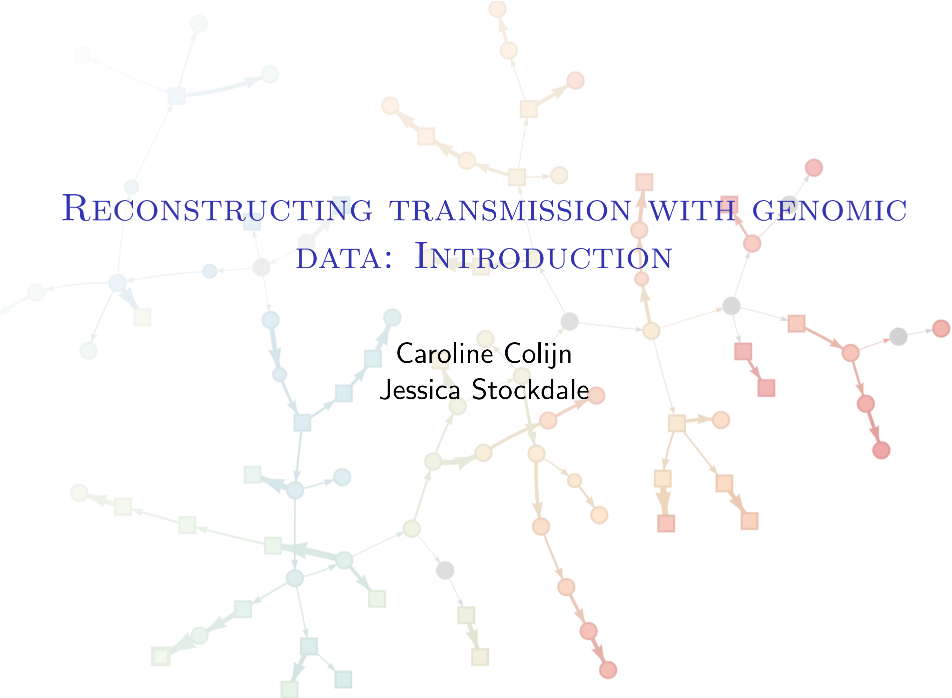


Reconstructing transmission with genomic data: Introduction

Caroline Colijn
Jessica Stockdale

What is “Reconstructing transmission with genomic data”?

Reconstructing transmission:

- inferring who infected whom
- estimating patterns among who infected whom: e.g. were vaccinated individuals infectious?
- inferring transmission routes among locations: farms, regions, even countries

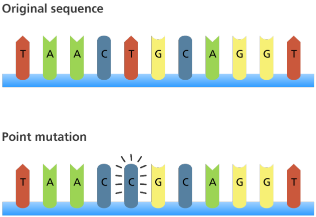
Genomic data:

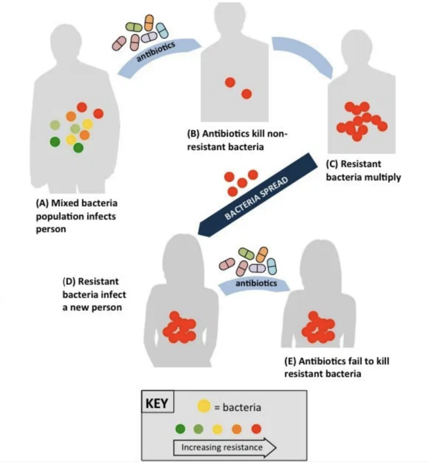
- sequence data (RNA, DNA), usually for viruses, bacteria (maybe other pathogens)
- usually consensus sequences: one sequence per isolate
- usually one isolate per individual (i.e. per host)

Here we focus on person-to-person transmission.

