Genomic data and the estimation of lineage- and cluster-specific epidemiological parameters.

Paul Tupper
with
Kurnia Susvitasari
Jessica Stockdale
Caroline Colijn

Simon Fraser University

Introduction

Aims

- Use genomic data to determine cluster-specific transmission dynamics
- Our example here: **Serial intervals**: relates to speed of transmission
- Other applications: transmission networks, other epi parameters including *R*

Team

- Postdocs Jessica Stockdale, Ben Sobkowiak: modelling, inference, genomic epi
- PhD students Kurnia Susvitasari, Nicola Mulberry (inference, modelling)
- Faculty Paul Tupper, Caroline Colijn
- Melbourne Team: Anders Gonçalves da Silva, Anne Watt, Courtney Lane

Questions about an infectious disease during a epidemic

How many people does each infectious person infect? (R_0, R)

How long after being infected can you infect others? (latent period)

When are you most infectious? When do you stop being infectious?

How long after being infected do you develop symptoms? (incubation period)

Where does transmission happen most?

Sources of Data

- Symptom onset times
- Test results and times
- Contact data (who were you with for how long)
- Clinical (symptoms and their severity, comorbidities)
- Location and demographic information

And an important new source:

Whole genome sequences of the pathogen.

Whole Genome Sequences

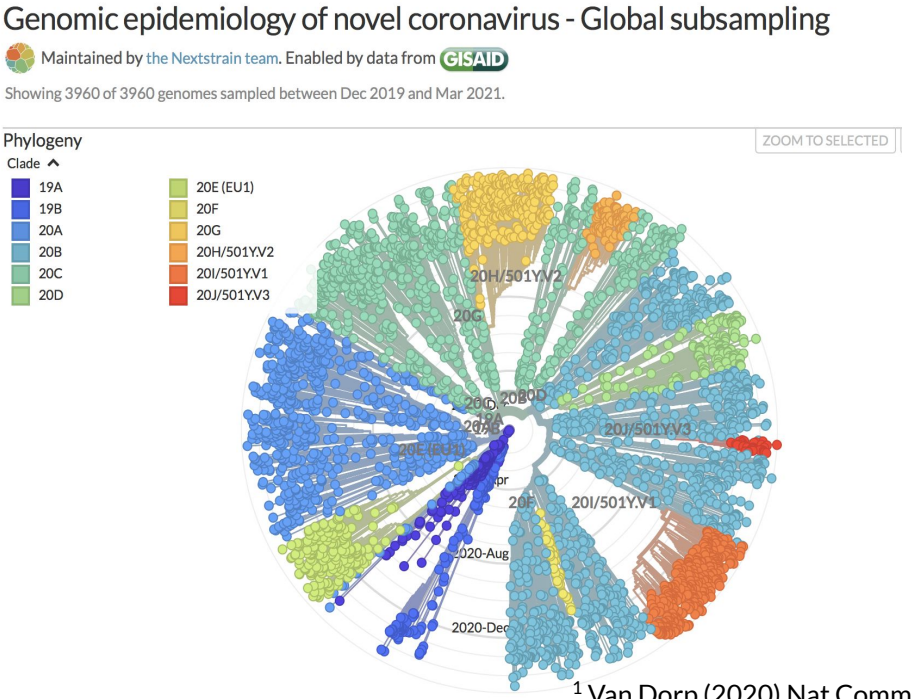
Sequence 1: A A C T A G G T A G

Sequence 2: A G T T A G G C G G

Suppose for sampled pathogens we have complete genetic material.

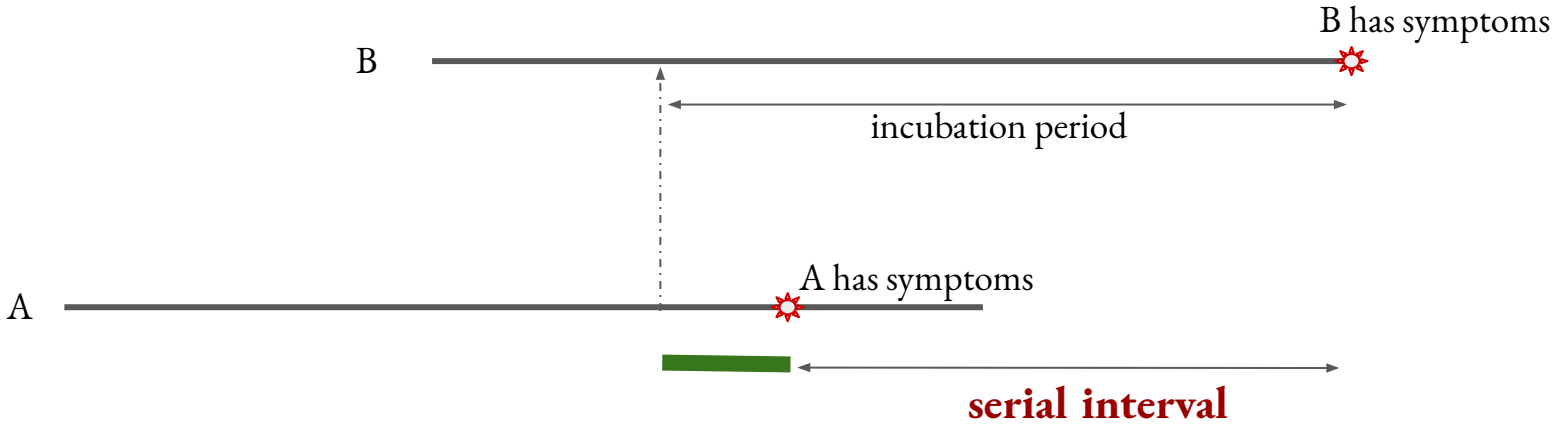
Can use this to estimate phylogenetic trees of all sampled pathogens.

Challenge: Use this information to make inferences about disease transmission, specific to lineage and cluster. Ideally fast!



[1] Van Dorp (2020) Nat Comm

Our motivating example: Serial interval distribution



- *Serial interval*: time between symptoms in A and symptoms in B (whom A infected)
- Can be negative, or ill-defined with asymptomatic transmission, but we neglect this.

