Detecting SARS-CoV-2 lineages and mutational load in municipal wastewater; a use-case in the metropolitan area of Thessaloniki, Greece

Nikolaos Pechlivanis[1,2], Maria Tsagiopoulou[1], Maria Christina Maniou[1], Anastasis Tougkousidis[1], Stamatia Laidou[1,2], Elisavet Vlachonikola[1,2], Evangelia Mouchtaropoulou[1], Taxiarchis Chassalevris[3], Serafeim Chaintoutis[3], Chrysostomos Dovas[3], Maria Petala[4], Margaritis Kostoglou[5], Thodoris Karapantsios[5], Anastasia Chatzidimitriou[1], Agis Papadopoulos[7], Nikolaos Papaioannou[3], Anagnostis Argiriou[1,6], Fotis Psomopoulos[1]

[1] Institute of Applied Biosciences, Centre of Research and Technology Hellas, Thermi, 57001 Thessaloniki, Greece
[2] Dept of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[3] School of Veterinary Medicine, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[4] Dept. of Civil Engineering, Aristotle University of Thessaloniki, Thessaloniki, 54 124, Greece
[5] Dept. of Chemistry, Aristotle University of Thessaloniki, 54 124 Thessaloniki, 54124, Greece
[6] Department of Food Science and Nutrition, University of the Aegean, 81400 Myrina, Lemnos, Greece
[7] EYATH S.A., Thessaloniki Water Supply and Sewerage Company S.A., Thessaloniki, 54636, Greece

Abstract

The SARS-CoV-2 pandemic represents an unprecedented global crisis necessitating novel approaches for, amongst others, early detection of emerging variants relating to the evolution and spread of the virus. Recently, the detection of SARS-CoV-2 RNA in wastewater has emerged as a useful tool to monitor the prevalence of the virus in the community. Here, we propose a novel methodology, called ***lineagespot***, for the detection of SARS-CoV-2 lineages in wastewater samples using next-generation sequencing. Our proposed method was tested and evaluated using NGS data produced by the sequencing of three wastewater samples from the municipality of Thessaloniki, Greece, covering three distinct time periods. The results showed a clear identification of trends in the presence of SARS-CoV-2 mutations in sewage data, and allowed for a robust inference between the variants evident through our approach and the variants observed in patients from the same area time periods. Lineagespot is an open-source tool, implemented in R, and is freely available on GitHub.

Introduction

Nearly a year after the first report of SARS-CoV-2 in Wuhan, China, the virus has spread at an unprecedented pace causing a global pandemic. As the main transmission process of the SARS-CoV-2 virus is through droplets and the contact between people, several testing strategies identify whether a person is infected and, in cases of a positive sample, what is the underlying virus variant. However, these methods are not easily scalable, especially in large urban areas. Interestingly, the viral RNA can also be detected in wastewater, with SARS-CoV-2 RNA levels in wastewater correlating with the COVID-19 epidemiology[1–3]. Indeed, in the previous work of Petala et al[3]. normalized viral copy levels

in Thessaloniki sewage were in agreement with the epidemiological conditions in the city. Thessaloniki is the second largest city in Greece with around 1 million inhabitants. The city is a chief gateway for entrepreneurs, traders, university students and tourists visiting Greece and, as such, it was the place where the Greek patient "zero" appeared in March 2020. Thessaloniki is also the place where the so-called South African mutation appeared in Greece for the first time. In other words, identifying new variants in the city sewage is critical to understand further scattering to the rest of the country.

The presence of SARS CoV-2 RNA in sewage provides us with a unique opportunity, i.e. to identify the most prevalent virus lineages through the analysis of the traces evident in wastewater samples. So far, although there are few studies exploring the SARS-CoV-2 diversity in sewage, it still remains an open issue as there are no widely accepted methods that can sufficient address this. The most common used approaches involve the sequencing the wastewater samples, and the consequent application of low frequency variant analysis methods[4] or metagenomic approaches[5,6]. In either case, the interpretation of the results focuses on the detection of specific variants[4] or lineages[2] such as B.1.1.7 and 501.V2, prevalent clades[6] (19A, 20A and 20B) or new uncharacterized mutations[6].

In this work we propose a novel methodology called ***lineagespot***, implemented as a software tool, that can facilitate the detection of SARS-CoV-2 lineages in wastewater samples using next generation sequencing. The method is tested and validated across three municipal wastewater samples retrieved in Thessaloniki, Greece in three different time periods, and correlated with the lineages observed in patients from the same area time points. Based on a variation of the Illumina Arctic pipeline for the identification of mutations at low frequencies (< 0.01), and the lineage assignments defined by Pangolin, this method identifies all SARS-CoV-2 mutations present in the wastewater of Thessaloniki, and attempts to infer the potential distribution of the SARS-CoV-2 lineages. The methodology is proven to be effective in detecting the mutational load in the wastewater, with the inferred lineages being roughly aligned to the predominant lineages identified through targeted (and therefore biased) patient-derived genotypes.

Results

Comparison of variant calling methods

The proposed methodology was applied on a selected sewage sample (corresponding to the 05-11 February 2021 time period), and for which three different variant callers were assessed: 1) *freebayes*, 2) *mpileup* and 3) *GATK Mutect2* (cancer only mode). In terms of parameters, *freebayes* was used with a low variant frequency parameter of 0.01, *mpileup* reported every position (either reference, or variant), and *GATK Mutect2* was used with the default parameters.

An example of the output produced by the methodology, regardless of the variant calling method, is shown in **Table 1**. In this table the overlap between the pangolin's rules and the rules generated by the tool for the input dataset is captured for each lineage. In order to quantify the overlap, three basic metrics are produced; the overlap by considering pangolin's rules as a decision tree (*Tree Overlap*), the total overlap regardless of the rules order (*Total Overlap*), and the overlap for the rules that are satisfied only by the identified mutations (i.e. explicitly listed in the variants' file), and therefore

excluding all rules based on the unmutated reference (*Total Overlap Var*). In addition to the previous metrics, the related ratio values are also calculated (*Tree ratio*, *Total Ratio*, *Total Ratio Var*). Finally, information regarding the read depth for each position (reference and variant) is also provided in the output file.

Table 1: Each row in the table corresponds to a single lineage rule defined by pangolin. The columns correspond to the different metrics captured, in order to perform the systematic evaluation.

Lineage	Rules	Total	Tree Overlap	Total Overlap	Total Overlap Var	Tree Ratio	Total Ratio	Total Ratio Var	Tree Avg AD	Total Avg AD	Total Avg DP	Total Sum AD	Total Sum DP	Avg DP	Total Run Reads	Avg. AF
B.1.177	26800=='C',...	17	0	11	3	0	0.647	0.176	0	6.6	19	33	95	25.20	104214	0.347
B.1.177	26800=='C',...	11	0	7	2	0	0.636	0.181	0	7	23.66	21	71	25.20	104214	0.295
B.1.1.7	26800!='C',...	13	3	11	1	0.230	0.846	0.076	0	2	9	2	9	25.20	104214	0.222

Based on this detailed table, a second output is generated as a simplified summary. In this case, all rows for which the *Total Overlap Var* column was equal to 0 were removed, and therefore potential lineages that would be assigned based only on the unmutated reference (i.e. no actual mutations detected) are excluded from the analysis. The remaining rows were collapsed (**Table 2**) through a process in which the average values of the basic metrics were calculated for each lineage; i.e. the mean of the *Tree Ratio*, the *Total Ratio*, the *Total Ratio Var*, the *Tree Av AD*, the *Total Av AD*, and the *Avg AF* columns.

