medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples

Katharina Jahn[1,2,*], David Dreifuss[1,2,*], Ivan Topolsky[1,2,*], Anina Kull[3], Pravin Ganesanandamoorthy[3], Xavier Fernandez-Cassi[4], Carola Bänziger[3], Elyse Stachler[3], Lara Fuhrmann[1,2], Kim Philipp Jablonski[1,2], Chaoran Chen[1,2], Catharine Aquino[5], Tanja Stadler[1,2], Christoph Ort[3], Tamar Kohn[4], Timothy R. Julian[3], Niko Beerenwinkel[1,2,+]

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland;
[2]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;
[3]Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland;
[4]Laboratory of Environmental Chemistry, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland;
[5]Functional Genomics Center Zurich, ETH Zurich, 8057 Zurich, Switzerland

*Equal contributions
[+]Correspondence to: niko.beerenwinkel@bsse.ethz.ch

Abstract

The SARS-CoV-2 lineages B.1.1.7 and 501.V2, which were first detected in the United Kingdom and South Africa, respectively, are spreading rapidly in the human population. Thus, there is an increased need for genomic and epidemiological surveillance in order to detect the strains and estimate their abundances. Here, we report a genomic analysis of SARS-CoV-2 in 48 raw wastewater samples collected from three wastewater treatment plants in Switzerland between July 9 and December 21, 2020. We find evidence for the presence of several mutations that define the B.1.1.7 and 501.V2 lineages in some of the samples, including co-occurrences of up to three B.1.1.7 signature mutations on the same amplicon in four samples from Lausanne and one sample from a Swiss ski resort dated December 9 - 21. These findings suggest that the B.1.1.7 strain could be detected by mid December, two weeks before its first verification in a patient sample from Switzerland. We conclude that sequencing SARS-CoV-2 in community wastewater samples may help detect and monitor the circulation of diverse lineages.

Background

Viral RNA of SARS-CoV-2 infected persons can be found in faeces (Gupta et al. 2020) and accordingly, at the community level, in the wastewater collected in wastewater treatment plants (WWTPs) (Kitajima et al. 2020; Medema et al. 2020; Gonzalez et al. 2020). In principle, wastewater samples can provide a snapshot of the circulating viral lineages and their diversity in the community (Izquierdo Lara et al. 2020; Nemudryi et al. 2020; Martin et al. 2020; Crits-Christoph et al. 2020) and serve as an efficient and complementary approach to genomic epidemiology based on individual patient samples (Nadeau et al. 2020).

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

However, it is challenging to analyze wastewater samples for SARS-CoV-2 genomic composition, because samples are enriched in PCR inhibitors during the virus concentration steps, viral genomes may be fragmented, and sewage contains large amounts of bacterial, human and other viral DNA and RNA genomes (Cantalupo et al. 2011; Fernandez-Cassi et al. 2018; Martínez-Puchol et al. 2020). Even if amplified and sequenced successfully, the read data obtained from the mixture of viral genomes from different infected persons is difficult to interpret, because of amplification biases, sequencing errors, and insufficient phasing information. Detecting a viral lineage that is present only in a small fraction of infected persons, such as, for example, a recently mutated or imported new viral variant into an area with an ongoing local outbreak, is therefore a major challenge, similar to identifying low-frequency variants within single-patient samples (Kuipers et al. 2020; Karamitros et al. 2020; Rose et al. 2020). On the other hand, tracking the genomic composition of viral RNA in the community in multiple samples over time increases our ability to make reliable calls and offers an opportunity to monitor the circulation of diverse SARS-CoV-2 lineages.

The novel SARS-CoV-2 variants B.1.1.7 (or VOC-202012/01), which was first found in the UK (Rambau et al. 2020), and 501.V2 (or S501Y.V2), which was identified in South Africa (Tegally et al. 2020), have recently generated a lot of interest as they may be associated with increased infectivity and hence accelerated spread in the human population (Volz et al. 2020, Vöhringer et al. 2020) (Tegally et al. 2020). Although some of the mutations defining these lineages arose independently before – without any evidence for increased transmissibility so far (van Dorp et al. 2020) – the combination of 17 and 12 genetic alterations, defining B.1.1.7 and 501.V2, respectively, is unique and may have unexpected consequences on the phenotype. In Switzerland, the B.1.1.7 lineage was first detected in two clinical samples on December 24 (Bundesamt für Gesundheit 2020a), and the 501.V2 lineage in another two samples on December 27 (Bundesamt für Gesundheit 2020b). To our knowledge, one of the 501.V2 cases has no known epidemiological links to abroad and several recent B.1.1.7 cases identified in Geneva have no known epidemiological links to abroad indicating local transmission (Département de la sécurité, de l'emploi et de la santé, Republic et Canton de Geneve 2021). However, it remains unknown to what extent the new variants are already circulating in Switzerland. Sequencing community wastewater samples offers an efficient and cost-effective way to address this epidemiologically important question.

