Population Genetics and Evolution – III

Speed of Adaptation – The Coalescent

Luca Peliti

Bengaluru / December 2017

SMRI (Italy)
luca@peliti.org

Outline

The Speed of Adaptation

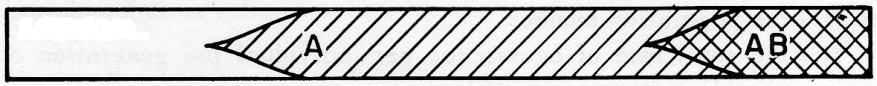
The Coalescent

The Speed of Adaptation

Adaptation Speed vs. Population Size

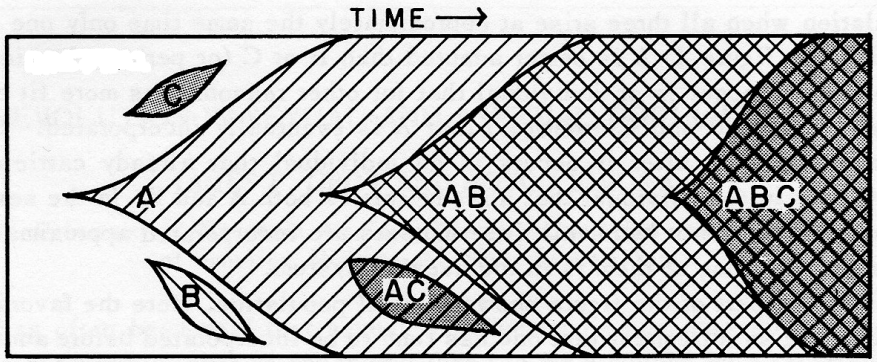
N: Population size; μ: Mutation rate (per geneation & individual)

- $N\mu \ll 1$:



"Periodic selection"

- $N\mu \gg 1$:



"Multiple mutations"

Fixation probability

s: Selection advantage for the mutant; $x_0 = n/N$: Fraction of the mutant population

$$p^{\mathrm{fix}}(x_0) = \frac{1 - \mathrm{e}^{-2Nsx_0}}{1 - \mathrm{e}^{-2Ns}} = \frac{1 - \mathrm{e}^{-2ns}}{1 - \mathrm{e}^{-2Ns}}$$

When $ns \gtrapprox 1$, $p^{\mathrm{fix}} = \mathrm{O}(1)$: the mutation is *established*
Probability of fixation agains genetic drift:

$$\pi(s) = p^{\mathrm{fix}}\left(\frac{1}{N}\right) = \frac{1 - \mathrm{e}^{-2s}}{1 - \mathrm{e}^{-2Ns}}$$

Clonal Interference

Gerrish & Lenski, 1998

Intermediate-size populations:

- Mutations arise independently each from the wild type
- Only advantageous mutations are considered
- The selective advantage s has a distribution $\rho(s)$ (typically exponential (Orr))
- Mutations go to fixation if they can do it before a new more advantageous mutation establishes

Rate of fixation of a mutation of advantage s:

$$\begin{aligned} p^{\text{fix}}(s) &= \pi(s)\, \mathrm{e}^{-\lambda(s)}\, \rho(s) \\ \lambda(s) &\simeq \frac{N\mu}{s} \ln N \int_s^\infty \mathrm{d}x\, \pi(x)\rho(x) \end{aligned}$$

Clonal Interference

Gerrish & Lenski, 1998

Expected fixation rate per generation:

$$E(k) = \int_0^\infty \mathrm{d}s \; p^{\mathrm{fix}}(s)$$

Expected fitness increment per substitution:

$$E(s) = \int_0^\infty \mathrm{d}s \; s \, p^{\mathrm{fix}}(s) / E(k)$$

Rate of adaptation (RA):

$$\lim_{t \to \infty} \frac{\langle W(t) \rangle}{t} = E(k) \log \left(1 + E(s)\right)$$

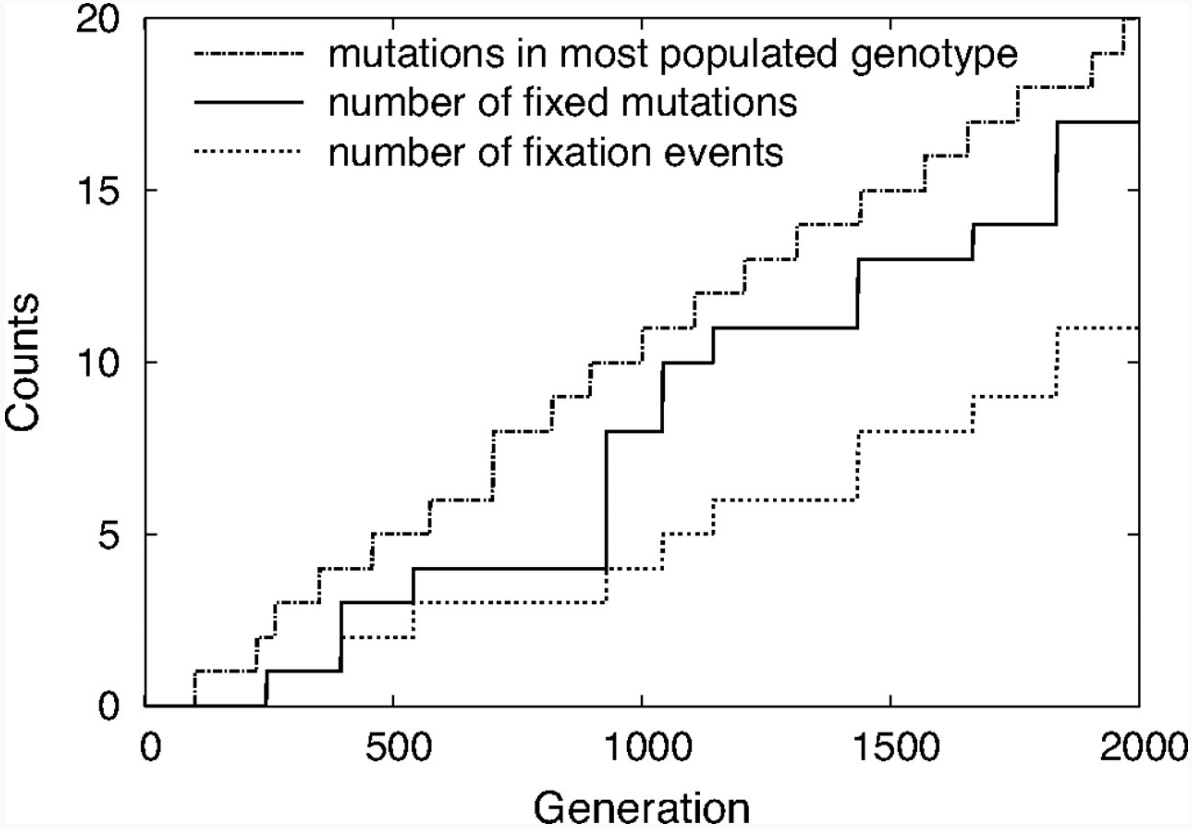
Multiple mutations

Park & Krug, 2007

- *Fixation* has a different meaning when multiple mutations can arise in a single genotype
- In clonal populations, *fixation* corresponds to the change in the last common ancestor
- At each fixation event, several mutations fix *simultaneously*
- Origination and fixation rates have to be distinguished
- The occurrence of simultaneous fixations leads to an *increase* of the rate of adaptation wrt to the Gerrish-Lenski (GL) estimate
- Adaptation events also occur more regularly in time

