Scalable and robust mechanistic integration of epidemiological and genomic data for phylodynamic inference

Hannah Waddel
Advisors: Max Lau, Lance Waller

EMORY | ROLLINS SCHOOL OF PUBLIC HEALTH

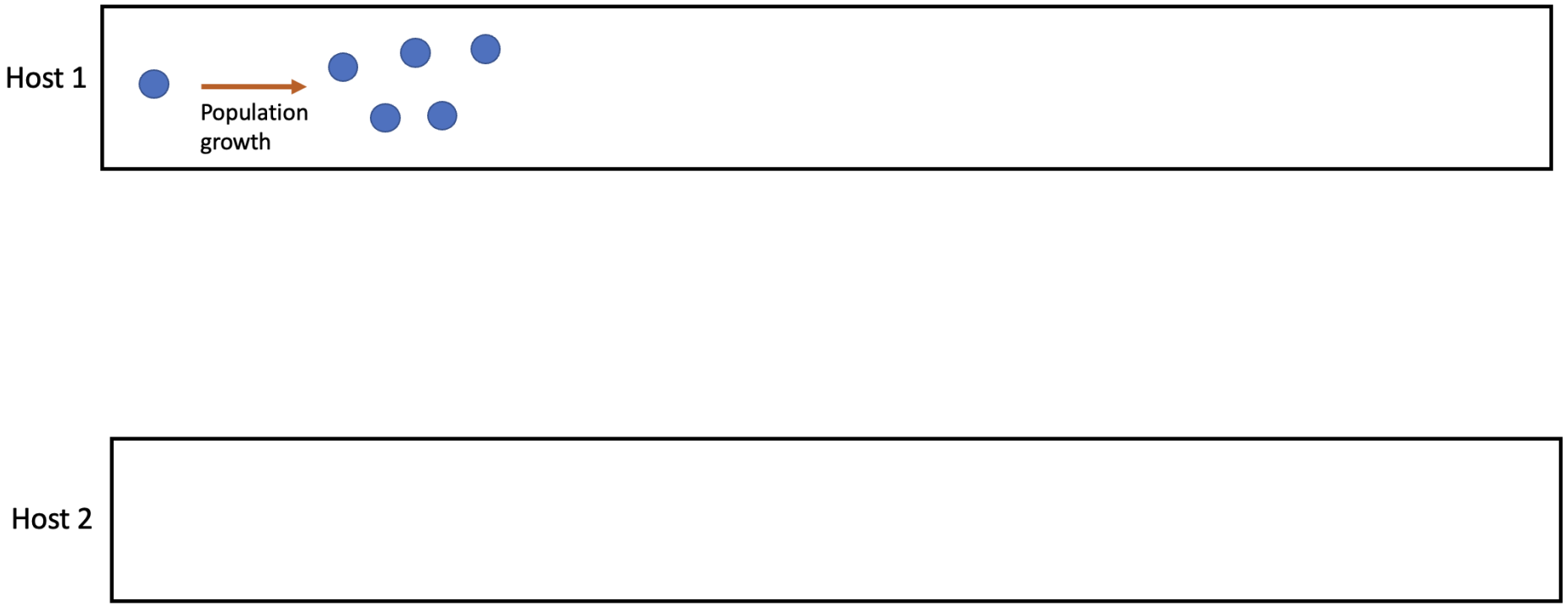
Problem setting

Introduction

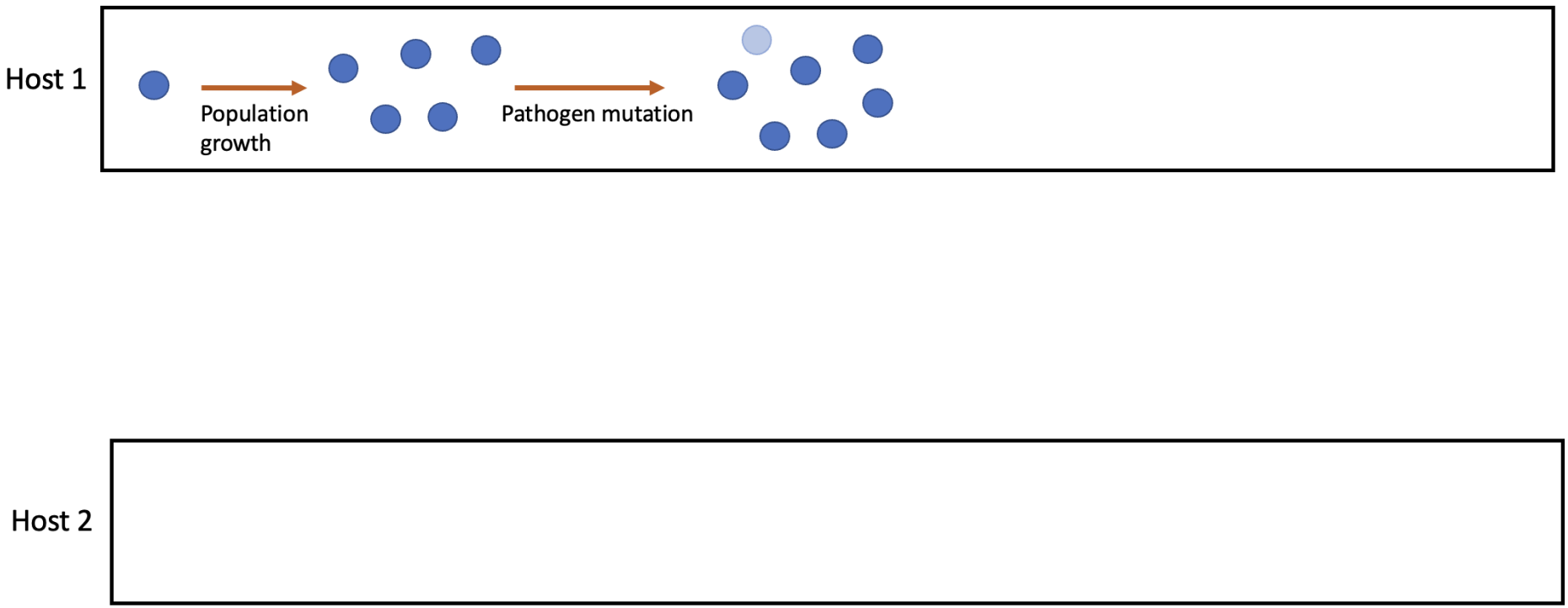
- **Mutation of pathogens** may occur on the **same time scale as disease transmission**
- Host carries a population of pathogens, which mutate as they replicate
- Transmission event moves a sample of the pathogen population to a new host
- Assumption: **More closely related** pathogen genetic samples are **more likely to be connected** by a transmission event

