Bioinformatics 1

Population variation

Claudia Acquisti

Evolutionary Functional Genomics
Institute for Evolution and Biodiversity, WWU Münster
claudia.acquisti@uni-muenster.de

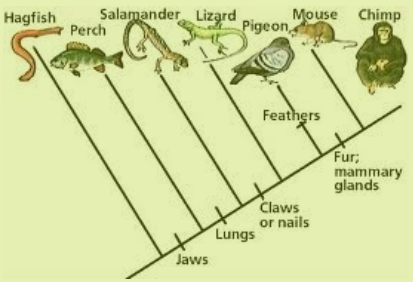
DNA damage

Copying errors

MUTATIONS: heritable changes to the genome, essential for evolution.

POLYMORPHISMS

?



SUBSTITUTIONS

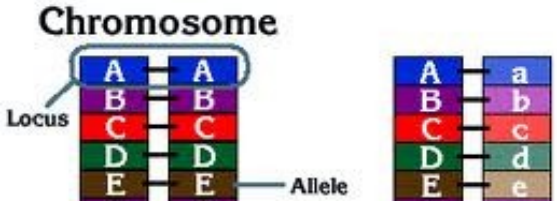
Phylogenetic trees

Test of selection: Dn/Ds



How much variation is out there, and how does it evolve over time ?

A useful way of looking at it: ALLELE FREQUENCIES

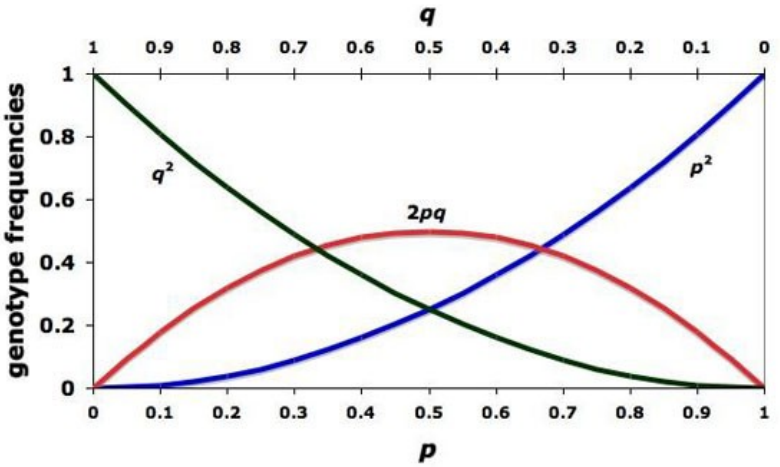


Do allele frequencies change over time? If so, how and why?

Ideal baseline: HARDY-WEINBERG EQUILIBRIUM MODEL

Hp) Infinite population
Random mating
No selection
No migration

Ts) Allele frequencies are constant



Nucleotide diversity: a measure of the degree of polymorphism within a population

```
subj0      ATAATAAAAA AATAATAAAA AAATAAAAAA AATAAAAAAA A
subj1      AAAAAAAATA AATAATAAAA AAATAAAAAA AAAAAAAAAA A
subj2      AAAATAAAAA TATAATAAAA AAATATAAAA AAAAAAAAAA A
subj3      AAAAAAAAAA AATAATAAAA AAATAAATAA ATAAAAAAAA A
subj4      AAAATAAAAA AAATATAAAA AAATAAAAAA AAAAAAAAAA A
subj5      AAAATAAAAA AAAAATAAAA AAAAAAAAAA AAAAATAAAA A
subj6      AAAAAATAAA AATAATAAAA AAATAAAAAA AAAAAAAAAA A
subj7      AAAAAAAAAA AAAAATAAAA AAATAAAAAA AAAAAAAAAT A
subj8      AAAAAAAAAA AAAAAAAAAA AAATAAAAAA AAAAAAAAAA A
subj9      AAAAAAAAAA AAAAATAAAA AAATAATAAA AAAAAAAAAA A
```

Average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population, and is denoted by π.

Tajima's D

$E[\pi] = \theta$
$E[S] = a1\ \theta$

$\theta = 4N\ u$ (2N= haploid population size, u= mutation rate/generation), and a1 is defined below.

$$D = \frac{\pi - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}}$$

A negative Tajima's D signifies an excess of low frequency polymorphisms relative to expectation, indicating population size expansion (e.g., after a bottleneck or a selective sweep) and/or purifying selection.

A positive Tajima's D signifies low levels of both low and high frequency polymorphisms, indicating a decrease in population size and/or balancing selection.

Pop size matters

$$N_e s < 1$$

Neutral mutations

At selectively neutral sites, the rate of substitution is equal to the rate of mutation

$$K = Nu \times 1/N = u$$

u = rate of mutation per generation
N = Pop size
K = rate of substitution per generation

Statistical measures of rate of Syn NonSyn substitutions

Approximation:

$1^{st} + 2^{nd}$ vs. 3^{rd} codon positions

Count number of Syn and NonSyn sites!

Exact calculation based on the genetic code

D_s = # Syn Subst / # Syn sites

D_n = # NonSyn Subst / # NonSyn sites

What do we expect?

