# Introduction

## Historical Background

The first recorded instance of the *Orthoflavivirus zikaense*, or colloquially known as the Zika virus, infecting a human is debated to be either in a 1954 diagnosis in Nigeria, or a 1964 diagnosis in Uganda (Wikan & Smith, 2017). Nevertheless, the first documented epidemic of Zika virus did not occur until 2007 on the island of Yap in Micronesia, where an estimated 73% of residents ages 3 or older were infected (Chang et al., 2016). The second outbreak started in 2015, across various countries in the Americas—Brazil alone had an estimated 440,000–1,300,000 cases (Hennessey et al., 2016). As a result, the World Health Organization (WHO) declared Zika virus on February 1, 2016, a "Public Health Emergency of International Concern" (WHO, 2022).

Although the explicit reason for the sudden rise in Zika virus cases between the early 20th to 21st century is unknown, it is generally believed to correlate with an increased rate of urbanization and global travel (Chang et al., 2016). Currently, there are no epidemics for the Zika virus, however, the Center for Disease Control and Prevention (CDC) reports that Zika virus infections continue at low rates in all continents today (CDC, 2025).

## Molecular Knowledge and Infectivity

The Zika virus is a single stranded, positive sense RNA virus with a genome length of around 11 thousand nucleotides (Mozūraitis et al., 2025). The virus typically spreads through mosquitoes, sexual intercourse, or perinatal routes (White et al., 2016). Its genome encodes a single polyprotein that comprises three structural and seven non-structural proteins (White et al., 2016).

There are currently no approved vaccinations or antiviral therapies for the Zika virus. Most preventative efforts are through controlling local mosquito populations and prompting safe sexual intercourse (Signh et al., 2018). The mortality rate of Zika virus infection is 0.1% in adults (Halani et al., 2021), however, the estimated mortality rate for newborns with congenital Zika virus infection is 5.26% (Paixiao et al., 2022).

## Objectives

Although there have been no recent epidemics, the Zika virus remains a public health concern. There are no available vaccines or antiviral medications to treat the disease. There is motivation to estimate the clock rate of the Zika virus to help understand the rate at which it could potentially develop mutations that may result in greater virulence. Subsequently, determining the reproduction number ($R_e$) of the Zika virus provides insight into its potential for subsequent outbreaks. Lastly, determining the date of the most recent common ancestor, can help to understand the historical and ecological factors that contributed to its emergence and spread, and could be a topic of scientific curiosity.

## Methods

### Data and Metadata Collection

I collected the Zika virus sequencing data from the National Center for Biotechnology Information nucleotide database with the query: "`esearch -db nucleotide -query 'txid64320[Organism] AND complete[Title] AND 9000:12000[slen]' | efetch -format gb > zika.gb`". To convert the genbank file into fasta format, I utilized the script from lab2, `parse_genbank.py`.

To extract metadata from the sequences into `zika_metadata.csv`, I ran the `get_metadata.py` script from the PoonLab/Bioplus repository. With the metadata file, I added the collection date to the sequence labels through the PoonLab/Bioplus script, `add_dates.py`. Additionally, I utilized this script to drop sequences with incomplete collection dates.

### Data Cleaning and Downsampling

According to the BEAST2 developer pages, accounting for extensive lengths of ambiguous nucleotides typically increases the computational time to calculate the likelihood of a tree (Bouckaert, 2022b). Although this same webpage states that small segments of ambiguous nucleotides will not affect the speed of the analysis, the developers never specified a threshold or guideline on determining the length of such "small" segments. For this study, I chose to remove sequences that had greater than 10 ambiguous nucleotides by developing and running the script `remove_ambiguous_seq.py`.

After removing samples with ambiguous nucleotides, I downsampled the data by randomly selecting 10 sequences per year through developing and using the script `downsample.py`. The 10 sample threshold was chosen to ensure that at least 100 sequences were selected for subsequent analysis. For years with fewer than 10 sampled sequences, all available sequences were included in the subsample.

