

Simulation of Markov genealogy processes

Aaron A. King

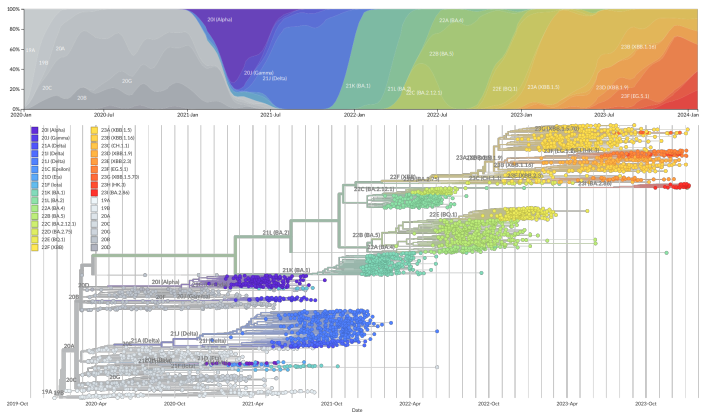
February 28, 2024

Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

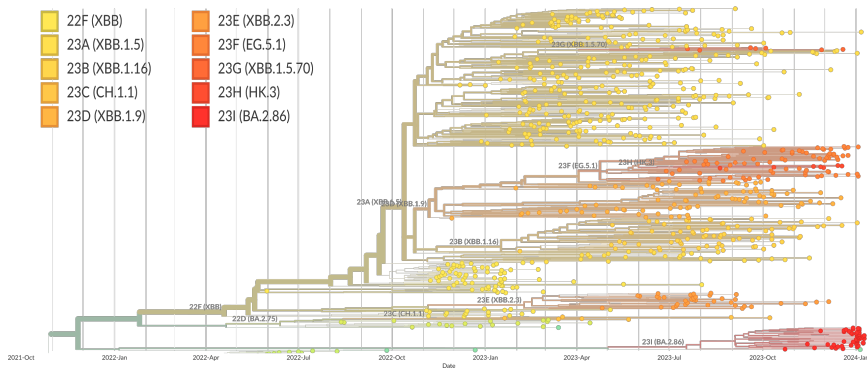
Objectives

Example: surveillance for emerging SARS-CoV-2 variants



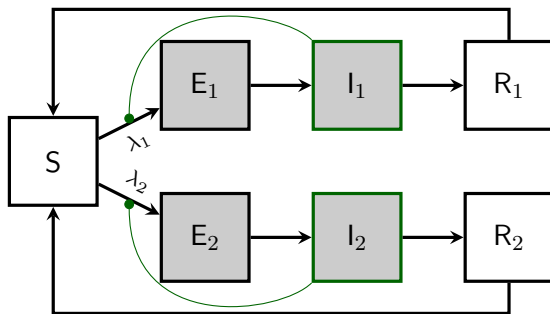
nextstrain.org (Hadfield *et al.*, 2018)

Example: surveillance for emerging SARS-CoV-2 variants



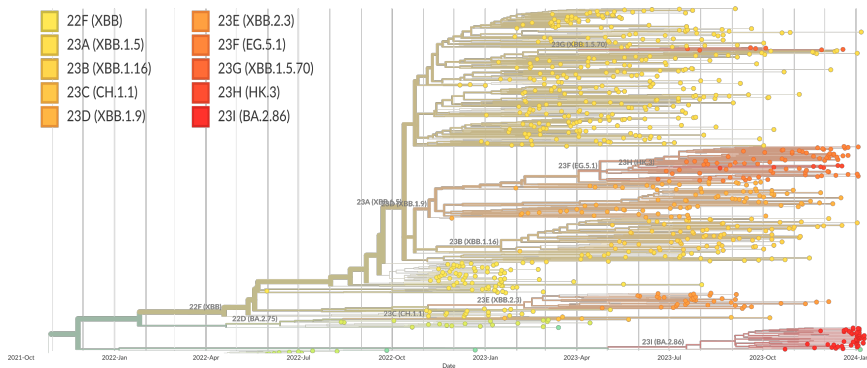
nextstrain.org (Hadfield *et al.*, 2018)

Example: surveillance for emerging SARS-CoV-2 variants



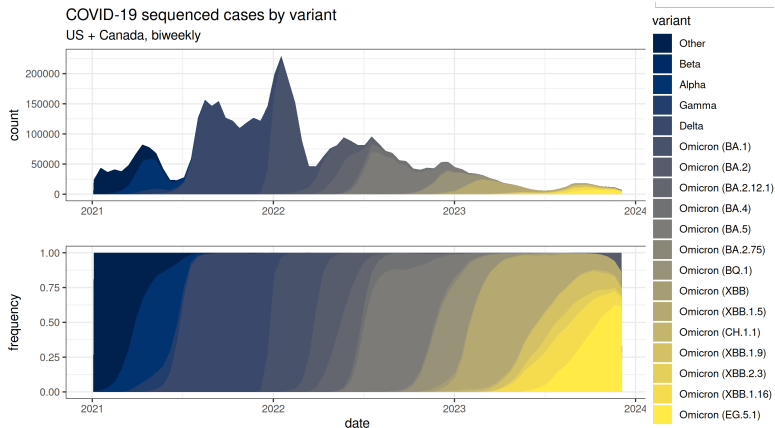
$$\lambda_1 = \beta_1 \frac{I_1}{N} \quad \lambda_2 = \beta_2 \frac{I_2}{N}$$

Example: surveillance for emerging SARS-CoV-2 variants



nextstrain.org (Hadfield *et al.*, 2018)

Example: surveillance for emerging SARS-CoV-2 variants



(Mathieu *et al.*, 2020)

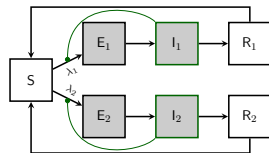
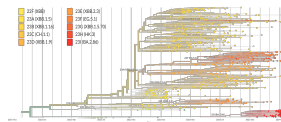
What is phylodynamics?

