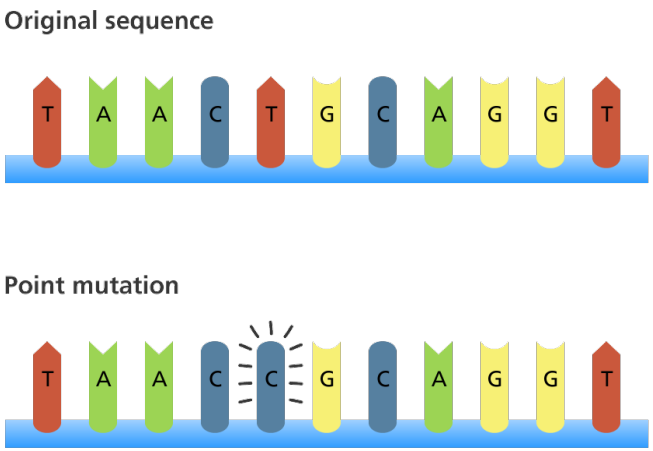
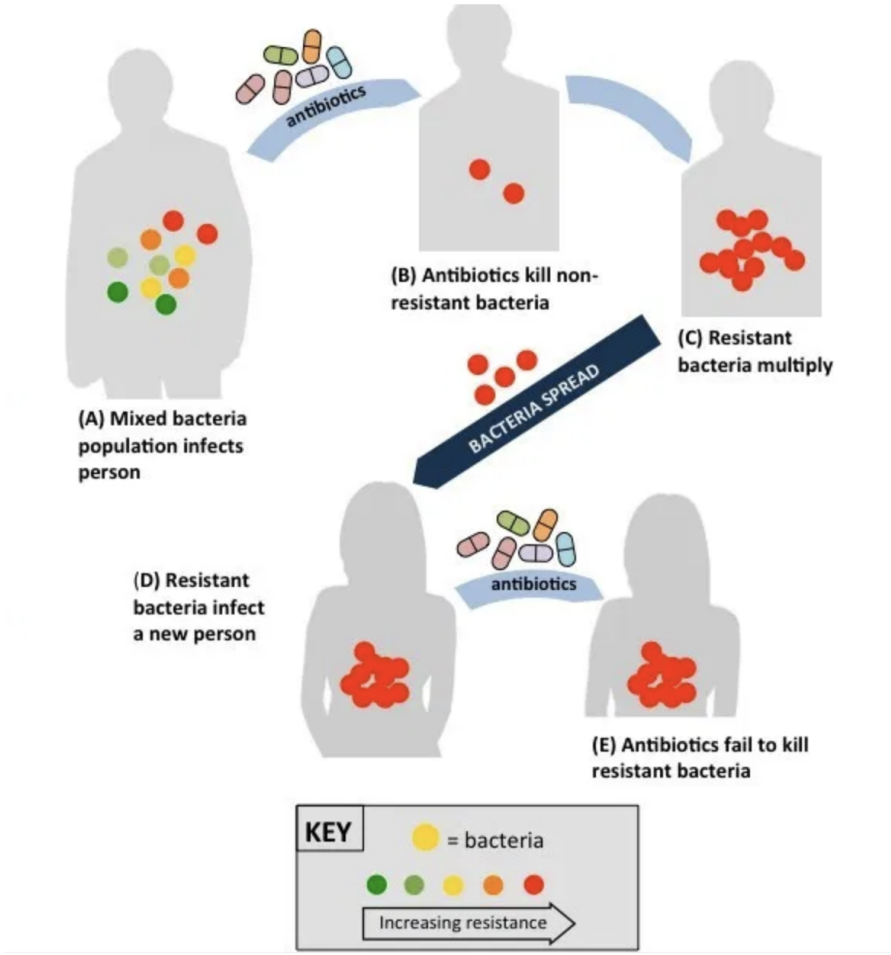
Reconstructing transmission with genomic data: Introduction

Caroline Colijn
Jessica Stockdale

Infections are evolving

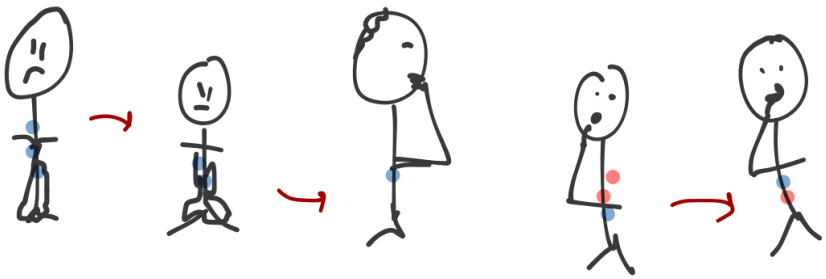
Take a look at the DNA of your favourite bacteria:

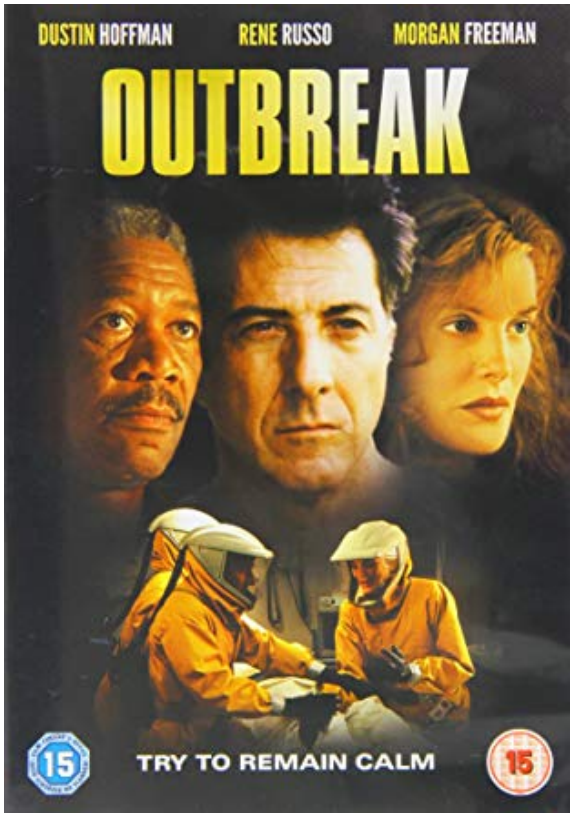




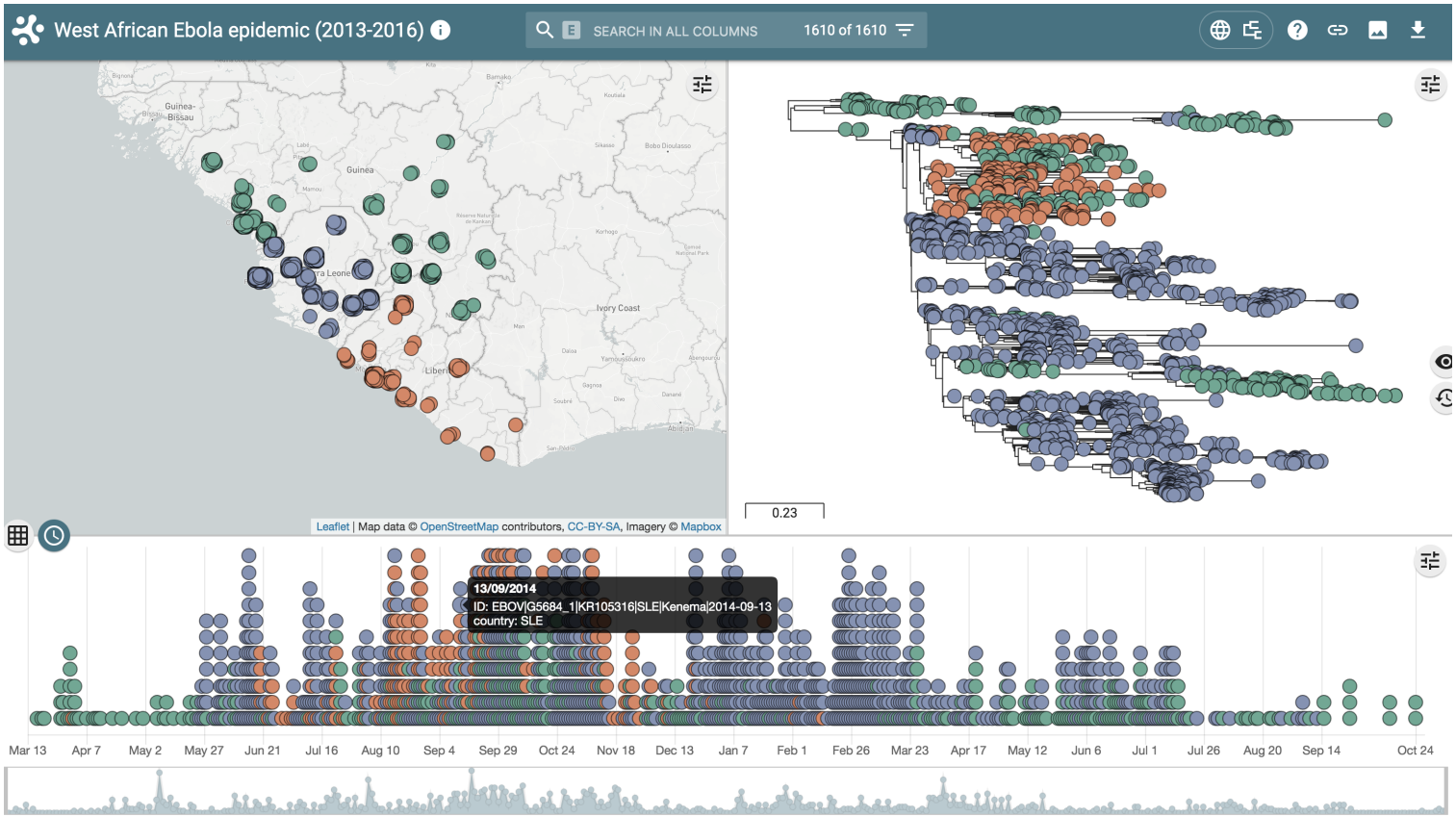
Vivian Chou, Harvard

It is hard to control outbreaks





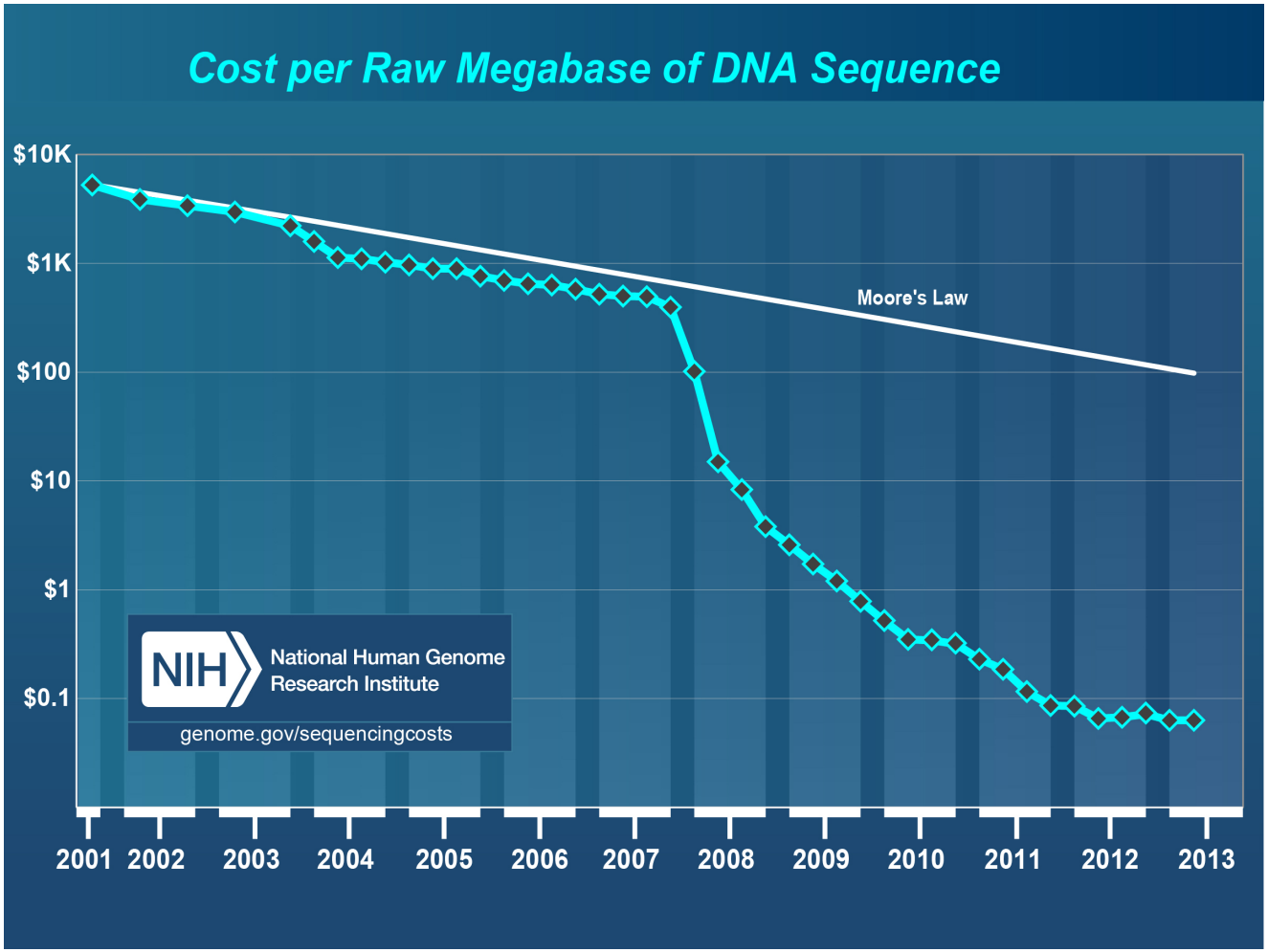
