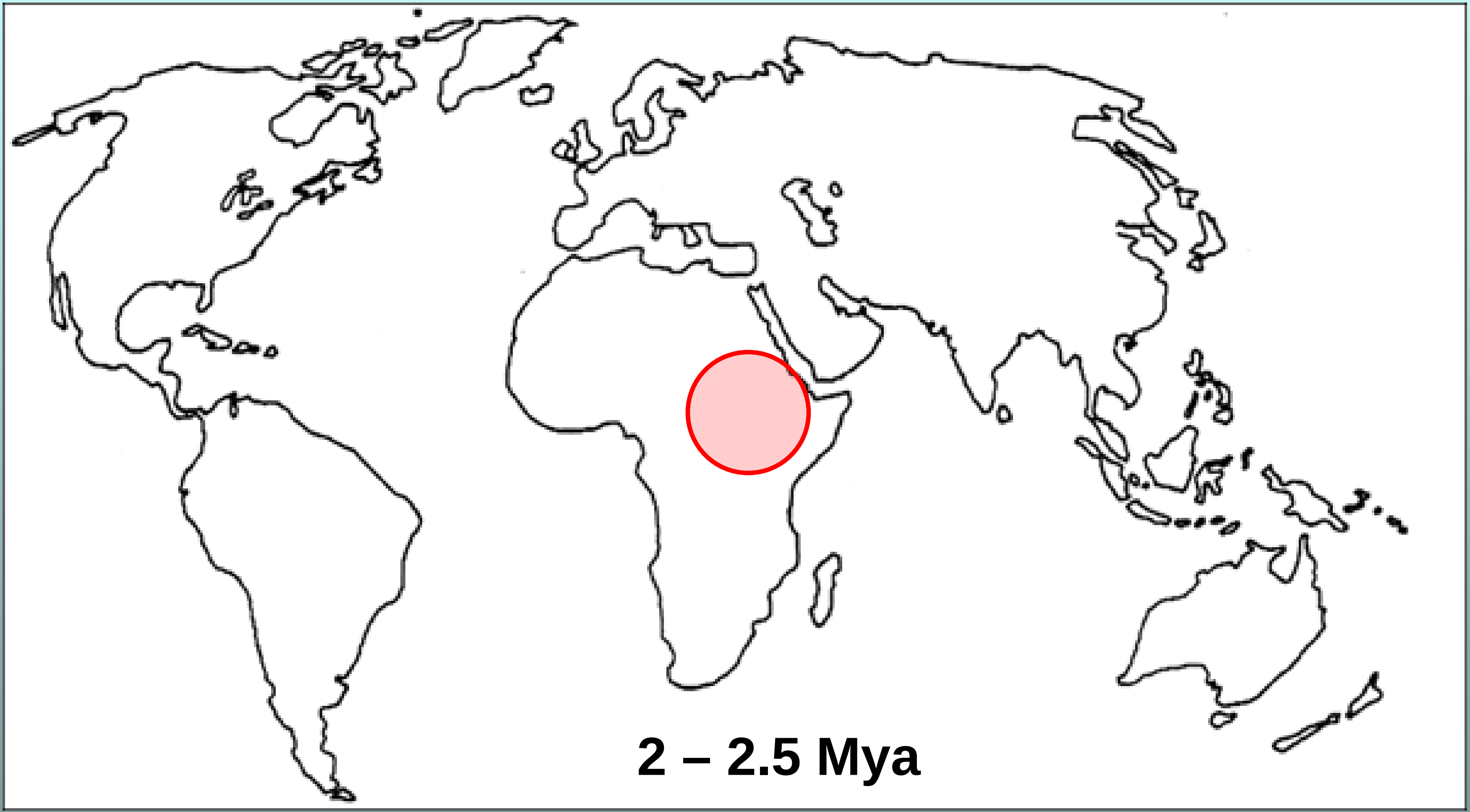
Detecting ancient admixture using DNA sequence data

October 10, 2008
Jeff Wall

Institute for Human Genetics
UCSF

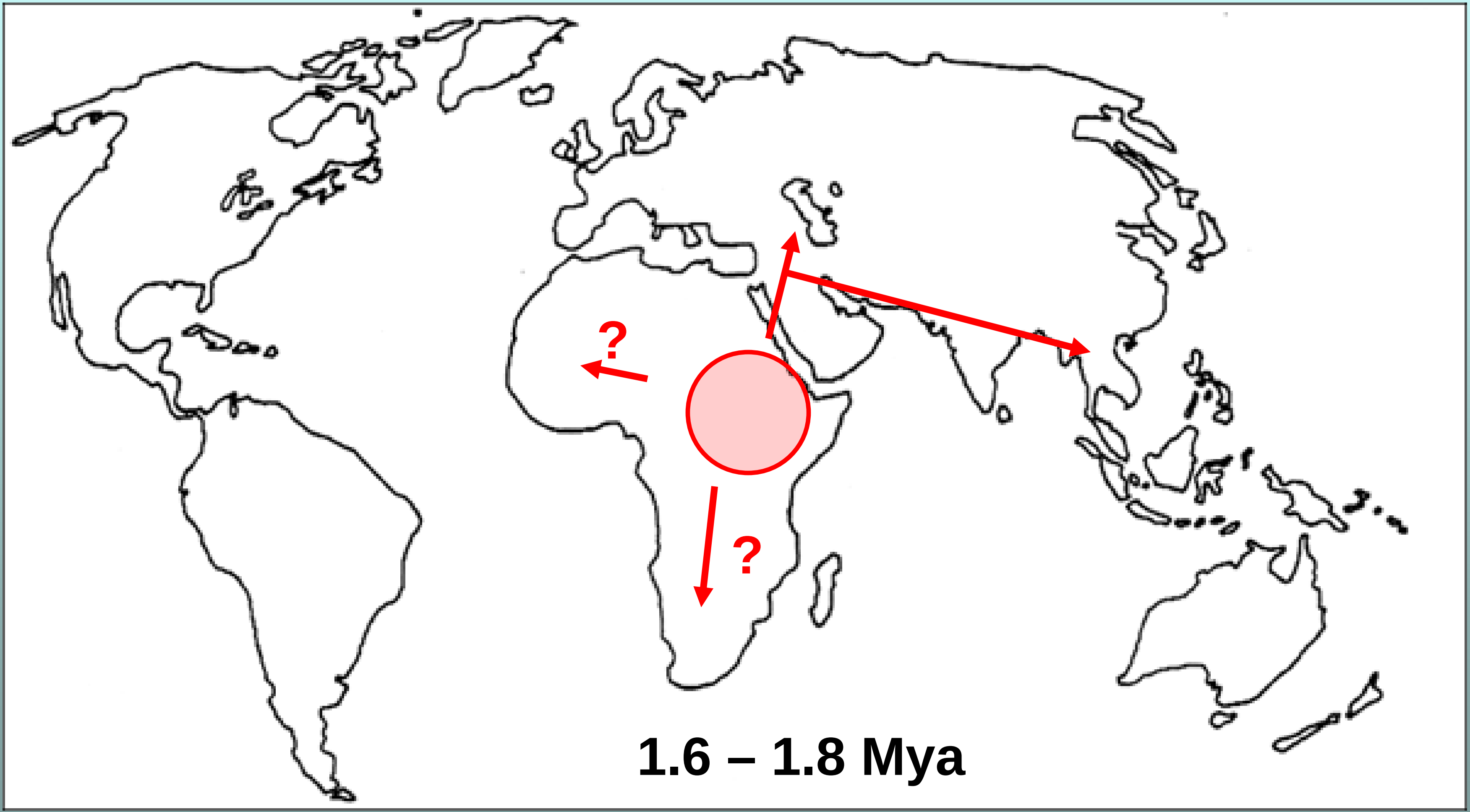
Background

Origin of genus *Homo*

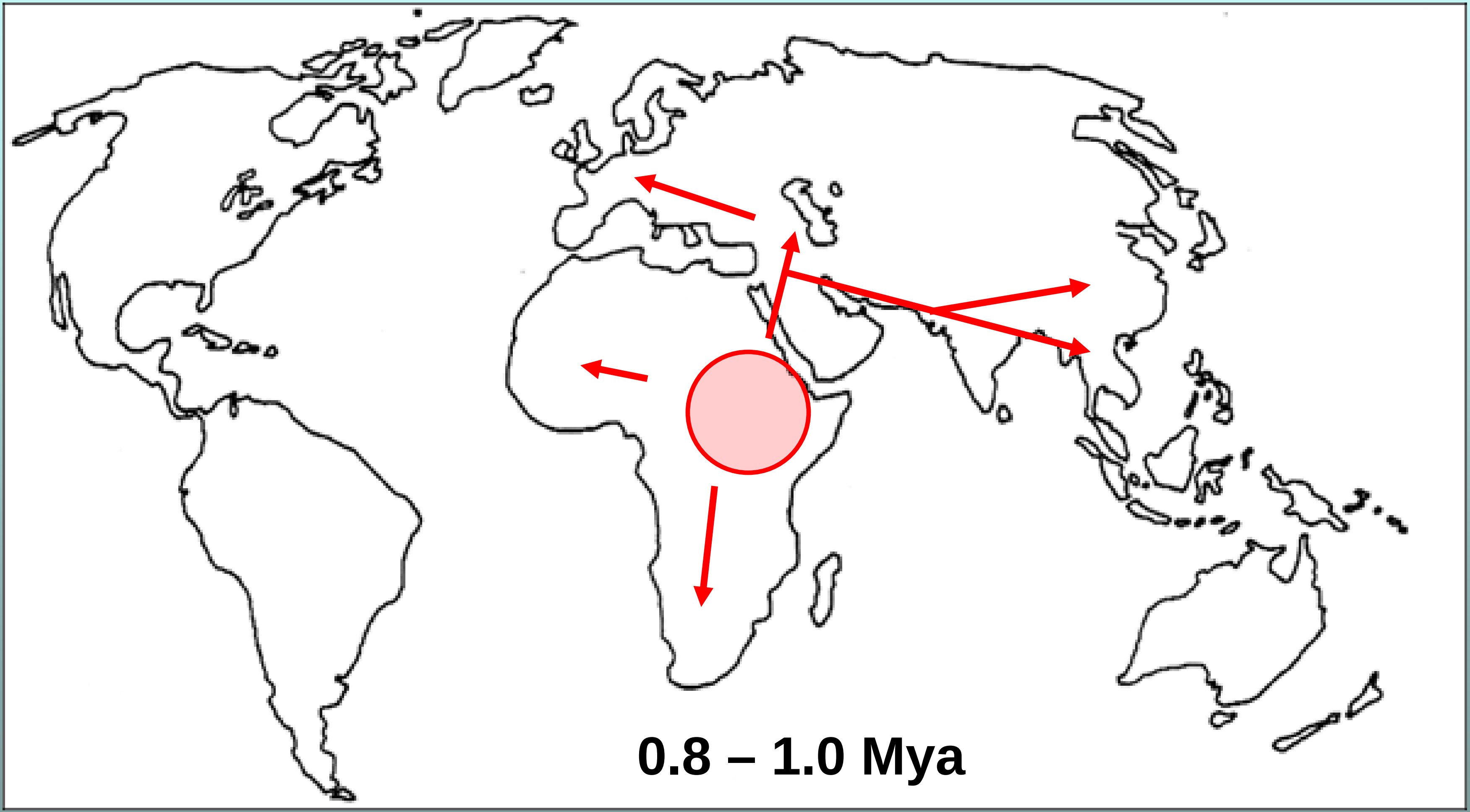


2 – 2.5 Mya

Out of Africa (part I)

