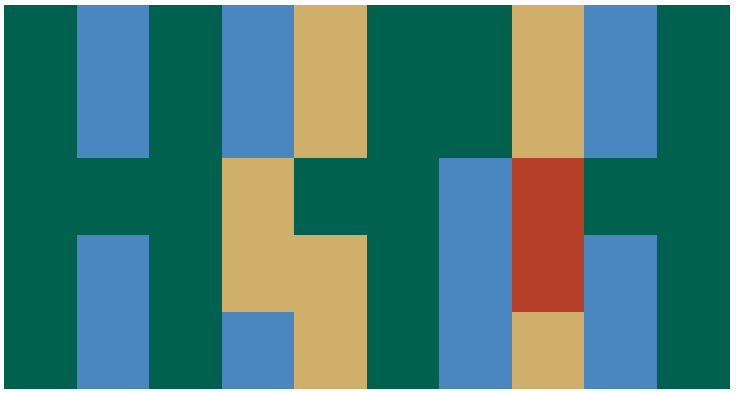
Estimation of demographic history in structured populations using a structured coalescent approach

Chieh-Hsi Wu

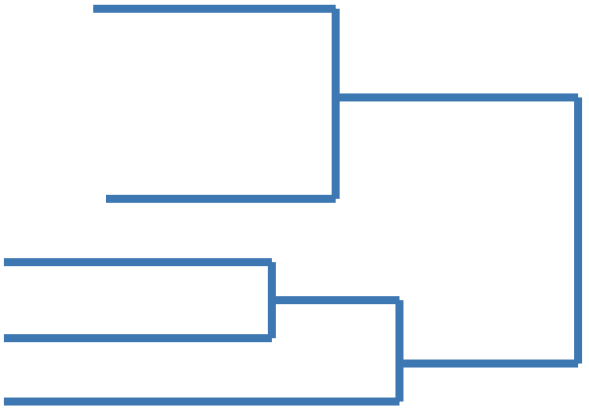
Nuffield Department of Medicine
University of Oxford

Estimating population history

Sequence alignment
of genetic data

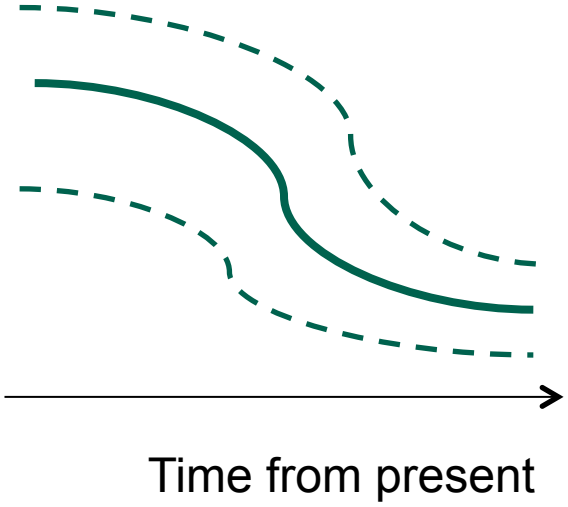


Genealogy



Coalescent

Population history

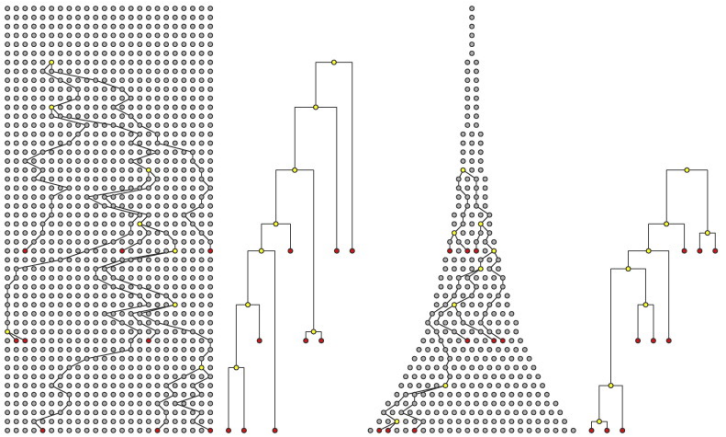


The coalescent

- The coalescent is a model that describes how the coalescent times in a genealogical tree of a random sample of individuals are related to the size of the population from which they have come.
- The Kingman's coalescent assumes an idealised Wright-Fisher model (Kingman, 1982).
- It has been generalized to any deterministically varying function for modelling the population trend (Griffiths and Tavare, 1994).
  - Parametric models
  - Skyline methods: direct estimation of the population trend.

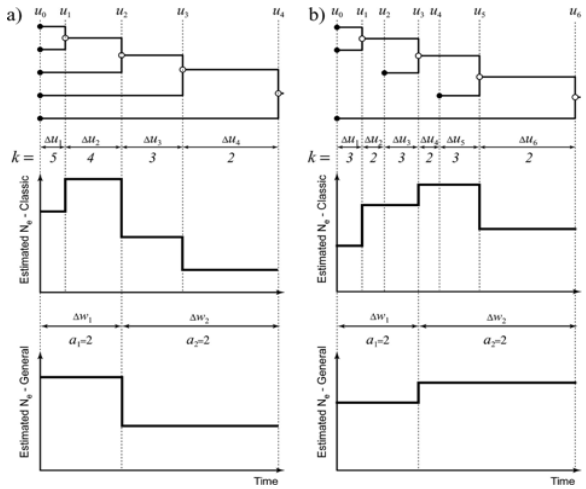
Constant (Wright-Fisher)

Exponential growth



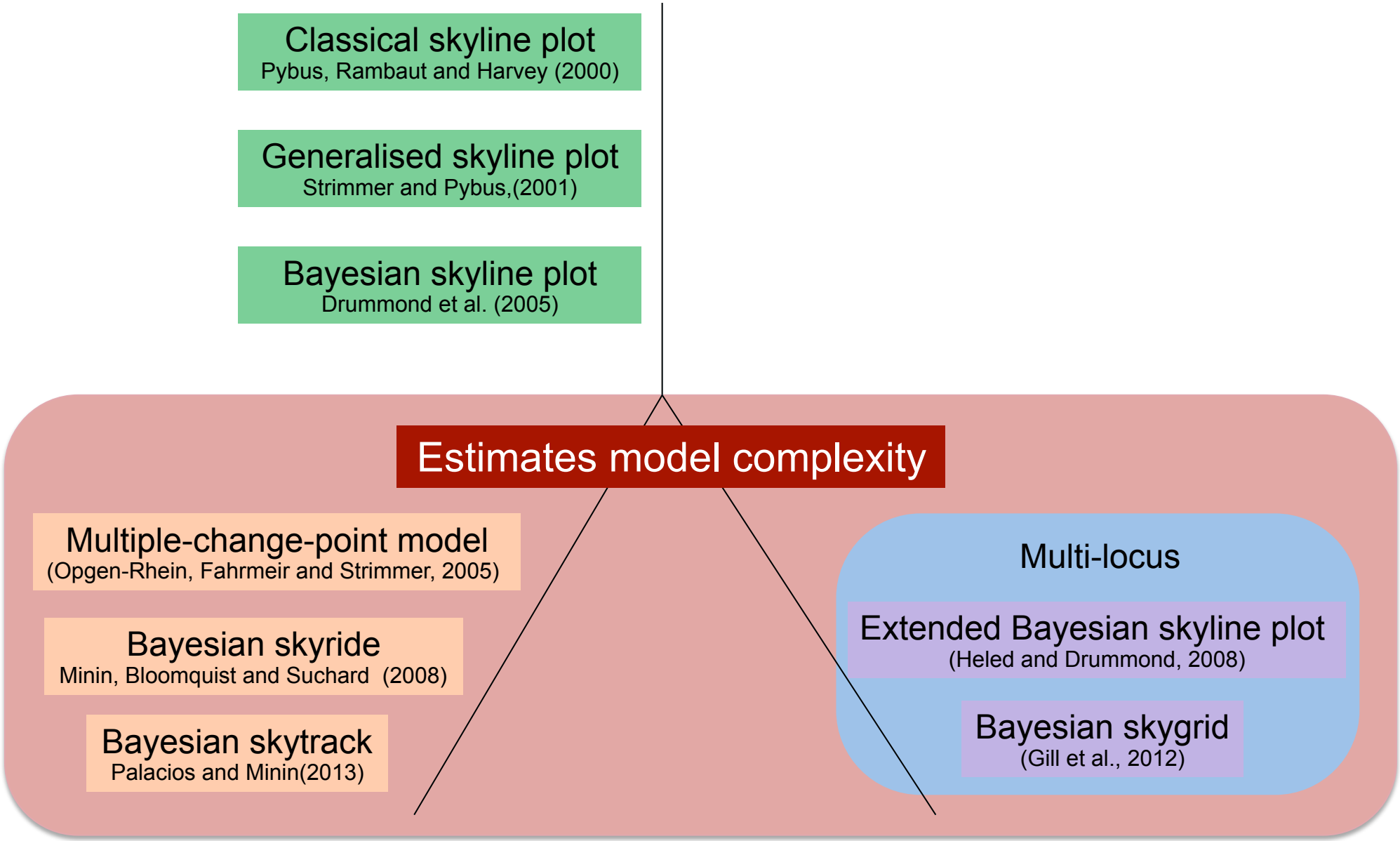
(Figure 1; Kühnert, Wu and Drummond, 2011)

