APPLICATION

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

nosoi: A stochastic agent-based transmission chain simulation framework in R

Sebastian Lequime[1,2] | **Paul Bastide**[1,3] | **Simon Dellicour**[1,4] | **Philippe Lemey**[1] | **Guy Baele**[1]

[1]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium

[2]Cluster of Microbial Ecology, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

[3]IMAG, CNRS, University of Montpellier, Montpellier, France

[4]Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium

Correspondence
Sebastian Lequime
Email: s.j.j.lequime@rug.nl

Abstract

1. The transmission process of an infectious agent creates a connected chain of hosts linked by transmission events, known as a transmission chain. Reconstructing transmission chains remains a challenging endeavour, except in rare cases characterized by intense surveillance and epidemiological inquiry. Inference frameworks attempt to estimate or approximate these transmission chains but the accuracy and validity of such methods generally lack formal assessment on datasets for which the actual transmission chain was observed.

2. We here introduce nosoi, an open-source R package that offers a complete, tunable and expandable agent-based framework to simulate transmission chains under a wide range of epidemiological scenarios for single-host and dual-host epidemics. nosoi is accessible through GitHub and CRAN, and is accompanied by extensive documentation, providing help and practical examples to assist users in setting up their own simulations.

3. Once infected, each host or agent can undergo a series of events during each time step, such as moving (between locations) or transmitting the infection, all of these being driven by user-specified rules or data, such as travel patterns between locations.

4. nosoi is able to generate a multitude of epidemic scenarios, that can—for example—be used to validate a wide range of reconstruction methods, including epidemic modelling and phylodynamic analyses. nosoi also offers a comprehensive framework to leverage empirically acquired data, allowing the user to explore how variations in parameters can affect epidemic potential. Aside from research questions, nosoi can provide lecturers with a complete teaching tool to offer students a hands-on exploration of the dynamics of epidemiological processes and the factors that impact it. Because the package does not rely on mathematical formalism but uses a more intuitive algorithmic approach, even extensive changes of the entire model can be easily and quickly implemented.

1 | INTRODUCTION

Infectious disease events, especially those resulting from novel emerging pathogens, have significantly increased over the past few decades, possibly as a result of alterations in various environmental, biological, socioeconomic and political factors (Chan et al., 2010). By definition, infectious agents need to spread through transmission between hosts. If successful, the resulting transmission process creates a connected chain of hosts linked by transmission events, usually called a transmission chain. Transmission is highly stochastic and can be influenced by a wide array of intrinsic and extrinsic factors, such as within-host dynamics and environmental or host behavioural factors. Reconstruction of transmission chains, however, remains difficult to achieve, except in certain rare cases characterized by intense surveillance and epidemiological inquiry (Mollentze et al., 2014; Worby et al., 2016).

Molecular data may represent a critical asset in reconstructing the transmission history of a pathogen (Campbell, Cori, Ferguson, & Jombart, 2019; De Maio, Worby, Wilson, & Stoesser, 2018; Didelot, Fraser, Gardy, & Colijn, 2017; Didelot, Gardy, & Colijn, 2014; Worby et al., 2016). Often, however, the relationship between individual cases is too distant to allow for the perfect reconstruction of a transmission chain. In that context, the study of infectious agents' genomic sequences can be used to reconstruct, under an evolutionary model, their likely evolutionary history. These reconstructions rely on evolution occurring on the same time-scale as the epidemic or transmission process, which is the case for most fast-evolving pathogens such as RNA viruses (Romero-Severson, Skar, Bulla, Albert, & Leitner, 2014; Ypma, van Ballegooijen, & Wallinga, 2013). The inferred evolutionary history has been used in recent years to estimate the timing, the origin or the effectiveness of mitigation measures of several epidemics (Dellicour et al., 2018; Dudas, Carvalho, Rambaut, & Bedford, 2018; Grubaugh et al., 2019; Hill et al., 2019).

The accuracy, validity or limitations of both currently available and future methods, however, generally lack formal assessment on datasets for which we have been able to observe the actual geographical spread and the complex factors that shaped its pattern. In that context, a simulated dataset is extremely useful as the exact transmission history is known and can be compared to the histories inferred from different software packages. The last decade has seen the development of several integrated epidemic and genetic simulation tools that can be used to assess the performance of some of these models, such as TreeSim (Stadler & Bonhoeffer, 2013), seedy (Worby & Read, 2015), outbreaker2 (Campbell et al., 2018) or favites (Moshiri, Ragonnet-Cronin, Wertheim, & Mirarab, 2019).