Pathogen Outbreak Process

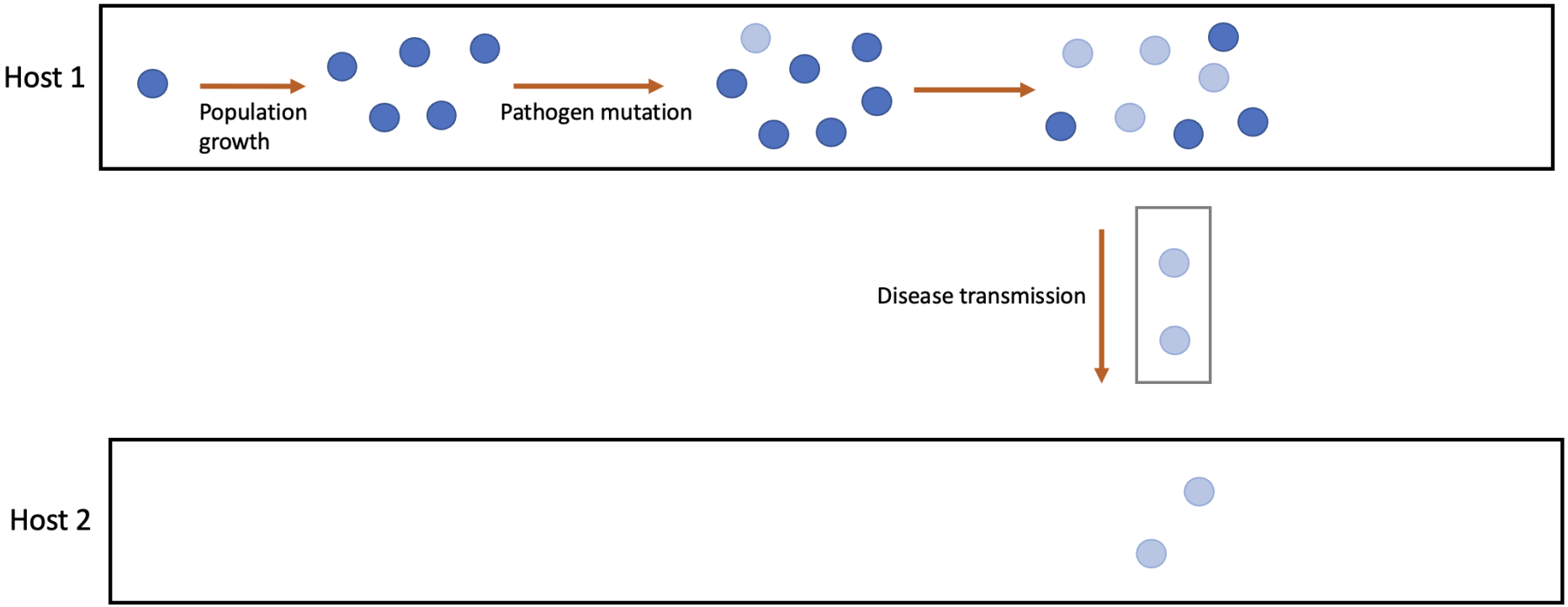
Pathogen Outbreak Process



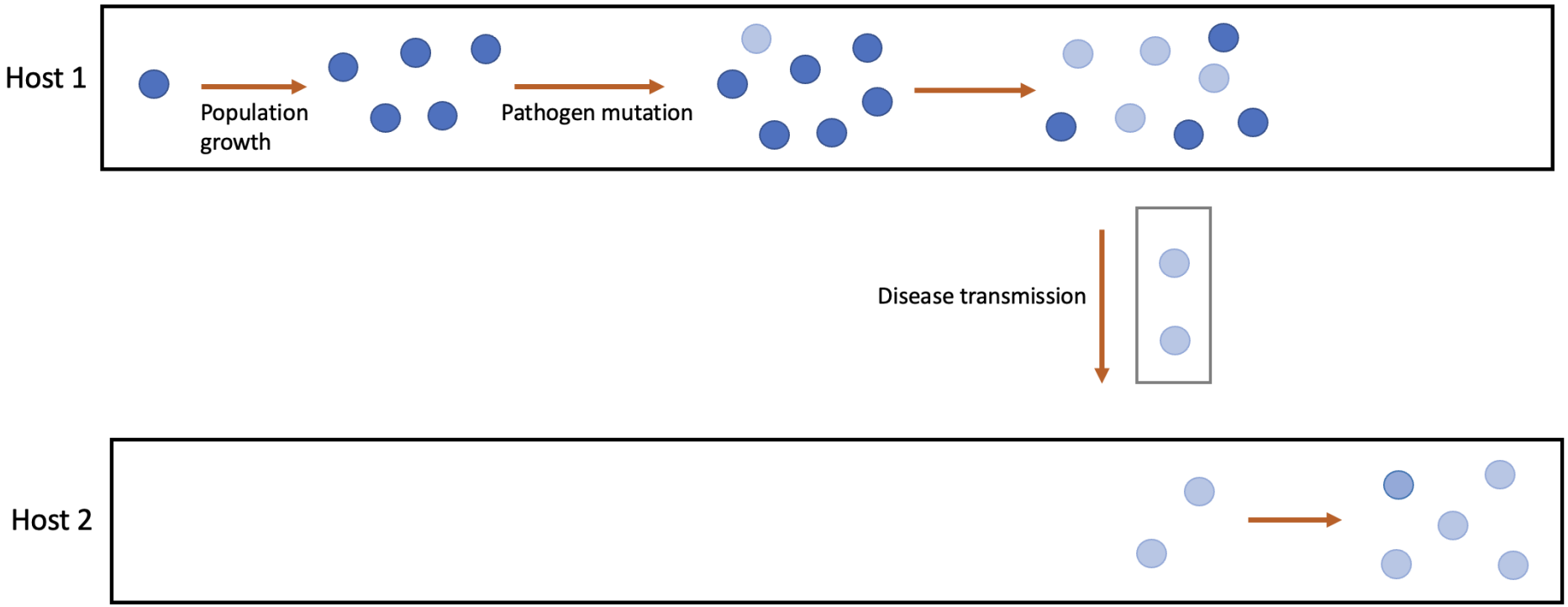
Pathogen Outbreak Process



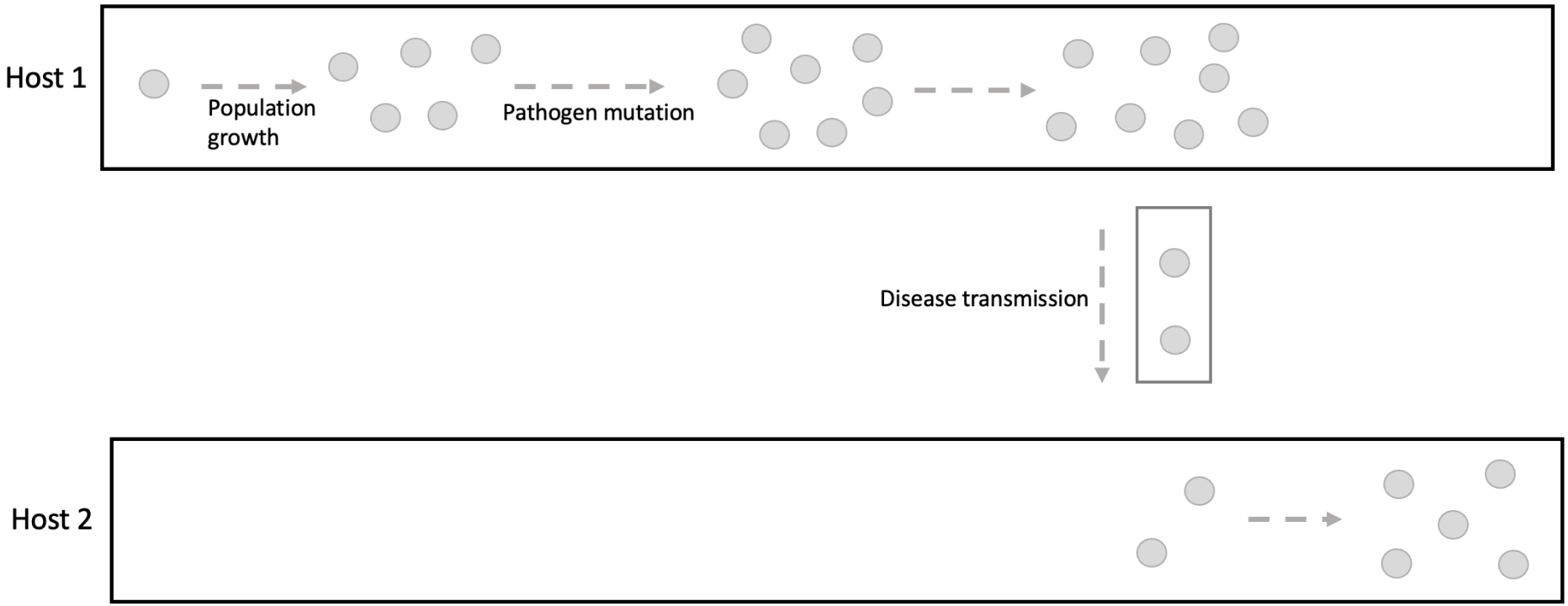
Pathogen Outbreak Process



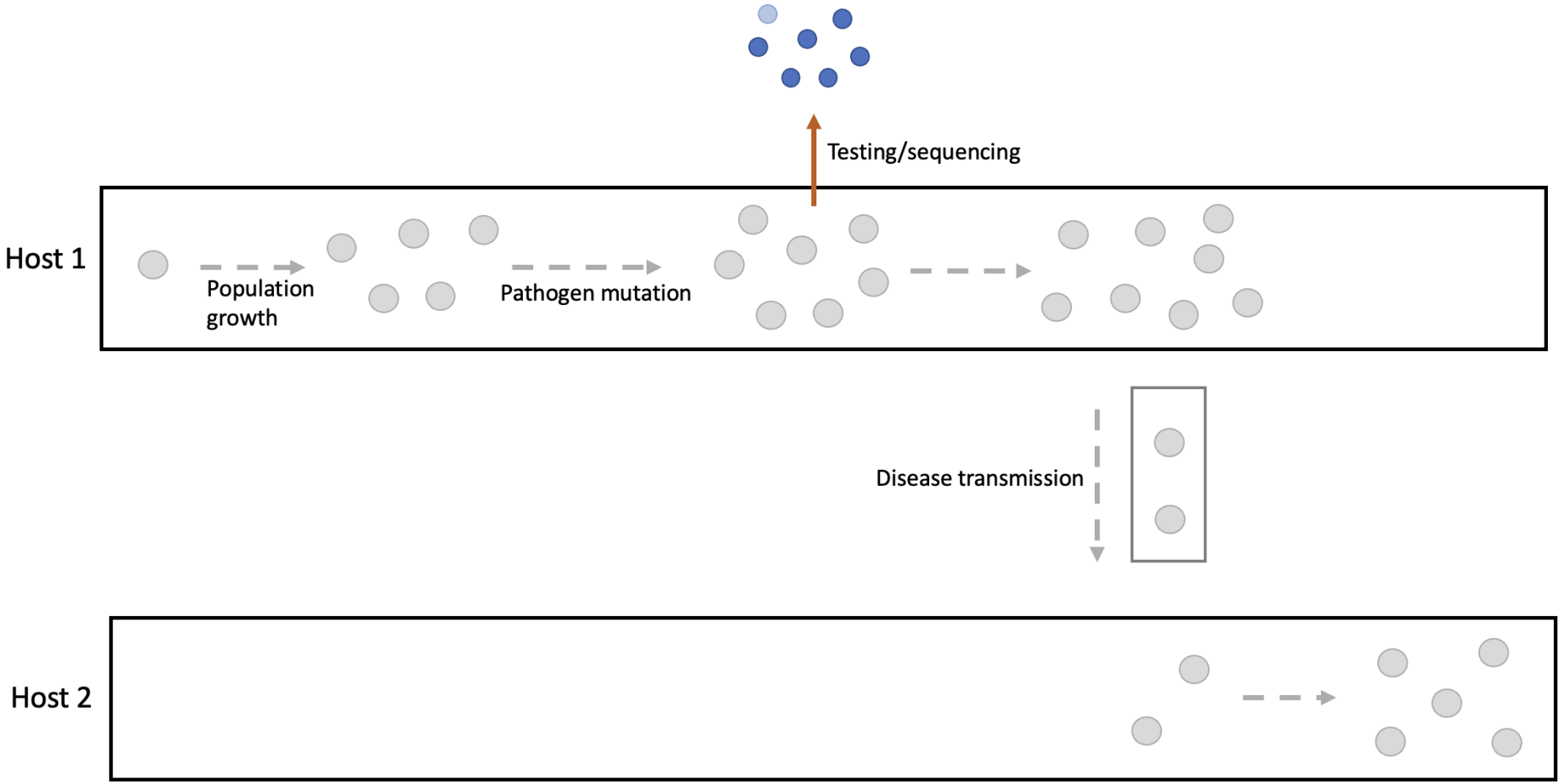
Pathogen Outbreak Process



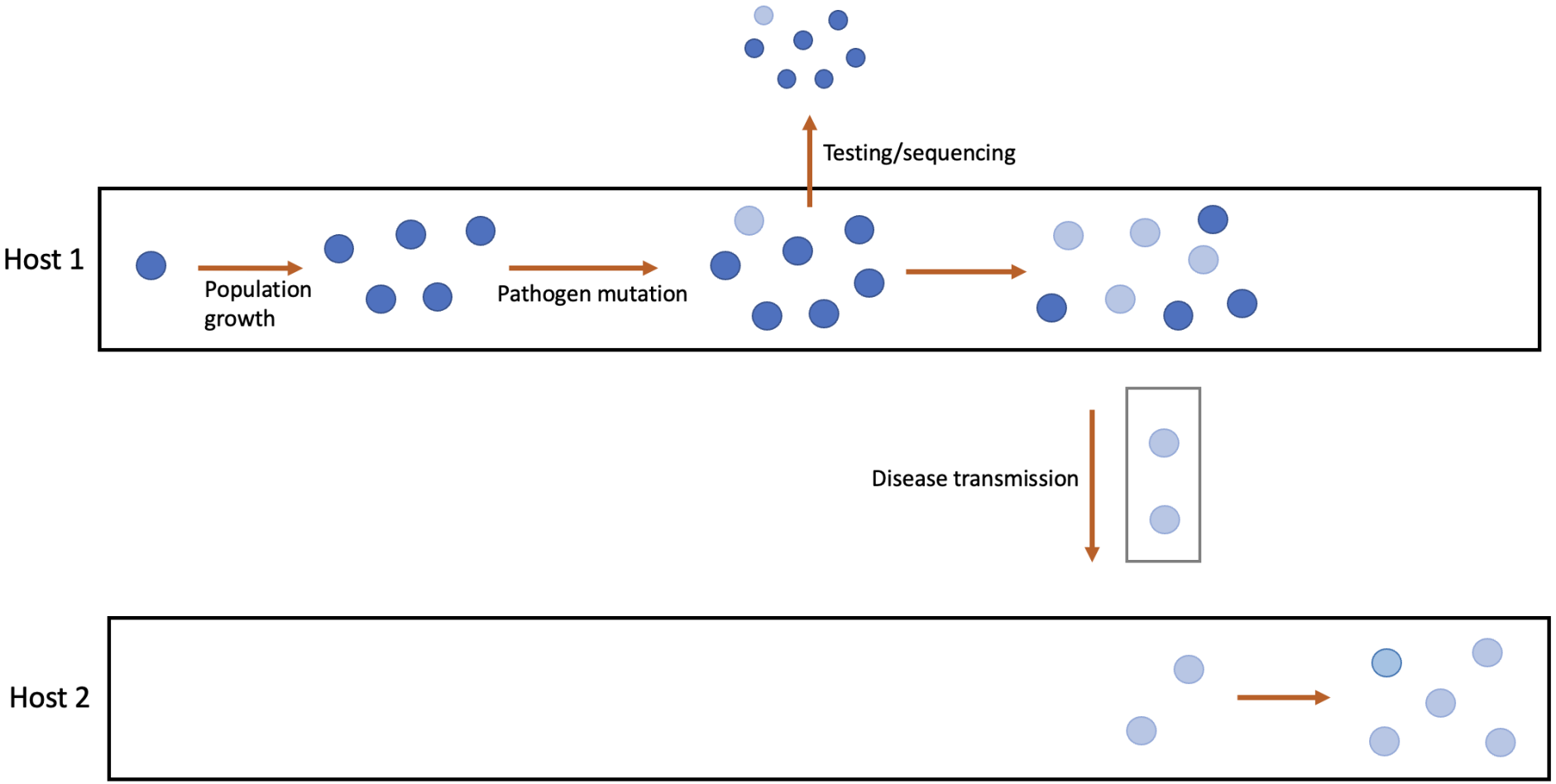
Pathogen Outbreak Process



Pathogen Outbreak Process

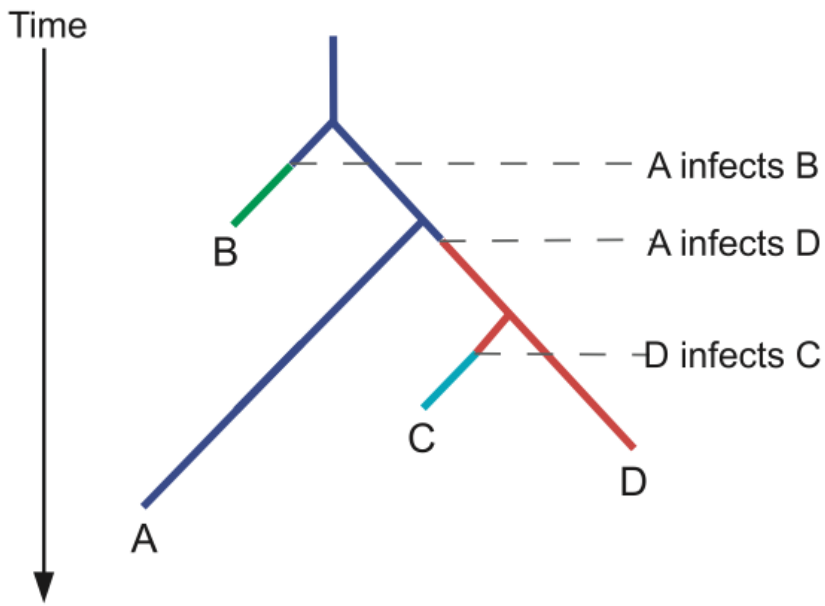


Pathogen Outbreak Process



Introduction

- **Phylogenetics** studies evolutionary history and relationship among organisms
- Simple phylogeny will not capture the transmission dynamics properly (Ypma et al., 2013)
  - Phylogeny does not show direction of transmission
  - Ancestors and descendants may be sampled



A phylogeny generated by a small outbreak. (Didelot et al., 2014)

Phylodynamics

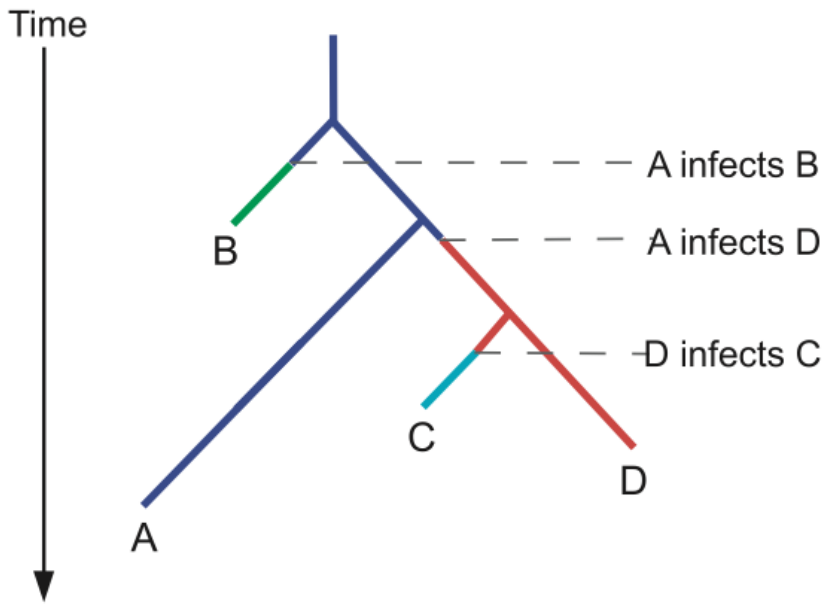
- **Phylodynamic** models unify **phylogenetics** and **epidemiology** (Grenfell et al., 2004)
- Explicitly account for the interacting dynamics of transmission and mutation
- Modeling of data-generating processes within host and pathogen populations (Klinkenberg et al., 2017):
  - Mutations in DNA/RNA sequence
  - Within-host evolution of pathogen population into variant subpopulations
  - **Transmission network** of "who-infected-whom" and **timing** of transmission
  - Case observations (unsampled hosts)

