Introduction to Markov genealogy processes

Aaron A. King

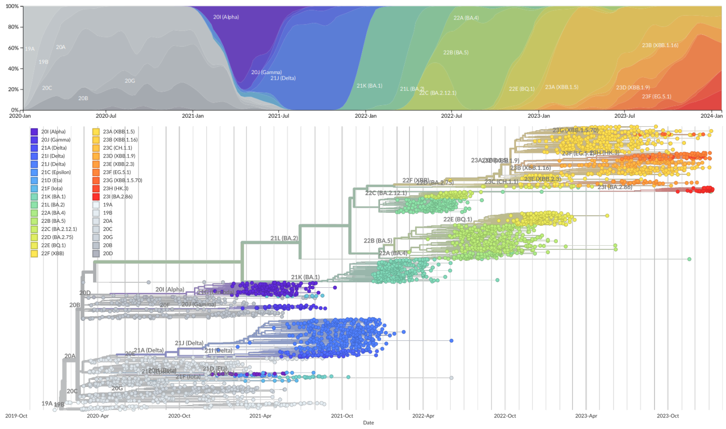
March 4, 2024

Contents

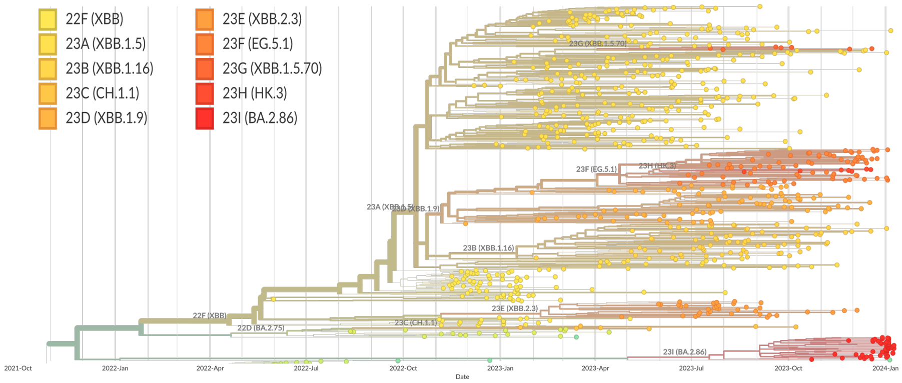
1 Context

1.1 Example: emerging variants

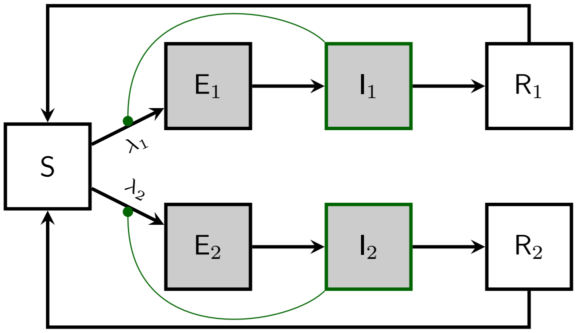
Example: surveillance for emerging SARS-CoV-2 variants



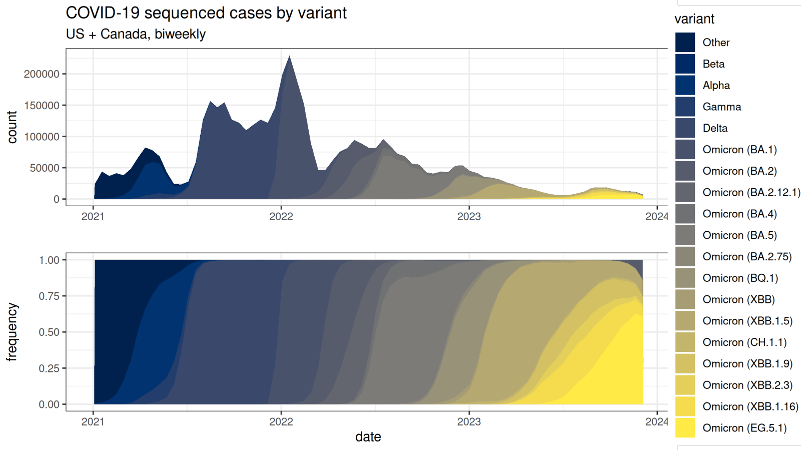
nextstrain.org (Hadfield *et al.*, 2018)



