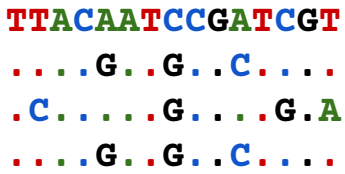
Definitions

Multiple sequence alignment (“MSA”): a representation of homologous positions from multiple genomes

Monomorphic site: a site that is invariant in a sample of sequences

Polymorphic site: a site that is variable in a sample of sequences (=segregating site)



See Fig 1.1 Hahn p. 2

Definitions

Locus: a location on a chromosome

Allele: one of two or more alternative forms of a locus

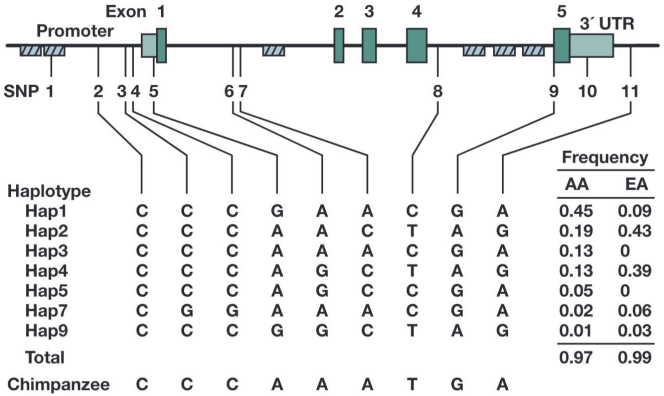
Definitions

Locus: a location on a chromosome

Allele: one of two or more alternative forms of a locus

Haplotype: an allele defined at the DNA sequence level and consisting of at least two positions

Example: haplotype frequencies in African Americans (AA) and European Americans (EA) at *PPIA* gene locus



Haplotype										Frequency	
										AA	EA
Hap1	C	C	C	G	A	A	C	G	A	0.45	0.09
Hap2	C	C	C	A	A	C	T	A	G	0.19	0.43
Hap3	C	C	C	A	A	A	C	G	A	0.13	0
Hap4	C	C	C	A	G	C	T	A	G	0.13	0.39
Hap5	C	C	C	A	G	C	C	G	A	0.05	0
Hap7	C	G	G	A	A	A	C	G	A	0.02	0.06
Hap9	C	C	C	G	G	C	T	A	G	0.01	0.03
Total										0.97	0.99
Chimpanzee	C	C	C	A	A	A	T	G	A		

An et al. (2007) PLOS Pathogens

Definitions

Locus: a location on a chromosome

Allele: one of two or more alternative forms of a locus

Haplotype: an allele defined at the DNA sequence level and consisting of at least two positions

Chromosome: a DNA sequence sampled from a population

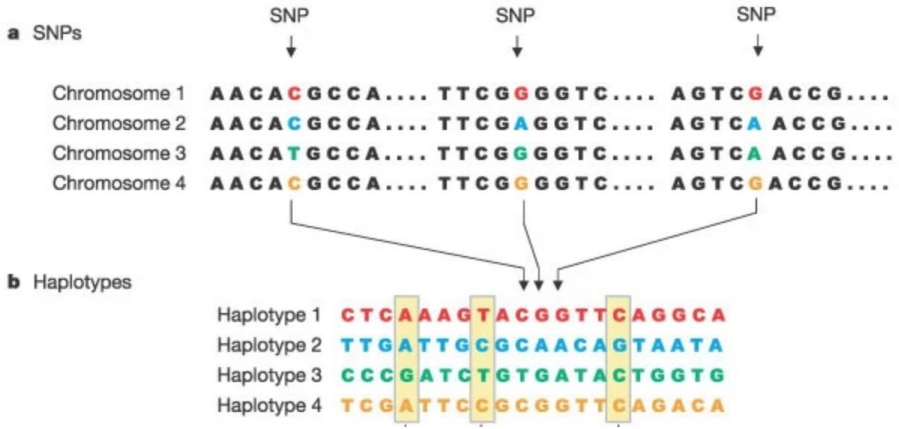
*note difference from conventional use of “chromosome”

