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

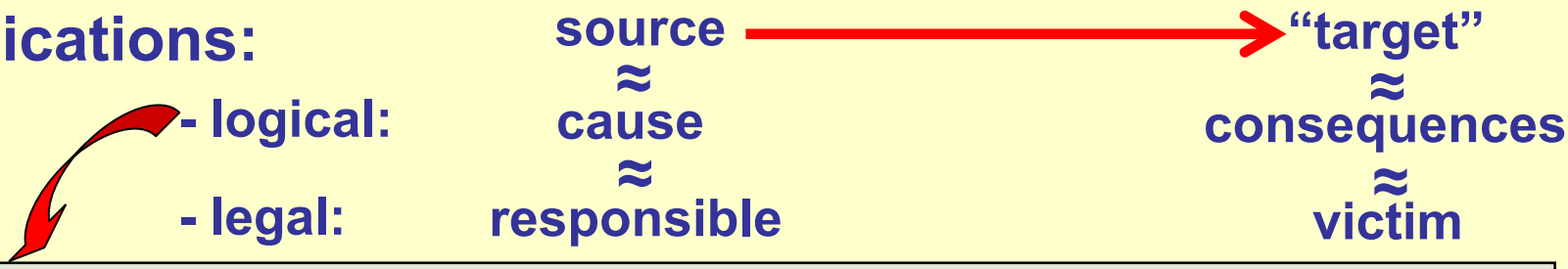
Molecular epidemiology and directionality

- **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:**



Accessible information	Use of the information
epidemiology	
• type of source and target individuals • transmission distances • important or missing sources • likely transmission modes	• parameterise epidemiological models (e.g., network models) • limit virus propagation
evolution	
• evolution during 1 transmission cycle	• multiscale models

→ In theory, complete description of the epidemic

→ In practice, data sets concerning few individuals

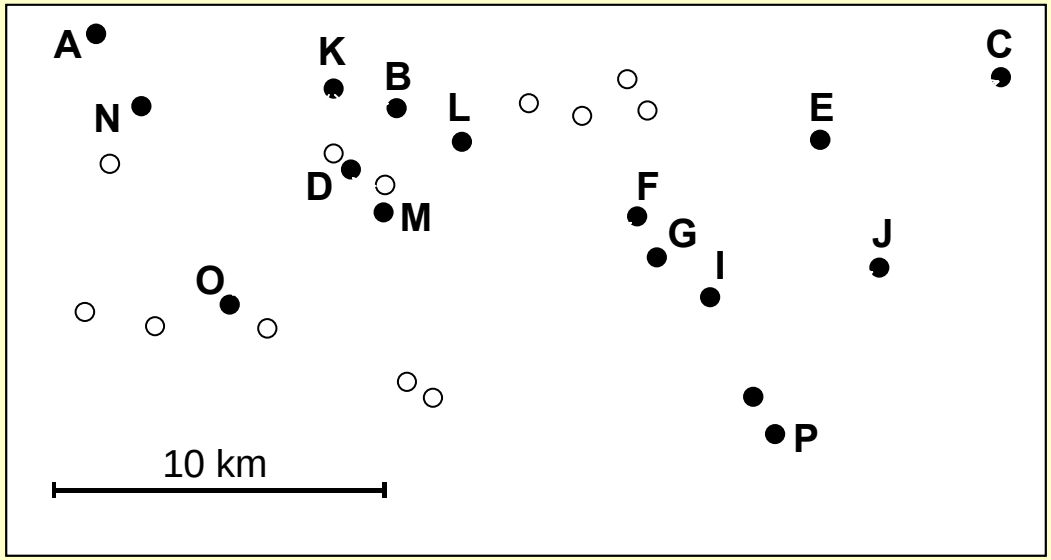
Introduction

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?

