Molecular Evolution

In 1952, Frederick Sanger and coworkers determined the complete amino acid sequence of insulin.

Since that time, the amount of sequence information has grown exponentially.

For example, Genbank contains all publicly available DNA sequences, which amounts to more than 3.8 billion basepairs from 4.8 million sequences!

In addition, the entire genomes of over thirty organisms have been sequenced, including two eukaryotes (the fungus, *Saccharomyces cerevisiae*, and the nematode, *Caenorhabditis elegans*).

The human genome is also well on its way to being sequenced, with an expected date of completion in the year 2003.

Molecular evolution is a new field born from this explosion of molecular information and from our desire to understand how and why molecular sequences have evolved to be the way they are.

Topics in Molecular Evolution

In the next two lectures, we will discuss a few examples of research in the field of molecular evolution:

- Detecting selection by examining

  (1) Substitution rates

  (2) Variability within a population

  (3) Replacement versus silent changes

- Detecting historical events by examining

  (1) Number of differences between sequences

- Detecting relatedness by examining

  (1) Similarity among sequences within a population

  (2) Similarity among sequences from different species

Evidence in the DNA: Selection

1. Substitution Rates

An early theme developed within this course is that, while it is easy to establish that evolutionary change has happened, it is difficult to establish whether selection has played a role in this change.

DNA can provide a record of selection.

If we compare DNA sequences from different organisms, we can estimate the rate at which mutations appear and fix, causing basepair substitutions.

Substitution rate = the rate at which mutant alleles rise to fix within a lineage.

(For neutral mutations, the substitution rate within a population equals the mutation rate, since $2N\mu$ mutations appear, each with a $1/(2N)$ chance of fixation.)

Variation in the rate of substitutions among regions of the genome is due, in part, to variation in the form of selection.

Recall that:

- A new beneficial mutation has a chance of fixing within a diploid population of ~ 2s.
- A new neutral mutation has a chance of fixing within a diploid population of ~ 1/(2N).
- A new deleterious mutation has almost no chance of fixing within a large population.

Therefore, the nucleotide substitution rate is expected to be:

- highest when mutations are beneficial,
- intermediate (μ) when mutations are neutral,
- lowest when mutations are deleterious.

→ Silent (or synonymous) mutations are more likely to be neutral.

→ Replacement (or non-synonymous) mutations are more likely to experience selection, but the form and strength of selection depends on the gene and its function.

Gene	Non-synonymous Substitution Rate (x 10^9)	Synonymous Substitution Rate (x 10^9)
Histone 3	0.00	3.94
Histone 4	0.00	4.52
Actin α	0.01	2.92
Myosin β	0.10	2.15
Insulin	0.20	3.03
Growth hormone	1.34	3.79
Immunoglobulin k	2.03	5.56

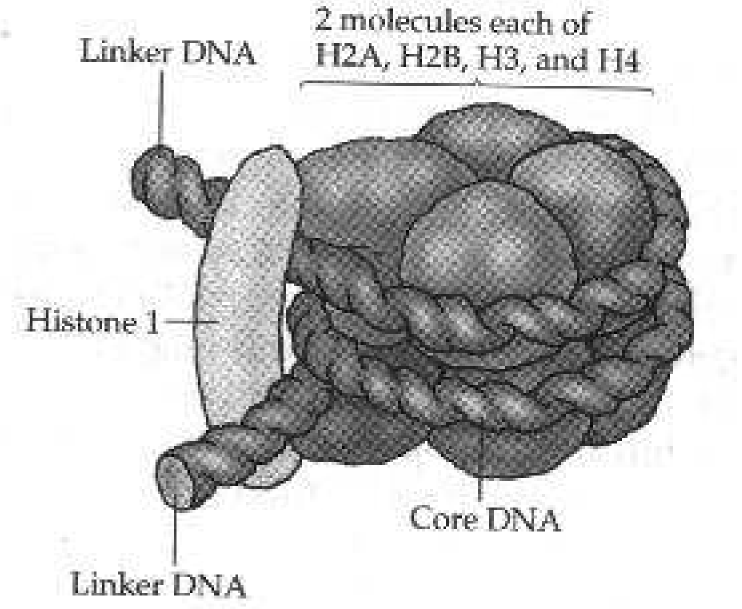
(From Li, 1997. Based on sequence differences between humans and rodents, estimated to have diverged 80MYA.)

