Research forefronts and open questions

Reconstructing transmission with genomic data

Estimation of serial intervals using pathogen genomic data

with an application to COVID-19

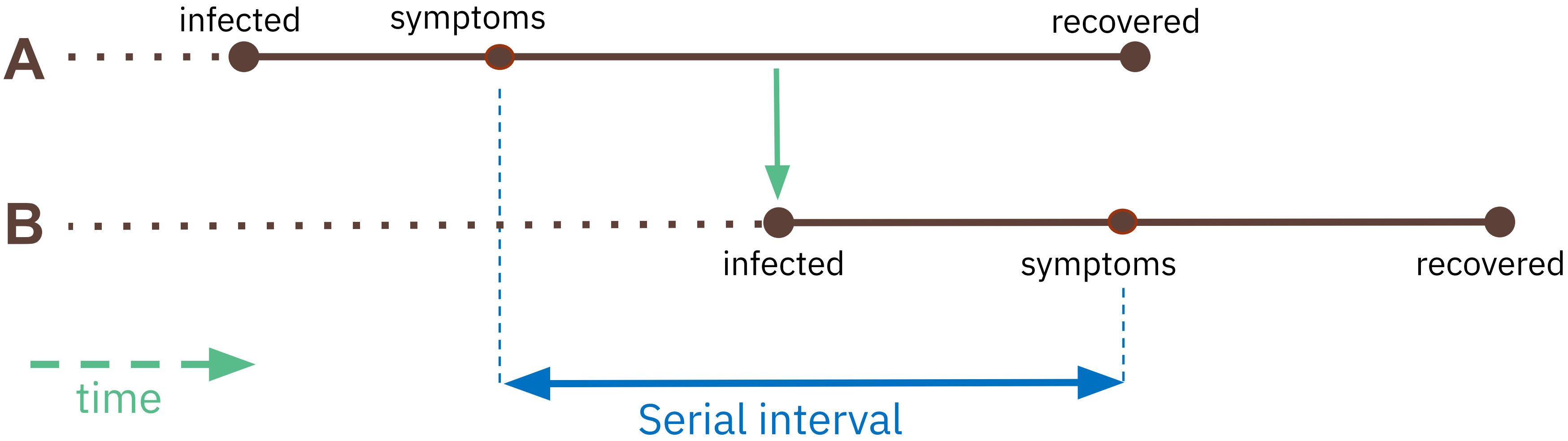
Sometimes, our data might not be sufficient to fully reconstruct the transmission tree

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?

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

□ Tells us about the speed of transmission...

□ ...this informs surveillance efforts.

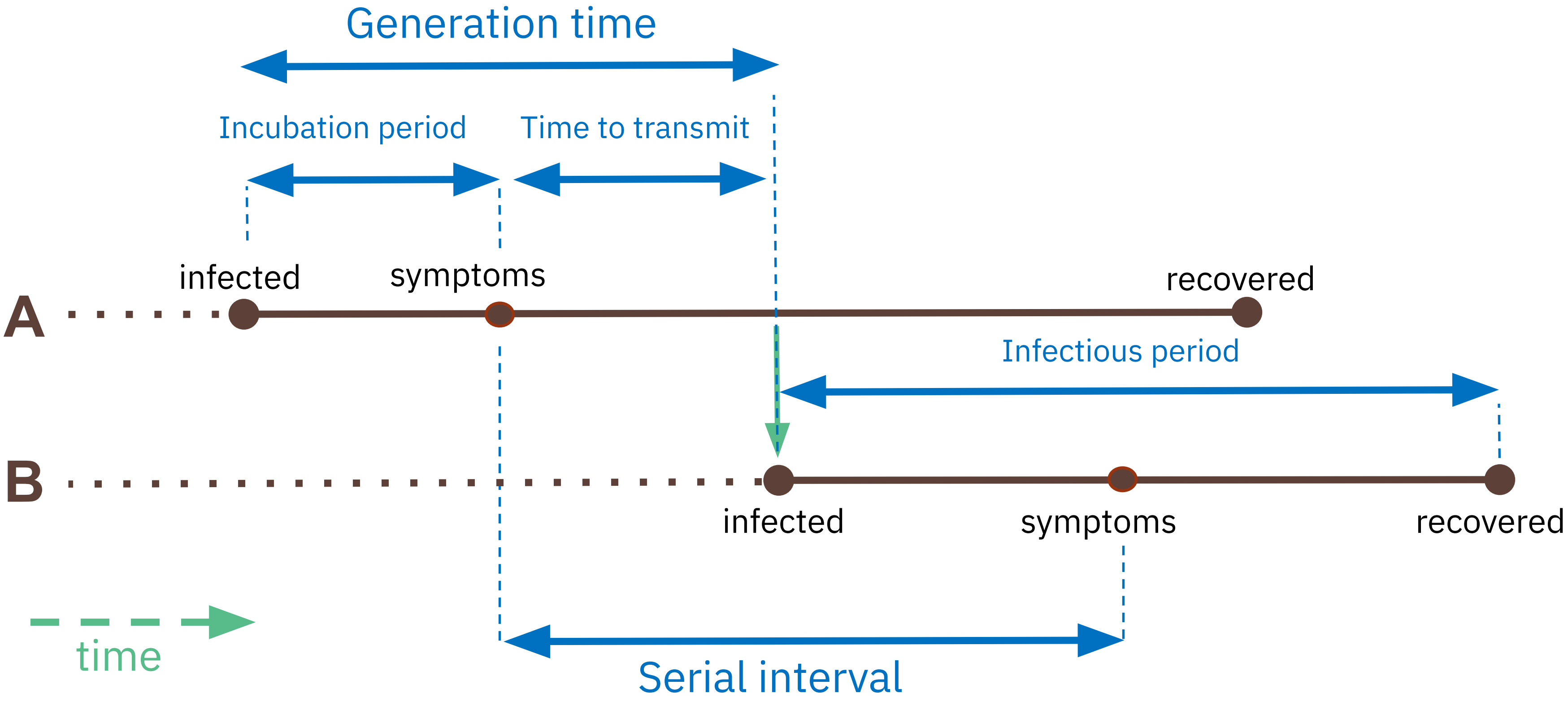
□ Used to calculate quantities like R$_0$, R$_t$ ...

$$R_0 \approx 1 + rS \longleftarrow$$ Epidemic exponential growth rate x serial interval *

□ ...and hence in understanding herd immunity thresholds and more.

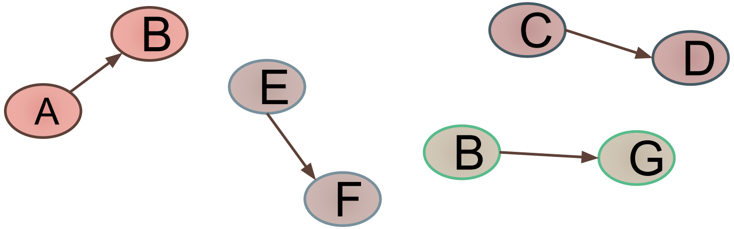
* How generation intervals shape the relationship between growth rates and reproductive numbers, Wallinga and Lipsitch (2006) *Proceedings of the Royal Society B*

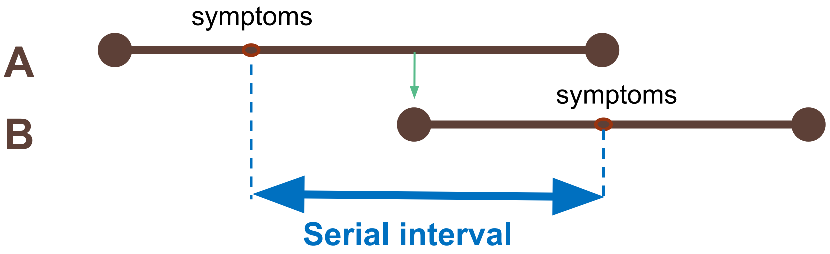
It’s also related to other disease intervals



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

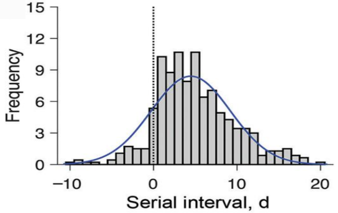
1. Contact trace pairs of cases which are assumed to represent direct transmission





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

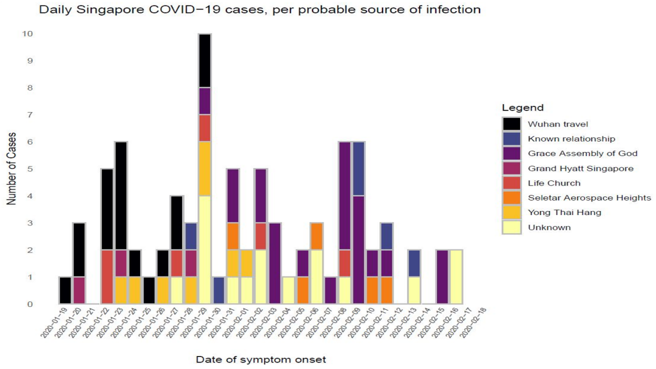
3. Parametric estimation of the serial interval given this observed data