While undoubtedly useful, these tools fall short in accommodating a wide range of epidemiological scenarios. In particular, arboviral (e.g. Zika, dengue or yellow fever) outbreaks, where two types of hosts participate in the epidemic process, are poorly modelled. These hosts are characterized by drastically different behaviour or infection dynamics and cannot be accurately modelled using a single host type. Furthermore, geographical location diffusion is simulated in these tools, when possible, on a contact network or in discrete space. Yet, recent years have seen the development of methods taking advantage of phylogeographical diffusion in continuous space (Dellicour, Rose, Faria, Lemey, & Pybus, 2016; Lemey, Rambaut, Welch, & Suchard, 2010), creating a need for epidemiological simulations in a continuous space.

To enable the performance assessment of these methods under complex and realistic scenarios, including spread in continuous space or arbovirus outbreaks, we present nosoi, a flexible agent-based transmission chain simulator implemented as an open-source R package (R Core Team, 2019).

2 | CHARACTERISTICS

nosoi generalizes and significantly extends a basic model that allowed individual humans and mosquitoes—each one being characterized by a unique set of infection parameters—to interact within a simulated environment (Fontaine et al., 2018). It was initially designed to model real-world arboviral epidemics unfolding under varying within-host dynamics (Fontaine et al., 2018).

nosoi employs agent-based modelling, which focuses on the individual active entities—known as (autonomous) agents—of a system and defines their behaviour and the interactions between them. The main interest then lies in the global dynamics of and the complex phenomena within the system that emerges from the interactions of the many individual behaviours. Within nosoi, the agents' behaviour is governed by user-specified rules that can accommodate high levels of stochasticity at each level of the epidemic process. Agents can experience dual-host dynamics, such as those from human and mosquito populations, and exist in structured populations, with different behaviours according to host type and/or structure. Population structure can either be absent, discrete (e.g. different categories) or continuous (such as geographical space). In these structures, agents can trigger a movement, a contact or a transmission event, with the probability of such an event occurring being potentially host-, individual-, structure- and/or time-dependent. These agents are recruited when infected and can either recover or die from the infection, resulting in their removal from the simulation. The status and location of each agent are assessed according to the model during each step of the discretized time of the simulation (Figure 1). The simulation ends when the user-specified value of the number of infected agents or when the targeted simulation time is reached.

In essence, `nosoi` allows the user to simulate and keep track of one or more transmission chains occurring during an infectious disease outbreak and, as such, to store and output a (collection of) transmission tree(s). Genetic data can be subsequently simulated along each transmission tree using sequence simulation software such as πbuss (Bielejec et al., 2014) or SantaSim (Jariani et al., 2019), which can then serve as input for phylodynamic

FIGURE 1 Schematic of status and location assessment for each agent (in case of a structured population), or host, during each discretized time step of the simulation. Optional steps in the simulation framework are shown in shades of green and are only performed in case of a structured (either discrete or continuous) population. Several factors (embedded in the gray box), either individually or globally set, can influence these steps according to user-specified settings

inference methods. `nosoi` is accompanied by extensive tutorials, helping the user to set up and visualize their simulation, available as documentation in the package, or at https://slequime.github.io/nosoi/.