Phylodynamic Applications

- **Basic reproduction number** R_0
- Original pathogen **source** and **timing**
- Effectiveness of **control efforts**
- Rate of **spread**
- Viral population size
- Transmission **risk** and **population heterogeneity**

Previous Phylodynamic Methods

- Two stage: infer epidemiological quantities after inferring phylogeny
- Phylogeny does not depend on epidemic transmission



(Didelot et al., 2014)

Motivating Data

- Swine influenza H1N1 and H3N2 outbreak among pigs at a county fair
- Pigs act as **"mixing vessels" for different subtypes** of influenza

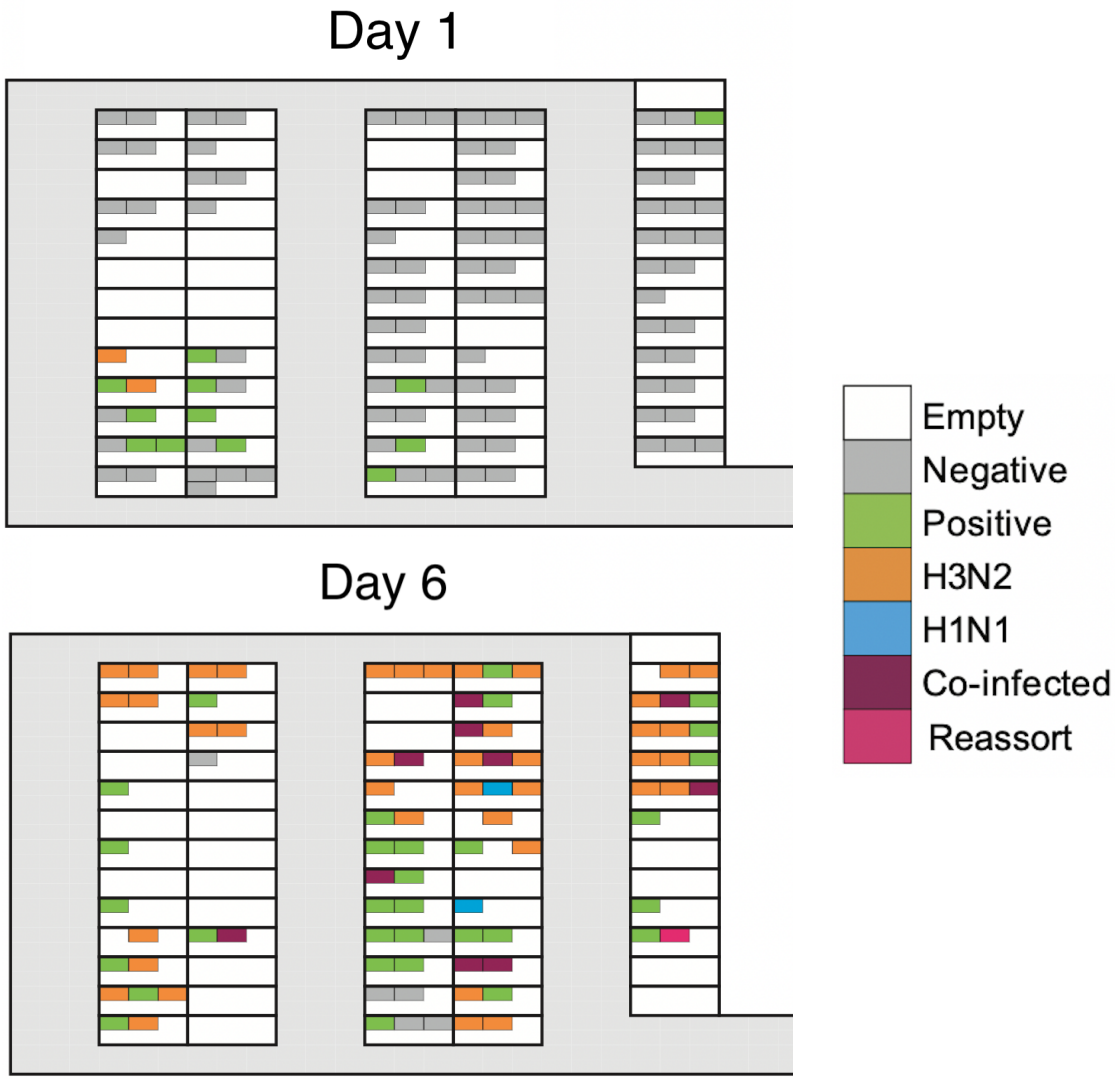
Andrew Bowman (OSU) sampling pigs. Source: *Science* Magazine

Motivating Data

- Pigs **densely sampled** with a nasal swab influenza test at weigh-in, nightly, and at auction
- 2,729 tests performed on 425 pigs over 7 days
- 408 pigs tested positive at least once before the end of the fair

A champion barrow (male) and gilt (female) from the fair.

Motivating Data



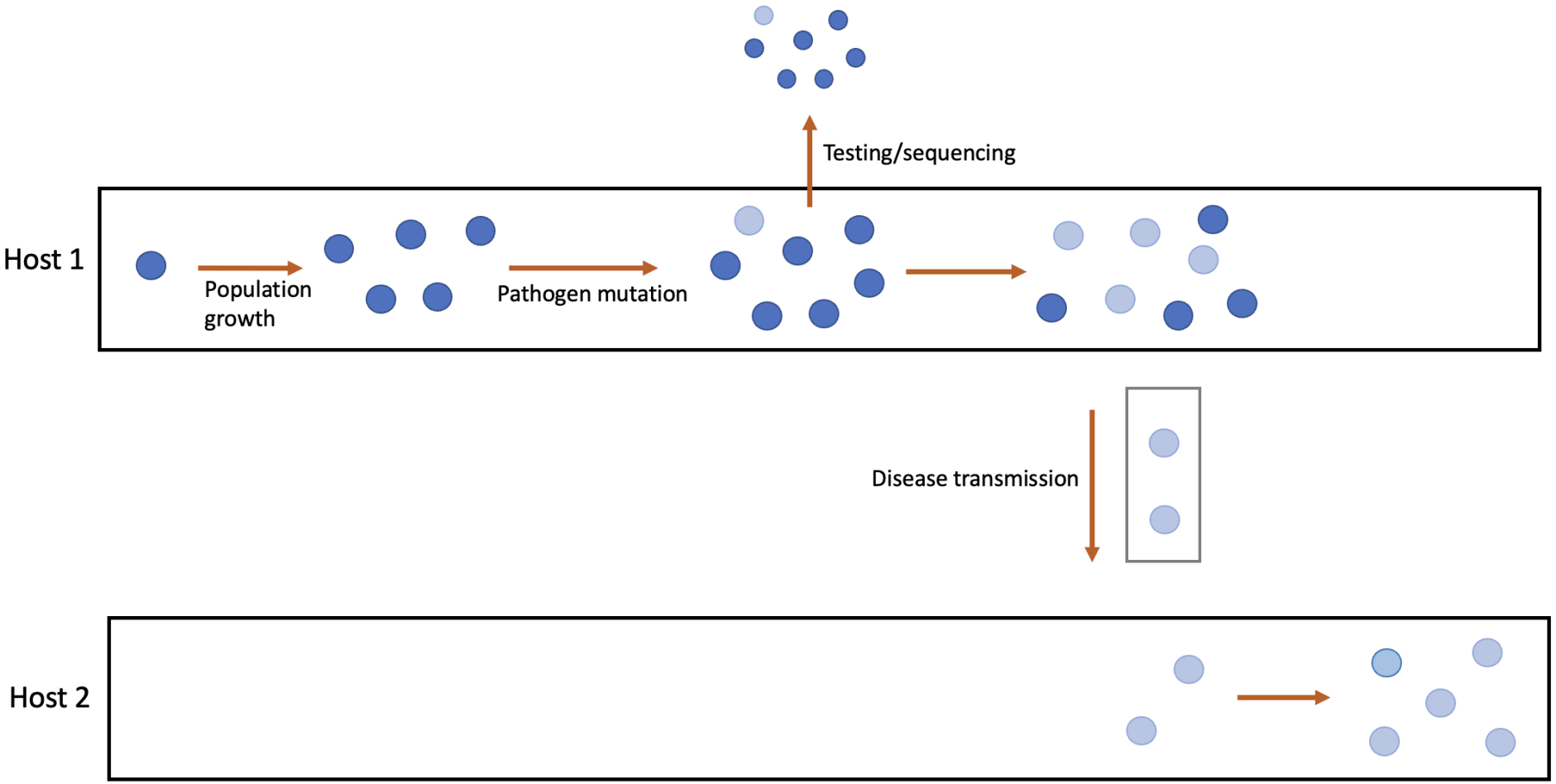
Big Picture Importance

- Variant viruses that originate and spread at county fairs can jump to humans
- Hypothesized 2009 H1N1 pandemic influenza origin in commercial swine farms in Mexico

Close interaction between humans and pigs at a county fair. Source: *Science* Magazine