Skyline methods

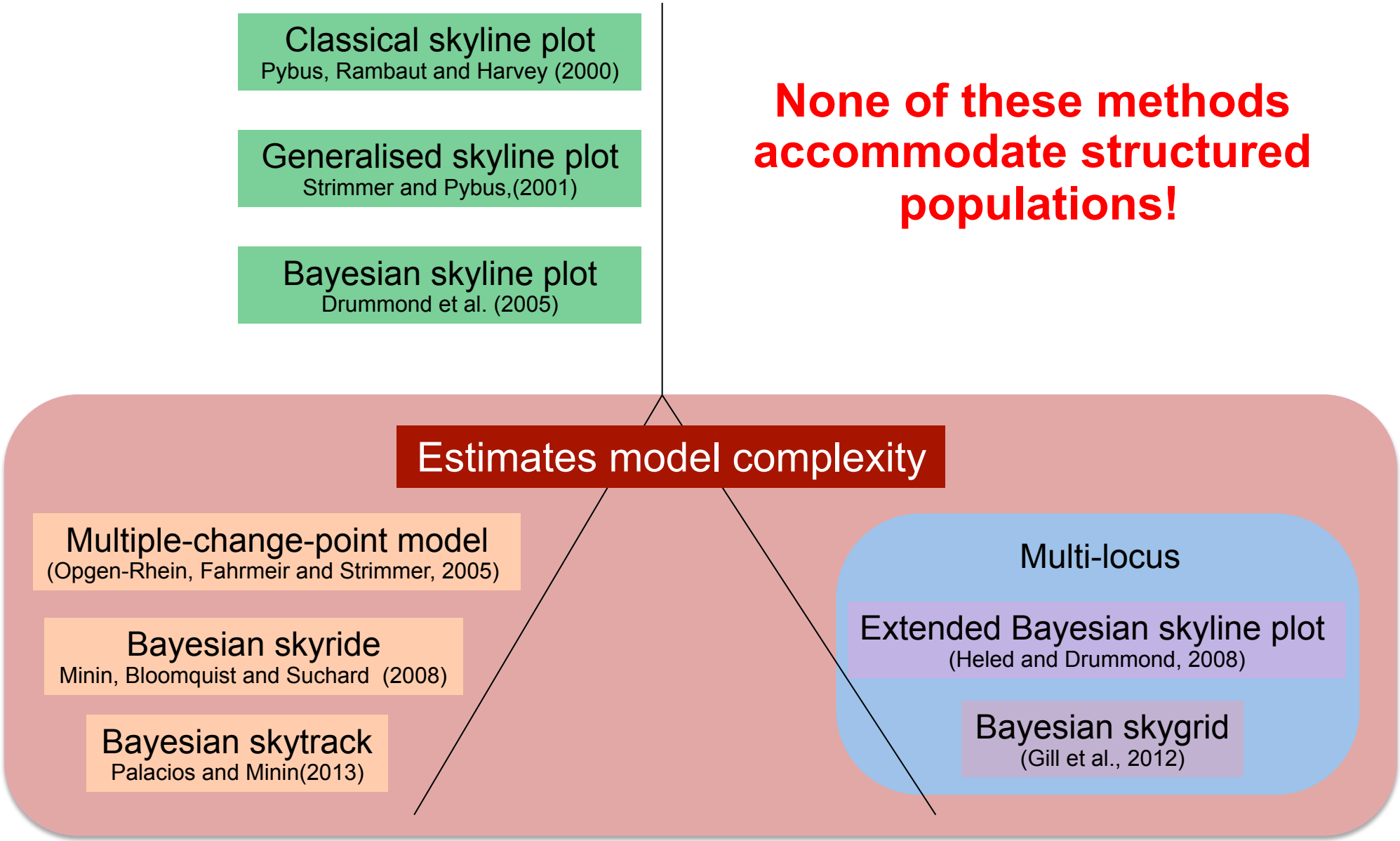


(Drummond et al. 2005)

Skyline methods



Skyline methods



Problems

- If the population is structured, the within deme population sizes produce coalescent patterns that are markedly different to those under the panmixia model
  - Wakeley (1999)
  - Pannell (2003)
  - Beaumont (2004)
  - Nielsen and Beaumont (2009)
  - Peter, Wegmann and Excoffier (2010)
- The estimated demographic trend is sensitive to biased sampling.

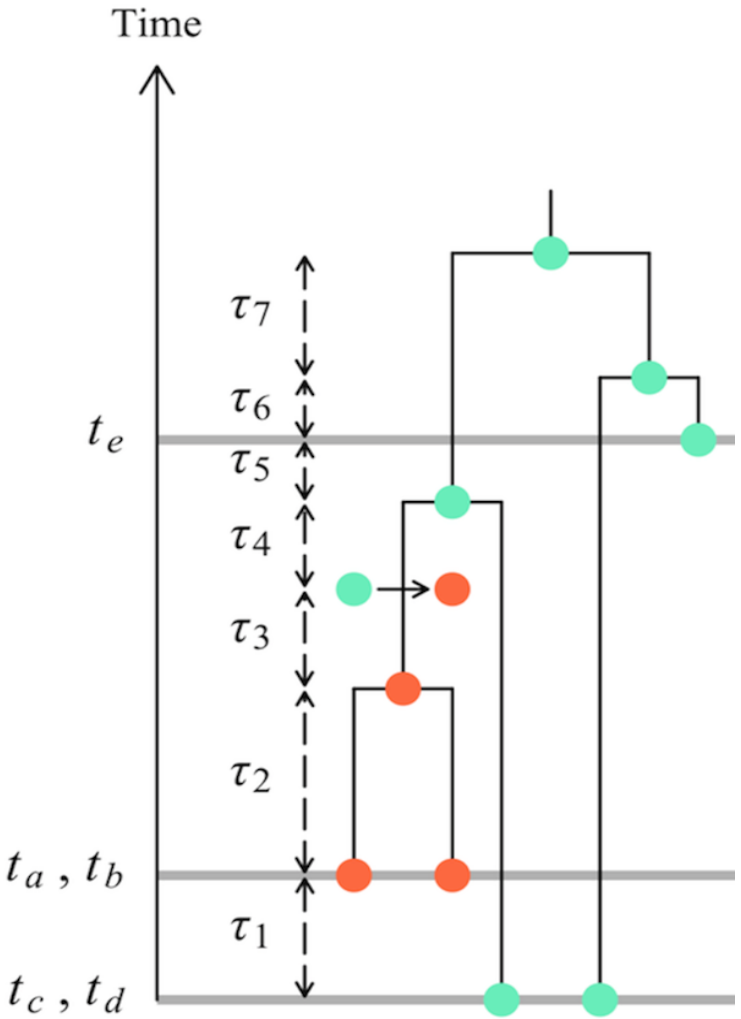
Problems

- Spurious population bottlenecks tended to be detected if the sampling scheme neglects some of the demes.
  - Städler et al. (2009)
  - Chikhi et al. (2010)
  - Heller, Chikhi and Siegismund (2013).
- Even if samples have been collected from all demes, disproportionate collection of recent isolates taken from the same deme also leads to spurious bottleneck effect.
  - Hall, Woolhouse and Rambaut (2015)

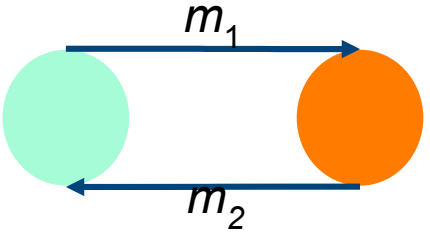
What if … ?