3 | PRACTICAL EXAMPLE

We here showcase `nosoi` with the starting scenario of a single human infected with an Ebolavirus-like pathogen in West Africa. The simulated epidemic unfolds in a geographically structured host population, specifically in a continuous geographic space, for 365 days or discrete time-steps. Within-host dynamics, influencing the probability of exiting the simulation (dying or recovering) and the between-host transmission probability, are modelled according to published literature that describes Ebolavirus infection in humans (Casillas, Nyamathi, Sosa, Wilder, & Sands, 2003; Skrip et al., 2017). The remaining parameters (number of daily contacts, probability of movement and standard deviation of the random walk in continuous space) were empirically set. The number of daily contacts is restricted by the number of people living in the area, as provided by spatial demographics data obtained from WorldPop (www.worldpop.org), to avoid reaching locally unrealistic counts of infected humans. The complete specification and accompanying code for this simulation are available as a document on
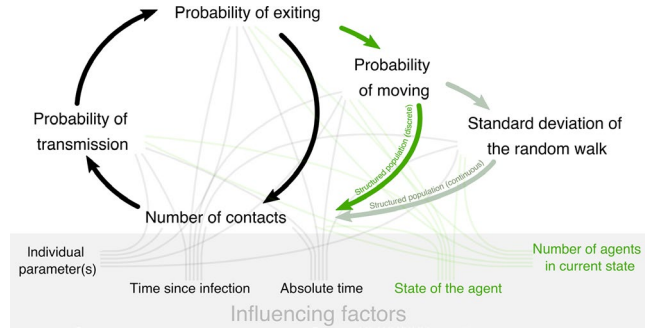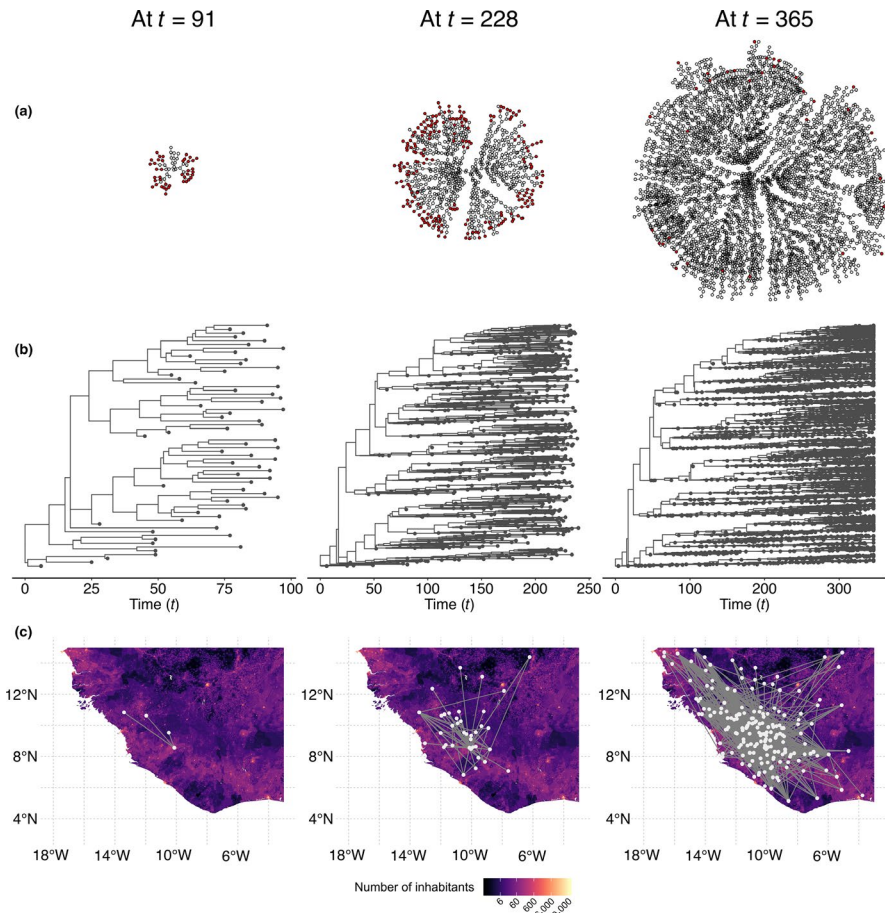

FIGURE 2 Visualization of a simulated Ebolavirus-like transmission chain in West Africa at three time-points (91, 228 and 365 days after the introduction of the first infected host), represented as (a) a network, (b) a tree or (c) a tree mapped on the continuous space the simulation took place in

`nosoi`'s website (https://slequime.github.io/nosoi/articles/examples/ebola.html).

Over the course of 365 days, the simulation has yielded 3,603 infected agents. The average number of secondary cases per agent is 1.12, which is roughly coherent with previous epidemiological estimates of R_0 for previous Ebolavirus outbreaks (Van Kerkhove, Bento, Mills, Ferguson, & Donnelly, 2015). The increase in infected agents' number is exponential, as would be expected considering the specifications of the model, that is, absence of intervention strategies or changes in the simulated environment.

