

Reconstructing Pathogen Spread by Incorporating Individual Travel Histories in BEAST

Samuel L. Hong

December 4, 2020

Abstract

Advances in sequencing technologies during the last decade have introduced quicker and cheaper sequencing, to the point where sequencing pathogen populations is now part of routine public health practices. Within this context, phylogenetics (and more specifically phylogeography) has become a popular method to study disease transmission using sequencing data. Here, Bayesian approaches have the benefit of being able to incorporate epidemiological and ecological data to achieve detailed reconstructions of the spatial spread of an epidemic while accounting for phylogenetic uncertainty. In this unit we describe how to reconstruct the spatial spread of a pathogen through Bayesian phylogenetic analysis in BEAST. The protocols focus on how to incorporate molecular, epidemiological and accommodate individual travel history data to perform Bayesian phylogeography in discrete space.

INTRODUCTION

Bayesian Evolutionary Analysis Sampling Trees (BEAST) [Suchard et al., 2018] is a software package that provides a general framework for phylogenetic inference and evolutionary hypothesis testing using molecular sequence data [Drummond and Rambaut, 2007, Drummond et al., 2012, Suchard et al., 2018]. As such, BEAST employs a combination of different types of models (including but not limited to molecular clock models, coalescent models and substitution models) to infer time-calibrated phylogenetic trees from an alignment of time-stamped sequences. Phylogenetic, phylogeographic and phylodynamic inference is performed using Bayesian inference through Markov chain Monte Carlo (MCMC) and in doing so makes use of BEAGLE, a high-performance computational library for statistical phylogenetics [?]. BEAST is widely used in the field of phylodynamics and molecular epidemiology of infectious disease to obtain insights from molecular data through an expanding array of statistical models and estimation procedures.

A specific type of analysis that can be performed within this context is pathogen phylogeographic inference. This type of analysis aims to answer the question of “how did an epidemic spread through space and time?” by jointly reconstructing the evolutionary and geographical history of a pathogen population in the form of annotated time-scaled phylogenies. We can further extend this framework to integrate epidemiological information as a way to identify potential predictors of spatial spread under a generalized linear model (GLM). Such an approach can allow us to, for example, assess the impact of air travel on the global spread of influenza [Lemey et al., 2014]. Because of the flexibility of this framework, Bayesian phylogeographic analyses in BEAST have enjoyed wide success in uncovering the origins of viral lineages [Worobey et al., 2016]

and characterizing pathogen spread to inform public health response [Dudas et al., 2017].

We can use BEAST to model the spatial spread of a pathogen between discrete locations using discrete trait analysis (DTA). This method estimates the probability of pathogen transmission between two locations using models analogous to those used for characterising nucleotide substitution probabilities [Lemey et al., 2009]. Recently, a new approach has been developed to integrate individual travel history data into DTA reconstructions in order to obtain more realistic reconstructions of pathogen spread and mitigate sampling bias [Lemey et al., 2020].

We here provide four related protocols to reproduce the travel history-aware phylogeographic reconstructions performed in [Lemey et al., 2020]. Protocol 1 introduces the GISAID database and provides the required steps to construct a SARS-CoV-2 multiple sequence alignment from the sequences available in GISAID. In Protocol 2, we provide instructions on how to set up a generalized linear model analysis for discrete state phylogeographic inference using BEAUti. Protocol 3 introduces an automated script to modify a BEAUti-generated XML file to incorporate travel history data, which can subsequently be run using BEAST. Finally in Protocol 4, we guide the user through the process of visualizing individual geographic dispersal histories over the posterior distribution of trees, as sampled during BEAST’s estimation process.

PROTOCOL 1: CREATING A SARS-CoV-2 MSA USING SEQUENCES FROM GISAID

The first step in any phylogenetic analysis is to obtain a high-quality multiple sequence alignment (MSA). For SARS-CoV-2 analyses, the largest repository of genomes is available through the GISAID database. The GISAID initiative provides a platform to openly share genomic data of influenza viruses as well as SARS-CoV-2 [Elbe and Buckland-Merrett, 2017]. Access to the database is free (<https://www.gisaid.org/registration/register/>), but requires the user to register and agree to GISAID’s terms of use in order to obtain access.

This protocol describes how to construct a SARS-CoV-2 MSA from sequences downloaded from GISAID. In this example, we will construct an alignment for the 282-taxa dataset used in [Lemey et al., 2020].

Necessary Resources

Hardware

Standard workstation running Linux, MacOS or Windows 10.

Software

A modern web browser (Google Chrome and Mozilla Firefox recommended).
The latest MAFFT [Katoh and Standley, 2013] version (v7.453 used in this protocol)
The latest Aliview [Larsson, 2014] version (v1.26 used in this protocol).
A terminal emulator running a standard Unix shell.

For Windows, a BASH shell can be installed through the Windows Subsystem for Linux. For more information on this, see <https://docs.microsoft.com/en-us/windows/wsl/>

Files

A list of GISAID accession identifiers.

A FASTA file with the SARS-CoV-2 reference genome UTR sequences

Example files can be found at

<https://github.com/hongsamL/travHistProtocol/tree/main/files/Protocol1>

1. Search and download the desired sequences in the GISAID database

Log in into GISAID and click on EpiCoV™'s Browse tab to access a table with all available SARS-CoV-2 sequences in the database. To bulk download sequences by accession ID, click on the Fulltext▲ button, paste your comma-separated list of accessions (see Files for an example) in the search box, and download the FASTA sequences using the download button (Figure 1). In this example, we will save the sequences as `gisaid_selection.fasta`.

We can also use the EpiCoV Browse portal to download a custom selection of genomes. On the header section of the table you will see multiple search fields and drop-down menus to filter sequences according to you desired criteria.

2. Remove white-spaces from the FASTA file

```
sed -i.bkp "s/ /_/g" gisaid_selection.fasta
```

To avoid potential issues when parsing the FASTA headers, we replace all white-spaces in the file with underscores using `sed`. We use the `-i` flag to find and replace the file in-place while keeping a backup of the original file.

3. Align the sequences using MAFFT

```
cat utr.fasta >> gisaid_selection.fasta  
mafft --thread -1 --nomemsave gisaid_selection.fasta > gisaid_aln.fasta
```

To remove potential sequencing errors in the error-prone 5' and 3' ends of the virus, we include the reference sequences for the 5' and 3' untranslated regions (UTR) of the SARS-CoV-2 genome (see example files). We do this so that we can later trim these regions from the final alignment. We concatenate these sequences to the FASTA file containing our genomes of interest, and align all sequences using MAFFT.

