Modeling the genealogy of populations using coalescents with multiple mergers

by Jason Schweinsberg
University of California at San Diego

partly joint work with Julien Berestycki and Nathanaël Berestycki

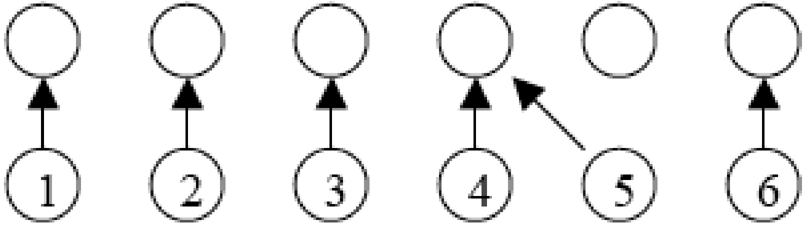
Outline of Talk

1. The Wright-Fisher model and Kingman's coalescent
2. Coalescents with multiple mergers
3. Populations with large families or selection
4. Quantities of interest
   a) Segregating sites
   b) Site frequency spectrum
   c) Allele frequency spectrum
5. Implications for statistical inference

The Wright-Fisher Model

One of the earliest models in population genetics, goes back to Fisher (1921) and Wright (1930).

- The population has fixed size N.
- Generations do not overlap.
- Each member of the population has one parent, chosen at random from the individuals in the previous generation.



Sample n individuals from generation 0. Let $\Psi_N(m)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim j$ if and only if ith and jth sampled individuals have the same ancestor in generation $-m$.

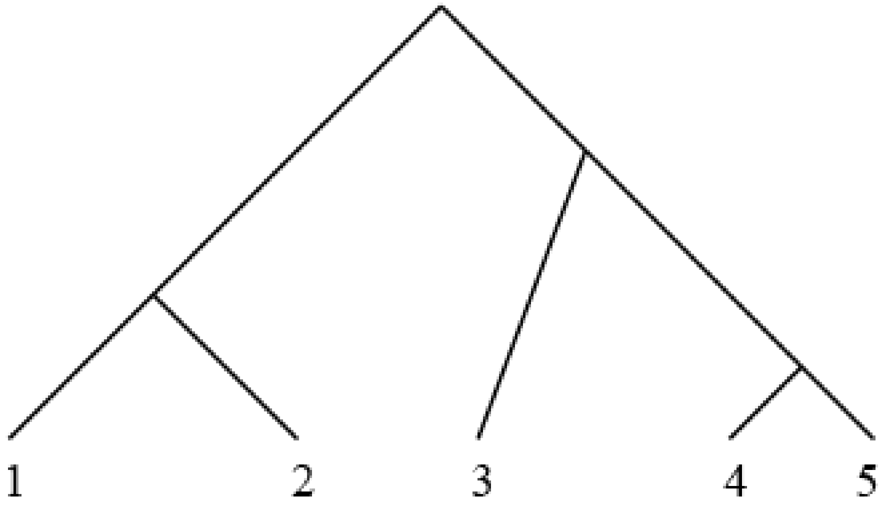
As $N \to \infty$, the processes $\Psi_N = (\Psi_N(\lfloor Nt \rfloor), t \geq 0)$ converge to Kingman's coalescent.

Kingman's Coalescent (Kingman, 1982)

Continuous-time Markov chain on set of partitions of $\{1, \ldots, n\}$.

Only two lineages merge at a time, each pair of lineages merges at rate one.

When there are k lineages, the distribution of the time until the next merger is exponential with rate $k(k-1)/2$. Then two randomly chosen lineages merge.