Here, we report a preliminary analysis of next-generation sequencing (NGS) data obtained from a total of 48 wastewater samples collected from three WWTPs in Switzerland between July 8 and December 21, 2020. We explore what SARS-CoV-2 wastewater genomic data can reveal about arising lineages and their circulation in the community and compare our findings to about 5000 SARS-CoV-2 patient sequences collected in Switzerland.

Methods

Patient sequences. Per-patient SARS-CoV-2 consensus sequences were downloaded from GISAID (Shu and McCauley 2017) for all samples collected in Switzerland between February 24 and December 24, 2020, and not identified as either B.1.1.7 or 501.V2 (see supplementary material for the list of accession numbers).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Wastewater sample collection and preparation. Raw wastewater samples were collected from three Swiss WWTPs: ARA Werdhölzli, Zurich (population connected: 450,000), STEP de Vidy, Lausanne (population connected: 240,000), and the WWTP of an alpine ski resort. Samples were concentrated and viral RNA was extracted using methods similar to those previously reported (Medema et al. 2020). In brief, 24-hour composite samples (Zurich and Lausanne) or grab samples (ski resort) were collected in 500ml polystyrene or polypropylene plastic bottles, shipped on ice, and stored at 4°C for up to 6 days before processing. Aliquots of 50 mL were clarified by filtration (2 µm glass fiber filter (Millipore) followed by a 0.22 µm filter (Millipore), Zurich samples), or by centrifugation (4,863 xg for 30 minutes, Lausanne and ski resort samples). Clarified samples were then concentrated using centrifugal filter units (Centricon Plus-70 Ultrafilter, 10kDa, Millipore, USA) by centrifugation at 3,000xg for 30 minutes. The resulting concentrate (up to 280 µL) was extracted in its entirety using the QiaAmp Viral RNA MiniKit (Qiagen, USA) according to the manufacturer's instructions, adapted to the larger volumes, and using 80 µL eluent volume. RNA extracts were stored at -80°C before sequencing.

Sequencing. Frozen cDNA transcripts of RNA extracts from wastewater samples were analyzed using NGS. Amplicons were produced and libraries prepared according to the COVID-19 ARTIC v3 protocol (R&D 2020) and sequenced using the Illumina NovaSeq 6000 platform, resulting in paired-ends reads of length 250 bp each.

Mutation calling. Mutation calls were produced using V-pipe (Posada-Céspedes et al. 2020), a bioinformatics pipeline for end-to-end analysis of viral sequencing reads obtained from mixed samples. Low-frequency mutation calling was based on local haplotype reconstruction with ShoRAH (Zagordi et al. 2010a, 2010b, 2011; McElroy et al. 2013; Posada-Céspedes et al. 2020). Custom scripts were used for detecting nearby co-occurring mutations in the multiple read alignments.

Mutation reporting and testing. Samples from the same location were grouped by sampling date into those collected before October 23, 2020, or afterwards. The cut-off date was chosen to ensure that the earlier samples were (likely) taken before the emergence of the B.1.1.7 and 501.V2 variants. Fisher's exact test was applied to test the null hypothesis that the total number of signature mutation observations is independent of sampling time. Contingency tables were obtained by counting in each group the number of times the signature mutations were observed to be present versus absent in the two groups. Missing observations were omitted from the test. The analysis was performed separately for the B.1.1.7 and 501.V2 signature mutations.

Results

The B.1.1.7 lineage is defined by 17 signature events (called mutations for simplicity), including 13 point mutations, 1 three-nucleotide change, and 3 deletions of length 3, 6, and 9 nucleotides each. The 501.V2 lineage is defined by 12 signature mutations, all of which are single point mutations (Table 1). The two lineages share one mutation, namely A23063T giving rise to the amino acid change N501Y in the spike protein. In order to interpret the observation of any of these signature mutations in wastewater samples, we first characterize

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

the mutational background distribution in Switzerland by analyzing patient samples not identified as B.1.1.7 nor 501.V2. Next, we identify signature mutations in wastewater samples, first individually and then, if possible, in combination.

