Forward-in-time Phylodynamics via Sequential Monte Carlo

Aaron A. King
R. A. (Alex) Smith
Edward L. Ionides

University of Michigan
EEB, Complex Systems, Mathematics,
Statistics, Bioinformatics

CENTER FOR
INFERENCE & DYNAMICS
OF INFECTIOUS DISEASES

Alex Smith

Ed Ionides

• Smith, R. A.; Ionides, E. L. and King, A. A. (2017). "Infectious disease dynamics inferred from genetic data via sequential Monte Carlo", *Molecular Biology and Evolution* 34: 2065-2084.
• Smith, Ionides, King (in prep)

Data courtesy of the Michigan Department of Community Health.
Funding from an NIH grant to the
Center for Inference & Dynamics of Infectious Disease,
a MIDAS Center of Excellence.

Phylodynamics

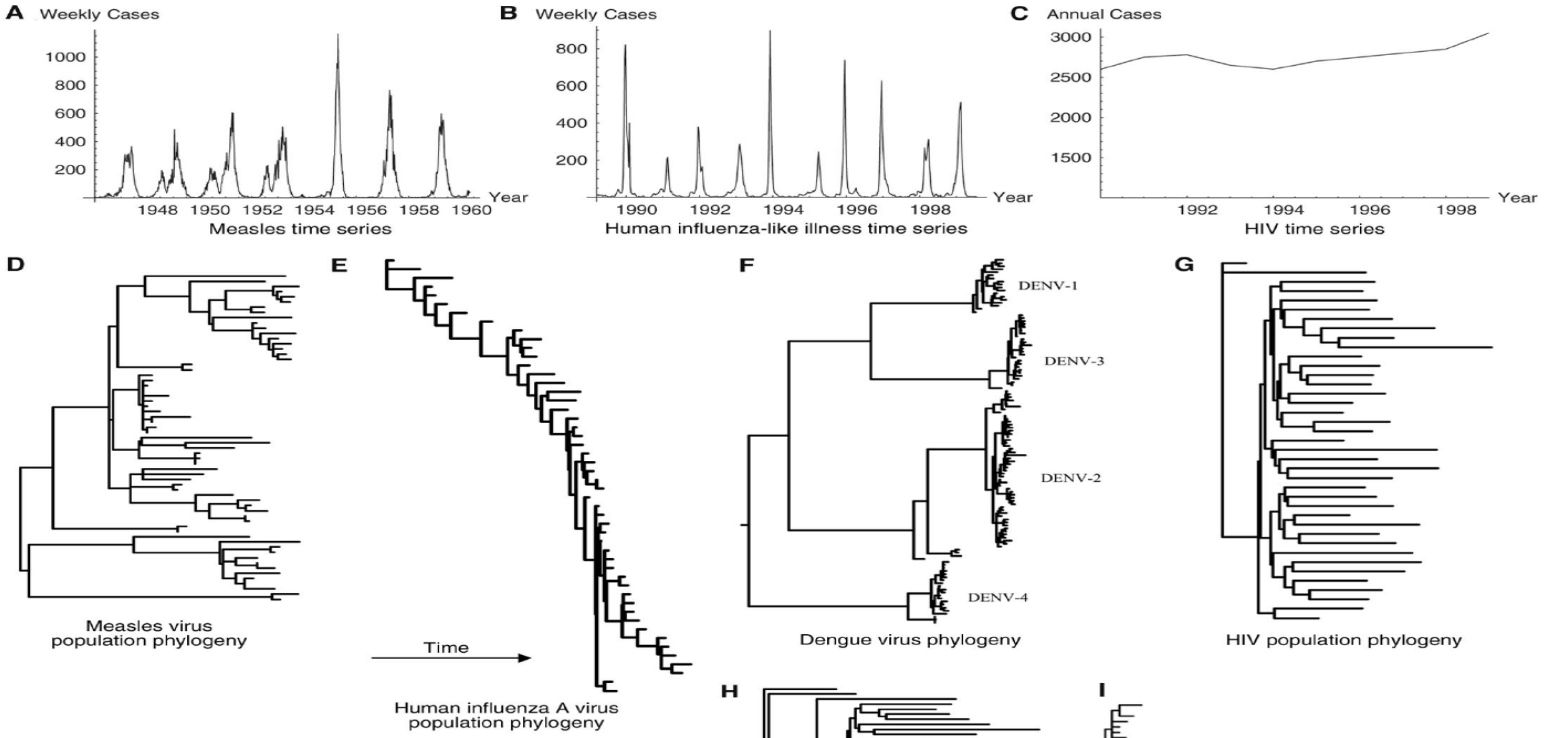


Fig. 1. (**A**) Prevaccination measles dynamics: weekly case reports for Leeds, UK (*7*). (**B**) Weekly reports of influenza-like illness for France (*44*). (**C**) Annual diagnosed cases of HIV in the United Kingdom (*45*). (**D**) Measles phylogeny: the measles virus nucleocapsid gene [63 sequences, 1575 base pairs (bp)]. (**E**) Influenza phylogeny: the human influenza A virus (subtype H3N2) hemagglutinin (*HA1*) gene longitudinally sampled over a period of 32 years (50 sequences, 1080 bp). (**F**) Dengue phylogeny: the dengue virus envelope gene from all four serotypes (DENV-1 to DENV-4, 120 sequences, 1485 bp). (**G**) HIV-1 population phylogeny: the subtype B envelope (E) gene sampled from different patients (39 sequences, 2979 bp). (**H**) HCV population phylogeny: the virus genotype 1b E1E2 gene sampled from different patients (65 sequences, 1677 bp). (**I**) HIV-1 within-host phylogeny: the partial envelope (E) gene longitudinally sampled from a single patient over 5.8 years [58 sequences, 627 bp; patient 6 from (*26*)]. All sequences were collected from GenBank and trees were constructed with maximum likelihood in PAUP* (*46*). Horizontal branch lengths are proportional to substitutions per site. Further details are available from the authors on request.

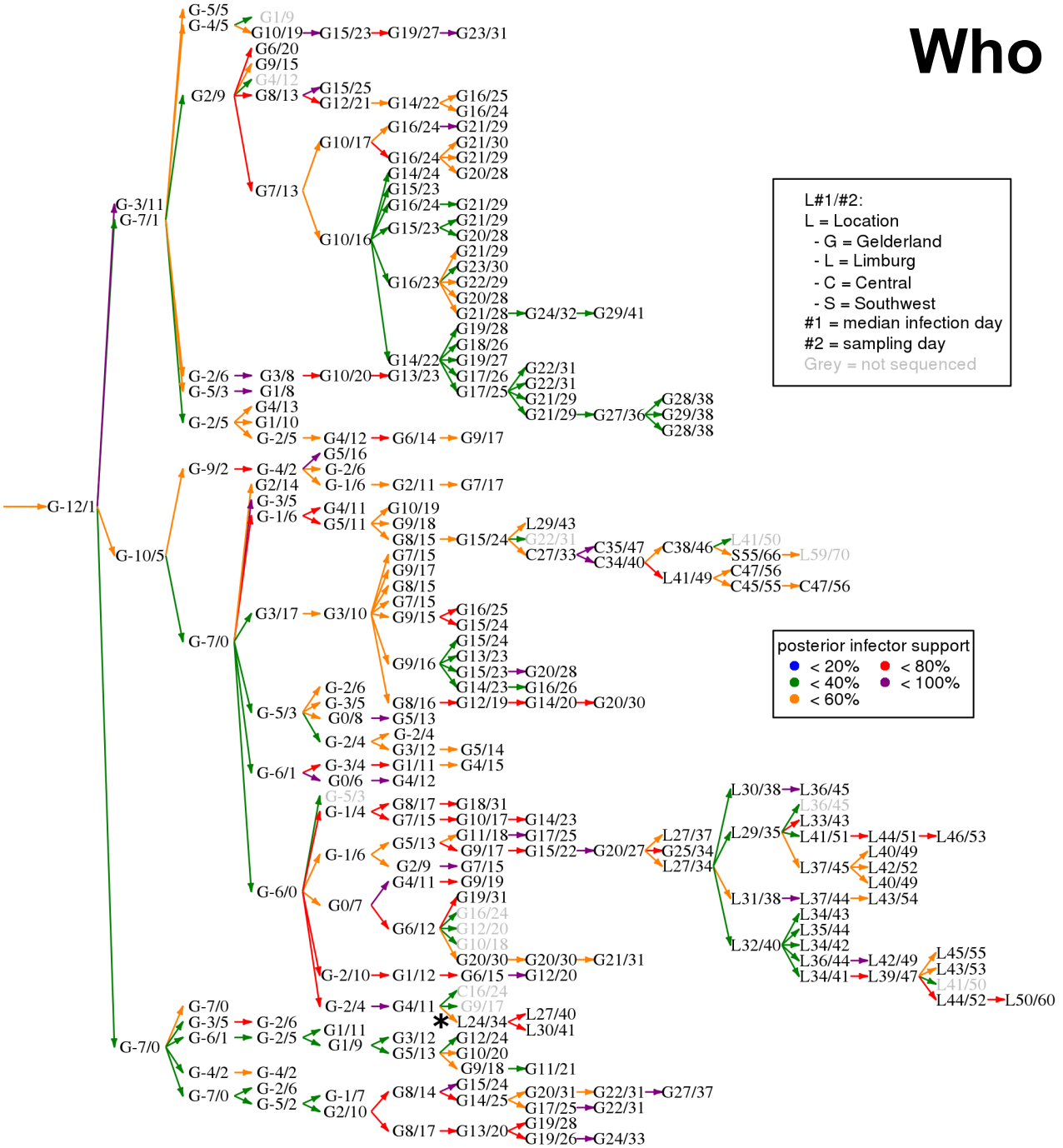
Grenfell et al. [2004]

Phylodynamics

Two paradigms:
- Small outbreaks: who acquires infection from whom?
- Model inference: assume phylogeny is generated by a stochastic transmission process

Current approaches commonly assume:
- *Neutral* evolution of sequences
- No *recombination* or *reassortment*
- Phylogenetic branchpoints *coincide* with transmission events

Who infected whom?

