Estimation of Proportions of SARS-CoV-2 Variants in a Mixed Sequencing Sample

Askar Gafurov[1], Andrej Baláž[2], Tomáš Vinař[2], and Broňa Brejová[1]

[1] Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia
[2] Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

Abstract: Estimation of proportions of SARS-CoV-2 genome variants (e.g. variant B.1.1.7 originating from Britain, variant B.1.351 originating from South-Africa) in a population is currently done by sequencing individual samples, which demands individual laboratory processing of each sample. This labor can be significantly reduced by mixing several samples together and processing them in one batch. Our project aims to estimate the proportion of samples with given variants from such mixtures using probabilistic modeling.

1 Introduction

Since December 2019, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is rapidly spreading across the world, causing the coronavirus infectious disease (COVID-19) pandemic. Up to April 5, 2021, there were 132 mil. cases of COVID-19 infection. Among these, 109 mil. cases have already had an outcome from which there was 2.8 mil. deaths, resulting in an estimate of the mortality rate around 2.5% [18].

Due to its worldwide spread and a mutation rate of about 2 mutations per month [9], the SARS-CoV-2 developed multiple variants. Some emerging variants have accumulated significantly more mutations and proved to be more dangerous, causing concerns around the globe [4]. Scientists estimate that continued transmission of SARS-CoV-2 and selective pressures, such as vaccines, are creating ideal conditions for additional significant virus evolution [11]. Therefore, it is critical to characterize the virus strains further and monitor the spread of the variants in the population in order to inform public policies and assess the effectiveness of containment strategies.

Monitoring of the SARS-CoV-2 variants requires conducting many sequencing experiments to determine the genome of the virus and to identify its mutations compared to a reference genome. Based on this, it is then possible to decide which variant was sequenced. Some part of the laboratory work is done individually for each sample, which is time-consuming and resource-intensive. In this work, we propose a model capable of estimating the proportions of different variants in a pooled sample, which allows to combine multiple samples into one sequencing experiment.

2 Problem description

2.1 Variants

In our experiments, we consider several variants of the SARS-CoV-2 virus, which are characterized by specific mutations described in the Table 1. For instance, if a particular SARS-CoV-2 genome has the nucleotide at position 3267 mutated from cytosine to thymine, it has one of the mutations characteristic for the variant which originated in the United Kingdom [14] and is known as a variant of concern alpha [16]. In the Pangolin SARS-CoV-2 lineage classification [15], it is denoted B.1.1.7. In this paper, we will denote this variant by acronym UK listed in the table. In our list of variants, we have included two additional variant of concern (beta and gamma) as well as several variants that had high prevalence in Slovakia in the fall of 2020 and early 2021, which is the time from which our data originate. For each variant, the table also contains the minimum number of these characteristic mutations which need to be observed in the genome in order to be considered as belonging to a given variant in our study. If a genome did not reach the number of characteristic mutations for any considered variant, it was assigned the label "other".

This characterization was used on all samples from the GISAID database (downloaded on Mar 4, 2021) [5], which provides a comprehensive resource of more than 650 thousands fully assembled SARS-CoV-2 genomes. Individual genomes were aligned to the SARS-CoV-2 reference `Wuhan/Hu-1/2019` by minimap2 [10]. Then, each sequence was assigned a label as described earlier. Finally, a matrix P was constructed, which consisted of entries $p_{k,i,a}$ representing the relative frequency of nucleotide a at genomic position i of variant k among aligned genomes.

This matrix thus characterizes each variant not only by the selected mutations listed in Table 1, but also captures any additional significant mutations present in the genomic sequences classified to the variant. This approach can be easily applied to a different set of variants by simply modifying the input table of variants and their characteristic mutations.

