Non-phylogenetic transmission reconstruction

Now that we understand a little more about genomic data, we can start to think about incorporating it into our analyses

Epidemiological outbreak reconstruction

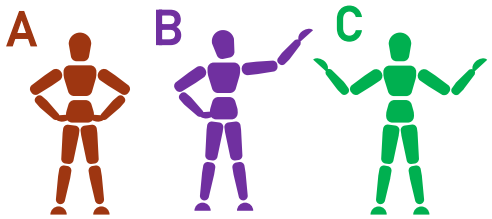
Epidemiological outbreak data alone can be used for outbreak reconstruction, but genetic data offer a high-resolution source of extra information

What can genomic data offer?

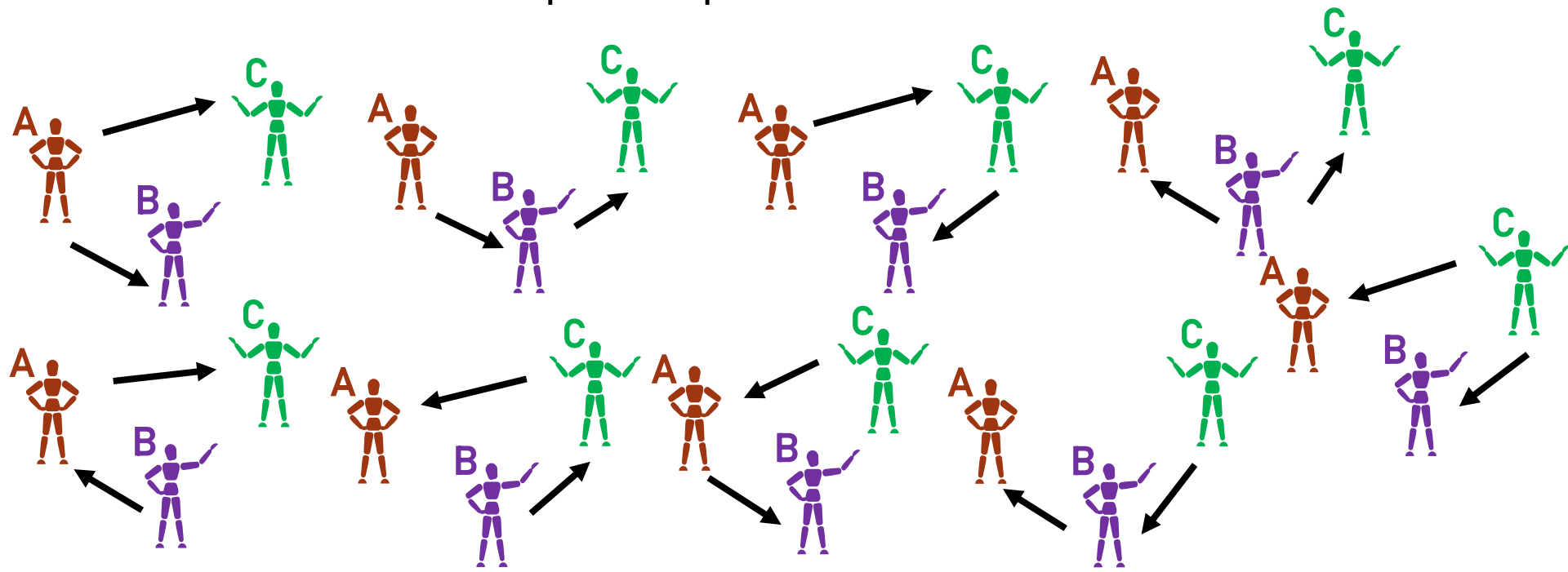
- Extra detail
- Resolve transmission where epi data are hard to get/have gaps
- Genomic data now much easier, cheaper and faster to get than ever before (real-time sequencing even becoming possible in the field)

Challenge: create a single framework/likelihood incorporating genomic + epidemiological data

Imagine we have 3 people infected in an outbreak…



We want to combine our genomic information and our epidemiological information, to best narrow down which possible path the infection took…



Challenge: create a single framework/likelihood incorporating genomic + epidemiological data

Early approaches included:

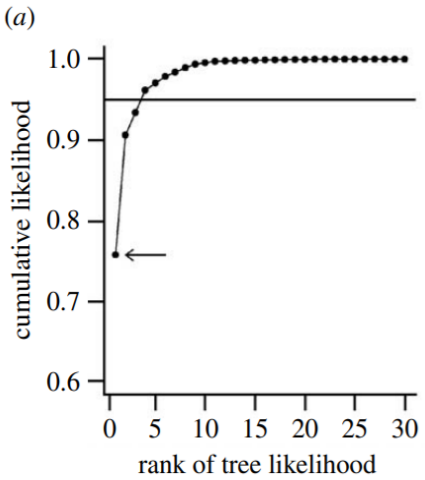
- Maximum likelihood approach
- First, restrict to all transmission trees which are consistent with **known** infections
- Then, use **genomic data to constrain** the set of possible transmission trees
- Then, calculate the **likelihood** of each remaining tree based on the **epi information** – e.g. the chance each individual (a farm, in this case) was infected on a given day/able to infect others on a given day.

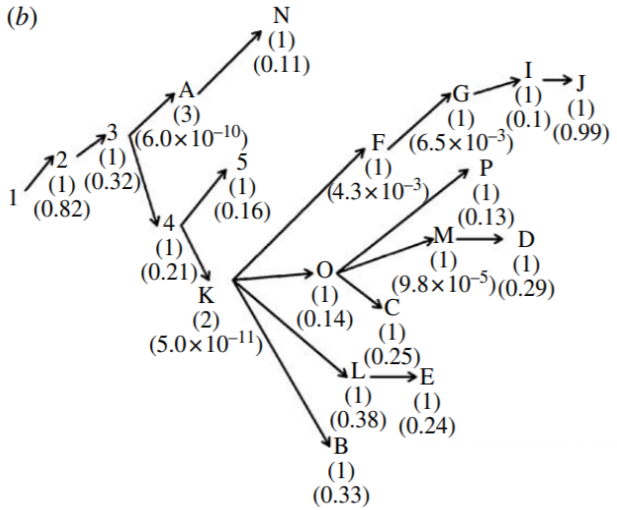
PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 887–895
doi:10.1098/rspb.2007.1442
Published online 29 January 2008

Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus

Eleanor M. Cottam[1,2], Gaël Thébaud[2,†], Jemma Wadsworth[1], John Gloster[3,‡], Leonard Mansley[4], David J. Paton[1], Donald P. King[1] and Daniel T. Haydon[2,*]





Challenge: create a single framework/likelihood incorporating genomic + epidemiological data

Early approaches included:

Heredity (2011) 106, 383–390

www.nature.com/hdy

ORIGINAL ARTICLE

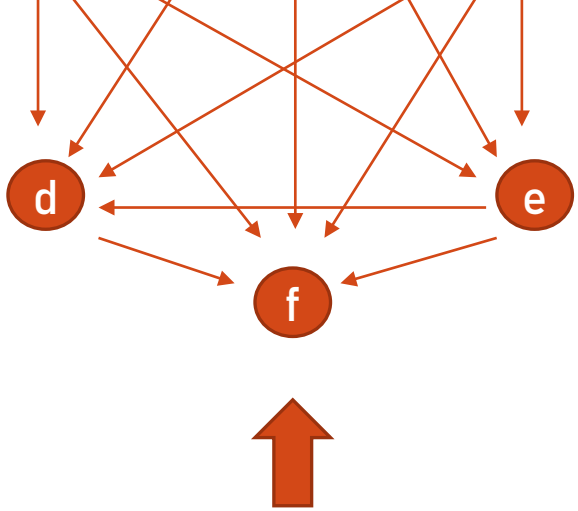
Reconstructing disease outbreaks from genetic data: a graph approach

T Jombart, RM Eggo, PJ Dodd and F Balloux
Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK

- Graph theory approach to find 'genetically parsimonious' transmission trees
- Algorithm *SeqTrack* finds the optimum branching in a directed graph

Distance matrix

	a	b	c	d	e	f
a	0	1	3	2	5	9
b		0	2	4	7	5
c			0	1	4	12
d				0	1	3
e					0	8
f						0



