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

Nikolaos Pechlivanis
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Maria Tsagiopoulou
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Maria Maniou
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Anastasis Togkousidis
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Evangelia Mouchtaropoulou
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Taxiarchis Chassalevris
School of Veterinary Medicine, Aristotle University of Thessaloniki
Serafeim Chaintoutis
School of Veterinary Medicine, Aristotle University of Thessaloniki
Maria Petala
Dept. of Civil Engineering, Aristotle University of Thessaloniki
Margaritis Kostoglou
Dept. of Chemistry, Aristotle University of Thessaloniki
Thodoris Karapantsios
Dept. of Chemistry, Aristotle University of Thessaloniki
Stamatia Laidou
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Elisavet Vlachonikola
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Anastasia Chatzidimitriou
Institute of Applied Biosciences, Centre of Research and Technology Hellas
Agis Papadopoulos
EYATH S.A., Thessaloniki Water Supply and Sewerage Company S.A.
Nikolaos Papaioannou
School of Veterinary Medicine, Aristotle University of Thessaloniki
Chrysostomos Dovas

School of Veterinary Medicine, Aristotle University of Thessaloniki

Anagnostis Argiriou

Institute of Applied Biosciences, Centre of Research and Technology Hellas

Fotis Psomopoulos (fpsom@certh.gr)

Institute of Applied Biosciences, Centre of Research and Technology Hellas

---

Research Article

Keywords: SARS-CoV-2, next-generation sequencing (NGS), novel methodology, called lineagespot, entrepreneurs, traders

Posted Date: July 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-677811/v1

Version of Record: A version of this preprint was published at Scientific Reports on February 17th, 2022. See the published version at https://doi.org/10.1038/s41598-022-06625-6.

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 Togkousidis[1], Evangelia Mouchtaropoulou[1], Taxiarchis Chassalevris[3], Serafeim C. Chaintoutis[3], Maria Petala[4], Margaritis Kostoglou[5], Thodoris Karapantsios[5], Stamatia Laidou[1,2], Elisavet Vlachonikola[1,2], Anastasia Chatzidimitriou[1], Agis Papadopoulos[6], Nikolaos Papaioannou[3], Chrysostomos I. Dovas[3], Anagnostis Argiriou[1,7], 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] EYATH S.A., Thessaloniki Water Supply and Sewerage Company S.A., Thessaloniki, 54636, Greece
[7] Department of Food Science and Nutrition, University of the Aegean, 81400 Myrina, Lemnos, Greece

[*]Correspondence to fpsom@certh.gr

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 monitoring of variants and the detection of SARS-CoV-2 lineages in wastewater samples using next-generation sequencing (NGS). Our proposed method was tested and evaluated using NGS data produced by the sequencing of fourteen wastewater samples from the municipality of Thessaloniki, Greece, covering a six month period. The results showed a clear identification of trends in the presence of SARS-CoV-2 mutations in wastewater data. Lineagespot was able to record the evolution and rapid domination of the B.1.1.7 lineage in the community, 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, which the underlying virus variant is. However, these methods are not easily scalable, especially in large urban areas, were a high number of individuals have to be tested to assess virus and variant spread among the population. 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 wastewater agreed 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 as well as the so-called South African mutation for the first time. In other words, the city of Thessaloniki is a suitable place as a case study for identifying new variants in the city wastewater before scattering to the rest of the country.

The presence of SARS CoV-2 RNA in wastewater 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 wastewater, 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 of 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 (NGS). The method is tested and validated across fourteen municipal wastewater samples retrieved in Thessaloniki, Greece in fourteen different time periods, and correlated with the variants and 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, 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 initially applied on a selected wastewater 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 Over-lap	Total Over-lap	Total Over-lap 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	25.2	104214	0.347
B.1.177	26800 =='C', …	11	0	7	2	0	0.636	0.181	0	7	23.66	21	71	25.2	104214	0.295
B.1.1.7	26800 !='C', …	13	3	11	1	0.230	0.846	0.076	0	2	9	2	9	25.2	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: Table shows the corresponding metrics to a unique lineage, after the merging process. The 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	0.231	0.846	0.077	2	0.223
B.1.177	0.003	0.611	0.057	9.769	0.488

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 (**Figure 1A**). 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) across 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

