Evolutionary genomics and molecular epidemiology of human influenza A virus*

Andrew Rambaut
Institute of Evolutionary Biology
University of Edinburgh

*and other viruses (if I have time)

the ideal phylogenetic tree

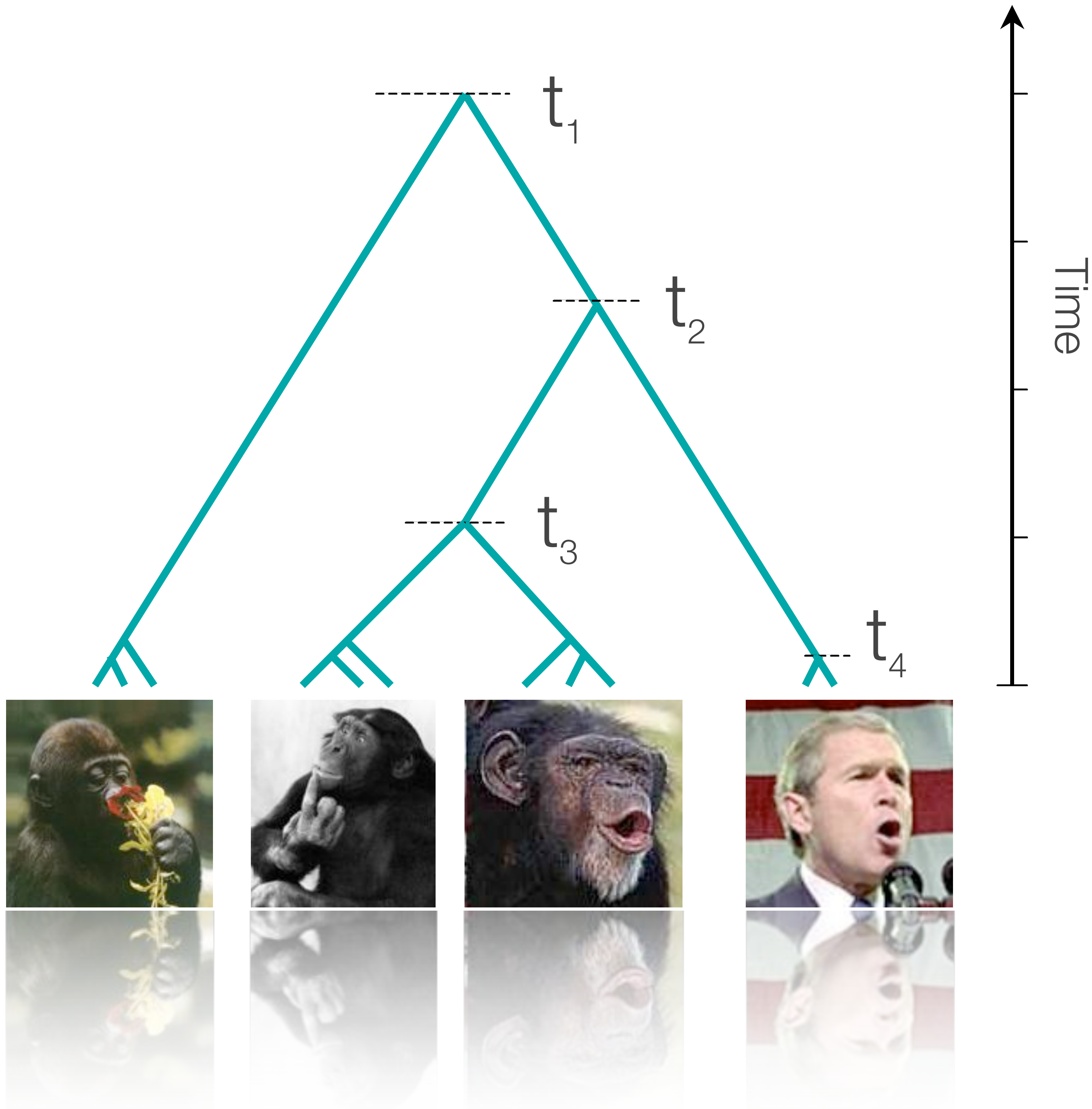
- is rooted

    implies a branching order

- has branch lengths in units of time

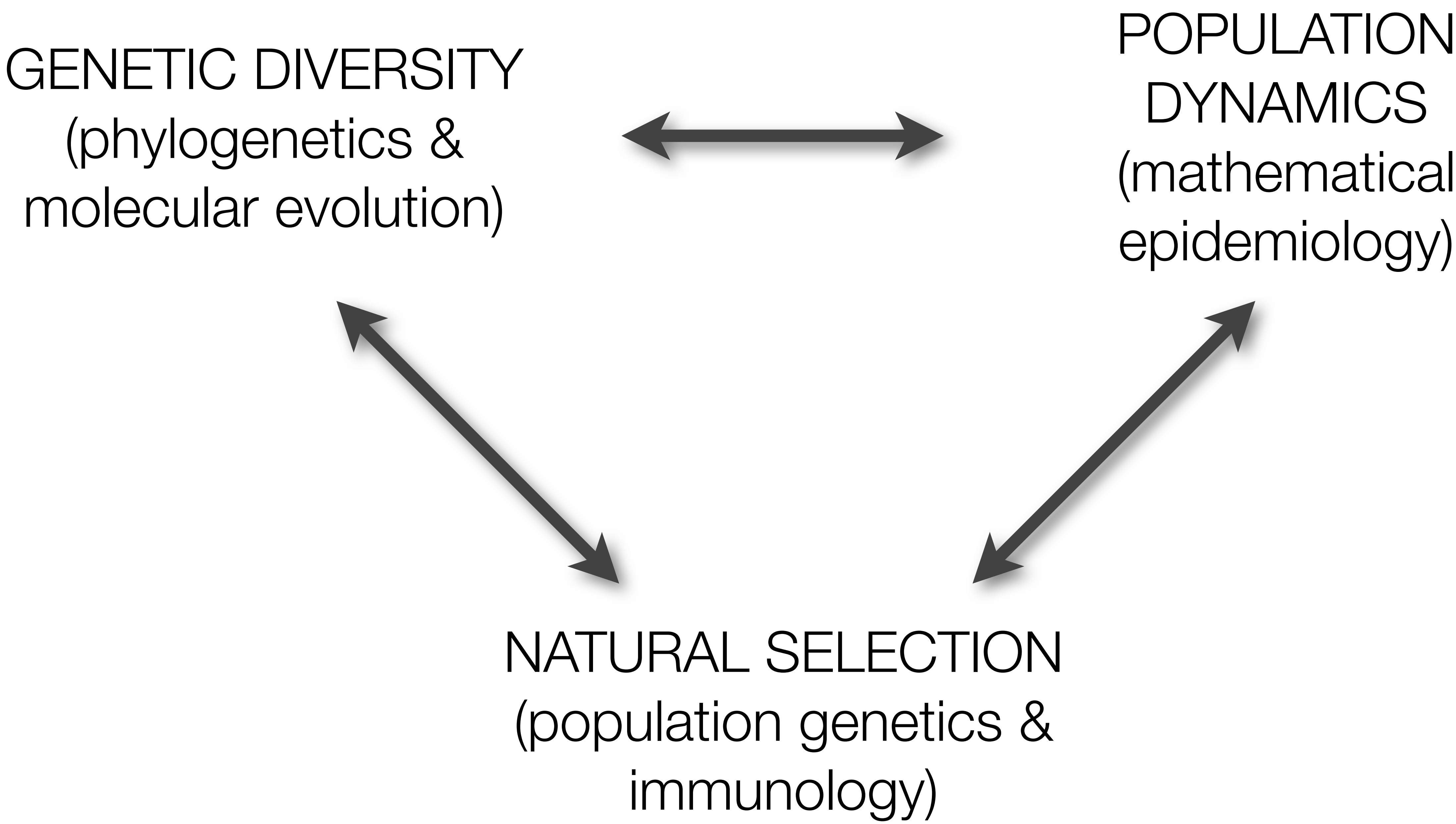
    an evolutionary history

- how do we construct one of these?
    - estimate tree
    - root tree
    - calibrate tree
    - estimate dates



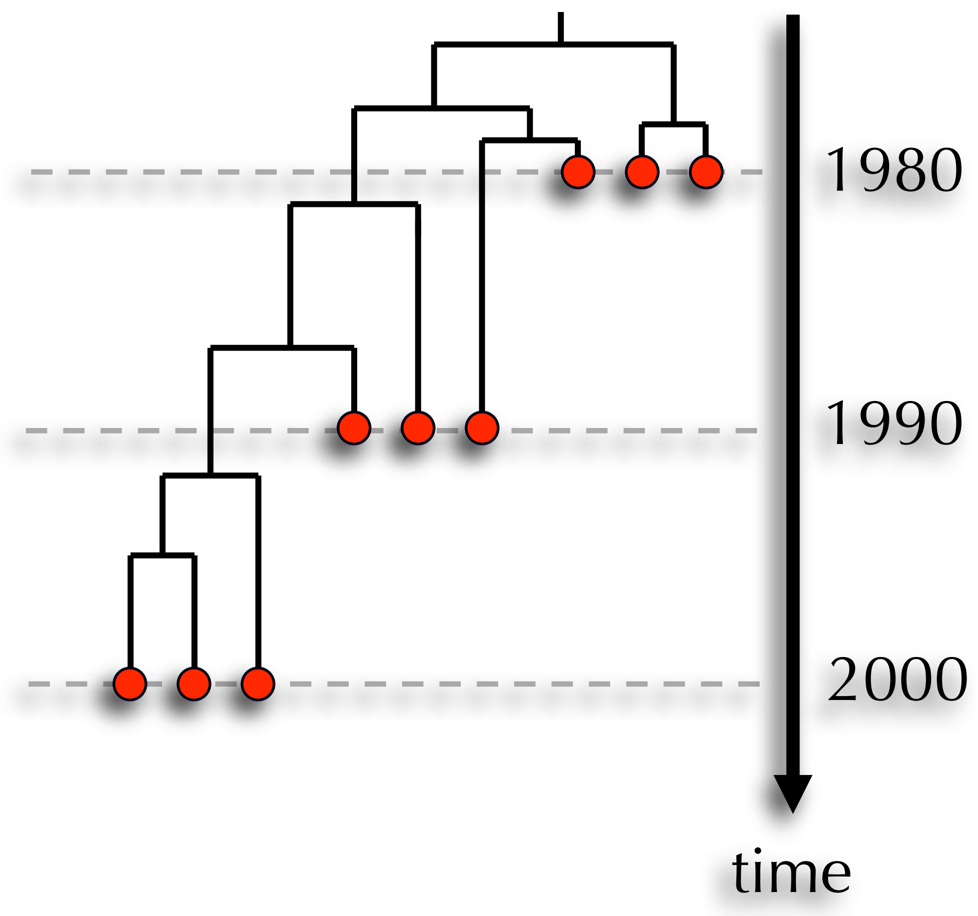
Phylodynamics

Modelling virus evolution

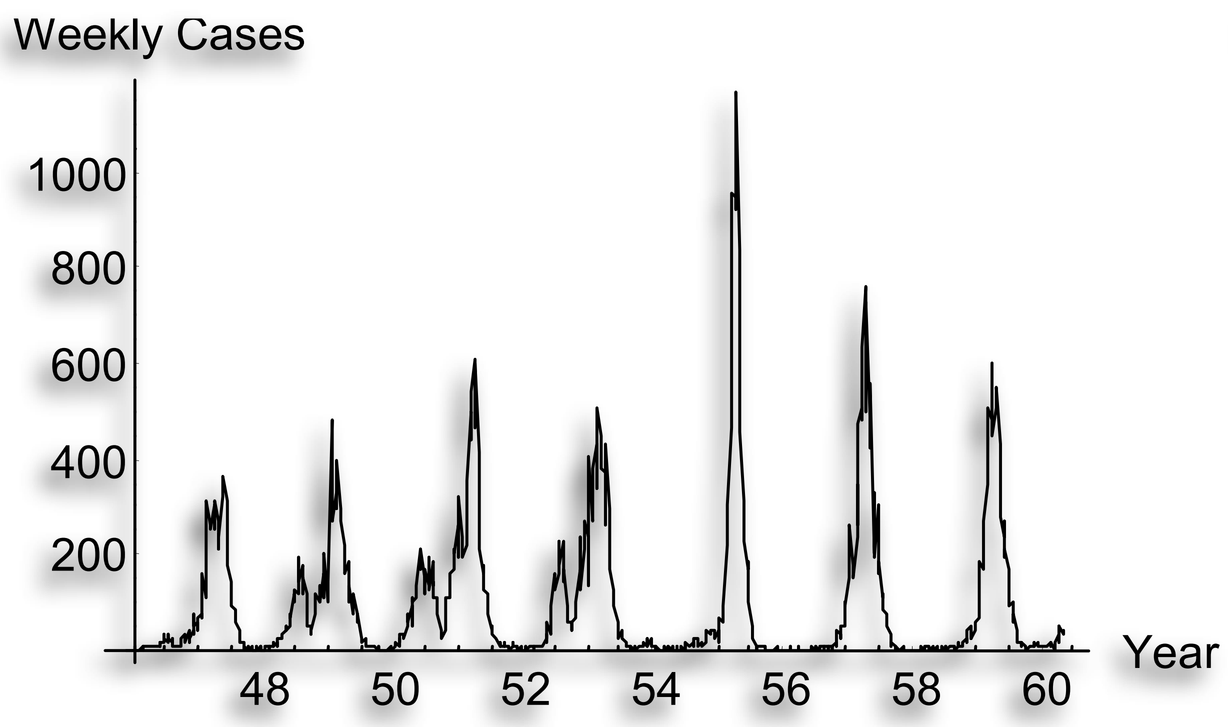


‘Phylodynamic’ Data

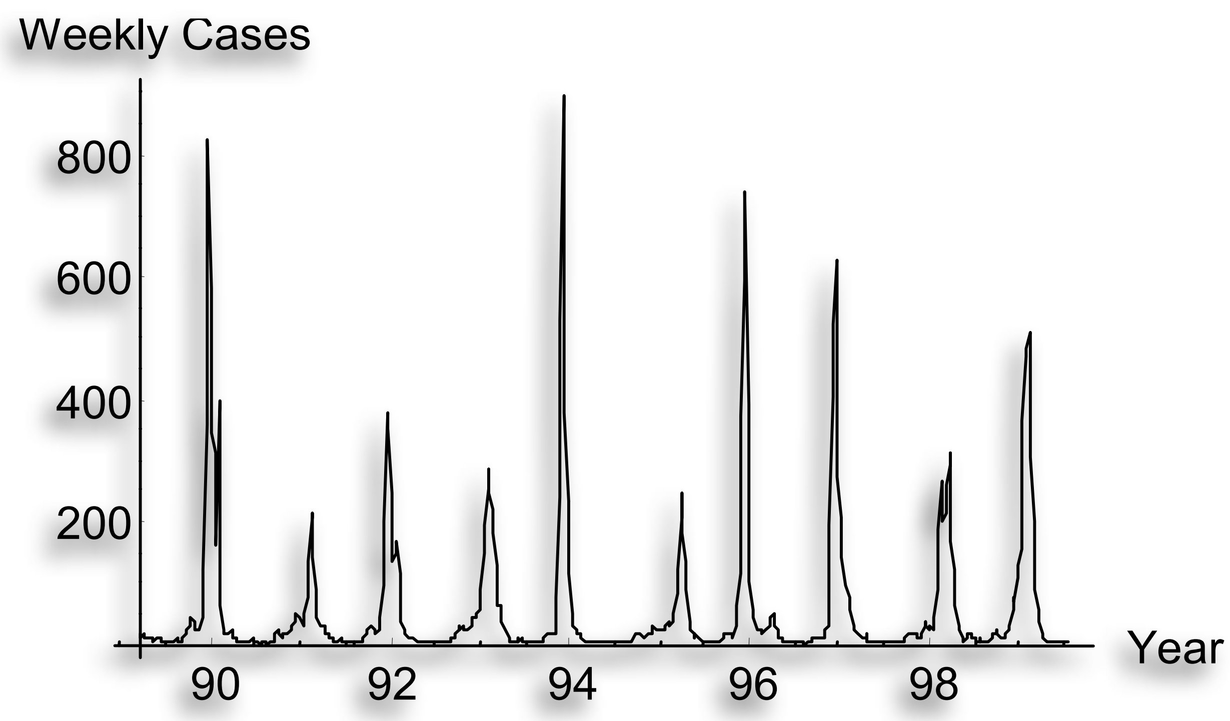
- Virus gene sequences sampled at different points in time.
- Immunological, epidemiological or medical data is often also available.
- Phylogenetic history is estimated on a real time-scale (e.g. years).



Measles & Influenza Dynamics

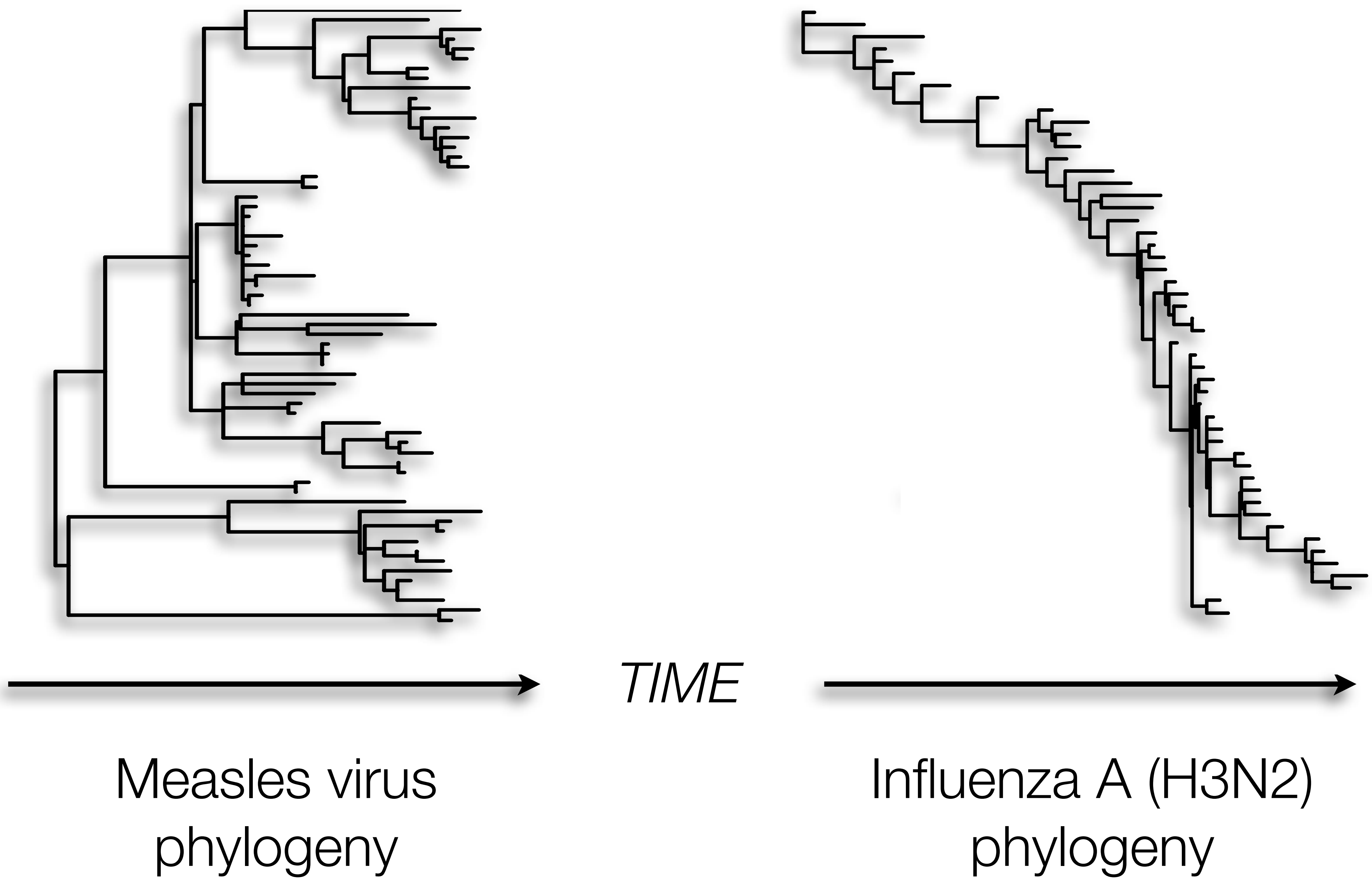


Measles, pre-vaccination
(1940 to 1960, UK)



Influenza-like illness
(1989-1999, France)

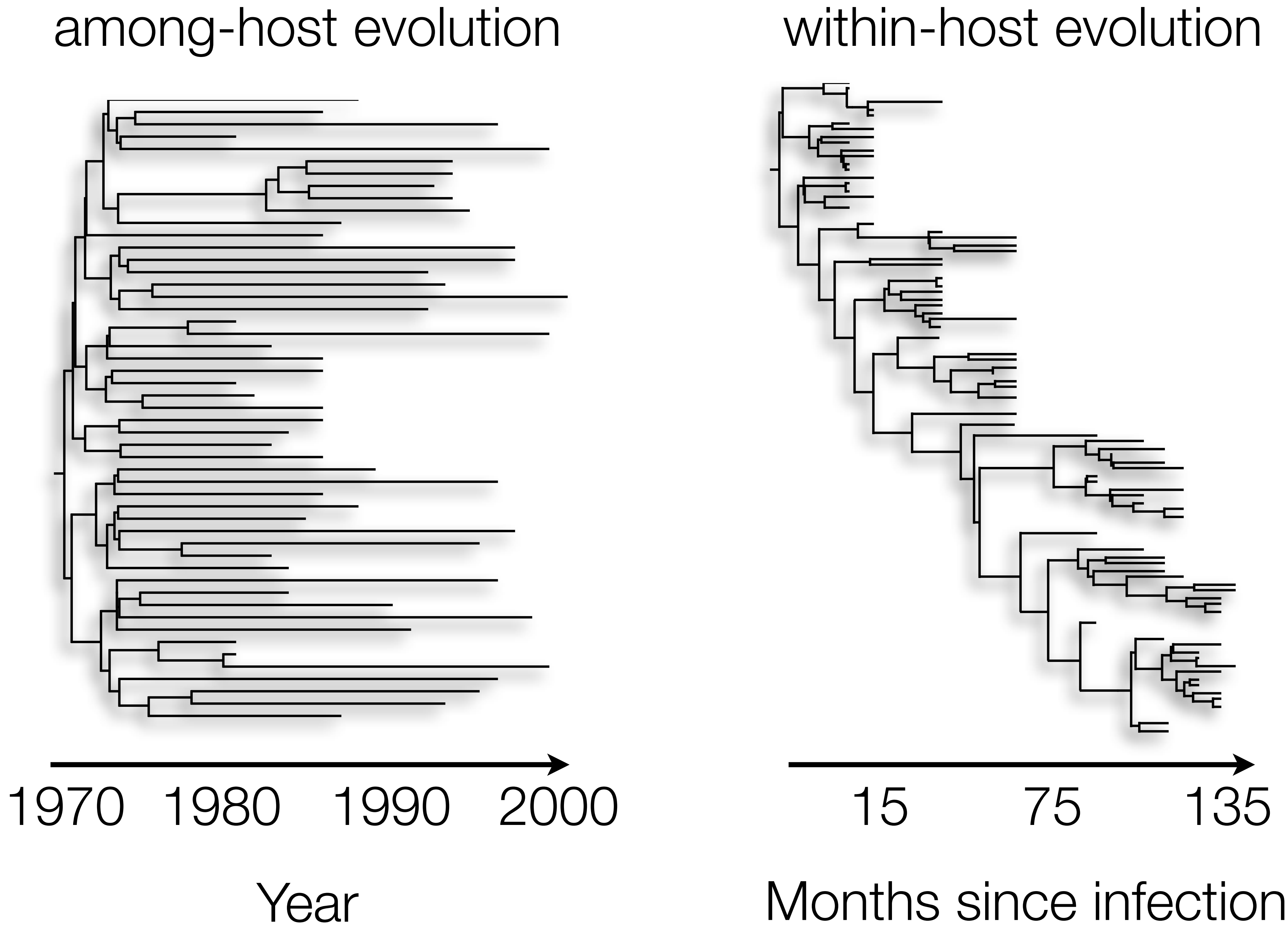
Measles & Influenza Evolution



Phylodynamic Patterns

	Continual Immune Selection	Weak/No Immune Selection	
Idealised Phylogeny Shapes			
Examples			

HIV-1 Has Two Phylodynamic Patterns



HIV within host evolution (Intra-host)

- Driven by continual and strong immune selection from neutralizing antibodies or cytotoxic T lymphocytes
- Temporal structure (clock-like)
- Mutation-fixation balance

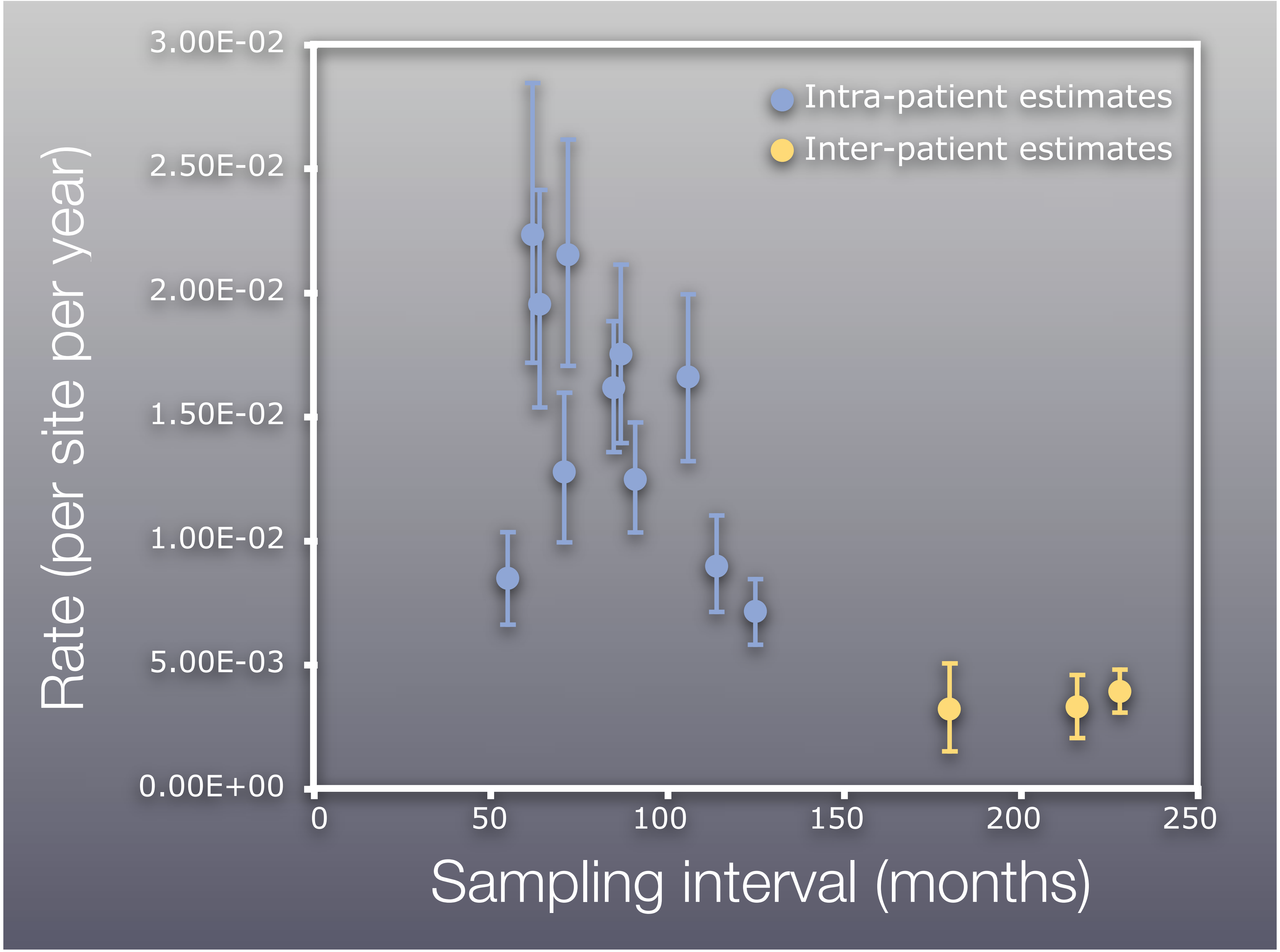
HIV among Host Evolution (Inter-host)

- High level of genetic variation.
- Less-temporal structure (better as older sequences become available)
- Little evidence for continuous selection.
- Neutral spatial and temporal structure of virus (Genetic drift)

HIV among Host Evolution (Inter-host)

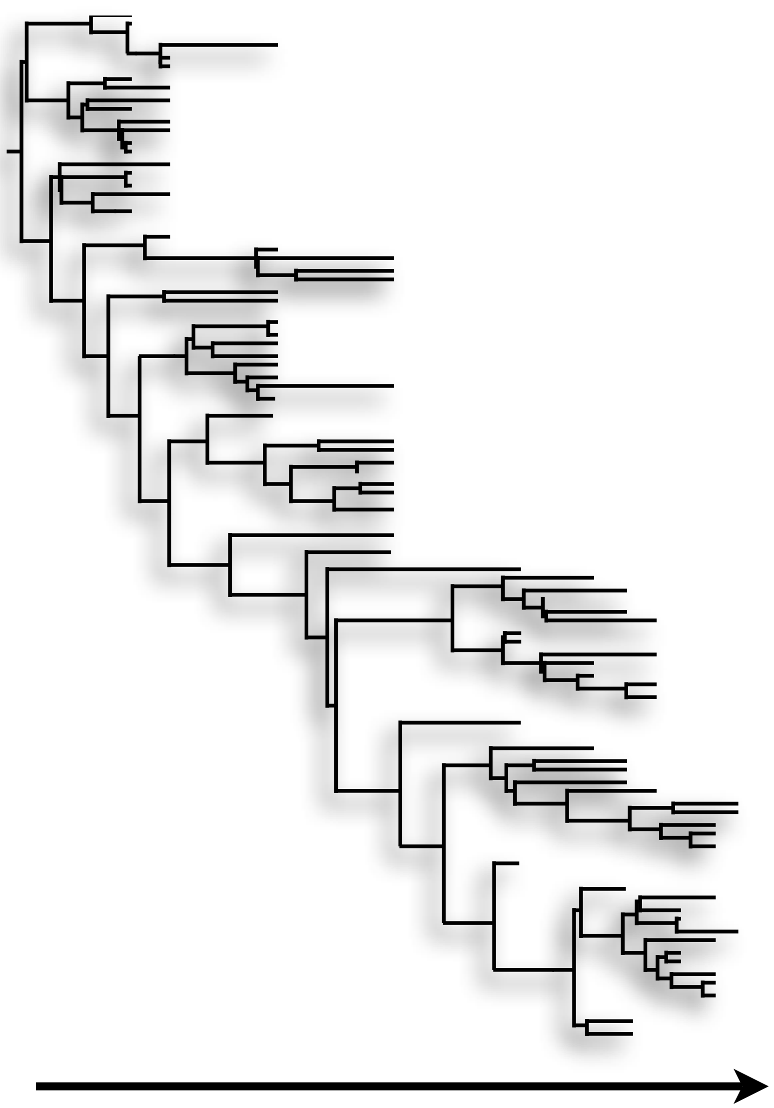
- Inter-host dynamics are slow because time between infections is long
- Little cross-protective immunity - superinfection with multiple strains relatively common
- Although intra-host adaptation to human lymphocyte antigen (HLA) types has been observed...
- ... little evidence for differential transmissibility among HIV-1 strains
- Phylogenetic structure reflects the demographic history of transmission.

