

Phylodynamics Workshop

Time-scales and BEAST Trees

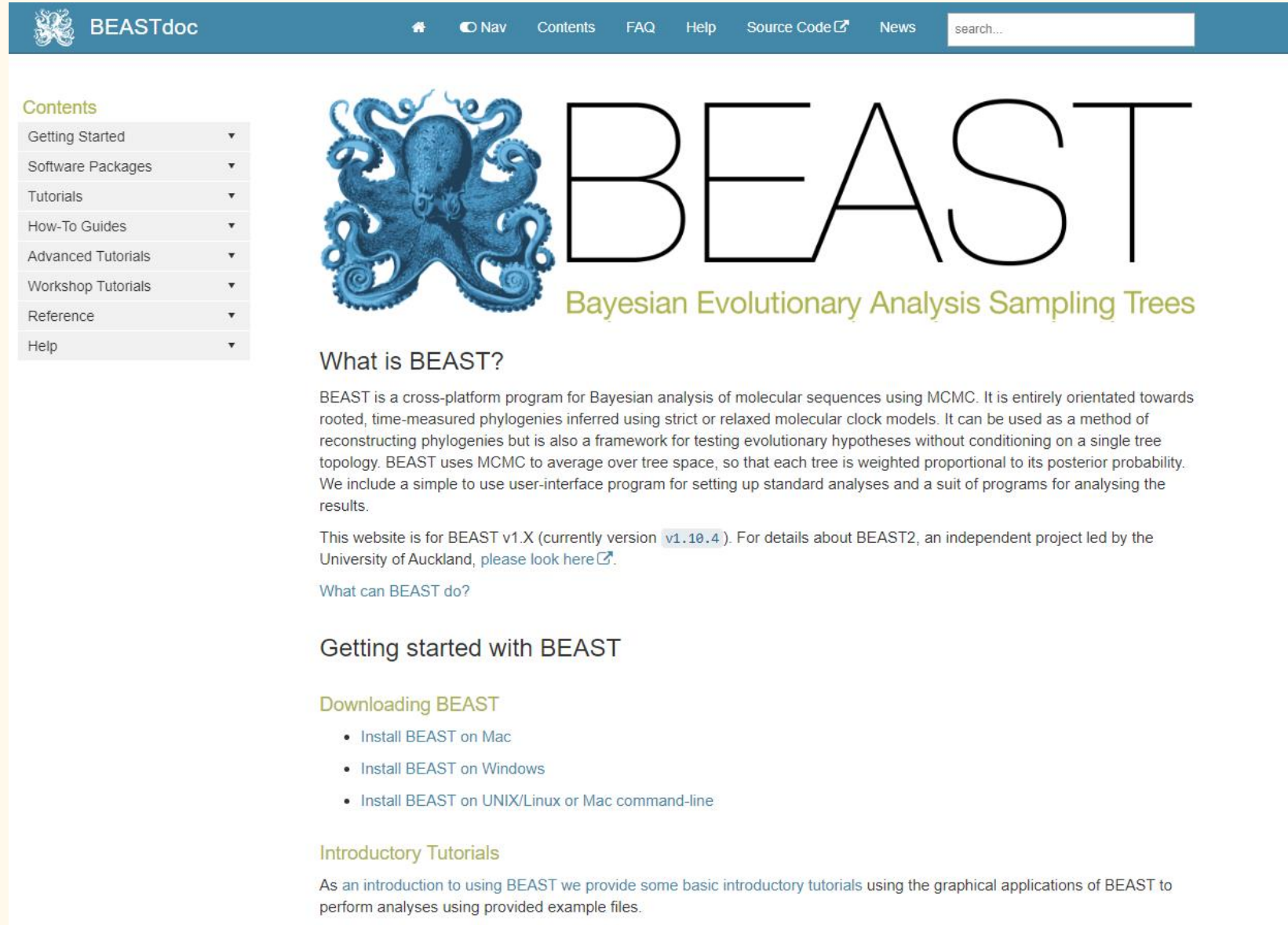
Inferring time-scaled trees using sophisticated evolutionary model
including clock rate

BEAST

What is BEAST ?

- Sequence data with times
=> Phylogenetic trees
- Uses Coalescent models –
i.e. also infer effective
population sizes over time

<http://beast.community/>



The screenshot shows the BEASTdoc website. At the top is a navigation bar with the BEASTdoc logo, a home icon, a navigation icon, and links for Contents, FAQ, Help, Source Code, and News. A search bar is on the right. On the left is a 'Contents' sidebar with a dropdown menu listing: Getting Started, Software Packages, Tutorials, How-To Guides, Advanced Tutorials, Workshop Tutorials, Reference, and Help. The main content area features a large blue octopus logo next to the word 'BEAST' in large, thin letters. Below this is the subtitle 'Bayesian Evolutionary Analysis Sampling Trees'. The text 'What is BEAST?' is followed by a paragraph explaining that BEAST is a cross-platform program for Bayesian analysis of molecular sequences using MCMC, oriented towards rooted, time-measured phylogenies. It mentions that BEAST uses MCMC to average over tree space and includes a user-interface program. Below this, it states the website is for BEAST v1.X (currently version v1.10.4) and provides a link for details about BEAST2. The section 'What can BEAST do?' is followed by 'Getting started with BEAST'. Under 'Downloading BEAST', there is a list of links: 'Install BEAST on Mac', 'Install BEAST on Windows', and 'Install BEAST on UNIX/Linux or Mac command-line'. The 'Introductory Tutorials' section states that basic introductory tutorials are provided using graphical applications of BEAST to perform analyses using provided example files.

BEAST

- BEAST = “Bayesian Evolutionary Analysis Sampling Trees”
- **What does this mean ?**
 - Previously you have come across **ONE tree** to describe the relationship between the sequences
 - This was generated by Neighbour Joining algorithm, or Maximum Likelihood
 - But, what if a slightly different tree was nearly as good / better ? (especially when sequences are just a few mutations different)
 - BEAST gives a collection of **MANY likely trees**

Tree Likelihood (1)

- A tree is a “model” – it has parameters:
 - Branch lengths
 - Topology (the branching order)
- Want to calculate the probability of a model (tree) given some data (sequences), in order to choose the best or collection of good models.

sub-models contribute to these,
e.g. clock rate model



Tree Likelihood (2)

- Use Bayes theorem:

$$p(\text{Model}|\text{Data}) = \frac{p(\text{Data}|\text{Model})p(\text{Model})}{p(\text{Data})}$$

Posterior

- we want a collection of high scores

Likelihood

- of seeing this data with those mutations, given that tree

Prior

- probability of the model before any data
- can be un-informative (anything goes)
- or can be based on other knowledge (biology, other data)

Markov Chain Monte Carlo (1)

- Start off with a parameter value
- Keep changing it until the Likelihood is at its highest
- But very many parameters and combinations !
- This means that “hill climbing” doesn’t work very well – beware the local optima !



