Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater

Artem Nemudryi[1], Anna Nemudraia[1], Kevin Surya, Tanner Wiegand, Murat Buyukyoruk, Royce Wilkinson, and Blake Wiedenheft[*]

Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717, USA

[1]These authors contributed equally.

Correspondence: bwiedenheft@gmail.com

ABSTRACT

SARS-CoV-2 has recently been detected in feces, which indicates that wastewater may be used to monitor viral prevalence in the community. Here we use two different sampling methods to monitor SARS-CoV-2 in wastewater over a 17-day period and sequencing is used to infer viral ancestry. While SARS-CoV-2 is detected over the entire time course, viral RNA has been steadily decreasing over the last week, suggesting that state mandated social isolation is having a measurable impact on containment of the outbreak.

In late December of 2019, authorities from the Peoples Republic of China (PRC) announced an epidemic of pneumonia (WHO, 2019). A novel coronavirus (Severe Acute Respiratory Syndrome coronavirus 2, SARS-CoV-2) was identified as the etiologic agent and the disease has been named “coronavirus disease 2019” (COVID-19). The virus spread rapidly, first to Thailand, Japan, Korea, and Europe, and now to

over 181 countries across all continents except Antarctica. The global total of infected individuals now exceeds 1.8 million (https://coronavirus.jhu.edu/).

Public health professionals around the world are working to limit the spread of SARS-CoV-2, and “flatten the curve”, which requires a reduction in cases from one day to the next. However, SARS-CoV-2 containment has been outpaced by viral spread and limited resources for testing. Moreover, mounting evidence suggests that the virus is not only spread by aerosols, but may also be transmitted via feces (Hindson, 2020; Lodder and de Roda Husman, 2020; Tang et al., 2020; Wang et al., 2020; Wu et al., 2020b; Xu, 2020). This has important implications for the spread of the virus and suggests that wastewater may be used to monitor progression or abetment of viral spread at the community level (Mallapaty, 2020). To test this hypothesis, we collected three 0.5-liter samples of pre-treated wastewater from the municipal wastewater treatment plant in Bozeman, Montana (USA). Samples were collected on 7 different days over the course of a 17-day period, using two different collection methods. The samples were filtered and concentrated prior to RNA extraction. Isolated RNA was used as a template for one-step reverse transcription quantitative polymerase chain reaction (RT-qPCR) according to CDC recommended guidelines (https://www.fda.gov/media/134922/download). Each RT-qPCR reaction was performed using two primer pairs (i.e., N1, and N2), each targeting distinct regions of the nucleocapsid (N) gene from SARS-CoV-2 (**Fig 1 and Supplemental Fig 1**). The first two samples were collected manually, using a sampling stick on the mornings of March 23rd and 27th, respectively. While samples collected on these two occasions all tested positive for SARS-CoV-2, the estimated

titers varied considerably between biological replicates (**Fig 1A**). We hypothesized that this variability was due to the sampling method, so subsequent sampling was performed using an autosampler that collects a volume proportional to flow, every hour for 24-hours. This 24-hour composite sample was then divided into three 0.5-liter aliquots. Like the previous samples, these samples also tested positive for SARS-CoV-2 (**Fig 1B**). Viral titers in composite samples were lower than from previous samples collected manually, but there was considerably less variation between replicates. This indicates that the previously noted variability was due to the sampling method, rather than inconsistencies associated with RNA extraction or the RT-qPCR assay.

To verify the RT-qPCR results and determine the phylogenetic origin of SARS-CoV-2 strains circulating in the Bozeman waste stream, we used 10 primer pairs that tile across the SARS-CoV-2 genome to PCR amplify phylogenetically informative regions (Quick, 2020). These primers were designed to target conserved regions of the SARS-CoV-2 genome that flank polymorphic sites that have been used to trace viral ancestry and geographic origins (Quick, 2020) (**Fig 1C and Supplemental Fig 2**). RNA isolated from the Bozeman waste stream on March 27th was used as a template for RT-PCR. All 10 primer pairs produced PCR products of the expected sizes (**Fig 1C**). PCR products were sequenced using Sanger methods and the reads were aligned to the reference genome using MUSCLE (Edgar, 2004; Wu et al., 2020a). We observed no sequence heterogeneity in redundant reads derived from each location of the genome, which suggests the predominance of a single SARS-CoV-2 genotype in the waste stream at the time of sampling.

