

outbreakeR

disease outbreak reconstruction using genetic data

Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling – Imperial College London

GenEpi — LSHTM, London
10-04-2013

Investigating disease outbreaks using genetic data

Background

- here, “*outbreak*” = small, localised epidemic
- **small-scale → dense sampling** possible
- **pathogen genomes sequences increasingly available**



Investigating disease outbreaks using genetic data

Background

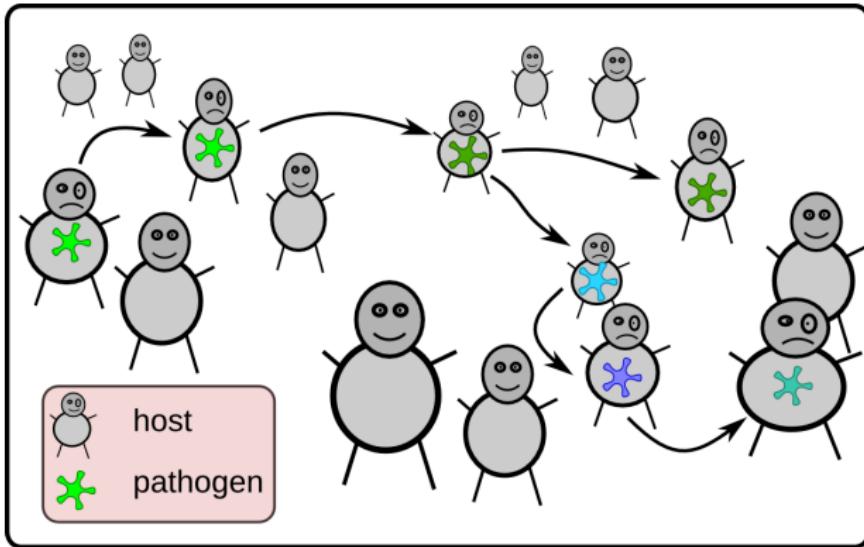
- here, “*outbreak*” = small, localised epidemic
- **small-scale** → **dense sampling** possible
- **pathogen genomes sequences increasingly available**



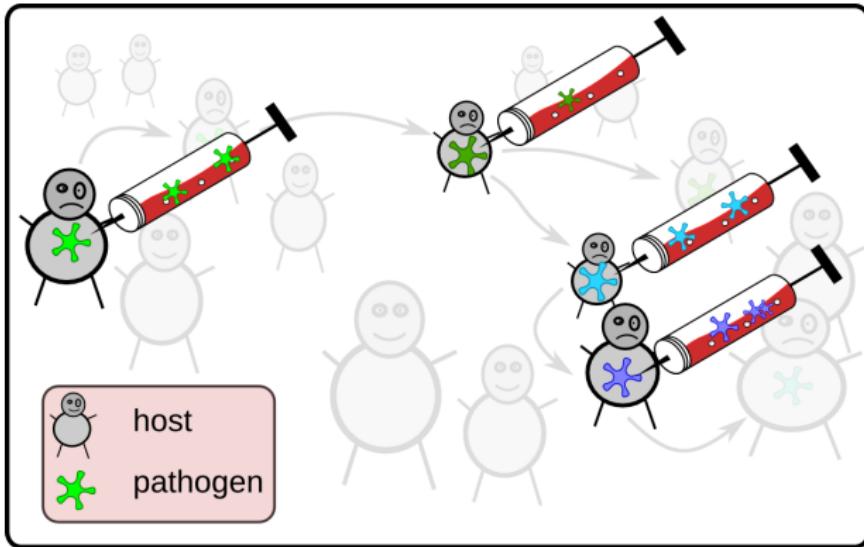
Objectives

- exploit genetic information to reconstruct outbreaks
- infer **transmission trees**, dates of infection, R , ...
- design a tool for **retrospective** or **near real-time** analysis

