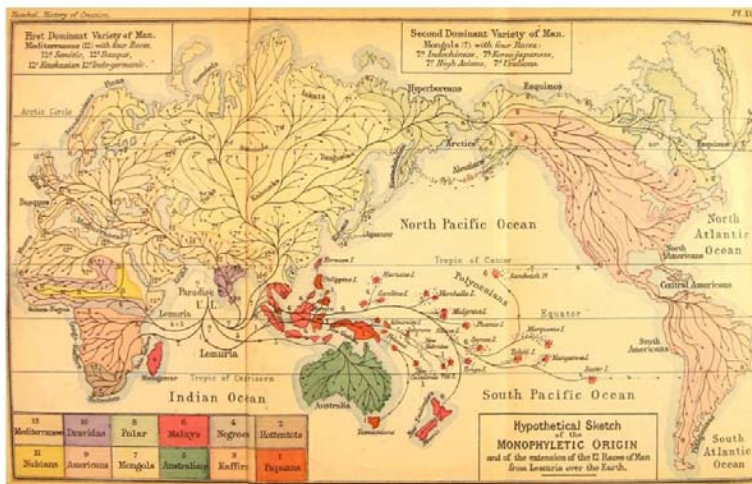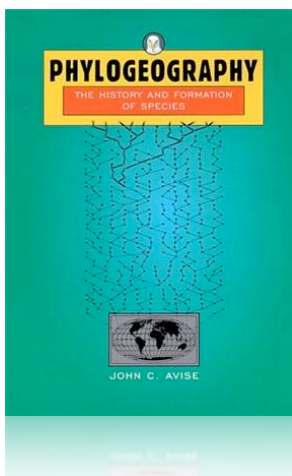# phylogeography

Philippe Lemey & Marc Suchard

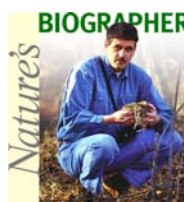# Lemuria and Darwin's origin of species

---

Phylogeography

"a field of study concerned with the principles and processes governing the geographic distribution of genealogical lineages, especially those within and among closely related species."
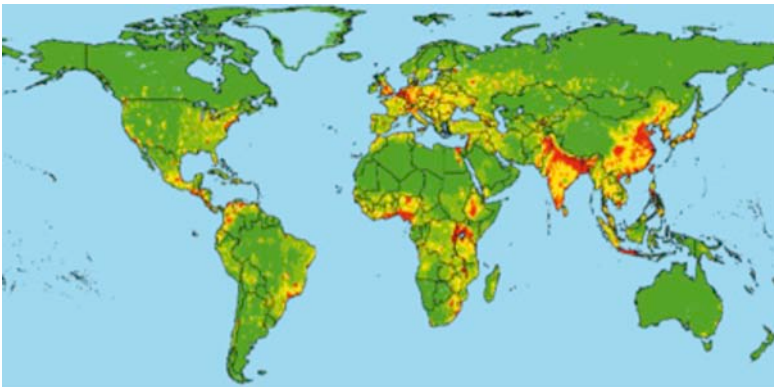
Avise, 2000

Phylogeography

Why bother?

Phylogeographic methods provide a means of examining the history of genetic exchange among populations, with the potential to distinguish biogeographic patterns of genetic variation caused by gene flow from those caused by common ancestry.

Epidemiology: emerging diseases go global
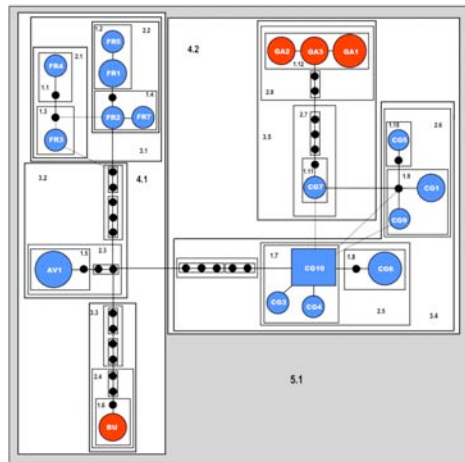

Jones et al. 2008. Nature 451: 990-994

phyloseminar

Phylogeographic inference

"Three roads diverged?"
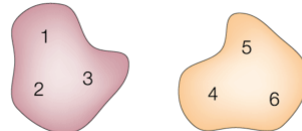
Phylogeographic inference (road I)

Nested clade phylogeographic analysis (NCPA)

Nested Clade Phylogeographic Analysis

1. make a haplotype tree
2. test whether there is a relationship between 'nested clades' and geography
3. Use inference key to find out the most likely explanation:
   - vicariance
   - isolation by distance
   - range expansion