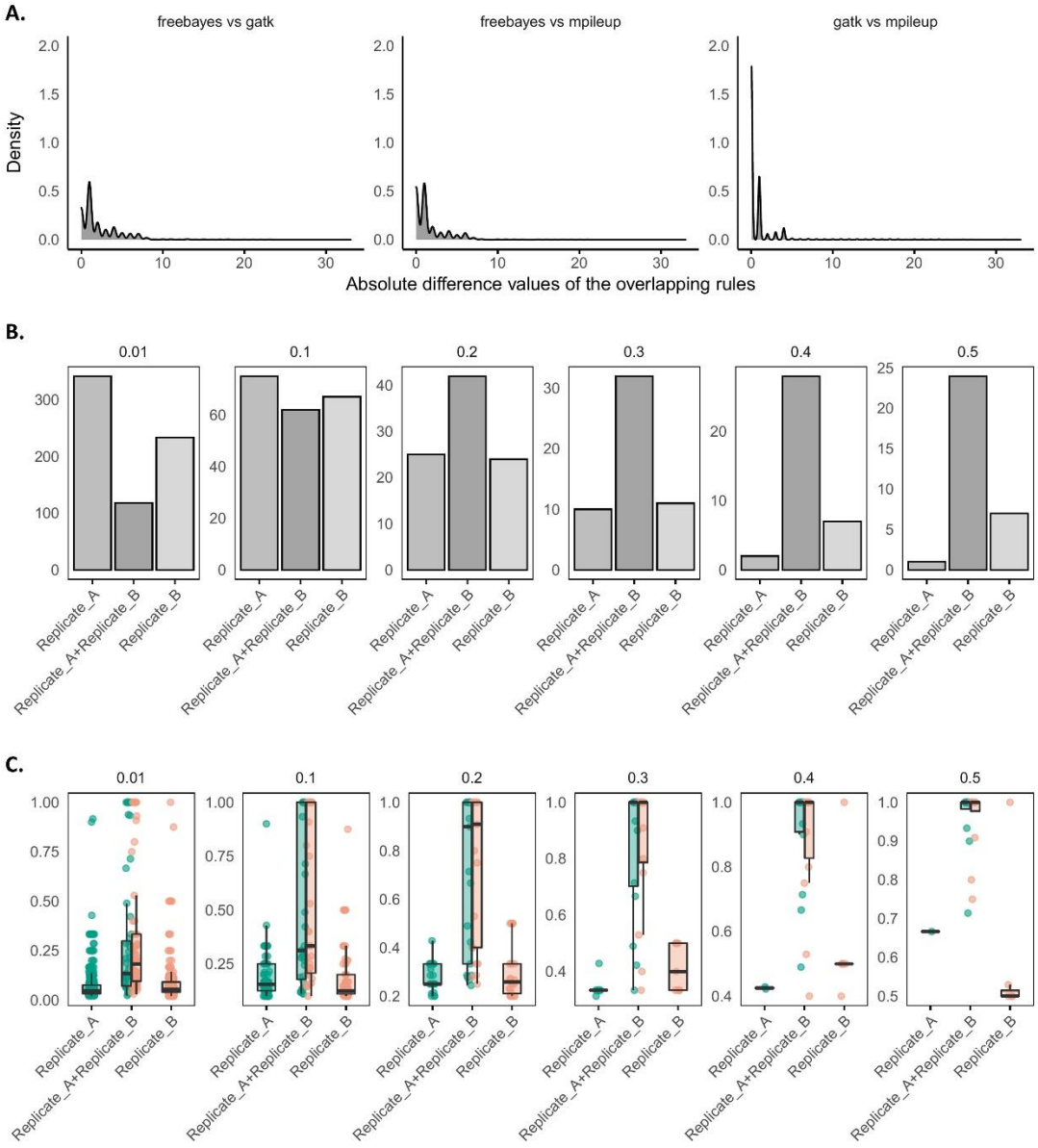


Figure 1: Evolution of variants across different low frequency parameters. **A.** Density plot of the absolute n_t difference values between the output of the three variant calling tools use (pairwise comparisons). **B.** Number of reads for each replicate and for the common variants **C.** The corresponding allele frequency for each replicate and for the common variants

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

Sensitivity

In order to investigate how specific variants, evolve over time, all variant files were merged into a combined table in which all the detected nucleotide variants along with the corresponding amino acid substitutions that have been identified are stored. Moreover, information about the read depth, the allele frequency, and the overlapping gene is also provided for each variant. **Table 5** gives an example of the overall information.

Table 5: Snapshot of table containing every mutation per sample along with the corresponding gene and the amino acid variant.

CHROM	POS	REF	ALT	DP	AD alt	TYPE	Gene Name	HGVS	AF	sample
NC_045512.2	326	T	A	7	1	snp	ORF1ab	Leu21Ile	0.143	Sample A
NC_045512.2	378	T	C	10	1	snp	ORF1ab	Val38Ala	0.1	Sample A
NC_045512.2	408	A	T	10	1	snp	ORF1ab	Asp48Val	0.1	Sample B
NC_045512.2	433	T	C	10	2	snp	ORF1ab	Val56Val	0.2	Sample C
NC_045512.2	442	C	T	10	1	snp	ORF1ab	Gly59Gly	0.1	Sample C

Having a structured format, we first examined if the number of reads that were produced during PCR amplification is affecting the number of variants that are identified in every sample. For this reason, two replicates of the same biological specimen were produced. The first replicate (*Replicate A*) contained 181,880 reads while the second replicate (*Replicate B*) contained 69,706 reads. The two replicates were analyzed as described in the "*Raw data analysis*" Section and two VCF files were produced which contained 401 and 293 variants respectively; of these, 59 variants were common in both replicates, 342 variants were unique for Replicate A and 234 unique for Replicate B.

A Student's t-test was performed in order to compare the mean values of the allele frequency between the two replicates and an F test to compare the variances. All tests resulted in p values higher than the 0.05 threshold, meaning that no statistically significant difference was found between the two replicates. On the contrary, when comparing two samples coming from different time periods there was a significant difference on the variants' allele frequency (p values from Student's t-test and F test were below 0.05).

