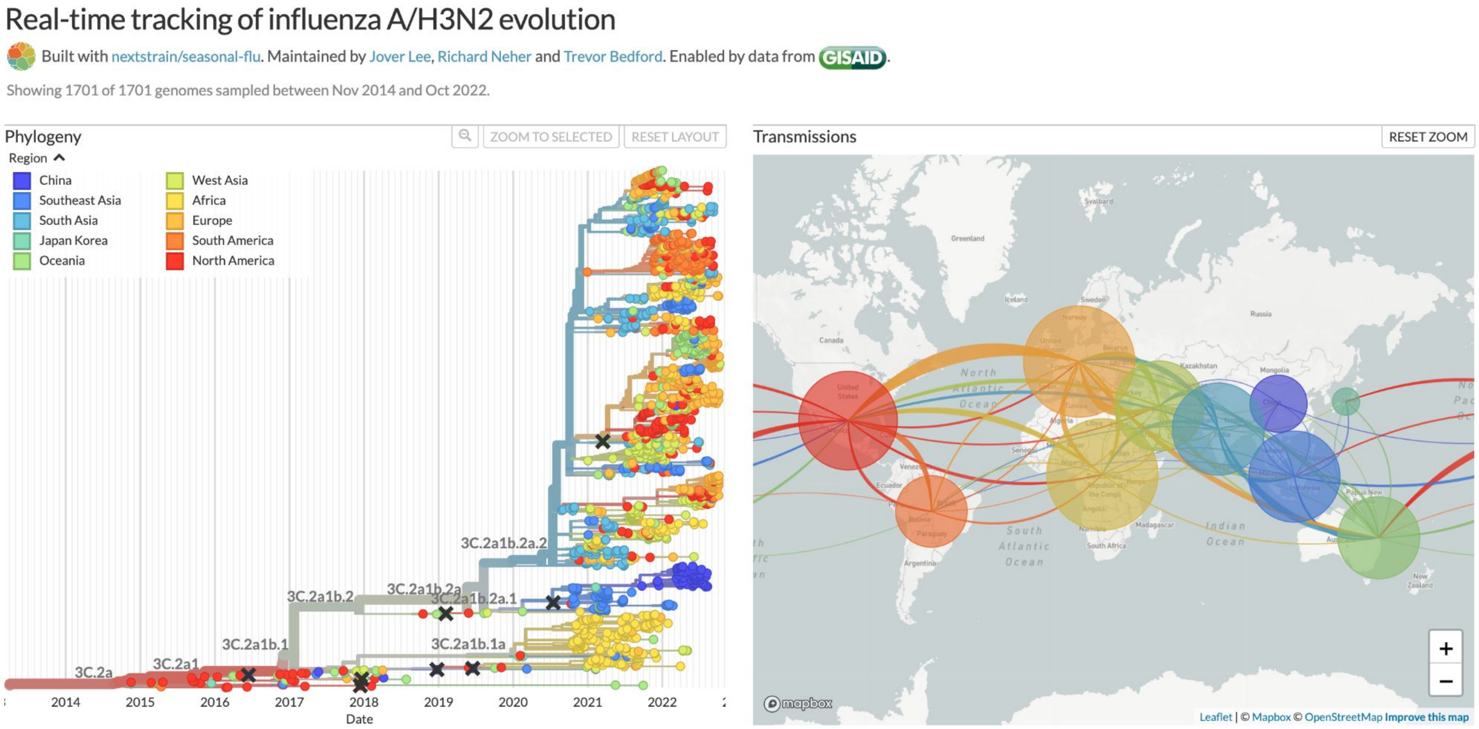
VIRAL LOCATIONS IN A TREE AND THE WORLD



Real-time tracking of influenza A/H3N2 evolution

Built with nextstrain/seasonal-flu. Maintained by Jover Lee, Richard Neher and Trevor Bedford. Enabled by data from GISAID.

Showing 1701 of 1701 genomes sampled between Nov 2014 and Oct 2022.

Figure from nextstrain.org

This section: work with **Yexuan Song, Pengyu Liu, Ailene MacPherson**

How standard (discrete) phylogeography works

Phylogeography*:* Using phylogenetic trees to infer past locations of organisms, like viruses.

Model: Location is a discrete trait. It changes along the tree under a continuous time Markov chain (CTMC) model.

