Research Forefronts: Estimation of serial intervals using pathogen genomic data

with an application to COVID-19

Sometimes, our data might not be sufficient to fully reconstruct the transmission tree

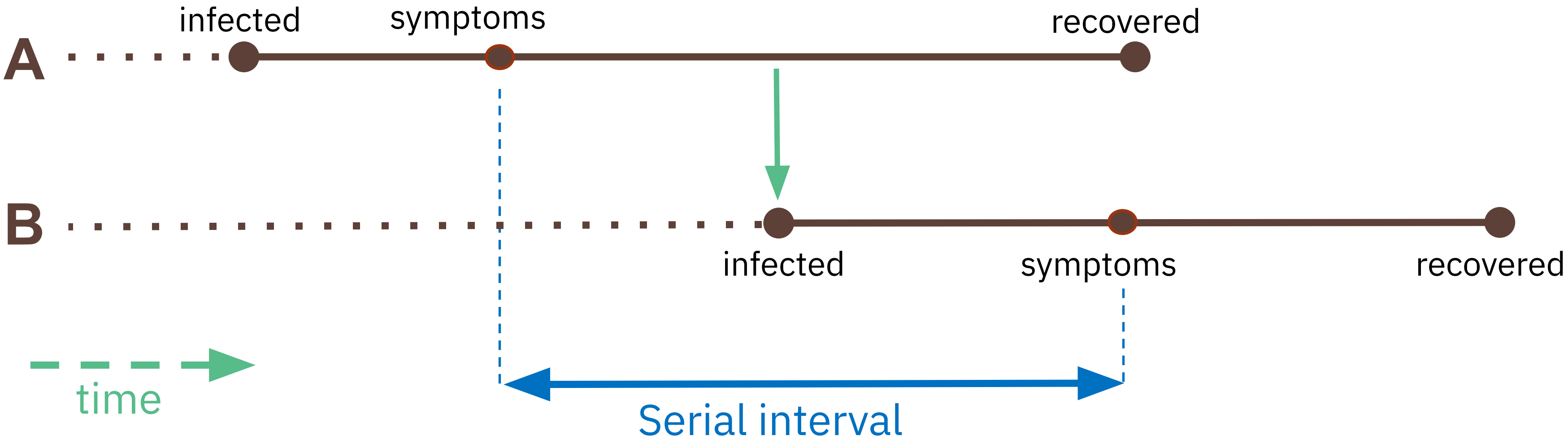
But that doesn't mean there's nothing we can learn…

We developed a method to estimate **serial intervals** using genomic data.

What is the serial interval?

Definition of the serial interval = length of time between successive cases in a chain of transmission

= length of time between symptom onset in an infector and infectee



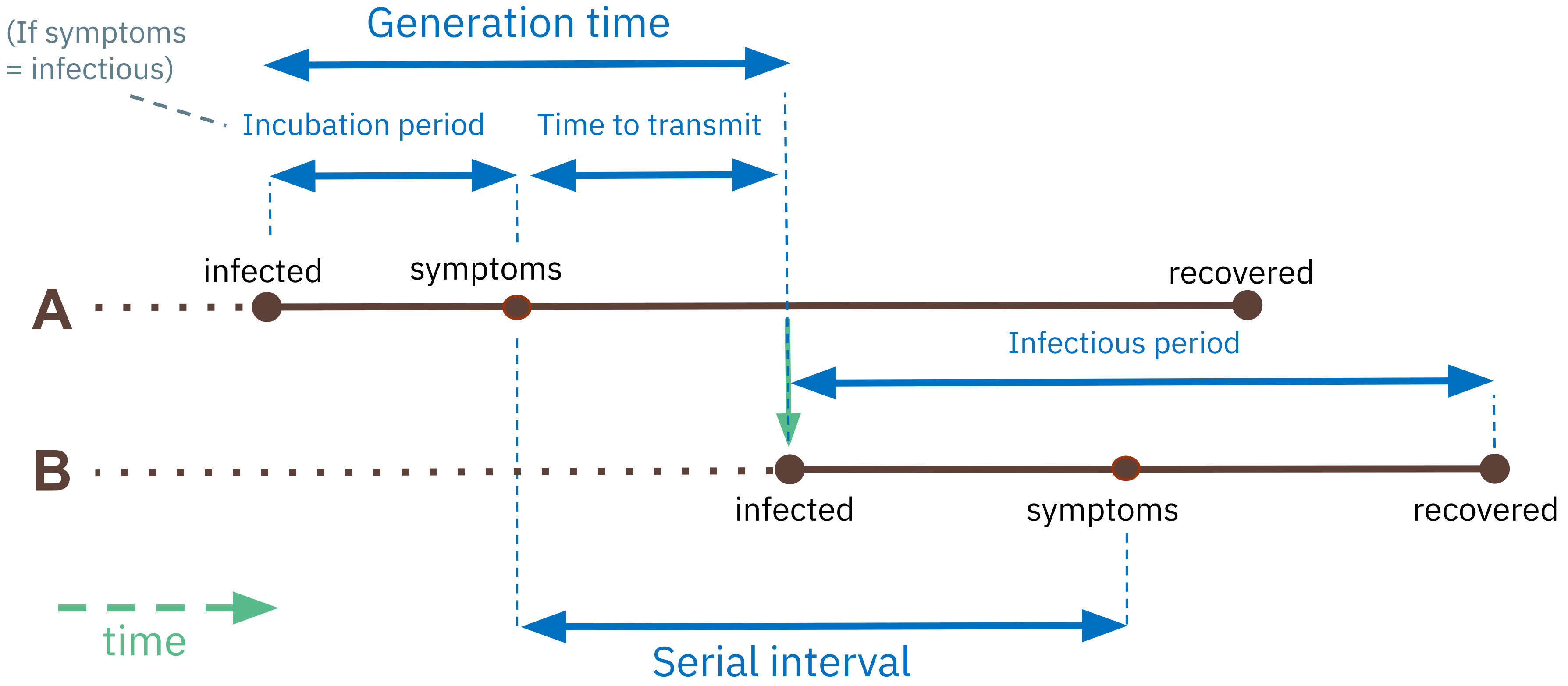
Why is the serial interval important?

- Tells us about the speed of transmission...
- ...this informs surveillance efforts.
- Used to calculate quantities like R_0, R_t ...
- $R_0 \approx 1 + rS$ ← Epidemic exponential growth rate x serial interval *
- ...and hence in understanding herd immunity thresholds and more.

R_0 = the average number of cases caused by a single infected individual, in a wholly susceptible population
R_t = the average number of cases caused by a single infected individual, at a specific time t

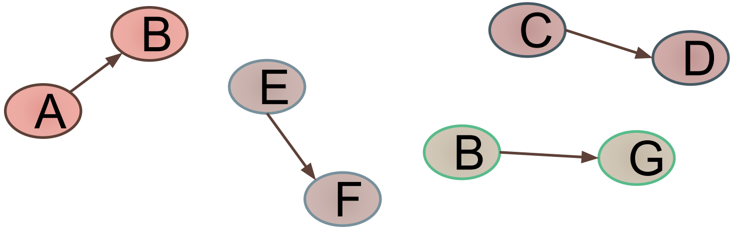
* How generation intervals shape the relationship between growth rates and reproductive numbers, Wallinga and Lipsitch (2006) *Proceedings of the Royal Society B*

It’s also related to other disease intervals

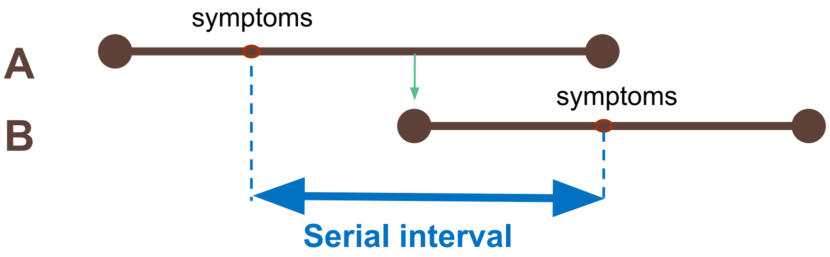


Most existing methods for serial interval estimation assume direct observation of transmission pairs (infectors & infectees)