I used mafft version 7.490 (Katoh & Standley, 2013) with the argument `--auto` to align the sequences. I then visualized the alignment on AliView version 1.30 (Larsson, 2014) and performed a manual inspection of the alignment to remove any gap-introducing sequences with a custom script, `remove_outlier_sequences.py`. Finally, I changed the date format of the alignment to decimal years to simplify BEAUTi's interpretation for tip dates through developing the script `change_dec_year.py`.

**Estimating Origin, $R_e$ and Clock Rate With BEAST2**

The developers of BEAST2 reported a good estimate for the starting tree will reduce the length of the burn-in period (Bouckaert, 2022b). To generate this estimate, I used IQ-TREE2 version 2.0.7 (Minh et al., 2020) with the argument `-m MFP` to construct a maximum likelihood tree while testing for the best substitution model based on the lowest Bayesian Information Criterion (BIC).

I implemented the Birth-Death Skyline Serial tree model from the BDSKY package version 1.5.1 (Stadler et al., 2012) in BEAST2 version 2.7.7 (Bouckaert, 2014) to estimate the values for origin, clock rate, and $R_e$ based on the sequence alignment. Before 2007, the Zika virus rarely infected humans, suggesting shifts in $R_e$ over time, which this tree model accounts for. Aside from variations in $R_e$, this model also allows for changes in becomeUninfectiousRate, the recovery rate plus the sampling rate, and SamplingProportion, the proportion of infected individuals sampled.

To select an appropriate substitution model, I used the model best determined by IQ-TREE2 according to BIC. For the clock model, I attempted to run the simulation with the optimized relaxed clock, as there is more biological evidence supporting varying clock rate between lineages (Bromham &

Penny, 2003). Additionally, given that some sequences contained ambiguous nucleotides, I used the ambiguities option.

Community resources for BEAST2, Taming the Beast (Barido-Sottani et al., 2017), reported that choosing the dimensions for $R_e$ is arbitrary and typically requires empirical testing. In their BDSKY serial tutorial, they inferred $R_e$ estimates for Hepatitis C Virus (HCV) using a becomeUninfectiousRate and SamplingProportion dimensions of one, and an $R_e$ dimension of 10. Their detected changes in $R_e$ were similar to Dr. Stadler's 2012 BDSKY serial tests with HCV, both showing a rise in $R_e$ in the 20th century. These results suggest that the reduced dimensions for becomeUninfectiousRate and SamplingProportion could be sufficient in BDSKY serial analyses. Utilizing the community defined parameters, I kept the dimensions for becomeUninfectiousRate and SamplingProportion to one, and the dimension for $R_e$ to 10.

As reported by Dr. Louca et al. (2021), the parameters within a birth-death model are confounded with one another, and thus, estimates are more accurate when some parameters are constrained to their true values. Finding the true values for the Zika virus is not biologically possible, however, I made attempts to create informative priors with constraints that matched previous findings in literature. I changed the origin prior distribution from uniform to log normal with a M=5 and S=0.5 (95% interval= 55.7 - 395 [equivalent to ~1968 - 1629], median=148 [equivalent to 1876]), to reflect the knowledge that the first detection of the Zika virus collection was in 1947. Additionally, previous studies estimated $R_e$ for the Zika virus to be between 1.6 to 6.6 (Lessler et al., 2016), and consequently, I updated the $R_e$ prior distribution to be a log normal distribution with M=0.5 and S=1 (95% interval=0.23 - 8.54, median=1.65). I set the Markov Chain Monte Carlo Chain (MCMC) to 100 million states to allow for sufficient sampling of parameter space.

Lastly, I ran a replica of this simulation in BEAST2 on a different seed to test for convergence. Taming the Beast reports that independent runs converging to the same posterior is a stronger indication that the simulations have truly converged than a longer MCMC (Barido-Sottani et al., 2017).

**Visualization**

To visualize the phylogenetic trees generated by IQTREE and BEAST2, I used FigTree version 1.4.4 (Rambaut, 2018). The developers of BDSKY included an Rscript, `bdsky_plot_1.3.3.R`, in their GitHub repository which visualized the $R_e$ estimates and 95% Highest Posterior Density (HPD) intervals across time. I adapted the Rscript to run with the latest version of BDSKY log files to read and plot $R_e$. Other estimated parameter values were visualized with Tracer version 1.7.2 (Rambaut et al., 2018).