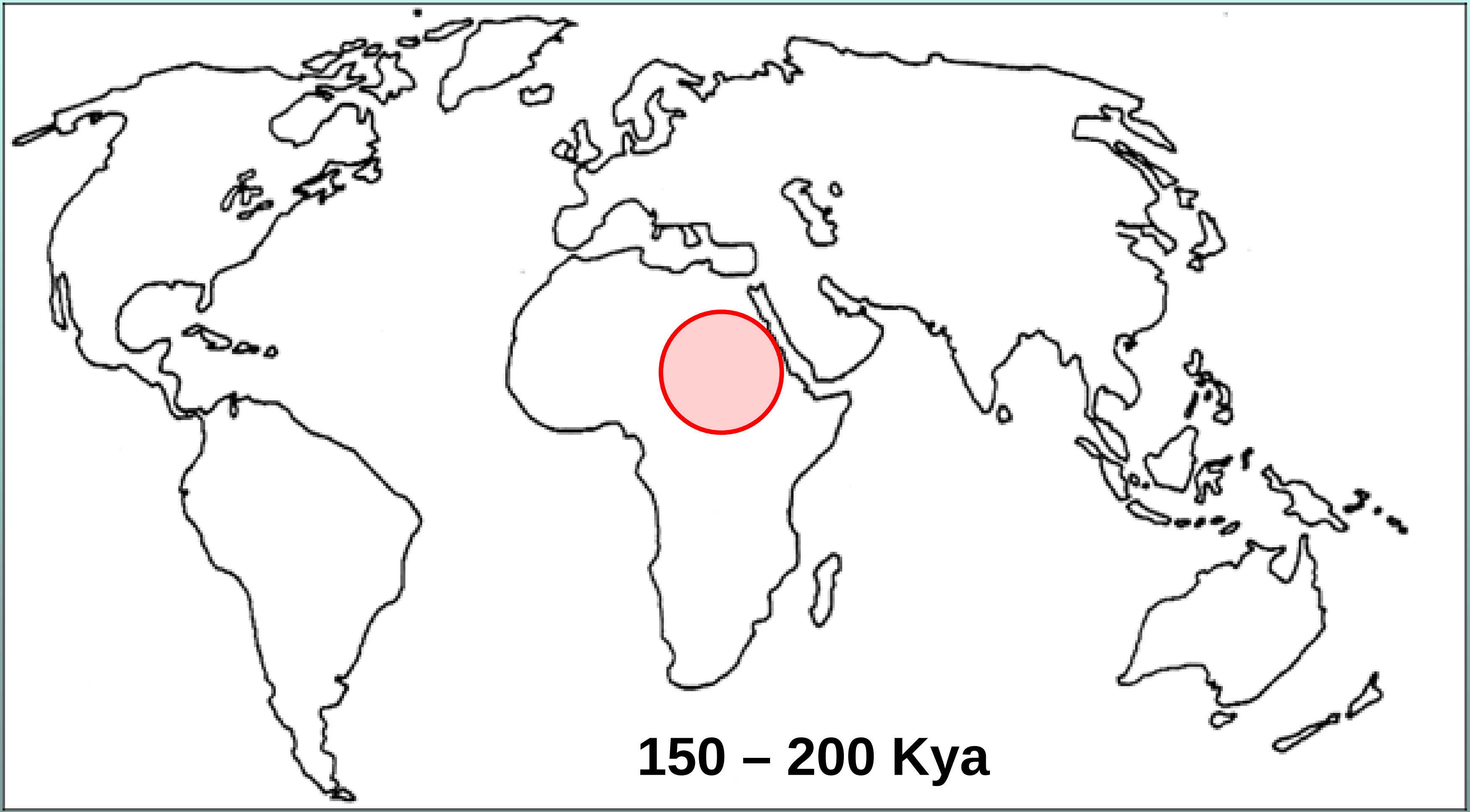


Further spread of *Homo erectus*



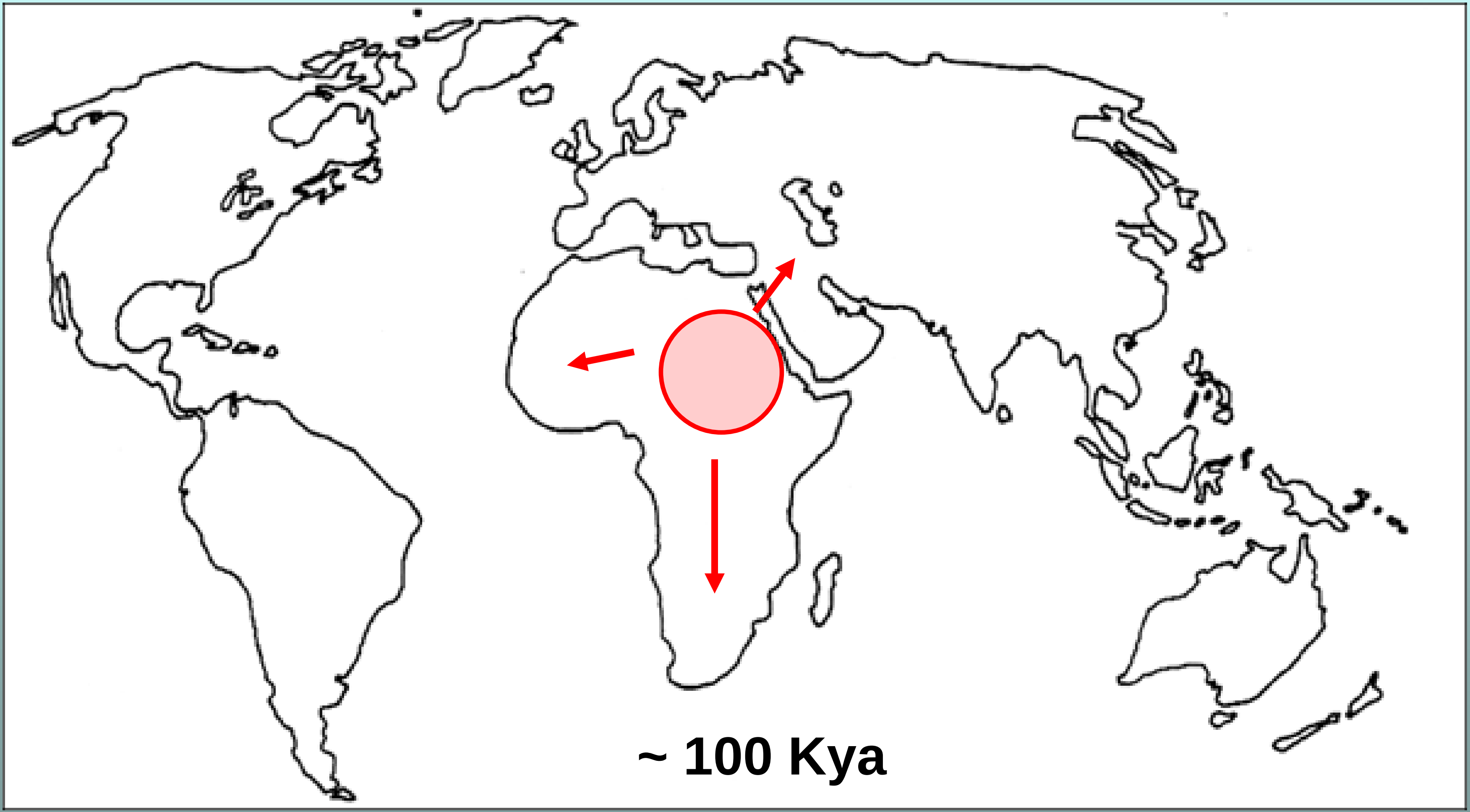
0.8 – 1.0 Mya

Origin of “modern” humans

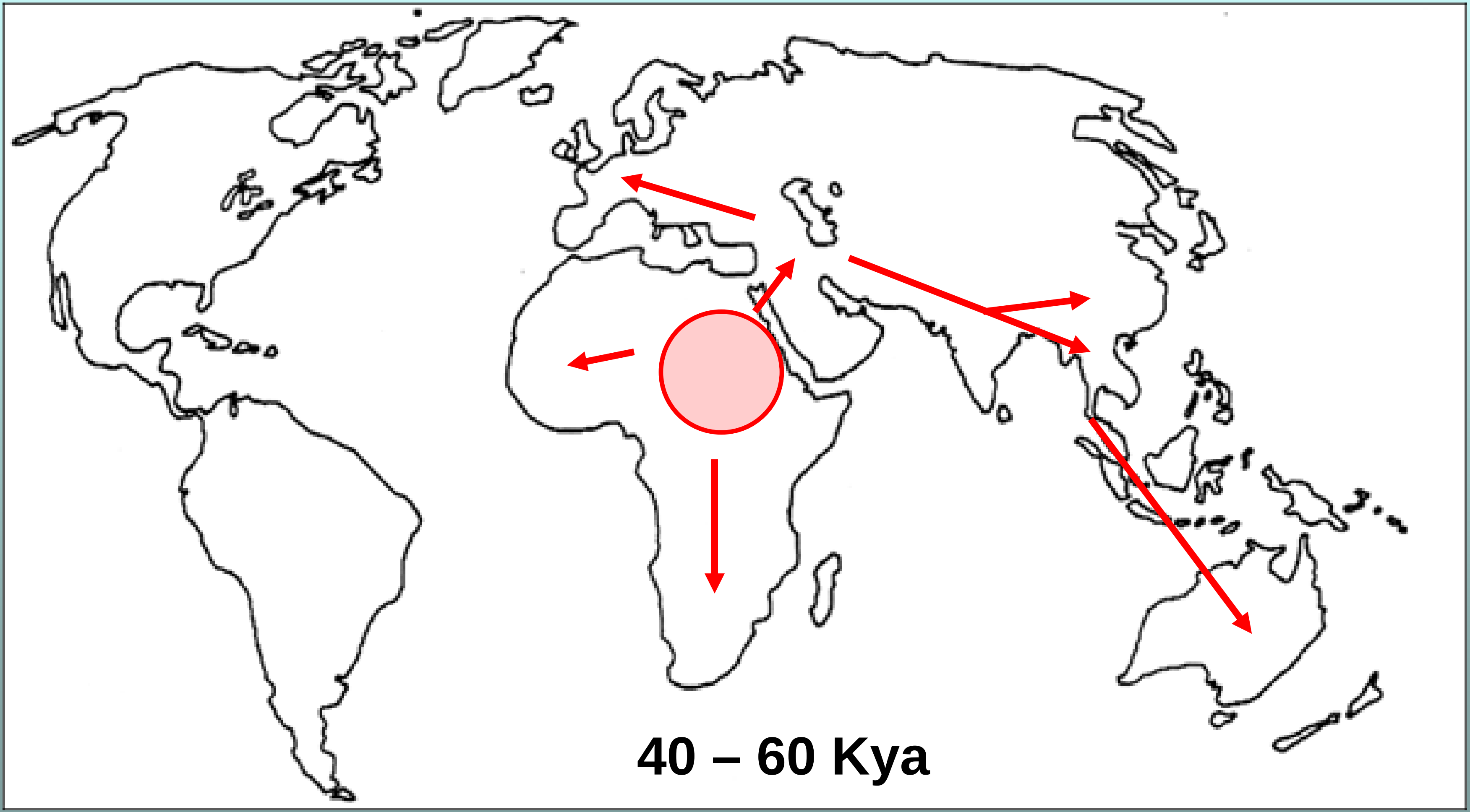


150 – 200 Kya

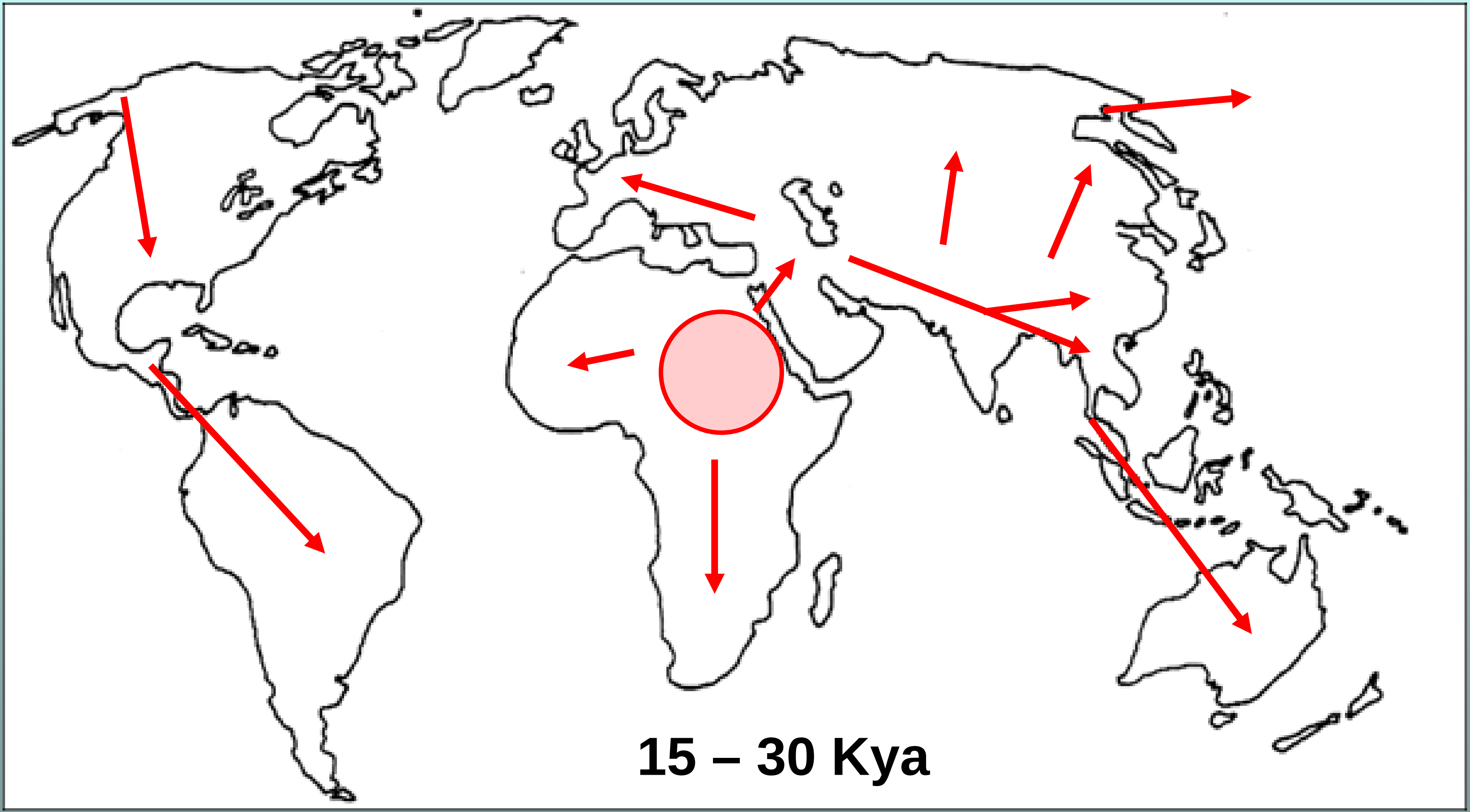
Out of Africa (part II)



