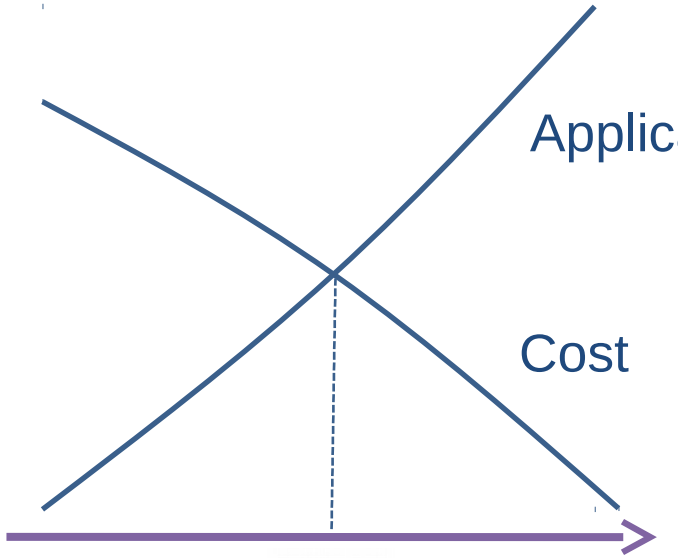
GACTGAGGAGTTTACGGGAGCAAAGCGGCGTCATTGCTATTCGTATCTGTTTAG

01010110001001010000101010101001101110011000110010100010101

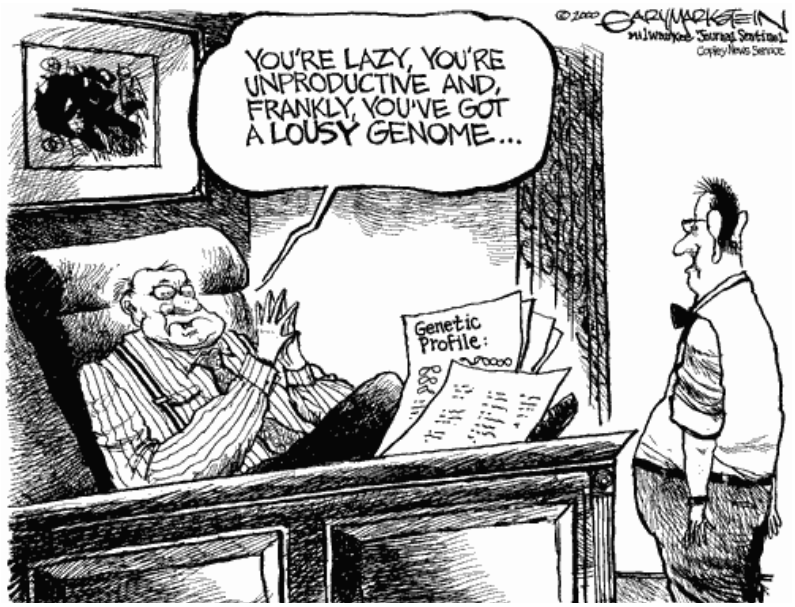
Human Population Genomics

How soon will we all be sequenced?

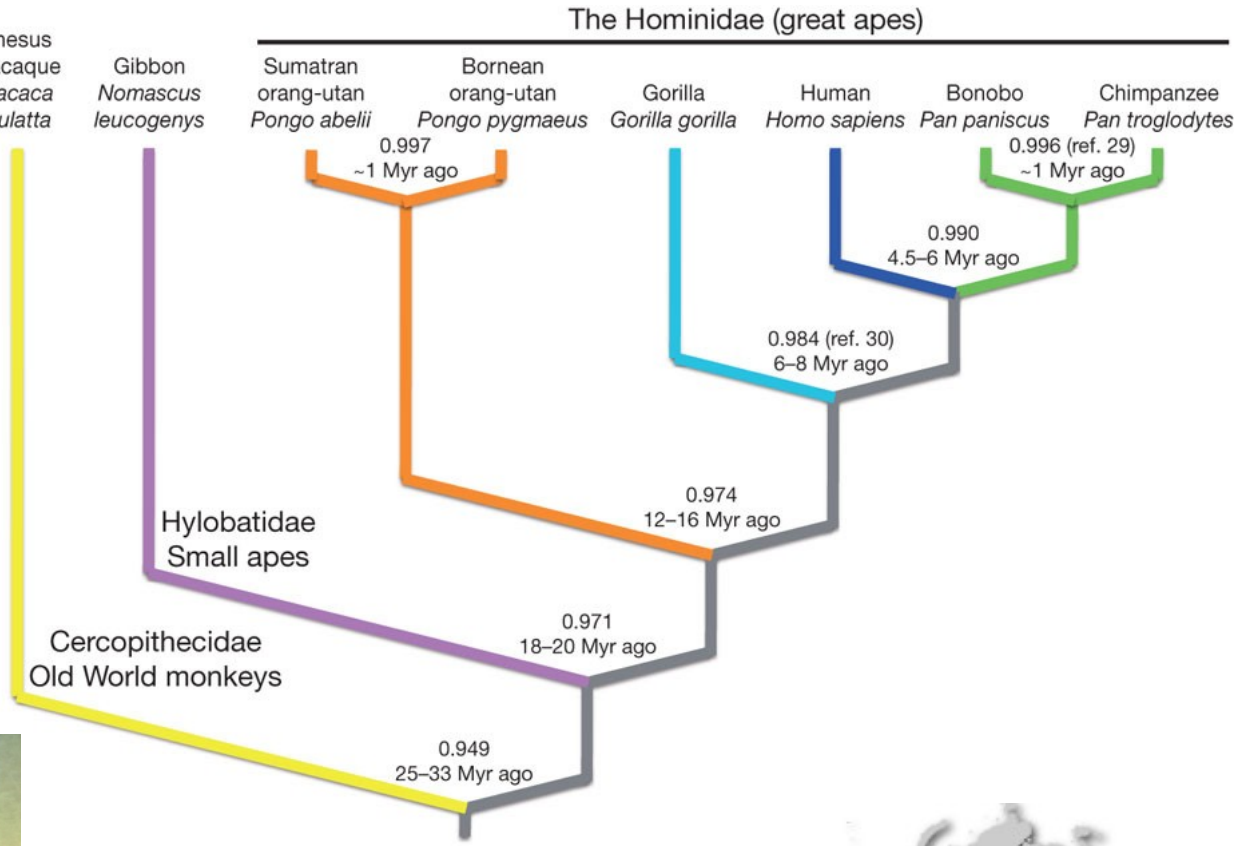
- Cost
- Killer apps
- Roadblocks?

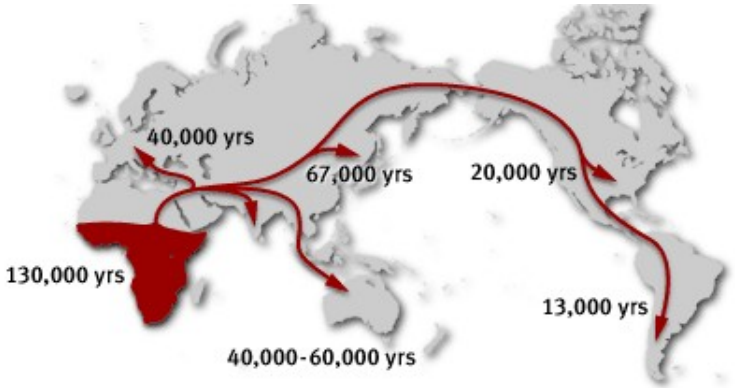


2015?
2020?

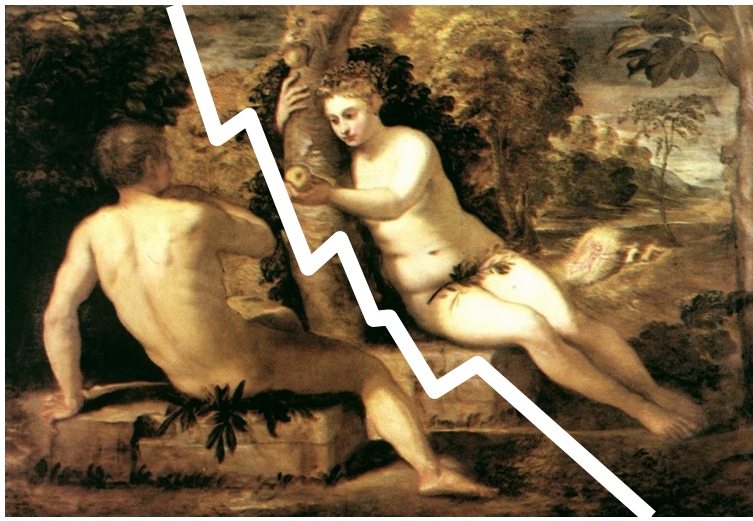


The Hominid Lineage



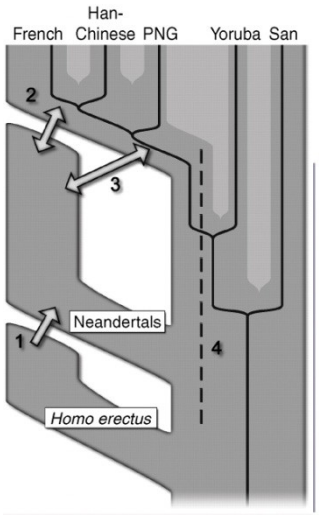


Human population migrations



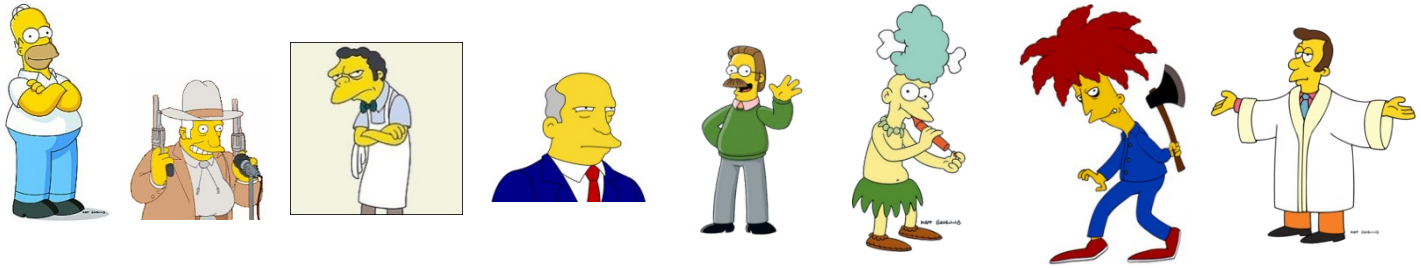
- Out of Africa, Replacement
  - Single mother of all humans (Eve) ~190,000yr
  - Single father of all humans (Adam) ~340,000yr
  - Humans out of Africa ~50000 years ago replaced others (e.g., Neandertals)

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



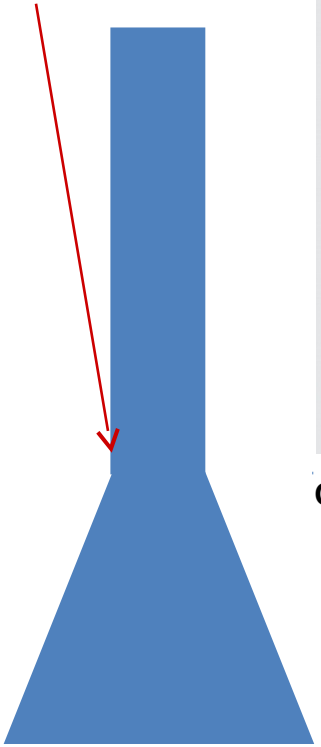
Coalescence

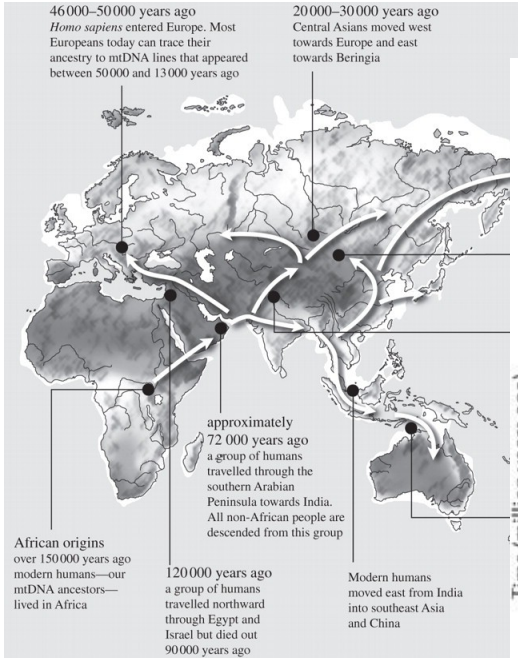
Y-chromosome coalescence



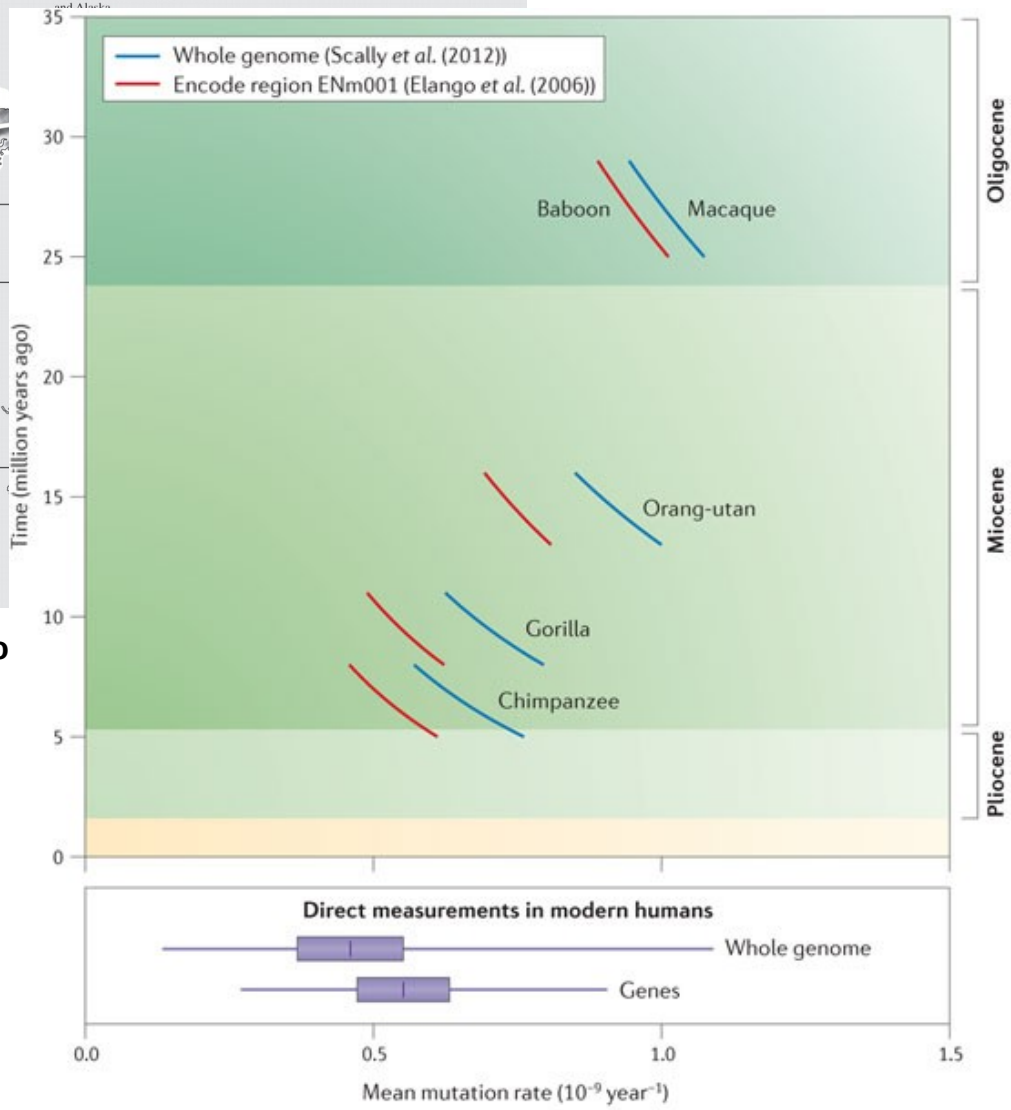
Why humans are so similar

Out of Africa





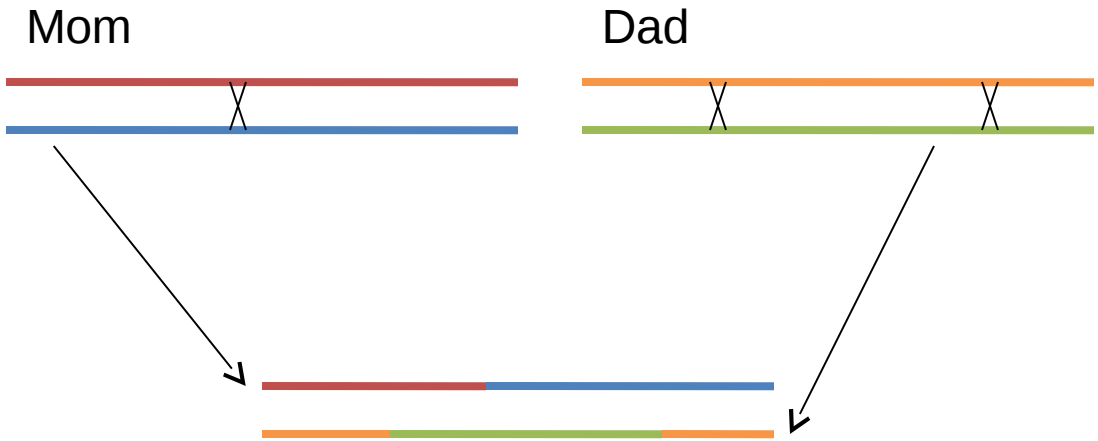
Oppenheimer S Phil. Trans. R. So



Nature Reviews | Genetics

Some Key Definitions

```
Mary: AGC( G/  ACG
John: AGC( G   ACG
Josh: AGC( G/  ACG
Kate: AGC( G   ACG
Pete: AGC( G/  ACG
Anne: AGC( T   ACG
Mimi: AGC( G/  ACG
Mike: AGC( G   ACG
Olga: AGC( G/  ACG
Tony: AGC( G   ACG
```



Alleles: G, T

Major Allele T

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

Human Genome Variation

SNP

TGCTGAGA
TGCCGAGA

Novel Sequence

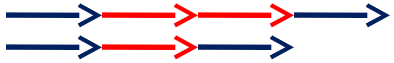
TGCTCGGAGA
TGC - - - GAGA

Inversion

Mobile Element or
Pseudogene Insertion

Translocation

Tandem Duplication



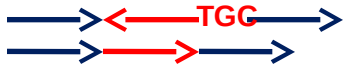
Microdeletion

TGC - - AGA
TGCCGAGA

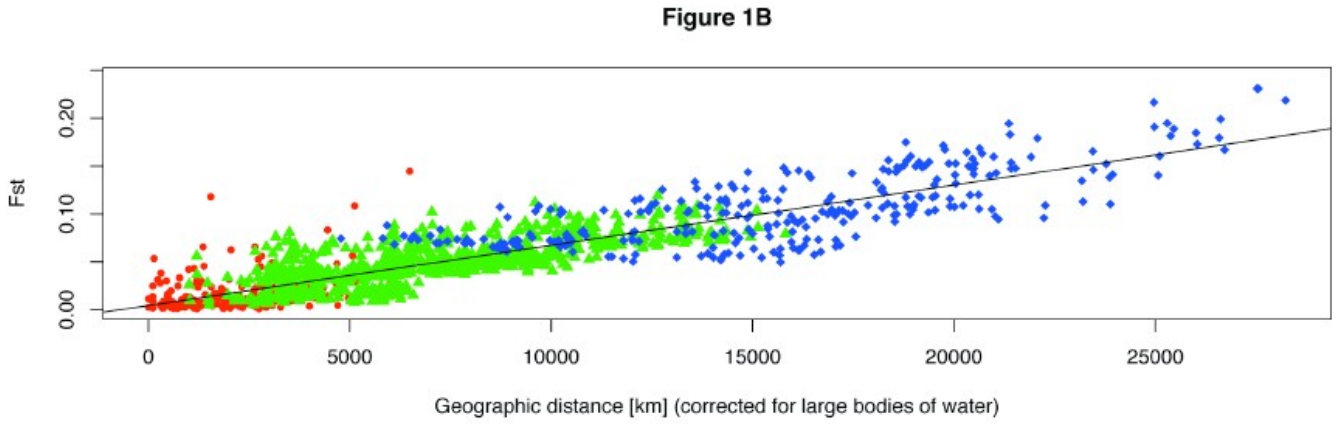
Transposition

Large Deletion

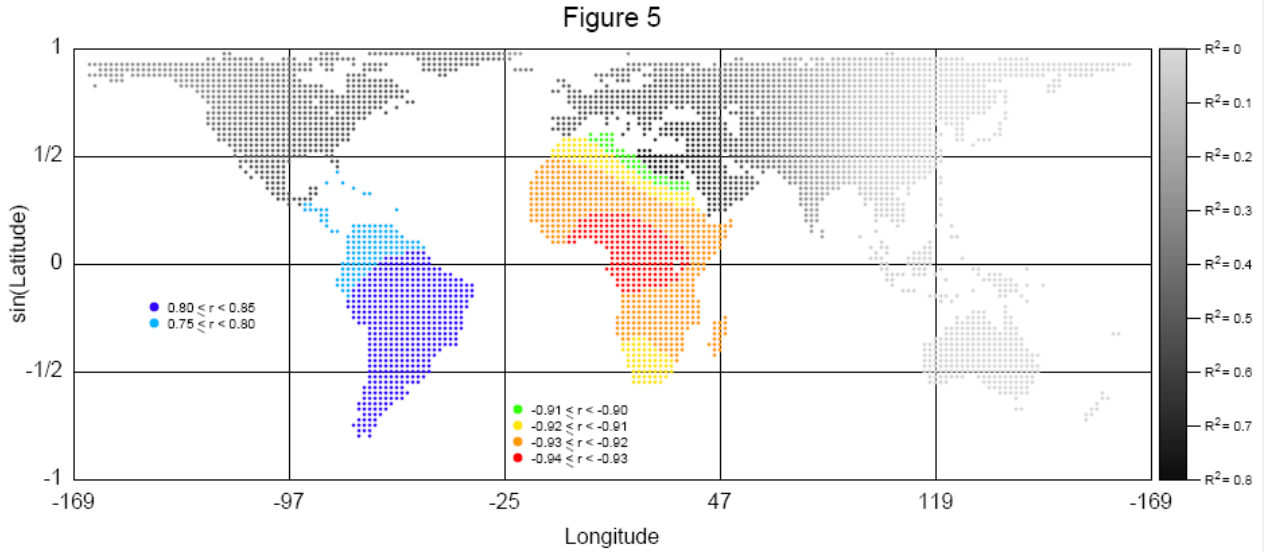
Novel Sequence
at Breakpoint



The Fall in Heterozygosity



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



The Neanderthal Genome

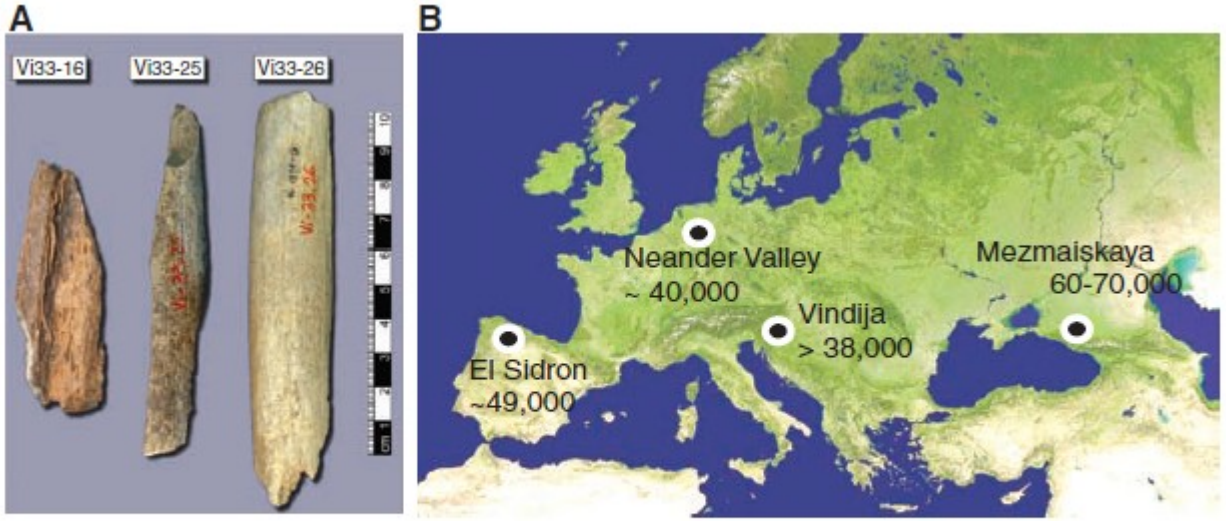


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

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

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

Neanderthal Genome

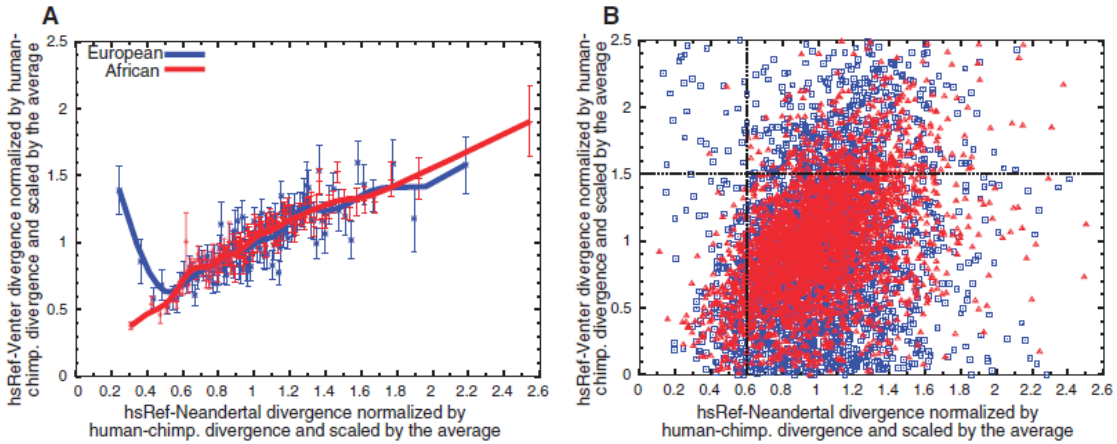


Fig. 5. Segments of Neandertal ancestry in the human reference genome. We examined 2825 segments in the human reference genome that are of African ancestry and 2797 that are of European ancestry. (**A**) European segments, with few differences from the Neandertals, tend to have many differences from 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.

