Infectious disease dynamics inferred from genetic data via sequential Monte Carlo

Edward Ionides
University of Michigan, Department of Statistics

Based on an article to appear in *Molecular Biology and Evolution* with coauthors R. Alex Smith and Aaron King.

Lecture 4 at Wharton Statistics Department
Tuesday 9th May, 2017

Slides are online at
`http://dept.stat.lsa.umich.edu/~ionides/talks/upenn`

Epidemiology via pathogen genetic sequences

- Growing amounts of genetic sequence data on pathogens should assist the study of infectious disease dynamics.
- Using this information effectively is a methodological challenge.
- The POMP/SMC approach enables general, statistically efficient, plug-and-play methodology that jointly estimates both disease transmission parameters and a phylogeny.
- We demonstrate the feasibility of our approach through simulation and apply it to estimate stage-specific infectiousness in a subepidemic of HIV in Detroit, Michigan.
- These methods may be applied in other biological systems where one seeks to infer population dynamics from genetic sequences, and they may also find application for evolutionary models with phenotypic rather than genotypic data.

Phylodynamics

"The study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies."

- From `en.wikipedia.org/wiki/Viral_phylodynamics`, paraphrasing Grenfell et al. (2004) who coined the word phylodynamics.
- "Phylodynamics" is also used to describe the inverse problem, using phylogenetic trees to infer dynamic mechanisms.
- Data are typically genetic sequences (partial or complete) on viral or bacterial pathogens.

Statistical methods for phylodynamics

- For successful phylodynamic inference, mechanisms of transmission must leave their signature in genetic sequences.
- This occurs when pathogen transmission and evolution occur on similar timescales.
- By explicitly relating models of disease dynamics to their predictions with respect to pathogen sequences, it is possible to estimate aspects of the mechanisms of transmission (reviewed by Frost et al., 2015).
- Most phylodynamic inference methods proceed in three stages.
  1. Estimate the pathogen phylogeny using sequence data.
  2. Fit models of disease dynamics to properties of the pathogen phylogeny, such as coalescent times or summary statistics on the tree.
  3. Assess robustness of the results to variation in the estimated phylogeny to account for phylogenetic uncertainty.
- It may happen that population dynamics, as estimated by the transmission model, are inconsistent with those assumed when estimating the phylogeny.

Previous uses of SMC for phylodynamic inference

- SMC techniques have previously been used for inferring phylogenies (Bouchard-Côté et al., 2012), and for phylodynamic inference conditional on a phylogeny (Rasmussen et al., 2011).
- These approaches avoid the high-dimensional, computationally challenging problem of joint inference through forward-in-time simulation of tree-valued processes.
- Several innovations were necessary to realize computationally feasible SMC on models and datasets of scientific interest.
  - Dimension reduction: constructing the POMP model with genetic sequences only in the measurement model to reduce the dimension of the latent variables.
  - Algorithm parallelization.
  - Hierarchical sampling.
  - Just-in-time construction of state variables.
  - Restriction to a class of physical molecular clocks.
  - Maximization of the likelihood using iterated filtering.
- These innovations have some general relevance to SMC-based inference for large and complex systems.

Dimension reduction

"A basic principle (or rule of thumb) in Monte Carlo computation: One should carry out analytical computation as much as possible." *Monte Carlo Strategies in Scientific Computing* (Liu, 2001, Section 2.3).

- A standard SMC variance reduction technique is to solve conditionally linear and Gaussian parts of the model by linear algebra.
- We have considerable flexibility how to put a model into the POMP framework, and we can sometimes choose to place parts of the model in either the measurement or the state process.
- Vanilla SMC requires tractability of the measurement density and does not scale well with dimension of the latent process. If possible, move tractable parts of the model into the measurement process.

Dimension reduction: Rewriting the measurement model

- Here, we can put the genetic sequences into the measurement model.
  - Formally, let a measurement consist of an assignment of a new sequence to an individual in the transmission tree.
  - The measurement density involves finding the likelihood of the new sequence given the old sequences and the tree. This likelihood can be computed efficiently by "peeling," the name given to the basic POMP identities on a tree.
  - Particles representing the latent process do not have to include the high-dimensional pathogen genome.

Parallelization

Every modern computer is designed for parallel computing. The only question is whether your algorithm is designed for it!

"Compute or concede," according to the IMS Presidential Address, "Let us own data science" by Yu (2014).

- We expect to have multiple **nodes** each with multiple **cores**. Cores within a node communicate quickly and share fast access memory.
- **Embarrassingly parallel** algorithms require communication between processes only to spawn jobs and collect results.
- Profile likelihood calculations are embarrassingly parallel
- If you expect to compute at least as many profile points as you have cores available, you probably should not parallelize the rest of your code - you have more to lose than to gain from this.
- Don't be embarrassed to use embarrassingly parallel algorithms!
- For Monte Carlo methods, there is no subsitute for replication!
- A Monte Carlo computation is an experiment. Fisher's design principles (comparisons, replication, randomization, blocking, orthogonality, factorial experiment) all apply.

Within-node parallelization

- An algorithm is **Within-node parallel** if it is embarrassingly parallel between nodes but uses communication and shared memory within nodes).
- This is appropriate if you have a number of nodes comparable to the number of profile points.
  - At a minimum, say, 3 replicates at each of 20 profile points = 60 nodes.
- What to do if you don't have that many nodes, but for speed reasons you think you need to parallelize?
- Remember Fisher: if you don't have resources to replicate you probably shouldn't do that experiment!
- Cautionary tale: To a first approximation, one imagines that all cores on a node have equal access to all the shared memory. This isn't true! Cores on the same socket have equal access, but large nodes have multiple sockets.
- `genPomp` uses within-node parallelization to investigate many alternatives for how a new genetic sequence attaches to an existing phylogeny.