Sample of locations from two populations

Sequence data

| 1 = ACTG | 3 = GTCG | 5 = GTCC |
| 2 = ATTG | 4 = GTCG | 6 = GTTG |

Haplotype network

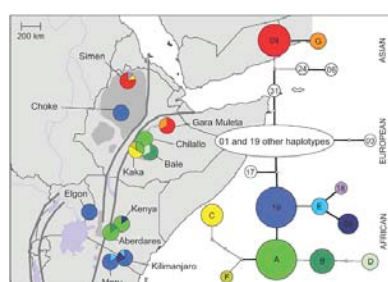*Inference key: http://darwin.uvigo.es/download/geodisKey_28Apr09.pdf*

Software: http://darwin.uvigo.es/software/geodis.html

Nested Clade Phylogeographic Analysis

Pro NCPA:

◉ first attempt to provide an exploratory method of phylogeography

Nested Clade Phylogeographic Analysis

Con NCPA:

◉ Lack of statistical rigor: "…NCPA does not assess error in its inferences about historical processes or contemporary gene flow…" (Knowles and Maddison, 2002)

◉ One study showed incorrect historical processes were inferred by NCPA in over 75% of the simulated datasets (Panchal and Beaumont 2007)

◉ Better alternative: Model-based inference in phylogeography (Nielsen and Beaumont, 2009; Beaumont et al 2010)

NCPA Pro and Con

The controversy has been battled out in the literature…

COMMENTARY

NEWS AND VIEWS

WHY DOES A METHOD THA
TO BE USED? THE ANSWER

REPLY

In defence of model-based inference in phylogeography

Alan R. Templeton
Department of Biology, Campus Box 1137, Washington University, St. L
E-mail: temple_a@wustl.edu

MARK A. BEAUMONT,[1] RASMUS NIELSEN,[2] CHRISTIAN ROBERT,[3] JODY HEY,[4] OSCAR GAGGIOTTI,[5] LACEY KNOWLES,[6] ARNAUD ESTOUP,[7] MAHESH PANCHAL,[8] JUKKA CORANDER,[9] MIKE HICKERSON,[10] SCOTT A. SISSON,[11] NELSON FAGUNDES,[12] LOUNÈS CHIKHI,[13] PETER BEERLI,[14] RENAUD VITALIS,[15] JEAN-MARIE CORNUET,[7] JOHN HUELSENBECK,[2] MATTHIEU FOLL,[16,17] ZIHENG YANG,[18] FRANCOIS ROUSSET,[19] DAVID BALDING[20] and LAURENT EXCOFFIER[16,17]

Coherent and incoherent
and human evolution

Alan R. Templeton[1]
Department of Biology, Washington University, St. Louis, MO 63130

Phylogeographic inference (road II)

Coalescent theory:

◉ is a statistical framework for the analysis of genetic polymorphism data

◉ is an extension of classical population-genetics theory and models

◉ one can estimate time (number of generations) for lineages to coalesce

◉ many applications (including migration analysis)

Discrete Generations
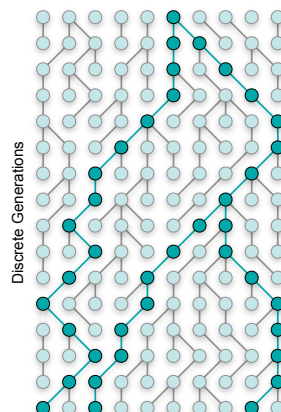
Phylogeographic inference (road II)

LAMARC - Likelihood Analysis with
Metropolis Algorithm using Random Coalescence

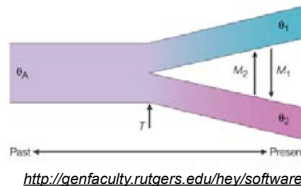*http://evolution.genetics.washington.edu/lamarc.html*

MIGRATE-N
estimation of population sizes
and gene flow using the coalescent

http://popgen.sc.fsu.edu/Migrate-n.html

http://genfaculty.rutgers.edu/hey/software

MDIV: http://www.biom.cornell.edu/Homepages/Rasmus_Nielsen/files.htm

Batwing: http://www.maths.abdn.ac.uk/~ijw/downloads/download.htm

phyloseminar

Viral phylogeography
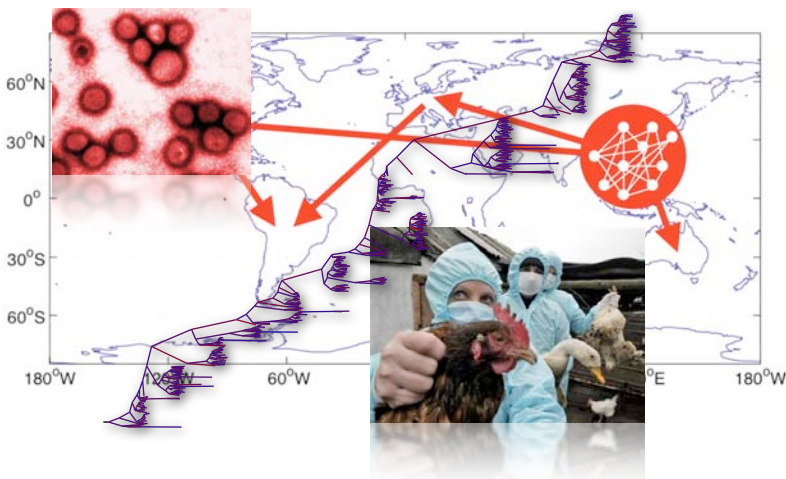
Phylogeographic patterns in RNA viruses