Goal: Estimate the distribution of serial interval, specific to a particular lineage or cluster.

Why serial intervals?

- Serial intervals are a key feature of transmission dynamics.
- They are used in R estimates.
- They are used in estimating R_0, the basic reproduction number -and hence in estimating the portion of the population we need to vaccinate, and parameters for *all modelling efforts*.
- It is important to understand cluster- and variant-specific transmission dynamics
- BUT: **the world is still using serial interval estimates from Tianjin and Singapore because there is so little alternative data available. Serial interval depends on setting, stage of the epidemic, and variant.**

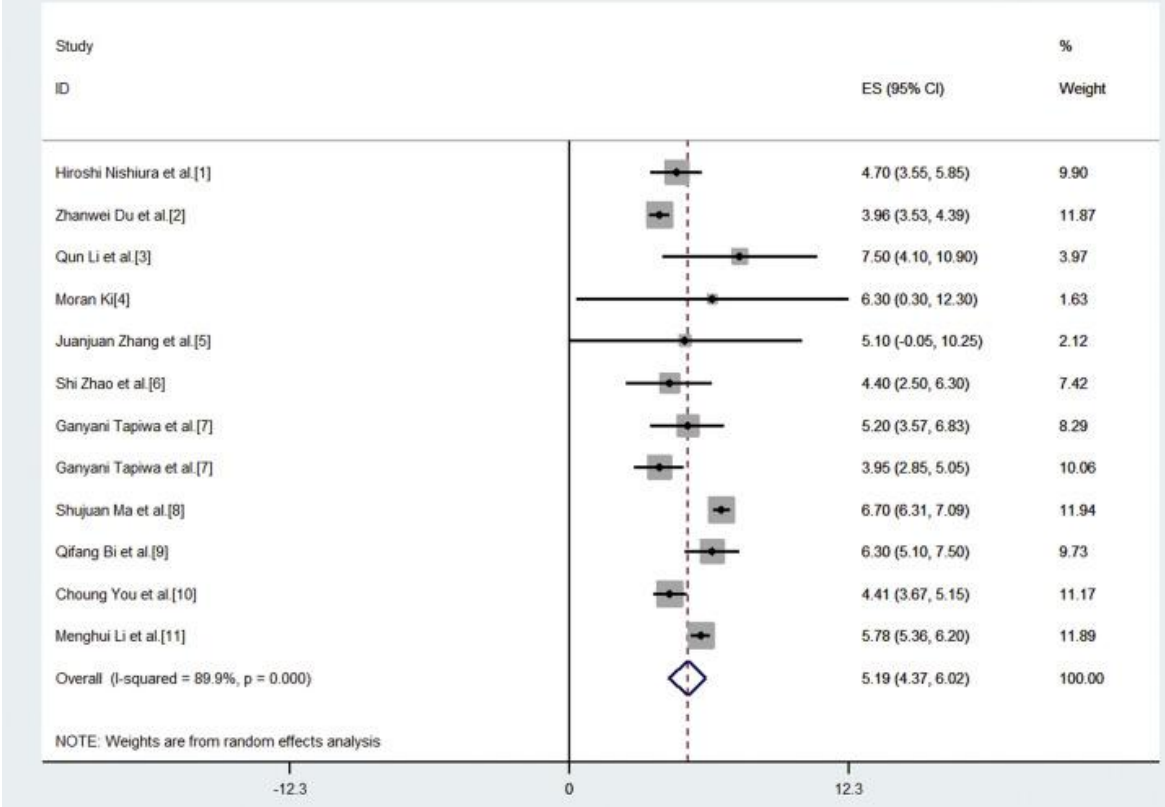
Systematic Review and Meta-analysis of Serial Intervals for COVID-19

Rai, Shukla, Dwivedi (2021)

Estimates mean serial interval of **5.19 days**

(95% CI 4.37,6.02).

All included studies use contact data.



Key innovation: use virus sequences as a proxy for contact data

If A and B's viruses are very closely related (indistinguishable, or nearly so) and they occur in the same place, with realistic timing, A and B are a "*plausible transmission pair*".

If A was observed or had symptoms first, A might have infected B (including indirectly, ie A → X → B where X is not observed).

This makes a link between virus sequence data and serial interval estimates.

Sequence 1: A A C T A G G T A G

Sequence 2: A G T T A G G C G G

Our approach in brief

- Start with clusters of cases based on genomic and epi data.
- Identify *plausible* transmission pairs using genomic distance and differences in symptom onset times.
- Sample possible transmission trees, composed of plausible pairs: each individual can have at most one infector
- For each tree, estimate parameters from times between symptom onsets in infector/infectee pairs.
- Average parameters over sampled trees.

Confidence intervals for parameter estimates incorporate both limited number of pairs as well as uncertainty in the transmission tree.

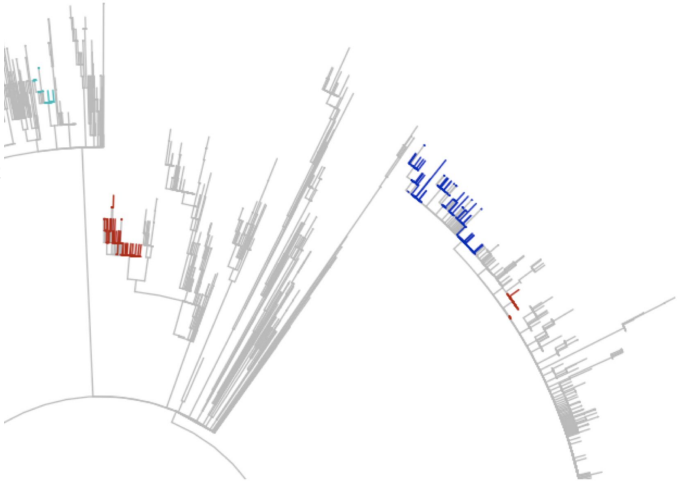


Illustration: clusters in BC on a phylogenetic tree

Data

Collaboration with Melbourne

We have a number of distinct “clusters”: groups of closely-related virus sequences, epidemiological links

