Recent advances on integrating epidemiological and whole genome sequence data for analysing infectious disease outbreak data.

Theo Kypraios

Mathematisches Forschungsinstitut Oberwolfach, February, 2023

School of Mathematical Sciences @ University of Nottingham

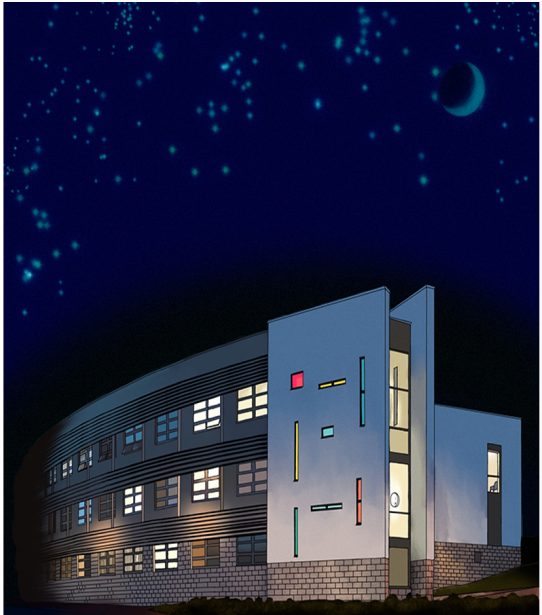
Acknowledgements

Joseph Marsh

Phil O'Neill

University of Nottingham

UK | CHINA | MALAYSIA



Motivation

- We consider disease outbreaks in settings where genomic sampling is done.
- We formulate stochastic epidemic models to investigate person-to-person transmission, based on observed genomic and epidemiological data.
- We develop computational statistics methodology for fitting these models to data within a Bayesian framework.
- Could be useful from both an infection control point of view (e.g. identify super spreading events, who-infected-whom) and a statistical point of view (e.g. reduce uncertainty in transmission rates).

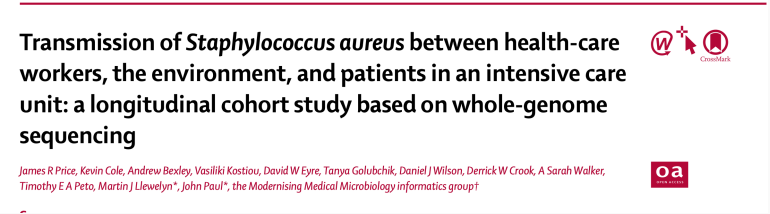
Outline

- Motivating dataset
- Modelling [epi + genetic data]
- Bayesian inference
- Results

Motivation

Motivating Dataset

- The data are taken from an Intensive Care Unit (ICU) and high-dependency unit (HDU) of the Royal Sussex County Hospital, a large acute hospital in Brighton[1].
- The pathogens of interest are Methicillin Resistant Staphylococcus Aureus (MRSA) and Meticillin-susceptible Staphylococcus Aureus (MSSA).
- Data tell us about MRSA and MSSA colonisation status of patients, healthcare workers and environment.



Transmission of *Staphylococcus aureus* between health-care workers, the environment, and patients in an intensive care unit: a longitudinal cohort study based on whole-genome sequencing

James R Price, Kevin Cole, Andrew Bexley, Vasiliki Kostiou, David W Eyre, Tanya Golubchik, Daniel J Wilson, Derrick W Crook, A Sarah Walker, Timothy E A Peto, Martin J Llewelyn, John Paul*, the Modernising Medical Microbiology informatics group†*

[1] Courtecy of James Price

Motivating Dataset

- Systematically sampled health-care workers, the environment, and patients over 14 months.
- Nasal swabs were taken from
  - health-care workers every 4 weeks,
  - bed spaces were sampled monthly,
  - and screening swabs were obtained from patients at admission, weekly thereafter, and at discharge.
- Isolates were cultured for MRSA and MSSA and their whole genome sequenced.

Data Overview

Patient-level data consisting of:

- Admission and discharge time;
- Dates and outcomes of any screening tests;
- Some sequenced isolates;
- Other clinical information.

Healthcare worker level data consisting of:

- Dates with positive tests that were sequenced.

Epi Data Summary

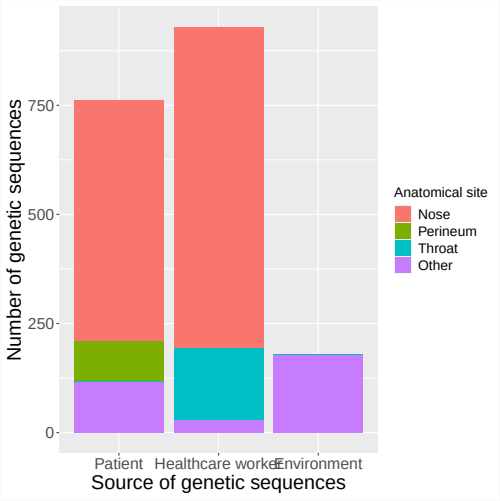
Patient data	
# Patient admitted	1919
# Unique patients	1759
# Patient episodes with ≥ 1 +ve swab	388
Total # patient +ve swabs collected	761
Total # patient +ve swabs sequenced	572
Mean stay (days)	5.02
Median stay (days)	3
Environment swabs	178
Unique # **healthcare workers** +ve	113
Total # **healthcare workers** +ve swabs sequenced	929