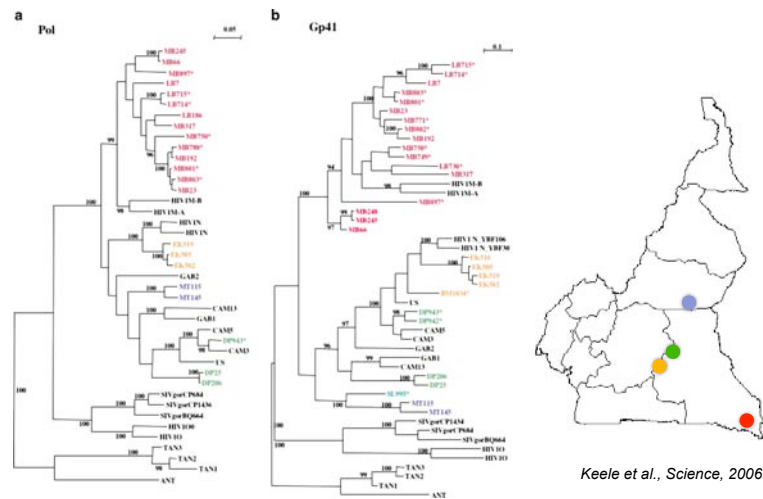a No spatial structure   b Wave-like transmission   c Sink-source transmission

d Gravity model transmission   e Population subdivision
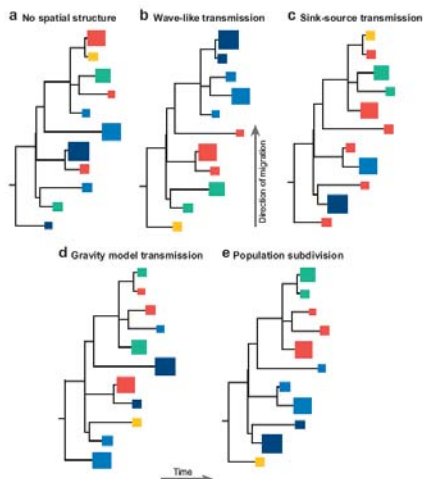
Understanding the evolutionary history of human viruses, along with the factors that have shaped their spatial distributions, is one of the most active areas of study in the field of microbial evolution.
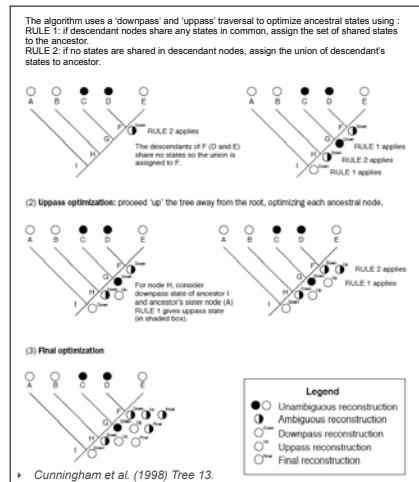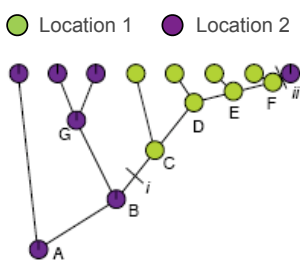
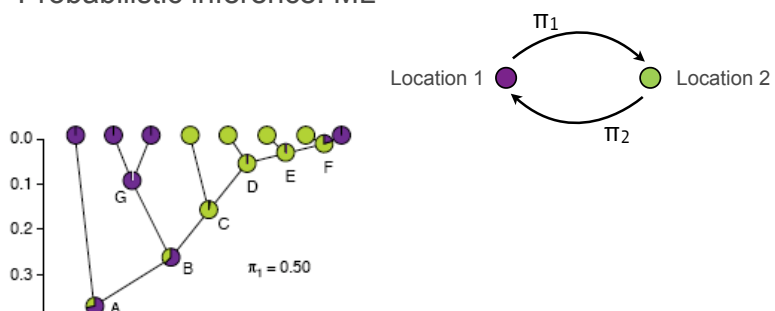**Eddie Holmes, Ann Rev Microbiol 2008**

Viral phylogeography

Keele et al., Science, 2006.