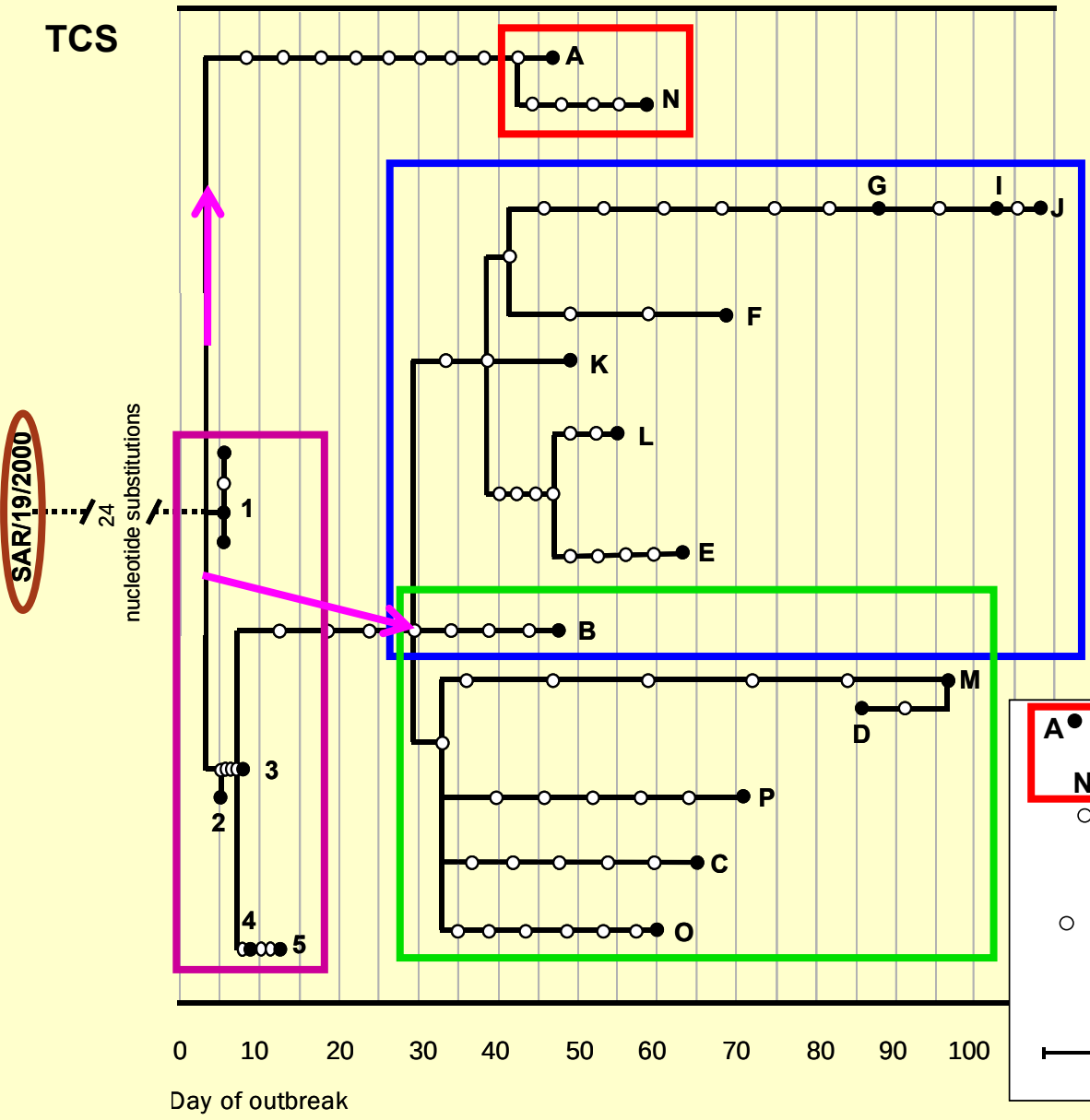
Biological system

- *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)

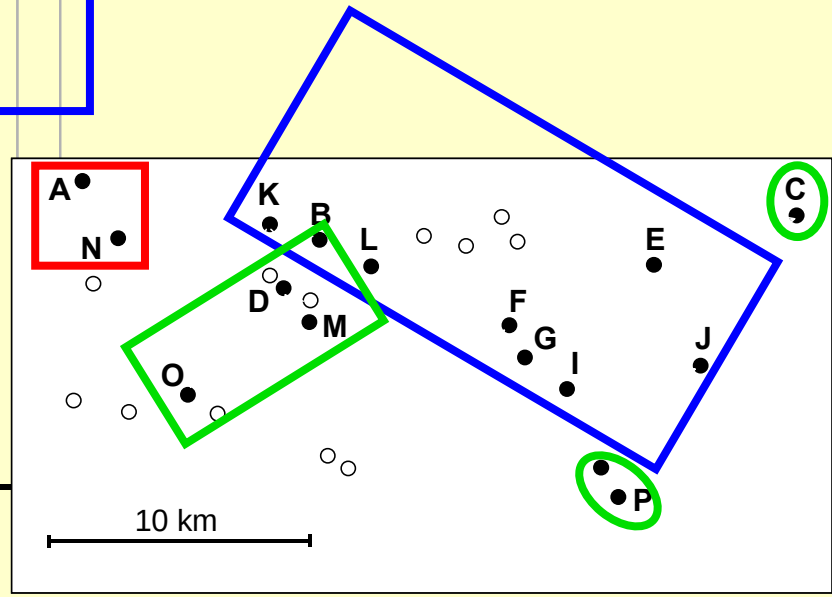


- Positive-strand RNA virus:
  - High mutation rate (~10^{-4} errors/nucleotide/replication)
  - Limited recombination

