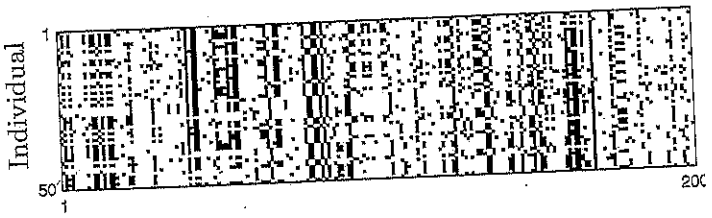
8 Population genetics

8.1. Introduction

ancestral type ■
derived type □

Pattern of mutation (genetic mosaic)



site index

Empirically observed patterns of genetic variation

Individual
ACTTTCGGAA ...
ACTTTCGCAA ...
ACTTTCGGAA ...
ACTTTCGCAA ...

site index (position along chromosome)

How does genetic mosaic reflect the the history of a population (its ancestry)?

Most recent common ancestor of Human population ca. 200 000 years ago.

To infer ancestry from genetic mosaic need a model for genealogies.

Which factors affect genetic evolution?

1. Inheritance
2. Mutations
3. Selection
4. recombination
5. demography (migration patterns, population size, ..)

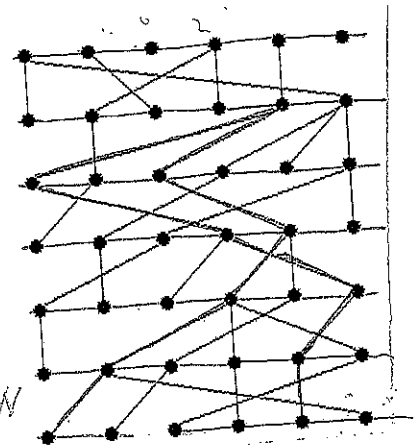
Hypothesis of neutral evolution

> Variation in large parts of genome can be explained without invoking Selection

- If selection not important, which factors affect variation in a neutral region?
- What difference does selection make where it matters?
- How can we find genomic regions where it matters?

Genes (regions expressed in protein sequences) are candidates. How to find genes?

5.2 Fisher-Wright model



Model for genealogy of selectively neutral locus

Assumptions

1. discrete non-overlapping generations $t=1,2,3,\ldots$
2. constant (haploid) population size N
3. freely mixing population
4. Mendelian inheritance (multinomial distribution of family sizes)

(3)+(4) ⇒ random sampling with replacement

Number of generations to most common recent ancestor = 5 for sample of size $n=2$ illustrated in the plot above.

random sampling with replacement ⇒ fixation

All genetic differences between individuals must eventually disappear (fixation).

However mutations cause differences to appear.

Different types of mutations.

- Single nucleotide polymorphism (SNP)

  ACCTGTT ⇩ ACCTCTT

- repeats in microsatellite loci (repetitive DNA sequence)

  ··· ATAG ATAG ATAG ··· ⇩ ··· ATAG ATAG ATAG ATAG ···

- inversion

Alleles: variants of sequences (genes) caused by mutation.

Neutral evolution mechanisms of copy & paste (differences disappear) and mutations (causing differences to appear) balance to create a steady state.

Quantify by population homozygosity F_2

F_2 = prob. that two alleles sampled from population are identical.

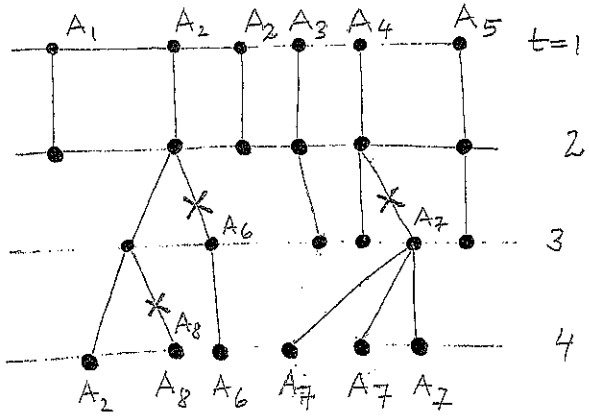
2.3 Infinite alleles model

Alleles = variants of a certain locus on chromosome.