Moreover, for the same replicates, 6 different pairs of VCF files were generated in order to investigate how does the number of variants located changes while increasing the low variant frequency parameter of 0.01 that was used at *freebayes* tool during the raw data analysis. The thresholds that were chosen were 0.01, 0.1, 0.2, 0.3, 0.4 and 0.5. In **Figure 1B**, the evolution of the number of variants found is shown while **Figure 1C** gives the corresponding allele frequency.

Mutational load detection

The proposed methodology was applied on the wastewater samples, across fourteen time periods, as show in Supplementary **Table 1.**

To this end, **Table 5** was collapsed at the gene and amino acid substitution level, therefore reducing any data noise that is introduced by nucleotide mutations that correspond to the same amino acid change. In **Figure 2A** all variants were clustered accordingly based on Euclidean distance, while **Figure 2B** highlights specific variants (cluster 1) that exhibit significant difference in behavior.

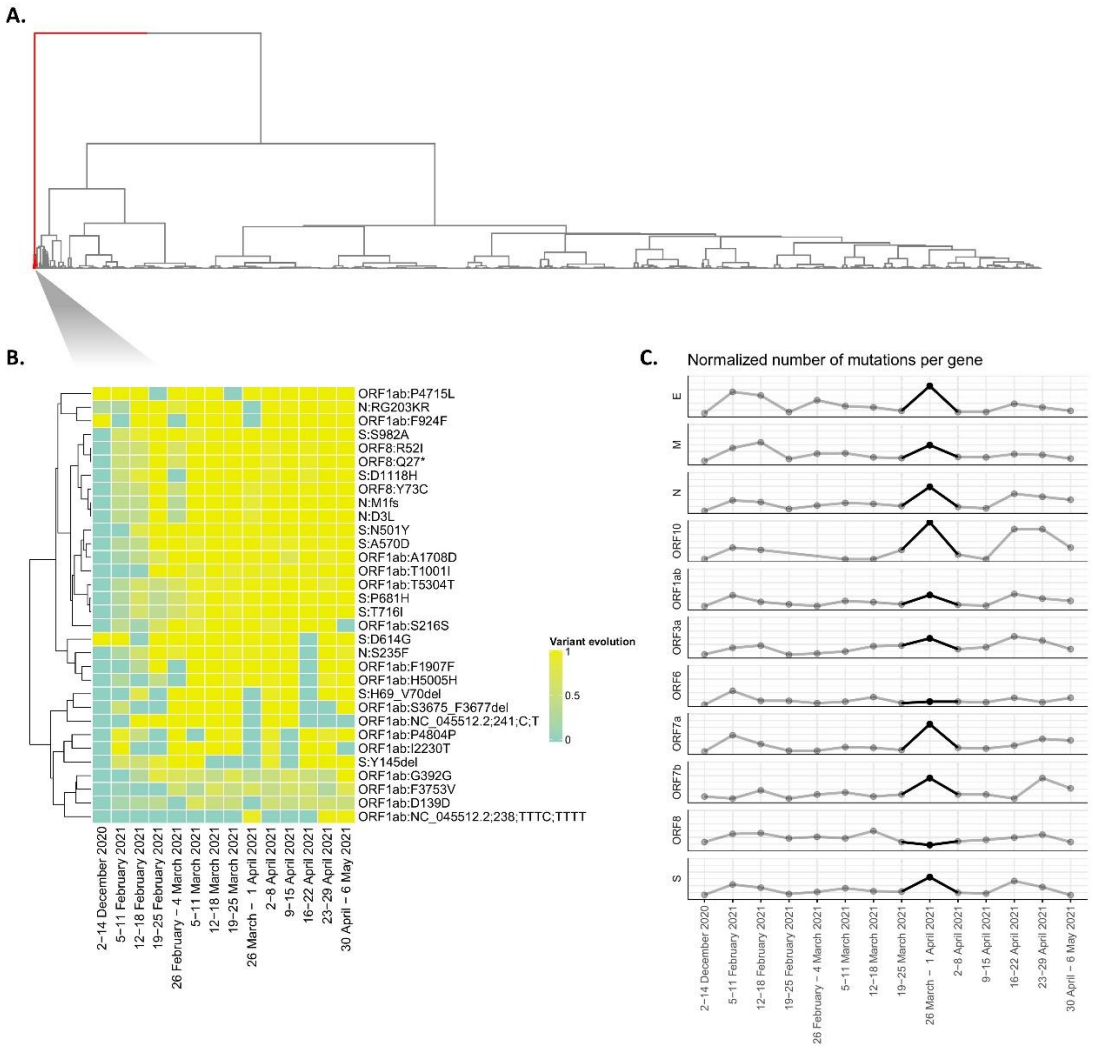


Figure 2: Unsupervised variant clustering was performed on a table containing all amino acid substitutions **A.** Hierarchical clustering shows the total variants using the Euclidean distance as a distance metric and ward.D as a clustering method. **B.** Hierarchical clustering based on the cluster 1 of the Fig.2A. The heatmap shows the variant evolution across the different periods. **C.** Number of mutations per gene across the different periods. The values of the plot were normalized based on the length of each gene.