2.2 Sequencing data

In order to identify the SARS-CoV-2 variant of a new patient, it is necessary to sequence an obtained sam-

Variant	Acronym	Required	Mutations
P.1 (gamma)	Brazil	8	T733C, C2749T, C3828T, A5648C, C12778T,
B.1.258	CZ	4	G12988T, G15598A, G18028T, T24910C, T26972C
B.1.177	EU1	4	C22227T, C28932T, G29645T, G21255C, C26801G
B.1.160	EU2	3	C4543T, G5629T, G22992A, T26876C
B.1.351 (beta)	South.Afr	3	G23012A, A21801C, A23063T, G22813T
B.1.1.7 (alpha)	UK	11	C3267T, C5388A, T6954C, A23063T, C23271A, C23604A, C23709T, T24506G, G24914C, C27972T, G28048T, A28111G, C28977T C13860T, G17259T, C21614T, C21621A
B.1.1.170	UKBA318	3	G12824A, C25777T, G26062T, C29754T

Table 1: Characteristic mutations of individual variants considered in this study.

ple. In this work, the used samples were sequenced using the ARTIC amplicon protocol [17] and the Oxford Nanopore MinION sequencer.

In the ARTIC protocol, the RNA of the virus is first reverse transcribed into DNA and specific regions of the DNA are amplified. These amplicons cover almost the entire length of the genome with slight overlaps between adjacent regions. In our data, the amplicons had length about 1700 bp. Amplified DNA was sequenced by the MinION sequencer, and the resulting sequences are called reads. The reads are then mapped to the reference genome by tool minimap2 [10] and stored in the BAM format.

Because each read comes from a genome of a specific variant, it is expected that it contains mutations that are characteristic for this variant. When analyzing a mixture of reads from different patient samples, the task at hand is to estimate the proportions of reads coming from individual variants, based on the presence of mutations characteristic for these variants.

3 Related work

There had been several attempts to identify SARS-CoV-2 variants from mixed samples. Most commonly, these were wastewater-based epidemiological studies.

Crits-Christoph et al. [3] showed that single nucleotide variants (SNVs)[1] detected from the sewage water from San Francisco Bay Area were significantly similar to local California-based patient-derived genotypes, thus demonstrating the possibility of identifying local SARS-CoV-2 variants in the wastewater samples. The SNVs were obtained via SNV caller inStrain v1.3.2 [12] and the similarity of genotypes was established via Fisher's exact test.

Another study [6] analyzed 91 wastewater samples from 11 states in the USA and identified 7973 SNVs, of which 5680 were "novel" at the time of the analysis with respect to the global clinically derived data. Interestingly, almost half of the "novel" variants were confirmed within the next 5 months after the analysis. This suggests that sewage samples may provide a more comprehensive snapshot of currently circulating SARS-CoV-2 variants in comparison with solely clinical cases.

A wastewater study conducted in Spain [13] identified 238 SNVs and 6 deletions in comparison with the reference genome of SARS-CoV-2 isolate Wuhan-Hu-1. The study used 40 samples, which were sequenced with ARTIC protocol v.3, analysed using the iVAR software.

In the wastewater study from Switzerland [7], the authors showed that it is possible to use Illumina reads to detect the variants before they appear in clinical cases. The 48 samples were collected as 24-hours composite or grab samples. Amplicons were created using the ARTIC v.3 protocol and sequenced using Illumina NovaSeq 6000, resulting in paired reads of length 250bp. The mutations were identified from these reads using the V-pipe bioinformatics pipeline. These mutations were then compared with the clinical cases from Switzerland in the GISAID database for the presence of characteristic mutations of the B.1.1.7 and B.1.351 variants using Fisher's exact test. The authors also looked at mutations co-occurring in the same read and found further evidence of the presence of the B.1.1.7 variant in Switzerland already in early December.

Mixed samples can also originate from environmental sources other than sewage water. The first report on recovering near-complete SARS-CoV-2 genome sequences from environmental surface swabs assessed the contamination with the virus in a hospital [2]. The authors of this study confirmed the low likelihood that SARS-CoV-2 contamination on hospital surfaces contains infectious virus.

Although many studies have identified SARS-CoV-2 variants from mixed environmental samples, all have qualitative results. In this paper, we estimate the proportions of the variants in the sample, which we believe is correlated with the proportions of the variants circulating in the community.

[1] Single nucleotide variants are simple mutations substituting one DNA base for another, not to be confused with virus variants, which are groups of evolutionarily related viruses.

