Integrating genetic and epidemiological data to determine virus transmission pathways

Eleanor COTTAM 2,3, Gaël THÉBAUD 1,2, Jemma WADSWORTH 3, John GLOSTER 3, Leonard MANSLEY 4, David PATON 3, Donald KING 3, Dan HAYDON 2

1 UMR BGPI (INRA-Montpellier)

2 Division of Environmental and Evolutionary Biology (University of Glasgow)

3 Institute for Animal Health (Pirbright)

4 Animal Health Divisional Office (Perth)
Introduction

Molecular epidemiology and directionality

- Genetic sequences:
  - phylogeny
  - clades / groups / types

- Comparison between genetic similarity and
  - geographic proximity
  - ecological zone
  - host species
  - ...

- Direction of transmission:
  - reference (more or less implicit)
    to additional information
Introduction

Why is directionality interesting?

- Implications:
  - logical: source ≈ cause ≈ consequences
  - legal: responsible ≈ victim

Accessible information

- type of source and target individuals
- transmission distances
- important or missing sources
- likely transmission modes

Use of the information

- parameterise epidemiological models (e.g., network models)
- limit virus propagation
- multiscale models

evolution

- evolution during 1 transmission cycle

In theory, complete description of the epidemic

In practice, data sets concerning few individuals
Questions on FMDV

• At which scale is there some viral genetic polymorphism?
  – animal, farm, disease focus?

• Can we use the observed polymorphism to identify transmission chains? How?

• What is the reliability of veterinary contact tracing?
Biological system

- **Foot-and-mouth disease virus outbreak (2001)**
  - 20 complete genomes (~10 kb each)
    - 5 initial infections with a known history
    - 15 farms from the same focus (Durham County)

- Positive-strand RNA virus:
  - High mutation rate (~10^{-4} errors/nucleotide/replication)
  - Limited recombination
- Known root
- 2 independent introductions
- 4 groups

Genetic data

Gaël Thébaud

10 km

Day of outbreak

MIEP08 • Montpellier, France • 10-12/06/2008
Genetic data

- Known root
- 2 independent introductions
- 4 groups

How to identify transmission history?
Genetic data

Which is the most likely transmission tree?

- Known root
- 1 known chain of transmissions
- 3 obvious transmissions
- What about the other ones??

Which is the most likely farm for each node?

Use of contact tracing data

Gaël Thébaud

MIEP08 • Montpellier, France • 10-12/06/2008
Epidemiology

- $L$: Probability density for latency ($\Gamma$)
- $I_i$: Probability density for the infection date of farm $i$
- $F_i(t)$: Probability for farm $i$ to be infectious at date $t$

\[ t \leq C_i : F_i(t) = \sum_{\tau=0}^{t} I_i(\tau) \cdot \left( \sum_{k=1}^{t-\tau} L(k) \right) \]
\[ t > C_i : F_i(t) = 0 \]
Epidemiology

Animal movement ban

- $L$: Probability density for latency ($\Gamma$)
- $I_i$: Probability density for the infection date of farm $i$
- $F_i(t)$: Probability for farm $i$ to be infectious at date $t$

\[
\lambda_{ij} = \frac{\sum_{t=0}^{\min(C_j, C_i)} I_i(t) \cdot F_j(t)}{\sum_{k=1}^{n} \sum_{k \neq i}^{\min(C_j, C_k)} \sum_{t=0}^{\min(C_j, C_k)} I_i(t) \cdot F_k(t)}
\]

- $\lambda_{ij}$: Likelihood of $i \leftarrow \{ j \}$ rather than another observed farm
• $\lambda_{ij}$: likelihood of $i \leftrightarrow \{ j \}$ rather than another observed farm
• $\lambda_{ij}$ can be computed for each transmission
• Thus, for a complete transmission tree ($k$), $\lambda_k = \prod \lambda_{ij}$
• And $\lambda_k$ can be computed for any tree
  … if all the possible trees can be enumerated
  → Algorithm defining the possible trees by recurrence from the leaves back to the root

All differing from contact tracing results
Rescaled likelihood: 
\[ \lambda'_k = \lambda_k / \Sigma \lambda_k \]

Which group of trees represent 95% of the rescaled likelihood?

Which is the most likely group of trees?
Genetics + epidemiology

Which is the most likely tree?

( # ) Number of distinct sources among the 4 most likely trees

[ # ] Likelihood of the most probable transmission

<table>
<thead>
<tr>
<th>Source farm</th>
<th>Recipient farm</th>
<th>Likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>A</td>
<td>27.3</td>
</tr>
<tr>
<td>4</td>
<td>K</td>
<td>5.1</td>
</tr>
<tr>
<td>A</td>
<td>N</td>
<td>2.1e+16</td>
</tr>
<tr>
<td>F</td>
<td>G</td>
<td>8.3e+11</td>
</tr>
<tr>
<td>K</td>
<td>B</td>
<td>168</td>
</tr>
<tr>
<td>K</td>
<td>F</td>
<td>4.5e+03</td>
</tr>
<tr>
<td>K</td>
<td>L</td>
<td>94.7</td>
</tr>
<tr>
<td>K</td>
<td>O</td>
<td>84.6</td>
</tr>
<tr>
<td>L</td>
<td>E</td>
<td>94.7</td>
</tr>
<tr>
<td>O</td>
<td>C</td>
<td>84.6</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>84.6</td>
</tr>
<tr>
<td>O</td>
<td>P</td>
<td>84.6</td>
</tr>
</tbody>
</table>

Gaël Thébaud                                MIEP08 • Montpellier, France • 10-12/06/2008
Genetics + epidemiology

Which is the most likely tree?
Short distance transmission

Mean distance

\[ P_{(1\text{-sided})} = 1.2 \times 10^{-3} \]

MeanDistSim

Frequency

\[
\begin{array}{cccc}
4000 & 6000 & 8000 & 10000 \\
0 & 5000 & 10000 & 15000 & 20000
\end{array}
\]

\(10\text{ km}\)
Conclusions

Summary

• The whole set of possible transmission trees is identified based on genetic data
• Their relative likelihood is evaluated based on epidemiological data
• Interesting method for real-time forensic applications

Difficulties

• Identifying the tree root
• Dealing with censoring / sampling issues
• Weighting different sources of information