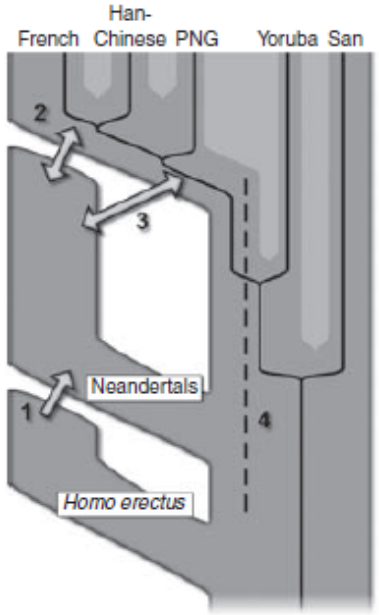
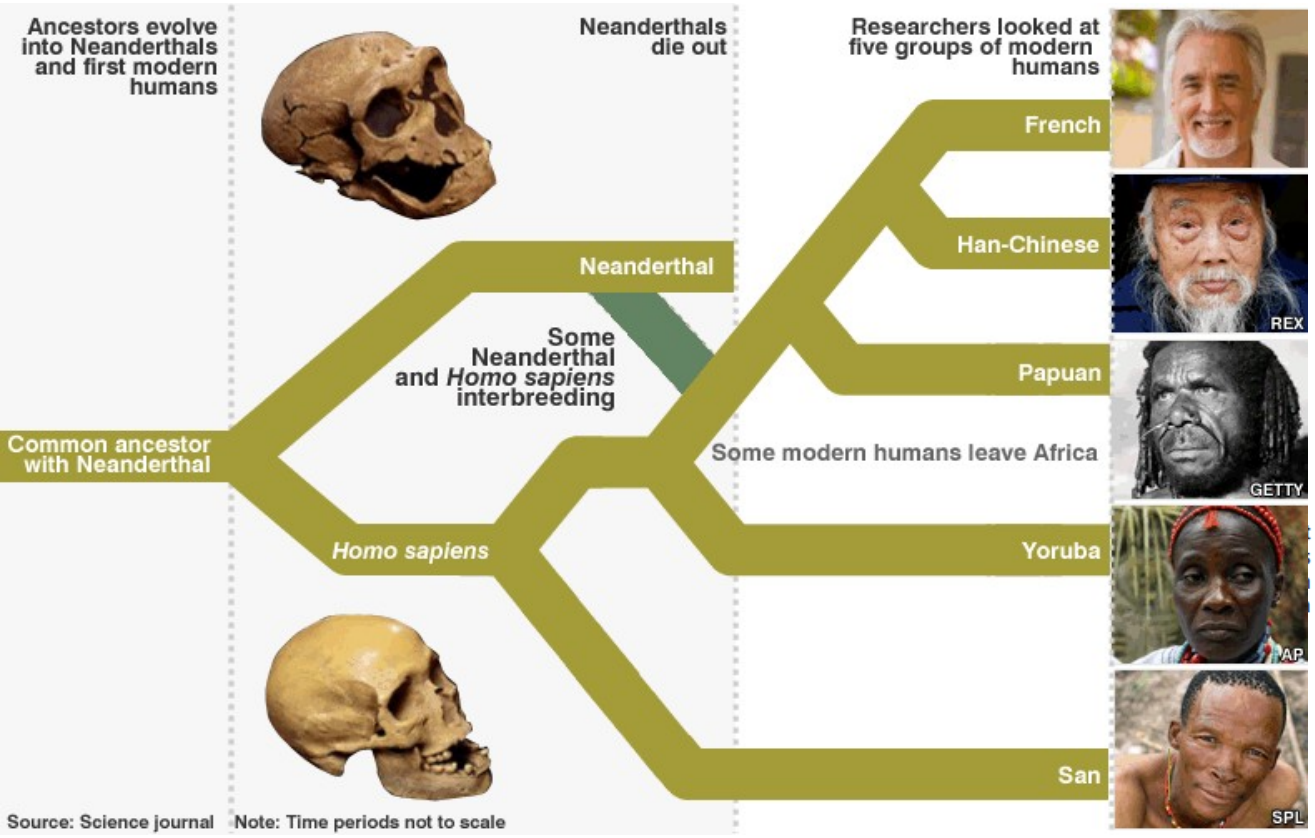


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from 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.

Neanderthal Genome



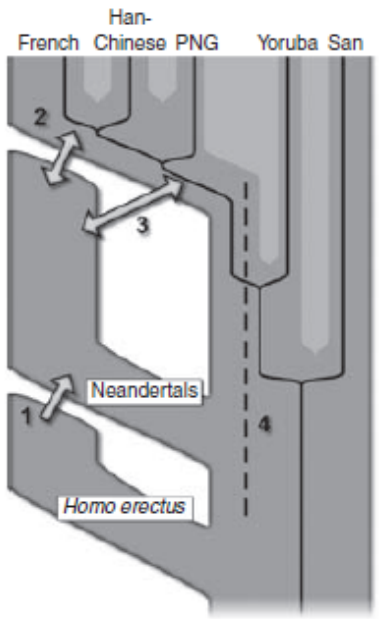


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from 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

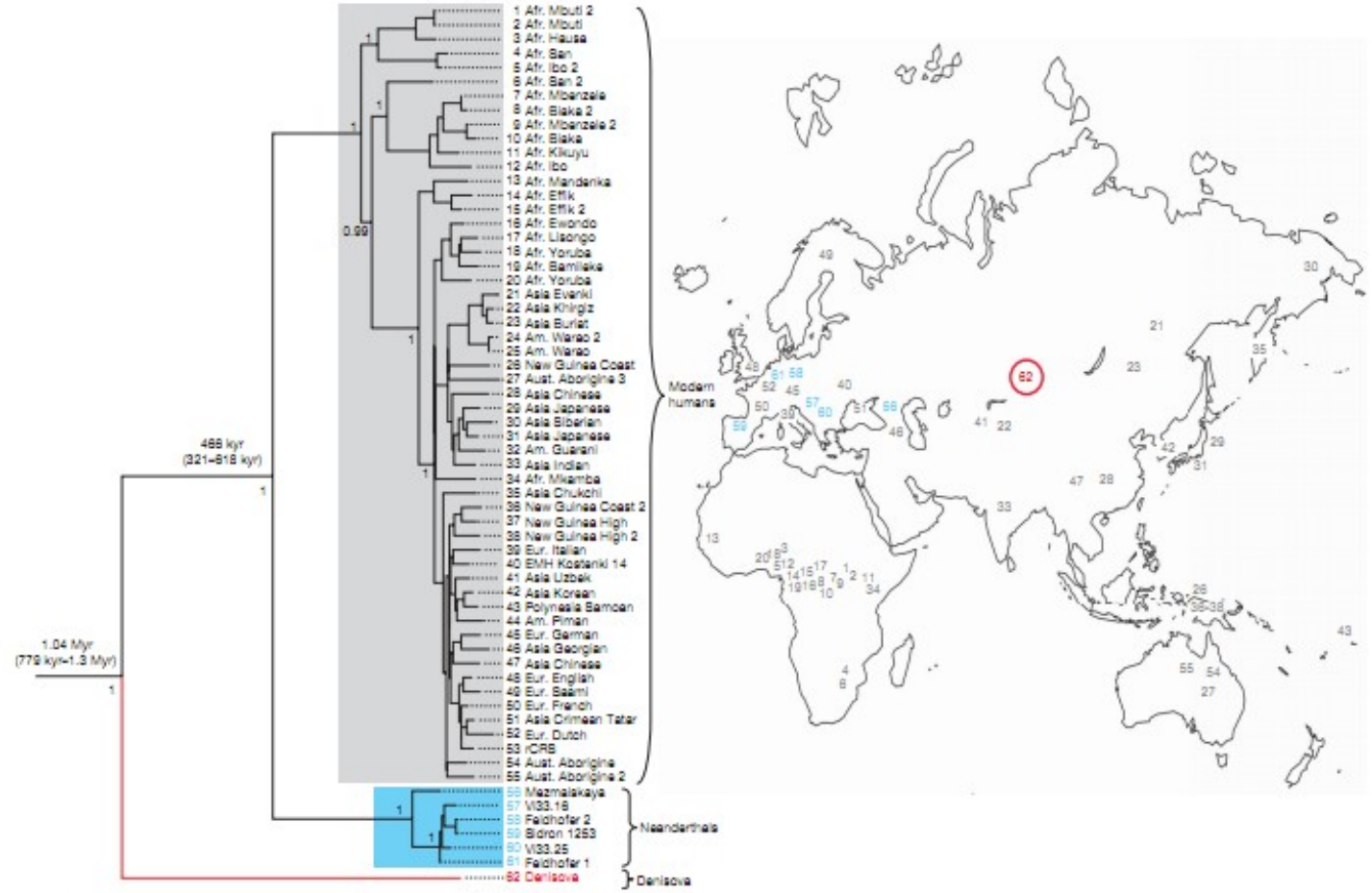
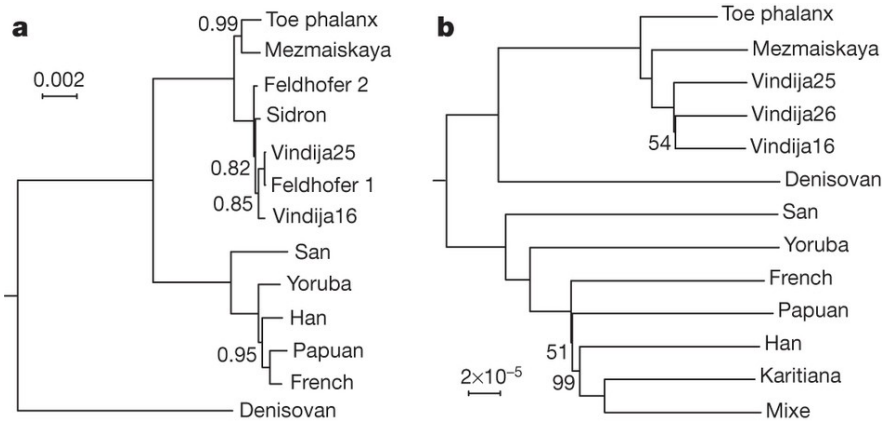
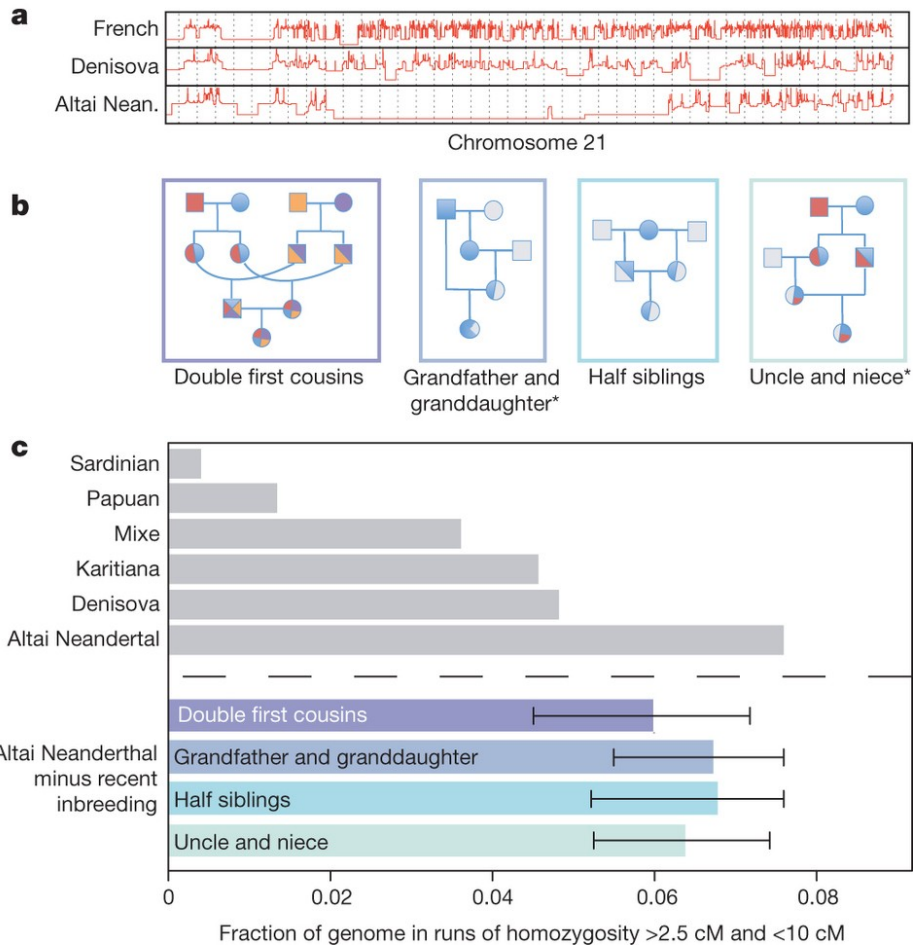


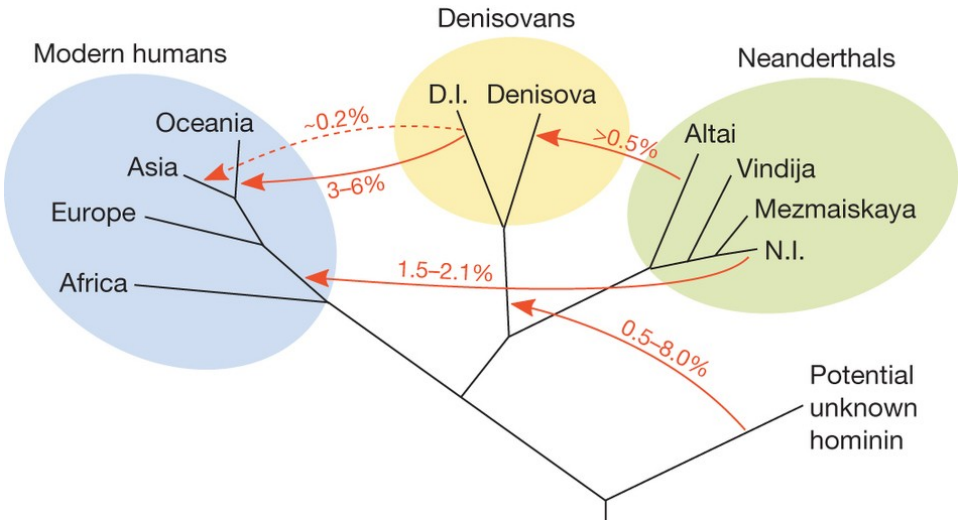
Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+1+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. 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

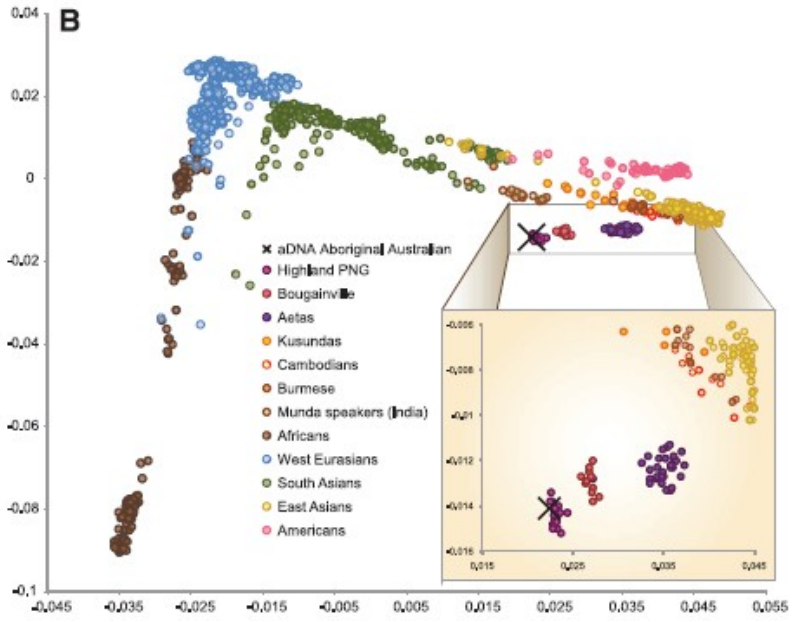


The Neanderthal Whole Genome





Aboriginal Australian



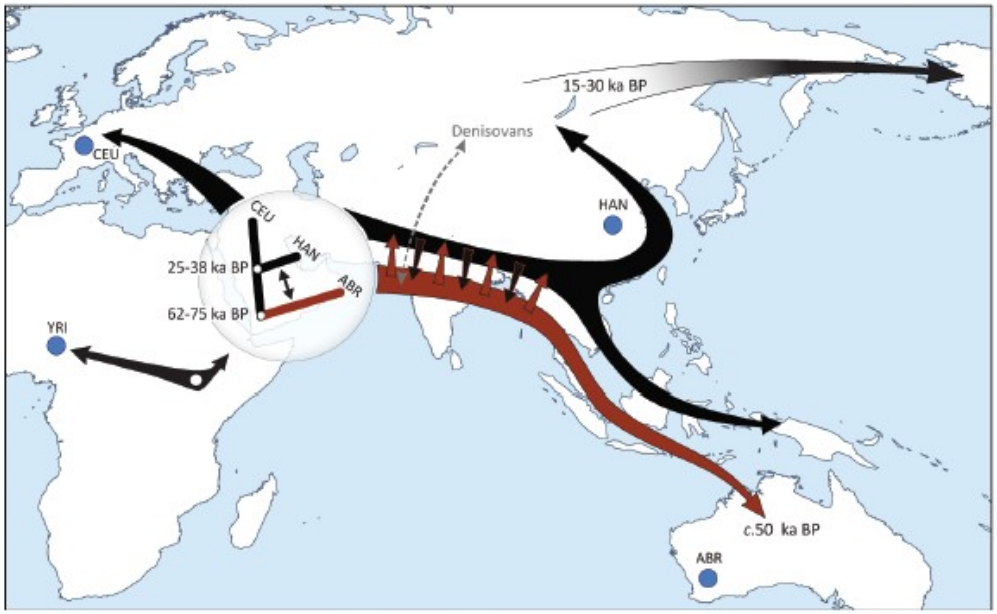
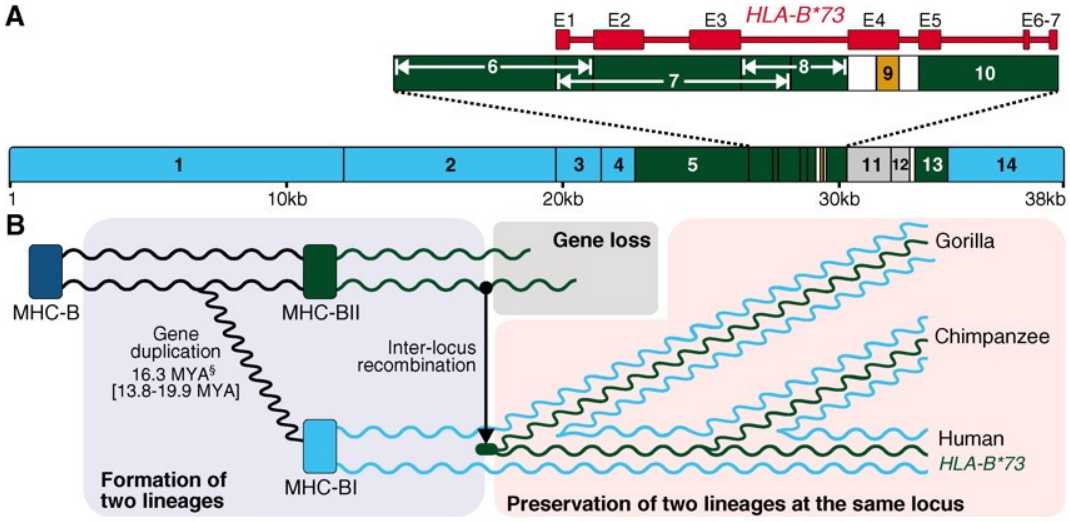
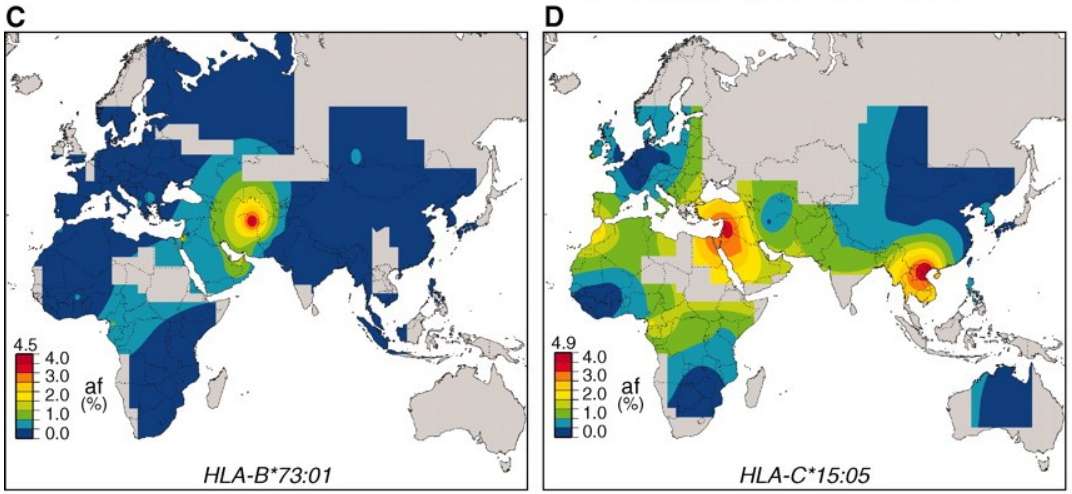


Fig. 2. Reconstruction of early spread of modern humans outside Africa. The tree shows the divergence of the Aboriginal Australian (ABR) relative to the CEPH European (CEU) and the Han Chinese (HAN) with gene flow between aboriginal Australasians and Asian ancestors. Purple arrow shows early spread of 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.

