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)
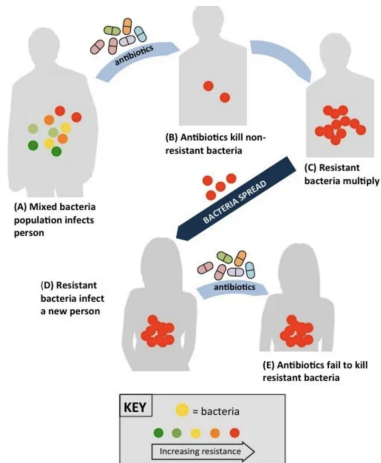
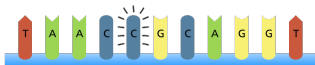Here we focus on person-to-person transmission.

Infections are evolving

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

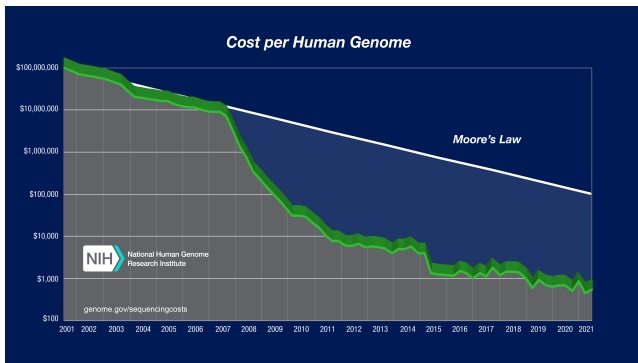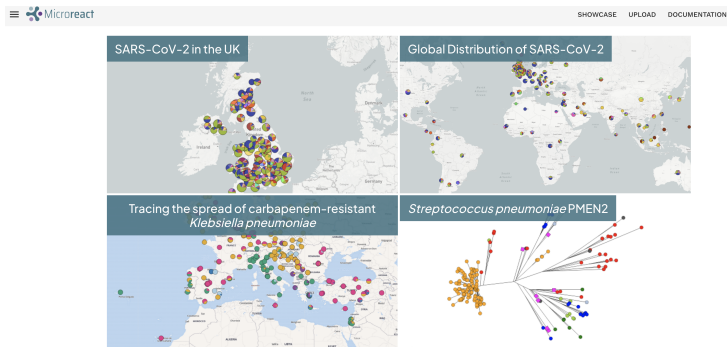Vivian Chou, Harvard

Cost per Human Genome

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.org, by D. Aanensen and colleagues

https://microreact.org
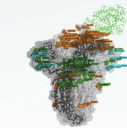
GISAID: Global Initiative on Sharing All Influenza Data

GISAID - July 11, 2022

https://gisaid.org

PATHOGENS ACQUIRE GENETIC VARIATION AS THEY SPREAD
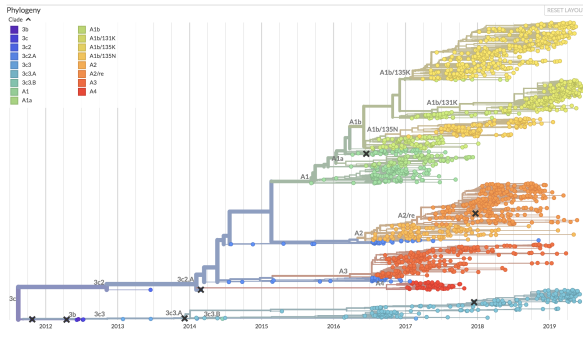
Visualization of flu evolution. Image: nextflu.org

https://nextstrain.org/ncov/global

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

WHY DO TRANSMISSION RECONSTRUCTION?

- We may not really 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
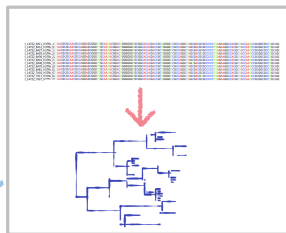
Applications of transmission reconstruction and genomic epidemiology

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?

$\leftarrow$ *GAP* $\rightarrow$
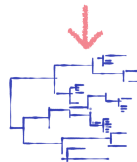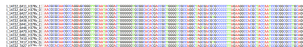
Sequence data
High-resolution picture

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)
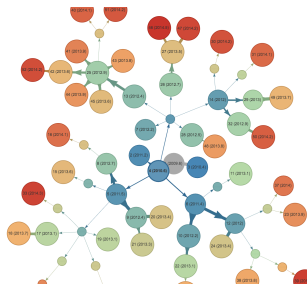
THREE MAIN ROLES FOR SEQUENCES IN TRANSMISSION DYNAMICS