The transmission chain can be represented either as a network (Figure 2a) or as a tree (Figure 2b) that can be mapped in the continuous space in which the epidemic took place (Figure 2c). The tree representation of the transmission chain can be seen as the genealogy of the pathogen population over which molecular evolution generates the observed sequence data, then used to reconstruct this same history. In this representation, each internal node is a transmission event, each tip represents the exit point in time of an agent, and the root is the starting point in time of the initially infected agent. Branches or sets of connected branches represent the life span of each agent. This tree is binary, counts as many tips as the total number of infected agents and as many internal nodes as transmission events.

Other examples are available on `nosoi`'s website illustrating various scenarios, such as spread of a dengue-like pathogen (dual-host) in a discrete space or an unstructured population of hosts. The tutorials also provide guidelines on how to set up simulations in various combinations of settings currently available.

4 | USES

Trends in globalization, including expansion in international travel and trade, have extended the reach and increased the pace at which infectious diseases spread (Chan et al., 2010). These trends provide infectious agents with ample opportunities to establish and spread successfully, but many practical difficulties remain in accurately inferring key aspects of an epidemic. Standard testing of models of spread typically focuses on using that same model to generate simulated data, which offers important but limited insights and mostly provides a test of proper implementation and a way to compare different methodologies. `nosoi`, however, is a phylogenetic model-independent agent-based simulation framework that offers realistic and complex epidemiological scenarios. As such, it enables accurate testing of popular inference methods in both discrete and continuous phylogeography using either maximum-likelihood (Ishikawa, Zhukova, Iwasaki, & Gascuel, 2019) or Bayesian inference (Lemey, Rambaut, Drummond, & Suchard, 2009; Lemey et al., 2010; Suchard et al., 2018), which are widely used in pathogen phylodynamics. In that regard, an interesting application of our proposed simulation framework could be to study the increasingly popular structured coalescent models (Bouckaert et al., 2019; De Maio, Wu, O'Reilly, & Wilson, 2015;

Müller, Rasmussen, & Stadler, 2017), and to compare their accuracy under realistic epidemiological transmission scenarios against discrete phylogeographical inference.

`nosoi` enables the simulation of real-life scenarios of viral outbreaks, and we provide several example scenarios to showcase its capabilities to generate a single transmission chain using different settings. An important aspect is that the resulting transmission tree, which describes the transmission events between infected hosts, differs from the phylogenetic tree, which describes the ancestral genetic relationships between pathogens sampled from these hosts. In that regard, it is crucial to acknowledge the growing number of methods that infer either phylogenetic trees, transmission trees or jointly estimate both (for an overview, we refer to Baele, Suchard, Rambaut, and Lemey (2017)).

Apart from assessing the performance of various methods in reconstructing geographical spread or the dynamics of an infectious agent, `nosoi` can prove useful for assessing the performance of classic deterministic SIR and SIRS compartmental models (Kermack & McKendrick, 1927). These epidemiological models estimate the theoretical number of people infected with a contagious illness in a closed population over time under some assumptions. For example, the original SIR model assumes that the population size is fixed, that the incubation period of the infectious agent is instantaneous and that the duration of infectivity is the same as the length of the disease. It also assumes a completely homogeneous population with no age, spatial or social structure. These assumptions can be matched as closely as possible by the user-defined settings in `nosoi` or be violated in more realistic settings, allowing to examine the sensitivity of the deterministic models to the assumptions under a complex and fine-tuned epidemiological scenario.

`nosoi` also offers, in line with its initial purpose (Fontaine et al., 2018), a comprehensive framework to leverage empirically acquired data. A pathogen's within-host dynamics characterized in laboratory settings can be embedded into a full stochastic epidemiological model, allowing the user to explore how variation can affect its epidemic potential.

Aside from research questions, `nosoi` can provide lecturers with a complete teaching tool to offer students a hands-on exploration of the dynamics of epidemiological processes and the factors that impact it. Because the package does not rely on mathematical formalism but uses a more intuitive algorithmic approach, even extensive changes of the entire model or part of it can be easily and quickly implemented. The documentation provides suggestions for visualization using well-known external R-packages, such as GGPLOT2 (Wickham, 2009) or GGTREE (Yu, Lam, Zhu, & Guan, 2018; Yu, Smith, Zhu, Guan, & Lam, 2016). The package is also fully integrated in the R and phylogenetic environments, and, through the use of the TREEIO and TIDYTREE R packages (Wang et al., 2019), simulated transmission trees can be exported in a wide variety of formats for downstream analyses, such as the BEAST (Suchard et al., 2018) or JPLACE (Matsen, Hoffman, Gallagher, & Stamatakis, 2012) formats.