What is (Usually) Not Known

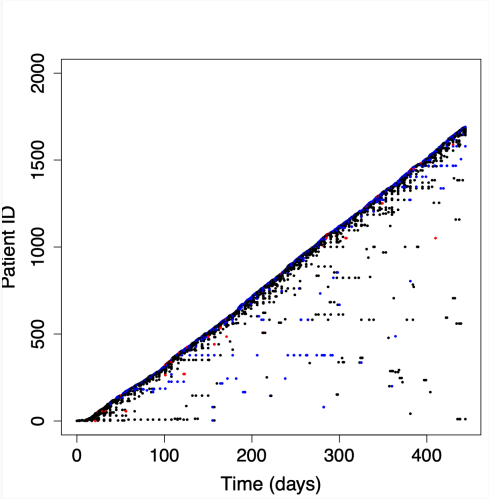
- If the patient was colonised on admission.
- When a patient became colonised (if ever)?
- How sensitive the swab test was?
- Which apparently uncolonised patients were colonised?
- Who colonised whom?
- True colonisation status of each healthcare worker.

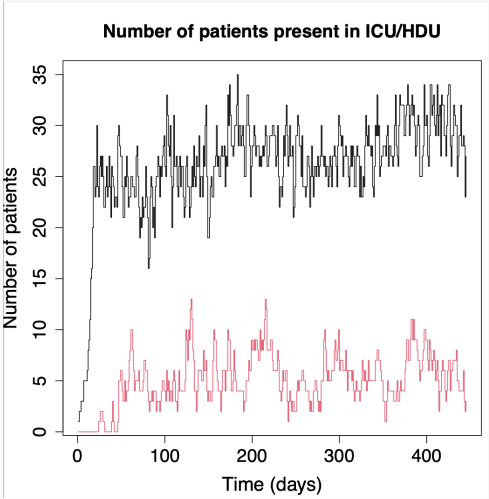
Sequencing Data Summary

- 1976 organisms were cultured and sequenced:
  - 867 from patients (43.9%),
  - 929 from healthcare workers (47.0%)
  - 180 from the environment (9.1%).



Visualising the Data



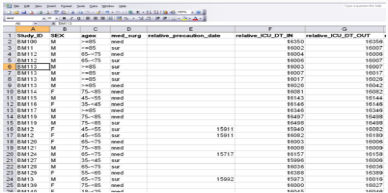


Modelling

The Basic Idea

The data have two parts:

- Epi part = patient entry, discharge, test results;



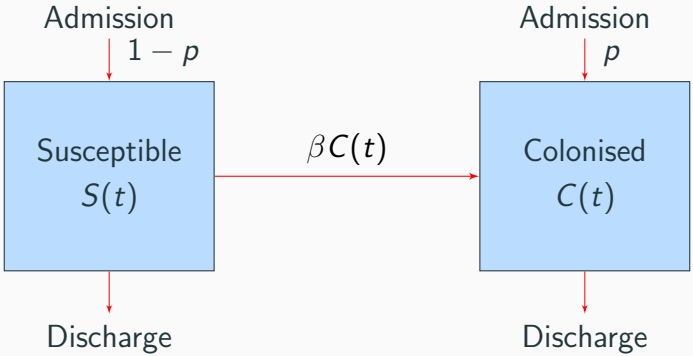
- WGS part = sequences from isolates.

The Basic Idea: Modelling the Epi Data

We construct an individual-based stochastic transmission model:

- Discrete time (days);
- Each individual is Susceptible or Colonised;
- P(patient colonised on admission) = p;
- Test sensitivity = z;
- Transmission rate (per (S, C) pair) = β;
- Once colonised a patieent stays colonised for the rest of their stay on the ward.
- Once colonised, a patient is able to colonise other patients in the following day.

The Basic Idea: Modelling the Epi Data



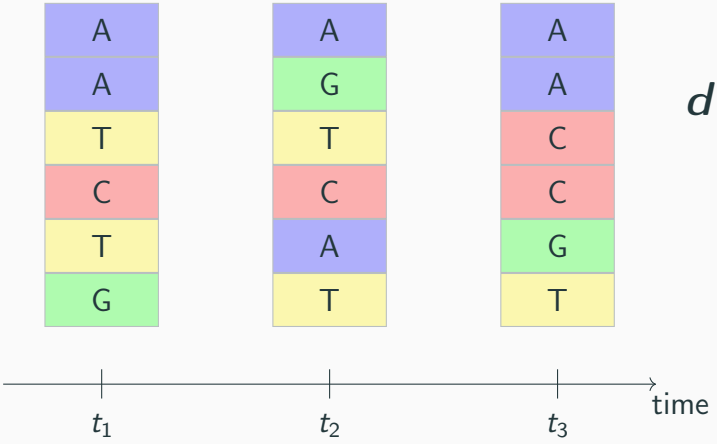
$$Pr(\text{patient } j \text{ is colonised on day } t) = 1 - \exp\left(-\beta C(t)\right))$$

The Basic Idea: Sequencing Part

- A colonised patient j who receives positive screening results **may** also **have** m_j **isolates** of the **pathogen sequenced**.
- The raw sequence data are often high dimensional.
- To reduce the dimensionality we measure the **genetic distance** between isolates by counting the number of horizontal differences of aligned sequences (i.e. SNPs).

$$\begin{bmatrix} 0 & d(1,2) & d(1,3) & d(1,4) & \dots \\ & 0 & d(2,3) & d(2,4) & \dots \\ & & 0 & d(3,4) & \dots \\ & & & 0 & \dots \end{bmatrix}$$

From Seqeuences to Distances: An Example



$$\boldsymbol{d} = \begin{bmatrix} 0 & 3 & 3 \\ & 0 & 3 \end{bmatrix}.$$

Building a Data Generating Model

We build on our earlier work [Worby et al (2016), Cassidy et al (2020)] where our individual-based stochastic transmission model also **generates such distances**.