4 Model description

In our model, we assume that in a set of aligned reads coming from one variant, the expected observed symbol counts at any given position in genome are proportional to the symbol frequencies among all sequences in a given variant, which are captured in matrix P described in subsection 2.1. The observed symbol counts at a particular position can be thus described by a multinomial distribution. To simplify the model, we further assume that individual positions in the genome are independent. When sequencing a mixed sample, we assume that individual variants are present at some unknown proportions which are represented as weights in a mixture model. We assume that these proportions stay the same at all positions in the genome. A more detailed description of the model is given in the next subsection.

4.1 Basic model

Let $\Sigma = \{A,C,G,T\}$ be the symbol alphabet. Let K denote the number of virus variants and L denote the length of the reference genome. Let $W = (w_1,\ldots,w_K)$ denote the (unknown) weights of individual variants in a mixture. Let $p_{k,i,a}$ denote the probability of observing symbol a at position i in the k-th variant and $O_{i,a}$ denote the count of symbol a at position i in the observed reads.

The probability of observing symbol a at position i in a mixture with weights W is then equal to $m_{i,a}(W) := \sum_{k=1}^{K} w_k \cdot p_{k,i,a}$.

The total probability of observations O given mixture weights W is then proportional to:

$$Pr[O|W] \sim \prod_{i=1}^{L}\prod_{a\in\Sigma} m_{i,a}(W)^{O_{i,a}}.$$

The inference of the mixture weights from observations can be then solved via maximisation of the log-likelihood function:

$$W^* := \arg\max_W \sum_{i=1}^{L}\sum_{a\in\Sigma} O_{i,a}\cdot\log m_{i,a}(W).$$

This task is equivalent to minimisation of cross-entropy between m and O:

$$W^* := \arg\min_W -\sum_{i=1}^{L}\sum_{a\in\Sigma} O_{i,a}\cdot\log m_{i,a}(W)$$

4.2 Minimisation and efficiency

The minimisation is done using the L-BFGS-B algorithm [19] implemented in Python library `scipy` [8]. Since the input data for the minimisation process are essentially two tables of fixed sizes $L\times|\Sigma|$ and $K\times L\times|\Sigma|$, the time and memory requirements of the optimisation process itself do not depend on the amount of sequencing reads and neither do the memory requirements of data preprocessing, allowing for the efficient processing of gigabytes of sequencing data even on common computers. All processes are easily parallelizable and could be potentially accelerated using a GPU.

4.3 Adding error model

Sequencing reads contain errors, where the symbol in the read differs from the actual genome being sequenced. We assume that these errors occur uniformly at random and add them to the model as follows. Let $\varepsilon \in (0,1)$ be the substitution error rate. Let $\bar{m}_{i,a}(W) := 1 - m_{i,a}(W)$ denote the probability of observing a base different from base a at position i in a mixture sample without sequencing errors. The probability of observing symbol a at position i in a mixture sample sequenced with errors is then equal to:

$$q_{i,a}(W,\varepsilon) = (1-\varepsilon)\cdot m_{i,a}(W) + \frac{\varepsilon}{3}\cdot\bar{m}_{i,a}(W).$$

The likelihood of observing O given mixture weights W and substitution rate ε is then proportional to:

$$Pr[O|W,\varepsilon] \sim \prod_{i=1}^{L}\prod_{a\in\Sigma} q_{i,a}(W,\varepsilon)^{O_{i,a}}.$$

The error rate ε can be set to a particular value, or be inferred simultaneously with the mixture weights W:

$$(W^*,\varepsilon^*) := \arg\min_{W,\varepsilon} -\sum_{i=1}^{L}\sum_{a\in\Sigma} O_{i,a}\cdot\log q_{i,a}(W,\varepsilon)$$

In our experiments, we used the latter option.

4.4 Evaluation of a posteriori probability of a given read belonging to a particular variant

We can represent a read as a data set consisting of only one read, i.e. $O_{i,a}$ is equal to 1 if read has symbol a aligned to reference's position i, and 0 otherwise. Let $V \in \{1,\ldots,K\}$ be a random variable representing the variant that the read belongs to. The likelihood of a read belonging to variant k is equal to $Pr[O|V=k] = Pr[O|W=e_k]$, where e_k is the vector of length K with value 1 at position k and value 0 elsewhere. The probability of the read belonging to variant k given its sequence is then, by the Bayes theorem, equal to:

$$Pr[V=k|O] = \frac{Pr[O|V=k]\cdot Pr[V=k]}{\sum_{j=1}^{K} Pr[O|V=j]\cdot Pr[V=j]}.$$

