KimTree: dealing with ascertainment bias and selection using SNP data

Seminar at CBGP, Montpellier

Florian Clemente, Mathieu Gautier, Renaud Vitalis

April 7th, 2016

Overview

- Introduction: KimTree

Overview

- Introduction: KimTree
- Ascertainment bias due to SNP data

Overview

- Introduction: KimTree
- Ascertainment bias due to SNP data
- Model improvements

Overview

- Introduction: KimTree
- Ascertainment bias due to SNP data
- Model improvements
- Application: estimation of sex-ratios in populations

Overview

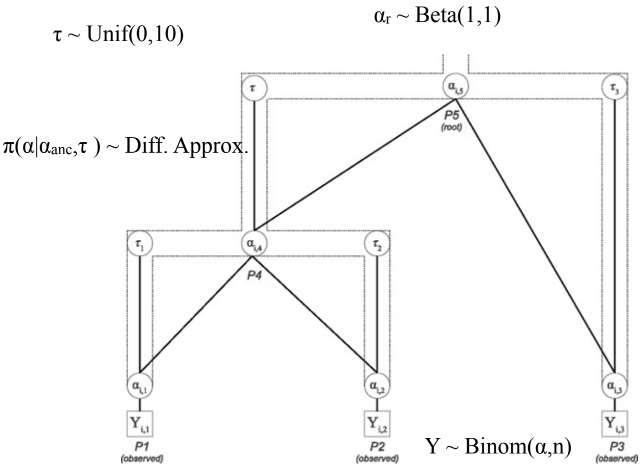
- Introduction: KimTree
- Ascertainment bias due to SNP data
- Model improvements
- Application: estimation of sex-ratios in populations
- Model extension: detection of selective sweeps

KimTree: Gautier and Vitalis, 2013

Assumptions:

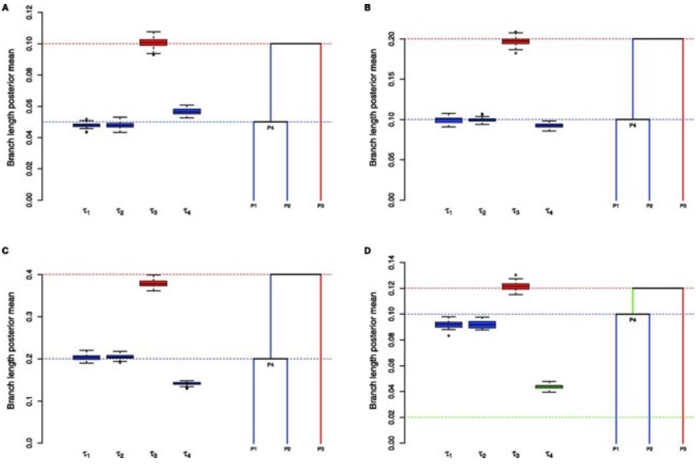
- known population history (tree topology)
- AF evolve according to WF-model (pure-drift process)
- SNPs are segregating independently in root population
- parameter of interest: $\tau_i = t_i/2N$

KimTree: Bayesian Framework



Mathieu Gautier, and Renaud Vitalis Mol Biol Evol 2013;30:654-668

Performance of the Kimura model for estimating branch lengths in population trees.



Mathieu Gautier, and Renaud Vitalis Mol Biol Evol 2013;30:654-668

Tataru et al., 2015 - beta with spikes model

- $f(x;t) = P(X_t = x | X_0 = x_0)$

Tataru et al., 2015 - beta with spikes model

- $f(x;t) = P(X_t = x | X_0 = x_0)$
- approximation: $f_B(x;t) = \frac{x^{\alpha_t - 1}(1-x)^{\beta_t - 1}}{B(\alpha_t, \beta_t)}$, [0,1]
- α_t and β_t are determined by mean and variance
- introduce spikes $\delta(x)$ for loss and fixation probabilities