Synonymous substitution to be more frequent than Nonsynonymous substitutions

Rate of Synonymous substitutions to be more similar between genes than the rate of Nonsynonymous substitutions

Dn/Ds measure of positive selection

Sliding window approach to detect local signal of positive selection

Neutral Evolution $D_N / D_s = 1$

Positive Darwinian selection $D_N / D_s > 1$

Negative purifying selection $D_N / D_s < 1$

Mc Donald Kreitmann test

Main idea: genetic variability between populations should be correlated with that within populations

	Fixed	Polymorphic
Synonymous	Ds	Ps
Nonsynonymous	Dn	Pn

Further levels....

Expression: allele specificity, tissue specificity

Epigenomics

new kinds of data: discrete (methylation states) and continuous (chromatin structure)

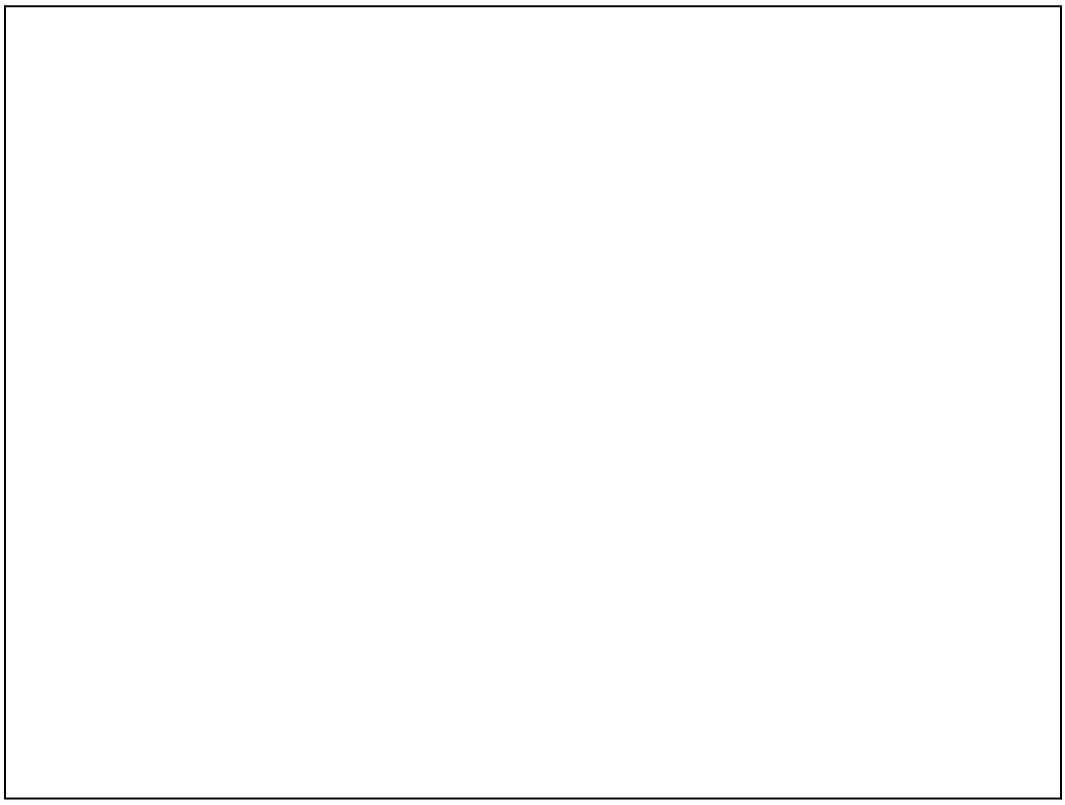
different relevant time scales

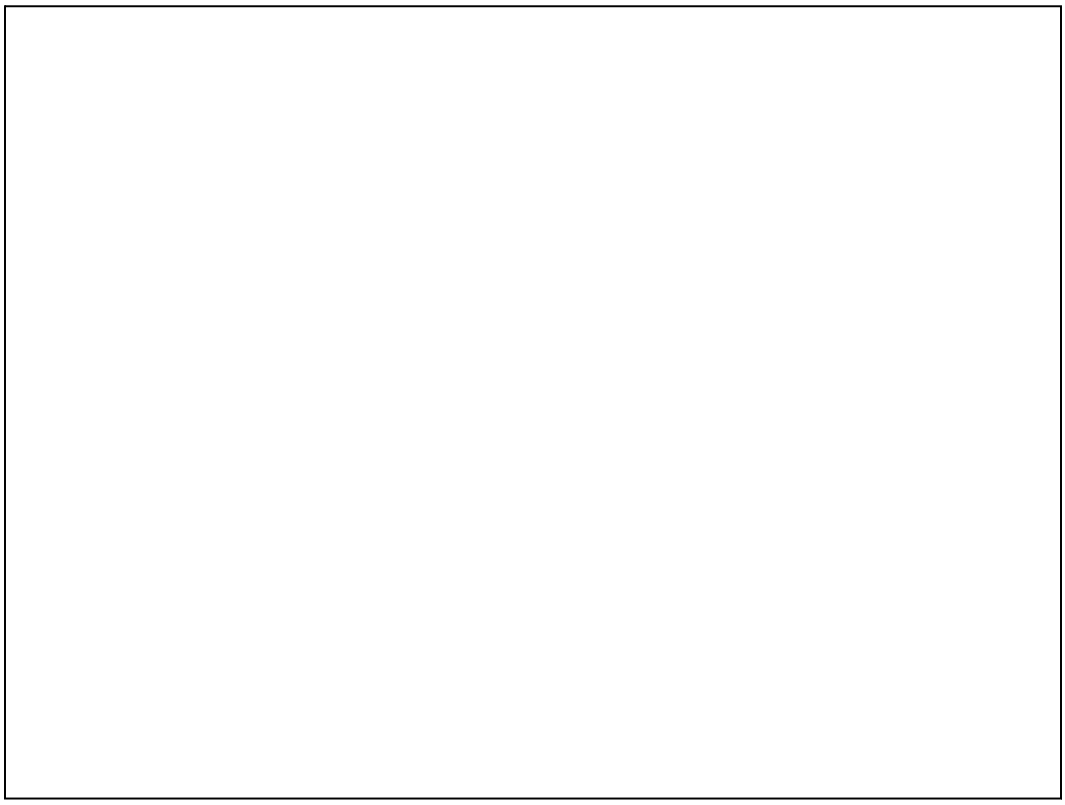
intra(!)-individual variability

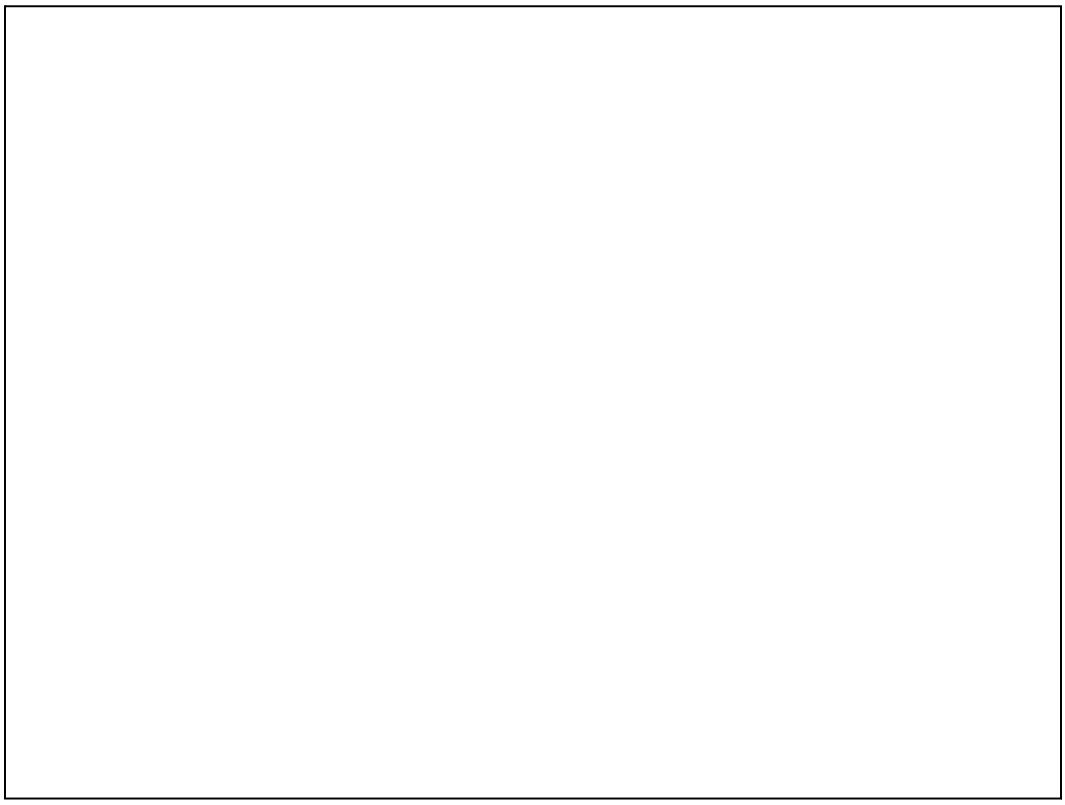
modes of heritability

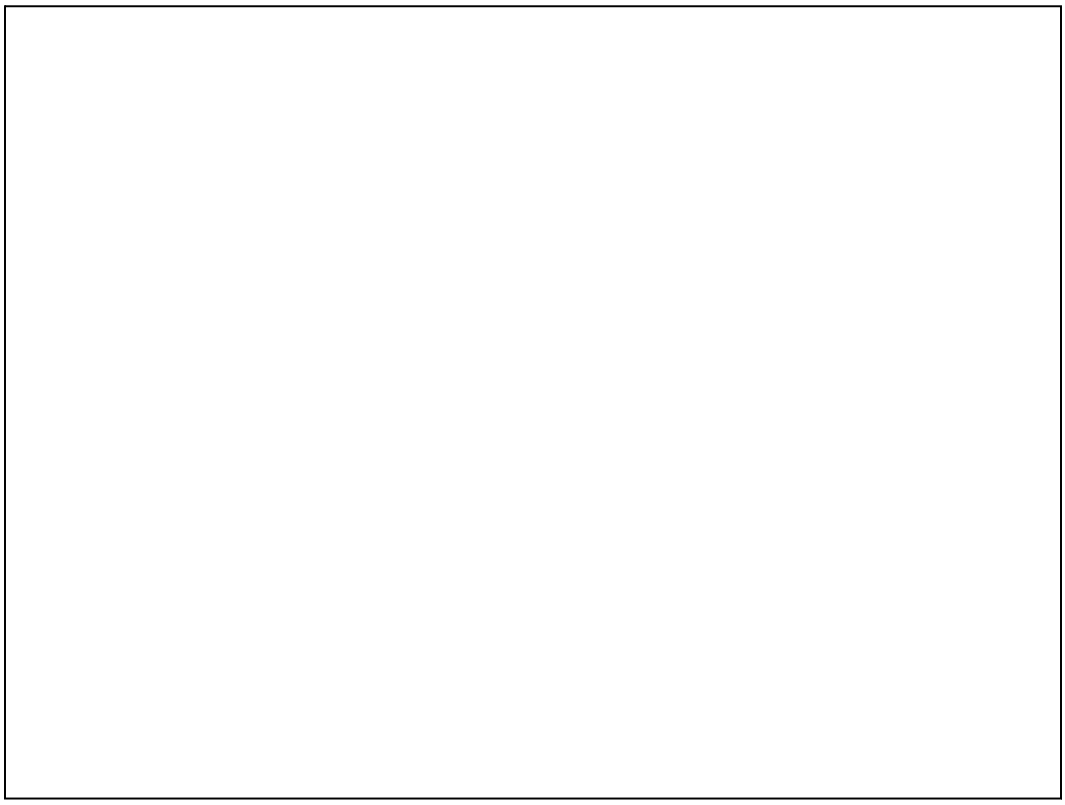
imprinting

non-Mendelian inheritance - meiotic drive



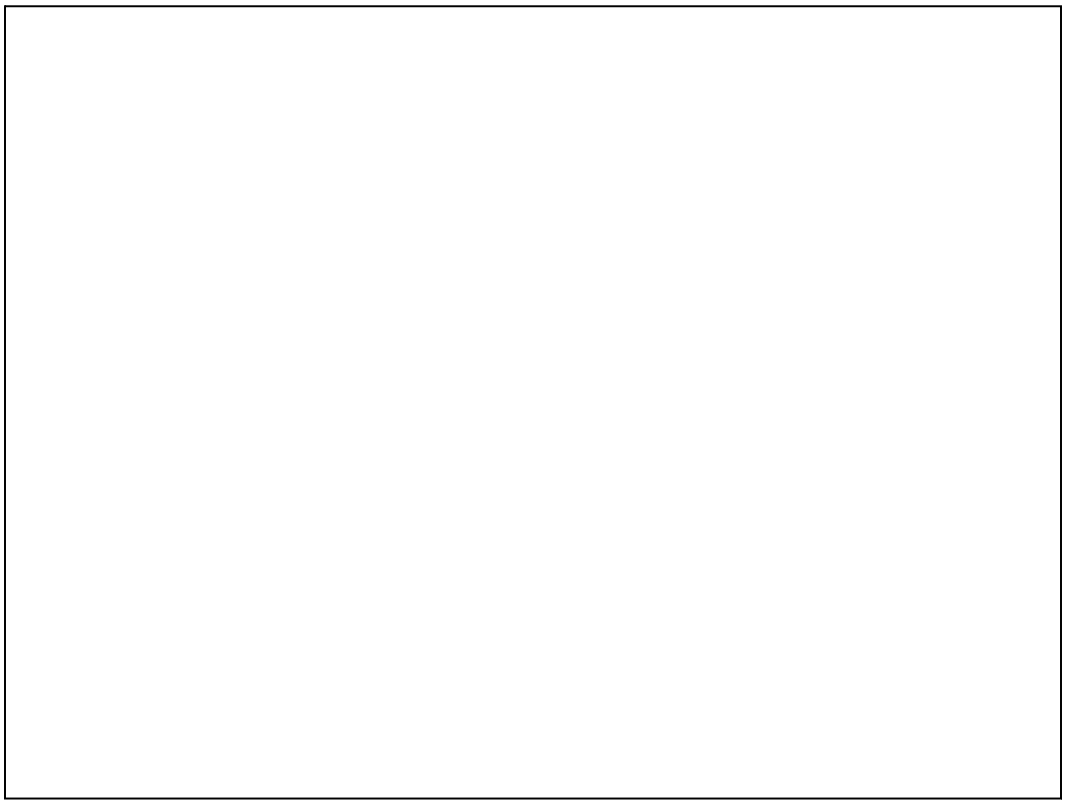