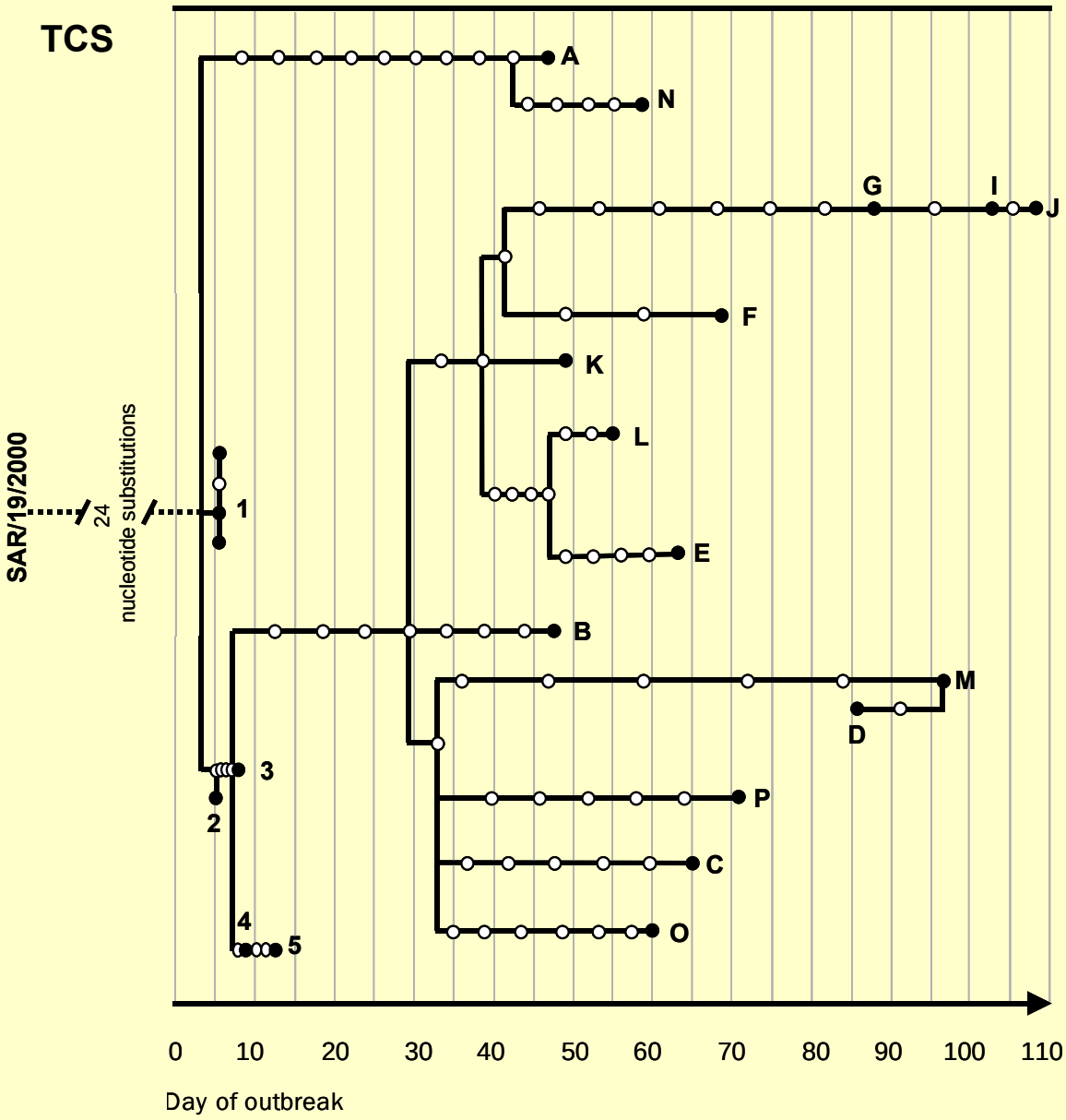
Genetic data



- Known root
- 2 independent introductions
- 4 groups



Genetic data

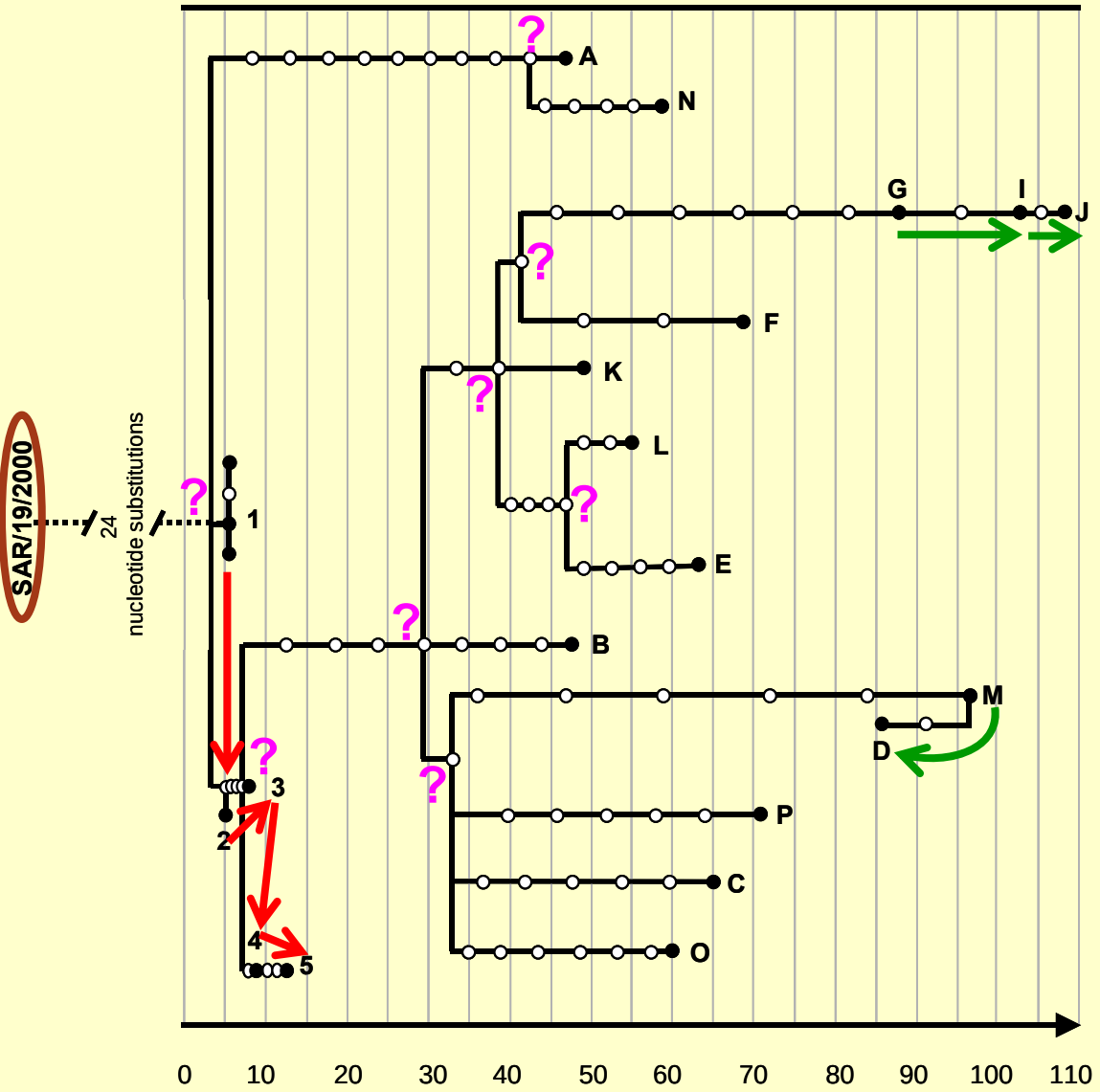


- Known root
- 2 independent introductions
- 4 groups

How to identify transmission history?