Broadly:

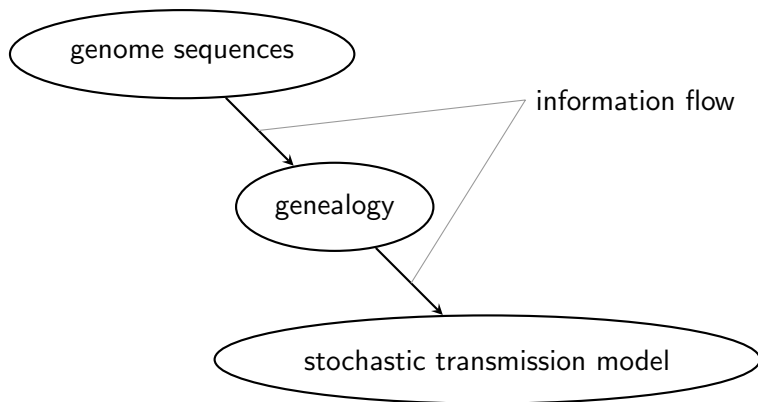
Phylodynamics is the project of inferring
determinants of epidemic spread
 using
genomic data from pathogen samples.

In this talk:

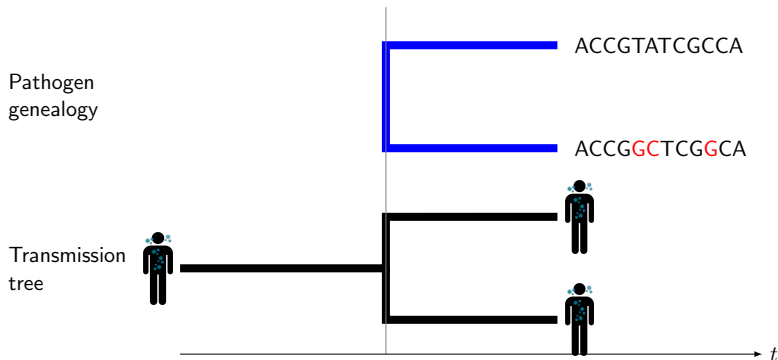
Phylodynamics means using
genomic data
 to infer
stochastic dynamic transmission models.



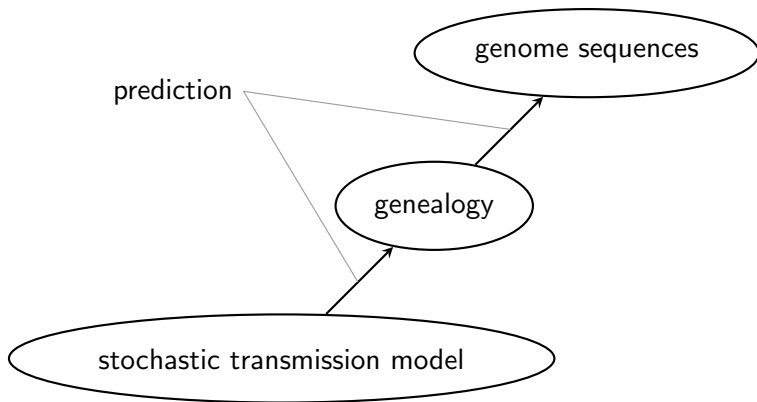
Core problems of phylodynamics



Core problems of phylodynamics



Core problems of phylodynamics



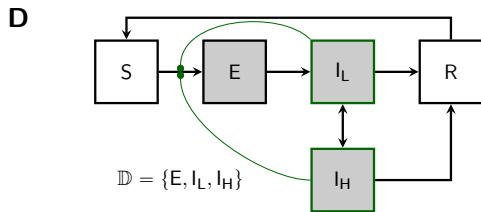
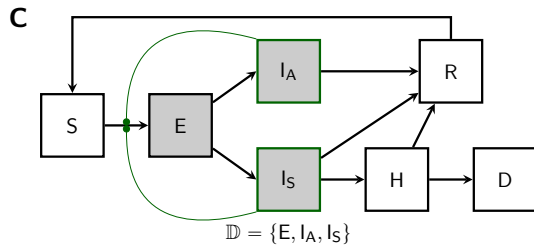
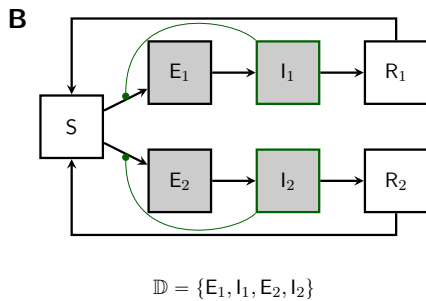
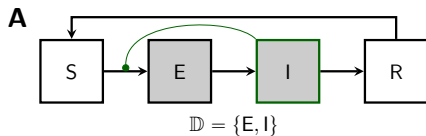
Overview

- We show how a given population process induces a unique genealogy process.
- *Pruning* and *obscuration* project a genealogy onto observable data.
- We derive the exact likelihood as the solution to a nonlinear filtering problem
- This equation can be solved by standard Monte Carlo methods.

Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

Population process



Population process

- Non-explosive Markov jump process, $\mathbf{X}_t \in \mathbb{X}$, $t \in \mathbb{R}_+$:
the *population process*.
- Initial-state distribution, p_0 :

$$\text{Prob}[\mathbf{X}_0 \in \mathcal{E}] = \int_{\mathcal{E}} p_0(x) \, dx$$

- Jump rates: $\alpha(t, x, x') = \text{rate of jump } x \rightarrow x'$

$$\alpha(t, x, x') \geq 0, \quad \int_{\mathbb{X}} \alpha(t, x, x') \, dx' < \infty$$

- Multiple events at each jump are allowed.

