Exploring the origin and spread of epidemics with phylogeography

Molecular Epidemiology of Infectious Diseases
Lecture 4

February 2nd, 2026

Phylogeography reconstructs the movement of lineages through time and space

Revisiting the origins of *P. infestans*

The oomycete pathogen *Phytophthora infestans* causes potato late blight.

The HERB-1 strain of *P. infestans* caused the 19thC Irish potato famine.

Two alternative origins have been intensely debated.

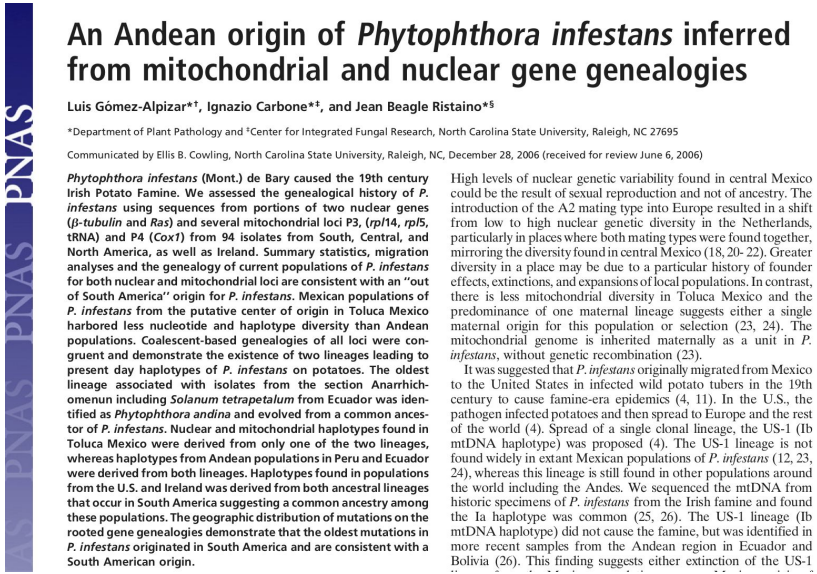
apsnet.org

An Andean origin

Native *Solanum* hosts are infected with *P. infestans* and its close relative *P. andina*.

Coalescent analysis shows ancestral lineages most likely in the Andean region of South America (Gómez-Alpizar et al., 2007).

Famine lineage may have first spread to the US and then to Europe.



PNAS PNAS PNAS

An Andean origin of *Phytophthora infestans* inferred from mitochondrial and nuclear gene genealogies

Luis Gómez-Alpizar*†, Ignazio Carbone*‡, and Jean Beagle Ristaino*§

*Department of Plant Pathology and ‡Center for Integrated Fungal Research, North Carolina State University, Raleigh, NC 27695

Communicated by Ellis B. Cowling, North Carolina State University, Raleigh, NC, December 28, 2006 (received for review June 6, 2006)

Phytophthora infestans **(Mont.) de Bary caused the 19th century Irish Potato Famine. We assessed the genealogical history of *P. infestans* using sequences from portions of two nuclear genes (*β-tubulin* and *Ras*) and several mitochondrial loci P3, (*rpl14*, *rpl5*, tRNA) and P4 (*Cox1*) from 94 isolates from South, Central, and North America, as well as Ireland. Summary statistics, migration analyses and the genealogy of current populations of *P. infestans* for both nuclear and mitochondrial loci are consistent with an "out of South America" origin for *P. infestans*. Mexican populations of *P. infestans* from the putative center of origin in Toluca Mexico harbored less nucleotide and haplotype diversity than Andean populations. Coalescent-based genealogies of all loci were congruent and demonstrate the existence of two lineages leading to present day haplotypes of *P. infestans* on potatoes. The oldest lineage associated with isolates from the section Anarrhichomenun including *Solanum tetrapetalum* from Ecuador was identified as *Phytophthora andina* and evolved from a common ancestor of *P. infestans*. Nuclear and mitochondrial haplotypes found in Toluca Mexico were derived from only one of the two lineages, whereas haplotypes from Andean populations in Peru and Ecuador were derived from both lineages. Haplotypes found in populations from the U.S. and Ireland was derived from both ancestral lineages that occur in South America suggesting a common ancestry among these populations. The geographic distribution of mutations on the rooted gene genealogies demonstrate that the oldest mutations in *P. infestans* originated in South America and are consistent with a South American origin.**

High levels of nuclear genetic variability found in central Mexico could be the result of sexual reproduction and not of ancestry. The introduction of the A2 mating type into Europe resulted in a shift from low to high nuclear genetic diversity in the Netherlands, particularly in places where both mating types were found together, mirroring the diversity found in central Mexico (18, 20- 22). Greater diversity in a place may be due to a particular history of founder effects, extinctions, and expansions of local populations. In contrast, there is less mitochondrial diversity in Toluca Mexico and the predominance of one maternal lineage suggests either a single maternal origin for this population or selection (23, 24). The mitochondrial genome is inherited maternally as a unit in *P. infestans*, without genetic recombination (23).

It was suggested that *P. infestans* originally migrated from Mexico to the United States in infected wild potato tubers in the 19th century to cause famine-era epidemics (4, 11). In the U.S., the pathogen infected potatoes and then spread to Europe and the rest of the world (4). Spread of a single clonal lineage, the US-1 (Ib mtDNA haplotype) was proposed (4). The US-1 lineage is not found widely in extant Mexican populations of *P. infestans* (12, 23, 24), whereas this lineage is still found in other populations around the world including the Andes. We sequenced the mtDNA from historic specimens of *P. infestans* from the Irish famine and found the Ia haplotype was common (25, 26). The US-1 lineage (Ib mtDNA haplotype) did not cause the famine, but was identified in more recent samples from the Andean region in Ecuador and Bolivia (26). This finding suggests either extinction of the US-1

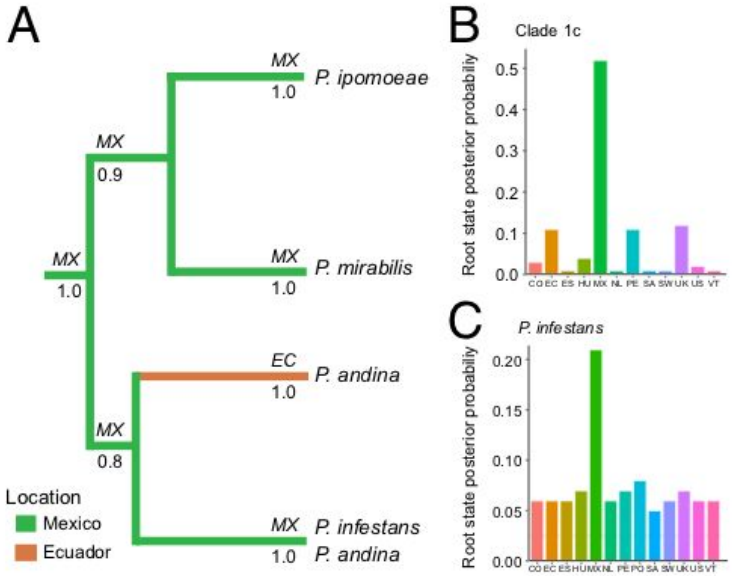
Gómez-Alpizar *et al.* (PNAS, 2007)

A central Mexican origin

Diverse, sexual populations of *P. infestans* are found in the Toluca Valley of C. Mexico.

