Multiple Sequence Alignment

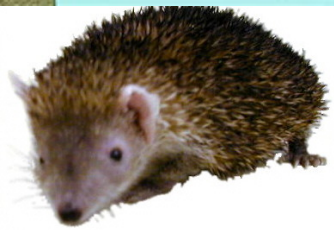








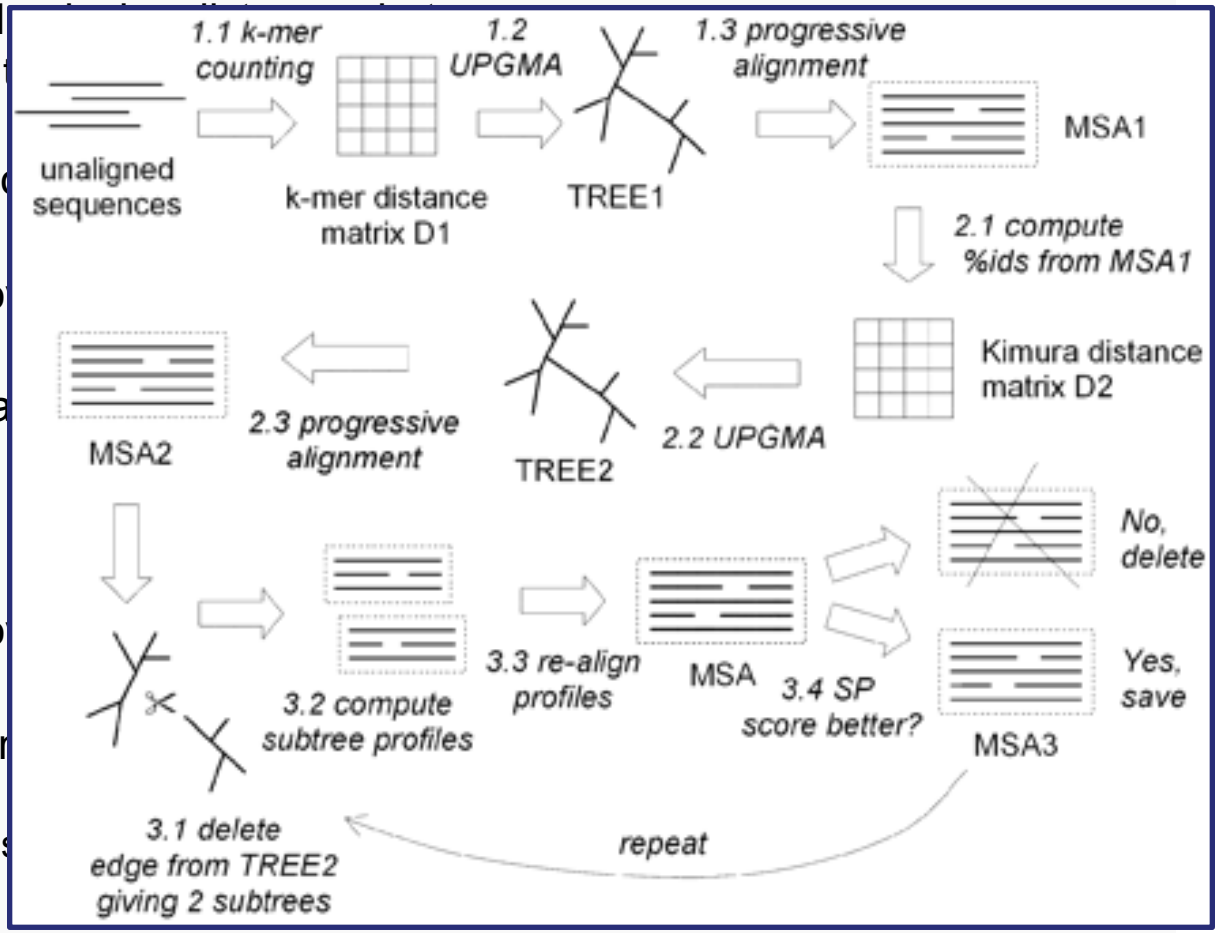




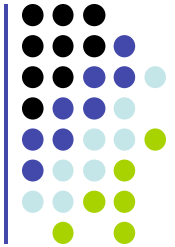



MUSCLE at a glance

1. Fast measurement of all
   - $D_{DRAFT}(x, y)$ defined in
2. Build tree T_{DRAFT} based
3. Progressive alignment o
4. Measure new Kimura-ba
5. Build tree T based on D
6. Progressive alignment o
7. Iterative refinement; for
   - *Tree Partitioning:* Split
   - If new alignment M' has



PROBCONS at a glance



1. Computation of all posterior matrices M_{xy} : $M_{xy}(i, j) = \text{Prob}(x_i \sim y_j)$, using a HMM

2. Re-estimation of posterior matrices M'_{xy} with ***probabilistic consistency***
   - $M'_{xy}(i, j) = 1/N \sum_{\text{sequence } z} \sum_k M_{xz}(i, k) \times M_{yz}(j, k)$; $M'_{xy} = \text{Avg}_z(M_{xz}M_{zy})$

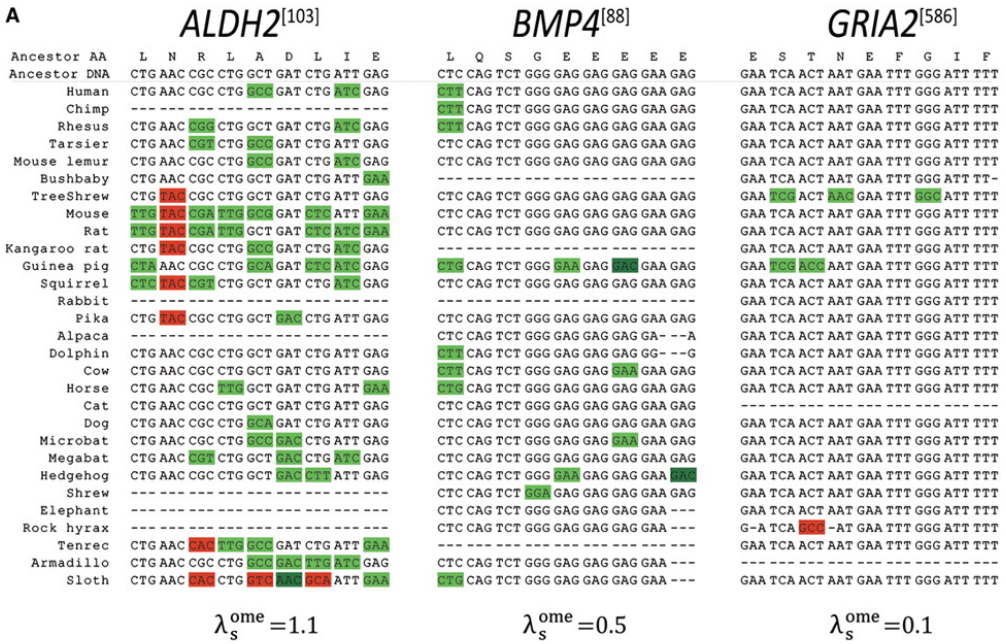
3. Compute for every pair x, y, the maximum expected accuracy alignment
   - A_{xy}: alignment that maximizes $\sum_{\text{aligned (i, j) in A}} M'_{xy}(i, j)$
   - Define $E(x, y) = \sum_{\text{aligned (i, j) in Axy}} M'_{xy}(i, j)$

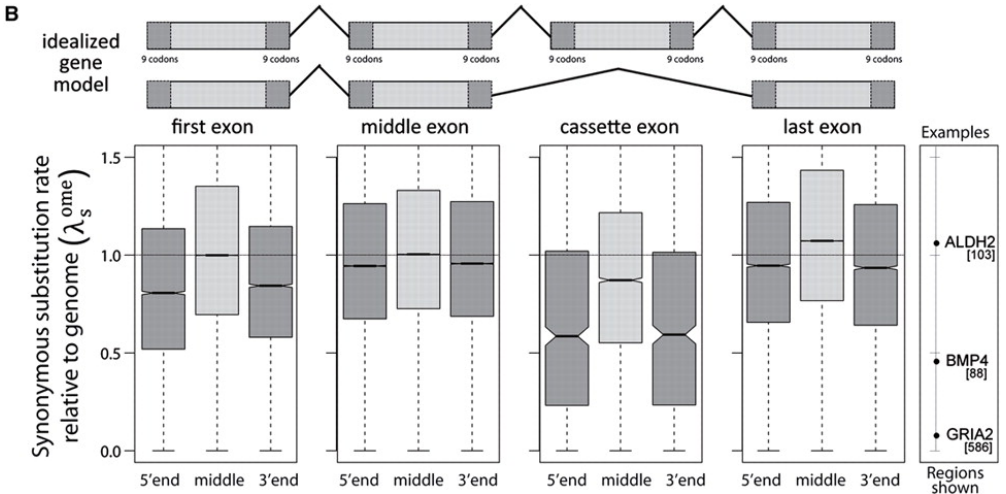
4. Build tree T with hierarchical clustering using similarity measure E(x, y)

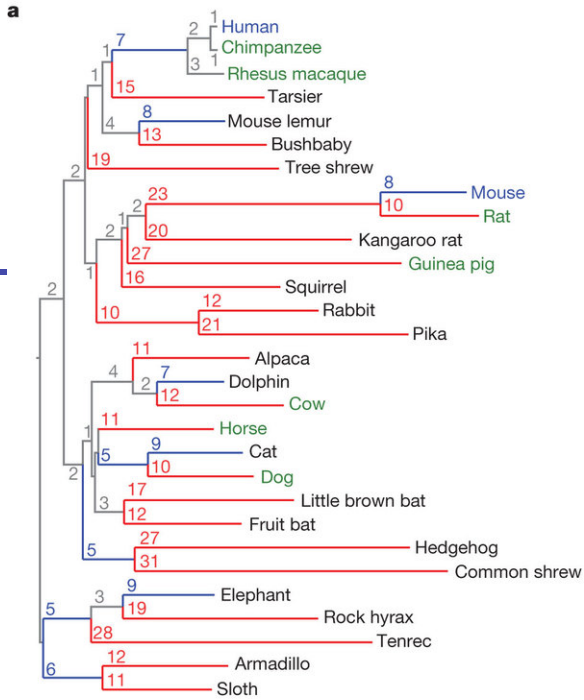
5. Progressive alignment on T to maximize E(.,.)

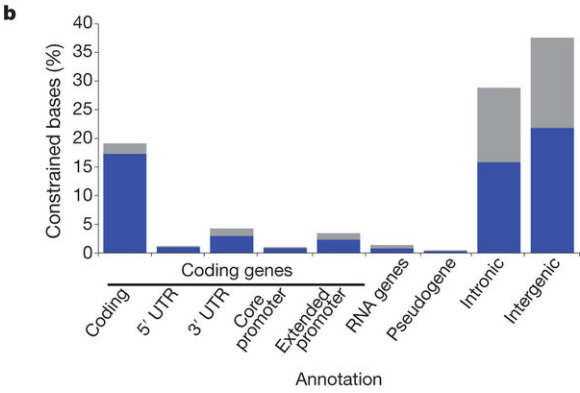
6. Iterative refinement; for many rounds, do:
   - *Randomized Partitioning:* Split sequences in M in two subsets by flipping a coin for each sequence and realign the two resulting profiles

Mammalian alignments





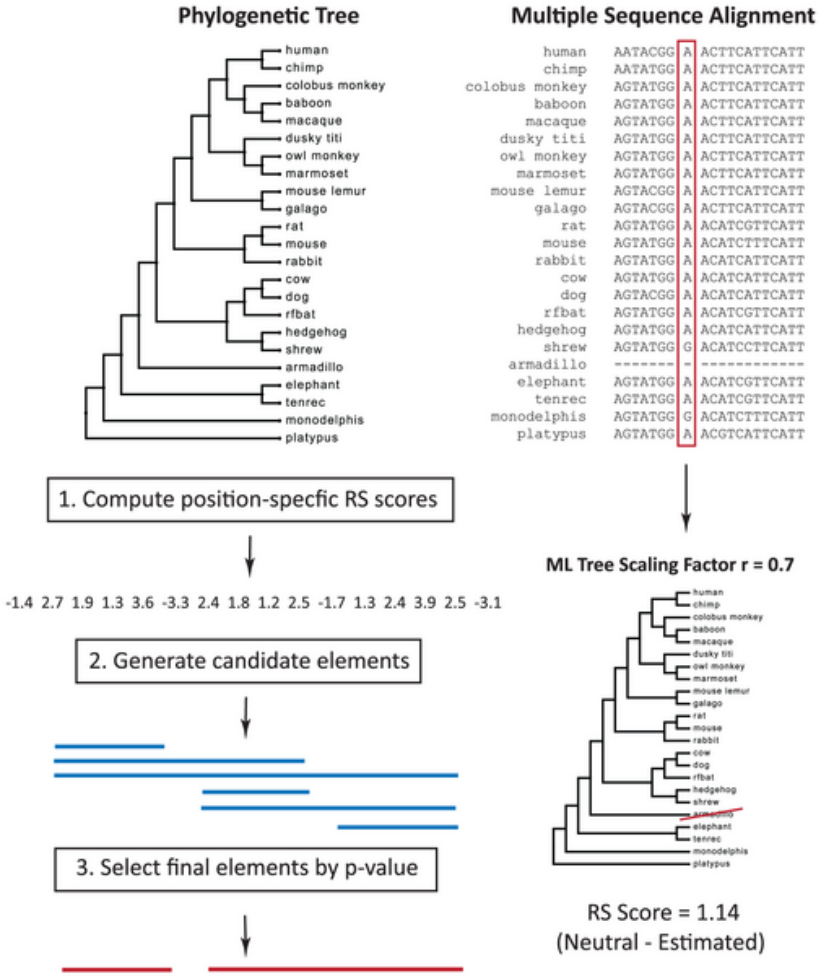


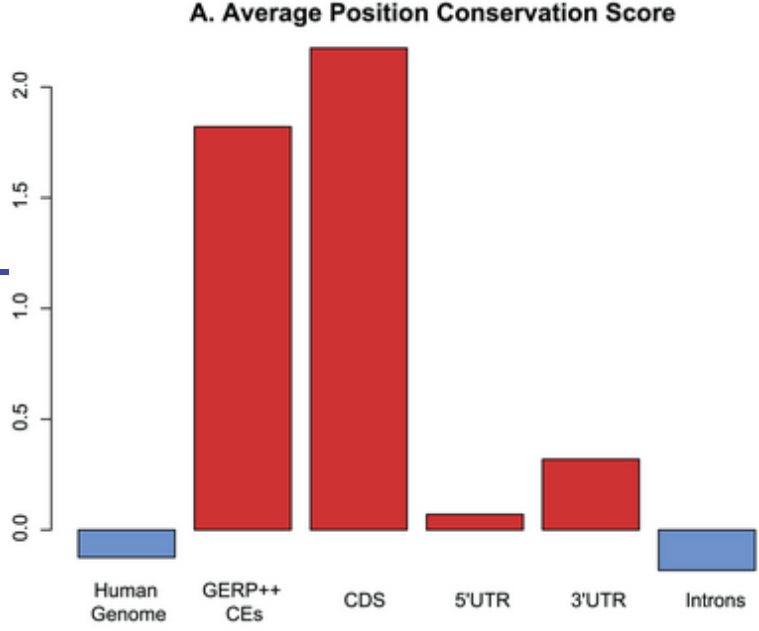


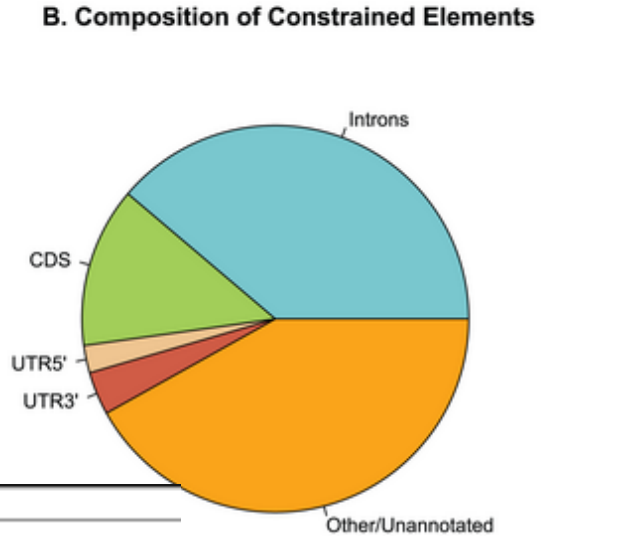
References

- Lindblad-Toh et al. Nature 478:476-482, 2011
- Lin et al. Genome Research 21:1916-1928, 2011

Genome Evolutionary Rate Profiling (GERP)





