

Genetic variation: the substrate of evolution

We now know that heritable phenotypic variation is caused by genetic variation.

But *how much* genetic variation is there in a typical species?

And what is its *origin*?

A complex morphological polymorphism in the medium ground finch (*Geospiza fortis*), controlled by alleles at many genetic loci, and also affected by gender and the environment.

From Egmond et al. (2001) *PNAS* 98, 6253-6255.

Biol 3410, 30 January 09

Genotype and Phenotype

Darwin labored under the then-traditional but incorrect theory of *blending inheritance*, which poses two major problems for his theory.

- It cannot explain the *maintenance* of heritable variation within populations.
- It is compatible with Lamarckian *inheritance of acquired characteristics*.

In the late 19th century, Weismann proposed the "germ-plasm theory" of inheritance, which sharply distinguishes between the *germ line* and the *soma*. On this theory only sex cells are passed from one generation to the next; the soma (body) then develops anew from the fertilized egg.

August Weismann (1834-1914)

Shortly after Mendel's work was rediscovered in the early 20th century, the modern terms "genotype" and "phenotype" were invented to distinguish the hereditary information transmitted through genes from its effect on the tangible physical characteristics of individuals.

DNA carries hereditary information in its base sequence

A double-helical complementary anti-parallel structure

Four nucleotide bases, paired Adenine-Thymine (A-T) and Guanine-Cytosine (G-C)

Sugar-phosphate backbone

Hydrogen bond

Strands can separate and serve as templates for synthesis of complementary daughter strands (passing identical sequences to daughter cells).

The Central Dogma is Weismann's theory writ very small!

Information flow Example

Germ plasm DNA → mRNA → Protein

Next generation DNA → mRNA → Protein

Soma Protein

Protein-coding DNA sequences are conventionally written as the "sense" strand (i.e., as if they were the mRNA sequence).

Thus:

TTT	Phenylalanine
TTC	Phenylalanine
TTA	Leucine
TTG	Leucine

First base	U	C	A	G	Third base
U	UUU Phenylalanine UUC Phenylalanine UUA Leucine UUG Leucine	UCU Serine UCC Serine UCA Serine UCG Serine	UAU Tyrosine UAC Tyrosine UAA Stop UAG Stop	UUU Cysteine UUC Cysteine UGU Stop UGG Tryptophan	U C A G
C	CUU Leucine CUC Leucine CUA Leucine CUG Leucine	CCU Proline CCC Proline CCA Proline CCG Proline	CAU Histidine CAC Histidine CAA Glutamine CAG Glutamine	CGU Arginine CGC Arginine CGA Arginine CGG Arginine	U C A G
A	AUU Isoleucine AUC Isoleucine AUA Isoleucine AUG Start/Methionine	ACU Threonine ACC Threonine ACA Threonine ACG Threonine	AAU Asparagine AAC Asparagine AAA Lysine AAG Lysine	AUU Serine AUC Serine AGA Arginine AGG Arginine	U C A G
G	GUU Valine GUC Valine GUA Valine GUG Valine	GGU Alanine GGC Alanine GGA Alanine GGG Alanine	GAU Aspartic Acid GAC Aspartic Acid GAA Glutamic Acid GAG Glutamic Acid	GGU Glycine GGC Glycine GGA Glycine GGG Glycine	U C A G

Codon: Amino acid

Some nucleotide-substitution mutations cause amino-acid polymorphisms

Information flow Example

DNA → mRNA → Protein

DNA: TTC → mutation → TGC

DNA: TTC → TGC

mRNA: AAG → ACG

Protein: Lysine slow (S) → Threonine fast (F)

The alcohol dehydrogenase enzyme in *Drosophila melanogaster* is polymorphic for alleles that differ by a single amino acid (positively charged lysine, or neutral threonine). As a result of this difference in charge, the proteins migrate at different speeds in an electric field.