A sequence of 100 nucleotides can have up to

$$4^{100} \simeq 10^{60}$$

different alleles. If a mutation strikes in this locus it is likely to create a new allele that did not exist in the population before.



Mutations with rate μ per individual per generation.

In this model two identical alleles must share the same history:

identity by state = identity by descent

Any given allele must eventually disappear from the population.
Classify genetic configuration of a population by allele frequencies (not types):

$$[\omega_1, \omega_2, \omega_3, \dots]$$

Where ω_1 is the frequency of alleles of one type, ω_2 that of another type, and so forth. The list is usually size ordered

$$\omega_1 \geq \omega_2 \geq \omega_3 \cdots$$

Population homozygosity satisfies recursion

$$F_2^{(t+1)} = (1-\mu)^2 \left[\frac{1}{N} + \left(1-\frac{1}{N}\right) F_2^{(t)}\right]$$

$F_2^{(t+1)}$: prob. to pick two identical alleles

$\frac{1}{N}$: pick same in random sampling

$1-\frac{1}{N}$: pick different

Steady state

$$F_2^{(t+1)} = F_2^{(t)} \Rightarrow F_2 = \frac{(1-\mu)^2}{1+2N\mu-N\mu^2+\mu^2-2\mu}$$

In the limit

$N \to \infty \quad \mu \to 0 \quad$ so that $\theta \equiv 2N\mu = \text{const.}$

(θ: population mutation rate)

find

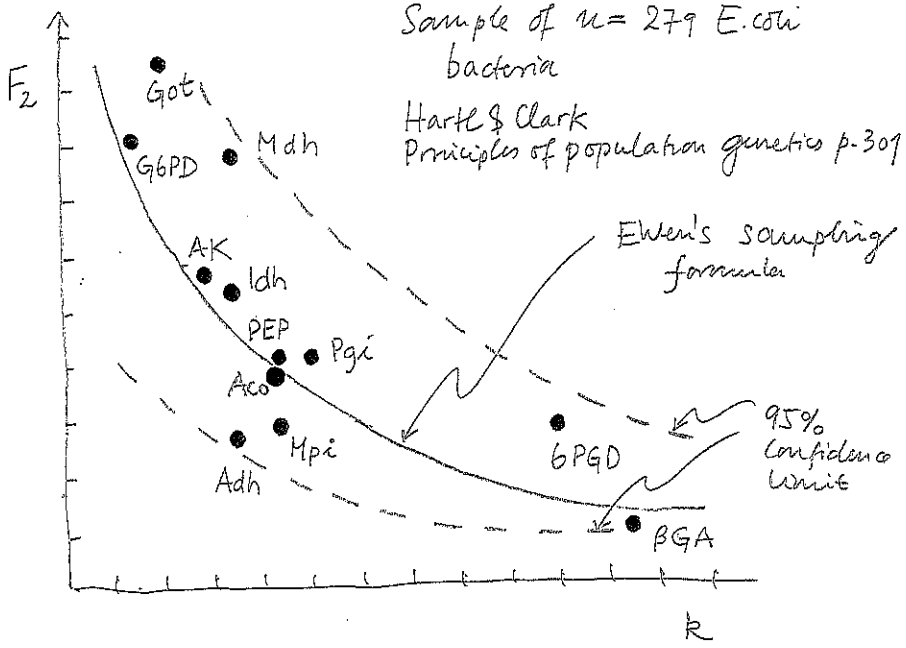
$$F_2 = \frac{1}{1+\theta}.$$

In a similar way

$$F_n = \frac{(n-1)!}{(1+\theta)(2+\theta)\cdots(n-1+\theta)}.$$

Expected number of allelic types in sample of size n (from Ewen's sampling formula)

$$\langle K \rangle = 1 + \frac{\theta}{1+\theta} + \frac{\theta}{2+\theta} + \cdots + \frac{\theta}{n-1+\theta}$$



Sample of $n = 279$ E. coli bacteria

Hartl & Clark Principles of population genetics p. 309

All loci fall into 95% confidence limit ⇒ neutral evolution, no selection.

