

# SUP: A Probabilistic Framework to Propagate Genome Sequence Uncertainty, with Applications

Devan Becker<sup>1,2,§,\*</sup>, David Champredon<sup>2,§</sup>, Connor Chato<sup>1</sup>, Gopi Gudan<sup>1</sup>, and Art Poon<sup>1</sup>

<sup>1</sup>Public Health Agency of Canada - National Microbiology Laboratory - Public Health Risk Sciences Division

<sup>2</sup>Department of Pathology and Laboratory Medicine, Schulich School of Medicine and Dentistry, Western University

§ equal contribution

## Abstract

Genetic sequencing is subject to many different types of errors, but most analyses treat the resultant sequences as if they are known without error. Next generation sequencing methods rely on significantly larger numbers of reads than previous sequencing methods in exchange for a loss of accuracy in each individual read. Still, the coverage of such machines is imperfect and leaves uncertainty in many of the base calls. In this work, we demonstrate that the uncertainty in sequencing techniques will affect downstream analysis and propose a straightforward method to propagate the uncertainty.

Our method (which we have dubbed Sequence Uncertainty Propagation, or SUP) uses a probabilistic matrix representation of individual sequences which incorporates base quality scores as a measure of uncertainty that naturally lead to resampling and replication as a framework for uncertainty propagation. With the matrix representation, resampling possible base calls according to quality scores provides a bootstrap- or prior distribution-like first step towards genetic analysis. Analyses based on these re-sampled sequences will include a more complete evaluation of the error involved in such analyses.

We demonstrate our resampling method on SARS-CoV-2 data. The resampling procedures adds a linear computational cost to the analyses, but the large impact on the variance in downstream estimates makes it clear that ignoring this uncertainty may lead to overly confident conclusions. We show that SARS-CoV-2 lineage designations via Pangolin are much less certain than the bootstrap support reported by Pangolin would imply and the clock rate estimates for SARS-CoV-2 are much more variable than reported.

1 Generating a genetic sequence from a biological sample is a complex process. Nucleic  
2 acids must be extracted from the sample while avoiding contamination by foreign mate-  
3 rial. If working with RNA, then we must use a reverse transcriptase reaction (which has  
4 a high base misincorporation rate) to convert the RNA into DNA. Polymerase chain reac-  
5 tion (PCR) amplification is often employed to enrich the sample for the target of interest.  
6 For next-generation sequencing (NGS) protocols, we have to generate a sequencing library,  
7 for instance by random shearing of nucleic acids into fragments that are ligated onto special  
8 “adaptors”. NGS procedures such as sequencing by synthesis suffer from greater error rate

1 relative to conventional Sanger dye-terminator sequencing, although these rates have contin-  
2 ued to improve with new technologies (Fuller et al., 2009; Goodwin et al., 2016; Salk et al.,  
3 2018). In addition, the short reads produced by NGS platforms need to be aligned — either  
4 by alignment against a reference genome, *de novo* assembly, or a combination of the two —  
5 to reconstruct a consensus sequence using one or more bioinformatic programs. Errors can  
6 be introduced in any one of these steps (Beerenwinkel and Zagordi, 2011; O’Rawe et al.,  
7 2015).

8 In some cases, naturally occurring variation, *i.e.*, genetic polymorphisms, or variation in-  
9 duced by experimental error is directly quantified and encoded into the output. For example,  
10 mixed peaks in sequence chromatograms produced from dye-terminator sequencing by cap-  
11 illary electrophoresis are assigned standard IUPAC codes (*e.g.*, Y for C or T) when the base  
12 calling program cannot determine which base is dominant (NC-IUB, 1986). Ewing and Green  
13 (1998) and Richterich (1998) both argued that estimates of the base call quality, quantified as  
14 Phred quality scores, can be an accurate estimate of the number of errors that the machines  
15 at the time would make, but improvements to these error probabilities have been proposed  
16 (Li et al., 2004, 2009b). Nevertheless, Phred scores remain the standard means of reporting  
17 the estimated error probabilities for current sequencing platforms. Generally, these scores  
18 are either used to censor the base calls (*i.e.*, label them “N” rather than A, T, C or G) if the  
19 estimated probability of error exceeds a predefined threshold or remove the sequence from  
20 further analysis if the total number of censored bases exceeds a maximum tolerance (*e.g.*,  
21 Doronina, 2005; Robasky et al., 2014; O’Rawe et al., 2015). Some authors/tools use more  
22 sophisticated models, such as Wu et al. (2017) who use statistical models that incorporate  
23 read depth to determine a probability of a sequencing error, but still use the resultant reads to  
24 form a consensus sequence with no measure of uncertainty. Furthermore, some studies have  
25 extended the concept of per-base error probabilities to calculate the joint likelihoods of partial  
26 or full sequences. For example, DePristo et al. (2011) and Gompert and Buerkle (2011) incor-  
27 porate adjusted Phred scores into a likelihood framework to generate more accurate estimates  
28 of genetic diversity within a population; this approach has subsequently been used to develop

1 new estimators of genetic diversity (Fumagalli et al., 2013). Kuo et al. (2018) recently used a  
2 similar approach to develop a statistical test of whether a given genome sequence is consistent  
3 with a specified alternative sequence. In general, the reported error probabilities from NGS  
4 technologies are primarily used for filtering low quality sequences and improving alignment  
5 algorithms (which both result in a consensus sequence that is assumed to be error-free) or for  
6 hypothesis tests concerning small collections (usually pairs) of sequences.

7 The uncertainty present in the sequences are most often ignored entirely. For example,  
8 methods for sequence alignment and homology searches generally employ heuristic algo-  
9 rithms that utilize similarity scores that do not explicitly incorporate the probabilities of  
10 sequencing errors. The problem of unacknowledged uncertainty is exacerbated when each  
11 sequence represents the consensus of diverse copies of a genome, such as rapidly evolving  
12 virus populations where genuine polymorphisms are confounded with sequencing error. See  
13 Schneider (2002) for more criticisms of the use of consensus sequences, along with visualiza-  
14 tions (Schneider and Stephens, 1990, called *sequence logos*) to display the deviations from a  
15 consensus.

16 Though rare, some studies have proposed methods for propagation of uncertainty from  
17 one step to later steps of an analysis. O’Rawe et al. (2015) suggest methods for propagation  
18 of sequence-level uncertainty into determining whether two subjects have the same alleles,  
19 as well as estimating confidence intervals for allele frequencies. Another exception can be  
20 found in Kuhner and McGill (2014), who incorporate an assumed or estimated error rate  
21 for the entire sequence into the calculation of a phylogenetic tree and found that incorpora-  
22 tion of errors makes the inferred branch lengths much closer to the true (simulated) branch  
23 lengths. Though they did not use nucleotide-level uncertainty, Gompert and Buerkle (2011)  
24 incorporate the coverage of NGS technologies as part of the uncertainty of estimates for the  
25 frequency of alleles in a population. Clement et al. (2010) present an alignment algorithm  
26 (called GNUMAP) that takes nucleotide-level uncertainty into account. Their method incor-  
27 porates Position Weight Matrices into a method of scoring multiple possible matches against  
28 a reference genome in order to choose the best alignment. These studies are the exceptions,

1 rather than the rules, and their methods have not yet attained widespread use.

2 We present a simple general-purpose framework that can be incorporated into any analy-  
3 sis of genetic sequence data. This framework involves converting the uncertainty scores into  
4 a matrix of probabilities, and repeatedly sampling from this matrix and using the resultant  
5 samples in downstream analysis. Unlike likelihood-based approaches, we do not make as-  
6 sumptions about the underlying patterns or distributions in the data. In so doing, we can gain  
7 more accurate estimation of the errors at the expense of computation time. Our technique is  
8 amenable to quality score adjustments prior to applying our methods. We demonstrate the  
9 impact of propagating sequence uncertainty by applying our methods to the problem of clas-  
10 sifying SARS-CoV-2 genomes into predefined clusters known as “lineages” (Rambaut et al.,  
11 2020), several of which correspond to variants carrying mutations that are known to confer an  
12 advantage to virus transmission or infectivity. We also analyse a collection of SARS-CoV-2  
13 sequences to demonstrate that the estimated rate of new mutations is much more variable than  
14 studies relying on deterministic sequences would conclude.