To determine viral ancestry, we aligned sequences isolated from the Bozeman wastewater to 2,399 SARS-CoV-2 genomes that have been determined from 59 different countries (Hadfield et al., 2018). While nine of the thirteen positions analyzed in our sample showed no mutations relative to the Wuhan-Hu-1/2019 sequence (**Fig 2C** and **Supplemental Fig 2**), four mutations are characteristic of sequences from Europe (**Fig 2A and Fig 2C**). One mutation in particular (G25563T) predominates in the SARS-CoV-2 isolates from France, but is absent in the majority of samples from any other country (**Fig 2A**). The conservation of these four mutations suggests a shared phylogenetic history between virus in the Bozeman wastewater and SARS-CoV-2 that is most recently derived from Europe.

In addition to position-specific mutation analysis, we also concatenated each of the ten sequences generated by PCR and then performed a phylogenetic analysis. Similar to the position-specific analysis, this approach indicates that the predominant SARS-CoV-2 strains in Bozeman's wastewater are most closely related to those from Europe, specifically strains from France and Iceland (**Fig 2B and C**). By scaling the phylogenetic data according to viral isolation dates, Hadfield *et al* previously showed that the G25563T mutation, which predominates specifically in France, arose more recently than the other mutations (Hadfield et al., 2018). This suggests that the predominate strain currently circulating in Bozeman may have originated in France. While our sequencing approach was designed to target phylogenetically informative positions, additional

sequencing is necessary to establish a more comprehensive understanding of SARS-CoV-2 diversity in the community.

Collectively, the results presented here demonstrate that municipal wastewater can be used to monitor the prevalence of SARS-CoV-2 in the community over time. Our data indicated that the sampling method is critical and that composite samples may provide the most reliable daily average of viral concentration in wastewater over time. Bozeman is a relatively small community (~50,000) and SARS-Cov-2 was not detected in Montana until March 12nd, 2020. While the time course presented here is relatively short, we note a statistically significant increase in SARS-Cov-2 RNA between the first (March 30th) and the second composite sample tested (April 1st, 2020), and then a gradual reduction of SARS-CoV-2 RNA over the next three time points (i.e., April 3rd, 6th and 8th, 2020). While temporal sampling is ongoing, the data collected thus far, reveal an encouraging trend, which suggests that community prevalence is currently on the decline.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (1R35GM134867), the Montana State University Agricultural Experimental Station, the MJ Murdock Charitable Trust, the Gianforte Foundation and the MSU Office of the Vice President for Research. We are grateful to Josh French, Justin Roberts, and the other dedicated wastewater technicians that made this work possible. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database (see Acknowledgement Table).

Declaration of interests

B.W. is the founder of SurGene, LLC, and is an inventor on patent applications related to CRISPR-Cas systems and applications thereof.