- Use discrete trait analysis (Lemey *et al.* 2008) to take care of the migration process among subpopulations while using one of the Bayesian skyline methods a tree prior.
  - The migration process is modelled by a continuous time Markov chain (CTMC) down the tree.
- The likelihood of this approach is independent of the coalescent process *a priori.*
  - The demes are effectively an additional nucleotide site with a different set of CTMC parameters.
  - The tree estimate is largely determined by the genetic data.
  - The inference of the population trend is not aware of the population structure.

Structured coalescent

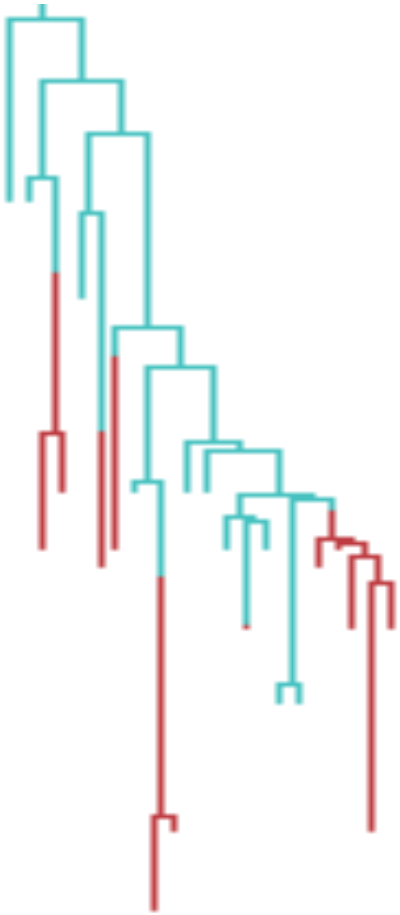


- The structured coalescent extends Kingman's coalescent to handle geographically structured populations with a given underlying migration rate matrix.
- Given the genealogy tree and the migration history, one could estimate the effective population size as well as the migration rates.



Structured coalescent

- The structured coalescent is described with the migration history known.
- The migration history can treated as unknown parameters when using the structured coalescent as a tree prior in a Bayesian inference.
- However co-estimating the tree, migration history and migration rates is computationally challenging.



Structured coalescent

- Recently, more efficient methods have been proposed to overcome the computational hurdle that hampered Bayesian analysis under the structured coalescent.
  - Multitype tree (Vaughan et al., 2014)
  - BASTA (de Maio et al., 2015)
- However these methods assume that subpopulation sizes remain constant through time.
- It would be ideal to have a method that could estimate the demographic trend in a structured coalescent framework.

Approximate structured coalescent

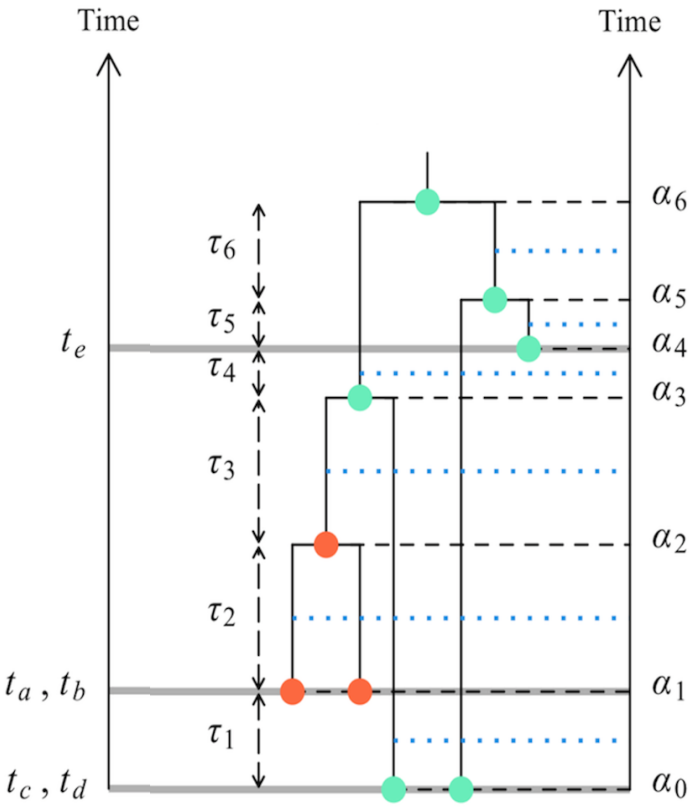
- More about BASTA …
  - BASTA (de Maio et al., 2015) is an approximation to the structured coalescent and computationally more efficient than the exact approach.
  - Instead of evaluating the joint probability density of the genealogy and migration history, BASTA evaluates an approximation of probability density of the genealogy under the structured coalescent integrated overall migration histories.

Approximate structured coalescent

- When integrating out the migration history, the exact probability density for an event interval (bounded by coalescent/sampling event at either ends) under the structured coalescent is give by

$$L_i' = \exp\left[-\int_{\alpha_{i-1}}^{\alpha_i} \sum_{d\in D}\sum_{l\in\Lambda}\sum_{l'\in\Lambda, l\neq l'} P(d_l = d, d_{l'} = d \mid t)\frac{1}{\theta_d}dt\right] E_i',$$

$$E_i = \begin{cases} 1 & \text{event is sampling} \\ \sum_{d\in D} P_{l,\alpha_i,d} P_{l',\alpha_i,d} \dfrac{1}{\theta_d} & \text{event is coalescence} \end{cases}$$



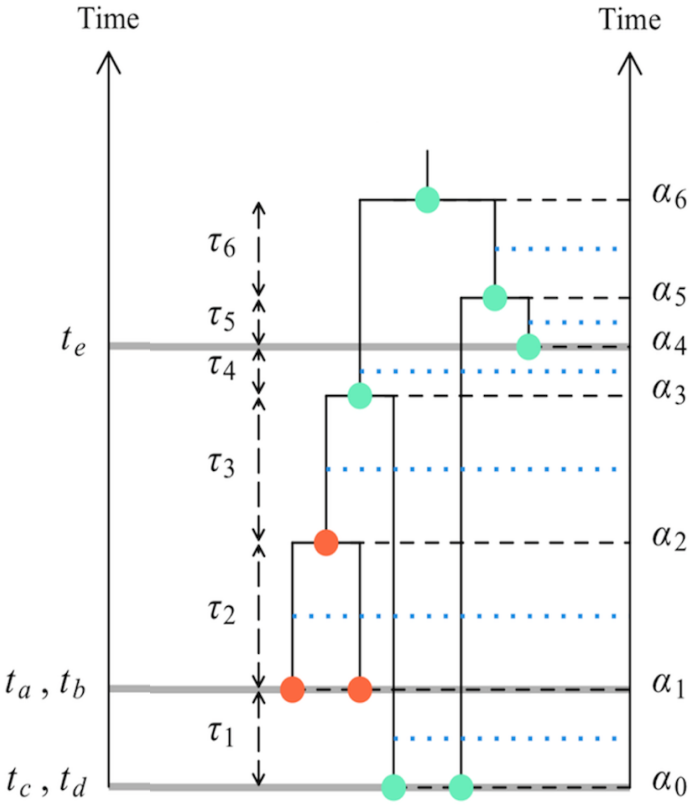
Approximate structured coalescent

Approximations:

- $P(d_l = d, d_{l'} = d|t) = P(d_l = d|t)P(d_{l'} = d|t)$
  - d_l = the deme lineage l belongs to
- $P(d_l = d|t)$ is only evaluated at points where there is a sampling or coalescent event.

$$\tilde{L}_{i1} = \exp\left[-\frac{\tau_i}{2}\sum_{d\in D}\sum_{l\in\Lambda}\sum_{l'\in\Lambda, l'\neq l} P_{l,\alpha_{i-1},d}P_{l',\alpha_{i-1},d}\frac{1}{\theta_d}\right]$$

$$\tilde{L}_{i2} = \exp\left[-\frac{\tau_i}{2}\sum_{d\in D}\sum_{l\in\Lambda}\sum_{l'\in\Lambda, l'\neq l} P_{l,\alpha_i,d}P_{l',\alpha_i,d}\frac{1}{\theta_d}\right]E_i'$$



New method

- We extend a BASTA to allow temporal variation of the effective population size in each deme.
- Definition: The relative size of a subpopulation *d* at coalescent interval *k* is the proportion

$$\frac{\theta_{d,k}}{\theta_{1,k} + \theta_{2,k} + \ldots + \theta_{|D|,k}}$$

where

$\theta_{d,k}$ = the effective population size of the subpopulation *d* at coalescent interval *k.*

D = a set of subpopulation

|D | = the number subpopulations

- Using that definition, the new methods can be classified into two categories
  - **Time-constant relative subpopulation size model**
  - **Time-heterogeneous relative subpopulation size models**

Relative subpopulation sizes

Time-constant relative subpopulation sizes

- If we constrain the relative subpopulation size to be constant across time, in other words

$$\theta_{d,k} = \theta_{\kappa} p_{d.}$$

where

θ_{κ} = the total population size during event interval *k*

p_d = the relative population size of subpopulation *d* for all coalescent intervals.