Infections are evolving

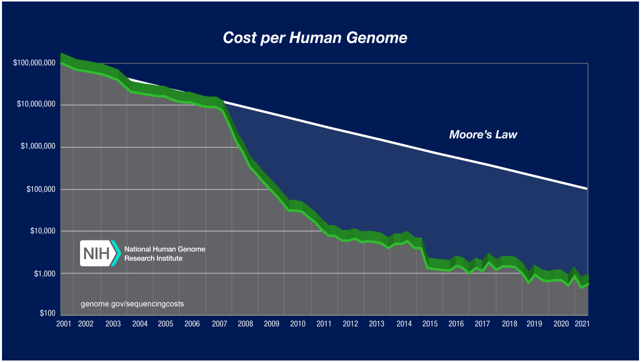
Take a look at the DNA of your favourite bacteria, for example:




antibiotics
(A) Mixed bacteria population infects person
(B) Antibiotics kill non-resistant bacteria
(C) Resistant bacteria multiply
BACTERIA SPREAD
(D) Resistant bacteria infect a new person
antibiotics
(E) Antibiotics fail to kill resistant bacteria
KEY
= bacteria
Increasing resistance

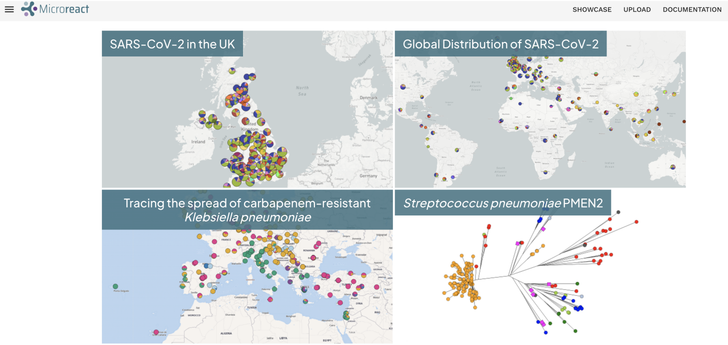
Vivian Chou, Harvard

Sequencing is less expensive now than ever



We can read the RNA, DNA of viruses and bacteria that cause infection, and see how they are changing as they spread.

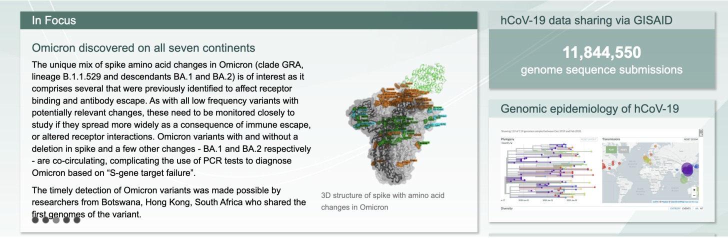
New data - a big opportunity


Microreact
SHOWCASE
UPLOAD
DOCUMENTATION
SARS-CoV-2 in the UK
Global Distribution of SARS-CoV-2
Tracing the spread of carbapenem-resistant *Klebsiella pneumoniae*
Streptococcus pneumoniae PMEN2

microreact.org, by D. Aanensen and colleagues

https://microreact.org

GISAID: Global Initiative on Sharing All Influenza Data



In Focus

Omicron discovered on all seven continents

The unique mix of spike amino acid changes in Omicron (clade GRA, lineage B.1.1.529 and descendants BA.1 and BA.2) is of interest as it comprises several that were previously identified to affect receptor binding and antibody escape. As with all low frequency variants with potentially relevant changes, these need to be monitored closely to study if they spread more widely as a consequence of immune escape, or altered receptor interactions. Omicron variants with and without a deletion in spike and a few other changes - BA.1 and BA.2 respectively - are co-circulating, complicating the use of PCR tests to diagnose Omicron based on "S-gene target failure".

The timely detection of Omicron variants was made possible by researchers from Botswana, Hong Kong, South Africa who shared the first genomes of the variant.

