Phylodynamics of Infectious Diseases (Part I)

Mandev Gill

Department of Statistics and Institute of Bioinformatics

University of Georgia

The term "phylodynamics" was introduced by Grenfell et al. in 2004:

Unifying the Epidemiological and Evolutionary Dynamics of Pathogens

Bryan T. Grenfell,[1]* Oliver G. Pybus,[2] Julia R. Gog,[1] James L. N. Wood,[3] Janet M. Daly,[3] Jenny A. Mumford,[3] Edward C. Holmes[2]

A key priority for infectious disease research is to clarify how pathogen genetic variation, modulated by host immunity, transmission bottlenecks, and epidemic dynamics, determines the wide variety of pathogen phylogenies observed at scales that range from individual host to population. We call the melding of immunodynamics, epidemiology, and evolutionary biology required to achieve this synthesis pathogen "phylodynamics." We introduce a phylodynamic framework for the dissection of dynamic forces that determine the diversity of epidemiological and phylogenetic patterns observed in RNA viruses of vertebrates. A central pillar of this model is the Evolutionary Infectivity Profile, which captures the relationship between immune selection and pathogen transmission.

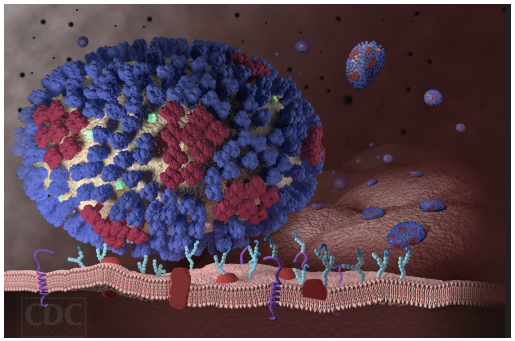
The population dynamics of many host-pathogen interactions are well character-

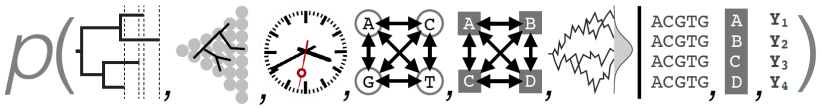
Observed Phylodynamic Patterns

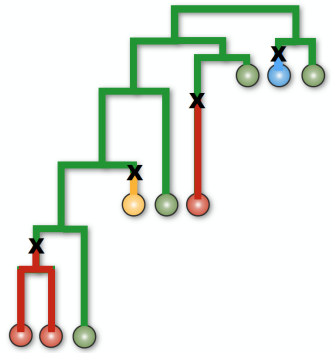
A major determinant of epidemic (and

However, the powerful, strain-transcending immunity driving these oscillations is not reflected directly in the measles virus phylogeny (Fig. 1D), because an immune response that is equally potent against all strains will not generate selection. Therefore, many strains coexist, with relative frequencies determined predominantly by nonselective epidemiological processes. This does not exclude the sporadic occurrence of selection, immunologically mediated or otherwise. Rather, the measles phylogeny indicates that selection is not operating sufficiently consistently to leave its imprint. Instead, the phylogeny is determined by global spatio-

What is Phylodynamics?







More generally, we can think of "phylodynamics" as the study of the interaction of evolutionary, epidemiological, ecological and immunological processes through the use of a phylogenetic inference framework.

This is possible for measurably evolving populations (such as rapidly evolving pathogens) for which these processes occur on approximately the same timescale.

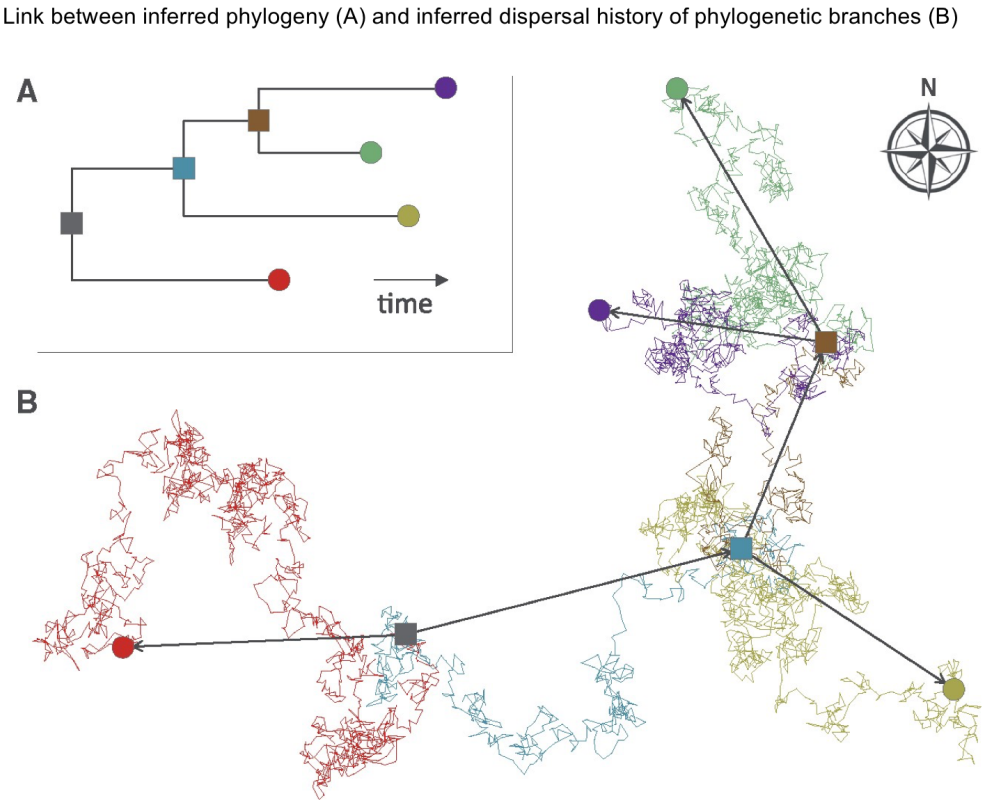
With serially sampled data, we can infer time-measured phylogenetic trees.

Pathogen Phylodynamics

Phylodynamic inference methods can reconstruct unobserved events and processes and provide valuable insights that can help answer a number of important questions in infectious disease research:

- How is a pathogen evolving and spreading?
- How far has it spread? When did it spread to different areas?
- What are its circulation patterns?
- What are the factors that are related to its evolution and spread?
- How have intervention and containment measures impacted its dispersal dynamics?
- What are the values of key epidemiological parameters?

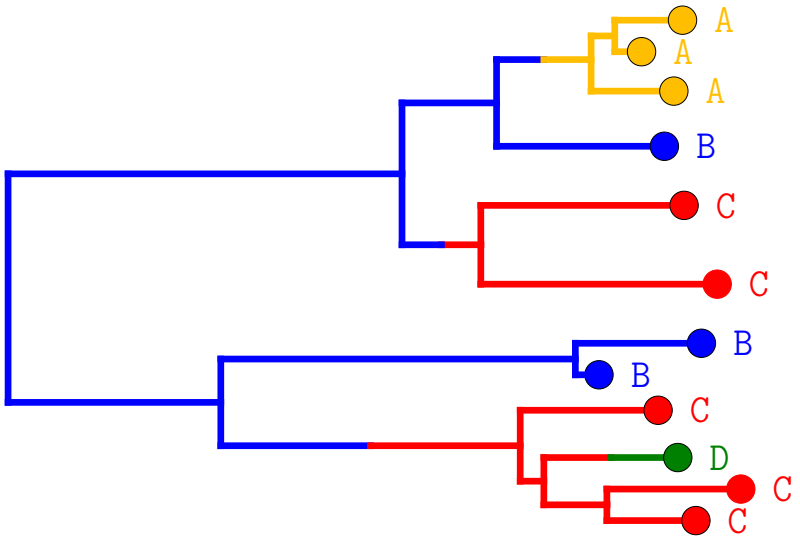
Phylogeographic Inference



(Dellicour et al., 2021)

- **Goal:** phylogeographic inference methods aim to connect the spatial and evolutionary history of a population
- By reconstructing the spatial and temporal spread of pathogens in an evolutionary context, we can better understand the origin, spread, and dynamics of infectious diseases

Discrete Trait Analysis



$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & \cdot & \lambda_{23} & \lambda_{24} \\ \lambda_{31} & \lambda_{32} & \cdot & \lambda_{34} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \cdot \end{bmatrix} \end{array}$$

- We can model the evolution of discrete "traits" on phylogenetic trees using a continuous time Markov chain (CTMC) model (Pagel, 1994; Lewis, 2001; Sanmartin et al., 2008; Lemey et al., 2009)
- Analogous to standard substitution models for molecular characters
- For phylogeography, we think of the "traits" as being discrete sampling locations (e.g., country, state, county, city).
- However, other scientific questions can be explored by taking the "traits" to be, for example, phenotypic traits, or host species, or body compartments within a host.

Discrete Trait Analysis

https://beast.community/

Lemey et al. (2009) implemented a Bayesian discrete trait analysis framework for phylogeography in BEAST X (previously called BEAST 1)

$$P(\tau, \boldsymbol{\phi}, \boldsymbol{\Lambda} | \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X} | \tau, \boldsymbol{\phi}) P(\mathbf{Y} | \tau, \boldsymbol{\Lambda}) P(\tau) P(\boldsymbol{\phi}) P(\boldsymbol{\Lambda})$$

$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ - aligned molecular sequence data

$\boldsymbol{\phi}$ - parameters that characterize CTMC that generates $\mathbf{X}$

$\mathbf{Y} = (Y_1, \ldots, Y_N)$ - sampling locations of sequences

$\boldsymbol{\Lambda}$ - infinitesimal rate matrix for CTMC that generates $\mathbf{Y}$

τ - phylogenetic tree

The transition probabilities between locations, $\mathbf{P}(t) = \exp[\mathbf{\Lambda} t]$, are characterized by the infinitesimal rate matrix $\mathbf{\Lambda}$:

$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & \lambda_{12} & \lambda_{13} & \lambda_{14} \\ \lambda_{21} & \cdot & \lambda_{23} & \lambda_{24} \\ \lambda_{31} & \lambda_{32} & \cdot & \lambda_{34} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \cdot \end{bmatrix} \end{array},$$

where we write "$\cdot$" in place of whatever it takes for a row to sum to 0. We can express $\mathbf{\Lambda}$ as $\mathbf{\Lambda} = \mathbf{S\Pi}$, where

$$\mathbf{S} = \begin{bmatrix} \cdot & s_{12} & s_{13} & s_{14} \\ s_{12} & \cdot & s_{23} & s_{24} \\ s_{13} & s_{23} & \cdot & s_{34} \\ s_{14} & s_{24} & s_{34} & \cdot \end{bmatrix}, \mathbf{\Pi} = \begin{bmatrix} \pi_A & & & \\ & \pi_B & & \\ & & \pi_C & \\ & & & \pi_D \end{bmatrix}$$

Bayesian Stochastic Search Variable Selection

In substitution models, most of the possible transitions have non-negligible probability of occurring over an evolutionary history.

By contrast, we expect most possible transitions between geographic locations to be rare (compare how each taxon corresponds to one observed location vs. hundreds or thousands of alignment sites).

Problem: Fitting overparameterized models to limited data can lead to poor estimates with high variance (for $\mathbf{\Lambda}$, and for inferred unobserved ancestral locations based on model)

Solution: Use Bayesian stochastic search variable selection to let observed data determine which of the possible transition rates should be nonzero.

$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & \overset{0}{\nearrow}\lambda_{12} & \lambda_{13} & \lambda_{14} \\ \overset{0}{\nearrow}\lambda_{21} & \cdot & \lambda_{23} & \overset{0}{\nearrow}\lambda_{24} \\ \lambda_{31} & \lambda_{32} & \cdot & \lambda_{34} \\ \lambda_{41} & \overset{0}{\nearrow}\lambda_{42} & \lambda_{43} & \cdot \end{bmatrix} \end{array}$$

Bayesian Stochastic Search Variable Selection

For each pair of locations j and k, introduce an auxiliary binary indicator variable δ_{jk}

- If $\delta_{jk} = 1$, there is a non-negligible chance of transitions between the locations, and λ_{jk} and λ_{kj} are included in the model
- If $\delta_{jk} = 0$, there is a negligible chance of transitions between the locations, and λ_{jk} and λ_{kj} are set to 0

Key point: the δ_{jk} are additional model parameters that are estimated from the data jointly with all other parameters

- A given δ_{jk} could be 0 for some MCMC iterations and nonzero for other iterations
- The posterior estimate of δ_{jk} thus represents the posterior probability of λ_{jk} and λ_{kj} being nonzero

Prior Specification

To incorporate this phylogeographic framework into our phylodynamic model, we must specify prior distributions for all parameters:

- Parameters that characterize the infinitesimal rate matrix $\mathbf{\Lambda} = \mathbf{S\Pi}$

$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & \pi_B s_{12} & \pi_C s_{13} & \pi_D s_{14} \\ \pi_A s_{12} & \cdot & \pi_C s_{23} & \pi_D s_{24} \\ \pi_A s_{13} & \pi_B s_{23} & \cdot & \pi_D s_{34} \\ \pi_A s_{14} & \pi_B s_{24} & \pi_C s_{34} & \cdot \end{bmatrix} \end{array}$$

- The binary indicator variables δ_{jk} that are required for Bayesian stochastic search variable selection

Prior Specification

We assign the equilibrium frequencies $(\pi_A, \pi_B, \pi_C, \pi_D)$ a uniform Dirichlet prior distribution

$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & \pi_B s_{12} & \pi_C s_{13} & \pi_D s_{14} \\ \pi_A s_{12} & \cdot & \pi_C s_{23} & \pi_D s_{24} \\ \pi_A s_{13} & \pi_B s_{23} & \cdot & \pi_D s_{34} \\ \pi_A s_{14} & \pi_B s_{24} & \pi_C s_{34} & \cdot \end{bmatrix} \end{array}$$

