Population divergence estimation

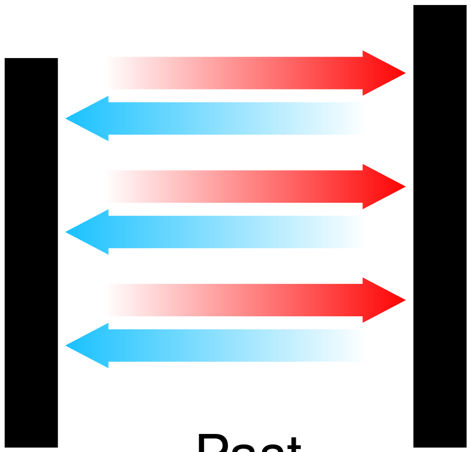
using individual lineage label switching

Peter Beerli Scientific Computing, Florida State University Twitter:@peterbeerli

Gene flow

Neanderthal

’Modern’ human



Twitter @peterbeerli

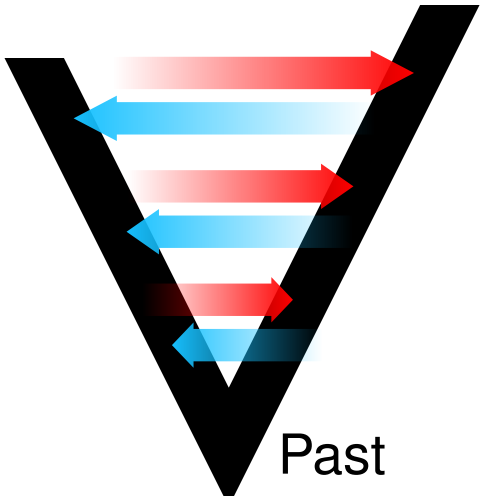
Data: Lego pictures from the net

Gene flow

Neanderthal

'Modern' human

-30,000 years

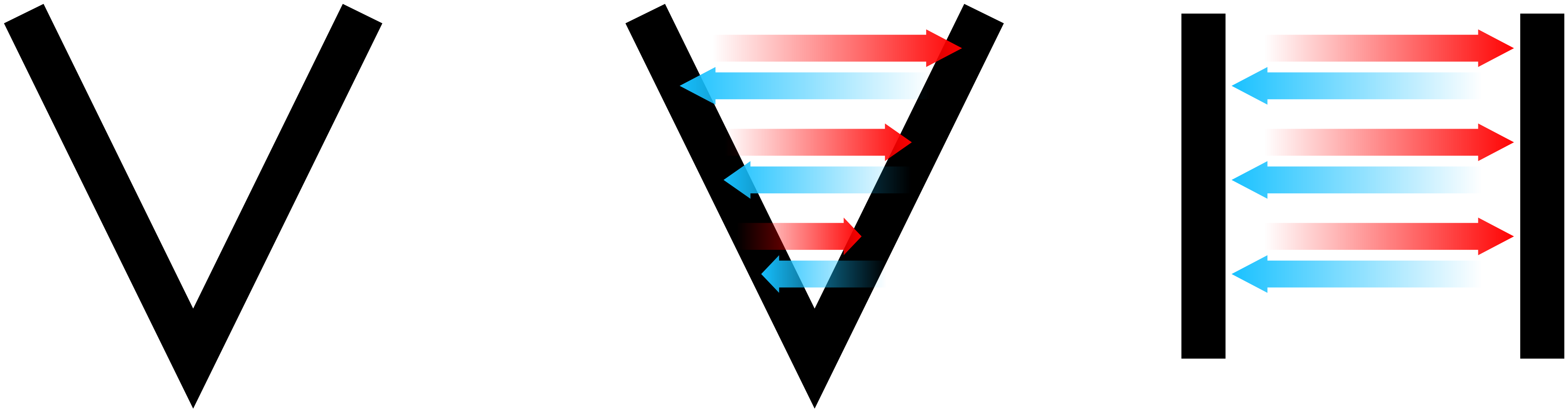


Present

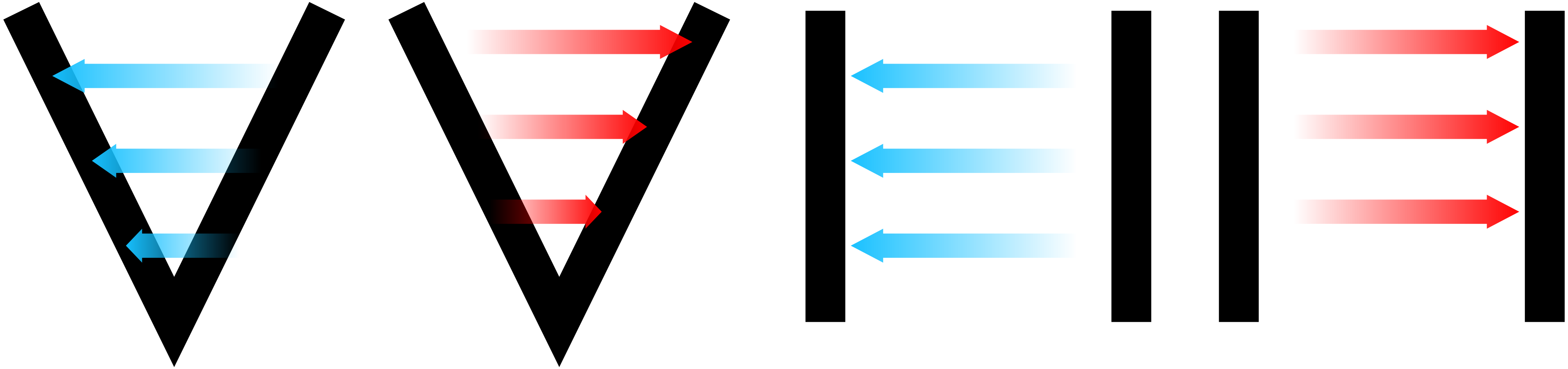
Past

Data: Lego pictures from the net

Different structural models

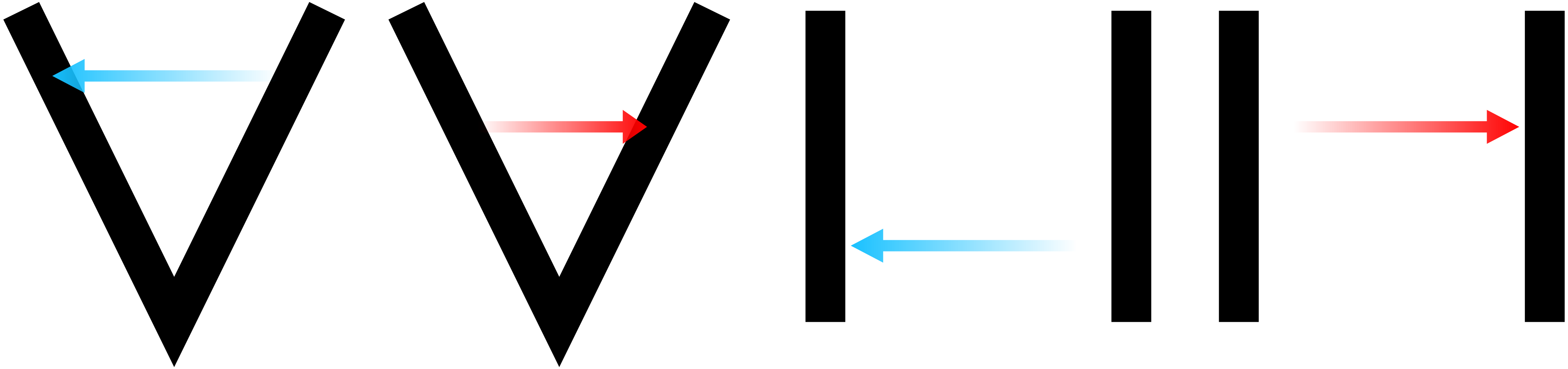


More different structural models



©2017 Peter Beerli
Twitter @peterbeerli

Even more different structural models

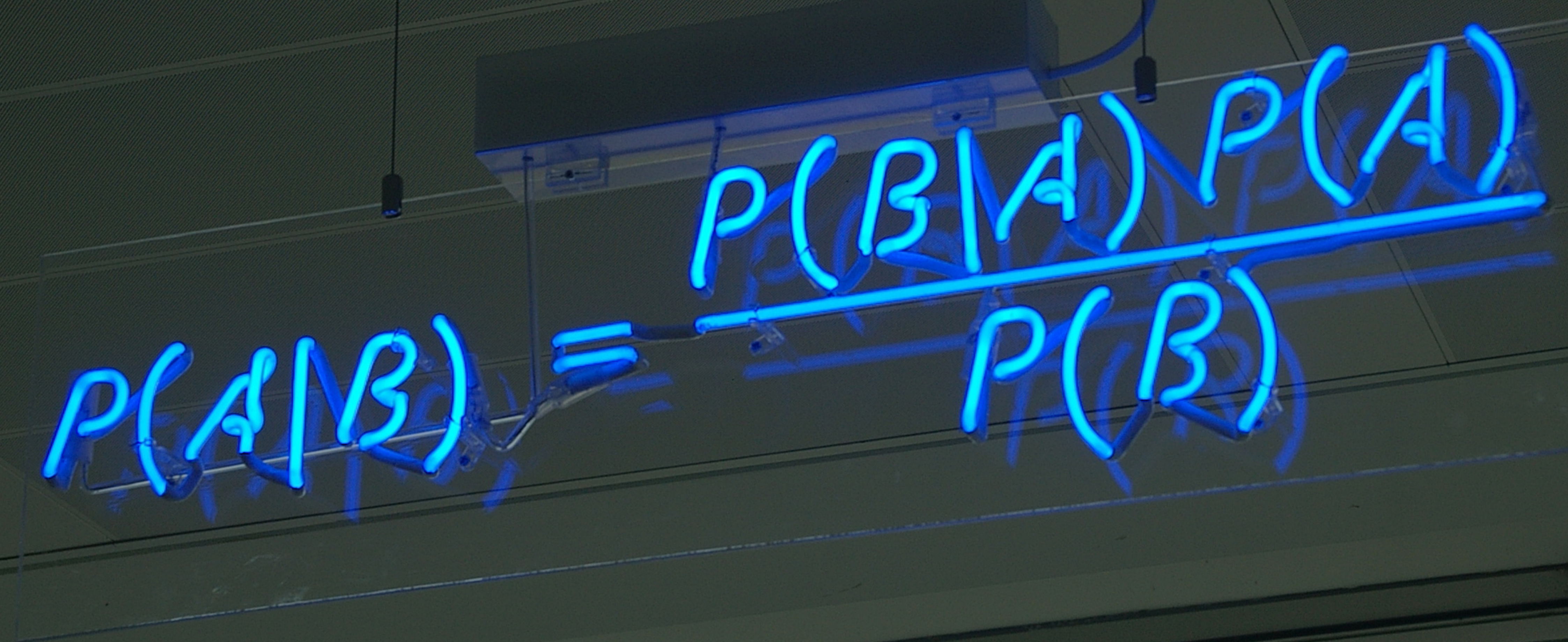


6 / 34 ©2017 Peter Beerli
Twitter @peterbeerli

So many models – so little time

Photo CC Wikimedia Wolfgang Sauber

Population Parameter Inference



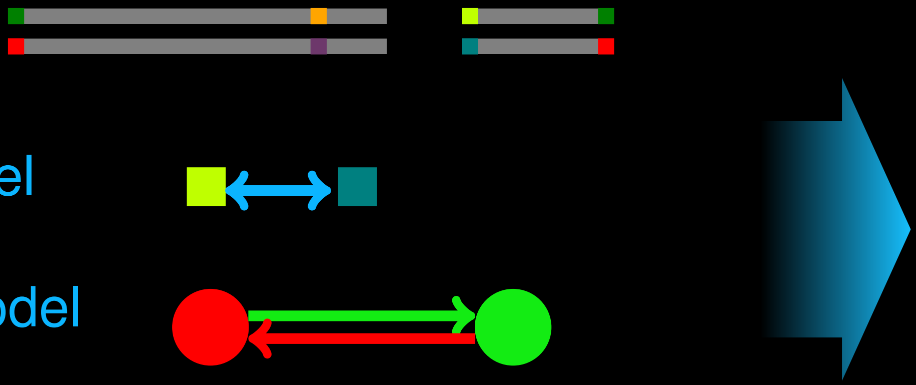
Wikimedia: Neon sign at Autonomy in Cambridge UK

