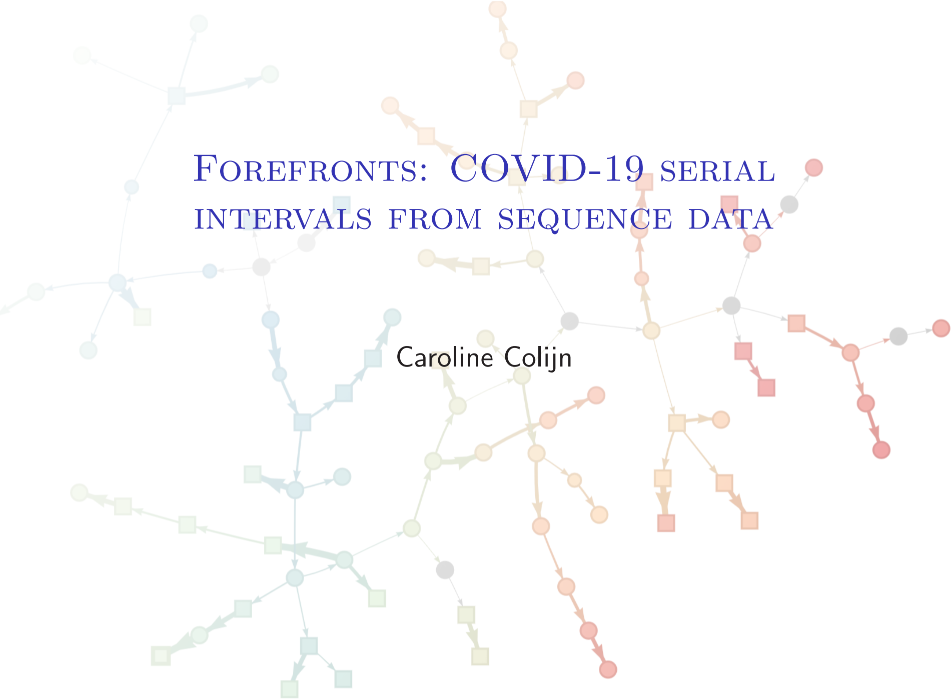
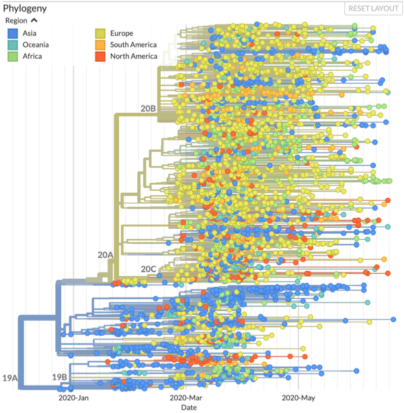


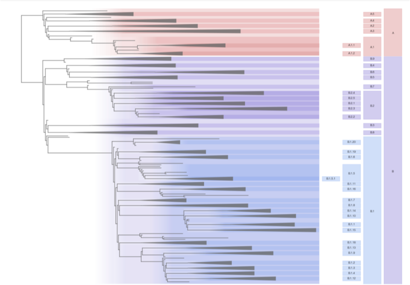
Forefronts: COVID-19 serial intervals from sequence data

Caroline Colijn

SARS-CoV-2 sequences global scale



SARS-CoV-2 lineages: PANGOLIN



Rambaut et al: Nomenclature proposal

https://www.biorxiv.org/content/10.1101/2020.04.17.046086v1

Three roles for sequences in immediate public health response

- Infer global routes of movement
- Infer population dynamics back through time
- Infer R0 is the mean number of new infections per case in a susceptible population. Phylodynamic methods provide coarse estimates of R0
- Group local cases into clusters. People with similar virus sequences might have infected each other

All of these have limited usefulness for immediate public health action. I think the last is the best. Recall introduction talk: Seemann et al.

Vision for better use of virus genomes in public health

- Connect sequences to epidemiological data
  - Time of symptom onset
  - Source of transmission (household, workplace, community, gathering)
- Connect patterns of transmission to patterns of genetic variation
  - Chains of undetected transmission look different than transmission at a mass gathering
- Together: high-resolution local pictures of transmission

This is an active project with funding from the Michael Smith Foundation for Health Research: operationalize this in BC.