Inter- vs. Intra-Patient Evolution

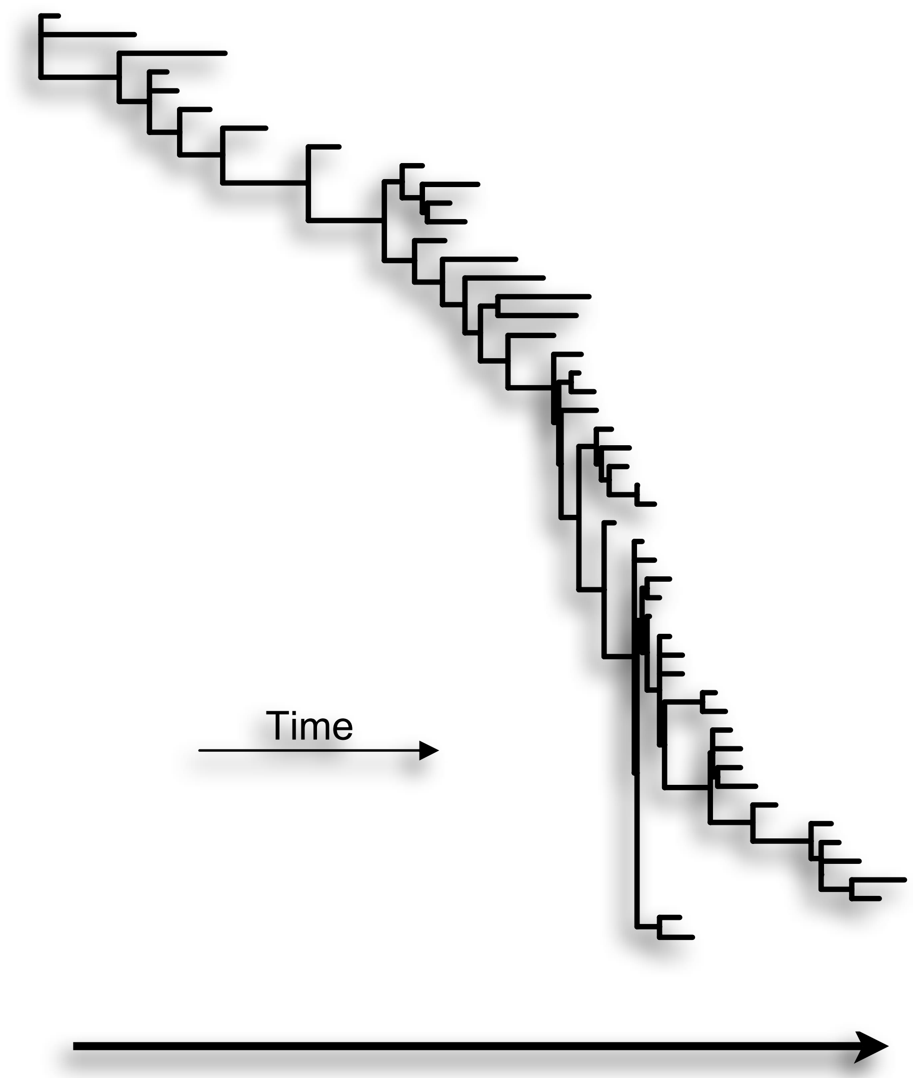


Influenza A

- Looks like HIV within patient evolution?

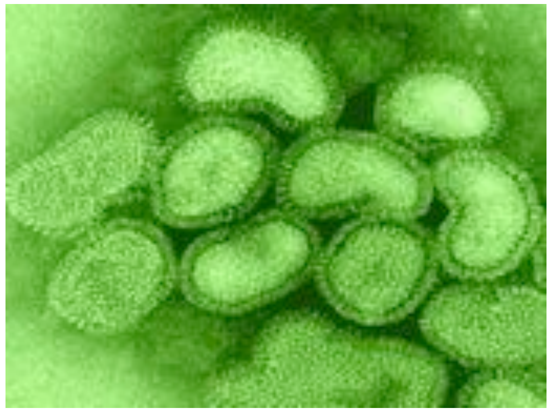


HIV within-patient
phylogeny



Influenza A (H3N2)
phylogeny

evolutionary dynamics of influenza A



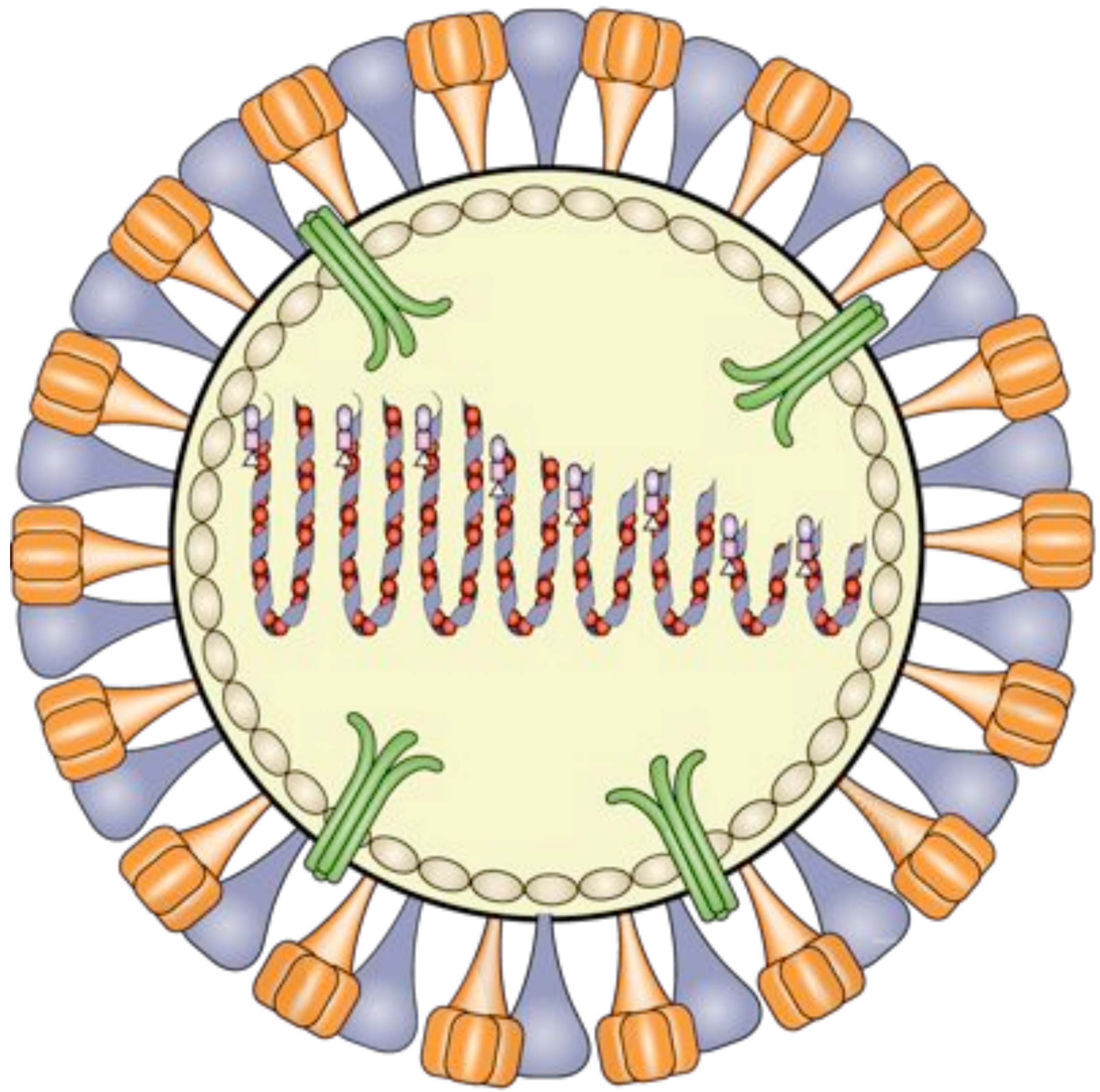
Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK & Holmes EC

influenza A

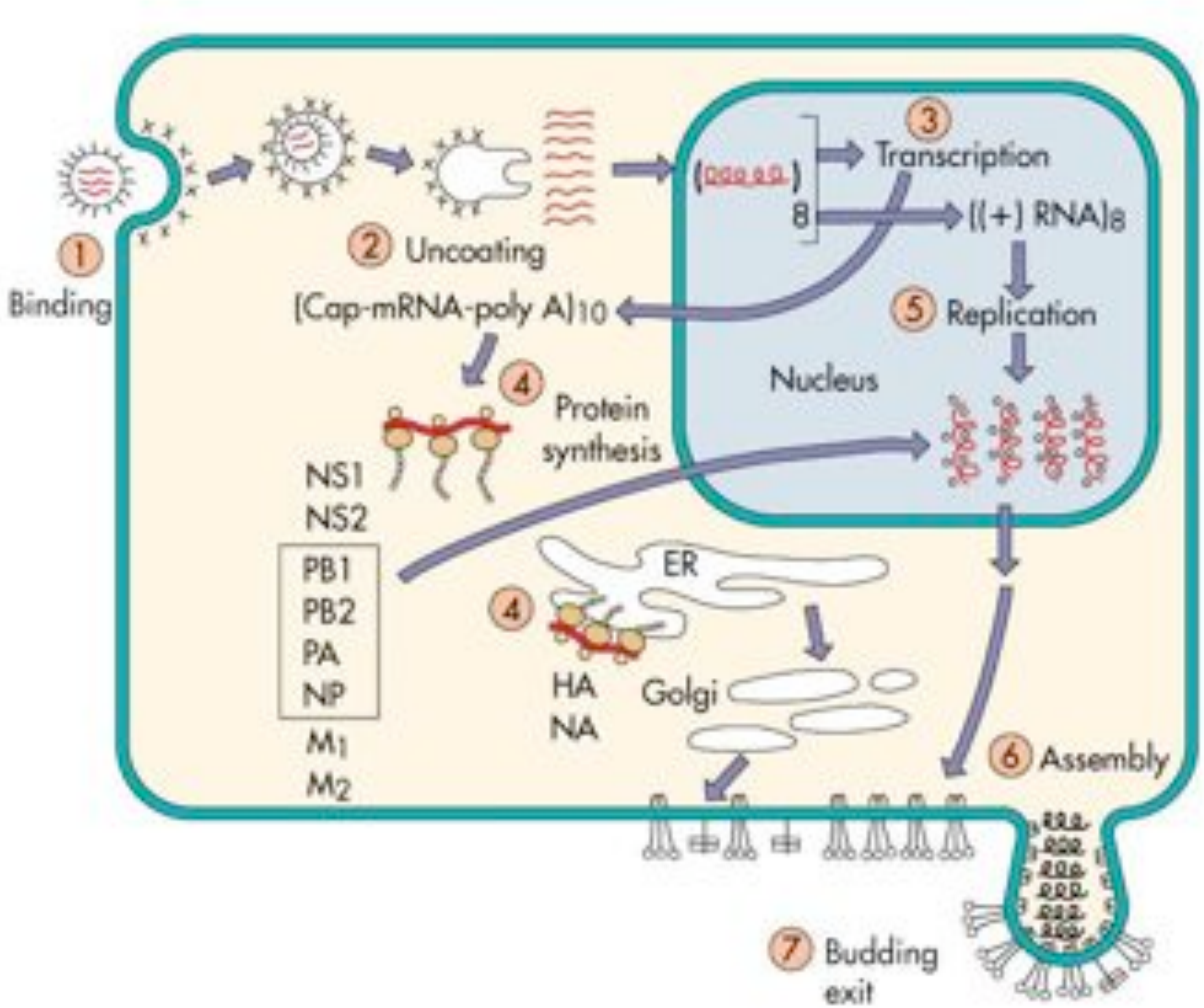
- one of the most important infectious diseases of humans
  - annual mortality of 250,000 to 500,000
- occasional global pandemics infect 20-40% of population
  - 1918-1919 pandemic probably caused 20-50 million deaths
    - the single most devastating disease outbreak in human history
- H3N2 genotype most prevalent but occasionally a season is dominated by H1N1 (the virus that caused the 1918 pandemic).

influenza A virus

- segmented RNA virus
- genome divided into 8 segments with 1 or 2 genes each
- 2 genes, HA & NA, are exposed on the virus and are targeted by immune system (and are thus vaccine targets).



Influenza replication



Major reassortment events in the history of IVA

- Since the global pandemic of 1918 caused by an influenza A subtype H1N1 virus, five genome segments have maintained an unbroken evolutionary history within humans –

  nucleocapsid protein (NP), matrix proteins (M1/2), the nonstructural proteins (NS1/2) and two polymerase proteins (PB2 and PA).

- new haemagglutinin (HA), neuraminidase (NA) and PB1 polymerase, acquired through reassortment with avian viruses.

  Acquisitions coincided with global pandemics:

  - HA subtype H2 and NA subtype N2 appeared in 1957, HA subtype H3 emerged in 1968, a new PB1 segment acquired in both 1957 and 1968.

Seasonal Human Influenza A

- H1N1 viruses re-emerged in 1977 and circulate to this day
- Seasonal epidemics of influenza A virus since 1968 have been dominated by H3N2 subtype viruses
- Characterized by punctuated antigenic evolution
- Reassortment amongst circulating strains also seems to be prevalent

questions

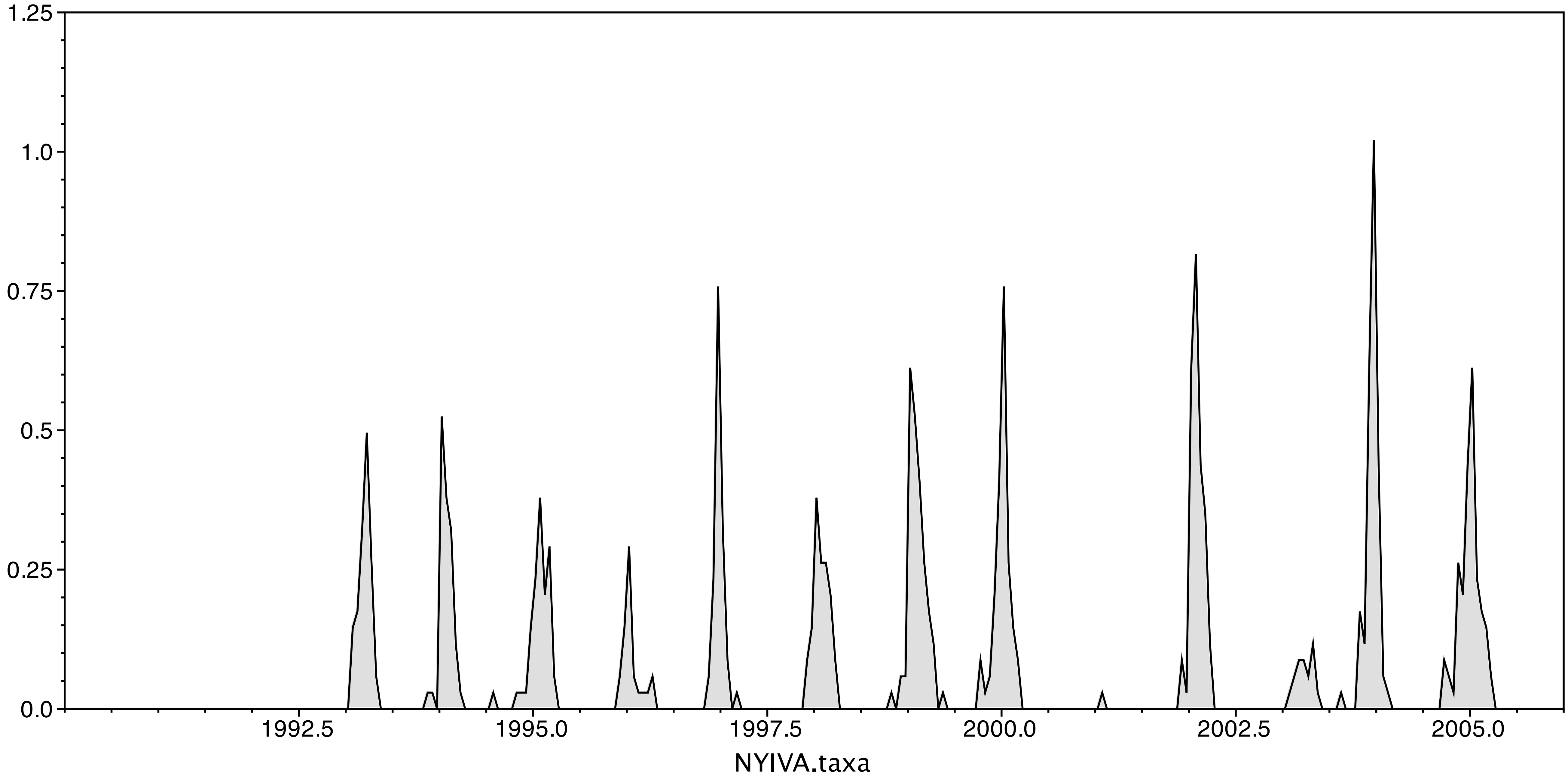
- Which is more important for influenza evolution?
  - mutation (antigenic drift)
  - reassortment (antigenic shift)
- How do different genomic segments interact?
  - Is reassortment random?
  - Are some combinations more likely than others?
- How do genetically distinct genotypes interact?
  - cross-protective immunity
  - antigenic fitness?
- What are the dynamics of influenza globally?

Human Influenza A genomic data

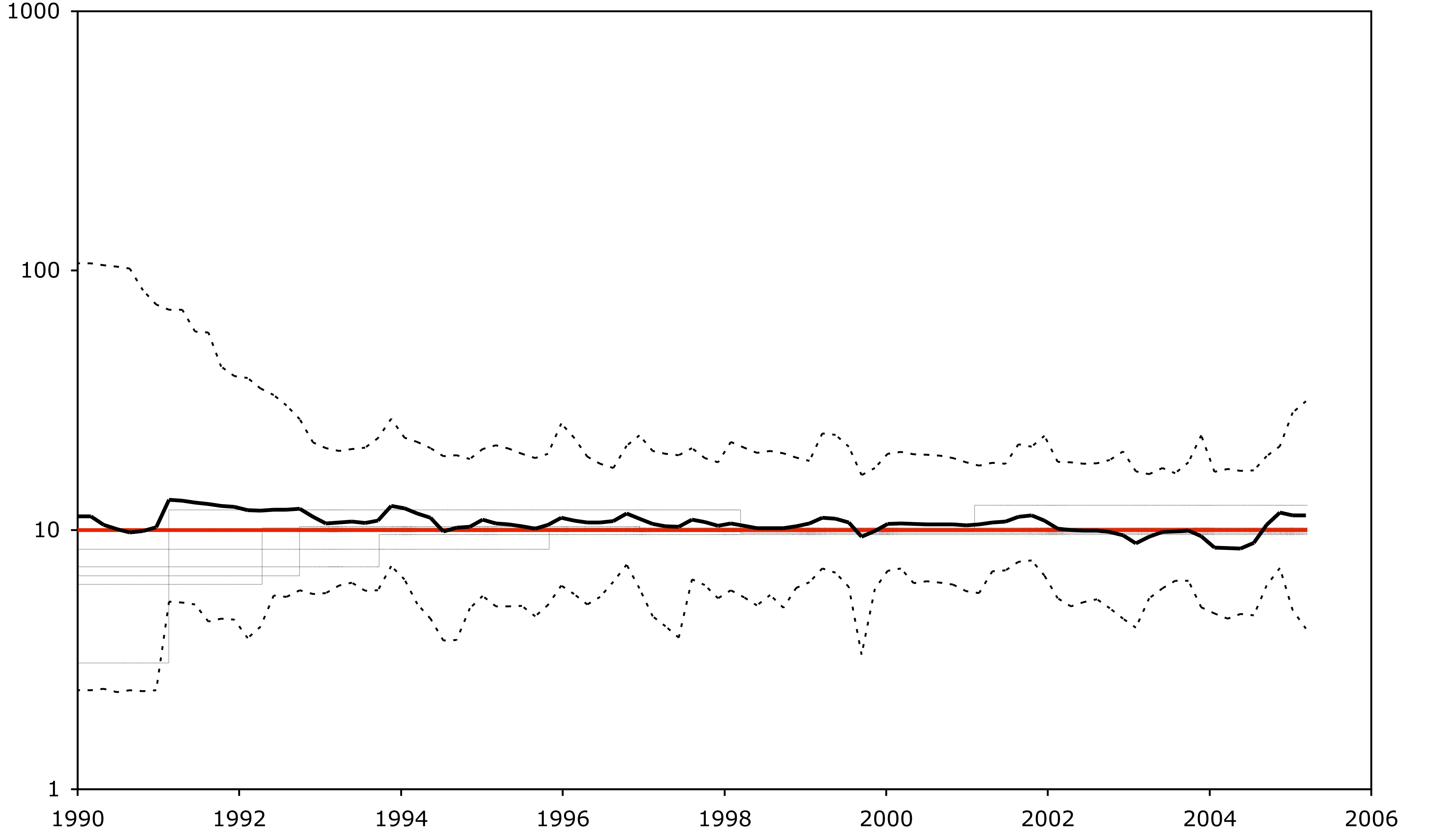
- 697 complete genomes (all 8 segments) from New York State sampled over 10 years.
- Genotypes H3N2 & H1N1
- Comparable data from the Southern Hemisphere (New Zealand + Australia)
- Precise dates of sampling

Heterogenous sampling through time

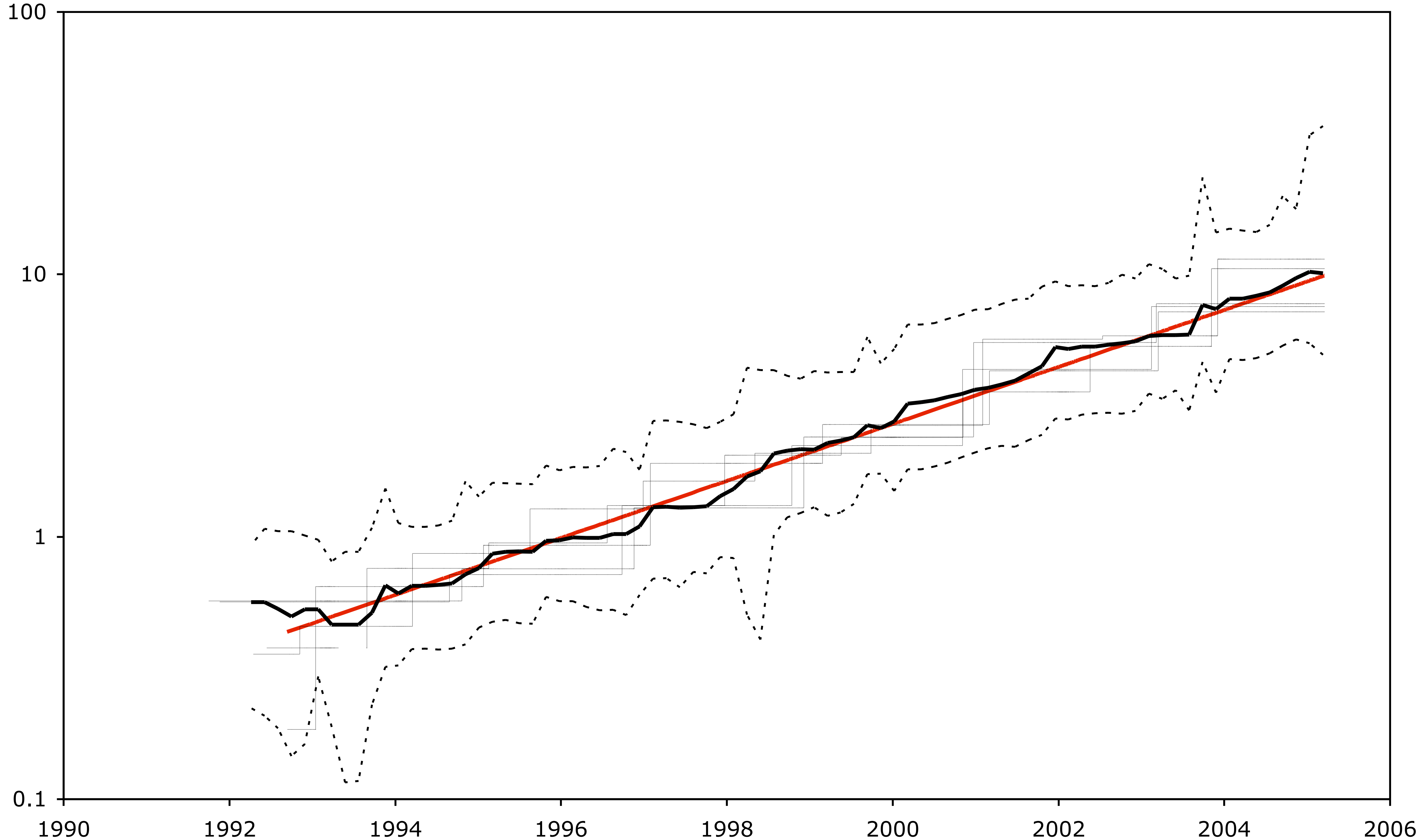
- 697 H3N2 samples from New York State 1993-2005.



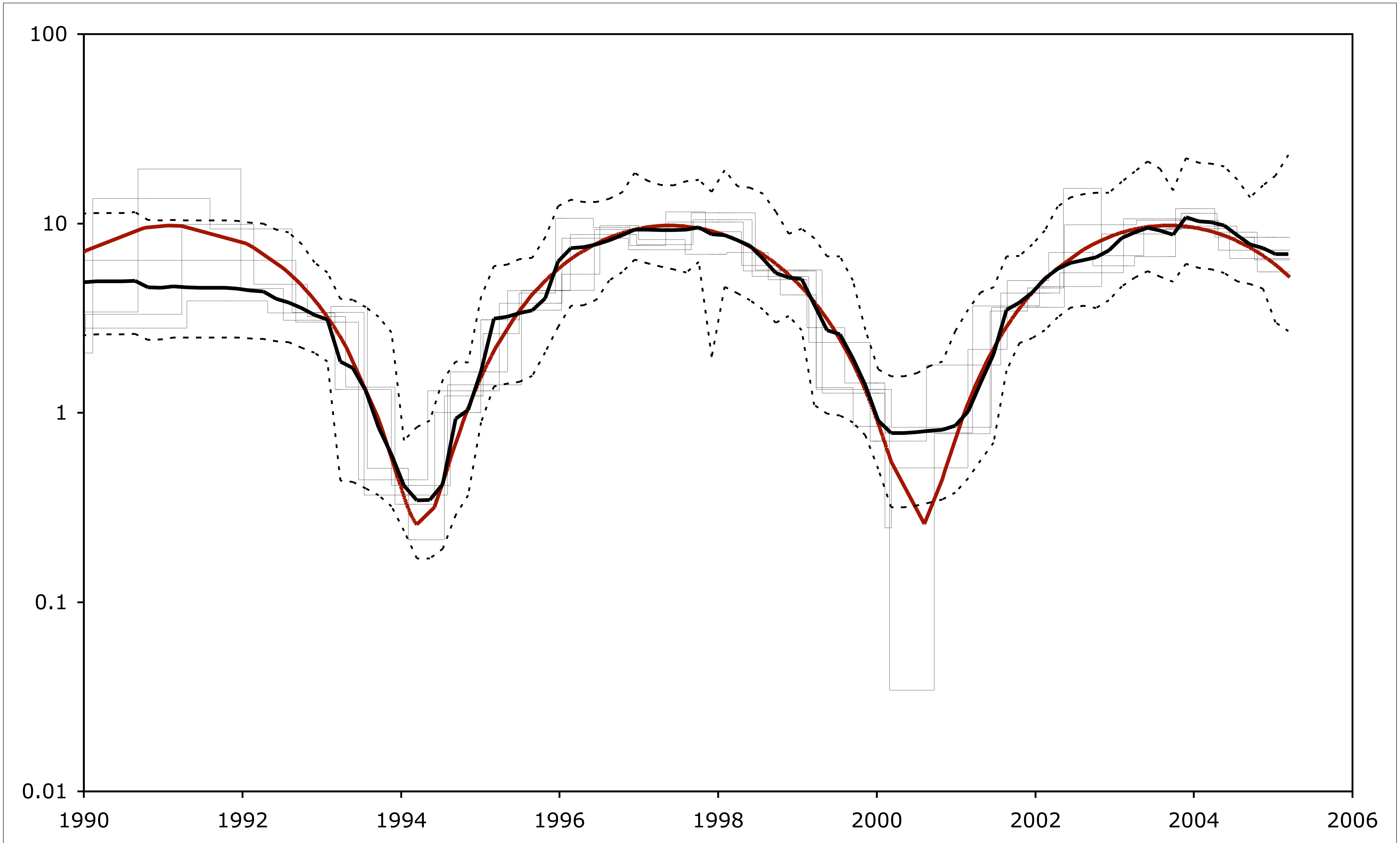
Does heterogenous sampling produce fluctuations?