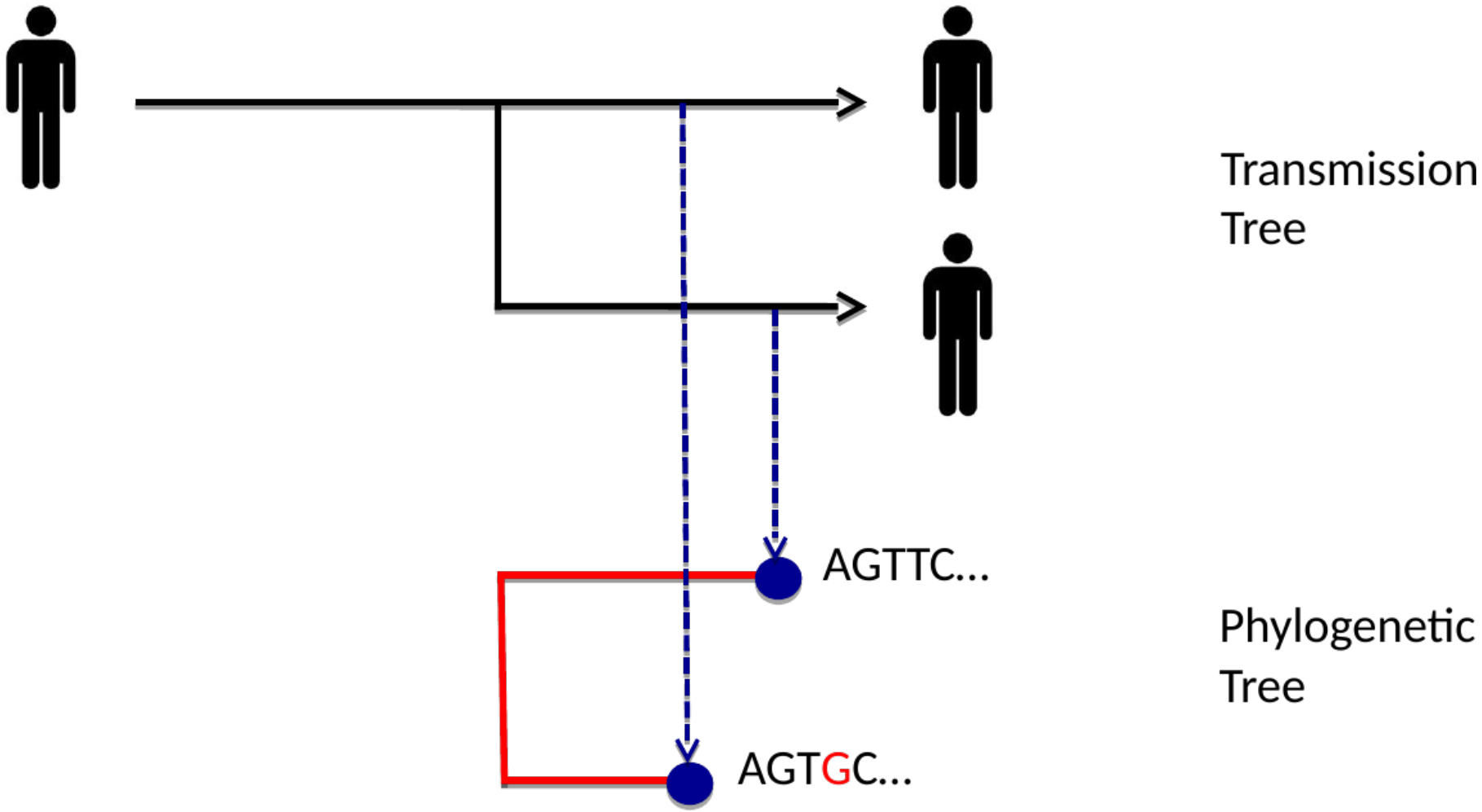


Phylodynamics

Two paradigms:
- Small outbreaks: who acquires infection from whom?
- Model inference: assume phylogeny is generated by a stochastic transmission process

Current approaches commonly assume:
- *Neutral* evolution of sequences
- No *recombination* or *reassortment*
- Phylogenetic branchpoints *coincide* with transmission events

Transmission trees and phylogenetic trees

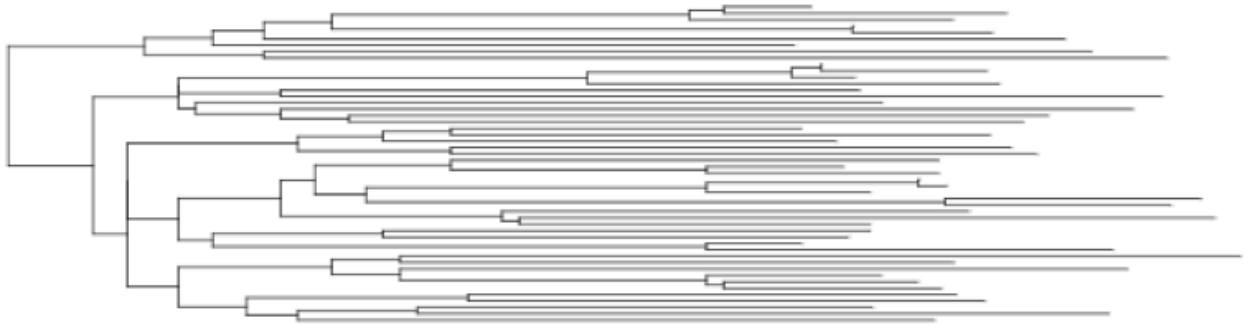


Phylodynamics

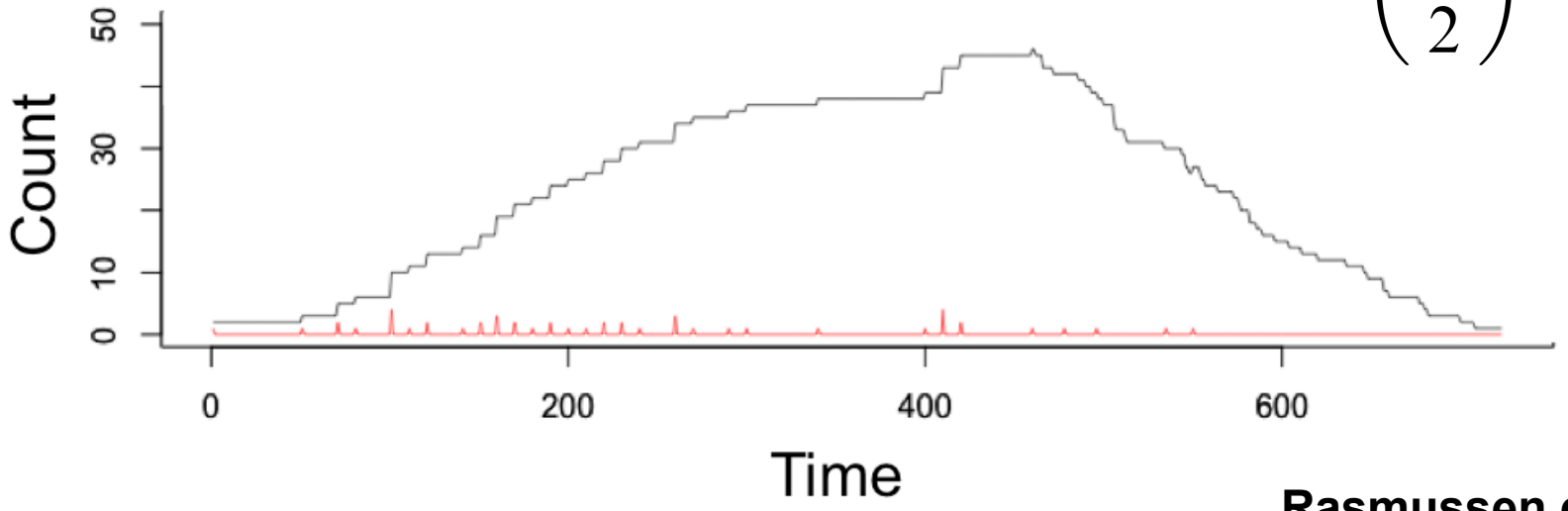
Two paradigms:
- Small outbreaks: who acquires infection from whom?
- Model inference: assume phylogeny is generated by a stochastic transmission process

Common approaches avoid the difficult problem of jointly inferring model and phylogeny by employing **two stages:**
1) estimate a phylogeny from the sequences
2) treating the phylogeny as data, fit the model to the phylogeny using variants of the *coalescent* process or *birth-death* processes to link model and phylogeny

