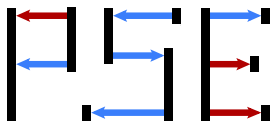
Multiple-merger coalescents - extended models, inference methods & evidence

F. Freund, (U. Hohenheim)
joint works w.
S. Matuszewski, M. Lapierre, E. Kerdoncuff (SMILE Paris), A. Lambert(SMILE Paris), J. Jensen (U. Arizona), G. Achaz (SMILE Paris)
A. Siri- Jégousse (UNAM Mexico City)
and F. Menardo (Swiss Tropical and Public Health Institute, Basel)

UNIVERSITY OF HOHENHEIM

200 1818 2018 YEARS



PROBABILISTIC STRUCTURES IN EVOLUTION

DFG SPP 1590

Probabilités et évolution biologique, Luminy, 28.06.2018

What do we want to infer?

Data:

Sample of size n of present-day genetic data (DNA sequences) from a population of a single species

- Can we pinpoint loci under positive selection?
- Can we infer the evolutionary history of the processes?

What do we want to infer?

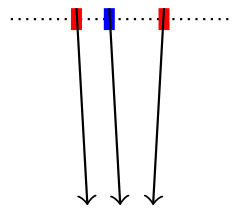
Data:

Sample of size n of present-day genetic data (DNA sequences) from a population of a single species

- Can we pinpoint loci under positive selection? Search for deviation of data from selectively neutral model (includes ALL other processes affecting genetic diversity)
- Can we infer the evolutionary history of the processes? Model selection between biologically reasonable models

Model the genetic diversity of n DNA SNP sequences

genetic locus (no recomb.)



0 1 1

0 1 0

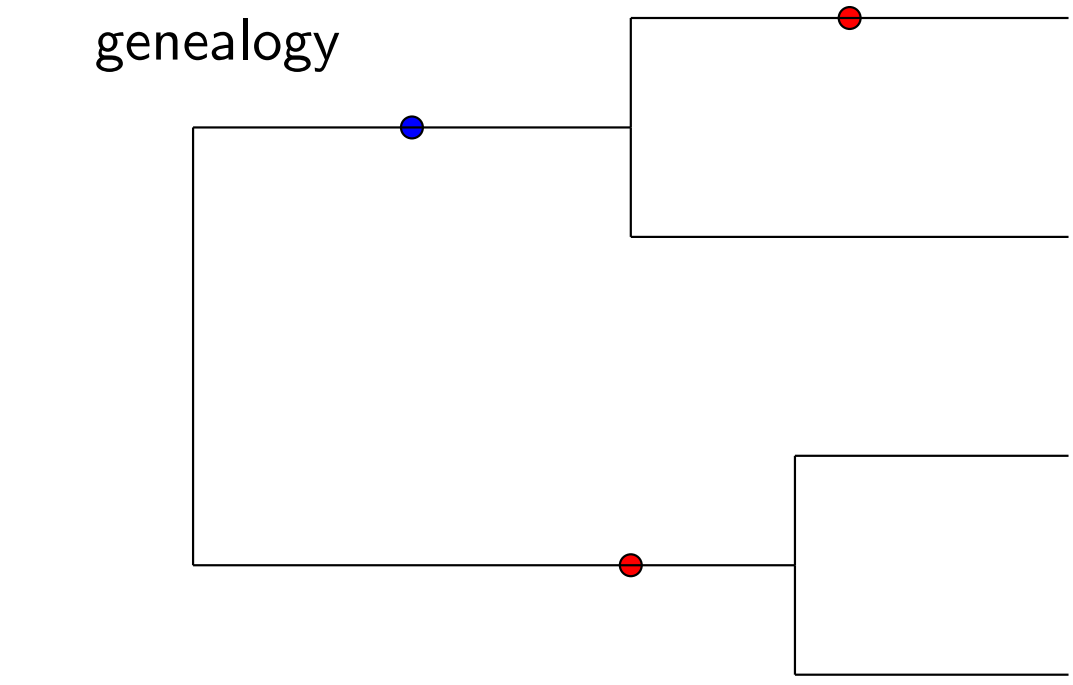
1 0 0

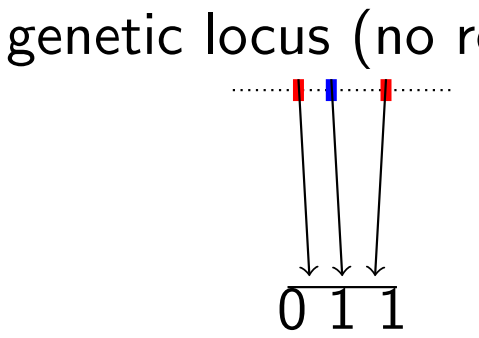
1 0 0

observed

Model the genetic diversity of n DNA SNP sequences

- Genealogy: random tree w. n leaves
- Mutation: Poisson PP w. rate $\frac{\theta}{2}$, infinite-sites m.
- Mutation is neutral: independent of genealogy





not observed

observed

Which evolutionary forces affect the genealogy?

Standard null model: Kingman's n-coalescent $\mathbb{KM}$

- Strictly bifurcating $\Leftrightarrow$ no multiple mergers
- Assumptions (robust to small deviations): Sample from large, randomly mating fixed size population, offspring distributions non-skewed (e.g. offspring/parent has bounded variance), no selection

Which evolutionary forces affect the genealogy?

Standard null model: Kingman's n-coalescent $\mathbb{KM}$

- Strictly bifurcating $\Leftrightarrow$ no multiple mergers
- Assumptions (robust to small deviations): Sample from large, randomly mating fixed size population, offspring distributions non-skewed (e.g. offspring/parent has bounded variance), no selection

Preserves bifurcation

- Moderate population size fluctuations (e.g. exponential growth)
- Population structure
- Moderate positive selection (affects locally)
- Seed banks
- (Recombination & HGT)

Which evolutionary forces affect the genealogy?

Standard null model: Kingman's n-coalescent $\mathbb{KM}$

- Strictly bifurcating $\Leftrightarrow$ no multiple mergers
- Assumptions (robust to small deviations): Sample from large, randomly mating fixed size population, offspring distributions non-skewed (e.g. offspring/parent has bounded variance), no selection

Preserves bifurcation

- Moderate population size fluctuations (e.g. exponential growth)
- Population structure
- Moderate positive selection (affects locally)
- Seed banks
- (Recombination & HGT)

May lead to multiple mergers

- Extreme, repeated bottlenecks
- Skewed offspring distributions (Reproduction sweepstakes)
- Rapid selection
- Recurrent selective sweeps

Genealogy: (Λ-)n-coalescents with multiple mergers

Random tree: n leaves, random branch lengths. Goes backwards in time.

Genealogy: (Λ-)n-coalescents with multiple mergers

Random tree: n leaves, random branch lengths. Goes backwards in time.

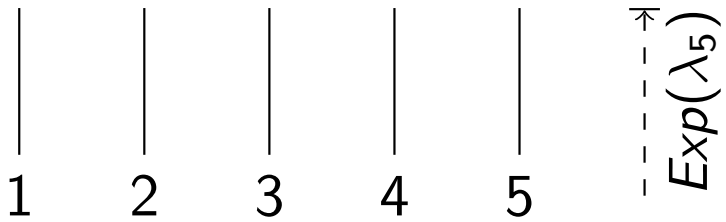
S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample

1 2 3 4 5

Genealogy: (Λ-)n-coalescents with multiple mergers

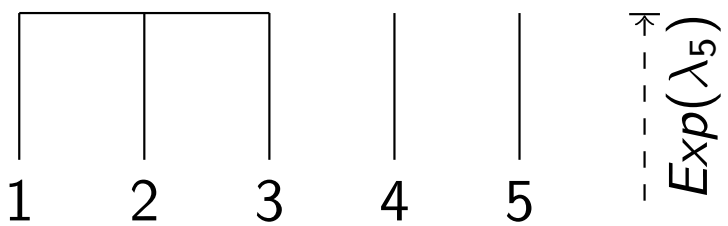
Random tree: n leaves, random branch lengths. Goes backwards in time.

- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$



Genealogy: (Λ-)n-coalescents with multiple mergers

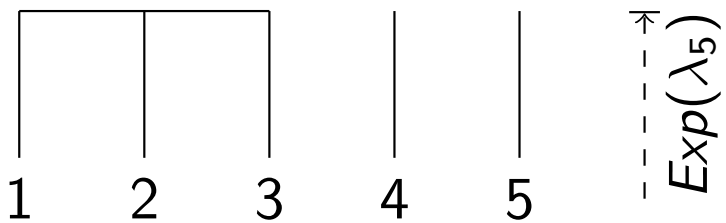
Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$

Genealogy: (Λ-)n-coalescents with multiple mergers

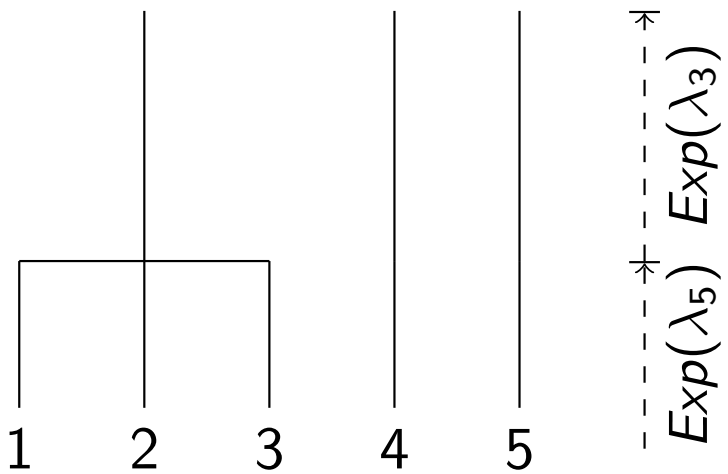
Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

Genealogy: (Λ-)n-coalescents with multiple mergers

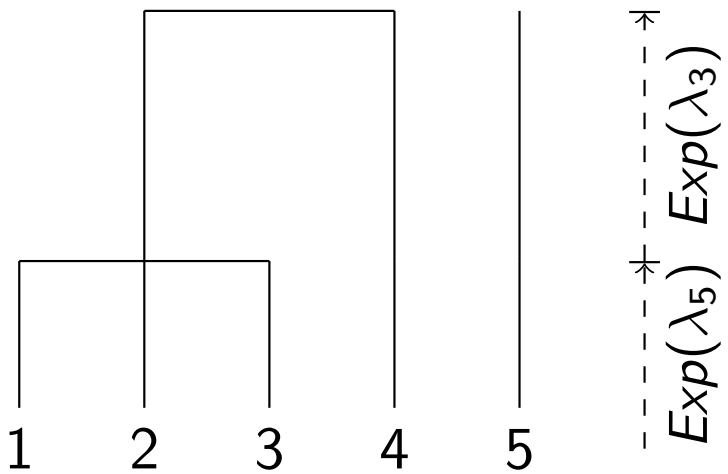
Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

Genealogy: (Λ-)n-coalescents with multiple mergers

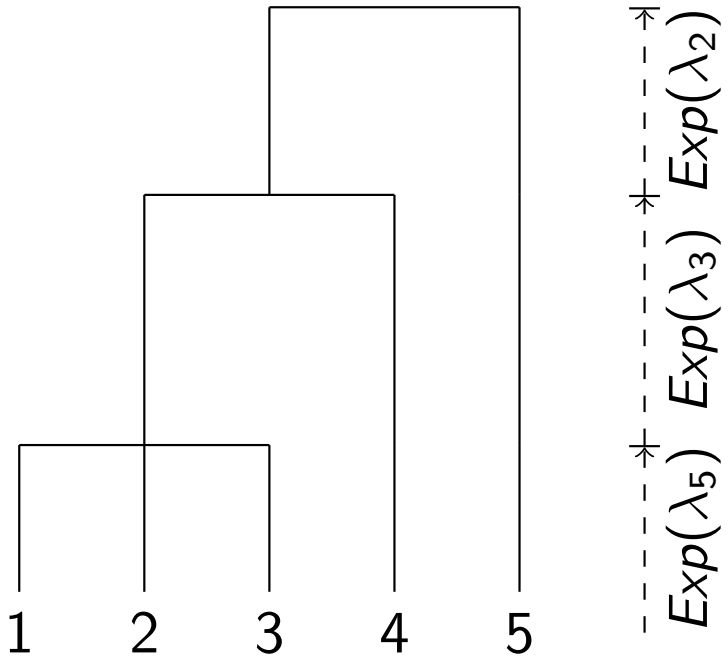
Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

