# Identifying Transmission Routes using High Resolution Genetic Data: An Application to Healthcare Associated Infections

Theo Kypraios
http://www.maths.nott.ac.uk/∼tk

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA
Mathematical Sciences

**SMIDDY, Zurich, September, 2013**

# Joint work with:

- **Dr Colin Worby** @ Harvard School of Public Health (previously at UoN)
- **Prof Phil O'Neill** @ University of Nottingham
- **Dr Ben Cooper** @ Mahidol University, Bangkok, Thailand (previously at the Public Health England, London)

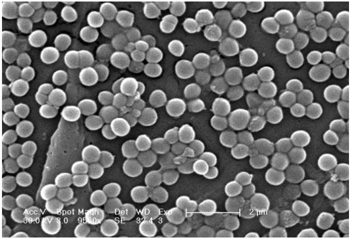
# Healthcare-associated infections

- Healthcare-associated infections (eg. MRSA, *C. difficile*, *E. coli*) are a major cause of illness and death in hospitals worldwide.

- It is of great interest to investigate transmission dynamics, in order to improve infection control strategies.

- The collection of high-resolution genetic data is becoming easier and cheaper.

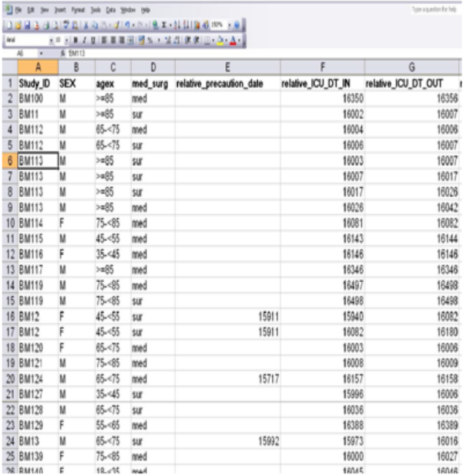
- High–resolution genetic data potentially offers new insights into the dynamics of a hospital disease outbreak.



# Typical Data sets

Typical data sets contain anonymised ward - level information on:

- Dates of patient admission and discharge.
- Dates/Outcomes of swab tests.
- Patient location (e.g. in isolation).
- Details of antibiotics administered.
- Typing data.

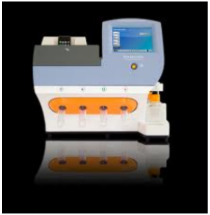


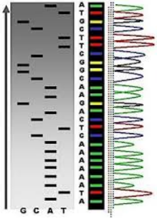
|  | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Study_ID | SEX | agex | med_surg | relative_precaution_date | relative_ICU_DT_IN | relative_ICU_DT_OUT |
| 2 | BM100 | M | >=85 | med |  | 16350 | 16356 |
| 3 | BM11 | M | >=85 | sur |  | 16002 | 16007 |
| 4 | BM112 | M | 65-<75 | med |  | 16004 | 16006 |
| 5 | BM112 | M | 65-<75 | sur |  | 16006 | 16007 |
| 6 | BM113 | M | >=85 | sur |  | 16003 | 16007 |
| 7 | BM113 | M | >=85 | sur |  | 16007 | 16017 |
| 8 | BM113 | M | >=85 | sur |  | 16017 | 16026 |
| 9 | BM113 | M | >=85 | med |  | 16026 | 16042 |
| 10 | BM114 | F | 75-<85 | med |  | 16081 | 16082 |
| 11 | BM115 | M | 45-<55 | med |  | 16143 | 16144 |
| 12 | BM116 | F | 35-<45 | med |  | 16146 | 16148 |
| 13 | BM117 | M | >=85 | med |  | 16346 | 16346 |
| 14 | BM119 | M | 75-<85 | med |  | 16497 | 16498 |
| 15 | BM119 | M | 75-<85 | sur |  | 16498 | 16498 |
| 16 | BM12 | F | 45-<55 | sur | 15911 | 15940 | 16082 |
| 17 | BM12 | F | 45-<55 | sur | 15911 | 16082 | 16180 |
| 18 | BM120 | F | 65-<75 | med |  | 16003 | 16006 |
| 19 | BM121 | M | 75-<85 | med |  | 16008 | 16009 |
| 20 | BM124 | M | 65-<75 | med | 15717 | 16157 | 16158 |
| 21 | BM127 | M | 35-<45 | sur |  | 15996 | 16006 |
| 22 | BM128 | M | 65-<75 | sur |  | 16036 | 16036 |
| 23 | BM129 | F | 55-<65 | med |  | 16388 | 16389 |
| 24 | BM13 | M | 65-<75 | sur | 15992 | 15973 | 16016 |
| 25 | BM139 | F | 75-<85 | med |  | 16000 | 16027 |
| 26 | BM140 | F | 18-<35 | med |  | 16045 | 16048 |

# High–Resolution Genetic Data (1)

"High–resolution genetic data": what are they?

