

# 1 Variant abundance estimation for SARS-CoV-2 in 2 wastewater using RNA-Seq quantification

3 Jasmijn A. Baaijens<sup>\*,1</sup>, Alessandro Zulli<sup>\*,2</sup>, Isabel M. Ott<sup>\*,3</sup>, Mary E. Petrone<sup>3</sup>, Tara Alpert<sup>3</sup>, Joseph R.  
4 Fauver<sup>3</sup>, Chaney C. Kalinich<sup>3</sup>, Chantal B.F. Vogels<sup>3</sup>, Mallery I. Breban<sup>3</sup>, Claire Duvallet<sup>4</sup>, Kyle McElroy<sup>4</sup>,  
5 Newsha Ghaeli<sup>4</sup>, Maxim Imakaev<sup>4</sup>, Malaika Mckenzie-Bennett<sup>5</sup>, Keith Robison<sup>5</sup>, Alex Plocik<sup>5</sup>, Rebecca  
6 Schilling<sup>5</sup>, Martha Pierson<sup>5</sup>, Rebecca Littlefield<sup>5</sup>, Michelle Spencer<sup>5</sup>, Birgitte B. Simen<sup>5</sup>, Yale SARS-CoV-2  
7 Genomic Surveillance Initiative, William P. Hanage<sup>6</sup>, Nathan D. Grubaugh<sup>†,3,7</sup>, Jordan Peccia<sup>†,2</sup>, Michael  
8 Baym<sup>†,1</sup>

9  
10 <sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

11 <sup>2</sup>Department of Chemical and Environmental Engineering, Yale University, New Haven, CT, USA

12 <sup>3</sup>Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

13 <sup>4</sup>Biobot Analytics, Inc., Cambridge, MA, USA

14 <sup>5</sup>Ginkgo Bioworks, Inc., Boston, MA, USA

15 <sup>6</sup>Center for Communicable Disease Dynamics and Department of Epidemiology, Harvard T.H. Chan  
16 School of Public Health, Boston, MA, USA

17 <sup>7</sup>Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

18 \*Denotes equal contributions

19 †Denotes co-senior authorship

20

21 Correspondence should be addressed to: [j.a.baaijens@tudelft.nl](mailto:j.a.baaijens@tudelft.nl)

22

## 23 Abstract

24 Effectively monitoring the spread of SARS-CoV-2 variants is essential to efforts to counter the ongoing  
25 pandemic. Wastewater monitoring of SARS-CoV-2 RNA has proven an effective and efficient technique  
26 to approximate COVID-19 case rates in the population. Predicting variant abundances from wastewater,  
27 however, is technically challenging. Here we show that by sequencing SARS-CoV-2 RNA in wastewater  
28 and applying computational techniques initially used for RNA-Seq quantification, we can estimate the  
29 abundance of variants in wastewater samples. We show by sequencing samples from wastewater and  
30 clinical isolates in Connecticut U.S.A. between January and April 2021 that the temporal dynamics of  
31 variant strains broadly correspond. We further show that this technique can be used with other  
32 wastewater sequencing techniques by expanding to samples taken across the United States in a similar  
33 timeframe. We find high variability in signal among individual samples, and limited ability to detect the  
34 presence of variants with clinical frequencies <10%; nevertheless, the overall trends match what we  
35 observed from sequencing clinical samples. Thus, while clinical sequencing remains a more sensitive  
36 technique for population surveillance, wastewater sequencing can be used to monitor trends in variant  
37 prevalence in situations where clinical sequencing is unavailable or impractical.

38

## 39 Introduction

40 As the SARS-CoV-2 pandemic continues, the virus is evolving in real time, challenging existing control  
41 measures. Increased infectivity, and potentially increased morbidity and immune evasion, have been  
42 observed in emerging *variant* lineages<sup>1,2</sup>. These variants are characterized by combinations of mutations  
43 compared to the original strain, some of which are likely to be selective adaptations of the SARS-CoV-2  
44 virus<sup>1</sup>. To adapt our approach to the pandemic with the virus, we need first to be able to observe which  
45 variants are present where, and critically how the rates of variants are changing in the population.

46 Genomic surveillance of SARS-CoV-2 enables early detection and clinical investigation of emerging  
47 variants. In late 2020, the Centers for Disease Control and Prevention (CDC) designated specific viral  
48 lineages as *variants of interest* or *variants of concern* based on potential changes in detectability,  
49 transmissibility, disease severity, therapeutic efficacy, and/or ability to evade control by natural or  
50 vaccine-induced immune responses<sup>3</sup>. Initially, this designation included lineage B.1.1.7 (corresponding to  
51 WHO designation *Alpha*), B.1.351 (*Beta*), and P.1 (*Gamma*). In early 2021, CDC added two new variants  
52 to this list: B.1.427 and B.1.429 (*Epsilon*), both of which were first identified in California<sup>4,5</sup>. B.1.617.2 and  
53 related sublineages (*Delta*) now pose a threat but were not observed at substantial rates in the United  
54 States until May 2021. Each variant is identified by a set of potentially overlapping amino acid mutations,  
55 which can be identified by genome sequencing. This is typically done by sequencing remnant clinical  
56 samples used for diagnostics (e.g., nasal swabs), but as infected patients excrete high levels of SARS-  
57 CoV-2 RNA, variant prevalences are potentially detectable from domestic wastewater.

58 Measuring the concentration of SARS-CoV-2 in domestic wastewater can be an efficient method for  
59 indicating infection dynamics in a population<sup>6</sup>. SARS-CoV-2 and fragments of its RNA genome are  
60 excreted by infected individuals through feces or urine<sup>7</sup> and collected in domestic wastewater. Viral RNA  
61 in wastewater can then be extracted and quantified via quantitative RT-qPCR. This approach has been  
62 used to measure SARS-CoV-2 abundance over time, across different regions<sup>8</sup> and wastewater RNA  
63 concentrations are correlated with COVID-19 case rates<sup>9,10</sup>. The genomes of SARS-CoV-2 in wastewater  
64 can also be sequenced, which can then be used to identify mutations present in an entire community with  
65 respect to the reference genome<sup>11,12</sup>. Genome analysis from wastewater sequencing is particularly  
66 challenging because of the low concentrations and poor, fragmented quality of RNA, and the presence of  
67 PCR inhibiting compounds which can interfere with library preparation in wastewater. This typically yields  
68 poor quality sequencing data where the sequencing depth is highly variable across the SARS-CoV-2  
69 genome and overall genome coverage is often incomplete. Despite these challenges, recent work has  
70 shown that it is feasible to observe mutations in wastewater sequencing data<sup>11,12</sup>, and suggests the  
71 possibility of monitoring the abundance of specific lineages. Throughout the world, SARS-CoV-2  
72 wastewater surveillance has been conducted for wastewater collection systems that serve populations

73 ranging from 10,000 to greater than 100,000 people<sup>13</sup>. A method for the quantitative measurement of  
74 variants in wastewater would provide a cost- and resource-efficient approach to population genome  
75 surveillance.

76 Here we introduce a technique to monitor for SARS-CoV-2 variants in a population by sequencing directly  
77 from wastewater and predicting abundances via a computational approach previously used for RNA-seq  
78 transcript quantification. We demonstrate the efficacy of this approach on wastewater data collected from  
79 Connecticut between January and April 2021, during the third wave of the SARS-CoV-2 pandemic, and  
80 compare these predictions to clinically observed variant frequencies from the same geographic area and  
81 time period. We then show the generality of the approach by expanding our analysis to samples collected  
82 across the United States from late December 2020 to January 2021.

83

## 84 Results

### 85 **Prediction of variant abundance is computationally analogous to RNA transcript abundance** 86 **estimation**

87 SARS-CoV-2 RNA fragments in wastewater originate from different infections, with potentially different  
88 viral lineages, pooled together into a single sample. After successfully extracting RNA from wastewater  
89 and sequencing SARS-CoV-2 genome fragments, the computational challenge is to assign reads to  
90 lineages and estimate relative abundance per lineage. This is analogous to RNA transcript quantification  
91 from RNA-Seq data, where the sequencing data consists of reads originating from different transcripts of  
92 a given gene, and the objective is to quantify the relative abundance per transcript (**Fig 1a**).

