Propagating Sequence Uncertainty into Downstream Analyses

David Champredon, Devan Becker, Art Poon, Connor Chato

Abstract

Genetic sequencing is subject to many different types of errors, but most analyses treat the resultant sequences as if they are perfect. Since the process of sequencing is very difficult, modern machines rely on significantly larger numbers of reads rather than making each read significantly more accurate. Still, the coverage of such machines is imperfect and leaves uncertainty in many of the base calls. Furthermore, there are circumstances around the sequencing that can induce further problems. In this work, we demonstrate that the uncertainty in sequencing techniques will affect downstream analysis and propose a straightforward (if computationally expensive) method to propagate the uncertainty.

Our method uses a probabilistic matrix representation of individual sequences which incorporates base quality scores and makes various uncertainty propagation methods obvious and easy. 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 an honest evaluation of the error involved in such analyses.

We demonstrate our resampling method on HIV and SARS-CoV-2 data. The resampling procedures adds computational cost to the analyses, but the large impact on the variance in downstream estimates makes it clear that ignoring this uncertainty leads to invalid conclusions. For HIV data, we show that phylogenetic reconstructions are much more sensitive to sequence error uncertainty than previously believed, and for SARS-CoV-2 data we show that lineage designations via Pangolin are much less certain than the bootstrap support would imply.

1 Intro

Extracting DNA/RNA from biological samples is a complex process that involves many steps: extraction of the genetic material of interest (avoiding contamination with foreign/unwanted genetic material); reverse transcription (if RNA); DNA fragmentation of the genome into smaller segments; amplification of the fragmented sequences using PCR; sequencing the fragments (*e.g.,* with fluorescent techniques); putting back the small fragments together by aligning them (de novo) or mapping them to benchmark libraries. Errors can be introduced at each of these steps for various reasons [Beerenwinkel and Zagordi, 2011, O'Rawe et al. [2015]] and some errors can be quantified (*e.g.,* sequencing quality scores from chromatographs).

When the phylogenic tree to infer is based on pathogen sequences infecting hosts, the potential genetic diversity of the infection adds a complexity in phylogeny reconstruction. Typical examples are epidemiological studies reconstructing transmission trees from viral genetic sequences (*e.g.,* HIV, HepC) sampled from infected patients.

The different sources of uncertainty described above impact our observations of the actual genetic sequences. There are standard approaches to deal with identifiable observation errors. Base calls that are ambiguous (from equivocal chromatograph curves or because of genuine polymorphisms) are assigned ambiguity codes (*e.g.,* Y for C or T, R for A or G, etc.). Alignment methods are heuristic methods based on similarity scores that generally do not quantify the uncertainty of alignment. Methods to reconstruct phylogenies usually leave out the uncertainty complexity and settle for sequences composed of the most frequent nucleotides and/or ignore ambiguity codes (with some exceptions, e.g. DePristo et al. [2011]).

In 1998, Ewing and Green [1998] and Richterich [1998] both showed that estimates of the base call quality (called Phred scores) can be an accurate estimate of the number of errors that the machines at the time would make. Modern machines still report these Phred scores, but methods for adjusting/recalibrating these scores for greater accuracy have been proposed [Li et al., 2004, DePristo et al. [2011], Li et al. [2009]] For

most analyses, these scores are used to censor the base calls (i.e., label them "N" rather than A, T, C, or G) if the base call error probability is too high or there are too few reads and a given location. It is commonplace to remove the sequence from analysis if the total sequence error probability is too high [see, e.g., Doronina, 2005, Robasky et al. [2014], O'Rawe et al. [2015]]. The error probability is deemed too high based on a strict threshold (e.g. 1% chance of error), but these thresholds aren't necessarily standard across studies.