- individual–level data on the pathogen;
- can be taken at single or multiple time points;
- high-dimensional e.g. whole genome sequences;
- proportion of individuals sampled could be high/low;
- becoming far more common due to cost reduction;





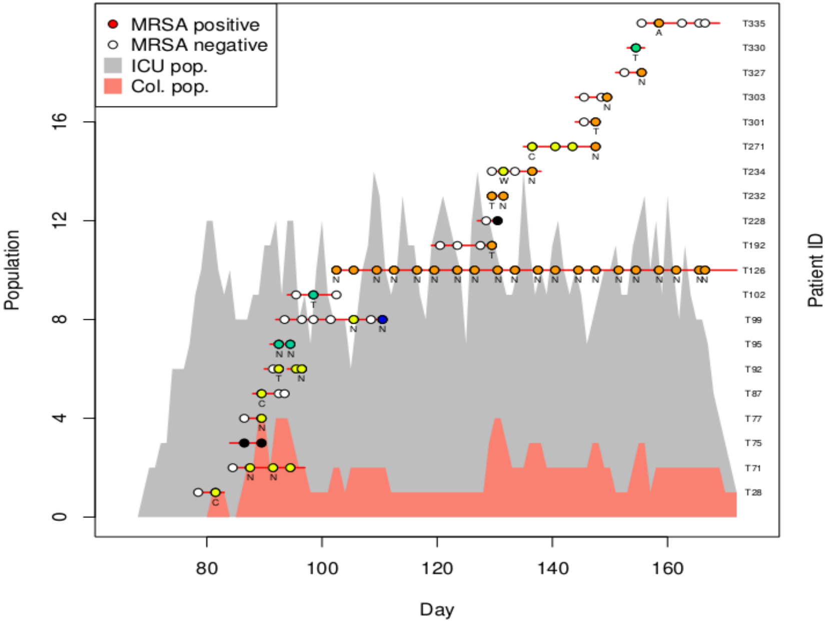
# High–Resolution Genetic Data (2)

"High–resolution genetic data": what use are they?

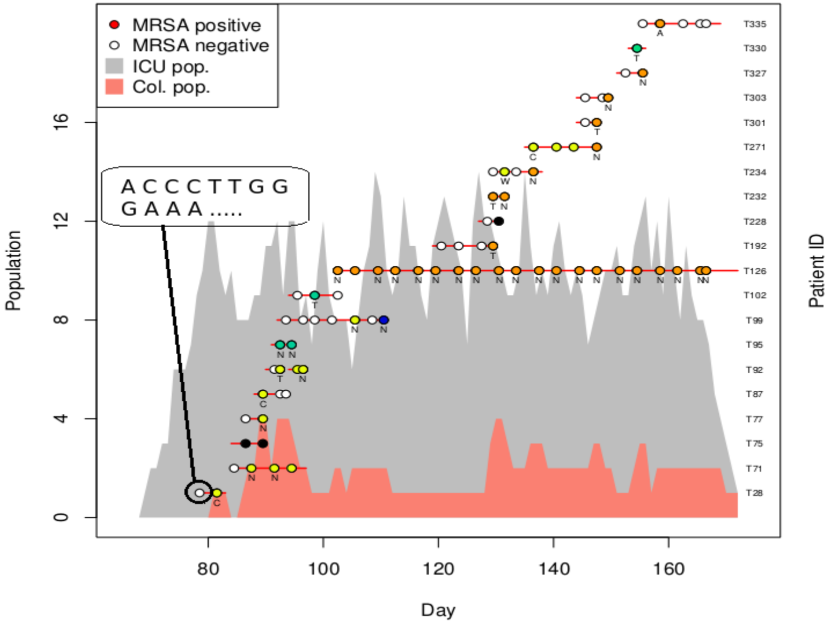
Can provide much insight into the dynamics of transmission:

- better inference about transmission paths
- more reliable estimates of epidemiological quantities (e.g. the effectiveness of infection control precautions)?;
- understand evolution of the pathogen.

# MRSA positive patients, ICU 1



# MRSA positive patients, ICU 1



# Existing Work/Studies

At least two kinds of approaches exist:

1. Separate genetic and epidemic components:For example,
   - estimate phylogenetic tree;
   - given the tree, fit epidemic model.

   or

   - cluster individuals into genetically similar groups;
   - given the groups, fit multi-type epidemic model.

   [See, for example, Volz *et al.* (2009), Rasmussen *et al.* (2011), . . .]

2. Combine genetic and epidemic components: For example,
   - model genetic evolution explicitly;
   - define model featuring both genetic and epidemic parts.