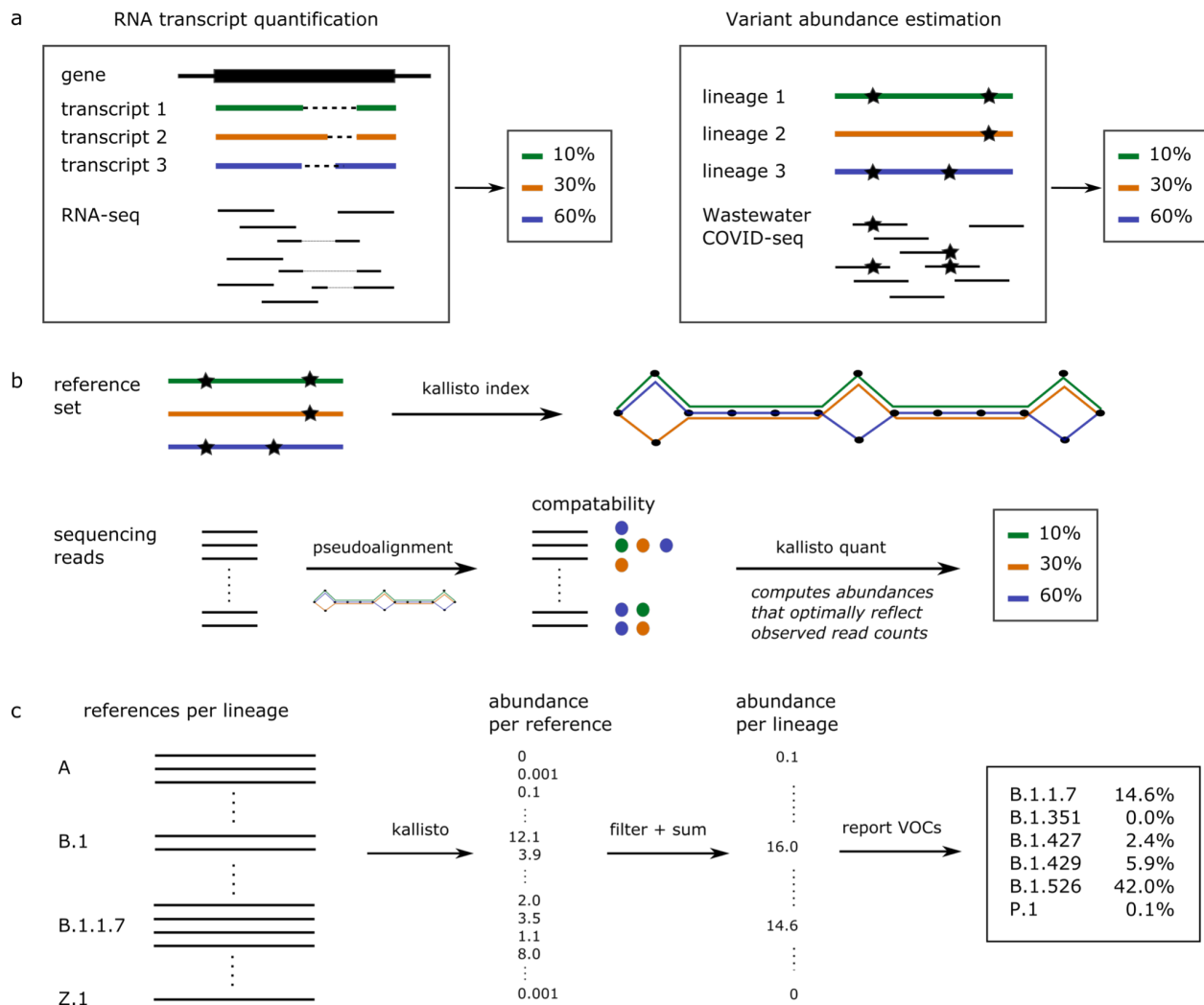
93

94 While pipelines for viral variant detection in clinical samples rely on individual mutation frequencies<sup>14</sup>  
95 using popular tools like V-pipe<sup>15</sup> or iVar<sup>16</sup>, RNA transcript quantification algorithms make use of the  
96 genome sequences without identifying specific mutations. This is a major advantage because identifying  
97 individual mutation frequencies from wastewater sequencing data is highly error-prone---any errors would  
98 be propagated into variant abundance estimates. Moreover, the fact that RNA transcript quantification  
99 tools have already been in use for several years in the RNA-Seq community has resulted in well-  
100 developed, user-friendly software that can be applied almost immediately to the variant abundance  
101 quantification problem.

102

103 Here, we predict variant abundance by applying kallisto<sup>17</sup>. This algorithm takes as input a set of reference  
104 sequences to be quantified: for RNA-Seq these would be the different transcripts, but for wastewater  
105 sequencing data we provide it with a collection of SARS-CoV-2 genomes representative of the

106 population. kallisto constructs an index from the reference sequences and subsequently matches  
 107 sequencing reads to references, which allows it to estimate the abundance of each reference transcript  
 108 provided (**Fig 1b**). SARS-CoV-2 lineages are characterized by a combination of mutations, but additional  
 109 variation is observed within lineages (**Fig S1**). We provide kallisto with a reference set consisting of  
 110 multiple genomic sequences per lineage, capturing the mutations specific to this lineage as well as within-  
 111 lineage variation (**Fig 1c**). Our constructed reference set includes 1-17 genome sequences per lineage,  
 112 with a total number of nearly 1500 sequences for the 881 unique SARS-CoV-2 lineages present in the  
 113 GISAID database at the time of download (9 March 2021). Including multiple sequences per lineage  
 114 reduces biases related to within-lineage variation and potentially identifies any additional genomic  
 115 signatures frequently seen in a given lineage. Finally, we filter out any predictions below a given percent  
 116 abundance threshold to reduce noise and sum all predictions per lineage, which gives predicted  
 117 abundances per lineage (**Fig 1c**). While in this study we used kallisto, we expect similar results with  
 118 comparable tools such as salmon<sup>18</sup>.  
 119



122 Figure 1. Computational approach to variant of concern (variant) abundance estimation. **a)** Computational similarity  
123 between RNA transcript quantification and variant abundance estimation. **b)** Key aspects of the kallisto algorithm in  
124 the context of variant abundance estimation. **c)** Our workflow uses multiple reference sequence per lineage to  
125 capture within-lineage variation. Applying kallisto (as in part b) results in abundance estimates per reference  
126 sequence. These abundances are filtered using a minimal abundance cutoff and subsequently summed per lineage  
127 to obtain abundance estimates per lineage. Finally, variant abundances are reported.

128

### 129 **Kallisto predictions of variant abundance are accurate on simulated data**

130 To evaluate the accuracy of the predictions obtained through our pipeline, we created a collection of  
131 benchmarking datasets that resemble real wastewater samples. For each variant (B.1.1.7, B.1.351,  
132 B.1.427, B.1.429, P.1) we created a series of 33 benchmarks by simulating sequencing reads from a  
133 variant genome, as well as a collection of background (non-variant of concern/interest) sequences, such  
134 that the variant abundance ranges from 0.05% to 100%. Analogously, we created a second series of  
135 benchmarks, simulating reads only from the Spike gene of each SARS-CoV-2 genome. We refer to the  
136 first set of benchmarks as "whole genome" and to the second set of benchmarks as "Spike-only". Finally,  
137 we performed these benchmarking experiments at different sequencing depths: 100x and 1000x  
138 coverage for the whole genome benchmarks, and 100x, 1000x, and 10,000x coverage for the Spike-only  
139 benchmarks (**Table 1 and Fig 2**).

140

141 Predicting variant abundance can be difficult when a variant is present at very low frequency, because of  
142 the high degree of similarity between lineages. On our simulated datasets, where we know the true  
143 frequency of each variant, we observe a background noise of 0.01--0.09% (**Fig S2**), meaning that some  
144 sequences are falsely predicted to be present at 0.01--0.09% abundance. These false positives are likely  
145 due to shared mutations or conserved sequences between lineages. The level of background noise tends  
146 to be higher for whole genome benchmarks than for Spike-only benchmarks, because the majority of  
147 defining variant mutations are in the Spike gene (**Figs S1 and 2**). In both cases, we apply a threshold of  
148 0.1% abundance to include a sequence in the lineage abundance computation and we only report the  
149 presence of a variant exceeding this threshold to avoid false positives. For this reason, we only report  
150 results for benchmarks with a true variant abundance of at least 0.1%. Note that this threshold applies to  
151 the overall sequence abundance and not to individual mutations, since the stochasticity of wastewater  
152 sequencing causes the abundance of mutations within a variant to vary significantly.

153

154 **Figure 2** (top) shows the predicted versus true frequencies per variant for two of the benchmarks;  
155 additional results are shown in **Figure S3**. In general, variant frequencies tend to be underestimated, in  
156 particular on whole genome data. This is another consequence of shared polymorphisms between  
157 relatively closely related lineages: a fraction of the reads is assigned to other, locally identical genomes,  
158 leading to an underestimated variant frequency. The more divergent a variant is in comparison with other

159 lineages and the more unique polymorphisms are associated with it, the lower the number of false  
160 positives and the smaller the underestimation of variant abundance. This explains why our predictions are  
161 most accurate for P.1, the most divergent lineage among the variants considered (**Fig S1**). **Figure 2**  
162 (bottom) shows the relative frequency estimation error, defined as the absolute difference between true  
163 and estimated frequency, relative to the true frequency. We observe that relative frequency estimation  
164 errors are highest at low frequencies, where a small deviation in the absolute sense makes a large  
165 relative difference but are relatively stable at true variant frequencies of at least 1%.

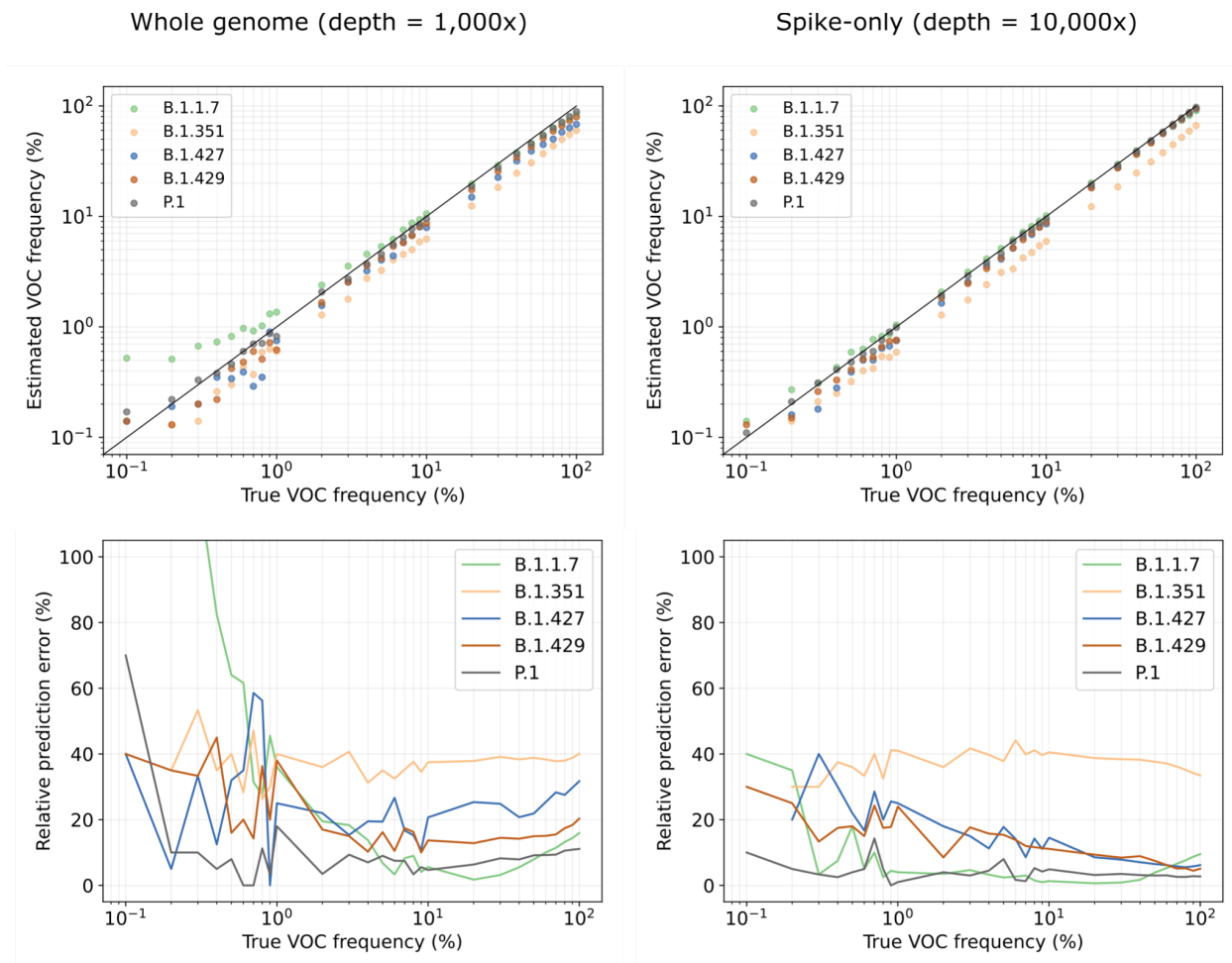
166

167 Besides underestimation, we also notice overestimation, particularly at low variant frequencies. This is  
168 due to shared sequence between the variant and the background lineages: in datasets where the variant  
169 is present at a low frequency, the background lineages must be present at relatively high frequencies.  
170 Any shared sequence between a background lineage and the variant will then lead to more reads being  
171 assigned to the variant, hence overestimation. This effect only applies when background lineages with  
172 shared sequence are more abundant than the variant. In **Figure 2**, we can clearly see how P.1 and  
173 B.1.1.7 abundances are overestimated at low frequencies, while being near-perfect at higher frequencies  
174 (>1%); for other variants, we see that this effect compensates some of the underestimation, resulting in  
175 better estimates at low variant frequencies.

