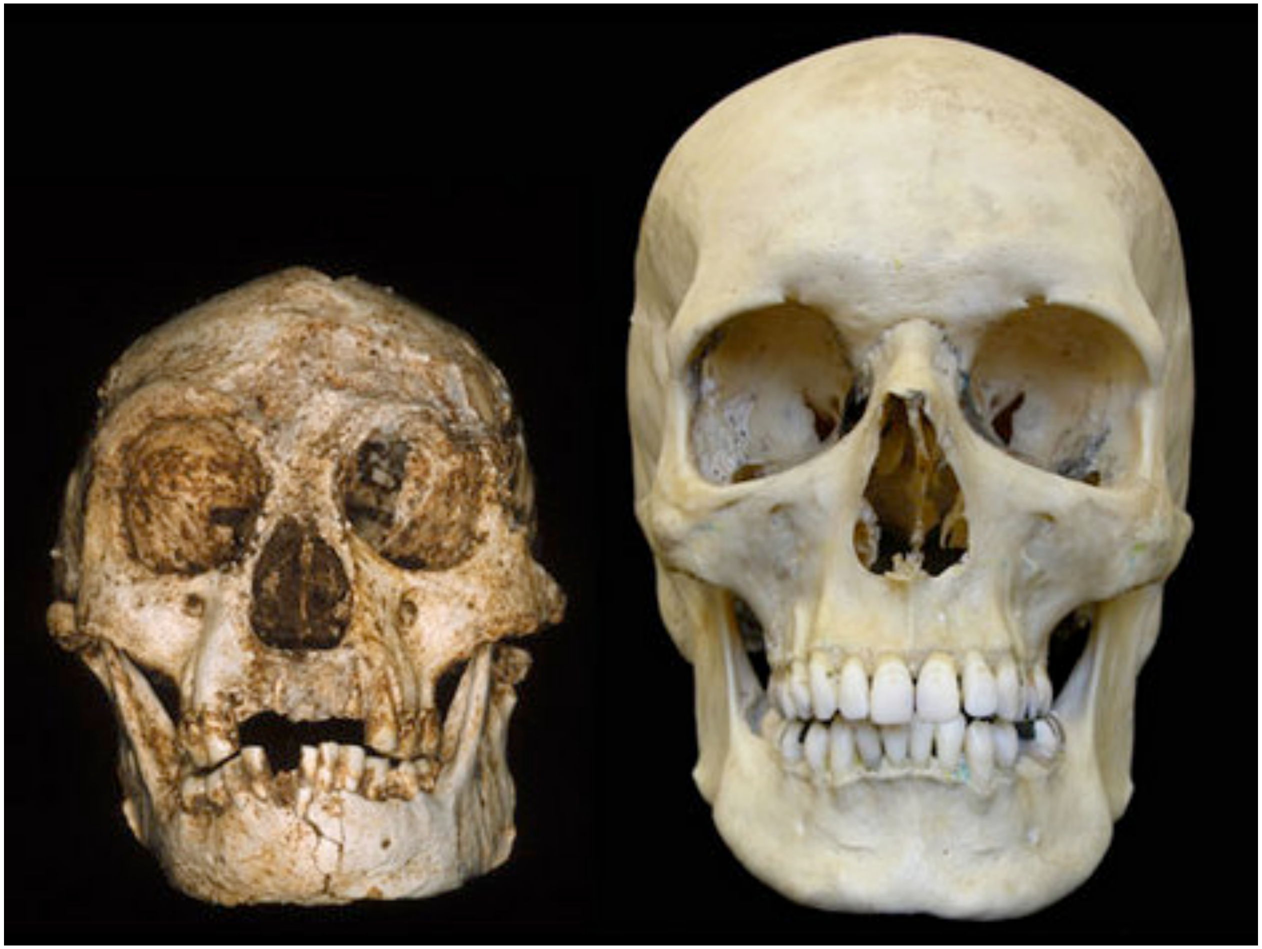
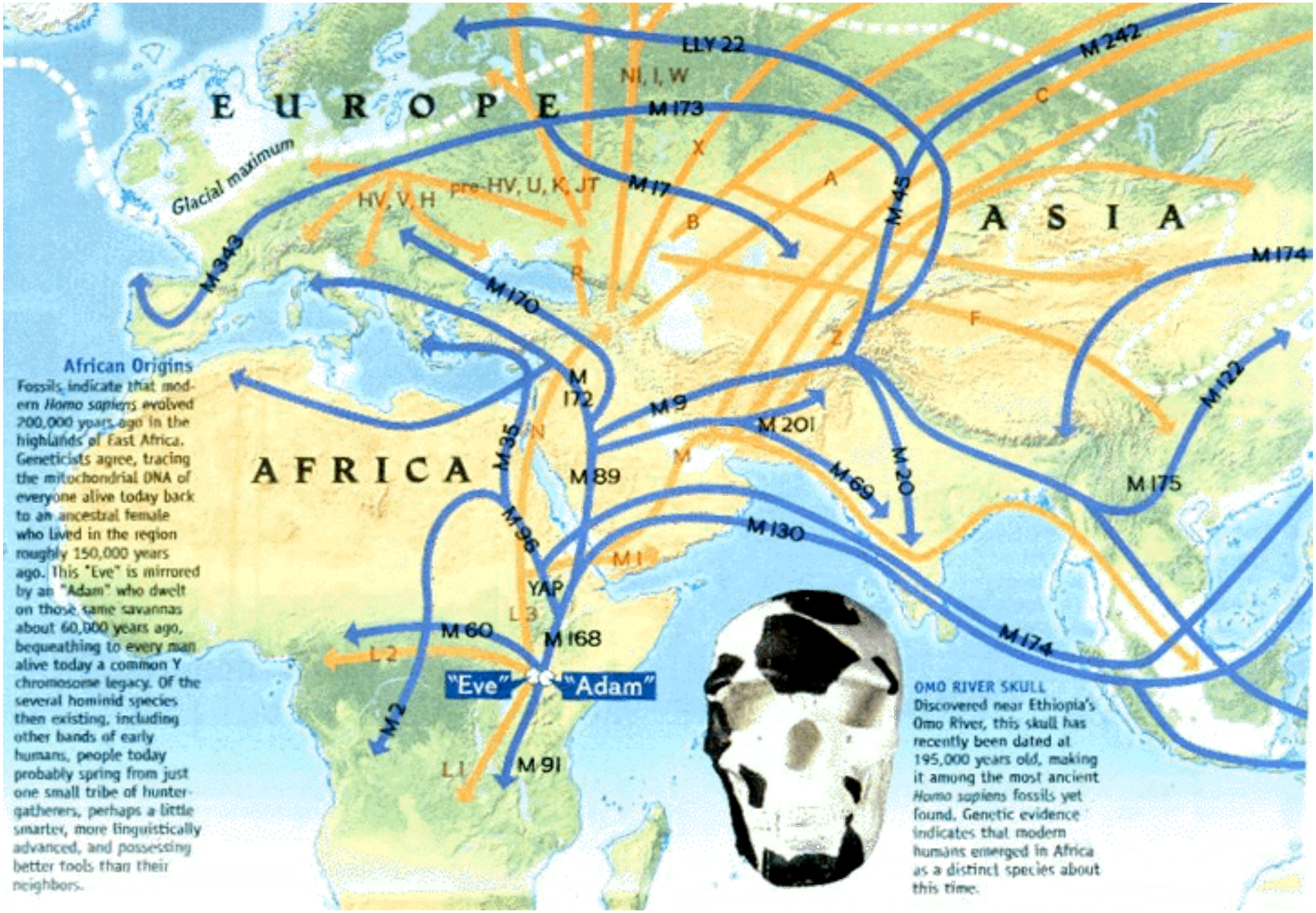
Natural selection on the molecular level



