Population Genetics and Evolution – IV

The Coalescent – Recombination

Luca Peliti

São Paulo / January 2019

SMRI (Italy)
luca@peliti.org

Outline

Introduction

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?

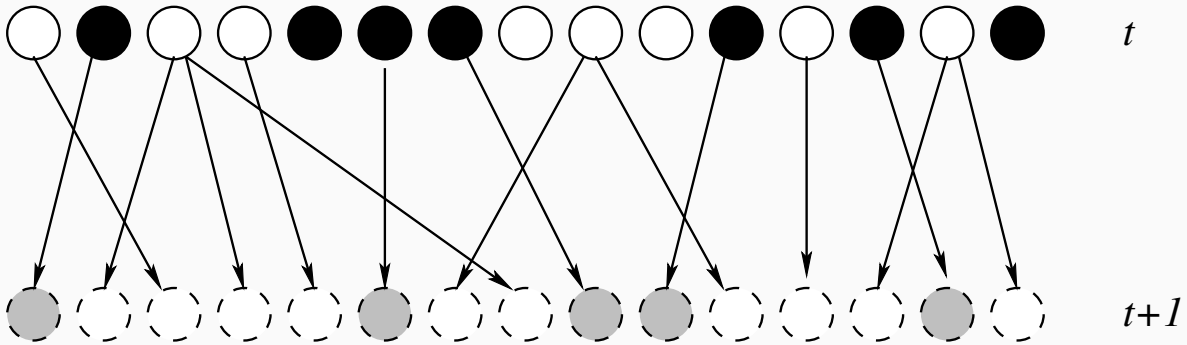
These questions can be addressed by using the concept of the *Coalescent*

The Coalescent

JFC Kingman

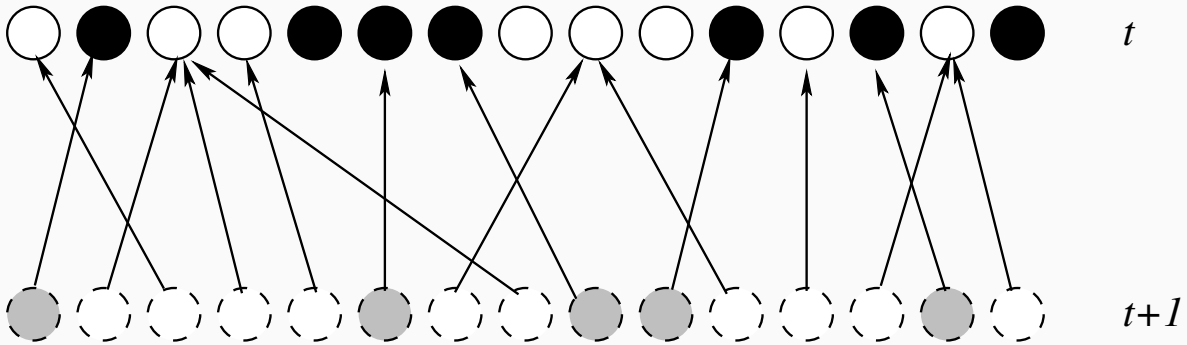
The Wright-Fisher model

Two ways of looking at the Wright-Fisher model:

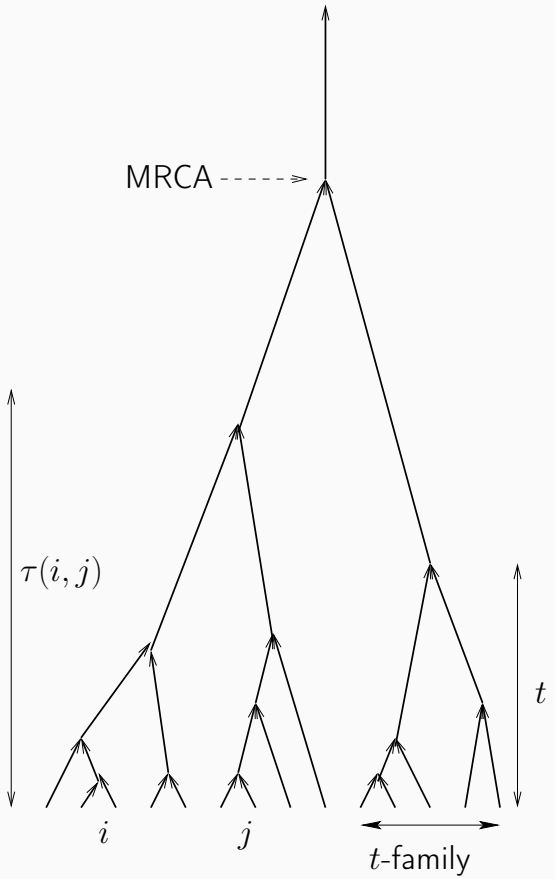


The Wright-Fisher model

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

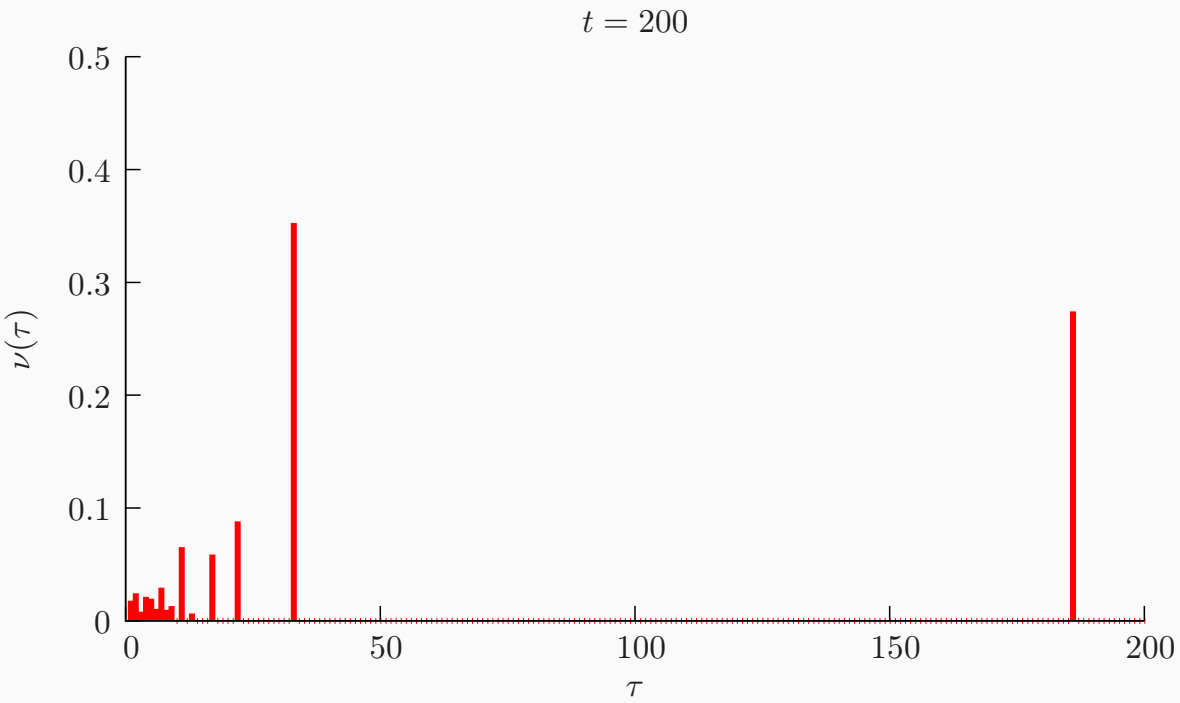
$$\Pi_n(t+1) = w_n \Pi_n(t) \simeq \left(1-\frac{n(n-1)}{2N}\right)\Pi_n(t)$$