176

177 Two variants which are particularly difficult to predict individually are B.1.427 and B.1.429; these lineages  
178 are highly similar and have the same characterizing mutations in the Spike gene<sup>19</sup>. However, because we  
179 include multiple sequences per lineage in our reference set, we capture within-lineage variation that  
180 allows us to distinguish between these two lineages using Spike-only data. This highlights the power of  
181 our approach using a complete reference set; it would not be possible to distinguish between B.1.427 and  
182 B.1.429 with an approach based on mutation frequencies alone.

183



184

185 Figure 2. Estimated variant abundances and relative prediction errors. Relative prediction errors are defined as the  
 186 absolute difference between true and estimated frequency, relative to the true frequency.

187

188

Benchmark	FPR	FNR	Precision	Recall	Relative estimation error (%)
Whole genome 100x	0.191 / 0.0	0.057 / 0.032	0.423 / 1.0	0.943 / 0.968	29.4 / 19.4
Whole genome 1,000x	0.163 / 0	0.007 / 0.042	0.470 / 1.0	0.993 / 0.958	27.1 / 18.5
Spike-only 100x	0.121 / 0.003	0.107 / 0.074	0.508 / 0.978	0.893 / 0.926	26.3 / 15.8

Spike-only 1,000x	0.041 / 0.003	0.043 / 0.042	0.753 / 0.978	0.957 / 0.958	17.3 / 14.0
Spike-only 10,000x	0.010 / 0	0.014 / 0.042	0.926 / 1.0	0.986 / 0.958	15.3 / 13.0

189 Table 1. Performance statistics per dataset. Results separated by a forward slash correspond to an abundance  
190 threshold of 0.1% and 1%, respectively. FPR = false positive rate; FNR = false negative rate; relative estimation error  
191 reflects the average relative frequency estimation error across all true positives.

192  
193 We observe that variant frequencies are consistently underestimated using our approach, except for  
194 slight overestimation of unusually divergent lineages (P.1, B.1.1.7) at low frequencies. Consistent variant  
195 bias as observed in our experiments is unlikely to be an issue in differential analysis, but in single-point  
196 evaluations it would be necessary to design a method to correct for these variant-specific biases.  
197 However, our benchmarks are generated using a single variant sequence, while real data consists of a  
198 mixture of different sequences for the same variant. This may have resulted in stronger underestimation  
199 on our benchmarks than would be seen on real data. To complement the benchmarking experiments  
200 presented here, it would be interesting to evaluate predictions on real data more thoroughly, e.g., by  
201 comparing to qPCR-based variant abundance estimates per variant. More extensive benchmarking  
202 experiments will make it possible to learn the variant-specific biases more accurately and adjust  
203 predictions accordingly.

204  
205 To evaluate false positive and false negative predictions, we computed for each experiment the overall  
206 false positive rate (FPR), false negative rate (FNR), precision, and recall (**Table 1**). We calculated these  
207 statistics for minimal variant abundance thresholds of 0.1% and 1%. Increasing the minimal abundance  
208 thresholds reduces the false positive rate but increases the false negative rate. We generally observe  
209 more false negatives for datasets of lower coverage, because low-frequency variants become harder to  
210 detect. In terms of precision and recall, we note that, unsurprisingly, increasing the number of reads  
211 (either by amplifying a larger region, or by increased sequencing depth) leads to better results. If a variant  
212 is uniquely defined by mutations on Spike, then sequencing depth is preferred over breadth, but if a  
213 variant is (nearly) identical to other lineages on Spike, e.g., B.1.427/B.1.429, then whole genome  
214 sequencing is preferable.

215  
216 While for this study we primarily used kallisto<sup>17</sup>, we also evaluated performance for the software package  
217 salmon<sup>18</sup>, which takes a slightly different algorithmic approach to the same problem (**Fig S4**) and found  
218 predictions were highly similar to those obtained with kallisto, the main difference being that salmon is  
219 slightly more conservative: it achieves higher precision (fewer false positives), at the expense of lower  
220 recall (more false negatives). Although salmon tends to miss variants at very low frequencies, one



221 potential advantage is that this method may also be applied to long reads (in alignment-based mode),  
222 while kallisto usage is limited to short reads.

223

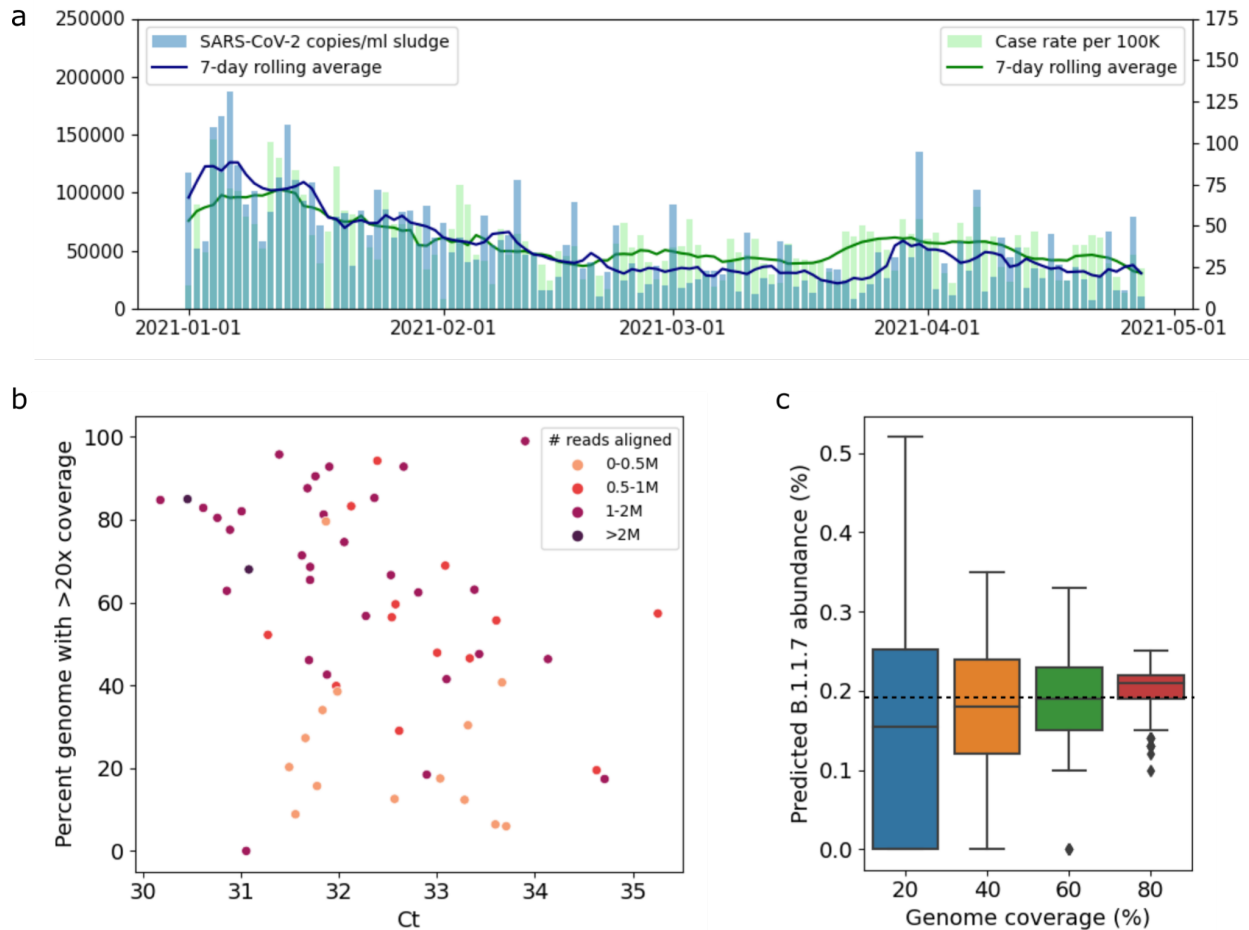
### 224 **Observed PCR values of wastewater correspond to genome coverage**

225 We obtained primary sewage sludge samples from the wastewater treatment plant serving New Haven,  
226 CT, USA, every 2 days between January 1, 2021 and April 27, 2021 (59 samples). The observed SARS-  
227 CoV-2 RNA levels in these samples follow the same trend as COVID-19 case rates in the same  
228 geographic region (**Fig 3a**, similarly found in <sup>10</sup>). PCR cycle threshold (Ct) values of SARS-CoV-2 from  
229 undiluted sludge samples ranged from 30.2 to 35.3 ( $7.1 \times 10^3$  -  $1.6 \times 10^5$  virus copies / mL), indicating  
230 that there was enough viral RNA to apply genomic sequencing. We generated 400nt tiled amplicons  
231 encompassing the SARS-CoV-2 genome using the Illumina COVIDSeq Test (RUO), modified to use  
232 NEBNext ARTIC V3 SARS-Cov-2 Primer Mixes 1 and 2 to improve genome coverage at low RNA  
233 concentrations<sup>20,21</sup>. The resulting sequencing data varied widely in terms of number of reads and genome  
234 coverage (**Fig S5**), with low Ct values (high SARS-CoV-2 RNA concentrations) generally leading to higher  
235 genome coverage (**Fig 3b**). For these datasets, Ct values < 31 yields at least 60% genome coverage;  
236 samples with a Ct value < 34 and at least 0.5M reads aligned reach a genome coverage of at least 40%.