Population process

Kolmogorov forward equation (KFE):

If

$$\frac{\partial w}{\partial t}(t, x) = \int w(t, x') \alpha(t, x', x) \, dx' - \int w(t, x) \alpha(t, x, x') \, dx'$$

and

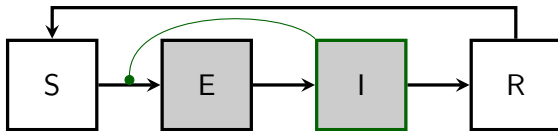
$$w(0, x) = p_0(x)$$

then

$$\int_{\mathcal{E}} w(t, x) \, dx = \text{Prob} [\mathbf{X}_t \in \mathcal{E}] .$$

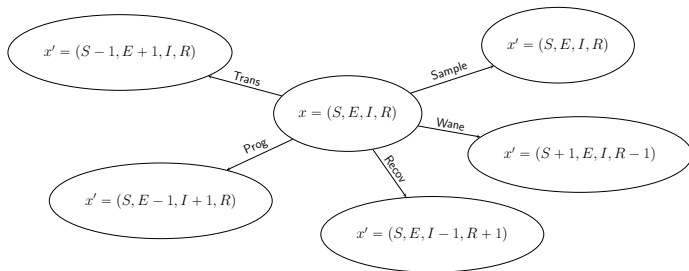
KFE is sometimes called the *master equation* for \mathbf{X}_t .

Population process



$$\frac{\partial w}{\partial t}(t, x) = \int w(t, x') \alpha(t, x', x) dx' - \int w(t, x) \alpha(t, x, x') dx'$$

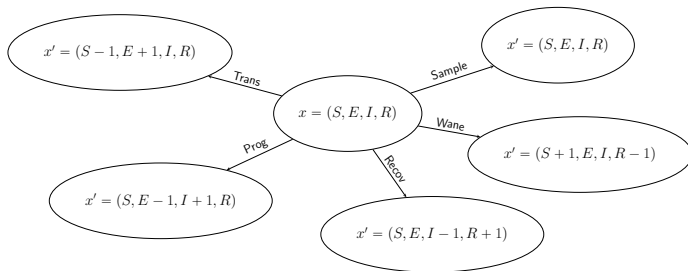
Population process



$$\mathbb{U} = \{\text{Trans, Prog, Recov, Wane, Sample}\}$$

$$\frac{\partial w}{\partial t}(t, x) = \sum_{u \in \mathbb{U}} \left\{ \int w(t, x') \alpha_u(t, x', x) dx' - \int w(t, x) \alpha_u(t, x, x') dx' \right\}$$

Population process



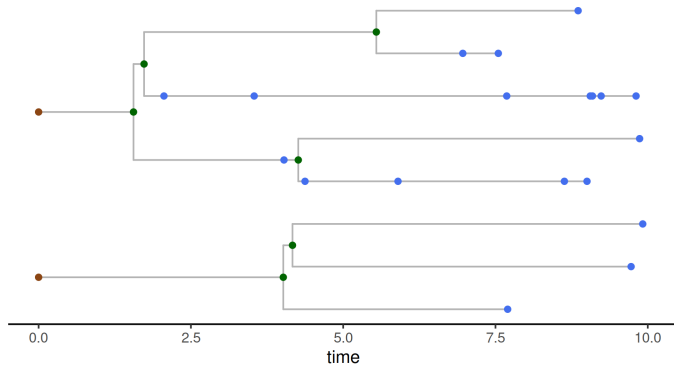
$$\mathbb{U} = \{\text{Trans, Prog, Recov, Wane, Sample}\}$$

$$\begin{aligned} \frac{\partial w}{\partial t}(t, S, E, I, R) = & \frac{\beta(t)(S+1)I}{N} w(t, S+1, E-1, I, R) - \frac{\beta(t)SI}{N} w(t, S, E, I, R) + \sigma(E+1) w(t, S, E+1, I-1, R) - \sigma E w(t, S, E, I, R) \\ & + \gamma(I+1) w(t, S, E, I+1, R-1) - \gamma I w(t, S, E, I, R) + \omega(R+1) w(t, S-1, E, I, R+1) - \omega R w(t, S, E, I, R) \end{aligned}$$

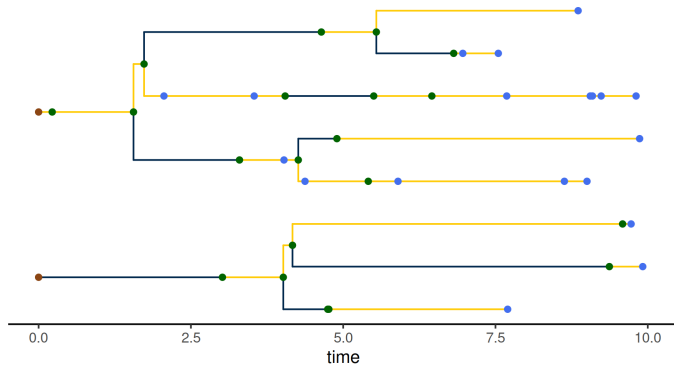
Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

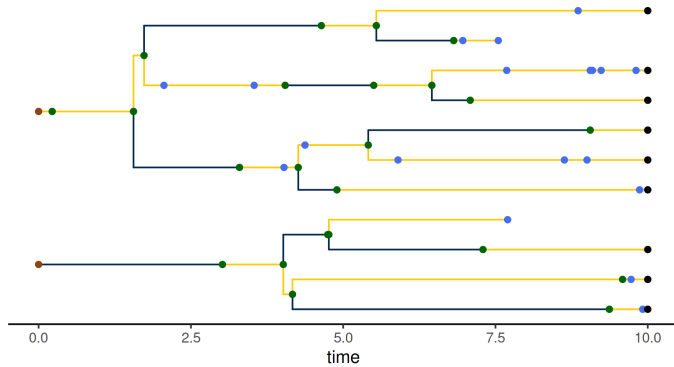
What is a genealogy?



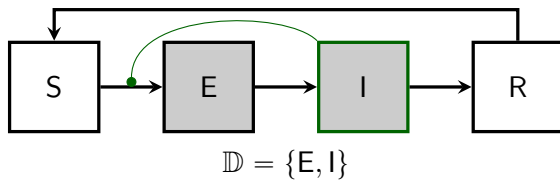
What is a genealogy?



