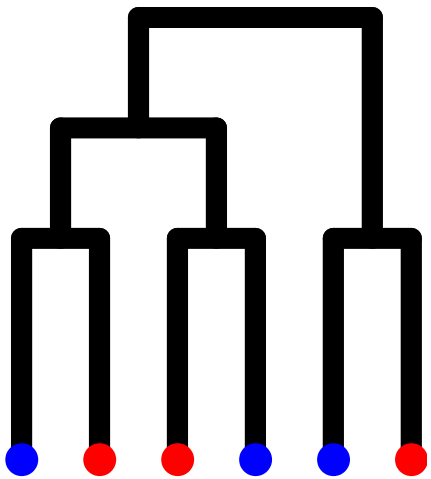
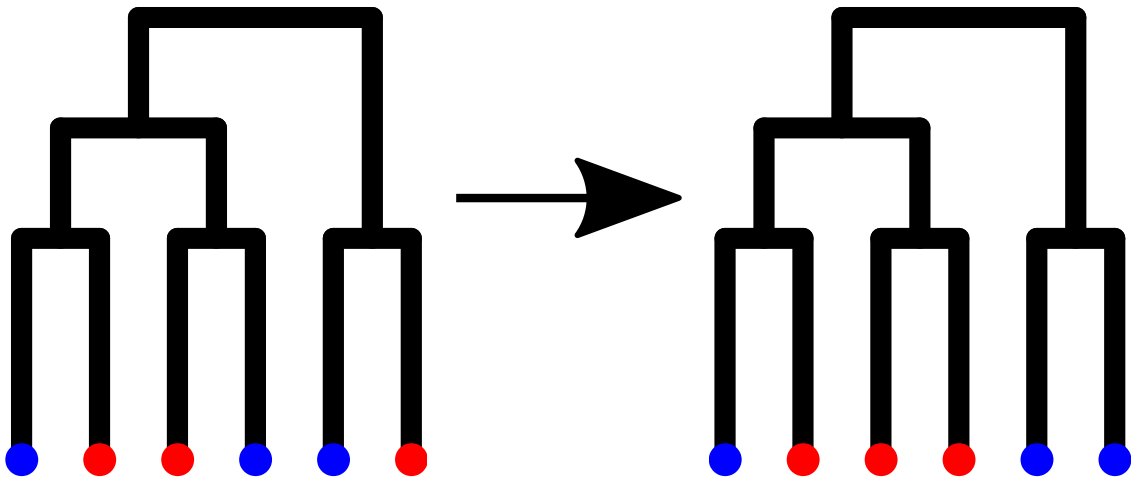
ETH zürich