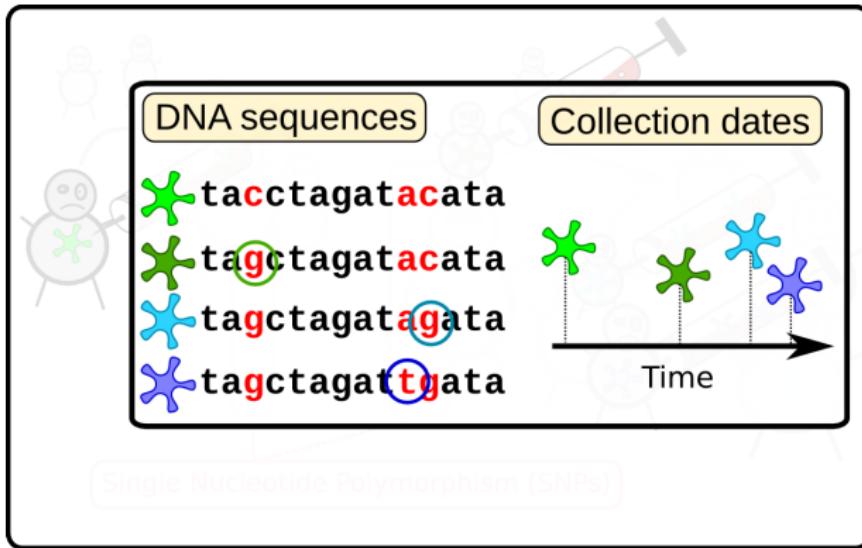
The data



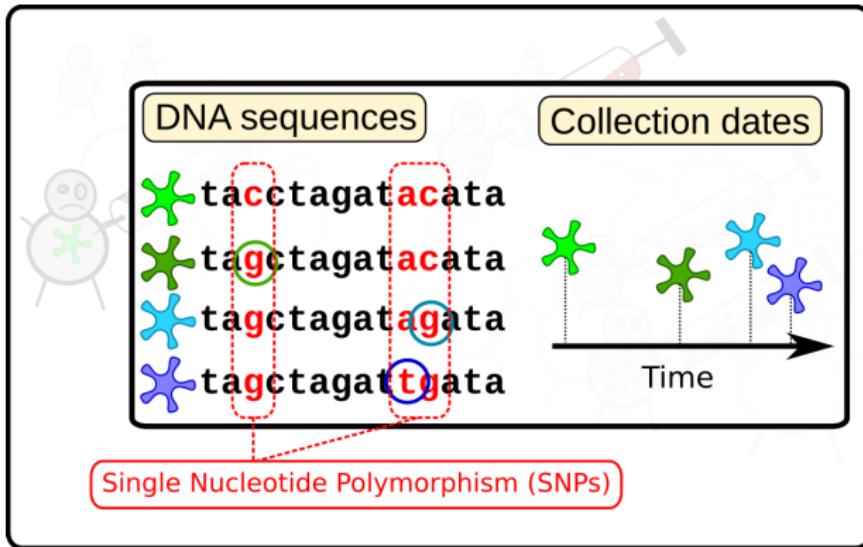
The data



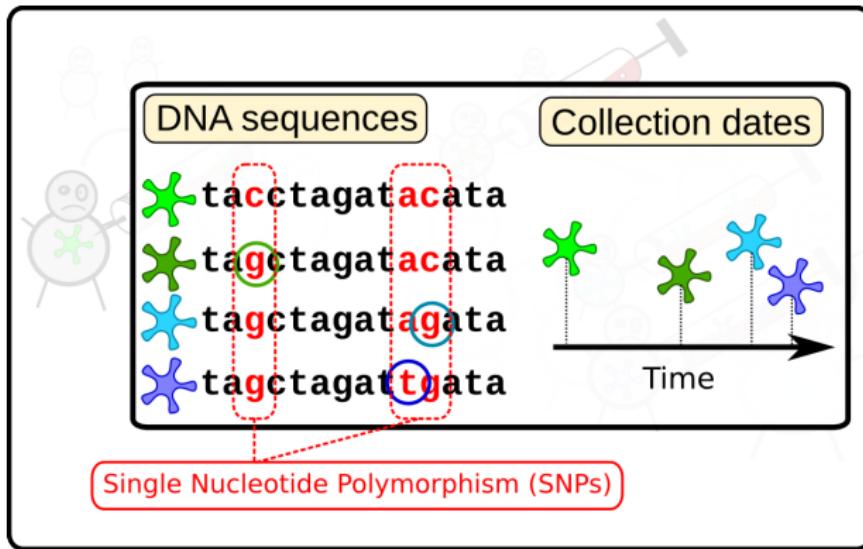
The data



The data



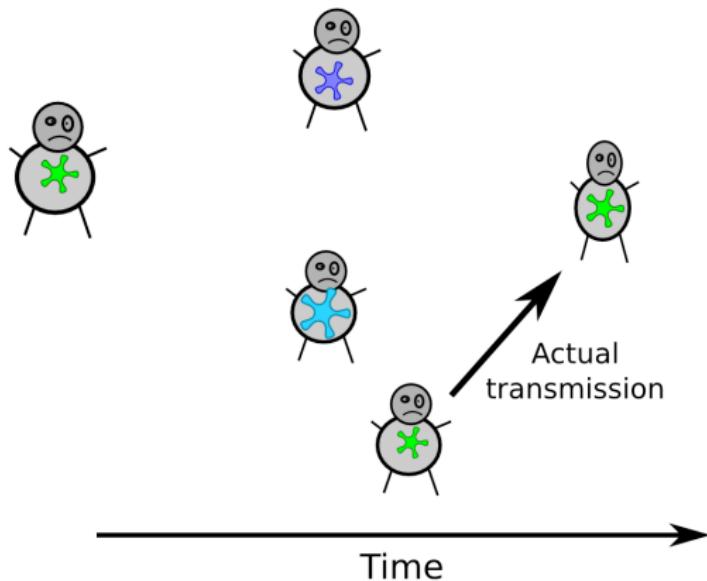
The data



Generic model integrating **genomic data** and **collection dates**

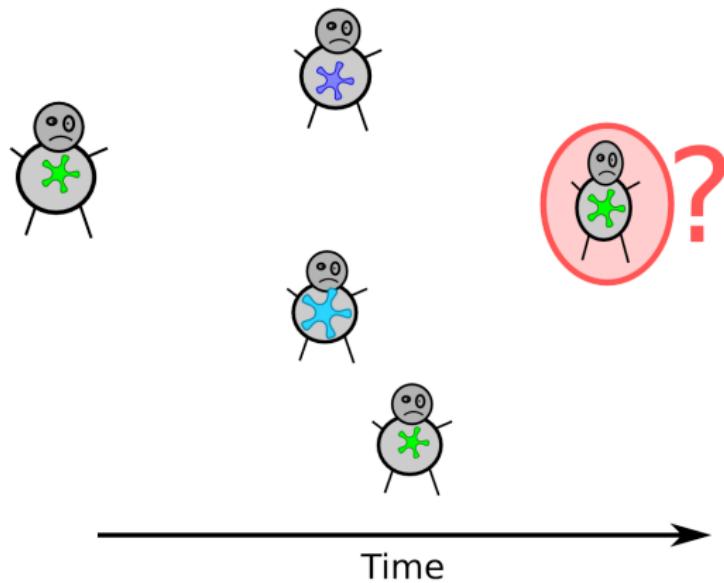
Rationale of the method

How to infer the infector of a given individual?



Rationale of the method