Serial Interval of COVID-19 among Publicly Reported Confirmed Cases

Du et al. (2020) *Emerging Infectious Diseases*

How we got thinking about serial intervals…

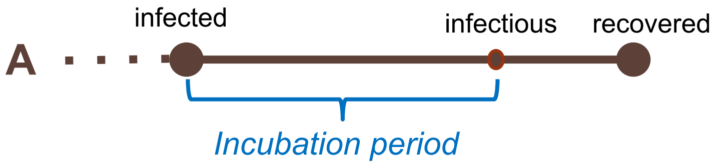


Evidence for transmission of COVID-19 prior to symptom onset. Tindale, Stockdale et al (2020) *eLife*

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

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

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



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

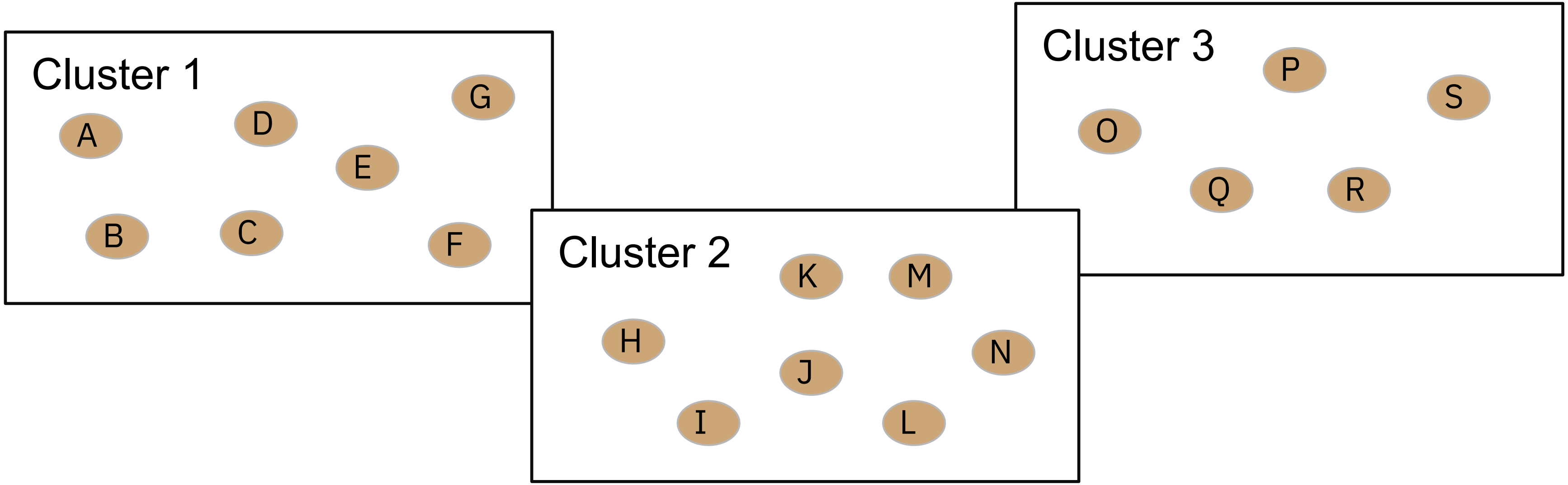
Contact tracing approaches require detailed personal data

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 under-sampling
- 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: instead of using contact data, 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

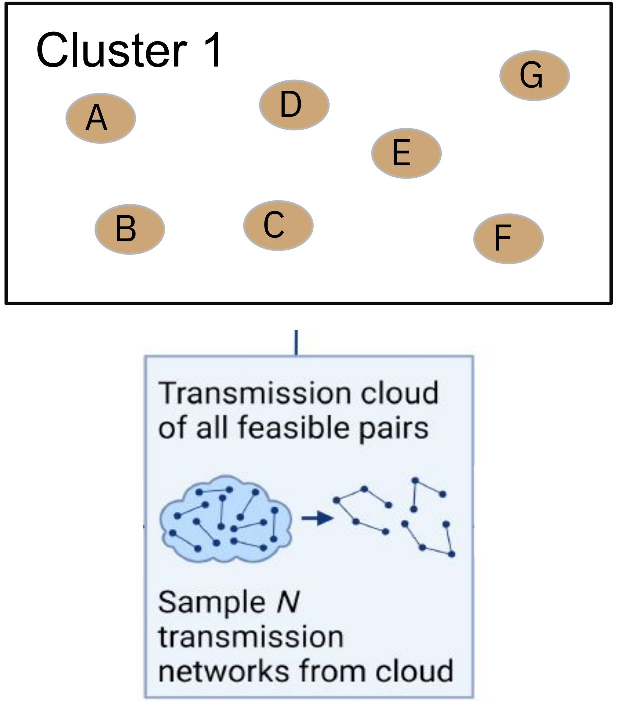
Take uncertainty in who infected whom into account, by sampling feasible transmission networks

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.

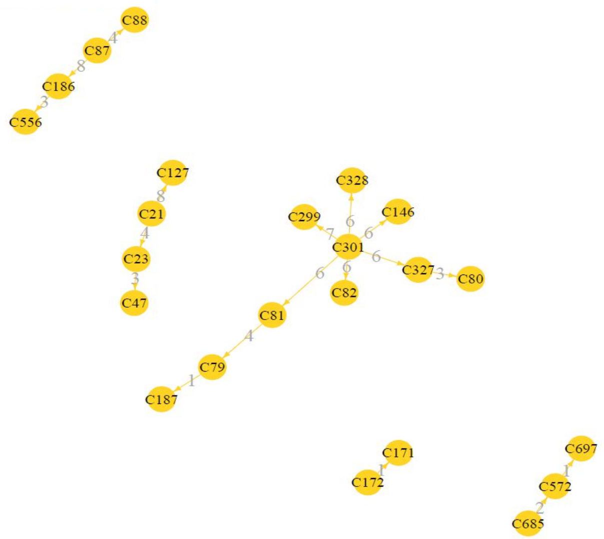


Transmission cloud of all feasible pairs

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**

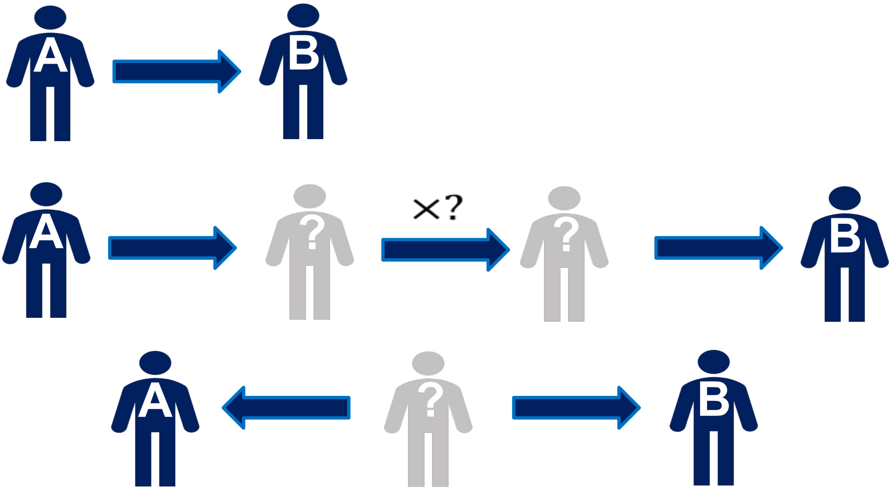
Estimate the serial interval distribution, taking indirect transmission into account

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:

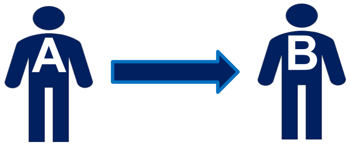
Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis
Vink et al. (2014)

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

Estimate the serial interval distribution: possible pathways

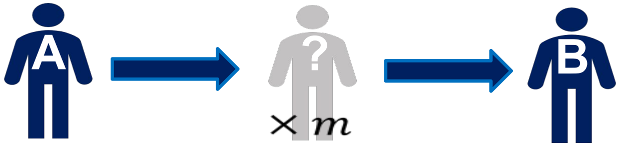
True serial interval distribution $\sim \Gamma(\mu, \sigma)$

Direct



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

Indirect



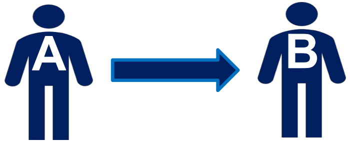
$T_{a,b} =$ sum of $m+1$ serial intervals, $m \sim \mathbf{Geo}(\pi)$
$\sim$Compound Geometric Gamma(μ, σ, π)

Proportion w of pairs

Estimate the serial interval distribution: possible pathways

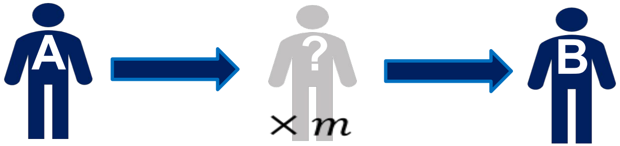
True serial interval distribution $\sim \Gamma(\mu, \sigma)$

Direct



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

Indirect



Probability of sampling an infectee

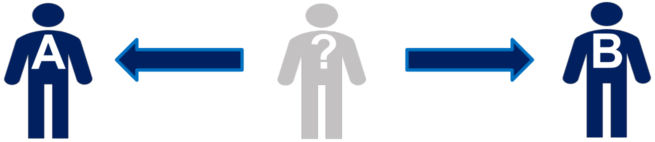
$T_{a,b}$ = sum of $m+1$ serial intervals, $m \sim \mathrm{Geo}(\pi)$
~Compound Geometric Gamma(μ, σ, π)

