Heterogeneity in *M. tuberculosis* transmission in the United States.

Sourya Shrestha
sourya@jhu.edu
Dept of Epidemiology,
Johns Hopkins School of Public Health

27th Annual Conference, Union-NAR, Vancouver
Feb 25, 2023

Collaboration:

US CDC/DTBE:

- **Andrew Hill, Suzanne Marks**
- **Kathryn Winglee, Steve Kammerer, Ben Silk**





Yale University:

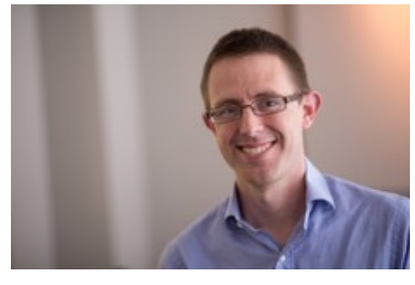
- **Jonathan Smith**

California Dept of Public Health:

- **Tambi Shaw**

Johns Hopkins:

- **David Dowdy**



Funding:

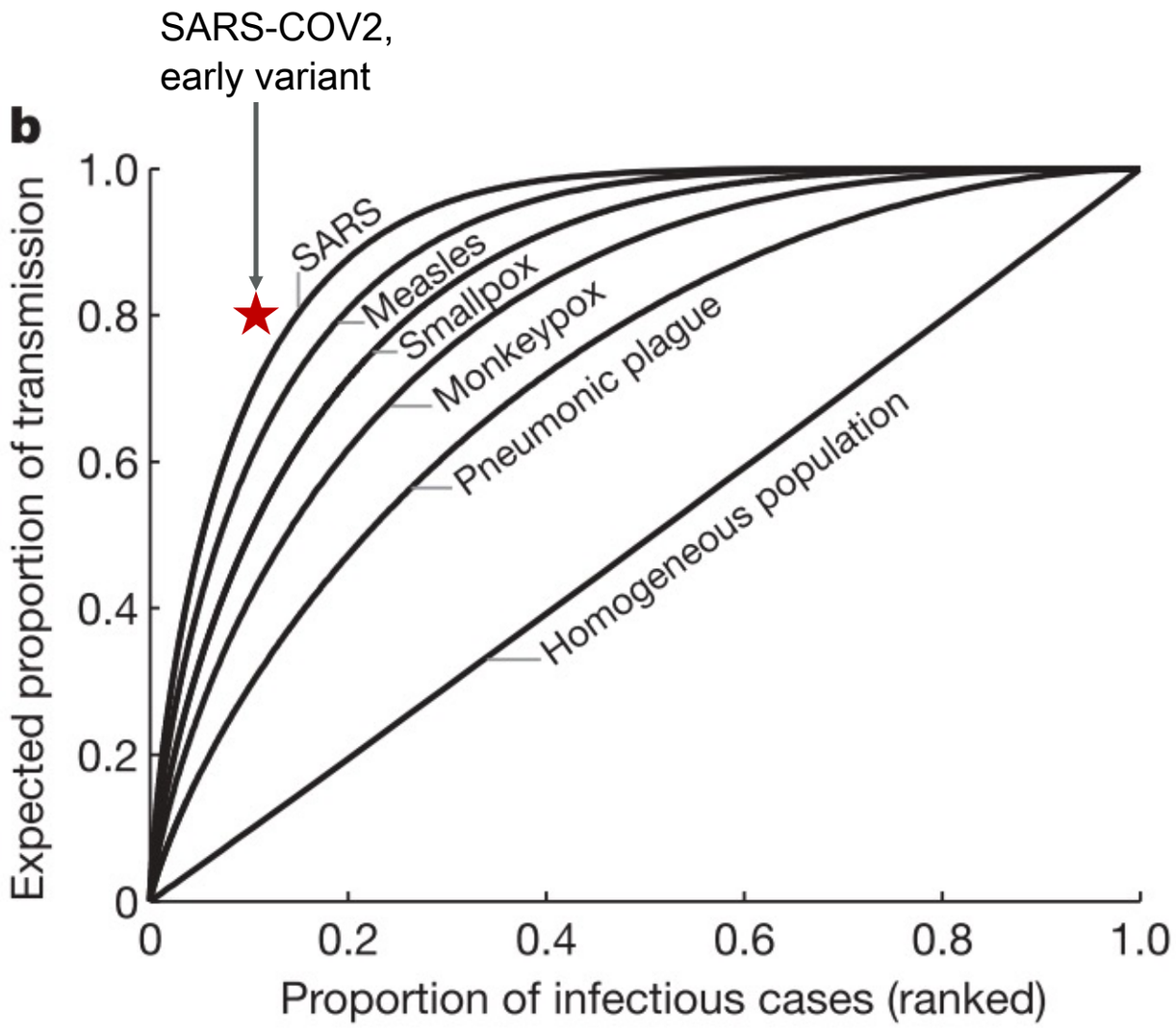
This project was funded by the CDC National Center for HIV, Viral Hepatitis, STD, and TB Prevention Epidemiologic and Economic Modeling Agreement (NEEMA 2.0).

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC, CDPH or other authors' affiliated institutions.

Background

Transmission of infectious pathogens is heterogenous.

- Small proportion of hosts contribute to large proportion of transmission.
- 20/80 rule, where 20% cases cause 80% of transmission, has been observed across many infectious diseases.