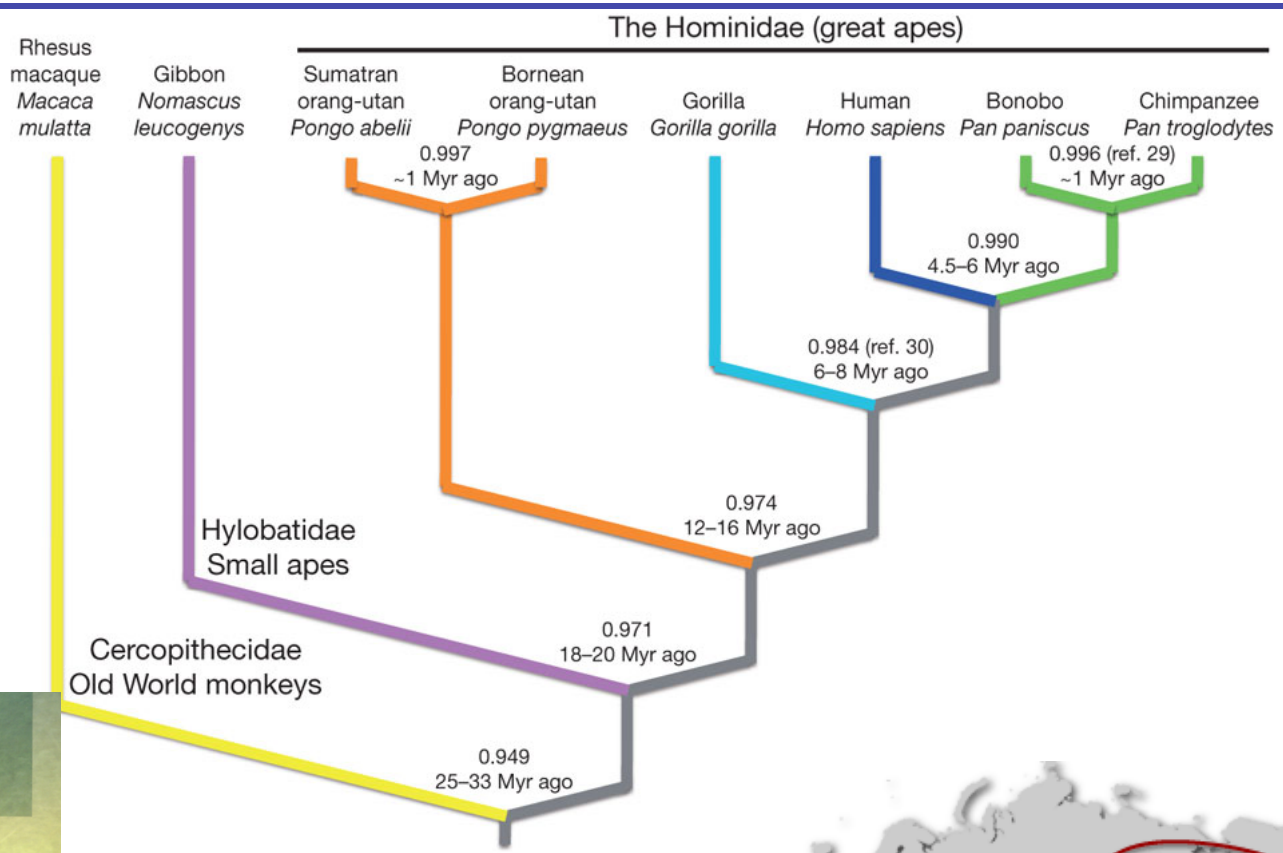


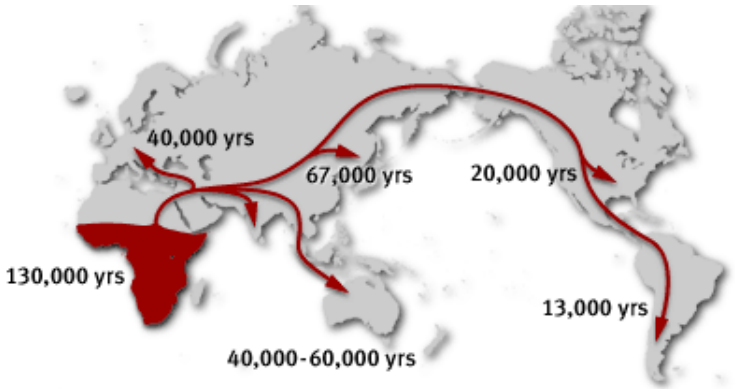
Annotation	% Coverage by CEs
Exons	84.6%
Introns	6.9%
UTR5'	23.7%
UTR3'	33.9%
ncRNA	10.1%

doi:10.1371/journal.pcbi.1001025.t001

Human Population Genomics

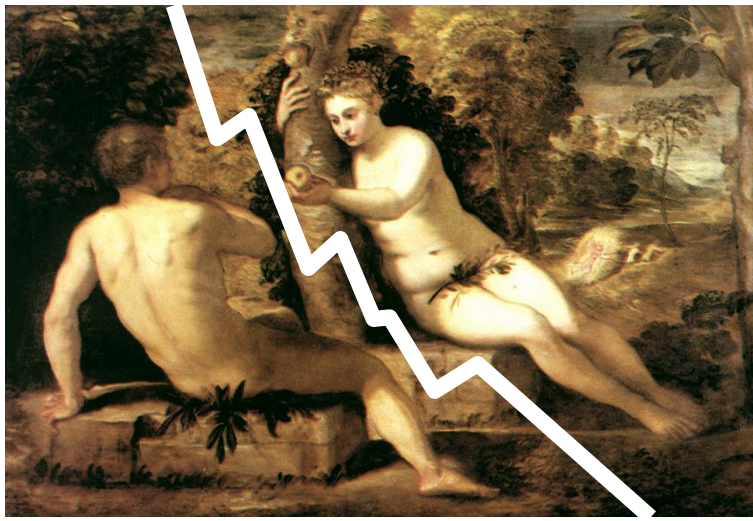
The Hominid Lineage



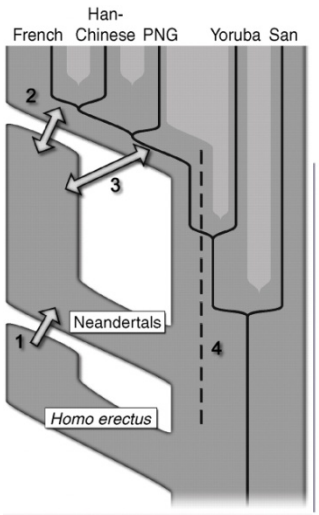


Human population migrations

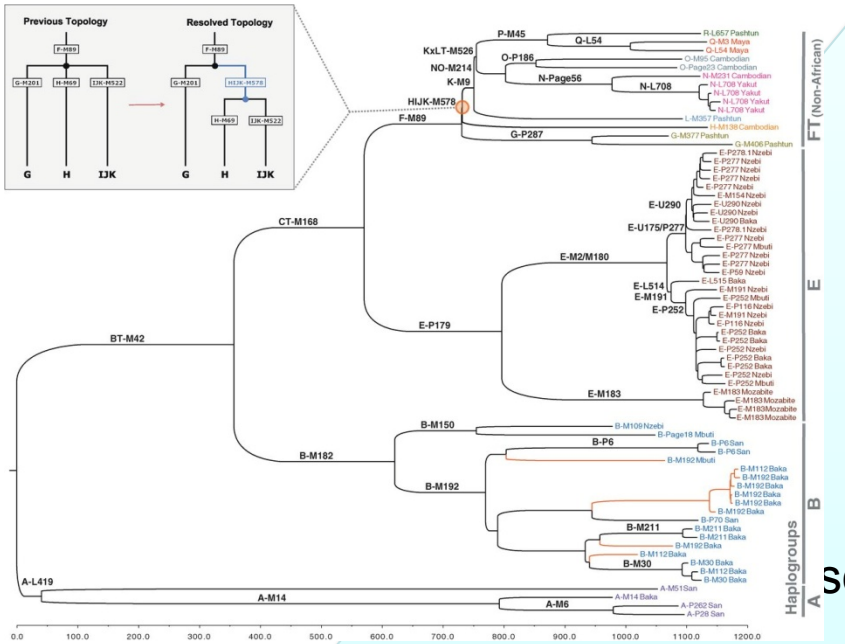
- Out of Africa, Replacement
  - Single mother of all humans (Eve) ~99,000 – 150,000yr
  - Single father of all humans (Adam) ~120,000 - 340,000yr
  - Humans out of Africa ~50000 years ago replaced others (e.g., Neandertals)



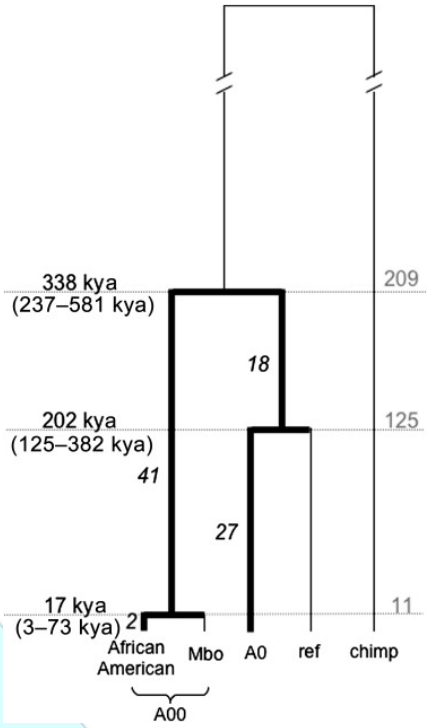
- Multiregional Evolution
  - Generally debunked, however,
  - ~5% of human genome in Europeans, Asians is Neanderthal, Denisova

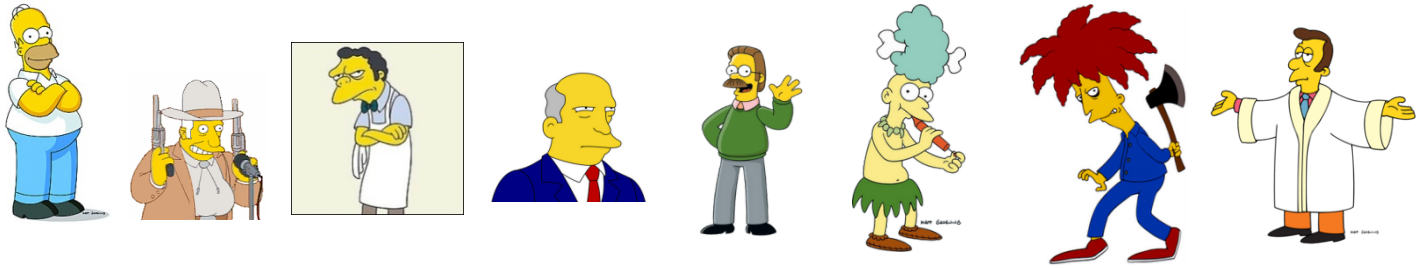


Coalescence



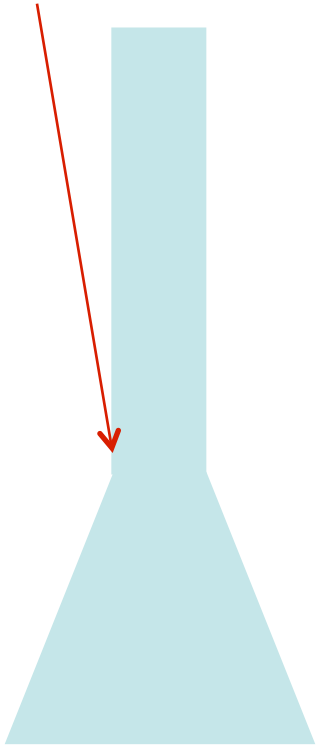
some coalescence

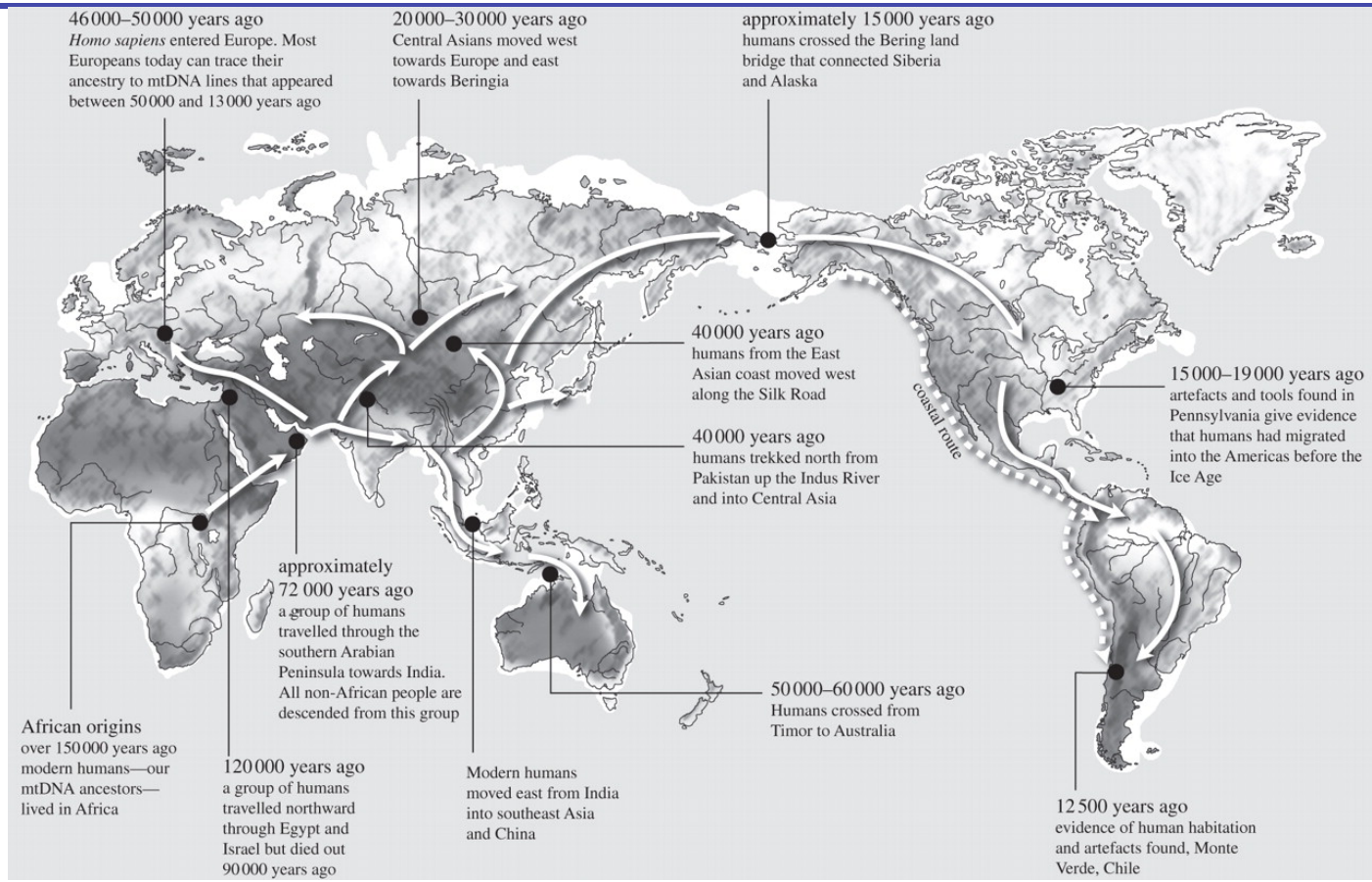




Why humans are so similar

Out of Africa

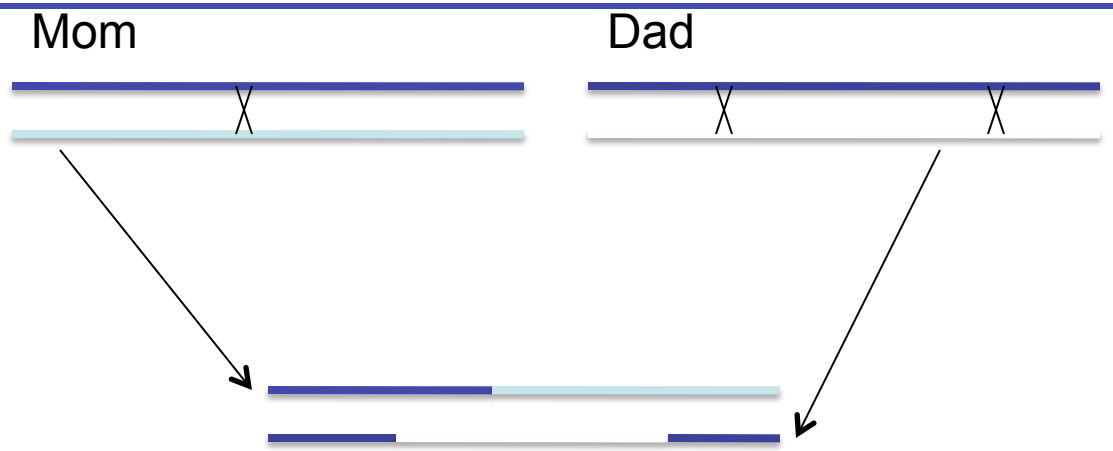




Oppenheimer S Phil. Trans. R. Soc. B 2012;367:770-784

Some Key Definitions

```
Mary: AGCC G/G CG
John: AGCC G/G CG
Josh: AGCC G/T CG
Kate: AGCC G/G CG
Pete: AGCC G/G CG
Anne: AGCC G/G CG
Mimi: AGCC G/G CG
Mike: AGCC T/T CG
Olga: AGCC T/G CG
Tony: AGCC T/G CG
```



Alleles: G, T

Major Allele: G
Minor Allele: T

Heterozygosity:
Prob[2 alleles picked at random with replacement are different]

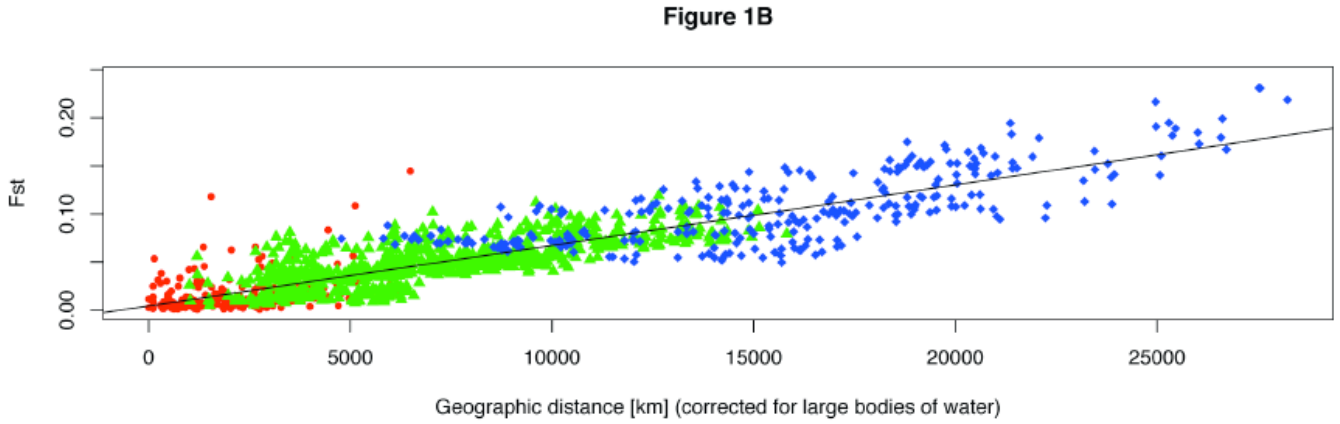
2*.75*.25 = .375

H = 4Nu/(1+4Nu)