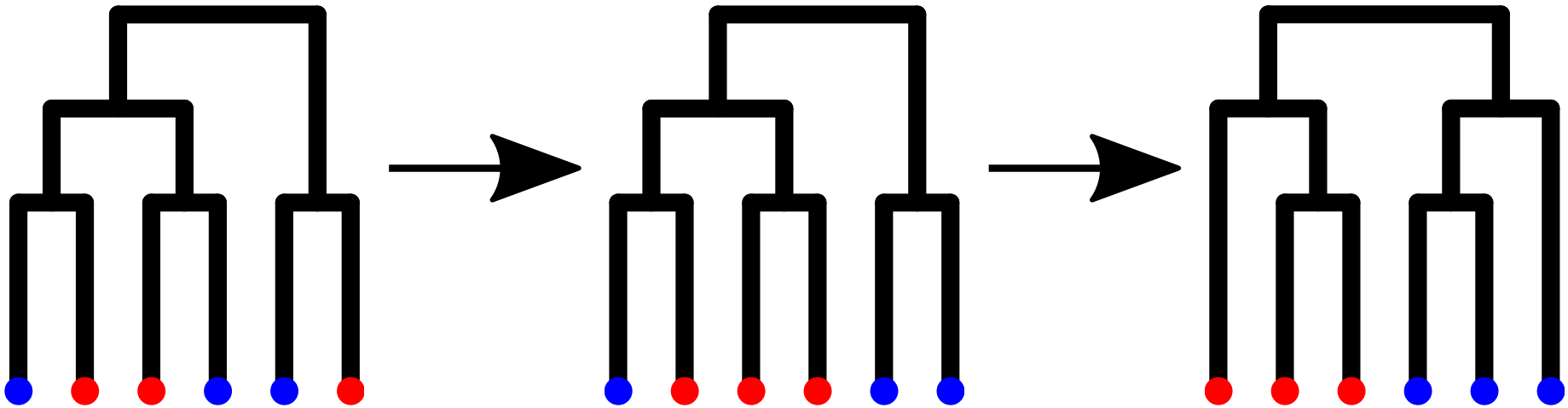
Structured coalescent approximations



- unstructured models allow to describe well mixed populations

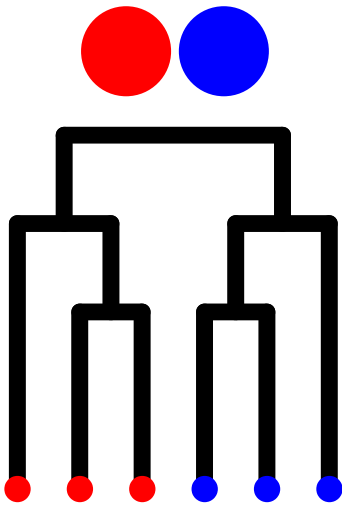


- unstructured models allow to describe well mixed populations



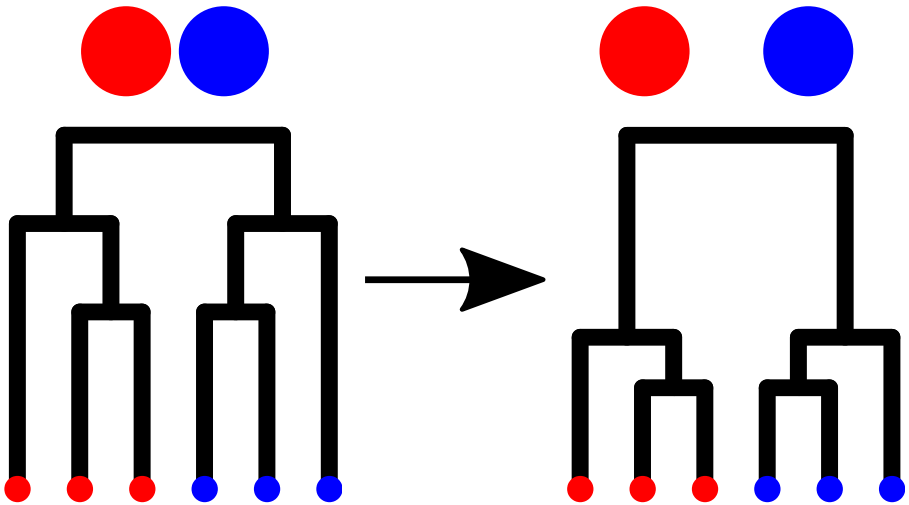
- unstructured models allow to describe well mixed populations

Why are structured models needed?



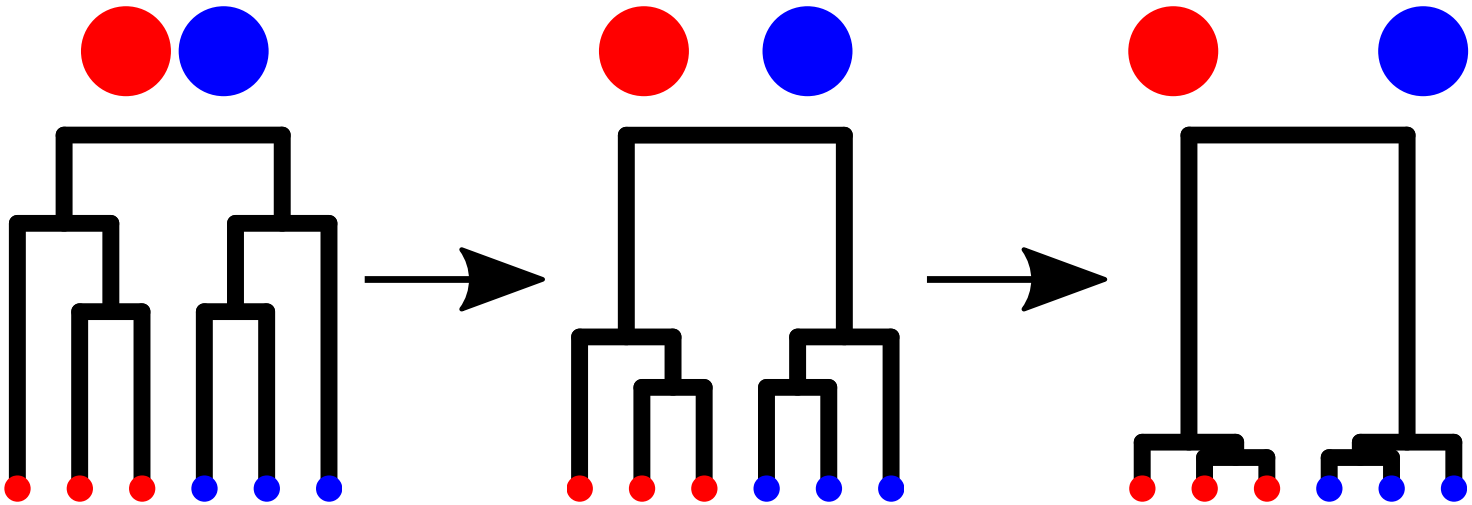
- the more dominant structure becomes...