Example: In the study of PPIA, 92 African Americans were sequenced at the PPIA gene. Therefore, since humans are diploid, 2 x 92 = 184 “chromosomes” were sequenced

Definitions

Single nucleotide polymorphism (“SNP”, pronounced “snip”): a polymorphism consisting of a single nucleotide base change

Tag SNP: a SNP that marks a block of correlated SNPs (i.e., a linkage disequilibrium block). Each tag SNP is uncorrelated with other tag SNPs



International HapMap Project (2003) Nature.

Mutation

Genetic variation in a population arises initially as a mutation in a single gamete

If there are N individuals in a diploid population, then the frequency of each new mutation must start at 1/2N

Generally, population genetic inferences are made from single base pair changes, or SNP (pronounced “snip”) that arise from point mutation

Population genetics is typically concerned with germline mutations only, somatic mutations (e.g., in tumors) are not considered

How many mutations per generation?

Human genome size: 3.1×10^{9} bp
Mutation rate: of $*1.2 \times 10^{-8}$ mutations per bp per generation
Number of copies of each chromosomes in diploid: 2

3.1×10^{9} bp * 1.2×10^{-8} per bp per generation * 2
= 74.4 mutations per generation

Kong et al. (2012) Nature

How many mutations per generation?

Human genome size: 3.1 x 10^9 bp
Mutation rate: of *1.2 x 10^{-8} mutations per bp per generation
Number of copies of each chromosomes in diploid: 2

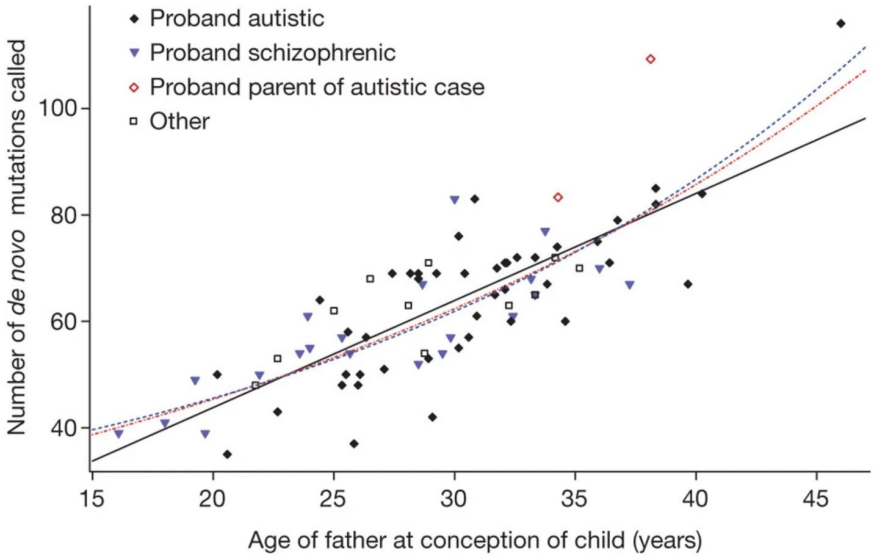
3.1 x 10^9 bp * 1.2 x 10^{-8} per bp per generation * 2
= 74.4 mutations per generation

Example:

Mutation rate study of Icelandic trios (3 genome sequences of parents + child) found average of 63 mutations

Strong dependency on father's age

Higher numbers of *de novo* mutations in older fathers may explain greater incidence of autism with increasing age of father



Kong et al. (2012) Nature

Infinite sites model of mutation

Infinite sites model makes simplifying assumption that each site has only mutated at most one time in a sample of sequences

Infinite sites model of mutation

Infinite sites model makes simplifying assumption that each site has only mutated at most one time in a sample of sequences

