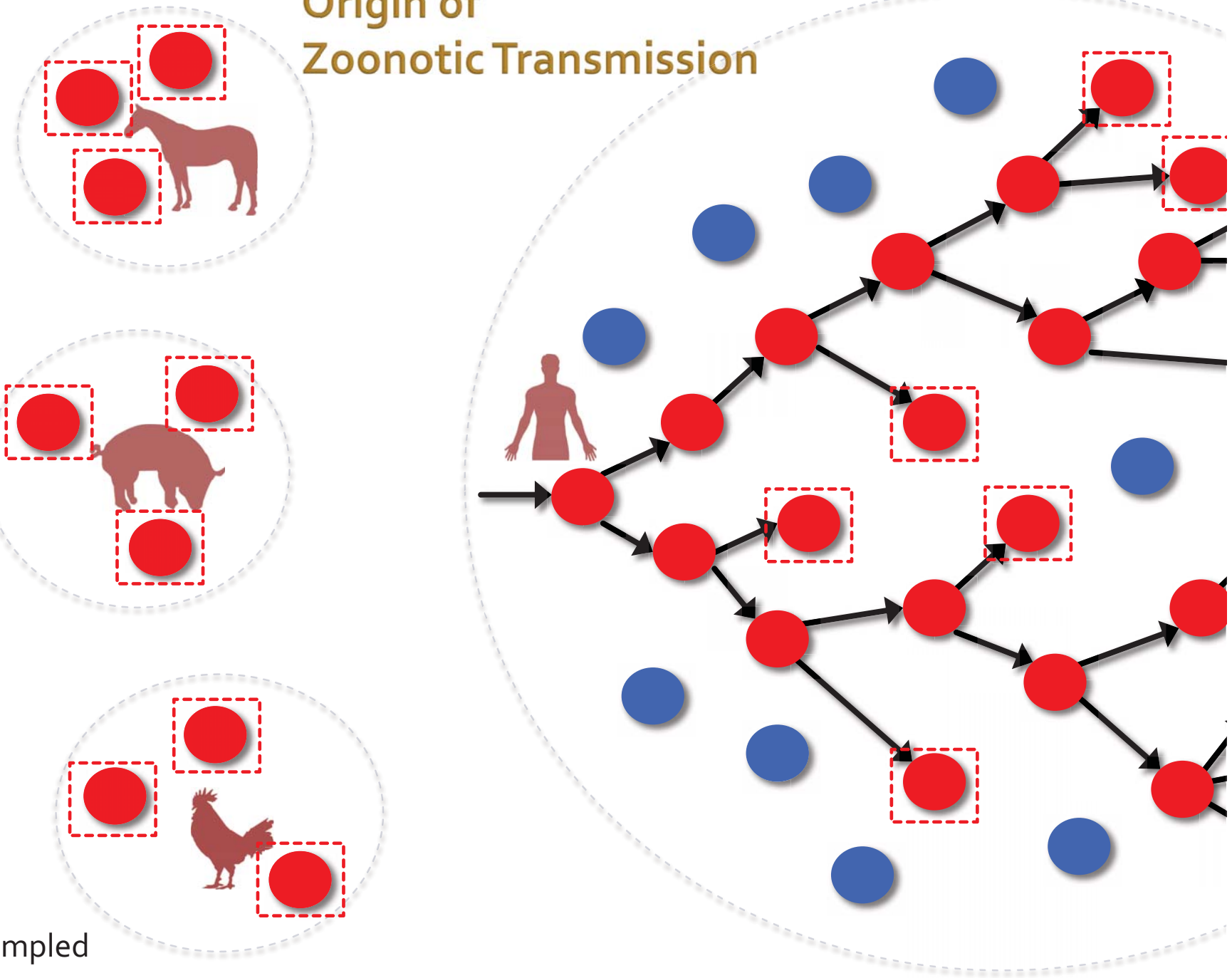
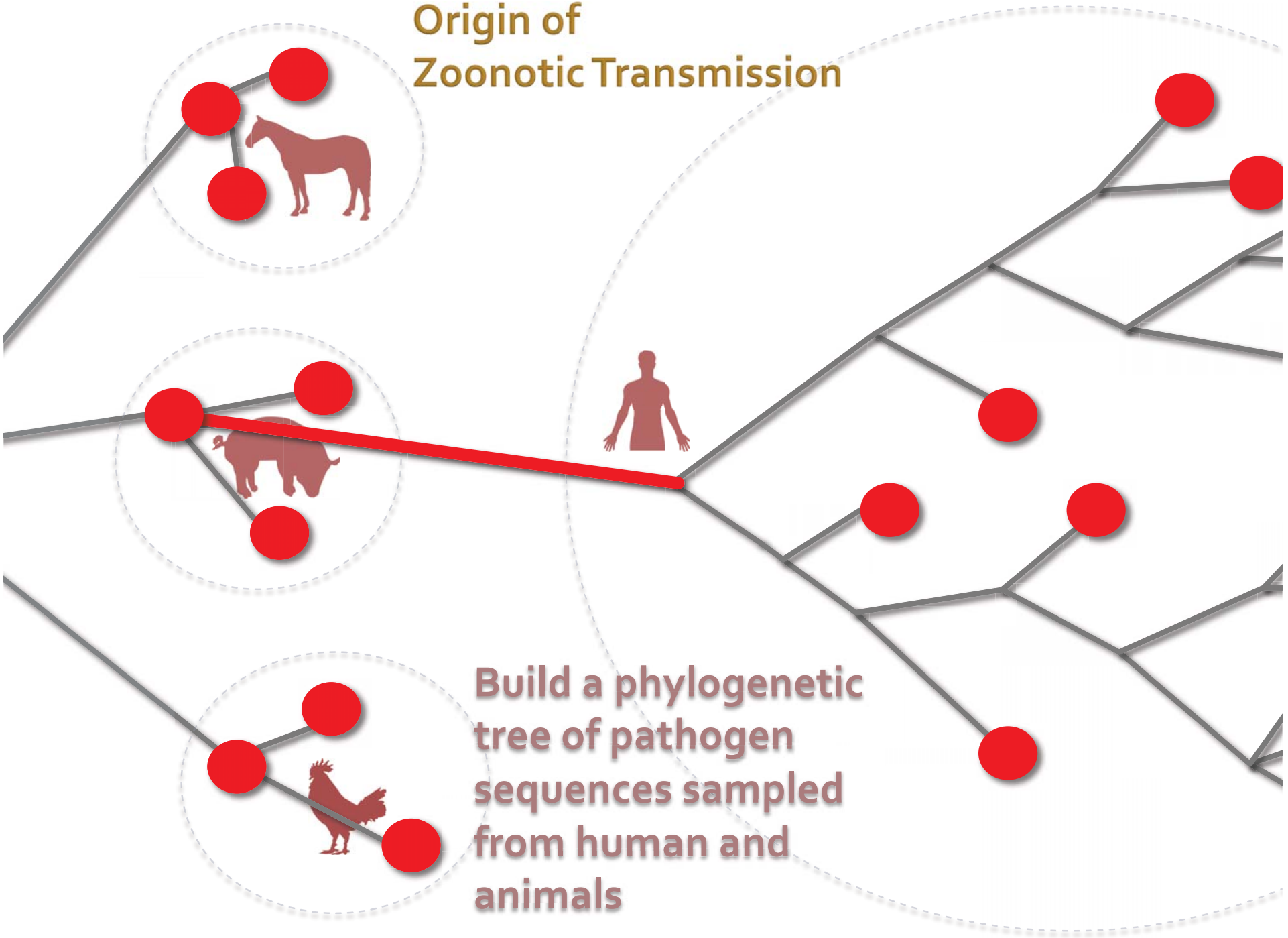
Qualitative interpretations of phylogenetic trees: **Examples**

Origin of Zoonotic Transmission



Sampled


Origin of
Zoonotic Transmission
Build a phylogenetic tree of pathogen sequences sampled from human and animals

Tracing the source or origin of diseases: H1N1 human influenza pandemic 2009

N Engl J Med *2009; 361:115-119*

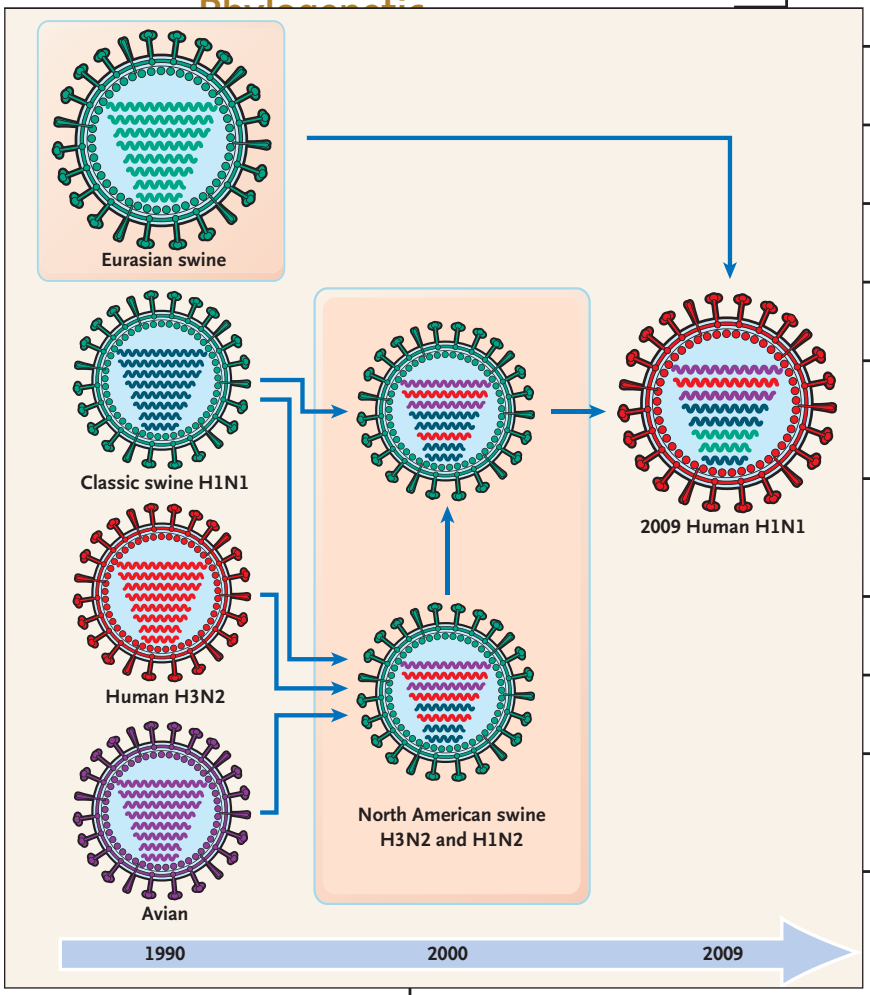


Figure 1. History of Reassortment Events in the Evolution of the 2009 Influenza A (H1N1) Virus.