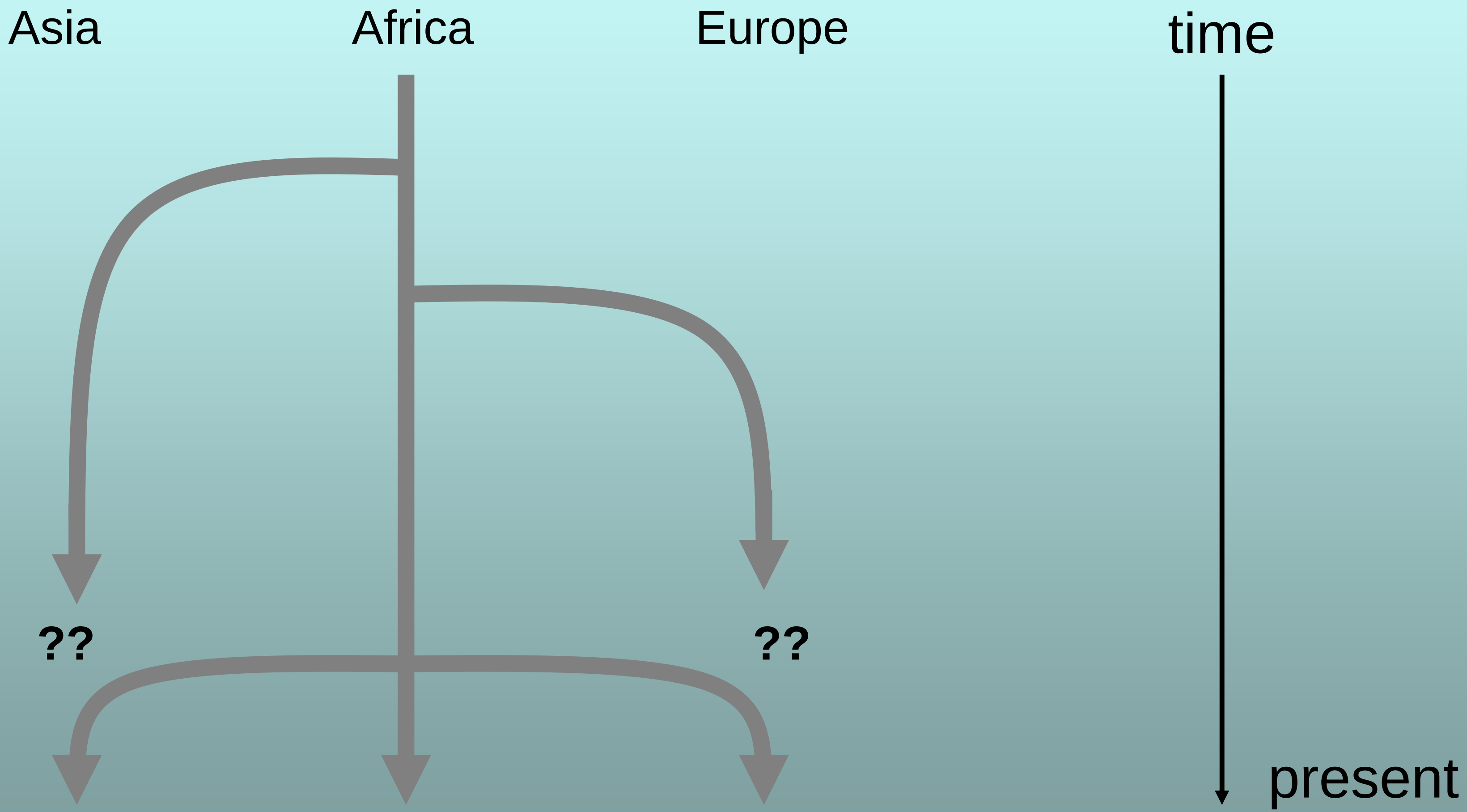
Re-colonization of Eurasia



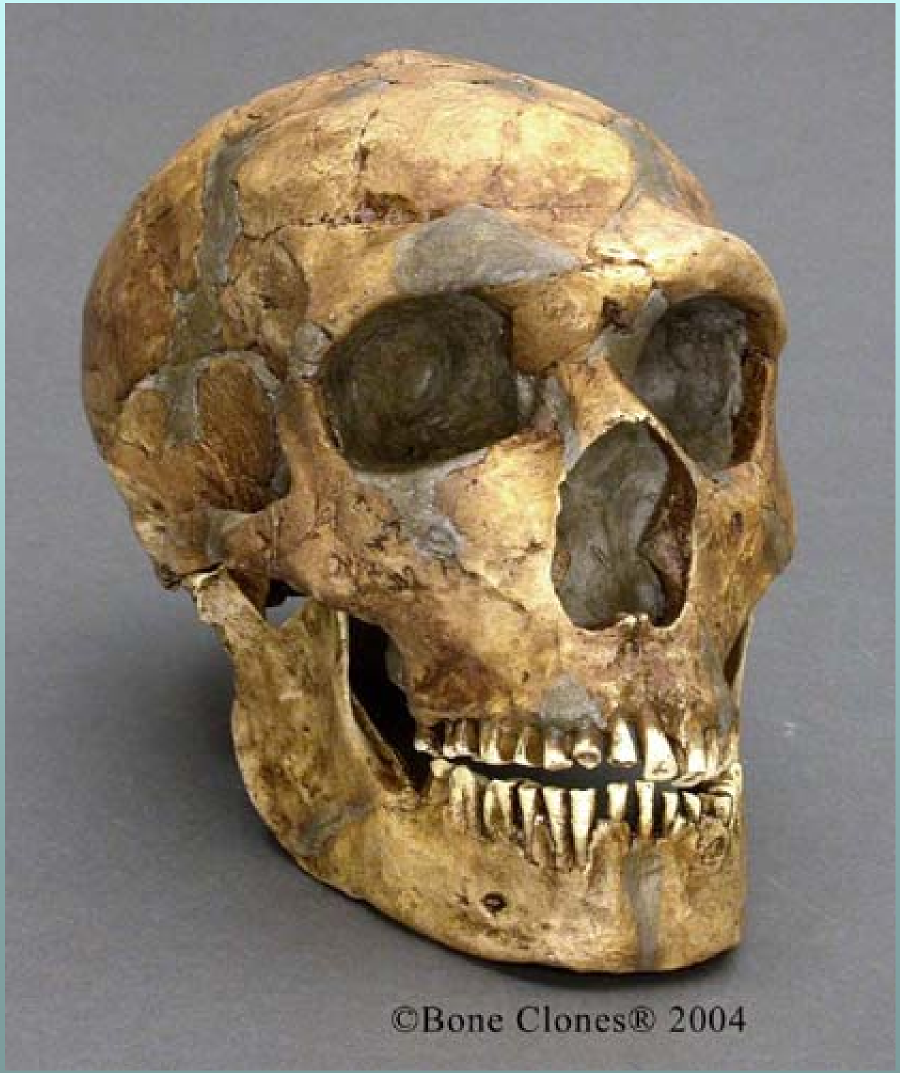
Colonization of the New World



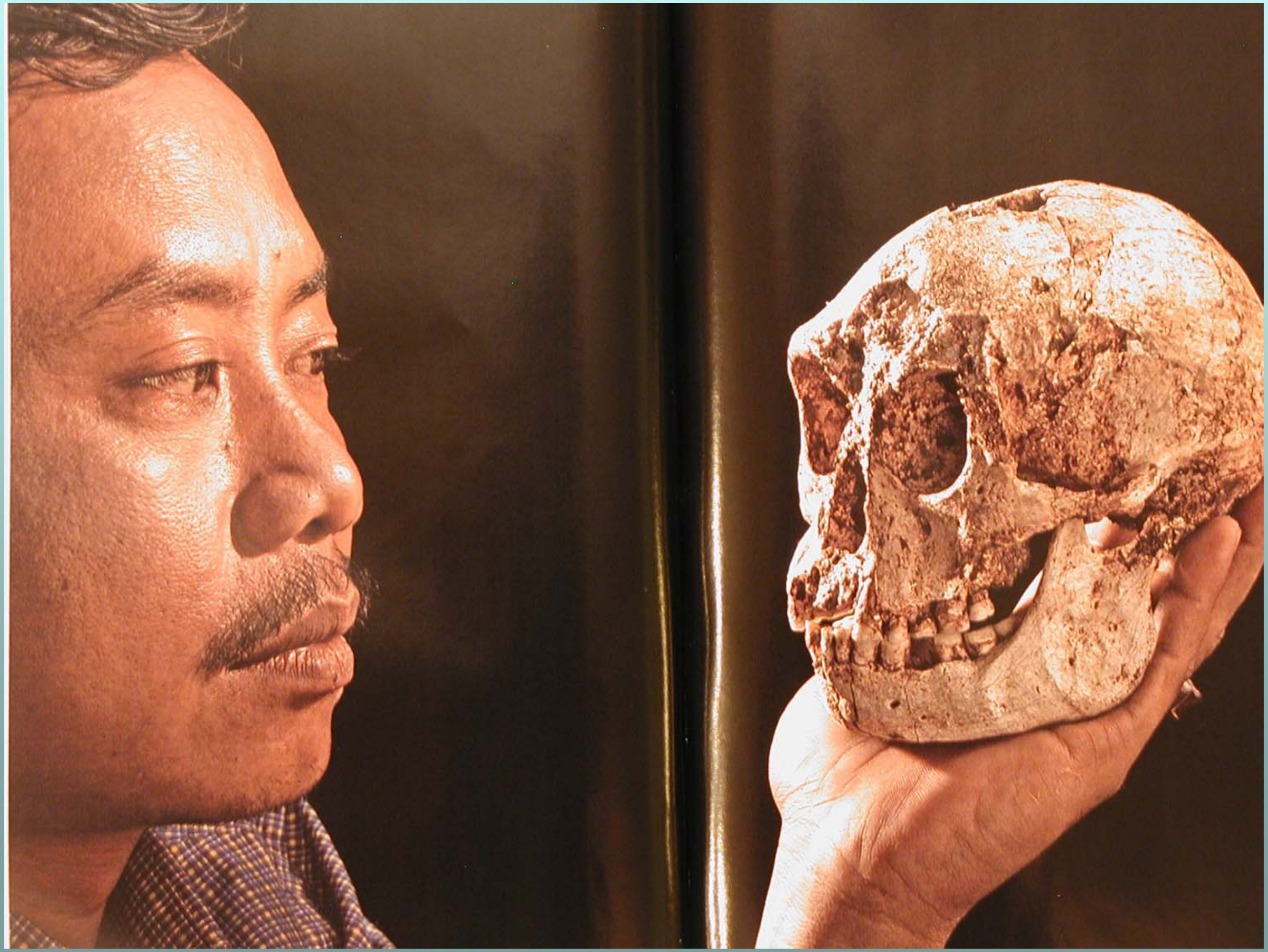
Modeling human demography



Neandertal skull


©Bone Clones® 2004

The “hobbit” (*Homo floresiensis*)



Modern Human Origins

- What contribution (if any) did archaic humans (e.g., Neanderthals or *Homo floresiensis*) make to the modern gene pool?
- How can we answer this from patterns of DNA sequence variation?

Two approaches

- Indirect approach
  - Examine patterns of genetic variation in extant humans to look for evidence of ancient admixture
- Direct approach
  - Compare recently recovered Neandertal (or other archaic human) DNA sequences with orthologous modern human sequences

Two approaches

- Indirect approach
  - Examine patterns of genetic variation in extant humans to look for evidence of ancient admixture

- **Direct approach**
  - Compare recently recovered Neandertal (or other archaic human) DNA sequences with orthologous modern human sequences

Neandertal mtDNA

Mitochondrial DNA from several different Neandertal fossils have been recovered.

All Neandertal mtDNA sequences are quite different from (and outside the range of) modern human mtDNA variation.

These results are consistent with admixture rates (i.e., contribution of Neandertals to the modern European gene pool) of 0 – 20 %.

Serre et al. (2004)

Neandertal nuclear DNA

Recent studies have also recovered Neandertal DNA from random nuclear regions (Green et al. 2006; Noonan et al. 2006) and targeted genes (Krause et al. 2007; Lalueza-Fox et al. 2007).

There is even a ‘Neandertal genome project’, which seeks to obtain a complete (though composite) Neandertal genome sequence.

Neandertal nuclear DNA

However, contamination with modern human DNA is a serious concern, and so far 2 out of 4 existing studies have been effectively debunked (Wall and Kim 2007; Coop et al. 2008).

Analyses of existing nuclear data are also consistent with Neandertal admixture rates of 0 – 20 % (Noonan et al. 2006).

Neanderthal sequencing

Two recent studies (Green et al. Nature 2006; Noonan et al. Science 2006) published nuclear DNA sequences obtained from a 38,000 year old Neandertal fossil.

However, the two studies came to completely different conclusions!

Comparing the two studies

We used the methods described in Noonan et al. (2006) to analyze the two data sets.

This ensures that the results are directly comparable.

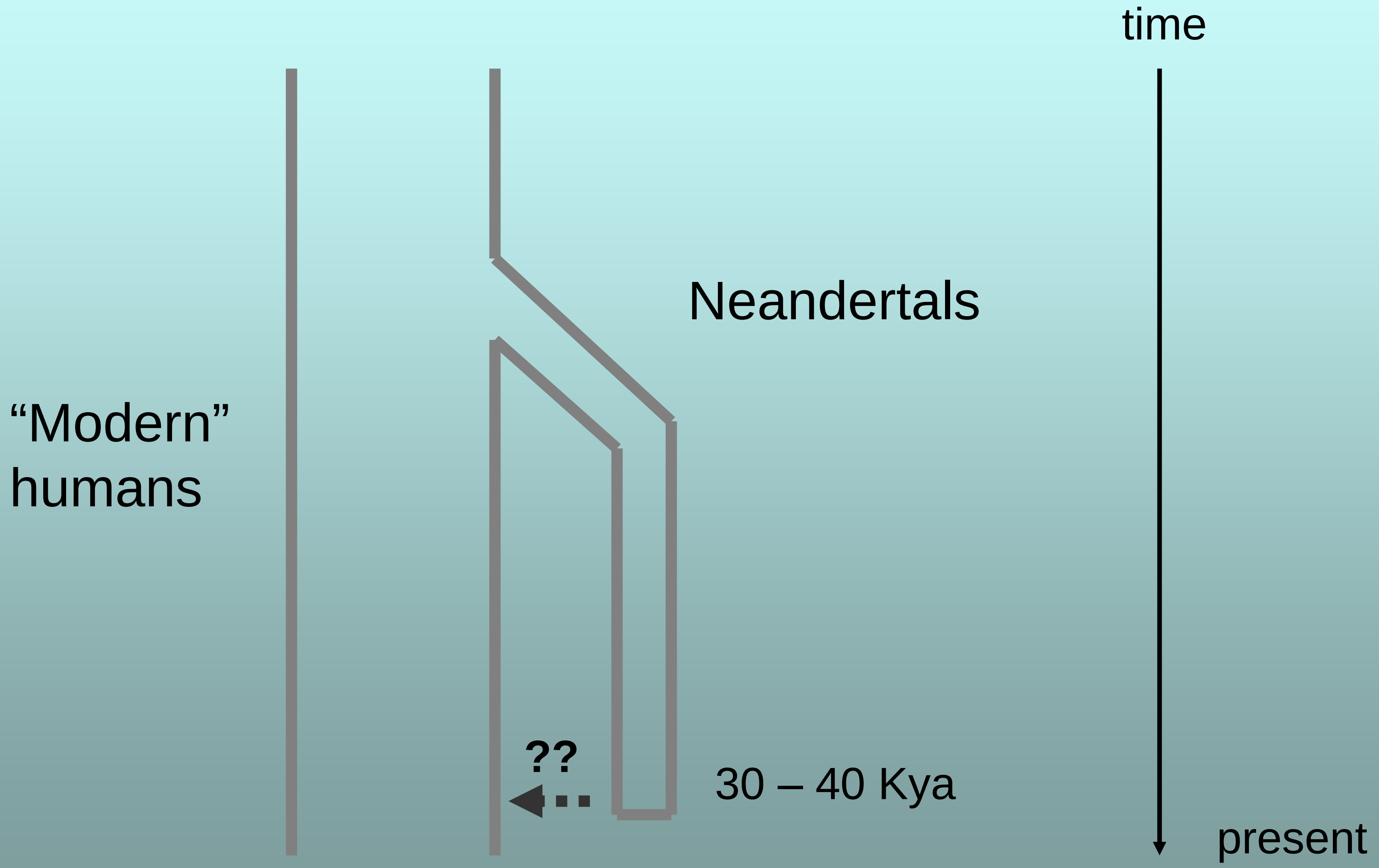
Wall & Kim (2007), PLoS Genetics

The data

For all sites where the Neandertal and human sequence differ, we tabulate

- Neandertal allele (ancestral vs. derived)
- Human reference sequence allele (ancestral vs. derived)
- Frequency of derived allele in Caucasian HapMap sample (if available)

Simple admixture model



Methods overview

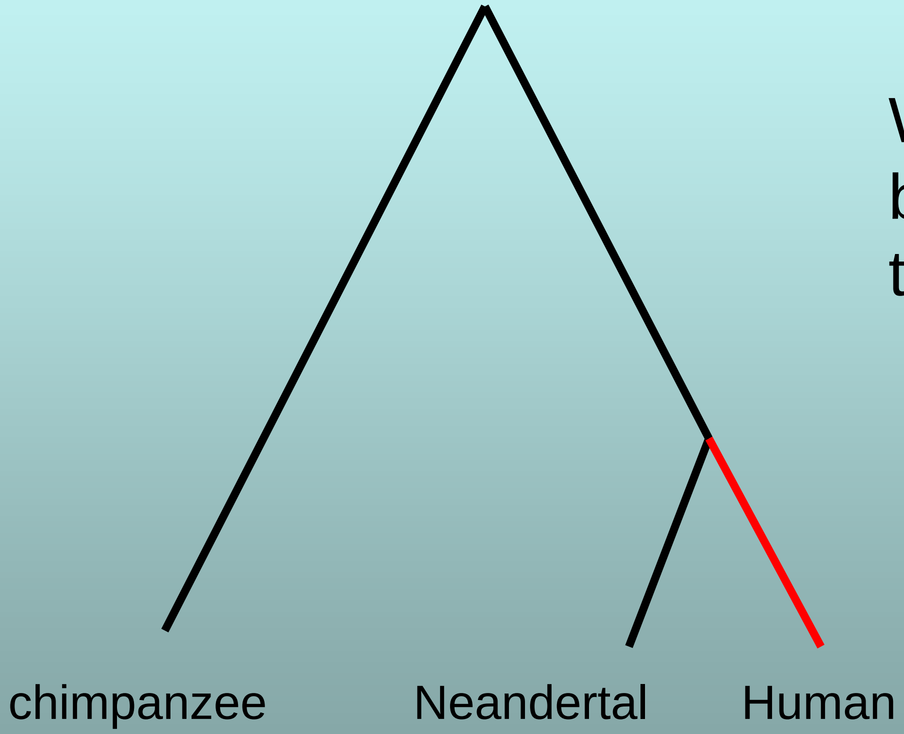
The likelihood of each particular configuration depends on model parameters, including the admixture proportion c.

We use a composite likelihood (assuming informative sites are independent) to estimate parameters.

Conflicting results

	Noonan et al. (2006)	Green et al. (2006)
Human-Neandertal DNA sequence divergence time	706 Kya (466 – 1028 Kya)	560 Kya (509 – 615 Kya)
Modern European-Neandertal population split time	325 Kya (135 – 557 Kya)	35 Kya (33 – 51 Kya)
Neandertal contribution to modern European ancestry	0 % (0 – 39%)	94% (81 – 100%)

Human branch divergence



What is the human-specific branch divergence for the two data sets?