In the setting we are considering, we need to allow for

- importations (patient arriving in the ward already colonised);
- multiple sequences for an individual.

Earlier Work [Worby et al 2016, Cassidy et al 2020]

- Such **distances** represent a **genetic difference** between the pathogen **between** two **individuals**, and its distribution depends on the relationship between the individuals in the transmission tree.
- The genetic distances between each pair of isolates in the models of Worby et al (2016) are assumed to be independent, and the distributions for these distances have been chosen somewhat arbitrarily (e.g. Geometric, Poisson) .
- Although the models of Cassidy et al (2020) address the issue of independence by considering the relatedness of host sequences, the distribution of distances are still arbitrary (e.g. Poisson).

This Work:

Assume an underlying mutation model and derive the joint distribution of the observed pairwise distances

Terminology

A *transmission network* is a directed acyclic graph describing the transmission of the pathogen through the population.

We define the term *transmission tree* to be the graph of all colonisations arising from a single imported case.

Multiple importations of the pathogen gives rise to disconnected transmission trees $\rightarrow$ transmission network is referred to as a *transmission forest.*

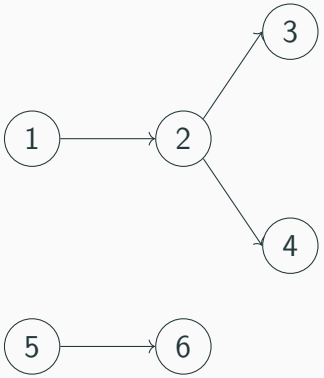


Figure 1: Tranmission Network

Genetic Network/Tree

- A *genetic network/tree* is a graph which describes the structure of genetic isolates;
- It provides a graphical representation of the genetic evolution of a pathogen through time.

- The genetic network is weighted by some edge weight function which assigns the absolute difference in time between genetic isolates.
- The *genetic network* is embedded within the *transmission network*.

Jukes and Cantor (1969) Mutation Model

Jukes and Cantor (1969) – JC69 – originally proposed the simplest model where the mutation rate for any transition is constant.

Define $\{X(t) : t \geq 0\}$ to be a homogeneous continuous-time Markov chain where $X(t)$ is a nucleotide base at time t on the state space $\mathcal{E} = \{A, G, C, T\}$ – for distinct $i, j \in \mathcal{E}$, let q_{ij} be the transition rate from state i to state j.

Let $q_{ij} = \lambda$, for $i \neq j$. The corresponding rate matrix is then given by

$$\mathbf{Q} = \begin{pmatrix} 1-3\lambda & \lambda & \lambda & \lambda \\ \lambda & 1-3\lambda & \lambda & \lambda \\ \lambda & \lambda & 1-3\lambda & \lambda \\ \lambda & \lambda & \lambda & 1-3\lambda \end{pmatrix}.$$

Jukes and Cantor (1969) Model

Let $p_{ij}(t)$ probability that a nucleotide base $i \in \mathcal{E}$ mutates to $j \in \mathcal{E}$ in t time units, i.e.

$$p_{ij}(t) = \Pr(X(t) = j \mid X(0) = i)$$

One can show that $p_{ij}(t)$ is given by

$$p_{ij}(t) = \begin{cases} \frac{1}{4}(1 + 3e^{-4\lambda t}) & i = j \\ \frac{1}{4}(1 - e^{-4\lambda t}) & i \neq j. \end{cases}$$

We now have a **probabilistic framework** for a nucleotide **observed** at **two distinct time points**,

JC69: Genetic Sequences of Length 1

- We first consider the pairwise genetic distance between genetic sequences of length $N = 1$.
- Consider a single nucleotide base evolving through time observed at times $t = t_1, ..., t_k$ denoted by $G = (B_1, ..., B_k)$, where $B_i \in \{A, G, C, T\}$ for $i = 1, ..., k$.

- Define $d_{ij} = 0$ if $B_i = B_j$ and $d_{ij} = 1$ otherwise, and $d_{ii} = 0$.

Aim

Derive the joint pmf of these distances d implied by the JC69 model

JC69: Genetic Sequences of Length 1

Aim is to **reconstruct the sequences from** the **distance matrix**; begin by labelling the set of nodes $S = \{1, ..., k\}$.

Lemma

Let $\mathbf{d} \in \mathcal{D}$ be a distance matrix for a sequence of length one for a known genetic network $\mathcal{G}$ with $k \geq 1$ connected nodes. Then there exists a unique corresponding sequence of nucleotides, up to a permutation of the bases $\mathcal{E} = \{A, G, C, T\}$.

Hence we **have a mapping from a distance matrix to nucleotides** up to a permutation of bases.

JC69: Genetic Sequences of Length 1

Let $\mathbf{d}$ be the pairwise distance matrix for a sequence of length one **under a genetic tree** $\mathcal{G} = (V, E)$ with $k \geq 1$ connected nodes;

Let $q(t)$ be the probability of observing a mutation in t time units under the JC69 model and $t_{ij} = |t_j - t_i|$ be the absolute difference in time between nodes i and j.