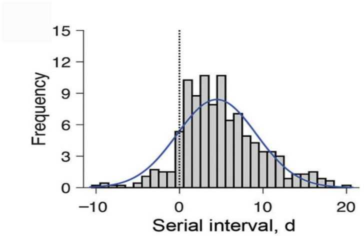
1. Contact trace pairs of cases which are assumed to represent direct transmission



2. This provides **direct observations** of the serial interval



3. Parametric estimation of the serial interval given this observed data

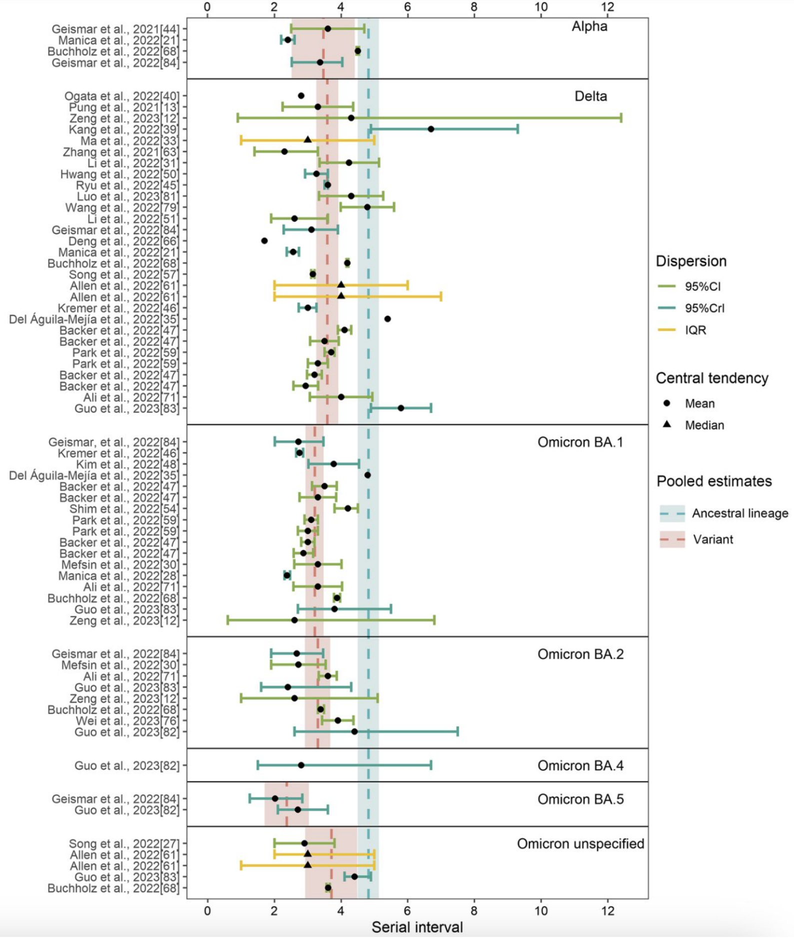


[Serial Interval of COVID-19 among Publicly Reported Confirmed Cases]()
Du et al. (2020) *Emerging Infectious Diseases*

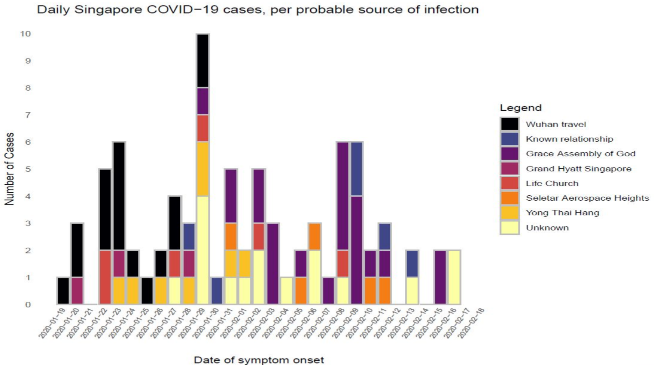
Systematic Review and Meta-analysis of Serial Intervals for COVID-19

Xu et al (2023) estimate a mean serial interval from ancestral lineage SARS-CoV-2 of **4.82 days (95% CI 4.5 - 5.14)**

All included studies (98) use contact data, the majority of which assume direct contact-traced pairs and many of which were household pairs.



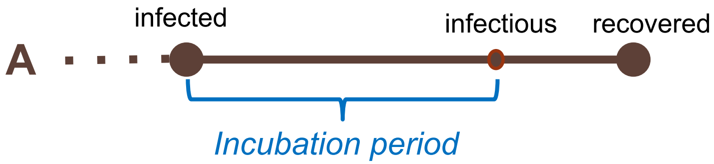
How we got thinking about serial intervals…



Evidence for transmission of COVID-19 prior to symptom onset. Tindale, Stockdale et al (2020) *eLife*

By collating contact data from outbreaks in Singapore and Tianjin, we estimated the amount of **pre-symptomatic transmission** of COVID-19.

This requires estimation of both the serial interval and incubation period.



“About 40% to 80% of the novel coronavirus transmission occurs two to four days before an infected person has symptoms”

We developed a new approach for estimating incubation periods whilst taking into account that observed pairs may not represent direct transmission

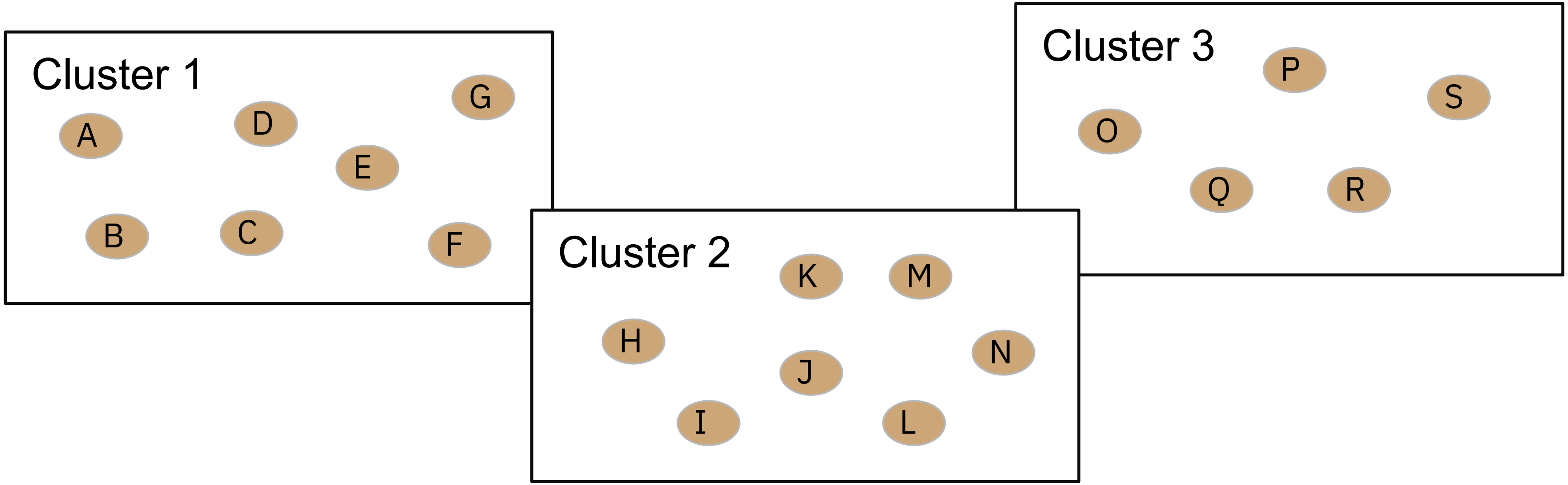
Contact tracing approaches require detailed personal data, and are often limited to household studies

This motivated a new genomic approach:

- Use pathogen whole genome sequence data as a proxy for contact data
- Use in broader clusters than e.g. households
- Incorporate possibility of missing cases
- Fast and cluster-specific estimates: track the serial interval through time and under different settings or variants

Estimating serial intervals with genomic data

Suppose we have a set of case clusters (perhaps genomic clusters, or clusters associated with e.g. schools, hospitals) from an outbreak of infectious disease. We wish to **estimate the serial interval in each cluster...**



We know each case's symptom onset time and pathogen sequence, but we don't know who infected whom.