Why are structured models needed?



- the more dominant structure becomes..

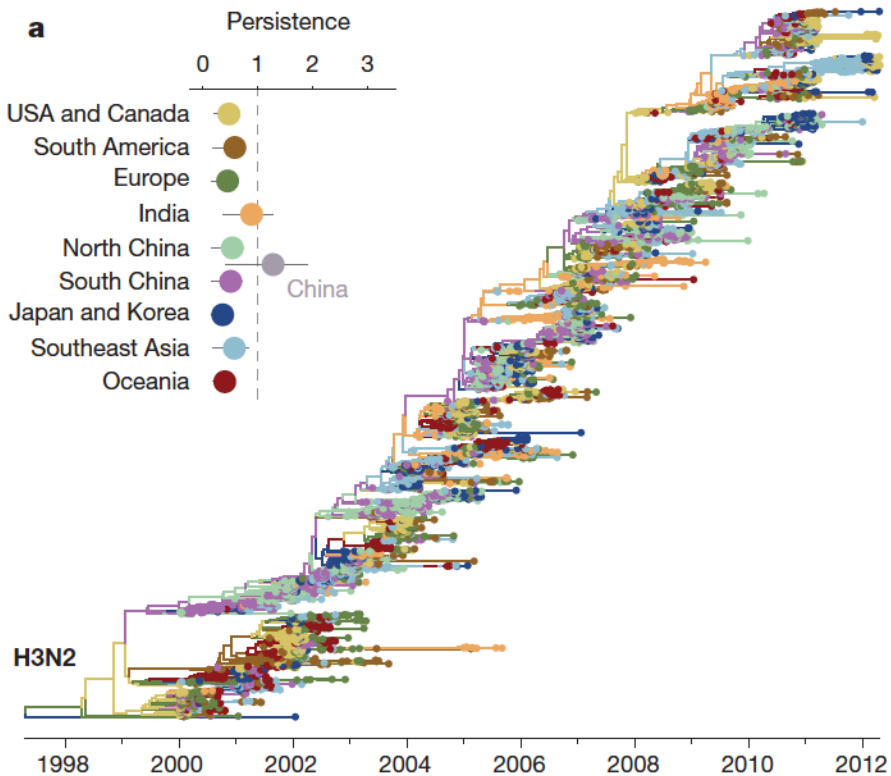
Why are structured models needed?



- the more dominant structure becomes..
- ... the less appropriate unstructured models become in describing the process that generated the tree