## 15 **1 Methods**

### 16 **1.1 Probabilistic representation of sequences**

17 Here, we describe two theoretical frameworks to model sequence uncertainty at the *nu-*  
18 *cleotide level* or at the *sequence level*. In both frameworks, the sequence of nucleotides  
19 from a biological sample is not treated as a single unambiguous observation (known without  
20 error), but rather as a collection of possible sequences weighted by their probability.

### 1.1.1 Nucleotide-level uncertainty

To represent the uncertainty at each position along the genome we introduce the following matrix, which we will refer to as a probabilistic sequence and denote as  $\mathcal{S}$ :

$$\mathcal{S} = \begin{matrix} & 1 & 2 & \dots & \ell \\ \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \\ - \end{matrix} & \begin{pmatrix} \mathcal{S}_{A,1} & \mathcal{S}_{A,2} & \dots & \mathcal{S}_{A,\ell} \\ \mathcal{S}_{C,1} & \mathcal{S}_{C,2} & \dots & \mathcal{S}_{C,\ell} \\ \mathcal{S}_{G,1} & \mathcal{S}_{G,2} & \dots & \mathcal{S}_{G,\ell} \\ \mathcal{S}_{T,1} & \mathcal{S}_{T,2} & \dots & \mathcal{S}_{T,\ell} \\ \mathcal{S}_{-,1} & \mathcal{S}_{-,2} & \dots & \mathcal{S}_{-,\ell} \end{pmatrix} \end{matrix} \quad (1)$$

Each column represents a position in a nucleotide sequence of length  $\ell$ . Each row represents one of the four nucleotides  $\text{A}, \text{C}, \text{G}, \text{T}$ , as well as an empty position “ $-$ ” that symbolizes a recorded deletion rather than missing data. Hence,  $\mathcal{S}$  is a  $5 \times \ell$  matrix.

The elements of the probability sequence represent the probability that a nucleotide exists at a given position, with a special case for the empty position  $-$ :

$$\mathcal{S}_{n,j} = \begin{cases} \mathbb{P}(\text{nucleotide } n \text{ is at position } j) & \text{if } n \in \{\text{A}, \text{C}, \text{G}, \text{T}\} \\ \mathbb{P}(\text{empty position } j) & \text{if } n = - \end{cases} \quad (2)$$

Note that we have for all  $1 \leq j \leq \ell$ :

$$\sum_n \mathcal{S}_{n,j} = 1 \quad (3)$$

Also, the sequence length is stochastic if  $0 < \mathcal{S}_{-,i} < 1$  for at least one  $i$ . The nucleotide (or deletion) drawn at each position is independent from all the others, so there are up to  $5^\ell$  possible different sequences for a given probabilistic nucleotide sequence, but these sequences are *not* equally probable.

A major limitation of this probabilistic representation of a sequence is that we lose all in-

1 formation on linkage disequilibrium. This is especially problematic for recording insertions  
2 because insertions with  $L \geq 2$  nucleotides are treated as  $L$  independent single nucleotide  
3 insertions. Instead, we assume that every nucleotide is an independent observation. For  
4 example, a probability sequence populated from short read data from a diverse population  
5 would not store the information that two polymorphisms were always observed in the same  
6 reads, *i.e.*, in complete linkage disequilibrium. We also lose information about autocorre-  
7 lation in sequencing error, such as clusters of miscalled bases associated with later cycles  
8 of sequencing-by-synthesis platforms. Sequence chromatograms and base quality scores are  
9 affected by the same loss of information.

10 We note that this representation is similar to the “CATG” file type as described in Kozlov  
11 (2018), which indicates the likelihoods of each nucleotide in an aligned mapping for multiple  
12 taxa. This file type is able to be used by RAXML-NG to estimate an overall error rate which is  
13 then used to estimate phylogenetic trees. A reviewer has pointed out that the `bio++` library  
14 contains parsers for a probabilistic version of the FASTA format, called PASTA. We have  
15 not found documentation for this format, but are hopeful that our methods promote greater  
16 use of probabilistic formats like this. Our probability sequence is also similar in concept  
17 to Position Weight Matrices (PWMs, Stormo et al., 1982) which are built according to the  
18 frequency of each base at each position of a multiple alignment. Our construction differs  
19 in that we are creating one matrix per sequence where the entries are weighted according to  
20 error probability within that sequence, rather than one matrix for a collection of sequences.  
21 However, methods that accept PWMs will be applicable to our probability sequences (and  
22 *vice-versa*).

23 It is also possible to determine the sequence-level uncertainty as the product of nucleotide  
24 uncertainties for all possible sequences. This could be useful for creating an ordered list of  
25 the most likely sequences or removing any sequences that are not biologically plausible (*e.g.*,  
26 sequences missing a crucial amino acid, especially a start or stop codon). A full discussion  
27 of this is in the supplementary materials.

### 1.1.2 Sequence-level uncertainty

A significant problem of storing probabilities at the level of individual nucleotides is that generating a sequence from this matrix requires drawing  $\ell$  independent outcomes. For example, the reference SARS-CoV-2 genome is 29,903 nucleotides, and a substantial number of naturally-occurring sequence insertions have been described. Thus it would not be surprising if  $\ell$  exceeded 30,000 nucleotides (nt). The majority of these technically possible  $5^\ell$  sequences are not biologically plausible. Therefore, we formulate an ordered subset  $\mathcal{B} = (\mathcal{B}_i)_{i \in \{1 \dots m\}}$  of the first  $m$  most likely sequences, which are ranked in descending order by the joint probability of nucleotide composition. Note that the sequences in  $\mathcal{B}$ ,  $\mathcal{B}_i$ , do not necessarily have the same length. The observed genetic sequence,  $s^*$ , is a sample from a specified discrete probability distribution  $a$ :

$$\mathbb{P}(s^* = \mathcal{B}_i | i \dots m) = a(i) \quad (4)$$

This compact and approximate representation drastically reduces the number of operations to one sample, after some pre-processing to calculate  $a$ . The observed plurality sequence  $s^*$  (the sequence consisting of the most likely base at each position) is guaranteed to be a member of  $\mathcal{B}$  if  $\mathcal{S}_{s(j),j} > 0.5 \ \forall j$  where  $s(j)$  is the  $j$ -th nucleotide of  $s^*$ ; indeed, it is guaranteed to be the highest ranked member  $i = 0$ . We refer to any member of the set  $\mathcal{B}$  as a *sequence-level probabilistic sequence*. Note that because  $a$  is a probability distribution, we must have  $\sum_{i=1}^m a(i) = 1$ . In other words, this probability is conditional on the sequence being in  $\mathcal{B}$ .

For example, suppose that we have the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \\ - \end{matrix} & \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0 & 0.3 & 0.9 & 0 \\ 0 & 0 & 0.01 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (5)$$

such that there are  $2 \times 3 \times 2^3 \times 3 = 144$  possible sequences. The most likely sequence has the highest joint nucleotide probability: **ACATGA** with probability 0.2694 ( $0.9 \times 0.8 \times 0.99 \times 0.7 \times 0.9 \times 0.6$ ). If there is a positive probability of deletion for at least one position, then the sequence has a variable length. Large genomes or sequencing targets will result in vanishingly small probabilities for all sequences, and thus calculations on the log scale may be necessary to reduce the chance of numerical underflow.

Table 1 demonstrates the calculation of sequence-level uncertainties using the values in (5). The probability column is the product of the matrix entries for each nucleotide. If the four sequences shown are the only biologically plausible sequences, then the normalized probabilities can be expressed as  $a(i)$ .

sequence	probability	$a(i)$
$\mathcal{B}_1 = \text{ACATGA}$	0.299	$a(1) = 0.467$
$\mathcal{B}_2 = \text{ACATGT}$	0.150	$a(2) = 0.233$
$\mathcal{B}_3 = \text{ACAGGA}$	0.128	$a(3) = 0.200$
$\mathcal{B}_4 = \text{ACAGGT}$	0.064	$a(4) = 0.100$

Table 1: Biologically plausible sequences with probabilities defined by (5)

In summary, sequence-level probabilistic sequences offer a convenient way to define a (much) smaller set of possible sequences than the potential  $5^\ell$  nucleotide-level probabilistic sequences. This set will be used to generate sequences randomly for downstream analyses. The size of this set (noted  $m$  above) is arbitrarily determined by users.