Theorem

The joint probability mass function is then given by

$$f(\mathbf{d}|\lambda, \mathcal{G}) = \begin{cases} 3^{\mathbb{1}_{\{h(\mathbf{d})>1\}}} 2^{\mathbb{1}_{\{h(\mathbf{d})>2\}}} \prod_{(i,j)\in E} (1 - q(t_{ij}))^{1-d_{ij}} \left(\frac{1}{3} q(t_{ij})\right)^{d_{ij}} & \mathbf{d} \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

where $h(\mathbf{d})$ is the of number of unique bases $\mathcal{G}$ given by $1 + \sum_{j=2}^{k} \left[\prod_{i=1}^{j-1} d_{i,j}\right]$.

JC69: Genetic Sequences of Length > 1

Our primary interest lies with whole genome sequences which vary in length depending on the organism.

- We **assume that each nucleotide site evolves independently to all other sites**, as such we wish to consider the joint distribution of genetic distances for each nucleotide.
- Suppose each of the N nucleotides evolve through time under the genetic tree $\mathcal{G}$, then let $\mathbf{d}^{[i]} \in \mathcal{D}$ denote the **observed pairwise distance matrix for the ith nucleotide** for $i = 1, ..., N$.

JC69: Genetic Sequences of Length > 1

Then the joint probability mass function is given by

$$f_{\mathbf{D}^{[1]},\ldots,\mathbf{D}^{[N]}}(\mathbf{d}^{[1]},\ldots,\mathbf{d}^{[N]}|\lambda,\mathcal{G}) = \Pr(\mathbf{D}^{[1]}=\mathbf{d}^{[1]},\ldots,\mathbf{D}^{[N]}=\mathbf{d}^{[N]}) = \prod_{i=1}^{N} f(\mathbf{d}^{[i]}|\lambda,\mathcal{G})$$

$$\text{[see Theorem 1]} = \prod_{i=1}^{N}\left[3^{\mathbb{1}_{\{h(\mathbf{d}^{[i]})>1\}}}2^{\mathbb{1}_{\{h(\mathbf{d}^{[i]})>2\}}}\prod_{(j,k)\in E}(1-q(t_{jk}))^{1-d_{jk}^{[i]}}\left(\frac{1}{3}q(t_{jk})\right)^{d_{jk}^{[i]}}\right]$$

$$= 3^a \times 2^b \times \prod_{i=1}^{N}\prod_{(j,k)\in E}(1-q(t_{jk}))^{1-d_{jk}^{[i]}}\left(\frac{1}{3}q(t_{jk})\right)^{d_{jk}^{[i]}}$$

$$= 3^a \times 2^b \times \prod_{(j,k)\in E}(1-q(t_{jk}))^{N-D_{jk}}\left(\frac{1}{3}q(t_{jk})\right)^{D_{jk}}$$

$$= H_1(\mathbf{d}^{[1]},\ldots,\mathbf{d}^{[N]})g_1(\lambda,\mathcal{G},T_1(\mathbf{d}^{[1]},\ldots,\mathbf{d}^{[N]})), \qquad (1)$$

JC69: Genetic Sequences of Length > 1

$$f_{\mathbf{D}^{[1]},\ldots,\mathbf{D}^{[N]}}(\mathbf{d}^{[1]},\ldots,\mathbf{d}^{[N]}|\lambda,\mathcal{G}) = 3^a \times 2^b \times \prod_{(j,k)\in E} (1-q(t_{jk}))^{N-D_{jk}} \left(\frac{1}{3}q(t_{jk})\right)^{D_{jk}}$$

$$= H_1(\mathbf{d}^{[1]},...,\mathbf{d}^{[N]}) \cdot g_1(\lambda, \mathcal{G}, T_1(\mathbf{d}^{[1]},...,\mathbf{d}^{[N]})),$$

- $a = \sum_{i=1}^{N} \mathbb{1}_{\{h(\mathbf{d}^{[i]})>1\}}$ and $b = \sum_{i=1}^{N} \mathbb{1}_{\{h(\mathbf{d}^{[i]})>2\}}$ are the number of distance matrices that contain more than one and two unique nodes respectively,
- $D_{jk} = \sum_{i=1}^{N} d_{jk}^{[i]}$ is the total number of differences between sequence j and th sequence k, where $(j,k) \in E$,
- T_1 is (matrix) statistic defined as $T_1(\mathbf{d}^{[1]},...,\mathbf{d}^{[N]}) = \sum_{i=1}^{N} \mathbf{d}^{[i]}$.

Some Remarks

- The quantities D_{ij} for $(i, j) \in E$ are straightforward to obtain by calculating a **matrix of pairwise distances for each observed sequence**.
- It follows from the Fisher-Neyman factorisation theorem that the (matrix) statistic T_1, which is the defined as the sum of distance matrices $\mathbf{d}^{[i]}$ for $i = 1, ..., N$, is **sufficient** for the underlying **parameter** λ.
- We are unable to calculate H_1 explicitly as we require knowledge of the actual sequences rather than the matrix of pairwise distances ...
- ... however since these are functions of the data alone these terms vanish in the posterior distribution when performing Bayesian inference.