Genealogy: (Λ-)n-coalescents with multiple mergers

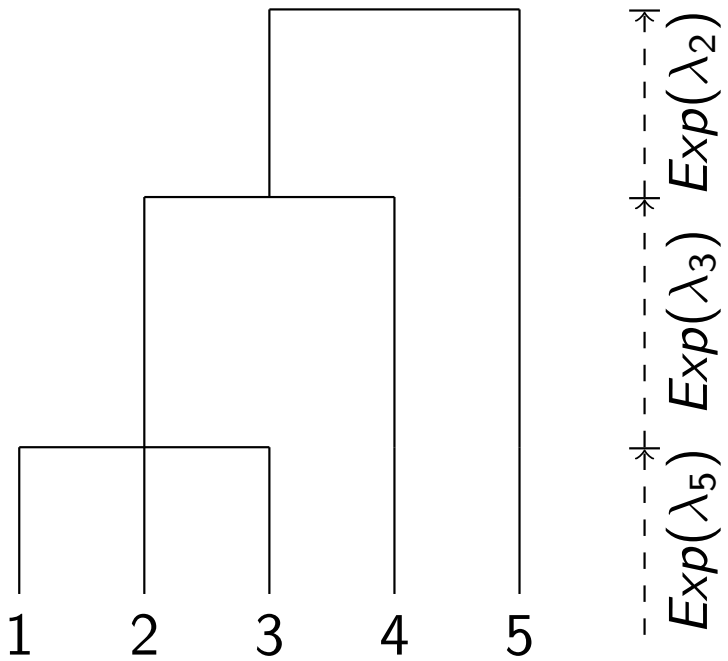
Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

Genealogy: (Λ-)n-coalescents with multiple mergers

Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

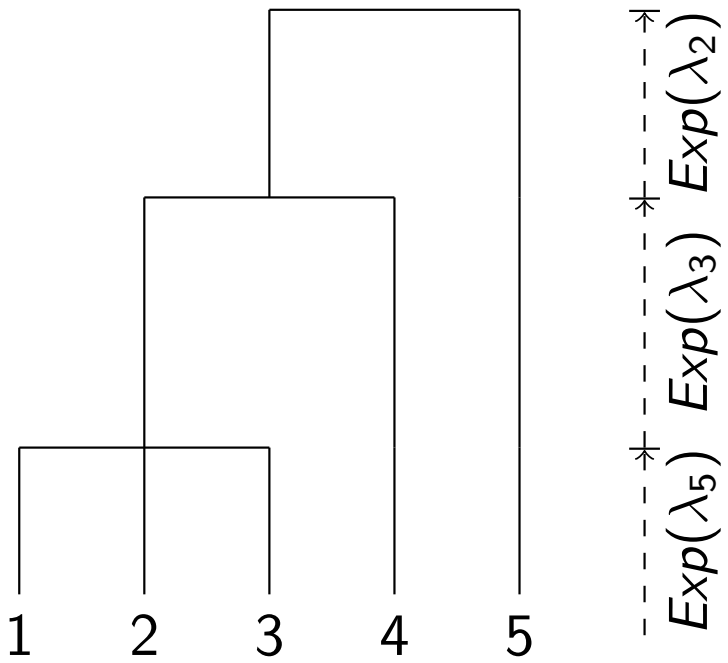
$\lambda_{n,k} = \int x^k (1-x)^{n-k} x^{-2} \Lambda(dx)$, Λ finite measure on $[0,1]$

Beta-n-coalescents $\mathbb{BETA}$ $\Lambda = \mathcal{B}(2-\alpha, \alpha)$, $1 \leq \alpha \leq 2$

Dirac-n-coalescents $\mathbb{DIRAC}$ $\Lambda = \delta_p$ (point mass in $p \in (0,1]$)

Genealogy: (Λ-)n-coalescents with multiple mergers

Random tree: n leaves, random branch lengths. Goes backwards in time.



- S0 We have l lineages (n at time 0) present. Only one lineage left: Reached MRCA of the sample
- S1 Prolongue/set each lineage by/to $Exp(\lambda_l)$
- S2 Add an ancestral lineage which connects $2 \leq k \leq l$ present lineages chosen randomly, k is chosen with probability $\frac{\lambda_{l,k}}{\sum_k \lambda_{l,k}} = \frac{\lambda_{l,k}}{\lambda_l}$
- S3 Return to Step $S0$

$$\lambda_{n,k} = \int x^k (1-x)^{n-k} x^{-2} \Lambda(dx),\ \Lambda = \mathcal{B}(2-\alpha, \alpha),\ \alpha \in \{0, 1\}$$

$\alpha = 1$: Bolthausen-Sznitman n-coalescent $\mathbb{BSZ}$

$\alpha = 2$: Kingman's n-coalescent ($\Lambda = \delta_0$)$\mathbb{KM}$

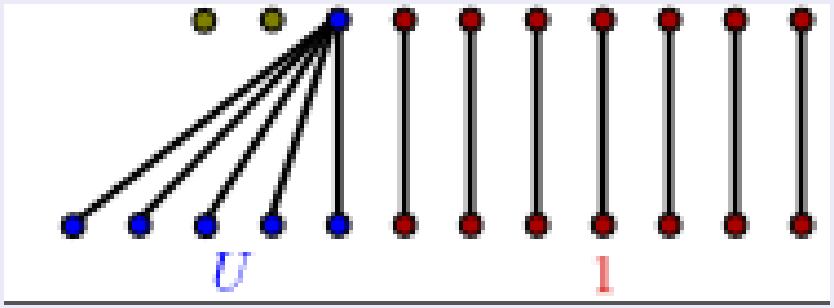
Some models with Λ-n-coalescent genealogy (Population size $N \to \infty$)

- $\mathbb{BSZ}$ (w. $+1$ to all external branch lengths): Clonal interference of equal-effect mutations [DWF13],[Sch17]
- $\mathbb{BSZ}$: Fixed-size population with increasing fitness given by a travelling wave, e.g. [BBS13],[BD13], [NH13]
- $\mathbb{BETA}$: Random sampling from a supercritical Galton-Watson process (offspring distribution heavy-tailed) [Sch03]
- $\mathbb{DIRAC}$: Modified Moran models ("lucky" individual produces 2 or U offspring) with fixed $U = \Psi N$ [EW06]
- Recurring extinction-recolonisation pattern (as # demes $\to \infty$) [TV09]

Any Λ-n-coalescent: limit genealogy of a family of modified Moran models (for $N \to \infty$) [HM13]

Adding exponential growth to multiple mergers

[MHAJ17]



from [MHAJ17]

Eldon-Wakeley model + exponential growth

- Population size back k generations: $N_k = \lfloor (1 - \frac{\Psi^2}{N^\gamma})^k \rfloor$
- Large N: $U_{N_k} = N_k \Psi 1_{\{BIG\}} + 2 \cdot 1_{\{small\}}$, $P(BIG) = 1 - P(small) = N_k^{-\gamma}$, $\Psi \in (0,1)$

For $N \to \infty$, $\gamma \in (0,2)$: Genealogy converges to a time-changed Dirac-n-coalescent $(\Pi_{\mathcal{G}_t})_{t \geq 0}$ w. $\Lambda = \delta_\Psi$ and $\mathcal{G}_t = (\rho\gamma)^{-1}(e^{\rho\gamma t} - 1)$ if one scales time in the discrete models by c_N^{-1} w. $c_N = P_N$(1,2 share parent) from the fixed model.

Proof idea: Use [Möh02]

- For the modified Moran models with changing pop. sizes, we want to lump # generations given by the pseudo-inverse $\mathcal{G}_N^{-1}(t)$ of $\sum_{r=1}^{[s]} c_{N,r}$, $c_{N,r} = P(1,2 \text{ have same parent})$ to end up at coalescent time t

Proof idea: Use [Möh02]

- For the modified Moran models with changing pop. sizes, we want to lump # generations given by the pseudo-inverse $\mathcal{G}_N^{-1}(t)$ of $\sum_{r=1}^{[s]} c_{N,r}$, $c_{N,r} = P(1,2 \text{ have same parent})$ to end up at coalescent time t
- Check that on this scale, the population changes moderately ($\sup c_{N,r} \to 0$, $\inf N_r \to \infty$) the transition

Proof idea: Use [Möh02]

- For the modified Moran models with changing pop. sizes, we want to lump # generations given by the pseudo-inverse $\mathcal{G}_N^{-1}(t)$ of $\sum_{r=1}^{[s]} c_{N,r}$, $c_{N,r} = P(1,2 \text{ have same parent})$ to end up at coalescent time t
- Check that on this scale, the population changes moderately ($\sup c_{N,r} \to 0$, $\inf N_r \to \infty$) the transition
- Convergence of the time-scaled discrete models follows if

$$\lim_{N\to\infty} \sum_{r=1}^{\mathcal{G}_N^{-1}(t)} \Phi_l^{(N)}(r; a_1, \ldots, a_l) < \infty$$

$\Phi_l^{(N)}(r; a_1, \ldots, a_l) = P_{\text{gen -r,}\sum_i a_i\text{ind.}}((a_i \text{ ind. common parent})_i)$. If it has the form $q_{a_1,\ldots,a_l} t$, the limit is a time-homogeneous Markov process with rates q.

Proof idea: Use [Möh02]

- For the modified Moran models with changing pop. sizes, we want to lump # generations given by the pseudo-inverse $\mathcal{G}_N^{-1}(t)$ of $\sum_{r=1}^{[s]} c_{N,r}$, $c_{N,r} = P(1,2 \text{ have same parent})$ to end up at coalescent time t
- Check that on this scale, the population changes moderately ($\sup c_{N,r} \to 0$, $\inf N_r \to \infty$) the transition
- Convergence of the time-scaled discrete models follows if

$$\lim_{N\to\infty} \sum_{r=1}^{\mathcal{G}_N^{-1}(t)} \Phi_l^{(N)}(r; a_1, \ldots, a_l) < \infty$$

  $\Phi_l^{(N)}(r; a_1, \ldots, a_l) = P_{\text{gen -r,}\sum_i a_i\text{ind.}}((a_i \text{ ind. common parent})_i)$. If it has the form $q_{a_1,\ldots,a_l} t$, the limit is a time-homogeneous Markov process with rates q.

For the modified Dirac coalescents, one gets the only non-zero limit $\lim_{N\to\infty} \sum_{r=1}^{\mathcal{G}_N^{-1}(t)} \Phi_l^{(N)}(r; a_1, \ldots, a_l) = \Psi^{k-2} t = \int x^{k-2}(1-x)^0 \delta_\Psi(dx)$ for $l = 1$ and $a_1 \geq 2$.

Proof idea: Use [Möh02]

- For the modified Moran models with changing pop. sizes, we want to lump # generations given by the pseudo-inverse $\mathcal{G}_N^{-1}(t)$ of $\sum_{r=1}^{[s]} c_{N,r}$, $c_{N,r} = P(1,2 \text{ have same parent})$ to end up at coalescent time t
- Check that on this scale, the population changes moderately ($\sup c_{N,r} \to 0$, $\inf N_r \to \infty$) the transition
- Convergence of the time-scaled discrete models follows if

$$\lim_{N\to\infty} \sum_{r=1}^{\mathcal{G}_N^{-1}(t)} \Phi_l^{(N)}(r; a_1, \ldots, a_l) < \infty$$

  $\Phi_l^{(N)}(r; a_1, \ldots, a_l) = P_{\text{gen -r,}\sum_i a_i\text{ind.}}((a_i \text{ ind. common parent})_i)$. If it has the form $q_{a_1,\ldots,a_l} t$, the limit is a time-homogeneous Markov process with rates q.