- Duck/Pennsylvania/446080
- Mallard/Washington/20010
- Chicken/California/6643
- Swine/Texas/01657
- Swine/Hong Kong/NS1890
- Swine/North Carolina/98225
- Texas/04
- California/04

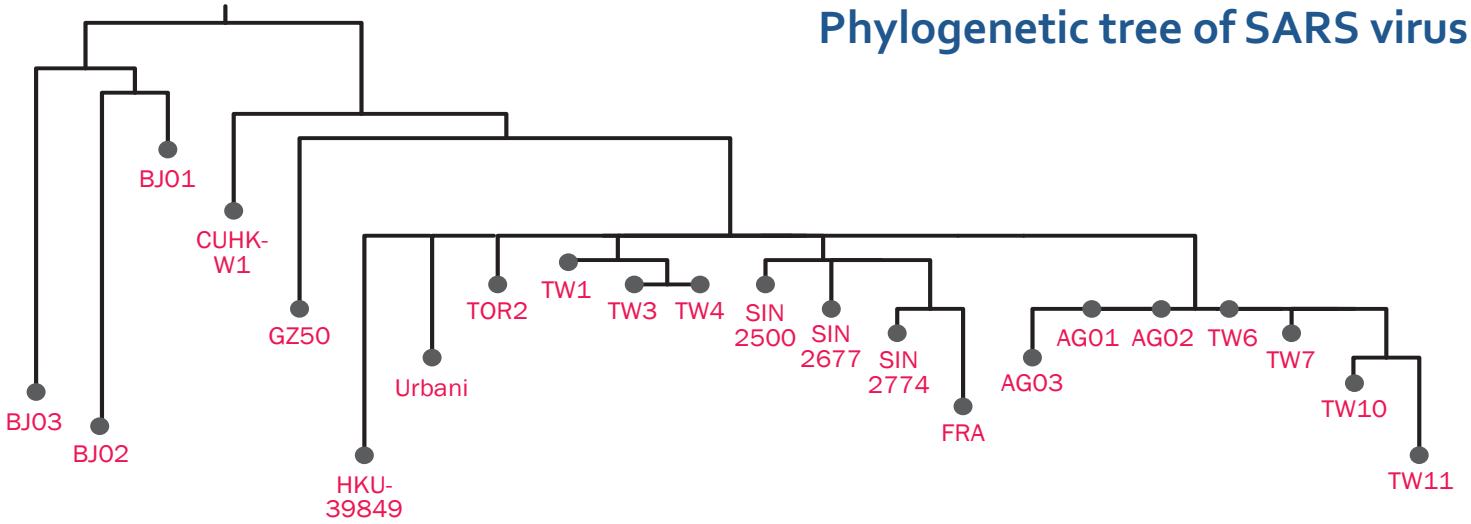
Novel H1N1 2009

- Equine/Xinjiang/3
- Equine/California/103
- Equine/Austria/421
- California/1243P
- Shanghai/LWS1
- Texas/1239P
- South Korea/1171P

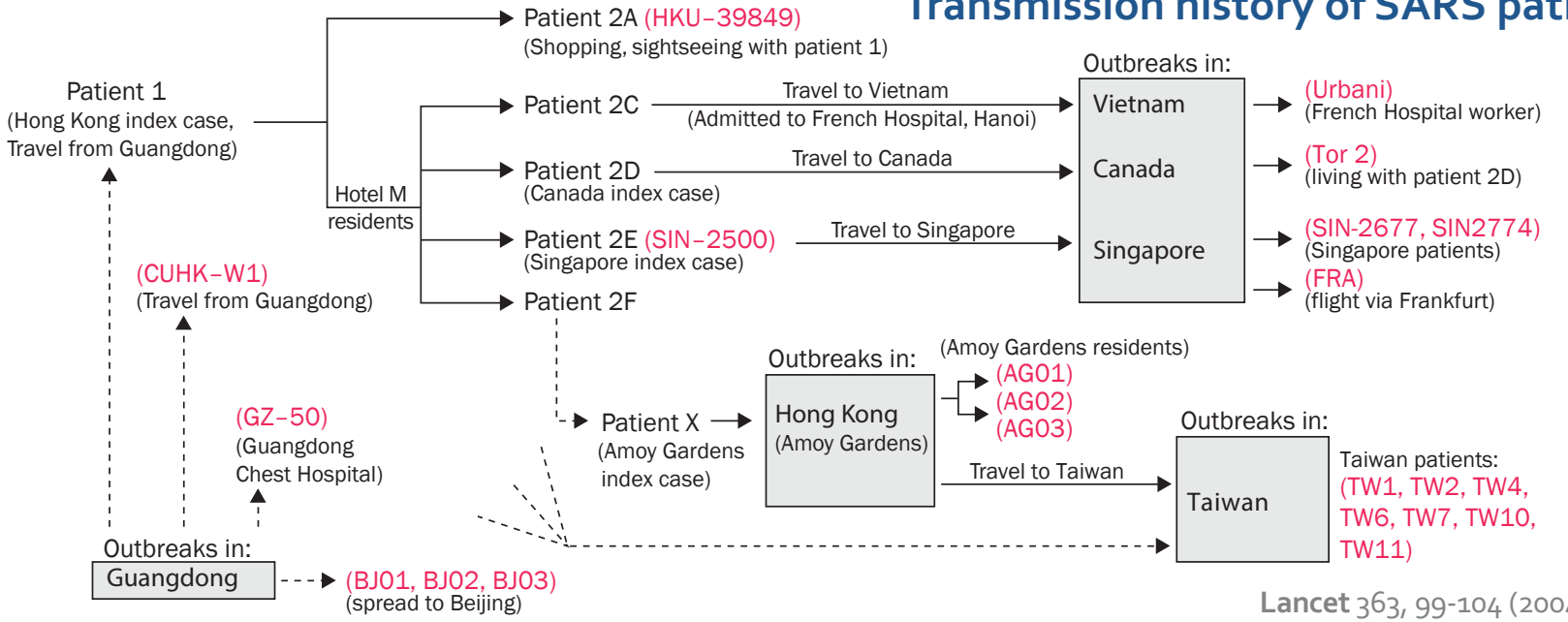
Seasonal H1N1 & H3N2

Examples – SARS Outbreak 2003

Phylogenetic tree of SARS virus genomes



Transmission history of SARS patients



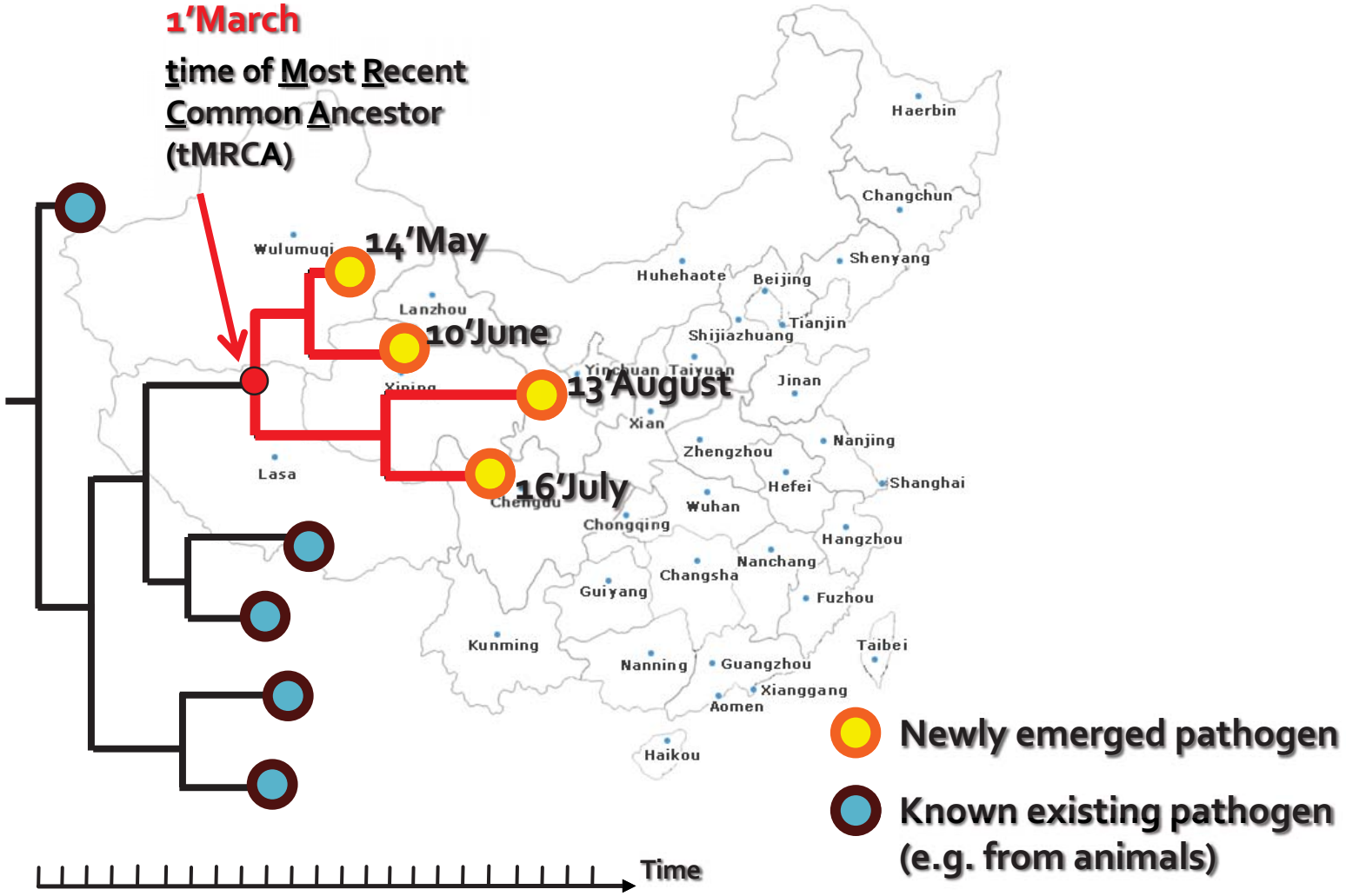
Lancet 363, 99-104 (2004).

Summary

- **Mutations accumulated** in the pathogen genome sequence along the **transmission chain.**
- Phylogenetic tree inferred from these sequences illustrate the **evolutionary history** of the pathogens, which is a reflection of the **disease transmission history.**
- Raw sequences must be **aligned** to build the phylogeny.
- Phylogenetic tree can be estimated with Markov model on formal **statistical framework** (e.g. **maximum likelihood**, **Bayesian**).
- **Qualitative interpretations** on disease origin and related transmission.

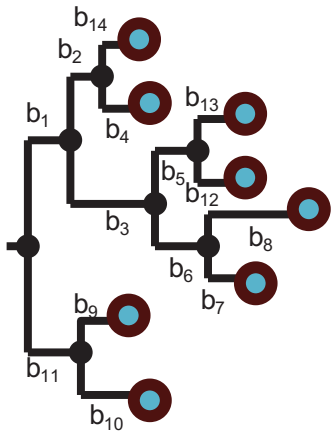
Quantitative inferences on phylogenetic trees – **Epidemic timescale**

'Dating' the starting time of epidemic

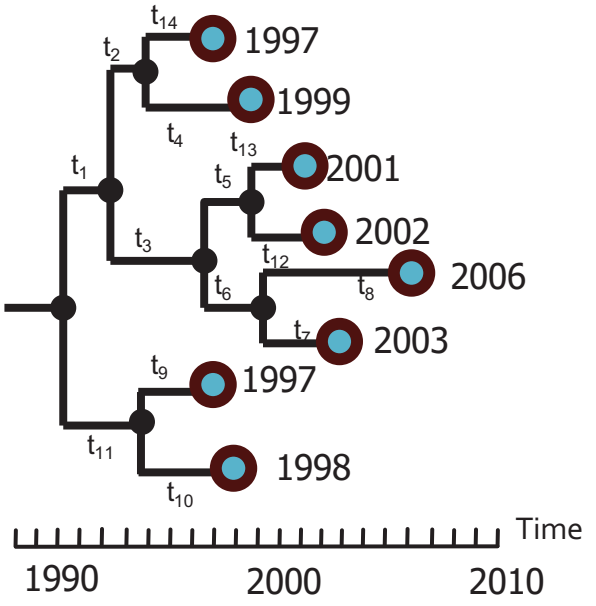


Time-scaled tree

- A tree measured in genetic distance is transformed to a tree in a time scale.
- $\mathbf{b_i}$ is known, but $\mathbf{t_i}$ is unknown.

