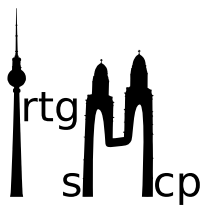
Computing likelihoods under Λ-coalescents

Matthias Birkner

LMU Munich, Dept. Biology II

Joint project with Jochen Blath and Matthias Steinrücken, TU Berlin

Probabilistic Models of Evolutionary Biology
CIRM, Luminy, 25–29 May 2009



Outline

- Introduction
- Beta$(2-\alpha, \alpha)$-coalescents
- Mutation models: Infinitely-many-alleles, infinitely-many-sites
- Computing likelihoods under (Λ-)coalescents
- Importance sampling methods
- Summary & Outlook

Genetic variability at the mitochondrial cyt *b* locus in Atlantic cod

	468	481	487	488	490	496	508	523	562	601	631	643	649	685	691
66	t	a	a	c	a	a	t	g	a	t	g	a	c	c	g
17	-	-	-	-	-	-	c	-	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-	a	-	-	-	-	-	t	-
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	t
1	-	-	-	-	-	-	-	-	-	-	-	-	-	t	-
2	-	-	-	t	-	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	a	-	-	-	g	-	t	-
1	-	-	-	-	-	-	-	-	-	-	-	-	t	-	-
1	-	-	-	-	g	-	c	-	-	-	-	-	-	-	-
1	-	-	-	-	-	g	-	-	-	-	-	-	-	-	-
1	-	-	g	-	-	-	-	-	-	-	a	-	-	-	t
1	-	-	-	-	-	-	c	-	g	-	-	-	-	-	-
1	g	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-	c	-	-	-	-	-
1	-	c	-	-	-	-	c	-	-	-	-	-	-	-	-

(a random subsample of the sample described in Árnason, *Genetics* 2004)

The Great Obsession of population geneticists (J. Gillespie)

What evolutionary forces could have lead to such divergence between individuals of the same species?

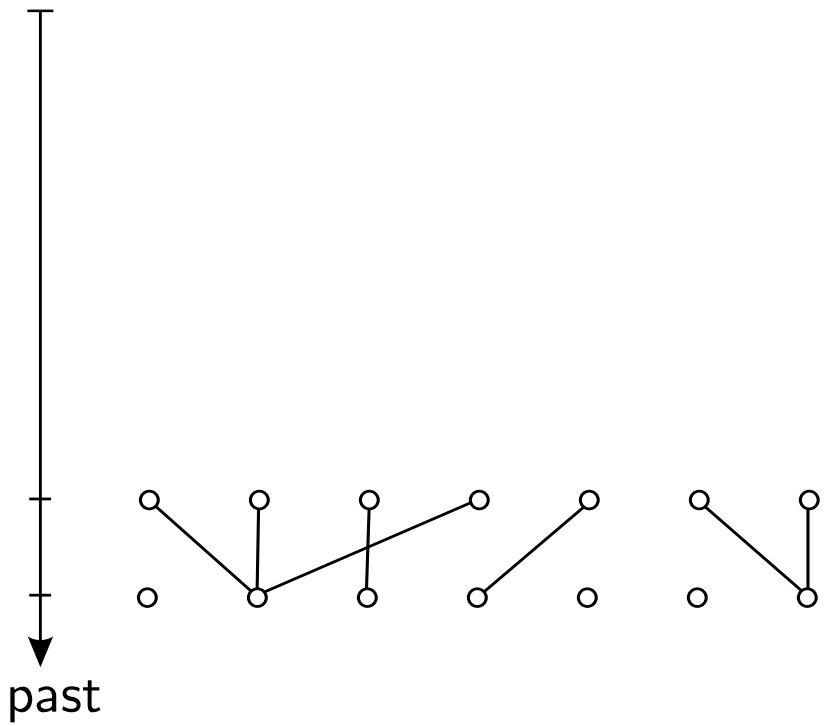
The Great Obsession of population geneticists (J. Gillespie)

What evolutionary forces could have lead to such divergence between individuals of the same species?

In this talk, we will focus on *neutral* genetic variation, and thus the interplay of *mutation* and *genetic drift*.

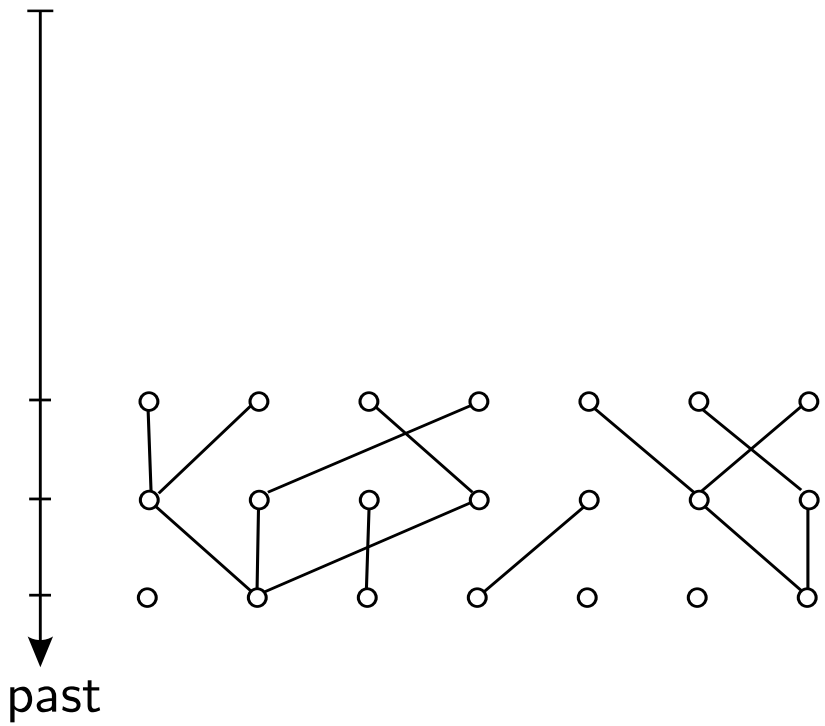
Wright-Fisher model: The fundamental model for 'genetic drift'

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



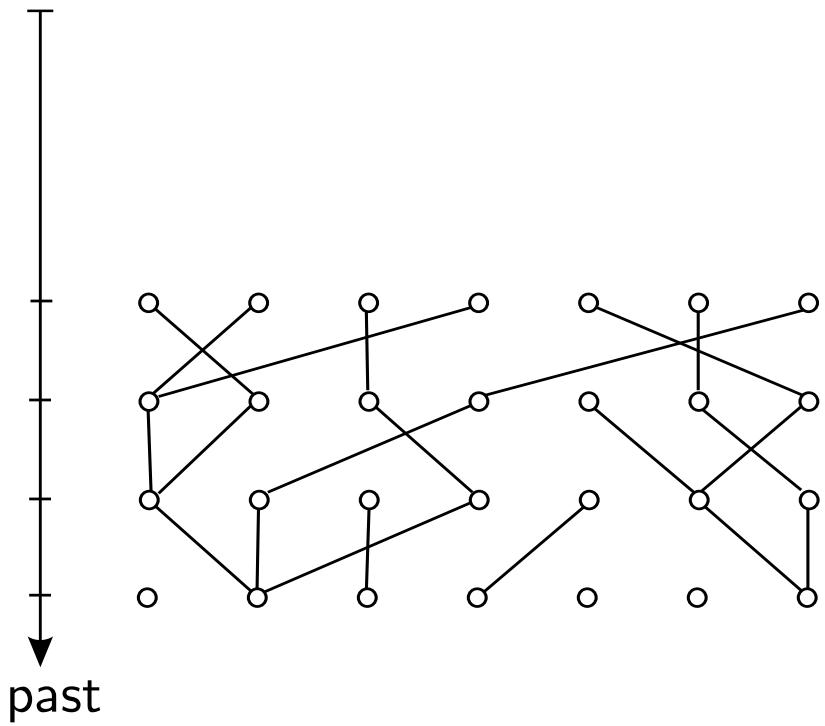
Wright-Fisher model: The fundamental model for 'genetic drift'

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



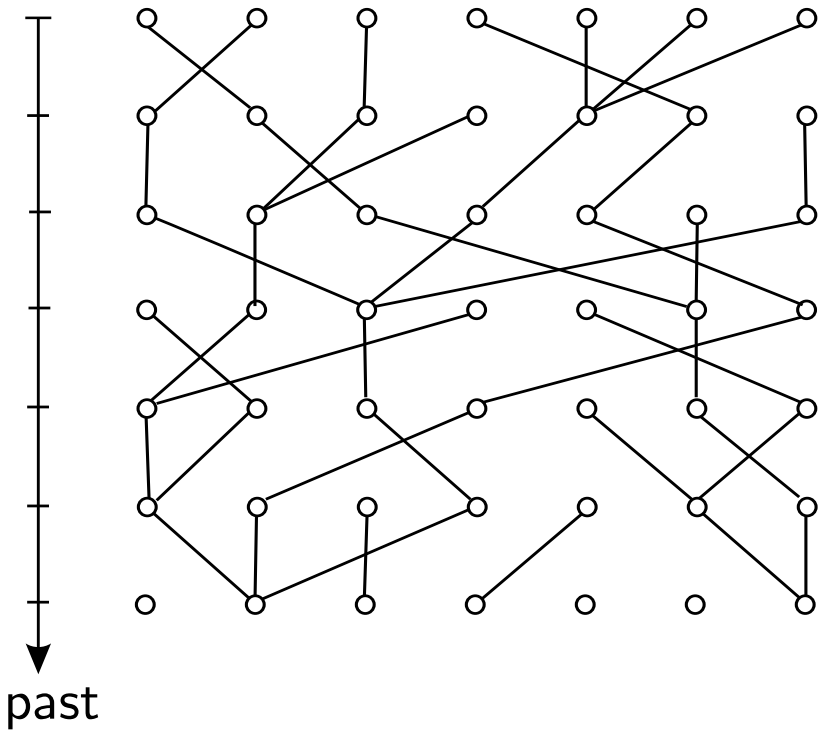
Wright-Fisher model: The fundamental model for 'genetic drift'

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



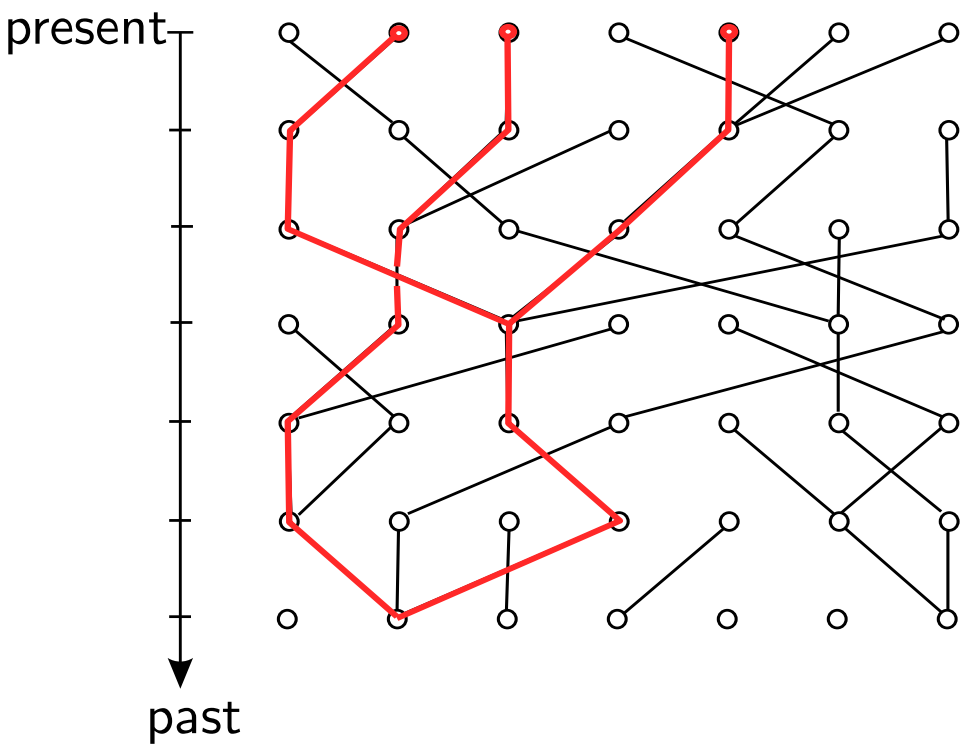
Wright-Fisher model: The fundamental model for 'genetic drift'

- A (haploid) population of N individuals per generation,
- each individual in the present generation picks a 'parent' at random from the previous generation,
- genetic types are inherited (possibly with a small probability of mutation).