   [See, for example, Ypma *et al* (2012), Worby (2013), . . .]

# Existing Work/Studies (Pros and Cons)

1. Separate genetic and epidemic components:
   - + “Simple” approach;
   - + Avoids complex modelling;
   - − Ignores any relationship between transmission and genetic information.
2. Combine genetic and epidemic components:
   - + “Integrated” approach.
   - − Is modelling too detailed? [mutation, recombination etc]
   - − Initial conditions: typical sequence?
   - +/− Model differences between individuals instead?

# Existing Work/Studies (Pros and Cons)

1. Separate genetic and epidemic components:

   + “Simple” approach;
   + Avoids complex modelling;
   − Ignores any relationship between transmission and genetic information.

2. Combine genetic and epidemic components:

   + “Integrated” approach.
   − Is modelling too detailed? [mutation, recombination etc]
   − Initial conditions: typical sequence?

   +/− **Model differences between individuals instead?**

# Our Proposed Framework

- Develop a more generalized approach to transmission network reconstruction;
- model the distribution of genetic distances observed between each pair of sampled isolates.
- allow multiple independent introductions of the pathogen;
- account for within-host diversity;
- make no assumptions about the evolutionary dynamics of the pathogen;
- do not consider the phylogenetic relationship between isolates.

# Genetic distance matrix

We define the genetic distance between isolates $X_1$ and $X_2$ to be the number of SNPs between the isolates, $\psi(X_1, X_2)$.

Since we are interested in the genetic distance between isolates, rather than the composition of the genome itself, we define $\Psi$ to be the matrix of pairwise genetic distances between all isolates.

In other words, that means that each new colonised patient ($i$, say) needs to have distance $\psi((i, k)$ to all existing colonised patients $k$.

We draw $\psi(i, k)$ from a probability distribution according to "type": Each new colonised patient is either:

1. An importation (i.e enter ICU already colonised)
2. An acquisition (i.e colonised by another patient)

# Genetic distance matrix (Cont)

1. **Importation structure model**: assigns each colonized patient a group where groups contain genetically similar sequences [groups are not pre-defined].

   - It is assumed that a patient acquires the same MRSA type as their source.
   - Importations may belong to the same group, which is realistic when there are common strain types circulating in the community, or a shared external source elsewhere in the hospital.
   - Under this model, any pair of isolates taken from patients within the same transmission chain have the same expected genetic distance (i.e. follow the same distribution) regardless of the network distance between the nodes.

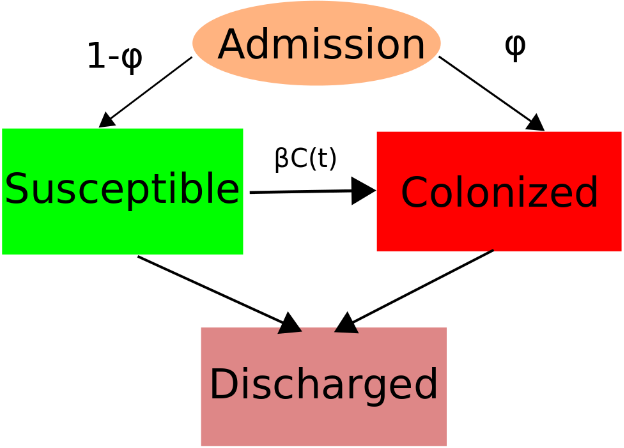
# Genetic distance matrix (Cont)

2. **Transmission diversity model**: assumes that the expected genetic diversity increases monotonically as sampled individuals are further apart in the network.

   - We assume that distances between isolates taken from individuals in unrelated transmission chains are drawn from a specified distribution, with an expected distance larger than within-chain distances.

   - Based on the idea that closely related individuals are likely to host genetically similar bacteria, while those who are part of independent outbreaks are likely to carry genetically diverse strains.

# Transition Model Dynamics



*$P$(susceptible patient avoids colonization on day $t$) $= \exp\{-\beta C_t\}$

*Screening tests (sensitivity $z\%$, specificity $100\%$)

# Data augmentation and MCMC

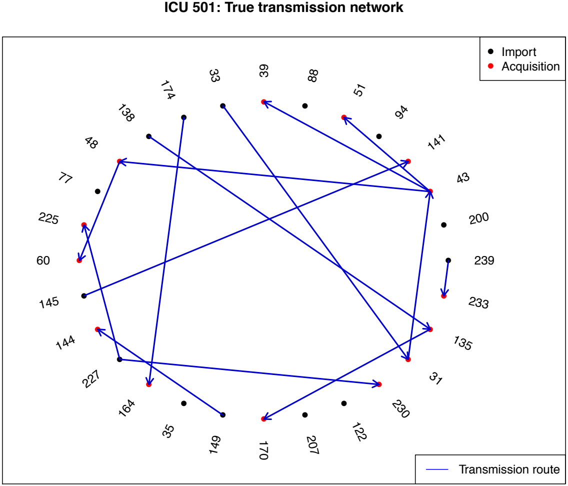
As such, we can write a likelihood function for the swab and sequencing data $X$, given the model, which is tractable provided the time and source of each positive individual is known.

