Demographic inference under the coalescent in a spatial continuum

Stéphane Guindon[1,2], *Hongbin Guo*[2] *& David Welch*[2]

[1] *LIRMM - CNRS UMR 5506 - Montpellier*
[2] *The University of Auckland - New Zealand*

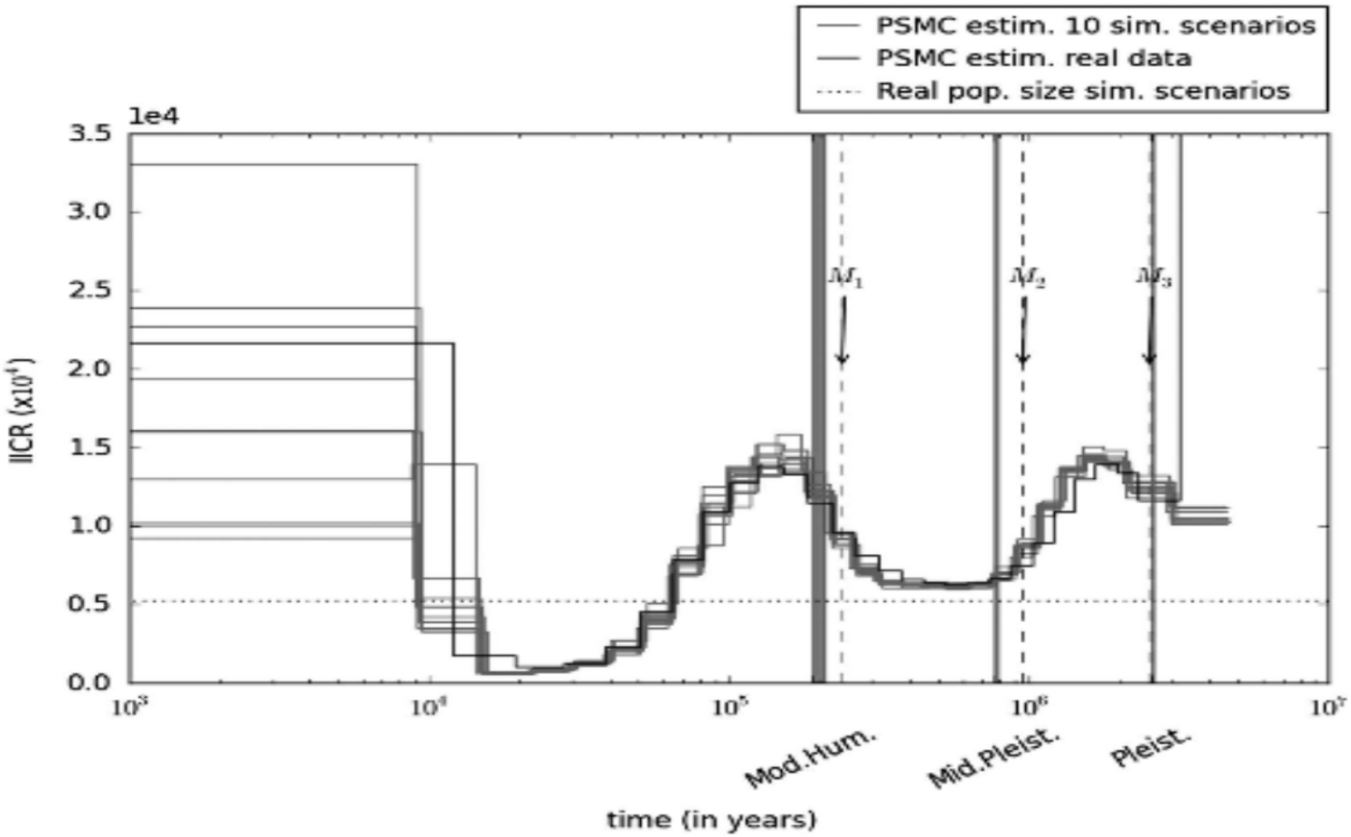
Heredity (2016) 116, 362–371

www.nature.com/hdy

ORIGINAL ARTICLE

On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference?

O Mazet[1,7], W Rodríguez[1,7], S Grusea[1], S Boitard[2,3] and L Chikhi[4,5,6]



Previous work: isolation-by-distance

- Wright, 1943 and Malécot, 1948.
- Each parent independently produces a Poisson number of offspring.
- Each descendant has its spatial location determined by a normal distribution *centered on the parental location* with variance σ^2.
- Problem: this model predicts the formation of clumps of individuals of increasing densities (Felsenstein, 1975).
- Tweaking the model to introduce regulation of population density in the continuous setting is not straightforward.

Previous work: stepping stone model (and dual)

- IBD model with regulation of population density.
- Habitat is approximated using a grid.
- Decreasing the mesh size (i.e., increasing the number of sub-populations) decreases the size of sub-populations $\mapsto$ coalescent breaks down?
- Dual process (backward in time): *structured coalescent*.
- Easy to calculate the joint probability density of a genealogy and the full migration "map" (i.e., time, position in the tree and type of migrations events considered as known).
- "Ghost demes" $\mapsto$ biased demographic estimates.

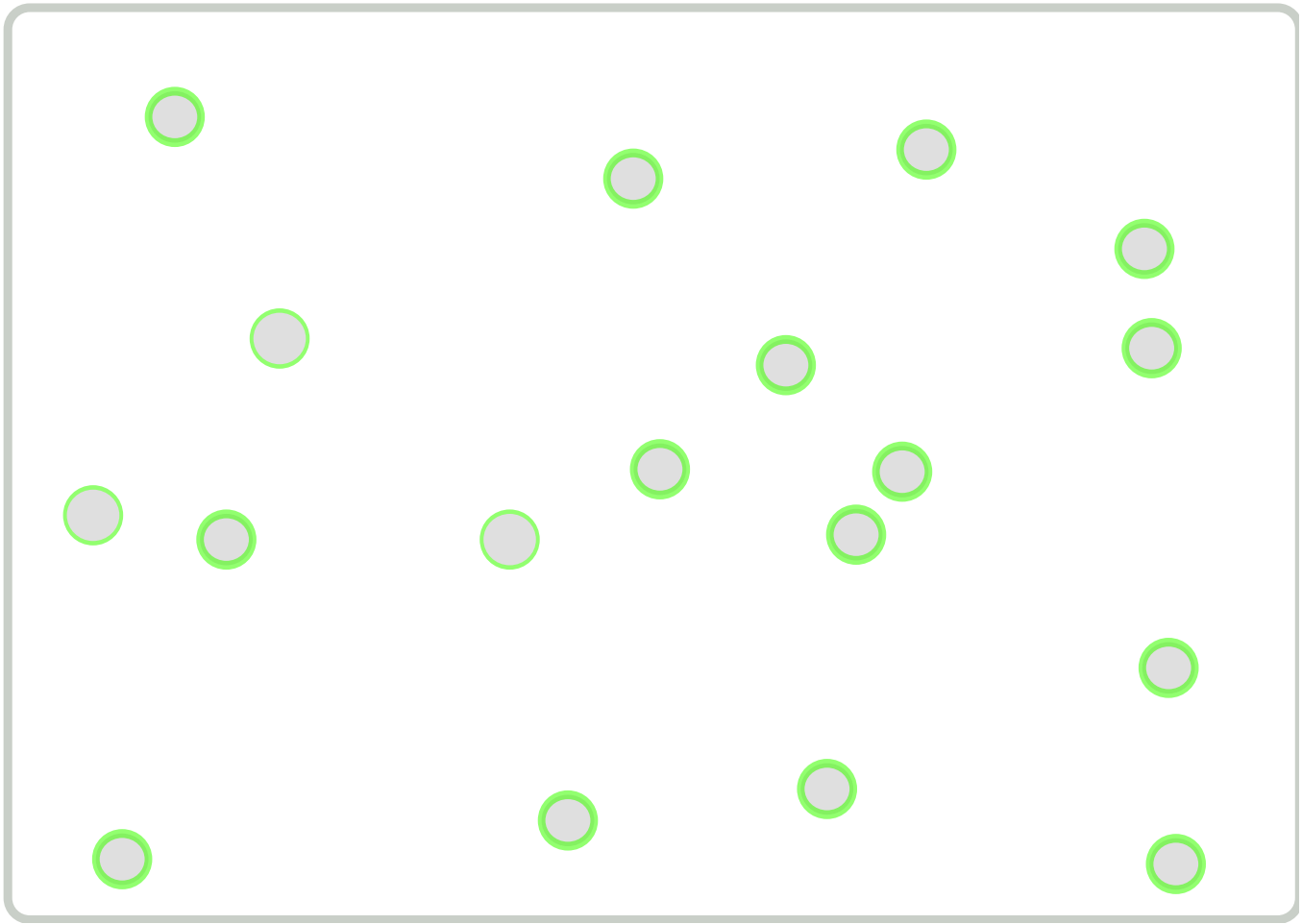
"Mugration" model

- Habitat still approximated as a grid.
- Continuous-time Markov chain running forward in time along the genealogy/phylogeny.
- Computationally efficient but statistically deficient.
- Migration rates reflect sampling intensity rather than actual biological process (De Maio et al., 2015), unequal deme sizes is not reflected in the calculation of the probability density of the genealogy.
- Sensitivity to mesh size?

