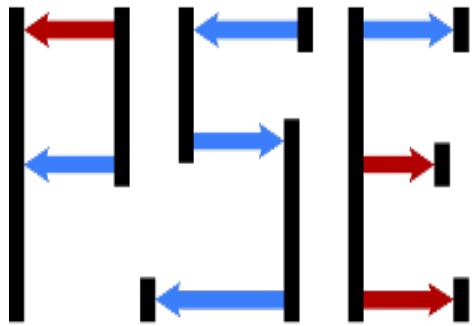
Gene genealogies in highly fecund populations with skewed offspring distribution

Bjarki Eldon
Museum für Naturkunde
Leibniz-Institut für Evolutions- und Biodiversitätsforschung an der Humboldt Universität, Berlin

museum für
naturkunde
berlin



PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

Overview

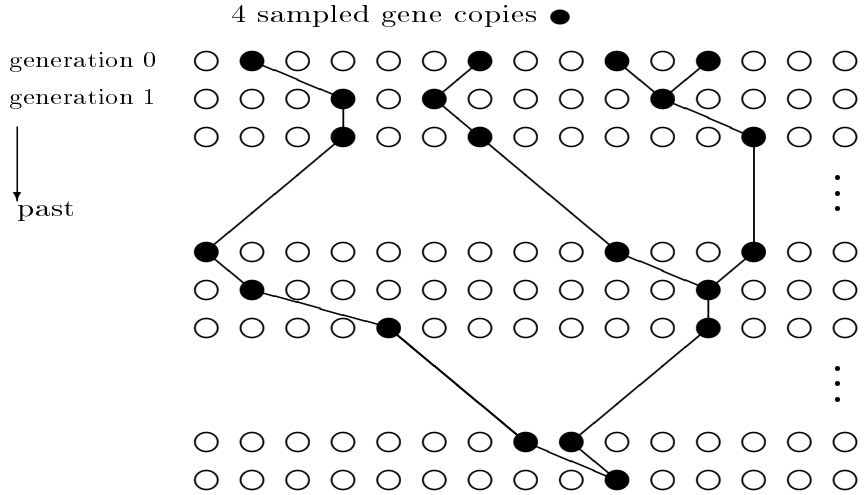
- models of high fecundity and skewed offspring distributions
- multiple merger coalescent processes
- site-frequency spectrum
- current projects

Models of small numbers of offspring

- **Arguably the most commonly employed population model - the Wright-Fisher model - may be classified as a *model of small numbers of offspring* which means that - in a large population - individuals have negligible chance of contributing huge numbers - on the order of the population size (N) - of offspring to the next generation**
- **Indeed, if all the moments of an exchangeable Cannings model are finite, one obtains the Kingman coalescent in the limit $N \to \infty$ (see e.g. Möhle & Sagitov (2001))**

The Kingman coalescent from the WF-model

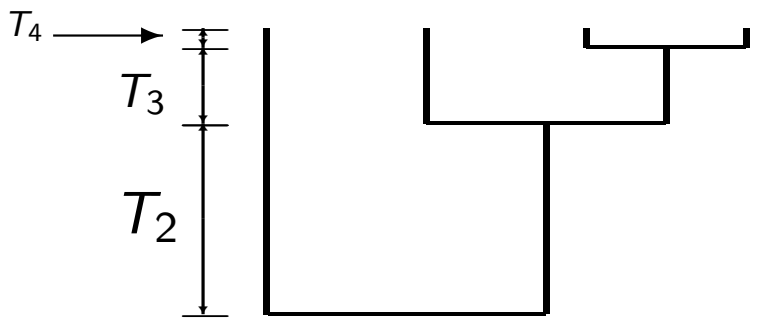
A realisation of ancestral relations in a WF population