Tataru et al., 2015 - beta with spikes model

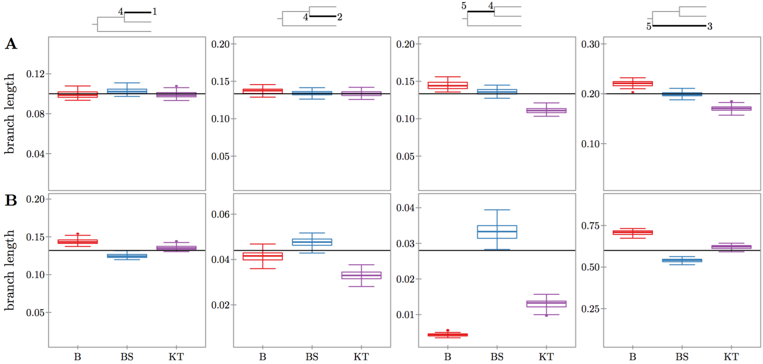
- $f(x;t) = P(X_t = x | X_0 = x_0)$
- approximation: $f_B(x;t) = \frac{x^{\alpha_t - 1}(1-x)^{\beta_t - 1}}{B(\alpha_t, \beta_t)}$, [0,1]
- α_t and β_t are determined by mean and variance
- introduce spikes $\delta(x)$ for loss and fixation probabilities

$$f_B^*(x;t) =$$

$$P(X_t = 0)\delta(x) +$$
$$P(X_t = 1)\delta(1-x) +$$
$$P(X_t \notin \{0,1\}) \frac{x^{\alpha_t^* - 1}(1-x)^{\beta_t^* - 1}}{B(\alpha_t^*, \beta_t^*)}$$

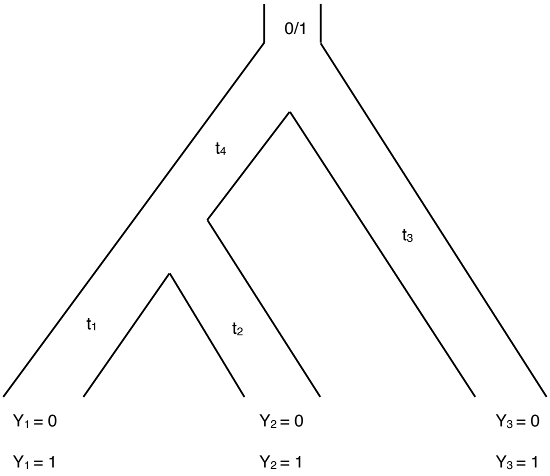
Tataru et al., 2015 - beta with spikes model vs KimTree

A Simulations: B(1.0,1.0) **B** Simulations: chimp exome B(0.0188, 0.0195)

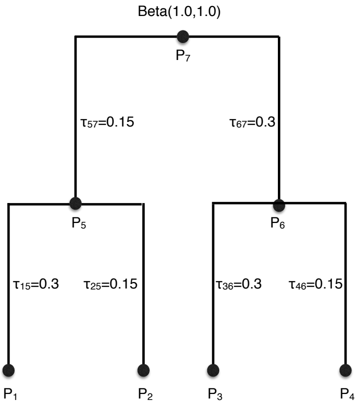


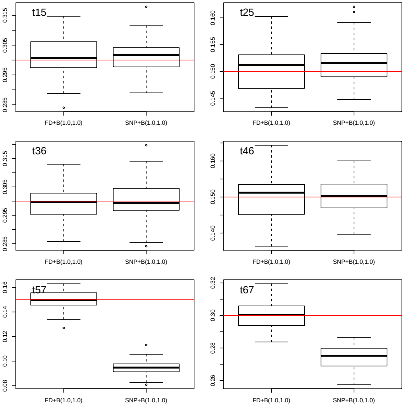
Ascertainment bias due to SNP data

Problem: mutations that get lost or become fixed in all populations



Full data check: 5000 markers simulated under the inference model

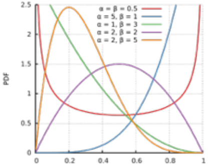




1st approach - flexible Beta(a,b)