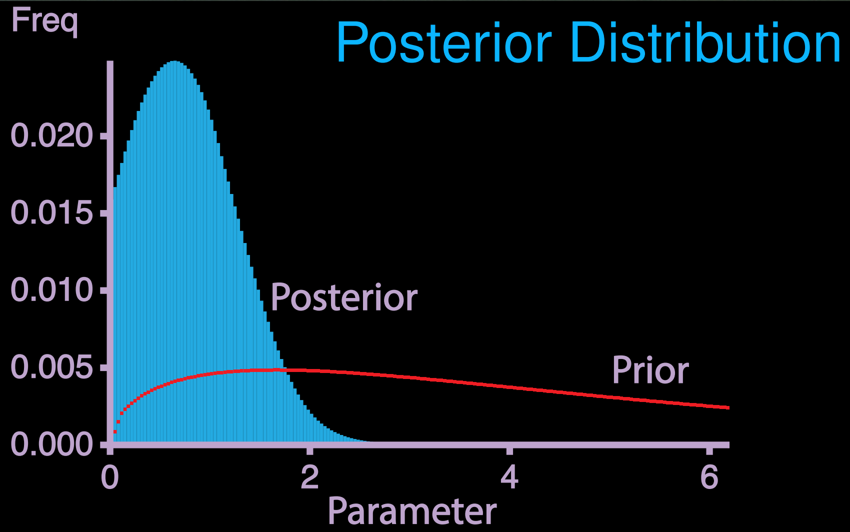
Population Parameter Inference

Genetic Data

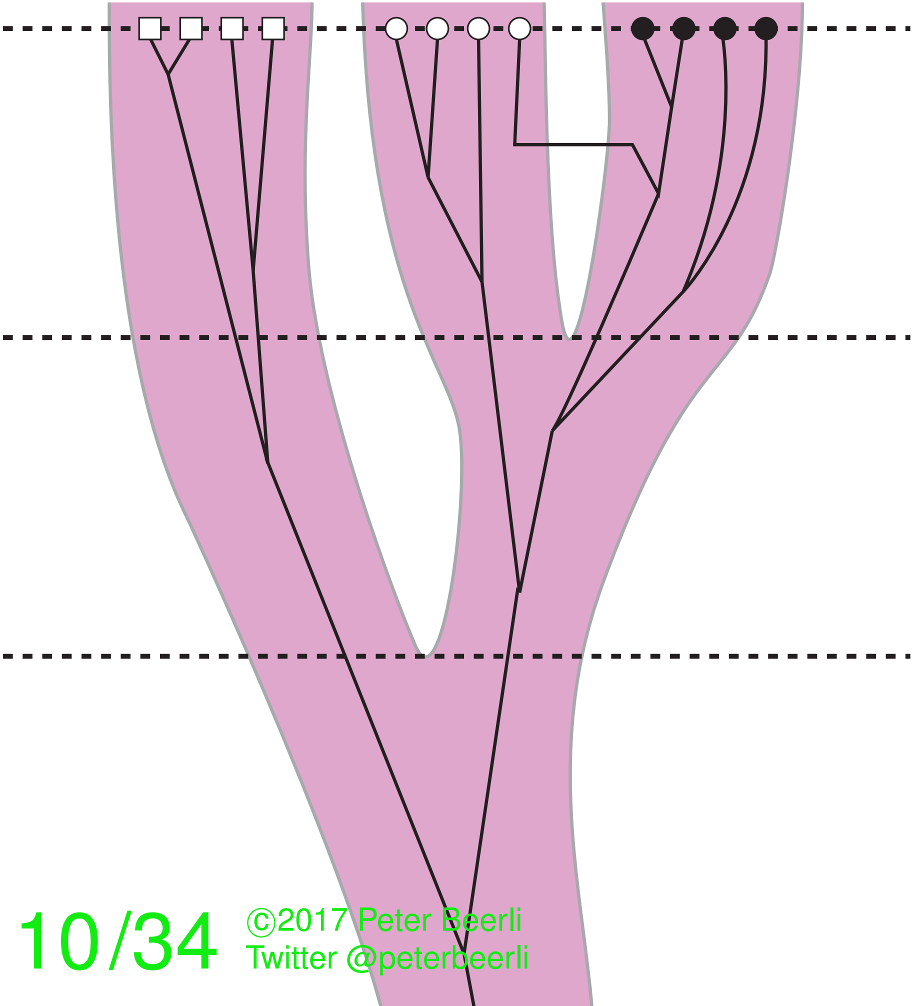
Mutation model

Population model





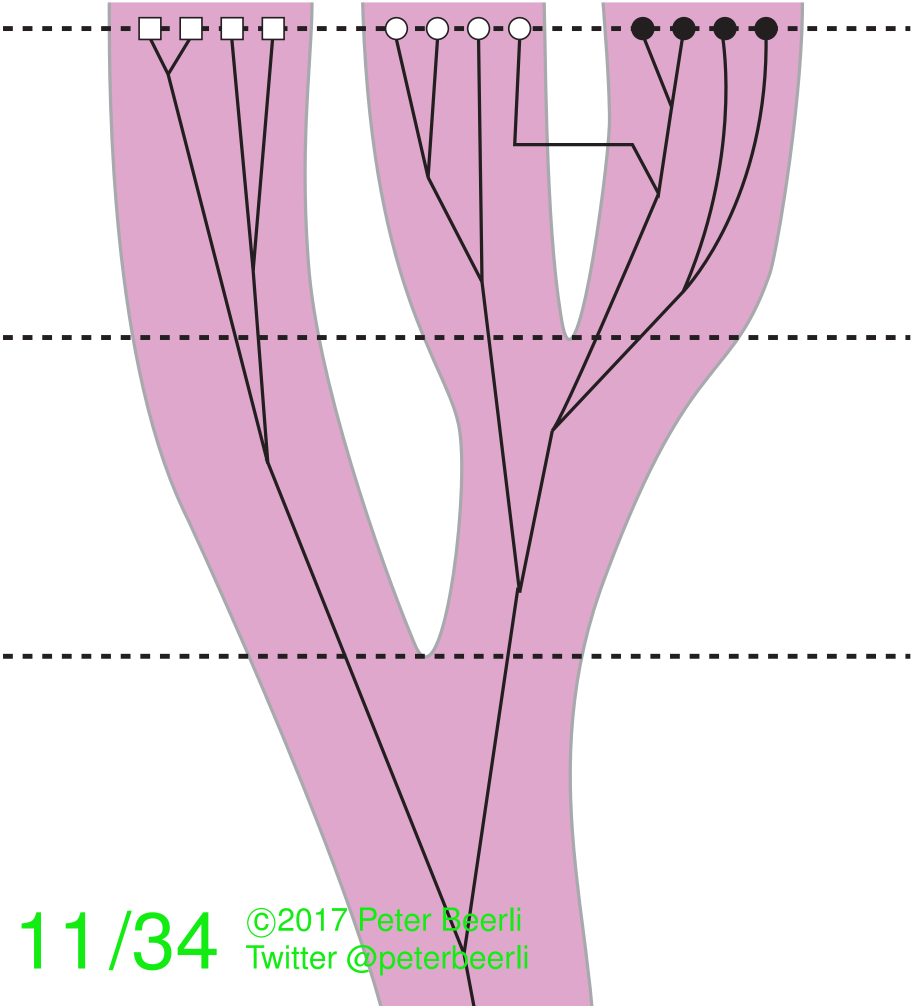
Population model



The relationship among individuals can be expressed, looking backward in time, by a waiting process where random lineages

- coalesce
- migrate between populations
- split off an ancestral population

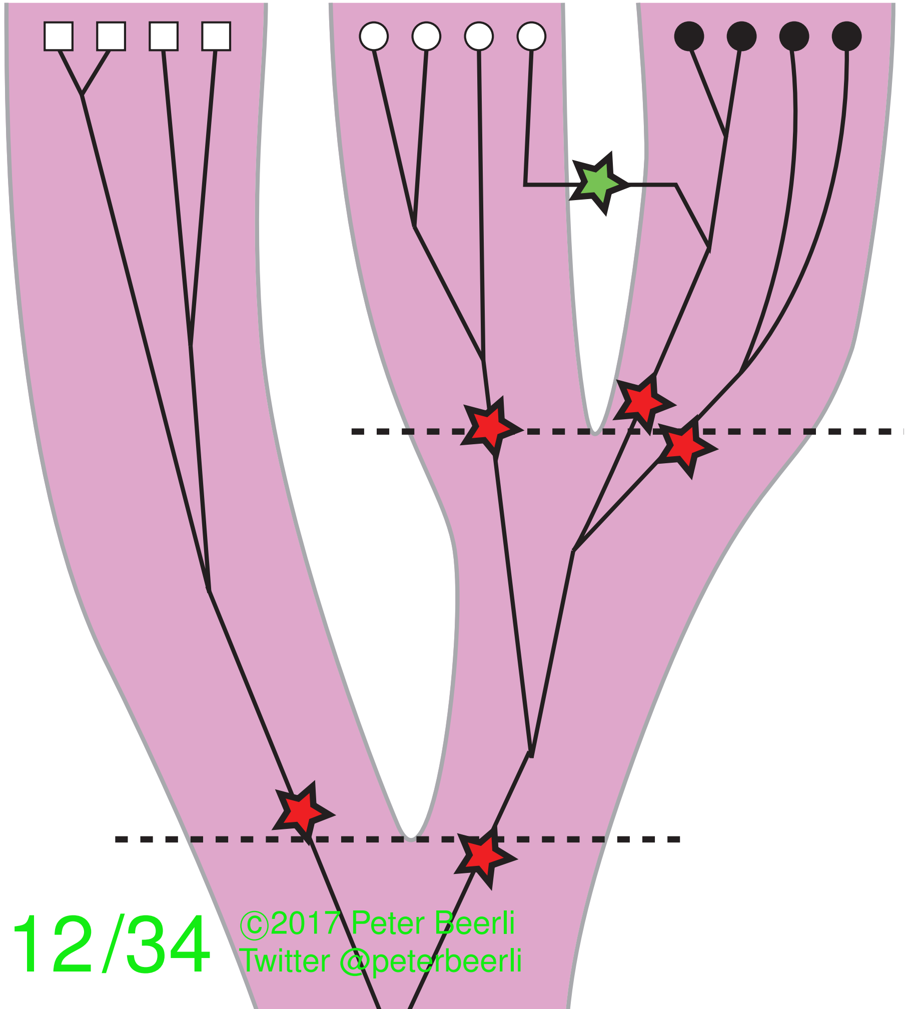
Current common population splitting model



The relationship among individuals can be expressed, looking backward in time, by a waiting process where random lineages

- coalesce
- migrate between populations
- split off an ancestral population

Alternative Population splitting model



The relationship among individuals can be expressed, looking backward in time, by a waiting process where random lineages