Hierarchical sampling

- SMC algorithms are analogous to sample surveys. Both are based on importance sampling.
- For sample surveys, a cluster sample (choosing a sample of cities, then sampling within those cities) is often a practical decision. This is because of communication costs.
- Communication costs are also a primary consideration in parallelization of code.
- Here, we say **hierarchical sampling** instead of cluster sampling, but the two are closely related.
- Parallelization of SMC is a research topic (Paige et al., 2014). Hierarchical sampling is one simple approach.

Just-in-time variables

- The **relaxed clock** allows evolutionary branch lengths flexibility to match genetic data.
- The relaxed clock for each individual must be part of the latent process, even though the evolutionary process is placed in the measurement model. Why?
- Relaxed clock variables on a branch of the full phylogenetic tree (including all hosts, whether or not they are diagnosed or their pathogens are sequenced) are not needed until that branch is included in the observed phylogenetic tree.
- There is some direct saving from avoiding computation of parts of the latent process that may never be observed or have any effect on an observable quantity.
- Perhaps more importantly, postponing assignment of evolutionary clock times until they are informed by data gives variance reduction.

Restriction to a class of physical molecular clocks

- Commonly used relaxed molecular clock specifications are hard to reconcile with underlying evolutionary dynamics.
- For example, log normal clock perturbations lack an additivity property: adding a node to split a branch must change the evolutionary process along that branch.
- We insist that the relaxed clock is part of a Markovian latent process, preventing such issues.
- We require the relaxed clock to be a non-decreasing continuous-valued Lévy process. In practice, we use a Gamma process clock.

Comments on variance reduction

- Variance reduction may sound boring and tedious work.

Comments on variance reduction

- Variance reduction may sound boring and tedious work.
- Often, it is.

Comments on variance reduction

- Variance reduction may sound boring and tedious work.
- Often, it is.
- Parallelization has more technological excitement but also adds to coding and debugging complexity.

Comments on variance reduction

- Variance reduction may sound boring and tedious work.
- Often, it is.
- Parallelization has more technological excitement but also adds to coding and debugging complexity.
- Sometimes, numerical tractability issues necessitate and inspire theoretical advances.

Comments on variance reduction

- Variance reduction may sound boring and tedious work.
- Often, it is.
- Parallelization has more technological excitement but also adds to coding and debugging complexity.
- Sometimes, numerical tractability issues necessitate and inspire theoretical advances.
- It is almost magical when we can carry out inference by accessing the likelihood surface relating a complex dynamic model to big data.

The latent process for a GenPOMP

- The latent Markov process, $\{X(t),\ t \in \mathbb{T}\}$, with $\mathbb{T} = [t_0,\ t_{\text{end}}]$, models the population dynamics and also includes any other processes needed to describe the evolution of the pathogen.
- Suppose we can write $X(t) = \big(\mathcal{T}(t),\ \mathcal{P}(t),\ \mathcal{U}(t)\big)$, where
  - $\mathcal{T}(t)$ is the *transmission forest*,
  - $\mathcal{P}(t)$ is the *pathogen phylogeny* equipped with a relaxed molecular clock,
  - $\mathcal{U}(t)$ represents the state of the pathogen and host populations.
- For example, $\mathcal{U}(t)$ may categorize each individual in the host population into classes representing different stages of infection.
- We suppose that $\{\mathcal{U}(t), t \in \mathbb{T}\}$ is itself a Markov process.