Two-stage methods

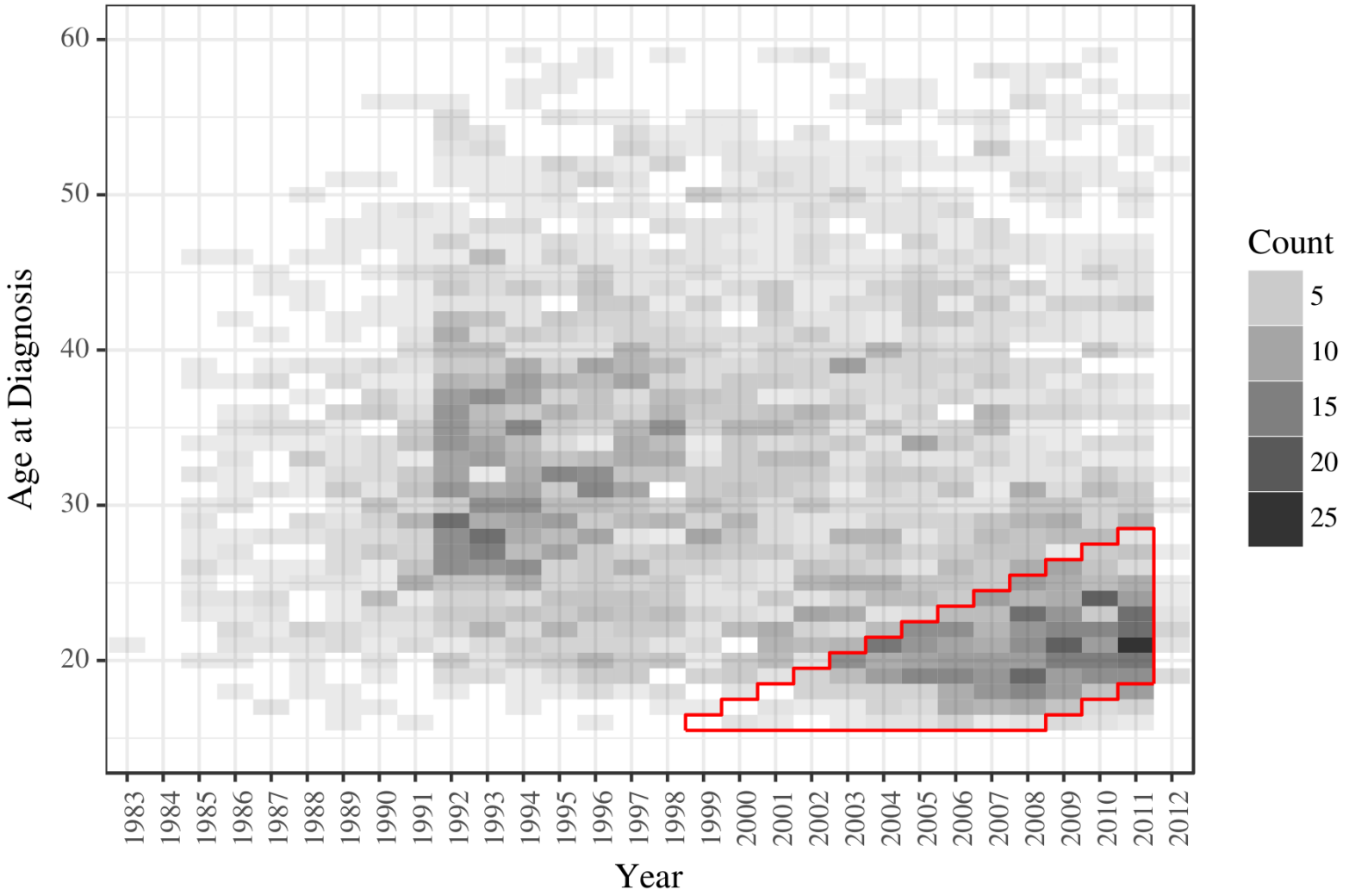


$$\lambda_t = \frac{\binom{i}{2}}{\binom{I_t}{2}} \beta(t) \frac{S_t}{N} I_t$$



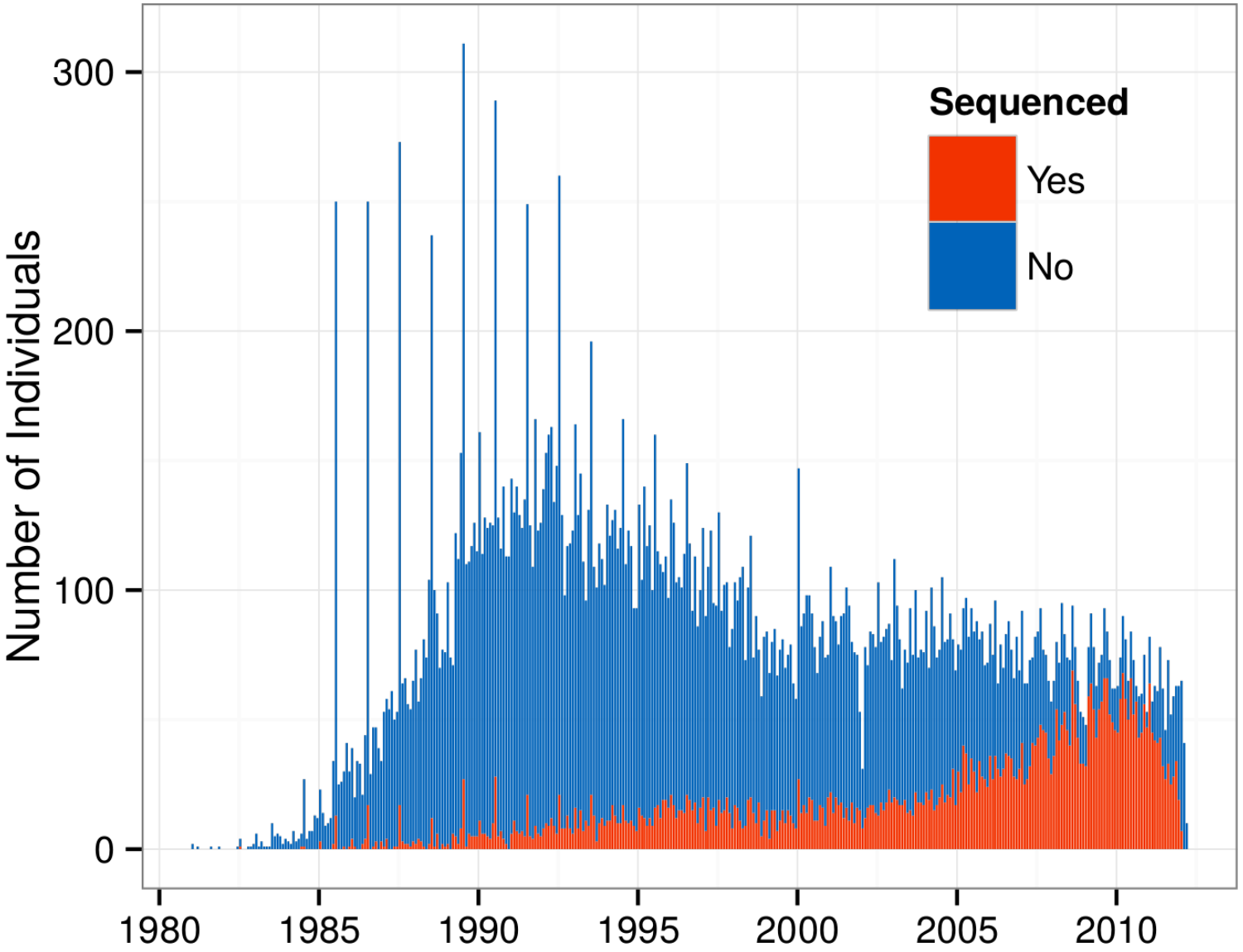
Rasmussen et al. [2011]

Example: HIV among young, black MSM in Detroit



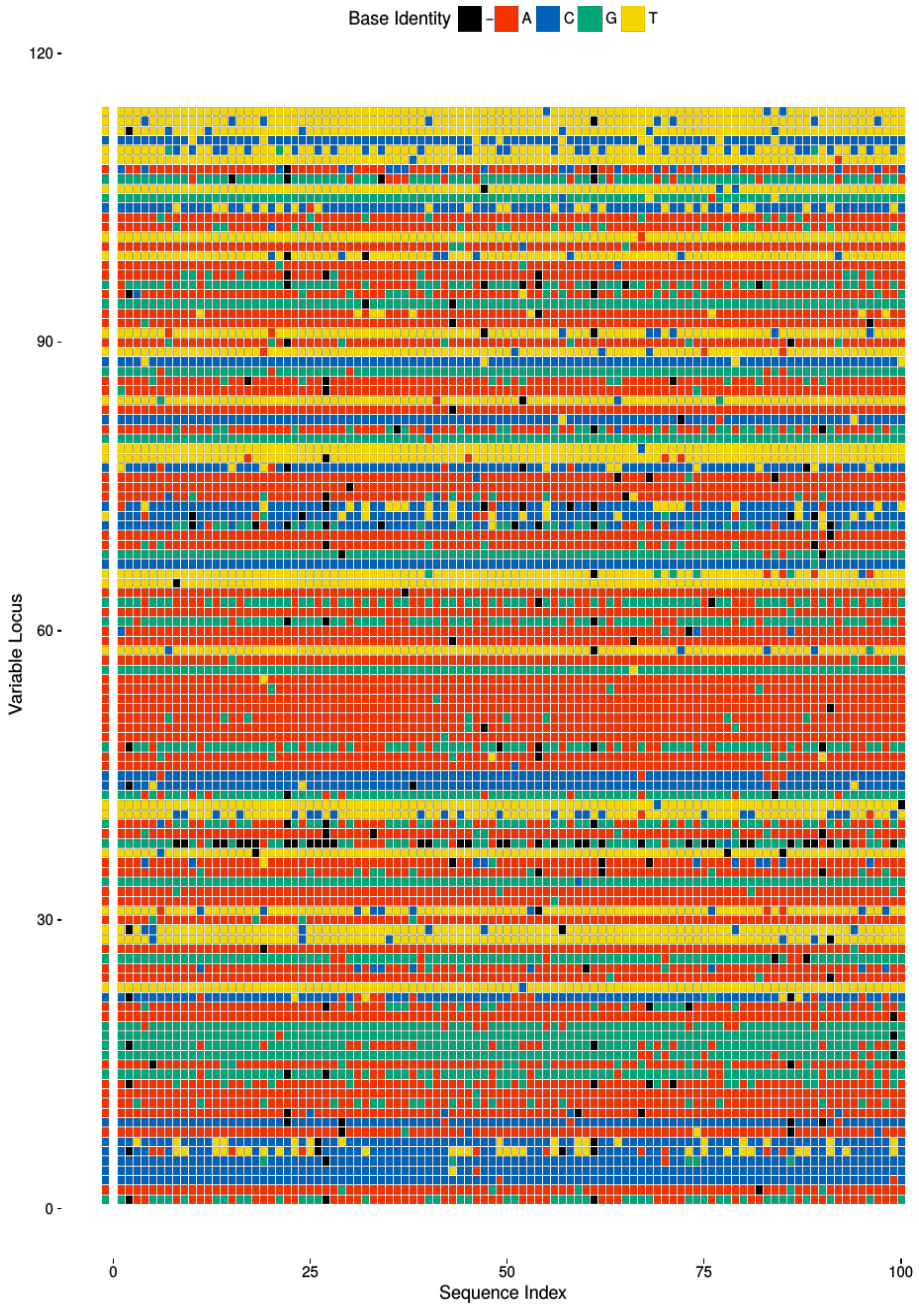
Smith, Ionides and King [2017]

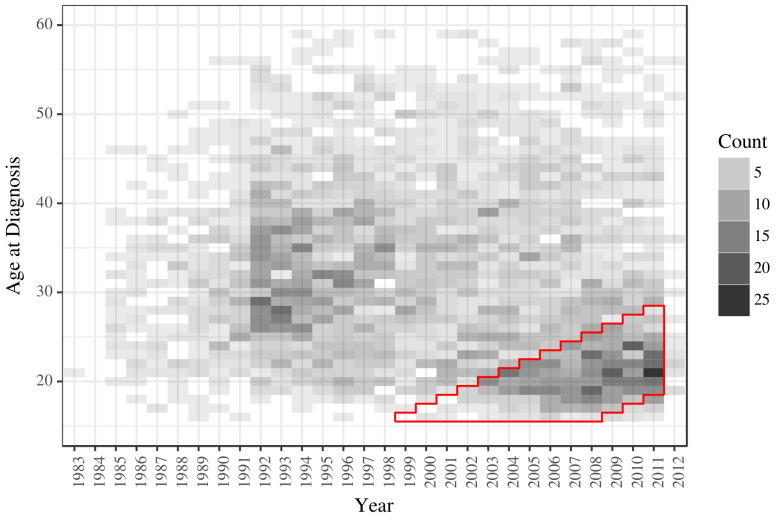
Example: HIV among young, black MSM in Detroit



Smith, Ionides and King [2017]

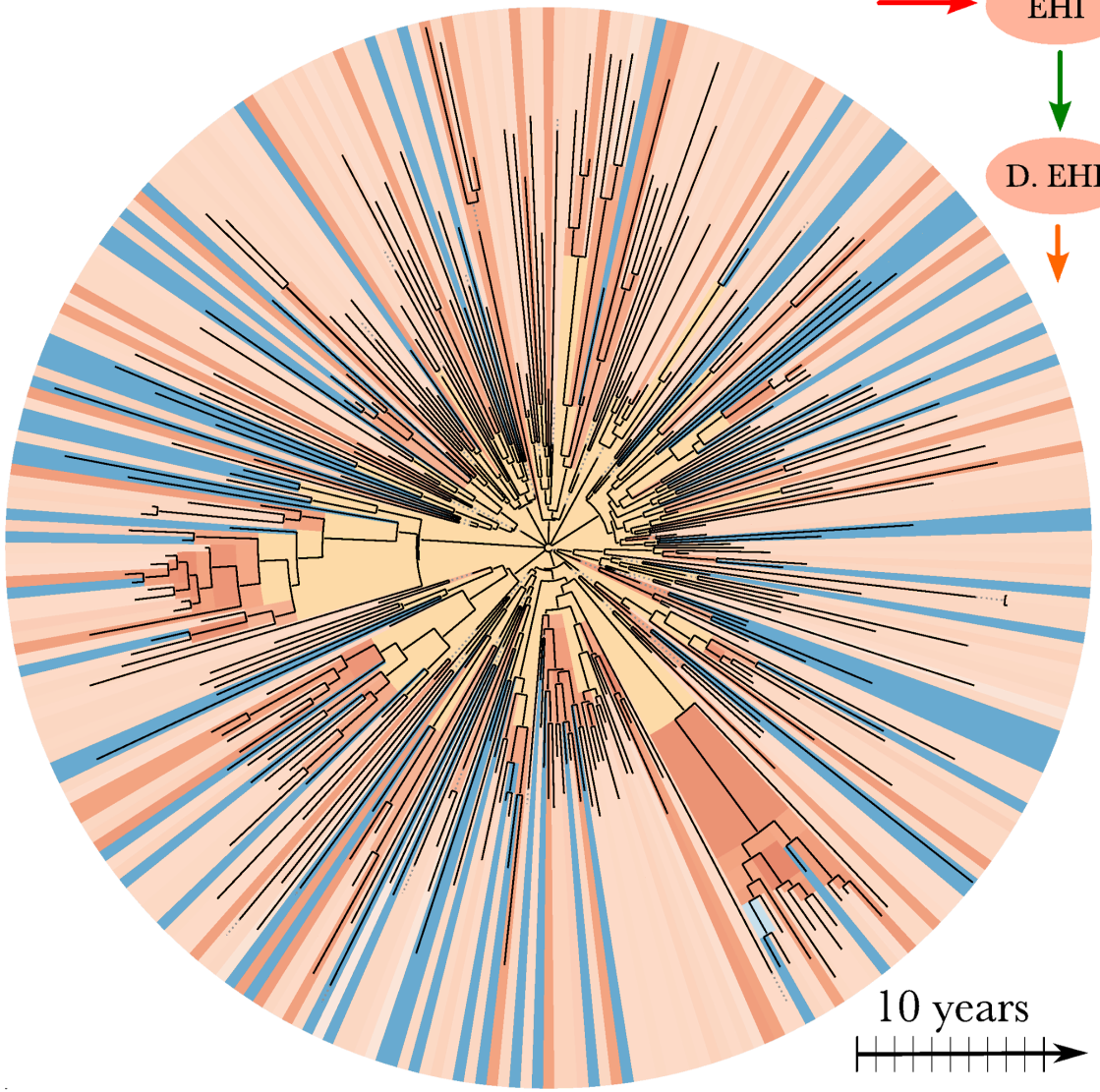
Example: HIV among young, black MSM in Detroit