237

238 To evaluate the impact of genome coverage on variant abundance predictions, we subsampled a dataset  
239 with maximal coverage (99% of the SARS-CoV-2 genome with >20x coverage, 1.9M paired-end reads) to  
240 obtain datasets with reduced genome coverage by randomly selecting 20%, 40%, 60%, and 80% of  
241 amplicons, respectively, each of which we repeated 100 times. **Figure 3c** shows the resulting abundance  
242 predictions for B.1.1.7 per coverage value. We observe that the median predicted abundance is close to  
243 the predicted abundance at full coverage (dashed line in **Fig 3c**) for all coverage values; however,  
244 variance is much larger in datasets with low coverage compared to datasets with high coverage,  
245 consistent with statistical predictions and prior work<sup>16</sup>. This indicates that datasets with low genome  
246 coverage can still result in accurate abundances, but the predictions are less reliable.

247



248

249 Figure 3. a) RNA levels in wastewater (copies/ml sludge, displayed on left vertical axis) follow the same trend as  
250 COVID-19 case rates (cases per 100K people, displayed on right vertical axis). b) Percent genome with >20x  
251 coverage versus sludge Ct values. c) Impact of genome coverage on predicted B.1.1.7 abundance for random  
252 subsamples of a sludge sample with full genome coverage. The horizontal dotted line indicated the predicted B.1.1.7  
253 abundance for the full sample (99% genome coverage).

254

255

### 256 Wastewater abundances of B.1.1.7 and B.1.526 in Connecticut broadly correspond to clinically 257 observed frequencies

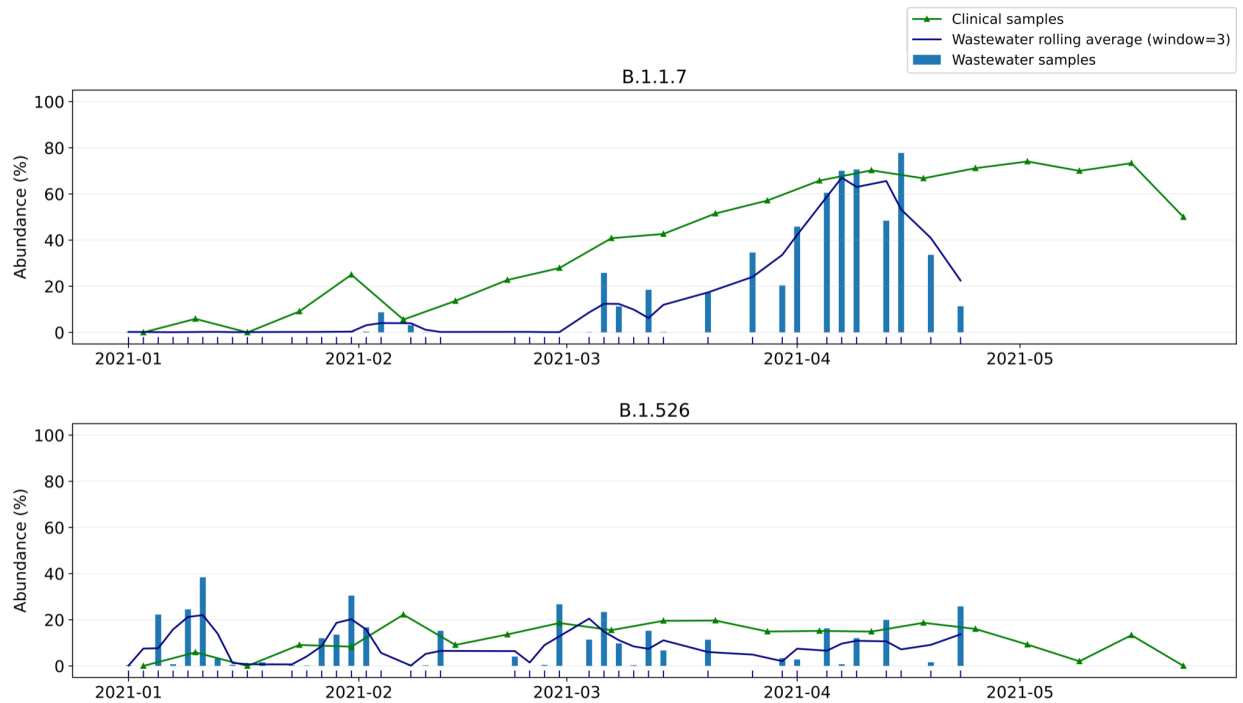
258 We applied our abundance prediction pipeline to the series of wastewater sequencing datasets described  
259 above. **Figure 4** shows the resulting predictions for lineages B.1.1.7 and B.1.526. There is a clear trend in  
260 B.1.1.7 abundance emerging in early February 2021, increasing in abundance through mid April 2021,  
261 while the abundance of other variants is relatively stable over time (see also **Fig S6**).

262

263 We then compared our wastewater abundance predictions to variant frequency estimates from data  
264 generated by sequencing remnant clinical diagnostic samples (mostly nasal swabs) in New Haven

265 County, CT (**Fig 4**). We observe that B.1.1.7 abundances predicted from wastewater are underestimated  
266 compared to the clinical abundance data. Based on our benchmarking experiments on whole genome  
267 data (**Fig 2**, left) we expect frequencies to be underestimated by 5-40% (relative to the actual frequency),  
268 with stronger underestimation for lower frequencies. This is consistent with what we see in **Figure 4**: the  
269 increase (and subsequent decrease) of B.1.1.7 abundance in wastewater is stronger than in clinical data  
270 because of this bias, while wastewater abundance predictions are very close to clinical predictions at  
271 frequencies of 60-70%. For B.1.526, both clinical and wastewater abundance is relatively stable over  
272 time. In theory, predictions will be closer to clinical abundances when sequencing only the Spike gene  
273 instead of the whole SARS-CoV-2 genome (**Fig 2**). In practice, however, using only reads aligning to the  
274 Spike gene captures too little information due to incomplete genome coverage and amplification bias.

275  
276 Kallisto offers a bootstrapping feature, through which the sequencing data is resampled at least 100 times  
277 and variant abundances are predicted for each of these resampled datasets. The resulting predictions  
278 can subsequently be analyzed to obtain confidence intervals for the predicted abundance on the original  
279 dataset. For the New Haven sludge samples discussed here we obtained narrow confidence intervals  
280 (upper and lower errors <1% abundance), suggesting that predictions are very consistent (**Fig S6**).  
281 However, this type of analysis captures only computational noise, and not technical noise (e.g. sampling  
282 bias). The fact that our predictions are more variable than the clinical data while confidence intervals from  
283 bootstrapping are narrow suggests that wastewater sequencing data is highly stochastic and not always  
284 representative of the infections in the population.



285

286 Figure 4. Wastewater versus clinical abundance estimates for B.1.1.7 and B.1.526 in New Haven from early January  
287 2021 to late April 2021. Dates of clinical sampling correspond to the date of specimen collection.

288

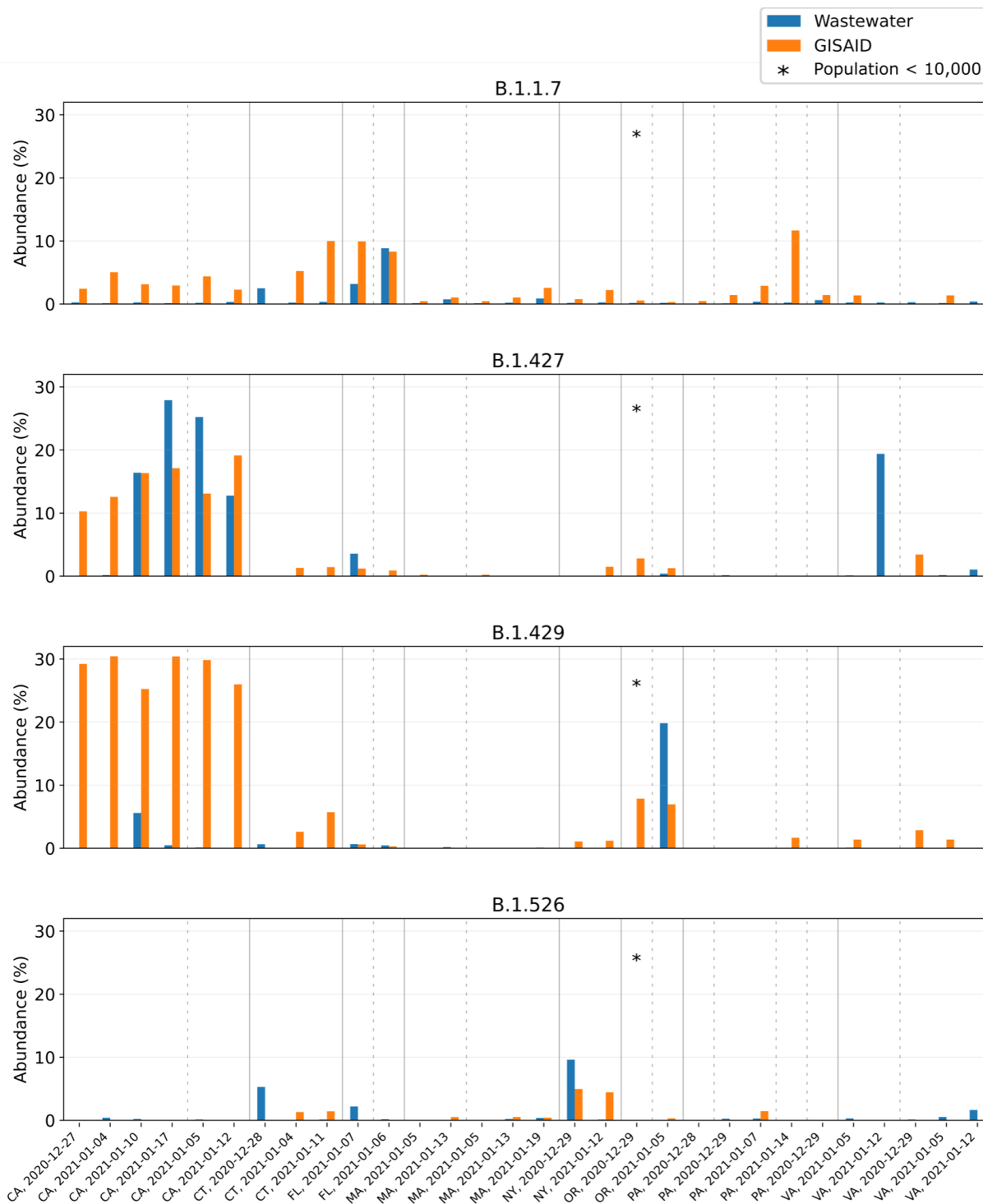
### 289 **Wastewater abundances for different variants across the US match expected patterns**

290 The quality of datasets obtained through wastewater sequencing varies widely. RNA levels and Ct values  
291 fluctuate with the number of infections in a sewershed, which impact overall genome coverage and  
292 prediction accuracy. Other factors such as sampling approach, PCR inhibition, and amplification bias add  
293 to this variability. To validate our approach for general use, we applied our pipeline to predict variant  
294 abundance on a diverse collection of composite influent wastewater samples obtained from 25 treatment  
295 plants across the United States between late December 2020 and late January 2021 (**Fig 5**). These  
296 samples had slightly higher average Ct values than the sludge samples analyzed above (33.1 vs 32.3),  
297 but reduced genome coverage compared to sludge samples. Nevertheless, the same quality filtering  
298 parameters apply: Ct < 31 or Ct < 34 with at least 0.5M reads aligned to select for samples with high  
299 coverage (**Fig S7**).