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



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



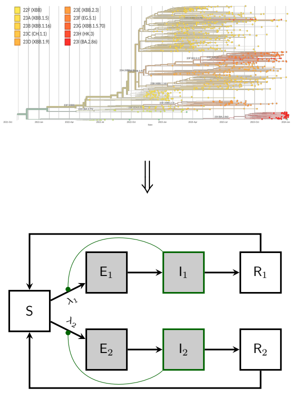
(Mathieu *et al.*, 2020)

1.2 Phylodynamics

What is phylodynamics?

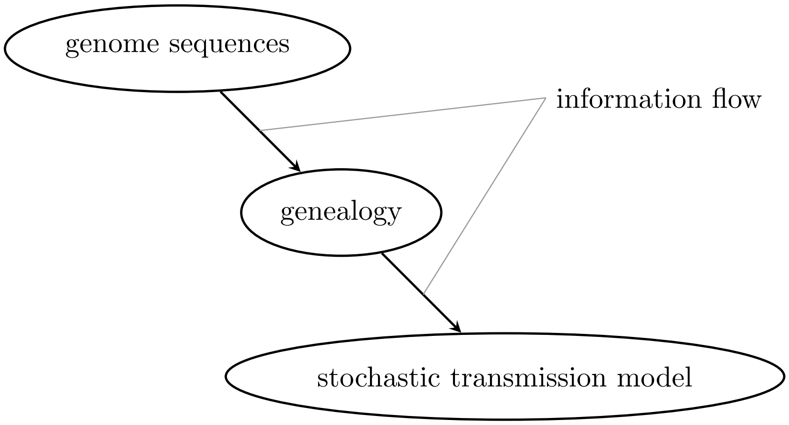
Broadly: Phylodynamics is the project of inferring *determinants of epidemic spread* using *genomic data from pathogen samples.*

In this talk: Phylodynamics means using *genomic data* to infer *stochastic dynamic transmission models.*