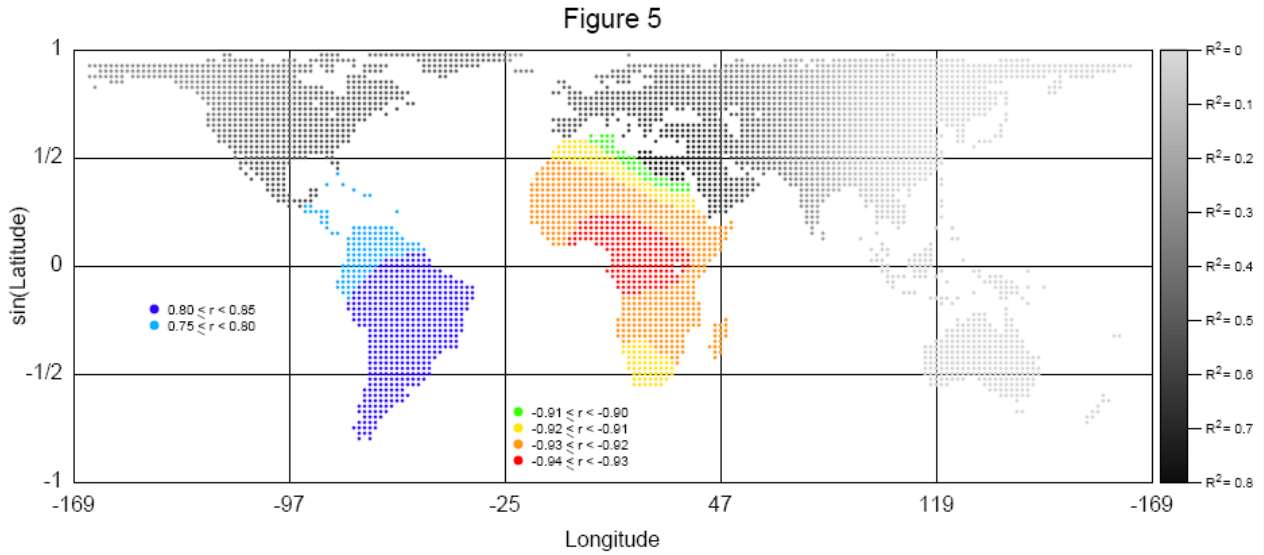
Recombinations:
At least 1/chromosome
On average ~1/100 Mb

Linkage Disequilibrium:
The degree of correlation between two SNP locations

The Fall in Heterozygosity



$$F_{ST} = \frac{H - H_{POP}}{H}$$



The Neanderthal Genome

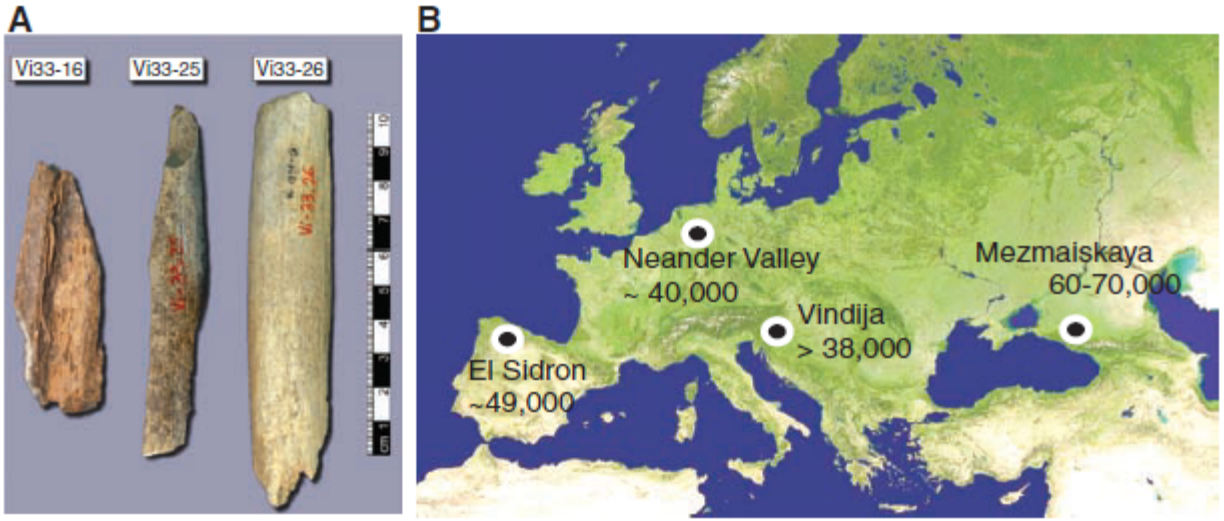


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Figure 1- R. E. Green et al., Science 328, 710-722 (2010)

- From bones, compared genomes of three different Neanderthals with five genomes from modern humans from different areas of the world

Neanderthal Genome

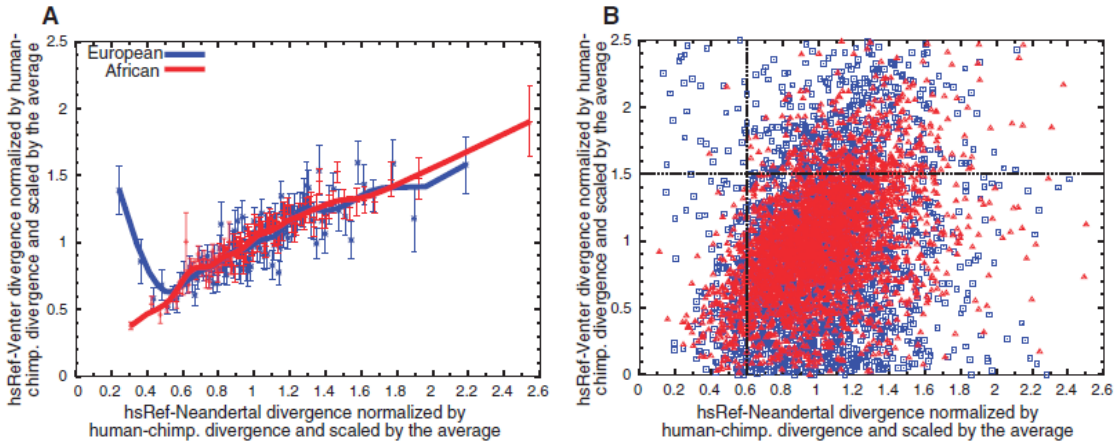


Fig. 5. Segments of Neandertal ancestry in the human reference genome. We examined 2825 segments in the human reference genome that are of African ancestry and 2797 that are of European ancestry. (**A**) European segments, with few differences from the Neandertals, tend to have many differences from other present-day humans, whereas African segments do not, as expected if the former are derived from Neandertals. (**B**) Scatter plot of the segments in (A) with respect to their divergence to the Neandertals and to Venter. In the top left quandrant, 94% of segments are of European ancestry, suggesting that many of them are due to gene flow from Neandertals.

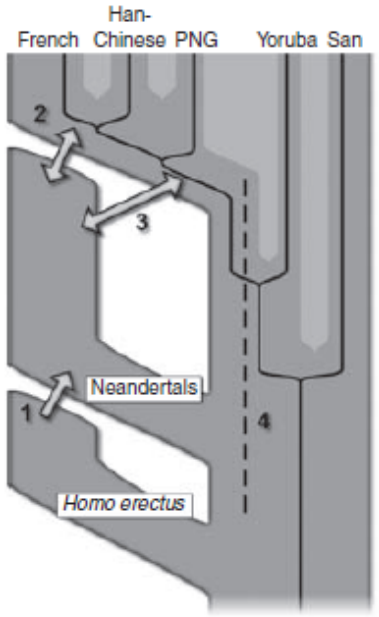


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from present-day humans. Scenario 2 represents gene flow between late Neandertals and early modern humans in Europe and/or western Asia. We see no evidence of this because Neandertals are equally distantly related to all non-Africans. However, such gene flow may have taken place without leaving traces in the present-day gene pool. Scenario 3 represents gene flow between Neandertals and the ancestors of all non-Africans. This is the most parsimonious explanation of our observation. Although we detect gene flow only from Neandertals into modern humans, gene flow in the reverse direction may also have occurred. Scenario 4 represents old substructure in Africa that persisted from the origin of Neandertals until the ancestors of non-Africans left Africa. This scenario is also compatible with the current data.

Denisovan – Another human relative

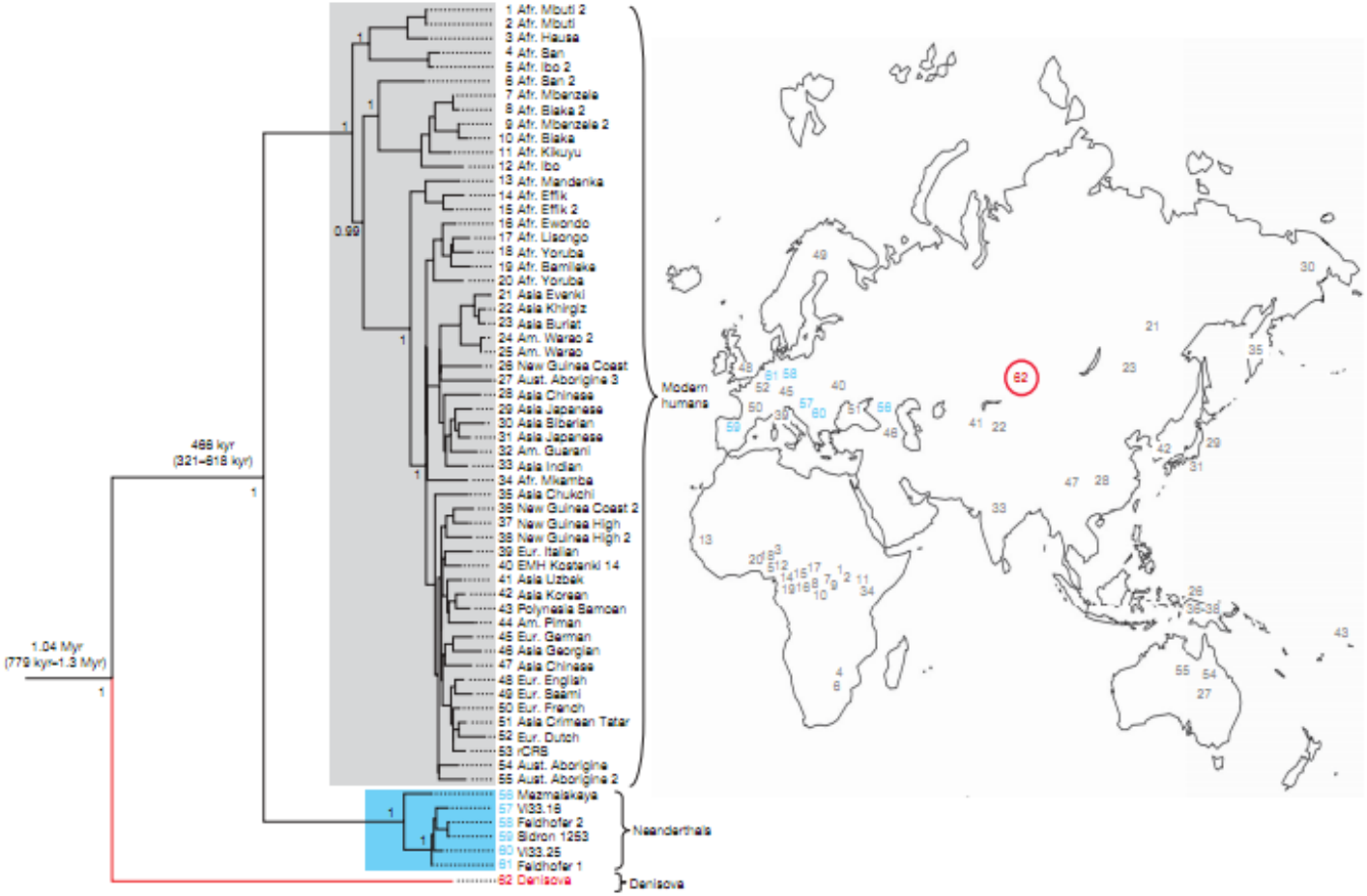
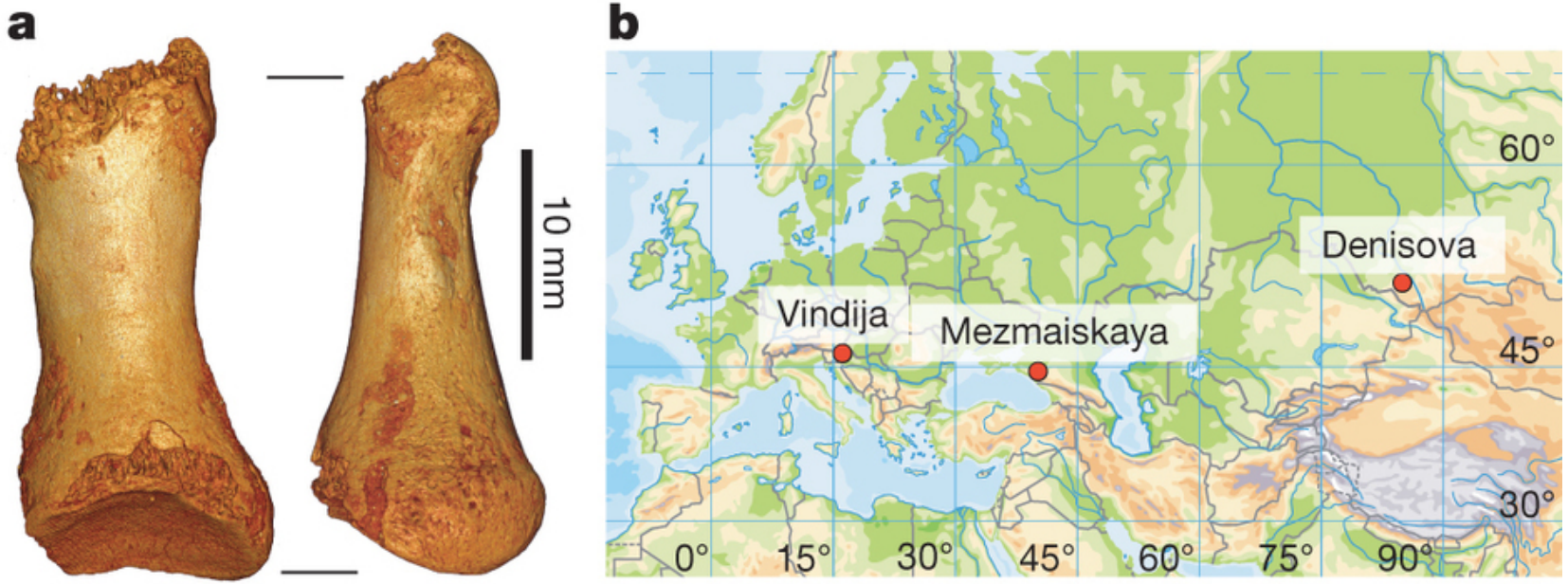
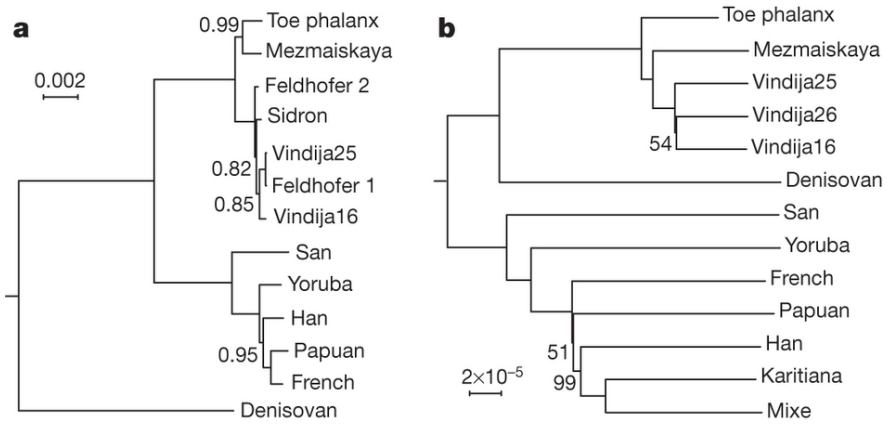


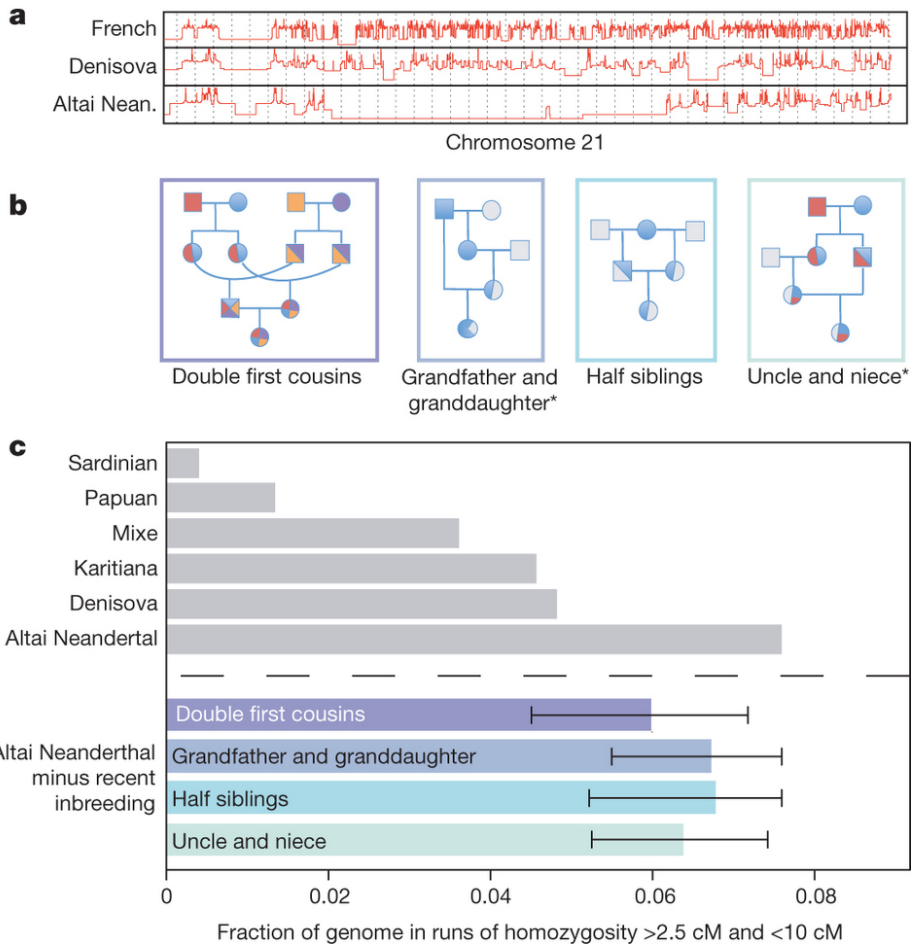
Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+1+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. Posterior probabilities are given for each major node. The map shows the geographical origin of the mtDNAs (24, 25, 32, 44 are in the Americas). Note that two partial mtDNAs sequenced from Teshik Tash and Okladikov Cave in Central Asia fall together with the complete Neanderthal mtDNAs in phylogenies[4] (not shown).