## Results

### Data Collection and Downsampling

The initial sample size for the NCBI query contained 1220 sequences. After running the `add_dates.py` script and dropping sequences which had incomplete collection dates, 998 sequences remained. An additional 205 sequences with at least 10 ambiguous nucleotides were removed, leaving 793 sequences before downsampling. Downsampling the dataset with respect to the year of sampling reduced the number of sequences to 155. Outlier identification of gap-introducing sequences with Aliview identified four problematic sequences. Sample KX198134.2 had over 240 additional nucleotides at the start of the alignment. Samples MT505350.1 and MT505349.1 introduced an internal 713 nucleotide gap. Sample KY766069.1 introduced a 256 nucleotide gap near the end of the alignment. These sequences were removed resulting in the final count of 151 sequences.

The final downsampled dataset is composed of sequences ranging from 10,272 to 10,808 nucleotides (mean=10,687.88), including 120 human-derived, 22 mosquito-derived, four monkey-derived, and five sequences from unknown sources. These samples were collected from 37 distinct geographic origins, with Thailand contributing the largest to the subset (n=45). The date range for the sequences span from 1947 to 2024.

### IQ-TREE2 and BEAST2

IQ-TREE2 reported the best-fit model for the data as `TIM2+F+G4`, which indicates substitution rates of AC=AT and CG=GT, empirical base frequencies, and 4 gamma rates. Furthermore, sequence

alignment analysis revealed three distinct sets of sequences with identical nucleotide composition: KY415989.1/KY415991.1, KY415987.1/KY415990.1/KY415986.1, and KU955593.1/MH158236.1. These seven sequences were kept in subsequent steps of the analysis.

An initial BEAST2 run with the optimized relaxed clock failed to converge after 10 million states. In an attempt to create a simpler model for faster convergence, subsequent runs were changed to utilize the strict clock model instead.

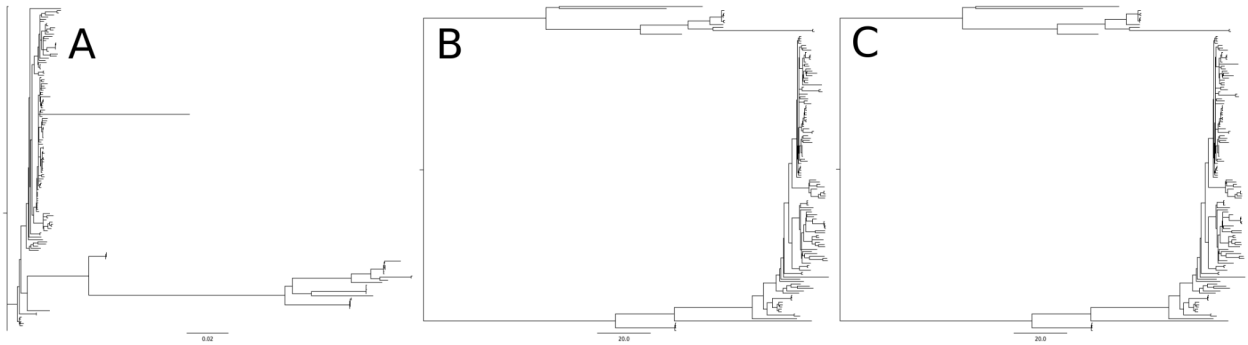


**Figure 1.** *Tree A is the unrooted maximum likelihood tree of the downsampled Zika virus sequences from IQ-TREE2, used as the starting tree in BEAST2. This tree was generated with substitution model TIM2 with empirical frequencies and 4 gamma rates. Tree B and C are the final rooted maximum clade credibility tree generated from the birth-death skyline simulation with a 10% burn-in.*

Both BEAST2 runs finished with 100 million states each. Figure 1 shows the original IQ-TREE2 unrooted start tree, compared to the final maximum clade credibility tree of both runs with a 10% burn-in. Figure 2 shows the posterior traces for both runs with a 10% burn-in, where both runs seemed to converge to -52039.72 (Estimated Sample Size [ESS]=3943) and -52039.59 (ESS=4210). Given that both runs converged to similar posteriors, subsequent parameter values were inferred from their combined log files.

