Exact phylodynamics via structured Markov genealogy processes

Edward Ionides

University of Michigan, Department of Statistics

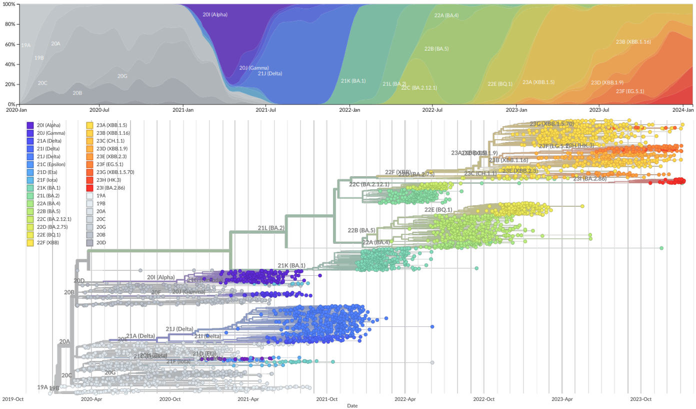
University of Waterloo, Department of Statistics and Actuarial Sciences
Seminar, 28 October 2025

Acknowledgments

This talk is based on King A. A., Lin, Q., & Ionides, E. L. (2025). Exact phylodynamic likelihood via structured Markov genealogy processes. *ArXiv:2405.17032*.

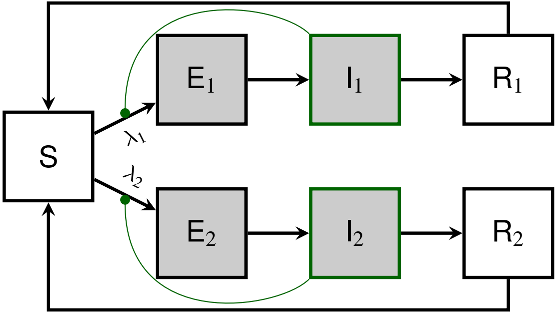
- Aaron King, Michigan. Project leader and credit for these slides.
  To see him present the topic:
  `https://www.youtube.com/watch?v=-KK7lTdYDYA`
- Qianying Lin, Ohio State
- Erik Volz, Imperial College
- Cris Moore & George Cantwell, Santa Fe Institute
- Ethan Romero-Seversen, Los Alamos
- Simon Frost, Microsoft
- NIH Grants 1R01AI143852, 1U54GM111274
- NSF/NIH Interface Grant 1761603
- NSF Grant DMS-2526827

Example: surveillance for emerging SARS-CoV-2 variants



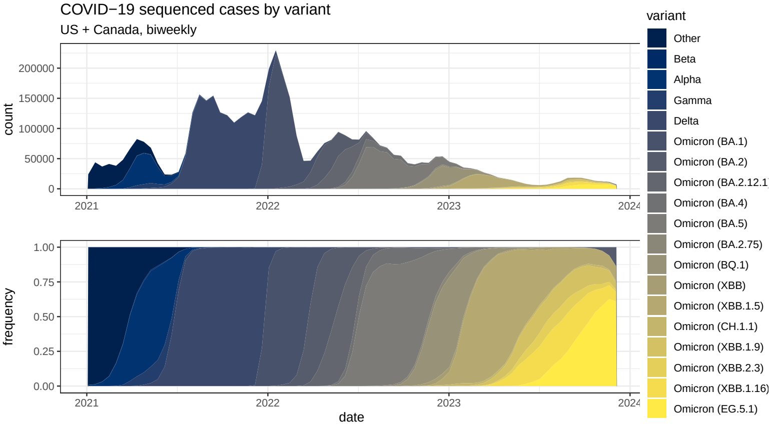
`nextstrain.org` (Hadfield et al., 2018)

Example: surveillance for emerging SARS-CoV-2 variants



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

Example: surveillance for emerging SARS-CoV-2 variants

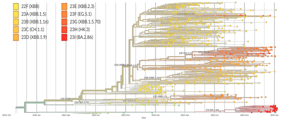


(Mathieu et al., 2020)

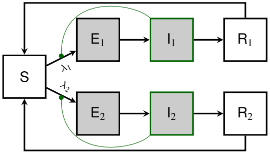
What is phylodynamics?

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

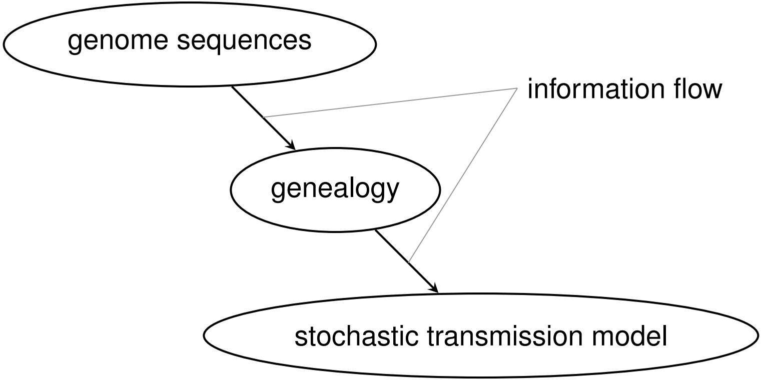
In this talk:
Phylodynamics means using
genomic data
to infer
stochastic dynamic transmission models.