As this information is typically unobserved, we proposed to augment the parameter space $\theta$ with latent data $T$.

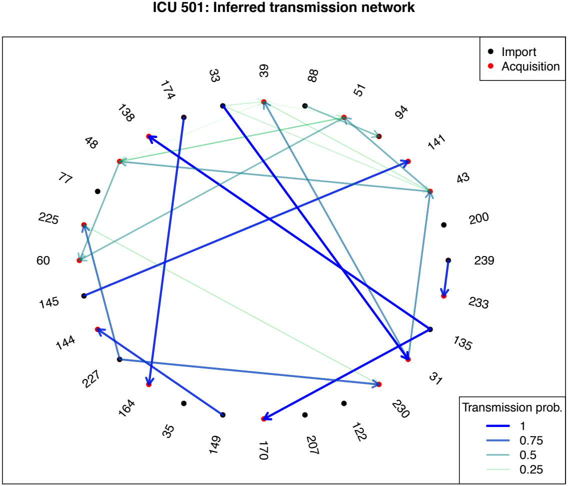
This results in a tractable likelihood, and we may explore the posterior density using a **Markov chain Monte Carlo** (MCMC) algorithm to sample unseen transmission dynamics $T$ and model parameters $\theta$.

$$\pi(\theta, T|X) \propto \pi(X|T,\theta)\pi(T|\theta)\pi(\theta)$$

# Simulated patient network

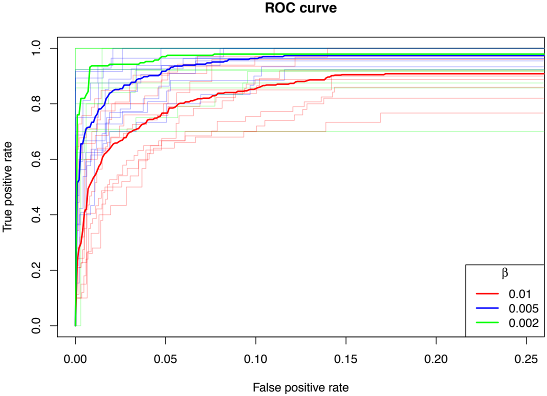


# Estimated patient network

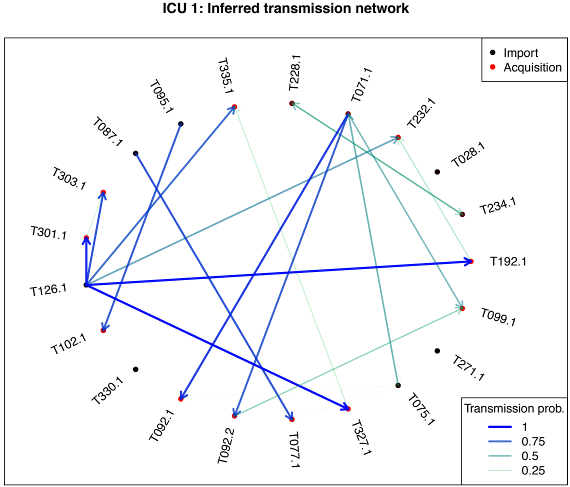


# Assessing network accuracy

We determined the accuracy of estimated networks using the ROC curve. Increased transmission, higher genetic diversity and lower sensitivity all resulted in reduced network accuracy.



# WGS samples from Thai ICU 1



# Limitations

- Only a small sample of sequences to work with — little indication of scale of within-host diversity.
- Imported strains may be related due to some external source.
- Multiple colonisation is not taken into account — it may be possible for a patient to acquire a second, genetically distinct colonisation which either replaces, or coexists with, the initial colonisation.

# Future work

- A generalized approach to reconstructing infection transmission routes using densely sampled genomic data.
- Although the model might be quite simplistic, provides a framework to incorporate additional complexity to the dynamics of transmission or genetic diversity.
- Within-host diversity makes it harder to resolve network.
- Mechanism to incorporate reinfection would be beneficial.

# Acknowledgments

- Outbreak Data in Thailand: Courtesy of S. Peacock, M. Holden, E.Nickerson, M. Hongsuwan & J. Parkhill who collected and processed the data set.

- **Funding**

The University of Nottingham

mosar
Mastering hospital antimicrobial resistance in Europe

THE ROYAL SOCIETY

EPSRC
Engineering and Physical Sciences Research Council