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

Coalescent statistics

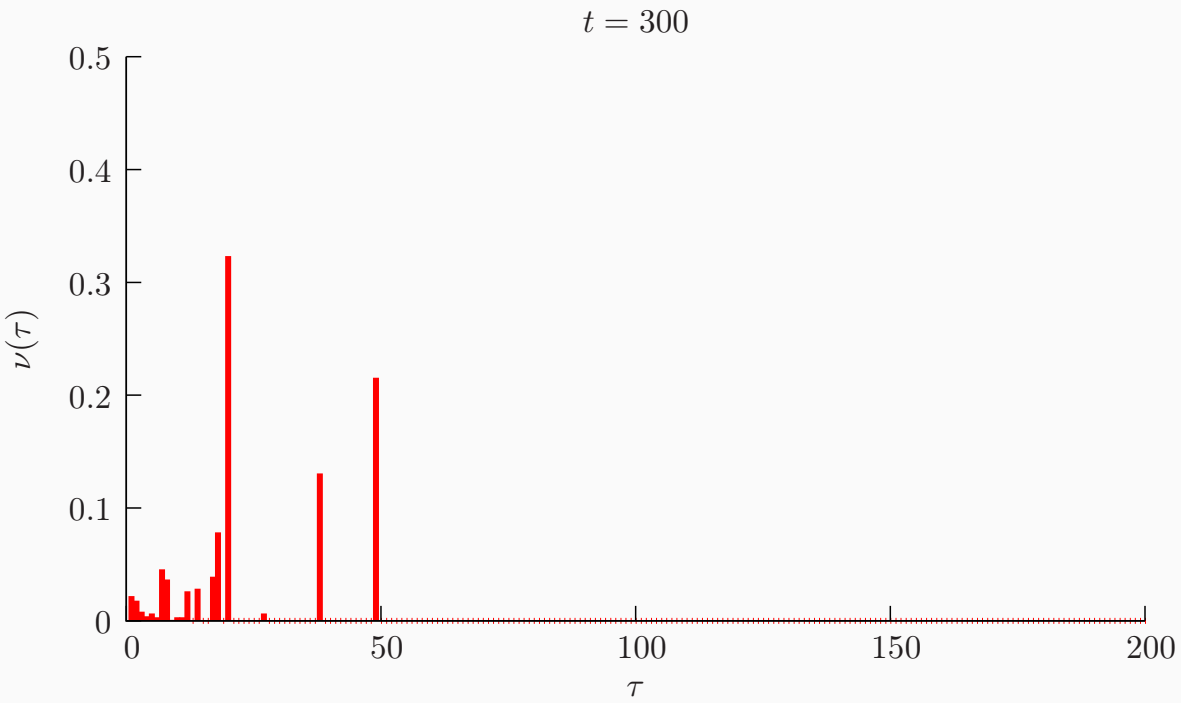
- Averages over the *process* are expressed by $\overline{\ldots}$
- Averages over the *population* are expressed by $\langle \ldots \rangle$
- Thus $\overline{\tau(i,j)} = N$
- Mutation rate u per genome and generation, infinite *site* model
- Expected # of mutations wrt the common ancestor: Nu
- Expected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

