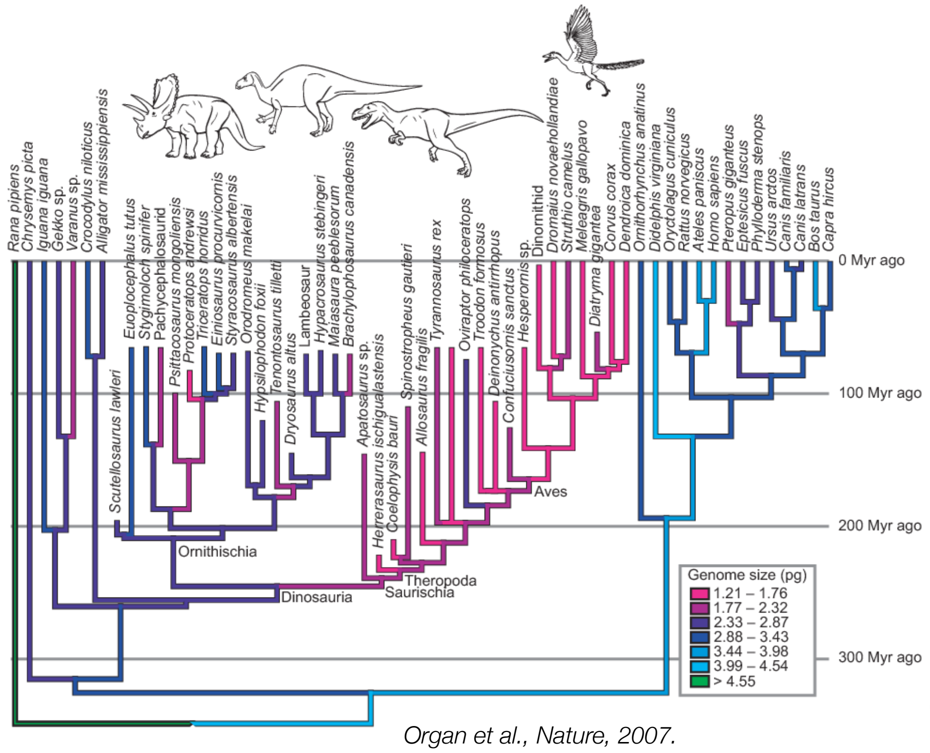
Clinical and Epidemiological Virology
Department of Microbiology
and Immunology
Rega Institute
KU Leuven, Belgium

Phylogenetic diffusion models

Philippe Lemey[1] and Marc Suchard[2]

1. Rega Institute, Department of Microbiology and Immunology, K.U. Leuven, Belgium.
2. Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA. Department of Biostatistics, UCLA School of Public Health

SISMID, Jul 10-12, 2019



Organ et al., Nature, 2007.

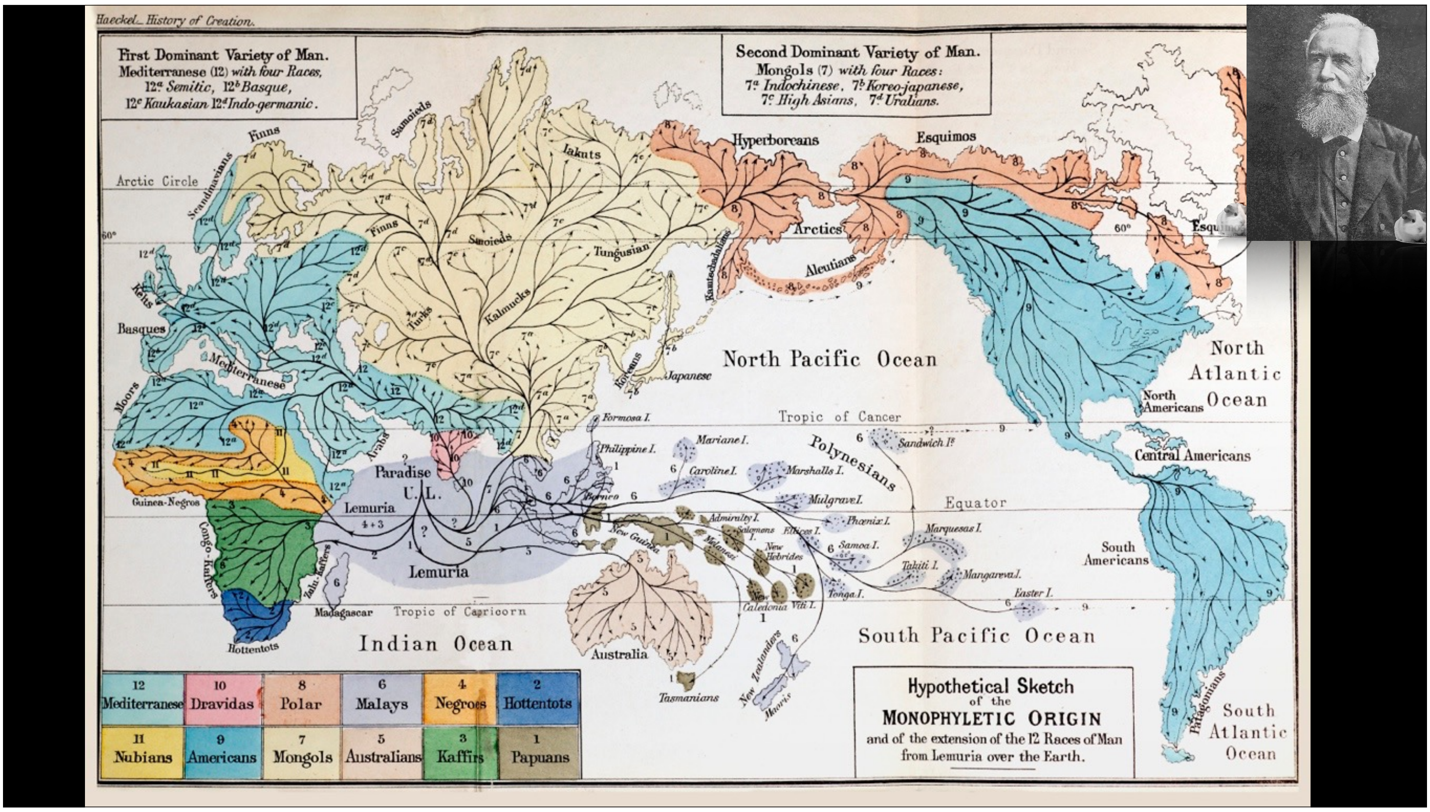

75¢

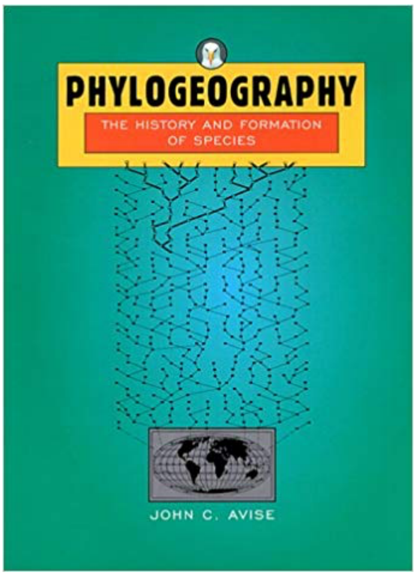
Vol. 7 — No. 1 January 3, 1989

Sun

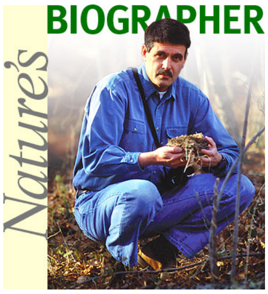
REVEALED:

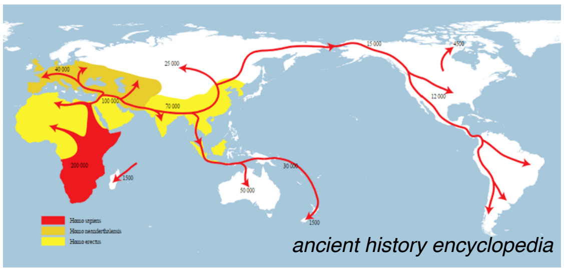
Hitler created AIDS virus to destroy U.S.





"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*



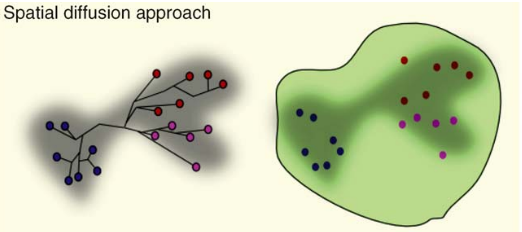


Three roads diverged? Routes to phylogeographic inference

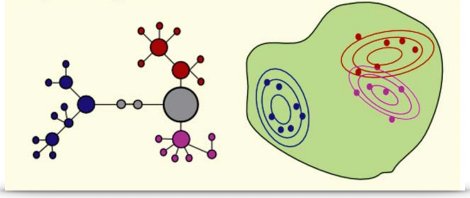
Erik W. Bloomquist[1], Philippe Lemey[2] and Marc A. Suchard[3,4]

TRENDS in Ecology & Evolution

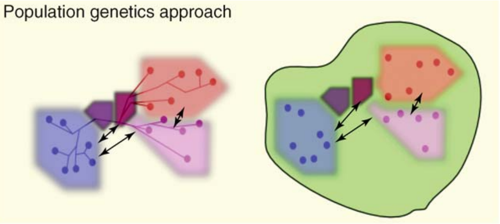
Spatial diffusion approach



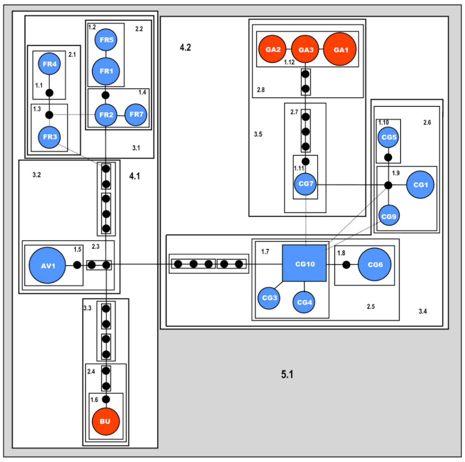
Comparative approach



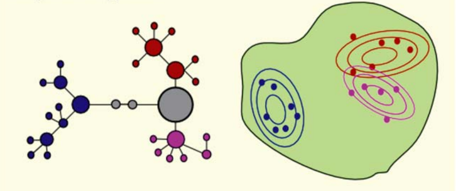
Population genetics approach

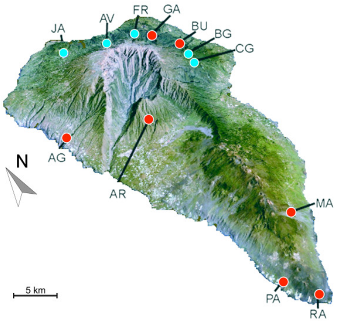


Nested clade phylogeographic analysis (NCPA)



Comparative approach





1. haplotype network
2. geographical structure test

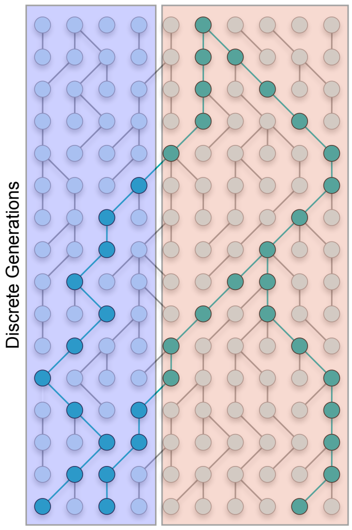
3. interpretation through inference key

1. Are all clades within the nesting clade found in separate areas with no overlap?
- NO – Go to step 2.
- YES - Go to step 19.

2. Is at least one of the following conditions satisfied?
   a. The D_c’s for one or more tips are significantly small and the D_c's for one or more of the interiors are significantly large or non-significant.
   b. The D_c’s for one or more tips are significantly small or non-significant and the D_c's for some but *not* all of the interiors are significantly small.
   c. The D_c’s for one or more interiors are significantly large and the D_c’s for the tips are either significantly small or non-significant
   d. The I-T D_c is significantly large.
- NO - Go to step 11.
- YES - Go to step 3.
- Tip/Interior Status Cannot be Determined - **Inconclusive Outcome.**

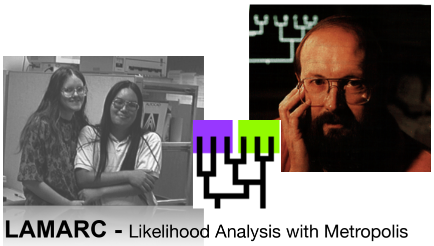
3. Is at least one of the following conditions satisfied?
   a. Some D_n and/or I-T D_n values are significantly reversed from the D_c values.
   b. One or more tip clades show significantly large D_n’s.
   c. One or more interior clades show significantly small D_n’s.
   d. I-T has a significantly small D_n with the corresponding D_c value non-significant.
- NO - Go to step 4.
- YES - Go to step 5.

4. Are both of the following conditions satisfied?
   a. The clades (or 2 or more subsets of them) with significantly small D_c or D_n values have ranges that are completely or mostly non-overlapping with the other clades in the nested group (particularly interiors), or if no clades have significantly small D_c or D_n values but an interior D_n or I-T D_n is significantly large, and that interior clade is completely or mostly non-overlapping with the other clades in the nested group.
   b. The pattern of completely or mostly non-overlapping ranges in the above condition represents a break or reversal from lower level trends within the nested clade series (applicable to higher-level clades only).
- NO - **Restricted Gene Flow with Isolation by Distance (Restricted Dispersal by Distance in Non-sexual species).** This inference is strengthened if the clades with restricted distributions are found in diverse locations, if the union of their ranges roughly corresponds to the range of one or more clades (usually interiors) within the same nested group (applicable only to nesting clades with many clade members or to the highest level clades regardless of number), and if the D_c values increase and become more geographically widespread with increasing clade level within a nested series (applicable to lower level clades only).
- YES - Go to step 9.



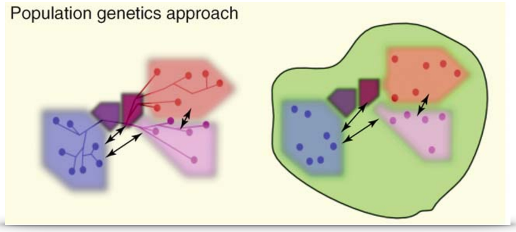
Coalescent theory:

- 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)



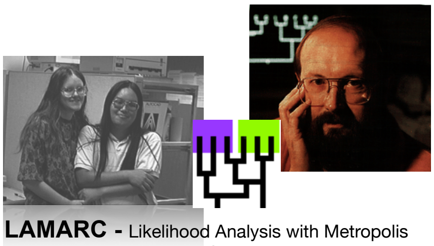
LAMARC - Likelihood Analysis with Metropolis Algorithm using Random Coalescence

MIGRATE-N



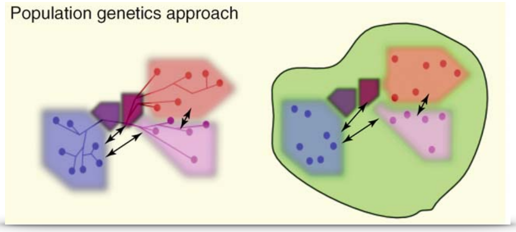
- Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the **structured coalescent**. Bioinformatics. 2014 Aug 15;30(16):2272-9. doi: 10.1093/bioinformatics/btu201.
- De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian **Structured Coalescent** Approximation. PLoS Genet. 2015 Aug 12;11(8):e1005421. doi: 10.1371/journal.pgen.1005421.
- Müller NF, Rasmussen DA, Stadler T. The **Structured Coalescent** and Its Approximations. Mol Biol Evol. 2017 Nov 1;34(11):2970-2981. doi: 10.1093/molbev/msx186.
- Kühnert D, Stadler T, Vaughan TG, Drummond AJ. **Phylodynamics** with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Mol Biol Evol. 2016 Aug;33(8):2102-16. doi: 10.1093/molbev/msw064.