Signature mutations in patient samples

We first assessed the prevalence of signature mutations defining the lineages B.1.1.7 and 501.V2 in all 5192 non-B.1.1.7 and non-501.V2 consensus sequences from GISAID obtained from clinical samples collected in Switzerland before December 24 (Table 1). Among the 17 B.1.1.7 signature mutations, seven occurred in patient samples, in total 121 times (2.33%), each in 2 - 90 samples (0.04 - 1.73%). Only two samples displayed co-occurrence of two mutations, namely (C23604A, C23709T), and no other co-occurrence of mutations was displayed. For 501.V2, 9 out of the 12 signature mutations occurred in patient samples, most of them in only 1 - 8 sequences (0.02 - 0.15%), but A10323G occurred 28 times (0.54%), C28887T 37 times (0.52%), C1059T 203 times (3.91%), and G25563T 1174 times (22.61%). Some of these mutations co-occurred in the consensus sequences as pairs or triplets, namely (C1059T, G25563T, C28887T), (C1059T, A10323G, G25563T) and (C23664T, G25563T) each co-occurred once, (G25563T, C28887T) seven times, (A10323G, G25563T) 9 times, and (C1059T, G25563T) 199 times. Overall, in non-B.1.1.7 and non-501.V2 patient samples from Switzerland, several of the signature mutations had not been detected at all before December 24 or only in very few samples and in isolation.

Table 1. Signature mutations of the B.1.1.7 and 501.V2 lineages and their frequencies among all 5192 non-B.1.1.7 and non-501.V2 consensus sequences available in GISAID and obtained from clinical samples in Switzerland until December 24, 2020.

Lineage	Gene	Nucleotide	Amino acid	Absolute frequency	Relative frequency
B.1.1.7	Orf1ab	C3267T	T1001I	4	0.08 %
B.1.1.7	Orf1ab	C5388A	A1708D	0	0 %
B.1.1.7	Orf1ab	T6954C	I2230T	0	0 %
B.1.1.7	Orf1ab	11288-11296 deletion	SGF 3675-3677 deletion	2	0.04 %
B.1.1.7	Spike	21765-21770 deletion	HV 69-70 deletion	90	1.73 %
B.1.1.7	Spike	21991-21993 deletion	Y144 deletion	0	0 %
B.1.1.7	Spike	A23063T	N501Y	0	0 %
B.1.1.7	Spike	C23271A	A570D	0	0 %
B.1.1.7	Spike	C23604A	P681H	12	0.23 %
B.1.1.7	Spike	C23709T	T716I	2	0.04 %
B.1.1.7	Spike	T24506G	S982A	0	0 %
B.1.1.7	Spike	G24914C	D1118H	0	0 %
B.1.1.7	Orf8	C27972T	Q27stop	7	0.13 %
B.1.1.7	Orf8	G28048T	R52I	0	0 %

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

B.1.1.7	Orf8	A28111G	Y73C	0	0 %
B.1.1.7	N	GAT28280CTA	D3L	4	0.08 %
B.1.1.7	N	C28977T	S235F	0	0 %
501.V2	Orf1ab	C1059T	T265I	203	3.91 %
501.V2	Orf1ab	G5230T	K1655N	8	0.15 %
501.V2	Orf1ab	A10323G	K3353R	28	0.54 %
501.V2	Spike	A21801C	D80A	0	0 %
501.V2	Spike	G22813T	K417N	0	0 %
501.V2	Spike	G23012A	EA84K	2	0.04 %
501.V2	Spike	A23063T	N501Y	0	0 %
501.V2	Spike	C23664T	A701V	1	0.02 %
501.V2	Orf3a	G25563T	Q57H	1174	22.61 %
501.V2	Orf3a	C25904T	S171L	4	0.08 %
501.V2	E	C26456T	P71L	2	0.04 %
501.V2	N	C28887T	T205I	37	0.52 %

Individual signature mutations in wastewater samples

In order to compare the patient-level findings to the mutations detectable in wastewater samples at the community level, we collected a total of 48 wastewater samples between July 9 and December 21, 2020, from three different WWTPs in Switzerland. In Lausanne, a total of 15 samples were taken between November 5 and December 21, in Zurich, 32 samples were collected between July 9 and December 21, and in the ski resort, one sample was taken on December 21. From all samples, viral RNA was extracted, DNA libraries prepared, and subjected to NGS using 2 x 250bp paired-end reads. After quality control, the median number of reads per sample was 1,269,169. The subsequent read mapping step aligned a median of 1,098,352 reads per sample, which resulted in a median per-sample median coverage of 2634.5 reads per position (range 0 - 28,958). The overall deep coverage allowed for calling low-frequency mutations in most samples down to a frequency of about 1/3000 on average.