3D structure of spike with amino acid changes in Omicron

hCoV-19 data sharing via GISAID

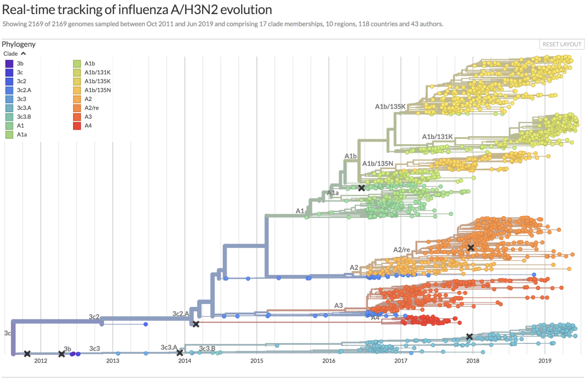
11,844,550
genome sequence submissions

Genomic epidemiology of hCoV-19

GISAID - July 11, 2022

https://gisaid.org

Pathogens acquire genetic variation as they spread



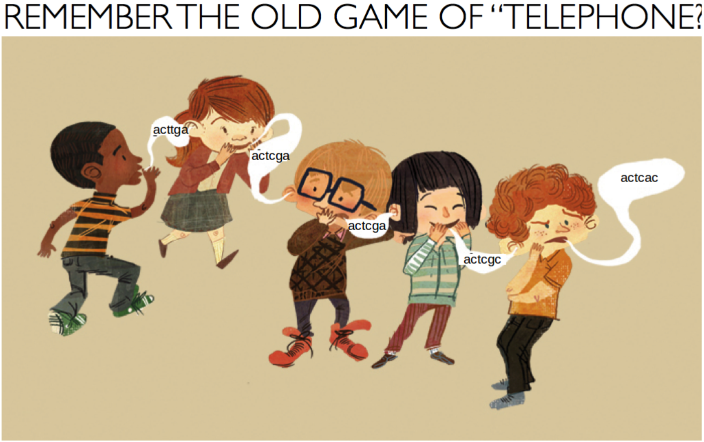
Real-time tracking of influenza A/H3N2 evolution

Showing 2169 of 2169 genomes sampled between Oct 2011 and Jun 2019 and comprising 17 clade memberships, 10 regions, 118 countries and 43 authors.

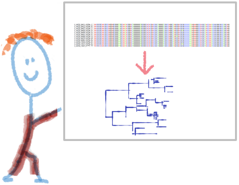
Visualization of flu evolution. Image: nextflu.org

https://nextstrain.org/ncov/global

Pathogens even vary from person to person



Transmission reconstruction



Reconstructing transmission using sequences is possible, but not with perfect accuracy

Why do transmission reconstruction?

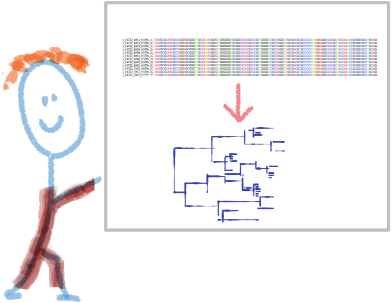
- We may or may not need to know individual transmission events: whether Alice infected Bob, Bob infected Eve, or Eve infected Chloe
- But we do want to know when, where and how transmission takes place
- Drilling into the details of transmission can help understand this, extract patterns, and project for future outbreaks
- Reconstructing transmission can help to identify when infection did not occur

Three reasons: (1) public health applications; (2) modelling applications; (3) mathematical and statistical challenges

Data on genetic variation can help

Data about this variation:

- can inform us about who infected whom - improve outbreak control
- bigger scale: help with choosing best vaccines, best antibiotics



Applications of transmission reconstruction and genomic epidemiology

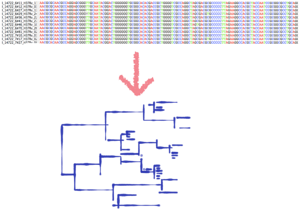
Outbreak questions:

- How fast do we have to find cases?
- How do we find missing cases?
- What are early signs of a big outbreak?

Large-scale questions

- How to choose what goes in a vaccine?
- How to use antibiotics most wisely?

← *GAP*→



Sequence data

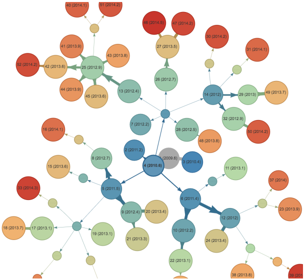
High-resolution picture

Relatedness is key

The answers to our questions aren't just in the data, but in the connections among data points.

The sequence
AACCATAGGT
doesn't mean much for transmission on its own.