For the modified Dirac coalescents, one gets the only non-zero limit $\lim_{N\to\infty} \sum_{r=1}^{\mathcal{G}_N^{-1}(t)} \Phi_l^{(N)}(r; a_1, \ldots, a_l) = \Psi^{k-2} t = \int x^{k-2}(1-x)^0 \delta_\Psi(dx)$ for $l = 1$ and $a_1 \geq 2$. In the end scale with c_N, $\lim_{N\to\infty} c_N \mathcal{G}_N^{-1}(t) = \mathcal{G}(t)$

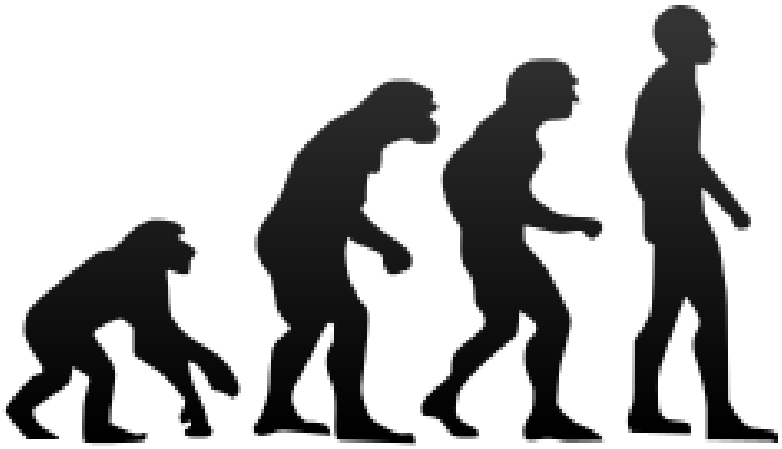
Works for other modified Moran models, too

w. S. Matuszewski, M. Lapierre, E. Kerdoncuff, A. Lambert, J. Jensen, G. Achaz, in prep.

- [HM13]: The fixed-size modified Moran model w. $P(U_N = j) = \binom{N}{j} \frac{B(j-\alpha,\alpha+N-j)}{B(2-\alpha,\alpha)}$, $j \geq 2$, has a Beta$(2-\alpha,\alpha)$-n-coalescent genealogy for $N \to \infty$
- Analogously, adding exponential growth ($N_k \sim N(1 - \frac{\rho}{N^\alpha})^k$) leads to a time-changed Beta coalescent genealogy, scaling with c_N^{-1} from the fixed N model
- [Sch03] (Fixed N): Each individual produces independent X_i offspring (heavy-tailed, $P(X_i \geq k) \sim Ck^{-\alpha}$). Next gen. randomly sampled from these.
- Adding exponential growth $N_k \sim N\left(1 - \frac{\rho}{N^{\alpha-1}}\right)^r$ again leads to a time-changed Beta$(2-\alpha,\alpha)$-n-coalescent $(\Pi_{\mathcal{G}_t})_{t\geq 0}$ with $\mathcal{G}_t = c^{-1}(e^{ct} - 1)$ for a constant c

Genealogy models $\leftrightarrow$ species

Species with non-skewed offspring distribution: Mammals, most plants,...



pictures from Wikimedia commons (users Tkgd2007,MrFrosty2)

Should lead to a bifurcating genealogy of (short) neutral loci ($\mathbb{KM}$,+growth,+pop. struct.). *Do they?*

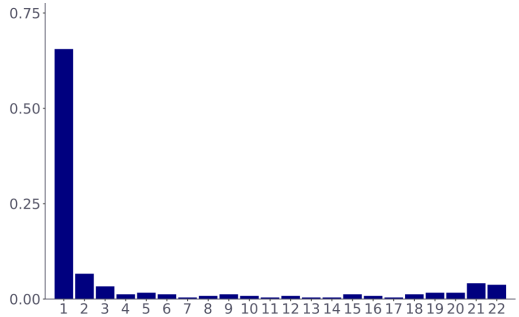
Genealogy models $\leftrightarrow$ species

Reproduction sweepstakes ($\mathbb{BETA}$):

Japanese sardine [NNY16]

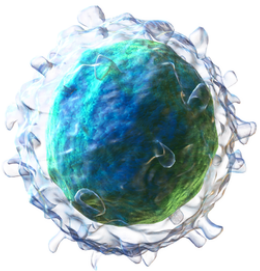
Atlantic cod [ÁH14]

with exp. growth:

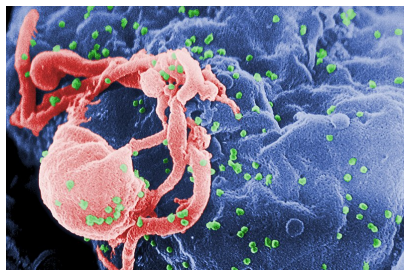


copy numbers in cancer cells [KVS$^+$17]

Candidate populations for rapid selection ($\Lambda = U_{[0,1]}$):



B-cells under HIV [NOŁ$^+$18]



HIV [ZBT$^+$15]

SeaFIC; H. Hillewaert; Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2); CDC/ C. Goldsmith, P. Feorino, E. L. Palmer, W. R. McManus

How to distinguish genealogy models

Usually, we would be interested in inferring the true (best-fitting) genealogy model, while we treat θ as a nuisance parameter

Inference approaches

- (Nearly) Full likelihood on SNP data (MCMC-based for moderate n) [SBB13],[BG00],[KJS15]...
  Slow, does not scale well w. large data sets
- Based on the site-frequency spectrum (SFS)
  $\xi_i^{(n)} :=$ # SNPs with derived allele frequency $\frac{i}{n}$, $i \in [n-1]$
  Quicker, but needs approximations and/or simulations

SFS-based inference between n-coalescents

Pseudo-likelihood approach [EBBF15], $s = \sum_i \eta_i$

Base approximate LRT on

$PsL(\xi_i^{(n)} = k_i, i \in [n-1])$

$= \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^{n} \xi_j^{(n)})} \right)^{k_i}$

Assumptions:

fixed-s, $\frac{\xi_i^{(n)}}{\sum_{j=1}^{n} \xi_j^{(n)}} \approx \frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^{n} \xi_j^{(n)})}$

$E(\xi_i^{(n)})$ can be computed analytically [SKS16],[PK03]

Multiple loci: Add up SFS

SFS-based inference between n-coalescents

Pseudo-likelihood approach [EBBF15],$s = \sum_i \eta_i$

Base approximate LRT on

$PsL(\xi_i^{(n)} = k_i, i \in [n-1])$

$= \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^n \xi_j^{(n)})} \right)^{k_i}$

Assumptions:

fixed-s, $\frac{\xi_i^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}} \approx \frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^n \xi_j^{(n)})}$

$E(\xi_i^{(n)})$ can be computed analytically [SKS16],[PK03]

Multiple loci: Add up SFS

Monte Carlo Likelihood approach [Kos17]

Perform approximate LRT for

$$\left(\frac{\xi_1^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}}, \frac{\sum_{i=k}^{n-1} \xi_i^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}} \right),$$

estimating the two-dimensional kernel density via simulation

Multiple loci: Use average over loci for LRT, estimate density of average

For distinguishing coalescent models, both methods are very robust if models tested misspecify θ

SFS-based inference between n-coalescents

Pseudo-likelihood approach [EBBF15],$s = \sum_i \eta_i$

Base approximate LRT on

$PsL(\xi_i^{(n)} = k_i, i \in [n-1])$

$= \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^n \xi_j^{(n)})} \right)^{k_i}$

Assumptions:

fixed-s, $\frac{\xi_i^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}} \approx \frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^n \xi_j^{(n)})}$

$E(\xi_i^{(n)})$ can be computed analytically [SKS16],[PK03]

Multiple loci: Add up SFS

Monte Carlo Likelihood approach [Kos17]

Perform approximate LRT for

$$\left(\frac{\xi_1^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}}, \frac{\sum_{i=k}^{n-1} \xi_i^{(n)}}{\sum_{j=1}^n \xi_j^{(n)}} \right),$$

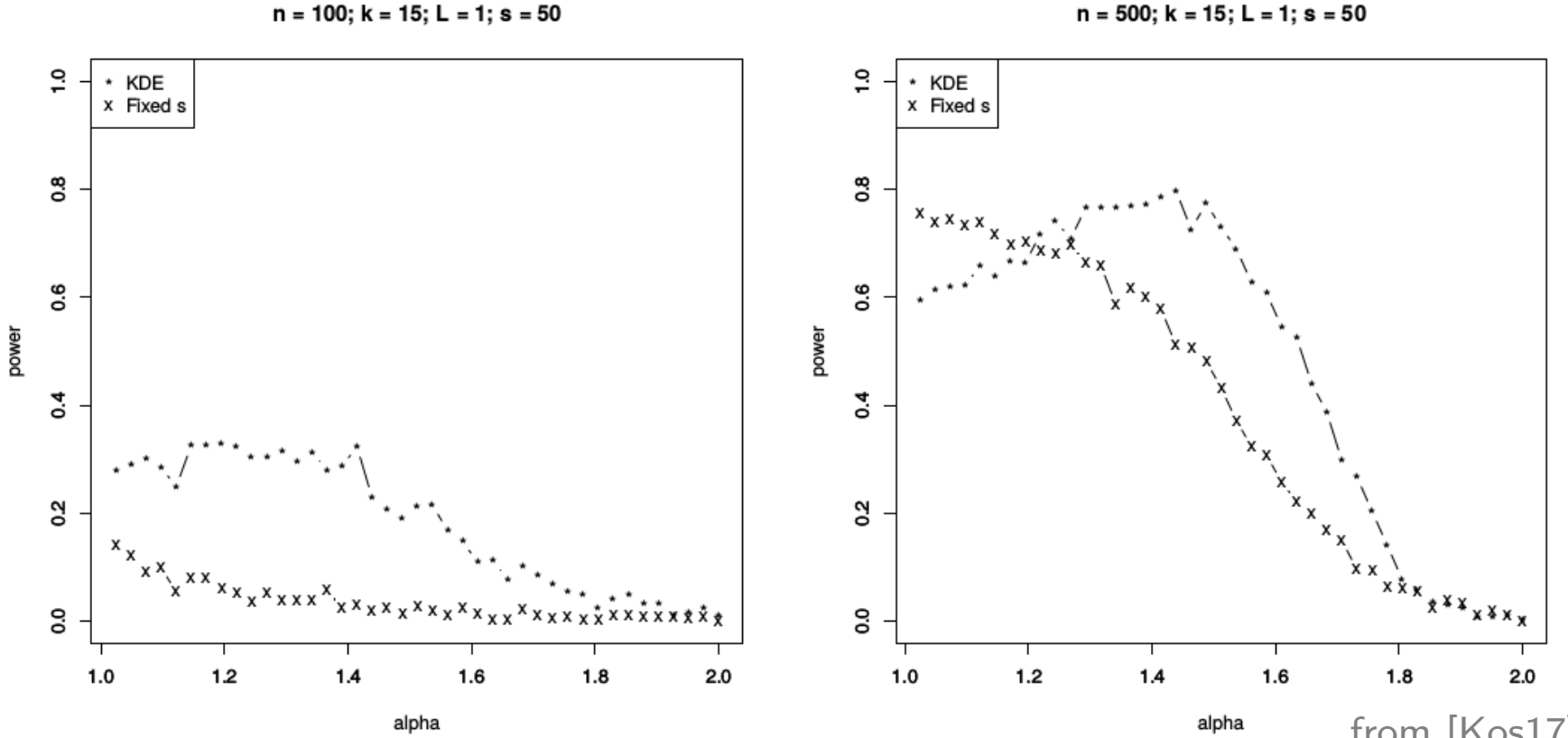
estimating the two-dimensional kernel density via simulation

Multiple loci: Use average over loci for LRT, estimate density of average

For distinguishing coalescent models, both methods are very robust if models tested misspecify θ