Markov Chain Monte Carlo (2)

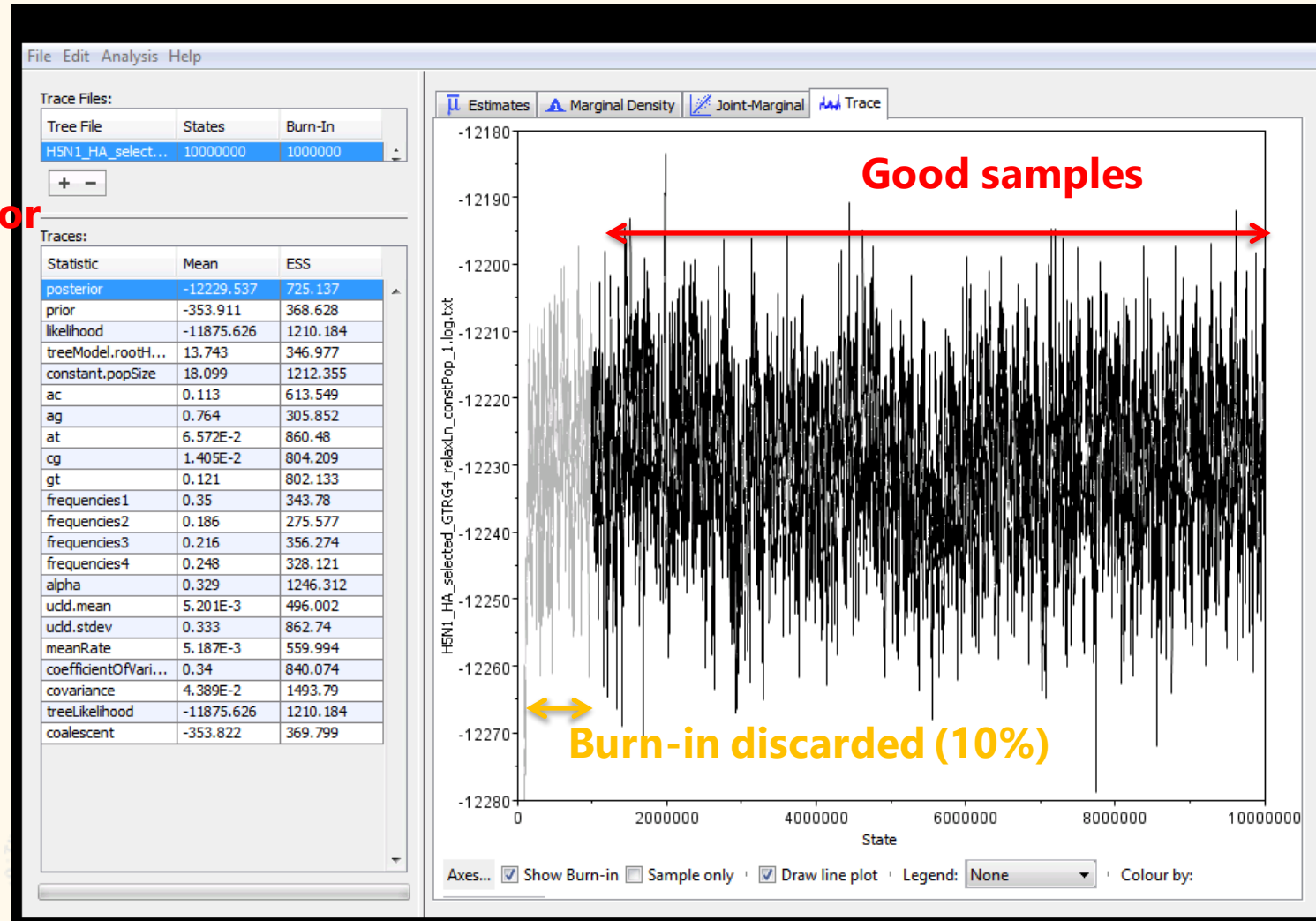
- Always accept a parameter value giving a better likelihood, but **sometimes** accept a parameter value with a worse likelihood
- “**Sometimes**” = with a probability proportional to the ratio of the new and original likelihoods x priors
 - (good mathematical reasons for this.. but beyond scope here)
- Results – perform several 1000 steps (e.g. 10^6)
 - **Initial climb, likelihood increases (“burn-in”)**
 - **Sampling around the peak of the posterior => this is the set of good answers**