1.3 Problems of phylodynamics

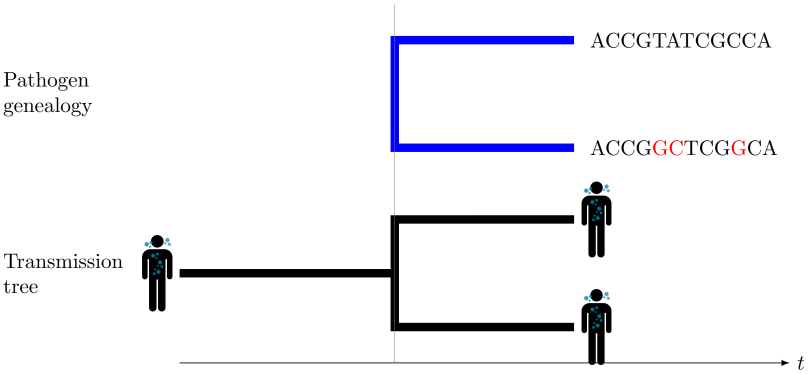
Core problems of phylodynamics

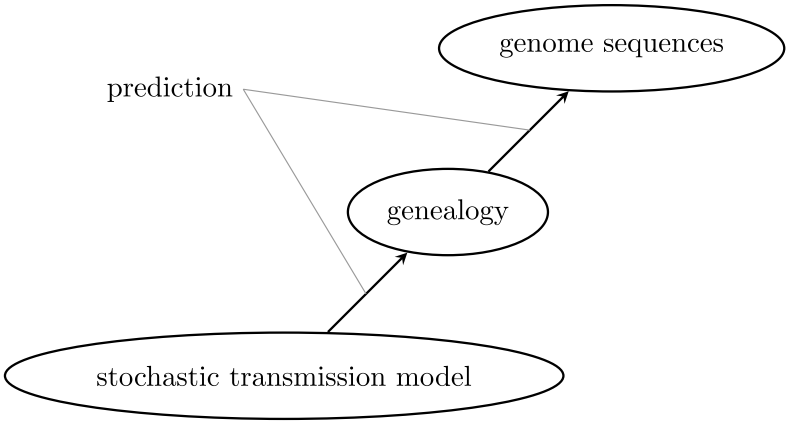

genome sequences

information flow

genealogy

stochastic transmission model





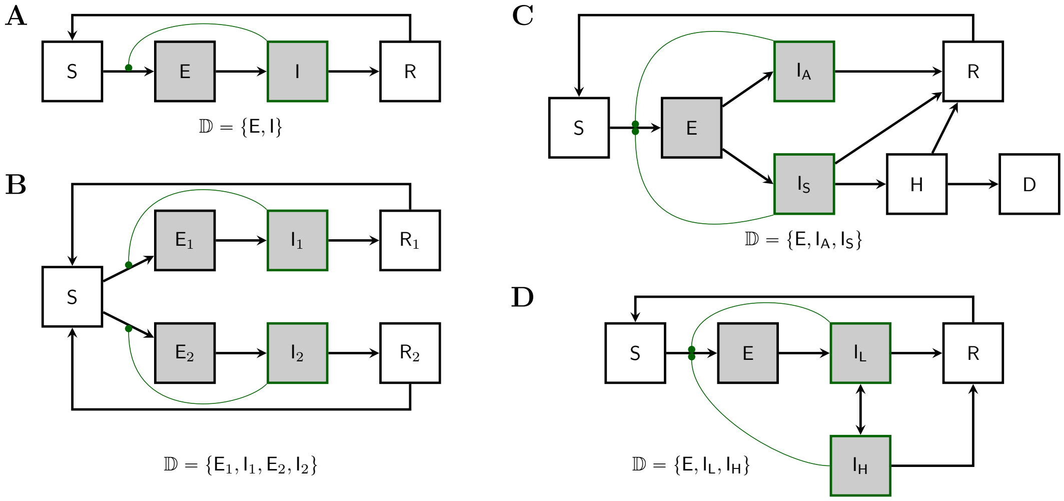
Overview

- We show how a given population process induces a unique genealogy process.
- *Pruning* and *obscuration* project a genealogy onto observable data.

2 Population process

2.1 Examples

Population process


A
S
E
I
R
$\mathbb{D} = \{\mathsf{E}, \mathsf{I}\}$
B
S
E_1
I_1
R_1
E_2
I_2
R_2
$\mathbb{D} = \{\mathsf{E}_1, \mathsf{I}_1, \mathsf{E}_2, \mathsf{I}_2\}$
C
S
E
I_A
I_S
R
H
D
$\mathbb{D} = \{\mathsf{E}, \mathsf{I}_\mathsf{A}, \mathsf{I}_\mathsf{S}\}$
D
S
E
I_L
R
I_H
$\mathbb{D} = \{\mathsf{E}, \mathsf{I}_\mathsf{L}, \mathsf{I}_\mathsf{H}\}$

2.2 Formalization

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:

$$\mathsf{Prob}\left[\mathbf{X}_0 \in \mathcal{E}\right] = \int_{\mathcal{E}} p_0(x)\,\mathrm{d}x$$

- Jump rates: $\alpha(t,x,x')$ = rate of jump $x \to x'$

$$\alpha(t,x,x') \geqslant 0, \qquad \int_{\mathbb{X}} \alpha(t,x,x')\,\mathrm{d}x' < \infty$$

- Multiple events at each jump are allowed.