(i) Create a connected, directed graph with weights w_{ij} equal to the genetic distance
(ii) Remove edge ij if $t_j < t_i$
(iii) Find the spanning directed tree optimizing (i.e. minimizing) $\sum w_{ij}$

Sample collection dates:

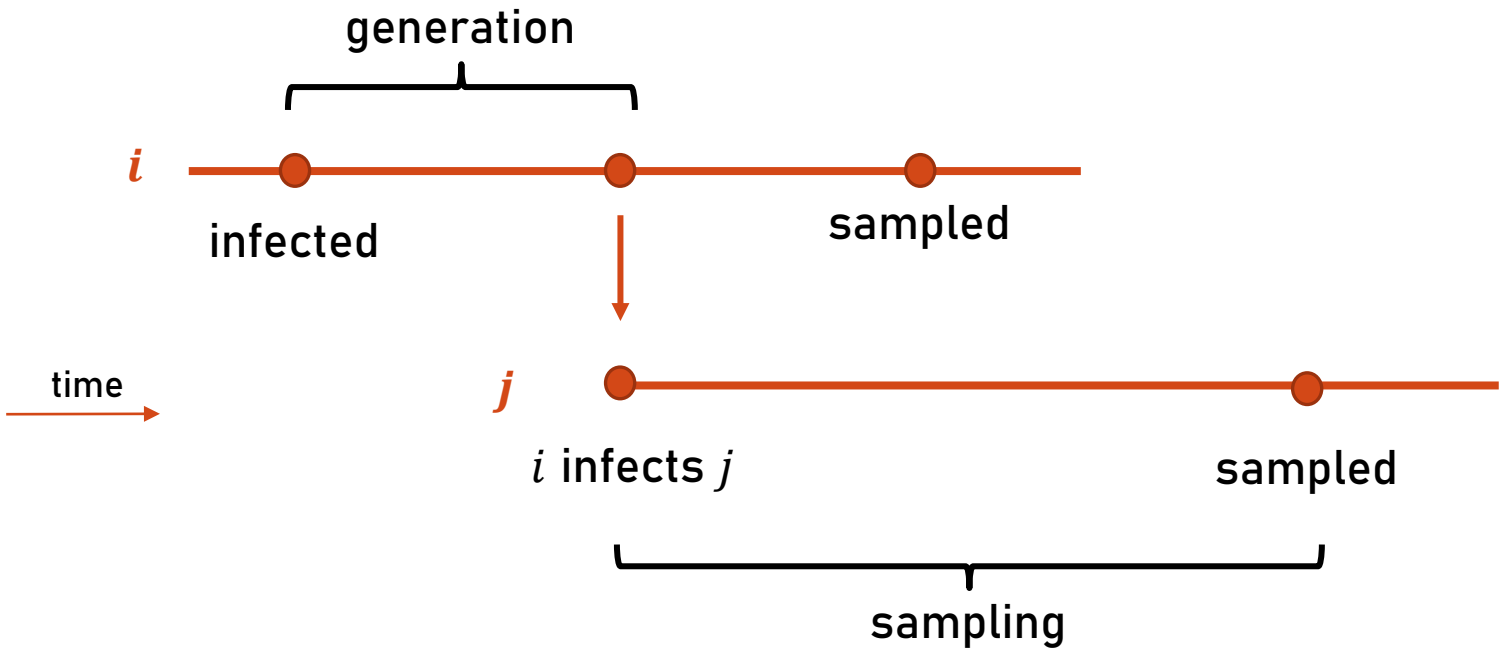
a: t=3
b: t=1
c: t=2
d: t=5
e: t=4
f: t=7

Challenge: create a single framework/likelihood incorporating genomic + epidemiological data

Some limitations:
- All cases come from single index case e.g. a single sampled ancestor
- All cases are known and sampled

SeqTrack also:
- Assumes that individuals became infectious in the order they are sampled
- Has no uncertainty in the output transmission tree or probabilistic parameters

A quick primer 1: generation time and sampling time



Generation time = the time interval between the infection of an individual and their seeding of new secondary cases.
Sampling time = the time interval between infection and collection of an isolate.

A quick primer 2: Markov Chain Monte Carlo (MCMC)

A popular method for exploring complex and/or high-dimensional spaces – e.g. transmission trees!