Human branch divergence

Data	Human branch div. (sites / Kb)
Noonan	7.43
Green	5.89

Human branch divergence

Data	Human branch div. (sites / Kb)
Noonan	7.43
Green	5.89
▪ Small fragments	7.32
▪ Med. fragments	5.97
▪ Large fragments	5.22

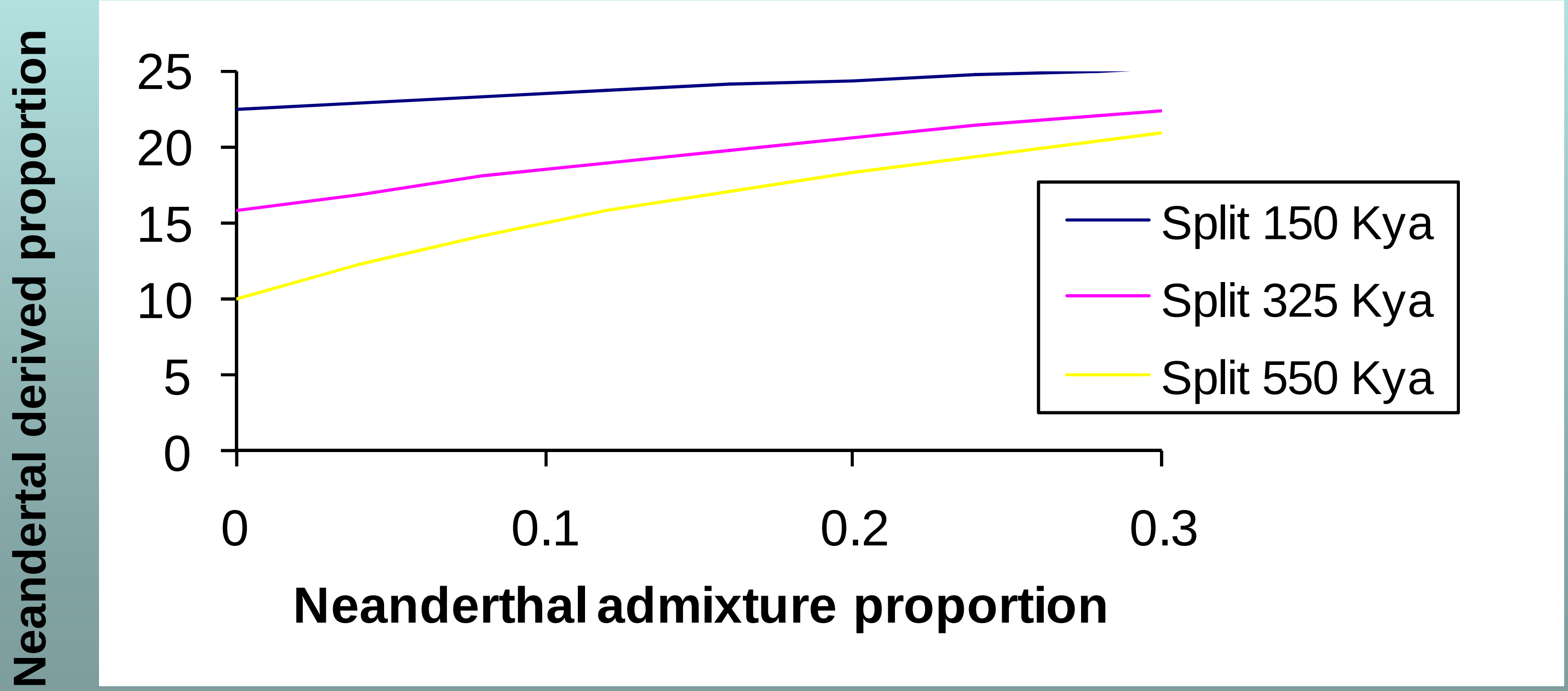
Human branch divergence

Data	Human branch div. (sites / Kb)
Noonan	7.43
Green	5.89
▪ Small fragments	7.32
▪ Med. fragments	5.97
▪ Large fragments	5.22
Human data	3.95 – 6.10

Neandertal derived alleles

(simulations)

We also tabulated the fraction of HapMap SNPs for which the Neandertal sequence has the derived allele:



Neandertal derived alleles

(data)

We also tabulated the fraction of HapMap SNPs for which the Neandertal sequence has the derived allele:

Data:	**Proportion**
Noonan	3.1
Green	32.9
▪ Small fragments	21.8
▪ Med. fragments	32.7
▪ Large fragments	37.2
Human ref. sequence	37.0

Conclusions

The most likely explanation is widespread contamination with modern human DNA in the Green et al. (2006) study.

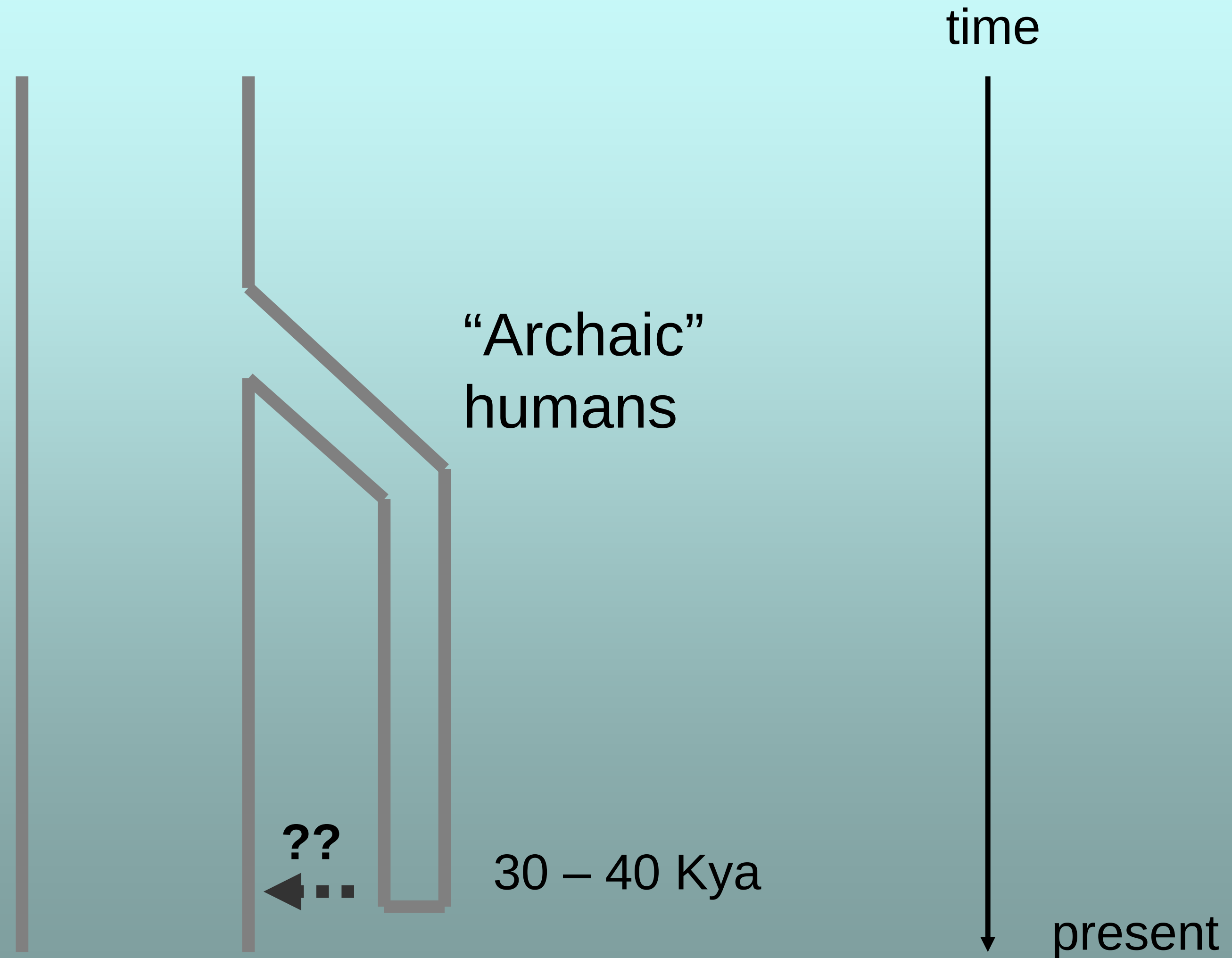
Two approaches

- **Indirect approach**
  - Examine patterns of genetic variation in extant humans to look for evidence of ancient admixture
- Direct approach
  - Compare recently recovered Neanderthal DNA sequences with orthologous modern human sequences

General overview

- Start with resequencing data from multiple human populations
- Use frequency spectrum data to estimate demographic parameters (assuming no admixture)
- Examine model fit using summaries of linkage disequilibrium (LD)
- Use a specially constructed measure of LD to estimate levels of ancient admixture