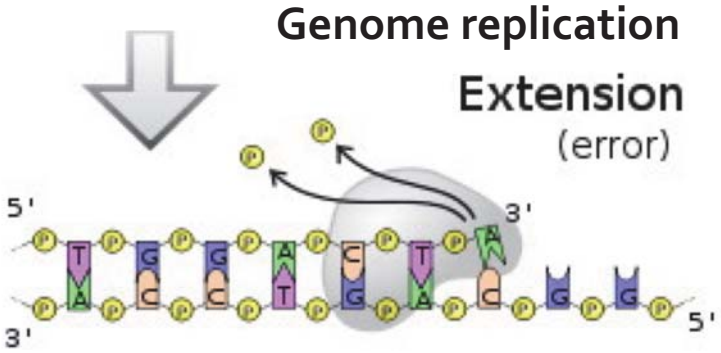


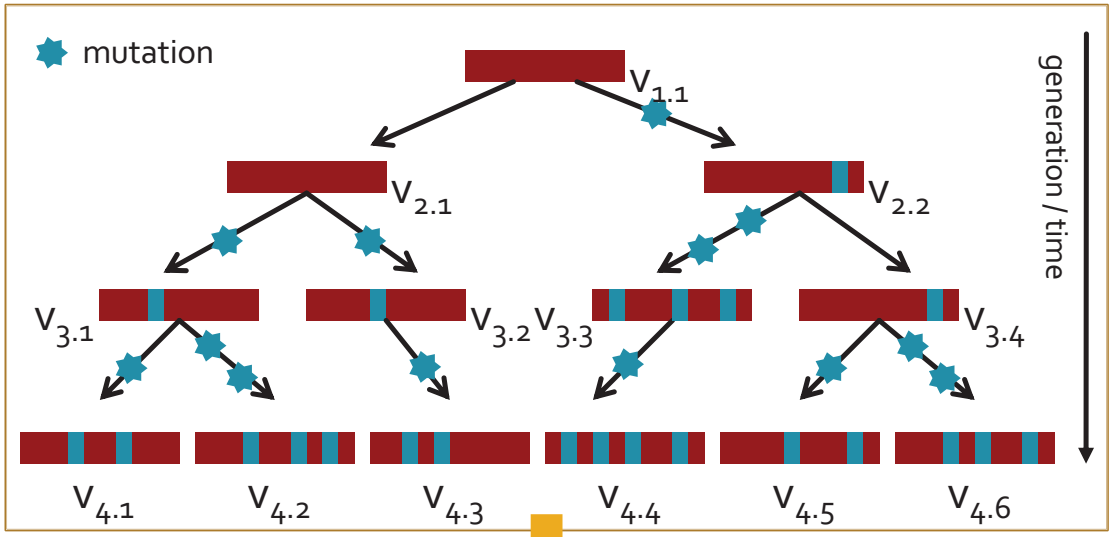
Branch lengths, $\mathbf{b_i}$, in unit of genetic substitutions

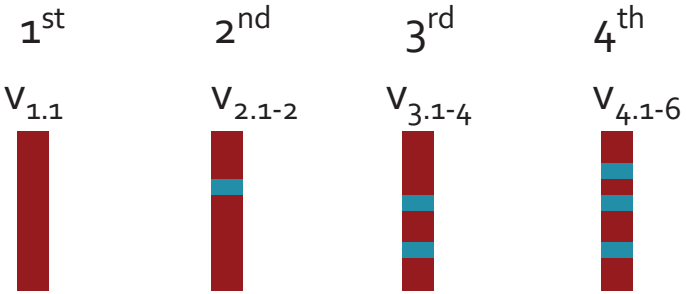


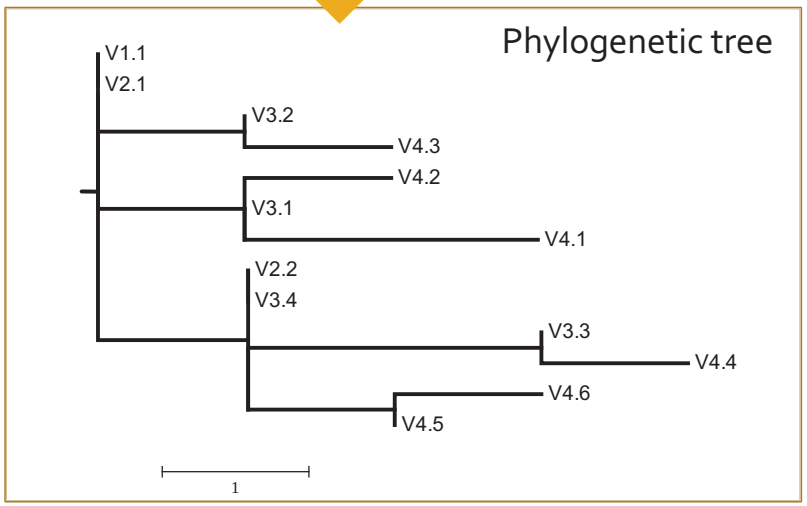
Branch lengths, $\mathbf{t_i}$, in unit of year.