The CTMC has a rate matrix, *Q,* specifying the per-unit time rate of transitions between locations *i* and *j*.

This means we can calculate $P_{ij}(t)$: probability that location *i* transitioned to location *j* on a branch of length *t*.

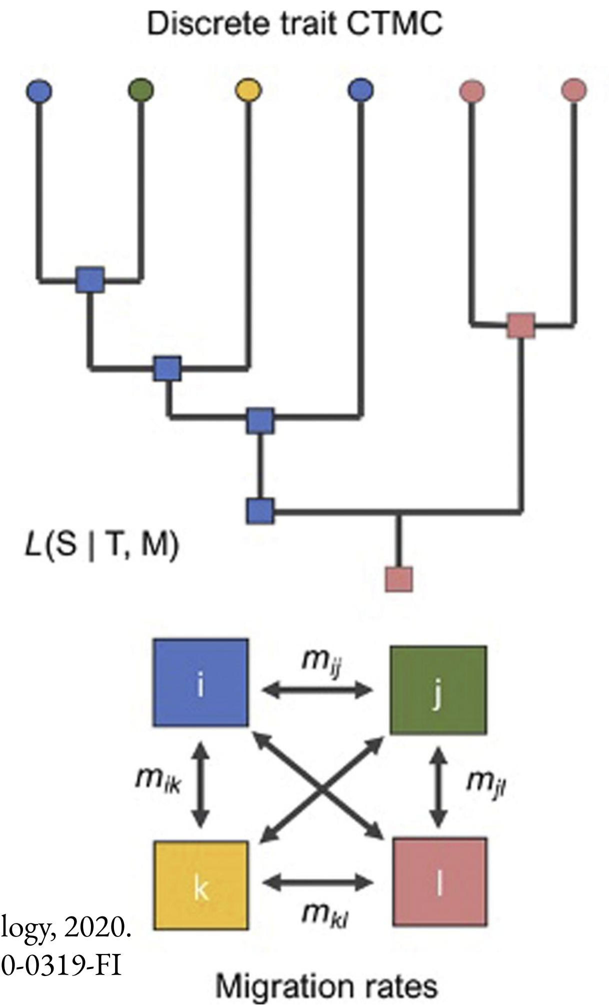
$P_{ij}(t)$ are used to assign locations to internal nodes, by maximizing the overall likelihood. *Q* can be estimated at the same time. (“Stochastic character mapping” -- Nielson, 2002).

Phylogeography illustration

Colour: location

Tip locations are observed.

Internal node locations are estimated with the CTMC.



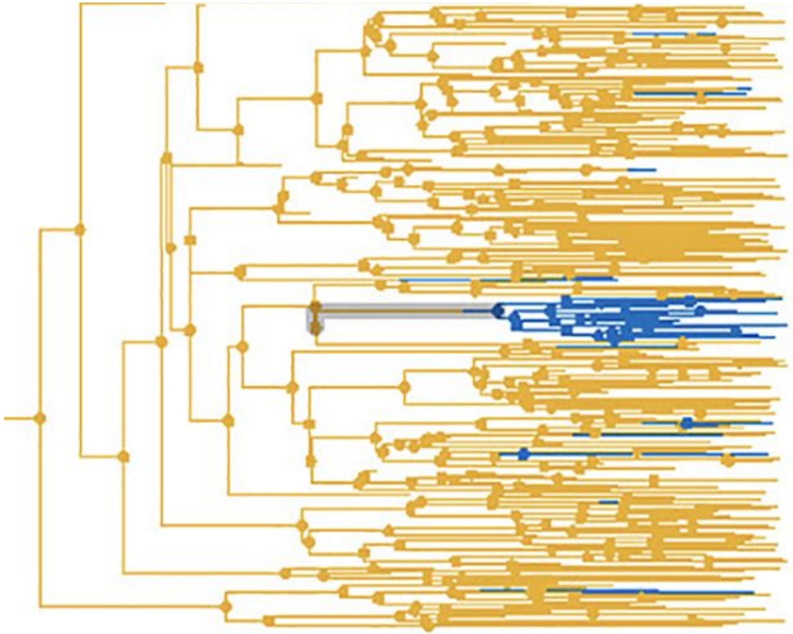
Rasmussen and Grunwald, Phytopathology, 2020.
https://doi.org/10.1094/PHYTO-07-20-0319-FI

Sampling: a challenge

What if some locations sample more than others? They get more tips.