[Kos17] Using KDE is superior, more loci is better

H_0 : $\mathbb{KM}$ w. exp. growth, $g \in [0, 1000]$, $+.1$ near 0, $+10$ for $g \geq 40$
H_1 : Beta$(2-\alpha,\alpha)$-n-coalescent, $\alpha \in \{1, 1.025, \ldots, 2\}$

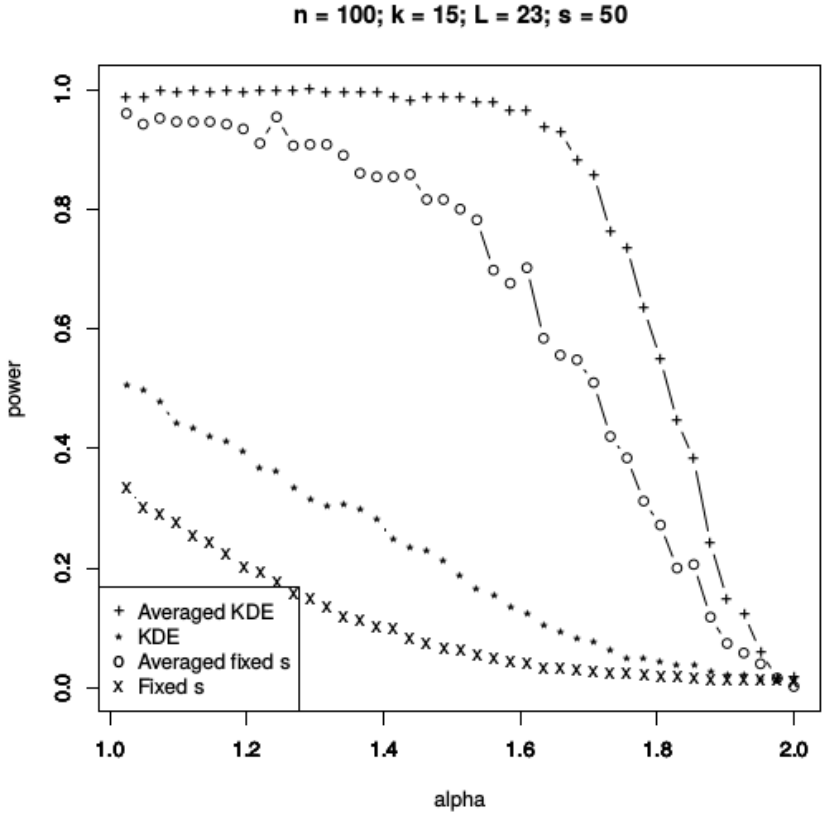


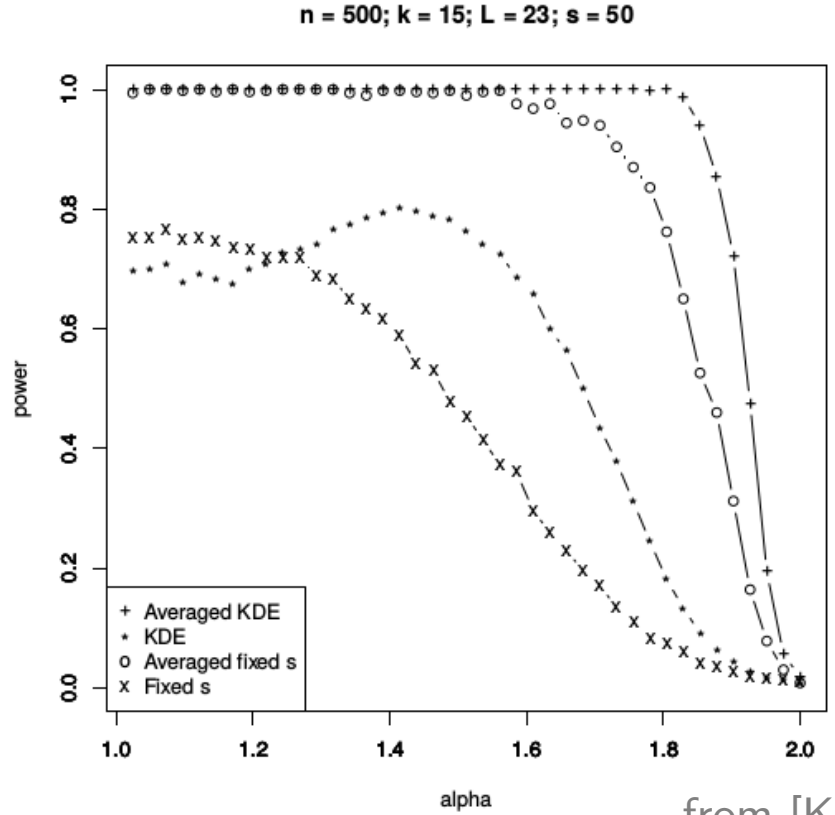
from [Kos17]

Freely recomb. loci in $\mathbb{MMC}$ are not ind. (though unlinked) [BBE13].
[Kos17]'s coalescent model accounts for this (PP-construction of Λ-n-coal.:
Given PPP, each locus uses it independently to construct its tree)

[Kos17] Using KDE is superior, more loci is better

H_0 : $\mathbb{KM}$ w. exp. growth, $g \in [0, 1000]$, $+.1$ near 0, $+10$ for $g \geq 40$
H_1 : Beta$(2-\alpha,\alpha)$-n-coalescent, $\alpha \in \{1, 1.025, \ldots, 2\}$



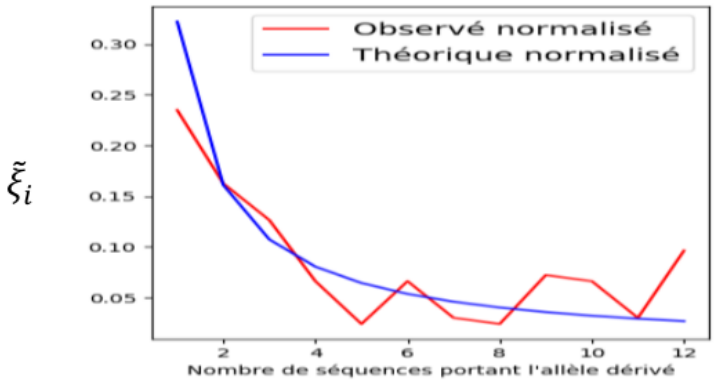


from [Kos17]

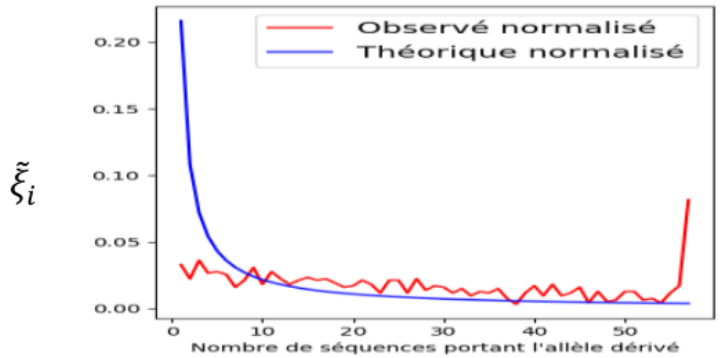
Freely recomb. loci in $\mathbb{MMC}$ are not ind. (though unlinked) [BBE13].
[Kos17]'s coalescent model accounts for this (PP-construction of Λ-n-coal.: Given PPP, each locus uses it independently to construct its tree)

U-shaped SFS vs. genealogy models

- Collection of U/J-shaped SFS from G. Achaz' group in a diverse set of species: pill-bug, *E. coli*, *Helicobacter pylori*, cuttlefish, *A. thaliana*, shark, *Drosophila melanogaster*, humans... ($18 \leq n \leq 1214$)



Caenorhabditis elegans



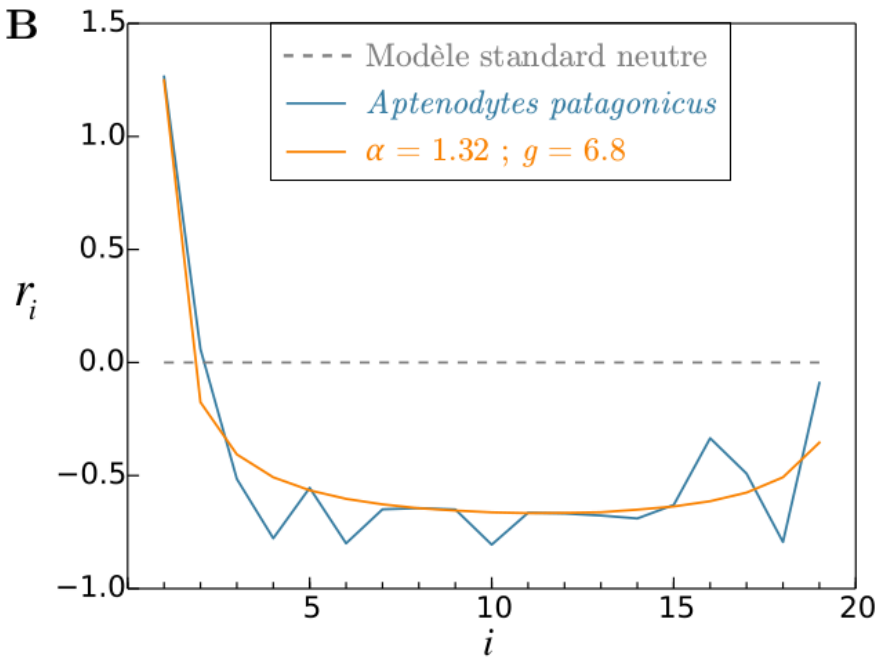
Glyphis garricki

(Graphs by R. Clodion)

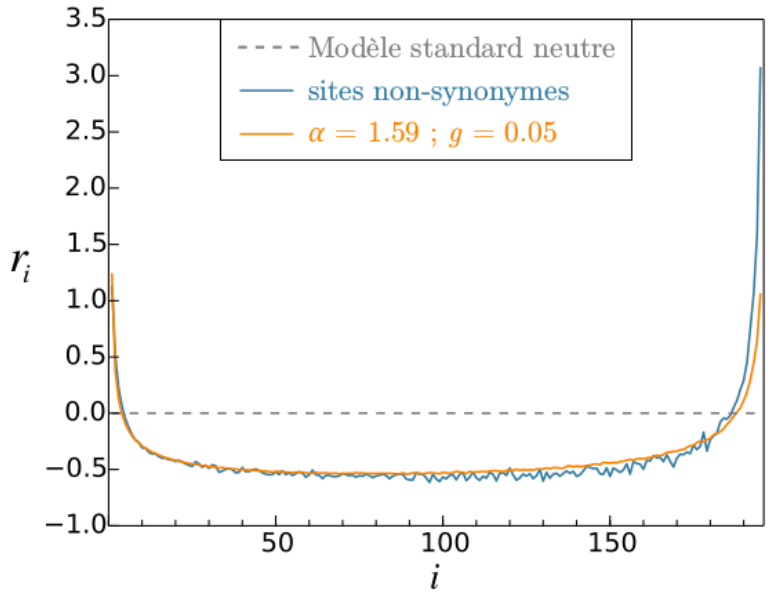
- Plots of $E(nSFS)$: expected scaled SFS $\tilde{\xi}_i := \frac{\xi_i^n}{\sum_j \xi_j^{(n)}}$
- [Lap17] Possible contributions to U/J-shape: Confusing derived and ancestral alleles (misidentification, *MI*), selection, demography, biased gene conversion, multiple-merger,...

E(nSFS) corrected for *MI*, fit to $\mathbb{BETA}$+exp. g.