Accumulation of mutations from generations to generations

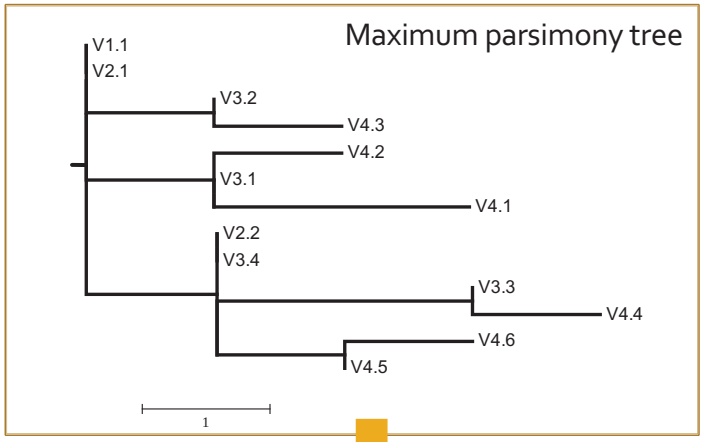


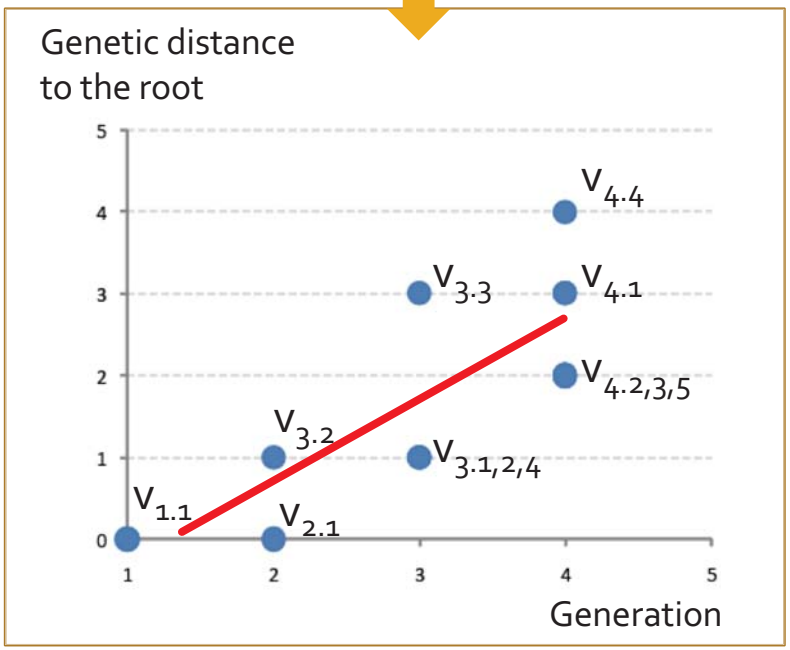






Tempo pattern of evolution

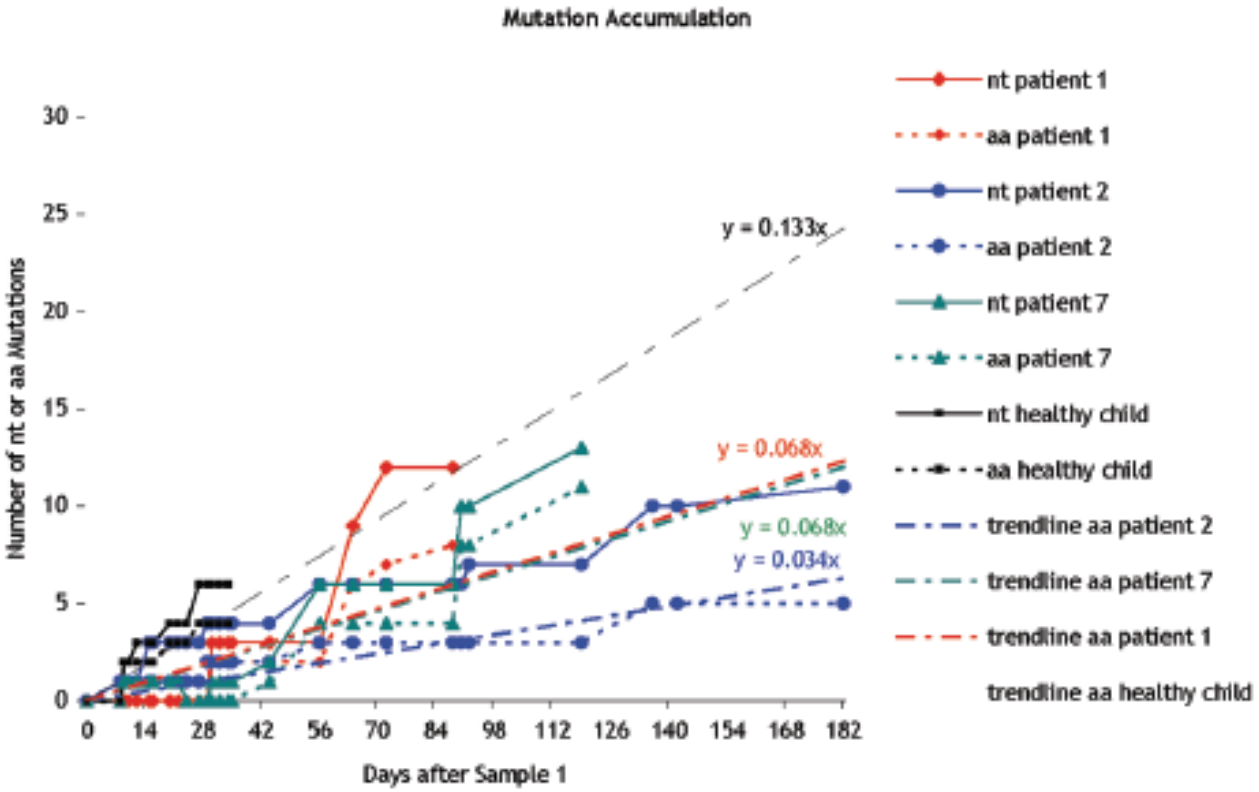




Slope (rate) = 0.98 mutations per generation

Tempo pattern of evolution
- observation from serial samples from patients

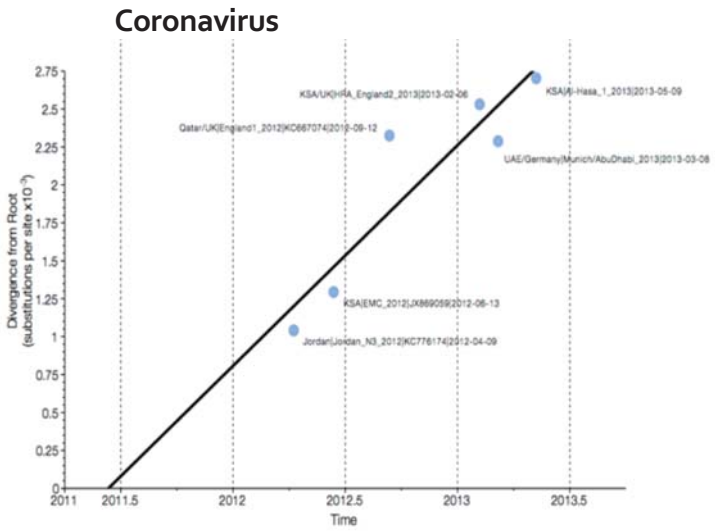
Mutations Accumulated since the 1st Day Sample of Norovirus GII.4

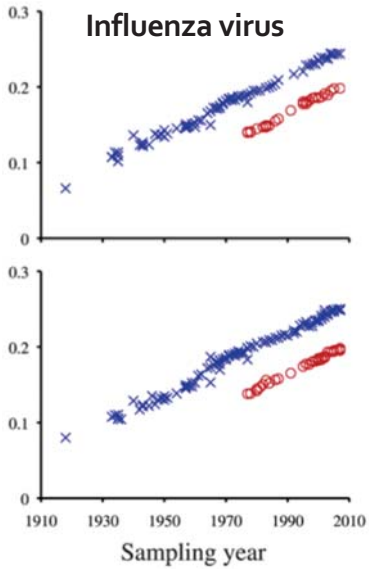


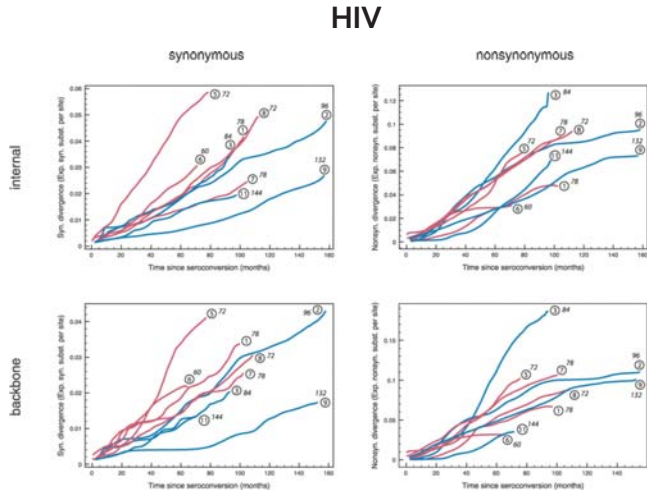
J. Siebenga. et al. (2008) *J. Infect. Dis*

Molecular clock

- **A simple molecular clock assumes mutations are accumulated in the pathogen genome in a constant rate.**
  - observed in many fast-evolving viruses, e.g. human and swine influenza, HIV, rabies, PRRSV, EV71, HCV, RSV, etc…..
  - **Substitution (*b*) = clock rate (*μ*) x Time (*t*)**
  - ***b* = *μt***
  - the expected distance between sequences increases linearly with their time of divergence
  - *(e.g. μ is in unit of substitution/year)*

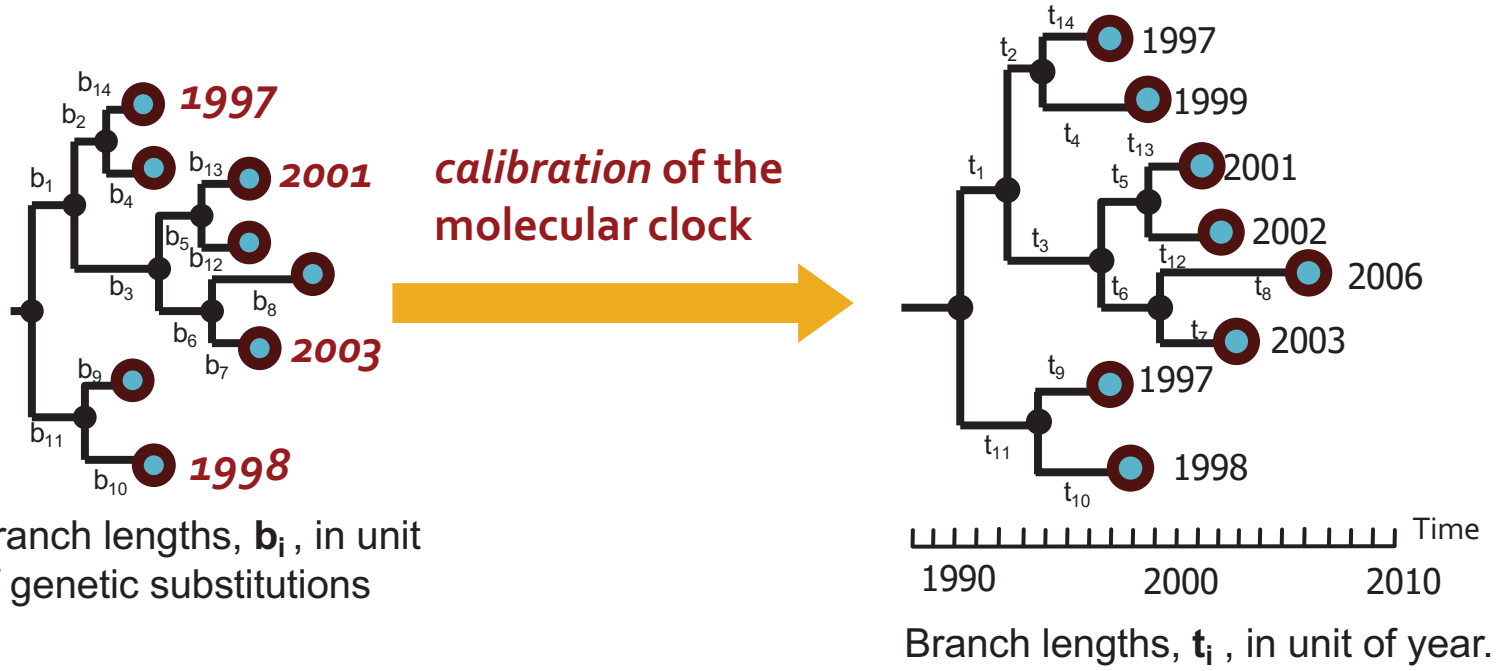






Molecular clock estimation of time-scaled tree

- **Molecular clock assumption** provides a simple yet powerful way of transforming a distance tree to time tree.
- If **external information** about the **ages of one or more nodes** on the phylogeny is available, **sequence distances or branch lengths** can be converted into **absolute calendar times**, and the clock rate

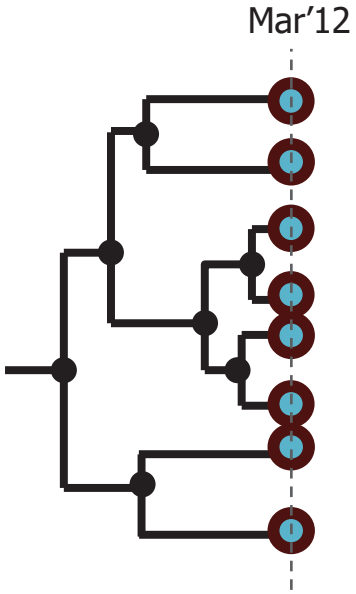


Branch lengths, $\mathbf{b_i}$, in unit of genetic substitutions

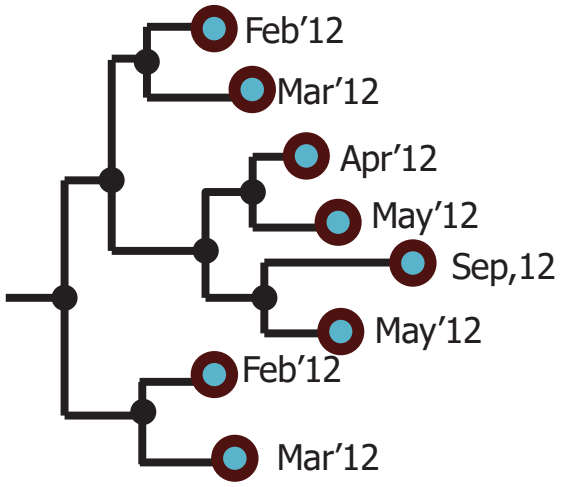
Branch lengths, $\mathbf{t_i}$, in unit of year.