The main idea:

likelihood

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

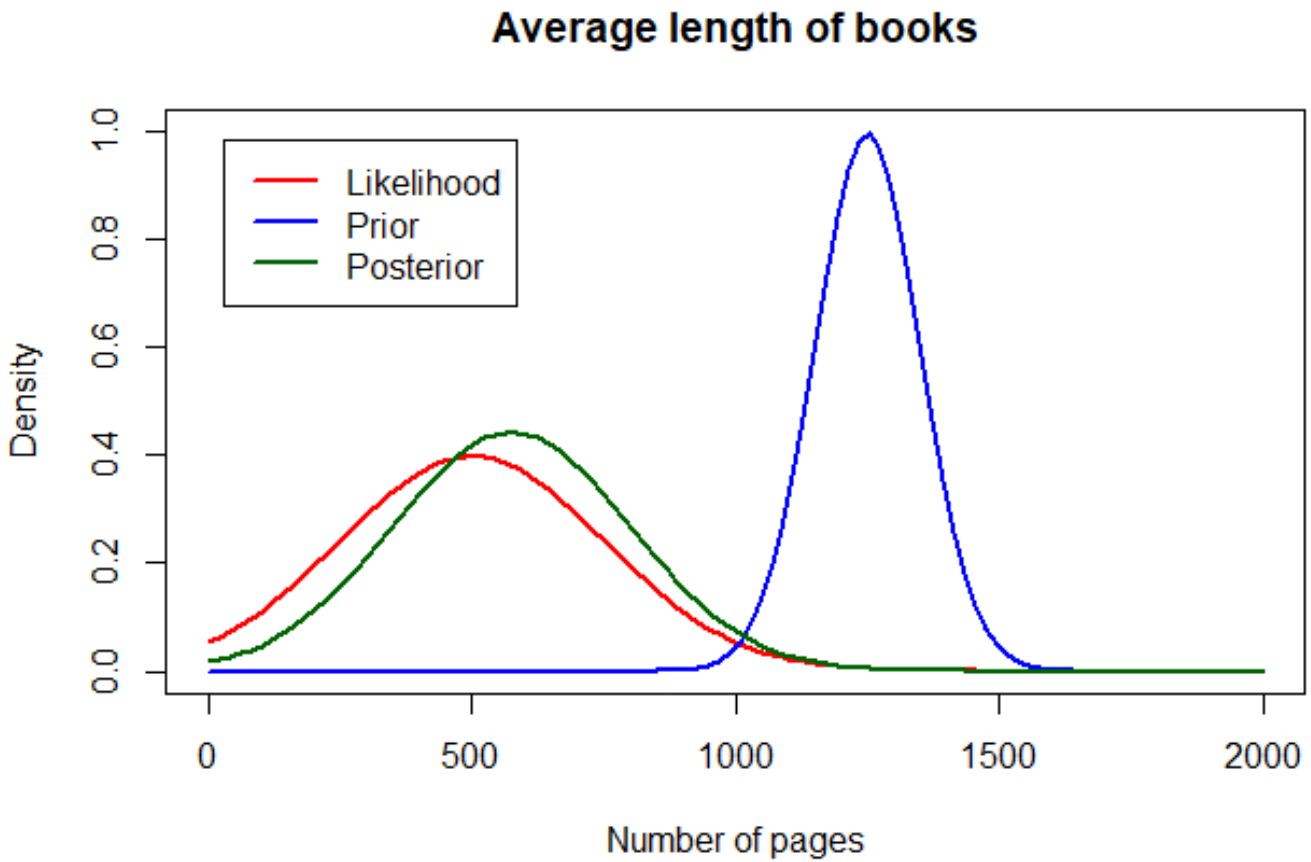
Posterior – the probability of our model parameters given the data