http://www.anthroblogs.org/nomadicthoughts/archives/2005/09/



African Origins

Fossils indicate that modern *Homo sapiens* evolved 200,000 years ago in the highlands of East Africa. Geneticists agree, tracing the mitochondrial DNA of everyone alive today back to an ancestral female who lived in the region roughly 150,000 years ago. This "Eve" is mirrored by an "Adam" who dwelt on those same savannas about 60,000 years ago, bequeathing to every man alive today a common Y chromosome legacy. Of the several hominid species then existing, including other bands of early humans, people today probably spring from just one small tribe of hunter-gatherers, perhaps a little smarter, more linguistically advanced, and possessing better tools than their neighbors.

OMO RIVER SKULL

Discovered near Ethiopia's Omo River, this skull has recently been dated at 195,000 years old, making it among the most ancient *Homo sapiens* fossils yet found. Genetic evidence indicates that modern humans emerged in Africa as a distinct species about this time.

Map from Genographic project

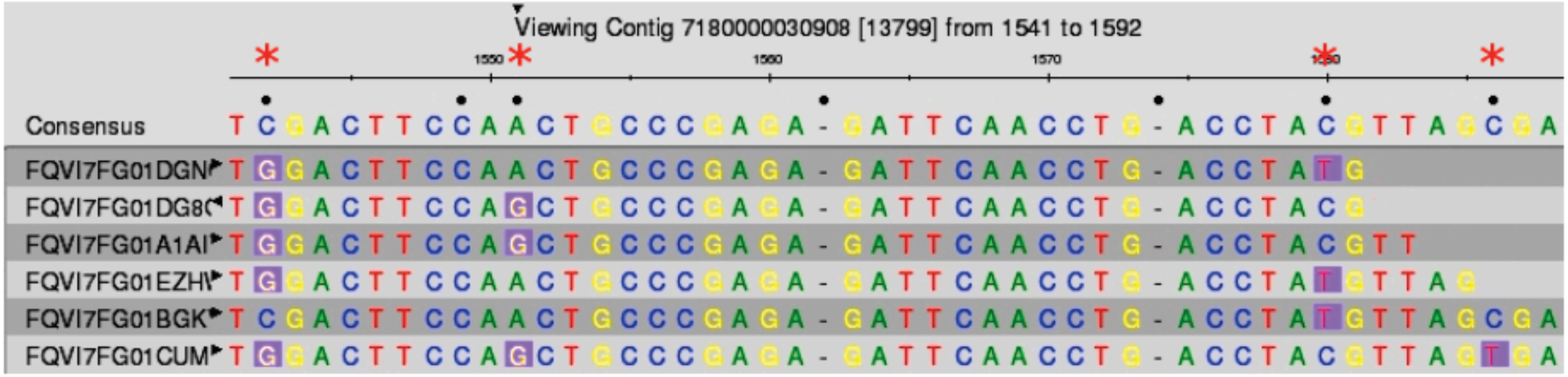
Review: some terms

- Polymorphism
  - Any difference between individuals in a population (e.g., humans)
- SNP
  - Single nucleotide polymorphism
  - Pronounced “snip”, usually single base differences (A or T polymorphism)
- Allele
  - An element of the set of possibilities at a specific location where there is a polymorphism
  - Mary has an “A” allele and John has a “T”

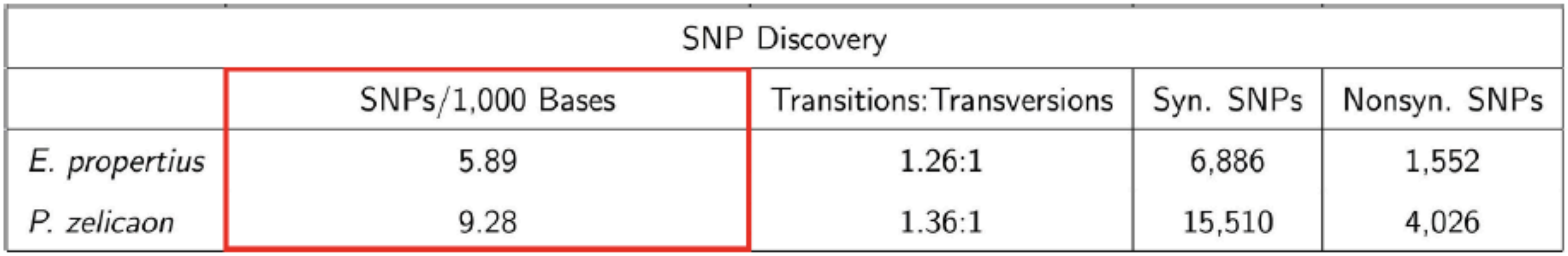
In class exercise

- Consider the following four individuals:
  - Calvin: AATGTA
  - Hobbes: ATTA
  - Snoopy: ACCATG
  - Tom: ATTA