4. Manually trim UTRs in the MSA using Aliview

In Aliview, visually identify the UTR sequences and manually select the corresponding sites. Remove the selected sites using the Edit menu. Remove the now-empty reference sequences and save the trimmed MSA (Figure 2).

PROTOCOL 2: SETTING UP A DISCRETE TRAIT PHYLOGEOGRAPHIC RECONSTRUCTION IN BEAUTI

Performing Bayesian phylogeographic inference while accommodating individual travel history data constitutes an extension of traditional Bayesian DTA [Lemey et al., 2009] available in BEAST. BEAUti is an interactive graphical application enabling to design the analysis you want to perform in BEAST and generates an XML file that will serve

The screenshot shows the EpiCoV GISAID portal interface. At the top, the GISAID logo is visible, along with navigation links for Registered Users, EpiFlu™, EpiCoV™, and My profile. The user is logged in as Samuel L Hong. Below the navigation bar, there are tabs for Browse, Downloads, Upload, and My Unreleased. A search bar contains a list of EPI_ISL accession numbers. Below the search bar, a table lists virus sequences with columns for Virus name, Passage date, Accession ID, Collection date, Submission date, Length, Host, Location, and Originating lab. A 'Download' modal is open, showing options for downloading sequences in FASTA format, Patient status metadata, Sequencing technology metadata, or Acknowledgement (Supplemental table). The modal also includes a 'Back' button and a 'Download' button. At the bottom of the table, it says 'Total: 282 viruses'. A footer note mentions the GISAID EpiFlu™ Database Access Agreement.

Figure 1: EpiCoV GISAID portal. Here, we search for all 282 sequences used in [Lemey et al., 2020], and download them by selecting the checkbox on the left of the table and pressing the download button. We can also download metadata for the sequences and the corresponding GISAID acknowledgement table using the same approach.

as the input file for BEAST. This XML file contains all the required data, the models (and priors) that were selected in BEAUti, and the computational settings which will be used to run the MCMC algorithm that will collect samples of all relevant parameters (including the phylogenetic tree with annotated ancestral locations) from the posterior.

This protocol describes how to set up a DTA analysis with a generalized linear model (GLM) extension to simultaneously reconstruct spatiotemporal history and test the contribution of potential predictors of spatial spread using BEAST [Lemey et al., 2014]. Such an approach enables parameterizing each rate of among-location movement in the phylogeographic model as a log linear function of various potential predictors. In this example, we will use the MSA generated in Protocol 1 to set up a DTA+GLM phylogeographic reconstruction using a flight matrix and a distance matrix as covariates (but see [Lemey et al., 2020] for more detailed information).

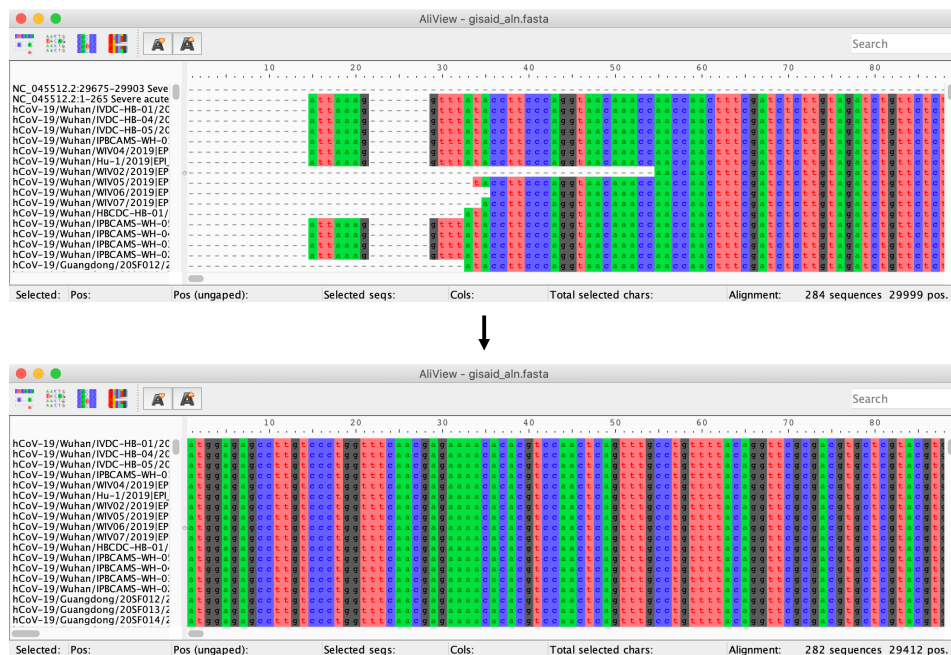


Figure 2: Alignment visualizations in Aliview before (top) and after (bottom) trimming the 5' and 3' UTR regions. Here we remove a total of 286 sites on the 5' side and 301 sites on the 3' side, bringing the alignment length to 29,412 sites.

Necessary Resources

Hardware

Standard workstation running Linux, MacOS, or Windows.

Software

Latest BEAUti version (v1.10.5)

Files

Multiple sequence alignment
Tab-delimited metadata file
Covariate matrices in CSV format

Example metadata and covariate files available at

<https://github.com/hongsamL/travHistProtocol/tree/main/files/Protocol2>

1. Import the MSA into BEAUti

Load the MSA into BEAUti by selecting Import Data from the File menu. You can also do this by dragging the FASTA file into the Partitions panel.

2. Specify the tip sampling dates

Select the Tips tab and check the “Use tip dates” box. By default all taxa will show as having a date of zero (i.e. all sequences were sampled at the same time in the present). To specify each sampling date, select “Import Dates” and load the metadata file. This metadata file contains a tab separated table mapping the FASTA header of each sequence in the alignment with the corresponding sampling date. When loading the file, select the “Parse calendar dates with variable precision” option. This allows for taxon dates to have different degrees of resolution (e.g. year-month-day vs. year-month). We estimate the sampling dates for those sequences without day-level resolution by selecting “Sampling uniformly from precision— in the “Tip date sampling” menu at the bottom of the table.

It is also possible to specify the sampling dates without using a metadata file. This is done by parsing the FASTA headers of each sequence. To do this, click on “Parse Dates”, and specify the rules for delimiting the date from all taxon labels. For an example of what this looks like, see [?] Figure 2a.

3. Specify the sampling location of each taxa as a discrete trait

Click on the Traits tab. To associate each taxon with a sampling location, click on “Import traits” and select the metadata file (Figure 3). This will create a new trait for each column on the metadata file. Delete all non-relevant traits by clicking on the “-” button at the bottom-left of the page (keep only the “location” trait for this example). Select the desired trait and click on “create partition from trait”. A new partition containing the trait data will be created under the Partitions tab.