For each s_{jk}, we specify a conditional prior that depends on δ_{jk}:

- Assume that $s_{jk}|\delta_{jk}$ is exponentially distributed with scale parameter equal to $\delta_{jk} m_{jk}$
- The mean and standard deviation are equal to $\delta_{jk} m_{jk}$

Bayesian Stochastic Search Variable Selection

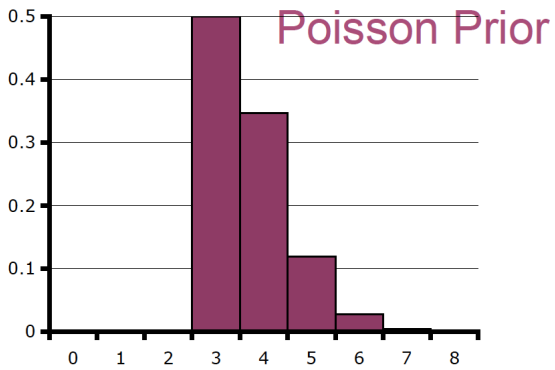
$$\mathbf{\Lambda} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{B} \\ \mathtt{C} \\ \mathtt{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{B} & \mathtt{C} & \mathtt{D} \end{array} \\ \begin{bmatrix} \cdot & 0 & \pi_C s_{13} & 0 \\ 0 & \cdot & \pi_C s_{23} & \pi_D s_{24} \\ \pi_A s_{13} & \pi_B s_{23} & \cdot & 0 \\ 0 & \pi_B s_{24} & 0 & \cdot \end{bmatrix} \end{array}$$

Assume that $s_{jk}|\delta_{jk}$ is exponentially distributed with scale parameter equal to $\delta_{jk} m_{jk}$:

- The mean and standard deviation are equal to $\delta_{jk} m_{jk}$
- When $\delta_{jk} = 0$, this prior forces $s_{jk} = 0$, which amounts to $\lambda_{jk} = \lambda_{kj} = 0$
- When $\delta_{jk} = 1$, the mean and standard deviation equal m_{jk}
- For m_{jk}, we can adopt a distance-based approach and set it to be inversely proportional to the distance between locations j and k

Prior Specification

We want to estimate the indicator variables δ_{jk} from the data, so they also require a prior distribution



- We can work with the sum of all indicators $W = \sum_{j<k} \delta_{jk}$ and assign it a Truncated-Poisson prior distribution with offset $K - 1$ (where K is number of distinct sampling locations) and mean η
- If we set the mean of the Truncated-Poisson prior $\eta = \log(2)$, it is equivalent to placing 50% prior probability on the minimal configuration of $K - 1$ nonzero rates that allows all locations to remain connected

Bayes Factor Test for Significant Diffusion Rates

How can we assess the support for migration between locations j and k?

- Compute the Bayes factor for the indicator δ_{jk}:

$$\text{BF} = \frac{\text{Posterior Odds}}{\text{Prior Odds}} = \frac{P(\delta_{jk} = 1|\mathbf{X}, \mathbf{Y})/(1 - P(\delta_{jk} = 1|\mathbf{X}, \mathbf{Y}))}{P(\delta_{jk} = 1)/(1 - P(\delta_{jk} = 1))}$$

- Under the Truncated-Poisson with mean $\eta = \log(2)$ prior for $W = \sum_{j<k} \delta_{jk}$, we have:

$$P(\delta_{jk} = 1) = \frac{\eta + K - 1}{K(K-2)/2}$$

- $P(\delta_{jk} = 1|\mathbf{X}, \mathbf{Y})$ is the posterior mean of δ_{jk}
- A Bayes factor greater than 3 can be considered substantial support

Phylogeographic Hypothesis Testing

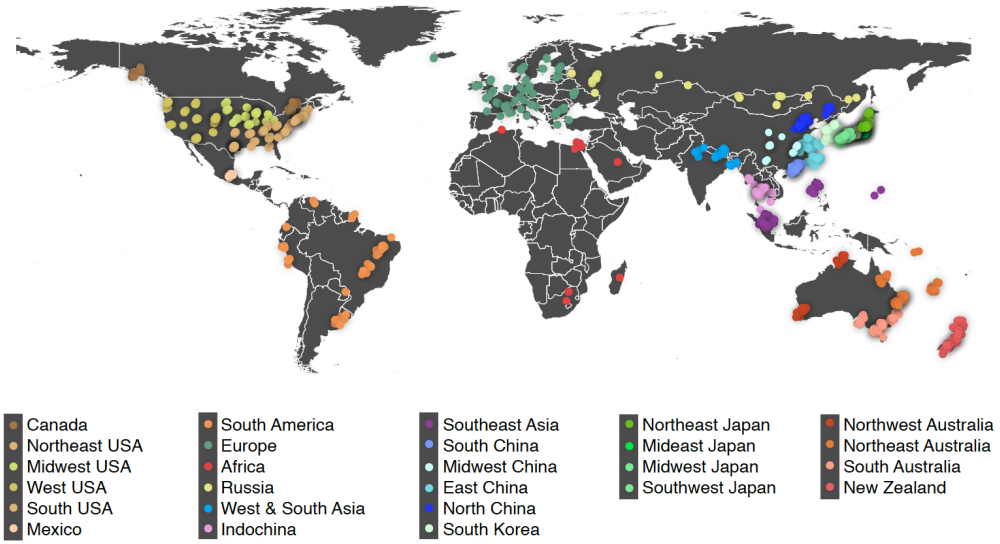
To gain insight into geographic dispersal of pathogens, want to test and quantify relationship between potential predictive variables and spatial spread

Lemey et al. (2014) extend discrete trait analysis phylogeographic framework by parameterizing instantaneous rates λ_{jk} as log-linear functions of predictors:

$$\log \lambda_{jk} = \beta_1 \delta_1 z_{j,k,1} + \beta_2 \delta_2 z_{j,k,2} + \cdots + \beta_P \delta_P z_{j,k,P}$$

- β_p is effect size coefficient that quantifies the relationship between instantaneous rates and predictor z_p
- $\mathbf{z}_p = (z_{1,2,p}, \ldots, z_{K-1,K,p})$ is a vector with corresponding values of the predictor z_p for each entry in instantaneous rate matrix $\mathbf{\Lambda}$
- $\delta_1, \ldots, \delta_P$ are binary indicator variables that determine the inclusion of the corresponding predictors in the model (estimated from the data via Bayesian stochastic search variable selection)

Global Circulation of Human Influenza H3N2



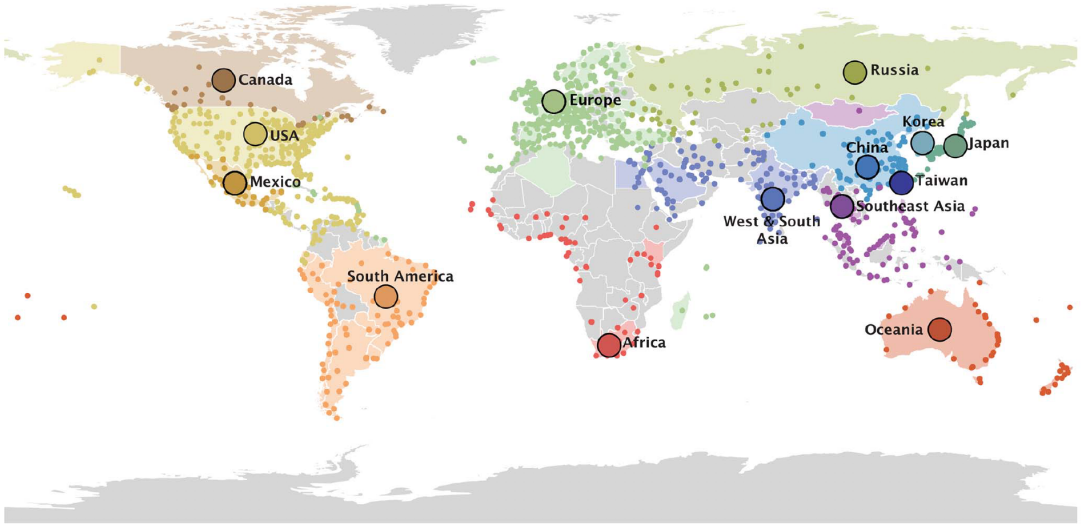
Partition of sampling locations into 26 different regions (Lemey et al., 2014)

Lemey et al. (2014) examine the global transmission dynamics of human influenza subtype H3N2 in an analysis of 1441 hemagglutinin sequences sampled globally from 2002-2007

Samples sizes can strongly influence phylogeographic reconstruction

- Depending on location-specific diversity, overrepresented locations may be more likely to be inferred as source/origin locations and underrepresented locations may be more likely to be inferred as sink/destination locations

Global Circulation of Human Influenza H3N2



Partition of sampling locations into 14 different global air communities (Lemey et al., 2014)

Three different geographic partitions for sampling locations:

- 26 geographic regions – attempt to keep number of samples per location as balanced as possible
- 15 geographic regions – reduce 26 regions by joining regions from single country, down-sample locations with large number of samples
- 14 discrete "air communities" that divide global air transportation network. Airports assigned to community for which it shows highest average affinity (high intra-community connectivity, low inter-community connectivity)

Global Circulation of Human Influenza H3N2

Predictors used in models:

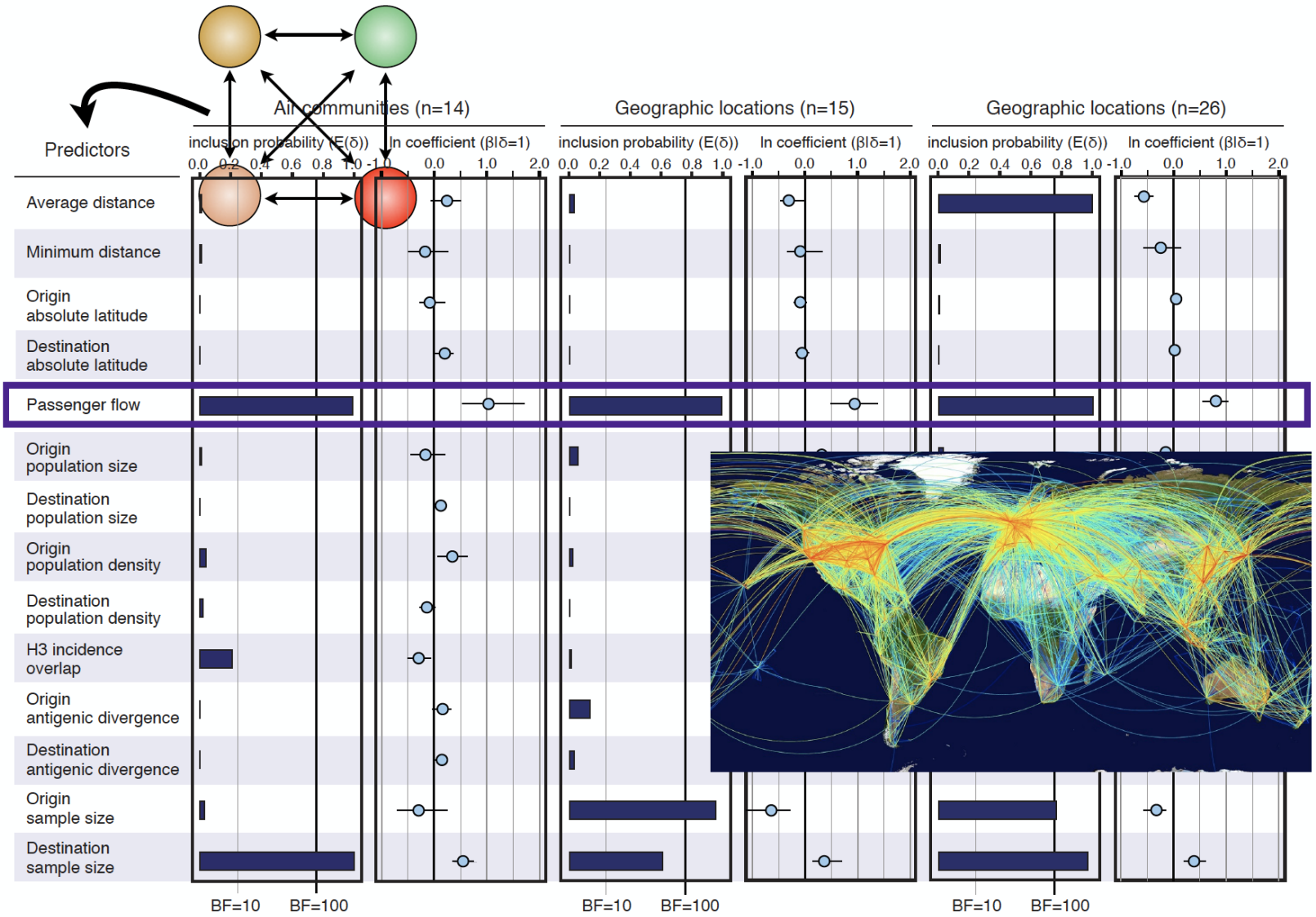
- Average distance (based on pairwise distances between all pairs of airports between two locations) and minimum distance among the pairwise distances
- Absolute latitude of destination and origin (based on absolute values of average latitudes of locations)
- Air passenger flow (total number of seats on flights between each pair of locations per day)

Global Circulation of Human Influenza H3N2

Predictors used in models (continued):

- Origin and destination population size and density (for air communities, based on urban population sizes for airport-associated cities)
- H3N2 Incidence overlap (overlapping area under the origin-destination incidence curves for each pair of locations). Data was unavailable for 26 region partition.
- Average origin and destination antigenic divergence (based on antigenic cartography data for strains in phylogeographic analyses)
- Origin and destination sample sizes (to test impact of sampling effects)

Global Circulation of Human Influenza H3N2



(Lemey et al., 2014)

Global Circulation of Human Influenza H3N2

Consistent and strong evidence that air passenger flow is the dominant driver of global spread of H3N2 influenza viruses

Inclusion of most sample size predictors, providing support that other predictors are not included in the model due to sampling bias

For the 26 location partition, there are high inclusion probabilities for average distance, origin population density and destination population density. In all of these cases, the effect size is negative.