2.4. Effective population size

In reality the population size is not constant

population expansions

bottlenecks

population-size fluctuations

Rapid population-size fluctuations ⇒ eff. pop. size N_{eff}.

Fisher-Wright model (no mutations) with constant N

$$1-F_2^{(t+1)} = \left(1-\frac{1}{N}\right)^t \left(1-F_2^{(1)}\right)$$

Now if N depends on time

$$1-F_2^{(t+1)} = \left(1-\frac{1}{N_t}\right)\left(1-\frac{1}{N_{t-1}}\right)\cdots\left(1-\frac{1}{N_1}\right)\left(1-F_2^{(1)}\right)$$

$$\equiv \left(1-\frac{1}{N_{eff}}\right)^t \left(1-F_2^{(1)}\right)$$

This defines

$$t \log\left(1-\frac{1}{N_{eff}}\right) = \sum_{j=1}^{t} \log\left(1-\frac{1}{N_j}\right)$$

Large N

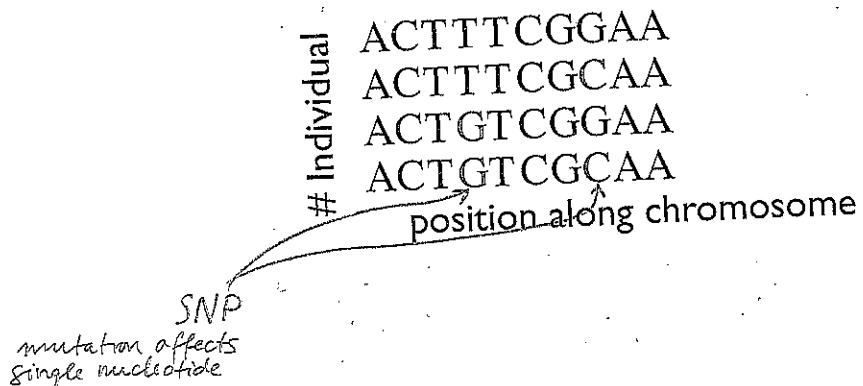
$$\frac{1}{N_{eff}} \approx \frac{1}{t}\sum_{j=1}^{t} \frac{1}{N_j}$$

When does this work?

geometric mean
⇒ small values of N_t matter

2.5. Single nucleotide polymorphisms (SNPs)

(infinite sites model)



Complete genetic sequences available for many organisms

Align sequences drawn randomly from population and compare.

In long DNA sequences most mutations (SNPs) occur at sites that were previously monomorphic.

Infinite-sites model assumes that

> every new single-site mutation occurs at a monomorphic site.

(closely related to infinite alleles model)

Distribution of number S_n of polymorphic sites in sample of size n.

$$P(S_n = j) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^j$$

Compare p. 7: $P(S_n = 0) = \frac{1}{1+\theta} = F_n$

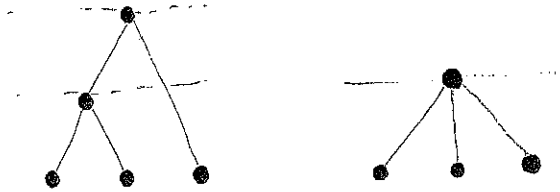
3.6. The coalescent process

Mathematical model for sample genealogy. In its simplest form coincident with Fisher-Wright model.

Goal: examine statistics of sample genealogies under different models (recombination, migration, selection, recombination, ...)

First consider gene genealogies in sample of size n from haploid population of size N. Assume neutral model.

Two examples for genealogies from p. 5



Question: What is the prob. that the gene sequences of the sampled individuals are the same?

In neutral model the statistics of sample genealogies is entirely determined by the random sequence of copy & paste events — independent of mutations. Can therefore answer the above question as follows

1. generate random sample genealogy with correct weight
2. scatter mutations randomly with rate μ
3. ask: What is the probability that no mutations fell on sample genealogy?