Table 2: Each row in the table corresponds to a unique lineage, after the merging process. The remaining metrics can be consequently used to assess the presence of the particular variant in the dataset.

Lineage	Mean Tree Ratio	Mean Total Ratio	Mean Total Ratio Var	Mean Total Av AD	Mean AF
B.1.1.7	0.230769231	0.846153846	0.07692307	2	0.222222222
B.1.177	0.002222635	0.610840685	0.05660585	9.76863806	0.488594472

Depending on the variant caller tool used (*freebayes*, *mpileup* and *GATK Mutect2)*, ***lineagespot*** generates a unique output. All outputs are compared pairwise, based on the decision tree rules (n_d), and the total number of rules satisfied (n_t). For each lineage, the absolute values of the differences between the two metrics (n_d, n_t) of the files are calculated. As an example, for lineage *A.1*, the output produced by using the *freebayes* variant caller tool in the first step of the methodology, returns $n_d = 2$ and $n_t = 10$ rules satisfied, while the output of the *GATK Mutect2* tool returns $n_d = 1$ and $n_t = 4$ rules satisfied. As a result, the two outputs exhibit a difference of $n_d = 1$ and $n_t = 6$ rules (**Table 3**).

In addition, the total number of lineages are shown in **Table 4**. In the same table, the maximum absolute n_d difference and the maximum n_t difference for each pair of files are calculated. The latter is used for an overall comparison of the output files.

Table 3: Snapshot of the difference between the two metrics (n_d, n_t) between the output files coming from *freebayes* variant caller and *mpileup*.

Lineage	Rules	N_t	$n_d^{freebayes}$	$n_t^{freebayes}$	$n_d^{mpileup}$	$n_t^{mpileup}$	$\lvert n_d^{freebayes} - n_d^{mpileup} \rvert$	$\lvert n_t^{freebayes} - n_t^{mpileup} \rvert$
B.1.1.7	26800!='C', …	13	3	72	3	71	0	1
B.1.177	26800=='C', …	66	3	72	3	69	0	3
B.1.177	26800=='C', …	6	3	72	3	69	0	3
B.1.351	26800!='C', …	38	3	162	3	160	0	2

Table 4: Summary table of the three output files produced by *freebayes*, *mpileup* and *gatk mutect2* variant caller. The three files are compared in pairs.

Files	Number of differences	max absolute n_d difference	max absolute n_t difference
freebayes - gatk	3791	4	31
freebayes - mpileup	3140	4	33
gatk - mpileup	1571	1	31

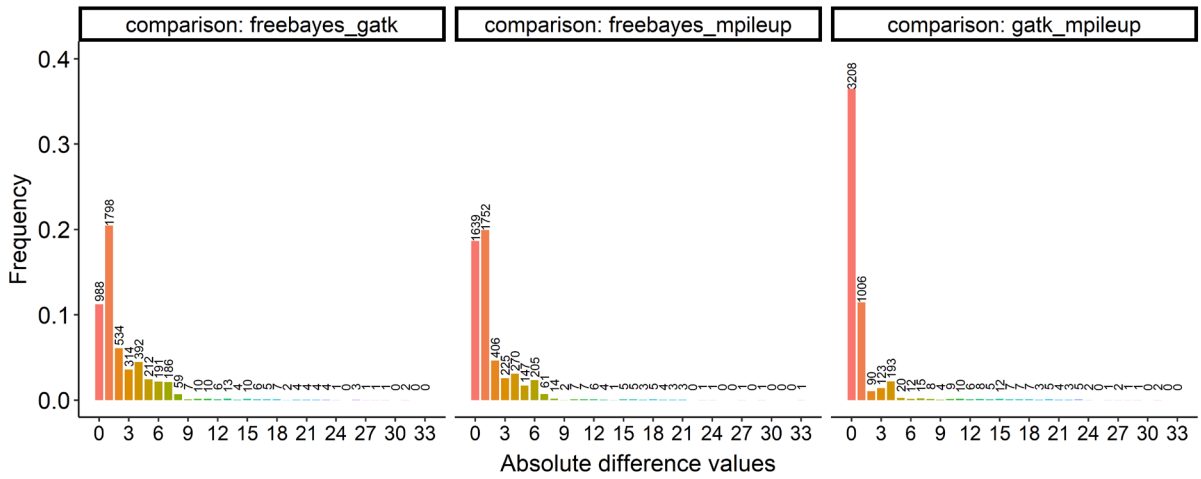


Figure 1: Distribution of the absolute n_t difference values between the output of the three variant calling tools use (pairwise comparisons).

Based on the above comparison, we consider that the most productive and informative approach is to utilize *freebayes* as the variant calling tool. The rest of the results shown below, are based only on the *freebayes* tool output.

Evaluating lineage-specific mutations across time periods

The proposed methodology was applied on three sewage samples, across three time periods: 02-14/12/2020 (*SampleA*), 05-11/02/2021 (*SampleB*) and 12-18/02/2021 (*SampleC*).

Table 5: The number of reads of each sample across the different stages of the applied bioinformatics analysis.

	SampleA	SampleB	SampleC
Total number of reads in raw fastq files	176380	466734	1567696
Total number of reads after adap-tor trimming	171312	453278	1565826
Total number of mapped reads	77444	133073	836105
Total number of sorted mapped reads	77105	131744	834085
Total number of input reads to ivar	75396	128013	811508
Total number of ivar shortened reads	4013	15853	37245
Reads outside primer region	1194	1937	12178
Total number of reads after primer trimming	71898	113954	784662
Sorted reads after primer trimming	69706	104214	769744
Mapped sorted reads after primer trimming (duplicates removal)	69706	104214	768820