- For smaller geographic areas, average distance predictor may capture human mobility other than air travel (such as workplace commuting)
- Significance of these predictors only for geographic partition with more confined areas shows that the key predictors of influenza spread will depend on scale
- Testing hypotheses at smaller scale may be possible, but requires adequate sampling. The need to adjust geographic partitions to balance sample sizes results in "locations" of widely varying areas and population sizes.

Incorporating Travel History in Discrete Trait Analysis

As SARS-CoV-2 was spreading in early 2020, there was an abundance of genomic data, but researchers had to confront several limitations:

- A relatively slow evolutionary rate and limited sequence diversity led to poorly resolved evolutionary reconstruction
- Large spatiotemporal biases existed in the available genome data

Lemey et al. (2020) sought to ameliorate these difficulties by incorporating individual travel history in a discrete trait analysis framework:

- Genomic data from returning travelers may help uncover pathogen diversity in undersampled locations

Incorporating Travel History in Discrete Trait Analysis

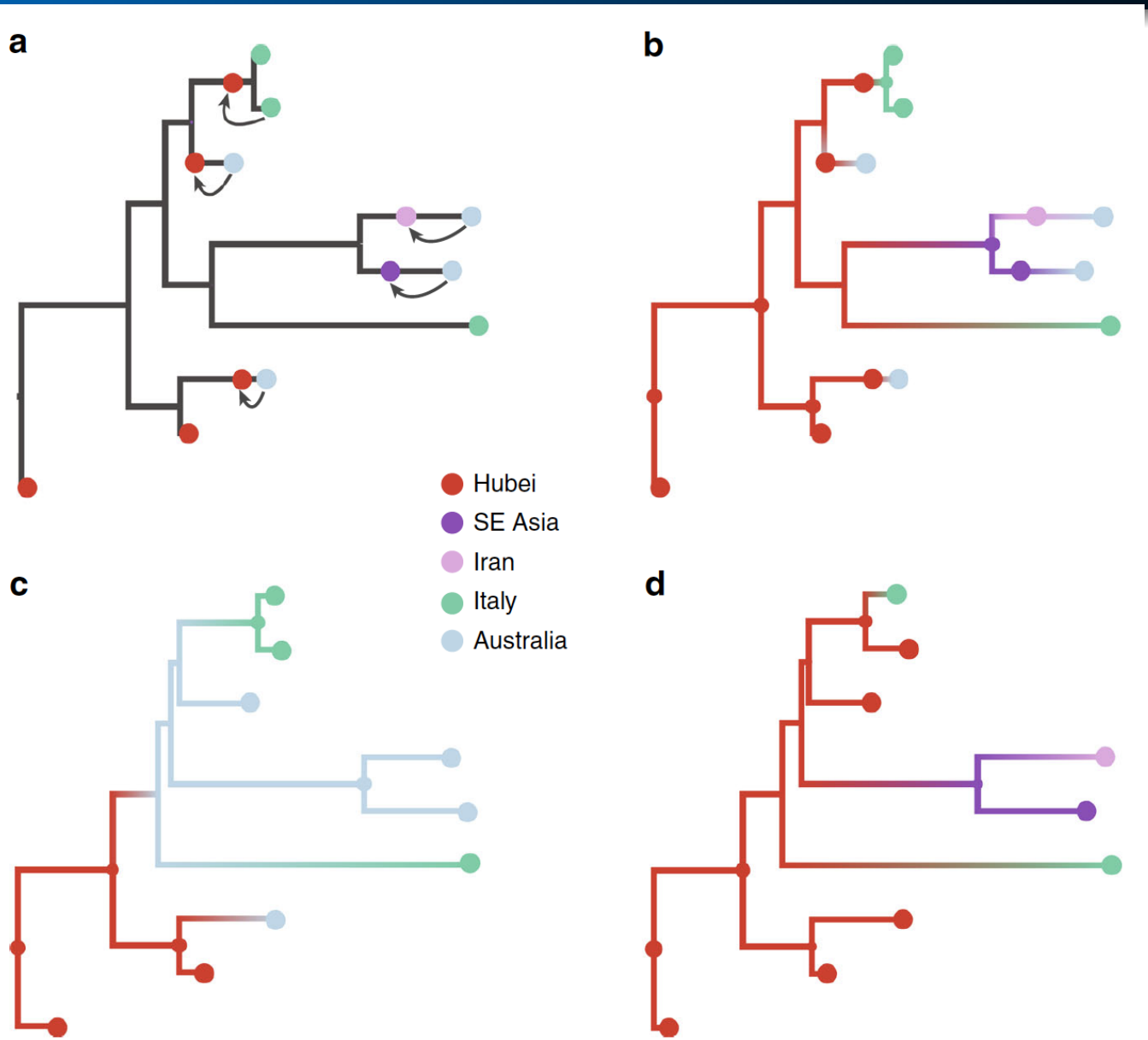
Problem: An individual returns home to location j (the "sampling location") after traveling to location i (the "travel origin location"). How do we set the tree tip location corresponding to the individual?

- Setting it to j ignores information about the ancestral location of the sequence.
- Setting it to i creates a data mismatch between location and time, and ignores final transition from i to j. This last transition is especially important when the infected traveler produces a transmission chain after returning home to j.

Solution: Augment phylogenetic tree with ancestral node that is associated with a location (but not with a known sequence)

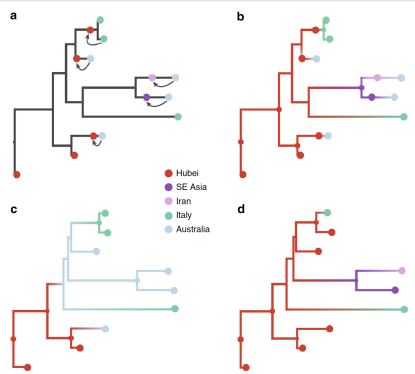
- Insert ancestral node at time at which traveler started journey from i to j

Incorporating Travel History in Discrete Trait Analysis



(Lemey et al., 2020) a) Ancestral nodes w/ travel origin locations b) Results w/ sampling locations and travel history c) Results w/ sampling locations only d) Results w/ travel origin locations at tips

Incorporating Travel History in Discrete Trait Analysis



(Lemey et al., 2020)

Example: 9 genomes (2 sampled from Hubei, 4 from Australia, 3 from Italy). Travel history is available for 5 genomes (4 sampled from Australia, 1 from Italy)

- Panel a) illustrates incorporation of ancestral nodes with travel origin locations, b) shows results using sampling locations and travel history, c) shows results using only sampling locations, d) shows results using travel origin locations for the tree tips for individuals with available travel history
- c) shows an unrealistic Australian ancestry and two transitions from Australia to Italy, likely because Australia is represented by the largest number of tips
- d) is closer to b), infers substantial history in Hubei, but misses transitions along 4 tip branches and suggests transition from Huebi for Italian patient without that travel history

Strengths and Weaknesses of Discrete Trait Analysis

The discrete trait analysis approach to phylogeography has many strengths:

- Computationally efficient (integrates over all migration histories via a pruning algorithm), scales to large data sets with a lot of different sampling locations
- Extension that incorporates explanatory variables enables formal phylogeographic hypothesis testing, extension that incorporates individual travel histories

However, discrete trait analysis has a number of limitations:

- Prior distribution of phylogenetic tree ignores information about migration process (does not take into account the impact of migration on branch lengths and shape of tree)
- When sampling intensity is not proportional to subpopulation size, we can get biased migration rate estimates

Structured Coalescent

These limitations can be overcome by phylogeographic modeling with the structured coalescent (Hudson, 1990; Notohara, 1990; Beerli and Felsenstein, 2001; Vaughan et al., 2014):

$$P(\tau, \mathbf{M}, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X} | \tau, \boldsymbol{\phi}) P(\tau, \mathbf{M} | \boldsymbol{\theta}, \mathbf{Y}) P(\boldsymbol{\phi}) P(\boldsymbol{\theta}).$$

Here, $\mathbf{M}$ denotes migration histories and $\boldsymbol{\theta}$ consists of migration rate and population size parameters for the structured coalescent.

Note the key differences with the discrete trait analysis model:

$$P(\tau, \boldsymbol{\phi}, \boldsymbol{\Lambda} | \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X} | \tau, \boldsymbol{\phi}) P(\mathbf{Y} | \tau, \boldsymbol{\Lambda}) P(\tau) P(\boldsymbol{\phi}) P(\boldsymbol{\Lambda})$$

Structured Coalescent

These limitations can be overcome by phylogeographic modeling with the structured coalescent (Hudson, 1990; Notohara, 1990; Beerli and Felsenstein, 2001; Vaughan et al., 2014):

$$P(\tau, \mathbf{M}, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X} | \tau, \boldsymbol{\phi}) P(\tau, \mathbf{M} | \boldsymbol{\theta}, \mathbf{Y}) P(\boldsymbol{\phi}) P(\boldsymbol{\theta}).$$

Here, $\mathbf{M}$ denotes migration histories and $\boldsymbol{\theta}$ consists of migration rate and population size parameters for the structured coalescent.

Note the key differences with the discrete trait analysis model:

$$P(\tau, \boldsymbol{\phi}, \boldsymbol{\Lambda} | \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{X} | \tau, \boldsymbol{\phi}) P(\mathbf{Y} | \tau, \boldsymbol{\Lambda}) P(\tau) P(\boldsymbol{\phi}) P(\boldsymbol{\Lambda})$$

Unfortunately, the structured coalescent is very computationally demanding

- Need to consider all possible migration histories
- Impractical in situations with large number of subpopulations

Approximations of the Structured Coalescent

Researchers have introduced phylogeographic inference frameworks based on approximations of the structured coalescent that

- Are more computationally efficient (De Maio et al., 2015; Müller et al., 2017)
- Incorporate explanatory variables (Müller et al., 2019)
- Feature flexible nonparametric modeling of subpopulation (deme) effective population sizes (Müller et al., 2025)

These models are available in BEAST 2

Beast2

Bayesian evolutionary analysis by sampling trees

`www.beast2.org`

Approximations of the Structured Coalescent

Structured coalescent approximations are more computationally efficient than the structured coalescent and do not suffer from some of the shortcomings exhibited by discrete trait analysis.

However,

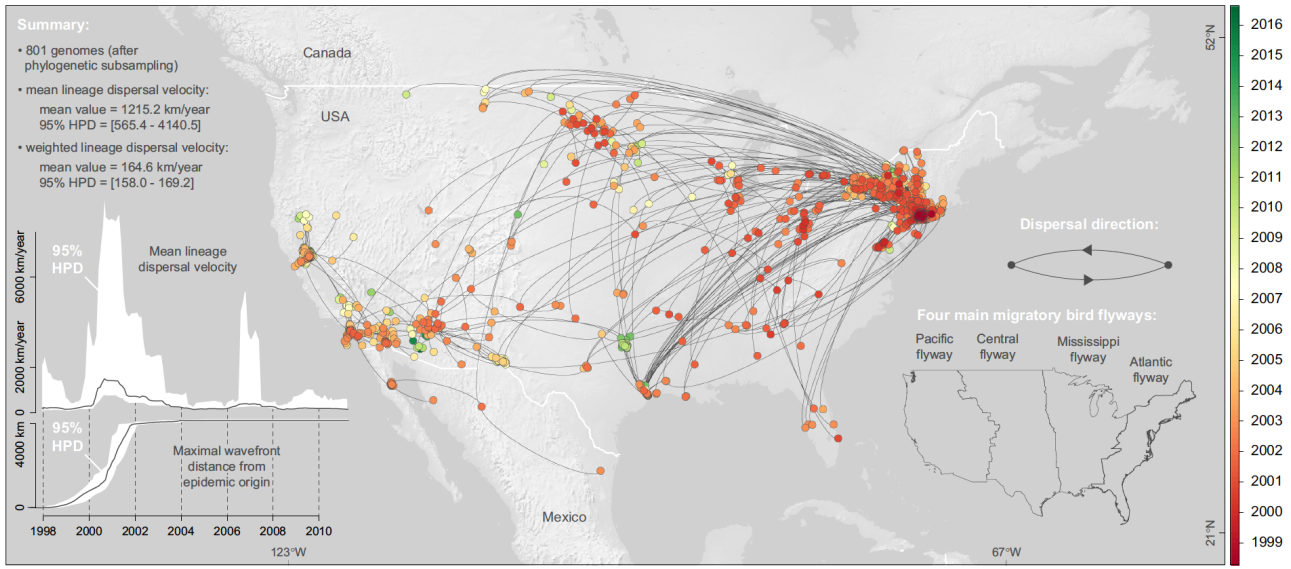
- The number of discrete locations that these structured coalescent approximations can accommodate is still substantially limited compared to discrete trait analysis
- The common assumption of subpopulations (demes) with constant effective population sizes can lead to biased inferences of within-deme and between-deme dynamics

Moving beyond Discrete Locations

Phylogeographic inference with a fixed set of discrete locations has several drawbacks:

- Often requires an arbitrary grouping of sampling locations that may lead to oversimplified abstraction or unrealistic subdivision of the study area
- Restriction that ancestral locations can only correspond to sampled locations can be unrealistic
- Sampling locations may be more continuously distributed

Moving beyond Discrete Locations



(Dellicour et al., 2020)