- How to find polymorphisms (SNPs, indels) and what are they?
- How many alleles are there?



```
Viewing Contig 7180000030908 [13799] from 1541 to 1592
Consensus     T C G A C T T C C A A C T G C C C G A G A - G A T T C A A C C T G - A C C T A C G T T A G C G A
FQVI7FG01DGN  T G G A C T T C C A A C T G C C C G A G A - G A T T C A A C C T G - A C C T A T G
FQVI7FG01DG8C T G G A C T T C C A G C T G C C C G A G A - G A T T C A A C C T G - A C C T A C G
FQVI7FG01A1AI T G G A C T T C C A G C T G C C C G A G A - G A T T C A A C C T G - A C C T A C G T T
FQVI7FG01EZH\ T G G A C T T C C A A C T G C C C G A G A - G A T T C A A C C T G - A C C T A T G T T A G
FQVI7FG01BGK  T C G A C T T C C A A C T G C C C G A G A - G A T T C A A C C T G - A C C T A T G T T A G C G A
FQVI7FG01CUM  T G G A C T T C C A G C T G C C C G A G A - G A T T C A A C C T G - A C C T A C G T T A G T G A
```



SNP Discovery				
	SNPs/1,000 Bases	Transitions:Transversions	Syn. SNPs	Nonsyn. SNPs
E. propertius	5.89	1.26:1	6,886	1,552
P. zelicaon	9.28	1.36:1	15,510	4,026

O'Neil et al.,2010, BMC Genomics.

Substitution rate

- Consider a new mutation in a diploid
  - 1 chromosome from mom and 1 from dad

- Any new mutation in a population of *N* individuals now is present at a frequency of 1/2*N*

- What are the chances this becomes the allele all members of a population have at random?

Substitution rate

- Let u be the mutation rate

- Substitution rate is:
  - $P = 2Nu(1/2N) = u$

- Note that the substitution rate is independent of population size!

Lecture today

- “Nothing makes sense except in the light of evolution”
  - Theodosius Dobzhansky

Selection

- "survival of the fittest"

- Two types in a genome:
  - – Negative, or purifying selection that removes bad mutations from a population
  - – Positive selection, that leads to an increase in allele frequency in a population

Case study: HIV virus

- Evolves during the course of infection, such that the virus that kills a patient is very different from the original strain

- Accomplished by a high mutation rate and short generation times.

- Example:
  - Resistance to AZT developing shortly after introduction in 1987

HIV

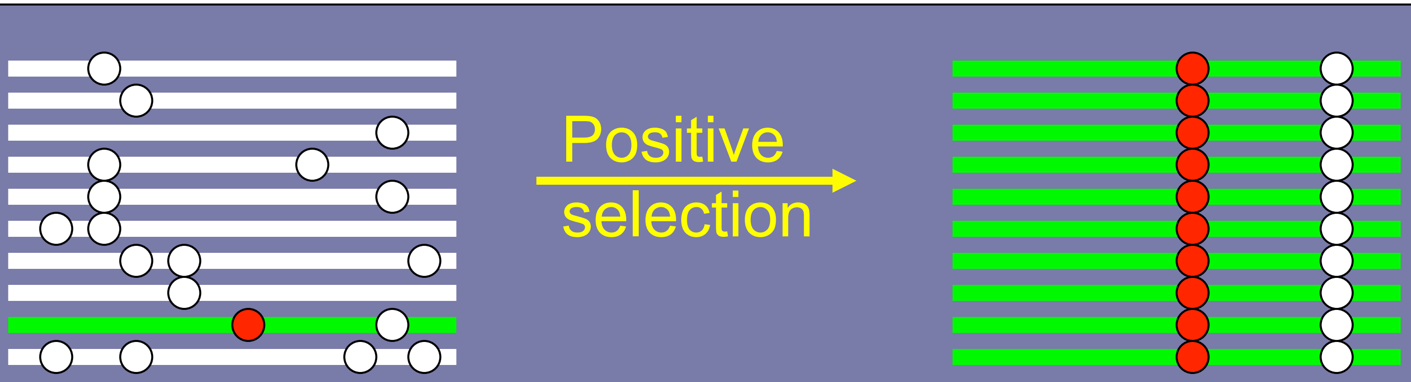
- Infects helper T cells, which are responsible for identifying infected targets.

- Infected cells are tagged by an epitope that sticks out and can be recognized by the immune system.

- These change quickly; mutation rate in HIV is 1000 times that of humans and takes 1.5 days for a single generation.