But with two:
AACCATAGGT
GACCATAGGT
we know we have two very similar things.
(Off the shelf machine learning:
not good enough)



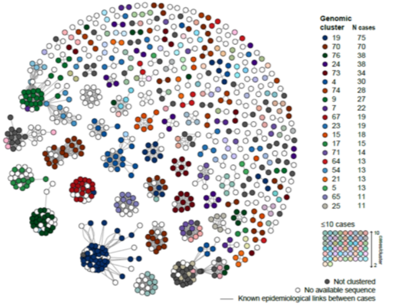
Three main roles for sequences in transmission dynamics

1. Infer global routes of movement of a virus
    - nextstrain.org
    - Challenges from different sampling and sequencing in different places
2. Infer virus population dynamics over time
    - Field of phylodynamics - see the *Evolutionary Dynamics and molecular epidemiology of viruses* module (N. Mueller, J. Palacios)
    - Often large-scale, can be limited in terms of direct public health action; more ecological
3. Reconstruction of transmission in localised outbreaks
    - Potentially useful for public health in the short term
    - High sampling required
    - Can be limited by low pathogen diversity at the outbreak scale

This module: focus on transmission in localised outbreaks

Identifying localized outbreaks: clustering often plays a role.

- Clustering: put similar sequences into groups
- Starting point for onward analysis
- Simple approach to identify groups of closely-related sequences
- Some limitations



Tracking the COVID-19 pandemic in Australia using genomics

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

doi: https://doi.org/10.1101/2020.05.12.20099929

Clustering: epidemiological view

A cluster (or outbreak):

- a set infections arising from a common source or through rapid chains of transmission
- usually in a defined location or population
- usually in a relatively short period of time
- distinguishable from the background epidemic dynamic

Sources: N. Oteko and PANGEA consortium, 2024 (draft)
Oster et al, Am. J. Prev. Med 2021
Context: HIV

Genomic clusters

- SNP: single nucleotide polymorphism. A change from (eg) A to C at a single site.
- Simplest clustering method: Sequences from Bob and Alice are placed in the same cluster if they differ by k SNPs or fewer
- Limitations:
  - what to do about 'N's in the multiple sequence alignment?
  - Should all sites "count" for the same distance?
  - Is there a "right" threshold, k? Why?
- Alternative: use a genetic distance threshold, with an estimated evolutionary model for your pathogen
  - same "what threshold?" problem

Note: none of this has any epidemiological information.

Key idea: sequence similarity reveals recent transmission.

Genetic clustering methods

- Account for time, substitution rate variation, selection: e.g. our own 'transcluster' (Stimson et al MBE 2019 “Beyond the SNP threshold”)
- Account for time, genetic distance, and epidemiological data (cov2clusters, B. Sobkowiak)
- Account for phylogenetic tree – joint evolution of a set of sequences
  - – examples; cluster picker, cluster picker II, cluster matcher (developed for HIV primarily)
  - – limitation: tree may not be known to high accuracy; tree and clusters may change as new data are added
- Integrated methods – combine diagnosis, genotyping, and genetic similarity (e.g. Poon et al, Lancet HIV)
- Modularity: use network science techniques, not strict cutoffs (Liu et al *Virus Evolution* 2023)

Use of clustering: an HIV example

Wilbourn et al, Characterization of HIV Risk Behaviors and Clusters Using HIV-Transmission Cluster Engine Among a Cohort of Persons Living with HIV in Washington, DC, *AIDS Res Hum Retroviruses*

- "Cluster analyses used HIV-Transmission Cluster Engine to identify linked pairs of sequences (defined as distance $\leq 1.5\%$). Twenty-eight clusters of ≥ 3 sequences (size range: 3-12) representing 108 (3%) participants were identified. None of the five largest clusters (size range: 5-12) included newly diagnosed [people living with HIV]. "

Use of clustering: a TB example

Nonghanphithak et al, Clusters of Drug-Resistant Mycobacterium tuberculosis Detected by Whole-Genome Sequence Analysis of Nationwide Sample, Thailand, 2014-2017, *Emerging Infectious Disease* 2021

- "We analyzed whole-genome sequence data for 579 phenotypically drug-resistant M. tuberculosis isolates (28% of available MDR/pre-XDR and all culturable XDR TB isolates collected in Thailand during 2014-2017). Most isolates were from lineage 2 (n = 482; 83.2%)."
- "Cluster analysis revealed that 281/579 isolates (48.5%) formed 89 clusters, including 205 MDR TB, 46 pre-XDR TB, 19 XDR TB, and 11 poly-drug-resistant TB isolates based on genotypic drug resistance."
- Researchers compare "clustered" to "unclustered" cases to learn about what might be drivers of local, recent transmission.