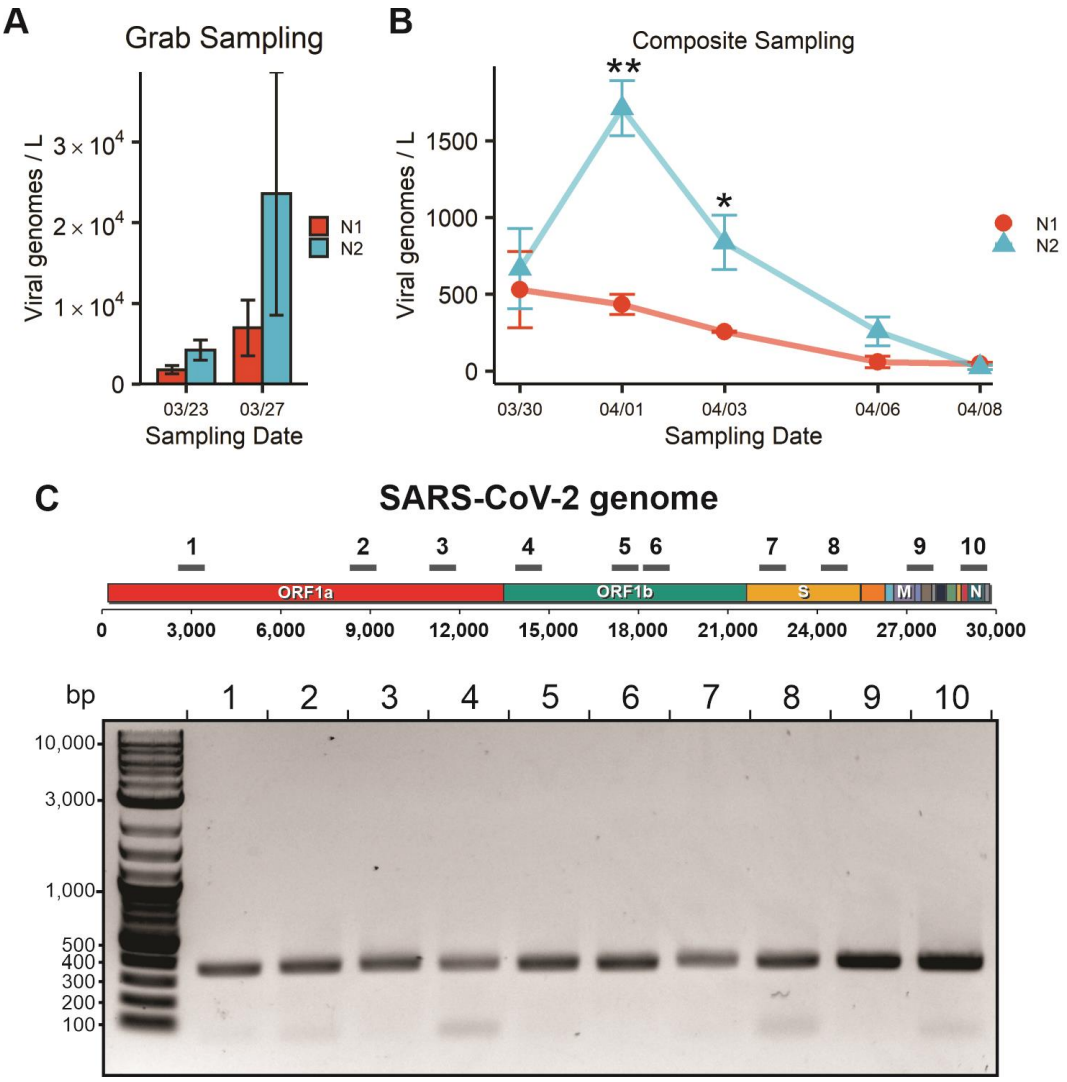


Fig 1. Detection and quantification of SARS-CoV-2 in wastewater. **A)** Three 0.5-liter samples were collected manually or **B)** subsampled from a 24-hour composite. SARS-CoV-2 RT-qPCR tests were performed according to CDC guidelines and protocols. Detection includes two primer pairs, each targeting distinct regions of the nucleocapsid (N) gene from SARS-CoV-2 (i.e., N1, and N2). Statistical analysis was conducted using ANOVA with Tukey`s post-hoc test; * $p < 0.05$, ** $p < 0.01$, compared to the previous sampling day. **C)** Map of the SARS-CoV-2 genome. The solid black lines indicate the approximate location of 10 primers pairs designed to amplify phylogenetically informative regions of the genome. Agarose gel of the corresponding PCR products.