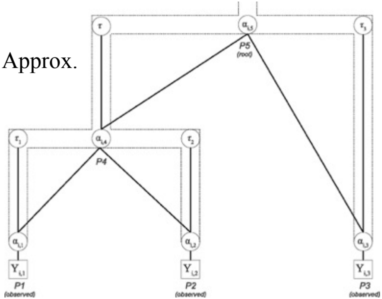
$\tau \sim \text{Unif}(0,10)$

$\alpha_r \sim \text{Beta}(\alpha,\beta)$



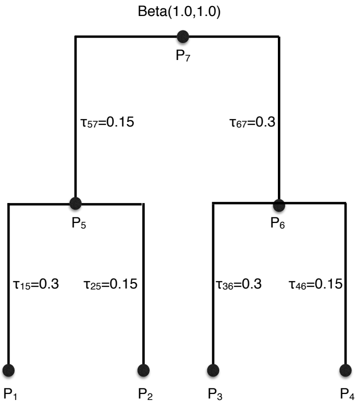
source: Wikipedia

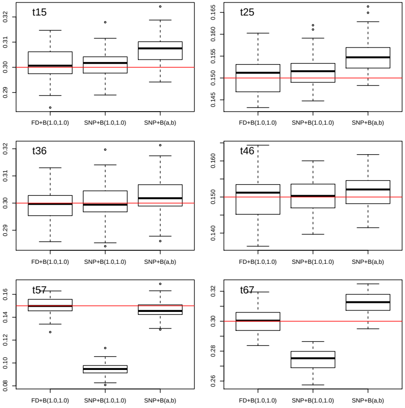
$\pi(\alpha|\alpha_{anc},\tau\ ) \sim$ Diff. Approx.



$Y \sim \text{Binom}(\alpha,n)$

Full data vs SNPs: flexible Beta(a,b)





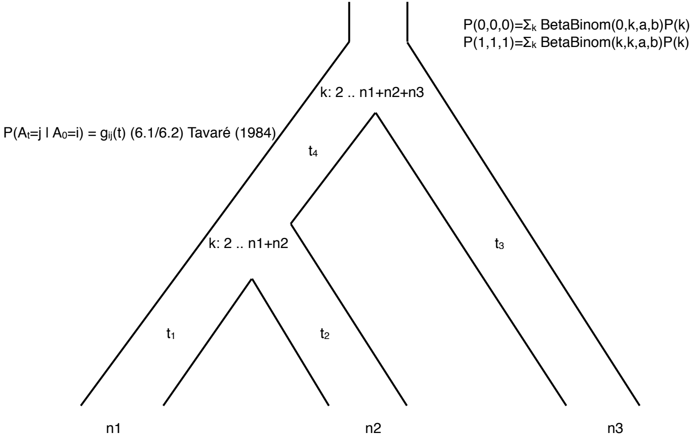
2nd approach - conditional likelihood

- $\prod_i L(Y_i; \Theta|\mathrm{poly}_i) = \prod_i L(Y_i; \Theta) / P(\mathrm{poly}_i|\Theta)$

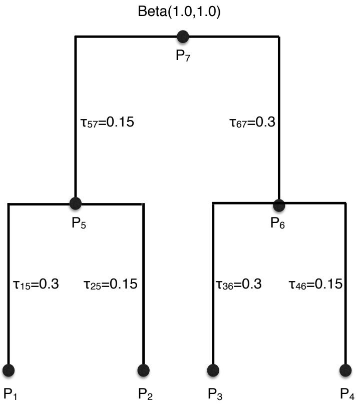
2nd approach - conditional likelihood

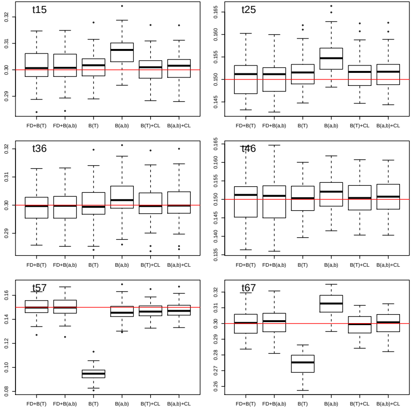
- $\prod_i L(Y_i;\Theta|\text{poly}_i) = \prod_i L(Y_i;\Theta)/P(\text{poly}_i|\Theta)$
- $P(\text{poly}_i|\Theta) = 1 - P(Y_i^{(N)} = 0|\Theta) - P(Y_i^{(N)} = 1|\Theta)$