In summary, `nosoi` provides a complete, tunable and expandable framework to simulate epidemiological processes based on transmission chains, in a user-friendly manner. Accessible through GitHub and the CRAN, the code is well covered by unitary tests and accompanied by extensive documentation, providing help and practical examples to users. Open-source and coded in the widely used R language, it allows users to customize their model by implementing new mechanisms for all or part of the core model. In addition, and contrary to other available tools, by decoupling sequence evolution from the epidemiological process, it can connect to any external sequence simulator, allowing the user to choose a tool and model that can address the biological question of interest.

AUTHORS' CONTRIBUTIONS
S.L. designed and conceived the package, and wrote its documentation; P.B. and S.D. provided editing and optimization to the package R code; P.L. and G.B. supervised and guided the project; S.L. and G.B. wrote the initial draft. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT
The package is available on GitHub (https://github.com/slequime/nosoi) and the CRAN (https://cran.r-project.org/package=nosoi). The reviewed version of the package presented in this manuscript is available through Zenodo (https://doi.org/:10.5281/zenodo.3860006). The complete specification and accompanying code for the simulation presented in this manuscript are available as a document on `nosoi`'s website (https://slequime.github.io/nosoi/articles/examples/ebola.html).

ORCID
Sebastian Lequime https://orcid.org/0000-0002-3140-0651
Paul Bastide https://orcid.org/0000-0002-8084-9893
Simon Dellicour https://orcid.org/0000-0001-9558-1052
Philippe Lemey https://orcid.org/0000-0003-2826-5353
Guy Baele https://orcid.org/0000-0002-1915-7732

REFERENCES
Baele, G., Suchard, M. A., Rambaut, A., & Lemey, P. (2017). Emerging concepts of data integration in pathogen phylodynamics. *Systematic Biology*, *66*, e47–e65.

Bielejec, F., Lemey, P., Carvalho, L. M., Baele, G., Rambaut, A., & Suchard, M. A. (2014). πBUSS: A parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics*, *15*, 133. https://doi.org/10.1186/1471-2105-15-133

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*, e1006650.

Campbell, F., Cori, A., Ferguson, N., & Jombart, T. (2019). Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Computational Biology*, *15*, e1006930. https://doi.org/10.1371/journal.pcbi.1006930

Campbell, F., Didelot, X., Fitzjohn, R., Ferguson, N., Cori, A., & Jombart, T. (2018). outbreaker2: A modular platform for outbreak reconstruction. *BMC Bioinformatics*, *19*, 320–328. https://doi.org/10.1186/s12859-018-2330-z

Casillas, A. M., Nyamathi, A. M., Sosa, A., Wilder, C. L., & Sands, H. (2003). A current review of Ebola virus: Pathogenesis, clinical presentation, and diagnostic assessment. *Biological Research for Nursing*, *4*, 268–275. https://doi.org/10.1177/1099800403252603

Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., … Brownstein, J. S. (2010). Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 21701–21706. https://doi.org/10.1073/pnas.1006219107

De Maio, N., Worby, C. J., Wilson, D. J., & Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Computational Biology*, *14*, e1006117–e1006123. https://doi.org/10.1371/journal.pcbi.1006117

De Maio, N., Wu, C.-H., O'Reilly, K. M., & Wilson, D. (2015). New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genetics*, *11*, e1005421.

Dellicour, S., Baele, G., Dudas, G., Faria, N. R., Pybus, O. G., Suchard, M. A., … Lemey, P. (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications*, *9*, 2222. https://doi.org/10.1038/s41467-018-03763-2

Dellicour, S., Rose, R., Faria, N. R., Lemey, P., & Pybus, O. G. (2016). SERAPHIM: Studying environmental rasters and phylogenetically informed movements. *Bioinformatics*, *32*, 3204–3206. https://doi.org/10.1093/bioinformatics/btw384

Didelot, X., Fraser, C., Gardy, J., & Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*, 997–1007. https://doi.org/10.1093/molbev/msw275

Didelot, X., Gardy, J., & Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, *31*, 1869–1879. https://doi.org/10.1093/molbev/msu121

Dudas, G., Carvalho, L. M., Rambaut, A., & Bedford, T. (2018). MERS-CoV spillover at the camel–human interface. *eLife*, *7*, 250.