Simple admixture model

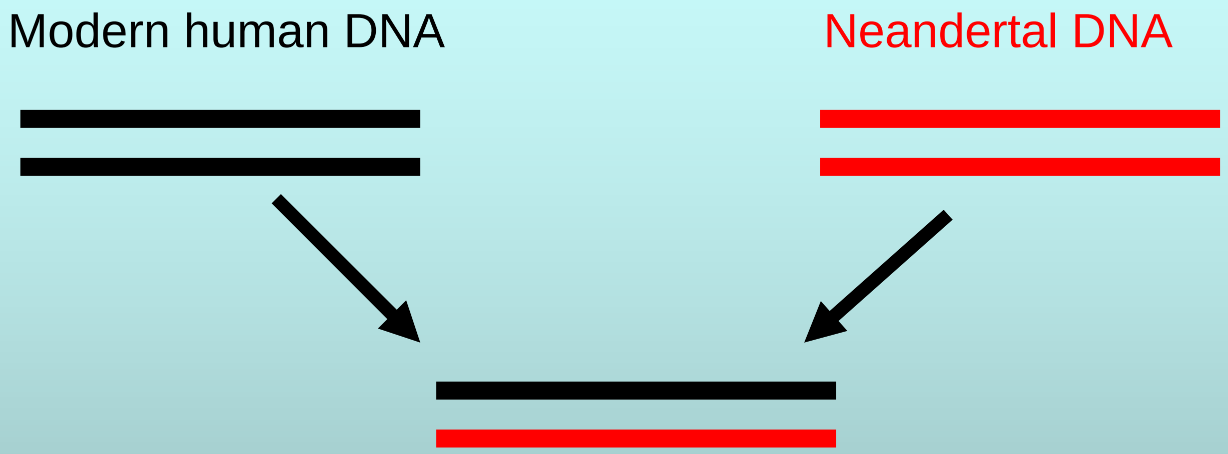


Admixture mapping

Modern human DNA

Neandertal DNA

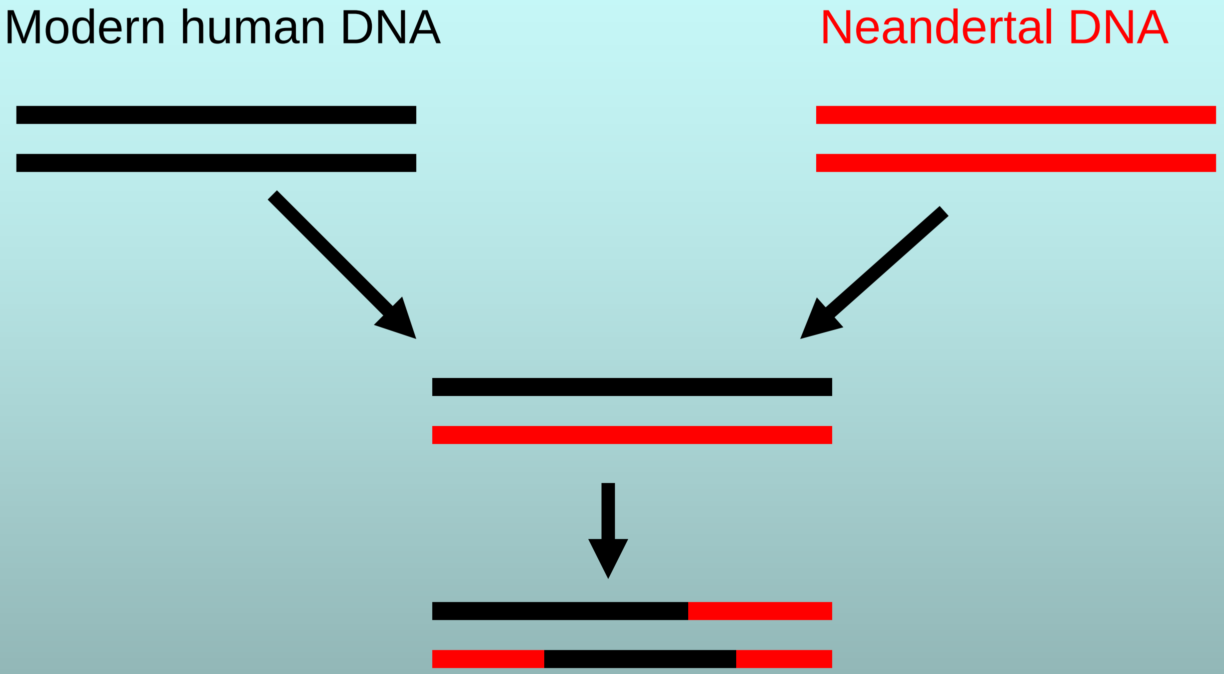
Admixture mapping



Modern human DNA

Neandertal DNA

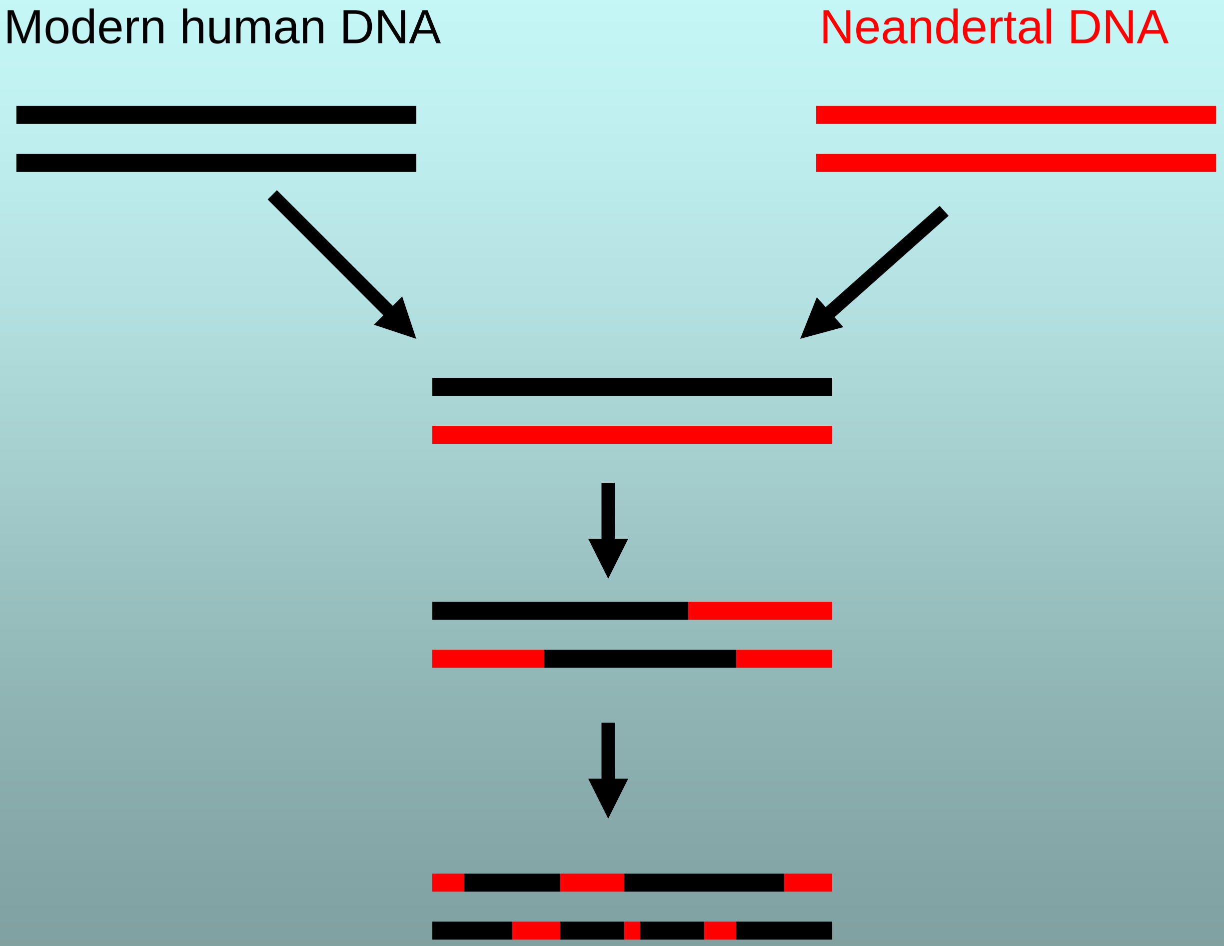
Admixture mapping



Modern human DNA

Neandertal DNA

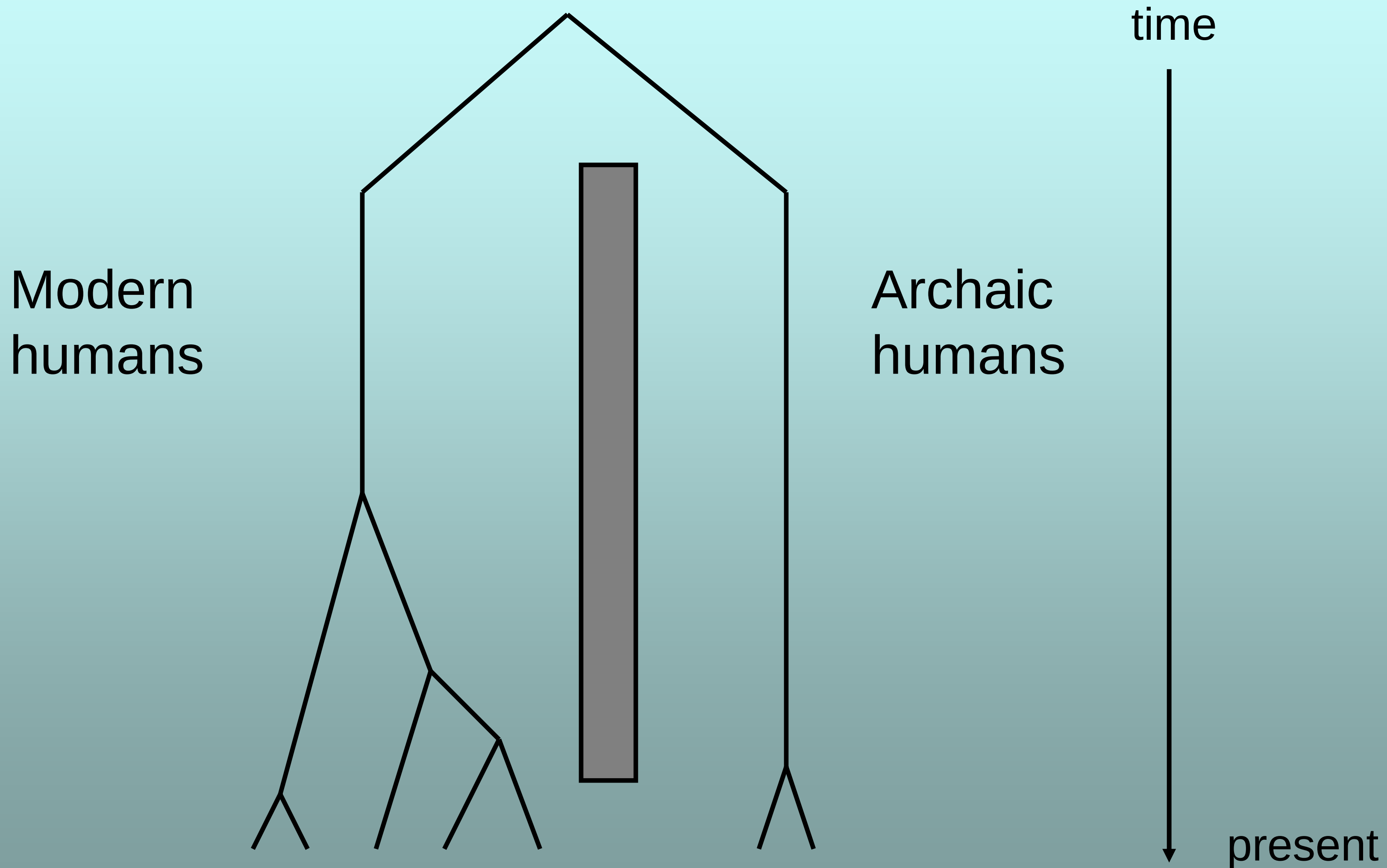
Admixture mapping



Modern human DNA

Neandertal DNA

Genealogy with archaic ancestry



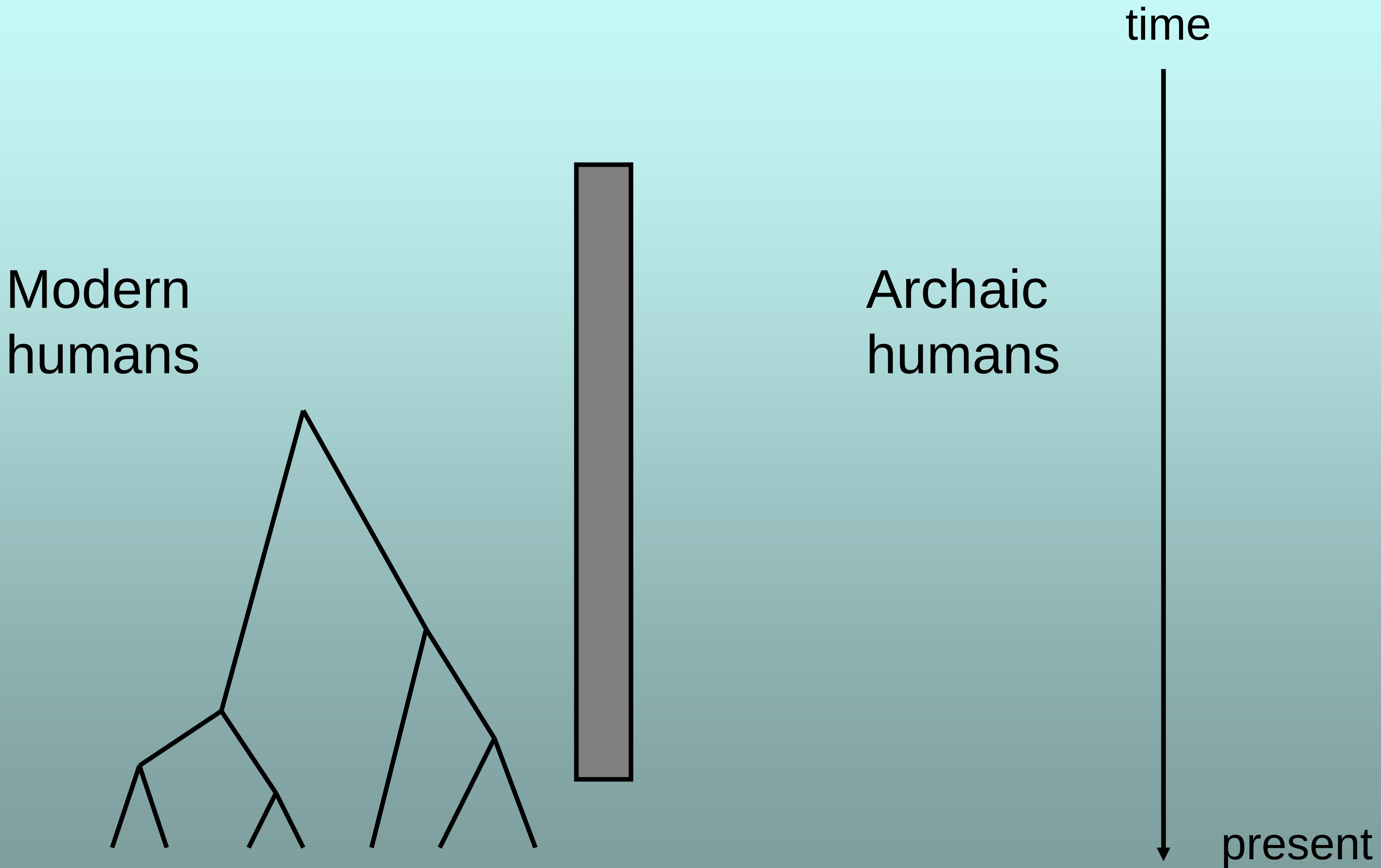
time

Modern
humans

Archaic
humans

present

Genealogy without archaic ancestry



time

Modern
humans

Archaic
humans

present

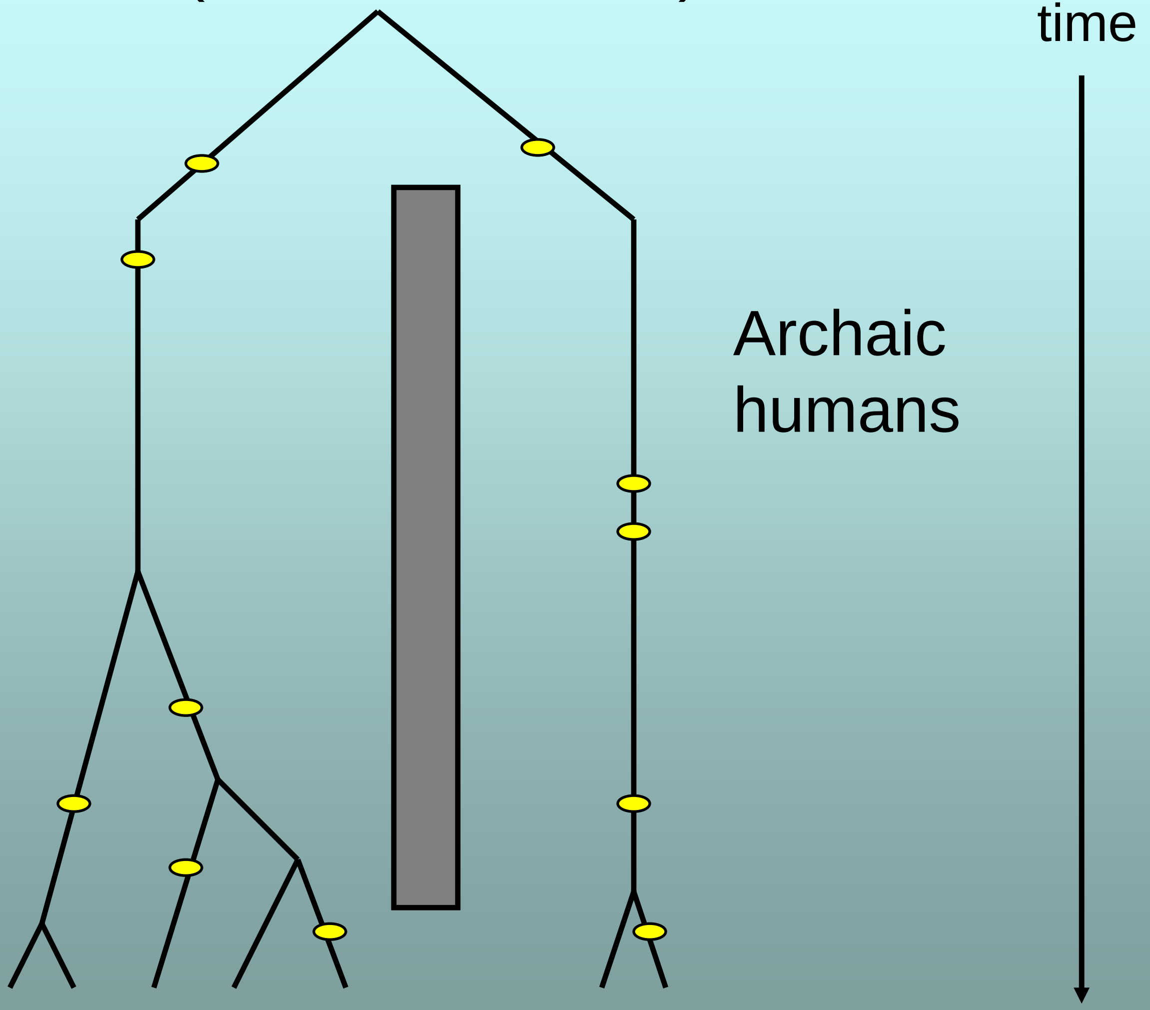
Our main questions

- What pattern does archaic ancestry produce in DNA sequence polymorphism data (from extant humans)?

- How can we use data to
  - estimate the contribution of archaic humans to the modern gene pool (c)?
  - test whether $c > 0$?