Kolmogorov forward equation (KFE): If

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

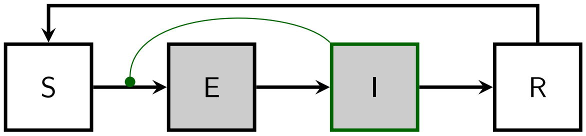
and

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

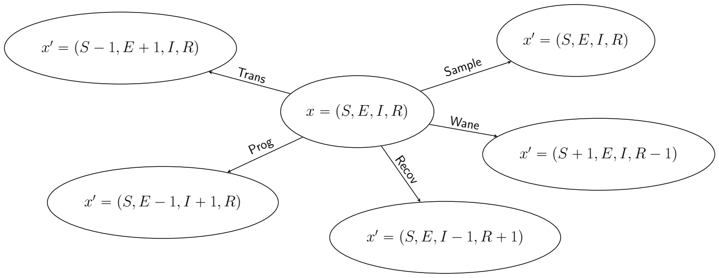
then

$$\int_{\mathcal{E}} w(t,x)\,\mathrm{d}x = \mathsf{Prob}\left[\mathbf{X}_t \in \mathcal{E}\right].$$

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



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



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

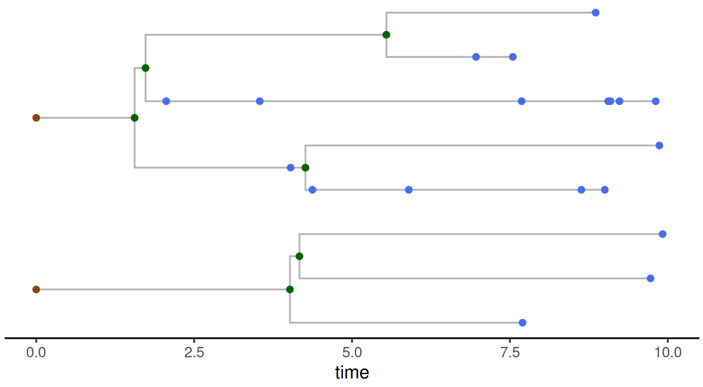
$$\frac{\partial w}{\partial t}(t,x) = \sum_{u\in\mathbb{U}} \left\{ \int w(t,x')\,\alpha_u(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\alpha_u(t,x,x')\,\mathrm{d}x' \right\}$$

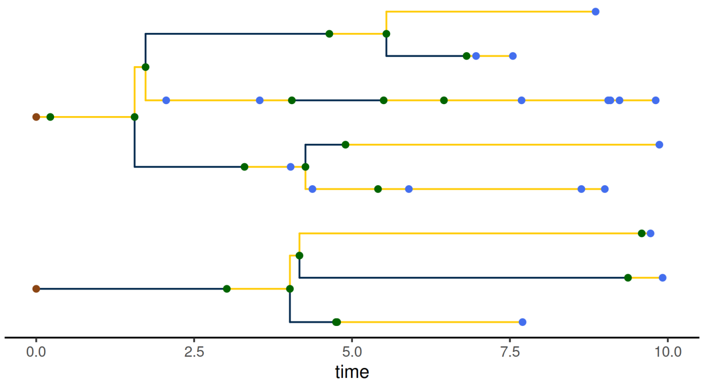
$$\begin{aligned}\frac{\partial w}{\partial t}(t,S,E,I,R) = {} & \frac{\beta(t)\,(S+1)\,I}{N}\,w(t,S+1,E-1,I,R) - \frac{\beta(t)\,S\,I}{N}\,w(t,S,E,I,R)\\ &+ \sigma\,(E+1)\,w(t,S,E+1,I-1,R) - \sigma\,E\,w(t,S,E,I,R)\\ &+ \gamma\,(I+1)\,w(t,S,E,I+1,R-1) - \gamma\,I\,w(t,S,E,I,R)\\ &+ \omega\,(R+1)\,w(t,S-1,E,I,R+1)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