Histones, for example, appear to have a very low rate of replacement substitutions.

This suggests that mutations causing basepair changes in histone genes may be deleterious.

Why?

Histones are DNA-binding proteins around which DNA is coiled to form chromatin. Many positions within the protein interact with the DNA or with other histones. In addition, histones are highly compact and alkaline.



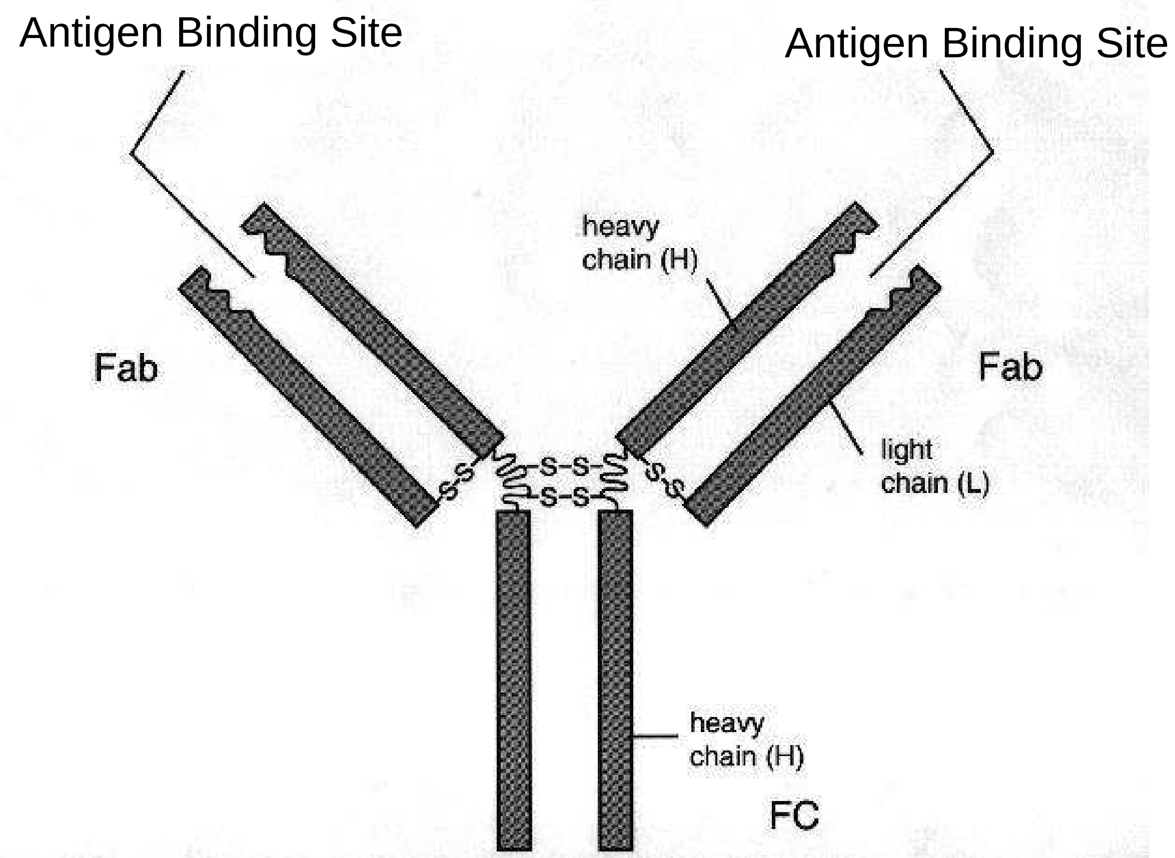
(From Li 1997)

Most amino acid changes in histone proteins may have negative or even disastrous consequences.