Genetic data

Which is the most likely transmission tree?

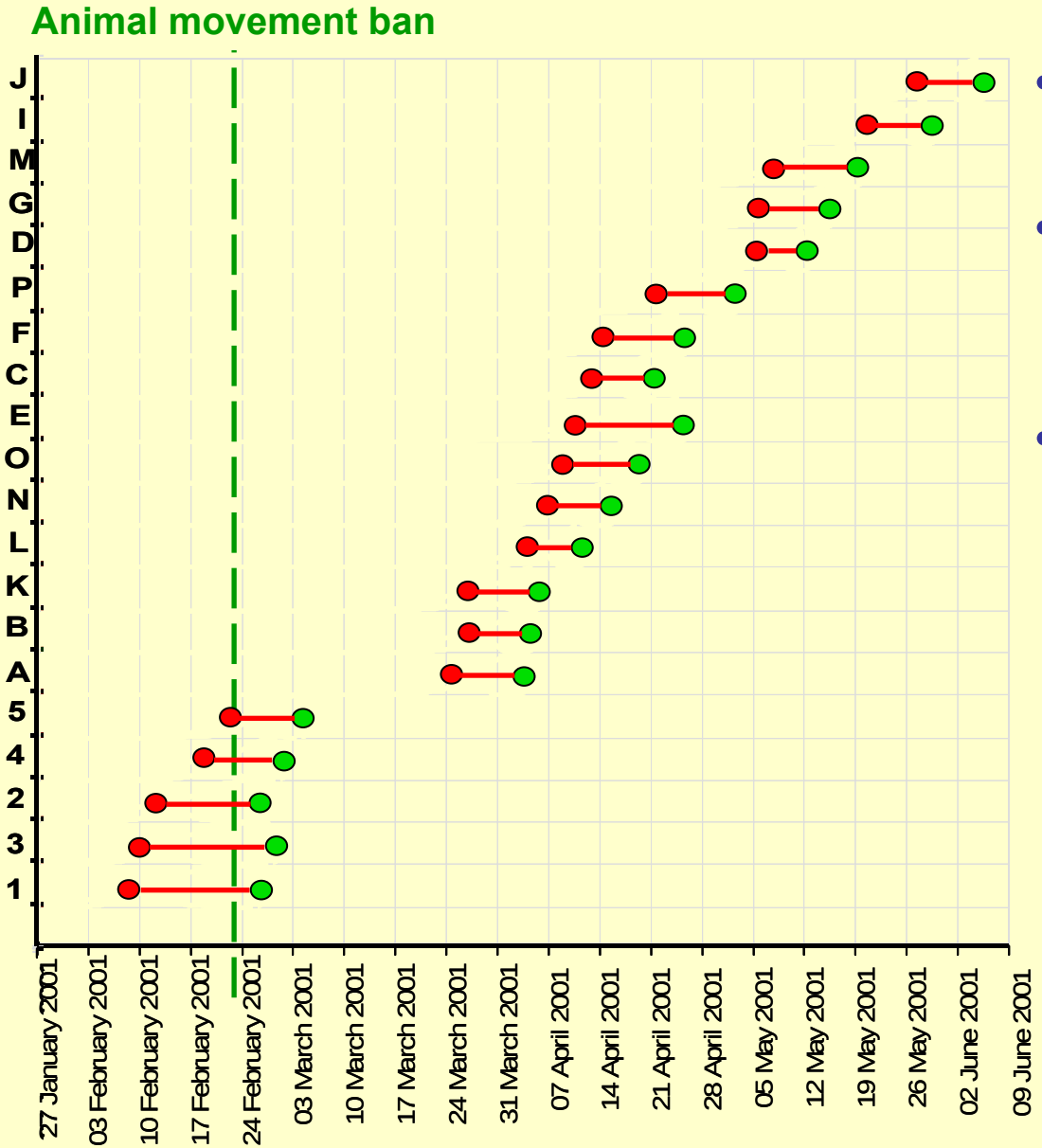


- Known root
- 1 known chain of transmissions
- 3 obvious transmissions
- What about the other ones??

Which is the most likely farm for each node ?