⋮ } steps where 'nothing' happens

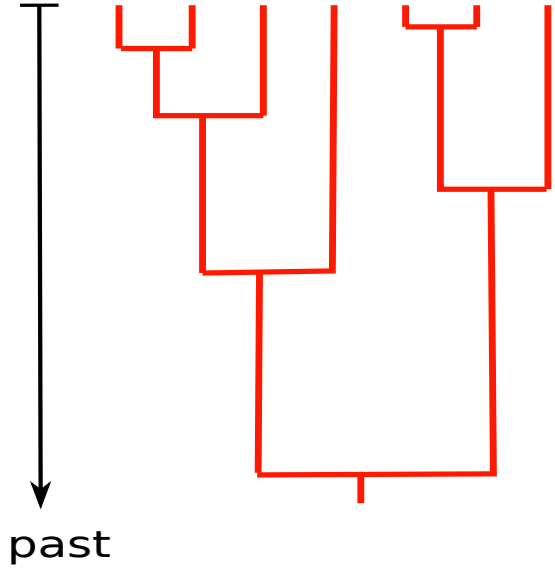
⋮ } steps where 'nothing' happens

The extracted gene genealogy



the times T_j are independent exponentials with rate $\binom{j}{2}$

Kingman coalescent



Each pair of ancestral lineages coalesces with rate 1. The genealogy of a class of Cannings exchangeable population models with finite moments and operating on *different timescales* converges (as $N \to \infty$) to the Kingman coalescent. The Kingman coalescent is the usual 'null model' in population genetics.

High fecundity and skewed offspring distributions

- ***High fecundity and skewed offspring distributions*** **- where individuals can have huge numbers of offspring with non-negligible probability (as $N \to \infty$) - give rise to multiple merger coalescent processes (e.g.** Hedgecock & Pudovkin (2011) **review in** *Bull. Marine Sci.* **)**
- **Potentially applicable to some marine organisms:** ***Pacific oysters*** **(Hedgecock, 1994; Beckenbach, 1994; Hedgecock, Chow, & Waples, 1992);** ***flat oyster*** **(Harrang etal, 2013);** ***white sea bream*** **(Planes & Lenfant, 2002);** ***bicolor damselfish*** **(Christie etal, 2010);** ***Atlantic cod*** **(Árnason, 2004)**
- **forest trees such as** ***European aspen*** **(Ingvarsson, 2010)**
- **other populations ?**

Skewed offspring distributions

Schweinsberg **(2003)**
The tail probability for the number of `viable` (haploid) offspring of an individual $i \in \{1, \ldots, N\}$, with $\mathbb{E}[X_i] > 1$, is

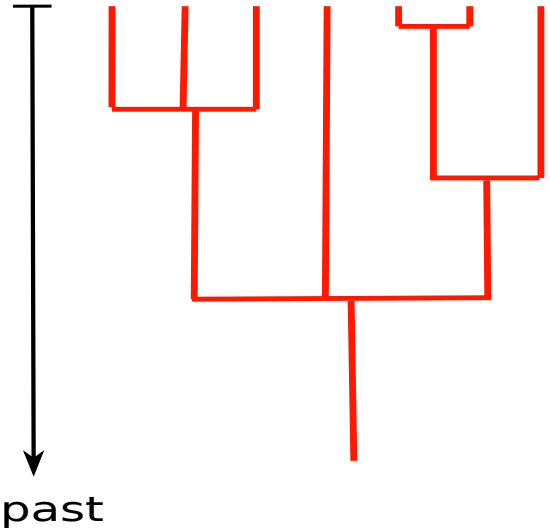
$$\mathbb{P}(X_i \geq k) \sim \frac{1}{k^\alpha}, \quad k \to \infty, \quad \alpha > 0$$

E. & Wakeley **(2006)**
A modified haploid Moran model, in which the number of offspring of the reproducing individual is

$$\mathbb{P}(X = k) = (1 - \varepsilon_N)\delta_1(k) + \varepsilon_N \delta_{\lfloor \psi N \rfloor}(k), \quad \psi \in (0, 1]$$

Also population models with skewed offspring distributions by Huillet & Möhle **(2011),** Sargsyan & Wakeley **(2008),** E. **(preprint)**

Λ coalescent



Donnelly and Kurtz (1999), Pitman (1999), Sagitov (1999) independently derive a multiple merger coalescent process, allowing each collection of $k \in \{2, \ldots, b\}$ active ancestral lineages to coalesce at the same time with rate

$$\lambda_{b,k}^{(\Lambda)} = \int_0^1 x^k (1-x)^{b-k} x^{-2} \Lambda(dx)$$

where Λ is a finite measure on $[0, 1]$.