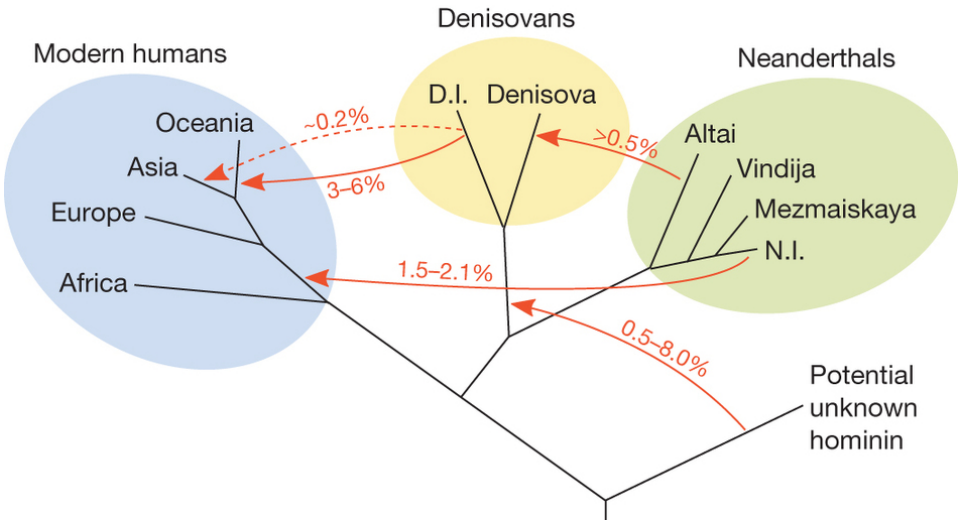
The Neanderthal Whole Genome



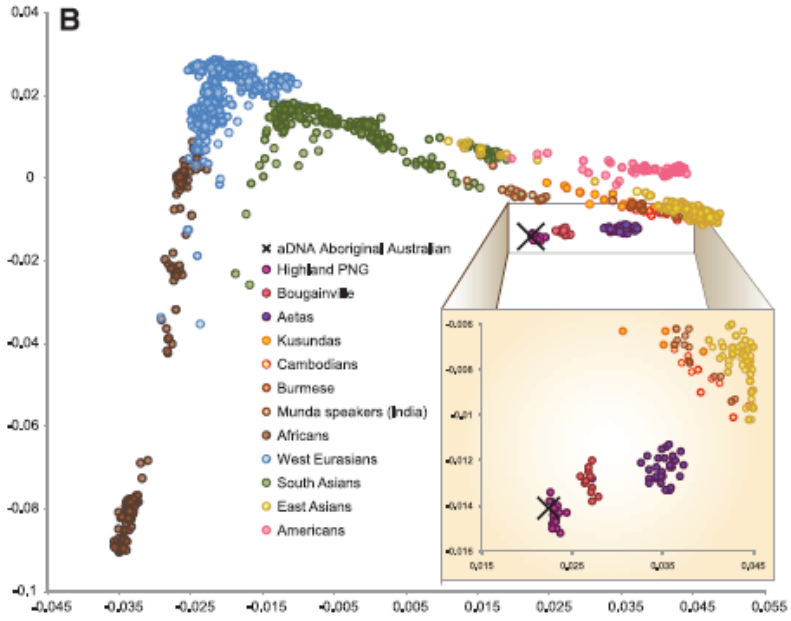


The Neanderthal Whole Genome





Aboriginal Australian



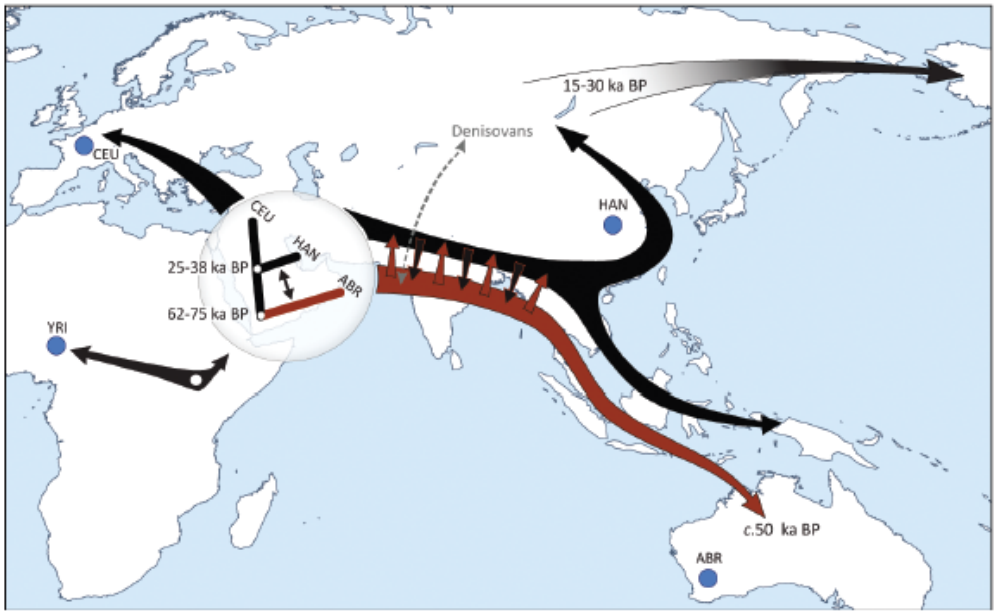
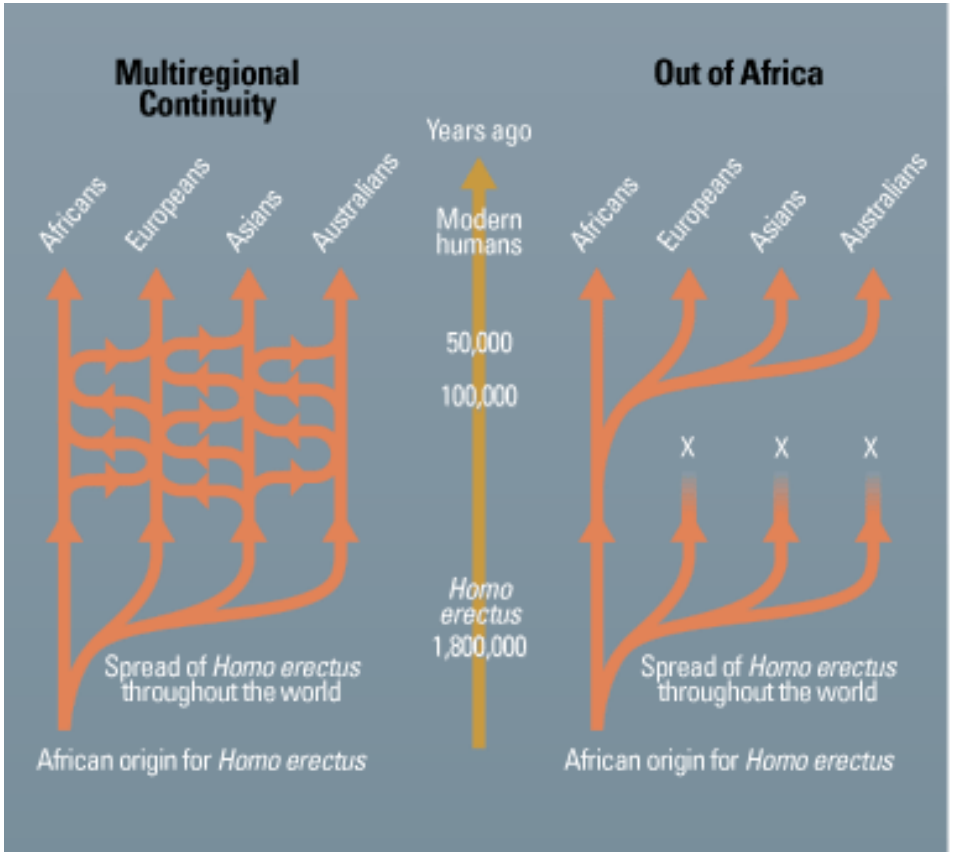
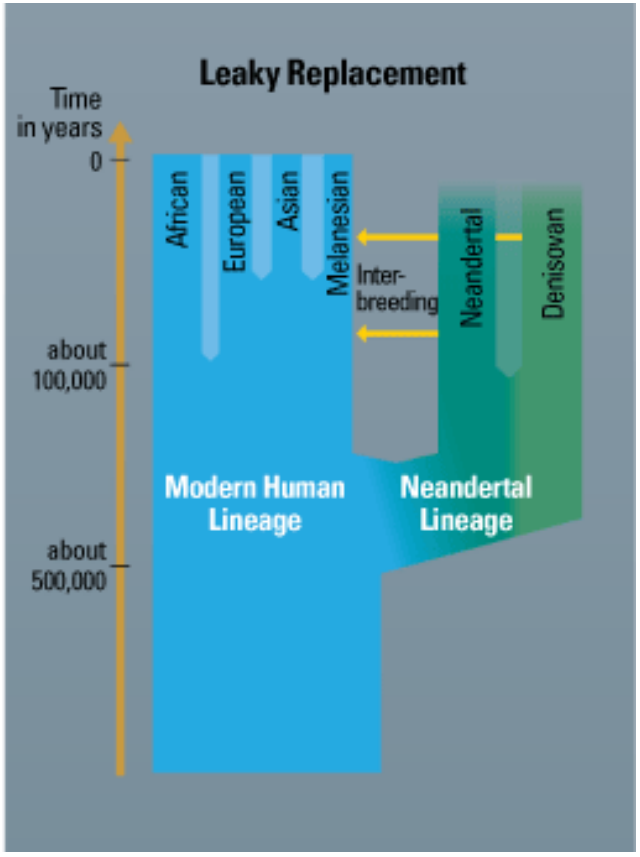


Fig. 2. Reconstruction of early spread of modern humans outside Africa. The tree shows the divergence of the Aboriginal Australian (ABR) relative to the CEPH European (CEU) and the Han Chinese (HAN) with gene flow between aboriginal Australasians and Asian ancestors. Purple arrow shows early spread of the ancestors of Aboriginal Australians into eastern Asia ~62,000 to 75,000 years B.P. (ka BP), exchanging genes with Denisovans, and reaching Australia ~50,000 years B.P. Black arrow shows spread of East Asians ~25,000 to 38,000 years B.P. and admixing with remnants of the early dispersal (red arrow) some time before the split between Asians and Native American ancestors ~15,000 to 30,000 years B.P. YRI, Yoruba.

Out of Africa Revisited

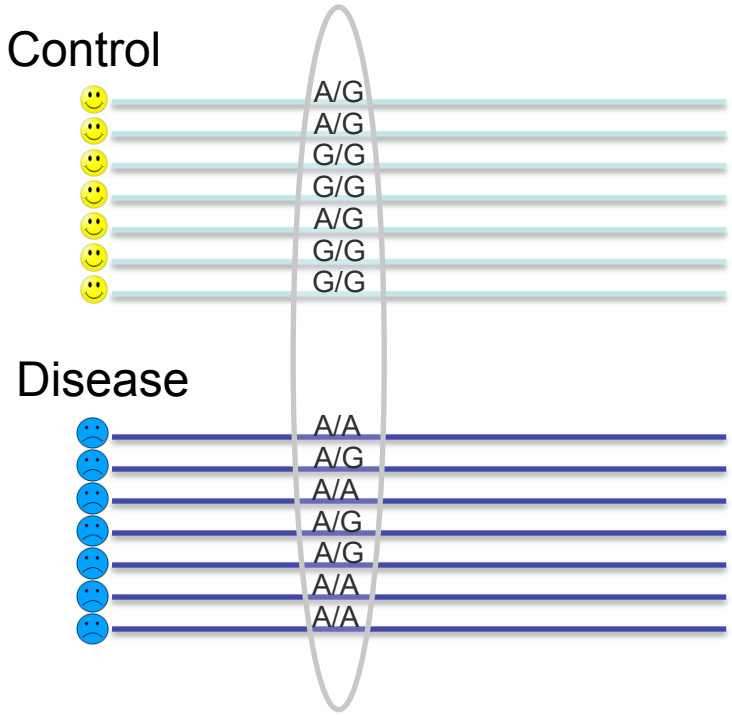
"Human uniqueness?"





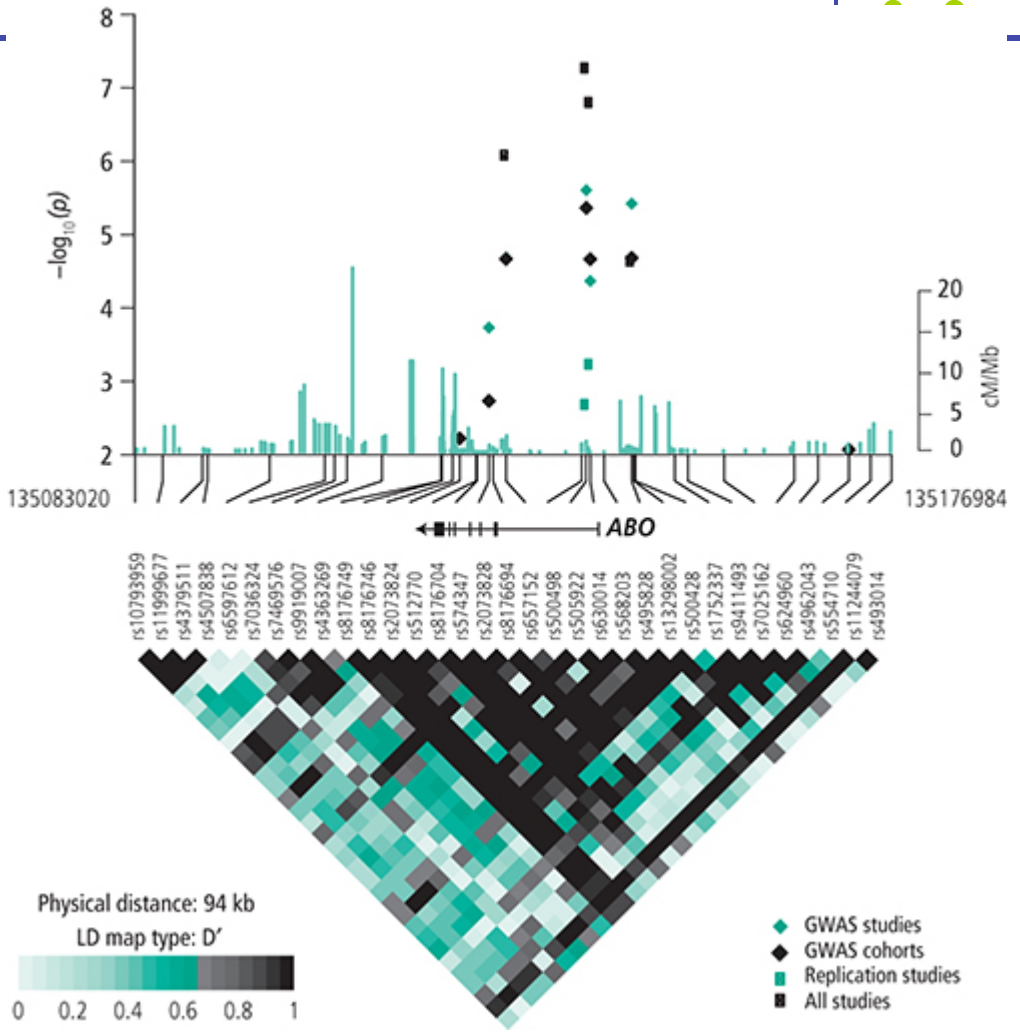
Ann Gibbons Science 28 January 2011:

Association Studies

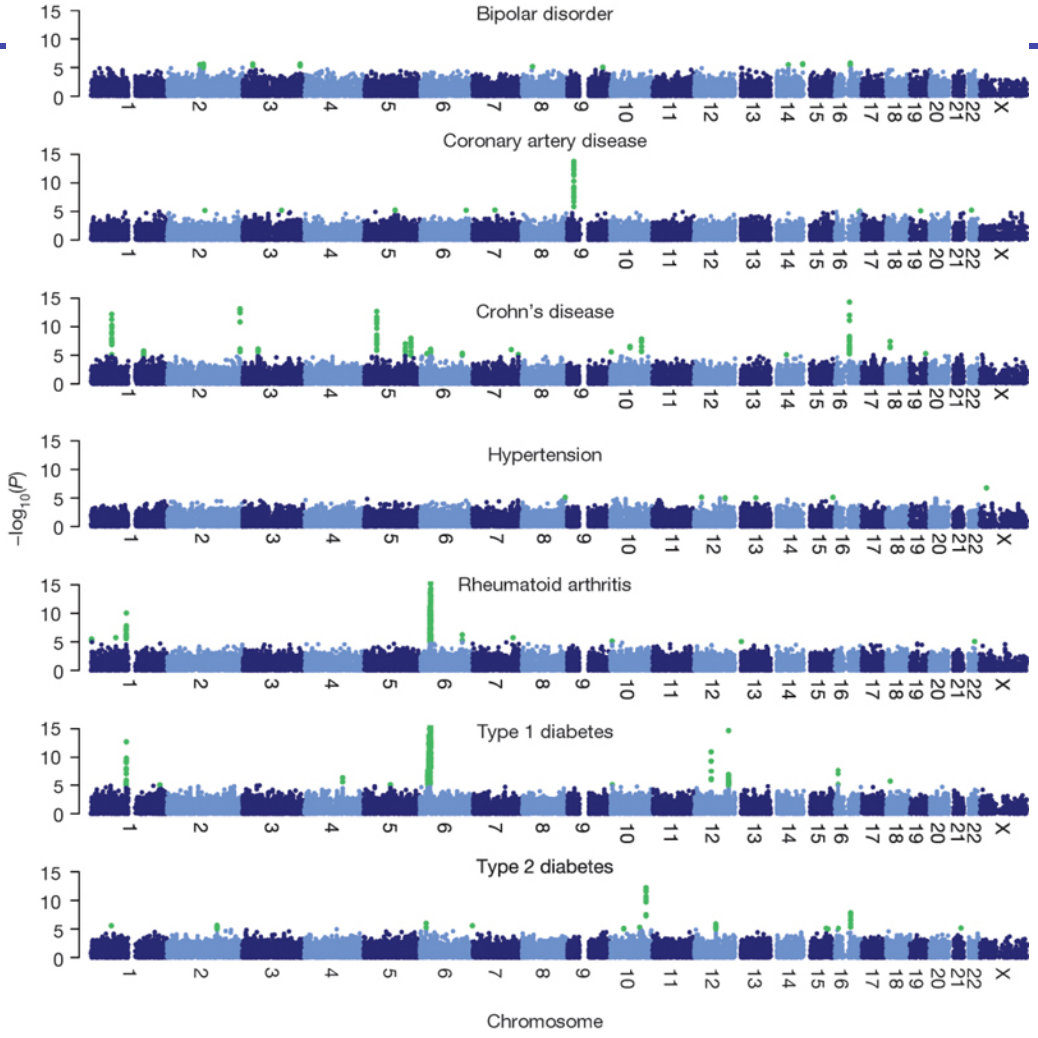


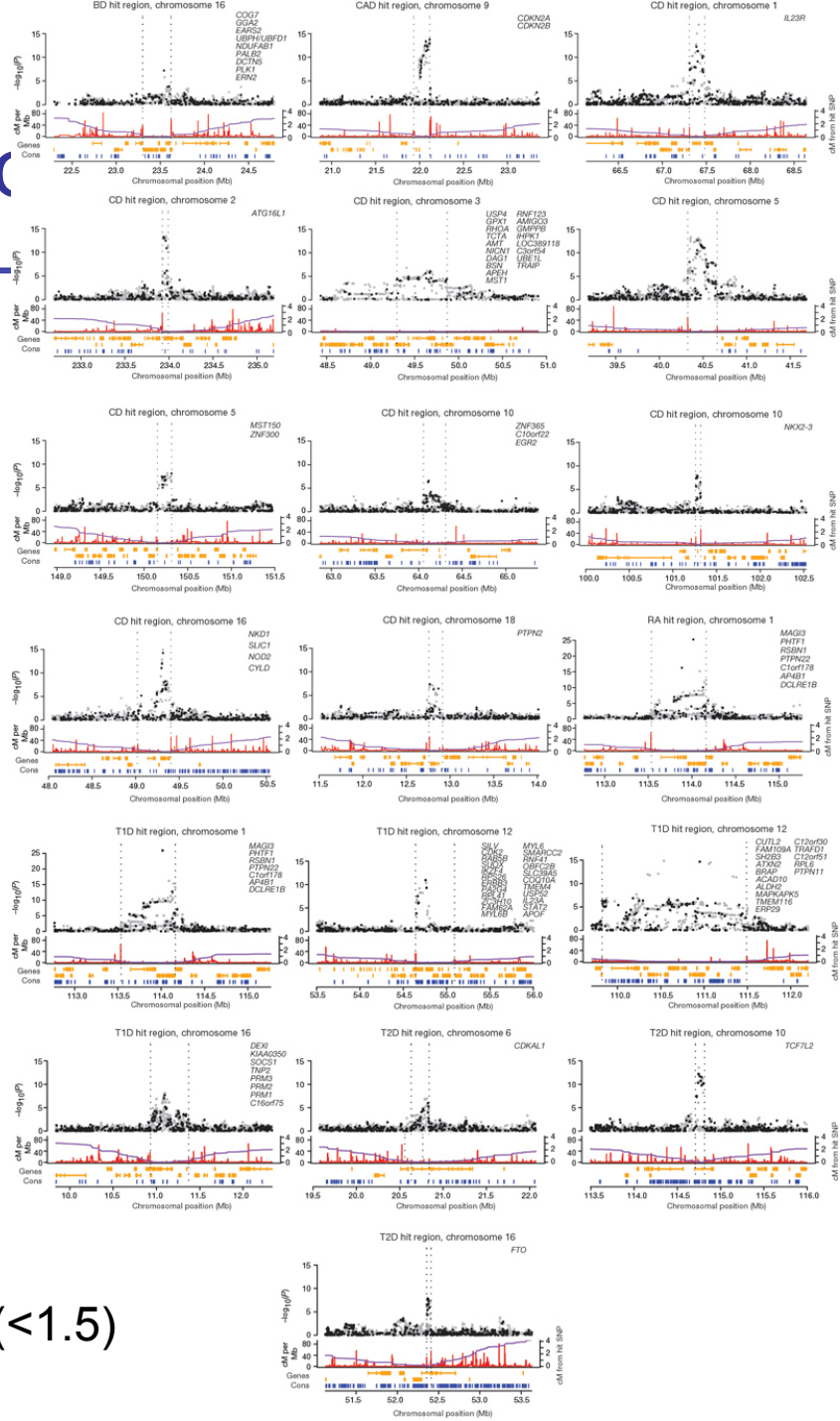
AA	0	4
AG	3	3
GG	4	0

→ p-value



Wellcome Trust Case Contro

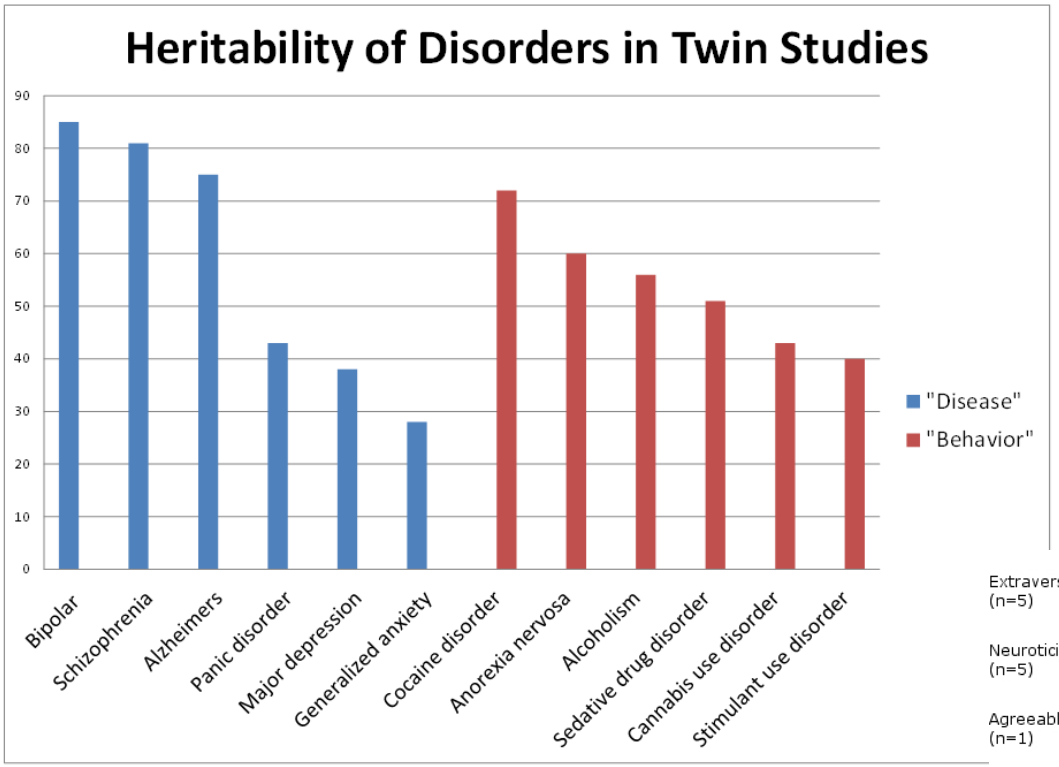


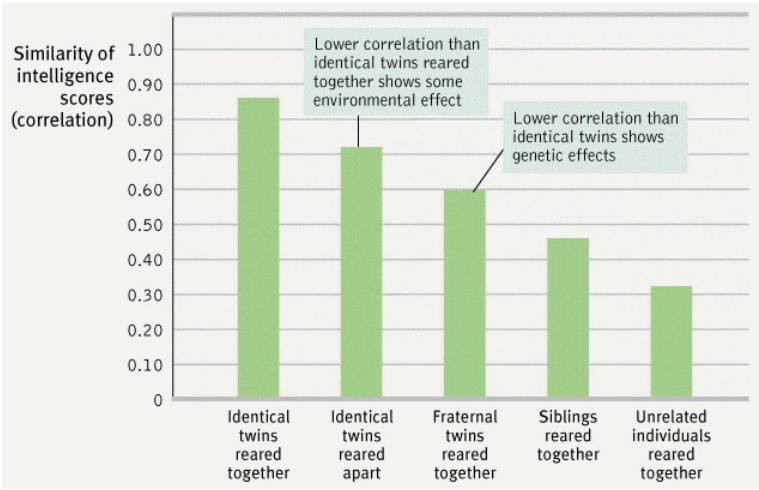


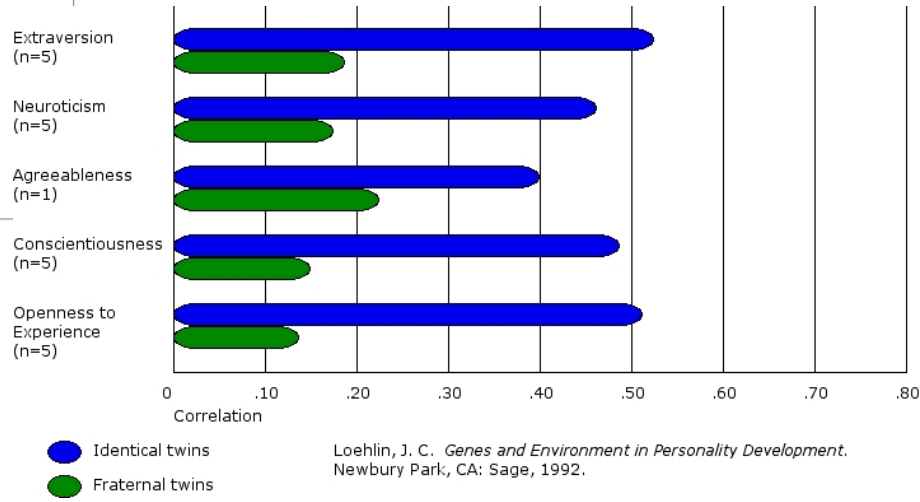
Many associations of small effect sizes (<1.5)

Nature **447**, 661-678(7 June 2007)

Heritability & Environment



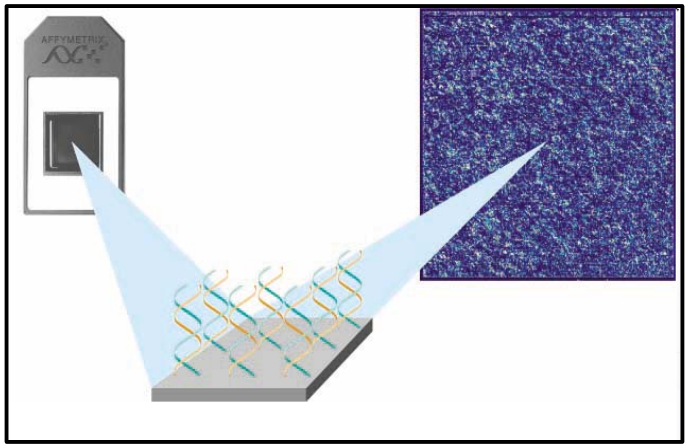


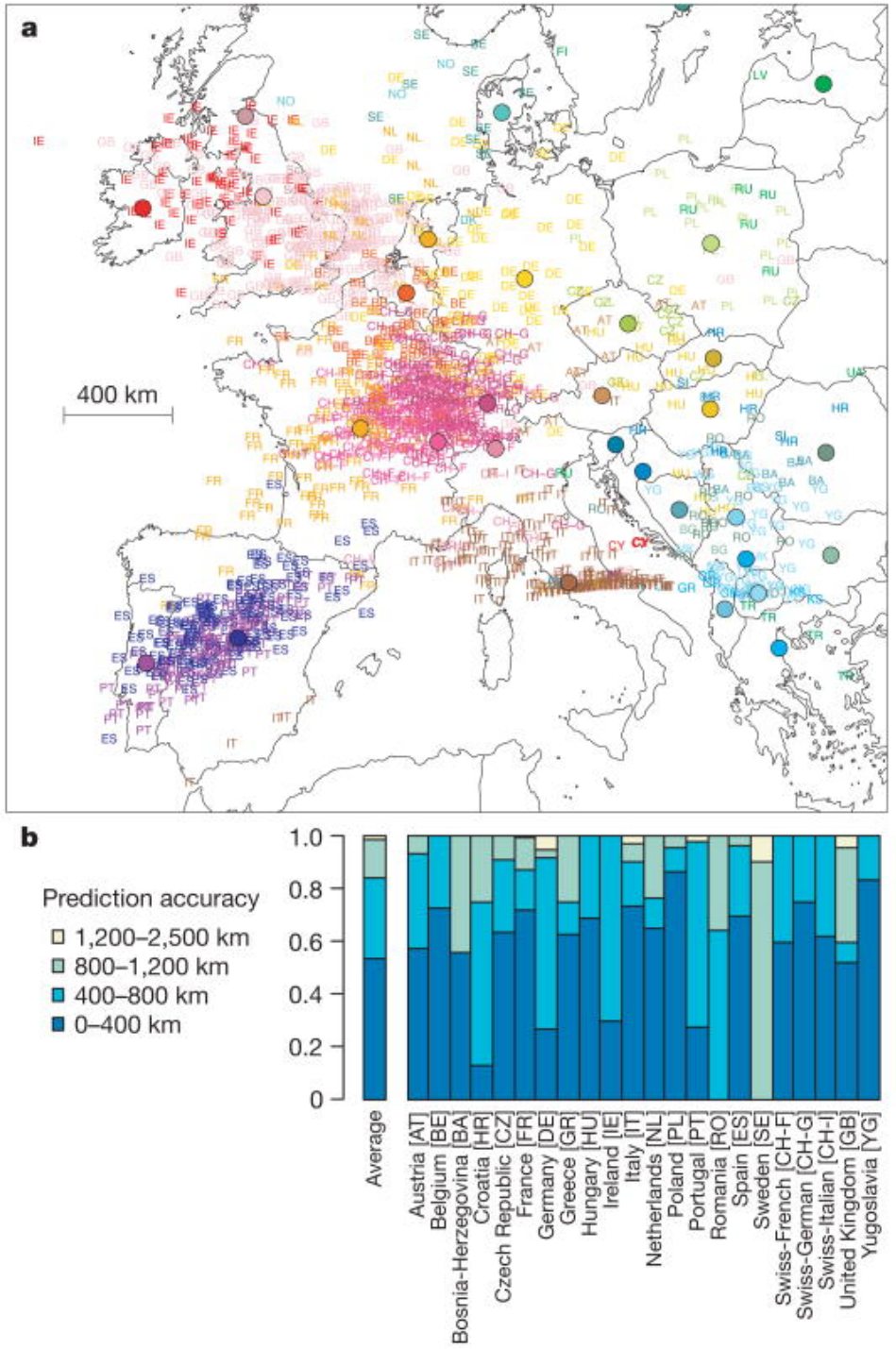


Loehlin, J. C. *Genes and Environment in Personality Development.* Newbury Park, CA: Sage, 1992.

Bienvenu OJ, Davydow DS, & Kendler KS (2011). *Psychological medicine, 41* (1), 33-40 PMID:

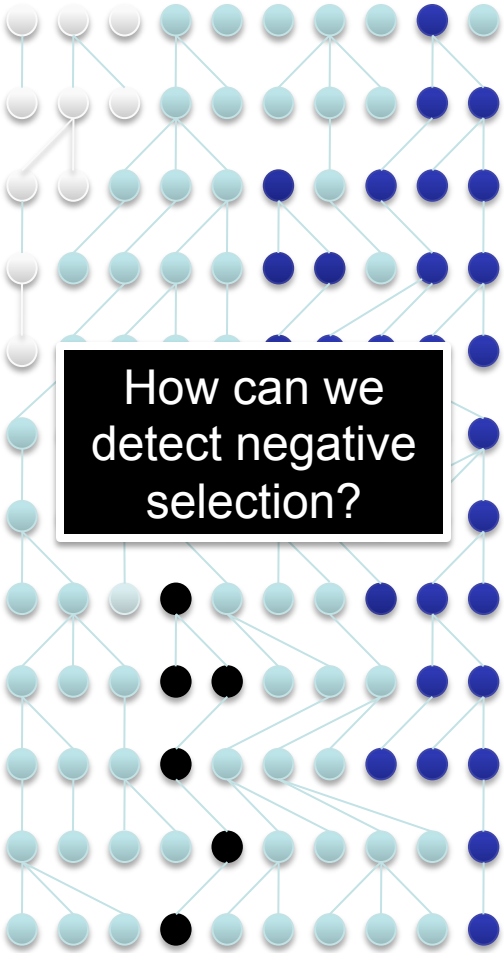
Global Ancestry Infere



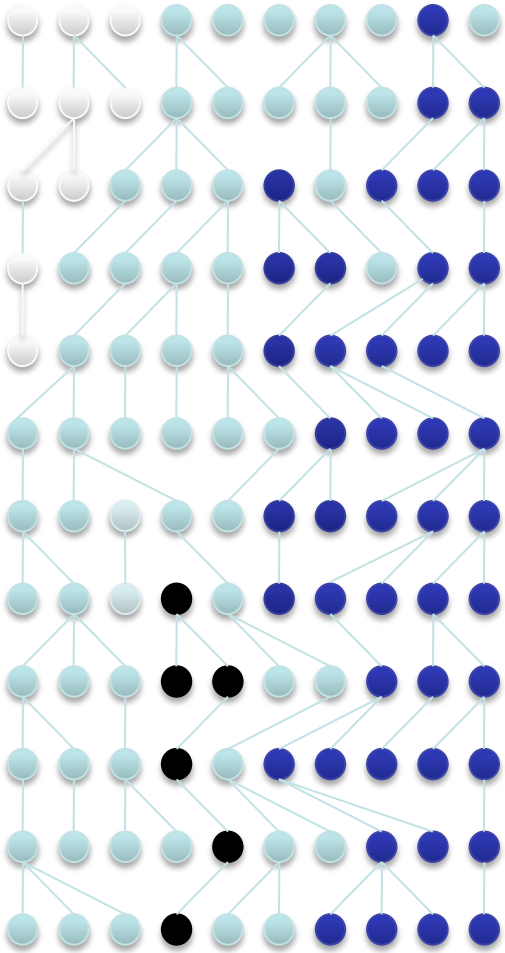


Nature. 2008 November 6; 456(7218): 98–101.