Coalescent theory



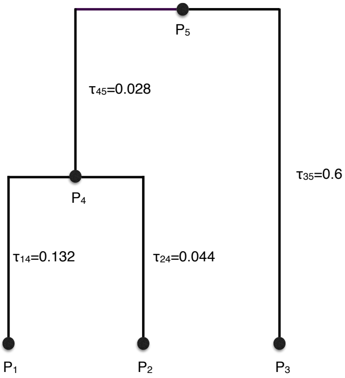
Full data vs SNPs: conditional likelihood

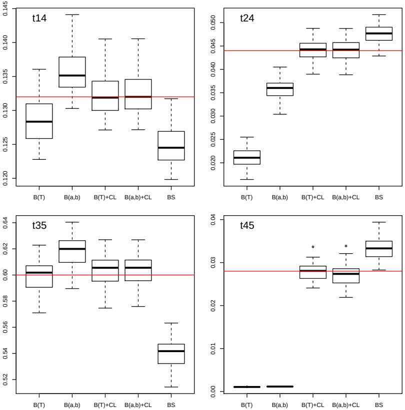




Tataru et al., 2015: KimTree vs beta with spikes model

chimp data B(0.0188, 0.0195):





KimTree: Limitations

- divergence times are in a diffusion time scale
- model does not use LD information
- model assumes no mutations after MRCA

Application: estimation of sex-ratios

Application: estimation of sex-ratios - Definitions

effective sex ratio: $\rho := \frac{N_e^f}{N_e^f + N_e^m}$

Application: estimation of sex-ratios - Definitions

effective sex ratio: $\rho := \frac{N_e^f}{N_e^f + N_e^m}$

- monogamy: $E[\rho] = 0.5$
- polygamy
  - polygyny: $E[\rho] > 0.5$
  - polyandry: $E[\rho] < 0.5$

Application: estimation of sex-ratios

contribution of males and females to strength of genetic drift differs on autosomes and sex-chromosomes

- if $N_e^f = N_e^m \Rightarrow N_e^X = \frac{3}{4} N_e^A, N_e^Y = \frac{1}{4} N_e^A$

Application: estimation of sex-ratios

contribution of males and females to strength of genetic drift differs on autosomes and sex-chromosomes

- if $N_e^f = N_e^m \Rightarrow N_e^X = \frac{3}{4} N_e^A, N_e^Y = \frac{1}{4} N_e^A$
- $N_e^A = \frac{4 N_e^f N_e^m}{N_e^f + N_e^m}, \qquad N_e^X = \frac{9 N_e^f N_e^m}{2 N_e^f + 4 N_e^m}$ Crow & Kimura (1971)
- $\rho = \frac{N_e^f}{N_e^f + N_e^m} = 2 - \frac{9}{8\lambda}, \lambda = \frac{N_e^X}{N_e^A}$

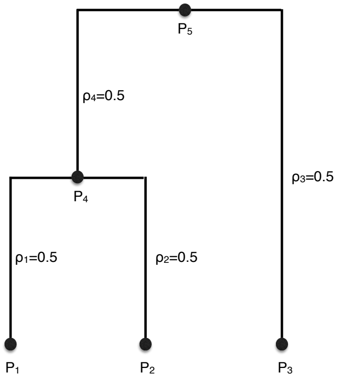
Application: estimation of sex-ratios

contribution of males and females to strength of genetic drift differs on autosomes and sex-chromosomes

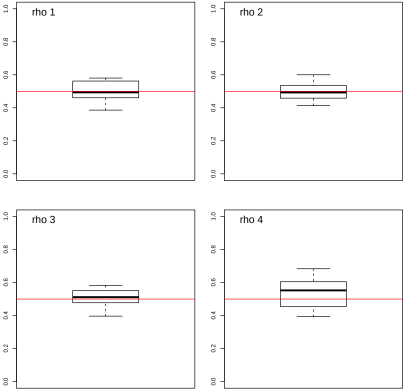
- if $N_e^f = N_e^m \Rightarrow N_e^X = \frac{3}{4}N_e^A, N_e^Y = \frac{1}{4}N_e^A$
- $N_e^A = \frac{4N_e^f N_e^m}{N_e^f + N_e^m}, \qquad N_e^X = \frac{9N_e^f N_e^m}{2N_e^f + 4N_e^m}$ Crow & Kimura (1971)
- $\rho = \frac{N_e^f}{N_e^f + N_e^m} = 2 - \frac{9}{8\lambda}, \lambda = \frac{N_e^X}{N_e^A}$
- KimTree: $\tau_A = \frac{t}{2N_e^A}; \tau_X = \frac{t}{2N_e^X}; \lambda = \frac{\tau_e^A}{\tau_e^X}$