➡ Histone proteins have strong *functional constaints.*

Conversely, the amino acid sequences of immunoglobulins (= antibody protein) evolve at a much higher rate.

In particular, the active sites (the complementarity-determining regions) of many immunoglobulins actually have higher rates of replacement changes than silent ones!



(From Ludwig Institute.)

It is thought that selection favors mutations in these regions, thereby increasing the diversity among antibodies produced by the body and improving the immune response.

2. Levels of variability within a population

Another method to detect selective events is to examine the level of variability currently present within a population.

If a beneficial mutation appears and sweeps through a population, what will happen to the level of polymorphism present at neighboring DNA sites?

For example, Berry et al (1991) sequenced 1.1 kilobases of the cubitus interruptus Dominant (ciD) locus in *Drosophila*.

ciD is located on a tiny fourth chromosome, which undergoes no recombination.

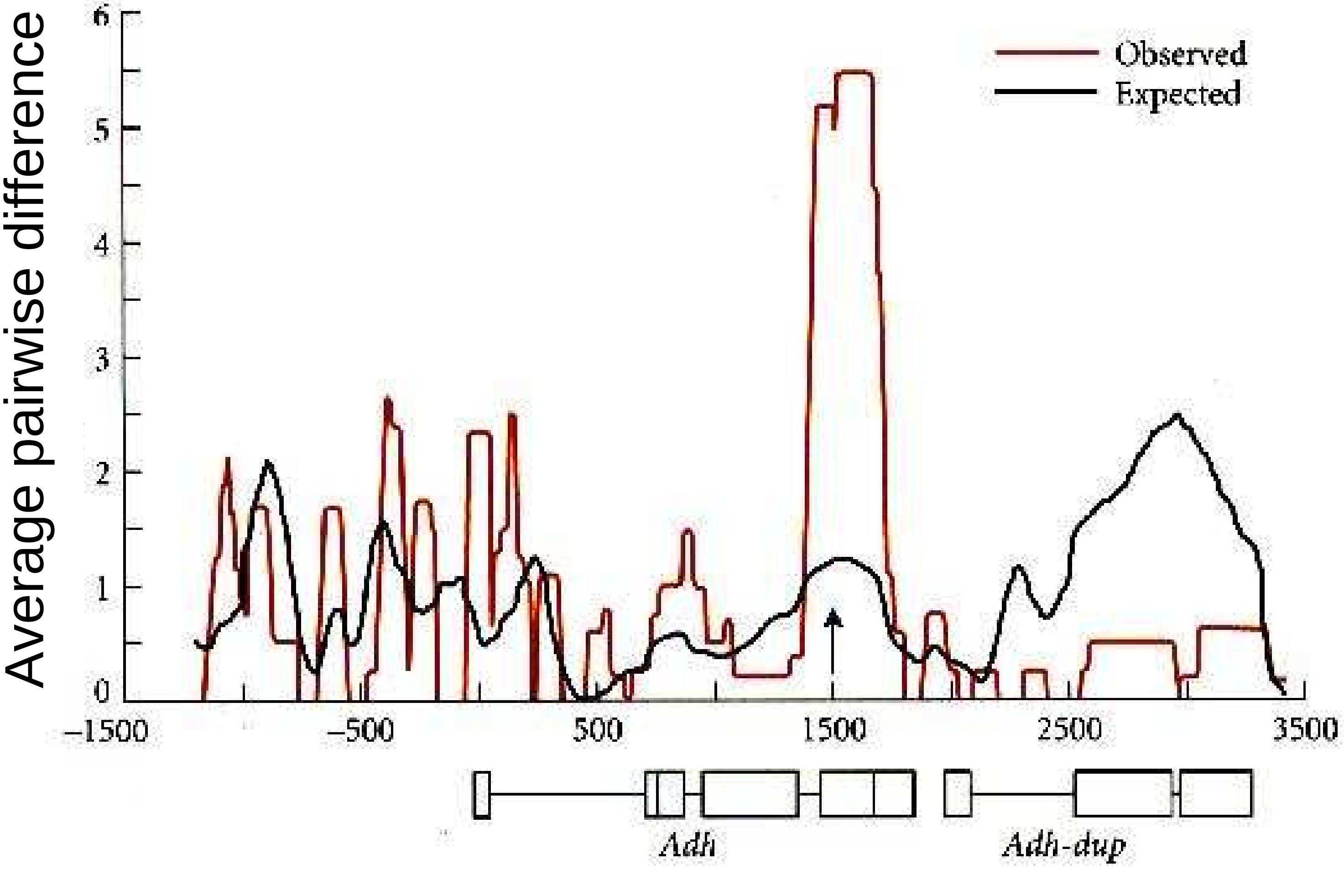
They found NO VARIATION among ten *D. melanogaster* sequences and only one basepair difference among nine *D. simulans* sequences, even though there were 54 differences between the species.