prior

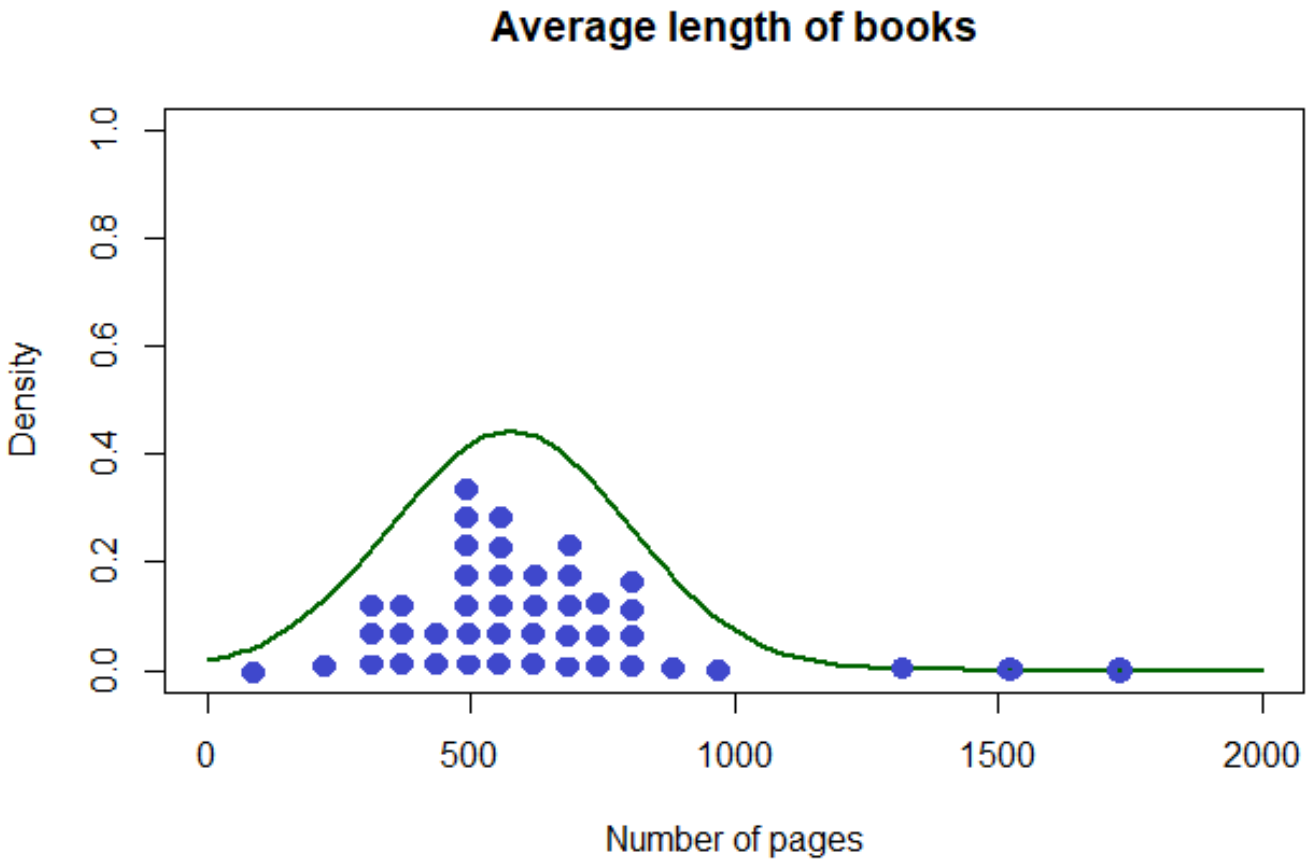
When this quantity is hard to e.g. maximise directly, we instead form a Markov chain with equilibrium distribution equal to the posterior distribution, and then take many samples from this chain.

Essentially, we approximate the posterior distribution by random sampling from a probabilistic space (of all possible transmission trees).

A quick primer 2: Markov Chain Monte Carlo (MCMC)



A quick primer 2: Markov Chain Monte Carlo (MCMC)



A quick primer 2: Markov Chain Monte Carlo (MCMC)

Data-augmented MCMC is a method for dealing with missing data within an MCMC algorithm. As well as sampling from the parameter space at each step of the Markov chain, we also sample values for the missing data.