  - The same trend is shared across all subpopulations

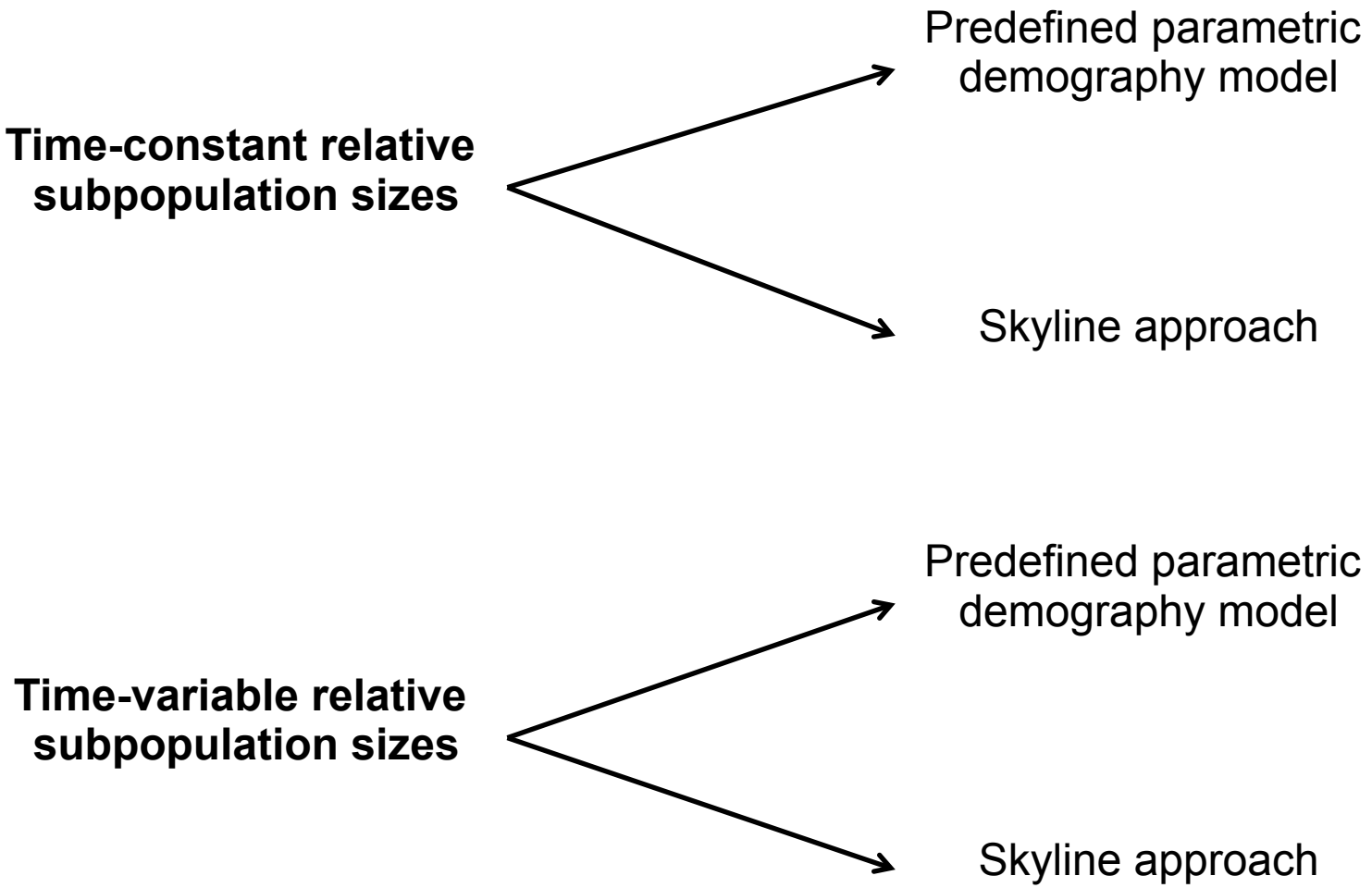
Time-variable relative subpopulation sizes

- If allow the relative subpopulation size to vary through time, the

$$\frac{\theta_{d,k}}{\theta_{1,k} + \theta_{2,k} + \ldots + \theta_{|D|,k}} \neq \frac{\theta_{d,k'}}{\theta_{1,k'} + \theta_{2,k'} + \ldots + \theta_{|D|,k'}}$$

  - The trend can vary across subpopulations.

Estimating population dynamics from structured populations

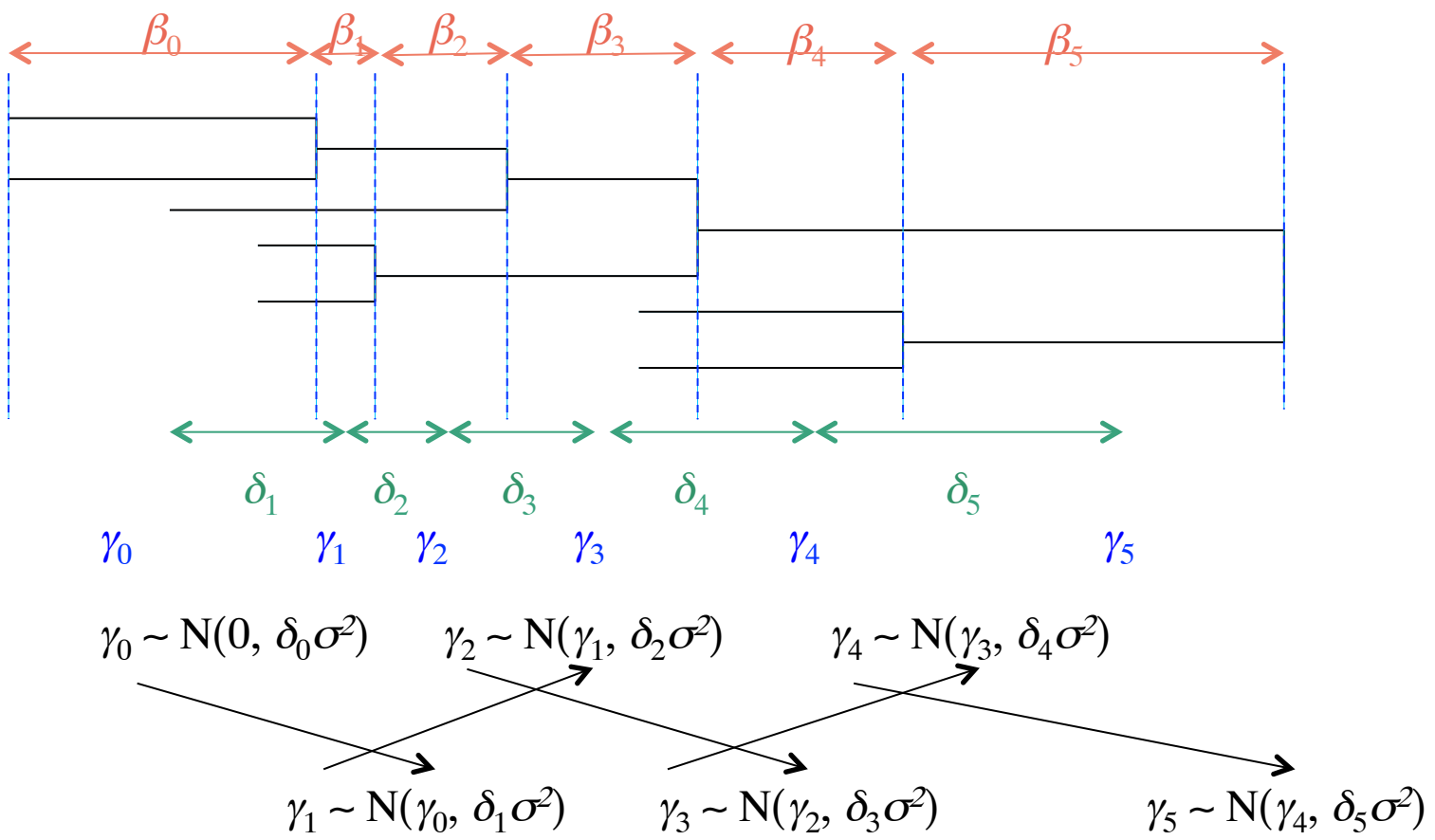


Piecewise parametric demography models

- The effective population size of each subpopulation varies across time in a step-wise fashion according to a pre-definition parametric function.
- The effective population size of a given coalescent interval is evaluated by calculating the value at the midpoint of the interval given a parametric function and its parameters.
- Parametric functions:
    - Constant (= BASTA)
    - Exponential growth
    - Constant-Exponential
    - Logistic
    - Constant-Logistic