- [Lap17] Correct $E(nSFS)$ for *MI*: Via outgroup, estimate *MI* prob. x, use $((1-\hat{x})\tilde{\xi}_i - \hat{x}\tilde{\xi}_{n-i})/(1-2\hat{x})$
- [Lap17], R. Clodion, E. Kerdoncuff: Multiple-merger coalescents (with exp. growth/decline) can match (*MI*-corrected) E(nSFS) rather well



D. melanogaster



$$r_i = (x_i - E_{\mathbb{KM}}(\xi_i^{(n)}))/E_{\mathbb{KM}}(\xi_i^{(n)}),$$

minimize $\sum_i \frac{(E_{\mathbb{B}exp}(\tilde{\xi}_i^{(n)}) - E_{\mathbb{KM}}(\tilde{\xi}_i^{(n)}))^2}{E_{\mathbb{KM}}(\tilde{\xi}_i^{(n)})}$

from [Lap17]

E(nSFS) corrected for *MI*, fit to $\mathbb{BETA}$+exp. g.

- [Lap17] Correct $E(nSFS)$ for *MI*: Via outgroup, estimate *MI* prob. x, use $((1-\hat{x})\tilde{\xi}_i - \hat{x}\tilde{\xi}_{n-i})/(1-2\hat{x})$
- [Lap17], R. Clodion, E. Kerdoncuff: Multiple-merger coalescents (with exp. growth/decline) can match (*MI*-corrected) E(nSFS) rather well

- Is the fit close enough? How close is an observed genome-wide SFS to the E(SFS) of the true model (# loci and dependence)? To better asses model fit: Estimate likelihood surface vs. PsL or MCL approach

E(nSFS) corrected for *MI*, fit to $\mathbb{BETA}$+exp. g.

- [Lap17] Correct $E(nSFS)$ for MI: Via outgroup, estimate MI prob. x, use $((1-\hat{x})\tilde{\xi}_i - \hat{x}\tilde{\xi}_{n-i})/(1-2\hat{x})$
- [Lap17], R. Clodion, E. Kerdoncuff: Multiple-merger coalescents (with exp. growth/decline) can match (MI-corrected) E(nSFS) rather well

- Is the fit close enough? How close is an observed genome-wide SFS to the E(SFS) of the true model (# loci and dependence)? To better asses model fit: Estimate likelihood surface vs. PsL or MCL approach
- Can be easily adjusted to assess misidentification:

$$PsL((\xi_i^{(n)} = k_i)_i) = \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(\xi_i^{(n)})}{E(\sum_{j=1}^{n} \xi_j^{(n)})} \right)^{k_i}$$

E(nSFS) corrected for *MI*, fit to $\mathbb{BETA}$+exp. g.

- [Lap17] Correct $E(nSFS)$ for *MI*: Via outgroup, estimate *MI* prob. x, use $((1-\hat{x})\tilde{\xi}_i - \hat{x}\tilde{\xi}_{n-i})/(1-2\hat{x})$
- [Lap17], R. Clodion, E. Kerdoncuff: Multiple-merger coalescents (with exp. growth/decline) can match (*MI*-corrected) E(nSFS) rather well

- Is the fit close enough? How close is an observed genome-wide SFS to the E(SFS) of the true model (# loci and dependence)? To better asses model fit: Estimate likelihood surface vs. PsL or MCL approach
- Can be easily adjusted to assess misidentification:

$$PsL((\xi_i^{(n)} = k_i)_i) = \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{(1-x)E(\xi_i^{(n)}) + xE(\xi_{n-i}^{(n)})}{E(\sum_{j=1}^{n} \xi_j^{(n)})} \right)^{k_i}$$

E(nSFS) corrected for *MI*, fit to $\mathbb{BETA}$+exp. g.

- [Lap17] Correct $E(nSFS)$ for *MI*: Via outgroup, estimate *MI* prob. x, use $((1-\hat{x})\tilde{\xi}_i - \hat{x}\tilde{\xi}_{n-i})/(1-2\hat{x})$
- [Lap17], R. Clodion, E. Kerdoncuff: Multiple-merger coalescents (with exp. growth/decline) can match (*MI*-corrected) E(nSFS) rather well

- Is the fit close enough? How close is an observed genome-wide SFS to the E(SFS) of the true model (# loci and dependence)? To better asses model fit: Estimate likelihood surface vs. PsL or MCL approach
- Can be easily adjusted to assess misidentification:

$$PsL((\xi_i^{(n)} = k_i)_i) = \frac{s!}{k_1! \cdots k_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{(1-x)E(\xi_i^{(n)}) + xE(\xi_{n-i}^{(n)})}{E(\sum_{j=1}^{n} \xi_j^{(n)})} \right)^{k_i}$$

w. S. Matuszewski, M. Lapierre, E. Kerdoncuff, A. Lambert, J. Jensen, G. Achaz

Inference using few loci: Use more statistics?

w. A. Siri-Jégousse

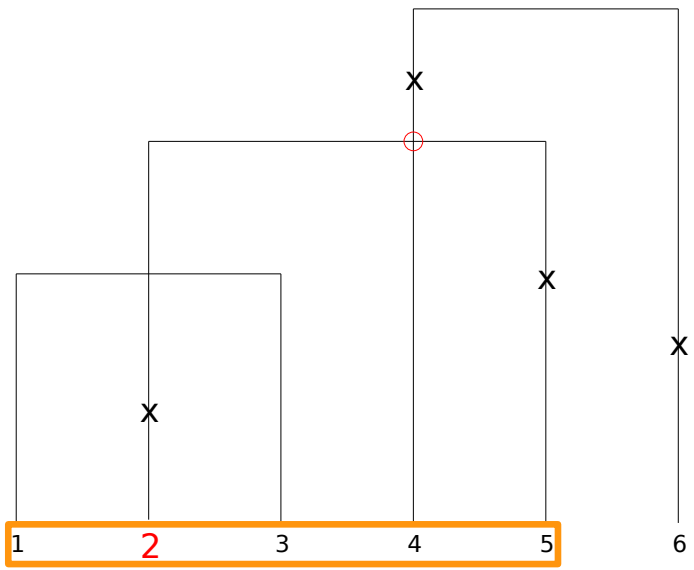
- SFS-based inference for few (1) loci is very noisy
- [KVS+17] report very low misclassification probabilities in an ABC approach for (essentially) $\mathbb{BETA}$+exp. growth vs. $\mathbb{KM}$+ exp. growth if additional statistics are used
  - Number of segregating sites
  - Quantiles .1,.3,.5,.7,.9 of allele frequencies
  - Quantiles .1,.3,.5,.7,.9 of pairwise Hamming distances (mismatch count)
  - Quantiles .1,.3,.5,.7,.9 of LD measured as squared correlation r^2 between SNP allele frequencies
  - Quantiles .1,.3,.5,.7,.9 of total branch length of reconstructed phylogeny (e.g. neighbor-joining tree)

Inference using few loci: Use more statistics?

w. A. Siri-Jégousse

- SFS-based inference for few (1) loci is very noisy
- [KVS+17] report very low misclassification probabilities in an ABC approach for (essentially) $\mathbb{BETA}$+exp. growth vs. $\mathbb{KM}$+ exp. growth if additional statistics are used
  - Number of segregating sites
  - Quantiles .1,.3,.5,.7,.9 of allele frequencies
  - Quantiles .1,.3,.5,.7,.9 of pairwise Hamming distances (mismatch count)
  - Quantiles .1,.3,.5,.7,.9 of LD measured as squared correlation r^2 between SNP allele frequencies
  - Quantiles .1,.3,.5,.7,.9 of total branch length of reconstructed phylogeny (e.g. neighbor-joining tree)
- Partly, low misclassification rates stem from models using identitical θ ranges (which can lead to stark differences in # segregating sites)
- If we compare models with comparable # segregating sites, do we see the same effect? Which statistics help to distinguish?

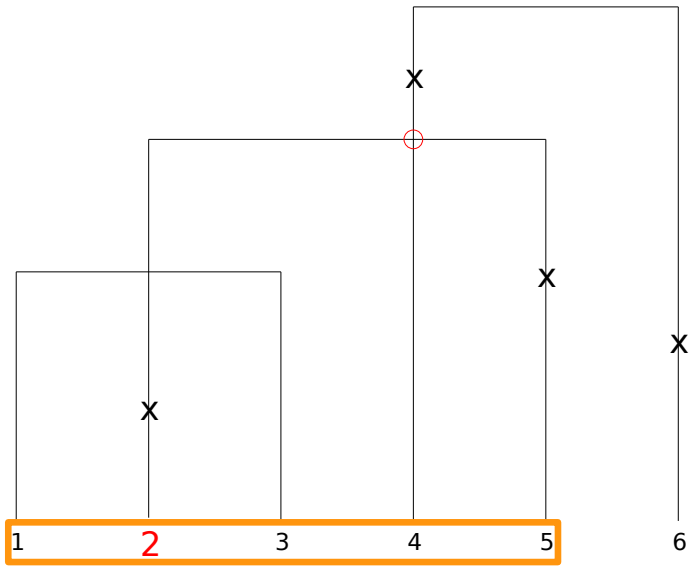
We also add a further statistic



$O_n(2)$=5

- $O_n(i) :=\#$ individuals sharing all non-private mutations of i
  Smallest family of i which can be genetically distinguished

We also add a further statistic

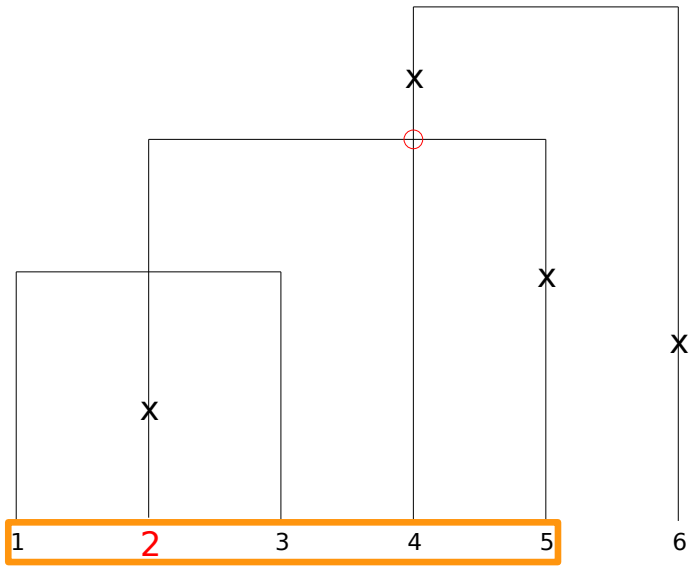


$O_n(2)$=5

- O_n observable from sequence data if ancestral base calls are known

- $O_n(i) :=$ # individuals sharing all non-private mutations of i
  Smallest family of i which can be genetically distinguished
- $\Leftrightarrow$ # descendants of the youngest ancestor of i with a mutation on the branch above it

We also add a further statistic



$O_n(2)$=5

- O_n observable from sequence data if ancestral base calls are known

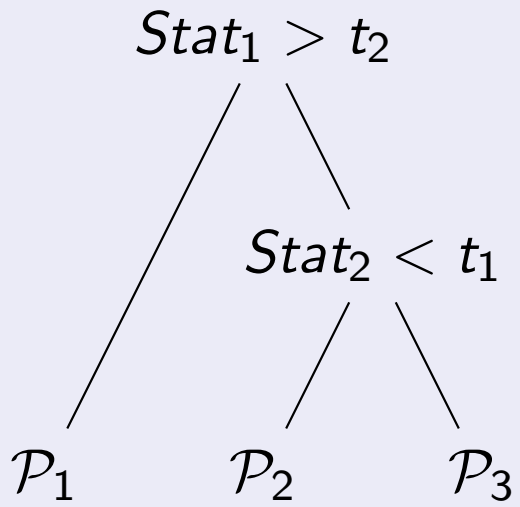
- $O_n(i)$:=# individuals sharing all non-private mutations of i
  Smallest family of i which can be genetically distinguished
- $\Leftrightarrow$ # descendants of the youngest ancestor of i with a mutation on the branch above it
- Use quantiles .1,.3,.5,.7,.9, the harmonic mean, sample mean and s.d.