In transmission inference, missing data might be the time of infection of the cases (since typically we only know sampling times) or the number of unsampled cases, for example.

outbreaker and outbreaker2

We're going to look at these in detail – and will be using them in the next exercise

Also creates a unified likelihood for genetic + epidemiological data, within a Bayesian framework which allows more estimation and greater flexibility.

outbreaker vs outbreaker2

outbreaker2 essentially a more customisable version of outbreaker

We're mainly going to focus on the core outbreaker model…

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$\mathrm{d}(s_i, s_j)$ = number of mutations (distance) between sequences i and j
$\mathrm{l}(s_i, s_j)$ = number of nucleotide positions which can be compared i and j
w = distribution of the generation time
f = distribution of the sampling time

```
> 1:1999-08-01
GCACCCATTCCCGCCTGGAGAT
> 2:2007-11-01
GCACCCATTCCCGCCTAGAGAT
```

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data:

N sampled cases, each with genetic sequence s_i and time of sampling t_i

Quantities:

$\mathrm{d}(s_i, s_j)$ = number of mutations (distance) between sequences i and j
$\mathrm{l}(s_i, s_j)$ = number of nucleotide positions which can be compared i and j
w = distribution of the generation time
f = distribution of the sampling time

Augmented data:

α_i = index of the most recent sampled ancestor of i
κ_i = number of (unsampled) generations between i and α_i
T_i^{inf} = date of infection of i

Parameters:

μ = mutation rate, per site per generation of infection
π = proportion of unsampled cases
are estimated as well as the transmission tree

Posterior distribution:

$$P(A,\theta\,|D) = \frac{P(D,A\,|\theta)P(\theta)}{P(D)} \propto p\left(\left\{s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\right\}_{i=1,\ldots,N}\,\middle|\,\mu,\pi\right)\times p(\mu,\pi).$$

All cases are assumed to be conditionally independent, given the identity of their most recent sampled ancestor, so the likelihood decomposes to:

$$p\left(\left\{s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\right\}_{i=1,\ldots,N}\,\middle|\,\mu,\pi\right) = \prod_{i=2}^{N} p\left(s_i,t_i,\alpha_i,\kappa_i,T_i^{\inf}\,|s_{\alpha_i},t_{\alpha_i},T_{\alpha_i}^{\inf},\mu,\pi\right)\times$$
$$p(t_1\,|T_1^{\inf})p(s_1)p\left(T_1^{\inf}\right)p(\alpha_1)p(\kappa_1)$$

The pseudo-likelihood is further decomposed into genetic and epidemiological components. For each case $i = 1, \ldots, \mathrm{N}$:

$$p\left(s_i, t_i, \alpha_i, \kappa_i, T_i^{\inf} | s_{\alpha_i}, t_{\alpha_i}, T_{\alpha_i}^{\inf}, \mu, \pi\right)$$
$$= p\left(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu\right) \times p\left(t_i | T_i^{\inf}\right) p\left(T_i^{\inf} \middle| \alpha_i, T_{\alpha_i}^{\inf}, \kappa_i\right) p(\kappa_i | \pi) p(\alpha_i)$$

Genetic part

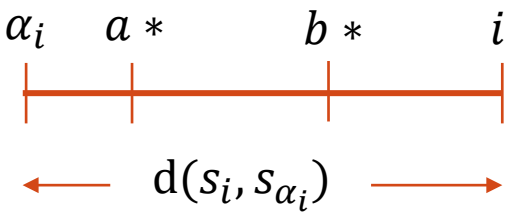
Epidemiological part

constant

Genetic part

The outbreaker genetic model assumes no within-host genetic diversity, and so mutations are direct features of transmission events. All transmission events are assumed independent, and the genetic pseudo-likelihood is very fast to compute.

Genetic pseudo-likelihood of case i = the probability of observing genetic distance $\mathrm{d}(s_i, s_{\alpha_i})$ between sequence s_i and the ancestral sequence s_{α_i} with i and α_i separated by κ_i generations.