Skyline approach

- The population trend of each subpopulation is modelled by a step-function that does not follow a predefined parametric function.
  - Use Gaussian Markov random field (GMRF) prior

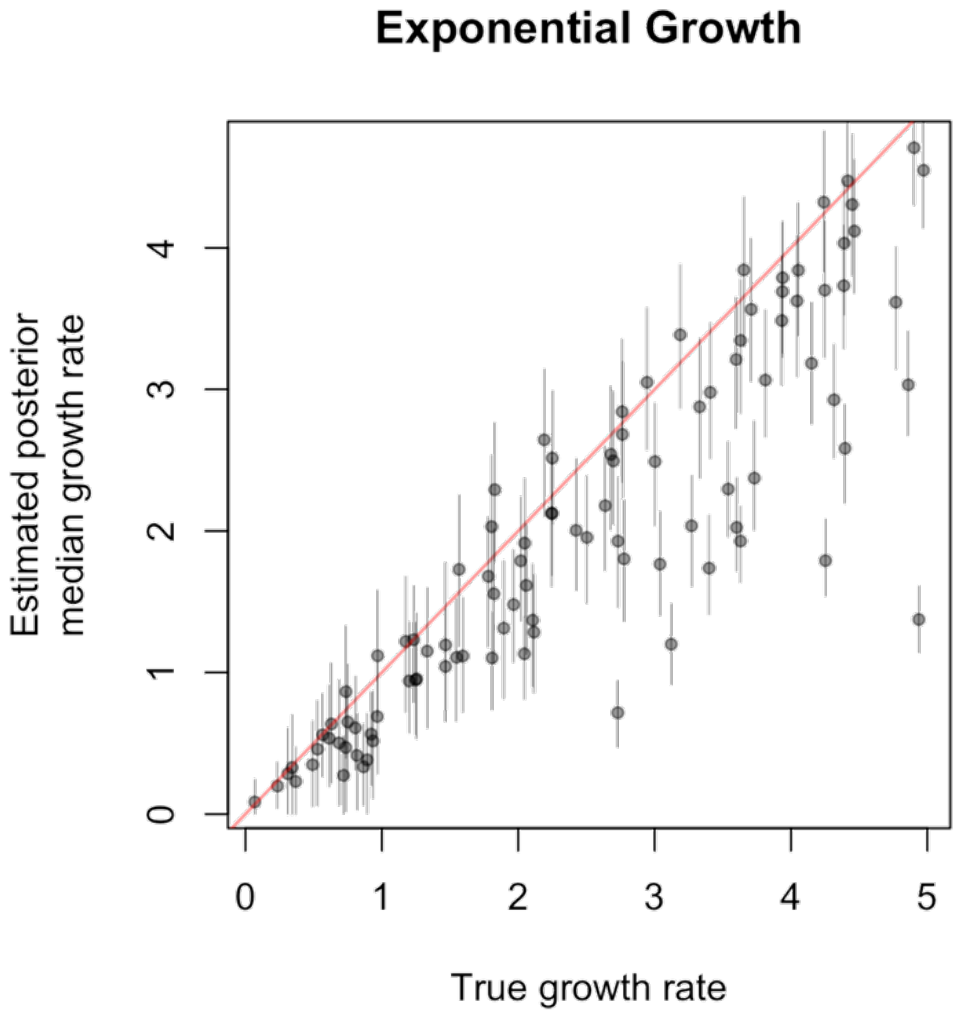


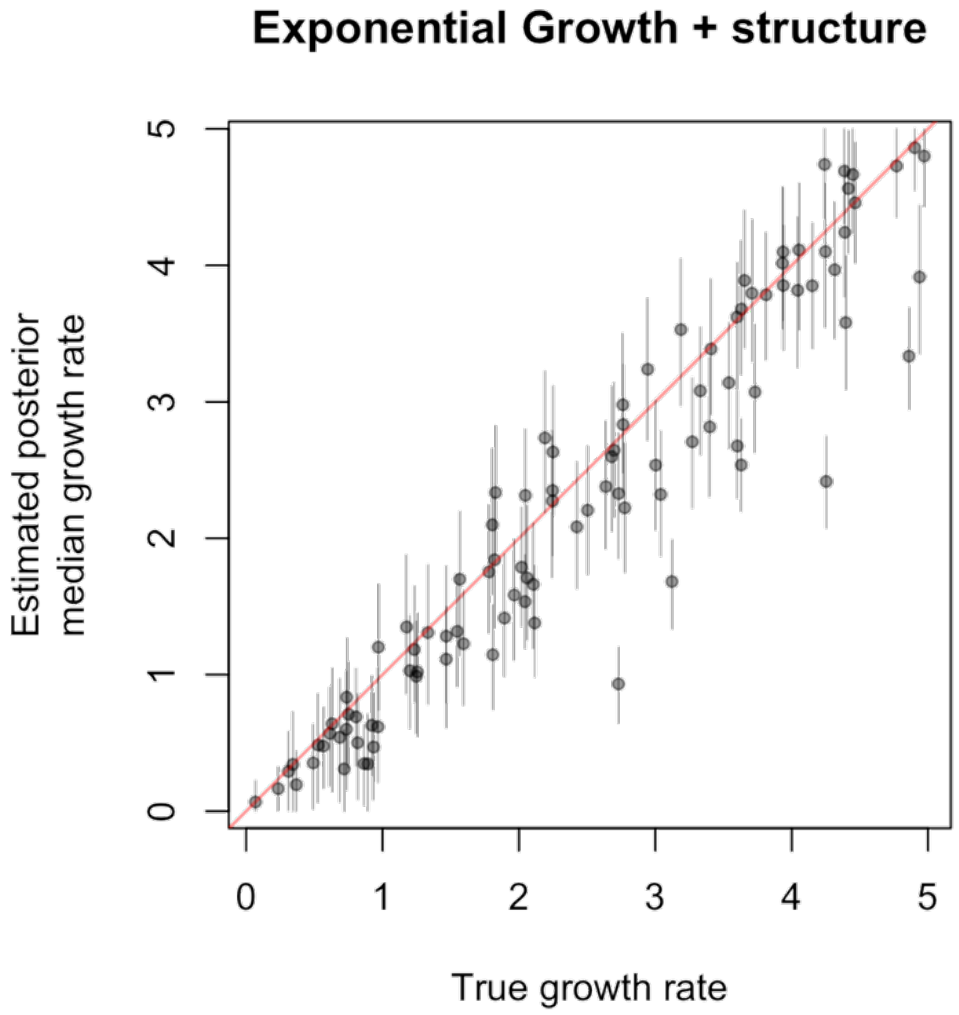
Simulations

- Simulated 100 trees under the structured coalescent with two subpopulations and an exponential growth model.
- For each replicate, first randomly generate
  - Exponential growth rate, r_g ~ Uniform(0, 5)
  - Current (total) effective population size,

    N_0 ~ Gamma(shape = 2, rate = 2)
  - Symmetric migration rate, m ~ Gamma(shape = 10, rate = 10)
- For all replications the relative population sizes for both subpopulations are set to 0.5.
- The sampling ratio of the subpopulations is 8:2.
- Simulate the trees according to the randomly generated parameters above.