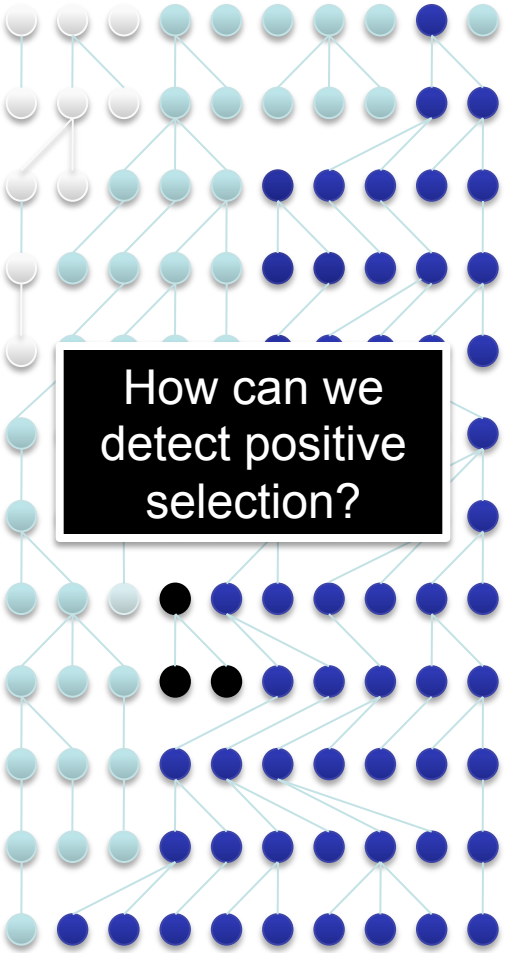
Fixation, Positive & Negative Selection



Negative Selection



Neutral Drift

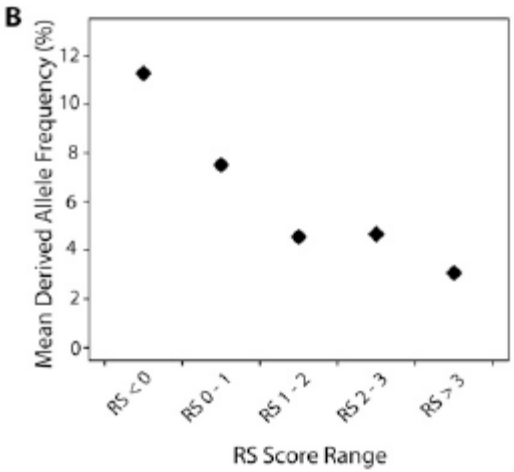


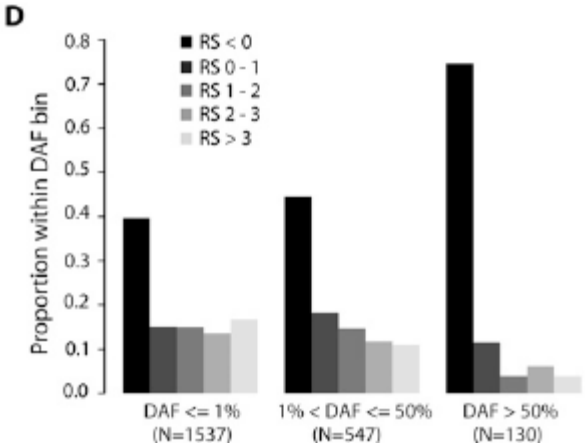
Positive Selection

Conservation and Human SNPs

CNSs have fewer SNPs

SNPs have shifted allele frequency spectra





How can we detect positive selection?

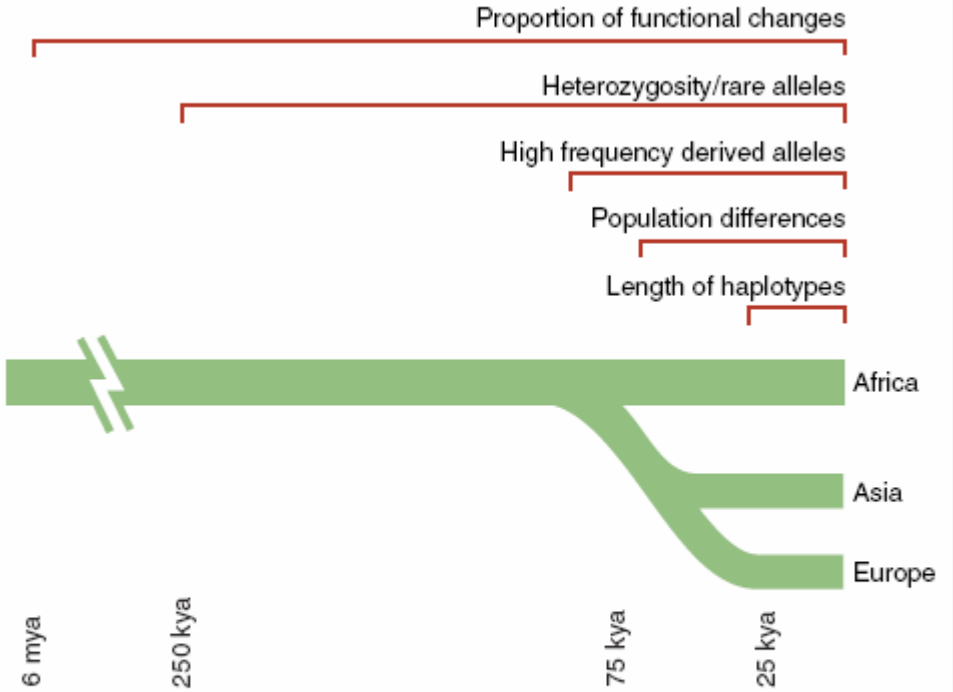


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

Ka/Ks ratio:
Ratio of nonsynonymous to synonymous substitutions

Very old, persistent, strong positive selection for a protein that keeps adapting

Examples: immune response, spermatogenesis

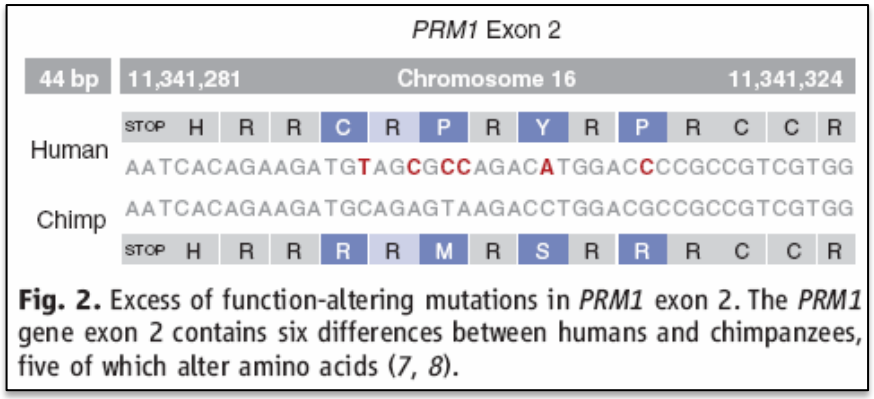


Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

How can we detect positive selection?

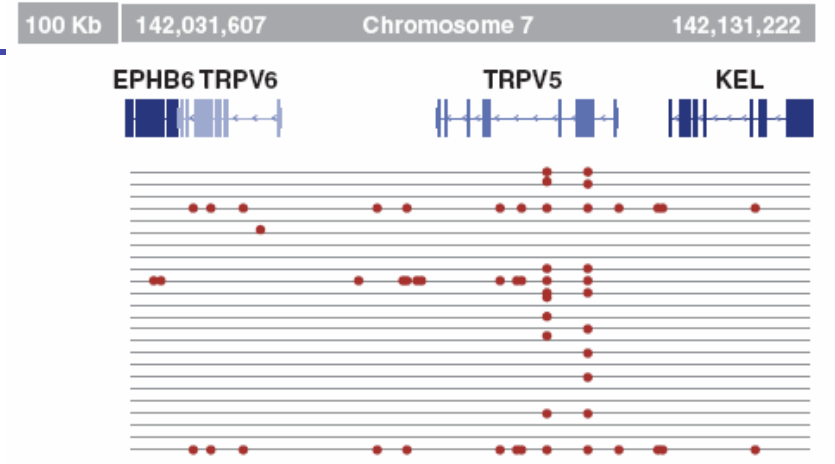


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

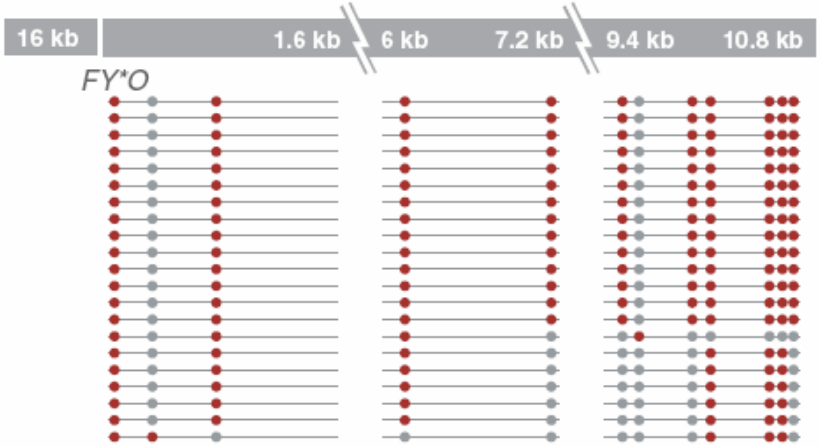


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (*34*). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

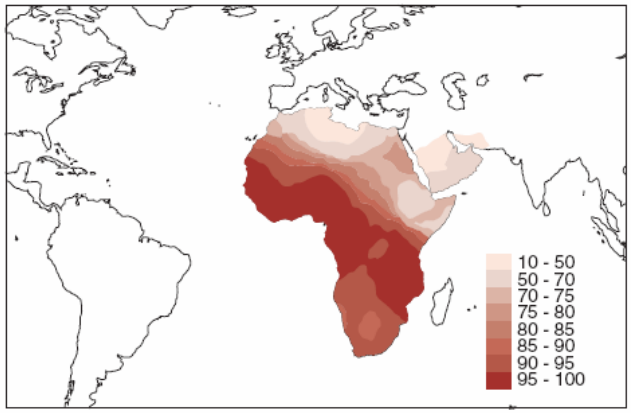


Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (*38*).

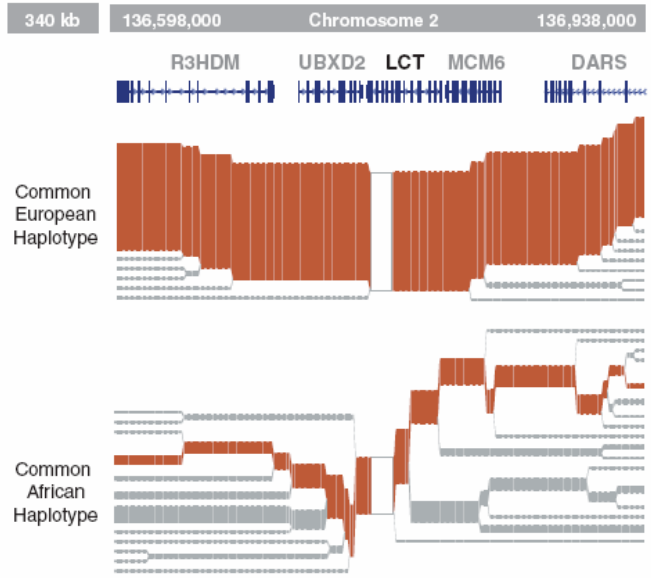
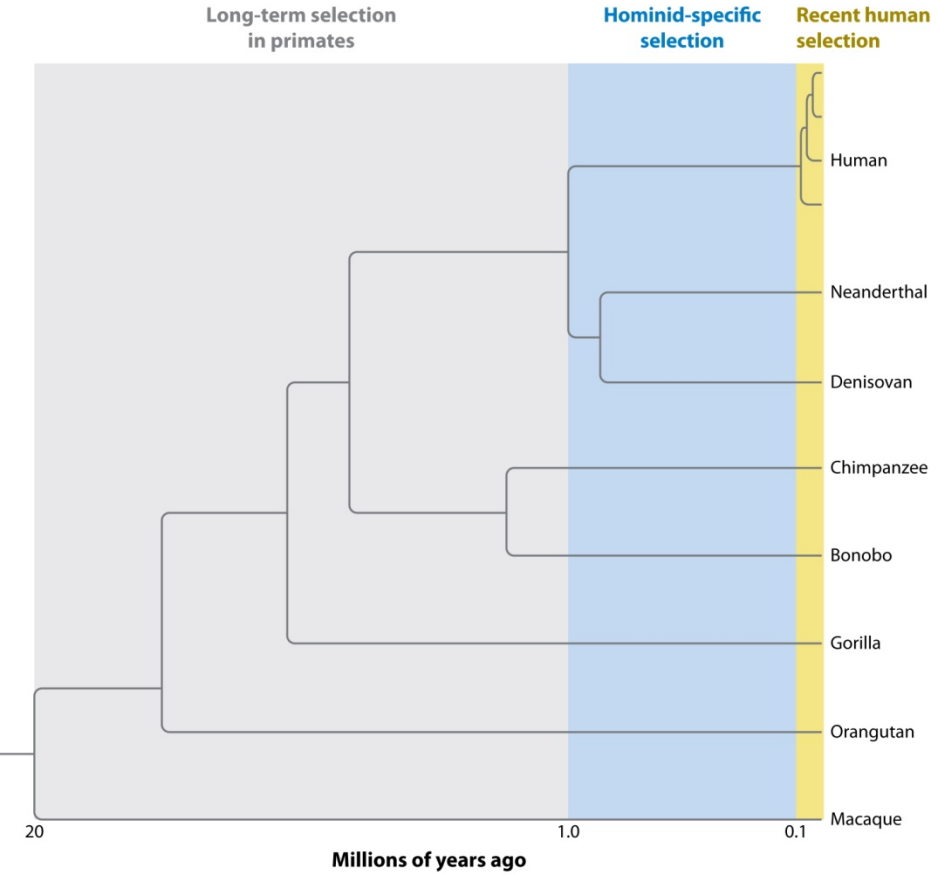
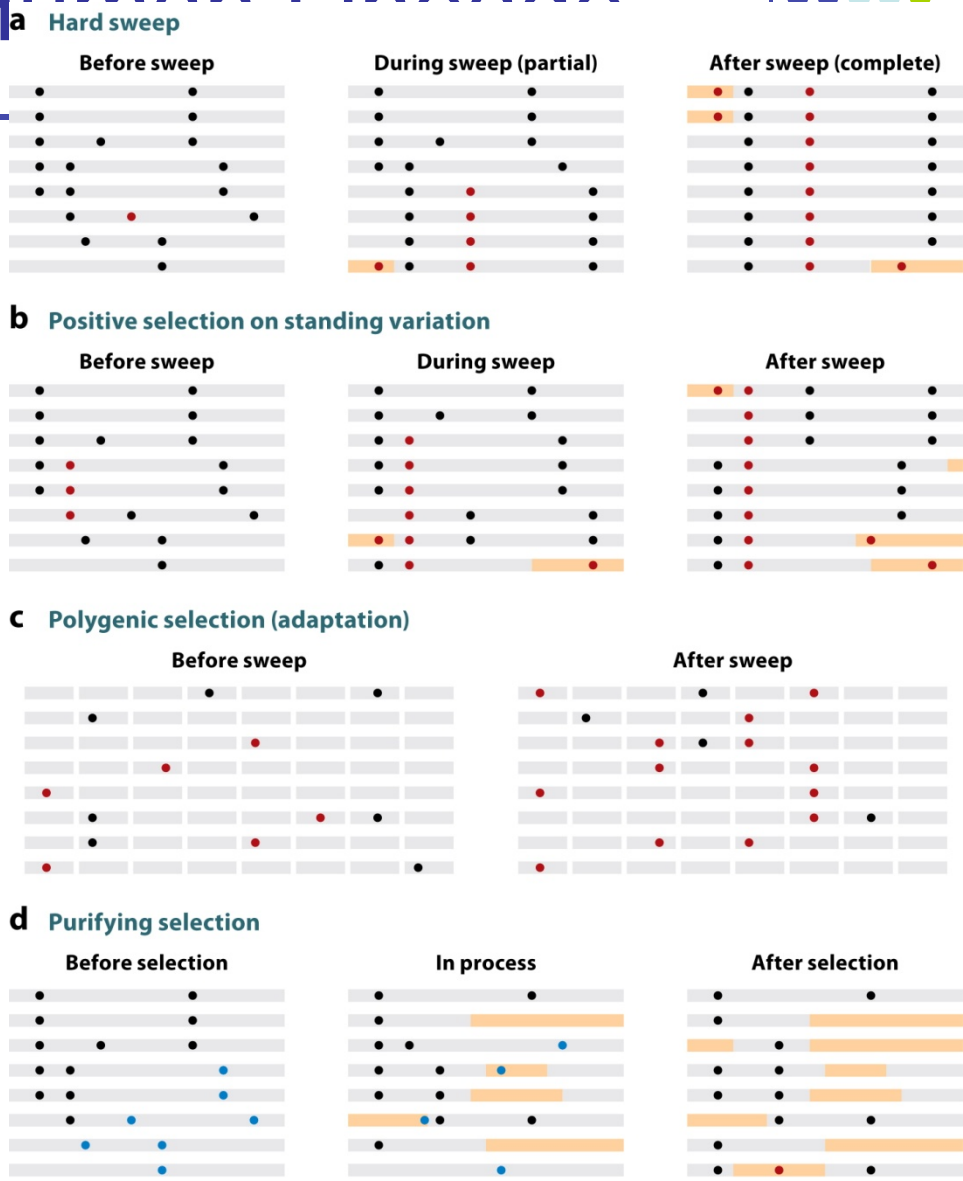


Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (*6*).