$\{1, 2, 3, 4, 5\}$

$\{1, 2\}, \{3, 4, 5\}$

$\{1, 2\}, \{3\}, \{4, 5\}$

$\{1\}, \{2\}, \{3\}, \{4, 5\}$

$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$

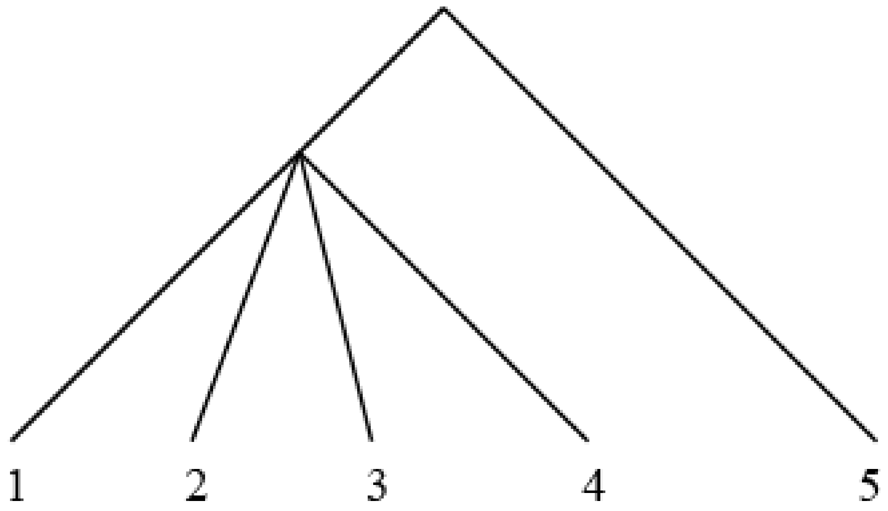
One time unit in Kingman's coalescent represents N generations.

Coalescents with multiple mergers (Λ-coalescents)

Continuous-time Markov chain on set of partitions of $\{1, \ldots, n\}$.

More than two ancestral lines can merge at a time.

First studied by Pitman (1999) and Sagitov (1999).



Applications of coalescents with multiple mergers:

- Large family sizes, as may occur with some marine species. (Sagitov, 1999; Möhle-Sagitov, 2001).
- Natural selection (Durrett-Schweinsberg, 2005).

Definition of Λ-coalescents

Let π be a partition of $\{1, \dots, n\}$ into blocks $B_1, \dots, B_j$. Let $p \in (0, 1]$. A p-merger of π is obtained as follows:

- Let $\xi_1, \dots, \xi_j$ be i.i.d. Bernoulli(p).
- Merge the blocks B_i such that $\xi_i = 1$.

Coalescents can be described in terms of a finite measure Λ on $[0, 1]$. Write $\Lambda = a\delta_0 + \Lambda_0$, where $\Lambda_0(\{0\}) = 0$. Transitions in the Λ-coalescent are as follows:

- Each pair of blocks merges at rate a.
- Construct a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity $dt \times p^{-2}\Lambda_0(dp)$. If (t, p) is a point of this Poisson process, then a p-merger occurs at time t.

When there are b blocks, let $\lambda_{b,k}$ denote the rate of a transition in which k blocks merge into one. Then, for $2 \leq k \leq b$,

$$\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k} \, \Lambda(dp).$$

Genealogy of Galton-Watson processes

Consider the following population model:

- Population size N in each generation.
- Numbers of offspring $\xi_1, \dots, \xi_N$ of the N individuals are i.i.d. with $P(\xi_i \geq k) \sim Ck^{-\alpha}$, where $\alpha \geq 1$, and $E[\xi_i] > 1$.
- Obtain the next generation by sampling N offspring.

Theorem (Schweinsberg, 2003):

- If $\alpha \geq 2$, genealogies converge to Kingman's coalescent.
- If $1 \leq \alpha < 2$, limit is the Λ-coalescent,

$$\Lambda(dx) = \frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}\,dx$$

  is the Beta$(2-\alpha, \alpha)$ distribution.

The case $\alpha = 1$ is the Bolthausen-Sznitman (1998) coalescent. Linked to random recursive trees (Goldschmidt-Martin, 2005) and Derrida's GREM (Bovier-Kurkova, 2007).

Idea of the proof $(1 < \alpha < 2)$

Let μ be the mean of the offspring distribution.

We get a p-merger with $p \geq x$ if

$$\frac{\xi}{\xi + N\mu} \geq x \qquad \Longleftrightarrow \qquad \xi \geq \frac{x}{1-x} \cdot N\mu$$

The probability of such a family in a given generation is

$$NP\left(\xi \geq \frac{x}{1-x} \cdot N\mu\right) \sim NC\left(\frac{x}{1-x} \cdot N\mu\right)^{-\alpha}.$$

The rate of such mergers in the Beta$(2-\alpha, \alpha)$-coalescent is

$$\frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \int_x^1 p^{-1-\alpha}(1-p)^{\alpha-1}\, dp = \frac{1}{\alpha\Gamma(\alpha)\Gamma(2-\alpha)} \left(\frac{x}{1-x}\right)^{-\alpha}.$$

A population model with selection

Brunet-Derrida-Mueller-Munier (2006, 2007)

- Population has fixed size N.
- Each individual has $k \geq 2$ offspring.
- The fitness of each offspring is the parent's fitness plus an independent random variable with distribution μ.
- Of the kN offspring, the N with the highest fitness survive to form the next generation.