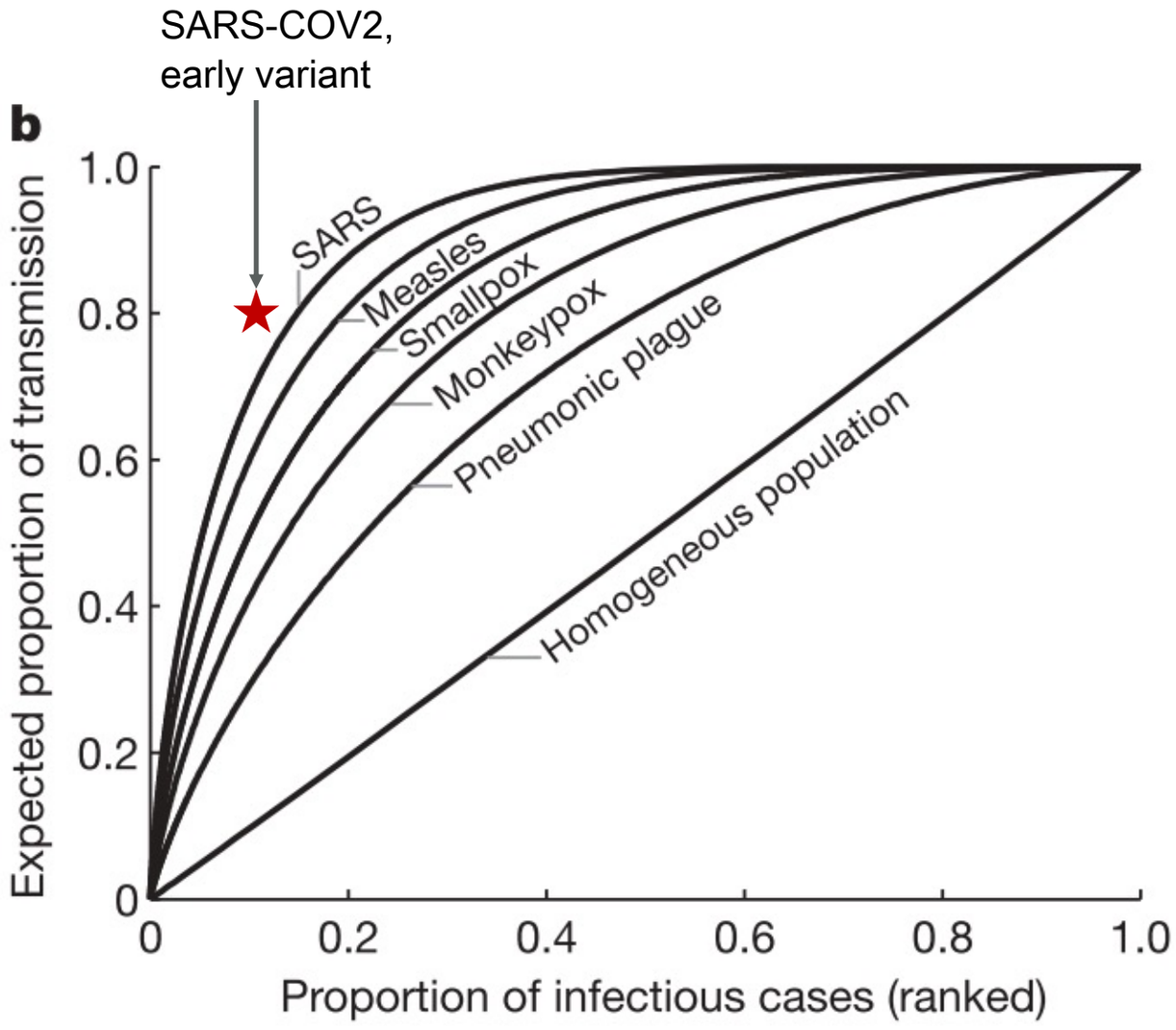


Woolhouse et. al., 1997; Lloyd-Smith et. al., 2005; Grantz et. al., 2020

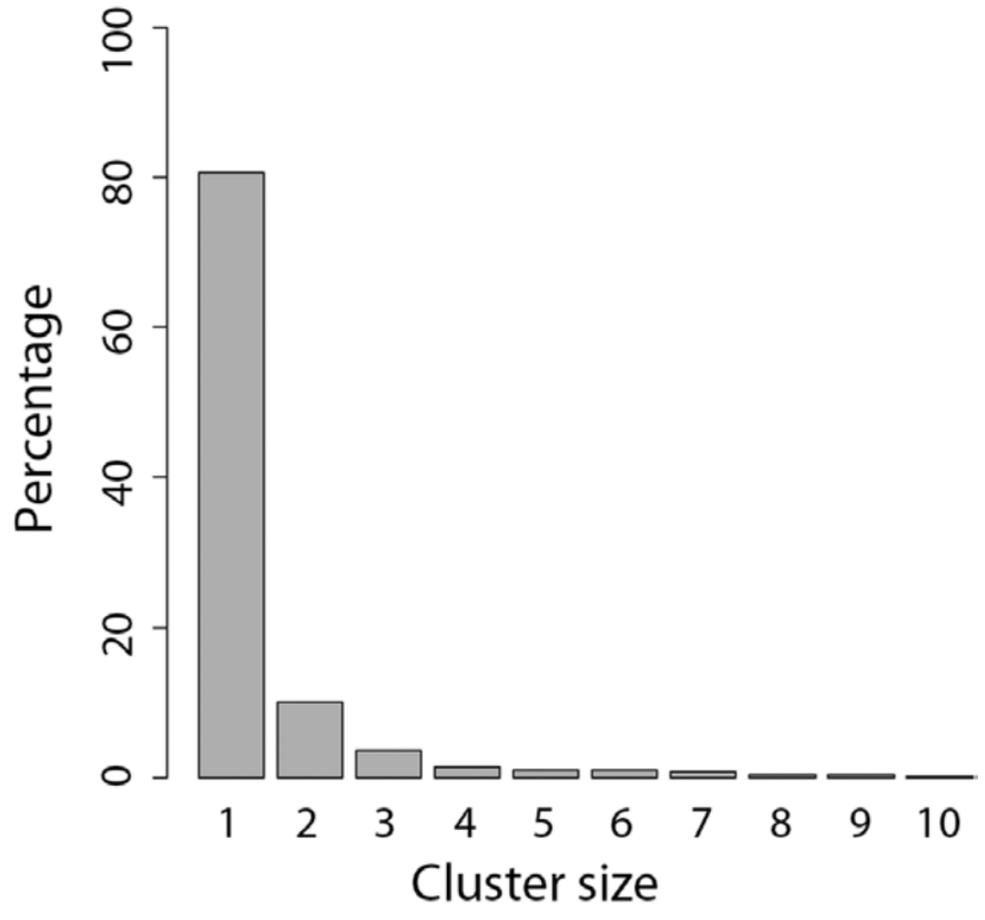
Background

Transmission of infectious pathogens is heterogenous.

- Small proportion of hosts contribute to large proportion of transmission.
- 20/80 rule, where 20% cases cause 80% of transmission, has been observed across many infectious diseases.
- Similar heterogeneity has been observed in tuberculosis transmission



Woolhouse et. al., 1997; Lloyd-Smith et. al., 2005; Grantz et. al., 2020



Based on TB cases diagnosed between 1993-2007 in the Netherlands (8,330 cases with RFLP)

Ypma et. al., 2013

Background

Understanding heterogeneity in transmission can:

- Help identify sources/settings where transmission risks are higher.
- Prioritize communities/settings/risk-factors, and potentially address disparities.
- Devote resources where most needed and help make TB-control most cost effective.

Background

Understanding heterogeneity in transmission can:

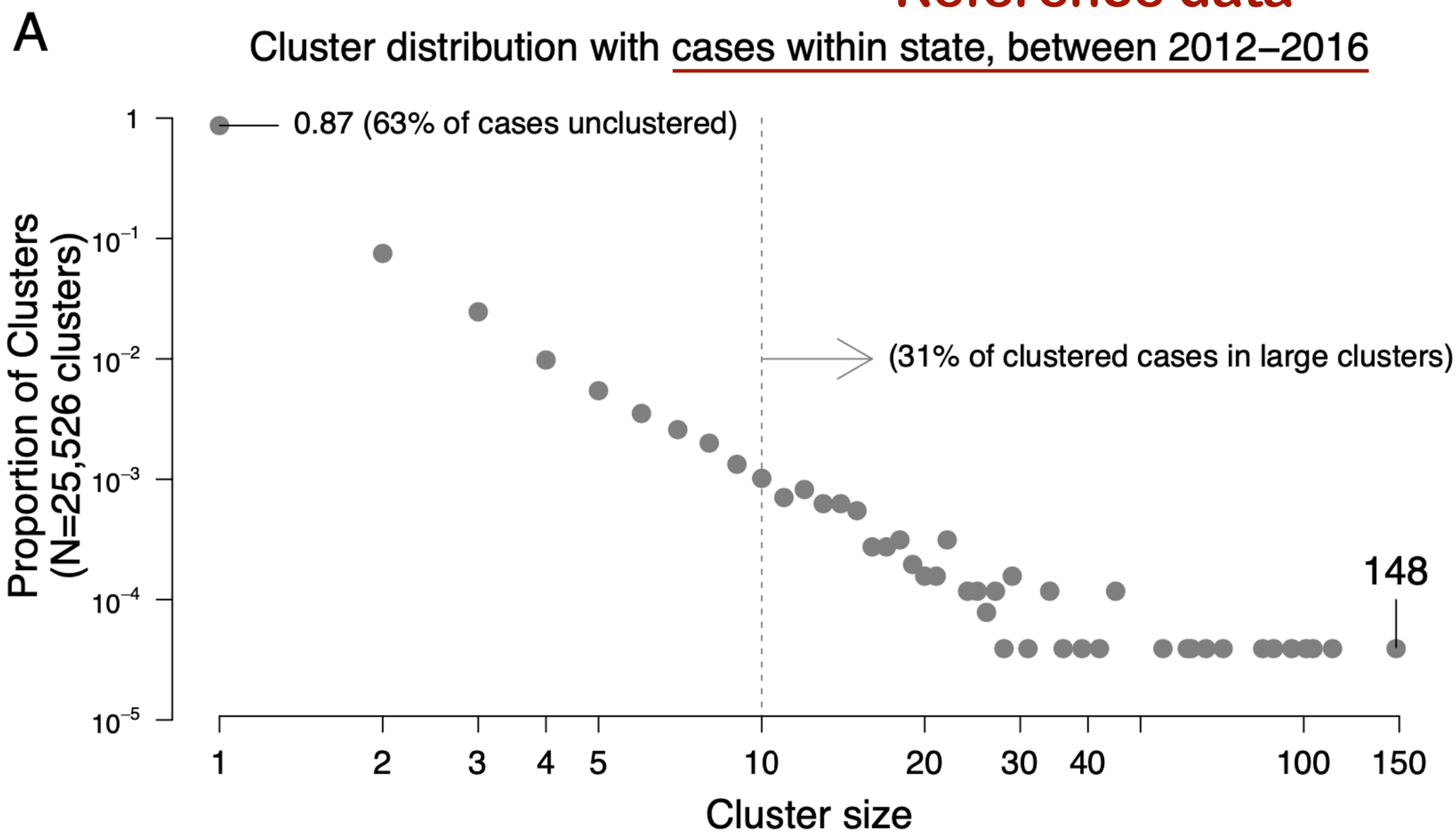
- Help identify sources/settings where transmission risks are higher.
- Prioritize communities/settings/risk-factors, and potentially address disparities.
- Devote resources where most needed and help make TB-control most cost effective.

We analyze transmission data from the United States (and key states, CA, FL, NY, and TX)

- Develop and fit mechanistic transmission models (branching process) to transmission clusters in the US.
- Estimate transmission parameters (e.g., R_0) and heterogeneity.
- Compare key states.
- Explore factors that affect these estimates.

Genotype cluster size distribution of TB cases in the US

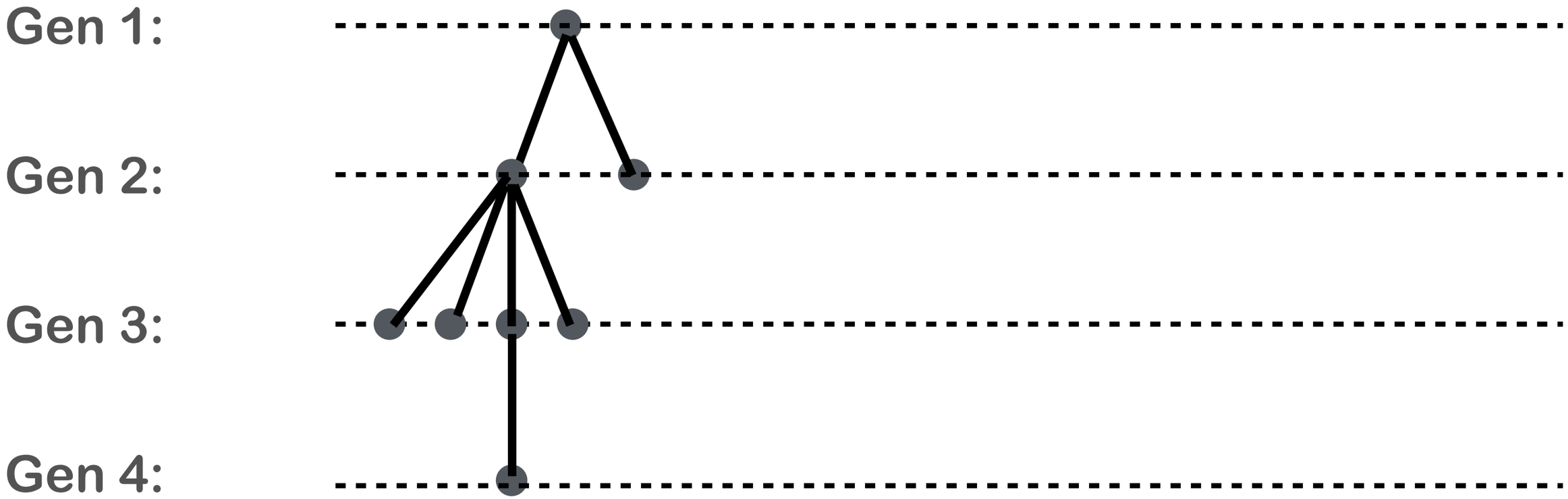
Reference data



Transmission links based on genotyping, i.e., matching isolates on the basis of spacer oligonucleotide typing (spoligotype) and 24-locus mycobacterial interspersed repetitive unit-variable number of tandem repeats (MIRU-VNTR)