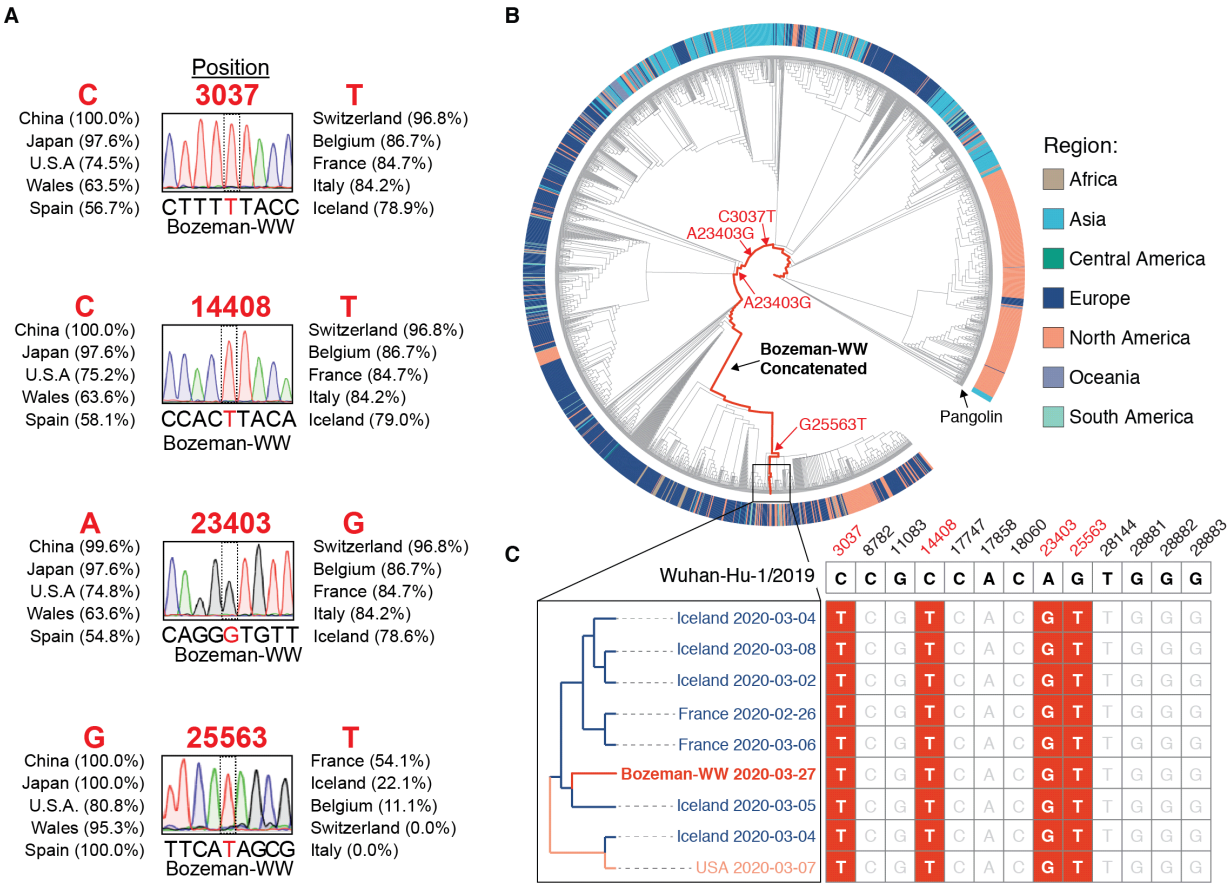


Fig 2. Phylogenetic analysis of SARS-CoV-2 sequences from wastewater. A) Sequence, position and geographic prevalence of mutations that have been previously used for mapping phylogenetic history. Mutations and positions shown are relative to the Wuhan-Hu-1/2019 sequence. Chromatograms are shown for SARS-CoV-2 sequences isolated from Bozeman wastewater. **B)** Maximum-likelihood phylogeny of the SARS-CoV-2 related lineage (n = 2,433 sequences). Outer ring colored according to the world region where samples were isolated. Tree is rooted relative to SARS-CoV-2 genome sequences isolated from pangolins. Points where mutations from panel A likely occurred in ancestral history are shown in red. **C)** Inset from panel B (left). Sequences isolated from Bozeman wastewater clade predominantly with sequences of French and Icelandic origin. Sequences are named according to geographic origin and viral isolation date. The Wuhan reference sequences for each of the 13 phylogenetically informative positions are shown across the top. The positions shown in red are deviations from the Wuhan reference that are conserved in viruses isolated from specified regions of the globe.

METHODS

Wastewater sampling and RNA extraction

Pre-treated wastewater samples were collected from influent at the Bozeman Water Reclamation Facility (WRF). Grab samples were collected in triplicates with 15 second intervals from the wastewater stream. Composite samples were collected with automatic flow proportional sampler to capture average characteristics of wastewater over a period of 24 hours. Each wastewater sample (500 mL) was sequentially filtered through 20 µM, 5 µM (Sartorius Biolab Products) and 0.45 µM (Pall Corporation) membrane filters and concentrated down to 150-200 µL using Corning Spin-X UF concentrators with 100 kDa molecular weight cut-off. Total RNA from concentrated samples was extracted with RNeasy Mini Kit (Qiagen) with 40 µL elution volume and used as a template for RT-qPCR.