Also studied by Bérard and Gouéré (2010), Durrett and Mayberry (2010), Durrett and Remenik (2009).

Conjectures of Brunet-Derrida-Mueller-Munier (2006, 2007):

- If two individuals are chosen from some generation, the number of generations back to their most recent common ancestor is $O((\log N)^3)$.
- If n individuals are sampled from some generation, their genealogy is governed by the Bolthausen-Sznitman coalescent.

Branching Brownian motion with absorption

- Begin with some configuration of particles in $(0, \infty)$.
- Each particle independently moves according to standard one-dimensional Brownian motion with drift $-\mu$.
- Each particle splits into two at rate 1.
- Particles are killed if they reach the origin.

Particles represent individuals in a population, and their position represents the fitness of the individual.

Theorem (Kesten, 1978): Starting with one particle at $x > 0$, this process dies out almost surely if $\mu \geq \sqrt{2}$. If $\mu < \sqrt{2}$, the number of particles grows exponentially with positive probability.

We take $\mu = \mu_N = \sqrt{2 - \frac{2\pi^2}{(\log N + 3 \log \log N)^2}}$.

Theorem (Berestycki-Berestycki-Schweinsberg, 2010): Fix a time $t > 0$. Choose n particles at random at time $t(\log N)^3$. Let $\Pi_N(s)$ be the partition of $\{1, \dots, n\}$ such that i and j are in the same block if and only if the ith and jth sampled particles have the same ancestor at time $(t - s/2\pi)(\log N)^3$. Under suitable initial conditions, the finite-dimensional distributions of $(\Pi_N(s), 0 \leq s \leq 2\pi t)$ converge as $N \rightarrow \infty$ to those of the Bolthausen-Sznitman coalescent.

Initial conditions are satisfied if $O(N)$ particles are placed in a relatively stable configuration.

Idea behind the multiple mergers:

- Occasionally, a particle gets very far to the right.
- This particle has a large number of surviving descendants, as the descendants avoid the barrier at zero.
- This leads to multiple mergers of ancestral lines.

Coalescents with mutations

Assume mutations happen along each lineage at rate θ.

Assume each mutation happens at different site on chromosome.