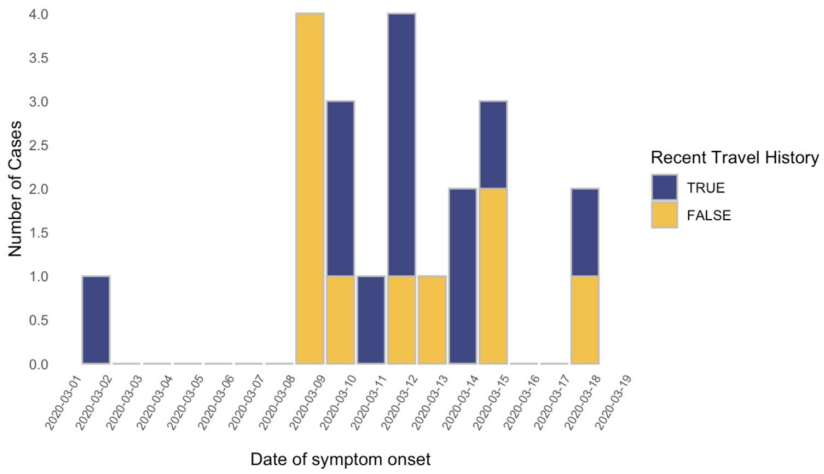
This is work in progress - demonstrative only

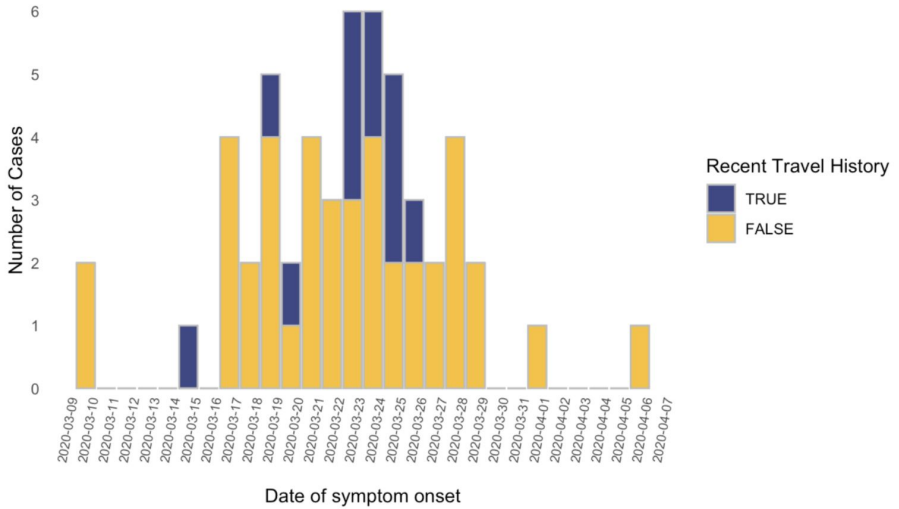
Article | Open Access | Published: 01 September 2020

Tracking the COVID-19 pandemic in Australia using genomics

Torsten Seemann, Courtney R. Lane, Norelle L. Sherry, Sebastian Duchene, Anders Gonçalves da Silva, Leon Caly, Michelle Sait, Susan A. Ballard, Kristy Horan, Mark B. Schultz, Tuyet Hoang, Marion Easton, Sally Dougall, Timothy P. Stinear, Julian Druce, Mike Catton, Brett Sutton, Annaliese van Diemen, Charles Alpren, Deborah A. Williamson & Benjamin P. Howden

Nature Communications **11**, Article number: 4376 (2020) | Cite this article





Previous work: Vink, Bootsma, Wallinga (2014)

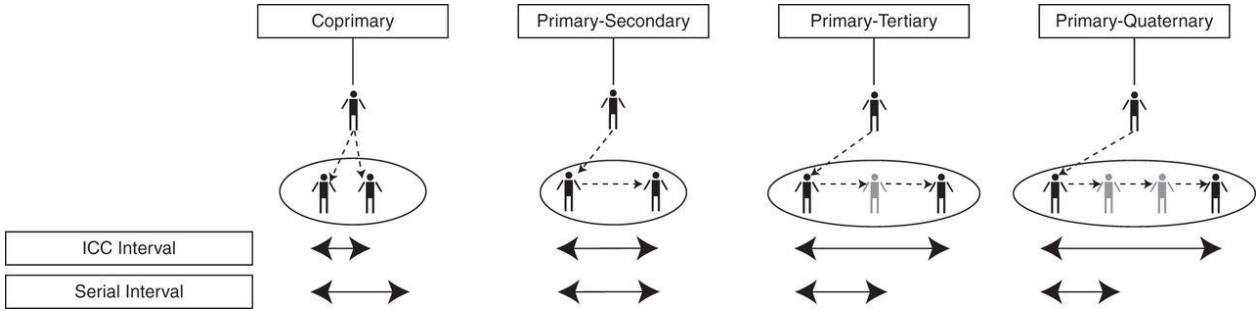
Use Index-Case to Case (ICC): in each cluster, index case is one with earliest symptom onset. All others are

