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.

Intro

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.

Extracting DNA/RNA from biological samples is a complex process that involves several 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] 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 error probability (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.

TODO: - Studies that incorporate the genome likelihood - O'Rawe et al. [2015]: suggests propogation methods - Also, fumagalliQuantifyingPopulationGenetic2013a and the studies they cite, which use Bayesian methods to get a posterior on the genome likelihoods. - Conclusion for this section

Methods

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.

Constructing The Uncertainty Matrix

(measureUnc)

(pairedReads)

Copy from David.

Insertions and Deletions

(missingnessProblem)

Propogation of Uncertainty via Resampling

Sequence-level Uncertainty (seqLevelUncertaint)

Reducing Computational Burden via Sequence-level Uncertainty

Application to SARS-CoV-2

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.

PANGOlearn

(Possibly) constructing trees

Variant hypothesis testing via MC

Conclusions

For Pangolin

For phylogenetics in general

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.

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.