Example: How much sequence evolution (divergence) has occurred between human and chimpanzee sequences?

```
Human       TTACAATCCGCTCAT
Chimpanzee  TTACGATGCGCTCGT
```

Step 1: Count the number of mutational differences (k)

Step 2: Count total sites in the alignment (L)

Step 3: Calculate the distance as the proportion of sites that differ

$$\text{p-distance} = k/L = 3/15$$

The p-distance is a measure of sequence divergence assuming infinite sites

Genetic drift

Genetic drift is the stochastic change in allele frequencies in a population

Genetic drift occurs because of chance inheritance of alleles (i.e., some chromosomes leave more descendants then others)

Allele frequencies may either rise or fall due to genetic drift

Genetic drift

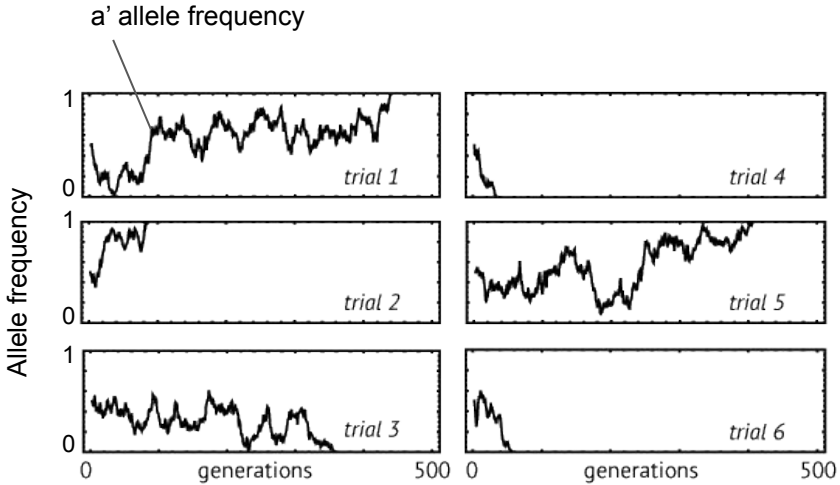
Example: Genetic drift in a simulated population

Consider a locus with two alleles, a and a’ in a population of 100 haploid individuals (i.e., 100 alleles)

At time zero, both alleles are at 50% frequency.

Each generation, the simulation generates a new sample of 100 alleles by drawing randomly from the pool of alleles in the prior generation (with replacement)

The figure shows the change in allele frequency of the a’ allele each generation (i.e., genetic drift)



Genetic drift

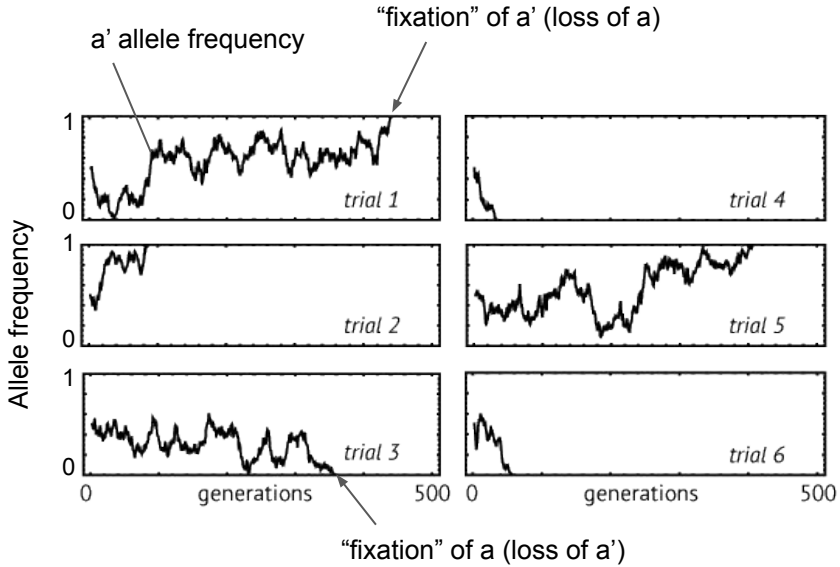
Example: Genetic drift in a simulated population