By contrast, other genes from the same individuals showed normal levels of polymorphism.

Berry et al (1991) argued that recent selective sweeps in both species may have eliminated most of the polymorphism on the fourth chromosome.

If there is overdominance at a site, what will happen to the level of polymorphism present at neighboring DNA sites?

Kreitman and Hudson (1991) sequenced a 4750 basepair region near the alcohol dehydrogenase (ADH) gene from 11 individuals of *D. melanogaster* and found higher than expected levels of polymorphism:



(From Futuyma 1998)

There is only one amino acid polymorphism (Adh^F/Adh^S) within this region, which occurs at site 1490 (see arrow).

Kreitman and Hudson (1991) hypothesize that there is selection maintaining a polymophism at or near this site.

ADH is an enzyme that breaks down ethanol.

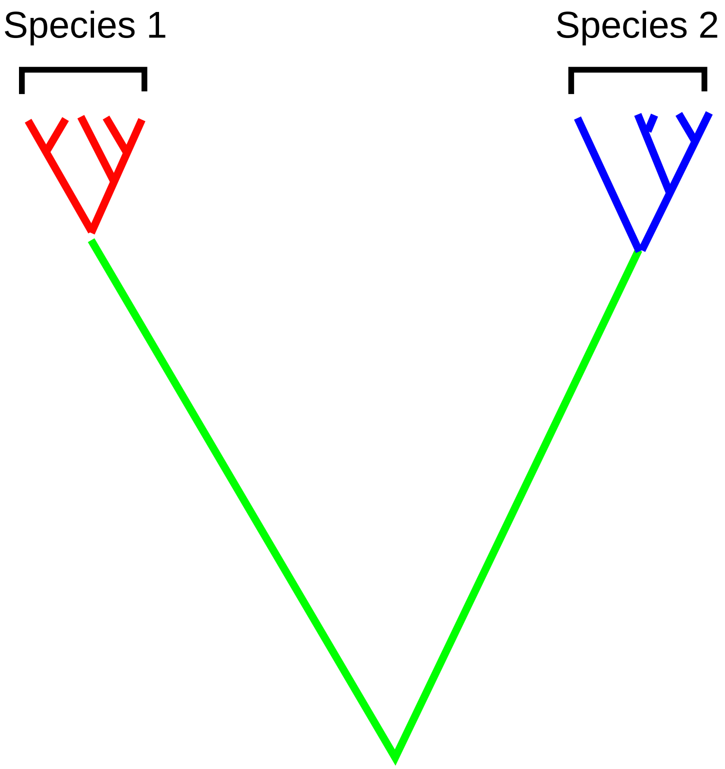
Flies carrying the Adh^F allele survive better when their food is spiked with ethanol than do flies carrying the Adh^S allele (Cavener and Clegg 1981).