Alternative: phylogeographic inference in continuous space

- Spatially explicit reconstruction overcomes many unrealistic modeling assumptions of discrete phylogeographic frameworks
- However, there are situations where discrete modeling works better (e.g., modeling pathogen dispersal between locations separated by geographic barriers)

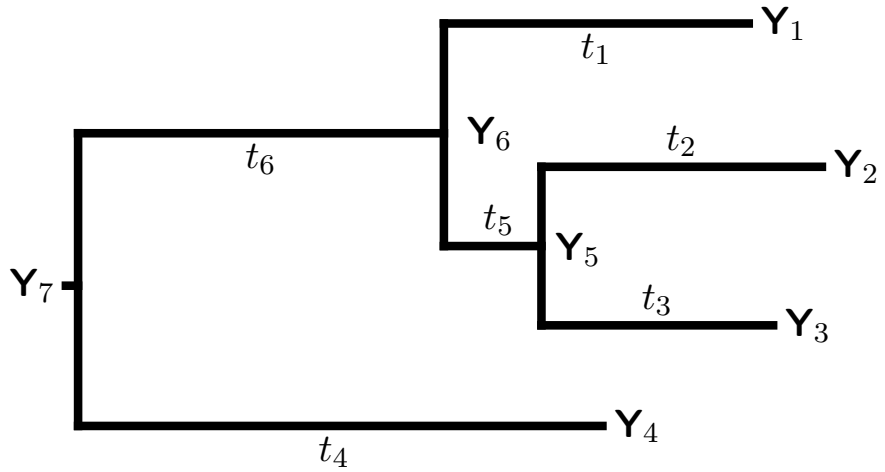
Brownian Diffusion on a Phylogenetic Tree

Continuously varying traits on a phylogenetic tree can be modeled as Brownian diffusion or, in other words, as a random walk (Felsenstein, 1985)

- Phylogeography: "traits" are bivariate latitude and longitude coordinates

We observe K-dimensional continuous trait values $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ corresponding to tips of phylogenetic tree τ

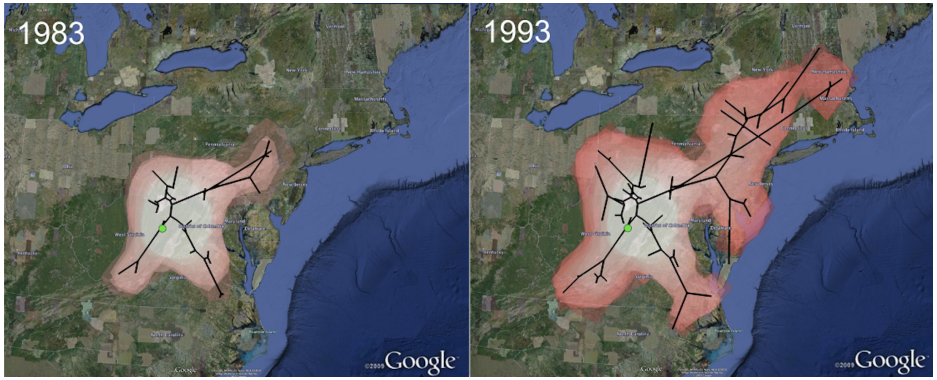
- Unobserved trait values $\mathbf{Y}_{N+1}, \ldots, \mathbf{Y}_{2N-1}$ at internal nodes and root



Multivariate correlated Brownian diffusion is characterized by the following "displacement" of the location along a phylogenetic tree branch:

$$\left(\mathbf{Y}_i - \mathbf{Y}_{pa(i)}\right) | \mathbf{Y}_{pa(i)} \sim N\left(\mathbf{0}, t_i \boldsymbol{\Sigma}\right)$$

Relaxed Random Walk



(Lemey et al., 2010)

Standard Brownian diffusion is widely used but can be overly restrictive. Lemey et al. (2010) introduced a "relaxed random walk" that models branch-specific variation of the displacement variance matrix $\mathbf{\Sigma}$ by introducing branch-specific scalars ϕ_i so that

$$\left(\mathbf{Y}_i - \mathbf{Y}_{pa(i)}\right) | \mathbf{Y}_{pa(i)} \sim N\left(\mathbf{0}, t_i \phi_i \mathbf{\Sigma}\right)$$

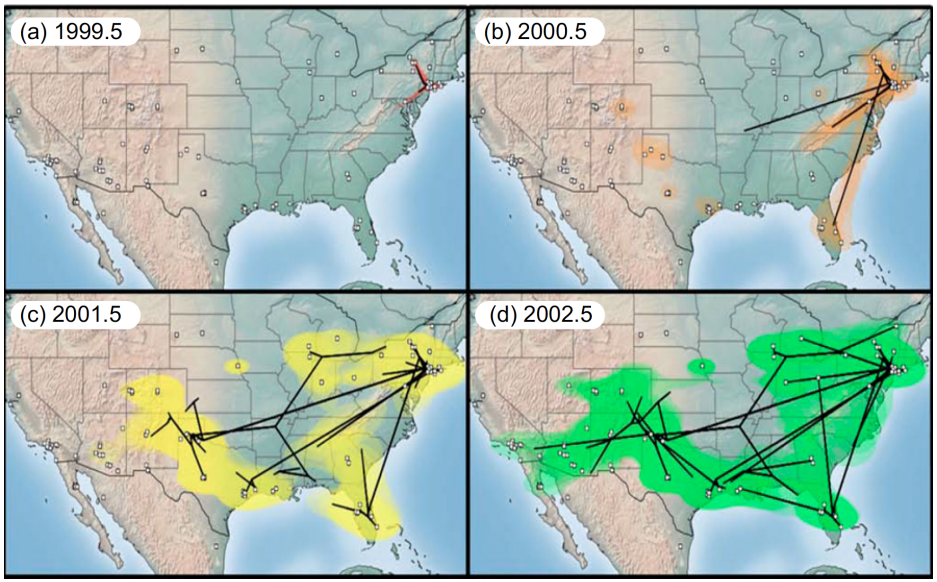
- The scalars ϕ_b can be drawn from different distributions. For example,

$$\phi_i \overset{\text{iid}}{\sim} \text{Gamma}(\nu/2, \nu/2),$$

or, if we wish to accommodate higher levels of overdispersion,

$$\phi_i \overset{\text{iid}}{\sim} \text{Log-Normal}(1, \sigma)$$

Example: West Nile Virus



(Pybus et al., 2012)

Pybus et al. (2012) use continuous phylogeography to study the spread of West Nile virus in North America

- Data: 104 West Nile virus complete genome sequences sampled between 1999 and 2007
- Compare homogeneous diffusion (standard random walk) to heterogeneous diffusion (relaxed random walk models with rates drawn from gamma or log-normal distributions)

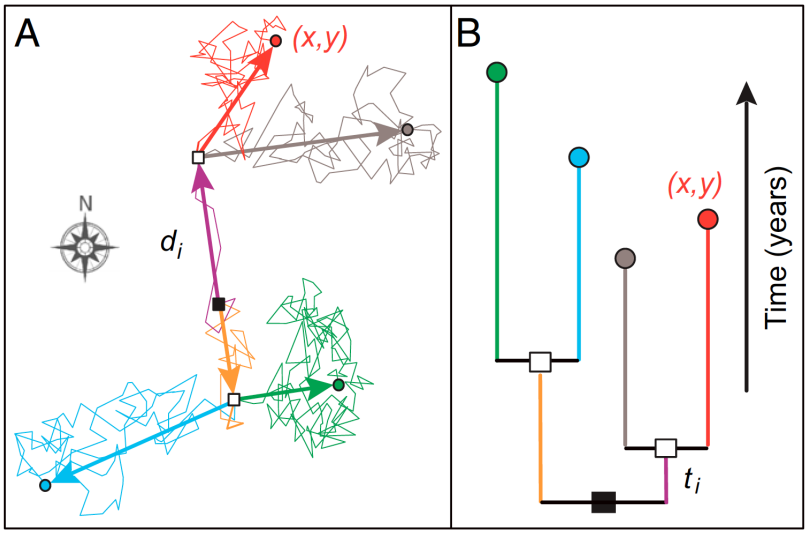
Example: West Nile Virus

	Homogeneous Diffusion	Heterogeneous Diffusion
Log Marginal Likelihood	-643.45	-399.43
Date of Epidemic Origin	1998.6 (1997.6, 1999.3)	1998.5 (1997.8, 1999.1)
Mean Evolution Rate	0.00058 (0.00049, 0.00066)	0.00057 (0.00051, 0.00064)
SD of Rates	0.38 (0.23, 0.53)	0.33 (0.21, 0.45)
Lat. of Origin	40.3 (37.1,43.7)	41.1 (40.4, 43.2)
Long. of Origin	-76.5 (-82.9, -70.5)	-74.6 (-76.1, -73.3)

(Pybus et al., 2012)

- Comparison of homogeneous diffusion (standard random walk) with heterogeneous diffusion (best-fitting relaxed random walk model that draws rates from gamma distribution)
- Under heterogeneous diffusion, we see a much better model fit (as measured by marginal likelihood) and more precise estimates of spatial parameters

Integration of Spatial Epidemiology and Phylogenetics



(Pybus et al., 2012)

The diffusion coefficient D is a fundamental ecological measure that reflects the area that an infected host will explore per unit time

$$D \approx \frac{1}{n} \sum_{i=1}^{n} \frac{d_i^2}{4t_i}$$

- n - number of phylogenetic tree branches under consideration (can be subset of all tree branches to explore how D varies)
- t_i - duration in years of branch i
- d_i - great circle distance moved along branch i away from its starting position

Integration of Spatial Epidemiology and Phylogenetics

The diffusion coefficient D together with the basic reproductive number R_0 (expected number of individuals infected by one infected individual) determines the wavefront velocity

D is generally difficult to estimate:

- Usually requires tracking movements of large number of infected hosts by time-consuming mark/recapture or telemetry. This approach will nevertheless fail to adequately capture spatial dynamics when dispersal behavior among individuals is highly variable.
- Alternatively, D can be inferred from relationship with wavefront velocity, but this requires knowing R_0 without error.

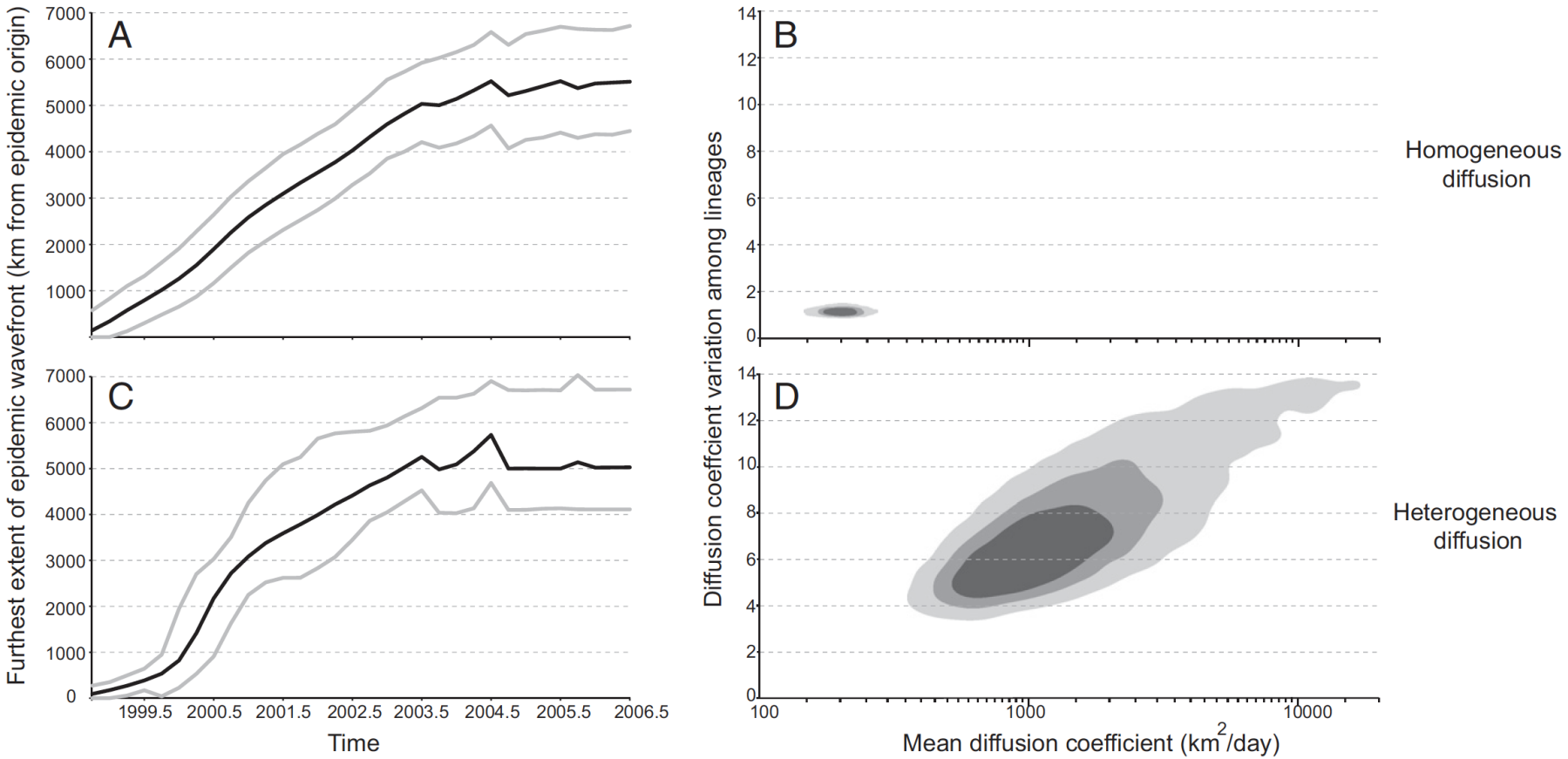
In addition to computing the diffusion coefficient,

$$D \approx \frac{1}{n} \sum_{i=1}^{n} \frac{d_i^2}{4t_i},$$

we can compute the mean lineage dispersal velocity as follows:

$$\text{mean lineage dispersal velocity} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i}{t_i}$$

Integration of Spatial Epidemiology and Phylogenetics



(Pybus et al., 2012) "Homogeneous diffusion" refers to a standard random walk model and "heterogeneous diffusion" refers to the best-fitting relaxed random walk model, which features branch rates drawn from a one-parameter Gamma distribution. The gradient of the curve in plots A and C yields the wavefront velocity. The contours in plots B and D indicate 50%, 75%, 95% HPD region of kernel density estimates of diffusion coefficient.