We first analyzed the wastewater samples collected in Zurich (Figure 1). The 32 samples from July to late December cover a period before the first documented occurrence of the strains and a period afterwards when they might have migrated into Swiss communities. Combining all 32 samples, we find six of the 17 B.1.1.7 signature mutations, with a total of 19 occurrences: a 6bp spike deletion in seven samples, a 9bp Orf1ab deletion in seven samples, a 3bp N gene substitution in two samples, a 3bp spike deletion in one sample, an Orf1ab point mutation in one sample, and a spike point mutation in one sample. When separating the occurrence of mutations before and after October 23 (assuming that the new lineages were not present in Switzerland before that date), we observe three signature mutations each occurring in one of the 12 samples before October 23 versus 16 occurrences of five different mutations in 20 samples after October 23. Association between sampling

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

time and number of signature mutations was assessed by comparing the number of times we found evidence for and against signature mutations between the early and the late group excluding missing data due to low coverage, and did not reach statistical significance (p = 0.093, Fisher's exact test). For the 12 501.V2 signature mutations, we find six occurring 37 times in total (in 21, 7, 4, 2, 1, and 1 samples), 10 before and 27 after October 23, and a significant association of their occurrence before versus after October 23 (p = 0.036). The most prevalent mutation found in every sample after October 23 is the Orf3a mutation G25563T, which is highly prevalent in the Swiss population at 22.6 % according to GISAID (Table 1). Stratification of the GISAID data by sampling time, shows that the prevalence of this mutation increased from 12.6% before October 23 to 34.6% after October 23 in the population, which matches the trend observed in the wastewater data. Hence, the increase in frequency in wastewater samples is likely not related to occurrence of the 501.V2 variant, because among patient samples, few other signature mutations were observed in combination with G25563T, most frequently the pair (C1059T, G25563T) in only 199 out of 5192 (3.8%) consensus sequences.

Figure 1. Occurrence of B.1.1.7 (A) and 501.V2 (B) signature mutations in wastewater samples collected in Zurich over time. Blue color shading encodes the estimated frequency of mutations detected in the sample, pink indicates absence of the mutation, and white indicates missing values (due to lack of sufficient coverage).

As compared to the data from Zurich, the time series data from Lausanne starts later, with one sample from November 5 and 14 daily samples between December 8 and 21 (Figure 2). Overall, we find a larger number of signature mutations in the Lausanne samples compared to the Zurich samples. We observe twelve of the 17 B.1.1.7 signature mutations with a total

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

of 28 observations across the 15 samples. The four B.1.1.7 mutations observed in Zurich are among them. For 501.V2, four out of twelve signature mutations occur, with the Orf3a mutation G25563T being present in all but one sample from Lausanne. For comparison, the same point mutation was found in Zurich in 2 out of 12 samples before and in all 19 samples after October 23 which had coverage at this position.