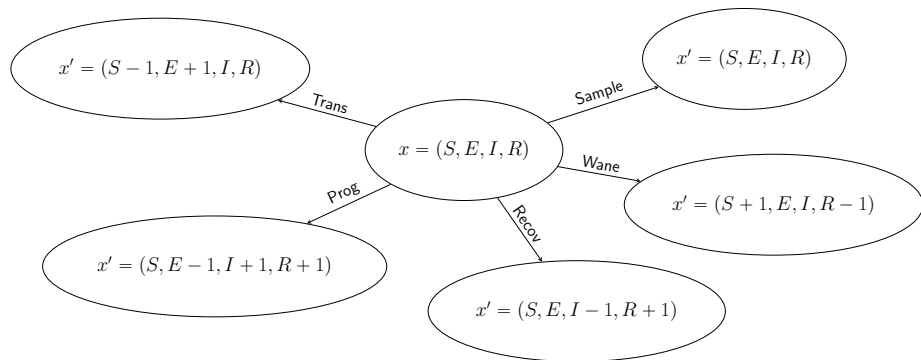
What is a genealogy?



Event types



Event types



$$\mathbb{U} = \{\text{Trans}, \text{Prog}, \text{Recov}, \text{Wane}, \text{Sample}\}$$

Event types

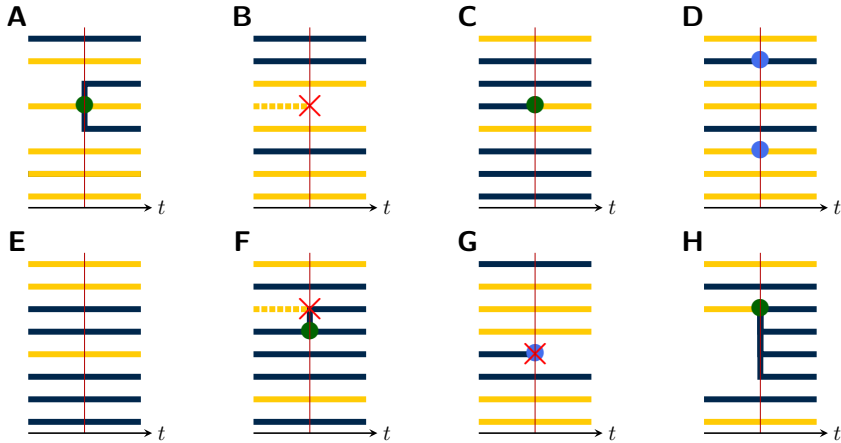
If we write

$$\alpha(t, x, x') = \sum_{u \in \mathbb{U}} \alpha_u(t, x, x'),$$

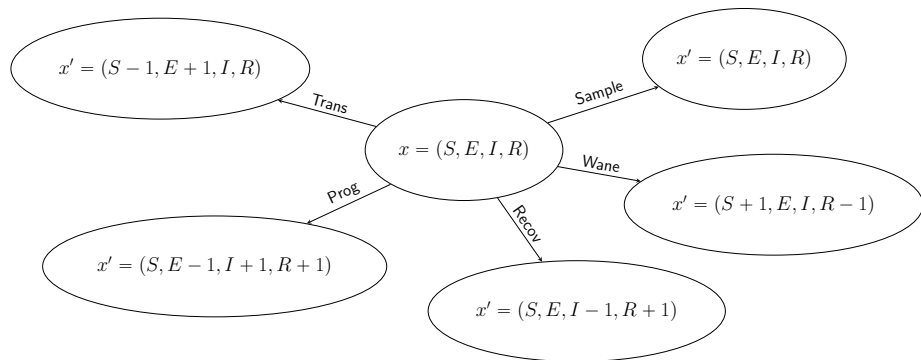
the KFE becomes

$$\frac{\partial w}{\partial t}(t, x) = \sum_u \int w(t, x') \alpha_u(t, x', x) \, dx' - \sum_u \int w(t, x) \alpha_u(t, x, x') \, dx'$$

Event types



Event types



$$\mathbb{U} = \{\text{Trans}, \text{Prog}, \text{Recov}, \text{Wane}, \text{Sample}\}$$

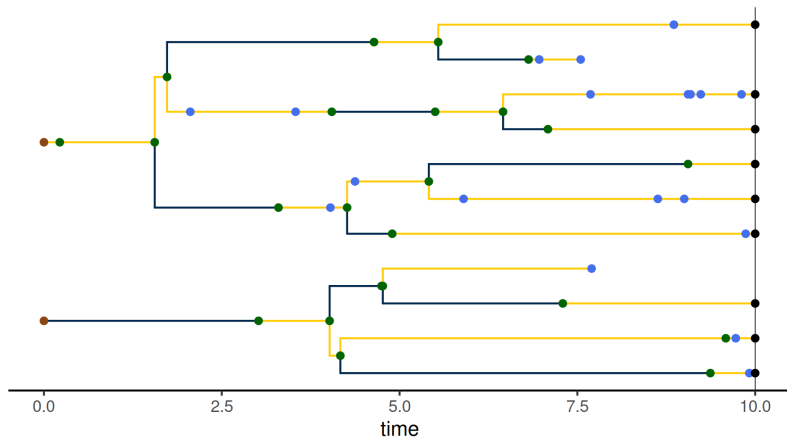
A population process induces a genealogy process

- G_t is a stochastic process on the space of genealogies.
- The map $\mathbf{X} \mapsto \mathbf{G}$ is random.
- **Key assumption:** Lineages within a deme are *exchangeable*.
There is no more structure than is implied by the population process.
- Simulation code on github.com/kingaa/phylopomp
- Animations at <https://kingaa.github.io/manuals/phylopomp/vignettes/>

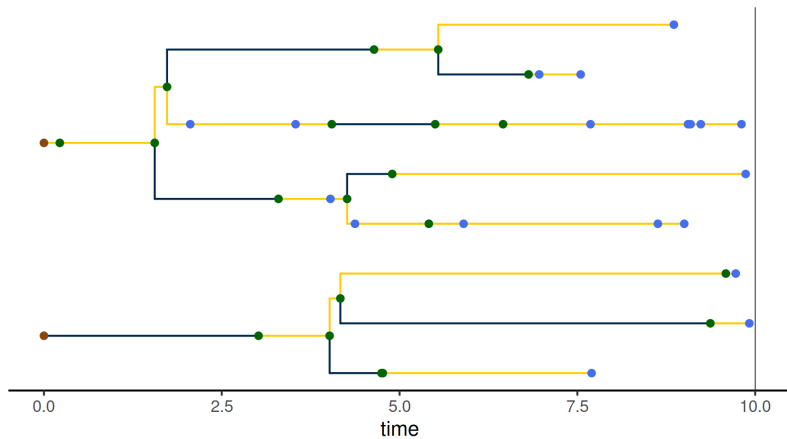
Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