## 1.2 Constructing the probability sequence

In most next-generation sequencing applications, the estimated probability of sequencing error is quantified with the quality (or “Phred”) score attributed to each base call produced by sequencing instrument. The quality score  $Q$  is directly related to this estimated error probability:  $\epsilon = 10^{-Q/10}$  (Ewing and Green, 1998), where  $Q$  typically ranges between 1 and 60 (with 60 being the lowest probability of error), depending on the sequencing platform and version of base-calling software. It is important to note that this quality score only



1 measures the probability of error from the machine;  $1 - \epsilon$  is an estimate of the probability of  
2 no sequencing errors and does not account for any other source of error.

3 More formally, the probability that the base call is correct is expressed as:

$$\mathbb{P}(\text{nucleotide} = X \mid \text{observed nucleotide} = X) = 1 - \epsilon \quad (6)$$

4 Unfortunately, quality scores have no information on the probabilities of the three other pos-  
5 sible nucleotides if the base call is incorrect. In the absence of information about the other  
6 bases (such as with consensus-level FASTQ or FASTA files), we assume that these other  
7 probabilities are uniformly distributed.

8 Raw short read data are typically recorded in a FASTQ format that stores both the se-  
9 quences (base calls) and base-specific quality scores for each short read. Since the reads  
10 often correspond to different positions of the target nucleic acid, *e.g.*, randomly sheared ge-  
11 nomic DNA, it is necessary to align the reads to identify base calls on different reads that  
12 represent the same genome position. This alignment step can be accomplished by mapping  
13 reads to a reference genome, by the *de novo* assembly of reads, or a hybrid approach that  
14 incorporates both methods. The aligned outputs are frequently recorded in the tabular Se-  
15 quence Alignment/Map (SAM) format (Li et al., 2009a). Each row represents a short read,  
16 including the raw nucleotide sequence and quality strings; the optimal placement of the read  
17 with respect to the reference sequence (as an integer offset); and the compact idiosyncratic  
18 gapped alignment report (CIGAR) string, an application-specific serialization of the edit op-  
19 erations required to align the read to the reference. The SAM format contains much more  
20 information (<https://samtools.github.io/hts-specs/SAMv1.pdf>), but for our purposes we only  
21 need the placement, sequence, quality, and CIGAR string.

22 We employed the following procedure to construct the nucleotide-level probabilistic se-  
23 quence from the contents of a SAM file. We initialize aligned sequence and quality strings  
24 with ‘-’ in all positions before the first read and after the last read, and ‘!’, which corresponds  
25 to a quality score of 0 ( $Q = 0$ ), to all other positions. Next, we tokenize the CIGAR string

into length-operation tuples, which determine how bases and quality scores from the raw strings are appended to the aligned versions. Deleted bases ('D' operations) are not assigned Phred scores, so we assume them to have 0 error probability. The overall process for constructing the probabilistic sequence is demonstrated in Figure 1, including our procedure for including paired-end reads which is explained in a subsequent section.

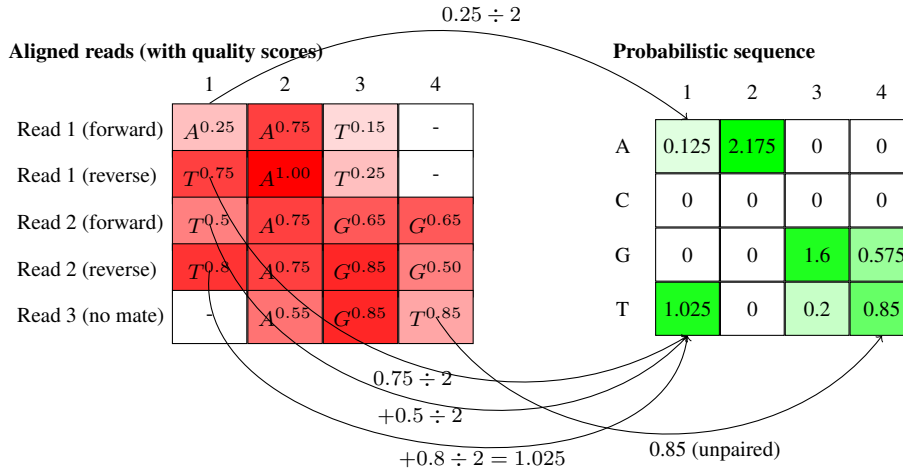


Figure 1: An illustration of constructing a probabilistic sequence from a SAM file. Each row in the matrix on the left is a graphical representation of a short read, and the superscript represents the quality score (from 0 to 1). Half of the quality score from paired end reads is added to the relevant cell in the matrix on the right. In both matrices, the column numbering represents a position on the reference genome. In the probabilistic sequence, the consensus sequence would be TAGT, but TAGG is also a very likely sequence given the quality scores.

### 1.3 Deletions and Insertions

By construction, the nucleotide-level probabilistic sequence would need to be defined with its longest possible length, *i.e.*, a multiple alignment for all reads. Deletions are naturally modelled with our representation but insertions would have to be modelled using deletion

1 probabilities.

$$\mathcal{S} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \\ - \end{matrix} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 1 \\ 0 & 0.01 & 0 & 0.99 & 0 & 0 \end{pmatrix} \end{matrix} \quad (7)$$

2 The low deletion probability for position 2 is straightforward to interpret: in about 1% of  
 3 the reads that contained this position, nucleotide **G** at position 2 is deleted. The high deletion  
 4 probability for position 4 means there is a 1% chance of a **T** insertion at this position (Table 2).

sequence	probability
$\mathcal{B}_1 = \text{CGAAT}$	$a(1) = 0.9799$
$\mathcal{B}_2 = \text{CAAT}$	$a(2) = 0.01$
$\mathcal{B}_3 = \text{CGATAT}$	$a(3) = 0.01$
$\mathcal{B}_4 = \text{CATAT}$	$a(4) = 0.0001$

Table 2: Sequence-level probabilistic sequence defined by (7)

5 This probability sequence is non-trivial to construct. Consider a short read with two bases  
 6 inserted at position  $j$  (say, an **A** at position  $j + 1$  and a **T** at position  $j + 2$ ) and a short read  
 7 with one insertion at position  $j$  (say, a **C**). It is entirely ambiguous whether the single insertion  
 8 (**C**) aligns with the first insertion (**A**) or the second insertion (**T**) of the first short read. This  
 9 is problematic for building up the matrix from reads aligned to the reference sequence. It  
 10 is conceptually and computationally simpler to start from a populated matrix and sampling  
 11 insertions. For our purposes, we only consider the pairwise alignment of these sequences  
 12 with a reference sequence and thus do not consider insertions.

## 13 1.4 Paired-End Reads

14 Some NGS platforms (*e.g.*, Illumina) use paired-end reads where the same nucleic acid tem-  
 15 plate is read in both directions. In these situations, we simply adjust all values by a factor of

one half. For bases where the paired-end reads overlap, this has the effect of averaging the  
 base probability  $1 - \epsilon$ . For example, if  $1 - \epsilon$  is 90% for **A** in one read and 95% **A** in its mate,  
 then 0.925 is added to the **A** row in  $S'$  (with the remaining 0.075 uniformly distributed across  
 the other nucleotides). If the two reads were 70% **A** and 55% **C** at the same position, then we  
 would increment the corresponding column vector (**A**, **T**, **C**, **G**) by  $(0.7/2, 0.1/2, 0.1/2, 0.1/2)$   
 for the first read and  $(0.15/2, 0.15/2, 0.55/2, 0.15/2)$  for the second, resulting in an addition  
 of  $(0.425, 0.125, 0.325, 0.125)$  for this pair. Bases outside of the overlapping region con-  
 tribute a maximum of 0.5 to  $S'$ , because the base call on the other read is missing data. This  
 approach has the advantage of making the parsing of SAM files trivially parallelizable since  
 we do not need to know how reads are paired. In addition, the coverage calculated from  $S'$  is  
 scaled to the number of templates rather than the number of reads.