It is also possible to add a discrete trait without using a metadata file. This is done by parsing the FASTA headers of each sequence. To do this, click on “Add Trait” to create a new trait with a corresponding data partition. Select all taxa and click on “Guess trait” values to parse the trait values from the taxon labels.

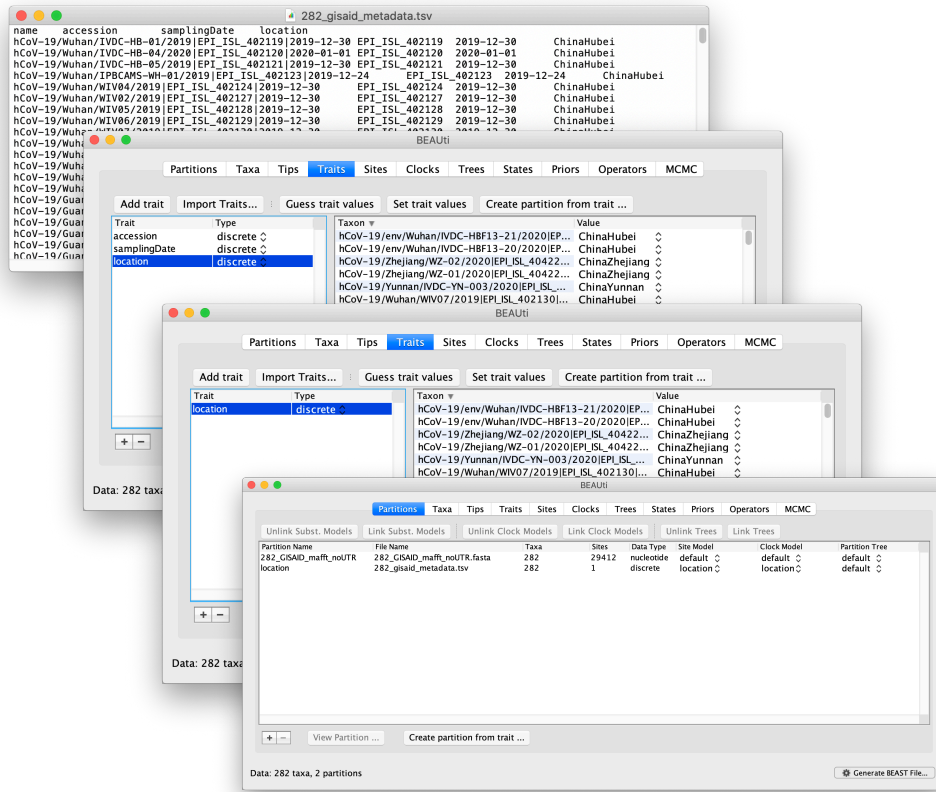


Figure 3: Adding the sampling locations as a trait. From back to front: *i*) tab-separated metadata file containing sequence names and sampling locations *ii*) columns in the metadata file other than “name” loaded as traits. *iii*) traits different from “location” removed from the analysis *iv*) partition corresponding to the “location” trait created.

4. Set up the nucleotide and trait substitution models

Click on the “Sites” tab. Following [Lemey et al., 2020], we will specify an HKY+ Γ nucleotide substitution model, and a GLM for the location trait. Load the covariates by clicking on “Setup GLM” and “Import Predictors” (Figure 4). In this example, we inform the rates of spread between locations using two predictors: a flight matrix containing air travel data between locations, and a distance matrix containing intra-continental distances. Check the “Log” and “Std” boxes to log-transform and standardize the GLM predictors.

In this context, the terms ‘predictor’ and ‘covariate’ are used interchangeably. By default, each predictor name will be the same as the name of the file it originates from. Non-pairwise covariates can also be setup as origin and destination predictors in this window. Predictor files are comma-separated files (CSV) formatted in two different ways depending on the nature of the predictor. For pairwise predictors, the CSV file is in the form of a square matrix with the different locations as column and row names alphabetically ordered. For origin-destination predictors, the CSV file is in the form of a two column table with location names alphabetically ordered and predictor values. You can also specify new predictor names by double clicking on each name.

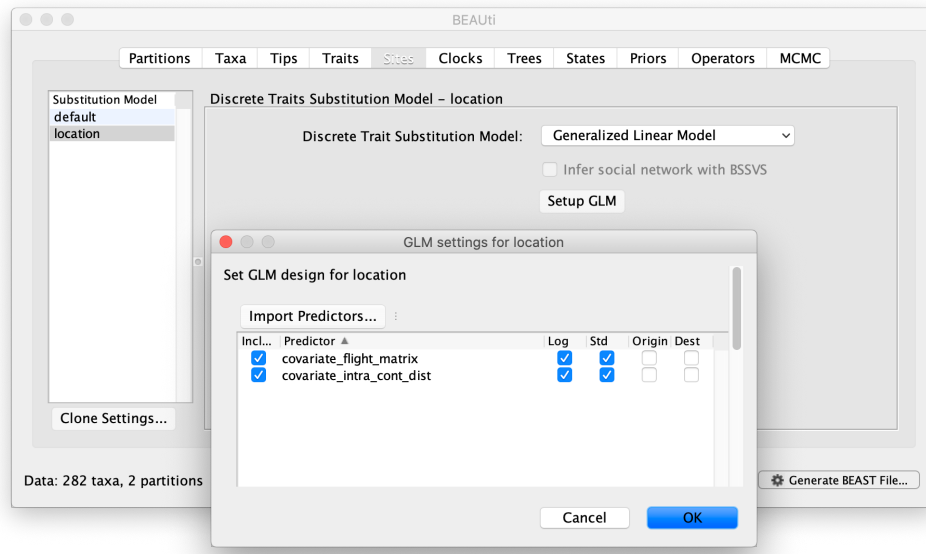


Figure 4: Setting up the predictors for the GLM-based spatial diffusion model. In this example, we load the air travel and intra-continental distance predictors, and log-transform and standardize them by checking the corresponding boxes.

5. Specify a clock model

Following the model specifications in [Lemey et al., 2020], click on the “Clocks” tab to specify a strict clock model for both the nucleotide and trait partitions.

BEAST analyses operate under the assumption of a molecular clock to estimate time-stamped phylogenies from the molecular sequence data and associated sampling times.