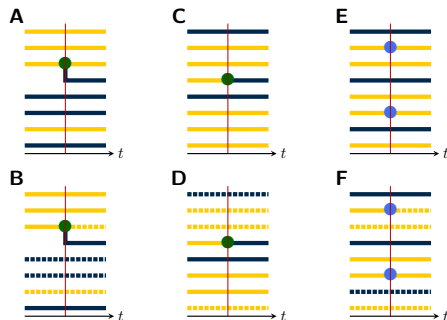
Full genealogy



Pruned genealogy

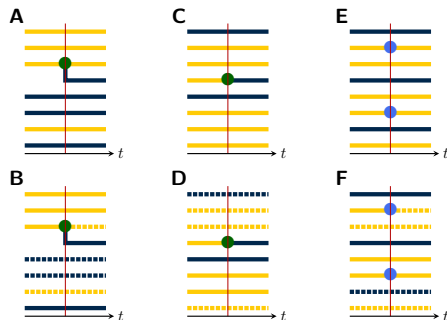


Local structure of a pruned genealogy



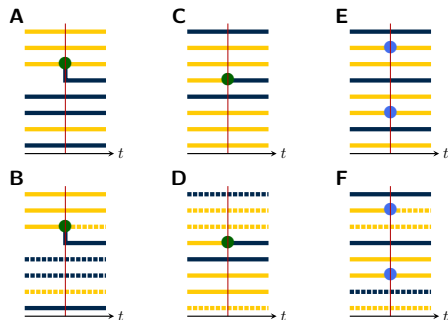
Top row shows the *unpruned genealogy* in neighborhood of an event.
 Bottom row shows the corresponding *pruned genealogy*.

Local structure of a pruned genealogy



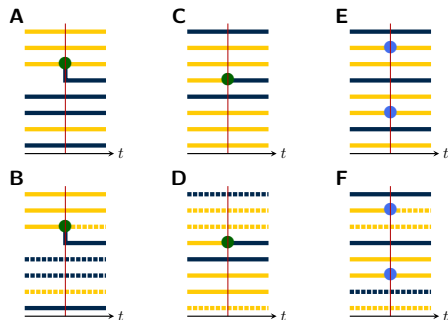
For $x \in \mathbb{X}$, $i \in \mathbb{D}$, $n_i(x)$ is the *occupancy* of deme i when the system is in state x .
 In panel A $n = (n_{\text{blue}}, n_{\text{yellow}}) = (4, 4)$; in panel C $n = (3, 5)$;

Local structure of a pruned genealogy



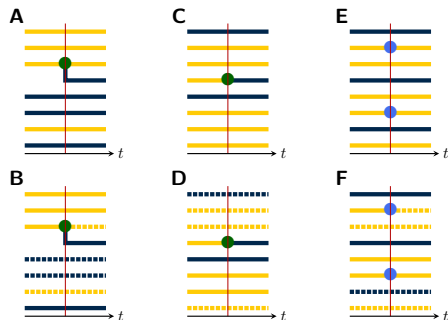
For $u \in \mathbb{U}$, $i \in \mathbb{D}$, r_i^u is the *production* of event u in deme i .
 In panel A, $r = (r_{\text{blue}}, r_{\text{yellow}}) = (1, 1)$; in panel E, $r = (0, 2)$.

Local structure of a pruned genealogy



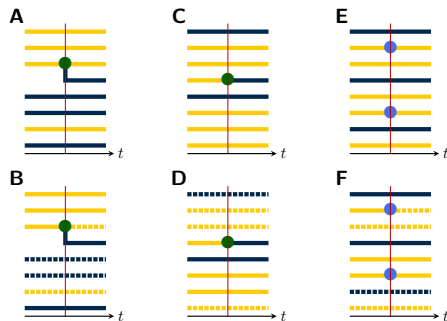
The *lineage count*, $\ell_i(t)$, is the number of unpruned lineages in deme i at time t .
In this case, for all panels, $\ell = (2, 2)$.

Local structure of a pruned genealogy



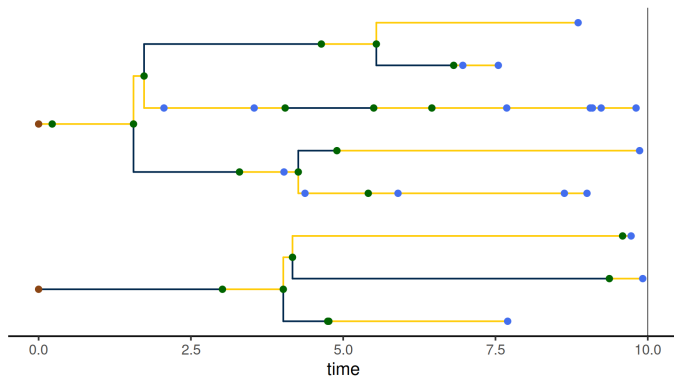
The *saturation*, s_i , is the number of unpruned lineages in deme i *descending* from the event. In panels B and D, $s = (1, 0)$; in panel F, $s = (0, 1)$.

Local structure of a pruned genealogy



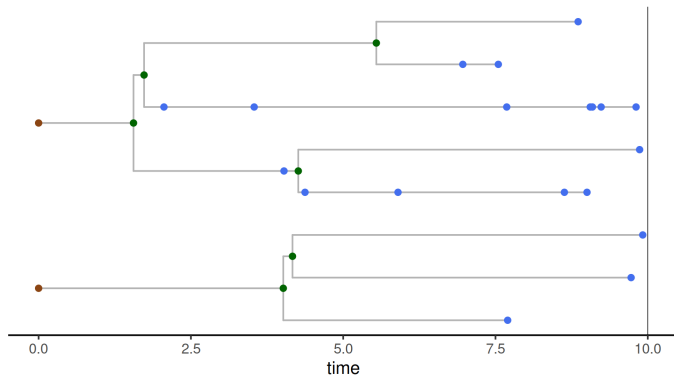
Obviously, $s_i \leq r_i \leq n_i$ and $s_i \leq \ell_i \leq n_i$.