Model Framework

Pathogen Outbreak Processes

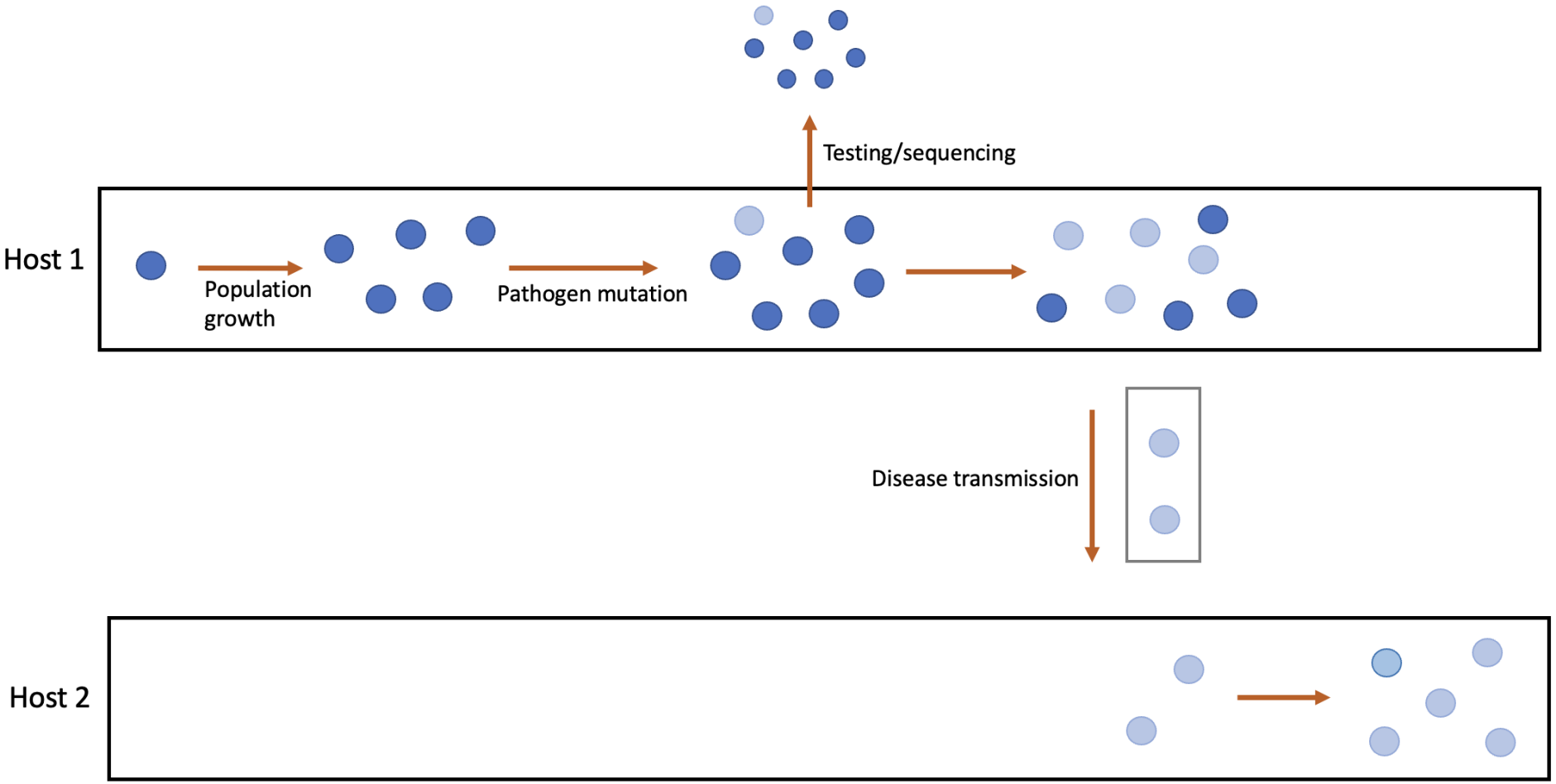


Complete-data Likelihood

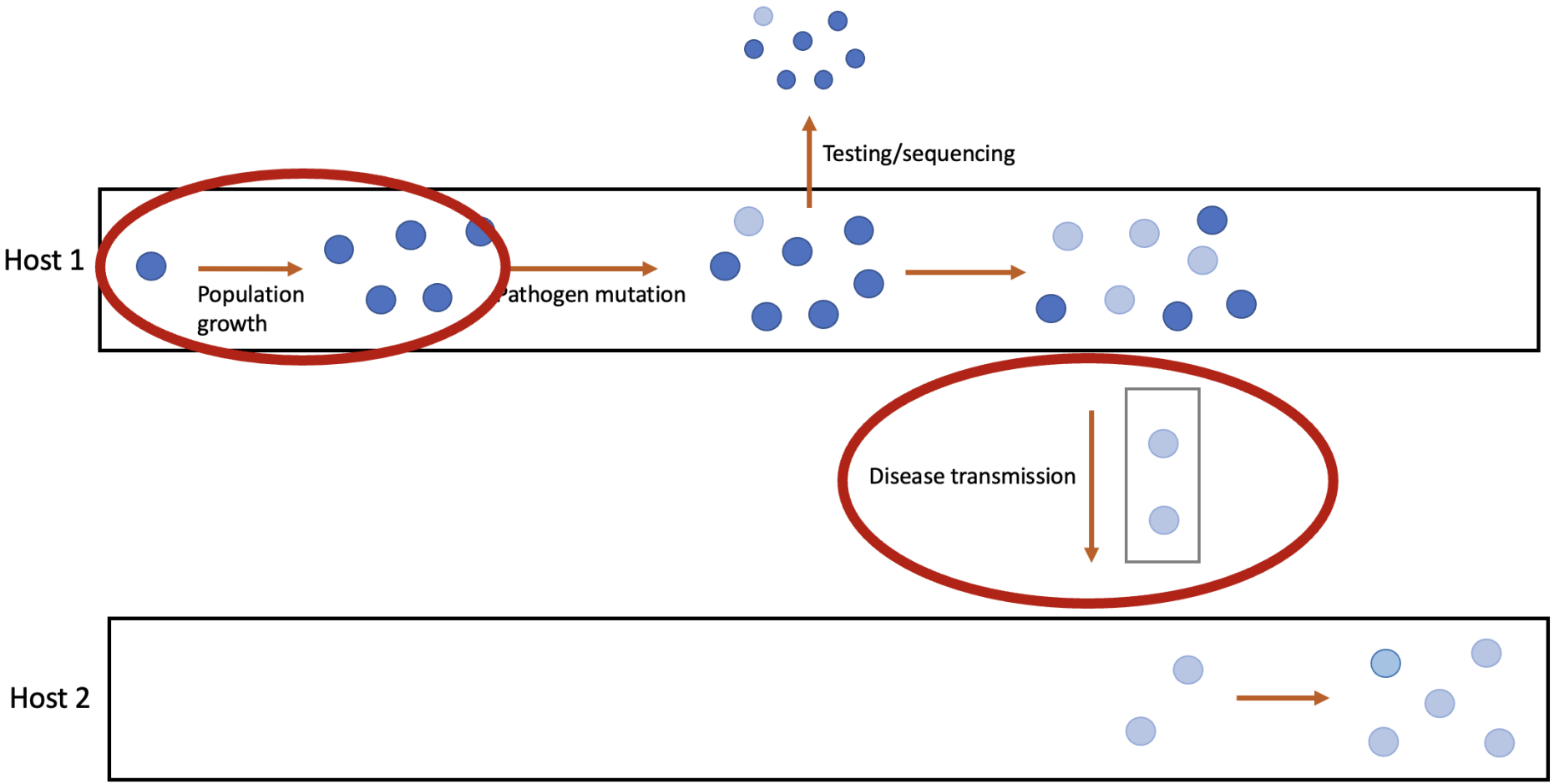
- Complete-data likelihood ties observed and unobserved processes together
- Epidemiological process: model parameters (Θ), transmission network (Ψ), transmission times $\mathbf{T}$
- Genomic process: transmitted and sampled genomic sequences ($\mathbf{G}$), mutation parameters in Θ
- Putting it all together, our complete-data likelihood is

$$L\left(\Theta; \mathbf{T}, \mathbf{G}, \Psi\right) = L\left(\Theta; \mathbf{T}, \Psi\right) \times L(\Theta; \mathbf{G} | \mathbf{T}, \Psi)$$

Pathogen Outbreak Processes

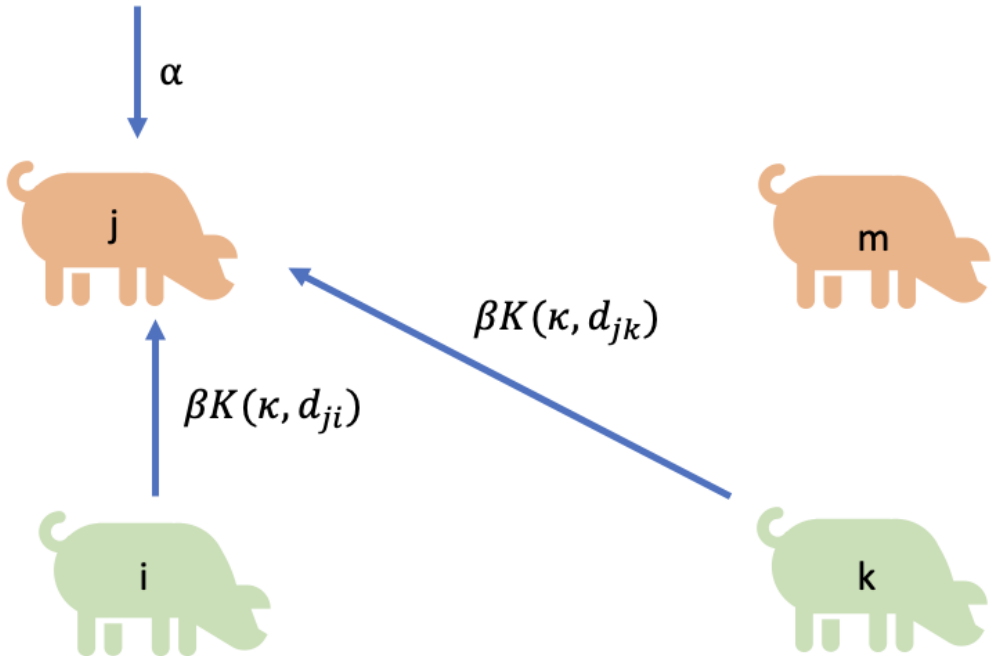


Pathogen Outbreak Processes



SEIR Model: Exposure

- Individuals begin in **susceptible** category
- Spatiotemporal process:
  - Exposure accumulates from currently infectious hosts
  - Closer hosts are a more probable source of infection, modeled via distance kernel $K(\kappa, d_{ij})$
- Small probability of exposure from the background (unobserved source, etc.)



SEIR Model: Accumulated Exposure

- Function $q_j(T)$ accumulates exposure to pathogen for individual j until time T:

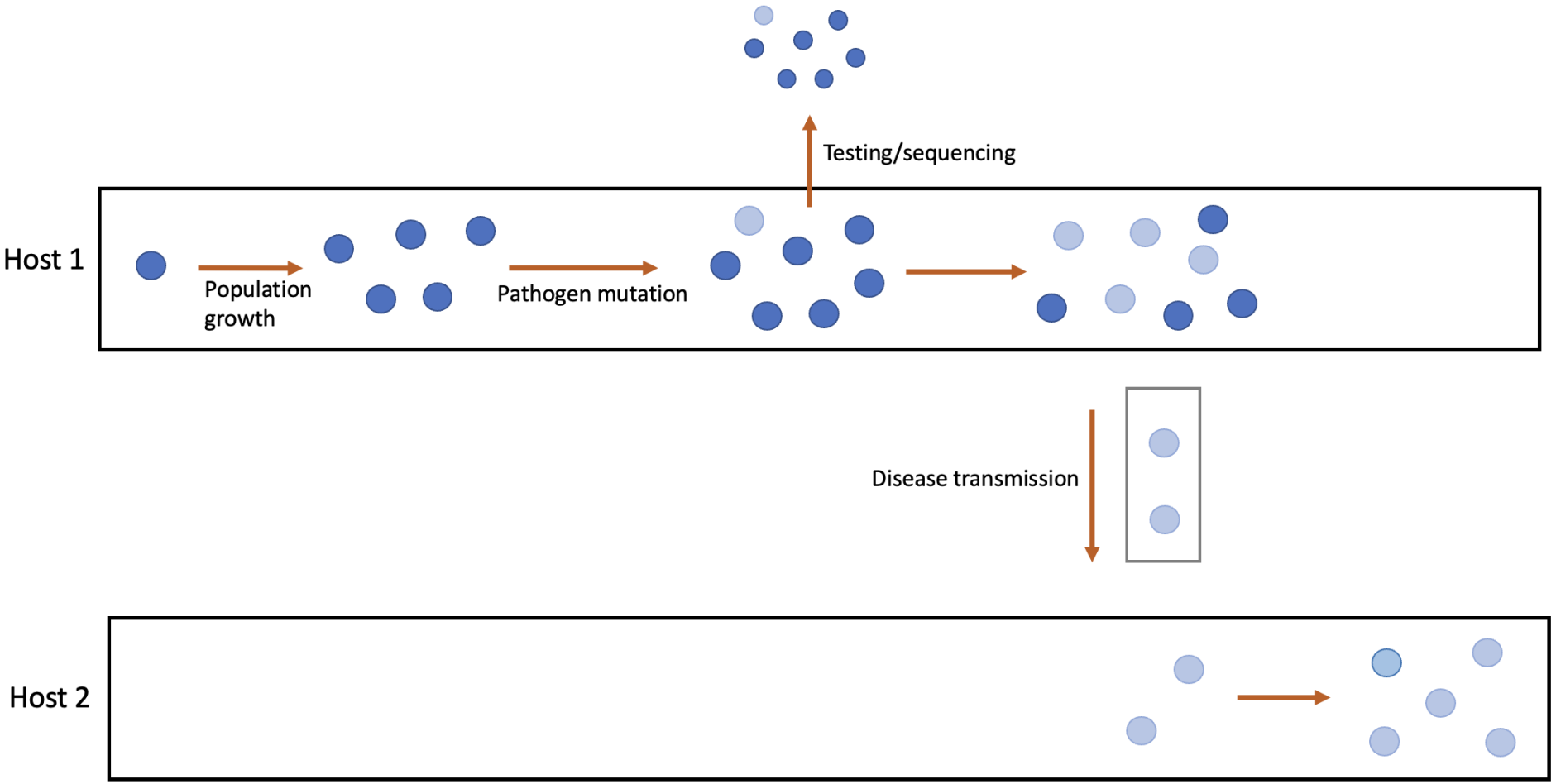
$$q_j(T) = \int_{t=0}^{T} \{\alpha + \sum_{i \in \chi_I(t), i \neq j} \beta * K(\kappa, d_{ij})\}$$

- When accumulated exposure reaches random, individual threshold, j switches from susceptible to **exposed**

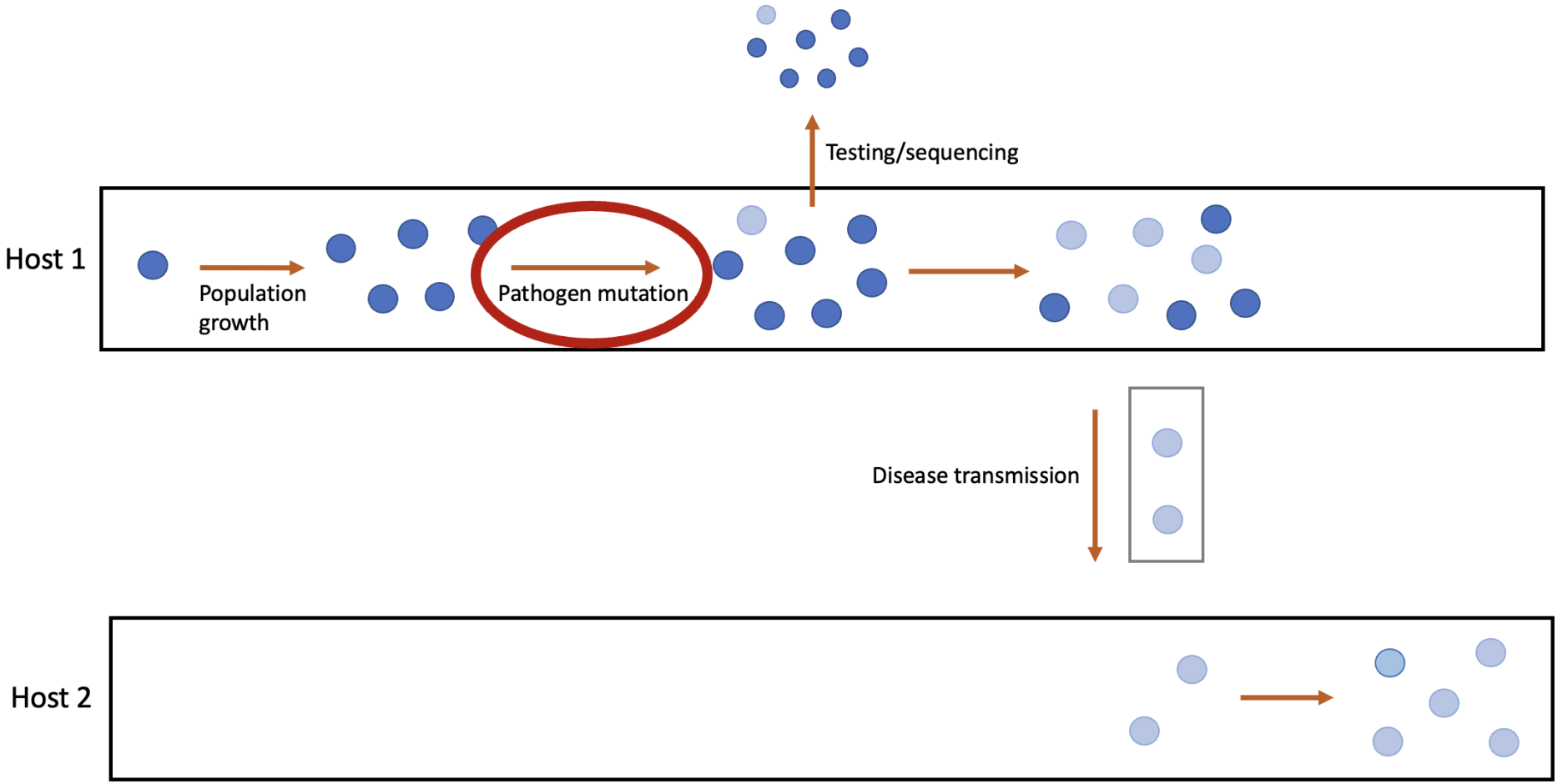
SEIR Model: Infection and Removal

- After an individual is exposed to a strain, it spends a sojourn time in the **exposed** compartment which follows $Gamma(a, b)$ distribution
- Individual spends sojourn time in the **infectious** compartment following a $Weibull(\gamma, \eta)$ distribution
- After infectious period, **recovery/removal** occurs