Fontaine, A., Lequime, S., Moltini-Conclois, I., Jiolle, D., Leparc-Goffart, I., Reiner, R. C., & Lambrechts, L. (2018). Epidemiological significance of dengue virus genetic variation in mosquito infection dynamics. *PLOS Pathogens*, *14*, e1007187–21. https://doi.org/10.1371/journal.ppat.1007187

Grubaugh, N. D., Saraf, S., Gangavarapu, K., Watts, A., Tan, A. L., Oidtman, R. J., … Andersen, K. G. (2019). Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. *Cell*, *178*, 1057–1071.e11. https://doi.org/10.1016/j.cell.2019.07.018

Hill, S. C., Vasconcelos, J., Neto, Z., Jandondo, D., Zé-Zé, L., Aguiar, R. S., … Faria, N. R. (2019). Emergence of the Asian lineage of Zika virus in Angola: An outbreak investigation. *The Lancet Infectious Diseases*, *19*, 1138–1147. https://doi.org/10.1016/S1473-3099(19)30293-2

Ishikawa, S. A., Zhukova, A., Iwasaki, W., & Gascuel, O. (2019). A fast likelihood method to reconstruct and visualize ancestral scenarios. *Molecular Biology and Evolution*, *36*, 2069–2085. https://doi.org/10.1093/molbev/msz131

Jariani, A., Warth, C., Deforche, K., Libin, P., Drummond, A. J., Rambaut, A., … Theys, K. (2019). SANTA-SIM: Simulating viral sequence evolution dynamics under selection and recombination. *Virus Evolution*, *5*, 301–308. https://doi.org/10.1093/ve/vez003

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *115*, 700–721.

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, *5*, e1000520. https://doi.org/10.1371/journal.pcbi.1000520

Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, *27*, 1877–1885. https://doi.org/10.1093/molbev/msq067

Matsen, F. A., Hoffman, N. G., Gallagher, A., & Stamatakis, A. (2012). A format for phylogenetic placements. *PLoS ONE*, *7*, e31009. https://doi.org/10.1371/journal.pone.0031009

Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., & Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20133251–20133251. https://doi.org/10.1098/rspb.2013.3251

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O., & Mirarab, S. (2019). FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, *35*, 1852–1861. https://doi.org/10.1093/bioinformatics/bty921

Müller, N. F., Rasmussen, D. A., & Stadler, T. (2017). The structured coalescent and its approximations. *Molecular Biology and Evolution*, *34*, 2970–2981. https://doi.org/10.1093/molbev/msx186

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., & Leitner, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Molecular Biology and Evolution*, *31*, 2472–2482. https://doi.org/10.1093/molbev/msu179

Skrip, L. A., Fallah, M. P., Gaffney, S. G., Yaari, R., Yamin, D., Huppert, A., … Galvani, A. P. (2017). Characterizing risk of Ebola transmission based on frequency and type of case-contact exposures. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*, 20160301. https://doi.org/10.1098/rstb.2016.0301

Stadler, T., & Bonhoeffer, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*, 20120198. https://doi.org/10.1098/rstb.2012.0198

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*. https://doi.org/10.1093/ve/vey016

Van Kerkhove, M. D., Bento, A. I., Mills, H. L., Ferguson, N. M., & Donnelly, C. A. (2015). A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Scientific Data*, *2*, 150019–10. https://doi.org/10.1038/sdata.2015.19

Wang, L.-G., Lam, T.-T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., … Yu, G. (2019). treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Molecular Biology and Evolution*, *60*, 291.

Wickham, H. (2009). *ggplot2*. New York, NY: Springer New York.

Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J. P., … Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Annals of Applied Statistics*, *10*, 395–417. https://doi.org/10.1214/15-AOAS898

Worby, C. J., & Read, T. D. (2015). 'SEEDY' (simulation of evolutionary and epidemiological dynamics): An R package to follow accumulation of within-host mutation in pathogens. *PLoS ONE*, *10*, e0129745. https://doi.org/10.1371/journal.pone.0129745

Ypma, R. J. F., van Ballegooijen, W. M., & Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, *195*, 1055–1062. https://doi.org/10.1534/genetics.113.154856

Yu, G., Lam, T.-T.-Y., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, *35*, 3041–3043.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T.-T.-Y. (2016). ggtree: An rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*, 28–36.