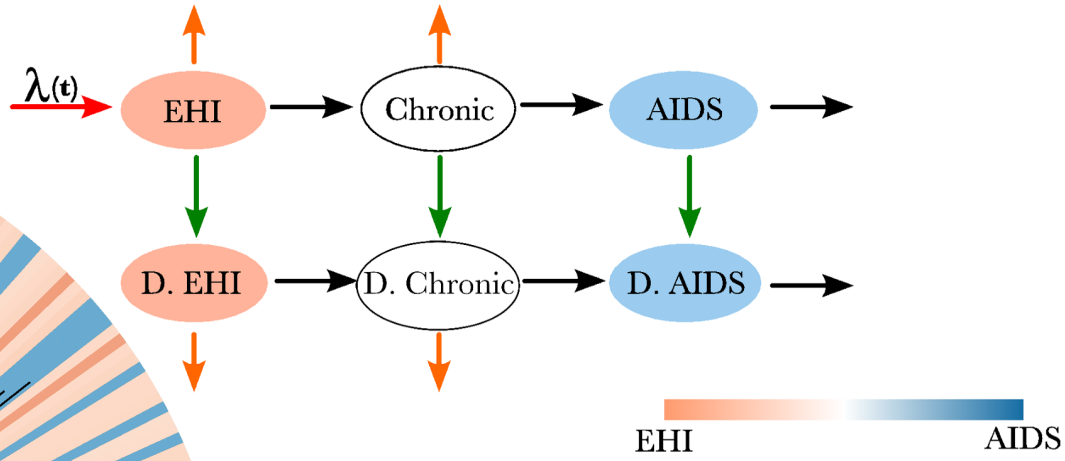




Smith, Ionides and King [2017]

Example: HIV among young, black MSM in Detroit



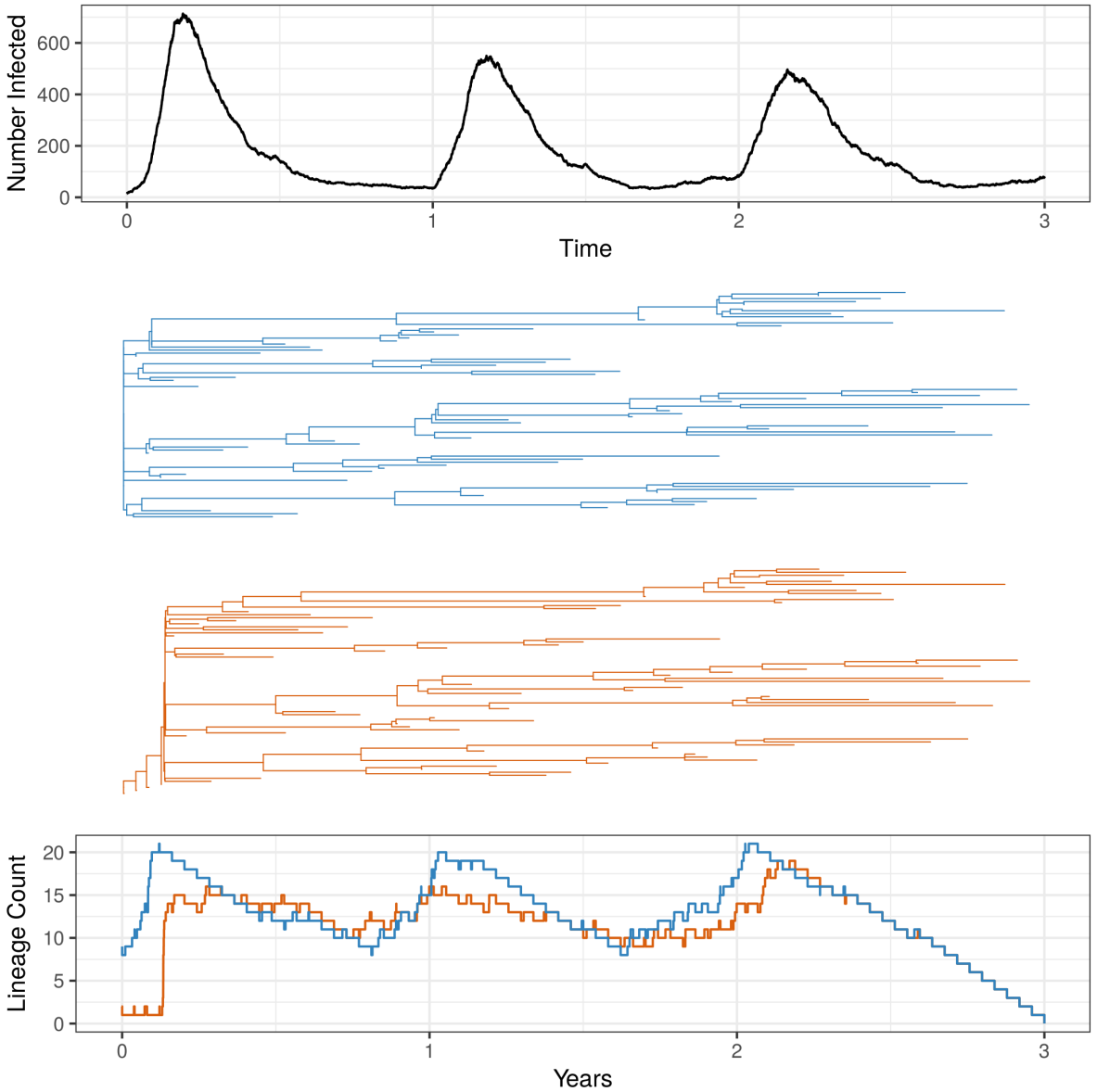


Volz et al. [2013b]

Problems with two-stage methods

- Model used to estimate phylogeny may be *logically inconsistent* with transmission model.
  - This leads to *bias.*
- Methods based on the *coalescent process* are most readily formulated in *backward time* while models for transmission processes can typically only be written at all in *forward time.*
- To get around this, *large population, small sample* assumptions must be made.
- As the models get more complicated (*e.g.,* heterogeneous populations, complex immunity, disease progression, etc.), the *structured coalescent* approaches become unwieldy.

Problems with two-stage methods



$$\lambda_t = \frac{\binom{i}{2}}{\binom{I_t}{2}} \beta(t) \frac{S_t}{N} I_t$$

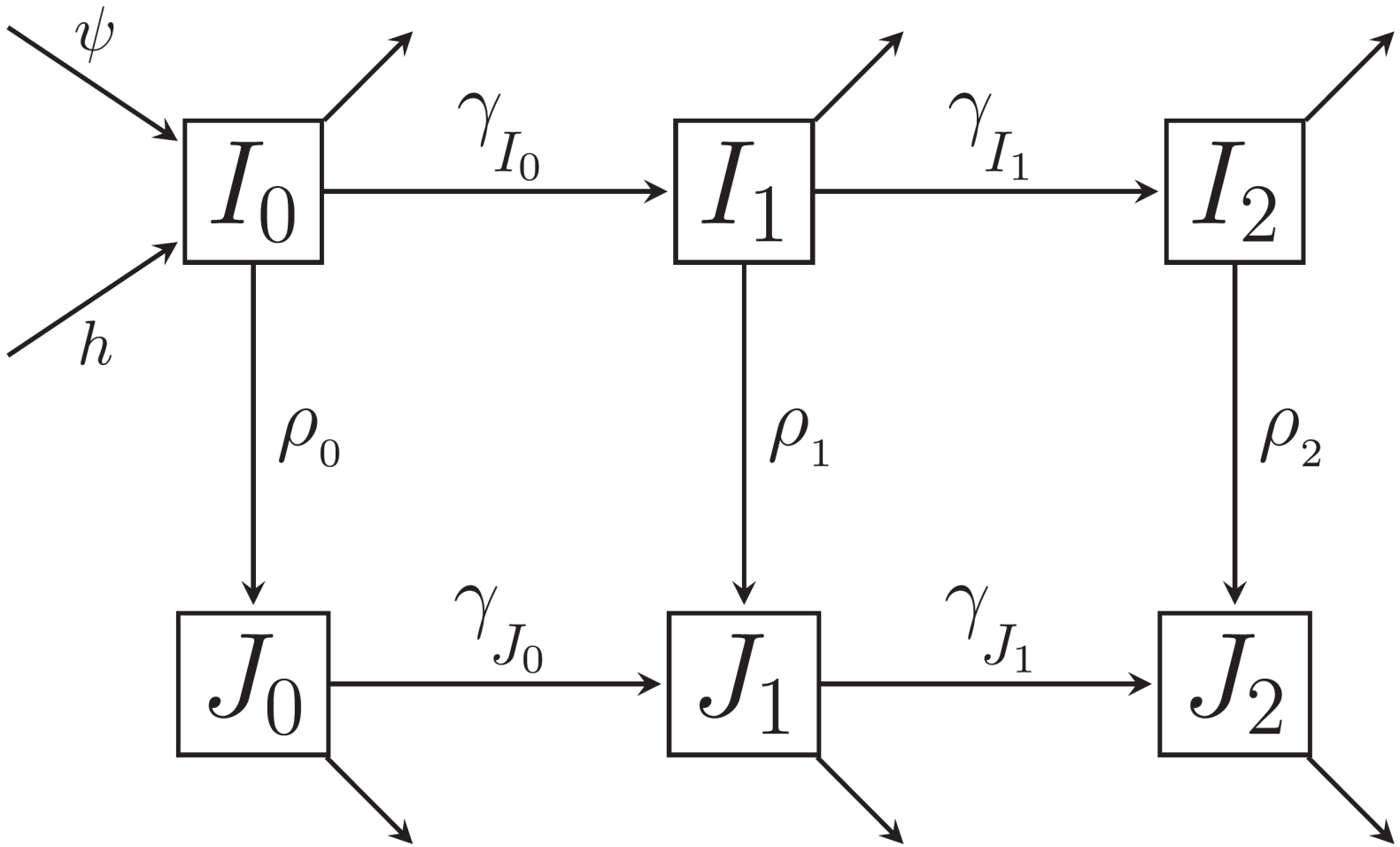
Smith, Ionides, King (in prep)

Phylodynamics done “properly”

We would like to:

- jointly estimate transmission model and phylogeny
- avoid questionable assumptions needed to apply reverse-time likelihoods to forward-time processes
- enjoy the *plug-and-play* property that affords freedom in investigating alternative hypotheses
- a method is plug-and-play if it requires only that one be able to *simulate* from the latent process, *i.e.,* transition densities need not be tractable

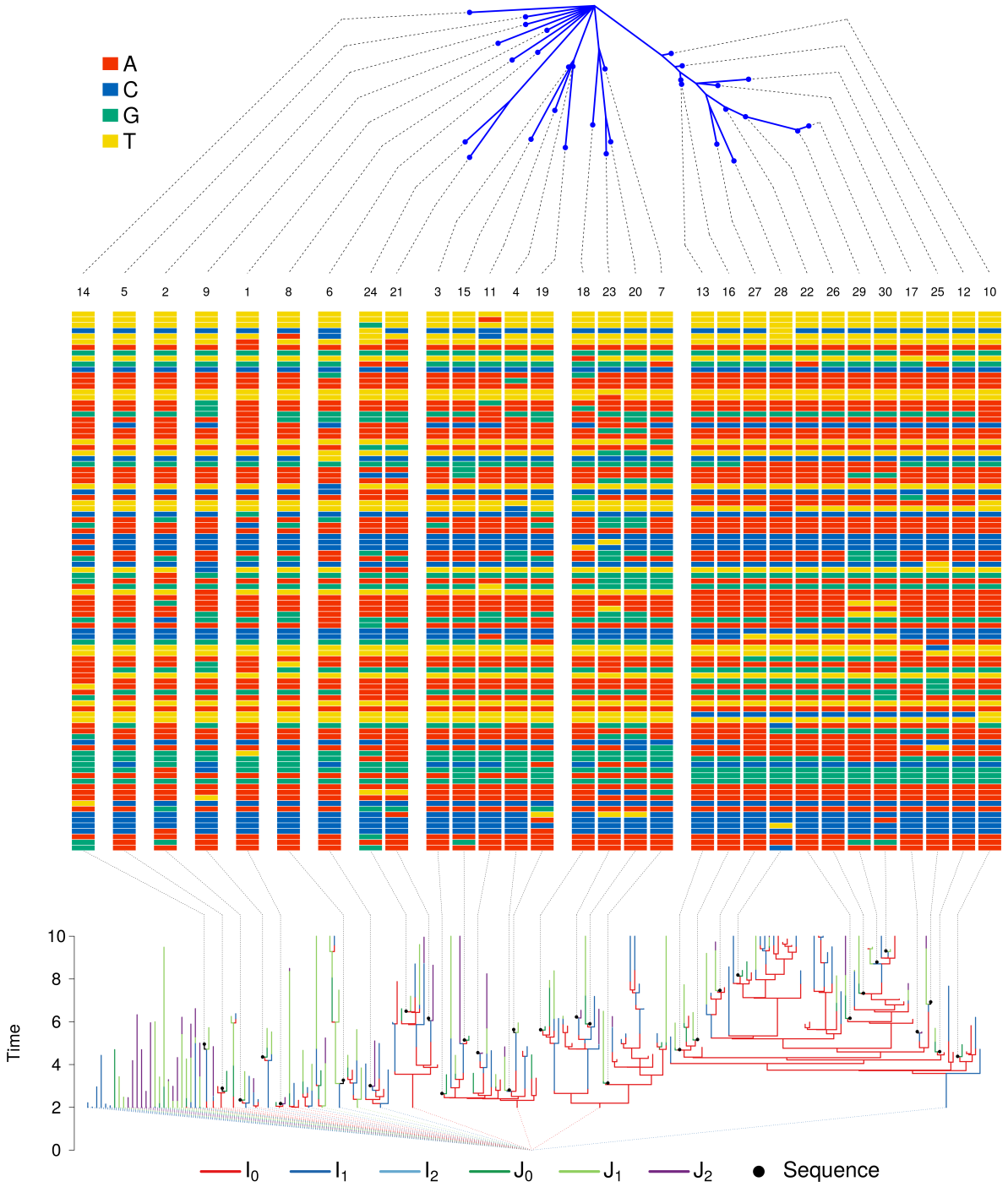
Example: HIV among young, black MSM in Detroit