Pruned genealogy



A pruned genealogy is specified by two functions of time, (Y, Z) :
 Z_t gives the local topological structure; Y_t gives the local coloring.

Obscured genealogy



An obscured genealogy is specified by (T, Z) .

Binomial ratio

For $n, r, \ell, s \in \mathbb{Z}_+^{\mathbb{D}}$, define the *binomial ratio*

$$\binom{n \quad \ell}{r \quad s} := \begin{cases} \prod_{i \in \mathbb{D}} \frac{\binom{n_i - \ell_i}{r_i - s_i}}{\binom{n_i}{r_i}}, & \text{if } \forall i \ n_i \geq \{\ell_i, r_i\} \geq s_i \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

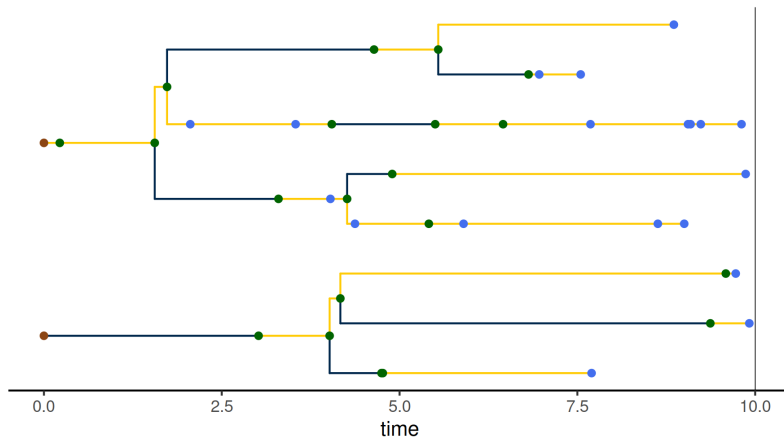
Observe that $\binom{n \quad \ell}{r \quad s} \in [0, 1]$. Moreover,

$$\sum_{s \in \mathbb{Z}_+^{\mathbb{D}}} \binom{n \quad \ell}{r \quad s} \binom{\ell}{s} = 1.$$

Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

Theorem: likelihood of a pruned genealogy



Theorem: likelihood of a pruned genealogy

Suppose that $P = (Y, Z)$ is a given pruned genealogy with depth T .

Define

$$\phi_u(x, y, y') := \begin{pmatrix} n(x) & \ell(y') \\ r^u & s(y, y') \end{pmatrix} Q_u(y, y').$$

Here, $Q = 1$ if the local structure of P is compatible with an event of type u at that time;
 $Q = 0$ otherwise.

Theorem: likelihood of a pruned genealogy

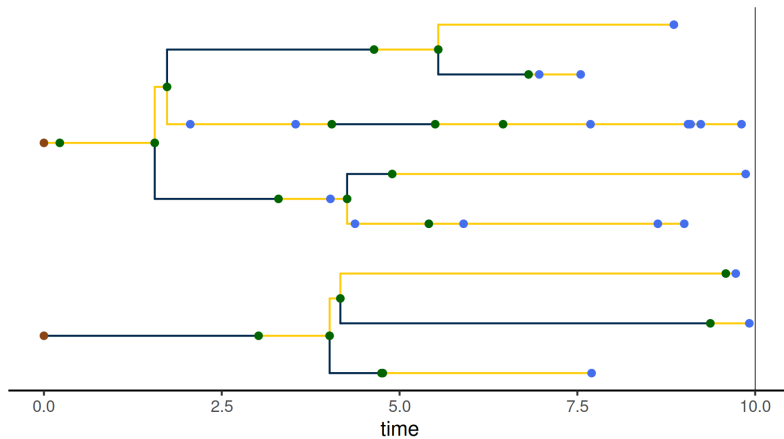
If $w = w(t, x)$ satisfies the initial condition $w(0, x) = p_0(x)$ and the filter equation

$$\begin{aligned} \frac{\partial w}{\partial t}(t, x) = & \sum_u \int w(t, x') \alpha_u(t, x', x) \phi_u(x, \tilde{Y}_t, Y_t) dx' - \sum_u \int w(t, x) \alpha_u(t, x, x') dx' \\ & + \sum_{e \in \text{ev}(\mathbf{P})} \delta(t, e) \left\{ \sum_u \int w(t, x') \alpha_u(t, x', x) \phi_u(x, \tilde{Y}_t, Y_t) dx' - w(t, x) \right\}, \end{aligned}$$

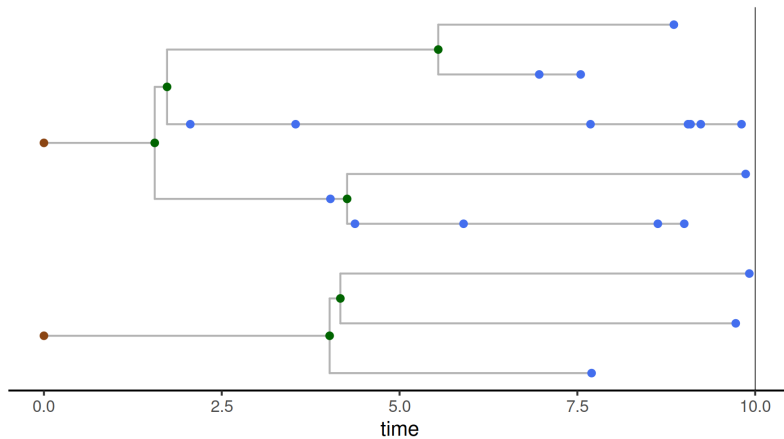
then the likelihood of \mathbf{P} is

$$\mathcal{L}(\mathbf{P}) = \int w(T, x) dx.$$

Theorem: likelihood of a pruned genealogy



Theorem: likelihood of an obscured genealogy



Theorem: likelihood of an obscured genealogy

Let (T, Z) , be a given obscured genealogy. Then there are probability kernels π and q such that if

$$\beta_u(t, x, x', y, y') = \alpha_u(t, x, x') \pi_u(t, x, x', y, y'), \quad \psi_u(t, x, x', y, y') = \frac{\phi_u(x', y, y')}{\pi_u(t, x, x', y, y')},$$