Pathogen Outbreak Processes



Pathogen Outbreak Processes

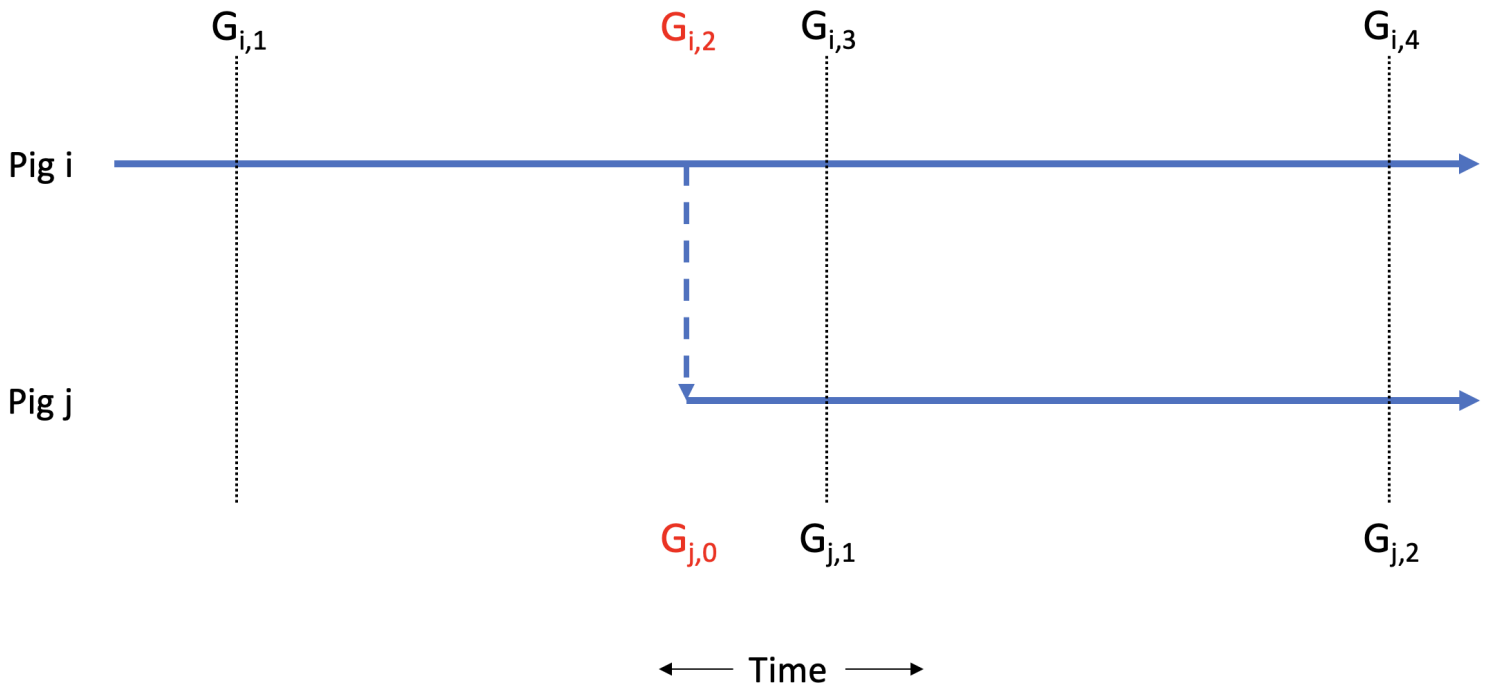


Lau 2015

- "A Systematic Bayesian Integration of Epidemiological and Genetic Data" (Lau et al., 2015)
- Joint single-stage inference of transmission network, exposure time, and genetic sequence of transmitted virus
- Genuine complete-data likelihood for data augmentation MCMC
- Performed best in methods comparison reconstructing Foot-and-Mouth Disease outbreak transmission network (Firestone et al., 2019)

Genomic Model: Lau 2015

- Pathogen genetic sequence mutates through time independently within each host
- Sequences $G_{i,1}, G_{i,2}, \ldots$ observed at sampling times
- Mutations from $G_{i,1} \to G_{i,2}$ and $G_{i,2} \to G_{i,3}$ are conditionally independent
- Calculating **complete data likelihood** requires $G_{i,2}$ and $G_{j,0}$



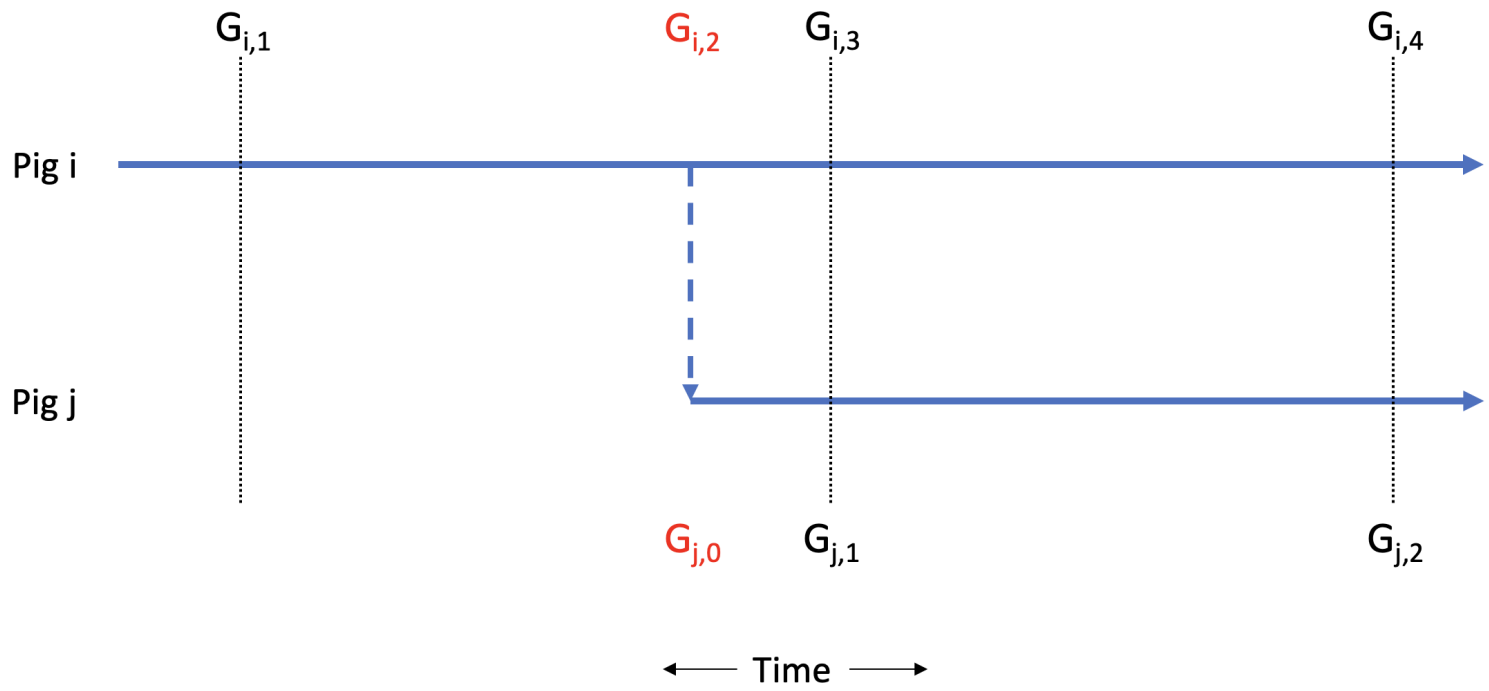
Kimura Mutation Model

- **Nucleotide bases** in sequence **mutate independently**
- Continuous-time Markov process: probability of any mutation increases through time
- Two-parameter Kimura Model (Kimura, 1980), rate of transition (μ_1) different than rate of transversion (μ_2)
  - **Transition**: Mutation within pyrimidines or purines (A to G or T to C, vice versa)
  - **Transversion**: Mutation between pyrimidines and purines (A to T/C, T to A/G, etc.)

$$P_{\mu_1,\mu_2}(y|x,\Delta t) = \begin{cases} 0.25 + 0.25e^{-4\mu_2\Delta t} + 0.5e^{-2(\mu_1+\mu_2)\Delta t}, \text{ for } x = y \\ 0.25 + 0.25e^{-4\mu_2\Delta t} - 0.5e^{-2(\mu_1+\mu_2)\Delta t}, \text{ transition} \\ 0.25 - 0.25e^{-4\mu_2\Delta t}, \text{ transversion} \end{cases}$$

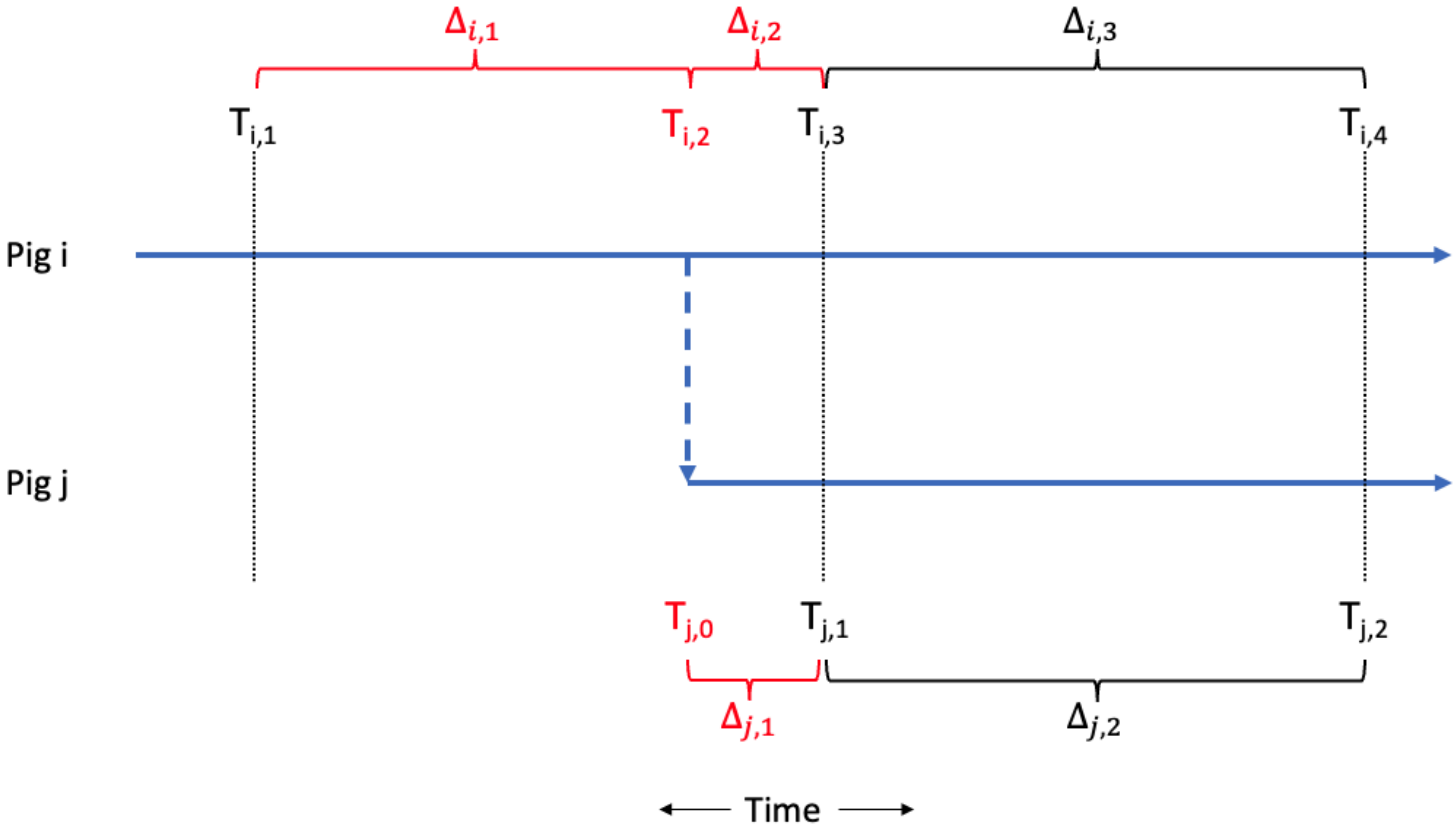
Genomic Model: Lau 2015

- Key innovation of model was imputing unobserved sequences
- **Problem**: Computation time and DNA storage hinders **scalability**



