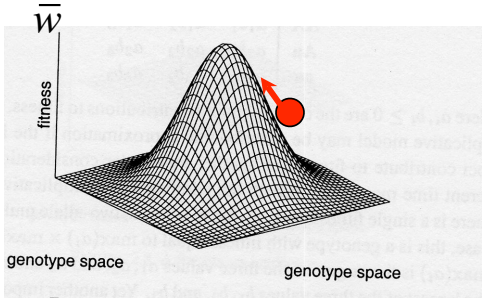
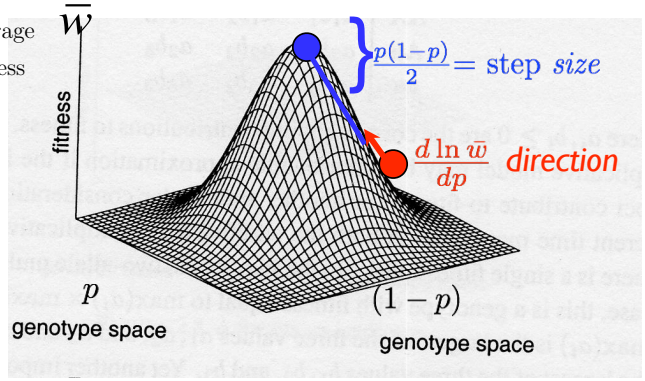
6.877: Computational Evolutionary Biology
Lecture 2: Climbing Mt. Improbable

Image removed due to copyright restrictions.



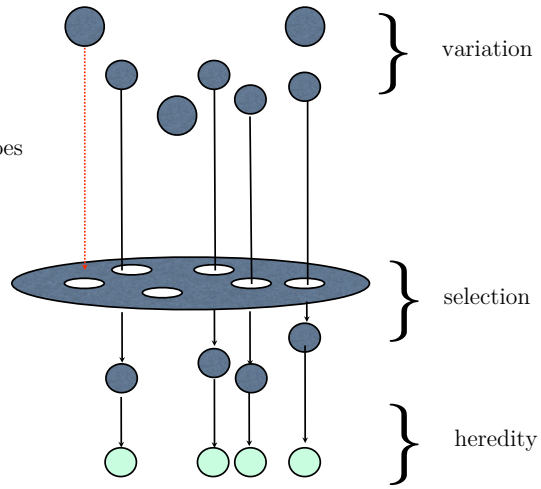
Goal: understand this model, the "F=ma"

Average fitness

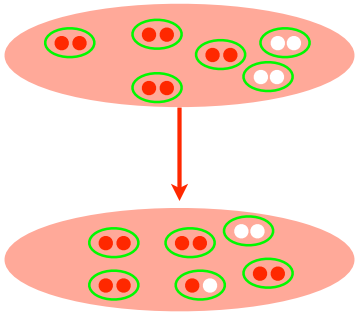


A selectional model of evolution

Q: What role does variation play?



Mutation
Migration
Population size
(drift)



Variation in the key
Keep track of it!

Fisher's proof of mud slides

x = first parent's deviation from mean value
y = second parent's deviation from mean value

variance = $E(x^2)$

What is the variance of $\frac{1}{2}(x+y)$?

$\text{var}\frac{1}{2}(x+y) = E[\{\frac{1}{2}(x+y)\}^2] =$
$E[\frac{1}{4}(x^2+2xy+y^2)] =$
$E[\frac{1}{4}(x^2+y^2)] = E[\frac{1}{4}(x^2+x^2)] = E[\frac{1}{4}(2x^2)] =$
$\frac{1}{2}E[x^2]$

The forces of evolution: a dynamical system model for computing a new state from the current state

- Statics: what's the model if we are *at equilibrium* – there are no forces acting? (And: what assumptions are required to maintain equilibrium?)
- Dynamics: what's the F=ma analog so we can compute p' from p?