Consider a locus with two alleles, a and a’ in a population of 100 haploid individuals (i.e., 100 alleles)

At time zero, both alleles are at 50% frequency

Each generation, the simulation generates a new sample of 100 alleles by drawing randomly from the pool of alleles in the prior generation (with replacement)

The figure shows the change in allele frequency of the a’ allele each generation (i.e., genetic drift)



Genetic drift in simulated populations

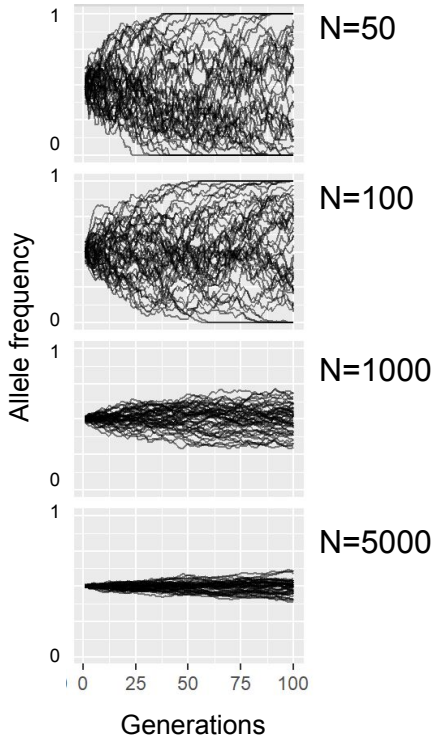
The rate of genetic drift (=the rate of allele frequency change) is dependent on the population size

Example: Simulation of a Wright-Fisher population with starting allele frequency of 0.5

Each panel represents a different constant population size

Each line represents an independent simulation

*Key point: Allele frequencies on average change more quickly in small populations compared to large



The Wright-Fisher model

To understand effects of drift, it is convenient to compare real populations to a hypothetical “idealized” population

A Wright-Fisher population, is a theoretical population with the following properties:

- *N* diploid hermaphrodites (self-compatible)
- Constant size (*N*) over time
- 2N chromosomes at each generation
- Random mating
- No individuals survive into the next generation (no overlapping generations)

Effective population size (N_e)

The effective size is the size of a Wright-Fisher population that experiences the same rate of genetic drift as a real population

N_e may be correlated with census population size, but knowledge of one may not inform the other

What factors affect N_e?

The number of breeding individuals

Variation in the number of offspring

Bottleneck/Founder effects

Migration and population structure

The concept of effective population size (N_e)

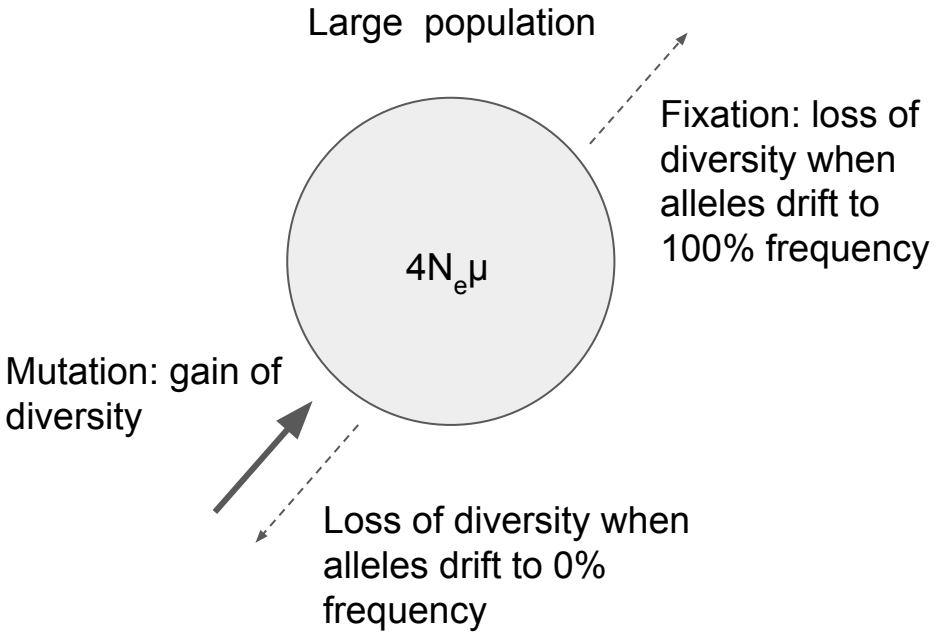
N_e has important implications for many populations properties