Use of contact tracing data

Epidemiology

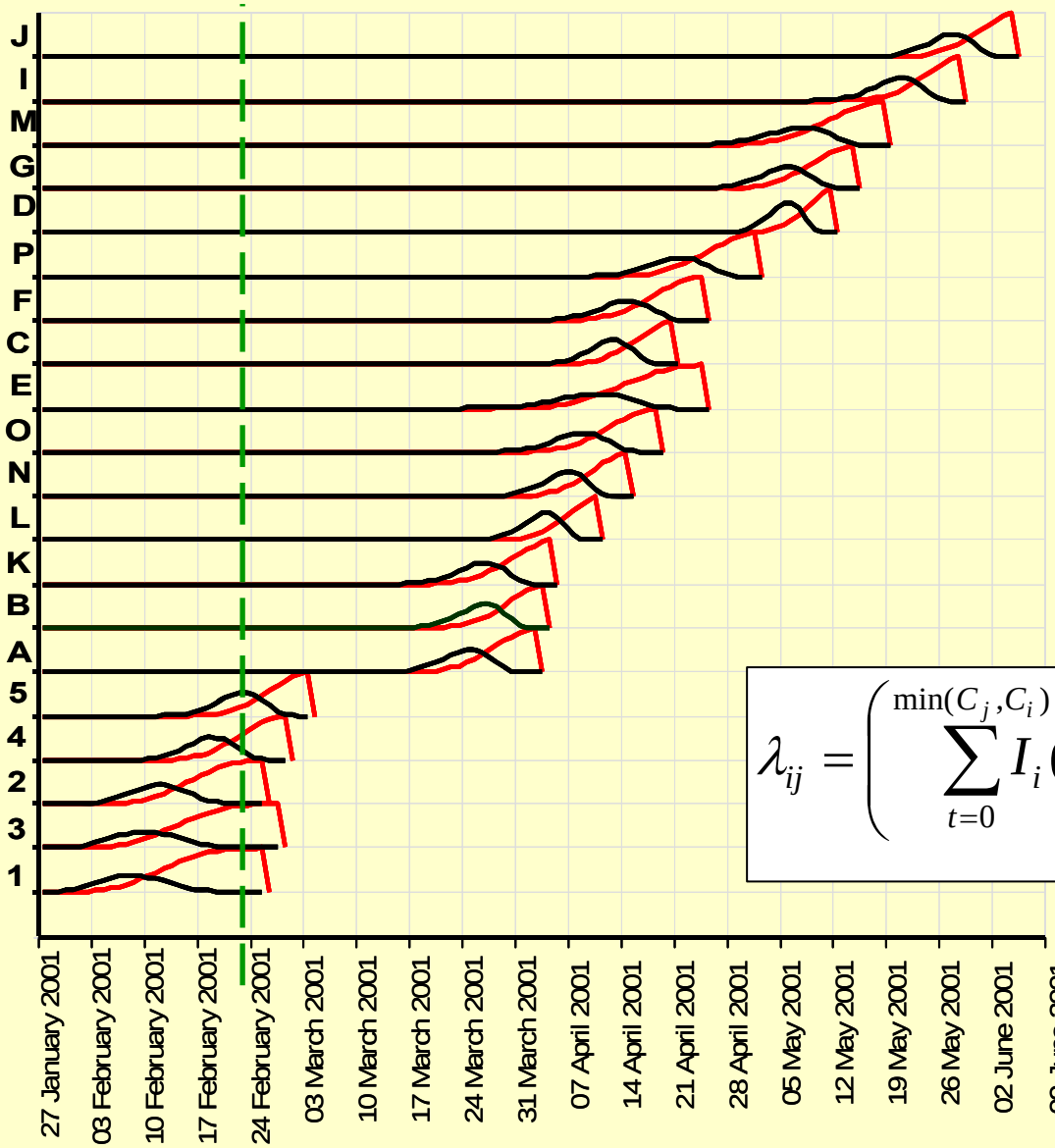


- L : Probability density for latency (Γ)
- I_i : Probability density for the infection date of farm *i*
- $F_i(t)$: Probability for farm i to be infectious at date t

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

$$t > C_i : F_i(t) = 0$$

Epidemiology



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

$$\lambda_{ij} = \left(\sum_{t=0}^{\min(C_j, C_i)} I_i(t) \cdot F_j(t) \right) \Big/ \sum_{\substack{k=1 \\ k \neq i}}^{n} \left(\sum_{t=0}^{\min(C_j, C_k)} I_i(t) \cdot F_k(t) \right)$$

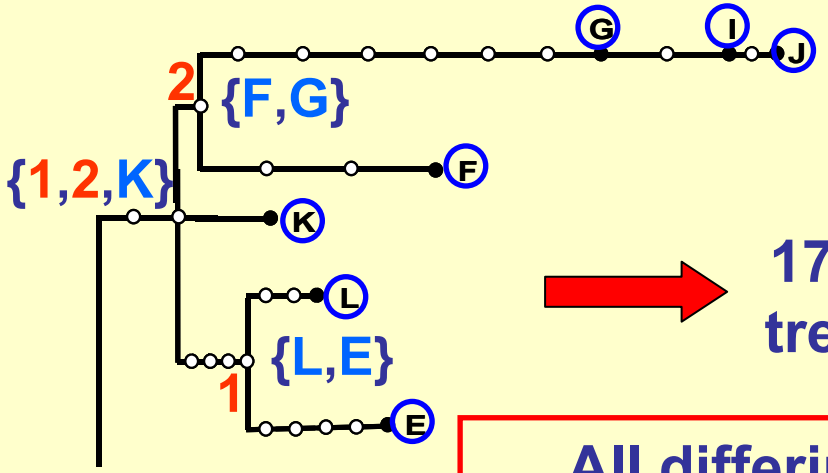
- λ_{ij}: Likelihood of i ← { j rather than another observed farm }

Genetics + epidemiology

- λ_{ij}: likelihood of i ← { j rather than another observed farm }
- λ_{ij} can be computed for each transmission
- Thus, for a complete transmission tree (*k*), $\lambda_k = \Pi\lambda_{ij}$
- And λ_k can be computed for any tree

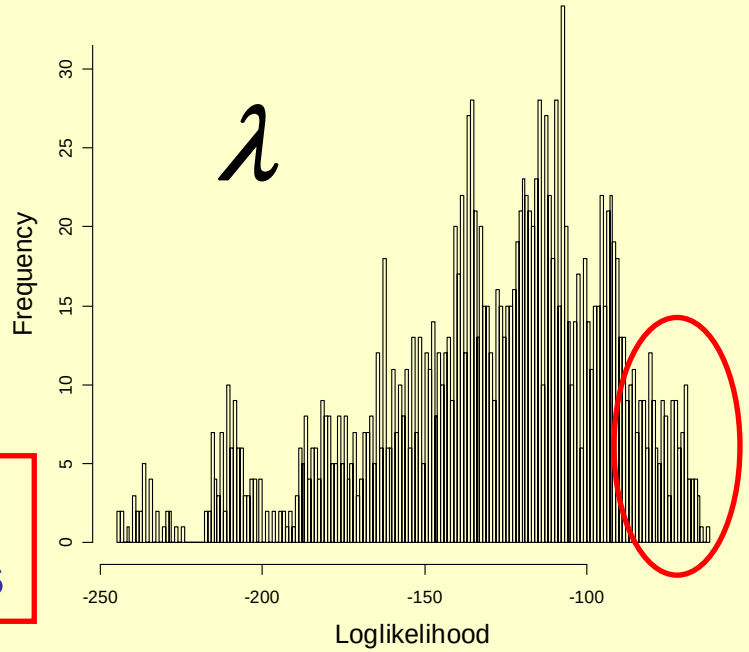
  … if all the possible trees can be enumerated