Some studies have incorporated the error probabilities using genome likelihoods. DePristo et al. [2011] and Gompert and Buerkle [2011] incoporate the adjusted Phred scores into a likelihood framework that allows them to more accurately estimate population-level allele distributions. Fumagalli et al. [2013] present a new estimator for population genetic diversity based on such analyses. Kuo et al. [2018] use genome likelihoods to perform hypothesis tests that evaluate whether a given genome sequence is consistent with a given alternative sequence, after accounting for the errors. As with any parametric model-based approach, genome likelihoods use assumptions about the errors in order to make conclusions about the underlying patterns.

Few other authors have considered the uncertainty present in the sequences in downstream analyses. One notable exception is O'Rawe et al. [2015], who suggest methods for propagating sequence-level uncertainty into determining whether two subjects have the same alleles as well as incorporating the sequence-level uncertainty into confidence intervals for allele frequency.

In our work, we present a simple, general framework that can be incorporated into any analysis of genetic sequence data. This framework involves converting the uncertainty scores into a matrix of probabilities, and repeatedly sampling from this matrix and using the resultant samples in downstream analysis. Unlike likelihood-based approaches, we do not make assumptions about the underlying patterns in the data. In doing so, we gain generalizability at the expense of computation time. Our technique is amenable to any appropriate quality score adjustments prior to building the uncertainty matrix.

2 Methods

2.1 Probabilistic Representation of Sequences

Here, we describe two theoretical frameworks to model sequence uncertainty at the *nucleotide level* or at the *sequence level*. In both frameworks, the sequence of nucleotides from a biological sample is not treated as a certain observation, but as a collection of possible sequences.

2.1.1 Nucleotide-level uncertainty

To represent the uncertainty at each location along the genome, we introduce following matrix:

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & \ell \end{array} \\ \begin{pmatrix} \mathcal{S}_{A,1} & \mathcal{S}_{A,2} & \dots & \mathcal{S}_{A,\ell} \\ \mathcal{S}_{C,1} & \mathcal{S}_{C,2} & \dots & \mathcal{S}_{C,\ell} \\ \mathcal{S}_{G,1} & \mathcal{S}_{G,2} & \dots & \mathcal{S}_{G,\ell} \\ \mathcal{S}_{T,1} & \mathcal{S}_{T,2} & \dots & \mathcal{S}_{T,\ell} \\ \mathcal{S}_{x,1} & \mathcal{S}_{x,2} & \dots & \mathcal{S}_{x,\ell} \end{pmatrix} \end{array} \tag{1}$$

Each column represents the nucleotide position, each row one of the four nucleotide `A,C,G,T` as well as an empty position "`-`"" that symbolizes a genuine deletion (not caused by missing data). Hence, $\mathcal{S}$ is a $5 \times \ell$ matrix. Its elements represent the probability that a nucleotide is at given position:

$$\mathcal{S}_{\texttt{n},j} = \mathbb{P}(\text{nucleotide } \texttt{n} \text{ is at position } j) \tag{2}$$

with the special case for a deletion:

$$\mathcal{S}_{\texttt{-},j} = \mathbb{P}(\text{empty position } j) \tag{3}$$

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

$$\sum_{n \in \{\texttt{A},\texttt{C},\texttt{G},\texttt{T},\texttt{-}\}} \mathcal{S}_{n,j} = 1 \tag{4}$$

Also, the sequence length is stochastic if $\mathcal{S}_{\texttt{-},i} > 0$ for at least one i. The probability that the sequence has the maximum length ℓ is $\prod_{i=1}^{\ell}(1 - \mathcal{S}_{\texttt{-},i})$. We call the matrix $\mathcal{S}$ the *nucleotide-level probabilistic sequence* of a biological sample. The nucleotide (or deletion) drawn at each position is independent from all the others, so there are 5^ℓ possible different sequences for a given probabilistic nucleotide sequence, but these sequences are *not* equally probable.