Genealogical point of view

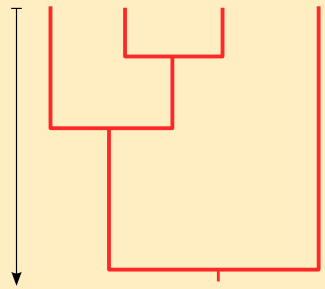
Sample $n\,(\ll N)$ individuals from the 'present generation'



Kingman's coalescent

Theorem (Kingman (& Hudson, Griffiths), 1982)

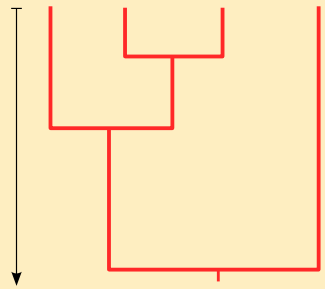
In the limit $N \to \infty$, the genealogy of an n-sample, measured in units of N generations, is described by a continuous-time Markov chain where each pair of lineages merges at rate 1.



Kingman's coalescent

Theorem (Kingman (& Hudson, Griffiths), 1982)

In the limit $N \to \infty$, the genealogy of an n-sample, measured in units of N generations, is described by a continuous-time Markov chain where each pair of lineages merges at rate 1.

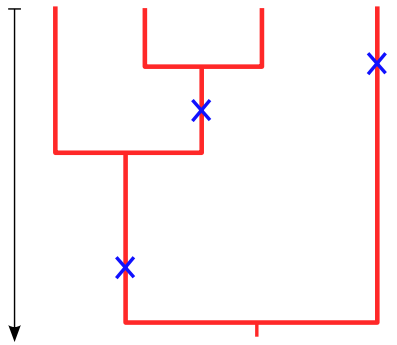


Robustness. The same limit appears for any *exchangeable* offspring vectors

$$(\nu_1, \ldots, \nu_N), \qquad \text{(independent over generations)},$$

if time is measured in units of $\dfrac{N}{\sigma^2}$ generations, where $\sigma^2 = \lim\limits_{N\to\infty} \mathrm{Var}(\nu_1)$ (under a third moment condition on ν_1).

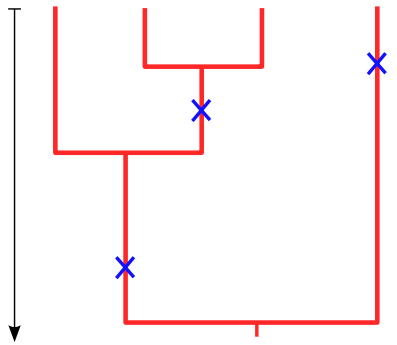
Modeling neutral variation: Superimposing types on the coalescent



Assume that the considered genetic types
do not affect their bearer's reproductive succes.

If as population size $N \to \infty$,
$\frac{N}{\sigma^2} \times$ mutation prob. per ind. per generation $\to r$,
the type configuration in the sample can be described by putting mutations with rate r along the genealogy.

Modeling neutral variation: Superimposing types on the coalescent



Assume that the considered genetic types
do not affect their bearer's reproductive succes.

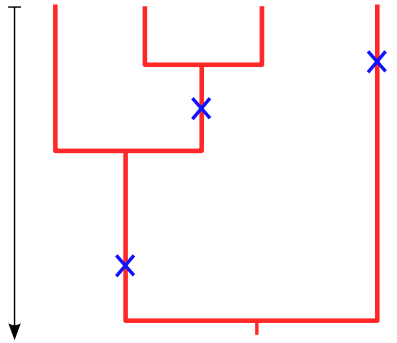
If as population size $N \to \infty$,
$\frac{N}{\sigma^2} \times$ mutation prob. per ind. per generation $\to r$,
the type configuration in the sample can be described by putting mutations with rate r along the genealogy.

Kingman's coalescent is the *standard model* of mathematical population genetics.

Modeling neutral variation: infinitely-many-alleles model

If mutations always generate a *completely new type*, information in n-sample is equivalent to allelic partition

$$(a_1, a_2, \ldots, a_n)$$

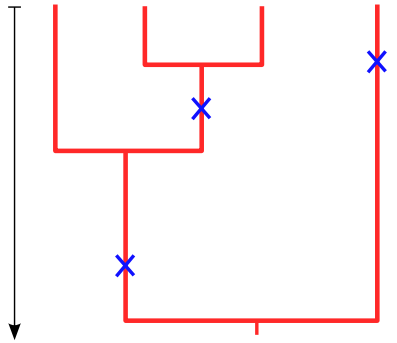


where a_j=no. of types with j representatives in the sample ($\sum_{i=j}^{n} j a_j = n$)

Modeling neutral variation: infinitely-many-alleles model

If mutations always generate a *completely new type*, information in n-sample is equivalent to allelic partition

$$(a_1, a_2, \ldots, a_n)$$



where a_j=no. of types with j representatives in the sample ($\sum_{i=j}^{n} j a_j = n$)

Then (Ewens' sampling formula, 1972)

$$p_r\big((a_1, a_2, \ldots, a_n)\big) = \frac{n!}{2r(2r+1)\cdots(2r+n-1)} \prod_{j=1} \frac{(2r/j)^{a_j}}{a_j!}.$$

Question

What if the variability of surviving offspring numbers across individuals is so large that reasonably

$$\text{individual offspring variance } \sigma^2 \approx \infty \text{ ?}$$

This might happen e.g. in marine species (so-called *reproduction sweepstakes*).

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate $\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$,

where Λ is a finite measure on $[0,1]$. (Sagitov, 1999; Pitman, 1999).

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate $\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$,

where Λ is a finite measure on $[0,1]$. (Sagitov, 1999; Pitman, 1999).

Interpretation:
re-write $\lambda_{n,k} = \int_{[0,1]} x^k(1-x)^{n-k}\,\frac{1}{x^2}\Lambda(dx)$ to see:
at rate $\frac{1}{x^2}\Lambda([x, x+dx])$, an '$x$-resampling event' occurs.

Thinking forwards in time, this corresponds to an event in which the fraction x of the total population is replaced by the offspring of a single individual.

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate $\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$, where Λ is a finite measure on $[0,1]$. (Sagitov, 1999; Pitman, 1999).

Interpretation:
re-write $\lambda_{n,k} = \int_{[0,1]} x^k(1-x)^{n-k}\,\frac{1}{x^2}\Lambda(dx)$ to see:
at rate $\frac{1}{x^2}\Lambda([x, x+dx])$, an '$x$-resampling event' occurs.

Thinking forwards in time, this corresponds to an event in which the fraction x of the total population is replaced by the offspring of a single individual.

Form of rates stems from $\lambda_{n,k} = \lambda_{n+1,k} + \lambda_{n+1,k+1}$ (consistency condition).

Coalescents with multiple collisions, aka 'Λ-coalescents'

While n lineages, any k coalesce at rate $\lambda_{n,k} = \int_{[0,1]} x^{k-2}(1-x)^{n-k}\,\Lambda(dx)$, where Λ is a finite measure on $[0,1]$. (Sagitov, 1999; Pitman, 1999).

Interpretation:
re-write $\lambda_{n,k} = \int_{[0,1]} x^k(1-x)^{n-k}\,\frac{1}{x^2}\Lambda(dx)$ to see:
at rate $\frac{1}{x^2}\Lambda([x, x+dx])$, an '$x$-resampling event' occurs.

Thinking forwards in time, this corresponds to an event in which the fraction x of the total population is replaced by the offspring of a single individual.

Form of rates stems from $\lambda_{n,k} = \lambda_{n+1,k} + \lambda_{n+1,k+1}$ (consistency condition).

Note: $\Lambda = \delta_0$ corresponds to Kingman's coalescent.

Cannings' models in the 'domain of attraction of a Λ-coalescent'

Fixed population size N, *exchangeable* offspring numbers in one generation

$$\big(\nu_1, \nu_2, \ldots, \nu_N\big).$$

Sagitov (1999), Möhle & Sagitov (2001) clarify under which conditions the genealogies of a sequence of exchangeable finite population models are described by a Λ-coalescent:

- $c_N :=$ pair coalescence probability over one generation $\to 0$
  ($c_N = \frac{1}{N-1}\mathbb{E}[\nu_1(\nu_1 - 1)]$)
- two double mergers asymptotically negligible compared to one triple merger
- $Nc_N \Pr\big(\text{a given family has size} \geq Nx\big) \sim \int_x^1 y^{-2}\Lambda(dy)$

Time is measured in $1/c_N$ generations (in general $\neq 1/\text{pop. size}$)

Note: There are many Λ-coalescents.

Maybe a natural "first candidate":

$$\Lambda = w\delta_0 + (1-w)\delta_\psi \quad \text{with} \quad w, \psi \in (0,1)$$

(as considered by Eldon & Wakeley, *Genetics* 2006)

A 'heavy-tailed' Cannings model and Beta-coalescents

Haploid population of size N. Individual i has X_i *potential offspring*,
$X_1, X_2, \ldots, X_N$ are i.i.d. with mean $m := \mathbb{E}\left[X_1\right] > 1$,
$\Pr\left(X_1 \geq k\right) \sim$ Const. $\times\, k^{-\alpha}$ with $\alpha \in (1, 2)$.
Note: infinite variance.

Sample N without replacement from all potential offspring to form the next generation.

A 'heavy-tailed' Cannings model and Beta-coalescents

Haploid population of size N. Individual i has X_i *potential offspring*,
$X_1, X_2, \ldots, X_N$ are i.i.d. with mean $m := \mathbb{E}\left[X_1\right] > 1$,
$\Pr\left(X_1 \geq k\right) \sim$ Const. $\times\, k^{-\alpha}$ with $\alpha \in (1,2)$.
Note: infinite variance.

Sample N without replacement from all potential offspring to form the next generation.

Theorem (Schweinsberg, 2003)

Let c_N = prob. of pair coalescence one generation back in N-th model. $c_N \sim$ const. $N^{1-\alpha}$, measured in units of $1/c_N$ generations, the genealogy of a sample from the N-th model is approximately described by a Λ-coalescent with $\Lambda = \mathrm{Beta}(2-\alpha,\alpha)$.

$\left(\mathrm{Beta}(2-\alpha,\alpha)(dx) = \mathbb{1}_{[0,1]}(x)\frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}\,dx \right)$

Why $\Lambda = \mathrm{Beta}(2-\alpha, \alpha)$?

Heuristic argument:

Probability that first individual's offspring provides

more than fraction y of the next generation,

given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$, for $y > \varepsilon$)

Why $\Lambda = \mathrm{Beta}(2-\alpha, \alpha)$?

Heuristic argument:

Probability that first individual's offspring provides

more than fraction y of the next generation,

given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$, for $y > \varepsilon$)

$$\approx \quad \mathbb{P}\Big(\frac{X_1}{X_1 + (N-1)m} \geq y \Big| X_1 \geq \varepsilon N\Big)$$

Why $\Lambda = \mathrm{Beta}(2-\alpha,\alpha)$?

Heuristic argument:

Probability that first individual's offspring provides

more than fraction y of the next generation,

given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$, for $y > \varepsilon$)

$$\begin{aligned} \approx \quad & \mathbb{P}\Big(\frac{X_1}{X_1+(N-1)m} \geq y \Big| X_1 \geq \varepsilon N\Big) \\ & = \mathbb{P}\Big(X_1 \geq (N-1)m\frac{y}{1-y}\Big| X_1 \geq \varepsilon N\Big) \end{aligned}$$

Why $\Lambda = \mathrm{Beta}(2-\alpha,\alpha)$?