## 1.5 Consensus Sequence FASTQ and FASTA Files

### 1.5.1 Consensus sequence FASTQ files

Full length or partial genome sequences are now frequently the product of next-generation  
 sequencing, by taking the consensus of the aligned or assembled read data. However, the  
 original read data are often not published alongside the consensus sequence. For example, on  
 September 30, 2022, there were nearly 390,000 SARS-CoV-2 consensus genome sequences  
 available in the Canadian VirusSeq Data Portal. None of the raw NGS data sets associated  
 with these consensus sequences are distributed in this database, however. Less than 6,700  
 (about 1.7%) raw SARS-CoV-2 FASTQ files for samples collected in Canada have been pub-  
 lished on the NCBI Sequence Read Archive. On the other hand, some consensus sequences  
 are released in a format where the bases are annotated with quality scores, *e.g.*, FASTQ. There  
 are several programs that provide methods to convert a SAM file into a consensus FASTQ  
 file (Li et al., 2004; Keith et al., 2002; Li et al., 2008). These programs use slightly differ-  
 ent methods for generating consensus quality scores, but filter quality scores for the majority  
 base. For example, suppose there are three reads with the following base calls at position  $j$ : **A**

1 with  $Q = 30$ , **A** with  $Q = 31$ , and **C** with  $Q = 15$ . Calculation of the consensus quality score  
 2 will thereby exclude the  $Q = 15$  value and report a quality score calculated from  $Q = 30$  and  
 3  $Q = 31$ , with the details of the calculation differing by software.

4 This omission makes it challenging for us to generate an  $\mathcal{S}$  matrix from a consensus  
 5 FASTQ file. Given the consensus base and its associated quality score at position  $j$ , we must  
 6 assume that the other bases are all equally likely with probability  $\epsilon_j/3$  (similar to Kuo et al.  
 7 (2018) and Chapter 5 of Kozlov (2018)). For example, let's assume the output sequence after  
 8 fragment sequencing and alignment is **ACATG** and its associated quality scores are respec-  
 9 tively  $Q = (60, 30, 50, 10, 40)$ . The probabilistic sequence is:

$$\mathcal{S} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \\ - \end{matrix} & \begin{pmatrix} 1 - 10^{-6} & 10^{-3}/3 & 1 - 10^{-5} & 10^{-1}/3 & 10^{-4}/3 \\ 10^{-6}/3 & 1 - 10^{-3} & 10^{-5}/3 & 10^{-1}/3 & 10^{-4}/3 \\ 10^{-6}/3 & 10^{-3}/3 & 10^{-5}/3 & 10^{-1}/3 & 1 - 10^{-4} \\ 10^{-6}/3 & 10^{-3}/3 & 10^{-5}/3 & 1 - 10^{-1} & 10^{-4}/3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (8)$$

10 Usually, the genetic sequence **ACATG** would be considered as certain and quality scores dis-  
 11 carded. In contrast, the probability of the sequence **ACATG** is only 0.899 within the proba-  
 12 bilistic sequence framework.

13 Incorporating deletions in the absence of raw data is also challenging. If one is willing to  
 14 assume a global deletion rate, then it is possible to extend the parameterization of  $\mathcal{S}$ . For ex-  
 15 ample, if the probability of a single nucleotide deletion is  $d$ , then the probability of the called  
 16 base is  $(1 - d_j)(1 - \epsilon_j)$  and the other three nucleotides have probability  $(1 - d)\epsilon_j/3$ . Hence,  
 17 if we assume the base call is **A**, the column of the nucleotide-level probabilistic sequence for

1 that position is

$$\mathcal{S}(, j) = \begin{matrix} & j \\ \text{A} & (1-d)(1-\epsilon_j) \\ \text{C} & (1-d)\epsilon_j/3 \\ \text{G} & (1-d)\epsilon_j/3 \\ \text{T} & (1-d)\epsilon_j/3 \\ - & d \end{matrix} \quad (9)$$

2 Since the FASTQ file only has a single sequence, we do have the same issues with align-  
 3 ment of differing lengths of insertions. In fact, insertions are only insertions relative to the  
 4 reference sequence; they can simply be treated as observed nucleotides with an associated  
 5 quality score. It would be possible to give insertions special treatment, however, by defining  
 6 a global insertion rate. This insertion rate can be expressed as a deletion rate relative to the  
 7 observed sequence, and thus one minus the insertion rate can be treated as the deletion rate  
 8 in the probabilistic sequence. As with the deletion rate, this requires an assumption about a  
 9 global rate which may be arbitrary.

10 A primary use of the probability sequence created from these FASTQ files would be to  
 11 construct a probability sequence as a reference genome for a given category. This would  
 12 entail collecting all available FASTQ files for a given lineage designation and using them in  
 13 the construction of a probability sequence as if they were short reads in a SAM file. From  
 14 here, lineage designation for a newly acquired sequence (and its probability sequence) could  
 15 be performed via a hypothesis test for whether the probability sequences are sufficiently  
 16 similar.

### 17 **1.5.2 Consensus sequence FASTA files**

18 If we do not have access to any base quality information, *e.g.*, the consensus sequence is  
 19 published as a FASTA file, then our ability to populate  $\mathcal{S}$  is severely limited. Any uncertainty  
 20 that we impose upon the data will be a principled assumption for the purpose of evaluating the  
 21 robustness of the results to potential or assumed sequence uncertainty. The error probability

1 at the  $j$  position of the consensus sequence can be simulated as a beta distribution, *i.e.*,

$$\epsilon_j \sim \text{Beta}(\alpha, \beta)$$

2 The called base at position  $j$  has probability  $1 - \epsilon_j$ , and the remaining bases are assigned  
3  $\epsilon_j/3$ . To incorporate deletions, another probability  $d$  can be generated as the *gap probability*.  
4 With these defined, the nucleotide-level probabilistic sequence at the  $j$ th column (assuming  
5 the base call at position  $j$  was **A**) can be written as above. This probabilistic sequence is  
6 completely fabricated, *i.e.*, not based on any empirical data. However, the sensitivity of an  
7 analysis can be evaluated by choosing different values of  $\alpha$ ,  $\beta$ , and  $d$  (*e.g.*, based on previous  
8 studies) and propagating these uncertainties into downstream analyses. The results from  
9 such an analysis would not indicate anything about the sequence itself but could be used to  
10 determine how robust the methods are to increased sequence uncertainty.

11 Figure 2 summarizes the various ways a probabilistic sequence can be obtained depending  
12 on the type of data available.

13 For both the FASTQ and FASTA format, the uniform distribution was chosen for illustra-  
14 tive purposes. We hope that future analyses take uncertainty into account, and each analysis  
15 will have unique needs. In the absence of available SAM files, alternate assumptions about  
16 the unknown uncertainties can be made. As noted by a reviewer, for viruses such as SARS-  
17 CoV-2 it is possible to calculate the per-position frequencies of each letter. In other contexts,  
18 there may be other potential assumptions that coincide with known features of the organism.

## 19 **1.6 Propagation of uncertainty via resampling**

20 The most general way to propagate uncertainty is through resampling. Given  $\mathcal{S}$  and assum-  
21 ing that individual nucleotides are independent outcomes we can propagate uncertainty by  
22 running downstream analyses on each set of sampled sequences.