Positive Selection in Human Lineage

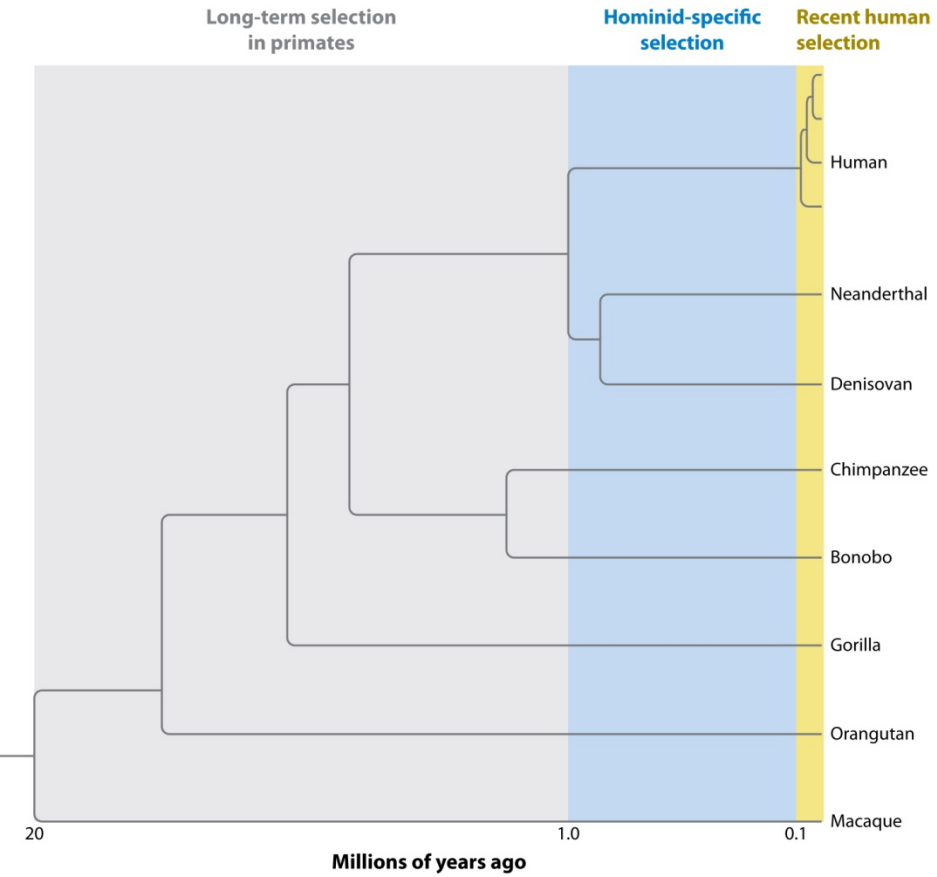


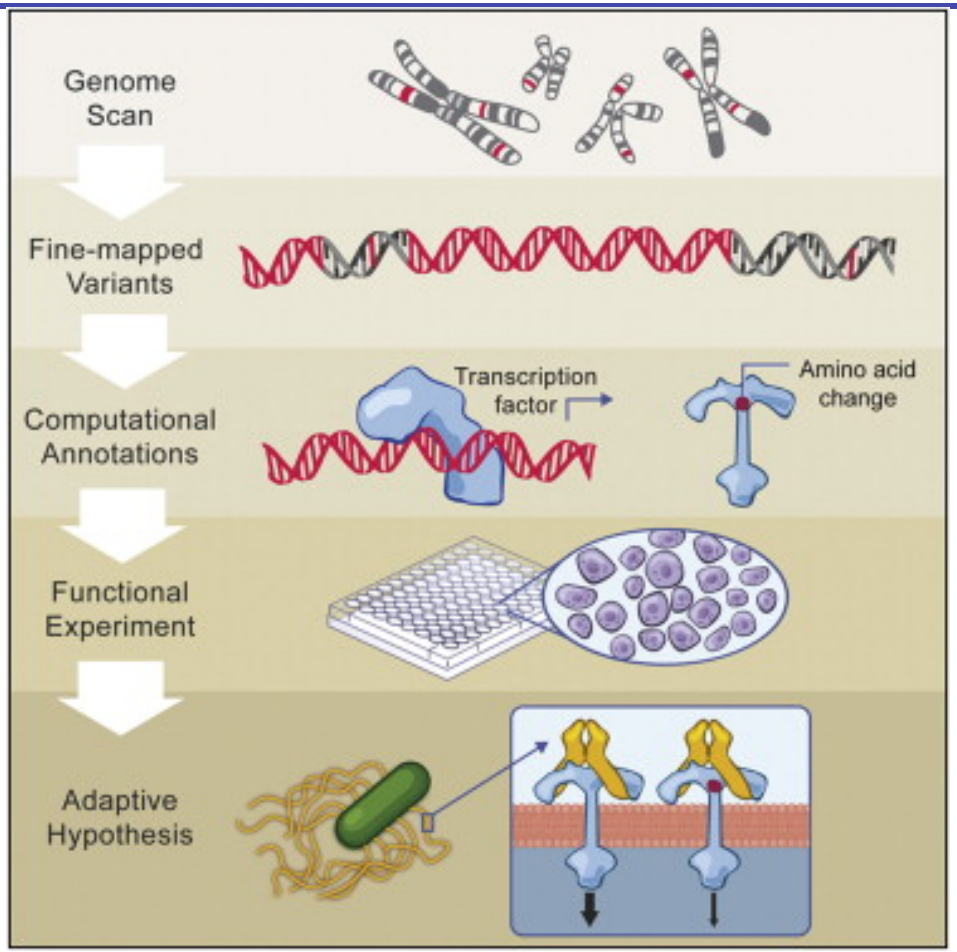
Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89



Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

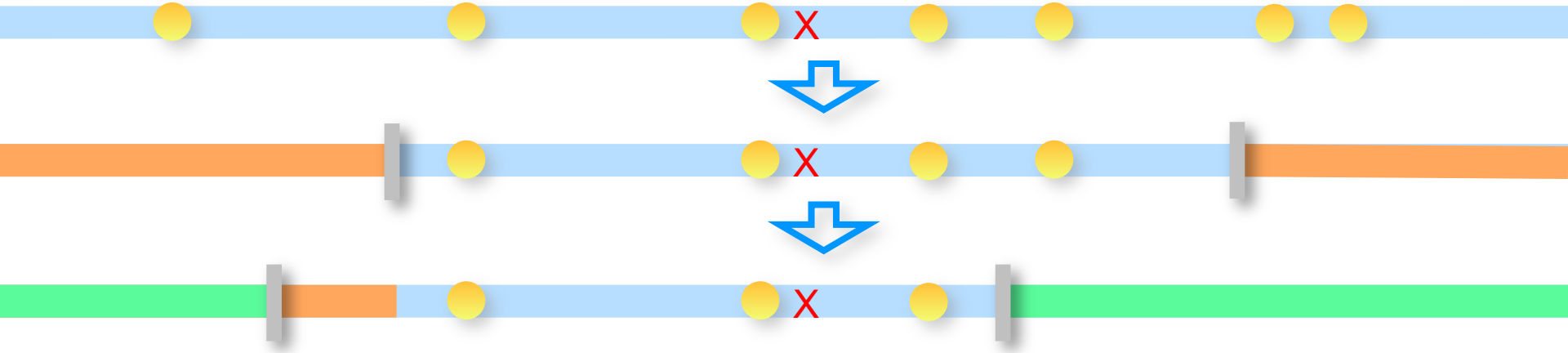
Positive Selection in Human Lineage





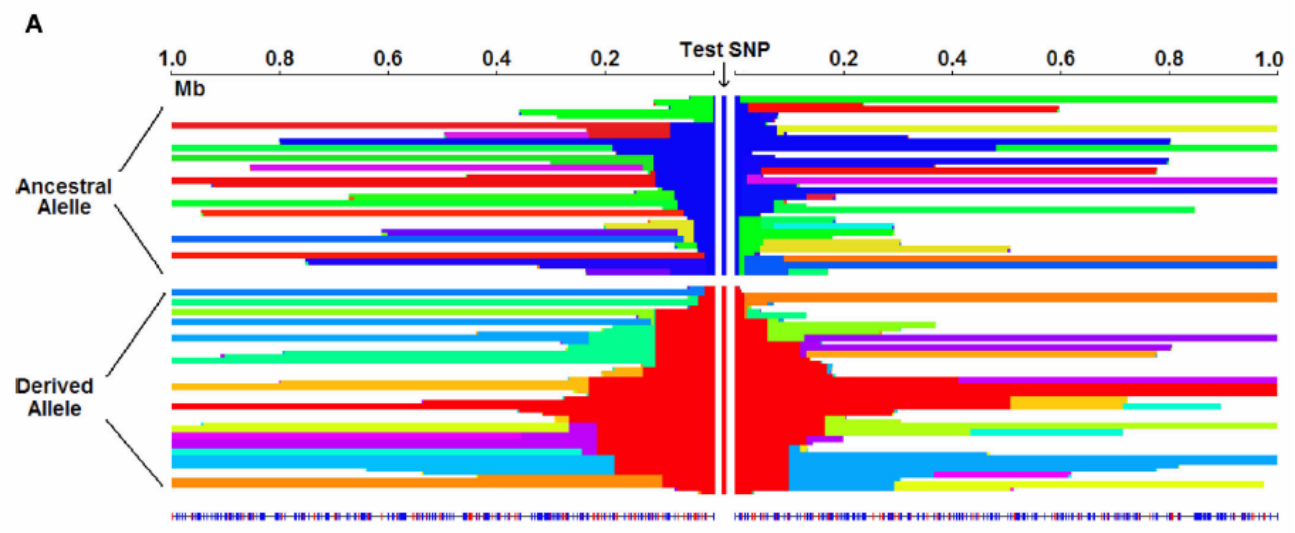
Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

Mutations and LD

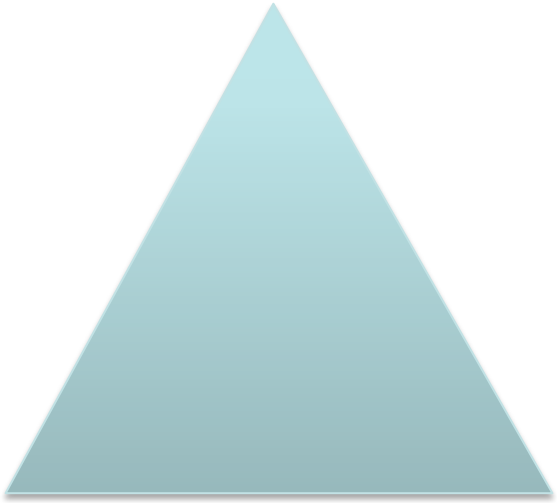


Slide Credits:
Marc Schaub

Long Haplotypes –iHS test



$$iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$



Less time:

- Fewer mutations
- Fewer recombinations

Application: Malaria

- Study of genes known to be implicated in the resistance to malaria.
- Infectious disease caused by protozoan parasites of the genus *Plasmodium*
- Frequent in tropical and subtropical regions
- Transmitted by the *Anopheles* mosquito

Image source: wikipedia.org

Slide Credits:
Marc Schaub

Application: Malaria

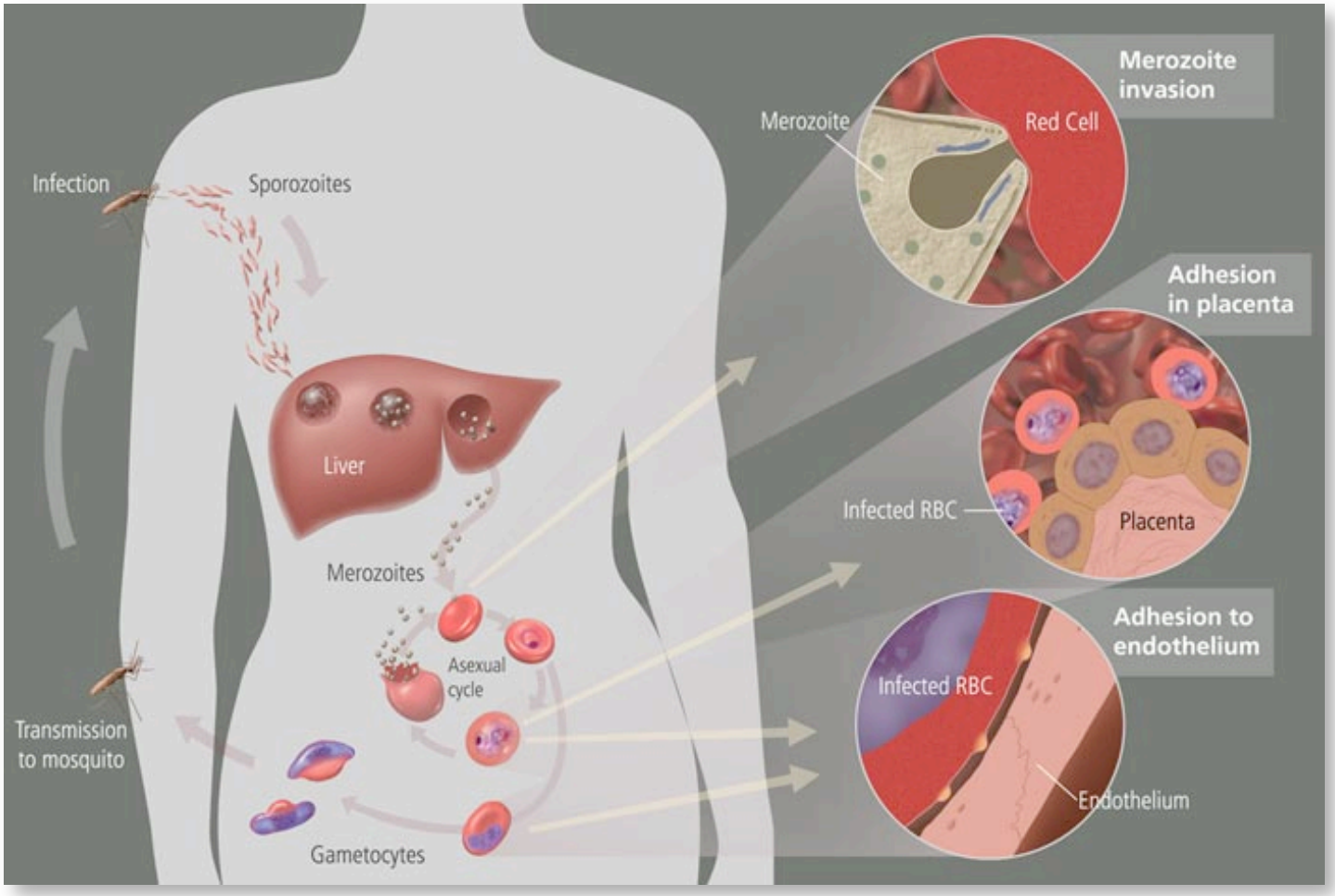


Image source:
NIH - http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg

Slide Credits:
Marc Schaub

Application: Malaria

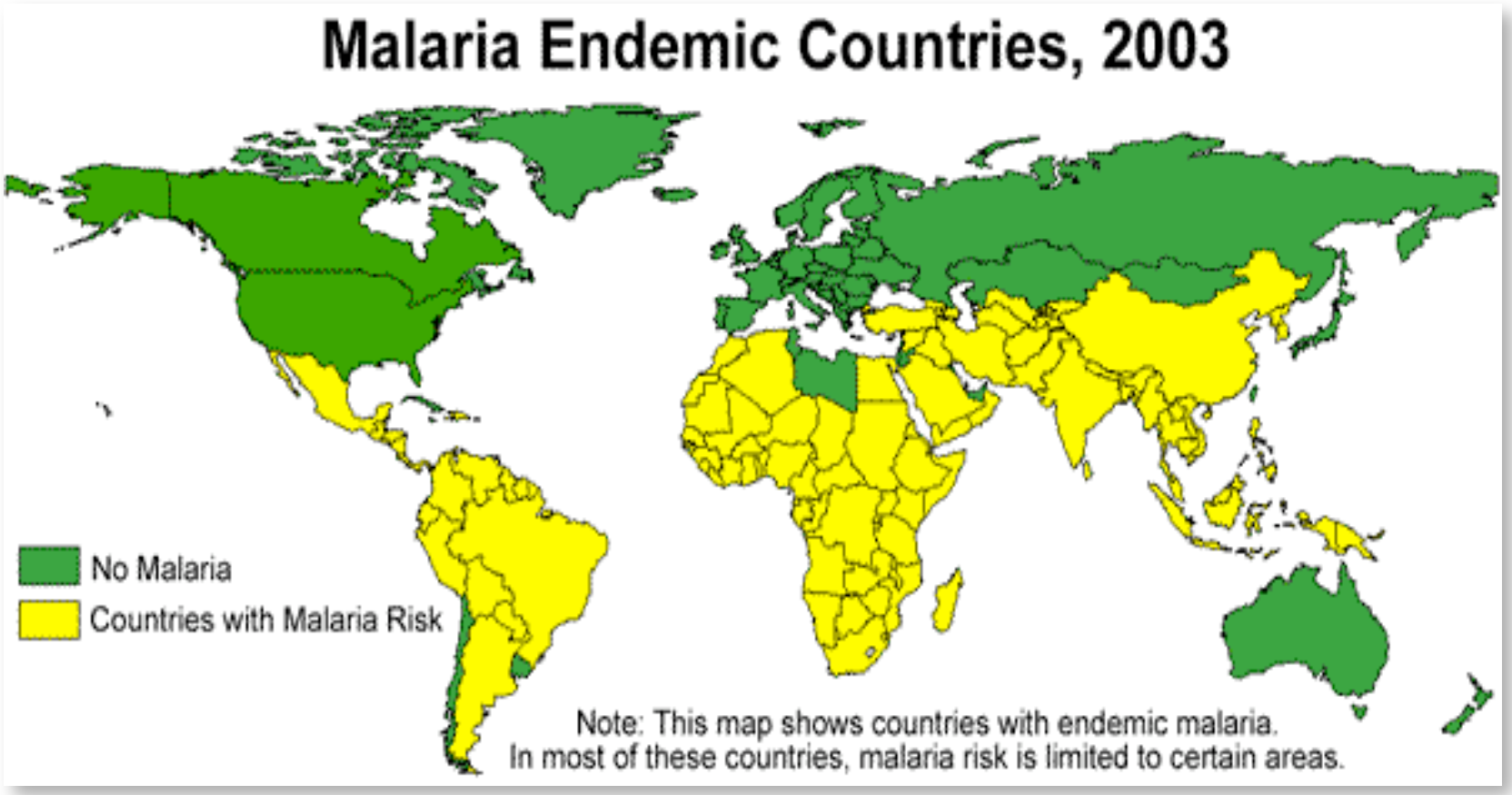
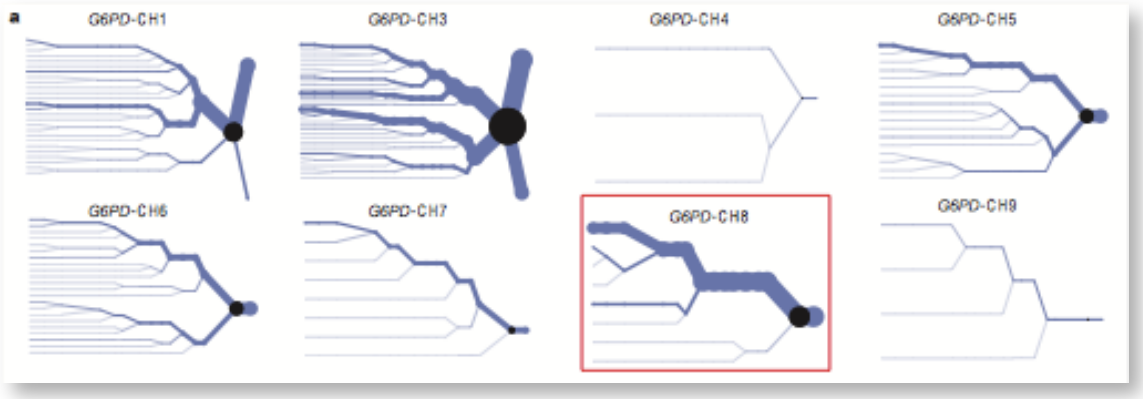
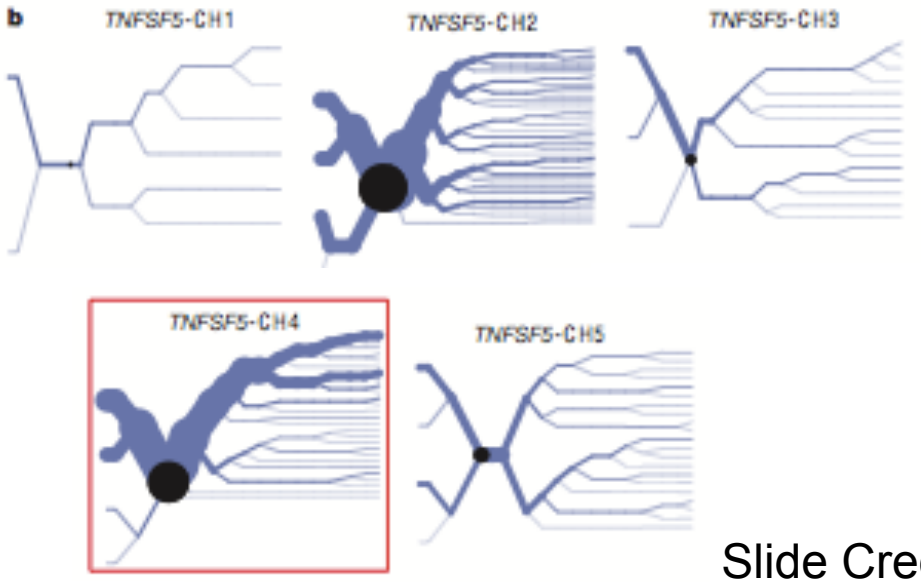