Branching Process Models

Branching process models capture transmission dynamics

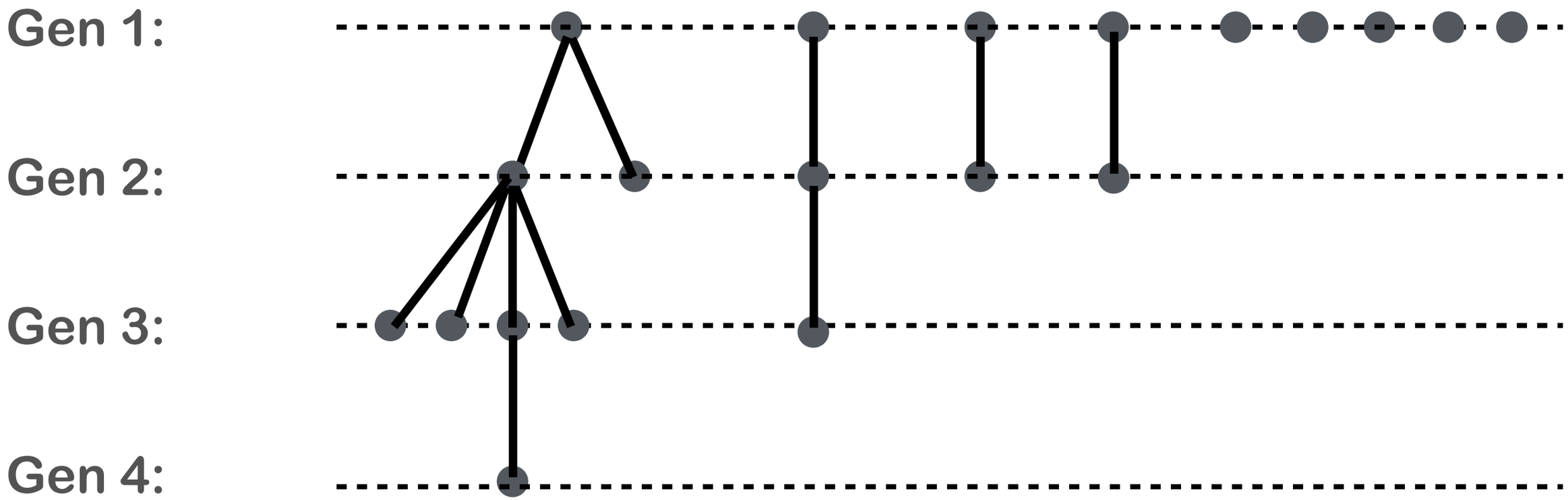


Compared to other kinds of transmission models (e.g., compartmental, individual based) branching process models:

- Focus on capturing transmission chains through several generations

Branching Process Models

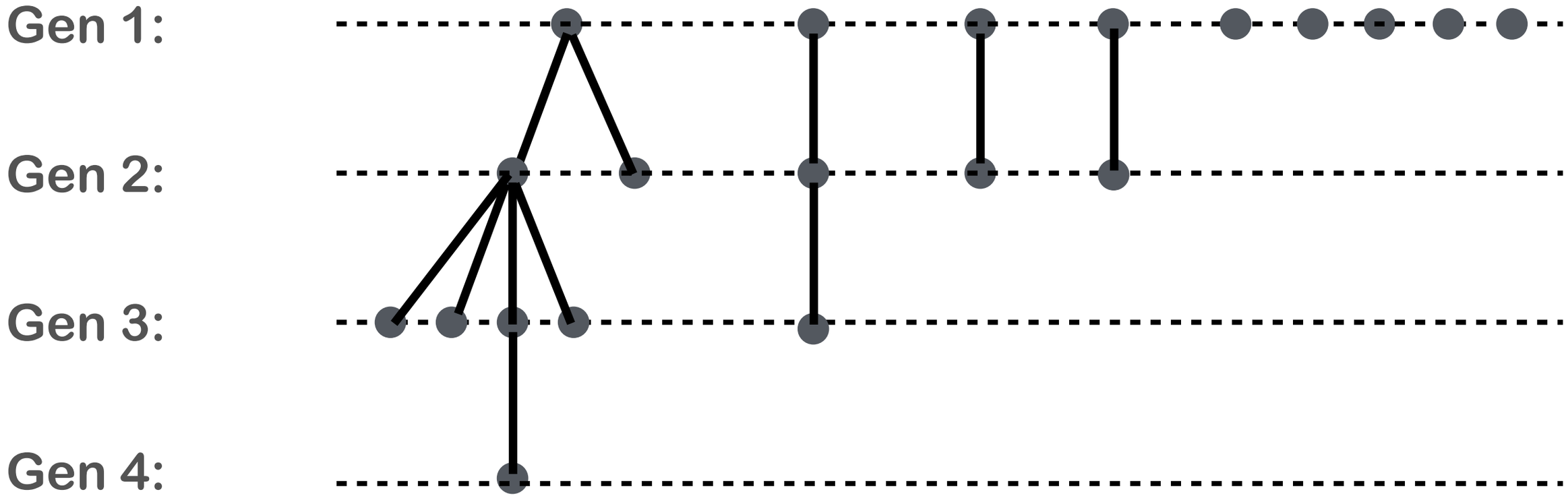
Branching process models capture transmission dynamics



Compared to other kinds of transmission models (e.g., compartmental, individual based) branching process models:

- Focus on capturing transmission chains through several generations
- Allow incorporating of heterogeneity at the individual level
- Have been used in the context of transmission of a range of infectious diseases including TB (Farrington et al, 2003; Lloyd-Smith et al, 2005; Ypma et al, 2013)

Branching Process Models



Incorporate individual-level heterogeneity.

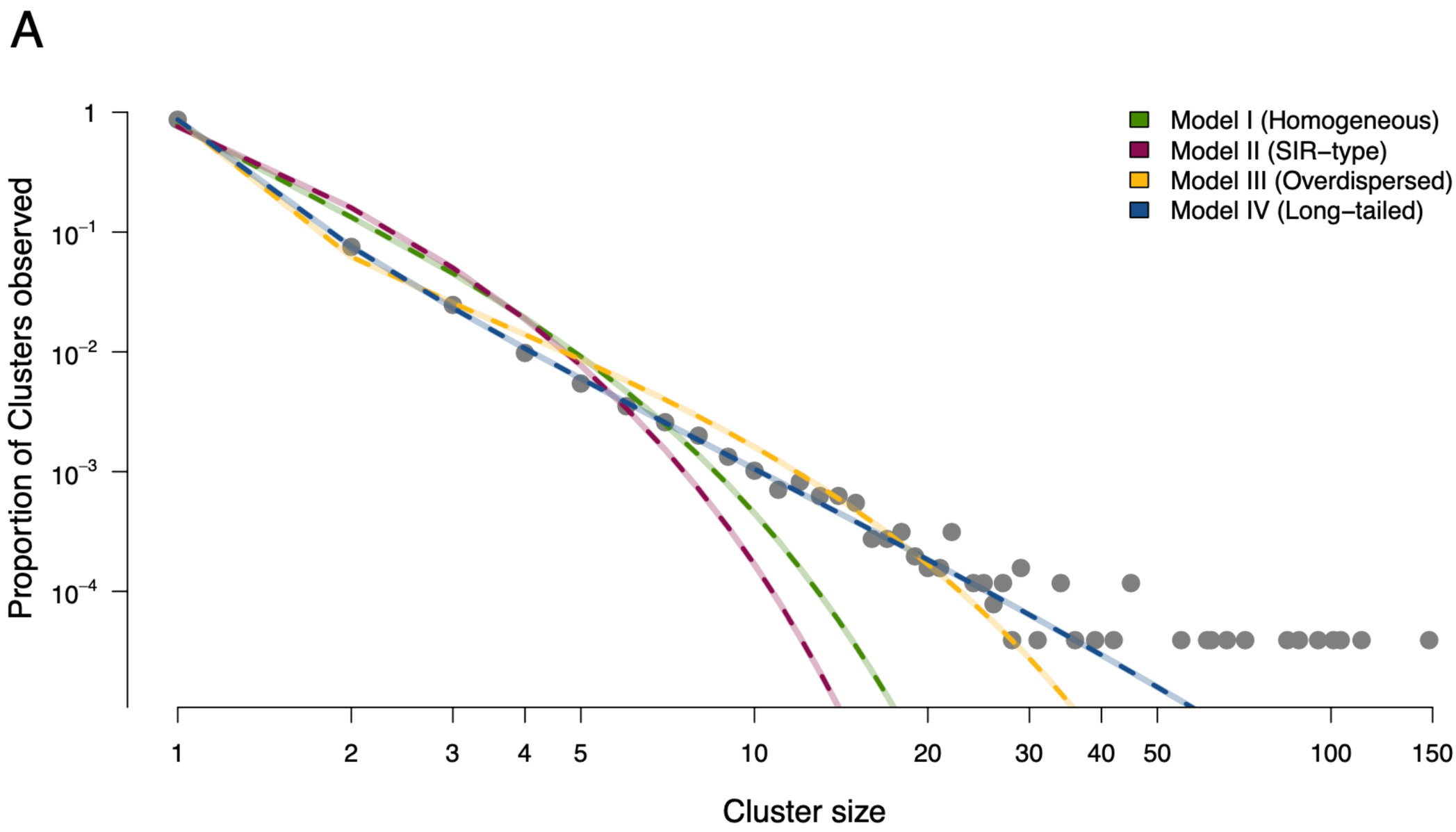
- Model 1: Homogenous model with no individual level variation
- Model 2: SIR-type model (Standard compartmental transmission model)
- Model 3: Overdispersed model (Ypma et al, 2013)
- Model 4: Long-tailed model (Poisson lognormal)

Use likelihood-based framework to evaluate the fit of the models.