  → Algorithm defining the possible trees by recurrence from the leaves back to the root



1728 trees

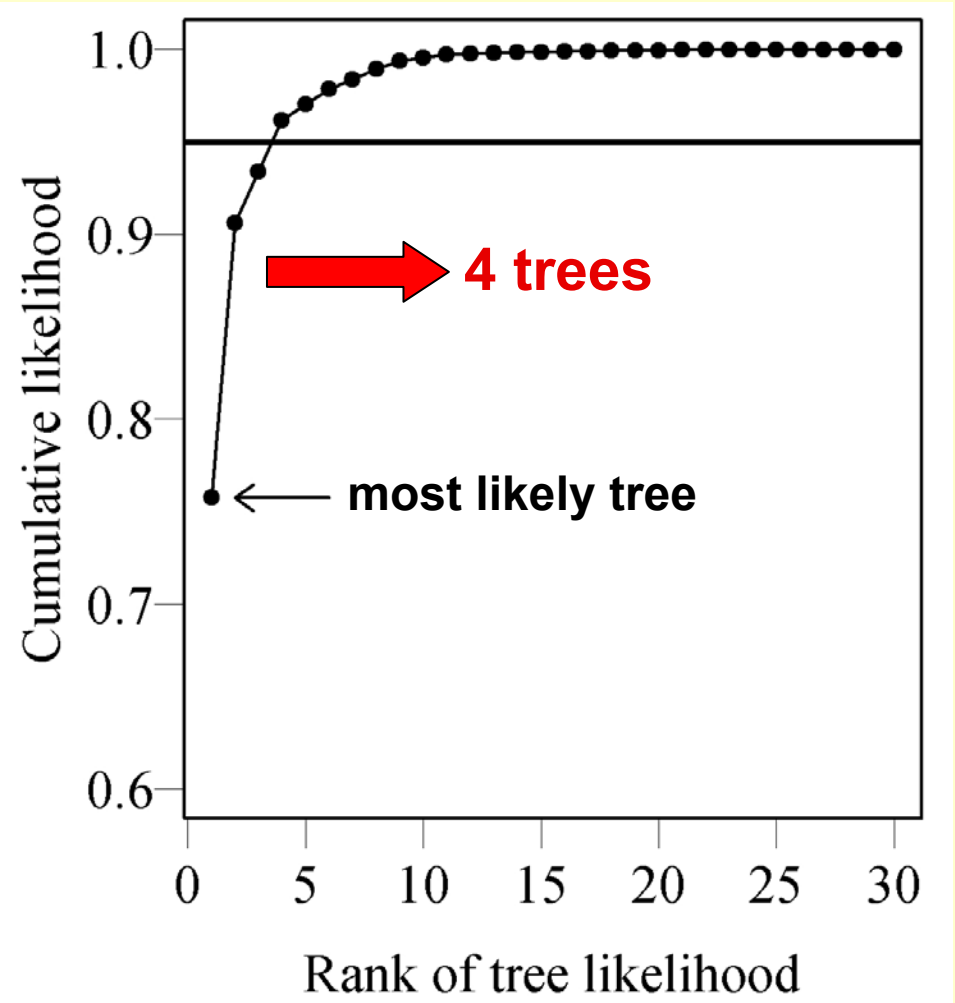
All differing from contact tracing results



Genetics + epidemiology

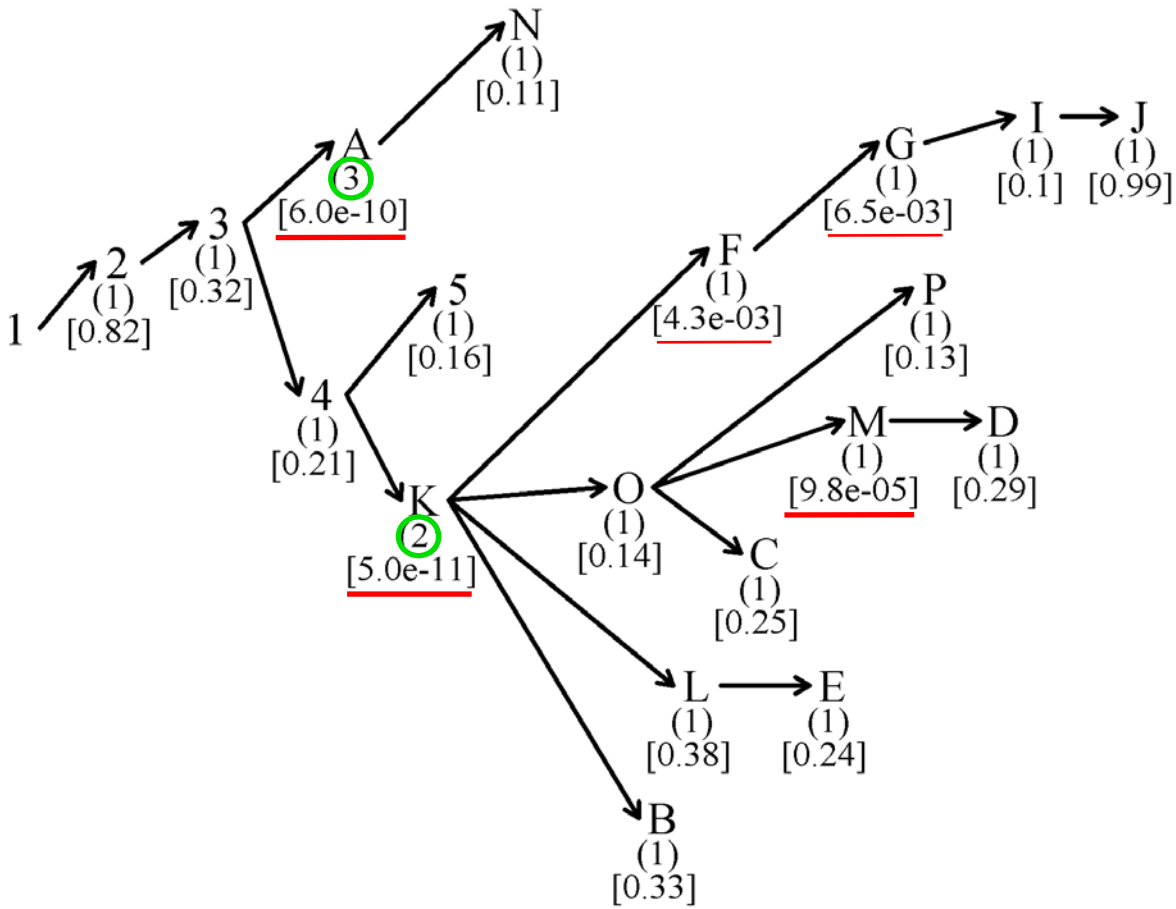
Which is the most likely group of trees?