Beast2
Bayesian evolutionary analysis by sampling trees

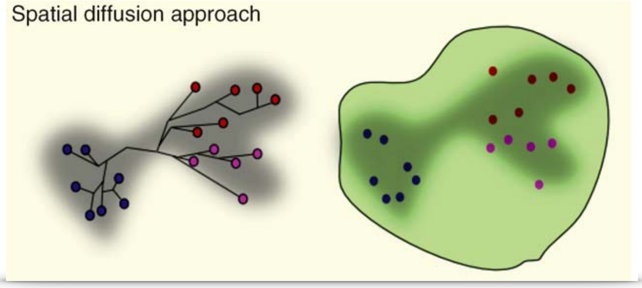


LAMARC - Likelihood Analysis with Metropolis Algorithm using Random Coalescence

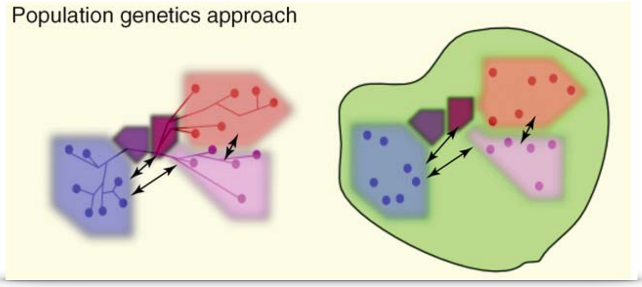
MIGRATE-N



Phylogenetic diffusion

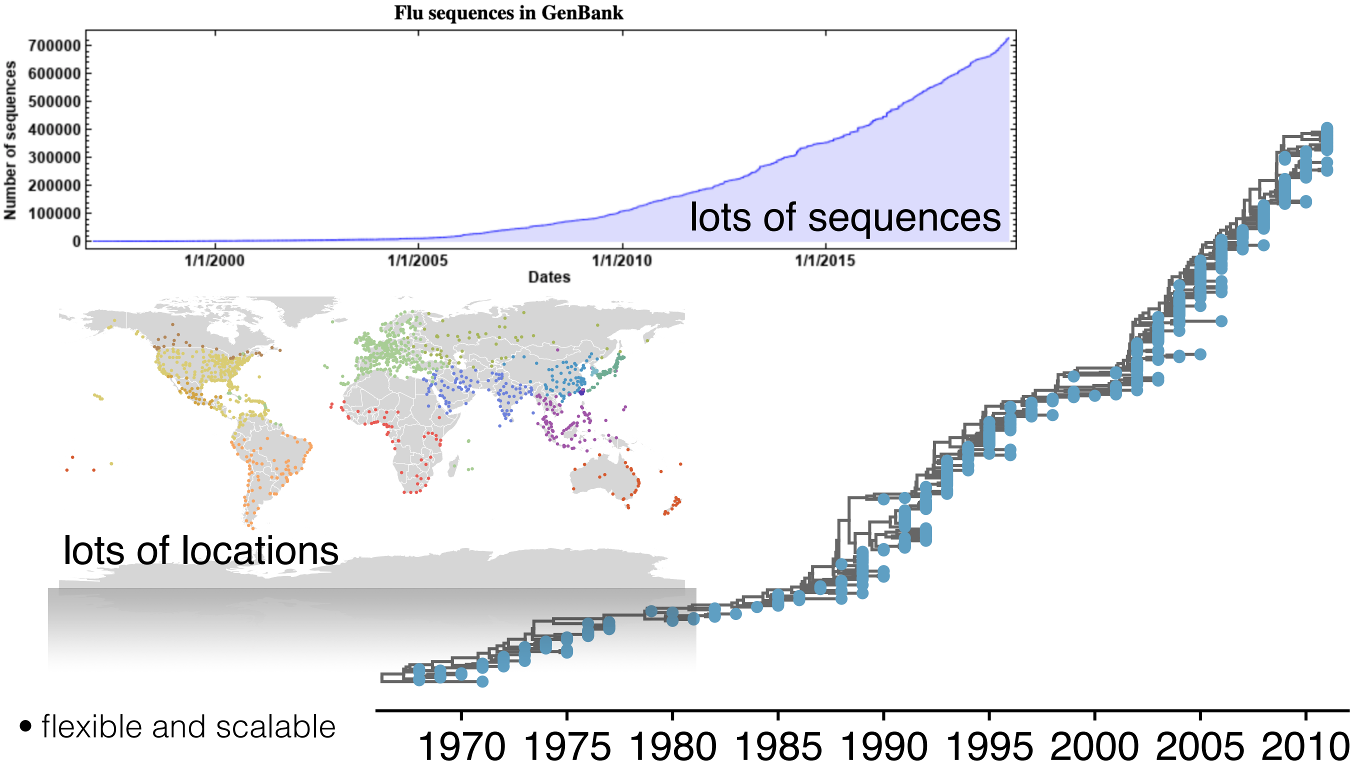


Structured coalescent



- Samples are assumed to be collected in a manner that is blind to their location.
- Sample location used as data
- Just as for genetic data, non-random sampling procedures will bias results
- flexible and scalable

- no assumptions about the manner in which samples are collected with respect to location
- Sample distribution not used as data
- Uneven sampling can reduce inference power, but will not bias results
- Computationally prohibitive

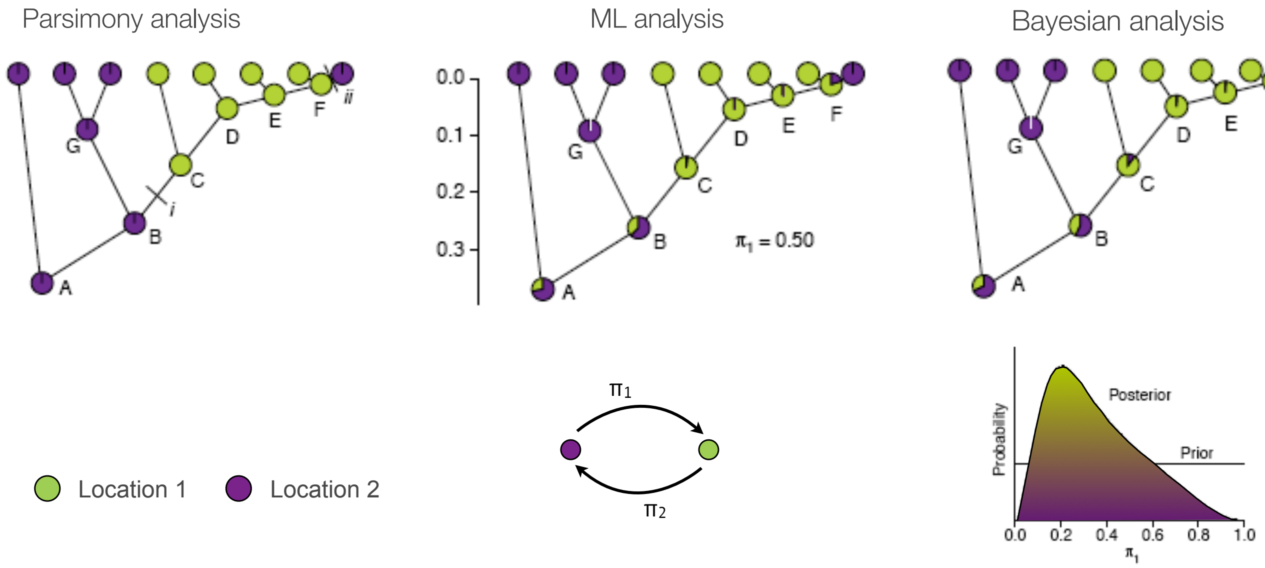


lots of sequences

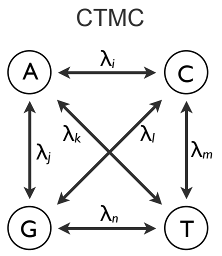
lots of locations

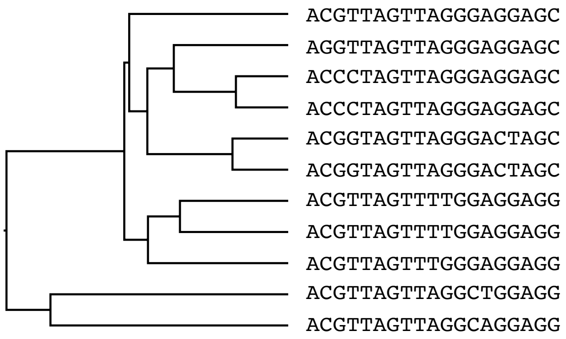
- flexible and scalable

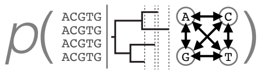
Discrete ancestral state reconstruction



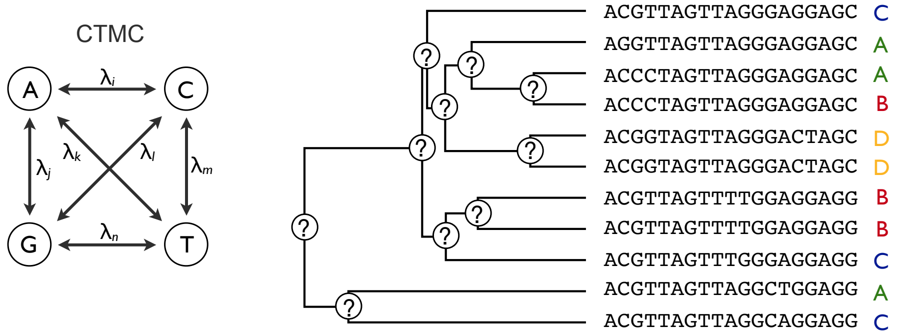
Discrete phylogenetic diffusion

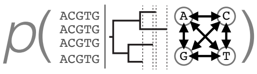




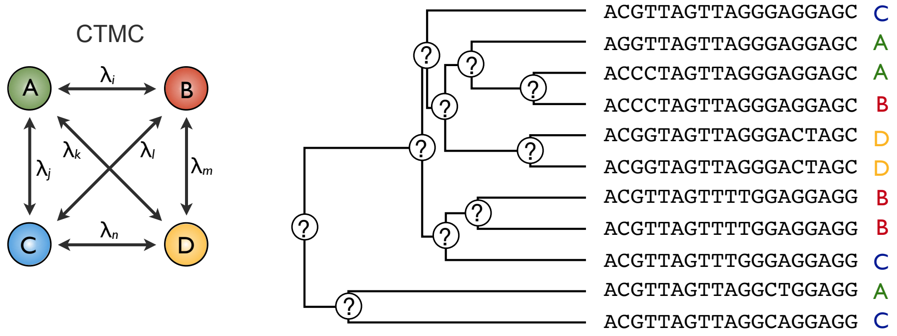


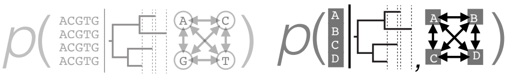
Discrete phylogenetic diffusion



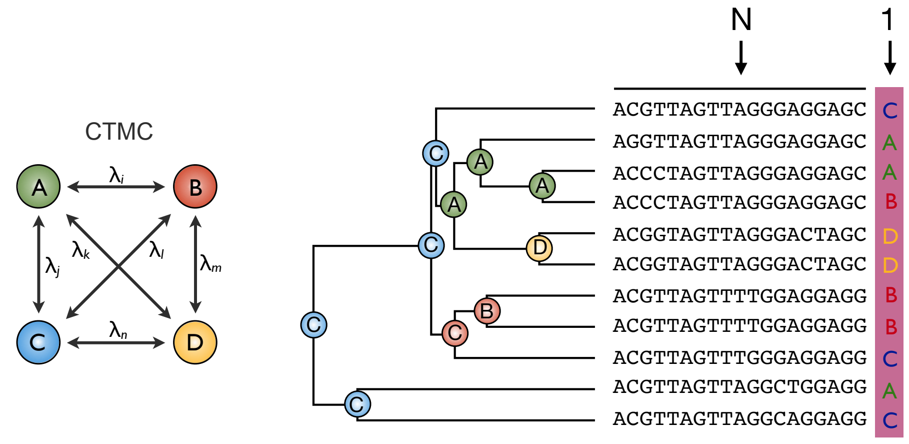


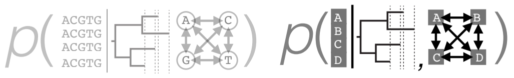
Discrete phylogenetic diffusion: *mu*gration





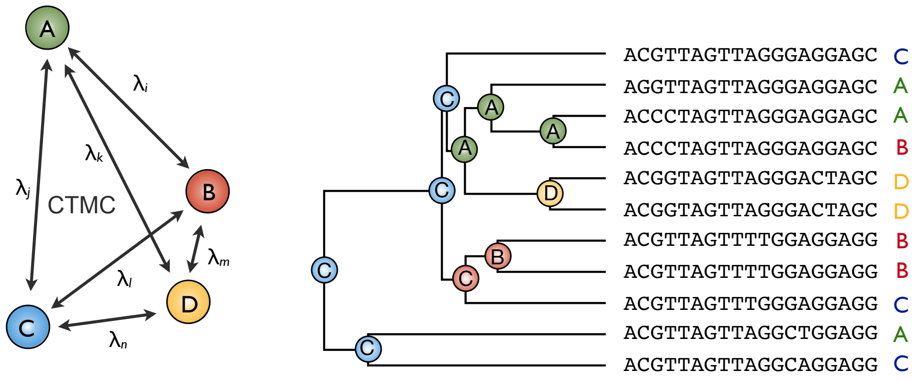
Discrete phylogenetic diffusion: *mu*gration

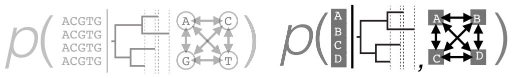




Discrete phylogenetic diffusion: *mu*gration

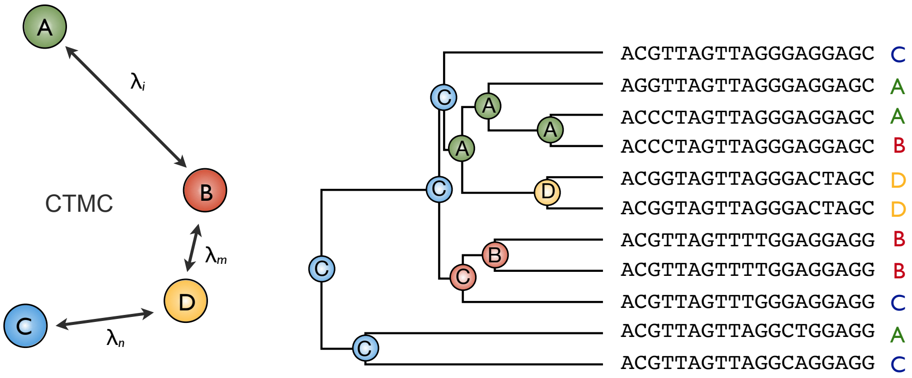
prior specification

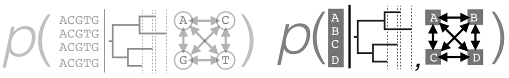




Discrete phylogenetic diffusion: *mu*gration

Bayesian variable selection - spike-and-slab



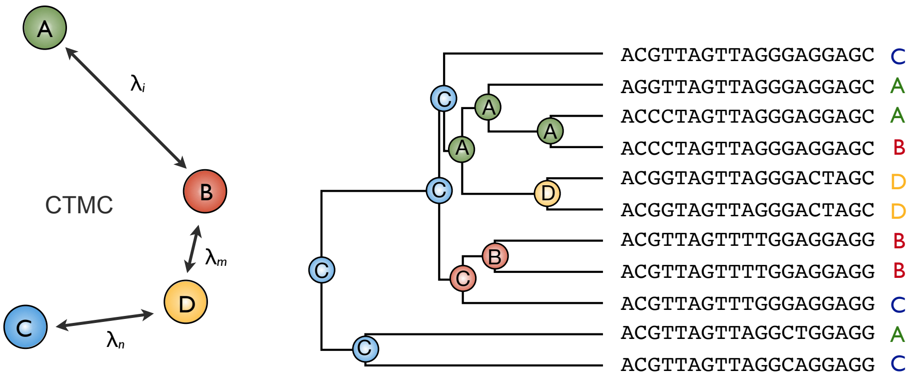


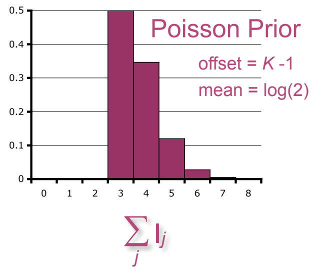
Discrete phylogenetic diffusion: *mu*gration

