INTERDISCIPLINARY QUICK TALKS

Open Data Open Possibilities

Open Genomic Data for Studying Infectious Diseases

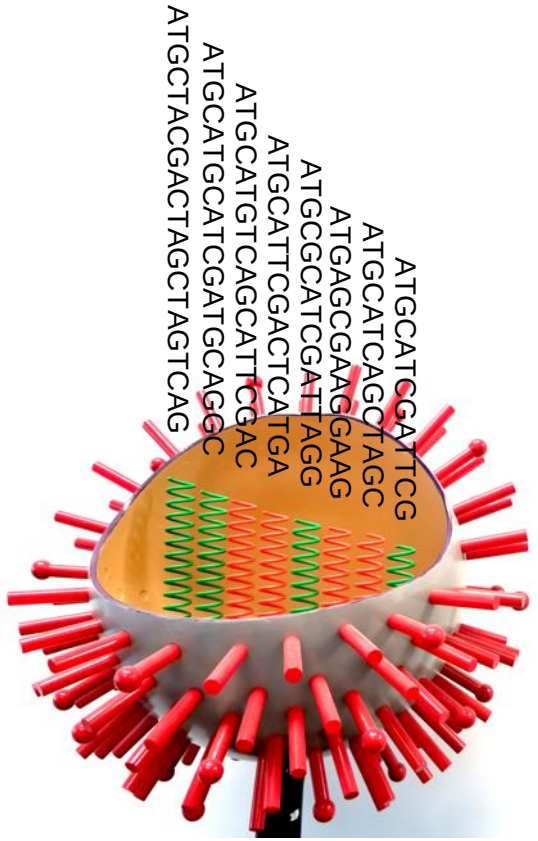
Dr Tommy Lam
School of Public Health

15 Feb 2019

Pathogen genomes for disease investigation

INTERDISCIPLINARY QUICK TALKS

Diagnosis	Precise pathogen identity, genotype, recombinant/reassortant
Predicting phenotype	Predict virulence, pathogenicity, transmissibility
Response to treatment	Antiviral drug resistance, vaccine escape, antimicrobial resistance
Source of infection	Human or animal origin, time and location of source
Transmission / epidemiology	Spatial dissemination pattern, outbreaks linkage, infected population size, transmission drivers