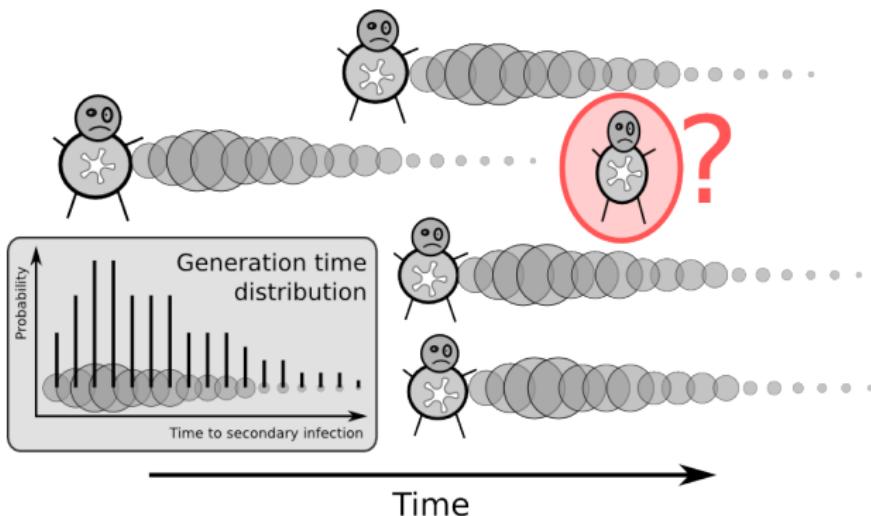
How to infer the infector of a given individual?



Rationale of the method

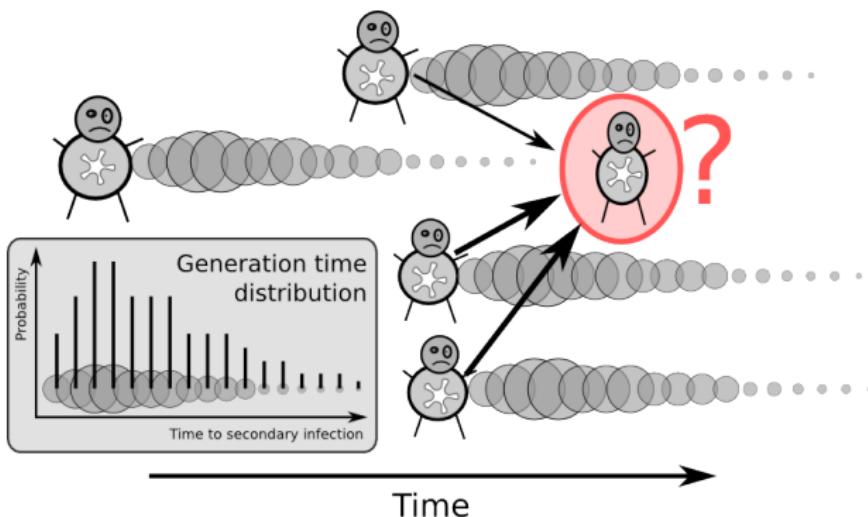
Approach based on generation time distribution

(Ferguson *et al.* 2001, Nature; Wallinga & Teunis 2004, Am. J. Epidemiol.)



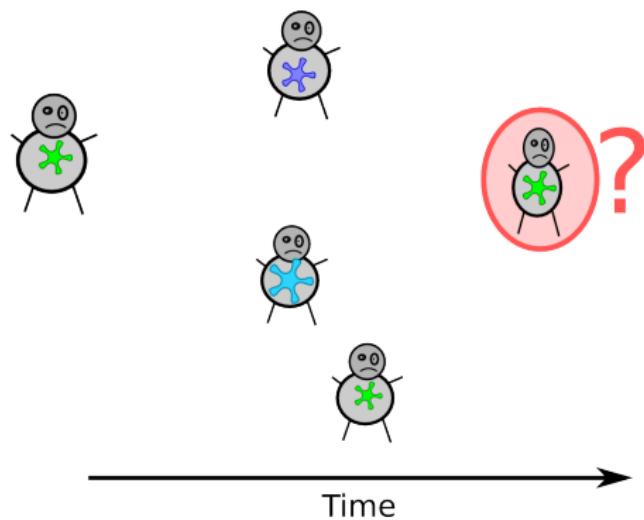
Rationale of the method

Approach based on generation time distribution
(Wallinga & Teunis 2004, Am. J. Epidemiol.)