Benefits of Admixture

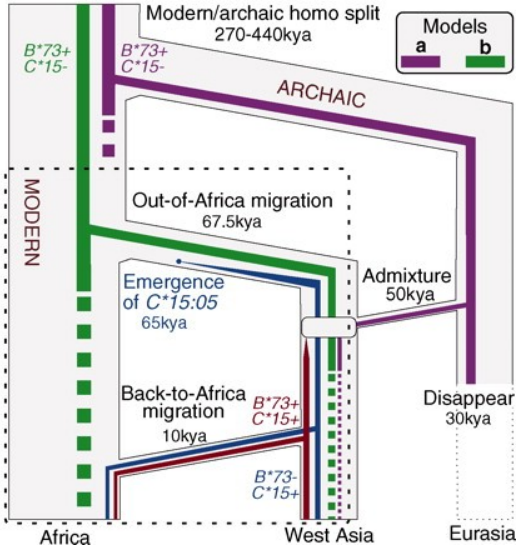




E

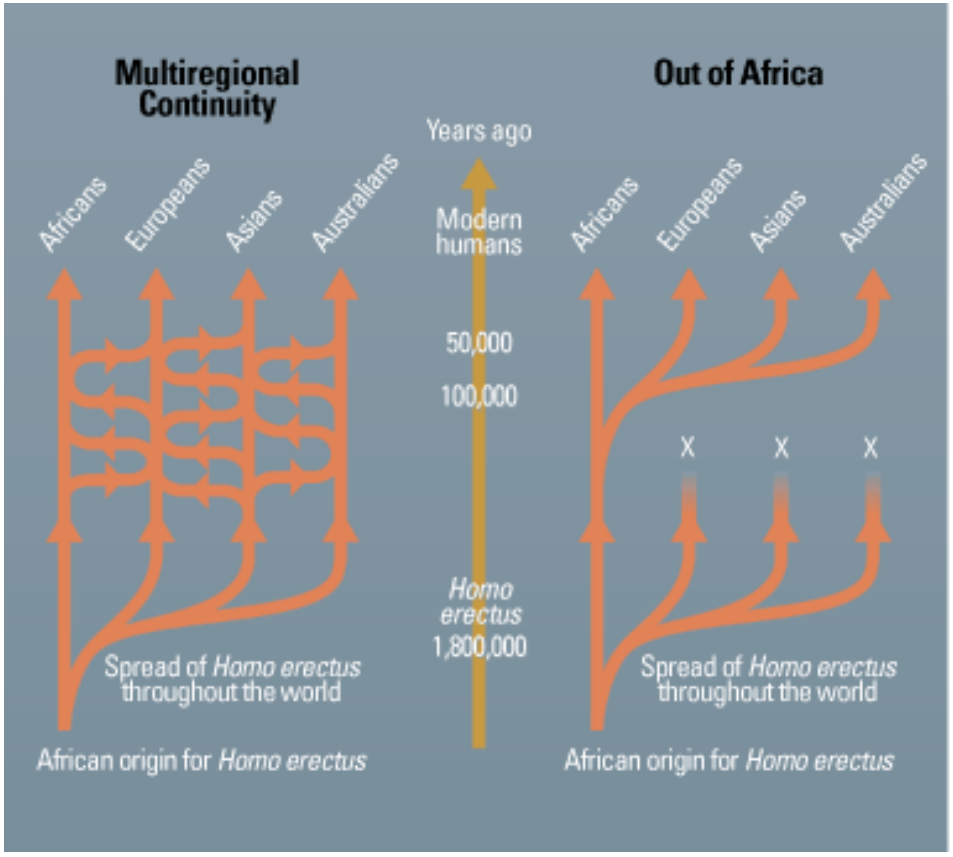
Geographic region	N	B*73+ individuals: Associated HLA-C alleles (%)			
		C*15		Not C*15	
		15:05	not 15:05	12:02	not 12:02
Europe	2,677	98.4	0.3	0.4	0.9
Europe#	2,907	98.5	0.3	0.3	0.9
Africa	39	100	0.0	0.0	0.0
Africa##	90	97.8	2.2	0.0	0.0
W Asia	128	89.8	5.5	0.8	3.9
N/S/E Asia	53	92.5	5.7	1.9	0.0
Other	498	99.0	0.0	0.4	0.6
Total	3,676	98.2	0.5	0.4	0.9

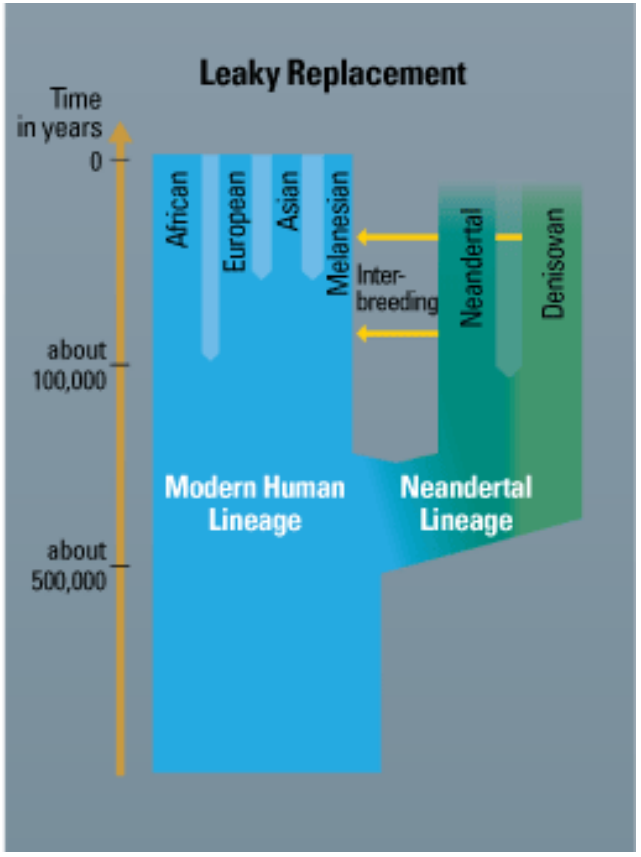
F



Out of Africa Revisited

"Human uniqueness?"

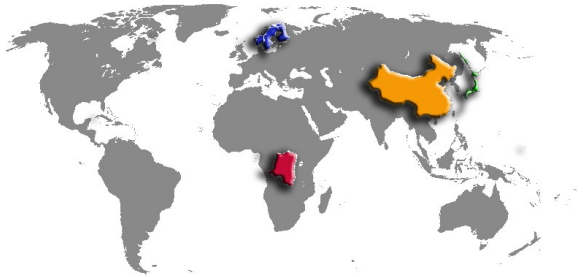


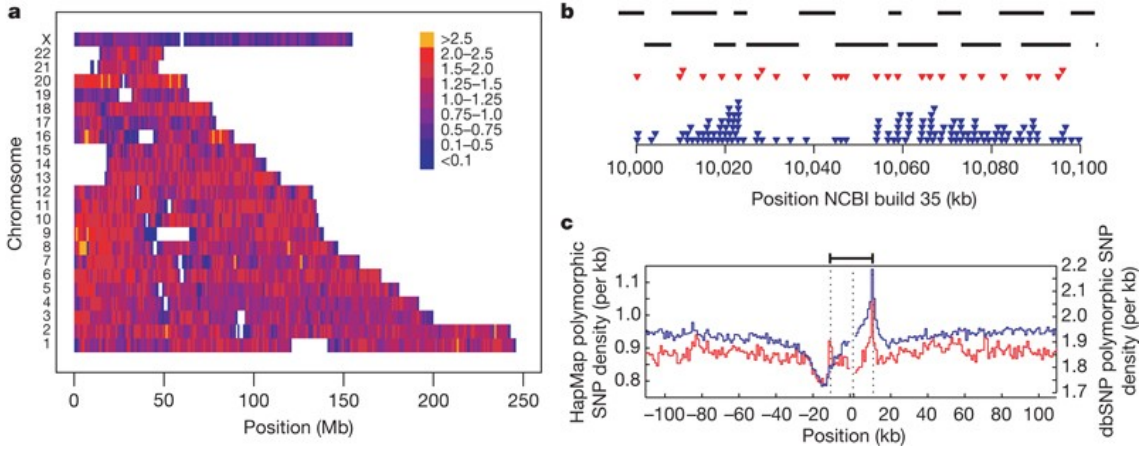


Ann Gibbons Science 28 January 2011:

The HapMap Project

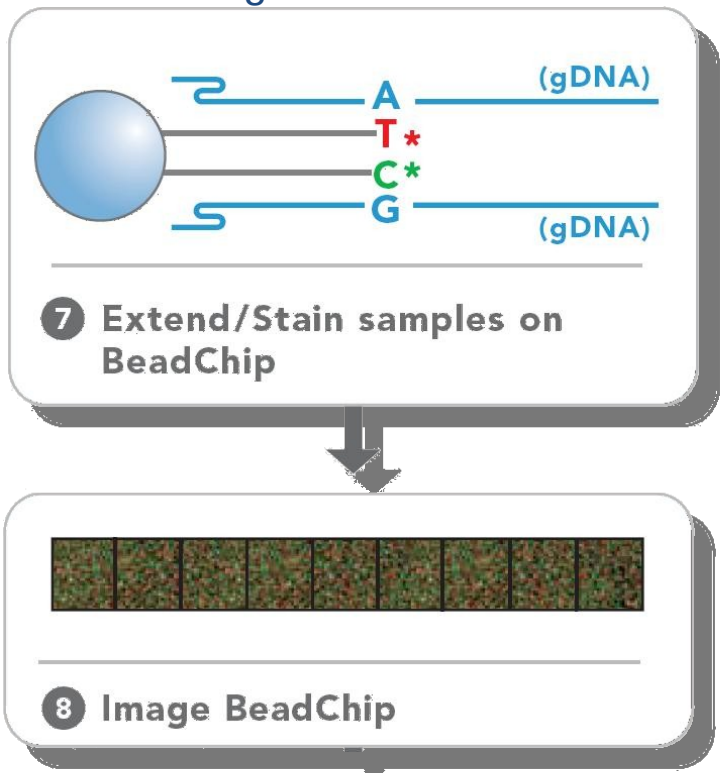
ASW	African ancestry in Southwest USA	90	
CEU	Northern and Western Europeans (Utah)		180
CHB	Han Chinese in Beijing, China	90	
CHD	Chinese in Metropolitan Denver	100	
GIH	Gujarati Indians in Houston, Texas	100	
JPT	Japanese in Tokyo, Japan	91	
LWK	Luhya in Webuye, Kenya	100	
MXL	Mexican ancestry in Los Angeles	90	
MKK	Maasai in Kinyawa, Kenya	180	
TSI	Toscani in Italia	100	
YRI	Yoruba in Ibadan, Nigeria	100	





Genotyping:

Probe a limited number (~1M) of known highly variable positions of the human genome



Linkage Disequilibrium & Haplotype Blocks

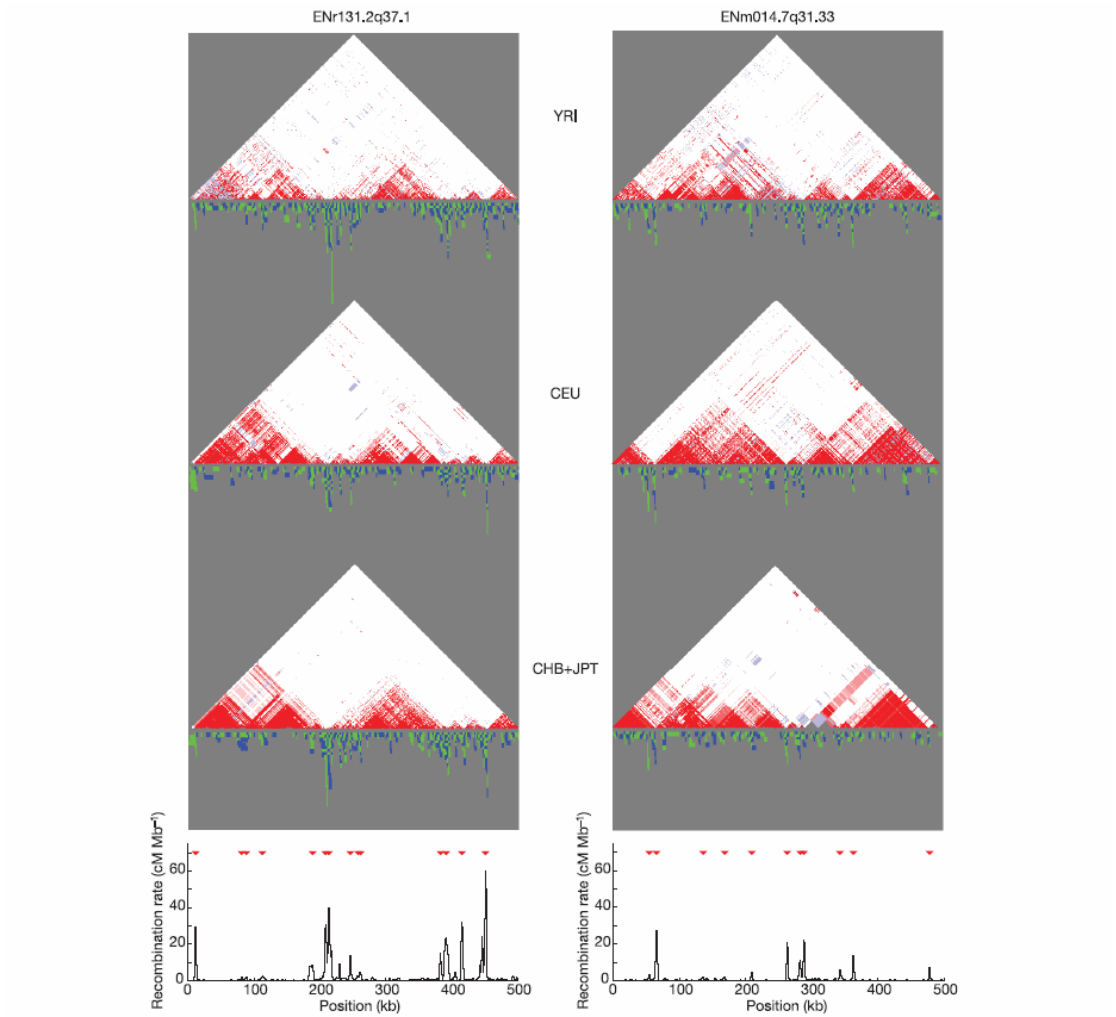


Figure 8 | Comparison of linkage disequilibrium and recombination for two ENCODE regions. For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB+JPT analysis panels are shown: white, $D' < 1$ and LOD < 2; blue, $D' = 1$ and LOD < 2; pink, $D' < 1$ and LOD ≥ 2; red, $D' = 1$ and LOD ≥ 2. Below each of these plots is shown the intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles[46].

Minor allele: A G

p_A p_G

Linkage Disequilibrium (LD):

$D = P(A \text{ and } G) - p_A p_G$

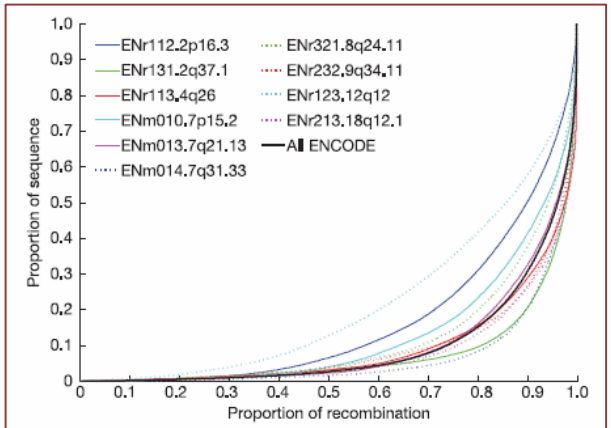
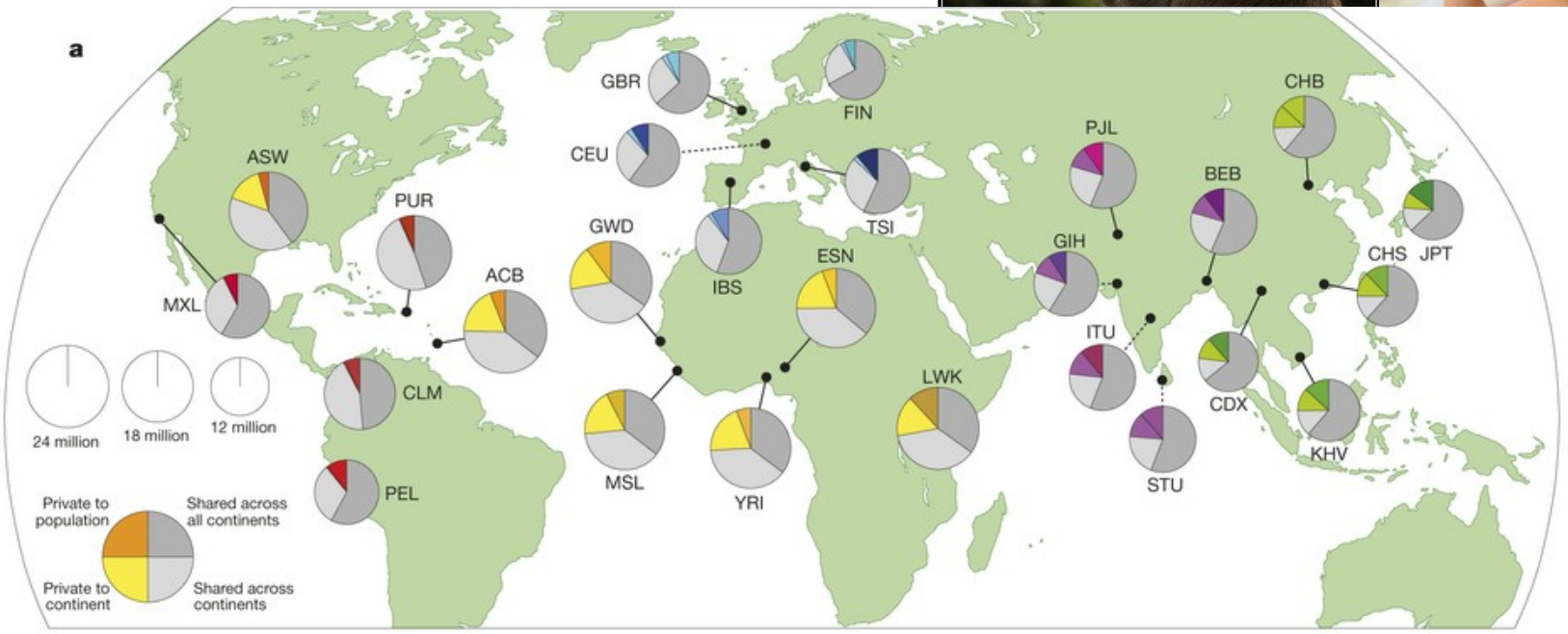
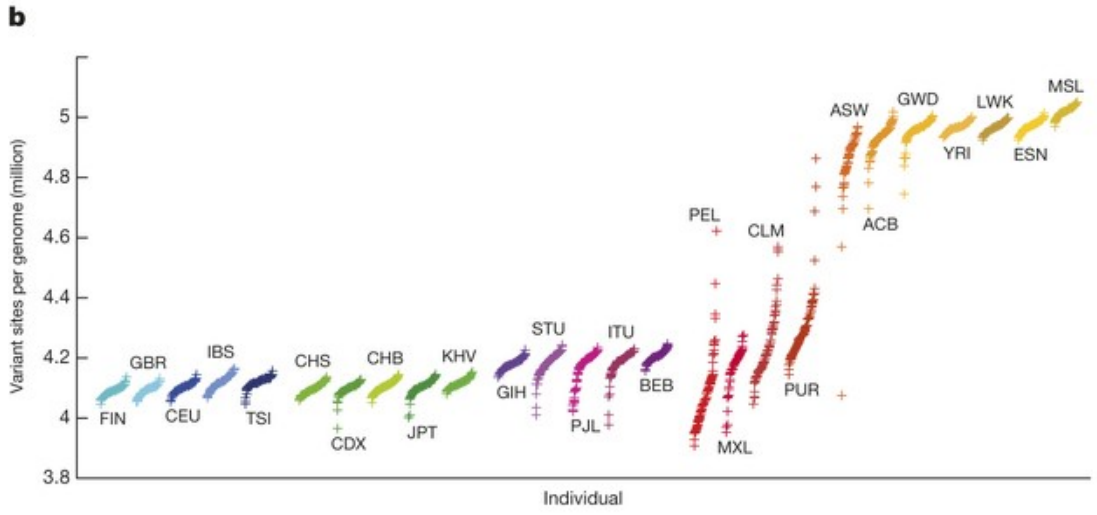
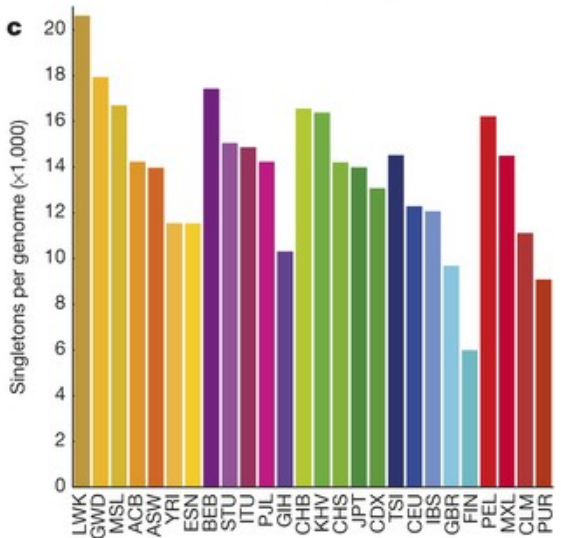


Figure 9 | The distribution of recombination events over the ENCODE regions. Proportion of sequence containing a given fraction of all recombination for the ten ENCODE regions (coloured lines) and combined (black line). For each line, SNP intervals are placed in decreasing order of estimated recombination rate[46], combined across analysis panels, and the cumulative recombination fraction is plotted against the cumulative proportion of sequence. If recombination rates were constant, each line would lie exactly along the diagonal, and so lines further to the right reveal the fraction of regions where recombination is more strongly locally concentrated.

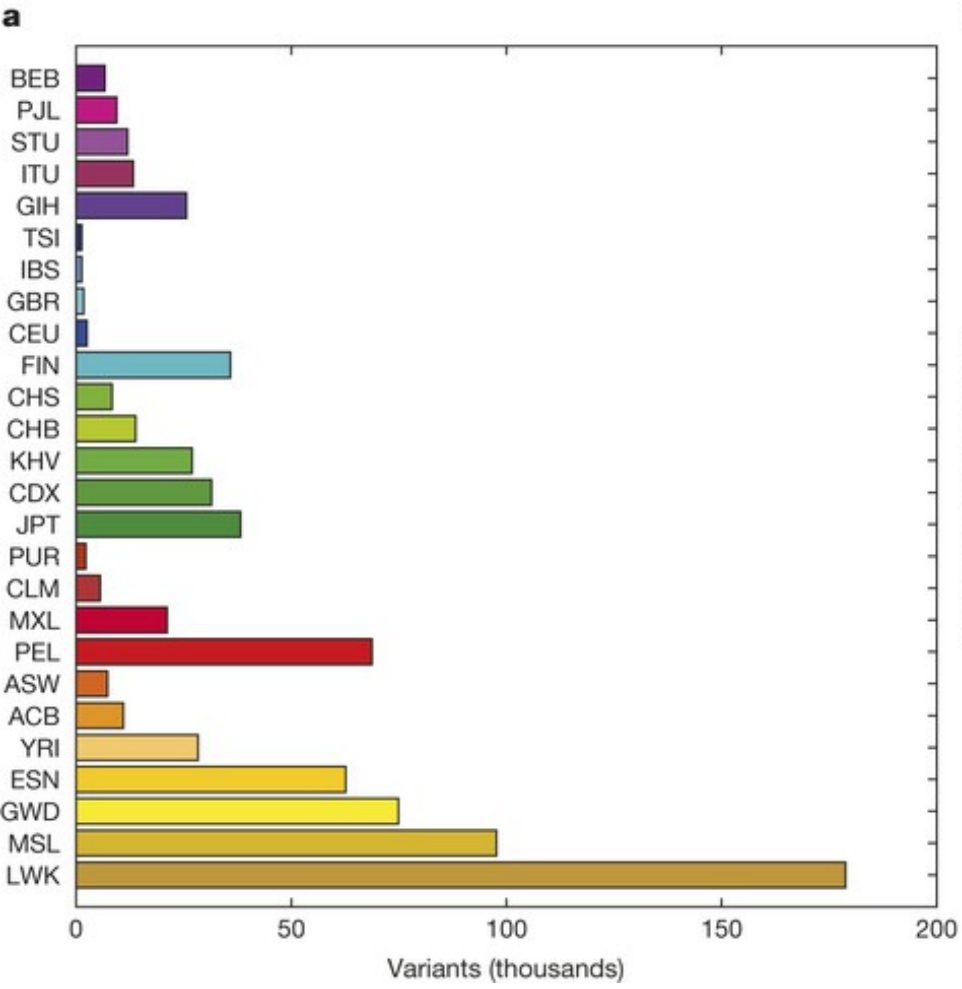
Population Sequencing – 1000 Genomes Project