- Coprimary
- Direct
- One intermediary
- Two intermediaries

Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis FREE

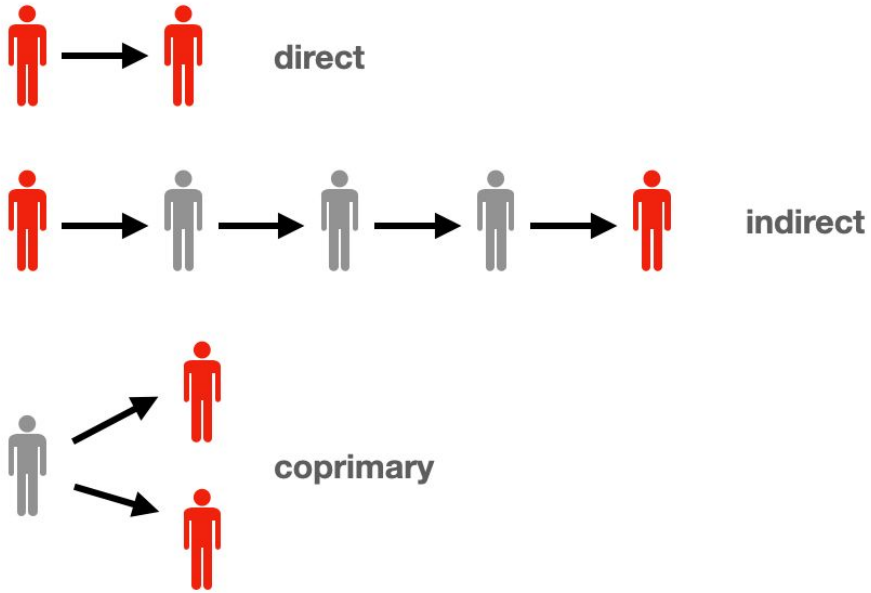
Margaretha Annelie Vink ✉, Martinus Christoffel Jozef Bootsma, Jacco Wallinga
Author Notes

American Journal of Epidemiology, Volume 180, Issue 9, 1 November 2014, Pages 865–875, https://doi.org/10.1093/aje/kwu209



Our Model for Time between symptom onset times of plausible pair

- *Direct:* Serial interval: Gamma distributed
- *Indirect:* Geometric number *m* of unobserved intermediates between the cases in the plausible pairs
- *Coprimary:* Account for possible unsampled infector of *both* A and B.
- Model with four parameters: determine with MLE and prior on sampling probability.



Comparison

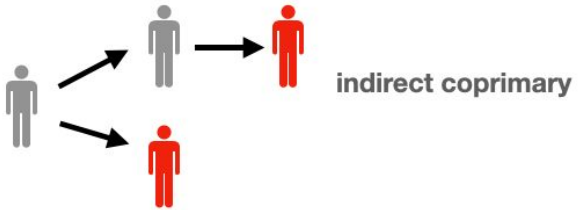
Our model:

- Geometric number of intermediaries
- Multiple index cases per cluster, but usually one with earliest symptom onset.
- Plausible link from genomic distance.
- Infector/infectee pairs determined by roughly computed transmission trees.

Vink, Bootsma, Wallinga (2014)

- One or two intermediaries.
- Index case unique earliest symptom onset.
- All cases linked with index case.
- Index case forms infector/infectee pair with all other cases.

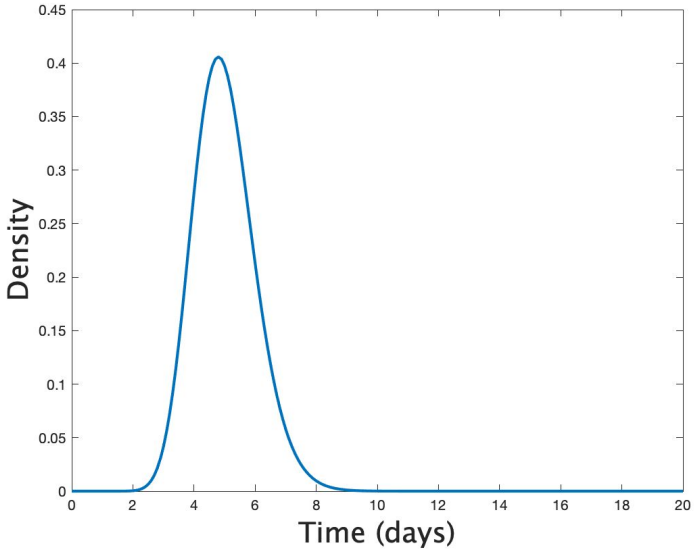
We both neglect “indirect comprimary” transmission.



Direct

Time between symptom onset times:

$t_{ij} \sim Gamma(\alpha, \beta)$ Gamma distribution: mean= $\mu = \frac{\alpha}{\beta}$,std= $\sigma = \frac{\sqrt{\alpha}}{\beta}$

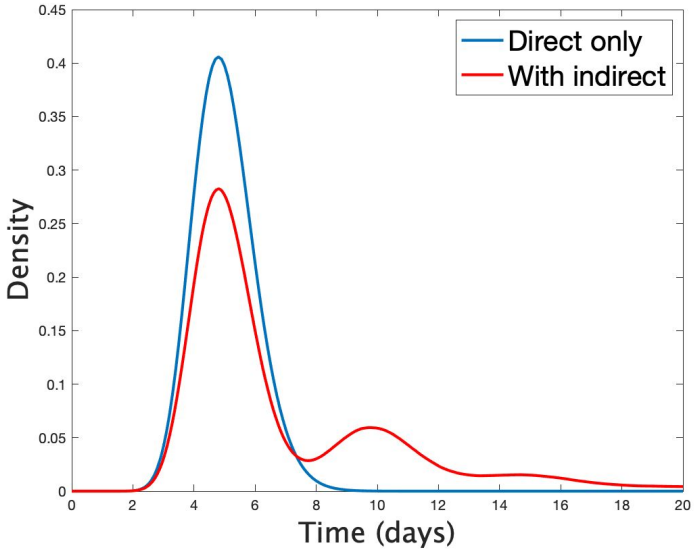


Direct and indirect

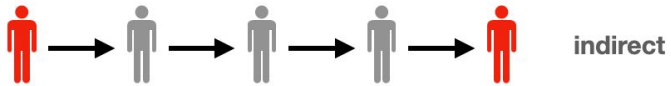
If we don't observe some cases, but don't have coprimary transmission

Use *Compound Geometric Gamma:* sum of geometric number of Gamma r.v.