Examples of Λ-coalescents

The most commonly employed examples of Λ-coalescents:

The Beta$(2-\alpha,\alpha)$-coalescent (Schweinsberg 2003) occurs when $\alpha \in [1,2)$:

$$\lambda_{b,k} = \frac{B(k-\alpha, b-k+\alpha)}{B(2-\alpha,\alpha)}$$

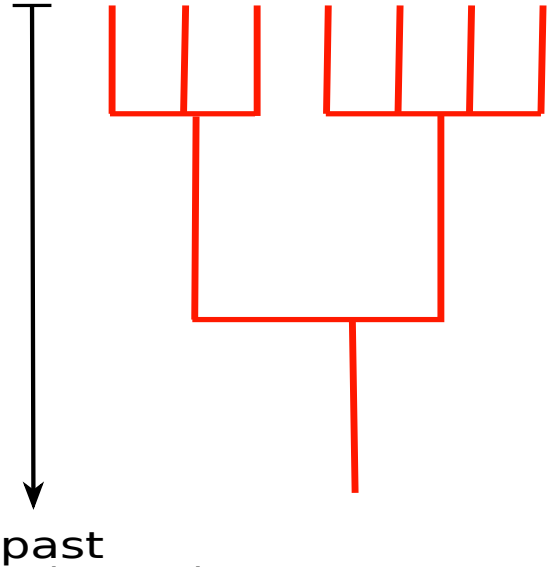
where $B(\cdot,\cdot)$ is the beta-function

The Dirac-coalescent (E. & Wakeley 2006) occurs when $N^2\varepsilon_N \to \infty$:

$$\lambda_{b,k} = \psi^{k-2}(1-\psi)^{b-k}$$

The *timescales* are different from the usually assumed Kingman timescale: 1 coalescent time unit corresponds to $1/c_N$ generations

Ξ-coalescent



Schweinsberg **(2000) introduces the Ξ-coalescents, where many multiple mergers may occur *simultaneously***

Example: **denote size of merger** $i \in [r]$, $r \in [M]$, $M \geq 2$, M **is level of ploidy, with** k_i, $\underline{k} = (k_1, \ldots, k_r)$, $|\underline{k}| = k_1 + \cdots + k_r$, **then the rate of a** $\underline{k}$**-simultaneous merger is given by**

$$\lambda_{b;\underline{k}}^{(\Xi)} = \sum_{\ell=0}^{n-|\underline{k}|} \binom{n-|\underline{k}|}{\ell} \frac{(M)_{r+\ell}}{M^{\ell+|\underline{k}|}} \lambda_{b,|\underline{k}|}^{(\Lambda)}, \quad 2 \leq |\underline{k}| \leq b$$

(Blath etal, 2015; preprint)

The site-frequency spectrum

Exact likelihoods are computationally inefficient (e.g. Birkner & Blath, 2008); hence we consider approximate likelihoods based on the *site-frequency spectrum*

The site-frequency spectrum $\underline{\xi}^{(n)} = \left(\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}\right)$ is a simple summary statistic of the full DNA sequence data, yet holds valuable information about genetic variation among individuals

Example of 3 DNA sequences with 5 segregating sites:

$$1 : 0020000000$$

$$2 : 0020001001$$

$$3 : 0000100100$$

The (unfolded) site-frequency spectrum for this example is

$$\underline{x}^{(3)} = \left(x_1^{(3)}, x_2^{(3)}\right) = (4, 1)$$

Exact expected site-frequency spectrum

Let $\theta > 0$ denote the appropriately scaled (note *different timescales*) mutation rate