Begin with step (1).

Idea: Create sample genealogies backwards in time.

Probability that two alleles have same ancestor in previous generation $= N^{-1}$ in (haploid) population of size N.

Probability that the two alleles have different ancestors

$$P_2 = 1 - \frac{1}{N}.$$

Probability that all three alleles have different ancestors

$$P_3 = P_2 \underbrace{\frac{N-2}{N}}_{\text{prob. that 3rd allele has ancestor different from the other two}} = \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)$$

Probability that n alleles have different ancestors in previous generation

$$P_n = \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)$$
$$= \prod_{j=1}^{n-1}\left(1-\frac{j}{N}\right) \underset{n\ll N}{\approx} 1-\frac{1}{N}\sum_{j=1}^{n-1} = 1-\frac{\binom{n}{2}}{N}$$

where $\binom{n}{2} = \frac{n(n-1)}{2}$.

Show this by considering logarithm

$$\log P_n = \sum_{j=1}^{n-1} \log\left(1-\frac{j}{N}\right) \underset{\text{for } n\ll N}{\approx} -\frac{1}{N}\sum_{j=1}^{j}$$

So

$$P_n = 1 - \frac{\binom{n}{2}}{N} \quad \text{when} \quad n \ll N$$

What is the meaning of higher-order terms, N^{-2} for instance?

Higher-order collisions. The probability that three alleles have the same ancestor in the previous generation $\propto N^{-2} \ll N$ when N is large

Conclusion. For $n \ll N$ the right genealogy on $p.$ is unlikely to occur.
So as you trace the genealogy back in time observe only <u>binary coalescences</u> of ancestral lines.

In other words: genealogies are binary trees.
Stochastic process generating these genealogies $\equiv$ <u>coalescent process</u>.

Probability that n alleles have distinct ancestors T generations back and that two alleles have the same ancestor in generation $T+1$

$$P_n^T (1 - P_n).$$

Equivalent interpretation: distribution of number of generations to first coalescence of ancestral lines.

When N is large, the distribution of T is approximately exponential

$$\log P_n^T = T \log P_n \approx -T \frac{\binom{n}{2}}{N}$$

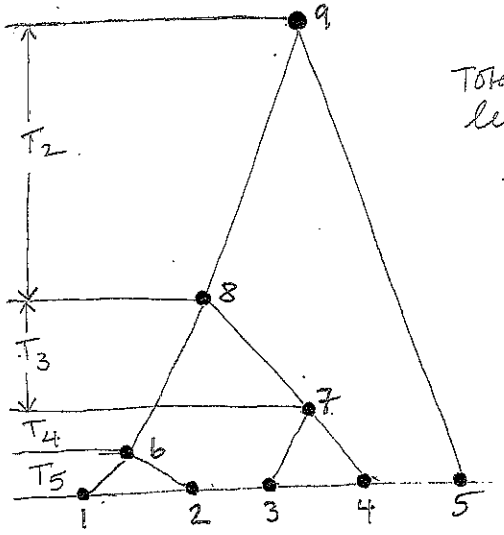
$$P_n^T \approx e^{-T \frac{\binom{n}{2}}{N}}$$

So

$$P_n^T (1 - P_n) \approx \frac{\binom{n}{2}}{N} e^{-T \frac{\binom{n}{2}}{N}}$$

In other words: time T_j to first coalescent event backward in time, starting with j lines

$$\text{Prob}(T_j) = \lambda_j e^{-\lambda_j T_j} \quad \text{with} \quad \lambda_j = \frac{\binom{j}{2}}{N}$$



Total branch length

$$T_c = \sum_{j=2}^{n} j\,T_j$$

Algorithm

Two arrays:

active lines

1	2	3	4	5	initially
1	–	3	4	5	after 1st coalescence
1	–	3	–	5	after 2nd coalescence