Distribution of coalescent times



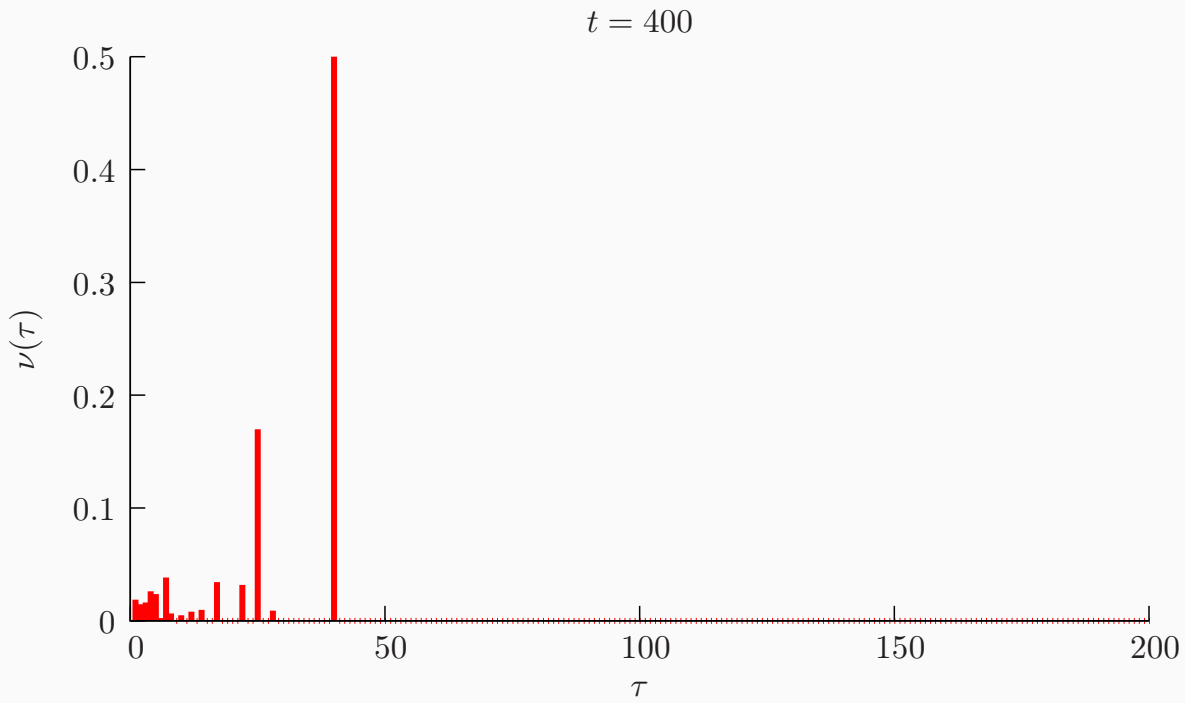
$N = 50$

Distribution of coalescent times



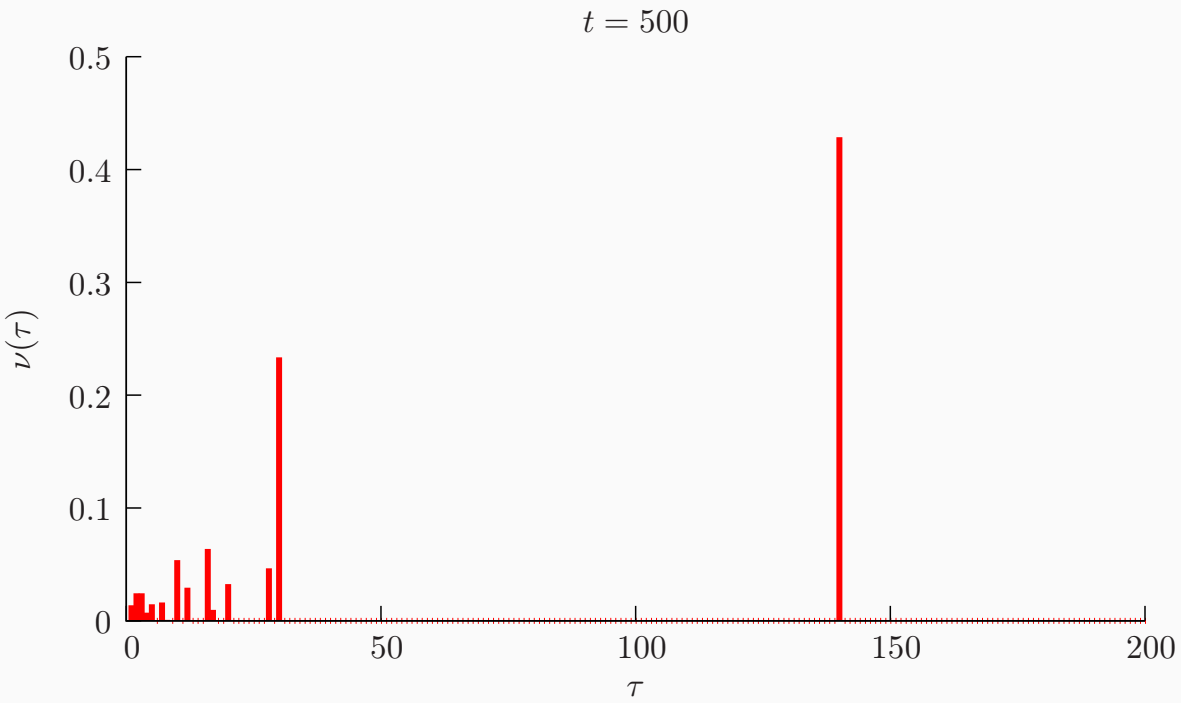
$N = 50$

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

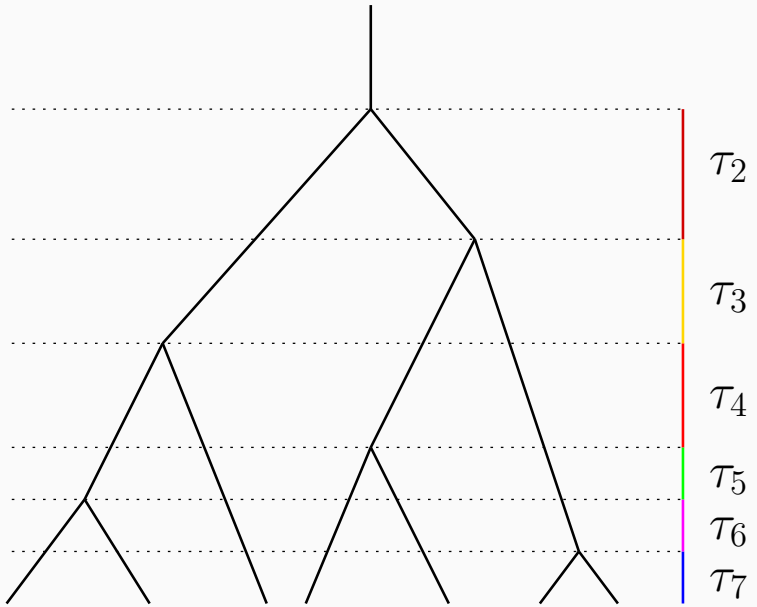
$$\begin{aligned} \text{WF model: } \pi_m &= \mathrm{e}^{-1}/m! \qquad \text{(Poisson)} \\ \text{Moran model: } \pi_0 &= \pi_2 = \frac{1}{N}\left(1-\frac{1}{N}\right),\ \pi_1 = 1-\frac{2}{N}\left(1-\frac{1}{N}\right) \end{aligned}$$

- $\overline{m} = \sum_m m\,\pi_m = 1$
- 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 $w_n = 1 - \frac{n(n-1)}{2}$

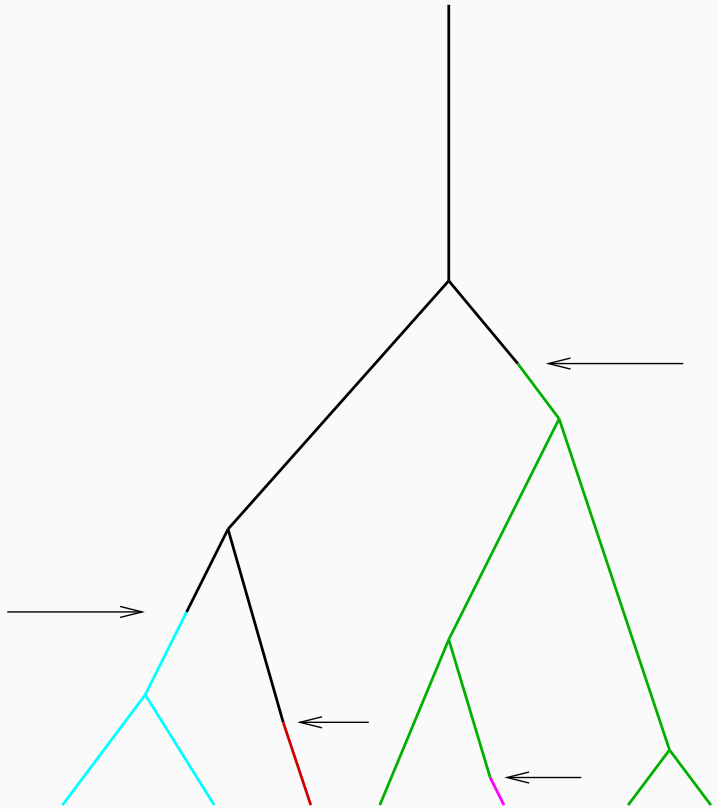
Probability of a genealogy



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

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

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

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

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

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

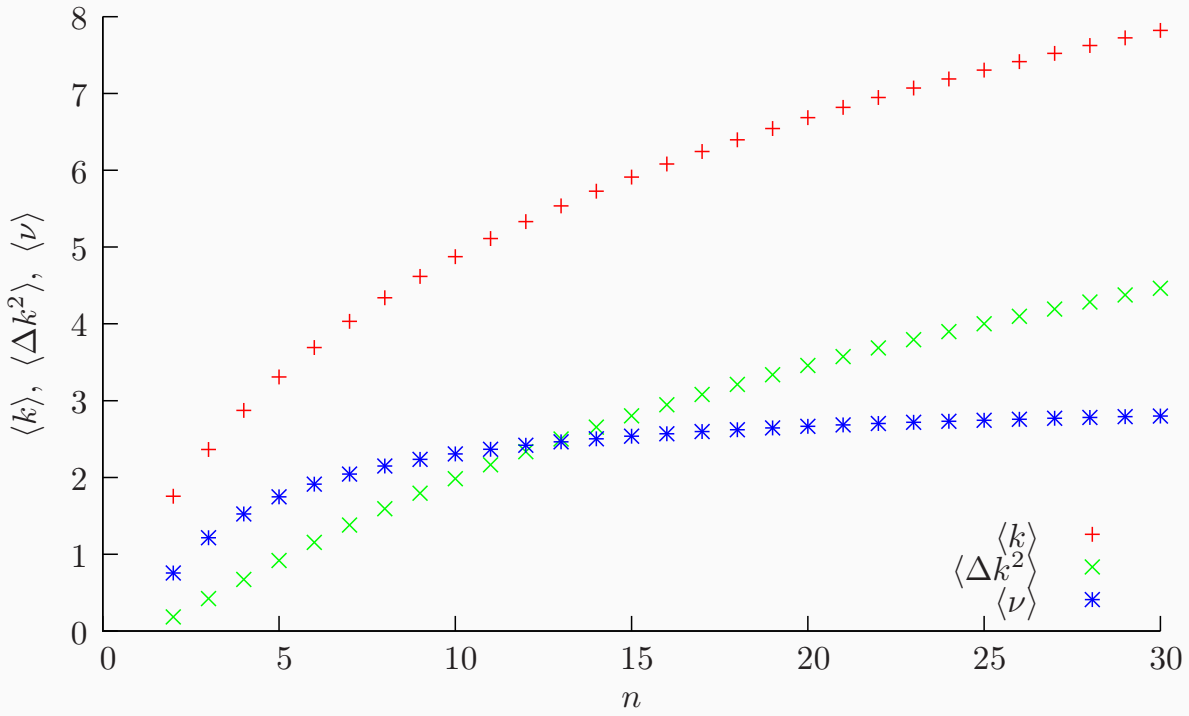
- Distribution of the number k of segregating alleles:

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

- Distribution of the number ν of singletons:

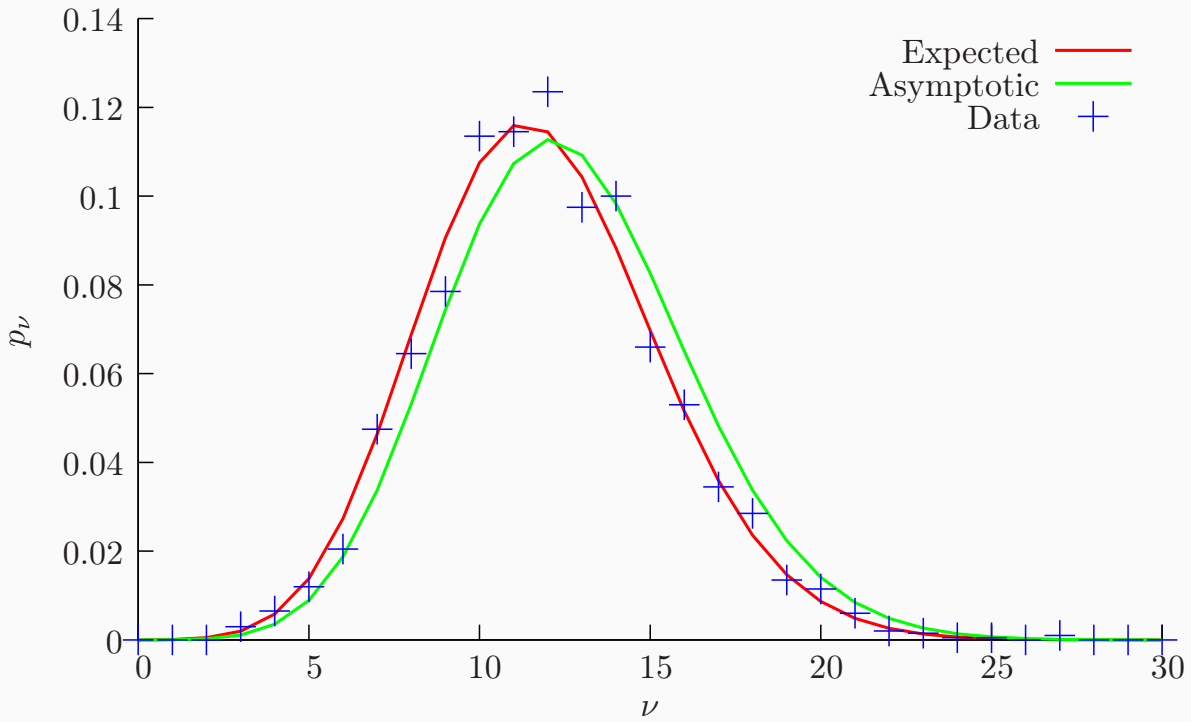
$$\begin{aligned} p_\nu(n+1) &= \frac{\theta}{\theta+n} p_{\nu-1}(n) + \frac{\nu}{\theta+n} p_{\nu+1}(n) + \frac{n-\nu}{\theta+n} p_\nu(n) \\ \overline{\nu(n)} &= \frac{n\theta}{\theta+n-1} \end{aligned}$$

Observables



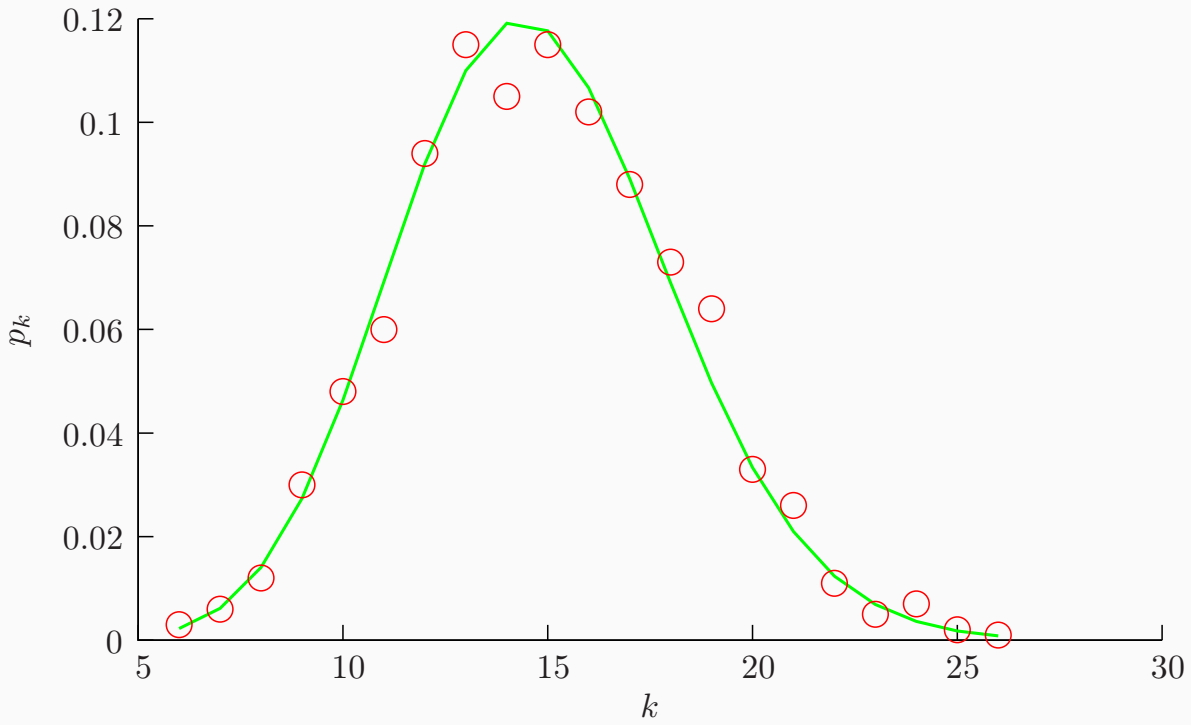
Average $\overline{k}$, variance $\overline{\Delta k^2}$ of segregating alleles and average $\overline{\nu}$ of singletons vs. n for $\theta = 3.1$

Observables



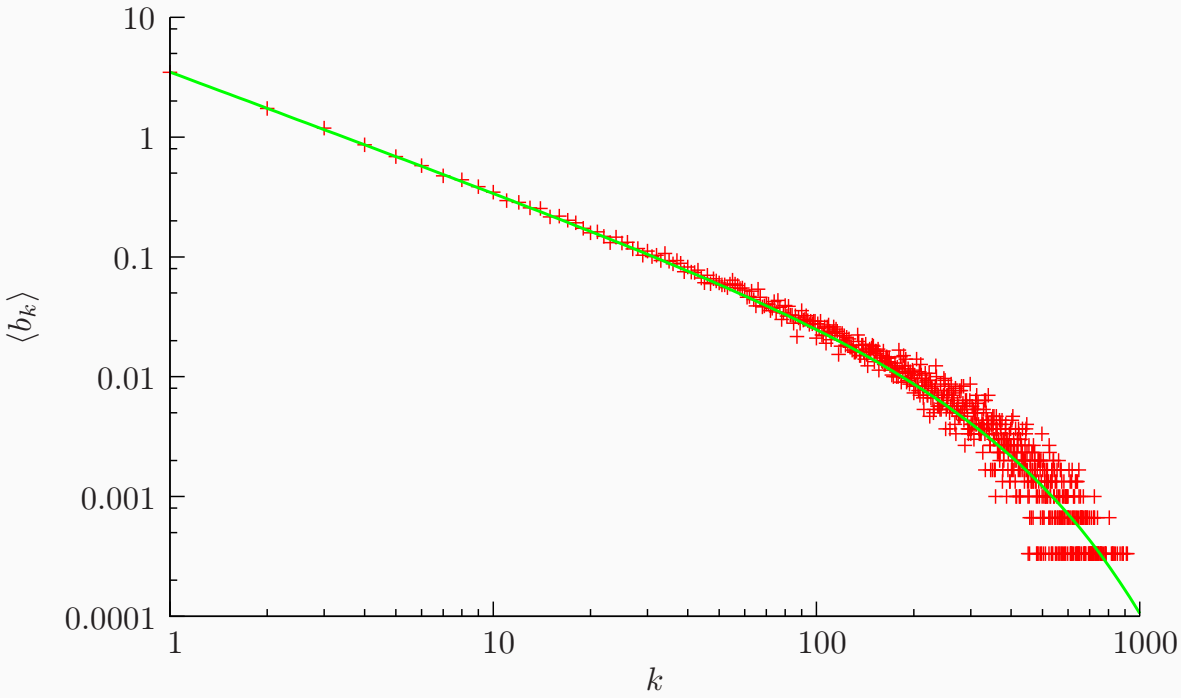
Distribution p_ν of the number of singletons for $n = 200$ and $\theta = 12.6$, together with the asymptotic distribution for $n \to \infty$ and simulation data over 1000 samples