Assuming a uniform prior $Pr[V] = \frac{1}{K}$, the formula reduces to:

$$Pr[V=k|O] = \frac{Pr[O|V=k]}{\sum_{j=1}^{K} Pr[O|V=j]}.$$

The notion of posterior probability enables us to classify individual reads into variants by choosing the variant

Variant	Isolate
CZ	UKBA-702, UKBA-722, UKBA-809, UKBA-818
EU1	UKBA-716, UKBA-717
EU2	UKBA-701
UK	UKBA-703, UKBA-704, UKBA-705, UKBA-706, UKBA-707, UKBA-708, UKBA-713, UKBA-714, UKBA-718, UKBA-719, UKBA-720, UKBA-801, UKBA-802, UKBA-803, UKBA-804, UKBA-805, UKBA-806, UKBA-807, UKBA-808, UKBA-814, UKBA-815, UKBA-816, UKBA-817
UKBA318	UKBA-709, UKBA-710, UKBA-711, UKBA-712, UKBA-723, UKBA-724
other	UKBA-715

Table 2: Table of used samples

with the maximum a posteriori probability (MAP). We use this read classification for visualization and filtering as discussed in the next section.

Unfortunately, this approach requires to set the substitution rate ε in advance. In our experiments, we estimated the substitution rate by first running the minimisation algorithm on the available data.

5 Experiments

5.1 Simulated data

To evaluate the model, we created simulated data sets, where the correct proportions of the variants were known. We have used 37 FASTQ files from the ARTIC sequencing experiments, conducted at Biomedical Research Center of the Slovak Academy of Sciences, using protocols described in [1], where each SARS-CoV-2 sample was sequenced separately (using a different barcode).

The list of samples and their classification to variants is shown in Table 2. Each file was mapped to the reference and the resulting BAM files were merged to create four mixed samples as follows:

1. 1x CZ, 1x EU1, 1x EU2, 1x UK (92 840 reads, average base coverage 5 858)
2. 4x CZ, 2x EU1, 1x EU2, 3x UK, 1x UKBA318 (264 113 reads, average base coverage 16 562)
3. 2x EU1, 1x EU2, 1x other (99 711 reads, average base coverage 6 273)
4. 3x CZ, 2x EU1, 9x UK, 1x other (335 601 reads, average base coverage 21 044)

The true proportions of variants in the samples is calculated as the proportion of the sums of read lengths for each variant.

5.2 Results on all reads

The results on the simulated mixes are shown in Table 3. We can see that the model is able to distinguish between present and absent variants, while significantly overestimating the "other" variant in all samples.

In order to find an explanation for this discrepancy, we analysed the posterior probabilities (see subsection 4.4) of individual reads (see Figure 1). The analysis shows that there are many "uncertain" reads (i.e. reads with no strong affinity towards any variant), particularly so among reads, classified into "other" variant.

We decided to formalize the notion of "uncertainty" and filter out such reads from the data set before running the model.

5.3 Filtering out uncertain reads

Let k-certainty of a read be defined as a sum of its k highest posterior probabilities of belonging to a particular variant.

We decided to use 1-certainty criterion in our experiments.

The pipeline works as follows: first we estimate the substitution error rate by running the model on the full data set. Then, using the estimated error rate, we evaluate the 1-certainty for each read. Reads with 1-certainty below a given threshold are then removed from the data set, and the observed counts $O_{i,a}$ are calculated. Finally, we run the model on these counts, reporting the mixture weights W as the final answer.

We ran our algorithm on the simulated mixes with different filtering thresholds. The estimated weights at different filtration thresholds are shown in Figure 2. The KL-divergence between the true and estimated weights are shown in Figure 3.

This heuristic works reasonably well on data sets without reads from "other" variant (Figures 2a and 2b). However, it fails on data sets with "other" variant present (Figures 2c and 2d) because reads from "other" variant have generally lower 1-certainty (Figure 1), and are consequently filtered out.

6 Discussion

We have demonstrated that our method can predict the proportions of individual SARS-CoV-2 variants in mixed samples when all variants in a sample are known to the model (i.e. there are no reads from "other" variant).