Sampling strategy to enable molecular clock dating

- **Contemporary sampling**
  - Samples collected at a **single** time point
  - No information for calibration
  - Explicit assumption on clock rate



- **Heterochronous sampling**
  - Samples are collected at **different time points**
  - Useful to **calibrate** molecular clock to estimate the time-scale of phylogeny



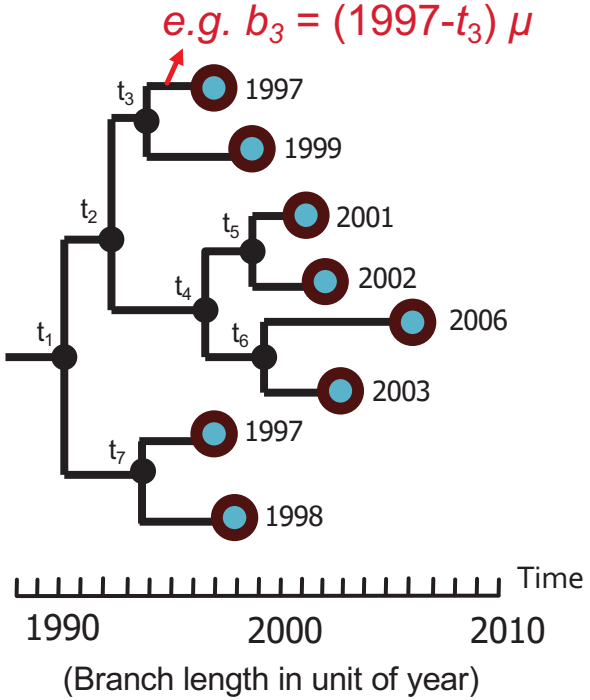
Molecular Clock Dating: Maximum likelihood

- In the clock model, s parameters: $t_{1\text{-}7}$ and μ

  Likelihood can be calculated because $b = \mu t$

  $L = P(D|\{b_{1\ldots 2s-2}, \lambda\})$

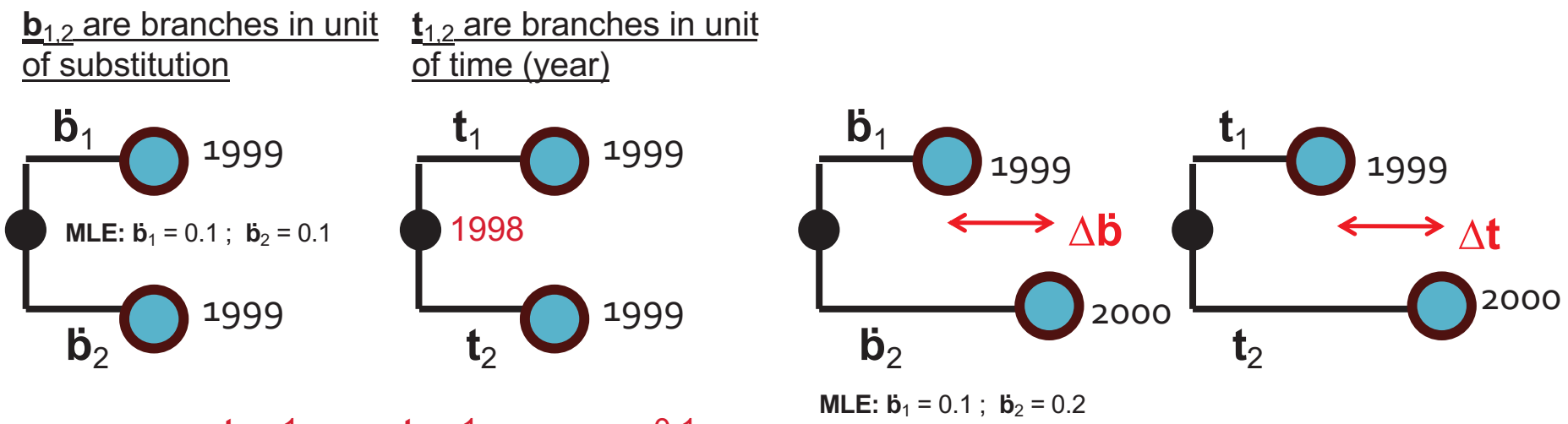
  $L = P(D|\{t_{1\ldots s-1}, \mu, \lambda\})$ *with clock model*



- The time parameters have to satisfy the **constraints** that any node should not be younger than any of its child nodes, e.g. t3 < min(1997, 1999)
- Numerical optimization of the likelihood function to be performed under such **constraints**
- Achieved by **constrained optimization** or through **variable transformations** (Yang and Yoder 2003).

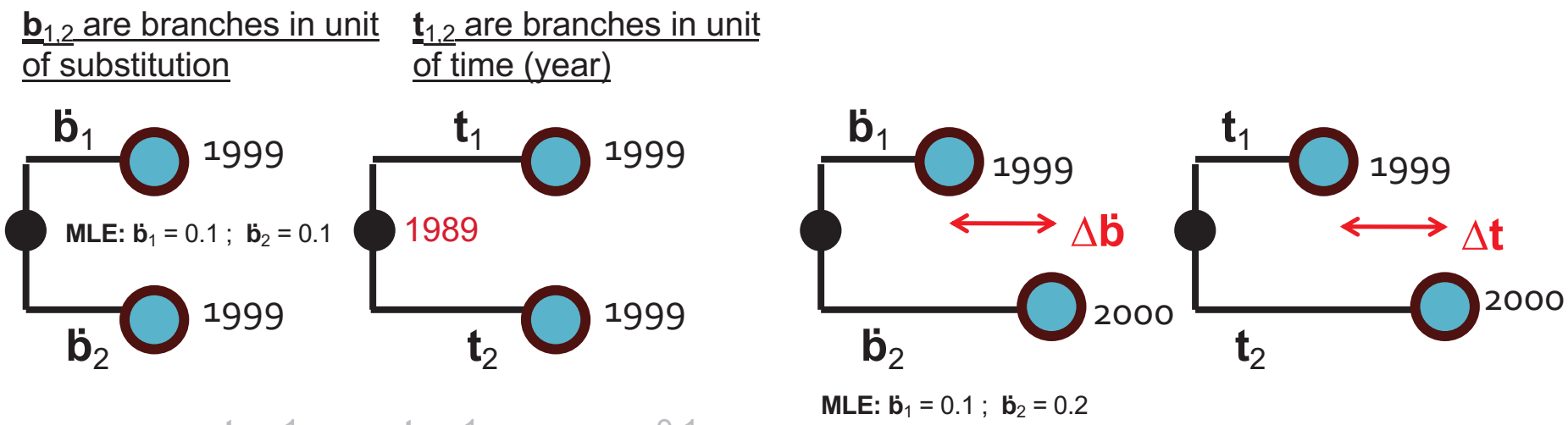
Molecular Clock Dating: Dated-Tips

- **Heterochronous** samples provide '**dated-tips**' with measurable time and genetic difference to estimate rates and timescale of the phylogeny



Molecular Clock Dating: Dated-Tips

- **Heterochronous** samples provide **'dated-tips'** with measurable time and genetic difference to estimate rates and timescale of the phylogeny



t_1 = 1 year ; t_2 = 1 year ; μ = 0.1

t_1 = 10 years ; t_2 = 10 years ; μ = 0.01

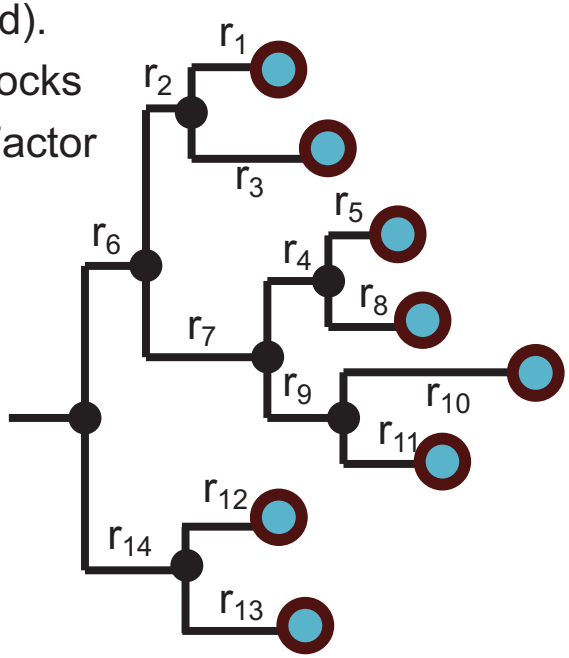
b_1 = 0.1 ; b_2 = 0.1

Give the same likelihood, hard to estimate the rate and timescale

Heterochronous samples give Δb & Δt, allowing calibration and estimation of clock rate: $\mu \approx \Delta b / \Delta t$

Different Clocks: 'Relaxed Clock'

- **More complicated clocks - 'Relaxed clock'** (Huelsenbeck et al. Genetics. 154 (2000); Kishino et al. MBE. 18. (2001); Drummond et al. PLoS Biol (5) (2006); Rannala et al. Syst. Biol. 56. (2007))
  - allow some rate variations in different lineages, which is more biologically realistic
  - *auto-correlated* and *uncorrelated* rate variation
  - variation follows some distributions, e.g. exponential, lognormal
  - In ML framework, clocks can be tested using likelihood ratio tests (if models are nested).
  - In BMCMC framework, clocks can be tested by Bayes Factor tests.



$r_{1\text{-}14} \sim \text{LogNormal}(\mu, \sigma^2)$

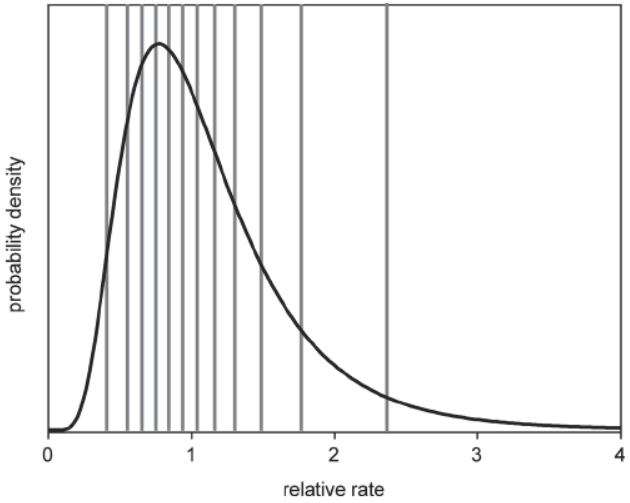


Figure 5. A Lognormal Distribution Discretized into 12 Rate Categories Each of the 12 categories has equal probability ($p = 1/12$). The i^{th} rate category (numbered from left to right) corresponds to the $(i - 0.5)/12$ quantile of the lognormal distribution.
DOI: 10.1371/journal.pbio.0040088.g005

Libyan HIV & HCV outbreak in a children hospital

- El-Fatih Children's Hospital, Libya.
- >400 children were tested positive for HIV and/or HCV in late 1998 – early 1999.
- Six foreign (1 Palestinian & 5 Bulgarian) medics, who started their work there in March 1998, were accused.
- In 2005/06, the children's HIV & HCV were sequenced and analyzed.

de Oliveira *et al.* (2006) ***Nature*** 444, 836-837

HCV sequences from the hospital formed **three separate clusters**: two from *genotype 4* and one from *genotype 1*.