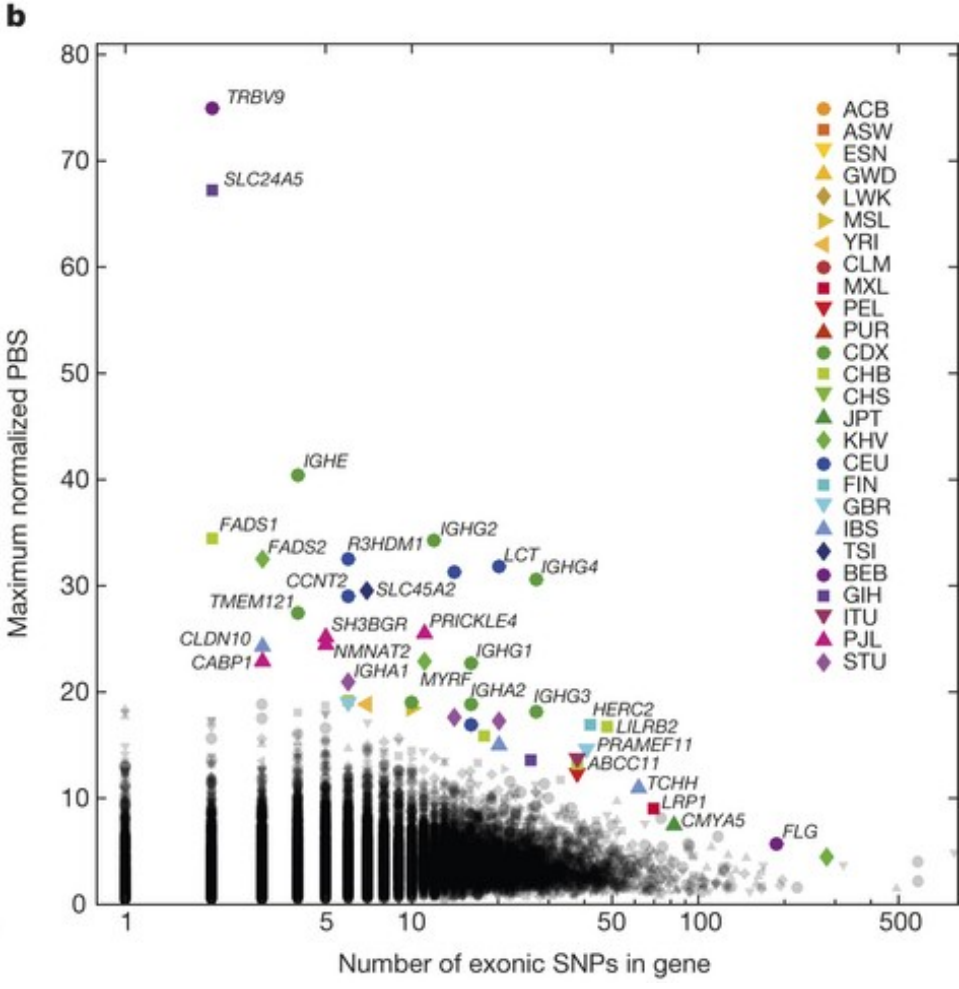




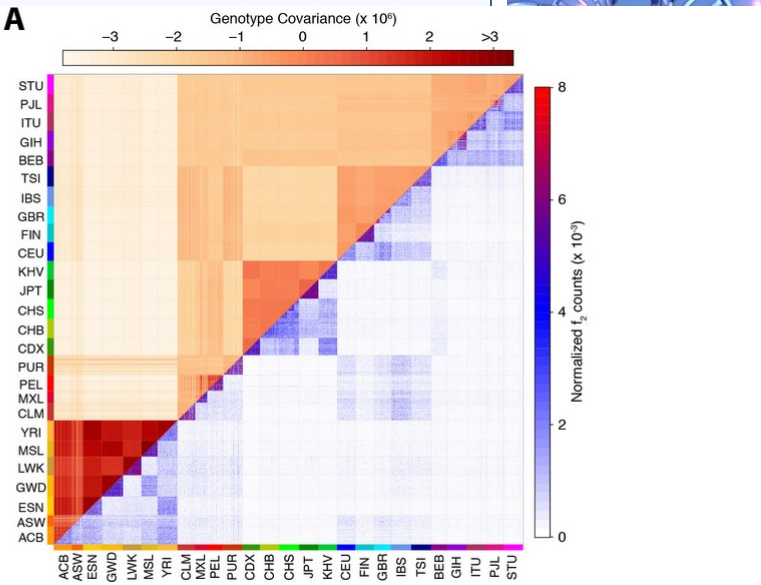


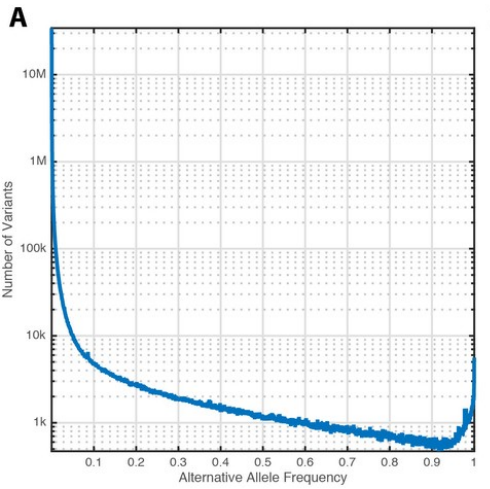
Population Sequencing – 1000 Genomes Project

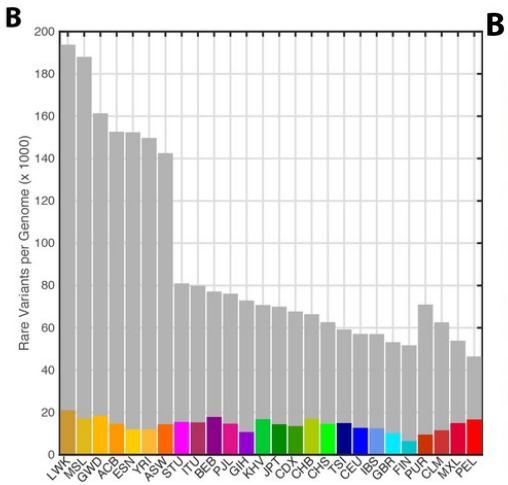


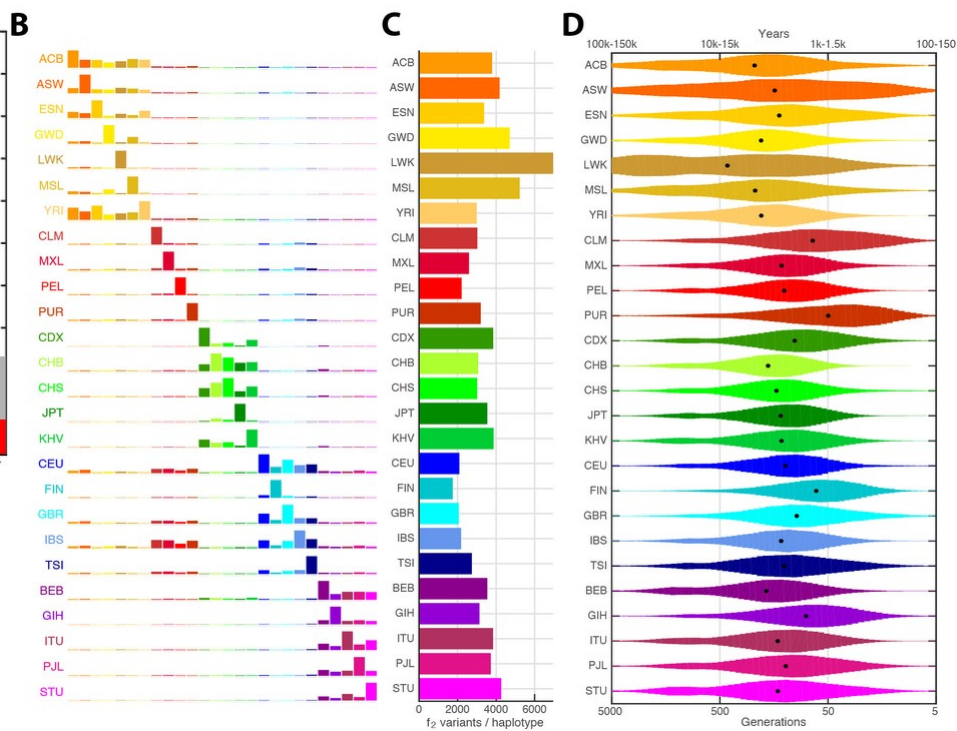


Population Sequencing – [1000 Genomes Project]

