investigate the origin of the virus on the state and national level, we performed a phylogenetic analysis using Nextstrain. We incorporated SARS-CoV-2 samples from North-Rhine-Westphalia (NRW; state where the outbreak took place) and in other regions in Germany that were uploaded to GISAID and collected between December 5, 2019 to April 4, 2020. Subsampling was performed based on samples collected from NRW and Germany (Fig. 2A and B). Our analysis confirmed that lineage B.3 was indeed the most prevalent strain in the beginning of the outbreak in the region. The analysis also revealed that B.3 and B.1 were competing strains around that time point. Thus, we hypothesize that the outbreak was not caused by an introduction of a single virus strain, but rather a series of at least two individual events which introduced different viral strains into this region and this fueled the spreading of the virus. Outbreaks involving multiple variants have been observed before. SARS-CoV-2 genomic diversity study from Brazil showed that lineage B.1 was the most prevalent one at the time point when it started to gain significance also in Europe. They also concluded, that a local transmission can be caused by multiple strains [30]. Another outbreak at a university in the USA from March 2021 – May 2021 was caused by multiple strains simultaneously, which was confirmed by the positive travel history of the infected individuals [31]. Another outbreak with multiple variants was linked to a single flight from New Delhi to Hong Kong in April 2021, in which 59 people were infected and the sequencing analyses revealed at least 3 sub-lineages [32]. Similarly to these, we identified two dominant lineages and we assume that their introduction did not occur simultaneously but rather distinctly in a timely manner.

We next used the same approach as above to identify the closest ancestor of the strain, which caused the outbreak. Performing a phylogenetic analysis, we used SARS-CoV-2 samples from Europe that were collected between December 5, 2019 to April 4, 2020. We observed that the B.3 samples cluster in one branch of the European SARS-CoV-2 phylogenetic tree. Additionally, the parent

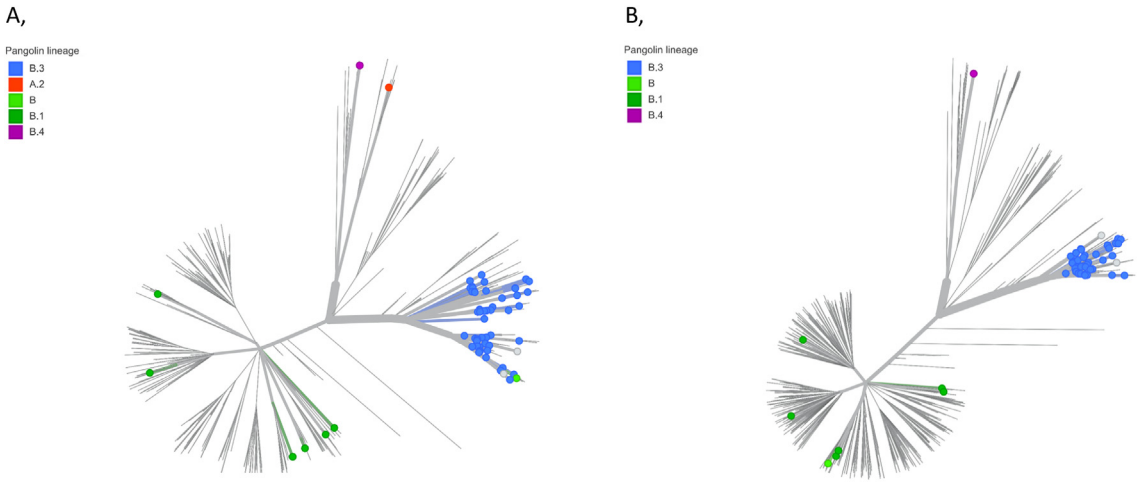

Fig. 2. Phylogenetic tree reconstruction using Nextstrain and the GISAID database- NRW and Germany. A) Nextstrain-based phylogenetic tree analysis using a subsampling schema based on state NRW level from December 2020 to March 2021. B) Nextstrain-based phylogenetic tree analysis using a subsampling schema based on national level from December 2020 to March 2021.