As a method designed for shorter timescale outbreaks, reverse mutations are considered negligible.

$$\mu^{\mathrm{d}(s_i, s_{\alpha_i})} (1-\mu)^{\kappa_i \times \mathrm{l}\left(s_i, s_{\alpha_i}\right) - \mathrm{d}(s_i, s_{\alpha_i})}$$

Epidemiological part

Time of sampling given time of infection

Time of infection given knowledge of infector

Number of missing cases given rate of missing cases

$$p\left(t_i \mid T_i^{\inf}\right) p\left(T_i^{\inf} \middle| \alpha_i, T_{\alpha_i}^{\inf}, \kappa_i\right) p(\kappa_i \mid \pi)$$

$$=$$

$$f\left(t_i - T_i^{\inf}\right) \times w^{\kappa_i}\left(T_i^{\inf} - T_{\alpha_i}^{\inf}\right) \times \mathrm{NB}(1 \mid \kappa_i - 1, \pi)$$

probability of obtaining one 'success' (sampling a case) after $\kappa_i - 1$ 'failures' (unobserved cases), with probability of success π.

That forms the core of the outbreaker model.

The likelihood expressions introduced in the previous slides are combined with priors for the mutation rate μ and proportion of unsampled cases π.

μ is given a uniform prior on [0,1] – corresponding to an assumption of scarce prior information on this
π is given a beta distributed prior with parameters controlled by the user of outbreaker. This is a flexible prior which can reflect different levels of prior knowledge for different datasets.

The authors also introduce a method for detecting **imported cases** – i.e. cases that are not descended from another case in the outbreak.

In an initial step of the model, genetic outliers are detected, relative to the other samples in the dataset. A 'global influence' GI_i is calculated for each sampled case, defined as

$$GI_i = \mathbb{E}\left(\sum_{j=1, j\neq i}^{n} GPL_J\right) - \mathbb{E}\left(\sum_{i=1}^{n} GPL_i\right)$$

where GPL is the genetic pseudo-likelihood. This is calculated over the first few samples of the MCMC, say 50.

A large value of the GI_i implies unlikely numbers of mutations i.e. a 'distant' sequence. Cases with a global influence more than 5 times the average across all cases are considered outliers.