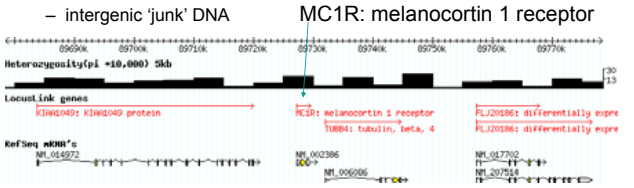
Mendelian genetics terminology review for "Evolutionary first law" (Hardy-Weinberg equillibrium)

- Gene or locus:
- Classical genetic: Chromosomal region to which a phenotypic mutation can be mapped

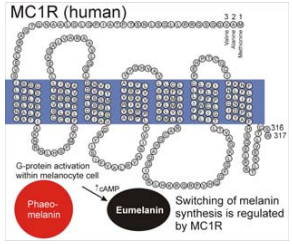
  Molecular: Open reading frame and associated regulatory elements

  Evolutionary: A stretch of hereditary material sufficiently small such that it is not broken up by recombination, and which can be acted on by natural selection
- Allele: One of two or more possible forms of a gene (locus)
- Genotype: The total complement of alleles present in an organism
- Allozyme: distinct protein form, corresponding to an allele
- Polymorphism: (Ford, 1940) working definition – a less common allele with a frequency $> 1\%$ (e.g., a mutation that has become common) *within* a species
  - Example: red hair color MC1R loss-of-function allele (the *only* pigmentation gene so far identified in human that explains substantial phenotypic variance

An example of a human gene variant with phenotypic effect

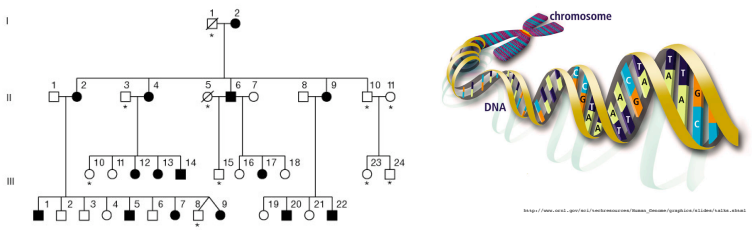


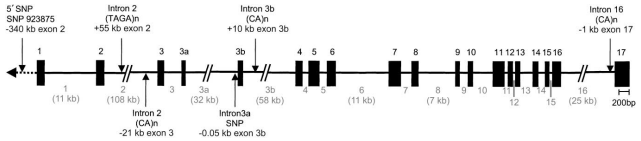
100 kb from chromosome 16 around the MC1R gene



AUG Met (M)
AUA Ile (I)
AUC
AUU
GCG Ala (A)
GCA
GCC
GCU
GUG Val (V)
GUA
GUC
GCU

Variation at all levels






```
          294
Human     300 TSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHT
Chimp     301 TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHT
Mouse     299 TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHT
```

How do we differ? – Let me count the ways

- Single nucleotide polymorphisms
  - 1 every few hundred bp, mutation rate* $\approx 10^{-9}$

```
TGCATTGCGTAGGC
TGCATTCCGTAGGC
```

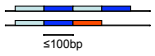
- Short indels (=insertion/deletion)
  - 1 every few kb, mutation rate v. variable

```
TGCATT---TAGGC
TGCATTCCGTAGGC
```

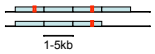
- Microsatellite (STR) repeat number
  - 1 every few kb, mutation rate $\leq 10^{-3}$

```
TGCTCATCATCATCAGC
TGCTCATCA------GC
```

- Minisatellites
  - 1 every few kb, mutation rate $\leq 10^{-1}$



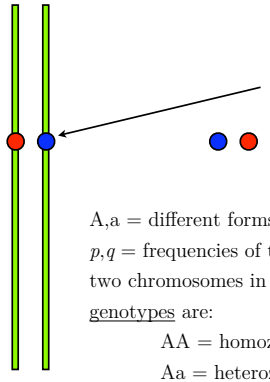
- Repeated genes
  - rRNA, histones



- Large inversions, deletions
  - Rare, e.g. Y chromosome

*per generation

What's a gene?