This can impact the inference of where viruses were in the past.

Example: COVID-19. Some locations test more than others. Some locations have more resources to sequence the virus. Some share their data more, or less.



Two questions

1. How does sampling bias impact phylogeographic results?

2. How can we adjust for sampling rates to improve phylogeographic results?

A simulation study: how much does sampling impact phylogeography?

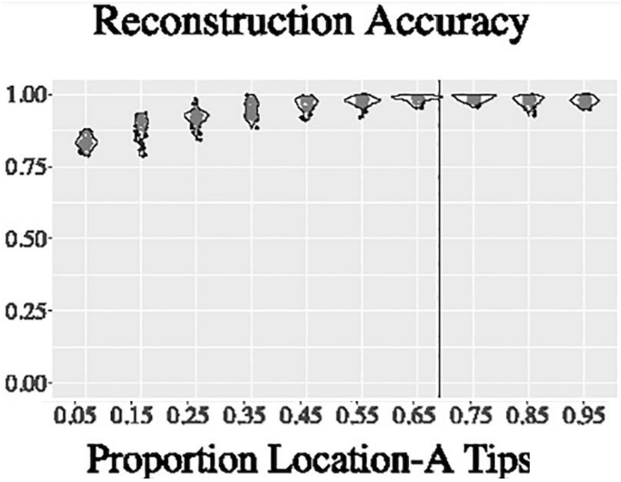
- We simulate two locations: yellow and blue. We know the true locations of all the nodes.
- We remove tips from the simulated trees to simulate different sampling fractions.
- We reconstruct the node locations using the standard CTMC.
- How wrong is the reconstruction, and in which ways?
- We examine an Ebola dataset: how much does sampling matter?

The impact of sampling bias on viral phylogeographic reconstruction. Pengyu Liu,Yexuan Song,Caroline Colijn,Ailene MacPherson. *PLOS Global Public Health,* 2022

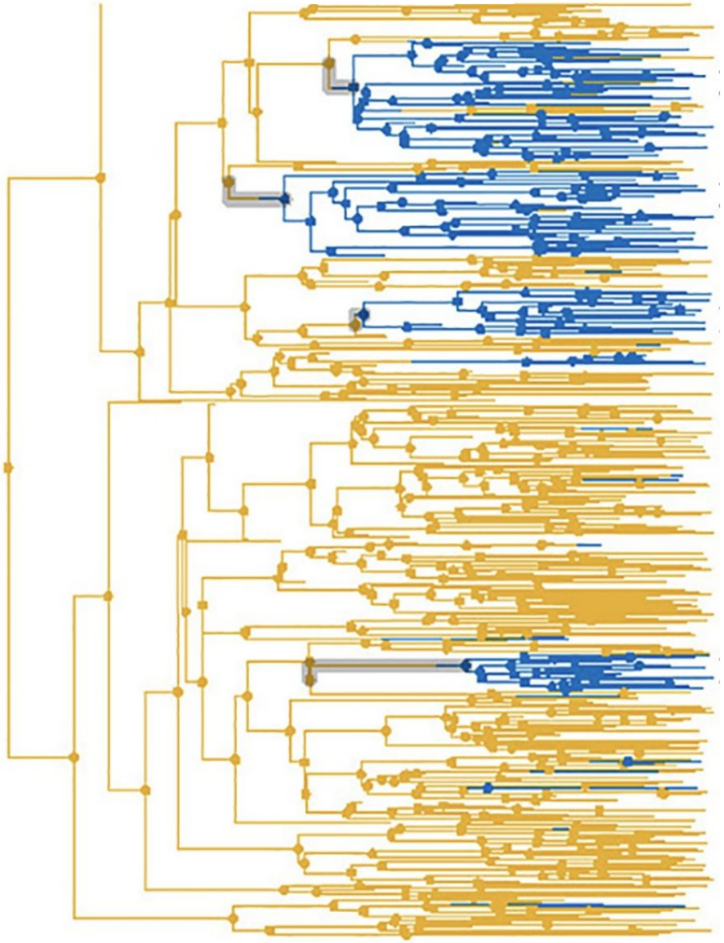
PLOS GLOBAL PUBLIC HEALTH

Low migration rate

When transitions are rare, the colours are “grouped” in the tree. The overall accuracy is high, and does not depend much on the sampling bias.