23 At a nucleotide level, we are sampling from a multinomial distribution. If the  $j$ th column  
24 of  $\mathcal{S}$  is (0.5, 0.2, 0.2, 0.09, 0.01), then we could sample **A** with 50% probability, **C** with 20%,  
25 etc. As with other sequence analyses, we can censor the positions that do not have enough

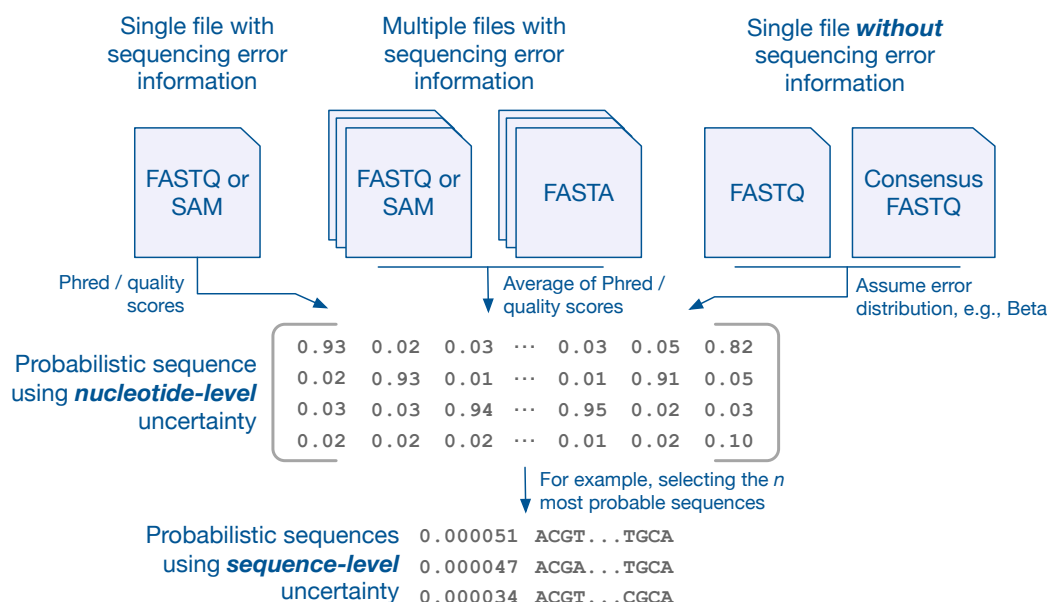


Figure 2: Summary of probabilistic sequences construction. Nucleotide-level probabilistic sequences can be generated from a single FASTQ or SAM file using the sequencing quality information (left). In the case of multiple FASTQ or SAM the user can average the sequencing quality information beforehand (center). When multiple FASTA files are available, the probabilities can be directly informed from the frequencies of nucleotides at each position (center). In the case of a single FASTA file or consensus FASTQ file, the user can assume a probability model (section 1.5.2) for the distribution of sequencing errors (right). Sequence-level probabilistic sequences may be obtained from the nucleotide-level ones, for example by selecting the  $n$  most probable sequences (bottom).



1 coverage. We arbitrarily chose to censor any position that had fewer than 10 reads.

## 2 **1.7 Implementation**

3 A C program has been written to convert SAM files into our matrix representation. The  
4 program assumes that the reads are aligned to a reference, then uses that reference to initiate  
5 the matrix. Because of our methods for handling paired reads, the program is able to stream  
6 the file line-by-line in a parallel computing environment.

7 The resampling algorithm defined above has been implemented in the R programming  
8 language. A shell script is used to repeatedly call the necessary R functions and apply the  
9 resampling algorithm to all outputs of the C program until the desired number of samples is  
10 obtained. All of the code for this project is available at <https://github.com/Poonlab/SUP>.

## 11 **2 Applications**

### 12 **2.1 SARS-CoV-2 lineage assignment**

13 In this section, we apply the re-sampling method to evaluate the impact of sequencing error  
14 on the lineage assignments of SARS-CoV-2. Sequences are sampled from  $\mathcal{S}$ , assigned a  
15 lineage based on the lineage designation algorithm described in Rambaut et al. (2020) using  
16 the pangoLEARN tool (Pangolin version 2.3.2, pangoLEARN version 2021-02-21) that the  
17 authors have made available ([github.com/cov-lineages/Pangolin](https://github.com/cov-lineages/Pangolin)). This tool uses a decision  
18 tree model to determine which lineage a given sequence is most likely to belong to. We  
19 demonstrate that even the best available tools are underestimating the variance and therefore  
20 producing overconfident conclusions.

#### 21 **2.1.1 Data**

22 The data for this application were downloaded from NCBI's SRA web interface (<https://www.ncbi.nlm.nih.gov/sra/?term=txid2697049>) on July 17th, 2021. Search results were  
23

1 filtered to only include records that had SAM files so that our alignments were consistent  
2 with the originating work. We note that the use of pre-aligned SAM files means that we do  
3 not have full control over the reference sequence, and thus there may be some difference  
4 in the choice of alignment which may lead to probabilistic sequences that are not aligned  
5 to each other. In our first application we do only make comparisons within re-samples of a  
6 sequence - not between sequences - and our second application involves a multiple sequence  
7 alignment in order to find mutations relative to each other. To select which runs to down-  
8 load, an arbitrary selection of 5-10 records from each of 20 non-sequential results pages were  
9 chosen. Once collecting the run accession numbers from the search results, an R script was  
10 run to download the relevant files and check that all information was complete. 23 out of  
11 275 files were incomplete due to technical errors during the download process and a further 4  
12 were rejected due to lack of CIGAR strings (the NCBI database automatically converts files  
13 uploaded as unaligned FASTQ into the SAM file format without performing alignment). The  
14 SRA accession numbers for the sequences we used are provided in the Appendix.

### 15 **2.1.2 Re-sampling the probabilistic sequence**

16 Since pangoLEARN is a pre-trained model, assigning lineage designations to a large number  
17 of resampled genome sequences is not computationally burdensome. Sampling 5,000 dif-  
18 ferent sequences from a probabilistic sequence can be done in a reasonable amount of time,  
19 even on a mid-range consumer laptop.

20 Figure 3 shows that the consensus sequence is almost always assigned to the same lin-  
21 eage as the majority of the resamples; the full results are in the Appendix. The proportion  
22 of resamples with the same lineage as the consensus sequence is very rarely 100% and can  
23 be as low as 32.86% (accession number ERR4440425). There were 52 cases where the pro-  
24 portion agreeing with the consensus sequence was either exactly 0 or less than 1%, and these  
25 cases occurred when the most common lineage sampled was labelled or "None" (sequences  
26 are labelled "None" when pangolin's classification does not reach a confidence threshold). It  
27 is noteworthy that the only times where 100% of resampled sequences agreed are when the

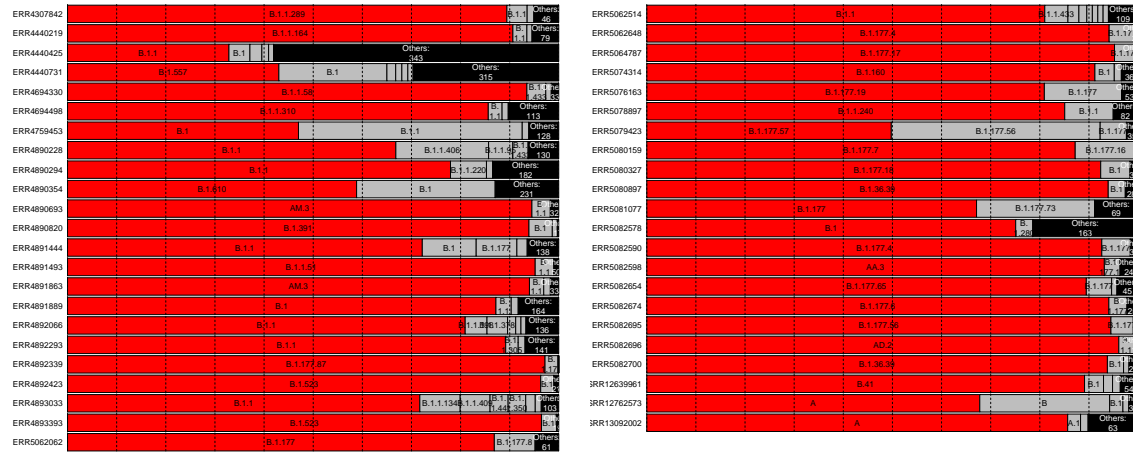


Figure 3: Visualization of called lineages from Pangolin. Red bars indicate the lineage of the most probable sequence and grey bars represent other sequences called from the same SAM file. Any lineage with fewer than 100 observations in the simulated sequences was grouped into the “Other” category. There were 95 sequences total, but we only plotted the ones where the second most common lineage designation had more than 250 observations.