Trace File

- Log file from BEAST displayed in Tracer

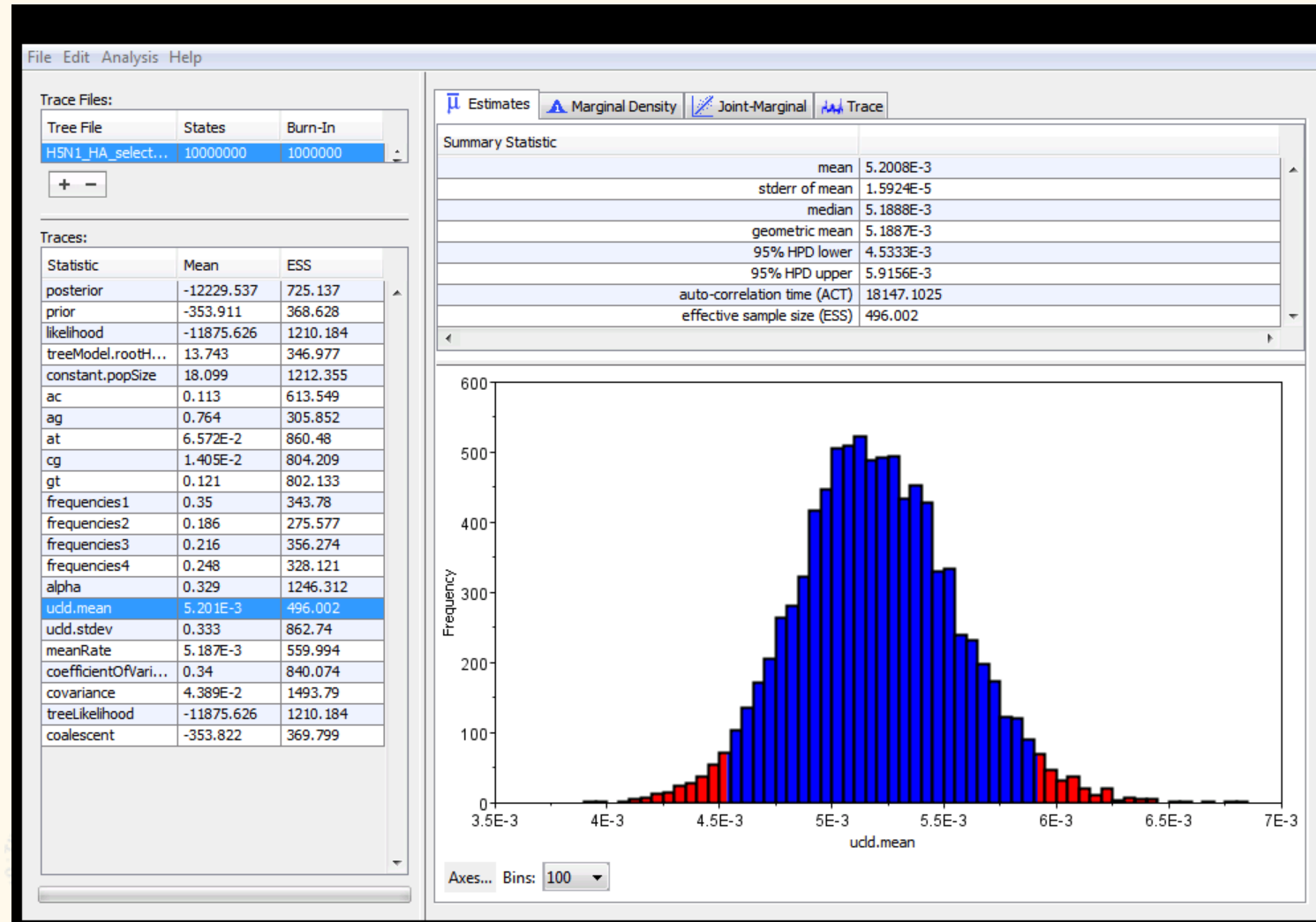
Posterior

Other
parameters

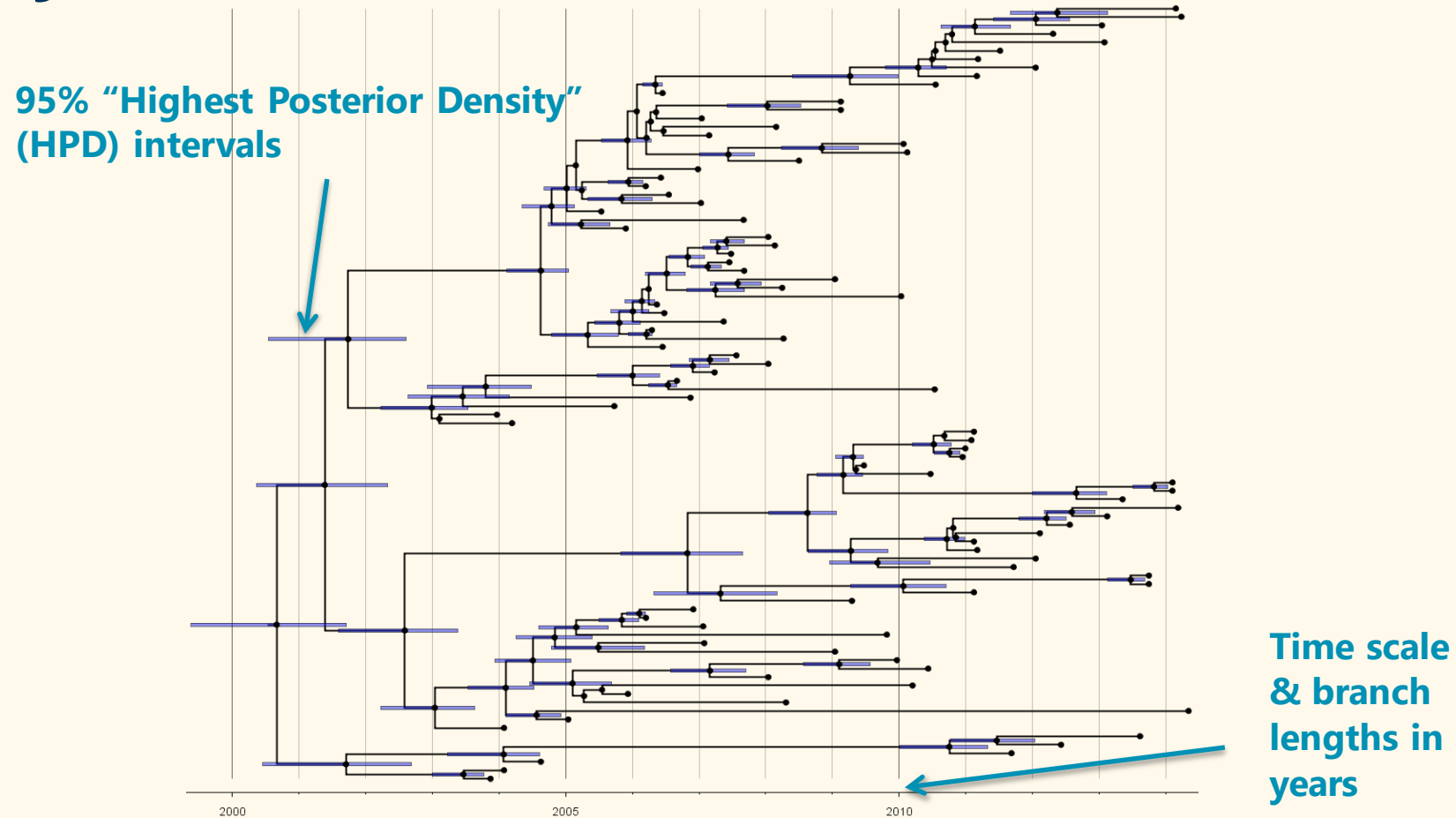


Clock Rate Estimate

- Log file from BEAST displayed in Tracer



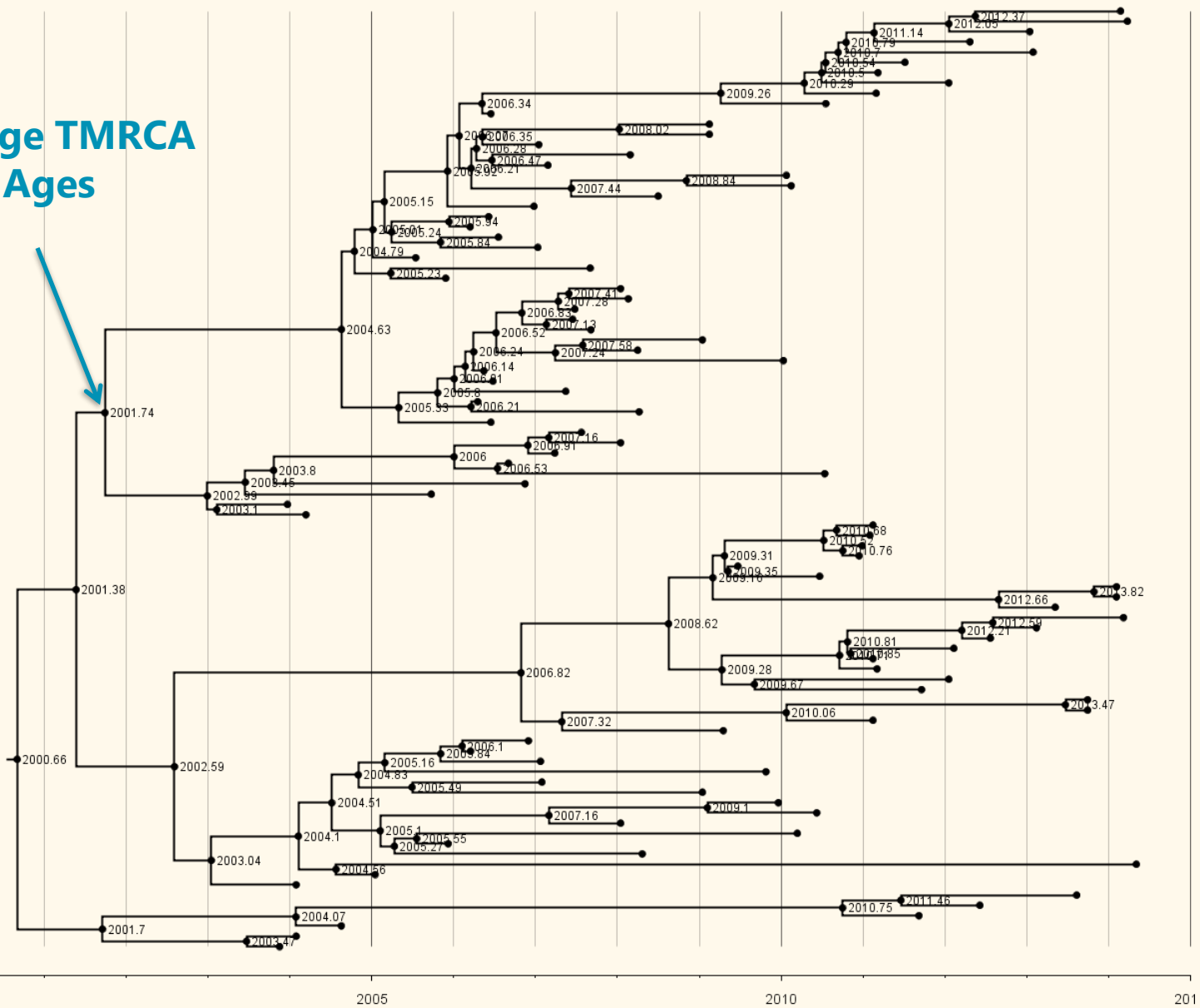
Summary Timescaled Tree



- This is the Maximum Clade Credibility Tree (MCC Tree)
- The "best" tree of the samples, with the intervals from the other trees mapped onto it

Summary Timescaled Tree

Average TMRCA
Node Ages



Basic BEAST model choices

- **Mutation model**

- Nucleotide: HKY or GTR