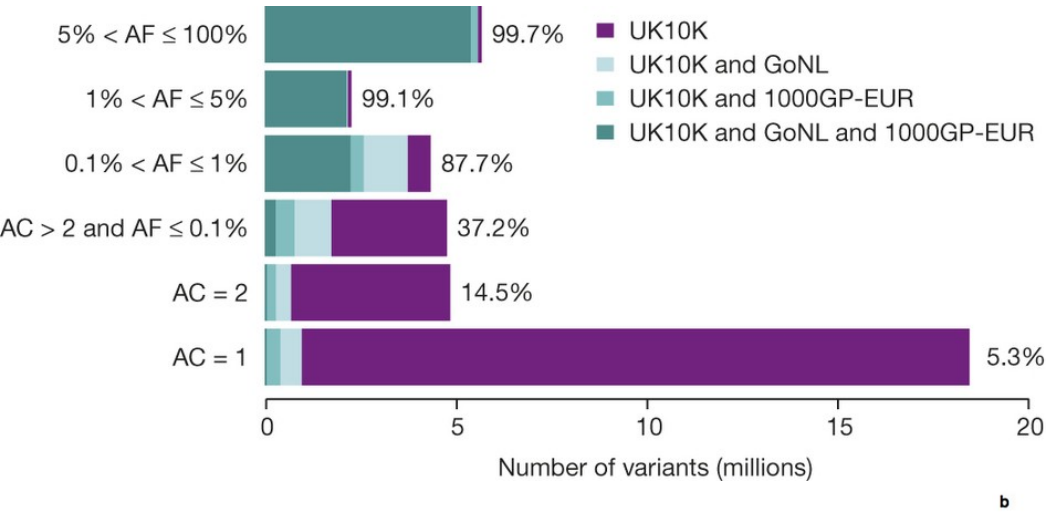




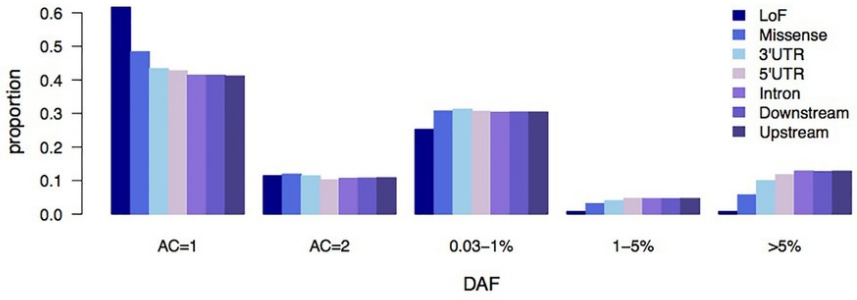




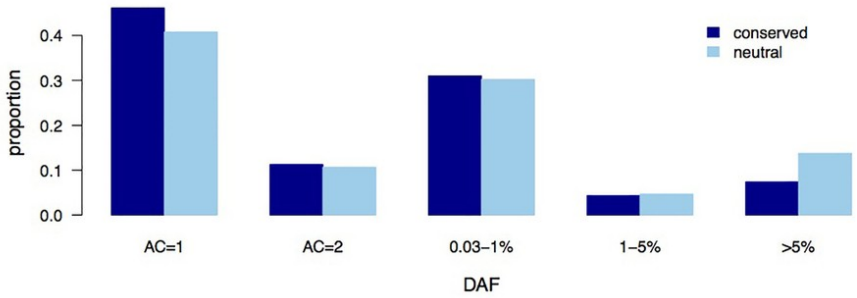
Population Sequencing – UK10K

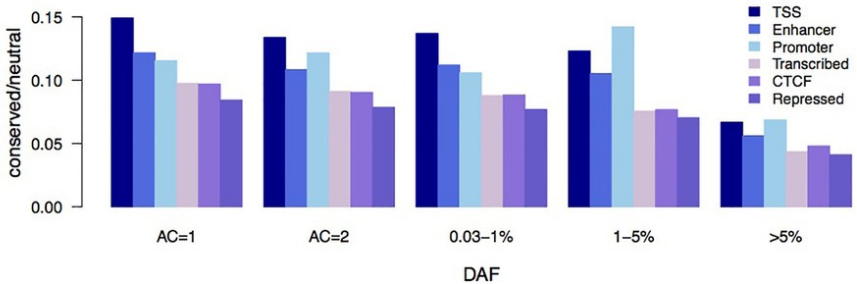


b

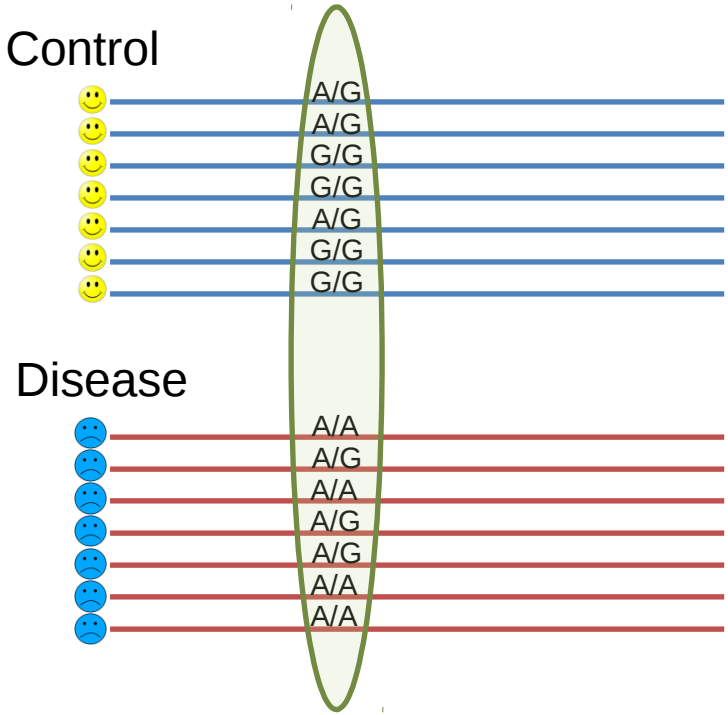


c



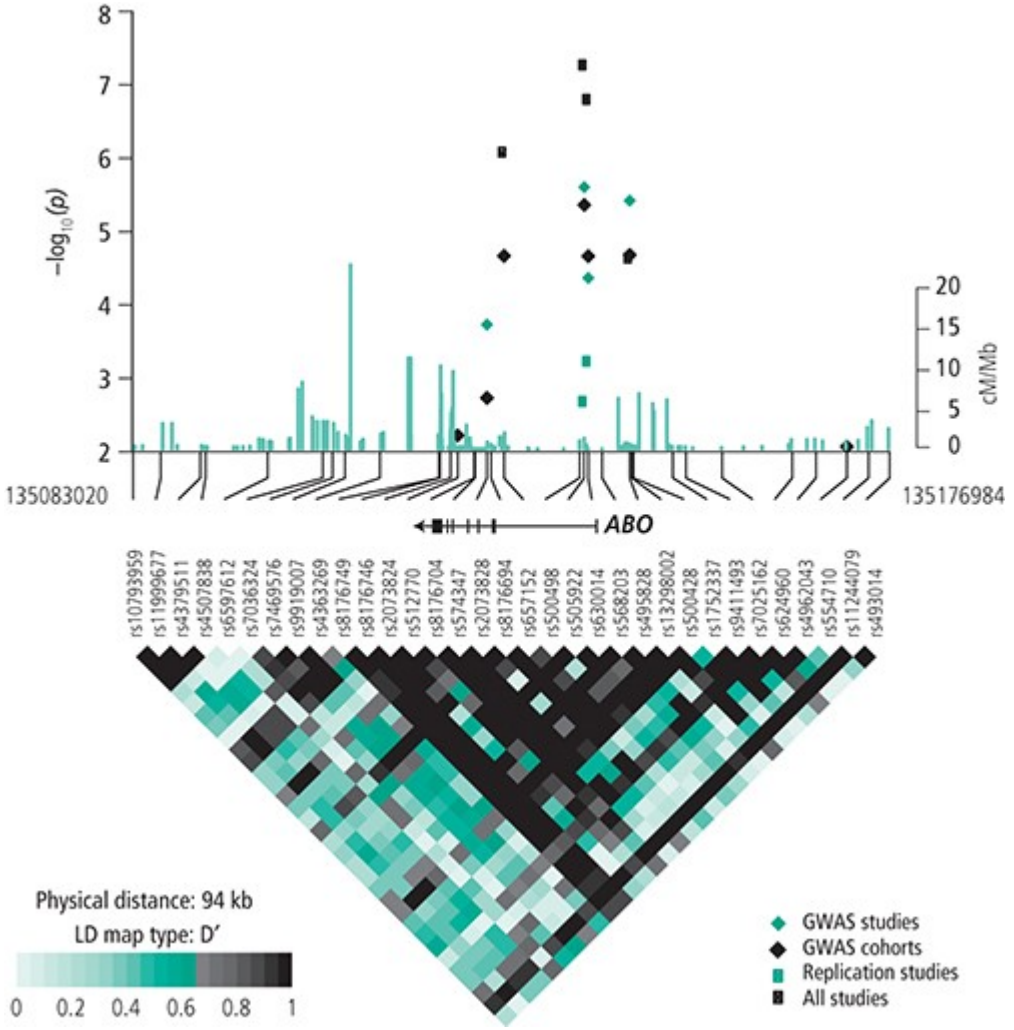


Association Studies

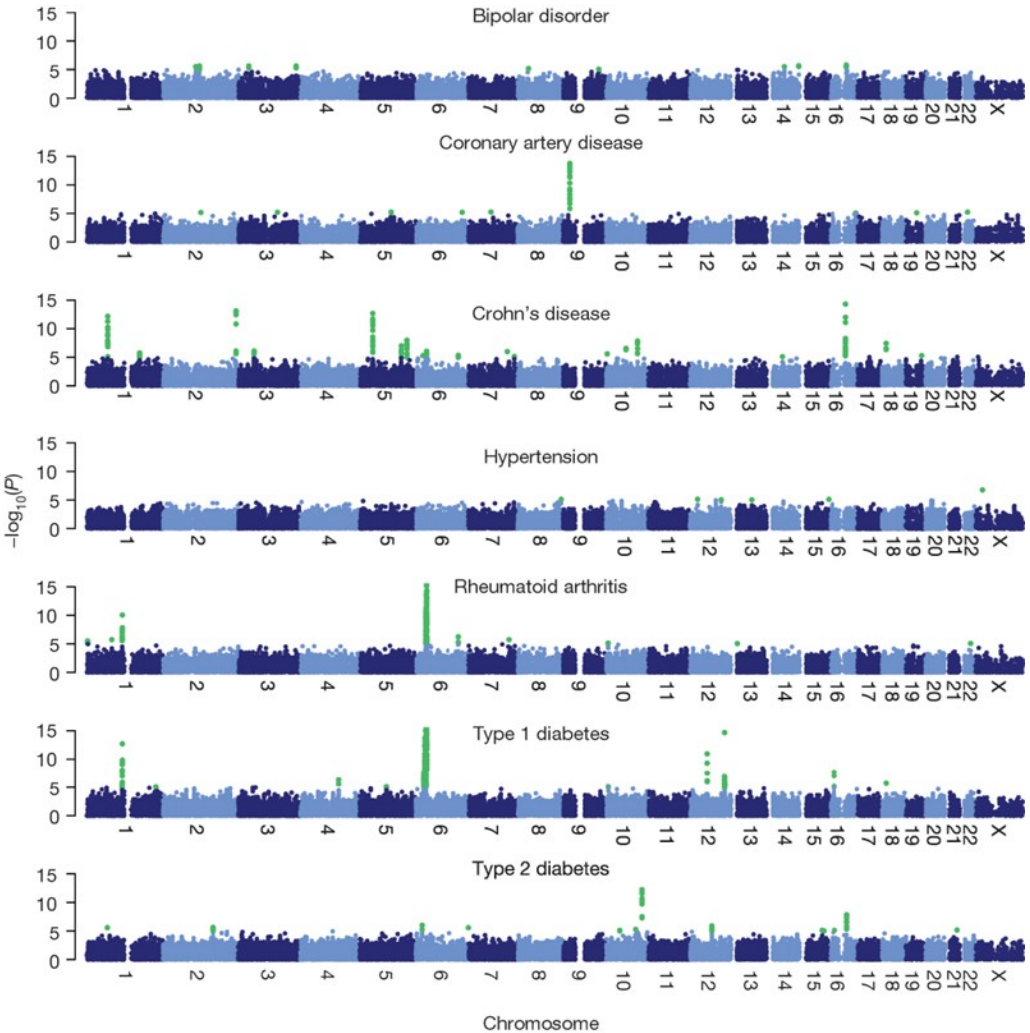


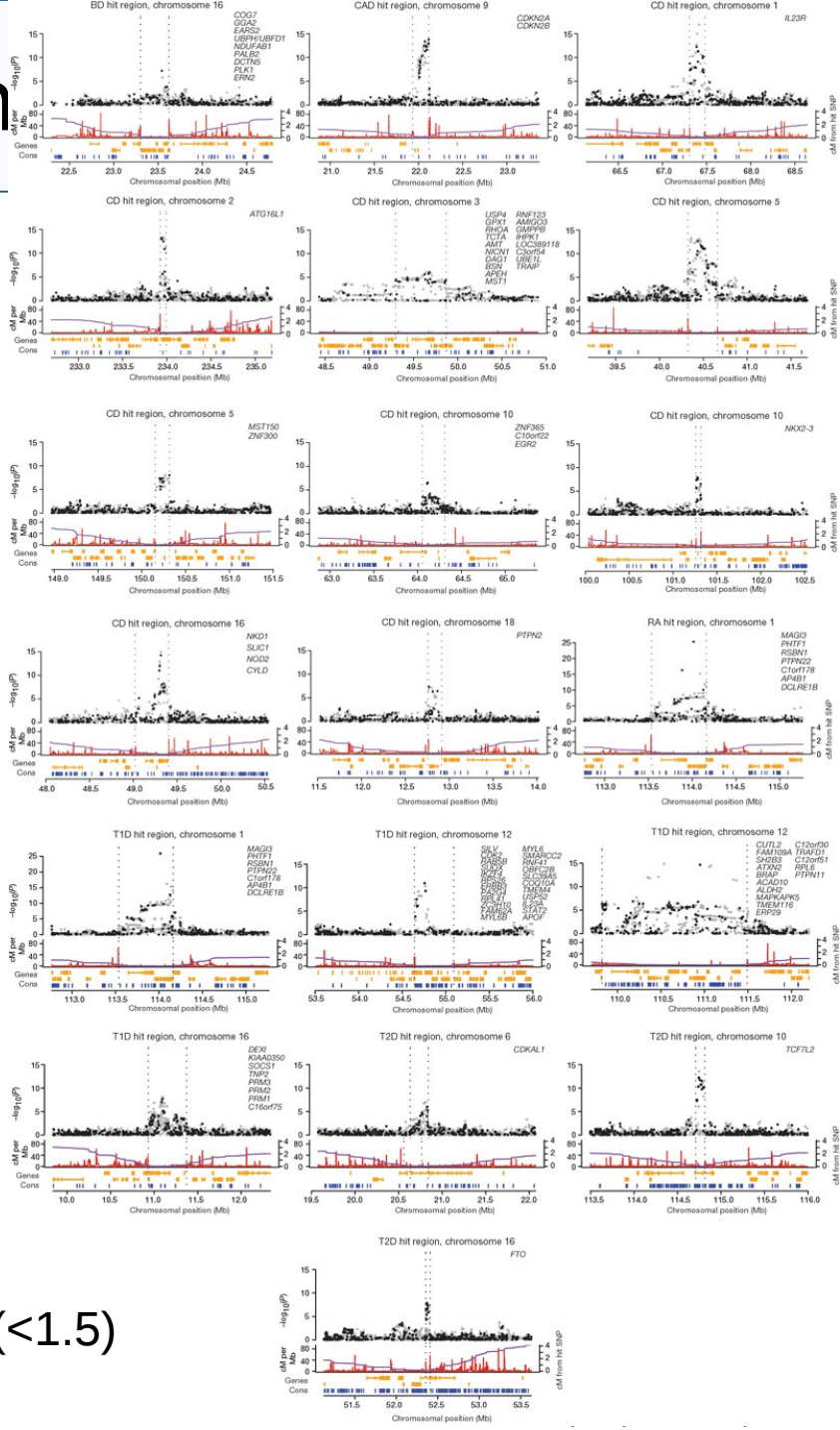
AA	0	4
AG	3	3
GG	4	0

p-value



Wellcome Trust Case Con

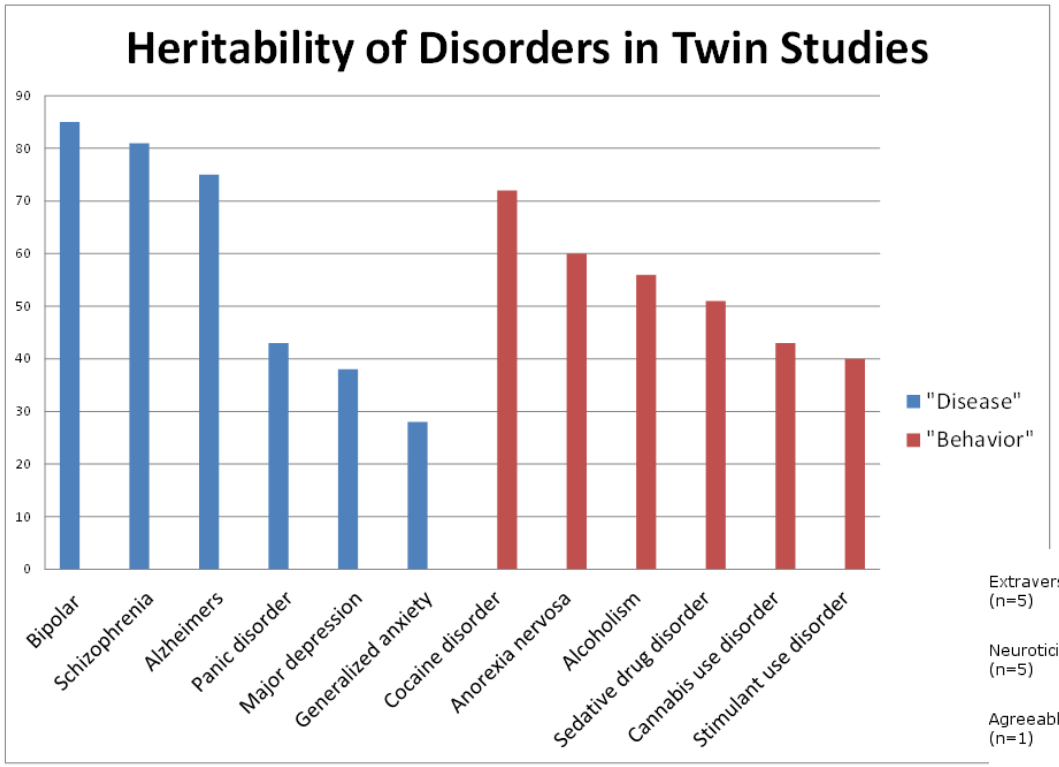


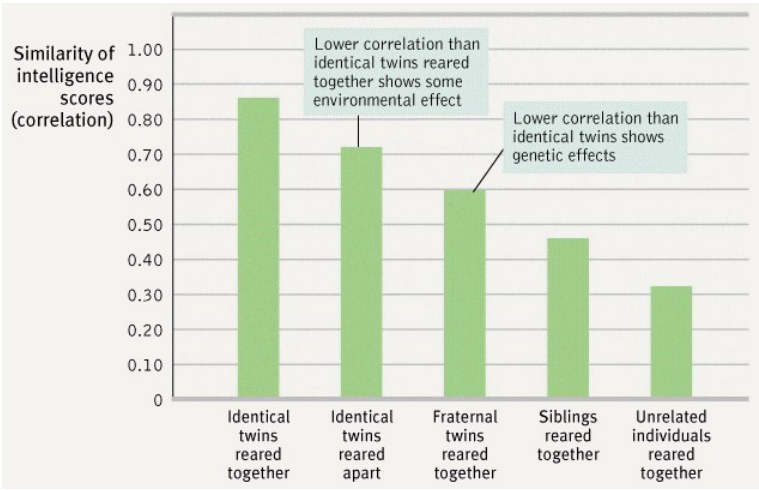


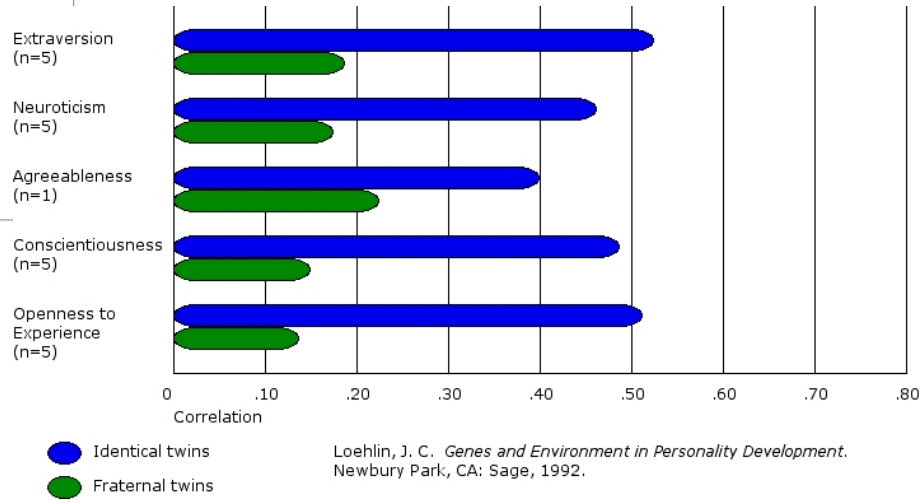
Many associations of small effect sizes (<1.5)

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

Heritability & Environment





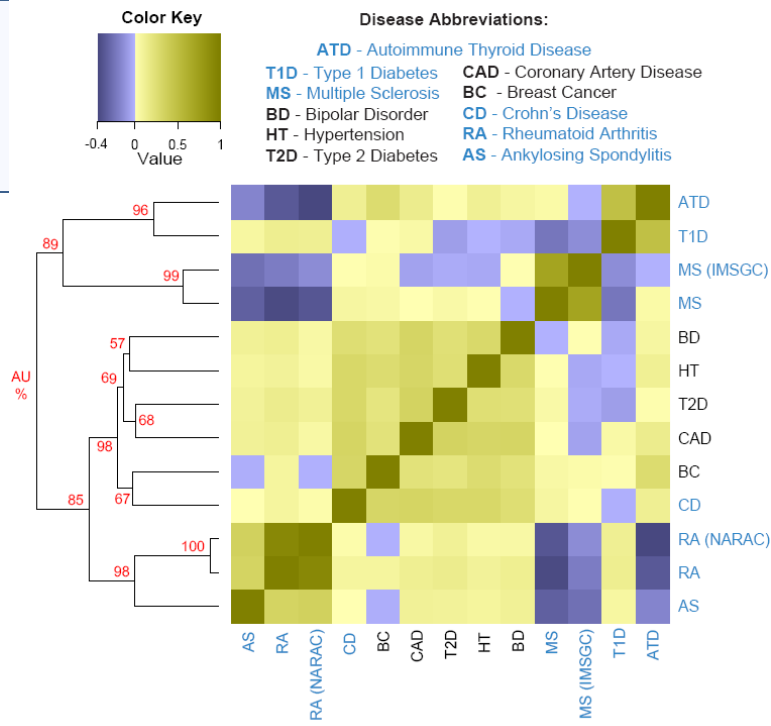


Loehlin, J. C. *Genes and Environment in Personality Development.* Newbury Park, CA: Sage, 1992.

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

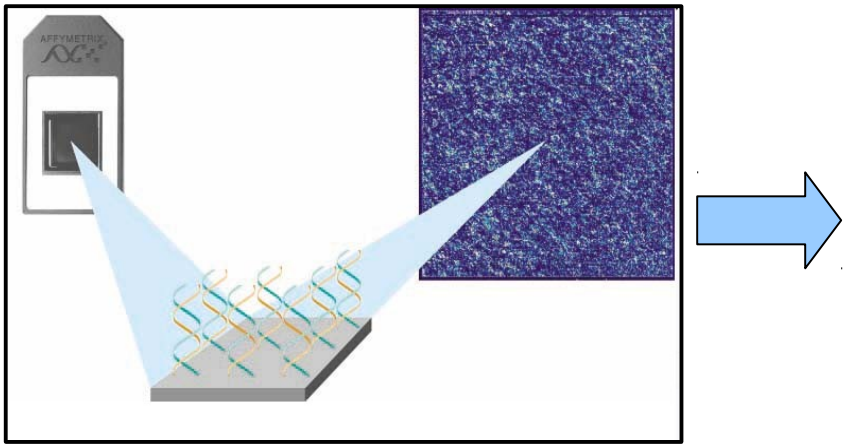
Disease Clustering

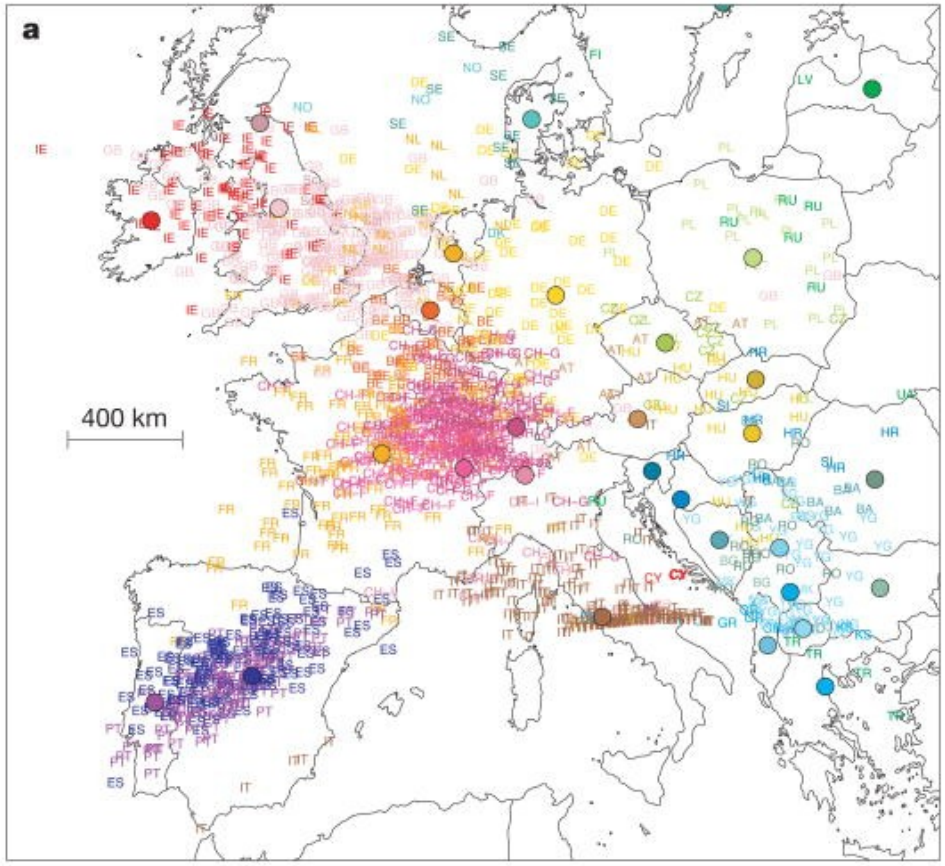
- RA vs. ATD
- RA vs. MS
  - No recorded co-occurrence of RA and MS

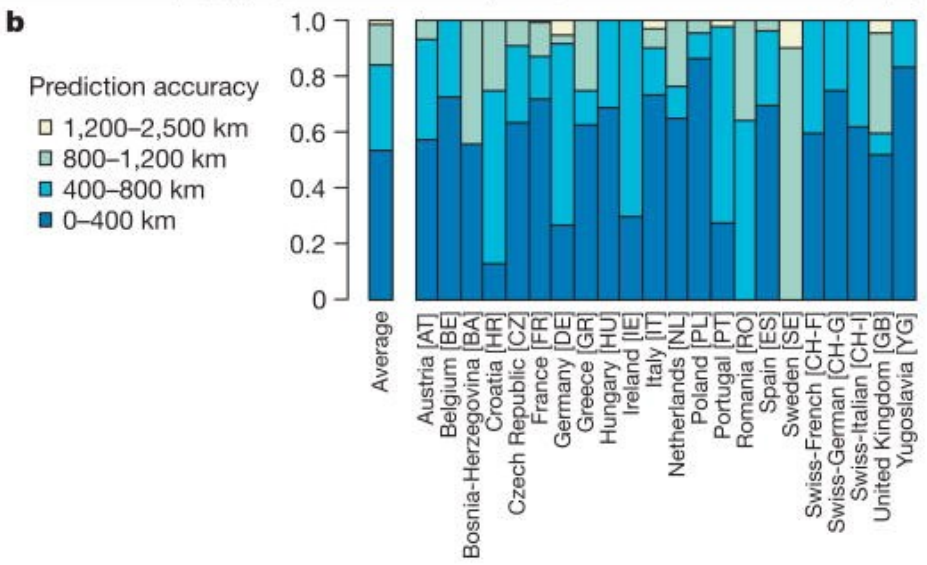


SNP - Allele	Gene Symbol	Genetic Variation Score (GVS)						
		RA (NARAC)	RA	AS	T1D	ATD	MS (IMSGC)	MS
rs11752919 - C	ZSCAN23	-3.48	-3.21	-9.39	1.10	0.70	3.25	2.99
rs3130981 - A	CDSN	-0.46	-1.00	-9.47	-4.94	0.33	10.00	13.41
rs151719 - G	HLA-DMB	-6.71	-4.77	-1.08	-13.63	0.34	8.58	17.76
rs10484565 - T	TAP2	25.52	8.37	1.34	15.74	-1.36	-0.56	-0.30
rs1264303 - G	VARS2	11.51	7.36	18.76	0.89	-1.76	-1.85	-1.75
rs1265048 - C	CDSN	6.59	2.97	50.13	6.34	-0.85	-2.39	-4.16
rs2071286 - A	NOTCH4	5.30	0.78	6.42	4.04	-0.03	-1.89	-2.45
rs2076530 - G	BTNL2	67.49	56.46	14.06	13.58	-6.41	-9.50	-18.52
rs757262 - T	TRIM40	14.58	9.11	6.27	1.56	-0.79	-2.05	-7.34

Global Ancestry Infe





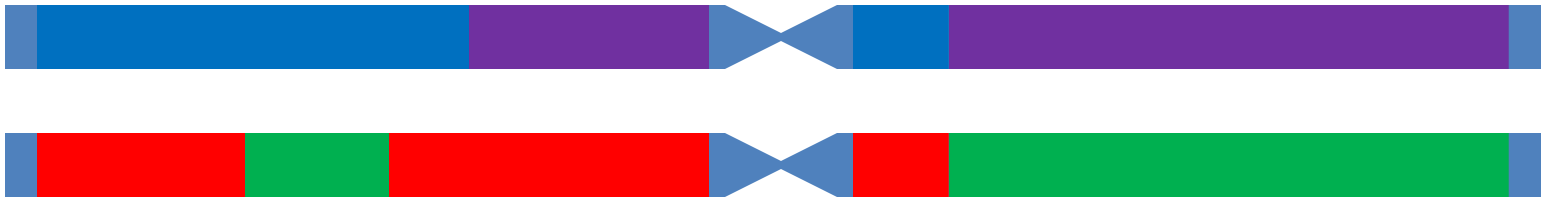


Nature. 2008 November 6; 456(7218): 98–101.

Ancestry Painting

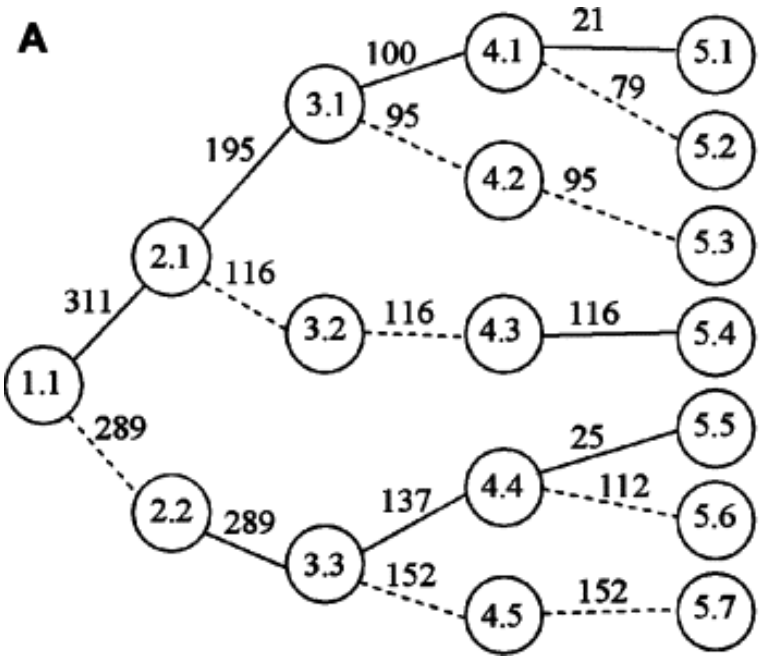
?

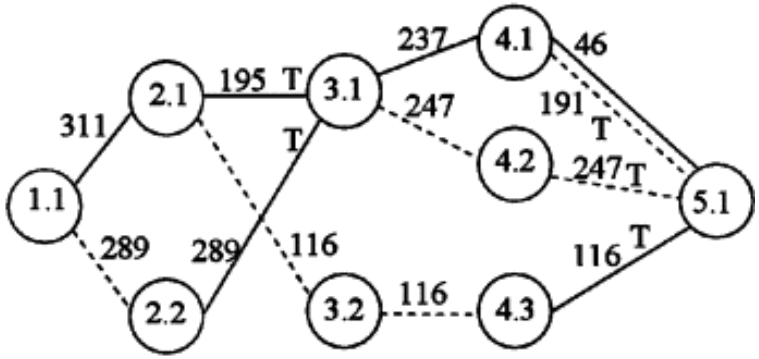
- Danish
- French
- Spanish
- Mexican



Modeling population haplotypes – VLMC

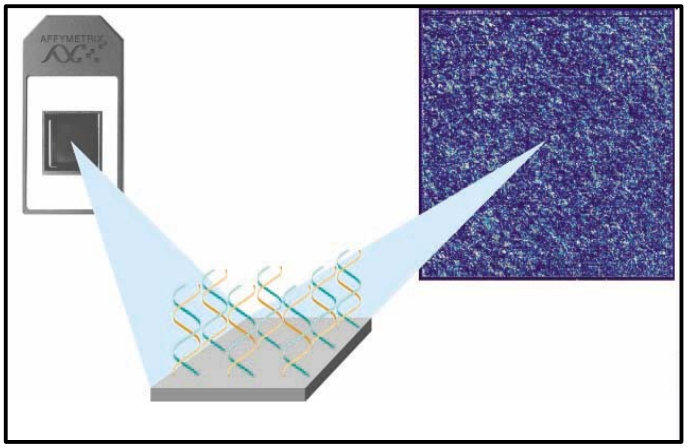
Haplotype	Count		
	Total	Case	Control
1111	21	12	9
1112	79	43	36
1122	95	43	52
1221	116	59	57
2111	25	14	11
2112	112	60	52
2122	152	69	83





Browning, 2006

Phasing



Haplotype Phasing

Haplotypes

ATCCGA

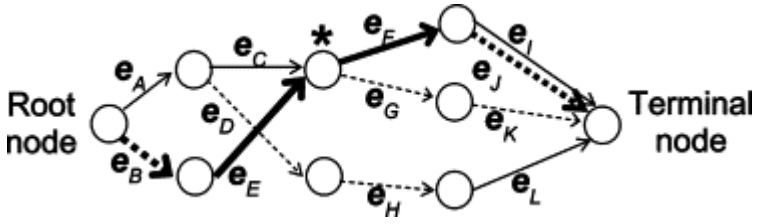
AGACGC

Genotype

$A\begin{Bmatrix}T\\G\end{Bmatrix}\begin{Bmatrix}C\\A\end{Bmatrix}CG\begin{Bmatrix}C\\A\end{Bmatrix}$

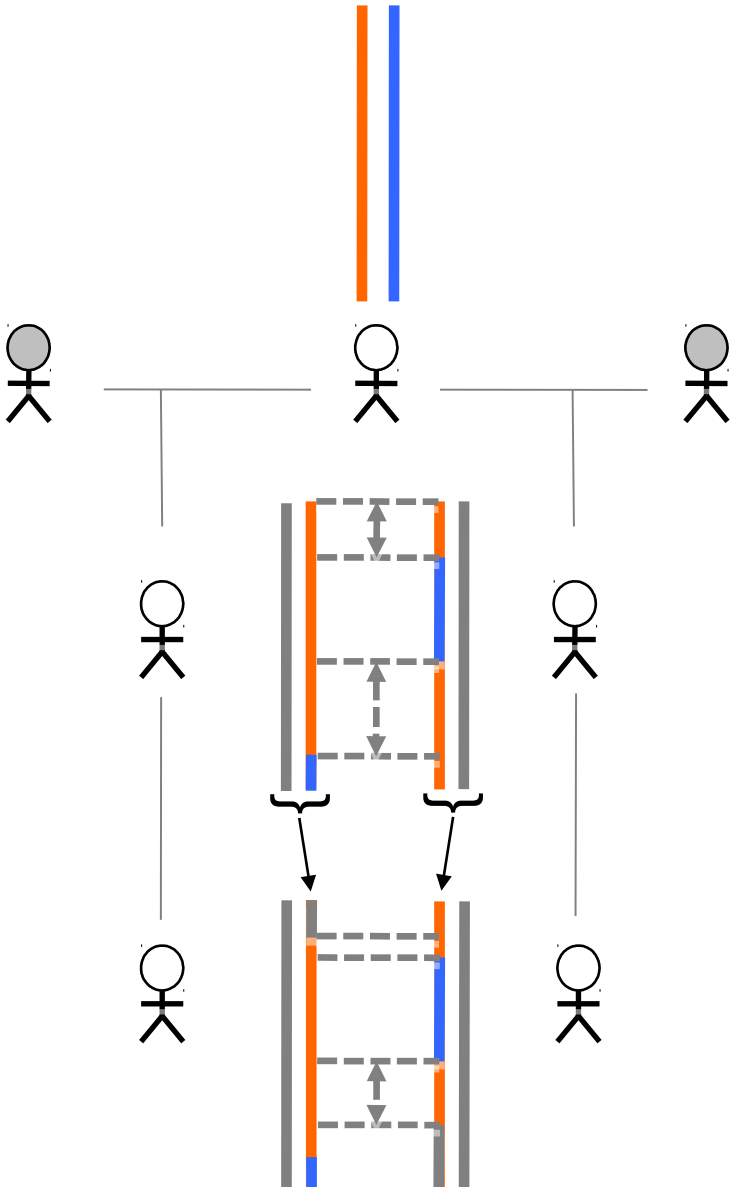
- High throughput cost effective sequencing technology gives genotypes and not haplotypes.

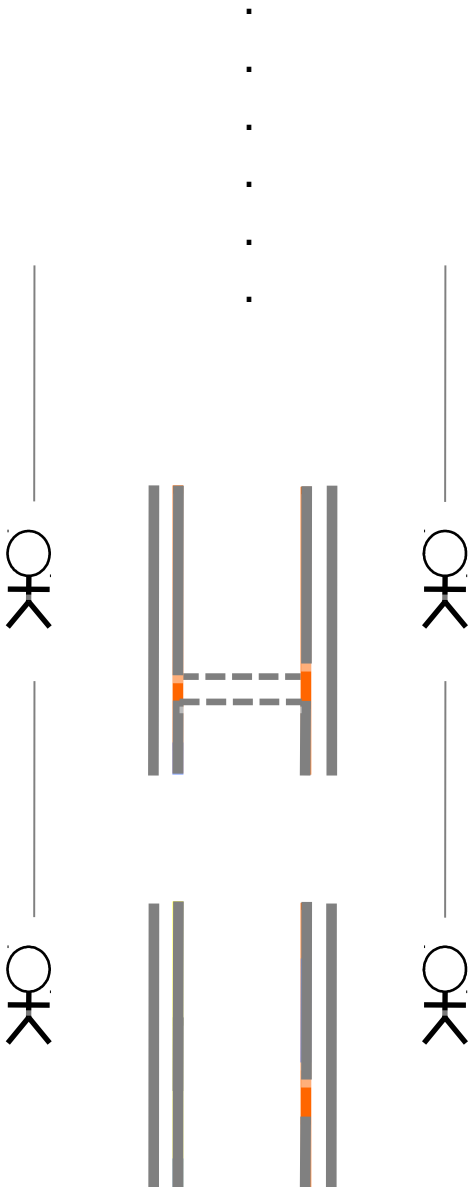
Possible phases: ATACGA AGACGA AGCCGC ATCCGC ….