Reverse Transcription quantitative PCR (RT-qPCR)

RT-qPCR was performed using two primers pairs (N1 and N2) and probes from 2019-nCoV CDC EUA Kit (IDT#10006606). SARS-CoV-2 in wastewater was detected and quantified using one-step RT-qPCR in ABI 7500 Fast Real-Time PCR System according to CDC guidelines and protocols (https://www.fda.gov/media/134922/download). 20 µL reactions included 8.5 µL of Nuclease-free Water, 1.5 µL of Primer and Probe mix, 5 µL of TaqPath 1-Step RT-qPCR Master Mix (ThermoFisher, A15299) and 5 µL of the template. Nuclease-free water was used as negative template control (NTC). Amplification was performed as follows: 25°C for 2 min, 50°C for 15 min, 95 °C for 2 min followed by 45 cycles of 95 °C for 3 s and 55 °C for 30 s. To quantify viral genome copy

numbers in the samples, standard curves for N1 and N2 were generated using a dilution series of a positive template control (PTC) plasmid (IDT#10006625) with concentrations ranging from 10 to 10,000 copies per reaction. Three technical replicates were performed at each dilution. The detection limit was down to 10 copies of the control plasmid. Run data was analyzed in SDS software v1.4 (Applied Biosystems). Threshold cycle (Ct) values were determined by manually adjusting threshold to fall within exponential phase of the fluorescence curves and above any background signal. Ct values of PTC dilutions were plotted against $\log_{10}$(copy number) to generate standard curves. Linear regression analysis was performed in RStudio v1.2.1335 and the trend line equation (Ct = [*slope*] × [$\log_{10}$(copy number)] + b) was used to calculate copy numbers from mean Ct values of technical replicates for each biological replicate. Primer efficiencies calculated as $E = (10^{(-1/[slope])} - 1) \times 100\%$ were 145.55 ± 6.76% for N1 and 116.58 ± 15.62% for N2 (n = 5, mean ± sd).

RT-PCR and Sanger sequencing

Reverse transcription was performed with 10 µL of RNA from SARS-CoV-2 positive grab wastewater samples using SuperScript™ III Reverse Transcriptase (Thermo Fisher Scientific) according to the supplier's instructions. Resulting cDNA was used as a template to amplify regions of SARS-CoV-2 genome with following previously published primer pairs (**Supplemental Table 1**) (dx.doi.org/10.17504/protocols.io.bbmuik6w) .

PCR reactions were performed using Q5 High-Fidelity DNA Polymerase (New England Biolabs) with following thermocycling conditions: 98°C for 2min, 35 cycles of 98°C for 15

s and 65°C for 5 min. PCR products were analyzed on 1 % agarose gels stained with SYBR Safe (Thermo Fisher Scientific). PCR products were purified using DNA Clean & Concentrator™ kit (Zymo Research) and sent to Psomagen for Sanger sequencing. Each PCR product was sequenced with both forward and reverse primer used for PCR.

Single Nucleotide Polymorphisms (SNP) and Phylogenetic Analysis

SNP analysis was conducted by aligning thirty-three Sanger reads to the genome of Wuhan-Hu-1/2019 (Wu et al., 2020a) and comparing nucleotides at positions that have previously been used to trace ancestry (Hadfield et al., 2018). SNP abundances for SARS-CoV-2 genomes from each country were calculated in R using the BioStrings package (Pagès et al., 2019), and chromatograms of Sanger sequencing reads were rendered in SnapGene (GSL Biotech; available at snapgene.com).