Whole genome sequences as a proxy for contact data

The main idea: As with the other methods in this course, differences in the sequences tell us how closely related people's infections are and therefore who might have infected whom.

```
Case A: ATCGGTATCAGTCAG

Case B: ATCAGTATCAGTCAG
```

However, since we want to work with broad clusters where we don't necessarily sample a large proportion of cases, we need to consider

1. There may be **large** uncertainty in who infected whom
2. Inferred pairs (infector & infectee) might not represent direct transmission

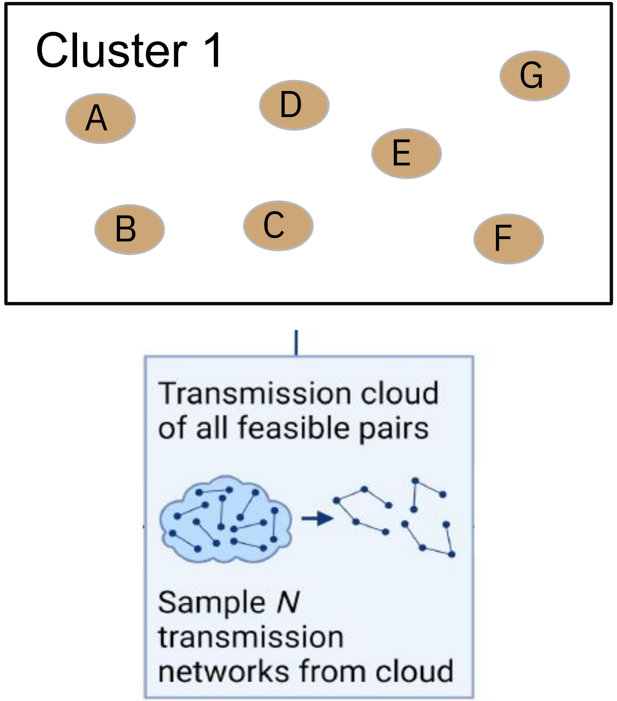
Take uncertainty in who infected whom into account, by sampling feasible transmission networks

1. Identify all plausible transmission pairs: Pairs *(i,j)* in the same cluster, with closely related sequences & realistic timing, where infector *i* showed symptoms first.

1. *Difference in symptom onset date ≤ T*
2. *Pairwise genomic distance ≤ G*

```
Case i: ATCGGTATCAG
Case j: ATCAGTATCAG
```

2. Sample plausible transmission networks: Built from the plausible pairs, by sampling an infector for each infectee.

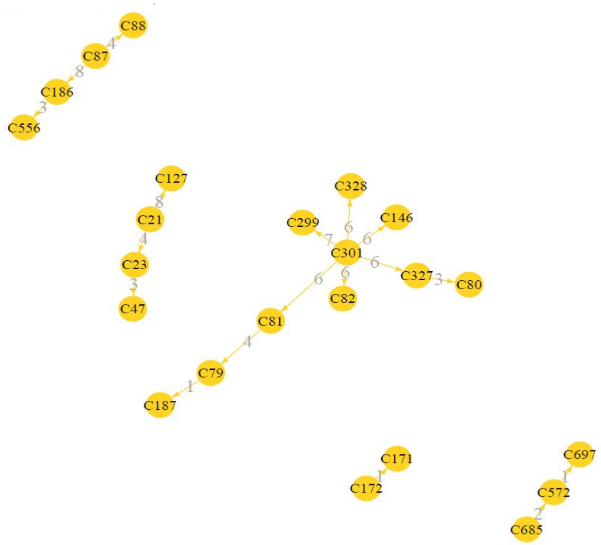


Transmission cloud of all feasible pairs

Sample *N* transmission networks from cloud

Take uncertainty in who infected whom into account, by sampling feasible transmission networks

Incorporate uncertainty by sampling a set of networks:

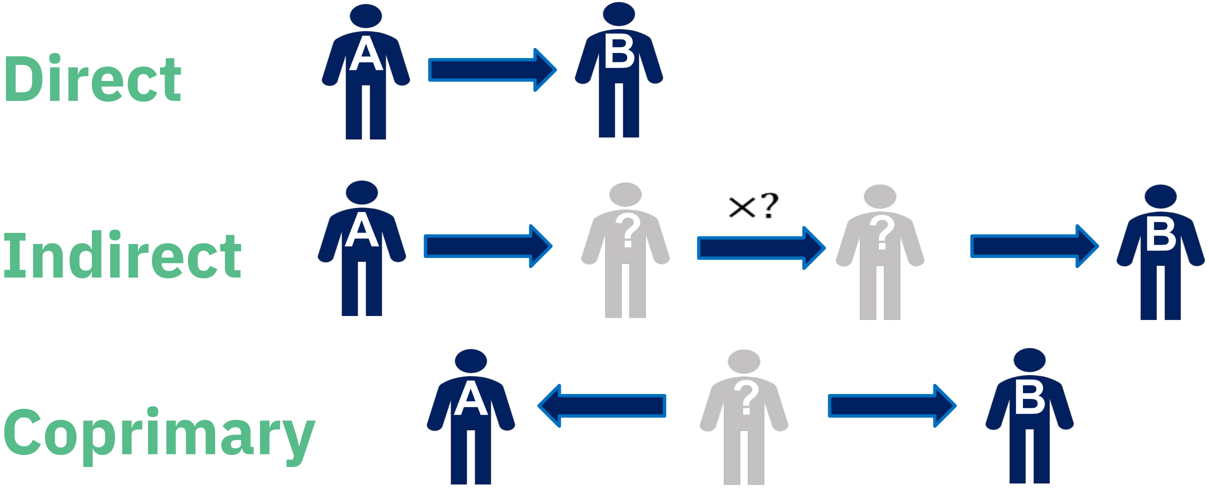


x *N*

We estimate the serial interval in each network independently – and then **average across networks**

Estimate the serial interval distribution, taking indirect transmission into account

To take under-sampling into account, we consider that, for every infector-infectee pair (A, B) in every sampled network, transmission may have been:



Direct

Indirect

Coprimary

Inspired by:

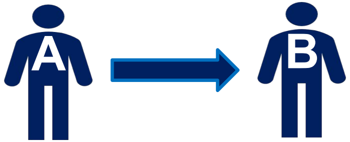
[Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis](#)
Vink et al. (2014)

We fit a mixture model to incorporate this idea...