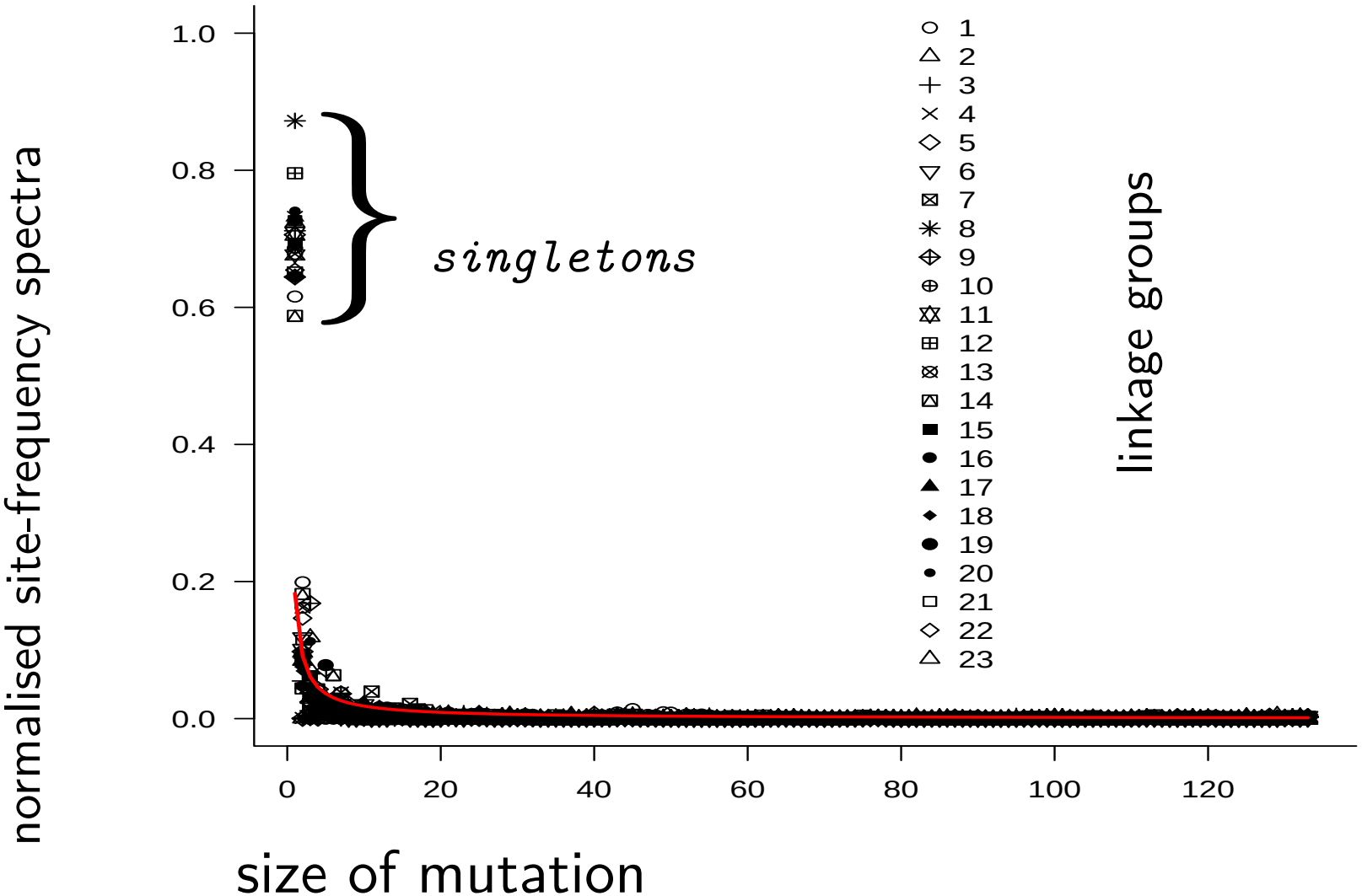
- **For the Kingman-coalescent, $\mathbb{E}^{(K)}\left[\xi_i^{(n)}\right] = \frac{\theta}{i}$ (Fu, 1995)**
- **For Λ-coalescents, $\mathbb{E}^{(\Lambda)}\left[\xi_i^{(n)}\right] =$ a recursion (Birkner etal, 2013)**
- **For Ξ-coalescents, $\mathbb{E}^{(\Xi)}\left[\xi_i^{(n)}\right] =$ more complicated recursion (Blath etal, (2016; preprint))**
- **An efficient method by Spence etal (2016; preprint) for Ξ-coalescents**

Current projects

- **Multi-loci ancestral selection graphs which admit simultaneous mergers to model the evolution of a trait affected by many loci in an effort to understand rapid adaptation**
- **Quantifying the association between the site-frequency spectra at unlinked loci in an effort to test for the presence of multiple mergers, and in a study design**
- **Tests for selection at loci in a background of multiple mergers**
- **Effect of large sample size and truncated offspring distributions on coalescent processes**
- **Extending the site-frequency spectrum (SFS) to *individual*-based SFS**

Normalised SFS from all linkage groups of Atlantic cod (Einar Árnason)