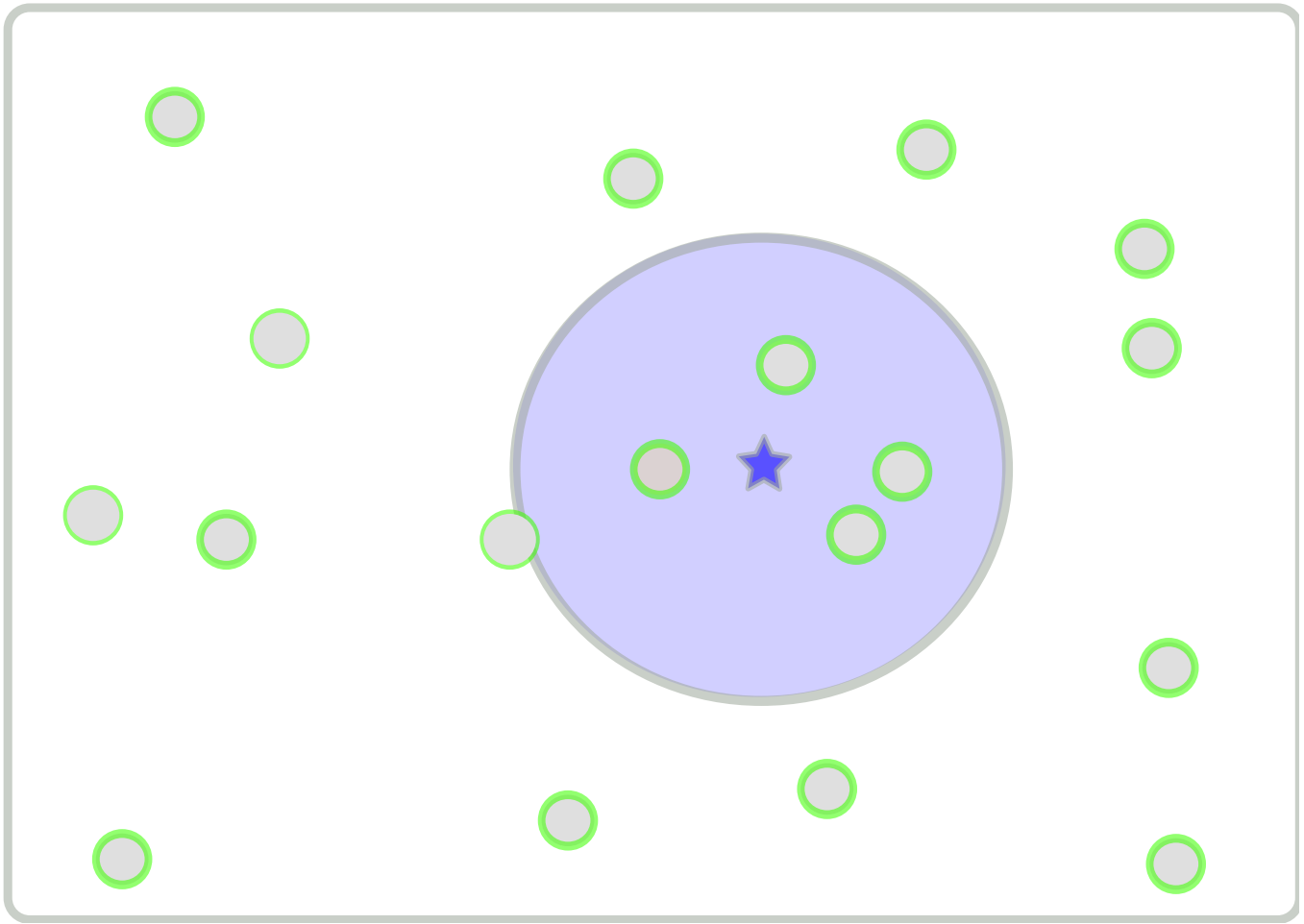
The spatial Λ-Fleming-Viot (ΛV) process

- Alison Etheridge, Nick Barton, Amandine Véber and colleagues.
- Related to Λ-coalescent, i.e., coalescent with multiple mergers.
- Space is continuous.
- Spatial distribution of individuals in population determined by a *stationary* Poisson point process.
- Coalescent $\mapsto$ sampling of individuals is dealt with.

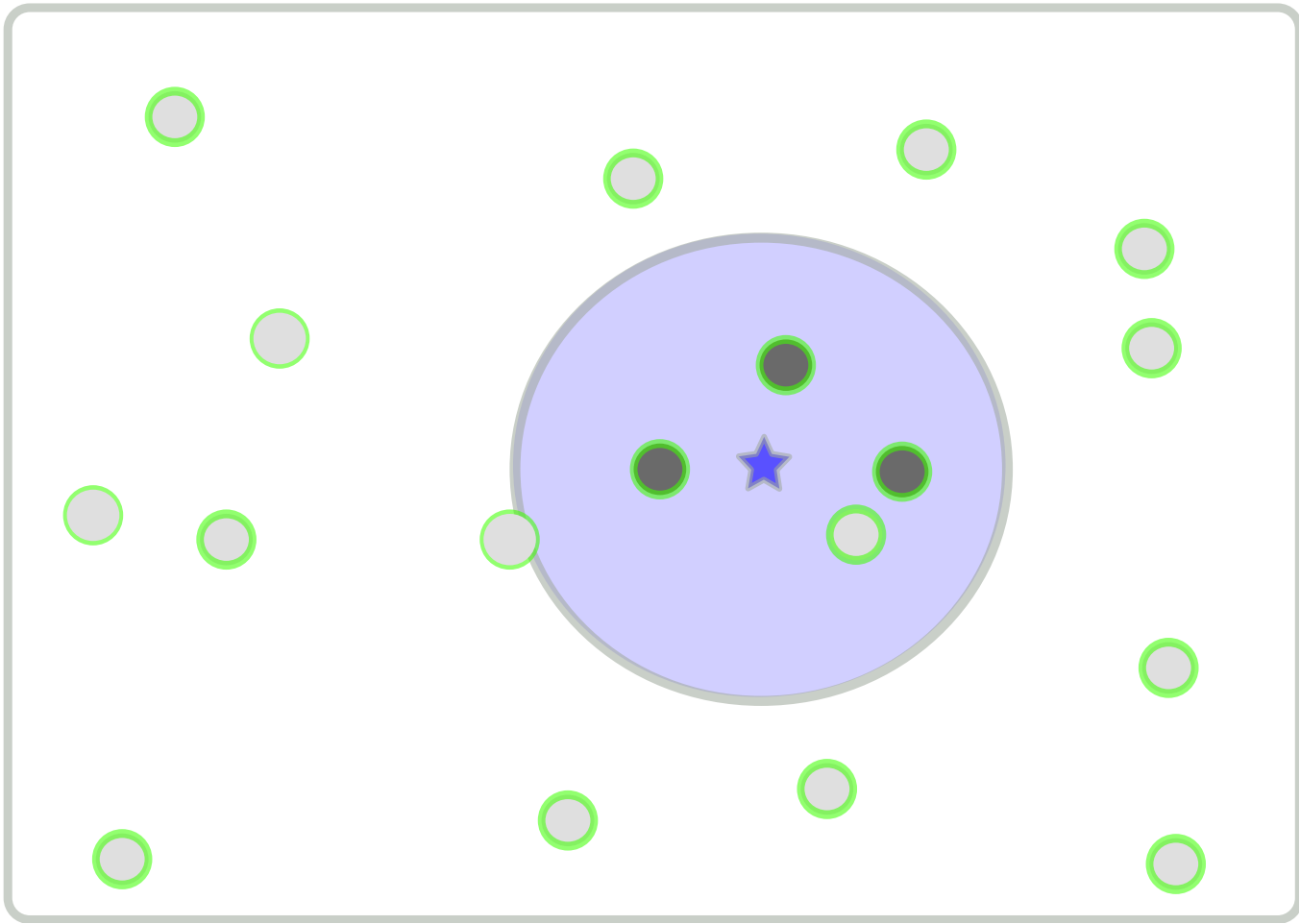
ΛV *forward-in-time dynamics (population)*



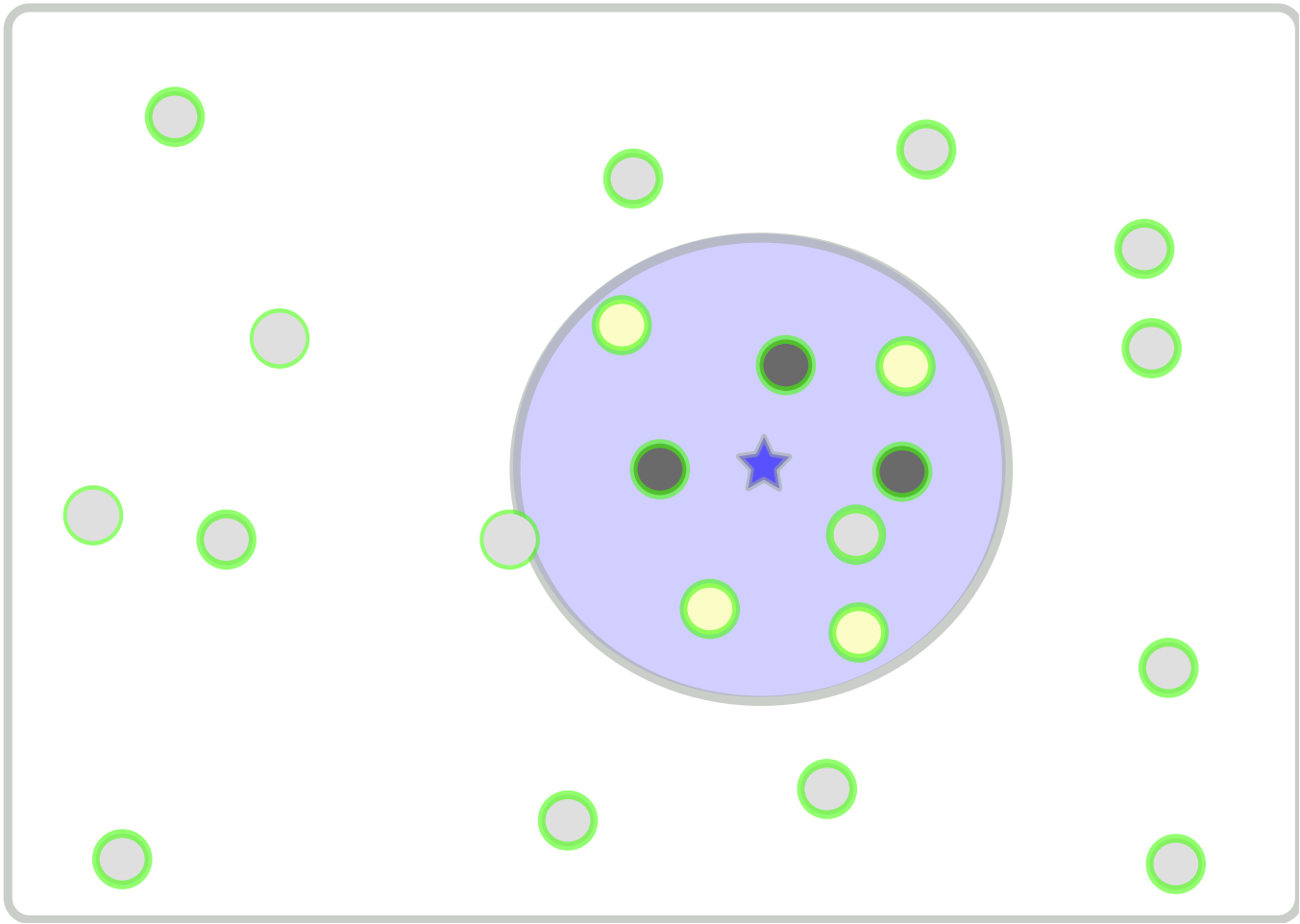
ΛV *forward-in-time dynamics (population)*