*For this to hold valid, there needs to be sufficient temporal signal in the data (see the section **Critical parameters and troubleshooting**). The presence of temporal signal in a data set can be assessed using TempEst [Rambaut et al., 2016] and more formally tested using Bayesian Evaluation of Temporal Signal (BETS) [Duchene et al., 2020]. For more information on the different molecular clock models available in BEAST: <https://beast.community/clocks>*

6. Specify a demographic model

Click on the “Trees” tab to specify the Tree prior. Following [Lemey et al., 2020], we specify an exponential growth coalescent model parameterized with a growth rate parameter. We refer to http://beast.community/tree_priors for a more detailed explanation on a subset of the available coalescent models in BEAST.

By default, BEAST will initialize the analysis with a randomly generated starting tree. It is also possible to start the analysis from a user-specified time-scaled phylogeny. This is usually done to reduce the burn-in time required for the tree topologies to converge. Specify a starting tree by clicking on “Import Data” under the “File” menu, to load a Nexus tree into BEAUti.

7. Set up the ancestral state reconstruction for the location trait

Click on the “States” tab and select the location partition. Check the “Reconstruct state change counts” and “Reconstruct complete change history on tree” boxes to save complete realizations of the spatial spread process on the output trees.

8. Specify the priors

Click on the “Priors” tab. Following [Lemey et al., 2020], we use default priors and only specify a Lognormal prior with `mu=0` and `sigma=10` for the effective population size, and a Laplace prior with `mean=0` and `scale=100` for the exponential growth rate parameter.

BEAST aims to offer sensible default priors when informative prior information is unavailable. A wide range of prior distributions is also available to customize each analysis as needed.

9. Set up the transition kernels

Click on the “Operators” tab. Identify the tip-date sampling transition kernels in the table. These have the description “Uniform sample from precision of age of this tip”. The parameters associated with these transition kernels tend to converge rapidly and have good mixing. Decrease the weight of these operators to 0.25, but leave the weights of the other transition kernels at their default values, to sample these parameters less frequently so that the analysis gets to spend more time on estimating other parameters of interest.

Transition kernels (called “operators” in BEAST) are used to propose new values for each parameter being estimated during the analysis. Different combinations of transition kernels can be used to fix certain parameters and customize the analysis as needed (e.g. estimating spatial spread on a fixed user-provided tree).

10. Set up the MCMC options and generate the BEAST XML file

Click on the “MCMC” tab. Set “Length of chain” to 200,000,000 states and “Log parameters every” to 100,000 states. This will thin the MCMC results so that only 2,000 samples are collected by the end of the run. Set your file name stem to generate the desired output file names (e.g. 282.GISAID_sarscov2), and click on “Generate BEAST File” to create the BEAST XML file for this analysis.

Thinning consists of storing only every n th sample from an MCMC analysis. This subsampling technique is done to decrease the autocorrelation in the samples collected and reduce the file size of the output. This is important since Bayesian phylogenetic analyses often require very long chains, and storing every single state would be prohibitive for file storage.

PROTOCOL 3: PHYLOGEOGRAPHIC RECONSTRUCTION INCORPORATING TRAVEL HISTORY INFORMATION

Phylogeographic reconstruction using DTA has been shown to be sensitive to spatiotemporal sampling bias. The ancestral reconstruction of locations will depend on the availability of samples from each location. In practice, this means that over/undersampling of sequences from a given location can greatly impact the ancestral locations being reconstructed. One way to mitigate sampling bias is through the incorporation of available travel history information from infected individuals. Travel history data can be used to correct for gaps in sampling by allowing for ancestral nodes to be in a given location even when molecular sequence data are not available.

This protocol explains how to augment a DTA analysis generated in BEAUti by incorporating individual travel history data. In this example, we will use the XML file generated in Protocol 2 and modify it to include the available travel history data (See [Lemey et al., 2020] for more detailed information). Importantly, BEAST requires its accompanying high-performance computational library for statistical phylogenetics – known as BEAGLE [?] – to be installed in order to optimise computational performance on a variety of hardware resources.

Necessary Resources

Hardware

Standard workstation running Linux, MacOS, or Windows.
CUDA-compatible Graphics Processing Unit (optional but recommended)

Software

Python v3.6+ with packages numpy and lxml
BEAGLE v3+
Latest BEAST jar file (v1.10.5) (provided with this protocol)
add_travel_history.py Python script (provided with this protocol)

Files

XML file for a DTA+GLM analysis set up in BEAUti
Travel history metadata CSV file

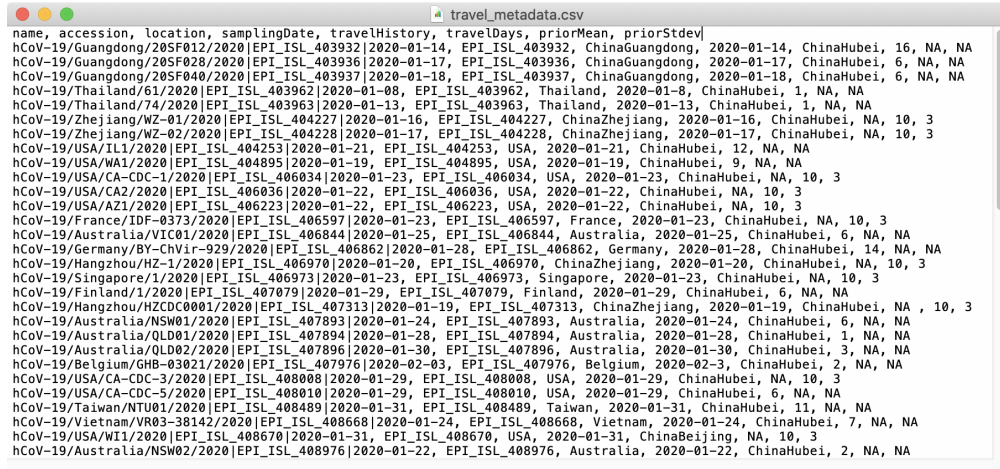
Augmented covariate data files

Example files are provided in the repository [github](#). The example in this protocol assumes an XML file for a DTA+GLM phylogeographic reconstruction with new travel history locations not present in the original BEAST XML file generated by BEAUti. Covariate files augmented to include the new locations are thus required to accommodate for the increase in state space.