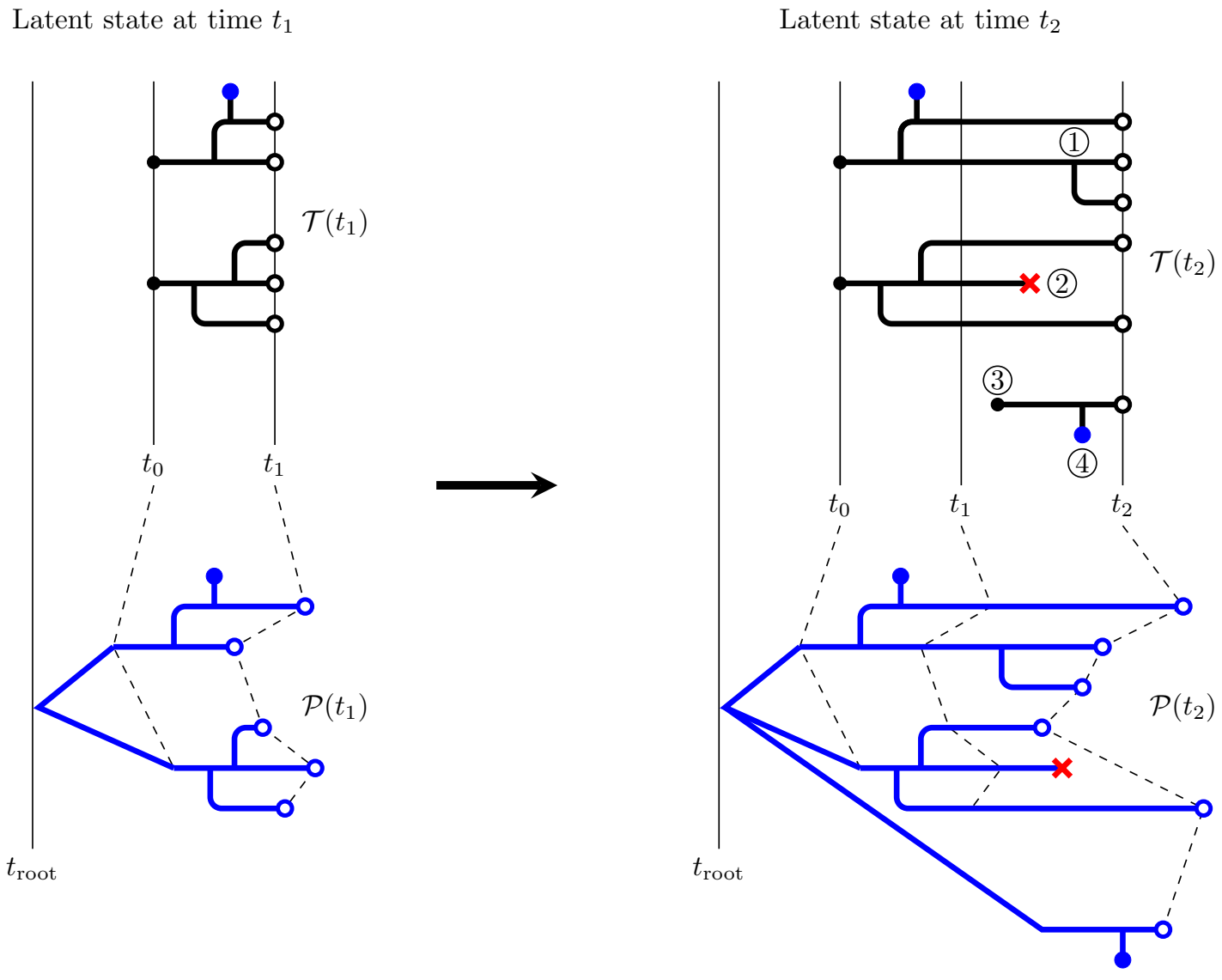
The transmission forest

- The transmission forest represents the history of transmission among hosts.
- Nodes in $\mathcal{T}(t)$ are time-stamped and of several types.
  Internal nodes represent transmission events.
  Terminal nodes are of three types:
  1. **active nodes** represent infections active at time t;
  2. **observed nodes** correspond to diagnosis events, possibly associated with genetic sequences;
  3. **dead nodes** correspond to death or emigration events.
- Root nodes at time t_0 correspond to infections present in the initial population.
- Root nodes at times $t > t_0$ correspond to immigration events.
- Since all nodes are time-stamped, edges of $\mathcal{T}(t)$ have lengths measured in units of calendar time.

The pathogen phylogeny

- The pathogen phylogeny $\mathcal{P}(t)$ represents the history of divergences of pathogen lineages.
- Internal nodes of $\mathcal{P}(t)$ represent branch-points of pathogen lineages, which, we assume, coincide with transmission events.
- The terminal nodes of $\mathcal{P}(t)$ are in 1-1 correspondence with the terminal nodes of $\mathcal{T}(t)$.
- The distinction between $\mathcal{P}(t)$ and $\mathcal{T}(t)$ allows for random variation in the rate of molecular evolution.
  - A model for sequence evolution is called a **molecular clock**
  - Data are over-dispersed. Including overdispersion gives a **relaxed molecular clocks**.
- The edge lengths of $\mathcal{T}(t)$ measure calendar time between events, whereas edge lengths in $\mathcal{P}(t)$ can have additional random variation describing non-constant rates of evolution.

Schematic diagram: advancing a GenPOMP from t_1 to t_2



- Black: transmission forest, $\mathcal{T}(t)$. Blue: pathogen phylogeny, $\mathcal{P}(t)$.

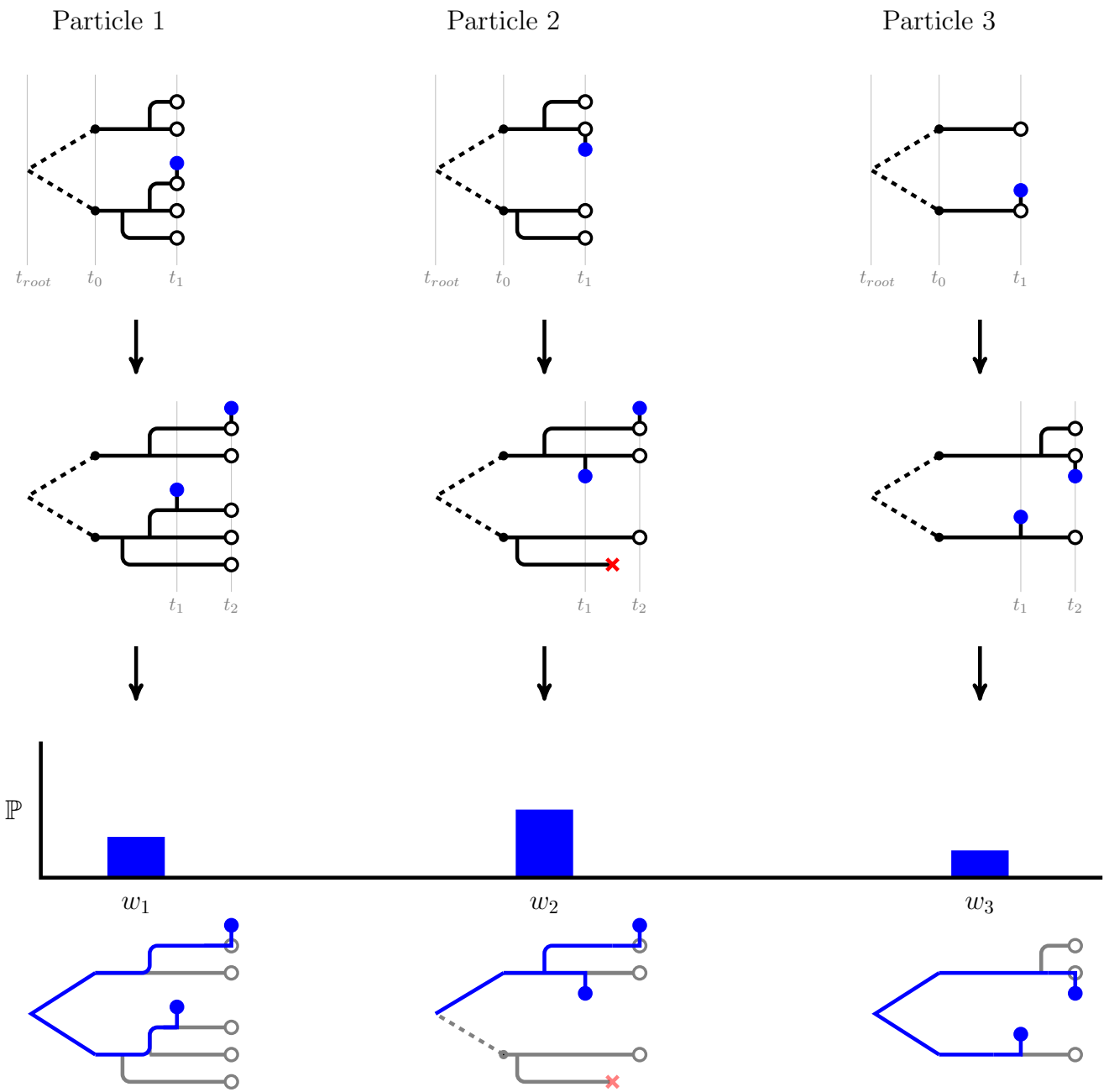
Annotations for the GenPOMP schematic diagram