Proportion w of pairs

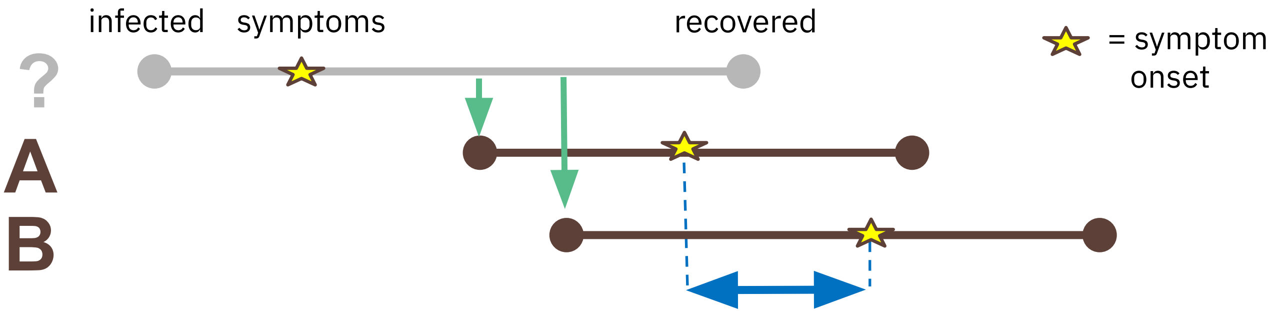
[*] A note on gamma difference distributions, Bernhard Klar, 2015

Estimate the serial interval distribution: possible pathways

Coprimary



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



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

Proportion 1-w of pairs

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

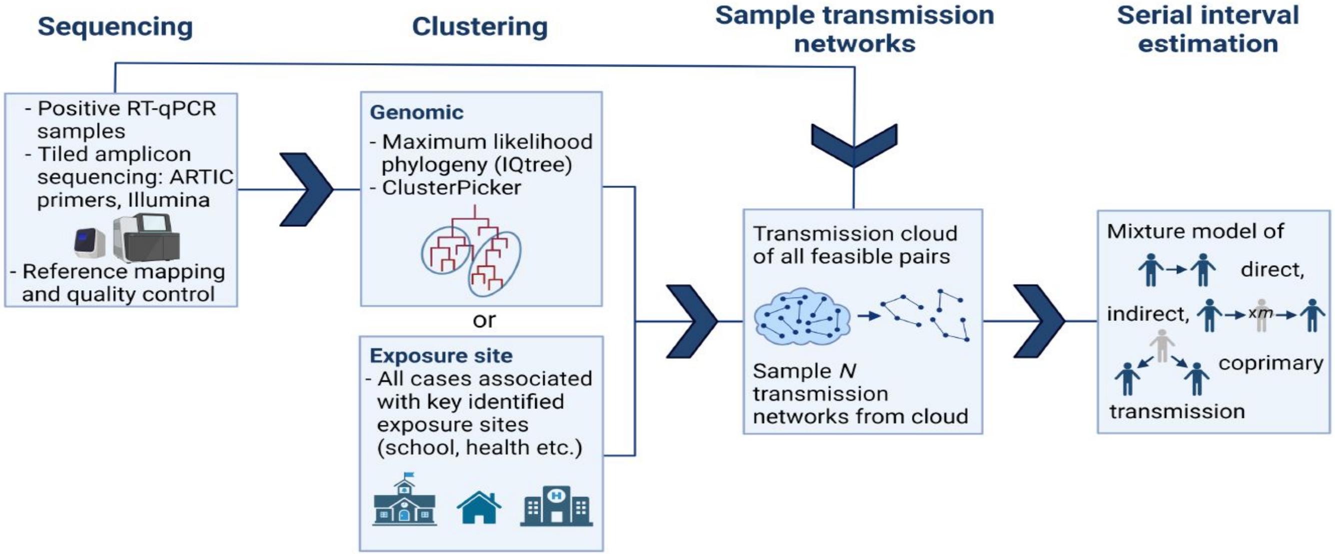
Instead of maximising this directly, we incorporate Beta distributed priors for *w* and *π*, and perform maximum a posteriori (MAP) estimation. We calculate the MAP for each sampled network, and then average over all networks

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

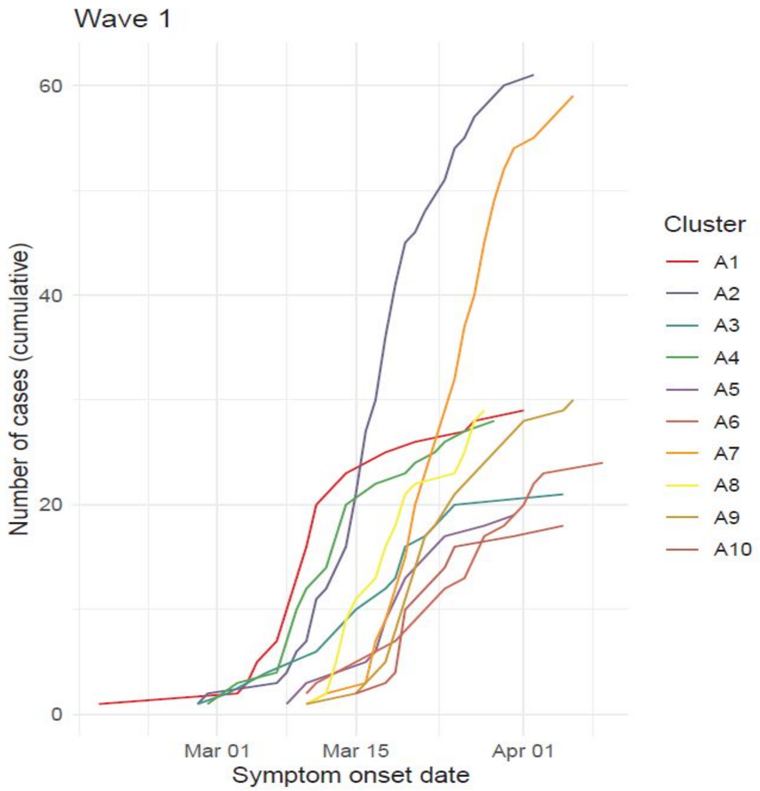
Transmission cloud of all feasible pairs