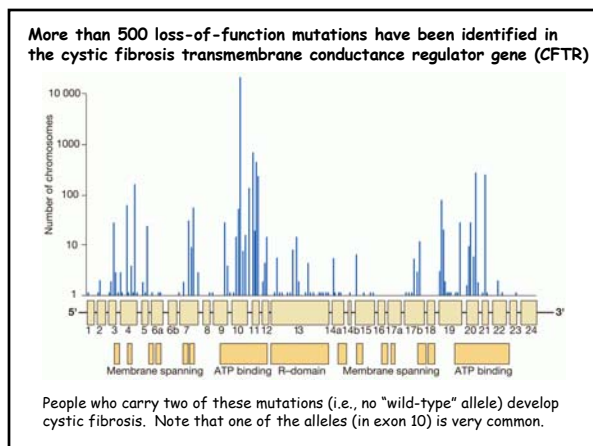
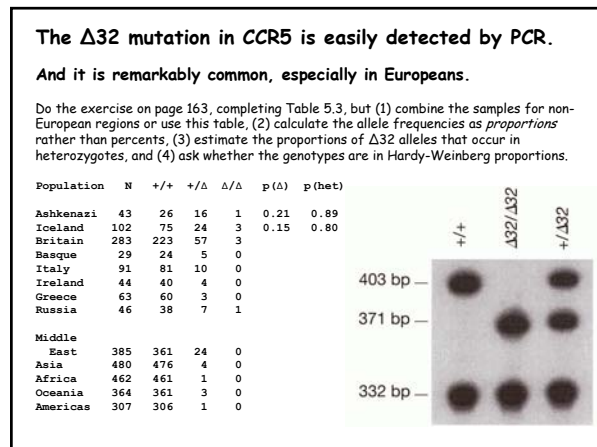
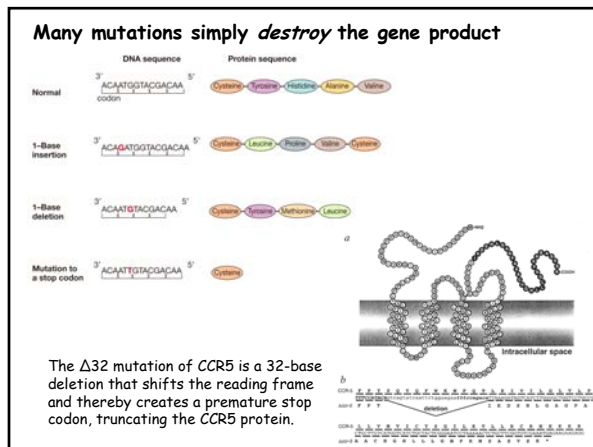
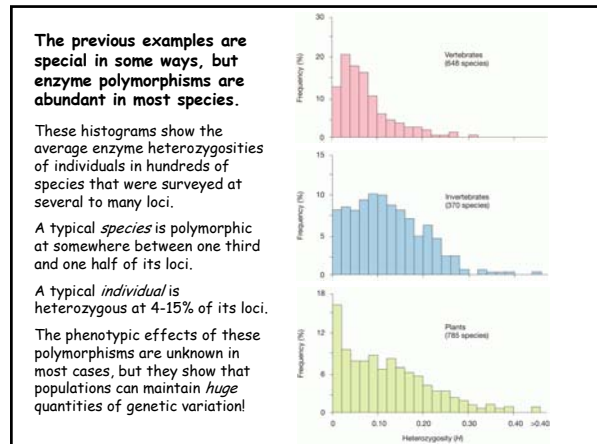
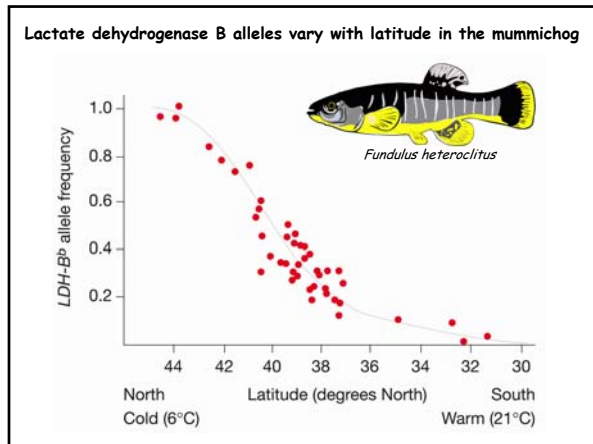
From FM Johnson & C Denniston (1964) *Nature* 204, 906-907.