⋮

nodes

1	2	3	4	5	6	7	8	9

for each node store T_j

index of ancestor

indices of descendants

Coalescence rate for j lines

$$\lambda_j = \frac{\binom{j}{2}}{N}$$ ← number of possible pairs of lines

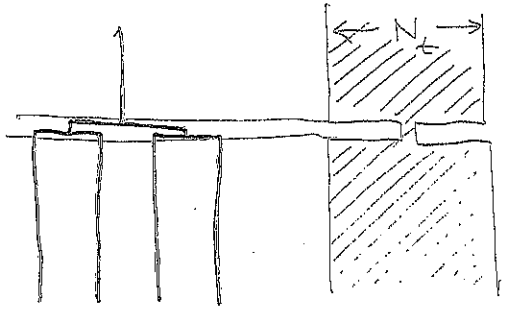
$$\langle T_j \rangle = \frac{N}{\binom{j}{2}}$$

Two conclusions:

(1) coalescent process is faster when there are more lines

(2) coalescent process is slower when the population size is larger

Population bottleneck



5.7. Adding mutations to the genealogy

Neutral model: mutations accumulate randomly with constant rate μ.

So the number of mutations on a genealogy is Poisson-distributed with rate μT_c

$$P(S=j) = \frac{(\mu T_c)^j}{j!} e^{-\mu T_c}$$

where $T_c = \sum_{j=2}^{n} j T_j$.

Average number of mutations

$$\langle j \rangle = \int_0^\infty dT_c \; \text{Prob}(T_c) \sum_{j=0}^{\infty} j \frac{(\mu T_c)^j}{j!} e^{-\mu T_c}$$

change summation index to $k = j-1$

$$\langle j \rangle = \mu \langle T_c \rangle \quad \underline{\text{molecular clock}}$$

In a sample of size $n=2$ $\quad \langle T_c \rangle = 2\langle T_2 \rangle$

Empirical data Humans $\quad \langle j \rangle \sim 10^{-3}$, $\mu \sim 5 \times 10^{-8}$ $\Rightarrow \langle T_2 \rangle \sim \frac{10^{-3} \cdot 10^8}{10}$

$\sim$ 10 000 gens.

$\sim$ 200 600 years

Since $\langle T_c \rangle = \sum_{j=2}^{n} j \langle T_j \rangle = 2N \sum_{j=1}^{n-1} \frac{1}{j}$

find

$$\langle j \rangle = \theta \sum_{j=1}^{n-1} \frac{1}{j}$$ weak dependence upon n

Example compute homozygosity F_2 with coalescent (p. 11)

$$P(S_2=0) = \langle e^{-2\mu T_2} \rangle$$

$$= \int_0^\infty \frac{dT_2}{N} e^{-2\mu T_2 - \frac{T_2}{N}} = \int_0^\infty dt \, e^{-(1+\theta)t}$$

$$= \frac{1}{1+\theta}$$

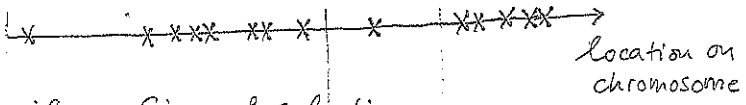
Exercise Derive an expression for $P(S_n=0)$ (p. 7)

Exercise: Compute the distribution of the number of SNPs in a sample of size $n=2$ in infinite-sites model. Answer