Inferring ancestral state locations

Parsimony analysis

Location 1 (green) Location 2 (purple)

The algorithm uses a 'downpass' and 'uppass' traversal to optimize ancestral states using:
RULE 1: if descendant nodes share any states in common, assign the set of shared states to the ancestor.
RULE 2: if no states are shared in descendant nodes, assign the union of descendant's states to ancestor.

Cunningham et al. (1998) Tree 13.

Inferring ancestral state locations

Probabilistic inference: ML

Location 1 ⟶ Location 2, π_1, π_2

$\pi_1 = 0.50$

Inferring ancestral state locations

Probabilistic inference: Bayesian inference
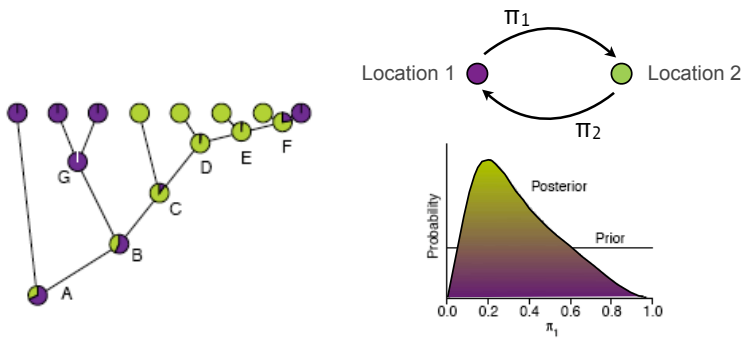
Location 1 π_1 Location 2

π_2

Probability — Posterior / Prior

0.0 0.2 0.4 0.6 0.8 1.0
π_1

A B C D E F G

Phylogenetic diffusion models

◉ Discrete Model:

CTMC

A ⟷ B
j, k, l, m, n
C ⟷ D

	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	$\pi_A k$	$\pi_B m$	$\pi_C n$	.

C
A
A
B
D
D
B
B
C
A
C
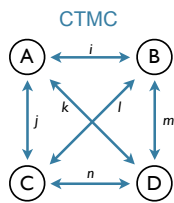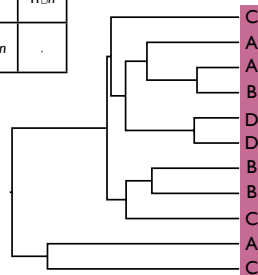
Phylogenetic diffusion models

◉ priors

	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	$\pi_A k$	$\pi_B m$	$\pi_C n$	.

B, C, A nodes with edges l, m, j, i, n, k

gamma(μ,σ)
$\mu \sim 1/d_k$
$\sigma \sim 1/d_k$

d_k

C
A
A
B
D
D
B
B
C
A
C
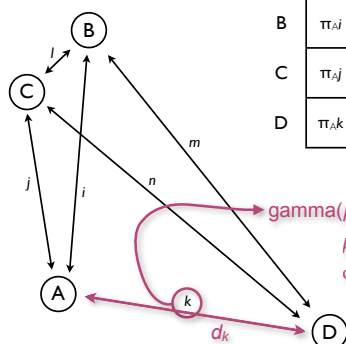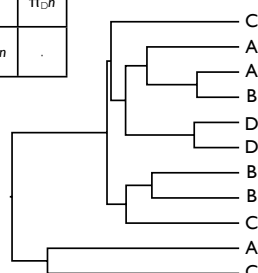
Bayesian stochastic search variable selection

◉ linear regression:

$X_1,...,X_p$ — potential predictors

$Y = [X_1,...,X_P]\,\beta + \varepsilon$ — N-dimensional outcome variable Y

β_p for $p = 1,...,P \neq 0$ $\longrightarrow$ X_P helps predicting Y
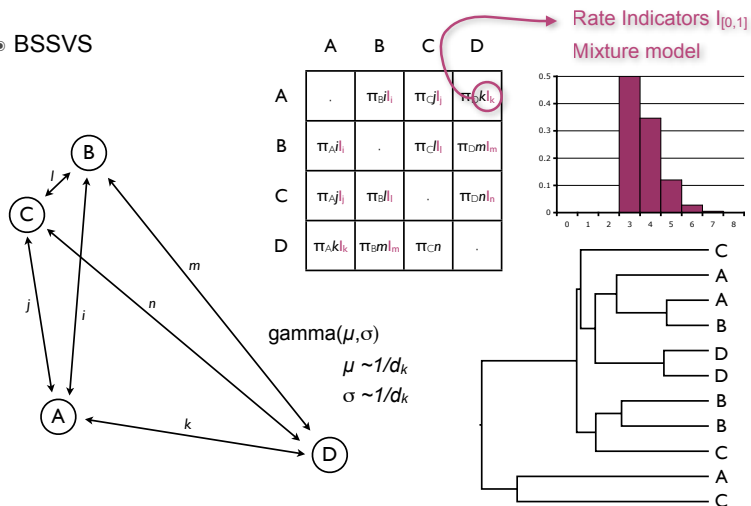
Optimal set of predictors?

