![JoSS logo](The Journal of Open Source Software)

SeqPanther: Sequence manipulation and mutation statistics toolset

James Emmanuel San [1,2], **Stephanie van Wyk** [2], **Houriiyah Tegally** [1,2], **Simeon Eche** [3], **Eduan Wilkinson** [1,2], **Aquillah M. Kanzi** [1], **Tulio de Oliveira** [1,2,4], **and Anmol M. Kiran** [5]

1 KwaZulu Natal Research and Innovation Sequencing Platform, KRISP, University of KwaZulu Natal, Durban, South Africa 2 Centre for Epidemic Response and Innovation, CERI, University of Stellenbosch, Stellenbosch, South Africa 3 Yale University School of Medicine, New Haven, Connecticut, USA 4 Department of Global Health, University of Washington, Seattle, WA, USA 5 University College Cork, Ireland

Keywords: Bioinformatics, sequence analysis, NGS, codon, amino acid substitution, nucleotide substitution, INDEL

Abstract

Pathogen genomes harbor critical information necessary to support genomic investigations that inform public health interventions such as treatment, control, and eradication. To extract this information, their sequences are analyzed to identify structural variations such as single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs) that may be associated with phenotypes of interest. Typically, this involves the assembly of reads, calling of variants, and calculation of a consensus sequence for downstream analysis. Several pipelines exist to perform quality control (QC) of raw reads and whole-genome assemblies. However, there is no easy-to-use, freely available bioinformatics QC tool to explore mappings for both positional codons and nucleotide distributions in mapped short reads of microbial genomes. To address this problem, we have developed a fast and accurate tool to summarise read counts associated with codons, nucleotides, and INDELs in mapped next-generation sequencing (NGS) short reads. The tool, developed in Python, also provides a visualization of the genome sequencing depth and coverage. Furthermore, the tool can be run in single or batch mode, where several genomes need to be analyzed. Our tool produces a text-based report that enables quick review or can be imported into any analytical tool for upstream analysis. Additionally, the tool provides the functionality to modify consensus sequences by adding, masking, or restoring to wild-type mutations specified by the user.

Availability: SeqPanther is available at https://github.com/codemeleon/seqPanther, along with the necessary documentation for installation and usage.

Introduction

Next-Generation Sequencing (NGS) platforms underlie an exciting era that facilitates the large-scale investigation of pathogen genomics. This in turn supports important aspects relating to public health including genomic epidemiology, pathogen surveillance, and pharmaceutical development of infectious diseases (Harvey et al., 2021). Despite these advances, quality control (QC) of sequences remains a concerning bottleneck to processing big data in a timely fashion to generate actionable information. A high-quality and accurate genome assembly provides insight into the occurrence and associated consequences of genetic changes within pathogenic microorganisms. Analyses to extract this information primarily involve the identification

of variations such as single nucleotide polymorphisms (SNPs), and insertions and deletions (INDELs) which may be associated with enhanced pathogen phenotypes, such as increased transmissibility, immune and vaccine escape, increased infectivity, and disease severity (Harvey et al., 2021). For SARS-CoV-2, and other pathogens of public health interest, several pipelines exist to map sequenced raw reads to a reference genome and generate the consensus sequence for downstream analyses (K.-K. Lam et al., 2015; Narvaez et al., 2022; Vilsker et al., 2018). However, often the consensus deviates from what is expected, requiring additional quality control (QC) and refinements prior to publication. Additionally, to track microbial and viral diversity, genomic sequences are classified into lineages comprising a constellation of mutations exclusive to the lineage (Rambaut et al., 2020). The absence of one or more lineage-defining mutations, some of which may have been implicated in the altering of viral phenotypes, calls for further investigation. This includes re-examining the sequenced and mapped reads to establish the underlying reasons for its absence, and occasionally, and if warranted restore it.