Types of mutations

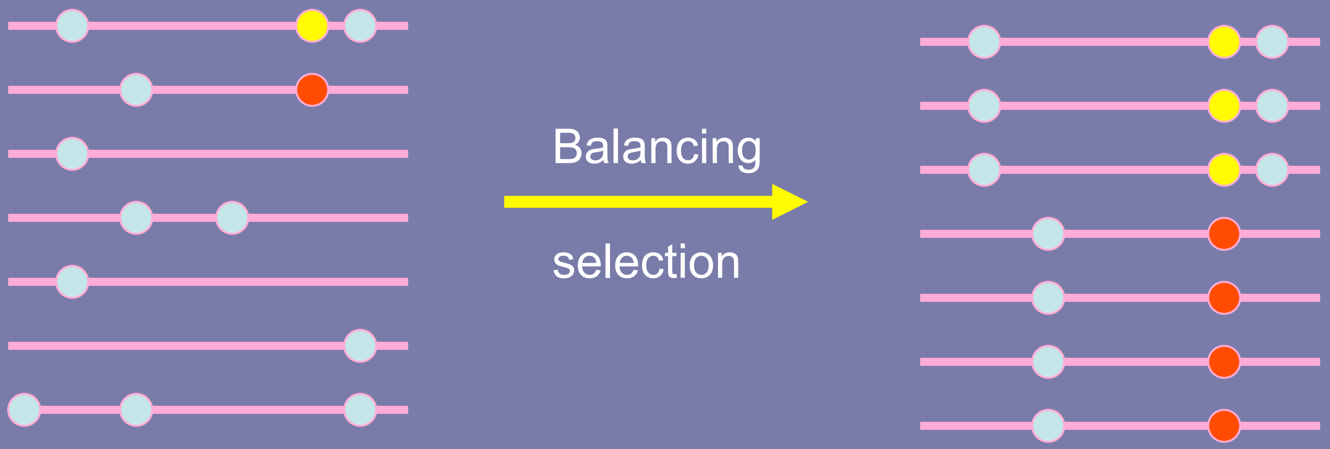
- Neutral
  - – No effect, usually in third “wobble” position
- Positive
  - – Leads to a fitness advantage
  - – FOXP2, olfactory receptors
- Negative
  - – Leads to a loss of fitness.
  - – 80-90% of changes that affect proteins are negative



Positive selection

🔴 Advantageous mutation

○ Neutral mutation



Balancing selection

●● ‘balanced’ mutation

○ Neutral mutation

Neutral model

- Proposed by Kimura in 1968

- Most changes have no effect under:
  - Random mating
  - Constant population size

- Two deviations:
  - Selection
  - Demography (population structure)

Adding selection to the system

- http://evolutiongenetics.georgetown.edu/simulations/driftselection/

Ka/Ks test

- Lets define two types of differences:
  - Synonmous, that do not change proteins (*Ks*)
  - Nonsynonmous, that DO change proteins (*Ka*)
- If both normalized rates are equal, then the ratio of the two is one.
- Interesting cases:
  - *Ka / Ks* > 1 = positive selection
  - *Ka / Ks* < 1 = purifying selection

Estimating *Ka/Ks*

- Count the number of possible synonymous and non-synonymous sites

- Then count the number of differences

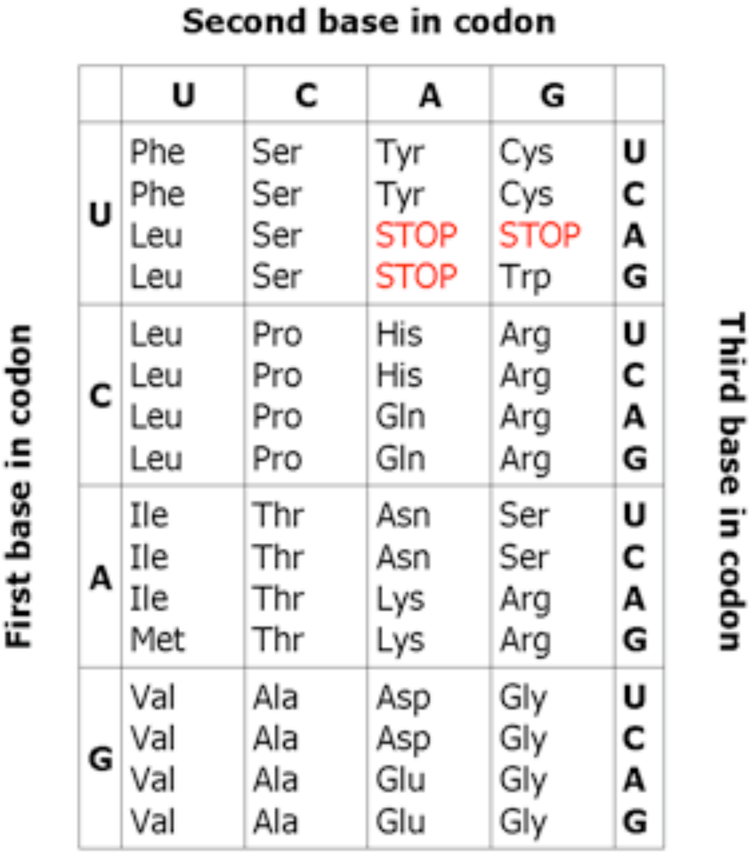
- After correction, we compute the number of non-synonymous and synonymous substitutions per site

Nei and Gojobori's algorithm

- Step 1: Count A and S sites

Consider: UUA that codes for Leucine

We will denote f_i as the fraction of changes that result in a synonomous site



Second base in codon

First base in codon

Third base in codon

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

- $f_1 = 1 / 3$ $f_2 = 0/3$ $f_3 = 1/3$

Calculating s_c and a_c

- Let s_c be:

$$s_c = \sum f_i$$

- And a_c be:

$$a_c = 3 - \sum f_i = 3 - s_c$$

Number of sites

- For a gene containing *r* codons:

$$S_c = \sum_{k=1}^{r} s_c(c_k) \qquad A_c = 3r - S_c$$

- Because these are defined on alignments, we use:

$$\widehat{S}_c = \left(S_{c1} + S_{c2}\right)/2 \qquad \widehat{A}_c = \left(A_{c1} + A_{c2}\right)/2$$