- coalesce
- migrate between populations
- split off an ancestral population

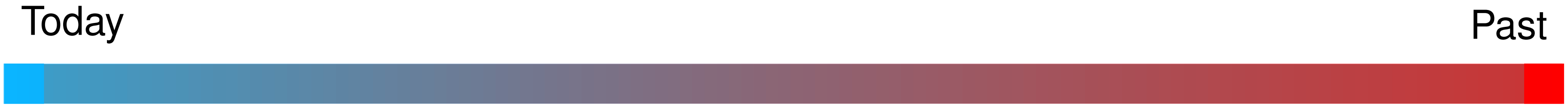
Each of these processes can be expressed as a waiting time process with rate λ for N populations and $k_j$lineages in population j.

Population splitting

if we consider only a single individual that is today in population **A**. We also know that its ancestor was a member of population **B**, then it will be only a matter of time to change the population label, but when?

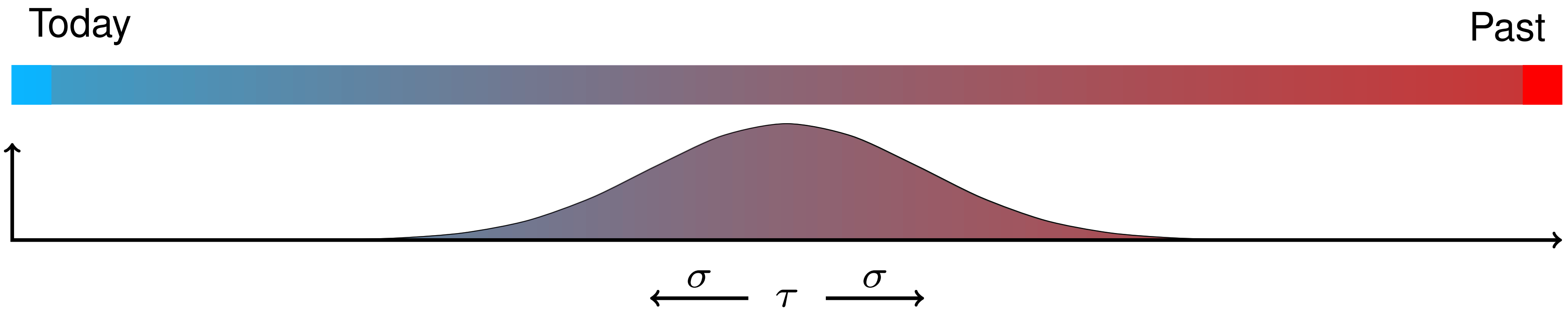
Population splitting

if we consider only a single individual that is today on population **A**. We also know that its ancestor was a member of population **B**, then it will be only a matter of time to change the population label, but when?