◉ BSSVS:

1. augment model-state space with 0,1-indicators $(\delta_1,...,\delta_P)$

   impose prior $p(\beta)$ with mean = 0, var $\propto PxP$ matrix
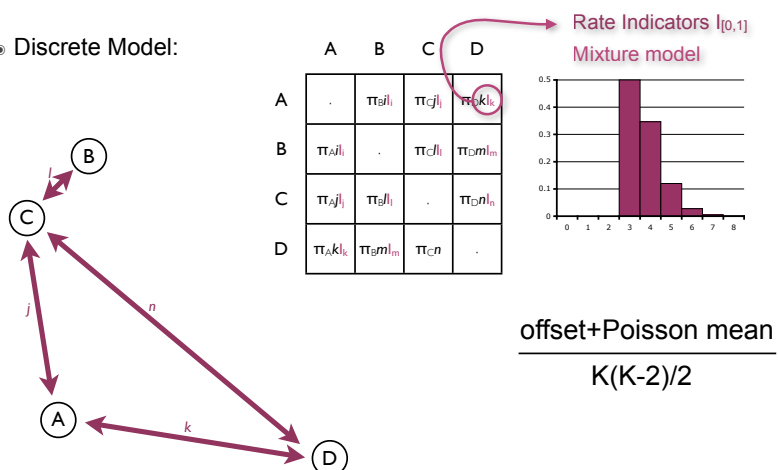
2. Explore joint state space (δ,β) using MCMC

Phylogenetic diffusion models

◉ BSSVS

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

gamma(μ,σ)
$\mu \sim 1/d_k$
$\sigma \sim 1/d_k$

Phylogenetic diffusion models

◉ Discrete Model:

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

$$\frac{\text{offset+Poisson mean}}{K(K-2)/2}$$

Implementation

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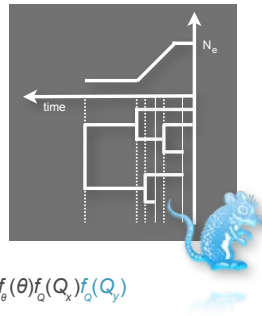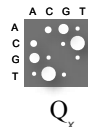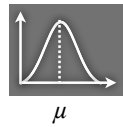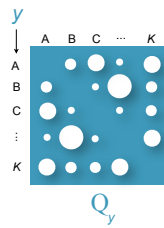   - substitution parameters (μ and Q_y)
    - ancestral genealogy ($g = E_g, t_Y$)
        - tree topology
        - dates of divergence
    - population history (θ)
    - discrete location states

μ Q_x

y

A B C $\cdots$ κ

A
B
C
$\vdots$
κ

Q_y

N_e

time

- Bayesian inference

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

Bayesian phylogeographic reconstruction

◉ Advantages:

✓ a discretized diffusion process that we can fit simultaneously with standard models for sequence evolution

✓ incorporates branch length differences

✓ handles rapid and unequal rates of state exchange

✓ accommodates uncertainty in the diffusion process

✓ accommodates uncertainty in the unknown phylogeny

✓ no need to fix Markov model to obtain ancestral states

✓ geography-based priors and BSSVS

✓ time-measured diffusion and demography co-estimation

H5N1 'bird flu'

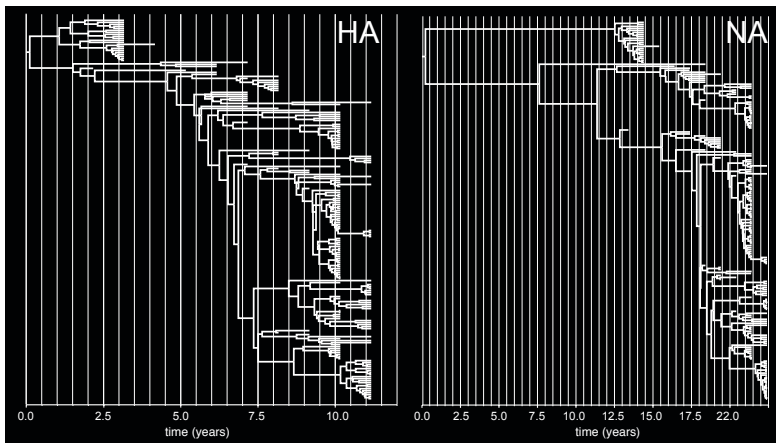◉ Wild fowl act as natural asymptomatic carriers