Fig. 2. Photograph of starch gel showing fast (FF), heterozygous (FS), and slow (SS) alcohol dehydrogenase.

D. melanogaster occurs on all continents, and all populations appear to contain both *Fast* and *Slow* alleles.

On all continents, *Adh^S* becomes more common toward the equator, while *Adh^F* becomes more common toward the poles.

The *Fast* allele appears to be fitter than *Slow* in cooler climates, but researchers still do not fully understand why!



At what rates do mutations happen?

DNA replication is unbelievably accurate.

Base substitutions typically occur at rates of 10^{-7} to 10^{-9} per site per generation.

But this can be several to many per genome, in genomes containing billions of bases.

And other kinds of mutations may occur at much higher rates.

- deletions (one to many bases)
- insertions (one or a few bases during DNA replication)
- insertions of transposable elements such as LINES and SINES
- duplications and rearrangements of existing sequences


But much of the genome is functionless "junk" in large eukaryotes like us.

The real question is, how many *significant* mutations occur per generation?

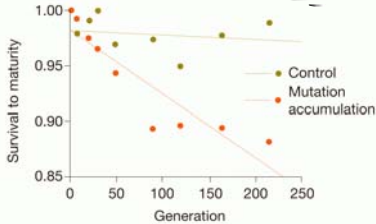
A mutation-accumulation experiment with *C. elegans*

Larissa Vassilieva and her colleagues set up more than 100 lines beginning with the offspring of a single homozygous individual. Then they propagated each line through a *single individual* each generation. This procedure effectively turns off natural selection and allows non-lethal mutations to accumulate.

Seventy-four lines survived beyond 200 generations. At intervals of 10-40 generations Vassilieva measured survival rates and other fitness-related characteristics of these lines.



Larissa Vassilieva



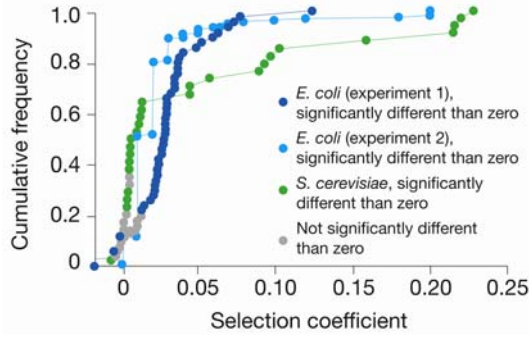
Survival to maturity

Control Mutation accumulation

Generation

Most genes make remarkably *modest* contributions to fitness

In these experiments, individual genes were "knocked out" and the resulting mutant line was then competed directly against the intact parental line.


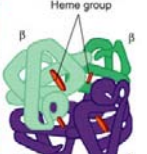


Cumulative frequency

Selection coefficient

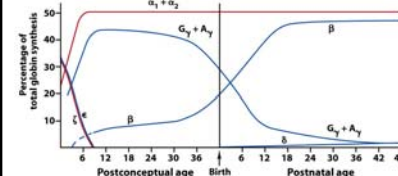
- $E. coli$ (experiment 1), significantly different than zero
- $E. coli$ (experiment 2), significantly different than zero
- $S. cerevisiae$, significantly different than zero
- Not significantly different than zero

Most genes belong to "families" of similar genes whose products perform similar functions (often on different substrates, or in different tissues, or at different times in development).

Heme group

The oxygen-carrying hemoglobins and myoglobin illustrate this principle, and show how gene families grow through duplication and divergence.



Percentage of total hemoglobin

Postconceptual age (weeks)

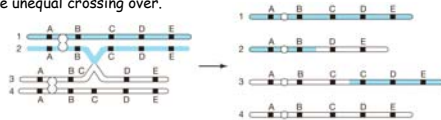
Birth

Postnatal age (weeks)

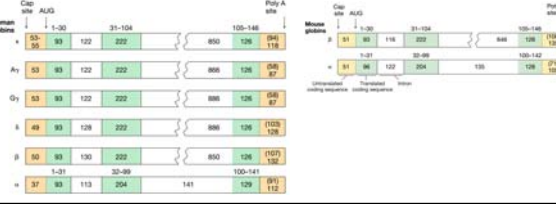
- α -like family includes three functional genes: $\alpha_1, \alpha_2, \zeta$ (zeta)
- β -like family includes five functional genes: β, ϵ (epsilon), δ (delta), γ (G-gamma), $A\gamma$ (A-gamma)

Where do new genes come from? Old genes, of course!