- Codon: SRD06 (HKY on positions 1 & 2, HKY on 3) or Yang (all GTR)
- All models can include site-site rate variation
- Typical choices:
 - Tb: HKY + Gamma x 4 (not much mutation)
 - Flu: SRD06 (+ Gamma x 4) (coding sequences)

- **Clock Models: strict or relaxed**

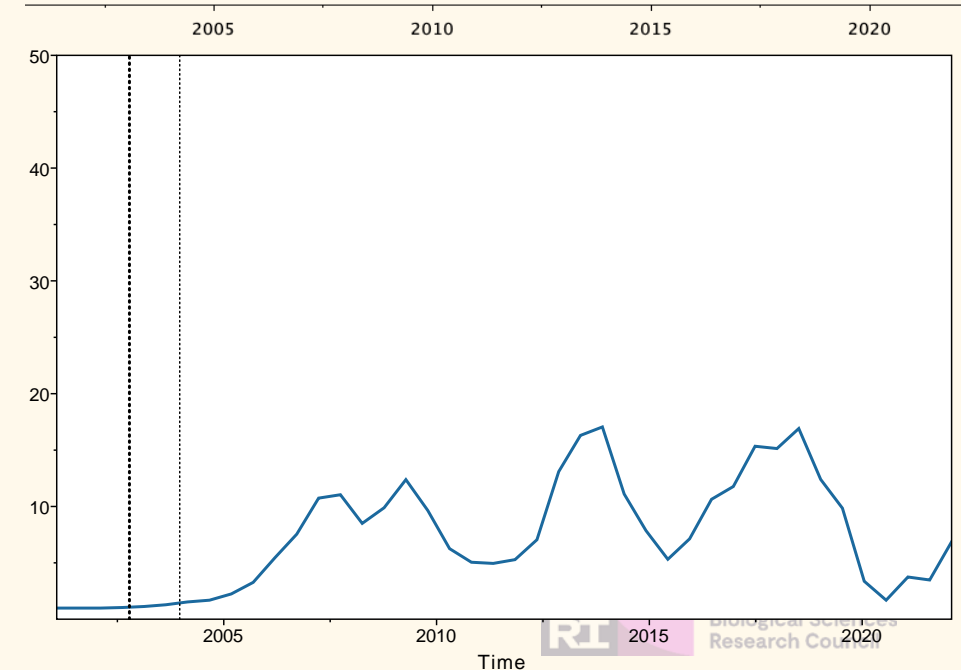
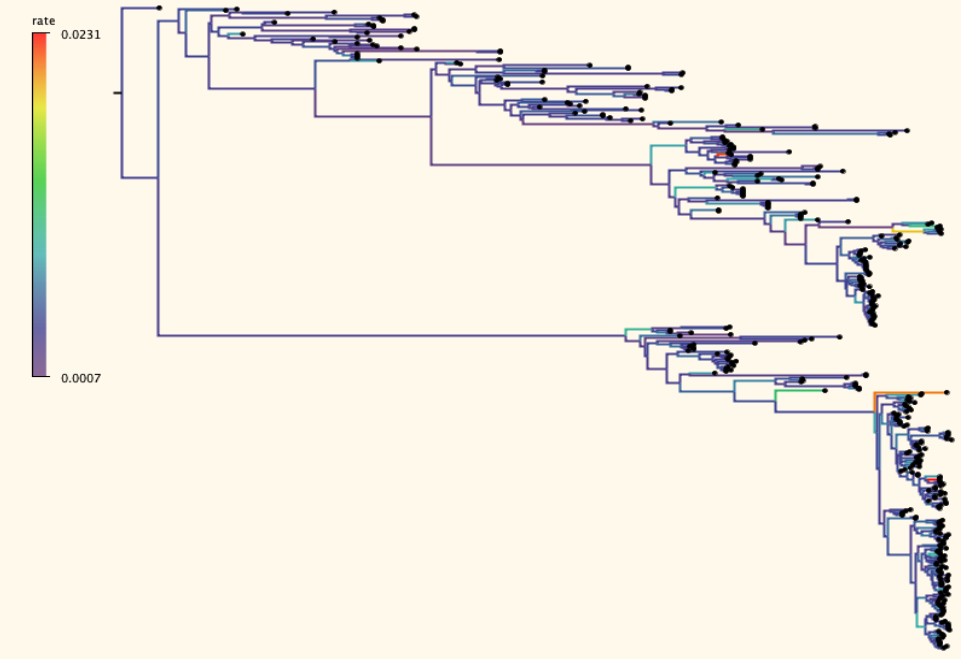
- Strict = one clock rate for the whole tree
- Relaxed = each branch has its own overall rate multiplier drawn from a log normal distribution (or exponential distribution). Parameters of the distribution are estimated as well as the overall rate.