2.1.2 Sequence-level uncertainty

Out of the 5^ℓ possible sequences, the nucleotide uncertainty may assign a positive probability to sequences that are not biologically possible. As an alternative representation and to reduce the space of possible sequences, let's assume we have enough information (either directly observed from data or simulated) to generate a set of sequences $\mathcal{B} = (\mathcal{B}_i)_{i \in \{1 \dots m\}}$ of all m biologically possible sequences. Note that the $\mathcal{B}_i$ do not have necessarily the same length. The observed genetic sequence, s, is a sample from a specified distribution a:

$$\mathbb{P}(s = \mathcal{B}_i) = a(i) \tag{5}$$

We call the set $\mathcal{B}$ the *sequence-level probabilistic sequence*. Note that, because a is a distribution, we must have $\sum_{i=1}^{m} a(i) = 1$.

If we have the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0 & 0.3 & 0.9 & 0 \\ 0 & 0 & 0.01 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

then there are $2 \times 3 \times 2^3 \times 3 = 144$ possible sequences. The most likely is the one having the highest nucleotides probabilities: 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. Let's take the same example as above, but adding one possible empty position:

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0 & 0.2 & 0.9 & 0 \\ 0 & 0 & 0.01 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0.1 & 0 & 0 \end{pmatrix} \end{array}$$

Like above, there is still a 0.2694 probability that the sequence is ACATGA, but now there is a chance that position 4 is deleted. For example, with probability 0.038 the sequence is ACA-GA.

Below is an example for a sequence-level probabilistic sequence $\mathcal{B}$:

sequence	a
ACATGA	0.60
ACATCA	0.12
AGATCA	0.15
ACAGA	0.05
GCATGA	0.08

% Sampling from $\mathcal{B}$, we will have for example ACATCA 12% of the time. Large genomes, such as SARS-CoV-2, will result in vanishly small probabilities for all sequences, and thus work on the log scale is often preferred.

2.2 Deletions and Insertions

By construction, the nucleotide-level probabilistic sequence must be defined with its longest possible length. Deletions are naturally modelled with our representation but insertions have to be modelled using deletion probability.

Consider the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 1 \\ 0 & 0.01 & 0 & 0.99 & 0 & 0 \end{pmatrix} \end{array} \tag{6}$$

The low deletion probability for position 2 is straightforward to interpret: about 1% of the time, nucleotide G at position 2 is deleted. The high deletion probability for position 4 means there is a 1% chance of a T insertion at this position. Table 1 illustrates this.

Table 1: Representation of insertions and deletions from $\mathcal{S}$ defined in (6)

sequence	**frequency**
CGAAT	common, 98% of the time
CAAT	rare (1% frequency) G deletion at position 2,
CGATAT	rare (1% frequency) T insertion at position 4
CATAT	very rare (0.01% frequency) deletion and insertion

The representation of deletions and insertions with a sequence-level probabilistic sequence (not nucleotide-level) is straightforward because in this framework the sequences are explicitly written out, so are their deletions/insertions.

2.3 Constructing The Uncertainty Matrix

Fragment sequencing error is an error that is quantified with quality (or "Phred") score attributed to each base call from sequencing instrument. The quality score Q is directly related to the error probability: $\epsilon = 10^{-Q/10}$. Ewing and Green [1998] (where Q typically ranges between 1 and 60). The FASTQ file format is the standard representation for combining sequence and observation error. Hence, the uncertainty associated to the base call is quantified by defining the probability that the observed nucleotide is the correct one:

$$\mathbb{P}(\text{nucleotide} = X \mid \text{observed nucleotide} = X) = 1 - \epsilon \tag{7}$$

Unfortunately, this base-call probability relates to only one *focal* nucleotide and we have no information on the probability for the three other possible nucleotides. Hence, we must make a modelling choice regarding the distribution of the remaining probabilities.