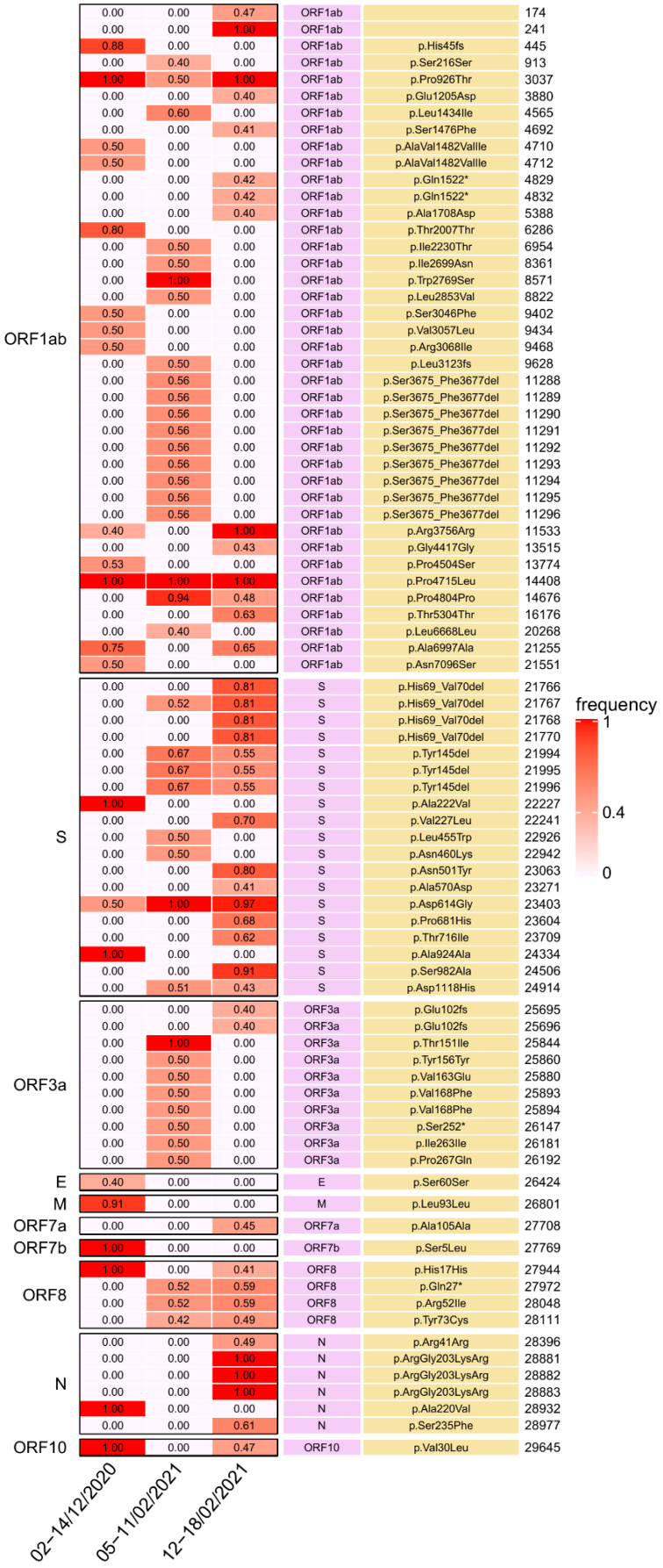


Figure 2: Mutation positions with allele frequency (AF) > 0.4.

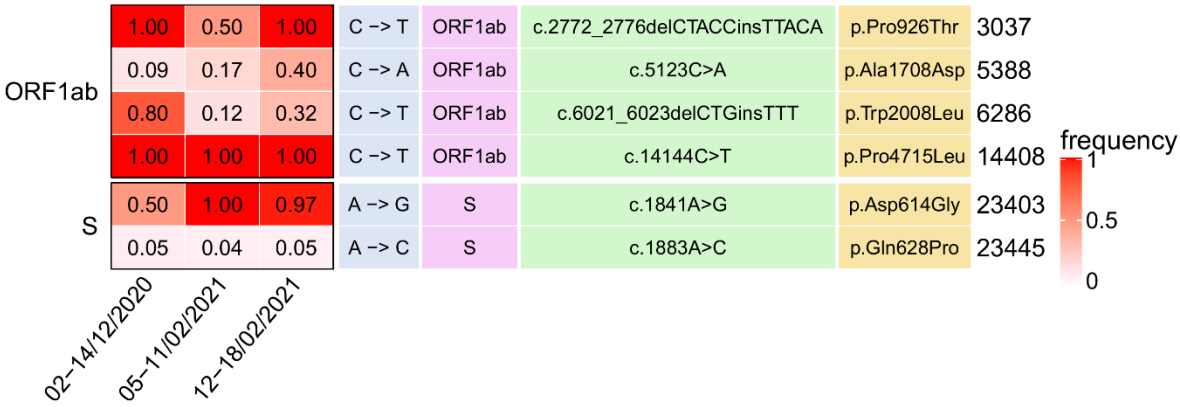


Figure 3: Common mutation positions across all time points.

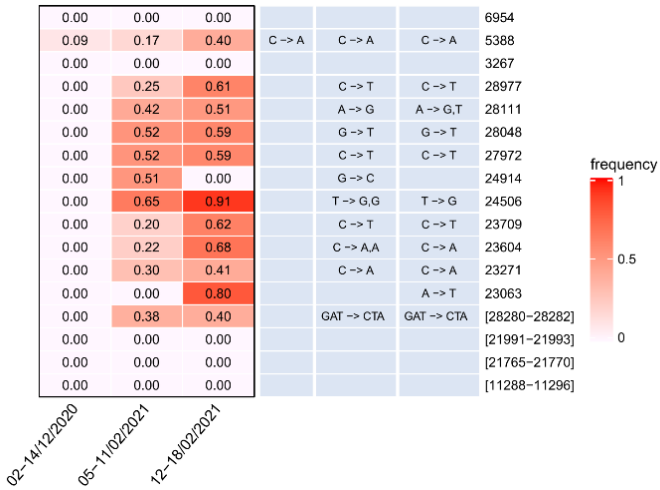


Figure 4: Detected B.1.1.7-detected mutations (UK Lineage)

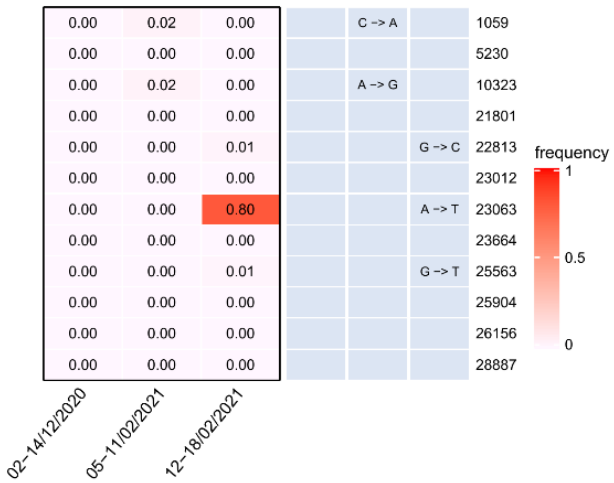


Figure 5: Detected B.1.351-detected mutations (South Africa lineage)

Assessing lineage assignment

Qualitative assessment between the major lineages found in the biased sampling and the three time periods.

Period 1: 02-14/12/2020
In this time period, the most frequent lineage detected in targeted sampling was B.1.177 (on average, 80% of the samples captured). The application of our method to the sewage sample of the same period, detected the same lineage (B.1.177) with the following characteristics:

Lineage	Mean Tree Ratio	Mean Total Ratio	Mean Total Ratio Var	Mean Total Av AD	Mean AF
B.1.177	0.002246535	0.605276332	0.042082807	5.626164875	0.513229318

Period 2: 05-11/02/2021
In this time period, the two most frequent lineages detected in targeted sampling were B.1.177 (~56% of the samples captured) and B.1.1.7 (~35% of the samples captured). The application of our method to the sewage sample of the same period, detected the same lineages (B.1.177 and B.1.1.7) with the following characteristics:

Lineage	Mean Tree Ratio	Mean Total Ratio	Mean Total Ratio Var	Mean Total Av AD	Mean AF
B.1.1.7	0.230769231	0.846153846	0.076923077	2	0.222222222
B.1.177	0.002222635	0.610840685	0.05660585	9.76863806	0.488594472

Period 3: 12-18/02/2021
In this time period, the two most frequent lineages detected in targeted sampling were B.1.177 (~54% of the samples captured) and B.1.1.7 (~38% of the samples captured). The application of our method to the sewage sample of the same period, detected the same lineages (B.1.177 and B.1.1.7) with the following characteristics:

Lineage	Mean Tree Ratio	Mean Total Ratio	Mean Total Ratio Var	Mean Total Av AD	Mean AF
B.1.1.7	0.230769231	0.846153846	0.153846154	31.33333333	0.5875
B.1.177	0.006394285	0.627302601	0.029552992	23.74139477	0.500008793

Discussion

Analyzing wastewater — used water that goes through the drainage system to a treatment facility — is a way that researchers and surveillance systems can track pathogens that are excreted in urine or feces, such as SARS-CoV-2. Monitoring effluents could provide better estimates on coronavirus spread

than sampling and testing the population, because wastewater surveillance can account for those who have not been tested and have only mild or no symptoms. Moreover, an effective and reliable methodology able to detect virus load and SARS-CoV2 variants from municipal wastewater samples could drastically, or at least help, decrease the cost of virus variant detection in the general population based on whole genome sequencing.