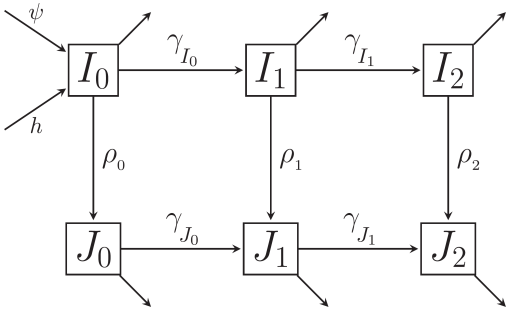


$$h(t) = \varepsilon_{I_0} N_{I_0}(t) + \varepsilon_{I_1} N_{I_1}(t) + \varepsilon_{I_2} N_{I_2}(t) + \varepsilon_{J_0} N_{J_0}(t) + \varepsilon_{J_1} N_{J_1}(t) + \varepsilon_{J_2} N_{J_2}(t)$$

Smith, Ionides, and King [2017];
cf. Volz et al. [2013b]

Ingredients





Smith, Ionides and King [2017]

Key innovations

Several innovations are needed:

1) realization of the process as a *partially observed Markov process* (POMP, AKA state space model)

2) concept of a *growing tree*

3) physical molecular clocks

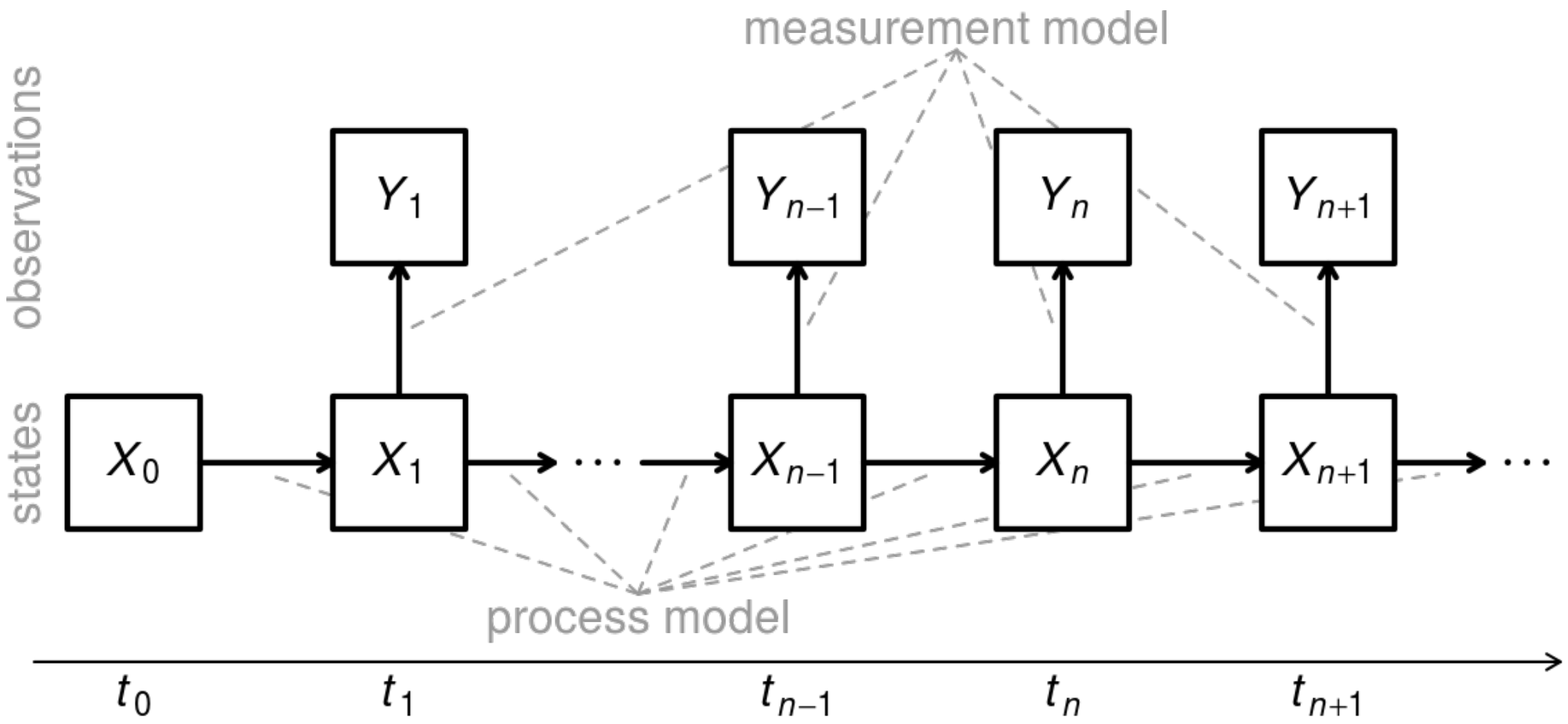
4) just-in-time construction of state variables

5) hierarchical sampling

6) efficient parallelization

Smith, Ionides and King [2017]

Partially observed Markov processes



Partially observed Markov processes

- Data: $y^*_{1:n} = \{y^*_1, \ldots, y^*_n\}$
- Modeled as a realization of a stochastic process $Y_{1:n}$
- Observation times: $t^*_{1:n} = \{t_1, \ldots, t_n\}$
- Latent Markovian state process: $X_{0:n} = \{X(t_0), X(t_1), \ldots, X(t_n)\}$
- Joint density:

$$f_{X_{0:n},Y_{1:n}}(x_{0:n}, y_{1:n}; \theta) = f_{X_0}(x_0; \theta) \prod_{k=1}^{n} f_{X_k|X_{k-1}}(x_k|x_{k-1}; \theta)\, f_{Y_k|X_k,Y_{1:k-1}}(y_k|x_k, y_{1:k-1}; \theta)$$

- Likelihood:

$$\mathcal{L}(\theta) = f_{Y_{1:n}}(y^*_{1:N}; \theta) = \int f_{X_{0:N},Y_{1:N}}(x_{0:N}, y^*_{1:N}; \theta)\, dx_{0:N}.$$

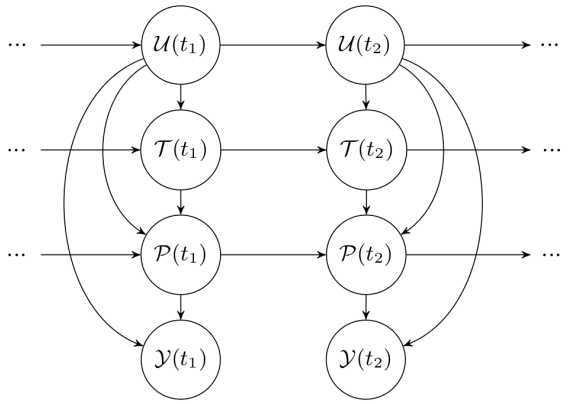
- Factorization:

$$\mathcal{L}(\theta) = \prod_{k=1}^{n} \int f_{Y_k|X_k,Y_{1:k-1}}(y^*_k \mid x_k, y^*_{1:k-1}; \theta)\, f_{X_k|Y_{1:k-1}}(x_k \mid y^*_{1:k-1}; \theta)\, \mathrm{d}x_k$$

Innovation 1: formulation as a POMP

- Data are:
  - Genetic sequences at known sampling times
  - Other information, *e.g.*, diagnoses without sequences
- Latent process: $X(t) = \big(\mathcal{T}(t),\, \mathcal{P}(t),\, \mathcal{U}(t)\big)$
- Transmission forest: $\mathcal{T}(t)$
- Pathogen phylogeny: $\mathcal{P}(t)$
- Auxiliary Markovian process: $\mathcal{U}(t)$

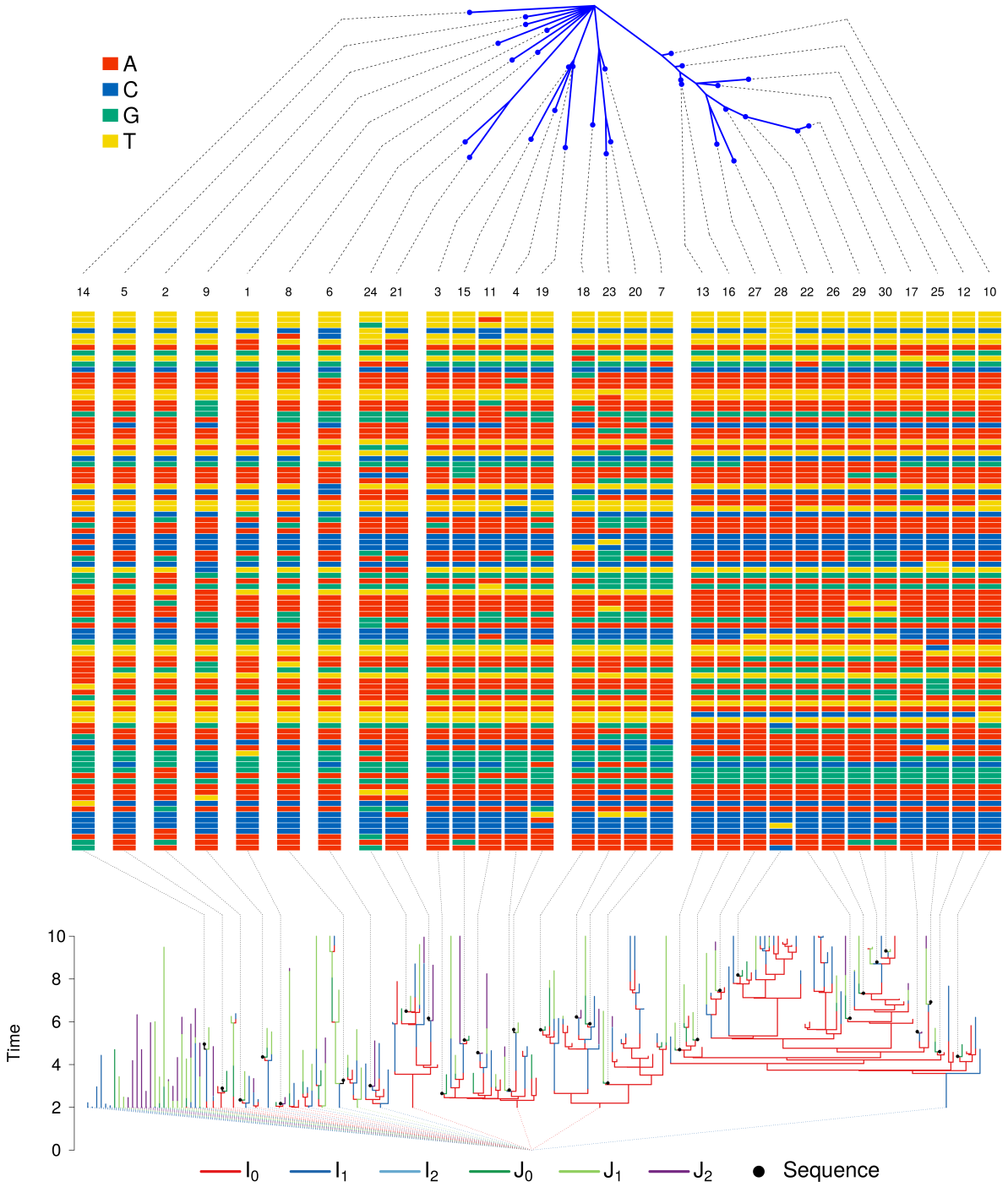
a *GenPOMP*

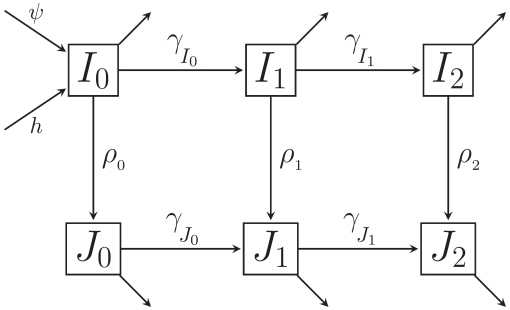


Dependency graph

Smith, Ionides and King [2017]