Moreover, we studied the number of amino acid variants per gene over each time period. In **Figure 2C** the number of variants were clustered based on Euclidean distance, after all count values were normalized by each gene's length.

Detection of Variants of Concern / Variants of Interest

Furthermore, the evolution of lineage-specific variants over the eleven time periods was studied. To this end, two Variants of Concern were selected; the B.1.1.7 (UK) and the B.1.351 (South Africa) lineages which are shown in **Figure 3**.

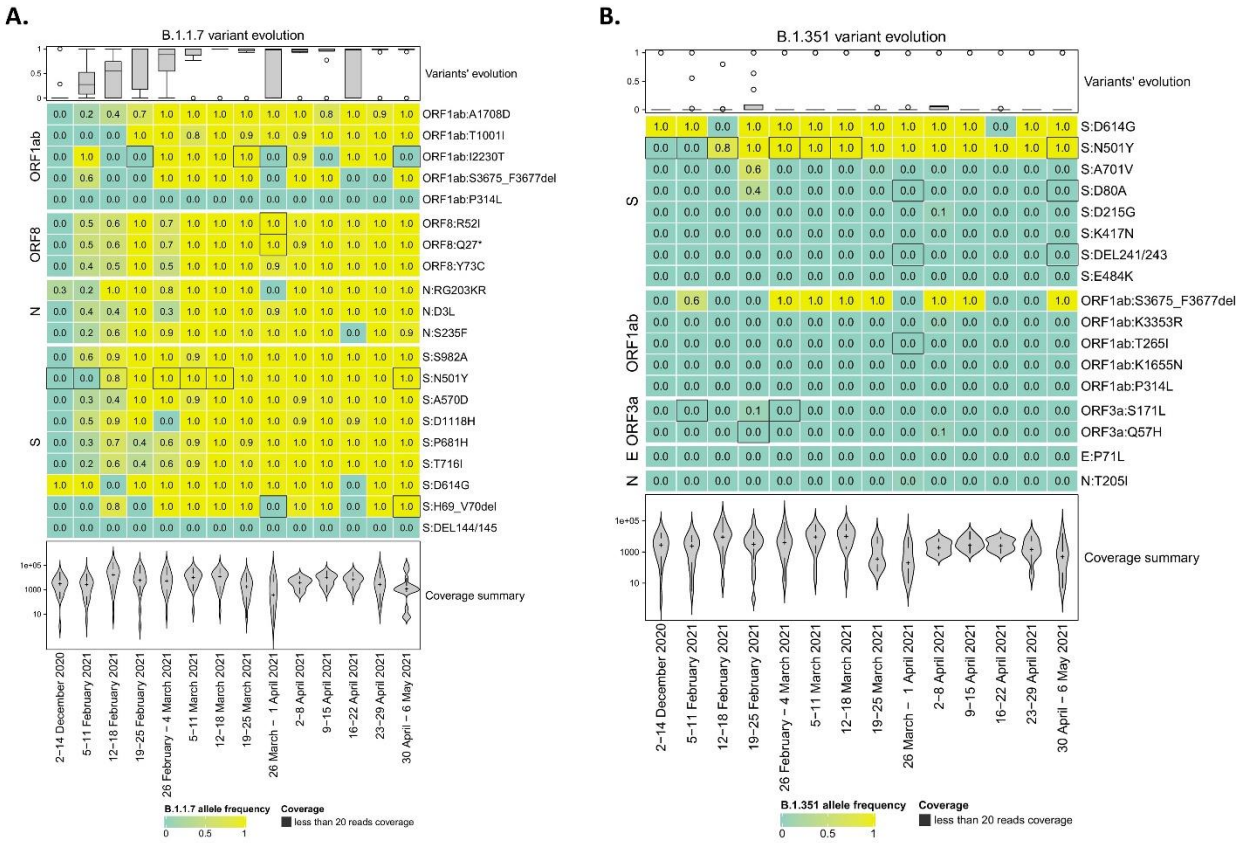


Figure 3: Clustering amino acid substitutions for B.1.1.7 (UK lineage) and B.1.351 (South Africa lineage) strains. Heatmap displays the corresponding allele frequency (AF) of each period per amino acid variant. **A.** Evolution of B.1.1.7-detected mutations (UK Lineage) **B.** Evolution of B.1.351-detected mutations (South Africa lineage). Positions with low coverage (less than 20 reads) are depicted with dark gray color.

Moreover, we examined the prevalence rate of the two variants of concern by calculating the average allele frequency of their mutations. The results for the 14 time periods are presented in **Table 6** and show that it does not lead to 100% sum per time period. The latter is caused, due to overlapping mutations among the lineages and implies that the average value of all the mutations is not a reliable metric to characterize a specific lineage's presence.

Table 6: Allele frequency metrics computed for the comparison with the clinical data. To this end the average allele frequency of all mutations, the average allele frequency of the unique mutations and the minimum allele frequency were calculated for each time period.