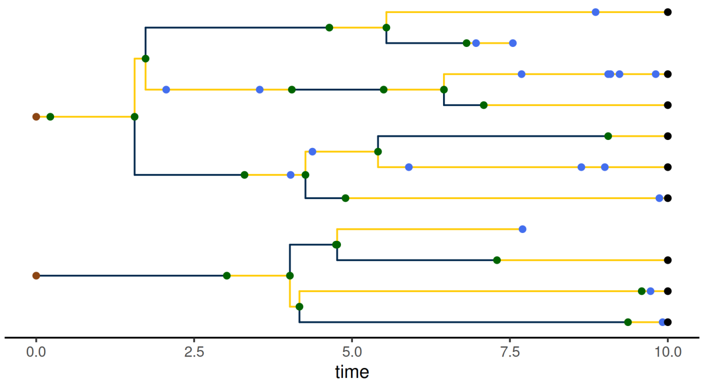
3 Genealogy process

3.1 Genealogies

What is a genealogy?

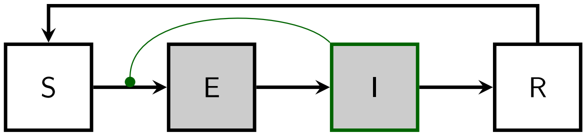




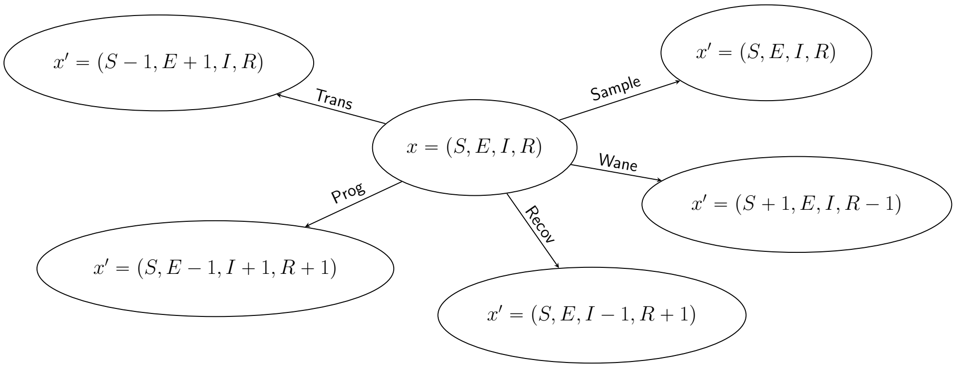


3.2 Induced genealogy process

Event types



$$\mathbb{D} = \{\mathsf{E}, \mathsf{I}\}$$



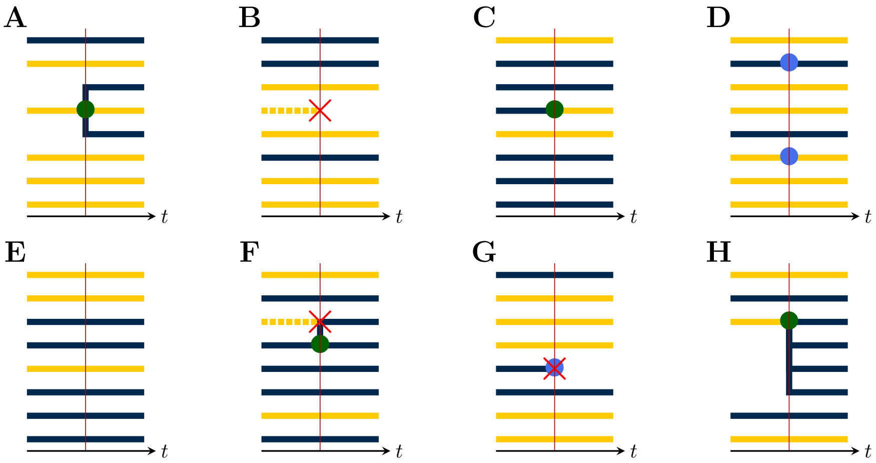
$\mathbb{U} = \{\mathsf{Trans}, \mathsf{Prog}, \mathsf{Recov}, \mathsf{Wane}, \mathsf{Sample}\}$