Integration of Spatial Epidemiology and Phylogenetics

Under the standard random walk, the wavefront velocity is approximately constant at 1000 km/year until it reaches the western seabord.

The best fitting relaxed random walk model indicates that the the invasion accelerated: between 1999 and 2003, the distance from the origin to wavefront doubled every 0.8 years on average.

- Notably, this acceleration rate (estimated from genomic data) is nearly identical to the acceleration rate that was independently estimated from spatiotemporal WNV incidence (Mundt et al., 2012).

The mean D under a standard random walk model is about 200 km^2/day. Under the best fitting relaxed random walk, D is much more variable and highly diffusive (has a mean of about 1000 km^2/day)

Integration of Spatial Epidemiology and Phylogenetics

To summarize, the relaxed random walk analysis uncovers spatiotemporal dynamics that are characterized by highly variable dispersal, with a few rapid long-range movements (consistent with strong correlation of mean and variation of D among lineages) and less-diffusive lineages that likely represent local transmission among hosts and vectors while they move within their typical home ranges.

Pybus et al. (2012) note that many mathematical models of WNV spread in North America that assumed homogeneous diffusion (and typically modeled host dispersal using data on short-term home range movements of birds) exhibited low mean diffusion coefficients (less than 14 km^2/day). This resulted in significant overestimates of R_0 (greater than 25).

On the other hand, Pybus et al. (2012) show that the observed high wavefront velocity of 1000 km/year does not have to be explained by a combination of a highly transmissible virus with a weakly diffusive host. A better explanation is a lower R_0 that transmits among hosts with highly variable dispersal.

Relaxed Directional Random Walk

A standard random walk can also be extended to a relaxed directional random walk that accounts for and quantifies directional trends via nonzero displacement means

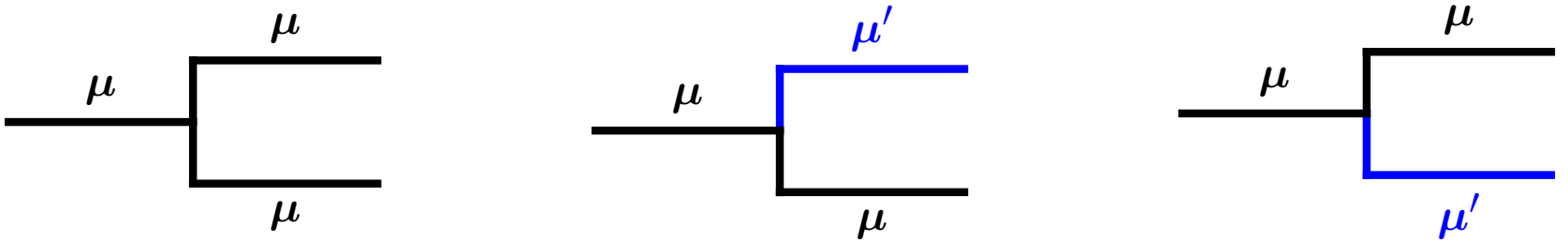
$$\left(\mathbf{Y}_i - \mathbf{Y}_{pa(i)}\right) | \mathbf{Y}_{pa(i)} \sim N\left(t_i \boldsymbol{\mu}_i, t_i \boldsymbol{\Sigma}\right)$$

Allowing each branch to assume a unique trend vector makes model unidentifiable: there can exist $\boldsymbol{\mu}_\tau^* \neq \boldsymbol{\mu}_\tau$ such that

$$P\left(\mathbf{Y} | \boldsymbol{\tau}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_\tau^*\right) = P\left(\mathbf{Y} | \boldsymbol{\tau}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_\tau\right)$$

To ensure identifiability while permitting trend variation, we allow two possibilities whenever a parent branch splits:

- both child branches assume the same trend as parent branch
- one child branch takes on a new trend, other assumes parent branch trend



Relaxed Directional Random Walk

We want to infer the number and types of trend changes that occur

Let γ_i denote a branch-specific trend change indicator, and consider

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_{pa(i)} + \boldsymbol{\alpha}_i.$$

Trend differences $\boldsymbol{\alpha}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}_{pa(i)}$ are *a priori* independent

$$\boldsymbol{\alpha}_i \sim N\left(\mathbf{0}, \gamma_i \sigma^2 \mathbf{I}\right)$$

If $\gamma_i = 0$ (no trend change) the prior variance on $\boldsymbol{\alpha}_i$ shrinks to zero forcing $\boldsymbol{\alpha}_i = \mathbf{0}$ in the posterior

The joint space $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is explored through MCMC

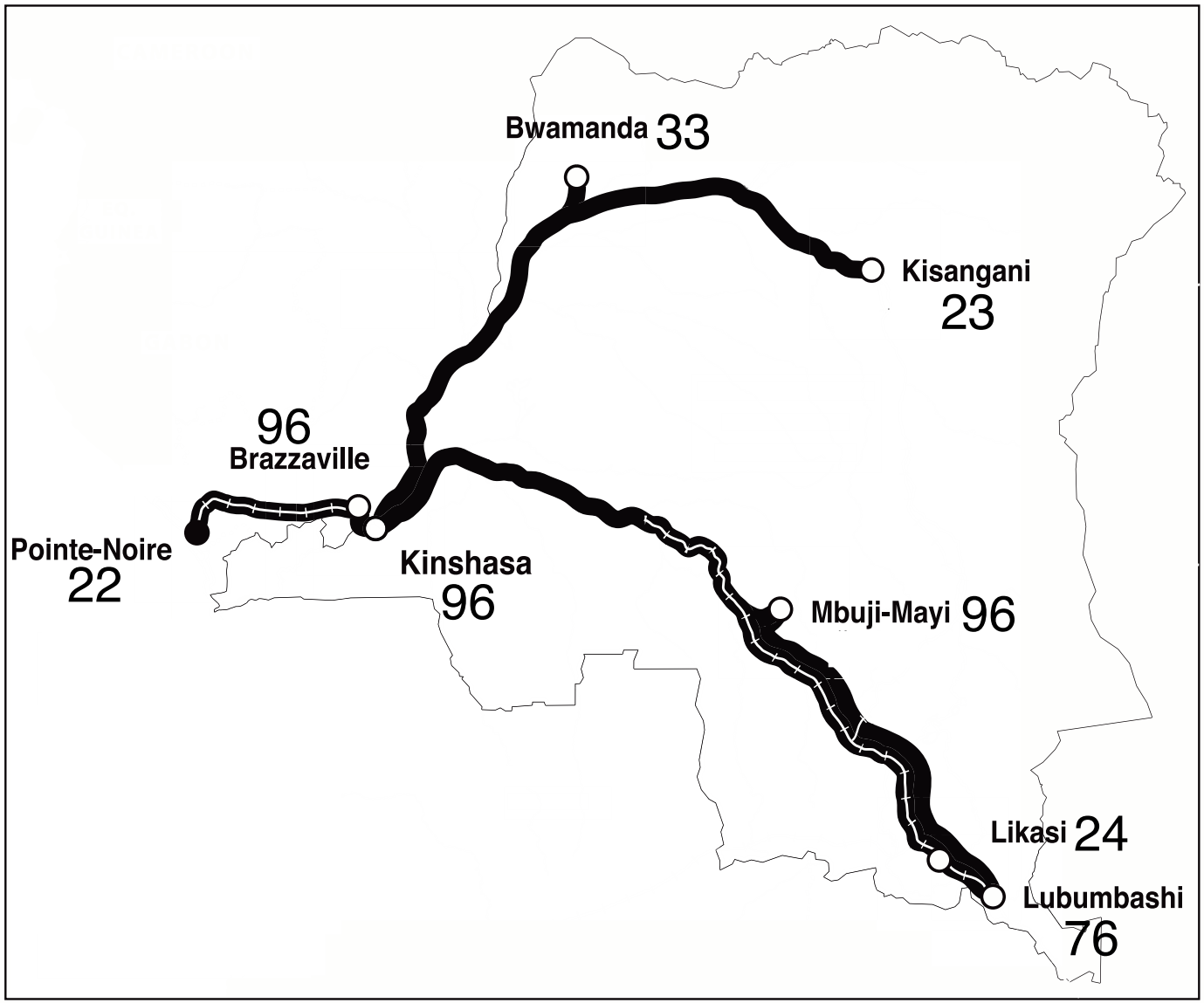
Spread of HIV-1 in Central Africa

Faria et al. (2014) explored the early spatial expansion and epidemic dynamics of HIV-1 in central Africa by analyzing sequence data sampled from countries in the Congo River basin.

They employed a discrete phylogeographic inference framework (Lemey et al., 2009) on 466 HIV-1 sequences sampled between 1985-2004. Among other things, they showed that the pandemic originated in Kinshasa in the 1920s.

We will compare the inferences of different Brownian diffusion models for continuous phylogeographic analysis on this data set.

Sampling Locations of 466 HIV-1 Sequences



(Faria et al., 2014) Numbers of samples from different locations are next to city names.

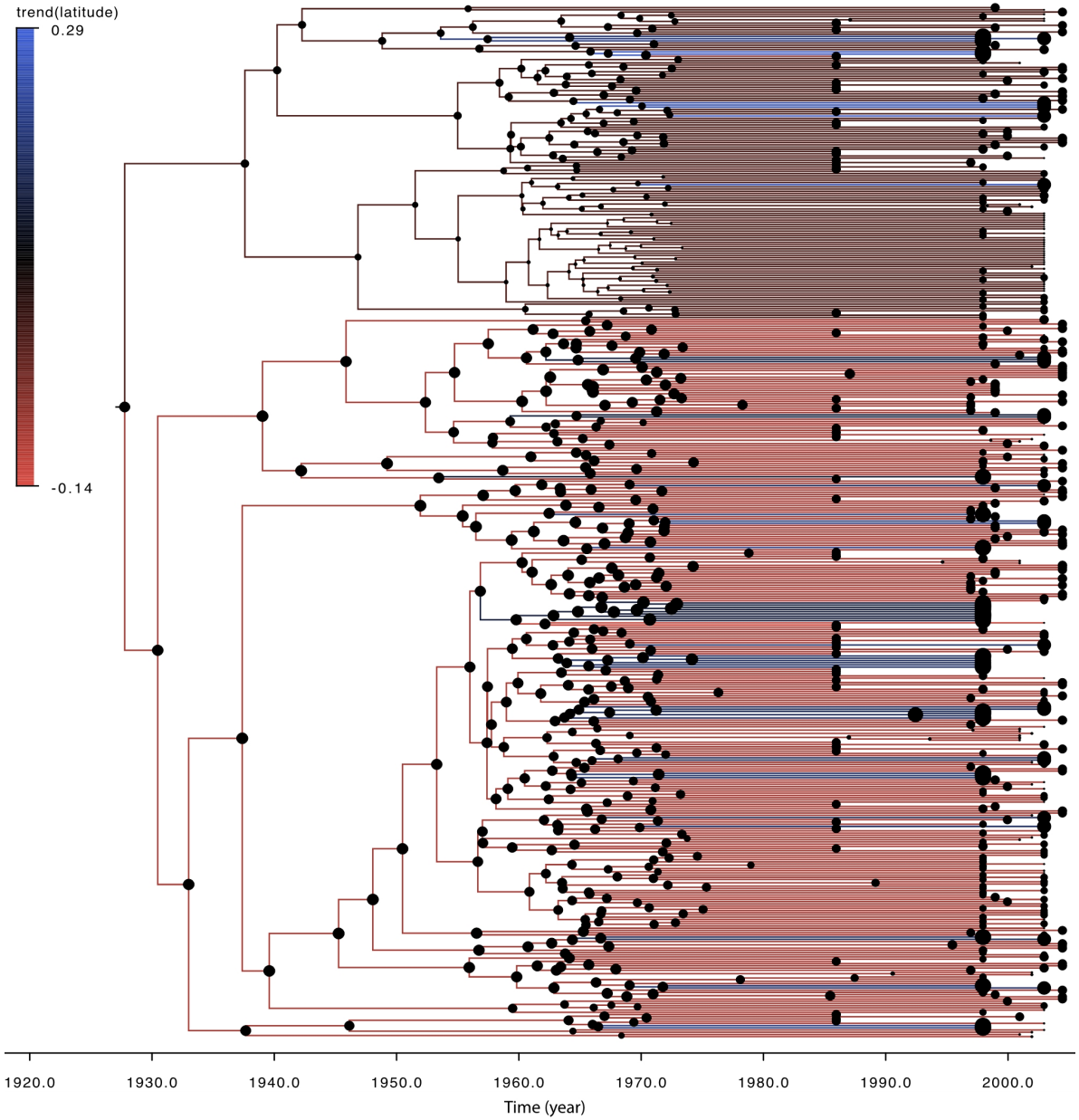
Comparison of Brownian Diffusion Models

	No Trend		Constant Trend		Relaxed Trend	
Trend (Long.)	-	-	0.30	(0.26, 0.33)	0.12	(0.08, 0.16)
Disp. Var. Rate (Long.)	**0.59**	**(0.55, 0.64)**	**0.37**	**(0.34, 0.40)**	0.43	(0.39, 0.45)
Trend (Lat.)	-	-	-0.09	(-0.11, -0.06)	-0.03	(-0.05, -0.01)
Disp. Var. Rate (Lat.)	0.25	(0.23, 0.27)	**0.23**	**(0.21, 0.25)**	**0.13**	**(0.12, 0.14)**
Correlation	-0.47	(-0.54, -0.40)	-0.40	(-0.47, -0.32)	-0.84	(-0.87, -0.82)

$$\begin{bmatrix} Y_i^{long} \\ Y_i^{lat} \end{bmatrix} - \begin{bmatrix} Y_{pa(i)}^{long} \\ Y_{pa(i)}^{lat} \end{bmatrix} \sim N\left(t_i \begin{bmatrix} \mu^{long} \\ \mu^{lat} \end{bmatrix}, t_i \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