Direct application of clustering – direct transmission didn't happen

Exclude suspected transmission events:

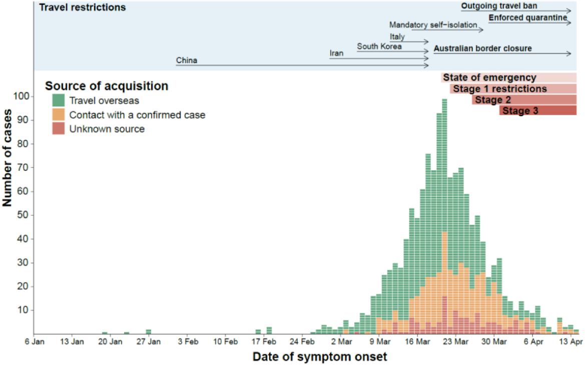
1. Example: consider 2 cases that are epidemiologically linked
2. If their viral (or bacterial) sequences are very different, there was likely no direct transmission
3. The apparent link was spurious ("false positive" epidemiological link)
4. Implication: there was another source of infection.

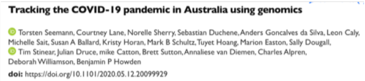
Direct application of clustering - find new links beyond epidemiological analysis

Identify sources of infection that did not have apparent epidemiological links

1. Sequences firmly place Bob in a cluster with Alice and Eve (for example, sequences are 1 SNP away)
2. Bob has no known epidemiological links to Alice, Eve or their contacts
3. Yet the genomes are so close – compared to other isolates in the dataset – that a link is very likely to exist
4. This can help identify previously unknown exposures: true links were missing ("false negative" epi link)

Example: COVID-19 genomic epidemiology in Australia





Tracking the COVID-19 pandemic in Australia using genomics

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

doi: https://doi.org/10.1101/2020.05.12.20099929

https://www.nature.com/articles/s41467-020-18314-x

SARS-CoV-2 data from Australia

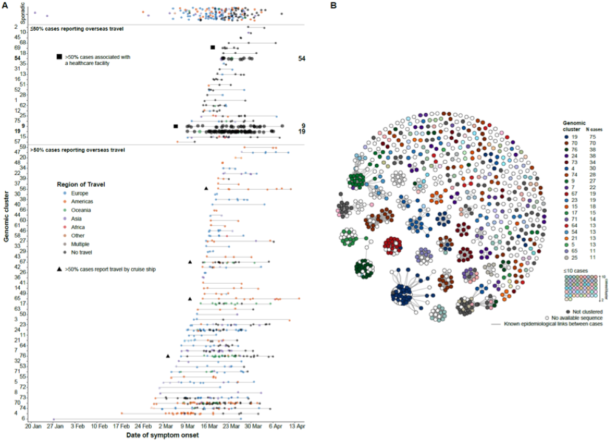
Brief data description

- 1388 lab-confirmed cases in Victoria
- 62% travellers
- 27% known contacts
- 10% unknown source of exposure
- 1242 sequenced
- 1085 passed quality control
- Maximum 15 SNPs compared to Wuhan 1

Clusters in Australian (Victoria) SARS-CoV-2 data

- The authors used ClusterPicker to divide the genomes into clusters
- 737 of 1085 were in any cluster
- 76 clusters: median size 5, median duration 13 days This suggests good control (and could provide a serial interval estimate too!)
- 34 clusters were entirely overseas travellers
- 34 were mixed. In these, typically the first case was a traveller
- 81 sequences with unknown exposure (from the epi point of view): in 24 clusters
- This gives information about the exposure

Figure 3.