- **Population Models – effective population size over time**

- Constant, Exponential Growth, Logistic Growth
- Skyline & Skygrid (variable)

Essence of Phylodynamics (1)

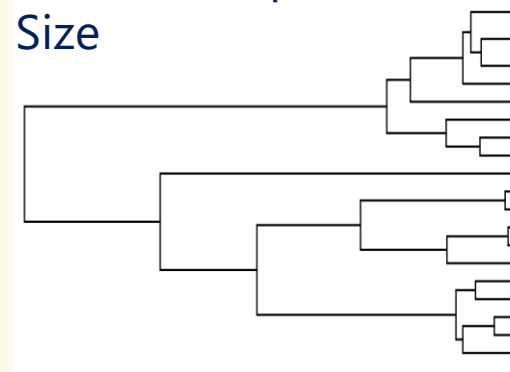
- Use virus sequence data to make phylogenies
 - Nucleotide differences => genetic distance
=> tree
- Add time-scale to phylogenies – require
 - time-stamped sequences and observed mutations
 - Concept of molecular clock (accumulate mutations at rate(s))
 - concept of viral diversity or effective population size over time
 - estimate phylodynamic growth rate or R_0 (or R) from surveillance samples



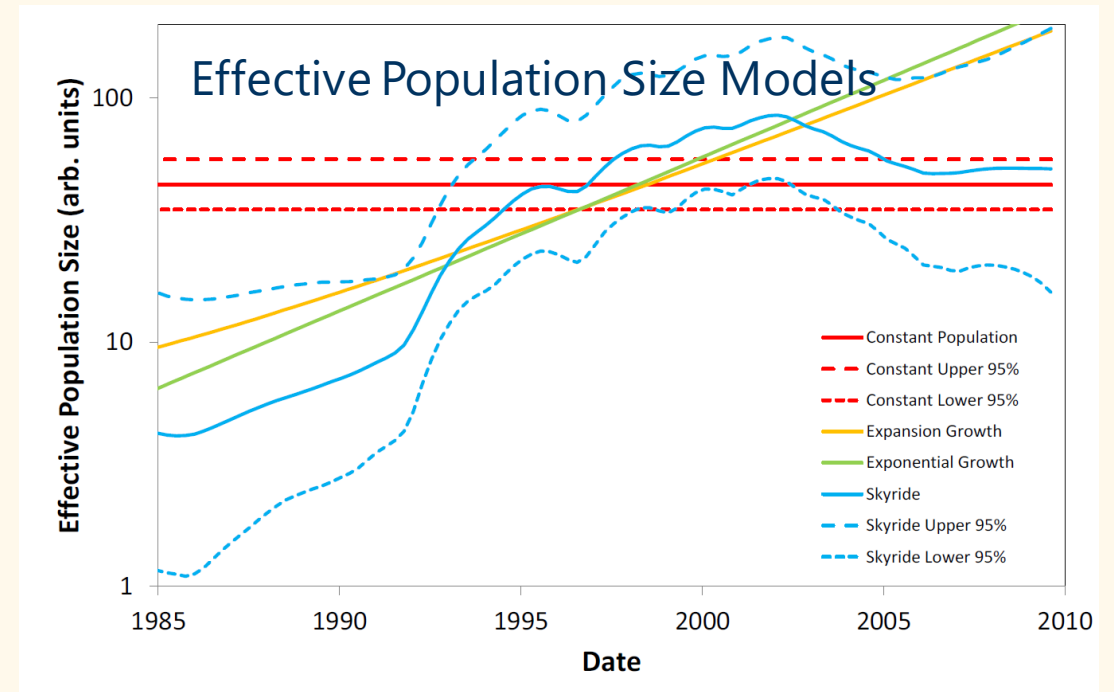
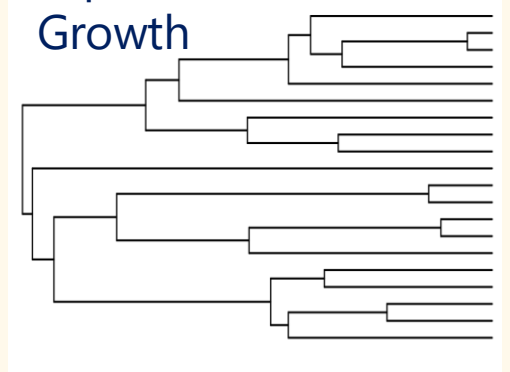
Viral Effective Population Size and Growth Rate

- Effective number of individuals (N_e) ~ Viral effective population size ~ viral diversity
- Distribution of branch lengths depends on effective population size model
- Exponential growth model applicable to within-host and epidemic situations
- TMRCA (origin time) and growth rate calculated from trees