Model selection via ABC with random forests

Using many test statistics: Curse of dimension, added noise

Random forest-based ABC [PME+15]

1) Build decision trees (CART) using bootstrap samples of simulated stats S (w. prior) to sort the latter into bins $\mathcal{P}_i$ from the same model.
2) For each tree, sort S_{obs} to $\mathcal{P}_i$



Randomised CART

At node, take stat from random subset w. minimal misclassification (Gini index)

- Misclassification measure: Out-of-the-bag error
- Model selection a) % trees: $S_{obs} \rightarrow$ model M b) Posterior probability
- Importance of stat S_i: Decrease in misclassification by all nodes of S_i, averaged over RF

Which statistics distinguish genealogy models?

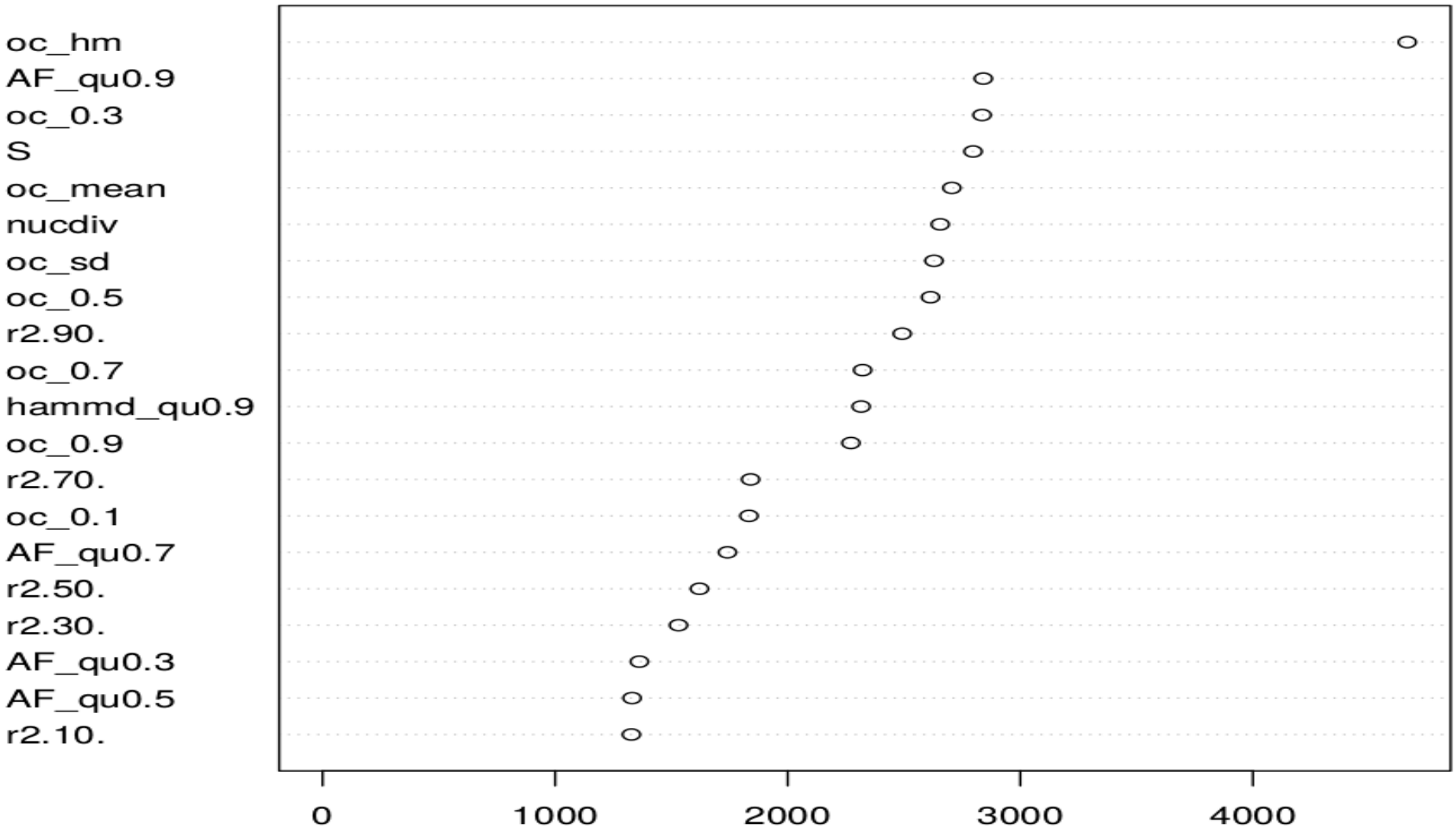
$M1 : \mathbb{KM}$ + exp. growth, $g \in \{0, .5, 1, 2.5, 4, 7, 10, 25, 50, 75, 100, 500, 1000\}$
$M2 : \mathbb{BETA}$, $\alpha \in \{1, 1.1, \ldots, 2\}$

$n = 100$, $\theta = 2s/E$(total coalescent length) for $s \in \{15, 20, 30, 40, 60, 75\}$, 175K sims/model (1x replicated), flat prior

Statistics: O_n, allele frequencies (SFS,fSFS), Hamming distances, r^2, phylogenetic branch lengths, nucleotide diversity π, # mutations S

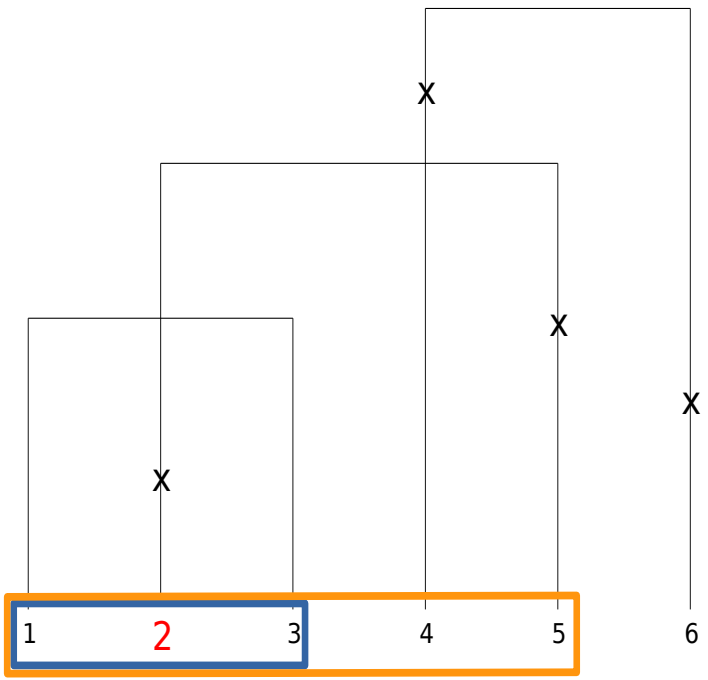
Stats	% $\mathbb{BETA}$ misclassified	% $\mathbb{KM}$+growth misclassified
All	16.9/16.8 %	23.2/23.3 %
No O_n	18.2/18.2 %	26.2/26.2 %
No r^2, phylo	17.2/17 %	23.3/23.5 %
AF, π, S	21.9/22.1 %	33.9/34.1 %
SFS, π, S	19.1/19.2 %	30.7/30.4 %
+ O_n, Hamming	17.7/17.6 %	23.8/23.6 %
fSFS, π, S	23.2/23.2 %	33/33.2 %
+ Hamming	19.8/20 %	27.6/27.5 %
+ r^2, phylo	19/19 %	26.6/26.7 %

Importance of statistics (full set)



measured by average decrease in Gini index over nodes of the statistic in the trees of the RF

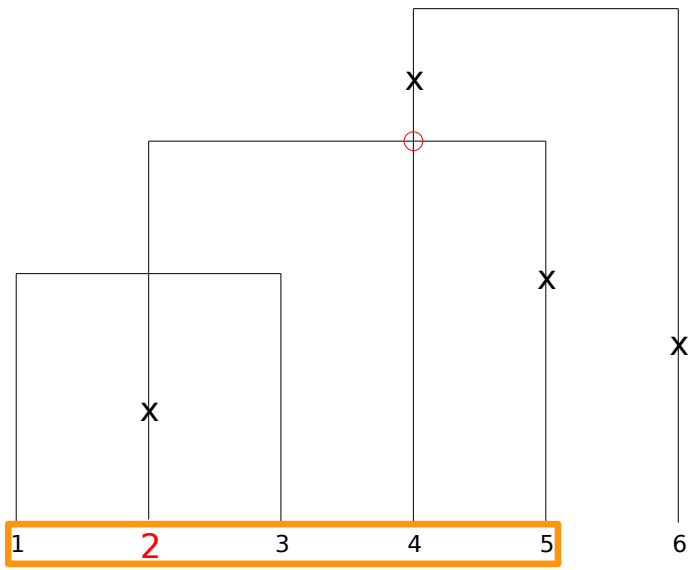
Why is the harmonic mean of $(O_n(i))_{i\in[n]}$ distinguishing well?



- $M_n(i)$: smallest family of i, # descendants of the most recent ancestor of i
- $M_n(i) \leq O_n(i)$, equality for $\theta \to \infty$
- $M_n(i)$ tends to be bigger for MMC than for KM [BF05], [FSJ14], [SJY16]
- $M_n(i)$'s law not changed if we make a time-change (to model pop.size changes)

Mathematical properties of O_n

n-coalescents: Processes in the partitions of $\{1, ..., n\}$
At time t: partition blocks = offspring of ancestral lines at t

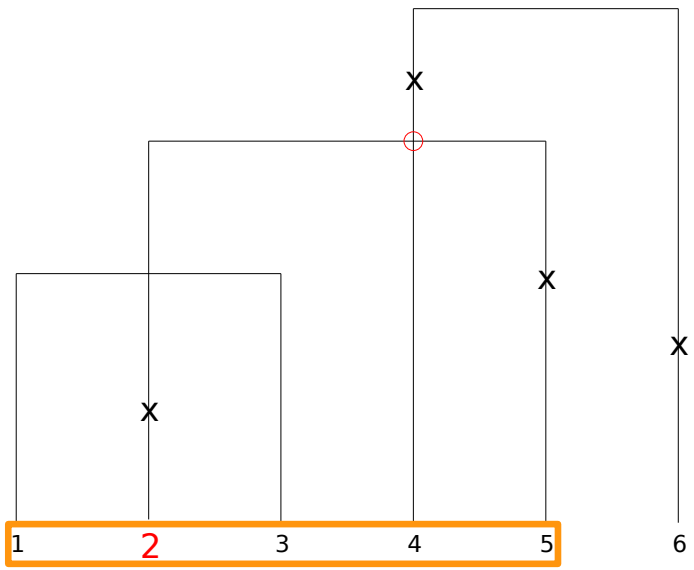


$O_n(2)$=5

- $O_n(i) := B_i^{(n)}(E_n(i) + T_n(i))$, where
  - $B_i^{(n)}(t)$ is the size of the block containing i at time t
  - $E_n(i)$ is the waiting time for the first merger of $\{i\}$
  - $T_n(i)$ is the waiting time for the first mutation affecting i after the first merger
  - $T_n(i)$ is independent of $B_i^{(n)}$, $E_n(i)$

Mathematical properties of O_n

n-coalescents: Processes in the partitions of $\{1, ..., n\}$
At time t: partition blocks = offspring of ancestral lines at t

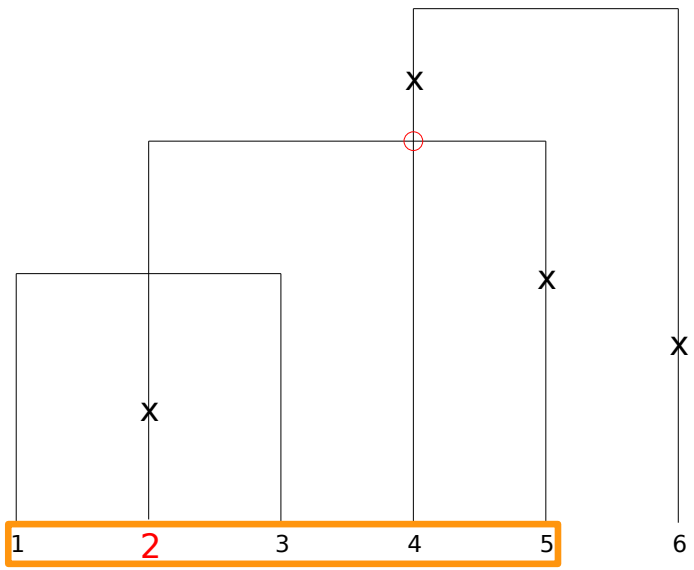


$O_n(2)$=5

- $O_n(i) := B_i^{(n)}(E_n(i) + T_n(i))$, where
  - $B_i^{(n)}(t)$ is the size of the block containing i at time t
  - $E_n(i)$ is the waiting time for the first merger of $\{i\}$
  - $T_n(i)$ is the waiting time for the first mutation affecting i after the first merger
  - $T_n(i)$ is independent of $B_i^{(n)}$, $E_n(i)$
- Exchangeability: $O_n(1) \stackrel{d}{=} O_n(i)$