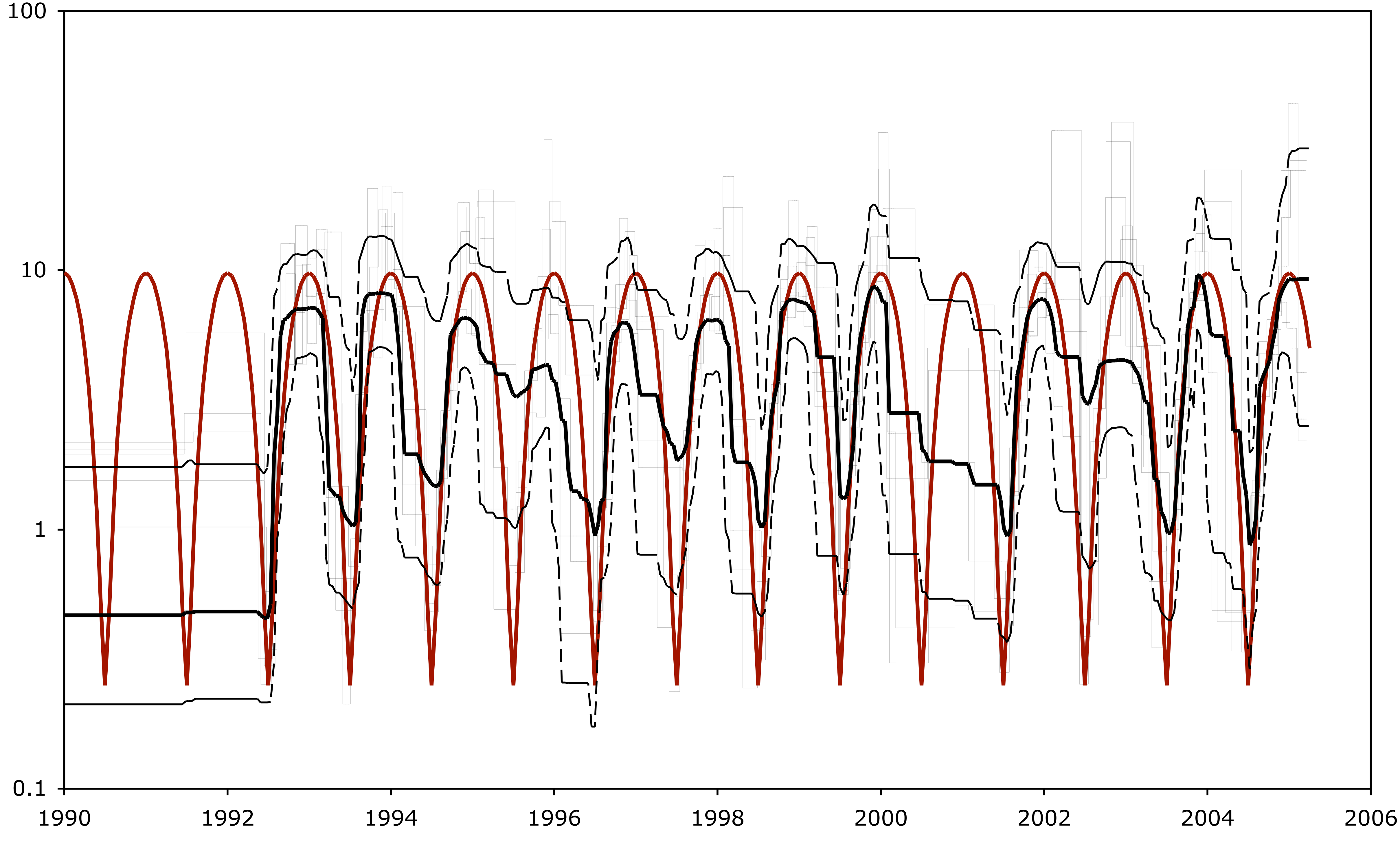
Exponential Growth



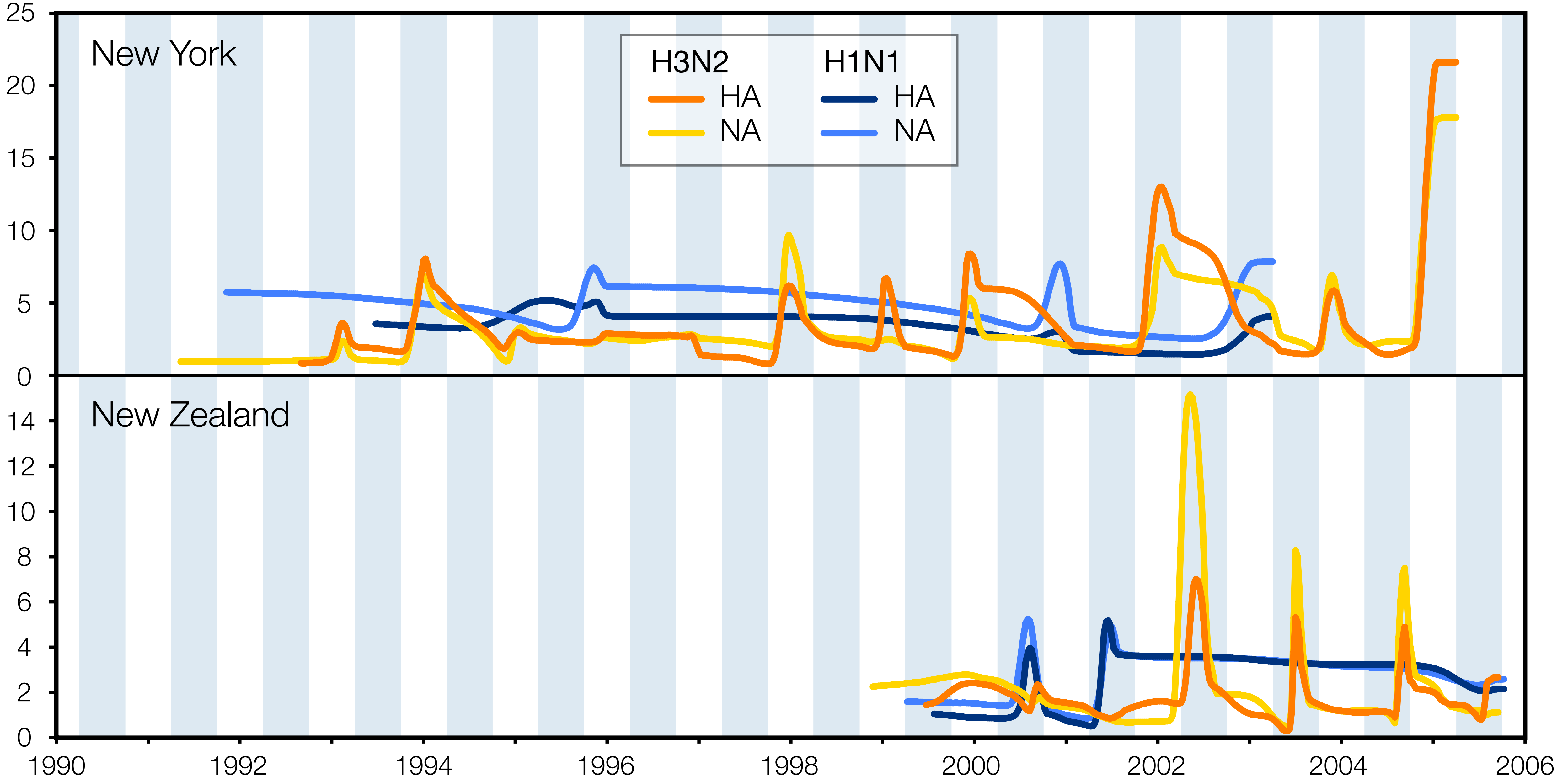
Sinusoidal Fluctuations - long period



Sinusoidal Fluctuations - annual period

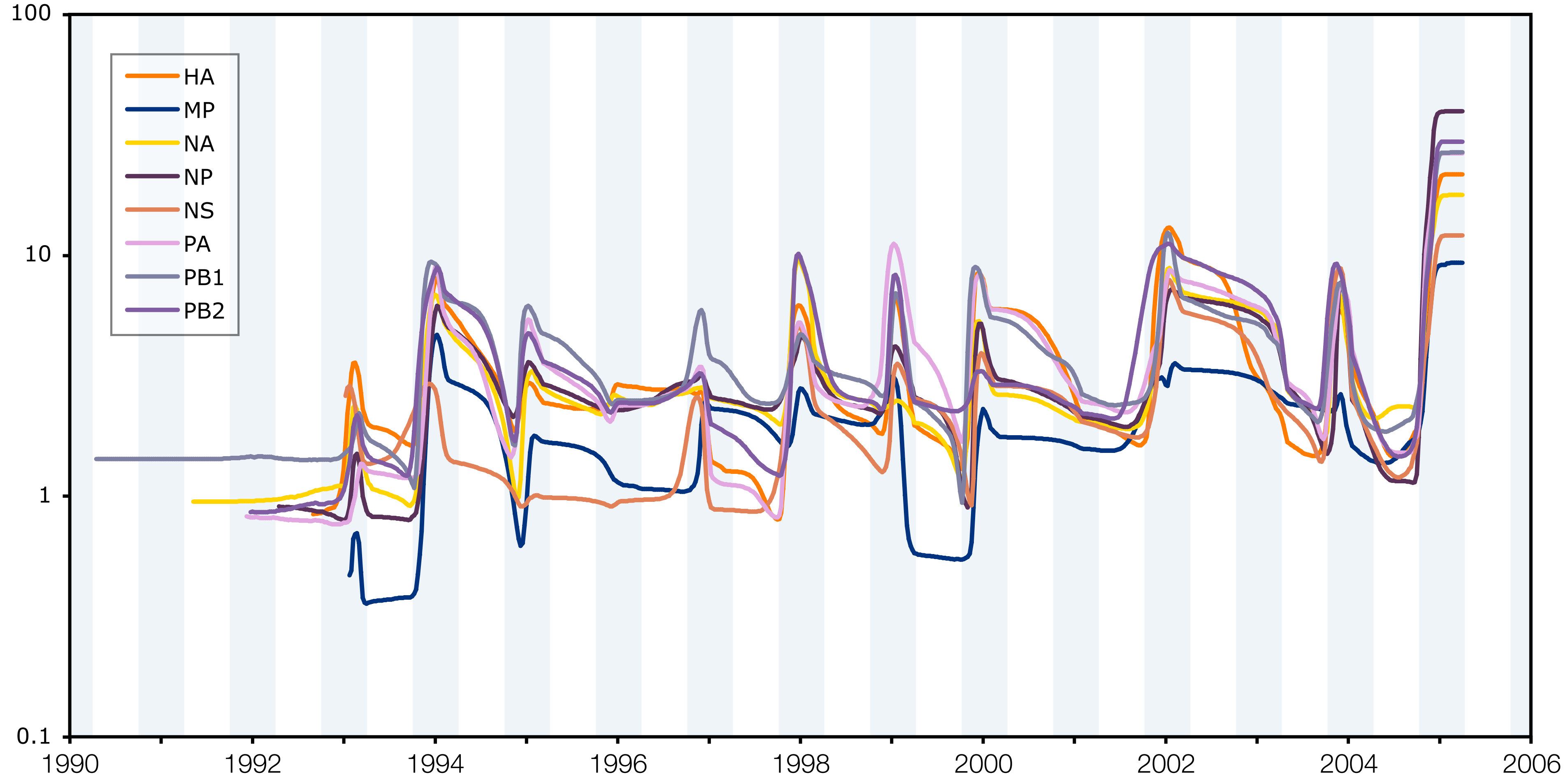


Seasonal dynamics of influenza A

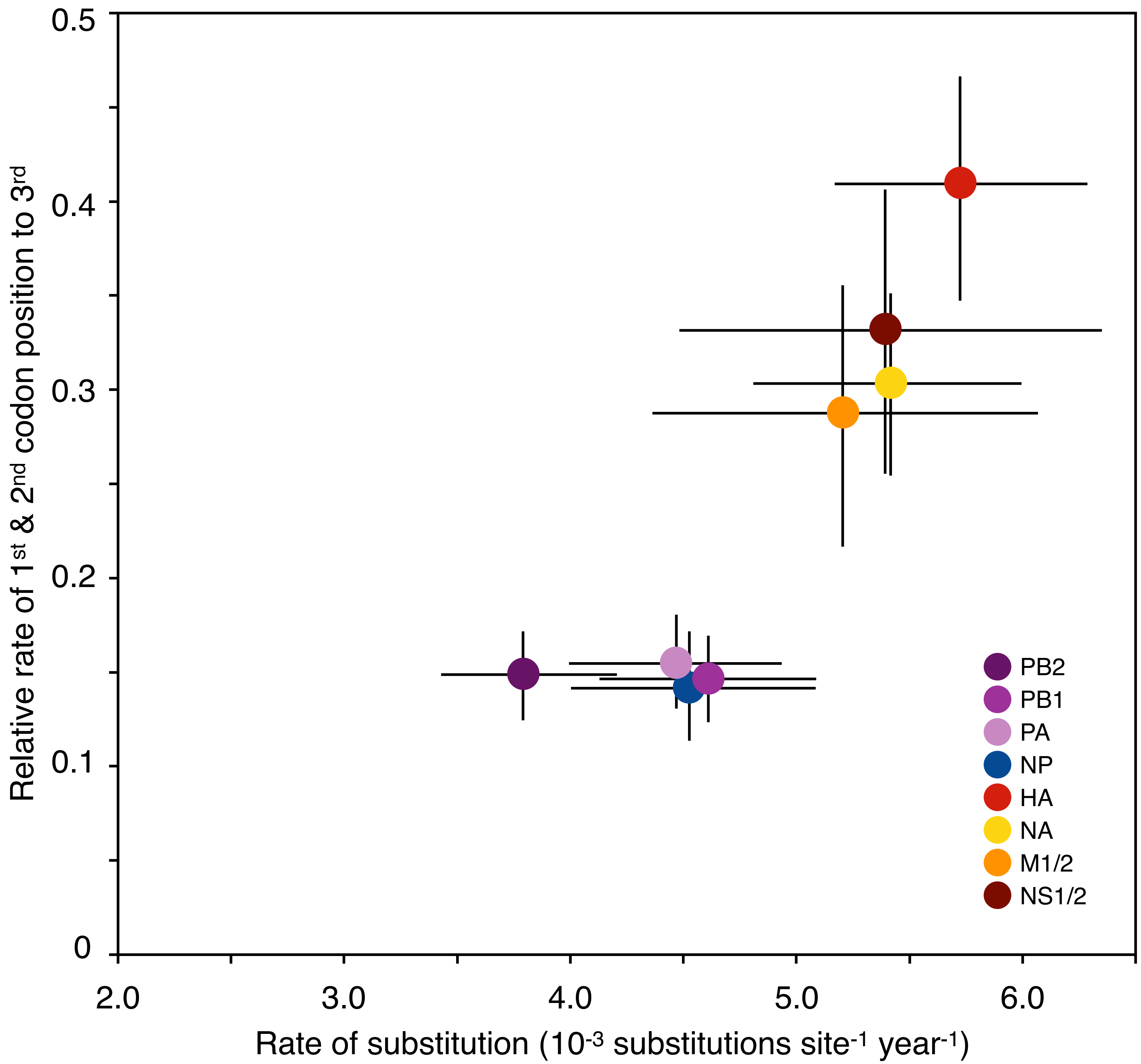


different evolutionary dynamics of segments

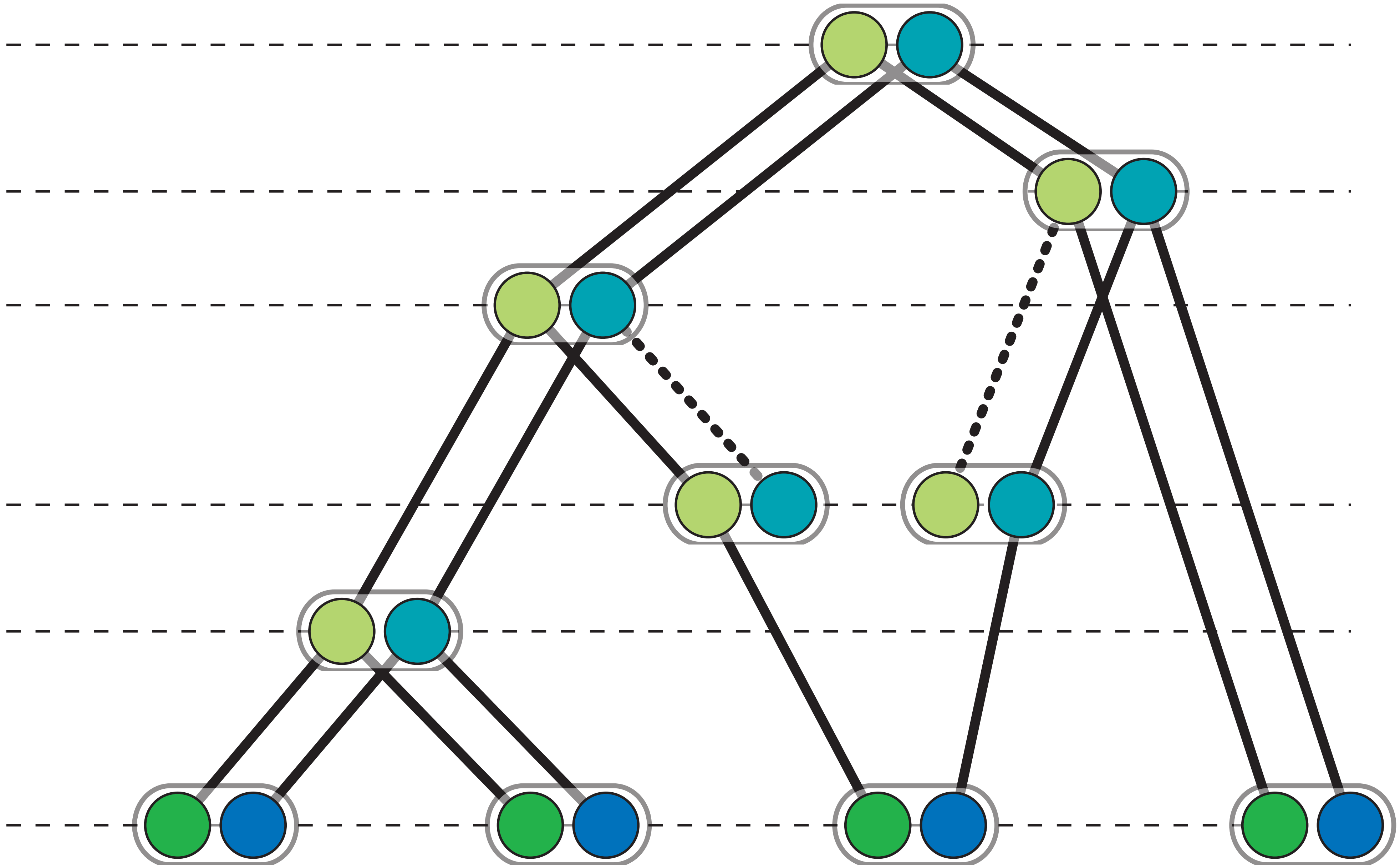
- estimate the population dynamics for each of the 8 segments individually