In our future work on this topic, we would like to address the following issues:

variant	Mix #1		Mix #2		Mix #3		Mix #4	
	true	estimated	true	estimated	true	estimated	true	estimated
Brazil	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
CZ	0.327	0.245	0.322	0.249	0.000	0.004	0.163	0.128
EU1	0.225	0.230	0.189	0.179	0.506	0.434	0.149	0.142
EU2	0.252	0.200	0.088	0.070	0.235	0.178	0.000	0.000
South.Afr	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000
UK	0.196	0.158	0.246	0.209	0.000	0.002	0.612	0.546
UKBA318	0.000	0.000	0.155	0.121	0.000	0.005	0.000	0.000
other	0.000	0.168	0.000	0.173	0.259	0.372	0.076	0.183

Table 3: Estimated weights for the simulated samples without any filtration. Note that the "other" variant is significantly overestimated in all samples. Aside from that, the method is able to distinguish between present and absent variants.

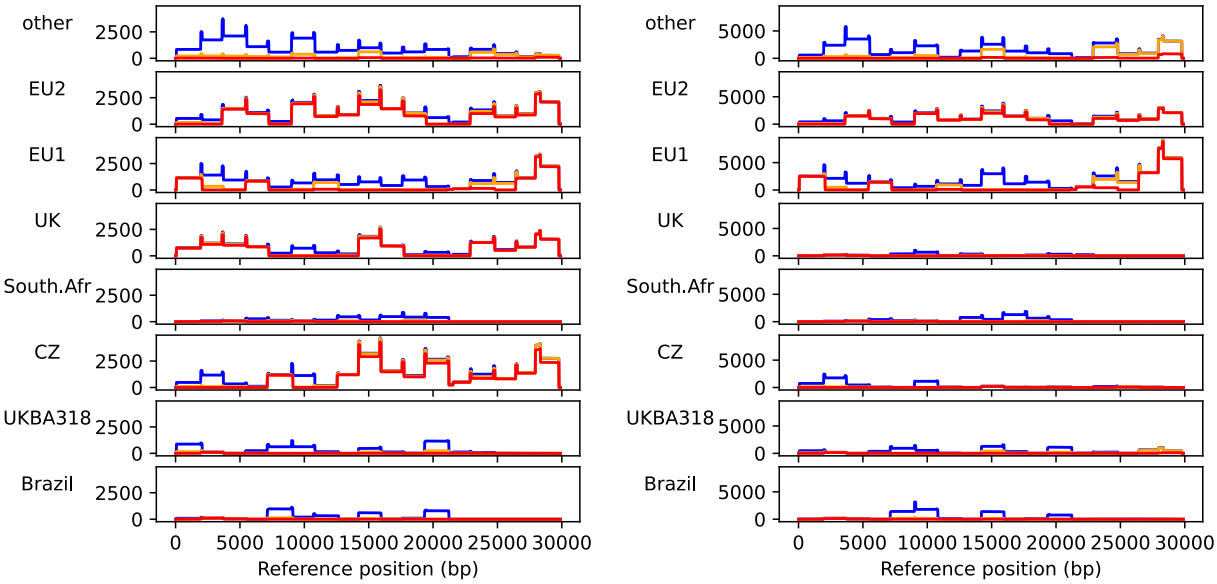
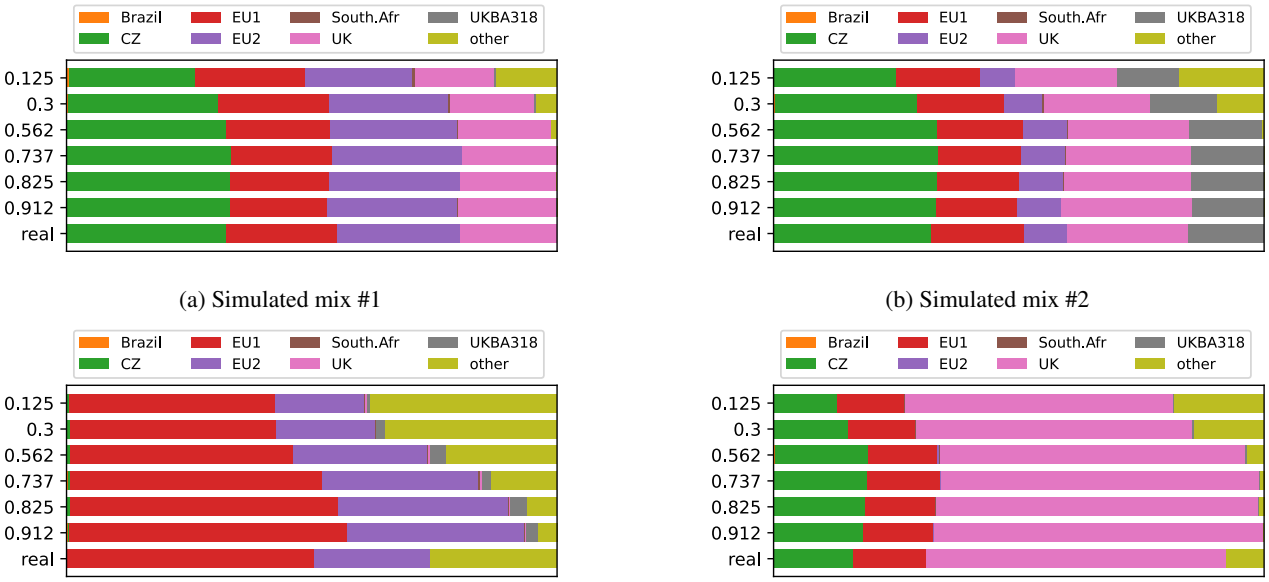