Our Proposal

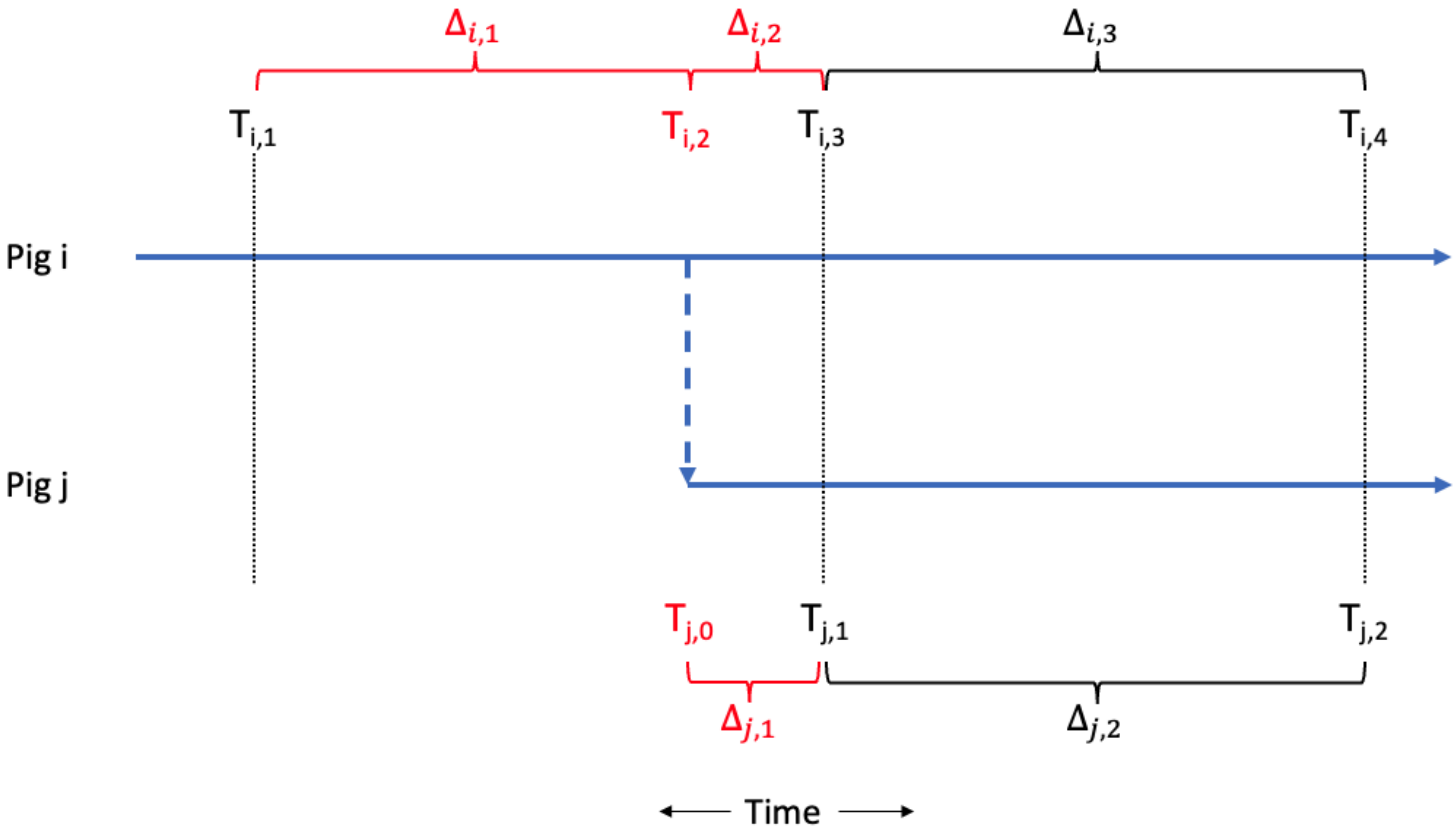
- Model genetic mutation at the **sequence level**
- Count of base pair mutations through time as a Poisson process
- "Infinite sites" model
  - Genetic sequence accumulates mutations, no reversion



Our Proposal

$$\Delta_{j,1} \sim \text{Poisson}(\lambda * \{T_{j,1} - T_{j,0}\})$$

$$\lambda = \text{average base pair mutations per unit time}$$



Implementation and Inference

Inference

- **Data augmentation Markov Chain Monte Carlo (MCMC)** to obtain posterior distributions on parameters θ and missing data y, given observed data x
  - Alternate update of θ and y
  - Update θ given x, y using $p(\theta|x, y)$
  - Update y given x, θ using $p(y|x, \theta)$
- **Challenge**: efficiently (and correctly!) propose values to **explore high-dimensional model space**

MCMC Algorithm

- Metropolis-Hastings algorithm framework (Hastings, 1970)
- Acceptance probability of proposed parameter θ':

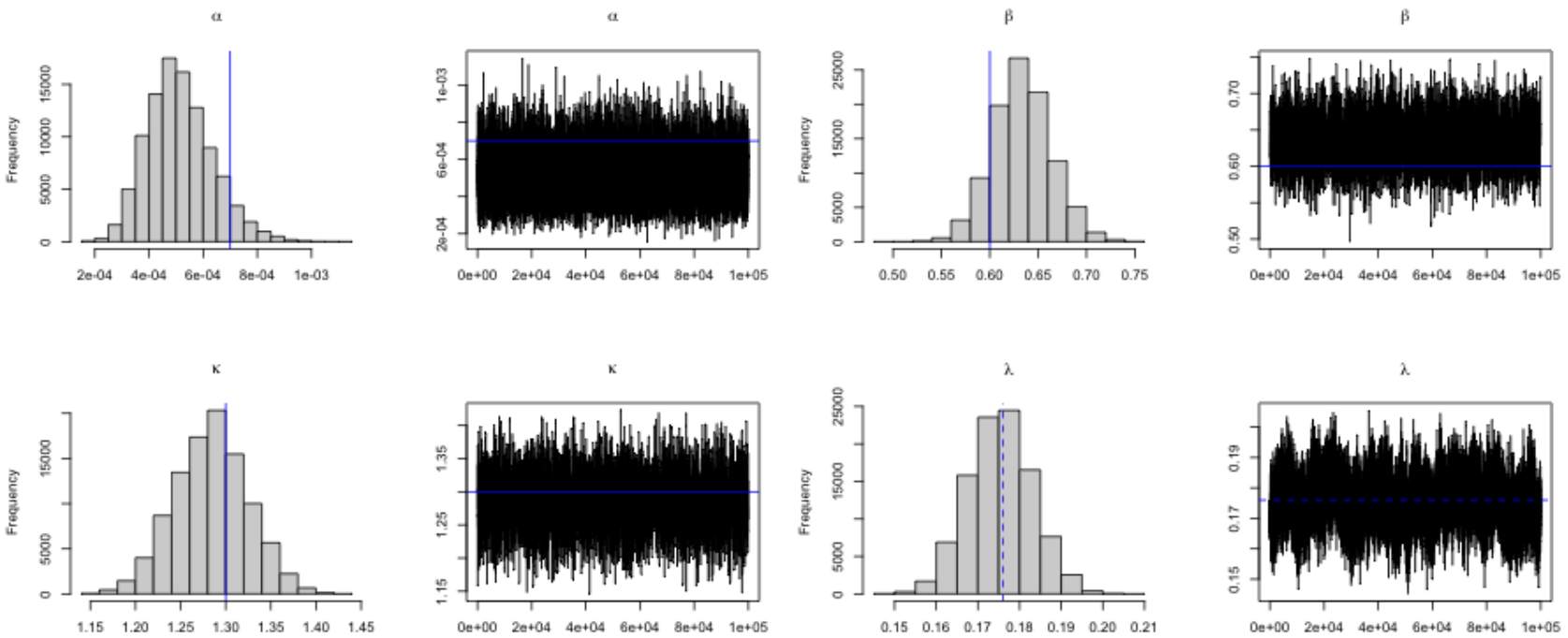
$$p_a = \min\{1, \frac{L(\theta'|z)}{L(\theta|z)} \times \frac{p(\theta')}{p(\theta)} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)}\}$$

- Where
  - L is the likelihood
  - p is the prior
  - q is the proposal distribution
- Scalar parameters $\{\alpha, \beta, \kappa, \lambda, ...\}$ proposed as random normal walk
- Everything else (transmission tree, infection time, genetic mutation, ...) is a custom algorithm **implemented in Rcpp for scalability**

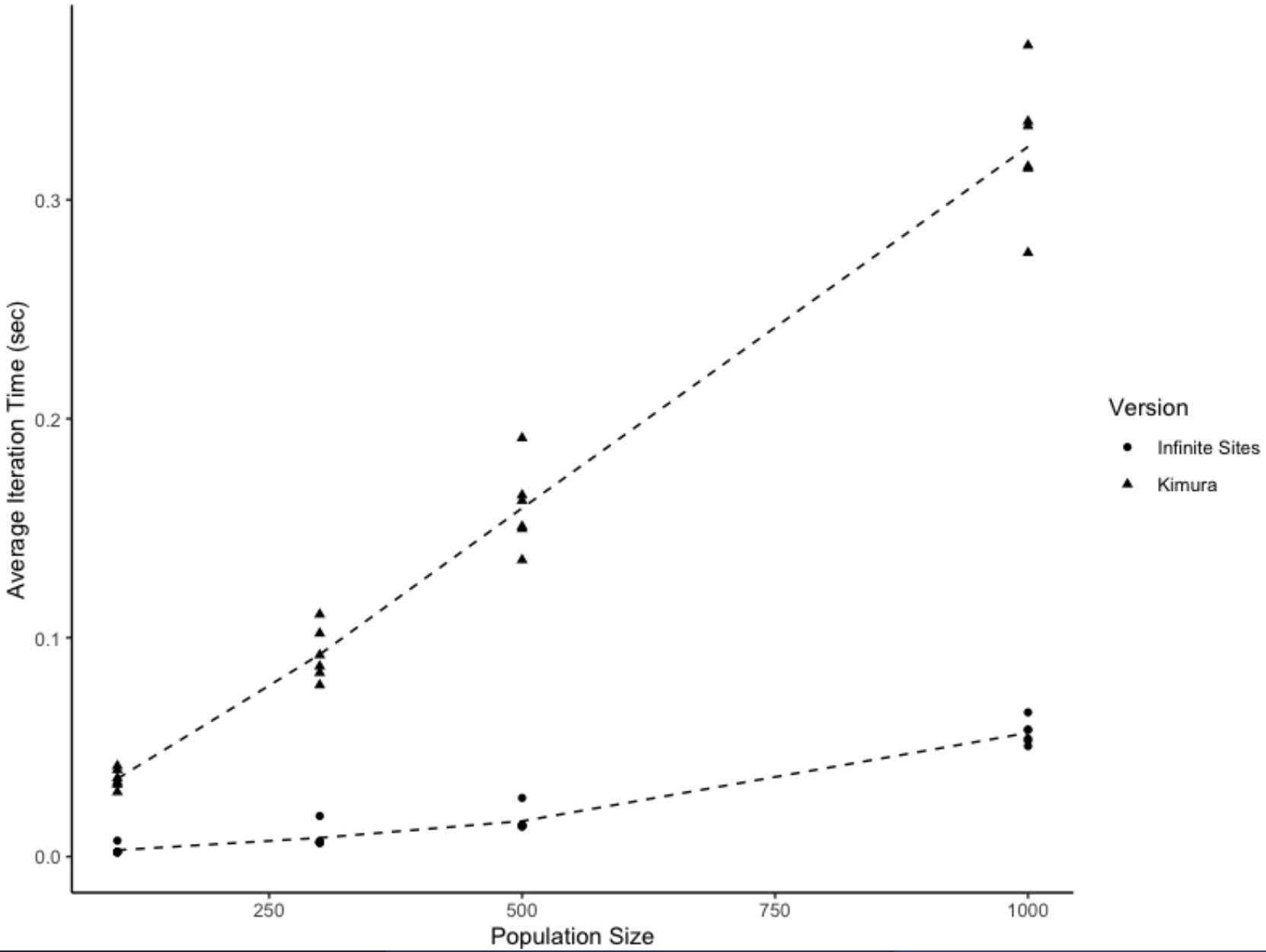
Preliminary Results

Model Fit

- Fit to full data simulated under the more complicated Kimura Model
- Estimation of key epidemic parameters generally robust to model misspecification