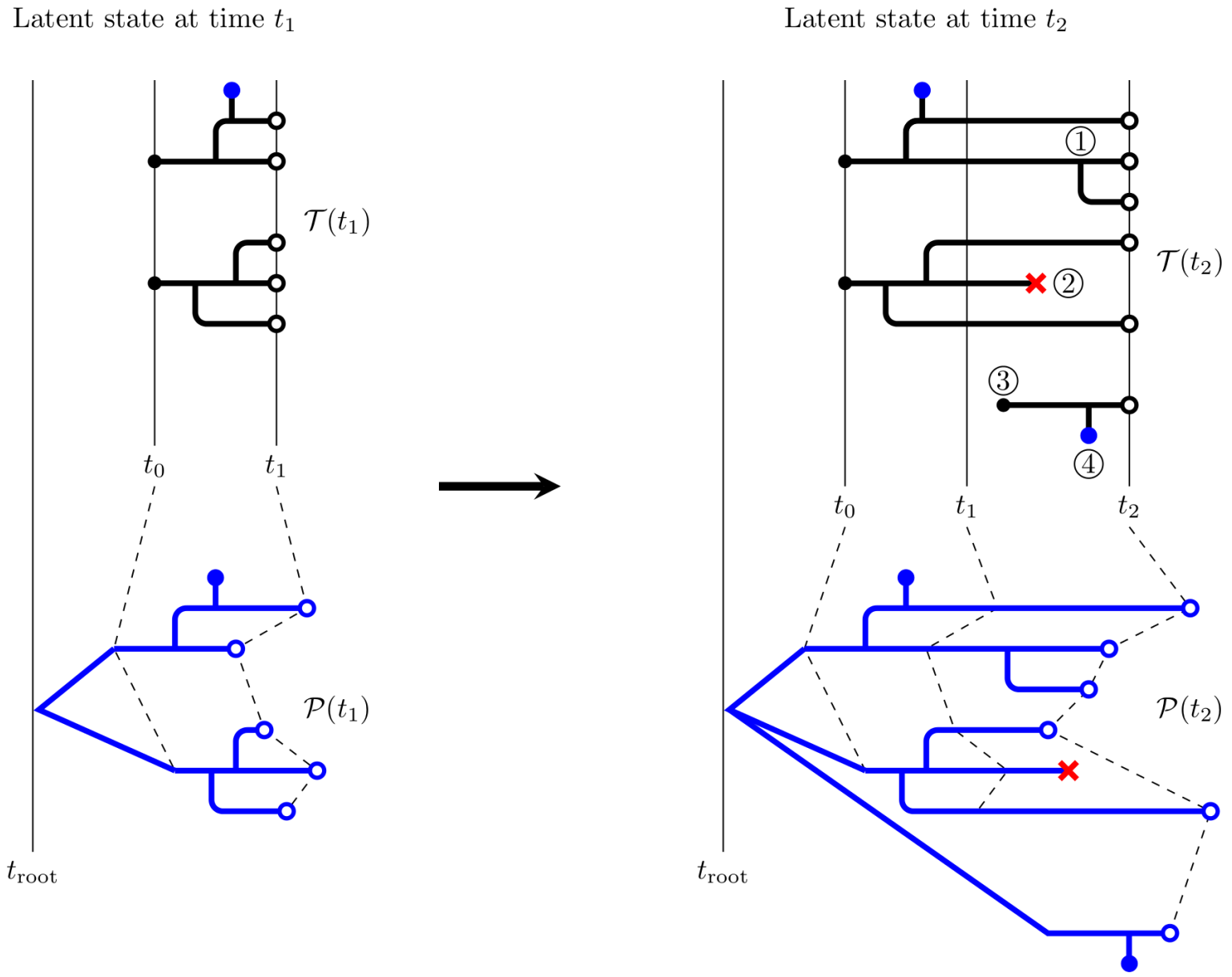
Ingredients



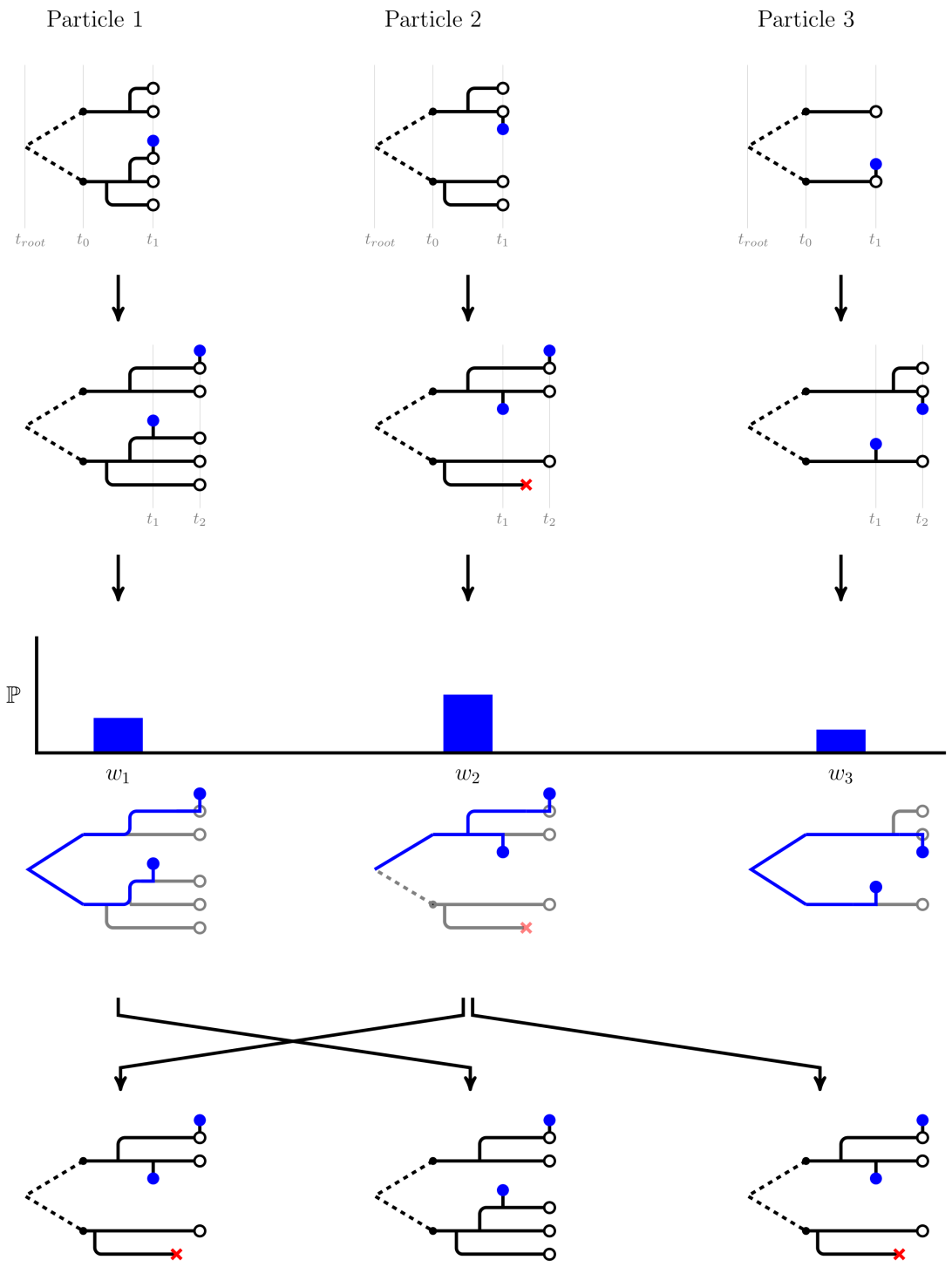


Smith, Ionides and King [2017]

Simulating the latent process



GenSMC: sequential Monte Carlo for a GenPOMP



1. **Proposal**. Simulate particles forward from time t_1 to time t_2. Then select an individual to be sequenced.

2. **Weighting**. Based on the structure of the proposed transmission forest, construct the subtree of the phylogeny that connects the observed sequences. Use this subtree to compute weight of the particle: the conditional probability of the new sequence.

3. **Resampling**. Resample particles with probability proportional to their weights.

Innovation 2: physical relaxed molecular clocks

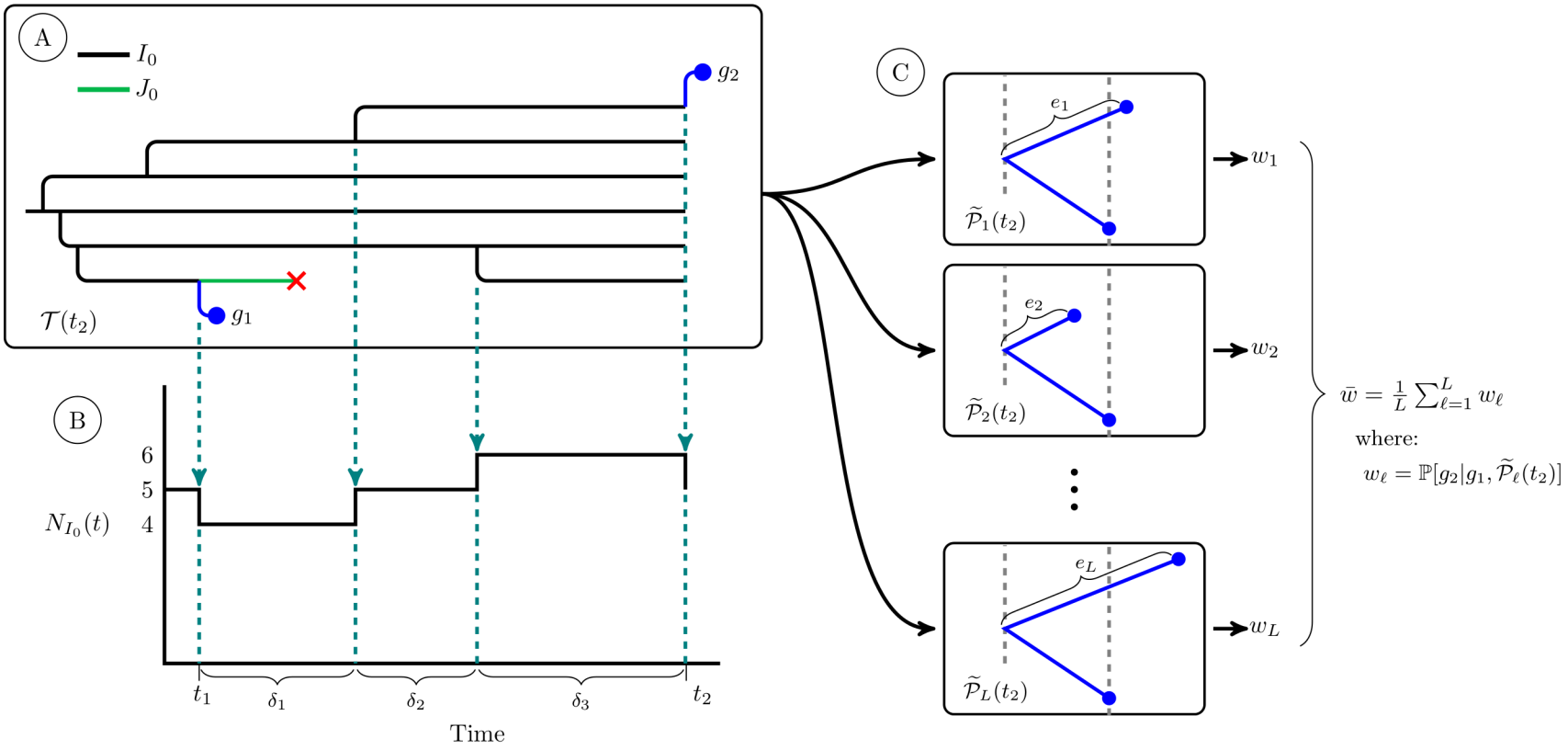
- Strict molecular clocks assume that the rate of evolution is constant through time and the mutation process is Poisson.