Simulations

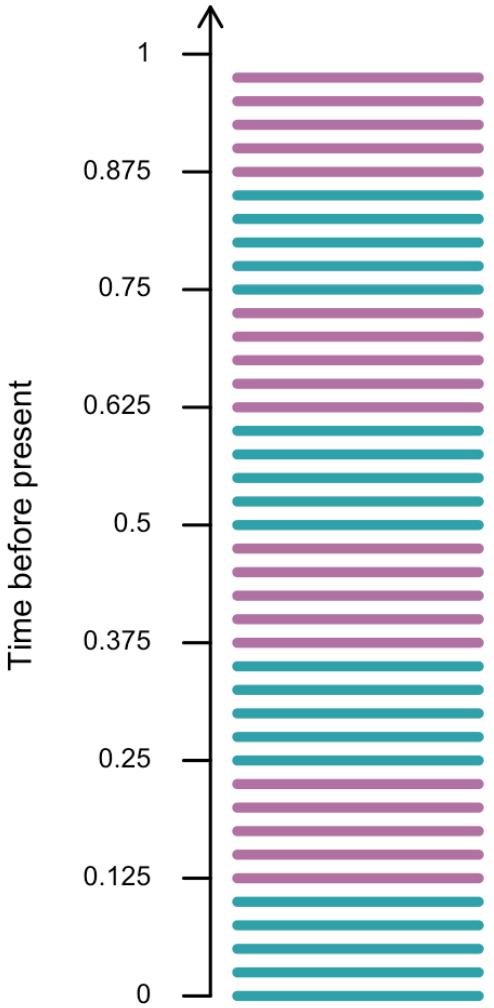




Simulations

Simulate 100 trees under the described demographic history, sampling scheme and each of the migration rate values

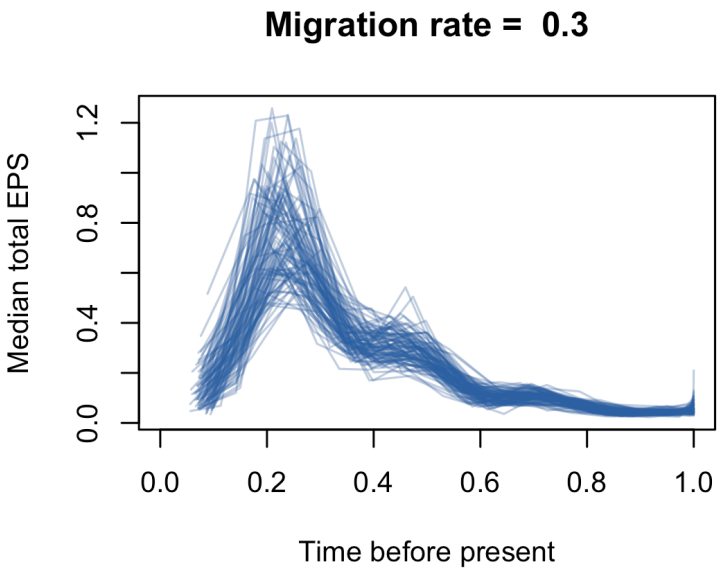
- **Demographic history**
  - Two exponentially growing subpopulations
  - Growth rates: 1 and 5
- **Sampling scheme**
  - Samples from each subpopulation are selected in a clustered manner through time.
  - The clustered samples alternative between the two subpopulations.
  - 100 samples from each subpopulations
- **Migrations rates**
  - Symmetric
  - 0.3, 0.5, 0.7 and 1.0