Results

Scenario 1: $N_e^f + N_e^m = 1000$, 50 data sets of 5000 SNPs for A and X (*IBD_sex*, Vitalis et al., in prep.)

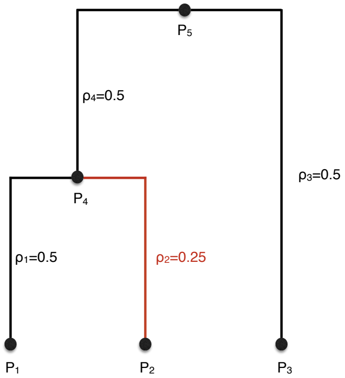


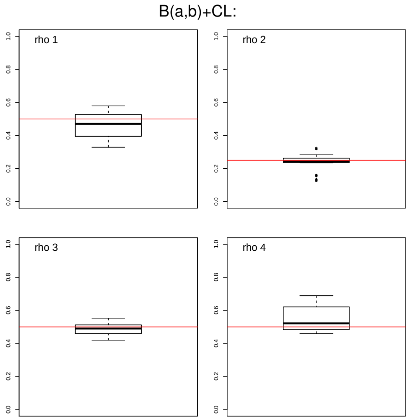
B(a,b)+CL:



Results

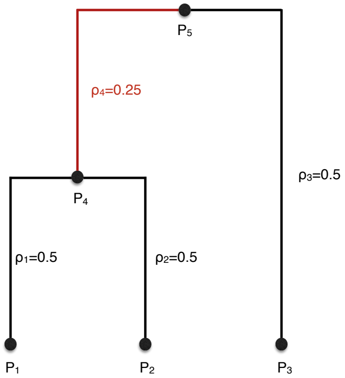
Scenario 2: $N_e^f + N_e^m = 1000$, 50 data sets of 5000 SNPs for A and X (*IBD_sex*, Vitalis et al., in prep.)

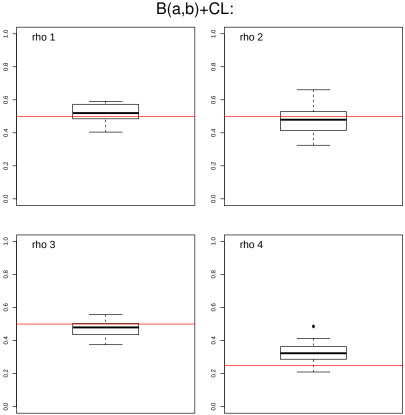




Results

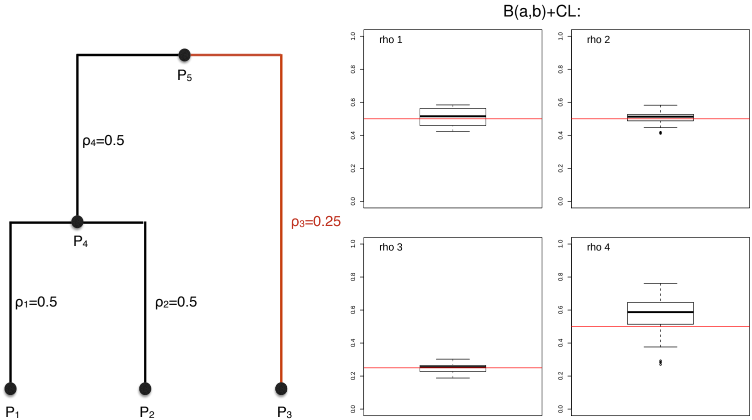
Scenario 3: $N_e^f + N_e^m = 1000$, 50 data sets of 5000 SNPs for A and X (*IBD_sex*, Vitalis et al., in prep.)