Assumptions

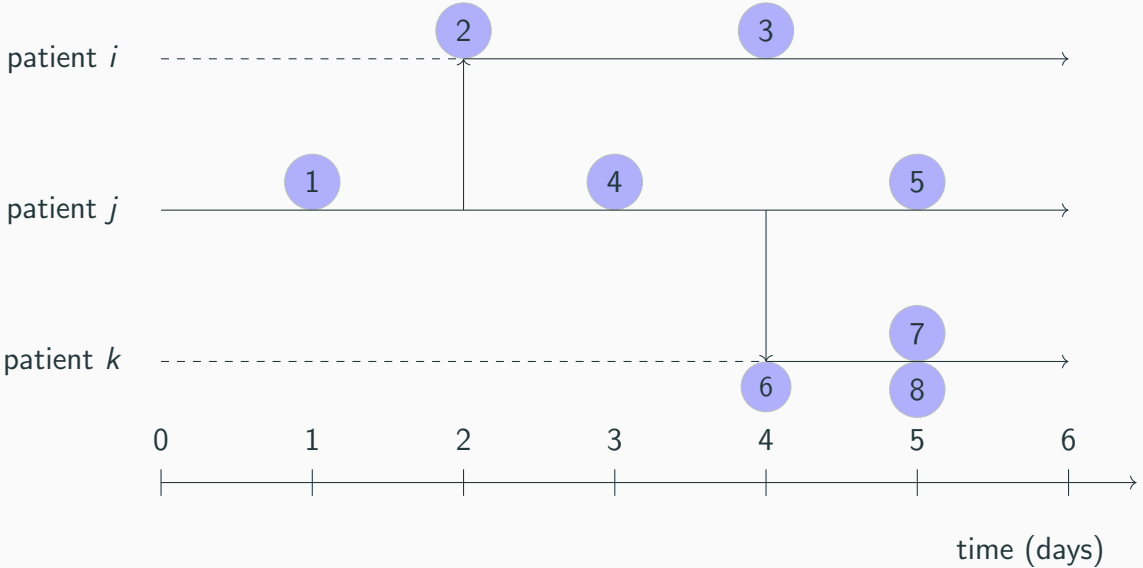
The *true* genetic network is unobserved and therefore must be constructed such that it is consistent with epidemiological data $\rightarrow$ following assumptions:

1. There is a **single dominant lineage** of the pathogen in the population at any point in time.
2. **Upon colonisation**, the genetic information of the pathogen is **transmitted** from the **source** of colonisation to the **recipient**. The transmitted pathogen then evolves independently from the pathogen in the source of colonisation.
3. Each individual has a genetic **sequence** at the **time** of **colonisation**, which may be either observed or unobserved.

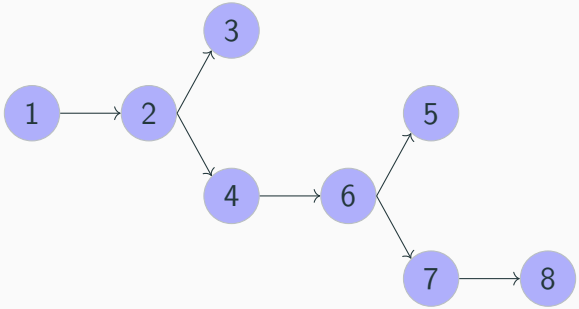
Constructing the Genetic Network

- The main idea behind constructing genetic networks is to determine the **genetic source** for each genetic sequence, i.e. the sequence that a sequence i is assumed to have evolved from.
- To determine the genetic source for a given sequence, at an individual level we **look backwards** in time to determine if there is a **previously sampled** genetic sequence.
- A genetic source sequence will either be a result of a swab from the same individual, or a swab at the time of colonisation from an infectee.

Constructing the Genetic Network (An Example)



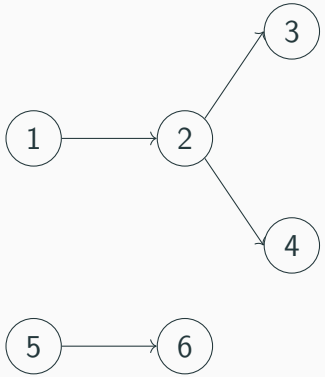
Constructing the Genetic Network (An Example)



Modelling Distance Between Imported Sequences

We are then able to construct the genetic network that is consistent with the epidemiological data, i.e. the current state of the transmission tree.

Need a model to explain the pairwise distances between imported sequences.



We assume that **imported** sequences in distinct transmission chains are **unrelated** and their distance $\sim$ Poisson(μ) s in Cassidy et al (2020).

Bayesian Inference

Bayesian Inference

Given the observed data (epi + genetic distances) we wish to make inference for the parameters of interest:

- transmission rate β;
- test sensitivity z;
- importation probability p;
- mutation rate λ;
- average imported sequence distance μ;
- transmission network.

Bayesian Inference via Data Augmentation

The joint likelihood of the transmission dynamics, screening results and genetic data given the model parameters is:

$$\pi(z^{obs}, \psi, F, x^s|\rho) = \pi(\psi, \tilde{\psi}|z^{obs}, T, \mathcal{G}, \rho)\pi(x^s|z^{obs}, T, \rho)\pi(T|z^{obs}, \rho), \qquad (2)$$

where

- the first term is the likelihood of the genetic data;
- the second term is the likelihood of observational data arising from the pathogen screening tests and
- the third term in is the likelihood of the transmission dynamics.