New data - a big opportunity



microreact.org, by D. Aanensen and colleagues

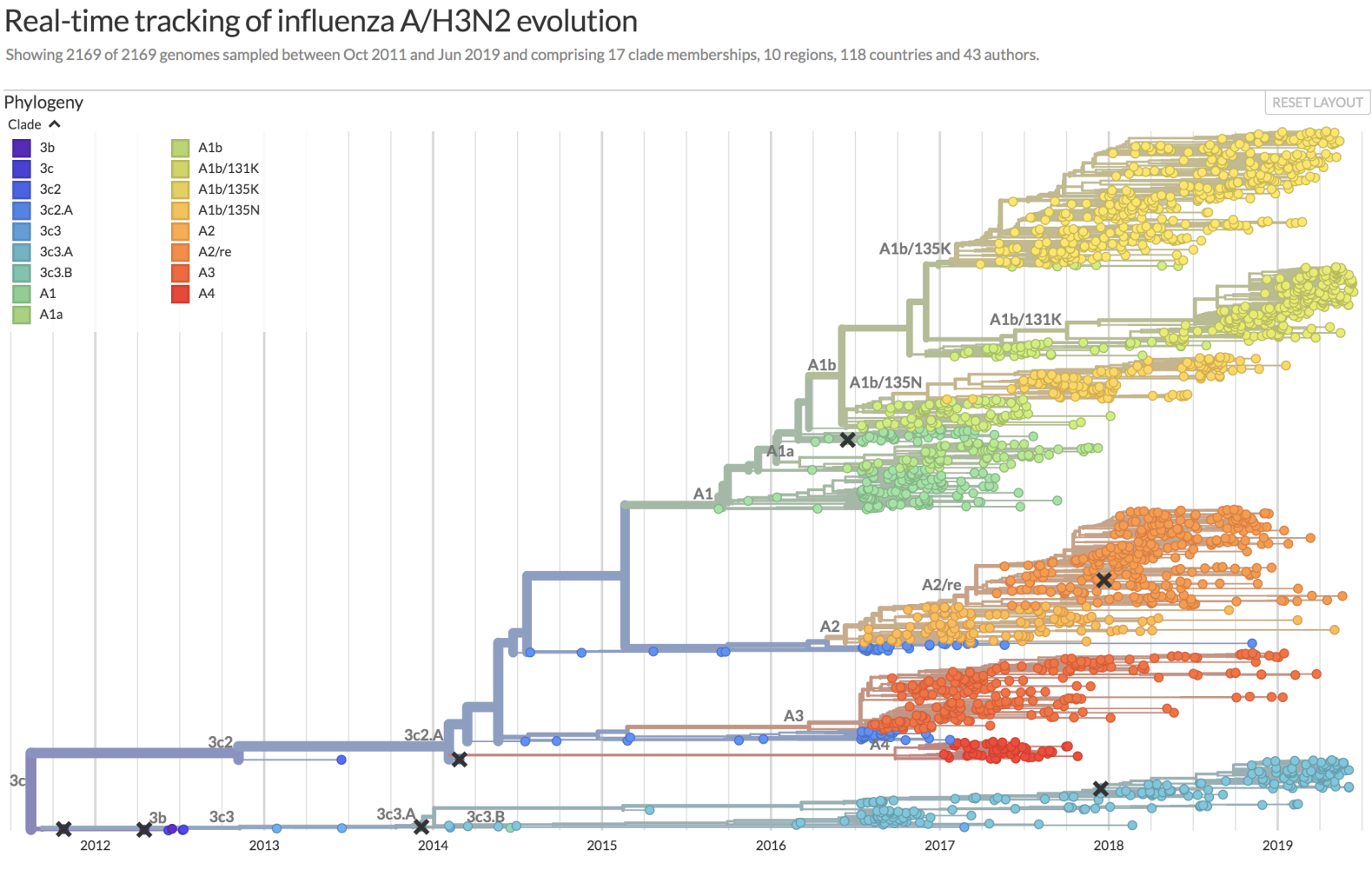
https://microreact.org/project/west-african-ebola-epidemic?tt=rc