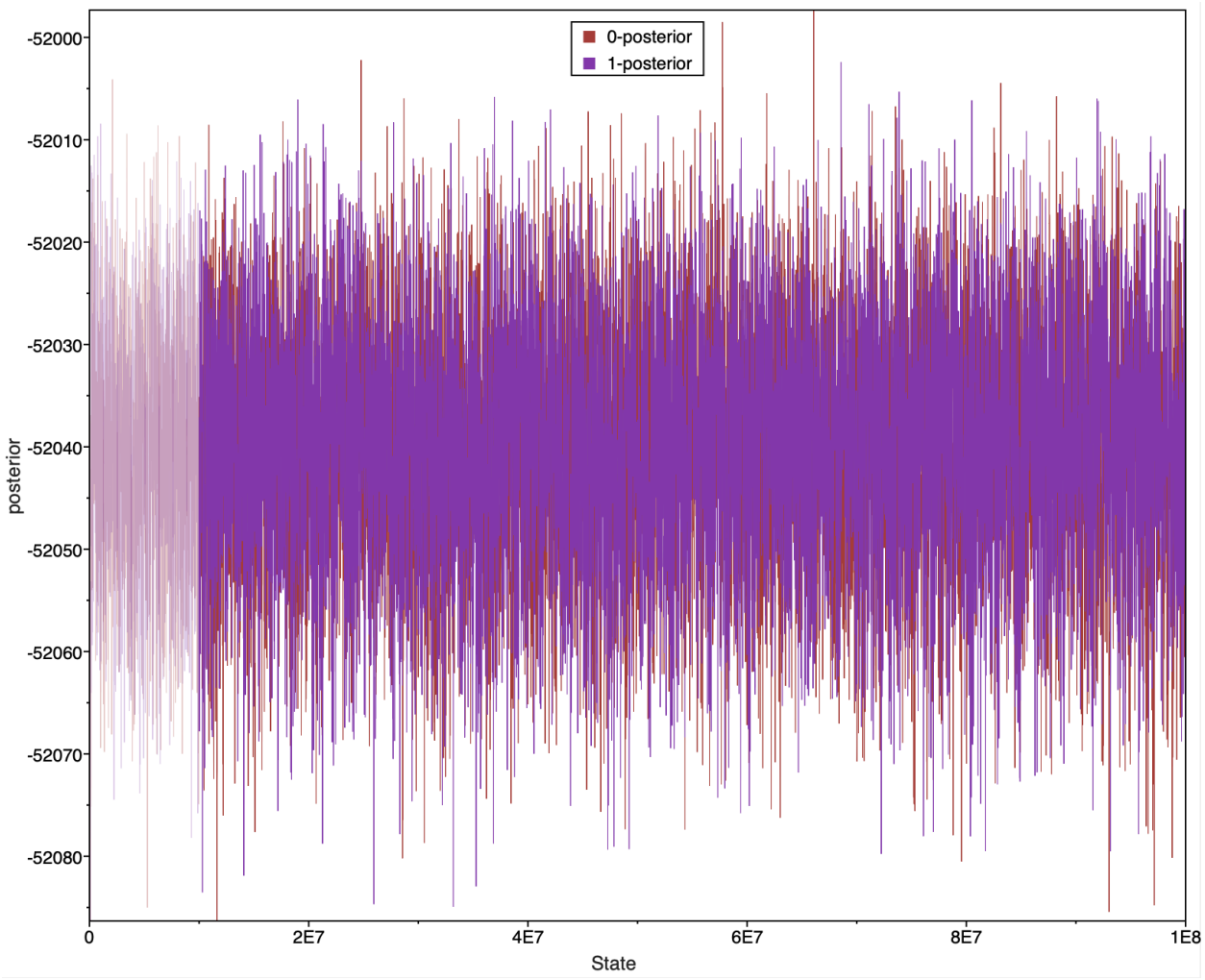


**Figure 2.** *Line plot depicting the combined posterior trace for both BEAST2 runs. The greyed out region is the 10% burn-in of 10 million states.*