Image source: CDC - http://www.dpd.cdc.gov/dpdx/images/ParasiteImages/M-R/Malaria/malaria_risk_2003.gif

Slide Credits:
Marc Schaub

Results: G6PD, TNFSF5

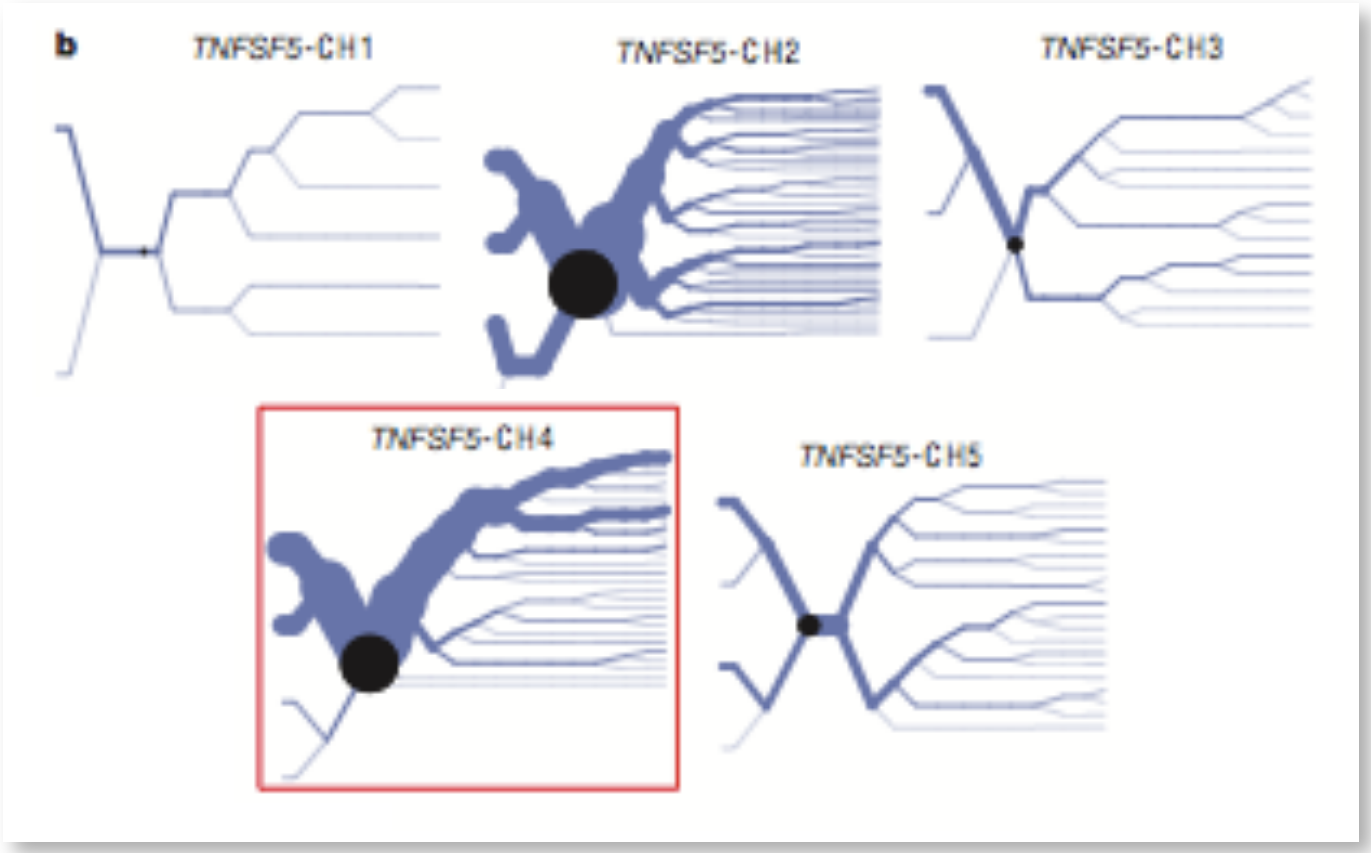




Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

Results: TNFSF5

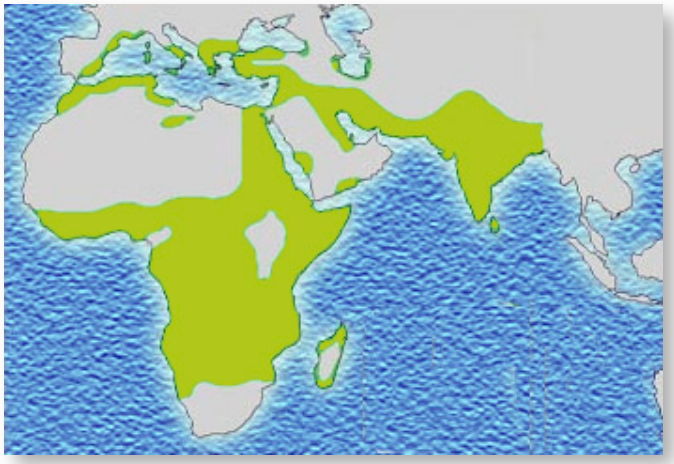


Source: Sabeti *et al.* Nature 2002.

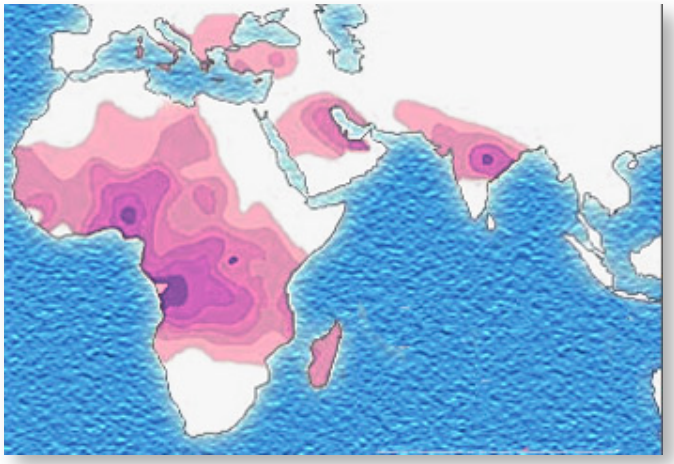
Slide Credits:
Marc Schaub

Malaria and Sickle-cell Anemia

- Allison (1954): Sickle-cell anemia is limited to the region in Africa in which malaria is endemic.



Distribution of malaria



Distribution of sickle-cell anemia

Image source: wikipedia.org

Slide Credits:
Marc Schaub

Lactose Intolerance

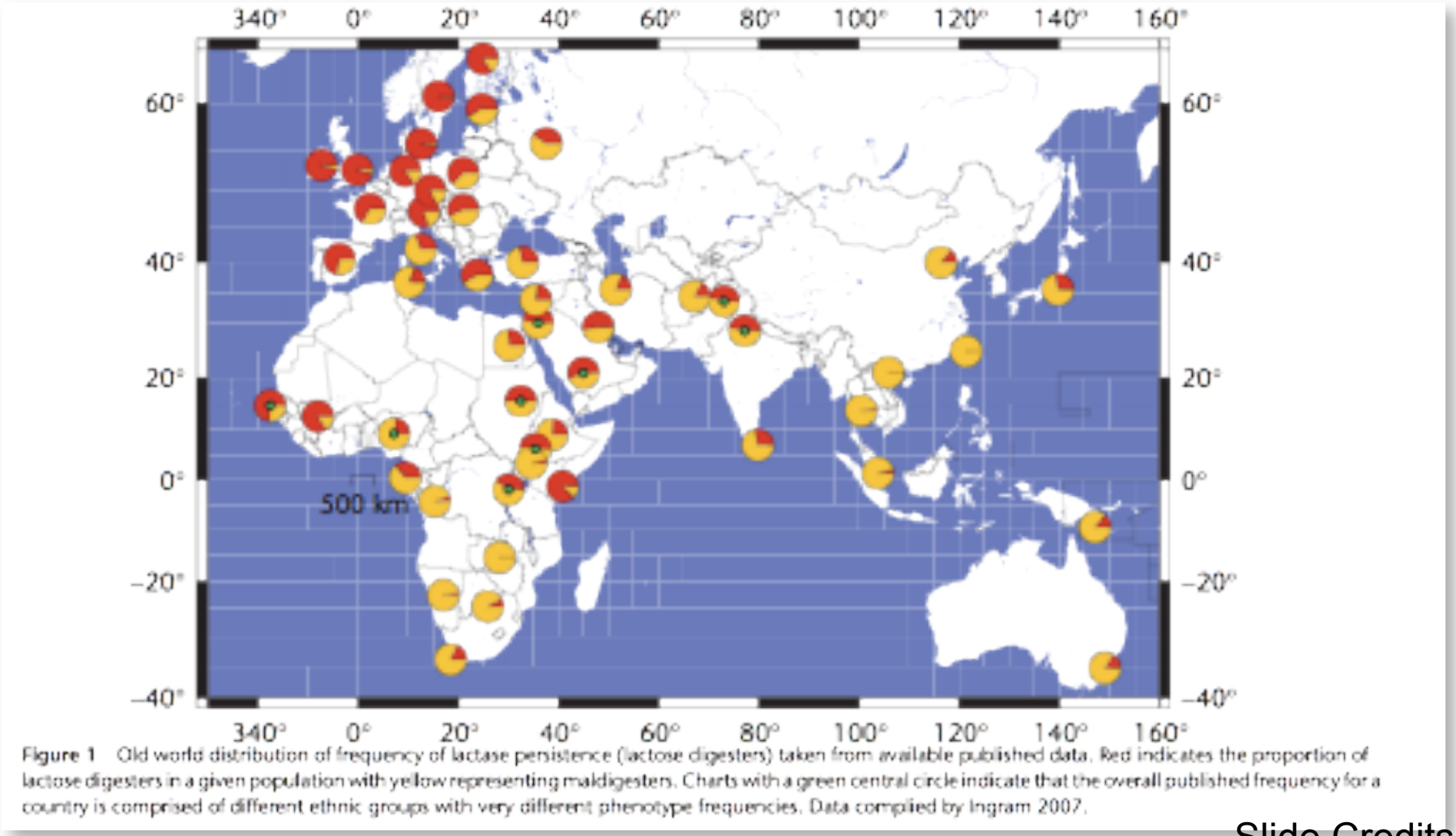
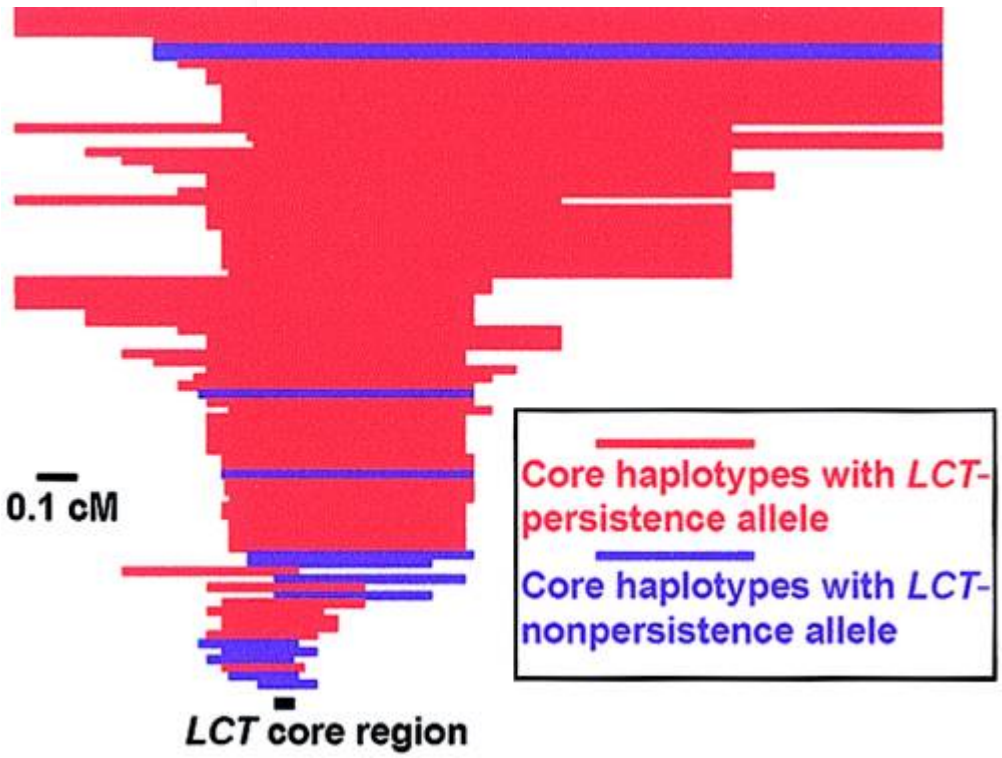


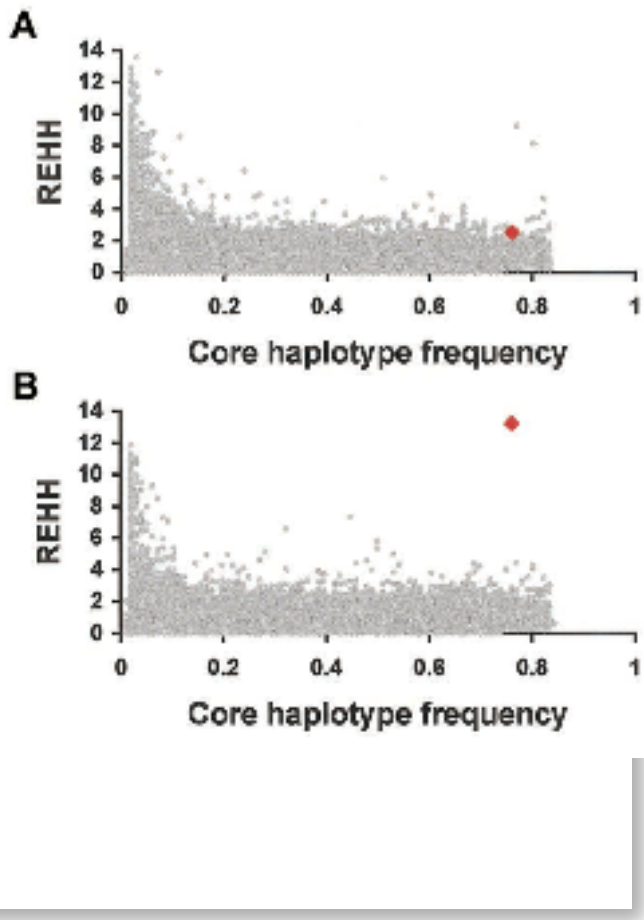
Figure 1 Old world distribution of frequency of lactase persistence (lactose digesters) taken from available published data. Red indicates the proportion of lactose digesters in a given population with yellow representing maldigesters. Charts with a green central circle indicate that the overall published frequency for a country is comprised of different ethnic groups with very different phenotype frequencies. Data compiled by Ingram 2007.

Source: Ingram and Swallow. Population Genetics of Encyclopedia of Life Sciences. 2007.

Slide Credits: Marc Schaub

Lactose Intolerance





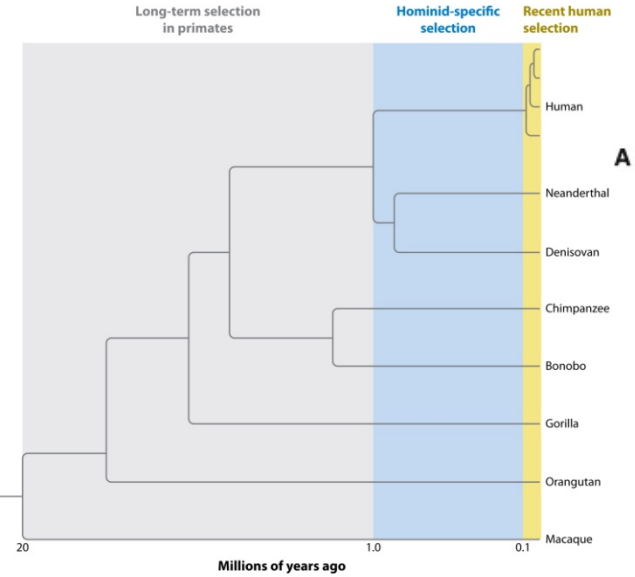
LCT, 5'

LCT, 3'

Source: Bersaglieri *et al.* Am. J. Hum. Genet. 2004.

Slide Credits:
Marc Schaub

Positive Selection in Human Lineage



Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89