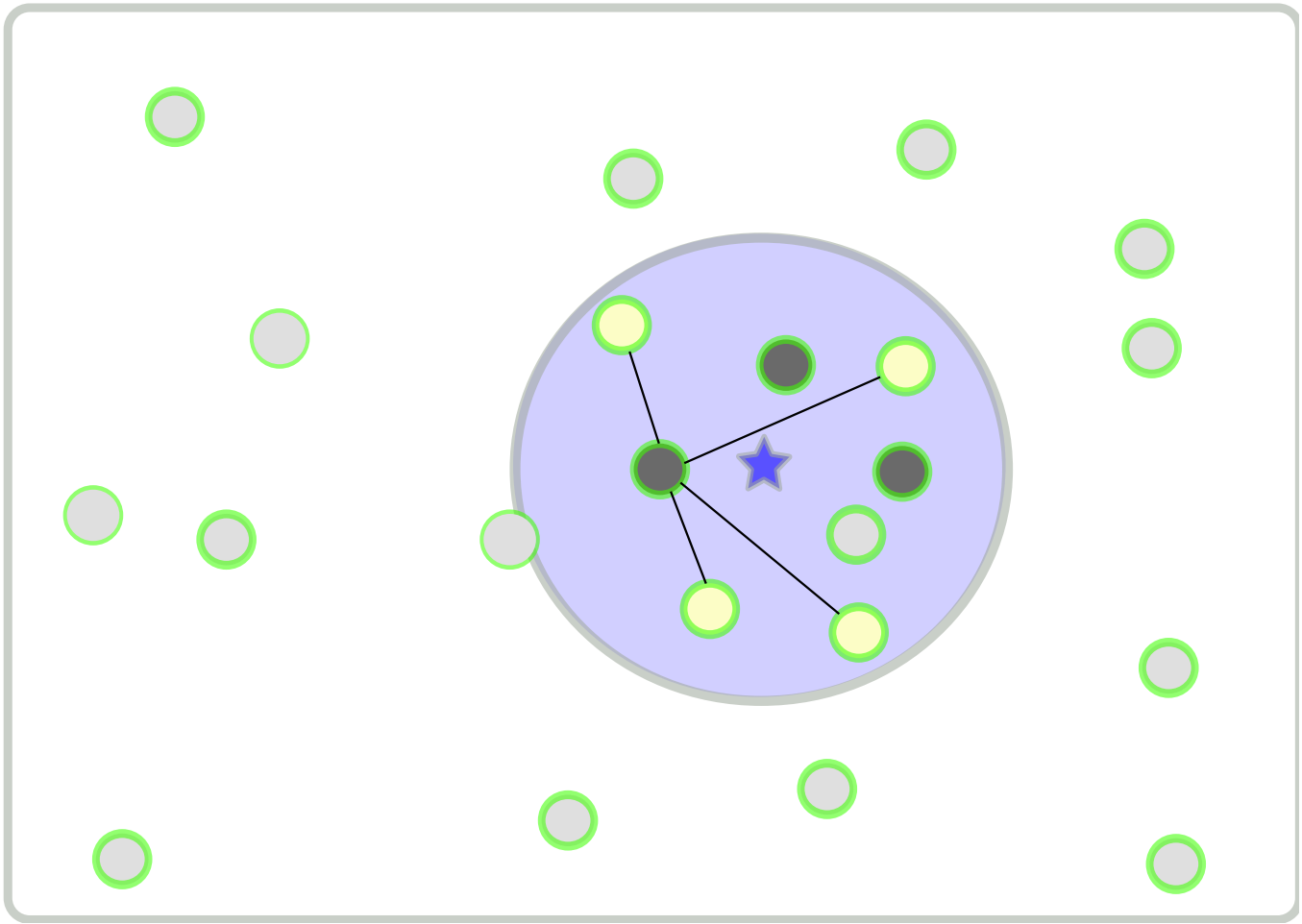
ΛV *forward-in-time dynamics (population)*



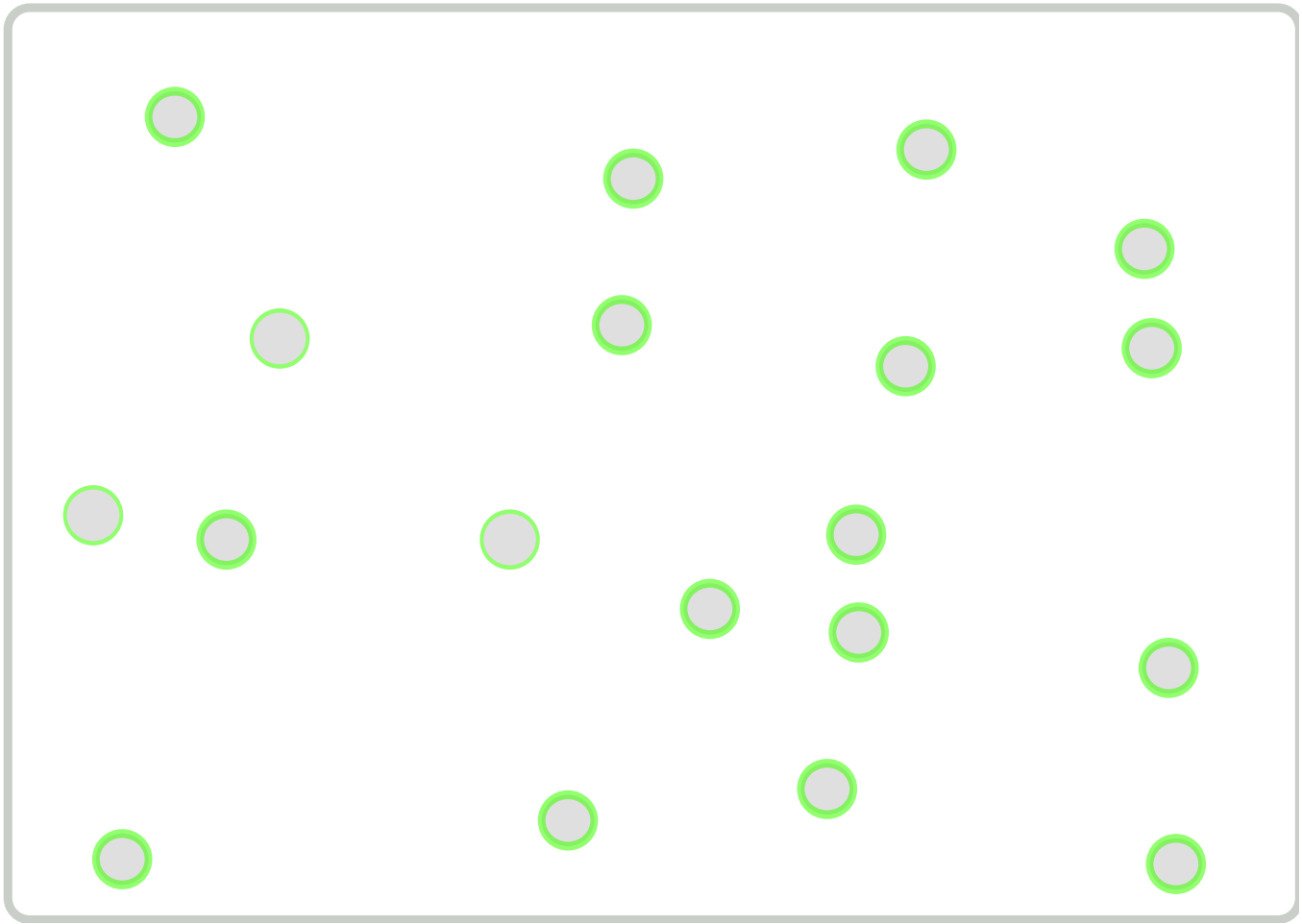
ΛV *forward-in-time dynamics (population)*