The most common mechanism of gene duplication is believed to be unequal crossing over.



A frequent result is gene families united by striking structural similarities.




Human genes

Mouse genes

And as you'd expect, their DNA and protein sequences are similar

This *amino-acid sequence alignment* shows the remarkable similarities of alpha-globins, beta-globins, and myoglobins in four distantly related vertebrate species. The pattern of pairwise similarities is *hierarchical*, as expected if these proteins evolved through a process of gene duplication and speciation.



Human α V L S P A D K T N V K A M G K V G A H G E Y A E A E I E M F L S F P T K T Y P H F - D L S H - - - - G S A Q V E G R G K R A D A D L T

Cow α V L S A A D K N K V K G I F T K I A G H E E Y A E T E M F I G F P T K T Y P H F - D L S H - - - - G S A Q I K G R G K R A D A L T

Chicken α V L S A A D K N K V K G I F T K I A G H E E Y A E T E M F I G F P T K T Y P H F - D L S H - - - - G S A Q I K G R G K R A D A L T

Shark α D - V S A A D R A E L A A L S K V L A G N A F A S A E A A M E T Y Y A A S E F K D Y K D F T A - - - - A A P S I E A H G A R V T A L A

Human β V H L T F E K S A V A L W G K V - - N V D E V G G A L G R L L V V P W O R F E S F G D L S T P D A V M N P K V K A H G K K L G A F S

Cow β M L I T A E E A A V T A F W K V - - N V D E V G G A L G R L L V V P W O R F E S F G D L S T A D A M M N P K V K A H G K K L G S D S

Chicken β V H W T A E K Q L I T G L W G K V - - N V A E G A T A A R L L V V P W O R F A S F G N L S S P T A I L G N P M M R A H G K K L T S F G

Shark β V H M S E V E L N E I E T T W K S I - - - D K S L S K A L A M F I G Y P W R T R Y F G R L K E F T A - - - - C S Y G R E A H K E S T G A L G

Human Mb G L S D G E W L V L N W G K V E A D I P G H Q E V E I R E F K G P E S L E K R D K F H L K S S D E M K A S E D K K H G A T V L T A L G

Cow Mb G L S D G E W L V L N W G K V E A D I P G H Q E V E I R E F K G P E S L E K R D K F H L K S S D E M K A S E D K K H G A T V L T A L G

Chicken Mb G L S D D E W O O L T I W G K V E A D I A G H E V E M L R H D H P E L D R D K F K G L K T E P M K G S E D E K H G G T V L T A L G

Shark Mb - - - - T E E H S N K V W A V E P D I P A S E L A E L E K E F H K E E K D I P F E K E I - P V O O L G N N E D L R K B V T M L R A L G

Human α N A V A R V D D M P N A E S A E S D L B A H K E L V D R V R F K L S S H C L L V T E A A R P A E F T R A V H A S E D E F E A S S T V E T S K R R

Cow α R A V E H L D D I P G A E S E S D L B A H K E L V D R V R F K L S S H S L V L T E A S H P S D F T R A V H A S L D F L A N S T V E T S K R R

Chicken α N A V E H A D D I S G A L S E S D L B A H K E L V D R V R F K L S S G O C L L V E V A H P A E L L A P R V H A S L D E F C A V G T E T A K R R

Shark α K A C D H L D D L K T H E L K L A T F F G S E L K V D R A N F O Y E S Y C L E V A L A V H L - T E F S P E T H C A L D K E E L T N V C H E S S P R R

Human β D G L A E D N L K G T E A T L E S L C D K L H V D P E N F R E L G N V L C V E A H F G K E R T P P Q A A Y O R V V A G A N A L A H K H H