$$P(S_2=j) = \frac{1}{1+\theta} \left(\frac{\theta}{1+\theta} \right)^j$$

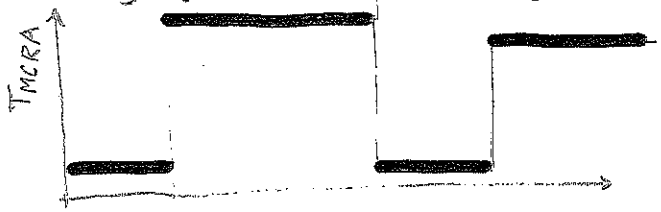
5.5. Recombination

Locations of SNPs in sample of size $n=2$

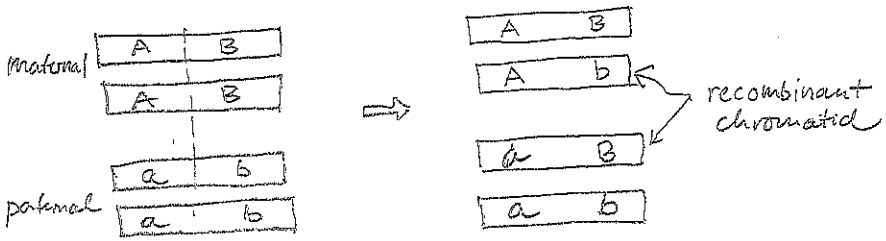


Not uniform. Sign of selection causing advantageous allele to fix?

Not necessarily. Recombination causes fluctuating gene histories along chromosome



Recombination: exchange of chromosome segments between paternal and maternal chromosomes during meiosis in eukaryotes (which usually carry two copies of each chromosome).



Consequence: different parts of a chromosome have different histories.

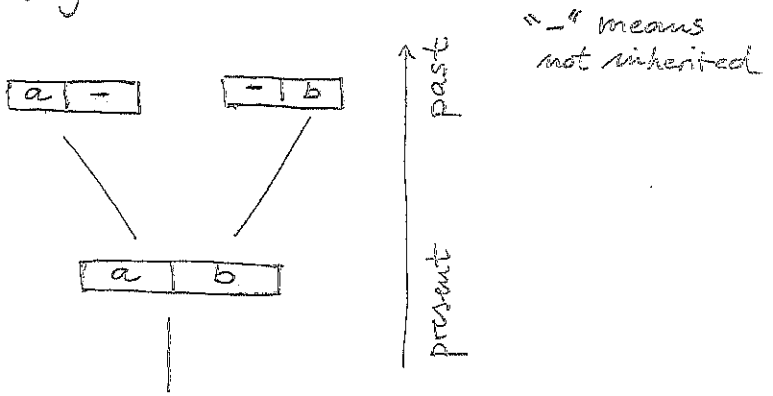
Model ① recombination events follow Poisson process in time, with rate r (between two loci, per individual, per generation).
② recombination sites uniformly distributed over chromosome.

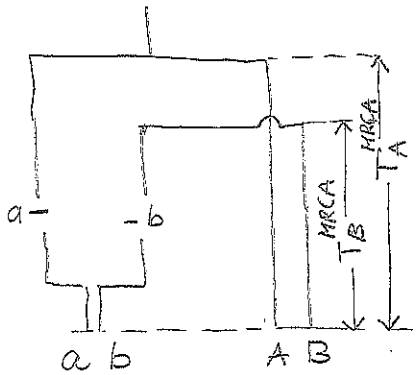
⇒ higher probability to observe recombination between two loci that are far apart.

linked loci: two closeby loci that have identical gene histories

linkage equilibrium: gene histories of two loci far apart from each other are independent.

ancestral recombination graph: coalescent no longer a tree





Coalescent with recombination.

For k ancestral lines, recombination occurs at rate kr

Coalescences occur at rate $\binom{k}{2}$

diploid $\longrightarrow \frac{\binom{k}{2}}{2N}$.

Measuring time in units of $2N$ generations

$$\lambda_C = \binom{k}{2} \qquad \lambda_R = 2Nkr \equiv \frac{kR}{2}$$

Both processes are Poisson, and independent.

If two times are independently exponentially distributed with rates λ_c and λ_R then the time to the first event

$$t_{min} = \min\{t_1, t_2\}$$

is exponentially distributed with rate $\lambda_c + \lambda_R$

$$\begin{aligned} P(t_{min} > T) &= P(\min\{t_1, t_2\} > T) \\ &= P(t_1 > T,\ t_2 > T) \\ &= P(t_1 > T)\, P(t_2 > T) \\ &= e^{-(\lambda_c + \lambda_R)T} \end{aligned}$$