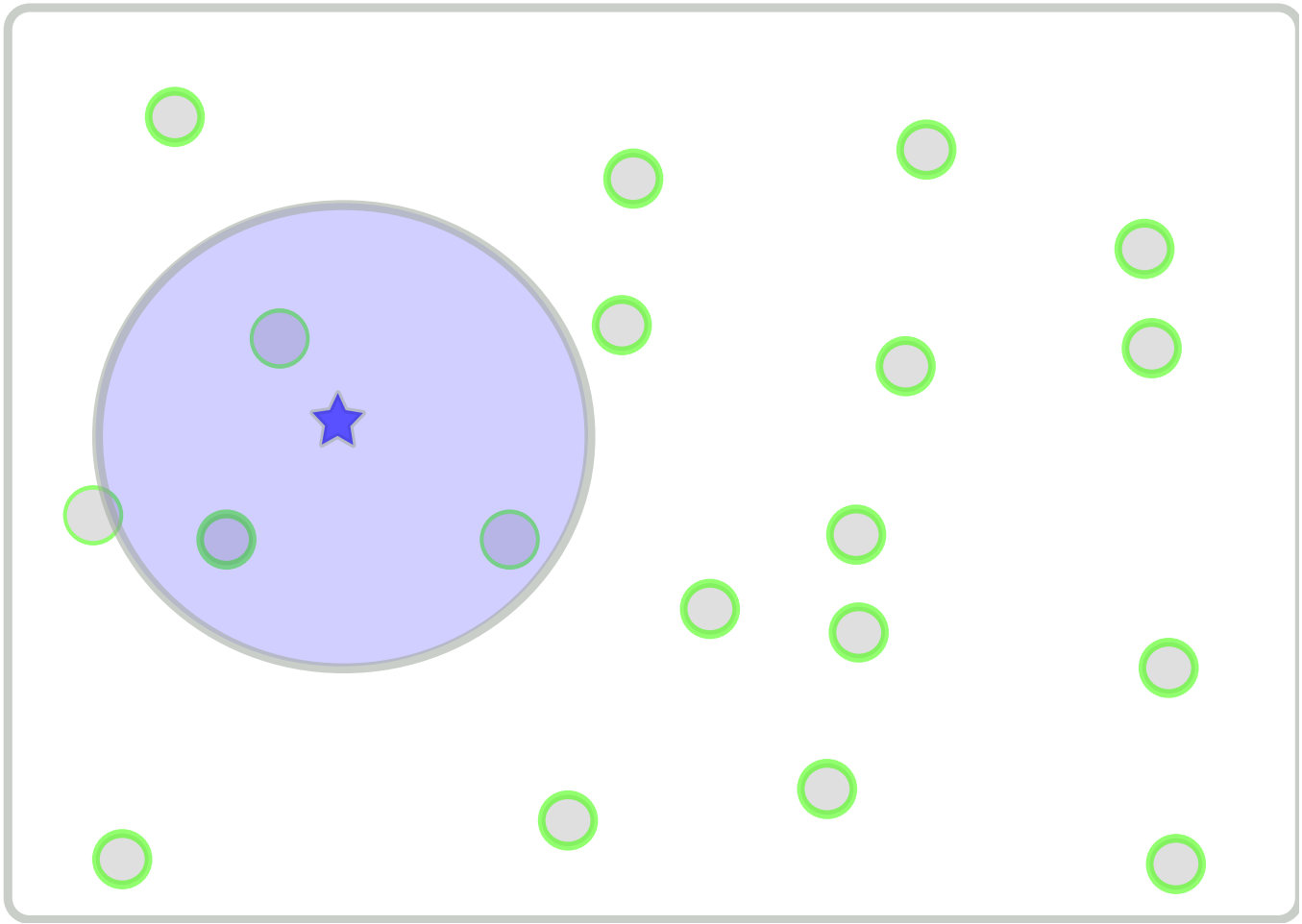
ΛV *forward-in-time dynamics (population)*



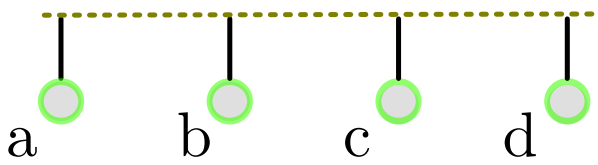
ΛV *forward-in-time dynamics (population)*