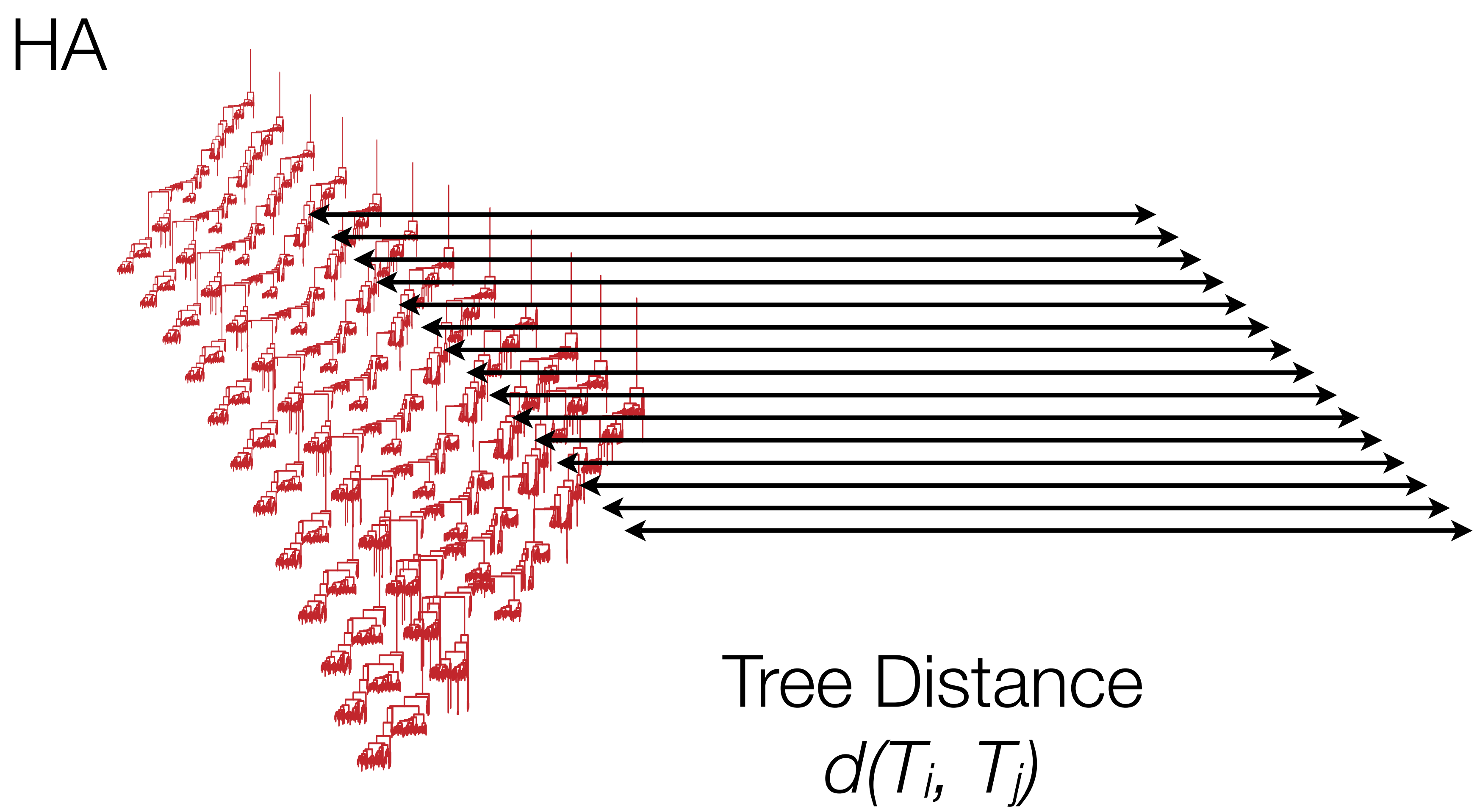
Evolutionary dynamics of influenza A



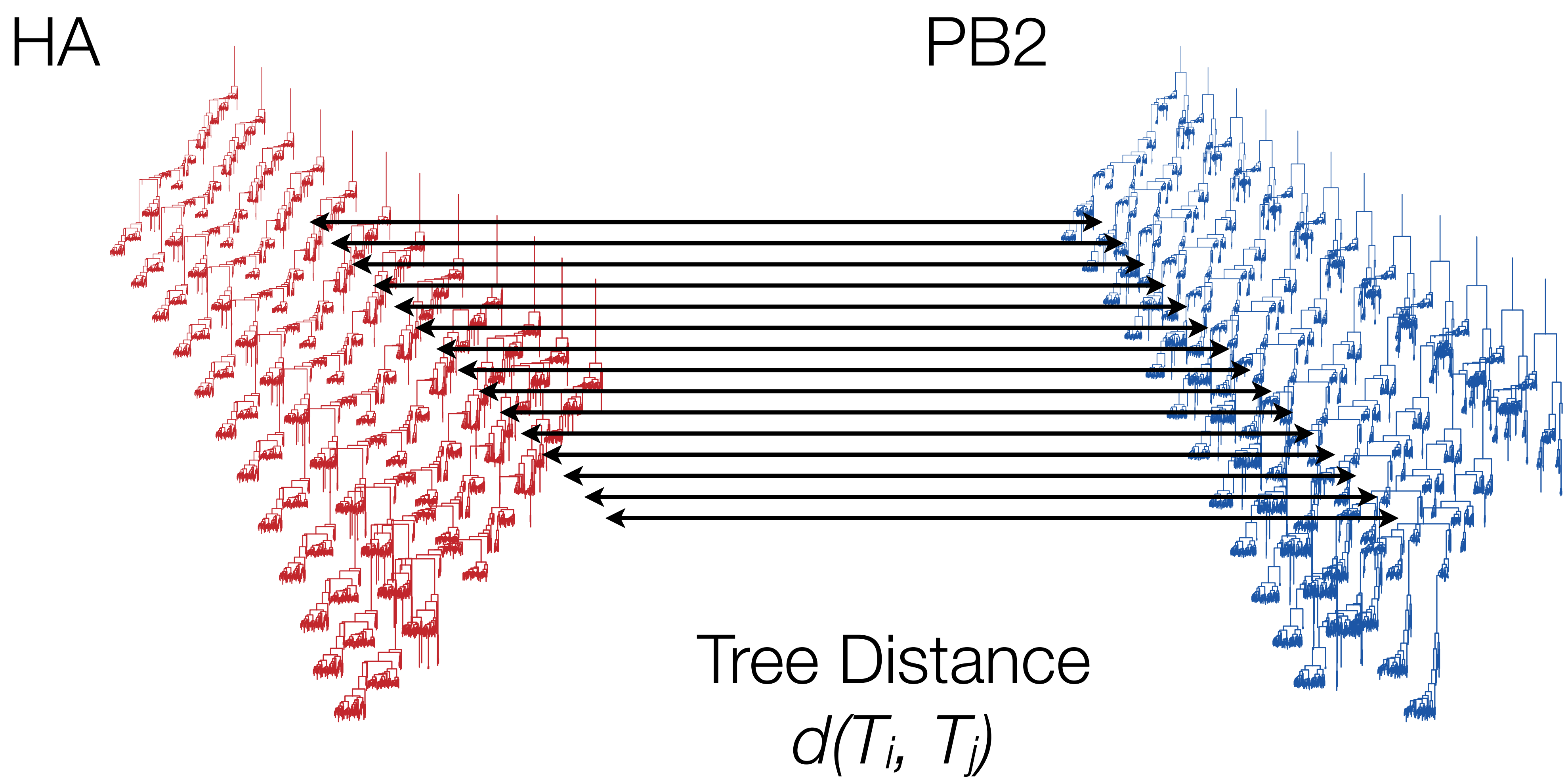
Reassortment



Phylogenetic uncertainty for a single segment

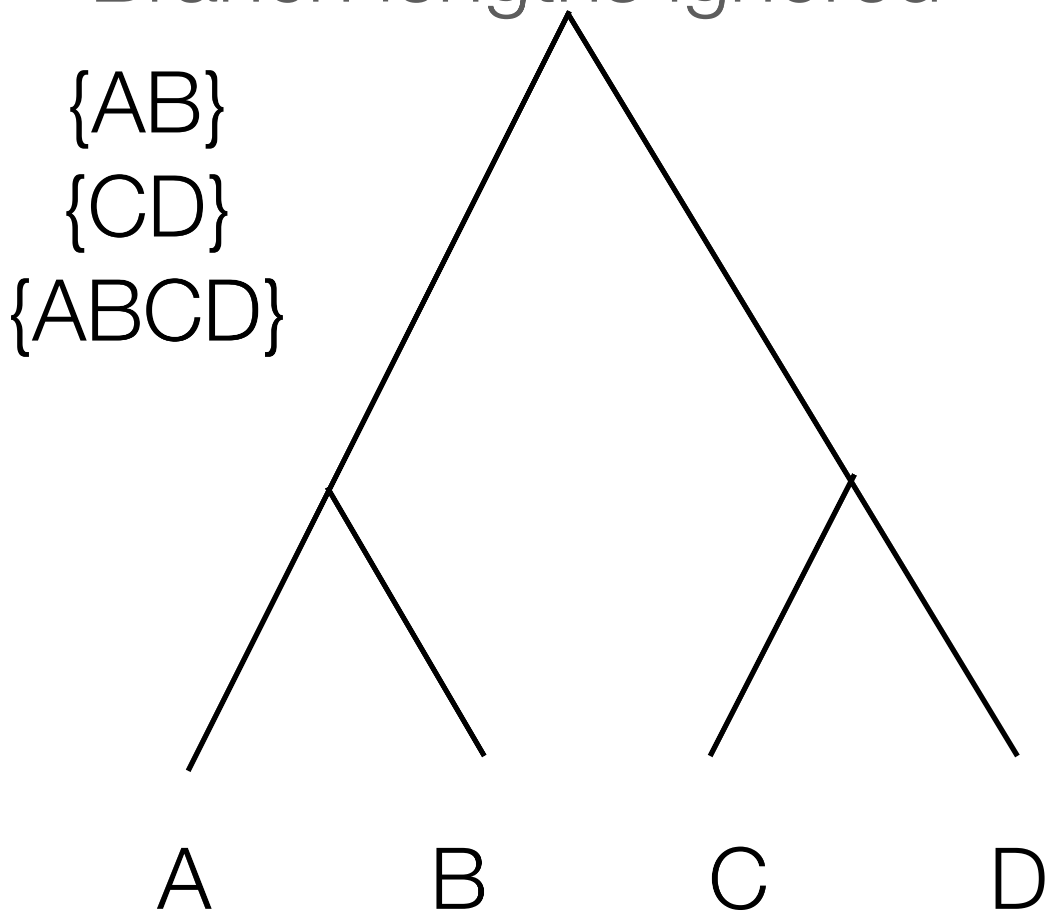


Reassortment of genomic segments

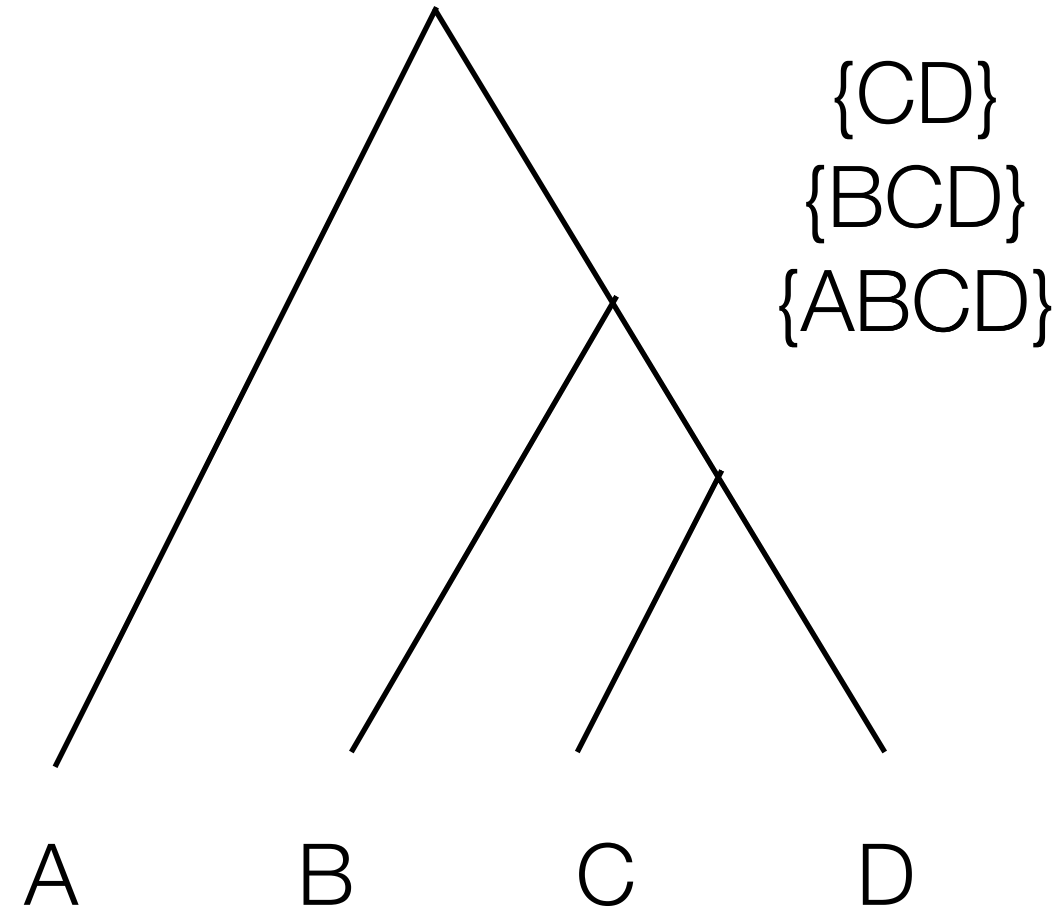


Robinson and Fould's Tree Distance

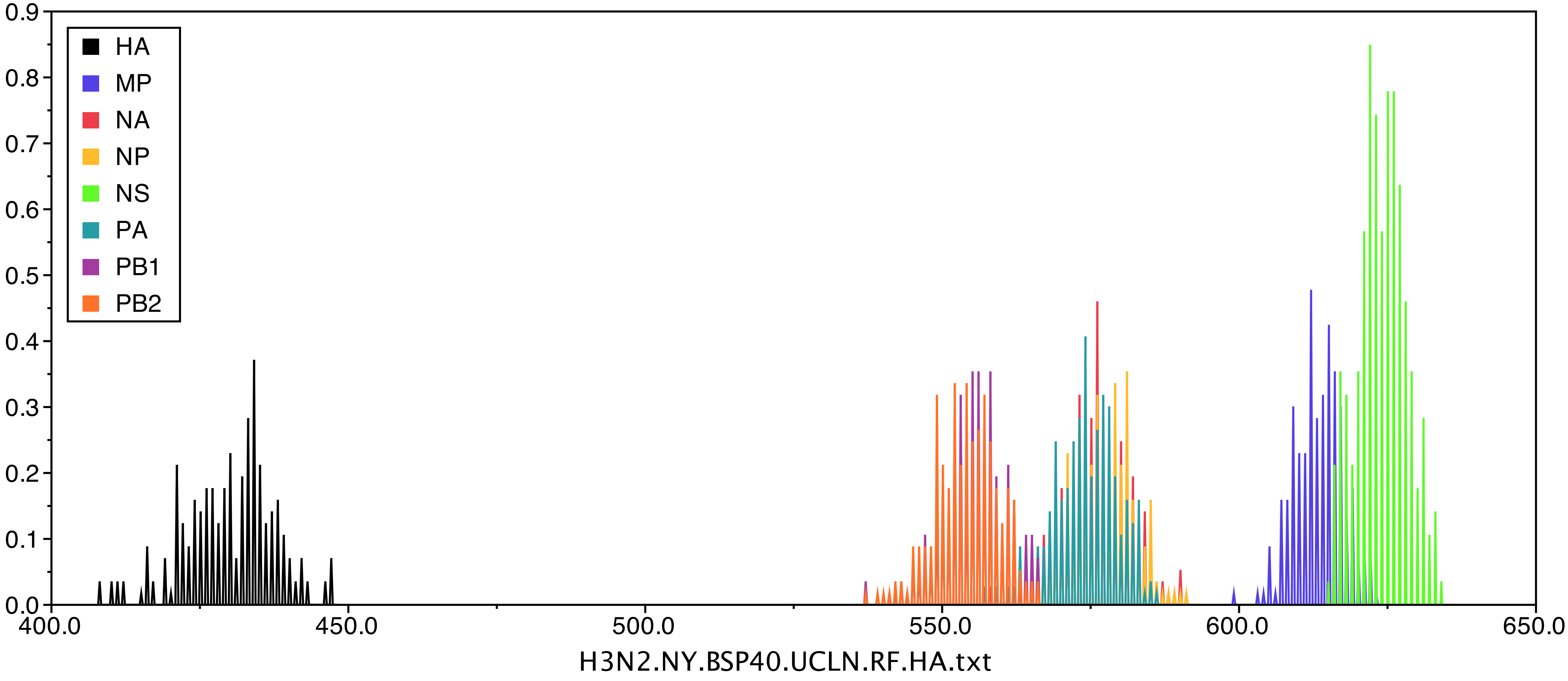
- Count of number of clades
- RF = number of clades in one tree but not the other.
- Clades weighted equally irrespective of size
- Branch lengths ignored



RF = 2

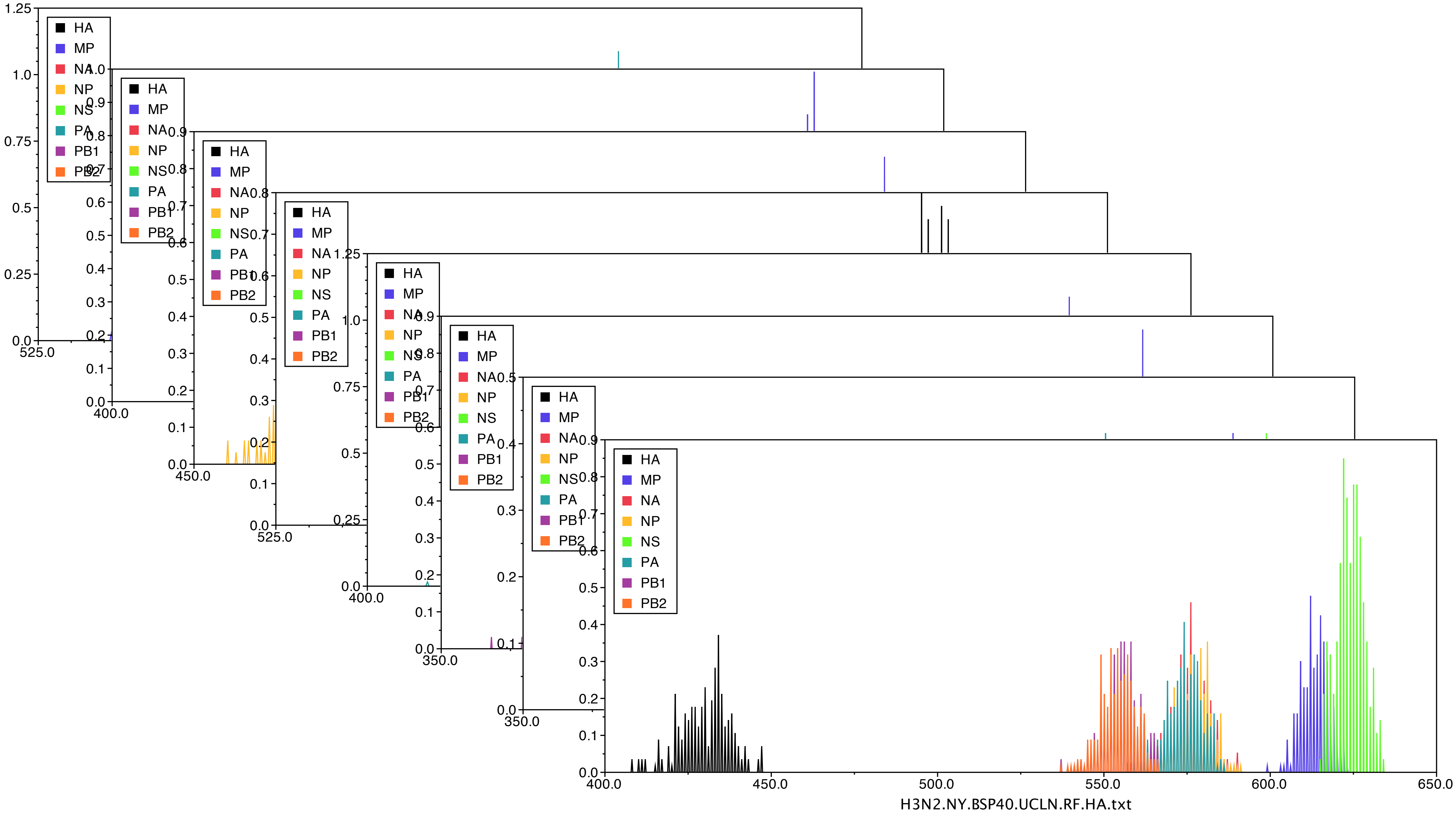


Reassortment of genomic segments - RF distances

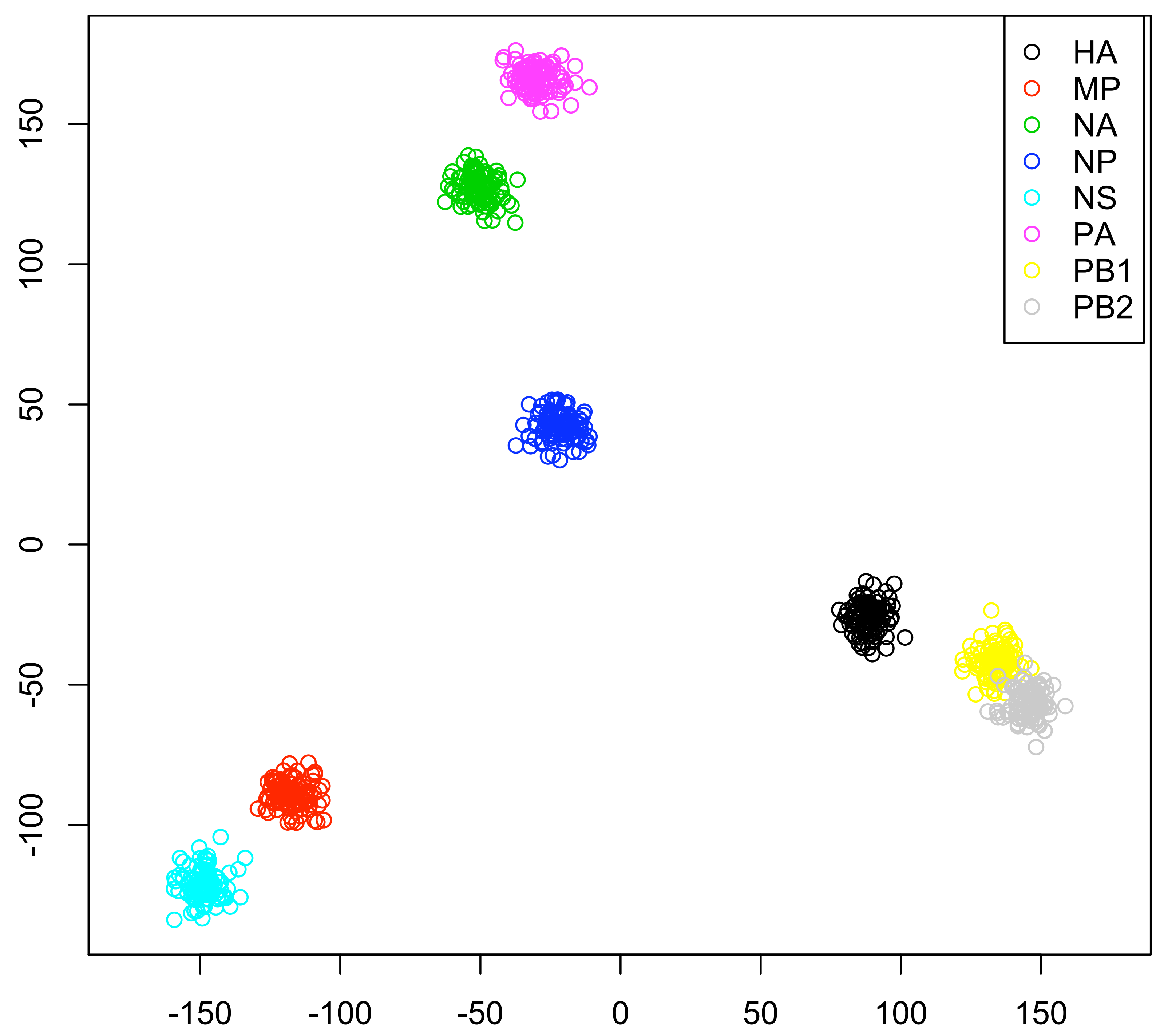


- HA segment vs others

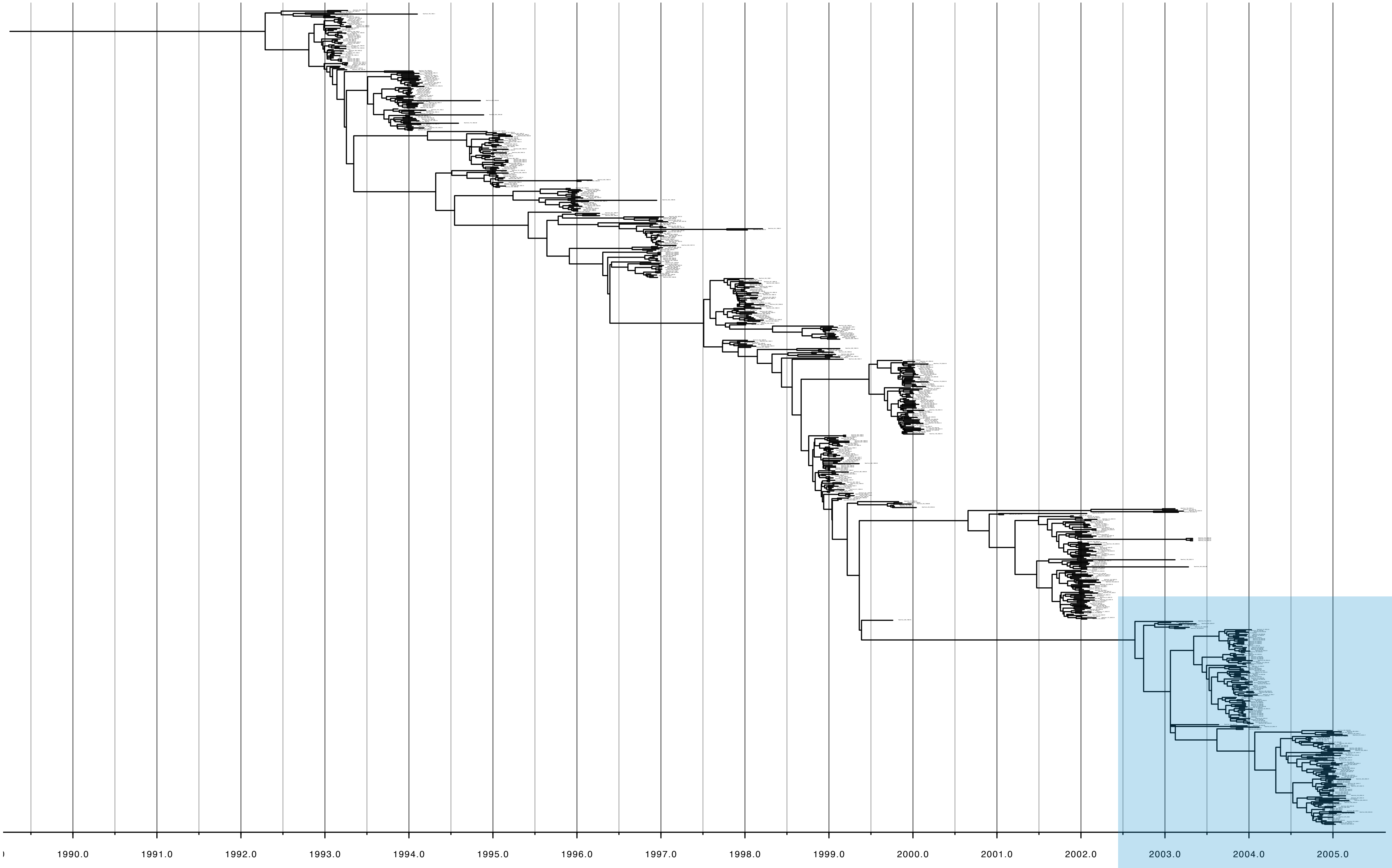
Reassortment of genomic segments - RF distances



RF distances using multi-dimensional scaling

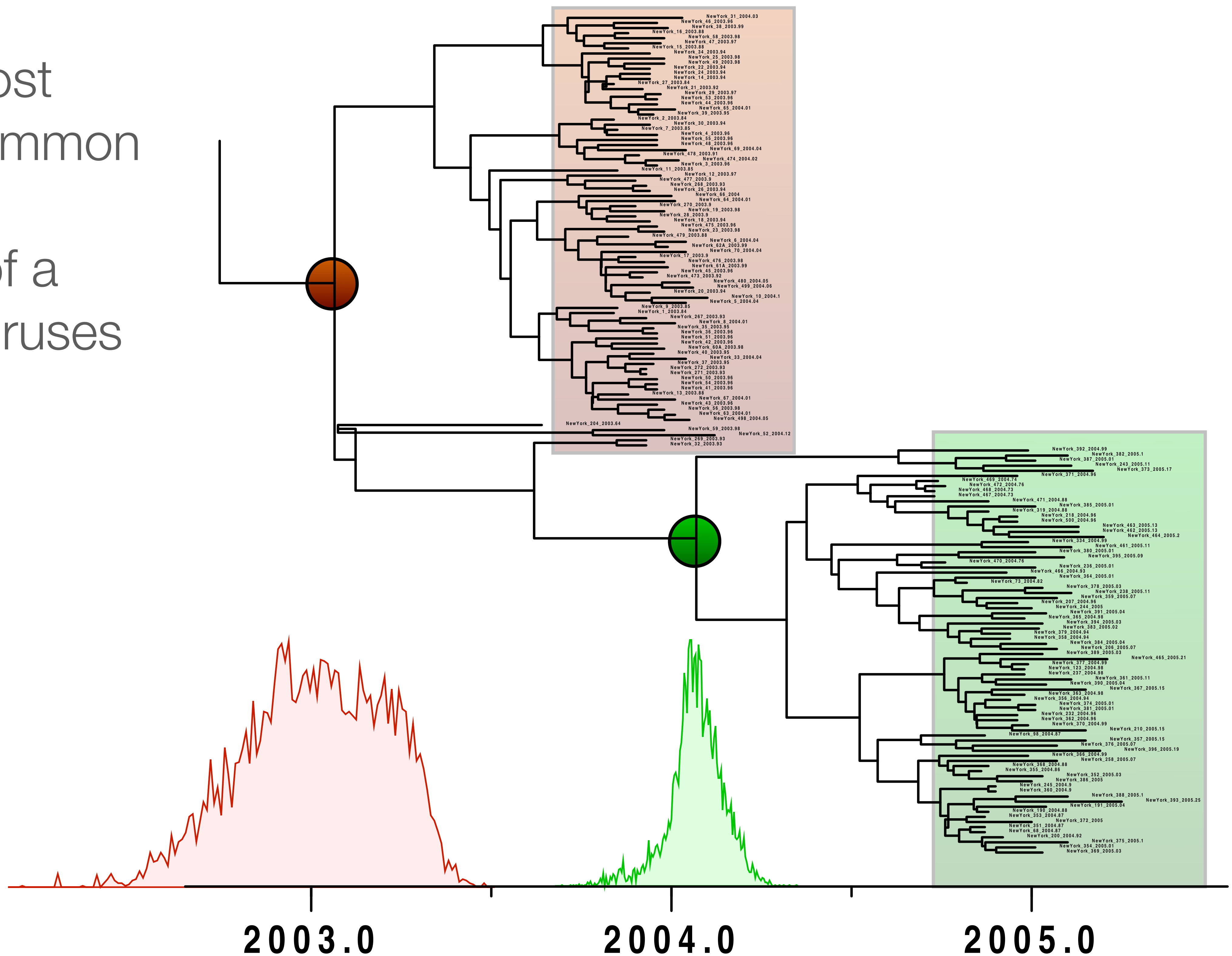


Genealogy of H3N2 HA genome segment (New York)

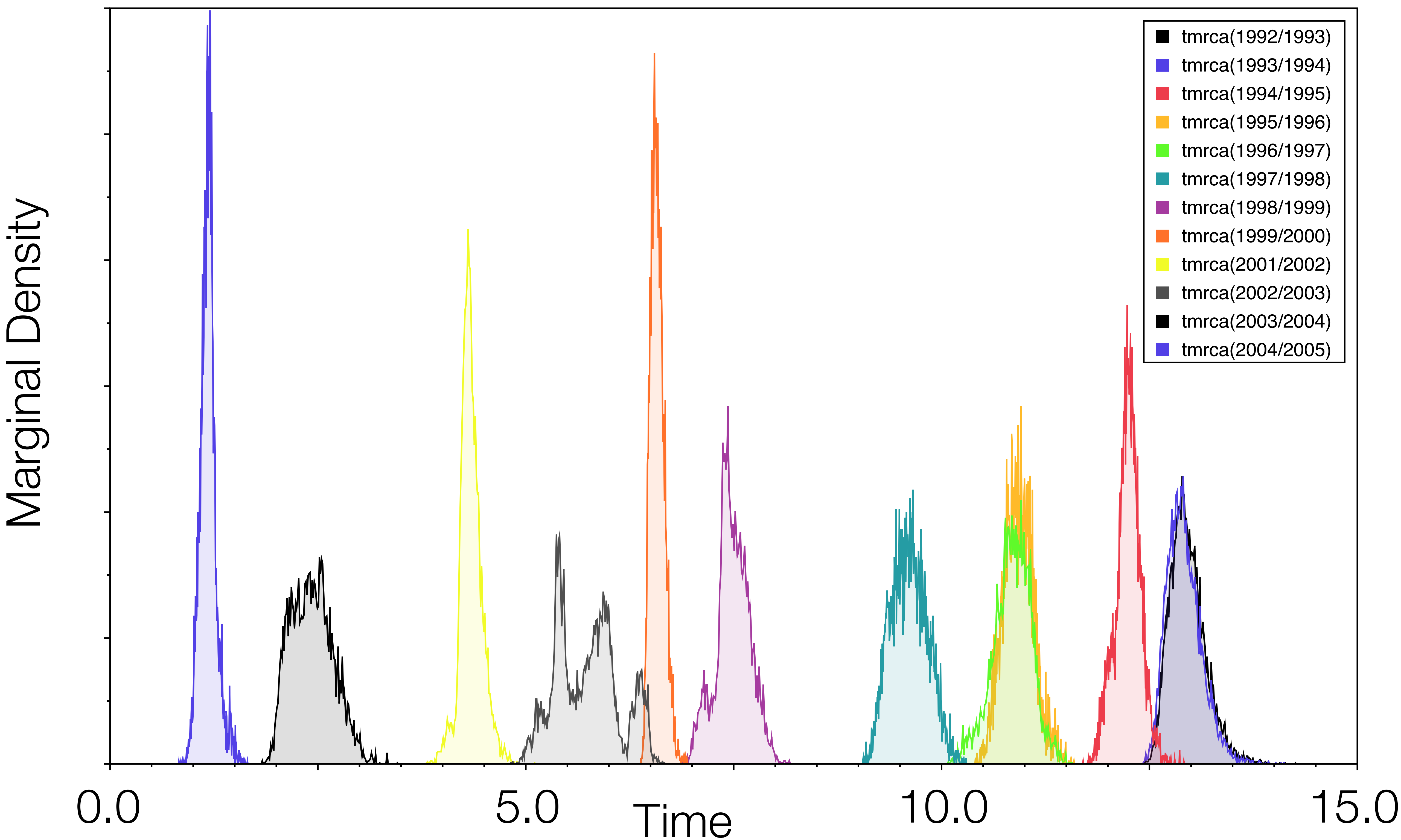


Genealogy of H3N2 HA genome segment (New York)