- Significant southward and eastward trends (units: degrees per year)
- Longitude: 1.48 trend changes with 95% BCI (1, 3)
- Latitude: 28.13 trend changes with 95% BCI (27, 29)
- Failure to adequately model trends can result in inflation of displacement variance rates (Σ_{11}, Σ_{22})

Latitudinal Trends



Prior Distributions for Phylogenetic Tree

$$P(\tau, \boldsymbol{\phi}|\mathbf{X}) \propto P(\mathbf{X}|\tau, \boldsymbol{\phi})P(\tau)P(\boldsymbol{\phi})$$

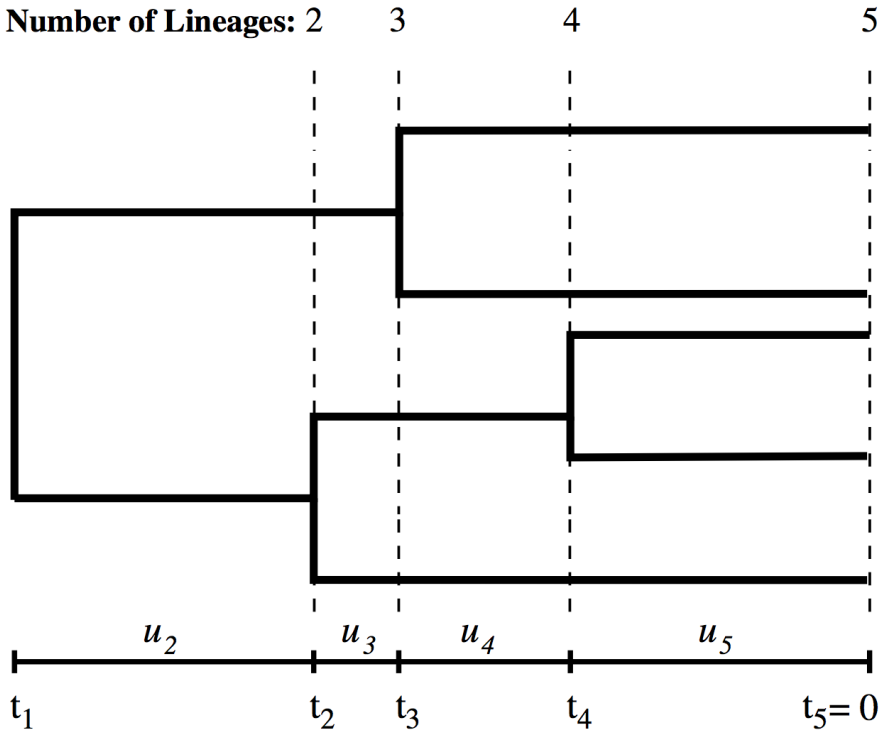
Bayesian phylogenetic inference models require a prior distribution $P(\tau)$ for the phylogenetic tree

Coalescent-based priors are very popular

- The coalescent generates genealogies that approximate the ancestry of a sample arising from an idealized Fisher-Wright population
- The effective population size parameter $N_e(t)$ is of fundamental interest in infectious disease epidemiology as a measure of pathogen genetic diversity and as a proxy for pathogen circulation

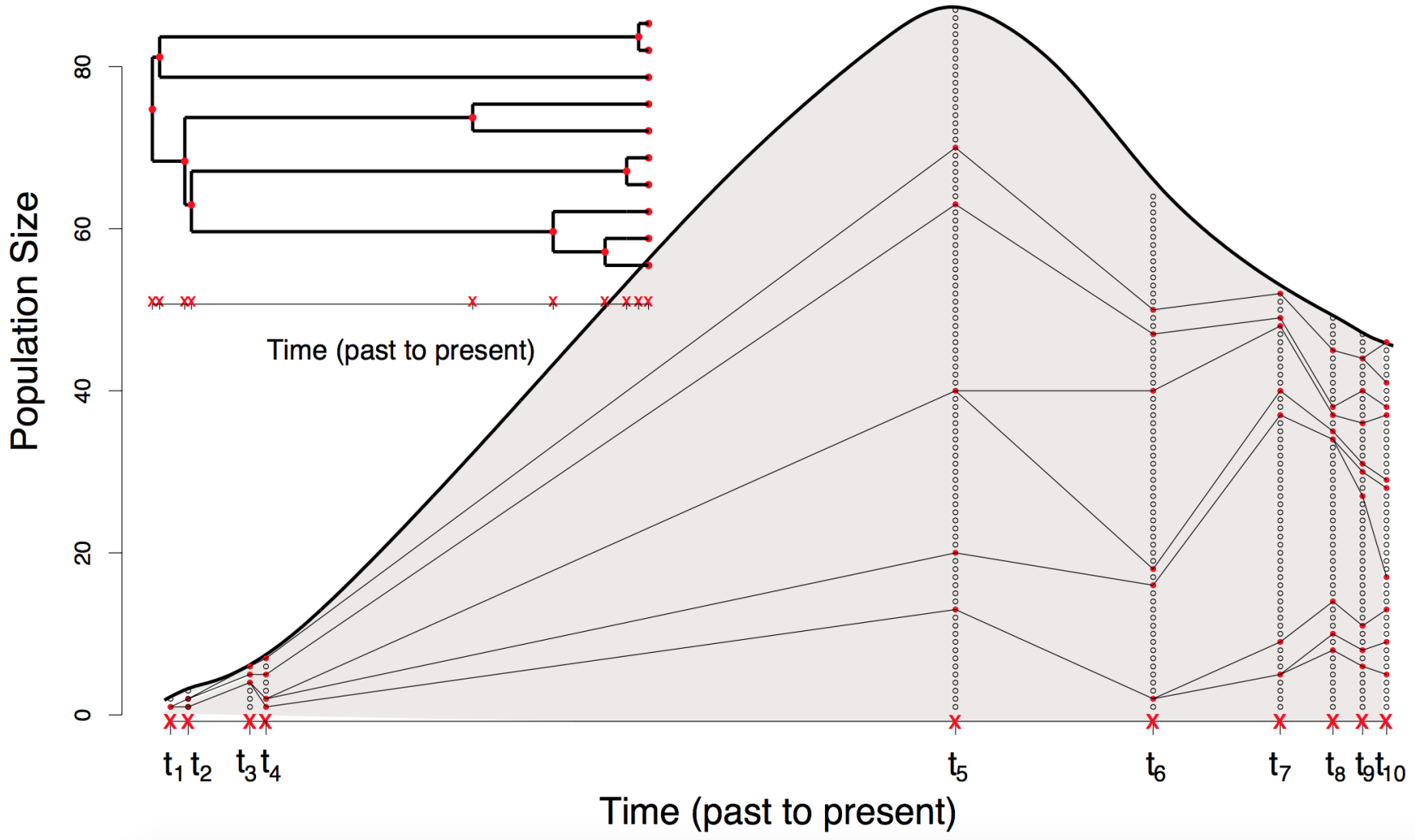
Review: Coalescent Theory

- Coalescent generates a genealogy relating a random sample of n individuals from a population of size $N_e(t)$
- Merging of two lineages is called a coalescent event



- Time increases into the past, starting with sampling time $t_n = 0$
- t_k is time of $(n-k)$-th coalescent event
- $$P(u_k|t_k) = \frac{k(k-1)}{2N_e(u_k+t_k)} \exp\left[-\int_{t_k}^{u_k+t_k} \frac{k(k-1)}{2N_e(t)} dt\right]$$

Genealogy and Population Dynamics



The effective population size $N_e(t)$ characterizes genetic diversity, and the coalescent enables inference of the effective population size over time from a genealogy

Bayesian Model

Bayesian Skygrid model: assume (scaled) $N_e(t)$ is piecewise constant, changes values at user-specified temporal change points $x_1, \ldots, x_K$

- $\log N_e(t) = \gamma_k$ for $x_{k-1} \leq t < x_k$ and $\log N_e(t) = \gamma_{K+1}$ for $t \geq x_K$

Need to estimate $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+1})$

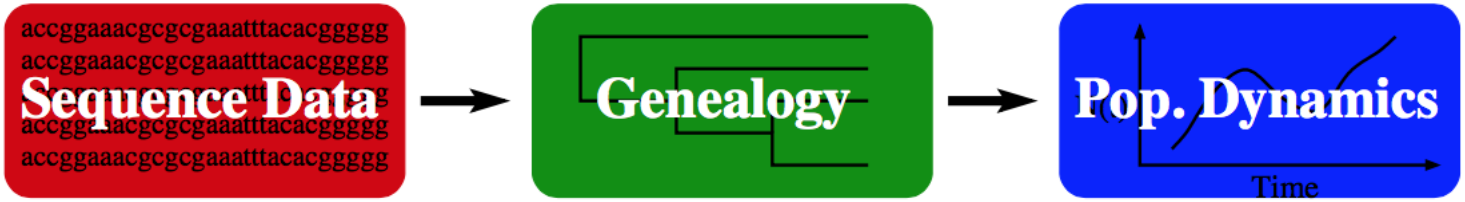
$$P(\boldsymbol{\gamma}, \rho|\tau) \propto P(\tau|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\rho)P(\rho)$$

- $P(\tau|\boldsymbol{\gamma})$ - coalescent likelihood
- $P(\boldsymbol{\gamma}|\rho)$ - Gaussian Markov random field (GMRF) smoothing prior

$$P(\boldsymbol{\gamma}|\rho) \propto \rho^{K/2} \exp\left[-\frac{\rho}{2}\sum_{i=1}^{K}(\gamma_{i+1} - \gamma_i)^2\right]$$

- $P(\rho)$ - diffuse Gamma prior on precision parameter ρ

Joint Inference from Molecular Sequence Data



$$P(\tau, \mathbf{Q}, \rho, \boldsymbol{\gamma}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Q}, \tau)P(\mathbf{Q})P(\tau|\boldsymbol{\gamma})P(\boldsymbol{\gamma}|\rho)P(\rho)$$

X - molecular sequence data

Q - mutation model parameters

τ - phylogenetic tree

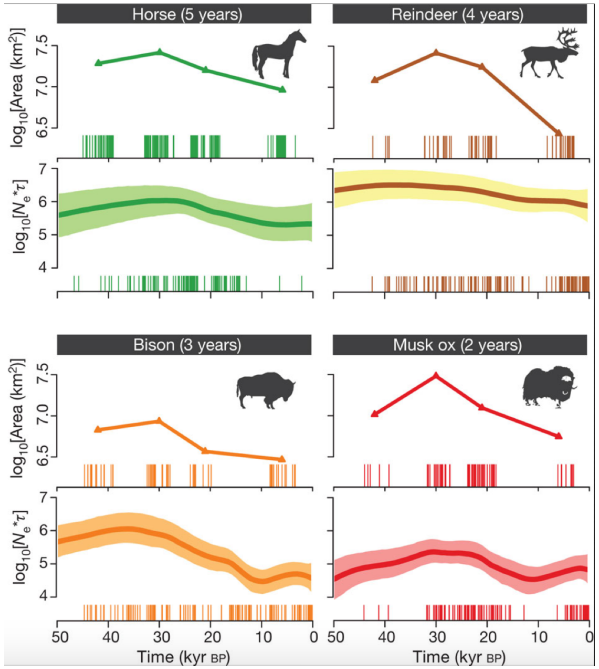
$\boldsymbol{\gamma}$ - effective population size trajectory

ρ - GMRF prior precision

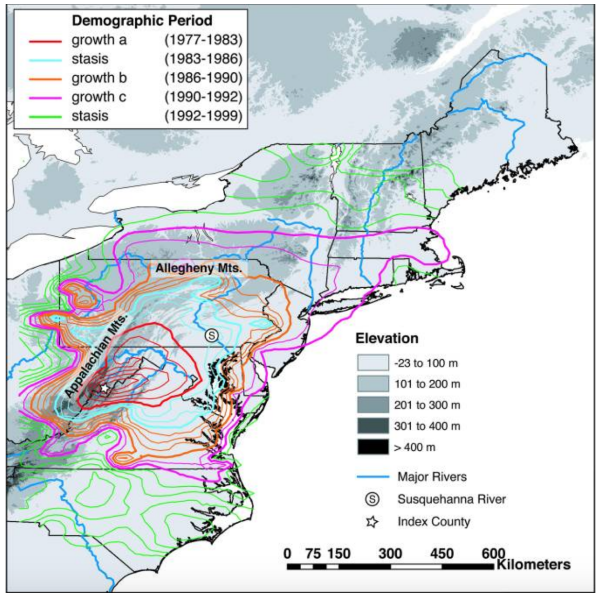
Sample from $P(\tau, \mathbf{Q}, \rho, \boldsymbol{\gamma}|\mathbf{X})$ using MCMC