If we write

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

the KFE becomes

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



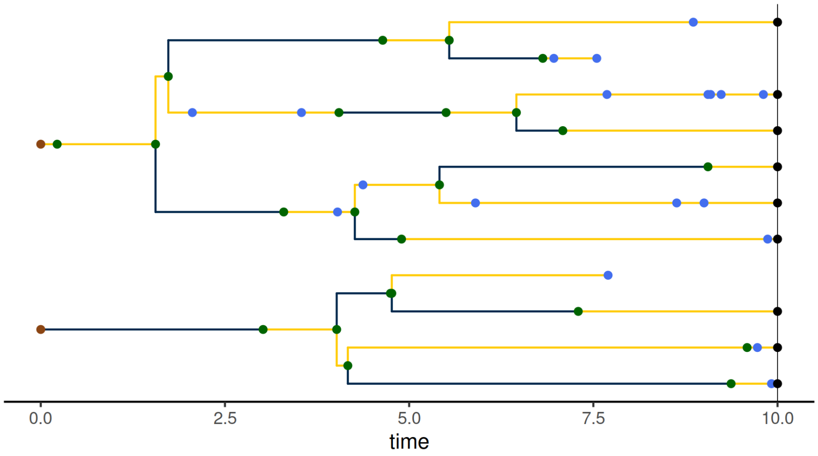
A population process induces a genealogy process

- $\mathbf{G}_t$ 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.

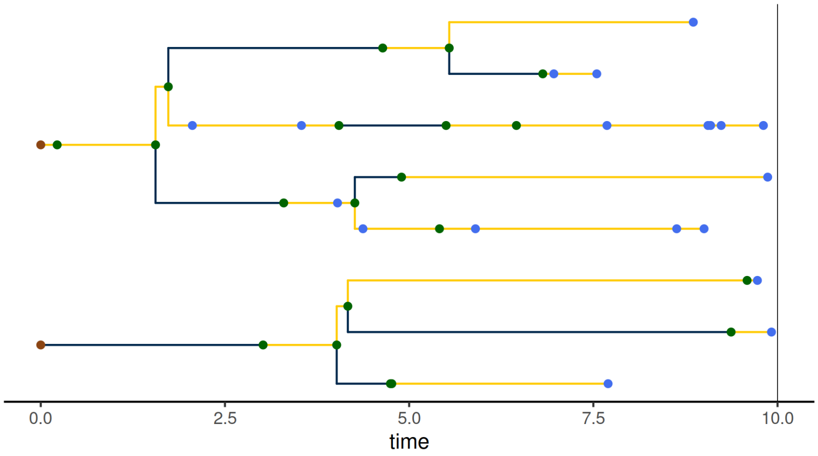
- Simulation code on github.com/kingaa/phylopomp
- Animations at https://kingaa.github.io/manuals/phylopomp/vignettes/

4 Pruned and obscured genealogies

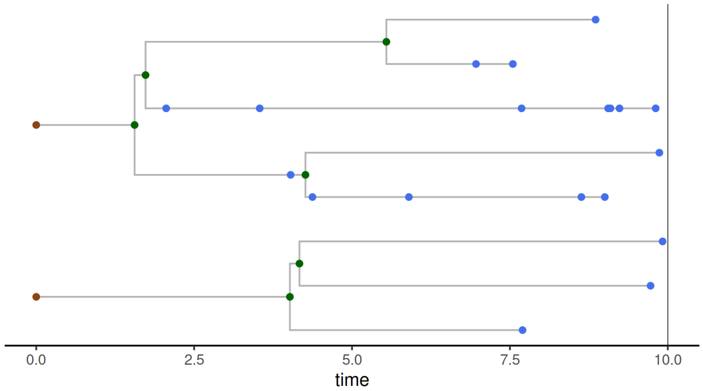
Full genealogy



Pruned genealogy



Obscured genealogy



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

Summary and outstanding challenges

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-Kaç (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, 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.

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.

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.

License, acknowledgments, and links

- The materials build on previous versions of this course and related courses.
-
- Produced with R version 4.3.2 and **pomp** version 5.6.

- Compiled on March 4, 2024.

Back to Lesson
R codes for this lesson