1. Update the BEAST XML file to incorporate travel history data

```
python add_travel_history.py --xml 282_GISAID_sarscov2.xml
--hist travel_metadata.csv
--covariate augmented_flight_matrix.csv
--covariate augmented_intra_cont_dist.csv
--out 282_GISAID_sarscov2_travelHist.xml
```

Although this example uses a DTA+GLM analysis, the `add_travel_history.py` script also works for symmetric and asymmetric DTA analyses. For such cases, run the same command without the `--covariate` flags. This script also requires for the travel history metadata to follow a specific format (Figure 5). The metadata file is in CSV format and must contain the following columns: “name” (taxon name), “travelHistory” (travel location), “travelDays” (date of travel as days before to sampling date), “priorMean” and “priorStdev” (prior specifications for the mean and standard deviation of a normal prior on travel dates when exact data is unavailable).



name	accession	location	samplingDate	travelHistory	travelDays	priorMean	priorStdev
hCoV-19/Guangdong/205F012/2020	EPI_ISL_403932	2020-01-14	EPI_ISL_403932	ChinaGuangdong	2020-01-14	ChinaHubei	16, NA, NA
hCoV-19/Guangdong/205F028/2020	EPI_ISL_403936	2020-01-17	EPI_ISL_403936	ChinaGuangdong	2020-01-17	ChinaHubei	6, NA, NA
hCoV-19/Guangdong/205F040/2020	EPI_ISL_403937	2020-01-18	EPI_ISL_403937	ChinaGuangdong	2020-01-18	ChinaHubei	6, NA, NA
hCoV-19/Thailand/61/2020	EPI_ISL_403962	2020-01-08	EPI_ISL_403962	Thailand	2020-01-08	ChinaHubei	1, NA, NA
hCoV-19/Thailand/74/2020	EPI_ISL_403963	2020-01-13	EPI_ISL_403963	Thailand	2020-01-13	ChinaHubei	1, NA, NA
hCoV-19/Zhejiang/WZ-01/2020	EPI_ISL_404227	2020-01-16	EPI_ISL_404227	ChinaZhejiang	2020-01-16	ChinaHubei	NA, 10, 3
hCoV-19/Zhejiang/WZ-02/2020	EPI_ISL_404228	2020-01-17	EPI_ISL_404228	ChinaZhejiang	2020-01-17	ChinaHubei	NA, 10, 3
hCoV-19/USA/IL1/2020	EPI_ISL_404253	2020-01-21	EPI_ISL_404253	USA	2020-01-21	ChinaHubei	12, NA, NA
hCoV-19/USA/WA1/2020	EPI_ISL_404895	2020-01-19	EPI_ISL_404895	USA	2020-01-19	ChinaHubei	9, NA, NA
hCoV-19/USA/CA-CDC-1/2020	EPI_ISL_406034	2020-01-23	EPI_ISL_406034	USA	2020-01-23	ChinaHubei	NA, 10, 3
hCoV-19/USA/CA2/2020	EPI_ISL_406036	2020-01-22	EPI_ISL_406036	USA	2020-01-22	ChinaHubei	NA, 10, 3
hCoV-19/USA/AZ1/2020	EPI_ISL_406223	2020-01-22	EPI_ISL_406223	USA	2020-01-22	ChinaHubei	NA, 10, 3
hCoV-19/France/IDF-0373/2020	EPI_ISL_406597	2020-01-23	EPI_ISL_406597	France	2020-01-23	ChinaHubei	NA, 10, 3
hCoV-19/Australia/VIC01/2020	EPI_ISL_406844	2020-01-25	EPI_ISL_406844	Australia	2020-01-25	ChinaHubei	6, NA, NA
hCoV-19/Germany/BY-ChVir-929/2020	EPI_ISL_406862	2020-01-28	EPI_ISL_406862	Germany	2020-01-28	ChinaHubei	14, NA, NA
hCoV-19/Hangzhou/HZ-1/2020	EPI_ISL_406970	2020-01-20	EPI_ISL_406970	ChinaZhejiang	2020-01-20	ChinaHubei	NA, 10, 3
hCoV-19/Singapore/1/2020	EPI_ISL_406973	2020-01-23	EPI_ISL_406973	Singapore	2020-01-23	ChinaHubei	NA, 10, 3
hCoV-19/Finland/1/2020	EPI_ISL_407079	2020-01-29	EPI_ISL_407079	Finland	2020-01-29	ChinaHubei	6, NA, NA
hCoV-19/Hangzhou/HZCD0001/2020	EPI_ISL_407313	2020-01-19	EPI_ISL_407313	ChinaZhejiang	2020-01-19	ChinaHubei	NA, 10, 3
hCoV-19/Australia/NSW01/2020	EPI_ISL_407893	2020-01-24	EPI_ISL_407893	Australia	2020-01-24	ChinaHubei	6, NA, NA
hCoV-19/Australia/QLD01/2020	EPI_ISL_407894	2020-01-28	EPI_ISL_407894	Australia	2020-01-28	ChinaHubei	1, NA, NA
hCoV-19/Australia/QLD02/2020	EPI_ISL_407896	2020-01-30	EPI_ISL_407896	Australia	2020-01-30	ChinaHubei	3, NA, NA
hCoV-19/Belgium/GBH-03021/2020	EPI_ISL_407976	2020-02-03	EPI_ISL_407976	Belgium	2020-02-03	ChinaHubei	2, NA, NA
hCoV-19/USA/CA-CDC-3/2020	EPI_ISL_408008	2020-01-29	EPI_ISL_408008	USA	2020-01-29	ChinaHubei	NA, 10, 3
hCoV-19/USA/CA-CDC-5/2020	EPI_ISL_408010	2020-01-29	EPI_ISL_408010	USA	2020-01-29	ChinaHubei	6, NA, NA
hCoV-19/Taiwan/NTU01/2020	EPI_ISL_408489	2020-01-31	EPI_ISL_408489	Taiwan	2020-01-31	ChinaHubei	11, NA, NA
hCoV-19/Vietnam/VR03-38142/2020	EPI_ISL_408668	2020-01-24	EPI_ISL_408668	Vietnam	2020-01-24	ChinaHubei	7, NA, NA
hCoV-19/USA/WI1/2020	EPI_ISL_408670	2020-01-31	EPI_ISL_408670	USA	2020-01-31	ChinaBeijing	NA, 10, 3
hCoV-19/Australia/NSW02/2020	EPI_ISL_408976	2020-01-22	EPI_ISL_408976	Australia	2020-01-22	ChinaHubei	2, NA, NA

Figure 5: Travel history metadata. Metadata file must contain the columns “name”, “travelHistory”, “travelDays”, “priorMean” and “priorStdev”. Other columns can be included but will not be parsed to update the XML. For sequences where either exact travel dates are not available, we set the “travelDays” column to NA, such that the MCMC samples from a range of possible travel dates with a Gaussian prior distribution of mean “priorMean” (in units of days) and standard deviation “priorStdev”. For sequences where exact travel dates are available, we set the prior columns to NA.