2.3.1 SAM Files

Massive parallel sequencing platforms (*e.g.,* Illumina, Oxford Nanopores, etc.) provide a large number of short reads sequences of the biological sample of interest. The length of those short reads are typically much smaller than the genome sequenced, so they have to be aligned and stitch together in order to re-assemble the full genome sequence. The short reads are typically stored in FASTQ files where the observation error of

each nucleotide (estimated by the sequencing platform itself)) is indicated by its Phred score. The alignment and assembly of the short reads is performed by software (internal to the sequencing platform or not *((check this. Examples?))*) and generates a SAM file *((ref))* that efficiently stores the alignment information. The assembly of the short reads in the SAM file can be represented in as an array where the columns are the nucleotide positions. The short reads are "stacked" vertically according to the alignment. The number of short reads stacked for a given nucleotide gives the "coverage" of that position. See Figure 1 for an illustration of this SAM file representation.

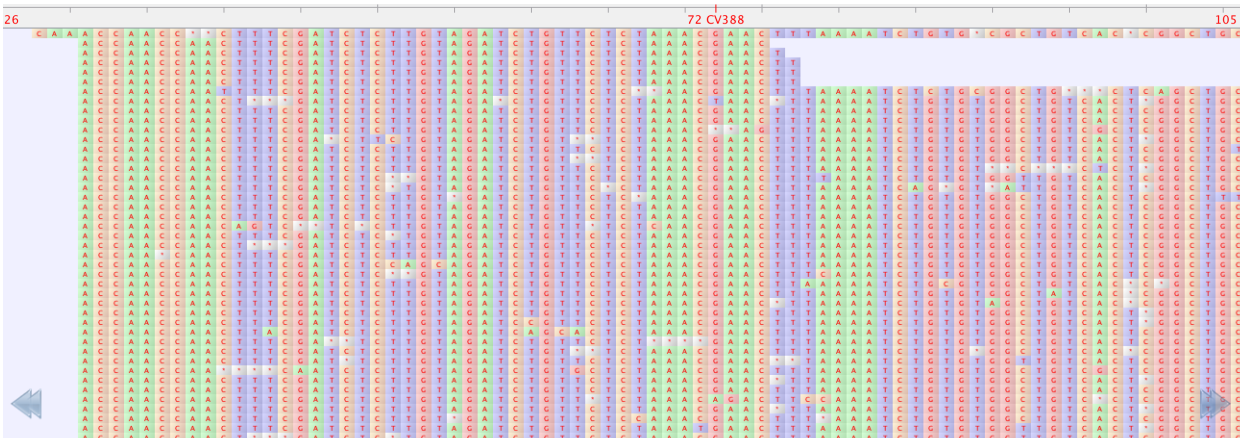

Figure 1: **SAM file graphical representation.** The software Tablet *((ref))* was used. Each base call is colour coded so that base calls that disagree with the consesnsu are obvious. The 72th nucleotide in this alignment has a coverage of 388 reads.

The nucleotide-level probabilistic sequence can be constructed from this alignment. The algorithm that we suggest is as follows.

We begin by initializing an empty $5 \times \ell$ matrix $\mathcal{S}$‘. Each label in Figure 1 has an associated probability score. These labels denote a row in $\mathcal{S}$, and their location on the genome denotes the columns, say $(\texttt{A}, j)$. At this entry, we add the corresponding quality score to whatever was there before.

By this construction, the sum of each column represents the coverage at that location. Dividing each entry of $\mathcal{S}'$ by the sum of that column results in $\mathcal{S}$. For our proposed propagation methods, it is convenient to work with $\mathcal{S}$‘ until probabilities are necessary.

2.3.2 FASTQ files

For most published genome sequences, the short read files are not available. It is common to find the so-called consensus sequence, which is the sequence that represents the most commonly called base at each location, along with a single quality score for each location. The methods for obtaining Phred scores when converting a SAM file to a consensus FASTQ differ by software [Li et al., 2004, Keith et al. [2002], Li et al. [2008]], but generally involve a computation that includes all of the Phred scores of the bases that agree with the consensus (*e.g.,* if the short reads have `A` with Phred of 30, `A` with a Phred of 31, and `C` with phred of 15, then the Phred scores of 30 and 31 are combined).