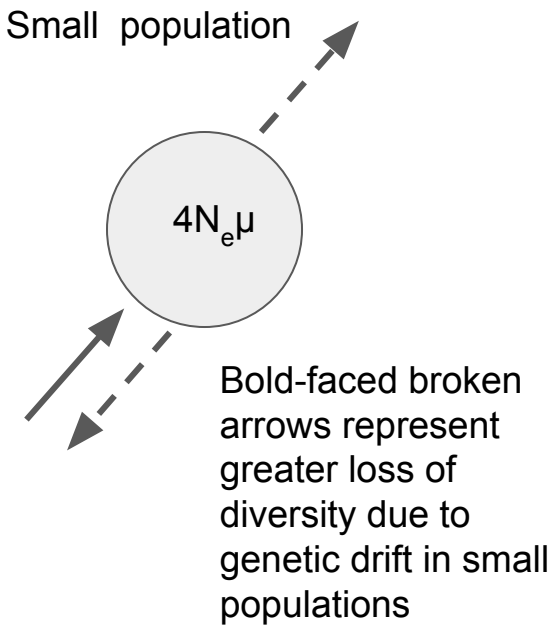
Some properties of populations that depend on N_e

- The probability of fixation (reaching 100% frequency) of neutral alleles
- The time to fixation of neutral alleles
- The coalescence time of two sequences
- The rate of loss of genetic diversity (e.g., heterozygosity)

Mutation-drift equilibrium

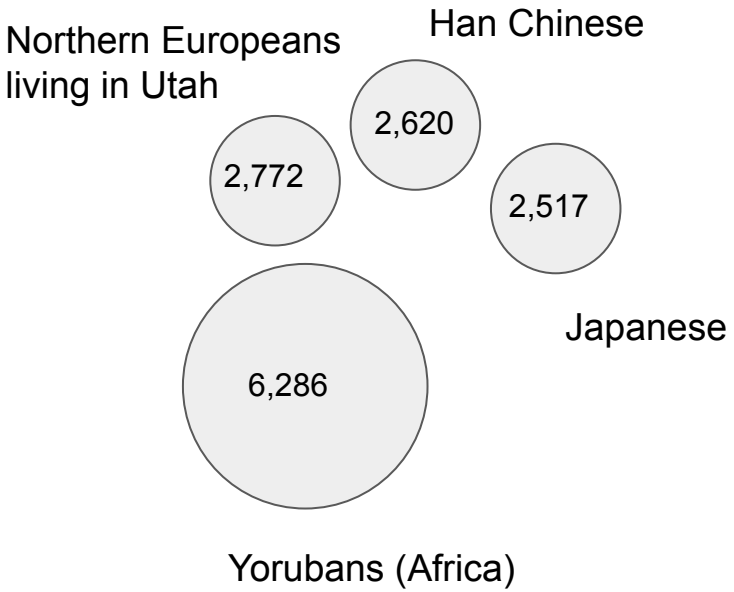
At mutation-drift equilibrium, genetic diversity should be correlated with the size of the population (e.g., in comparisons of different populations)





Effective population size (N_e) in comparative contexts

Estimates of N_e reflect major demographic events (e.g., population bottlenecks)