1 lineage call was “None” (13 cases) or for the lineage labelled B.1.1.7 (16 cases). This lineage  
 2 represents 6% of our data and is a significantly more infectious lineage that is of special con-  
 3 cern to health authorities (Wise, 2020; European Centre for Disease Prevention and Control,  
 4 2021).

## 5 2.2 Clock rate estimation for SARS-CoV-2

6 The molecular clock rate (the number of mutations per site per unit of time) of a phylogenetic  
 7 tree is found by considering both the number of mutations for each observed sequence relative  
 8 to the root of the tree and the sample dates of those sequences. Assuming heterochronous  
 9 sampling dates, the rate of mutations can be estimated by regressing the number of mutations  
 10 against the sampling date. In the simplest case the clock rate is the slope estimate from a linear  
 11 regression, thus assuming a fixed clock rate. Polynomial and non-linear clock rates can be  
 12 estimated (Sagulenko et al., 2018), as well as Bayesian non-parametric estimates (Drummond  
 13 and Bouckaert, 2015).

14 The clock rate for SARS-CoV-2 is commonly estimated as a fixed rate near 0.001 mu-  
 15 tations per site per year (Duchene et al., 2020; Choudhary et al., 2021; Song et al., 2021;

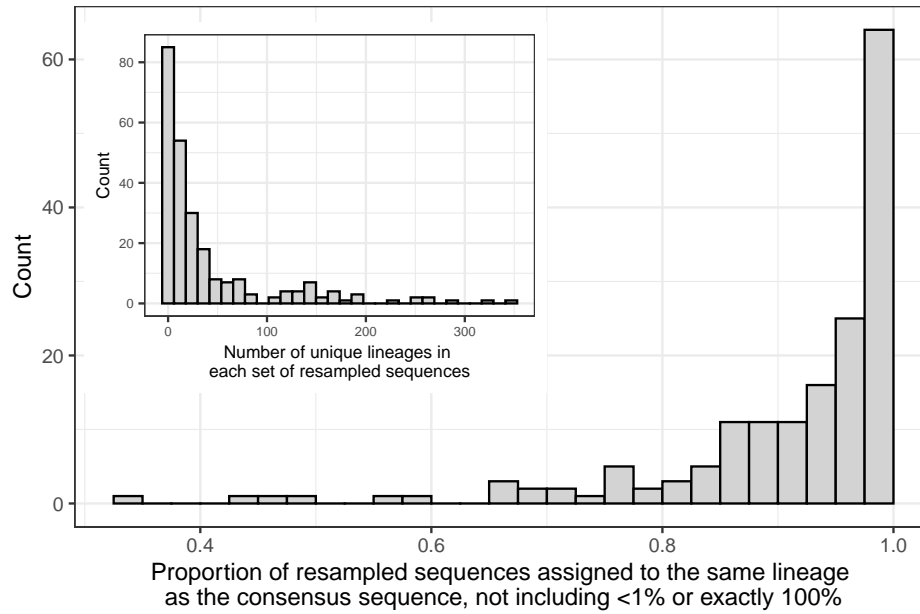


Figure 4: **Main plot:** Proportion of resampled sequences that are assigned to the same lineage as the consensus sequence. One proportion is calculated for each SAM file. The sets of resampled sequences where the proportion was less than 1% or exactly 100% are explained in Section 2.1.2. **Inset:** The number of distinct lineage assignments within each set of resampled sequences.

1 Nie et al., 2020; Geidelberg et al., 2021). Using the same resampling methods as above, we  
 2 estimate a clock rate for trees estimated from each of 50 resamples and for the tree estimated  
 3 based on the consensus sequences.

4 To obtain the data, we sampled genomes uniformly from each month of recorded data  
 5 in GenBank, using filters to ensure that the genomes were complete and had an associated  
 6 SAM file. We further had to filter out SAM files that were incomplete or did not contain the  
 7 CIGAR strings necessary for alignment, leaving us with 244 sequences. The associated SRA  
 8 accession numbers are provided in the Appendix.

9 Our re-sampling method will, by definition, introduce other possible mutations beyond  
 10 what the consensus sequence suggests. Because of this, the apparent number of mutations  
 11 between a re-sampled genome and the estimated root is a function of the coverage, with more  
 12 positions read or more uncertainty in the sequence leading to artificially inflated terminal  
 13 branch lengths. Furthermore, we are sampling nucleotides at each position independently of  
 14 other positions as well as independently of ancestral sequences. This implies that the esti-

1 mates of the time for the most recent common ancestor are not reliable. However, assuming  
2 that the sequences have comparable levels of uncertainty, each branch increases by a similar  
3 amount and the clock rate should not be affected.

4 The sequences that we acquired did not have comparable levels of uncertainty; the viruses  
5 sampled early in the pandemic had considerably higher uncertainty, most likely due to a lack  
6 of consistent laboratory guidelines for sequencing this new virus. To account for this, we  
7 calculated the sum of  $\mathcal{S}'$  for each sequence and applied Statistical Process Control techniques  
8 to ensure that all of the sequences had a similar level of coverage. In particular, we calculated  
9 the mean coverage of the sequences in our data set,  $\bar{c}$ , and the standard deviation of the  
10 coverages,  $s$ . We removed any sequences outside of  $\bar{c} \pm 3s$ , recalculated  $\bar{c}$  and  $s$ , and iterated  
11 the removal process until all sequence coverages were within the bounds, amounting to 20  
12 removed sequences.

13 The clock rate was estimated using TreeTime Sagulenko et al. (2018). We recorded the  
14 clock rate and standard error from the time tree constructed using the consensus sequences  
15 and compared this to the clock rate and standard deviations of the estimated clock rates in  
16 the resampled sequences. The tree built from consensus sequences had a clock rate of  $6.5 \times$   
17  $10^{-4}$  with a standard error of  $8.01 \times 10^{-5}$ . The mean of the clock rates for all of the sets  
18 of resampled sequences was  $8.6 \times 10^{-4}$  with standard deviation of  $5.3 \times 10^{-4}$ , which is  
19 approximately 1.6 times as large as the standard error for the consensus sequences.

20 The estimates of the clock rate are shown in Figure 5. The red line and shaded region are  
21 the clock rate for the tree built from consensus sequences along with  $\pm 1.96$  standard errors.  
22 Rate estimates from Duchene et al. (2020) (n=122), Choudhary et al. (2021) (n=261), Song  
23 et al. (2021) (n=29), Nie et al. (2020) (n=112), and Geidelberg et al. (2021) (n=77) are also  
24 labelled on the plot with purple error bars for 95% Bayesian Credible Intervals (BCI) or 95%  
25 Highest Posterior Density (HPD), indicating that the rates and errors from each root-to-tip  
26 regression are in line with other published results. Figure 5 demonstrates that the estimated  
27 evolutionary rates have an average close to the rate estimated from our tree estimated from  
28 consensus sequences as well as the rates from other studies, but each of the individual er-