Heuristic argument:

Probability that first individual's offspring provides

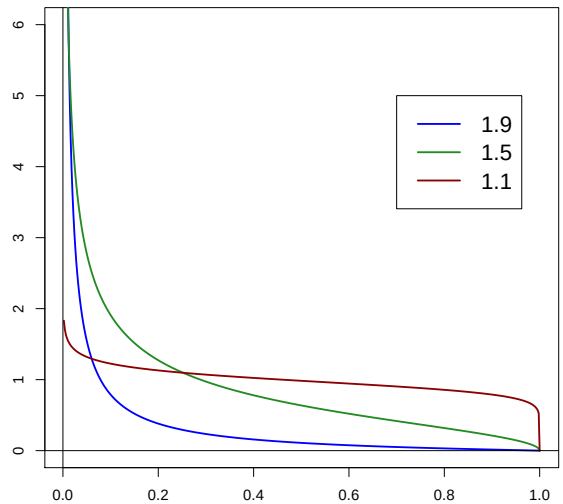
more than fraction y of the next generation,

given that the family is substantial (i.e. given $X_1 \geq \varepsilon N$, for $y > \varepsilon$)

$$\begin{aligned} \approx \quad & \mathbb{P}\Big(\frac{X_1}{X_1+(N-1)m} \geq y \Big| X_1 \geq \varepsilon N\Big) \\ & = \mathbb{P}\Big(X_1 \geq (N-1)m\frac{y}{1-y}\Big| X_1 \geq \varepsilon N\Big) \\ & \sim \text{const.}\frac{(1-y)^\alpha}{y^\alpha} = \text{const.}'\,\mathrm{Beta}(2-\alpha,\alpha)([y,1]). \end{aligned}$$

The family $\mathrm{Beta}(2-\alpha,\alpha)$, $\alpha \in (1,2]$

- Kingman's coalescent included as boundary case: $\mathrm{Beta}(2-\alpha,\alpha) \to \delta_0$ weakly as $\alpha \to 2$.
- Smaller α means tendency towards more extreme resampling events.
- For $\alpha \leq 1$, corresponding coalescents *do not* come down from infinity.

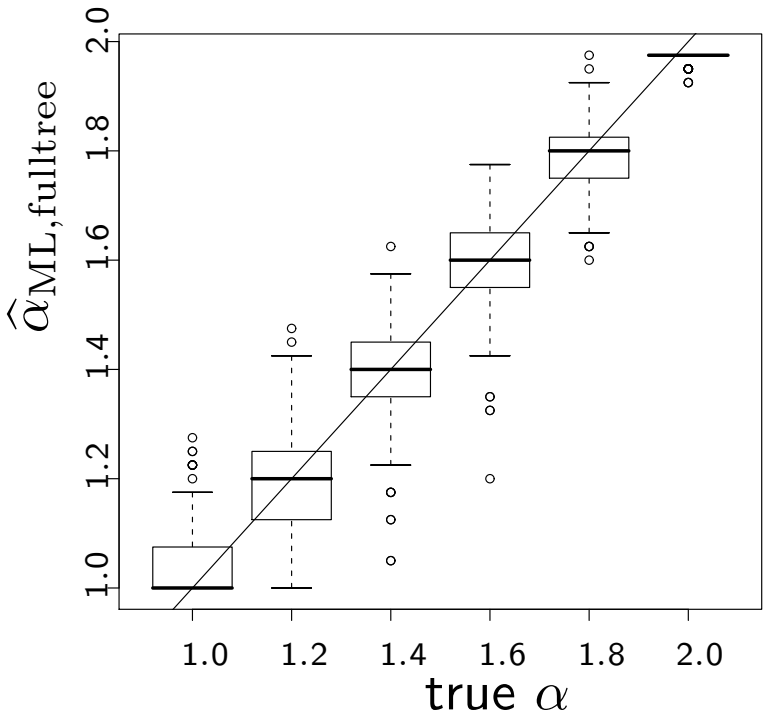


- $\mathrm{Beta}(2-\alpha,\alpha)$-coalescents appear as genealogies of α-stable continuous mass branching process (via a time-change).
- Scaling relation of mutation rate per generation relative to population size depends on α!

'Meta-mathematic' associations

Brownian motion	$\leftrightarrow$	Kingman's coalescent
$\cap$		$\cap$
Stable processes	$\leftrightarrow$	Beta$(2-\alpha, \alpha)$-coalescents
$\cap$		$\cap$
General Lévy processes	$\leftrightarrow$	General Λ-coalescents

Playing god with simulated "full trees"



ML estimates of α for simulted datasets with sample size $n = 100$, estimate based on *full genealogical tree*
(400 replicates for each value of α).

Consider n-Λ-coalescent with mutation rate r per line (and infinite alleles mutation model). $\mathbf{n} = (n_1, \ldots, n_\ell)$, possible type configuration

Theorem (Möhle 2005)

The probability $p(\mathbf{n})$ of observing a type configuration $\mathbf{n} = (n_1, \ldots, n_\ell)$ satisfies the recursion given by $p(1) = 1$ and

$$p(\mathbf{n}) = \frac{nr}{\sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} + nr} \sum_{\substack{j=1 \\ n_j=1}}^{\ell} \frac{1}{\ell} p(\tilde{\mathbf{n}}^{(j)})$$

$$+ \frac{1}{\sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} + nr} \sum_{k=2}^{n} \sum_{\substack{j=1 \\ n_j \geq k}}^{\ell} \binom{n}{k} \lambda_{n,k} \frac{n_j - k + 1}{n - k + 1} p(\mathbf{n} - (k-1)\mathbf{e}_j).$$

$(\tilde{\mathbf{n}}^{(j)} = (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k))$

Infinitely many sites model

Model genetic locus as infinite sequence of completely linked sites, mutations always hit a new site.

Mathematical abstraction:

- a gene is $[0, 1]$
- a type is a configuration of points on $[0, 1]$

Ethier & Griffiths (1987) parametrisation:

- type space $E = [0, 1]^{\mathbb{N}}$
- mutation operator

$$Bf\big((x_1, x_2, \dots)\big) = r \int_0^1 f\big((u, x_1, x_2, \dots)\big) - f\big((x_1, x_2, \dots)\big)\, du$$

Asymptotics of the frequency spectrum

Consider an n-$\mathrm{Beta}(2-\alpha,\alpha)$-coalescent, mutations at rate r according to the *infinitely-many-sites* model (assuming known ancestral types). Let

$$\begin{aligned} M(n) &:= \#\text{total number of mutations in the sample,} \\ M_k(n) &:= \#\text{number of mutations affecting exactly } k \text{ samples,} \end{aligned}$$

$k = 1, 2, \ldots, n-1$.

Theorem (Berestycki, Berestycki & Schweinsberg 2007)

$$\frac{M(n)}{n^{2-\alpha}} \to r\frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}, \qquad \frac{M_k(n)}{n^{2-\alpha}} \to r\alpha(\alpha-1)^2\frac{\Gamma(k+\alpha-2)}{k!}$$

in probability as $n \to \infty$.

Asymptotics of the frequency spectrum

Consider an n-$\mathrm{Beta}(2-\alpha,\alpha)$-coalescent, mutations at rate r according to the *infinitely-many-sites* model (assuming known ancestral types). Let

$$\begin{aligned} M(n) &:= \#\text{total number of mutations in the sample,} \\ M_k(n) &:= \#\text{number of mutations affecting exactly } k \text{ samples,} \end{aligned}$$

$k = 1, 2, \ldots, n-1$.

Theorem (Berestycki, Berestycki & Schweinsberg 2007)

$$\frac{M(n)}{n^{2-\alpha}} \to r\frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}, \qquad \frac{M_k(n)}{n^{2-\alpha}} \to r\alpha(\alpha-1)^2\frac{\Gamma(k+\alpha-2)}{k!}$$

in probability as $n \to \infty$.

Thus $M_1(n)/M(n) \approx 2-\alpha$ for n large, which suggests

$$\widehat{\alpha}_{\mathrm{BBS}} := 2 - \frac{M_1(n)}{M(n)} \qquad \text{as an estimator for } \alpha.$$

Infinitely-many-sites model

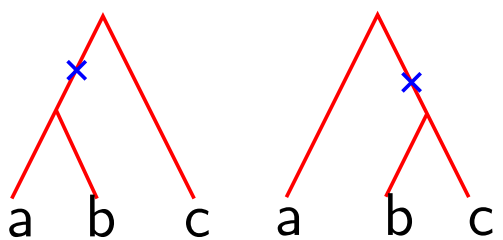
Model genetic locus as infinite sequence of completely linked sites, mutations always hit a new site

Example:

	segr. site			
Seq.	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	0	0	1	1
4	0	0	1	1
5	0	0	1	0

(0=wild type, 1=mutant
assume known ancestral types)

Obs. fit IMS $\iff$ no sub-matrix $\begin{matrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{matrix}$
(and no row permutation).



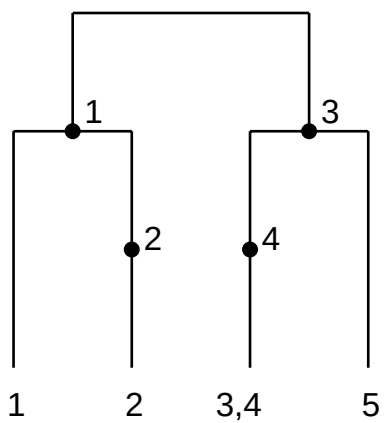
Infinitely-many-sites model, II

If the infinitely-many-sites model applies, the observations correspond to a unique rooted perfect phylogeny (or 'genetree').

Sequences,

	segr. site			
Seq.	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	0	0	1	1
4	0	0	1	1
5	0	0	1	0

Genetree,



obs. types

type	multiplicity
(1, 0)	1
(2, 1, 0)	1
(4, 3, 0)	2
(3, 0)	1

Construct e.g. using GUSFIELD's (1991) algorithm.

Note: purely combinatorial, does not depend on a probabilistic model for the observations.

Simulating samples under the IMS model

The Ethier-Griffiths urn (1987) can be used to generate a random sample of size n under Kingman's coalescent
(with mutation rate r per line):

- Start with 2 leaves.
- When there are k leaves:

  Add a mutation to a leaf w. prob. $\frac{2r}{2r+(k-1)}$,

  split one leaf w. prob. $\frac{k-1}{2r+(k-1)}$

  (leaf picked uniformly among the k).
- Stop when $n+1$ leaves, delete last leaf.

Simulating samples under the IMS model: Λ-case

$(Y_t^{(n)})_{\geq 0}$ *block counting process* of Λ-coalescent starting from n blocks:

- Jump from i to $j \in \{1, 2, \ldots, i-1\}$ at rate $q_{ij} := \binom{i}{i-j+1} \lambda_{i,i-j+1}$.
- $\tau_1 := \inf\{t \geq 0 : Y_t^{(n)} = 1\}$.

$\widetilde{Y}_t^{(n)} := Y_{(\tau_1 - t)-}^{(n)}$ *time-reversed* block counting process

- $(\widetilde{Y}_t^{(n)} = \partial$ for $t \geq \tau_1)$.
- Jump rates $\widetilde{q}_{ji}^{(n)} = \frac{g_{ni} q_{ij}}{g_{nj}}$, $\widetilde{q}_{n\partial}^{(n)} = -q_{nn} = \sum_{j=1}^{n-1} q_{nj}$,
- $\mathbb{P}(\widetilde{Y}_0^{(n)} = k) = \mathbb{P}(Y_{\tau_1 -}^{(n)} = k) = g_{nk} q_{k1}$.

- $g_{ni} := \mathbb{E} \int_0^\infty \mathbf{1}(Y_t^{(n)} = i)\, dt$ is the Green function (in general, not known explicitly, but easy recursion).