Nevertheless, the factors maintaining the Adh^F/Adh^S polymorphism remain unknown.

3. Replacement versus silent changes

McDonald and Kreitman (1991) compared substitutions between species and polymorphisms within a species to construct a test to detect selection.

Imagine that five sequences are obtained from each of two species and that the tree relating these sequences is:



Any mutation that happens on a red branch will appear as a polymorphism within species 1.

Any mutation that happens on a blue branch will appear as a polymorphism within species 2.

Any mutation that happens on a green branch will appear as a fixed difference between species 1 and 2.

If mutations occur randomly over time and if the chance that a mutation does or does not cause an amino acid change remains constant, then the ratio of replacement to silent changes should be the same along any of these branches.

If mutations are neutral, any of these mutations has an equal chance of being preserved.

McDonald-Kreitman test

H_0: The ratio of replacement to silent changes among polymorphic sites (within a species) should equal the ratio among fixed differences (between species) in the absence of selection.

If new mutations are advantageous, they will fix rapidly and cause more fixed differences between species.

If new mutations are deleterious, they will rarely fix, but they will temporarily create polymorphisms.

These effects of selection should be stronger on mutations that change the amino acid sequence (replacement) than ones that don’t (silent).

➡ An excess of amino acid differences between species should be seen when replacement mutations have been beneficial and fixed by selection.

➡ A lack of amino acid differences between species should be seen when replacement mutations have been deleterious and eliminated by selection.

	Fixed differences	Polymorphic sites
ADH gene		
Replacement	7	2
Silent	17	42
G6PD gene		
Replacement	21	2
Silent	26	36

(From Li, 1997. ADH study by McDonald and Kreitman, 1991: 12 *D. melanogaster,* 6 *D. simulans,* and 24 *D. yakuba* sequences. G6PD study by Eanes et al, 1993: 32 *D. melanogaster* and 12 *D. simulans*.)

For both genes, the ratio of replacement to silent substitutions is significantly lower among polymorphic sites within species (2 : 42 for ADH and 2 : 36 for G6PD) than among fixed differences between species (7 : 17 for ADH and 21 : 26 for G6PD).

➡ The null hypothesis that selection is absent is rejected in both cases.

The excess of replacement differences between species suggests that mutations have been positively favored.

Indeed, G6PD is an important enzyme in the metabolism of pentose sugars, and it has been argued that amino acid changes may have been selectively favored in changing environments.