- Time of Most Recent Common Ancestor (TMRCA) of a season's viruses

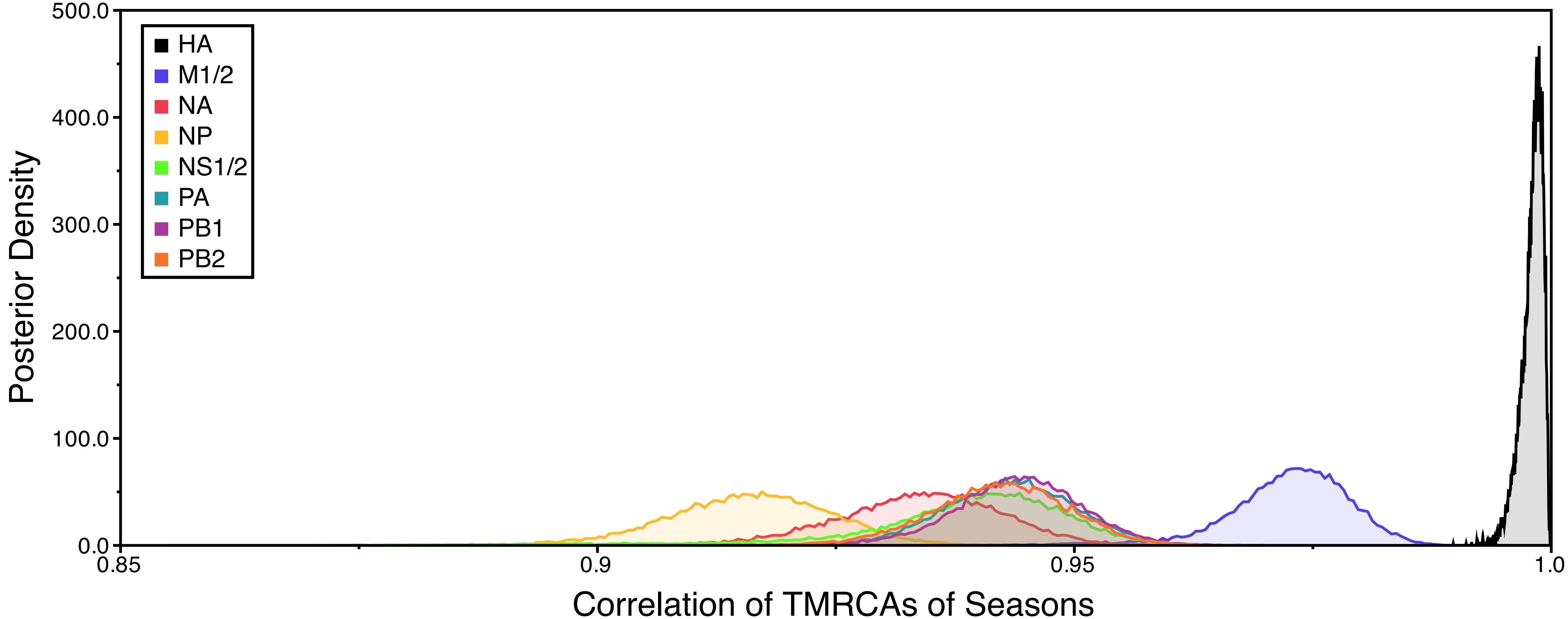


Posterior density of TMRCA of HA by season



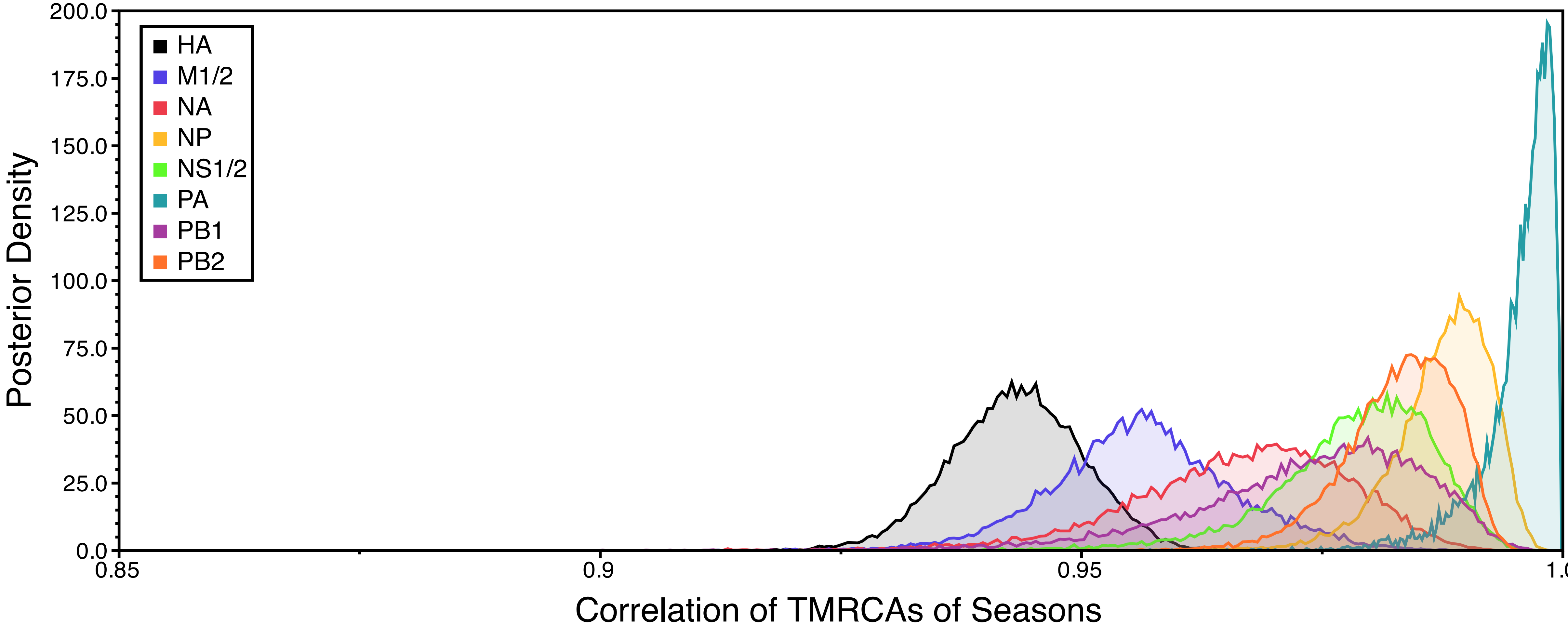
Reassortment of genomic segments

- Vector of TMRCAs for each season for each tree
- $d(T_1, T_2) = 1 - C(V_1, V_2)$



- HA segment vs others

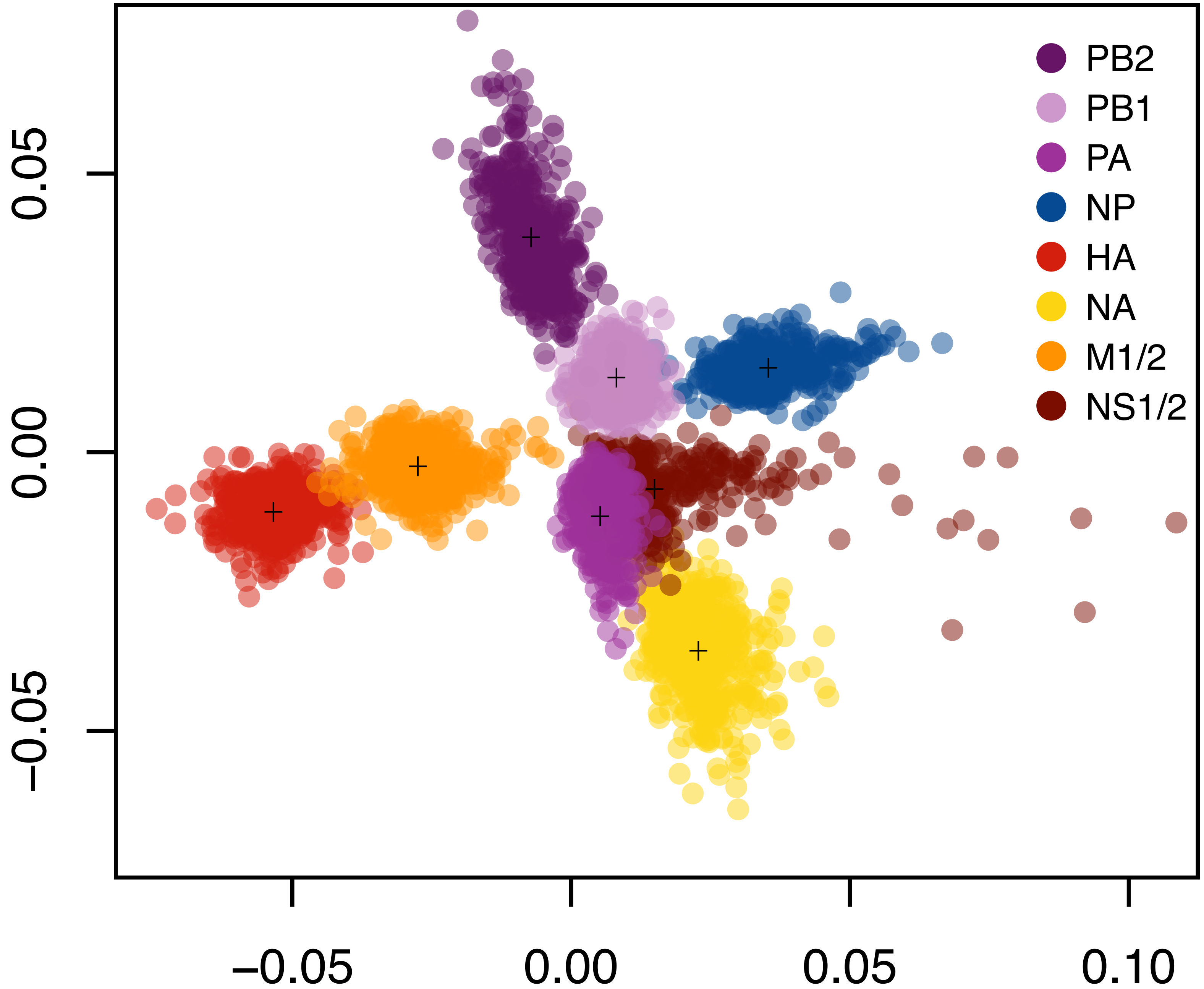
Reassortment of genomic segments



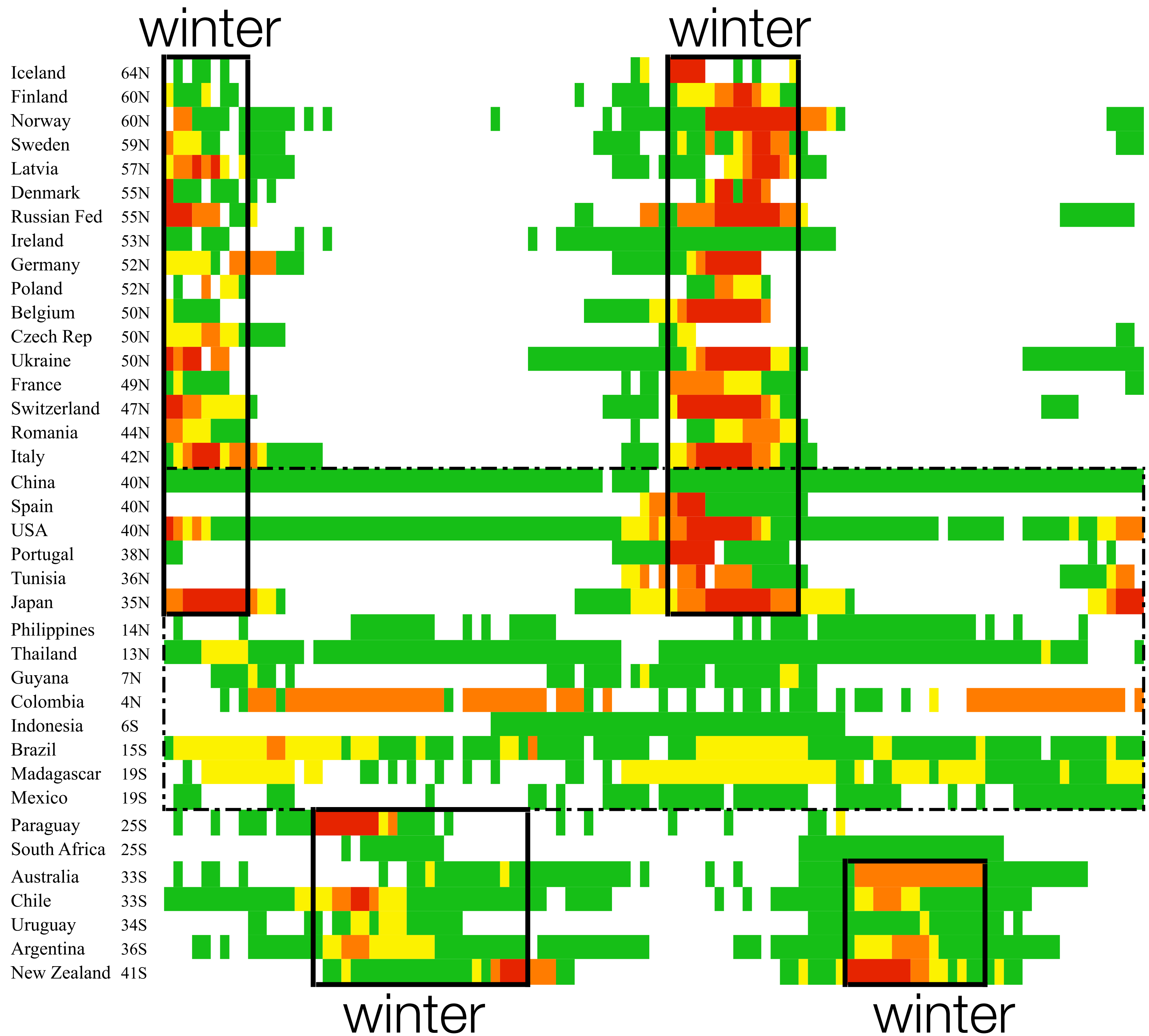
- PA segment vs others

Reassortment of genomic segments

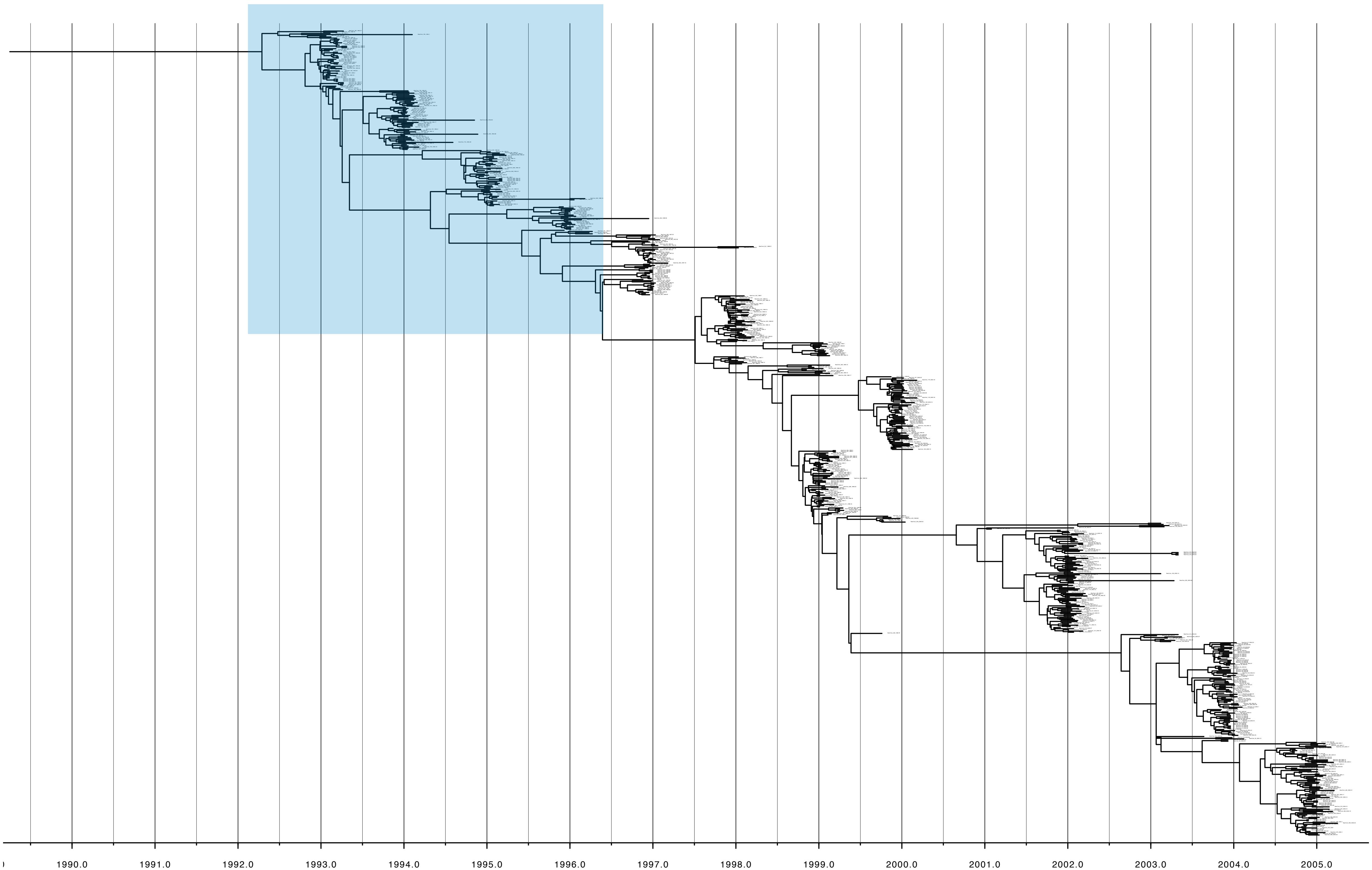
- distance metric between all pairs of trees
- plotted using MDS
- close linkage between structural and functional genes
    - PA, PB1, NP & NS1/2
    - HA & MA
    - NA
    - PB2



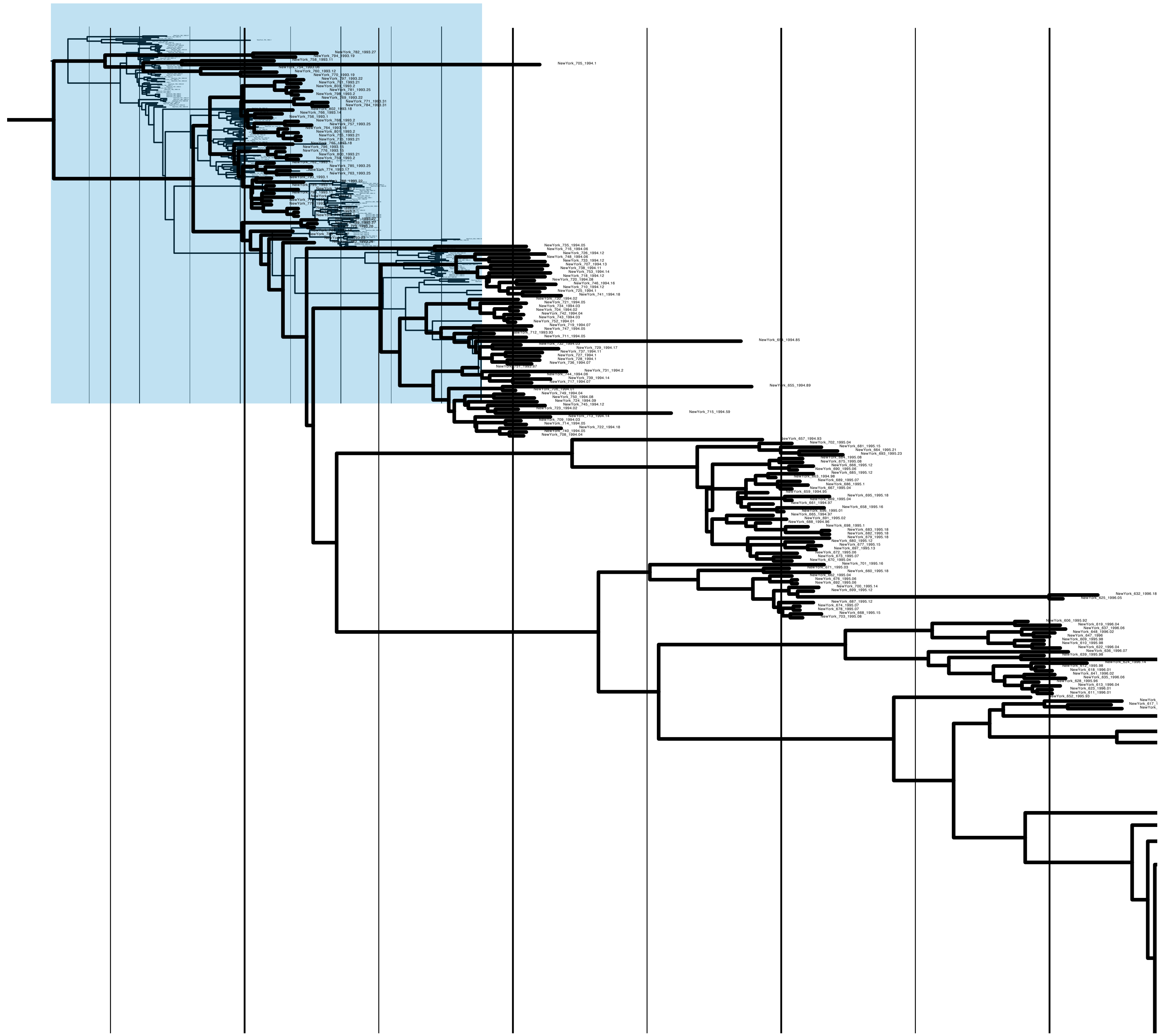
Does influenza A persist in temperate summers?



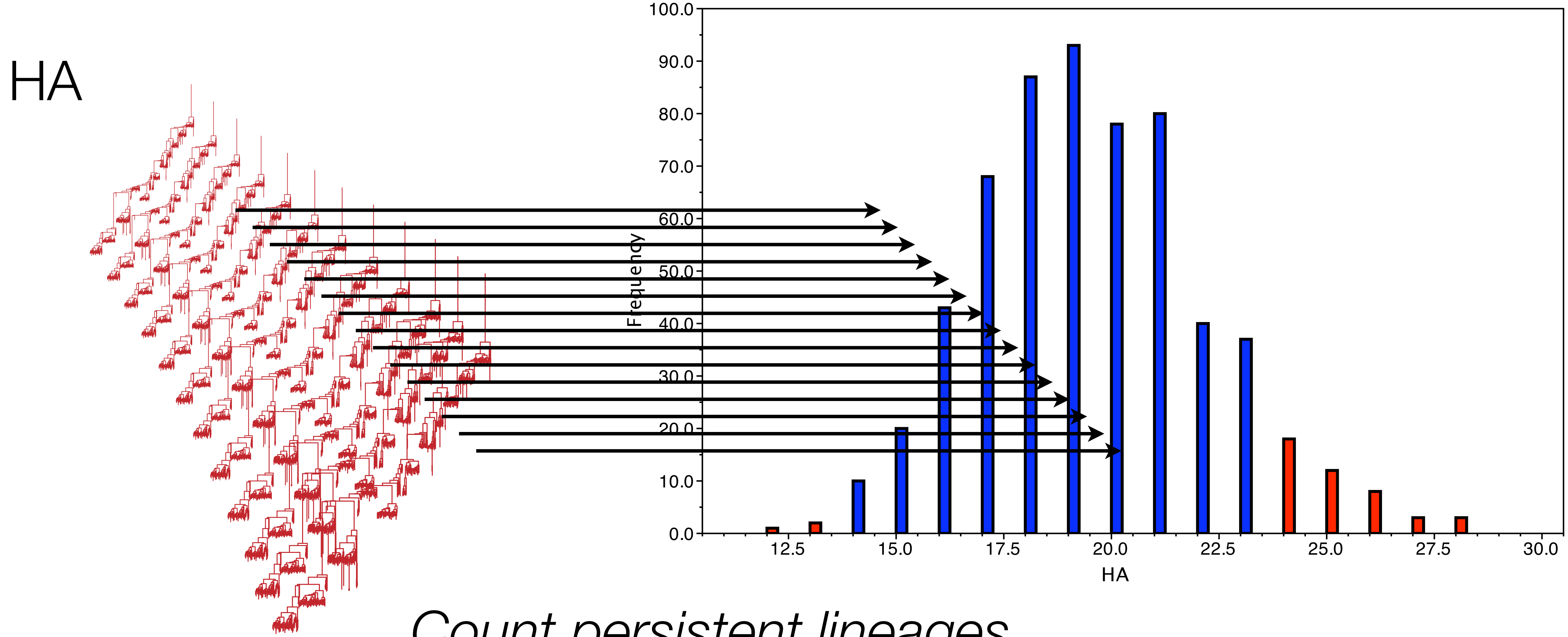
Does influenza A persist in temperate summers?



Does influenza A persist in temperate summers?

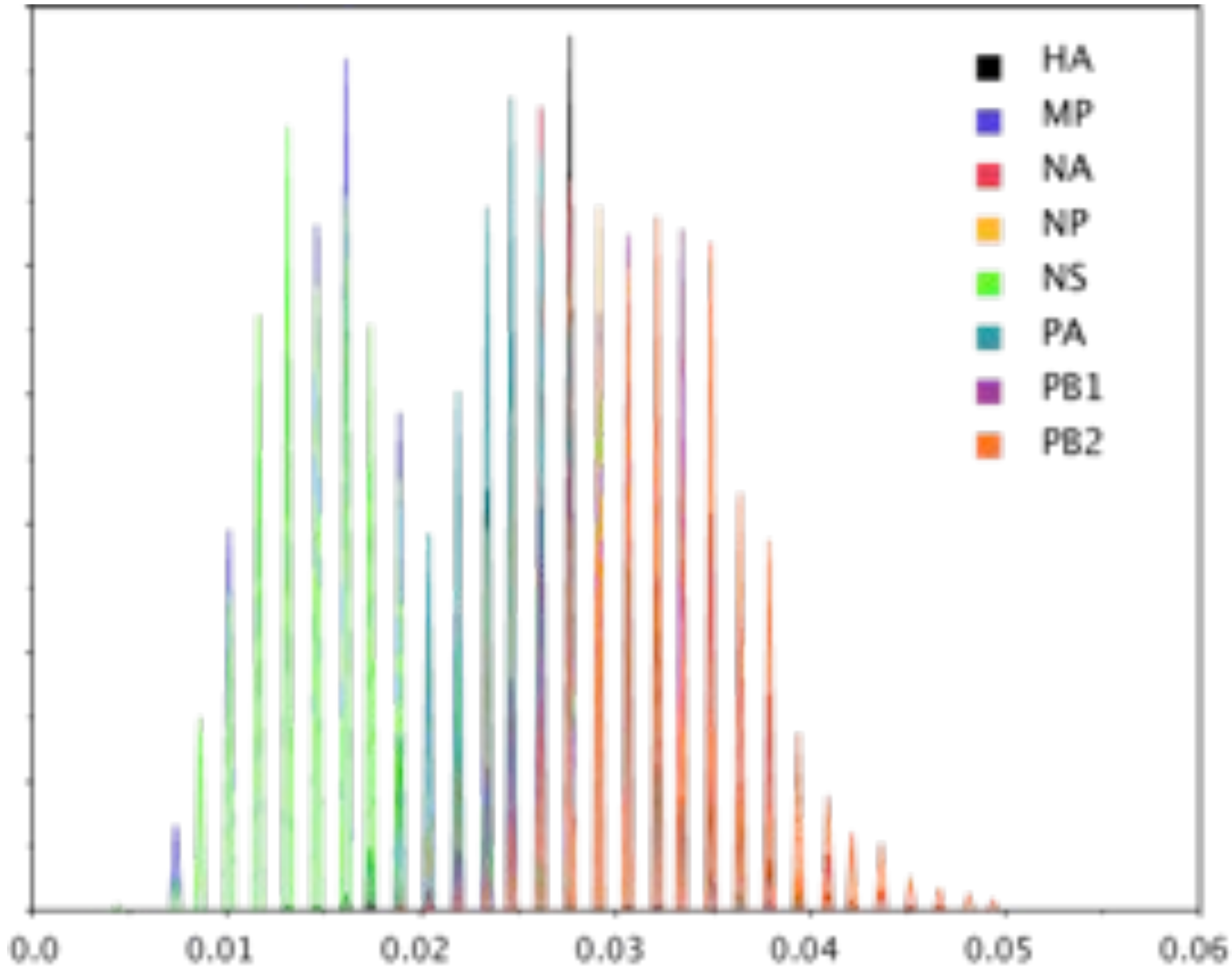


Phylogenetic uncertainty for a single segment

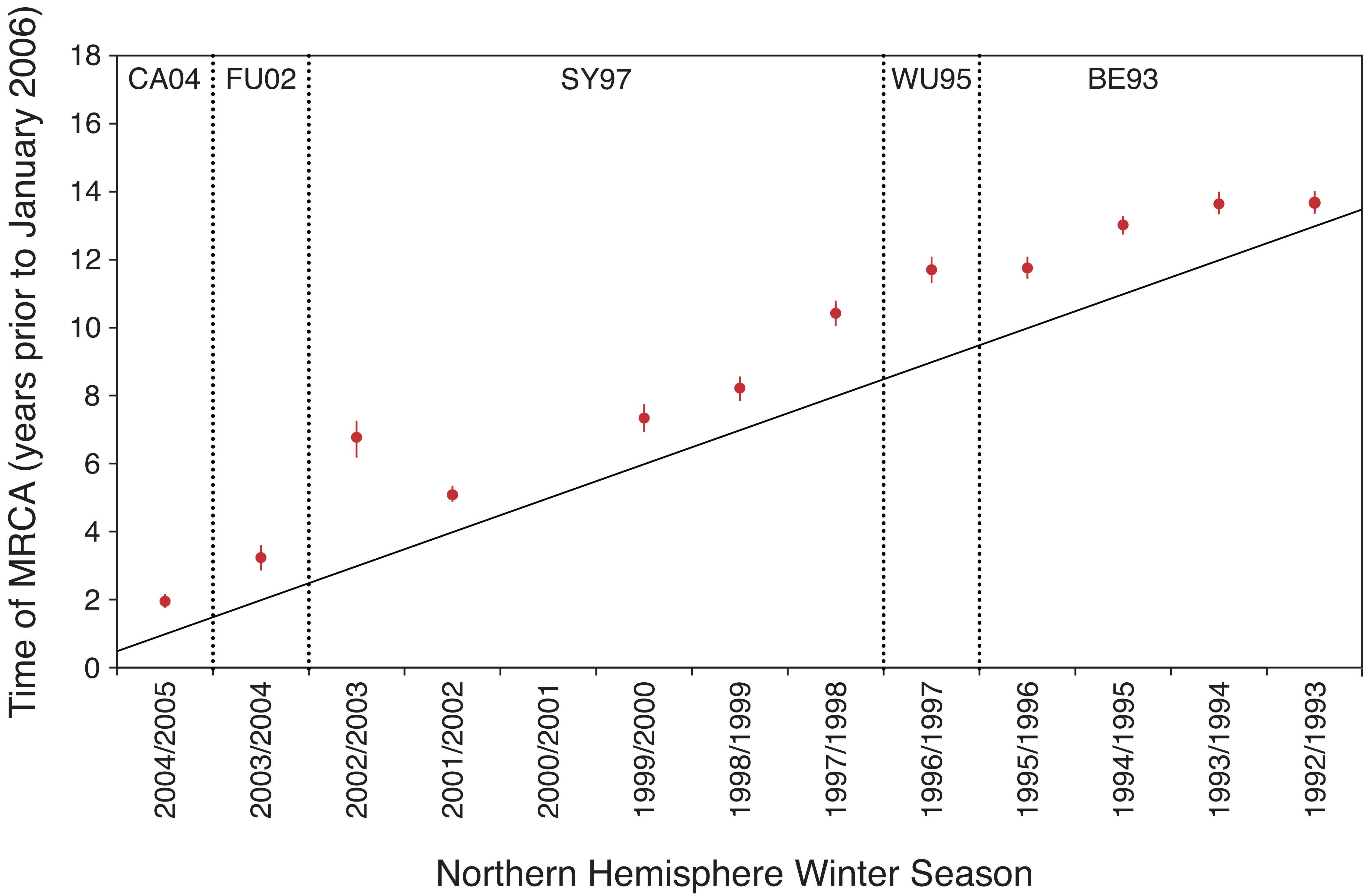


Count persistent lineages

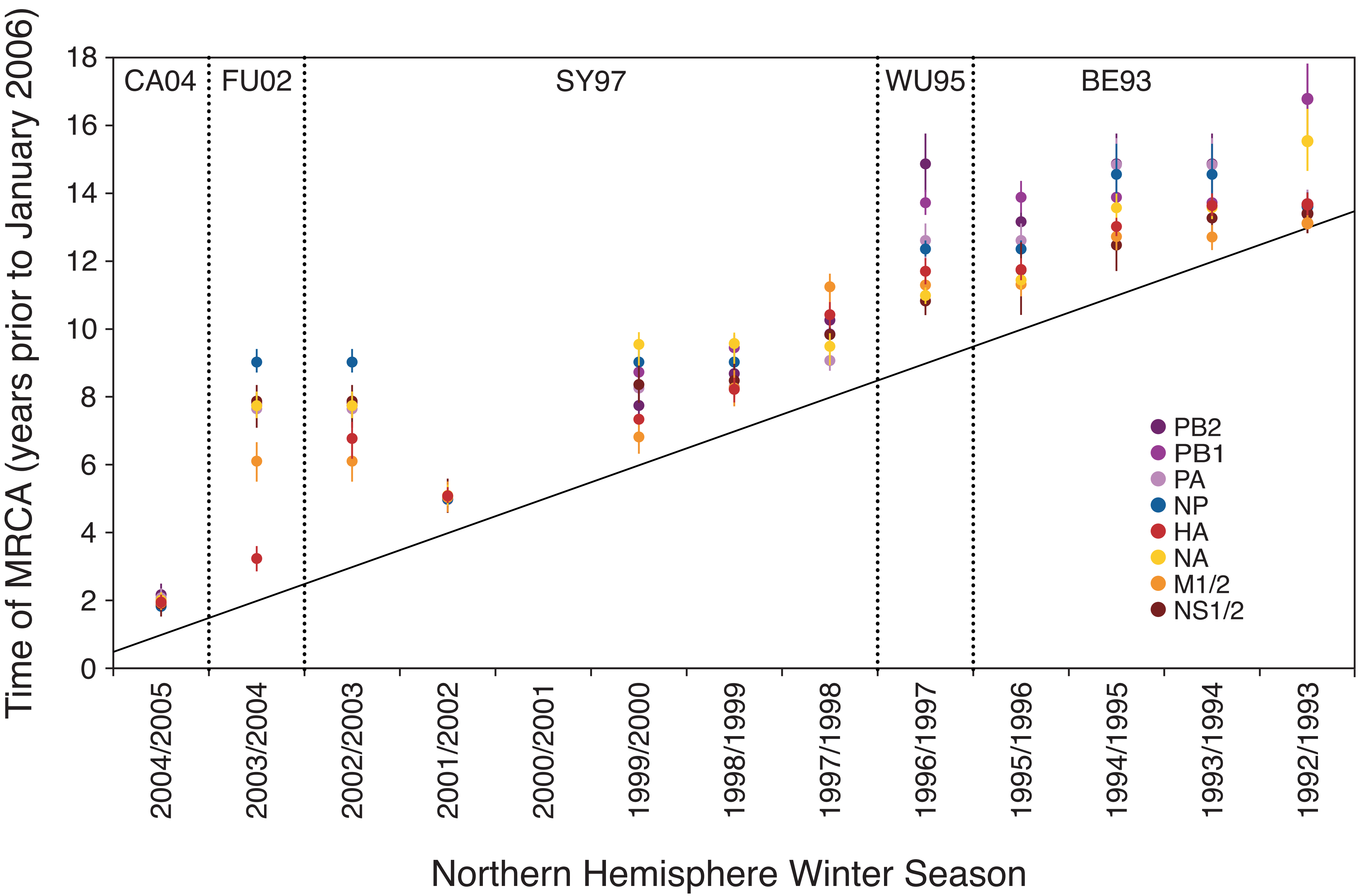
Proportion of viruses persisting from previous season