◉ first HPAI outbreak in Guangdong, China in 1996

◉ 'Bird flu' outbreak in Hong Kong in 1997

◉ the A/goose/Guangdong/ 1/96 (Gs/GD) virus lineage has become the longest recorded HPAI virus to remain endemic in poultry
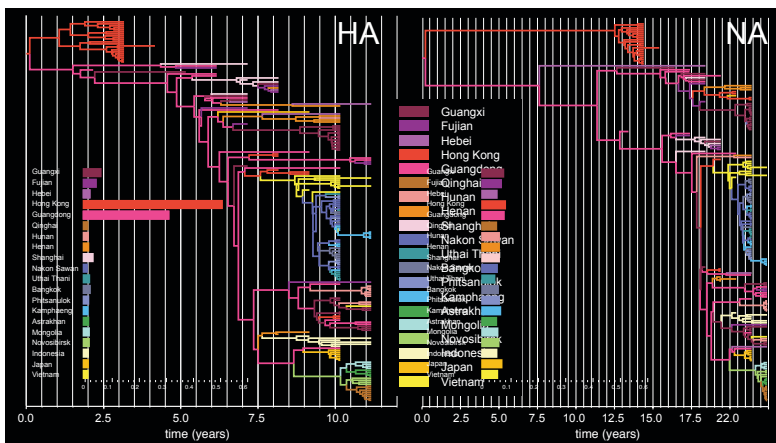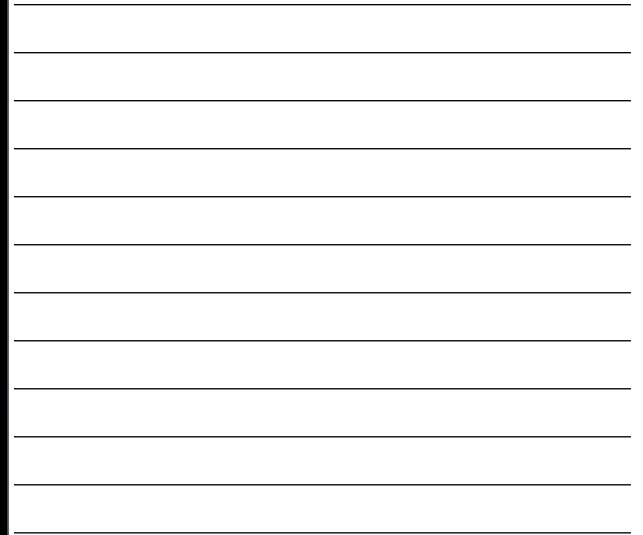
Influenza A H5N1

Wallace et al., PNAS, 2007
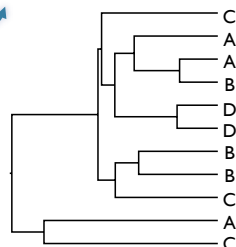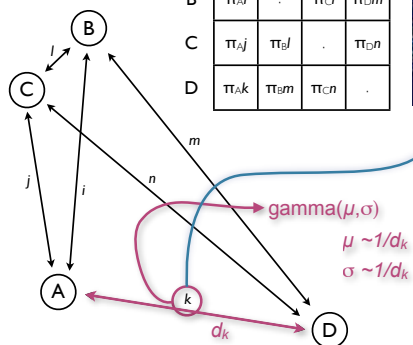
Influenza A H5N1: discrete model

Influenza A H5N1: BSSVS

- A Bayes factor test to identify epidemiological linkage
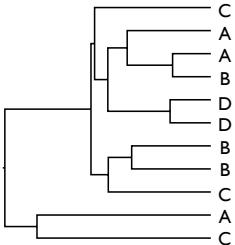
Swine-origin H1N1

Hypothesis testing

◉ 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	$\pi_A k$	$\pi_B m$	$\pi_C n$	.

gamma(μ,σ)
$\mu \sim 1/d_k$
$\sigma \sim 1/d_k$

Hypothesis testing

● GLM approach:

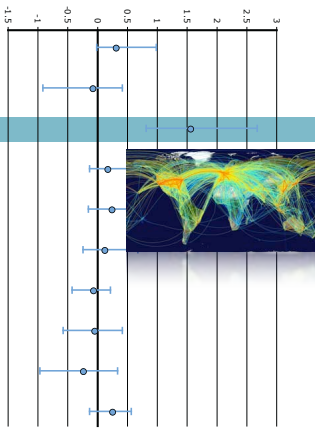|	A	B	C	D
A	.	$\pi_B i$	$\pi_C j$	$\pi_D k$
B	$\pi_A i$	.	$\pi_C l$	$\pi_D m$
C	$\pi_A j$	$\pi_B l$	.	$\pi_D n$
D	$\pi_A k$	$\pi_B m$	$\pi_C n$	.