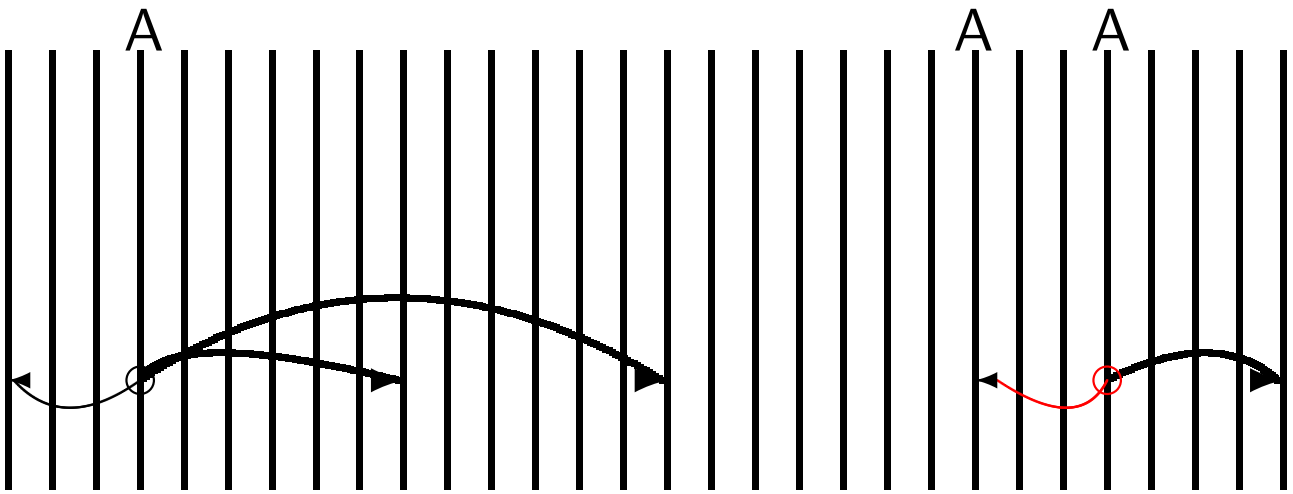
sample size 134, segregating sites $\approx 10^4$



Multi-loci ancestral selection graphs (ASG)

(joint work with Wolfgang Stephan)

Two allelic types (a, A); A is beneficial



we derive multi-loci ancestral selection graphs which admit (simultaneous) multiple mergers of ancestral lineages for various forms of selection

(see Etheridge *etal* (2010) for a single-locus Λ-ASG)

Quantifying association between SFS at unlinked loci

We use the multivariate association statistic $\mathcal{I}_n$ of Bakirov etal (2006): Let $X_j \in \mathbb{R}^p$ and $Y_j \in \mathbb{R}^q$ for $j \in [n]$ be a random sample from $(X, Y) \in \mathbb{R}^{p+q}$ with $X \sim F_1$, $Y \sim F_2$ and $(X, Y) \sim F$ and we test the hypothesis

$$F = F_1 F_2 \quad \textbf{vs} \quad F \neq F_1 F_2$$

We form all unordered pairs of concatenated SFS $\left(\underline{\xi}^{(i)}, \underline{\xi}^{(j)}\right)$, $i < j$, of SFS for ℓ unlinked loci, and with $n = \binom{\ell}{2}$, compute the statistic (Bakirov etal (2006))

$$\mathcal{I}_n = \sqrt{\frac{2\overline{z} - z_d - z}{x + y - z}} \in [0, 1]$$

where $\overline{z}, z_d, z, x, y$ are averages of the Euclidean norm in $\mathbb{R}^{p+q}$.

Extending the site-frequency spectrum (SFS)

The standard site-frequency spectrum (SFS) $\underline{\xi}^{(n)} = \left(\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}\right)$ **can be extended into the *sequence-specific* SFS**