Estimate the serial interval distribution: possible pathways

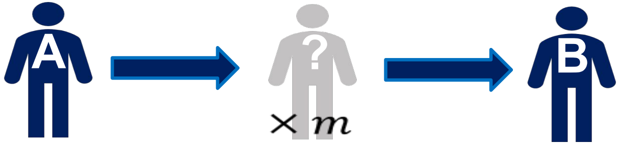
True serial interval distribution $\sim \Gamma(\mu, \sigma)$

Direct



Observed time difference between A & B, $T_{a,b} \sim \Gamma(\mu, \sigma)$

Indirect



$T_{a,b} =$ sum of $m+1$ serial intervals, $m \sim \mathbf{Geo}(\pi)$

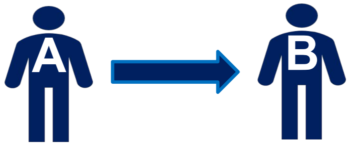
$\sim$Compound Geometric Gamma(μ, σ, π)

Proportion w of pairs

Estimate the serial interval distribution: possible pathways

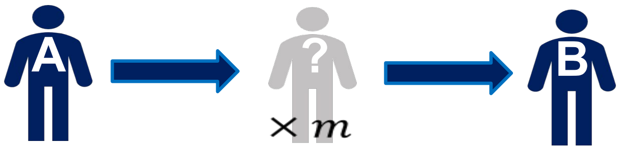
True serial interval distribution $\sim \Gamma(\mu, \sigma)$

Direct



Observed time difference between A & B, $T_{a,b} \sim \Gamma(\mu, \sigma)$

Indirect



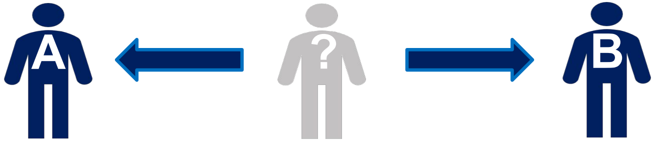
Probability of sampling an infectee

$T_{a,b} =$ sum of $m+1$ serial intervals, $m \sim \mathrm{Geo}(\pi)$
$\sim$Compound Geometric Gamma(μ, σ, π)

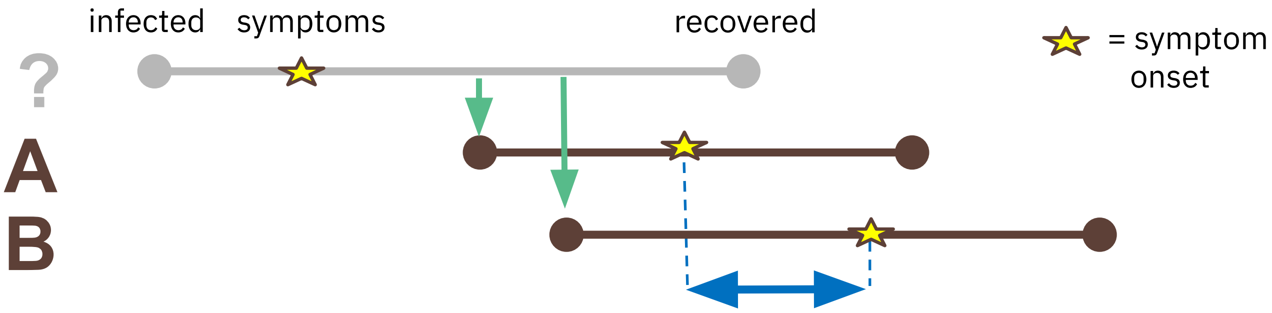
Proportion w of pairs

Estimate the serial interval distribution: possible pathways

Coprimary



$T_{a,b}$ is the strictly non-negative difference of two $\Gamma(\mu, \sigma)$ distributions



time

$$T_{a,b} \sim \text{Folded Gamma difference}(\mu, \sigma)^{*}$$

Proportion 1-w of pairs

* A note on gamma difference distributions, Bernhard Klar, 2015

Estimate the serial interval distribution: mixture model taking possible pathways into account

We combine these possible transmission pathways into a mixture model with log-likelihood:

$$l(\mu, \sigma, \pi, w|D) = \sum_{k=1}^{n} \log[\mathrm{wf}_{\mathrm{CGG}}(T_{a_k,b_k}|\mu, \sigma, \pi) + (1 - w) f_{\mathrm{FGD}}(T_{a_k,b_k}|\mu, \sigma)]$$

Sum over all pairs in the network

Direct or indirect

Coprimary

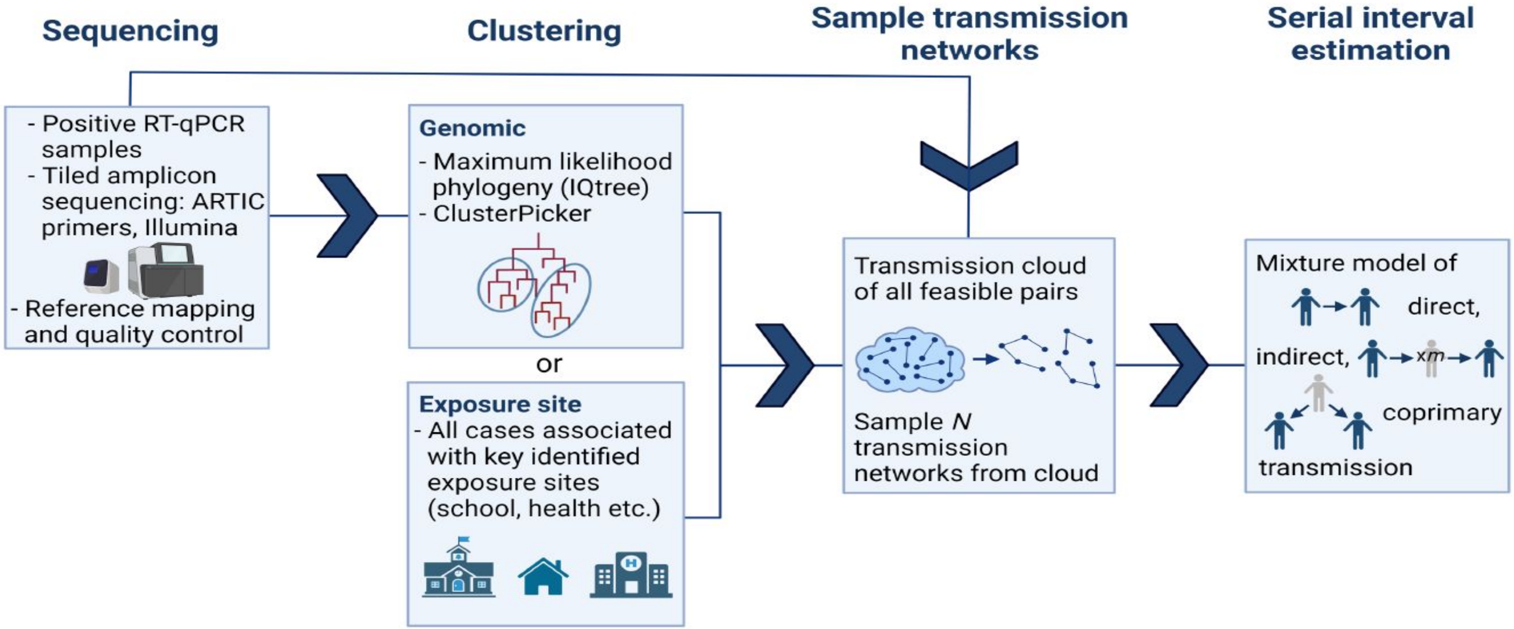
Instead of maximising this directly, we incorporate Beta distributed priors for *w* and *π*, and perform maximum a posteriori (MAP) estimation. We calculate the MAP for each sampled network, and then average over all networks

Our confidence intervals need to take into account uncertainty in each network, as well as uncertainty when combining across networks:

$$\hat{\mathrm{Var}}(\hat{\mu}_c) = \mathbb{E}_{\tau}\left(\hat{\mathrm{se}}(\hat{\mu}_{c,\tau_k})^2\right) + \mathrm{Var}_{\tau}\left(\hat{\mu}_{c,\tau_k}\right).$$

For cluster c and each network τ_k

A schematic view of the method



Sequencing

- Positive RT-qPCR samples
- Tiled amplicon sequencing: ARTIC primers, Illumina
- Reference mapping and quality control

Clustering

Genomic

- Maximum likelihood phylogeny (IQtree)
- ClusterPicker

or

Exposure site

- All cases associated with key identified exposure sites (school, health etc.)