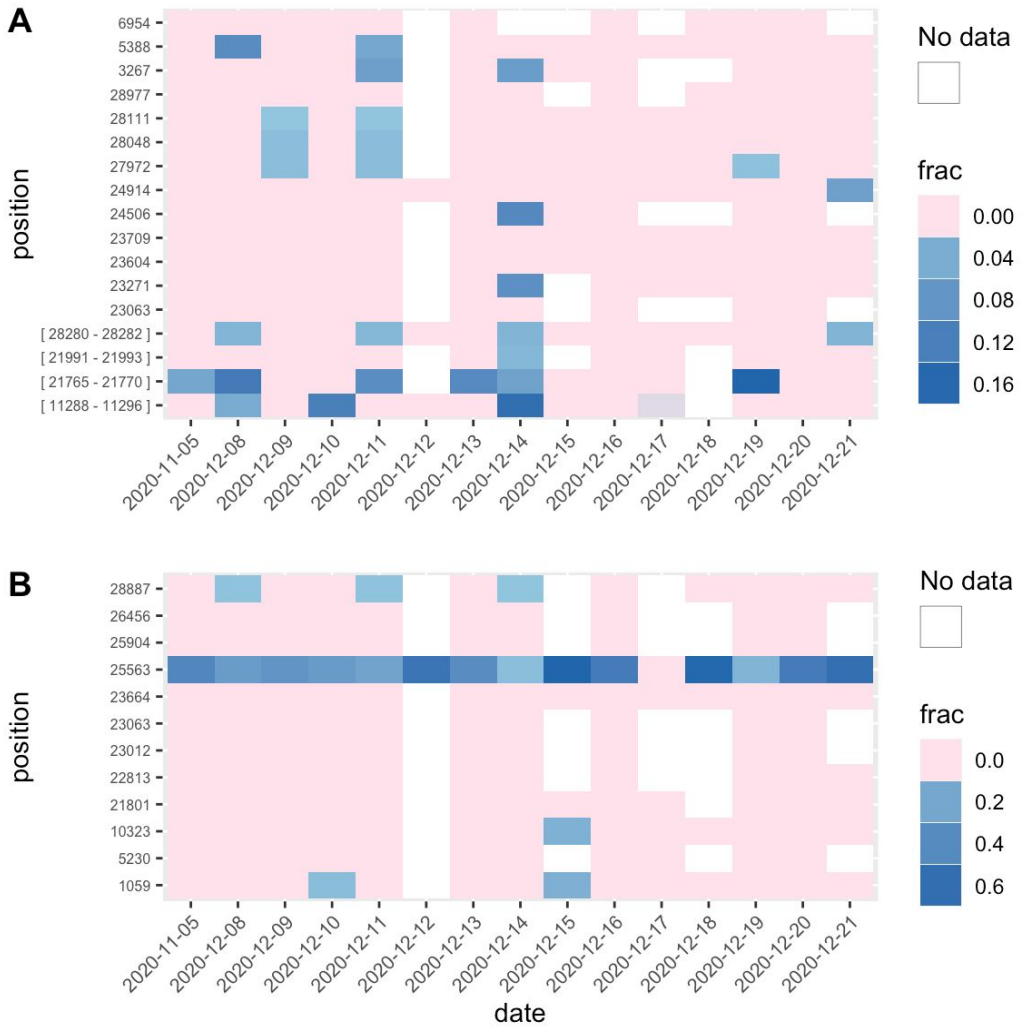

Figure 2. Occurrence of B.1.1.7 (A) and 501.V2 (B) signature mutations in wastewater samples collected in Lausanne over time. Blue color shading encodes the estimated frequency of mutations detected in the sample, pink indicates absence of the mutation, and white indicates missing values (due to lack of sufficient coverage).

Finally, the sample from the ski resort is striking: here, we detected 10 out of the 17 B.1.1.7 mutations and one of the 12 501.V2 mutations (Figure 3). Given the patient background frequencies of individual mutations (Table 1), many of which are zero for the B.1.1.7 variant, and the very few combinations of at most three of them that were observed previously, it seems likely that the B.1.1.7 strain, which was first detected in patient samples from Geneva of December 22, was present in the December 21 wastewater sample.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

In summary, the comparative analysis of individual signature mutations in both patient samples and in community wastewater samples showed their consistency and provided some indication for the presence of B.1.1.7 in a sample from the ski resort of December 21.

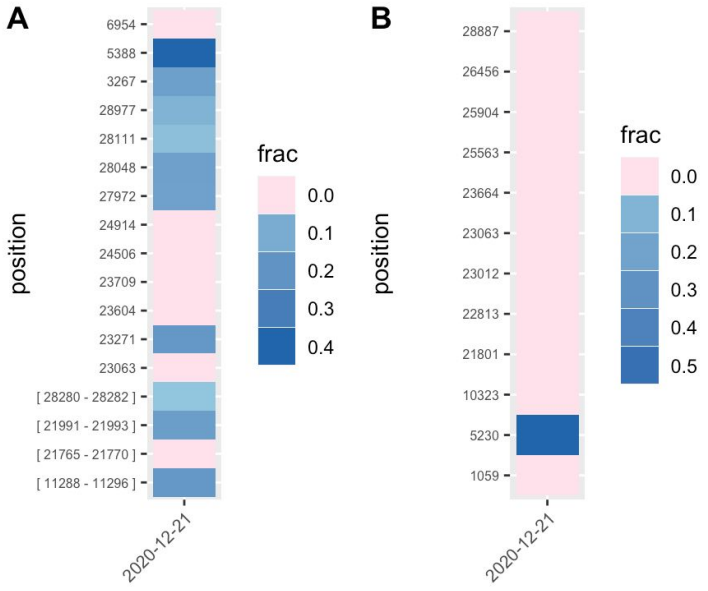

Figure 3. Occurrence of B.1.1.7 (A) and 501.V2 (B) signature mutations in wastewater samples collected in a Swiss ski resort on December 21, 2020. Blue color shading encodes the estimated frequency of mutations detected in the sample and pink indicates absence of the mutation.