How does structure arise

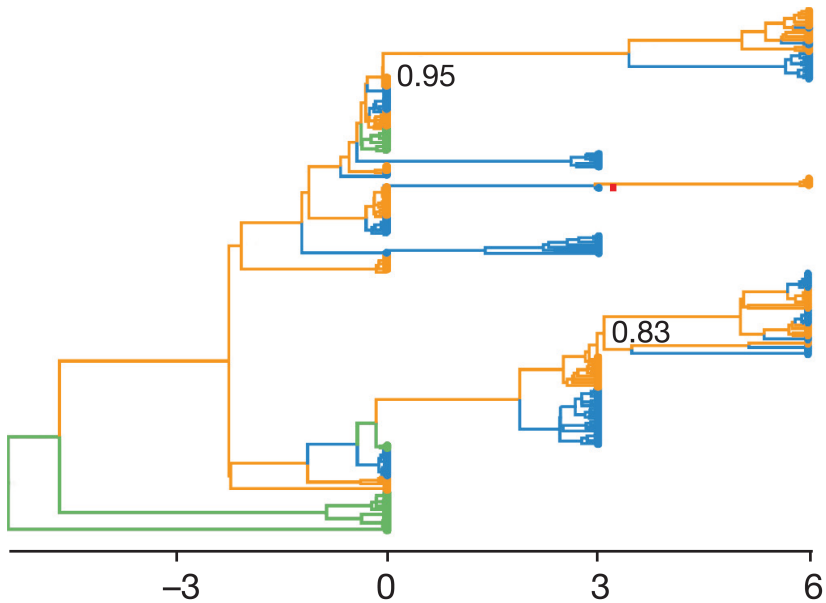
some examples: Global migration of H3N2



Bedford 2015 et al., Nature (2015)

How does structure arise

some examples: Within host clustering of HIV



Lorenzo-Redondo & Fryer et al., Nature (2015)

How does structure arise

other possible reasons:

- different host species
- different age groups
- different stages of a disease

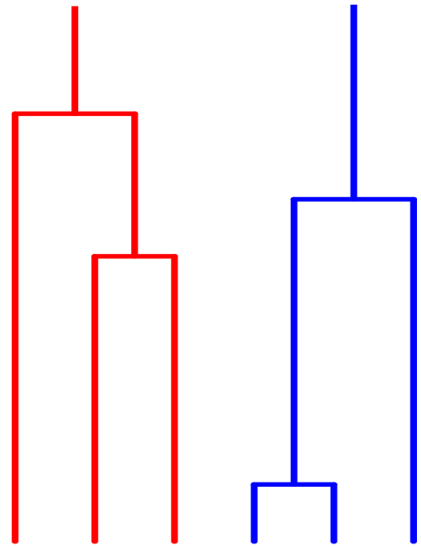
...

The structured coalescent

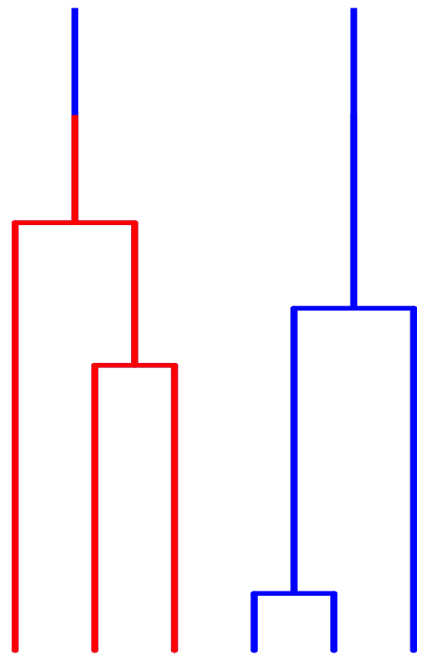
The structured coalescent describes a coalescent process in sub-populations between which individuals can migrate.

Coalescence in sub-population

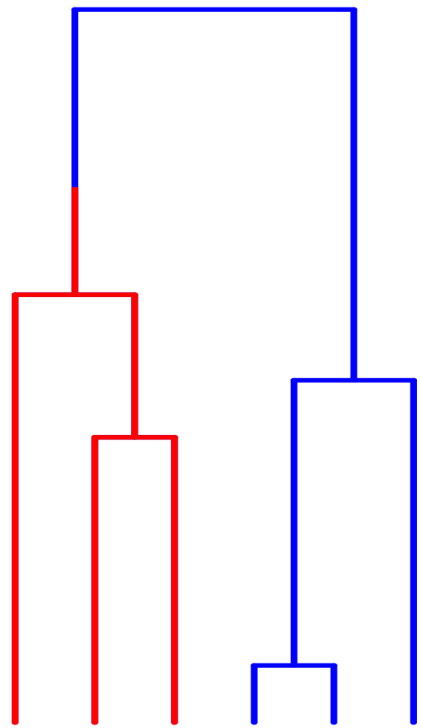
Coalescence in sub-population



Migration between sub-populations



Coalescence in sub-population



Calculation of the probability of a tree under the structured coalescent

- Given a coloured tree, a set of effective population sizes, migration rates and sampling states, the probability of a phylogeny under the structured coalescent can be calculated.