Constant Population Size

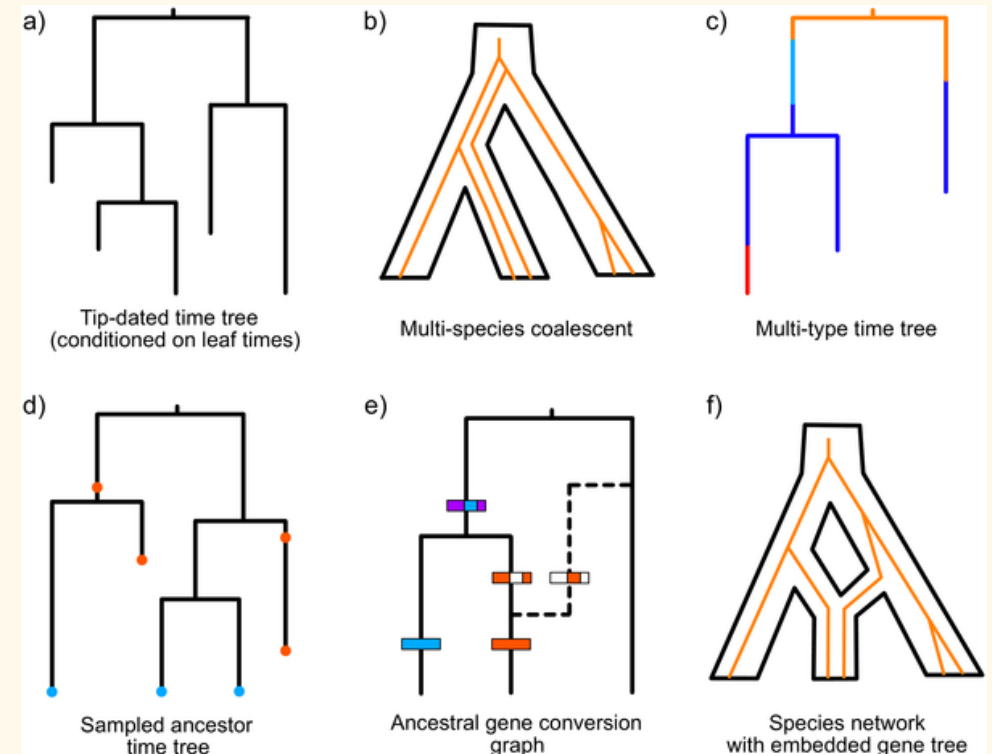


Exponential Growth



Current computational possibilities with BEAST

- BEAST uses sequence data, time scales, and other 'trait' data
- Infers mutation rates, population size over time etc (see next)
- BEAST 1 <http://beast.community/>
- BEAST 2 <http://www.beast2.org/>

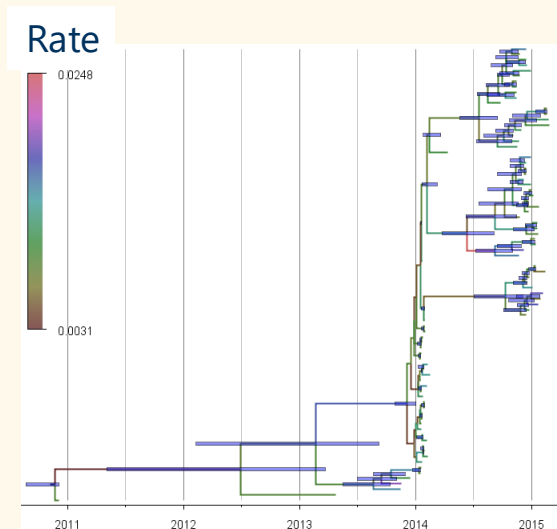


Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology 15(4): e1006650. <https://doi.org/10.1371/journal.pcbi.1006650>

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650>

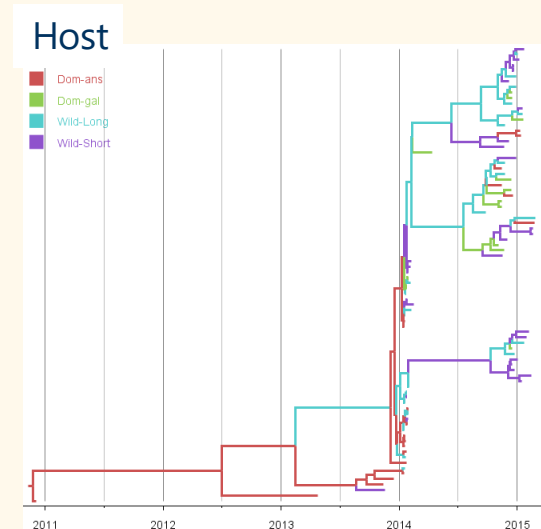
Current Methods Menu (BEAST)

Time scaled trees



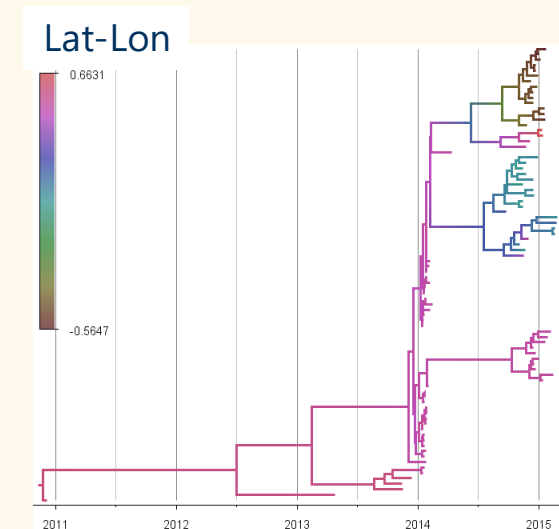
Clock models
Population models
Tree Priors
Structured coalescents