de Oliveira *et al.* (2006) *Nature* 444, 836-837

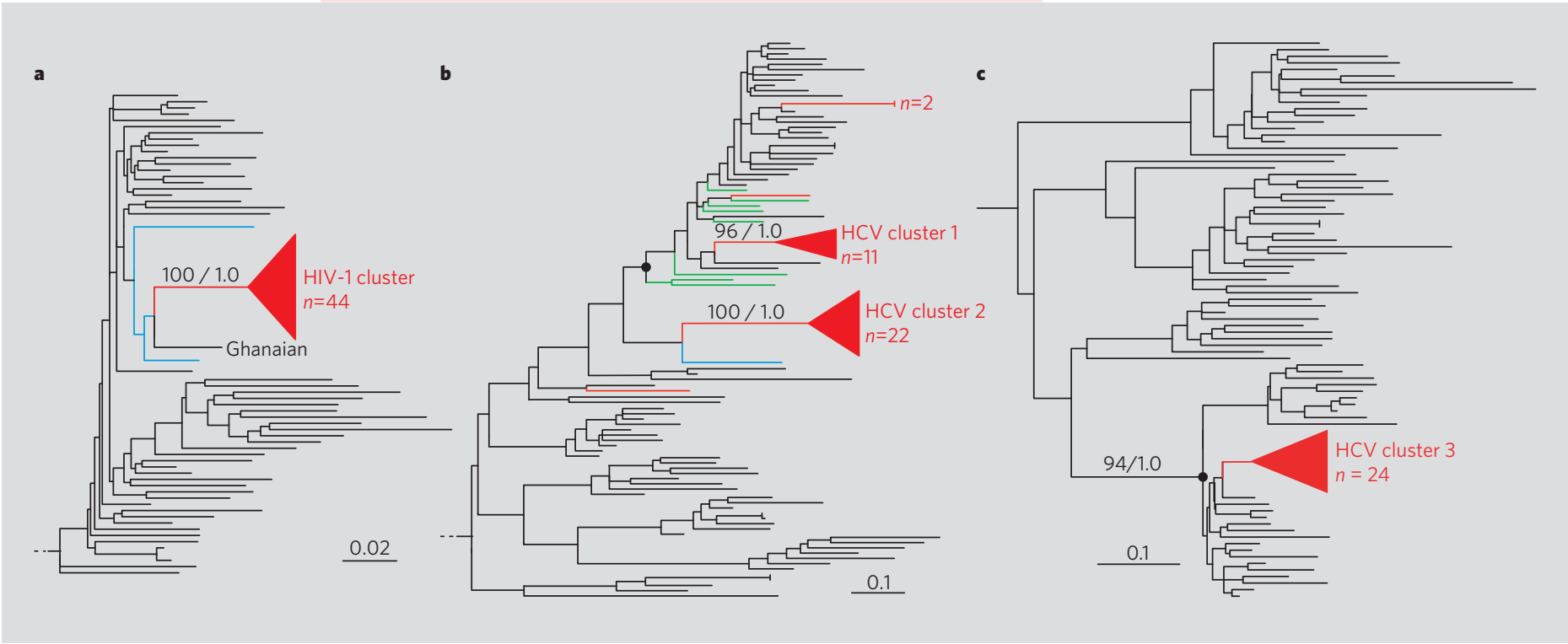


Figure 1 | HIV-1 and HCV sequences from 1998 Al-Fateh Hospital (AFH) outbreak. a–c, Estimated maximum-likelihood phylogenies for HIV-1 CRF02_AG (**a**), HCV genotype 4 (**b**) and HCV genotype 1 (**c**). Source of sequences used for analysis: AFH, red; Egypt, green; Cameroon, blue. Black circles mark the common ancestor of HCV subtype 4a and 1a; numbers above AFH lineages give clade support values using bootstrap and bayesian methods, respectively. Scale bar units are nucleotide substitutions per site. For visual clarity, AFH clusters are represented by triangles and some non-informative reference strains are excluded.

Libyan HIV & HCV outbreak in a children hospital

de Oliveira *et al.* (2006) *Nature* 444, 836-837

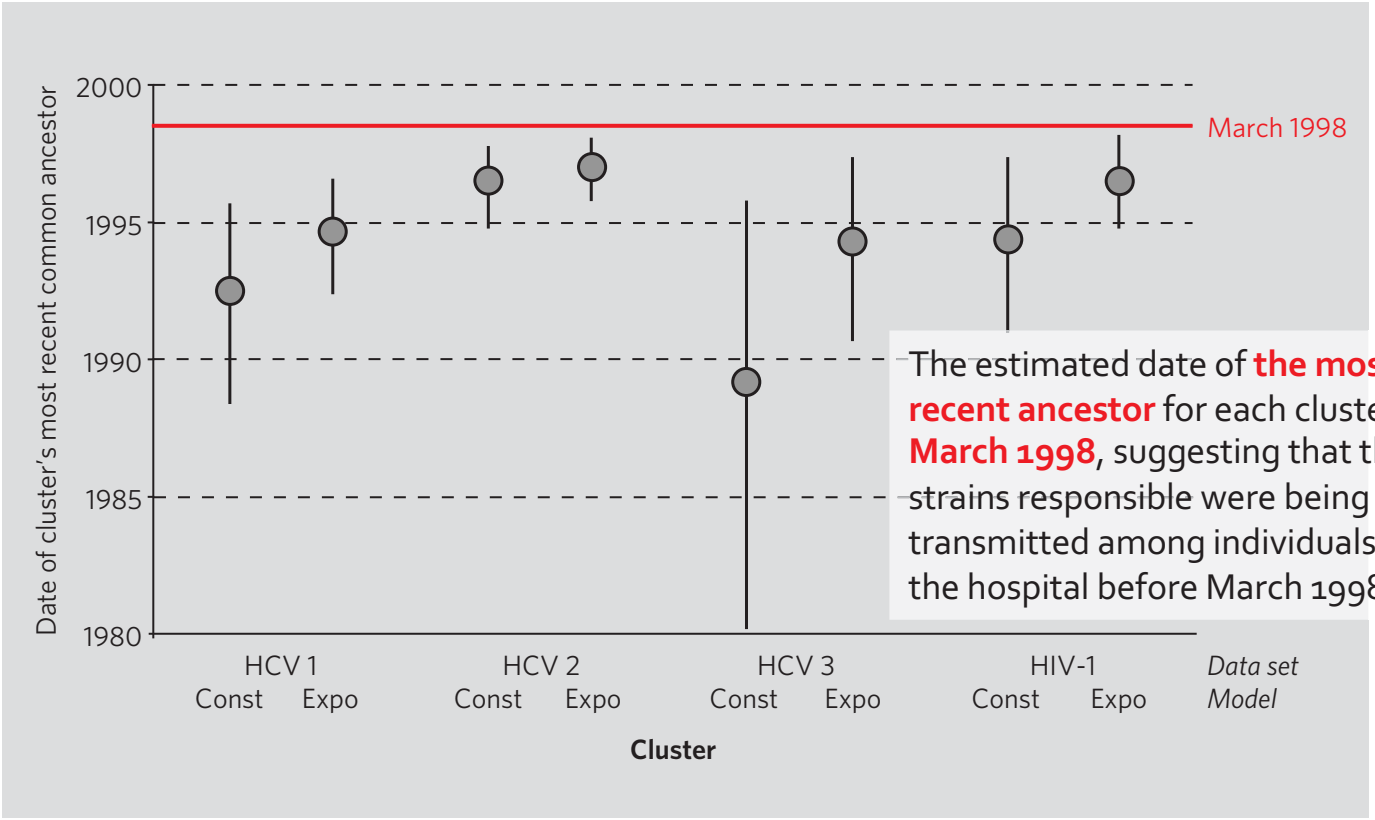


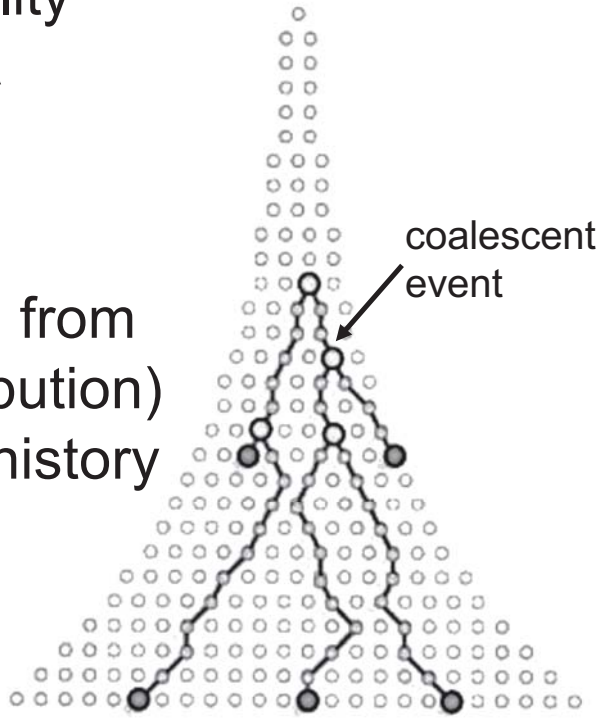
Figure 2 | **Estimated dates of the most recent common ancestor for each cluster.** Results obtained by using different evolutionary models. Vertical lines show the 95% highest posterior density intervals. Red line shows time of arrival of the foreign staff in March 1998. For further details, see supplementary information. 'Const', constant size; 'Expo', exponential growth.

Quantitative inferences on phylogenetic trees – **Epidemic growth**

Coalescent inference: genealogy-based method in Population Genetics

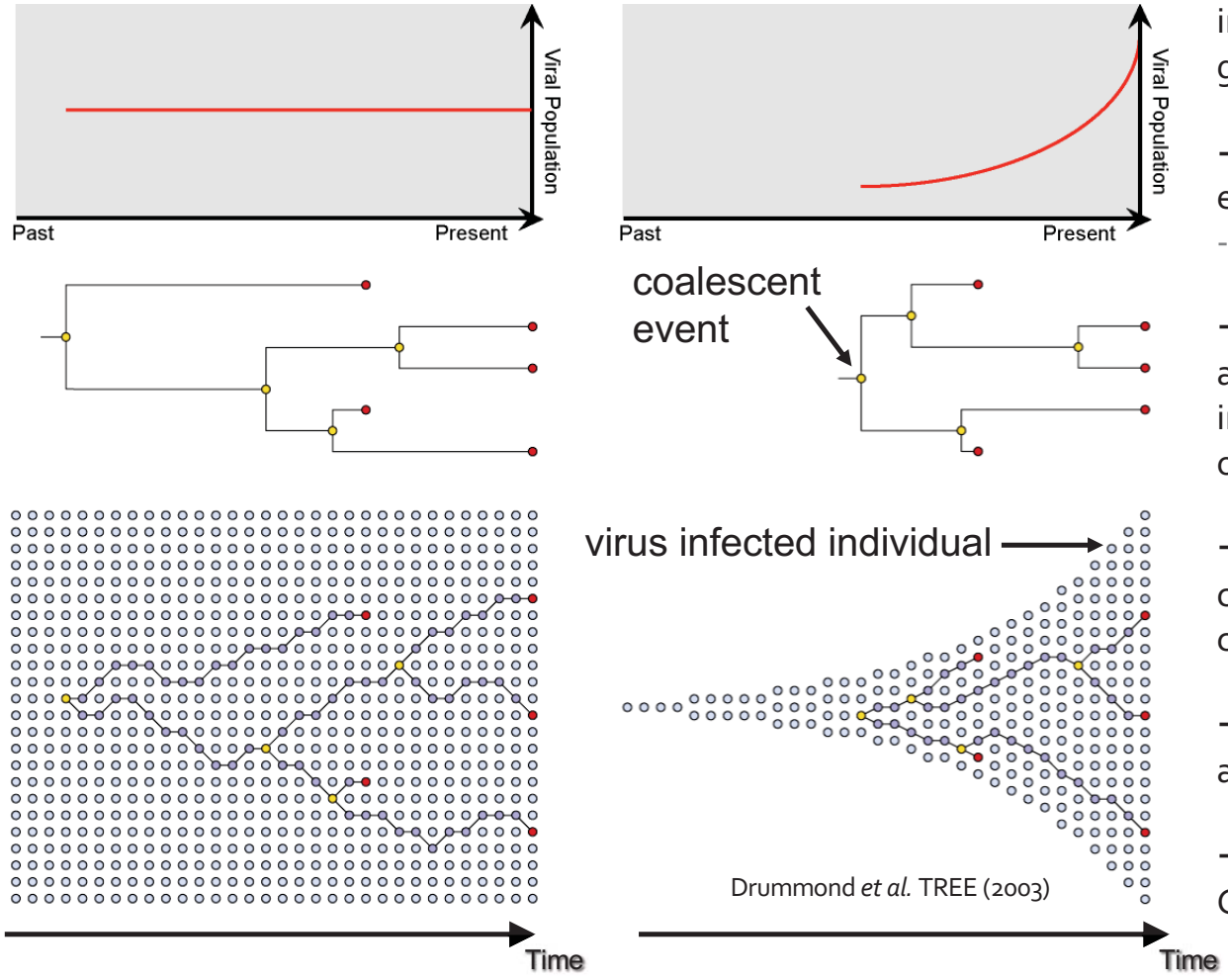
- In a genealogy(tree), if you trace the ancestry of some individuals, you will always find a common ancestor of them. A coalescent event is where two lineages merge as a common ancestor.
- The coalescent model describes the probability distribution on the coalescent events given a population history.