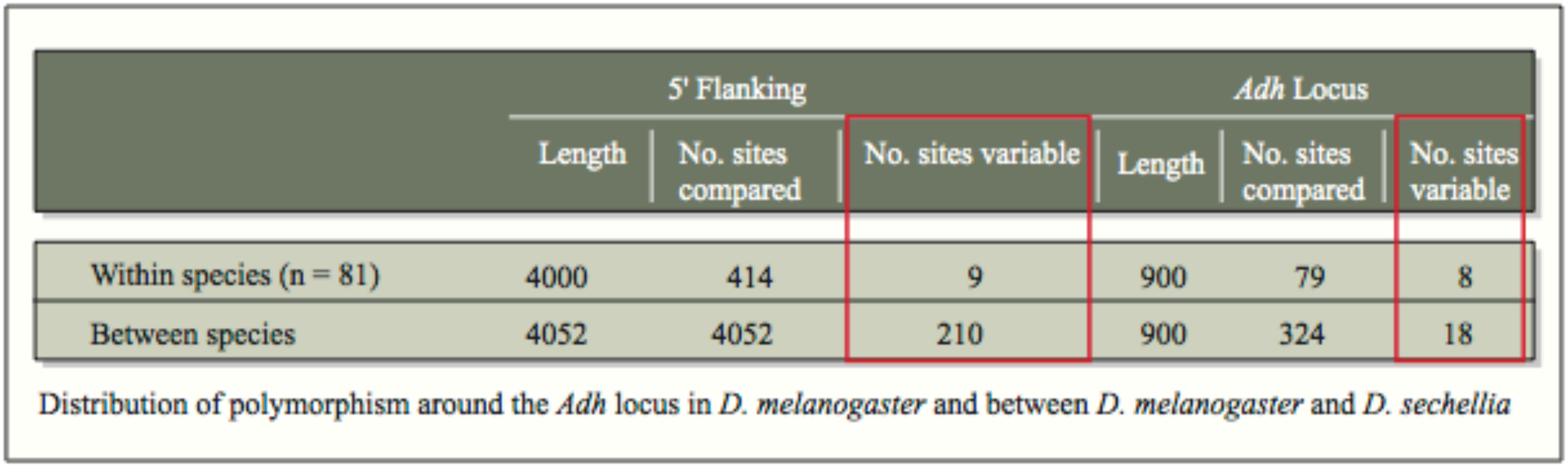
Use of *Ka / Ks* test

- One of the first tests proposed and to show positive selection in genomes.

- Problems:
  - Very conservative; often misses interesting regions that other tests find

HKA test

Hudson, Richard, Martin Kreitman, and Montserrat Aguade. "A Test of Neutral Molecular Evolution Based on Nucleotide Data." *Genetics* 116, no. 1 (1987): 153-159.



	5' Flanking			*Adh* Locus		
	Length	No. sites compared	No. sites variable	Length	No. sites compared	No. sites variable
Within species (n = 81)	4000	414	9	900	79	8
Between species	4052	4052	210	900	324	18

Distribution of polymorphism around the *Adh* locus in *D. melanogaster* and between *D. melanogaster* and *D. sechellia*

Figure by MIT OpenCourseWare, based on paper cited above.

apply chi-squared test to summary statistics of polymorphism, divergence

Conclusion: *Adh* exhibits excessive polymorphism

One more thing

- There is some confusion of substitution rate (new alleles are fixed in a population) as previously defined and genetic distance (# differences between two sequences)

- If we know the divergence time, T, the “other” substitution rate can be defined as
  - P = K / (2 T)
  - We divide by 2 T as there are two lineages

Jukes-Cantor model

- If the distances between two sequences are large, we may be missing some.
- On the other hand, if the distance is short then most differences should be real.
- The model of Thomas Jukes and Charles Cantor (1969) addresses this.

Correction

- We want to correct for the possibility of multiple substitutions.

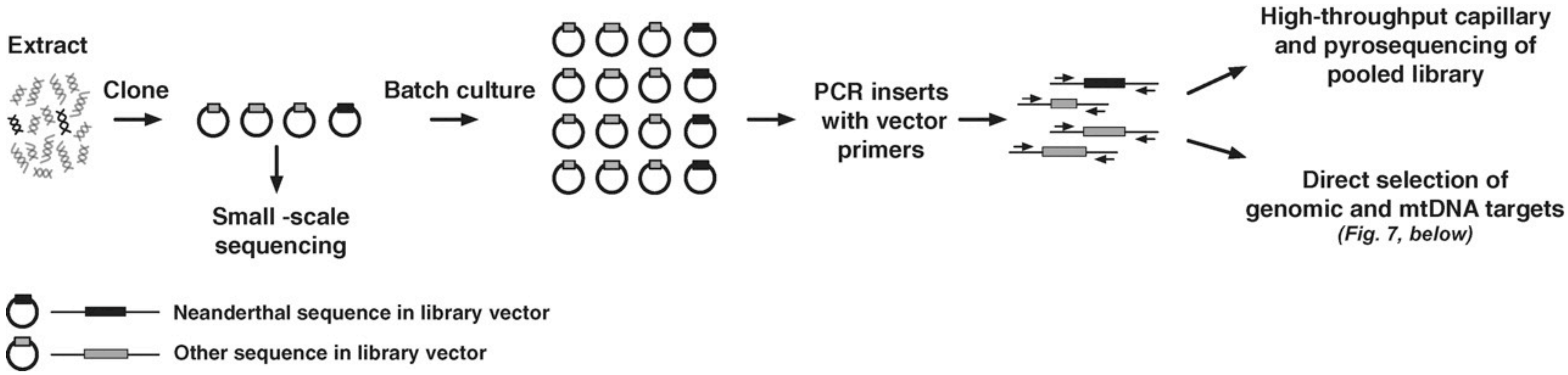
- Let *d* be the fraction of sites that differ.

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}d\right)$$

- This states that the substitutions per site (K) can be estimated from observed differences (d)

Neanderthal + humans

- Coexisted with humans as late as 30,000 years ago in Europe and western Asia.
- Previous to the Noonan et al. study, little or no admixture, or human-neanderthal hybrids, was reported.
- However, for reasons discussed in the text, most of these results were based on mitochondrial DNA.