Reconstruction Accuracy

True tree

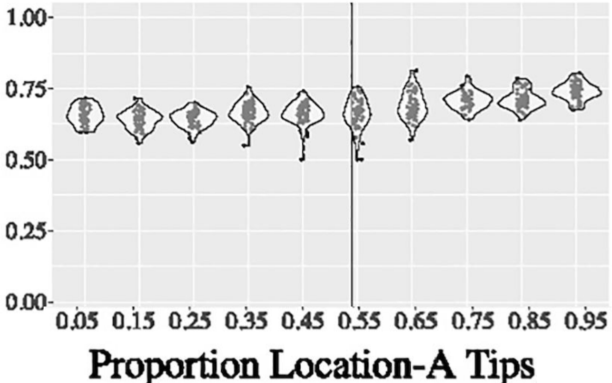


High migration rate

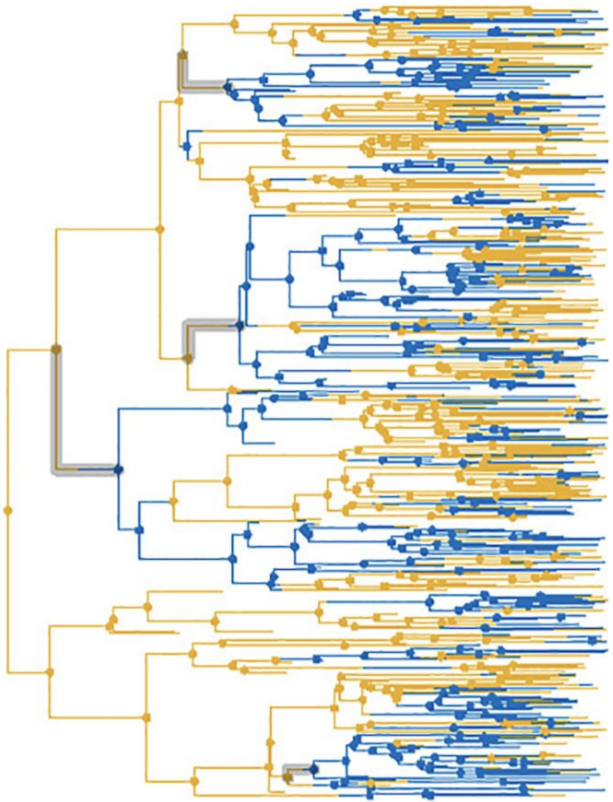
If the transitions are frequent, the tip locations don’t carry as much information about the internal node locations.

The accuracy is worse, but it does not depend strongly on sampling.