300

301 After this filtering step, we predicted variant abundance for the 30 remaining datasets (corresponding to  
302 16 different locations across 8 states). **Figure 5** shows the predictions for lineages B.1.1.7, B.1.427,  
303 B.1.429, and B.1.526, along with the clinical lineage frequencies in the corresponding state (calculated  
304 from GISAID in the 7-day window centered at the wastewater sampling date). Although data uploaded to  
305 GISAID has its own biases and individual towns do not necessarily reflect state-wide variant abundances,  
306 this is the only statistic we can compare against across all states. We observe that, while individual

307 samples are unreliable, the predicted variant abundances match expected patterns across the US from  
308 the times of sampling: B.1.1.7 was predicted most abundantly in Florida; B.1.427 and B.1.429 were  
309 primarily found in California; and B.1.526 was predicted most abundantly in New York and Connecticut.  
310 Other variants (B.1.351, P.1) were not observed in GISAID for these states at the time of sampling and  
311 our predictions for these variants agree: B.1.351 was predicted to be present at very low frequency in 4  
312 samples and absent in all other samples; P.1 was predicted present in a single dataset at 1% abundance  
313 and absent in all others (**Fig S8**). Although these predictions may be false positives, at the time P1 was  
314 thought to be likely at such low prevalence that these cases were not picked up by the sequencing efforts  
315 in place.



316

317 Figure 5. Wastewater versus GISAID abundance estimates for B.1.1.7, B.1.427, B.1.429 and B.1.526 at 16 locations

318 across 8 states of the US. Samples were collected between late December 2020 and late January 2021; sampling

319 date and location are indicated on the horizontal axis. Samples are sorted by location, with different locations

320 separated by a dotted line and different states separated by a solid line.

## 321 Discussion

322 Our results show that methods for RNA transcript quantification can be applied to wastewater sequencing  
323 data to obtain consistent and relevant variant abundance estimates. This technique can be readily applied  
324 to a wide range of data types, from Spike-only amplicon sequencing to whole genome sequencing; it is  
325 not appreciably more difficult than setting up a “reference set” of potential variants and running existing  
326 tools on whatever the sequencing data happens to be. While this reference set approach allows easy  
327 updating as new variants appear, it also means that this approach cannot be used to detect new variants,  
328 but only to near-optimally impute the mixture of known variants most likely responsible for the observed  
329 data. Further, the approach may be readily extended to the detection and quantification of other pathogen  
330 lineages present in wastewater.

331 Detecting variants in this fashion appears to be near-optimal on simulated data, with a detection limit  
332 under 1%, though when handling real data which is subject to multiple factors that can alter the quality of  
333 the RNA and the resulting signal, this may be closer to 10%. Wastewater abundances generally follow the  
334 abundance trends seen in clinical data, though with sufficient noise that individual timepoints should not  
335 be considered reliable abundance estimates. Predicting variant abundance from wastewater sequencing  
336 is challenging when viral titers are low, as a result of low prevalence of SARS-CoV-2, but in high-  
337 prevalence regions this approach can be extremely effective.

338 When comparing wastewater abundances to population-level clinical frequencies of variants, there are  
339 three distinct potential sources of errors, all of which are conflated in the accuracy of the final estimate:  
340 (1) how well clinical frequencies match the rates in the population as a whole, (2) how representative the  
341 RNA in a given wastewater is of the infections in the population as a whole, and (3) how well the  
342 predicted variant abundances from sequencing accurately represent what is in the wastewater sample  
343 itself. Clinical frequencies and the GISAID data used for the national data above have strong biases and  
344 so are not themselves ground truth. Individual wastewater samples can be unreliable: the catchment and  
345 composition of an individual wastewater sample can include hospital or industrial inputs and is not  
346 necessarily representative of the population, and low infection levels, inhibitory compounds, and  
347 degradation of RNA can result in higher Ct values and associated genome coverage. Further, the  
348 potential that different variants have different viral shedding in waste is not taken into account nor are  
349 differences in shedding rates for vaccine breakthrough cases. The data here is consistent with high-  
350 coverage sequencing being well representative of the sample and the computation faithfully  
351 reconstructing it, therefore we believe the primary source of observed noise is the underlying noisiness  
352 correspondence between a given wastewater sample, the population as a whole, and the clinical cases.

353 These results do offer lessons for future development of wastewater sequencing methods. With perfect  
354 coverage (simulated data), Spike-only sequencing gives better predictions than whole genome

355 sequencing. More broadly, as the variant-discriminating mutations are restricted to a subset of the  
356 amplicons in tiled sequencing, a strategy focusing sequencing depth proportionally to the potentially  
357 informativity of the tiles would likely yield the more accurate predictions per sequence read. In practice,  
358 whole genome sequencing is often necessary to obtain enough information versus restricting to Spike.  
359 These differences in sequencing protocols, as well as differences in population size and sampling  
360 techniques, make it challenging to compare data between locations. Another important lesson is the  
361 improved coverage with higher virus RNA concentrations (lower PCR Ct values). While Ct is largely  
362 controlled by the infection rate in a community and excretion into wastewater, sample concentration and  
363 PCR inhibition removal approaches can be deployed to lower the Ct value and improve coverage in most  
364 samples.

365 In applying RNA-seq tools to variant prediction, reference set composition is central. We were able to  
366 reduce variant-specific biases substantially by including multiple reference sequences per lineage, thus  
367 capturing within-lineage variation to identify variants with highly similar genomes. Additional experiments  
368 (data not shown) comparing a US-specific reference set to a global reference set showed that,  
369 unsurprisingly, the US-specific reference set gives more accurate predictions for benchmarks with  
370 sequences from US origin. This suggests that using a state- or county-specific reference set construction  
371 could further improve results and that identification of variants is likely being aided by the presence of  
372 hitchhiker mutations which are locally over-represented or non-defining.

373 In conclusion, we present a computational approach for estimating the percent abundances of SARS-  
374 CoV-2 variants in wastewater. Temporal patterns in wastewater of variant abundances in a mid-size  
375 municipality in Connecticut matched those defined by compiled clinical sequence data, and sequencing  
376 metrics were interrogated to define the Ct value and other confidence thresholds that ensure optimal  
377 performance. In settings across the world where strong clinical variant sequencing programs do not exist,  
378 wastewater sequencing can be an effective tool for low cost, efficient monitoring of variant abundance. As  
379 this is unlikely to be the last viral pandemic, nor the last with variants of concern, extending these  
380 approaches to other viruses and other sample types may allow broader monitoring of real-time pandemic  
381 evolution.

382

383



## 384 Methods

### 385 **Constructing a reference set**

386 We selected representative genomes per lineage from the GISAID database<sup>4</sup>, downloaded on 9 March  
387 2021. As our samples are from wastewater collection systems across the US, we considered only  
388 reference sequences of US origin. After removing low-quality sequences (defined as having less than  
389 29,500 non-ambiguous nucleotides) we randomly selected 1000 sequences per lineage for further  
390 analysis. We used minimap2 and pafutils to align each of these sequences to the reference genome  
391 (MN908947.3) and subsequently identify variation with respect to this reference<sup>22</sup>. We then used  
392 VCFtools to compute allele frequencies within each lineage. Based on these allele frequencies, we  
393 selected sequences per lineage such that all mutations with an allele frequency of at least 50% were  
394 captured at least once. This resulted in a final reference set of 1,488 complete SARS-CoV-2 genome  
395 sequences.

### 396 397 **Designing benchmarking experiments**

398 To evaluate the accuracy of our variant abundance predictions, we created benchmarks consisting of a  
399 selection of non-variant genomes (background) and one variant. In order to build benchmarks that reflect  
400 our real data as closely as possible, we selected the background genomes by taking all 11 sequences in  
401 GISAID collected in Connecticut at 2021-02-11 (the most recent collection date). For Spike-only  
402 benchmarks, we trimmed these sequences to keep only the Spike gene and simulated paired-end (2x150  
403 bp) Illumina sequencing reads at equal abundance using ART<sup>23</sup>. In addition, we randomly selected variant  
404 sequences from GISAID and simulated sequencing reads for the Spike gene of each variant (B.1.1.7,  
405 B.1.351, P.1, B.1.427, B.1.429) at varying frequencies (0.05, 0.06 ..., 0.1, 0.2, ..., 1, 2, ..., 10, 20, ...,  
406 100%) to create 33 data sets per variant, hence 165 data sets in total. We performed these simulations at  
407 a total coverage of 100x and 1000x for whole genome benchmarks, and at 100x, 1000x, and 10,000x for  
408 Spike-only benchmarks.