2. Run the updated XML file using BEAST

```
java -cp beast.jar dr.app.beast.BeastMain -seed 2020
```

```

-beagle_gpu
-save_every 1000000
-save_state travelHist.checkpoint
282_GISAID_sarscov2_travelHist.xml

```

Here, we execute BEAST on the command-line using the latest BEAST jar file. We specify a starting seed with the `-seed` flag, and use the `-beagle.cuda` flag to accelerate the likelihood computations using a GPU (only applicable if you have a powerful GPU with sufficient double precision – or FP64 – compute performance available). This option is recommended when available, as using a GPU reduces runtime by accelerating likelihood computations when performing phylogeographical analyses on large datasets. We also take advantage of the BEAST checkpointing functionality [?] to save a snapshot of the MCMC run into the `travelHist.checkpoint` file every 1,000,000 states. This allows us to resume the analysis from the checkpoint in case the run becomes interrupted, or more iterations of the chain are required.

PROTOCOL 4: VISUALIZING INDIVIDUAL-SPECIFIC SPATIAL TRAJECTORIES

Necessary Resources

Hardware

Standard workstation running Linux, MacOS, or Windows.

Software

Latest BEAST jar file (v1.10.5) (provided with this protocol)
 R with package MarkovJumpR

Files

Trees output from a BEAST travel history DTA+GLM analysis

1. Extract all Markov jump histories for an isolate of interest

```

java -cp beast.jar dr.app.tools.TaxaMarkovJumpHistoryAnalyzer
      -taxaToProcess "hCoV-19/Brazil/SP-02/2020|EPI_ISL_413016|2020-02-28"
      -stateAnnotation location
      -burnin 100
      -msrd 2020.1748633879781
      282_GISAID_GLM.location.history.trees EPI_ISL_412975_MJhist.csv

```

The BEAST jar file packages a number of standalone applications that can be accessed by using the `-cp` flag when calling Java from the command line. Here, we use the `TaxaMarkovJumpHistoryAnalyzer` application to extract all Markov jump histories for isolate `EPI_ISL_413016`, into a CSV file. This application takes a trees file with complete state change history as an input (which is being generated by running the XML constructed in the protocol), and outputs all spatial trajectories for a selection of taxon/taxa. We specify the desired taxon labels through the `-taxaToProcess` flag, and specify the annotation name of the discrete trait that was reconstructed using `-stateAnnotation`. We can also remove a number of trees corresponding to the burn-in using the `-burnin` flag, and scale the output results to

reflect chronological time instead of node heights using the `-msrd` flag by specifying the most recent sampling date. An example output file can be found in <https://github.com/hongsamL/travHistProtocol/tree/main/files/Protocol1>

2. Load spatial trajectories into R

```
library(MarkovJumpR)
spatial_paths <- loadPaths(fileName = "EPI_ISL_413016_MJhist.csv")
```

3. Inspect spatial trajectories reconstructed

```
spatial_paths$minTime
```

yields the earliest time along a spatial path across all trees

```
2019.892
```

To look at the frequency of locations visited across all spatial paths we type

```
loc_freq <- table(spatial_paths$paths$location)
loc_freq[order(loc_freq,decreasing = T)]
```

which yields a frequency table of locations in descending order

Italy	Brazil	ChinaHubei	Switzerland	Finland
444	437	436	85	23
ChinaBeijing	UK	Australia	Germany	USA
9	8	7	6	6
France	Singapore	Spain	ChinaHongKong	Japan
5	5	5	4	3
Netherlands	Sweden	Vietnam	ChinaGuangdong	ChinaShandong
3	3	3	2	2
NewZealand	Thailand	Belgium	Cambodia	Canada
2	2	1	1	1
ChinaChongqing	ChinaFujian	ChinaYunnan	India	Iran
1	1	1	1	1
Mexico	Nepal	Portugal	SouthKorea	Taiwan
1	1	1	1	1

In this example, we see that Italy, Brazil, ChinaHubei and Switzerland appear most commonly across the spatial paths.

4. Set up plot colors

We here specify four colors of choice corresponding to the four locations of interest that make up the spatial trajectory of isolate EPI_ISL_413016.

```
locations <- c("ChinaHubei","Italy","Brazil","Switzerland")
locationColors <-c("#E3272F","#31B186","#931ECF","#C695BD")
locationMap <- data.frame(location = locations,
                           position = c(1, 2, 3, 4))
locationMap$color <- sapply(locationColors,as.character)
```

5. Set up plot labels

```
dateLabels <- c("01-Dec-19", "15-Dec-19", "01-Jan-20", "15-Jan-20",
               "01-Feb-20", "15-Feb-20", "01-Mar-20")
```

6. Plot path spatial trajectories

```
plotPaths(travelHistPaths$paths, locationMap = locationMap,
          yJitterSd = 0.1, alpha = 0.1, minTime = spatial_paths$minTime,
          addLocationLine = TRUE,
          xAt = decimal_date(dmy(dateLabels)),
          xLabels = dateLabels,
          mustDisplayAllLocations = TRUE)
```

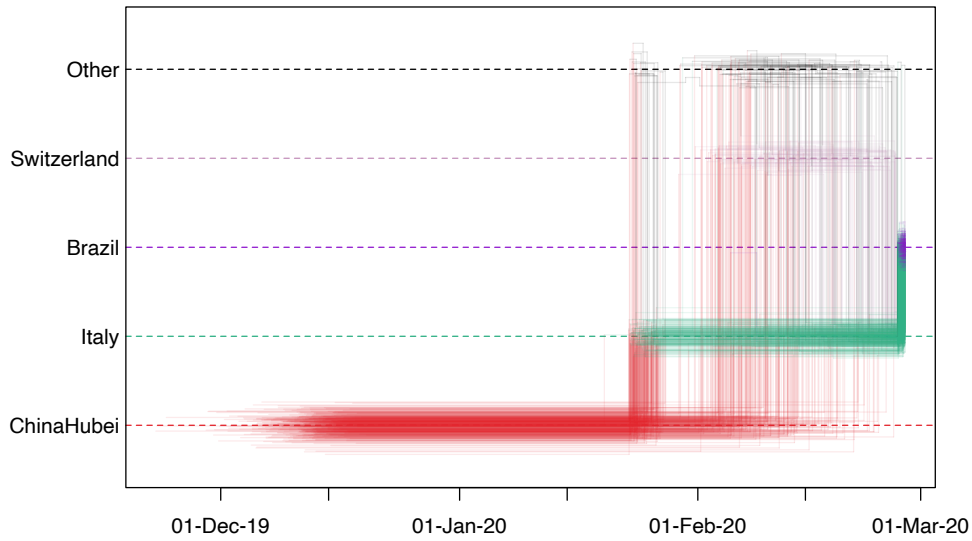


Figure 6: Spatial trajectory plot of isolate EPI_ISL_413016. The plot depicts the posterior spatiotemporal ancestral transition history for a single isolate. Each line represents a single Markov jump history in the posterior distribution. The time spent in each location is denoted in the horizontal dimension, and transitions between two locations are depicted with vertical lines. The relative density of lines reflects the posterior uncertainty in location state and transition time between states. Here we see that the most supported ancestral history for isolate EPI_ISL_413016 is that of an origin in ChinaHubei, with a jump into Italy late January, and an introduction into Brasil close to March 1st.