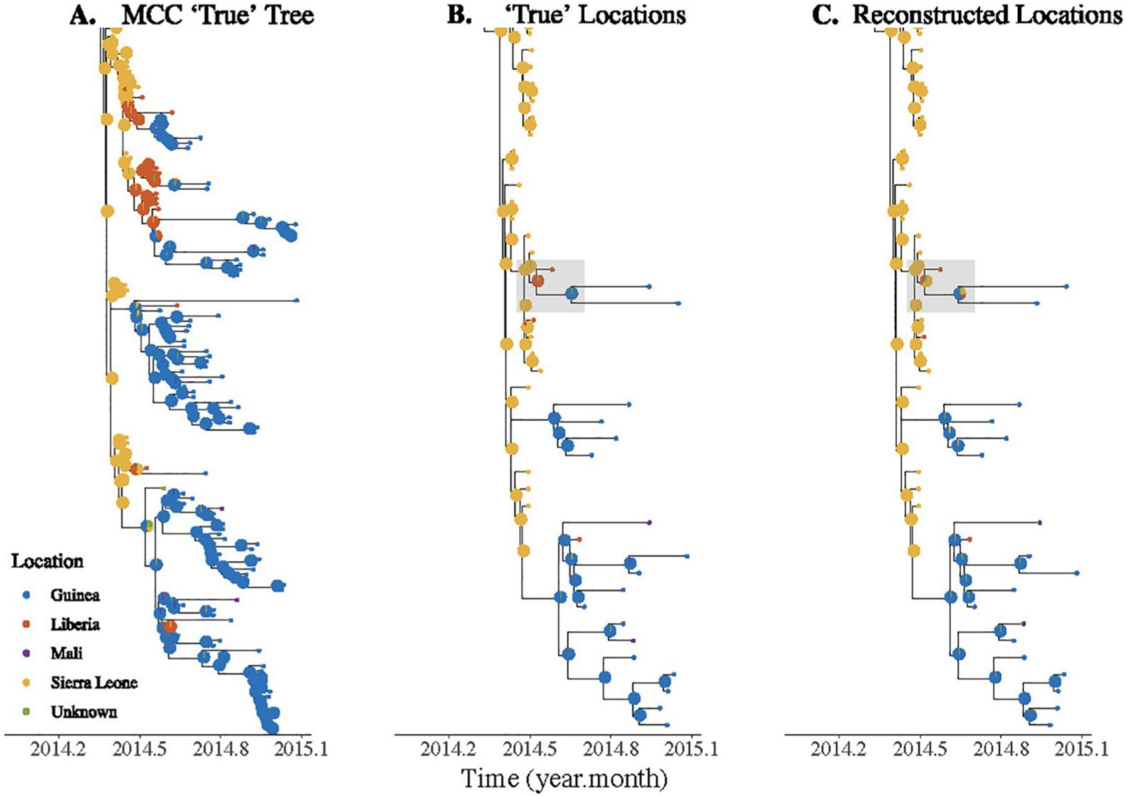
Reconstruction accuracy



True tree



Sampling can affect whether we detect “key migration events”



Simulation results

1. Standard methods did a good job with overall accuracy, which didn’t depend much on the sampling bias
2. Low migration rate: overall accuracy is really high at whatever sampling bias
3. High migration rate: accuracy is lower, but not very dependent on sampling bias
4. **However,** you can really get key importation events wrong.

How can we adjust for sampling bias?

We need a new mathematical model.

Background: Binary state-dependent speciation and extinction (BiSSE) family of models (Maddison, Midford, Otto. *Systematic Biology* 2007)

Underlying process: multi-type branching process.

Each “state” (here, location; general: value of a trait) has its own branching rate and death rate, which can be estimated from data.

Y. Song MSc: We combined two advances: (1) extension to incomplete observation (Fitzjohn et al); (2) description of how to estimate node locations in the model (Freyman & Hohna).

This section: work with **Yexuan Song, Ailene MacPherson**

The mathematical method

Key ingredients:

- D_{BNi} - probability that a lineage *N* in state *i* at time *t* gives the observed descendants.
- D_{FNi} - probability that a lineage *N* in state *i* at time *t* arose from the observed ancestor.
- E_{Ni}- probability that a lineate *N* in state *i* at time *t* dies out (not observed by the present)

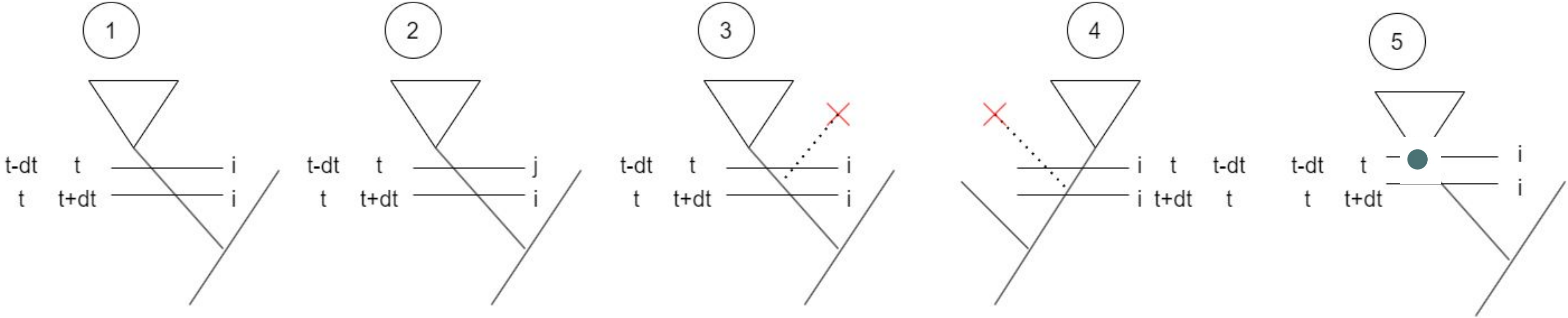
Approach: derive differential equations for these terms.