Simulating samples under the IMS model: Λ-case, cont.

The n-Λ-"Ethier-Griffiths urn" (mutation rate r).

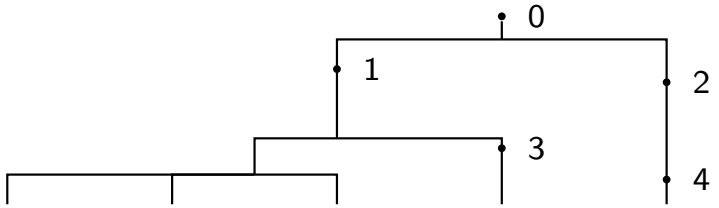
- Begin with K leaves, $\mathbb{P}(K = k) = \mathbb{P}(\widetilde{Y}_0^{(n)} = k)$.
- While there are k leaves:

Add a mutation to a leaf	w. prob.	$\frac{r}{kr - \widetilde{q}_{kk}^{(n)}}$,
split one leaf into ℓ	w. prob.	$\frac{\widetilde{q}_{k,k+\ell-1}^{(n)}}{\widetilde{q}_{kk}^{(n)}}$,
if $k = n$ goto stop	w. prob.	$\frac{-\widetilde{q}_{nn}^{(n)}}{kr - \widetilde{q}_{nn}^{(n)}}$

  (leaf picked uniformly among the k).
- Stop.

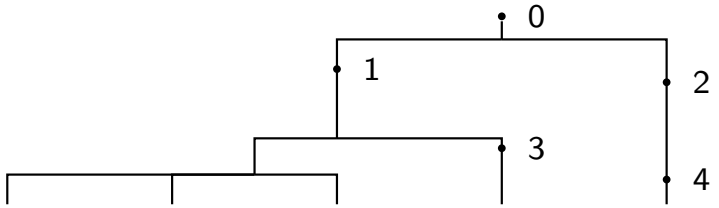
Recursion for tree probabilities

Can calculate $\mathbb{P}_{(r,\Lambda)}\big(\text{observed sequence data } (\mathbf{t},\mathbf{n})\big) =: p(\mathbf{t},\mathbf{n})$ via



Recursion for tree probabilities

Can calculate $\mathbb{P}_{(r,\Lambda)}\big(\text{observed sequence data } (\mathbf{t},\mathbf{n})\big) =: p(\mathbf{t},\mathbf{n})$ via



$$p(\mathbf{t},\mathbf{n}) = \frac{1}{rn+\lambda_n}\sum_{i:n_i\geq 2}\sum_{k=2}^{n_i}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1}p\big(\mathbf{t},\mathbf{n}-(k-1)\mathbf{e}_i\big)$$
$$+\frac{r}{rn+\lambda_n}\sum_{\substack{i:n_i=1,x_{i0}\text{unique},\\ \mathfrak{s}(\mathbf{x}_i)\neq\mathbf{x}_j\,\forall j}} p\big(\mathfrak{s}_i(\mathbf{t}),\mathbf{n}\big)$$
$$+\frac{r}{rn+\lambda_n}\frac{1}{d}\sum_{\substack{i:n_i=1,\\ x_{i0}\text{unique}}}\sum_{j:\mathfrak{s}(\mathbf{x}_i)=\mathbf{x}_j}(n_j+1)p\big(\mathfrak{r}_i(\mathbf{t}),\mathfrak{r}_i(\mathbf{n}+\mathbf{e}_j)\big).$$

Extends Ethier & Griffiths (1987) to Λ-coalescents and Möhle's recursion (2005) to IMS model.

Compute probabilities

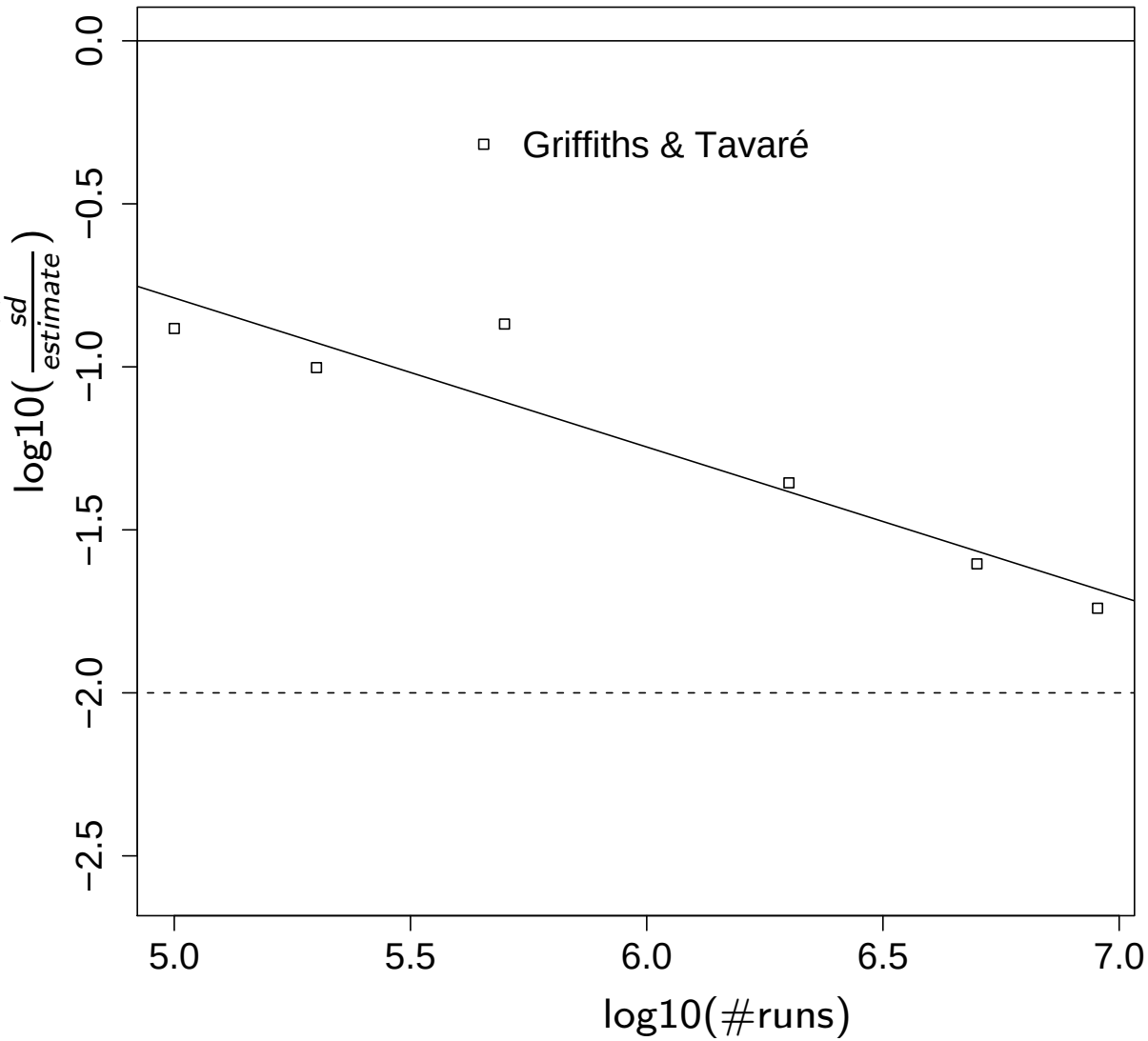
Use exact recursions for moderate sample complexities.
Approach more complex samples by version of Griffiths & Tavaré's (1994)

Monte Carlo method

$$p(\mathbf{t}, \mathbf{n}) = \mathbb{E}_{(\mathbf{t},\mathbf{n})}\Big[\prod_{i=0}^{\tau-1} f_{(r,\Lambda)}(X_i)\Big]$$

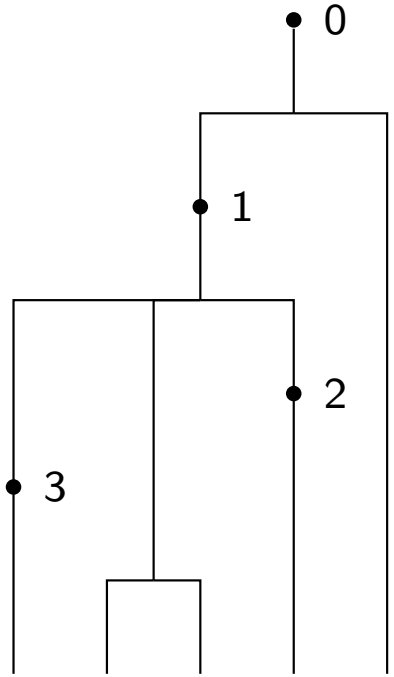
- For suitable Markov chain X_i on sample configurations.
- Estimate expectation via empirical mean of independent runs.
- extension to Λ-coalescents by B. & Blath (2008)

Artifical sample of size 12 analysed with $r = 1$ and $\alpha = 1.5$:



Histories

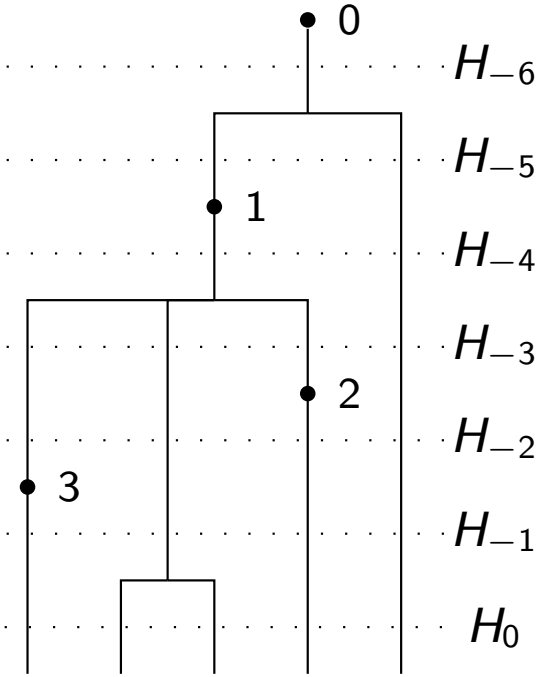
Interpret genealogy as sequency of historical states:
$\mathcal{H} = \left(H_{-\tau} = ((1),(0)), H_{\tau-1}, \ldots, H_{-1}, H_0 = (\mathbf{t}, \mathbf{n})\right)$



$$\Big(\big((3,1,0),(1,0),(2,1,0),(0)\big),(1,2,1,1)\Big)$$

Histories

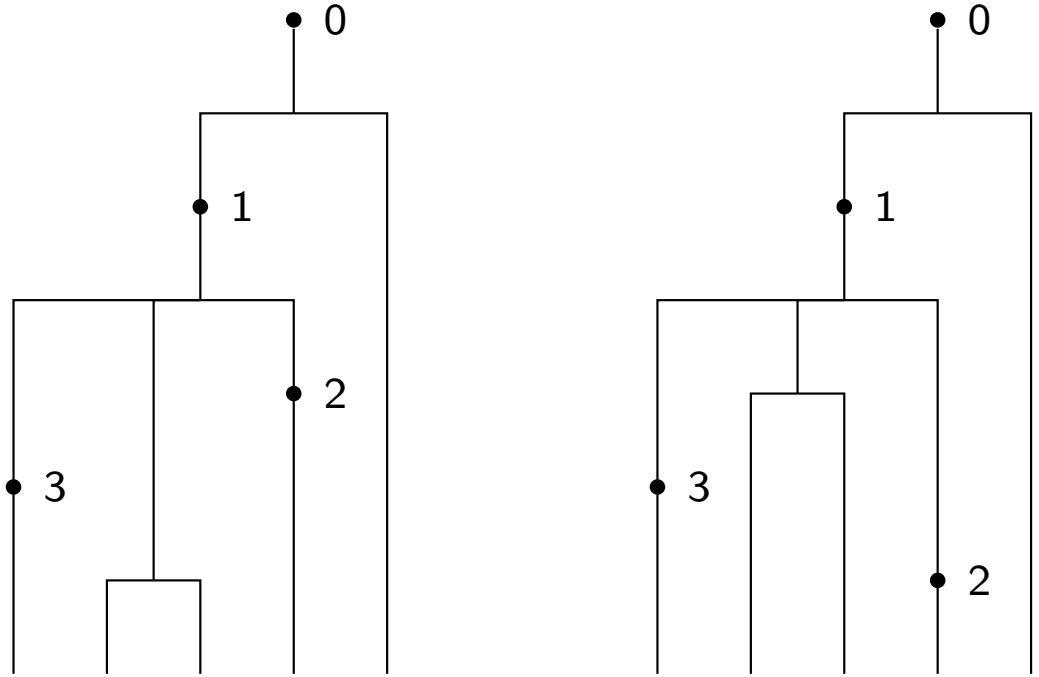
Interpret genealogy as sequency of historical states:
$\mathcal{H} = \left(H_{-\tau} = ((1),(0)), H_{\tau-1}, \ldots, H_{-1}, H_0 = (\mathbf{t}, \mathbf{n})\right)$