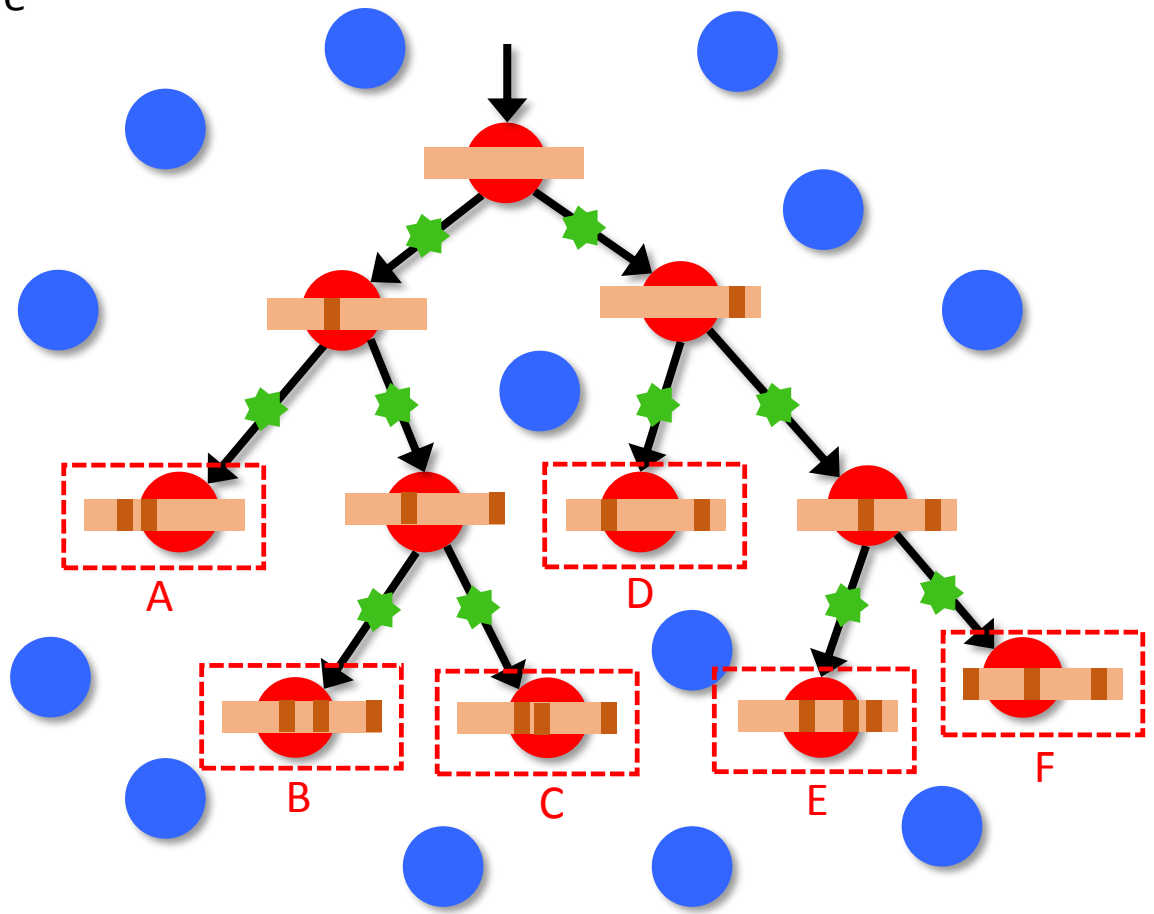


Transmission Imprinted in Pathogen Genome

Susceptible

Infected

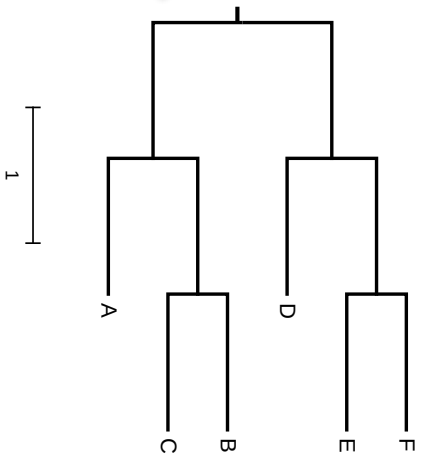
Sampled

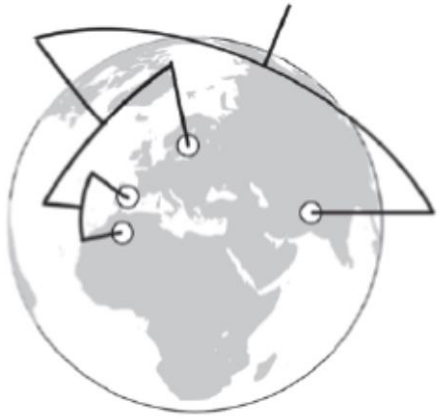


Pathogen genome sequences:

```
A|  AATTAAAAAAA
B|  ATATAAAAAAT
C|  AAATTAAAAAT
D|  AAAAATAATAA
E|  AAAAAATTTAA
F|  TAAAAATATAA
```