Genealogy with archaic ancestry

(Mutations added)



time

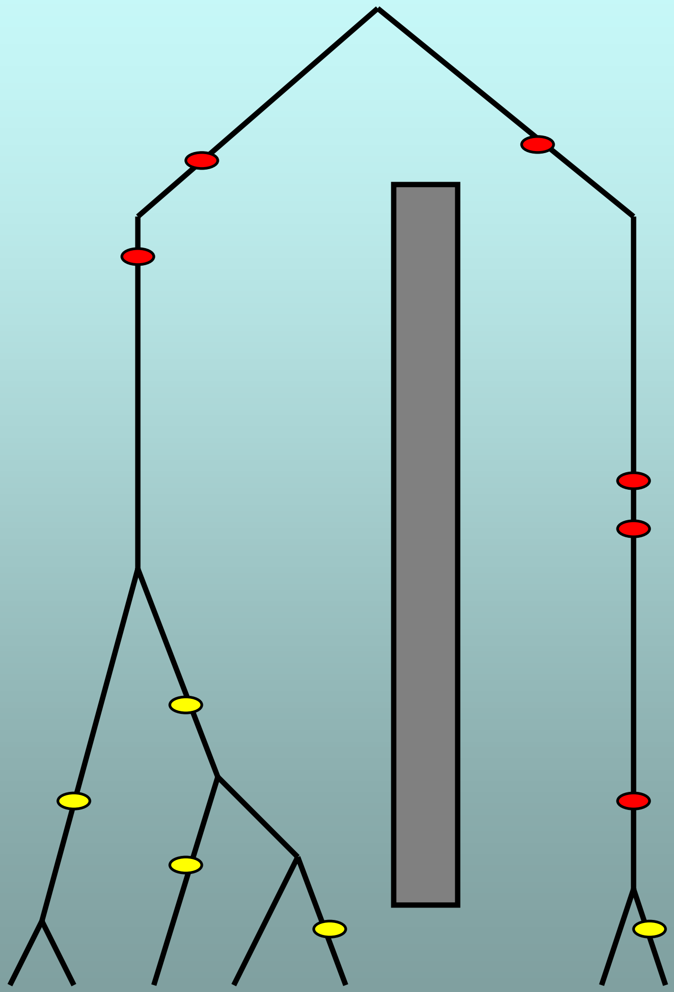
Modern
humans

Archaic
humans

present

Genealogy with archaic ancestry

(Mutations added)



time

Modern humans

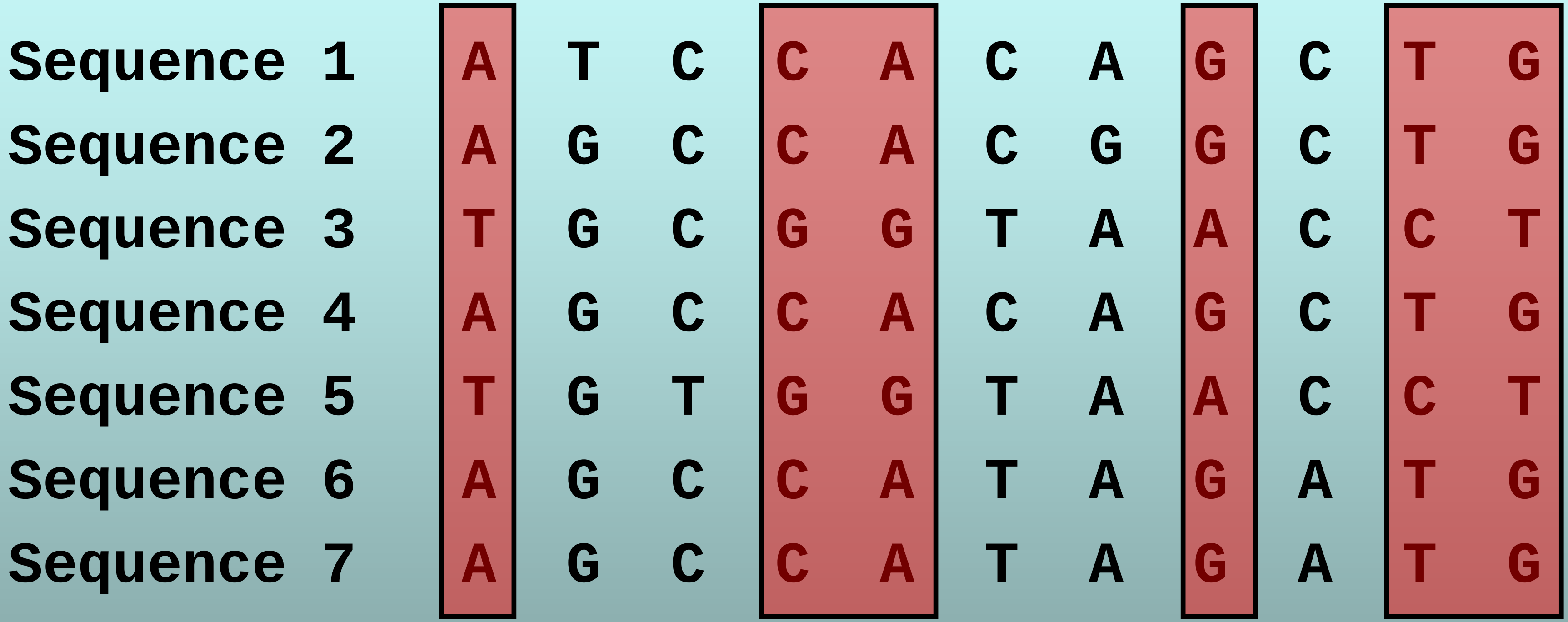
Archaic humans

present

Patterns in DNA sequence data

Sequence 1	**A**	**T**	**C**	**C**	**A**	**C**	**A**	**G**	**C**	**T**	**G**
Sequence 2	**A**	**G**	**C**	**C**	**A**	**C**	**G**	**G**	**C**	**T**	**G**
Sequence 3	**T**	**G**	**C**	**G**	**G**	**T**	**A**	**A**	**C**	**C**	**T**
Sequence 4	**A**	**G**	**C**	**C**	**A**	**C**	**A**	**G**	**C**	**T**	**G**
Sequence 5	**T**	**G**	**T**	**G**	**G**	**T**	**A**	**A**	**C**	**C**	**T**
Sequence 6	**A**	**G**	**C**	**C**	**A**	**T**	**A**	**G**	**A**	**T**	**G**
Sequence 7	**A**	**G**	**C**	**C**	**A**	**T**	**A**	**G**	**A**	**T**	**G**

Patterns in DNA sequence data



Sequence 1	A	T	C	C	A	C	A	G	C	T	G
Sequence 2	A	G	C	C	A	C	G	G	C	T	G
Sequence 3	T	G	C	G	G	T	A	A	C	C	T
Sequence 4	A	G	C	C	A	C	A	G	C	T	G
Sequence 5	T	G	T	G	G	T	A	A	C	C	T
Sequence 6	A	G	C	C	A	T	A	G	A	T	G
Sequence 7	A	G	C	C	A	T	A	G	A	T	G

We call the sites in red **congruent** sites – these are sites inferred to be on the same branch of an unrooted tree.

Measuring ‘congruence’

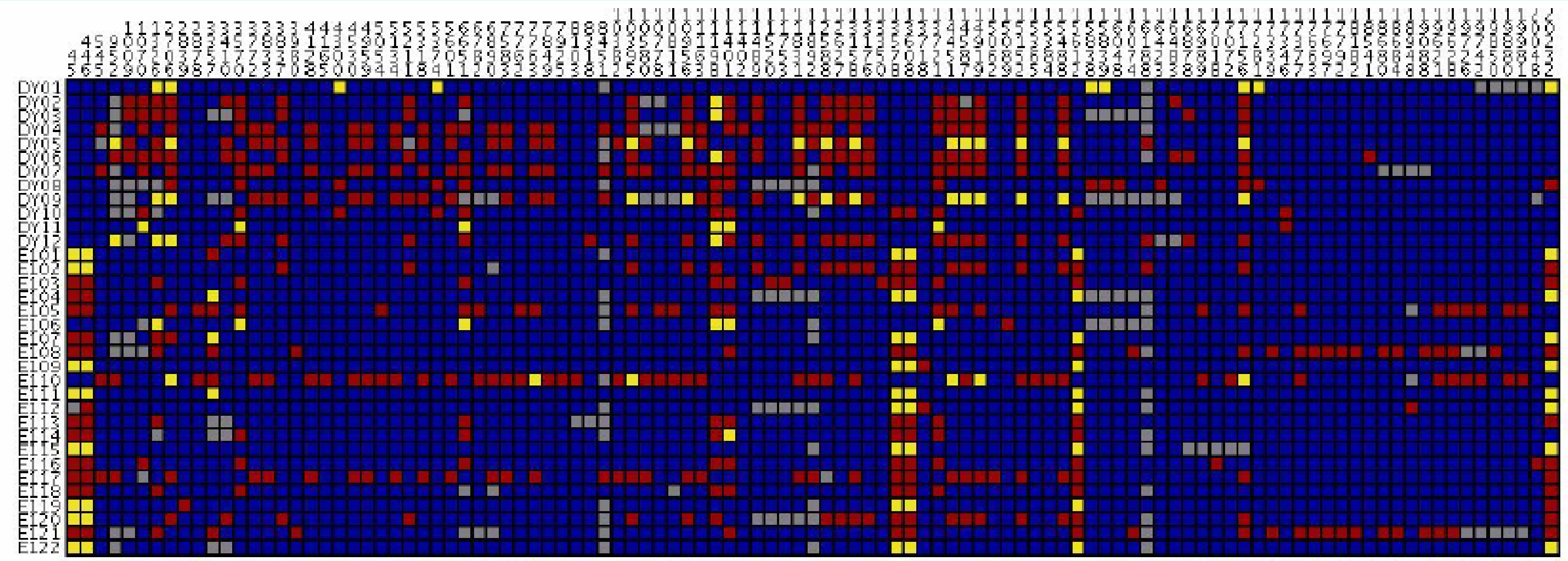
To measure the level of ‘congruence’ in SNP data from larger regions we define a score function

$$S^* = \max_{I \subset \{1,2,\ldots n\}} S(I)$$

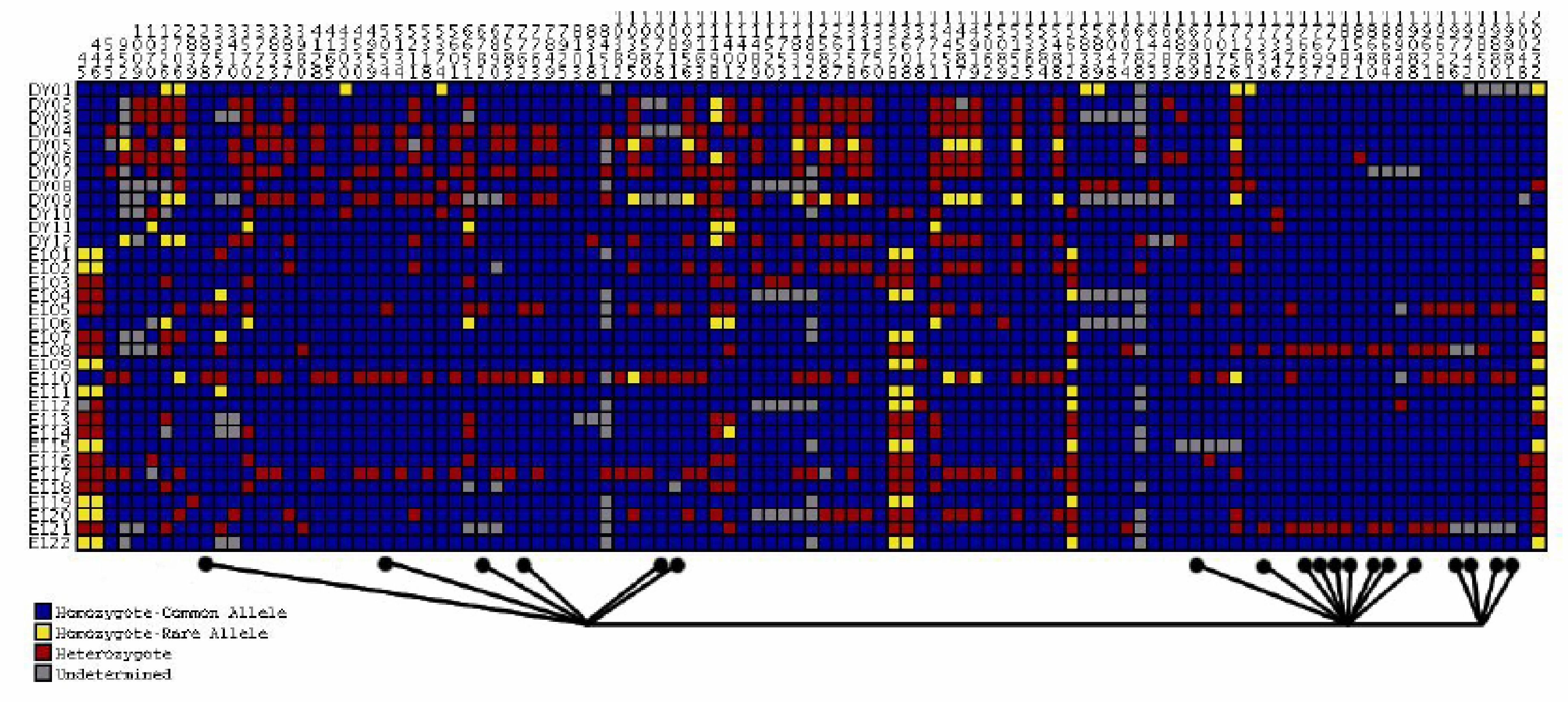
where $S\,(i_1, \ldots i_k) = \sum_{j=1}^{k-1} S\,(i_j, i_{j+1})$

and $S\,(i_j, i_{j+1})$ is a function of both congruence (or near congruence) and physical distance between i_j and i_{j+1}.

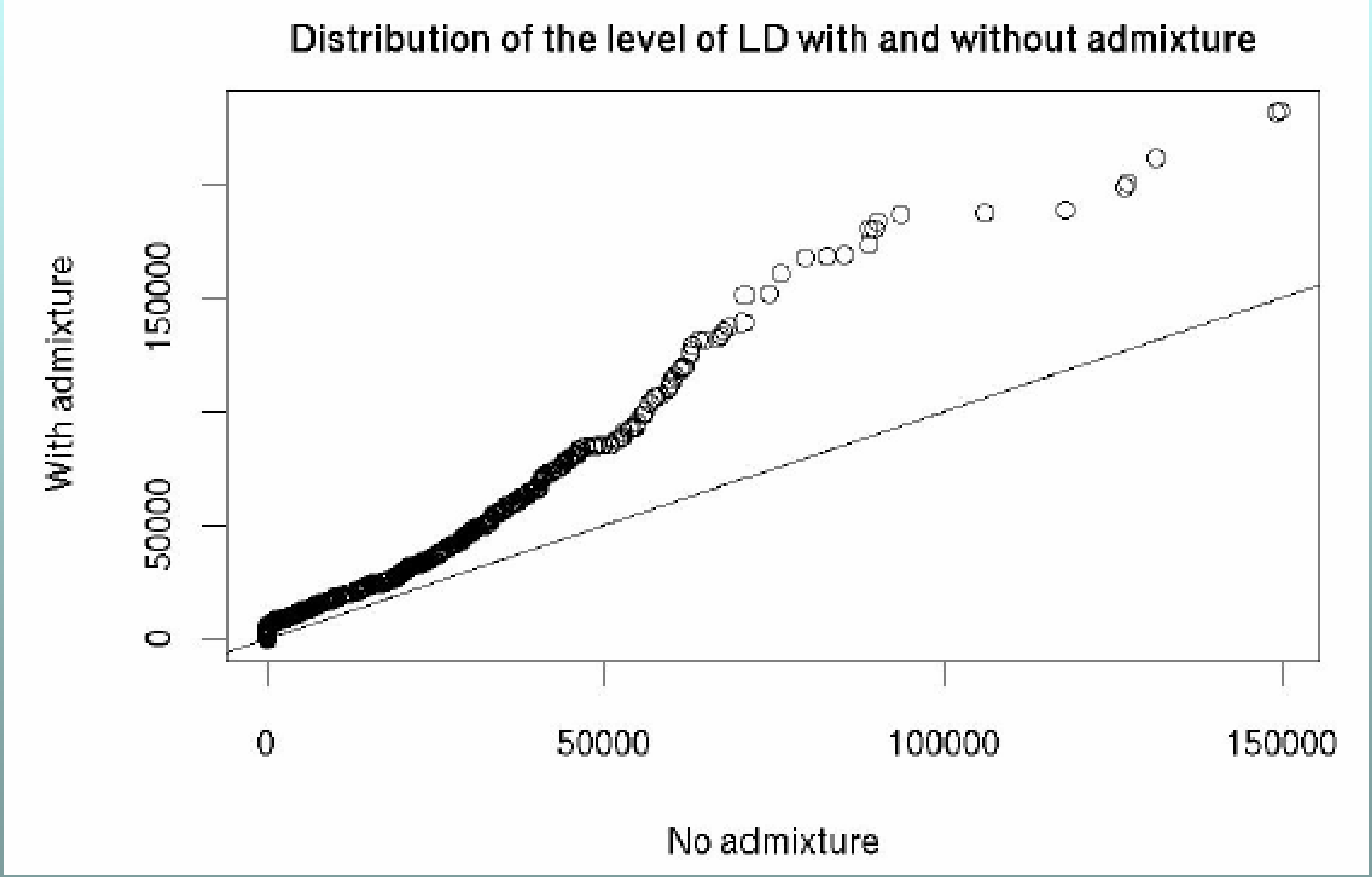
An example


DY01
DY02
DY03
DY04
DY05
DY06
DY07
DY08
DY09
DY10
DY11
DY12
E101
E102
E103
E104
E105
E106
E107
E108
E109
E110
E111
E112
E113
E114
E115
E116
E117
E118
E119
E120
E121
E122