- It is commonly necessary to allow for overdispersion in this process, which leads to *relaxed* molecular clocks.

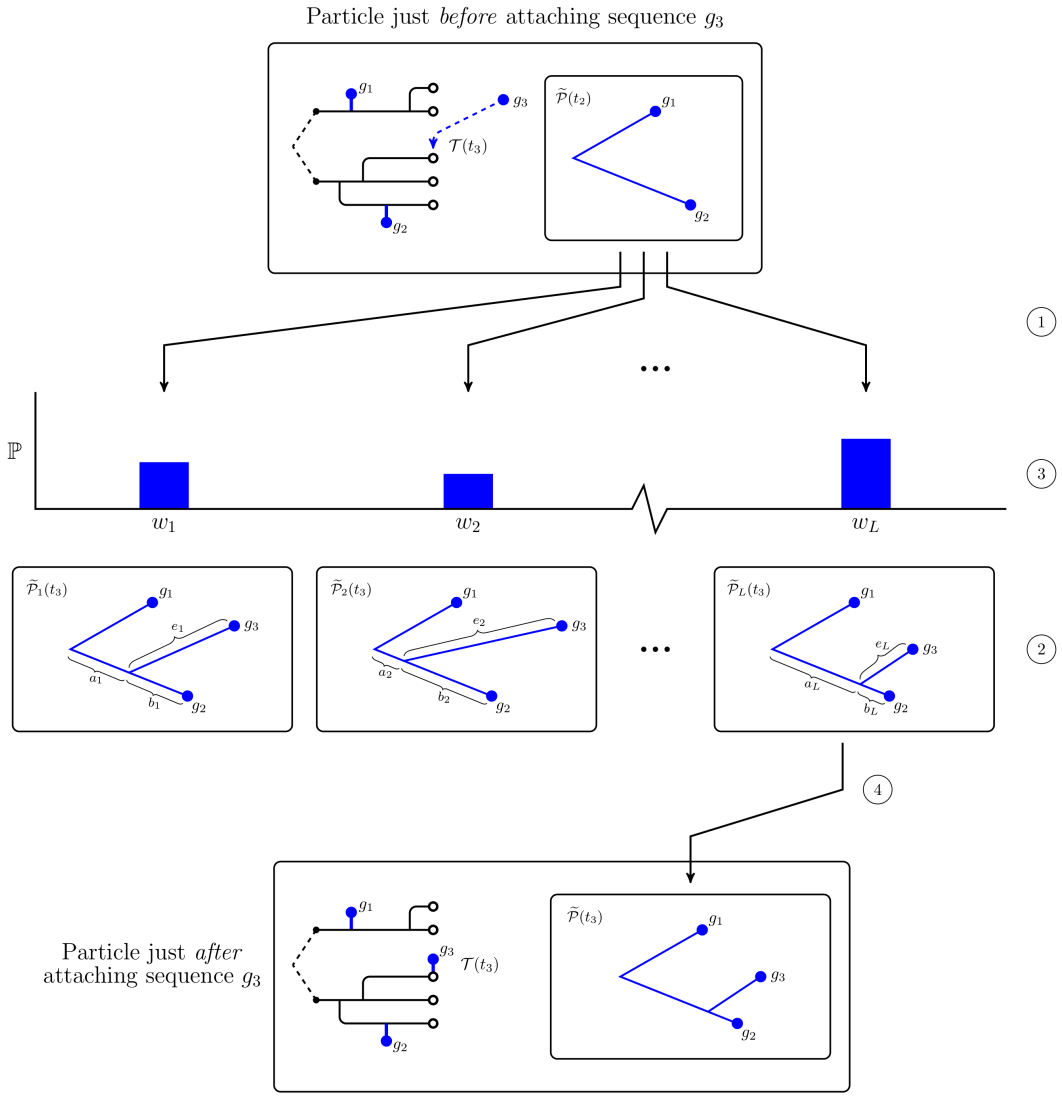
- Most relaxed clocks employed in practice are incompatible with Markovian assumptions.

- We require that the molecular clock is a non-decreasing, continuous-valued Lévy process, *e.g.,* a Gamma clock.

Innovation 2: physical relaxed molecular clocks



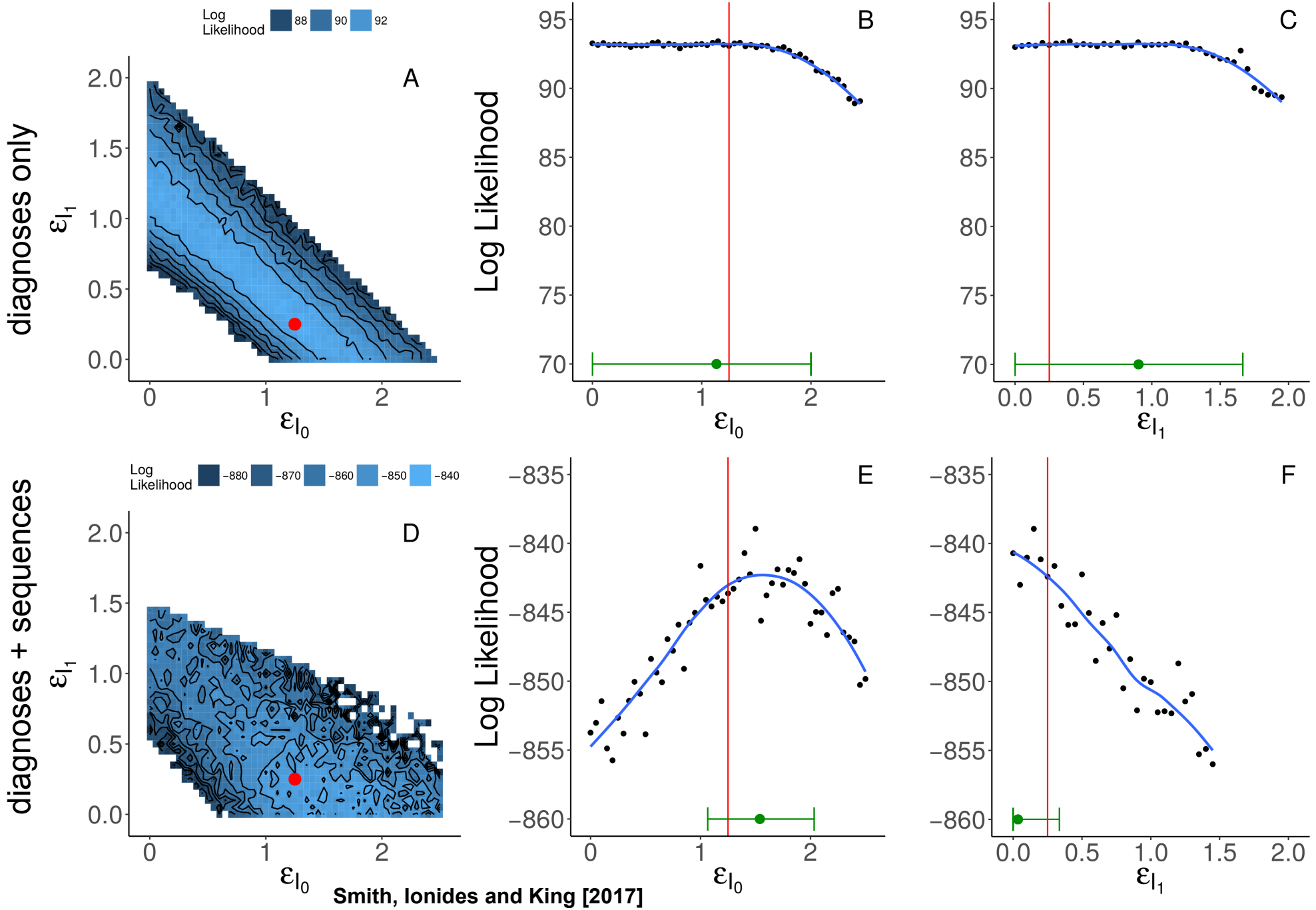
Innovation 2: physical relaxed molecular clocks



Innovation 3: Just-in-time state-variable construction

- The evolutionary process for the sequences goes into the measurement model.
- Formally, a measurement is the assignment of a new sequence to an individual in the transmission tree.
- Evaluating the measurement density involves finding the likelihood of the new sequence given the old sequences and the tree.
- This likelihood is computed efficiently by the Felsenstein peeling (pruning) recursion.
- The high-dimensional pathogen genome need not be included in the latent state.