$$t_{ij} \sim CGG(\alpha, \beta, \pi)$$



Prob each case is missed= π

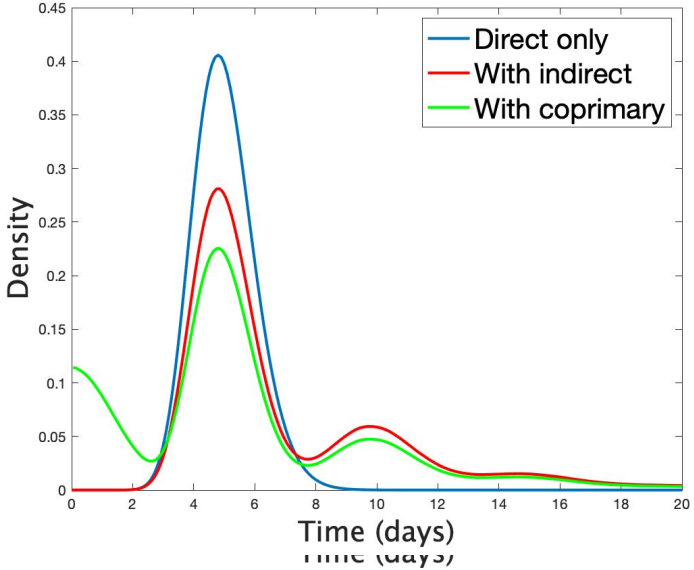


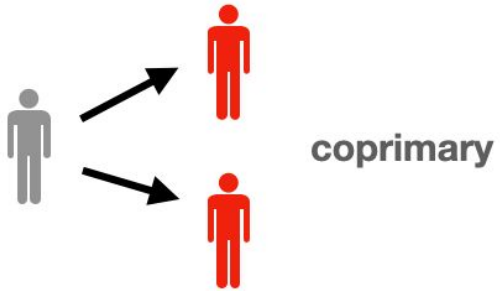
Direct, indirect, and comprimary

Include coprimary transmission: Use *Gamma Difference Distribution* (GDD)

$$t_{ij} \sim w\ CGG(\alpha, \beta, \pi)\ +\ (1 - w)\ GDD(\alpha, \beta)$$

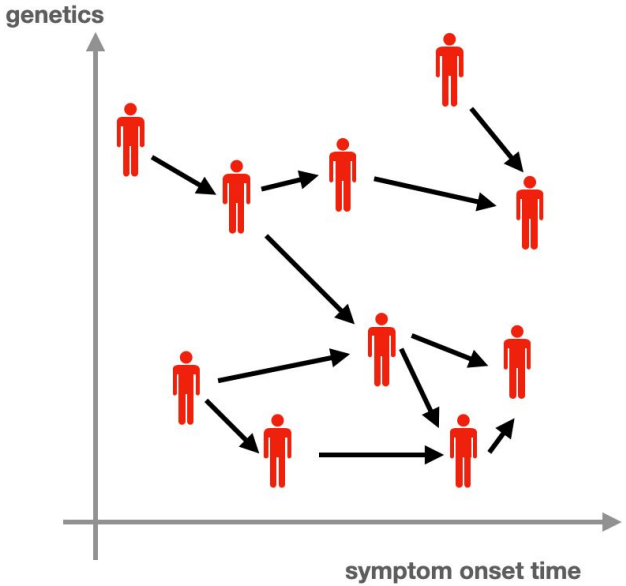
Fraction non-coprimary= w





Sampling plausible transmission trees

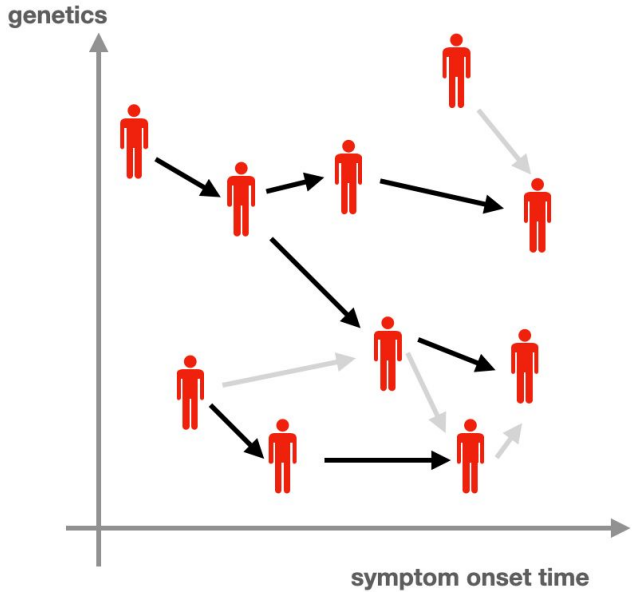
Start by linking all cases that are close enough in symptom-onset time, and close enough genetically. Earlier case is infector, later is infectee.



Problem: cases can have multiple infectors. Not a tree.

Sampling plausible transmission trees

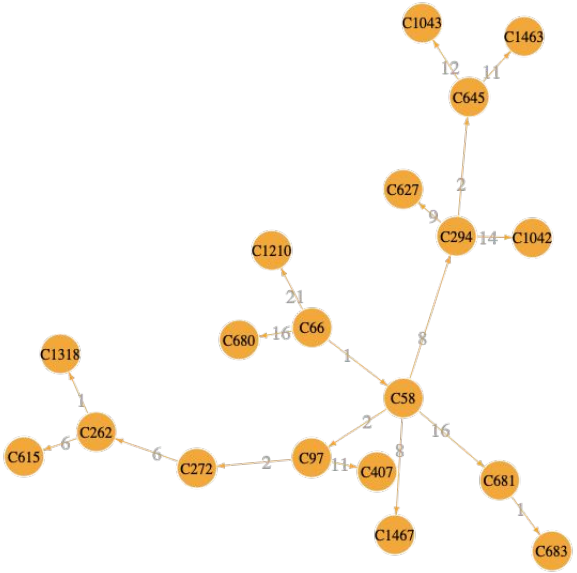
Select uniformly at random infector from all possible infectors.