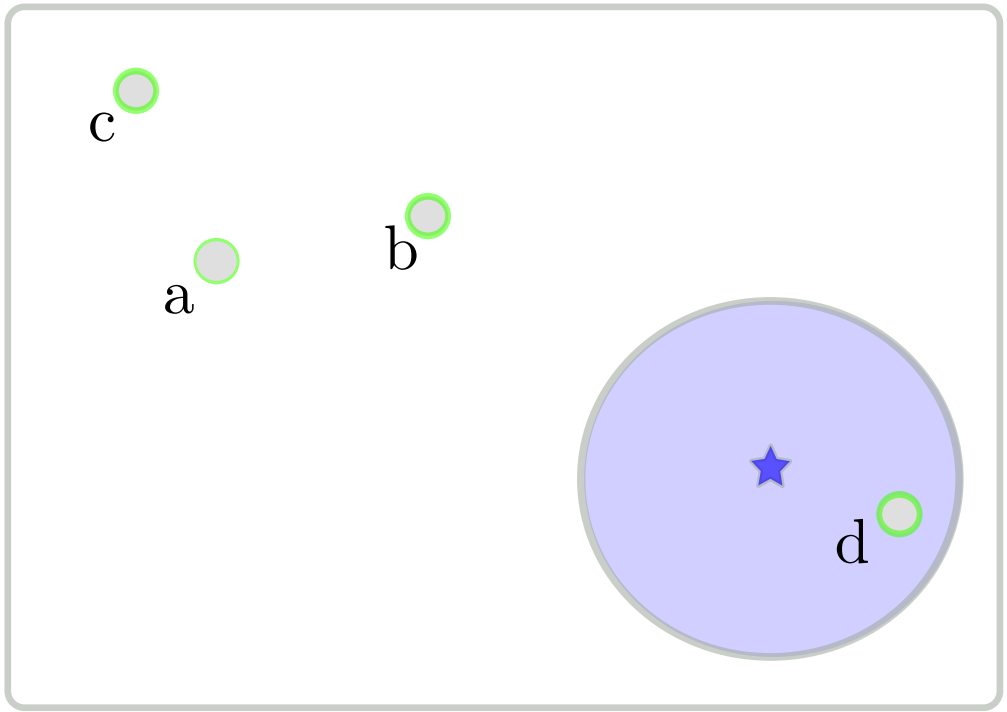


ΛV *forward-in-time dynamics (population)*

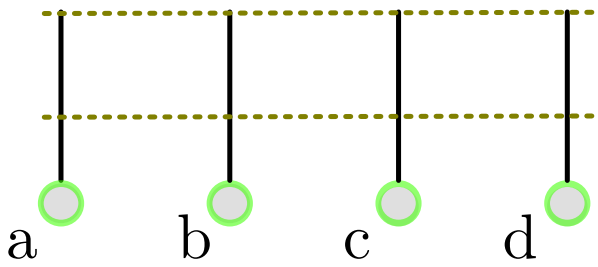


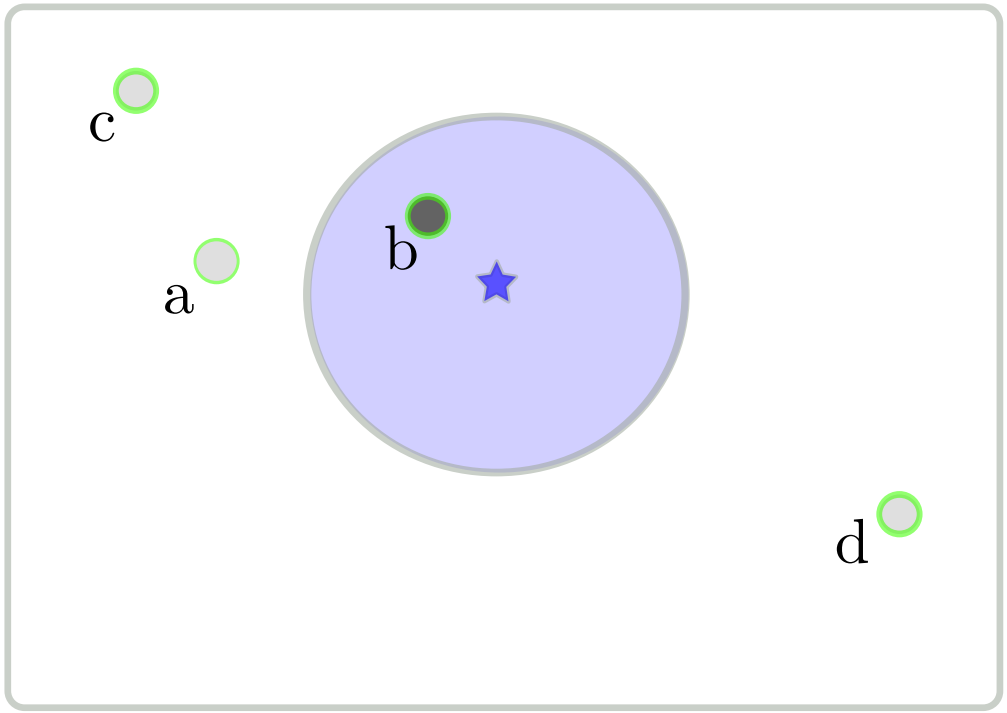
ΛV *backward-in-time dynamics (sample)*



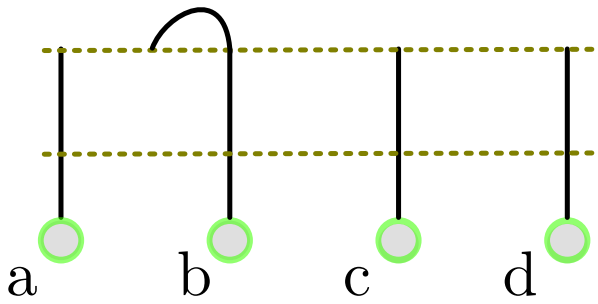


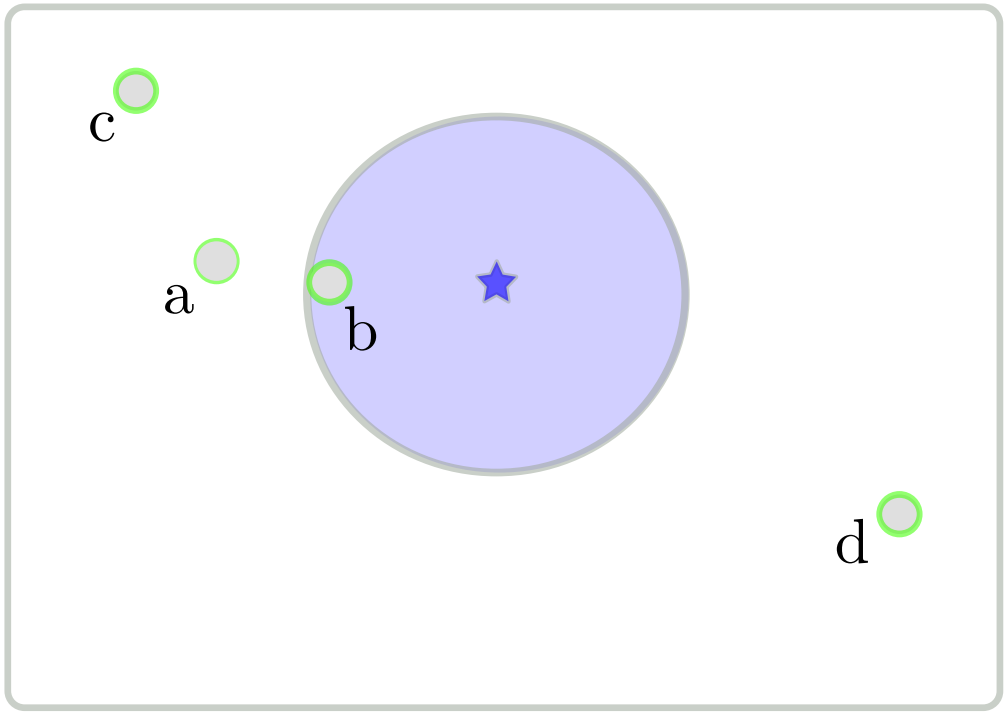
ΛV *backward-in-time dynamics (sample)*



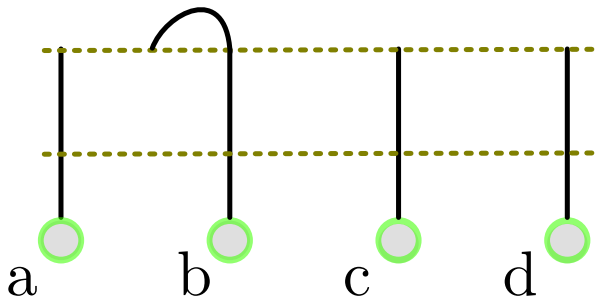


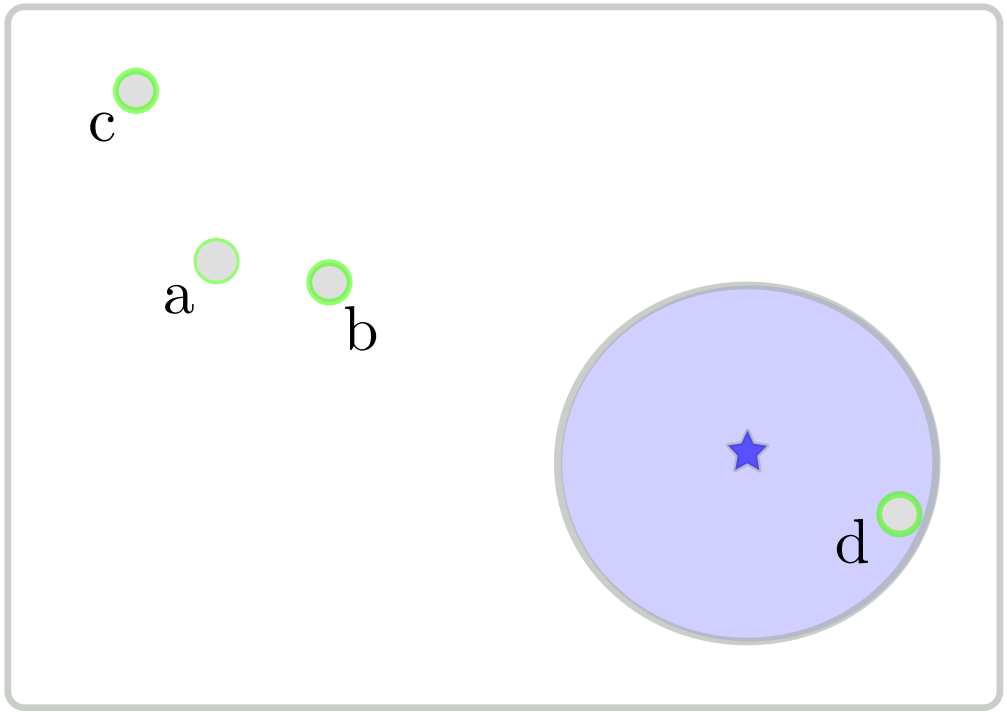
ΛV *backward-in-time dynamics (sample)*



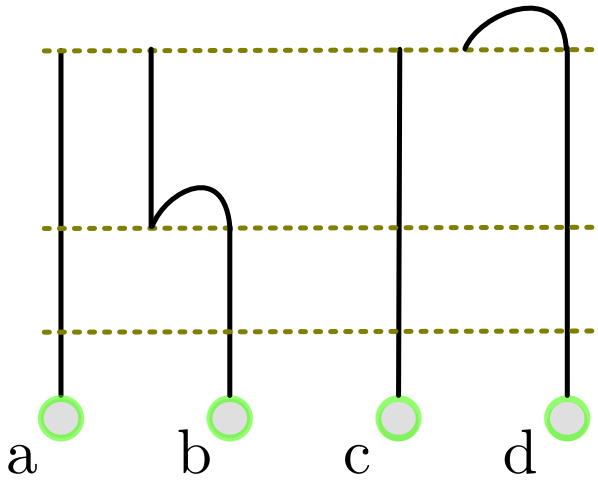


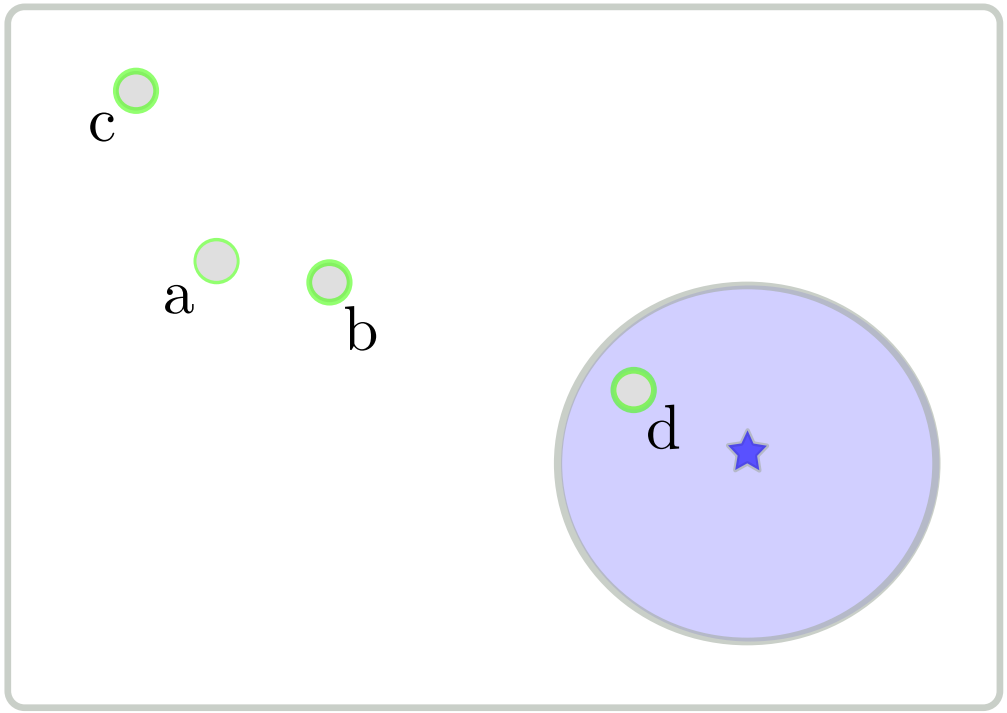
ΛV *backward-in-time dynamics (sample)*



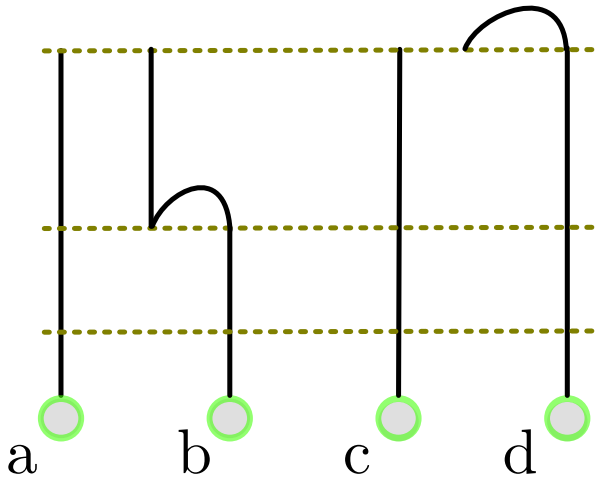


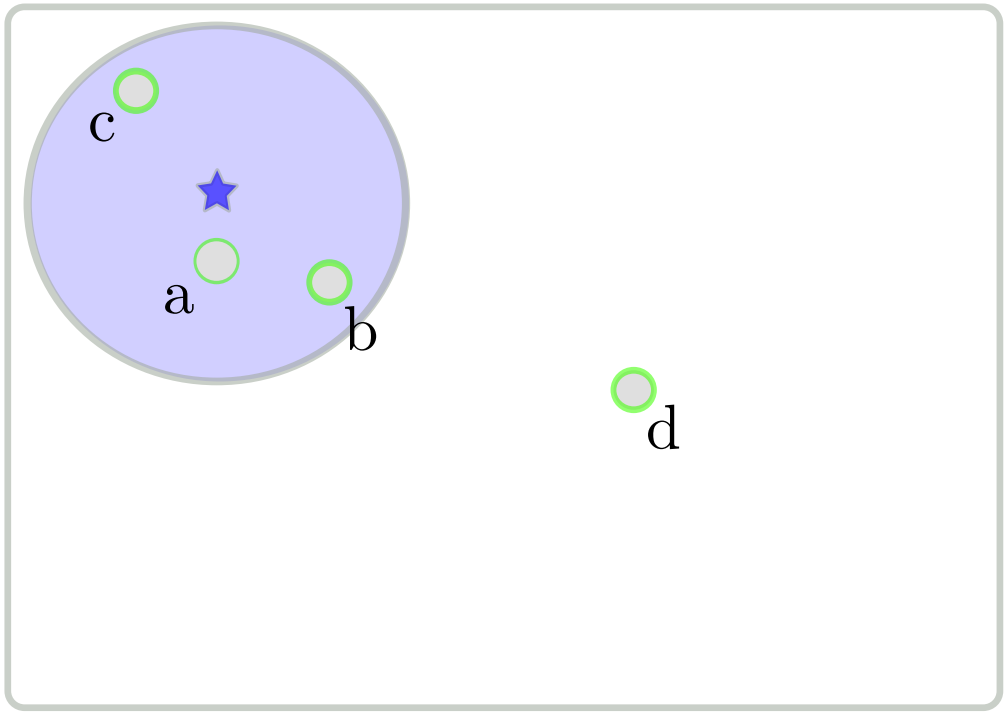
ΛV *backward-in-time dynamics (sample)*



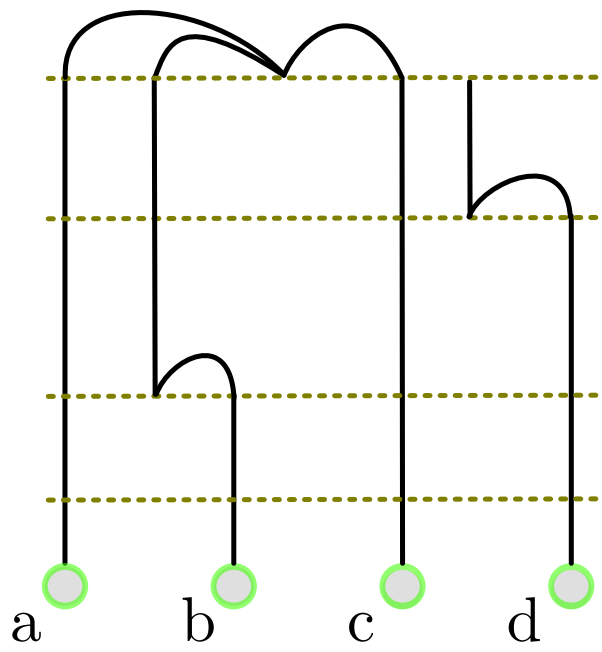


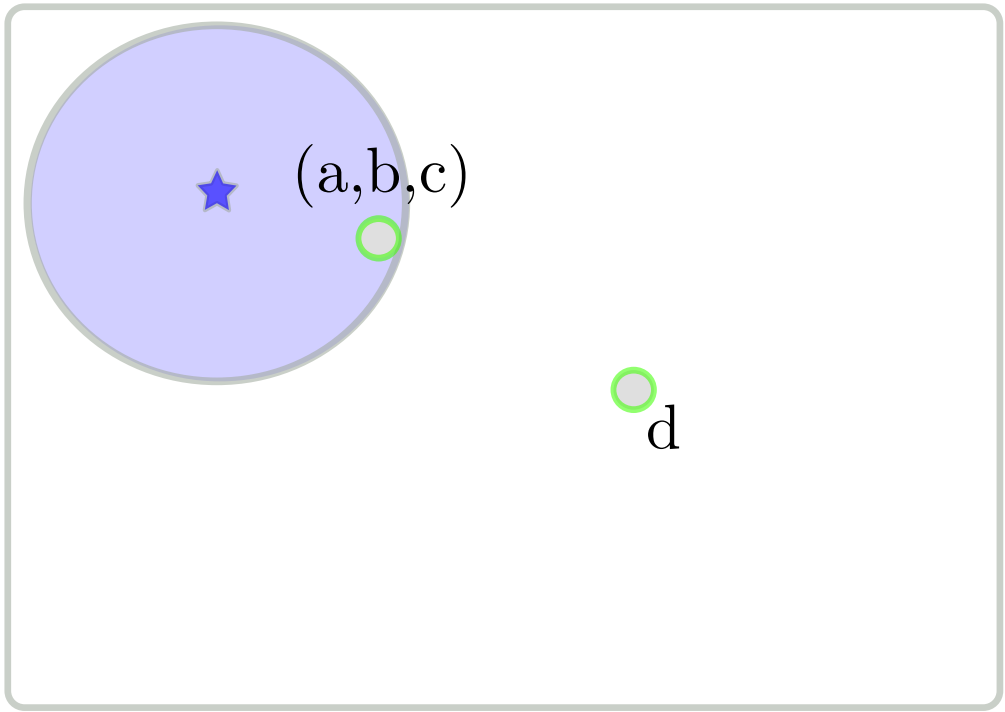
ΛV *backward-in-time dynamics (sample)*