A
Sporadic
≤50% cases reporting overseas travel
>50% cases associated with a healthcare facility
54
9
19
>50% cases reporting overseas travel
Region of Travel
Europe
Americas
Oceania
Asia
Africa
Other
Multiple
No travel
>50% cases report travel by cruise ship
Genomic cluster
2
10
45
68
69
18
54
35
31
13
16
51
52
28
1
62
12
25
75
9
19
15
57
59
47
20
48
34
22
39
37
56
30
38
40
66
44
60
61
46
58
27
33
29
43
67
42
26
36
41
14
49
65
17
63
50
3
23
24
21
64
11
7
76
32
53
71
55
5
72
8
73
70
74
4
6
20 Jan
27 Jan
3 Feb
10 Feb
17 Feb
24 Feb
2 Mar
9 Mar
16 Mar
23 Mar
30 Mar
6 Apr
13 Apr
Date of symptom onset
B
Genomic cluster
N cases
19 75
70 70
76 38
24 38
73 34
4 30
74 28
9 27
7 22
67 19
23 19
15 18
17 15
71 14
64 13
54 13
21 13
5 13
65 11
25 11
≤10 cases
10
2
Not clustered
No available sequence
Known epidemiological links between cases

Learning from epidemiological and genomic clusters

- Epidemiological (epi) clusters: groups of cases thought to be linked together on the basis of epidemiological (not genome) data, eg where people live, timing, health care, suspected exposure
- Genomic clusters: similar genomes grouped together on the basis of (sort of) genetic distance
- Four distinct epidemiological clusters → one genomic cluster -find links they didn't know about
- One big epi cluster separated into 4 distinct genomic clusters: exclude links they thought they knew about

Data required for this nice work: SARS-CoV-2 sequences, suspected exposure times and sources from epidemiology.

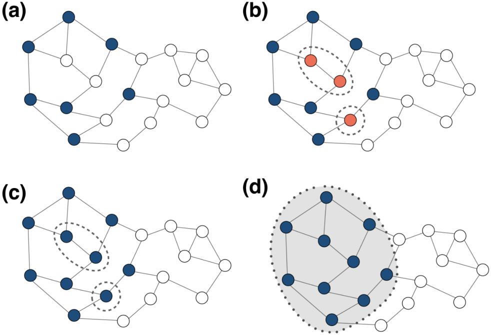
Punch line: Just with sequences and clustering, we can learn things about transmission. But we can do better.

Why do more than clustering?

If a pathogen spreads person-to-person, what do “unclustered” cases mean? At least one of:

- Transmission that is ancestral to the sample but not sampled (we don’t have the infector, or their infector)
- The individual did not infect anyone (that we identified before the relevant time period ended)
- There was a lot of evolution since the last transmission event (e.g. long latency): a form of transmission not being sampled
- Low cutoffs: cluster boundaries are strict, so some (true) links are rejected
- Epi-identified clusters: as above and/or incorrect epidemiological assumptions, e.g. contact data are not perfect

Clustering and sampling in a network



Min et al, Chaos, Solitons and Fractals, 2024

Sampling and cutoffs can have a large impact on clustering patterns.

Clusters and too many pairs

- In a cluster of 10 people there are 45 pairs.
- Each person is infected once: only 9 true pairs at most.
- Cluster of 50: 1225 pairs! Only 49 can be transmission events.
- Some papers interpret “being in the same cluster” or “being within 3 SNPs” as indicating transmission, but this includes a lot of incorrect transmission events

Beyond clusters: the challenge of transmission reconstruction

- Sequences offer much more than just really good typing
- Sequences and their distances don't tell us directly who infected whom
- But they provide information about it
- Additional information in time, location, shared mutations
- Challenge: extract transmission information, capturing uncertainty

In the next part of this lecture: some key terms, and previous methods to meet this challenge

Key terms and concepts

- Transmission tree: who infected whom (and often, at what time)
- Phylogenetic tree, or phylogeny: more on this later.
  - Tips: sampled taxa.
  - Internal nodes: inferred common ancestors.
  - A phylogeny is kind of like the tree of life.
- Genetic distance: a measure of how distant two sequences are. Could be number of single nucleotide differences, or distance in an evolutionary model
- SNP: single nucleotide polymorphism
- More broadly: we use probability, maximum likelihood, trees, evolutionary models, Bayesian inference; serial intervals, generation times.