Mathematical properties of O_n

n-coalescents: Processes in the partitions of $\{1, ..., n\}$
At time t: partition blocks = offspring of ancestral lines at t



$O_n(2)$=5

- $O_n(i) := B_i^{(n)}(E_n(i) + T_n(i))$, where
  - $B_i^{(n)}(t)$ is the size of the block containing i at time t
  - $E_n(i)$ is the waiting time for the first merger of $\{i\}$
  - $T_n(i)$ is the waiting time for the first mutation affecting i after the first merger
  - $T_n(i)$ is independent of $B_i^{(n)}$, $E_n(i)$
- Exchangeability: $O_n(1) \stackrel{d}{=} O_n(i)$
- All moments of $O_n(1)$ can be computed recursively for any Λ-n-coalescent

Asymptotics of $O_n(i)$ for $\mathbb{BETA}$, $n \to \infty$

$O_n(1) := B_1^{(n)}(E_n(1) + T_n(1))$

- $B_1^{(n)}(t)$: size of block of 1 at time t
- $E_n(1)$: waiting time for first merger of $\{1\}$
- $T_n(1)$: waiting time for first mutation affecting 1 after $E_n(1)$
- $T_n(1)$: independent of $B_1^{(n)}$, $E_n(1)$

$E_n(1) \xrightarrow{d} Exp(\mu_{-1})$, $n \to \infty$

$T_n(1) \stackrel{d}{=} Exp(\theta/2)$

$\mu_{-1} = \infty$, dust-free for Λ-coalescents

- $f_1(t) := \lim_{n\to\infty} n^{-1} B_1(t)$
- $\lim_{n\to\infty} n^{-1} O_n(1) = f_1(T(1))$

Asymptotics of $O_n(i)$ for $\mathbb{BETA}$, $n \to \infty$

$O_n(1) := B_1^{(n)}(E_n(1) + T_n(1))$

- $B_1^{(n)}(t)$: size of block of 1 at time t
- $E_n(1)$: waiting time for first merger of $\{1\}$
- $T_n(1)$: waiting time for first mutation affecting 1 after $E_n(1)$
- $T_n(1)$: independent of $B_1^{(n)}$, $E_n(1)$

$E_n(1) \xrightarrow{d} Exp(\mu_{-1})$, $n \to \infty$

$T_n(1) \stackrel{d}{=} Exp(\theta/2)$

$\mu_{-1} = \infty$, dust-free for Λ-coalescents

- $f_1(t) := \lim_{n\to\infty} n^{-1} B_1(t)$
- $\lim_{n\to\infty} n^{-1} O_n(1) = f_1(T(1))$
- All moments of $f_1(t)$ known from [Pit99, Prop 29]
- $E[(f_1(T(1)))^k] = 1 - \sum_{r=2}^{k+1} a_{k,r} \frac{\theta/2}{\lambda_r + \theta/2}$, where λ_r is the total rate of the Λ-coalescent in a state with r blocks and $a_{k,r}$ is a rational function of $\lambda_2, \ldots, \lambda_k$.
- $E[f_1(T(1))] = \frac{\Lambda([0,1])}{\Lambda([0,1]) + \theta/2}$

Asymptotics of $O_n(i)$ for $\mathbb{BETA}$, $n \to \infty$

$O_n(1) := B_1^{(n)}(E_n(1) + T_n(1))$

- $B_1^{(n)}(t)$: size of block of 1 at time t
- $E_n(1)$: waiting time for first merger of $\{1\}$
- $T_n(1)$: waiting time for first mutation affecting 1 after $E_n(1)$
- $T_n(1)$: independent of $B_1^{(n)}$, $E_n(1)$

$E_n(1) \xrightarrow{d} Exp(\mu_{-1})$, $n \to \infty$

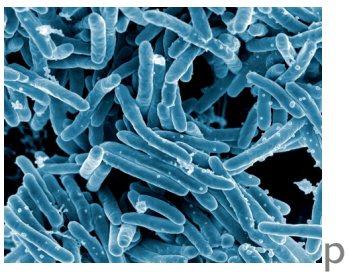
$T_n(1) \stackrel{d}{=} Exp(\theta/2)$

$\mu_{-1} = \infty$, dust-free for Λ-coalescents

- $f_1(t) := \lim_{n\to\infty} n^{-1} B_1(t)$
- $\lim_{n\to\infty} n^{-1} O_n(1) = f_1(T(1))$
- All moments of $f_1(t)$ known from [Pit99, Prop 29]
- $E[(f_1(T(1)))^k] = 1 - \sum_{r=2}^{k+1} a_{k,r} \frac{\theta/2}{\lambda_r + \theta/2}$, where λ_r is the total rate of the Λ-coalescent in a state with r blocks and $a_{k,r}$ is a rational function of $\lambda_2, \ldots, \lambda_k$.
- $E[f_1(T(1))] = \frac{\Lambda([0,1])}{\Lambda([0,1]) + \theta/2}$
- $\mathbb{BSZ}$ $f_1(T(1)) \stackrel{d}{=} Beta\left(\frac{1}{1+\frac{\theta}{2}}, \frac{\frac{\theta}{2}}{1+\frac{\theta}{2}}\right)$

Mycobacterium tuberculosis - multiple-merger genealogy?

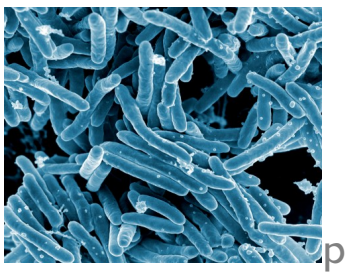
- Bacterial agent of tuberculosis, haploid



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

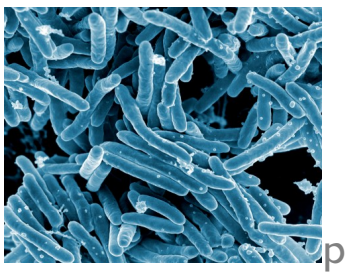
- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

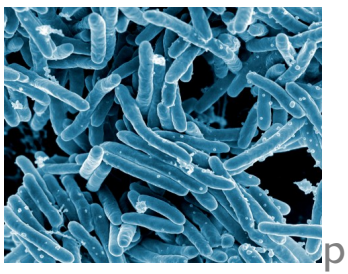
- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus
- Data sequenced with high coverage, mutations can be well-polarized (phylogeny very clear $\Rightarrow$ practically no misorientation [Lap17])



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

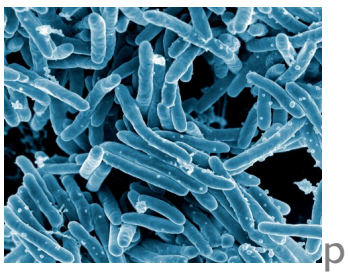
- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus
- Data sequenced with high coverage, mutations can be well-polarized (phylogeny very clear $\Rightarrow$ practically no misorientation [Lap17])
- Could be under strong selection pressure $\Rightarrow$ $\mathbb{BSZ}$ genealogy ($\mathbb{BSZ}+1$ also tested, fits clearly worse that other models)



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

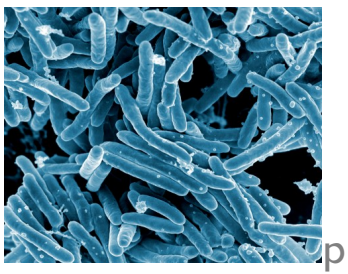
- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus
- Data sequenced with high coverage, mutations can be well-polarized (phylogeny very clear $\Rightarrow$ practically no misorientation [Lap17])
- Could be under strong selection pressure $\Rightarrow \mathbb{BSZ}$ genealogy ($\mathbb{BSZ}+1$ also tested, fits clearly worse that other models)
- Reproduction also potentially skewed* $\Rightarrow \mathbb{BETA}$, other multiple-merger coalescents



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

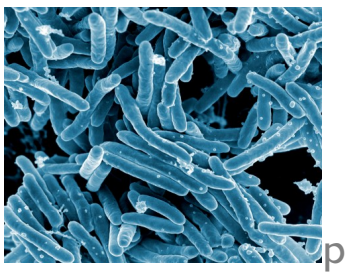
- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus
- Data sequenced with high coverage, mutations can be well-polarized (phylogeny very clear $\Rightarrow$ practically no misorientation [Lap17])
- Could be under strong selection pressure $\Rightarrow \mathbb{BSZ}$ genealogy ($\mathbb{BSZ}+1$ also tested, fits clearly worse that other models)
- Reproduction also potentially skewed* $\Rightarrow \mathbb{BETA}$, other multiple-merger coalescents
- We use data sets from outbreaks and local samples to control population structure



pic from NIAID

Mycobacterium tuberculosis - multiple-merger genealogy?

- Bacterial agent of tuberculosis, haploid
- Clonal reproduction (gen.$\approx$ 15-20 h), rather small 4 Mb genome, very few recombination events $\Rightarrow$ treated as a single locus
- Data sequenced with high coverage, mutations can be well-polarized (phylogeny very clear $\Rightarrow$ practically no misorientation [Lap17])
- Could be under strong selection pressure $\Rightarrow \mathbb{BSZ}$ genealogy ($\mathbb{BSZ}+1$ also tested, fits clearly worse that other models)
- Reproduction also potentially skewed* $\Rightarrow \mathbb{BETA}$, other multiple-merger coalescents
- We use data sets from outbreaks and local samples to control population structure
- Genealogy model (usually) proposed in the literature: Kingman's n-coalescent with exponential growth



pic from NIAID

ABC with random forest approach w. F. Menardo, in prep.

Models

$\mathbb{KM}$ + exp. growth, $g \in \{0, \ldots, 5000\}, \{0, \ldots, 20000\}$
$\mathbb{BETA}$ $\alpha \in \{1, 1.025, \ldots, 1.975, 0.\}$
$\mathbb{DIRAC}$ $\Psi \in \{0.025, 0.05, \ldots, 0.975\}$

Setup

- $\theta \in [\hat{\theta}_w/5, 5\hat{\theta}_w]$
- Sequential ABC (2x) to fit growth range
- ABC w. RF can also be used for parameter estimation
- Many mutations, large samples: Misclassification $\leq 5\%$

PROBLEM: We treat the sequences as sampled at the same time!

ABC with random forest results w. F. Menardo, in prep.

sample	n	S_{obs}	best model (post. prob.)	.1/.9 quant. posterior g, α
(Inuit, '11,'13	147	454	$\mathbb{BETA}$ (1)	1.2/1.425
Hamburg, 99-'10	61	74	$\mathbb{BETA}$ (.96)	1.075/1.35
Argentinia 96-'09	248	497	$\mathbb{BETA}$ (.998)	1.1/1.3
subset '01-'05	137	312	$\mathbb{BETA}$ (.95)	1.125/1.35
subsubset '01-'03	91	205	$\mathbb{BETA}$ (.98)	1.075/1.375
Ethiopia '06-'10	21	1334	$\mathbb{BETA}$ (.78)	1.3/1.725
East Europe/Russia	176	1164	$\mathbb{KM} + exp$ (0.98)	(1535, 3629)

data from [LRP+15],[RDK+13],[EMR+15],[CHK+15],[SKM+17] ([SBJ+16], Uganda: not completely analysed, but subsets fit better to $\mathbb{KM} + exp$)

Different sampling times, but we assume an ultrametric tree

ABC with random forest results w. F. Menardo, in prep.