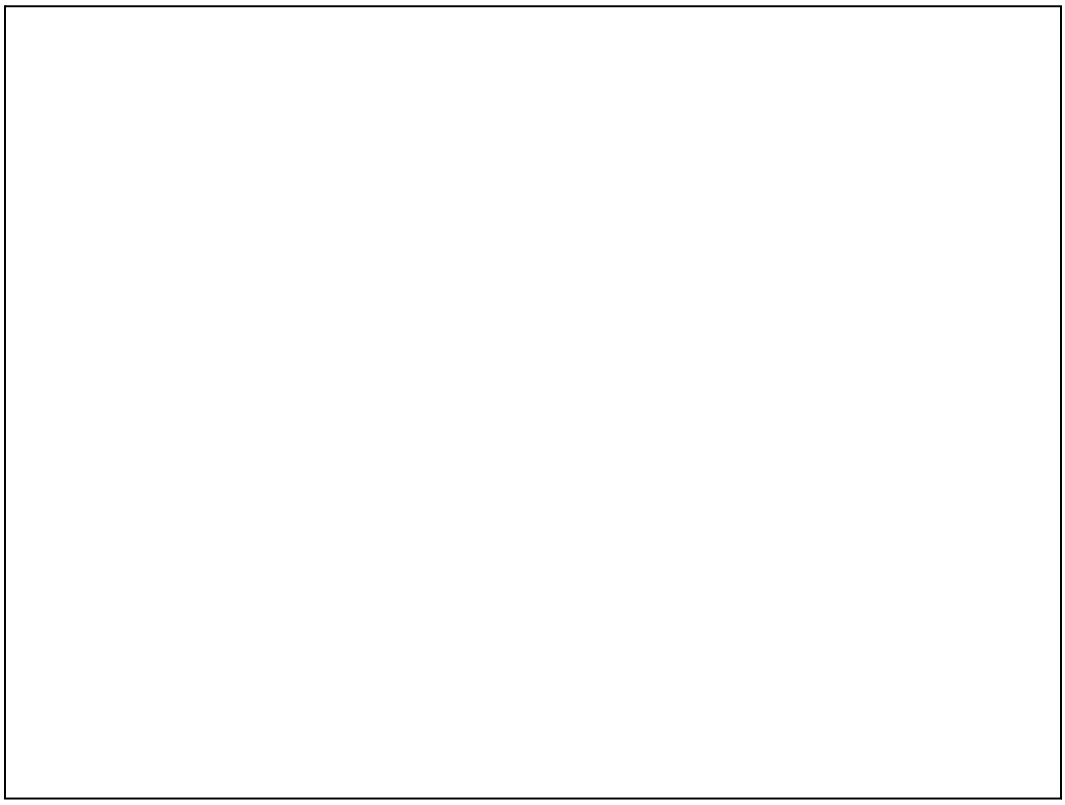
The Hardy-Weinberg theorem characterizes the distributions of genotype frequencies in populations that are not evolving, and is thus the fundamental null model for population genetics.

Evolution can be defined as the change in allele frequencies (i.e., genotypes) over time. In the early 1900s, G. Hardy and W. Weinberg developed a mathematical model that predicts genotype frequencies when a population is not being affected by evolutionary forces. The model is known as the Hardy Weinberg Equilibrium Model. The Hardy-Weinberg equilibrium is a baseline against which the evolution of populations can be measured and is the foundation for the genetic theory of evolution. When a population is in Hardy-Weinberg equilibrium, there is no change in gene frequencies and, therefor, no evolution of the population. The conditions required for Hardy-Weinberg equilibrium are almost never met in nature, however. As a result, gene frequencies change and evolution occurs. If a population is not at Hardy-Weinberg equilibrium, then we know that one or more of the conditions required for the equilibrium are not being met.