With Discrete trait



A/symmetric
BSSVS
GLM

With Continuous Trait

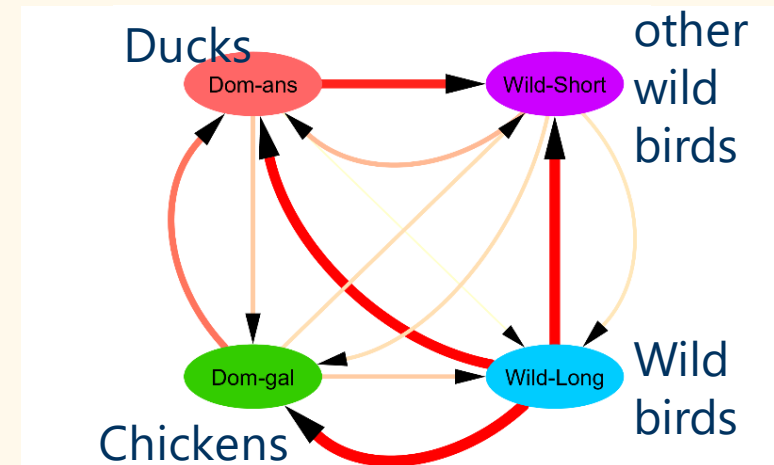
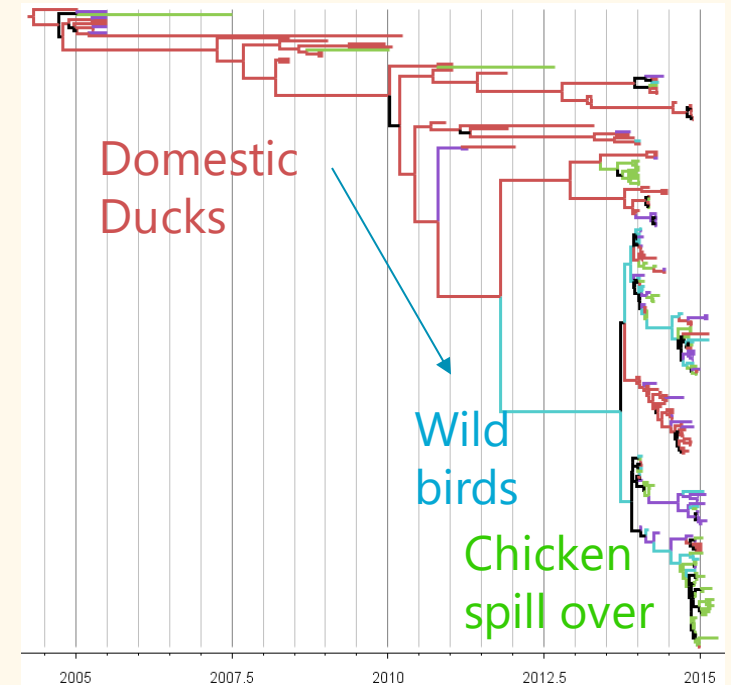


Diffusions in space
Relaxed diffusions
Warped space

Posterior set of many likely trees (MCMC)

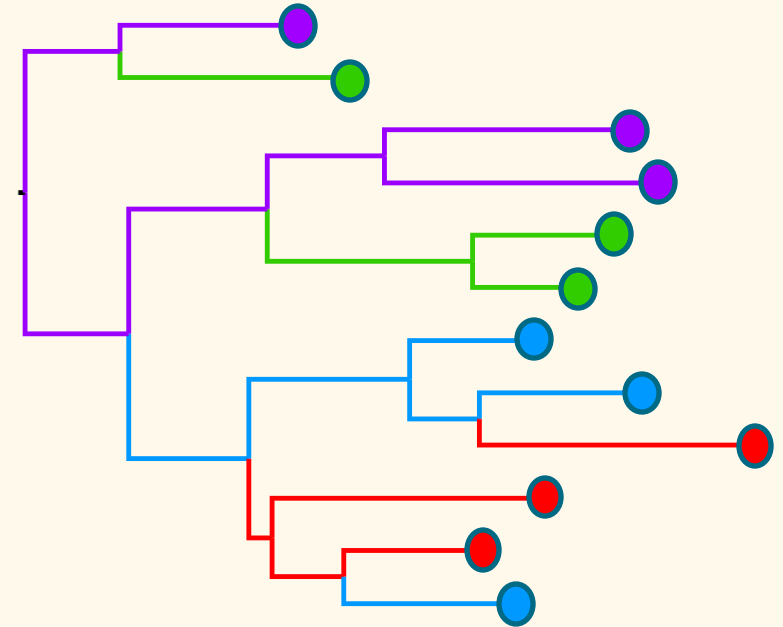
Essence of Phylodynamics (2)

- Hosts, Populations and Demes:
 - Population scale: between hosts, one sequence per individual but many individuals in e.g. a farm
 - Transmission experiment scale: per individual or pen
- 'Plain': map mutations, discrete traits, continuous traits onto trees;
 - traits do not affect the trees, but correlation can be calculated
- 'Enhanced': traits and mutations do affect the trees,
 - E.g. mutation => more pathogenicity; or vaccine / immune escape



Phylogeography and spreading patterns

- Use time scaled phylogenetic trees to infer 'who infected whom'
- Add location traits to time-scaled tree
- For discrete locations:
- Estimate transition rates between locations along branches
- Transmission pattern represented by rate matrix
- Additional: model the rate matrix as being a combination of other driving factor rate matrices (Phylogenetic Generalised Linear Model)



Transition Rate Matrix (M)

	A	B	C	D
A	-	A -> B	A -> C	A -> D
B	B -> A	-	B -> C	B -> D
C	C -> A	C -> B	-	C -> D
D	D -> A	D -> B	D -> C	-

Probability of Ancestral state (x'),
given branch length t and child state x :

$$p(x'|t) \sim e^{Mt}x$$

What else ?

(not in the practical today)

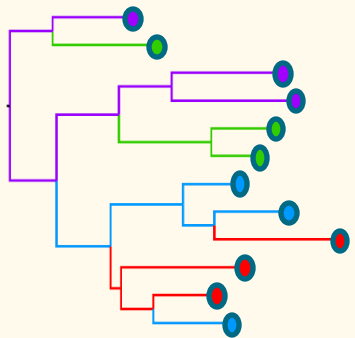
Further analyses involving BEAST

Essence of Phylodynamics (3) – Predictive Factors

- Diffusion rates between places and/or species (or other traits) can be modelled using a phylodynamic generalized linear models (GLM)

Discrete trait

- Rate matrix & predictors of rate matrix (phylogenetic GLM)
- Time-varying predictors
- Can use structured coalescent



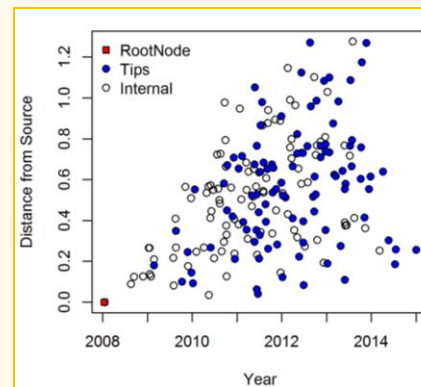
Transition Rate Matrix (M)