In this manuscript we present and validate a methodology named ***lineagespot***, making use of Next Generation Sequencing data, able to detect lineages and mutational load of SARS-CoV2. The methodology aims to aid the epidemiological system for the monitoring of COVID 19 pandemic in urban areas.

The method has been tested in different time point samples taken from the main Municipal Wastewater Treatment Plant of Thessaloniki - Greece, where effluents from approx. 750.000 inhabitants are collected. The ***lineagespot*** method demonstrated to be sensitive enough to identify and quantify differences in the mutational load, across various time points. Moreover, the quantitative data obtained using ***lineagespot*** are in accordance with the trends of well-known mutations (such as Asp614Gly) in the same period with the overall epidemiological status of the municipal area. The application of ***lineagespot*** in such complex samples, like those from Wastewater Treatment Plants, was able to assign lineages and in agreement with the trend of the major lineages detected in the area of Thessaloniki, in the three time points by whole genome sequencing of samples from the general population.

Overall, the method compared to other ones (Sanger sequencing)[4,7] resulted more informative, sensitive enough to detect mutations with low frequency and able to assign with good approximation the correct lineage present in the municipality.

Methods

Sampling and isolation

Wastewater samples were collected from the entrance of the main Municipal Wastewater Treatment Plant of the city which accommodates sewerage of about 750.000 inhabitants. Wastewater entering this Plant refers exclusively to citizens from urban districts of the city. Typical values of certain physicochemical properties of wastewater samples tested in this study are displayed in **Table 1.** These properties demonstrate, among others, the existence of suspended solids, dissolved organic matter, dissolved oxygen and salinity that may have strong impact on viral adsorption and decay because of oxidation and increased metabolic activity of microorganisms in sewage. The residence time of sewage until the entrance of the Plant is between 2 and 7 hours (depending on the area), which is more than enough to allow viral adsorption and decay. Identification of mutations may be hindered by viral adsorption and decay and for this reason the present effort is particularly significant.

Table 1. Main quality characteristics of wastewater samples

Parameter/ Sample period	02-14/12/2020	05-11/02/2021	12-18/02/2021
pH	7.5 ± 0.15	7.8 ± 0.10	7.8 ± 0.15
Electrical Conductivity (S/cm)	8.5 ± 1.5	9.6 ± 2.7	4.6 ± 1.0
Total Suspended Solids (mg/L)	620 ± 80	930 ± 115	1200 ± 125
BOD_5 (mg/L)	385 ± 75	525 ± 40	650 ± 65
COD (mg/L)	960 ± 120	1250 ± 150	1570 ± 170
Dissolved Organic Carbon (mg/L)	35.0 ± 6.5	49.0 ± 10.5	44.0 ± 5.0
UV absorption at 254 nm (1/cm)	0.35 ± 0.05	0.40 ± 0.06	0.45 ± 0.05
Total Nitrogen (mg/L)	62.0 ± 15.0	76.0 ± 11.5	95.0 ± 12.0
Ammonium Nitrogen (mg/L)	28.5 ± 8.0	33.0 ± 3.0	38.0 ± 4.0
Total Phosphorus (mg/L)	11.5 ± 4.5	11.5 ± 1.5	12.0 ± 2.0

Sampling and handling of the sewage samples were performed according to Petala et al[3]. Briefly, samples were obtained using a refrigerated autosampler (6712 Teledyne ISCO) programmed to deliver a 24-hours composite sample by mixing consecutive half-hour samples. Samples were transported to the lab on ice and were processed immediately. Three 50-mL aliquots of each untreated wastewater sample were subjected to centrifugation at 4000xg for 30 min. Afterwards, a composite sample was obtained from supernatants and pH was adjusted to 4 using 2 M HCl solution. Then, three aliquots of 40 mL each, were filtered through 0.45-µm pore-size, 47-mm diameter electronegative membranes (HAWP04700; Merck Millipore, Ireland) followed by RNA extraction as described in Ahmed et al[8].

RNA extraction and SARS-CoV2 detection

Extracted RNAs originating from 12 processed electronegative membranes and spanning different days of sampling were filtered through OneStep PCR inhibitor removal kit (Zymoresearch) according the manufacturer' s instructions, pooled and concentrated using the NucleoSpin® RNA XS, Micro kit (MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany). Concentrated RNAs were subjected to real-time RT-PCR testing for SARS-CoV-2 quantification, utilizing the N2 protocol proposed by the Centers for Disease Control and Prevention (CDC) for the diagnosis of COVID-19 in humans (CDC, 2020). The assay was performed on a CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad Laboratories, Hercules, CA, USA). A calibration curve was generated and SARS-CoV-2 viral loads were expressed as genome copies per µl of RNA extract. In total, three concentrated RNA pools from respective time periods were quantified; 02-14/12/2020 (36 copies/µl), 05-11/02/2021 (68 copies/µl), and 12-18/02/2021 (53 copies/µl).

Library preparation and sequencing

The targeted sequencing method was applied by preparing 400nt amplicons using the ARTIC v3 protocol developed by Wellcome Sanger Institute[9], with some modifications. First, cDNA synthesis was prepared from 10 ul of RNA using SuperScript II reverse transcriptase (Invitrogen - Thermo Fisher Scientific, USA) and 50 ng/ul of random primers according to the protocol guidelines. For subsequent cDNA amplification, 2.5 ul of the generated cDNA was used instead of 6 ul, using ARTIC PCR primer pools (v3). Finally, the NEBNext adaptor (New England Biolabs, US, #7335) was used in the ligation reaction, diluted with adaptor dilution buffer at 10µM final concentration. All purification steps were

performed according to the ARTIC protocol. The samples were paired-end sequenced on a MiSeq platform (Illumina, USA) with a read length of 2 × 300 bp.

Raw data analysis

The initial phase of the bioinformatics analysis is to produce an alignment of the sequencing reads, while maintaining extremely strict criteria, in order to remove any potential contaminants and/or sequencing errors. The first step is the adaptor removal process, where any adaptors are removed from the raw *fastq* sequences, with the cleaned reads mapped to the SARS-CoV-2 reference genome (Wuhan variant, NC_045512), using minimap2 tool[10] with a minimal chaining score (matching bases minus log gap penalty) equal to 40. From this process, only the paired-end sequences are retained, while any other (unmatched, multiple mappings, etc.) are removed. In the next step, the primer sequences are excluded using the *ivar* tool[11], setting a minimum of 200 length in nucleotides for a read to be retained after trimming, and a minimum threshold for sliding window of 15 quality to pass (width of sliding window equal to 4). The final sequences are then remapped to the same reference genome (minimal chaining score equal to 40). Finally, and in order to be able to detect low frequency variants, the *freebayes* variant caller was used with a low variant frequency parameter of 0.01. Ultimately, all identified mutations were annotated using the *SnpEff* tool[12] and the NC_045512.2 (version 5.0) database.