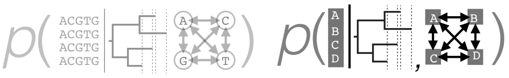
Bayesian variable selection - spike-and-slab

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

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & . & \pi_B i I_i & \pi_C j I_j & \pi_D k I_k \\ B & \pi_A i I_i & . & \pi_C l I_l & \pi_D m I_m \\ C & \pi_A j I_j & \pi_B l I_l & . & \pi_D n I_n \\ D & \pi_A k I_k & \pi_B m I_m & \pi_C n & . \end{array}$$

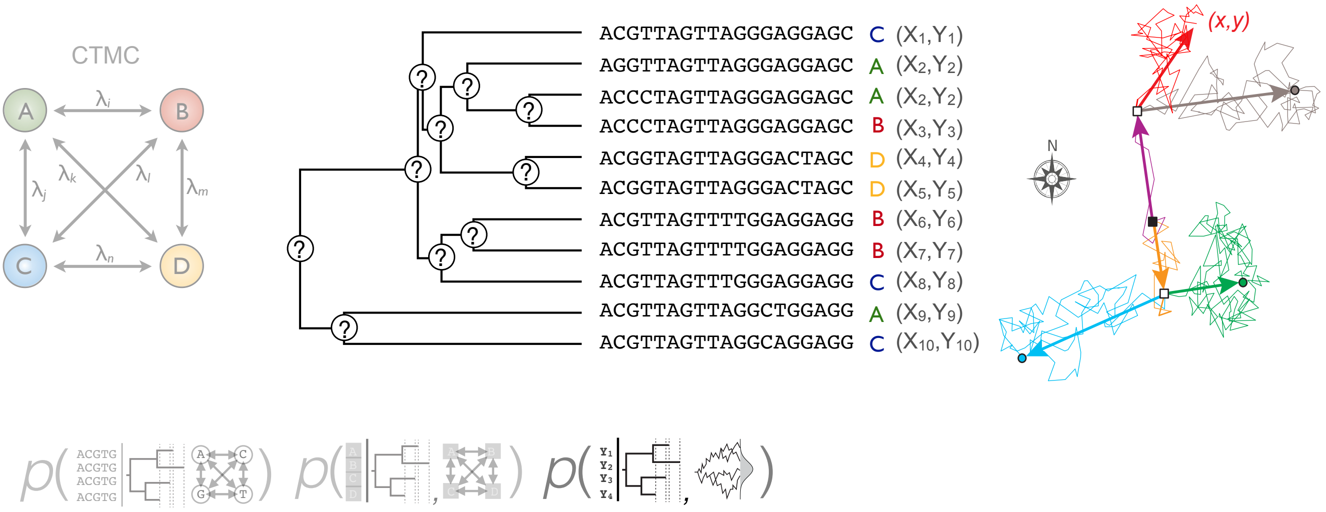


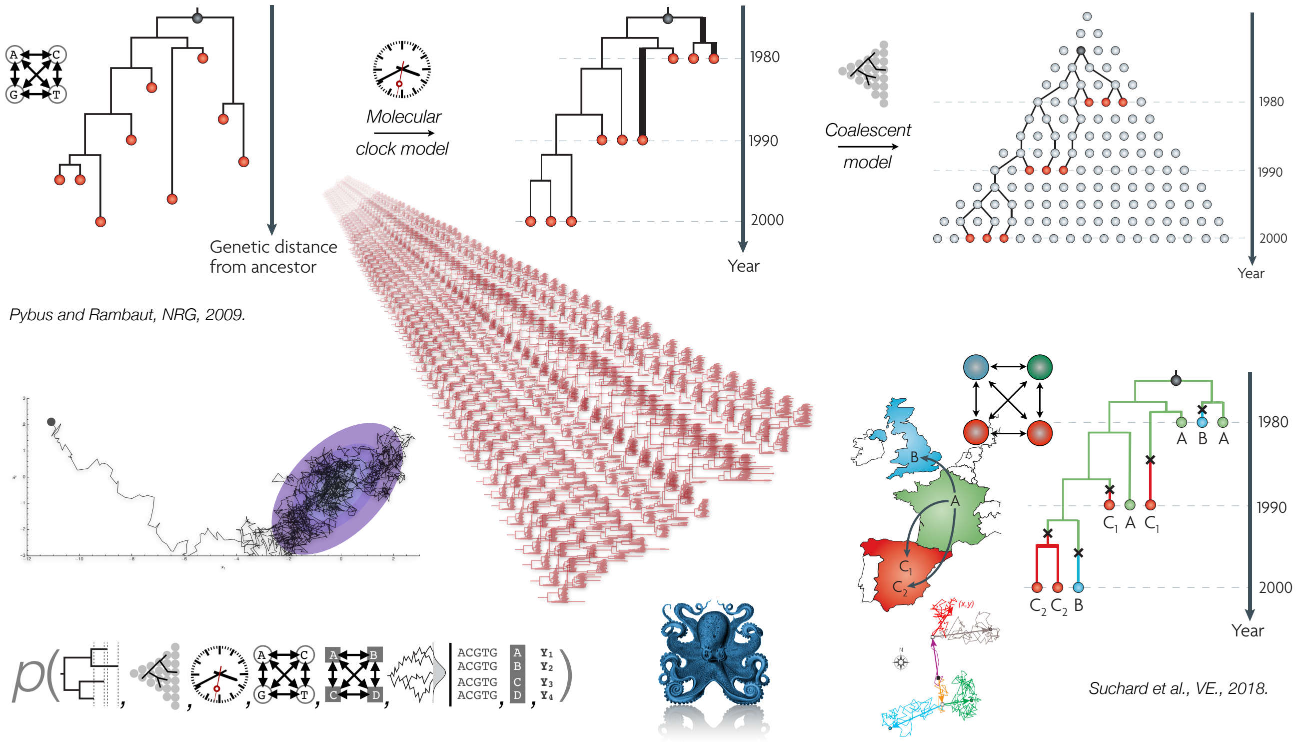




$$\text{Bayes factor} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\Pr(I=1|\ D)/(1-\Pr(I=1|\ D)}{\Pr(I=1)/(1-\Pr(I=1)}$$

Continuous phylogenetic diffusion: Brownian motion process



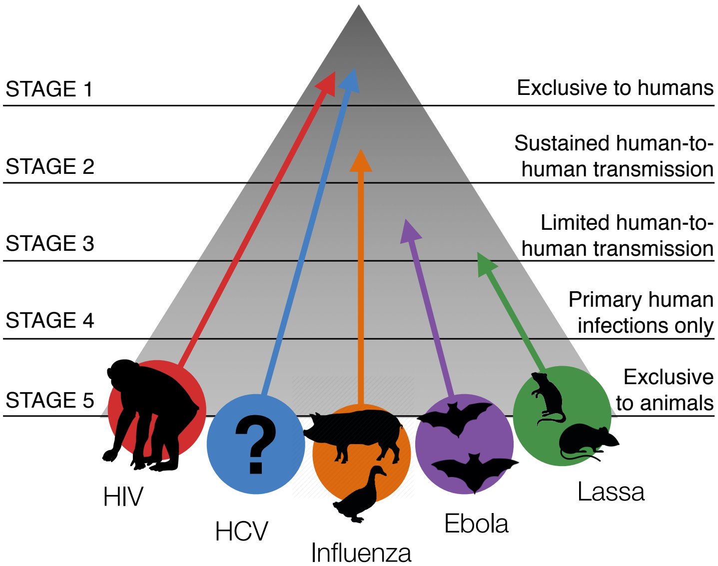


Pybus and Rambaut, NRG, 2009.

Suchard et al., VE., 2018.

Phylodynamics of pathogen emergence and spread

- *Where do they come from?*
- *How do they get established in humans*
- *What is the role of adaptation vs. epidemiology in shaping pathogen genetic diversity*
- *How to support public health interventions*



STAGE 1	Exclusive to humans
STAGE 2	Sustained human-to-human transmission
STAGE 3	Limited human-to-human transmission
STAGE 4	Primary human infections only
STAGE 5	Exclusive to animals

Wolfe et al., Nature, 2007.

Phylodynamics of pathogen emergence and spread

- *Where do they come from?*
- *How do they get established in humans*
- *f adaptation n shaping diversity*
- *ublic health*

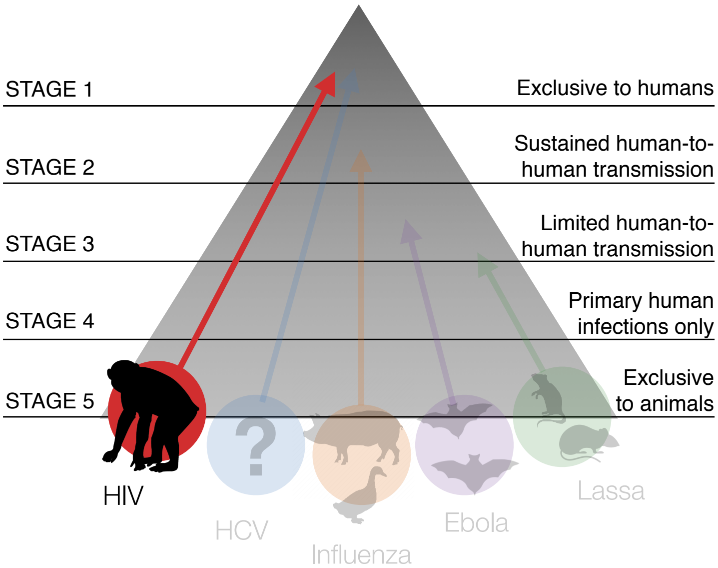
75¢

Vol. 7 — No. 1 January 3, 1989

Sun

REVEALED:

Hitler created AIDS virus to destroy U.S.



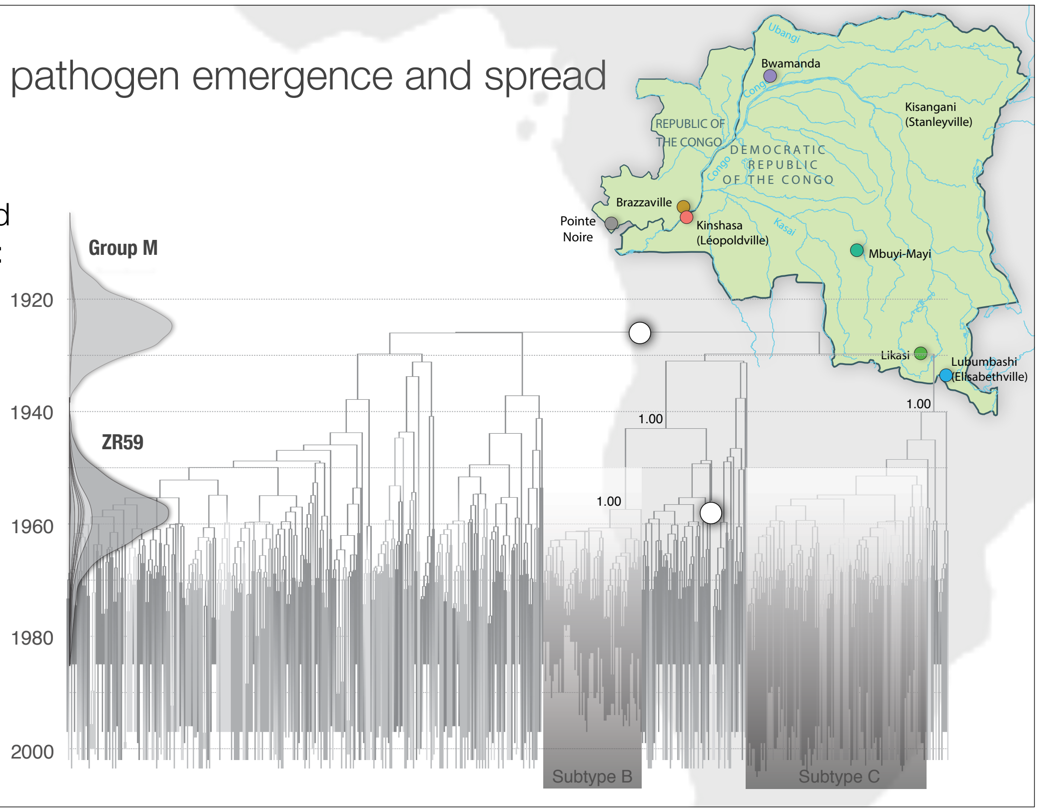
STAGE 1	Exclusive to humans
STAGE 2	Sustained human-to-human transmission
STAGE 3	Limited human-to-human transmission
STAGE 4	Primary human infections only
STAGE 5	Exclusive to animals

Wolfe et al., Nature, 2007.

Phylodynamics of pathogen emergence and spread

HIV-1 group M emerged around 1920 (95% BCI: 1909-30)

Faria et al., Science, 2014.

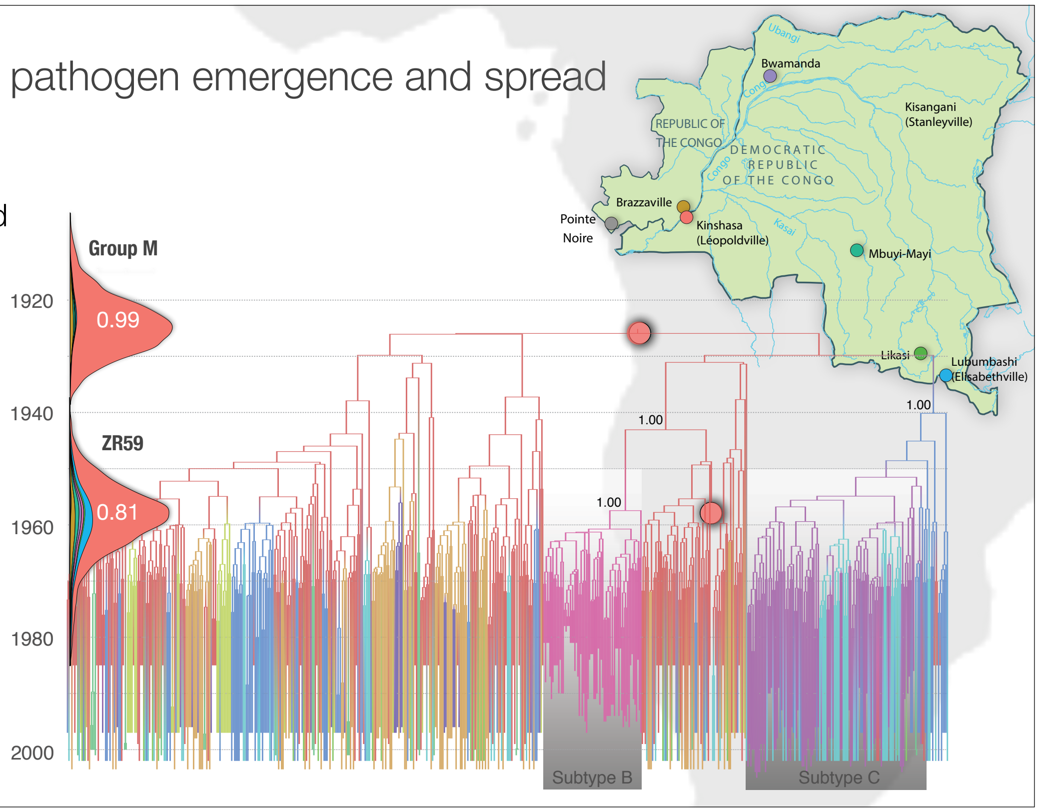



HIV

Phylodynamics of pathogen emergence and spread

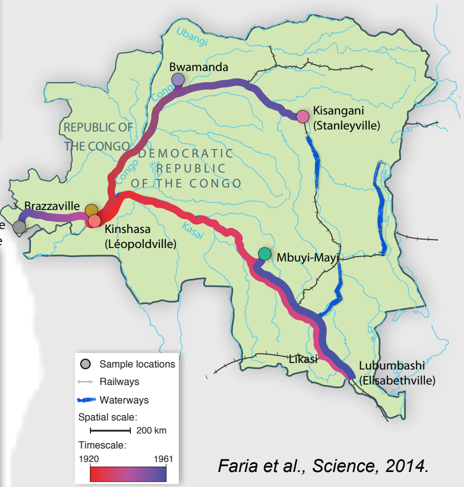
HIV-1 group M emerged around 1920 (95% BCI: 1909-30)
in Leopoldville

Faria et al., Science, 2014.