The estimate for the origin is 157.47, roughly corresponding to the year of 1866 (ESS=3960, 95% HPD interval=145.24 - 172.67 [equivalent to ~1879 - ~1851]), which is around 10 years earlier than the mean of the prior. The clock rate estimate is 7.34E-4 nucleotide substitutions per site per year (ESS=5548, 95% HPD interval=6.82E-4 - 7.92E-4) from a uniform prior distribution. Figure 3 plots the mean values for $R_e$ estimates 1-10, 0.12, 0.67, 0.45, 0.49, 0.52, 0.37, 0.32, 0.70, 1.01, 1.31, and 2.96 respectively (all ESS>11000), from the origin, 1866, to the latest sampling date, 2024. Most of these estimates were far from the $R_e$ prior mean of 1.65.

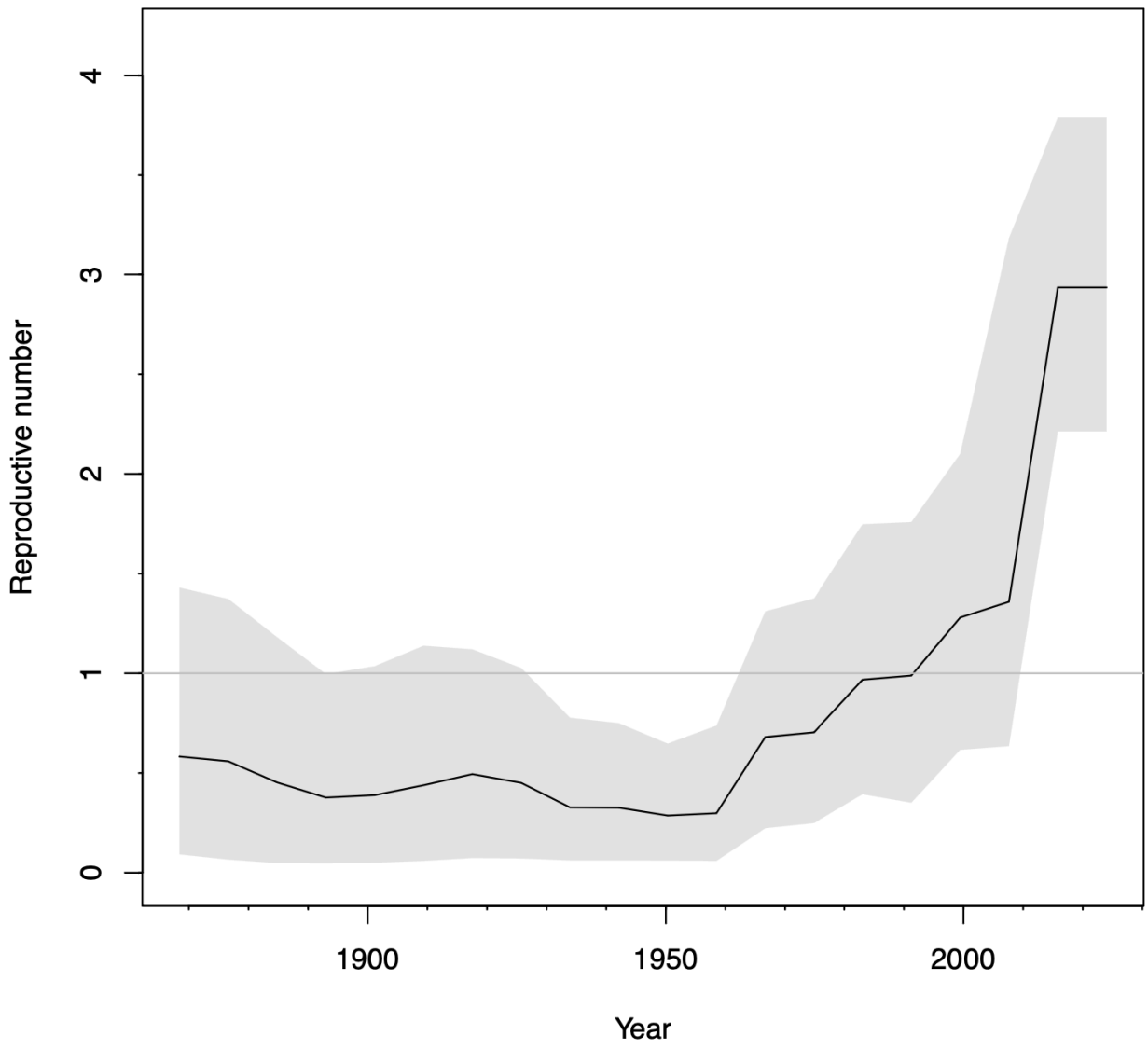


**Figure 3.** *Smoothed line plot depicting 10 equally temporally spaced $R_e$ estimates of the Zika virus from 1866 to 2024, (~16 years per $R_e$ estimate). Data retrieved from the combined log file of both BEAST2 runs with a 10% burn-in. Grey banded regions indicate the 95% highest posterior density interval for each estimate.*

## Discussion

### Broader Implications and Contextualization

From 1866 to around the 1960s, my estimated $R_e$ remained around 0.5. It then rose above 1.0 in the 1990s. These estimated $R_e$ changes are aligned with the global epidemiologic data for the spread of

the Zika virus. Wikan & Smith (2017) suggests that the first zoonotic event occurred in 1966, which correlates to the rise in the $R_e$ estimates. The increase in Re above 1.0 in the early 2000s is consistent with the first outbreak in Micronesia in 2007 (Chang et al., 2016). Finally, the highest Re value of 2.96 aligns with the largest epidemic in 2015–2016. This consistency between our estimates and epidemiological data supports the reliability of our $R_e$ estimations.

Figure 1 shows two clades in Trees B and C diverging from the root node. Current literature identifies two distinct strains of the Zika virus: the Asian and African variant (White et al., 2016). It is possible that these variants have different $R_e$ values, and our estimates represent an average of the two. Repeating this analysis while separating the variants could help identify differences in their transmission dynamics. Separate estimates of $R_e$ for each strain could inform public health strategies in regions where each variant predominates.