Build a **phylogenetic tree**





Pathogen & its genome

Mutation

Transmission history ≈ Evolutionary history

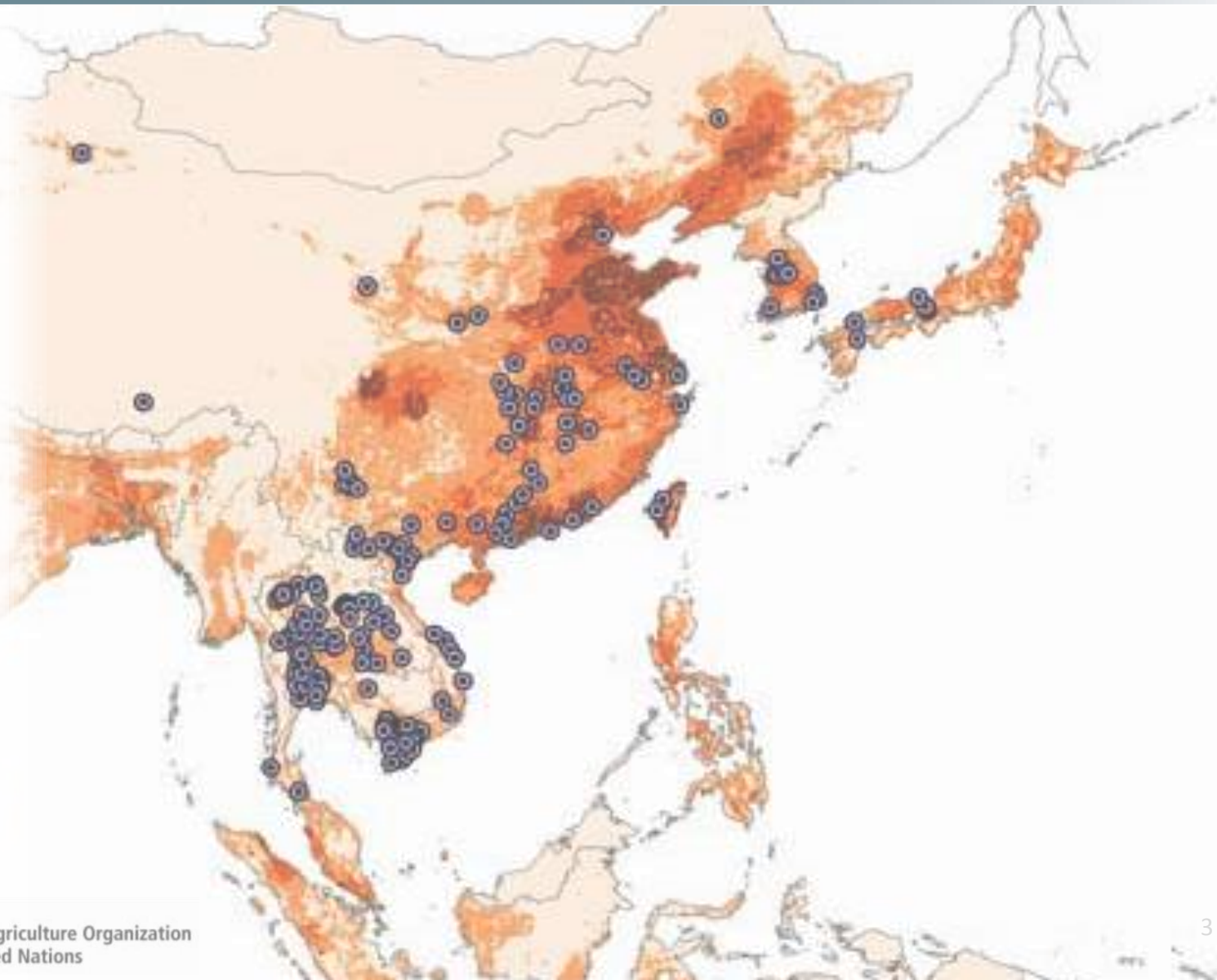
Spatial spread of infectious disease

Data: disease reports – geospatial data

Analysis methods:

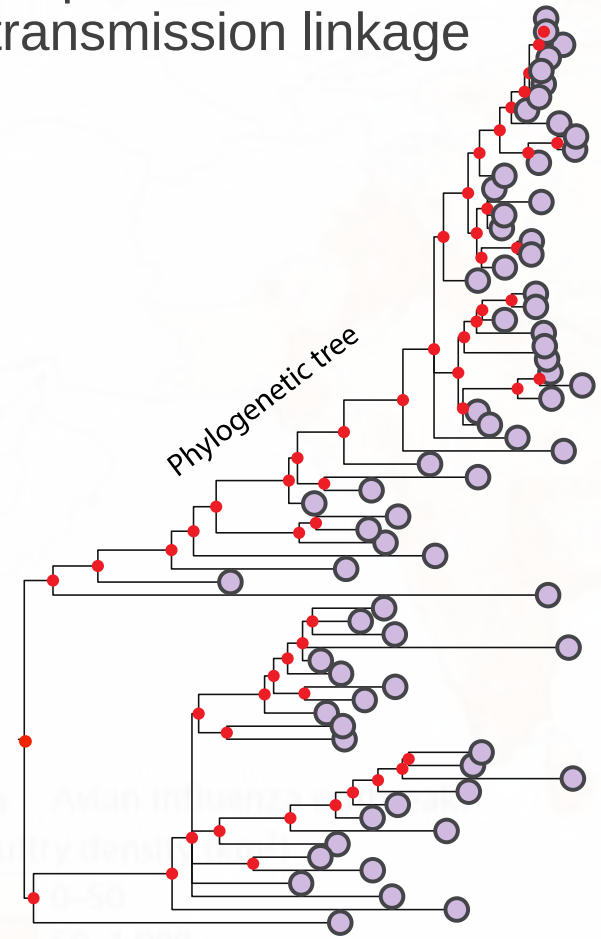
- Spatial clustering (Moran's I, Geary's c, Ripley's K, EMM, etc)
- Risk factor modeling (regression, discriminant analyses)
- Simulation (spatial-epidemic models)