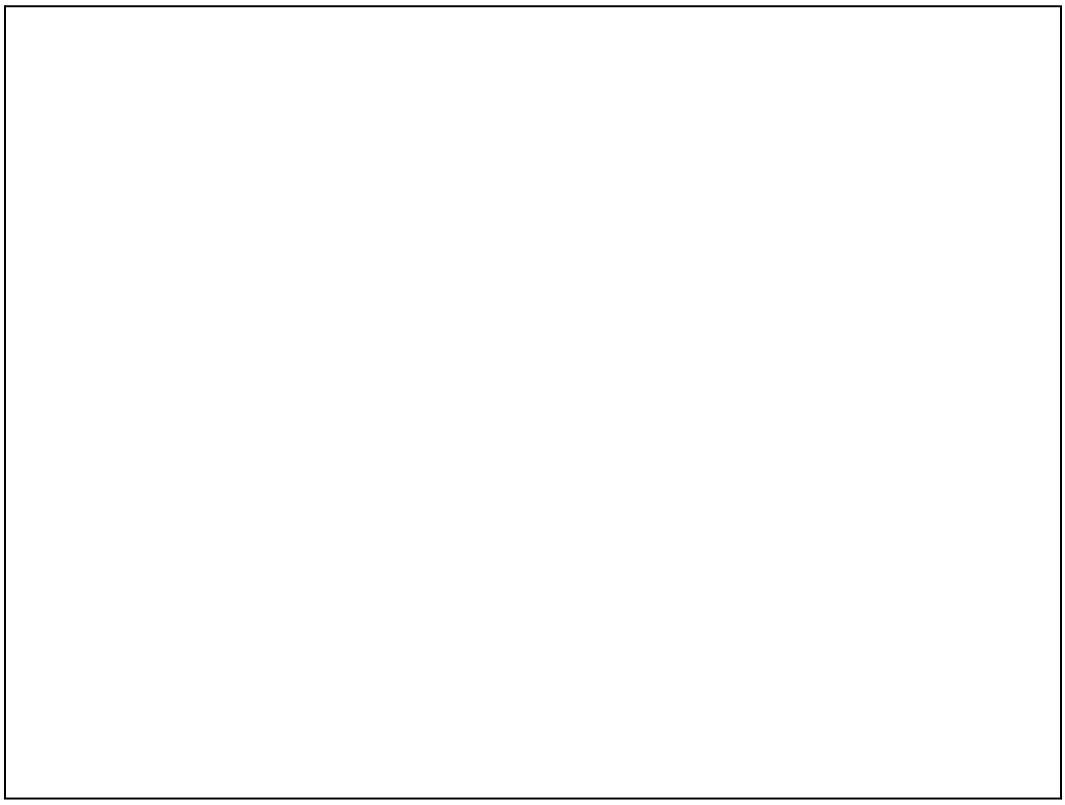
The Hardy-Weinberg theorem characterizes the distributions of genotype frequencies in populations that are not evolving, and is thus the fundamental null model for population genetics.

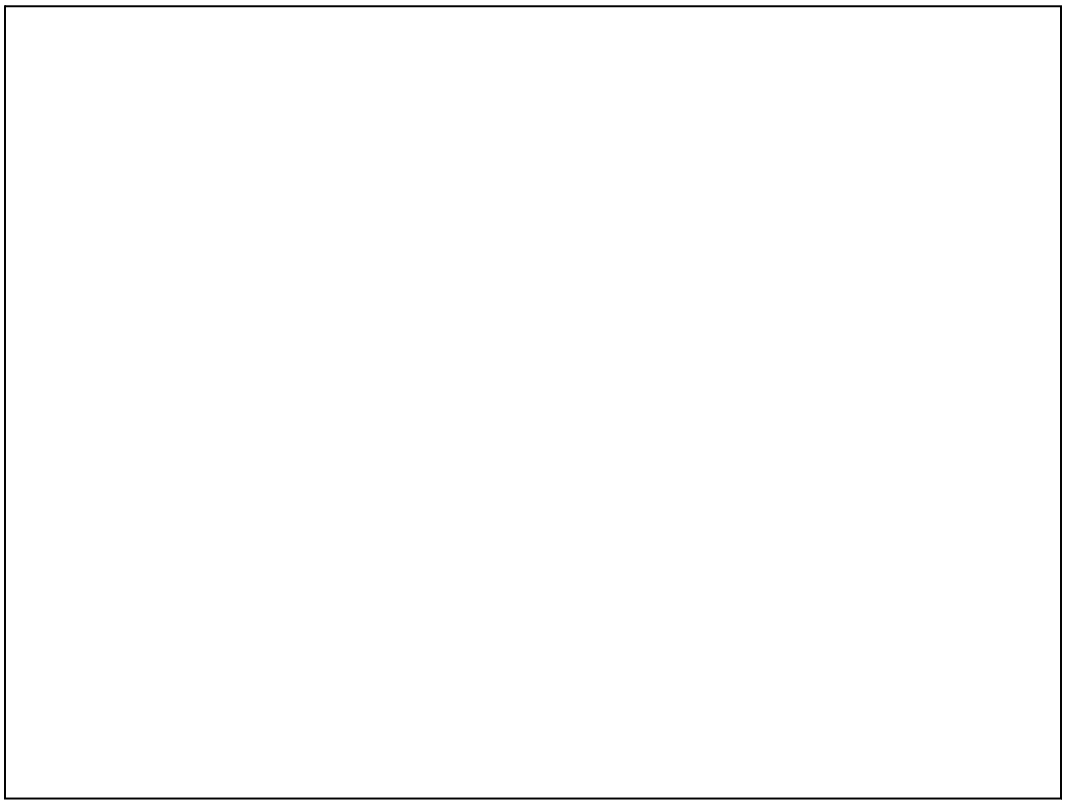
Evolution can be defined as the change in allele frequencies (i.e., genotypes) over time. In the early 1900s, G. Hardy and W. Weinberg developed a mathematical model that predicts genotype frequencies when a population is not being affected by evolutionary forces. The model is known as the Hardy Weinberg Equilibrium Model. The Hardy-Weinberg equilibrium is a baseline against which the evolution of populations can be measured and is the foundation for the genetic theory of evolution. When a population is in Hardy-Weinberg equilibrium, there is no change in gene frequencies and, therefor, no evolution of the population. The conditions required for Hardy-Weinberg equilibrium are almost never met in nature, however. As a result, gene frequencies change and evolution occurs. If a population is not at Hardy-Weinberg equilibrium, then we know that one or more of the conditions required for the equilibrium are not being met.