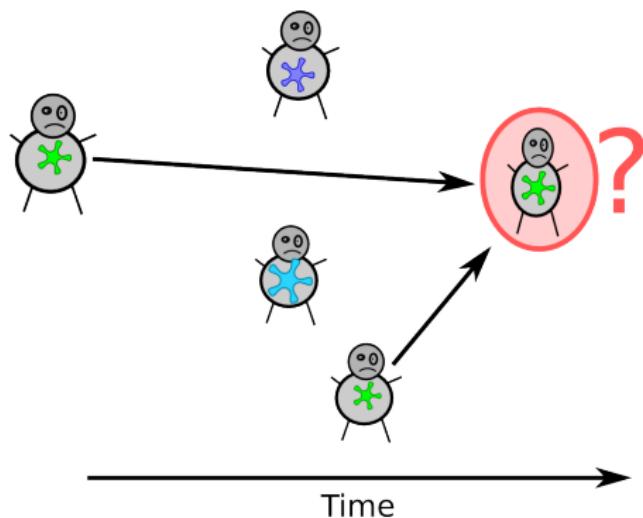
Rationale of the method

Approach based on genetic data (*SqTrack*)
(Jombart *et al.* 2010, Heredity.)



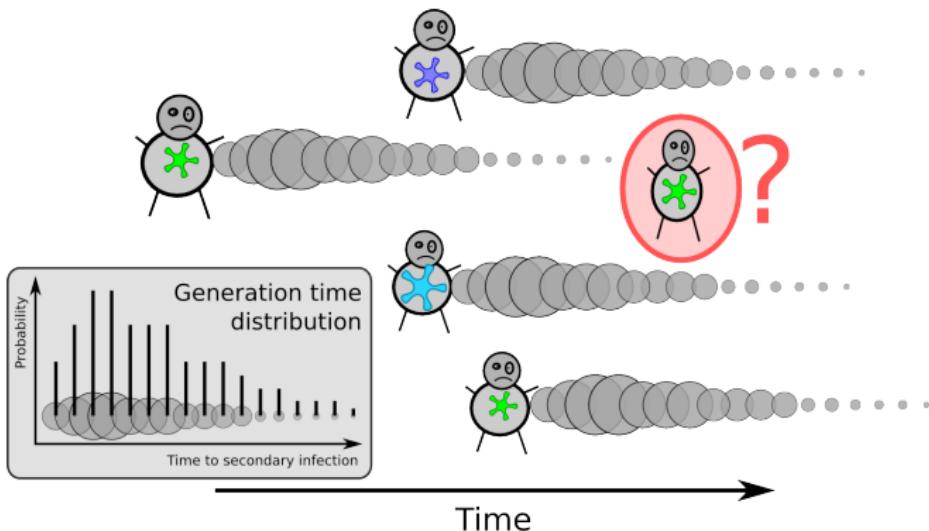
Rationale of the method

Approach based on genetic data (*SqzTrack*)
(Jombart *et al.* 2010, Heredity.)



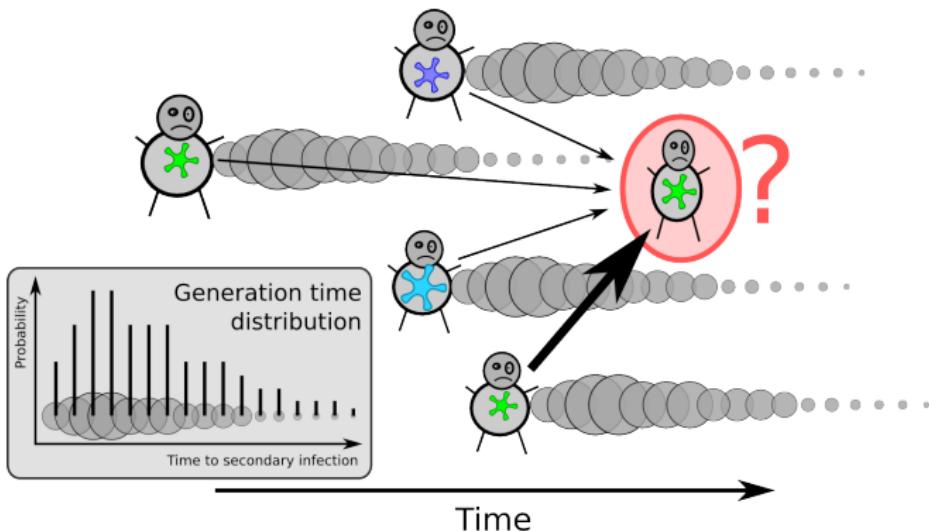
Rationale of the method

outbreaker: use generation time distribution and genetic data:



Rationale of the method

outbreaker: use generation time distribution and genetic data:



Outline of the model

Likelihood

- branching process:

$$p(\text{transmission tree}) = \prod_{\text{all branches}} p(\text{branches})$$

- $p(\text{branch}) = p(\text{infection/collection dates}) \times p(\text{genetic differences})$