Empirical evidence of transmission linkage between infections



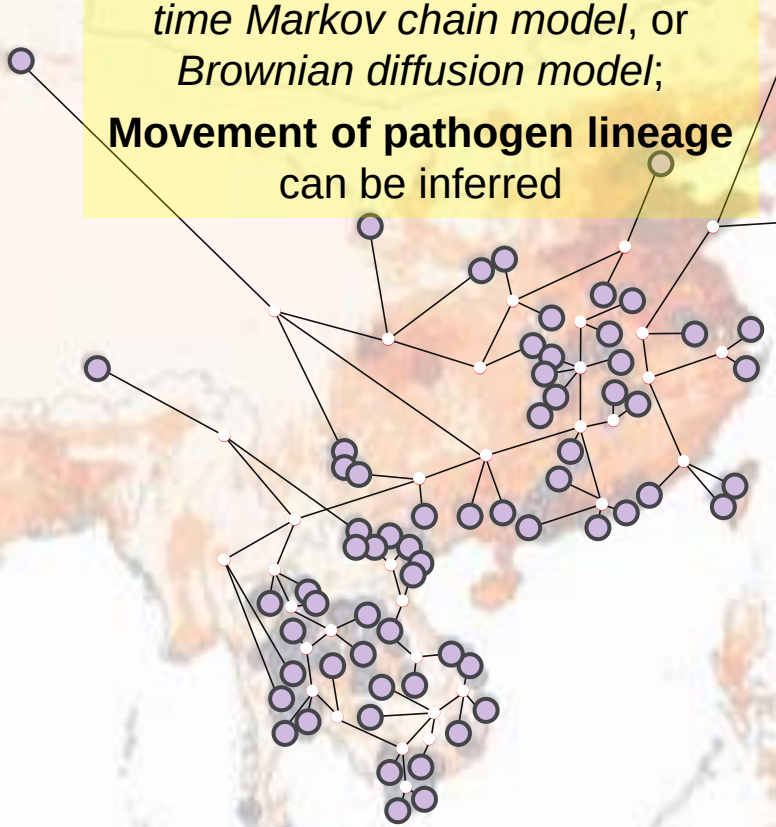
Phylogeographic inference

Phylogenetic tree provides empirical evidence of transmission linkage



Ancestral geographic locations are estimated with *continuous-time Markov chain model*, or *Brownian diffusion model*;

Movement of pathogen lineage can be inferred



Lemmon et al. Syst. Biol (2008)57

$$\delta_i = 2\,R \arcsin \times \left[\sqrt{\sin^2\left(\frac{\phi_k^r-\phi_j^r}{2}\right)+\cos(\phi_j^r)\cos(\phi_k^r)\sin^2\left(\frac{\lambda_k^r-\lambda_j^r}{2}\right)}\right],$$

$$L(\phi_e, \lambda_e, \psi|\phi_o, \lambda_o) = \prod\left(\frac{\delta_i}{b_i\psi^2}e^{-\delta_i^2/(2b_i\psi^2)}\right)$$

$$\frac{\partial}{\partial\psi}L(\phi_e, \lambda_e, \psi|\phi_o, \lambda_o) = \frac{(\xi-2)e^{-\frac{1}{2}\xi}}{\psi^{2n+1}}\prod\frac{\delta_i}{b_i}$$

Lemey et al. MBE (2010)27:8

$$f(g,\Theta,\Phi,\Omega|D) = \frac{1}{Z}\Pr\{D|g,\Phi,\Omega\}f_G(g|\Theta)f_{\Theta\Omega\Phi}(\Theta,\Omega,\Phi).$$

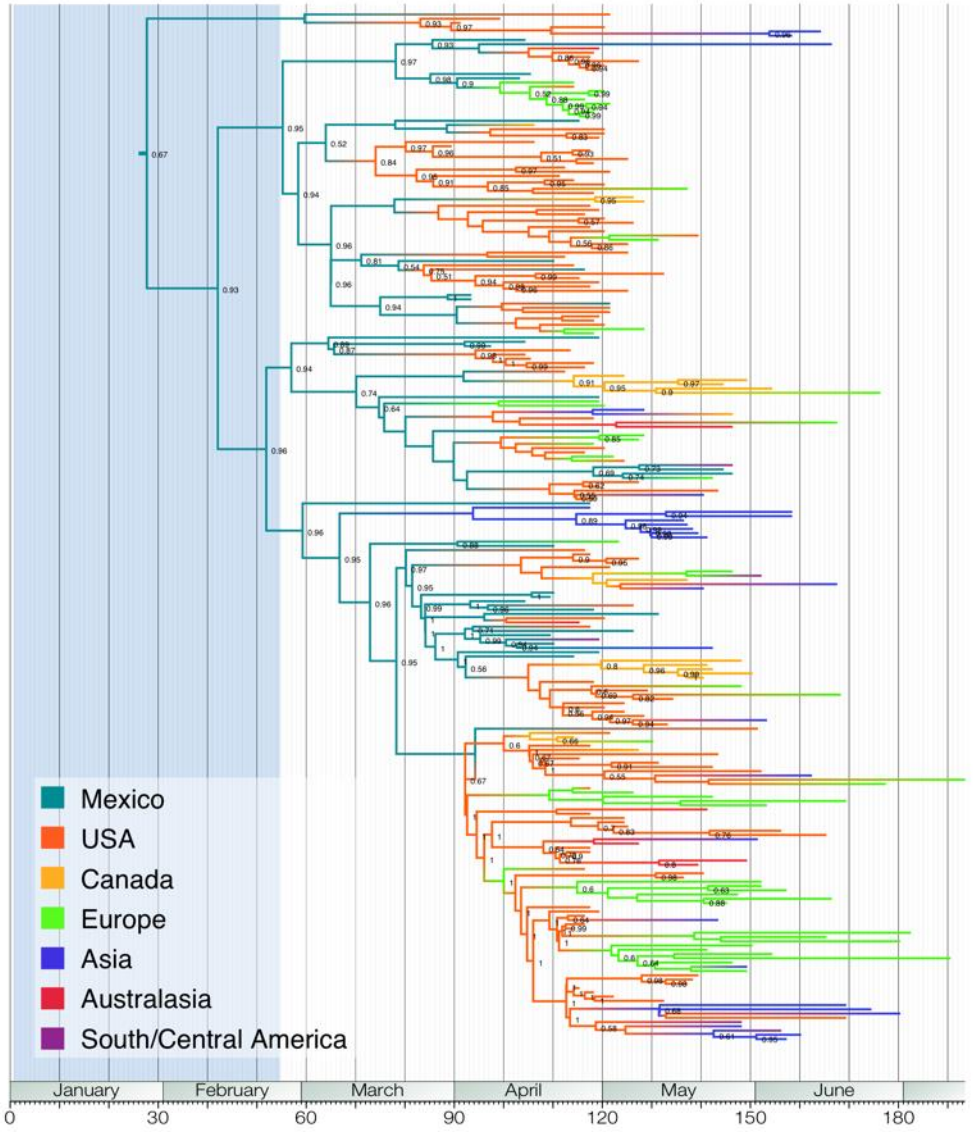
$$\phi_b \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\nu/2,\nu/2)$$