Wrong and/or uncalled mutations, representing false positives and negatives, could arise due to several factors that negatively affect the sequencing, mapping, and assembly outcomes. These include primer drop-outs and algorithmic issues (Heng Li, 2018). Algorithmic issues occur when expected parameter values significantly differ from the values encountered, for example, when depth is lower than the expected minimum due to low viral loads (C. Lam et al., 2021). Low viral loads are typical of samples collected at the later stages of infection or tail ends of an outbreak commonly characterized by high cycle threshold (Ct) values (> 30) or low viral loads (Sutton et al., 2022). Sequences from these are usually defined by large numbers of frameshifts, INDELs, clustered and private mutations, i.e. mutations that are unique to a strain compared to their nearest neighbor in the global phylogeny for supported pathogens (Nextstrain, 2020). Primer drop-outs, on the other hand, are often caused by hypermutation in primer binding regions, reducing the amplification and sequencing for the targeted regions. This results in sections of the genome with low or no coverage. Primer drop-outs were commonly reported throughout the SARS-CoV-2 pandemic, especially in the variants of concern (VOCs)(Davis et al., 2021; Nextstrain, 2020; Sutton et al., 2022). Mutations that have resulted in primer drop-outs for VOCs include the G142D (Delta and Omicron) in the 2_Right primer, the 241/243del (Beta) that occurs in the 74_Left primer, and the K417N (Beta) or K417T (Gamma) which occurs in the 76_Left primer (Ahmed et al., 2022; Davis et al., 2021). The Delta/B.1.1.672 variant has also been associated with ARTIC v3 drop-out of primers 72R and 73L (Borcard et al., 2022).

Tools such as Nextclade (Aksamentov et al., 2021) can capture and report these sequence anomalies. Reports generated through Nextclade include detailed information on the excess number of gaps, mixed bases, private mutations, and frameshifts. However, there is no easy-to-use bioinformatics QC tool to further explore and report codon-affecting alterations (INDELs and substitutions) in the mapped short reads from a mixed bacterial/viral population or batch update of consensus sequences. Moreover, such tools are often taxonomically limited and remain optimized for a select set of reference genomes.

Here, we introduce SeqPanther, a Python application that provides the user with a suite of tools to further interrogate the circumstances under which these mutations occur and to modify the consensus as needed for non-segmented bacterial and viral genomes where reads are mapped to a reference. SeqPanther generates detailed reports of mutations identified within a genomic segment or positions of interest, including visualization of the genome coverage and depth. Our tool is particularly useful in the examination of multiple NGS short-read samples. Additionally, we have integrated Seqpatcher (Singh et al., 2022) which supports the merging of Sanger sequences into their respective NGS consensus sequences.

Implementation

We utilized Python (v3.9.9) as the base programming language to develop our pipeline. Pysam (a wrapper around HTSlib and the samtools package, 0.18.0) is the core module (Danecek et
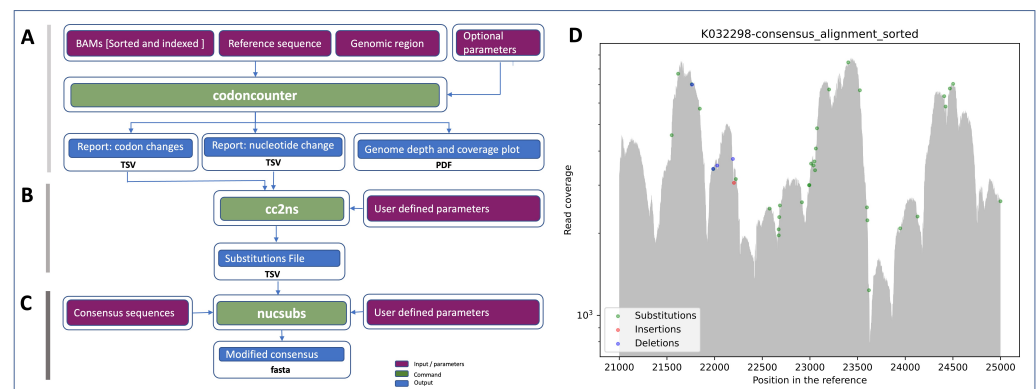
al., 2021; H. Li et al., 2009) of the pipeline which allows for the exploration and manipulation of BAM files generated by sequence mapping tools to extract the distribution of reads. The tool also uses the click package (v8.0.3) to capture user inputs, Pandas (v1.3.5) (Reback et al., 2020) to store and arrange data in data frames and to generate tables, and Matplotlib (v3.6.2) (Hunter, 2007) for plotting read distributions and highlighting the changes in a given region. The outputs from the tool were visually validated using Geneious Prime 2023.0.1 (https://www.geneious.com/). The manuscript was typeset following (Price-Whelan, 2017).