$$\Big(((3,1,0),(1,0),(2,1,0),(0)),(1,2,1,1)\Big)$$

Histories

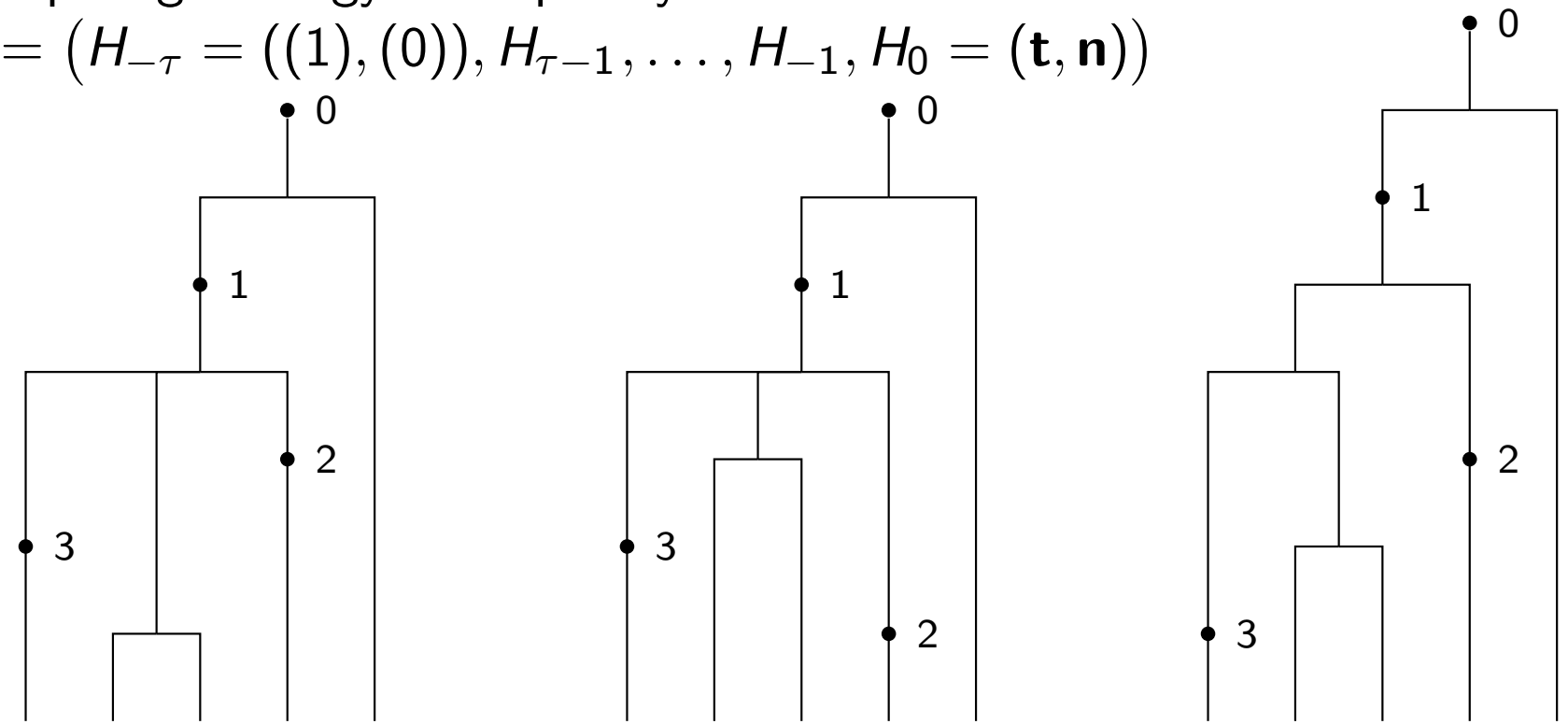
Interpret genealogy as sequency of historical states:
$\mathcal{H} = \left(H_{-\tau} = ((1),(0)), H_{\tau-1}, \ldots, H_{-1}, H_0 = (\mathbf{t}, \mathbf{n})\right)$



$$\Big(\big((3,1,0),(1,0),(2,1,0),(0)\big),(1,2,1,1)\Big)$$

Histories

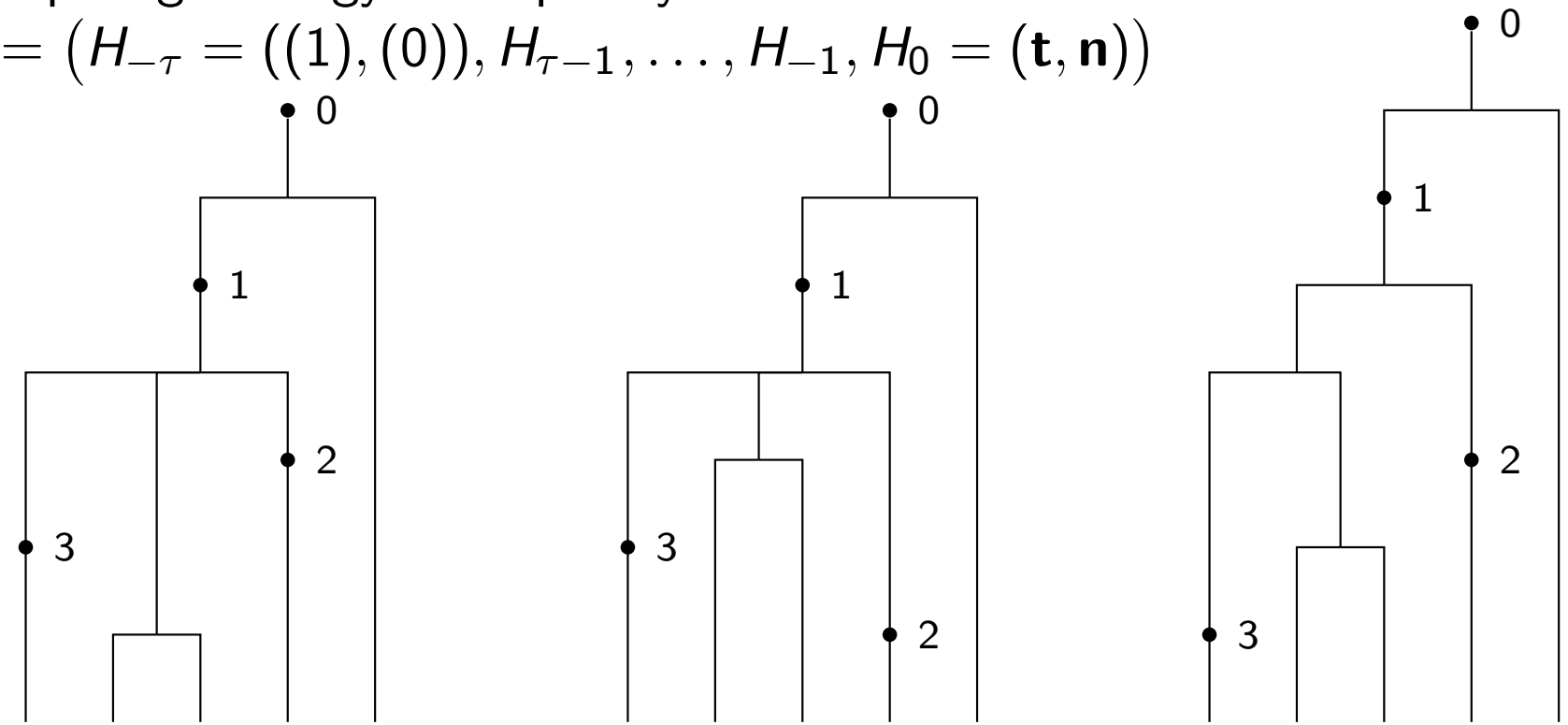
Interpret genealogy as sequency of historical states:
$\mathcal{H} = \left(H_{-\tau} = ((1),(0)), H_{\tau-1}, \ldots, H_{-1}, H_0 = (\mathbf{t}, \mathbf{n})\right)$



$$\left(((3,1,0),(1,0),(2,1,0),(0)),(1,2,1,1)\right)$$

Histories

Interpret genealogy as sequency of historical states:
$\mathcal{H} = \left(H_{-\tau} = ((1),(0)), H_{\tau-1}, \ldots, H_{-1}, H_0 = (\mathbf{t}, \mathbf{n})\right)$



different histories can lead to same sample

$$\Big(\big((3,1,0),(1,0),(2,1,0),(0)\big),(1,2,1,1)\Big)$$

Importance sampling

We have

$$p(\mathbf{t},\mathbf{n}) = \mathbb{P}_{(r,\Lambda)}\big(H_0 = (\mathbf{t},\mathbf{n})\big) = \sum_{\mathcal{H}:H_0=(\mathbf{t},\mathbf{n})} \mathbb{P}_{(r,\Lambda)}\big(\mathcal{H}\big)$$

$$= \sum_{\mathcal{H}:H_0=(\mathbf{t},\mathbf{n})} \underbrace{\frac{\mathbb{P}_{(r,\Lambda)}\big(\mathcal{H}\big)}{\mathcal{Q}\big(\mathcal{H}\big)}}_{\substack{=:w(\mathcal{H}) \\ \text{importance weight}}} \mathcal{Q}(\mathcal{H}),$$

for any law $\mathcal{Q}$ on histories s.th. $\mathbb{P}_{(r,\Lambda)}\Big|_{\{H_0=(\mathbf{t},\mathbf{n})\}} \ll \mathcal{Q}$.

Thus,

$$p(\mathbf{t},\mathbf{n}) \approx \frac{1}{R}\sum_{i=1}^{R} w(\mathcal{H}^{(i)}),$$

where $\mathcal{H}^{(1)},\ldots,\mathcal{H}^{(R)}$ are independent samples from $\mathcal{Q}$

(Theoretical) optimal solution

$$p(\mathbf{t},\mathbf{n}) = \sum_{\mathcal{H}:H_0=(\mathbf{t},\mathbf{n})} \frac{\mathbb{P}_{(r,\Lambda)}(\mathcal{H})}{\mathcal{Q}(\mathcal{H})}\mathcal{Q}(\mathcal{H}) \approx \frac{1}{R}\sum_{i=1}^{R} w(\mathcal{H}^{(i)})$$

$\mathcal{Q}_{\mathrm{opt}}(\cdot) := \mathbb{P}_{(r,\Lambda)}\big(\,\cdot\,\big|H_0 = (\mathbf{t},\mathbf{n})\big)$ is optimal (Stephens & Donnelly 2000):

- Variance of estimator is zero since $w(\mathcal{H}^{(i)}) \equiv p(\mathbf{t},\mathbf{n})$.
- Finding $\mathcal{Q}_{\mathrm{opt}}$ is as hard as the original problem.
- $H_0, H_{-1}, \ldots$ is Markov chain under $\mathcal{Q}_{\mathrm{opt}}$.