	Average allele frequency of the mutation		Average allele frequency of the unique mutation		Minimum allele frequency of the present (non-zero) unique mutation	
	B.1.1.7	B.1.351	B.1.1.7	B.1.351	B.1.1.7	B.1.351
2 – 14 Dec. 2020	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
5 – 11 Feb. 2021	32.71%	0.00%	35.15%	0.00%	16.67%	0.00%
12 – 18 Feb. 2021	51.92%	9.05%	49.20%	0.21%	39.53%	1.47%
19 – 25 Feb. 2021	79.78%	22.15%	85.84%	14.20%	58.06%	35.48%
26 Feb. – 4 Mar. 2021	78.18%	11.11%	77.51%	0.00%	31.88%	0.00%
5 – 11 Mar. 2021	95.54%	11.11%	96.28%	0.00%	76.71%	0.00%
12 – 18 Mar. 2021	94.12%	11.11%	92.31%	0.00%	100.0%	0.00%
19 – 25 Mar. 2021	92.58%	11.41%	91.14%	0.56%	92.75%	3.92%
26 Mar. – 1 Apr. 2021	74.09%	11.51%	74.27%	0.63%	89.49%	4.44%
2 – 8 Apr. 2021	96.89%	12.30%	95.98%	1.56%	92.14%	2.03%
9 – 15 Apr. 2021	86.20%	11.10%	81.99%	0.00%	77.15%	0.00%
16 - 22 Apr. 2021	74.32%	10.97%	74.50%	0.00%	89.22%	0.00%
23 - 29 Apr. 2021	81.60%	11.11%	83.81%	0.00%	93.41%	0.00%
30 Apr. - 6 May 2021	93.54%	11.11%	91.55%	0.00%	93.85%	0.00%

Further in the analysis, we attempted to examine the minimal lineage support per time period. In this step we calculated the average allele frequency of the mutations unique to each variant of concern and produced a table which is more comparable with the clinical data metric. In this context, the minimum allele frequency of the present (greater than zero) mutations indicates the level of our confidence. The results of the two metrics are shown in **Table 6**.

Assessing lineage assignment

Qualitative assessment between the major lineages found in the targeted clinical samples, retrieved from the ENA study ID: PRJEB44141 (ERP128154), and across the fourteen time periods. The clinical samples exhibit the lineage distribution shown in **Table 7**.

Table 7: Lineage distribution of samples across all periods. The prevalent strain is depicted with bold font

Period	No. of patients / period	Lineages Detected	Percentage (%)
2 – 14 Dec. 2020	3	**B.1.177**	66.7
		B.1.1.189	33.3
5 – 11 Feb. 2021	57	B.1.1.7	43.9
		B.1.177	54.4
		B.1.351	1.75

		B.1.1.7	53.1
12 – 18 Feb. 2021	32	B.1.177	37.5
		B.1.351	9.38
19 – 25 Feb. 2021	21	**B.1.1.7**	81
		B.1.177	19
		B.1	0.935
		B.1.1.7	83.2
26 Feb. – 4 Mar. 2021	214	B.1.177	11.7
		B.1.258	2.8
		B.1.351	1.4
5 – 11 Mar. 2021	13	**B.1.1.7**	100
12 – 18 Mar. 2021	22	**B.1.1.7**	100
19 – 25 Mar. 2021	5	**B.1.1.7**	100
26 Mar. – 1 Apr. 2021	3	**B.1.1.7**	100
2 – 8 Apr. 2021	3	**B.1.1.7**	100
9 – 15 Apr. 2021	35	**B.1.1.7**	100
16 – 22 Apr. 2021	32	B.1.1	3.12
		B.1.1.7	96.9
23 – 29 Apr. 2021	66	B.1.1.318	1.52
		B.1.1.7	98.5
30 Apr. – 6 May 2021	45	**B.1.1.7**	100

Based on the clinical results, as derived from targeted and non-randomized sampling of the Thessaloniki area, we can perform a direct comparison (**Figure 4**) between the level of presence in a particular variant of concern (B.1.1.7) in clinical and wastewater data.