### 409 410 **Wastewater collection and sequencing from New Haven, CT**

411 Primary sewage sludge samples were collected from the New Haven, CT, USA Wastewater Treatment  
412 Plant. The plant serves 200,000 residents in the towns of New Haven, Hamden, East Haven and  
413 Woodbridge, CT. Primary sludge samples were collected from the effluent pump of the plant's gravity  
414 thickener. Samples were collected every other day starting January 3, 2021 and ending April 27, 2021.  
415 RNA was extracted from 500  $\mu$ L sample using a Zymo *Quick-RNA* Fecal/Soil Microbe Microprep Kit  
416 modified by the addition of 100  $\mu$ L of phenol-chloroform to the bead beating step, and eluted in 50  $\mu$ L of  
417 nuclease-free water. The SARS-CoV-2 whole-genome sequencing library was prepared from extracted  
418 RNA using the Illumina COVIDSeq Test (RUO), modified to use NEBNext ARTIC V3 SARS-Cov-2 Primer  
419 Mixes 1 and 2 instead of the included primer mixes, for cDNA synthesis, amplicon generation,

420 tagmentation, and cleaning. The pooled and cleaned library was sequenced on an Illumina NovaSeq at  
421 the Yale Center for Genomic Analysis; each sample was given at least 1 million reads. Negative controls  
422 were included at the cDNA synthesis and amplicon generation steps.

423

#### 424 **Wastewater collection and sequencing from across the US**

425 Composite influent samples were collected by participating wastewater treatment facilities using  
426 equipment that these facilities already had in-house. Composite samples were aliquoted into three 50-mL  
427 conical tubes and shipped within 24 hours of collection overnight with ice packs to the Biobot Analytics  
428 laboratory (Cambridge, MA). Received samples were immediately pasteurized at 60°C for 1h.

429

430 One of the three tubes was then filtered to remove large particulate matter using a 0.2µm vacuum-driven  
431 filter (EMD-Millipore SCGP00525 or Corning 430320, depending on sample turbidity). We then used  
432 Amicon Ultra-15 centrifugal ultrafiltration units (Millipore UFC903096) to concentrate 15mL of wastewater  
433 approximately 100x. We lysed viral particles in the concentrate by adding AVL Buffer containing carrier  
434 RNA (Qiagen 19073) to the Amicon unit before transfer and >10 minute incubation in a 96-well 2mL  
435 block. To adjust binding conditions, 100% ethanol was added to the lysate, and samples were applied to  
436 RNeasy Mini columns or RNeasy 96 cassettes (Qiagen 74106 or 74181). For a subset of samples (all  
437 from locations within Massachusetts) we processed 45mL of wastewater by loading the same Amicon  
438 Ultra-15 unit three times.

439

440 The RNA samples resulting from the extraction process described above were used as the template for  
441 reverse-transcription (RT) reactions performed with LunaScript RT SuperMix enzyme mix (NEB) to  
442 generate cDNA. Reaction conditions were as follows: primer annealing at 25°C for 2 min, cDNA synthesis  
443 at 55°C for 10 min and heat inactivation at 95°C for 1 min. Multiplexed polymerase chain reaction (PCR)  
444 amplification of cDNA was performed with Q5 Hot Start High-Fidelity 2X Master Mix (NEB) and ARTIC v3  
445 primers (0.015 µM each, final) in two non-overlapping pools with the following cycling conditions: heat  
446 activation at 98°C for 30 sec, followed by 35 cycles of 15 sec denaturation at 98°C, 5 min  
447 annealing/elongation at 65°C.

448

449 The non-overlapping amplicon pools were combined and sequencing libraries for Illumina platform were  
450 prepared using tagmentation with bead-linked transposomes (Illumina) and a modified amplification  
451 protocol with KAPA HiFi HotStart ReadyMix (Roche) and combinatorial dual-indexed adapter sequences.  
452 Libraries were sequenced with NextSeq550 (Illumina).

453

454

455

456

## 457 **Wastewater data preprocessing**

458 Before processing with kallisto, we first removed adapter sequences from the reads using Trimmomatic<sup>24</sup>,  
459 aligned the trimmed reads to a reference genome (GenBank MN908947.3) with BWA-MEM v0.7.17<sup>25</sup>,  
460 and subsequently identified primer sequences using iVar v.1.3.1<sup>16</sup> and removed these with jvarkit  
461 (<http://lindenb.github.io/jvarkit/Biostar84452>).

462

## 463 **Clinical sequencing and data processing from New Haven, CT**

### 464 Ethics statement

465 The Institutional Review Board from the Yale University Human Research Protection Program determined  
466 that the RT-qPCR testing and sequencing of de-identified remnant COVID-19 clinical samples obtained  
467 from clinical partners conducted in this study is not research involving human subjects (IRB Protocol ID:  
468 2000028599).

469

### 470 Sequencing and consensus generation

471 Residual routine testing samples from confirmed SARS-CoV-2 positive individuals were provided by Yale  
472 New Haven Hospital, Yale Pathology Laboratory, “Yale Campus Study”, Connecticut Department of  
473 Public Health, and Murphy Medical Associates. Sample types included nasal swabs in viral transport  
474 media, raw saliva, and extracted RNA. Samples not arriving as RNA were processed using the MagMAX  
475 viral/pathogen nucleic acid isolation kit; RNA was extracted from 300 µL of sample and eluted in 75 µL  
476 elution buffer. All products were tested using a locally developed assay for variants to determine viral  
477 RNA concentration<sup>26</sup>. Samples with sufficient RNA for sequencing (defined as a viral target cycle  
478 threshold value <35) were prepared using the Illumina COVIDSeq Test RUO for cDNA synthesis,  
479 amplicon generation, tagmentation, and cleaning. Pooled and cleaned libraries were sequenced using a  
480 2x100 or 2x150 approach on an Illumina NovaSeq at the Yale Center for Genomic Analysis; each sample  
481 was given at least 1 million reads. Negative controls were included at RNA extraction, cDNA synthesis,  
482 and amplicon generation steps.

483

484 Reads were aligned to a reference genome (GenBank MN908937.3) using BWA-MEM v.0.7.15<sup>25</sup>.

485 Adaptor trimming, primer sequence masking, and simple majority base calling were conducted using iVar  
486 v1.2.1<sup>16</sup> and SAMtools<sup>27</sup>. Lineages were assigned using pangolin v.2.4.2<sup>28</sup>.

487

488

### 489 **Software availability**

490 All code used for the analysis presented in this manuscript is publicly available at

491 [https://github.com/baymlab/wastewater\\_analysis](https://github.com/baymlab/wastewater_analysis).

492

493

494 **Data availability**

495 The raw SARS-CoV-2 sequencing data from New Haven wastewater (.fastq files) are available on NCBI  
496 SRA under Bioproject PRJNA741211. The clinical sequencing data can be accessed via  
497 covidtrackerct.com. The raw SARS-CoV-2 sequencing data from across the U.S. (.fastq files) are  
498 available on NCBI SRA under Bioproject PRJNA759260. The simulated wastewater sequencing data  
499 (.fastq files) for benchmarking are available on Zenodo (DOI: 10.5281/zenodo.5307070).

500

501 **Yale SARS-CoV-2 Genomic Surveillance Initiative authors**

502 Ahmad Altajar, Anderson F. Brito, Anne E. Watkins, Anthony Muyombwe, Caleb Neal, Chen Liu,  
503 Christopher Castaldi, Claire Pearson, David R. Peaper, Eva Laszlo, Irina R. Tikhonova, Jafar Razeq,  
504 Jessica E. Rothman, Jianhui Wang, Kaya Bilguvar, Linda Niccolai, Madeline S. Wilson, Margaret L.  
505 Anderson, Marie L. Landry, Mark D. Adams, Pei Hui, Randy Downing, Rebecca Earnest, Shrikant Mane,  
506 Steven Murphy

507

508 **Acknowledgements**

509 This work was supported in part by the Pew Charitable Trusts, the David and Lucile Packard Foundation,  
510 NIH NIGMS award R35GM133700, and the Alfred P. Sloan Foundation (J.A.B. and M.B); CTSA Grant  
511 Number TL1 TR001864 (M.E.P. and T.A.); Fast Grant from Emergent Ventures at the Mercatus Center at  
512 George Mason University (N.D.G.); CDC Contract #75D30120C09570 (N.D.G.); Yale CoReCT pilot  
513 award (J.P. and N.D.G.); and NIH NIGMS award U54GM088558 (W.P.H.). James McGann and Jim  
514 Griffin were involved in developing the sequencing methodology at Ginkgo Bioworks.

515

516 **Competing interests**

517 N.D.G. is an infectious diseases consultant for Tempus Labs. W.P.H. is a scientific advisory board  
518 member to Biobot Analytics and has received compensation for expert witness testimony on the expected  
519 course of the pandemic. N.G. is co-founder of Biobot Analytics; C.D., K.A.M., and M.I. are employees of  
520 Biobot Analytics.