Outline of the model

Likelihood

- branching process:

$$p(\text{transmission tree}) = \prod_{\text{all branches}} p(\text{branches})$$

- $p(\text{branch}) = p(\text{infection/collection dates}) \times p(\text{genetic differences})$

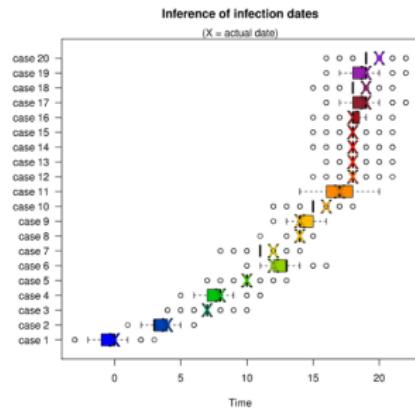
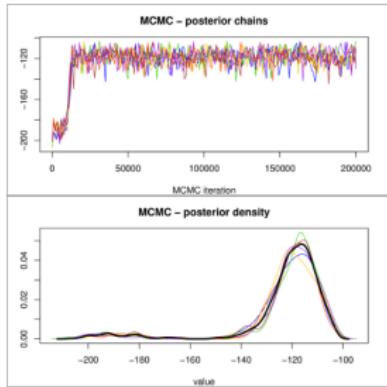
Implementation

- Bayesian framework
- augmented data for ancestries and unobserved cases
- MCMC (Metropolis-Hastings) for sampling from posterior

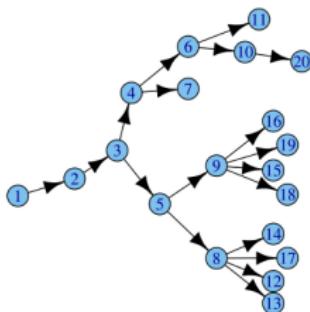
The package `outbreakeR`

- C implementation embedded within  package
- multi-platform: linux, MacOS X, Windows, Solaris, . . .
- supports parallelization
- post-processing of MCMC, simulations, graphics
- tested on wide range of simulations

outbreaker : some examples

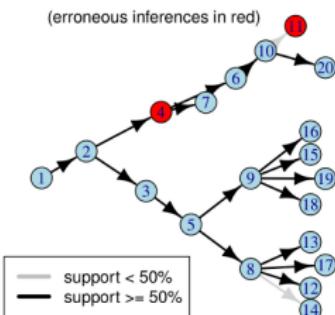


Simulated data



Reconstructed outbreak

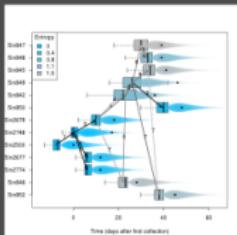
(erroneous inferences in red)



Looking ahead: applications and developments

Applications

SARS 2003 Singapore outbreak

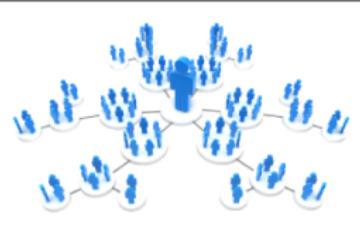


S. pneumoniae carriage Maela camp, Thailand



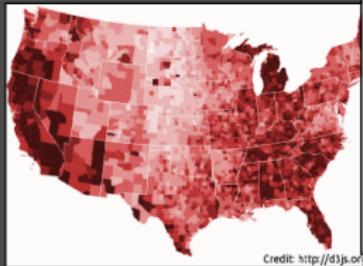
Model improvement

Spatial information, contact networks

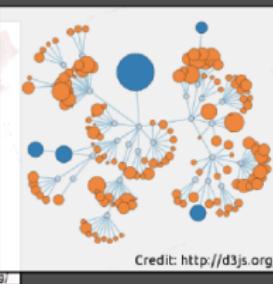


New visualization tools

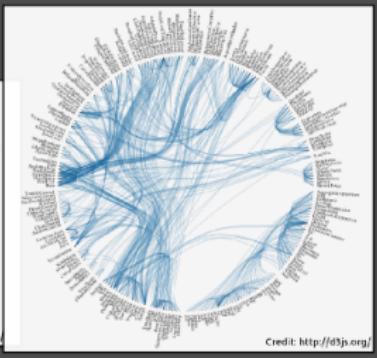
Web-based interactive graphics



Credit: <http://d3js.org>



Credit: <http://d3js.org>



Credit: <http://d3js.org/>

Looking ahead: the bigger picture

The collage consists of five screenshots arranged in a grid-like fashion:

- outbreaker**: A web-based application for disease outbreak analysis. It features sections for "Applications" (SARS 2003 Singapore outbreak, S. pneumoniae carriage Maela camp, Thailand) and "Model improvement" (Spatial information, contact networks). The main interface shows a map of the United States with red dots indicating outbreaks and a network graph below it.
- EpiEstim**: A package for estimating the basic reproduction number and the effective reproduction number over time. The screenshot shows the package's description page on CRAN, highlighting its use for real-time data analysis to inform public health responses.
- Hackout**: A hackathon for the analysis of disease outbreaks in R. The screenshot shows the event's website with a large "R" logo, logos for MRC, Imperial College London, and SourceForge, and a photo of several people working at laptops.
- epibase**: A package for the analysis of disease outbreaks based on epidemiological and genetic data. The screenshot shows the package's description page on CRAN, featuring a large "epibase" logo and a photo of researchers working on laptops.
- ... (three placeholder boxes)**: Three empty rectangular boxes with three ellipsis symbols each, likely placeholders for additional tools or packages.