GUIDELINES FOR UNDERSTANDING RESULTS

The BEAST software package offers a flexible approach for combining demographic, molecular clock, nucleotide and trait evolution models to infer time-scaled trait-annotated phylogenies. BEAST employs Bayesian inference through MCMC to sample trees and all of the model parameters from the joint density (often simply called the posterior). Protocols 2 and 3 show how to set up the different models and run the corresponding BEAST analyses to collect samples from the posterior. The phylogenetic trees that are sampled from the posterior are stored in a `.trees` file and samples of the model parameters are stored in a `.log` file. Here we present some of the standard applications that are commonly used to interpret the output that BEAST generates.

Assessing convergence

The MCMC sampling strategy is to construct a Markov chain that (eventually) converges to a stationary distribution, which is the posterior in the case of Bayesian inference. Given enough time we know that the chain will converge, but it may take considerable time for this to happen. We can visually assess the convergence of a BEAST run by inspecting the sampled parameter values across an MCMC analysis. To do so, we can load a .log file into Tracer (<https://beast.community/tracer>), and visually inspect the trace plot, which shows a time series of the parameter values sampled throughout the analysis. A detailed guide on how to use Tracer can be found at https://beast.community/tracer_convergence.html

Parameter estimates and effective sample size

Loading a .log file in Tracer will also show the estimates and effective sample size (ESS) values for each model parameter. Note that ESS values are only defined for continuous parameters that are being estimated. A characteristic of inference through MCMC is that the samples collected tend to be correlated. This in turn poses a challenge, since having a large number of samples does not guarantee a considerable reduction of the uncertainty in our posterior estimates. A way to control for this is to look at the ESS values of an estimate. The ESS of a parameter sampled from an MCMC method is the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to. ESS values will increase as more and more samples are collected by the sampler. Larger ESS values will result in more precise posterior estimates but require higher computational resources. Tracer will automatically calculate the ESS for all parameters in a log file, and flag values above 100 (acceptable) and 200 (ideal).

Summarizing trees and phylogeographic estimates

Individually inspecting every tree from the posterior obviously constitutes an impractical way to interpret the MCMC results. We are thus required to summarize the distribution of trees sampled as a point estimate with associated uncertainties. The TreeAnnotator application in BEAST enables creating a maximum clade credibility (MCC) tree to summarize the sampled trees for this purpose. For every tree in the distribution, a posterior clade probability for each node (i.e. the support for a node) is calculated by computing the frequency of the clustering that is defined by the node in question. The MCC tree is then defined as the tree that maximizes the product of the posterior clade probabilities across the tree. Instructions on how to use TreeAnnotator can be found at beast.community/second_tutorial.

In some data sets, the posterior support for all nodes in a tree is such that many topologies do not end up represented in the MCC tree. When that is the case, a point estimate of a tree is unable to capture the diverse phylogeographic histories compatible with the data. Protocol 4 allows us to inspect individual spatial trajectories by summarizing across all possible phylogenetic ancestries in the posterior. Each spatial trajectory in the posterior is represented by a stepwise curve, where vertical lines represent transitions between two locations, and horizontal lines time intervals where the pathogen remains in the same location. The relative density of lines reflects the posterior uncertainty in spatiotemporal ancestry.

COMMENTARY

Background Information

Models for phylogeographic inference can be broadly categorized into two classes or

types, depending on the assumptions used to model the spatial spread of a pathogen. Structured coalescent approaches model

movement in terms of a migration matrix, and relate migration rates with a location’s population size, which involves population size dynamics having to be estimated for each location [De Maio et al., 2015, Müller et al., 2018]. On the other hand, diffusion approaches consider sampling locations as observed traits independent from the tree-generative process, and model movement across space with a continuous-time Markov chain (CTMC) [Lemey et al., 2009, Lemey et al., 2010]. While structured coalescent approaches are theoretically a more robust, the lack of computationally efficient implementations (especially for larger datasets) has made diffusion methods more widely adopted. Currently, BEAST v1.10.5 focuses on providing inference using CTMC models in the case of discrete location data.

Diffusion models for phylogeographic inference can be further categorized depending on the spatial resolution being considered. Discrete locations are modeled using discrete trait analysis (DTA), with transition rates between locations in the form of a CTMC matrix analogous to those used for nucleotide substitution [Lemey et al., 2009]. In contrast, continuous locations are modeled using Brownian diffusion based random walk models [Lemey et al., 2010]. Much of the genomic data collected has a spatial resolution coarser than latitude and longitude, which makes DTA the only practical approach to study the spread of an epidemic.

Under the DTA formulation, movement between K discrete locations is parameterized in terms of a $K \times K$ infinitesimal rate matrix Λ , where Λ_{ij} is the instantaneous movement rate from location i to j . This model has been extended to allow for symmetrical and asymmetrical transition rates, and uses Bayesian Stochastic Search Variable Selection (BSSVS) to limit the number of rates to only those that adequately explain the phylogenetic diffusion process. An alternative formulation of this model parameterizes the rate matrix using a generalized linear model [Lemey et al., 2014]. Here, the transition rates are defined as a linear combination of P of potential ex-

planatory predictors $(x_{ij1}, \dots, x_{ijP})$, with corresponding coefficients $(\beta_1, \dots, \beta_P)$ and indicator variables $(\delta_1, \dots, \delta_P)$ such that $\text{Log}(\Lambda_{ij}) = \sum_{p=1}^P \beta_p \delta_p x_{ijp}$. This model specification allows us to use BSSVS to explore the space of 2^P predictor combinations and obtain a posterior probability on the indicator variables to determine the support for inclusion of each predictor in the model.