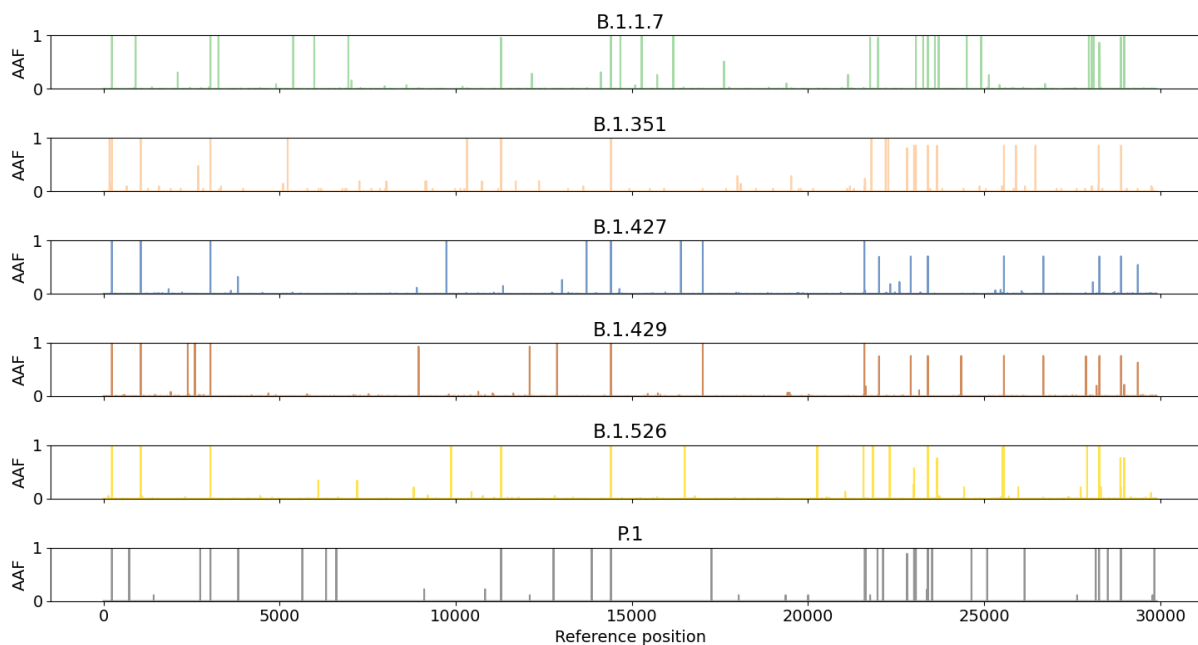
## 521 References

- 522 1. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England.  
523 *Science* **372**, (2021).
- 524 2. Lucas, C. *et al.* Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity in  
525 uninfected and previously infected individuals. *bioRxiv* (2021) doi:10.1101/2021.07.14.21260307.
- 526 3. CDC. SARS-CoV-2 Variant Classifications and Definitions. [https://www.cdc.gov/coronavirus/2019-](https://www.cdc.gov/coronavirus/2019-nCoV/variants/variant-info.html)  
527 [ncov/variants/variant-info.html](https://www.cdc.gov/coronavirus/2019-nCoV/variants/variant-info.html) (2021).
- 528 4. GISAID - Initiative. <https://www.gisaid.org/>.
- 529 5. Zhang, W. *et al.* Emergence of a Novel SARS-CoV-2 Variant in Southern California. *JAMA* **325**,  
530 1324–1326 (2021).
- 531 6. Nemudryi, A. *et al.* Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal  
532 Wastewater. *Cell Rep Med* **1**, 100098 (2020).
- 533 7. Peng, L. *et al.* SARS-CoV-2 can be detected in urine, blood, anal swabs, and oropharyngeal swabs  
534 specimens. *J. Med. Virol.* **92**, 1676–1680 (2020).
- 535 8. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2  
536 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the  
537 Epidemic in The Netherlands. *Environ. Sci. Technol. Lett.* **7**, 511–516 (2020).
- 538 9. Wolfe, M. K. *et al.* Scaling of SARS-CoV-2 RNA in Settled Solids from Multiple Wastewater  
539 Treatment Plants to Compare Incidence Rates of Laboratory-Confirmed COVID-19 in Their  
540 Sewersheds. *Environmental Science & Technology Letters* vol. 8 398–404 (2021).
- 541 10. Zulli, A. *et al.* Predicting daily COVID-19 case rates from SARS-CoV-2 RNA concentrations across a  
542 diversity of wastewater catchments. *medRxiv* (2021).
- 543 11. Crits-Christoph, A. *et al.* Genome sequencing of sewage detects regionally prevalent SARS-CoV-2  
544 variants. doi:10.1101/2020.09.13.20193805.
- 545 12. Jahn, K. *et al.* Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater  
546 samples. *medRxiv* (2021).
- 547 13. COVID-19 Wastewater Epidemiology SARS-CoV-2. <https://www.covid19wbec.org/>.
- 548 14. Ellmen, I. *et al.* Alcov: Estimating Variant of Concern Abundance from SARS-CoV-2 Wastewater  
549 Sequencing Data. *medRxiv* (2021).
- 550 15. Posada-Céspedes, S., Seifert, D., Topolsky, I., Metzner, K. J. & Beerenwinkel, N. V-pipe: a  
551 computational pipeline for assessing viral genetic diversity from high-throughput sequencing data.  
552 doi:10.1101/2020.06.09.142919.
- 553 16. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost  
554 virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
- 555 17. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification.  
556 *Nat. Biotechnol.* **34**, 525–527 (2016).
- 557 18. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware  
558 quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
- 559 19. Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry  
560 Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su,  
561 Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology. [outbreak.info](https://outbreak.info).  
562 [outbreak.info](https://outbreak.info).
- 563 20. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus  
564 genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
- 565 21. Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome  
566 sequencing using nanopore. *bioRxiv* (2020) doi:10.1101/2020.09.04.283077.
- 567 22. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.

- 568        *Bioinformatics* **32**, 2103–2110 (2016).
- 569    23. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator.  
570        *Bioinformatics* **28**, 593–594 (2012).
- 571    24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.  
572        *Bioinformatics* **30**, 2114–2120 (2014).
- 573    25. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-  
574        bio.GN]* (2013).
- 575    26. Vogels, C. B. F. *et al.* Multiplex qPCR discriminates variants of concern to enhance global  
576        surveillance of SARS-CoV-2. *PLoS Biol.* **19**, e3001236 (2021).
- 577    27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079  
578        (2009).
- 579    28. O'Toole, A. *et al.* Pangolin: lineage assignment in an emerging pandemic as an epidemiological tool.  
580        (2020).
- 581
- 582
- 583

584 **Supplementary figures**

585

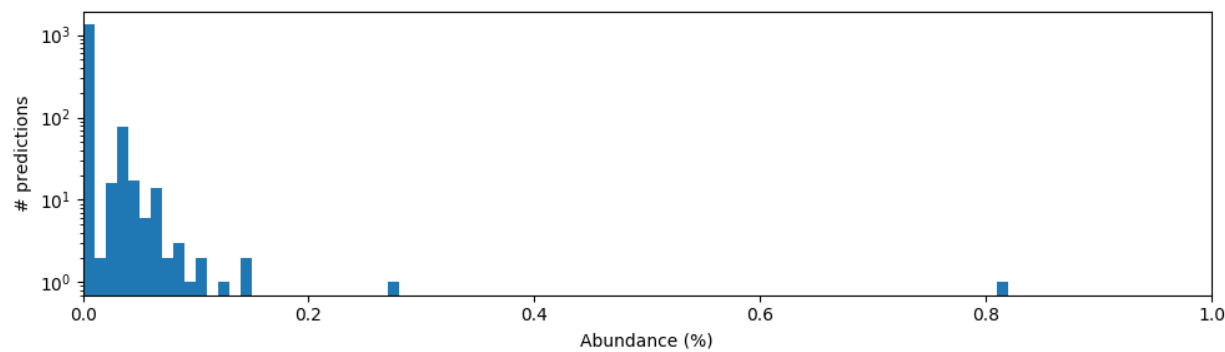


586

587 *Figure S1: Within-lineage diversity observed in SARS-CoV-2 genomes on GISAID (downloaded 9 March 2021). The*  
588 *horizontal axis shows the position (in base pairs) on the reference genome (accession MN908947.3). The y-axis*  
589 *shows the alternative allele frequency (AAF), i.e. the fraction of genomes with a different nucleotide at a given*  
590 *position than the reference genome. This plot was computed by randomly selecting 1000 genomes of US origin per*  
591 *lineage.*

592

593



594

595 *Figure S2. Histogram of raw abundances predicted by kallisto on a simulated dataset consisting of 100% B.1.1.7. The*  
596 *majority of false positives (background noise) can be filtered out by applying a minimal abundance threshold of 0.1%.*  
597 *True predictions occur at higher abundances (beyond the x-axis limit of 1.0%).*

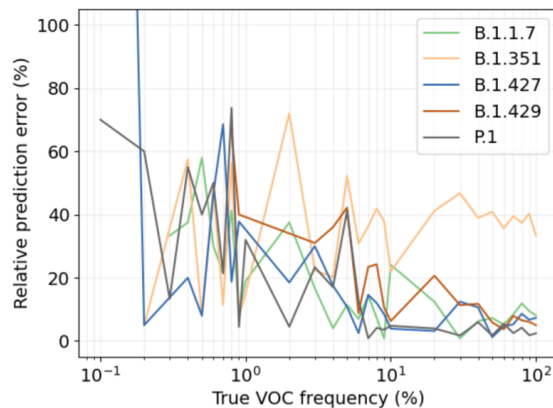
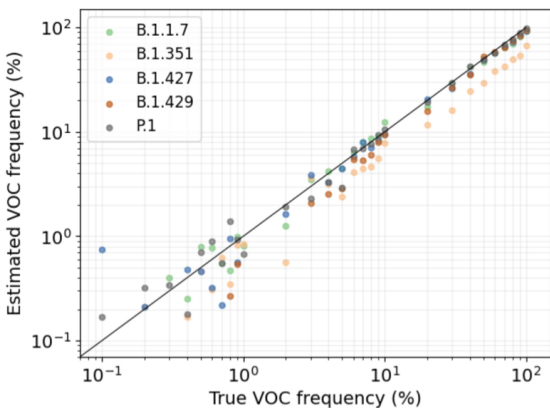
598

599

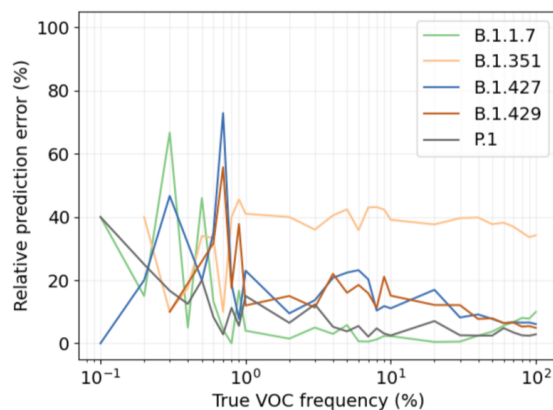
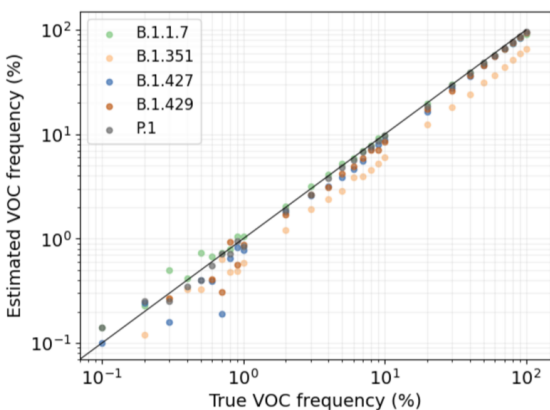
600



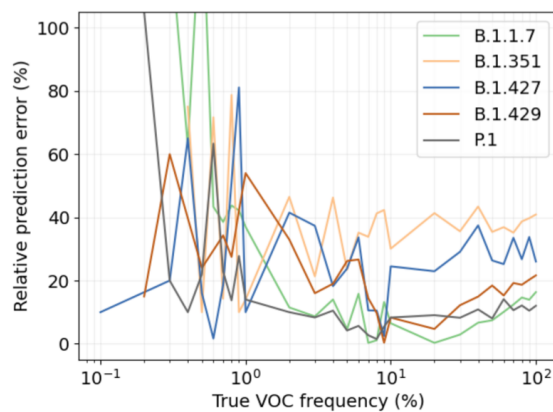
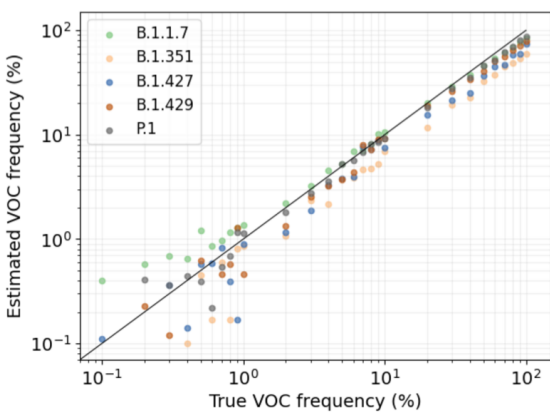
*Spike-only, depth = 100x*