Results

Scenario 4: $N_e^f + N_e^m = 1000$, 50 data sets of 5000 SNPs for A and X (*IBD_sex*, Vitalis et al., in prep.)



Sex-ratio estimation: Limitations

- A and X-linked variation depend on:
  - population size changes Pool and Nielsen, 2007
  - positive selection, background selection Hammer et al, 2008
  - sex-specific migration
  - sex-specific mutation rates Labuda et al, 2010

Sex-ratio estimation: Limitations

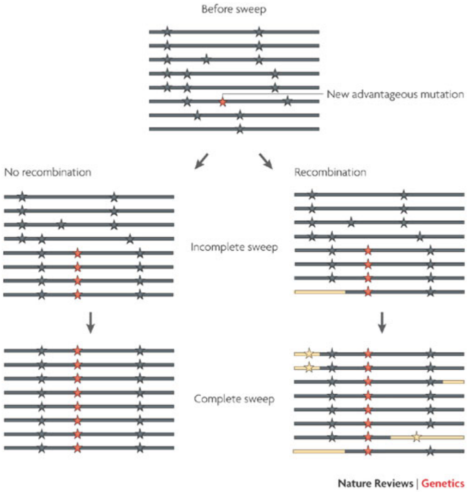
- A and X-linked variation depend on:
  - population size changes Pool and Nielsen, 2007
  - positive selection, background selection Hammer et al, 2008
  - sex-specific migration
  - sex-specific mutation rates Labuda et al, 2010
- methodological differences (F_{st} vs π) Emery et al, 2010

Sex-ratio estimation: Future perspective

- test effects of population size changes and demographies in general
- apply model to real data

Model extension: detection of selective sweeps

Selective Sweep



Nature Reviews Genetics 8, 857-868 (November 2007)

Selection model: Chen et al. (2010), Genome Research

- Nicholson model: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{\frac{(x-p_0)^2}{2\sigma^2}}$, $\sigma^2 = \omega p_0(1-p_0)$

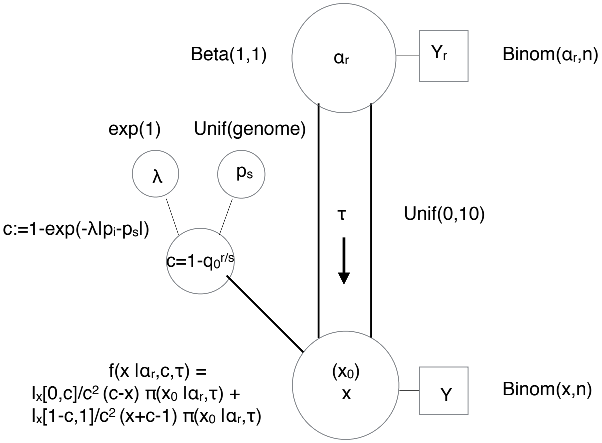
Selection model: Chen et al. (2010), Genome Research

- Nicholson model: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{\frac{(x-p_0)^2}{2\sigma^2}}$, $\sigma^2 = \omega p_0(1 - p_0)$
- joint effect of selection & recombination: Smith & Haigh (1974)
  - $x_{AB} = 1 - c + cx_0$; $x_{aB} = cx_0$
  - $c \approx 1 - q_0^{r/s}$

Selection model: Chen et al. (2010), Genome Research

- Nicholson model: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{\frac{(x-p_0)^2}{2\sigma^2}}$, $\sigma^2 = \omega p_0(1-p_0)$
- joint effect of selection & recombination: Smith & Haigh (1974)
  - $x_{AB} = 1 - c + cx_0$; $x_{aB} = cx_0$
  - $c \approx 1 - q_0^{r/s}$
- $f(p_1|r, s, p_2, \omega) =$
  $\frac{1}{\sqrt{2\pi\sigma}} \frac{p_1+c-1}{c^2} \mathrm{e}^{\frac{(p_1+c-1-cp_2)^2}{2c^2\sigma^2}} \mathrm{I}_{[1-c,1]}(p_1) + \frac{1}{\sqrt{2\pi\sigma}} \frac{c-p_1}{c^2} \mathrm{e}^{\frac{(p_1-cp_2)^2}{2c^2\sigma^2}} \mathrm{I}_{[0,c]}(p_1)$