HIV

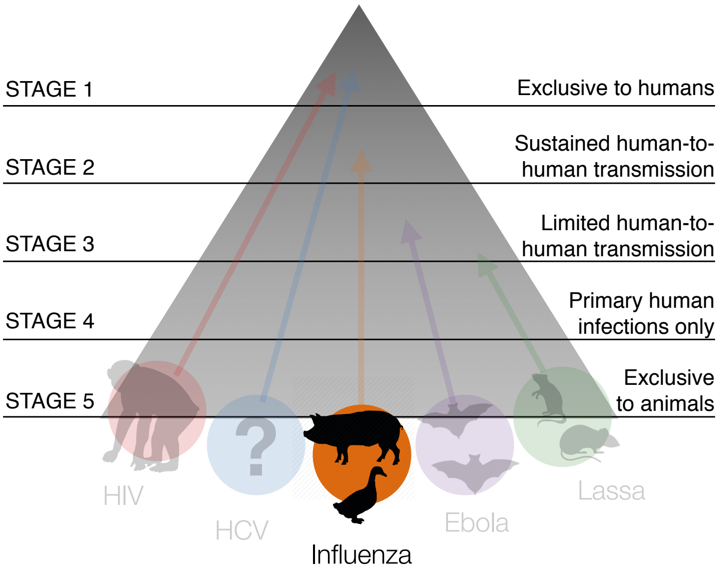
From early on, Kinshasa was the best connected city in central Africa



Faria et al., Science, 2014.

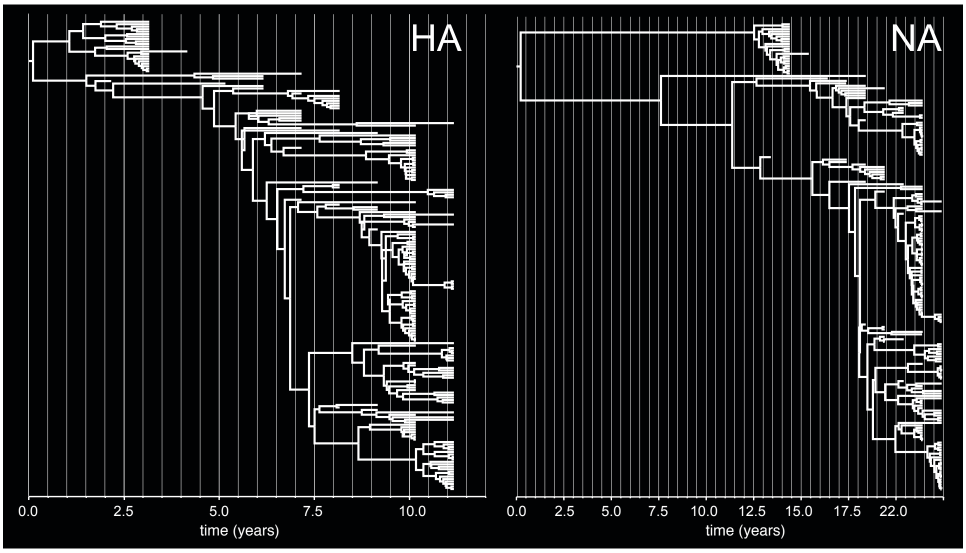
Phylodynamics of pathogen emergence and spread

- *Where do they come from?*
- *How do they get established in humans*
- *What is the role of adaptation vs. epidemiology in shaping pathogen genetic diversity*
- *How to support public health interventions*



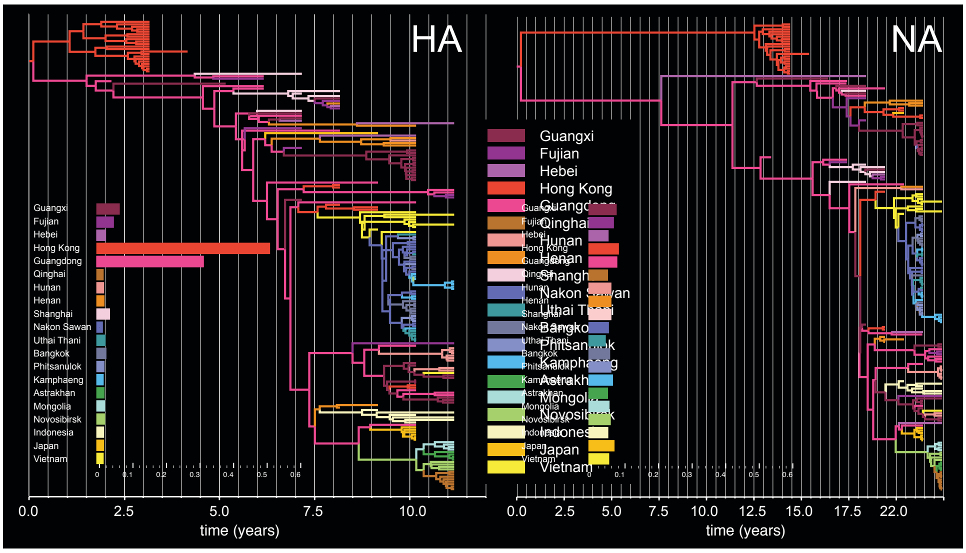
Wolfe et al., Nature, 2007.

Avian influenza H5N1



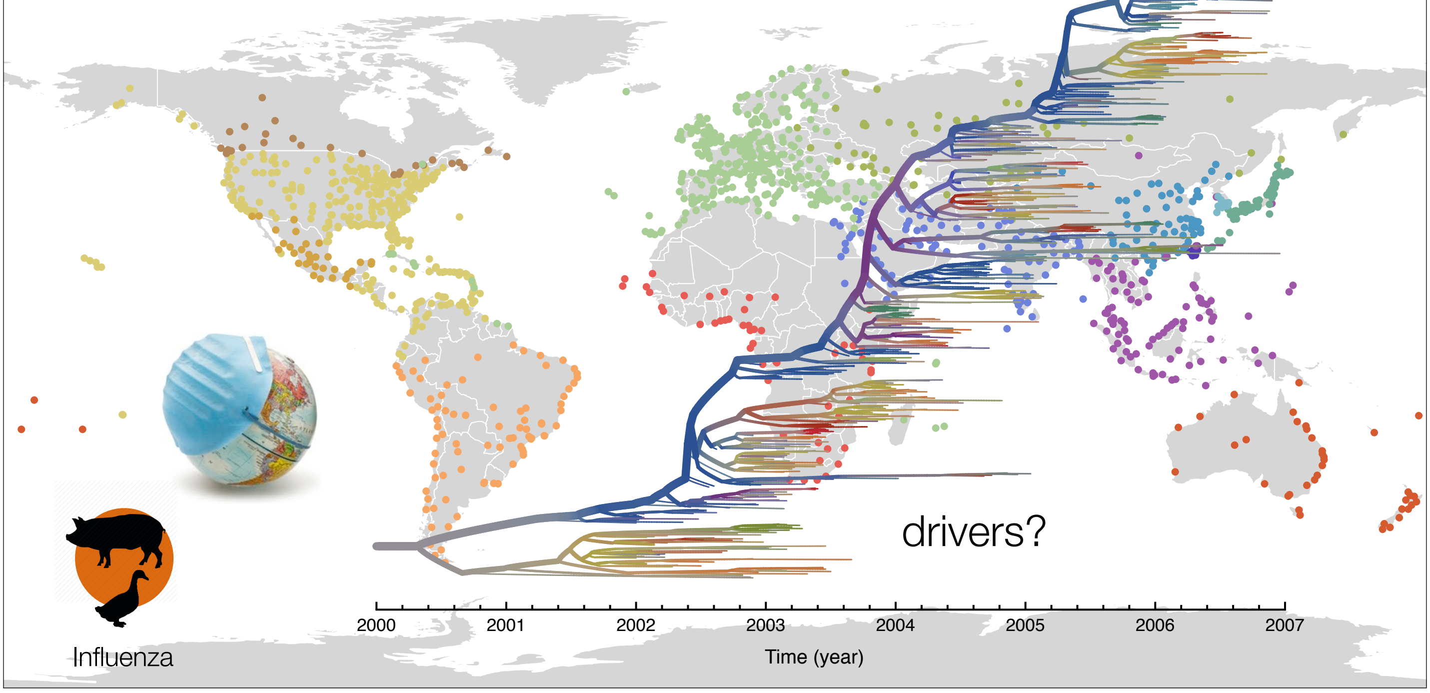
Wallace et al., PNAS, 2007

Avian influenza H5N1



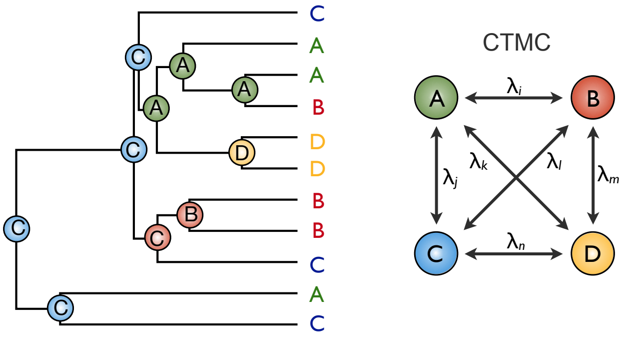
Wallace et al., PNAS, 2007

The global circulation of human influenza A H3N2

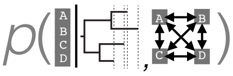


drivers?

Discrete phylogenetic diffusion: *mu*gration

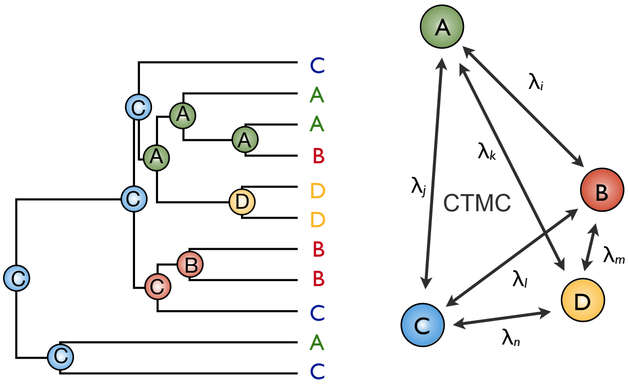


drivers?

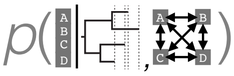


Lemey et al., PLoS Path, 2014

Discrete phylogenetic diffusion: *mu*gration

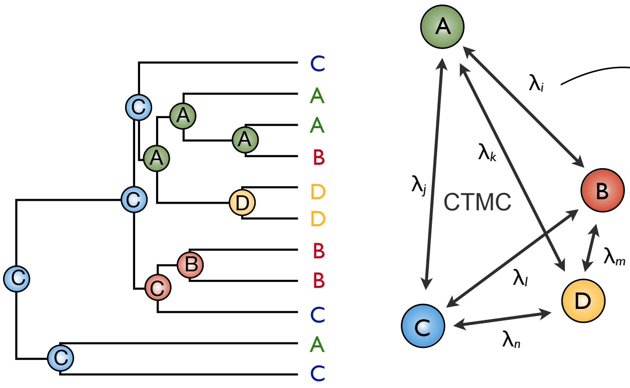


drivers?



Lemey et al., PLoS Path, 2014

Discrete phylogenetic diffusion: *mu*gration

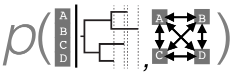


$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{i,j,1} + \beta_2 \delta_2 x_{i,j,2} + \ldots + \beta_P \delta_P x_{i,j,P}$$

Predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$

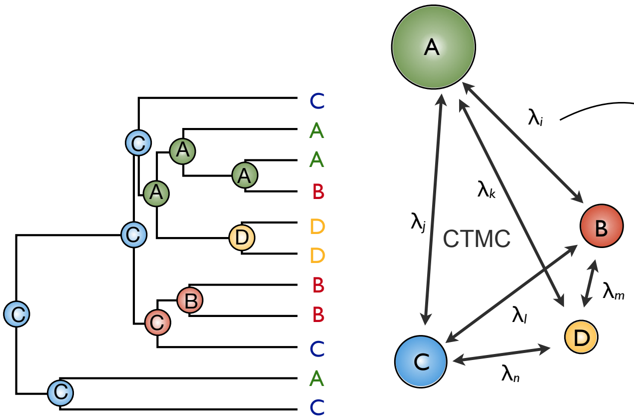
$$\mathbf{x}_p = (x_{1,2,p}, \ldots x_{K-1,K,p})$$

drivers?



Lemey et al., PLoS Path, 2014

Discrete phylogenetic diffusion: *mu*gration



inclusion probability

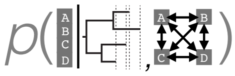
contribution

$$\log \Lambda_{ij} = \beta_1\delta_1 x_{i,j,1} + \beta_2\delta_2 x_{i,j,2} + \ldots + \beta_P\delta_P x_{i,j,P}$$

Predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$

$\mathbf{x}_p = (x_{1,2,p}, \ldots x_{K-1,K,p})$

drivers?