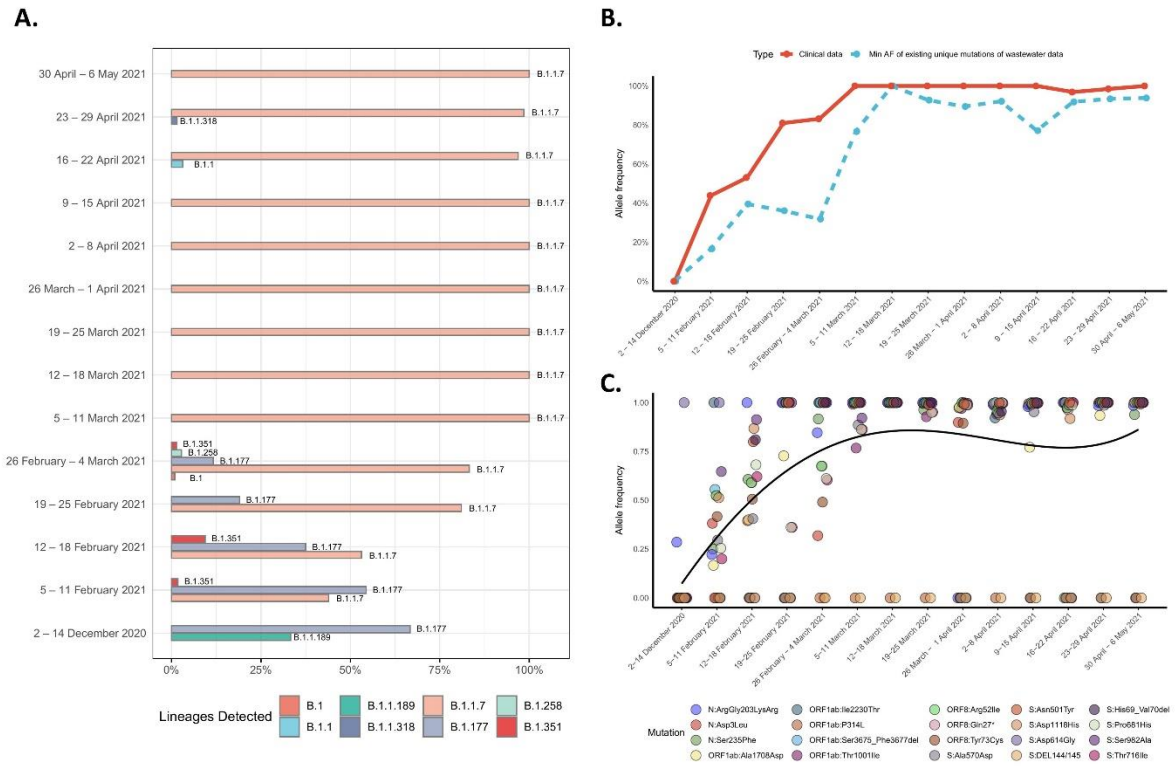


Figure 4: Comparison between wastewater samples and clinical data. **A.** SARS-CoV-2 lineages detected on clinical samples over all time periods **B.** The percentage of presence of the B.1.1.7 variant of concern (VoC) in the clinical samples (red line) and the estimated minimum level of presence of the same VoC in the wastewater data (blue line). **C.** Average percentage of the presence of each characteristic mutation of the B.1.1.7 variant of concern (VoC). The line corresponds to the average value per time period. Mutations that are not found in a particular timepoint are detected in the neighbouring ones, thus leading to the variations in the average.

By performing the analysis on the wastewater data, we identified the profiles presented in **Table 8**, that are representative of the most prevalent variants identified in each period and based on the lineage assignment model offered by the Pangolin tool.

Table 8: Number of Pangolin's rules satisfied by the mutations detected in wastewater samples across all time periods.

Lineage	Mean (Tree Ratio)	Mean Total Ratio	Mean Total Ratio Var	Mean Total Av AD	Mean AF	Period
B.1.177	0	0.564	0.0101	12.51734	0.562649	2 – 14 Dec. 2020
B.1.1.7	0.022619	0.618	0.032266	7.050461	0.480622	
B.1.1.7	0.022619	0.651	0.067954	7.033142	0.346846	5 – 11 Feb. 2021
B.1.177	0	0.556	0.028353	11.35913	0.52117	

B.1.1.7	0.022619	0.634453	0.051619	29.74789	0.482472	12 – 18 Feb. 2021
B.1.177	0	0.576864	0.046912	14.04043	0.349998	
B.1.1.7	0.029478	0.633467	0.059038	16.58363	0.505483	19 – 25 Feb. 2021
B.1.177	0	0.563693	0.02373	17.74751	0.508294	
B.1.1.7	0.02381	0.628576	0.051367	17.47607	0.487571	26 Feb. – 4 Mar. 2021
B.1.177	0	0.571146	0.009646	11.87439	0.337707	
B.1.1.7	0.022619	0.630203	0.054135	38.854	0.514036	5 – 11 Mar. 2021
B.1.177	0	0.558882	0.022893	30.16625	0.512106	
B.1.1.7	0.022619	0.627289	0.071746	24.92428	0.557327	12 – 18 Mar. 2021
B.1.177	0	0.563743	0.018941	18.61508	0.446698	
B.1.1.7	0.027891	0.633067	0.05223	469.058	0.516416	19 – 25 Mar. 2021
B.1.177	0	0.568151	0.025152	762.8515	0.47739	
B.1.1.7	0.041931	0.645114	0.122677	1487.607	0.52834	26 Mar. – 1 Apr. 2021
B.1.177	0	0.559649	0.047282	1920.774	0.55107	
B.1.1.7	0.05285	0.611235	0.111091	938.4528	0.515389	2 – 8 Apr. 2021
B.1.177	0	0.5658	0.020531	1172.69	0.527678	
B.1.1.7	0.022619	0.64304	0.085195	1750.538	0.400779	9 – 15 Apr. 2021
B.1.177	0	0.557766	0.01395	1538.826	0.445925	
B.1.1.7	0.257666003	0.724088929	0.067766428	28.4705103	0.455354099	16 – 22 Apr. 2021
B.1.177	0	0.549734606	0.028445651	43.86612546	0.563500191	
B.1.1.7	0.132953798	0.688224115	0.066440254	33.76142848	0.495087239	23 – 29 Apr. 2021
B.1.177	0.000575705	0.557364992	0.036746317	47.34227275	0.51233052	
B.1.177	0	0.550166357	0.02576803	24.96695208	0.429484878	30 Apr. – 6 May 2021
B.1.1.7	0	0.644013622	0.026942757	43.19821767	0.645142425	