Model comparison

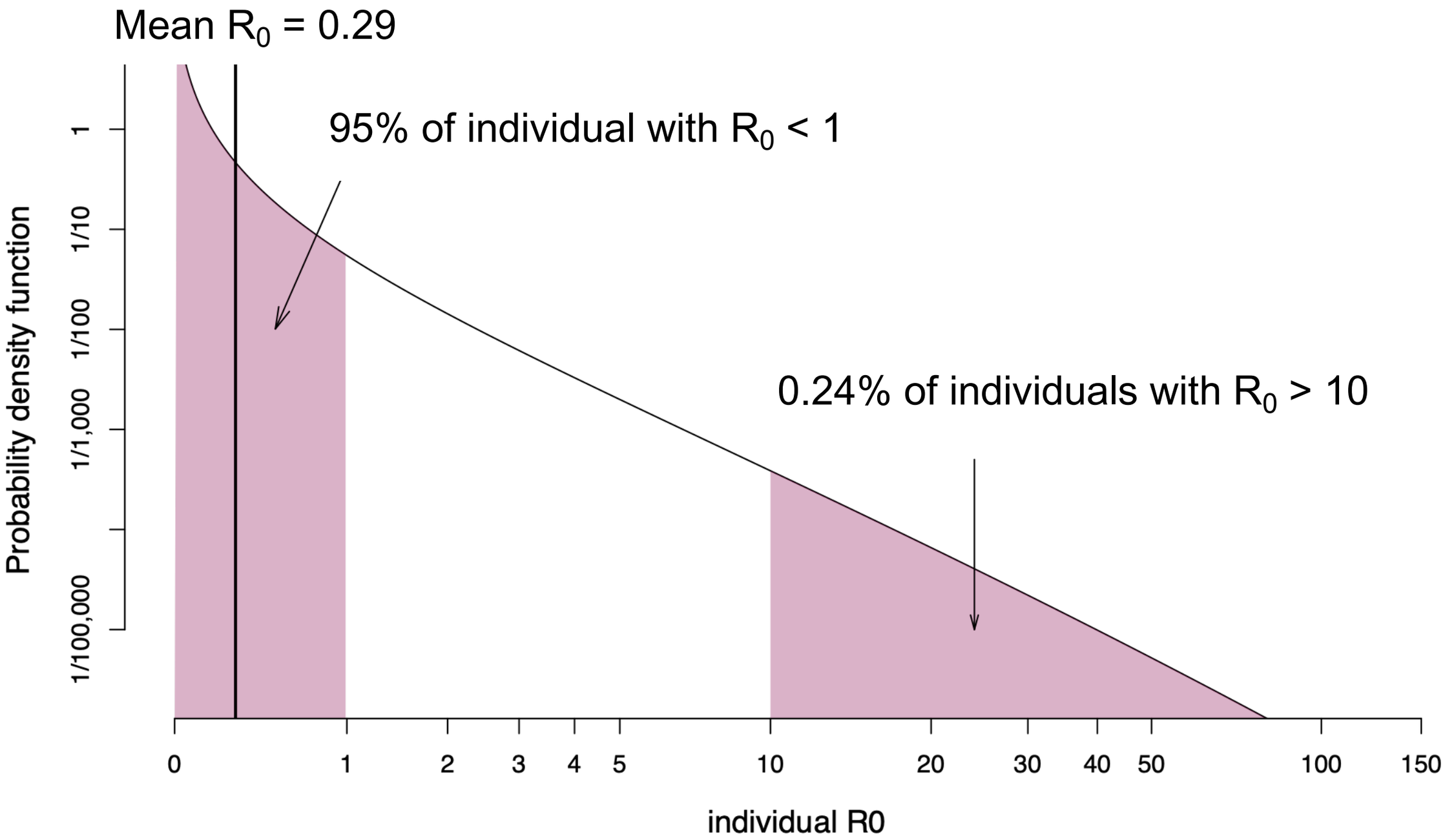
Models	Model description	Underlying distribution of individual reproductive number, ν; the resulting distribution of secondary cases, $\boldsymbol{Z}$; variance of $\boldsymbol{Z}$	Maximum likelihood estimate, MLE, log scaled (difference in log likelihood units relative to the highest estimate)	Relative likelihood compared to the best model **
Model I: Homogeneous model*	Assumes no individual-level heterogeneity, i.e., all individuals have the reproductive number.	ν is constant; $Z \sim \text{Poisson}(R_0)$; R_0	-16,787.68 (-1,450.19)	< 1/1000
Model II: SIR-type model*	Reflecting assumption in standard SIR-type compartmental models, assumes exponentially distributed individual reproductive numbers.	ν is exponentially distributed; $Z \sim \text{geometric}(R_0)$; $R_0(1 + R_0)$	-17,804.98 (-2,468.19)	< 1/1000
Model III: Overdispersed model	Assumes that the number of secondary cases from an individual are over dispersed, and the degree of overdispersion is estimated.	ν is gamma distributed; $Z \sim \text{negative binomial}(R_0, k)$ k is the dispersion parameter, smaller values relate to larger heterogeneity; $R_0(1 + \frac{R_0}{k})$	-15,507.78 (-170.99)	< 1/1000
Model IV: Long-tailed model	Assumes that individual-level heterogeneity is lognormally distributed (allowing for even larger heterogeneity).	ν is lognormally distributed; $Z \sim \text{Poisson lognormal}(\mu, \sigma^2)$ μ, σ^2 are, respectively, mean variance of the underlying normal distribution; $R_0\,[1 + R_0\,(exp(\sigma^2) - 1)]$	-15,336.79 (Ref)	–

Fitting branching process models to cluster distributions.



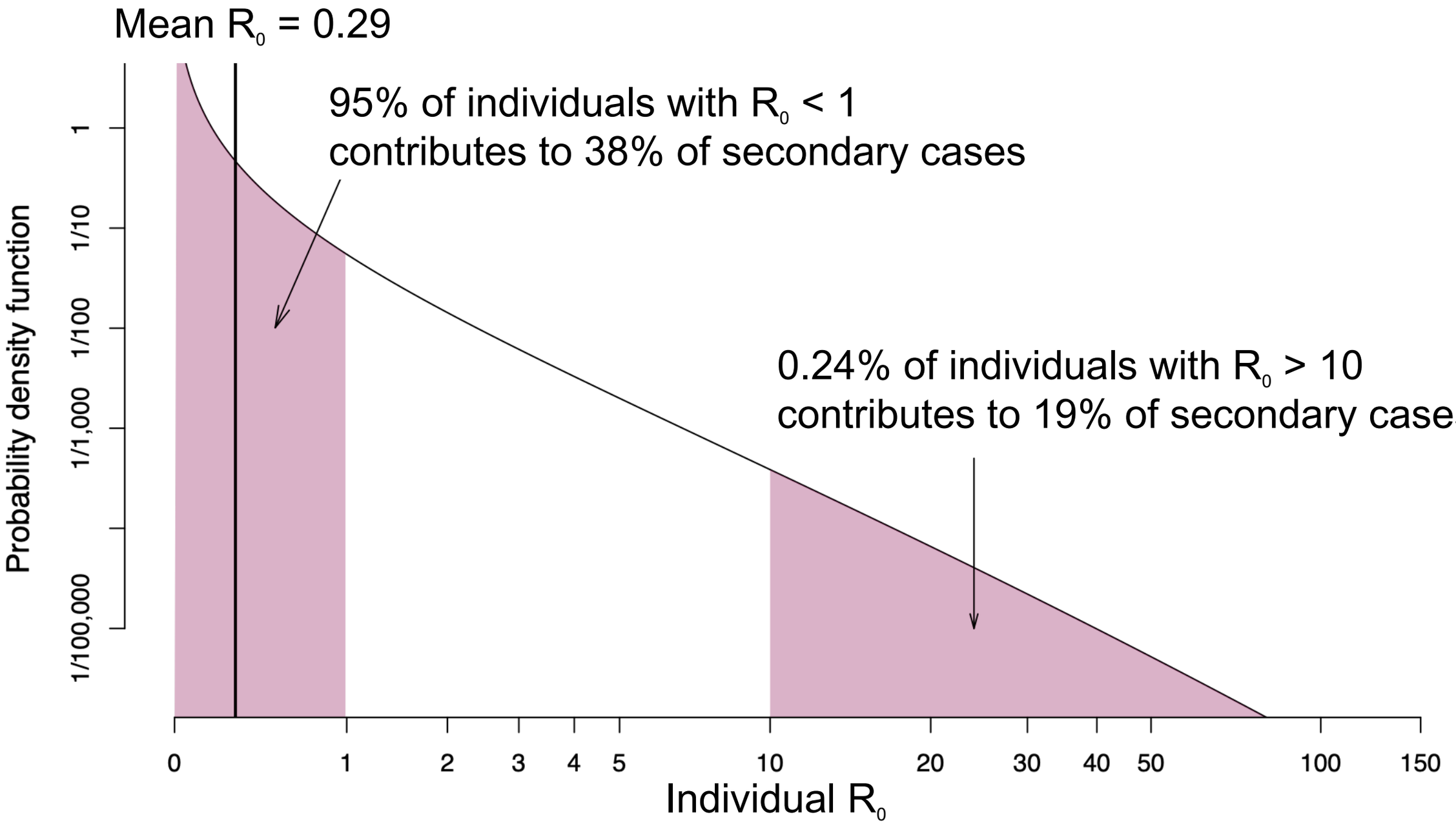
- SIR-type, homogeneous, and over-dispersed model fail to capture the “long tail” in the cluster distribution
- Long-tailed model captures the frequency of large clusters, and is statistically a better fit