- The branching pattern of the pathogen phylogeny mirrors that of $\mathcal{T}(t)$ over the interval $[t_0, t_1]$, so pathogen lineages are assumed to branch exactly at transmission events. This simplifying assumption can be changed.
- Randomness in the rate of evolution—a relaxed molecular clock—results in random edge lengths in $\mathcal{P}(t)$.
- At ①, an active node splits in two when a transmission event occurs.
- At ②, an active node becomes a dead node ($\times$) when an infected host emigrates, recovers, or dies.
- At ③, an immigration event gives rise to a new active node with its own root.
- At ④, a sequence node ($\bullet$) is spawned when a sample is taken.

More on relaxed molecular clocks

- A strict molecular clock models the rate of evolution as constant through time and across lineages, assuming (i) sequence evolution is Markovian; (ii) no simultaneous mutations.
- These assumptions imply a Poisson-like mean-variance relationship (Bretó and Ionides, 2011).
- Overdispersion (known as a relaxed clock) has been shown to improve the fit of phylogenetic models to observed genetic sequences in many cases (Drummond et al., 2006).
- In our approach, this corresponds to constructing each edge length of $\mathcal{P}(t)$ as a stochastic process on the corresponding edge of $\mathcal{T}(t)$.
- Various forms of such processes have been assumed in the literature, but not all are self-consistent under Markovian assumptions. The class of suitable random processes includes the class of nondecreasing Lévy processes, i.e., continuous-time processes with independent, stationary, non-negative increments.

Schematic diagram: SMC 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.

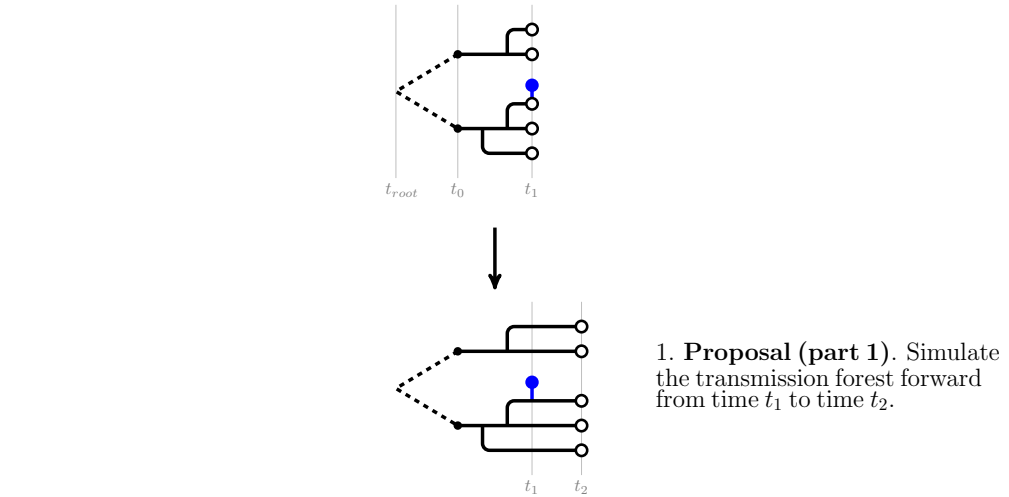
Just-in-time construction of state variables

- Although the model of the latent process includes the full phylogeny of the pathogen, $\mathcal{P}(t)$, for the purposes of computation we need only store $\widetilde{\mathcal{P}}(t)$ in memory.
- In our implementation, we add new edges to $\widetilde{\mathcal{P}}(t)$ at the time of measurement; it is not until a sequence is placed on a lineage of $\mathcal{T}(t)$ that we have enough information to update $\widetilde{\mathcal{P}}(t)$.
- We call this approach **just-in-time** construction of state variables because simulation of part of the state is postponed until the last moment.