My estimate of the current $R_e$ for the Zika virus is 2.96. This result suggests the possibility of another outbreak, which supports this virus being a public health concern. However, the final $R_e$ estimate spans from approximately 2008 to 2024. This estimate may be inflated by the higher transmission of the 2015-2016 epidemic and may not accurately reflect current transmission dynamics. Future analyses should continue with larger dimensions of $R_e$ to get a higher temporal resolution for greater accuracy. Still, the countries that currently have high mosquito populations that carry the Zika virus, such as Brazil (CDC, 2025), should maintain programs that attempt to control the spread of the disease.

Additionally, BEAST2 estimated the origin of the Zika virus to be around 1866. This date predates the first recorded primate infection in 1954, suggesting that the virus existed in animal reservoirs for nearly a century spilling over into the human population. These findings highlight the need for further research into the evolutionary history of Zika virus, including the ecological and molecular factors that facilitated its eventual emergence as a human pathogen.

Lastly, BEAST2 reported an average clock rate of 7.249E-4 nucleotide substitutions per site per year. This estimate is lower than some RNA viruses, such as SARS-COV-2 with an estimated clock rate of 8.47E-3 nucleotide substitutions per site per year for the alpha variant (Tay et al., 2022). This

comparatively lower evolutionary rate could indicate that Zika virus mutations require adaptation to both mosquito vectors and human hosts (Yu et al., 2021). These selective pressures may constrain mutations in viral protein. Understanding these evolutionary constraints could inform vaccine development by identifying conserved genomic regions that are less prone to variation, warranting further investigation.

**Limitations**

Due to computational and time limitations, I downsampled the data with respect to the sampling year. Although sequences were chosen randomly, this removes biologically relevant information, which could change the estimations of BEAST2. Subsequent analysis would benefit from utilizing a larger dataset of available sequences.

The BDSKY model assumes that sampled sequences are removed (Stadler, 2012). Although people diagnosed with the Zika virus may be more cautious as to not transmit the disease, the assumption that they are removed from the infectious pool is false. Additionally, the model I used only had 10 dimensions for $R_e$, where each estimate spanned around 16 years. With the WHO declaring and concluding the Zika virus epidemic within the span of 9 months (WHO, 2022), the rate of change in $R_e$ is likely in much finer 16 years.