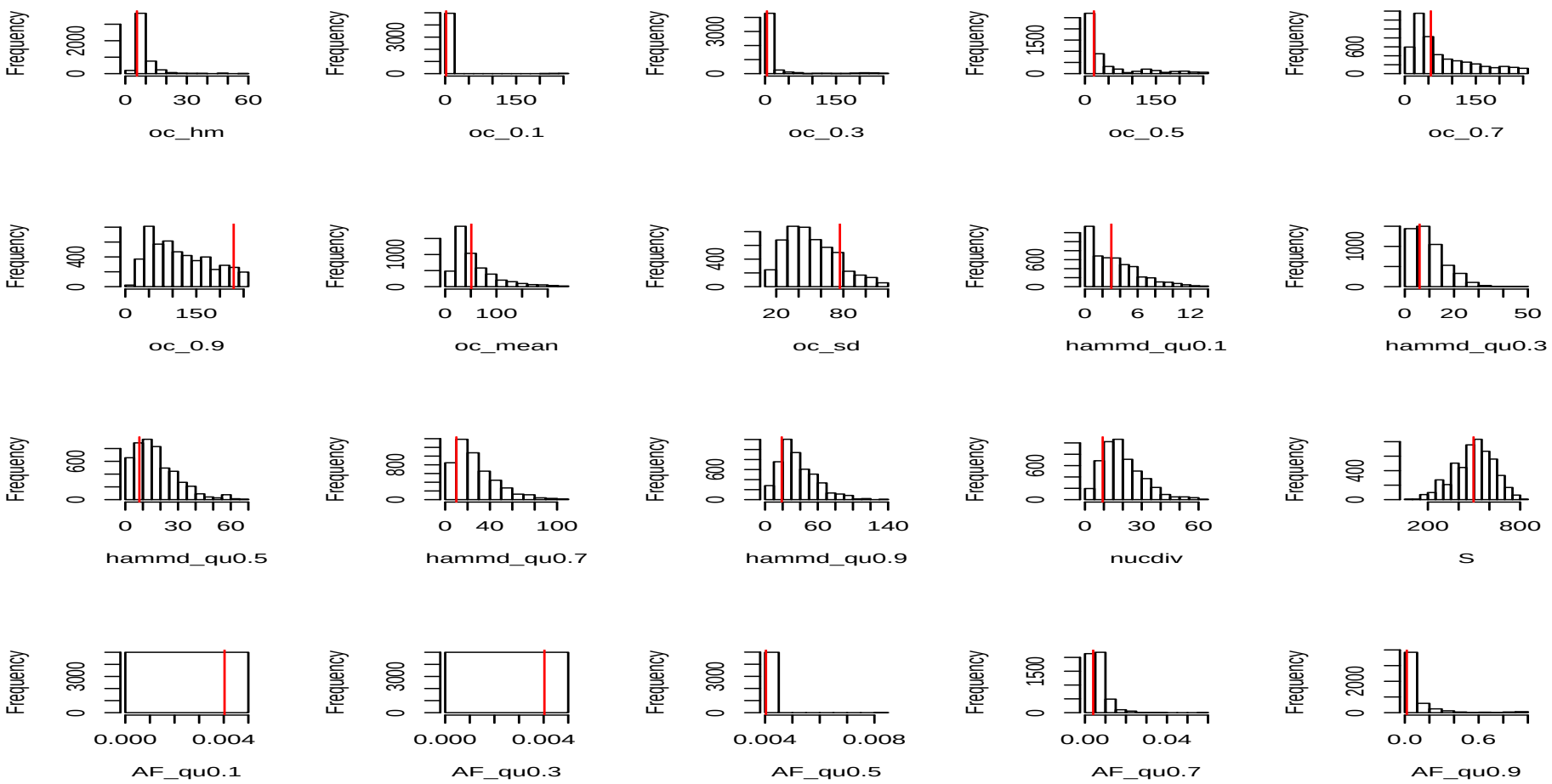
sample	n	S_{obs}	best model (post. prob.)	.1/.9 quant. posterior g, α
(Inuit, '11,'13	147	454	$\mathbb{BETA}$ (1)	1.025/1.25)
Hamburg, 99-'10	61	74	$\mathbb{BETA}$ (.98)	1/1.4
Argentinia 96-'09	248	497	$\mathbb{BETA}$ (.95)	1/1.225
subset '01-'05	137	312	$\mathbb{BETA}$ (.9)	1.025/1.3
subsubset '01-'03	91	205	$\mathbb{BETA}$ (.95)	1.025/1.3
Ethiopia '06-'10	21	1334	$\mathbb{BETA}$ (.77)	1.175/1.8
East Europe/Russia	176	1164	$\mathbb{KM} + exp$ (1)	(2536, 4867)

data from [LRP+15],[RDK+13],[EMR+15],[CHK+15],[SKM+17] ([SBJ+16], Uganda: not completely analysed, but subsets fit better to $\mathbb{KM} + exp$)

Different sampling times, but we assume an ultrametric tree

Method is rather robust $\Rightarrow$ Leave out all private mutations: $\leq 6\%$ misclassification, same results for classification

Posterior predictive checks



Multiple mergers vs. *Mycobacterium tuberculosis*

Still many questions!

- Magnitude of bias on non-singleton mutations by assuming equal sampling times?
- Account for different sampling times by modifying the coalescent trees as suggested in [HP18]?
- Better fitting model: $\mathbb{BETA}$+growth? Others?

Multiple mergers vs. *Mycobacterium tuberculosis*

Still many questions!

- Magnitude of bias on non-singleton mutations by assuming equal sampling times?
- Account for different sampling times by modifying the coalescent trees as suggested in [HP18]?
- Better fitting model: $\mathbb{BETA}$+growth? Others?
- For the majority (2/3) of data sets analysed, $\mathbb{BETA}$ reasonable alternative null model. What is a reasonable reproduction model underlying it?
- $\mathbb{BSZ}$: Some doubt (or noise), signal has to be clarified (more simulations, adjusted $\mathbb{BSZ}$?)
- Nearly all sets fit clearly to conceptually very different models: biological reasons?

Thanks for the attention!

Any questions?

Bibliography I

[ÁH14] Einar Árnason and Katrín Halldórsdóttir. Nucleotide variation and balancing selection at the ckma gene in atlantic cod: Analysis with multiple merger coalescent models. *PeerJ PrePrints*, 2, 2014.

[BBE13] Matthias Birkner, Jochen Blath, and Bjarki Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

[BBS13] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. The genealogy of branching brownian motion with absorption. *The Annals of Probability*, 41(2):527–618, 2013.

[BD13] Éric Brunet and Bernard Derrida. Genealogies in simple models of evolution. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01006, 2013.

Bibliography II

[BF05] Michael G. B. Blum and Olivier François. Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.*, 37(3):647–662, 06 2005.

[BG00] M Bahlo and RC Griffiths. Inference from gene trees in a subdivided population. *Theoretical population biology*, 57(2):79–95, 2000.

[CHK+15] Iñaki Comas, Elena Hailu, Teklu Kiros, Shiferaw Bekele, Wondale Mekonnen, Balako Gumi, Rea Tschopp, Gobena Ameni, R Glyn Hewinson, Brian D Robertson, et al. Population genomics of mycobacterium tuberculosis in ethiopia contradicts the virgin soil hypothesis for human tuberculosis in sub-saharan africa. *Current Biology*, 25(24):3260–3266, 2015.

[DWF13] Michael M Desai, Aleksandra M Walczak, and Daniel S Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

Bibliography III

[EBBF15] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[EMR+15] Vegard Eldholm, Johana Monteserin, Adrien Rieux, Beatriz Lopez, Benjamin Sobkowiak, Viviana Ritacco, and Francois Balloux. Four decades of transmission of a multidrug-resistant mycobacterium tuberculosis outbreak strain. *Nature communications*, 6:7119, 2015.

[EW06] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006.

[FSJ14] Fabian Freund and Arno Siri-Jégousse. Minimal clade size in the Bolthausen-Sznitman coalescent. *Journal of Applied Probability*, 51(3):657–668, 2014.

Bibliography IV

[HM13] Thierry Huillet and Martin Möhle. On the extended moran model and its relation to coalescents with multiple collisions. *Theoretical population biology*, 87:5–14, 2013.

[HP18] Patrick Hoscheit and Oliver G. Pybus. The multifurcating skyline plot. *submitted*, 2018.

[KJS15] Jere Koskela, Paul Jenkins, and Dario Spanò. Computational inference beyond kingman's coalescent. *Journal of Applied Probability*, 52(2), 2015.

[Kos17] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *arXiv preprint arXiv:1701.07787*, 2017.

Bibliography V

[KVS+17] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society Open Science*, 4(9), 2017.

[Lap17] Marguerite Lapierre. *Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2017.

[LRP+15] Robyn S. Lee, Nicolas Radomski, Jean-Francois Proulx, Ines Levade, B. Jesse Shapiro, Fiona McIntosh, Hafid Soualhine, Dick Menzies, and Marcel A. Behr. Population genomics of mycobacterium tuberculosis in the inuit. *Proceedings of the National Academy of Sciences*, 112(44):13609–13614, 2015.

Bibliography VI

[MHAJ17] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 2017.

[Möh02] M Möhle. The coalescent in population models with time-inhomogeneous environment. *Stochastic processes and their applications*, 97(2):199–227, 2002.

[NH13] Richard A Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, 2013.

[NNY16] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

Bibliography VII

[NOŁ+18] Armita Nourmohammad, Jakub Otwinowski, Marta Łuksza, Thierry Mora, and Aleksandra M Walczak. Clonal competition in b-cell repertoires during chronic hiv-1 infection. *arXiv preprint arXiv:1802.08841*, 2018.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

[PK03] A Polanski and M Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, 2003.

[PME+15] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable abc model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Bibliography VIII

[RDK+13] Andreas Roetzer, Roland Diel, Thomas A. Kohl, Christian Rckert, Ulrich Nbel, Jochen Blom, Thierry Wirth, Sebastian Jaenicke, Sieglinde Schuback, Sabine Rsch-Gerdes, Philip Supply, Jrn Kalinowski, and Stefan Niemann. Whole genome sequencing versus traditional genotyping for investigation of a mycobacterium tuberculosis outbreak: A longitudinal molecular epidemiological study. *PLOS Medicine*, 10(2):1–12, 02 2013.

[SBB13] Matthias Steinrücken, Matthias Birkner, and Jochen Blath. Analysis of dna sequence variation within marine species using beta-coalescents. *Theoretical population biology*, 87:15–24, 2013.

[SBJ+16] David Stucki, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, Liliana Rutaihwa, Sonia Borrell, Tao Luo, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics*, 48(12):1535, 2016.

Bibliography IX

[Sch03] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic processes and their applications*, 106(1):107–139, 2003.

[Sch17] Jason Schweinsberg. Rigorous results for a population model with selection ii: genealogy of the population. *Electronic Journal of Probability*, 22, 2017.

[SJY16] Arno Siri-Jégousse and Linglong Yuan. Asymptotics of the minimal clade size and related functionals of certain beta-coalescents. *Acta Applicandae Mathematicae*, 142(1):127–148, 2016.

[SKM+17] Egor Shitikov, Sergey Kolchenko, Igor Mokrousov, Julia Bespyatykh, Dmitry Ischenko, Elena Ilina, and Vadim Govorun. Evolutionary pathway analysis and unified classification of east asian lineage of mycobacterium tuberculosis. *Scientific reports*, 7(1):9227, 2017.

Bibliography X

[SKS16] Jeffrey P. Spence, John A. Kamm, and Yun S. Song. The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561, 2016.

[TV09] Jesse E Taylor and Amandine Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab*, 14:242–288, 2009.

[ZBT+15] Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. Population genomics of intrapatient hiv-1 evolution. *Elife*, 4:e11282, 2015.