Underlying individual-level heterogeneity



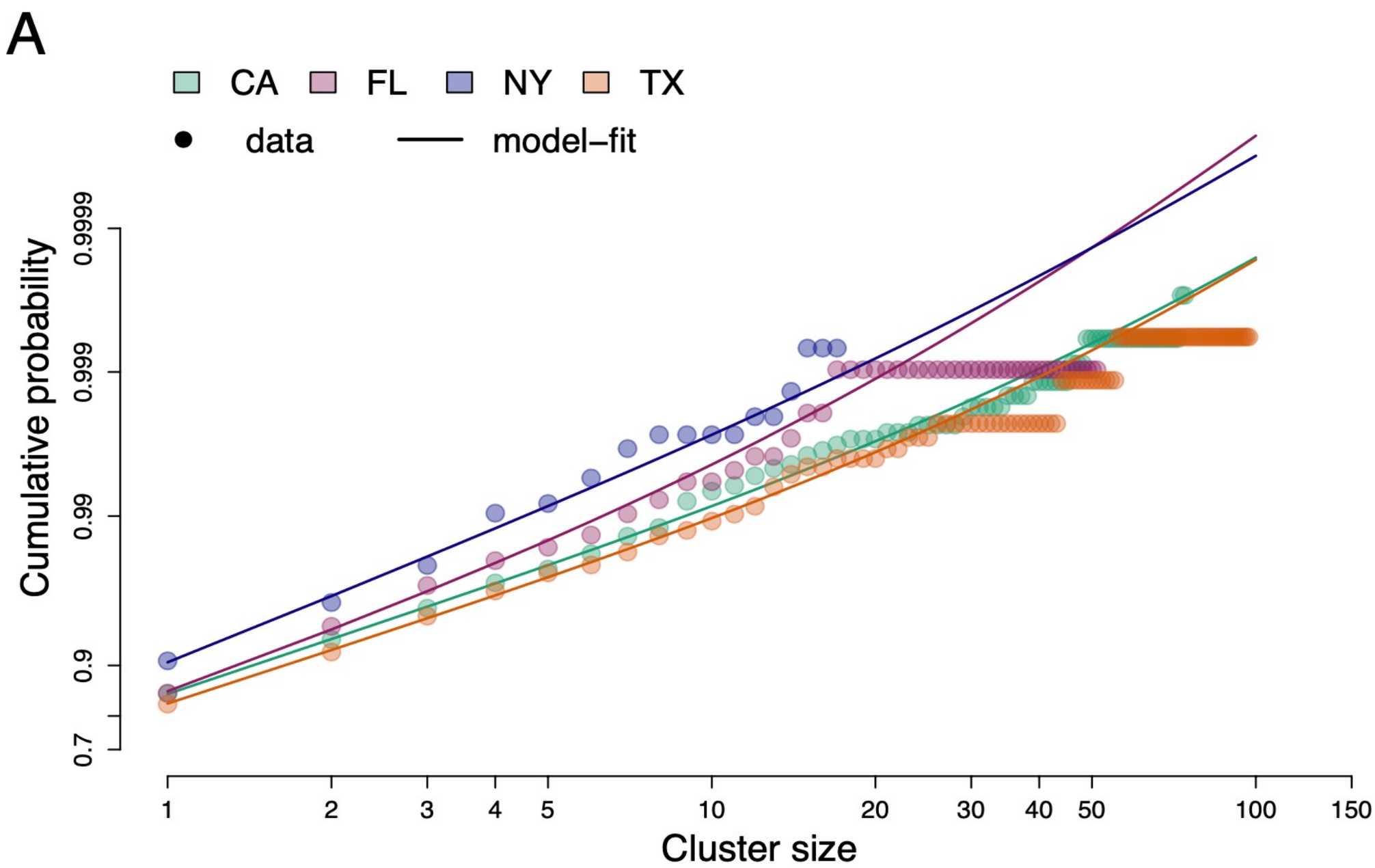
- Underlying individual-level R_0 distribution, corresponding to best fit Long-tailed model

Underlying individual-level heterogeneity



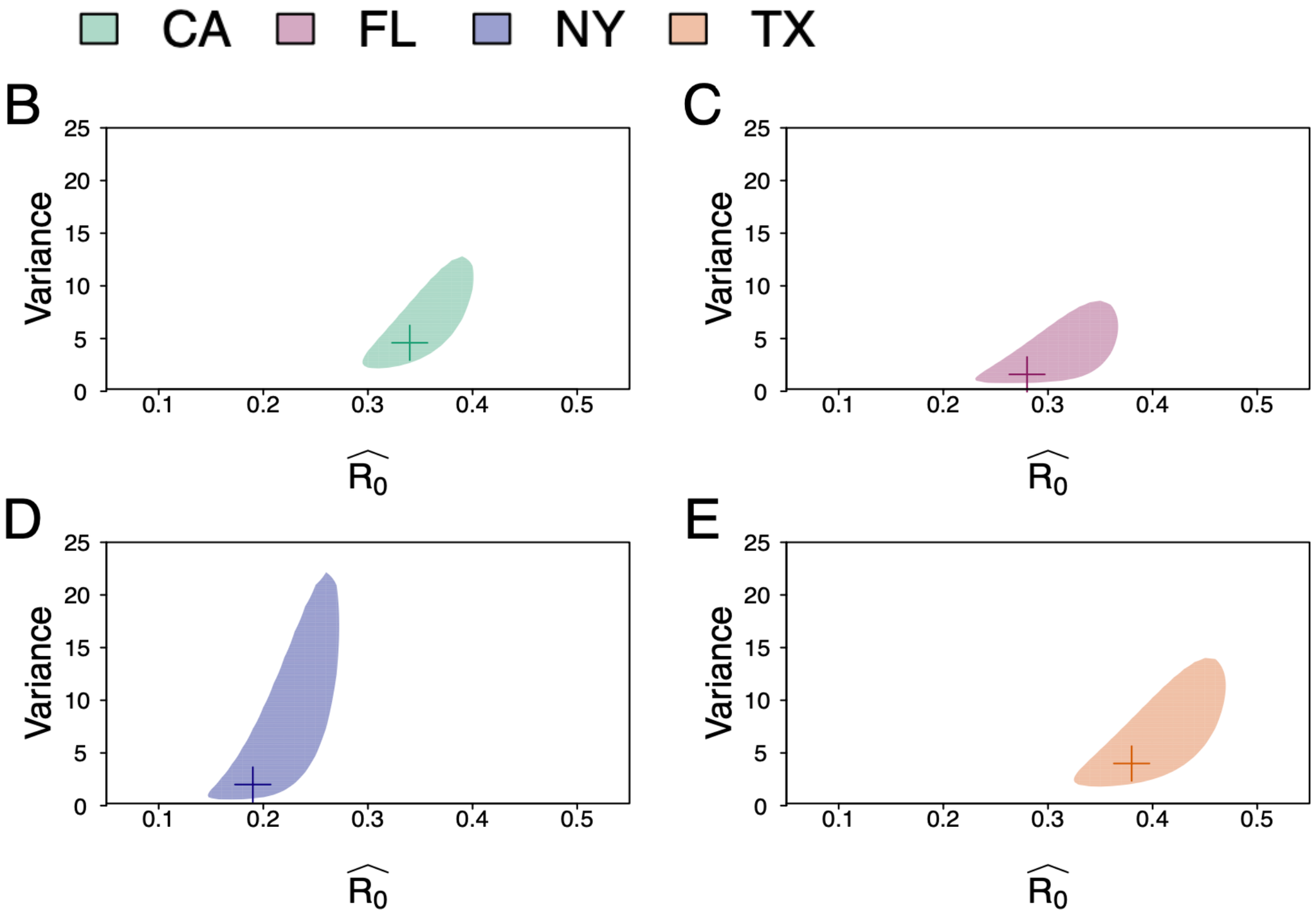
- Underlying individual-level R_0 distribution, corresponding to best fit long-tailed model shows:
  - Low transmission rate: mean R_0 = 0.29
  - Incredible heterogeneity: 95% of individuals have R_0 < 1 and contribute to only 38% of secondary cases, but very few individuals with high R_0 (contribute substantially).

State-level differences across CA, FL, NY and TX



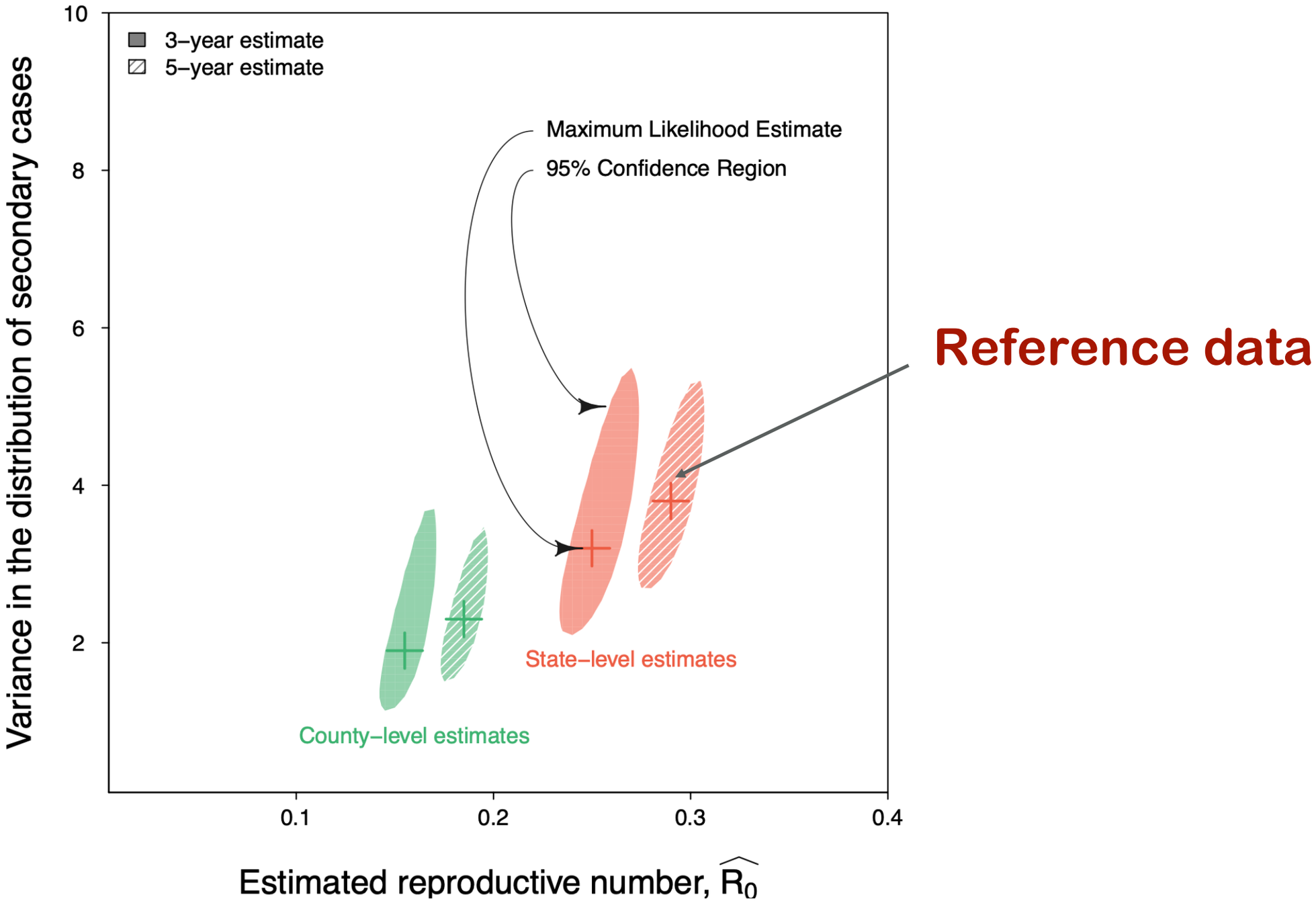
- Fit to state-level data in four states (2014-2016)
- Long-tailed models are better fits to the data.