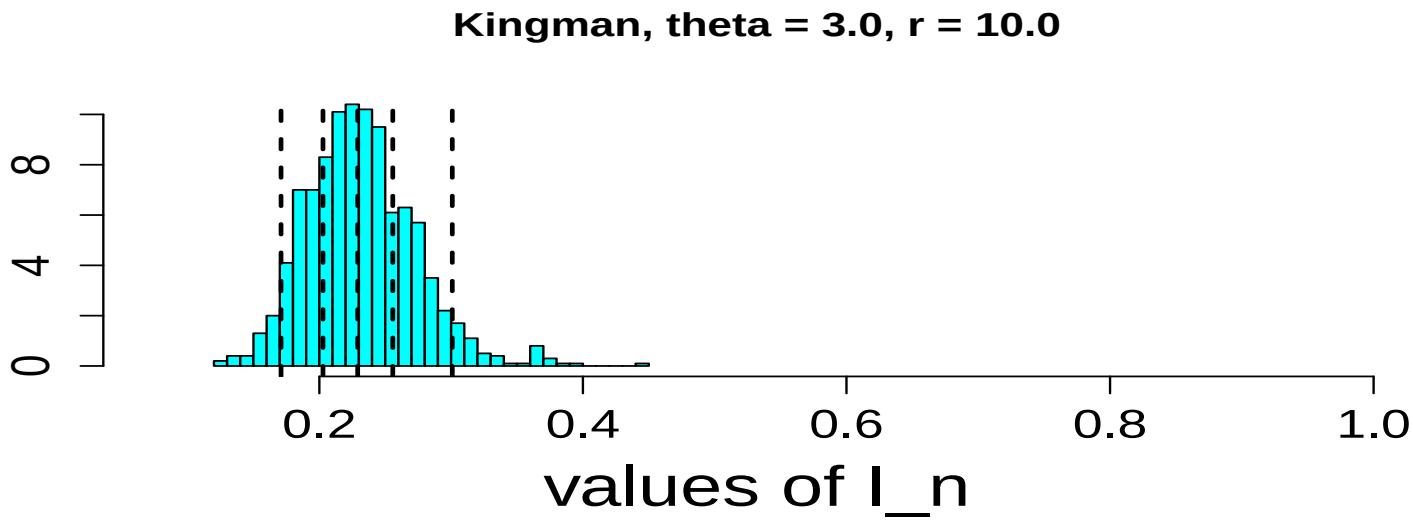
$$\left(\xi_{j,i}^{(n)}\right) = \text{ the number of polymorphic sites on sequence } j \text{ shared with } i-1 \text{ other sequences}$$

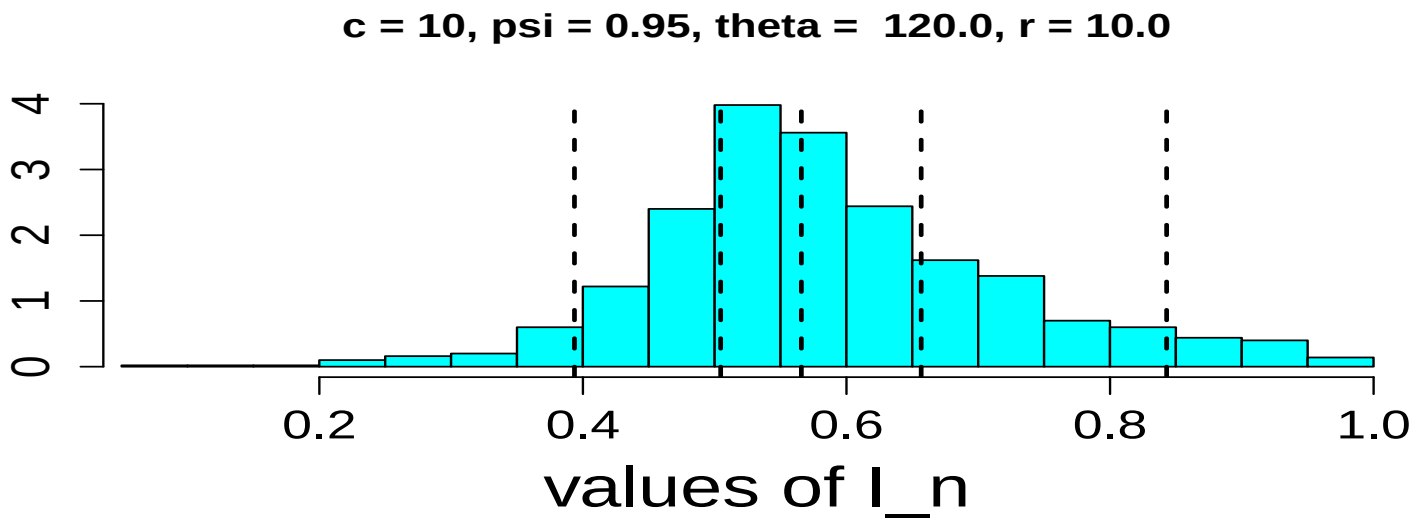
$$\xi_i^{(n)} = \frac{1}{i} \sum_{j \in [n]} \xi_{j,i}^{(n)}$$