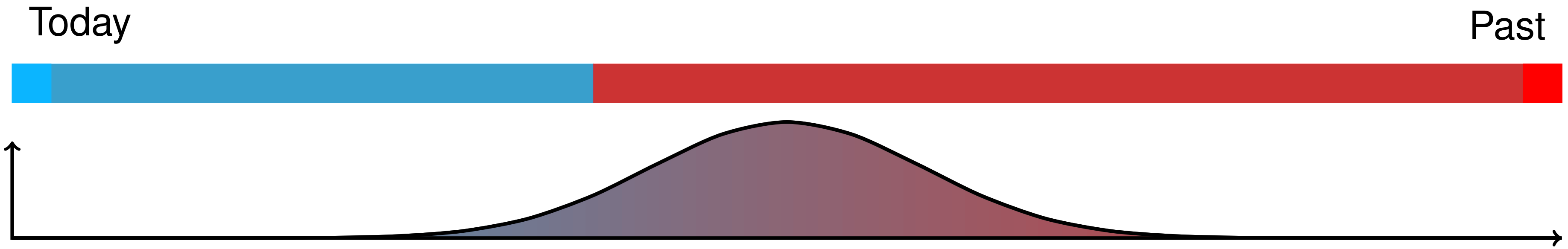
Population splitting

We could assume that the label change is drawn from a Normal distribution with known mean τ and standard deviation σ.



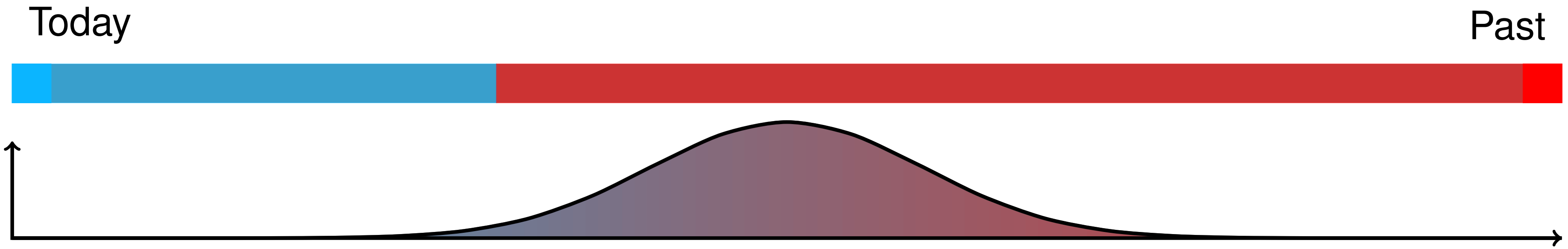
Population splitting

We could assume that the label change is drawn from a Normal distribution with known mean τ and standard deviation σ. For example like this:



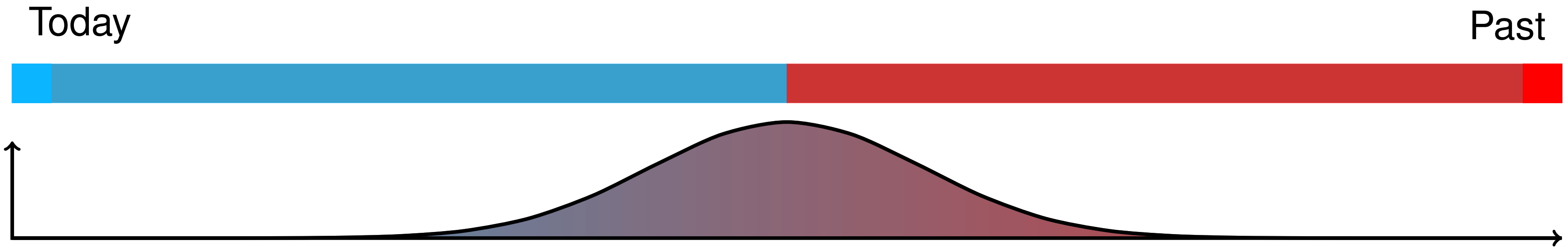
Population splitting

We could assume that the label change is drawn from a Normal distribution with known mean τ and standard deviation σ. For example like this:



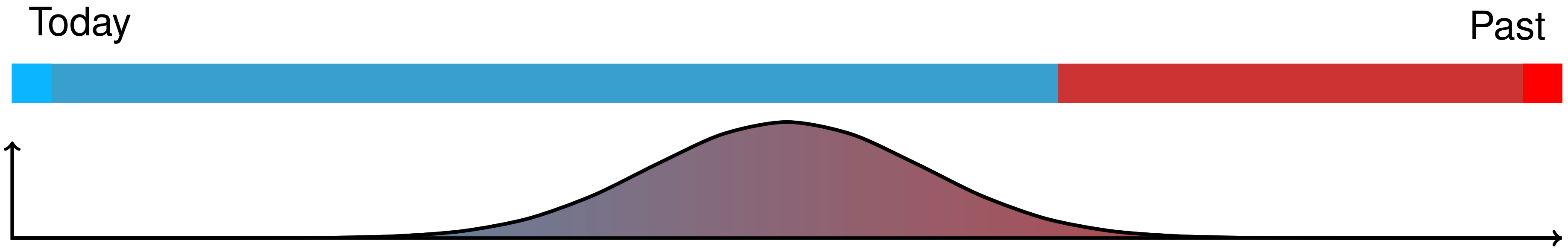
Population splitting

We could assume that the label change is drawn from a Normal distribution with known mean τ and standard deviation σ. For example like this:



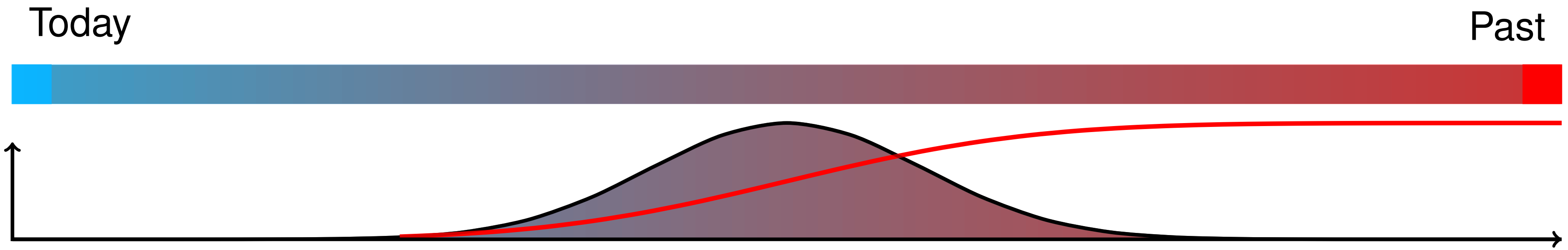
Population splitting

We could assume that the label change is drawn from a Normal distribution with known mean τ and standard deviation σ. For example like this:



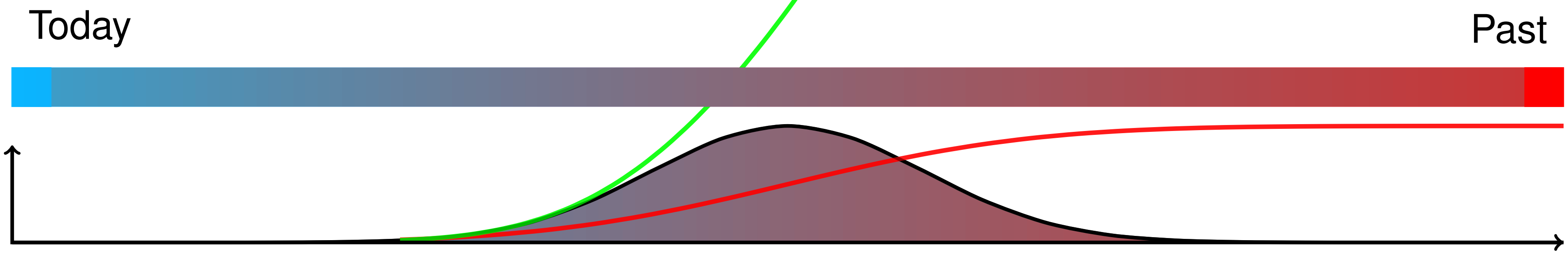
Population splitting

Looking at the cumulative density function of the Normal(τ, σ), then once $t > \tau$ being in the **B** state is more common than being in the **A** state.