- **Rescaled likelihood:**
  $\lambda'_k = \lambda_k / \Sigma\lambda_k$
- **Which group of trees represent 95% of the rescaled likelihood?**



Genetics + epidemiology

Which is the most likely tree?



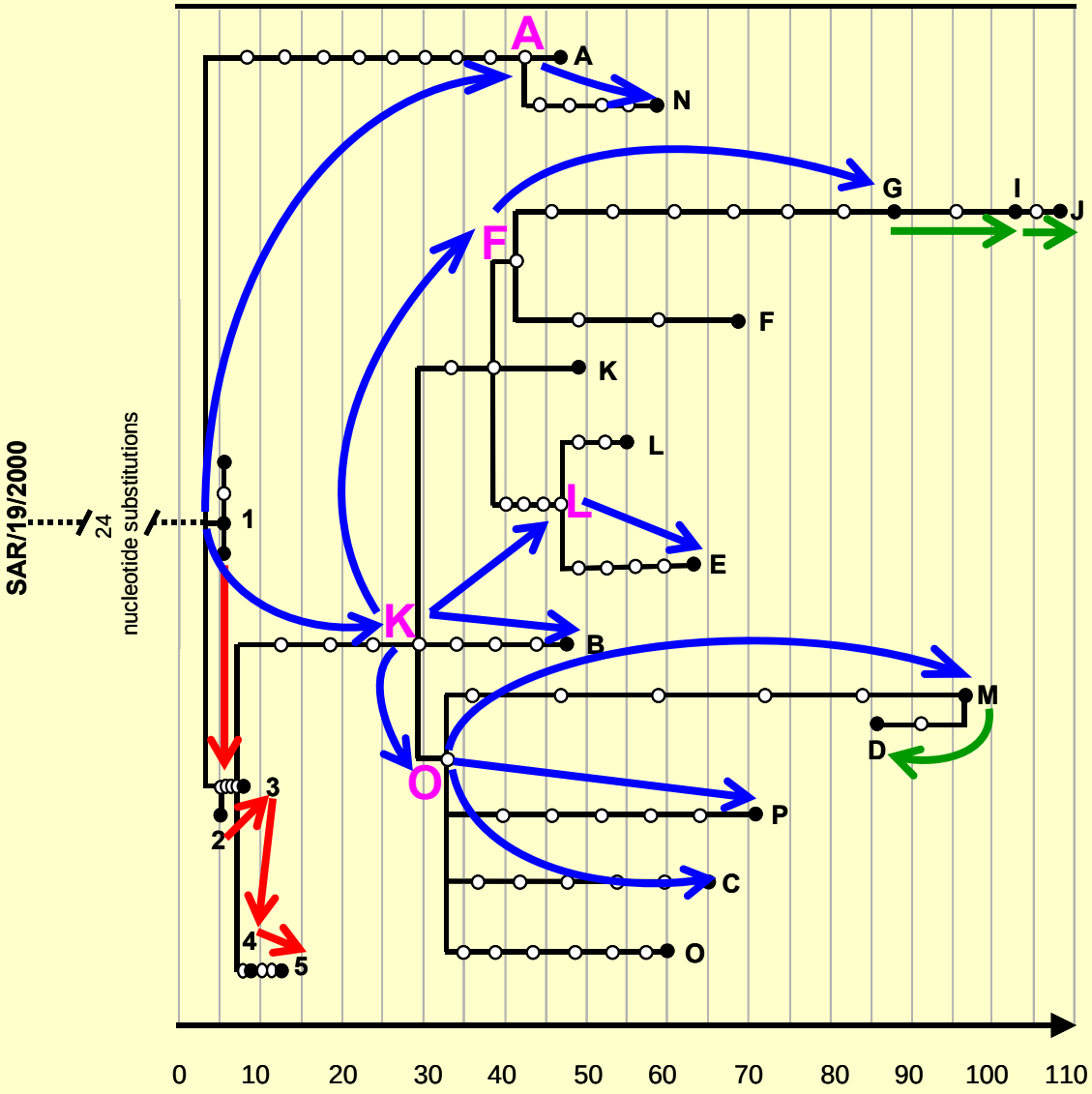
Source farm	Recipient farm	Likelihood ratio
3	A	27.3
4	K	5.1
A	N	2.1e+16
F	G	8.3e+11
K	B	168
K	F	4.5e+03
K	L	94.7
K	O	84.6
L	E	94.7
O	C	84.6
O	M	84.6
O	P	84.6