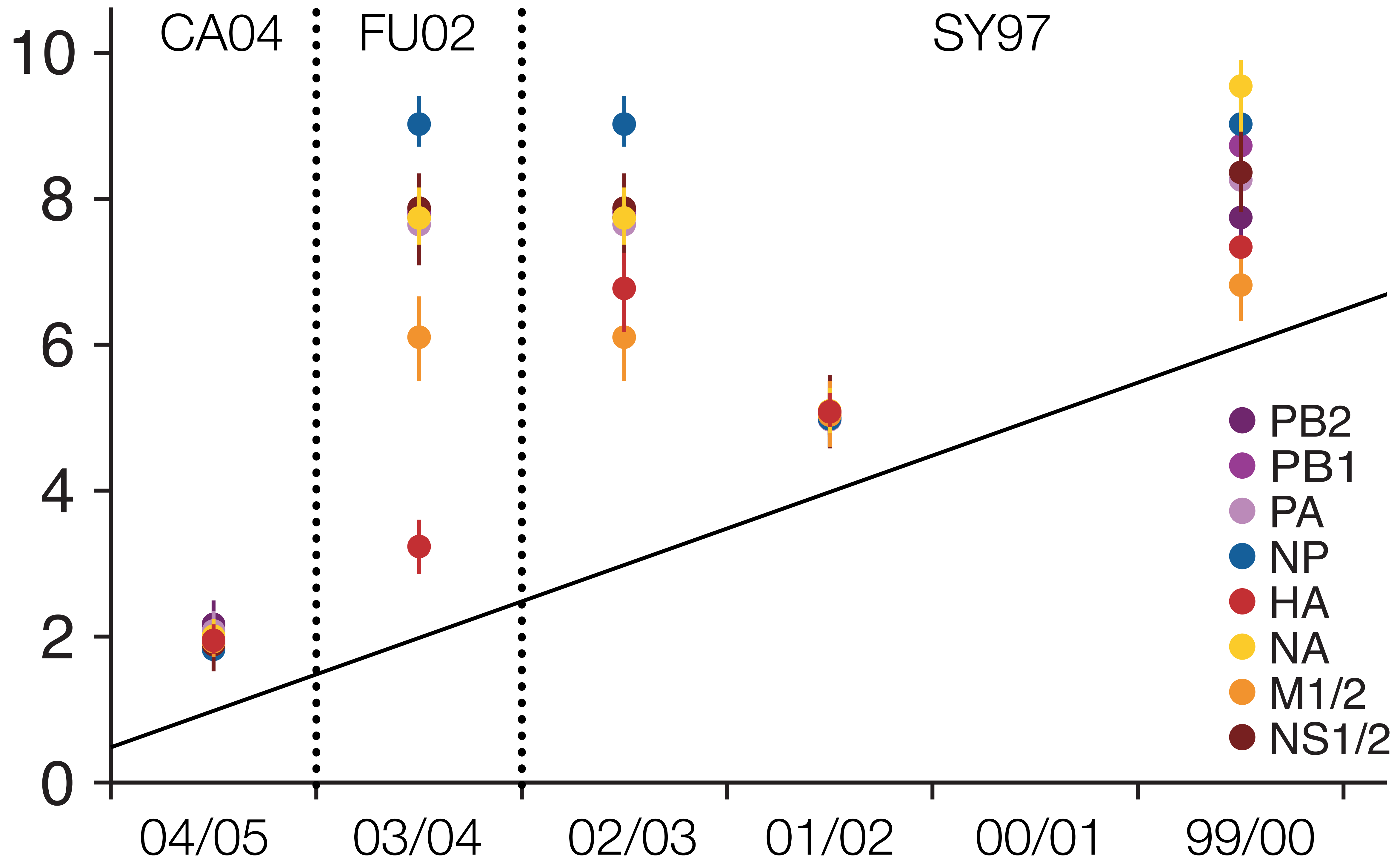
TMRCA of HA by season



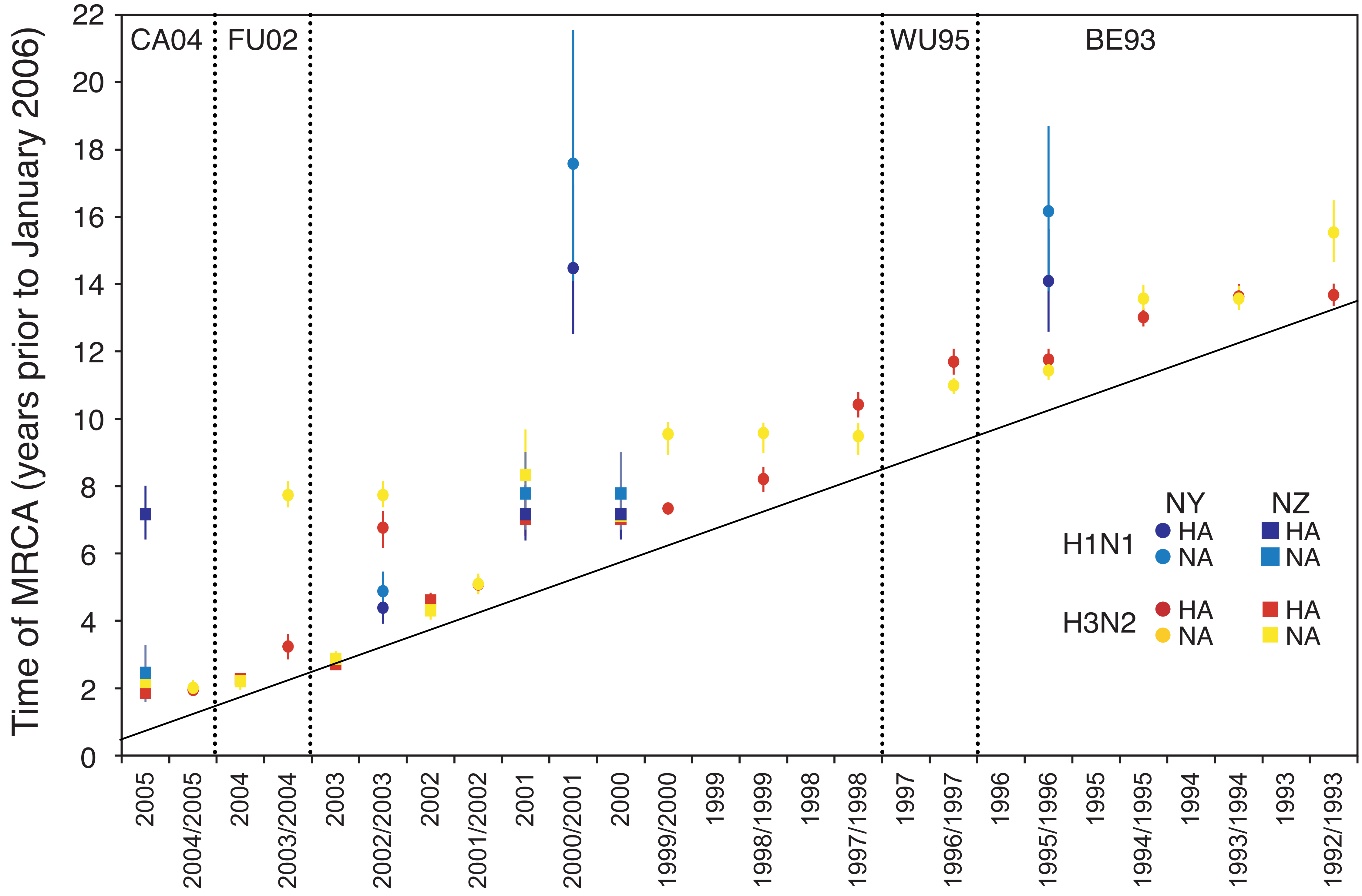
TMRCA of seasons by genomic segment



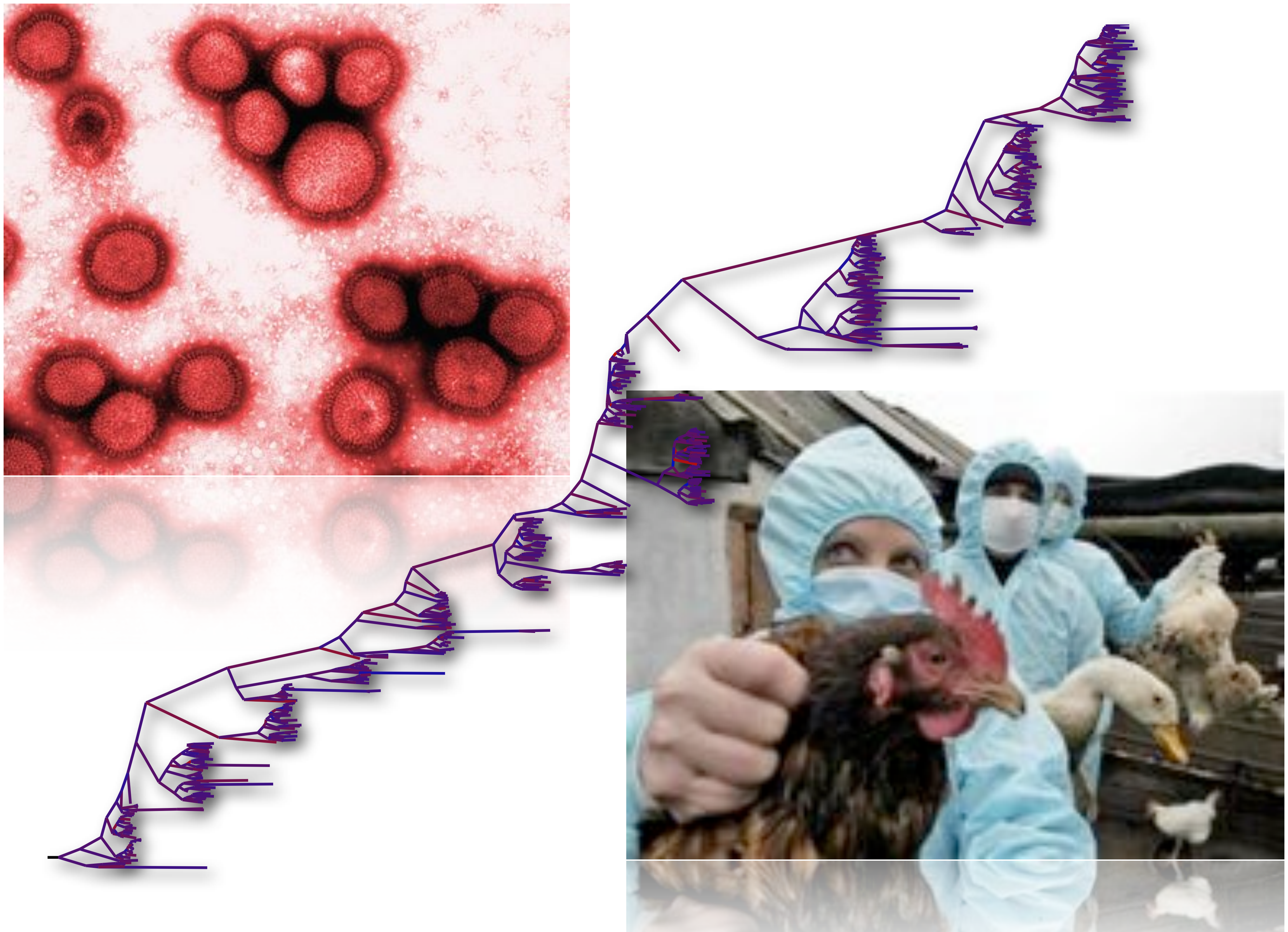
TMRCA of seasons by genomic segment



TMRCA of seasons by genomic segment



Viral phylogeography

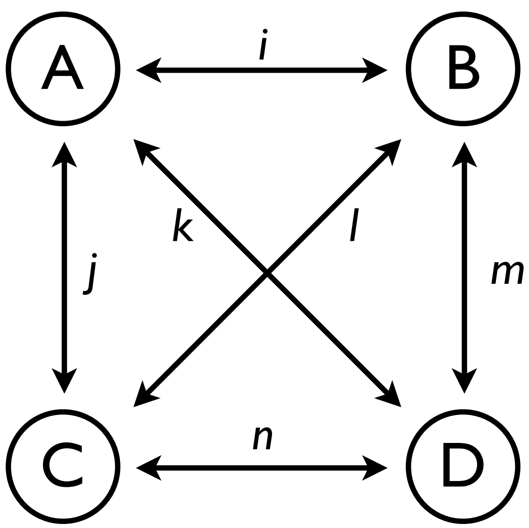


Phylip Lemey, Marc Suchard and Alexei Drummond

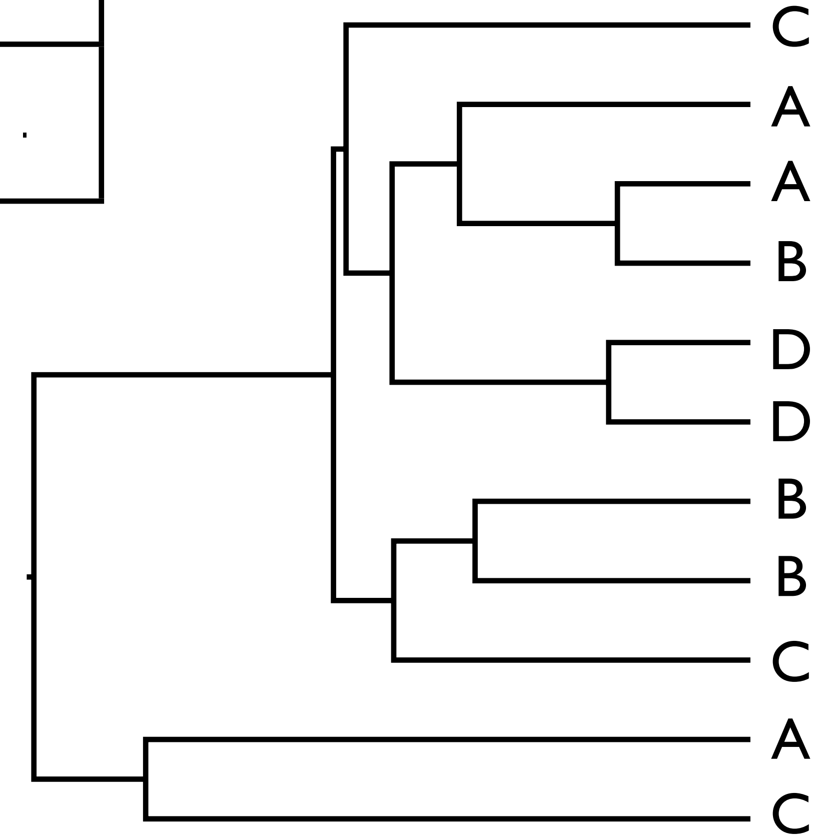
Phylogenetic diffusion models

- Discrete Model:

CTMC



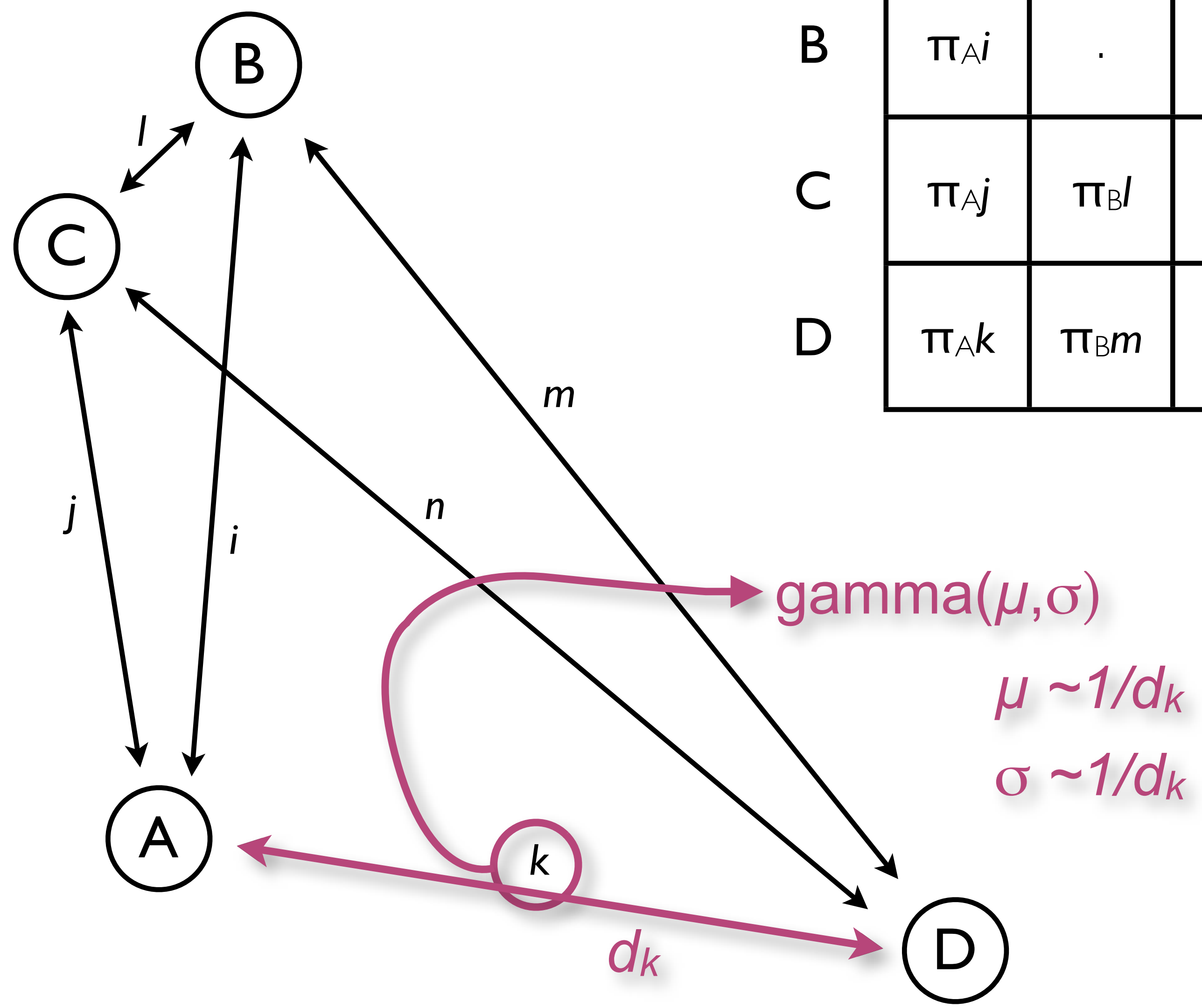
	A	B	C	D
A	.	$\pi_B i$	$\pi_C j$	$\pi_D k$
B	$\pi_A i$	.	$\pi_C l$	$\pi_D m$
C	$\pi_A j$	$\pi_B l$	.	$\pi_D n$
D	$\pi_A k$	$\pi_B m$	$\pi_C n$	.

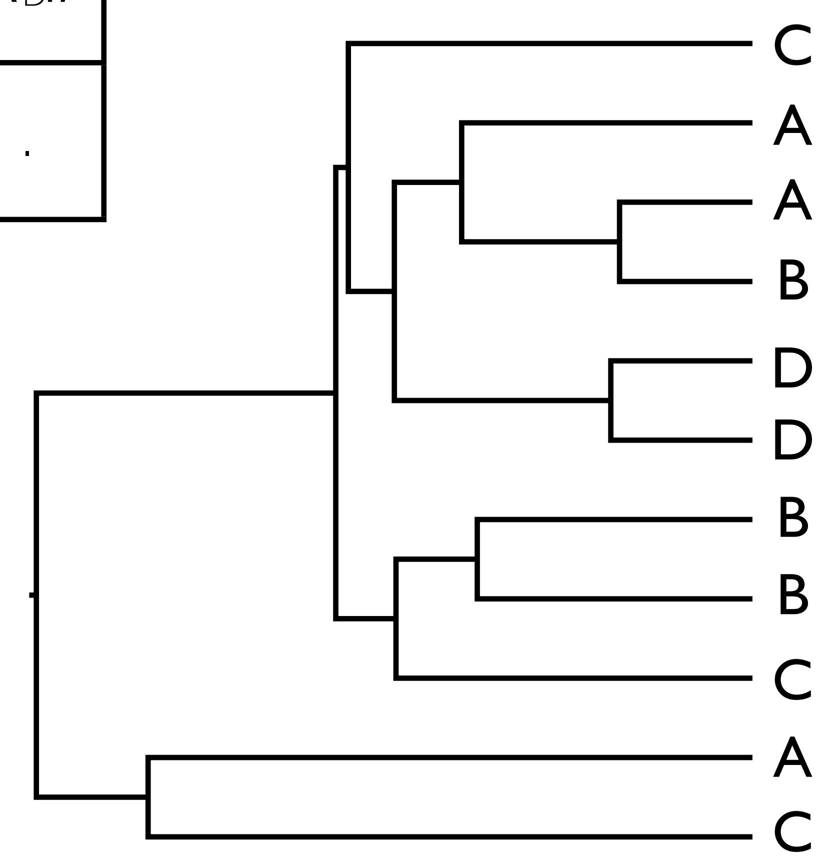


Phylogenetic diffusion models

- Discrete Model:

	A	B	C	D
A	.	$\pi_B i$	$\pi_C j$	$\pi_D k$
B	$\pi_A i$	.	$\pi_C l$	$\pi_D m$
C	$\pi_A j$	$\pi_B l$	.	$\pi_D n$
D	$\pi_A k$	$\pi_B m$	$\pi_C n$	.





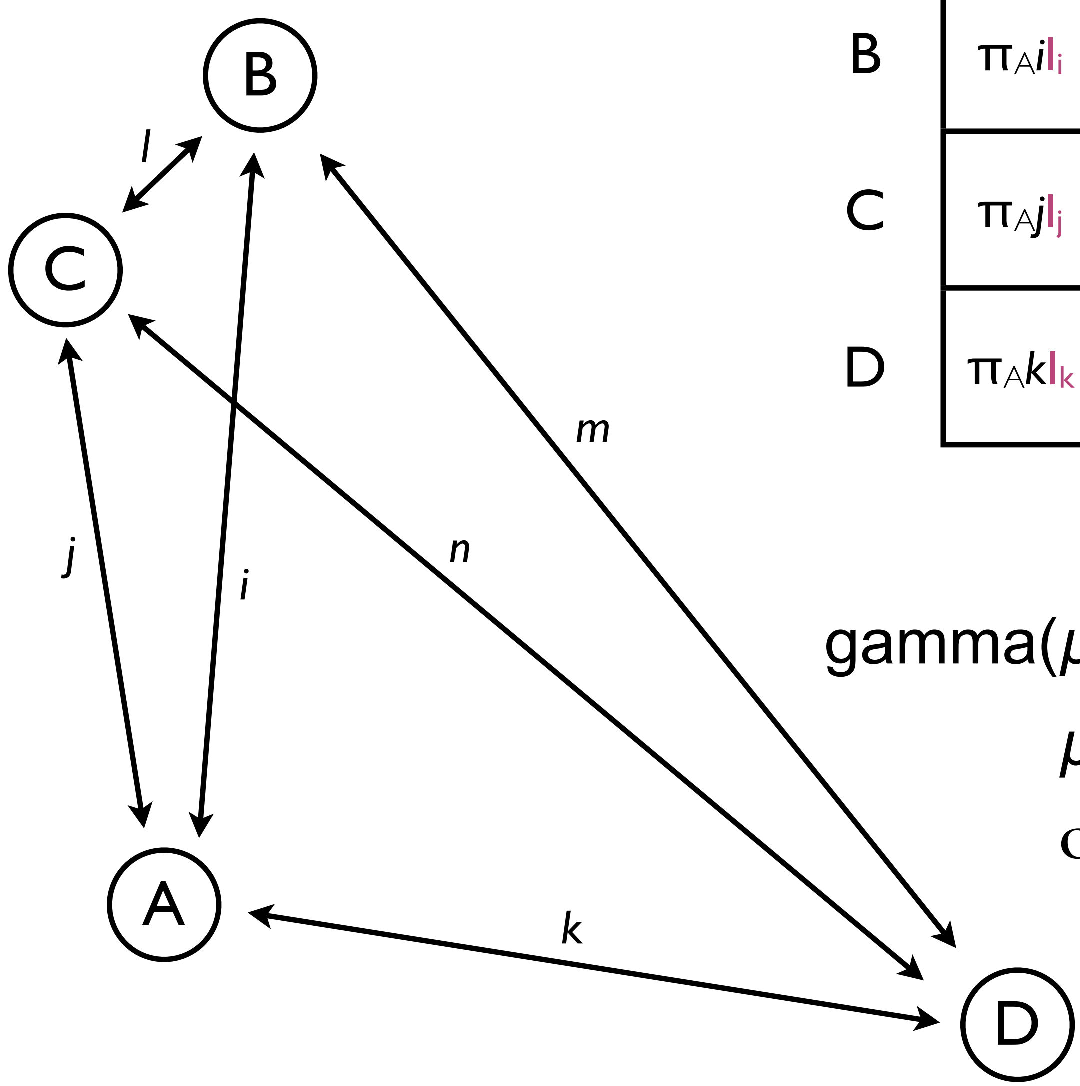
Phylogenetic diffusion models

- Discrete Model:

	A	B	C	D
A	.	$\pi_B i I_i$	$\pi_C j I_j$	$\pi_D k I_k$
B	$\pi_A i I_i$	.	$\pi_C l I_l$	$\pi_D m I_m$
C	$\pi_A j I_j$	$\pi_B l I_l$	.	$\pi_D n I_n$
D	$\pi_A k I_k$	$\pi_B m I_m$	$\pi_C n$	.