Selected early methods - you will hear more

- 2008 Cottam et al Cottam et al: Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. (Roy. Soc. Proc. B)
- 2012 Jombart et al, Reconstructing disease outbreaks from genetic data: a graph approach, Heredity (SeqTrack)
- 2012 Ypma RJ, Bataille AM, Stegeman A, Koch G, Wallinga J, et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc Biol Sci 279:
- 2012 Morelli MJ, Thebaud G, Chadoeuf J, King DP, Haydon DT, et al. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput Biol 8
- 2014 Outbreaker: Jombart T, Cori A, Didelot Z, Cauchemez A, Fraser C, Ferguson N, Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data, PLOS Computational Biology

How is the genetic distance between two pathogen isolates related to identifying transmission?

The genetic distance between transmitting pairs depends on

- timing of cases
  - TB, a chronic disease, has highly variable generation time (latency). Other pathogens are less variable (e.g. SARS-COV-2)
- How diverse is the pathogen population within an individual?
  - is diversity higher in high-burden places?
  - how does it depend on the individual?
  - how does it depend on treatment, pre-existing subpopulations, selection, re-infection
- The bioinformatics pipeline and culture method can matter.

Genetic distance and transmission

You have two cases. Are they (likely to be) linked by transmission?

Ideally, we integrate various sources of data and local information:

- Transmission $\leftrightarrow$ genetic distance depends on:
  - in-host diversity
  - are polymorphisms related to drug resistance, treatment history?
- Transmission $\leftrightarrow$ timing depends on
  - natural history / timing of transmissibility after infection
  - sampling - individual (e.g. health care worker sampled sooner than 'hidden' population
- location
- social contacts

When is person-to-person transmission reconstruction from genomic data feasible?

Person-to-person transmission reconstruction requires:

- Enough sampling density that there are likely some transmission pairs in the data.
- Inference methods targeted for the sampling you have. For our methods, over about 25%. (if less, there are options, but not in our main focus)
- Clustering: reduce numbers of very unlikely pairs (100s of SNPs) or very long branches in the phylogeny (with many unsampled cases)
- Sufficient diversity that the genomic data tells you something. SARS-CoV-2: borderline with approximately 1 mutation every 2 weeks.

Real-time interventions and applications

Transmission reconstructions can be used to predict:

- the most likely infector(s) for individuals, and how uncertain this is
- where there are unsampled cases (in the phylogeny, or in the epidemiology: connected to which individuals, for example)
- for whom we have not found a likely infector: could be used to direct further investigation, if the outbreak is ongoing
- how long before secondary infections?
- correlates of “transmitter” status.
    - this is in contrast to standard molecular epidemiology, which often asks: what is correlated with “being in a cluster” (transmitter or no)
    - Side note: interpreting “clustering” is hard – missing transmission, ancestral to the data, but not in the data.

Challenges and options

- Genetic distances may be higher if there is a lot of drug resistance (TB) or sites under selection
- Need closely-related clusters: probability or distance cutoffs still play a role in many analyses
- Many tools exist and the field is growing
- As yet, very limited tools for pathogens that move among humans and in the environment. Would need strong baseline data for different populations.

Challenges and options, continued

- Some packages and tools:
  - TransPhylo: R, unsampled cases and in-host diversity
  - phybreak (Klinkenburg): R, no unsampled cases
  - Beastlier (Hall): Beast, simultaneous transmission and phylogenetic trees, but no unsampled cases
  - Outbreaker: R, no phylogenetic trees, and no in-host diversity, but flexible.
  - SCOTTI (de Maio): Beast2, unsampled $\approx$ environmental samples
  - New methods are coming!

Some key components of transmission reconstruction

- Clustering, or choosing which sequences to include. Importations?
- Sampling– can we account for unsampled hosts?
- How do we account for the shared evolutionary history of our set of viruses?
  - Phylogeny?
  - Parsimony?
  - Pairs?
- How do we account for additional epidemiological data, like:
  - timing of infection, clearance
  - location, exposure site, likely exposure time
- Transmission bottleneck: how many pathogens are transmitted from one host to another?

What's next?

- Reconstructing transmission trees!
- Introduction to genomics for genomic epidemiology
- Non-phylogenetic outbreak reconstructions in outbreaker
- Phylogenetic trees: theory and practice
- TransPhylo: genomic epi with trees
- Research forefronts: SARS-CoV-2
- Discussion