Discussion

Analyzing wastewater, i.e. used water that goes through the drainage system to a treatment facility, is a way that researchers and surveillance systems can track pathogens, such as SARS-CoV-2, or biomarkers that are excreted in urine or feces. Monitoring effluents could be a reliable and more effective tool to estimate SARS-CoV-2 spread compared to 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 viral load and SARS-CoV-2 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, since only a few samples must be processed and analyzed.

In this manuscript we present and validate a methodology named ***lineagespot***, making use of NGS 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 and was capable of recording the evolution of the B.1.1.7 lineage in the community. 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 fourteen time points by whole genome sequencing of samples from the general population.

Overall, the method developed herein was proven superior compared to other methodologies (Sanger sequencing)[4,7], since it was more informative and 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 9.** 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 wastewater. The residence time of wastewater 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 9: Main quality characteristics of wastewater samples

Parameter/ Sample period	pH	Electrical Conductivity (S/cm)	Total Suspended Solids (mg/L)	BOD_5 (mg/L)	COD (mg/L)	Dissolved Organic Carbon (mg/L)	UV absorption at 254 nm (1/cm)	Total Nitrogen (mg/L)	Ammonium Nitrogen (mg/L)	Total Phosphorus (mg/L)	Copies/µl
02-14 Dec. 2021	7.5	8.5	620	385	960	35	0.35	62	28.5	11.5	36
05-11 Feb. 2021	7.8	9.6	930	525	1,250	49	0.4	76	33	11.5	68
12-18 Feb. 2021	7.8	4.6	1,200	650	1,570	44	0.45	95	38	12	53
19-25 Feb. 2021	7.9	3.5	1,225	684	1,610	56	0.49	95	38.2	15.2	82
26 Feb. – 4 Mar. 2021	7.8	2.9	1,225	535	1,383	53.5	0.47	78.5	36.7	12.4	179
5-11 Mar. 2021	7.6	2.8	1,017	540	1,373	50.2	0.47	71.7	36.8	12.1	102
12-18 Mar. 2021	7.8	4.5	852	580	1,285	66.3	0.48	76.4	37.4	11.7	277
19-25 Mar. 2021	7.6	4.1	926	582	1,467	60.8	0.48	79.3	39.6	11.6	467
26 Mar. – 3 Apr. 2021	7.6	3.4	1,095	660	1,708	52.1	0.44	85.9	41.4	12.3	494
2-08 Apr. 2021	7.6	4.2	1,054	667	1,537	52	0.48	88.1	40.5	13.2	498
09-15 Apr. 2021	7.7	4.1	1,025	579	1,464	55.6	0.5	80	32.1	11.4	505

Sampling and handling of the wastewater 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 4,000 × *g* 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 respective 0.45-µm pore-size, 47-mm diameter electronegative membranes (HAWP04700; Merck Millipore, Ireland). Each membrane filter was rolled into a Falcon™ 15-mL conical centrifuge tube with the top side facing inward, and was subjected to RNA extraction.

RNA extraction and SARS-CoV-2 quantification

Each electronegative membrane was subjected to RNA extraction process based on a phenol-chloroform-method[8] coupled with magnetic bead binding. The following reagents were added sequentially: a) 900 µL of guanidinium isothiocyanate-based "Lysis buffer I" [5 M guanidinium isothiocyanate, 25 mM EDTA, 25 mM sodium citrate (pH 7.0), 25 mM phosphate buffer (pH 6.6)] containing 1% N-Lauroylsarcosine, 2% Triton X-100, 2% CTAB and 2% PVP, b) 18 µl β-mercaptoethanol, c) 300 µl H_2O. Tubes mixed thoroughly by inversion and were stored at 4 °C for 10-30 min. Subsequently, 1.2 mL of "Lysis buffer II" [prepared by mixing 152.5 gr guanidinium hydrochloride, 31.25 mL of 2 M acetate buffer (pH 3.8) and water-saturated phenol stabilized (pH 4), at a final volume of 500 ml] was added, followed by incubation on a horizontal rotator (150 rpm, 10 min, RT). The liquid phase was transferred into a 2-mL microcentrifuge tube, was clarified by centrifugation (21,000 × *g*, 5 min, 4 °C) and 1.6 mL of the supernatant were transferred to a new 2-mL tube, wherein 200 µl chloroform-isoamyl alcohol (24:1) were added, followed by vigorous shaking for 30 s, incubation (−20 °C, 30 min) and centrifugation (21,000 × *g*, 10 min, 4 °C). The upper aqueous phase (800 µL) was transferred and mixed with 667 µL isopropanol and 20 µL of magnetic beads (IDEXX Water DNA/RNA Magnetic Bead Kit; IDEXX Laboratories Inc., Westbrook, ME, USA), followed by incubation on a horizontal rotator (150 rpm, 15 min, RT). The beads were washed according to the manufacturer's protocol. RNA was eluted in 100 µl buffer, and eluates were subjected to filtration, using the OneStep™ PCR Inhibitor Removal Kit (Zymo Research Corporation, Irvine, CA, USA) and were stored at -80 °C. Extracted RNAs originating from 12 processed electronegative membranes and spanning 6 different days of sampling were pooled (1.1 mL total RNA extract) and mixed with 2.2 mL binding buffer containing isopropanol (IDEXX Water DNA/RNA Magnetic Bead Kit). Half of the mixture (1.65 mL) was incubated with 20 µL of magnetic beads on a horizontal rotator (150 rpm, 15 min, RT). After the magnetic separation of beads, the supernatant was removed and the procedure was repeated by adding the remaining 1.65 mL of the mixture. The beads were washed according to the manufacturer's protocol and RNA was eluted in 60 µL buffer.