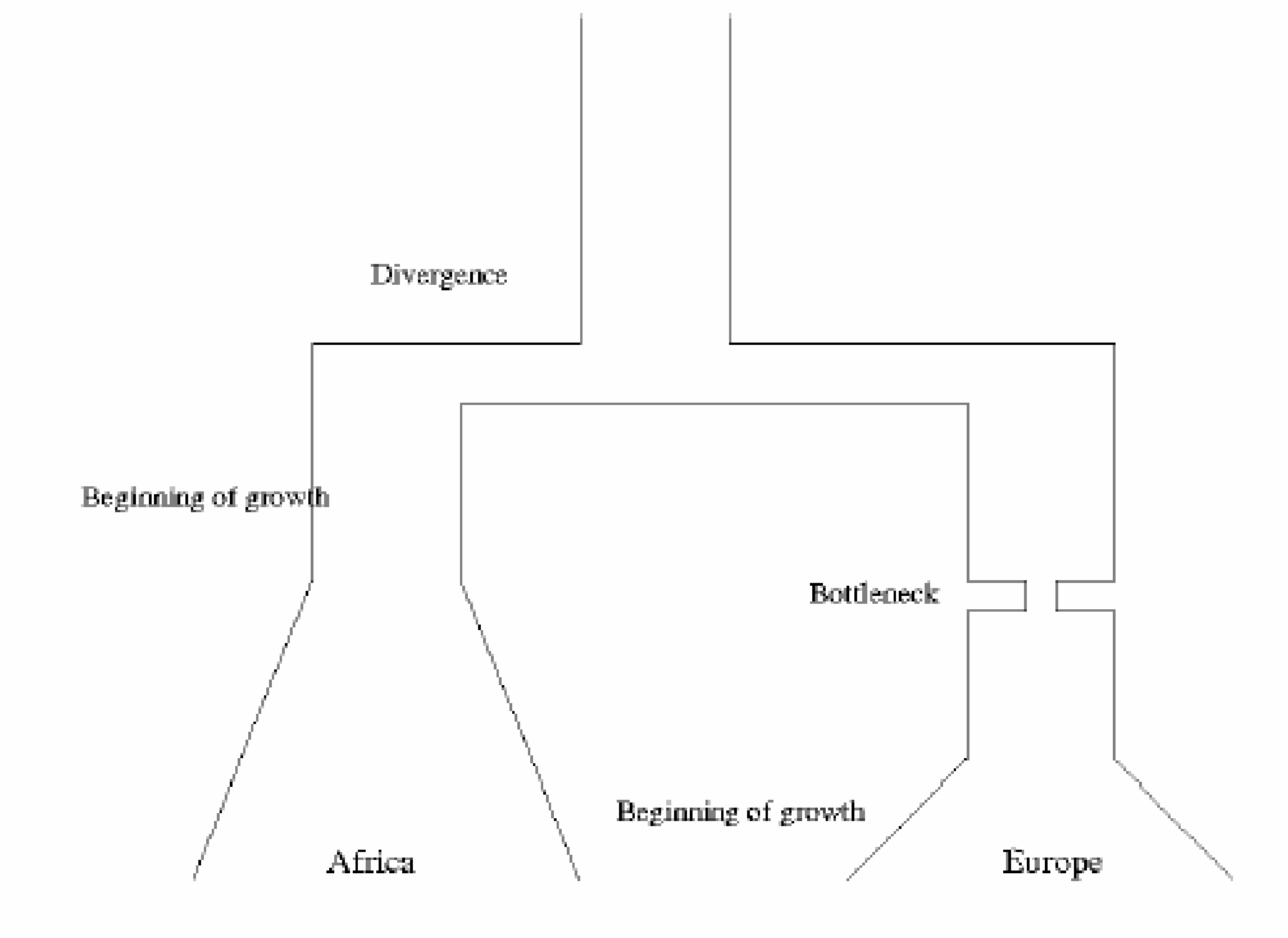
An example


DY01
DY02
DY03
DY04
DY05
DY06
DY07
DY08
DY09
DY10
DY11
DY12
E101
E102
E103
E104
E105
E106
E107
E108
E109
E110
E111
E112
E113
E114
E115
E116
E117
E118
E119
E120
E121
E122
🟦 Homozygote-Common Allele
🟨 Homozygote-Rare Allele
🟥 Heterozygote
⬜ Undetermined

*S** is sensitive to ancient admixture



A more realistic null model



Existing data

To test our methods, we analyzed data from the Environmental Genome Project (from D. Nickerson’s laboratory at the Univ. of Washington).

We analyzed data from 222 genes in 12 Yoruba and 22 European-Americans. In all there was over 5 Mb of sequence data and over 20,000 SNPs.

Plagnol and Wall 2006, PLoS Genetics;
Plagnol et al., unpublished results

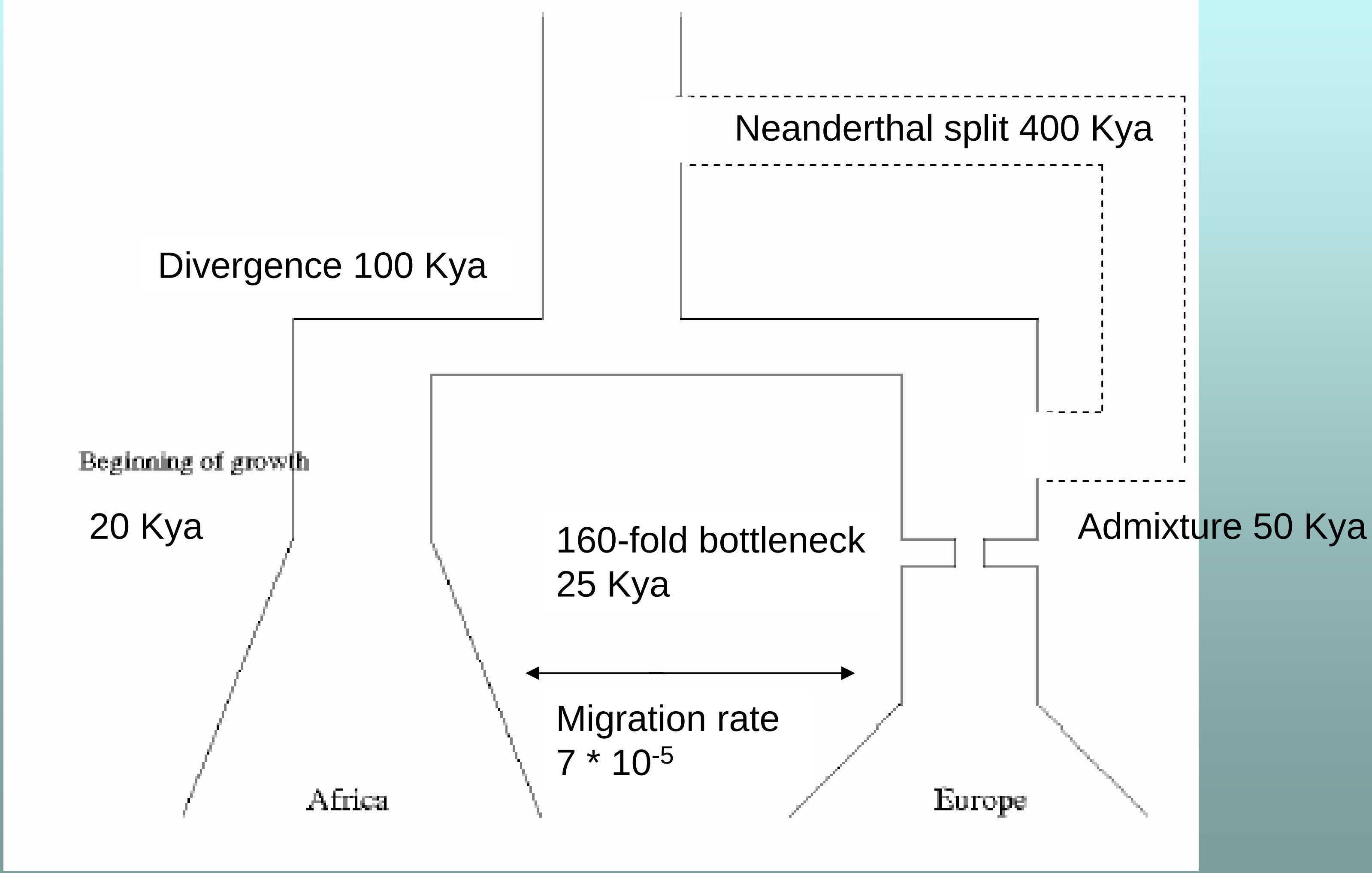
Demographic null model

We used a summary likelihood approach to estimate basic demographic parameters, including

- Time of split between African and European populations
- Timing and strength of bottleneck in European populations
- Timing of recent population growth in both populations
- Migration rate between the two populations

These parameters were then used to construct a demographic null model (without ancient admixture).

Demographic null model



Neanderthal split 400 Kya

Divergence 100 Kya

Beginning of growth

20 Kya

160-fold bottleneck
25 Kya

Admixture 50 Kya

Migration rate
$7 * 10^{-5}$

Fit of the null model

How well does the demographic null model fit the patterns of genetic variation found in the actual data?

Fit of the null model

How well does the demographic null model fit the patterns of genetic variation found in the actual data?

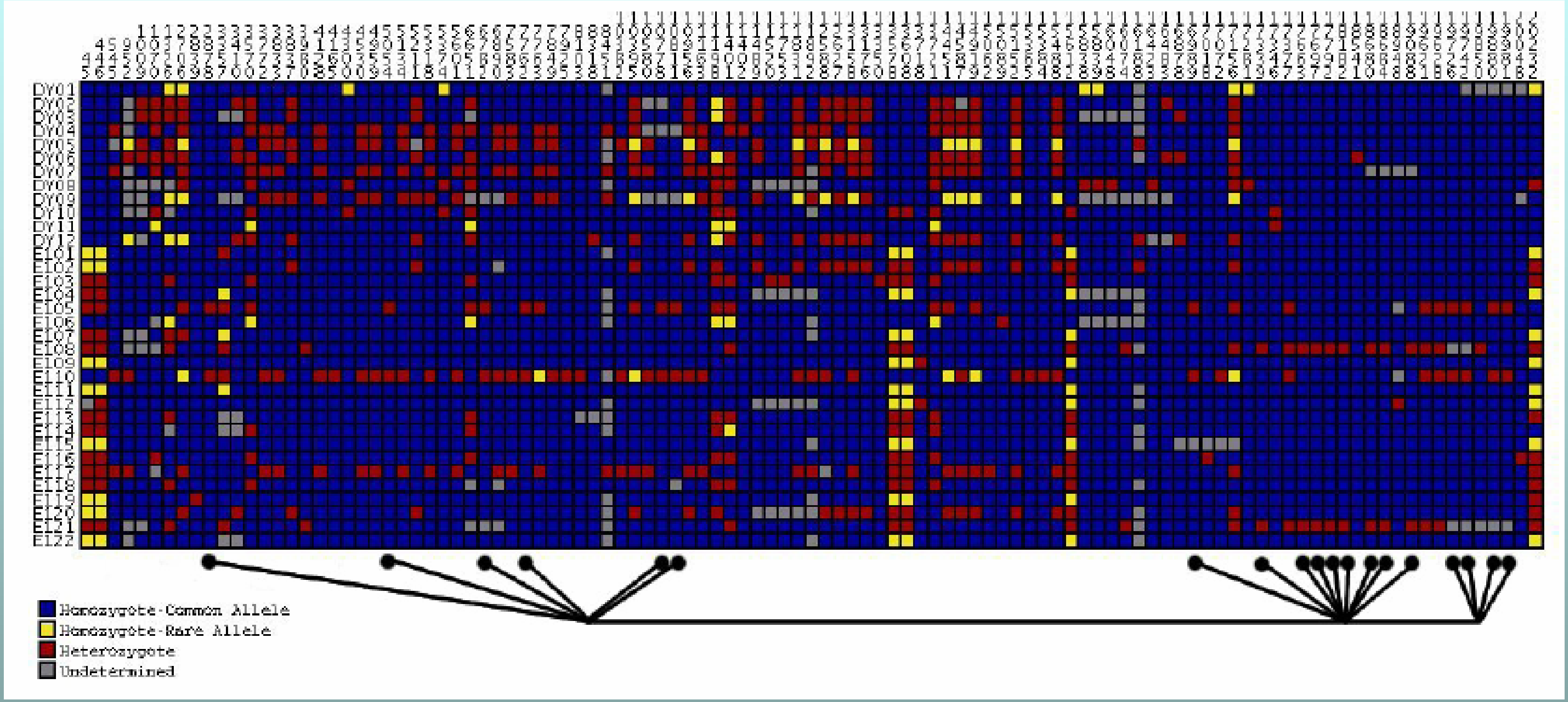
Quite well. The model accurately reproduces both parameters used in the original fitting (e.g., Tajima's *D* in each population) as well as other aspects of the data (e.g., estimates of $\rho = 4Nr$)

General approach

Is our null model sufficient to explain the patterns of LD in the data?