In regards to becomeUninfectiousRate and SamplingProporition, the model I used had their dimension set to 1. A constant value for these parameters is likely an incorrect assumption. Public health agencies have now implemented preventative plans, likely reducing the becomeUninfectiousRates from the height of the pandemic (Signh et al., 2018). Additionally, as technology improves and the awareness of the Zika virus increases, the samplingProportion likely increased from the time it was first sequenced in 1947. Future analyses should attempt to include a higher number of dimensions for all three parameters while providing more sequence data to help with convergence.

**Bibliography**


Barido-Sottani, J., Bošková, V., Plessis, L. D., Kühnert, D., Magnus, C., Mitov, V., Müller, N. F., PečErska, J., Rasmussen, D. A., Zhang, C., Drummond, A. J., Heath, T. A., Pybus, O. G., Vaughan, T. G., & Stadler, T. (2017). Taming the BEAST—A Community Teaching Material Resource for BEAST 2. *Systematic Biology* (Vol. 67, Issue 1, pp. 170–174). Oxford University Press (OUP). https://doi.org/10.1093/sysbio/syx060

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. A. Prlic (Ed.), *PLoS Computational Biology* (Vol. 10, Issue 4, p. e1003537). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pcbi.1003537

Bouckaert, R. (2022a, February). *BEAST2 Help Me Choose Standard Template -- Partitions Panel.* Beast2. https://beast2-dev.github.io/hmc/hmc/Standard/Partitions/

Bouckaert, R. (2022b, February). *BEAST2 Help Me Choose Standard Template -- Starting Tree Panel.* Beast2. https://beast2-dev.github.io/hmc/hmc/Standard/Starting_tree/index.html

Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics* (Vol. 4, Issue 3, pp. 216–224). Springer Science and Business Media LLC. https://doi.org/10.1038/nrg1020

Chang, C., Ortiz, K., Ansari, A., & Gershwin, M. E. (2016). The Zika outbreak of the 21st century. In *Journal of Autoimmunity* (Vol. 68, pp. 1–13). Elsevier BV. https://doi.org/10.1016/j.jaut.2016.02.006

Center For Disease Control and Prevention. (2025, May 15). *Countries & Territories at Risk for Zika virus.* Zika virus. https://www.cdc.gov/zika/geo/index.html

Gubler, D. J., Vasilakis, N., & Musso, D. (2017). History and Emergence of Zika virus. In The Journal of Infectious Diseases (Vol. 216, Issue suppl_10, pp. S860–S867). *Oxford University Press (OUP)*. https://doi.org/10.1093/infdis/jix451

Halani, S., Tombindo, P. E., O'Reilly, R., Miranda, R. N., Erdman, L. K., Whitehead, C., Bielecki, J. M., Ramsay, L., Ximenes, R., Boyle, J., Krueger, C., Willmott, S., Morris, S. K., Murphy, K. E., &

Sander, B. (2021). Clinical manifestations and health outcomes associated with Zika virus infections in adults: A systematic review. In P. F. C. Vasconcelos (Ed.), *PLOS Neglected Tropical Diseases* (Vol. 15, Issue 7, p. e0009516). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pntd.0009516

Hennessey, M., Fischer, M., Staples, E. (2016, January). *Zika Virus Spreads to New Areas — Region of the Americas, May 2015–January 2016*. Center for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/65/wr/mm6503e1.htm

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* (Vol. 30, Issue 4, pp. 772–780). Oxford University Press (OUP). https://doi.org/10.1093/molbev/mst010

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* (Vol. 30, Issue 22, pp. 3276–3278). Oxford University Press (OUP). https://doi.org/10.1093/bioinformatics/btu531

Lessler, J., Chaisson, L. H., Kucirka, L. M., Bi, Q., Grantz, K., Salje, H., Carcelen, A. C., Ott, C. T., Sheffield, J. S., Ferguson, N. M., Cummings, D. A. T., Metcalf, C. J. E., & Rodriguez-Barraquer, I. (2016). Assessing the global threat from Zika virus. *Science* (Vol. 353, Issue 6300). American Association for the Advancement of Science (AAAS). https://doi.org/10.1126/science.aaf8160