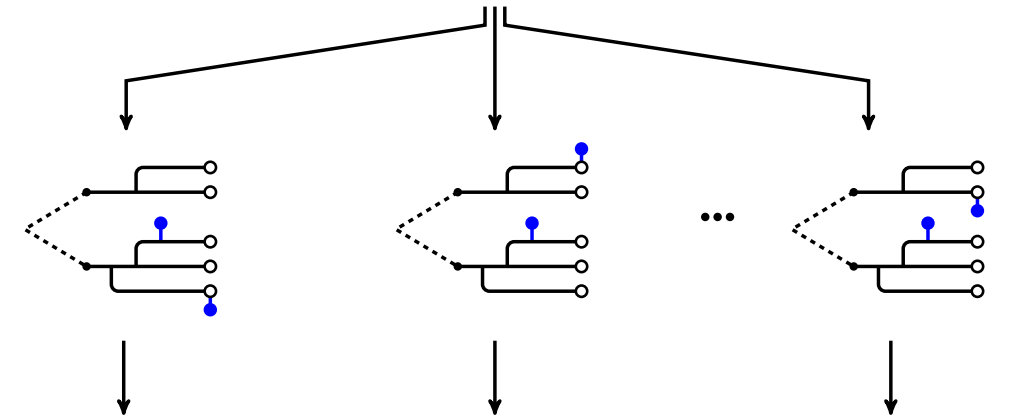
A hierarchical sampling scheme

- We developed a hierarchical sampling scheme to allow for scaling the effective number of particles while holding only a fraction of the effective number of particles in memory.
- For each of J particles in memory, we consider K nested proposals. In this hierarchical scheme, we split the proposal into two steps:
  - proposal of an updated transmission forest.
  - proposal of the location of the sampled sequence on the transmission forest.
- Each of J particles first proposes an updated transmission forest, then calculates the likelihood of the next observed sequence for K possible locations of the observed sequence.
- One of the K nested particles is kept, sampled with weight proportional to its conditional likelihood, and the remaining $K - 1$ particles are discarded.
- The weight of the surviving particle is the average of the conditional likelihoods of the K nested particles.

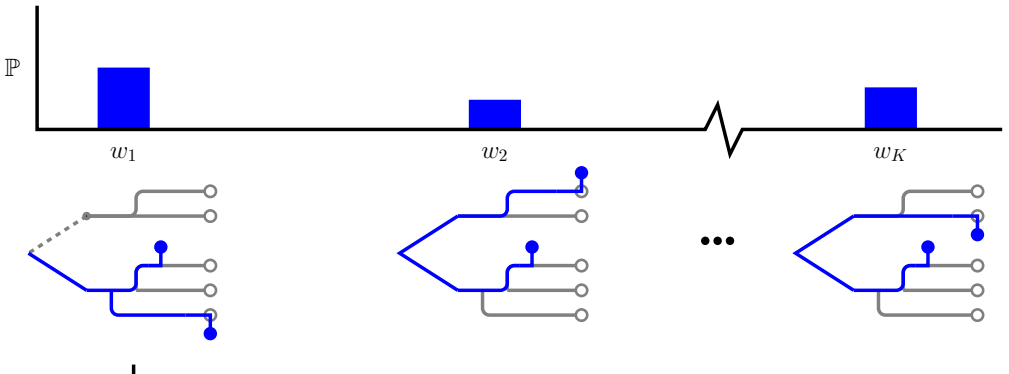
A schematic of our hierarchical sampling scheme.



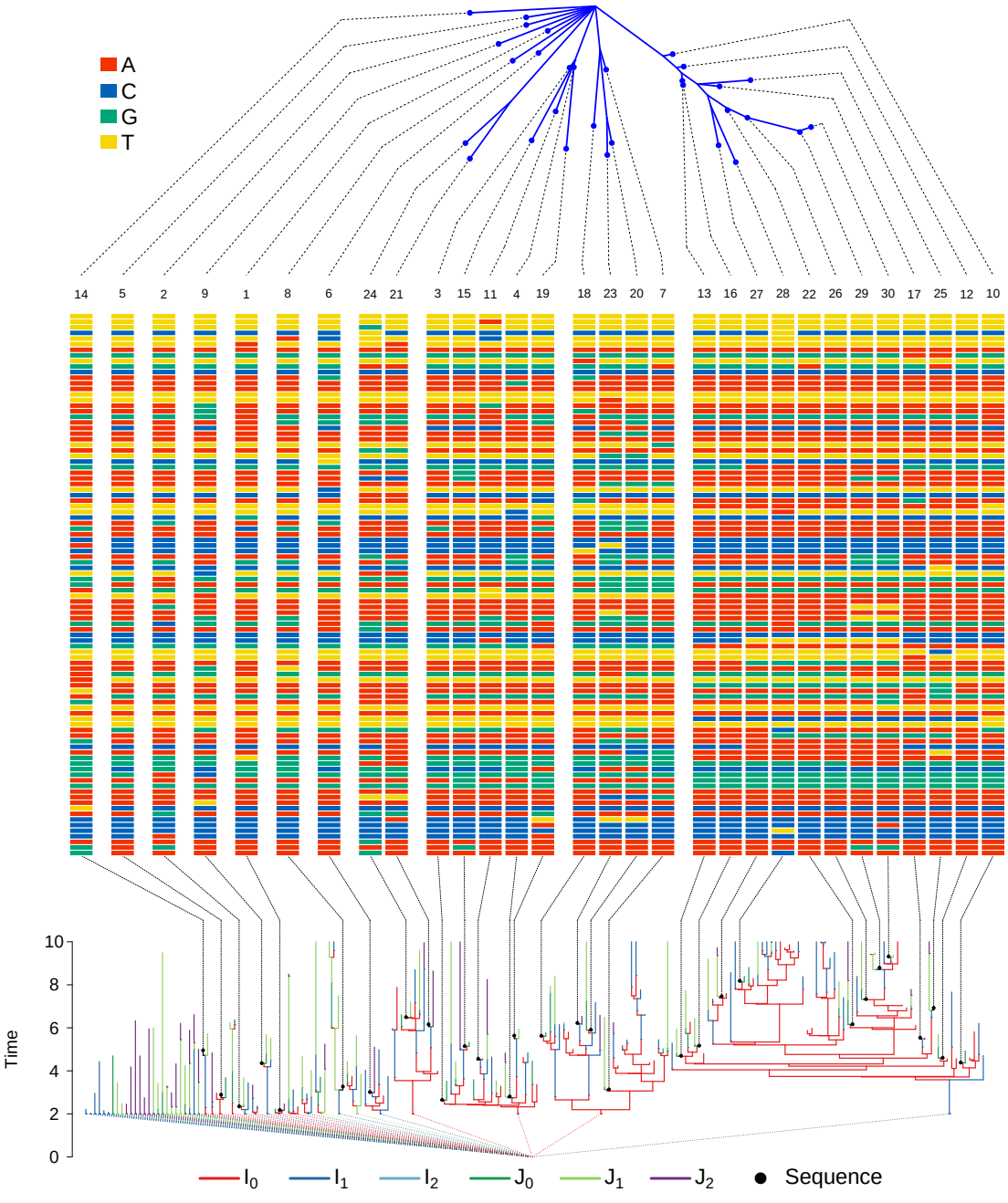
1. **Proposal (part 1)**. Simulate the transmission forest forward from time t_1 to time t_2.