Remark: Transistion probabilities $q_{\mathrm{GT}}(H_i|H_{i+1}) \propto \mathbb{P}_{(r,\Lambda)}(H_{i+1}|H_i)$ gives (Λ-)Griffiths-Tavaré method.

Stephens and Donnelly's (2000) IMS candidate

Kingman case: Choose individual uniformly:

- If type is unique in sample, remove "outmost" mutation,
- if at least two individuals with this type, merge two lines.

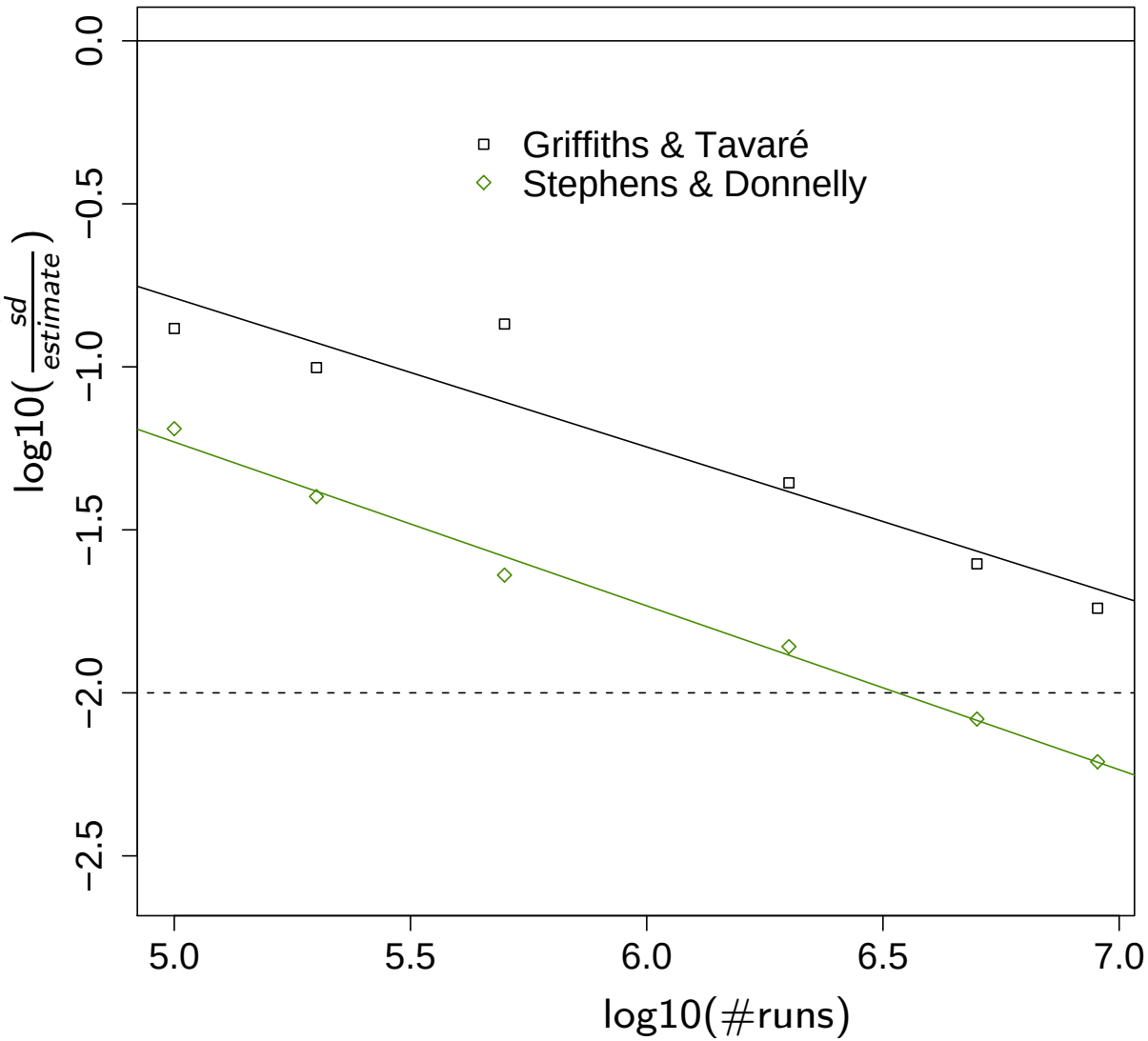
(this *would be* optimal for parent-independent mutations)

Heuristic extension to Λ case:

$$(\mathbf{t},\mathbf{n}) \to \begin{cases} \left(\mathfrak{s}_i(\mathbf{t}),\mathbf{n}\right) & \text{w.p. } \propto 1 \text{ if } n_i = 1, x_{i0} \text{ unique, } \mathfrak{s}_i(\mathbf{x}_i) \neq \mathbf{x}_j \forall j \\ \left(\mathfrak{r}_i(\mathbf{t},\mathfrak{r}_i(\mathbf{n}+\mathbf{e}_j)\right) & \text{w.p. } \propto 1 \text{ if } n_i = 1, x_{i0} \text{ unique, } \mathfrak{s}_i(\mathbf{x}_i) \neq \mathbf{x}_j \\ \left(\mathbf{t},\mathbf{n}-(k-1)\mathbf{e}_i\right) & \text{w.p. } \propto n_i \bar{q}_{n_i}(k) \text{ if } 2 \leq k \leq n_i, \end{cases}$$

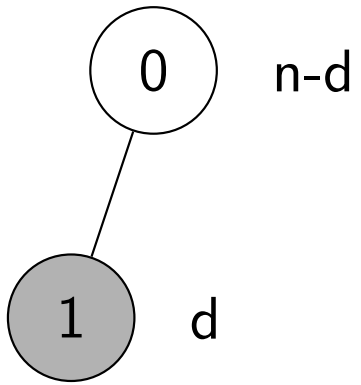
where $\bar{q}_{n_i}(k) = \frac{q_{n,n-k+1}}{\sum_{l=2}^{n_i} q_{n,n-l+1}}$, jump probabilities of block counting process.

Artifical sample of size 12 analysed with $r = 1$ and $\alpha = 1.5$:



Hobolth, Uyenoyama & Wiuf's (2008) idea

Sample of size n where exactly one mutation is visible (in d copies).



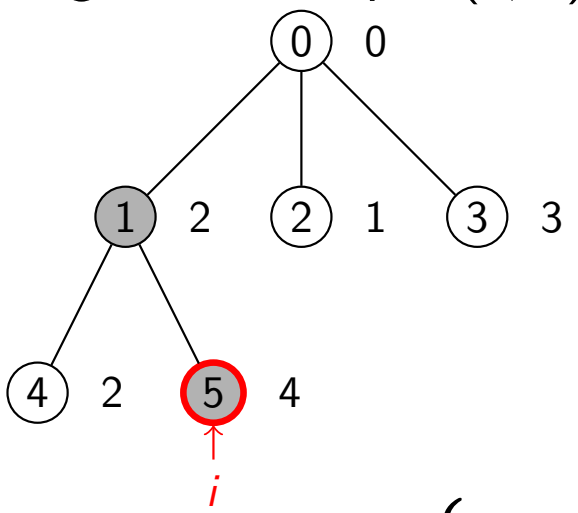
$$p^{(1)}_{(r,\Lambda)}(n,d) = \mathbb{P}_{(r,\Lambda)}\left\{\begin{array}{l}\text{most recent event involves indi-}\\ \text{vidual bearing mutation}\end{array}\right\}$$

Probability can be computed

- Kingman case: explicit formula (HUW (2008))
- Λ case: numerically, using recursion

Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$



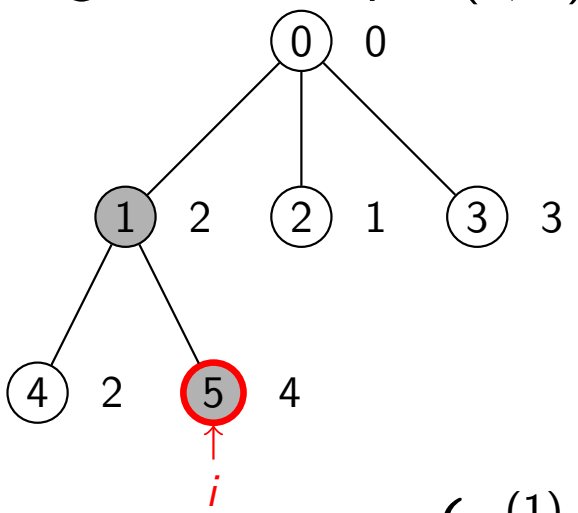
put

$$u_{i,m} = \begin{cases}$$

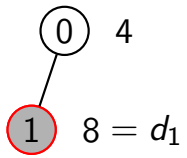
where mutation m is present in d_m individuals.

Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$



carrying:



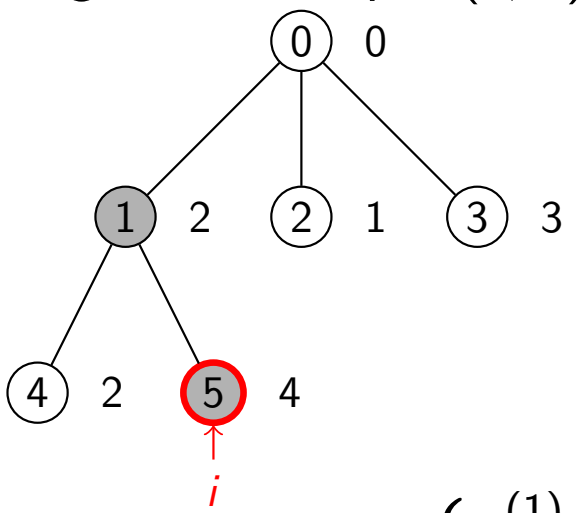
put

$$u_{i,m} = \begin{cases} p^{(1)}_{(r,\Lambda)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{if } i \text{ bears } m \\ \end{cases}$$

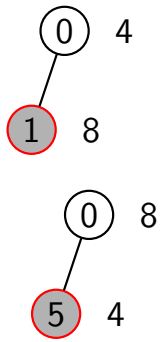
where mutation m is present in d_m individuals.

Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$



carrying:



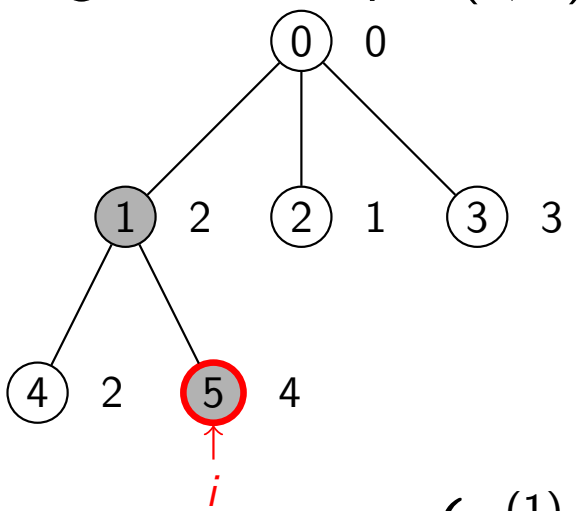
put

$$u_{i,m} = \begin{cases} p^{(1)}_{(r,\Lambda)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{if } i \text{ bears } m \\ \end{cases}$$

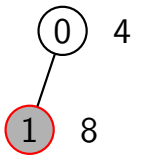
where mutation m is present in d_m individuals.