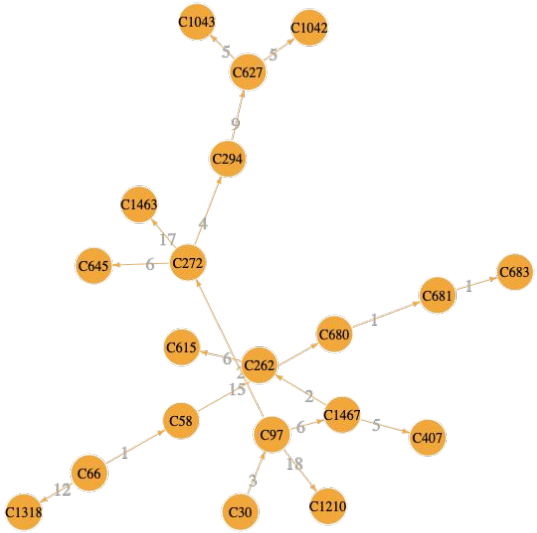
This gives a plausible transmission tree (or forest).
Estimate parameters using infector/infectee pairs in this tree.
Repeat for many trees.

Sampled trees

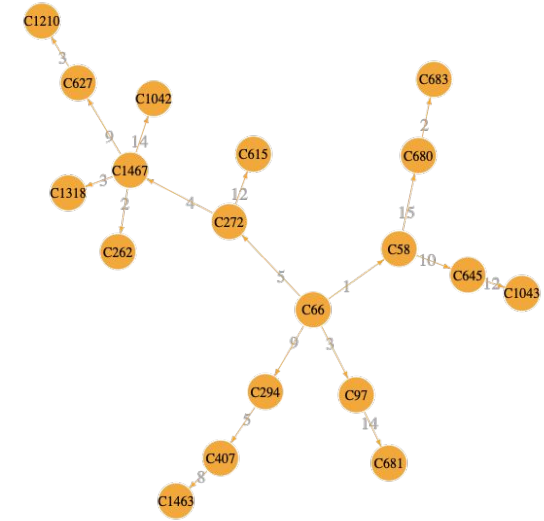
Cluster 9, Tree 1



Cluster 9, Tree 3



Cluster 9, Tree 4



Parameter estimation with uncertain trees

Each tree τ gives parameters Θ_τ with standard error SE_τ (from MLE + inv Hessian).

Our estimate of parameters is $\Theta = \mathbb{E}_\tau \Theta_\tau$ (average estimates over all trees)

Standard error of Θ is (from Law of Total Variance.)

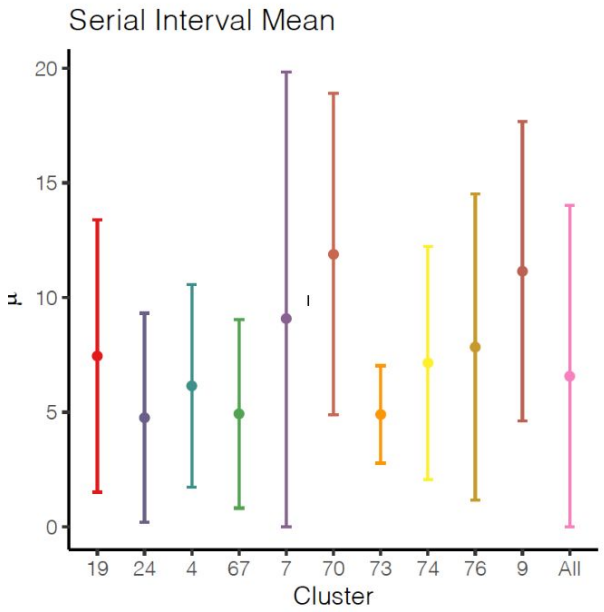
$$SE^2 = \mathbb{E}_\tau (SE_\tau)^2 + \mathrm{Var}_\tau \Theta_\tau$$

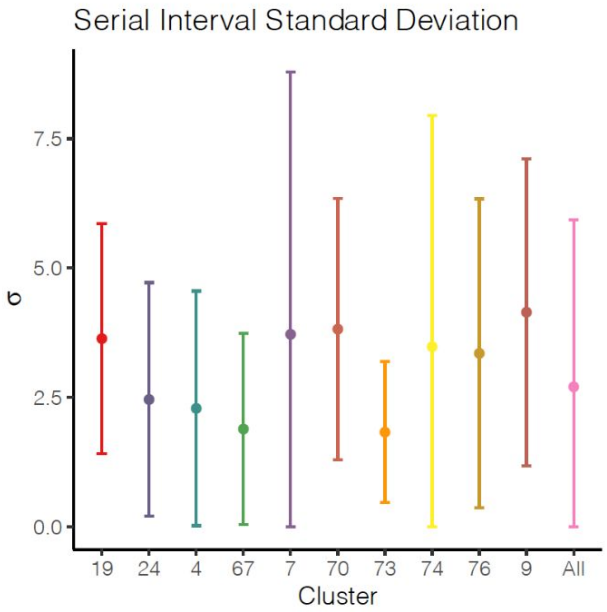
Uncertainty in parameter estimate.

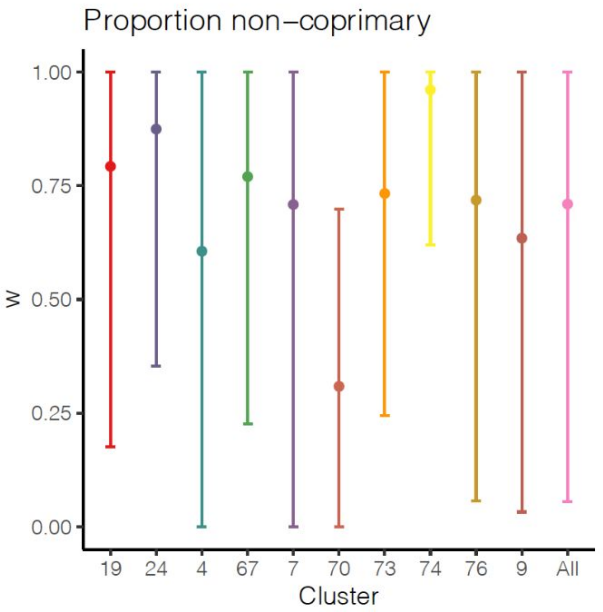
Uncertainty for each tree, averaged over trees. (finite data)

Uncertainty due to uncertainty in tree.