Many tuber-bearing *Solanum* species like potato are native to the Toluca Valley.

Phylogeographic analysis of a larger dataset rooted *P. infestans* in Mexico



Goss *et al.* (PNAS, 2014)

How do we reconstruct ancestral locations?

Two ways of thinking about space

- **Discrete trait models** track the movement of lineages as jumps between isolated populations

- **Continuous trait models** track the diffusion of lineages across a connected landscape

Discrete trait phylogeographic models

Discrete trait methods model transitions between locations as a **continuous time Markov chain**, i.e. the same way we model sequence evolution at a single site.

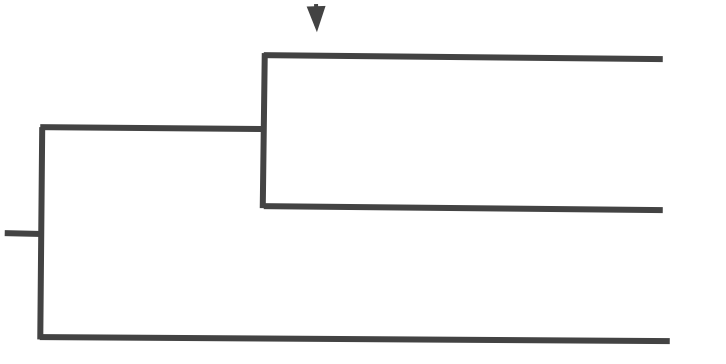
Instead of a substitution rate matrix, we have a migration rate matrix *M*:

$$M = \begin{bmatrix} -\sum_{i \neq 1}^{n} m_{1,i} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & -\sum_{i \neq 2}^{n} m_{2,i} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & -\sum_{i \neq n}^{n} m_{n,i} \end{bmatrix}$$

Because migration is modeled the same we model mutations, these models are sometimes referred to as “**mugation” models**.

Computing ancestral state probabilities

$$p(t) = \begin{bmatrix} ? \\ ? \end{bmatrix}$$



$$p(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

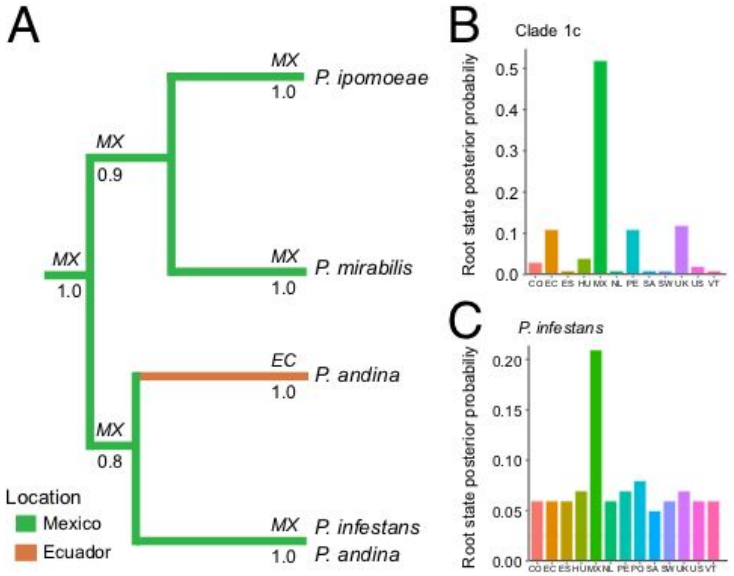
$$p(t) = e^{Mt}p(0)$$

We can easily compute ancestral state probabilities under a CTMC given our migration rate matrix ***M*** and the time elapsed along a branch ***t***.

Bayesian phylogeography

In Bayesian phylogeography, MCMC is used to sample trees with migration histories mapped onto the tree so that each node is associated with an ancestral state.

Posterior probabilities for ancestral states can then be estimated from the fraction of posterior trees sampled with a node in a given state.

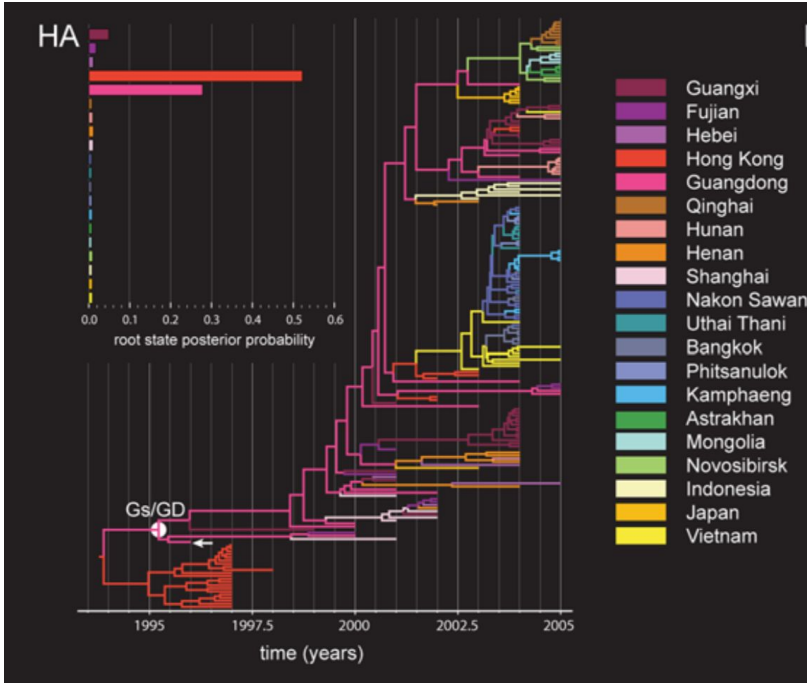


Goss *et al.* (PNAS, 2014)

Bayesian phylogeography

Bayesian methods based on CTMC efficiently reconstruct migration histories even for relatively large trees and many different sampling locations.