	A	B	C	D
A	-	A → B	A → C	A → D
B	B → A	-	B → C	B → D
C	C → A	C → B	-	C → D
D	D → A	D → B	D → C	-

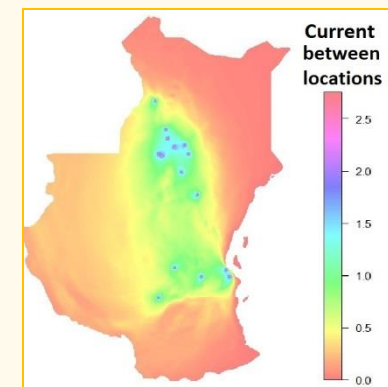
Tree with Location Traits

Continuous trait

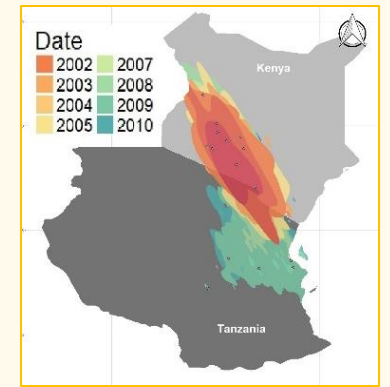
- Random walk dispersion, Brownian motion
- Correlation of branch lengths with ‘distances’ in trait space
- ‘Distances’ calculated as a path – ‘Resistance’ or ‘Conductance’



Diffusion from source



Resistance surface
(e.g. from Elevation Map)

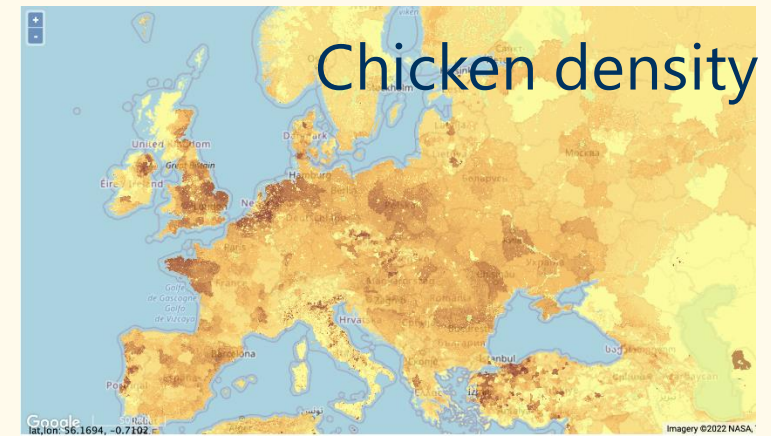


FMDV Diffusive spread
Kenya <-> Tanzania

Initial results for Avian Influenza 2020-2022 H5NX

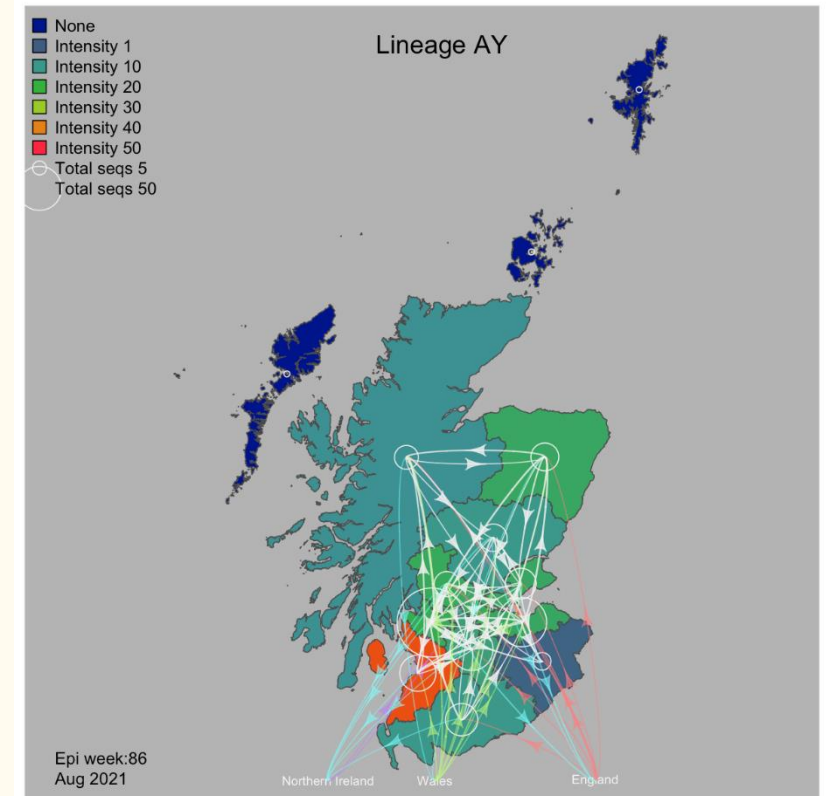
- Correlate phylodynamic dispersion with risk factors as gridded datasets (0.5 deg)
- Test virus remain in, and/or disperse towards
- Consider 33 unique risk factors in 10 groups
 - Biodiversity, Bird flyway, Climatic, Elevation, Forest, Land use, Socio-economic, Vegetation and Water
- AIV tended to remain in and to leave from areas with higher poultry and human density
- Other significant predictors:
 - Flyways of anseriformes and passeriformes, Vegetation, cropland use, urban land, broadleaf trees, and wetlands

Example predictors



Conclusions re-iterated !

- Tracking spread of infection using pathogen sequences and phylodynamics
 - Near real time surveillance sequencing
 - Global surveillance and data sharing important
 - Applicable to many measurably evolving systems: Influenza, SARS-CoV-2, FMDV..
- To make predictions or forecasts for viral spread in populations:
 - where are the current infections ? – surveillance & field
 - Imports and spreading patterns in the population ?
 - Fitness of (new) variants ? - Integrate experimental results with population scale growth rate estimates



Example of SARS-CoV-2 Delta AY lineages in Scotland calculated using whole genomes, time-scaled trees and discrete trait models

Arrow width: number of imports/exports in 7 days
Circle size: max sequences of AY Lineages in 7 days.

Background colour: "intensity" within healthboard transmissions

A flock of birds, likely swallows, is captured in flight against a clear, pale blue sky. The birds are arranged in a loose V-formation, with some leading and others following. Their wings are spread wide, and their bodies are angled downwards, suggesting they are in a steady, powerful flight. The lighting is even, highlighting the dark feathers of their wings and the lighter plumage on their chests.

Thank you

Now to the Practical !

samantha.lycett@ed.ac.uk