Co-occurrence of signature mutations in wastewater samples

Some of the signature mutations analyzed above are located in close proximity to each other on the SARS-CoV-2 genome. In this case, in addition to detecting individual mutations independently, co-occurring mutations can be observed directly on the same sequencing read or read pair coming from the same amplicon produced by the ARTIC v3 protocol. Specifically, we analyzed two regions on amplicons 92 and 93 which contain three and two B.1.1.7-defining mutations, respectively, and two 501.V2-specific mutations on amplicon 76 (Table 2). Co-occurrence of mutations provides higher confidence in the presence of the respective strain, as both independent biological generation and technical artifacts are much less likely to produce such mutational patterns. To validate this approach, we first searched for the patterns in two B.1.1.7-positive and two 501.V2-positive control patient samples and indeed detected co-occurrence of the respective B.1.1.7 and 501.V2 mutations (Table 2, last four rows). In addition, all patient and wastewater samples that we analyzed displayed, as expected, on amplicon 77 the A23403G signature mutation of the B.1 lineage, which is the most prevalent lineage (>99%) in Switzerland at the time of sampling (Alm et al. 2020; Stange et al. 2020).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Table 2. Co-occurrence of signature mutations. For 7 selected samples, genomic regions containing two or more B.1.1.7 or 501.V2 signature mutations observable on individual read pairs were analyzed. The first column indicates the sample date and location, or ID. Columns 2-5 are labeled with the respective genomic region and the signature mutations it contains and show the number of amplicons with all signature mutations present simultaneously, (separated by a dash) the total coverage by single amplicons, and the fraction of co-occurring mutant reads. Column five reports the signature mutation of lineage B.1, as a control. All wastewater samples were included with at least 0.1% of all amplicons and at least 5 read pairs carrying all signature mutations in at least one of the regions. The first 5 rows show wastewater samples labeled by date and location. The last four rows contain patient samples known to derive from the B.1.1.7 and 501.V2 lineages, respectively, as positive controls.

Sample	**Amplicon 92 27809-28144** **C27972T, G28048T, A28111G (B.1.1.7)**	**Amplicon 93 28105-28441** **A28111G GAT28280CTA (B.1.1.7)**	**Amplicon 76 22822-23188** **G23012A, A23063T (501.V2)**	**Amplicon 77 23145-23499** **A23403G (B.1)**
2020-12-21 Ski resort	514 / 3689 13.93%	0 / 20672	0 / 165	36208 / 36209 100.00%
2020-12-21 Lausanne	0 / 10	93 / 3393 2.74%	0 / 0	10 / 10 100.00%
2020-12-14 Lausanne	0 / 4858	816 / 35838 2.28%	0 / 177	20280 / 20284 99.98%
2020-12-11 Lausanne	154 / 13504 1.14%	0 / 82020	0 / 802	93625 / 93659 99.96%
2020-12-09 Lausanne	5 / 457 1.09%	0 / 40213	0 / 76	12846 / 12847 99.99%
Patient sample 410256 (501.V2 positive)	0 / 2601	0 / 3526	8 / 8 100.00%	6570 / 6574 99.94%
Patient sample 410279 (501.V2 positive)	0 / 20487	0 / 16822	156 / 156 100.00%	32633 / 32699 99.80%
Patient sample 420389 (B1.1.7 positive)	389 / 389 100.00%	1498 / 1501 99.80%	0 / 3	3184 / 3184 100.00%
Patient sample 420394 (B1.1.7 positive)	207 / 207 100.00%	739 / 742 99.60%	0 / 7	2067 / 2068 99.95%

In the wastewater sample from the ski resort of December 21, which, as reported above, shows a high prevalence of individual B.1.1.7 mutations, we found co-occurrence of the three B.1.1.7 signature mutations C27972T, G28048T, and A28111G on amplicon 92 in 514 out of 3689 reads (13.93%), but no reads from amplicon 93 with the two B.1.1.7 mutations

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

A28111G and GAT28280CTA. Nevertheless, the large fraction of triple-mutated amplicon-92 reads indicates the presence of the B.1.1.7 lineage, consistent with reports of British tourists staying in the geographic region around the ski resort covered by the WWTP. We analyzed an additional four samples from Lausanne of December 9, 11, 14, and 21 that displayed some of the individual mutations. We found co-occurrences of the following B.1.1.7 signature mutations: two samples carried the signature combination (C27972T, G28048T, A28111G) on 1.09% and 1.14% of reads from amplicon 92, and two samples displayed the signature pair (A28111G, GAT28280CTA) on 2.28% and 2.74% of reads from amplicon 93.