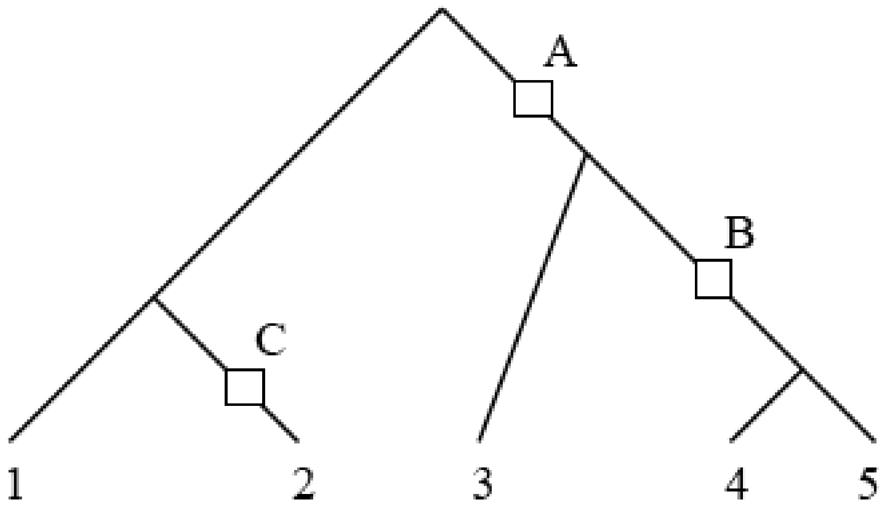
1: AC**G**CTAA**T**AGC**A**

2: AC**G**CTAA**T**AGC**T**

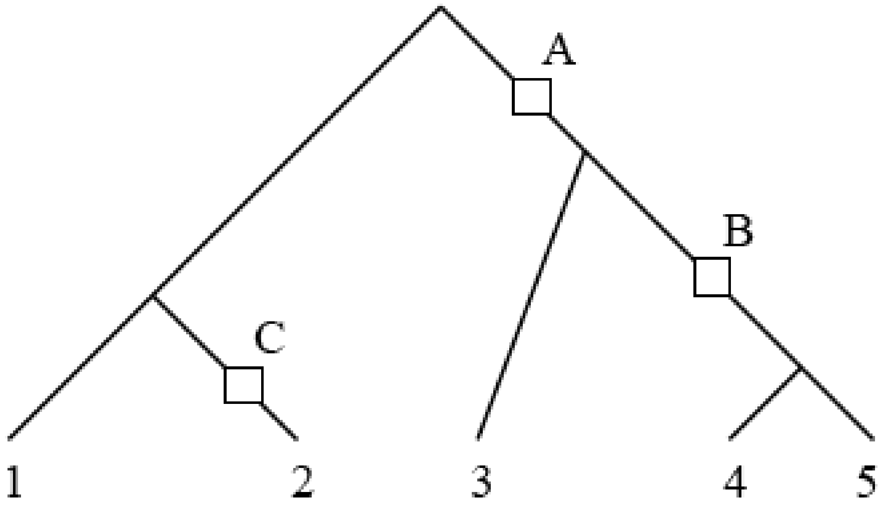
3: AC**C**CTAA**T**AGC**A**

4: AC**C**CTAA**C**AGC**A**

5: AC**C**CTAA**C**AGC**A**



Quantities of Interest



Segregating sites: $S_n =$ number of sites at which not all members of sample agree. Example: $S_n = 3$.

Allelic partition: blocks represent groups of individuals that got the same mutations. Example: $\Pi_n = \{\{1\}, \{2\}, \{3\}, \{4, 5\}\}$.
$A_n =$ number of blocks of Π_n (haplotypes). Example: $A_n = 4$.

Allele frequency spectrum: $N_{k,n} =$ number of blocks of size k in allelic partition. Example: $N_{1,5} = 3$, $N_{2,5} = 1$.

Site frequency spectrum: $M_{k,n} =$ number of mutations affecting k individuals. Example: $M_{1,5} = 1$, $M_{2,5} = 1$, $M_{3,5} = 1$.

Segregating sites

Kingman's coalescent:

$$E[S_n] = \theta \sum_{b=2}^{n} b \binom{b}{2}^{-1} = 2\theta \sum_{b=2}^{n} \frac{1}{b-1} \sim 2\theta \log n$$

and $(S_n - E[S_n])/\sqrt{\mathrm{Var}(S_n)} \Rightarrow N(0,1)$.

Beta$(2-\alpha,\alpha)$-coalescent with $1 < \alpha < 2$ (Berestycki-Berestycki-Schweinsberg, 2008):

$$\frac{S_n}{n^{2-\alpha}} \to_p \frac{\theta\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}.$$

- Results extended by Berestycki-Berestycki-Limic (2011).
- Limiting distribution unknown, conjectured to be normal if and only if $\alpha > \sqrt{2}$ (Delmas, Dhersin, Siri-Jegousse, 2008).

Bolthausen-Sznitman case (Drmota-Iksanov-Möhle-Rösler, 2007):

$$\frac{\log n}{n} S_n \to_p \theta, \qquad \frac{(\log n)^2}{\theta n}\left(S_n - \frac{\theta n}{\log n} - \frac{\theta n \log\log n}{(\log n)^2}\right) \Rightarrow X,$$

where $E[e^{itX}] = \exp\left(-\frac{\pi}{2}|t| + it\log|t|\right)$.

Site and allele frequency spectrum (Kingman's coalescent)

Ewens sampling formula (Ewens, 1972): The probability that the allelic partition has a_j blocks of size j for $j = 1, \dots, n$ is

$$\frac{n!}{2\theta(1+2\theta)\dots(n+2\theta)} \prod_{j=1}^{n} \left(\frac{2\theta}{j}\right)^{a_j} \frac{1}{a_j!}.$$

Site frequency spectrum: $E[M_{k,n}] = 2\theta/k$.

Allele frequency spectrum: $E[N_{k,n}] \sim 2\theta/k$.

Site and allele frequency spectrum (Beta coalescent)

Theorem (Berestycki-Berestycki-Schweinsberg, 2007): For the Beta$(2-\alpha, \alpha)$-coalescent with $1 < \alpha < 2$, we have

$$\frac{M_{k,n}}{S_n} \rightarrow_p \frac{(2-\alpha)\Gamma(k+\alpha-2)}{\Gamma(\alpha-1)k!} = a_k$$

and $N_{k,n}/A_n \rightarrow_p a_k$.

We have $a_1 = 2 - \alpha$ and $a_k \sim Ck^{\alpha-3}$. Smaller α means more low frequency mutants.