Probability that coalescence occurs first is

$$\frac{\lambda_c}{\lambda_c + \lambda_R} = \frac{k-1}{k-1+R}$$

Probability that recombination occurs first

$$\frac{\lambda_R}{\lambda_c + \lambda_R} = \frac{R}{k-1+R}$$

Correlation of gene histories

$$\langle T_A^{MRCA} T_B^{MRCA} \rangle - \langle T_A^{MRCA} \rangle \langle T_B^{MRCA} \rangle$$

$$\sim R_{AB}^{-1} \quad \text{for large } R_{AB}.$$

3.4. Selection

Consider one locus, two allelic types a (new mutation) and A (ancestral type)

Assume that a has higher fitness

$$W_a = 1+s \text{ with } s>0 \qquad W_A = 1$$

Model effect of selection as bias in Fisher-Wright model: initial frequencies:

$$X_a^{(0)} \text{ and } X_A^{(0)} \qquad \text{in generation } t=0$$

Sample not with $X_a^{(0)}$ and $X_A^{(0)}$ but with bias

$$\frac{W_a}{W_a+W_A} X_a^{(0)} \text{ and } \frac{W_A}{W_a+W_A} X_A^{(0)}$$

In the limit of infinite population size

$$X_a^{(1)} = \frac{W_a X_a^{(0)}}{W_a X_a^{(0)} + W_A X_A^{(0)}} = \frac{(1+s) X_a^{(0)}}{(1+s) X_a^{(0)} + (1-X_a^{(0)})}$$

Change in allele frequency to next generation

$$X_a^{(1)} - X_a^{(0)} = \frac{(1+s) X_a^{(0)} - X_a^{(0)} \left[1+s X_a^{(0)}\right]}{1+s X_a^{(0)}}$$

$$\approx s X_a^{(0)} (1-X_a^{(0)}) \qquad \text{when } s \ll 1$$

When s is small and population size large then this change is small, so that

$$\frac{dx_a}{dt} \approx s\, y_a (1-y_a)$$

Logistic equation (p. 20) describes spreading of advantageous gene in population (<u>selective sweep</u>)

Duration of selective sweep

$$t_s = \frac{1}{s} \int_{N^{-1}}^{1-N^{-1}} \frac{dx}{x(1-x)} = \left[\log \frac{x}{x-1} \right]_{N^{-1}}^{1-N^{-1}} \approx \frac{2}{s} \log(N-1)$$

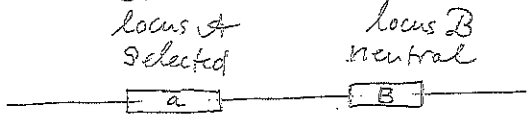
(I have set upper boundary to $1-N^{-1}$ because the time to reach $x_a = 1$ diverges in stochastic approximation)

3.10. Genetic hitchhiking

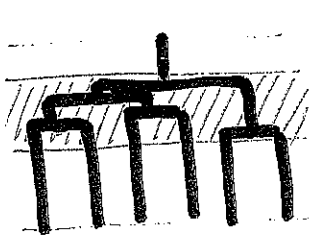
Recombination shapes the effect of mechanisms such as selection (migration, ...)

How to detect selection from genetic mosaics?

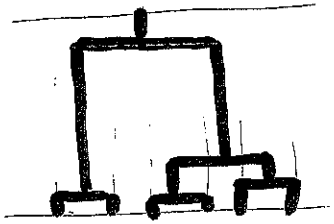
Signature of recent selective sweep on gene genealogy of neutral locus nearby?



Assume no recombination. Then genealogy of B looks like that of A



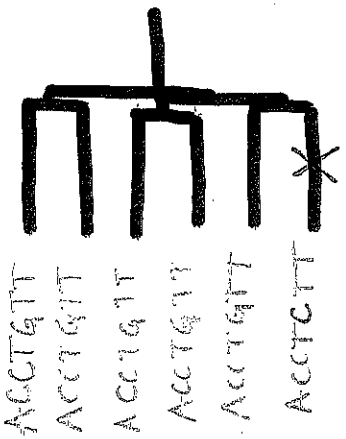
Sweep



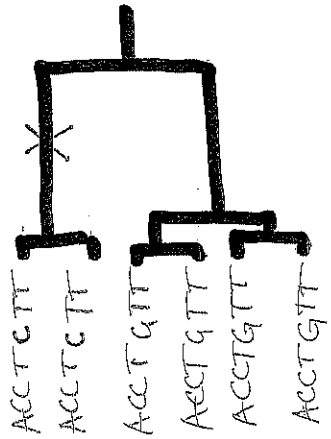
genealogy of locus B linked to selected locus