Parameter Estimates from clusters in Melbourne's first wave





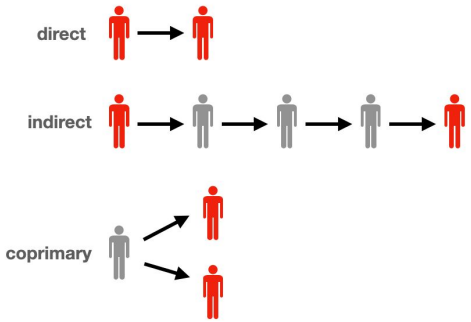


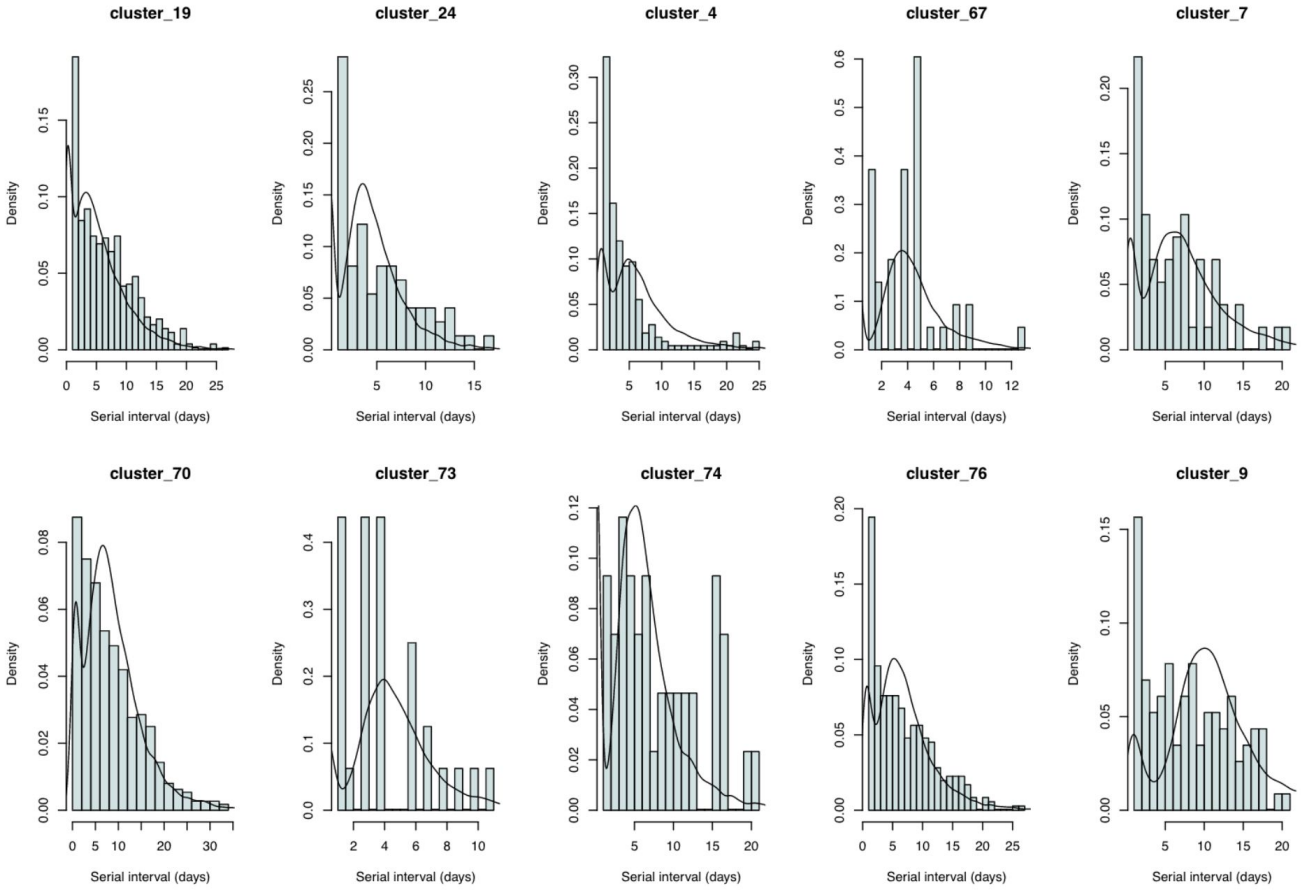
Estimates consistent with international estimates.
No evidence for differences in mean serial interval between clusters.

Cluster comparison

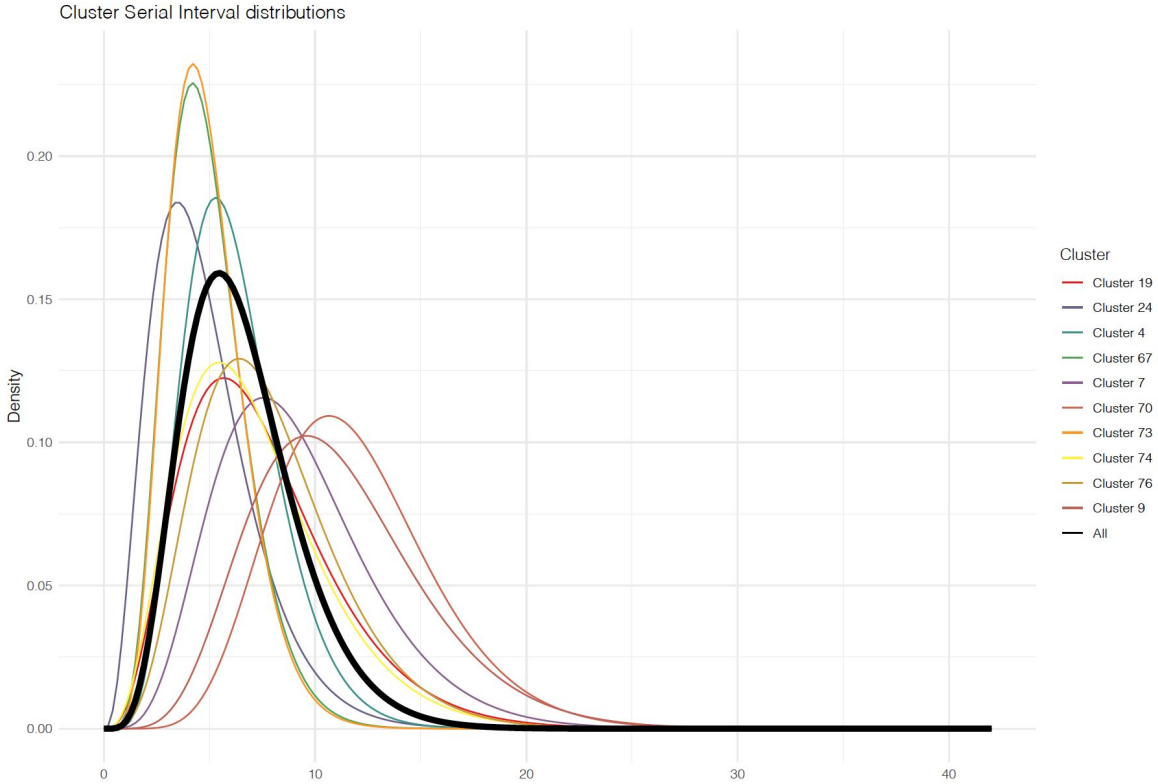
Black lines: density for the “plausible pair” time intervals in each cluster