Sequencing is less expensive now than ever



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

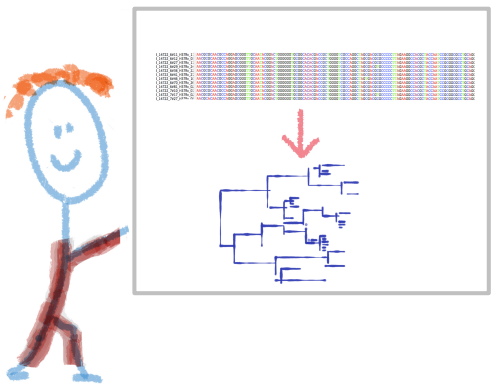
Bugs acquire small genetic variation as they spread



Visualization of flu evolution. Image: nextflu.org

Bugs even vary from person to person

Can we use sequences to understand transmission?



Can we learn about who infected whom and when from sequence data?

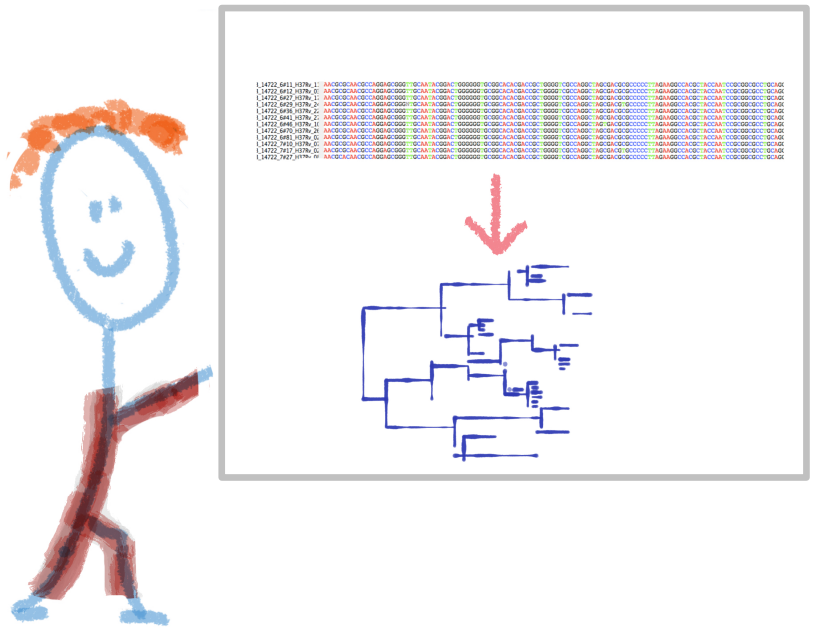
Why do we want to know who infected whom?

- We don't really need to know whether Alice infected Bob, Bob infected Eve, or Eve infected Jackson
- But we do want to know when, where and how transmission takes place
- Drilling into the details of transmission can help understand this
- And it can help to know when infection did not occur

Data on genetic variation can help!

Data about this variation:

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

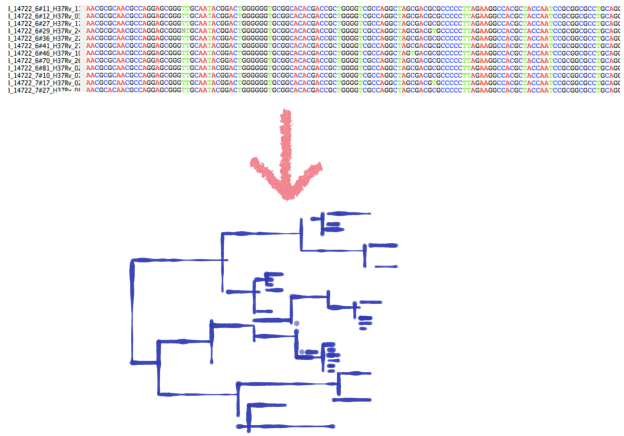


Challenges in genomic epidemiology and beyond

Outbreak questions:

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

$\leftarrow GAP \rightarrow$



Big fancy data

Large-scale questions

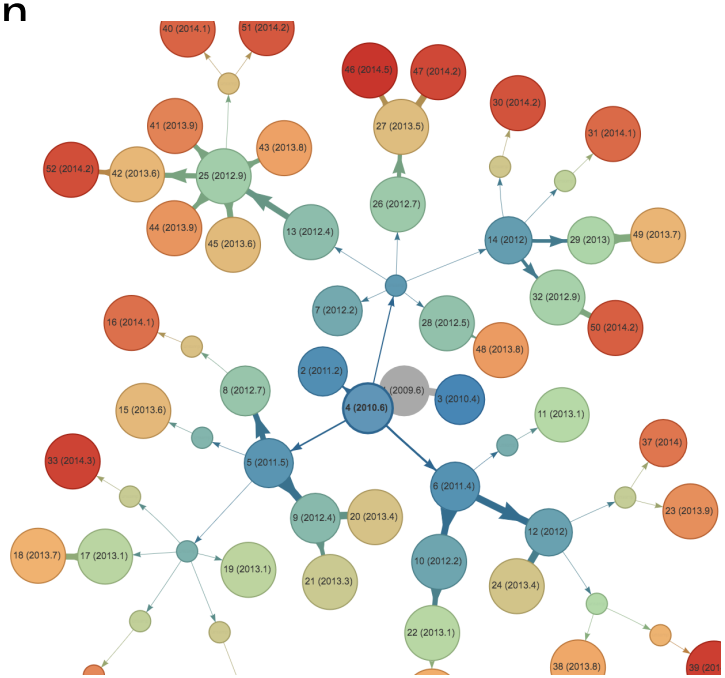
- How to choose what goes in a vaccine?
- How to use antibiotics most wisely?

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.

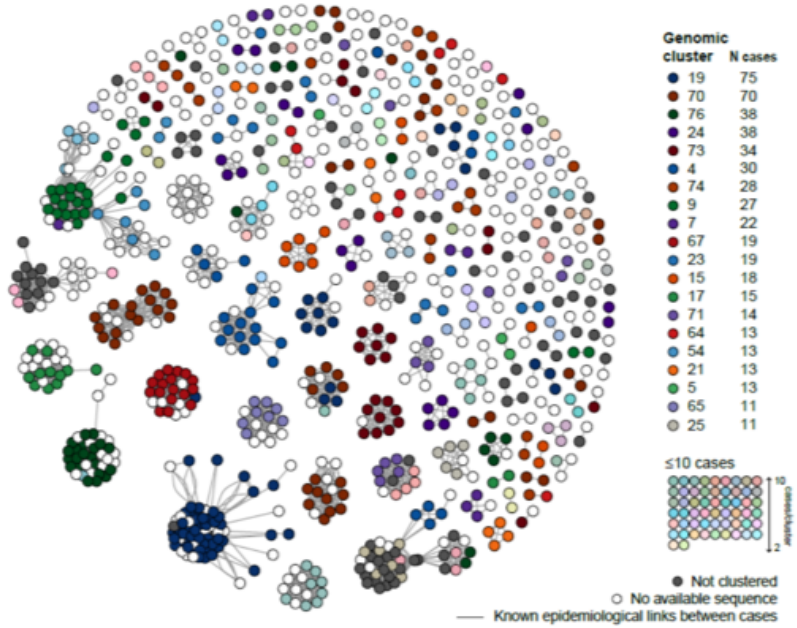


Three roles for sequences in public health

1. Infer global routes of movement
   - nextstrain.org
   - Challenges from different sampling and sequencing in different places
2. Infer population dynamics back through time
   - Field of phylodynamics - see Julia Palacios' module!
   - Limited in terms of direct public health action; more ecological
3. Analysis of transmission in localised outbreaks
   - Potentially directly useful for public health in the short term

Transmission in localised outbreaks

- The first step is to broadly find out what you have.
- Clustering: put similar sequences into groups



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

How to put sequences into groups: clustering

A cluster: 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
- More sophisticated clustering methods:
  - Account for time, rate variation, selection: eg our own 'transcluster' (Stimson et al MBE 2019 Beyond the SNP threshold")
  - Phylogenetic methods: cluster picker, cluster picker II, cluster matcher (developed for HIV primarily)
  - Integrated methods – combine diagnosis, genotyping, and genetic similarity (Poon et al Lancet HIV)

Sequence clusters in public health 1

Exclude suspected transmission events:

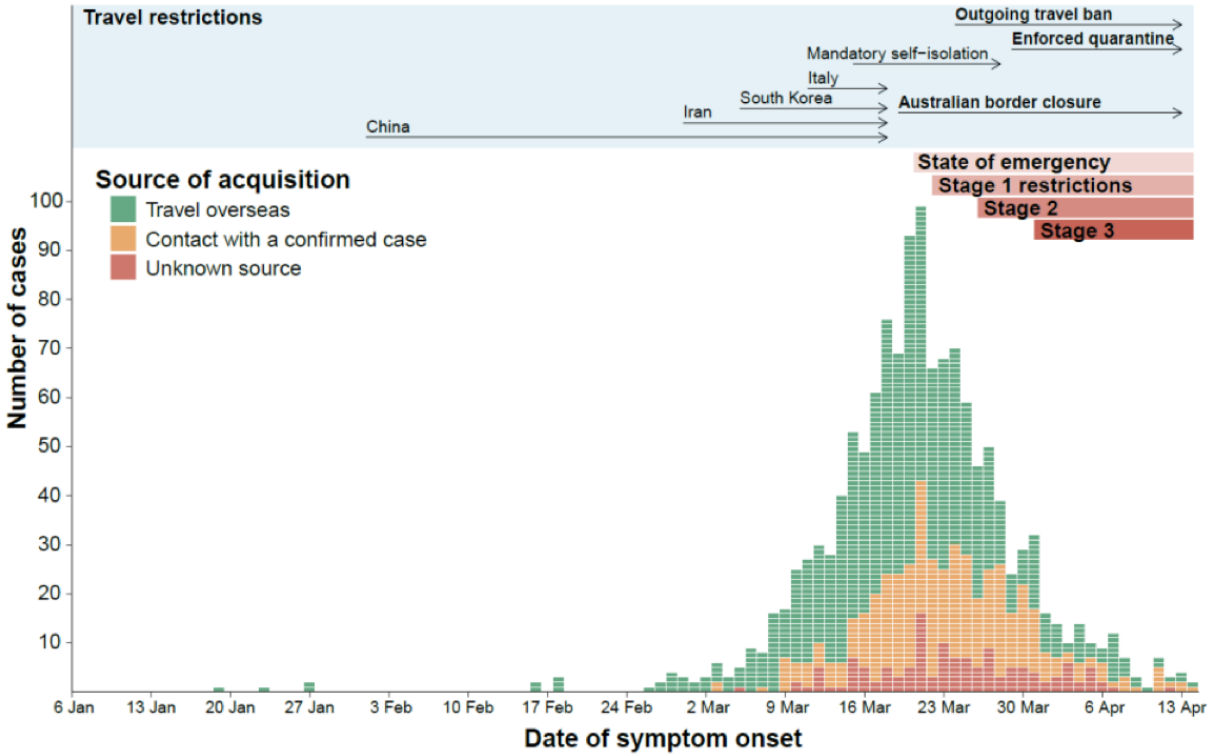
1. For example: 2 cases that are epidemiologically linked
2. If their viral (or bacterial) sequences are very different, there was likely no direct transmission
3. The apparent link was spurious ("false positive" epi link)