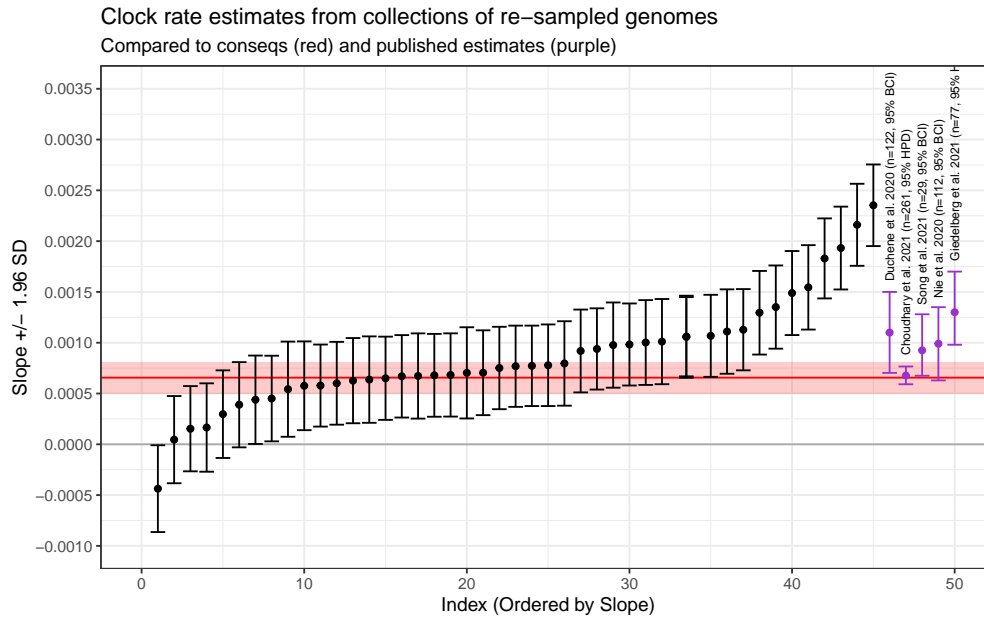


Figure 5: Clock rates (slope) and 95% Confidence Intervals for the collections of re-sampled sequences. The red line and red shaded region are the clock rate and 95% CI for the consensus sequences. The purple points and error bars are the clock rates and error intervals (either Bayesian Credible Interval or Highest Posterior Probability) from published studies, as labelled. The re-sampled sequences are in line with the consensus sequences as well as the published sequences, but represent a much larger variation due to the uncertainty in the original genome sequences.

1 ror bars (from the five studies identified above) miss the excess variation due to sequence  
2 uncertainty.

### 3 Conclusions

4 The files produced by NGS platforms include valuable information about the quality of base  
5 calls which should be propagated into analyses. In this study, we have demonstrated that  
6 these errors in base calling can lead to different conclusions when determining a lineage via  
7 Pangolin and that the variance in clock rate estimates is larger than previously shown due to  
8 these errors. Both of these situations could lead to incorrect conclusions, such as missing  
9 a variant of interest or making overconfident conclusions about the date of the first case of  
10 COVID-19. The potential for errors in base calls should always be taken into account when  
11 making decisions based on genetic sequencing data.

1 Our analysis of Pangolin lineage classification demonstrates that the uncertainty in the  
2 base calls has a non-trivial effect on the potential lineage calls. The reported lineage classi-  
3 fications are based on a sophisticated classification algorithm which has high confidence in  
4 the predicted category, but this assumes that the input sequence is known without error. We  
5 are not aware of any classification system that incorporates per-base error, so we suggest that  
6 interpretations of the output of any classification system be interpreted with reference to the  
7 uncertainty in their sequence.

8 Our clock rate estimation suggest that the confidence/credible intervals for the published  
9 clock rates are underestimated. As with lineage classification, we are not aware of any clock  
10 rate estimation procedures that incorporate the uncertainty in the base calls of the sequences.  
11 Researchers should be conscious of this potential source of currently unacknowledged error  
12 when reporting any results from sequenced genomes.

## 13 **4 Discussion**

14 The primary contribution of this research is the construction of the probability sequence,  
15 which allows for a wide variety of future research directions. The direction we described  
16 here is focused on re-sampling, which allows a more complete appraisal of the variance in the  
17 estimates (or provides a reasonable prior distribution in a Bayesian setting), while comparing  
18 results for the most likely sequences provide a measure of robustness to sequence uncertainty.

19 Our proposed methods can result in a linear increase in computational expense. Even  
20 the method based on ordering the sequences by likelihood inevitably requires re-running  
21 the analysis numerous times. However, we have demonstrated that the uncertainty in the  
22 sequences themselves can lead to major changes to the interpretations of the results. The so-  
23 called “consensus sequence” is simply the most likely sequence, and the reported uncertainty  
24 is not merely an academic curiosity. Ideally individual analyses would be constructed to take  
25 nucleotide-level uncertainty into account. For instance, phylogenies have been estimated  
26 based on uncertain sequence information in Ross and Markowitz (2016), Jahn et al. (2016)

1 and Zafar et al. (2017), but the uncertainty is not derived from base quality scores. An  
2 extension of these methods to incorporate the base quality scores is a worthwhile research  
3 direction.

4 As noted by a reviewer, De Maio et al. (2013) presents a method to construct phyloge-  
5 netic trees such that each tip is associated with a collection of species. It uses a multiple  
6 sequence alignment for each of a collection of species and incorporates the polymorphisms  
7 for each species. Our method could re-purpose this paradigm to apply to re-samples from the  
8 probabilistic sequence in place of multiple sequence alignments, with the separate genomes  
9 acting as species. Alternatively, the method could be altered to directly incorporate sequence  
10 uncertainty, possibly using values from our construction of the probabilistic sequence as allele  
11 proportions. This combination of methods would improve the estimation of the variance and  
12 allow for an improved estimate of error rate (analogous to the within-species evolution rate).

13 Computational burden can also be reduced by sorting the sequences in decreasing uncer-  
14 tainty. It is possible to devise an algorithm that puts the sequences in (approximate) order of  
15 their uncertainty without calculating the uncertainty for every sequence (specifically, by start-  
16 ing with the consensus and at each step changing the base call that had the lowest quality).  
17 Any model that uses sequence data could be re-fit with each sequence in order of uncertainty  
18 to investigate the robustness of that model to sequence uncertainty.

19 Our analysis focused on lineage classification according to the Pangolin model as well  
20 as estimation of the clock rate. The importance of incorporating sequence uncertainty is not  
21 confined to these applications; any analysis involving sequenced genomes would benefit from  
22 some method of incorporating the uncertainty or including some measure of robustness. For  
23 example, the estimated frequency of alleles in the population could be used as the probability  
24 sequence, then propagated into further analyses. We also included a section on assumptions  
25 about errors that are not quantified (consensus-level FASTQ and FASTA files), but we have  
26 not implemented an example of this. Evaluating particular methods was not part of our scope,  
27 but such a study would be a valuable research direction.

28 Within SARS-CoV-2, there are many potential use-cases for our methods. As noted by



1 a reviewer, one potential use-case is to use simulated reads (with known lineage) with vary-  
2 ing levels of uncertainty in order to estimate the potential variance around a given lineage  
3 assignment. It is likely that, due to different amounts of mutations used to define lineages  
4 and differences in average read depth at different locations, different lineages may be subject  
5 to different levels of variability. We stress that re-sampling is a general method, and devel-  
6 opment of methods that incorporate uncertainty — *e.g.*, incorporating uncertainty directly in  
7 the inference procedure, perhaps directly in the formulation of the likelihood — should be a  
8 priority for future research in particular applications of uncertainty propagation.

9 Our method does not preclude tertiary analyses to test for systematic errors. For instance,  
10 De Maio et al. (2020) suggest that some errors arise due to issues in the sequencing protocol  
11 in particular laboratories. Our method allows for adjustments of the base call quality score,  
12 such as in Brockman et al. (2008), correcting for laboratory-specific errors, as well as more  
13 sophisticated definitions of genome likelihoods (*e.g.*, Li et al., 2004; DePristo et al., 2011; Li  
14 et al., 2009b).

15 We have evaluated an algorithm to include insertion events in a re-sampling scheme,  
16 but many of the resultant sequences were not mappable to known sequences. The Pangolin  
17 lineage assignment system appears to treat insertions differently from single nucleotide poly-  
18 morphisms, and our method of sampling insertions is incompatible with their treatment of  
19 them. This is potentially because the sampled base pair at any given position is independent  
20 of each other position, and the insertions observed in real-world data are possibly always  
21 associated with particular mutations elsewhere. However, insertions in the SARS-CoV-2  
22 genome have been relatively rare.

23 This study should not be taken in any way as a criticism of the Pangolin lineage assign-  
24 ment procedure. Rather, Pangolin was chosen as it is the state-of-the art tool for lineage  
25 classification. The phylogeny created by this team has been a vital resource for researchers  
26 and for public health professionals. In particular, the PANGO label for the current Variants of  
27 Concern (VOCs), especially B.1.1.7, are the labels being used worldwide by news organiza-  
28 tions. The output from Pangolin and many other bioinformatics tools are usually interpreted

1 as *deterministic* results. This study is an argument that inherent uncertainty in sequencing  
2 warrants propagation into downstream analyses.