An application from

OPEN ACCESS Freely available online

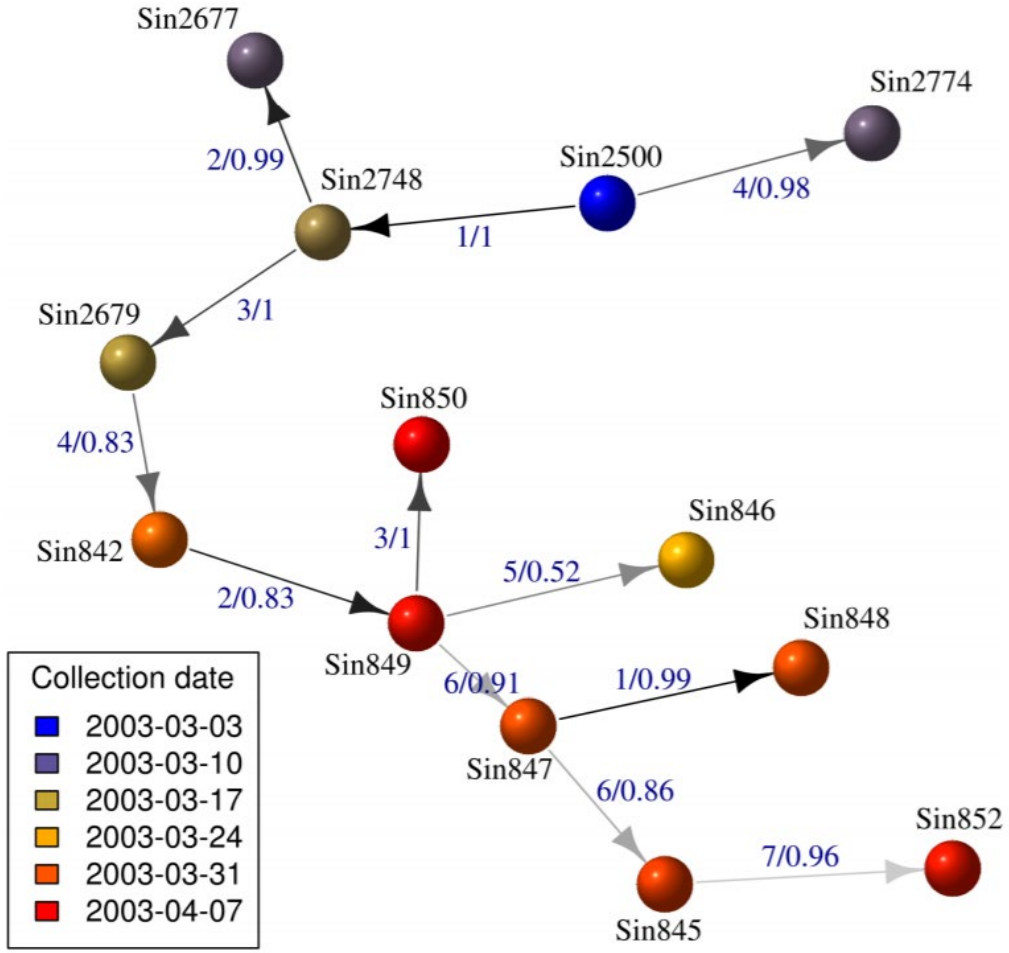
PLOS | COMPUTATIONAL BIOLOGY

Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data

Thibaut Jombart*, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser*, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, United Kingdom

Data from 2003 Singaporean Severe Acute Respiratory Syndrome (SARS) outbreak.
13 genomes with <15 mutations between all pairs.



outbreaker2: extensions

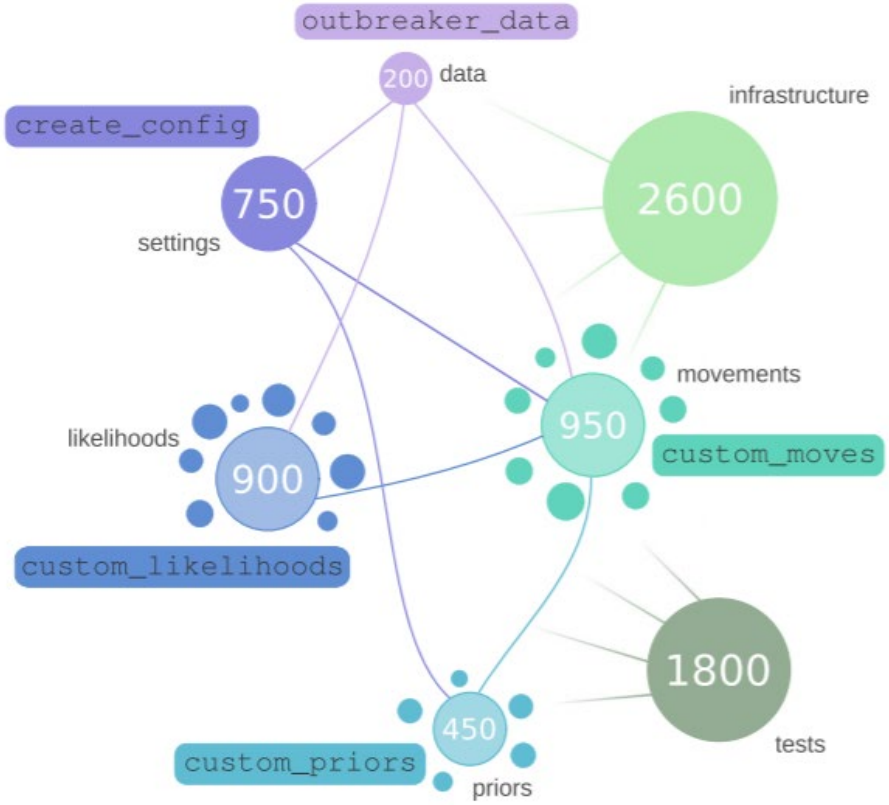
Campbell *et al. BMC Bioinformatics* 2018, **19**(Suppl 11):363
https://doi.org/10.1186/s12859-018-2330-z

BMC Bioinformatics

SOFTWARE Open Access

outbreaker2: a modular platform for outbreak reconstruction

Finlay Campbell, Xavier Didelot, Rich Fitzjohn, Neil Ferguson, Anne Cori and Thibaut Jombart*



- Combines an R package with C++ code for efficiency, through Rcpp
- Can customise all these facets of the package
- For example, they implemented the TransPhylo methodology, which we will work with tomorrow, in outbreaker2.