Bayesian Inference via Data Augmentation

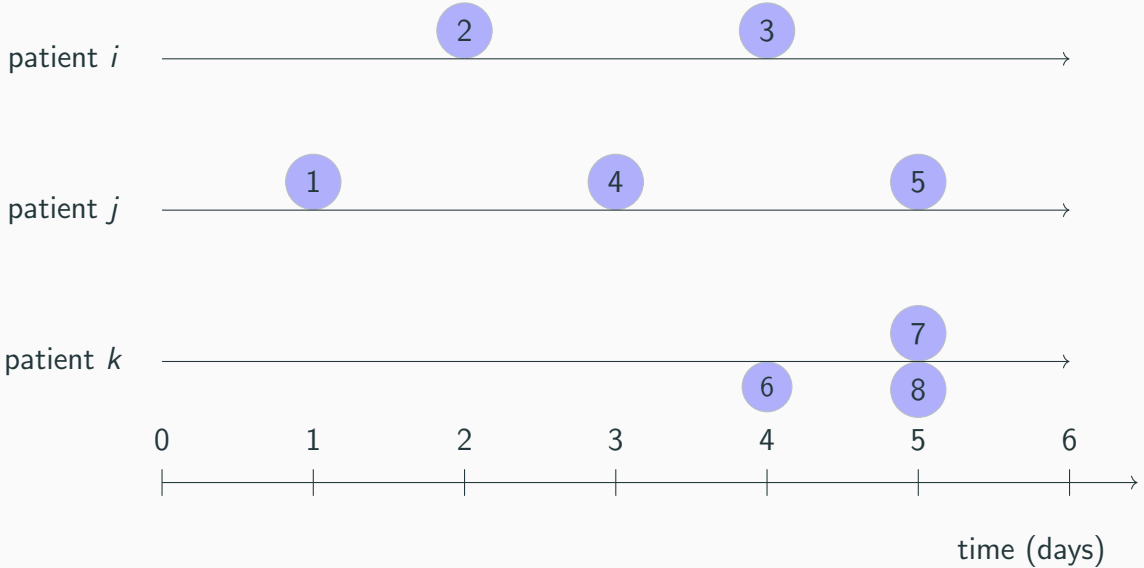
- Data Augmentation (colonisation times, admission status, genetic network, transmission network);
- Bayesian Inference;
- (Bespoke) Markov Chain Monte Carlo (MCMC).

MCMC

Constructing an efficient MCMC algorithm is not a straightforward task!

Involves sampling sources, adding/deleting/moving colonisation times, sampling unobserved distances ...

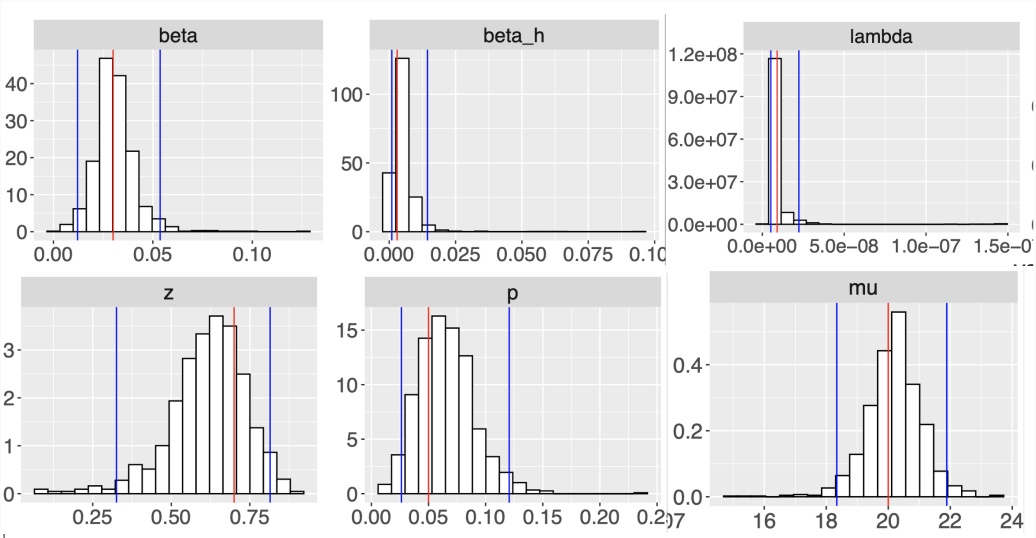
Bayesian Inference via Data Augmentation



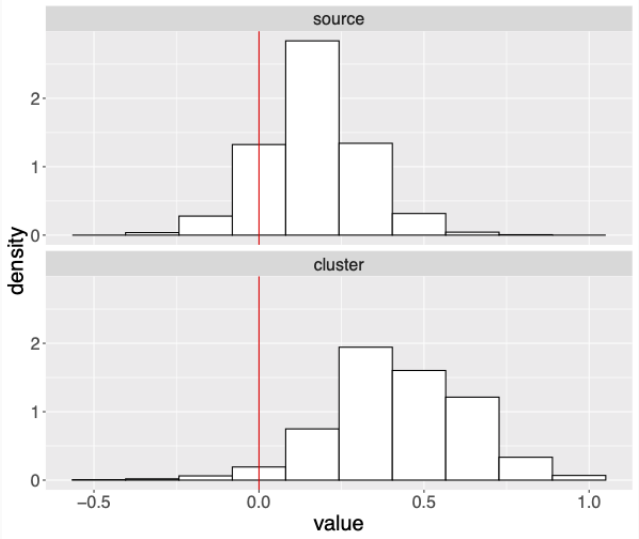
Results

Results on Simulated Data

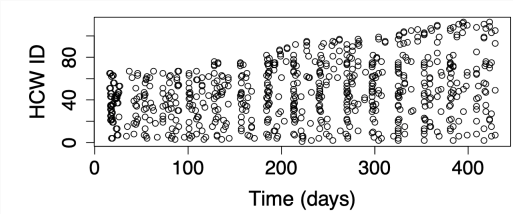
100 people; 5 healthcare workers



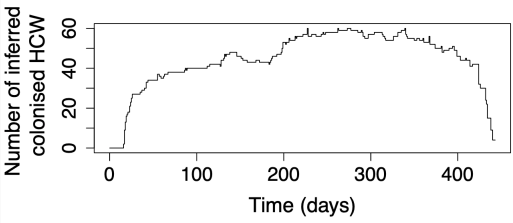
Results on Simulated Data



Data on Healthcare Workers (HCW)