Incorporating individual travel history data can also be done to any of the other DTA variants (symmetric or asymmetric rates [Lemey et al., 2009]) by augmenting the available dataset to include ancestral nodes associated with a known state but not with a known sequence [Lemey et al., 2020]. These “ghost samples” enable exploiting individual travel cases for which molecular sequence data was not obtained. Ambiguous ancestral locations can also be allowed by integrating over the all possible locations with equal or user specified weights [Scotch et al., 2019]. An example would be the case where an individual traveled to multiple locations prior to being sampled. Given that the individual could have gotten infected in any of the visited countries, we can integrate this uncertainty by marginalizing over all potential locations for the unsampled ancestor when performing the phylogeographic inference.

CRITICAL PARAMETERS AND TROUBLESHOOTING

A critical assumption for any BEAST analysis is that the data set under consideration constitutes a samples from a measurably evolving population (MEP). MEPs [Drummond et al., 2003, Biek et al., 2015] refer to time-stamped sequence data where a sufficient amount of molecular evolution has occurred throughout the sampling period to establish a statistical relationship between genetic divergence and time. A data set conforming to this criteria is said to contain sufficient temporal signal. A lack of temporal signal will ultimately result in flawed analyses with unreliable divergence time estimates. Popular ways to assess temporal signal include doing a root-to-tip re-

gression of genetic divergence and sampling times on a maximum likelihood tree [Rambaut et al., 2016] and permutating tip date labels through date-randomization [Ramsden et al., 2009]. Recently, a formal way to assess temporal signal has been developed under a Bayesian framework using model comparison through Bayes factors [Duchene et al., 2020]. In cases where the temporal signal is deemed insufficiently strong, one can resort to adding more data to increase temporal coverage or using prior knowledge to inform the molecular clock rate or the time to the most recent common ancestor.

Another commonly encountered issue is that of low ESS values for parameters relevant to the analysis. At the end of a BEAST analyses, some parameters may have much higher ESS values associated to them compared to others. One way to increase the ESS values of a parameter is to decrease the weight of other parameters that have been able to attain sufficiently high (or very high) ESS values to increase the sampling frequency of the (problematic) parameter of interest (e.g. Protocol 2, step 9). Other ways to obtain more samples – and as a result increase the ESS value – include increasing the MCMC chain length and combining the output of multiple independent BEAST analyses, i.e. analysing the same XML file using BEAST but with different starting seeds.

SUGGESTIONS FOR FURTHER ANALYSIS

This section is optional but I’m open for suggestions.

References

- [Biek et al., 2015] Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.*, 30(6):306–313.
- [De Maio et al., 2015] De Maio, N., Wu, C.-H., O’Reilly, K. M., and Wilson, D. (2015). New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet.*, 11(8):e1005421.
- [Drummond et al., 2003] Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends Ecol. Evol.*, 18(9):481–488.
- [Drummond and Rambaut, 2007] Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214.
- [Drummond et al., 2012] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, 29(8):1969–1973.
- [Duchene et al., 2020] Duchene, S., Lemey, P., Stadler, T., Ho, S. Y. W., Duchene, D. A., Dhanasekaran, V., and Baele, G. (2020). Bayesian evaluation of temporal signal in measurably evolving populations. *Mol. Biol. Evol.*
- [Dudas et al., 2017] Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D’Ambrozio, J., Dellicour, S., Di Caro, A., Diclario, J. W., Duraffour, S., Elmore, M. J., Fakoli, L. S., Faye, O., Gilbert, M. L., Gevaio, S. M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D. S., Haagmans, B. L., Hiscox, J. A., Jah, U., Kugelman, J. R., Liu, D., Lu, J., Malboeuf, C. M., Mate, S., Matthews, D. A., Matranga, C. B., Meredith, L. W., Qu, J., Quick, J., Pas, S. D., Phan, M. V. T., Pollakis, G., Reusken, C. B., Sanchez-Lockhart, M., Schaffner, S. F., Schieffelin, J. S., Sealfon, R. S., Simon-Loriere, E., Smits, S. L., Stoecker, K., Thorne, L., Tobin, E. A., Vandi, M. A., Watson, S. J., West, K., Whitmer, S., Wiley, M. R., Winnicki, S. M., Wohl, S., Wölfel, R., Yozwiak, N. L., Andersen, K. G., Blyden, S. O., Bolay, F., Carroll, M. W., Dahn, B., Diallo, B.,

- Formenty, P., Fraser, C., Gao, G. F., Garry, R. F., Goodfellow, I., Günther, S., Happi, C. T., Holmes, E. C., Kargbo, B., Keita, S., Kellam, P., Koopmans, M. P. G., Kuhn, J. H., Loman, N. J., Magassouba, N., Naidoo, D., Nichol, S. T., Nyenswah, T., Palacios, G., Pybus, O. G., Sabeti, P. C., Sall, A., Ströher, U., Wurie, I., Suchard, M. A., Lemey, P., and Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315.
- [Elbe and Buckland-Merrett, 2017] Elbe, S. and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID’s innovative contribution to global health. *Glob Chall*, 1(1):33–46.
- [Katoh and Standley, 2013] Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780.
- [Larsson, 2014] Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278.
- [Lemey et al., 2020] Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., O’Toole, Á., McCrone, J. T., Andersen, K. G., Worobey, M., Nelson, M. I., Rambaut, A., and Suchard, M. A. (2020). Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.*, 11(1):5110.
- [Lemey et al., 2014] Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., and Suchard, M. A. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.*, 10(2):e1003932.
- [Lemey et al., 2009] Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, 5(9):e1000520.
- [Lemey et al., 2010] Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.*, 27(8):1877–1885.
- [Müller et al., 2018] Müller, N. F., Rasmussen, D., and Stadler, T. (2018). MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848.
- [Rambaut et al., 2016] Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*, 2(1):vew007.
- [Ramsden et al., 2009] Ramsden, C., Holmes, E. C., and Charleston, M. A. (2009). Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.*, 26(1):143–153.
- [Scotch et al., 2019] Scotch, M., Tahsin, T., Weissenbacher, D., O’Connor, K., Magge, A., Viente, M., Suchard, M. A., and Gonzalez-Hernandez, G. (2019). Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography. *Virus Evol*, 5(1):vey043.
- [Suchard et al., 2018] Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*, 4(1):vey016.
- [Worobey et al., 2016] Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., Koblin, B. A., Heneine, W., Lemey, P., and Jaffe, H. W. (2016). 1970s and ‘patient 0’ HIV-1 genomes illuminate early HIV/AIDS history in north america. *Nature*, 539(7627):98–101.