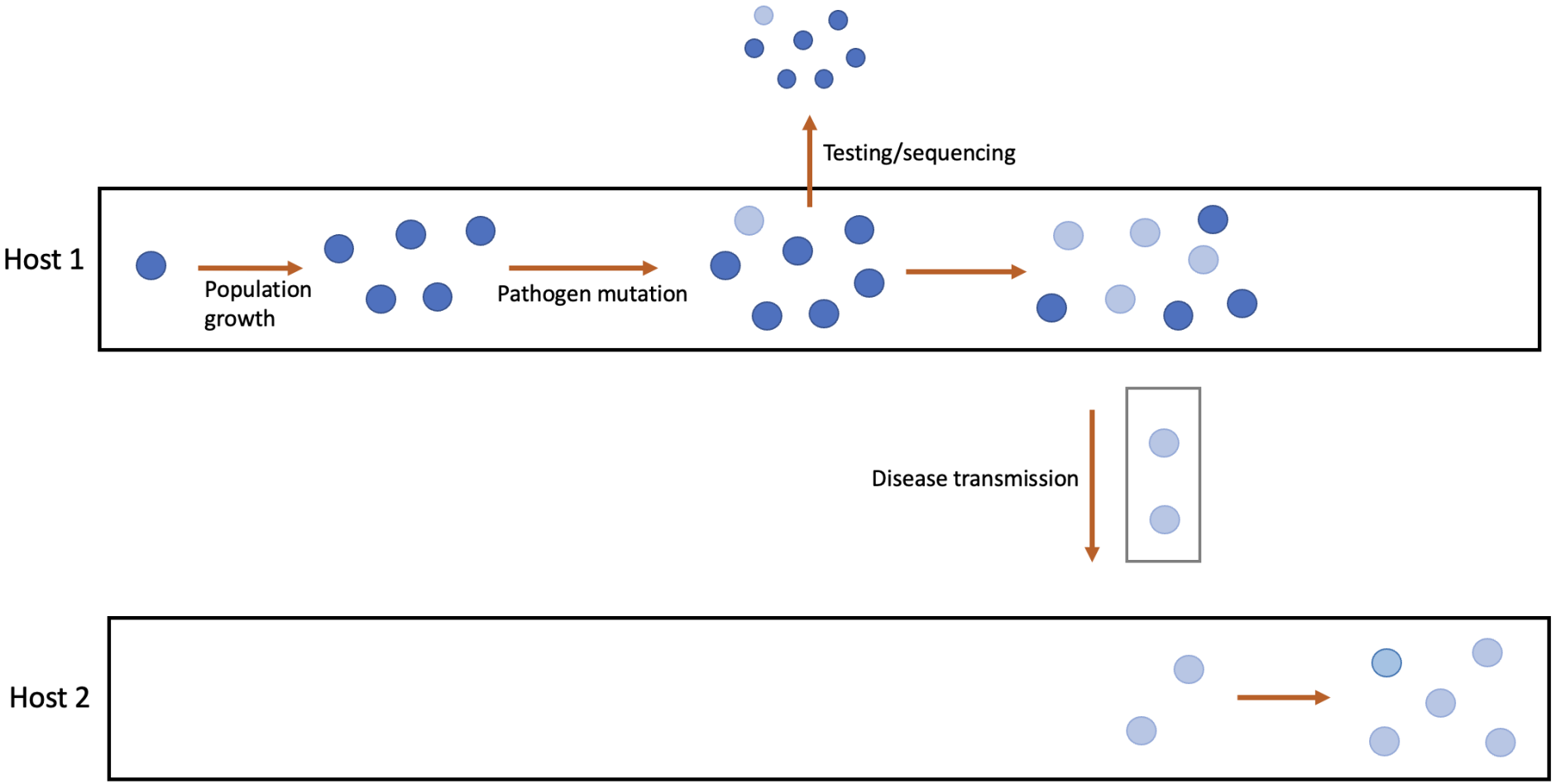


Computation Time

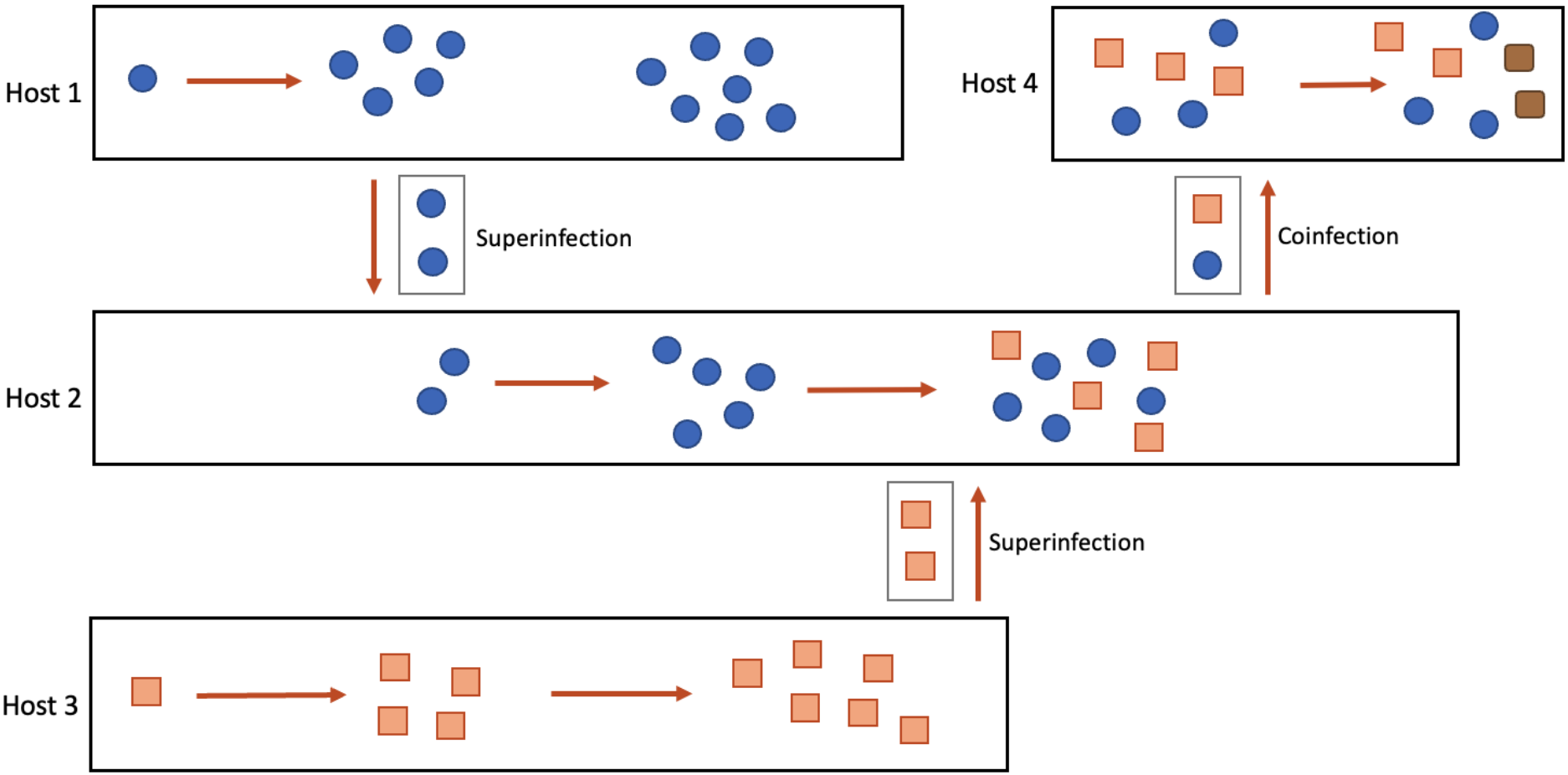


Topic 2: Multiple Circulating Subtypes

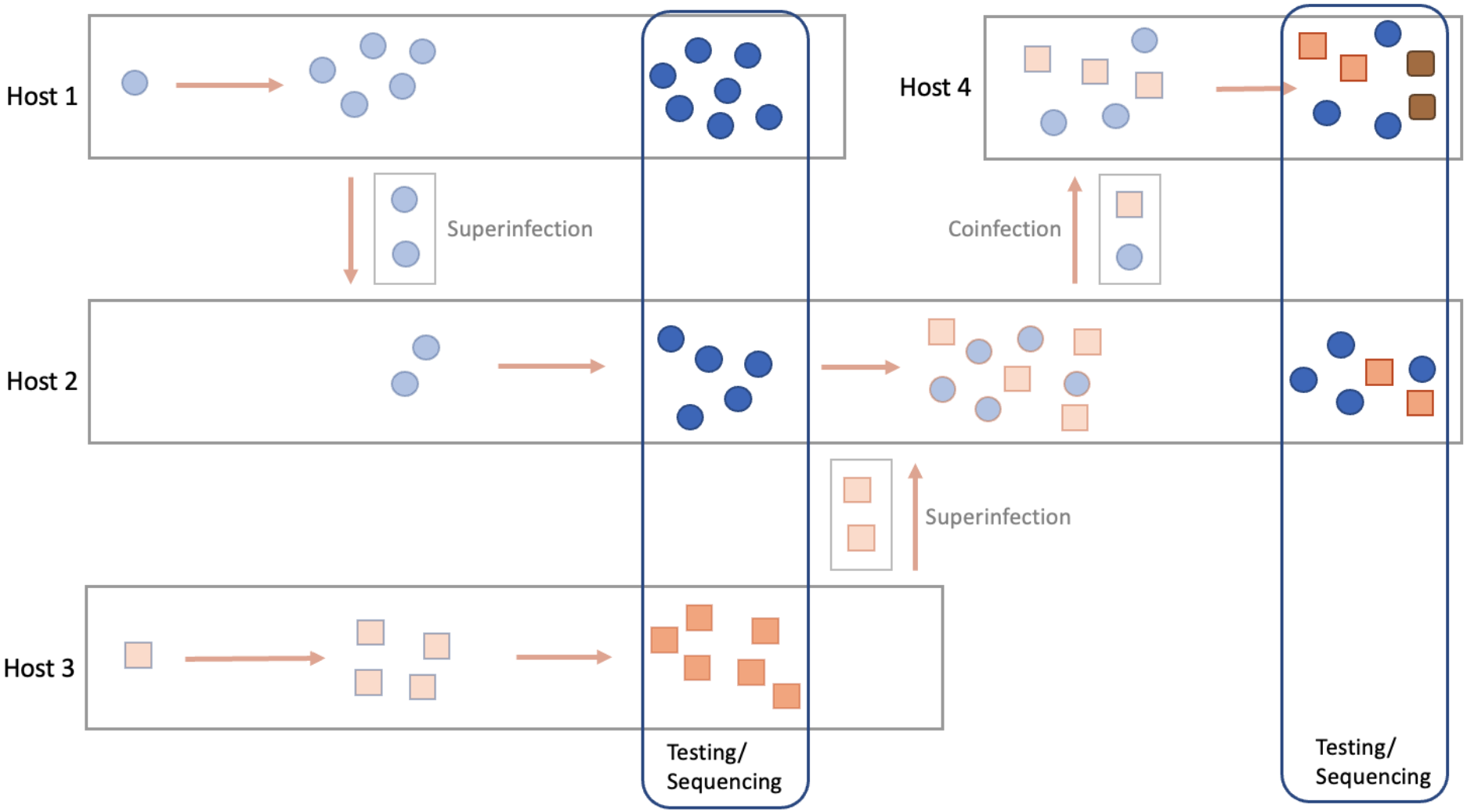
Pathogen Outbreak Processes



Multi-subtype Scenario

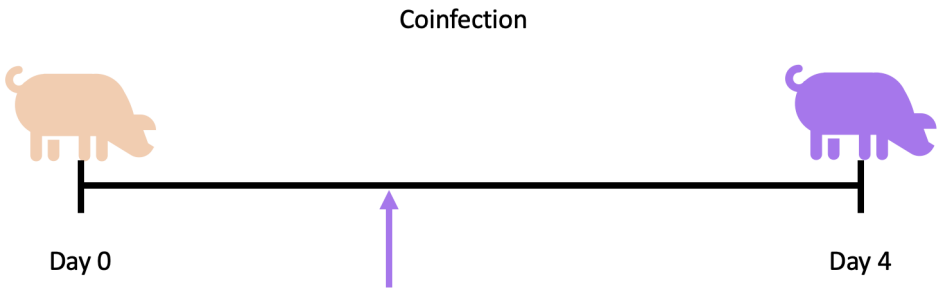


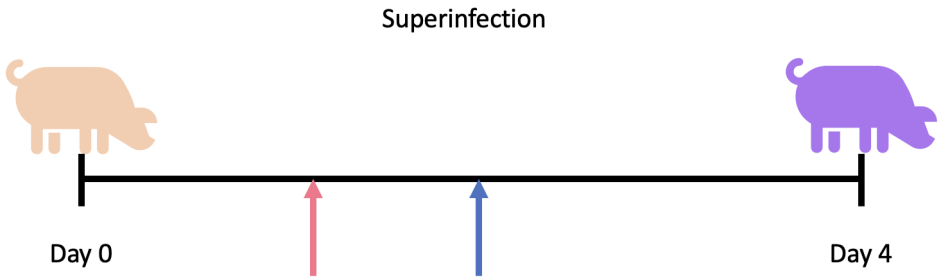
Multi-subtype Scenario



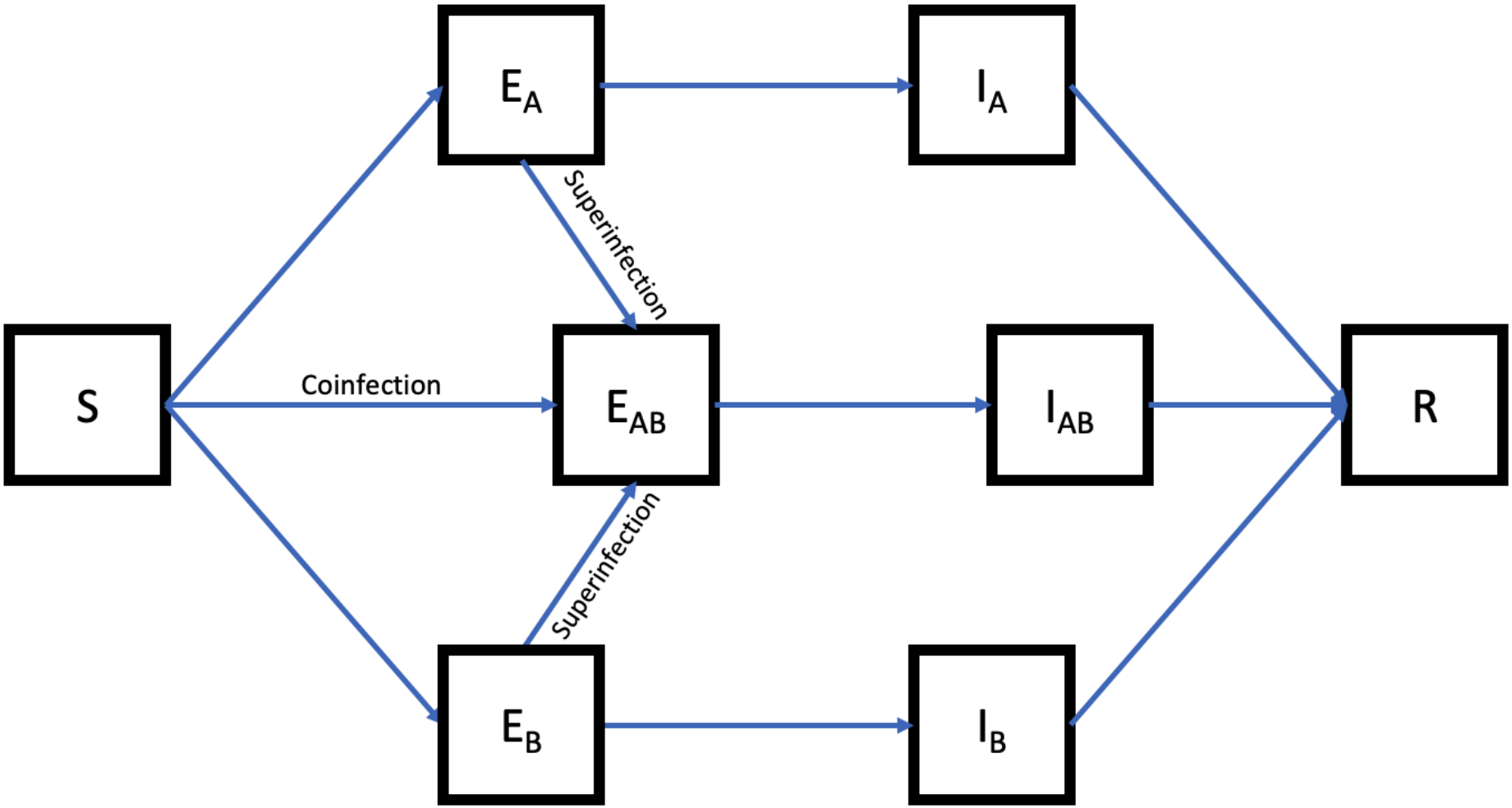
Multi-subtype Scenario

- More than one subtype of a pathogen can circulate and interact within hosts (**Influenza**, SARS CoV-2...)
- **Transmission network** complicated by coinfection and superinfection
  - **Coinfection**: Two subtypes of a pathogen transmitted by one host carrying both
  - **Superinfection**: Two subtypes of a pathogen transmitted by two hosts to one recipient



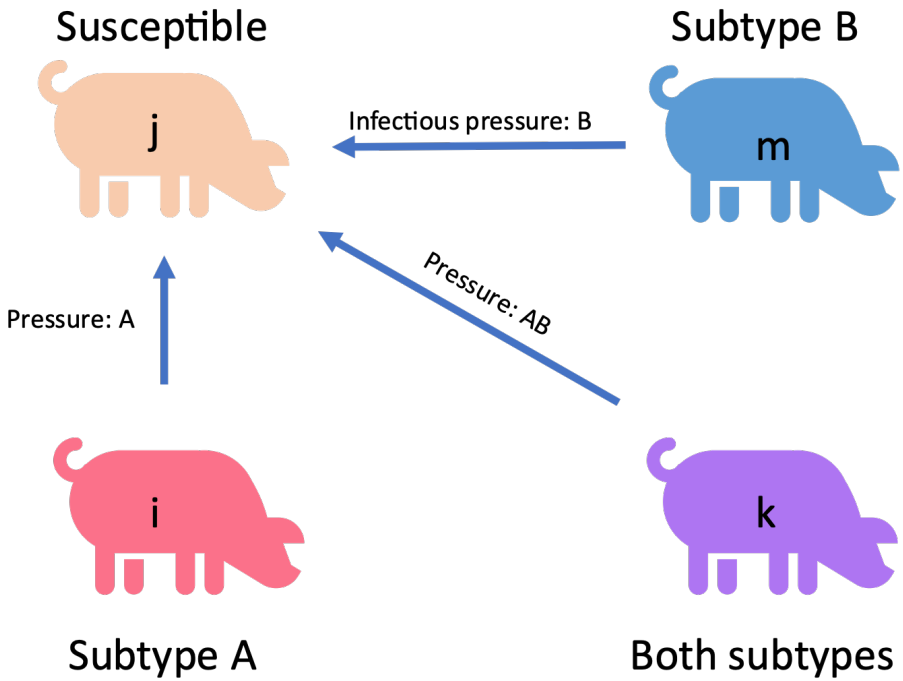


Proposed Model



Multi-subtype Exposure

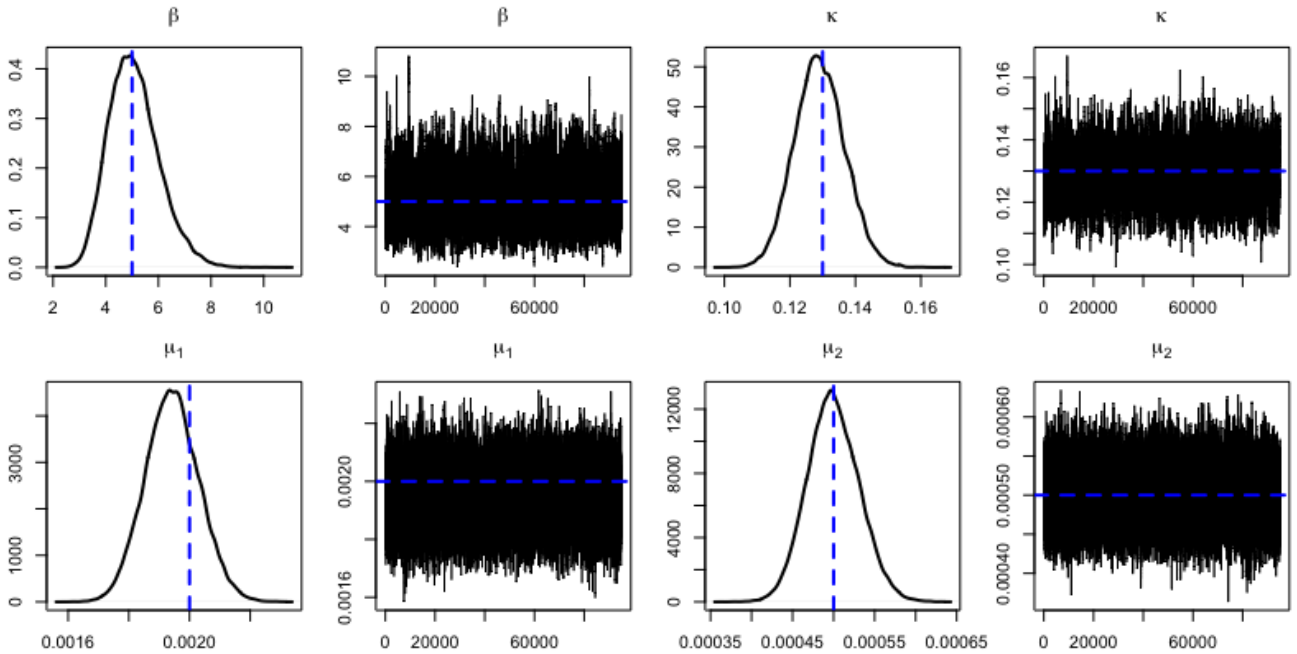
- A host may be exposed to subtype A, subtype B, or both at one time



MCMC Implementation

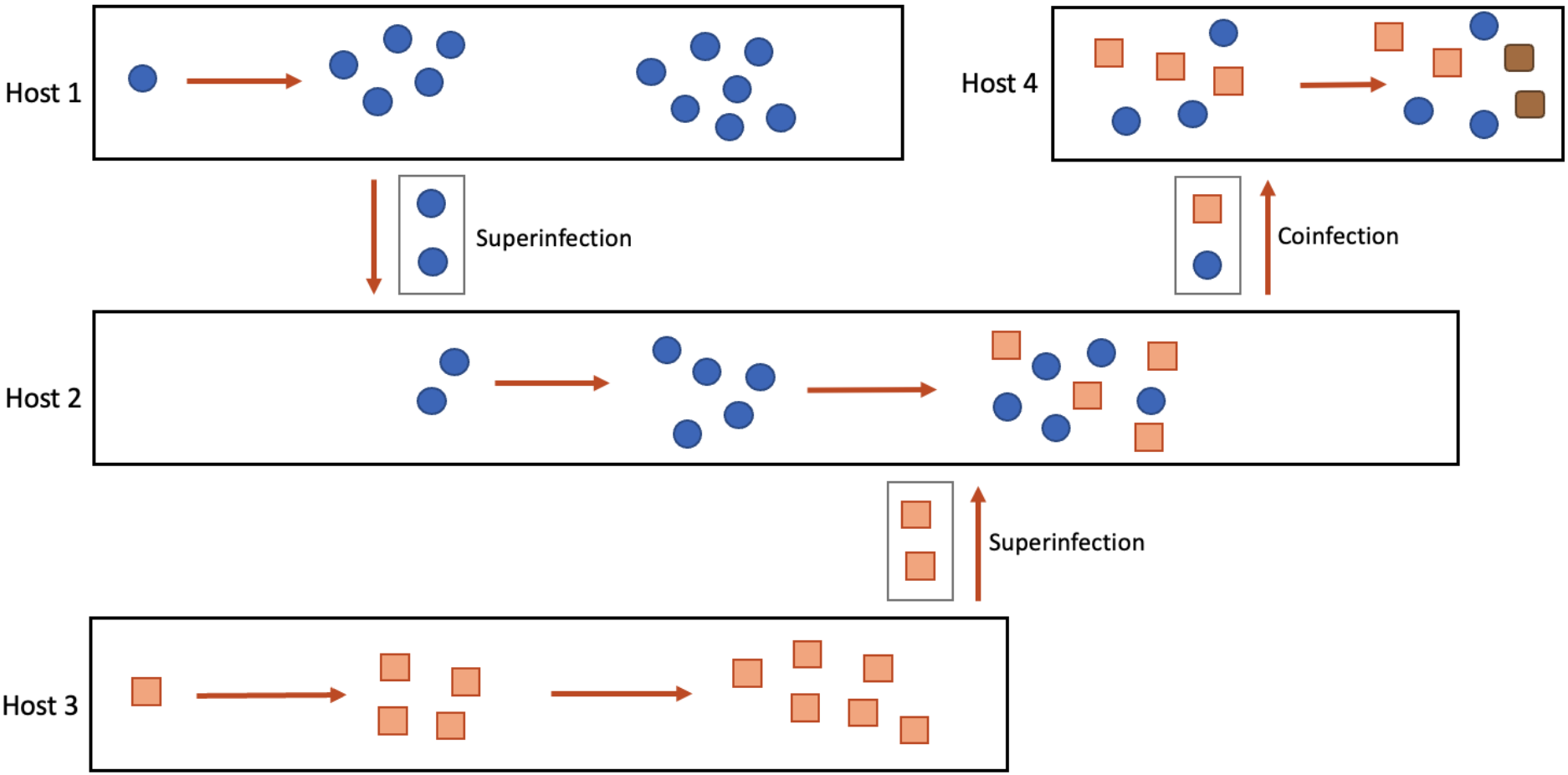
- Data Augmentation Markov Chain Monte Carlo
- Challenge: Switching coinfected vs. superinfected status for individuals in proposal q distribution
- High-dimensional model with changing dimensions of transmission source if status switches
- Developing and implementing Reversible-Jump MCMC algorithm to switch models (Green, 1995)

Preliminary Results



- Simulation and complete-data likelihood are implemented correctly and recapture scalar parameters with known transmission network