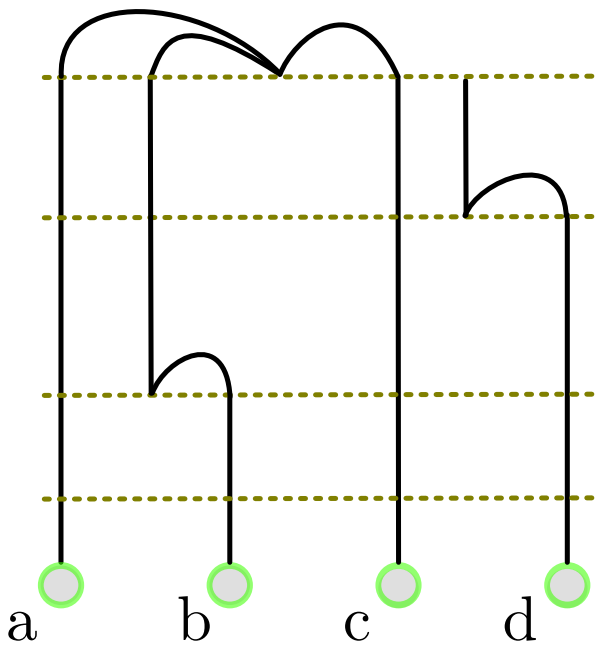


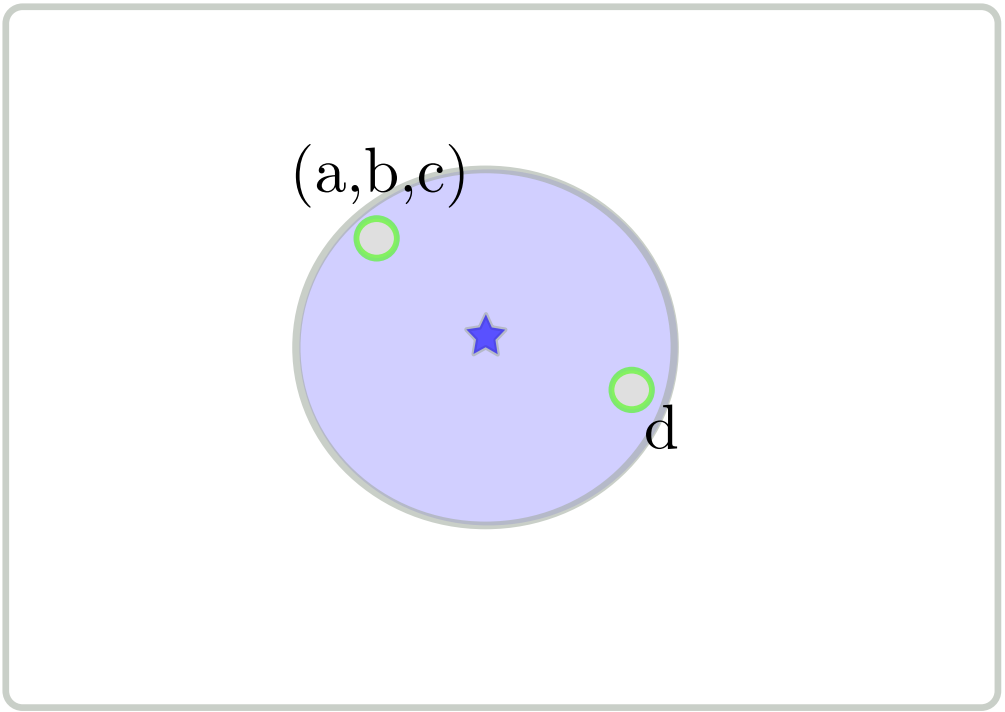
ΛV *backward-in-time dynamics (sample)*



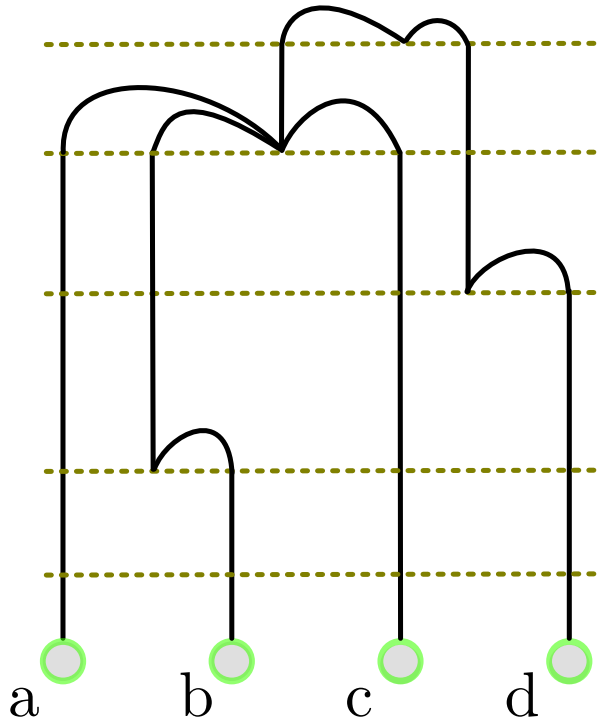


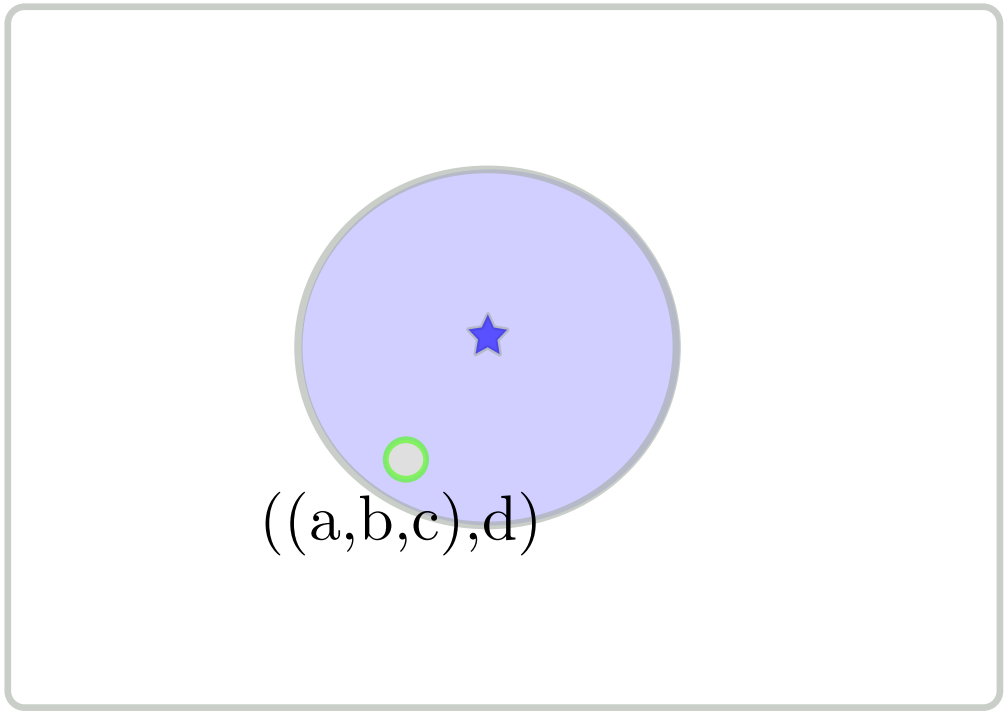
ΛV *backward-in-time dynamics (sample)*





ΛV *backward-in-time dynamics (sample)*





Backward-in-time process

- REX events occur at rate λ.
- c_i chosen uniformly at random.
- Lineage k, with location $l^+_{i,k}$ at time t^+_i, is "hit" by the event with probability:

$$u(l^+_{i,k}, c_i) = \mu \exp\left(-\frac{\| l^+_{i,k} - c_i \|^2}{2\theta^2}\right)$$

- It jumps to (parental) location $l^-_{i,k}$ with probability density:

$$v(l^-_{i,k}, c_i) \propto \exp\left(-\frac{\| l^-_{i,k} - c_i \|^2}{2\theta^2}\right)$$

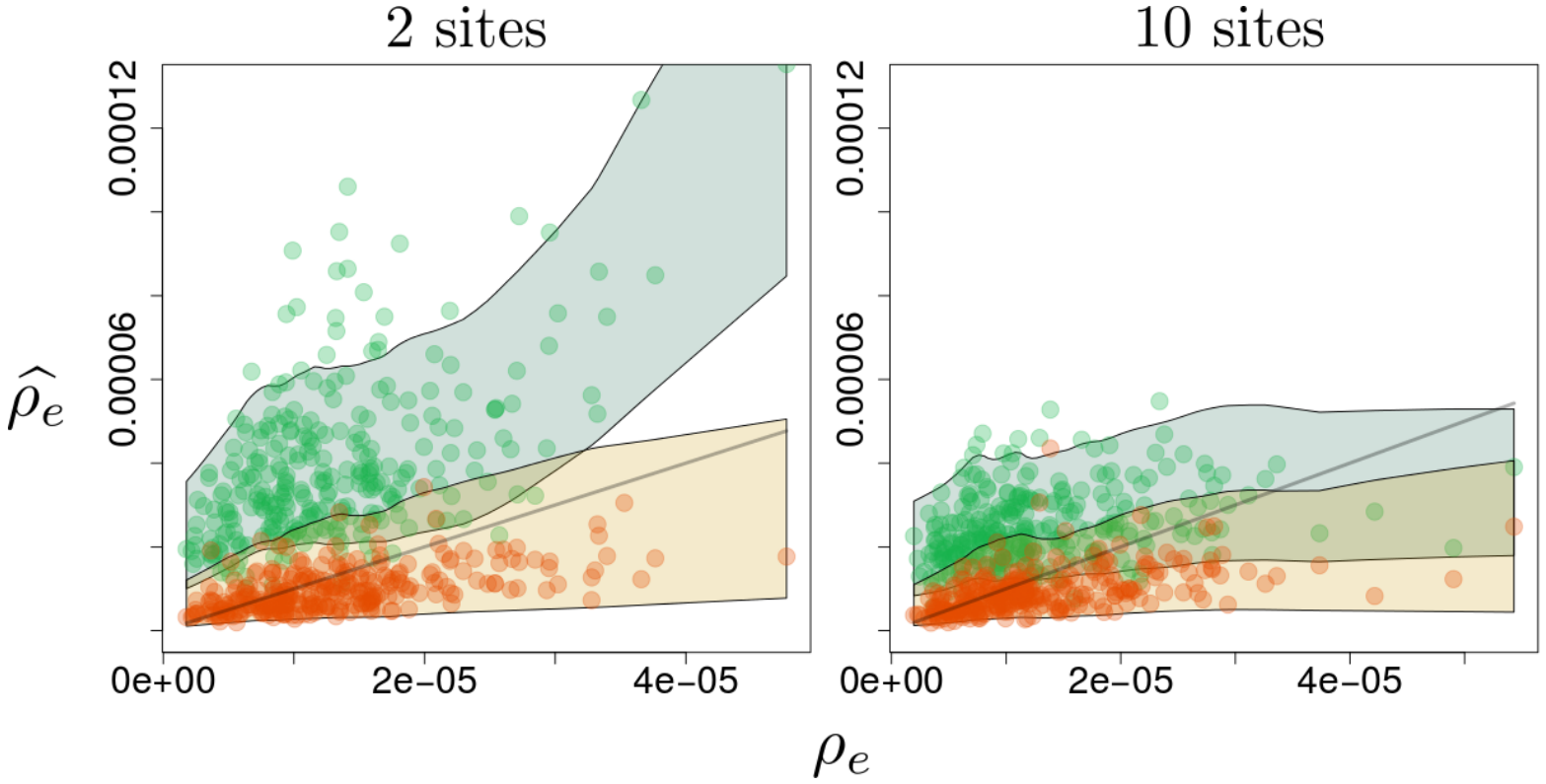
Parameter inference

- Unable to evaluate the likelihood of the ΛV model given the available data (i.e., geo-referenced genetic sequences)