Lemey et al., PLoS Path, 2014

Predictors

Average distance

Minimum distance

Origin absolute latitude

Destination absolute latitude

Passenger flow

Origin population size

Destination population size

Origin population density

Destination population density

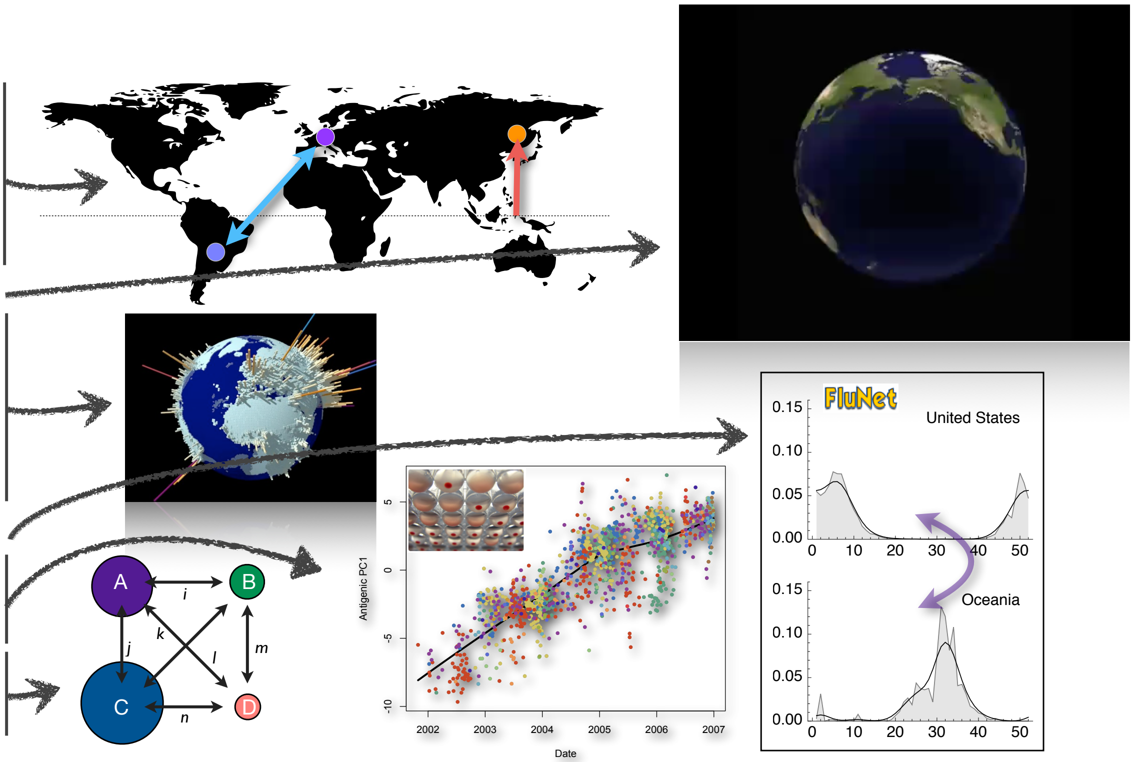
H3 incidence overlap

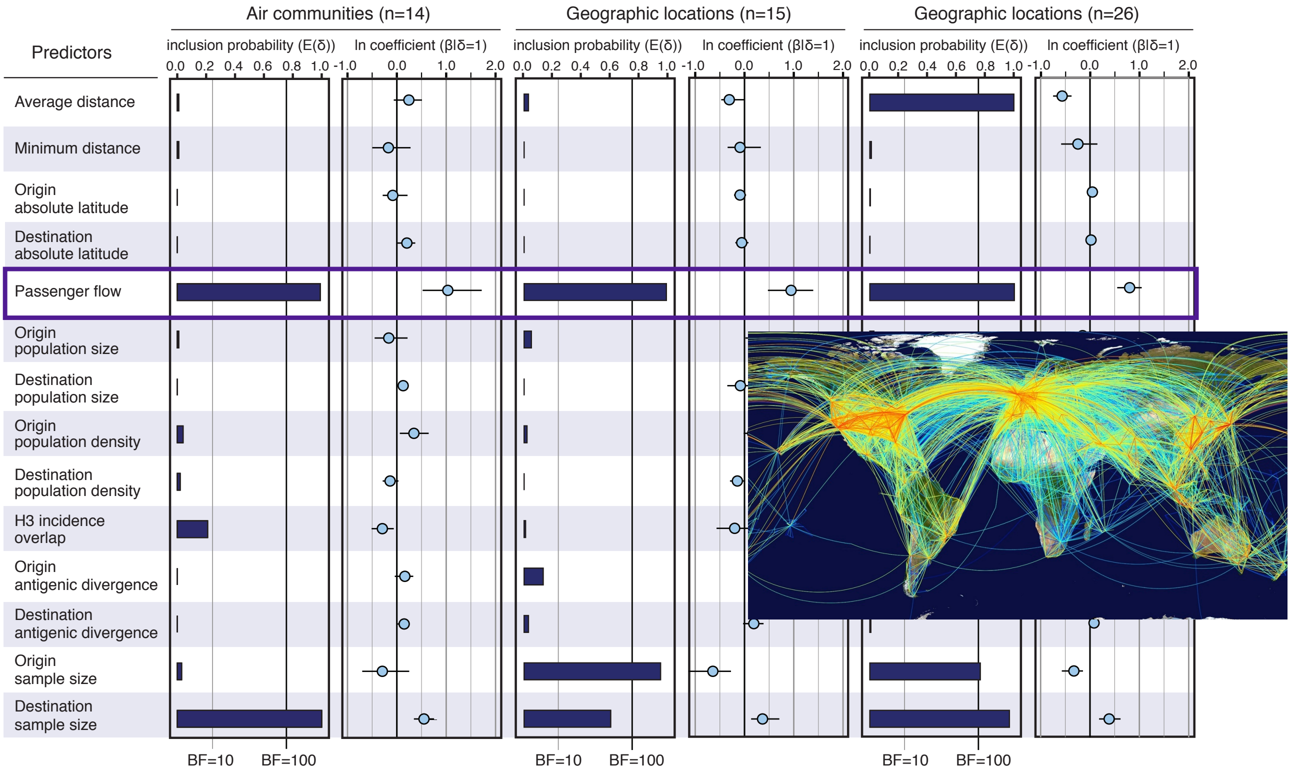
Origin antigenic divergence

Destination antigenic divergence

Origin sample size

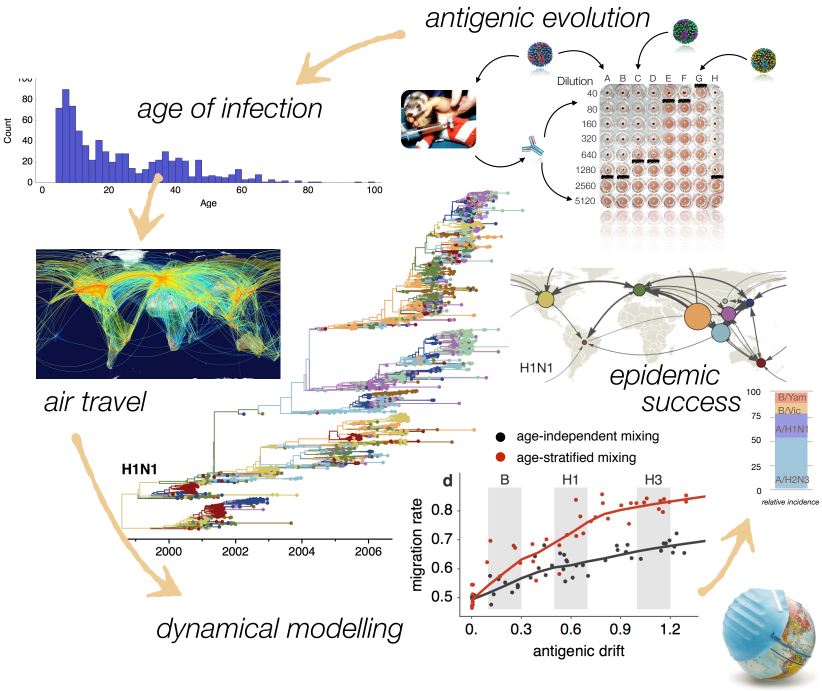
Destination sample size





The global circulation of human influenza A and B

- Where do they come from?
- How do they get established in humans
- *What is the role of adaptation vs, epidemiology in shaping pathogen genetic diversity*
- How to support public health interventions

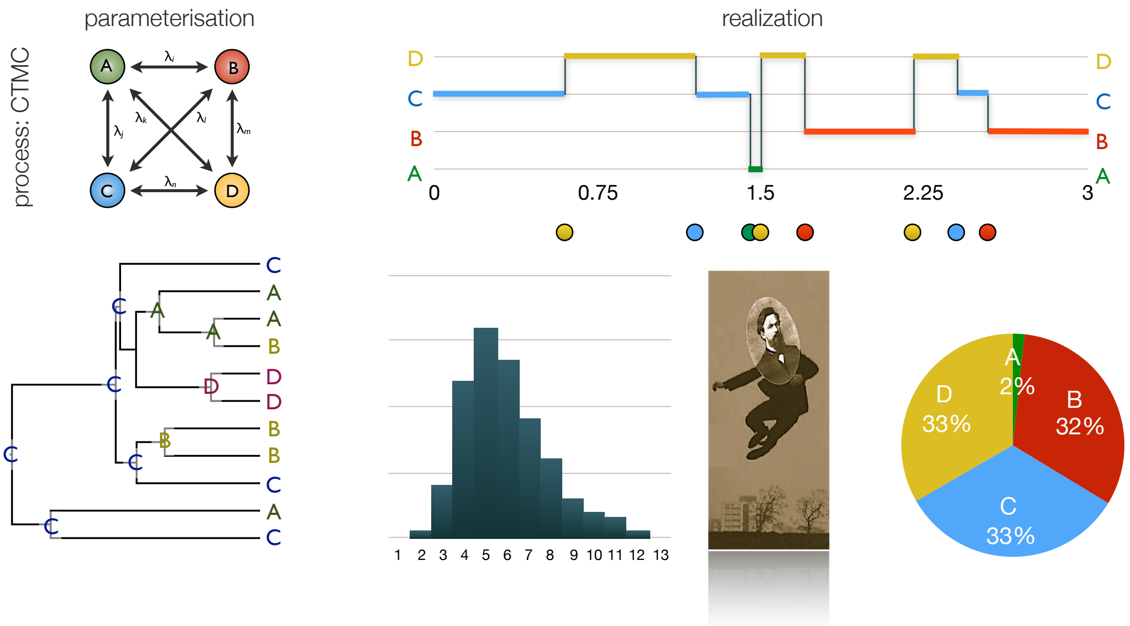


Influenza

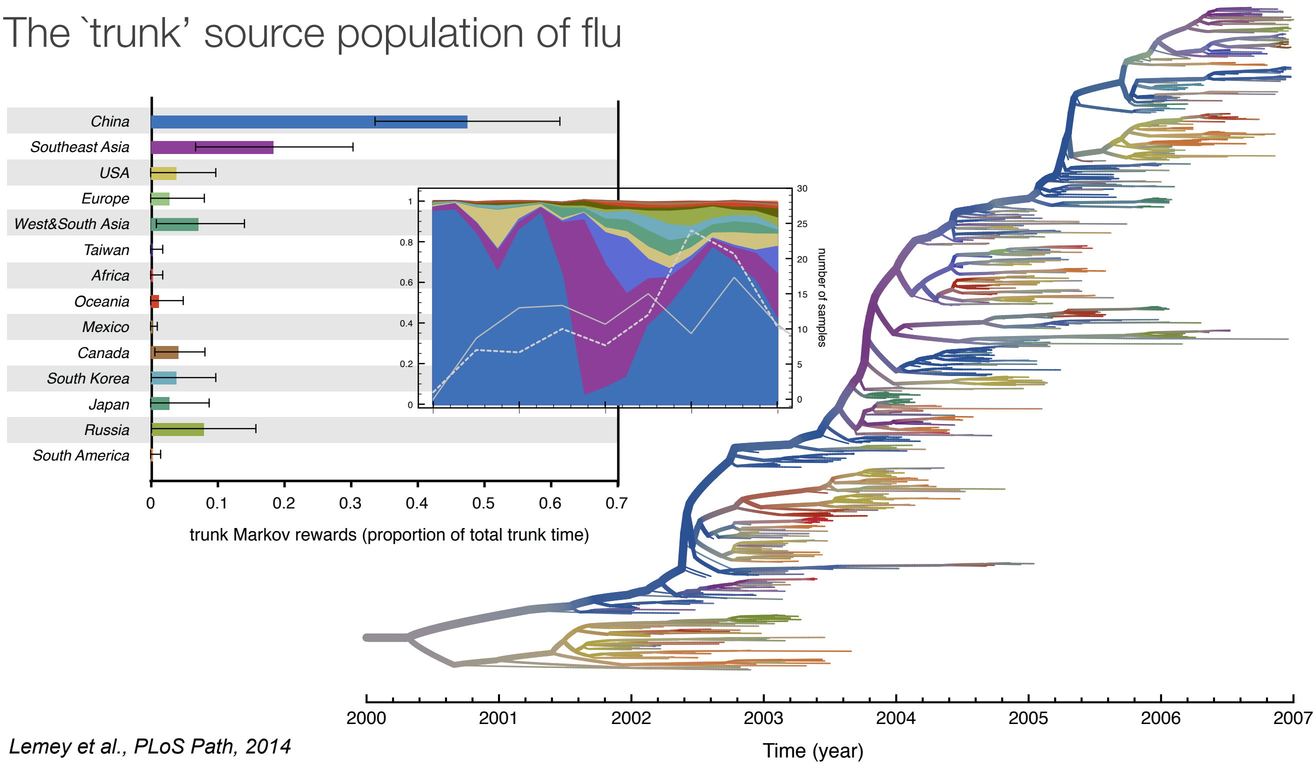
Bedford *et al.*, elife, 2014

Bedford *et al.*, Nature, 2015

From parameters to realisations: Markov jumps and rewards



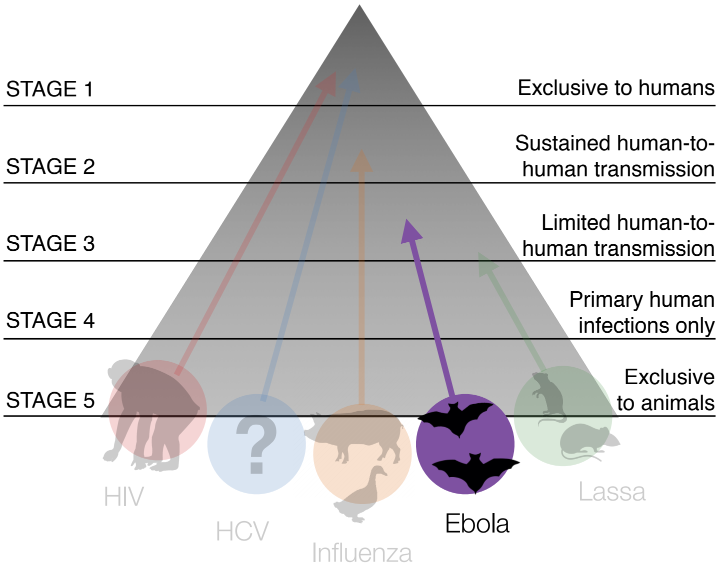
The `trunk' source population of flu



Lemey et al., PLoS Path, 2014

Phylodynamics of pathogen emergence and spread

- Where do they come from?
- How do they get established in humans
- What is the role of adaptation vs. epidemiology in shaping pathogen genetic diversity
- *How to support public health interventions*



STAGE 1 — Exclusive to humans

STAGE 2 — Sustained human-to-human transmission

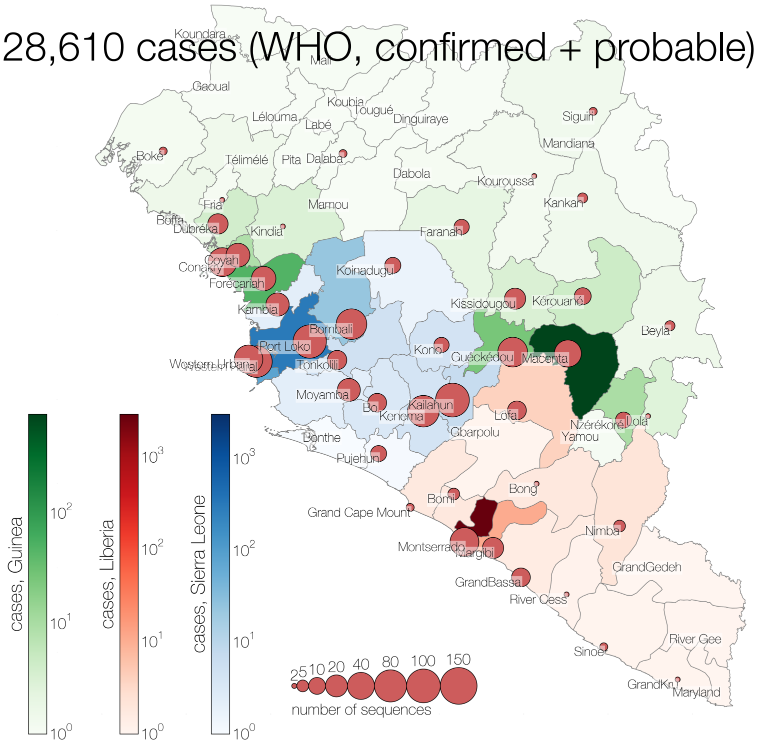
STAGE 3 — Limited human-to-human transmission

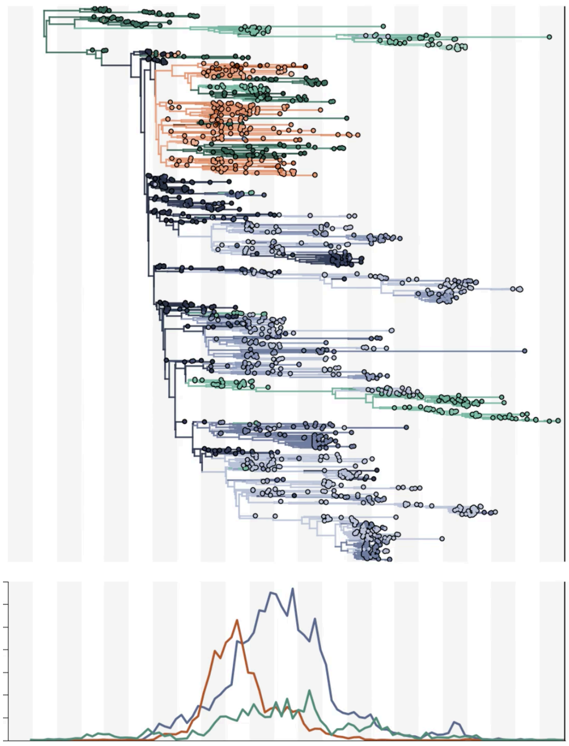
STAGE 4 — Primary human infections only

STAGE 5 — Exclusive to animals

Wolfe et al., Nature, 2007.