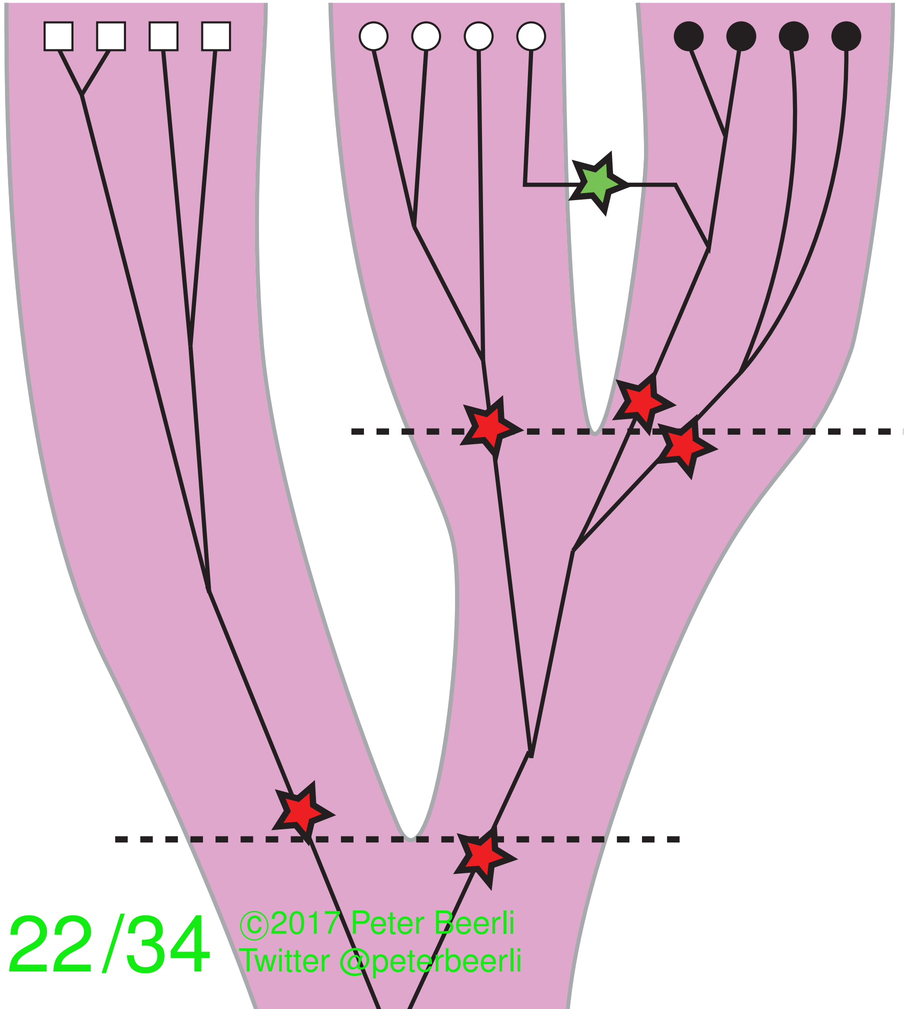
Population splitting

Looking backwards in time we could think about the risk of **A** turning into **B** which becomes larger and larger the further back in time the lineage goes.



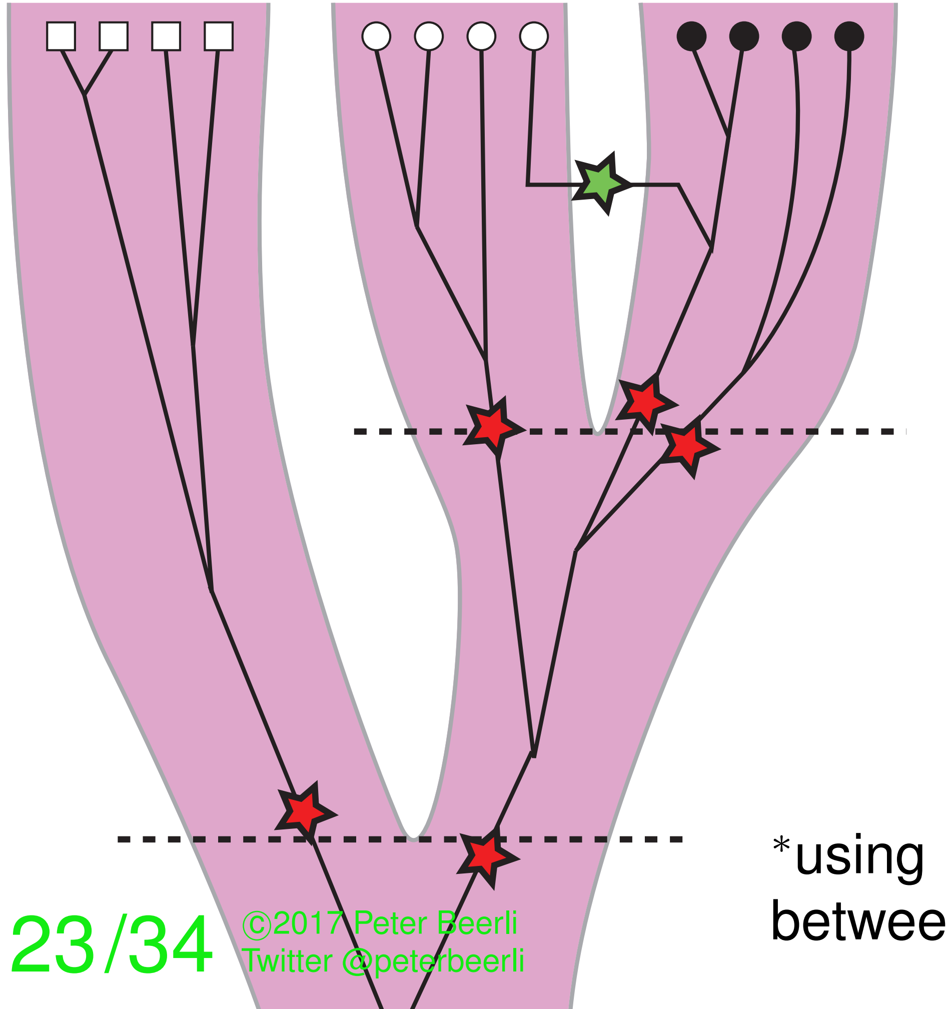
In the coalescence framework we are well accustomed to that thinking: we use the risk of a coalescent or the risk of a migration event. This risk can be expressed using the hazard function (or failure rate). Here we use the hazard function of the Normal distribution.

Population genetics



One lineage is easy, but what about the genealogy? Each lineage is at risk of being in the ancestral population, thus we need to consider coalescences, migration events, and population label changing events. This results in genealogies that are realizations of migration and population splitting events.

Population genetics



Each of these processes can be expressed as a waiting time process with rate λ for N populations and k_j lineages in population j:

$$\lambda_{\text{two lineages coalesce}} = \sum_{j=1}^{N} \frac{k_j(k_j-1)}{4N}$$

$$\lambda_{\text{lineages migrate}} = \sum_{j=1}^{N} \sum_{i=1, i\neq j}^{N} k_j m_{ij}$$

$$\lambda_{\text{a lineage splits off}^*} = \frac{\sqrt{\frac{2}{\pi}}\, e^{\frac{(t-\mu)^2}{2b^2}}}{b\left(1-\text{erf}\left(\frac{t-\mu}{\sqrt{2}b}\right)\right)}$$

*using a Normal distribution to model the splitting time between two populations.

Combining the parts

$$P(\boldsymbol{\Theta}|\mathbf{D_1},\mathbf{D_2},...,\mu) = \frac{P(\boldsymbol{\Theta})P(\mathbf{D_1},\mathbf{D_2},...|\boldsymbol{\Theta})}{P(\mathbf{D_1},\mathbf{D_2},...)} = \frac{P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n\mathrm{Loci}} P(\mathbf{D_i}|\boldsymbol{\Theta},\mu)dG}{\int_\Theta P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n\mathrm{Loci}} P(\mathbf{D_i}|\boldsymbol{\Theta},\mu)dGd\Theta}$$

$$P(G|\boldsymbol{\Theta}) = \prod_{i=1}^{K} \lambda_x \exp(-t_i[\lambda_{\text{coalescence}} + \lambda_{\text{migration}} + \lambda_{\text{splitting}}])$$

$\boldsymbol{\Theta}$ vector of parameters for population size, migration and splitting parameters.

$\mathbf{D_1},\mathbf{D_2},...$ independent genetic sequence data,

μ mutation model,

G nuisance genealogies that we integrate out (we are interested in the parameters not the trees).

x the particular event on the genealogy

K number of total events on the genealogy

Combining the parts

$$P(\boldsymbol{\Theta}|\mathbf{D_1},\mathbf{D_2},...,\mu) = \frac{P(\boldsymbol{\Theta})P(\mathbf{D_1},\mathbf{D_2},...|\boldsymbol{\Theta})}{P(\mathbf{D_1},\mathbf{D_2},...)} = \frac{P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n_{\text{Loci}}} P(\mathbf{D_i}|\boldsymbol{\Theta},\mu)dG}{\int_\Theta P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n_{\text{Loci}}} P(\mathbf{D_i}|\boldsymbol{\Theta},\mu)dGd\Theta}$$

$$P(G|\boldsymbol{\Theta}) = \prod_{i=1}^{K} \lambda_x \exp(-t_i[\lambda_{\text{coalescence}} + \lambda_{\text{migration}} + \lambda_{\text{splitting}}])$$

$\boldsymbol{\Theta}$ vector of parameters for population size, migration and splitting parameters.

$\mathbf{D_1},\mathbf{D_2},...$ independent genetic sequence data,

μ mutation model,

G nuisance genealogies that we integrate out (we are interested in the parameters not the trees). **Really?**

x the particular event on the genealogy

K number of total events on the genealogy

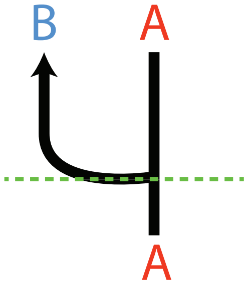
Finally....