Multiple mutations

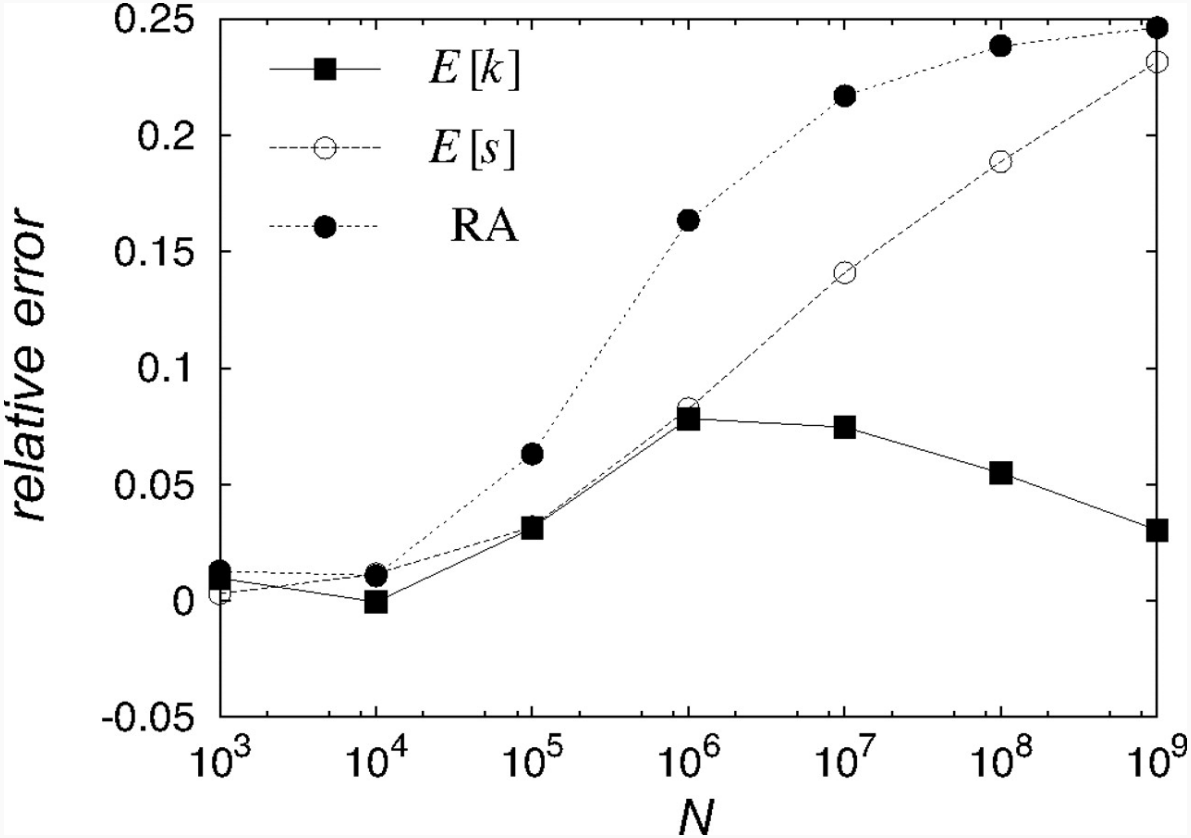
Park & Krug, 2007



An example plot showing different quantities characterizing the rate of substitution. Population size is $N = 10^9$

Multiple mutations

Park & Krug, 2007



Semilogarithmic plot of the relative error of the GL prediction (= (GLprediction-simulation)/simulation) vs. N for $E(k)$, $E(s)$, and the RA $\log \langle W(t) \rangle / t \sim (E(k) \log(1 + E(s)))$, respectively.

The Coalescent

Genealogies

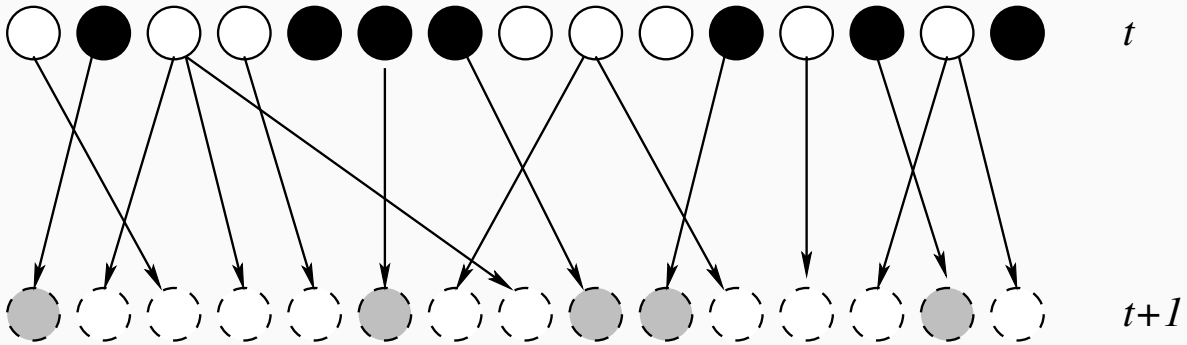
- How far in the past must we go to reach the last common ancestor of n individuals? of the whole population?
- How many different genotypes can we expect to find by sampling n individuals?
- How do the times to the last common ancestor depend on the particular chosen sample? on the population size?
- How do they fluctuate as the population evolves in time?
- How are they affected by selection?

These questions can be addressed by using the concept of the *Coalescent*

JFC Kingman

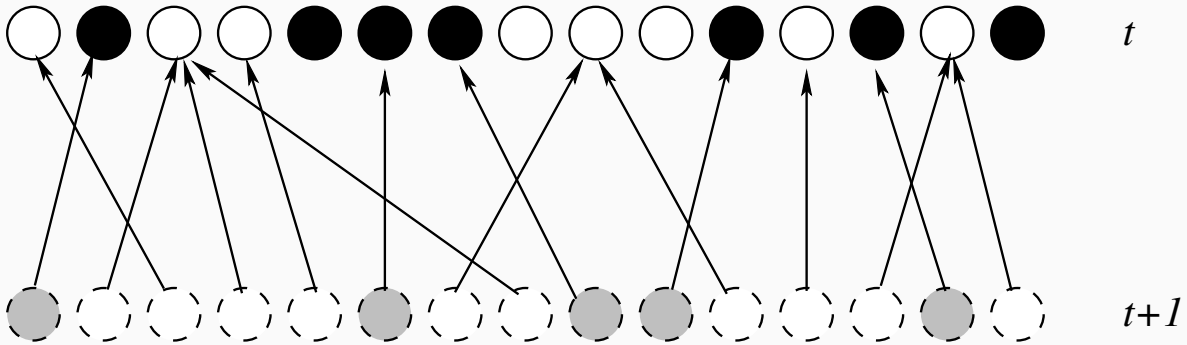
The Wright-Fisher model

Two ways of looking at the Wright-Fisher model:

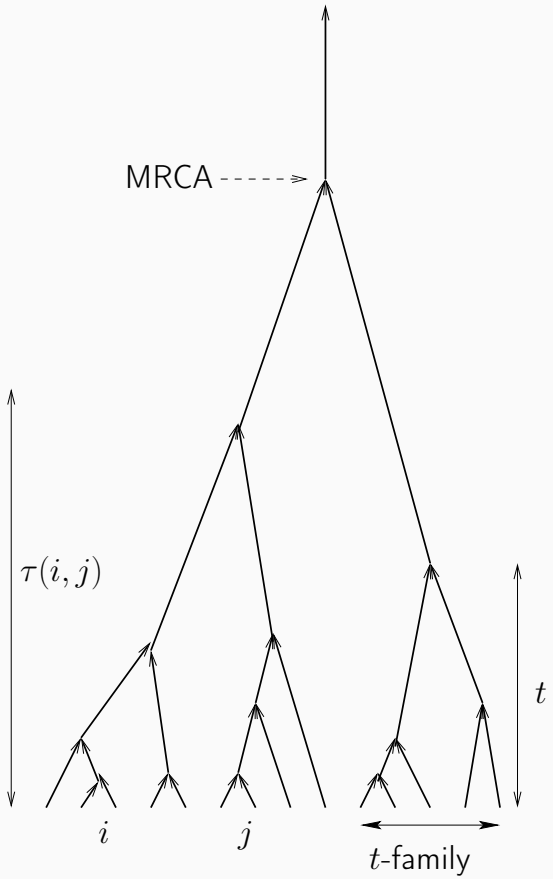


The Wright-Fisher model

Two ways of looking at the Wright-Fisher model:



Iterating the process



Iterating the process

Neutral Wright-Fisher process:

- Set $t = 0$ for the present, and count generations *backward* from the present
- Individual labels: $\{1, \ldots, N\}$
- At each generation, define the application $p : \ i \mapsto p_t(i)$ from i to its parent
- $p_t(i)$ is extracted at random, independently for each i and each t
- Ancestor: $a_t(i) = \underbrace{p_t(p_{t-1}(\cdots p_2(p_1}_{t \text{ times}}(i))))$
- Lineage: $\mathsf{L}(i) = (a_0(i) = i, a_1(i), a_2(i), \ldots)$
- Lineage coalescence: $a_t(i) = a_t(j)$, $i \neq j$
- Coalescence time: $\tau(i, j)$: $a_\tau(i) = a_\tau(j)$, $a_{\tau-1}(i) \neq a_{\tau-1}(j)$

Iterating the process

Disclaimer:

> *In this [lecture] gene genealogies will sometimes be referred to simply as genealogies. It should be understood that this refers to the genetic ancestry of a sample at some locus in the genome and not to the usual definition of a genealogy, being the family relationship of a set of individuals.*

J. Wakeley, 2009

Iterating the process

Questions:

- How many generations to the MRCA?
- What is the distribution of $\tau(i, j)$?
- What are the consequences for quantities we can measure?

N.B.: When treating *diploids*, set $N = 2 \cdot$ population size
Discussion of the *effective* population size: later!

Coalescent statistics

Hypotheses:

1. Equal fitness for all types (neutral process)
2. No subdivisions in the population (geographical or otherwise)
3. Constant population size

Assumptions 1. and 2. lead to *exchangeability*: the number of offspring of any individual is statistically the same random variable as for any other individual

Coalescent statistics

- Probability that n individuals have all different parents:

$$\begin{aligned} w_n &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right) \\ &\simeq 1-\frac{n(n-1)}{2N} \qquad\qquad n \ll N \end{aligned}$$

- $\Pi_n(t)$: probability of n independent lineages at time t

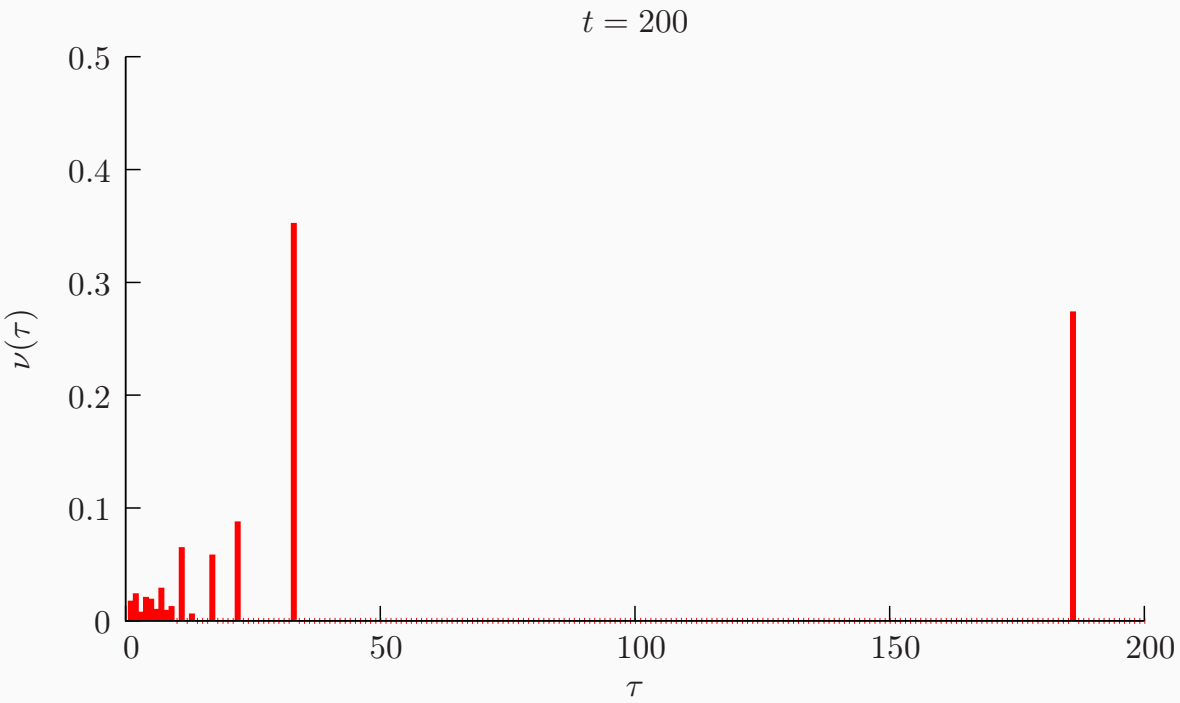
$$\Pi_n(t+1) = w_n \Pi_n(t) \simeq \left(1-\frac{n(n-1)}{2N}\right)\Pi_n(t)$$

- $\Pi_n(t) = \left(1-\frac{n(n-1)}{2N}\right)^t \simeq \mathrm{e}^{-n(n-1)t/(2N)}$
- In particular $\Pi_2(t) \simeq \mathrm{e}^{-t/N}$