Understanding Past Population Dynamics

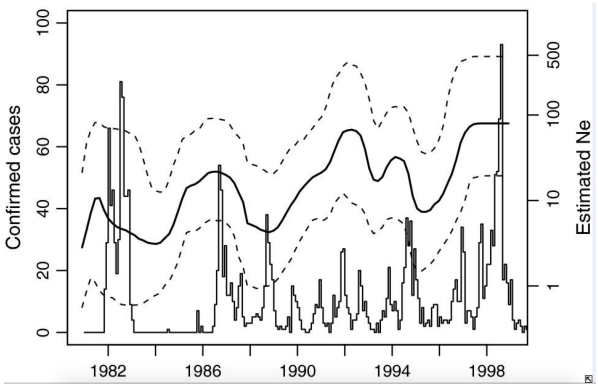
A major theme in the literature:



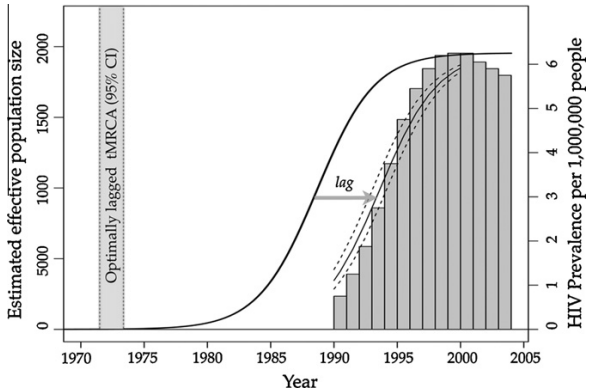
(Lorenzen et al., 2011)



(Biek et al., 2007)



(Bennett et al., 2010)



(Faria et al., 2012)

Association of Pop. Dynamics w/ Time-Varying Covariates

The relationships between population dynamics and time-varying covariates have typically been examined in *post hoc* fashions:

- Informal visual comparisons
- More formally, employ a generalized linear model (GLM) framework with mean estimate of effective population size as response variable

  Can estimate effect size coefficient β to quantify relationship and test for statistical significance

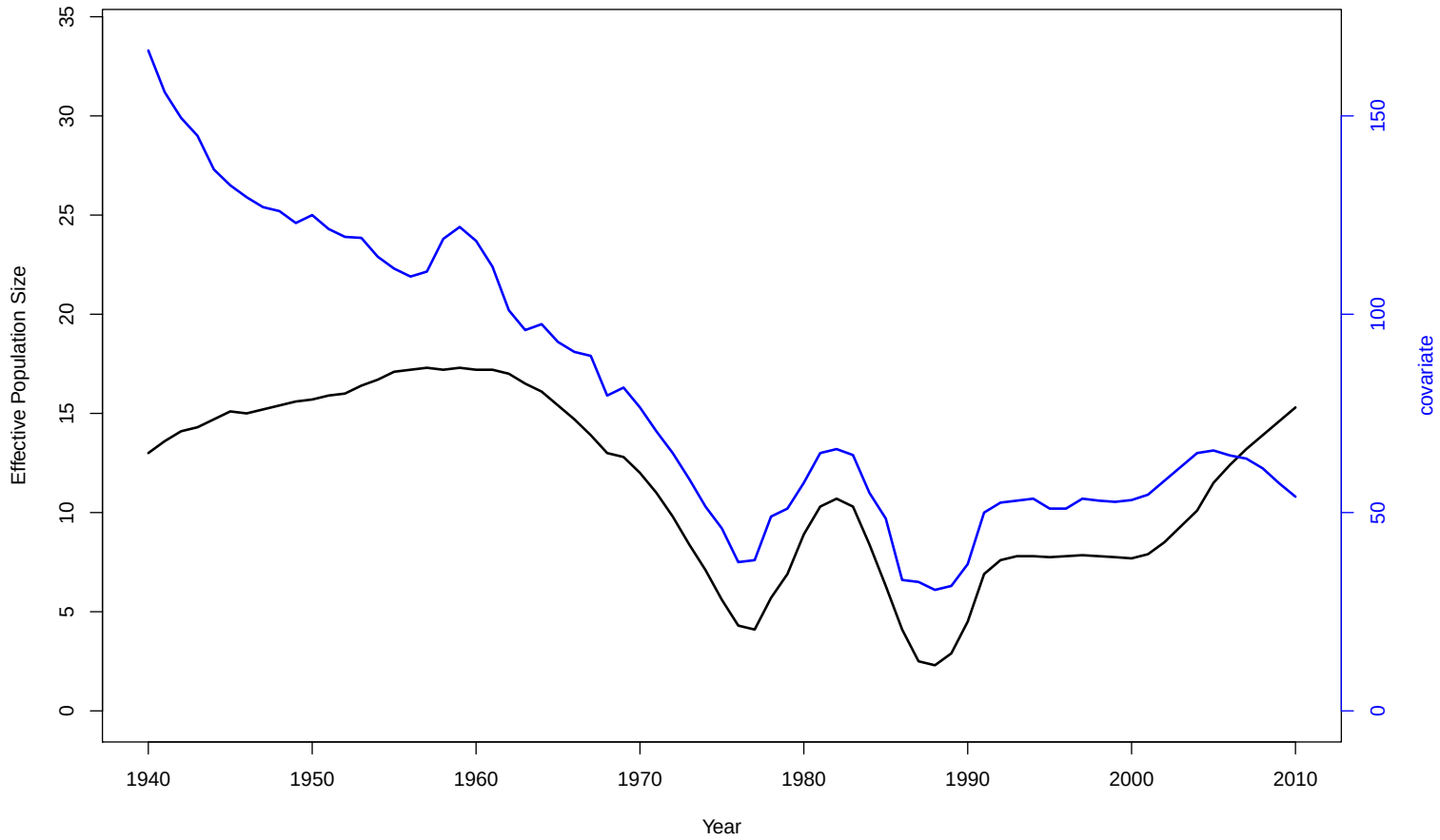
  Limitation: ignores uncertainty in demographic reconstructions

  Missed opportunity: does not make use of covariates to inform effective population size estimates

Drawbacks of Post Hoc Approach

Ignoring uncertainty in effective population size trajectory can:

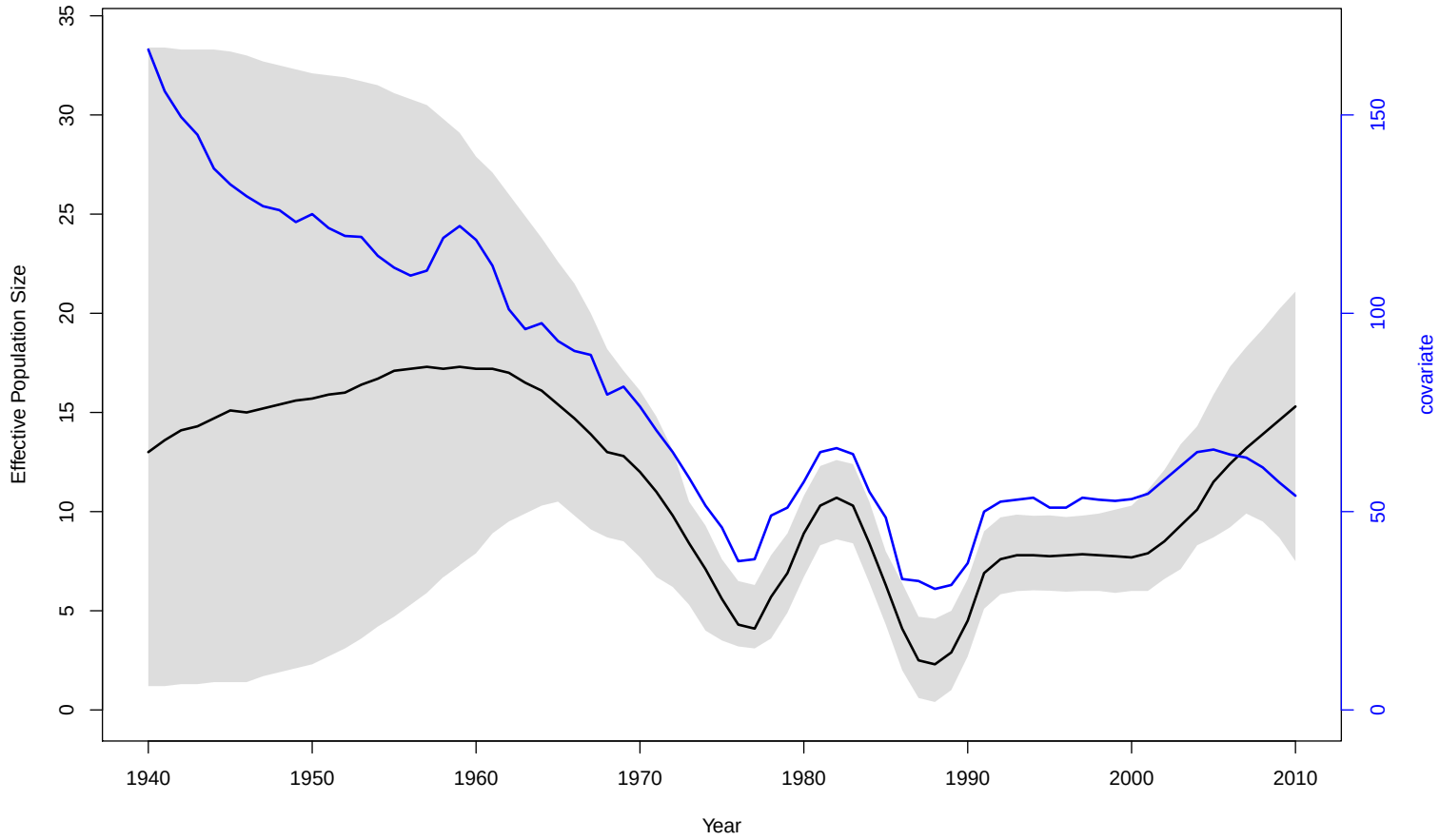
- Inflate precision of effect size coefficient estimates
- Erroneously rule out significant associations



Drawbacks of Post Hoc Approach

Ignoring uncertainty in effective population size trajectory can:

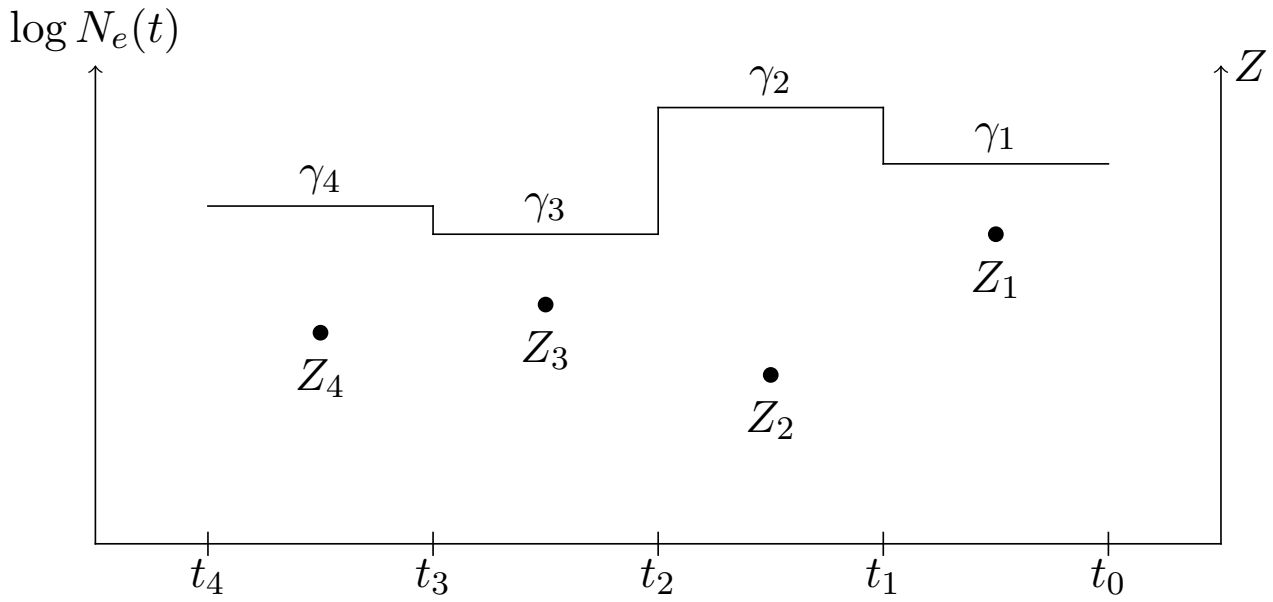
- Inflate precision of effect size coefficient estimates
- Erroneously rule out significant associations



Incorporating Covariates in the Skygrid

Alternative to post-hoc comparisons: Skygrid extension incorporate covariates in model and jointly infer effective population size γ and covariate effect size coefficients β

- Accounts for effective population size uncertainty
- Covariates may provide information about pop. dynamics and improve estimates



- $\log N_e(t) = \gamma_k = \beta_1 Z_{k1} + \cdots + \beta_P Z_{kP} + w_k$
- Model **w** as GMRF to enforce temporal dependence between adjacent intervals

Incorporating Covariates

We incorporate covariates in the model through a GMRF prior:

$$P(\boldsymbol{\gamma}|\mathbf{Z},\boldsymbol{\beta},\tau) \propto \rho^{(K+1)/2} \exp\left[-\frac{\rho}{2}(\boldsymbol{\gamma}-\mathbf{Z}\boldsymbol{\beta})'\mathbf{Q}(\boldsymbol{\gamma}-\mathbf{Z}\boldsymbol{\beta})\right]$$

$\mathbf{Q}$ is $(K+1)\times(K+1)$ tri-diagonal matrix with off-diagonal elements equal to -1, $Q_{11} = Q_{K+1,K+1} = 1$, and $Q_{ii} = 2$ for $i = 2,\ldots,K$

The individual components of $\boldsymbol{\gamma}$ have full conditional distributions

$$\begin{aligned}
\gamma_1|\boldsymbol{\gamma}_{-1} &\sim N\left(\mathbf{Z}_1'\boldsymbol{\beta} - \mathbf{Z}_2'\boldsymbol{\beta} + \gamma_2, \frac{1}{\rho}\right), \\
\gamma_i|\boldsymbol{\gamma}_{-i} &\sim N\left(\mathbf{Z}_i'\boldsymbol{\beta} + \frac{\gamma_{i-1}+\gamma_{i+1} - \mathbf{Z}_{i-1}'\boldsymbol{\beta} - \mathbf{Z}_{i+1}'\boldsymbol{\beta}}{2}, \frac{1}{2\rho}\right) \\
&\qquad \text{for } i = 2,\ldots,M, \\
\gamma_{M+1}|\boldsymbol{\gamma}_{-(M+1)} &\sim N\left(\mathbf{Z}_{M+1}'\boldsymbol{\beta} - \mathbf{Z}_M'\boldsymbol{\beta} + \gamma_M, \frac{1}{\rho}\right)
\end{aligned}$$

Example: West Nile Virus

Dellicour et al. (2020) examine the dynamics of the West Nile virus in North America by analyzing a data set of 801 genomic sequences sampled between 1999-2016. They consider temperature of the affected area as a covariate.