$$\phi_b \overset{\text{i.i.d.}}{\sim} \text{Lognormal}(1,\sigma),$$

Pandemic (H1N1) 2009 influenza

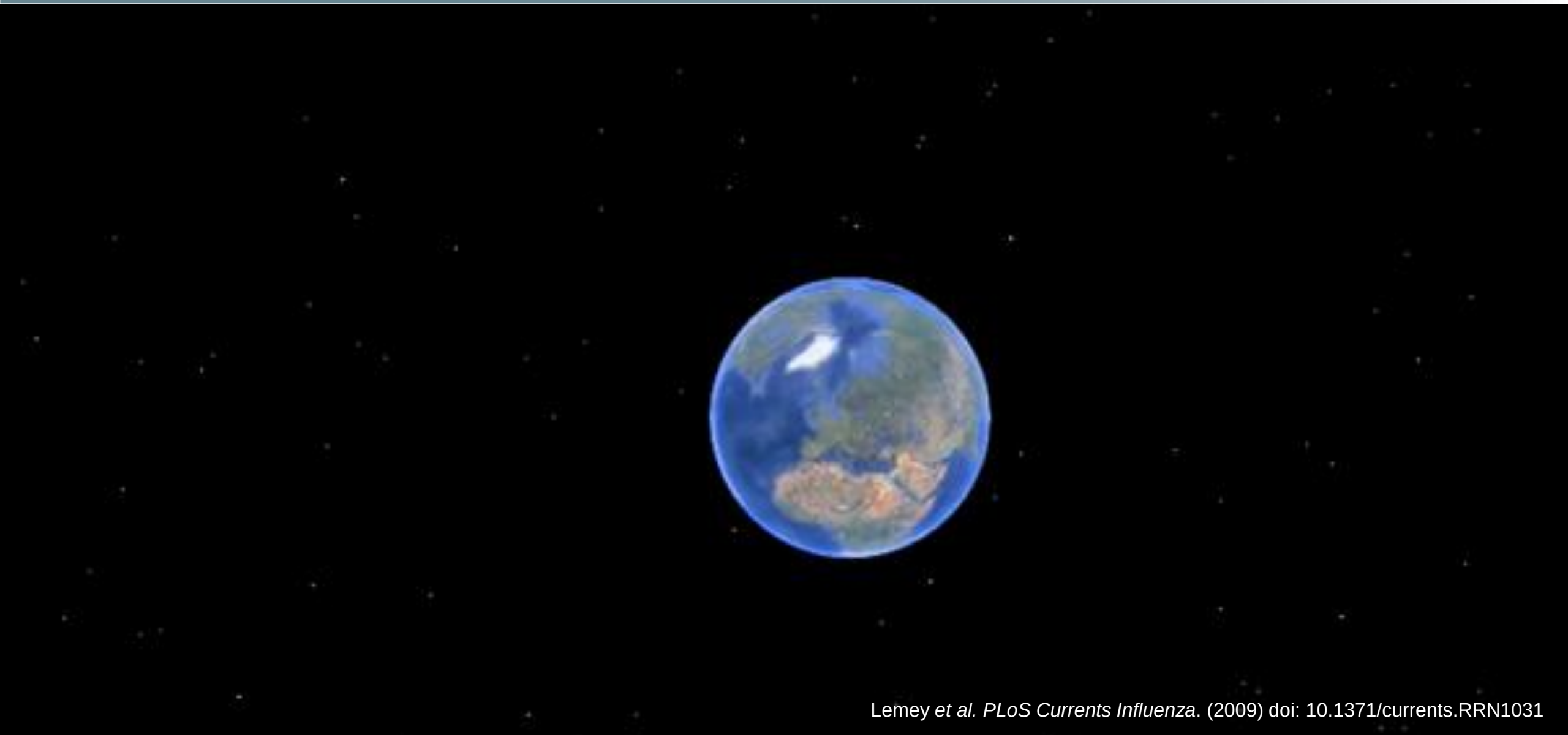
- A **novel** influenza A (H1N1) infection started in **California, 15th April 2009**
- Soon **spread to the world** within months
- Many **virological labs** sequenced the viruses, and deposited to **public databases**
- Analyzed the disease emergence and spread using **242 open viral sequences** (40 locations; Mar-Jul 2009), with **geographic labels**; using *relaxed molecular clock models* and *discrete continuous-time Markov chain* as the spatial diffusion model
- Need **bioinformatics specialist**