Coalescent statistics

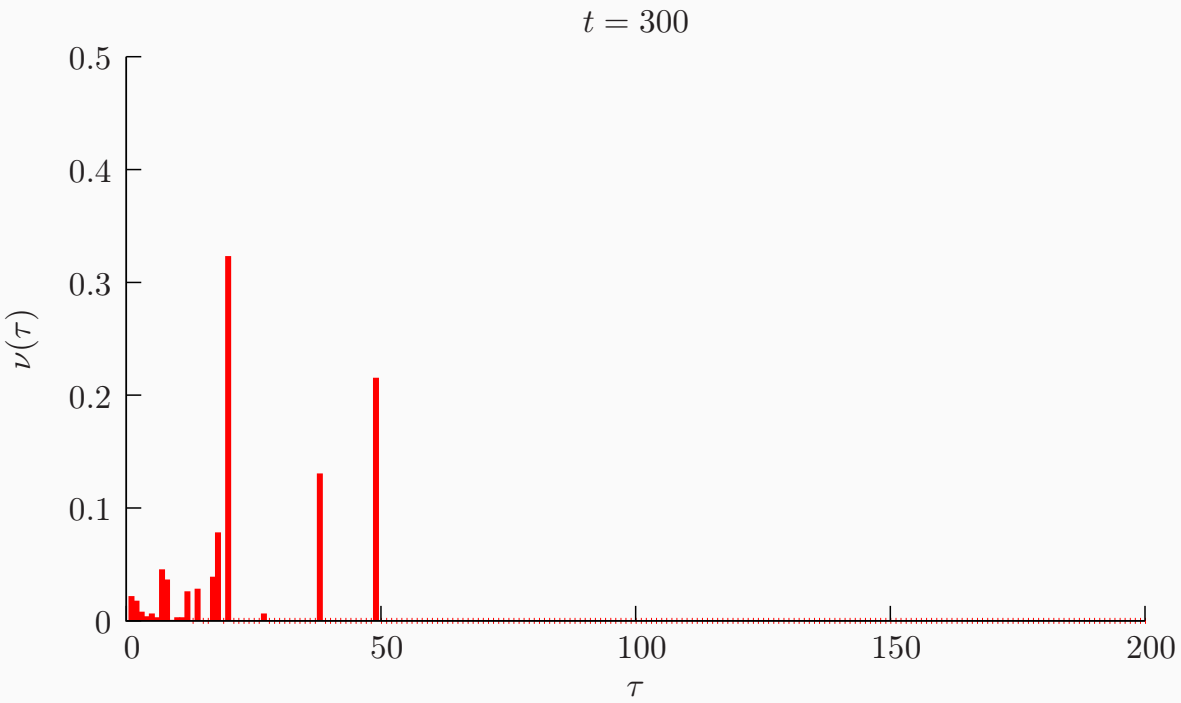
- Averages over the *process* are expressed by $\overline{\ldots}$
- Averages over the *population* are expressed by $\langle \ldots \rangle$
- Thus $\overline{\tau(i,j)} = N$
- Mutation rate u per genome and generation, infinite *site* model
- Expected # of mutations wrt the common ancestor: Nu
- Expected # of mutations between i and j: $2Nu = \theta$