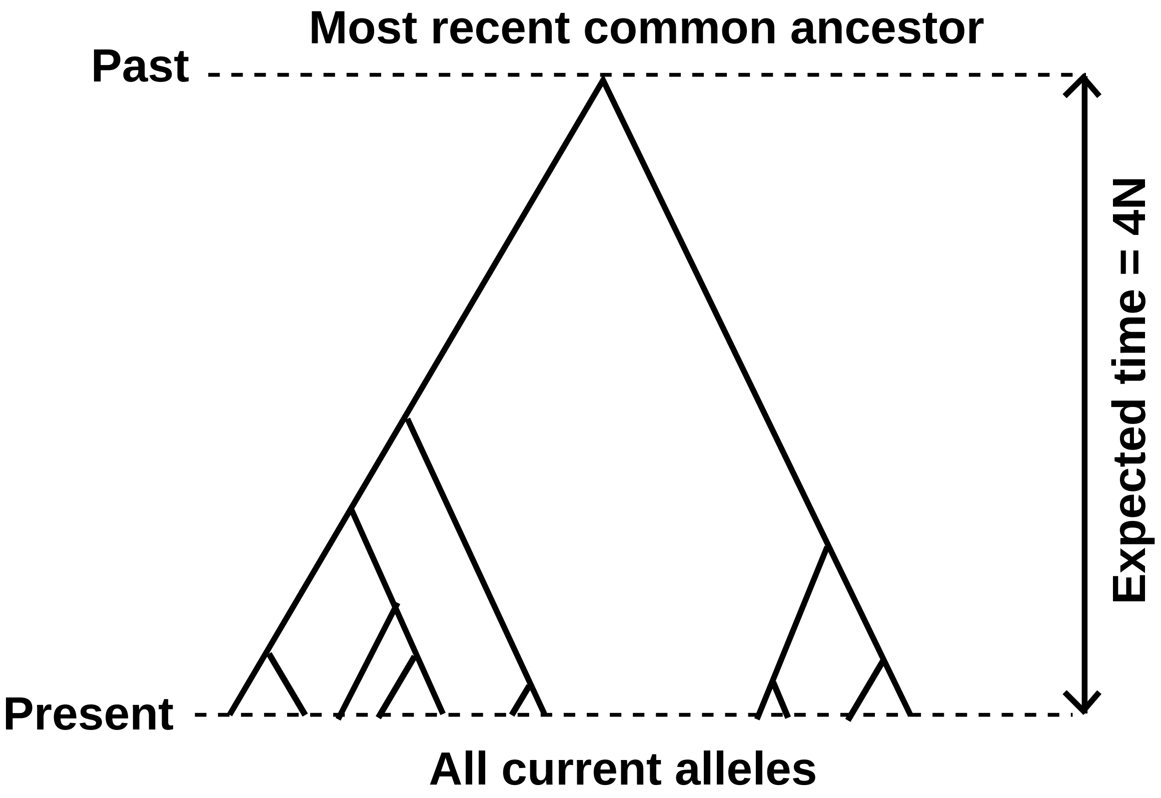
Evidence in the DNA: History

DNA sequences record other historical events besides selection.

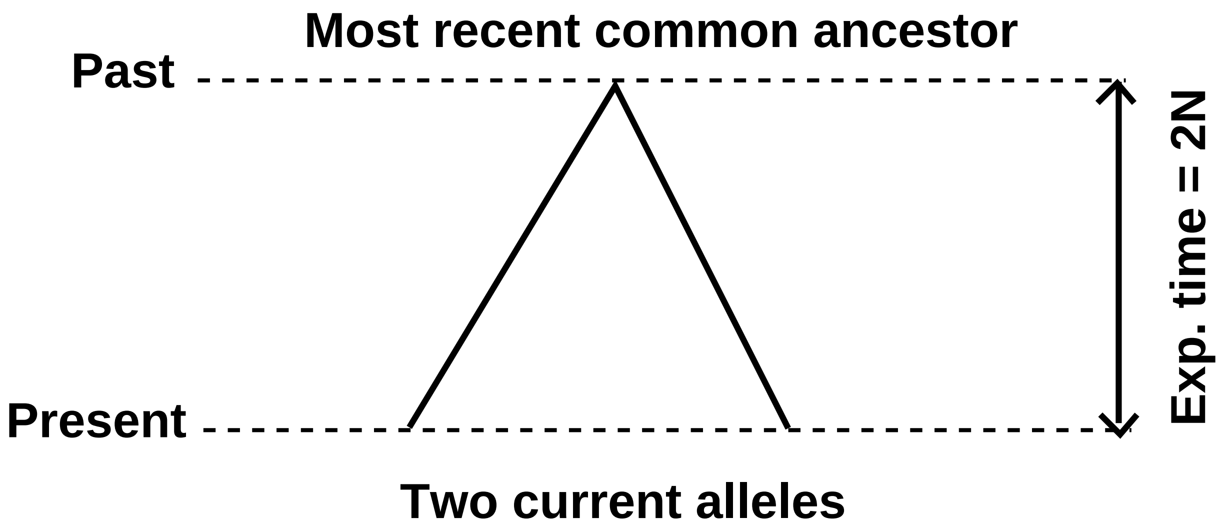
Sequence comparisons have been used to trace past migration events and also past changes in population size.