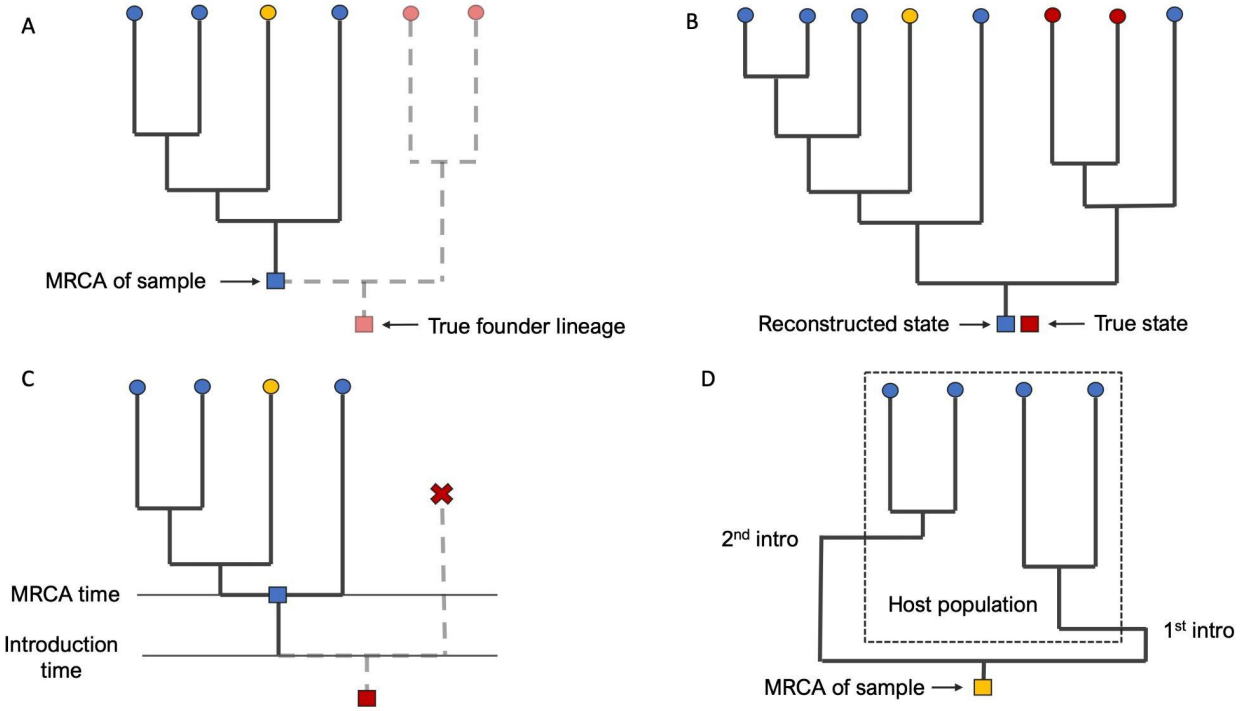
Posterior probabilities directly quantify our uncertainty about ancestral locations.



MCC tree for avian H5N1 influenza virus

Lemey *et al.* (2009)

Inferences about root state can mislead



Rasmussen and Grunwald (Phytopathology, 2021)

The curse of dimensionality

The number of migration rates we need to estimate grows quickly with *n*, the number of locations.

n = 4

$$M = \begin{bmatrix} \cdot & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & \cdot & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & \cdot & m_{3,4} \\ m_{4,1} & m_{4,2} & m_{4,3} & \cdot \end{bmatrix}$$

$$D = 4(4-1) = 12$$

n = 40

$$M = \begin{bmatrix} \cdot & m_{1,2} & \cdots & m_{1,40} \\ m_{2,1} & \cdot & \cdots & m_{2,40} \\ \vdots & \vdots & \ddots & \vdots \\ m_{40,1} & m_{40,2} & \cdots & \cdot \end{bmatrix}$$

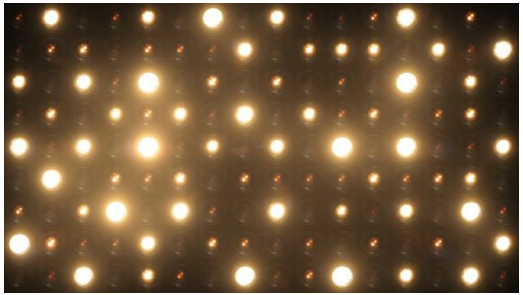
$$D = 40(40-1) = 1560$$

Moreover, it is unlikely that a migration event between each pair of locations has occurred in the history of the sample.

BSSVS

Lemey *et al.* (2009) introduced the idea of using **Bayesian Stochastic Search Variable Selection** to “sparsify” to the migration rate matrix.

$$M = \begin{bmatrix} \cdot & \delta_{1,2}m_{1,2} & \cdots & \delta_{1,n}m_{1,n} \\ \delta_{2,1}m_{2,1} & \cdot & \cdots & \delta_{2,n}m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n,1}m_{n,1} & \delta_{n,2}m_{n,2} & \cdots & \cdot \end{bmatrix}$$



Here the □$_{ij}$’s represent indicators that can either be on (□$_{ij}$ = 1) or off (□$_{ij}$ = 0).

The idea being that most transitions rarely, if ever, occur so that most migration rates can be set to zero, resulting in a much lower-dimensional model.

BSSVS and model selection

We can use BSSVS to perform model selection for us! Say we have two competing models:

M_1: Contains a non-zero rate for a specific migration rate m_{kl}

M_0: Assumes $m_{kl} = 0$, and is therefore a simpler, reduced version of M_1.

For BSSVS implemented in MCMC, the proportion of time the chain spends in the "on" state ($p_{\square=1}$) provides the posterior support (i.e. evidence) for the model containing the rate parameter over the model that does not.

We can use this information to compute **Bayes factors**.

Bayes factors

Bayes factors summarize the evidence provided by the data in favor of one model over an alternative model, similar to a likelihood ratio test:

$$B_{01} = \frac{p(D|M_1)}{p(D|M_0)}$$

Bayes factors can be thought of as the posterior odds for preferring one model over another after observing the data.

$2 \log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

Kass and Raftery (1995)

Bayes factors

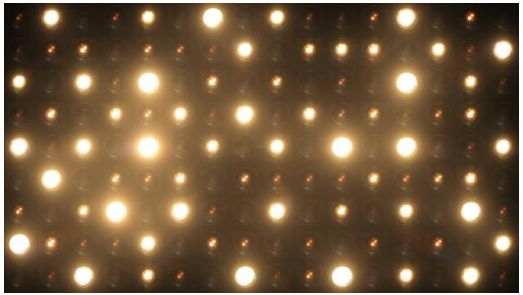
We can directly compute Bayes factors from the fraction of time the MCMC spends in the “on” state ($p_{\square=1}$) versus the “off” state ($1 - p_{\square=1}$):

$$B_{01} = \frac{p(D|M_1)}{p(D|M_0)} = \frac{p_{\delta=1}}{1 - p_{\delta=1}}$$

BSSVS

Lemey *et al.* (2009) introduced the idea of using **Bayesian Stochastic Search Variable Selection** to “sparsify” to the migration rate matrix.

$$M = \begin{bmatrix} \cdot & \delta_{1,2}m_{1,2} & \cdots & \delta_{1,n}m_{1,n} \\ \delta_{2,1}m_{2,1} & \cdot & \cdots & \delta_{2,n}m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n,1}m_{n,1} & \delta_{n,2}m_{n,2} & \cdots & \cdot \end{bmatrix}$$



Here the □$_{ij}$’s represent indicators that can either be on (□$_{ij}$ = 1) or off (□$_{ij}$ = 0).

The idea being that most transitions rarely, if ever, occur so that most migration rates can be set to zero, resulting in a much lower-dimensional model.

Phylogeography can also identify drivers of spatial spread

GLMs with predictor variables

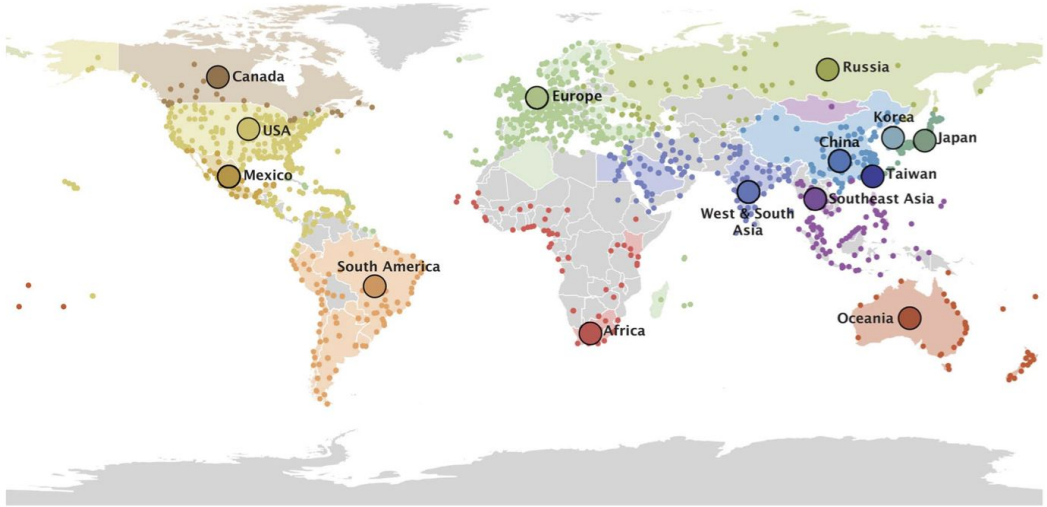
Generalized Linear Models can be used to predict migration rates based on predictor variables:

$$m_{i,j} = g(\beta_1 x_{i,j,1} + \beta_2 x_{i,j,2} + \cdots + \beta_P x_{i,j,P})$$

Here, the ***x***'s represent explanatory/predictor variables and the **β**'s regression coefficients as in a classic linear regression model.

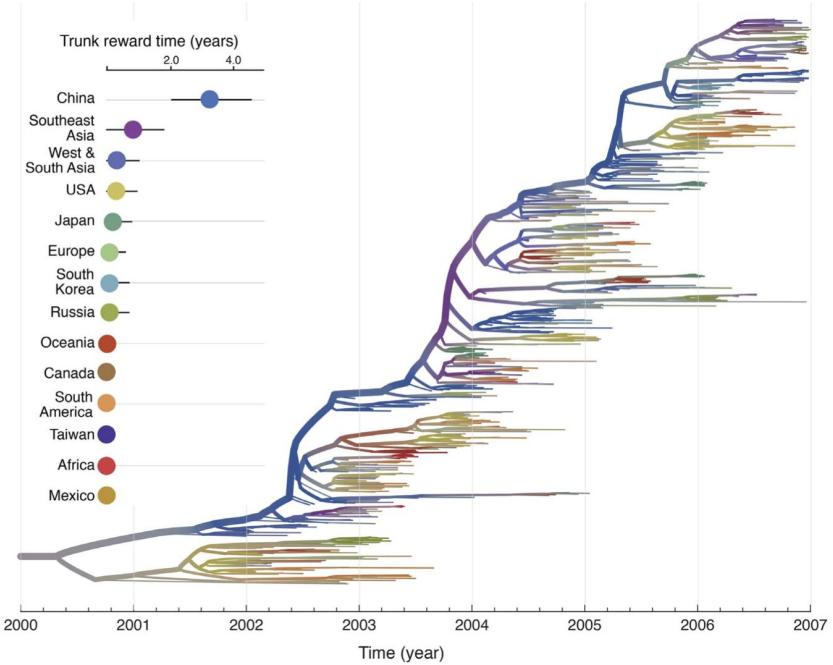
Generalized because additive linear effects can be transformed using some other function ***g*** (e.g. log transformed).

Air traffic predicts global spread of flu



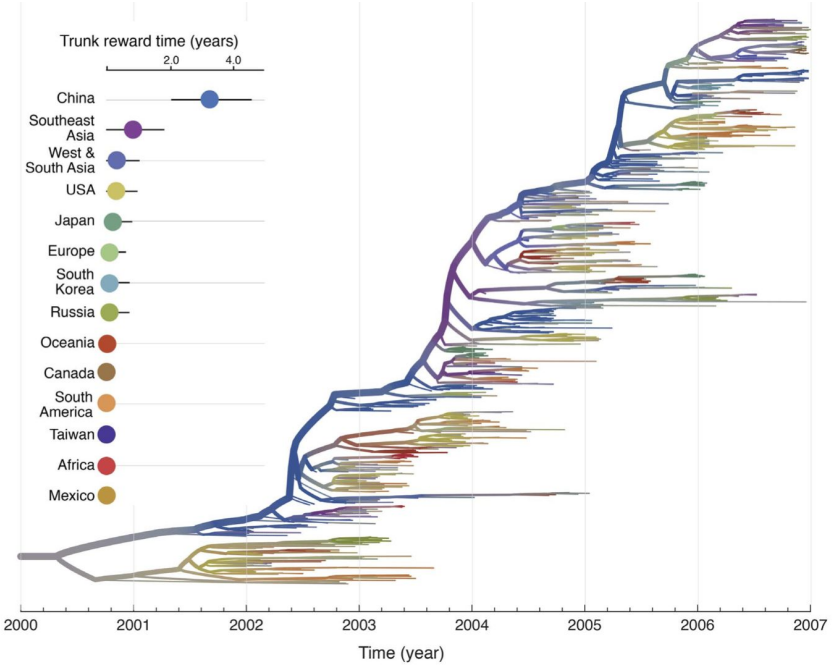
Lemey *et al.* (PLoS Pathogens, 2014)

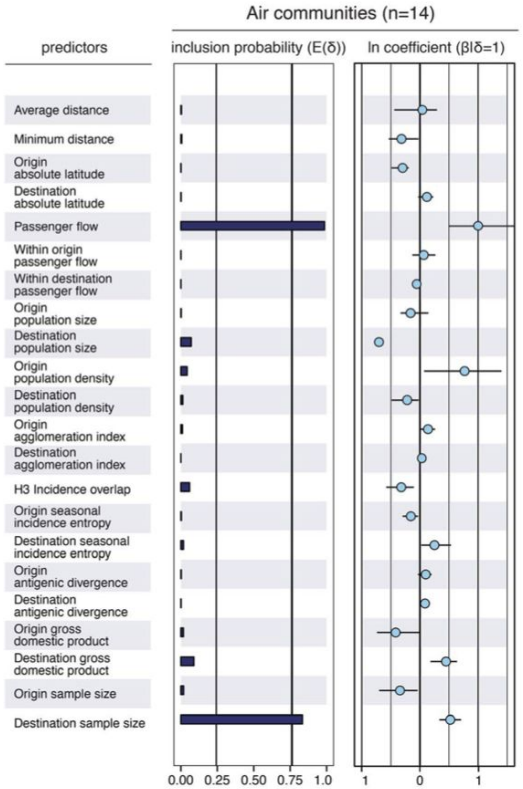
Air traffic predicts global spread of flu



Lemey *et al.* (PLoS Pathogens, 2014)

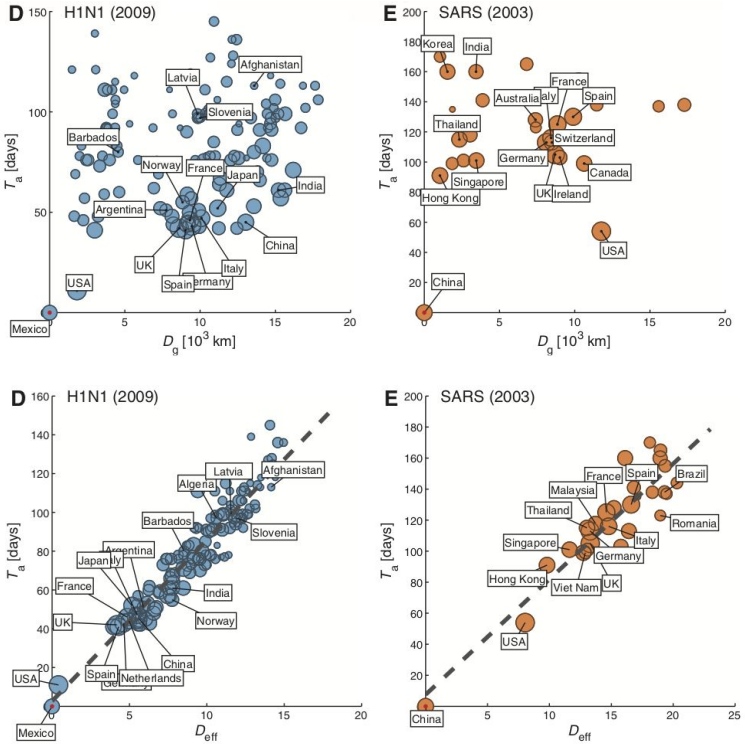
Air traffic predicts global spread of flu

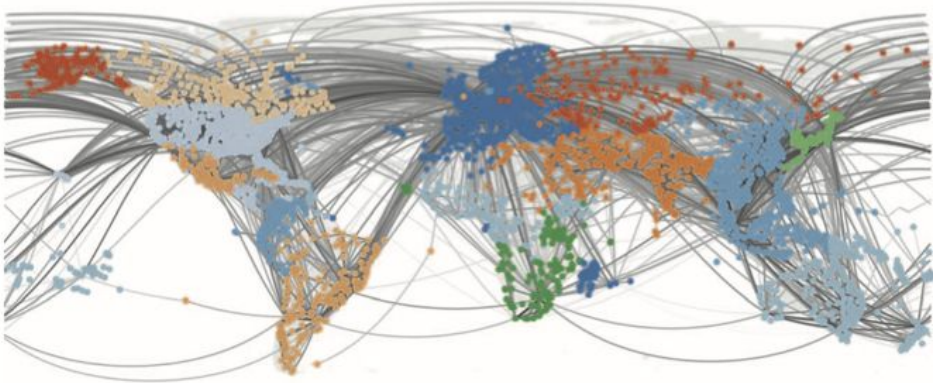




Lemey *et al.* (PLoS Pathogens, 2014)

Global transit networks predict spread





Brockmann and Helbing (Science, 2013)

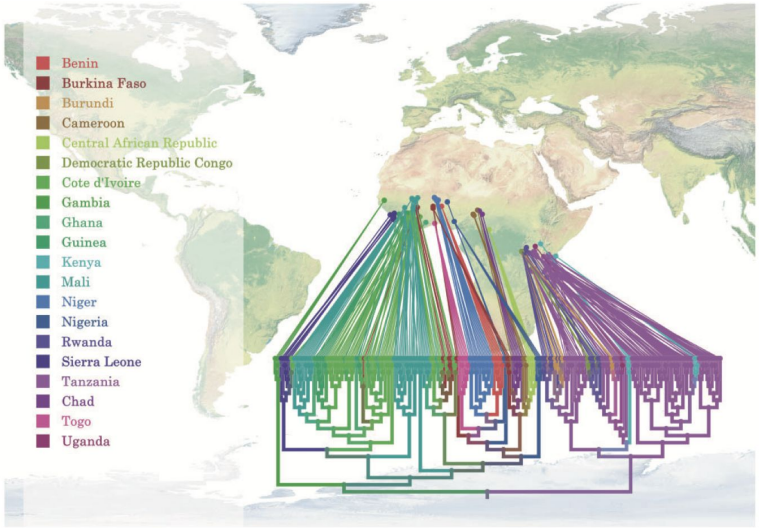
Predicting spatial spread

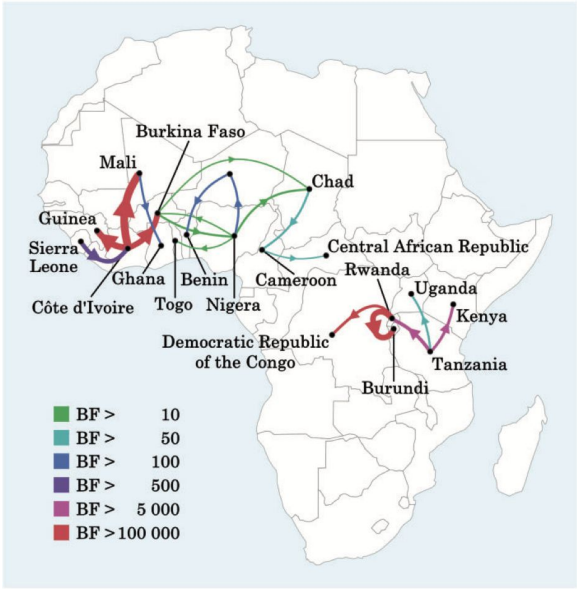
For human respiratory pathogens, global transportation networks (e.g. air traffic) strongly influence the global spread of pathogens.

What about other pathogens? What predicts pathogen spread through agricultural landscapes?

Spatial diffusion of Rice yellow mottle virus

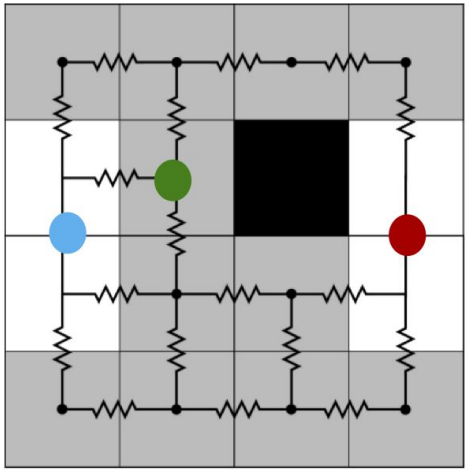
Analysis of RYMV shows how Bayes factors can be used to identify significant diffusion routes through mixed landscapes.

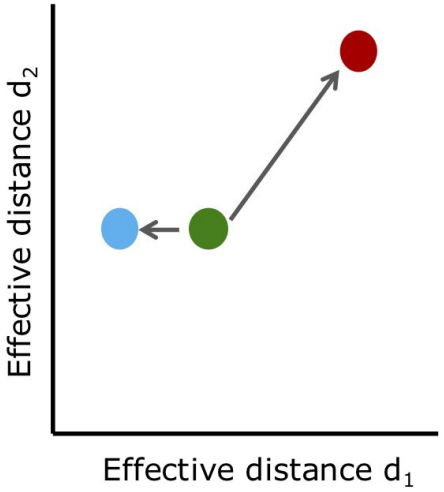




Trovão *et al.* (Virus Evo., 2015)

Effective resistance distances





Areas of high rice production = low resistance

Areas of low rice production = high resistance

Effective distance = resistance along path

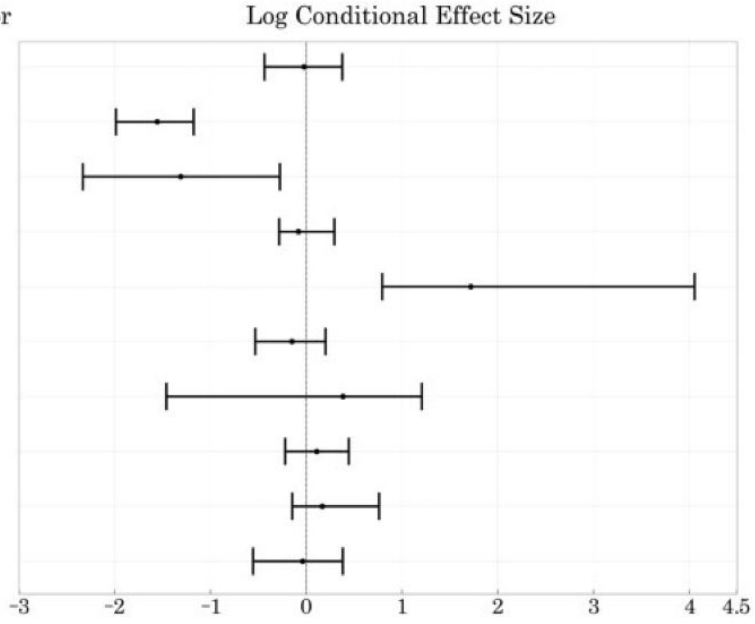
Predictors of RYMV spatial spread

Harvested rice intensity and connectivity are the main determinants of spread

Low connectivity

Production intensity at source

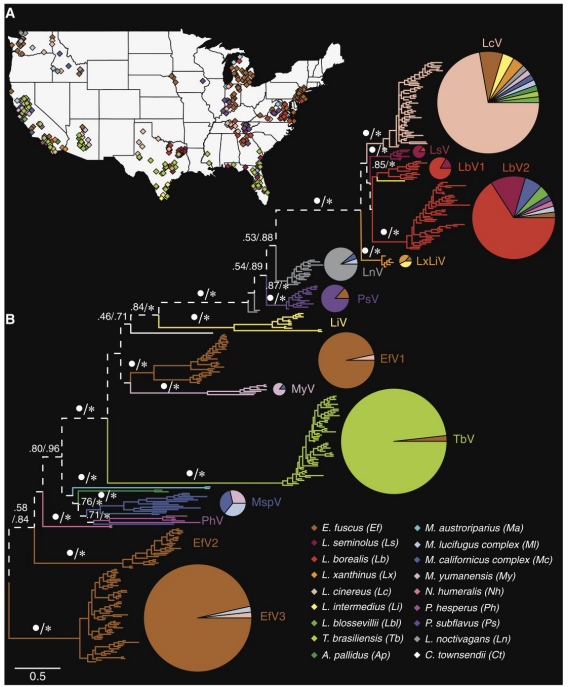
Predictors	Bayes Factor
Distance	0.0032
Resistance on harvested rice area	>9000
Origin of rice intensity in 1960	0.1873
Destination of rice intensity in 1960	0.0028
Origin of rice intensity in 1990	14.26
Destination of rice intensity in 1990	0.0038
Origin of precipitation	0.0660
Destination of precipitation	0.0019
Origin of sample size	0.0041
Destination of sample size	0.0032



Trovão *et al.* (Virus Evo., 2015)

Non-geographic phylogeography

- Subpopulations within a larger host population (e.g. risk groups)
- Different host species for multi-host pathogens



Streicker *et al.* (Science, 2010)

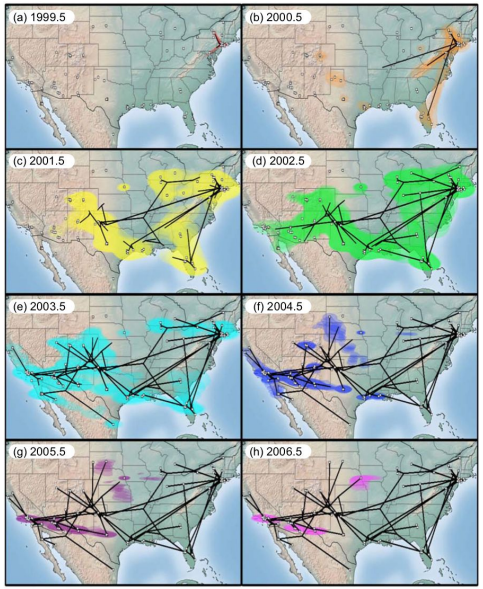
Two main approaches

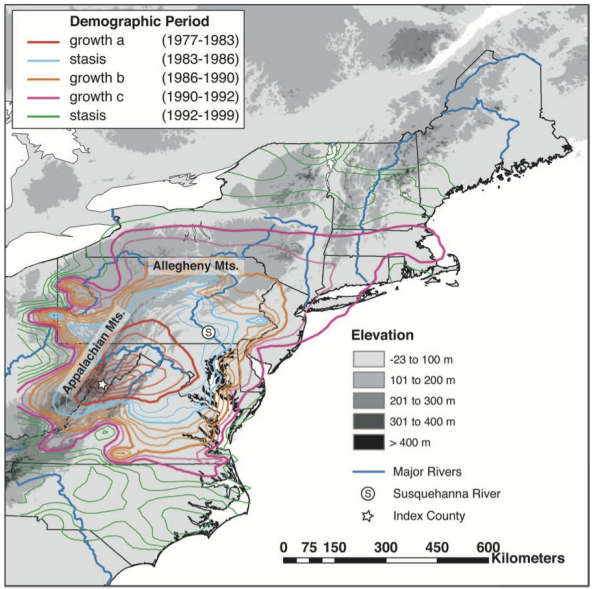
- Discrete trait methods
- Continuous trait methods

Continuous spatial diffusion

Treating space or distances as a continuous variable often makes sense for epidemics with wave-like diffusion across a landscape (e.g. wildlife diseases).

West Nile Virus spread in North America (Pybus *et al.*, 2012)

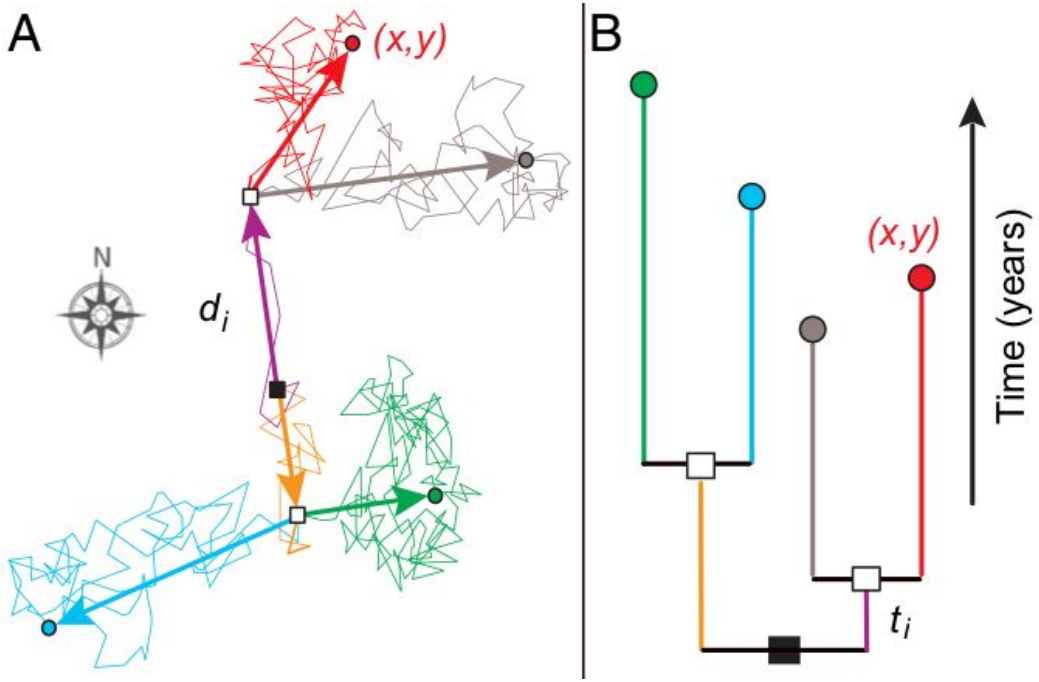




Racoon rabies in the mid-Atlantic states (Beik *et al.*, 2007)

The basic idea of spatial phylogeography

“If the dates and locations of all phylogenetic nodes are known or posited, then each branch represents a conditionally independent trajectory of viral movement, defined by a start location, end location, and duration.”



Pybus *et al.* (PNAS, 2012)

Quantifying spatial spread

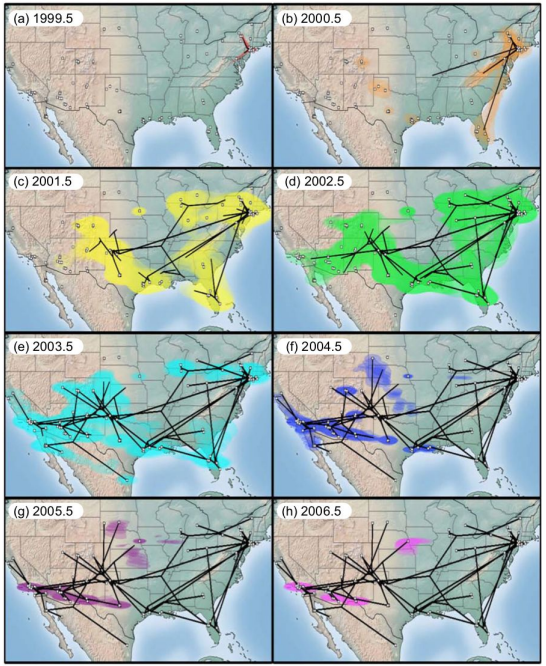
We can then estimate the wavefront velocity, i.e. the rate at which the epidemic front spreads from the origin to its furthest extent.

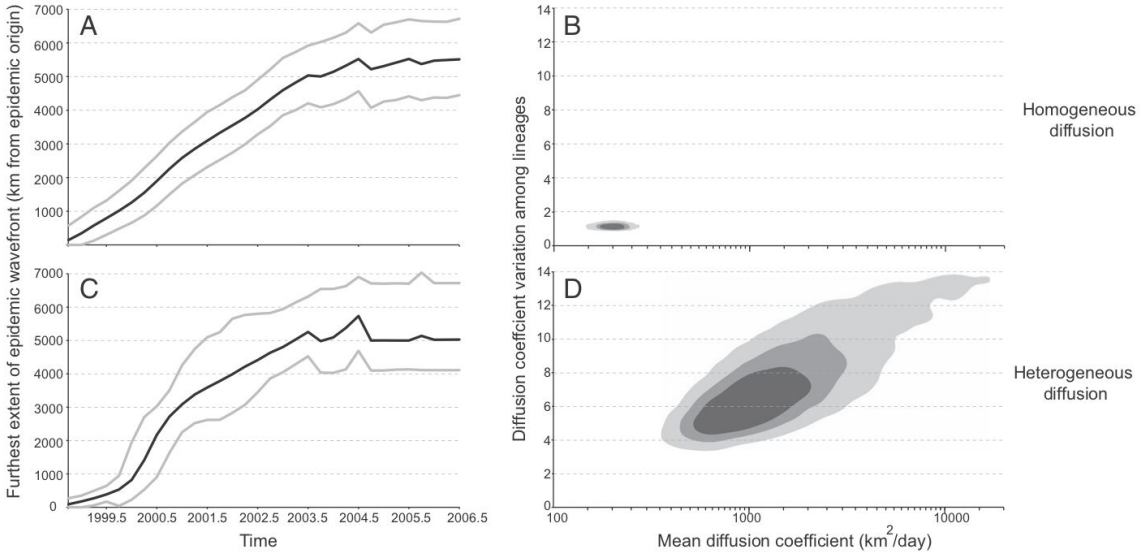
We can also estimate the mean diffusion coefficient *D* based on the distance d_i that each lineage *i* has traveled over time *t* along each branch:

$$D \approx \frac{1}{n} \sum_{i=1}^{n} \frac{d_i^2}{4t_i}$$

Pybus *et al.* (2012)