A simulation study

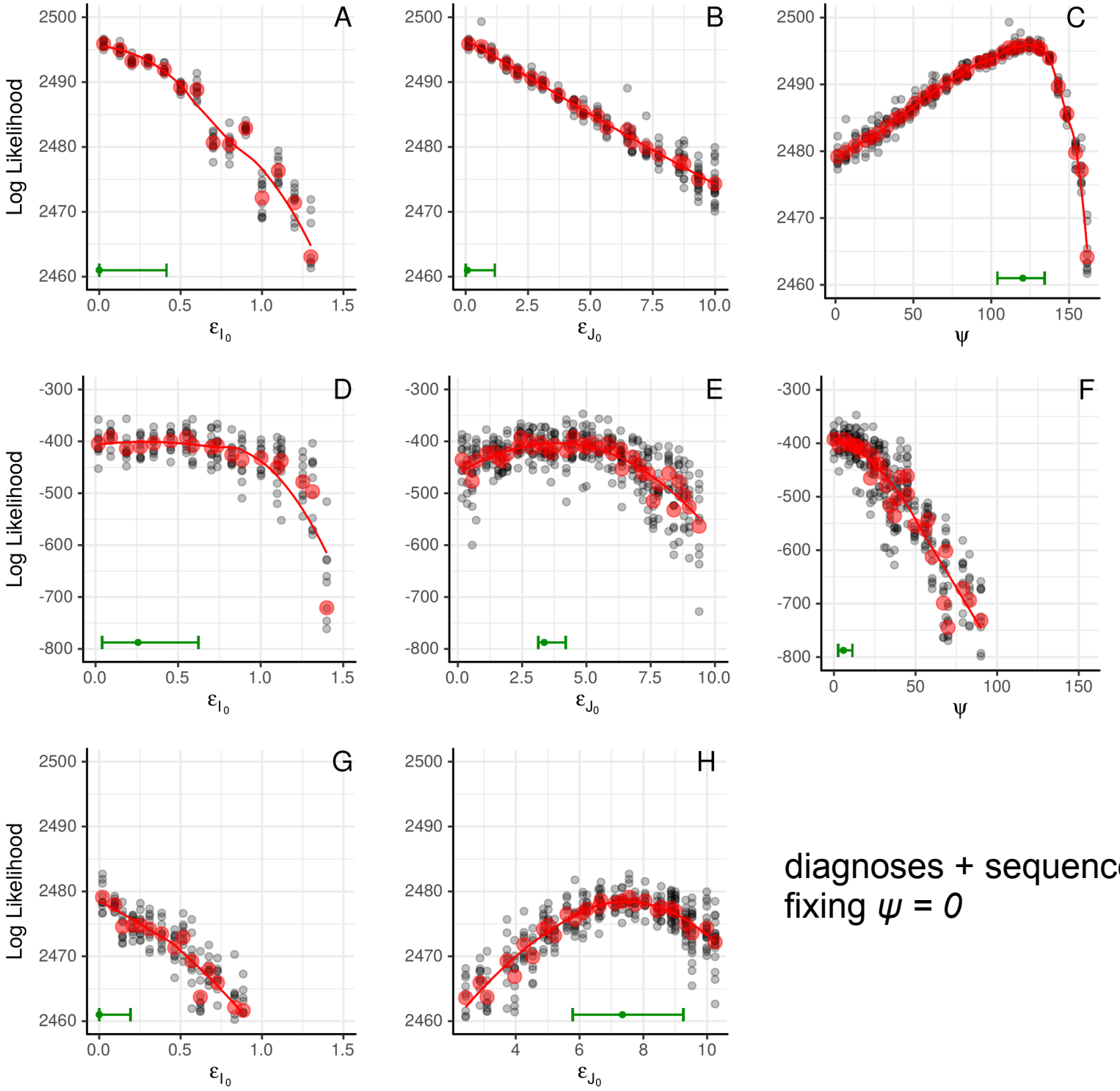


Smith, Ionides and King [2017]

Results of HIV study

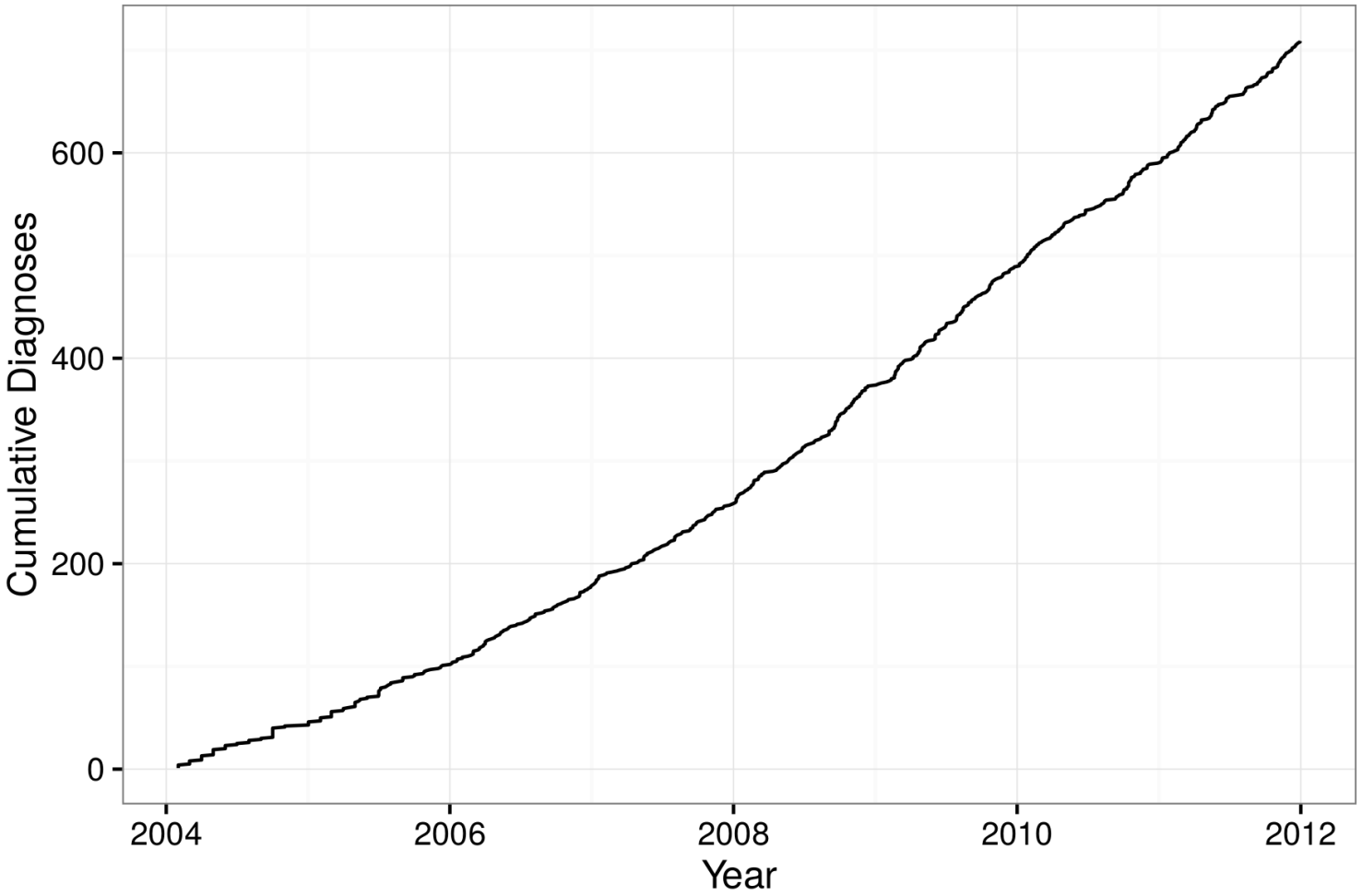
diagnosis only

diagnoses + sequences



diagnoses + sequences, fixing $\psi = 0$

Results of HIV study



Conclusions

- Joint inference is possible with order 10^2 sequences and order 10^3 infections
- We are continuing to investigate how the algorithms scale, but further work is needed to scale to much larger problems
- Being able to compute (even noisy) estimates of the likelihood is useful, to evaluate bias and loss of information in other methods
- Simulation-based methods can reveal modeling errors hidden by other methods
- A promising arena for these approaches is hospital infections

References

Grenfell, B. T.; Pybus, O. G.; Gog, J. R.; Wood, J. L. N.; Daly, J. M.; Mumford, J. A. and Holmes, E. C. (**2004**). *Unifying the epidemiological and evolutionary dynamics of pathogens*, Science 303 : 327-332.

Rasmussen, D. A.; Boni, M. F. and Koelle, K. (**2014a**). *Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam*, Molecular Biology and Evolution 31 : 258-271.

Rasmussen, D. A.; Ratmann, O. and Koelle, K. (**2011**). *Inference for nonlinear epidemiological models using genealogies and time series.*, PLoS Computational Biology 7 : e1002136.

Rasmussen, D. A.; Volz, E. M. and Koelle, K. (**2014b**). *Phylodynamic inference for structured epidemiological models*, PLoS Computational Biology 10 : e1003570.

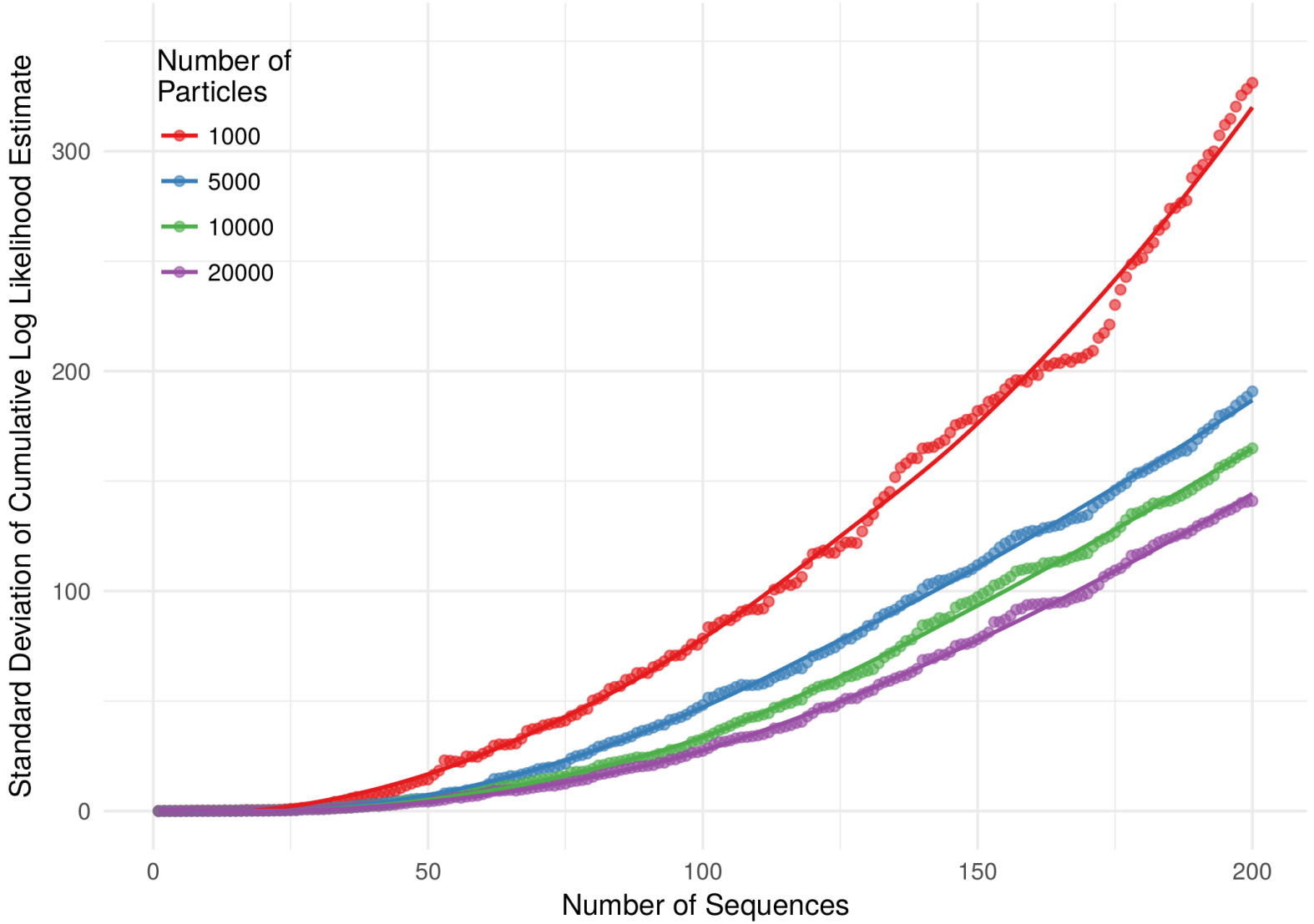
Smith, R. A.; Ionides, E. L. and King, A. A. (**2017**). *Infectious disease dynamics inferred from genetic data via sequential Monte Carlo*, Molecular Biology and Evolution 34 : 2065-2084.

Volz, E. M. (**2012**). *Complex population dynamics and the coalescent under neutrality*, Genetics 190 : 187-201.

Volz, E. M.; Ionides, E.; Romero-Severson, E. O.; Brandt, M.-G.; Mokotoff, E. and Koopman, J. S. (**2013b**). *HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis*, PLoS Medicine 10 : e1001568.

Volz, E. M.; Koelle, K. and Bedford, T. (**2013a**). *Viral phylodynamics*, PLoS Computational Biology 9 : e1002947.

Volz, E. M.; Kosakovsky Pond, S. L.; Ward, M. J.; Leigh Brown, A. J. and Frost, S. D. W. (**2009**). *Phylodynamics of Infectious Disease Epidemics*, Genetics 183 : 1421-1430.


Number of Particles
1000
5000
10000
20000
Standard Deviation of Cumulative Log Likelihood Estimate
300
200
100
0
0
50
100
150
200
Number of Sequences