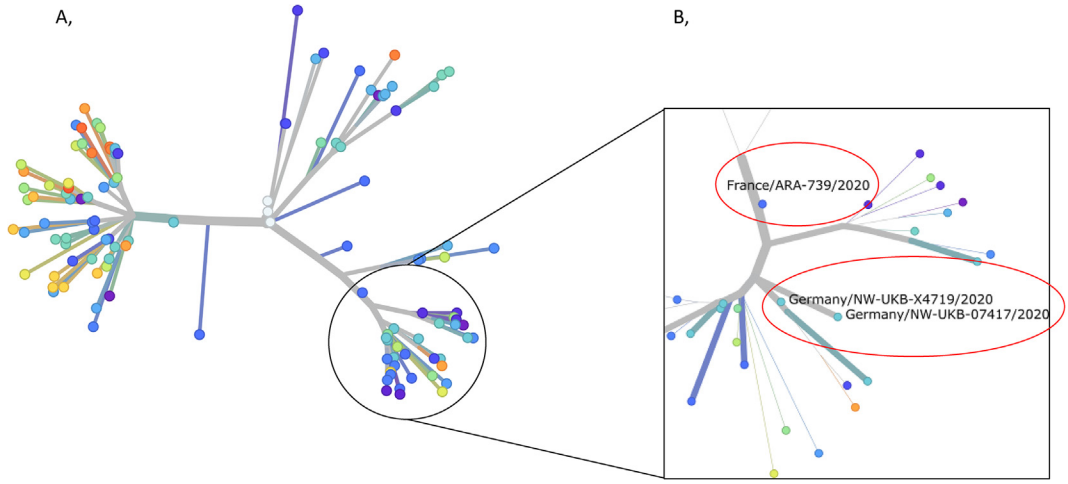

Fig. 3. Phylogenetic tree reconstruction using Nextstrain and the GISAID database. A) Nextstrain-based phylogenetic tree analysis using a subsampling schema based on European country level from January to March 2020. The node colors indicate the exposed countries and each dot represents a genome from the GISAID database. B) Zoom into our cohort revealed that the internal nodes prior to the cohort was assigned to France. The red circles indicate the representative genomes from our cohort and a closely related strain from France. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

branch is assigned to France. The European level analysis revealed a closely related strain located in France (Fig. 3). Taking into consideration that the first reported cases of SARS-CoV-2 in France were in the east of the country, a region neighbor to Germany [33], it is possible that the virus was introduced from there. However, the lack of information on the travel history and the subsampling approach applied in the Nextstrain workflow limits the analysis. Moreover, the majority of the early samples before mid-February 2020 were collected in France. This may bias the phylogenetic analysis on the European level, but it is consistent with the sample collection dates.

Lastly, we characterized and assessed the genetic differences between the two lineages which were associated with the outbreak. We were able to identify a prominent missense mutation in the spike protein D614G in the B.1 lineage. Overall, 10% of the SARS-CoV-2 isolates carried exclusively this mutation in the spike protein which differentiates B.1 from B.3. A SARS-CoV-2 variant carrying the spike protein amino acid change D614G has later become the most prevalent form in Europe and it was identified in early March 2020 [34]. Although we observed that the B.3 lineage has a higher representation in our cohort, B.1 lineage could be the predominant one which spread from this area to the rest of the country. As described by Korber et al., and also seen from our data, at that time point (March 2020) the B.1 lineage carrying D614G mutation was rare globally but gaining prominence in Europe. A similar observation was made in Basel, Switzerland where they also experienced a massive-spreading event with dominating B.1 lineage [35].

A recent study has shown that the first major outbreak in Germany, which we are describing in this study, started shortly after carnival festivities [36]. A study from Netherlands compared the number of new COVID-19 cases in regions that celebrate carnival and those which do not. They found that the number of new SARS-CoV-2 infections exceeded those in the non-carnival region about 1 week after the first case was reported [37].

4. Conclusion

In summary, we identified the B.3 lineage probably causing the first major outbreak in Germany, with the B.1 lineage probably being introduced at a later time point. We identified a closely related strain of the circulating B.3 lineage, as a strain located to France. The strain introduced at a later time point (B.1) in the course of the outbreak has become the dominant one in Germany, but also in the rest of Europe. The virus may adjust to infection and replication in humans, therefore the constant monitoring of all SARS-CoV-2 lineages, strains, and variants that are present in the population worldwide is very important to quickly and efficiently determine the ongoing virus evolution. This study demonstrates the power of sequence analysis of SARS-CoV-2 to reconstruct viral spread.