Sample transmission networks

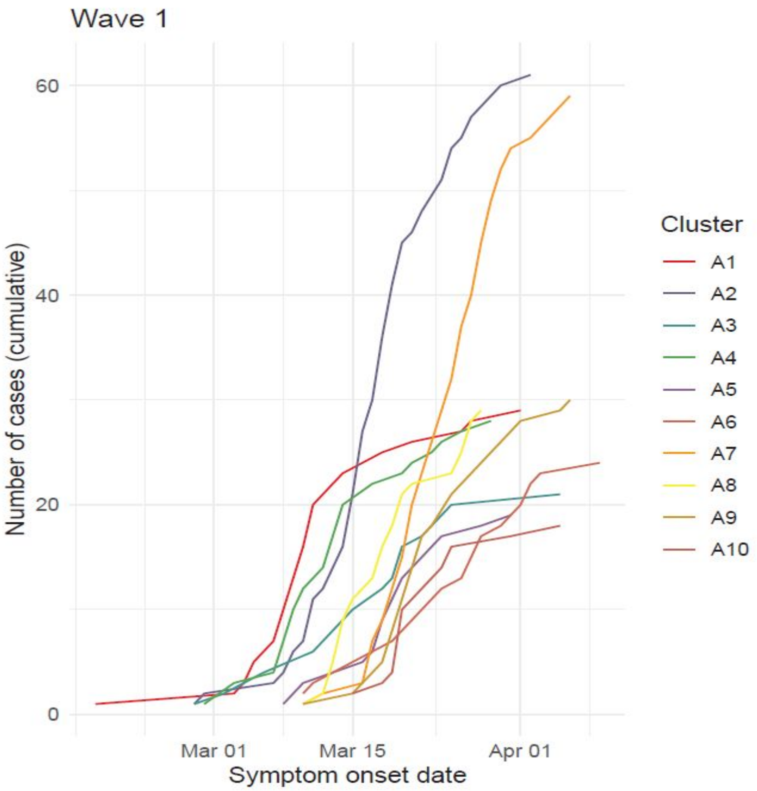
Transmission cloud of all feasible pairs

Sample *N* transmission networks from cloud

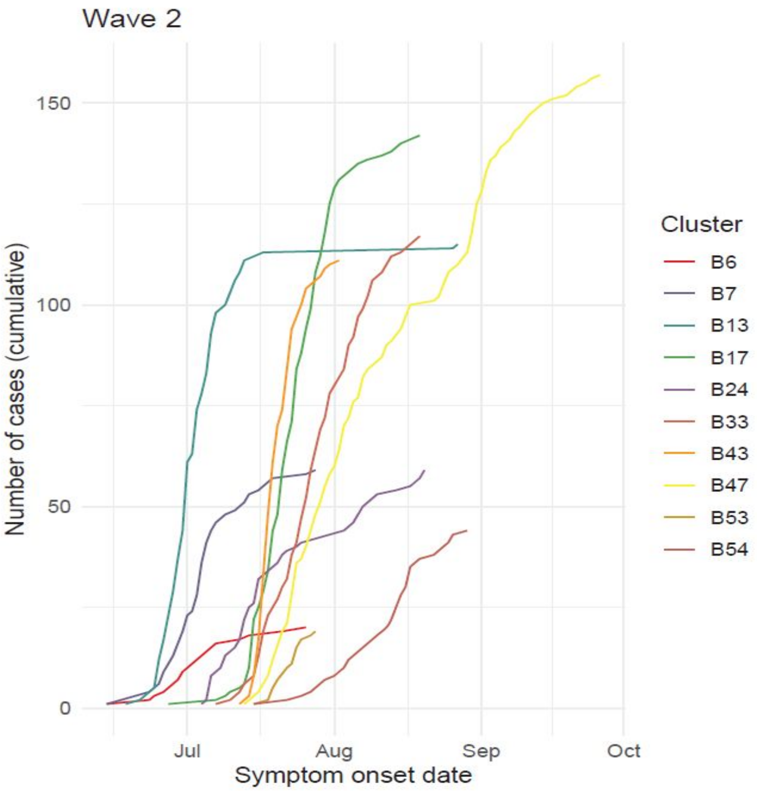
Serial interval estimation

Mixture model of direct, indirect, coprimary transmission

Application: COVID-19 clusters in Victoria, Australia

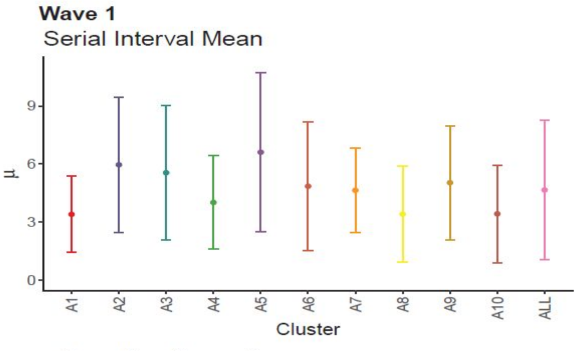


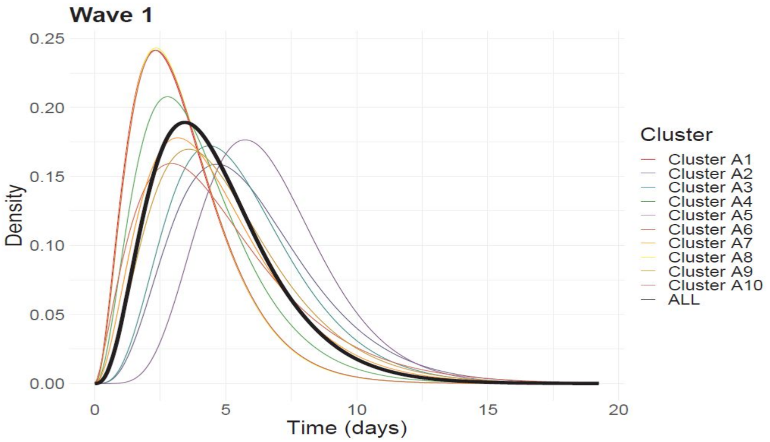
6 January–14 April 2020

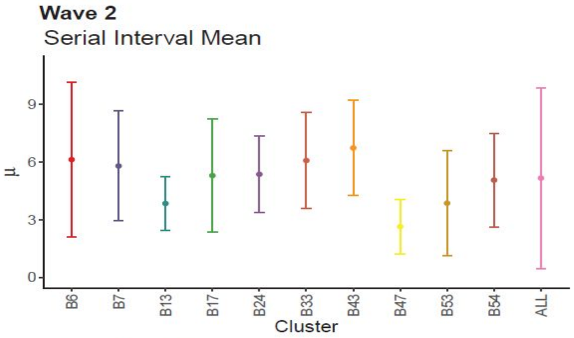


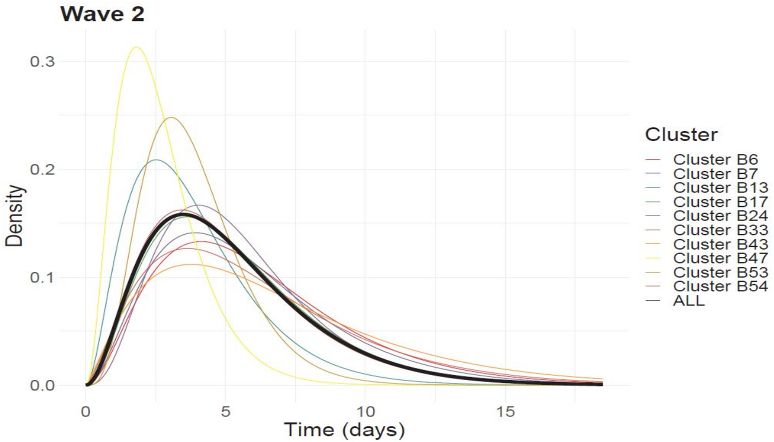
1 June–28 October 2020

Cluster-specific serial intervals: in line with published estimates, with some variation by cluster