Figure 1: The base coverage of a sequence by reads classified into individual variants by maximum a posteriori predictor (MAP). The first simulated mix is shown in the left, the third in the right. The blue line represents the coverage by all reads classified into a particular variant, the orange and red lines only reads with 1-certainty above 50% and 90%, respectively. Reads that have been, both erroneously (first mix) and correctly (third mix) classified into "other" variant, have mostly low 1-certainty. Even for correct variants, there are regions in the reference where the model cannot capture any reads. These regions correspond well with the regions without characteristic mutations (see Table 1).



(a) Simulated mix #1

(b) Simulated mix #2

(c) Simulated mix #3

(d) Simulated mix #4

Figure 2: The estimated weights for each simulated mix at different thresholds of filtration of uncertain reads. The vertical axis shows the filtration threshold. Lengths of individual color stripes represent their estimated weight for a given threshold of filtration. The last row (denoted "real") shows the true weights for each simulated mix. The lowest filtration threshold does not filter out any reads (because it's equal to $1/K$, which is the lowest possible value for the 1-certainty).

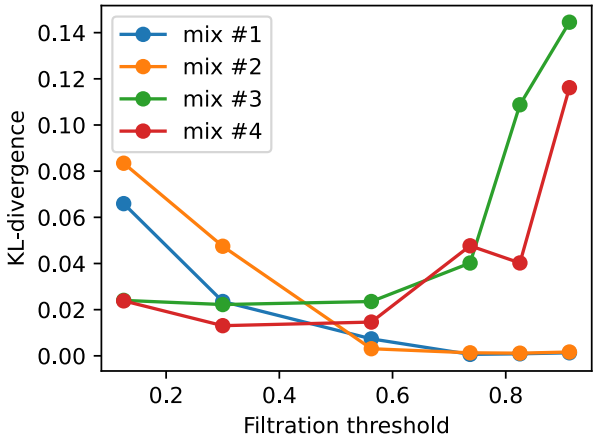


Figure 3: The KL-divergence between the true and estimated weights for different read filtration thresholds. The horizontal axis shows the filtration threshold. The vertical axis shows the resulting KL-divergence (lower is better). It can be seen that the filtration method works reasonably well for the mixes without "other" variant present (the first and second mixes).