Downstream analysis of lineages detection

In order to identify and assign different SARS-CoV-2 lineages based on the mutations detected from a single sewage sample, we implemented the proposed methodology in a tool named ***lineagespot***. The tool accepts as input a VCF file, which contains all mutations identified in the sample, along with the reference SARS-CoV-2 genome file, and a file containing all lineage-assignment rules, as retrieved from the pangolin tool[13] repository. After analyzing all inputs, a tab-delimited file (TSV file) is produced containing the most probable lineages that have been found. Figure 1 shows an overview of the tool's functionalities, which can be described in 2 phases:

i) Creating rules from variants

In this phase all rules that can be derived from the VCF file are constructed. Initially, a vector of all genome positions is created, for which each position is set to be equal to the reference genome nucleotides. Then, the VCF file is read and a set of new rules is formed by setting each position of the file with the reported variant or multiple variants (in case there is more than one reported variant at the same position). It should be noted that positions that have been detected with more than one variant, should include all of them at the VCF's ALT column in a comma-delimited format. Most of the variant caller tools (*freebayes*, *GATK*, etc.) are doing this by default. Finally, positions with reference read depth equal to zero are removed from the first vector. The remaining two vectors are merged into one.

In addition to finding all positions that need to be set equal to the base that has been allocated, four more rules are added for each genome position. These rules contain all bases *not* equal to the nucleotide of the original rule. For example, if position 5388 is equal to base 'A' (representing rule *5388=='A'*), then four new rules are added containing all bases not equal to 'A', e.g., *5888!='T',*

5388!='G', 5388!='C', 5388!='-' (where the '-' symbol stands for a gap in the referred sequence). Finally, all rules are merged into a single vector representing this particular lineage.

A.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
NC_045512.2	326	.	T	A	0.0001	.	DP=6;I16=6,0,0,0,187,5875,0,0,360,21600,0,0,19,73,0,0;QS=1,0;MQ0F=0
NC_045512.2	378	.	T	C	1.3e-06	.	DP=7;I16=6,0,1,0,197,6591,18,324,360,21600,60,3600,20,92,5,25;QS=0.916279,0.0837209,0;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQ0F=0
NC_045512.2	408	.	A	T	2.5e-06	.	DP=8;I16=8,0,0,0,268,9088,0,0,480,28800,0,0,32,174,0,0;QS=1,0;MQ0F=0
NC_045512.2	433	.	T	C	0.0048	.	DP=9;I16=9,0,0,0,296,10046,0,0,540,32400,0,0,40,246,0,0;QS=1,0;MQ0F=0

B.

Rules
"326=='T'", "326=='A'", "326!='G'", "326!='C'", "326!='-'"
"378=='T'", "378=='C'", "378!='G'", "378!='A'", "378!='-'"
"408=='A'", "408!='T'", "408!='G'", "408!='C'", "408!='-'"
"433=='T'", "433!='C'", "433!='A'", "433!='G'", "433!='-'"

C.

Lineage	Rules
B.1.177.17	28931!='A',25613!='G',407!='-',22087!='A'
B.1.177	28931!='A',25613!='G',407!='-'...
B.1	28931!='A',25613!='G',407!='-'...
B.1.177.13	28931!='A',25613!='G',407!='-'...

D.

Lineage	Rules	Total	Tree Overlap	Total Overlap	Total Overlap Var	Tree Ratio	Total Ratio	Total Ratio Var	Tree Avg AD	Total Avg AD	Tota Avg DP	Total Sum AD	Total Sum DP	Avg DP	Total Run Reads	Avg.AF
B.1.177	26800=='C',...	17	0	11	3	0	0.647	0.176	0	6.6	19	33	95	25.20	104214	0.347
B.1.177	26800=='C',...	11	0	7	2	0	0.636	0.181	0	7	23.66	21	71	25.20	104214	0.295
B.1.1.7	26800!='C',...	13	3	11	1	0.230	0.846	0.076	0	2	9	2	9	25.20	104214	0.222

Figure 6: Snapshots of the intermediate steps. **A.** A VCF file produced by the chosen variant caller. **B.** The rules as they are generated by the VCF file. **C.** Pangolin's decision rules. **D.** A tab-delimited file as produced by ***lineagespot***.

ii) Comparing with pangolin rules

The second phase aims to compare the rules derived from the VCF file with the assignment rules provided by the pangolin tool. Specifically, all decision rules are read from the input pangolin file, and for each lineage, the related rules are compared with the final rule vector.

Three metrics are computed and stored in the output file; the total number of rules leading to the related lineage (N_r), the number of rules satisfied by the created rule vector, considering pangolin's rules as a decision tree (n_d), and the total number of rules satisfied (n_t). Also, the related ratio values are being computed, giving a satisfaction percentage of each lineage:

$$R_d = \frac{n_d}{N_r}, R_t = \frac{n_t}{N_r}$$

Underlying assumptions of the method

It should be noted that the methodology relies on the following assumption. Given a group of reads that satisfy a rule A of lineage L, and another group of reads that satisfy rule B from the same lineage L, then the lineage L is incorrectly assigned. As an example, suppose that a group of reads satisfy only the first two rules from lineage *B.1.177.17* (*28931!='A', 25613!='G'*), and another group of reads satisfy the next two rules from the same lineage (*407!='-', 22087!='A'*). Based on the method description

above, lineage B.1.177.17 will be marked as an identified lineage, even though none of the reads satisfy all the rules of the lineage.

In order to mitigate this risk, we are taking into consideration a number of different indicators, that reflect the number of total rules satisfied, the number of rules that are satisfied based on the detected mutations, and the overall number of reads that support both reference and allele for each of the rules.

References

1. Nemudryi, A. *et al. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater*. http://medrxiv.org/lookup/doi/10.1101/2020.04.15.20066746 (2020) doi:10.1101/2020.04.15.20066746.
2. Jahn, K. *et al. Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples*. http://medrxiv.org/lookup/doi/10.1101/2021.01.08.21249379 (2021) doi:10.1101/2021.01.08.21249379.
3. Petala, M. *et al.* A physicochemical model for rationalizing SARS-CoV-2 concentration in sewage. Case study: The city of Thessaloniki in Greece. *Science of The Total Environment* **755**, 142855 (2021).
4. Martin, J. *et al.* Tracking SARS-CoV-2 in Sewage: Evidence of Changes in Virus Variant Predominance during COVID-19 Pandemic. *Viruses* **12**, 1144 (2020).
5. Crits-Christoph, A. *et al.* Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *mBio* **12**, e02703-20, /mbio/12/1/mBio.02703-20.atom (2021).
6. Izquierdo-Lara, R. *et al. Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing*. http://medrxiv.org/lookup/doi/10.1101/2020.09.21.20198838 (2020) doi:10.1101/2020.09.21.20198838.
7. Daughton, C. G. Wastewater surveillance for population-wide Covid-19: The present and future. *Science of The Total Environment* **736**, 139631 (2020).
8. Ahmed, W. *et al.* First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of The Total Environment* **728**, 138764 (2020).
9. Pipelines, D. *et al.* COVID-19 ARTIC v3 Illumina library construction and sequencing protocol v4 (protocols.io.bgxjjxkn). doi:10.17504/protocols.io.bgxjjxkn.
10. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
11. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 8 (2019).
12. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Gene.* **3**, (2012).
13. *pangolin; Phylogenetic Assignment of Named Global Outbreak LINeages, https://github.com/cov-lineages/pangolin*.