Extract

Clone

Small -scale sequencing

Batch culture

PCR inserts with vector primers

High-throughput capillary and pyrosequencing of pooled library

Direct selection of genomic and mtDNA targets *(Fig. 7, below)*

Neanderthal sequence in library vector

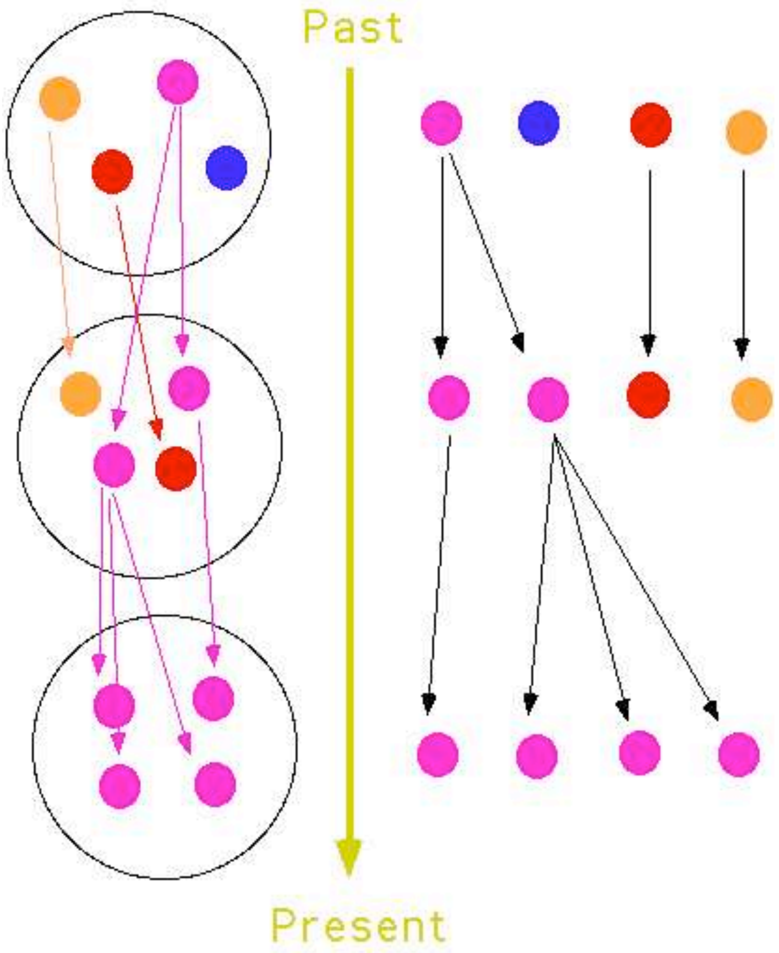
Other sequence in library vector

Noonan et al. (2006)

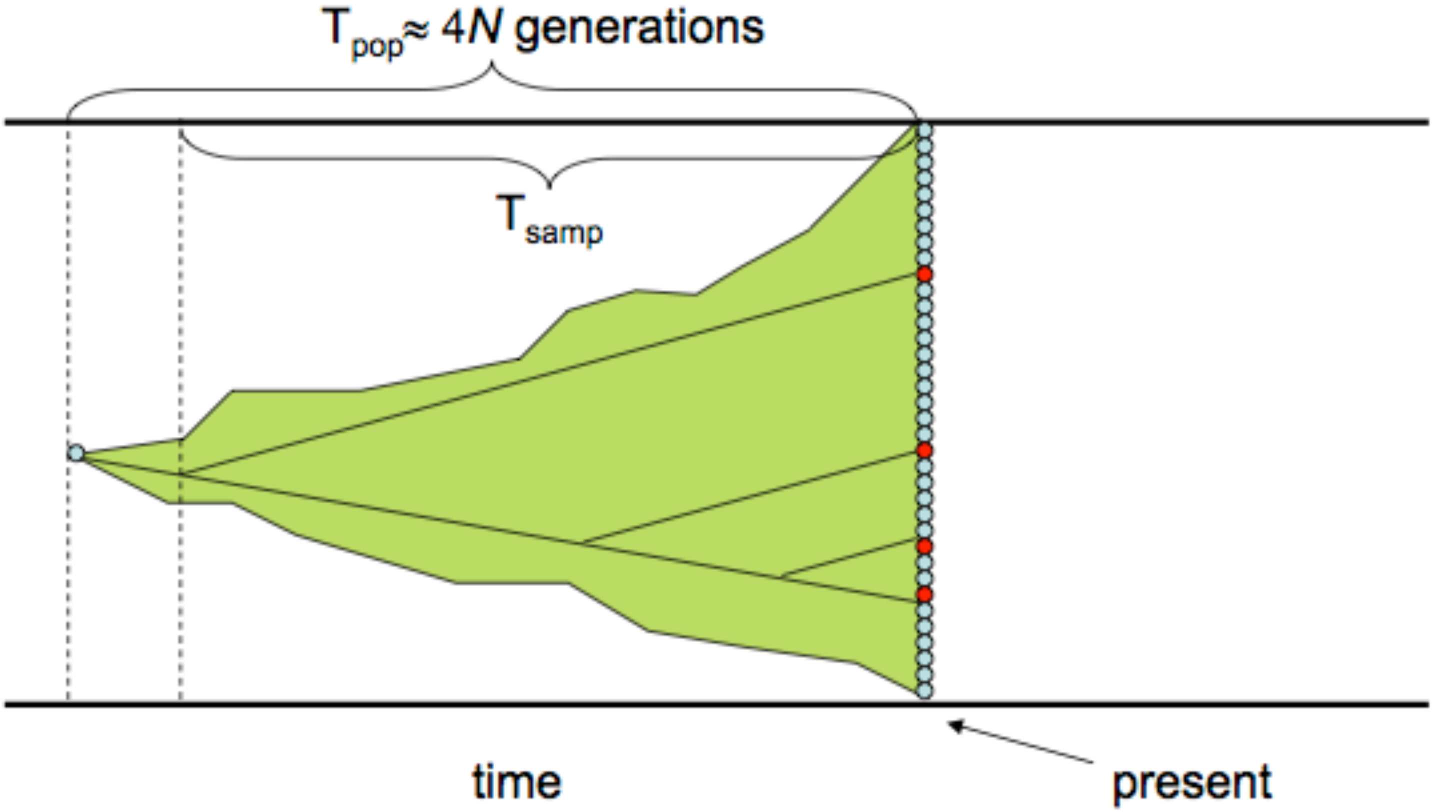
Coalescence time

- Based on statistical modeling, the most recent common ancestor existed 706,000 years ago