## 3 **References**

- 4 Beerenwinkel, N. and Zagordi, O. (2011). Ultra-deep sequencing for the analysis of viral  
5 populations. *Current Opinion in Virology*, 1(5):413–418.
- 6 Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C.,  
7 Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and SNP detection in  
8 sequencing-by-synthesis systems. *Genome Research*, 18(5):763–770.
- 9 Choudhary, M. C., Crain, C. R., Qiu, X., Hanage, W., and Li, J. Z. (2021). Severe Acute  
10 Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Sequence Characteristics of Coro-  
11 navirus Disease 2019 (COVID-19) Persistence and Reinfection. *Clinical Infectious Dis-*  
12 *eases*, (ciab380).
- 13 Clement, N. L., Snell, Q., Clement, M. J., Hollenhorst, P. C., Purwar, J., Graves, B. J., Cairns,  
14 B. R., and Johnson, W. E. (2010). The GNUMAP algorithm: Unbiased probabilistic map-  
15 ping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1):38–45.
- 16 De Maio, N., Schlötterer, C., and Kosiol, C. (2013). Linking Great Apes Genome Evolution  
17 across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology*  
18 *and Evolution*, 30(10):2249–2262.
- 19 De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkiewicz, G., and Goldman, N. (2020).  
20 Issues with SARS-CoV-2 sequencing data. [https://virological.org/t/issues-with-sars-cov-](https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473)  
21 [2-sequencing-data/473](https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473), Accessed 2021-11-24.
- 22 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philip-  
23 pakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernyt-  
24 sky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J.

- 1 (2011). A framework for variation discovery and genotyping using next-generation DNA  
2 sequencing data. *Nature Genetics*, 43(5):491–498.
- 3 Doronina, N. V. (2005). Phylogenetic position and emended description of the genus  
4 *Methylovorus*. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY*  
5 *MICROBIOLOGY*, 55(2):903–906.
- 6 Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*.  
7 Cambridge University Press.
- 8 Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and  
9 Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus*  
10 *Evolution*, 6(2).
- 11 European Centre for Disease Prevention and Control (2021). SARS-CoV-2 variants of con-  
12 cern as of 26 November 2021. <https://www.ecdc.europa.eu/en/covid-19/variants-concern>,  
13 Accessed 2021-11-26.
- 14 Ewing, B. and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using *Phred*.  
15 II. Error Probabilities. *Genome Research*, 8(3):186–194.
- 16 Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jo-  
17 vanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., and Vezenov, D. V. (2009).  
18 The challenges of sequencing by synthesis. *Nature Biotechnology*, 27(11):1013–1023.
- 19 Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderroth, T., Huerta-Sánchez, E., Al-  
20 brechtsen, A., and Nielsen, R. (2013). Quantifying Population Genetic Differentiation  
21 from Next-Generation Sequencing Data. *Genetics*, 195(3):979–992.
- 22 Geidelberg, L., Boyd, O., Jorgensen, D., Siveroni, I., Nascimento, F. F., Johnson, R.,  
23 Ragonnet-Cronin, M., Fu, H., Wang, H., Xi, X., Chen, W., Liu, D., Chen, Y., Tian, M.,  
24 Tan, W., Zai, J., Sun, W., Li, J., Li, J., Volz, E. M., Li, X., and Nie, Q. (2021). Genomic  
25 epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evolution*, 7(1).

1 Gompert, Z. and Buerkle, C. A. (2011). A Hierarchical Bayesian Model for Next-Generation  
2 Population Genomics. *Genetics*, 187(3):903–917.

3 Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of  
4 next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

5 Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data.  
6 *Genome Biology*, 17(1):86.

7 Keith, J. M., Adams, P., Bryant, D., Kroese, D. P., Mitchelson, K. R., Coachran, D. A. E.,  
8 and Lala, G. H. (2002). A simulated annealing algorithm for finding consensus sequences.  
9 *Bioinformatics*, 18(11):1494–1499.

10 Kozlov, O. (2018). *Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference,*  
11 *Handling Sequence Uncertainty*. PhD thesis, Karlsruhe Insititute of Technology.

12 Kuhner, M. K. and McGill, J. (2014). Correcting for Sequencing Error in Maximum Likeli-  
13 hood Phylogeny Inference. *G3 Genes—Genomes—Genetics*, 4(12):2545–2552.

14 Kuo, T., Frith, M. C., Sese, J., and Horton, P. (2018). EAGLE: Explicit Alternative Genome  
15 Likelihood Evaluator. *BMC Medical Genomics*, 11(2):28.

16 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abeca-  
17 sis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools.  
18 *Bioinformatics*, 25(16):2078–2079.

19 Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling  
20 variants using mapping quality scores. *Genome Research*, 18(11):1851–1858.

21 Li, M., Nordborg, M., and Li, L. M. (2004). Adjust quality scores from alignment and  
22 improve sequencing accuracy. *Nucleic Acids Research*, 32(17):5183–5191.

1 Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detec-  
2 tion for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124–  
3 1132.

4 NC-IUB (1986). Nomenclature for incompletely specified bases in nucleic acid sequences.  
5 Recommendations 1984. Nomenclature Committee of the International Union of Biochem-  
6 istry (NC-IUB). *Proceedings of the National Academy of Sciences of the United States of*  
7 *America*, 83(1):4–8.

8 Nie, Q., Li, X., Chen, W., Liu, D., Chen, Y., Li, H., Li, D., Tian, M., Tan, W., and Zai,  
9 J. (2020). Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Research*,  
10 287:198098.

11 O’Rawe, J. A., Ferson, S., and Lyon, G. J. (2015). Accounting for uncertainty in DNA  
12 sequencing data. *Trends in Genetics*, 31(2):61–66.

13 Rambaut, A., Holmes, E. C., O’Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L.,  
14 and Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to  
15 assist genomic epidemiology. *Nature Microbiology*.

16 Richterich, P. (1998). Estimation of Errors in “Raw” DNA Sequences: A Validation Study.  
17 *Genome Research*, 8(3):251–259.

18 Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitiga-  
19 tion in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62.

20 Ross, E. M. and Markowitz, F. (2016). OncoNEM: Inferring tumor evolution from single-cell  
21 sequencing data. *Genome Biology*, 17(1):69.

22 Sagulenko, P., Puller, V., and Neher, R. A. (2018). TreeTime: Maximum-likelihood phylo-  
23 dynamic analysis. *Virus Evolution*, 4(1):vex042.

- 1 Salk, J. J., Schmitt, M. W., and Loeb, L. A. (2018). Enhancing the accuracy of next-  
2 generation sequencing for detecting rare and subclonal mutations. *Nature reviews. Ge-*  
3 *netics*, 19(5):269–285.
- 4 Schneider, T. D. (2002). Consensus Sequence Zen. *Applied bioinformatics*, 1(3):111–119.
- 5 Schneider, T. D. and Stephens, R. (1990). Sequence logos: A new way to display consensus  
6 sequences. *Nucleic Acids Research*, 18(20):6097–6100.
- 7 Song, N., Cui, G.-L., and Zeng, Q.-L. (2021). Genomic Epidemiology of SARS-CoV-2  
8 From Mainland China With Newly Obtained Genomes From Henan Province. *Frontiers*  
9 *in Microbiology*, 12:673855.
- 10 Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'perceptron'  
11 algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*,  
12 10(9):16.
- 13 Wise, J. (2020). Covid-19: New coronavirus variant is identified in uk. *BMJ*, 371.
- 14 Wu, S. H., Schwartz, R. S., Winter, D. J., Conrad, D. F., and Cartwright, R. A. (2017).  
15 Estimating error models for whole genome sequencing using mixtures of Dirichlet-  
16 multinomial distributions. *Bioinformatics*, 33(15):2322–2329.
- 17 Zafar, H., Tzen, A., Navin, N., Chen, K., and Nakhleh, L. (2017). SiFit: Inferring tumor trees  
18 from single-cell sequencing data under finite-sites models. *Genome Biology*, 18(1):178.