KimTree with selection (simplified model)

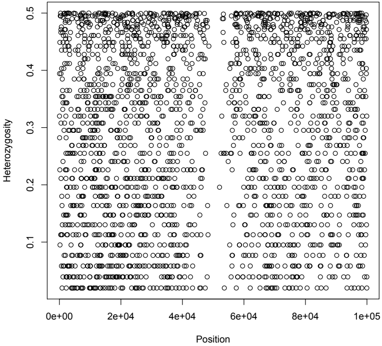


SLiM: Simulating Evolution with Selection and Linkage Philipp W. Messer, 2013

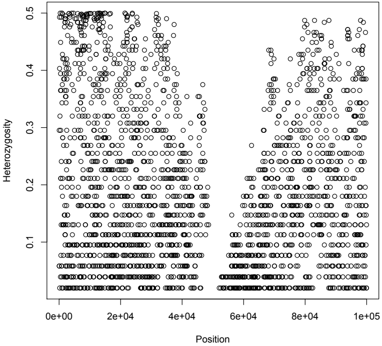
- Neutral phase:
  - 10000 generations
  - locus $L = 100000$ bp
  - effective popSize $N_e = 1000$
  - mutation rate $\mu = 2.5e - 6$
  - recombination rate $r = 2.5e - 5$
- Selection phase:
  - 101 generations
  - $pos_s = 50000$
  - selection coeff. $s = 5$
  - mutation rate $\mu = 0$

SLiM: time series data

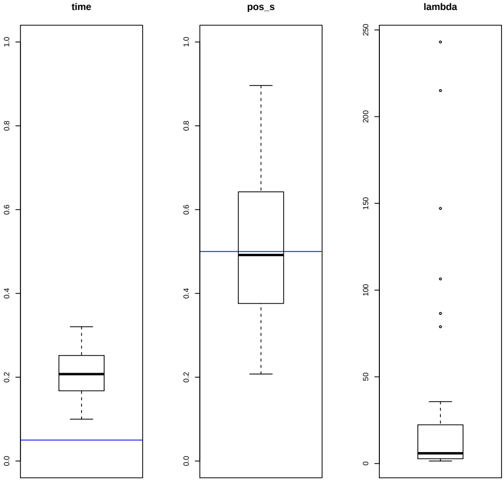
before sweep:



after sweep:



Results: 50 posterior means



Selection Model: Future perspective

$\alpha_r \sim \mathrm{Beta}(\alpha,\beta)$

$\tau \sim \mathrm{Unif}(0,10)$

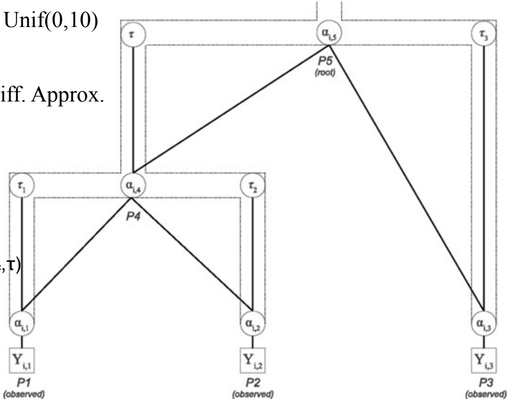
$\pi(\alpha|\alpha_{anc},\tau) \sim$ Diff. Approx.

$$f(x\,|\alpha_{anc},c,\tau) = I_x[0,c]/c^2\,(c-x)\,\pi(x_0\,|\alpha_{anc},\tau) + I_x[1-c,1]/c^2\,(x+c-1)\,\pi(x_0\,|\alpha_{anc},\tau)$$

$\lambda \sim \exp(1)$

$p_s \sim \mathrm{Unif(genome)}$

$c := 1-\exp(-\lambda|p_i - p_s|)$



$Y \sim \mathrm{Binom}(\alpha,n)$

Selection Model: Future perspective

- include information from fixed sites or LD
- estimate strength of selection s and recombination rate r
- apply model to real data

Thank you!