28,610 cases (WHO, confirmed + probable)





1610 EBOV genomes (>5% of cases)

geographical distance

country boundaries

accessibility (time to travel to 100K city)

economic activity (GCP)

population density

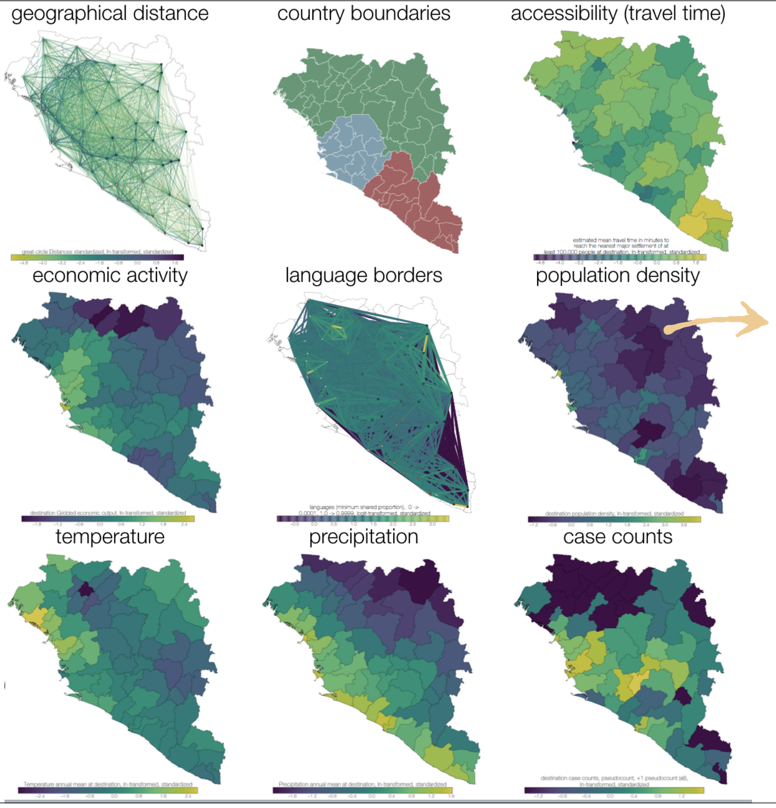
number of cases (WHO)

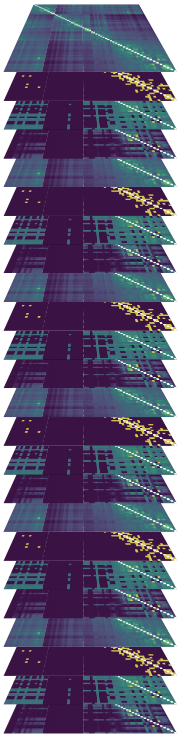
languages

temperature

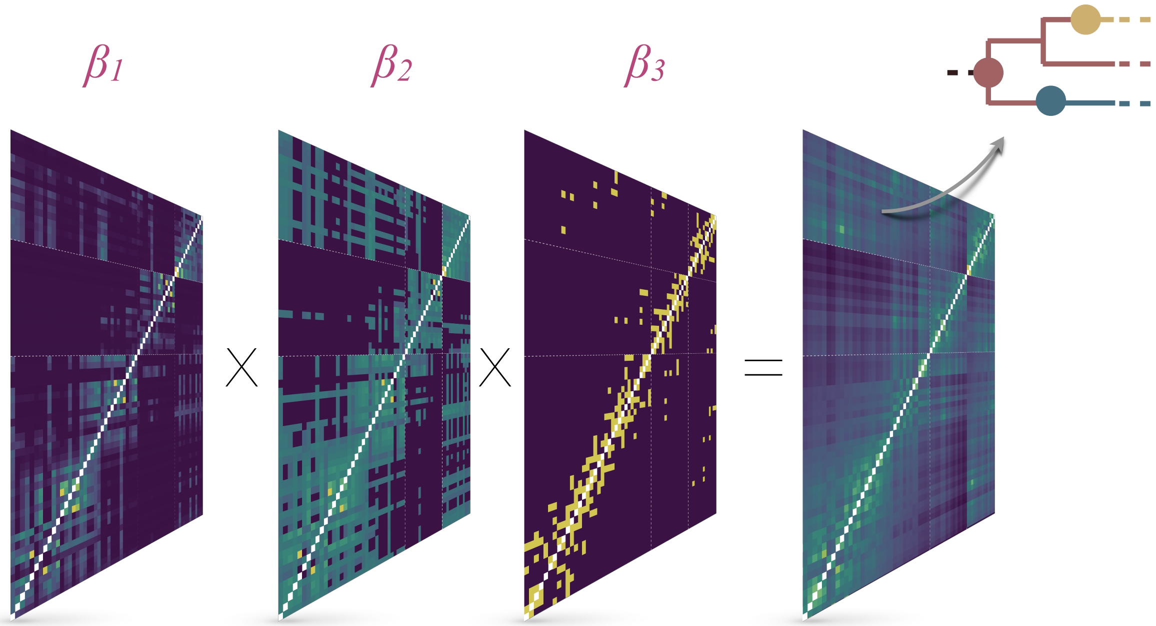
precipitation

climate seasonality

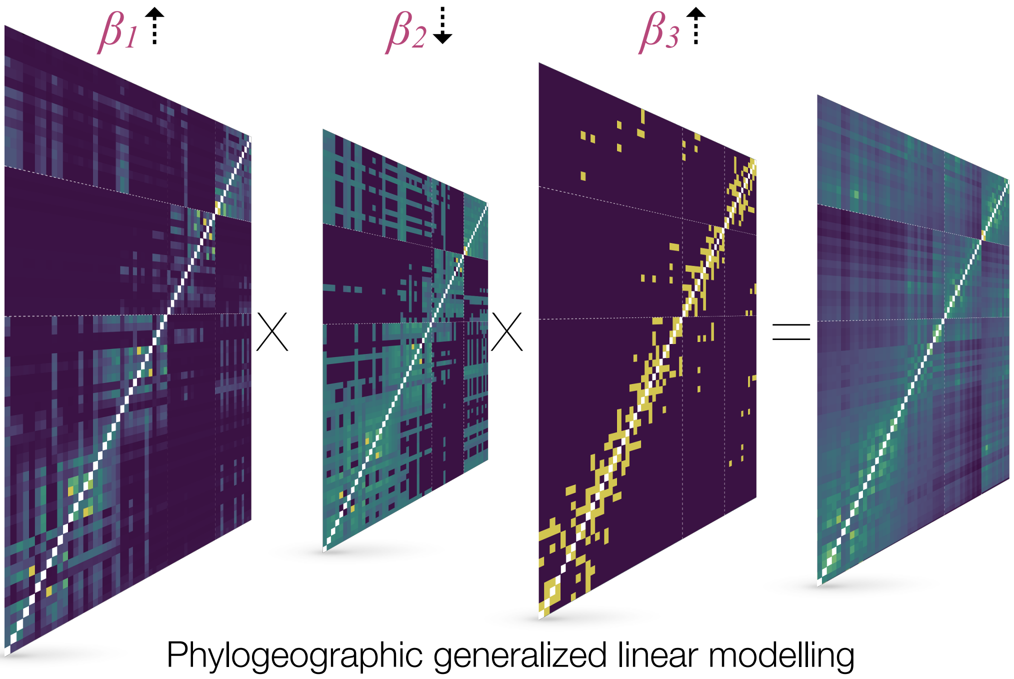




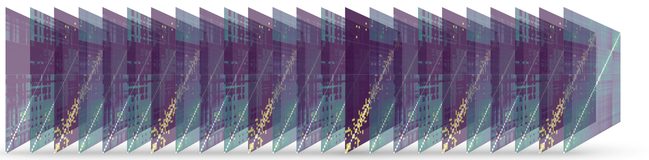
Dudas et al., Nature, in 2017.

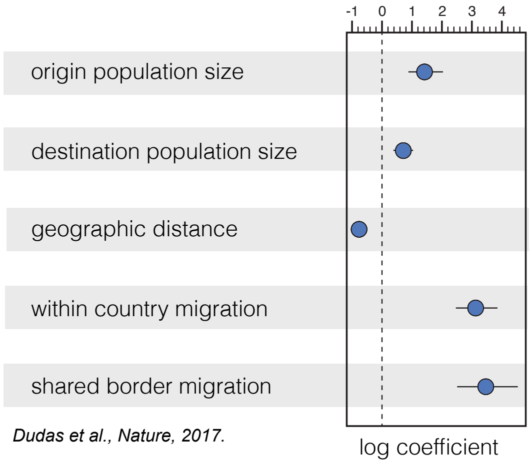


Phylogeographic generalized linear modelling

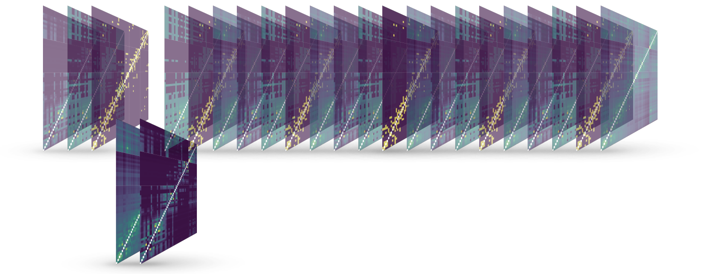


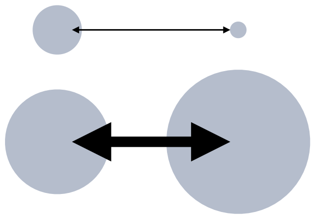
Phylogeographic generalized linear modelling

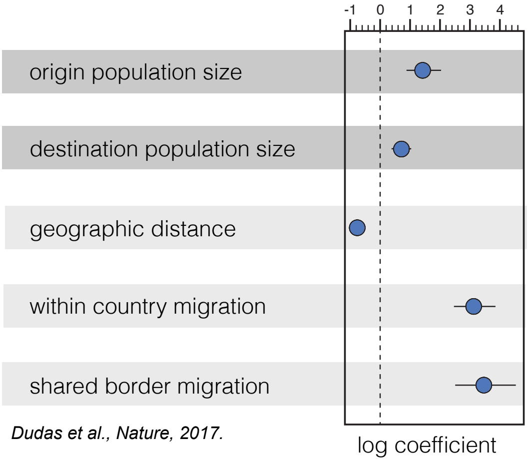




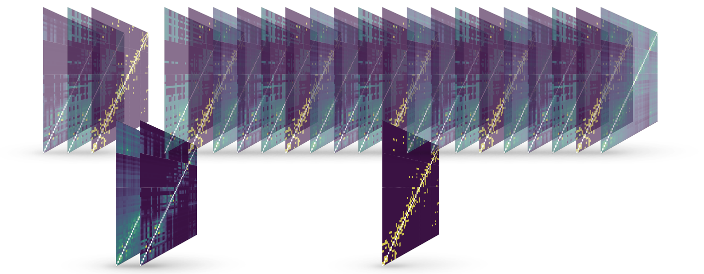
Dudas et al., Nature, 2017.

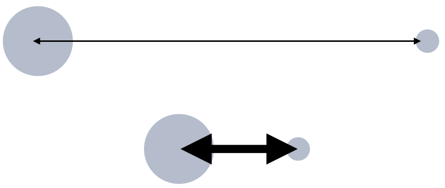


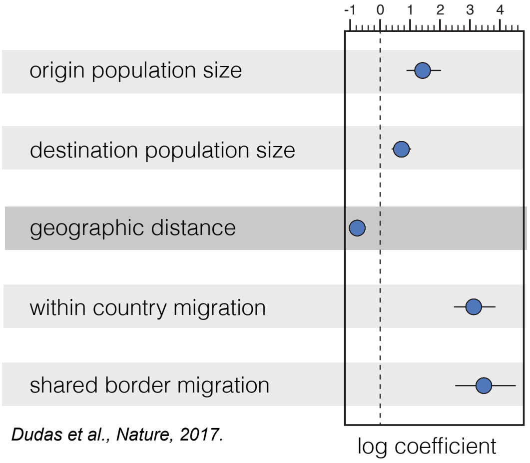




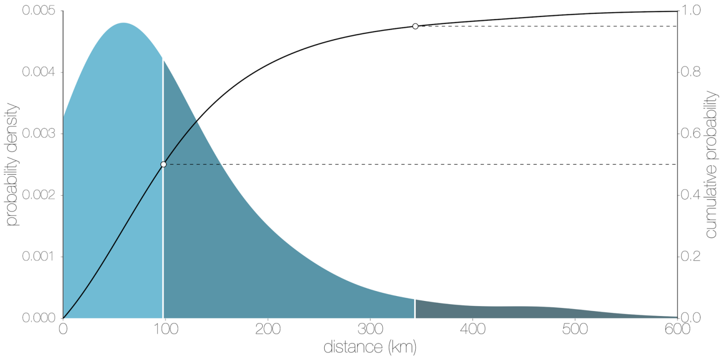
Dudas et al., Nature, 2017.





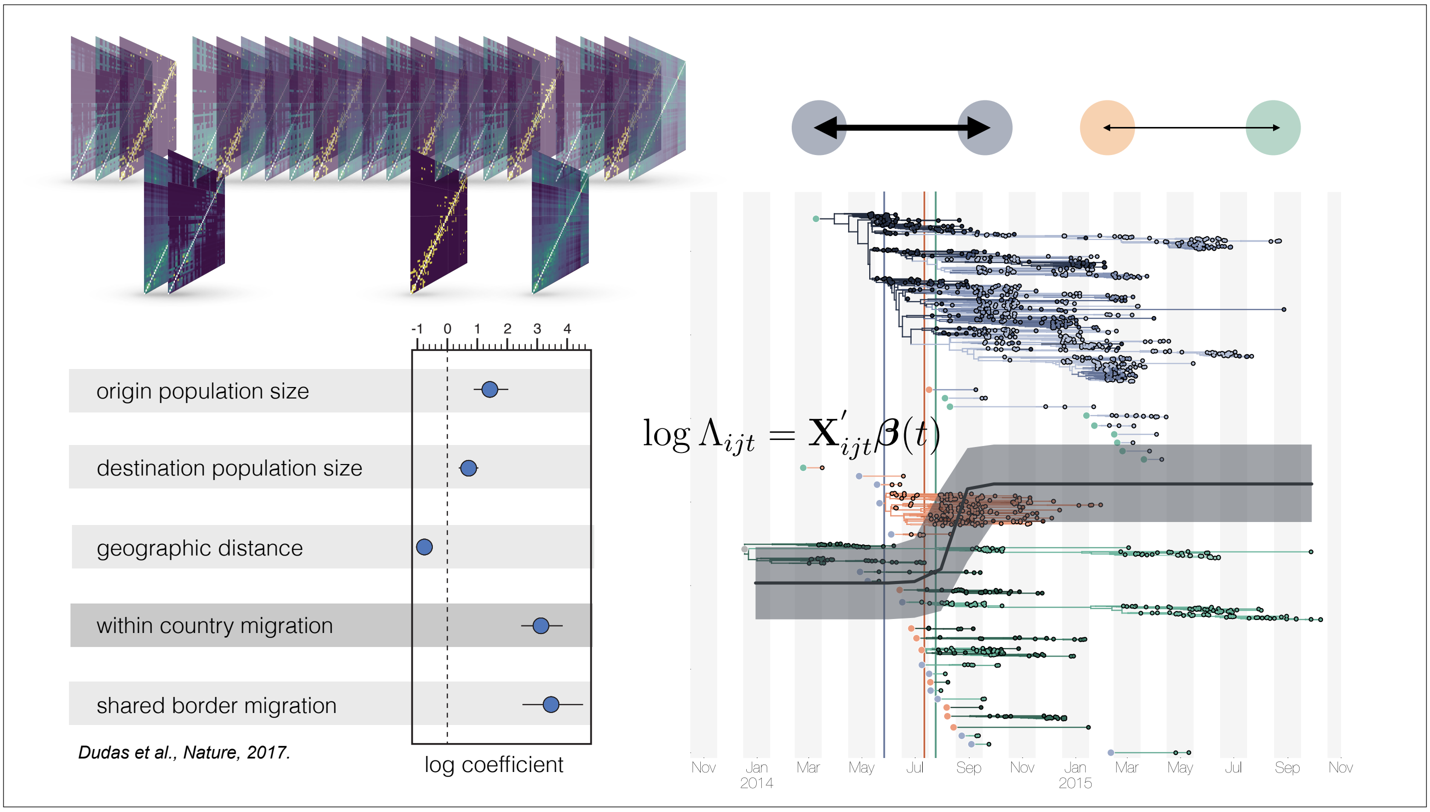


Dudas et al., Nature, 2017.