Sequence clusters in public health 2

Identify sources of infection that did not have 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 epidemiological links to Alice, Eve or their contacts
3. 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

SARS-CoV-2 data from Australia

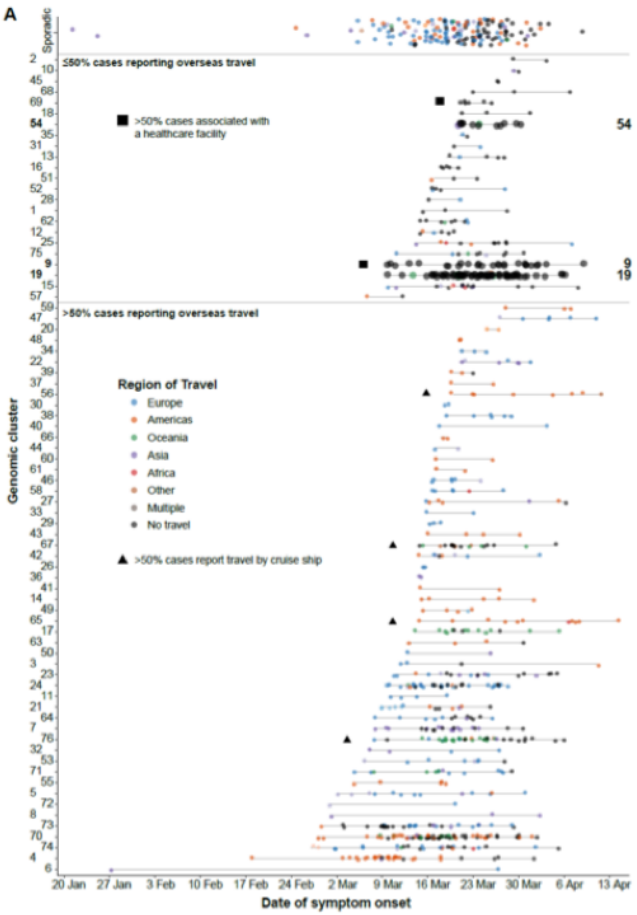
Brief data description

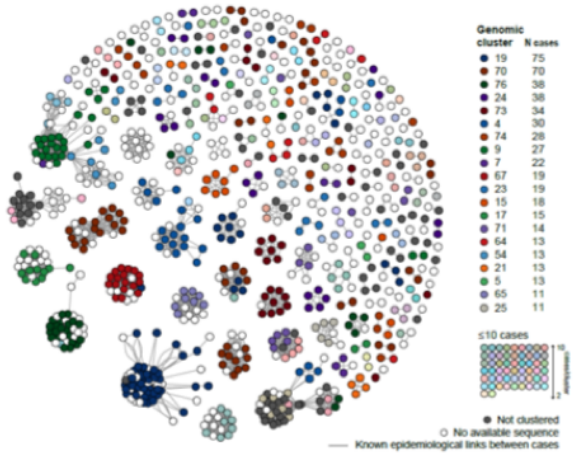
- 1388 lab-confirmed cases in Victoria
- 62% travellers
- 27% known contacts
- 10% unknown source of exposure
- 1242 sequenced
- 1085 passed quality control
- Maximum 15 SNPs compared to Wuhan 1

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, typically the first case was a traveller
- 81 sequences with unknown exposure (from the epi point of view) land in 24 clusters
- This suggests what the exposure was

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
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
Not clustered
No available sequence
Known epidemiological links between cases

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 one genomic cluster -find links they didnt 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.

Early work: transmission with sequences

- Cottam et al: Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc. Biol. Sci. 2008.
  - Phylogenetic trees constrain transmission events
- Jombart T, Eggo RM, Dodd PJ, Balloux F (2011) Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106: 383390
  - Parsimonious: transmission tree with fewest mutations

Early work continued

- 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
  - Both use a unified likelihood of genetic and epidemiologica data - needs full sampling
- 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 CB

Outbreaker and SARS

The data:

- Sequences
- Times of sample collection
- Number of cases

Additional inputs (derived from data):

- Generation time: time from infection to symptom onset
- Time from infection to sample collection

The math behind outbreaker

Outbreaker is a Bayesian MCMC method, and it uses *augmentation*.
" Augmented" quantities for each case i:

- The infector for each case, or most recent sampled infector
- The number of unknown intermediates between i and i's ancestor
- The date when i was infected

The likelihood relies on these augmented data.