Potential predictors for rate of diffusion:

$$P_1, P_2, ...$$

$$\log \lambda_{AB} = \delta_{\log P_1} \beta_{\log P_1} \log P_{1AB} +$$
$$\delta_{\log P_2} \beta_{\log P_2} \log P_{2AB} + ... +$$
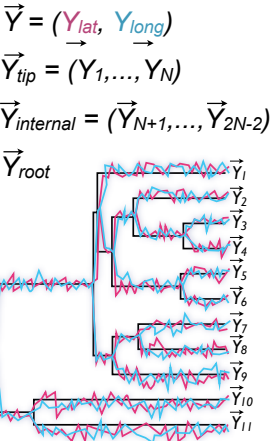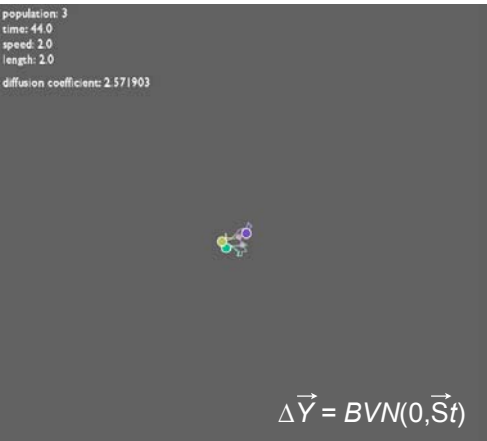$$\delta_{\log P_n} \beta_{\log P_n} \log P_{nAB}$$

Hypothesis testing

	Bayes factor	Coefficient
average distance	0,09	
minimum distance	0,04	
passenger flow	>2699	
donor population size	0,05	
recipient population size	0,04	
donor population density	0,03	
recipient population density	0,02	
H3N2 surveillance data: 8 weeks	0,04	
H3N2 surveillance data: 16 weeks	0,05	
H3N2 surveillance data: 24 weeks	0,12	

Continuous phylogenetic diffusion

population: 3
time: 44.0
speed: 2.0
length: 2.0
diffusion coefficient: 2.571903

$$\vec{Y} = (Y_{lat}, Y_{long})$$
$$\vec{Y}_{tip} = (\vec{Y_1}, ..., \vec{Y_N})$$
$$\vec{Y}_{internal} = (\vec{Y}_{N+1}, ..., \vec{Y}_{2N-2})$$
$$\vec{Y}_{root}$$

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

Implementation

- Given sequence data *(D_X)* that is temporally *and geographically (D_Y)* spaced estimate true values of:

  ‣ substitution parameters (μ and Q_y)

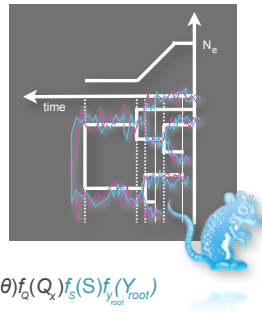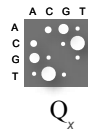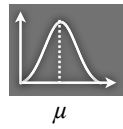 ‣ ancestral genealogy (g = E_g, t_v)

    tree topology

    dates of divergence $\vec{y} = \{lat, long\}$

  ‣ population history (θ)