Lemey et al. PLoS Currents Influenza. (2009) doi: 10.1371/currents.RRN1031



Pandemic (H1N1) 2009 influenza

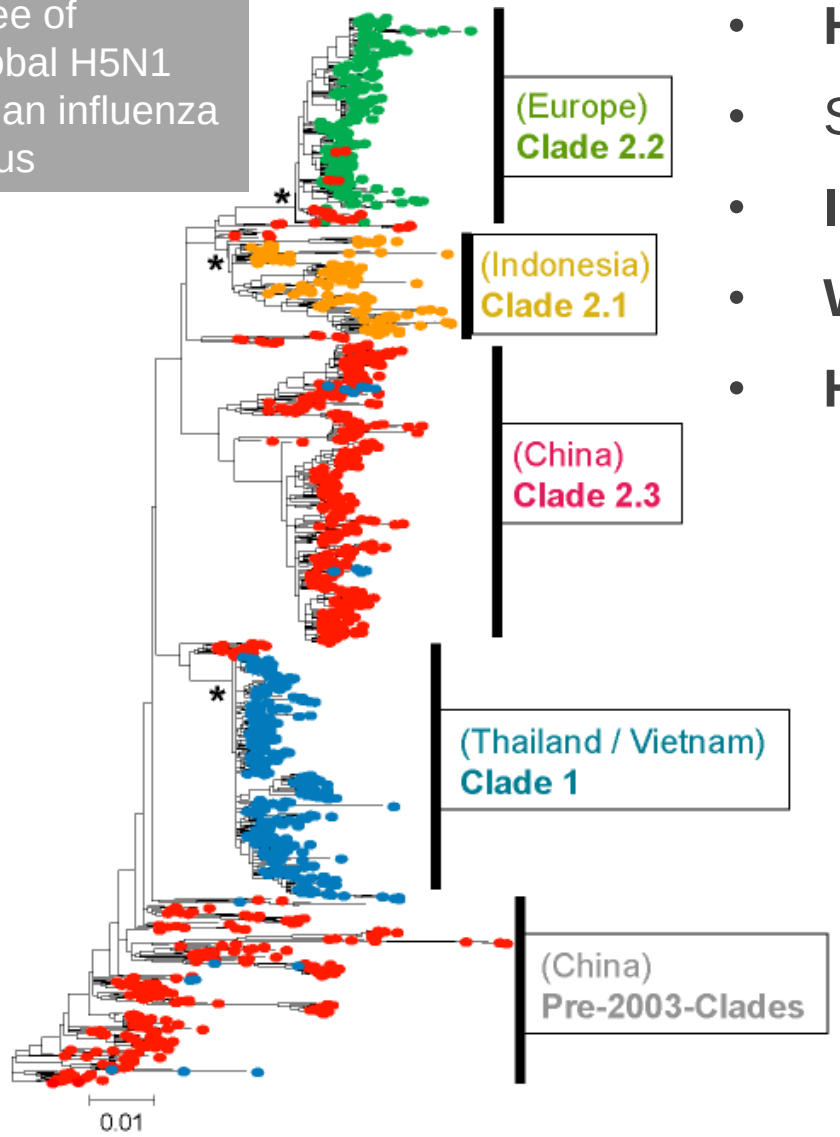
INTERDISCIPLINARY QUICK TALKS



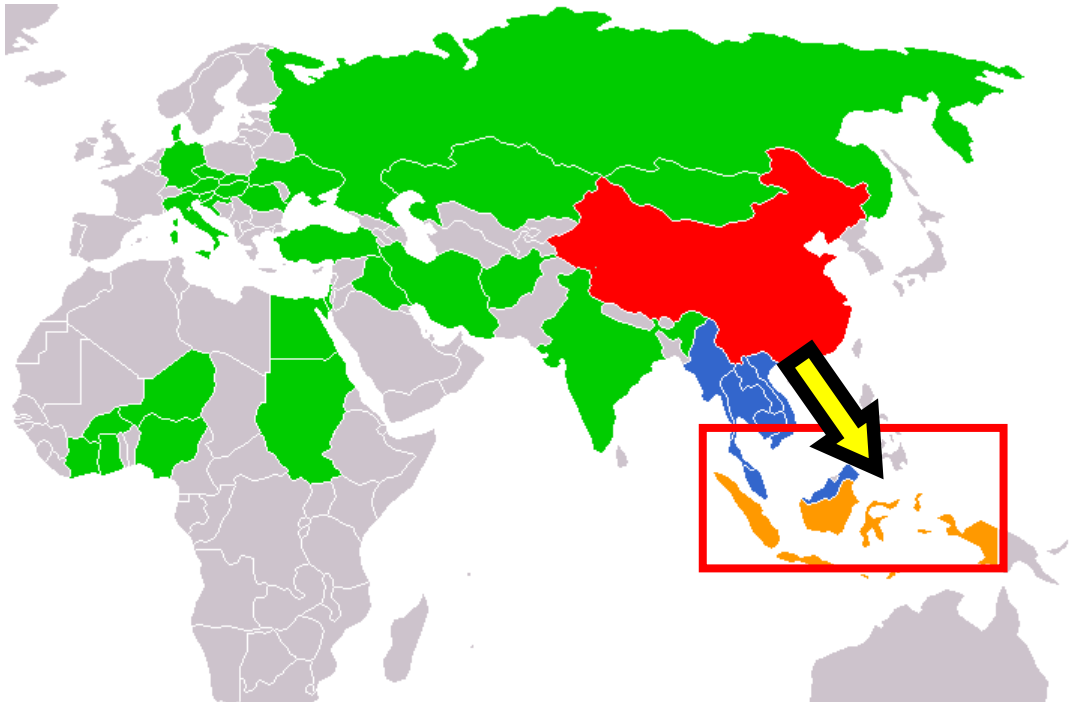
Lemey *et al. PLoS Currents Influenza*. (2009) doi: 10.1371/currents.RRN1031