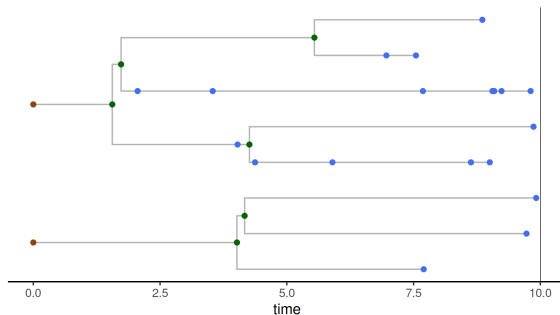
and if $w = w(t, x, y)$ satisfies the initial condition $w(0, x, y) = p_0(x) q(x, y)$ and the filter equation

$$\begin{aligned} \frac{\partial w}{\partial t} = & \sum_{uy'} \int w(t, x', y') \beta_u(t, x', x, y', y) \psi_u(t, x', x, y', y) dx' - \sum_{uy'} \int w(t, x, y) \beta_u(t, x, x', y, y') dx' \\ & + \sum_{e \in \text{ev}(Z)} \delta(t, e) \left\{ \sum_{uy'} \int w(t, x', y') \beta_u(t, x', x, y, y') \psi_u(t, x', x, y', y) dx' - w(t, x, y) \right\}, \end{aligned}$$

Theorem: likelihood of an obscured genealogy

then the likelihood of Z is

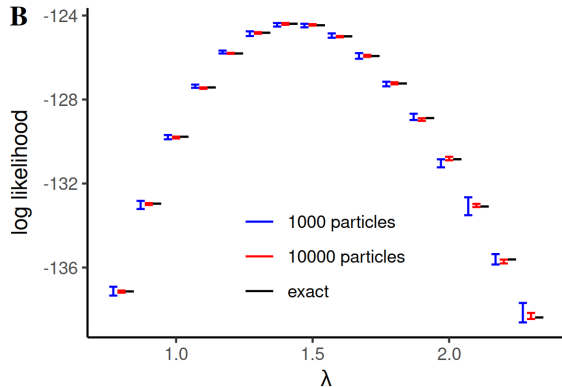
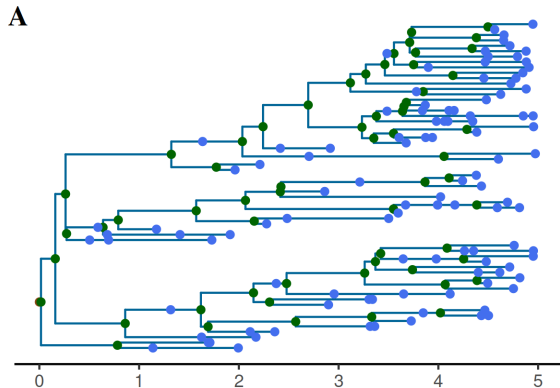
$$\mathcal{L}(Z) = \sum_y \int w(T, x, y) dx.$$



Outline

- 1 Context
 - Example: emerging variants
 - Phylodynamics
 - Problems of phylodynamics
- 2 Population process
 - Examples
 - Formalization
- 3 Genealogy process
 - Genealogies
 - Induced genealogy process
- 4 Pruned and obscured genealogies
- 5 Theorems
 - Pruned genealogies
 - Obscured genealogies
- 6 Examples

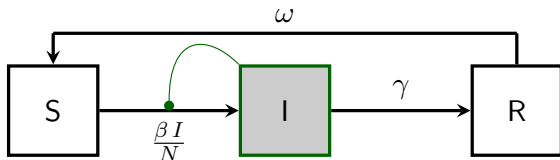
Linear birth-death model



Uniform sampling.

Exact likelihood is available in closed form.

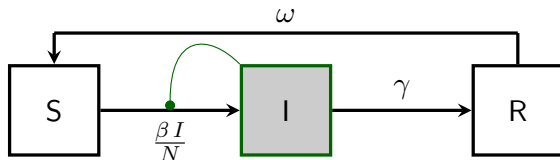
SIRS model



Between genealogical events:

$$\begin{aligned} \frac{\partial w}{\partial t} = & \frac{\beta (S+1)(I-1)}{N} \left(1 - \frac{\binom{\ell}{2}}{\binom{I}{2}} \right) w(t, S+1, I-1, R) + \gamma (I+1) w(t, S, I+1, R-1) \\ & + \omega (R+1) w(t, S-1, I, R+1) - \left(\frac{\beta S I}{N} + \gamma I + \omega R + \psi I \right) w(t, S, I, R). \end{aligned}$$

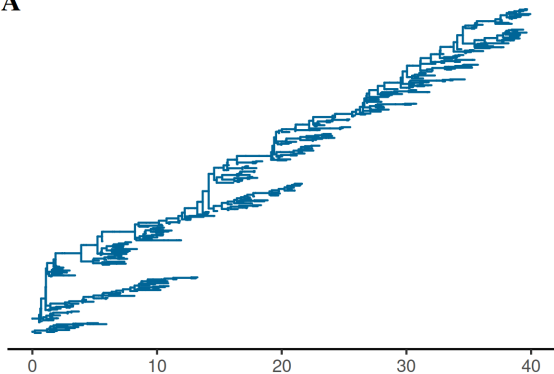
SIRS model



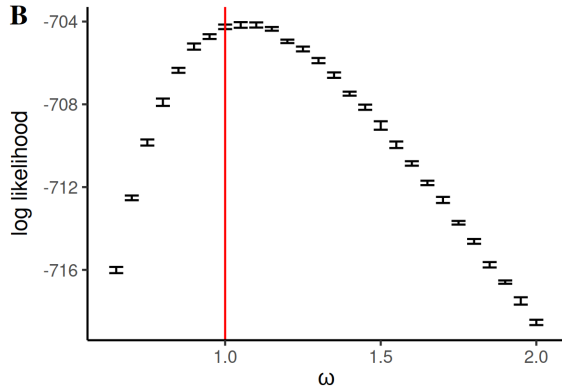
At genealogical events:

$$w(t, S, I, R) = \begin{cases} \frac{2\beta(S+1)}{IN} \tilde{w}(t, S+1, I-1, R), & \text{branch point at } t, \\ \psi \tilde{w}(t, S, I, R), & \text{internal sample at } t, \\ \psi (I - \ell) \tilde{w}(t, S, I, R), & \text{terminal sample at } t. \end{cases}$$