Acknowledgements

- **Organisers:** Anton Camacho & Marc Baguelin
- **Imperial College London:** Anne Cori, Simon Cauchemez, Xavier Didelot, Christophe Fraser, Neil Ferguson, David Aanensen
- **Wellcome Trust Sanger Institute:** Claire Chewapreecha, Sephen Bentley
- **Funding:** MIDAS
- **Thanks for your attention.**

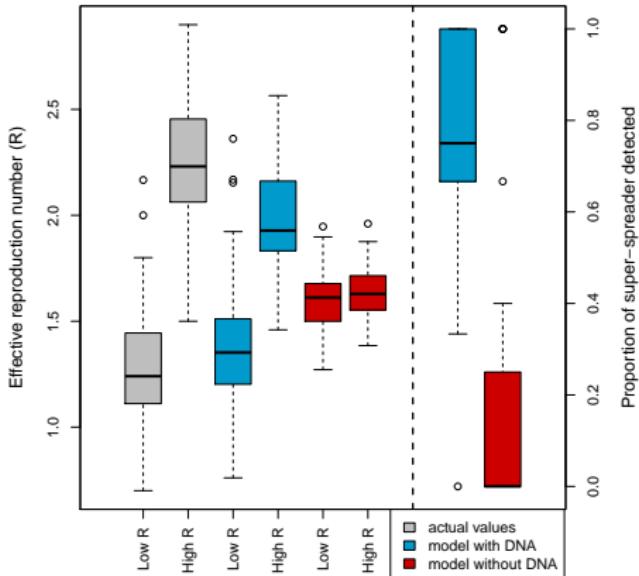
Getting your hands dirty

Google “adegenet” → “adegenet on the web” → “Documents”
(<http://adegenet.r-forge.r-project.org/>)

Genetic data reveal heterogeneous infectivity

Simulations with structured infectivity

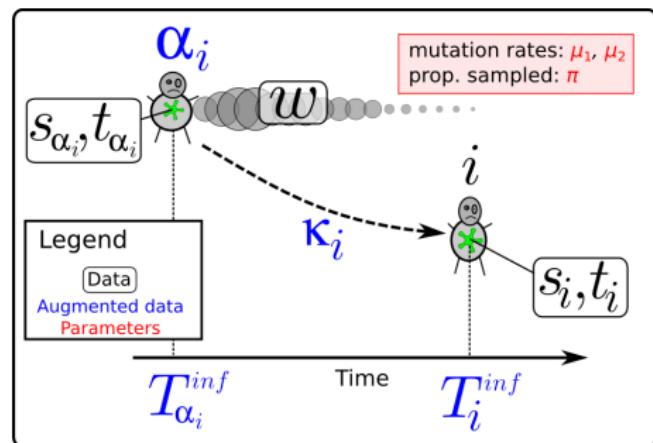
- left: 2 groups of hosts, low/high infectivity ($R_0 = 1.5/3$)
- right: rare super-spreaders ($R_0 = 1.5/20$)



outbreaker : model notations

Data

- i : index of cases ($i = 1, \dots, n$)
- s_i : genetic sequence of i
- t_i : collection date for s_i
- w : generation time distribution



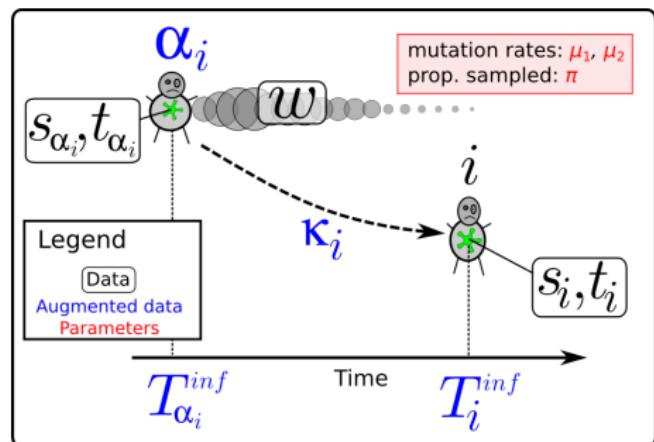
outbreaker : model notations

Data

- i : index of cases ($i = 1, \dots, n$)
- s_i : genetic sequence of i
- t_i : collection date for s_i
- w : generation time distribution

Augmented data

- T_i^{inf} : date of infection of i
- α_i : infector of i
- κ_i : number of generations between i and α_i