Avian influenza (H5N1)

Tree of global H5N1 avian influenza virus



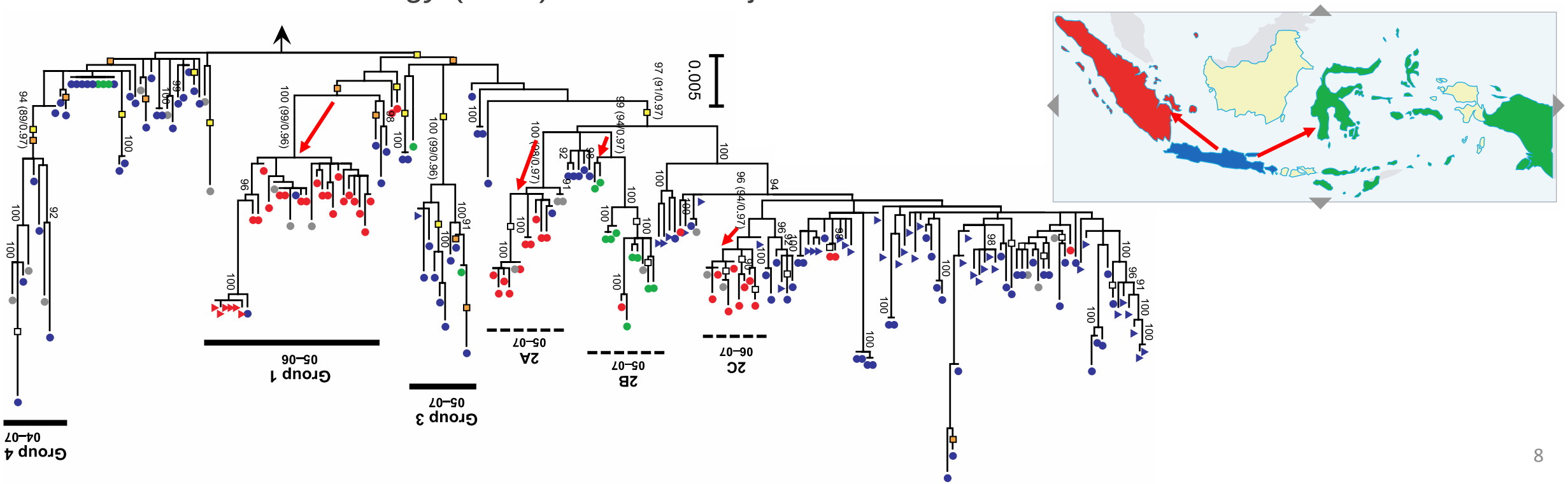
- **H5N1 highly pathogenic avian influenza** emerged in China, 1996.
- Spread to **Southeast Asia, including Indonesia** in 2003.
- **Indonesia** - caused big economic and health burden
- **Where** was it first introduced?
- **How** did it spread across the Indonesian islands?