For example, Takahata et al (1995) estimated the population size of early humans, using coalescence theory.

Recall that all the alleles currently present within a population are descended from a common ancestor that lived, on average, 4N generations ago.



Any two alleles chosen at random from the current population will share a common ancestor 2N generations ago, on average:



These predictions also hold for two sequences of DNA.

Therefore, if mutations occur at a rate μ per generation per basepair on each of the two branches leading to the two current sequences, the proportion of sites that differ between the two sequences is expected to be $4N\mu$.

Takahata et al (1995) used a more sophisticated version of this idea to estimate N from human gene sequences.

They studied 49 different loci from human populations.

The total number of differences between two randomly chosen sequences varied from zero at 37 loci to five at one locus.

Using an estimated mutation rate of $2\ 10^{-8}$ per basepair, Takahata et al estimated an effective population size for humans of 10,000!

[NOTE: There may have been more individuals alive. 10,000 represents the "effective" population size -- the size of an ideal population of constant size that would have led to the observed amount of sequence divergence.]

Even though the current population size of humans is nearly 6 billion, the molecular sequence divergence among humans reflects a much smaller historical population size (~ 400,000 to 50,000 years ago).

➡ Humans are genetically very similar, due in part to a recent population explosion from a relatively small number of individuals within the last few hundreds of thousands of years.

Evidence in the DNA: Relatedness

Beyond looking for clues to our past, molecular data can be used to tell us about the diversity within and among species currently alive.

1. Human genetic diversity

> "Accustomed as we are to noticing variations in skin color or facial structure, we tend to assume that the differences between Europeans, Africans, Asians, and so on must by large...This simply is not so: the remainder of our genetic makeup hardly differs at all."
>
> -- Cavalli-Sforza and Cavalli-Sforza (1995) p. 124

In a major study of human polymorphisms, Cavalli-Sforza and collaborators studied the allele frequencies of different populations at 110 genes.

In all cases, the differences between "races" were quantitative *not* qualitative.

That is, there was not a single gene for which two races were totally different.

Instead, slight differences in allele frequencies were observed at most loci, e.g.:

Allele Frequency	European	African	Asian
GC-1	72%	88%	76%
HP-1	38%	57%	23%

In a similar study, Nei and Roychoudhury (1982) found that 85% of the genetic variation in the human species exists *within* populations and that only 8% is among the major "races".

> "If everyone on earth became extinct except for the Kikiyu of East Africa, about 85% of all human variation would still be present in the reconstituted species."
>
> -- Lewontin et al. (1984)

Similarly, Brown (1980) studied the mitochondrial DNA from 21 humans of diverse origin.

868 nucleotide sites were examined and only a few differences were observed between any pair of individuals.

Overall, the sequences differed from a postulated ancestral mtDNA sequence at only 0.18% of the sites.

Using a substitution rate estimate of 10^{-8} per basepair per year, Brown concluded that humans passed through a severe population bottleneck ~180,000 years ago.

2. Mammalian genetic diversity

How much do we differ genetically from other mammals?

Species comparison	% sequence difference
Human-Chimp	1.45
Human-Gorilla	1.51
Human-Orangutan	2.98
Human-Rhesus Monkey	7.51

(From Li 1997. Based on 5.3 kb of non-coding DNA.)

Surprisingly little!

Tables, such as the above, provide information about the relatedness of different species.

This data can be used to reconstruct the phylogenetic relationships among the species involved.

In the next few lectures, we'll discuss how phylogenetic reconstruction is accomplished and look at some interesting phylogenies.

SOURCES:

- **Li (1997) Molecular Evolution. Sinauer Associates, MA.**
- Futuyma (1998) Evolutionary Biology. Sinauer Associates, MA.
- Cavalli-Sforza and Cavalli-Sforza (1995) The Great Human Diasporas. Addison-Wesley, NY.