Browning & Browning, 2007

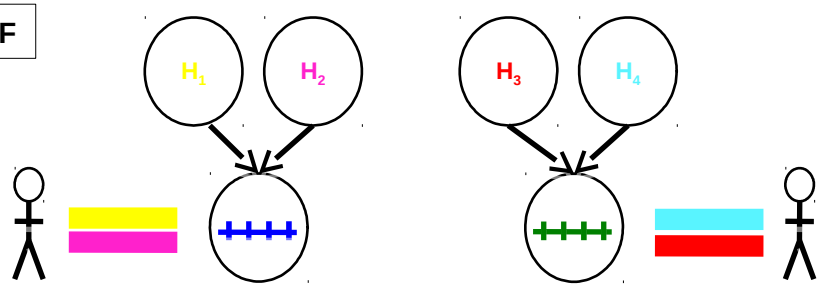
Identity By Descent



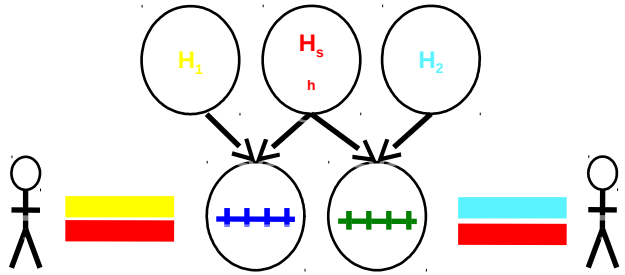


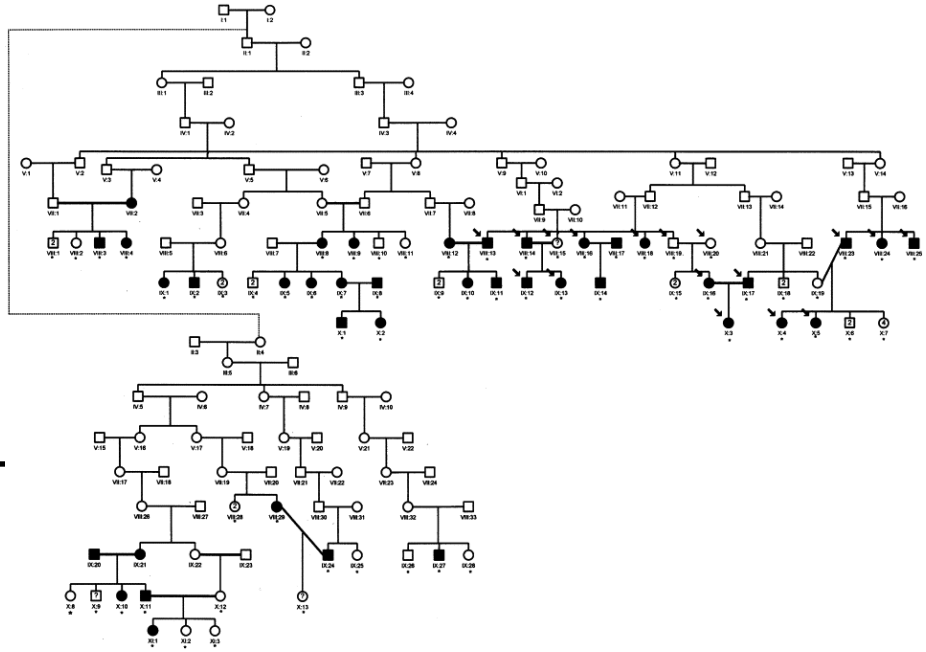
IBD detection

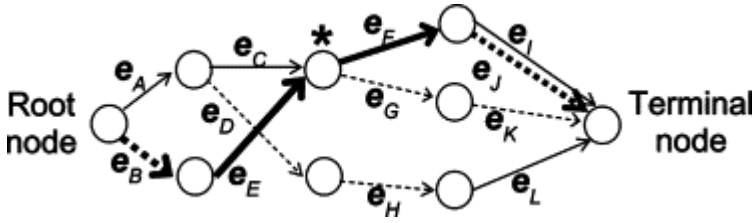
IBD = F



IBD = T







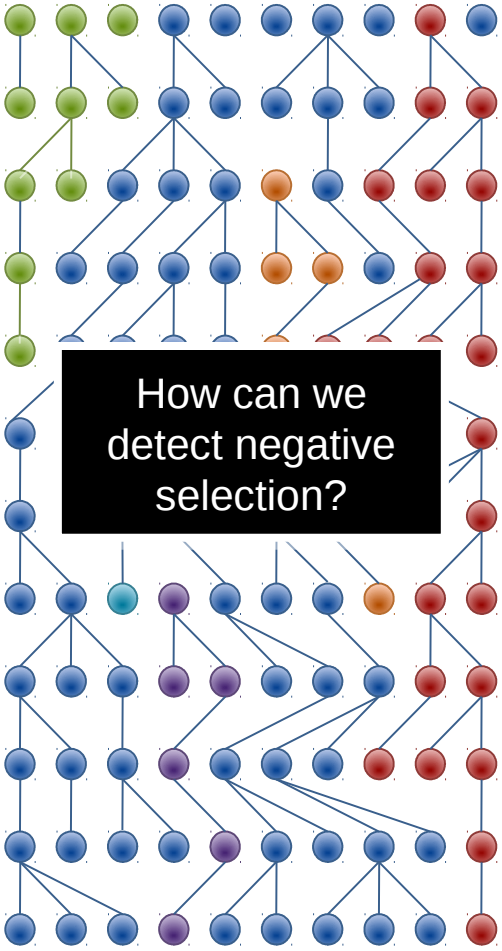
Parente

[Rodriguez et al. 2013]

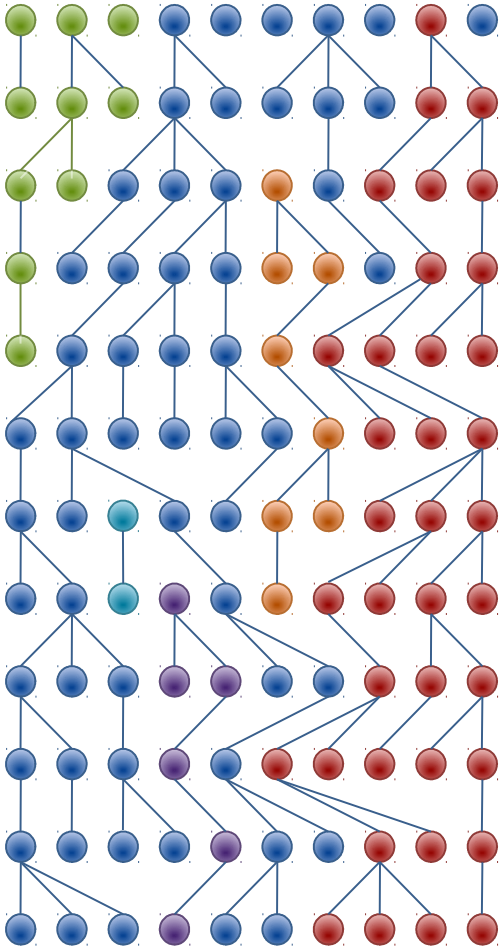
FastIBD: sample haplotypes for each individual, check for IBD

[Browning & Browining 2011]

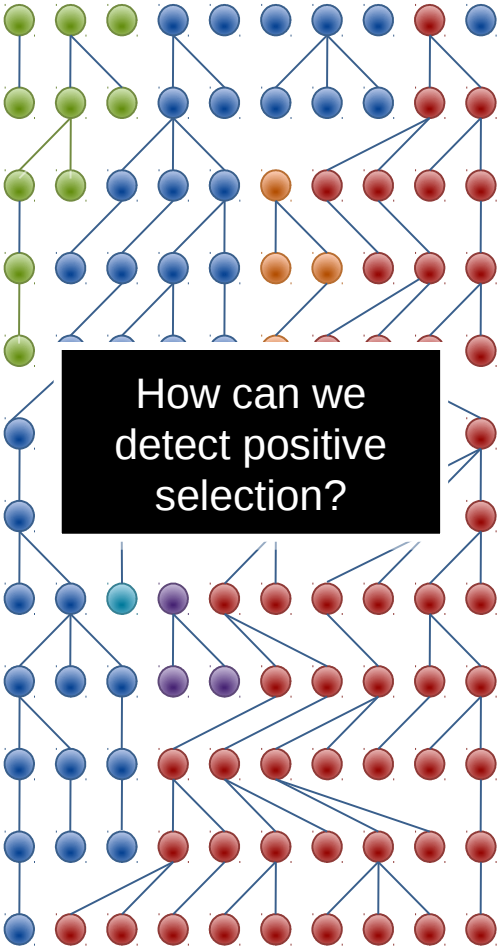
Fixation, Positive & Negative Selection



Negative Selection



Neutral Drift



Positive Selection

How can we detect positive selection?

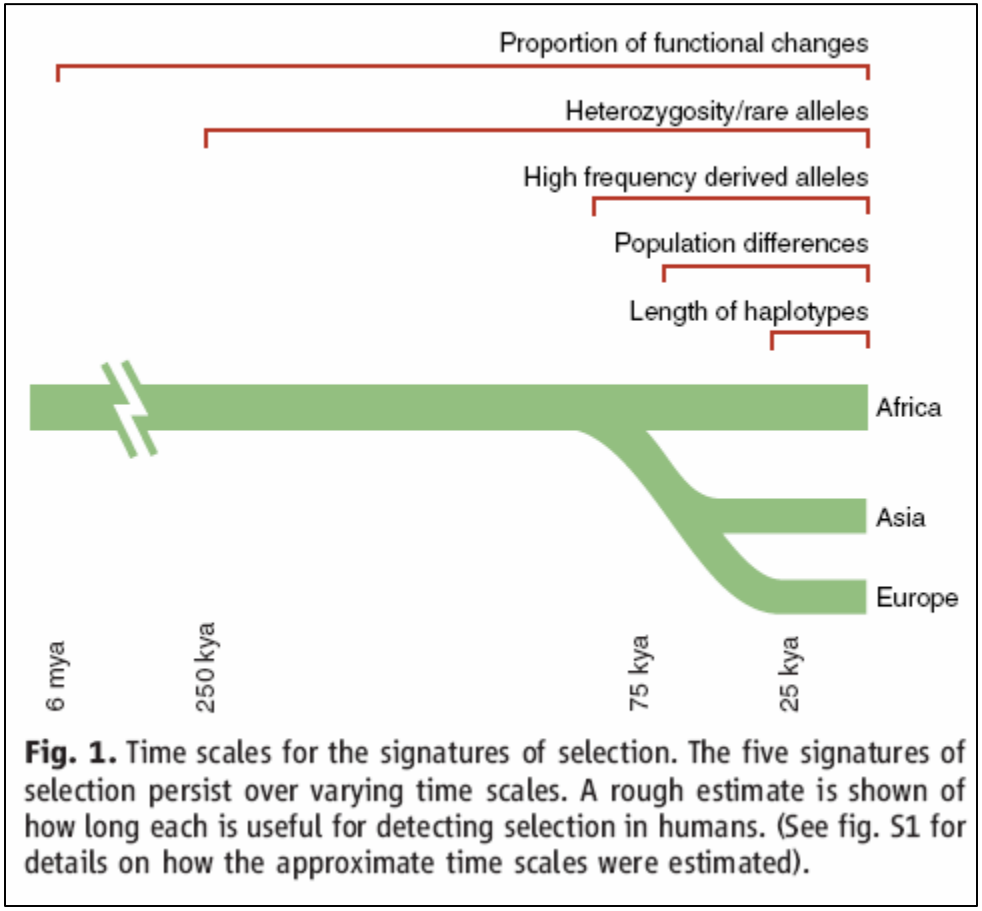


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

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

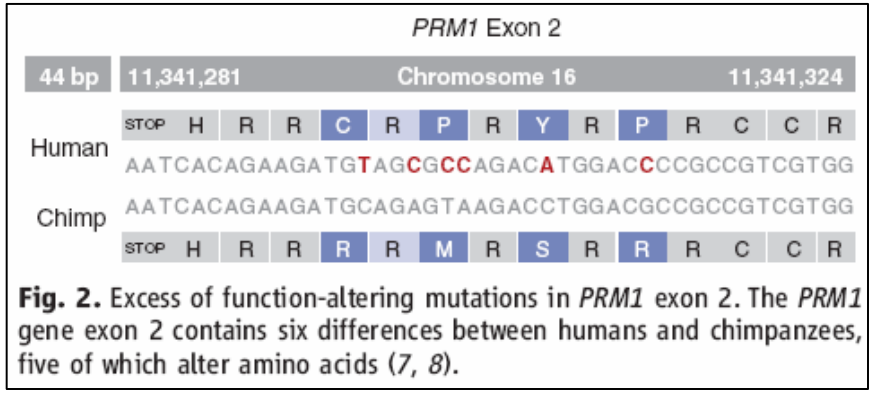


Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

How can we detect positive selection?

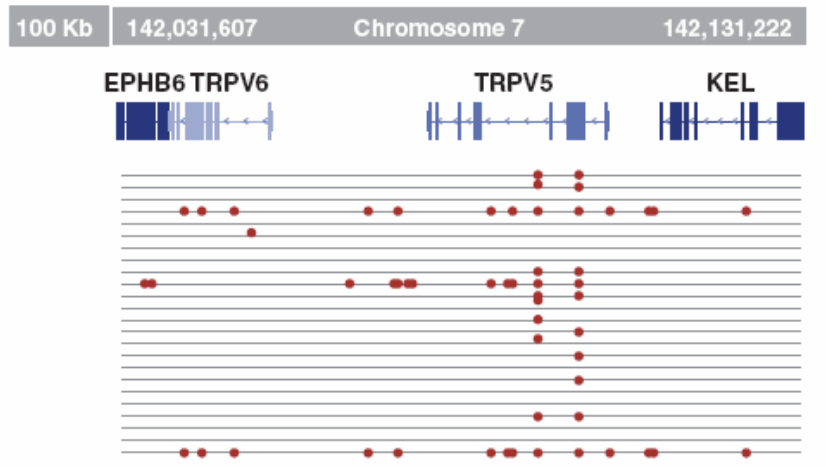


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

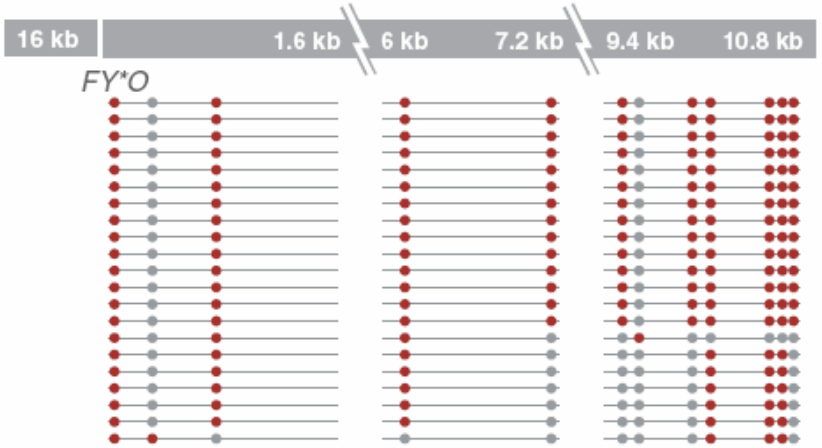


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (*34*). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

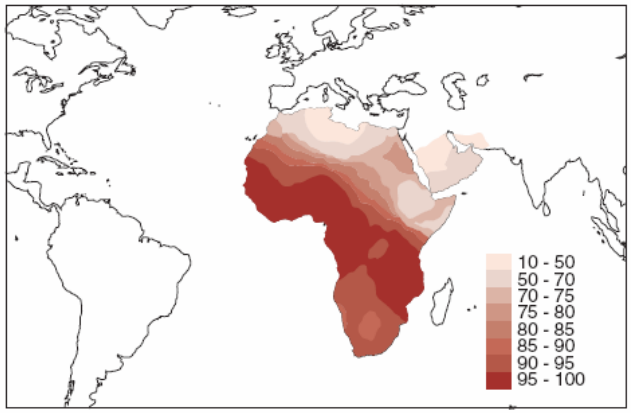


Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (*38*).

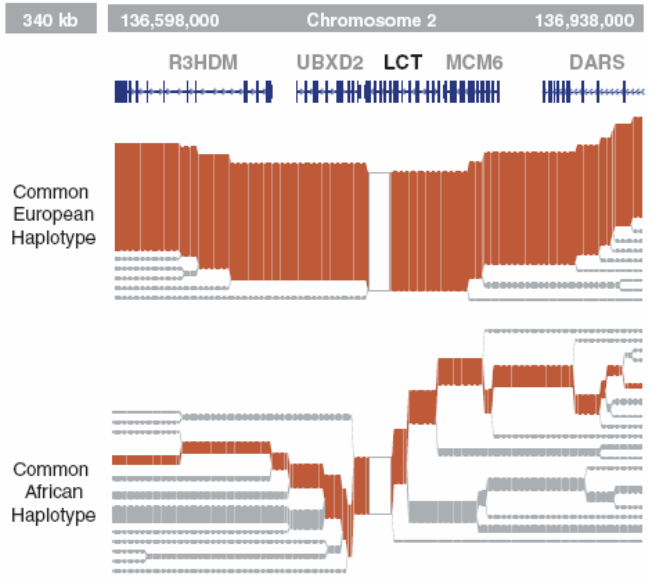
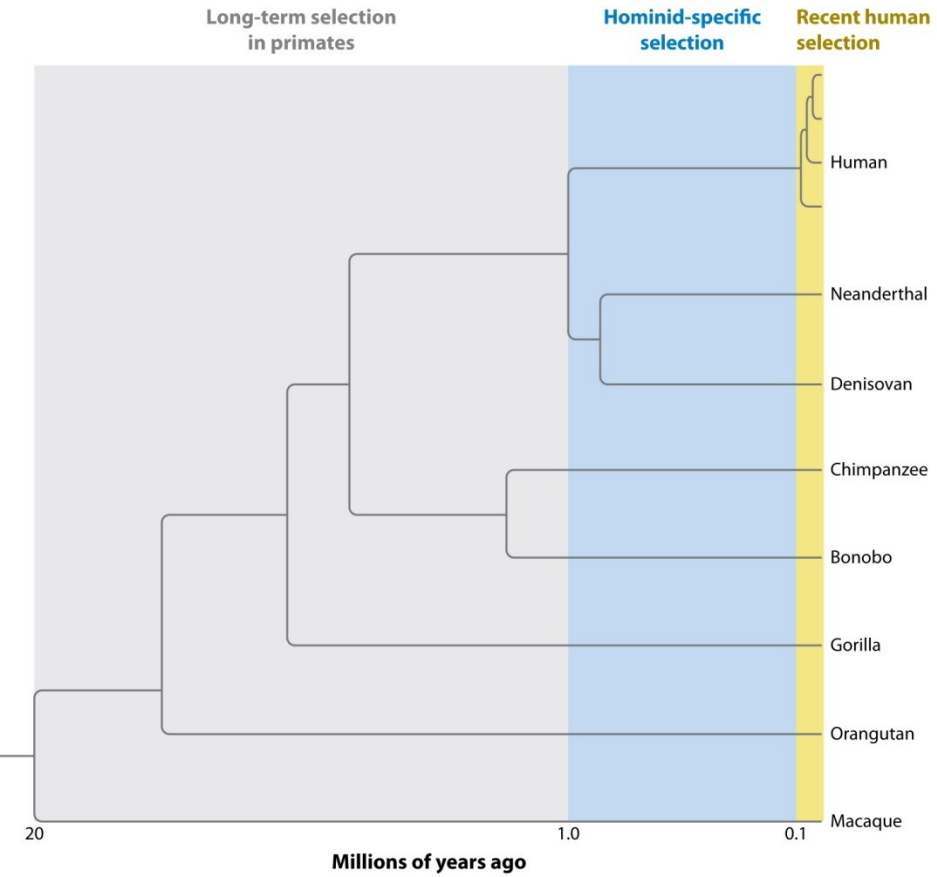
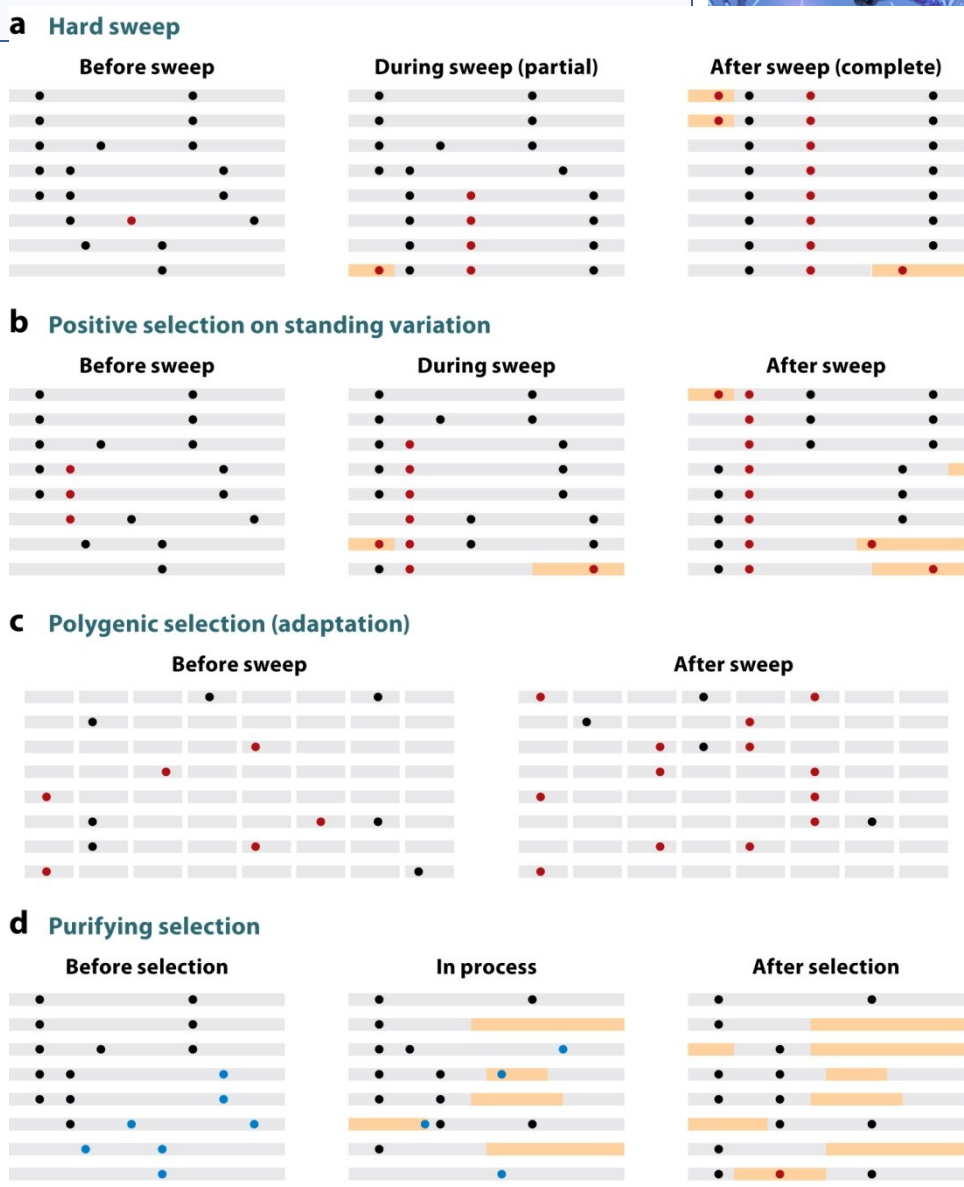


Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (*6*).