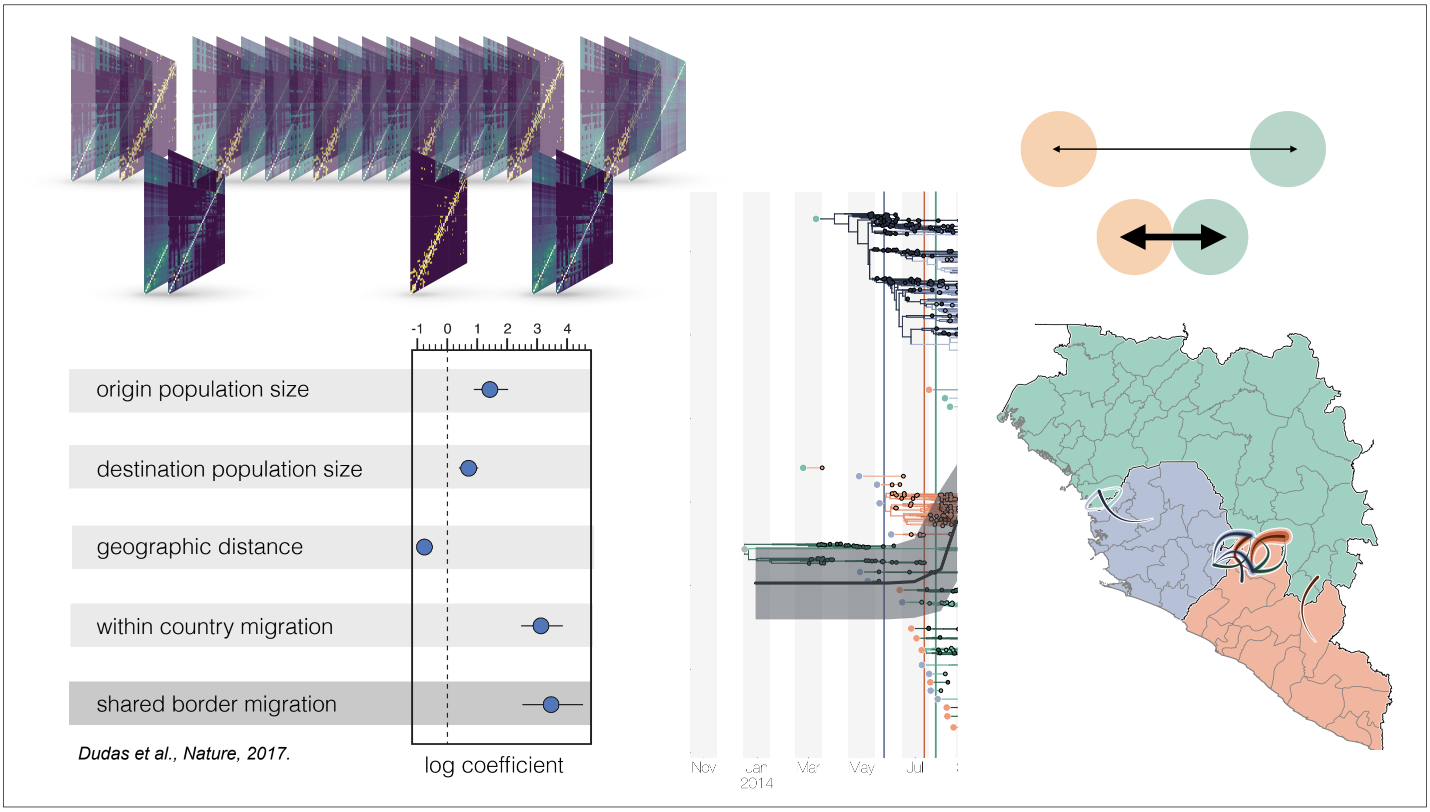
50% of virus migrations are by fewer than 100km

5% are by greater than 340km

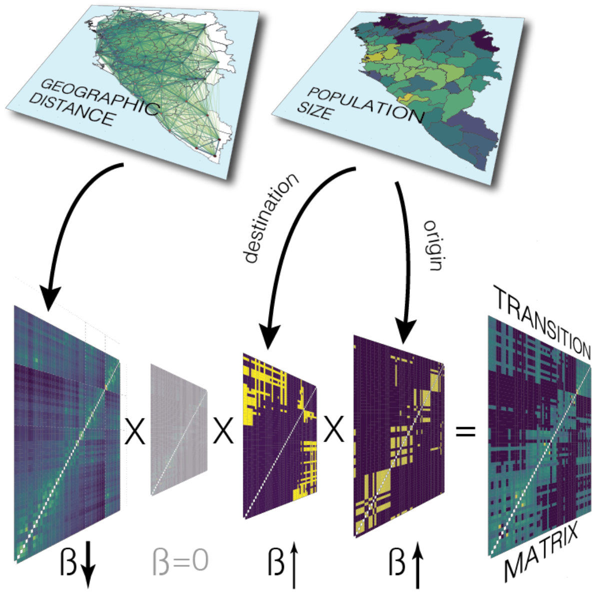


$$\log \Lambda_{ijt} = \mathbf{X}'_{ijt}\boldsymbol{\beta}(t)$$

Dudas et al., Nature, 2017.



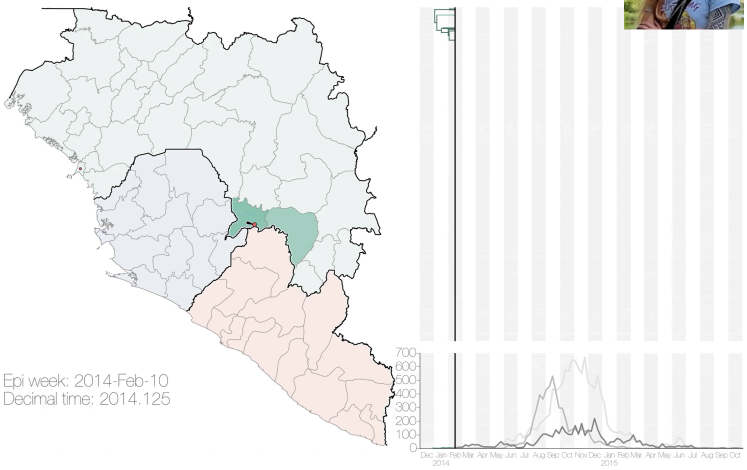
Dudas et al., Nature, 2017.



Grubaugh et al., Nat Micr, 2019.

SciPy John Hunter Excellence in Plotting Contest

winning entry



CTMC: parameters ~ (locations)2

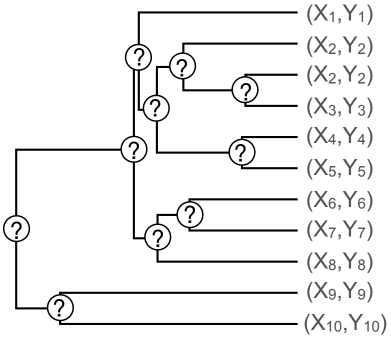
GLM-CTMC: parameters ~ predictors

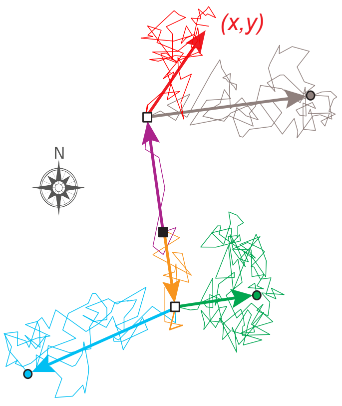
overfitting: $\log \Lambda_{ij} = \mathbf{X}'_{ij}\boldsymbol{\delta}\boldsymbol{\beta} + \epsilon_i + \epsilon_j$

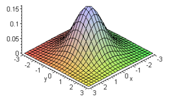
computation: BEAGLE

Dudas et al., Nature, 2017.

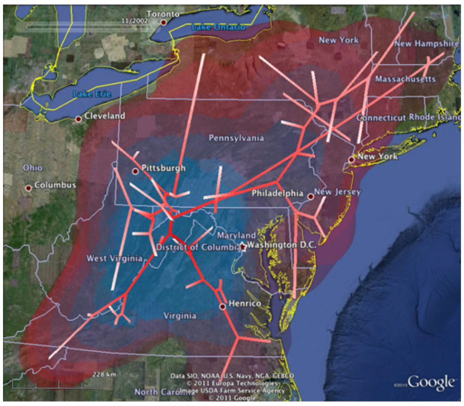
Continuous phylogenetic diffusion: Brownian motion process



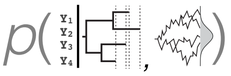




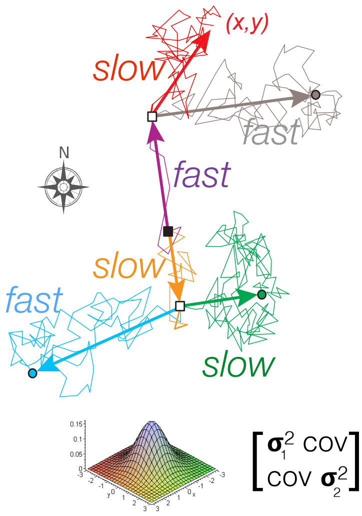
$$\begin{bmatrix} \boldsymbol{\sigma}_1^2 & \text{cov} \\ \text{cov} & \boldsymbol{\sigma}_2^2 \end{bmatrix}$$



Lemey et al., MBE, 2010.



Continuous phylogenetic diffusion: *relaxed* Brownian motion process



$$\begin{bmatrix} \boldsymbol{\sigma}_1^2 & \mathrm{cov} \\ \mathrm{cov} & \boldsymbol{\sigma}_2^2 \end{bmatrix}$$

FIGHT the BITE!

WEST NILE VIRUS

Take the bite out of camping

Evening through morning:
- Use bug spray with DEET
- Cover up when bugs are biting
- Wear long pants and long sleeves
- Use a tent with screened doors and windows

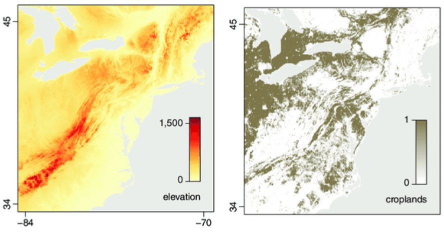
Call your doctor if you have fever, headache or stiff neck

For more information, visit www.health.utah.gov/wnv

Utah Department of Health ©2003

Pybus et al. 2012. PNAS.

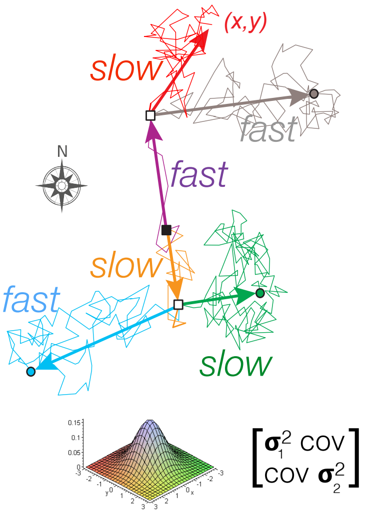
Continuous phylogenetic diffusion: landscape phylogeography



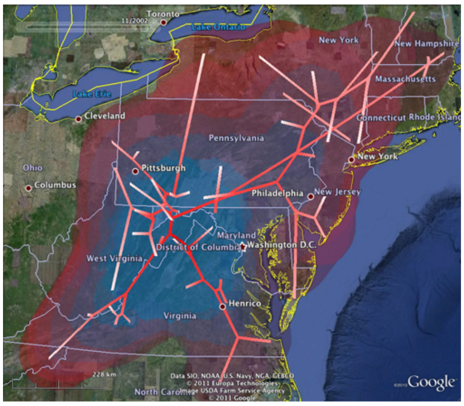
Genetics and population analysis

SERAPHIM: studying environmental rasters and phylogenetically-informed movements

Simon Dellicour[1,*], Rebecca Rose[2], Nuno R. Faria[3], Philippe Lemey[1], Oliver G. Pybus[3,*]



$$\begin{bmatrix} \boldsymbol{\sigma}_1^2 & \mathrm{cov} \\ \mathrm{cov} & \boldsymbol{\sigma}_2^2 \end{bmatrix}$$

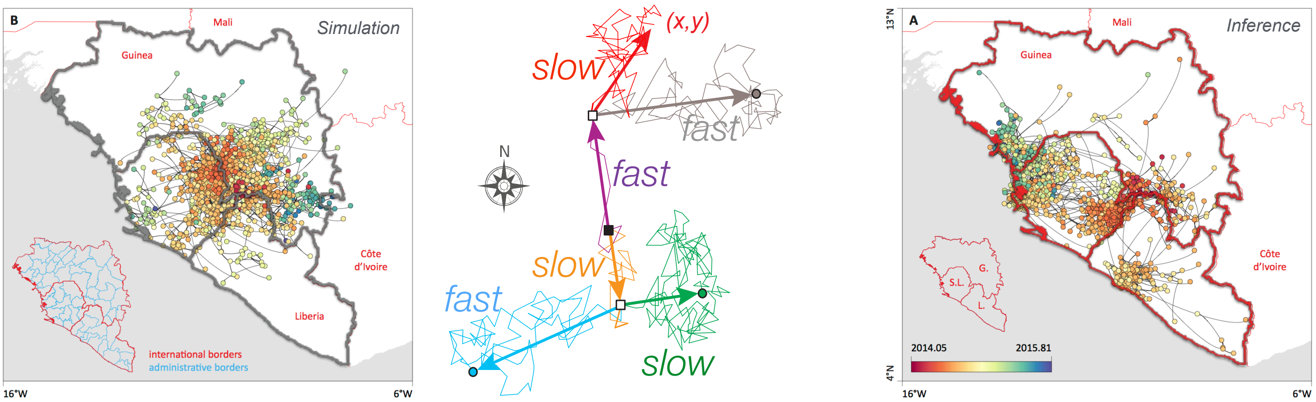


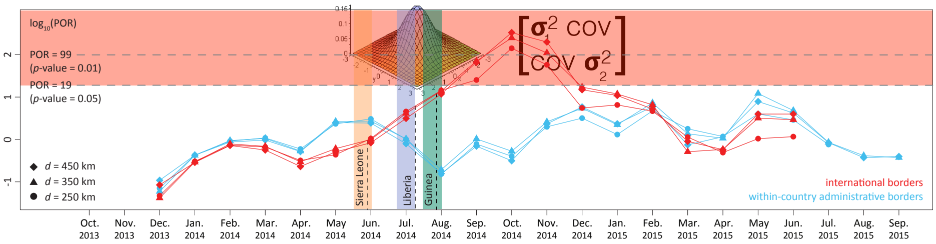
Lemey et al., MBE, 2010.

drivers?

https://github.com/sdellicour/seraphim

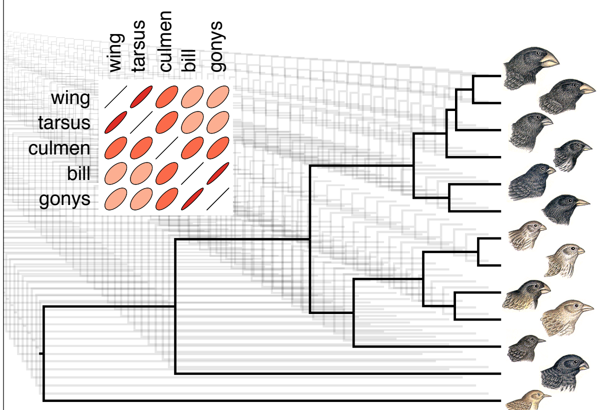
Continuous phylogenetic diffusion: landscape phylogeography





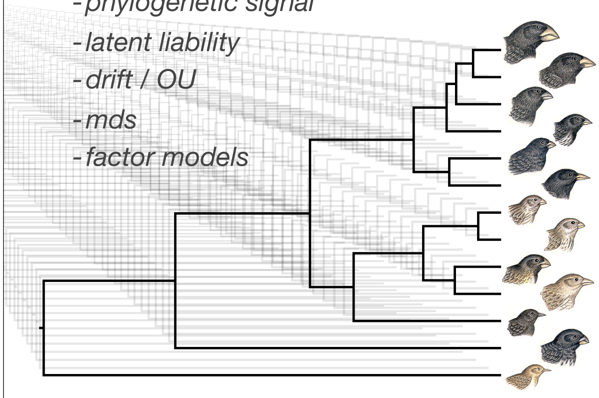
Dellicour et al.,
Nat. Comm, 2018

Continuous phylogenetic diffusion: comparative methods



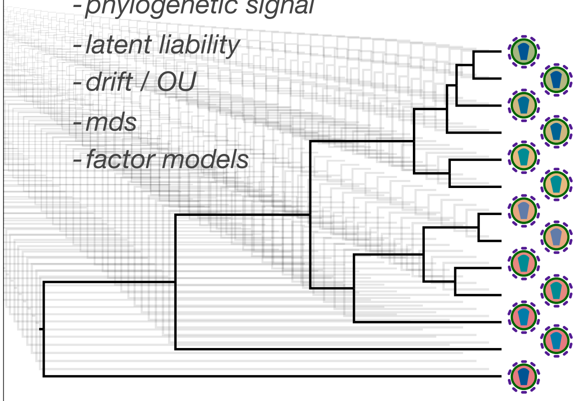
Continuous phylogenetic diffusion: comparative methods