Rate Indicators $I_{[0,1]}$

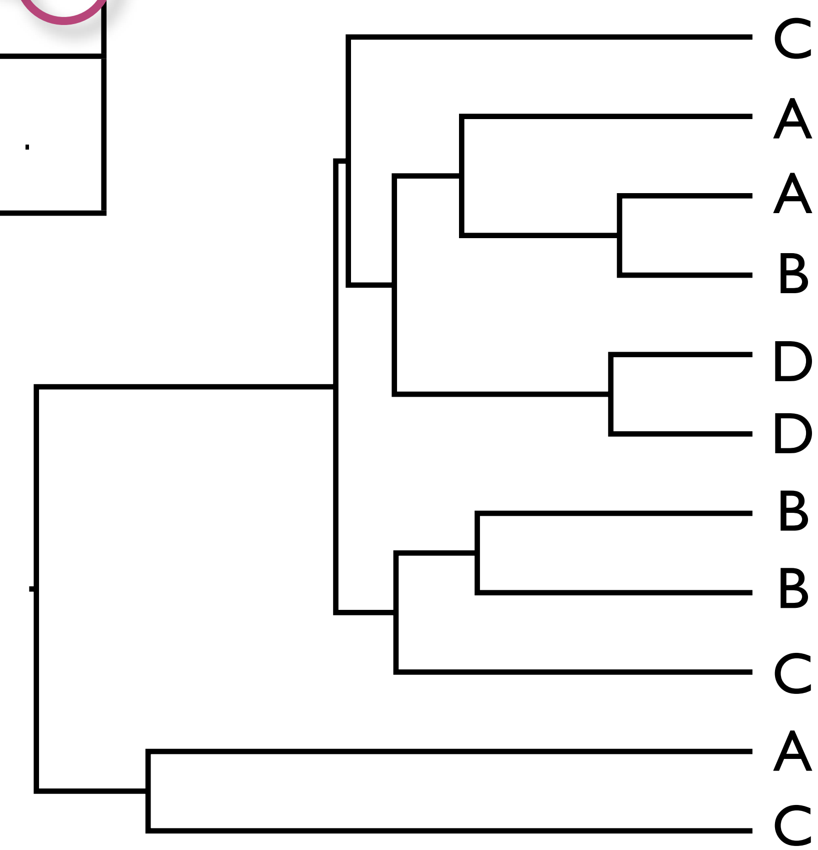
Mixture model



gamma(μ,σ)

$\mu \sim 1/d_k$

$\sigma \sim 1/d_k$



Phylogenetic diffusion models

- Continuous Model:

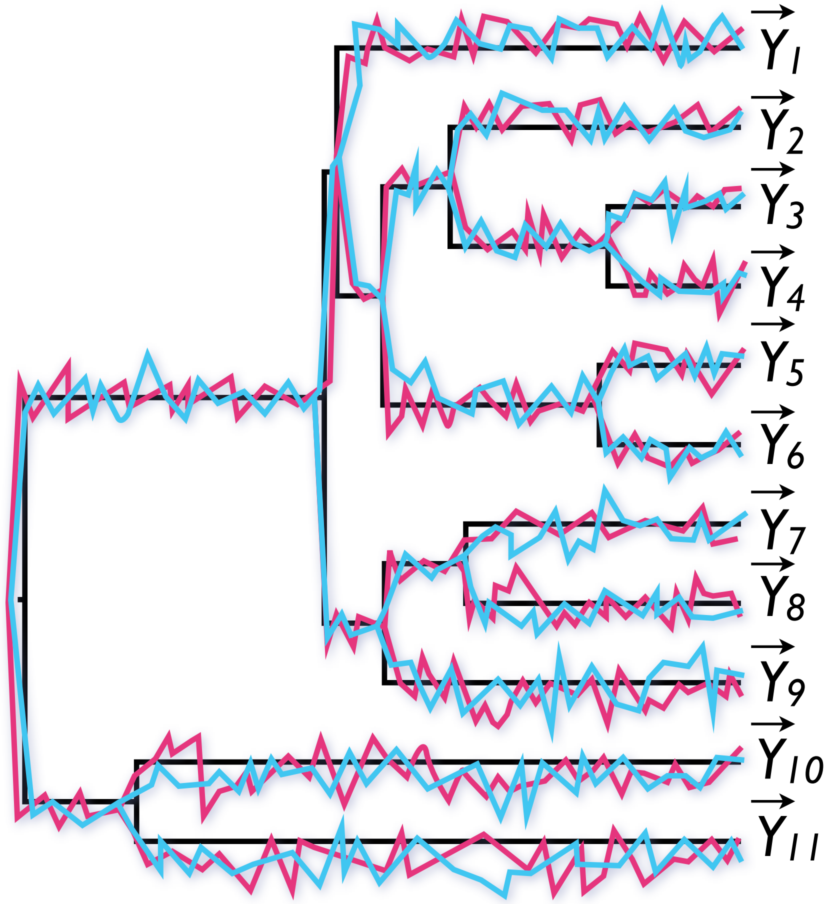
$$\vec{Y} = (Y_{lat}, Y_{long})$$

$$\vec{Y}_{tip} = (\vec{Y}_1, \ldots, \vec{Y}_N)$$

$$\vec{Y}_{internal} = (\vec{Y}_{N+1}, \ldots, \vec{Y}_{2N-2})$$

$$\vec{Y}_{root}$$

$$\Delta\vec{Y} = BVN(0, \vec{S}t)$$

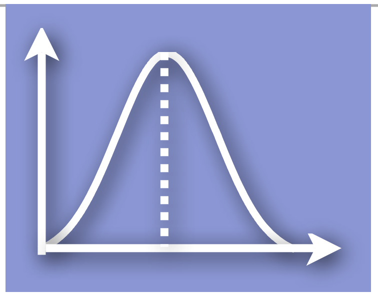


Implementation

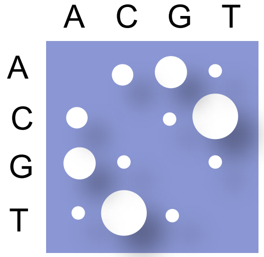
- Given sequence data that is temporally spaced estimate true values of:
  - substitution parameters (μ and Q)
  - ancestral genealogy ($g = E_g, t_Y$)

    tree topology

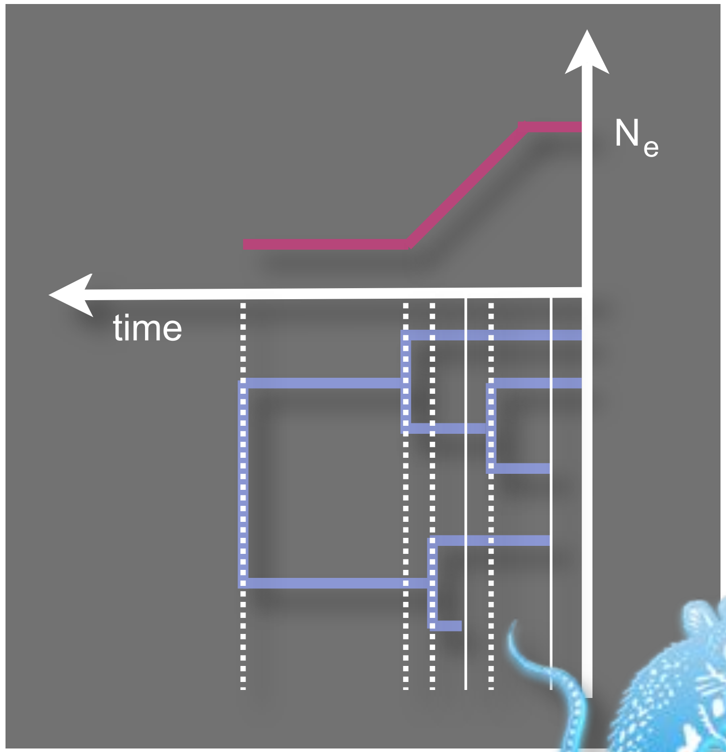
    dates of divergence
  - population history (θ)



μ



Q

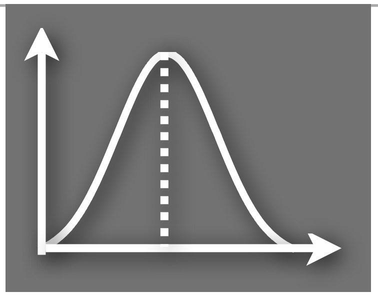


- Bayesian inference

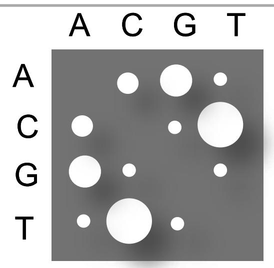
$$P(g,\mu,\theta,Q|D) = \frac{1}{Z}\Pr\{D|g,\mu,Q\} f_g(g|\theta) f_\mu(\mu) f_\theta(\theta) f_Q(Q)$$

Implementation

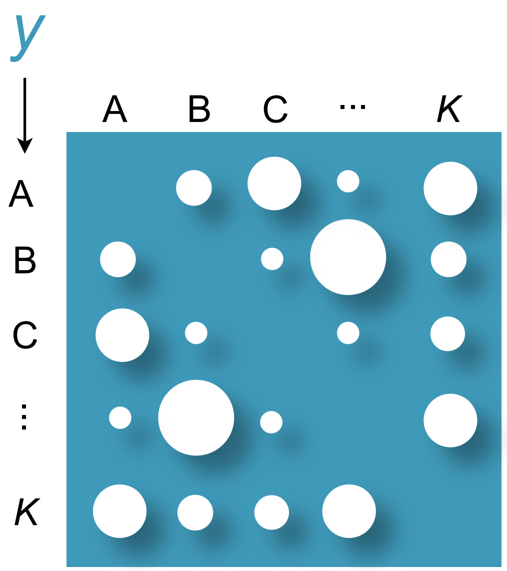
- Given sequence data *(D_X)* that is temporally *and geographically (D_Y)* spaced estimate true values of:
  - substitution parameters (μ and Q_y)
  - ancestral genealogy (g = E_g, t_Y)
    - tree topology
    - dates of divergence
  - population history (θ)
  - discrete location states



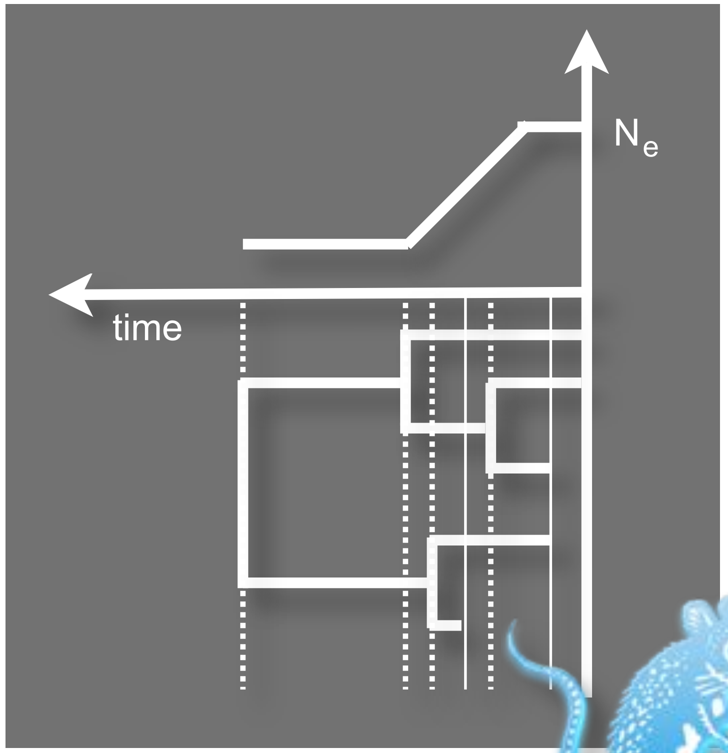
μ



Q_x



Q_y



- Bayesian inference

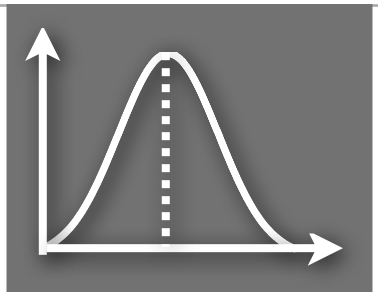
$$P(g,\mu,\theta,Q_x,Q_y|D_x,D_y) = \frac{1}{Z} \Pr\{D_x|g,\mu,Q\}\Pr\{D_y|g,Q_y\} f_g(g|\theta) f_\mu(\mu) f_\theta(\theta) f_Q(Q_x) f_Q(Q_y)$$

Implementation

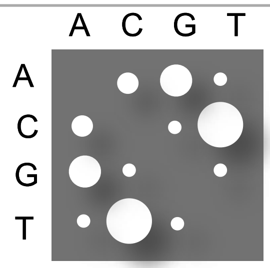
- Given sequence data *(DX)* that is temporally *and geographically (DY)* spaced estimate true values of:
  - substitution parameters (μ and Q_y)
  - ancestral genealogy (g = E_g, t_Y)

    tree topology

    dates of divergence
  - population history (θ)
  - continuous location traits

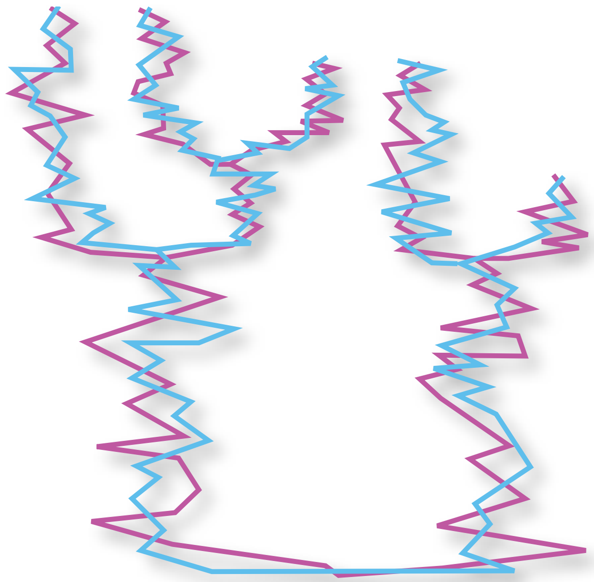


μ

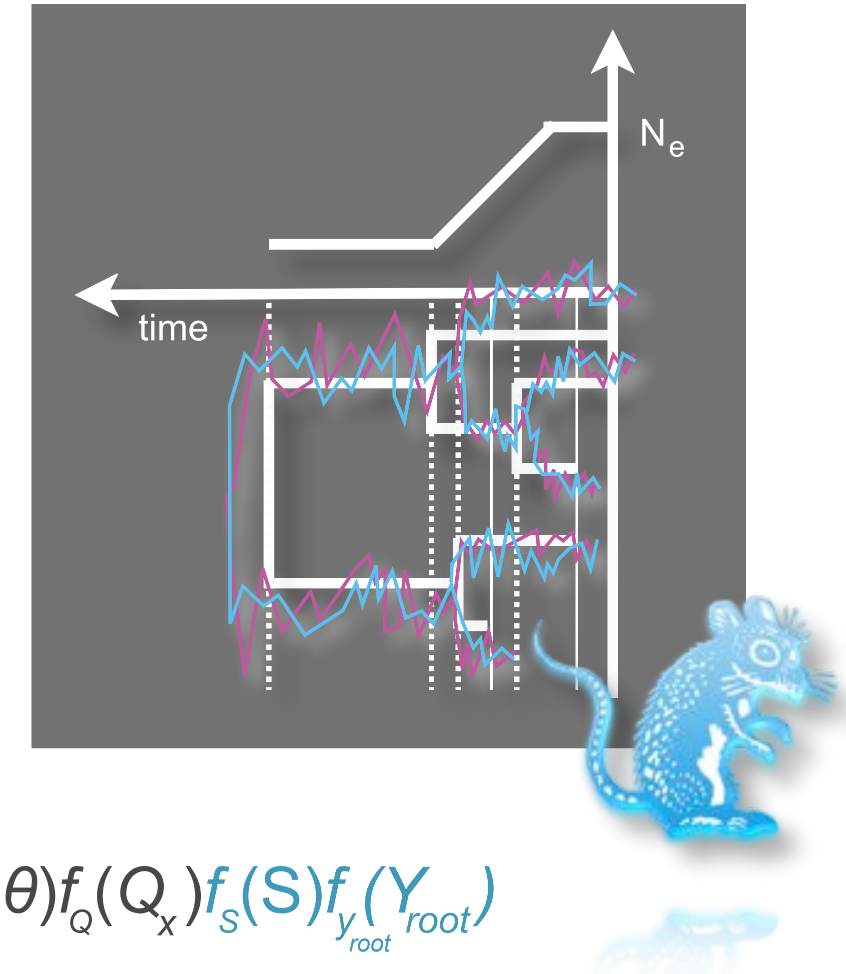


Q_x

$\vec{y} = \{lat, long\}$



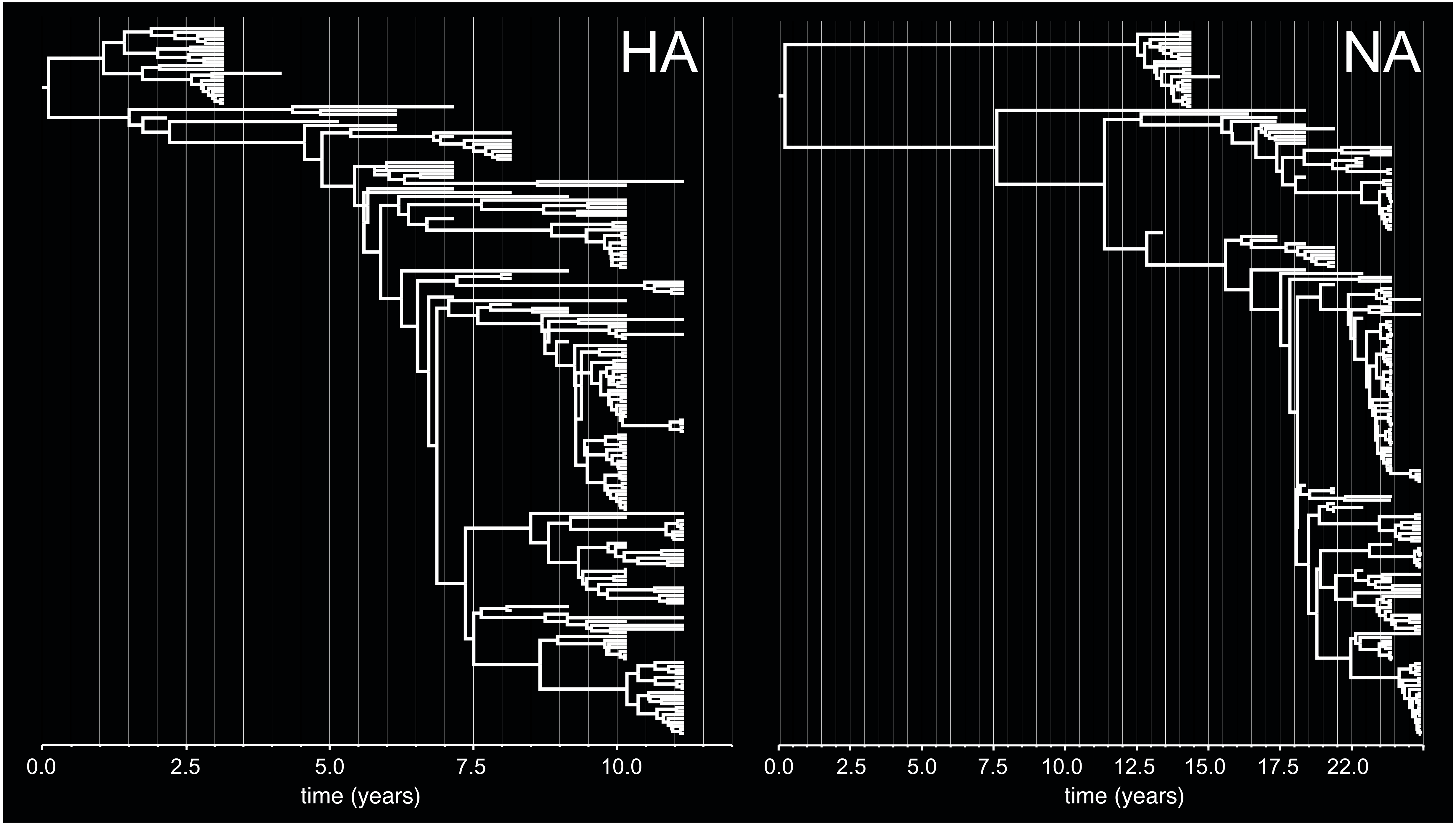
$\Delta\vec{Y} = BVN(0,\vec{S}t)$



- Bayesian inference

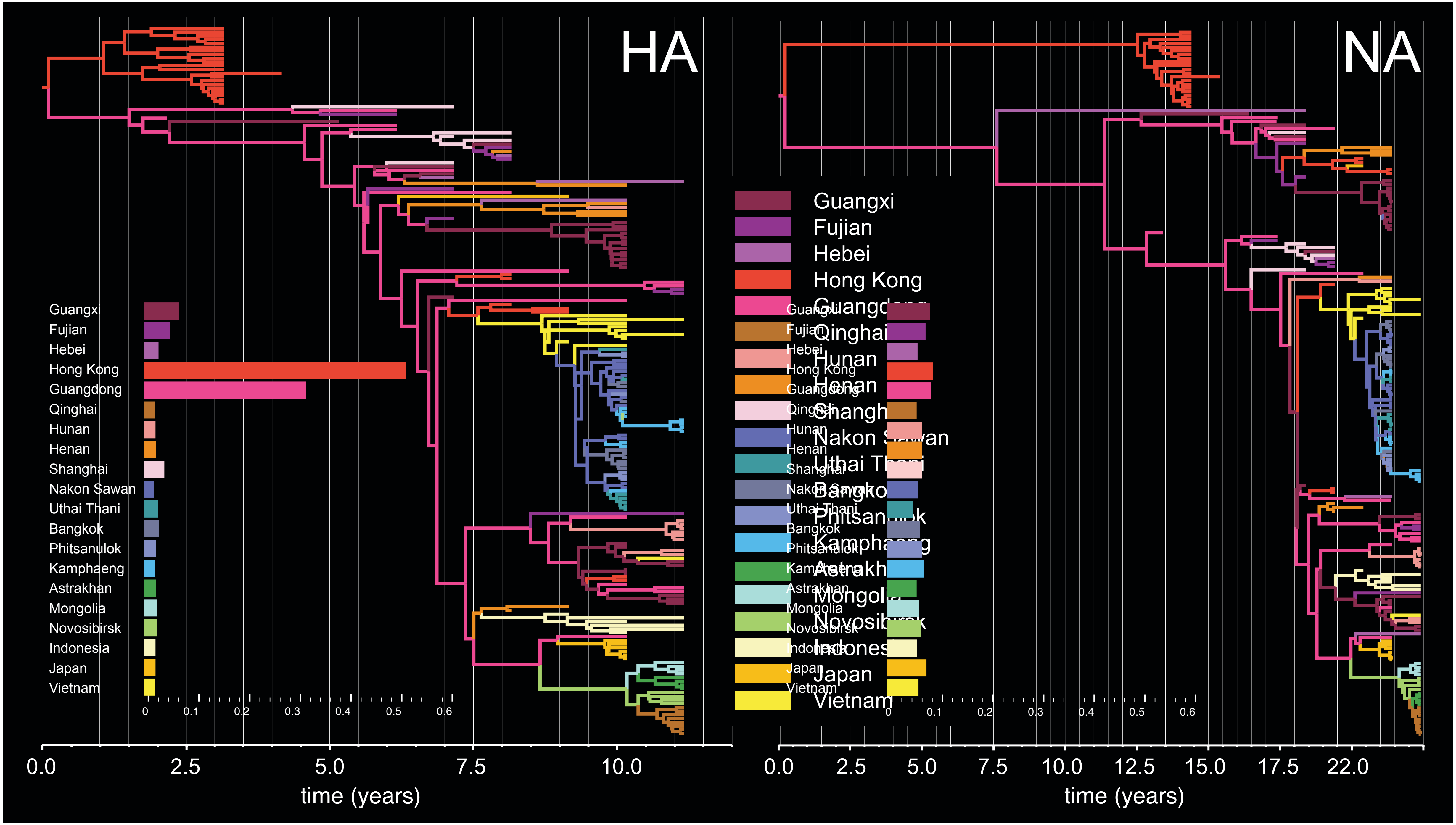
$$P(g,\mu,\theta,Q_x,S|D_x,D_y)= \frac{1}{Z} \Pr\{D_x|g,\mu,Q_x\}\Pr\{D_y|g,S\}f_g(g|\theta)f_\mu(\mu)f_\theta(\theta)f_Q(Q_x)f_S(S)f_{y_{root}}(Y_{root})$$