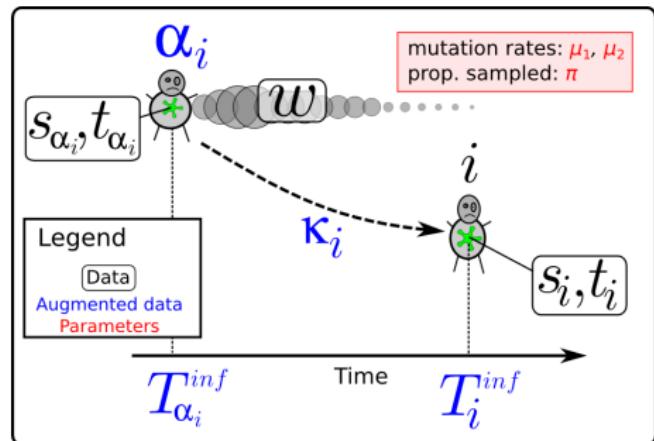
outbreaker : model notations

Data

- i : index of cases ($i = 1, \dots, n$)
- s_i : genetic sequence of i
- t_i : collection date for s_i
- w : generation time distribution

Augmented data

- T_i^{inf} : date of infection of i
- α_i : infector of i
- κ_i : number of generations between i and α_i



Parameters

- μ_1, μ_2 : rates of transitions and transversions ($\mu_2 = \gamma \times \mu_1$)
- π : proportion of the outbreak sampled

outbreakeR : model definition

- Posterior proportional to joint distribution:

$$\begin{aligned} & p(\{s_i, t_i, T_i^{inf}\}_{(i=1,\dots,n)}, \alpha, \kappa, w, \mu_1, \gamma, \pi) \\ = & \underbrace{p(\{s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i\}_{(i=1,\dots,n)} | w, \mu_1, \gamma, \pi)}_{likelihood} \times \underbrace{p(w, \mu_1, \gamma, \pi)}_{prior} \end{aligned}$$

outbreakeR : model definition

- Posterior proportional to joint distribution:

$$\begin{aligned} & p(\{s_i, t_i, T_i^{inf}\}_{(i=1,\dots,n)}, \alpha, \kappa, w, \mu_1, \gamma, \pi) \\ = & \underbrace{p(\{s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i\}_{(i=1,\dots,n)} | w, \mu_1, \gamma, \pi)}_{likelihood} \times \underbrace{p(w, \mu_1, \gamma, \pi)}_{prior} \end{aligned}$$

- Likelihood of case i decomposed as:

$$\underbrace{p(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu_1, \gamma)}_{genetic} \times \underbrace{p(t_i | T_i^{inf}, w) p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) p(\kappa_i | \pi)}_{epidemiological}$$

outbreakeR : model definition

- Posterior proportional to joint distribution:

$$\begin{aligned} & p(\{s_i, t_i, T_i^{inf}\}_{(i=1,\dots,n)}, \alpha, \kappa, w, \mu_1, \gamma, \pi) \\ = & \underbrace{p(\{s_i, t_i, T_i^{inf}, \alpha_i, \kappa_i\}_{(i=1,\dots,n)} | w, \mu_1, \gamma, \pi)}_{likelihood} \times \underbrace{p(w, \mu_1, \gamma, \pi)}_{prior} \end{aligned}$$

- Likelihood of case i decomposed as:

$$\underbrace{p(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu_1, \gamma)}_{genetic} \times \underbrace{p(t_i | T_i^{inf}, w) p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) p(\kappa_i | \pi)}_{epidemiological}$$

- Sampling from posterior distribution using MCMC (Metropolis-Hasting)

Model - detail of likelihoods

- Genetic likelihood:

$$\underbrace{\mathcal{B}(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i}) \kappa_i, \mu_1)}_{transitions} \times \underbrace{\mathcal{B}(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i}) \kappa_i, \gamma \mu_1)}_{transversions}$$

Model - detail of likelihoods

- Genetic likelihood:

$$\underbrace{\mathcal{B}(d(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i}) \kappa_i, \mu_1)}_{transitions} \times \underbrace{\mathcal{B}(g(s_i, s_{\alpha_i}) | l(s_i, s_{\alpha_i}) \kappa_i, \gamma \mu_1)}_{transversions}$$

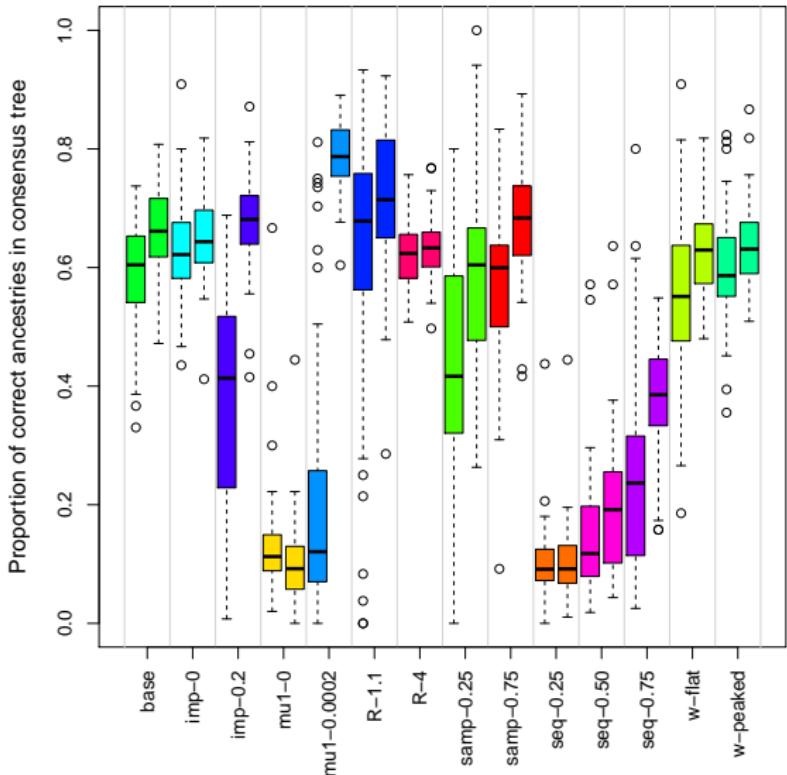
- Epidemiological likelihood:

$$p(t_i | T_i^{inf}, w) \times p(T_i^{inf} | \alpha_i, T_{\alpha_i}^{inf}, \kappa_i, w) \times p(\kappa_i | \pi) \\ w(t_i - T_i^{inf}) \times w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf}) \times f_{NB}(1 | \kappa_i - 1, \pi)$$

with:

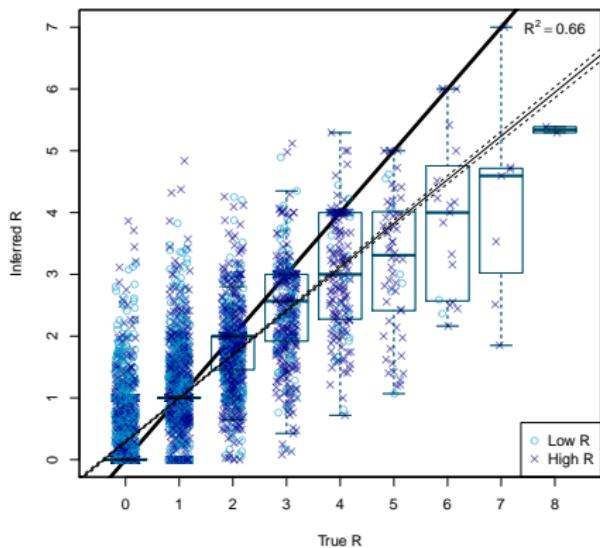
- $w(t_i - T_i^{inf})$: probability of sampling date (assumed prop. to infectiousness)
- $w^{(\kappa_i)}(T_i^{inf} - T_{\alpha_i}^{inf})$: probability of infection date ($w^{(\kappa_i)}$ denotes κ_i convolutions of w)
- $f_{NB}(1 | \kappa_i - 1, \pi)$: probability of $\kappa_i - 1$ unsampled cases (f_{NB} : density of the Negative Binomial distribution)

Simulation results: inference of correct ancestries

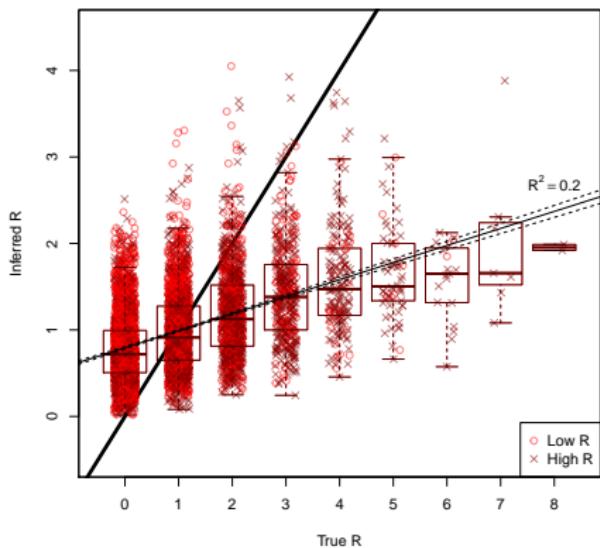


Inferring heterogeneous R : 2-groups simulations

With genetic data:

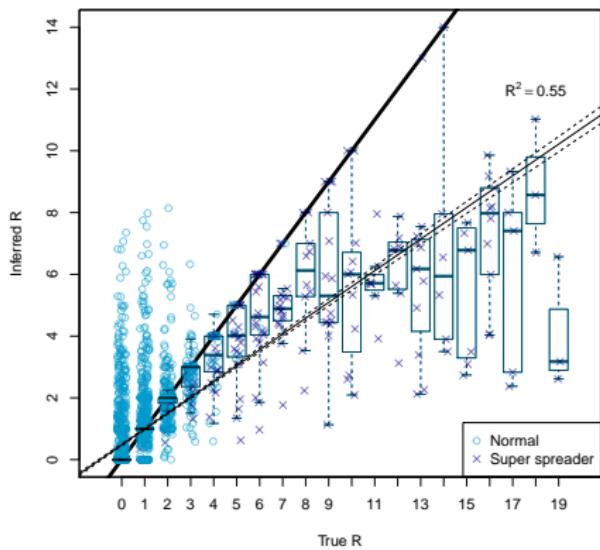


Without genetic data:



Inferring heterogeneous R : super-spreader simulations

With genetic data:



Without genetic data:

