Bayesian Phylogeography

Remco R. Bouckaert
remco@cs.auckland.ac.nz

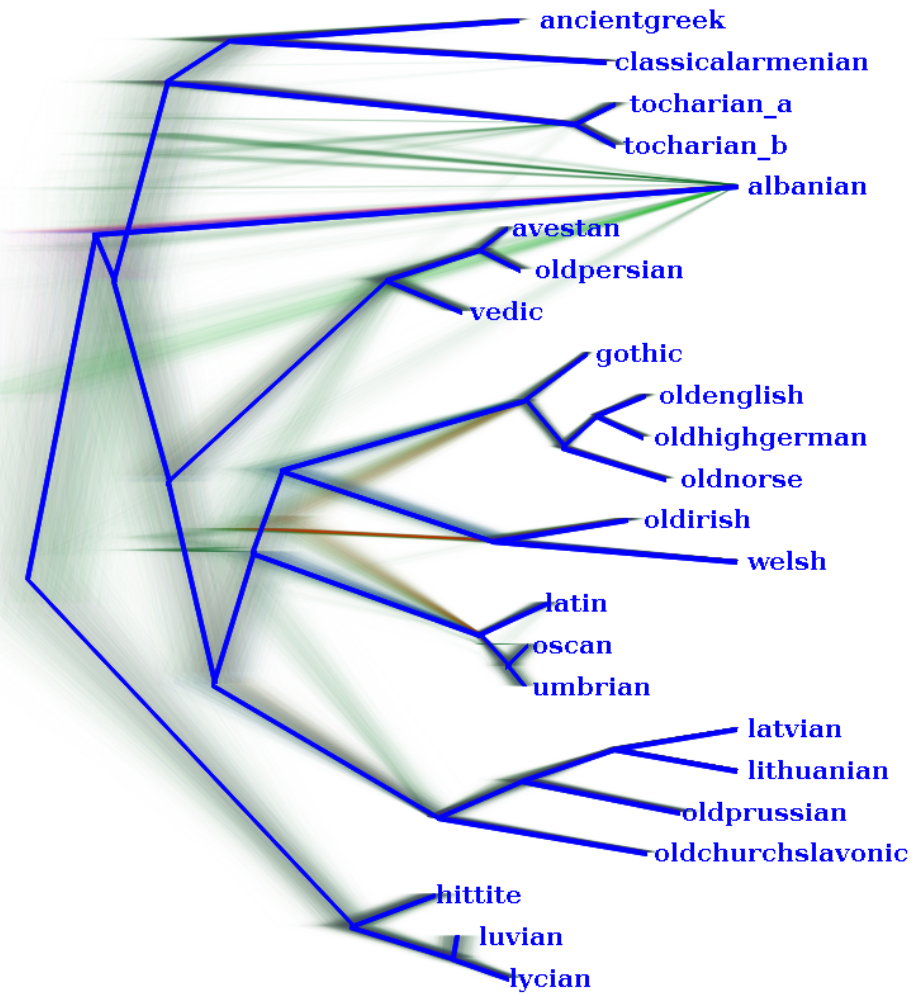
Centre of Computational Evolution
Department of Computer Science, University of Auckland
&
Max Planck Institute for the Science of Human History

Seili, September 2017

Bayesian phylogenetics

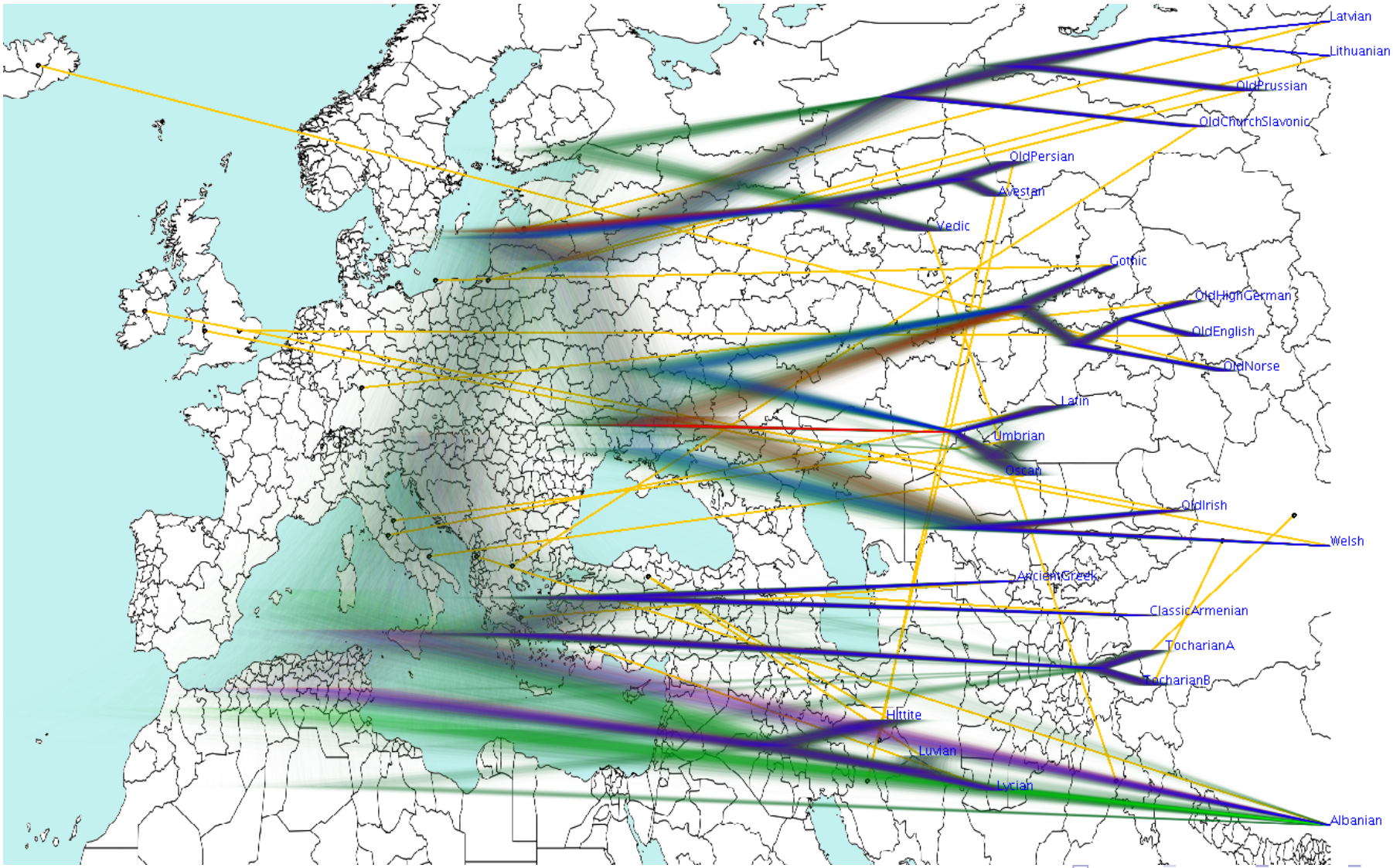
- Posterior $\propto$ prior $\times$ likelihood
- Prior info easy to incorporate
- Framework for adding sources of information from different sources – geography, WALS, cognate data, etc.
- Think in distributions: there is no "The Tree", only a distribution over trees

Ringe DensiTree

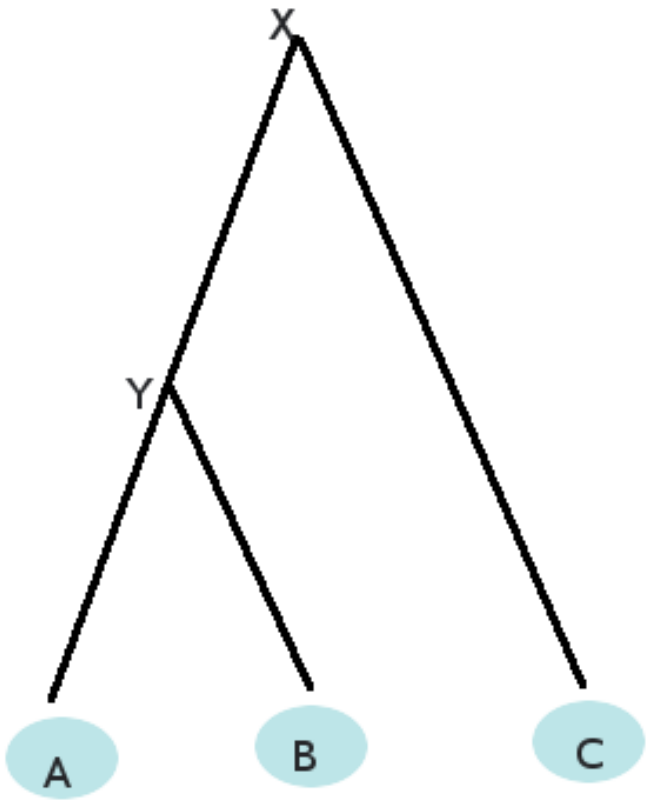


Basic Phylogeography

Link tree tips to locations + story telling

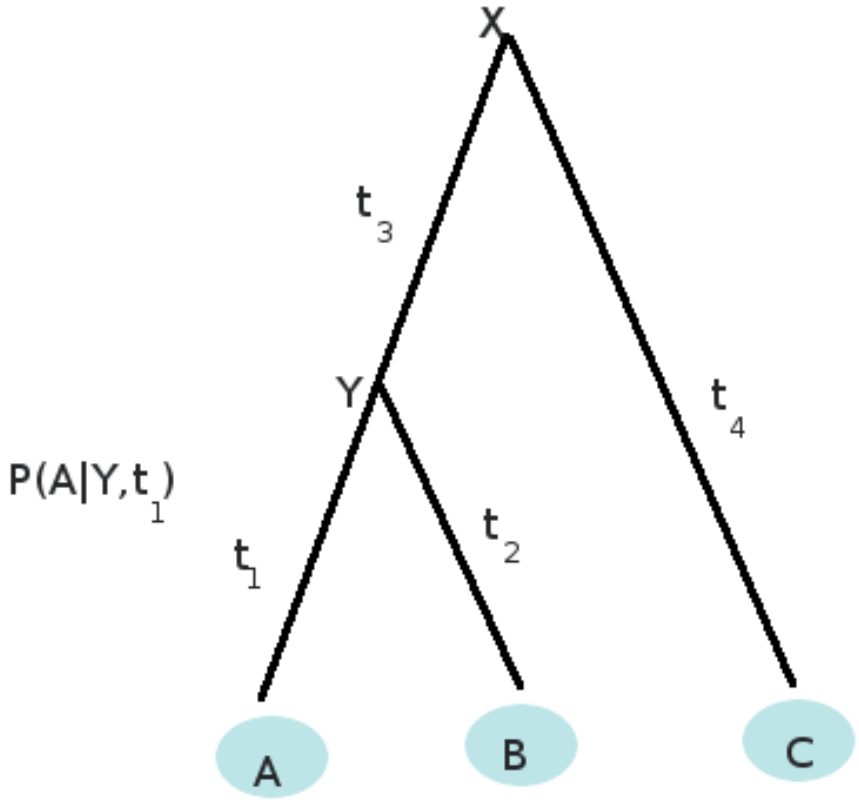


Phylogeography conditioned on tree



Locations: A, B, C, X, Y

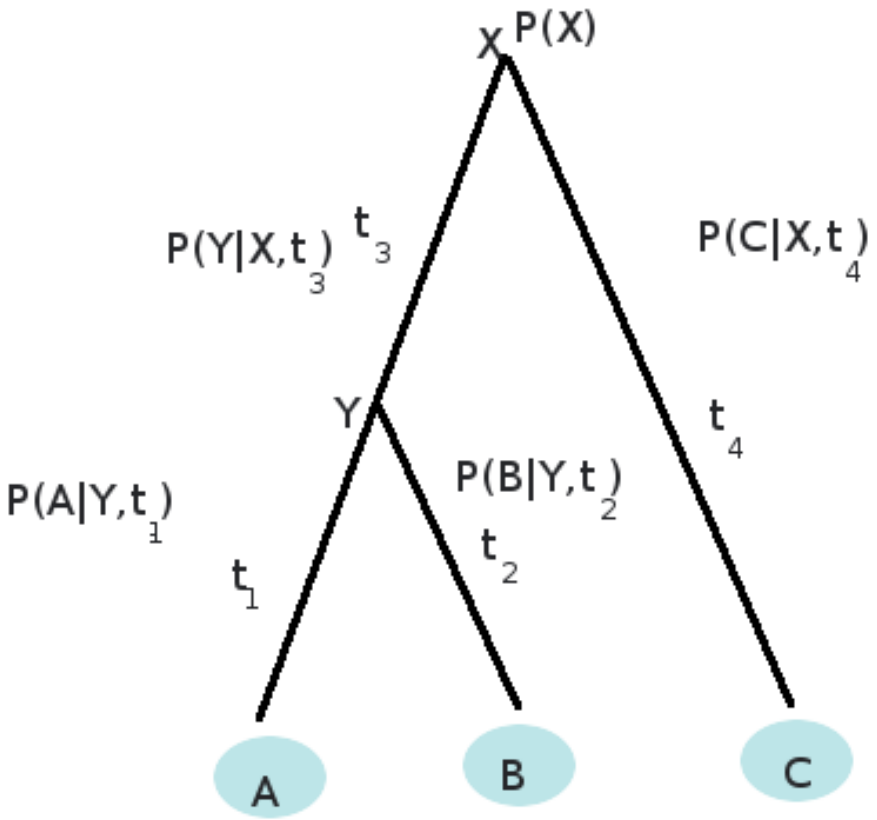
Phylogeography conditioned on tree



Locations: A, B, C, X, Y

Branch lengths: t_1, t_2, t_3, t_4

Phylogeography conditioned on tree



Locations: A, B, C, X, Y

Branch lengths: t_1, t_2, t_3, t_4

Probabilities: $\forall x \in \{A, B, C, Y\} P(x|parent(x), t_x)$ and $P(X)$

Ancestral state reconstruction (BEAST_CLASSIC package)

- requires specifying regions
- many parameters ($O(n^2)$ with n states)

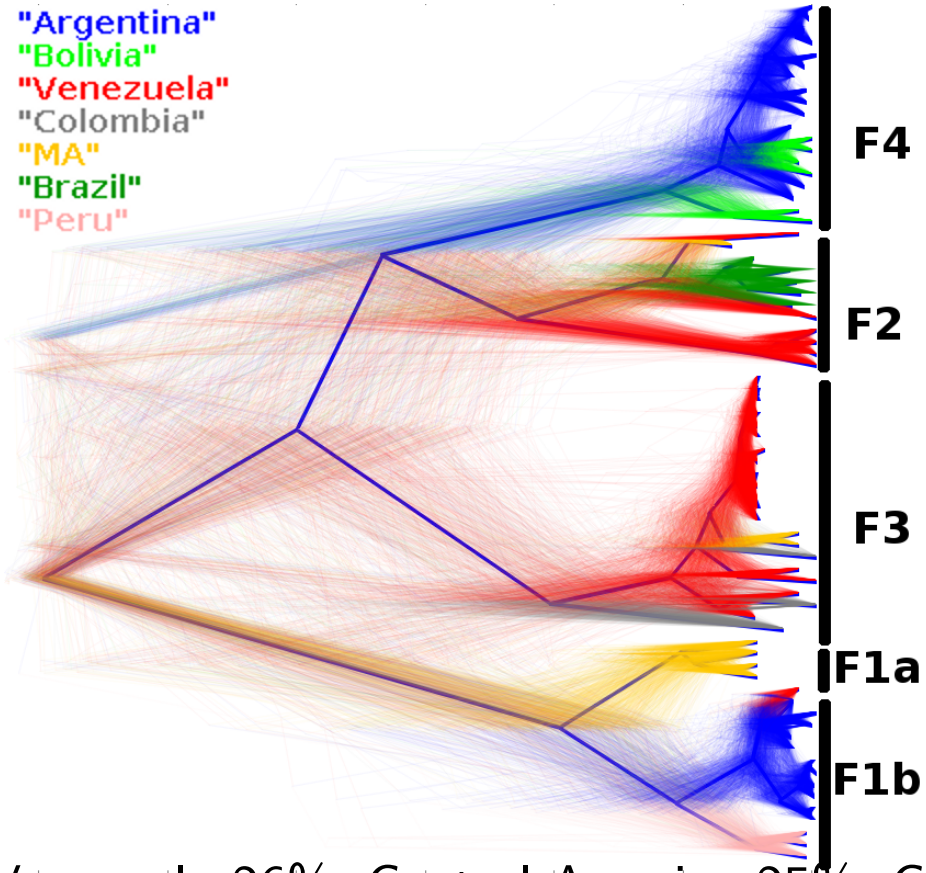
	Argentina	Bolivia	Brazil	Central America	Colombia	Peru	Venezuela
Argentina	–	$r_{A\to Bo}$	$r_{A\to Br}$	$r_{A\to CA}$	$r_{A\to Co}$	$r_{A\to Pe}$	$r_{A\to Ve}$
Bolivia	$r_{Bo\to A}$	–	$r_{Bo\to Br}$	$r_{Bo\to CA}$	$r_{Bo\to Co}$	$r_{Bo\to Pe}$	$r_{Bo\to Ve}$
Brazil	$r_{Br\to A}$	$r_{Br\to Bo}$	–	$r_{Br\to CA}$	$r_{Br\to Co}$	$r_{Br\to Pe}$	$r_{Br\to V}$
Central America	$r_{CA\to A}$	$r_{CA\to Bo}$	$r_{CA\to Br}$	–	$r_{CA\to Co}$	$r_{CA\to Pe}$	$r_{CA\to V}$
Colombia	$r_{Co\to A}$	$r_{Co\to Bo}$	$r_{Co\to Br}$	$r_{Co\to CA}$	–	$r_{Co\to Pe}$	$r_{Co\to V}$
Peru	$r_{Pe\to A}$	$r_{Pe\to Bo}$	$r_{Pe\to Br}$	$r_{Pe\to CA}$	$r_{Pe\to Co}$	–	$r_{Pe\to V}$
Venezuela	$r_{V\to A}$	$r_{V\to Bo}$	$r_{V\to Br}$	$r_{V\to CA}$	$r_{V\to Co}$	$r_{V\to Pe}$	–

transition probabilities $P(t) = e^{Qt}$
so use Bayesian Stochastic Variable Selection (BSVS)

- breaks structured coalescent model => biased estimates
- weak – 95% HPD of root location often contains all states

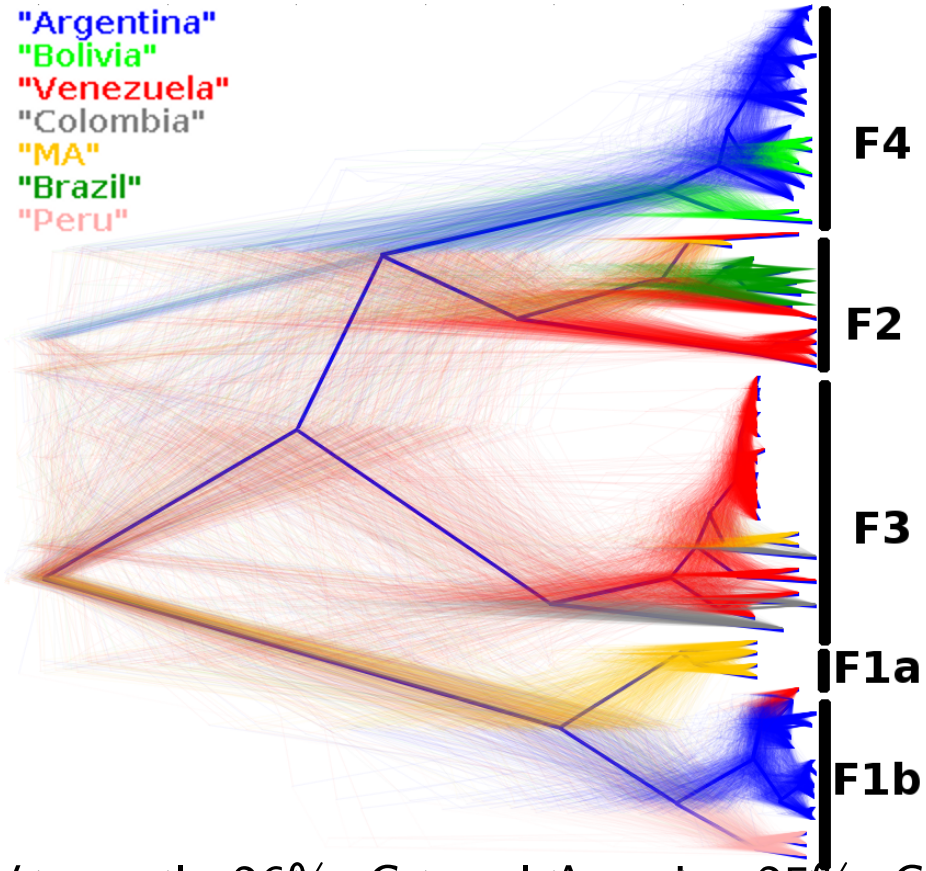
Lemey et al, PLOS Comp Biology, 2009

Ancestral state reconstruction – HBV-F in South America



- Root location Venezuela 26%, Central America 25%, Colombia 18%, Argentina 12%, Bolivia 6%, Brazil 6%, Peru 6%

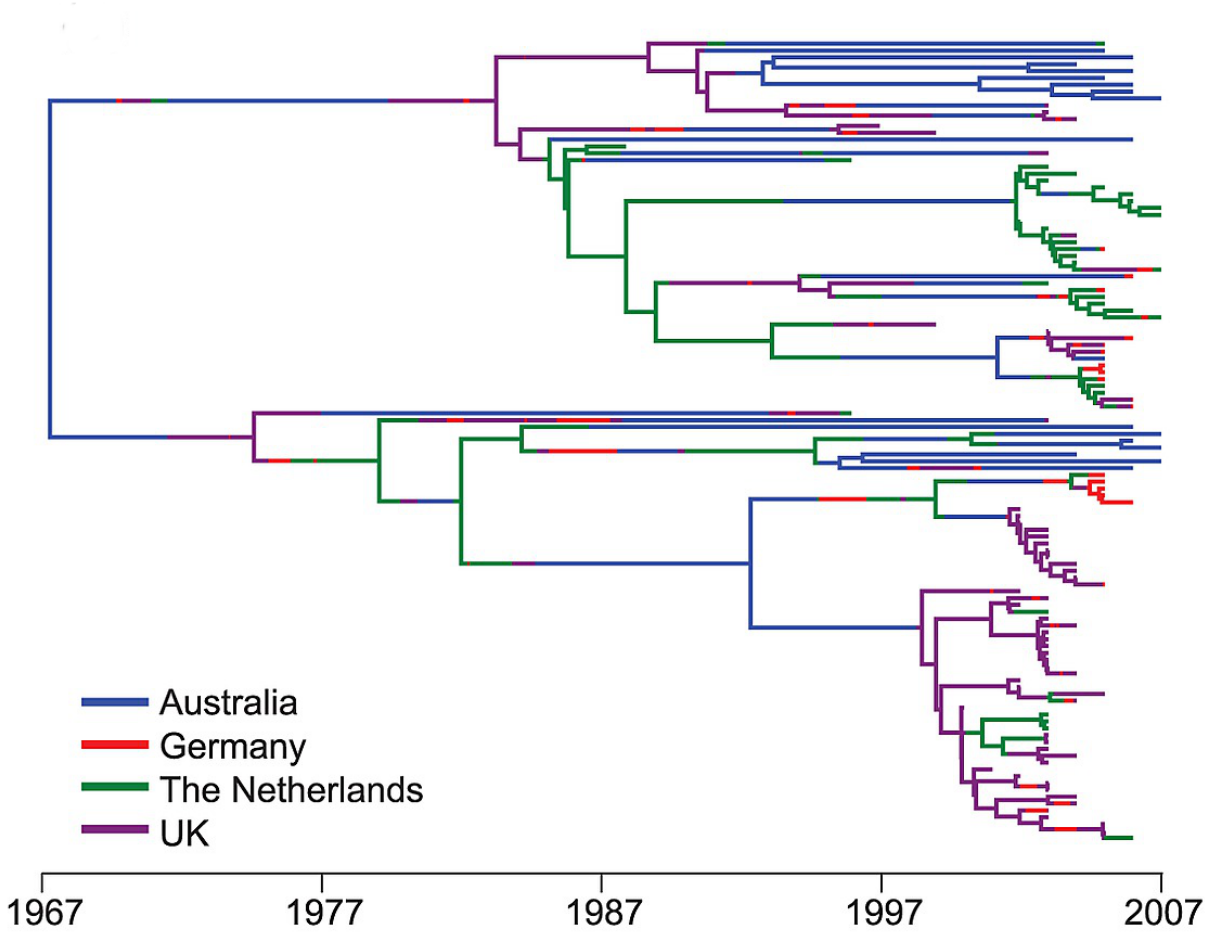
Ancestral state reconstruction – HBV-F in South America



- Root location Venezuela 26%, Central America 25%, Colombia 18%, Argentina 12%, Bolivia 6%, Brazil 6%, Peru 6%
- Initially HBV spread from the North into South America
- Argentina obtained from Central America or Venezuela, not the closer by Brazil or Bolivia and HBV spread from Argentina to West of South America
- Brazil obtained HBV from Central America

Structured Coalescent (`MultiTypeTree` package)

Requires specifying regions (demes)



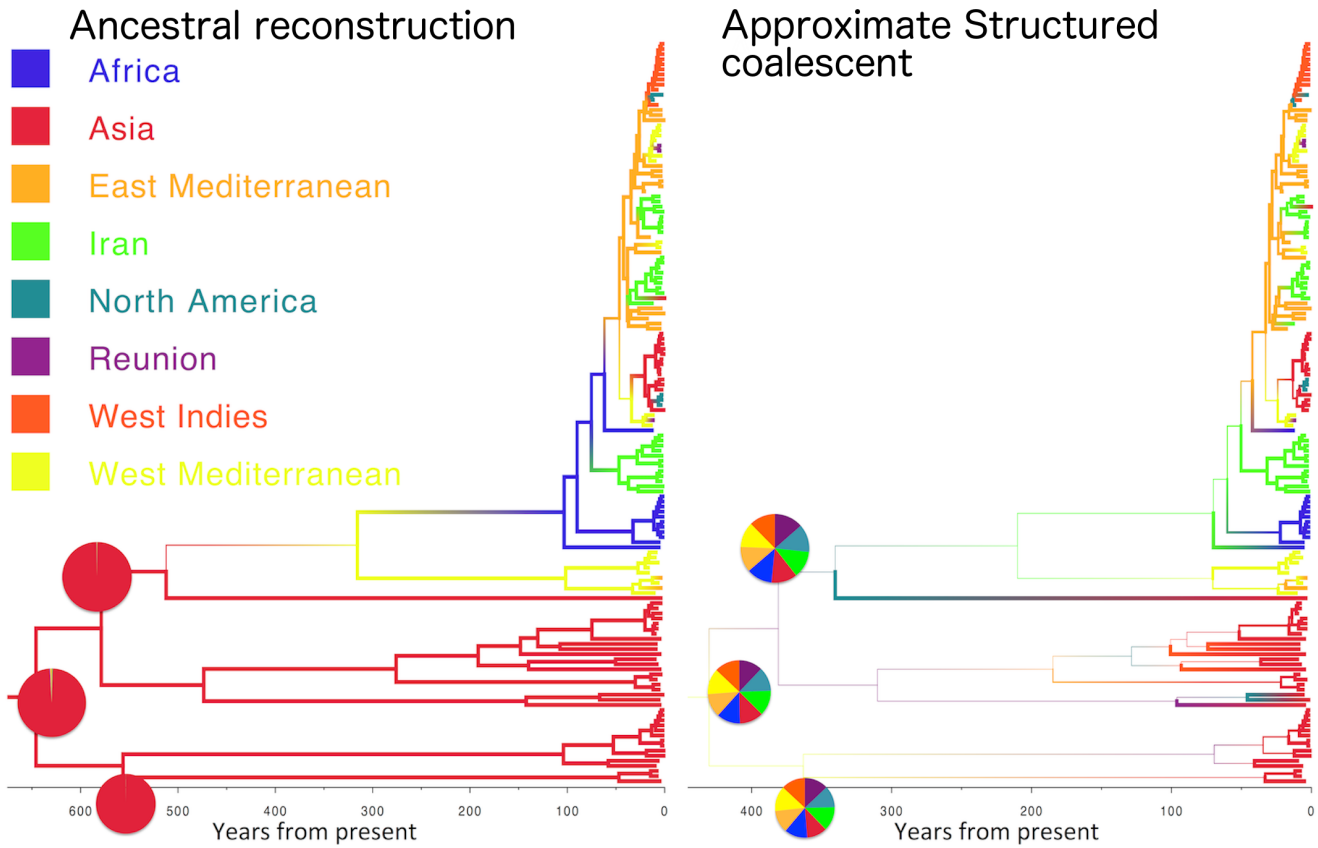
Principled way of migrating/coalescence

Does not scale (max 4 demes)

Vaughan et al. BioInformatics, 2014

Approximate structured coalescent (BASTA package)

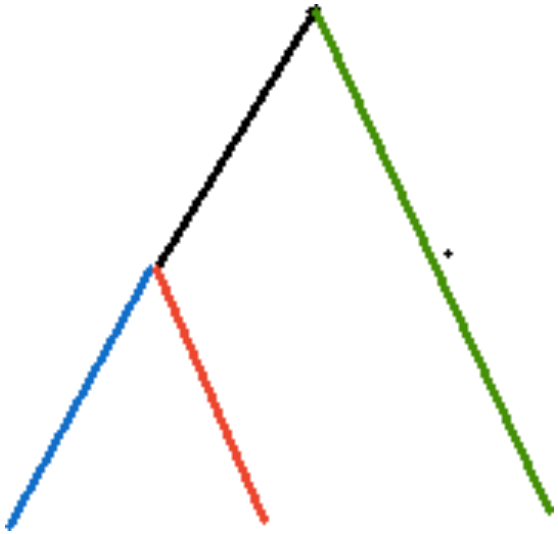
- Scales slightly better (max 12? demes) than structured coalescent
- Results differ significantly from ancestral reconstruction



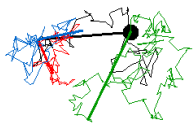
Tomato Yellow Leaf Curl Virus

De Maio et al, PLOS Genetics, 2015

Random walks – types

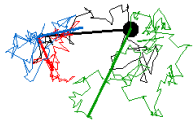


Simple

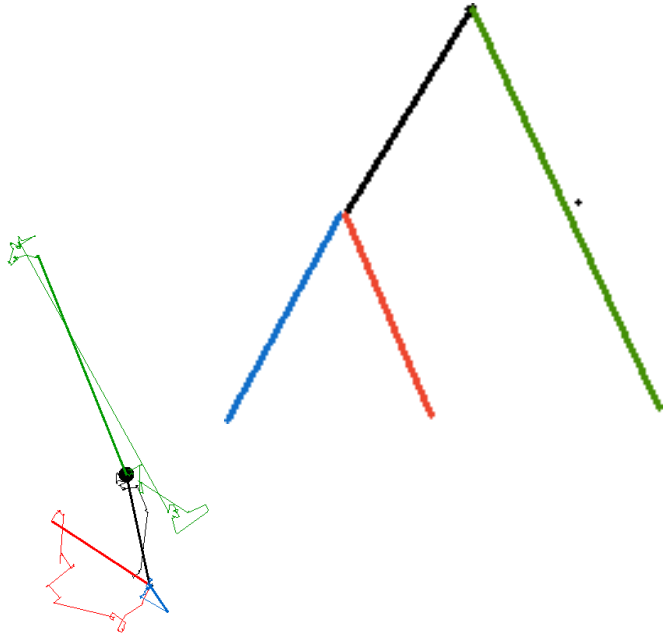


Random walks – types

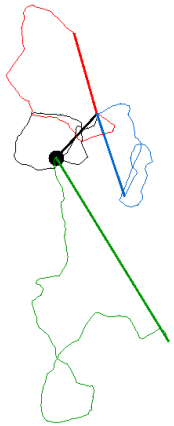
Simple



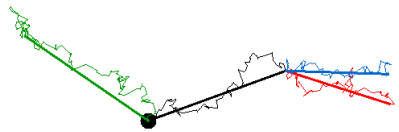
Levy



Correlated

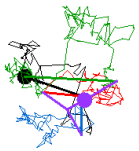


Biased

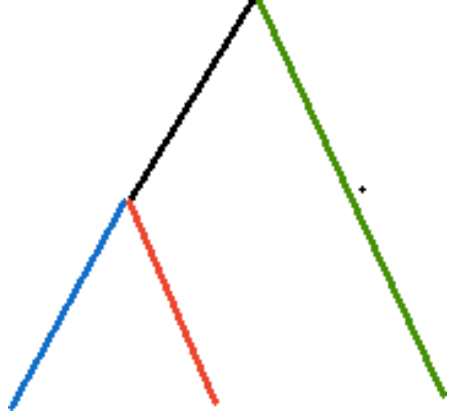


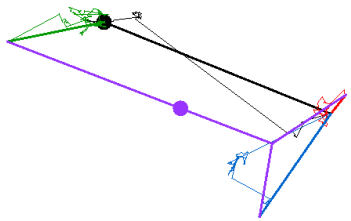
Random walks – tree reconstruction

Simple

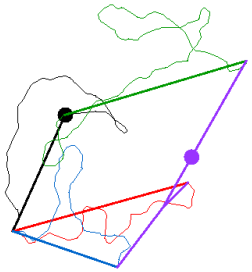


Levy

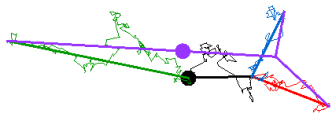




Correlated

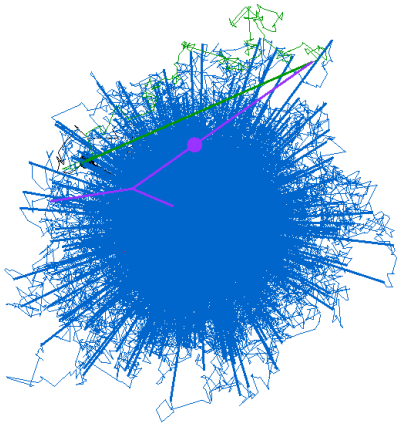


Biased

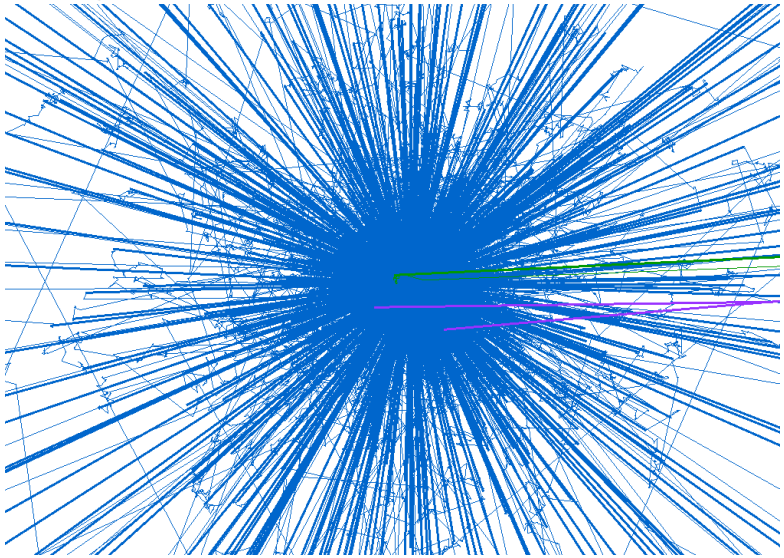


Random walks – distributions

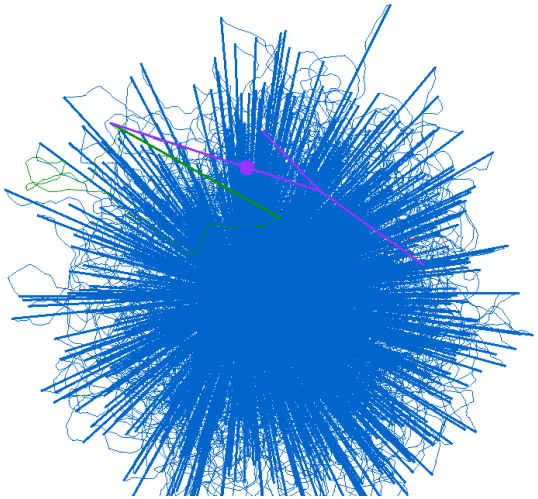
Simple



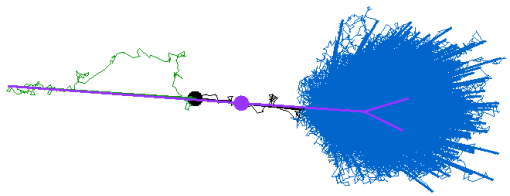
Levy



Correlated



Biased



Random walks

Diffusion on a plane (`BEAST_CLASSIC` package)

- more power than ancestral reconstruction
- limited landscape awareness
- does not deal with Mercator projection distortion

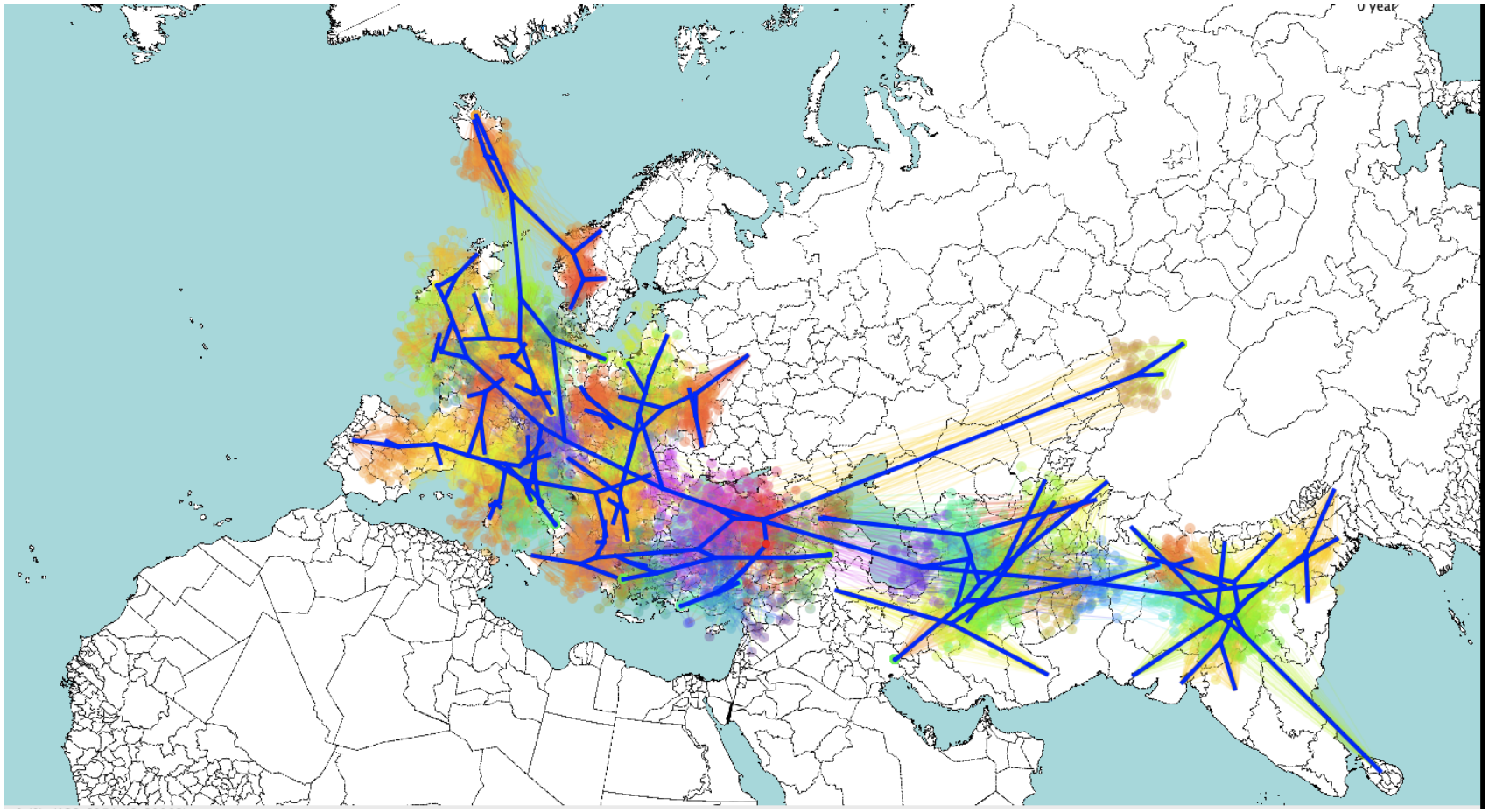
Lemey et al. MBE, 2010 Pybus et al. PNAS, 2012

Diffusion on a sphere (`GEO_SPHERE` package)

- as above, but deals with Mercator projection distortion

Bouckaert Peerj 2016

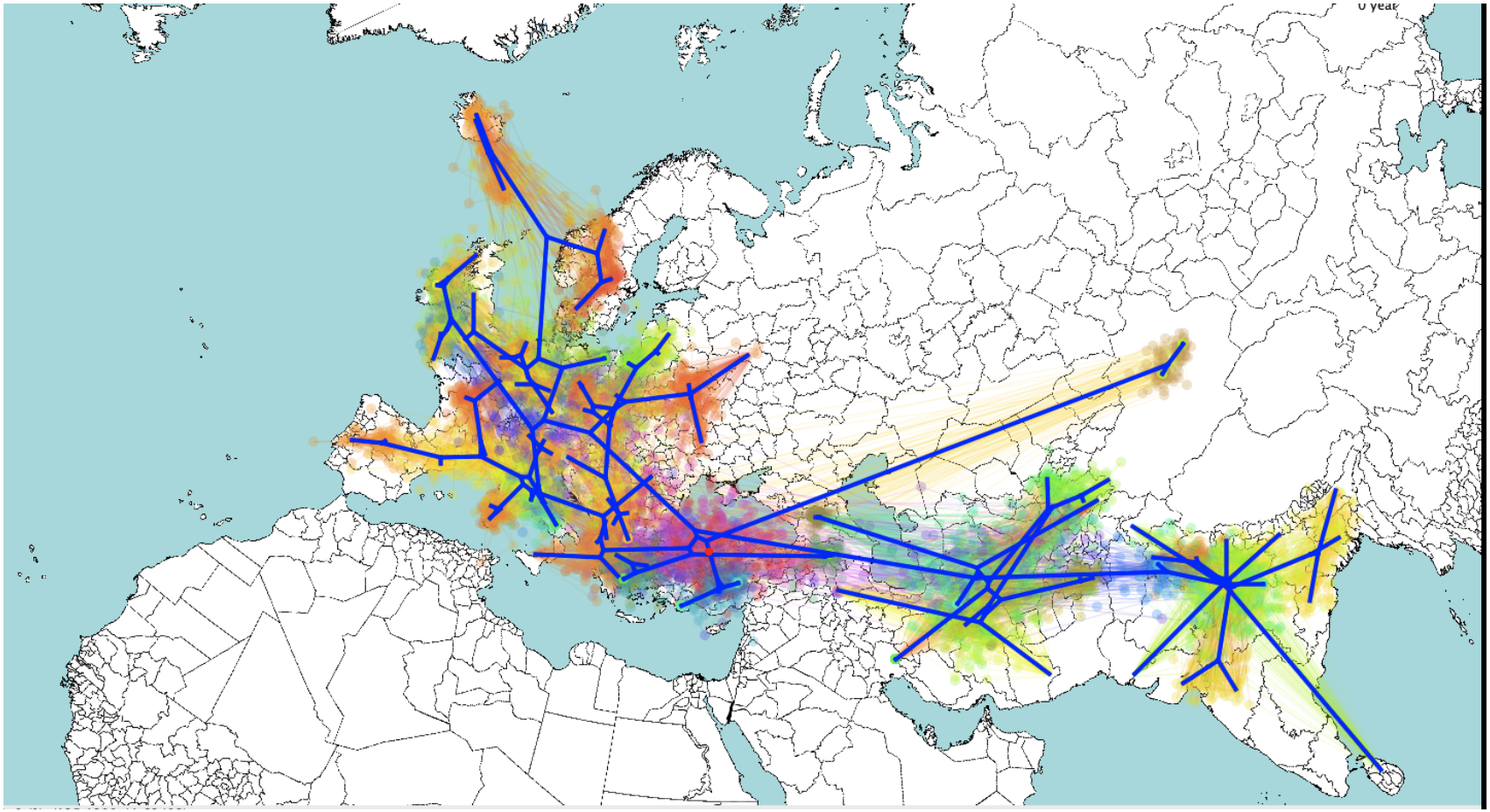
Reconstructing the origin or Indo European



Simple random walk on sphere, fixed tree, strict clock

`GEO_SPHERE` package IEstrict.mp4

Indo European – relaxed random walks*



Simple random walk on sphere, fixed tree, relaxed clock, much better fit

`GEO_SPHERE` package IERC.mp4

* Lemey et al. MBE, 2010

Geographical priors

- root location
- tip location, e.g. Russian, sampled from Russian speaking regions
  - use uniform prior on region `GEO_SPHERE` package Bouckaert et al Science, 2012

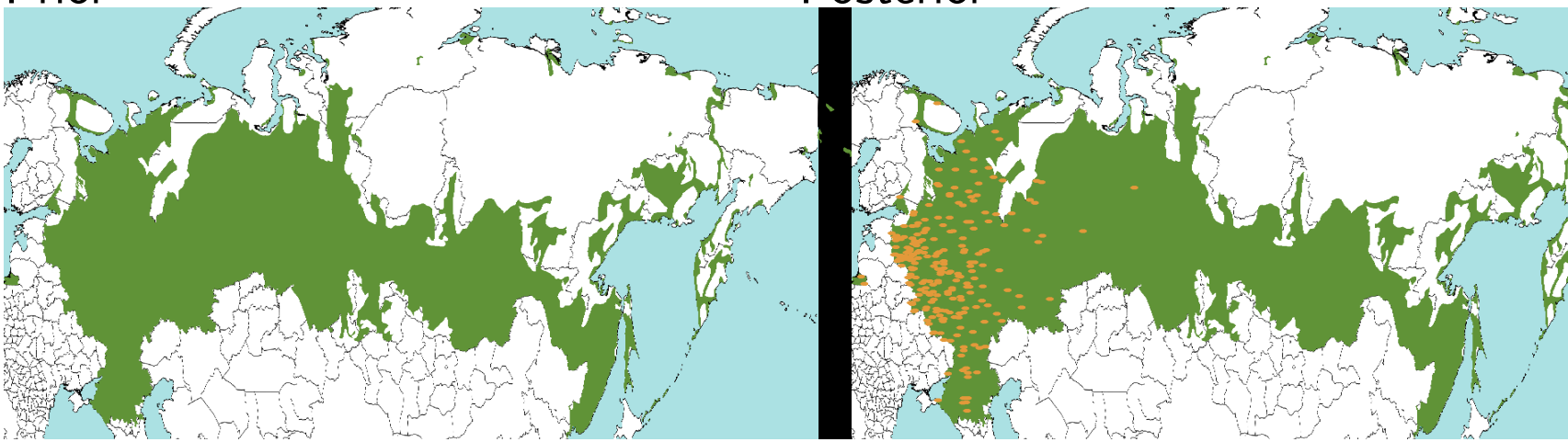
Prior

Posterior

Geographical priors

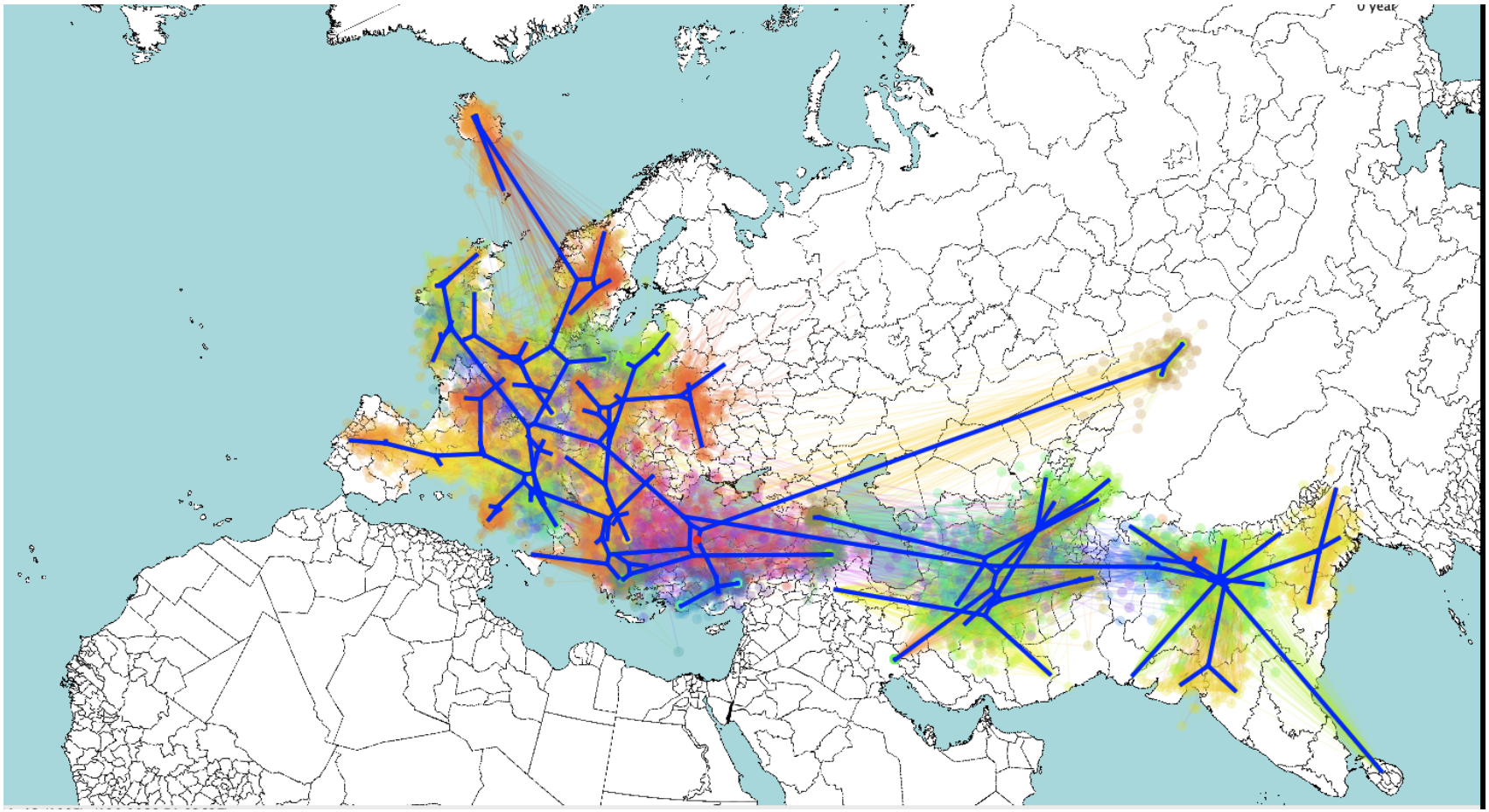
- root location
- tip location, e.g. Russian, sampled from Russian speaking regions
  - use uniform prior on region `GEO_SPHERE` package Bouckaert et al Science, 2012

Prior Posterior



  - use non-uniform prior on region Nylinder et al SysBio 2014 BEAST 1
  - Integrate out regions Quintero et al SysBio 2015 fixed tree, R package only
- MRCA of clade location, e.g origin of American samples in region around Bering strait
- restrict all internal nodes, e.g. no node in water

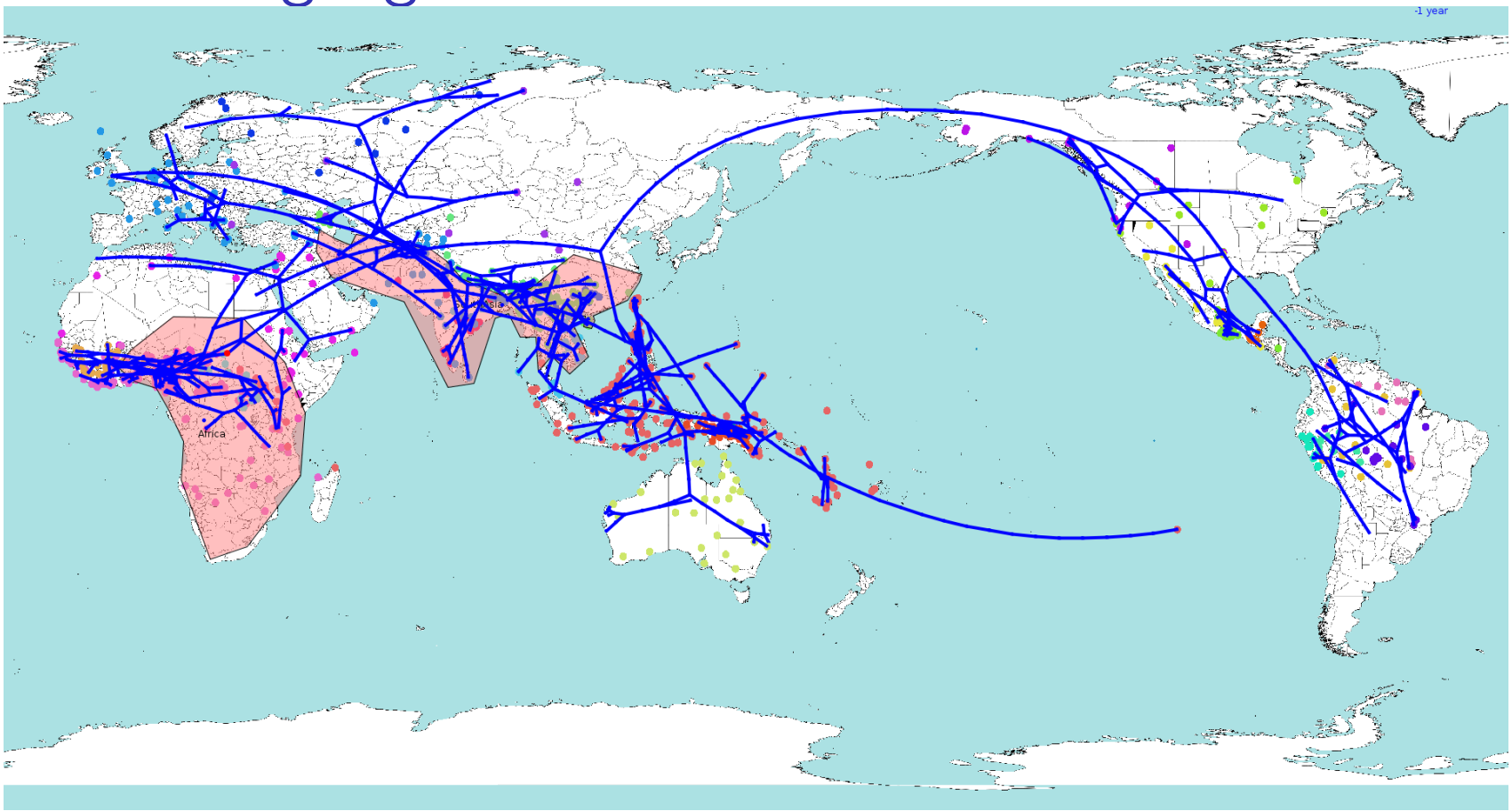
Indo European – Sampling tip locations



Simple random walk on sphere, fixed tree, relaxed clock, more uncertainty

`GEO_SPHERE` package IERCtips.mp4

Global language tree

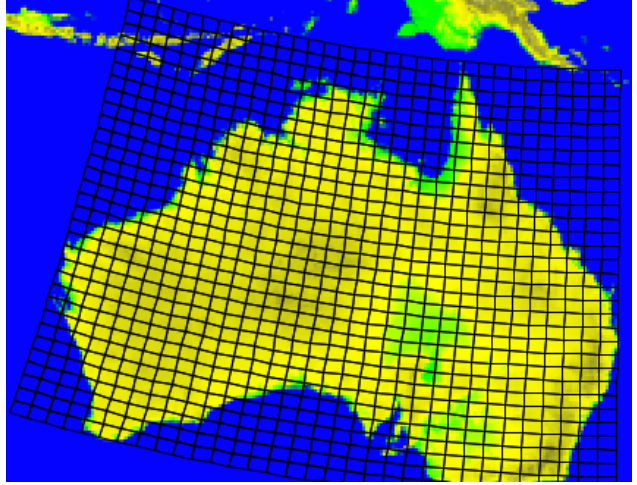


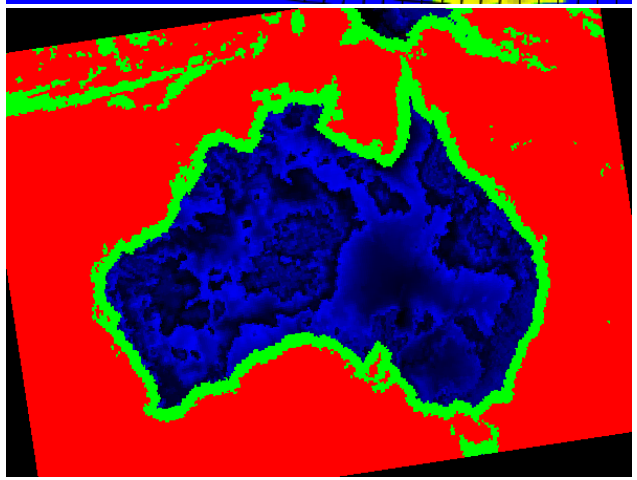
- prior on root location and 'rest of world' origin
- age constraint on all families
- age constraint on root
- glottolog constraints on tree glotto.mpg

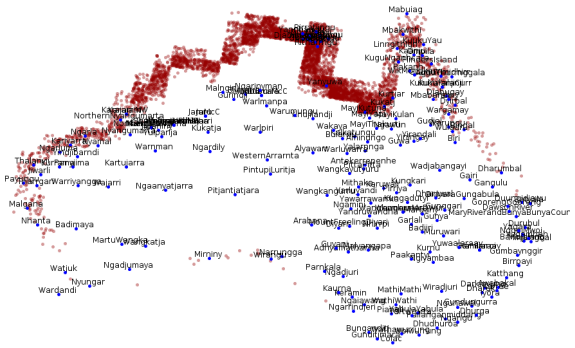
Landscape aware models

- divide map in a 32x32 grid
- distinguish between land, water, interior
- solve ODE on underlying pixel map using Euler's method (la.mpg)
- relaxed clock

- allows landscape feature modelling (water, mountains, barriers)
- computationally expensive
- cannot sample model parameters during MCMC/needs pre-computation
- allows nice visualisation (la2.mpg)





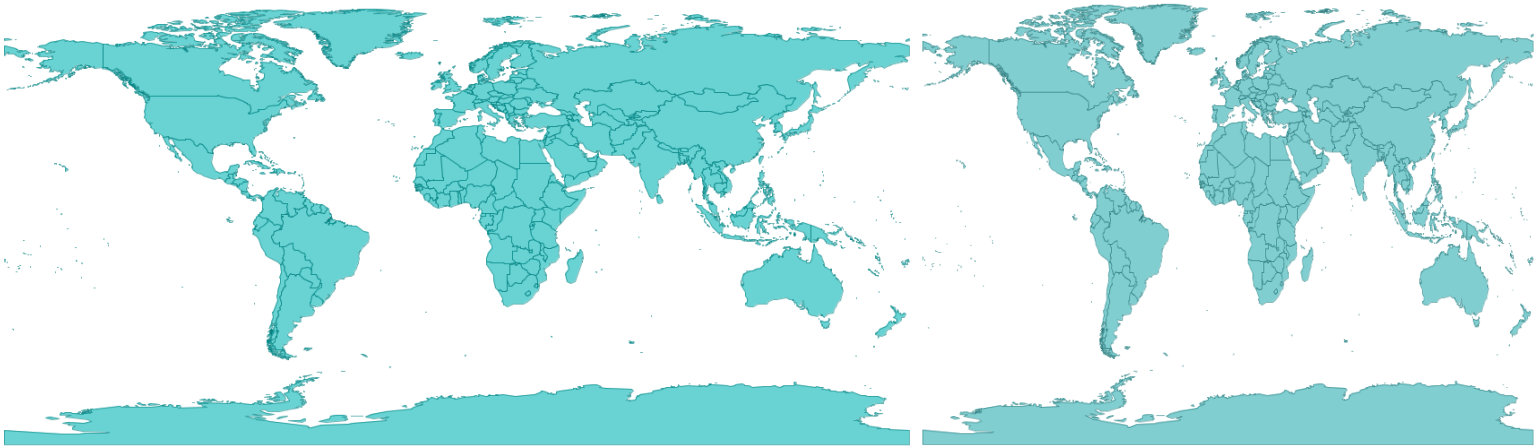


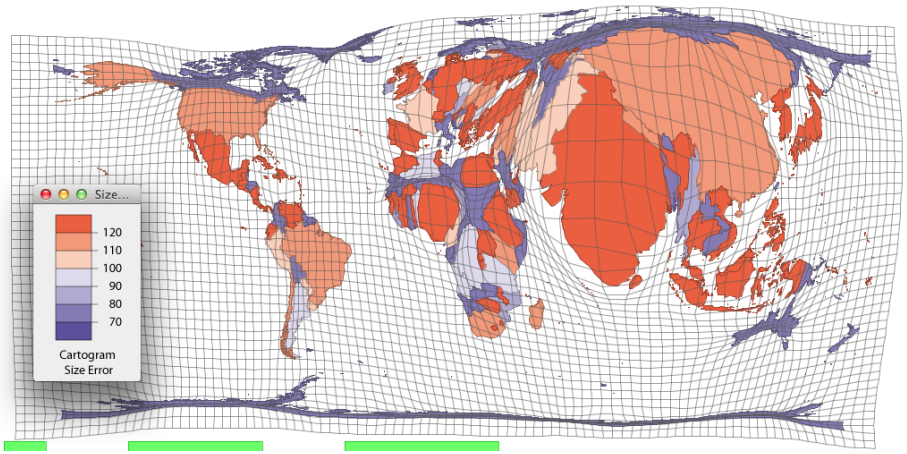
Bouckaert et al. Science, 2012 (package in development)

Landscape awareness

By distorting the map + run homogeneous random walk model

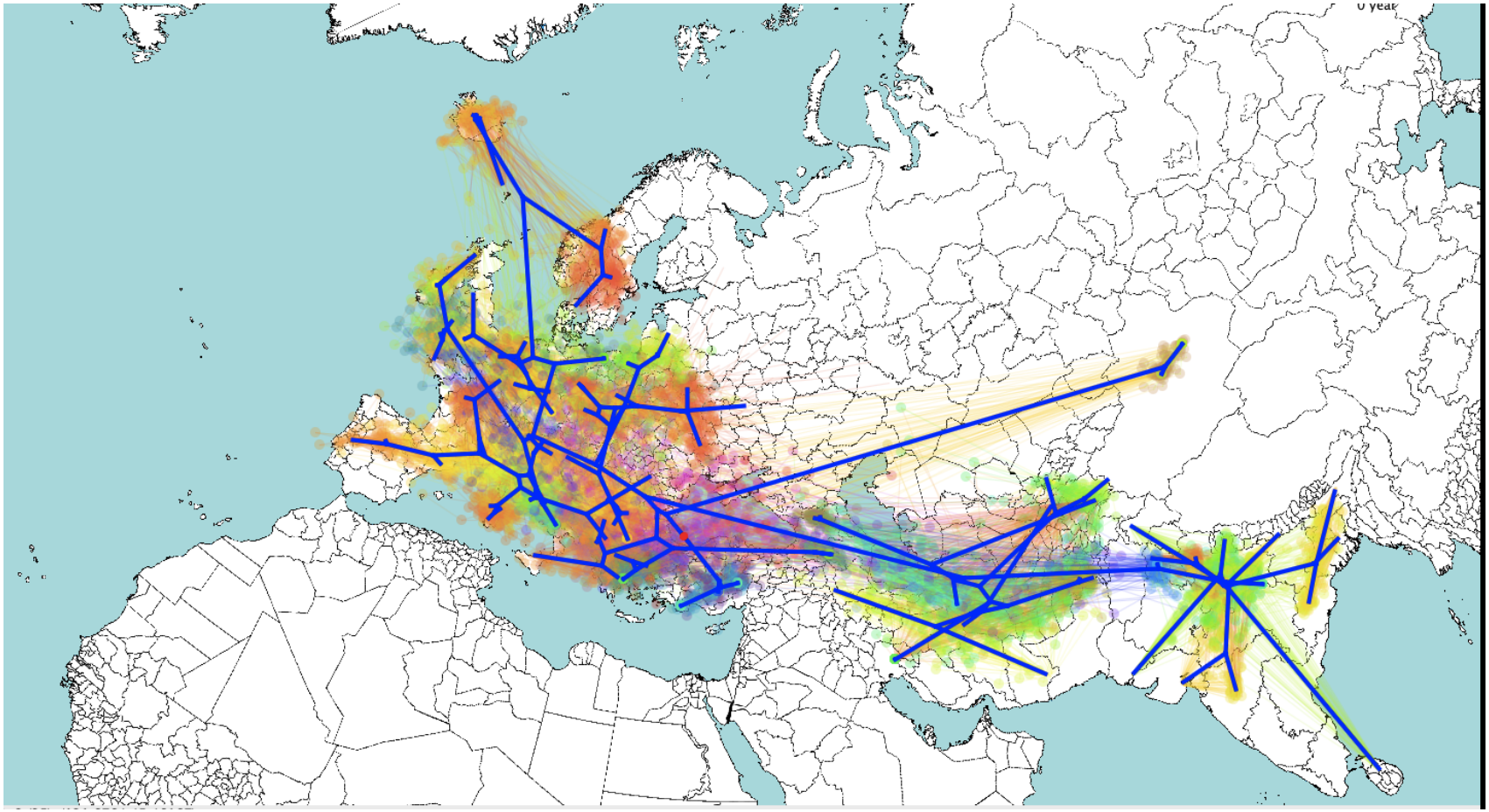
- affine transforms
- cartogram





Bouckaert BioArchiv 2015

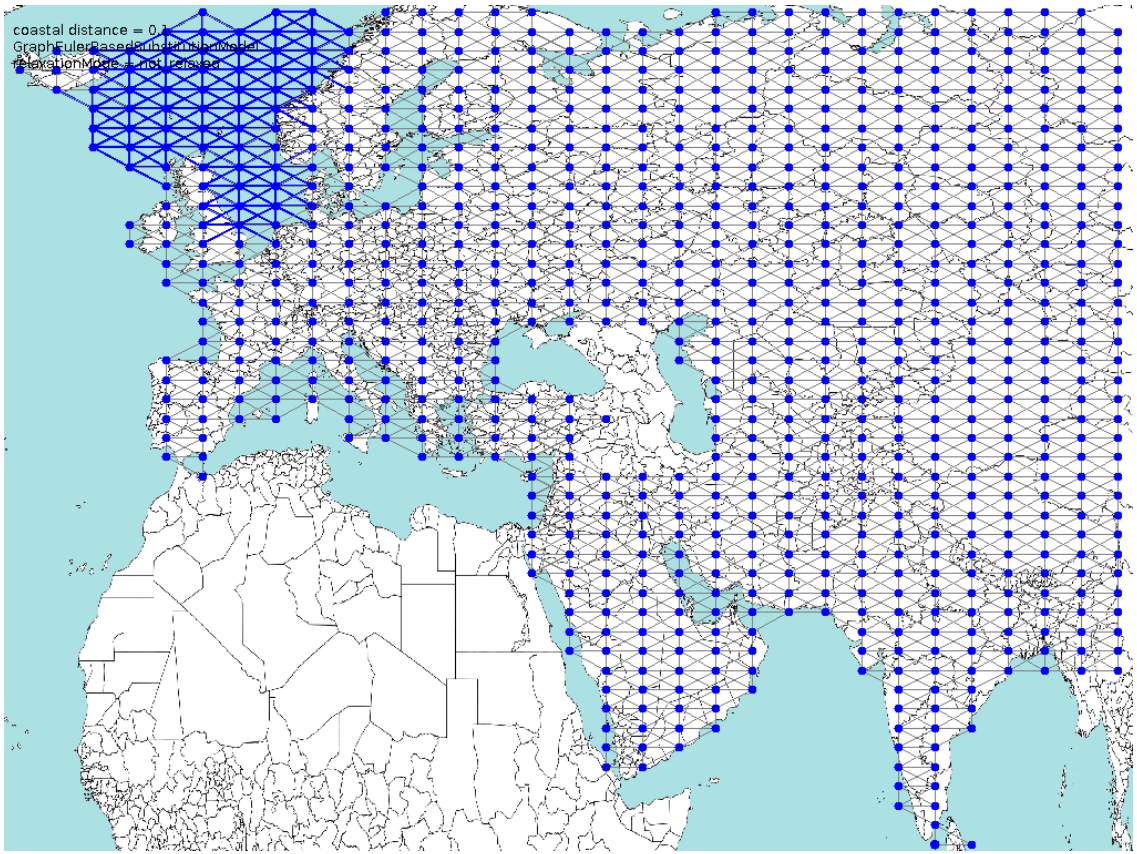
Indo European – population size distorted map



Origin shifts around a slightly IERCtipsTrans.mp4 `GEO_SPHERE` package

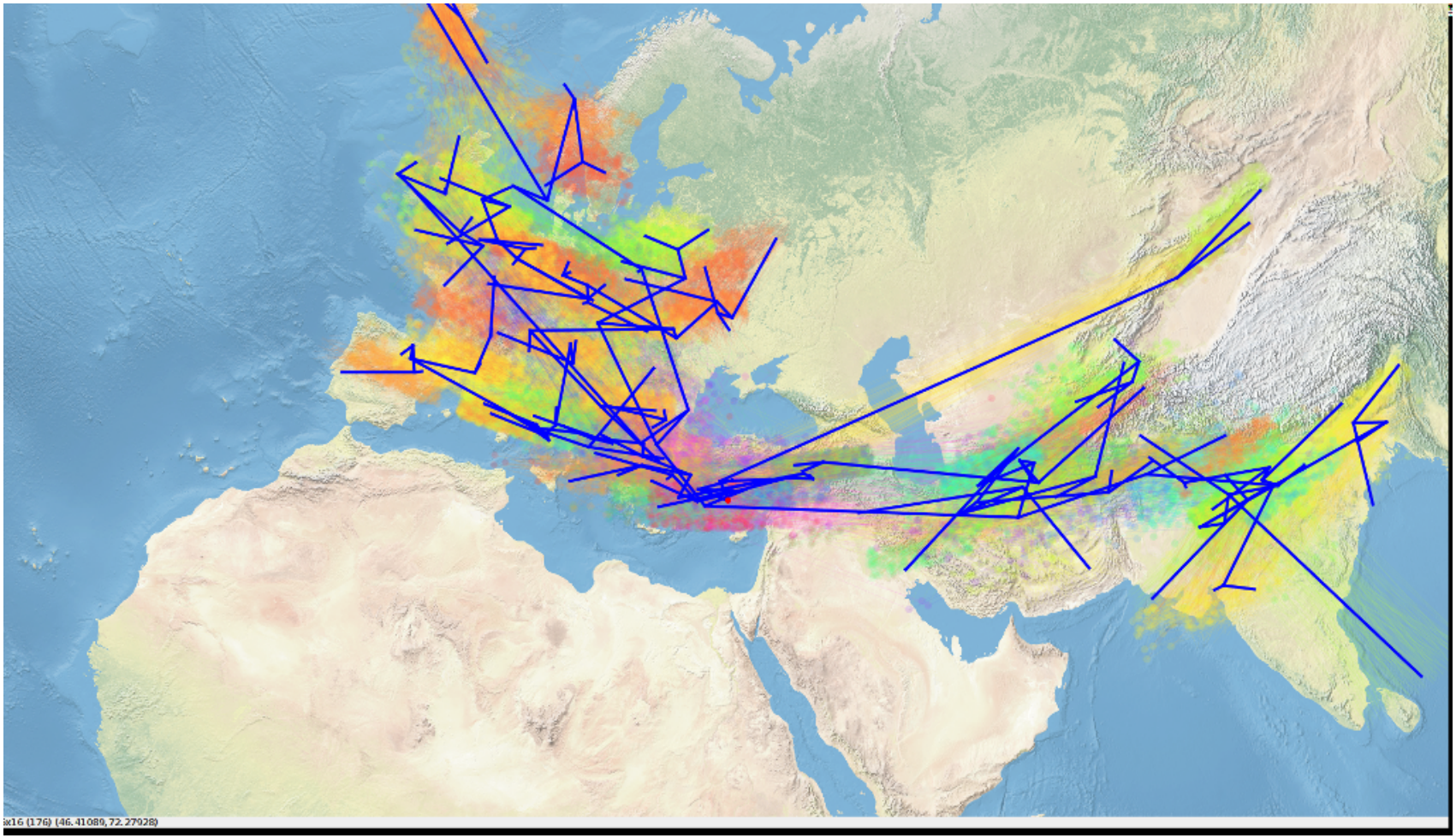
Random walk on graph

- arbitrary neighbourhood structure
- arbitrary rates
- allow testing hypothesis of barriers, e.g. no migration through Caucasus



show SphereTesselation app Bouckaert BioArchiv 2016

Indo European – no migration through Caucasus



IEnoCaucasus.mp4

Visualisation

- Summary tree on map connecting internal nodes
  - by straight lines
  - by shortest great circle distance lines
  - MAP route in underlying landscape model NNA2.mp4
- Posterior distribution on map
  - by dots
  - be DensiTree
  - by 80%HPD regions ie.mov

Methods of inference

- Data augmentation
  - sample locations L for nodes in tree
  - $P(L|T,\theta) = \prod_{i \in L} P(i|parent(i),\theta)$
  - may have trouble converging
- Integrate out internal node locations (peeling algorithm)
  - works well with standard distributions (e.g. normal diffusion)
  - converges fast
  - can deal with *a few* constraints by data augmentation for affected nodes
  - cannot deal constraints on all nodes (e.g. restriction to being on land)
- Particle filtering
  - flexible in dealing with constraints
  - slow

Ideal questions

- Precondition: Requires a tree
  - genetic data
  - morphological data
  - anything with evolutionary mechanism

Some questions that can be answered with phylogeography

- What is the origin of a group (95%HPD)?
- Test two (or more) origin hypotheses
  What is the Bayes factor in favour of one?
- Reconstruct migration routes
- Estimate migration rates
- Test transition hypotheses (discrete space methods)
- Test landscape heterogeneity hypotheses

Summary

- Discrete space methods
  - Ancestral reconstruction
  - (Approximate) Structured Coalescent
- Continuous space diffusion models
  - various random walks
  - tip distributions
  - priors on nodes
  - somewhat heterogeneous
- Random walk on graph
  - flexible tip distributions
  - flexible priors
  - flexible heterogeneity

+Strict/Relaxed clocks for all