Acknowledgements / CRediT author statement

This work was supported by the "*Greece vs Corona: Flagship Action to address the SARS-CoV-2 crisis. Epidemiological study in Greece through extensive testing for virus and antibody detection, viral genome sequencing and genetic analysis of patients*" project, which is funded by the General

Secretariat for Research and Innovation, under the Public Investments Program (PIP). Additionally, this work was supported by *ELIXIR*, the research infrastructure for life-science data.

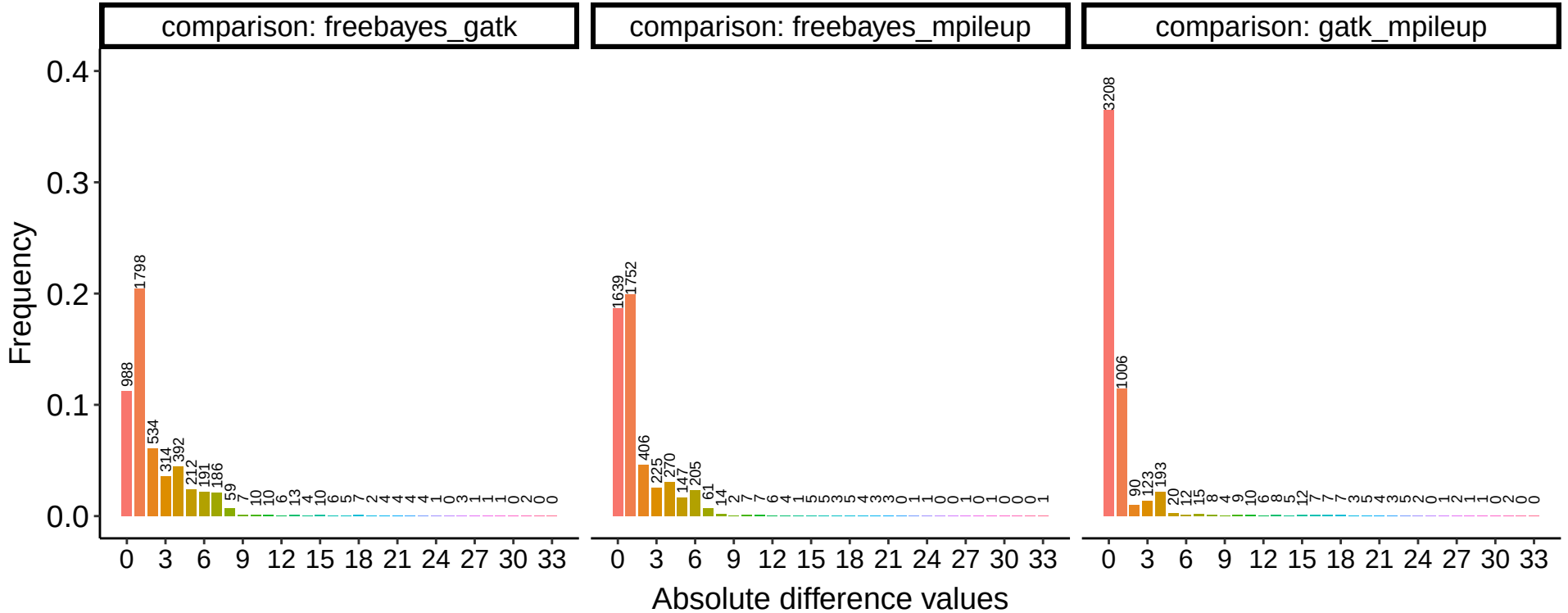
NP, MT, MCM and AT developed the ***lineagespot*** code and performed the downstream analysis. NP and MT performed the analysis of the raw NGS data. EM, SL and EV did the library preparation. CD and AC provided the data and reviewed the submitted version. MP and MK participated in experimental investigation. AP, NP, and TK participated in project conceptualization, management and funding. FP and AA designed, supervised the study, and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Supplementary Material

S1: Code and data

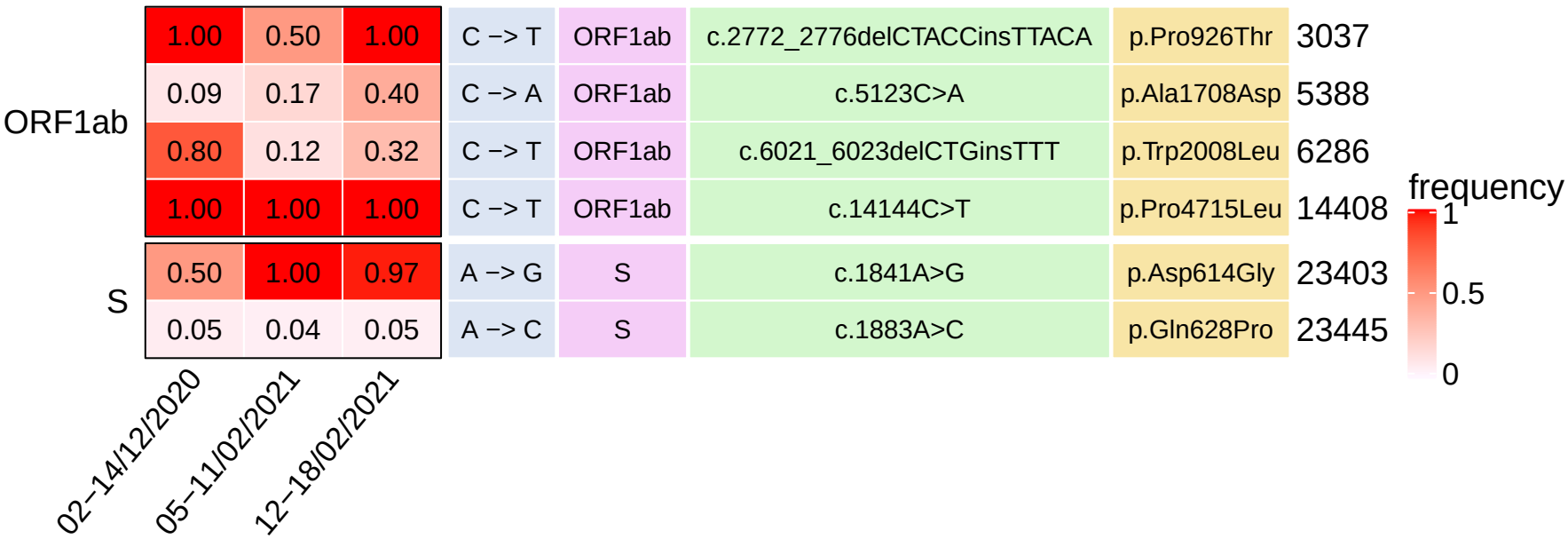
The implemented code that produces the results of this paper, starting from the VCF files, is available on the GitHub repository: https://github.com/BiodataAnalysisGroup/lineagespot.