State-level differences across CA, FL, NY and TX



- Substantial variation in estimated R_0 and heterogeneity across states.

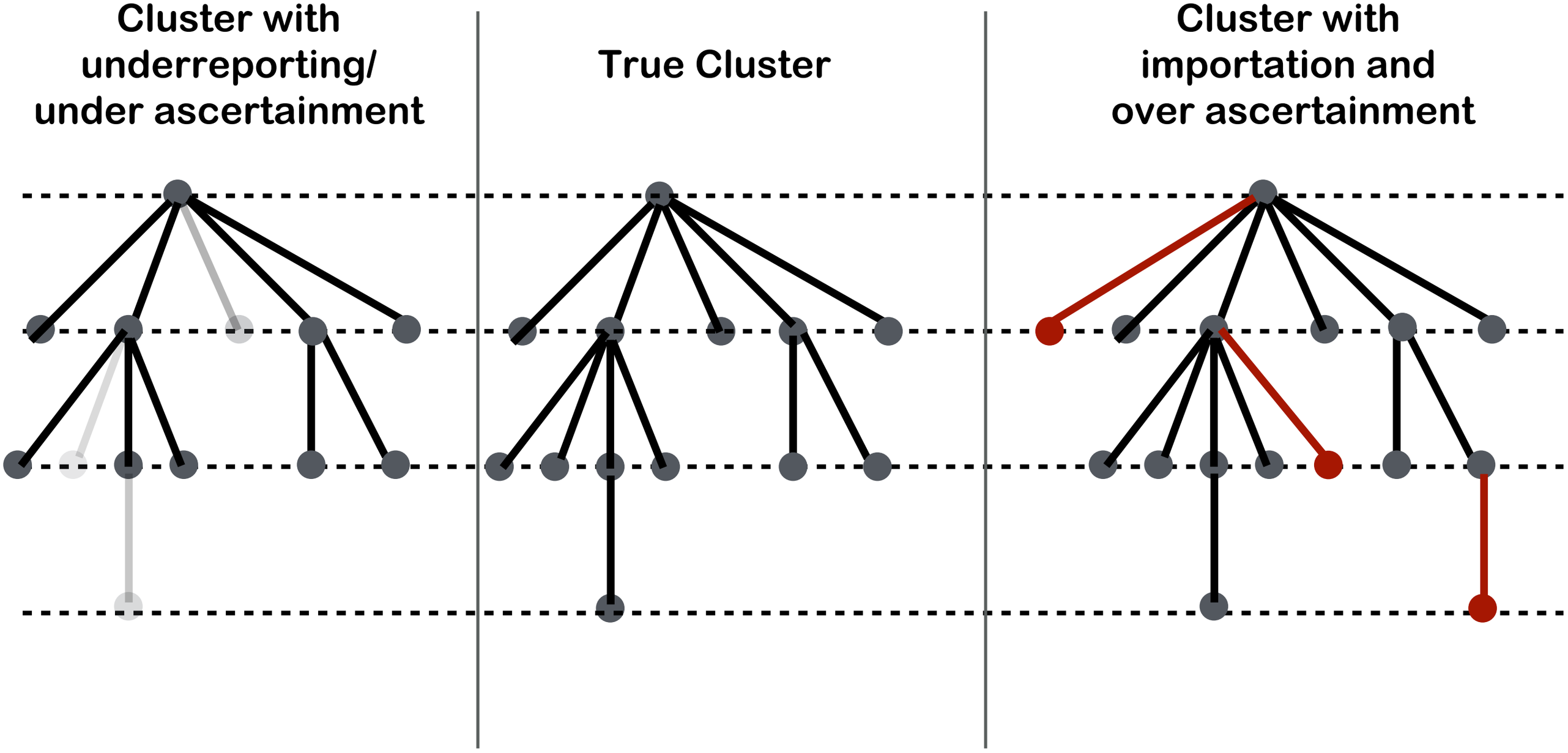
Comparing inferences under different cluster definitions



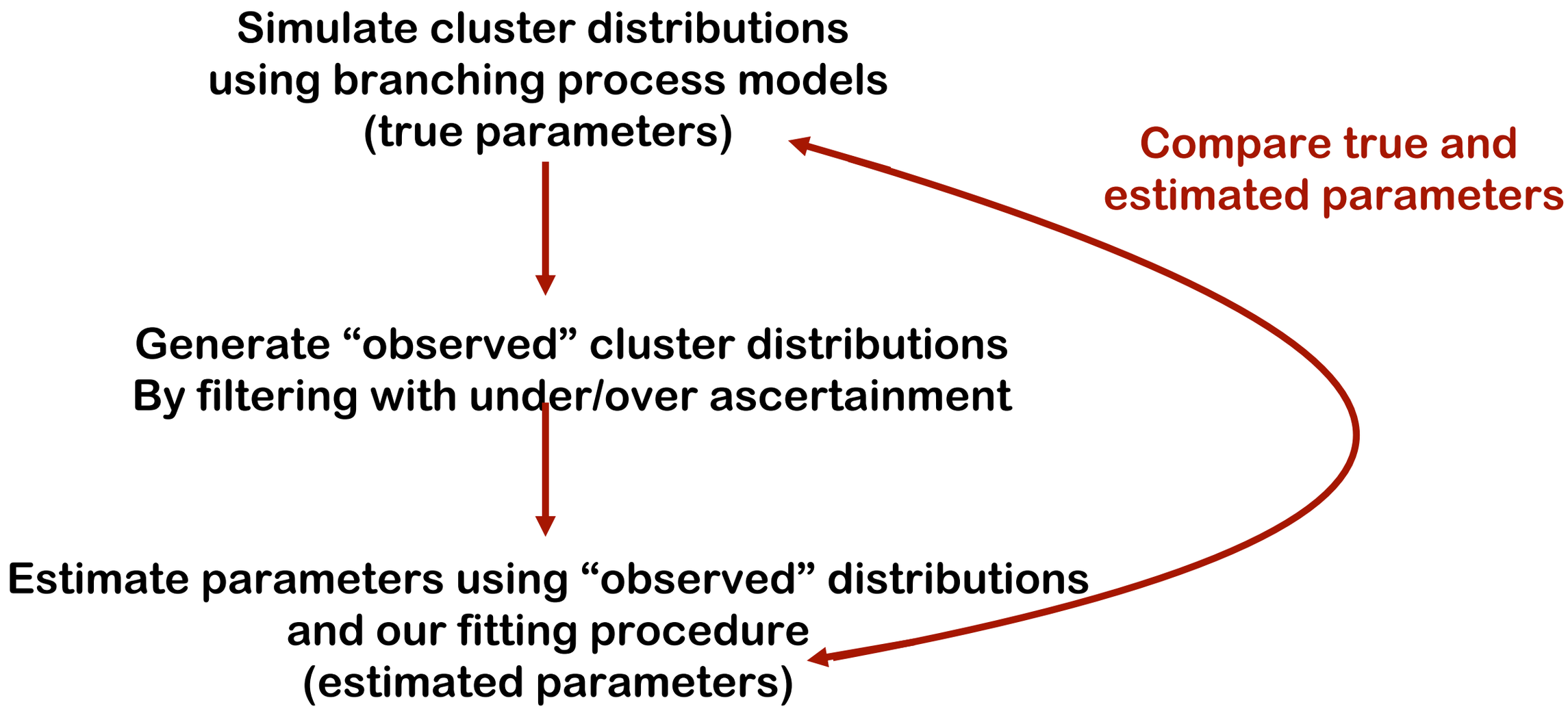
Inference of R_0 are sensitive to cluster definition.

- More sensitive to using a cluster definition consisting of state vs. county, compared to using 3-year vs. 5-year.