  ‣ continuous location traits

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

μ Q_x
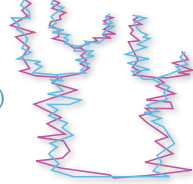
- Bayesian inference

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

A random rabies walk through time and space
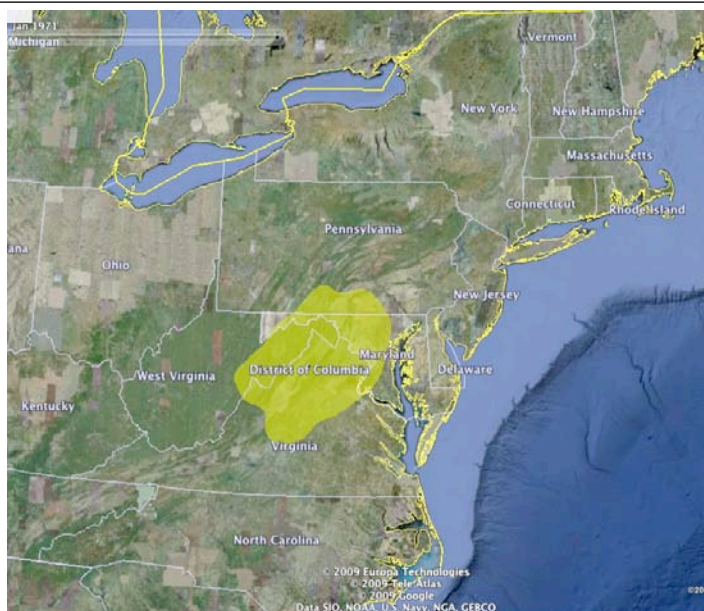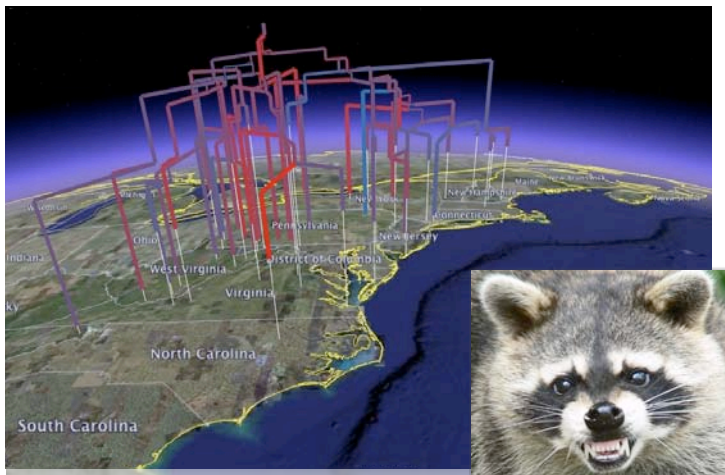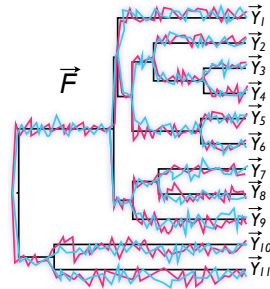

A relaxed random walk

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

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

$\quad \vec{Y}_{internal} = (\vec{Y}_{N+1}, ..., \vec{Y}_{2N-2})$
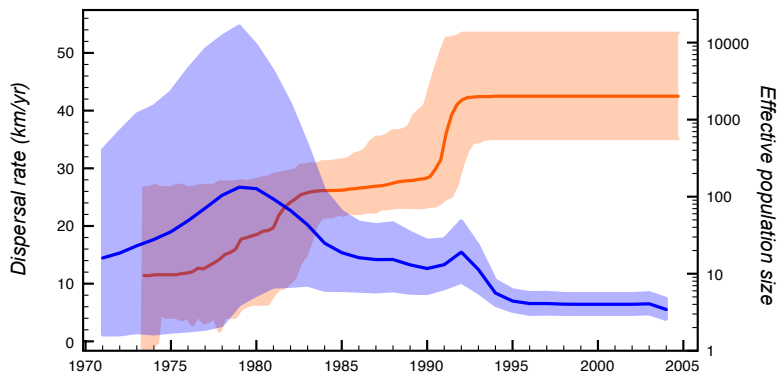
$\quad \vec{Y}_{root}$
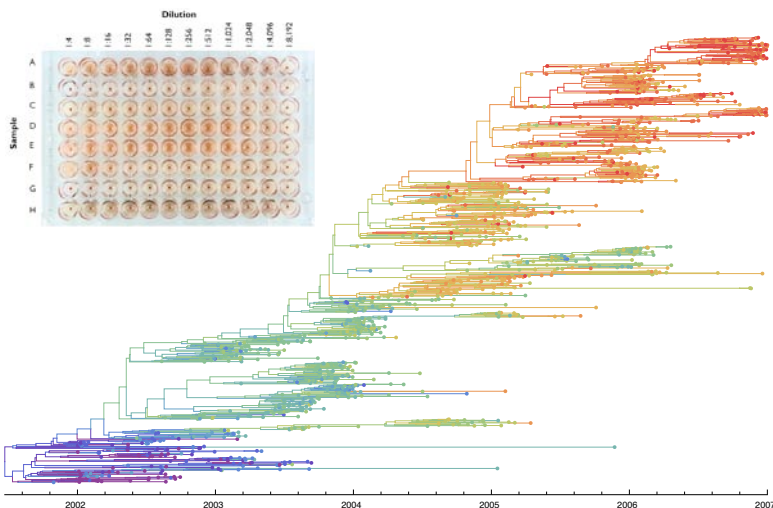
$\Delta \vec{Y} = BVN(0, \vec{S}t)$

Instantaneous diffusion rate ~ $1/\sigma^2$

Drummond et al., 2006 $1/\sigma^2 \times \phi_b$ *where b is a branch in* $\vec{F}$

A random rabies walk through time and space

Phylogeographic? diffusion

Acknowledgements

◉ Marc Suchard, University of California
◉ Andrew Rambaut, University of Edinburgh
◉ Alexei Drummond, University of Auckland
◉ John Welch, University of Edinburgh
◉ Beth Shapiro, Penn State University
◉ Eric Bloomquist, The Ohio State University

Influenza A
◉ Dirk Brockmann, Northwestern University
◉ Christian Thiemann, Northwestern University

◉ FWO
◉ Marsden Grant