Bars: pair intervals in the cluster, for one tree.



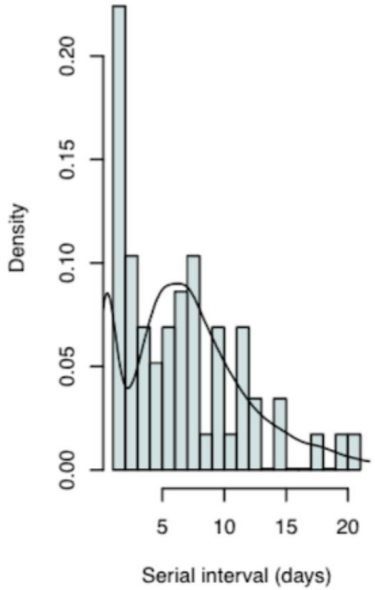


Estimated serial interval distribution by cluster

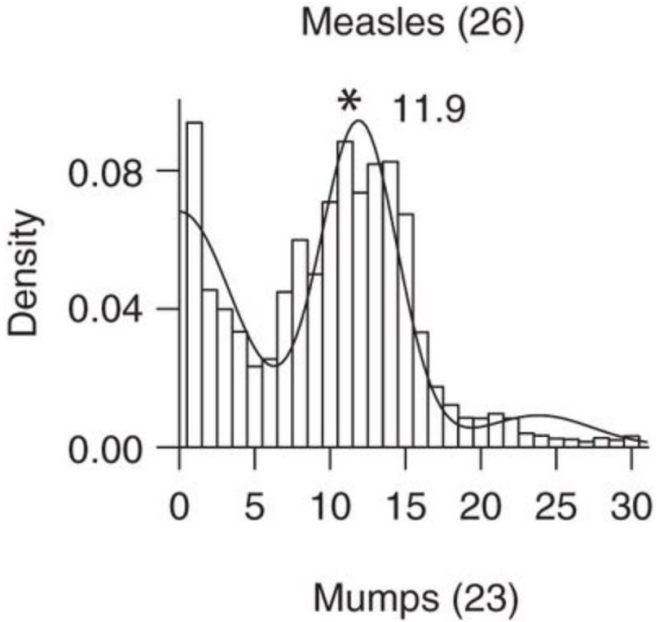


Comparison of results

Our model: COVID-19, Melbourne



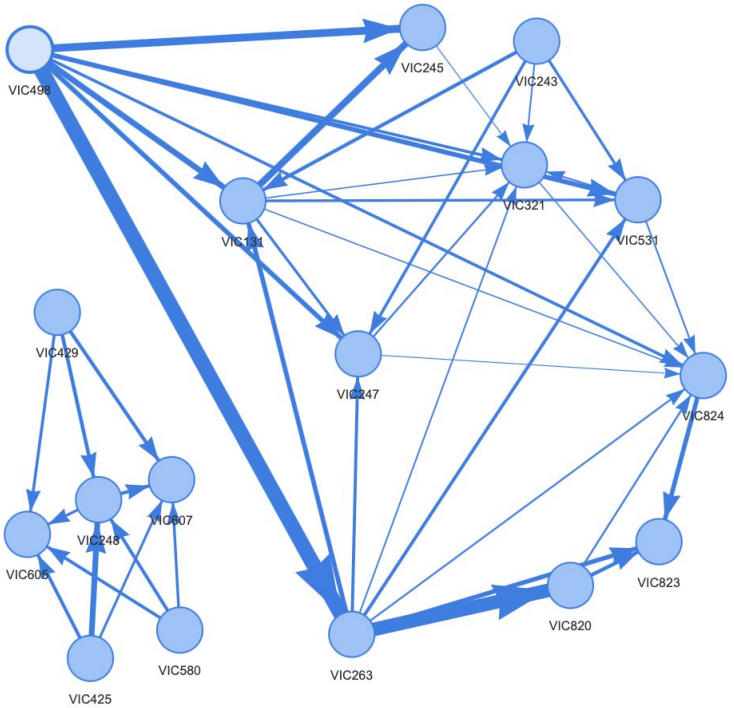
Vink et al 2014: Measles, Providence R.I.



Higher quality data gives Vink et al more precise estimates.

Next steps: Applications

- Wave 2 Melbourne data -- underway
- VOC comparisons where possible
- Sampled transmission trees: who might have infected whom?
- Other transmission dynamics: R_0, R.
- Data from household studies: compare with Vink et al.



Next steps: Methodology

- Simulation studies to understand systematic errors in our approach.
- Incorporating prior information on ascertainment rate and coprimary transmission
- Using with full phylogenetic reconstruction
- Incorporating contact-tracing data.

Thanks for your attention!