The serial interval

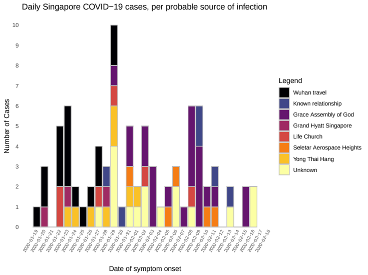
Definition: Time between symptom onset in the infector and the infectee

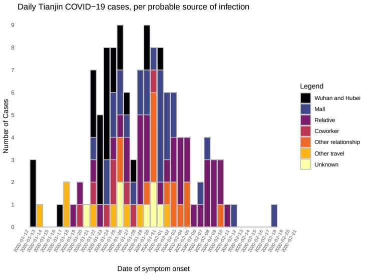
- Fundamental to understanding the epidemiology of an infectious disease
- Used in all those Rt estimates you see on twitter
- Used in estimating R0 from case data
- R0 in turn: epidemic size, portion to vaccinate, all modelling work
- Widely used in public health planning

How is the serial interval estimated?

- Time between symptom onset in A and B, where A infected B
- Requires knowing many A, B pairs where A infected B
- Relies on contact tracing, data availability
- Challenges that methods try to address
  - Unsampled / unknown cases: A $\rightarrow$ X $\rightarrow$ B instead of A $\rightarrow$ B
  - Bias: easier to find pairs where the symptom onsets were close together
  - Bias: easier to find pairs where people know each other
  - Right truncation: we only see those whove had symptom onset already

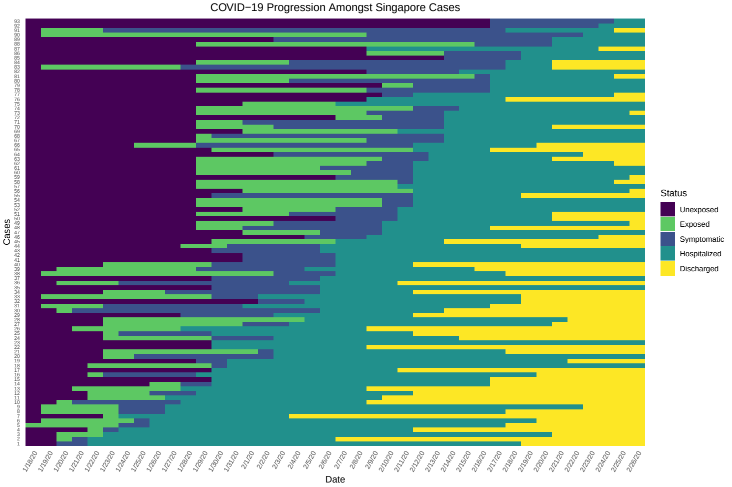
Estimating the serial interval with conventional data



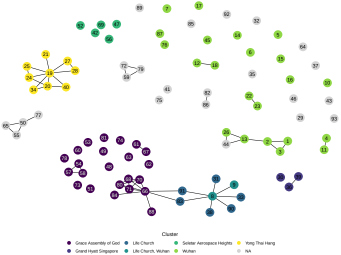


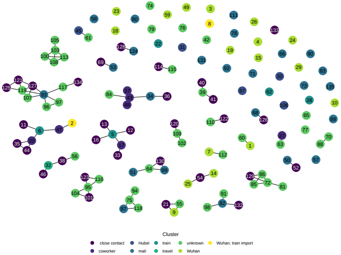
Two cities, Singapore and Tianjin, China, made contact data – who was exposed to whom and when – publicly available early on.

Individual case timing



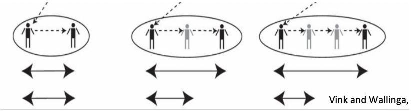
Contact data: a network visualization





Serial interval estimates

These data are hard-won. We used them to estimate the serial interval. Vink and Wallinga introduced a mixture model to account for missing intermediates.



Vink and Wallinga, AJE, 2014

Direct transmission One intermediate More intermediates

Result: The *mean* COVID-19 serial interval was 4.17 (95%CI 2.44, 5.89) in Singapore, and 4.31 (95%CI 2.91, 5.72) days in Tianjin.

Tindale, Stockdale et al, eLife 2020

https://elifesciences.org/articles/57149

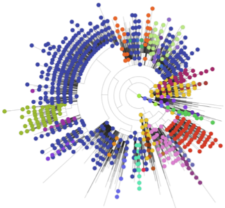
Higher-resolution serial intervals: use sequences!