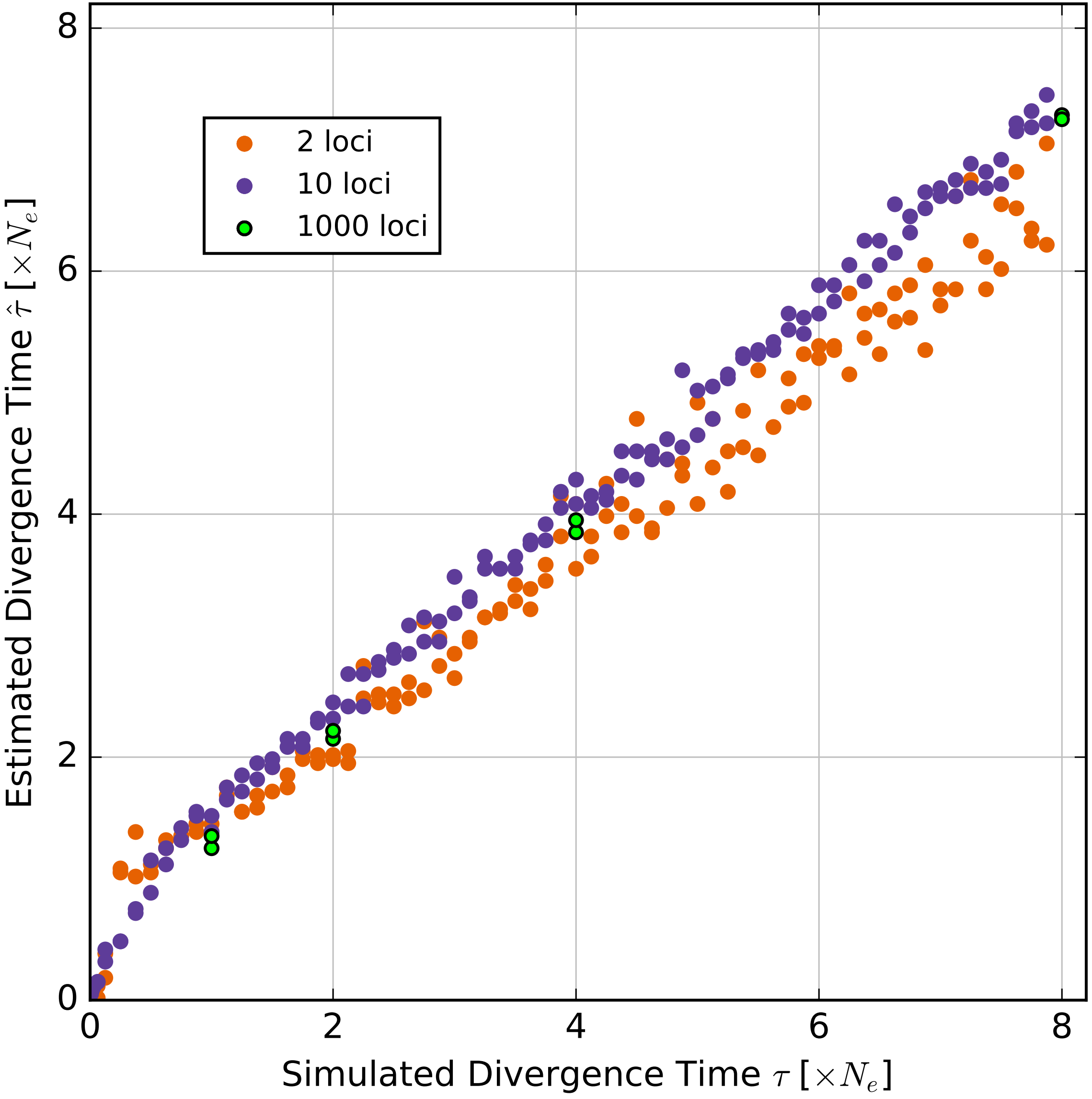
$$p(D|\boldsymbol{\Theta}) = \int_G p(G|\boldsymbol{\Theta})p(D|G)dG$$

The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.

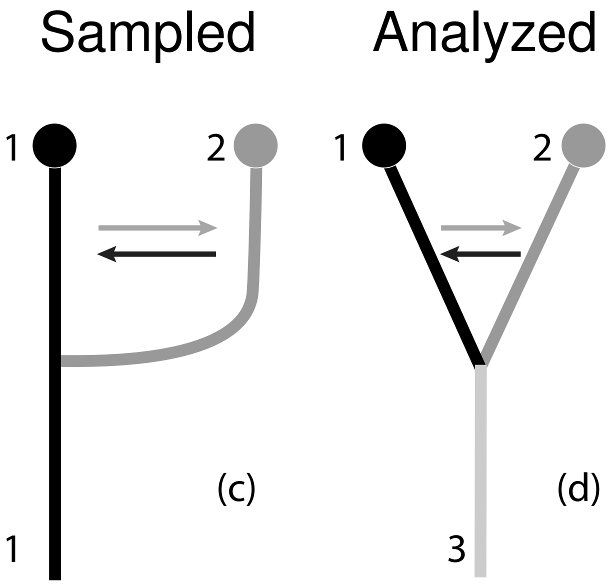
Population splitting

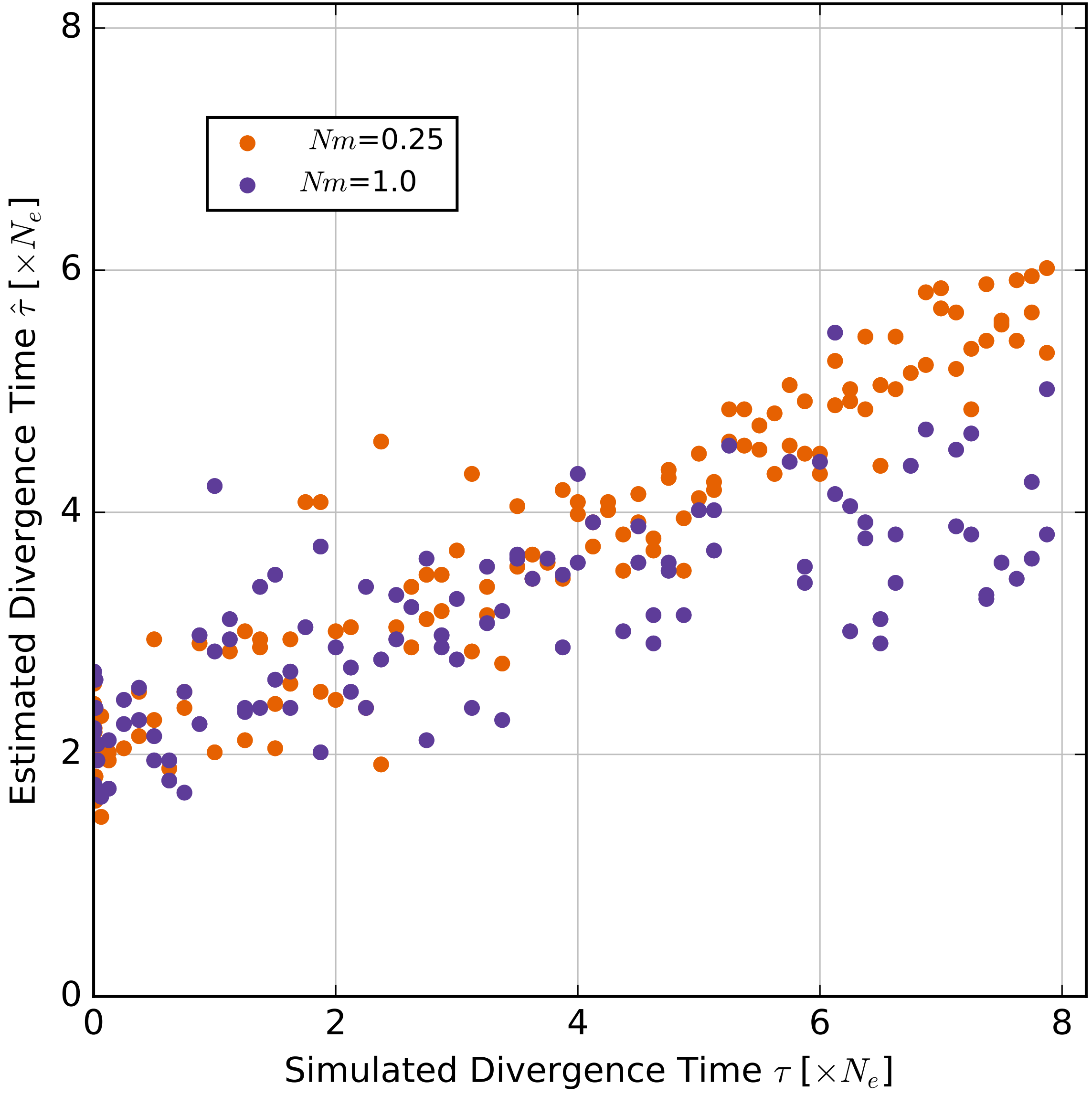
Comparison of estimated versus simulated divergence times for different number of loci