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}

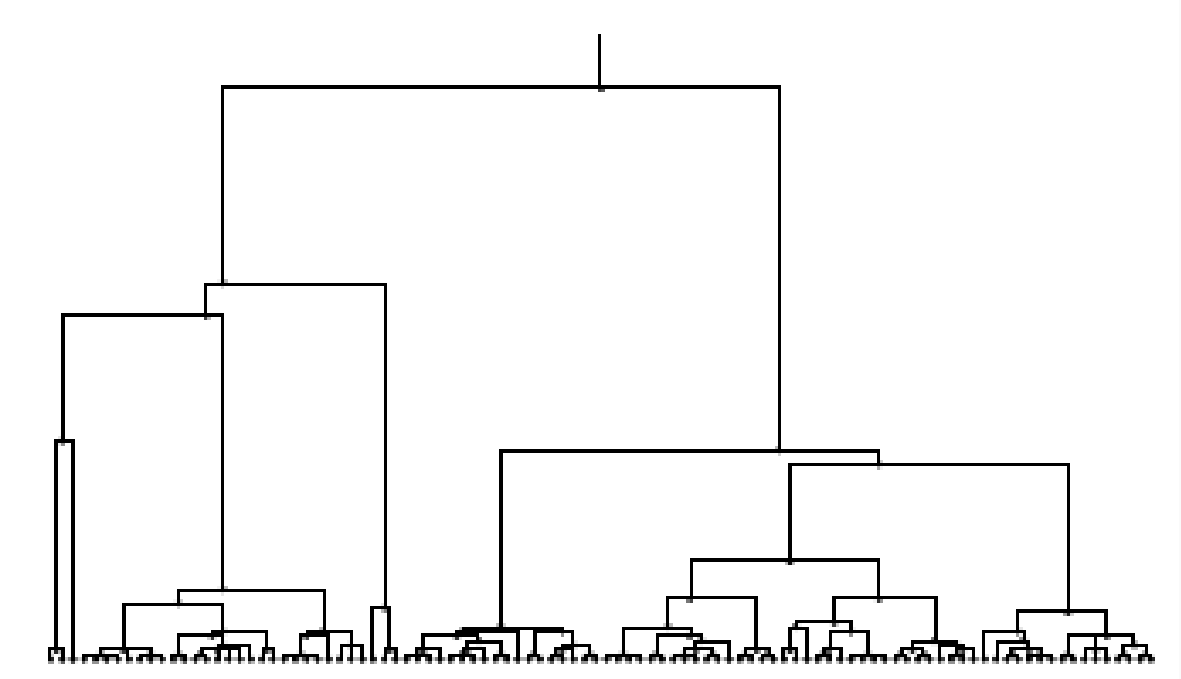
In practice, N_{e} is chosen to fit the data:

- 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!)

The Coalescent with selection

The Coalescent in the presence of selection

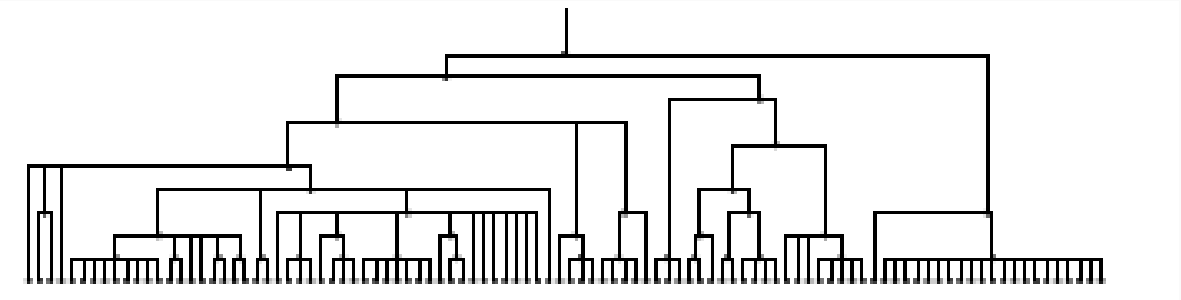
Brunet, Derrida *et al.*, 2006–2012



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

The Coalescent in the presence of selection

Brunet, Derrida *et al.*, 2006–2012



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

Coalescent models

A general coalescence model (Λ-coalescent):

- One starts with N points: in each interval of duration $\mathrm{d}t$ there is a probability $\pi_k \, \mathrm{d}t$ for every subset of k points to coalesce into one
- Then for some measure Λ one has

$$\pi_k = \int_0^1 x^k \, \Lambda(\mathrm{d}x)$$

- Rate $\lambda_{b,k}$ at which k ($2 \leq k \leq p$) points out of p coalesce into one is given by

$$\lambda_{p,k} = \int_0^1 x^{k-2}(1-x)^{p-k}\lambda(\mathrm{d}x) = \sum_{n=0}^{p-k} \frac{(p-k)!}{n!(p-k-n)!}(-1)^n \pi_{n+k}$$

- $r_p(\ell) \, \mathrm{d}t$: probability of having ℓ lineages at time $t + \mathrm{d}t$ if there are p lineages at time t:

$$r_p(\ell) = \frac{p!}{(\ell-1)!(p-\ell+1)!}\lambda_{p,p-\ell+1}$$

Coalescent models

- The Kingman coalescent:

$$\pi_2 \neq 0 \qquad \pi_k = 0, \quad \forall k > 2$$

- The *Bolthausen-Sznitman coalescent*:

$$\pi_k = \frac{\pi_2}{k-1}$$

Selection models

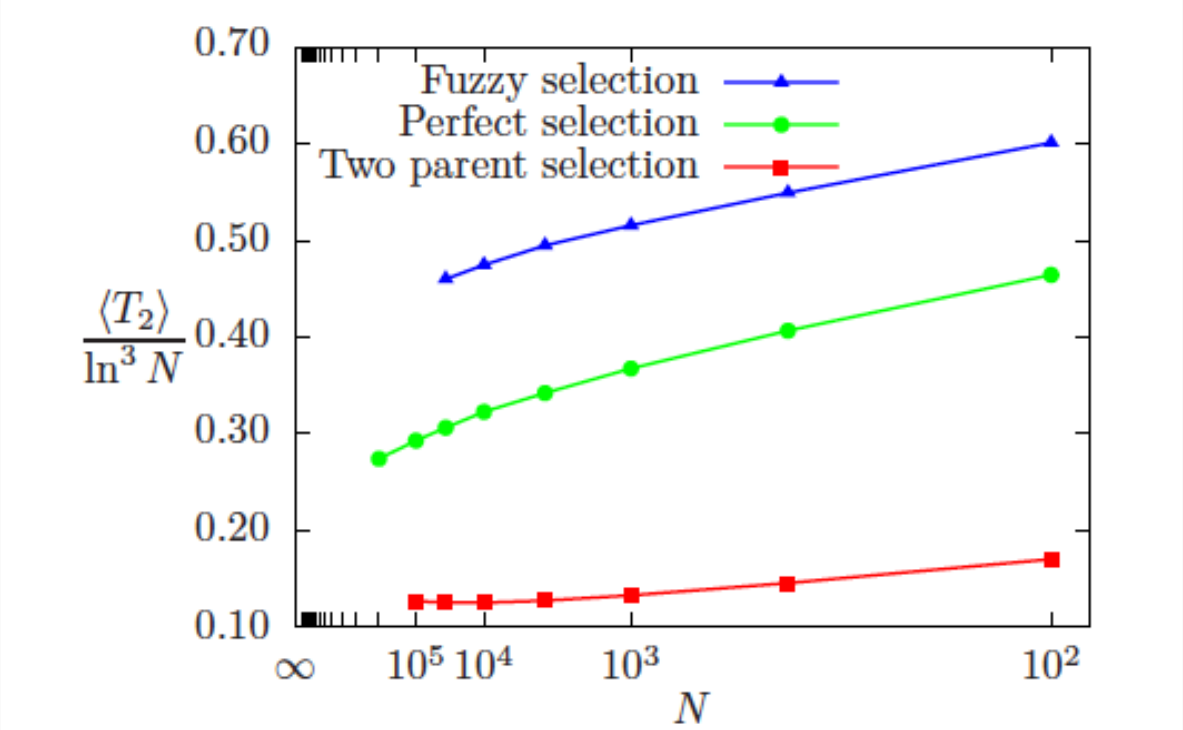
Brunet et al., 2006–2012

- Each individual has two potential offspring
- The fitness of each offspring is shifted by z wrt to the parent's one, with pdf $\rho(z)$ (flat in the simulations)
- Selection modes:
  - *Perfect selection*: The best N are retained
  - *Fuzzy selection*: Random choice among the $3N/2$ best
  - *Two-parent selection*: Each individual chooses two parents, but only the better one is kept