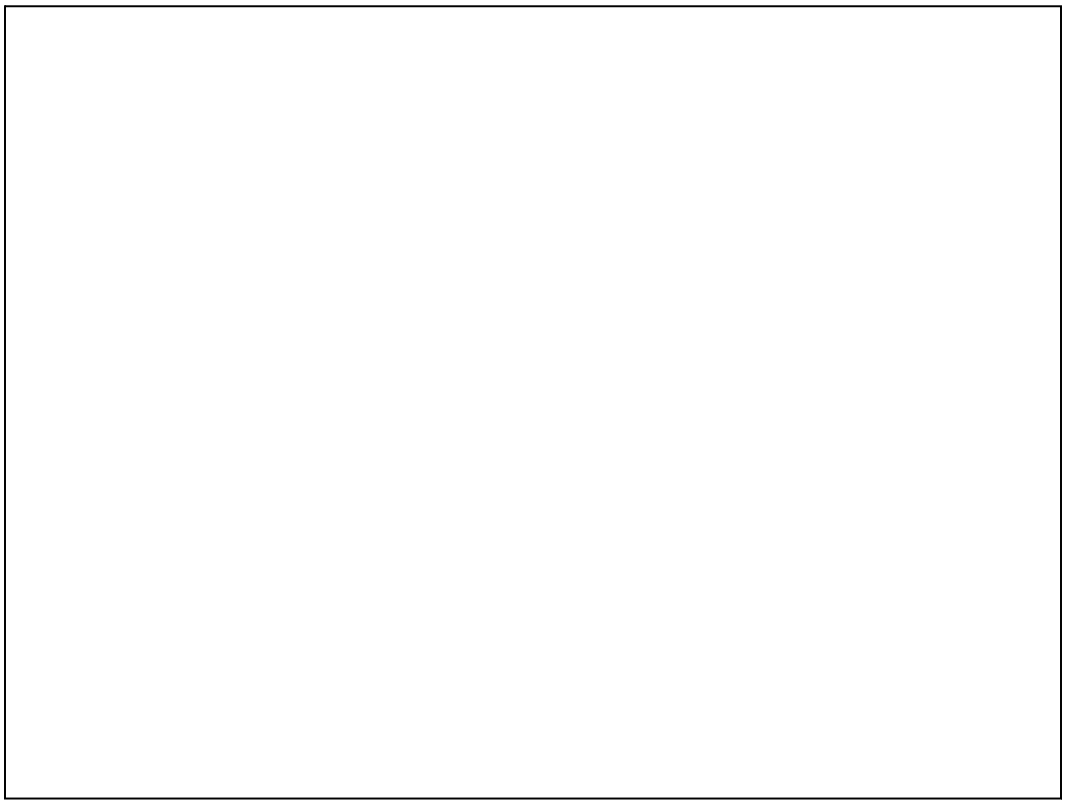


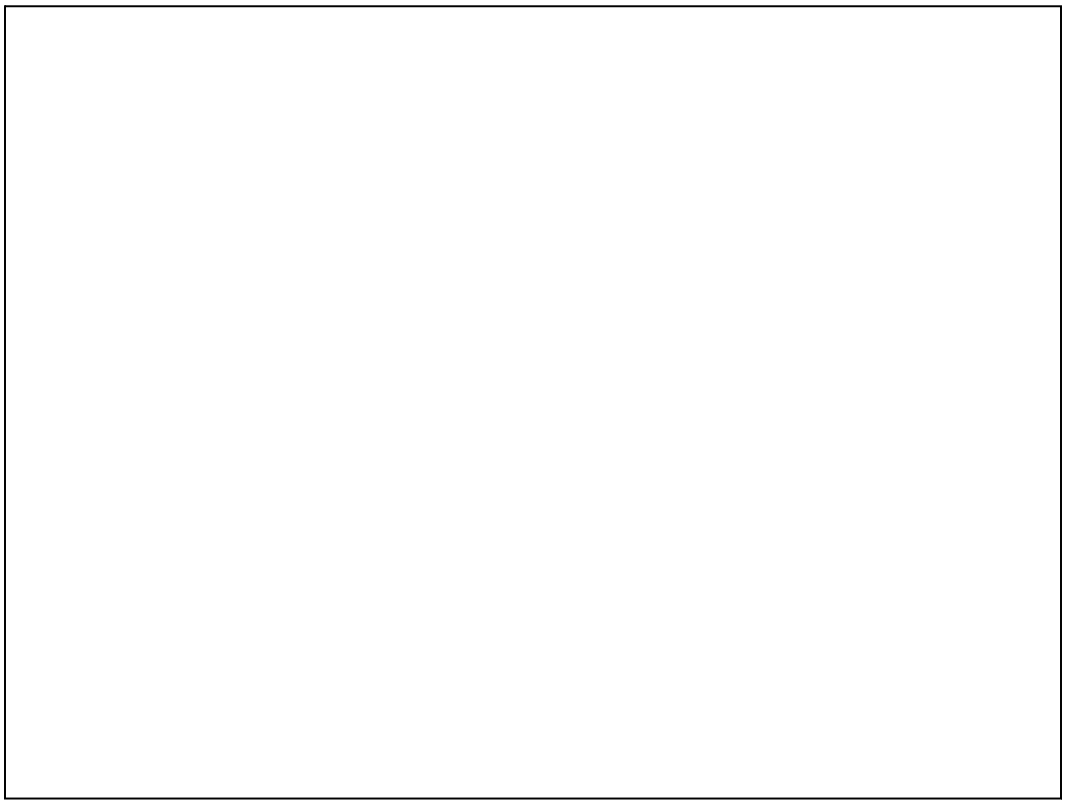
The Hardy-Weinberg theorem characterizes the distributions of genotype frequencies in populations that are not evolving, and is thus the fundamental null model for population genetics.