2. **Proposal (part 2)**. Copy the particle K times. For each copy, propose an individual to sequence.



3. **Weighting**. For each of the K copies, 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.



A simulated HIV epidemic (Smith et al., 2017)

Top: phylogeny of observed sequences.

Middle: simulated sequence data. Actual data are confidential.

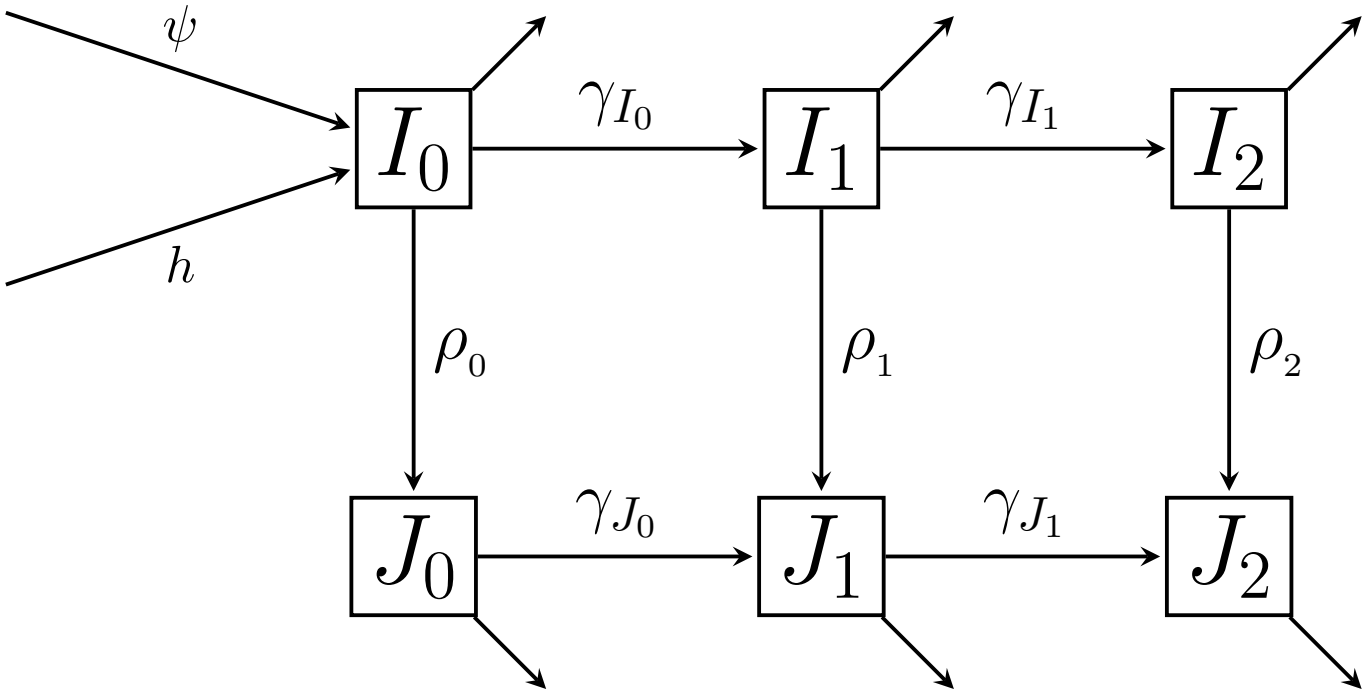
Bottom: Transmission forest for the full epidemic.