  Kingman (1982); Griffiths and Tavare (1994)
- Therefore the model can convert information from ancestral relationships (i.e. coalescent distribution) into information about the actual population history and vice versa.



Population Inference using Coalescent Theory

- Coalescent theory [Kingman (1982); Griffiths and Tavare (1994)]



In a smaller population,

→ Higher probability for two individuals to coalesce in the previous generation

→Age of common ancestor is expected to be younger.

→ Distribution of coalescent nodes and their time depth can be used to infer *effective* population size (N_e) over time.

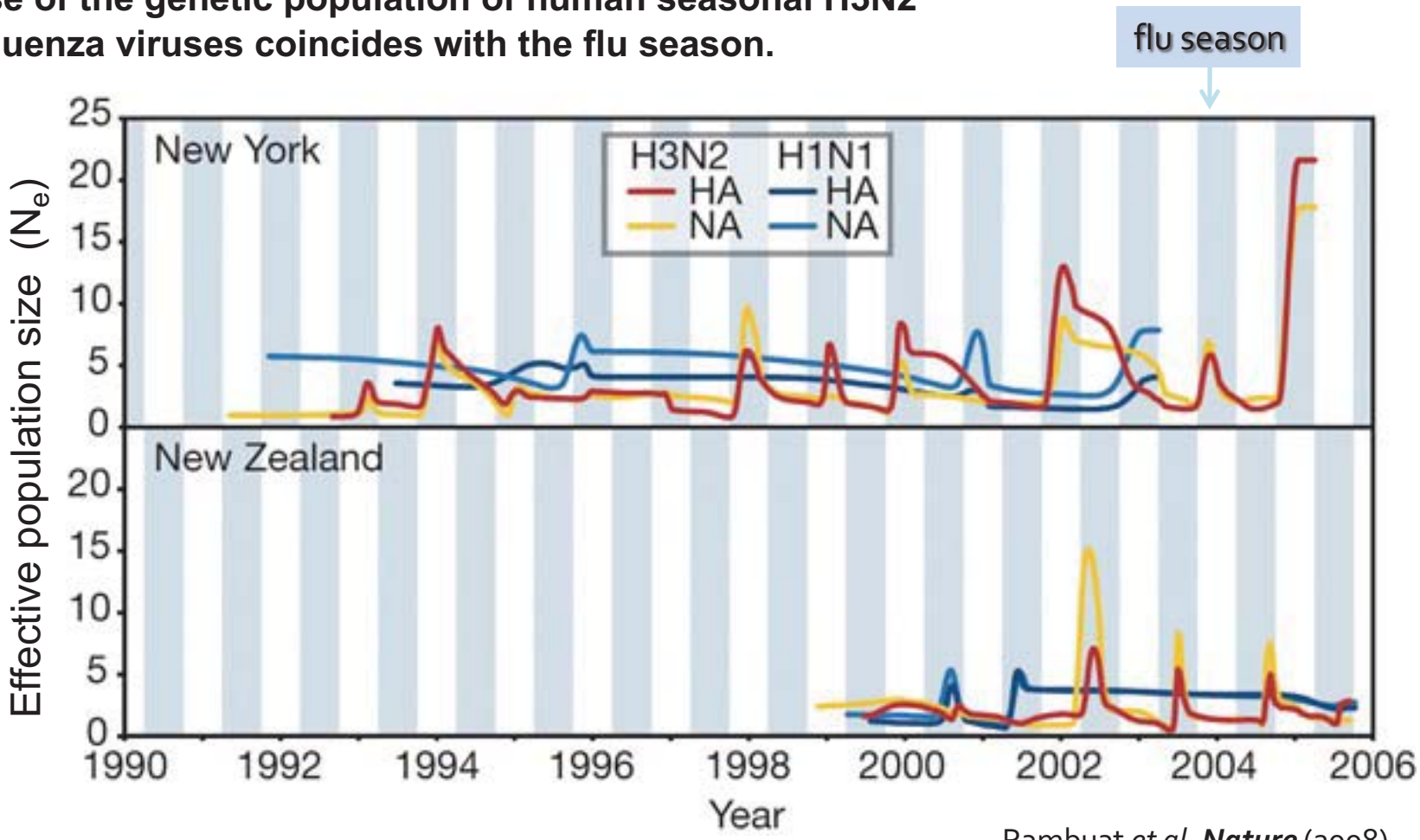
→N_e is an abstract parameter, but changes in N_e reflect changes in the census infected population size

→The time scale was estimated by assuming a molecular clock

→ Details can be found in Pybus *et al.* Genetics (2000)

Population fluctuation of human influenza virus

- **Rise of the genetic population of human seasonal H3N2 influenza viruses coincides with the flu season.**

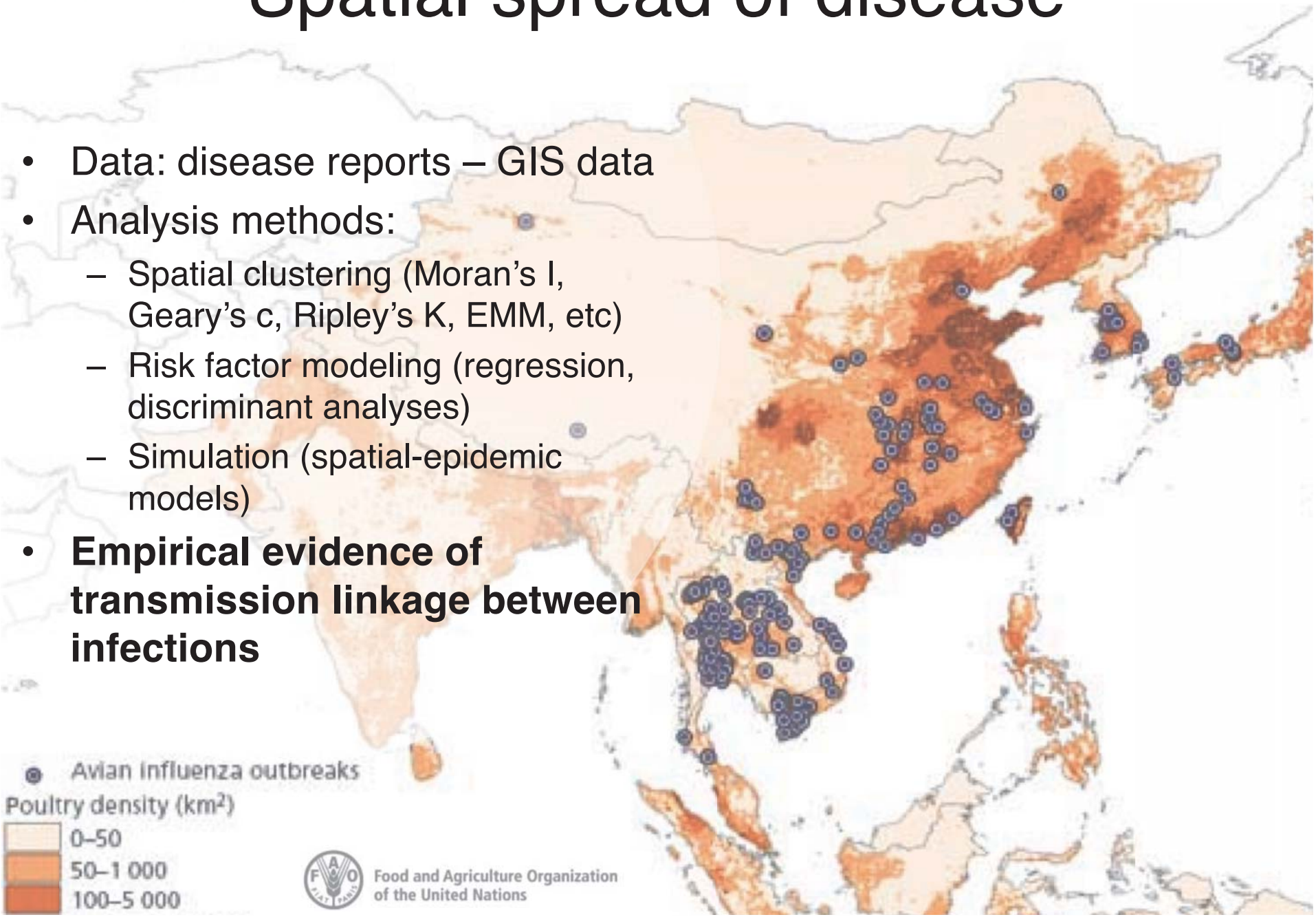


Rambuat *et al.* ***Nature*** (2008)

Quantitative inferences on phylogenetic trees – **Spatial diffusion**

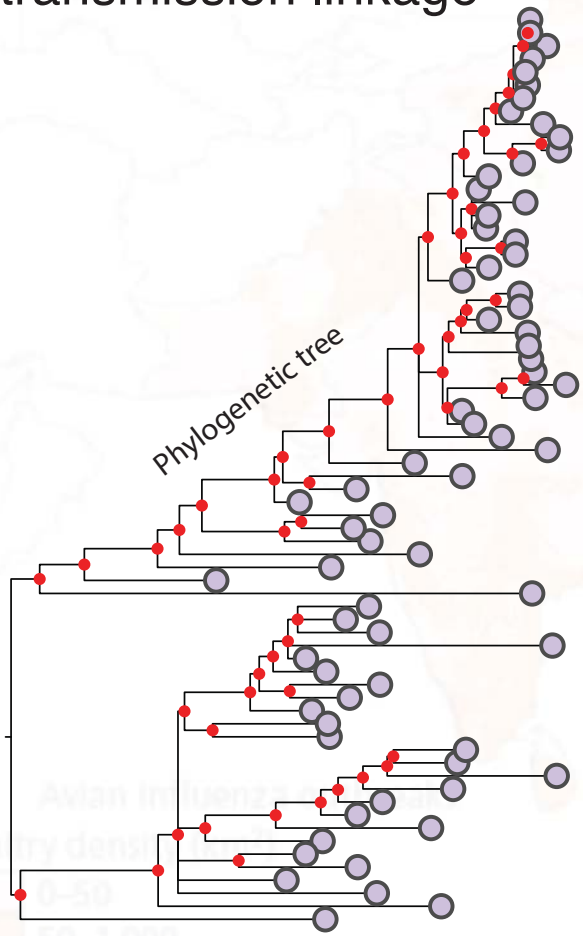
Spatial spread of disease

- Data: disease reports – GIS 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**