We expect a significant relationship based on what we know about the West Nile virus and its hosts:

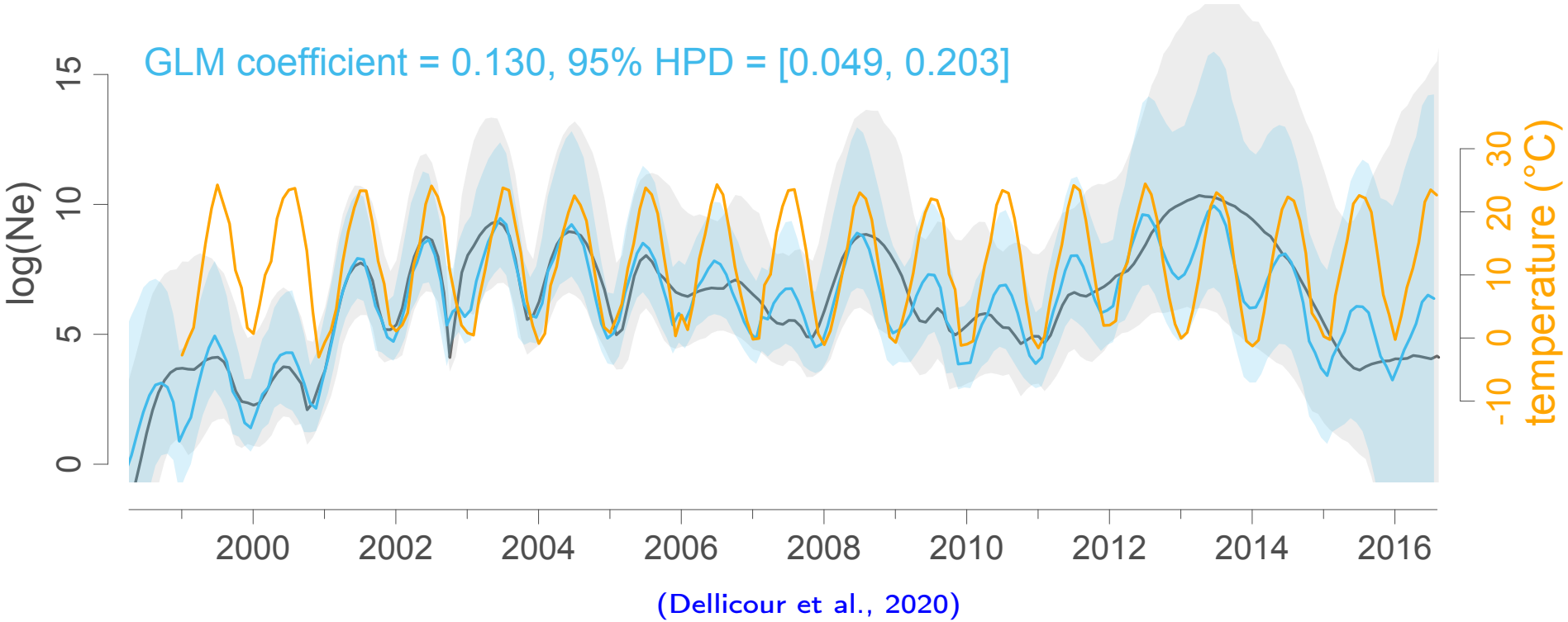
- For instance, mosquitos are one of the main hosts for the West Nile virus, and higher temperatures have been shown to directly impact the mosquito life cycle by accelerating larval development, decreasing the interval between blood meals, and prolonging the mosquito breeding season.

https://www.cdc.gov/west-nile-virus/

Example: West Nile Virus

We consider temperature of the affected area as a covariate (shown in orange)

We can then reconstruct the effective population size with and without the temperature of the affected area (shown in orange) as a covariate



(Dellicour et al., 2020)

Estimated mean trajectories and 95% Bayesian credibility interval regions (shaded)

- Black/gray - inference based strictly on sequence data
- Blue - inference based on sequence and covariate data

Accounting for Preferential Sampling

Standard coalescent-based priors do not account for the relationship between sequence sampling times and the effective population size

- However, sequence samples may be collected more frequently when pathogen circulation is high and less frequently when it is low
- The standard implicit assumption of no relationship between sampling times and population dynamics can lead to biased inferences

Coalescent-based priors $P(\tau|\boldsymbol{\gamma})$ have been constructed by implicitly conditioning on the sampling times $\mathbf{s}$ (and should really be written $P(\tau|\mathbf{s}, \boldsymbol{\gamma})$).

- What if we consider the distribution of $\mathbf{s}$?

Accounting for Preferential Sampling

Karcher et al. (2016) introduced a framework that models sequence sampling times as an inhomogeneous Poisson process with intensity $\lambda(t)$.

- The number of samples in a time interval $[t_0, t_1]$ follows a Poisson distribution with mean $\int_{t_0}^{t_1} \lambda(t)dt$
- To model possible dependence of the sampling intensity on the effective population size, they posit

$$\log \lambda(t) = \beta_0 + \beta_1 \gamma(t),$$

where $\lambda(t)$ is assumed to be piecewise constant, and $\gamma(t)$ is the piecewise constant log effective population size.

Accounting for Preferential Sampling

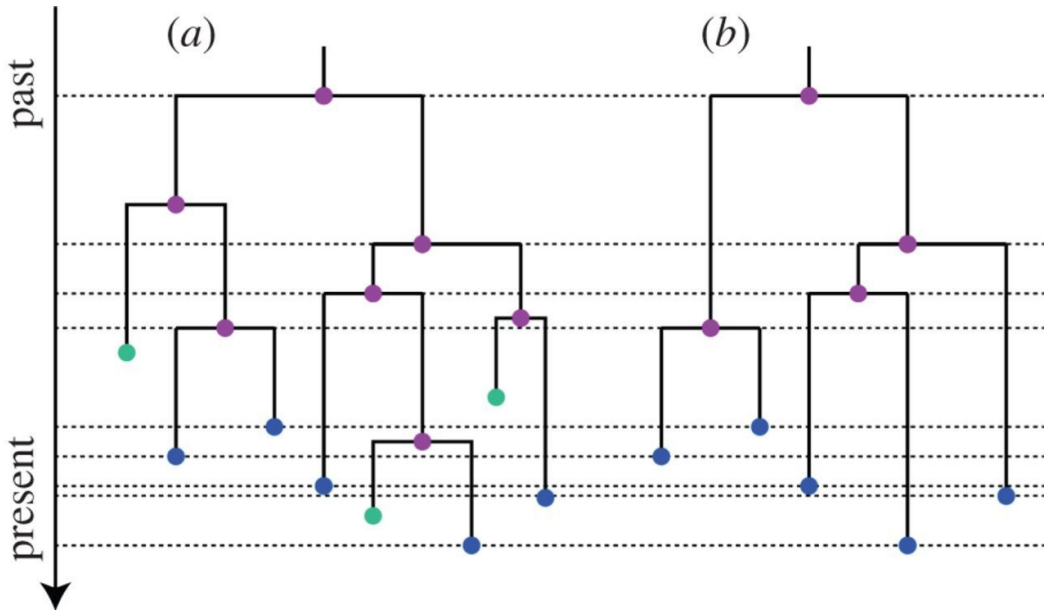
However, the relationship between the sampling intensity and effective population size could change over time and could depend on other factors

- There could be a decrease in sequencing costs over time
- In the context of endemic infectious disease surveillance, it is likely important to account for seasonality

Karcher et al. (2020) extended their framework to allow for the sampling intensity $\lambda(t)$ to depend on time-varying covariates $f_2(t), \ldots, f_m(t)$ and their interactions with the log effective population size $\gamma(t)$:

$$\log \lambda(t) = \beta_0 + \beta_1 \gamma(t) + \beta_2 f_2(t) + \cdots + \beta_m f_m(t) + [\delta_2 f_2(t) + \cdots + \delta_m f_m(t)]\gamma(t)$$

Birth-Death Processes

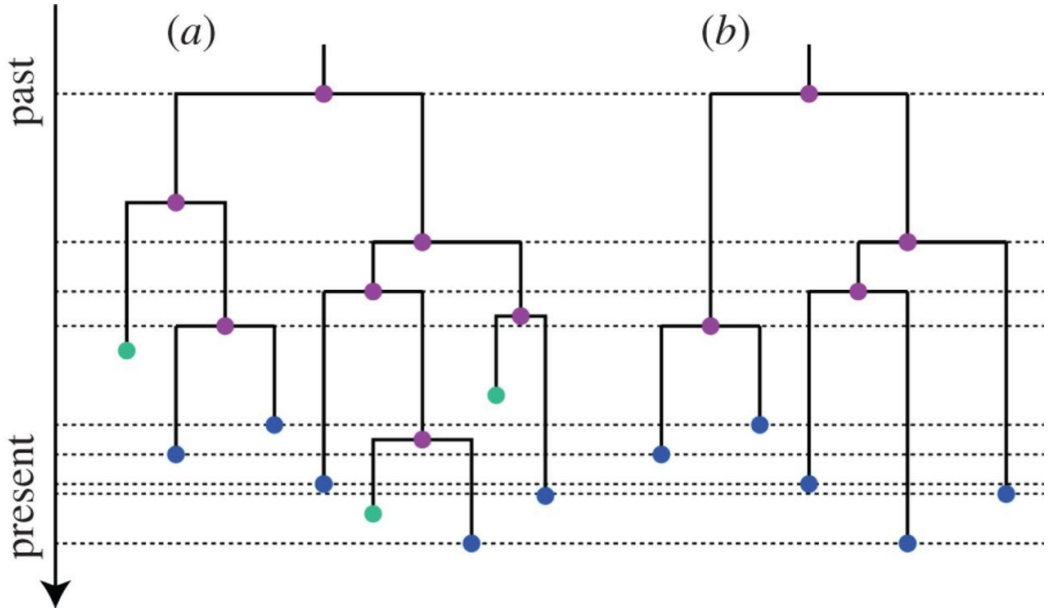


(Kühnert et al., 2014) Here, a) depicts a full transmission tree with birth nodes (purple), sampling events (blue) and death events (green), and b) depicts the tree pruned to include only observed individuals.

As an alternative to the coalescent, phylogenetic tree priors can be specified via birth-death processes (Rannala and Yang, 1996; Stadler, 2009; Stadler et al., 2013; Heath et al., 2014; Kühnert et al., 2014) that can be characterized by the following parameters:

- λ - speciation\transmission\birth rate
- μ - extinction\death\recovery rate
- ψ - fossil observation rate or sampling rate for individual

Birth-Death Processes



(Kühnert et al., 2014) Here, a) depicts a full transmission tree with birth nodes (purple), sampling events (blue) and death events (green), and b) depicts the tree pruned to include only observed individuals.

In the context of infectious disease modeling, a "birth" corresponds to the infection of an individual, and a "death" corresponds to an individual becoming noninfectious (which can be due to treatment, change in behavior, or death).

We can think of an infected individual that transmits at rate λ and becomes noninfectious with rate μ. There is also the possibility of being "sampled" with rate ψ, and this is usually linked to treatment or behavioral changes, so we can categorize an individual as noninfectious immediately after they are sampled.

Birth-Death Processes

Birth-death model parameters can be used to estimate key epidemiological parameters.

For example, Stadler et al. (2012) propose estimating the basic reproductive number R_0 by the ratio

$$R_0 = \frac{\lambda}{\mu + \psi}.$$

- Here, the duration of infection, $1/\delta$, is determined by the total rate of becoming noninfectious $\delta = \mu + \psi$.

Kühnert et al. (2014) link a birth-death tree prior to a stochastic compartmental susceptible-infected-removed (SIR) model

- Enables joint inference of phylogenetic tree and SIR trajectories

Birth-Death SIR Model

Under the Kühnert et al. (2014) birth-death SIR framework,

- An infected individual infects a susceptible individual with rate β and recovers with rate γ
- Some recoveries are observed (or sampled) whereas other recoveries are hidden/unobserved. The probability of a recovery being observed is represented by the sampling proportion s
- Can simulate the numbers of individuals in each compartment $n_S(t), n_I(t), n_R(t)$ at time points $t = t_1, \ldots, t_m$

To connect the compartmental SIR model to the birth-death phylogenetic tree prior:

- Assume time-varying birth rates λ_i for interval $[t_i, t_{i+1})$ and set $\lambda_i = \beta n_S(t_i)$
- Set sampling rate $\psi = s\gamma$
- Set death rate $\mu = (1 - s)\gamma$