red: undiagnosed early infection

blue: undiagnosed chronic infection

green: diagnosed

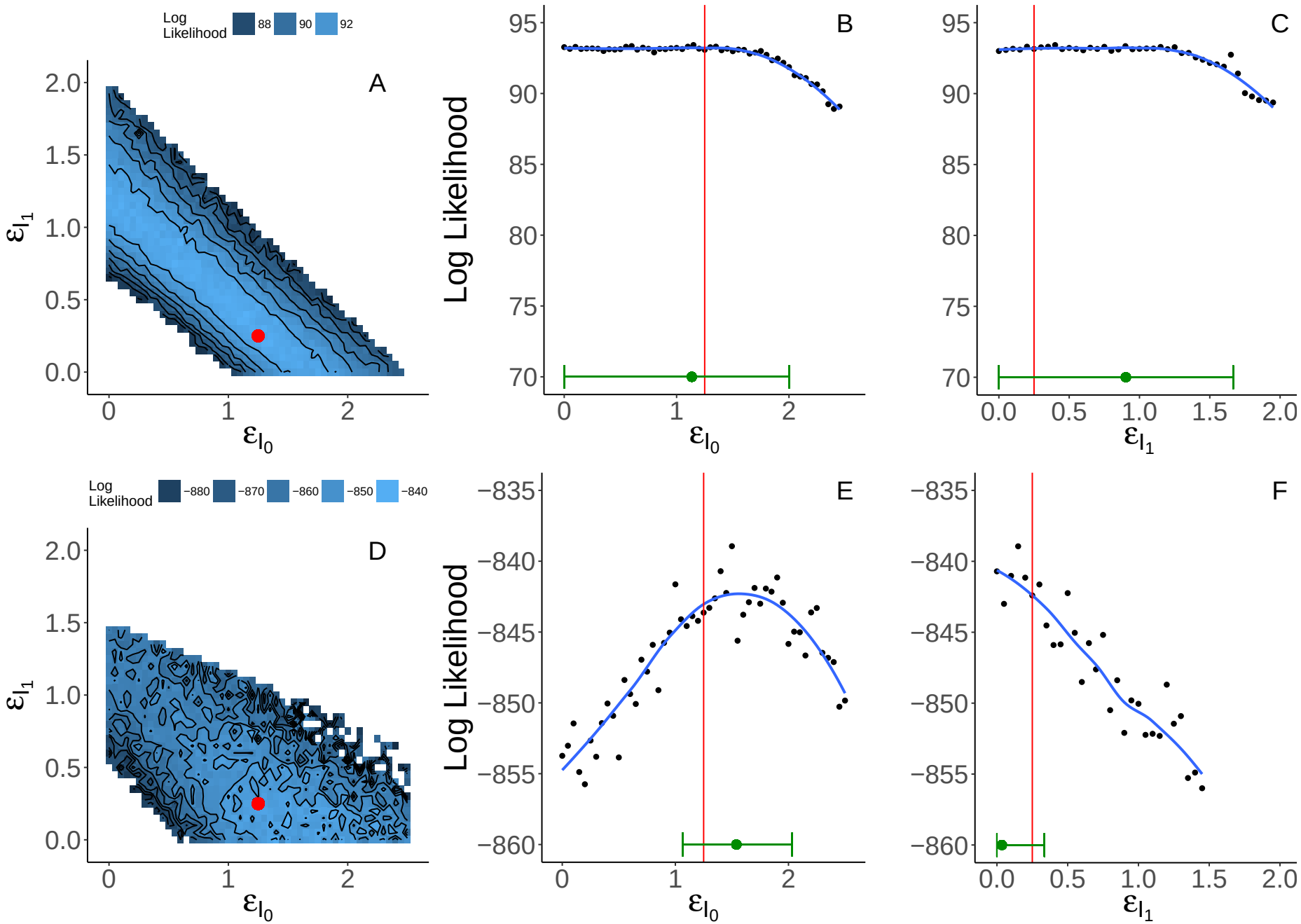
Model for infection and disease progression



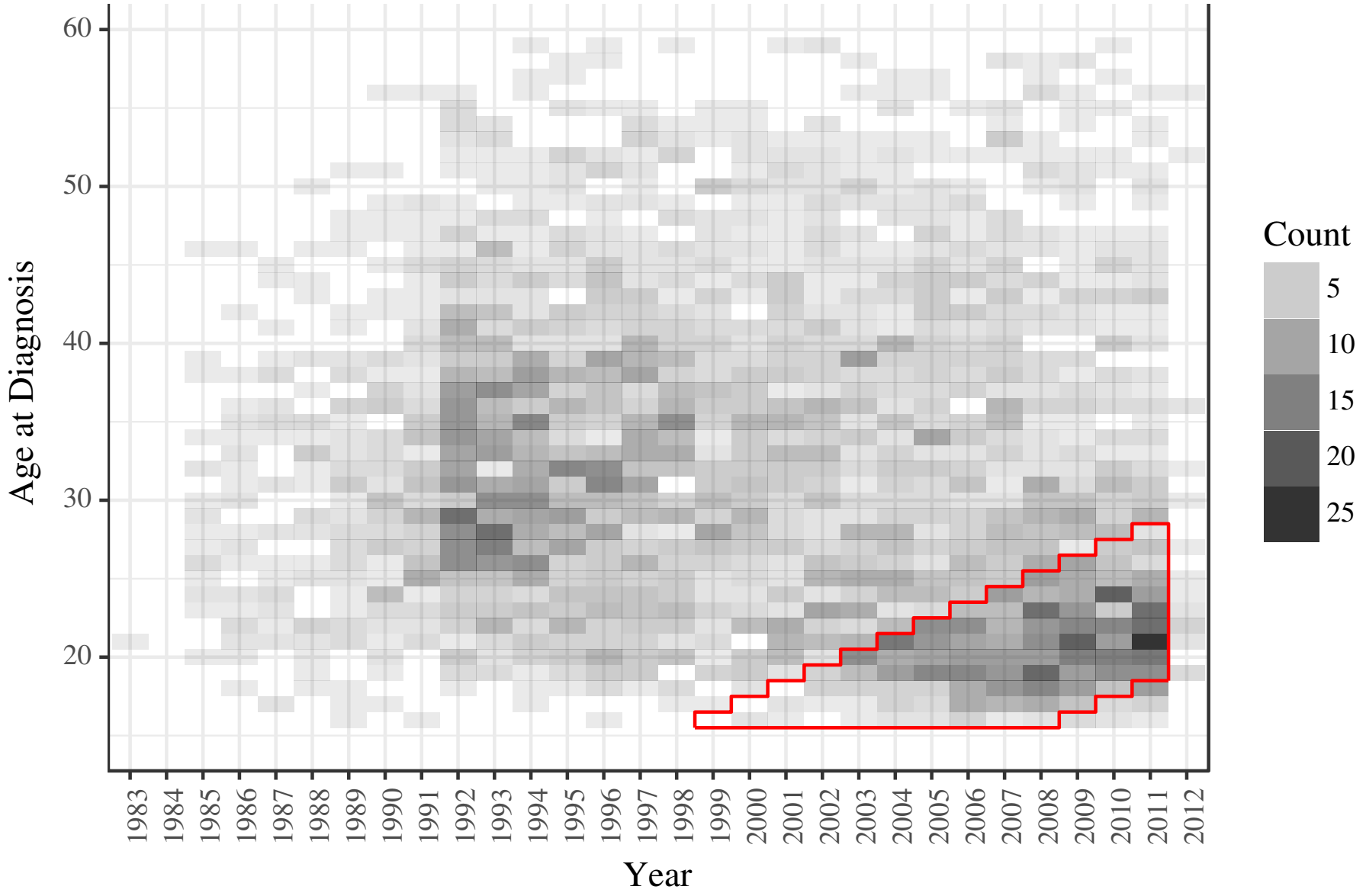
A flow diagram for HIV.