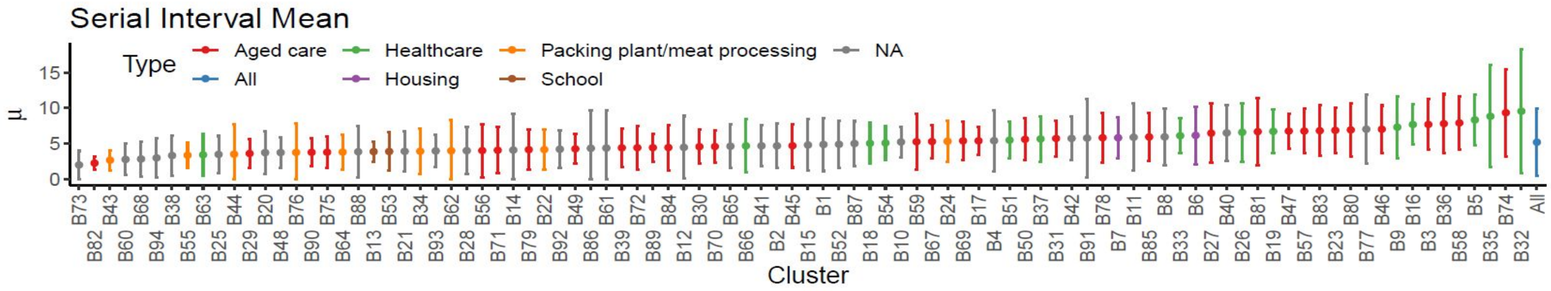




Context: Early published estimates ~5 days

Using a larger range of 2nd wave clusters, we can compare across different exposure settings

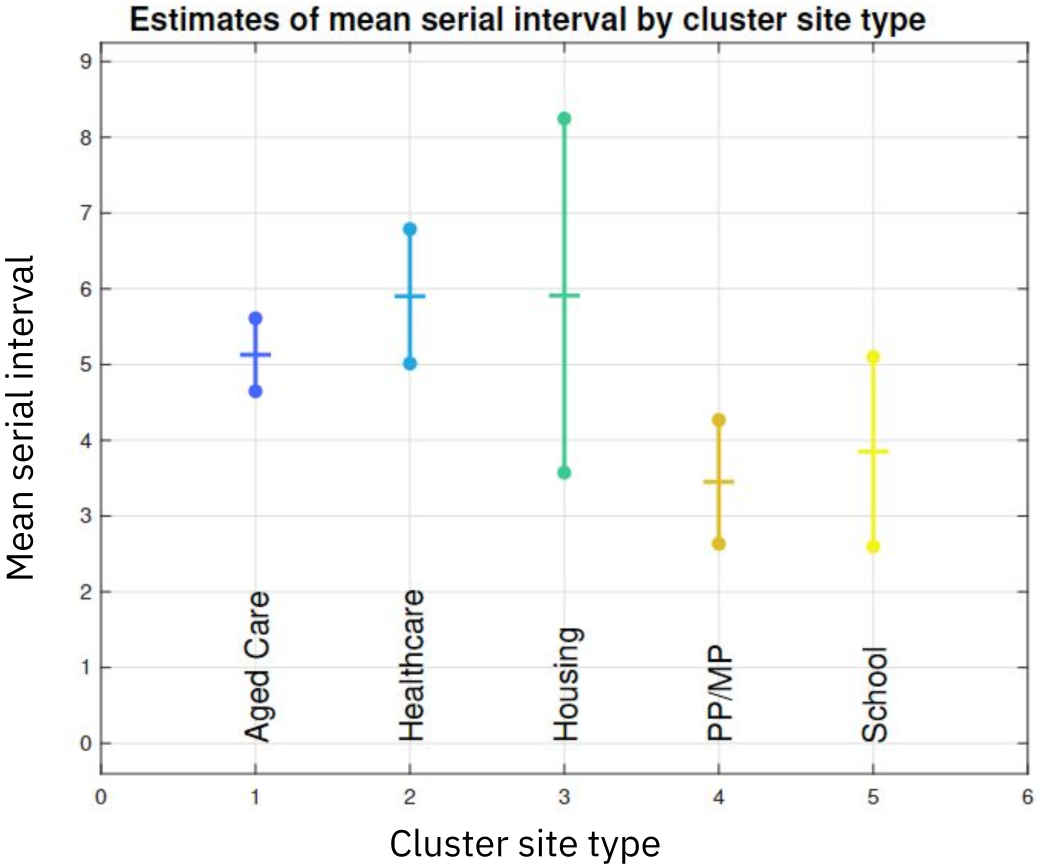
Estimates of the mean range from **2 to 9.5 days** (compared to standard estimates ~5 days)



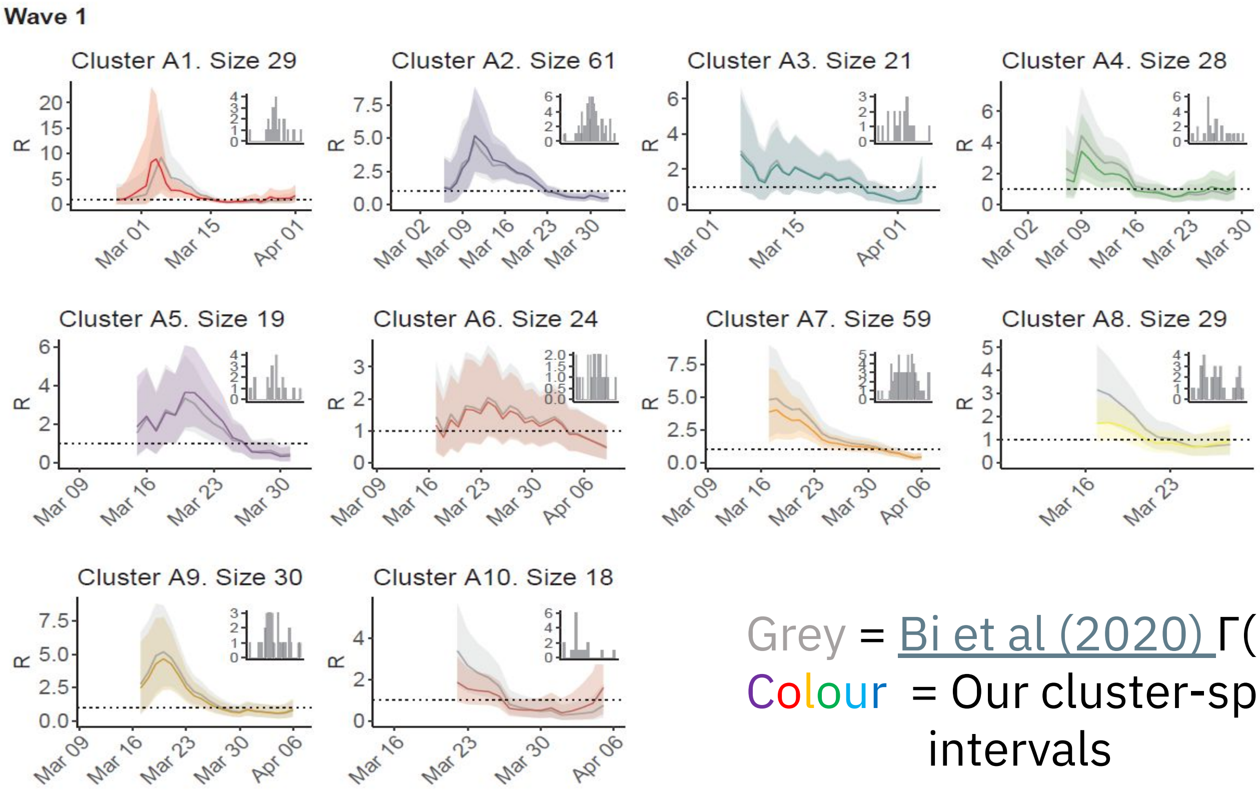
Among the shorter: packing plants, schools