Features

SeqPanther is an easy-to-use and accurate command-line tool for generating summary statistics from BAM files relating to structural changes (SNPs and INDELs) relative to a reference sequence and modifying the consensus sequences calculated from the BAM files to add, remove, or change the variations at specified positions within the sequences. SeqPanther features a suite of tools that perform various functions including codoncounter, cc2ns, and nucsubs. Figure 1 shows the inputs, outputs, and dependencies between the tools.

Codoncounter

The Codoncounter module accepts a BAM file, reference FASTA, and General Feature Format (GFF) genome annotation file as input and produces four output files: 1) a table summarizing nonsynonymous substitutions (Table S1); 2) a table listing the synonymous substitutions (Table S2); 3) a list of the nucleotide INDELs identified (Table S3); and 4) a PDF file with coverage plots and color-coded substitutions at different genomic coordinates that occur relative to the reference **??**. In the case of unsorted and unindexed BAM files, the command first generates a temporary sorted and indexed BAM file that is used to complete the analysis. Amino acid changes are displayed relative to the strand where the gene is present. A codon-aware alignment of the nucleotides is inferred and used to determine the amino acid changes. Codon changes at INDEL sites are not included in the substitution file but are explored independently in the INDELs file.

Figure 1: Visual illustration of the SeqPanther components. A) codoncounter accepts a BAM file or directory containing BAM files (if running batch mode), a reference sequence, and a genomic region or set of positions and generates two reports (codon and nucleotide) and a genome coverage and depth map. B) The cc2ns command accepts INDEL(s) and substitution reports as inputs and generates a substitution file containing all the nucleotide substitutions and INDELs as well as their frequency. C) The nucsubs module takes as input the substitutions file, that has been reviewed and edited by the user and modifies the consensus sequences. D) Example depth and coverage plot generated by codoncounter. Colored circles represent different types of alterations across the length of the assembly, which are supported by more than 5% of total reads.

For each substitution reported by codoncounter, the command provides additional statistics.

These include the total number of reads at the position, the number of reads mapping to the reference base (or bases), the number of reads mapping to the alternate alleles at the position, and the percentage of the total reads that they each comprise. Typically, the codon results are tied to a user-specified region such as an open reading frame (ORF) or protein-coding gene with coordinates defined in a general feature format (GFF) file. The focus on a user-specified genomic segment is based on the fact that some samples might have an extremely large mutation load and quality control will only be focused on just a small subset of mutations in a region of interest, for example, in the case of SARS-CoV-2 mutations in the Spike gene or the Receptor Binding Domain (RBD). The user can also specify the set of nucleotide positions or a genomic range of interest or both. The user-provided reference sequence ID must match the ID in the BAM, GFF, and reference FASTA file. The codons report only contains non-synonymous substitutions or 3x nucleotide INDELs in the mapped reads (Supplementary Table 1). Nucleotide changes that are not associated with an amino acid change (synonymous changes) are reported in a separate file (Supplementary Table 2). The report is thus also useful to investigate synonymous mutations which, although not changing the encoded amino acid, play a crucial role in the estimation of the molecular clock (Yu et al., 2021) and designation of strains and lineages. Examples of lineage-defining synonymous mutations include the T23341C and T24187A in the P320 variant and C26801T in the Alpha/B.1.1.7 variant (cov-lineages.org, 2020).