Selection models

Brunet et al., 2006–2012

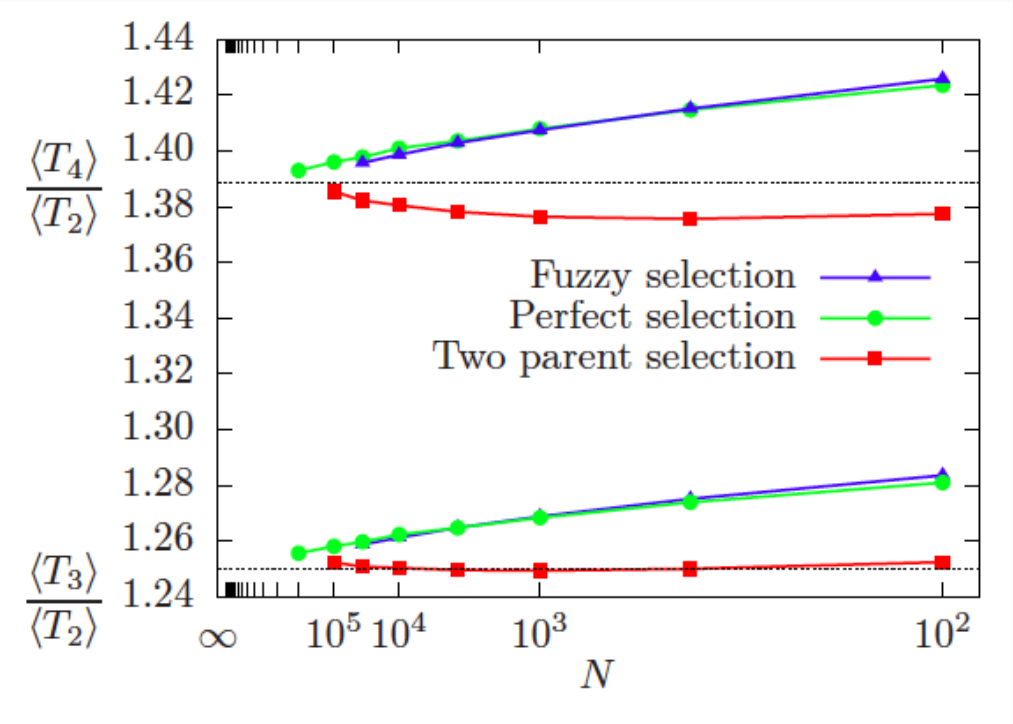
T_2: coalescence time for 2 lineages



Selection models

Brunet et al., 2006–2012

T_p: coalescence time for p lineages



- Kingman: $\langle T_4 \rangle / \langle T_2 \rangle = 3/2$; $\langle T_3 \rangle / \langle T_2 \rangle = 4/3$
- Bolthausen-Sznitman: $\langle T_4 \rangle / \langle T_2 \rangle = 25/18$; $\langle T_3 \rangle / \langle T_2 \rangle = 5/4$

Selection models

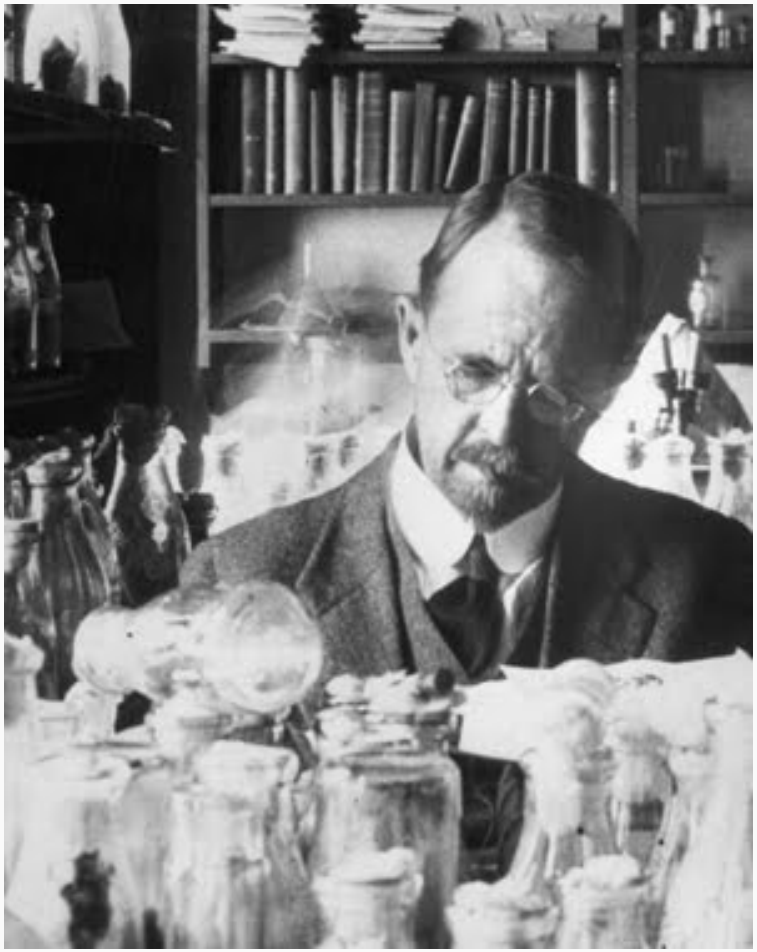
Brunet et al., 2006–2012

Coalescence time scale: $\overline{T_2} \sim \log^3 N$
Phenomenological theory

- The population looks like an advancing Kolmogorov-Fisher wave in “fitness” space
- Most of the time its motion is deterministic
- At intervals $\sim \log^3 N$ exceptionally “adapted” individuals arise
- These individual “sweep” a finite fraction of the population in a short time (multiple coalescence!)
- The distribution of the “sweep” sizes corresponds to the Bolthausen-Sznitman coalescent

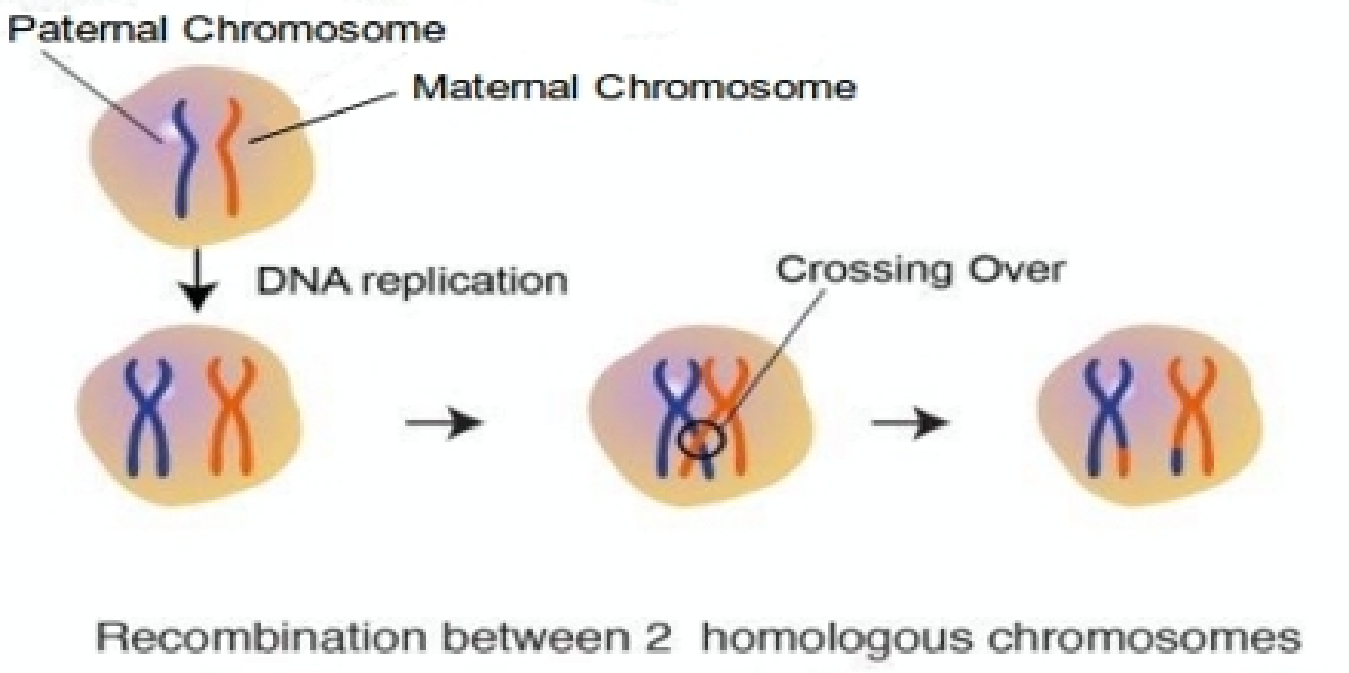
Recombination

Thomas Hunt Morgan



Recombination

- *Recombination* is a process leading to different assortments of genetic materials in life forms undergoing sexual reproduction
- It takes place in *meiosis* via the exchange of DNA segments between homologous chromosomes



Recombination between 2 homologous chromosomes

Linkage equilibrium

- Two loci, A and B, two alleles: A, a and B, b, random mating, no selection
- Allele frequencies: x_i ($i \in \{\text{A,a,B,b}\}$)
- Recombination *does not change* allele frequencies
- Change in genotype frequencies in one generation, e.g.:

$$x'_{\text{AB}} = \underbrace{(1-r)x_{\text{AB}}}_{\text{no recombination}} + \underbrace{r\ x_{\text{A}}x_{\text{B}}}_{\text{recombination}}$$

- *Linkage Equilibrium*: set $x'_{\text{AB}} = x_{\text{AB}}$

$$x_{\text{AB}} = x_{\text{A}}x_{\text{B}}$$

Linkage disequilibrium

- Deviation from equilibrium: $x_{\mathsf{AB}} = x_{\mathsf{A}} x_{\mathsf{B}} + D$, $x_{\mathsf{Ab}} = x_{\mathsf{A}} x_{\mathsf{b}} - D$, etc.

$$D = x_{\mathsf{AB}} x_{\mathsf{ab}} - x_{\mathsf{Ab}} x_{\mathsf{aB}}$$

- After one round of mating, one has

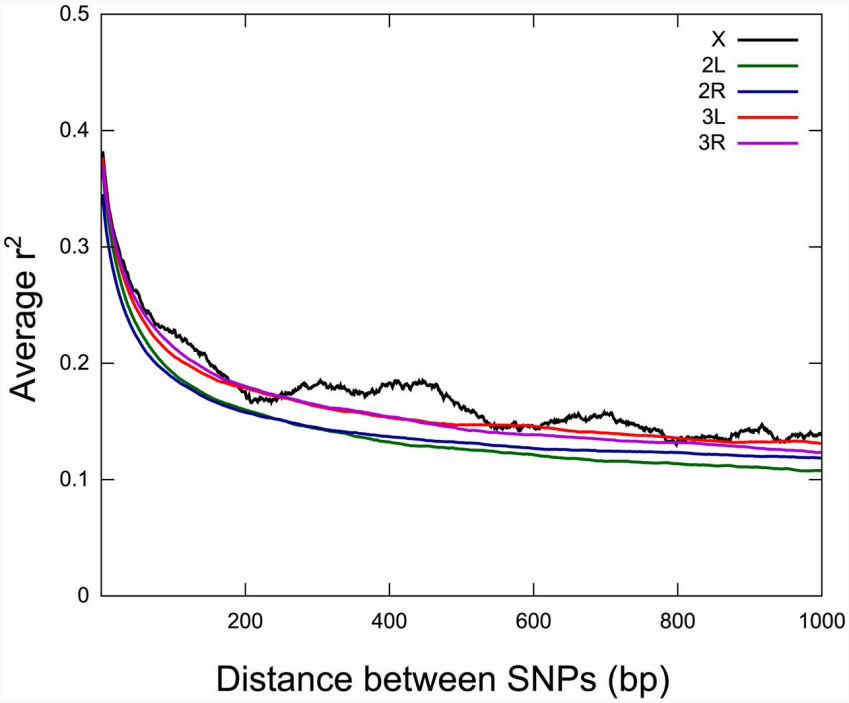
$$D' = (1 - r) D$$

- Thus $D(t) = (1 - r)^t D_0 \approx \mathrm{e}^{-rt} D_0$
- Empirical measure of linkage disequilibrium (LD) (unfortunately also denoted by r):

$$\mathsf{r}^2 = \frac{D^2}{x_{\mathsf{AB}} x_{\mathsf{ab}} x_{\mathsf{Ab}} x_{\mathsf{aB}}}$$

LD decay

LD r^2 vs. distance along the genome in *Anopheles arabiensis*



Marsden et al. 2014

The recombination rate r between two loci increases (roughly linearly) with the distance

Hitchhiking

Hitchhiking: Effect of positive selection of one allele at one locus has on alleles at neighboring loci

- Two loci, two alleles: A, a and B, b
- Fitness table (haploid):

	B	b
A	$1+s$	$1+s$
a	1	1

- Genotype frequency: $x_{\alpha\beta}$, $\alpha \in \{\mathsf{A}, \mathsf{a}\}$, $\beta \in \{\mathsf{B}, \mathsf{b}\}$
- Allele frequency: $x_\alpha = \sum_\beta x_{\alpha\beta}$, $x_\beta = \sum_\alpha x_{\alpha\beta}$
- Conditional allele frequency: $\xi_{\alpha\beta} = x_{\alpha\beta}/x_\alpha$
- Genotype frequencies (haploid) (forget about dominance/recessivity!!!):

	B	b
A	$x_\mathsf{A}\xi_\mathsf{AB}$	$x_\mathsf{A}\xi_\mathsf{Ab}$
a	$x_\mathsf{a}\xi_\mathsf{aB}$	$x_\mathsf{a}\xi_\mathsf{ab}$

- Mean fitness: $\langle w \rangle = 1 + x_\mathsf{A} s$

Hitchhiking

- Evolution equation for x_A: $x'_\mathsf{A} = x_\mathsf{A}(1+s)/(1+x_\mathsf{A}s) \Rightarrow$ $x_\mathsf{A}(t) = x_\mathsf{A}(0)(1+s)^t/(1-x_\mathsf{A}(0)(1-(1+s)^t))$
- Evolution equation for x_AB:

$$(1+x_\mathsf{A}s)^2 x'_\mathsf{AB} = (1+s)\left[x_\mathsf{AB}(1+x_\mathsf{A}s) + r\left(x_\mathsf{A}x_\mathsf{aB} - x_\mathsf{a}x_\mathsf{AB}\right)\right]$$

- Evolution equation for x_aB:

$$(1+x_\mathsf{A}s)^2 x'_\mathsf{aB} = (1+x_\mathsf{A}s)x_\mathsf{AB} + r(1+s)\left(x_\mathsf{a}x_\mathsf{AB} - x_\mathsf{A}x_\mathsf{aB}\right)$$