- Work out the likelihood of *augmented* data: geo-referenced genetic sequences + number and position of all REX events + coordinates of ancestral lineages.
- Use MCMC to sample from the joint posterior of the ΛV model parameters and nuisance parameters arising as a consequence of data augmentation (i.e., treat ancestral "data" as parameters).

Simulations

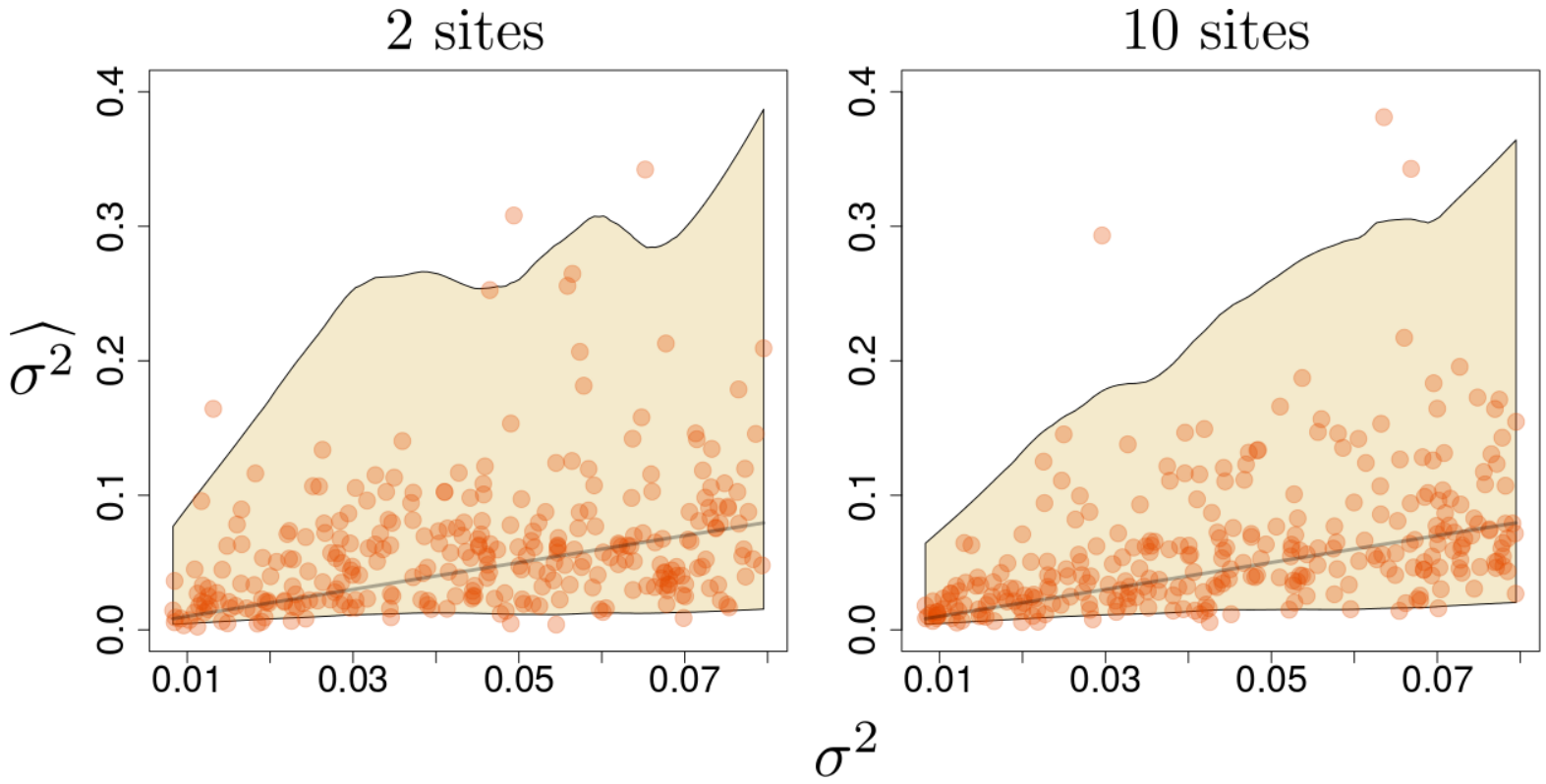
- Habitat: 10×10 square.
- Effective sample size, N_e, uniform in $[100, 5000]$.
- $\mathcal{N}$ uniform in $[N_e \times 10^{-3}, N_e \times 10^{-2}]$.
- θ uniform in $[1.5, 4]$.
- Simulate a population forward in time.
- Throw two/ten triangles uniformly at random on the square (covering $\sim$4% to $\sim$10% of the total area).
- Trace the genealogy of the sample backward in time.
- Simulate sequences along the tree.