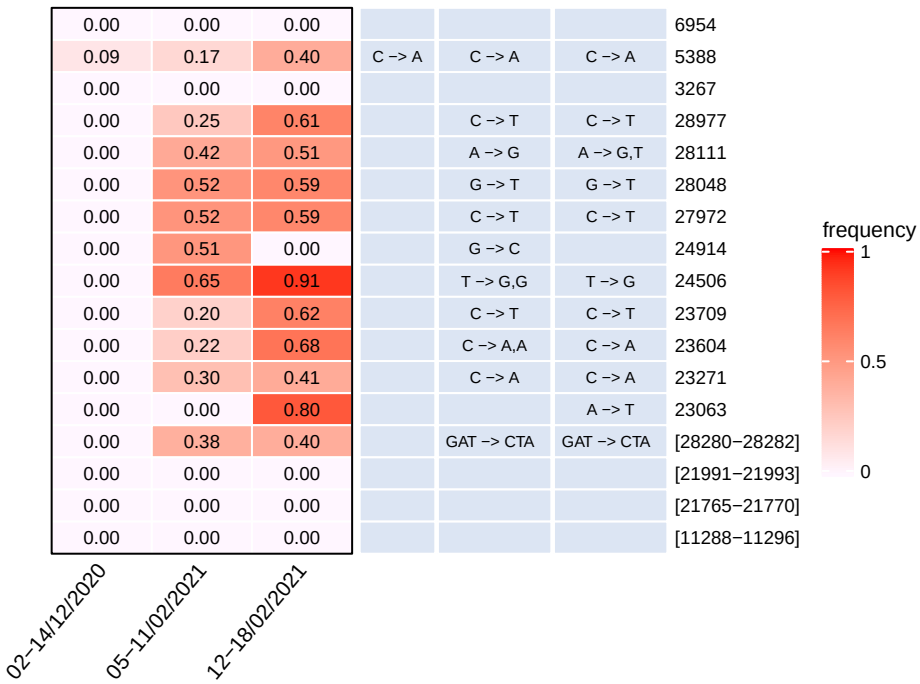
The raw FASTQ files are available through Zenodo: https://zenodo.org/record/4564182

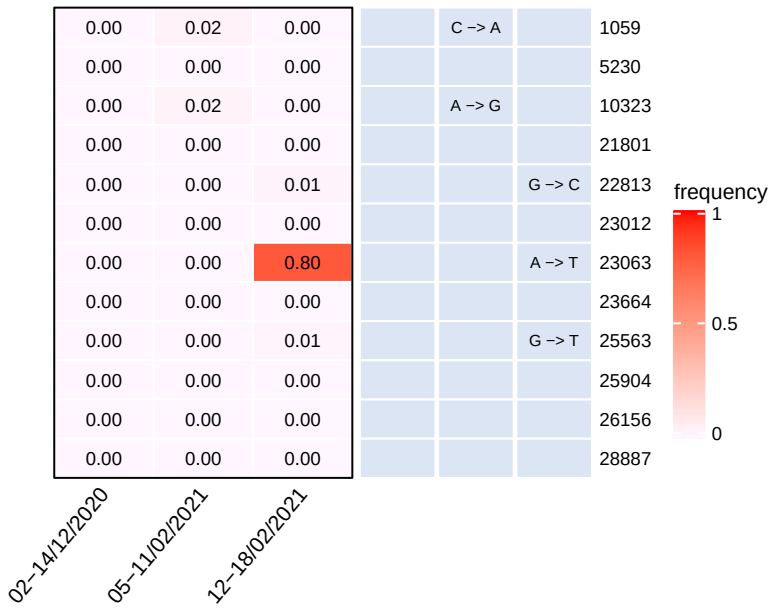

comparison: freebayes_gatk
comparison: freebayes_mpileup
comparison: gatk_mpileup
Frequency
0.4
0.3
0.2
0.1
0.0
988
1798
534
314
392
212
191
186
59
7
10
10
6
13
4
10
6
5
7
2
4
4
4
4
1
0
3
1
1
1
0
2
0
0
1639
1752
406
225
270
147
205
61
14
2
7
7
6
4
1
5
5
3
5
4
3
3
0
1
1
0
0
1
0
1
0
1
0
0
0
1
3208
1006
90
123
193
20
12
15
8
4
9
10
6
8
5
12
7
7
7
3
5
4
3
5
2
0
1
2
1
1
0
2
0
0
0 3 6 9 12 15 18 21 24 27 30 33
Absolute difference values

	02–14/12/2020	05–11/02/2021	12–18/02/2021			
ORF1ab	0.00	0.00	0.47	ORF1ab		174
	0.00	0.00	1.00	ORF1ab		241
	0.88	0.00	0.00	ORF1ab	p.His45fs	445
	0.00	0.40	0.00	ORF1ab	p.Ser216Ser	913
	1.00	0.50	1.00	ORF1ab	p.Pro926Thr	3037
	0.00	0.00	0.40	ORF1ab	p.Glu1205Asp	3880
	0.00	0.60	0.00	ORF1ab	p.Leu1434Ile	4565
	0.00	0.00	0.41	ORF1ab	p.Ser1476Phe	4692
	0.50	0.00	0.00	ORF1ab	p.AlaVal1482ValIle	4710
	0.50	0.00	0.00	ORF1ab	p.AlaVal1482ValIle	4712
	0.00	0.00	0.42	ORF1ab	p.Gln1522*	4829
	0.00	0.00	0.42	ORF1ab	p.Gln1522*	4832
	0.00	0.00	0.40	ORF1ab	p.Ala1708Asp	5388
	0.80	0.00	0.00	ORF1ab	p.Thr2007Thr	6286
	0.00	0.50	0.00	ORF1ab	p.Ile2230Thr	6954
	0.00	0.50	0.00	ORF1ab	p.Ile2699Asn	8361
	0.00	1.00	0.00	ORF1ab	p.Trp2769Ser	8571
	0.00	0.50	0.00	ORF1ab	p.Leu2853Val	8822
	0.50	0.00	0.00	ORF1ab	p.Ser3046Phe	9402
	0.50	0.00	0.00	ORF1ab	p.Val3057Leu	9434
	0.50	0.00	0.00	ORF1ab	p.Arg3068Ile	9468
	0.00	0.50	0.00	ORF1ab	p.Leu3123fs	9628
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11288
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11289
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11290
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11291
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11292
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11293
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11294
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11295
	0.00	0.56	0.00	ORF1ab	p.Ser3675_Phe3677del	11296
	0.40	0.00	1.00	ORF1ab	p.Arg3756Arg	11533
	0.00	0.00	0.43	ORF1ab	p.Gly4417Gly	13515
	0.53	0.00	0.00	ORF1ab	p.Pro4504Ser	13774
	1.00	1.00	1.00	ORF1ab	p.Pro4715Leu	14408
	0.00	0.94	0.48	ORF1ab	p.Pro4804Pro	14676
	0.00	0.00	0.63	ORF1ab	p.Thr5304Thr	16176
	0.00	0.40	0.00	ORF1ab	p.Leu6668Leu	20268
	0.75	0.00	0.65	ORF1ab	p.Ala6997Ala	21255
	0.50	0.00	0.00	ORF1ab	p.Asn7096Ser	21551
S	0.00	0.00	0.81	S	p.His69_Val70del	21766
	0.00	0.52	0.81	S	p.His69_Val70del	21767
	0.00	0.00	0.81	S	p.His69_Val70del	21768
	0.00	0.00	0.81	S	p.His69_Val70del	21770
	0.00	0.67	0.55	S	p.Tyr145del	21994
	0.00	0.67	0.55	S	p.Tyr145del	21995
	0.00	0.67	0.55	S	p.Tyr145del	21996
	1.00	0.00	0.00	S	p.Ala222Val	22227
	0.00	0.00	0.70	S	p.Val227Leu	22241
	0.00	0.50	0.00	S	p.Leu455Trp	22926
	0.00	0.50	0.00	S	p.Asn460Lys	22942
	0.00	0.00	0.80	S	p.Asn501Tyr	23063
	0.00	0.00	0.41	S	p.Ala570Asp	23271
	0.50	1.00	0.97	S	p.Asp614Gly	23403
	0.00	0.00	0.68	S	p.Pro681His	23604
	0.00	0.00	0.62	S	p.Thr716Ile	23709
	1.00	0.00	0.00	S	p.Ala924Ala	24334
	0.00	0.00	0.91	S	p.Ser982Ala	24506
	0.00	0.51	0.43	S	p.Asp1118His	24914
ORF3a	0.00	0.00	0.40	ORF3a	p.Glu102fs	25695
	0.00	0.00	0.40	ORF3a	p.Glu102fs	25696
	0.00	1.00	0.00	ORF3a	p.Thr151Ile	25844
	0.00	0.50	0.00	ORF3a	p.Tyr156Tyr	25860
	0.00	0.50	0.00	ORF3a	p.Val163Glu	25880
	0.00	0.50	0.00	ORF3a	p.Val168Phe	25893
	0.00	0.50	0.00	ORF3a	p.Val168Phe	25894
	0.00	0.50	0.00	ORF3a	p.Ser252*	26147
	0.00	0.50	0.00	ORF3a	p.Ile263Ile	26181
	0.00	0.50	0.00	ORF3a	p.Pro267Gln	26192
E	0.40	0.00	0.00	E	p.Ser60Ser	26424
M	0.91	0.00	0.00	M	p.Leu93Leu	26801
ORF7a	0.00	0.00	0.45	ORF7a	p.Ala105Ala	27708
ORF7b	1.00	0.00	0.00	ORF7b	p.Ser5Leu	27769
ORF8	1.00	0.00	0.41	ORF8	p.His17His	27944
	0.00	0.52	0.59	ORF8	p.Gln27*	27972
	0.00	0.52	0.59	ORF8	p.Arg52Ile	28048
	0.00	0.42	0.49	ORF8	p.Tyr73Cys	28111
N	0.00	0.00	0.49	N	p.Arg41Arg	28396
	0.00	0.00	1.00	N	p.ArgGly203LysArg	28881
	0.00	0.00	1.00	N	p.ArgGly203LysArg	28882
	0.00	0.00	1.00	N	p.ArgGly203LysArg	28883
	1.00	0.00	0.00	N	p.Ala220Val	28932
	0.00	0.00	0.61	N	p.Ser235Phe	28977
ORF10	1.00	0.00	0.47	ORF10	p.Val30Leu	29645