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

[#] Likelihood of the most probable transmission

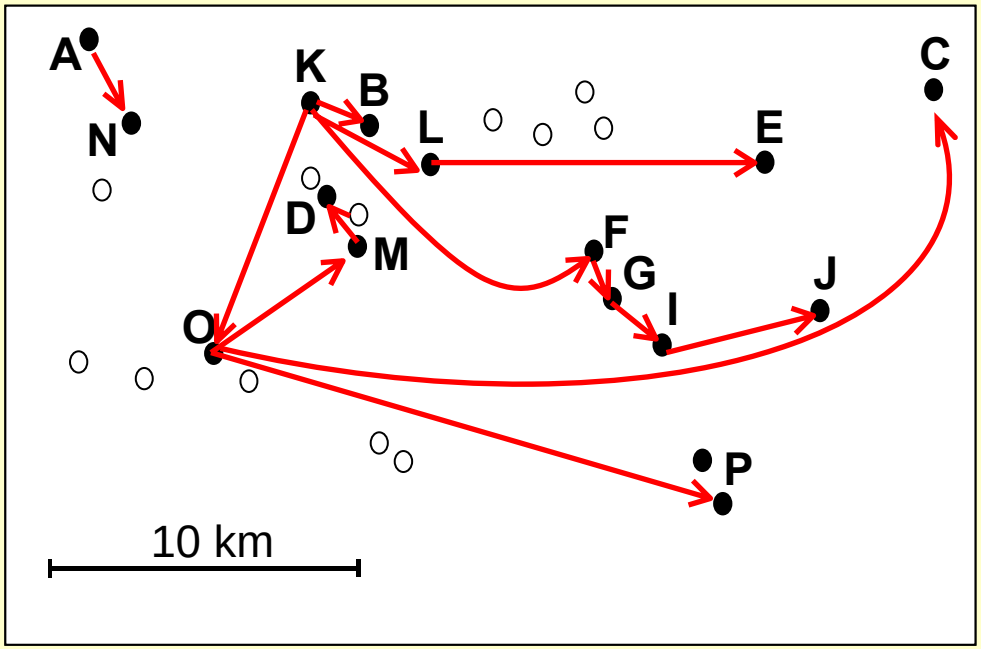
Genetics + epidemiology

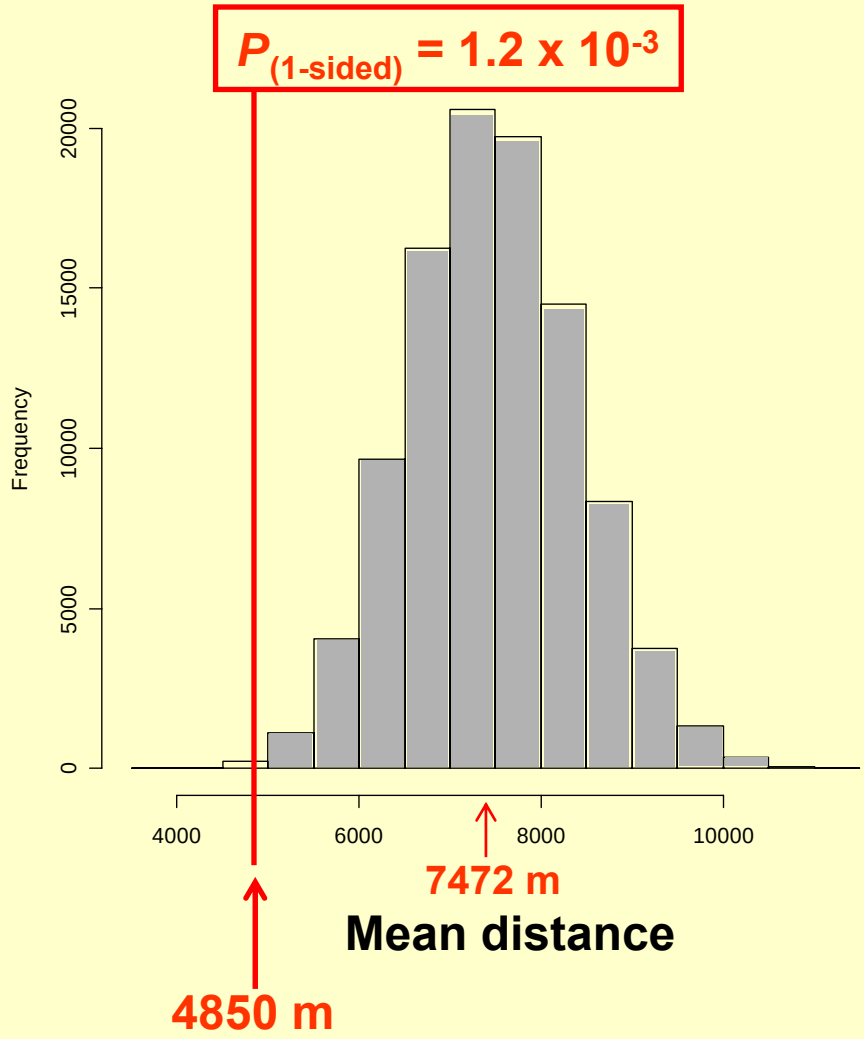
Which is the most likely tree?



Genetics + epidemiology

Spatial pattern





Short distance transmission

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

Cottam E.M. *et al*. (2008) Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* 275: 887-895.