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
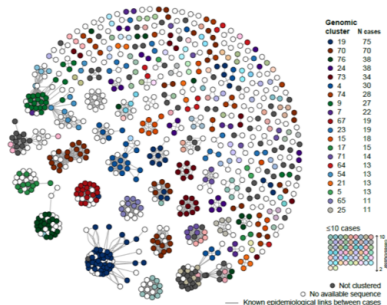
Tracking the COVID-19 pandemic in Australia using genomics

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

A cluster: a group of sequences that are more similar to each other than to objects outside of the group

- 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 or why not?
- Alternative: use a genetic distance threshold, with an estimated evolutionary model for your pathogen
  - same "what threshold?" problem

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)

DIRECT APPLICATION OF CLUSTERING - WHERE 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 - there was a link that was not found by 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)
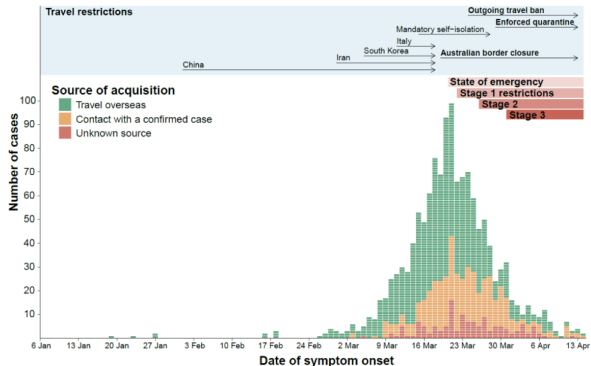
Tracking the COVID-19 pandemic in Australia using genomics

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.
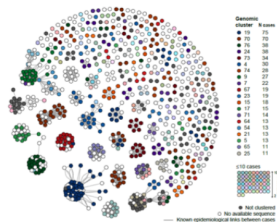
SPECIAL CLUSTERS

- 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 $\rightarrow$ 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 more!

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 WAS EACH CASE INFECTED?

The TransPhylo method (more later) combines the priors for generation and sampling time, plus the genetic data, to give posterior times of infection for each case.
Data:

- Cluster of closely-related cases in Norway
- Cases occur among people who have moved to Norway from abroad
- It is often assumed that they were infected before arriving
- But genomic data show signs of recent transmission

We compared time of arrival in Norway to the estimated time of infection for 13 closely-related cases.
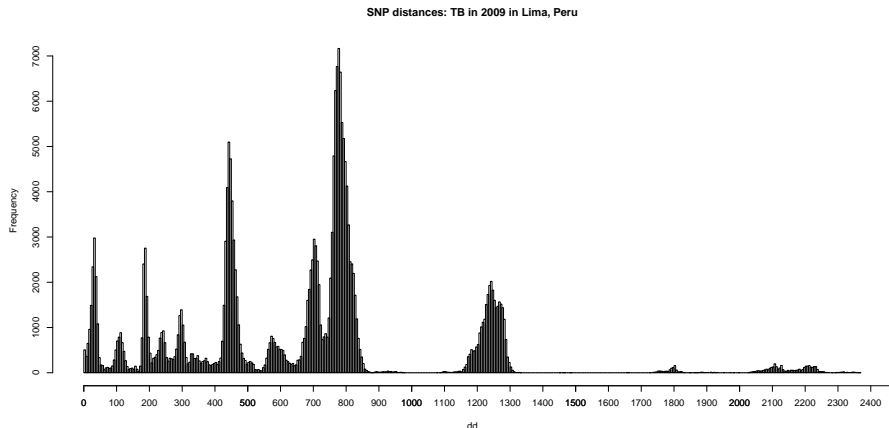
WHEN MAY TELL US WHERE

Red: Estimated time of infection. Blue: arrival in Norway.

Some cases were very likely infected in Norway.

PERU: A DIFFERENT CONTEXT FOR TB

SNP distances: TB in 2009 in Lima, Peru

Year: 2009. Population-wide survey, random sample, plus some isolates
from an MDR household study (L. Grandjean).

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

- Peru: many pairs have very large SNP distances - 100s of SNPs!
- Compare: closely-related clusters (e.g. Norway): closer to 5 SNPs diameter. Sometimes less!

Person-to-person transmission reconstruction in this course 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 approx 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.
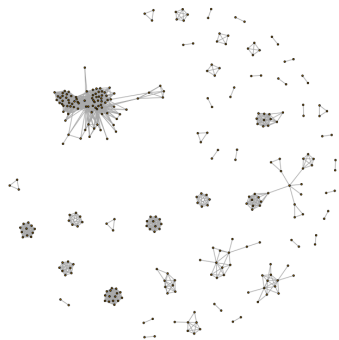
- Some packages and tools:
  - ▶ TransPhylo: R, unsampled cases and in-host diversity
  - ▶ phybreak (Klinkenburg): R, no unsampled cases
  - ▶ Beastlier (Hall): Beast, simultaneous transmission and phylogenetic trees, but no unsampled cases
  - ▶ Outbreaker: R, no phylogenetic trees, and no in-host 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