frequency
1
0.4
0


ORF1ab
S
1.00 0.50 1.00 C –> T ORF1ab c.2772_2776delCTACCinsTTACA p.Pro926Thr 3037
0.09 0.17 0.40 C –> A ORF1ab c.5123C>A p.Ala1708Asp 5388
0.80 0.12 0.32 C –> T ORF1ab c.6021_6023delCTGinsTTT p.Trp2008Leu 6286
1.00 1.00 1.00 C –> T ORF1ab c.14144C>T p.Pro4715Leu 14408
0.50 1.00 0.97 A –> G S c.1841A>G p.Asp614Gly 23403
0.05 0.04 0.05 A –> C S c.1883A>C p.Gln628Pro 23445
02–14/12/2020
05–11/02/2021
12–18/02/2021
frequency
1
0.5
0


0.00 0.00 0.00 6954
0.09 0.17 0.40 C –> A C –> A C –> A 5388
0.00 0.00 0.00 3267
0.00 0.25 0.61 C –> T C –> T 28977
0.00 0.42 0.51 A –> G A –> G,T 28111
0.00 0.52 0.59 G –> T G –> T 28048
0.00 0.52 0.59 C –> T C –> T 27972
0.00 0.51 0.00 G –> C 24914
0.00 0.65 0.91 T –> G,G T –> G 24506
0.00 0.20 0.62 C –> T C –> T 23709
0.00 0.22 0.68 C –> A,A C –> A 23604
0.00 0.30 0.41 C –> A C –> A 23271
0.00 0.00 0.80 A –> T 23063
0.00 0.38 0.40 GAT –> CTA GAT –> CTA [28280−28282]
0.00 0.00 0.00 [21991−21993]
0.00 0.00 0.00 [21765−21770]
0.00 0.00 0.00 [11288−11296]
02−14/12/2020
05−11/02/2021
12−18/02/2021
frequency
1
0.5
0


0.00 0.02 0.00 C –> A 1059
0.00 0.00 0.00 5230
0.00 0.02 0.00 A –> G 10323
0.00 0.00 0.00 21801
0.00 0.00 0.01 G –> C 22813
0.00 0.00 0.00 23012
0.00 0.00 0.80 A –> T 23063
0.00 0.00 0.00 23664
0.00 0.00 0.01 G –> T 25563
0.00 0.00 0.00 25904
0.00 0.00 0.00 26156
0.00 0.00 0.00 28887
02−14/12/2020
05−11/02/2021
12−18/02/2021
frequency
1
0.5
0

A.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
NC_045512.2	326	.	T	A	0.0001	.	DP=6;I16=6,0,0,0,187,5875,0,0,360,21600,0,0,19,73,0,0;QS=1,0;MQ0F=0
NC_045512.2	378	.	T	C	1.3e-06	.	DP=7;I16=6,0,1,0,197,6591,18,324,360,21600,60,3600,20,92,5,25;QS=0.916279,0.0837209,0;SGB=-0.379885;RPB=1;MQB=1;BQB=1;MQ0F=0
NC_045512.2	408	.	A	T	2.5e-06	.	DP=8;I16=8,0,0,0,268,9088,0,0,480,28800,0,0,32,174,0,0;QS=1,0;MQ0F=0
NC_045512.2	433	.	T	C	0.0048	.	DP=9;I16=9,0,0,0,296,10046,0,0,540,32400,0,0,40,246,0,0;QS=1,0;MQ0F=0

B.

Rules
"326=='T'", "326=='A'", "326!='G'", "326!='C'", "326!='-'"
"378=='T'", "378=='C'", "378!='G'", "378!='A'", "378!='-'"
"408=='A'", "408!='T'", "408!='G'", "408!='C'", "408!='-'"
"433=='T'", "433!='C'", "433!='A'", "433!='G'", "433!='-'"

C.

Lineage	Rules
B.1.177.17	28931!='A',25613!='G',407!='-',22087!='A'
B.1.177	28931!='A',25613!='G',407!='-'...
B.1	28931!='A',25613!='G',407!='-'...
B.1.177.13	28931!='A',25613!='G',407!='-'...

D.

Lineage	Rules	Total	Tree Overlap	Total Overlap	Total Overlap Var	Tree Ratio	Total Ratio	Total Ratio Var	Tree Avg AD	Total Avg AD	Tota Avg DP	Total Sum AD	Total Sum DP	Avg DP	Total Run Reads	Avg.AF
B.1.177	26800=='C',...	17	0	11	3	0	0.647	0.176	0	6.6	19	33	95	25.20	104214	0.347
B.1.177	26800=='C',...	11	0	7	2	0	0.636	0.181	0	7	23.66	21	71	25.20	104214	0.295
B.1.1.7	26800!='C',...	13	3	11	1	0.230	0.846	0.076	0	2	9	2	9	25.20	104214	0.222