A method for estimating serial intervals in local outbreaks

1. Obtain sequences and symptom onset dates
   - Challenging: kept siloed for privacy, and because of expertise differences
   - For now, use sample collection dates as a proxy
   - Getting set up in BC to join symptom onset dates and genomics
2. Group sequences into clusters – we use a SNP cutoff but could do more
3. Sample large numbers of possible transmission trees
4. Estimate serial intervals in each cluster, accounting for intermediates

Step 1: the data

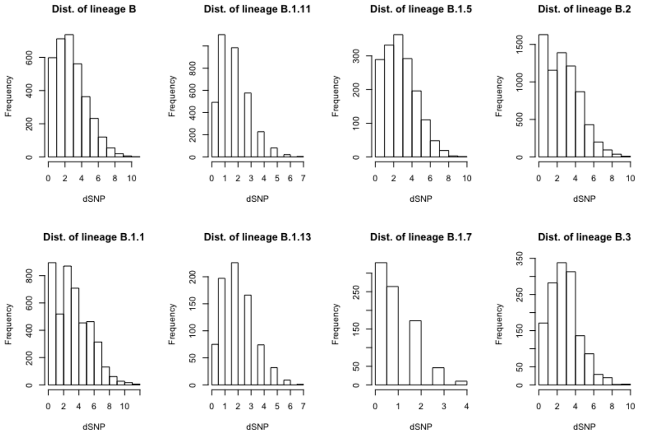
- COVID-19 genomics UK consortium (COG-UK)
- We use several lineages in the first 10,000 genomes from the COG-UK Consortium (also in GISAID)
- You can explore them in microreact: https://microreact.org/project/cogconsortium-2020-06-26/b2735aed/
- As far as we know, sampling was not lineage-specific
- This is demonstrative only.
- We use sample collection date in place of symptom onset



The phylogenetic tree indicates the position of the UK genomes within a global context.

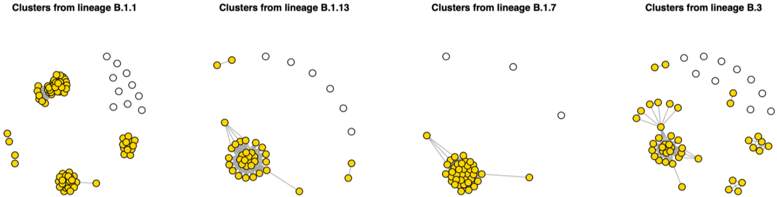
The data, continued

We chose several defined lineages for which there were reasonable numbers of samples in the cog consortium data. Genetic diversity is quite low.



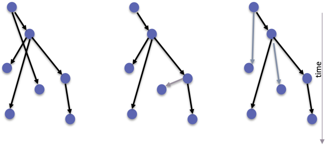
Step 2: group into clusters

We use a SNP cutoff (group two into the same cluster if there are fewer than 3 single nucleotide polymorphisms between the two)


Clusters from lineage B.1.1
Clusters from lineage B.1.13
Clusters from lineage B.1.7
Clusters from lineage B.3

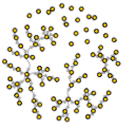
Step 3: Sample large numbers of transmission trees

- We sample transmission trees within the clusters
- Each individual gets an infector
- The infector *has to have had symptoms before the infectee*
- There are many ways to choose!
- We choose many. We use sequences to help choose.

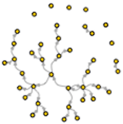


Examples of sampled transmission trees

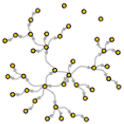
TTree from lineage B.1.1



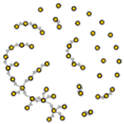
TTree from lineage B.1.13



TTree from lineage B.1.7



TTree from lineage B.3



Step 4: estimate serial intervals in each cluster

We assume a parametric function for the serial interval:

$$\tau \sim Gamma(a, b)$$

where a and b are the shape and scale parameters.

Then we use a gamma mixture model to allow for the unknown intermediates.

- If τ has density $Gamma(a, b)$ then $\tau_1 + \tau_2$ has density $Gamma(2a, b)$).
- If τ has density $Gamma(a, b)$ then $\tau_1 + \tau_2 + \tau_3$ has density $Gamma(3a, b)$).

Mixture model

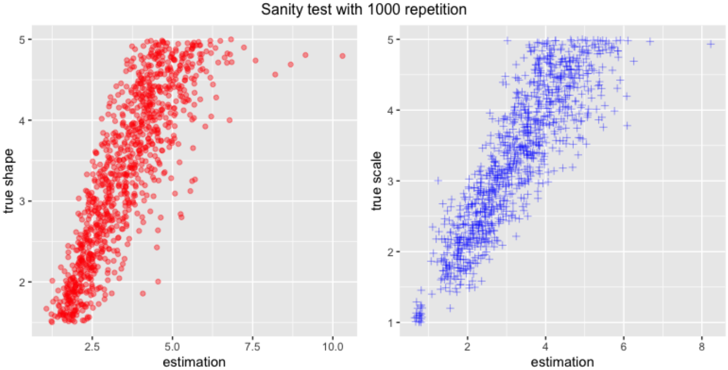
The likelihood of an observed interval has to take into account that the interval might be a “true” sample of the serial interval, or it might be a few of them together, because of the unknown intermediates.