neutral genealogy

Star-like topology: high prob. of singletons



Neutral tree by contrast much much smaller prob. of singletons

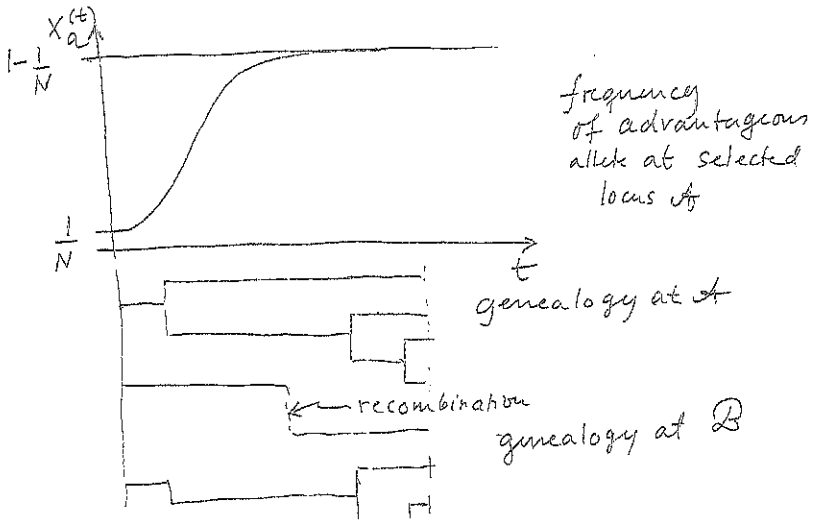


Statistical tests to infer recent selective sweeps from singleton frequency.

Problems: ① Singletons can be due to sequencing error

② Bottlenecks (reduction of population size during short time in the past) result in star-shaped genealogies.

Distinguish bottlenecks from recent selective sweeps by analysing how genealogies vary with distance from selected locus.
Recombination allows lines of neutral locus to escape sweep (avoid genetic hitchhiking)



To quantify how the effect of selection decays as distance to selected locus increases compute prob. Q of a line at locus B to escape the sweep

$$Q = \int_0^{t_s} dt\, r e^{-rt} \left(1 - X_a^{(t)}\right)$$

- $r e^{-rt}$: prob. that recombination at time t in the past per sequence
- $1 - X_a^{(t)}$: frequency of sequences with unfavoured allele A at A (recombination with a does not escape)

To compute Q, assume deterministic model

$$X_a^{(t)} = \frac{1}{1 + e^{-st}(N-1)} = \frac{1}{1 + e^{-s(t - \frac{t_s}{2})}}$$

with $t_s = \frac{2}{s} \log(N-1)$

So

$$Q = \int_0^{t_s} dt\, r e^{-rt} \left[1 - \frac{1}{1 + e^{-s(t - \frac{t_s}{2})}}\right]$$

Take limit

$N \to \infty$

$\frac{r}{s} \log N$ —— constant

In this limit the first term in $[\cdots]$ integrates to unity, so that

$$Q \approx 1 - \int_0^{t_s} dt\, r e^{-rt} \frac{1}{1 + e^{-s(t - \frac{t_s}{2})}}$$

Main contribution to integral comes from $t < \frac{t_s}{2}$ where one can approximate

$$\approx 1 - \int_0^{t_s} dt\, r e^{-rt} \cdot \frac{1}{e^{-s(t - \frac{t_s}{2})}}$$

$$\approx 1 - e^{-\frac{r}{s} \log N}$$

Heterozygosity reduced by hitchhiking:

$$\frac{H}{H_{\text{initial}}} \approx 1 - e^{-\frac{r}{s} \log N}$$