Sample *N* transmission networks from cloud

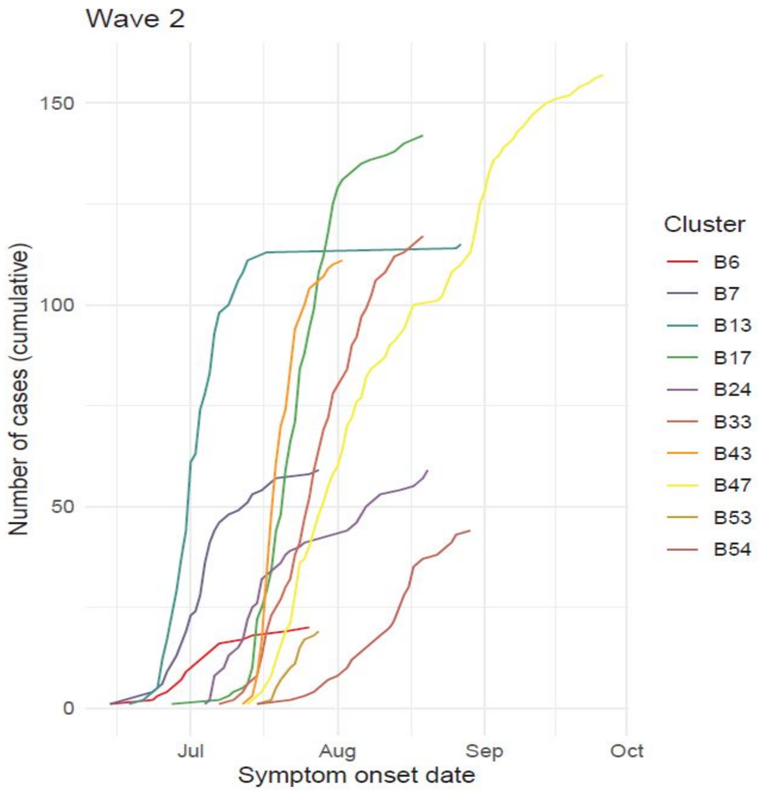
Serial interval estimation

Mixture model of direct, indirect, coprimary transmission

Application: COVID-19 clusters in Victoria, Australia

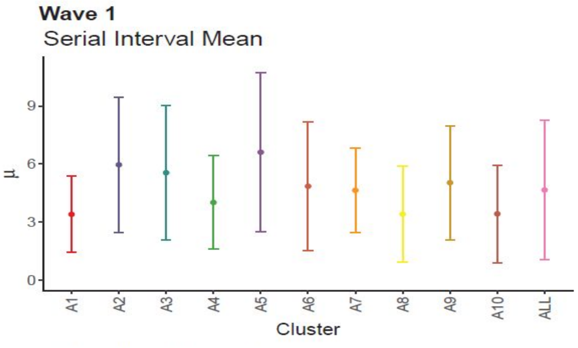

Wave 1

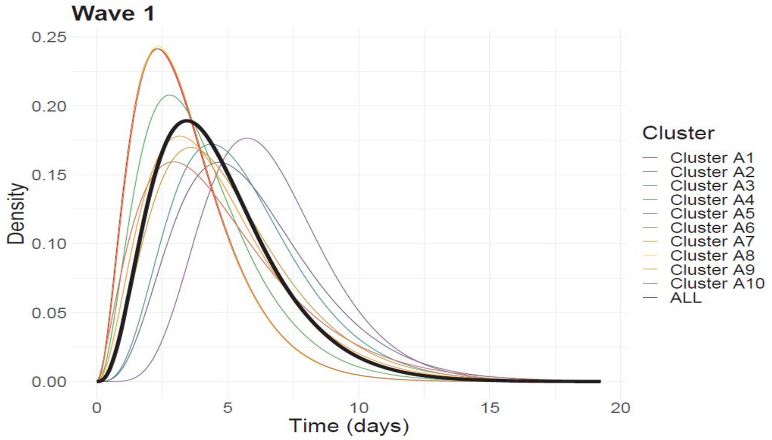
6 January–14 April 2020

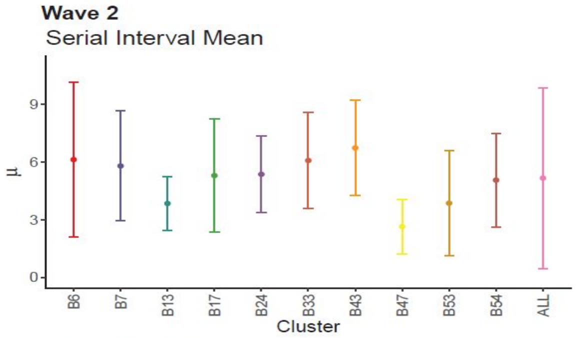

Wave 2

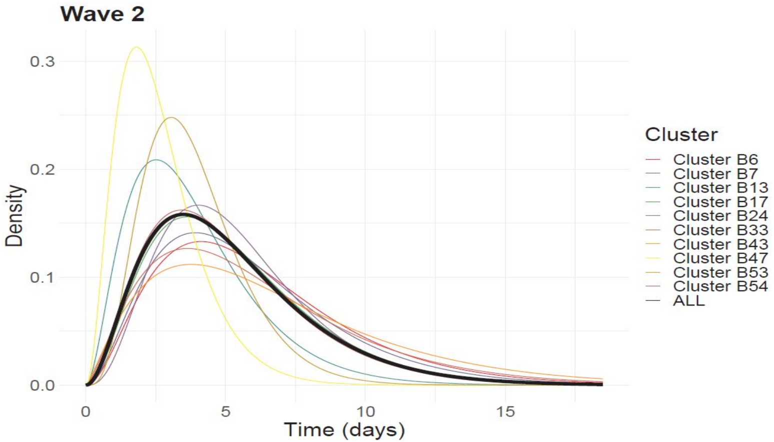
1 June–28 October 2020

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





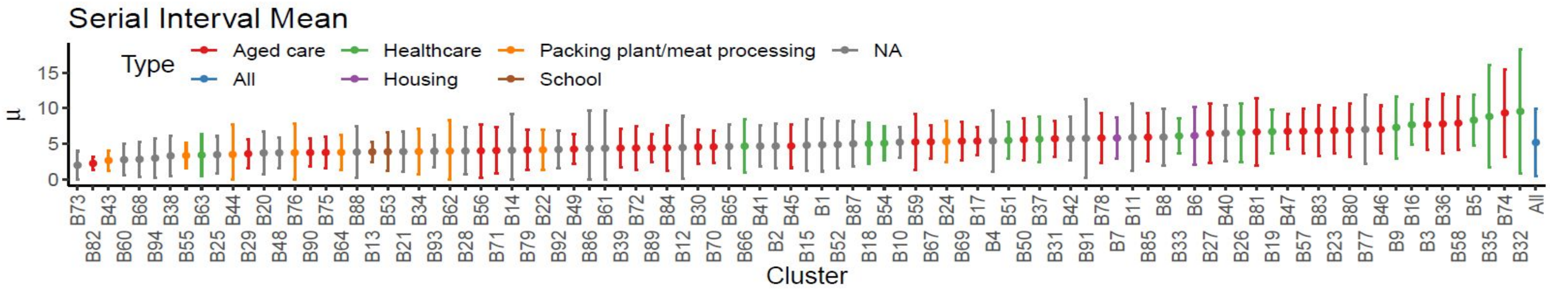




Context: Early published estimates ~5 days

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

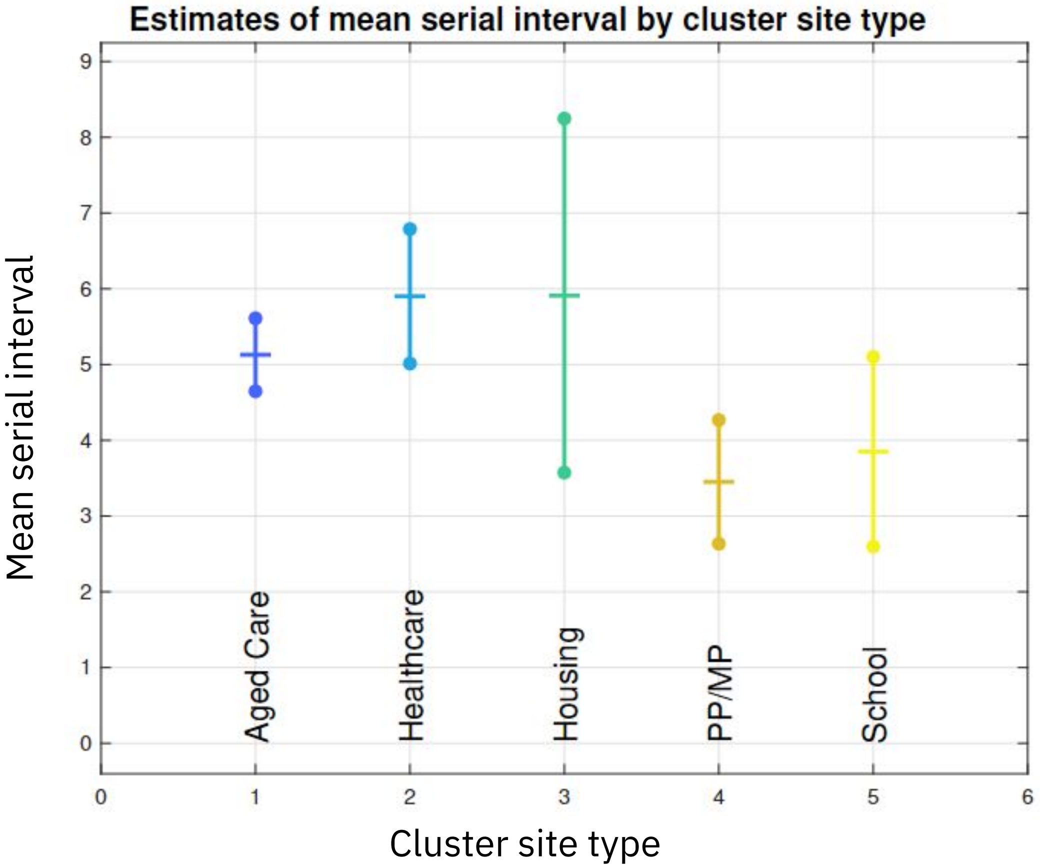
Estimates of the mean range from **2 to 9.5 days** (compared to standard estimates ~5 days)



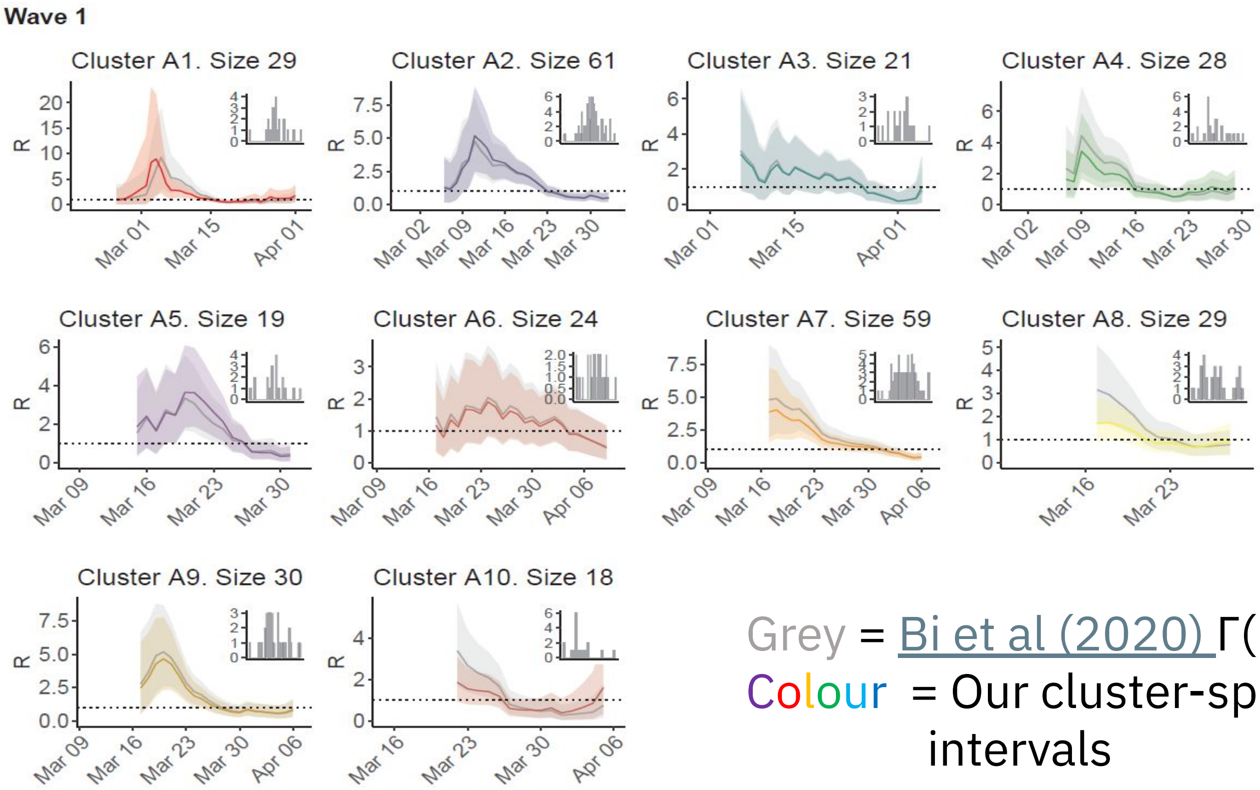
Among the shorter: packing plants, schools