Original proof used connections with CSBPs.

Results improved by Berestycki-Berestycki-Limic (2011).

Theorem (Basdevant-Goldschmidt, 2008) For the Bolthausen-Sznitman coalescent,

$$\frac{\log n}{n} N_{1,n} \rightarrow_p \theta, \qquad \frac{(\log n)^2}{n} N_{k,n} \rightarrow_p \frac{\theta}{k(k-1)}, \; k \geq 2.$$

Example 1: Pacific Oyster

Data on 141 Pacific Oysters from British Columbia.
Data from Boom, Boulding, and Beckenbach (1994).
Analyzed by Eldon-Wakeley (2006), Sargsyan-Wakeley (2008).

There were 48 segregating sites.
$M_{1,n} = 29$, $M_{2,n} = 12$, $M_{3,n} = 4$, $M_{6,n} = 2$, and $M_{67,n} = 1$.

Predictions with Kingman's Coalescent: to estimate θ, set

$$48 = 2\widehat{\theta} \sum_{j=1}^{140} \frac{1}{j},$$

which gives $\widehat{\theta} \approx 4.35$. Then predict $M_{k,n} = 2\widehat{\theta}/k$.

Predictions with beta coalescent: predict $M_{k,n} = 48a_k$. Choose the α that gives the best fit to the data.

Comparision of predictions from Kingman's coalescent and from the beta coalescent with $\alpha = 1.35$.

Site Frequency Spectrum

k	Observed	Kingman	beta
1	29	8.7	31.2
2	12	4.3	5.5
3	4	2.9	2.5
4	0	2.2	1.4
5	0	1.7	1.0
6	2	1.4	0.7
7+	1	26.7	5.7

Neither fit is good. The fit from the beta coalescent is better.

Example 2: Atlantic Cod

Data on 1278 Atlantic Cod, segment 250 base pairs long.

Data from Arnason (2004), analyzed by Birkner and Blath (2007) and Birkner, Blath, and Steinrücken (2011).

There were 59 haplotypes (blocks of allelic partition).

Estimate $\alpha = 1.43$ for the beta coalescent.

Allele Frequency Spectrum

k	Observed	Kingman	beta
1	32	7.6	33.6
2	7	3.8	7.2
3	6	2.5	3.4
4	2	1.9	2.1
5	3	1.5	1.4
6	1	1.3	1.0
7	1	1.1	0.8
8+	7	39.2	9.3

Statistical analysis in Birkner and Blath (2007) allows one to reject the Kingman's coalescent hypothesis.

Limitations to this analysis

1. Violations of assumptions. Example: Atlantic Cod data had only 39 segregating sites, but 59 haplotypes.

2. It seems that
$$\frac{M_{k,n}}{S_n} = a_k + O\left(\frac{1}{\log n}\right),$$
so the a_k are not precise for finite values of n (Durrett, Huerta-Sanchez).

3. Different coalescent processes can lead to similar values for the site frequency spectrum and allele frequency spectrum. It is difficult to distinguish the effects of selection, large family sizes, changing population size.

Block sizes of exchangeable random partitions

Let Π be an exchangeable random partition of $\mathbb{N}$.
Let Π_n be the restriction of Π to $\{1, \dots, n\}$.
Let N_n be the number of blocks of Π_n, and let $N_{k,n}$ be the number of blocks of size k.

Theorem (Karlin, 1967; Gnedin-Hansen-Pitman, 2007; Schweinsberg, 2010): Suppose $1 < \alpha < 2$. If $N_n/n^{2-\alpha} \rightarrow_p c > 0$, then

$$\frac{N_{k,n}}{N_n} \rightarrow_p a_k = \frac{(2-\alpha)\Gamma(k+\alpha-2)}{\Gamma(\alpha-1)k!}.$$

Example (Schweinsberg, 2010): Suppose the population size in generation $-t$ is $\lceil Nt^{-\gamma} \rceil$, where $\gamma > 0$. Genealogy is a time-changed Kingman's coalescent with merger rate $r(t) = t^{\gamma}$. Let $\alpha = (2+\gamma)/(1+\gamma)$. Then

$$\frac{N_n}{n^{2-\alpha}} \rightarrow_p \frac{\theta 2^{\alpha-1}(\alpha-1)^{2-\alpha}\pi}{\sin(\pi(2-\alpha))}.$$

Thus, $N_{n,k}/N_n \rightarrow_p a_k$.