Distribution of coalescent times



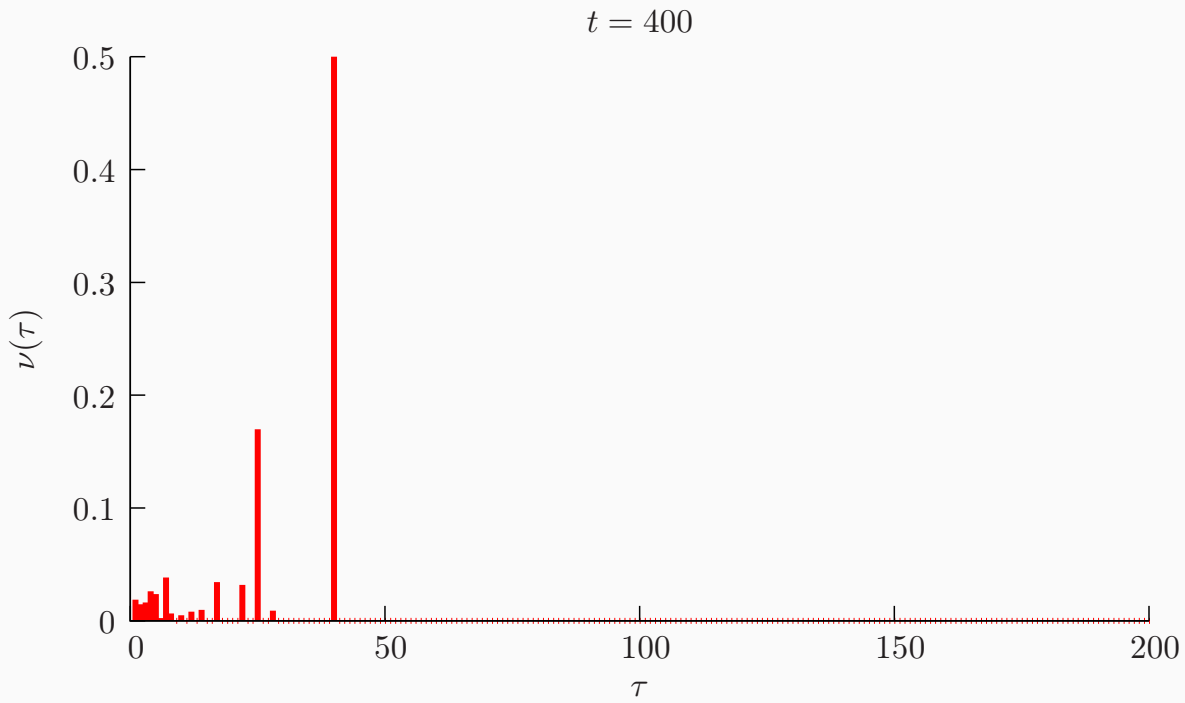
$N = 50$

Distribution of coalescent times



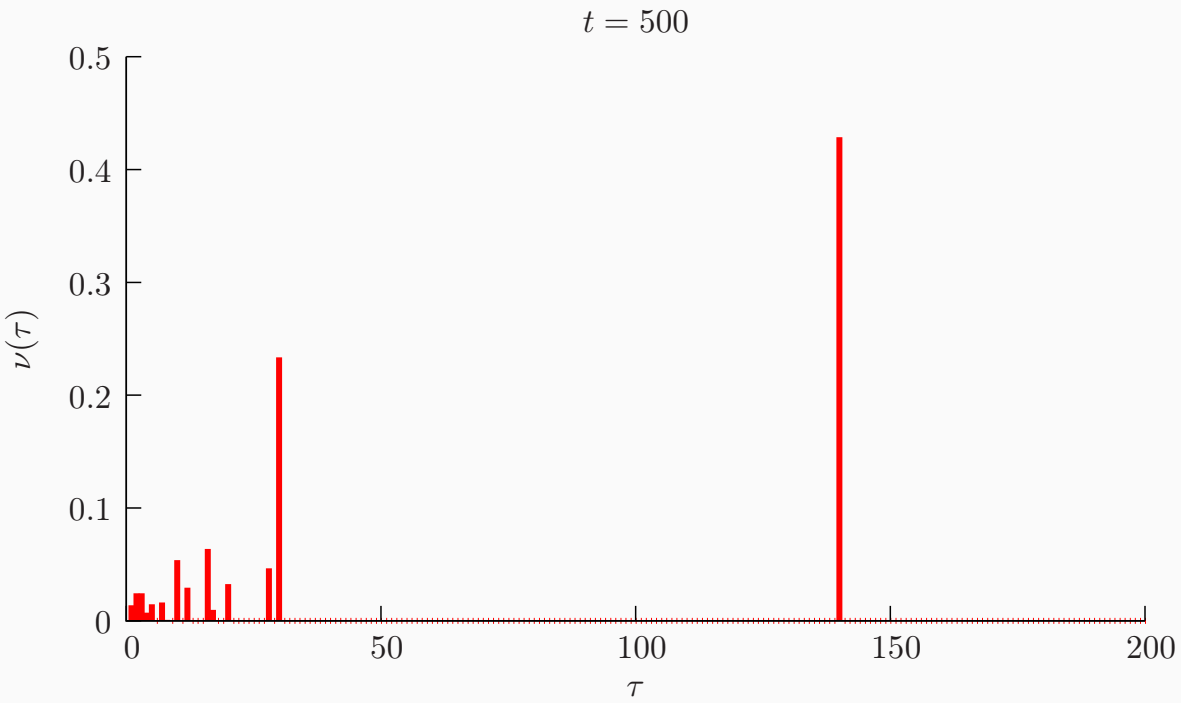
$N = 50$

Distribution of coalescent times



$N = 50$

Distribution of coalescent times



$N = 50$

Universality of the coalescent

- Reproduction model: Distribution of offspring size m: π_m

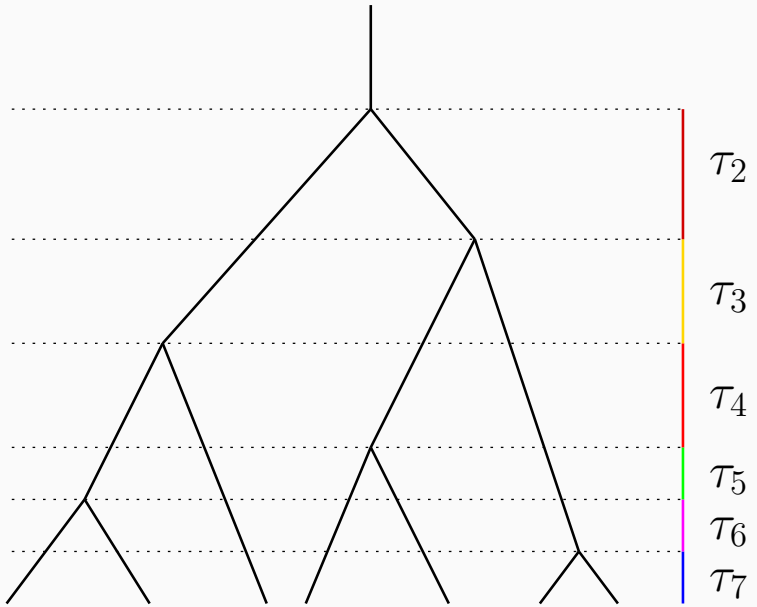
$$\text{WF model: } \pi_m \quad = \quad \mathrm{e}^{-1}/m! \qquad \text{(Poisson)}$$

- $\overline{m} = \sum_m m\,\pi_m = 1$
- Probability of coalescence for n lineages:

$$1 - w_n = \binom{n}{2} \frac{1}{N} \sum_m m(m-1)\,\pi_m = \frac{n(n-1)}{2N} \left(\overline{m^2} - 1\right)$$

- Define $\overline{m(m-1)} = \overline{m^2} - 1 = \kappa$
- Thus $w_n = 1 - \frac{n(n-1)}{2} \frac{\kappa}{N}$
- If $\overline{m^2} < \infty$, all results hold, up to a time rescaling
- Choose time units so that $w_n = 1 - \frac{n(n-1)}{2}$

Probability of a genealogy



$$P(\tau_2, \ldots, \tau_7) = \exp\left\{-\frac{1}{2}\left[7 \cdot 6 \cdot \tau_7 + 6 \cdot 5 \cdot \tau_6 + \cdots + 2 \cdot 1 \cdot \tau_2\right]\right\}$$

Each τ_k is independent, with distribution $\mathcal{P}_k(\tau) = \binom{k}{2} \mathrm{e}^{-\binom{k}{2}\tau}$

Distribution of the age of the MRCA

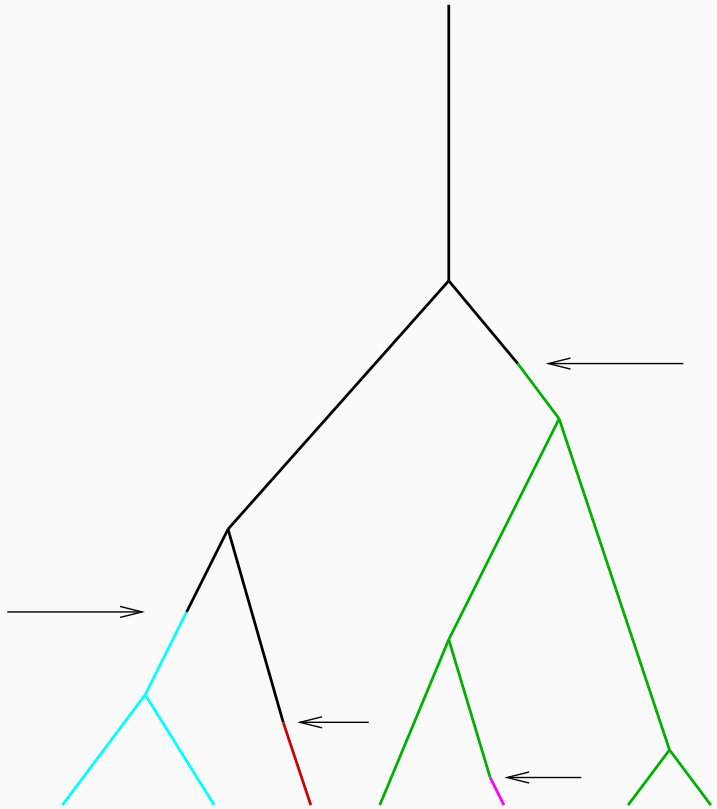
- Define T_{MRCA} as the age of the MRCA of n samples
- Then $T_{\mathrm{MRCA}} = \sum_{k=2}^{n} \tau_k$
- Each τ_k is exponentially distributed, with average $\overline{\tau_k} = \left[\binom{k}{2}\right]^{-1}$

Distribution of the age of the MRCA

$$
\begin{aligned}
\mathcal{P}_{\mathrm{MRCA}}(T) &= \mathsf{Prob}(T_{\mathrm{MRCA}} = T) \\
&= \sum_{k=2}^{n} \binom{k}{2} \mathrm{e}^{-\binom{k}{2}T} \prod_{j\ (\neq k)} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{k}{2}} \\
&= \sum_{k=2}^{n} \binom{k}{2} (-1)^k (2k-1) \frac{n(n-1)\cdots(n-k+1)}{n(n+1)\cdots(n+k-1)} \mathrm{e}^{-\binom{k}{2}T}
\end{aligned}
$$

Tavaré, 1984; Takahata and Nei, 1985

Coalescence and mutations



The probability of a mutation occurring is uniform per unit length of the genealogy

Coalescence and mutations

- Assume mutation rate u per genome and generation, infinite *allele* model
- Two individuals carry the same allele if they encounter no mutation before their last common ancestor
- The probability of *not* having a mutation in a generation in a lineage is $1 - u$
- The probability that *neither* lineage exhibits a mutation is $(1-u)^{2\tau(i,j)} \simeq \exp\left(-2u\tau(i,j)\right)$
- Thus the probability that two individuals have the same allele is

$$\begin{aligned} p_{\text{same}} &= \frac{1}{N}\int_0^\infty \mathrm{d}\tau\ \mathrm{e}^{-2u\tau-\tau/N} \\ &= \frac{1}{1+2uN} = \frac{1}{1+\theta} \end{aligned}$$

Ewens' sampling formula

- Infinite-allele model
- Take n samples from a large population with $\theta = 2Nu$
- Samples belong to the same group if they exhibit the same allele
- What is the probability that there are b_1 groups with 1 element, b_2 groups with 2 elements,... b_k with k elements,... ?