Among the longer: aged-care, healthcare, housing

Using a larger range of 2nd wave clusters, we can compare across different exposure settings



Estimates of *Rt* are (sometimes) impacted by the underlying serial interval distribution



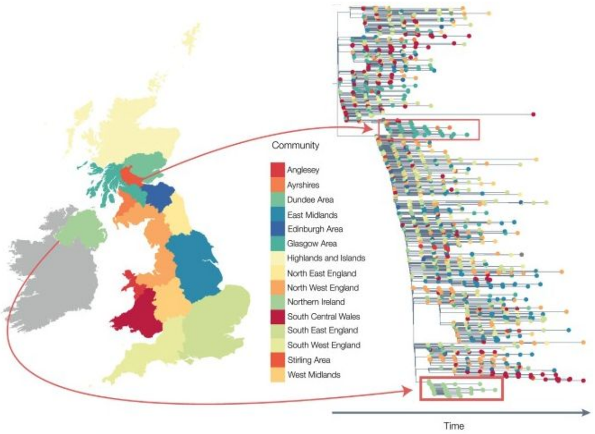
Grey = [Bi et al (2020)](#) Γ(μ = 6.3, σ = 4.2)

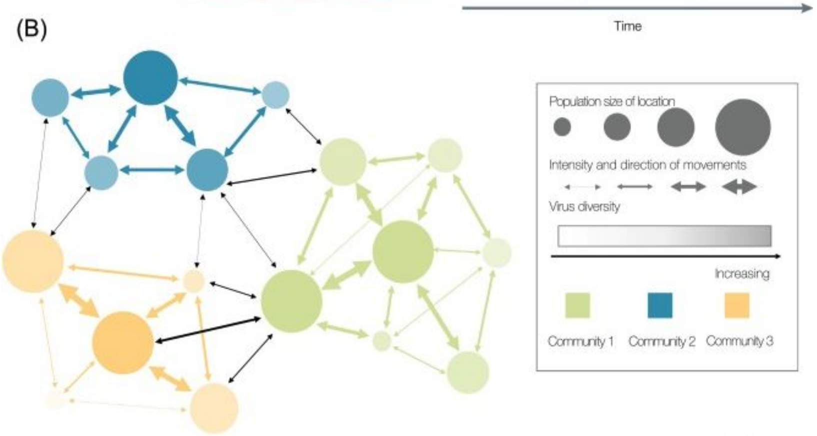
Colour = Our cluster-specific serial intervals

In conclusion

- It would have been difficult to do full transmission reconstruction (outbreaker, TransPhylo) for the Victoria data: **low diversity sequences, lots of missing cases (wave 2 particularly)**
- Even still, pathogen sequence data can help us learn about aspects of transmission. Here, we estimate serial intervals, without the need for contact studies
- Broad population sequencing makes it easier to compare serial intervals across time, space, setting, or Variants of Concern (VOC)

Genomic epi at different scales



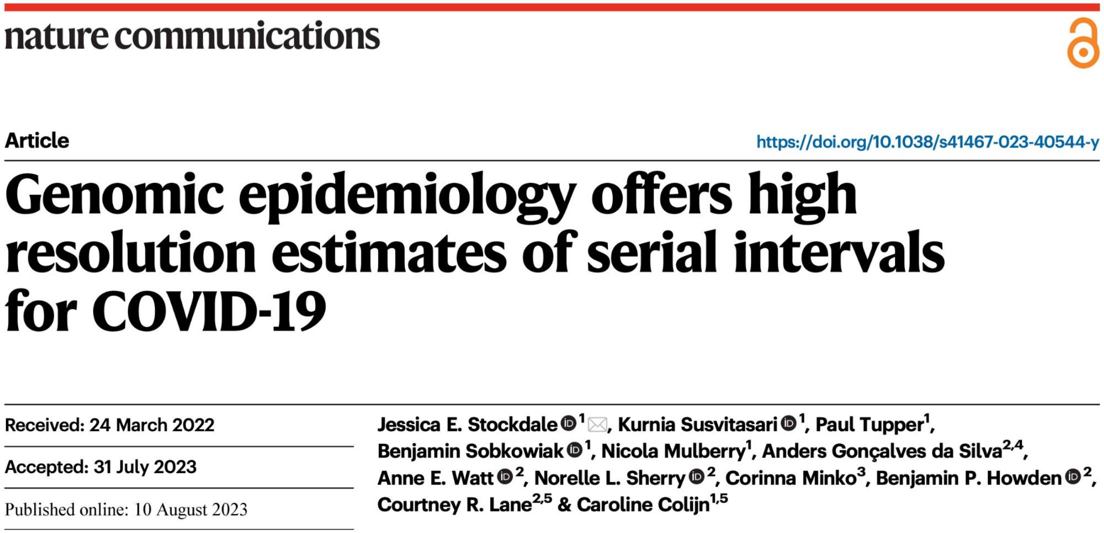


Population scale - phylogenetics, phylodynamics, phylogeography

Intermediate scale

Local scale - methods from this course

Graphics from Progress and challenges in virus genomic epidemiology, Hill et al (2021) *Trends in Parasitology*



nature communications

Article

https://doi.org/10.1038/s41467-023-40544-y

Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19

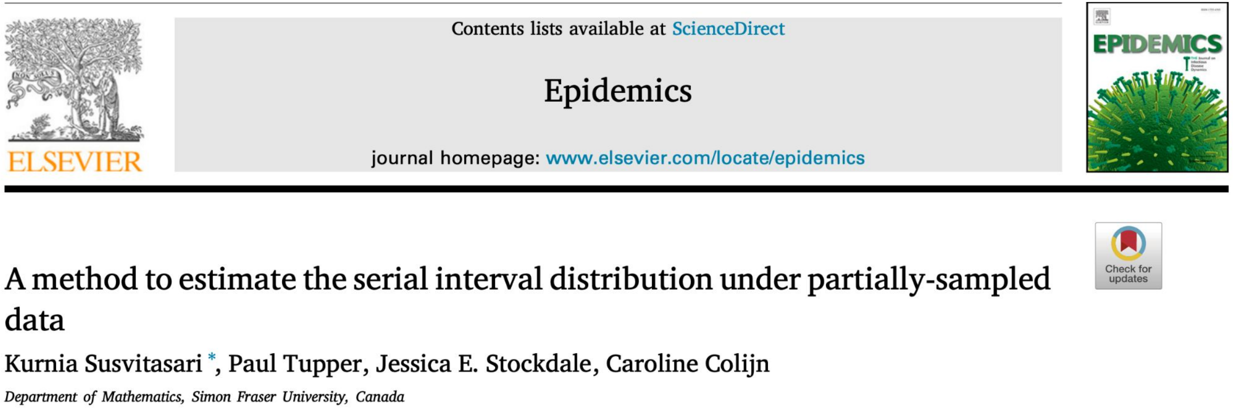
Received: 24 March 2022

Accepted: 31 July 2023

Published online: 10 August 2023

Jessica E. Stockdale[1], Kurnia Susvitasari[1], Paul Tupper[1], Benjamin Sobkowiak[1], Nicola Mulberry[1], Anders Gonçalves da Silva[2,4], Anne E. Watt[2], Norelle L. Sherry[2], Corinna Minko[3], Benjamin P. Howden[2], Courtney R. Lane[2,5] & Caroline Colijn[1,5]