2.4 Propogation of Uncertainty via Resampling

2.5 Sequence-level Uncertainty (seqLevelUncertaint)

2.5.1 Reducing Computational Burden via Sequence-level Uncertainty

3 Application to SARS-CoV-2

3.1 Data

The data for this application were downloaded from NCBI's SRA web interface. Results were filtered to only include runs that had bam files. To select which runs to download, a selection of 5-10 files from each of 20 non-sequential search result pages was chosen. Once collecting the run accession numbers from the search results, an R script was run to download the relevant files and check that all information was complete. 23 out of 300 files were labelled incomplete due to having too few reads (possibly because the download timed out) or not containing a CIGAR string.

There was no particular reason for choosing any given file, but the resulting data should not be viewed as a random sample. Each result page likely includes several runs from the same study, and runs were chosen arbitrarily within each result page. We were not attempting a completely random sampling strategy, we simply wanted a collection of runs on which to demonstrate our methods.

3.2 PANGOlearn

3.3 (Possibly) constructing trees

3.4 Variant hypothesis testing via MC

4 Conclusions

4.1 For Pangolin

4.2 For phylogenetics in general

4.3 For analysis of genetic data

- Our method does not preclude tertiary analyses to test for systematic errors or deviations from a Mendelian inheritance pattern assumption.

References

Niko Beerenwinkel and Osvaldo Zagordi. Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1(5):413–418, November 2011. ISSN 1879-6257. doi: 10.1016/j.coviro.2011.07.008.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.806.

N. V. Doronina. Phylogenetic position and emended description of the genus Methylovorus. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 55(2):903–906, March 2005. ISSN 1466-5026, 1466-5034. doi: 10.1099/ijs.0.63111-0.

Brent Ewing and Phil Green. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. *Genome Research*, 8(3):186–194, March 1998. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.8.3.186.

Matteo Fumagalli, Filipe G Vieira, Thorfinn Sand Korneliussen, Tyler Linderoth, Emilia Huerta-Sánchez, Anders Albrechtsen, and Rasmus Nielsen. Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics*, 195(3):979–992, November 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.154740.

Zachariah Gompert and C. Alex Buerkle. A Hierarchical Bayesian Model for Next-Generation Population Genomics. *Genetics*, 187(3):903–917, March 2011. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.110.124693.

Jonathan M. Keith, Peter Adams, Darryn Bryant, Dirk P. Kroese, Keith R. Mitchelson, Duncan A. E. Coachran, and Gita H. Lala. A simulated annealing algorithm for finding consensus sequences. *Bioinformatics*, 18(11):1494–1499, November 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.11.1494.

Tony Kuo, Martin C. Frith, Jun Sese, and Paul Horton. EAGLE: Explicit Alternative Genome Likelihood Evaluator. *BMC Medical Genomics*, 11(2):28, April 2018. ISSN 1755-8794. doi: 10.1186/s12920-018-0342-1.

Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, January 2008. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.078212.108.

Ming Li, Magnus Nordborg, and Lei M. Li. Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic Acids Research*, 32(17):5183–5191, 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh850.

Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124–1132, January 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.088013.108.

Jason A. O'Rawe, Scott Ferson, and Gholson J. Lyon. Accounting for uncertainty in DNA sequencing data. *Trends in Genetics*, 31(2):61–66, February 2015. ISSN 0168-9525. doi: 10.1016/j.tig.2014.12.002.

Peter Richterich. Estimation of Errors in "Raw" DNA Sequences: A Validation Study. *Genome Research*, 8(3):251–259, January 1998. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.8.3.251.

Kimberly Robasky, Nathan E. Lewis, and George M. Church. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62, January 2014. ISSN 1471-0064. doi: 10.1038/nrg3655.