Table 4: Types of changes resulting in amino acid alterations. In the case of INDELs, only multiples of 3 consecutive nucleotide INDELs were considered for integration in consensus sequences.

Event #	Reference	Alternative	Event	Amino Acid Change	Additional Notes
1	ACT	ATT	Substitution	T-to-I	
2	ACT CCA	ACT TCA CCA	Insertion between codon	TP-to-TSP	Insertion of an amino acid
3	ACA	ACT TCA	Insertion within codon	T-to-TS	Substitution of one amino acid with two
4	ACT	ACG T	Insertion resulting frame-shift	Depends on down-stream nucleotides	Alters all amino acids downstream
5	ACT TCA CCA	ACT CCA	Deletion of codon	TSP-to-TP	Deletion of one amino acid
6	ACT TCA	ACA	Deletion of triple in two codons	TS-to-T	Substitution of two amino acids with one
7	ACT	AT	Deletion resulting frame-shift	Depends on down-stream nucleotides	Alters all amino acids downstream

The output from this module also provides insight into sub-consensus mutations that could potentially become fixed in the population. By default, the tool implements a threshold of 5% to report mutations, however, this can be adjusted up or down to the desired detection threshold by the user. The tool can be run in batch mode or on a single BAM file. To execute the batch mode, the user simply provides a folder containing the BAM files to be analyzed. The output tabular reports can be imported into other tools for further downstream analyses.

Figure 1: Visual illustration of the `SeqPanther` components. A) codoncounter which accepts a BAM file or directory containing BAM files (if running batch mode), a reference sequence, and a genomic region or set of positions and generates two reports (codon and nucleotide) and a genome coverage and depth map. B) The cc2ns command accepts INDEL and substitution reports as inputs and generates a substitution file containing all the nucleotide substitutions and INDELs as well as their frequency. C) The nucsubs module then takes the substitutions

file, that has been reviewed and edited by the user, and modifies the consensus sequences. D) Coverage plot generated by codoncounter. Colored circles represent different types of alterations across the length of the assembly, which are supported by more than 5% of total reads.

Cc2ns

The cc2ns command takes the output from the codoncounter command (nucleotide substitution and INDEL reports) and generates a tabular file according to the user-specified filters that contain the changes which could be integrated into the strain-specific consensus sequences. The output file contains one change per row. In case of multiple changes to a coordinate, the changes are provided in multiple rows. The user can edit the file by adding additional changes or by removing suggested alterations. In case of multiple changes in a position, the first instance is accepted. The reviewed file is passed to the nucsubs command to modify the consensus sequences. The file contains a substitution percentage (i.e. of the reads supporting the substitution). Although this value is optional, it can be used to specify a threshold for all mutations to be restored. This is particularly useful when mutations are not called due to higher thresholds in the variant calling pipeline and is recommended mostly for known, fixed mutations within a constellation.

Nucsubs

The nucsubs command modifies consensus sequences by integrating the user-defined base substitutions, deletions, and insertions by adding additional nucleotides at a given position relative to a reference sequence. As consensus sequences usually differ in length from the reference, to integrate the changes, coordinates are specified relative to the reference. To acquire relative alteration positions, consensus sequences are aligned to the reference using MAFFT aligner v7.508 (Katoh, 2002) in auto mode. The MAFFT alignment is then imported into a Pandas data frame for coordinate manipulation and integration of changes. The modified sequences are saved as a FASTA file. The functionality expedites the often manual process of modifying consensus sequences that may introduce new artifacts when modifying a large number of sequences. The module is sufficiently intelligent to only update the sequences of interest within the alignment or sequence file as required by the user.

These modules were tested on South African SARS-CoV-2 strains and the results were validated manually with the Geneious Prime program (version 2022.2.2). Manual editing of alignments is a common practice in molecular analyses to improve the quality of sequences prior to phylogenetic inference (Barson & Griffiths, 2016).