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). Calibration curves were generated using the synthetic single-stranded RNA standard "EURM-019" (Joint Research Centre, European Commission) and SARS-CoV-2 viral loads were expressed as genome copies per µL of RNA extract.

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 Super Script 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 µl of the generated cDNA was used instead of 6 µl, 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 10 µM 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 were 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 were retained, while any other (unmatched, multiple mappings, etc.) were removed. In the next step, two different computational workflows were employed, corresponding to the two different sequencing protocols employed. For the first seven samples and the last three (2 December 2020 – 18 March 2021 and 16 April 2021 – 06 May 2021), 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). For the four samples between 19 March 2021 and 15 April 2021 primer trimming and remapping to reference genome was not applied, as the updated protocol used did not necessitate this step. Finally, 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.

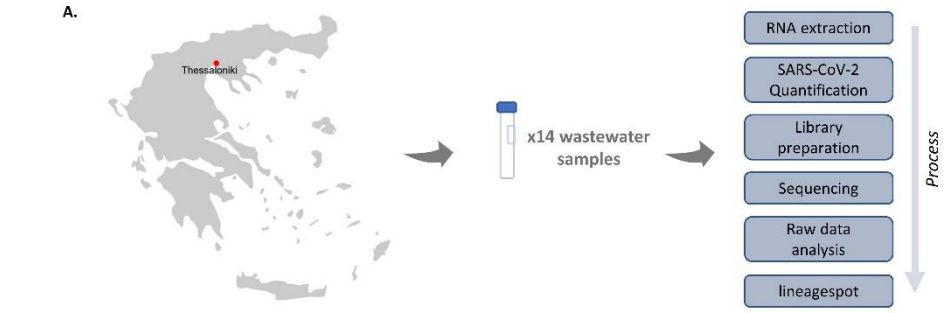
Downstream analysis of lineages detection

In order to identify and assign different SARS-CoV-2 lineages based on the mutations detected from a single wastewater 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 5** 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.



B.

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=0.91 6279,0.0837209,0;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

C.

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!='-'"

D.

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!='-'...

E.

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

Figure 5: Snapshots of the intermediate steps. **A.** A summary plot showing the overall process from the sampling to LineageSpot **B.** A VCF file produced by the chosen variant caller **C.** The rules as they are generated by the VCF file. **D.** Pangolin's decision rules. **E.** 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 of the lineage's rules.

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.

Source of lineage-specific mutations

For the detection of lineage-specific mutations, three different data sources were used; pangolin, VEO and *outbreak.info*. An example of the differences of the three sources is provided in **Supplementary Figure 1**, which compares the detection of B.1.1.7 using **A** outbreak.info, **B** pangolin and **C** VEO data.

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. Chaintoutis, S. C., Papadopoulou, E., Melidou, A., Papa, A. & Dovas, C. I. A PCR-based NGS protocol for whole genome sequencing of West Nile virus lineage 2 directly from biological specimens. *Molecular and Cellular Probes* **46**, 101412 (2019).
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*.

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). Support was also received by the Region of Central Macedonia through the project on "*Epidemiological status of COVID-19 disease based on viral loading monitoring in wastewater*", as well as through the "*SARS-*

CoV-2 RNA monitoring in untreated wastewater" Initiative of the Hellenic National Public Health Organization. Finally, this work was supported by *ELIXIR*, the research infrastructure for life-science data.

NPe, MT, MCM and AT developed the ***lineagespot*** code and performed the downstream analysis. NPe and MT performed the analysis of the raw NGS data. EM, SL and EV did the library preparation. CID and AC provided the data and reviewed the submitted version. CID, TC, SCC, MP and MK participated in experimental investigation. CID, AP, NPa, FP 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.

All raw FASTQ files are deposited on ENA: Project IDs **PRJEB44141** (*for patient samples*) and **PRJEB44548** (*for wastewater samples*).

Supplementary Figure legend

Supplementary Figure 1: Detected mutations of B.1.1.7 (UK Lineage) using data provided by **A.** outbreak.info, **B.** pangolin and **C.** VEO.

Supplementary Table caption

Supplementary Table 1: Wastewater samples used for the SARS-CoV-2 variant analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplTable1.pdf
- SupplementaryFigure1.jpg