A,a = different forms of the same gene, or "alleles"

p, q = frequencies of these alleles

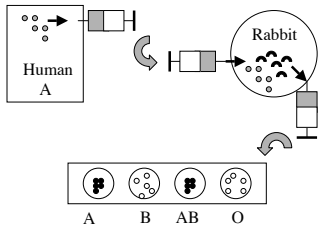
two chromosomes in each eukaryotic cell - diploid - so possible genotypes are:

AA = homozygote A

Aa = heterozygote A,a

aa = homozygote a

Serological techniques for detecting variation

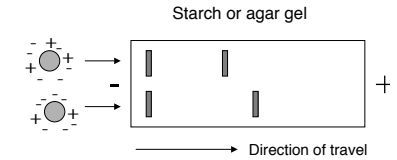


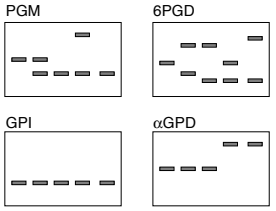
Polymorphic blood groups in the white English population (no. types)

ABO	(4)	Kidd	(3)
Rh	(7)	Dombrock	(2)
MNS	(6)	Auberger	(2)
P	(3)	Xg	(2)
Secretor	(2)	Sd	(2)
Duffy	(3)	Lewis	(2)

Pr{2 people same blood type} $\approx$ 3 in 10,000

Protein electrophoresis

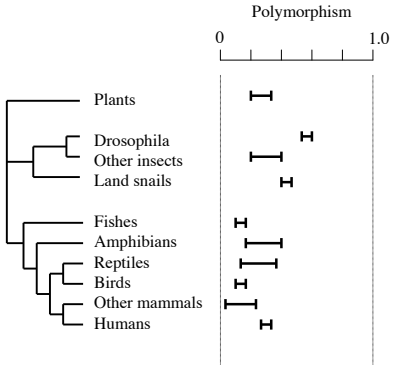




Polymorphism = 0.75

Heterozygosity = 0.30

The phylogenetic distribution of allozyme variation



Humans Polymorphism = 0.31

Heterozygosity = 0.06

Two haploid genomes are expected to differ at c. 6,000 loci

Patterns of variation at the DNA level

- Synonymous & nonsynonymous mutations

```
Arg Gln Val      Arg Gln Val
AGA CAA GTA      AGA CAA GTA
    ↓                ↓
CAG CGA GTA      AGA CAG GTA
Arg Arg Val      Arg Gln Val
```

D. simulans

π_{total} = 0.010 per site

π_{silent} = 0.038

$\pi_{noncoding}$ = 0.023

- Nucleotide variation v. protein variation?

	Humans	*D. melanogaster*
Allozyme	6%	14%
Nucleotide	0.1%	1%

Alleles & genotypes: Genetic composition of a population…has 3 components

1. The number of alleles at a locus
2. The frequency of alleles at the locus
3. The frequency of genotypes at the locus (not the same as 2!)

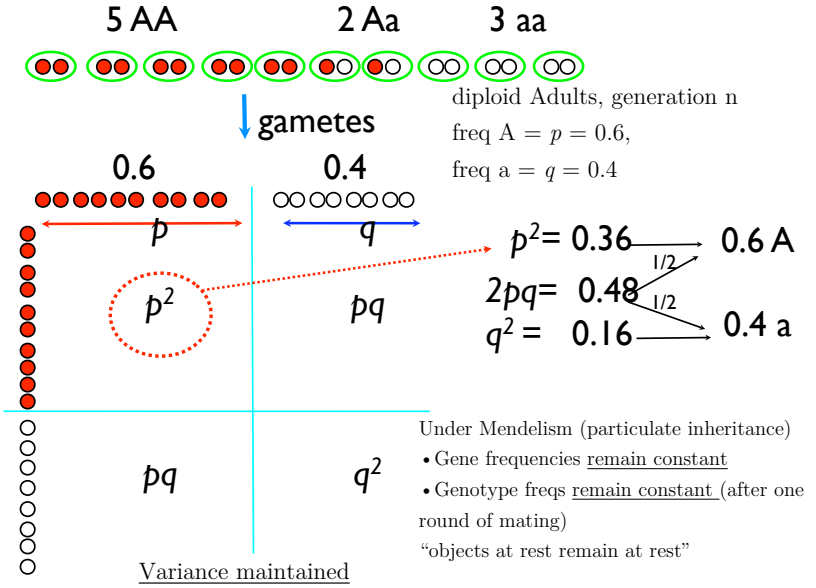
	AA	Aa	aa
Population 1	50	0	50
Population 2	25	50	25

freq(A)=0.5 in both;

but when can we compute genotype freqs from allele freqs?