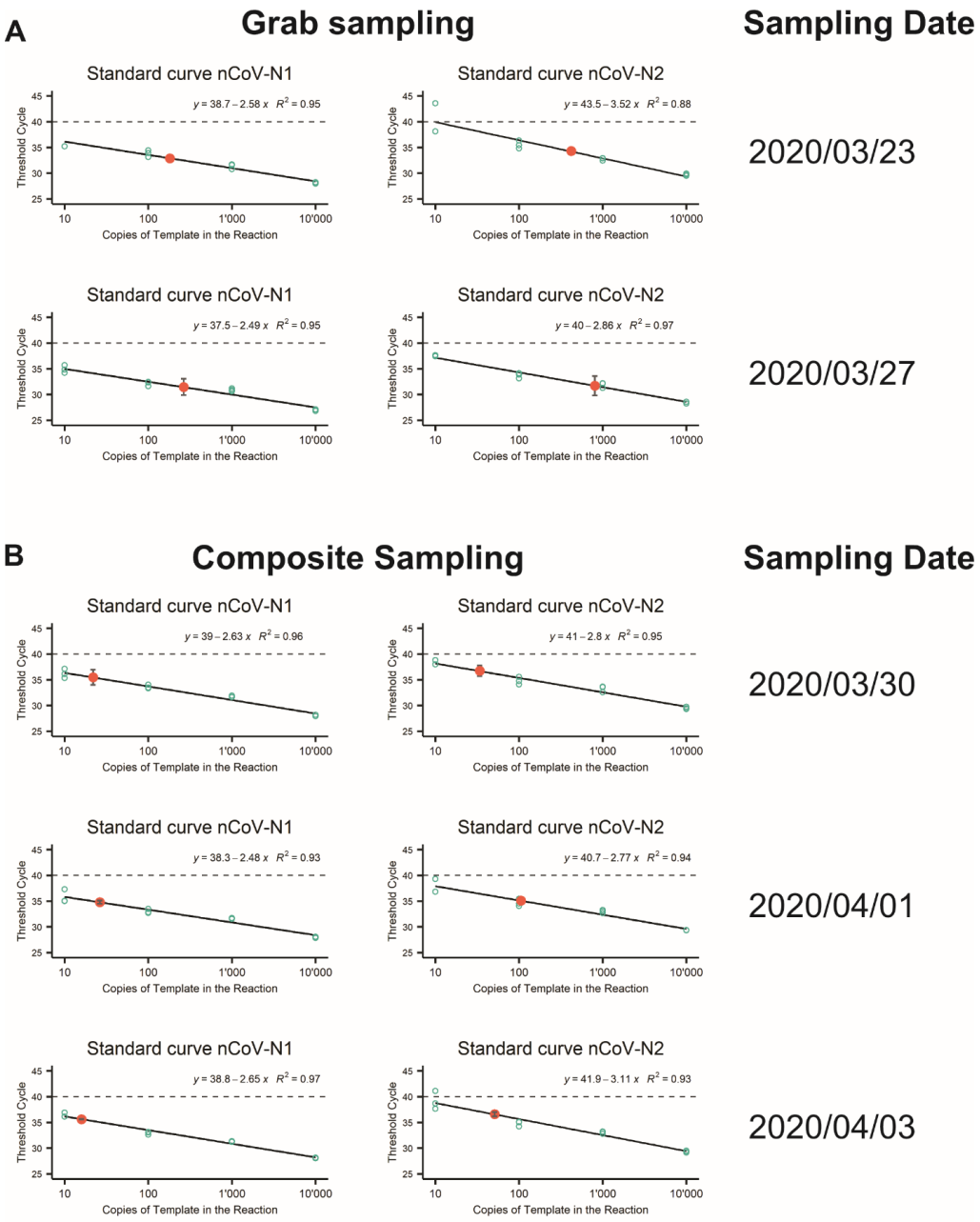
Phylogenetic analysis was performed by concatenating Sanger reads covering each of the thirteen positions of interest. This concatemer, and its component thirty three sequences, were aligned to 2,550 SARS-CoV-2 genomes retrieved from GISAID on 3/29/2020, 1:25:22 PM (https://www.gisaid.org/), using the FFT-NS-2 setting in MAFFT v7.429 (Katoh et al., 2019; Shu and McCauley, 2017). The alignment was optimized using MaxAlign v1.1, which removed 150 gapped sequences (Gouveia-Oliveira et al., 2007). Columns composed of more than 50% gaps were removed with trimAl v1.2rev59 (Capella-Gutierrez et al., 2009).

A maximum-likelihood phylogenetic tree was constructed from this alignment using IQ-Tree in the Augur utility of Nextstrain (Hadfield et al., 2018; Minh et al., 2020). Metadata obtained from GISAID was used to refine the tree in TreeTime (Augur) (Sagulenko et al., 2018; https://www.gisaid.org/). The APE v5.3 package in R was used to re-root the tree relative to pangolin sequences (Paradis and Schliep, 2019), and the tree was plotted using ggtree v3.10 package in R (Yu et al., 2017). The subtree, visualized in Figure 2C, was rendered in FigTree v1.4.4 (Rambaut, 2017).