$$\mathbf{L}(\text{interval}) = w_1 \mathit{Gamma}(a, b) + w_2 \mathit{Gamma}(2a, b) + w_3 \mathit{Gamma}(3a, b)$$

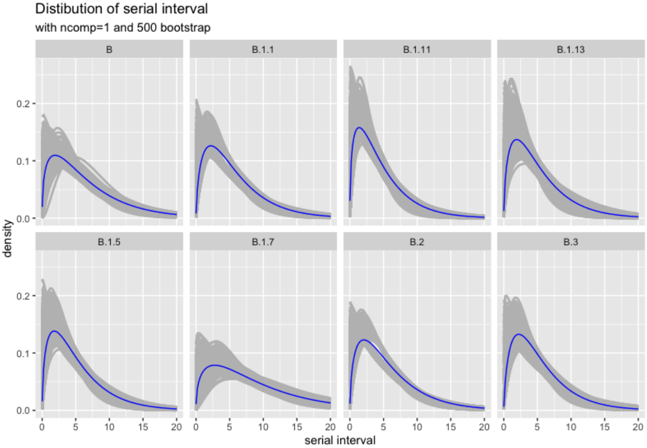
We use an expectation-maximization algorithm.

It estimates the weights given a, b, then a, b given the weights, and so on.

Performance with simulated data

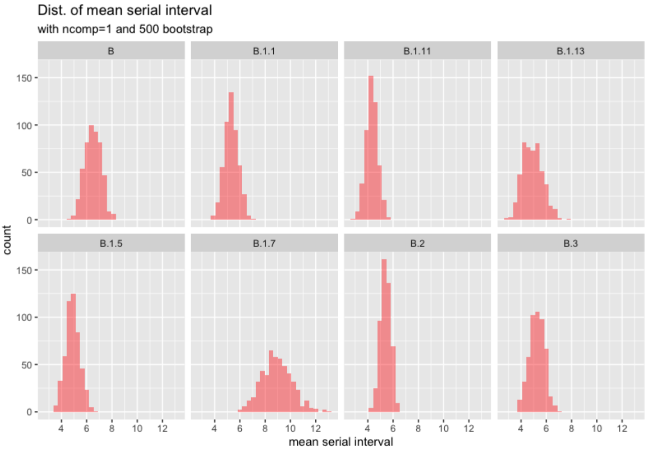


Serial interval distributions by lineage

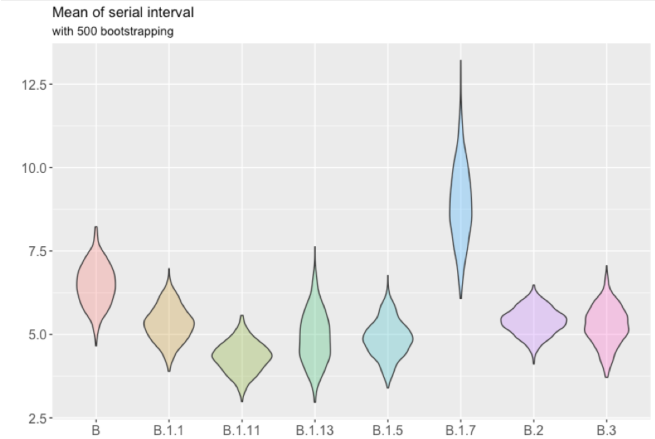


Grey: 500 bootstrap replicates. We used 3 mixture components

Mean serial intervals: some more uncertain than others



Some lineages have longer serial intervals than others



Take home messages

- Serial interval estimates differ by lineage. If true, then:
- Estimates of the time-dependent reproductive number Rt should account for this
- Estimates of the basic reproductive number R0 should account for this

This is potentially a higher-resolution view of transmission dynamics that can inform public health

Discussion points

- The result could be confirmed or refuted with line list data if these could be linked to viral sequences or at least to lineage
- Sampling effects, local epidemiology probably play a role
- I would not claim now that it is evolution!
- Limitation that we used collection date instead of onset date
- unknown true component number
- This should not affect our lineage comparisons, even if the delay between onset and sample collection differs between lineages.
- Within-lineage effects would be interesting to explore, too - higher resolution!
- Serial intervals can change over time

Thank you

This work is led by Kurnia Susvitasari at SFU

Thank you to Genome BC, the Michael Smith Foundation for Health Research, NSERC and the Canada 150 Research Chairs program

Thank you to the COG Consortium UK!