we may also consider joint probabilities, such as

$$\mathbb{P}\left(\xi_{j,1}^{(n)} > 0, \quad \xi_{j,2}^{(n)} > 0\right)$$

Estimating the distribution of $\mathcal{I}_n$ for 2 loci





Tests of selection in a multiple merger background

Given data for ℓ unlinked loci in two populations, assuming the loci are independent the likelihood function on segregating sites (S) and pairwise differences (Δ):

$$L(\underline{\vartheta}, \underline{\theta}; \underline{s}, \underline{d}) = \prod_{j \in [\ell]} \mathbb{P}^{(\vartheta_j, \theta_j)} (S_j = s_j) \, \mathbb{P}^{(\vartheta_j, \theta_j, \tau)} (\Delta_j = d_j) \quad (*)$$

but $(*)$ is a function of both $\underline{\vartheta}$, $\underline{\theta}$, and τ the time of separation -
we replace segregating sites with normalised site-frequency spectrum $\underline{\zeta}^{(j)}$, and consider normalised joint site-frequency spectrum between the two populations:

$$L(\underline{\vartheta}; \underline{z}, \underline{y}) = \prod_{j \in [\ell]} \mathbb{P}^{(\vartheta_j)} \left(\underline{\zeta}^{(j)} = \underline{z}_j \right) \mathbb{P}^{(\vartheta_j, \tau)} \left(\underline{\chi}^{(j)} = y_j \right) \quad (**)$$

and $(**)$ is only a function of $\underline{\vartheta}$ and τ

Large sample size and truncated offspring distribution

(joint work with Alison Etheridge and Jerome Kelleher in Oxford, and Matthias Hammer at TU Berlin)

- **The standard approach when deriving coalescent processes from population models is to assume that sample size is fixed, and much smaller than the (effective) population size; what happens when sample size is assumed to be *on the order of* the (effective) population size?**
- **we also consider truncated offspring distributions**

$$\mathbb{P}(X_i \geq k) = Ck^{-\alpha}\mathbb{I}(1 \leq k \leq \phi(N)), \quad k \geq 1$$

and, depending on ϕ, obtain Lambda-coalescents where Λ is associated with truncated beta-distributions

- **Extend the model to polyploidy and multiple loci with recombination**

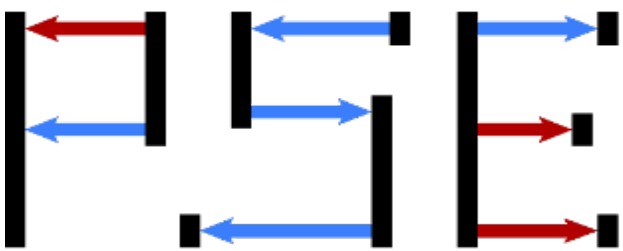
References

- Blath, Cronjäger, E., Hammer: **The site-frequency spectrum associated with Xi-coalescents. Preprint (biorxiv)**
- E.: **Age of an allele and gene genealogies of nested subsamples for populations admitting large offspring numbers. Preprint (arxiv)**
- E., Birkner, Blath, Freund: **Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?, *Genetics 199*, 841-856, (2015)**
- Birkner, Blath, E.: **Statistical properties of the site-frequency spectrum associated with Lambda-coalescents, *Genetics 195*, 1037-1053, (2013)**
- Birkner, Blath, E.: **An ancestral recombination graph for diploid populations with skewed offspring distribution, *Genetics 193*, 255-290, (2013)**

Funding and collaborators

Joint work with *Matthias Birkner* (JGU Mainz)
Jochen Blath, Matthias Hammer (TU Berlin)
Fabian Freund (University of Hohenheim)
Mathias Cronjäger (University of Oxford)

DFG SPP1819: Rapid Evolutionary Adaptation



PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

museum für
naturkunde
berlin