Use the *Ds* to assign the states at the internal nodes.

Differential equations for the *D* terms

There are 4 possibilities:

- 1. No state change, no speciation, no sampling.
- 2. There is a stage change but no speciation or sampling.
- 3, 4. Speciation: only the left or right lineage survives.
- 5. The lineage gets sampled at time t.



$$
\begin{aligned}
D_{BNi}(t+\Delta t) \approx D_{BNi}(t) & \\
&+ (1-\mu\Delta t)[(1-q_{ij}\Delta t)(1-\lambda\Delta t)D_{BNi}(t) \quad \text{no change} \\
&+ q_{ij}\Delta t(1-\lambda\Delta t)D_{BNj}(t) \quad \text{state change} \\
&+ 2(1-q_{ij}\Delta t)\lambda\Delta t E_i(t) D_{BNi}(t)] \quad \text{Speciation; one goes extinct} \\
&+ \mu\Delta t(0) + O(\Delta t^2) \quad \text{Extinction}
\end{aligned}
$$

$$
\frac{d}{dt}D_{BNi}(t) = -(\lambda+\mu+q_{ij})D_{BNi}(t) + 2\lambda E_i D_{BNi}(t) + q_{ij}D_{BNj}(t)
$$

where λ: branching rate; μ: death rate; f_i: location-specific sampling fraction; q_{ij}: transition rate.

Initial conditions:

Tips: $D_{BNi}(0) = f_i$ if the tip is in state i. Internal nodes: probability of giving rise to clades C_1 and C_2 is the product of probabilities D.

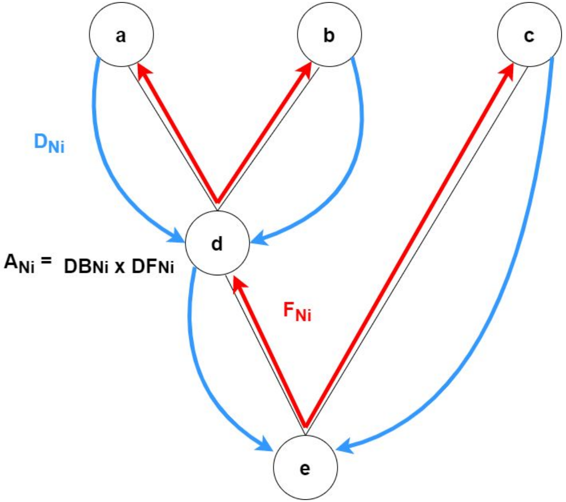
Method overview

Take the same approach for forward equations for $D_{FNi.}$ and for the extinction probability $E_{i.}$

Use Freyman and Hohna's approach to "stochastic character mapping"-- assigning states to the internal nodes:

A_{Ni} is the probability that a node is in state *i:* $A_{Ni} = D_{FNi}D_{BNi}$

Assign the max-probability state to each internal node.