Tenesa et al. (2007) Genome Research

Natural selection

A new mutation can have the different effects on fitness

- advantageous (=beneficial)
- deleterious
- neutral

Loci with advantageous mutations are subject to “**positive”** selection

Those with deleterious mutations are subject to negative (**“purifying”**) selection

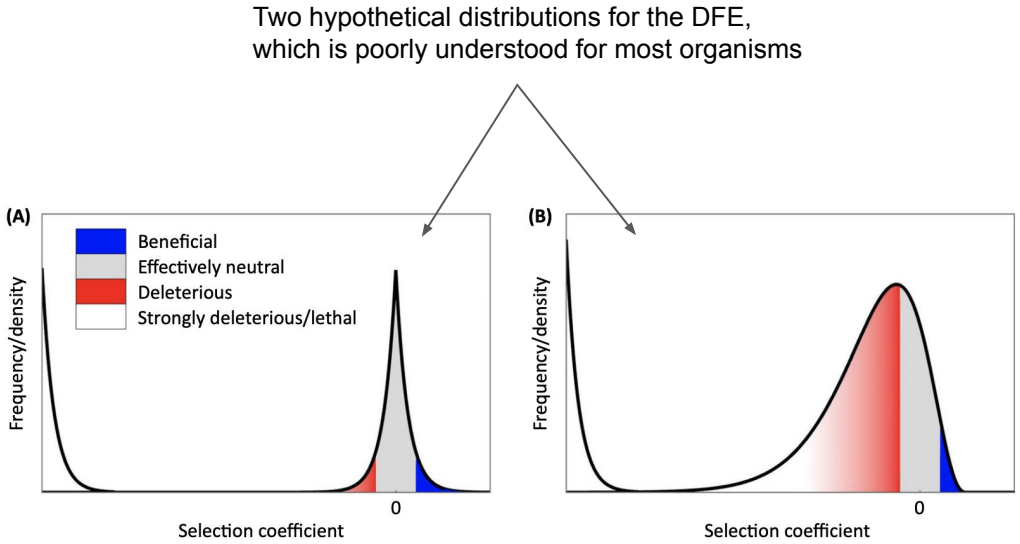
Natural selection

A major current area of research is to estimate the distribution of fitness effects (DFE) of new mutations

Selection coefficient = 0 indicates no effect on fitness

Selection coefficient > 0 indicates beneficial mutations that increase fitness

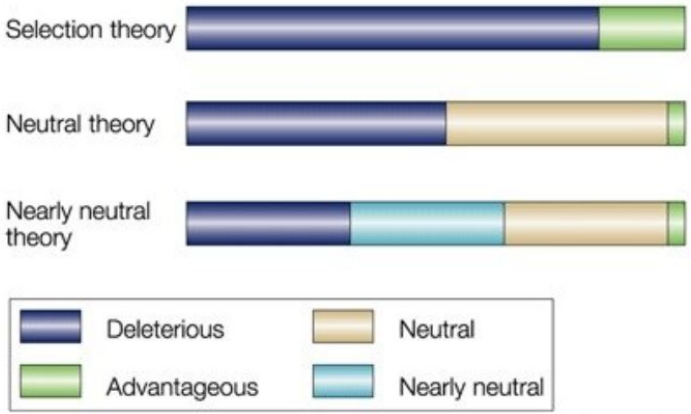
Selection coefficient < 0 indicates deleterious mutations that decrease fitness



The neutralist-selectionist debate

The distribution of fitness effects of new mutations (DFE) describes the fraction of mutations that are neutral, nearly neutral (i.e., weakly deleterious), and advantageous

Models of molecular evolution are distinguished based on different assumptions about this distribution



Bromham and Penny (2003) Nature Reviews Genetics

What’s coming up:

Week 1 Hahn Chapters 1 and 2

Week 2 (next week) Hahn Chapter 3

Week 2 recitation: Introduction to large-scale sequencing, the human reference genome, and parsing VCF with R