Influenza A H5N1

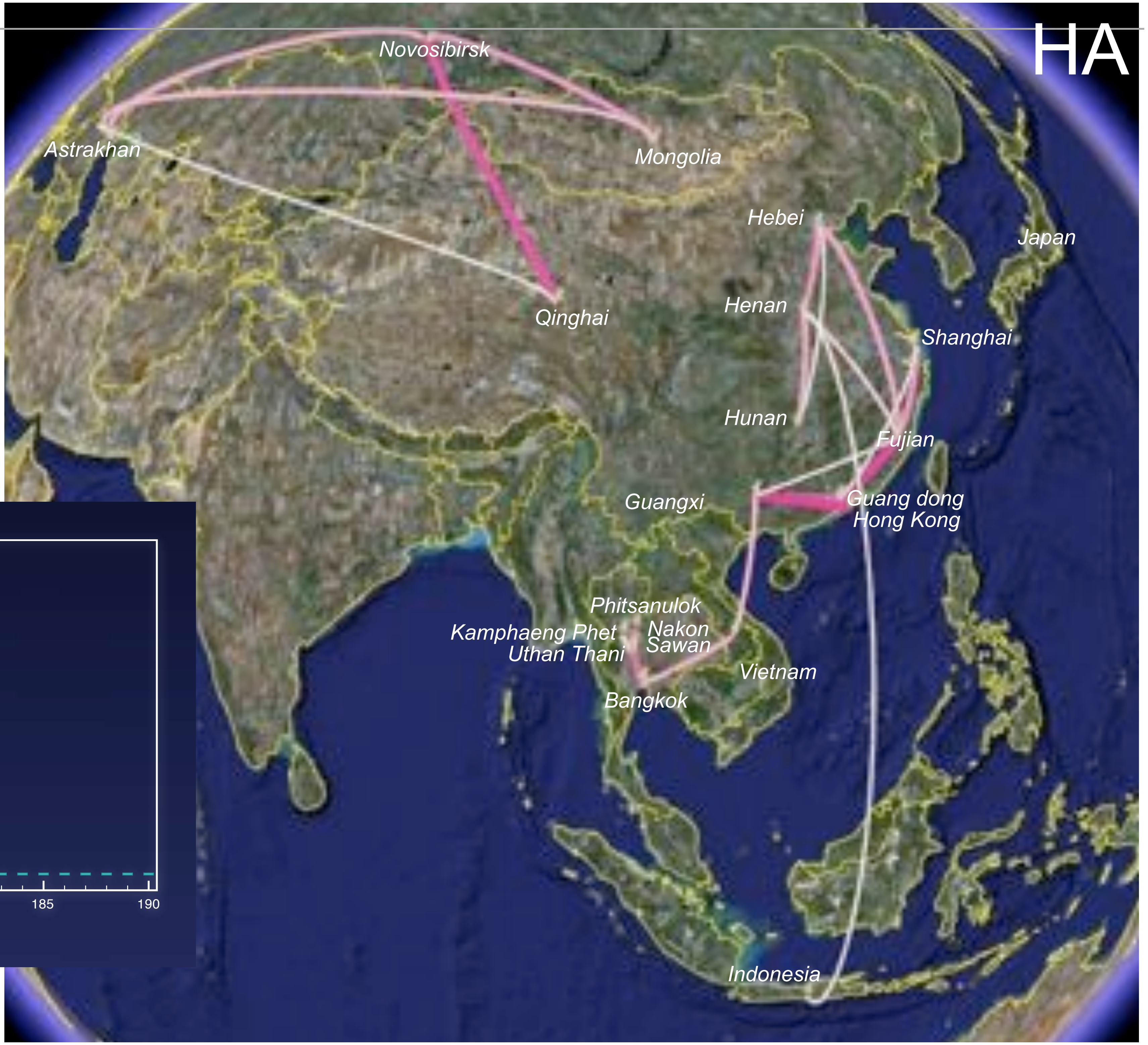


Wallace et al., PNAS, 2007

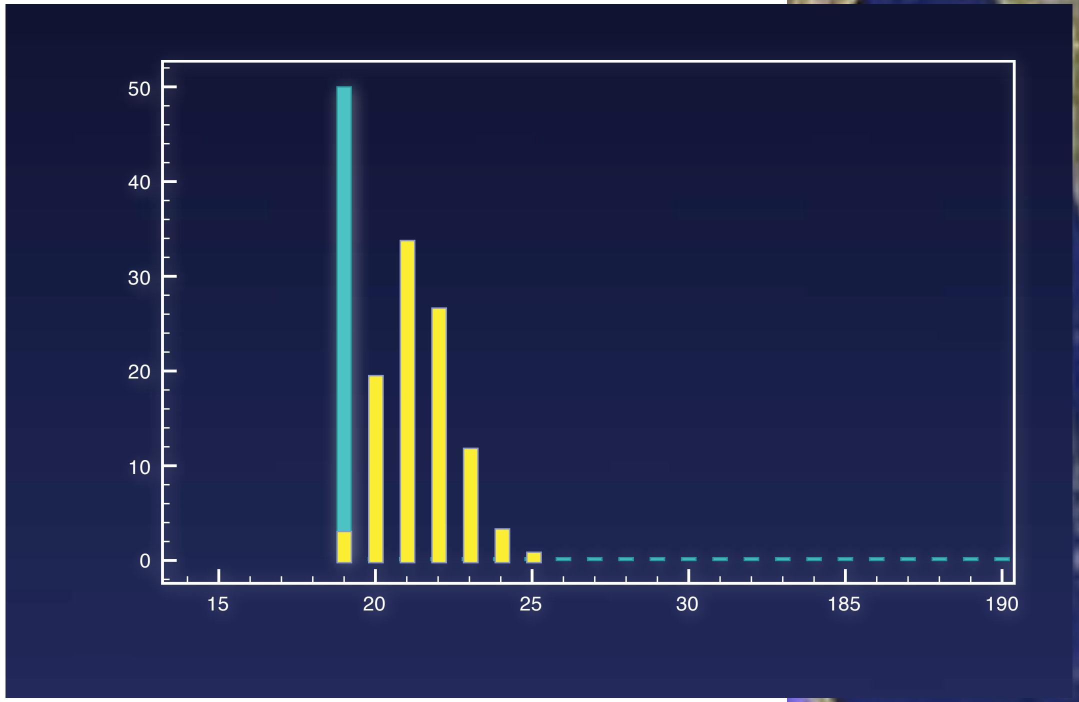
Influenza A H5N1: discrete model



Influenza A H5N1: discrete mixture model

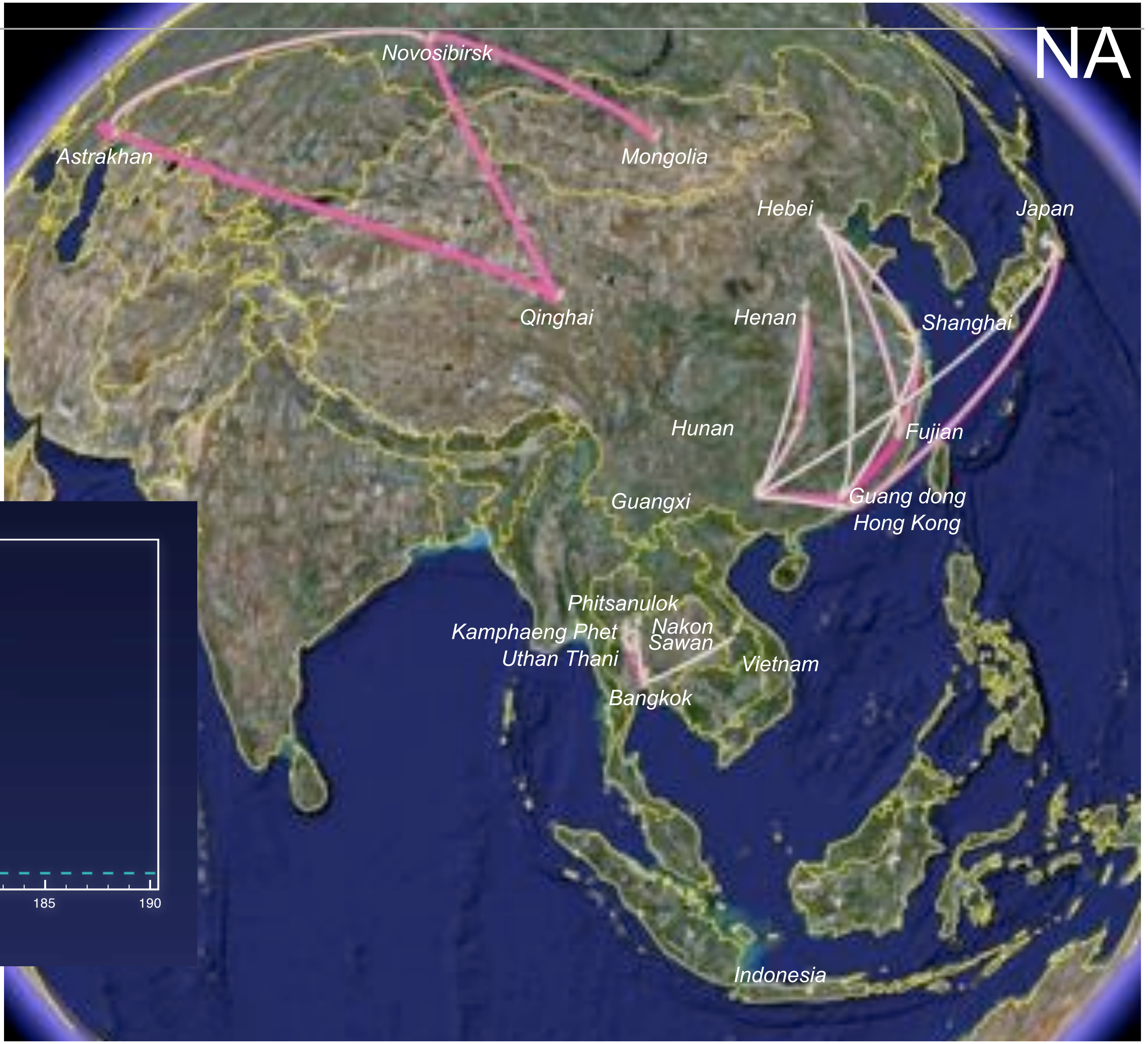


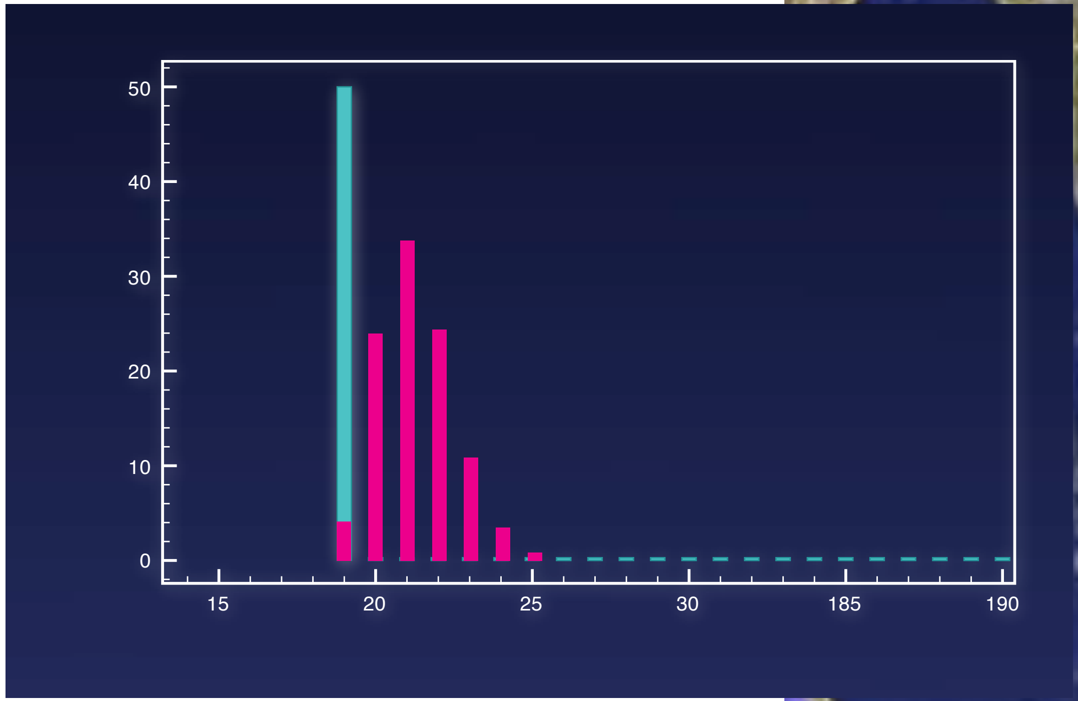
- A Bayes factor test for significant rates



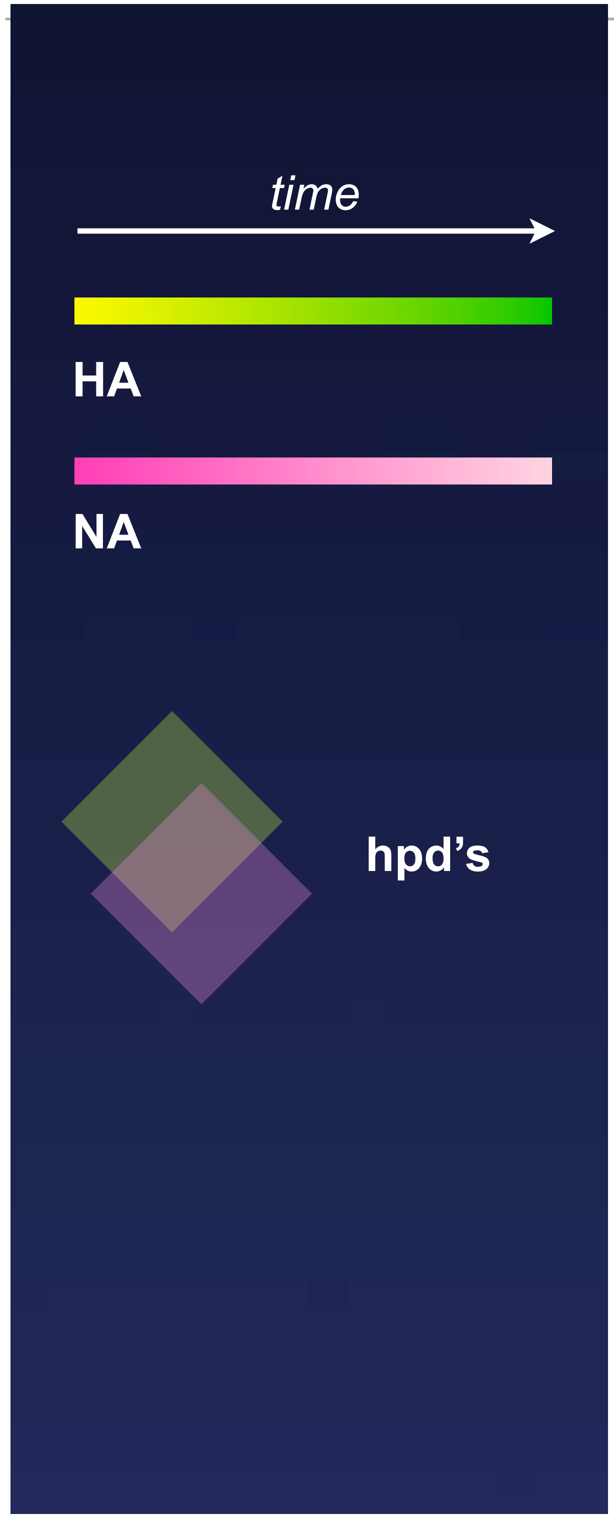
Influenza A H5N1: discrete mixture model

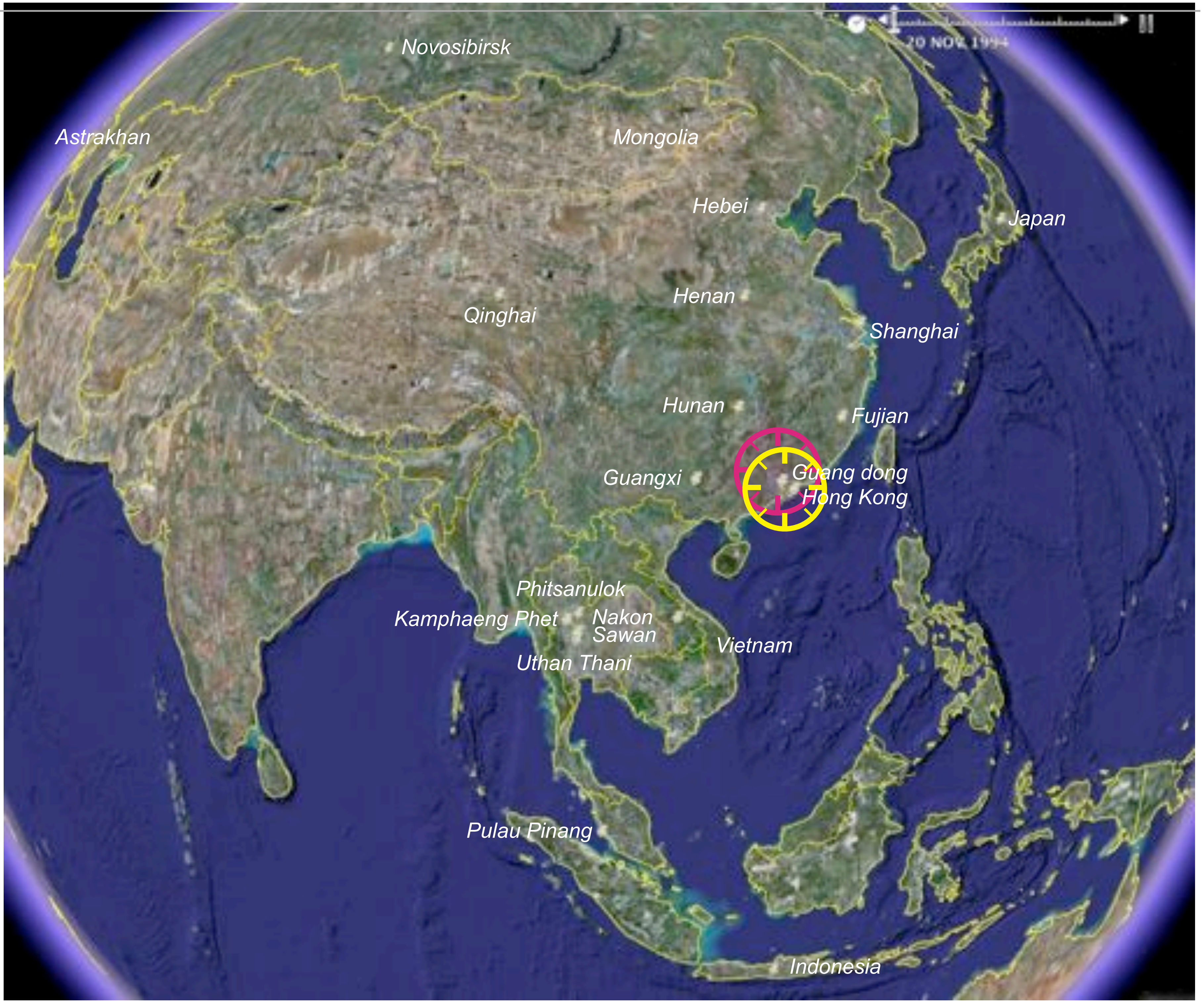
- A Bayes factor test for significant rates



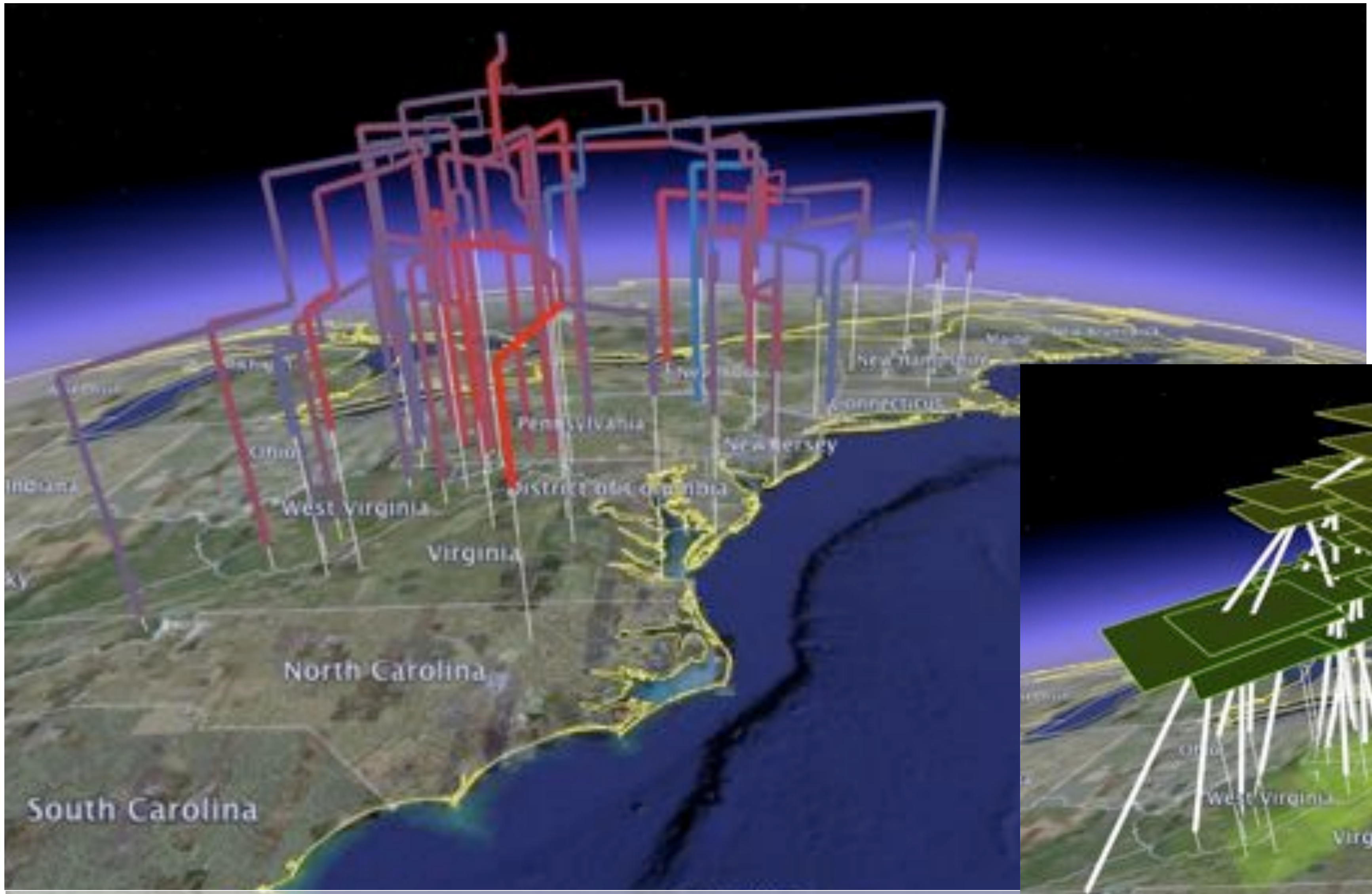


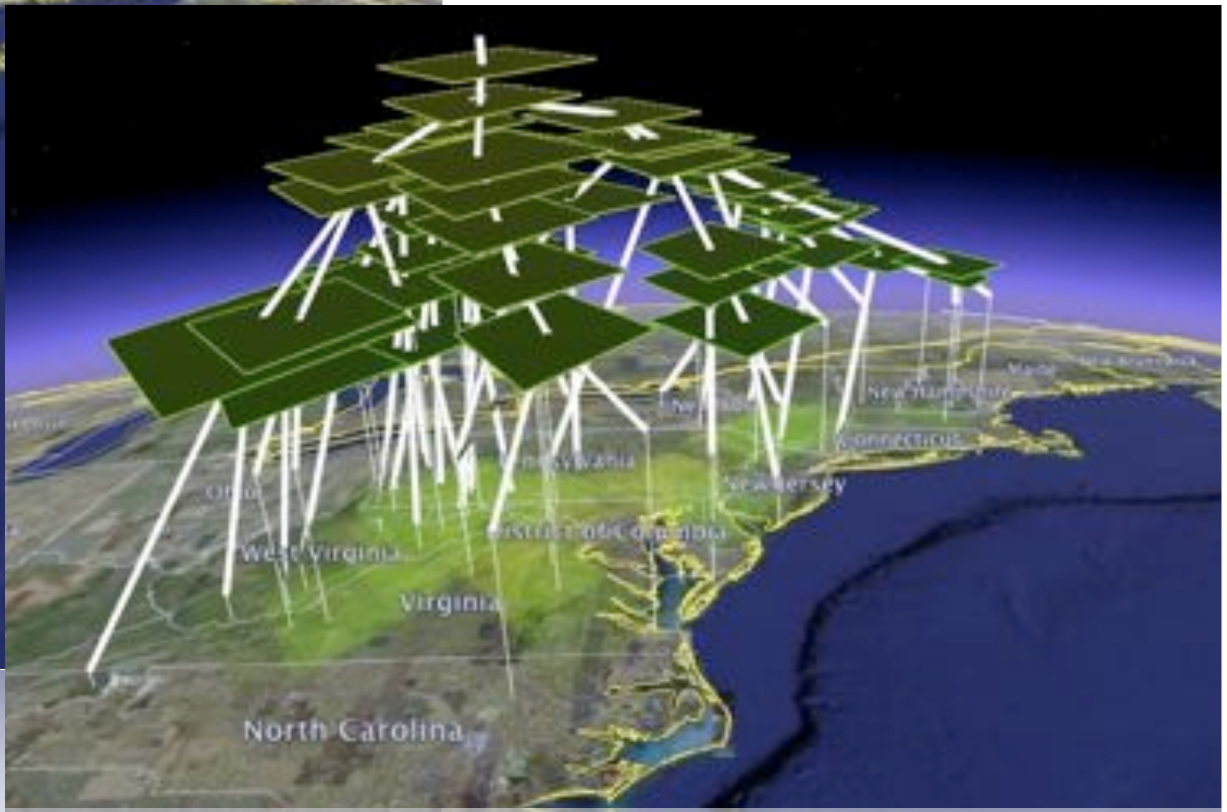
Influenza A H5N1: continuous model





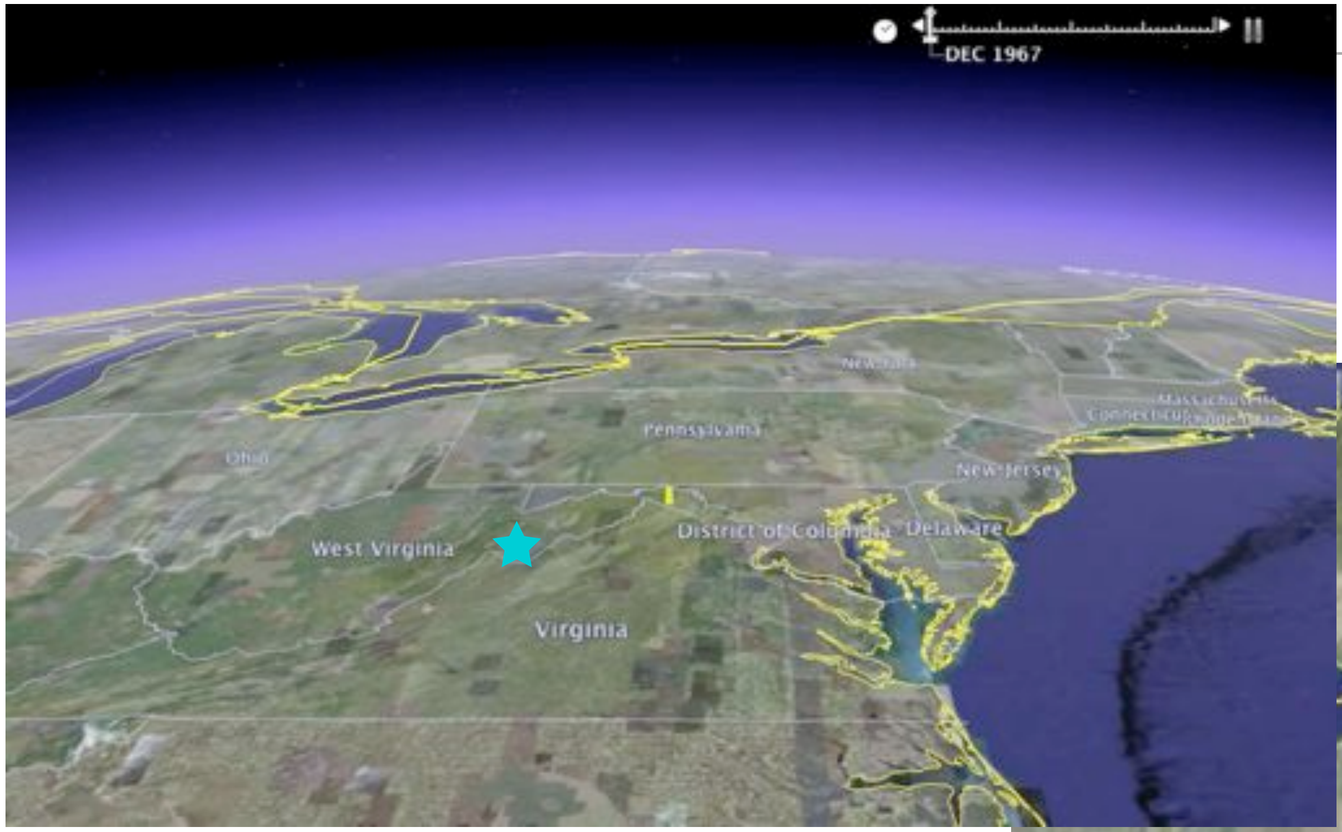
Raccoon dog Rabies: continuous model

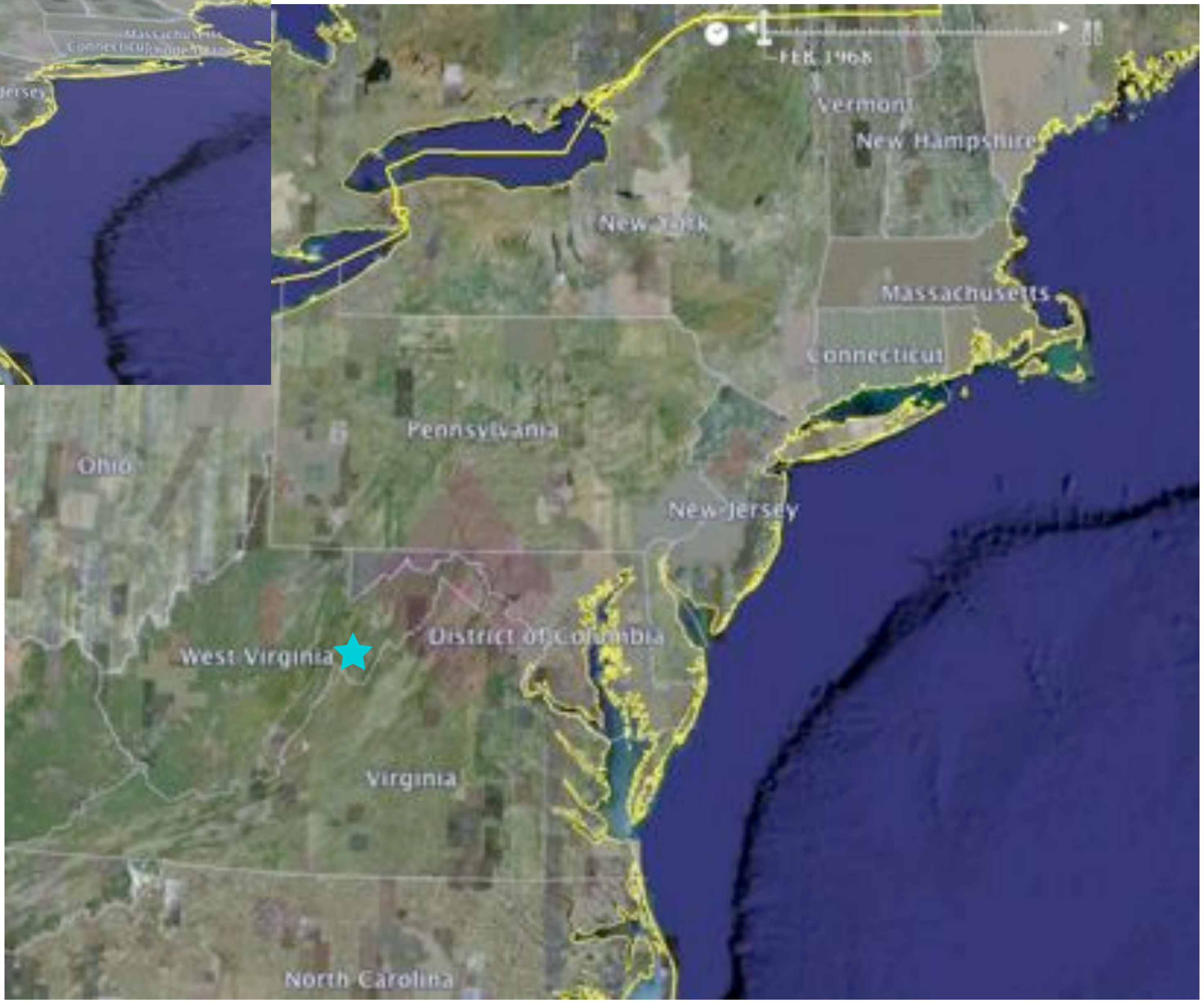




Biek et al., PNAS, 2007

Raccoon dog Rabies: continuous model





★ First reported case,

Pembleton County, WV, 1977

the future of RNA virus evolutionary dynamics

- currently:

  nearly 4000 influenza genomes sequenced

  over 1500 HIV genomes sequenced

- analyses of the type described here take CPU months with an upper limit of about 1000 sequences

the future of RNA virus evolutionary dynamics

- Roche 454 Genome sequencer
- sequences 20 million nucleotides per 5 hour run
- equivalent of about 1000 RNA virus genomes per run
- potential for studying complete viral genome diversity in a patient over time on a fine scale
- new computational techniques will be required to study the complex interactions between virus and host

BEAST

Bayesian Evolutionary Analysis Sampling Trees

- Can be used for estimating:
  - phylogenies
  - rates of evolution, dates of divergence
  - demographic history
- Averages estimates over plausible trees
  - http://beast.bio.ed.ac.uk

Acknowledgments

- University of Auckland:
  Alexei Drummond
- Pennsylvania State:
  Eddie Holmes
- Katholieke Universiteit, Leuven:
  Philippe Lemey
- University of Oxford:
  Oliver Pybus
- UCLA:
  Marc Suchard