Cow β D G M K E D D L K G T F A A S E L R C D K L H V D P E N F K L G N V L V V V L A R N F G E F T P V L Q A D P Q R V V A G A N A L A H K H H

Chicken β D A V K N S D I K N T F S G L S L C D K L H V D P E N F L G D I L I L V E A M P S K D F T R E C Q A M W O L V R V V A G A N A L A H K H H

Shark β V A Y T H L G D V K S Q E T L S K K A E E L H V D V E S K E L A K C F V V E G I L L D K F A R O T A I W E Y F V V D V I S K E E H

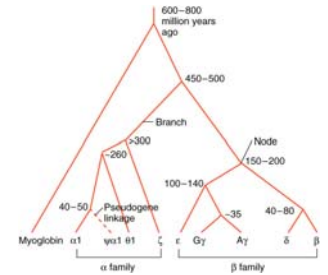
Human Mb G I E K K G H R E A E I K P L A Q S A T E K H I P V K Y L E F S E C I I Q V L O S K H P G D G A D A G A M N K A L E F R K B M A S N X K E L G F O G

Cow Mb G I E K K G H R E A E V E H A S B A N K H K V P F Y L E F S D A I H V L K A K H P I N F A A D A G A M S A L L F R N D A A K K V L G F G H G

Chicken Mb A Q L E K K G H R E A D E F P A Q T B A T E K H I P V K Y L E F S E V I I K Y I A K H A D F G A D G A M M A L L F R D M A L X R E E F G F O G

Shark Mb N I E K O K G H S T N Y E L A D T I R I N K H I P P K N V L E T N I A V K V L T E M P S D M I G M Q E S P S V F T V I C S D L E T L E K E A D F O G

The amino-acid sequences of the globins can be used to infer the history of the gene duplications that gave rise to the whole family!



600-800 million years ago

450-500

Branch

Node

>300

260

100-140

150-200

40-50

Pseudogene linkage

35

40-80

Myoglobin $\alpha 1$ ζ ϵ γ $A\gamma$ δ β

α family

β family

α -globin genes on human chromosome 16

LCR ζ ϵ γ $\gamma 1$ $\gamma 2$ δ β

β -globin genes on human chromosome 11

LCR $\psi 2$ ϵ γ $A\gamma$ $\psi 1$ δ β

mb

mb

And their chromosomal locations also reflect this history (more closely related members tend to be located nearer to each other).

Summary: Stuff Happens (a lot), giving rise to variation

Mutations of many different kinds occur every generation in the genomes of most species.

Most of these are either harmless (e.g., those in junk) or harmful (most of those in genes).

Rates of significantly deleterious mutation have been estimated at 0.01-1 mutation per genome per generation, in various different species.

A small minority of mutations are beneficial, at least under some conditions.

Some of these are maintained as polymorphisms by selection.

But most polymorphism is thought to be "nearly neutral" (i.e., neither strongly favored nor strongly opposed by selection, at least on average).

New genes begin as polymorphic mutations of a special kind (duplications). Like other mutations, most are probably harmless or harmful, but a very few increase the fitnesses of their carriers and rise in frequency, eventually adding a novel genetic locus to the genome of their species.

Evolution can be viewed as the movement of mutations on gene genealogies or "gene trees".

The rate at which new alleles are introduced is governed by μ (the mutation rate).

The rate at which they move on the tree is governed by N (the population size), and possibly also by selection.

Definition: the allele frequency p is the proportion of all gene copies that are "A" or "A₁" or whatever we choose to identify with p .

IN THE BEGINNING, PURITY...

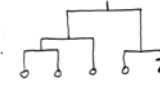


$$P = 1$$

$$q = 0$$

$$H = 2pq = 0$$

... THEN A MUTATION...

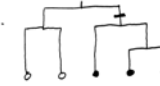


$$P = 0.8$$

$$q = 0.2$$

$$H = 2pq = 0.32$$

... LATER...



$$P = 0.4$$

$$q = 0.6$$

$$H = 2pq = 0.48$$

... AND FINALLY, PURITY AGAIN



$$P = 0$$

$$q = 1$$

$$H = 2pq = 0$$