Positive Selection in Human Lineage

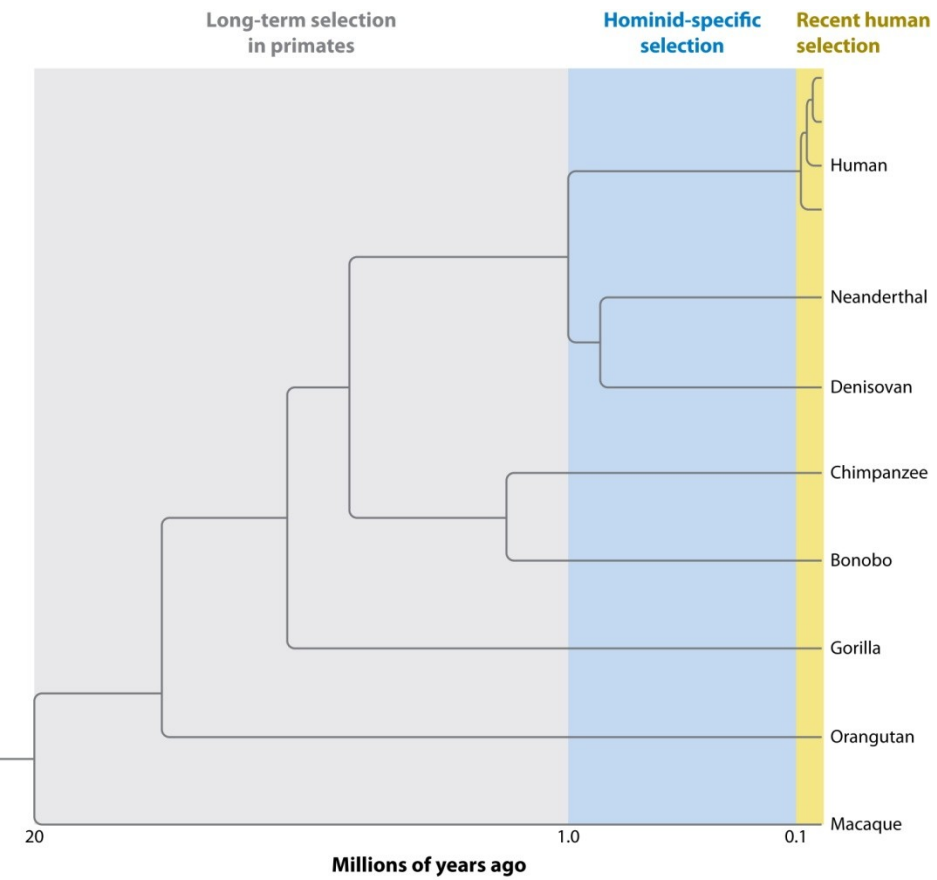


Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

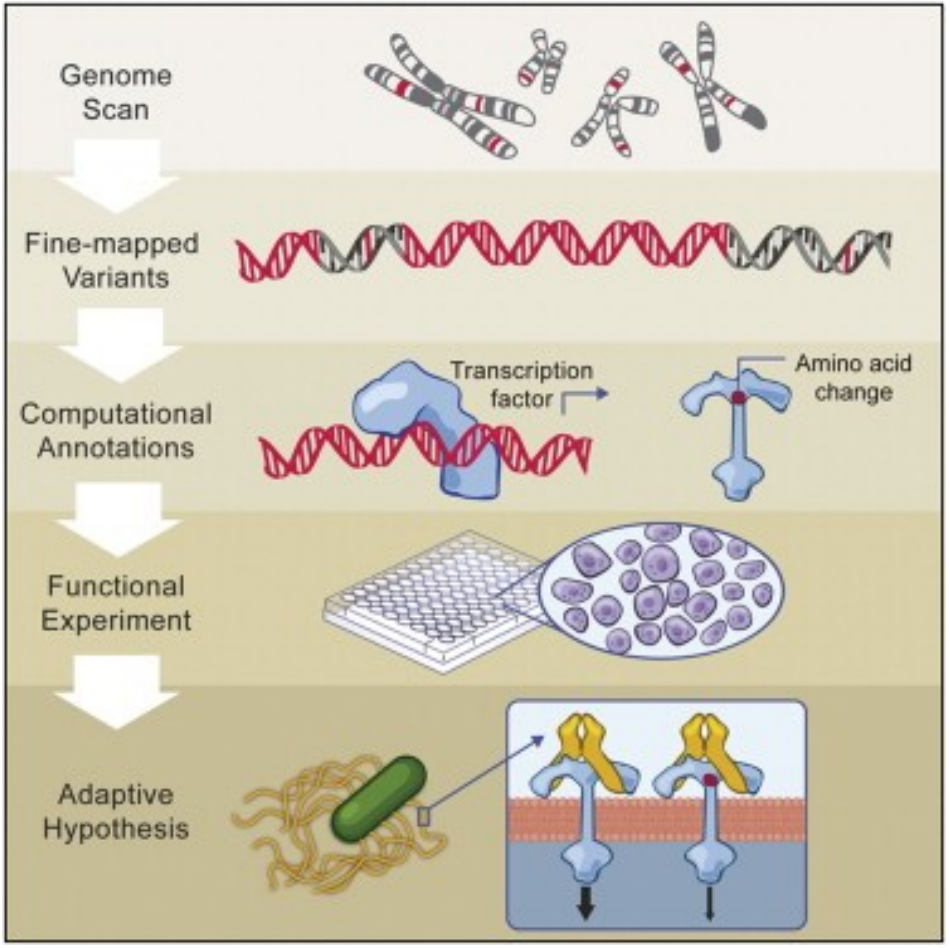


Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

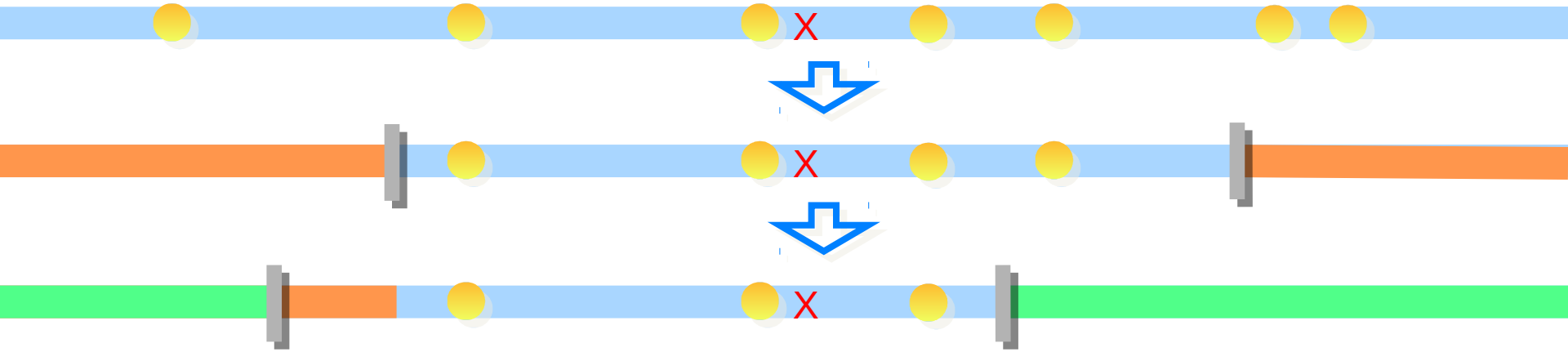
Positive Selection in Human Lineage



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

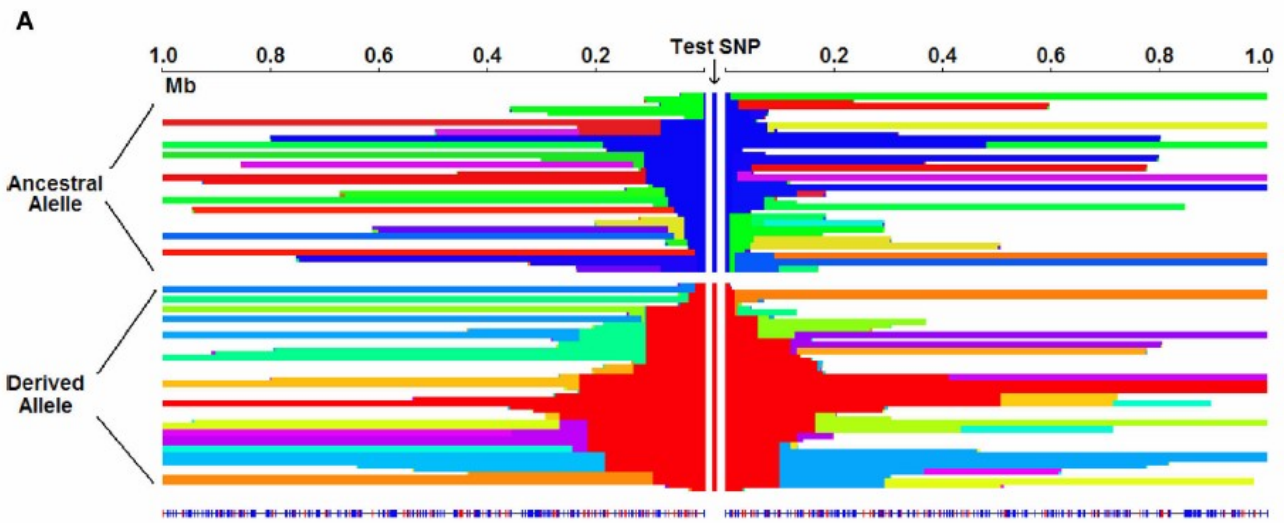


Mutations and LD

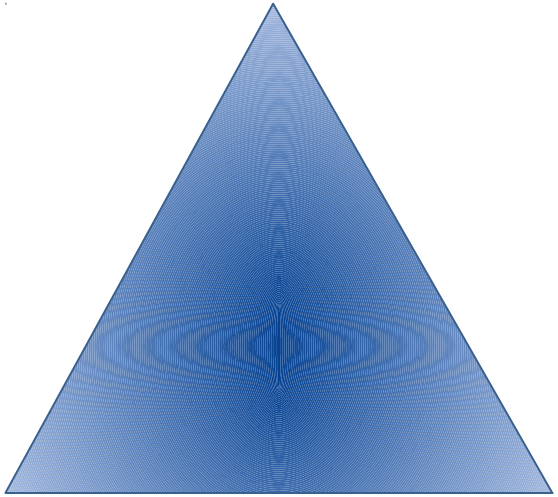


Slide Credits:
Marc Schaub

Long Haplotypes –EHS, iHS tests

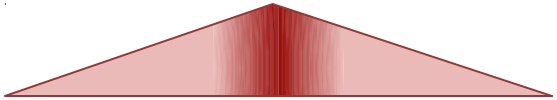


$$iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$



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

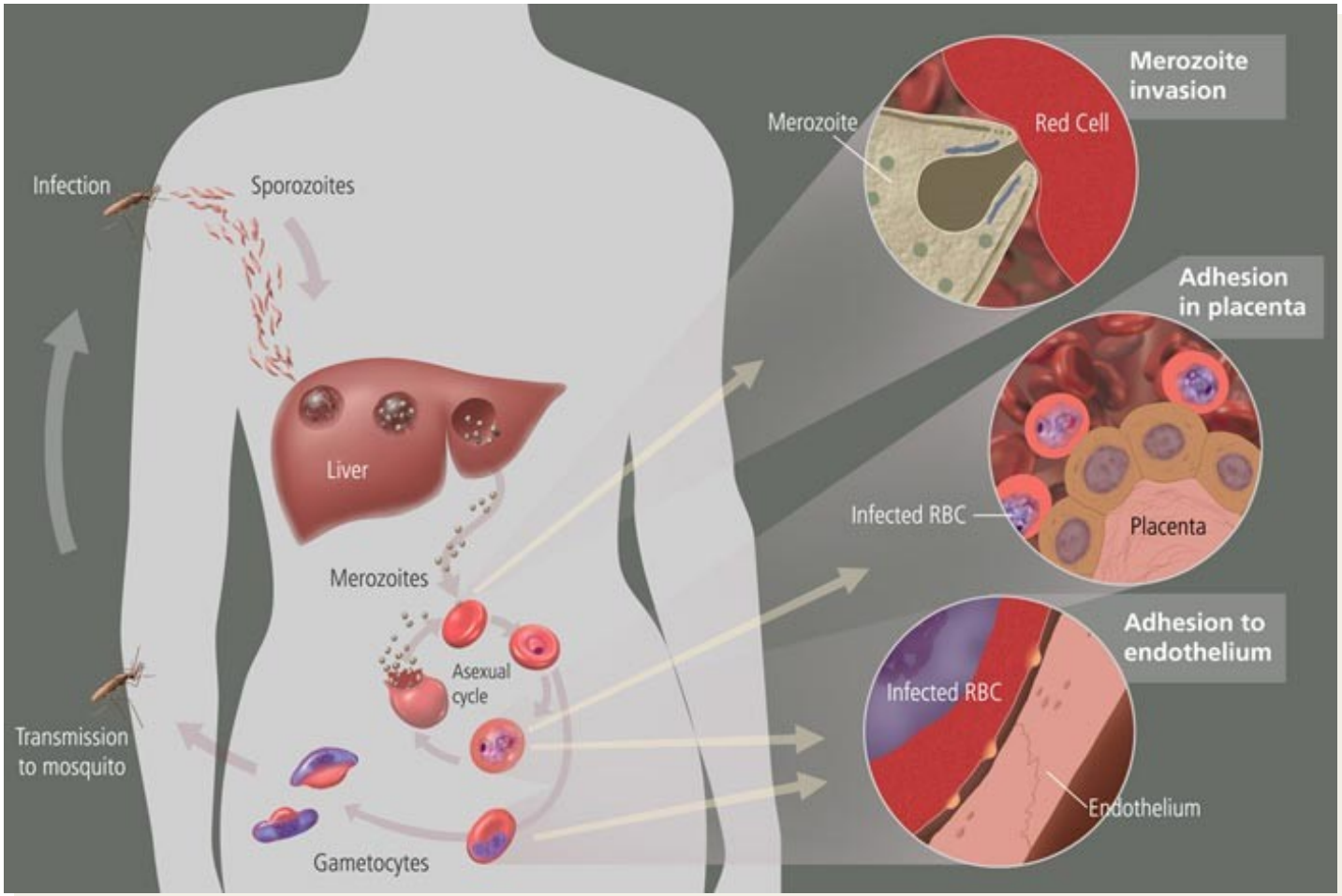


Image source: [NIH - http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg](http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg)

Slide Credits: Marc Schaub

Application: Malaria

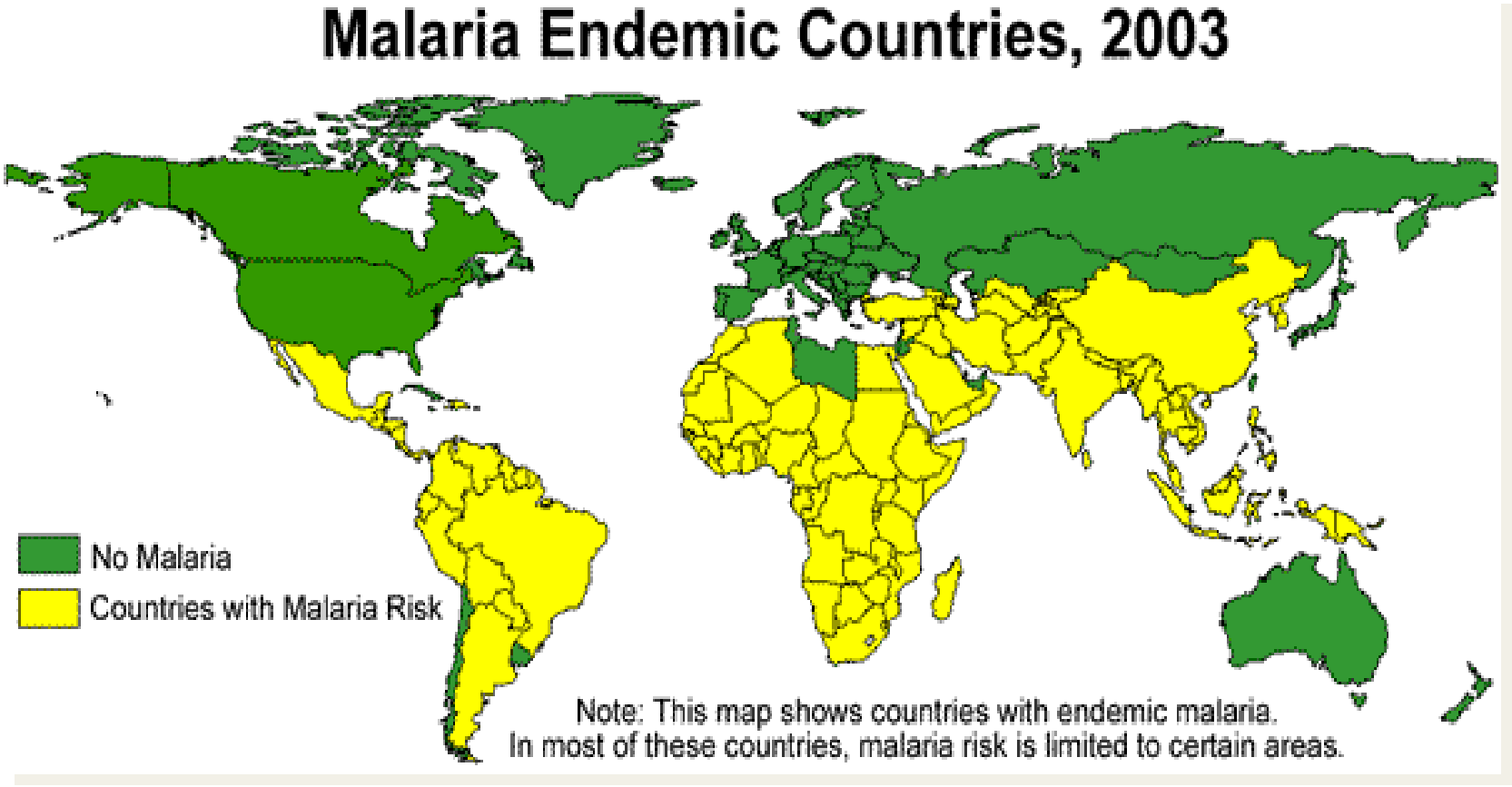
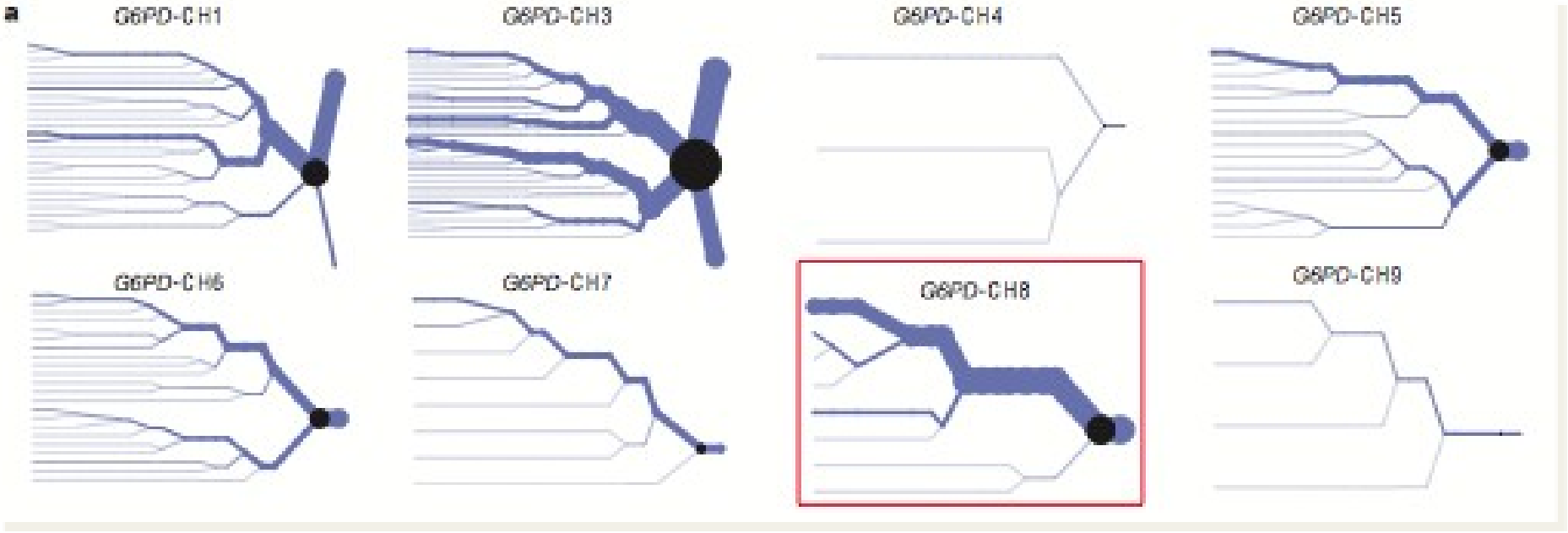


Image source: CDC - http://www.dpd.cdc.gov/dpdx/images/ParasiteImages/M-R/Malaria/malaria_risk_2003.gif

Slide Credits: Marc Schaub

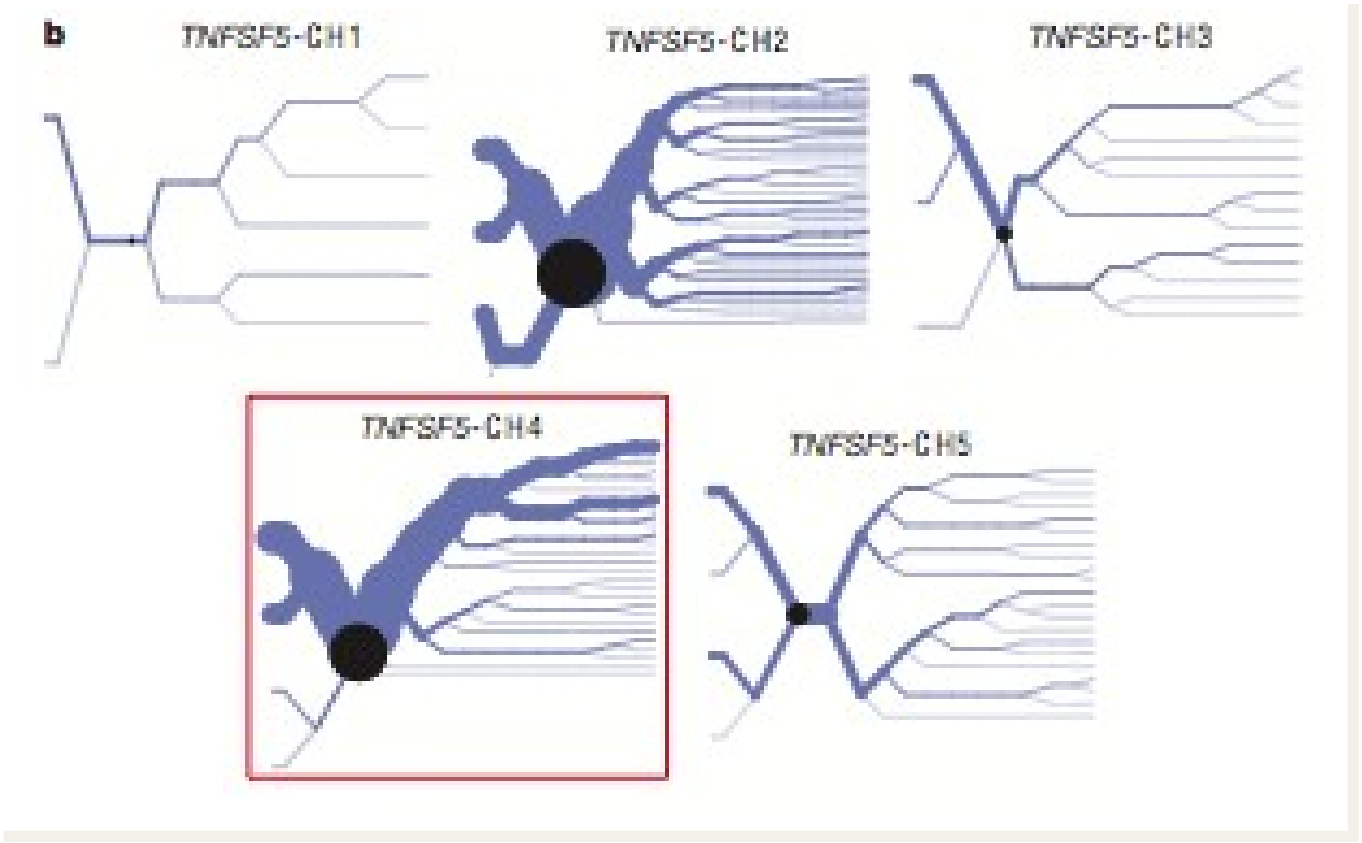
Results: G6PD



Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

Results: TNFSF5



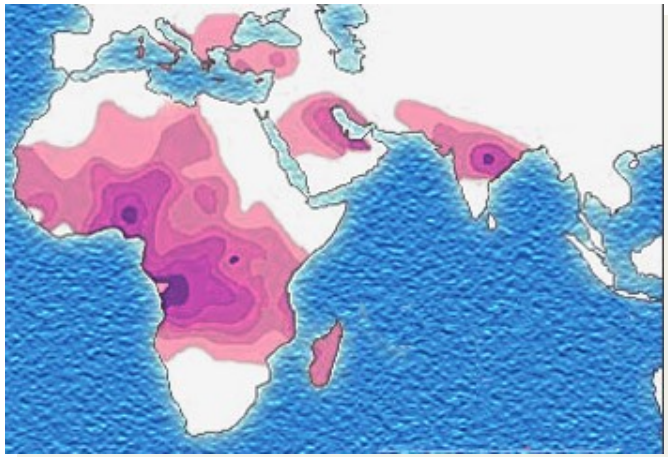
Source: Sabeti *et al.* Nature 2002.

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

Malaria and Sickle-cell Anemia

- Single point mutation in the coding region of the Hemoglobin-B gene (glu → val).

- Heterozygote advantage:
  - Resistance to malaria
  - Slight anemia.