Seqpatcher

In addition to the new features, we have integrated Seqpatcher (Singh et al., 2022), our bioinformatics tool to merge Sanger sequence fragments into NGS-based consensus sequences. Seqpatcher is particularly useful in the case of primer drop-outs resulting in large gaps spanning an entire protein, open reading frame or gene. Seqpatcher accepts Sanger sequences both in FASTA (pre-processed) or chromatograph (raw .ab1) formats. Sanger sequencing to cover such gaps that are less than a thousand bases is effective in both cost and time relative to re-sequencing the whole genome.

Conclusion

Genomic epidemiology and associated public health interventions entirely depend on the quality of pathogenic genomic data. False positive or negative sequencing outputs can have dire consequences on downstream analyses and subsequent public health decisions. Quality control and assurance are therefore critical components of the sequence generation process.

SeqPanther provides a unique suite of tools to explore short reads and modify consensus sequences. The functionalities provided by SeqPanther can significantly simplify the process of and improve the speed of sequence quality control in small to medium-sized sequencing laboratories, thus in turn reducing the turnaround time to provide quality genomic sequences. Although the tools have been primarily developed and tested using SARS-CoV-2 data, they can be applied to most non-segmented bacterial and viral genomes where reads are mapped to a reference.

Future prospects:

Many microorganisms have fragmented genes (Introns separated exons), overlapped genes or genes with frameshift events such as Influenza A with a segmented genome, however, at present the tool only supports single-segment genomes without other stated events. We plan to extend it to support complex gene and genome structures. We wish to report INDELs whose length is not a multiple of 3 and yet they are present in the same read and summation leads to a multiple of three. Furthermore, as the cost of sequencing continues to decrease and the capacity to sequence increases, we expect to see a further influx in the number of microbial genomes. To this effect, we intend to implement performance enhancements to reduce the batch analyses running time given a fairly large number of genomes.

Acknowledgements

Conflict of interest

The authors declare no competing interests.

Funding sources

References

Ahmed, W., Bivins, A., Smith, W. J. M., Metcalfe, S., Stephens, M., Jennison, A. V., Moore, F. A. J., Bourke, J., Schlebusch, S., McMahon, J., Hewitson, G., Nguyen, S., Barcelon, J., Jackson, G., Mueller, J. F., Ehret, J., Hosegood, I., Tian, W., Wang, H., … Simpson, S. L. (2022). Detection of the omicron (b.1.1.529) variant of SARS-CoV-2 in aircraft wastewater. *Science of The Total Environment*, *820*, 153171. https://doi.org/10.1016/j.scitotenv.2022.153171

Aksamentov, I., Roemer, C., Hodcroft, E., & Neher, R. (2021). Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*,

6(67), 3773. https://doi.org/10.21105/joss.03773

Barson, G., & Griffiths, E. (2016). SeqTools: Visual tools for manual analysis of sequence alignments. *BMC Research Notes*, *9*(1). https://doi.org/10.1186/s13104-016-1847-3

Borcard, L., Gempeler, S., Miani, M. A. T., Baumann, C., Grädel, C., Dijkman, R., Suter-Riniker, F., Leib, S. L., Bittel, P., Neuenschwander, S., & Ramette, A. (2022). Investigating the extent of primer dropout in SARS-CoV-2 genome sequences during the early circulation of delta variants. *Frontiers in Virology*, *2*. https://doi.org/10.3389/fviro.2022.840952

cov-lineages.org. (2020). *Constellations*. https://cov-lineages.org/constellations.html

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Davis, J. J., Long, S. W., Christensen, P. A., Olsen, R. J., Olson, R., Shukla, M., Subedi, S., Stevens, R., & Musser, J. M. (2021). Analysis of the ARTIC version 3 and version 4 SARS-CoV-2 primers and their impact on the detection of the G142D amino acid substitution in the spike protein. *Microbiology Spectrum*, *9*(3). https://doi.org/10.1128/spectrum.01803-21