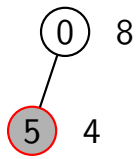
Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$

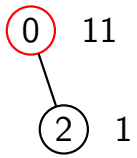


carrying:





not carrying:



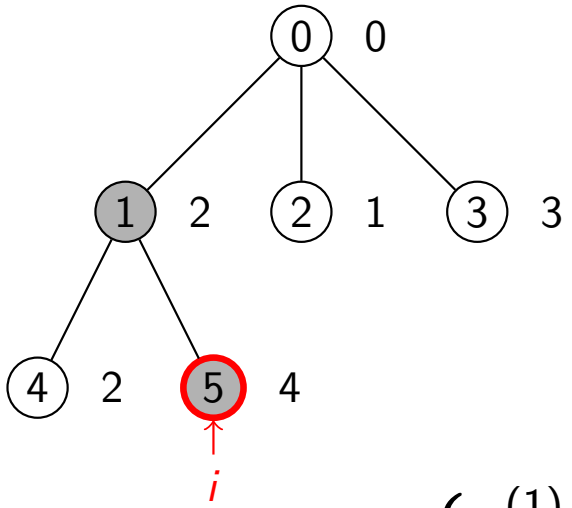
put

$$u_{i,m} = \begin{cases} p^{(1)}_{(r,\Lambda)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{if } i \text{ bears } m \\ \left(1 - p^{(1)}_{(r,\Lambda)}(n, d_m)\right) \cdot \frac{n_i}{n - d_m} & \text{otherwise,} \end{cases}$$

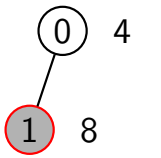
where mutation m is present in d_m individuals.

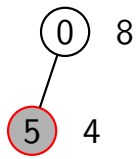
Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$

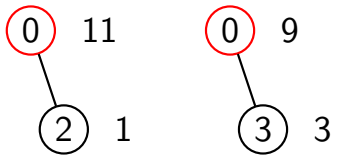


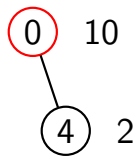
carrying:





not carrying:





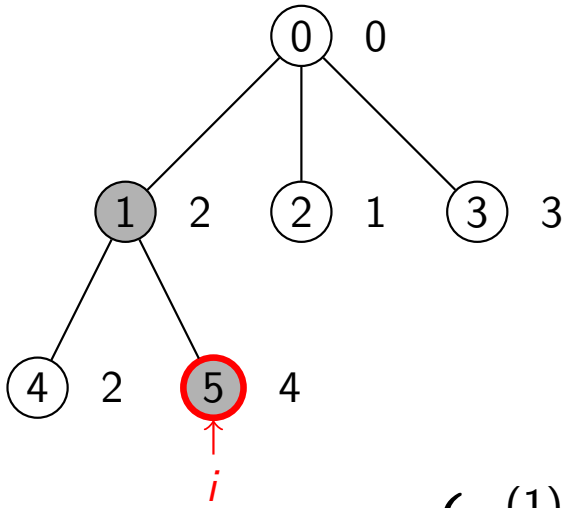
put

$$u_{i,m} = \begin{cases} p^{(1)}_{(r,\Lambda)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{if } i \text{ bears } m \\ \left(1 - p^{(1)}_{(r,\Lambda)}(n, d_m)\right) \cdot \frac{n_i}{n - d_m} & \text{otherwise,} \end{cases}$$

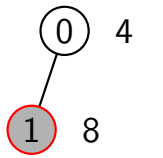
where mutation m is present in d_m individuals.

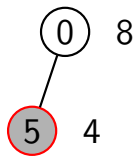
Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

For a general sample $(\mathbf{t}, \mathbf{n})$

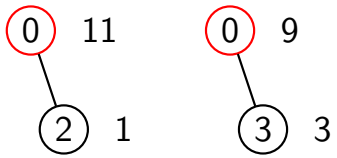


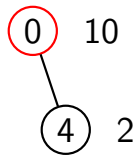
carrying:





not carrying:





put

$$u_{i,m} = \begin{cases} p^{(1)}_{(r,\Lambda)}(n, d_m) \cdot \frac{n_i}{d_m} & \text{if } i \text{ bears } m \\ \left(1 - p^{(1)}_{(r,\Lambda)}(n, d_m)\right) \cdot \frac{n_i}{n - d_m} & \text{otherwise,} \end{cases}$$

where mutation m is present in d_m individuals. Propose type i according to

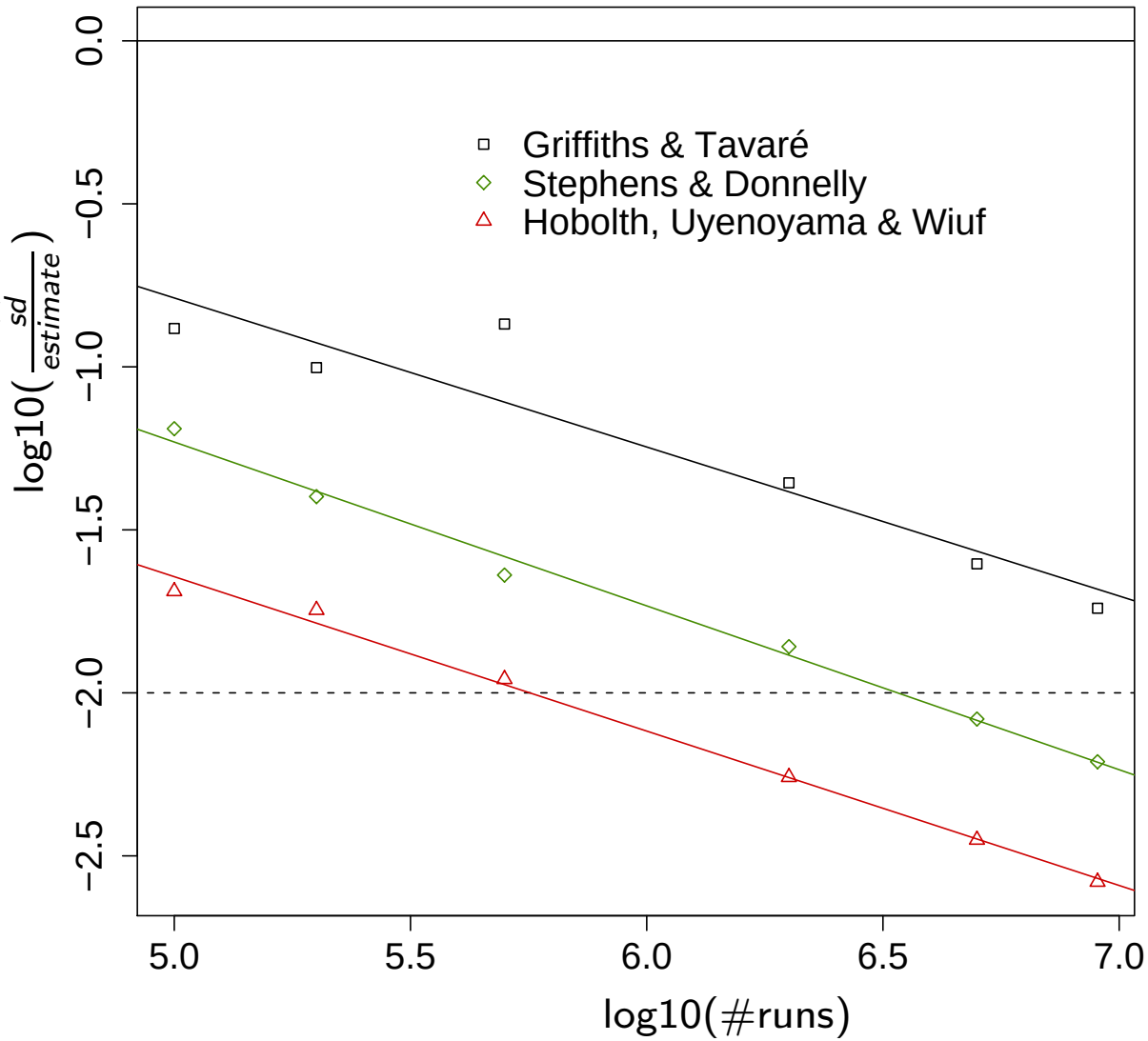
$$q_{\Lambda\text{-HUW}}\left(i|(\mathbf{t}, \mathbf{n})\right) \propto \begin{cases} \sum_m u_{i,m} & \text{if } i \text{ is allowed to act} \\ 0 & \text{otherwise.} \end{cases}$$

Hobolth, Uyenoyama & Wiuf's (2008) idea contd.

If proposed type i

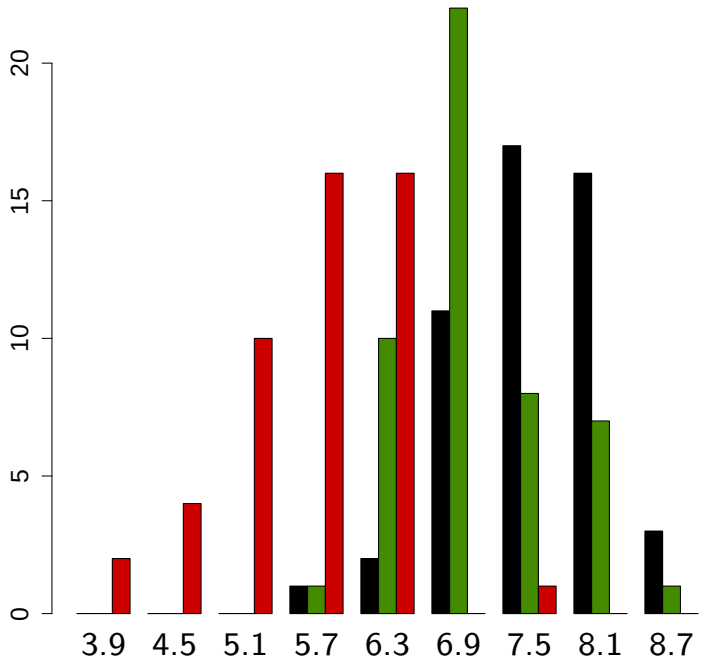
- is singleton: remove "outmost" mutation,
- has $n_i \geq 2$: merger inside type i.
  - Kingman case: merge two lines
  - Λ-case: propose $\ell + 1$-merger w.p. $\propto \mathbb{P}_{r,\Lambda}\left\{ \begin{array}{c|c} \begin{array}{cc} 0 & n \\ 1 & d_o - l \end{array} & \begin{array}{cc} 0 & n \\ 1 & d_o \end{array} \end{array} \right\}$

Artifical sample of size 12 analysed with $r = 1$ and $\alpha = 1.5$:

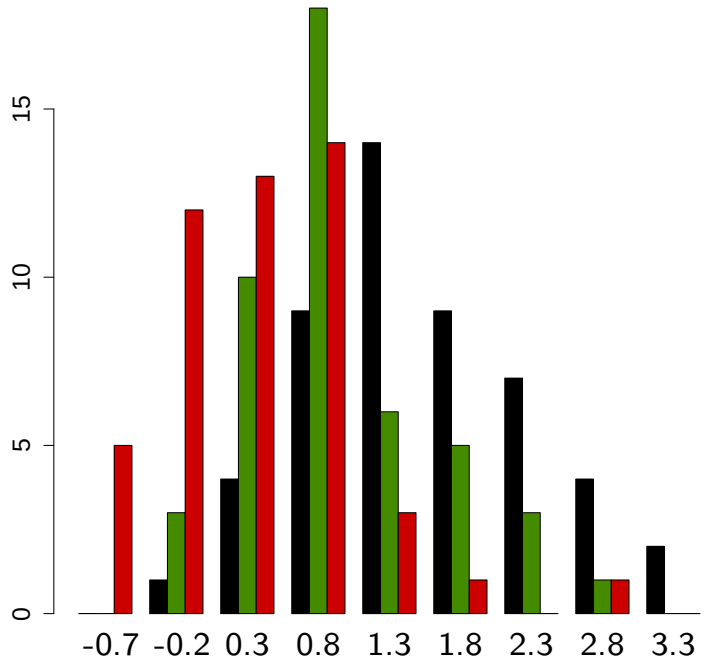


Performance

Simulated 50 samples of size 15 with $r = 2$ and $\alpha = 1.5$. Analysed with $r = 1$ and $\alpha = 1.5$. Time needed to get relative error below 0.01:



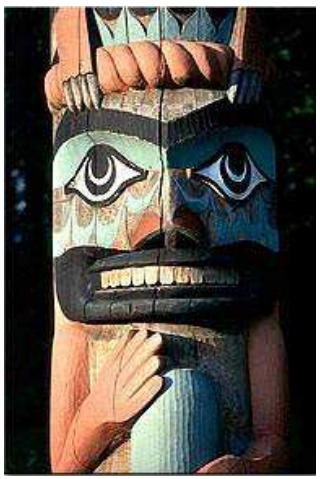
(a) measured in log10(# runs of MC)

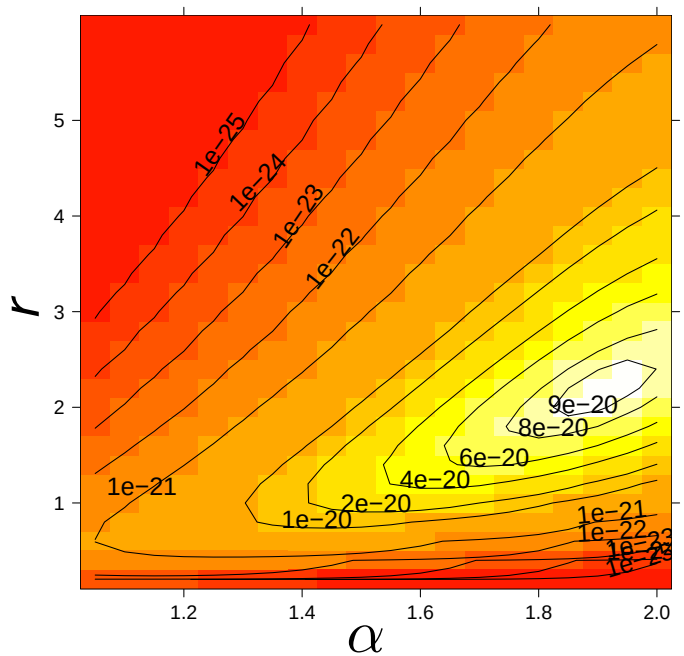


(b) measured in log10(seconds)

Dataset from Ward et al, Extensive Mitochondrial Diversity Within a Single Amerindian Tribe, *PNAS* 1991

Analysis with Beta-Coalescent:

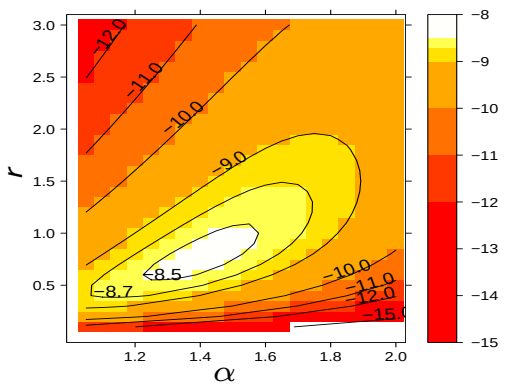




Mitochondrial control region from 55 female Nuu-Chah-Nulth:
$\widehat{\alpha}_{\mathrm{ML}} = 1.9$, $\widehat{r}_{\mathrm{ML}} = 2.2$
(Sample as edited in Griffiths & Tavaré, *Stat. Sci.*, 1994)

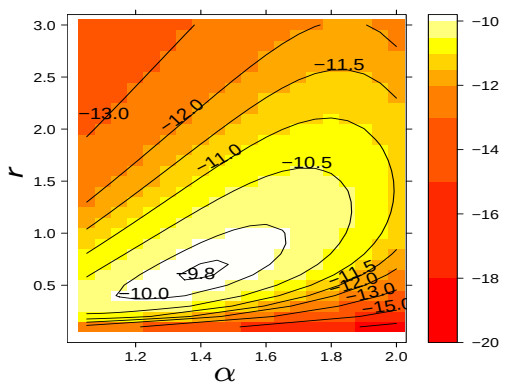
Genetic variation at the mitochondrial cyt *b* locus of Atlantic cod: log-likelihood surfaces

Carr & Marshall 1991



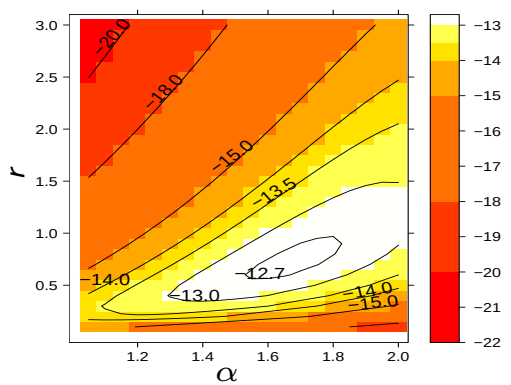
$\widehat{\alpha}_{\mathrm{ML}} = 1.4, \widehat{r}_{\mathrm{ML}} = 0.8, \widehat{\alpha}_{\mathrm{BBS}} = 1.4$

Pepin & Carr 1993



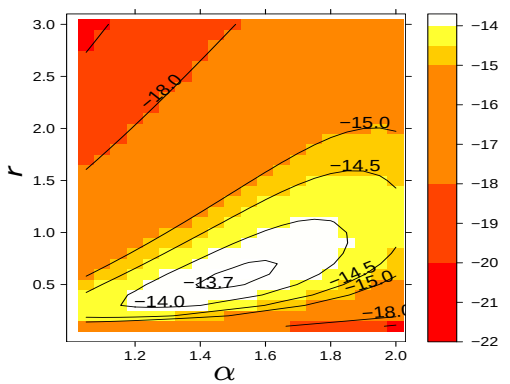
$\widehat{\alpha}_{\mathrm{ML}} = 1.4, \widehat{r}_{\mathrm{ML}} = 0.6, \widehat{\alpha}_{\mathrm{BBS}} = 1.5$

Árnason & Palsson 1996



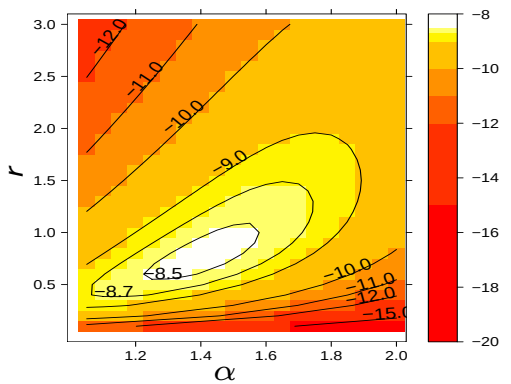
$\widehat{\alpha}_{\mathrm{ML}} = 1.65, \widehat{r}_{\mathrm{ML}} = 0.7, \widehat{\alpha}_{\mathrm{BBS}} = 1.5$

Árnason et al 1998



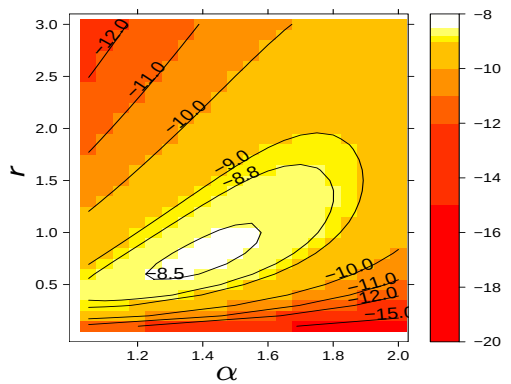
$\widehat{\alpha}_{\mathrm{ML}} = 1.55, \widehat{r}_{\mathrm{ML}} = 0.6, \widehat{\alpha}_{\mathrm{BBS}} = 1.8$

Árnason et al 2000



$\widehat{\alpha}_{\mathrm{ML}} = 1.4, \widehat{r}_{\mathrm{ML}} = 0.8, \widehat{\alpha}_{\mathrm{BBS}} = 1.5$

Sigurgíslason & Árnason 2003



$\widehat{\alpha}_{\mathrm{ML}} = 1.4, \widehat{r}_{\mathrm{ML}} = 0.8, \widehat{\alpha}_{\mathrm{BBS}} = 1.44$

"α-effective population size" — do the figures make sense?

In Schweinsberg's model, we have

$$\text{pair coalescence prob. } c_N \sim C \times N^{1-\alpha}$$

($C = \alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}$, where $m =$ mean of X_1)

"α-effective population size" — do the figures make sense?

In Schweinsberg's model, we have

$$\text{pair coalescence prob. } c_N \sim C \times N^{1-\alpha}$$

($C = \alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}$, where $m =$ mean of X_1)

Let $\mu =$ mutation rate per gen. at the considered locus, then $\dfrac{\mu}{c_N} \approx r$

"α-effective population size" — do the figures make sense?

In Schweinsberg's model, we have

$$\text{pair coalescence prob. } c_N \sim C \times N^{1-\alpha}$$

($C = \alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}$, where $m =$ mean of X_1)

Let $\mu =$ mutation rate per gen. at the considered locus, then $\dfrac{\mu}{c_N} \approx r$,

hence
$$N_{\text{eff},\alpha} \approx \left(\frac{r\alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}}{\mu}\right)^{1/(\alpha-1)}.$$

"α-effective population size" — do the figures make sense?

In Schweinsberg's model, we have

$$\text{pair coalescence prob. } c_N \sim C \times N^{1-\alpha}$$

($C = \alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}$, where $m =$ mean of X_1)

Let $\mu =$ mutation rate per gen. at the considered locus, then $\dfrac{\mu}{c_N} \approx r$, hence

$$N_{\text{eff},\alpha} \approx \left(\frac{r\alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}}{\mu}\right)^{1/(\alpha-1)}.$$

Using $\mu = 250 \times 1.85 \cdot 10^{-7}$ (Árnason, 2004), $\widehat{\alpha} = 1.5$, $\widehat{r} = 1$ (and, ad hoc, $m = 2$), this gives

$$\widehat{N}_{\text{eff},\alpha=1.5} \approx 3.2 \cdot 10^8$$

Árnason (2004) writes: "... the actual population size [of atlantic cod] is not $< 10^9$ and probably one or two orders of magnitude larger."

"α-effective population size" — do the figures make sense?

In Schweinsberg's model, we have

$$\text{pair coalescence prob. } c_N \sim C \times N^{1-\alpha}$$

($C = \alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}$, where $m =$ mean of X_1)

Let $\mu =$ mutation rate per gen. at the considered locus, then $\dfrac{\mu}{c_N} \approx r$,

hence $$N_{\text{eff},\alpha} \approx \left(\frac{r\alpha\Gamma(\alpha)\Gamma(2-\alpha)m^{-\alpha}}{\mu}\right)^{1/(\alpha-1)}.$$

Using $\mu = 250 \times 1.85 \cdot 10^{-7}$ (Árnason, 2004), $\widehat{\alpha} = 1.5$, $\widehat{r} = 1$ (and, ad hoc, $m = 2$), this gives

$$\widehat{N}_{\text{eff},\alpha=1.5} \approx 3.2 \cdot 10^8$$

Árnason (2004) writes: "... the actual population size [of atlantic cod] is not $< 10^9$ and probably one or two orders of magnitude larger."

By contrast, using a Wright-Fisher model and $N_{\text{eff},\alpha=2} \times \mu = r$, we have (using $\widehat{r}_{\alpha=2} = 2$): $$\widehat{N}_{\text{eff},\alpha=2} \approx 4.3 \cdot 10^4.$$

Summary & Outlook

Eldon & Wakeley, *Genetics* 2006, wrote

> For many species, *the* coalescent with multiple mergers might be a better null model than Kingman's coalescent.

Summary & Outlook

Eldon & Wakeley, *Genetics* 2006, wrote

> For many species, *the* coalescent with multiple mergers might be a better null model than Kingman's coalescent.

- For panmictic fixed-size discrete generations populations, haploid neutral one-locus theory is "mathematically complete".
- Tools for estimation exist, results point towards "non-Kingman-ness" in certain cases.
- Statistical properties of estimators?
- speed-up of computer-intensive methods?
  - combinations between IS-methods possible
  - "Double-HUW" scheme: ask all pairs of mutations what to do
- A good class of alternative models? In particular, true diploid models?
- Application to scenarios with selection?

Thank you for your attention!