Link



Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

ELSEVIER

A method to estimate the serial interval distribution under partially-sampled data

Kurnia Susvitasari *, Paul Tupper, Jessica E. Stockdale, Caroline Colijn

Department of Mathematics, Simon Fraser University, Canada

Link

Methods comparison and opportunities

Methods comparison (these are abilities, not quality)

Method	**Unsampled hosts**	**Phylogeny vs pairs**	**Multiple sequences per host**	**Simultaneous phylogeny and transmission**	**Bottleneck >1**	**Environmental organism**	**Incorporate epidemiological data (beyond times of collection, infectious period)**
BEASTLIER	✖	phylogeny	✅	✅	✖	✖	✖
TransPhylo	✅	phylogeny	✖	✖	✖	✖	✖
Outbreaker 2	✅	pairs	✖	✖	✖	✖	✅(readily)
Phybreak	✖	phylogeny	✖	✅	✖	✖	✖
SCOTTI	✅ (limited)	phylogeny	✅	✅	✅	✅(limited)	✖
BREATH	✅	phylogeny	✖ (in progress)	✅	✖	✖	✖

Some recent methods and studies

Ke and Vikalo, *Graph-Based Reconstruction and Analysis of Disease Transmission Networks Using Viral Genomic Data*, Journal of Computational Biology (2023)

Clustering + transmission reconstruction within clusters via graphs and host importance scores

Lindsey *et al. Characterising within-hospital SARS-CoV-2 transmission events using epidemiological and viral genomic data across two pandemic waves*, Nature Communications (2022).

Adapted *Outbreaker2* for hospital settings. Includes ward occupancy data

Junhang Pan *et al, TransFlow: a Snakemake workflow for transmission analysis of Mycobacterium tuberculosis whole-genome sequencing data*, Bioinformatics (2023)

Pipeline from raw sequences to clustering to transmission reconstruction, combining various existing methods

Van der Roest *et al, A Bayesian inference method to estimate transmission trees with multiple introductions; applied to SARS-CoV-2 in Dutch mink farms*, PLoS Comp Bio (2023)

Extension to Phybreak allowing for multiple pathogen introductions

Areas that need more methods

Intermediate sampling: between 10-40%

Plasmids and bacteria together

Variable sampling:

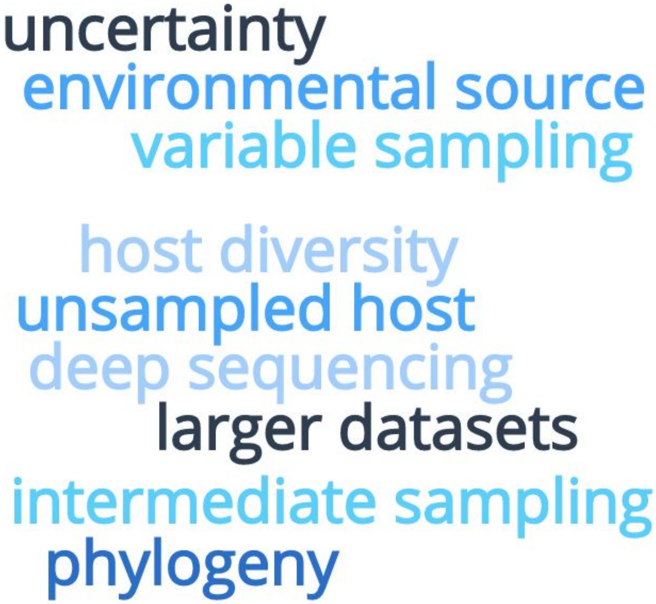
- over time (in principle ok in TransPhylo implemention underway)
- across a dataset

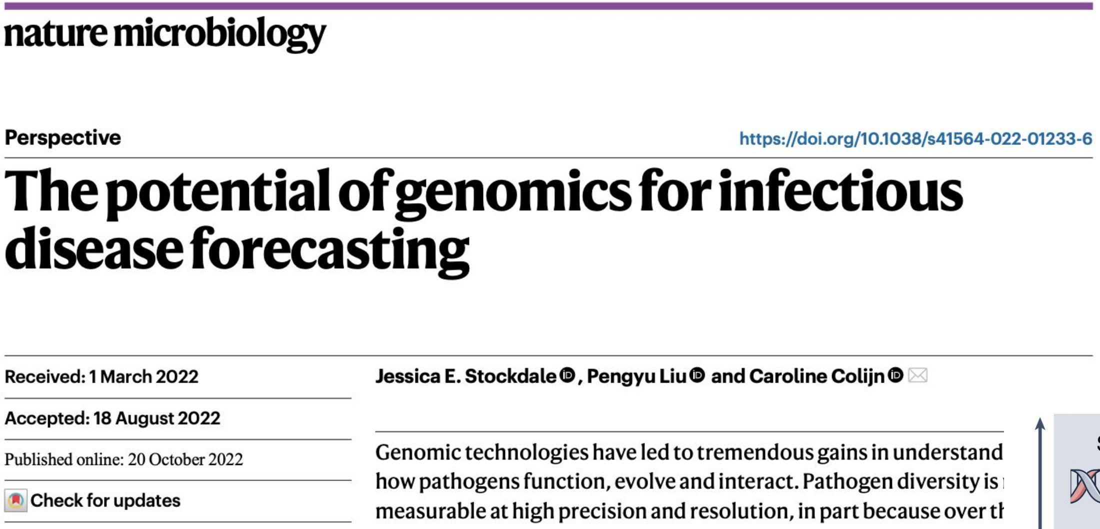
Reinfections *and* coinfections

Environmental transmission

Incorporate more epidemiological data

Connect to forecasting





nature microbiology

Perspective

https://doi.org/10.1038/s41564-022-01233-6

The potential of genomics for infectious disease forecasting

Received: 1 March 2022

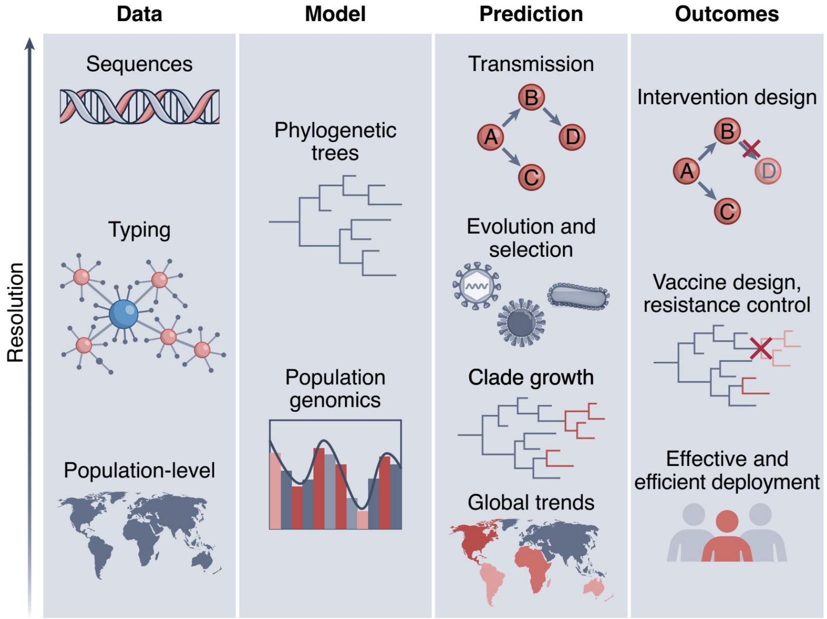
Accepted: 18 August 2022

Published online: 20 October 2022

Check for updates

Jessica E. Stockdale, Pengyu Liu and Caroline Colijn

Genomic technologies have led to tremendous gains in understand
how pathogens function, evolve and interact. Pathogen diversity is
measurable at high precision and resolution, in part because over th



Questions and discussion