Ewens' sampling formula

$$n = \sum_{k=1}^{n} k\, b_k \qquad \text{\# of samples}$$

$$\boxed{P(b_1, \ldots, b_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \frac{1}{1^{b_1} \cdot 2^{b_2} \cdots n^{b_n}} \frac{\theta^{\sum_k b_k}}{b_1! b_2! \cdots b_n!}}$$

The Chinese Restaurant Process

The Chinese Restaurant Process

At each step, when there are n customers:

- The customer sits at a new empty table with probability $\theta/(\theta + n)$, or
- The customer picks up one of the customers at random and sits at the same table

The Chinese Restaurant Process

- At each step, we get a factor $1/(\theta+n)$ $(n=0,1,\ldots)$
- Each new table gets a factor θ
- In going from k to $k+1$, each table gets a factor k
- Thus the probability that the (labeled) customers sit at ℓ tables, $i=1,\ldots,\ell$ of size k_i, $\sum_{i=1}^{\ell} k_i = n$ is given by

$$P^{\mathrm{lab}}(k_1,\ldots,k_\ell) = \frac{\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell}(k_i-1)!$$

- There are $n!/(k_1!\cdots k_\ell!)$ distributions of the customers compatible with $(k_1,\ldots,k_\ell)$, thus

$$\begin{aligned} P(k_1,\ldots,k_\ell) &= \frac{n!}{k_1!\cdots k_\ell!}\,\frac{\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell}(k_i-1)! \\ &= \frac{n!\,\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{\ell}\frac{1}{k_i} \end{aligned}$$

The Chinese Restaurant Process

- Labelling the tables has introduced an overcounting: only the sizes of the tables matter! Thus defining

$$b_j = \sum_{i=1}^{\ell} \delta_{k_i,j}$$

we obtain

$$P(b_1, \ldots, b_n) = \frac{n!\,\theta^\ell}{\theta(\theta+1)\cdots(\theta+n-1)} \frac{1}{1^{b_1}\cdots n^{b_n}} \underbrace{\frac{1}{b_1!\cdots b_n!}}_{\text{Table permutations}}$$

Observables

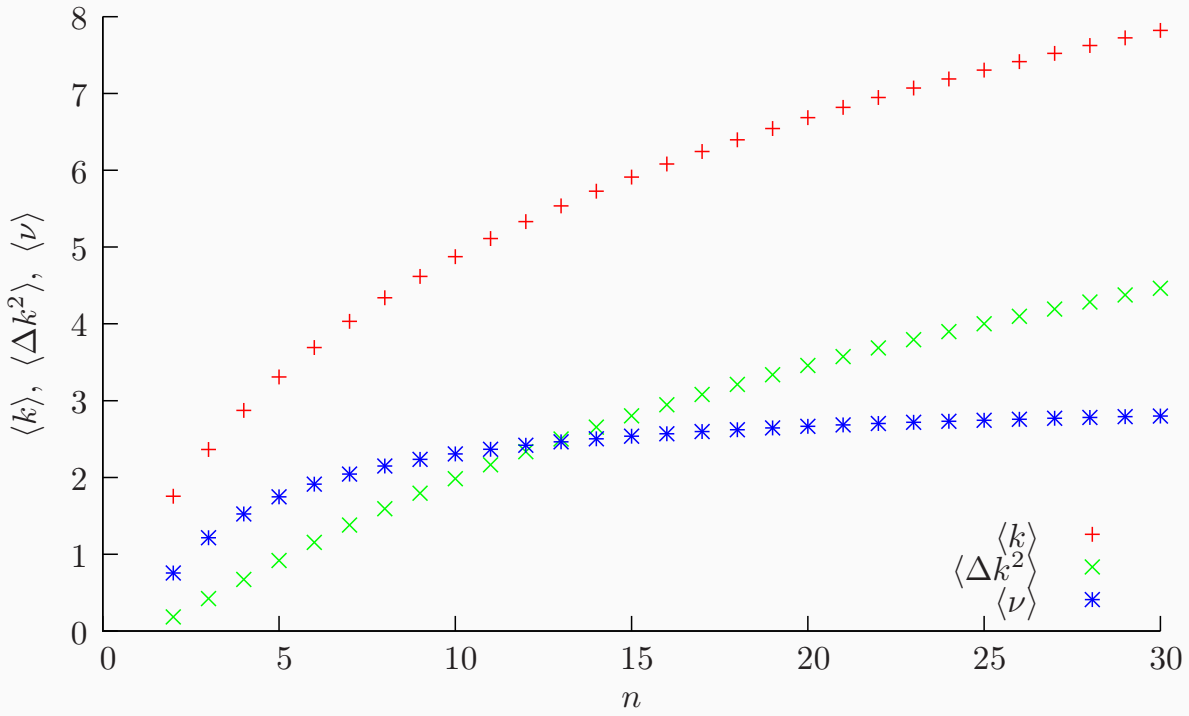
- Distribution of the number k of segregating alleles:

$$\begin{aligned} p_k(n+1) &= \frac{n}{\theta+n} p_k(n) + \frac{\theta}{\theta+n} p_{k-1}(n) \\ \overline{k(n+1)} &= \overline{k(n)} + \frac{\theta}{\theta+n} = \theta \sum_{j=1}^{n-1} \frac{1}{\theta+j} \\ \overline{\Delta k^2(n+1)} &= \overline{k^2(n)} - \overline{k(n)}^2 = \overline{\Delta k^2(n)} + \frac{n\theta}{(\theta+n)^2} \end{aligned}$$

- Distribution of the number ν of singletons:

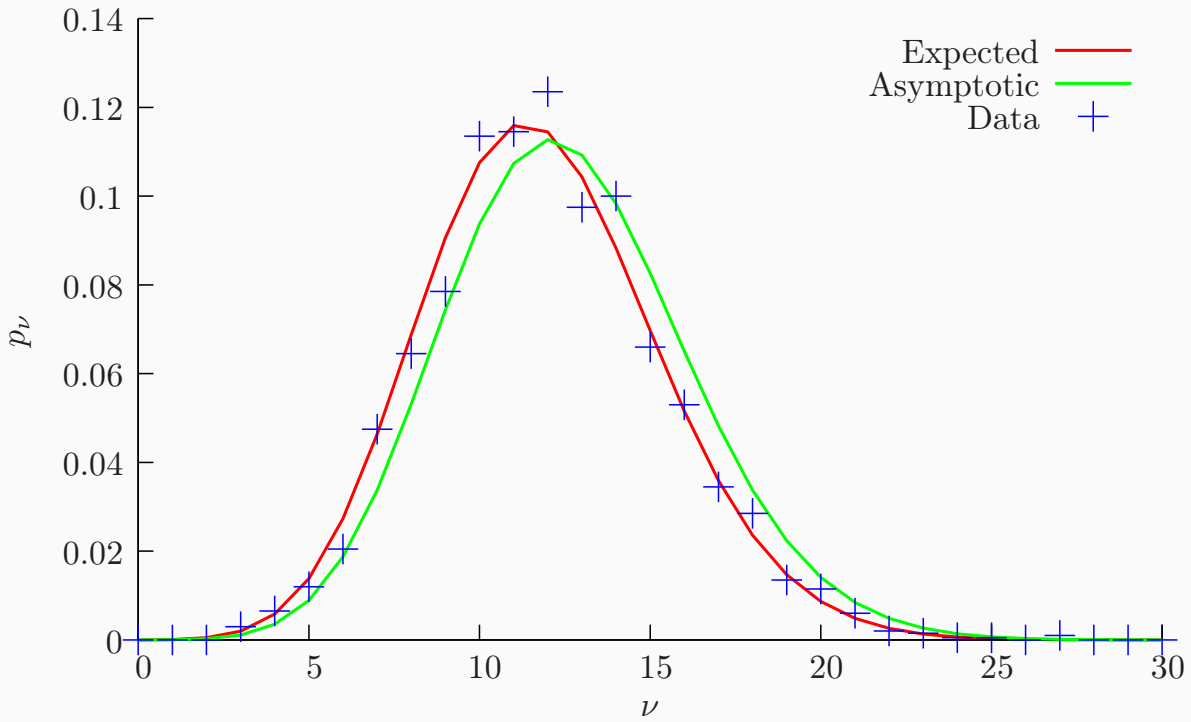
$$\begin{aligned} p_\nu(n+1) &= \frac{\theta}{\theta+n} p_{\nu-1}(n) + \frac{\nu}{\theta+n} p_{\nu+1}(n) + \frac{n-\nu}{\theta+n} p_\nu(n) \\ \overline{\nu(n)} &= \frac{n\theta}{\theta+n-1} \end{aligned}$$