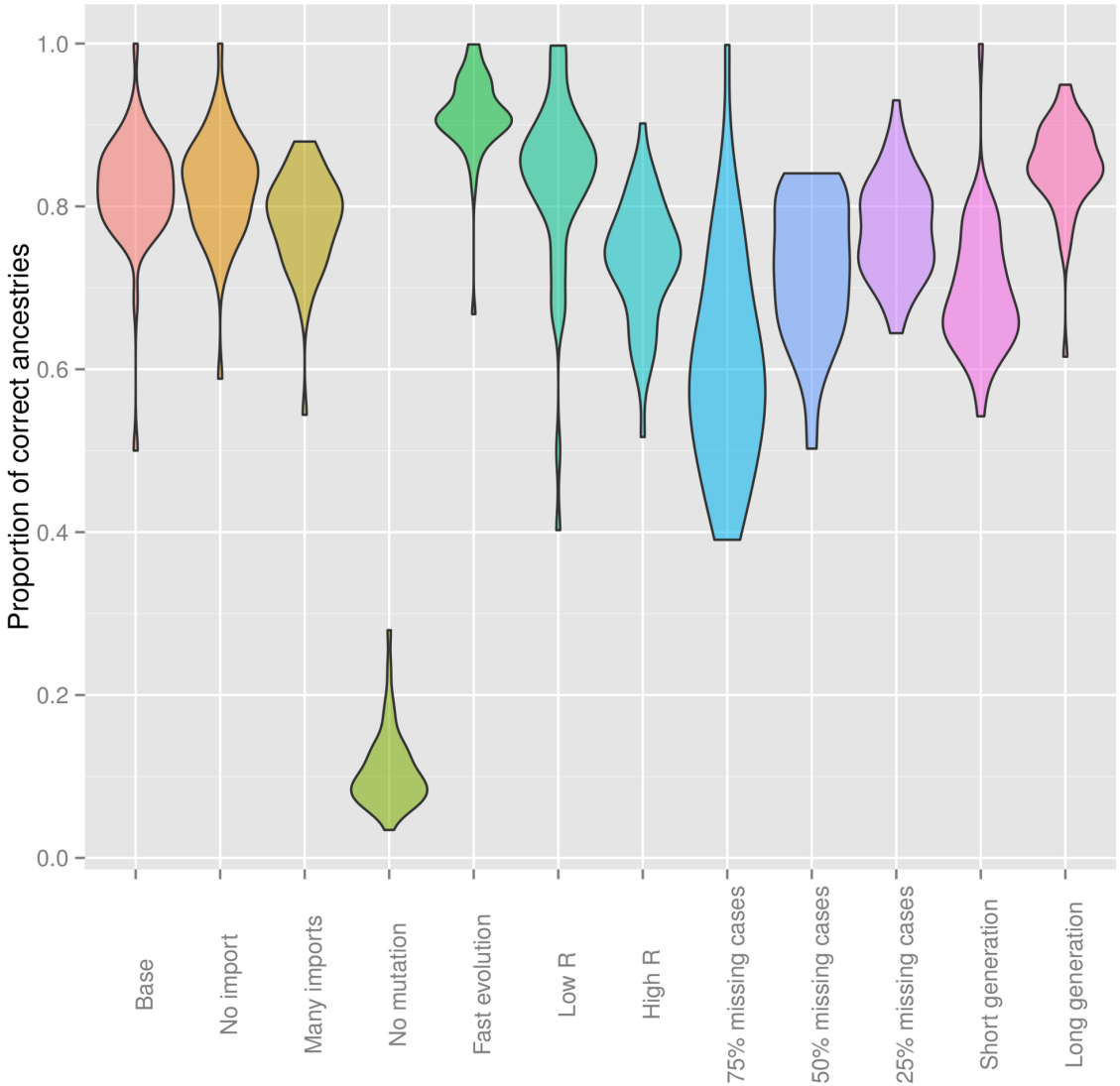
Outbreaker's likelihood in brief

Likelihood for case $i \propto$
P(sequence | ancestor's sequence, intermediates, evolution model) $\times$
P(sampled time | infection time) $\times$
P(infection time | parent's infection time, intermediates)

A 60-second MCMC briefing:

- Propose augemented data: transmission tree, intermediates, times of infection
- Compute the above likelihood
- Accept or reject (in a principled way, depending on the proposal method and the likelihood)
- Run for a long time. Collect "posterior samples" of these quantities