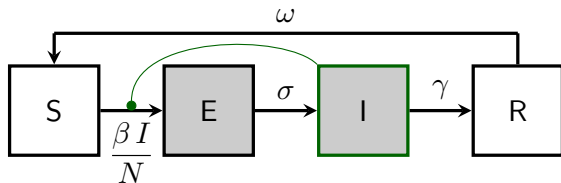
SIRS model

A

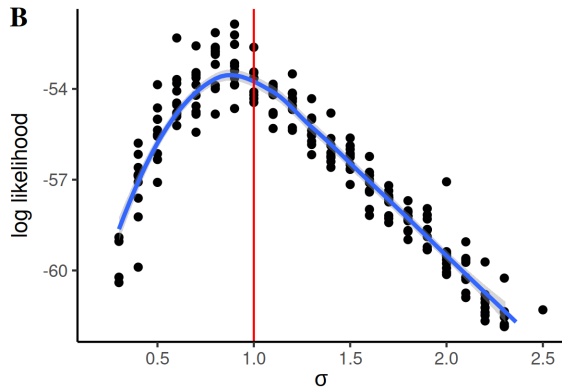
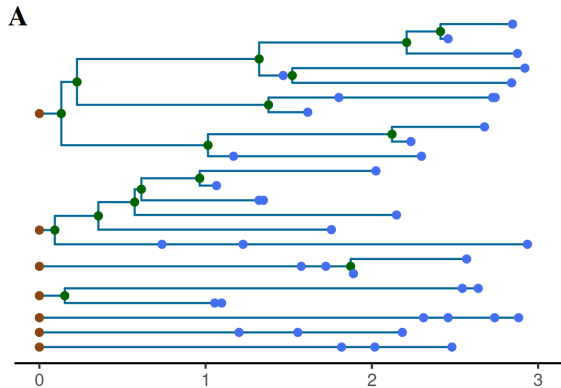
Uniform sampling.
One deme only.

B

SEIRS model



SEIRS model



Concluding remarks

- The theory *corrects* and *strictly extends* all existing likelihood-based phylodynamic methods (e.g., Volz *et al.*, 2009; Rasmussen *et al.*, 2011; Stadler, 2010; Volz, 2012; Volz and Frost, 2014; Rasmussen *et al.*, 2014; Vaughan *et al.*, 2019).
- It eliminates the need for large population-size and small sample-fraction assumptions, as well as any dependence on linearization.
- All computations can be carried out forward in time.
- This greatly expands the class of models that can be entertained.
- Great flexibility in sampling model
- Applications in disease ecology and beyond
- The unstructured case can be found in King *et al.* (2022).

Outstanding challenges

- Choice of importance-sampling kernel
- Borrowing information from future is allowed.
- Phylogenetic uncertainty
- Efficient algorithms
- Reassortment and recombination

Summary

- A discretely structured Markov population process uniquely induces a genealogy-valued Markov process.
- The likelihood of an observed genealogy satisfies a nonlinear filtering equation, which can be efficiently computed via Feynman-Kač (sequential Monte Carlo) algorithms.
- In principle, these results liberate us to entertain models that more closely match our biological questions, without less hindrance from inference methodology.

References

- Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA (2018). “Nextstrain: real-time tracking of pathogen evolution.” *Bioinformatics*, **34**(23), 4121–4123. doi: 10.1093/bioinformatics/bty407.
- King AA, Lin Q, Ionides EL (2022). “Markov genealogy processes.” *Theor Popul Biol*, **143**, 77–91. doi: 10.1016/j.tpb.2021.11.003.
- King AA, Nguyen D, Ionides EL (2016). “Statistical inference for partially observed Markov processes via the R package pomp.” *J Stat Softw*, **69**(12), 1–43. doi: 10.18637/jss.v069.i12.

References II

- Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M (2020). "Coronavirus pandemic (COVID-19)." *Our World in Data [Online resource]*. URL <https://ourworldindata.org/coronavirus>.
- Rasmussen DA, Ratmann O, Koelle K (2011). "Inference for nonlinear epidemiological models using genealogies and time series." *PLoS Comput. Biol.*, **7**(8), e1002136.
doi: 10.1371/journal.pcbi.1002136.
- Rasmussen DA, Volz EM, Koelle K (2014). "Phylogenetic inference for structured epidemiological models." *PLoS Comput. Biol.*, **10**(4), e1003570.
doi: 10.1371/journal.pcbi.1003570.


References III

- Stadler T (2010). "Sampling-through-time in birth-death trees." *J. Theor. Biol.*, **267**(3), 396–404. doi: 10.1016/j.jtbi.2010.09.010.
- Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T (2019). "Estimating epidemic incidence and prevalence from genomic data." *Mol. Biol. Evol.*, **36**, 1804–1816. doi: 10.1093/molbev/msz106.
- Volz EM (2012). "Complex population dynamics and the coalescent under neutrality." *Genetics*, **190**(1), 187–201. doi: 10.1534/genetics.111.134627.
- Volz EM, Frost SDW (2014). "Sampling through time and phylodynamic inference with coalescent and birth-death models." *J. R. Soc. Interface*, **11**(101), 20140945. doi: 10.1098/rsif.2014.0945.

References IV

Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009).
“Phylodynamics of infectious disease epidemics.” *Genetics*, **183**(4), 1421–1430.
doi: 10.1534/genetics.109.106021.

License, acknowledgments, and links

- The materials build on previous versions of this course and related courses.
- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin. 
- Produced with R version 4.3.2 and **pomp** version 5.6.
- Compiled on February 28, 2024.

[Back to Lesson](#)

[R codes for this lesson](#)