Population density



Green: structured coalescent – Yellow: ΛV

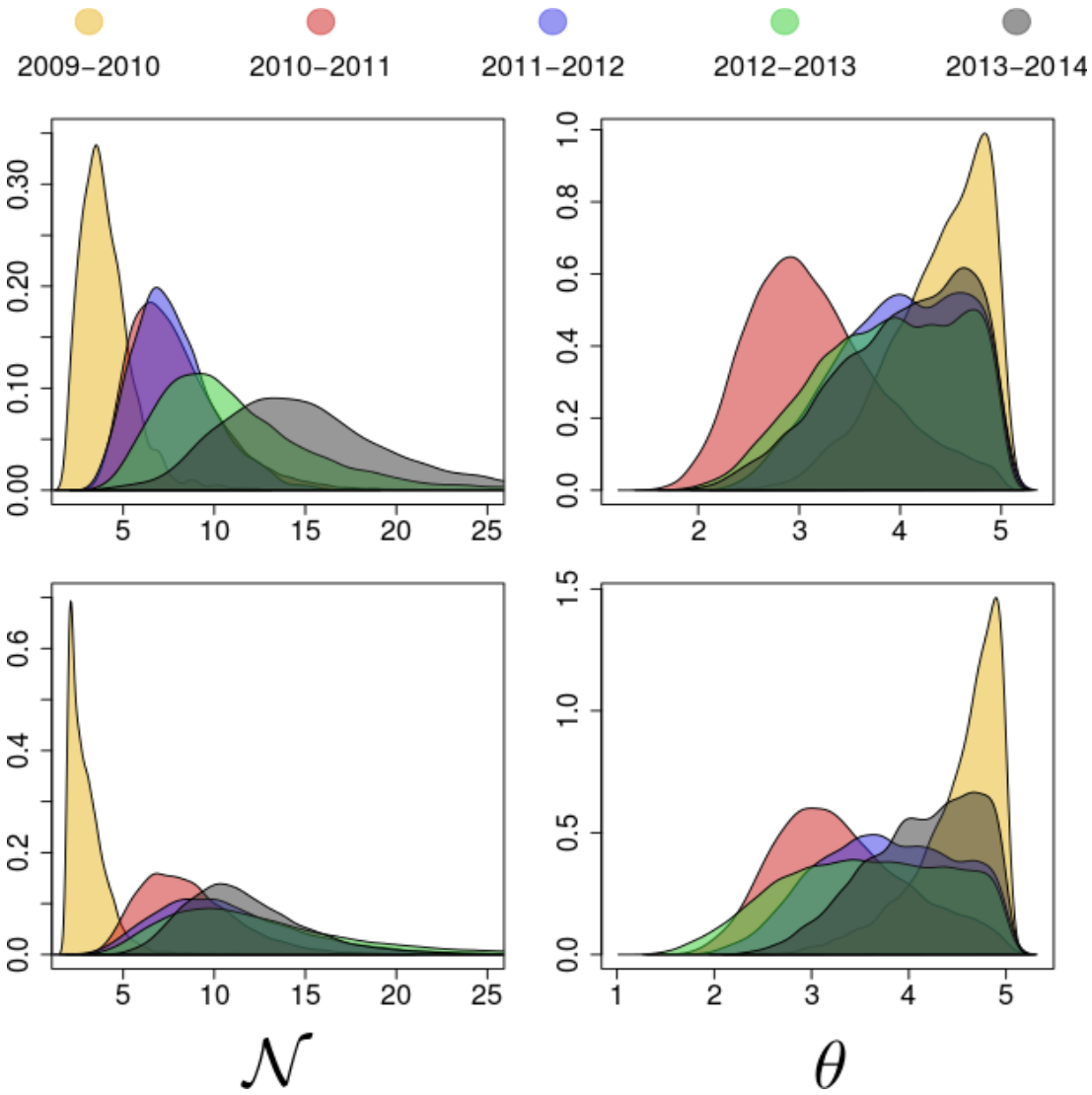
Dispersal intensity



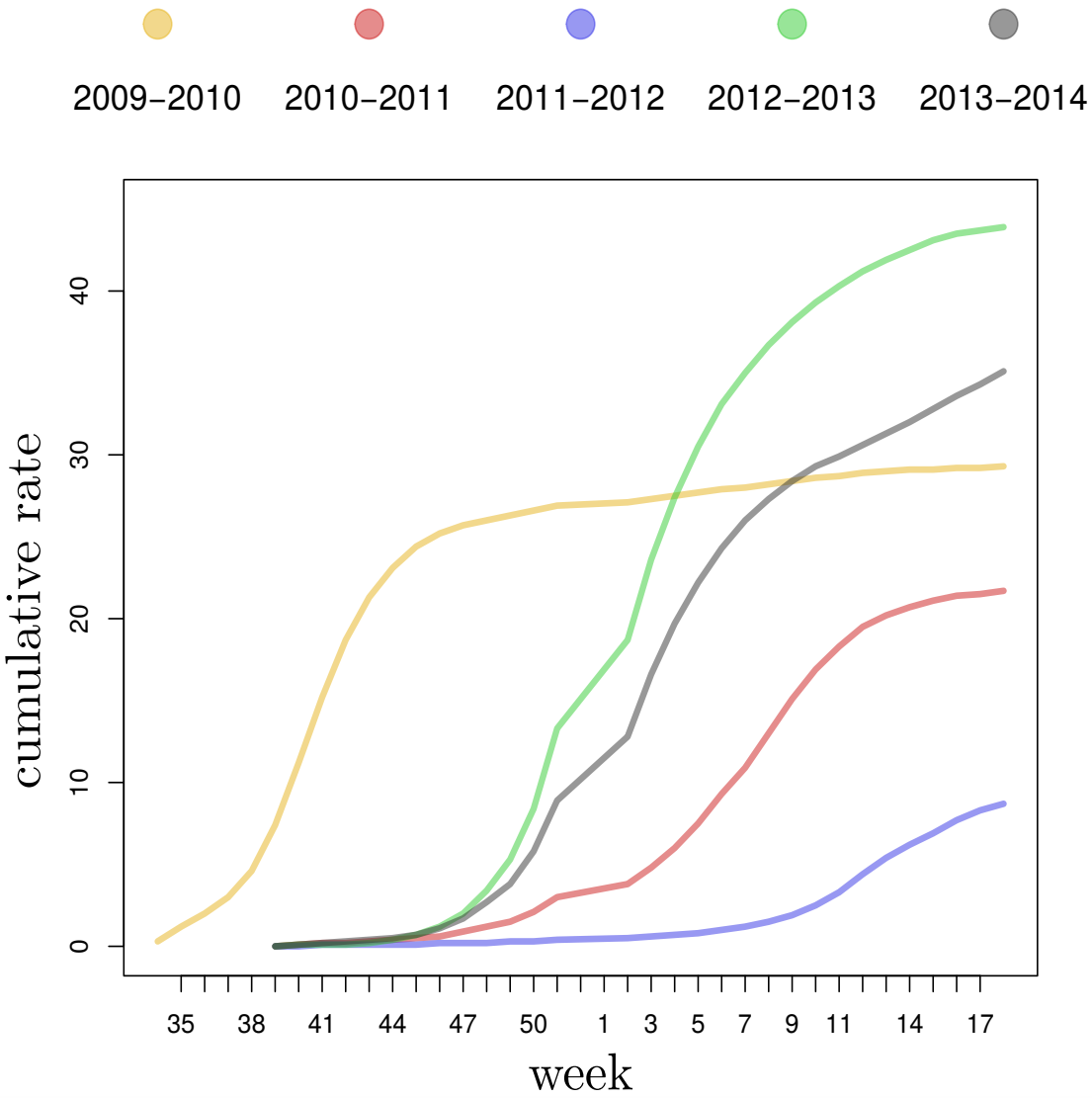
Influenza data

- NA segment of the Influenza A virus (H1N1 sub-type)
- Five flu seasons (2009-2010 to 2013-2014) in the USA.
- Hawaii and Alaska excluded in order to approximate the shape of the habitat with a rectangle.
- Two non-overlapping sets of sequences analysed for each season $\mapsto$ two biological replicates (with a single sequence per state for each set).

Influenza data



Influenza data



Future work

- Multiple unlinked loci.
- Analytical shortcuts: integrate out "empty" events.
- Non-isotropic dispersals (model/test barriers to gene flow).
- Expanding/shrinking populations and/or habitat.

Thank you

- Bayesian inference under ΛV model implemented in **phyrex** software as part of the PhyML package:

  https://github.com/stephaneguindon/phyml

- Article to appear in *Theoretical Population Biology*.