The first law: Hardy-Weinberg equillibrium - 8 assumptions!

1. Genotype frequencies are the same in both males and females
2. Genotypes mate at random *with respect to their genotyhpe at this particular locus*
3. Meiosis is fair
4. No input of new genetic material (no mutation, migration)
5. Population is of arbitrarily large size s.t. actual frequency of matings is equal to their expected frequency, and the actual frequency of offspring from each mating is equal to the Mendelian expectations
6. All matings produce the same # of offspring, on average
7. Generations do not overlap
8. There are no differences among genotypes in pr of survival (no selection)



diploid Adults, generation n
freq A = $p = 0.6$,
freq a = $q = 0.4$

Under Mendelism (particulate inheritance)
- Gene frequencies remain constant
- Genotype freqs remain constant (after one round of mating)

"objects at rest remain at rest"

H-W

freq(AA in zygotes) = p^2
freq(Aa in zygotes) = $2pq$
freq(aa in zygotes) = q^2

1. If assumptions #1-#8 are true, then equations must be true
2. If genotypes are in H-W proportions, then one or more of assumptions #1-#8 may still be violated
3. If genotypes are *not* in H-W proportions, one or more of Assumptions #1-#8 must be false

An example: testing whether a population is in H-W equillibrium

Data: 1000 individuals
90 are AA
420 are Aa
490 are aa
Q: is this population in H-W equillibrium?

Step 1: calculate allele frequencies.
freq A allele = total # A alleles/total # alleles = (90*2+420)/2000 = 0.3
freq a allele = 1-0.3 = 0.7, i.e., (490*2+420)/2000

Step 2: calculate genotype frequencies.
p = freq AA = 90/1000 = 0.09; freq Aa = 420/1000 = 0.42; q = freq aa = 490/1000=0.49

Step 3: calculate expected H-W genotype proportions, in ratio p^2: $2pq$: q^2
$p^2 = 0.3^2 = 0.09$
$2pq$ = 2 x 0.3 x 0.7 = 0.42
$q^2 = 0.7^2 = 0.49$

The genetics of natural selection: the simplest case

- Which H-W assumptions involve selection?

Assumption # 3: *Meiosis is fair.*
But: suppose the alleles are not
equally frequent in gametes produced. Example: *t*-allele in mouse, 95% in heterozygotes. Or: gamete competition (sperm, pollen)

Assumption #6: *All matings produce the same # of offspring.* But: suppose # offspring depends on maternal genotype or parental genotype, or both – *fertility selection*

Assumption #8: *Survival does not depend on genotype.*
But: suppose prob of survival from zygote to adult depends on genotype – *viability selection*

The algebra of viability selection - J.B.S. Haldane, 1924

1 gene in 2 different forms (alleles)

genotype	AA	Aa	aa
frequency	p^2	$2pq$	q^2
relative fitness	w_{11}	w_{12}	w_{22}
after selection	$w_{11}\, p^2$	$w_{12}\, 2pq$	$w_{22}\, q^2$

survivors

Intuitively, w is a 'growth rate' – the expectation that an individual with a particular genotype will survive and reproduce – factor altering H-W proportions

Note that if N_t = # before selection, the total # after selection is:

$N_{t+1} = \bar{w} N_t$ where

$\bar{w} = w_{11}p^2 + w_{12}2pq + w_{22}q^2$

What is the average (marginal) fitness of A's?