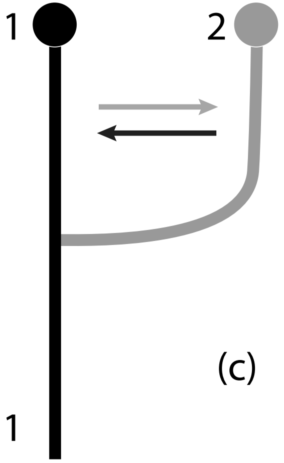
Population splitting

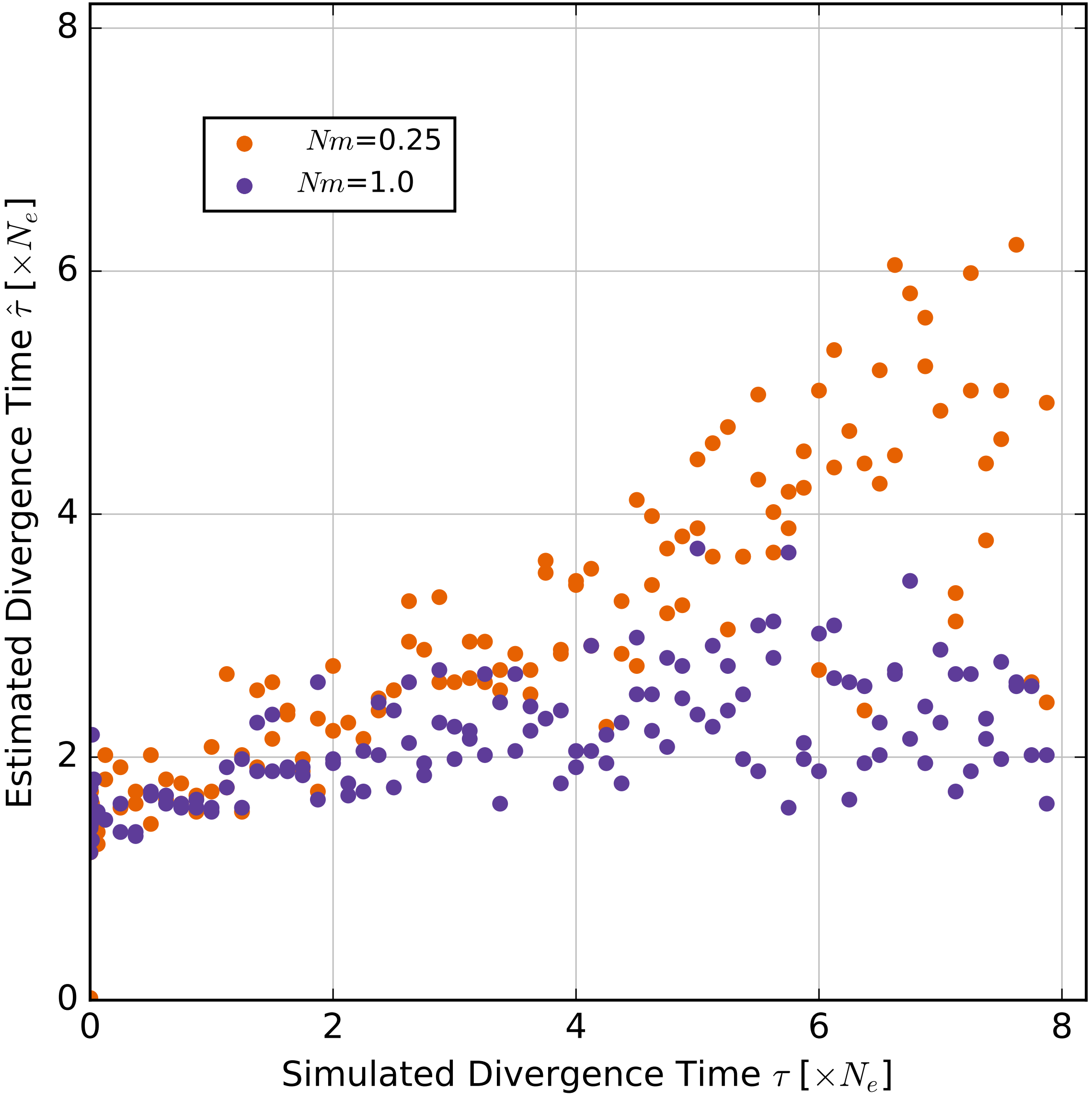




Population splitting

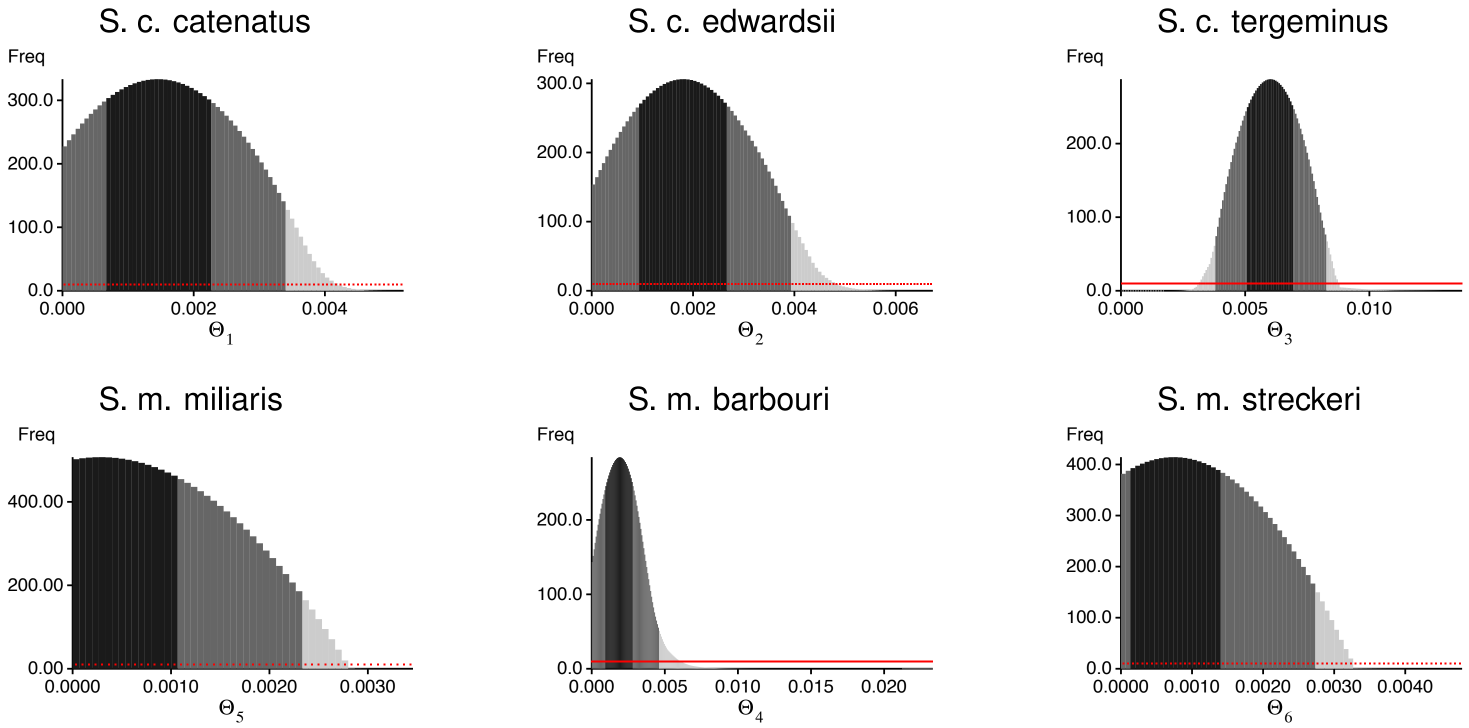
Sampled and Analyzed





Twitter @peterbeerli

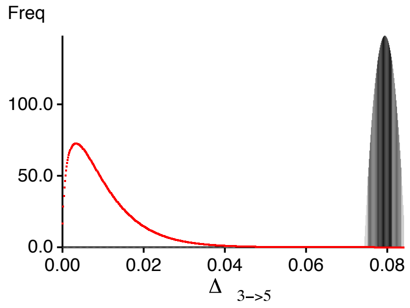
Population sizes



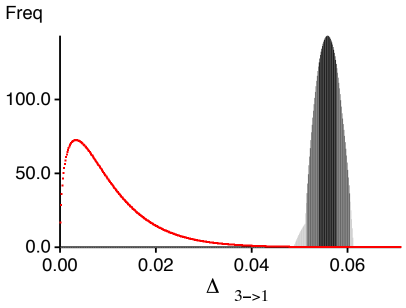
Estimation of splitting dates of 6 subspecies of pygmy rattle snakes using MIGRATE (data from Kubatko et al. 2011)