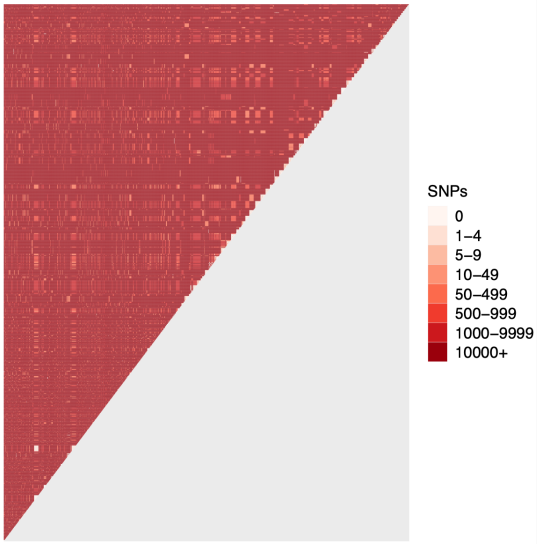


Positive test results (black circle) for each healthcare worker (top).



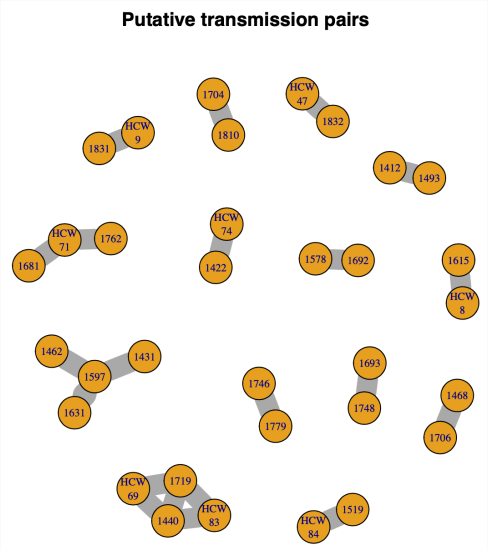
Inferred number of colonised healthcare workers on the ward by assuming constant carriage in-between test results and for a fixed period of 14 days thereafter.

Heatmap of Distances



- There are extremely diverse genetic sequences with many distances $> 10,000$ SNPs (1800 years of evolution).
- Albeit extremely diverse, are still variations of the pathogen *S. aureus.*
- There area small areas of the heat map with colours indicating < 5 SNPs.

Putative Transmission Pairs



- We define the term *putative transmission pair* to be any pair of individuals that have observed distances with < 5 SNPs and have spent time on the ward together at the same time.
- Focus on patient-patient and healthcare worker-patient pairs.
- 14 potential clusters containing a total of 22 individuals that are assumed to be epidemiologically related.

Transmission Model Incorporating HCW

- Let $C(t)$ and $H(t)$ denote the number of colonised patients and healthcare workers at time t respectively.
- The total colonisation pressure A exerted on susceptible patients is given by

$$A = \beta C(t) + \beta_H H(t)$$

and

$$\Pr(\text{patient j is colonised on day t}) = 1 - \exp(-A).$$

where β is the rate of contact between colonised and susceptible patient pairs and β_H describes the rate of contact between susceptible patient and colonised healthcare worker pairs.

Introducing Genetic Subtypes

- We are not able to perform any meaningful analysis with the extremely diverse genetic distances.
- Instead we focus on the **similar genetic isolates**.
- In a similar fashion to Price et al. (2017) we define the term *genetic subtype* to refer to "similar" sequences.
- Sequences within the same genetic subtype can be thought of as **snapshots of the same organism**, therefore the distances may be described by the mutation model.

Modelling Multiple Genetic Sequences

The (previous) assumption that at any point in time there exists a single dominant strain or subtype in the population appears unreasonable in this setting.

Hence we build a model that readily incorporates genetic diversity in a flexible and efficient manner, which builds upon the model for a single dominant strain:

- Label sequences pre-analysis with *genetic groups* where each genetic group contains sequences that are assumed to have evolved from separate organisms.
- Each group is modelled by a separate stochastic process where each of these stochastic processes are assumed to evolve independently of one another.

Determing the Genetic Groups

The process to classify and cluster genetic sequences to groups is typically a non trivial task.

We have considered two ways to do so:

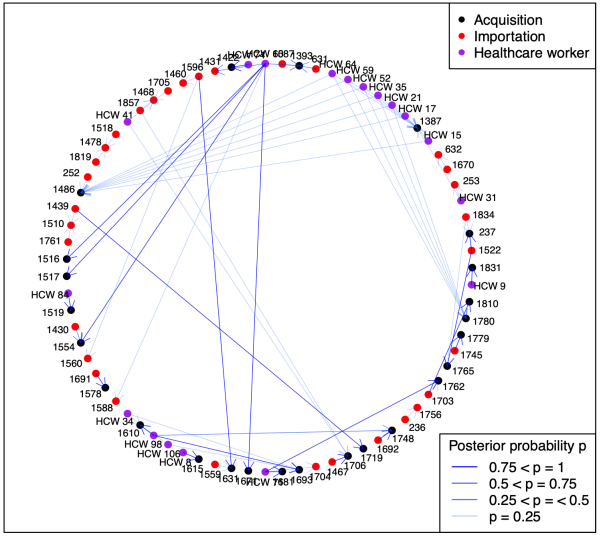
- **Group by threshold**: choose an arbitrary threshold, α say, and group similar genetic sequences that have pairwise distances $< \alpha$.
- **k-means clustering**: seeks to partition objects into groups such that the objects within the group are sufficiently close.