$w_1^* = P$(paired with another A)w_{11} + P(paired with an a)w_{12}=
$w_1^* = pw_{11} + qw_{12}$ or if just 2 alleles:
$w_1^* = pw_{11} + (1-p)w_{12}$

genotype	AA	Aa	aa
frequency	p^2	$2pq$	q^2
relative fitness	w_{11}	w_{12}	w_{22}
after selection	$w_{11}\, p^2$	$w_{12}\, 2pq$	$w_{22}\, q^2$

w_1^* This is the *expectation* that A will survive

Two allele case: we can now calculate $p - p'$ *i.e.*, the change in allele frequency, or *evolution*

In this generation, freq $A = p_t$ = # A's/total # alleles
In next generation, freq $A = p_{t+1}$ = expected # A survivors/total expected # survivors
Expected # A's = $w_1^* n_A$

Expected # all alleles = $\bar{w} n_{total}$

$$p_{t+1} = \frac{w_1^* n_A}{\bar{w} n_{total}} = \frac{p_t w_1^*}{\bar{w}}$$

$$p_{t+1} - p_t = \frac{p_t w_1^*}{\bar{w}} - \frac{p_t \bar{w}}{\bar{w}}$$

$$\triangle p = \frac{p_t (w_1^* - \bar{w})}{\bar{w}}$$

Think about what this means: what if w_l is *greater* than average fitness? *Less*?

"The company you keep"

Understanding the basic $\triangle p$ formula

Fitness of organisms in which A finds itself

$$\triangle p = p\frac{\bar{w}_1 - \bar{w}}{\bar{w}}$$

Divided by fitness of all organisms

To derive the rest of the 'jet fuel' formula

$$\triangle p = \frac{p_t(w_1^* - \bar{w})}{\bar{w}}$$

Substitute: $\bar{w} = pw_1^* + (1-p)w_2^*$

$$\triangle p = \frac{p_t(w_1^* - pw_1^* - (1-p)w_2^*)}{\bar{w}} \text{ or}$$

$$\triangle p = \frac{p(1-p)(w_1^* - w_2^*)}{\bar{w}}$$

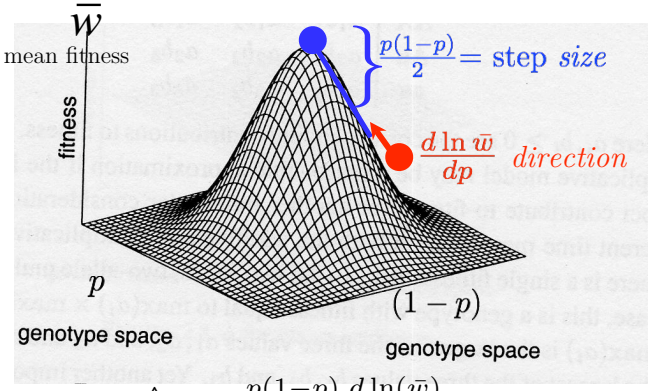
Now note that derivative of $\bar{w}$ wrt p (assuming what?) can now be calculated from:
$\bar{w} = w_{11}p^2 + p(1-p)w_{12} + (1-p^2)w_{22}$ as:

$$\begin{aligned}\frac{d(\bar{w})}{dp} &= 2pw_{11} + 2w_{12} - 4pw_{12} - 2w_{22} + 2pw_{22} \\ &= 2[pw_{11} + (1-p)w_{12}] - 2[pw_{12} + (1-p)w_{22}] \\ &= 2(w_1^* - w_2^*)\end{aligned}$$

$$\triangle p = \frac{p(1-p)}{2}\frac{d\ln(\bar{w})}{dp}$$

Sewall Wright's adaptive landscape:

Understanding the formula



$$\triangle p = \frac{p(1-p)}{2}\frac{d\ln(\bar{w})}{dp}$$

Some dissection...

$$\triangle p = \frac{p(1-p)}{2}\frac{d(\bar{w})}{\bar{w}dp}$$

Variance component of allele A within genotype

Why variance? Draw from pool of A, a gametes many many times: binomial sampling – frequency of A within a genotype is either 1, 1/2, or 0; variance is p(1-p)/2 ("heterozygosity")

Slope of fitness function divided by mean population fitness – a *potential function*?