Outbreaker results: simulated data



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003457

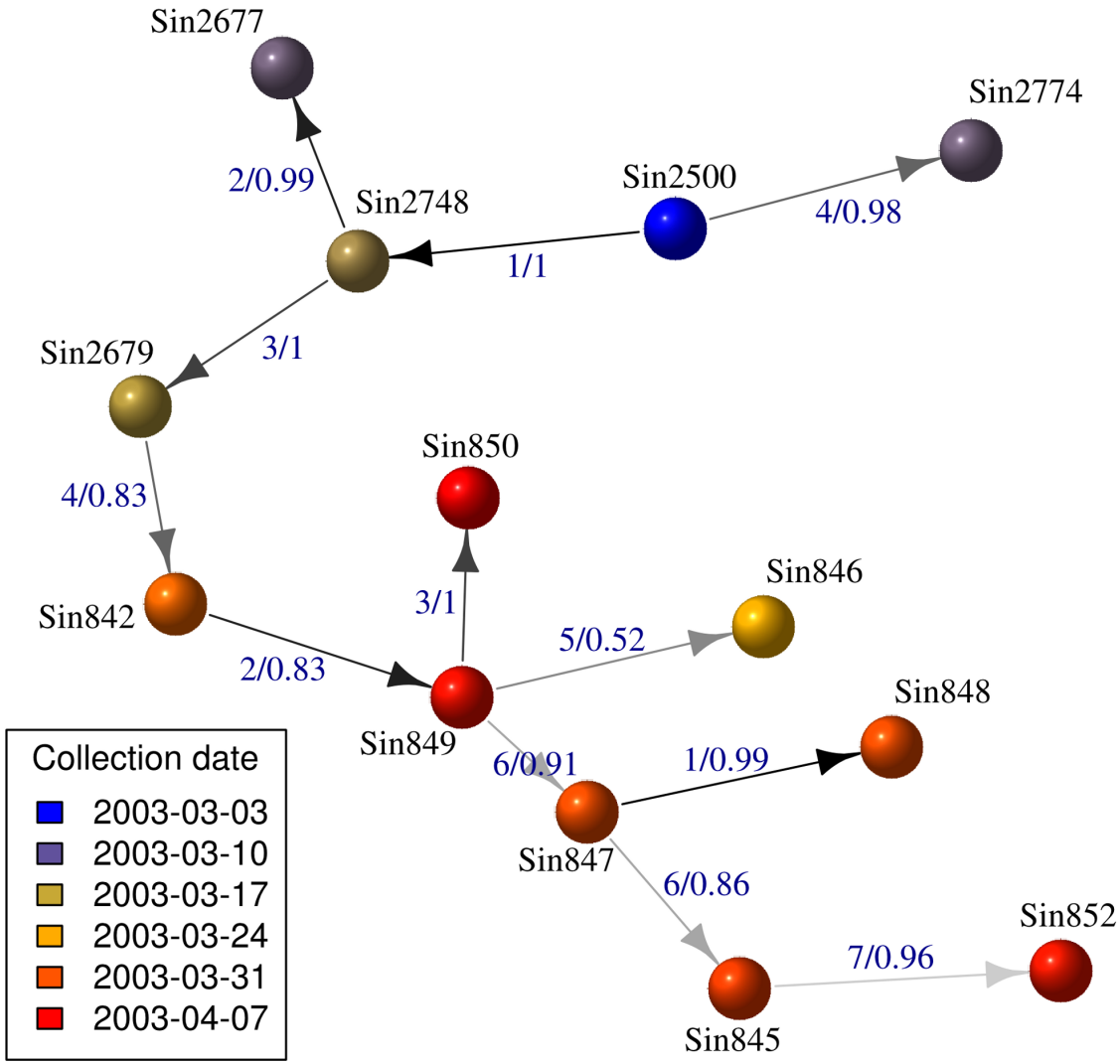
Application: SARS (the first SARS!)

- 13 SARS "genomes" – but note: sequencing was *very* different then
- Ruan et al, *Lancet*, Volume 361, Issue 9371, 24 May 2003
  https://www.sciencedirect.com/science/article/pii/S0140673603134149
- Liu et al *PLOS Medicine* https://doi.org/10.1371/journal.pmed.0020043
- Generation time:
  gamma with mean 8.4 days
- Time to sampling: same
- Lots more variation
  than TB or SARS-CoV-2

cm

	Singapore cases					Overseas cases											
	Index case		Primary		Secondary	Hanoi		Hong Kong		N China		N China					
	SIN2500		SIN2774		SIN2679	URBANI		CUHKW1		BJ01		BJ03					
	29 711		29 711		29 711	29 727		29 712		28 920		29 291					
Genome position		Primary SIN2748		Primary SIN2677		Canada TOR2	Hong Kong HKU39849		S China GZ01		N China BJ02		N China BJ04		Variations frequency	ORF/protein	AA change (SIN2500→ variation)
Length		29 706		29 705		29 712	29 727		29 429		29 430		24 774				
2601	T	T	T	T	T	C	T	C	T	T	T	T	T	_	2	Orf1a (MHV p65)	Silent
7919	C	C	C	C	C	C	T	C	C	C	C	C	T	C	2	Orf1a (nsp1)	A→V
8559	T	T	T	T	T	T	T	T	T	C	T	T	T	A	2	Orf1a (nsp1)	Silent
8572	G	G	G	G	G	G	G	G	G	T	G	T	G	G	2	Orf1a (nsp1)	V→L
9404	T	T	T	T	T	T	T	T		C	C	C	C	_	5	Orf1a (nsp1)	V→A
9479	T	T	T	T	T	T	T	T	C	C	T	T	T	_	2	Orf1a (nsp1)	V→A
9854	C	C	C	C	C	C	C	C	C	C	T	T	T	_	3	Orf1a (nsp1)	A→V
17 564	T	T	T	T	T	T	T	T		G	G	G	G	G	6	Orf1ab (nsp10, helicase)	D→E
19 064	A	A	A	A	A	A	G	A	G	A	A	A	A	A	2	Orf1ab (nsp11)	Silent
19 084	T	T	T	T	C	C	C	C	C	C	C	C	C	C	4	Orf1ab (nsp11)	I→T
19 838	A	A	A	A	A	A	A	A	A	A	G	G	G	G	4	Orf1ab (nsp12)	Silent
21 721	G	G	G	G	G	G	G	G	A	_	_	A	_	_	2	Spike glycoprotein	G→D
22 222	T	T	T	T	T	T	T	T		C	C	C	C	N	5	Spike glycoprotein	I→T
27 243	C	C	C	C	C	C	C	C	C	C	T	T	N	T	3	Putative protein	P→L
27827	T	T	T	T	T	T	T	T		C	C	C	C	C	6	Non-coding	
29279	A	A	A	A	A	A	A	A	A	C	A	C	A	A	2	Nucleocapsid	Q→P
	Linked to Hotel M								No link to Hotel M								

SARS (1) results from outbreaker



Discussion

- You'll hear more about the advent of sequencing - much more data now!
- Outbreaker is now in the broader package 'outbreaker2'
- there was more in these papers that I did not describe
- Both these methods have limitations:
  - they do not capture the shared ancestry of the pathogens
  - for that, we need phylogenetic trees!
  - they do not accommodate variation within hosts

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: bringing in more data
- Research forefronts: SARS-CoV-2 and COVID-19
- Discussion