Avian influenza (H5N1) - Indonesia

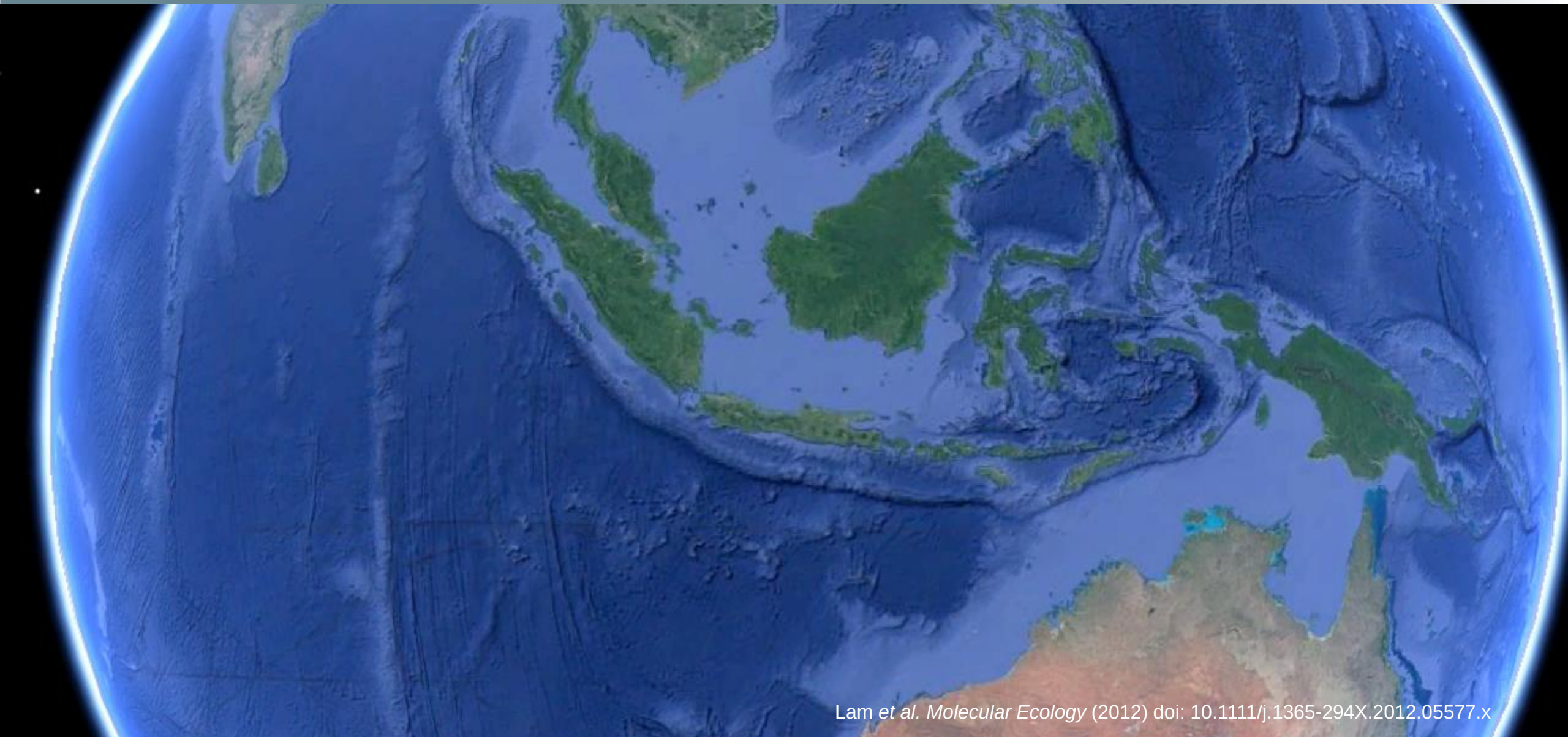
- Analyzed **213** open Indonesian H5N1 avian influenza gene sequences, with time and location information
- Introduced to **Java Island** and spread to other islands
- How did it **spread inside** Java Island?

Lam *et al. Molecular Ecology* (2012) doi: 10.1111/j.1365-294X.2012.05577.x



Avian influenza (H5N1) - Indonesia

INTERDISCIPLINARY QUICK TALKS



Lam *et al. Molecular Ecology* (2012) doi: 10.1111/j.1365-294X.2012.05577.x

Benefits of Open Genomic Data

- Encourage **Multidisciplinary** Collaboration, **Full Use** of Data
  - Microbiologist and Bioinformatician
- Allow **Massive Analyses**
  - First Swine Flu genome released on 24th April 2009, many analyses published in personal blogs
- Increase **Completeness** in Analyses
  - Genetic data from every region are important for powerful analytics and interpretations

More Insights, Better Prevention & Control, Healthier City

Open Genomic Data for Hong Kong

- There were pathogen genome sequences published by HK Gov Labs (e.g. influenza, norovirus) - Are all data open?
- Time is key for infectious disease control and prevention
- Less / solvable privacy issue for pathogen data
- E.g. Dengue virus genome sequences
  - Determine local transmission or foreign introductions
  - Source of introductions
  - Time of local transmission started

Links to Further Information

Analysis on open *Swine Flu* genome data:

- Lemey *et al.* PLoS Currents Influenza. (2009) http://currents.plos.org/influenza/index.html%3Fp=4725.html

Analysis on open Indonesian *Avian Flu* genome data:

- Lam *et al.* PLoS Pathogen (2008) 4 https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000130
- Lam *et al.* Molecular Ecology (2012) 12 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2012.05577.x

Figure Credits

- Slide#2: https://www.cell.com/trends/ecology-evolution/pdf/S0169-5347(11)00354-5.pdf
- Slide#3,4: http://www.fao.org/docrep/007/y5537e/y5537e03.htm

THANK YOU