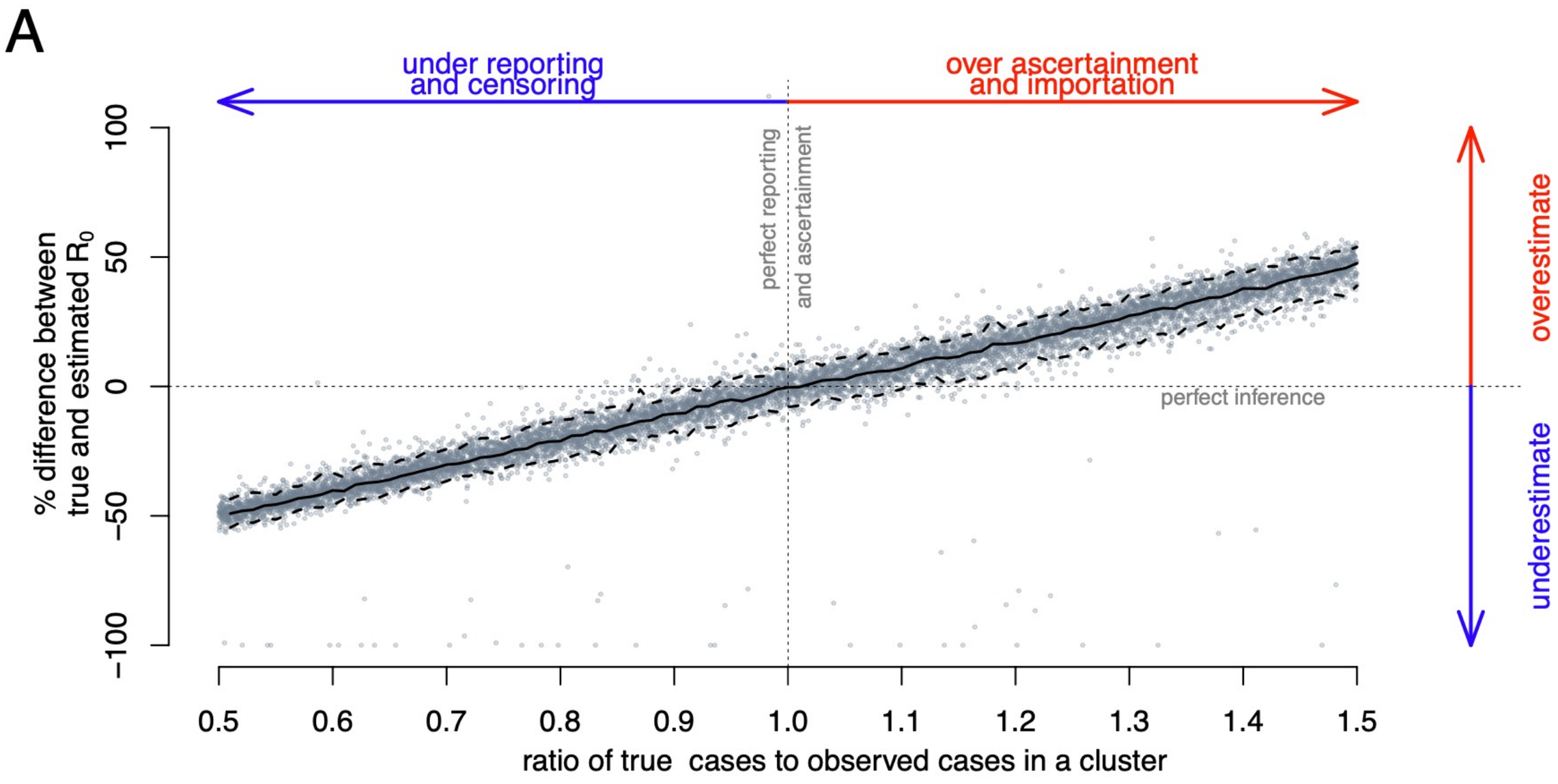
Model for under and over ascertainment



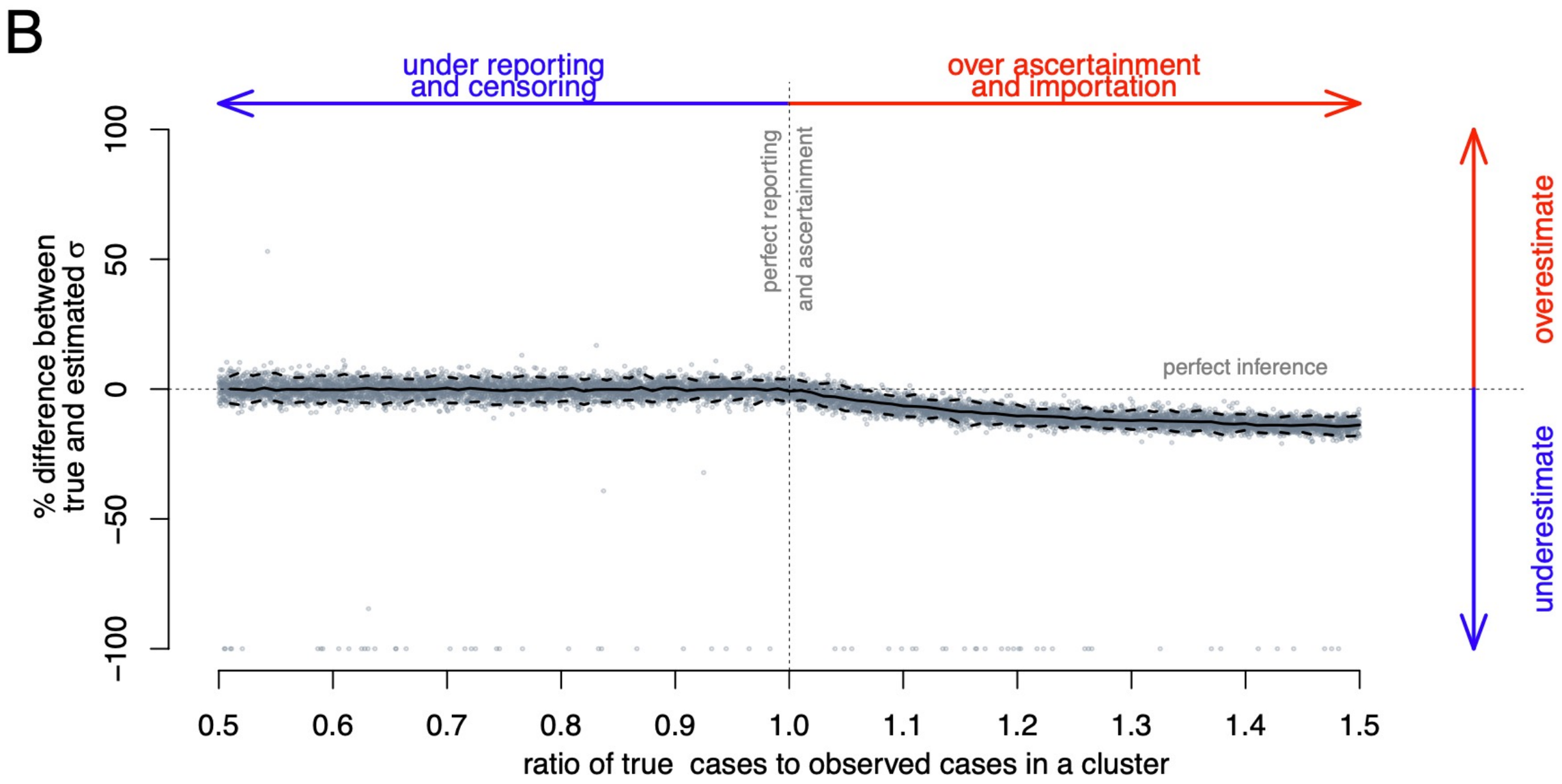
Simulation Study for Sensitivity of Model-based inference



Sensitivity of model inference



Sensitivity of model inference



- Estimates of R_0 sensitive to ascertainment, but in a predictable manner.
- Estimated heterogeneity didn't appear to be very sensitive.

Summary

- Estimated transmission rates in the United States were low.
  - R_0 estimates similar to other low burden countries (UK: 0.41; Netherlands: 0.24; Brooks-Pollock et al, 2020)
- Transmission highly heterogeneous.
  - Degree of heterogeneity better captured by long-tailed distribution (Brooks-Pollock et al, 2020)
  - Most simulated cases (95%) had individual reproductive number < 1
  - Very few cases (0.24%) contributed to 19% of secondary cases of recent transmission
- Transmission varied across states.
  - R_0 estimates were twice as large in Texas compared to New York
- Definition of genotype cluster, and imperfection in cluster ascertainment affected estimates of R_0
  - More conservative definitions of cluster resulted in smaller estimates of R_0
  - The effect on heterogeneity estimates were generally smaller

Limitations and next steps

- Conventional genotyping can be prone to both under and over ascertainment
  - Underreporting/missing cases, lack of specimen culture, left/right censoring —> Under ascertainment
  - Transmission in past from an endemic strain, importation, detection of deeper ancestry —> False attribution or over ascertainment (~60% confirmed via WGS)
  - Can vary between Mtb strains (differences in diversity)

- State-level differences could be driven by other factors
  - Difference in circulating strains
  - Demography and size of state and counties

- Estimated individual-level heterogeneity are not entirely individual-specific
  - Societal, environmental, pathogen-specific, TB-program related factors can drive heterogeneity.
  - Understanding the drivers can help prioritize programs/interventions.

Thank you!