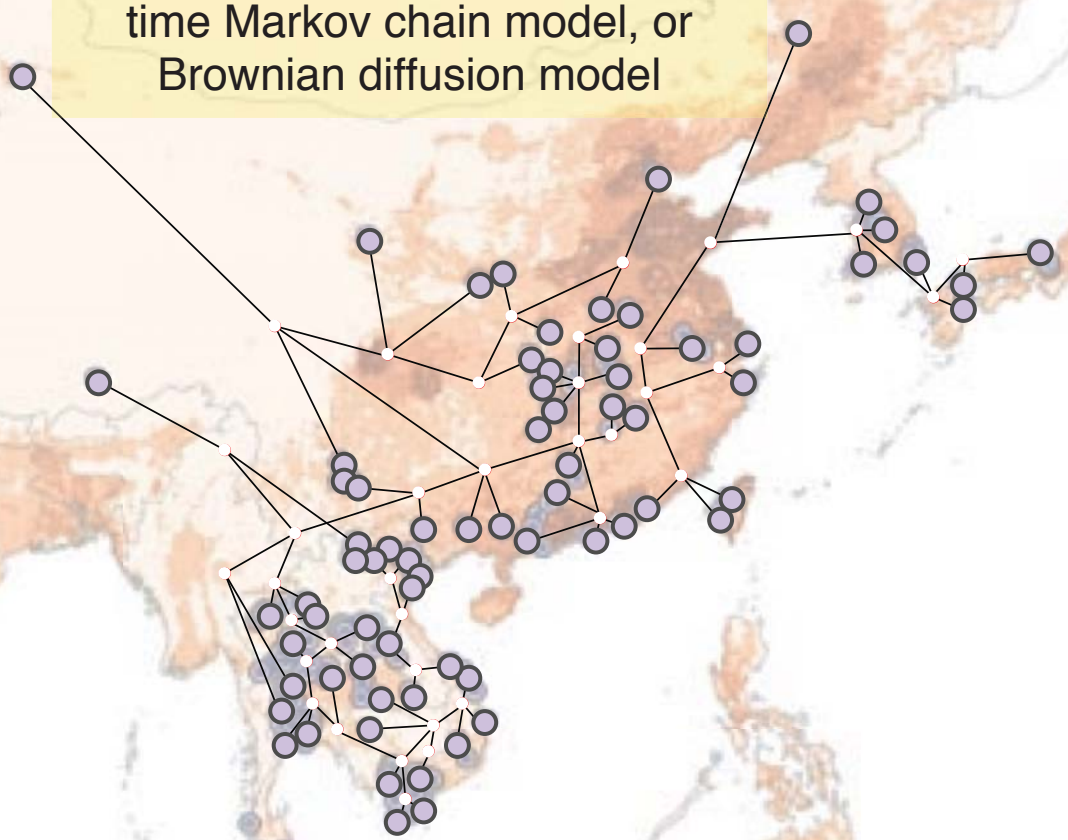


Phylogeographic inference

Phylogenetic tree provides empirical evidence of transmission linkage



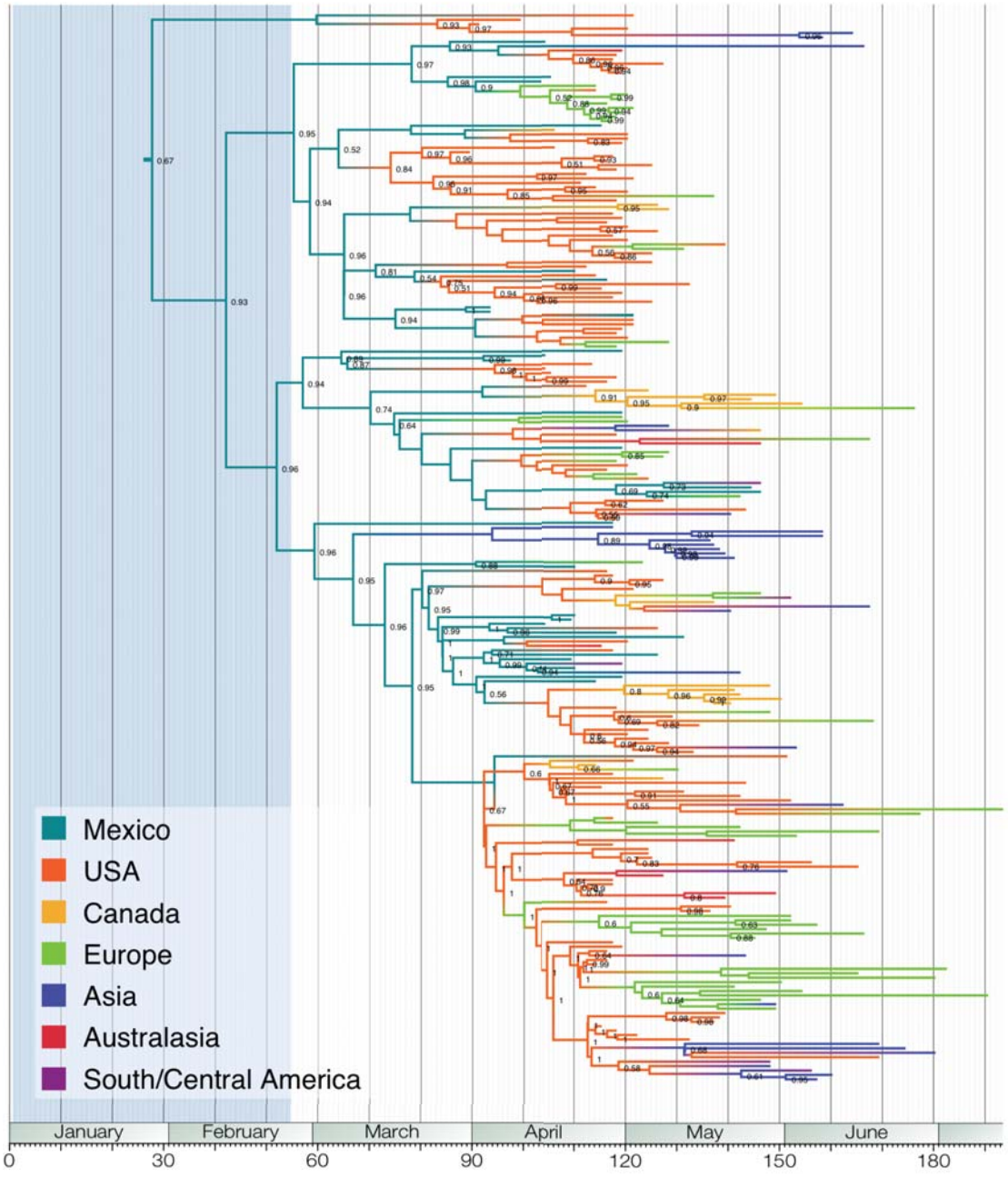
Ancestral geographic locations are estimated with continuous-time Markov chain model, or Brownian diffusion model



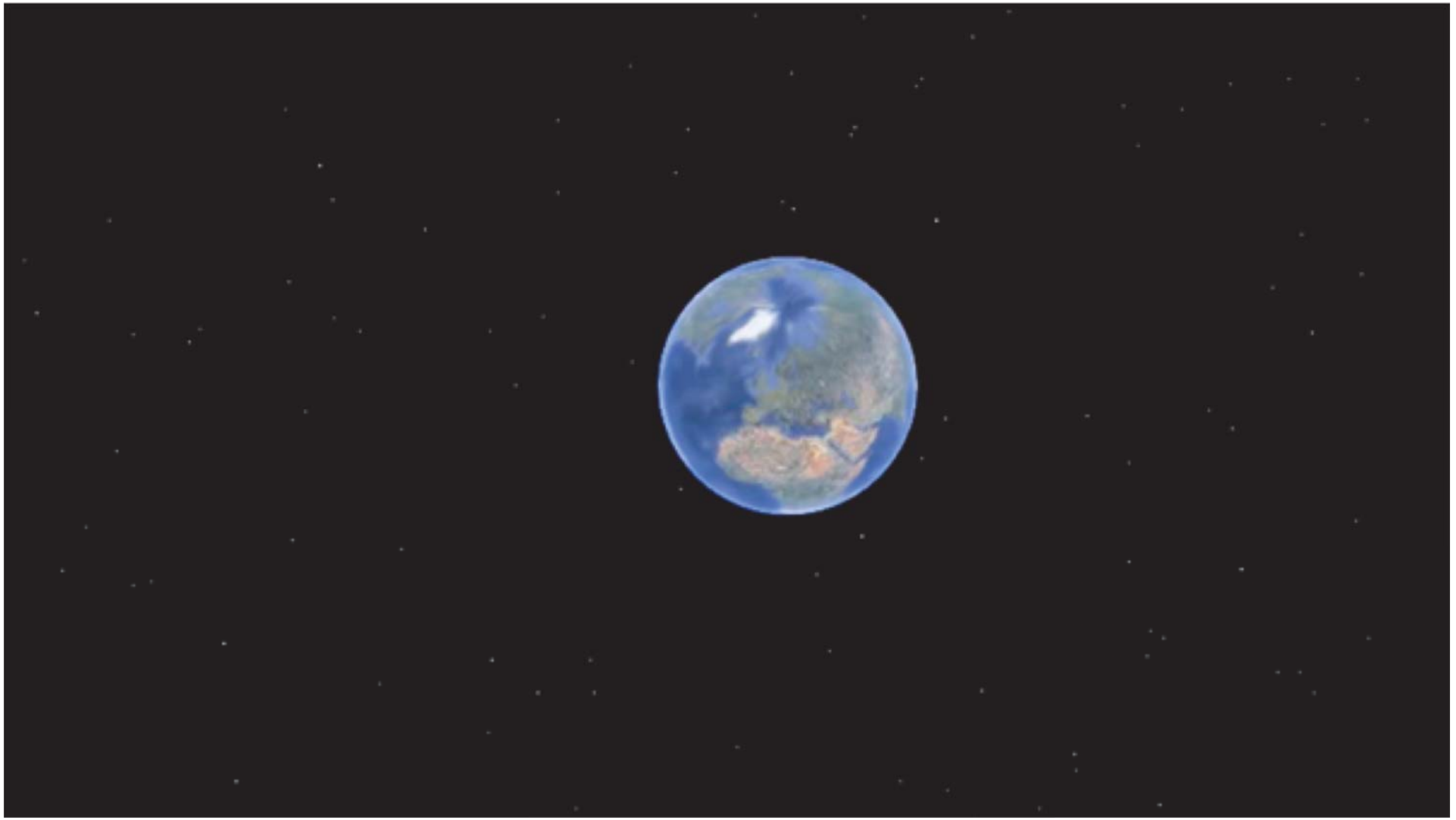
Lemmon et al. Syst. Biol (2008)57
Lemey et al. MBE (2010)27:8

Pandemic (H1N1) 2009 influenza

- Lemey et al. PLoS Currents Influenza. (2009) doi: 10.1371/currents.RRN1031
- Phylogenetic tree of 242 human pdmH1N1 sequences (40 locations; Mar-Jul 2009)
- Relaxed molecular clock model
- Spatial diffusion is modeled as discrete continuous-time Markov chain



Pandemic (H1N1)
2009 influenza



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

Summary

- Some **epidemic parameters** (timescale, infected population, spatial diffusion rate) can be **co-estimated** with phylogenetic tree in **statistical frameworks**
- Require **external information** about the samples, e.g. time and spatial locations of the samples, to **calibrate** the analyses
- Extend the study insights in both **temporal** and **spatial** scales.
- Mostly applicable to **fast-evolving pathogens** such as RNA viruses.