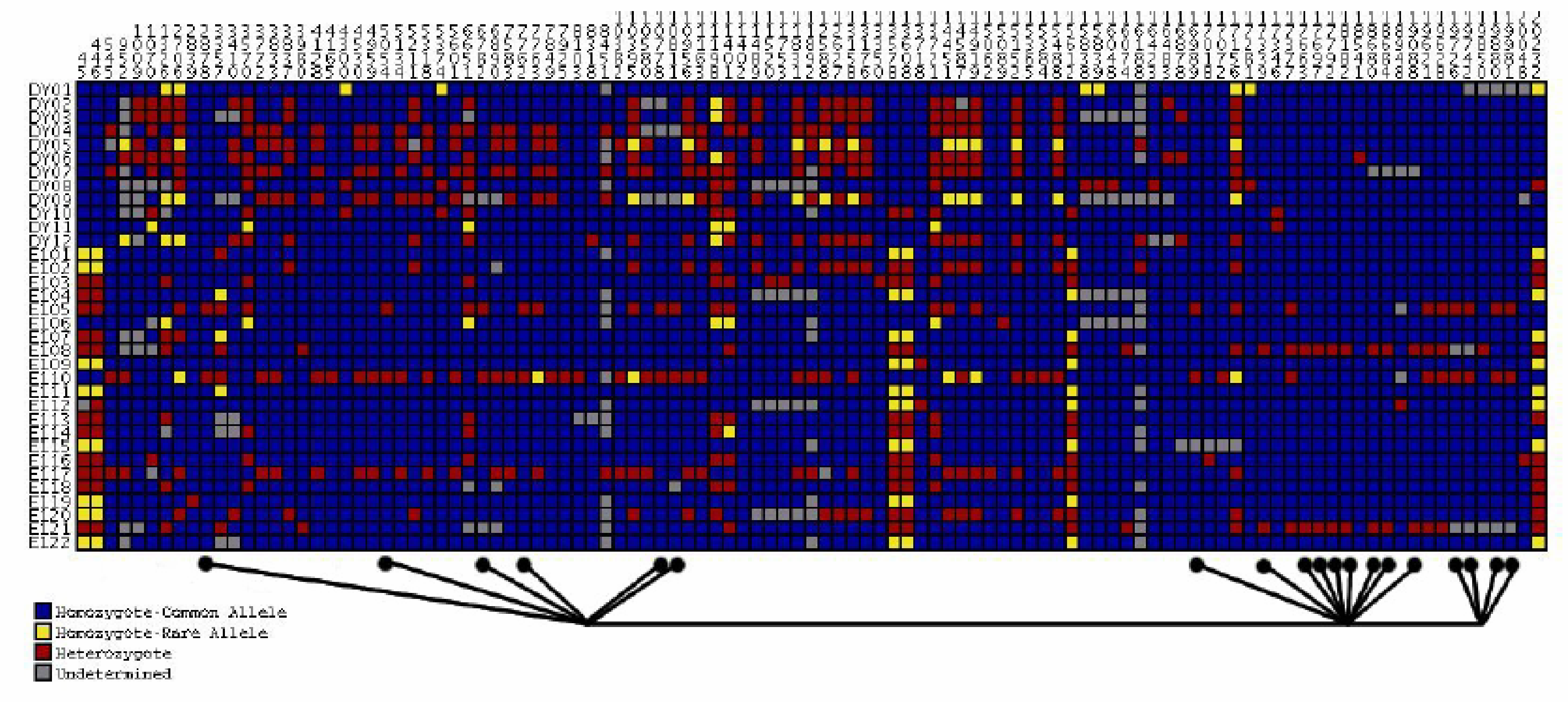
We test this by comparing the observed *S** values with the distribution of *S** values calculated from data simulated under the null model.

An example (CHRNA4)



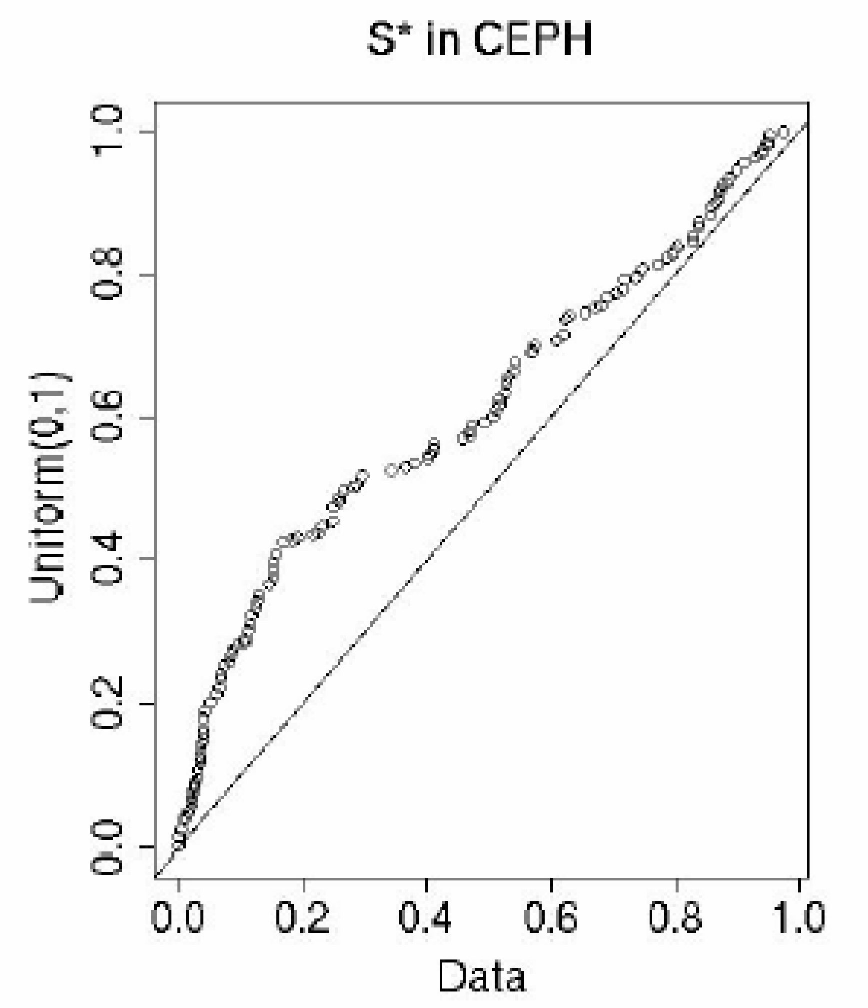
How often is *S** from simulations greater than or equal to the *S** value from the actual data?

An example (CHRNA4)



How often is *S** from simulations greater than or equal to the *S** value from the actual data? $p = 0.025$

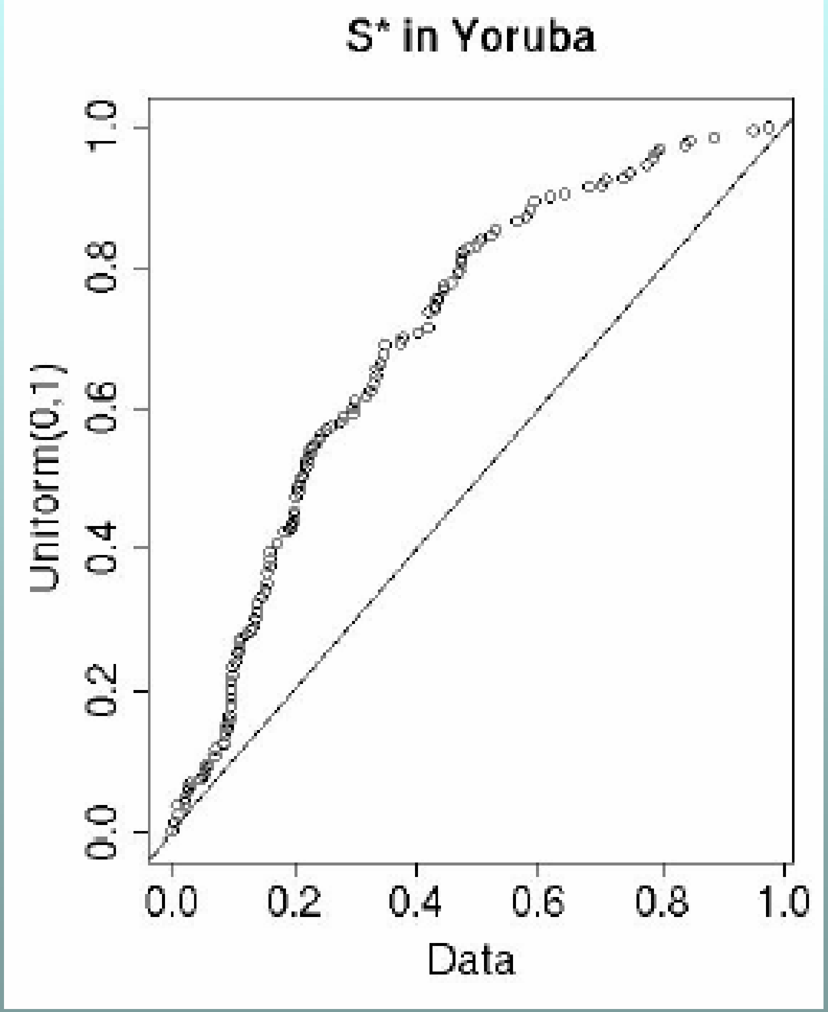
*S** for the NIEHS data (European)



Q – Q plot

$p < 5. * 10^{-3}$

*S** for the NIEHS data (African)



Q – Q plot

$p < 1. * 10^{-7}$

Source populations

We have found some evidence ($p < 5 * 10^{-3}$) for ancient admixture in European, East Asian and West African populations.

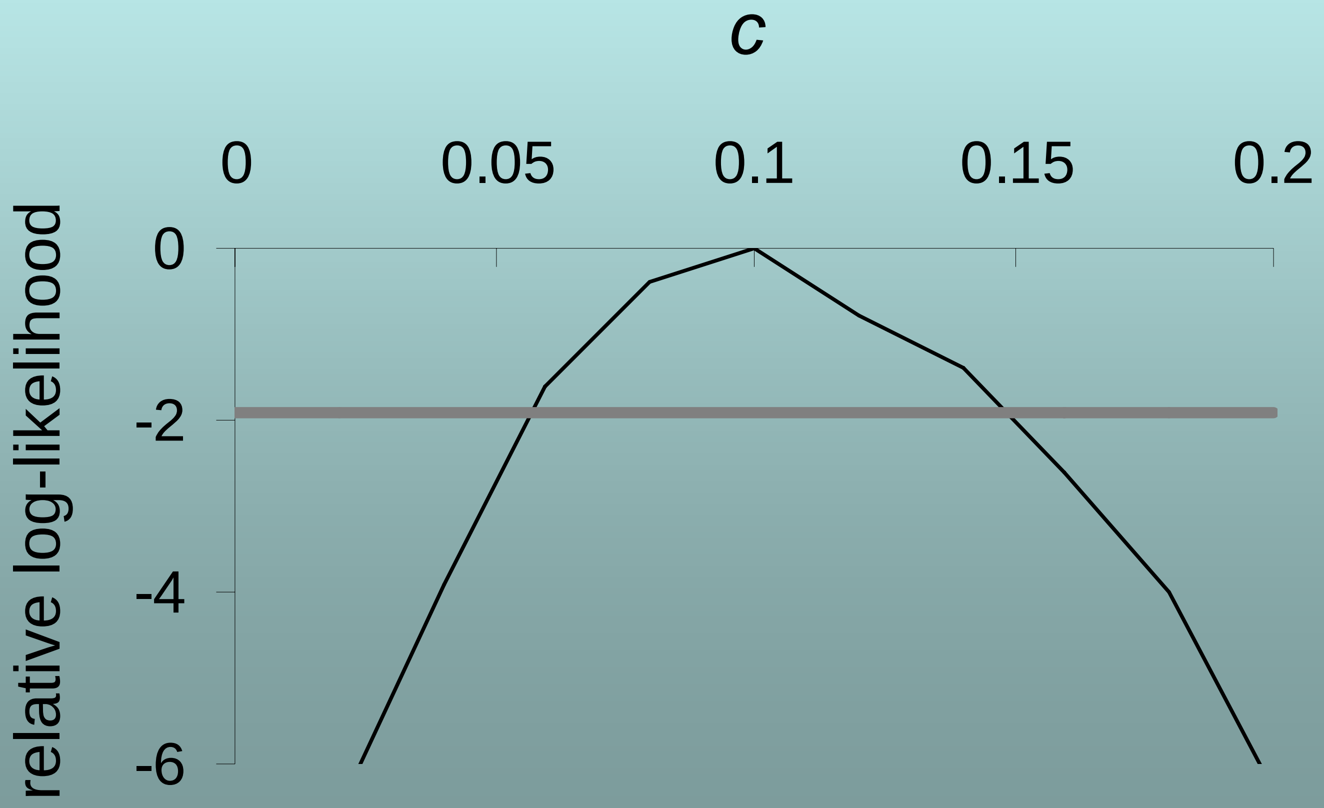
While Neandertals form an obvious archaic source population candidate in Europe, there is not yet a clear source population candidate in West Africa.

Admixture levels

We examine profile likelihood curves for *c* using the summary likelihood method described before.

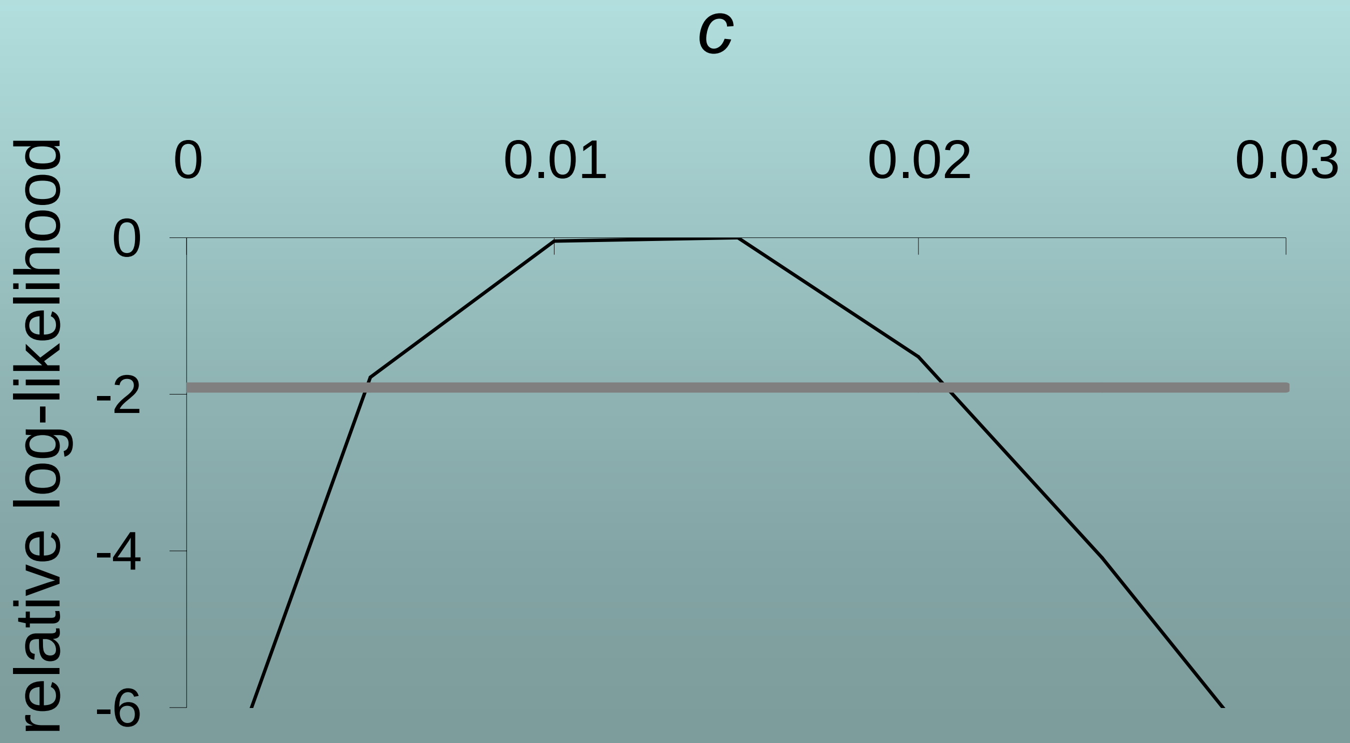
Admixture levels in Europe

We examine profile likelihood curves for *c* using the summary likelihood method described before.



Admixture levels in Asia

The same type of analysis can be done for putative admixture with *Homo erectus* in Asia.



Conclusions

- Using patterns of LD, there seems to be evidence of ancient admixture in all populations studied

- Note that this cannot necessarily be tied to specific fossil groups, such as Neandertals

- These results may reflect ancestral structure predating modern humans’ expansion out of Africa

Acknowledgments

People:
Vincent Plagnol
Kirk Lohmuller
Sung Kim

Data:
Environmental Genome Project
Rubin lab
Pääbo lab

Funding:
National Science Foundation (HOMINID)
Alfred Sloan Foundation