$$P(T, C|\bar{N}, \bar{M}, L)$$

Calculation of the probability of a tree under the structured coalescent

- Given a coloured tree, a set of effective population sizes, migration rates and sampling states, the probability of a phylogeny under the structured coalescent can be calculated.
- However the coloured trees must be sampled using MCMC

Calculation of the probability of a tree under the structured coalescent

- Given a coloured tree, a set of effective population sizes, migration rates and sampling states, the probability of a phylogeny under the structured coalescent can be calculated.
- However the coloured trees must be sampled using MCMC

$$P(T|\bar{N},\bar{M},L) = \int_C P(T,C|\bar{N},\bar{M},L)dC,$$

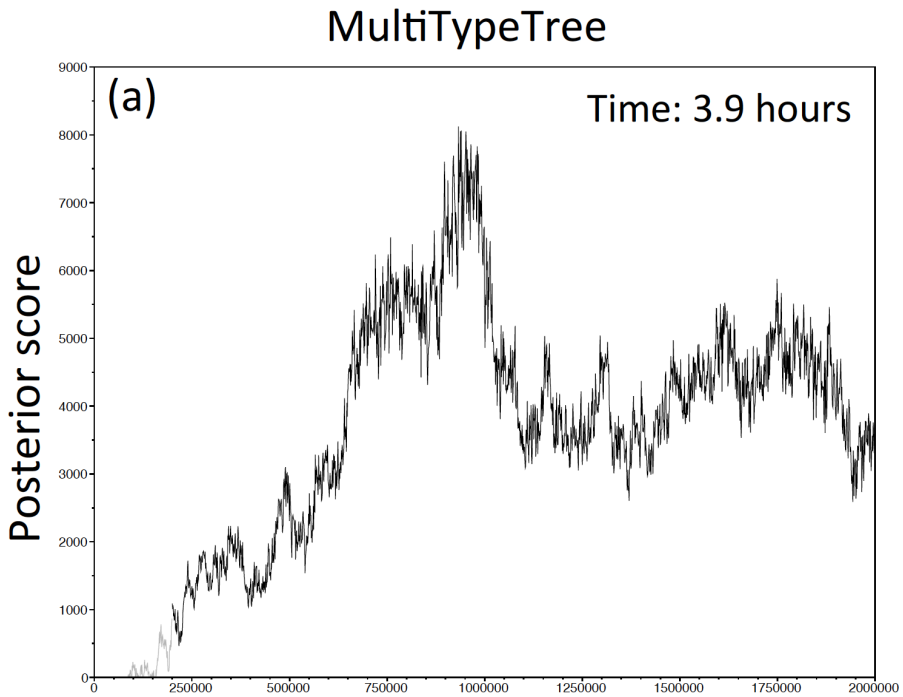
Calculation of the probability of a tree under the structured coalescent

- Given a coloured tree, a set of effective population sizes, migration rates and sampling states, the probability of a phylogeny under the structured coalescent can be calculated.
- However the coloured trees must be sampled using MCMC

$$P(T|\bar{N},\bar{M},L) = \int_C P(T,C|\bar{N},\bar{M},L)dC,$$

- With increasing number of events or different colors this sampling of colors on a tree becomes problematic.

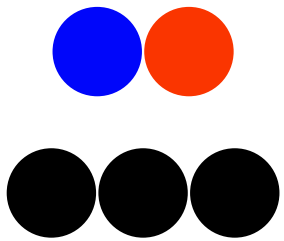
Calculation of the probability of a tree under the structured coalescent



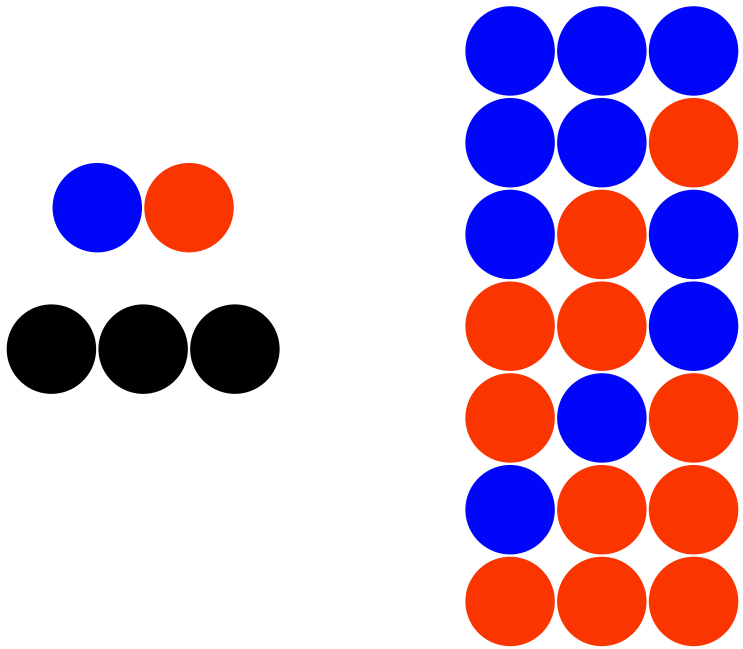
De Maio et al., PLoS Genetics (2016)