- 95% CI is 468,000 to 1,015,000 years



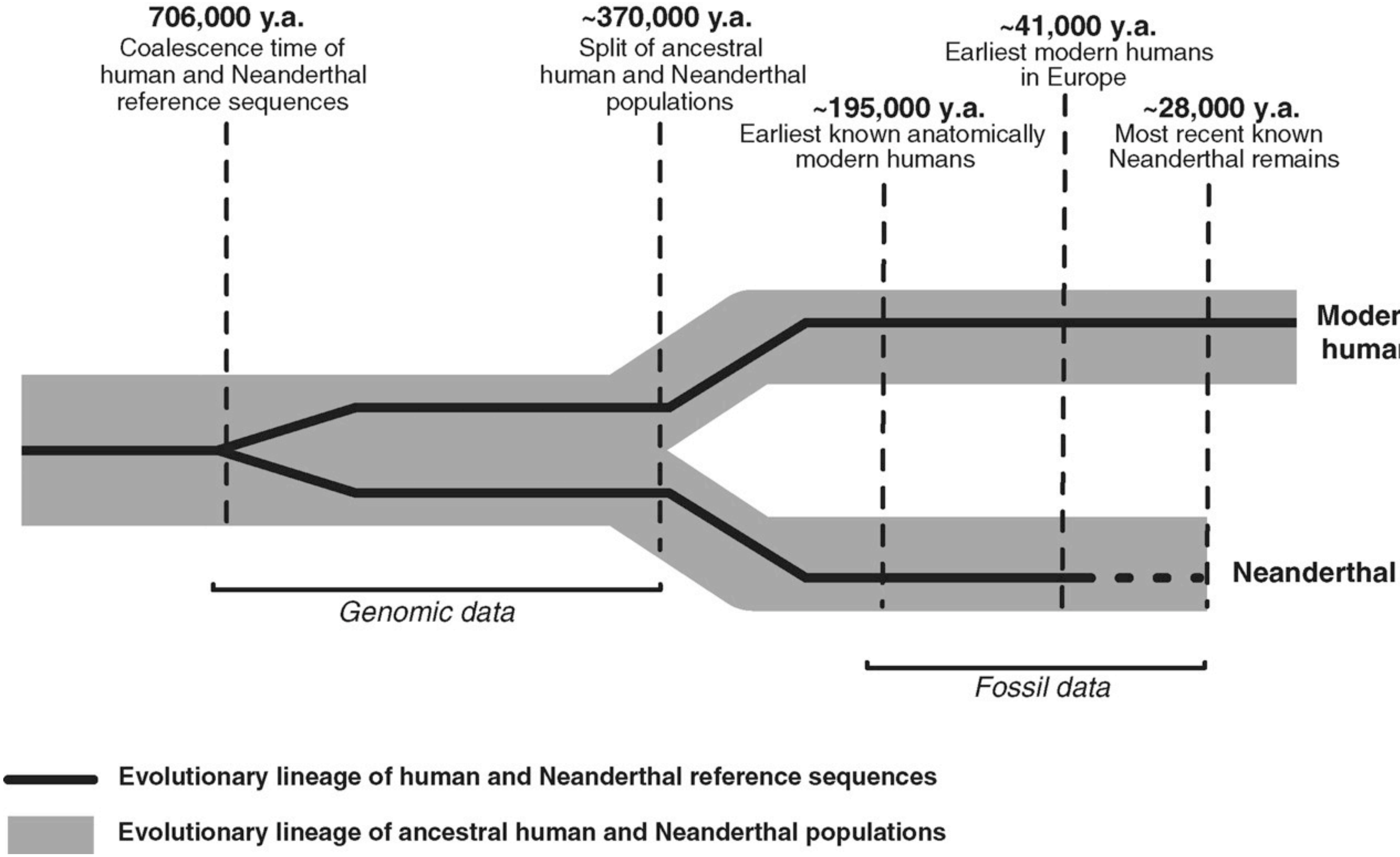
http://nitro.biosci.arizona.edu/courses/EEB320-2001/Lecture47/Lecture47.html


$T_{pop} \approx 4N$ generations
T_{samp}
time
present

Admixture results

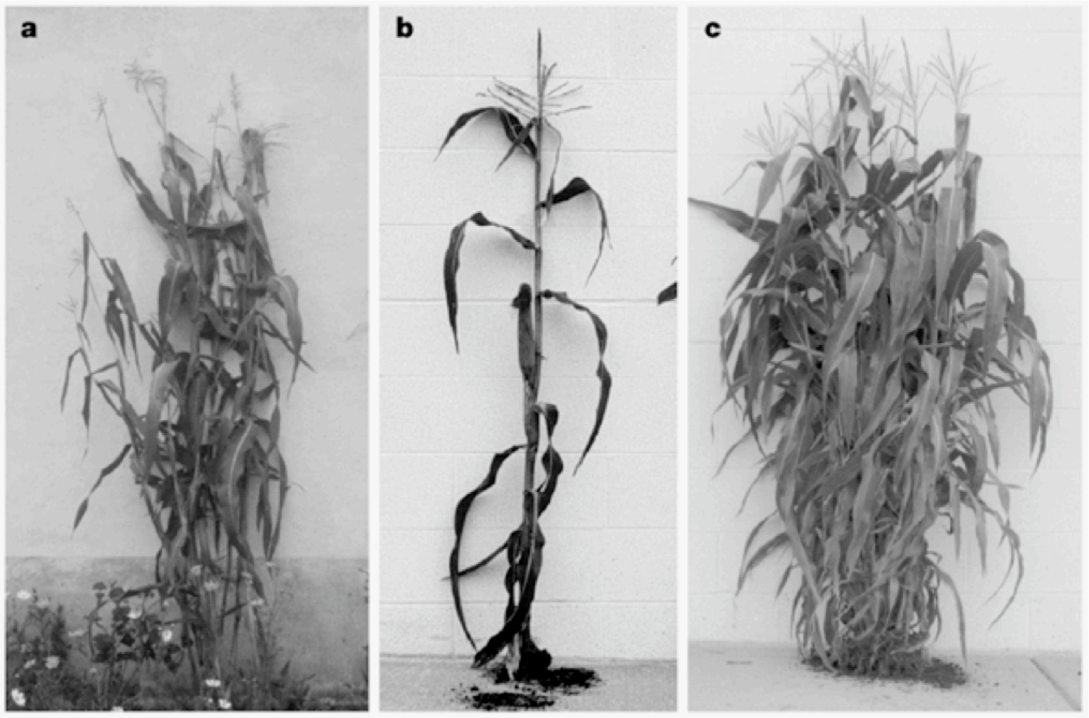
- Based on these ~1000 loci, the estimate is no admixture between modern humans and neanderthals (ca. 2006)