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] Zhou P et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3.
[2] Grasselli G et al. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region. *Italy* JAMA 2020;323(16):1574–81.
[3] Farooq HZ et al. Real-world SARS CoV-2 testing in Northern England during the first wave of the COVID-19 pandemic. J Infect 2021;83(1):84–91.
[4] Chang YS et al. Transmission Dynamics of Large Coronavirus Disease Outbreak in Homeless Shelter, Chicago, Illinois, USA, 2020. Emerg Infect Dis 2022;28 (1):76–84.
[5] Yu JB et al. Epidemiological characteristics of imported COVID-19 cases in Tianjin. Zhonghua Liu Xing Bing Xue Za Zhi 2021;42(12):2082–7.
[6] Luo Z et al. Epidemiological Characteristics of Infectious Diseases Among Travelers Between China and Foreign Countries Before and During the Early Stage of the COVID-19 Pandemic. Front Public Health 2021;9:739828.
[7] Arora M et al. Airport Pandemic Response: An Assessment of Impacts and Strategies after One Year with COVID-19. Transp Res Interdiscip Perspect 2021:100449.
[8] Streeck H et al. Infection fatality rate of SARS-CoV2 in a super-spreading event in Germany. Nat Commun 2020;11(1):5829.
[9] Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol 2018;16 (8):e3000003.
[10] Callaway E. The coronavirus is mutating – Does it matter? Nature 2020;585 (7824):174–7.
[11] Rambaut A et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5(11):1403–7.
[12] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011. **17**(1). 3.
[13] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–100.
[14] Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[15] Grubaugh ND et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol 2019;20(1):8.

[16] Gurevich A et al. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29(8):1072–5.
[17] Andrews, S. *FASTQC. A quality control tool for high throughput sequence data*. 2010 [cited 2021; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[18] Ewels P et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32(19):3047–8.
[19] Katoh K et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30(14):3059–66.
[20] Nicola De Maio, C.W., Rui Borges, Lukas Weilguny, Greg Slodkowicz, Nick Goldman. *Issues with SARS-CoV-2 sequencing data*. 2020 [cited 2021; Available from: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
[21] Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 2010;5(3):e9490.
[22] Rambaut, A. *FigTree*. 2007. Available from: http://tree.bio.ed.ac.uk/software/figtree/.
[23] Hadfield J et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121–3.
[24] Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol 2018;4(1):vex042.
[25] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – From vision to reality. Euro Surveill 2017;22(13).
[26] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164.
[27] Zheng X et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 2012;28(24):3326–8.
[28] Cella E et al. SARS-CoV-2 Lineages and Sub-Lineages Circulating Worldwide: A Dynamic Overview. Chemotherapy 2021;66(1–2):3–7.
[29] Alteri C et al. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. Nat Commun 2021;12(1):434.
[30] Varela APM et al. SARS-CoV-2 introduction and lineage dynamics across three epidemic peaks in Southern Brazil: massive spread of P.1. Infect Genet Evol 2021;96:105144.
[31] Doyle K et al. Multiple Variants of SARS-CoV-2 in a University Outbreak After Spring Break – Chicago, Illinois, March-May 2021. MMWR Morb Mortal Wkly Rep 2021;70(35):1195–200.
[32] Dhanasekaran, V., et al. *Air travel-related outbreak of multiple SARS-CoV-2 variants*. medRxiv. 2021. 2021.07.22.21260854.
[33] Gerbaud L et al. Hospital and Population-Based Evidence for COVID-19 Early Circulation in the East of France. Int J Environ Res Public Health 2020;17(19):7175.
[34] Korber B et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 2020;182(4):812–827 e19.
[35] Stange M et al. SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to a mass gathering event. PLoS Pathog 2021;17(3):e1009374.
[36] Wessendorf L et al. Dynamics, outcomes and prerequisites of the first SARS-CoV-2 superspreading event in Germany in February 2020: a cross-sectional epidemiological study. BMJ Open 2022;12(4):e059809.
[37] Group L-C-R et al. Why crowding matters in the time of COVID-19 pandemic? – A lesson from the carnival effect on the 2017/2018 influenza epidemic in the Netherlands. BMC Public Health 2020;20(1):1516.