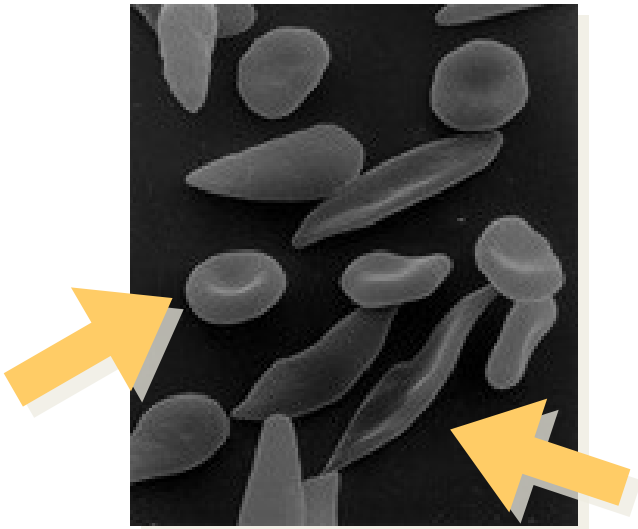


Image source: wikipedia.org

Slide Credits:
Marc Schaub

Lactose Intolerance

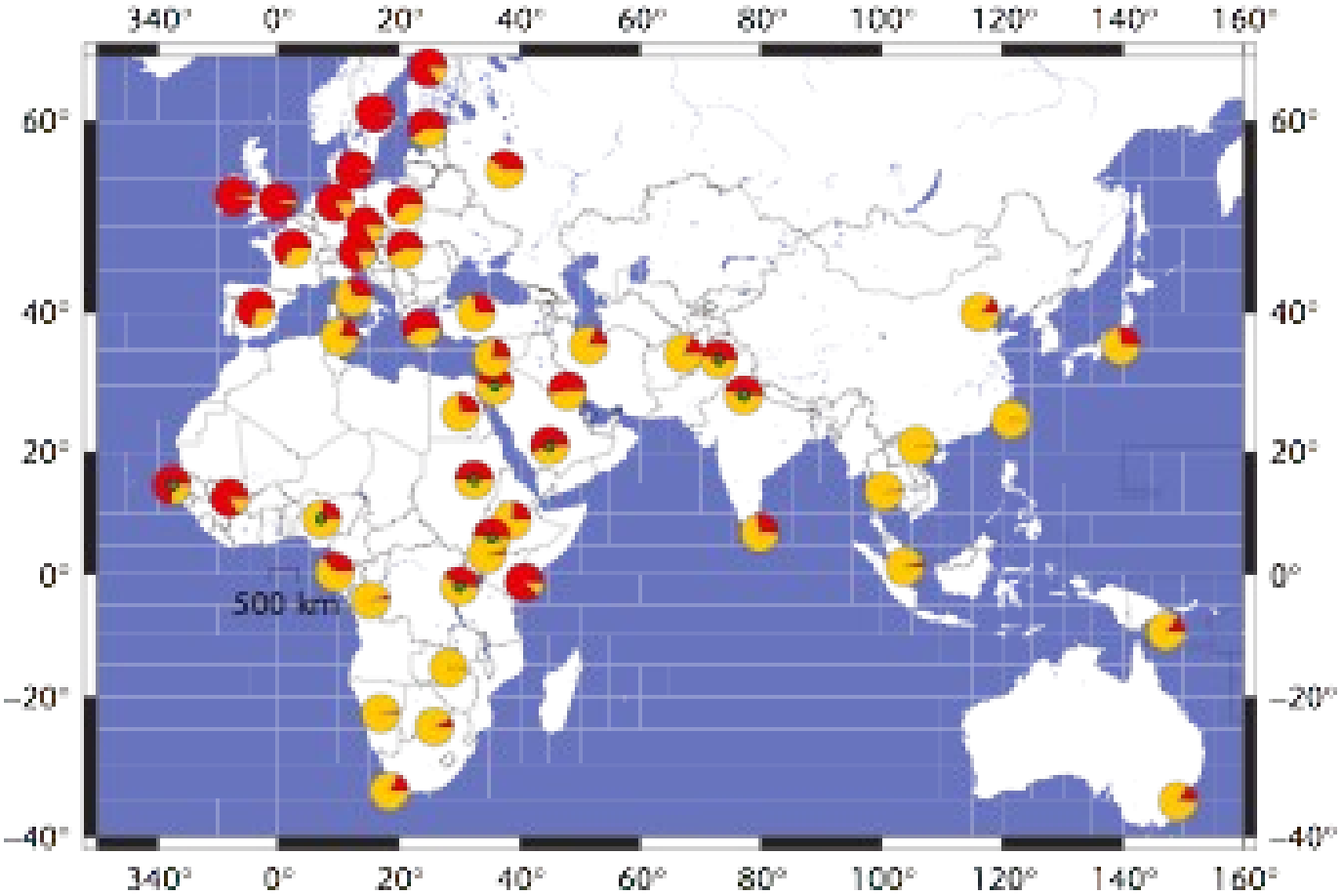
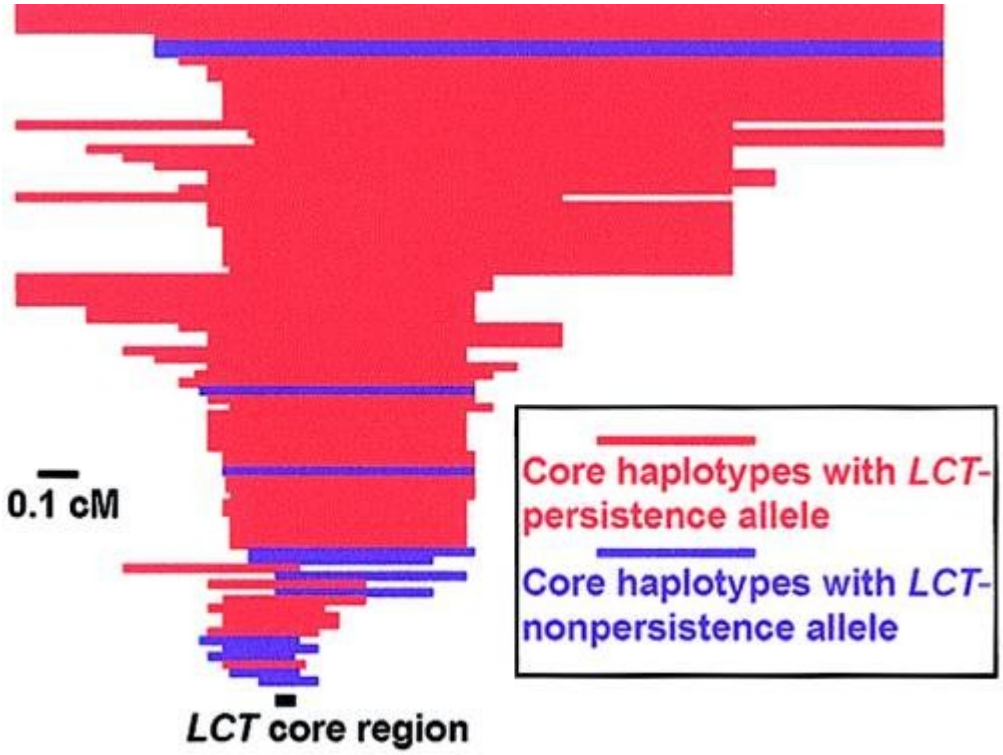


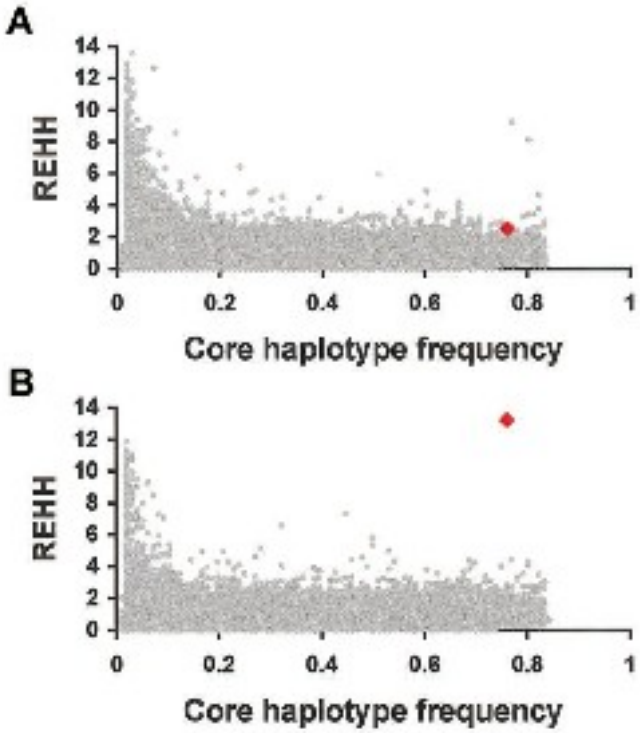
Figure 1 Old world distribution of frequency of lactase persistence (lactose digesters) taken from available published data. Red indicates the proportion of lactose digesters in a given population with yellow representing maldigesters. Charts with a green central circle indicate that the overall published 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





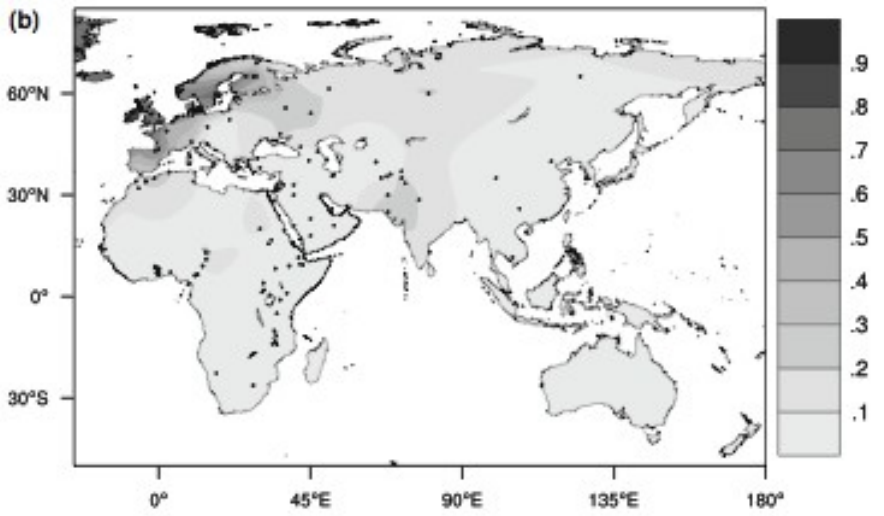
LCT, 5’

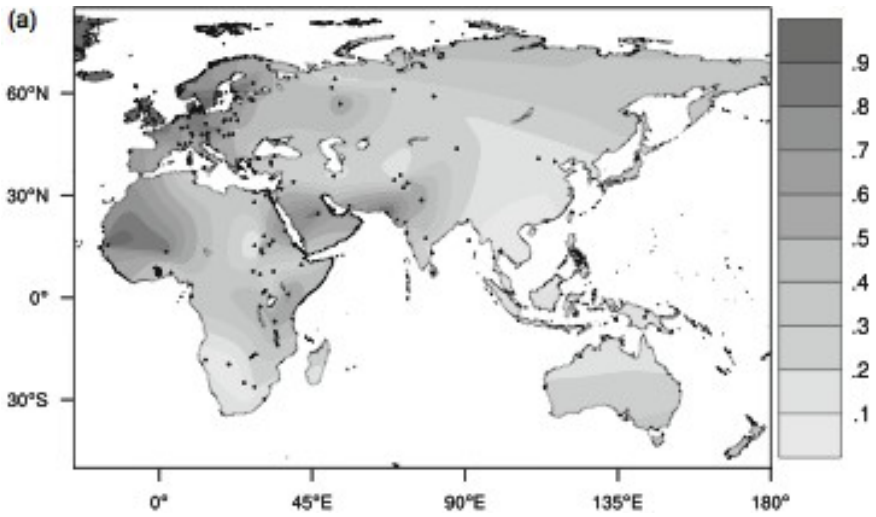
LCT, 3’

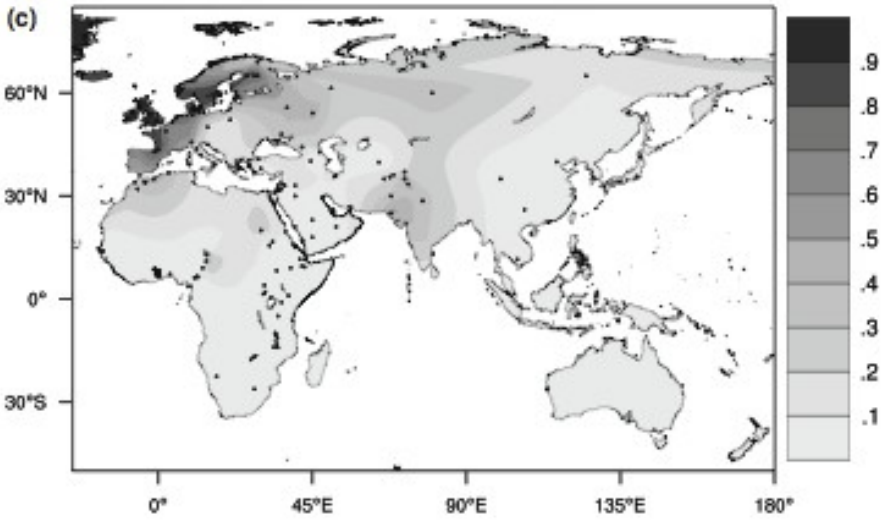
Source: Bersaglieri *et al.* Am. J. Hum. Genet. 2004.

Slide Credits:
Marc Schaub

13910*T distribution







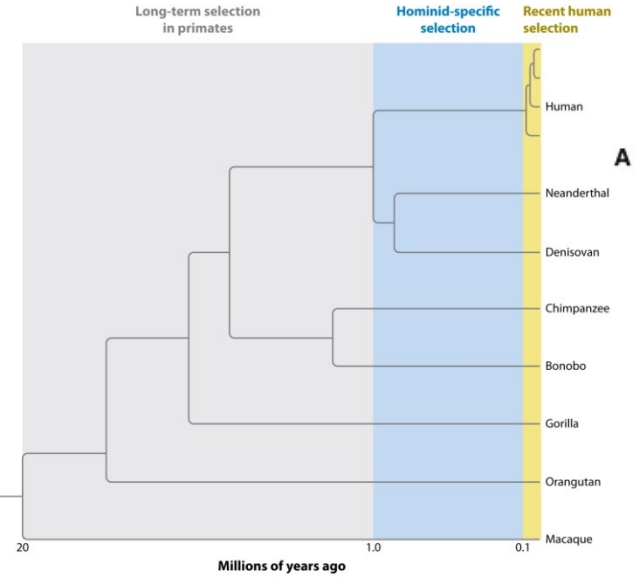
Lactase persistence (litterature)

Predicted lactase persistence

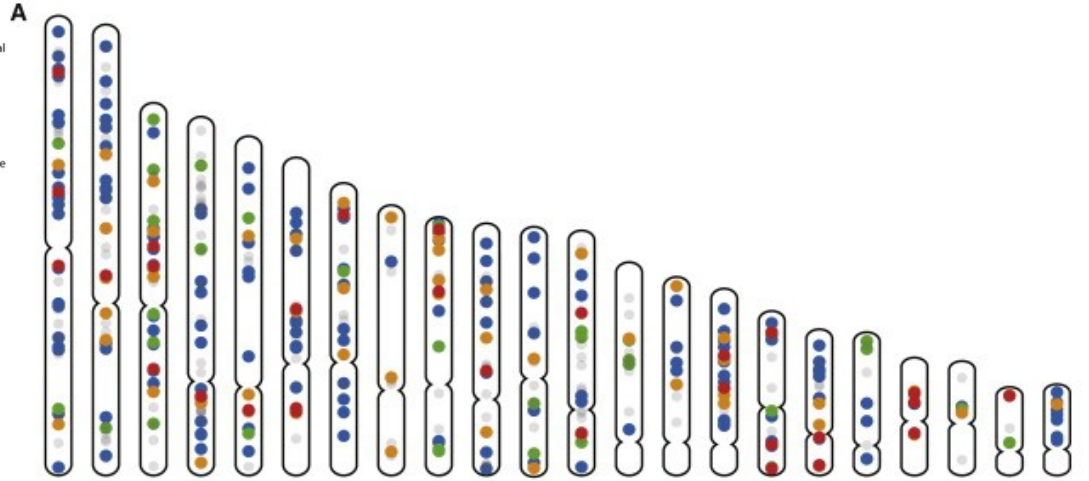
Source: Ingram *et al.* Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet. 2009 Jan;124(6):579-91

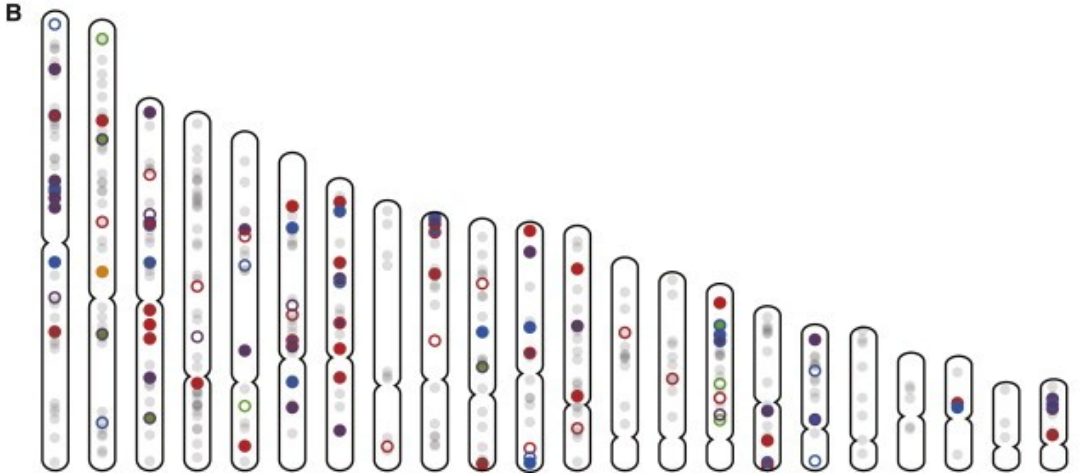
Slide Credits: Marc Schaub

Positive Selection in Human Lineage

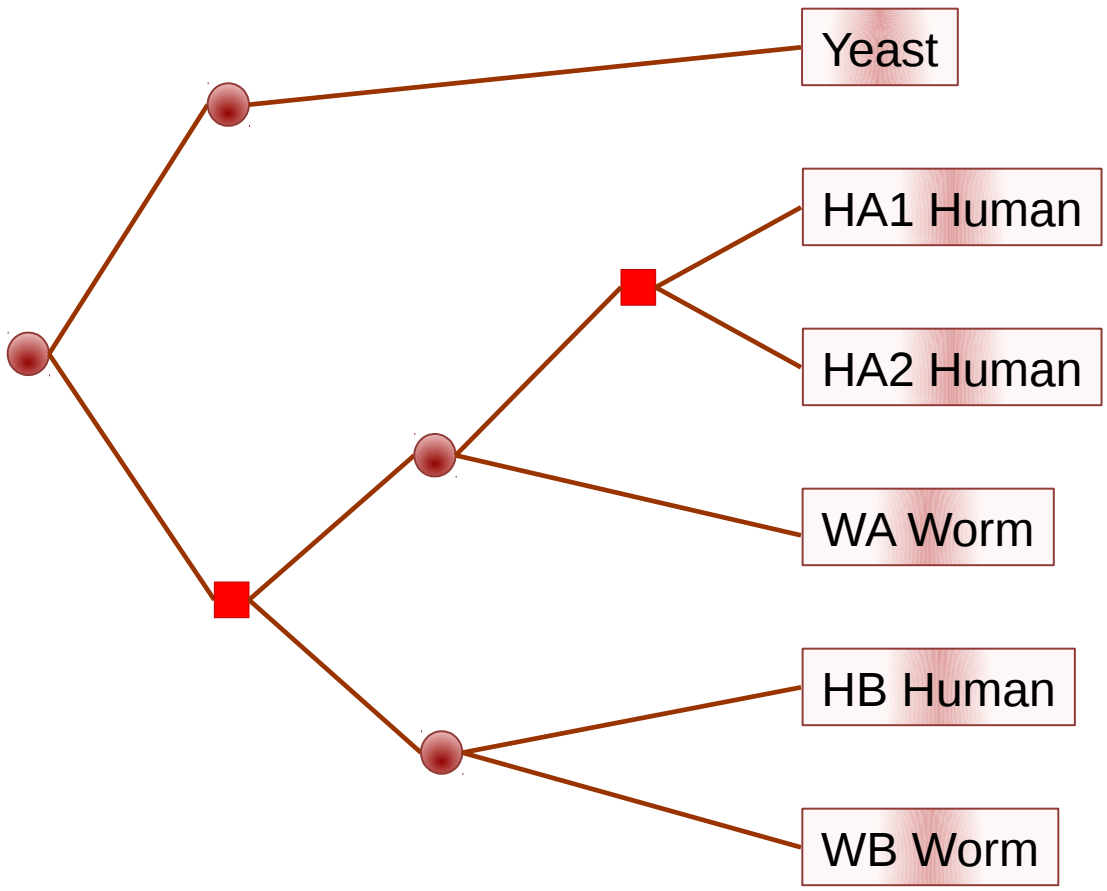


Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89





Orthology and Paralogy



Orthologs: *Derived by speciation*

Paralogs: *Everything else*

Orthology, Paralogy, Inparalogs, Outparalogs

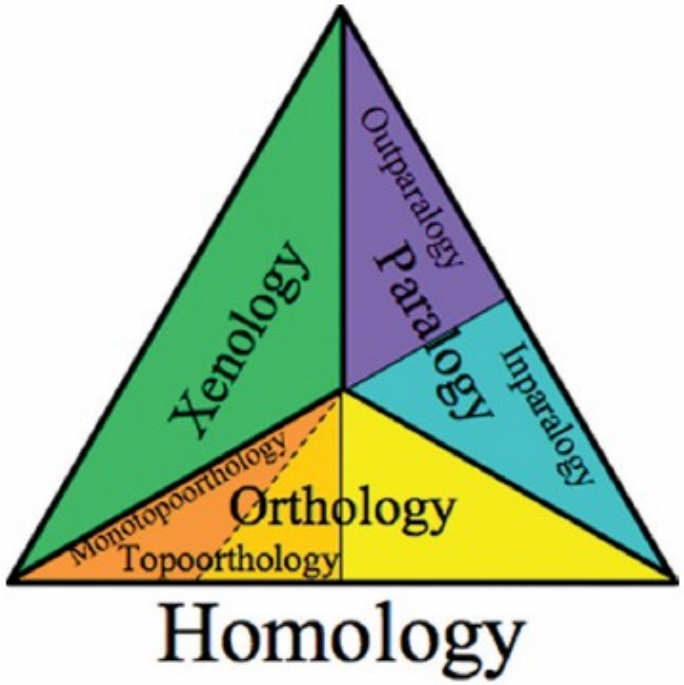


Figure 1. Refinements of homology.

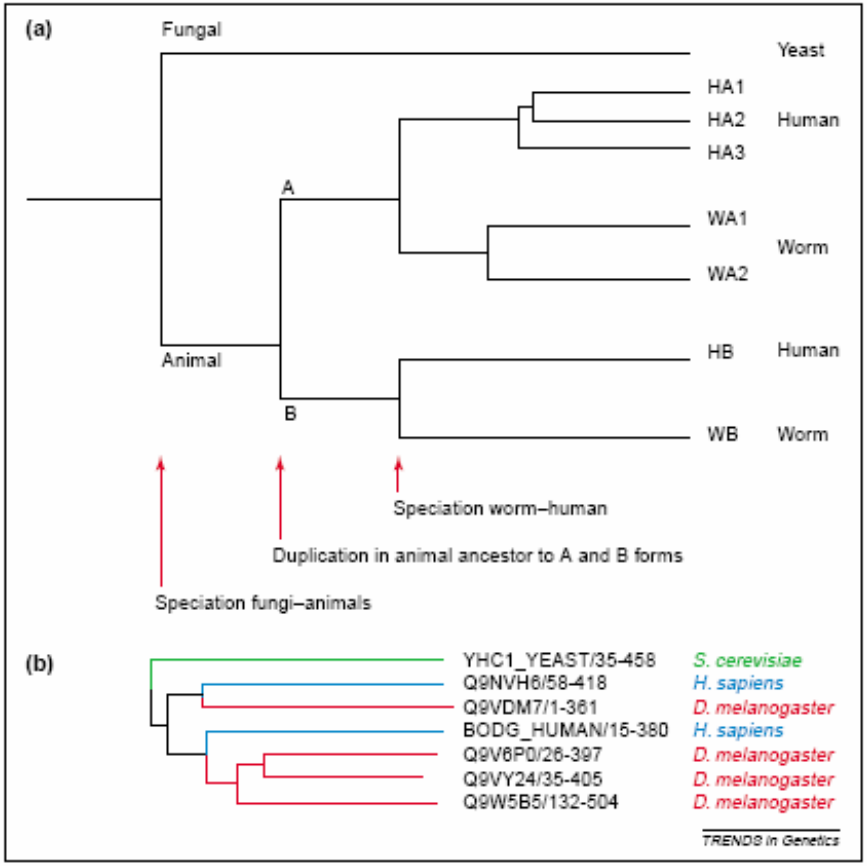


Fig. 1. The definition of inparalogs and outparalogs. (a) Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'inparalogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'outparalogs' when comparing human with worm..However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication. (b) Real-life example of inparalogs: γ-butyrobetaine hydroxylases. The points of speciation and duplication are easily identifiable. The alignment is a subset of Pfam:PF03322 and the tree was generated by neighbor-joining in Belvu. All nodes have a bootstrap support exceeding 95%.