- *missing data*
- *repeated measurements*
- *phylogenetic signal*
- *latent liability*
- *drift / OU*
- *mds*
- *factor models*



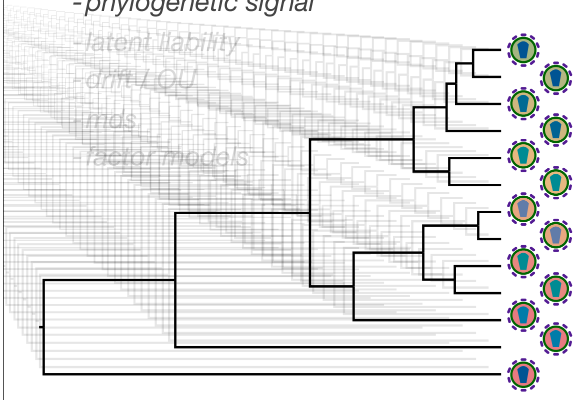
Continuous phylogenetic diffusion: comparative methods

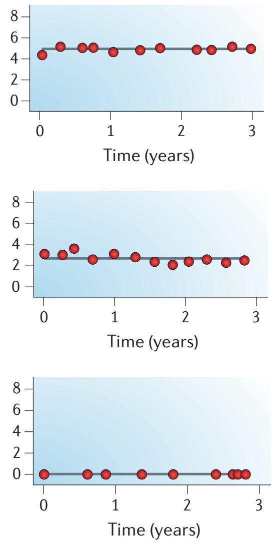
- *missing data*
- *repeated measurements*
- *phylogenetic signal*
- *latent liability*
- *drift / OU*
- *mds*
- *factor models*



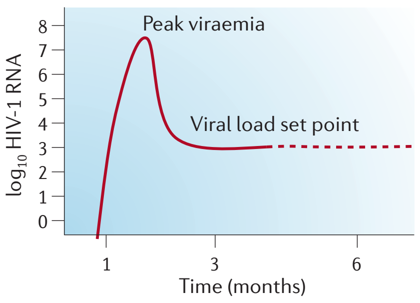
Continuous phylogenetic diffusion: comparative methods

- missing data
- repeated measurements
- *phylogenetic signal*
- latent liability
- drift / OU
- mds
- factor models





‘heritability’ of infection traits?

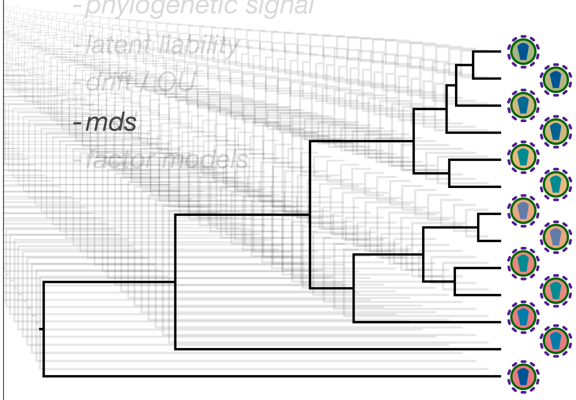


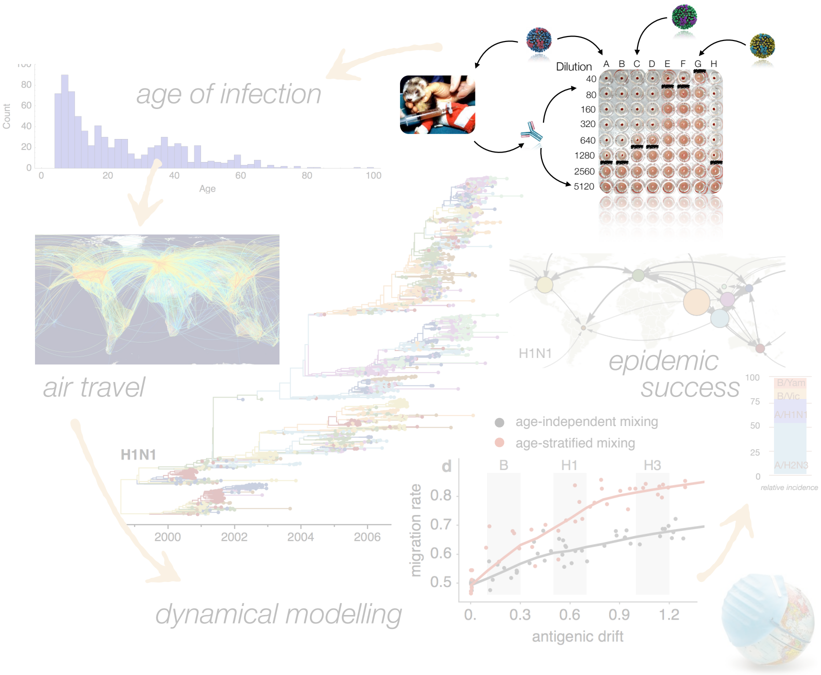
Vrancken et al., MEE, 2015

Continuous phylogenetic diffusion: comparative methods

antigenic evolution

- missing data
- repeated measurements
- phylogenetic signal
- latent liability
- drift / OU
- *mds*
- factor models





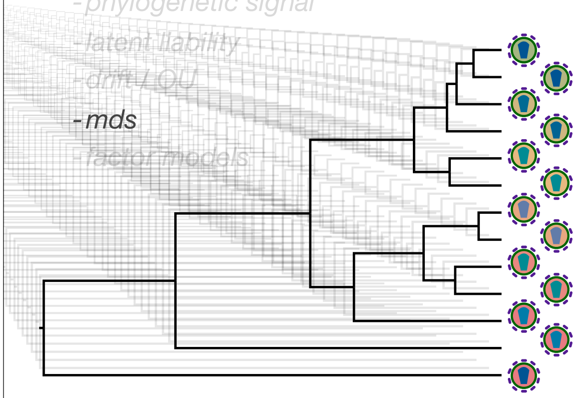
Bedford *et al.*, elife, 2014

Bedford *et al.*, Nature, 2015

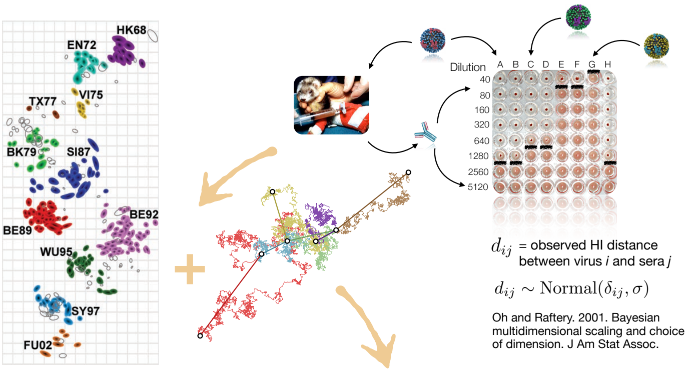
Continuous phylogenetic diffusion: comparative methods

antigenic evolution

- *missing data*
- *repeated measurements*
- *phylogenetic signal*
- *latent liability*
- *drift / OU*
- *mds*
- *factor models*



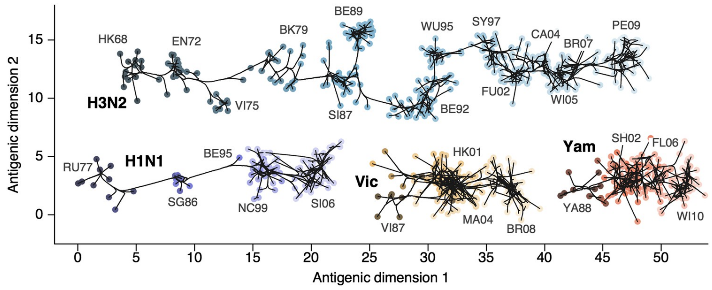
Bedford *et al.*, elife, 2014



d_{ij} = observed HI distance between virus *i* and sera *j*

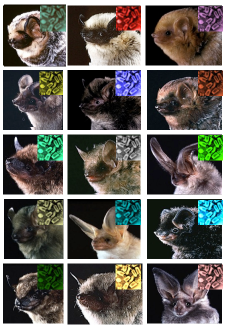
$d_{ij} \sim \mathrm{Normal}(\delta_{ij}, \sigma)$

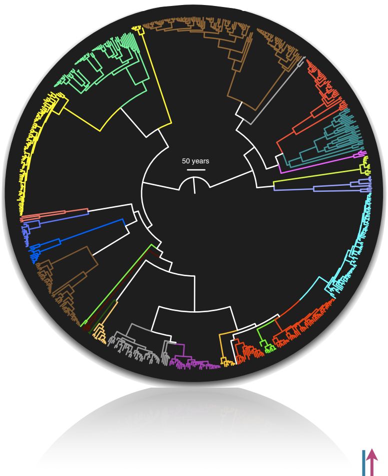
Oh and Raftery. 2001. Bayesian multidimensional scaling and choice of dimension. J Am Stat Assoc.



---

discrete phylogenetic diffusion: other traits

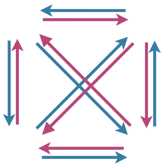




$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{i,j,1} + \beta_2 \delta_2 x_{i,j,2} + \ldots + \beta_P \delta_P x_{i,j,P}$$

predictor	***support***
host divergence	45.23
range overlap	0.35
roost structure overlap	1.45
Δ wing aspect ratio	0.14
Δ wing loading	0.23
Δ body size	0.42

Streicker et al., Science, 2012
Faria et al., Phil Roy Soc B, 2014



Phylogeography: discrete or continuous?

sampling discrete ⟷ continuous

ancestral locations
sampling locations
anywhere

geography
abstraction
flexible

dispersal ~ f(distance)
distance = euclidean

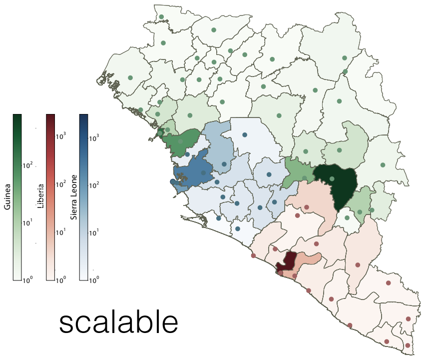
hypothesis testing
GLM

Dellicour,
Bioinformatics,
2016

Phylogeography: discrete or continuous?

sampling discrete ⟷ continuous

BIAS!!



scalable

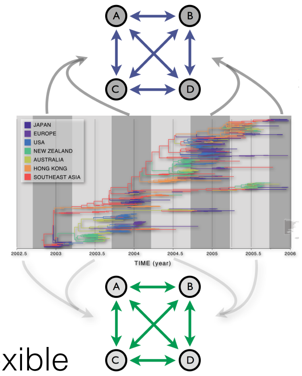
ancestral locations
sampling locations
anywhere

geography
abstraction
flexible

dispersal ~ f(distance)
distance = euclidean

hypothesis testing
GLM

Dellicour,
Bioinformatics,
2016

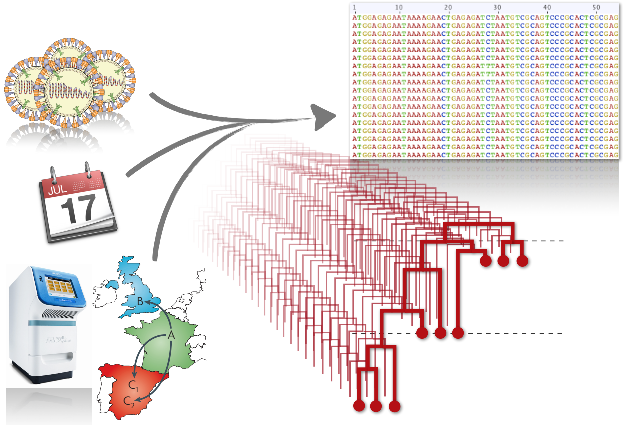


flexible

Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10

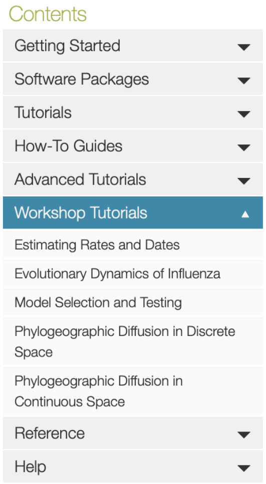
Marc A Suchard ✉, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond ✉, Andrew Rambaut ✉ Author Notes

Virus Evolution, Volume 4, Issue 1, 1 January 2018, vey016,
https://doi.org/10.1093/ve/vey016
Published: 08 June 2018



website: *beast.community*

source code: *https://github.com/beast-dev/beast-mcmc*

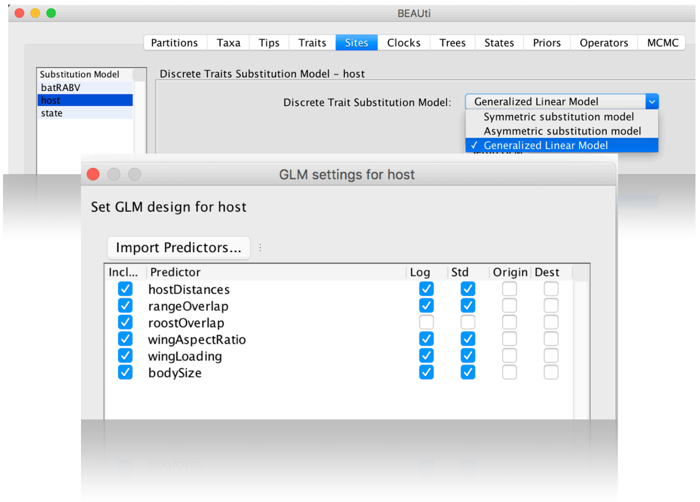

Contents
- Getting Started
- Software Packages
- Tutorials
- How-To Guides
- Advanced Tutorials
- Workshop Tutorials
  - Estimating Rates and Dates
  - Evolutionary Dynamics of Influenza
  - Model Selection and Testing
  - Phylogeographic Diffusion in Discrete Space
  - Phylogeographic Diffusion in Continuous Space
- Reference
- Help

veme: *https://rega.kuleuven.be/cev/veme-workshop/2019*

Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10

Marc A Suchard ✉, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond ✉, Andrew Rambaut ✉ Author Notes

Virus Evolution, Volume 4, Issue 1, 1 January 2018, vey016,
https://doi.org/10.1093/ve/vey016
Published: 08 June 2018



Discrete Traits Substitution Model – host

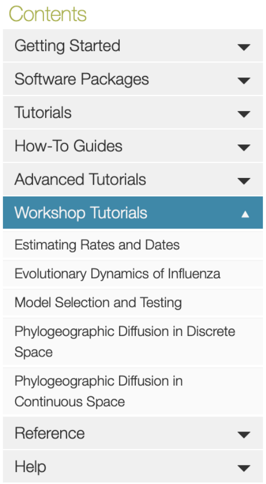
Discrete Trait Substitution Model: ✓ Generalized Linear Model

Set GLM design for host

Incl...	Predictor	Log	Std	Origin	Dest
☑	hostDistances	☑	☑	☐	☐
☑	rangeOverlap	☑	☑	☐	☐
☑	roostOverlap	☐	☐	☐	☐
☑	wingAspectRatio	☑	☑	☐	☐
☑	wingLoading	☑	☑	☐	☐
☑	bodySize	☑	☑	☐	☐

website: *beast.community*

source code: *https://github.com/beast-dev/beast-mcmc*


Contents
- Getting Started
- Software Packages
- Tutorials
- How-To Guides
- Advanced Tutorials
- Workshop Tutorials
  - Estimating Rates and Dates
  - Evolutionary Dynamics of Influenza
  - Model Selection and Testing
  - Phylogeographic Diffusion in Discrete Space
  - Phylogeographic Diffusion in Continuous Space
- Reference
- Help

veme: *https://rega.kuleuven.be/cev/veme-workshop/2019*