Why not just calculate $P(T|\bar{N}, \bar{M}, L)$ directly?

Direct calculation of $P(T|\bar{N}, \bar{M}, L)$



Direct calculation of $P(T|\bar{N}, \bar{M}, L)$



- Requires $states^{lineages}$ number of ODEs to be solved. (1048576 for 2 states and 10 lineages)

Calculation of $P(T|S, M, \Lambda)$ using ODEs

$$\frac{dP_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)}{dt}$$

$$= \sum_{i=1}^{n} \sum_{a=1}^{m} \Big(\mu_{a l_i} P_t(L_1 = l_1, ..., L_i = a, ..., L_n = l_n, T)$$

$$- \mu_{l_i a} P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T) \Big)$$

$$- \sum_{a=1}^{m} \lambda_a \binom{k_a}{2} P_t(L_1 = l_1, ..., L_i = l_i, ..., L_n = l_n, T)$$

Müller et al., MBE (2017)

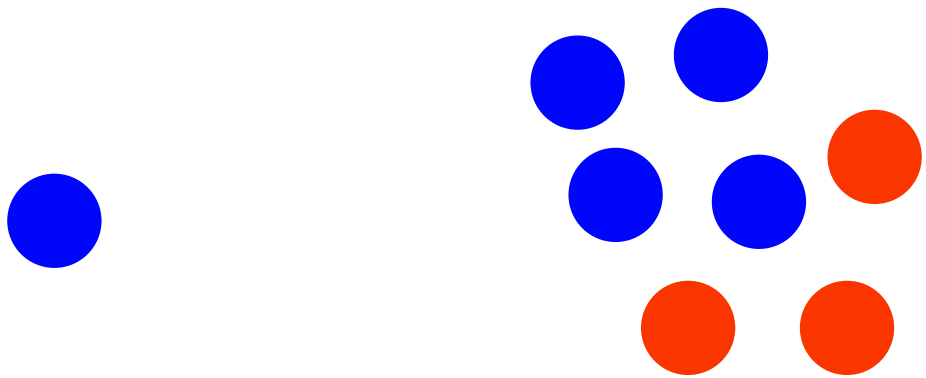
- MCMC sampling of coloured trees is not feasible for dataset with many states or many migration events

- MCMC sampling of coloured trees is not feasible for dataset with many states or many migration events
- ODE approach is not feasible for more than 10 lineages and three or four states.

- MCMC sampling of coloured trees is not feasible for dataset with many states or many migration events
- ODE approach is not feasible for more than 10 lineages and three or four states.
- Use approximations instead.

What can be approximated?

- Assume that the location of one lineage does not depend on the location of one other lineage, but just on the sum of lineages in other states.



What can be approximated?

- Assume that the location of one lineage does not depend on the location of one other lineage, but just on the sum of lineages in other states.
- This allows to describe the marginal probability of a lineage being in a location over time by solving differential equations

What can be approximated?