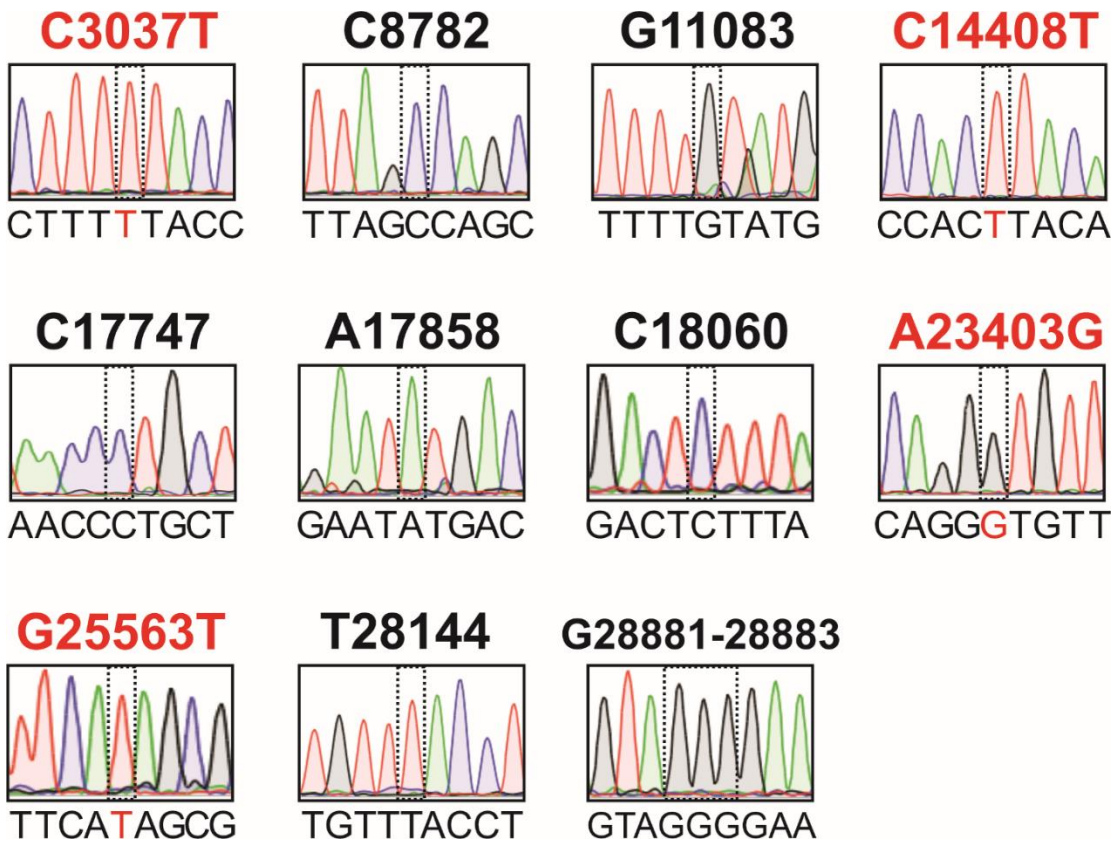
Quantification and Statistical Analyses

Data in figures are shown as mean of three biological replicates (each with two technical replicates) ± standard error of mean (sem). Estimated copy numbers in RT-qPCR reactions were used to calculate titers per liter of wastewater for each biological replicate. One-way analysis of variance (ANOVA) with Tukey`s post-hoc test used to determine statistical significance. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. All statistical analyses were performed in RStudio v1.2.1335.

Supplementary Data:



Supplemental Figure S1. Standard curves for absolute quantification of SARS-CoV-2 titers. Standard curves were generated for each RT-qPCR run. A 10-fold dilution series of positive control plasmid (IDT#10006625) encoding for SARS-CoV-2 *N* gene were used, three technical replicates at each dilution (green circles). Data was plotted as Threshold Cycle (Ct) on y-axis versus $\log_{10}$(copy number) on x-axis. Trend lines were fitted to the data by linear regression analysis in RStudio v1.2.1335, Linear equations and R^2 values are shown for each standard curve. Red dots show mean Ct values for wastewater samples. Error bars represent standard deviation of Ct values.