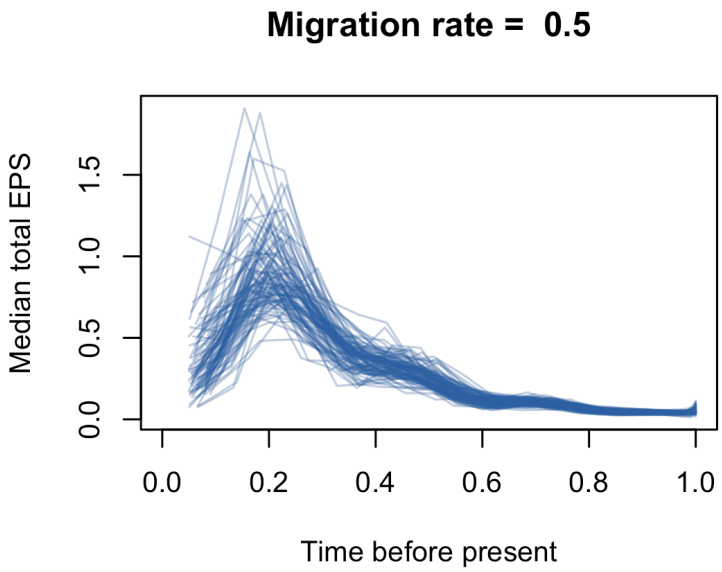


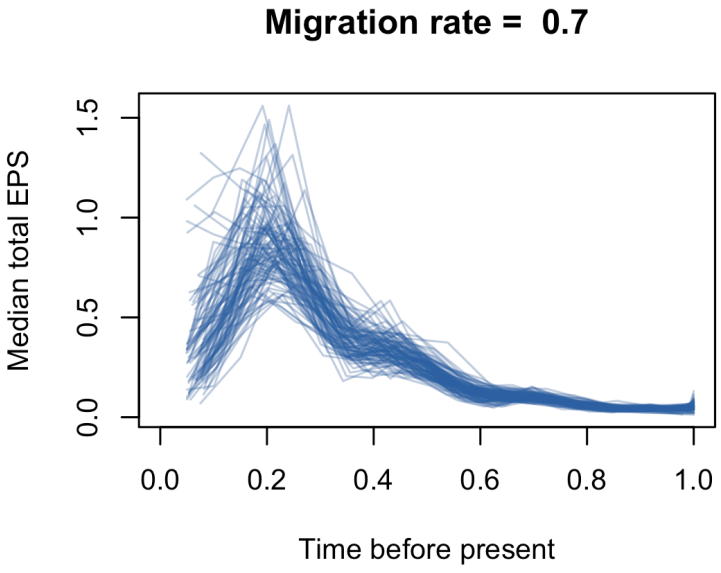
Simulations

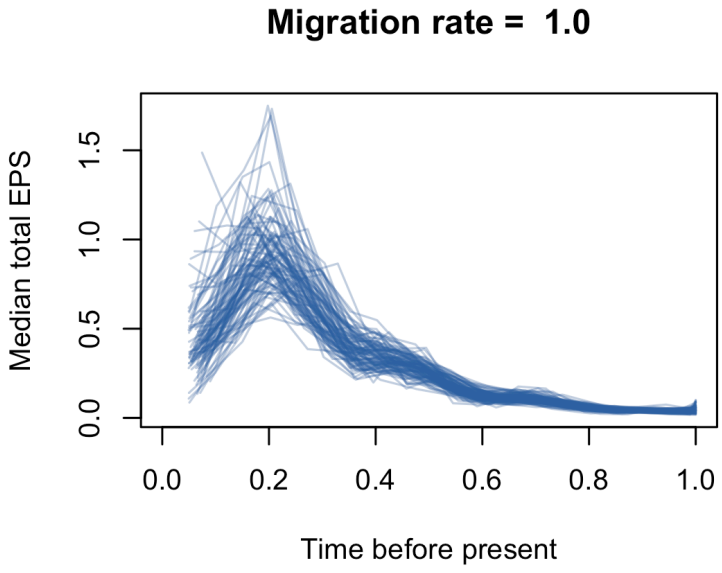
Method:
Bayesian skygrid

grids = 10









Simulations

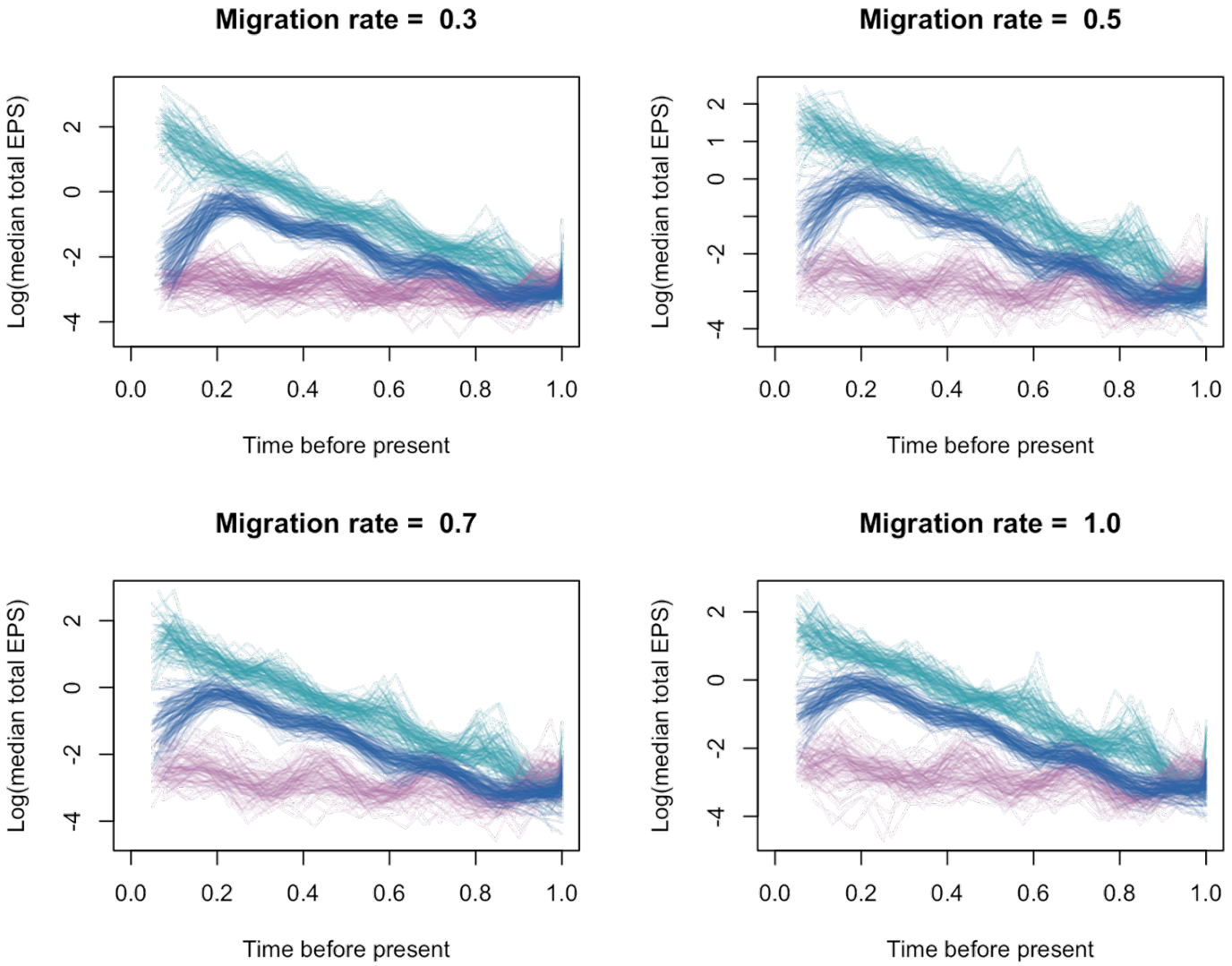
Estimates from Bayesian skygrid skyline

Blue: overall EPS

Estimates from the structured skyline

Purple: EPS estimate of subpopulation 2

Green: EPS estimate of subpopulation 2



Avian influenza

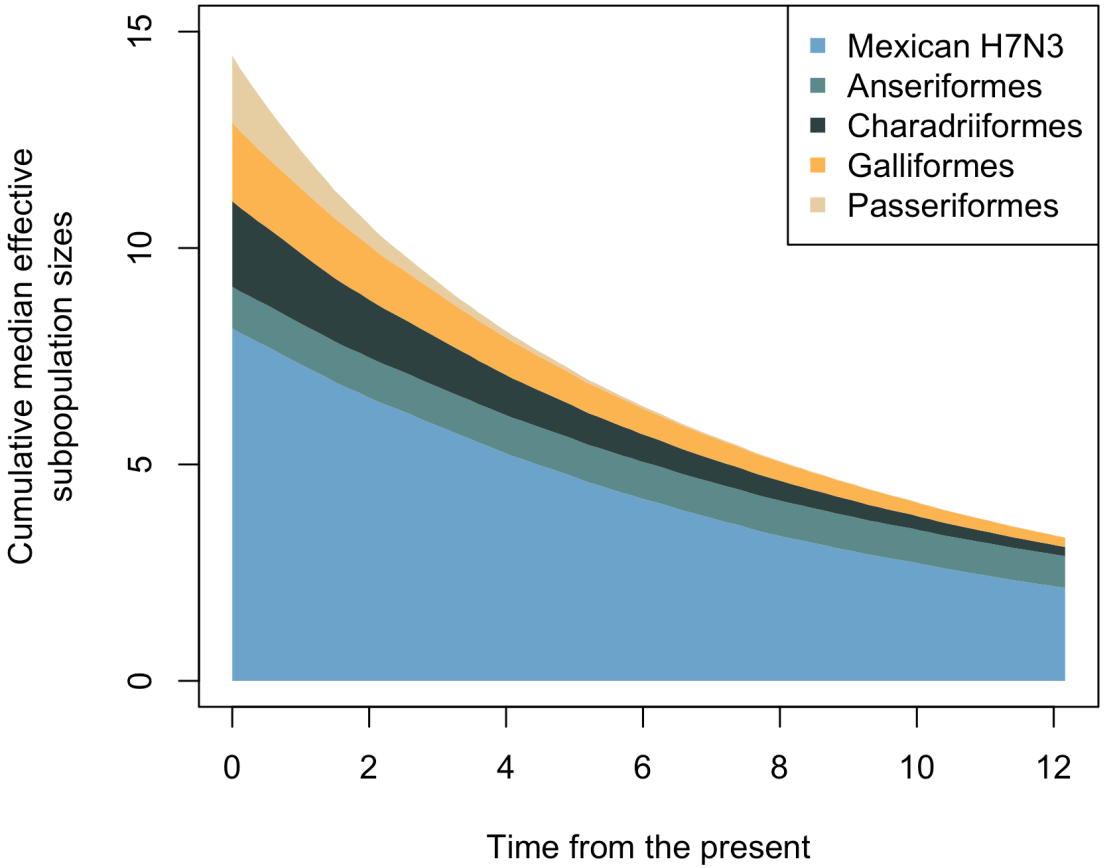
- **Data**
  - This data set was compiled in a study to determine the phylogenetic and phylogeographic origin of H7N3 avian influenza in Mexico (Lu, Lycett and Brown, 2014).
  - Contains HA sequences from 133 isolates
  - The sampling time period starts from June 2001 to June 2012.
  - Host types:
    - Anseriformes = 93
    - Charadriiformes = 30
    - Galliformes = 6
    - Mexico outbreak = 2
    - Passeriformes = 2