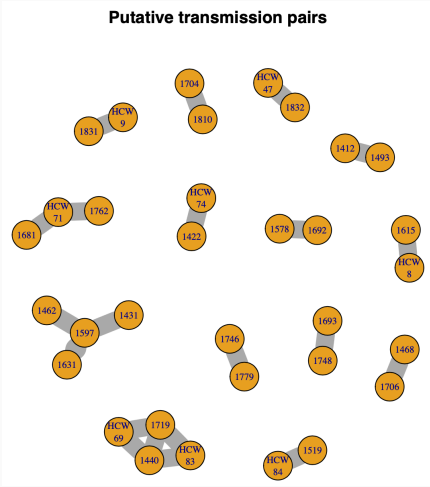
Results for MRSA + MSSA Combined

Parameter	Estimate	95% Credible Interval
z	0.61	(0.56, 0.66)
p	0.23	(0.20, 0.26)
β	0.00077	(0.00037, 0.00134)
β_H	0.00012	(0.000068, 0.00019)
$\lambda \times 10^{-9}$	5.67	(5.28, 6.04)
μ	12.83	(12,13.66)

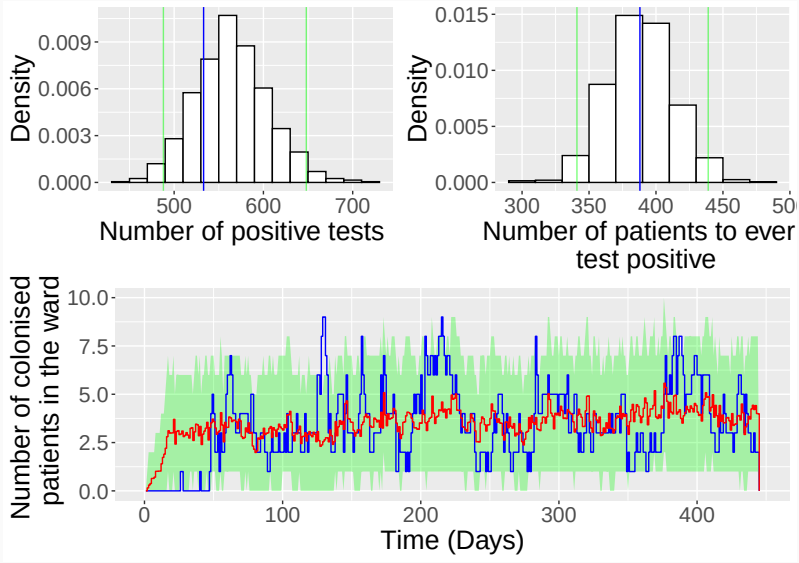
Table 1: Posterior mean estimates for the model parameters along with 95% (equal-tailed) credible intervals.

Inferred Transmission Network

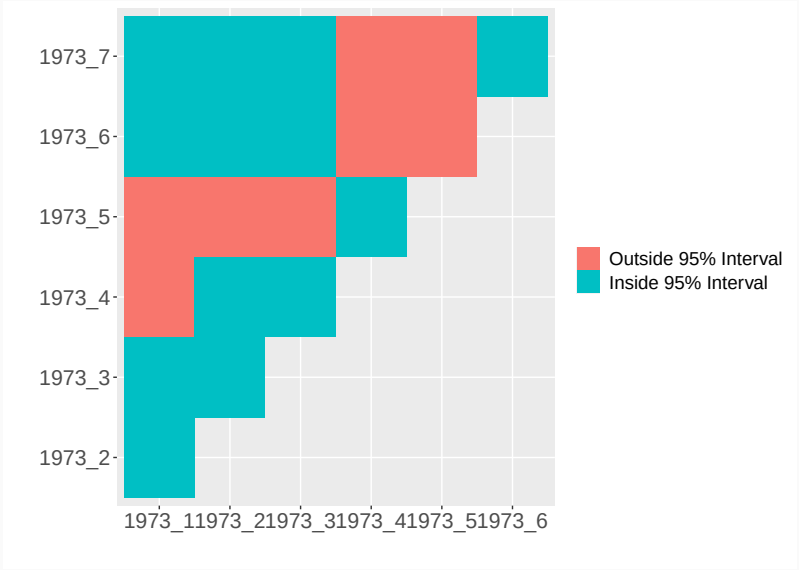




Model Assessment (Epi Data)



Model Assessment (Genetic Data)



Conclusions

Conclusions

- Principled generative modelling approach, i.e. we have a model that can generate the data $\rightarrow$ enabling model assessment.
- Bayesian inference via bespoke MCMC algorithms;
- Key idea is to model differences between sequences rather than sequences themselves $\rightarrow$ dimension reduction.
- Allows for importations, multiple and/or diverse genetic sequences.
- Can be **adapted** to other **mutation models** (e.g. Kimura model).
- **Address** the issues of **arbitrariness** of **earlier work** by deriving the distribution of the pairwise distances under the assumption of the a mutation model.

Future Work / Work in Progress

- Settings outside hospital infections.
- Systematic comparison against the work of Lau et al (2015).
- Alternative model for imported sequences.
- Model assessment methods.
- Improve the efficiency of the MCMC sampler for large scale inference.