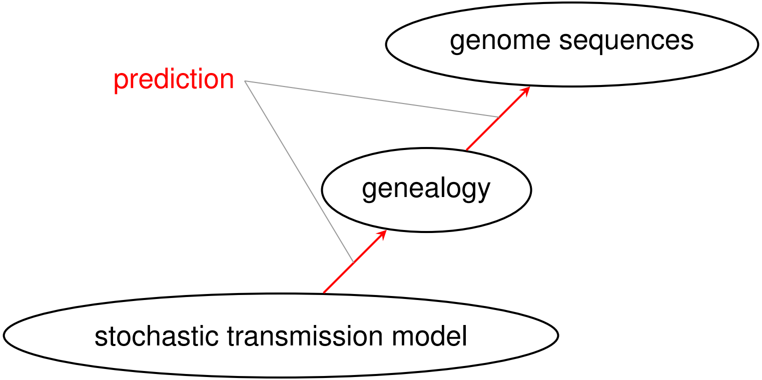
⇓



Core problems of phylodynamics



Core problems of phylodynamics



Core problems of phylodynamics

S = set of genome sequences

Φ = genealogical tree relating the sequences

E = sequence evolution model

D = dynamic, stochastic transmission model

Y = other time series data

$$\mathcal{L} = f(S, Y|D, E) = \int f(S|\Phi, E)\, f(\Phi, Y|D)\, \mathrm{d}\Phi$$

$f(S|\Phi, E)$ = phylogenetic likelihood

$f(\Phi, Y|D)$ = phylodynamic likelihood

Current approaches

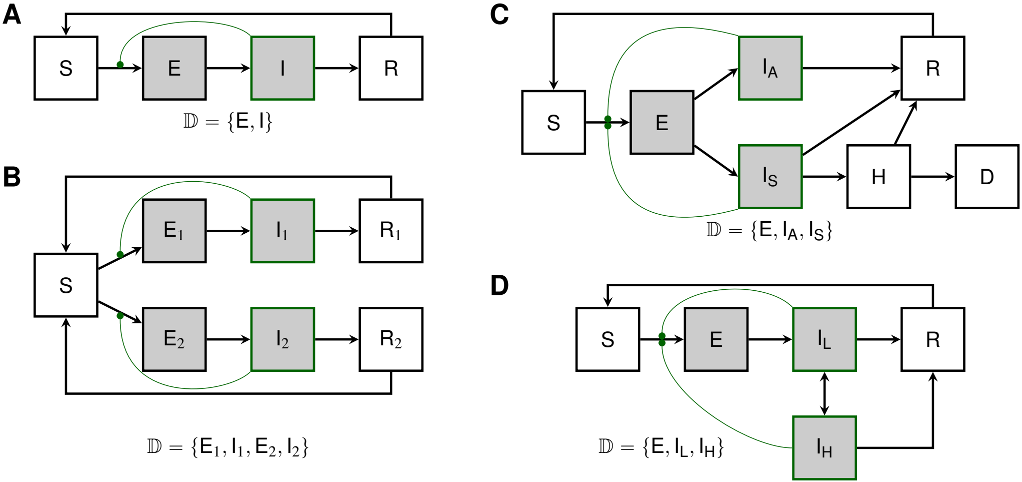
- Coalescent models
    - asymptotic large-population justification
    - naturally studied backwards in time
    - hard to relate formally to small-population models specified forwards in time, except in special cases.
- Linear birth-death processes
    - tractability stems from independence of lineages
    - simple branching models struggle to explain nonlinear phenomena such as susceptible depletion.
    - linearization is possible under large-population, small sample-fraction assumptions
- How can we investigate scientifically motivated nonlinear models?
- An exact likelihood-based method would be statistically efficient.

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.

Details on the arXiv (King et al., 2024)

Population process



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

$\mathbb{D} = \{\mathrm{E}_1, \mathrm{I}_1, \mathrm{E}_2, \mathrm{I}_2\}$

$\mathbb{D} = \{\mathrm{E}, \mathrm{I}_\mathrm{A}, \mathrm{I}_\mathrm{S}\}$

$\mathbb{D} = \{\mathrm{E}, \mathrm{I}_\mathrm{L}, \mathrm{I}_\mathrm{H}\}$

Population process

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

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

- 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.

Population process

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

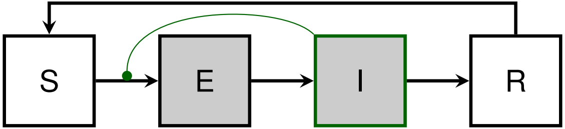
$$w(0,x) = p_0(x)$$

then

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

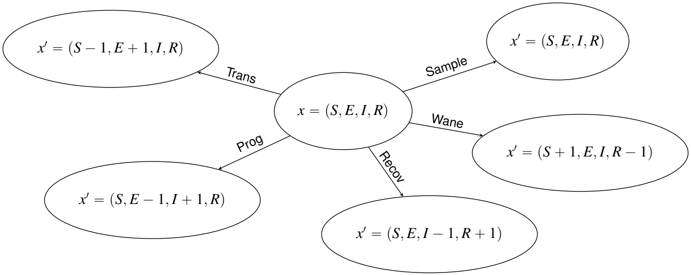
KFE is sometimes called the *master equation* for $\mathbf{X}_t$.

Population process



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

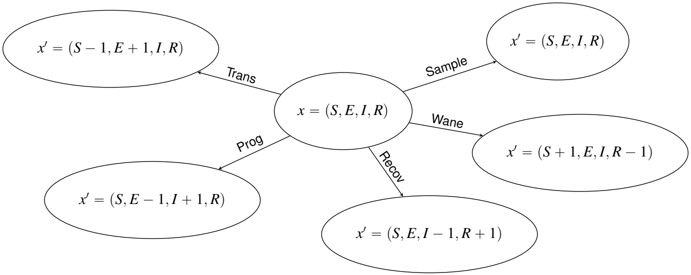
Population process



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

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

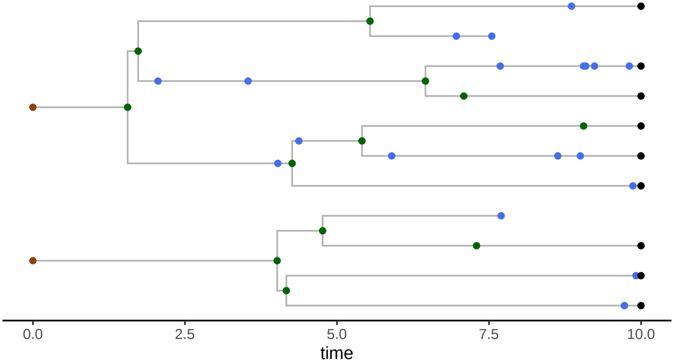
Population process



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

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

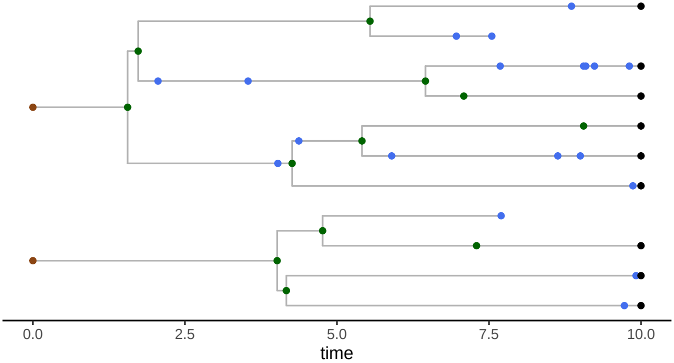
What is a genealogy?



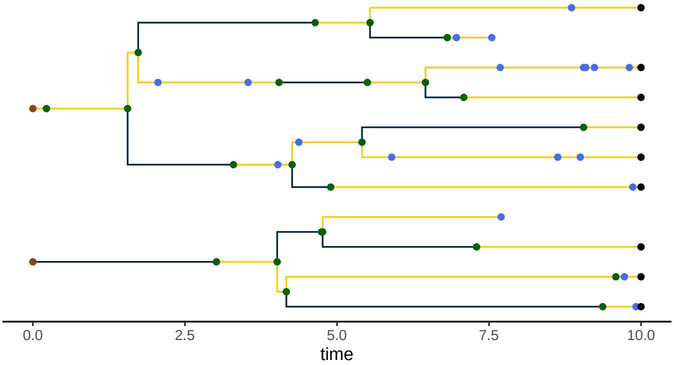
What is a genealogy?

- $\mathbb{L}$: countable set of labels
- partit$(\mathbb{L})$: set of collections of finite, mutually-disjoint subsets of $\mathbb{L}$.
- partition *fineness* defines a partial order, $\preccurlyeq$, on partit$(\mathbb{L})$.
- The tree structure of a genealogy is a monotone, càdlàg map
  $Z : [0, T] \to$ partit$(\mathbb{L})$ such that $t_1 \leqslant t_2$ implies $Z_{t_1} \preccurlyeq Z_{t_2}$.

What is a genealogy?



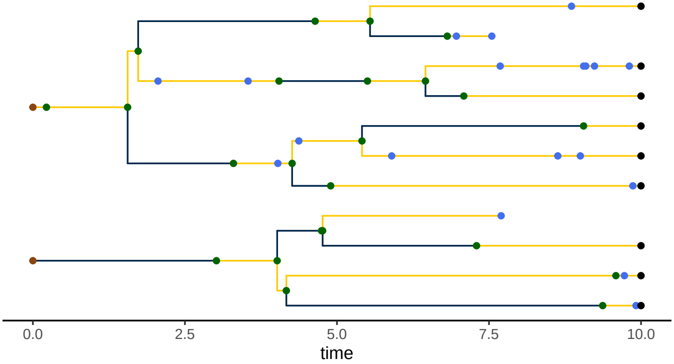
What is a genealogy?



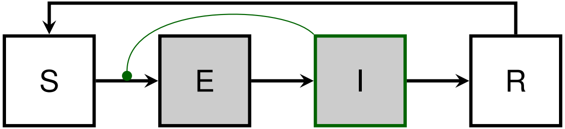
What is a genealogy?

- A *coloring*, Y, is an assignment of a deme to each point of the genealogy.
- For $t \in [0, T]$, a a label, $Y_t(a) = (Y_t^{\mathsf{d}}(a), Y_t^{\mathsf{m}}(a)) \in \mathbb{D} \times \mathbb{Z}_+$
- $Y_t^{\mathsf{d}}(a)$ is the deme in which the lineage of a is located at time t.
- $Y_t^{\mathsf{m}}(a)$ is the number of nodes encountered along the lineage a in going from time 0 to t.
- $Y_t^{\mathsf{m}}(a)$ is a simple counting process.
- Given a tree Z, let Y(Z) denote the set of colorings Y that are compatible with Z
- Formally, a genealogy is a triple, (T, Z, Y).

What is a genealogy?

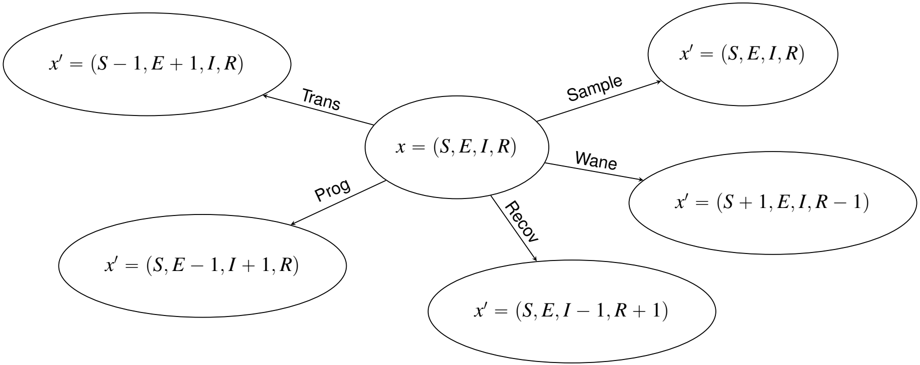


Event types



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

Event types



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

Event types

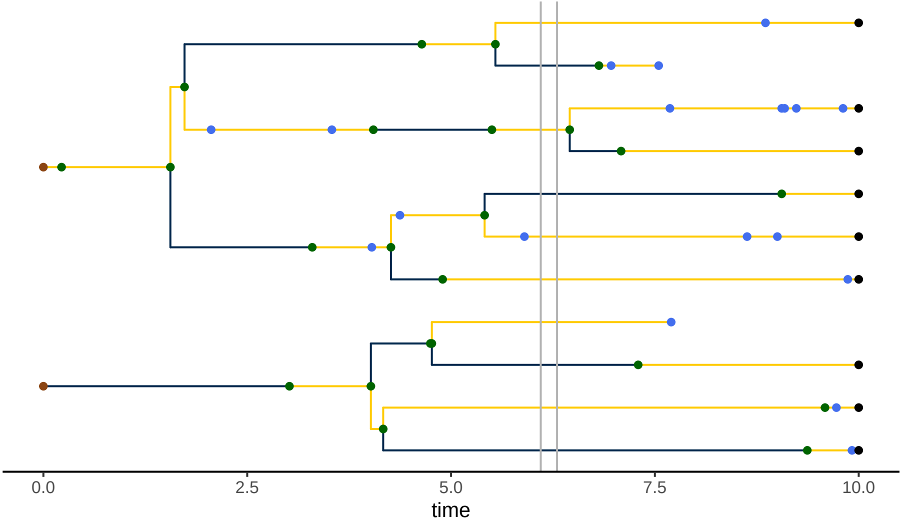
If we write

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

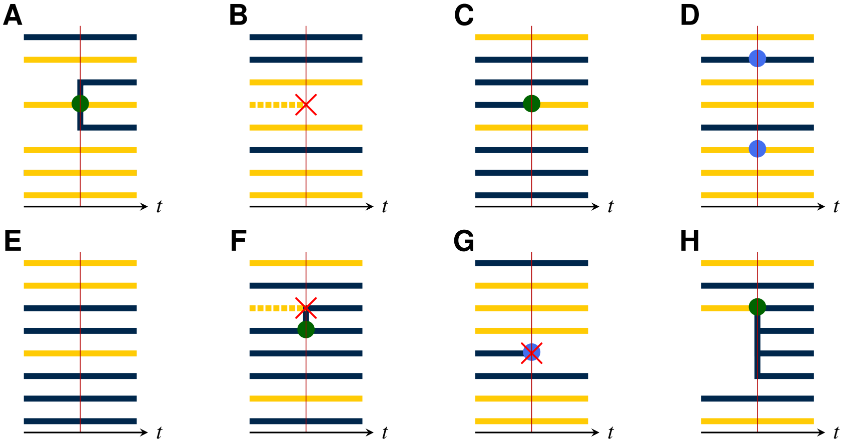
the KFE becomes

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

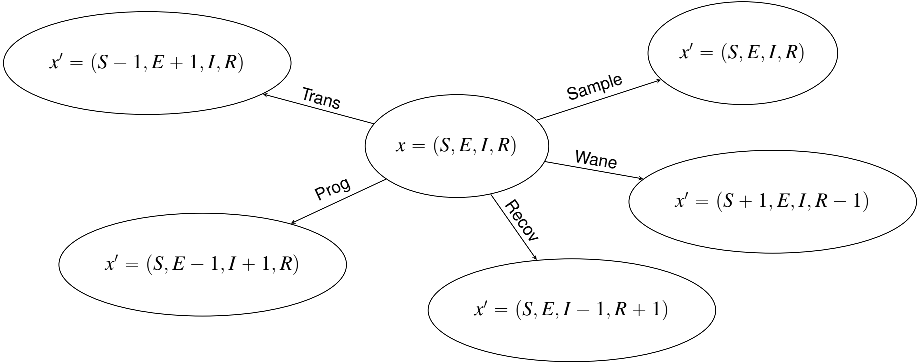
Event types



Event types



Event types



$x = (S, E, I, R)$

Trans

$x' = (S-1, E+1, I, R)$

Prog

$x' = (S, E-1, I+1, R)$

Recov

$x' = (S, E, I-1, R+1)$

Wane

$x' = (S+1, E, I, R-1)$

Sample

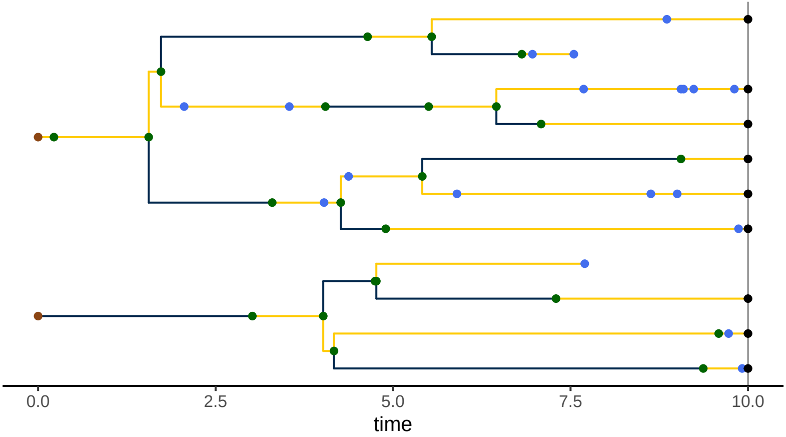
$x' = (S, E, I, R)$

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

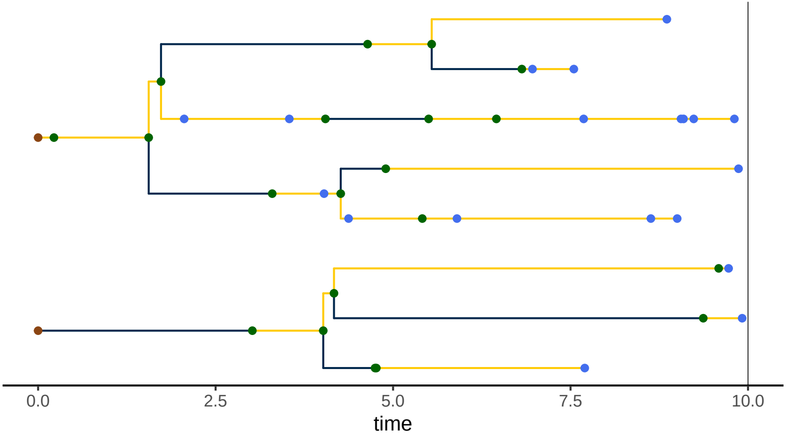
A population process induces a genealogy process

- $\mathbf{G}_t$ 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/`

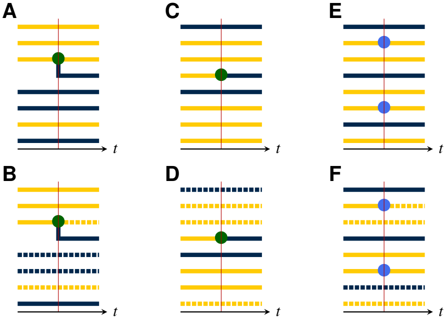
Full genealogy



Pruned genealogy

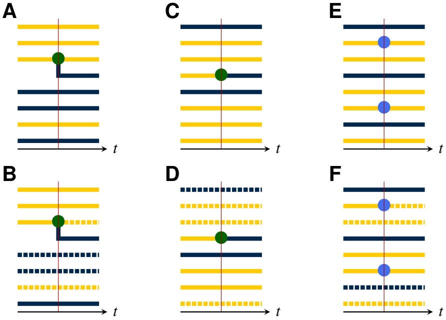


Local structure of a pruned genealogy



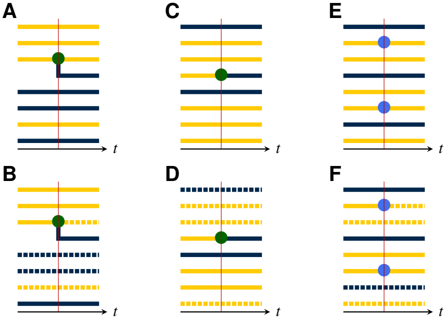
Top row shows the *unpruned genealogy* in neighborhood of an event.
Bottom row shows the corresponding *pruned genealogy*.

Local structure of a pruned genealogy



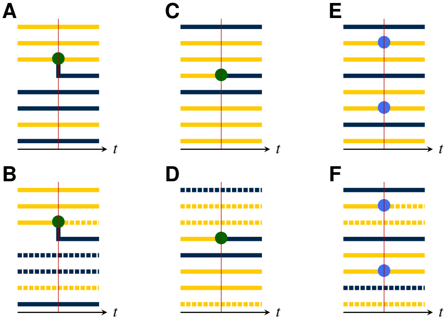
For $x \in \mathbb{X}$, $i \in \mathbb{D}$, $n_i(x)$ is the *occupancy* of deme i when the system is in state x.
In panel A $n = (n_{\text{blue}}, n_{\text{yellow}}) = (4, 4)$; in panel C $n = (3, 5)$;

Local structure of a pruned genealogy



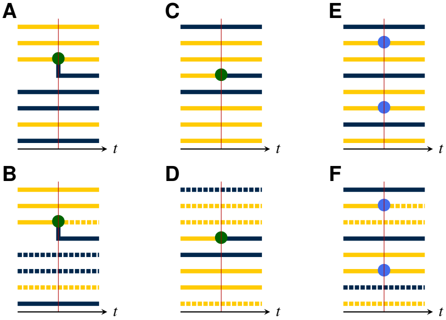
For $u \in \mathbb{U}$, $i \in \mathbb{D}$, r_i^u is the *production* of event u in deme i.
In panel A, $r = (r_{\text{blue}}, r_{\text{yellow}}) = (1, 1)$; in panel E, $r = (0, 2)$.

Local structure of a pruned genealogy



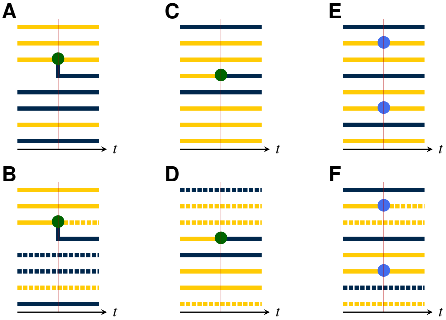
The *lineage count*, $\ell_i(t)$, is the number of unpruned lineages in deme i at time t.
In this case, for all panels, $\ell = (2, 2)$.

Local structure of a pruned genealogy



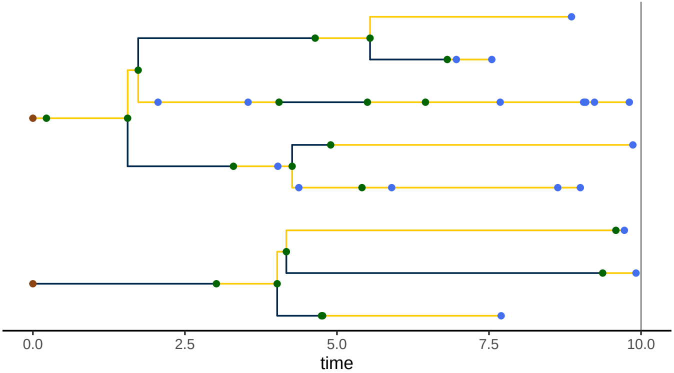
The *saturation*, s_i, is the number of unpruned lineages in deme i *descending* from the event. In panels B and D, $s = (1, 0)$; in panel F, $s = (0, 1)$.

Local structure of a pruned genealogy



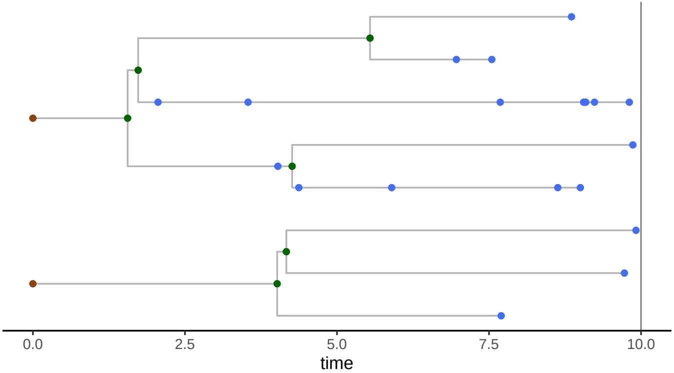
Obviously, $s_i \leqslant r_i \leqslant n_i$ and $s_i \leqslant \ell_i \leqslant n_i$.

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

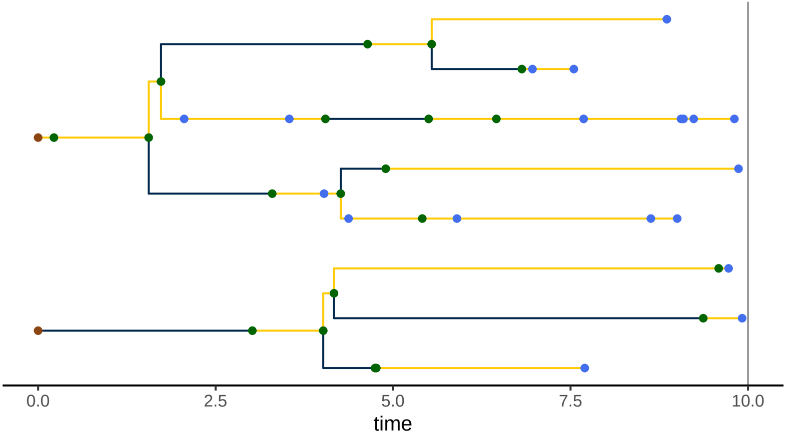
For $n, r, \ell, s \in \mathbb{Z}_+^{\mathbb{D}}$, define the *binomial ratio*

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

Observe that $\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \in [0, 1]$. Moreover,

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

Theorem: likelihood of a pruned genealogy



Theorem: likelihood of a pruned genealogy

Suppose that $\mathrm{P} = (\mathrm{Y}, \mathrm{Z})$ is a given pruned genealogy with depth T.
Define

$$\phi_u(x, y, y') \coloneqq \binom{n(x) \quad \ell(y')}{r^u \quad s(y, y')} Q_u(y, y').$$

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

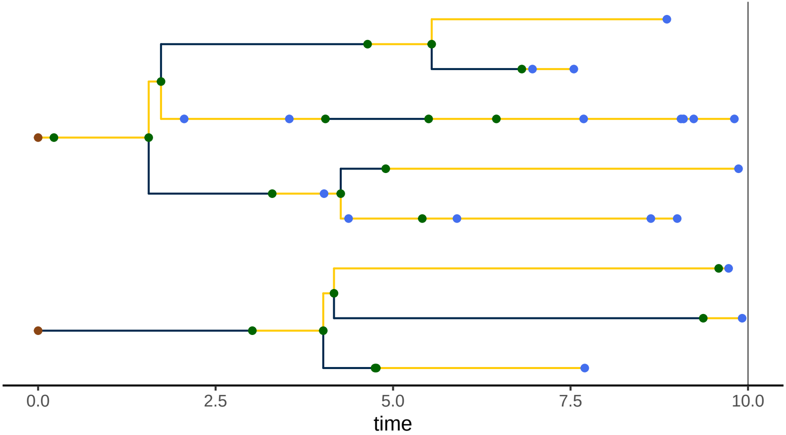
If $w = w(t,x)$ satisfies the initial condition $w(0,x) = p_0(x)$ and the filter equation

$$\begin{aligned}\frac{\partial w}{\partial t}(t,x) &= \sum_u \int w(t,x')\,\alpha_u(t,x',x)\,\phi_u(x,\tilde{\mathrm{Y}}_t,\mathrm{Y}_t)\,\mathrm{d}x' - \sum_u \int w(t,x)\,\alpha_u(t,x,x')\,\mathrm{d}x', && t \notin \text{ev}(\mathrm{P}),\\ w(t,x) &= \sum_u \int \tilde{w}(t,x')\,\alpha_u(t,x',x)\,\phi_u(x,\tilde{\mathrm{Y}}_t,\mathrm{Y}_t)\,\mathrm{d}x', && t \in \text{ev}(\mathrm{P}),\end{aligned}$$

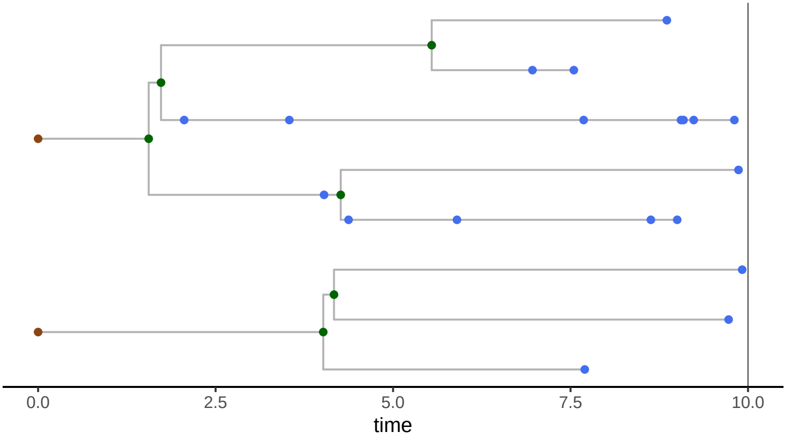
then the likelihood of P is

$$\mathcal{L} = \int w(\mathrm{T},x)\,\mathrm{d}x.$$

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

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

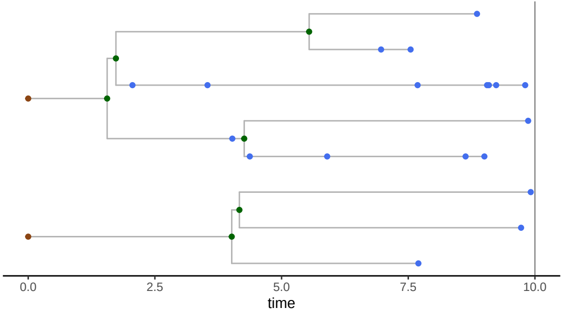
and if $w = w(t,x,y)$ satisfies the initial condition $w(0,x,y) = p_0(x)\,\mathbb{1}\{q(x,y) > 0\}$ and the filter equation

$$\begin{aligned} \frac{\partial w}{\partial t} &= \sum_{uy'} \int w(t,x',y')\,\beta_u(t,x',x,y',y)\,\psi_u(t,x',x,y',y)\,\mathrm{d}x' - \sum_{uy'} \int w(t,x,y)\,\beta_u(t,x,x',y,y')\,\mathrm{d}x', && t \notin \mathrm{ev}(\mathrm{Z}), \\ w(t,x,y) &= \sum_{uy'} \int \widetilde{w}(t,x',y')\,\beta_u(t,x',x,y',y)\,\psi_u(t,x',x,y',y)\,\mathrm{d}x', && t \in \mathrm{ev}(\mathrm{Z}), \end{aligned}$$

then the likelihood of (T, Z) is

$$\mathcal{L} = \sum_{y} \int w(T,x,y)\,\mathrm{d}x.$$

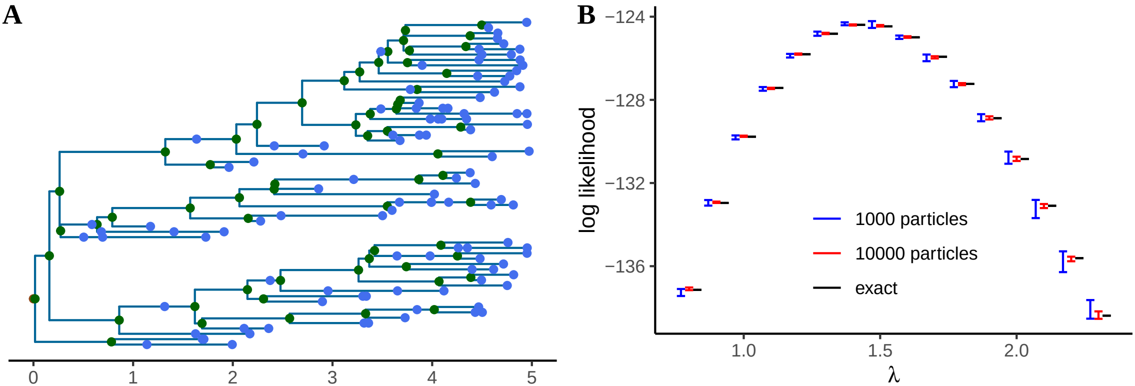
Theorem: likelihood of an obscured genealogy



$$\frac{\partial w}{\partial t} = \sum_{uy'} \int w(t,x',y')\, \beta_u(t,x',x,y',y)\, \psi_u(t,x',x,y',y)\, \mathrm{d}x' - \sum_{uy'} \int w(t,x,y)\, \beta_u(t,x,x',y,y')\, \mathrm{d}x', \qquad t \in \mathsf{ev}(\mathrm{Z}),$$

$$w(t,x,y) = \sum_{uy'} \int \tilde{w}(t,x',y')\, \beta_u(t,x',x,y',y)\, \psi_u(t,x',x,y',y)\, \mathrm{d}x', \qquad t \in \mathsf{ev}(\mathrm{Z}),$$

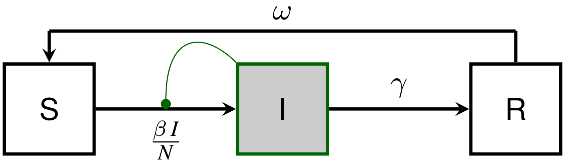
Linear birth-death model



Uniform sampling.

Exact likelihoood is available in closed form.

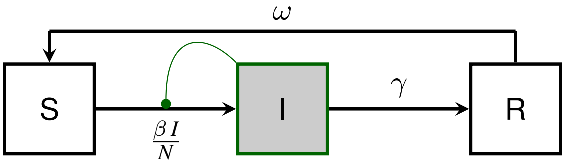
SIRS model



Between genealogical events:

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

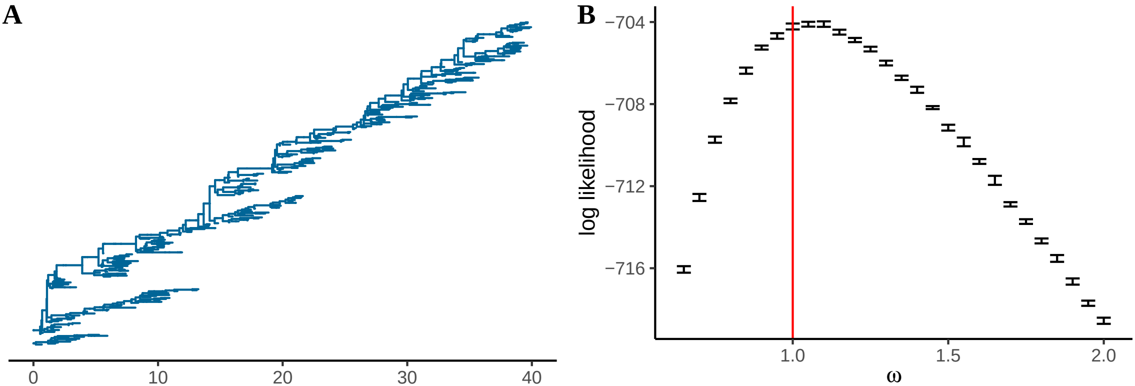
SIRS model



At genealogical events:

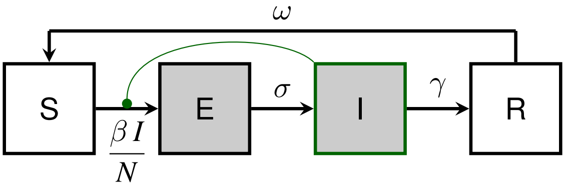
$$w(t,S,I,R) = \begin{cases} \frac{2\,\beta\,(S+1)}{I\,N}\,\tilde{w}(t,S+1,I-1,R), & \text{branch point at } t, \\ \psi\,\tilde{w}(t,S,I,R), & \text{internal sample at } t, \\ \psi\ (I-\ell)\ \tilde{w}(t,S,I,R), & \text{terminal sample at } t. \end{cases}$$

SIRS model

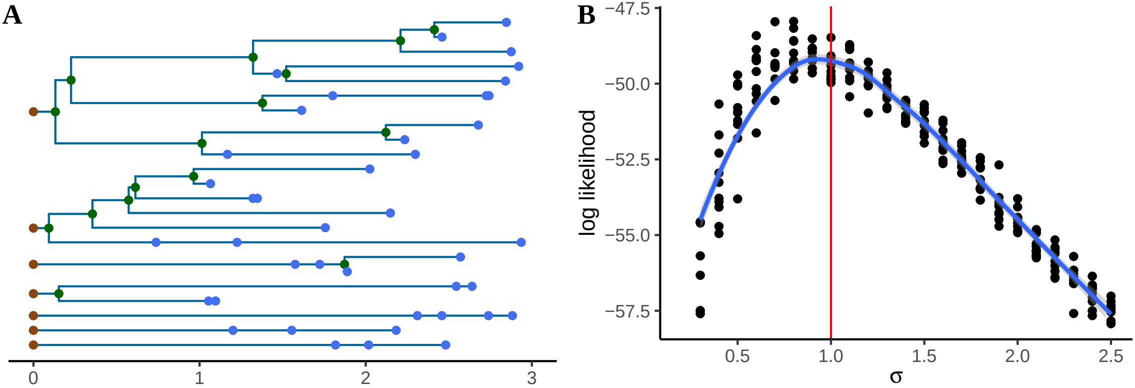


Uniform sampling.
One deme only.

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 & Frost, 2014; Rasmussen et al., 2014; Vaughan et al., 2019).
- All computations can be carried out forward in time.
  This expands the class of models that can be entertained.
- There is great flexibility in the sampling model.
- Other data streams can be readily and simultaneously assimilated.
- Applications beyond infectious disease epidemiology.
- Full details in King et al. (2024).

Outstanding challenges

- There is some way to go before these results translate into algorithms.
- Key issues: scalability and expense
- Efficient choice of importance-sampling kernel
  (Borrowing information from future is allowed.)
- Phylogenetic uncertainty
- Efficient simulation 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.
- Existing tree-based phylodynamic approaches are special cases.
- Various approaches to solving this equation are possible and have yet to be fully explored.
- These results liberate us to entertain models that more closely match our scientific questions, with less hindrance from inference methodology.

References

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**:4121–4123.
DOI: `10.1093/bioinformatics/bty407`

King, A. A., Lin, Q., & Ionides, E. L. (2022) Markov genealogy processes. *Theoretical Population Biology* **143**:77–91.
DOI: `10.1016/j.tpb.2021.11.003`

King, A. A., Lin, Q., & Ionides, E. L. (2024) Exact phylodynamic likelihood via structured Markov genealogy processes. *arXiv* 2405.17032.
DOI: `10.48550/arxiv.2405.17032`

References II

King, A. A., Nguyen, D., & Ionides, E. L. (2016) Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software* **69**: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]* . `https://ourworldindata.org/coronavirus`

Rasmussen, D. A., Ratmann, O., & Koelle, K. (2011) Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology* **7**:e1002136. DOI: `10.1371/journal.pcbi.1002136`

References III

Rasmussen, D. A., Volz, E. M., & Koelle, K. (2014) Phylodynamic inference for structured epidemiological models. *PLoS Computational Biology* **10**:e1003570.
DOI: `10.1371/journal.pcbi.1003570`

Stadler, T. (2010) Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* **267**:396–404.
DOI: `10.1016/j.jtbi.2010.09.010`

Vaughan, T. G., Leventhal, G. E., Rasmussen, D. A., Drummond, A. J., Welch, D., & Stadler, T. (2019) Estimating epidemic incidence and prevalence from genomic data. *Molecular Biology and Evolution* **36**:1804–1816.
DOI: `10.1093/molbev/msz106`

References IV

Volz, E. M. (2012) Complex population dynamics and the coalescent under neutrality. *Genetics* **190**:187–201.
DOI: `10.1534/genetics.111.134627`

Volz, E. M. & Frost, S. D. W. (2014) Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of the Royal Society, Interface* **11**:20140945.
DOI: `10.1098/rsif.2014.0945`

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., & Frost, S. D. W. (2009) Phylodynamics of infectious disease epidemics. *Genetics* **183**:1421–1430.
DOI: `10.1534/genetics.109.106021`