- Current work: Implementing data augmentation portion of MCMC (estimating exposure time and transmission network with co- and superinfection)

Further Work



Acknowledgements

- Committee
  - Max Lau
  - Lance Waller
  - Steve Qin
  - Katia Koelle (PBEE)
- Swine MP3 Group
  - Anice Lowen (Microbiology and Immunology)
  - Katia Koelle (PBEE)
  - Dave VanInsberghe (PBEE) (Genetic data and dataset graphics)
- Molecules and Pathogens to Populations and Pandemics (MP3) Initiative (Funding)
- The Ohio State College of Veterinary Medicine, Animal Influenza Ecology and Epidemiology Research Program (Dataset)

References I

Didelot, X., Gardy, J., and Colijn, C. (2014). Bayesian Inference of Infectious Disease Transmission from Whole-Genome Sequence Data. *Molecular Biology and Evolution*, 31(7):1869–1879.

Firestone, S. M., Hayama, Y., Bradhurst, R., Yamamoto, T., Tsutsui, T., and Stevenson, M. A. (2019). Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Scientific Reports*, 9(1):4809.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Grenfell, B., Pybus, O., Gog, J., Wood, J., Daly, J., Mumford, J., and Holmes, E. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327–332.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C., and Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):1–32.

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. (2015). A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLOS Computational Biology*, 11(11):e1004633. Publisher: Public Library of Science.

Ypma, R. J. F., van Ballegooijen, W. M., and Wallinga, J. (2013). Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics*, 195(3):1055–1062.