Louca, S., McLaughlin, A., MacPherson, A., Joy, J. B., & Pennell, M. W. (2021). Fundamental Identifiability Limits in Molecular Epidemiology. In K. Crandall (Ed.), *Molecular Biology and Evolution* (Vol. 38, Issue 9, pp. 4010–4024). Oxford University Press (OUP). https://doi.org/10.1093/molbev/msab149

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. In E. Teeling (Ed.), *Molecular Biology and Evolution* (Vol. 37, Issue 5, pp. 1530–1534). Oxford University Press (OUP). https://doi.org/10.1093/molbev/msaa015

Mozūraitis, R., Cirksena, K., Raftari, M., Hajkazemian, M., Mustapha Abiodun, M., Brahimi, J., Radžiutė, S., Apšegaitė, V., Bernotienė, R., Ignatowicz, L., Hick, T., Kirschning, A., Lenman, A., Gerold, G., & Emami, S. N. (2025). Zika virus modulates human fibroblasts to enhance transmission success in a controlled lab-setting. *Communications Biology* (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s42003-025-07543-9

Paixao, E. S., Cardim, L. L., Costa, M. C. N., Brickley, E. B., de Carvalho-Sauer, R. C. O., Carmo, E. H., Andrade, R. F. S., Rodrigues, M. S., Veiga, R. V., Costa, L. C., Moore, C. A., França, G. V. A., Smeeth, L., Rodrigues, L. C., Barreto, M. L., & Teixeira, M. G. (2022). Mortality from Congenital Zika Syndrome — Nationwide Cohort Study in Brazil. *New England Journal of Medicine* (Vol. 386, Issue 8, pp. 757–767). Massachusetts Medical Society. https://doi.org/10.1056/nejmoa2101195

Rambaut, A. (2018) FigTree v1.4.4. Institute of Evolutionary Biology, *University of Edinburgh*, Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. In E. Susko (Ed.), *Systematic Biology* (Vol. 67, Issue 5, pp. 901–904). Oxford University Press (OUP). https://doi.org/10.1093/sysbio/syy032

Singh, R. K., Dhama, K., Khandia, R., Munjal, A., Karthik, K., Tiwari, R., Chakraborty, S., Malik, Y. S., & Bueno-Marí, R. (2018). Prevention and Control Strategies to Counter Zika Virus, a Special Focus on Intervention Approaches against Vector Mosquitoes—Current Updates. *Frontiers in Microbiology* (Vol. 9). Frontiers Media SA. https://doi.org/10.3389/fmicb.2018.00087

Stadler, T., Kühnert, D., Bonhoeffer, S., & Drummond, A. J. (2012). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences* (Vol. 110, Issue 1, pp. 228–233). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1207965110

Tay, J. H., Porter, A. F., Wirth, W., & Duchene, S. (2022). The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate. In T. Leitner (Ed.), *Molecular*

*Biology and Evolution* (Vol. 39, Issue 2). Oxford University Press (OUP). https://doi.org/10.1093/molbev/msac013

White, M. K., Wollebo, H. S., David Beckham, J., Tyler, K. L., & Khalili, K. (2016). Zika virus: An emergent neuropathological agent. *Annals of Neurology* (Vol. 80, Issue 4, pp. 479–489). Wiley. https://doi.org/10.1002/ana.24748

Wikan, N., & Smith, D. R. (2017). First published report of Zika virus infection in people: Simpson, not MacNamara. *The Lancet Infectious Diseases* (Vol. 17, Issue 1, pp. 15–17). Elsevier BV. https://doi.org/10.1016/s1473-3099(16)30525-4

World Health Organization. (2022, December 8). *Zika Virus.* https://www.who.int/news-room/fact-sheets/detail/zika-virus

Yu, X., Shan, C., Zhu, Y., Ma, E., Wang, J., Wang, P., Shi, P.-Y., & Cheng, G. (2021). A mutation-mediated evolutionary adaptation of Zika virus in mosquito and mammalian host. *Proceedings of the National Academy of Sciences* (Vol. 118, Issue 42). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2113015118