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & and, D. L. R. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), 409–424. https://doi.org/10.1038/s41579-021-00573-0

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &Amp; Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Katoh, K. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Lam, C., Gray, K., Gall, M., Sadsad, R., Arnott, A., Johnson-Mackinnon, J., Fong, W., Basile, K., Kok, J., Dwyer, D. E., Sintchenko, V., & Rockett, R. J. (2021). SARS-CoV-2 genome sequencing methods differ in their abilities to detect variants from low-viral-load samples. *Journal of Clinical Microbiology*, *59*(11). https://doi.org/10.1128/jcm.01046-21

Lam, K.-K., LaButti, K., Khalak, A., & Tse, D. (2015). FinisherSC: A repeat-aware tool for upgrading<i>de novo</i>assembly using long reads. *Bioinformatics*, *31*(19), 3207–3209. https://doi.org/10.1093/bioinformatics/btv280

Li, Heng. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & and, R. D. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Narvaez, S. A., Shen, Z., Yan, L., Stenger, B. L. S., Goodman, L. B., Lim, A., Nissly, R. H., Nair, M. S., Zhang, S., & Sanchez, S. (2022). Optimized conditions for listeria, salmonella and escherichia whole genome sequencing using the illumina iSeq100 platform with point-and-click bioinformatic analysis. *PLOS ONE*, *17*(11), e0277659. https://doi.org/10.1371/journal.pone.0277659

Nextstrain. (2020). *Quality control (QC)*. https://docs.nextstrain.org/projects/nextclade/en/stable/user/algorithm/07-quality-control.html

Price-Whelan, A. M. (2017). Gala: A python package for galactic dynamics. *The Journal of Open Source Software*, *2*(18). https://doi.org/10.21105/joss.00388

Rambaut, A., Holmes, E. C., O'Toole, Áine, Hill, V., McCrone, J. T., Ruis, C., Plessis, L. du, & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2

lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Reback, J., McKinney, W., Jbrockmendel, Bossche, J. V. D., Augspurger, T., Cloud, P., Gfyoung, Sinhrks, Klein, A., Roeschke, M., Hawkins, S., Tratner, J., She, C., Ayd, W., Terji Petersen, Garcia, M., Schendel, J., Hayden, A., MomIsBestFriend, … Mortada Mehyar. (2020). *Pandas-dev/pandas: Pandas 1.0.3*. Zenodo. https://doi.org/10.5281/ZENODO.3715232

Singh, L., San, J. E., Tegally, H., Brzoska, P. M., Anyaneji, U. J., Wilkinson, E., Clark, L., Giandhari, J., Pillay, S., Lessells, R. J., Martin, D. P., Furtado, M., Kiran, A. M., & Oliveira, T. de. (2022). Targeted sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing. *Microbial Genomics*, *8*(3). https://doi.org/10.1099/mgen.0.000774

Sutton, M., Radniecki, T. S., Kaya, D., Alegre, D., Geniza, M., Girard, A.-M., Carter, K., Dasenko, M., Sanders, J. L., Cieslak, P. R., Kelly, C., & Tyler, B. M. (2022). Detection of SARS-CoV-2 b.1.351 (beta) variant through wastewater surveillance before case detection in a community, oregon, USA. *Emerging Infectious Diseases*, *28*(6). https://doi.org/10.3201/eid2806.211821

Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L. C., Eynden, E. V., Vandamme, A.-M., Deforche, K., & Oliveira, T. de. (2018). Genome detective: An automated system for virus identification from high-throughput sequencing data. *Bioinformatics*, *35*(5), 871–873. https://doi.org/10.1093/bioinformatics/bty695

Yu, Y., Li, Y., Dong, Y., Wang, X., Li, C., & Jiang, W. (2021). Natural selection on synonymous mutations in SARS-CoV-2 and the impact on estimating divergence time. *Future Virology*, *16*(7), 447–450. https://doi.org/10.2217/fvl-2021-0078