- However, statistically as much as 20% could be, providing a rationale for additional sequencing.


706,000 y.a.
Coalescence time of human and Neanderthal reference sequences
~370,000 y.a.
Split of ancestral human and Neanderthal populations
~41,000 y.a.
Earliest modern humans in Europe
~195,000 y.a.
Earliest known anatomically modern humans
~28,000 y.a.
Most recent known Neanderthal remains
Modern human
Neanderthal
Genomic data
Fossil data
Evolutionary lineage of human and Neanderthal reference sequences
Evolutionary lineage of ancestral human and Neanderthal populations

Noonan et al. (2006)

The evolution of maize (corn)



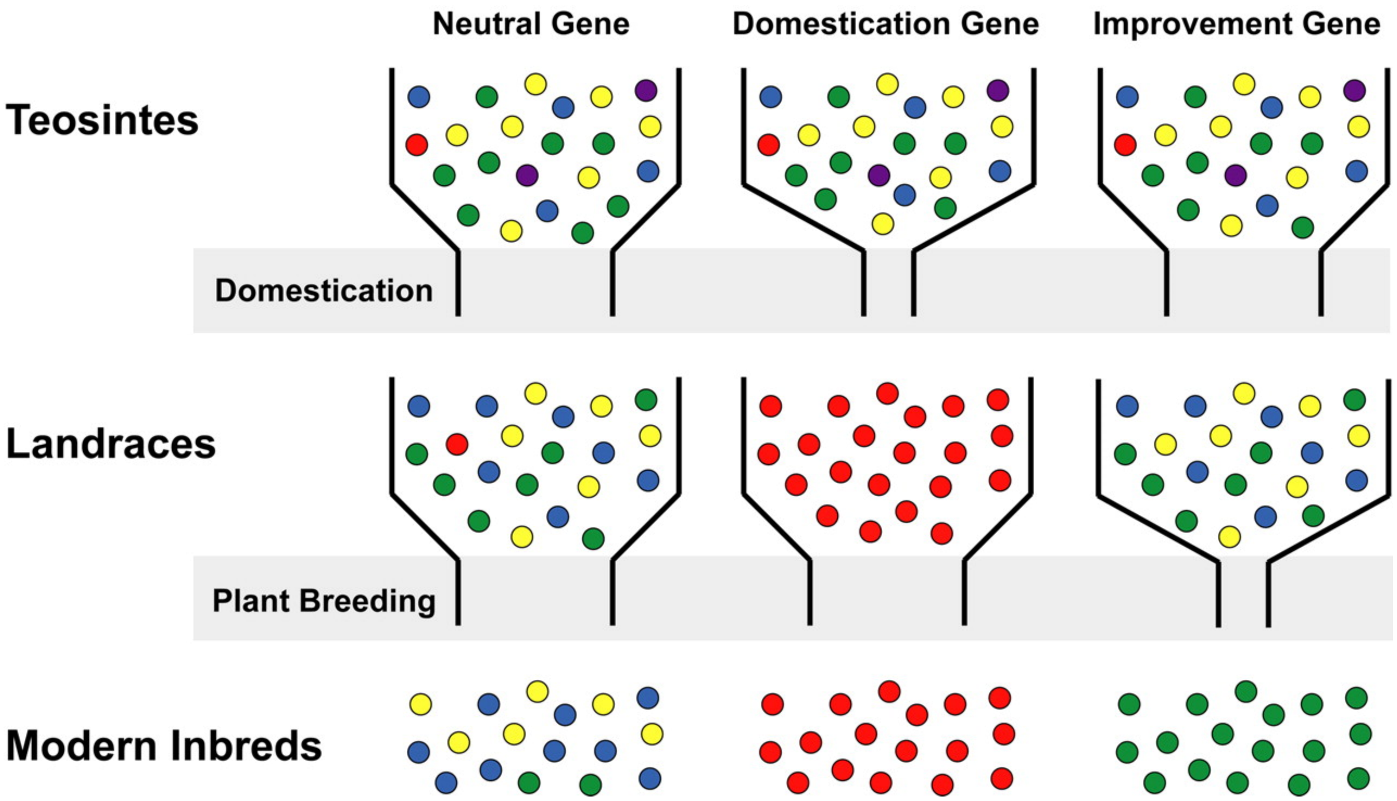
A: teosinte plant B: maize plant C: maize plant carrying mutation in *tb1*

Nature Reviews Genetics 2, 370-381 (May 2001)

Domestication in maize

- Yamasaki and colleagues looked at other 1000 genes in 14 diverse lines of corn.
- 35 of these had no differences over at least 200bp
- Additional testing (HKA test and coalescent simulation) predicted 6 domestication genes and 11 “improvement” genes.

Effect of Domestication and Plant Breeding on Genetic Diversity of Maize Genes



Neutral Gene

Domestication Gene

Improvement Gene

Teosintes

Domestication

Landraces

Plant Breeding

Modern Inbreds

Yamasaki, M., et al. Plant Cell 2005;17:2859-2872

Tests and their value

- HKA test
  - Compares diversity of gene of interest to one that is neutral
- Coalescent simulation
  - Looks at reduction on diversity relative to estimated demographic history
  - Relies on sophisticated models