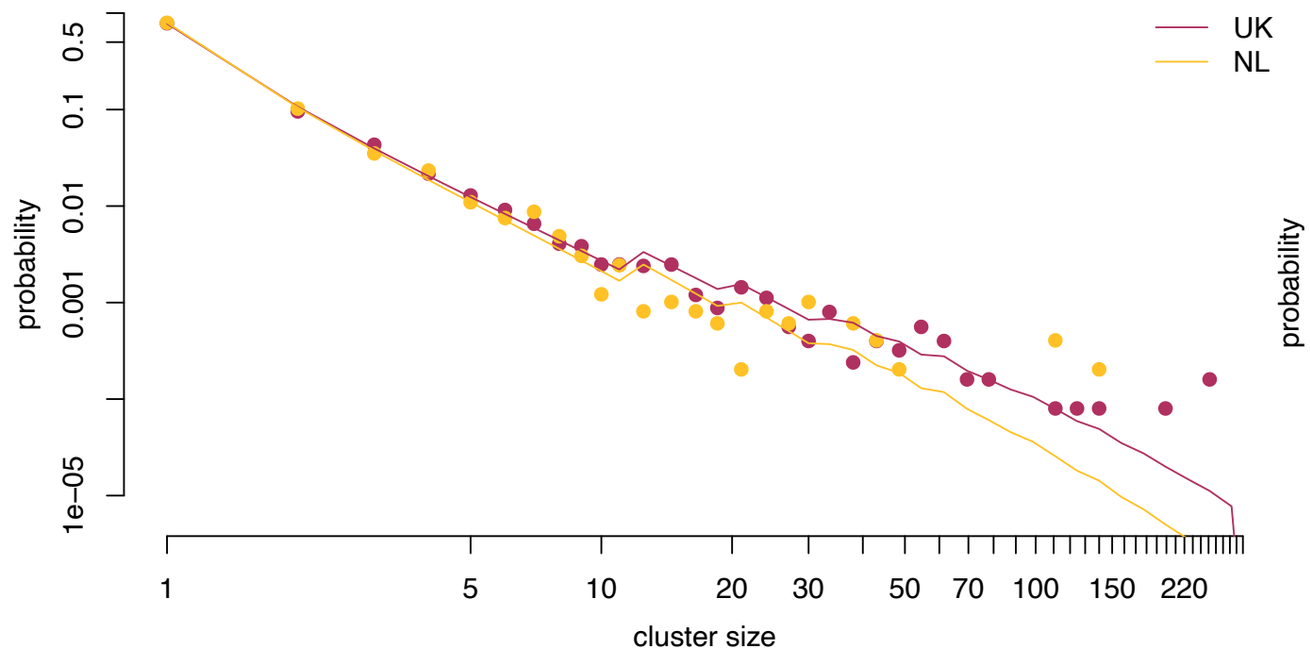
Clinical Infectious Diseases

MAJOR ARTICLE

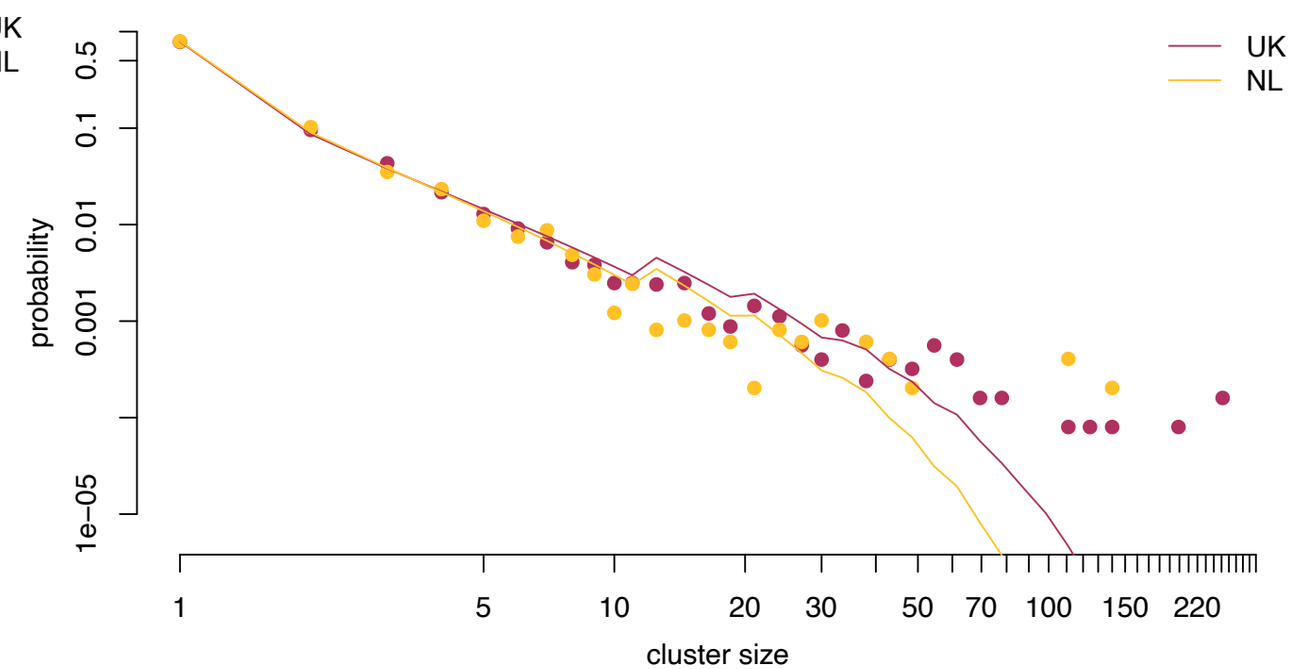
Model-based Analysis of Tuberculosis Genotype Clusters in the United States Reveals High Degree of Heterogeneity in Transmission and State-level Differences Across California, Florida, New York, and Texas

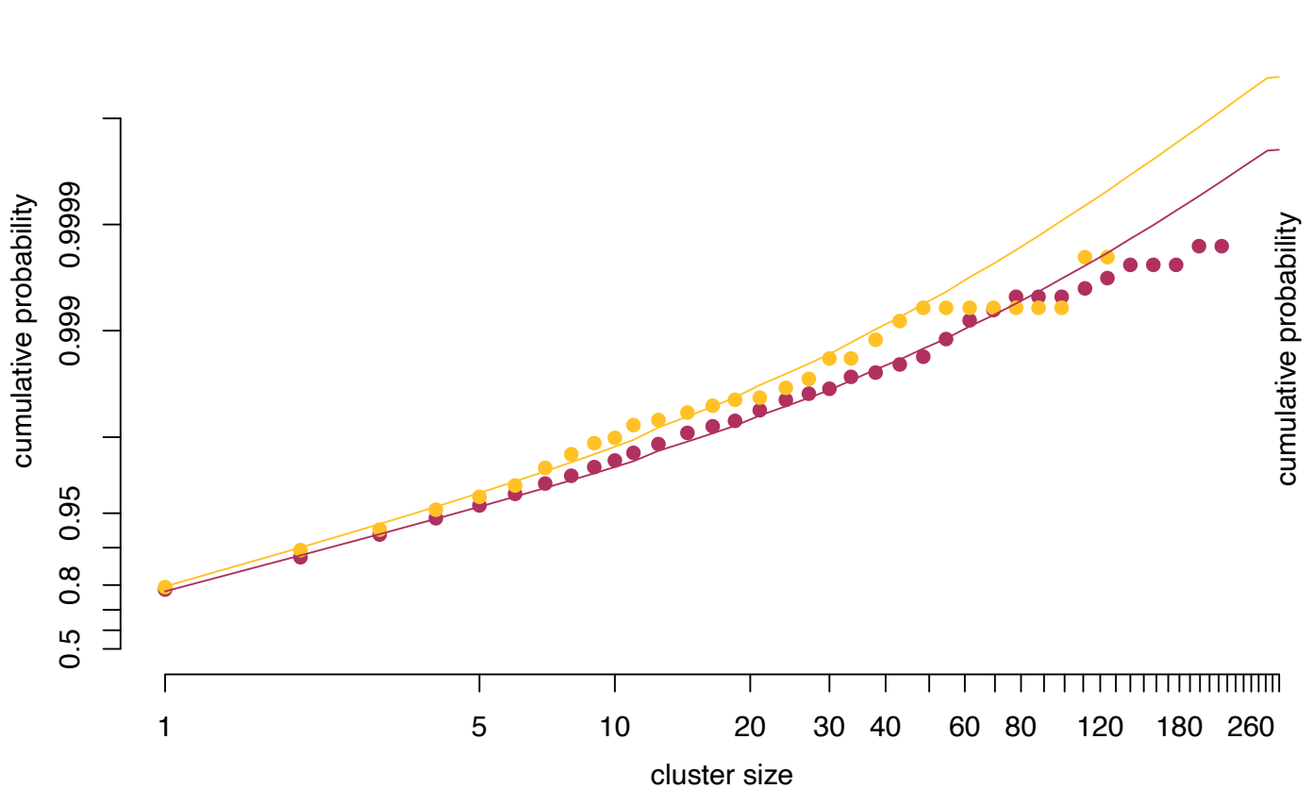
The Netherlands and the UK

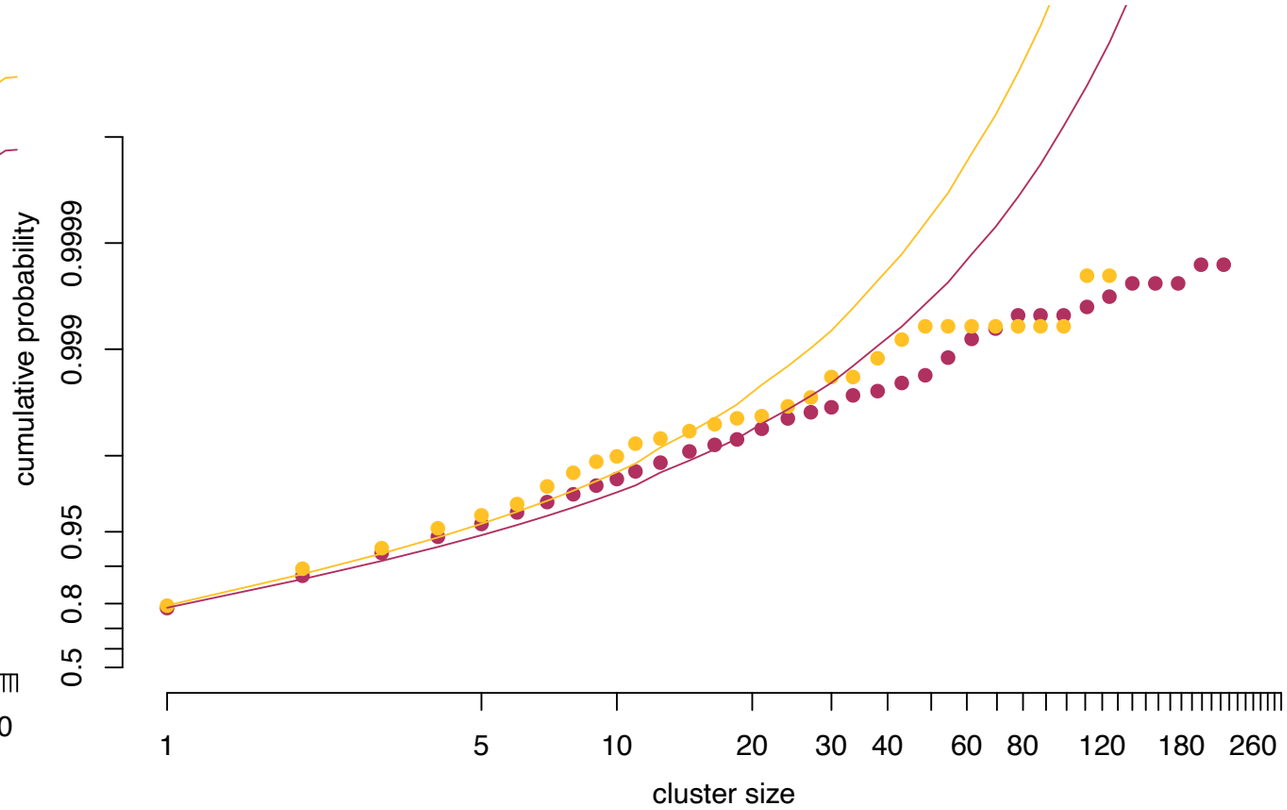
Poisson lognormal



Negative binomial







Data source: Brooks-Pollock et. al., 2020