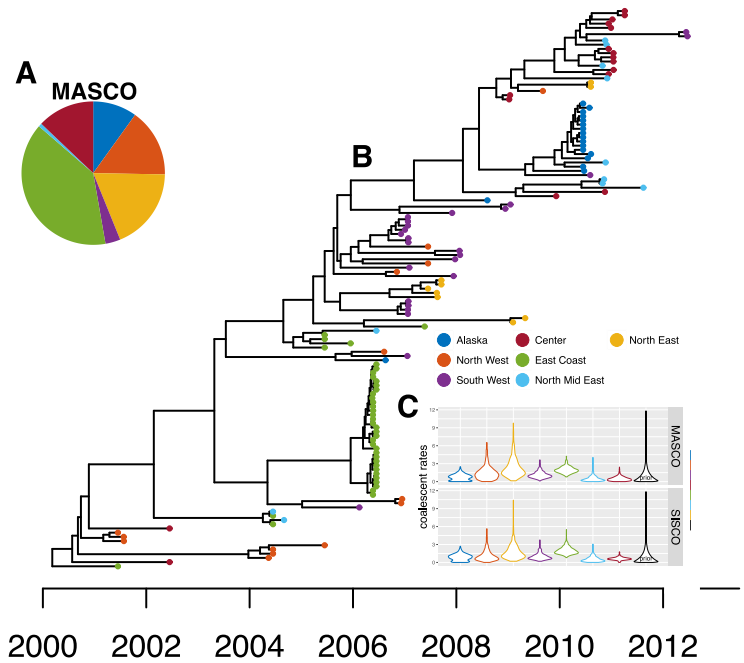
- Assume that the location of one lineage does not depend on the location of one other lineage, but just on the sum of lineages in other states.
- This allows to describe the marginal probability of a lineage being in a location over time by solving differential equations
- **MASCOT: M**arginal **A**pproximation of the **S**tructured **CO**alescen**T**

MASCOT differential equation

$$\frac{d}{dt}P(L_i = l_i, T) = \sum_{a=1}^{m}\Big(\mu_{al_i}P_t(L_i = a, T) - \mu_{l_i a}P_t(L_i = l_i, T)\Big)$$
$$- P_t(L_i = l_i, T)\Big(\lambda_{l_i}\sum_{\substack{k=1\\k\neq i}}^{n} P_t(L_k = l_i \,|\, T)$$
$$+ \sum_{a=1}^{m}\frac{\lambda_a}{2}\sum_{\substack{j\neq i\\j=1}}^{n}\sum_{\substack{k\neq j,i\\k=1}}^{n} P_t(L_j = a \,|\, T)P_t(L_k = a \,|\, T)\Big)$$

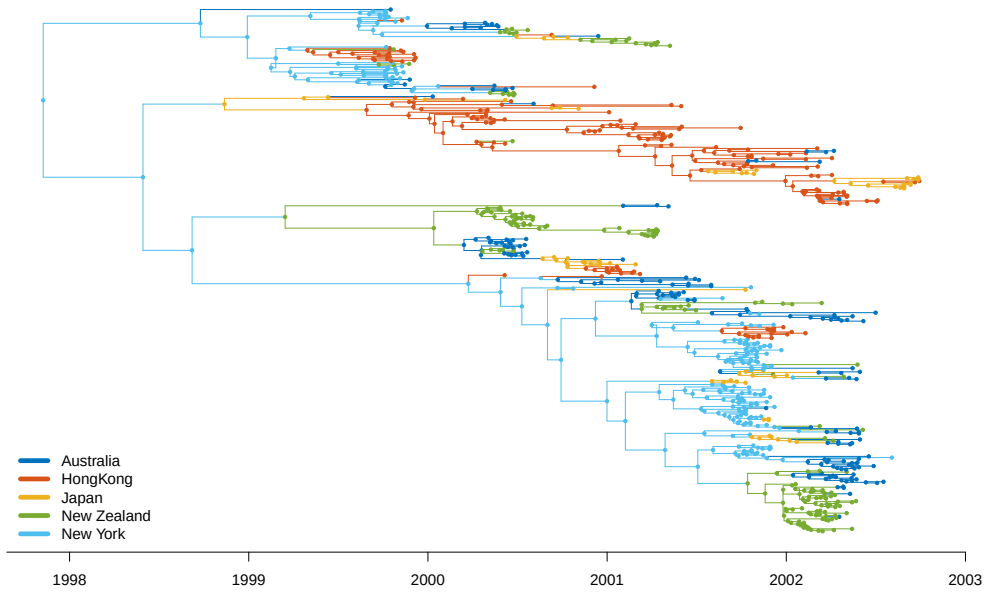
Müller et al., MBE (2017)

MASCOT allows to infer migration an coalescent rates with many states: Avian Influenza



Müller et al., MBE (2017)

MASCOT allows to infer migration an coalescent rates with many states: H3N2



Müller et al., in preperation (2017)

Tutorial

https://taming-the-beast.github.io/tutorials/Mascot-Tutorial/