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 impacted by the underlying serial interval distribution

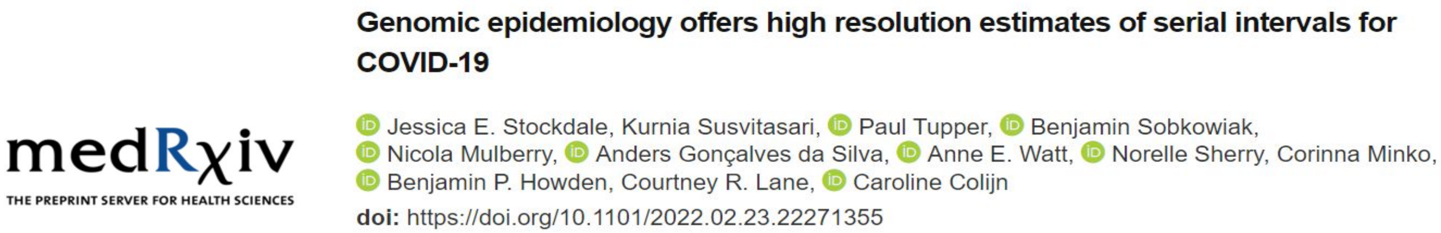


Grey = Bi et al (2020) Γ(μ = 6.3, σ = 4.2)

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)



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

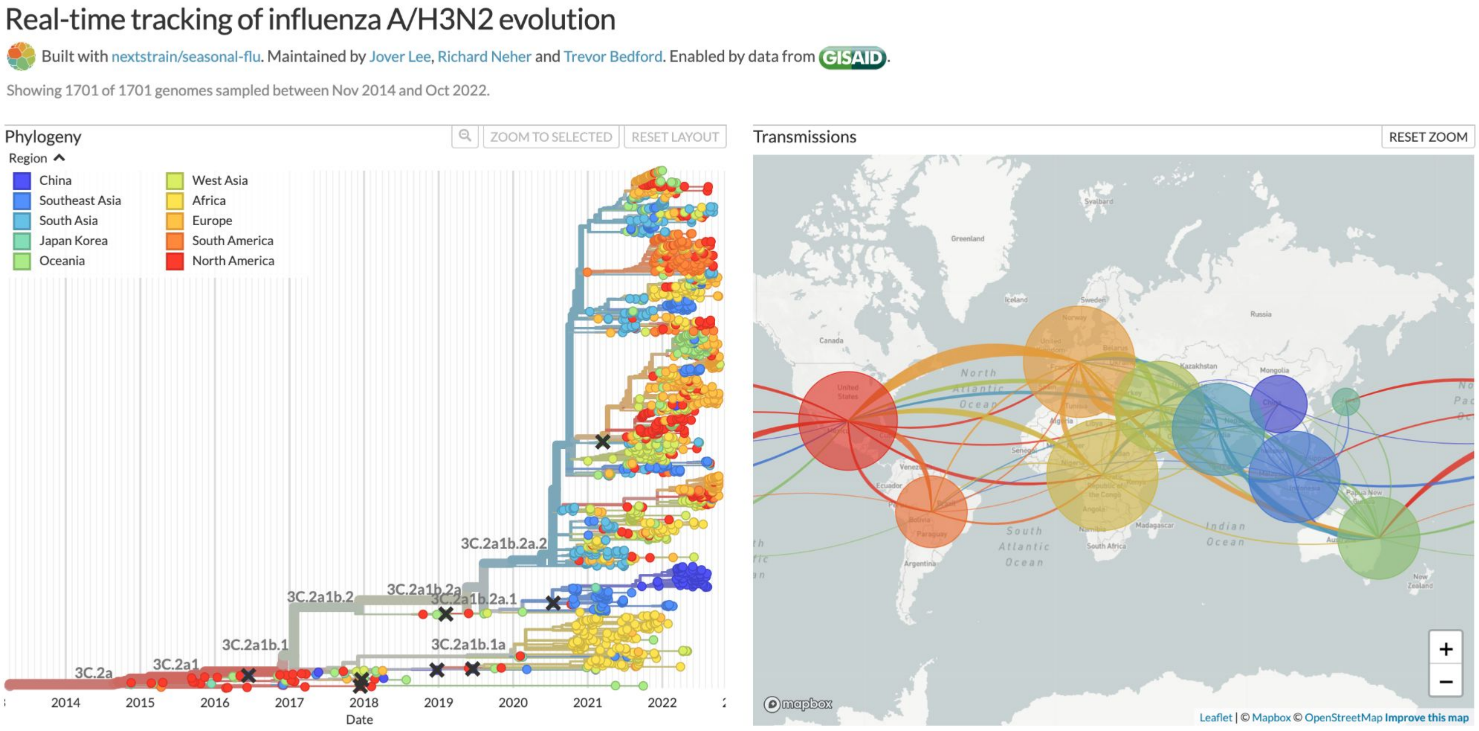
Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19

Jessica E. Stockdale, Kurnia Susvitasari, Paul Tupper, Benjamin Sobkowiak, Nicola Mulberry, Anders Gonçalves da Silva, Anne E. Watt, Norelle Sherry, Corinna Minko, Benjamin P. Howden, Courtney R. Lane, Caroline Colijn

doi: https://doi.org/10.1101/2022.02.23.22271355

https://www.medrxiv.org/content/10.1101/2022.02.23.22271355v1

VIRAL LOCATIONS IN A TREE AND THE WORLD



Real-time tracking of influenza A/H3N2 evolution

Built with nextstrain/seasonal-flu. Maintained by Jover Lee, Richard Neher and Trevor Bedford. Enabled by data from GISAID.

Showing 1701 of 1701 genomes sampled between Nov 2014 and Oct 2022.

Figure from nextstrain.org

This section: work with **Yexuan Song, Pengyu Liu, Ailene MacPherson**

How standard (discrete) phylogeography works

Phylogeography*:* Using phylogenetic trees to infer past locations of organisms, like viruses.

Model: Location is a discrete trait. It changes along the tree under a continuous time Markov chain (CTMC) model.

The CTMC has a rate matrix, *Q,* specifying the per-unit time rate of transitions between locations *i* and *j*.

This means we can calculate $P_{ij}(t)$: probability that location *i* transitioned to location *j* on a branch of length *t*.

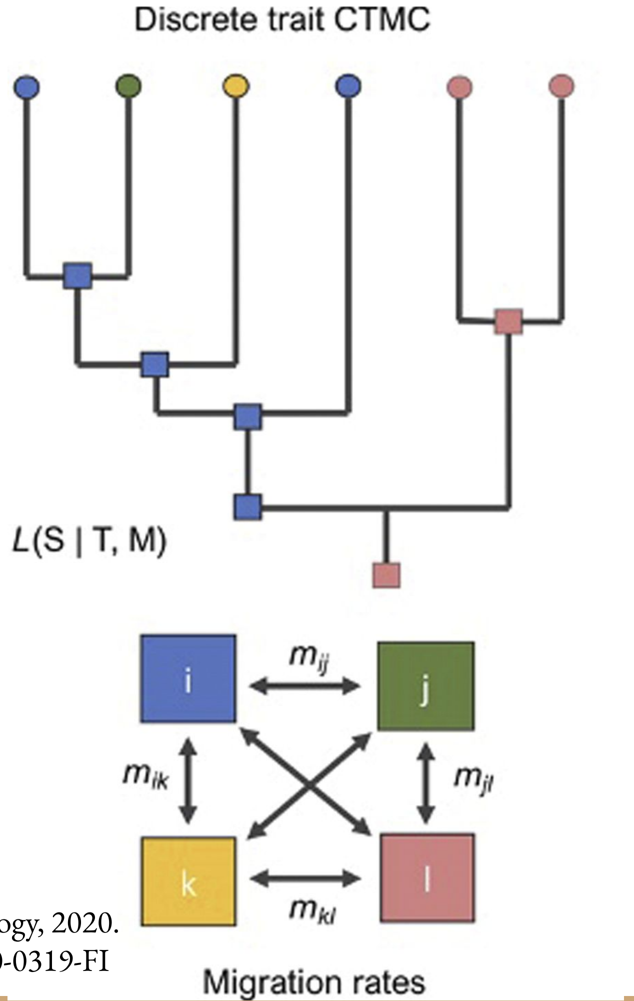
$P_{ij}(t)$ are used to assign locations to internal nodes, by maximizing the overall likelihood. *Q* can be estimated at the same time. (“Stochastic character mapping” -- Nielson, 2002).