Population splitting

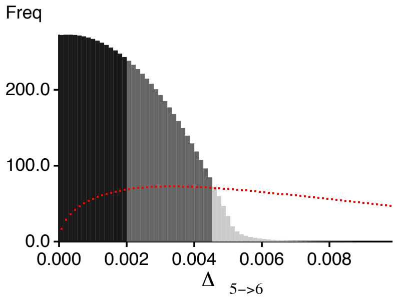
S. catenatus versus S. miliaris



S. c. edwardsii/tergeminus versus S. c. catenatus

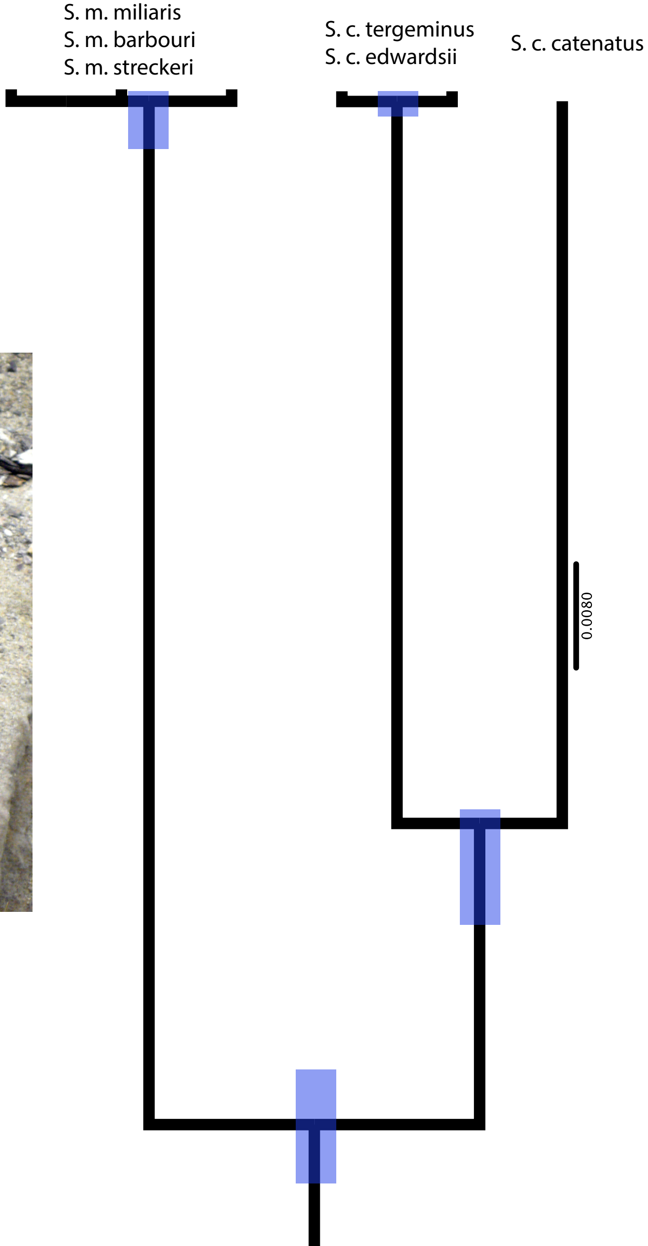


S. m. streckeri/barbouri versus S, miliaris

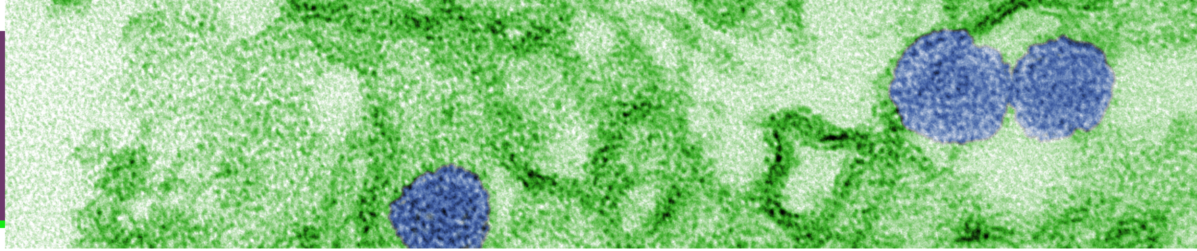


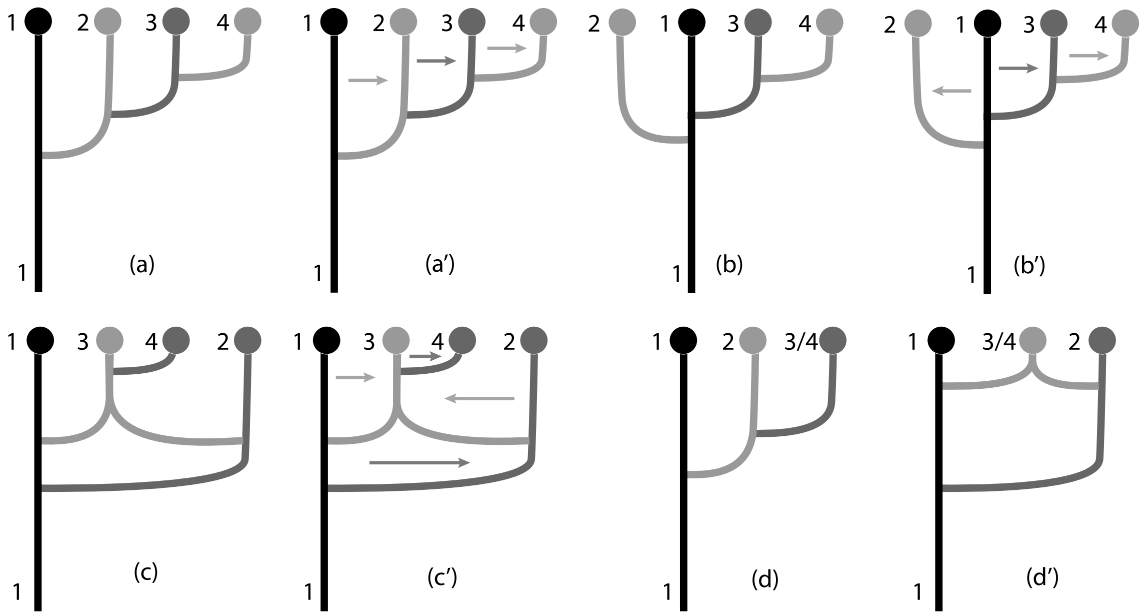
Estimation of splitting dates of 6 subspecies of pygmy rattle snakes using MIGRATE (data from Kubatko et al. 2011)

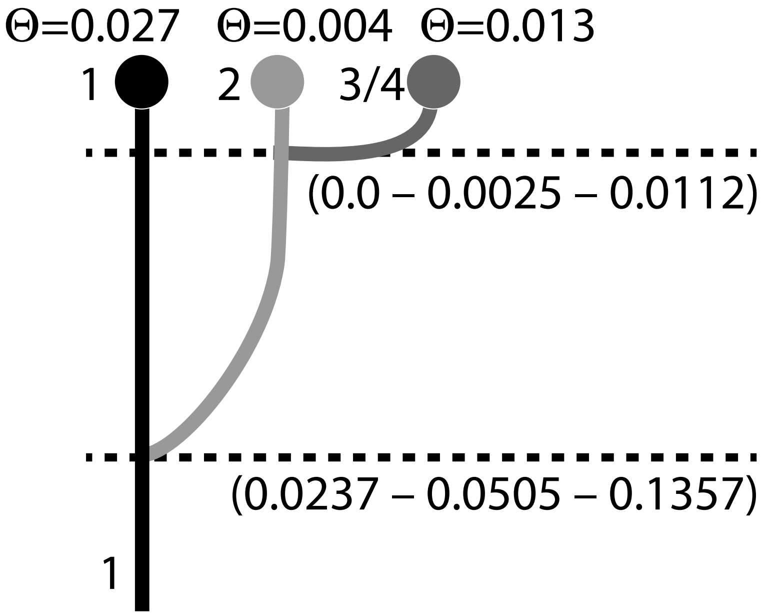
Population splitting



Model comparison







1=Africa, 2=Asia, 3=Brazil, 4='Central' America

Twitter @peterbeerli

Thank you

Lucrezia Bieler

National Science Foundation

Michal Palzcewski, Haleh Ashki, Justin Bricker, Somayeh Mashayekhi, Kyle Shaw

http://popgen.sc.fsu.edu

Credit: ESO/C. Malin