*Spike-only, depth = 1000x*



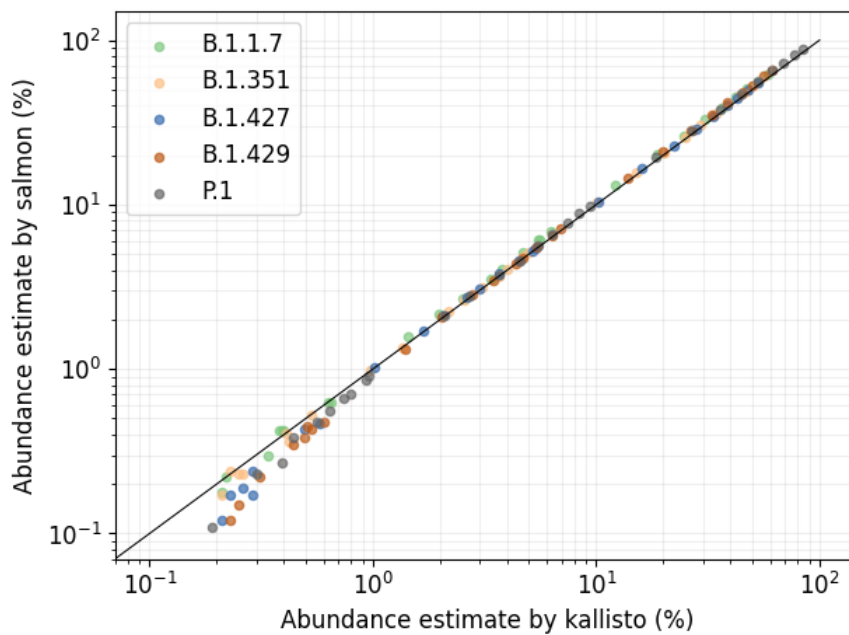
*Whole genome, depth = 100x*



601  
602  
603  
604  
605

Figure S3: Additional benchmarking results

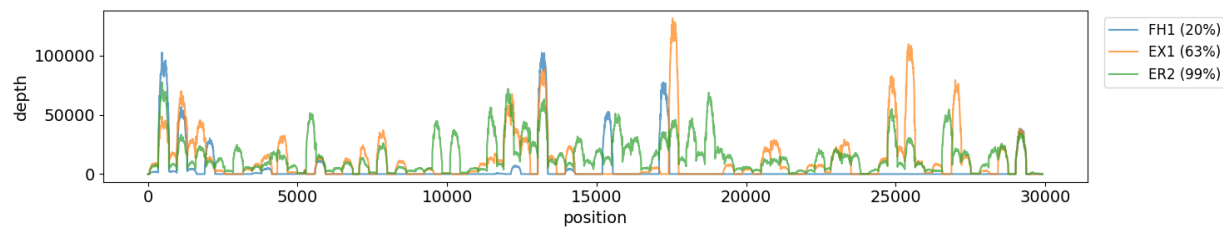
606  
607  
608



609  
610  
611

Figure S4: salmon versus kallisto abundance estimates per variant.

612

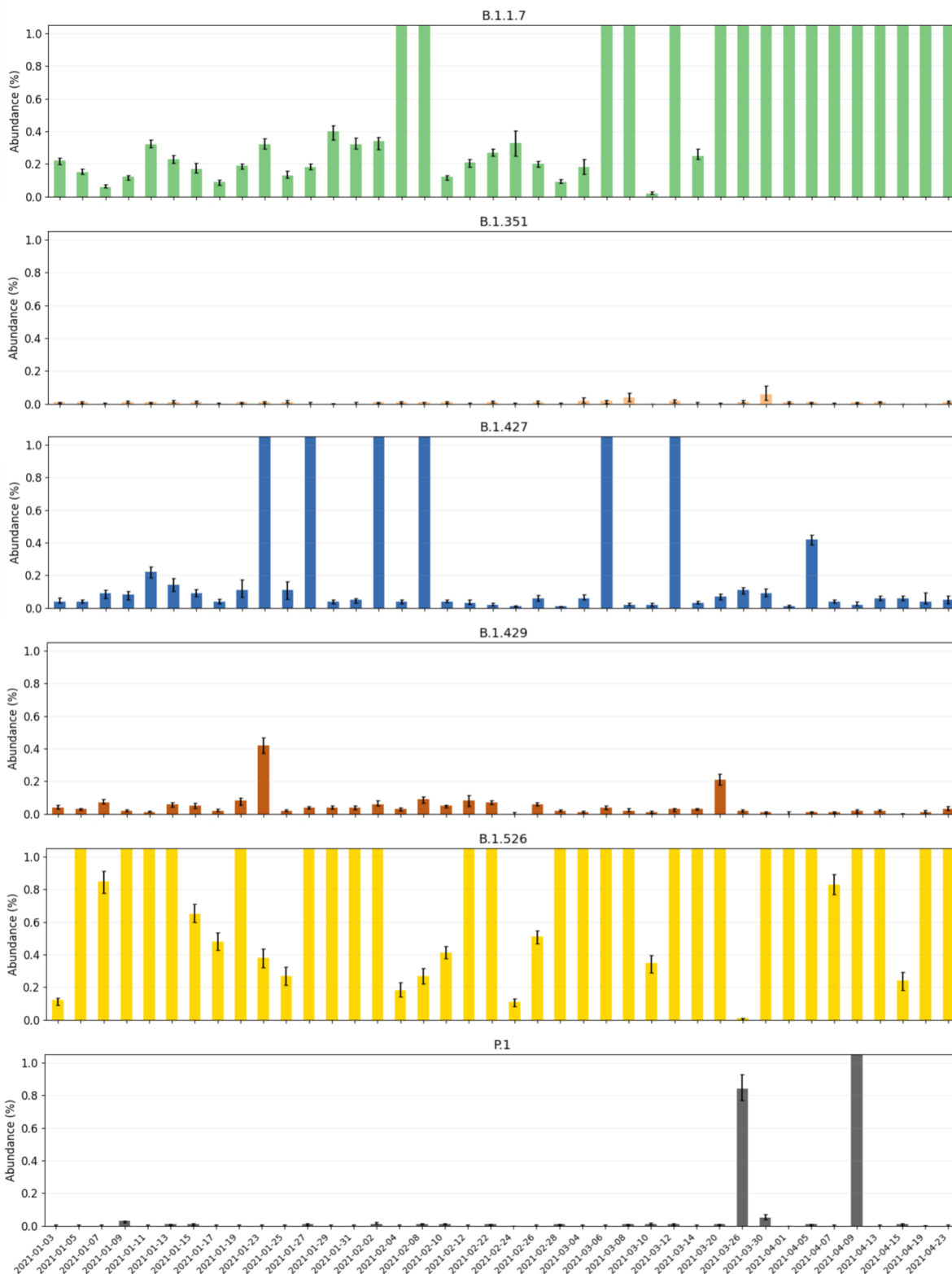


613

614 *Figure S5: Sequencing depth from wastewater samples is highly uneven between amplicons. This figure shows*  
615 *sequencing depth along the genome for three samples collected in New Haven, CT. The first sample (FH1) has low*  
616 *genome coverage (20%), with very few amplicons reaching high sequencing depth. The second sample (EX1) has*  
617 *moderate genome coverage (63%), with roughly half of the amplicons reaching high sequencing depth. The third*  
618 *sample (ER2) has high genome coverage (99%), with nearly all amplicons reaching high sequencing depth.*

619

620



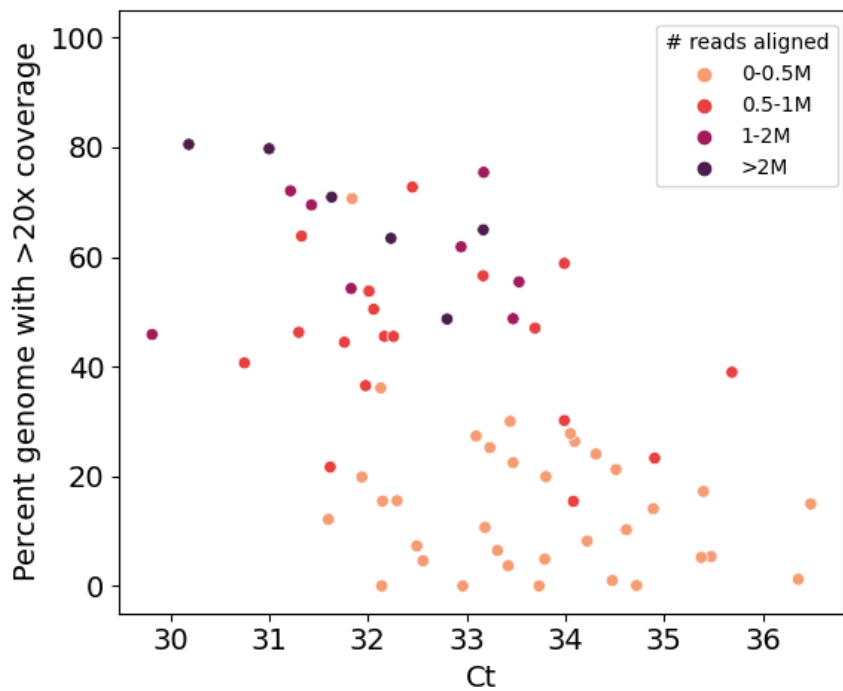
621

622

623 *Figure S6: Raw predictions per variant with confidence intervals based on bootstrap analysis for New Haven*

624 *samples. Note that in all subplots the y-axis is capped at 1% for improved readability.*

625  
626



627  
628 *Figure S7: Percent genome coverage versus Ct values for samples across the US*

629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640



641

642

643

Figure S8: Raw predictions per variant with confidence intervals based on bootstrap analysis for samples across the US. Note that in all subplots the y-axis is capped at 1% for improved readability.