Phylogeography illustration

Colour: location

Tip locations are observed.

Internal node locations are estimated with the CTMC.



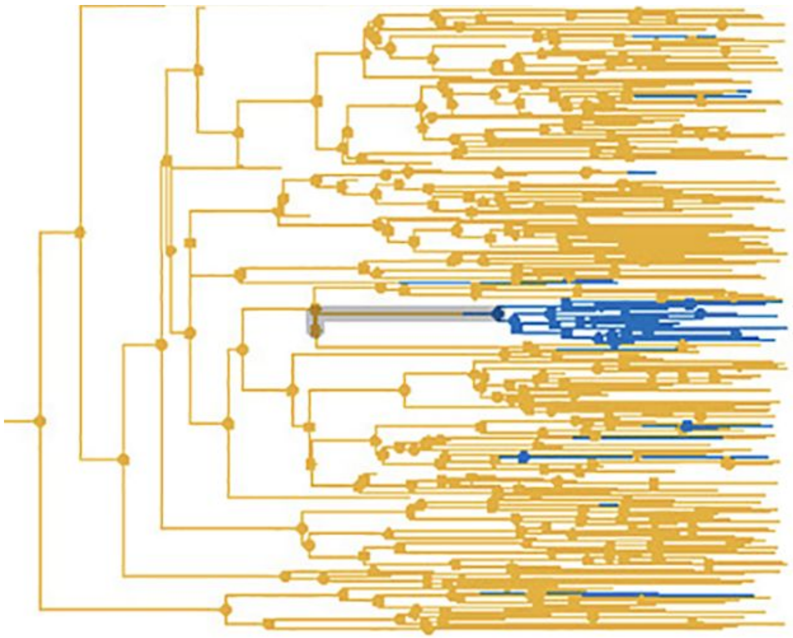
Rasmussen and Grunwald, Phytopathology, 2020.
https://doi.org/10.1094/PHYTO-07-20-0319-FI

Sampling: a challenge

What if some locations sample more than others? They get more tips.

This can impact the inference of where viruses were in the past.

Example: COVID-19. Some locations test more than others. Some locations have more resources to sequence the virus. Some share their data more, or less.



Two questions

1. How does sampling bias impact phylogeographic results?

2. How can we adjust for sampling rates to improve phylogeographic results?

A simulation study: how much does sampling impact phylogeography?

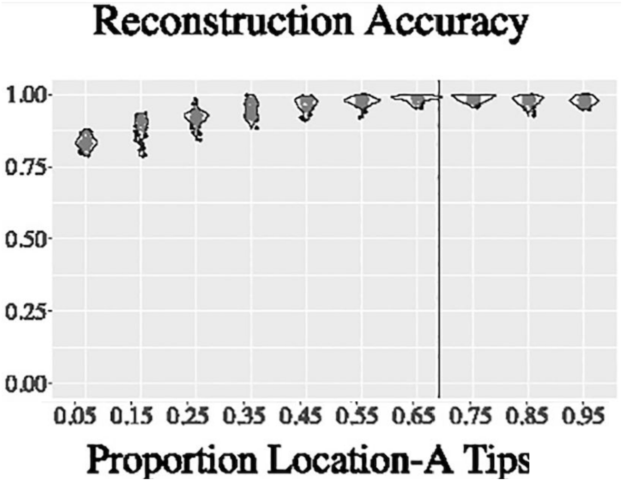
- We simulate two locations: yellow and blue. We know the true locations of all the nodes.
- We remove tips from the simulated trees to simulate different sampling fractions.
- We reconstruct the node locations using the standard CTMC.
- How wrong is the reconstruction, and in which ways?
- We examine an Ebola dataset: how much does sampling matter?

The impact of sampling bias on viral phylogeographic reconstruction. Pengyu Liu,Yexuan Song,Caroline Colijn,Ailene MacPherson. *PLOS Global Public Health,* 2022

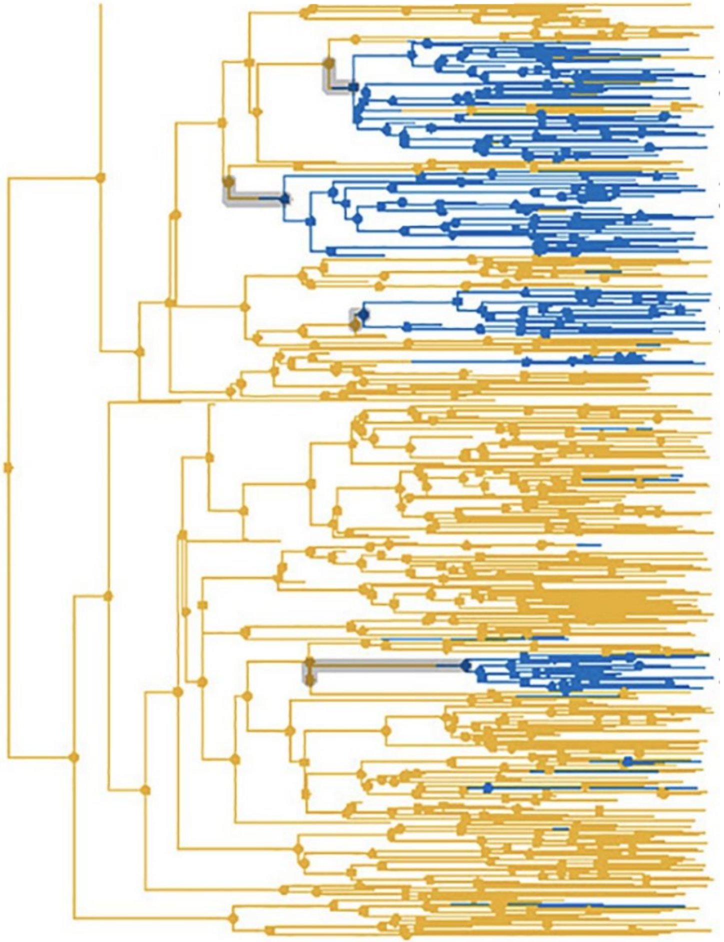
PLOS GLOBAL PUBLIC HEALTH

Low migration rate

When transitions are rare, the colours are “grouped” in the tree. The overall accuracy is high, and does not depend much on the sampling bias.



True tree

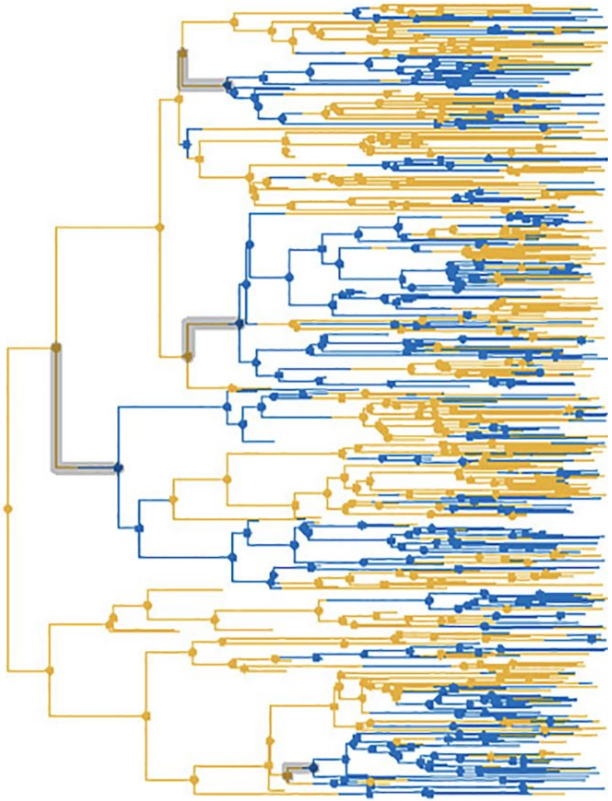


High migration rate

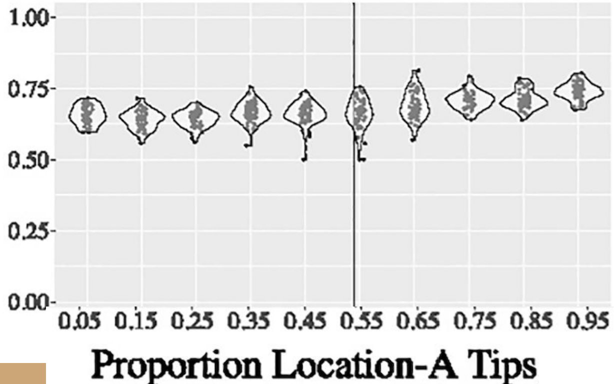
If the transitions are frequent, the tip locations don’t carry as much information about the internal node locations.

The accuracy is worse, but it does not depend strongly on sampling.