Supplemental Figure S2. Sanger sequencing of 13 polymorphic sites in SARS-CoV-2 genome. Ten primer pairs were used to amplify and sequence 13 polymorphic sites in SARS-CoV-2 genome (Hadfield et al., 2018; Tang et al., 2020). Sanger traces (9-bp windows) are shown for each polymorphic site (dotted boxes). Nucleotide substitutions are highlighted in red. Nucleotide numbering is based on Wuhan-Hu-1/2019 sequence as a reference (Wu et al., 2020a).

Table S1. Primer design

PCR Product	Primer	Sequence (5'-3')
1	nCoV-2019_10_LEFT	TGAGAAGTGCTCTGCCTATACAGT
	nCoV-2019_10_RIGHT	TCATCTAACCAATCTTCTTCTTGCTCT
2	nCoV-2019_29_LEFT	ACTTGTGTTCCTTTTTGTTGCTGC
	nCoV-2019_29_RIGHT	AGTGTACTCTATAAGTTTTGATGGTGTGT
3	nCoV-2019_37_LEFT	ACACACCACTGGTTGTTACTCAC
	nCoV-2019_37_RIGHT	GTCCACACTCTCCTAGCACCAT
4	nCoV-2019_48_LEFT	TGTTGACACTGACTTAACAAAGCCT
	nCoV-2019_48_RIGHT	TAGATTACCAGAAGCAGCGTGC
5	nCoV-2019_59_LEFT	TCACGCATGATGTTTCATCTGCA
	nCoV-2019_59_RIGHT	AAGAGTCCTGTTACATTTTCAGCTTG
6	nCoV-2019_60_LEFT	TGATAGAGACCTTTATGACAAGTTGCA
	nCoV-2019_60_RIGHT	GGTACCAACAGCTTCTCTAGTAGC
7	nCoV-2019_77_LEFT	CCAGCAACTGTTTGTGGACCTA
	nCoV-2019_77_RIGHT	CAGCCCCTATTAAACAGCCTGC
8	nCoV-2019_84_LEFT	TGCTGTAGTTGTCTCAAGGGCT
	nCoV-2019_84_RIGHT	AGGTGTGAGTAAACTGTTACAAACAAC
9	nCoV-2019_93_LEFT	TGAGGCTGGTTCTAAATCACCCA
	nCoV-2019_93_RIGHT	AGGTCTTCCTTGCCATGTTGAG
10	nCoV-2019_95_LEFT	TGAGGGAGCCTTGAATACACCA
	nCoV-2019_95_RIGHT	CAGTACGTTTTTGCCGAGGCTT

References

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.
Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.
Gouveia-Oliveira, R., Sackett, P.W., and Pedersen, A.G. (2007). MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics *8*, 312.
Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121-4123.
Hindson, J. (2020). COVID-19: faecal-oral transmission? Nat Rev Gastroenterol Hepatol.
Katoh, K., Rozewicki, J., and Yamada, K.D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform *20*, 1160-1166.
Lodder, W., and de Roda Husman, A.M. (2020). SARS-CoV-2 in wastewater: potential health risk, but also data source. Lancet Gastroenterol Hepatol.
Mallapaty, S. (2020). How sewage could reveal true scale of coronavirus outbreak. Nature.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol.
Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings.
Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526-528.
Quick, J. (2020). nCoV-2019 sequencing protocol. protocolsio.
Rambaut, A. (2017). FigTree-version 1.4.4, a graphical viewer of phylogenetic trees.
Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol *4*, vex042.
Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill *22*.
Tang, A., Tong, Z.D., Wang, H.L., Dai, Y.X., Li, K.F., Liu, J.N., Wu, W.J., Yuan, C., Yu, M.L., Li, P.*, et al.* (2020). Detection of Novel Coronavirus by RT-PCR in Stool Specimen from Asymptomatic Child, China. Emerg Infect Dis *26*.
Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., and Tan, W. (2020). Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA.
WHO (2019). SITUATION REPORT - 1.
Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y.*, et al.* (2020a). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265-269.
Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., Yin, H., Xiao, Q., Tang, Y., Qu, X.*, et al.* (2020b). Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol Hepatol.
Xu, Y.e.a. (2020). Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat Med.

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol *8*, 28–36.