Observables



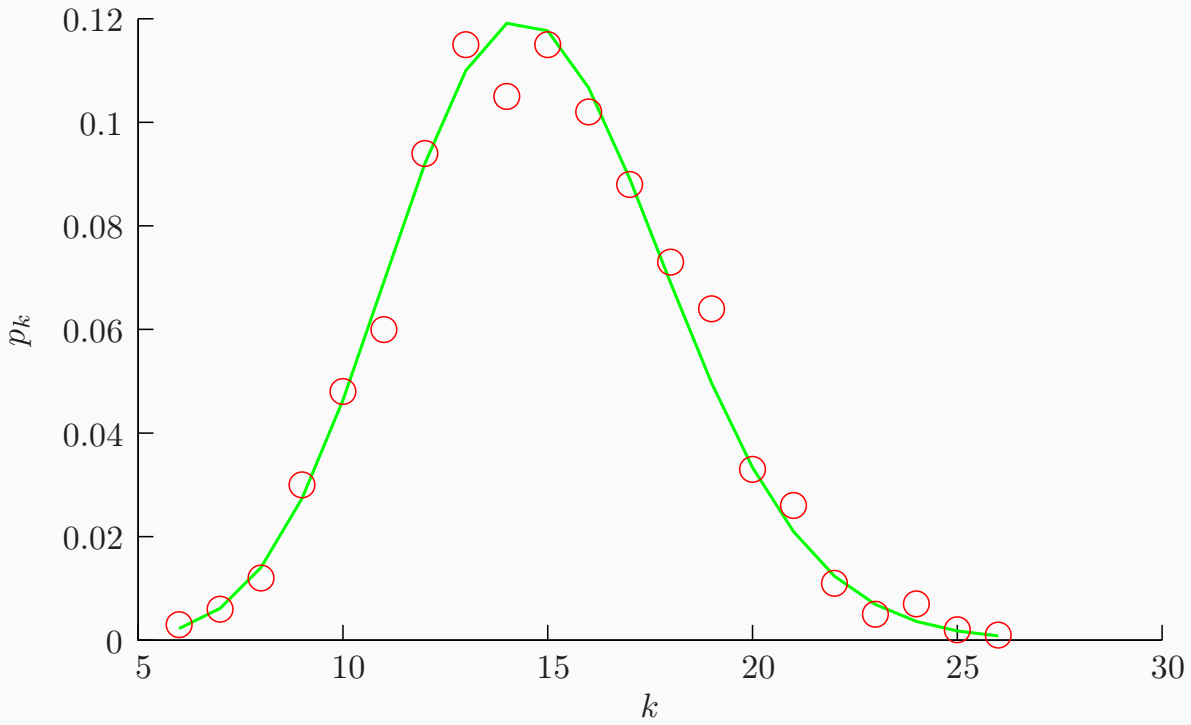
Average $\overline{k}$, variance $\overline{\Delta k^2}$ of segregating alleles and average $\overline{\nu}$ of singletons vs. n for $\theta = 3.1$

Observables



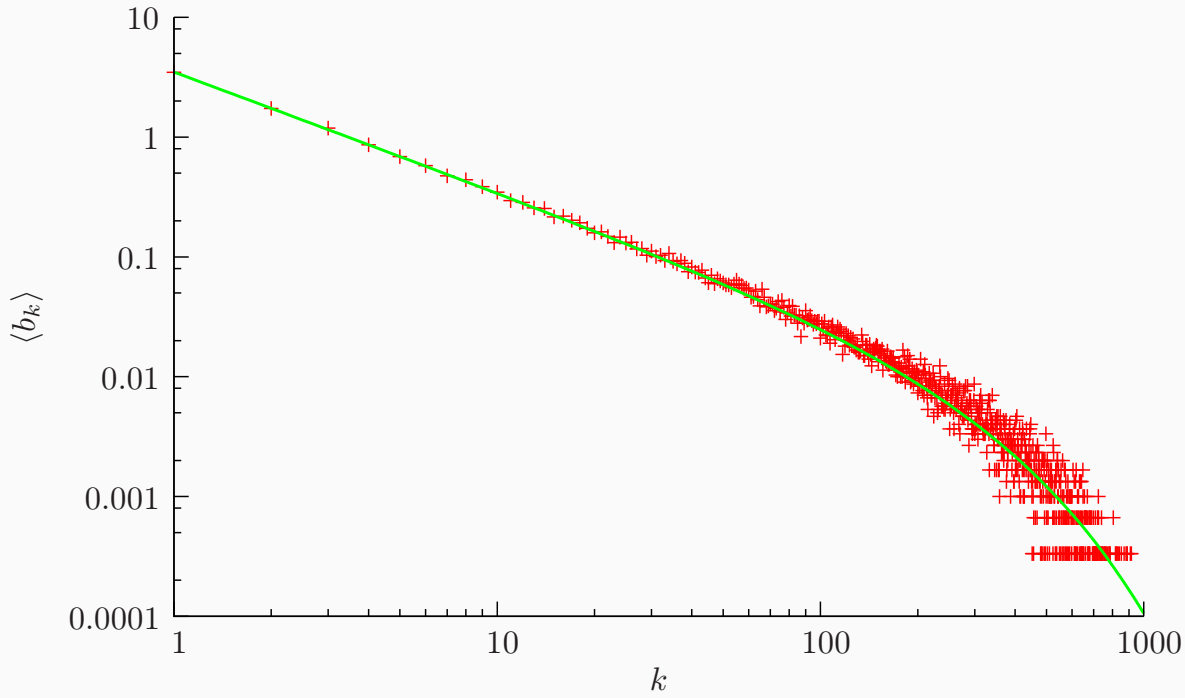
Distribution p_ν of the number of singletons for $n = 200$ and $\theta = 12.6$, together with the asymptotic distribution for $n \to \infty$ and simulation data over 1000 samples

Observables



Distribution p_k of the number of segregating alleles for $n = 300$ and $\theta = 3.1$, together with simulation data averaged over 1000 samples

Frequency spectrum



Average number $\overline{b_k}$ of groups of size k with $n = 1000$ and $\theta = 3.5$. The average is taken over 3000 realizations of the process.