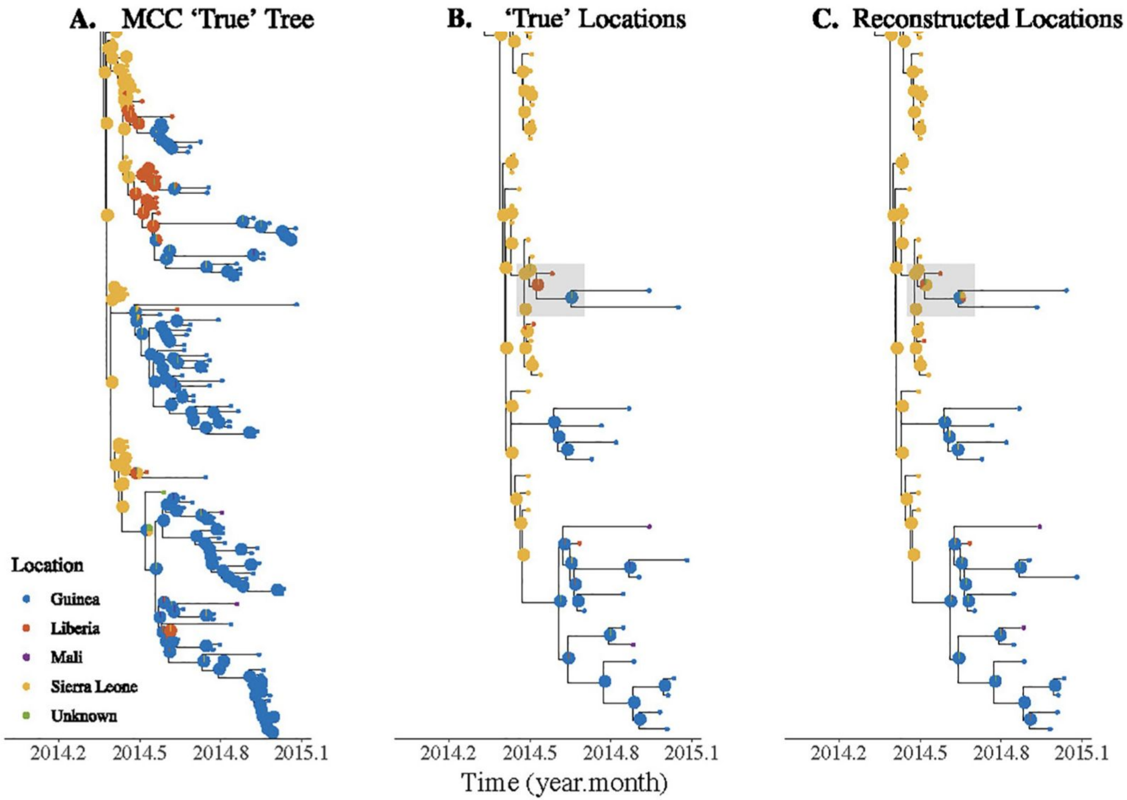
True tree



Reconstruction accuracy



Sampling can affect whether we detect "key migration events"



Simulation results

1. Standard methods did a good job with overall accuracy, which didn’t depend much on the sampling bias
2. Low migration rate: overall accuracy is really high at whatever sampling bias
3. High migration rate: accuracy is lower, but not very dependent on sampling bias
4. **However,** you can really get key importation events wrong.

How can we adjust for sampling bias?

We need a new mathematical model.

Background: Binary state-dependent speciation and extinction (BiSSE) family of models (Maddison, Midford, Otto. *Systematic Biology* 2007)

Underlying process: multi-type branching process.

Each “state” (here, location; general: value of a trait) has its own branching rate and death rate, which can be estimated from data.

Y. Song MSc: We combined two advances: (1) extension to incomplete observation (Fitzjohn et al); (2) description of how to estimate node locations in the model (Freyman & Hohna).

This section: work with **Yexuan Song, Ailene MacPherson**

The mathematical method

Key ingredients:

- D_{BNi} - probability that a lineage *N* in state *i* at time *t* gives the observed descendants.
- D_{FNi} - probability that a lineage *N* in state *i* at time *t* arose from the observed ancestor.
- E_{Ni}- probability that a lineate *N* in state *i* at time *t* dies out (not observed by the present)

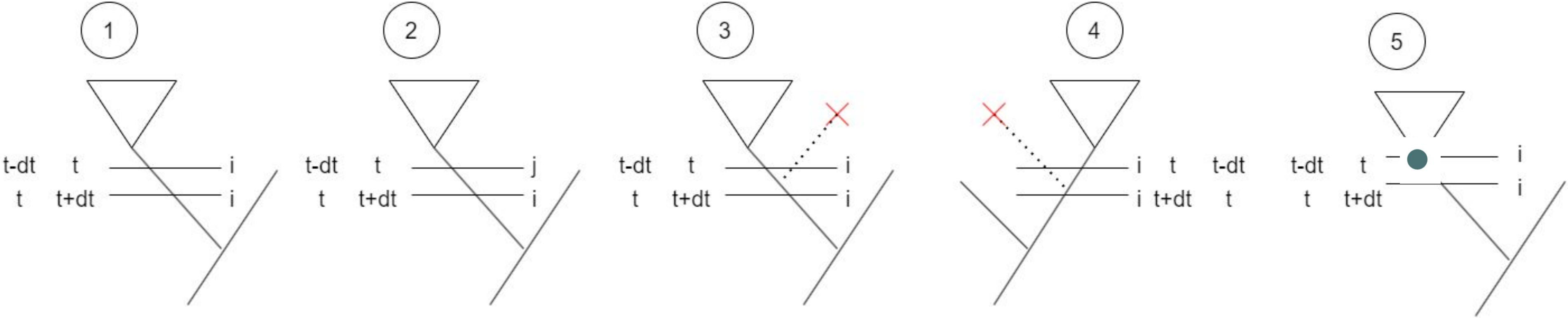
Approach: derive differential equations for these terms.

Use the *Ds* to assign the states at the internal nodes.

Differential equations for the *D* terms

There are 4 possibilities:

- 1. No state change, no speciation, no sampling.
- 2. There is a stage change but no speciation or sampling.
- 3, 4. Speciation: only the left or right lineage survives.
- 5. The lineage gets sampled at time t.



$$
\begin{aligned}
D_{BNi}(t+\Delta t) \approx D_{BNi}(t) & \\
& + (1-\mu\Delta t)[(1-q_{ij}\Delta t)(1-\lambda\Delta t)D_{BNi}(t) \quad \text{no change} \\
& + q_{ij}\Delta t(1-\lambda\Delta t)D_{BNj}(t) \quad \text{state change} \\
& + 2(1-q_{ij}\Delta t)\lambda\Delta t E_i(t)D_{BNi}(t)] \quad \text{Speciation; one goes extinct} \\
& + \mu\Delta t(0) + O(\Delta t^2) \quad \text{Extinction}
\end{aligned}
$$

$$
\frac{d}{dt}D_{BNi}(t) = -(\lambda+\mu+q_{ij})D_{BNi}(t) + 2\lambda E_i D_{BNi}(t) + q_{ij}D_{BNj}(t)
$$

where λ: branching rate; μ: death rate; f_i: location-specific sampling fraction; q_{ij}: transition rate.

Initial conditions:

Tips: $D_{BNi}(0) = f_i$ if the tip is in state i. Internal nodes: probability of giving rise to clades C_1 and C_2 is the product of probabilities D.

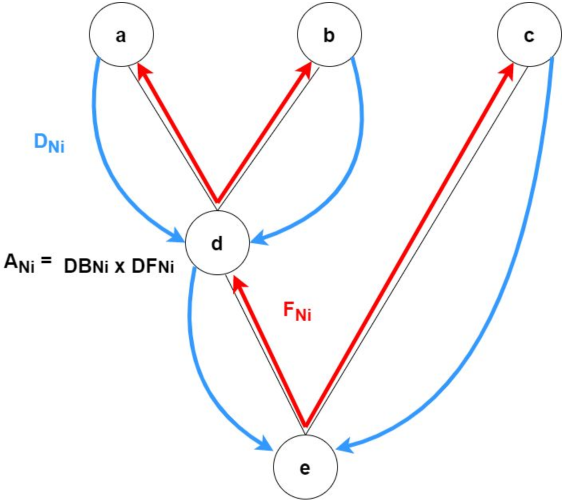
Method overview

Take the same approach for forward equations for $D_{FNi.}$ and for the extinction probability $E_{i.}$

Use Freyman and Hohna's approach to "stochastic character mapping"-- assigning states to the internal nodes:

A_{Ni} is the probability that a node is in state i: $A_{Ni} = D_{FNi}D_{BNi}$

Assign the max-probability state to each internal node.