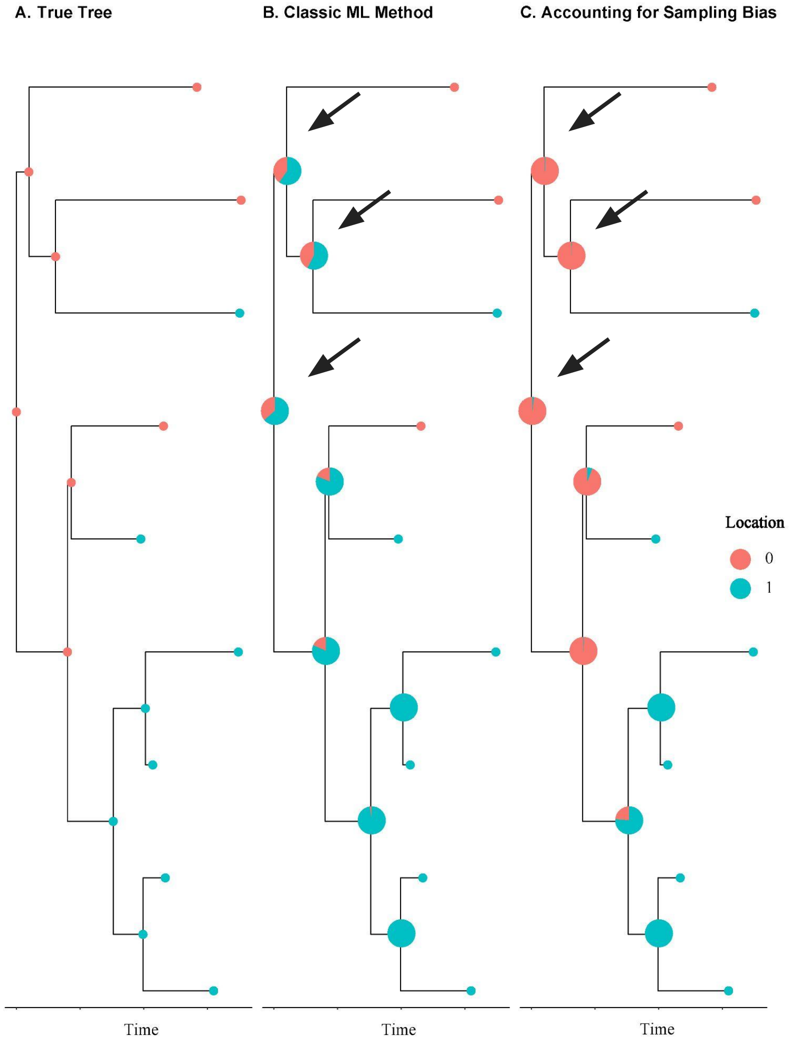
Results: better than standard meth

Simulate: blue location sampled 1.7x more than the red location.

Standard phylogeography over-estimates the number of blue nodes.

The new method gets the right locations for the red internal nodes.

This is a proof of principle: working on large-scale implementation and testing.



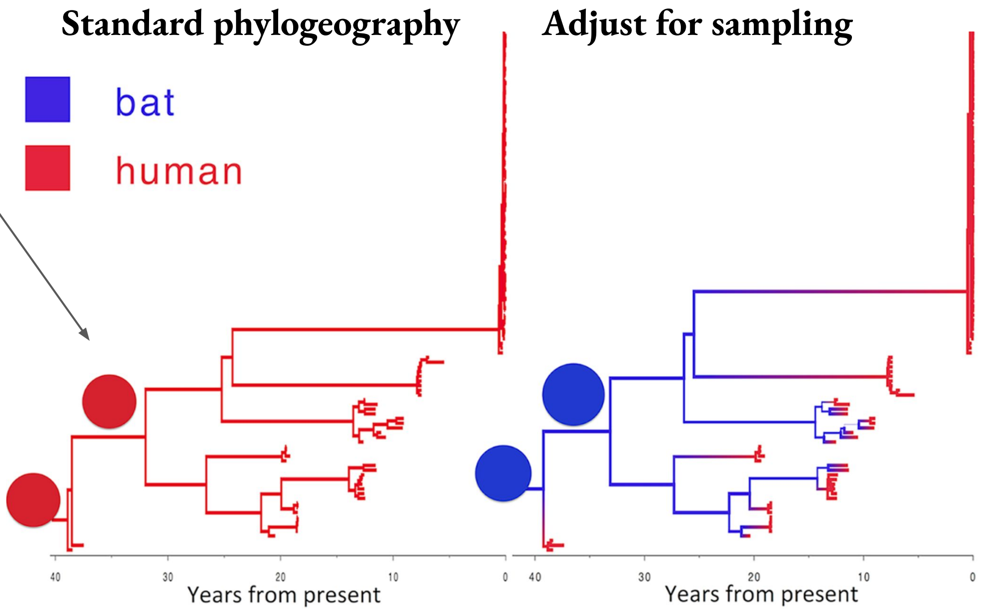
Intuitively, why do all these differential equations help?

Standard phylogeography: humans all the way back.

This doesn't account for the fact that if it *had* been in humans, it would have been sampled over all those years.

Adjusting for sampling fraction: few observations means higher likelihood that it was in the bats.

Location: which animal is the Ebola virus in? Bats or humans?



Shamelessly taken from: de Maio et al, *New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation.* PLOS Genetics, 2015