The line corresponds to $\overline{b_k} = \overline{b_1} \mathrm{e}^{-\theta k/n}/k$, with $\overline{b_1} = n\theta/(\theta + n - 1)$

Effective population size N_{e}

The *effective population size* N_{e} can be different from the *census population* N:

- In sexual populations, because only some males actually reproduce(*leks*)
- Generally due to fluctuating population size:

$$\frac{1}{N_{\mathrm{e}}} \simeq \frac{\overline{1}}{N} > \frac{1}{\overline{N}}$$

- If fitness is nonuniform N_{e} is reduced wrt N:

$$N_{\mathrm{e}} = \frac{N}{1 + \mathrm{var}(\#\mathrm{offspring})}$$

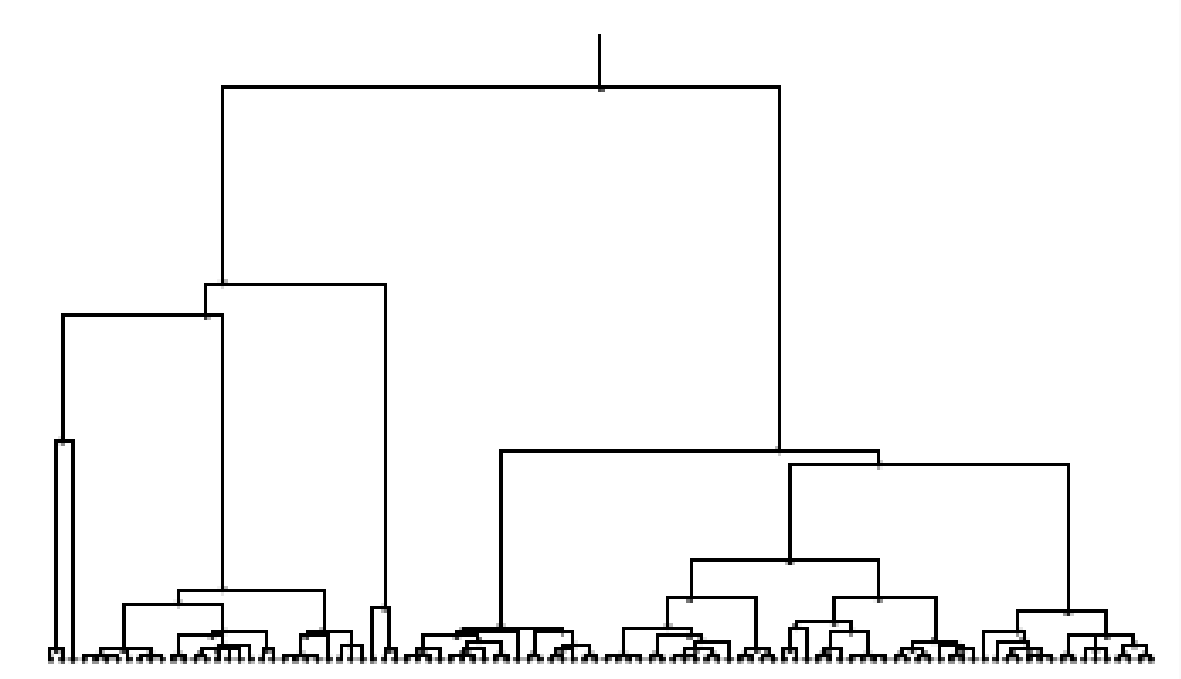
Effective population size N_{e}

In practice, N_{e} is chosen to fit the data:

- For several human genes, $T_{\mathrm{MRCA}} \simeq 400\,000\ \mathrm{yrs}$
- One generation $\simeq 20\ \mathrm{yrs}$
- Assuming neutrality, $N_{\mathrm{e}} \simeq 10\,000$ (diploidy!)
- “Out-of-Africa” bottleneck?

The Coalescent in the presence of selection

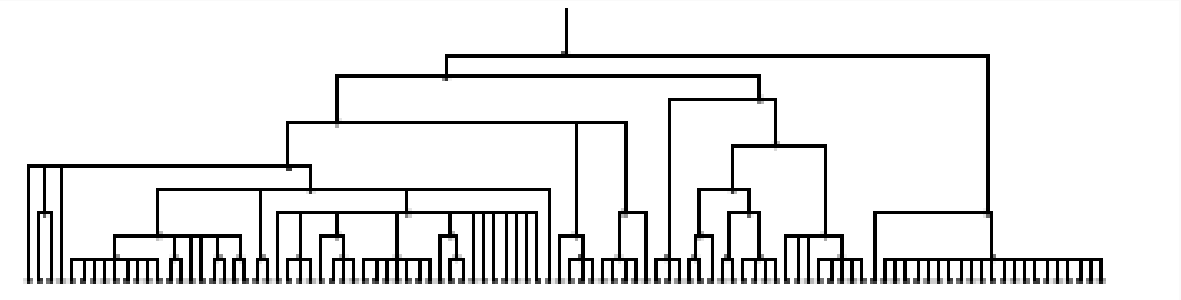
Brunet, Derrida *et al.*, 2006–2012



Neutral genealogy: $N = 100$, $T_{\mathrm{MRCA}} = 125$

The Coalescent in the presence of selection

Brunet, Derrida *et al.*, 2006–2012



Genealogy with selection: $N = 100$, $T_{\mathrm{MRCA}} = 10$

Thank you!

References i

1. J. H. Gillespie, *Population Genetics: A Concise Guide* (2nd ed.) (Baltimore: Johns Hopkins U. P., 2004)
2. J. Wakeley, *Coalescent Theory – An Introduction* (Greenwood Village, Co.: Roberts & Co., 2009)
3. J. Hein, M. H. Schierup and C. Wiuf, *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory* (Oxford: Oxford U. P., 2005)
4. J. F. C. Kingman, The coalescent, *Stochastic Process. Appl.* **13** 135-248 (1982)
5. N. Berestycki, Recent progress in coalescent theory, *Ensaios Matemáticos* **16** 1-193 (2009)

References ii

6. S. Tavaré, Lines-of-descent and genealogical processes, and their application in population genetic models, *Theor. Pop. Biol.* **26** 119-164 (1984)
7. C. Wiuf and J. Hein, On the number of ancestors to a DNA sequence, *Genetics* **147** 1459-1468 (1997)
8. N. Takahata and M. Nei, Gene genealogy and variance of interpopulational nucleotide differences, *Genetics* **122** 325-344 (1985)
9. W. Ewens, The sampling theory of selectively neutral alleles, *Theor. Pop. Biol.* **3** 87–112 (1972)
10. H. Crane. The Ubiquitous Ewens Sampling Formula, *Statistical Science* **31** 1 (2016)

References iii

11. E. Brunet, B. Derrida, A. H. Mueller, and S. Munier, Noisy traveling waves: effect of selection on genealogies, *Europhys. Lett.* **76** 1-7 (2006)
12. E. Brunet, B. Derrida, A. H. Mueller, and S. Munier, Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization, *Phys. Rev. E* **76** 041104 (2007)
13. E. Bolthausen and A.-S. Sznitman, Ten lectures on random media, DMV Seminar, Band 32, Birkhäuser (2001)
14. P. J. Gerrish, and R. E. Lenski, The fate of competing beneficial mutations in an asexual population, *Genetica* **102–103** 127–144 (1998).
15. P. Gerrish, The rhythm of microbial adaptation, *Nature* **413** 299–302 (2001)
16. S.-C. Park and J. Krug, Clonal interference in large populations, *Proc. Natl. Acad. Sci. USA* **104** 18135–18140 (2007)