- As we discussed in the previous section, our model tends to overestimate the "other" variant, and our current filtering heuristics is not a robust solution. The filtering may introduce bias into the estimation. This issue may be addressed by a more systematic approach to defining the "other" variant.
- We would like to be able to evaluate confidence intervals on our estimations. That would give us, for example, the ability to rule out the presence of a variant with a low estimated weight.
- There are many non-informative reads coming from the areas of the genome with no significant alterations between individual variants. The proportion of such reads is even higher when the reads are shorter (e.g. obtained by Illumina sequencing). The goal is to extend the model so that it could handle such data.
- The sequencing data shows severe non-uniformity of the coverage. This issue may be addressed by a selective weighting of individual reads based on the coverage at their position.
- The current model does not take into account insertions and deletions, both in the individual variants and during the sequencing process.
- The way we characterize the individual variants using the averaging of all available data from the GISAID database may incur a bias toward specific subvariants (e.g. some countries, such as the UK, submit a much higher number of sequence samples).
- Our approach splits all reads into individual bases, thus potentially losing all long-range information. In the future, we would like to adjust the model, so that it would keep the reads intact.

- In our model we assume that the individual positions are independent. But, the way the variant profiles are created (by demanding at least n mutations to be present in a sequence) introduces some dependency between them. In the future, we would like to resolve that inconsistency either by changing of the way the variant profiles are estimated or by relaxing the assumption in the model.

Code availability. The code is available at https://github.com/fmfi-compbio/covid-pooling.

Acknowledgements. We would like to thank the reviewers for their thorough reviews and helpful suggestions.

Funding. This research was supported by the Operational Program Integrated Infrastructure within project: Pangenomics for personalized clinical management of infected persons based on identified viral genome and human exome (Code ITMS:313011ATL7, 80%) co-financed by the European Regional Development Fund. It was also supported by VEGA 1/0458/18 (10%) and Comenius University Grant UK/336/2021 (10%).

References

[1] B. Brejova, K. Borsova, V. Hodorova, V. Cabanova, A. Gafurov, D. Fricova, M. Nebohacova, T. Vinar, B. Klempa, and J. Nosek. Nanopore sequencing of SARS-CoV-2: Comparison of short and long PCR-tiling amplicon protocols. *medRxiv*, 2021.

[2] D. A. Coil, T. Albertson, S. Banerjee, G. Brennan, A. J. Campbell, S. H. Cohen, S. Dandekar, S. L. Díaz-Muñoz, J. A. Eisen, T. Goldstein, et al. SARS-CoV-2 detection and genomic sequencing from hospital surface samples collected at UC Davis. *medRxiv*, 2021.

[3] A. Crits-Christoph, R. S. Kantor, M. R. Olm, O. N. Whitney, B. Al-Shayeb, Y. C. Lou, A. Flamholz, L. C. Kennedy, H. Greenwald, A. Hinkle, et al. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *medRxiv*, 2020.

[4] ECDC. Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom—20 december 2020. *ECDC: Stockholm*, 2020.

[5] S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.

[6] R. S. Fontenele, S. Kraberger, J. Hadfield, E. M. Driver, D. Bowes, L. A. Holland, T. O. Faleye, S. Adhikari, R. Kumar, R. Inchausti, et al. High-throughput sequencing of SARS-CoV-2 in wastewater provides insights into circulating variants. *medRxiv*, 2021.

[7] K. Jahn, D. Dreifuss, I. Topolsky, A. Kull, P. Ganesanandamoorthy, X. Fernandez-Cassi, C. Bänziger, E. Stachler, L. Fuhrmann, K. P. Jablonski, et al. Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples. *medRxiv*, 2021.

[8] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[9] K. Kupferschmidt. The pandemic virus is slowly mutating. but does it matter?, 2020.

[10] H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[11] NCIRD. Science brief: Emerging SARS-CoV-2 variants, 2021.

[12] M. R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B. A. Firek, M. J. Morowitz, and J. F. Banfield. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, pages 1–10, 2021.

[13] A. Pérez Cataluña, Á. Chiner-Oms, E. Cuevas Ferrando, A. Díaz-Reolid, I. Falcó, W. Randazzo, I. Girón-Guzmán, A. Allende, M. A. Bracho, I. Comas, et al. Detection of genomic variants of SARS-CoV-2 circulating in wastewater by high-throughput sequencing. 2021.

[14] A. Rambaut et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, 2020.

[15] A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5(11):1403–1407, 2020.

[16] The World Health Organization. Tracking SARS-CoV-2 variants, 2021. `https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/`.

[17] J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv*, 2020.

[18] M. Wasiuta. Worldometers, 2009.

[19] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.