To put the wastewater co-occurrence observations into the context of the background mutation distribution, in the GISAID patient consensus sequences (Table 1), we did not observe any co-occurrence of these signature mutations neither for the B.1.1.7 nor 501.V2 lineages. However, among the 5192 consensus sequences, mutation C27972T alone occurred in a total of 7 (0.13%) samples, 2 of them from the canton Zurich (1: 2020-08-20, 1: 2020-08-21), 4 from Berne (1: 2020-11-23, 1: 2020-12-06, 2: 2020-12-14) and one from Thurgau (2020-11-30), but never in co-occurrence with G28048T or A28111G. The signature mutation G23012A from the 501.V2 lineage only occurred twice in GISAID patient samples, namely in samples from Basel-Land (1: 2020-03-21) and Vaud (1: 2020-09-30). Mutation A23063T was never observed in any of the 5192 Swiss patients. Hence, the co-occurrences found in the wastewater samples (Table 2) cannot be explained by the background mutation distribution. They may have either been generated recently *de novo* or be the result of incomplete sampling of the genomes of B.1.1.7 viruses in the wastewater samples. Given that 5 out of 61 samples (all sampled in mid or late December, at the suspected time of B.1.1.7 occurrence) display these co-occurrence patterns and the high prevalence of individual B.1.1.7 signature mutations in the December-21 ski resort sample, it seems more likely that the B.1.1.7 variant was present in the four samples from Lausanne and the one sample from the ski resort.

Conclusion

The preliminary analysis of deep sequencing data from 48 community wastewater samples collected between July and December 2020 in Switzerland has identified several of the signature mutations that define the B.1.1.7 and 501.V2 lineage in several of the samples. In five samples, one dating as early as December 9, we have found up to three co-occurring B.1.1.7 signature mutations. These co-occurrences have not been observed in any clinical samples and hence any known circulating strains until December 24. This observation may be explained either by an unobserved new strain in Switzerland which is not part of B.1.1.7, or by B.1.1.7 having circulated already in the first part of December in Switzerland. Given that B.1.1.7 was identified first in Switzerland in a sample dated December 22 and none of the non-B.1.1.7 clinical samples carry the co-occurring mutations, the latter explanation appears more likely suggesting that the B.1.1.7 lineage was present in Switzerland already in early December.

We have also observed an increase in the number of individual signature mutations in wastewater samples after October in Zurich, consistent with the same trend in clinical samples. When analyzing mutations independently, only the December-21 ski resort finding

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

with 10 out of 17 B.1.1.7 signature mutations provided some indication for the presence of this strain. By contrast, the co-occurrence analysis appears to be more powerful in detecting the new strains, owing to the genomic location of some of the signature mutations on the same amplicon.

Additional work is required to assess the sensitivity and specificity of both approaches in detecting new genomic variants. Although NGS data from mixed samples can only provide indirect evidence for the presence of a lineage and not a definite proof, wastewater-based SARS-CoV-2 genomics may support epidemiology by providing timely, inexpensive, non-invasive, unbiased community-level samples. Close-mesh time series data are particularly informative, not only for detailed monitoring of circulating lineages, but also for increased accuracy in detecting genomic alterations.

Author Contributions

Conceptualization: NB, TK, TRJ, CO, TS, XFC, IT
Data curation: IT, CC, KPJ
Formal Analysis: KJ, LF, DD
Funding acquisition: NB, TK, TRJ, CO, TS
Investigation: XFC, AK, PG, ES, CB, CA
Methodology: NB, XFC, KJ, DD, IT
Project administration: NB, TK, TRJ, CO, TS
Software: IT, KPJ, LF, DD, KJ
Supervision: NB, TK, TRJ, CO, TS, IT
Validation: NB, IT, KJ, KPJ
Visualization: KJ
Writing – original draft: NB, IT, KJ, DD
Writing – review & editing: NB, TK, TRJ, CO, TS, KJ, KPJ, LF, XFC, DD, CA

Funding

This work was supported by the SIB Swiss Institute of Bioinformatics as a Competitive Resource, the Swiss National Science Foundation (project 31CA30_196538), Eawag discretionary funds and an EPFL COVID19 grant. XFC was a fellow of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie Grant Agreement No. 754462.

Acknowledgements

The supplementary material contains the detailed acknowledgements of the originating and submitting laboratories of the GISAID data.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Bibliography

Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, Melidou A, Neher RA, O'Toole Á, Pereyaslov D, et al. 2020. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill* **25**.

Bundesamt für Gesundheit. 2020a. Covid-19: Neue Variante des Coronavirus in zwei Proben in der Schweiz entdeckt. https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/medienmitteilungen.msg-id-81829.html (Accessed January 4, 2021).

Bundesamt für Gesundheit. 2020b. Covid-19: Zwei Proben mit Variante aus Südafrika entdeckt. https://www.bag.admin.ch/bag/de/home/das-bag/aktuell/medienmitteilungen.msg-id-81830.html (Accessed January 4, 2021).

Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW, Girones R, Wang D, et al. 2011. Raw sewage harbors diverse viral populations. *MBio* **2**.

Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, Lou YC, Flamholz A, Kennedy LC, Greenwald H, Hinkle A, et al. 2020. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *medRxiv* **2020.09.13.20193805**.

Département de la sécurité, de l'emploi et de la santé, Republic et Canton de Geneve. 2021. COVID-19 - Nouvelle souche de coronavirus plus contagieuse provenant du Royaume-Uni. https://www.ge.ch/actualite/covid-19-nouvelle-souche-coronavirus-plus-contagieuse-provenant-du-royaume-uni-2-01-2021 (Accessed January 4, 2021).

Fernandez-Cassi X, Timoneda N, Martínez-Puchol S, Rusiñol M, Rodriguez-Manzano J, Figuerola N, Bofill-Mas S, Abril JF, Girones R. 2018. Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. *Sci Total Environ* **618**: 870–880.

Gonzalez R, Curtis K, Bivins A, Bibby K, Weir MH, Yetka K, Thompson H, Keeling D, Mitchell J, Gonzalez D. 2020. COVID-19 surveillance in Southeastern Virginia using wastewater-based epidemiology. *Water Res* **186**: 116296.

Gupta S, Parker J, Smits S, Underwood J, Dolwani S. 2020. Persistent viral shedding of SARS-CoV-2 in faeces - a rapid review. *Colorectal Dis* **22**: 611–620.

Izquierdo Lara RW, Elsinga G, Heijnen L, Oude Munnink BB, Schapendonk CME, Nieuwenhuijse D, Kon M, Lu L, Aarestrup FM, Lycett S, et al. 2020. Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing. *medRxiv*.

Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. 2020. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *J Clin Virol* **131**: 104585.

Kitajima M, Ahmed W, Bibby K, Carducci A, Gerba CP, Hamilton KA, Haramoto E, Rose JB. 2020. SARS-CoV-2 in wastewater: State of the knowledge and research needs. *Sci Total Environ* **739**: 139076.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Kuipers J, Batavia AA, Jablonski KP, Bayer F, Borgsmüller N, Dondi A, Drăgan M-A, Ferreira P, Jahn K, Lamberti L, et al. 2020. Within-patient genetic diversity of SARS-CoV-2. *BioRxiv*.

Martin J, Klapsa D, Wilton T, Zambon M, Bentley E, Bujaki E, Fritzsche M, Mate R, Majumdar M. 2020. Tracking SARS-CoV-2 in Sewage: Evidence of Changes in Virus Variant Predominance during COVID-19 Pandemic. *Viruses* **12**.

Martínez-Puchol S, Rusiñol M, Fernández-Cassi X, Timoneda N, Itarte M, Andrés C, Antón A, Abril JF, Girones R, Bofill-Mas S. 2020. Characterisation of the sewage virome: comparison of NGS tools and occurrence of significant pathogens. *Sci Total Environ* **713**: 136604.

McElroy K, Zagordi O, Bull R, Luciani F, Beerenwinkel N. 2013. Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. *BMC Genomics* **14**: 501.

Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. 2020. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environ Sci Technol Lett*.

Nadeau S, Beckmann C, Topolsky I, Vaughan T, Hodcroft E, Schaer T, Nissen I, Santacroce N, Burcklen E, Ferreira P, et al. 2020. Quantifying SARS-CoV-2 spread in Switzerland based on genomic sequencing data. *medRxiv*.

Nemudryi A, Nemudraia A, Wiegand T, Surya K, Buyukyoruk M, Vanderwood KK, Wilkinson R, Wiedenheft B. 2020. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. *medRxiv*.

Posada-Céspedes S, Seifert D, Topolsky I, Metzner KJ, Beerenwinkel N. 2020. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput sequencing data. *BioRxiv*.

R&D DP. 2020. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol.

Rose R, Nolan DJ, Moot S, Feehan A, Cross S, Garcia-Diaz J, Lamers SL. 2020. Intra-host site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. *medRxiv*.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**: 30494.

Stange M, Mari A, Roloff T, Seth-Smith HMB, Schweitzer M, Brunner M, Leuzinger K, Søgaard KK, Gensch A, Tschudin-Sutter S, et al. 2020. SARS-CoV-2 phylogeny during the early outbreak in the Basel area, Switzerland: import and spread dominated by a single B.1 lineage variant (C15324T). *medRxiv*.

Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, et al. 2020. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*.

van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. 2020. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun* **11**: 5986.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.08.21249379; this version posted January 9, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* **12**: 119.

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. 2010a. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* **17**: 417–428.

Zagordi O, Klein R, Däumer M, Beerenwinkel N. 2010b. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* **38**: 7400–7409.