- I_k classes represent undiagnosed infections.
- J_k classes represent diagnosed infections.
- $k = 0, 1, 2$ denotes early, chronic and AIDS stages.
- Infection can come from within, or outside, the study population.

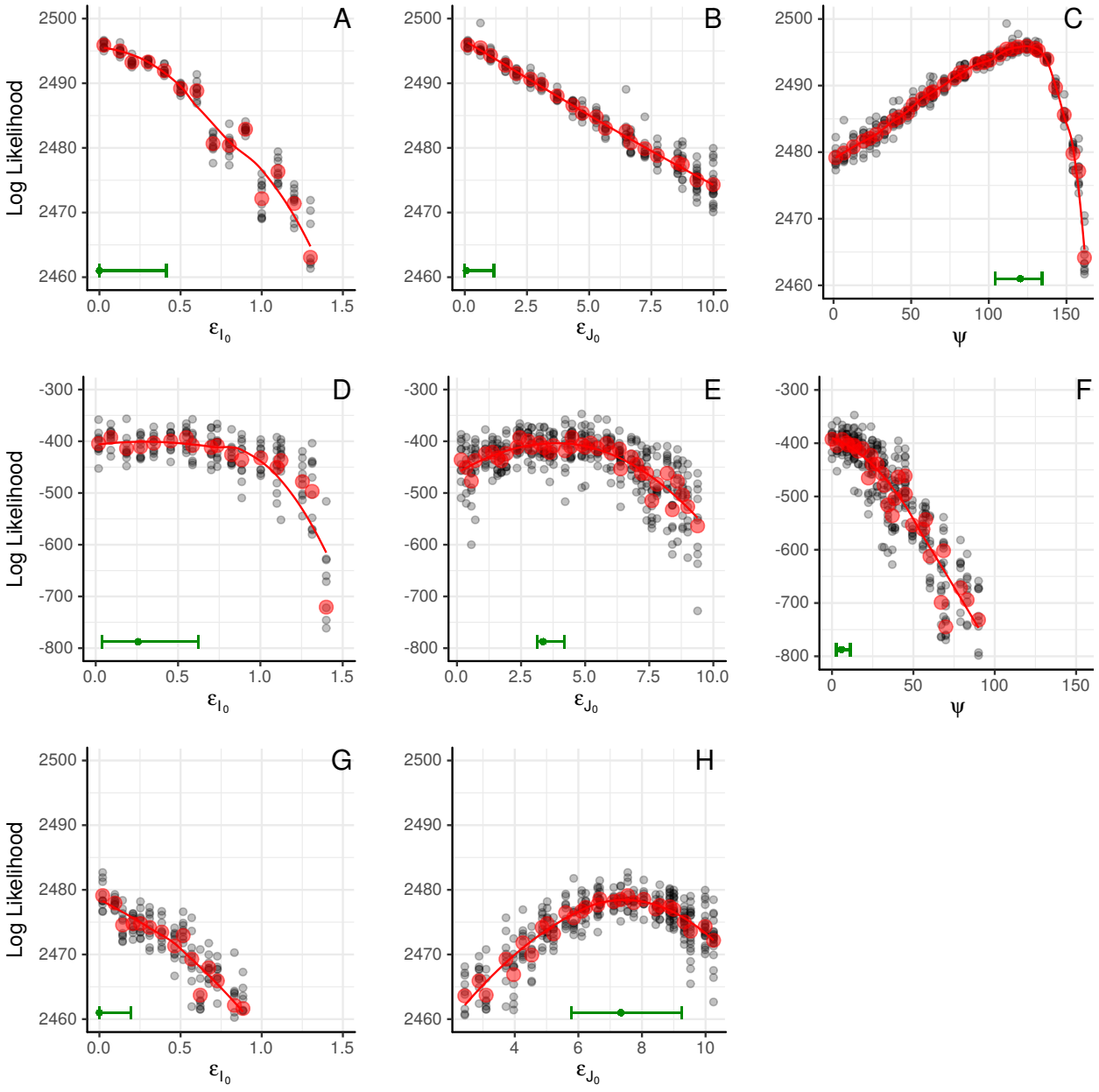
Top: diagnosis data only. Bottom: including sequence data



Detroit data: a young black MSM epidemic



Detroit data: a young black MSM epidemic



A-C. Diagnosis only.

D-F. Including sequence data.

G-H. Diagnosis only, fixing $\psi = 0$.

References I

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174 – 188.

Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Systematic Biology*, 61(4):579–593.

Bretó, C. and Ionides, E. L. (2011). Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Processes and their Applications*, 121:2571–2591.

Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005*, pages 64–69. IEEE.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., and Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92.

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.

References II

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Paige, B., Wood, F., Doucet, A., and Teh, Y. W. (2014). Asynchronous anytime sequential Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 3410–3418.

Rasmussen, D. A., Ratmann, O., and Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, 7(8):e1002136.

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*, pre-published online, doi:10.1093/molbev/msx124.

Yu, B. (2014). IMS presidential address: Let us own data science. *IMS Bulletin*, 43.