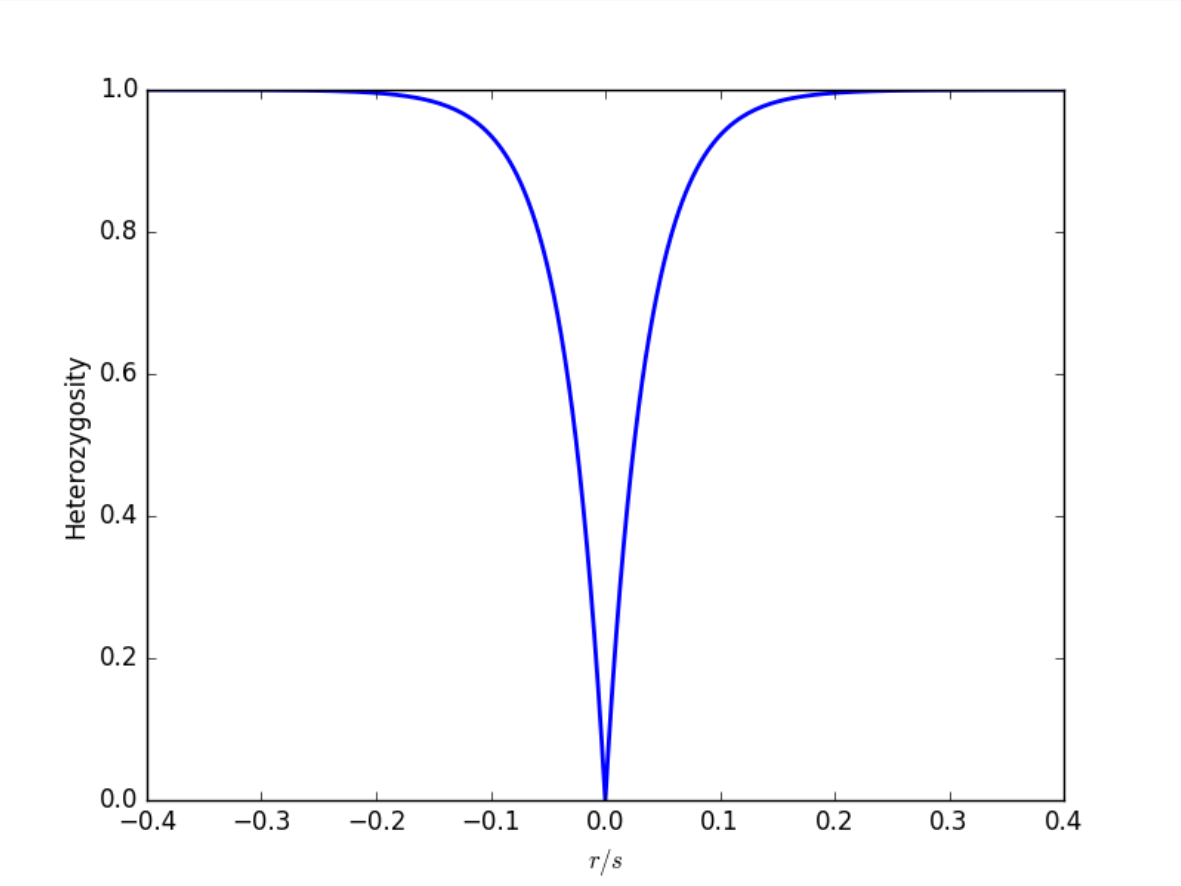
- This implies

$$\xi'_\mathsf{AB} - \xi'_\mathsf{aB} = (1-r)(\xi_\mathsf{AB} - \xi_\mathsf{aB})$$

- Assume that initially $\xi_\mathsf{AB} = 0$, i.e., that A originates in a b background, then when A is fixated, we have

$$\xi_\mathsf{AB}(\infty) = r\xi_\mathsf{aB}(0)(1-x_\mathsf{A}(0))\sum_{t=0}^{\infty}\frac{(1-r)^t}{1-x_\mathsf{A}(0)+x_\mathsf{A}(0)(1+s)^{t+1}}$$

Hitchhiking



Heterozygosity $4x_{\mathrm{B}}(\infty)x_{\mathrm{b}}(\infty)$ as a function of r/s for $s = 0.1$, $x_{\mathrm{B}}(0) = 0.5$ and $x_{\mathrm{A}}(0) = 10^{-6}$

Genetic draft

- Which allele hitchhikes on an advantageous allele going to fixation is “chosen” at random
- This introduces an additional random factor called *Genetic draft*
- Assume $r = 0$ (for simplicity) and an initial frequency $x_{\mathrm{B}}(0) = p$
- Then $x_{\mathrm{B}} = 1$ with probability p and 0 with prob. $1 - p$
- We have of course $\langle x_{\mathrm{B}} \rangle (\infty) = p$, $\left\langle \Delta x_{\mathrm{B}}^2(\infty) \right\rangle = p(1-p)$
- If the “sweep” takes place with prob. ρ we have the same average, but $\left\langle \Delta x_{\mathrm{B}}^2(\infty) \right\rangle = \rho p(1-p)$
- This is reminiscent of neutral drift, with effective population $N_{\mathrm{e}} = 1/2\rho$

Genetic draft

- Assume fixation takes place in a time *short* wrt the time between fixations
- Then successive sweeps are independent (Bernoulli) and fixation times are Poissonian
- Recombination: mutation arises in a single copy of the genome, that eventually reaches frequency y
- Then we have

$$x_{\mathrm{B}}(\infty) = \begin{cases} p, & \text{with probability } 1-\rho \qquad \text{no sweep} \\ p(1-y)+y, & \text{with probability } \rho p \\ p(1-y), & \text{with probability } \rho(1-p) \end{cases}$$

- These effects tend to increase variability, i.e., to reduce the effective population size
- For large population, genetic draft dominates: the effective population size due to draft is given by

$$N_{\mathrm{e}} = \frac{N}{1 + 2N\rho\,\langle y^2 \rangle}$$

Recombination and Epistasis

- Deviations from linkage equilibrium arise due to selective effects involving two (or more) loci
- Set, e.g., $\sigma_{\mathsf{A}} = +1$, $\sigma_{\mathsf{a}} = -1$, $\sigma_{\mathsf{B}} = +1$, $\sigma_{\mathsf{b}} = 1$, and assume $w_{\alpha\beta}$ has the form

$$w_{\alpha\beta} = f_0 + f_\alpha \sigma_\alpha + f_\beta \sigma_\beta + \underbrace{f_{\alpha\beta}\sigma_\alpha\sigma_\beta}_{\text{epistasis}}$$

- Then selection introduces correlations between loci
- Define

$$R = \frac{x_{++}x_{--}}{x_{+-}x_{-+}}$$

then one can show that

$$\langle w \rangle \, \Delta \log R = 4f_{12} - r(R-1)H,$$

where

$$H = \frac{x_{+-}w_{+-}x_{-+}w_{-+}}{\langle w \rangle^2} \sum_{\sigma_\alpha \sigma_\beta} \frac{1}{x_{\alpha\beta}}$$

Recombination and Epistasis

- If $0 < f_{12} \ll r$ R will reach values close to 1 very quickly, and then evolve slowly on the scale of f_{12}
- If $R \approx 1$ then $H \approx 1$ and we obtain a quasi-stationary state with

$$R \approx 1 + \frac{f_{12}}{r}$$

- This state has been called *Quasi-linkage equilibrium*
- It has been generalized to many interacting loci by Neher and Shraiman

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. J. Maynard Smith and J. Haigh, The hitch-hiking effect of a favourable gene, *Genet. Res. Camb.* **23** 23-56 (1974)

References iv

15. Clare Diana Marsden, Yoosook Lee, Katharina Kreppel, Allison Weakley, Anthony Cornel, Heather M. Ferguson, Eleazar Eskin and Gregory C. Lanzaro, Diversity, differentiation, and linkage disequilibrium: Prospects for association mapping in the malaria vector *Anopheles arabiensis*, *G3: Genes, Genomes, Genetics* **4** 121-131 (2014)
16. J. H. Gillespie, Genetic drift in an infinite population: The pseudohitchhiking model, *Genetics* **155** 909–919 (2000)
17. R. A. Neherand B.I. Shraiman, Genetic draft and quasi-neutrality in large facultatively sexual populations, *Genetics* **188** 975–996 (2011)
18. M. Kimura, Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection, *Genetics* **52** 875-890 (1965)

References v

19. R. A. Neher and B. I. Shraiman, Statistical genetics and evolution of quantitative traits, *Rev. Mod. Phys.* **83** 1283 (2011)