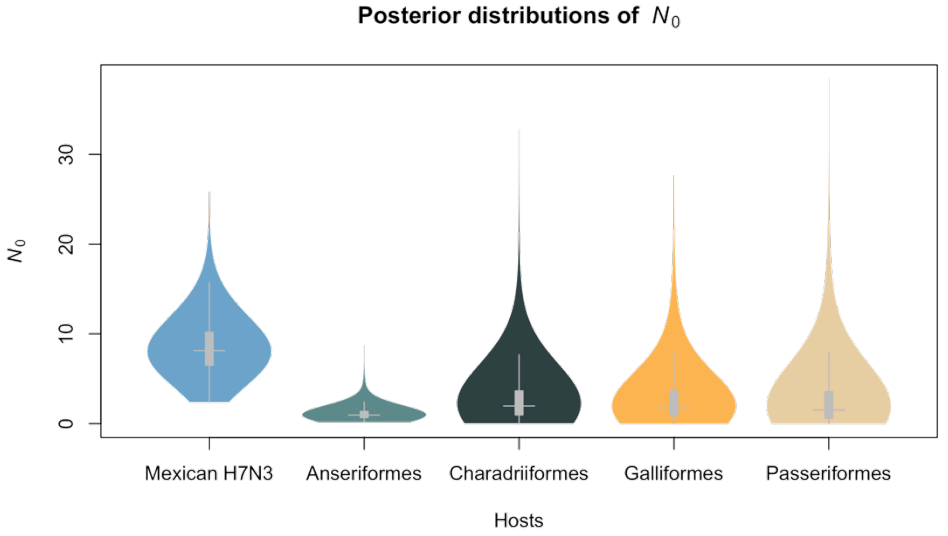
Avian influenza

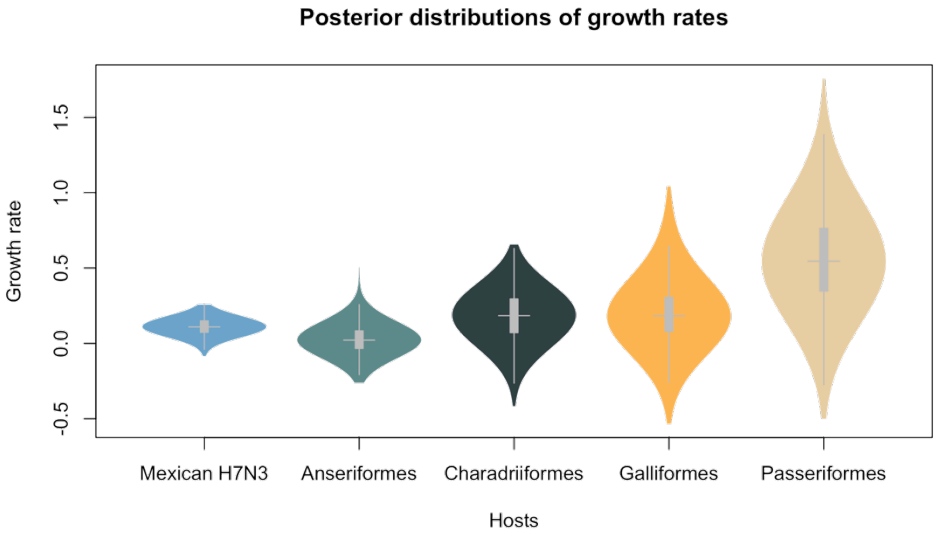
- **Analysis**
  - Assume that each subpopulation follows an exponential growth model and has its own growth rate and current effective population size.
    - $f_d(t) = \theta_{d,0}\exp(-r_d t)$
    - $\theta_d \sim \text{Exp}(\theta)$; $\theta \sim \text{Exp}(1)$
    - $r_d \sim \text{N}(\mu, \sigma^2)$; $\mu \sim \text{N}(\text{mean} = 0, \text{variance} = 100)$; $\sigma^{-2} \sim \text{Exp}(1)$
  - Symmetric migration rates
    - $m_{dd'} \sim \text{Exp}(\theta_m)$; $\theta_m \sim \text{Exp}(1)$

Avian influenza

- **Results**





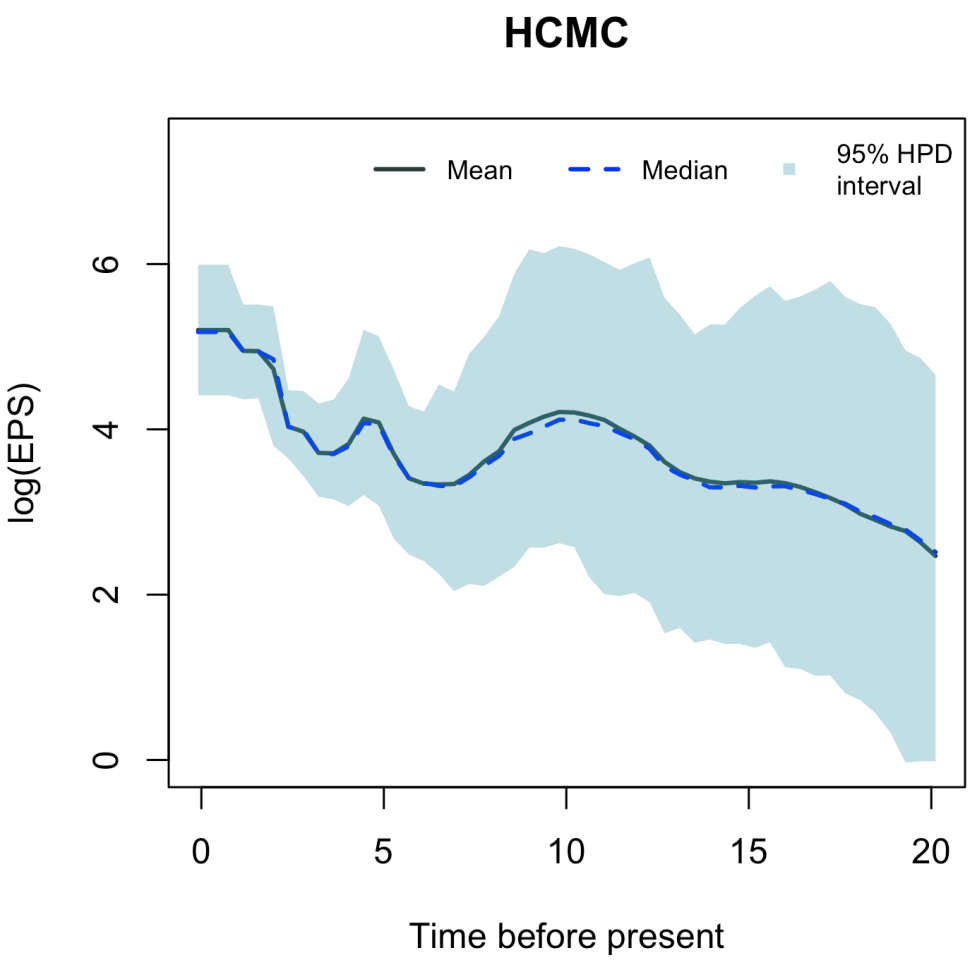


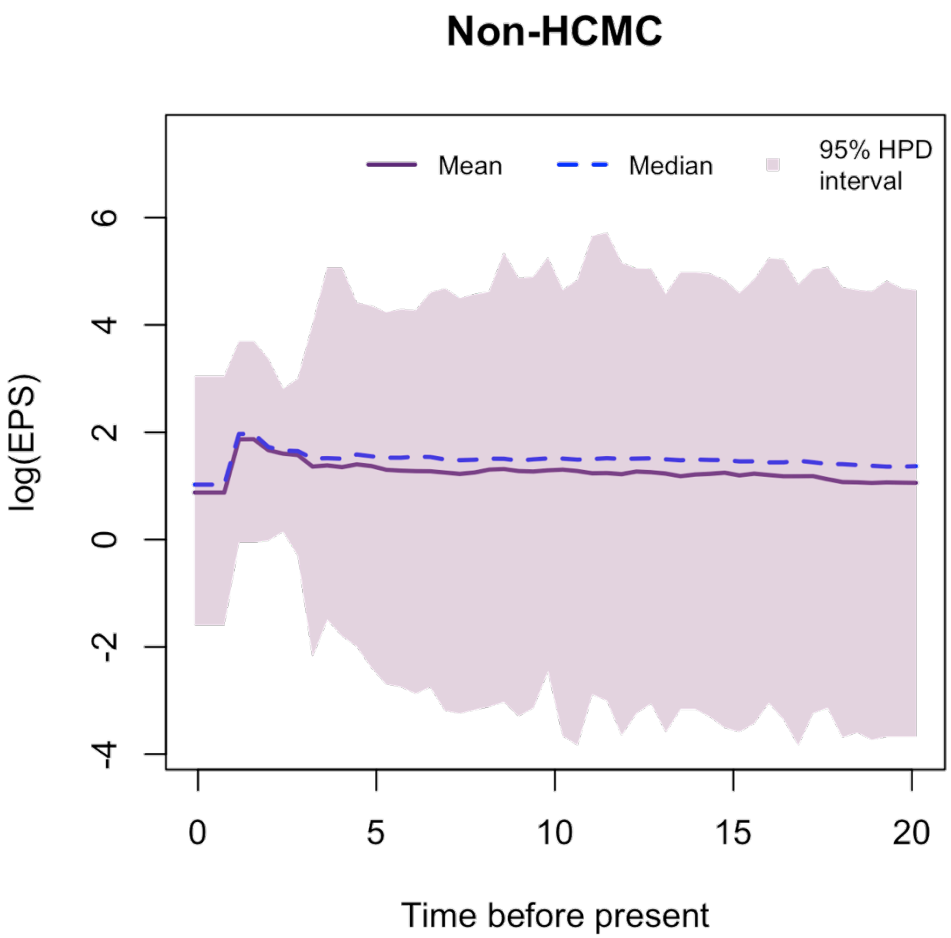
Dengue virus

- **Data**
  - Contains 170 genomic sequences of dengue type 1 isolates
  - The sampling time period starts from February 2003 to February 2008.
  - Locations:
    - Ho Chi Minh city = 85
    - Outside Ho Chi Minh city = 85

- **Analysis**
  - The population trends are estimated using piecewise constant functions (skyline approach)

Dengue virus





Summary

- Ignoring population structure, when it exists, can result in misleading inference of the population trends.
- We present a new method extending an efficient structured coalescent approximation (BASTA; de Maio et al., 2015) to allow inference of the subpopulation trends in structured populations.
  - The new method allows the population size to vary according to a pre-defined parametric function or directly estimated from the data.
  - It also permits the trend of the population sizes to vary across different subpopulations.

Acknowledgements

Nicola de Maio

Daniel Wilson

This project is supported by the Sir Henry Dale fellowship