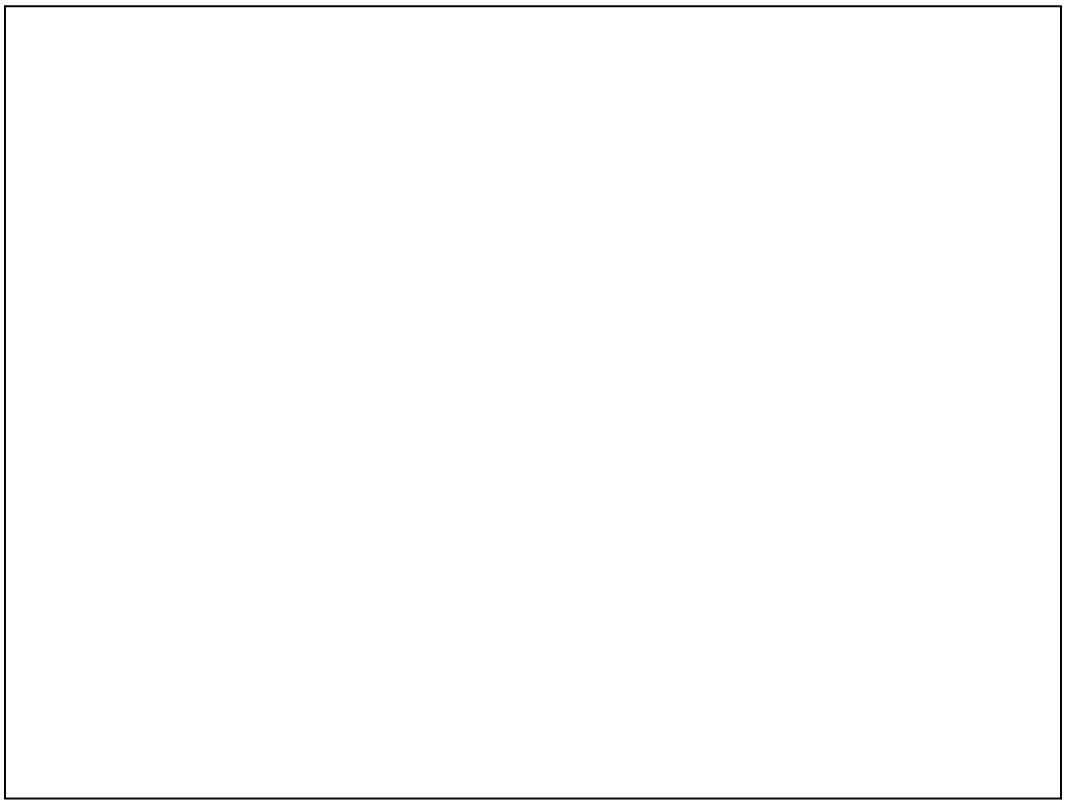
Evolution can be defined as the change in allele frequencies (i.e., genotypes) over time. In the early 1900s, G. Hardy and W. Weinberg developed a mathematical model that predicts genotype frequencies when a population is not being affected by evolutionary forces. The model is known as the Hardy Weinberg Equilibrium Model. The Hardy-Weinberg equilibrium is a baseline against which the evolution of populations can be measured and is the foundation for the genetic theory of evolution. When a population is in Hardy-Weinberg equilibrium, there is no change in gene frequencies and, therefor, no evolution of the population. The conditions required for Hardy-Weinberg equilibrium are almost never met in nature, however. As a result, gene frequencies change and evolution occurs. If a population is not at Hardy-Weinberg equilibrium, then we know that one or more of the conditions required for the equilibrium are not being met.











Neutral Evolution $D_N / D_s = 1$

Positive Darwinian selection $D_N / D_s > 1$

Negative purifying selection $D_N / D_s < 1$

Mc Donald Kreitmann test

Main idea: genetic variability between populations should be correlated with that within populations

	Fixed	Polymorphic
Synonymous	Ds	Ps
Nonsynonymous	Dn	Pn

Further levels....

Expression: allele specificity, tissue specificity

Epigenomics

new kinds of data: discrete (methylation states) and continuous (chromatin structure)

different relevant time scales

intra(!)-individual variability

modes of heritability

imprinting

non-Mendelian inheritance - meiotic drive