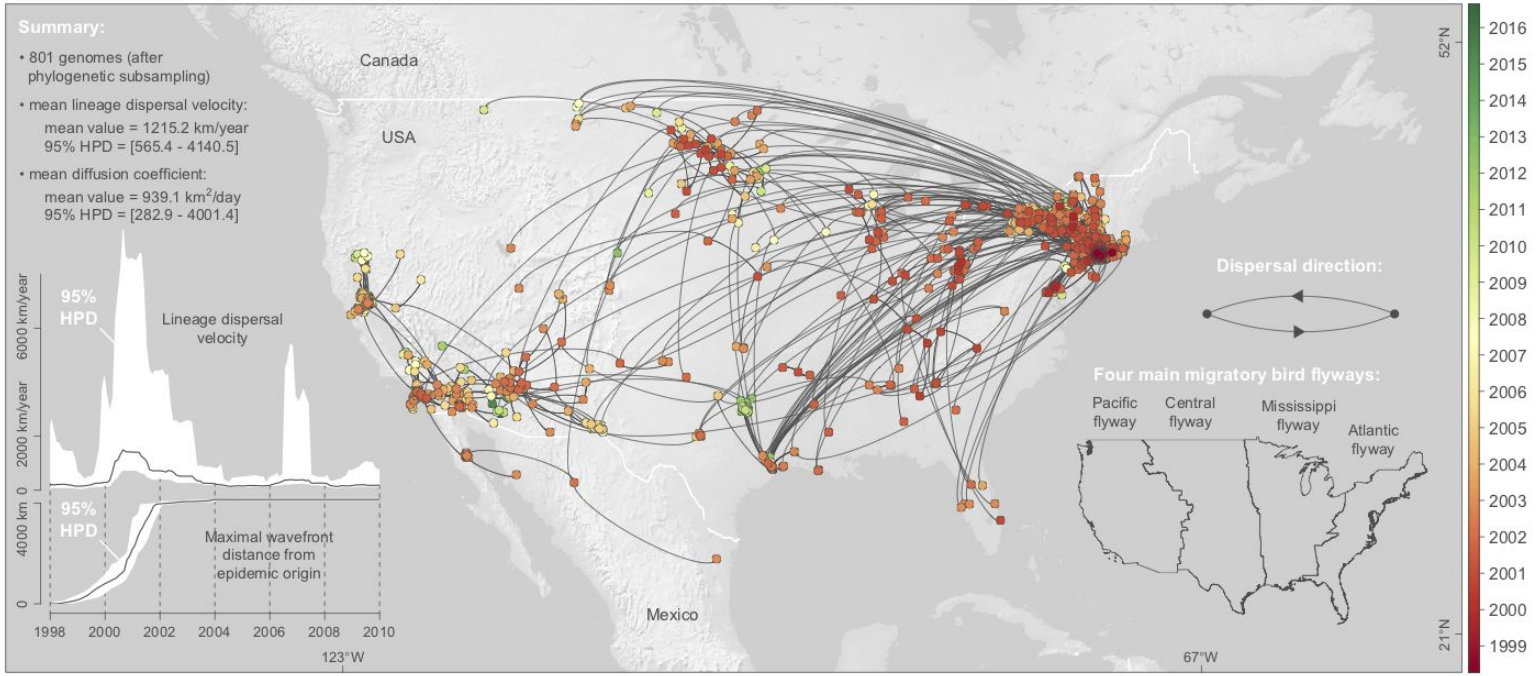
Spatial diffusion of West Nile Virus





Pybus *et al.* (2012)

Spatial diffusion of West Nile Virus



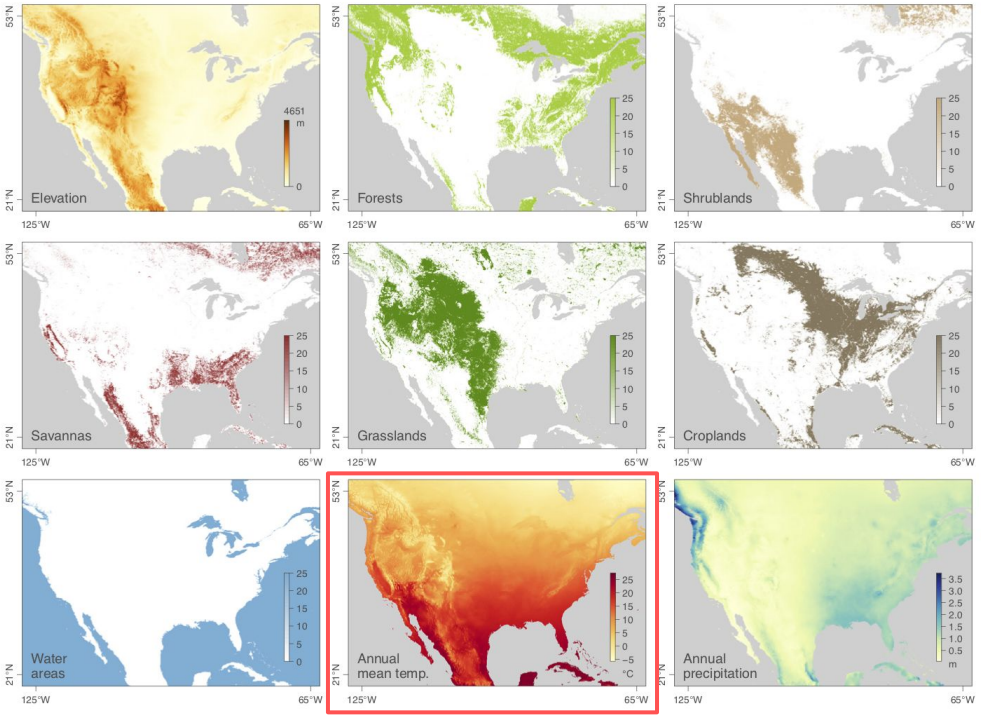
Dellicour *et al.* (2019)

Explaining variation in diffusion rates

Annual mean temperature was the only significant explanatory variable for spatial diffusion rates.

Higher temps likely accelerate WNV transmission by mosquitoes due to shorter extrinsic incubation periods.

Shows how continuous models can also be used to explore the drivers of spread.



Dellicour *et al.* (2019)

Conclusions

Phylogeography can be used to reconstruct the historical origins of epidemics and new pathogen strains.

Inferences about ancestral locations are often plagued by high uncertainty due to limited and non-representative sampling. Appropriate measures of uncertainty are therefore extremely important.

Phylogeography can also be used to identify factors driving spatial spread in contemporary pathogen populations.