Yexuan Song's MSc thesis

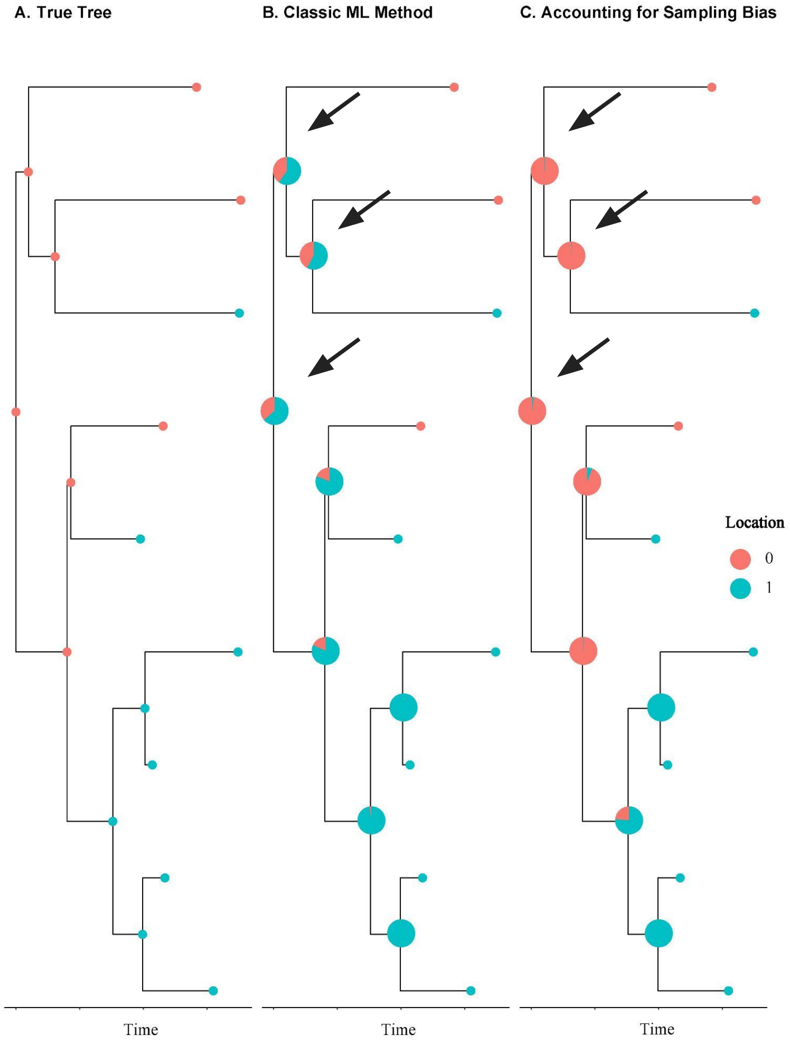
Results: better than standard

Simulate: blue location sampled 1.7x more than the red location.

Standard phylogeography over-estimates the number of blue nodes.

The new method gets the right locations for the red internal nodes.

This is a proof of principle: working on large-scale implementation and testing.



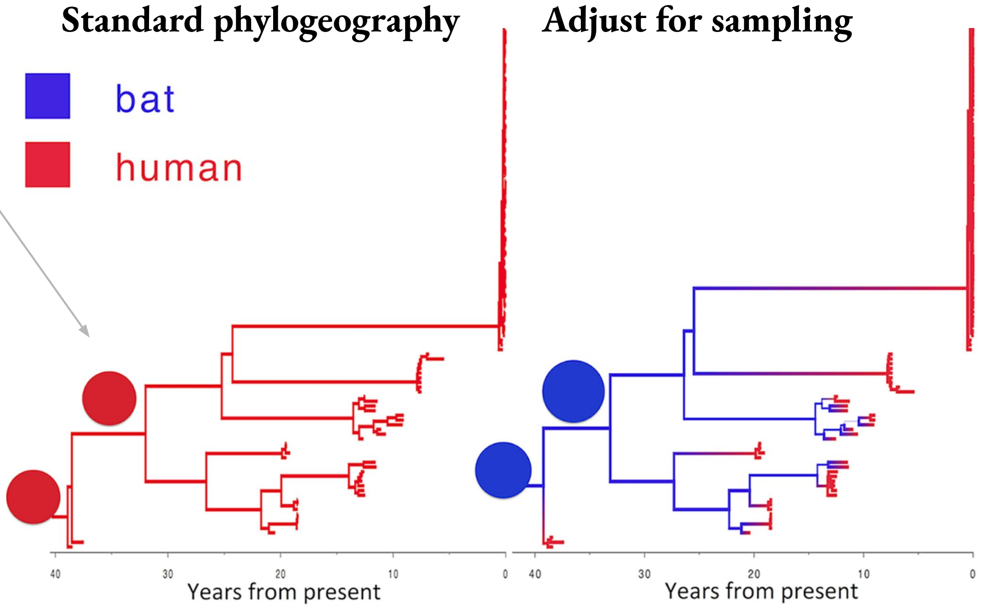
Intuitively, why do all these differential equations help?

Standard phylogeography: humans all the way back.

This doesn’t account for the fact that if it *had* been in humans, it would have been sampled over all those years.

Adjusting for sampling fraction: few observations means higher likelihood that it was in the bats.

Location: which animal is the Ebola virus in? Bats or humans?



Shamelessly taken from: de Maio et al, *New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation.* PLOS Genetics, 2015

3. 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	✖	✅	✖ (in progress?)	✖	✖
SCOTTI	✅ (limited)	phylogeny	✅	✅	✅	✅(limited)	✖

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*, bioRxiv (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:

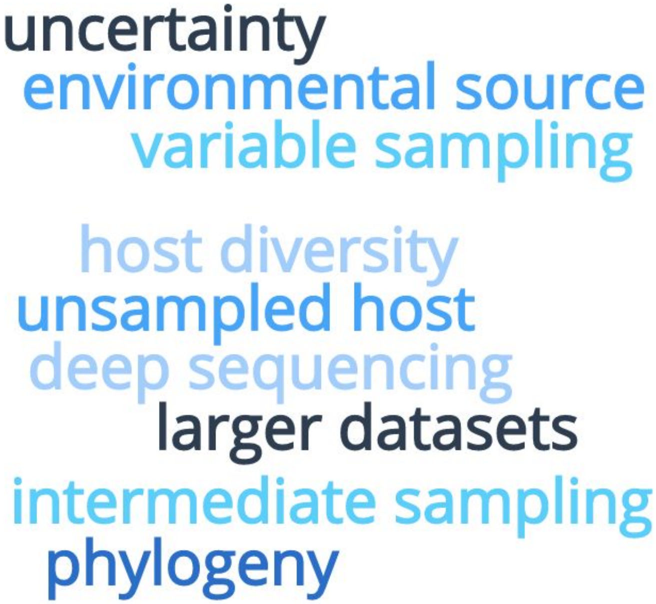
- over time (in principle ok in TransPhylo implemention underway)
- across a dataset

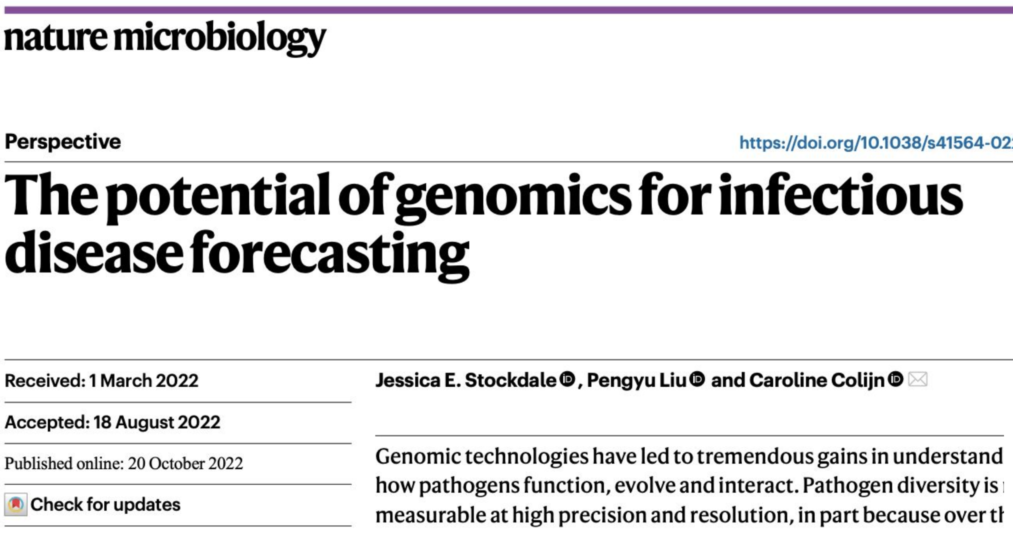
Reinfections *and* coinfections

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

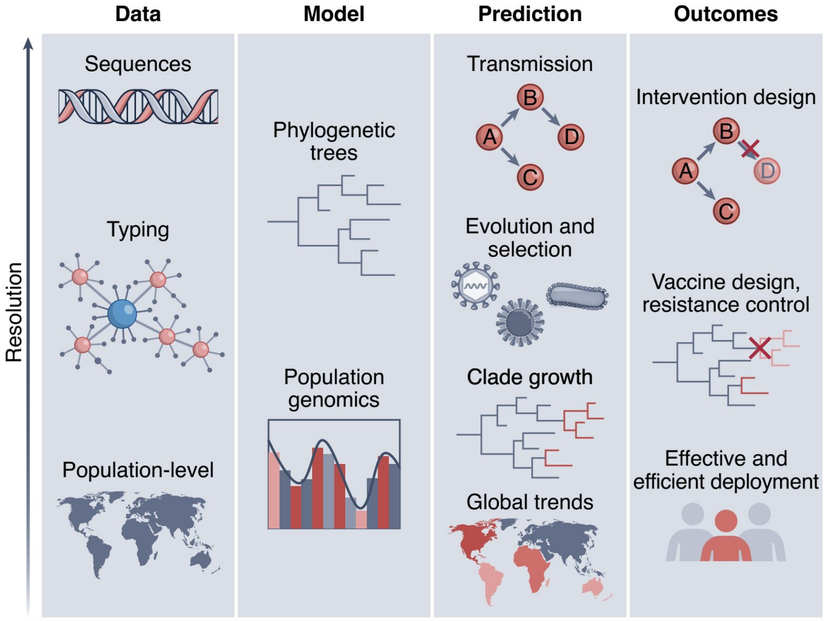
Accepted: 18 August 2022

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



4. Remaining questions and discussion