



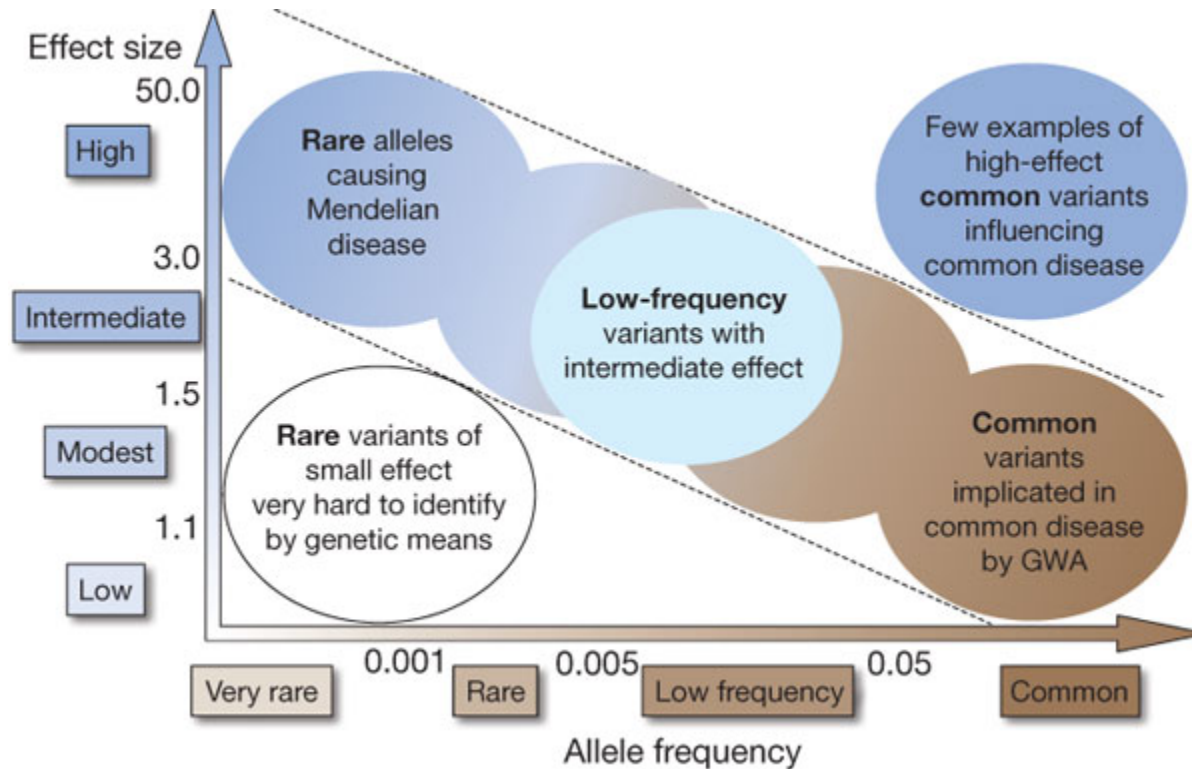
TTGACTGAGGAGTTTACGGGAGCAAAGCGGCGTCATTGCTATTCGTATCTGTTTAG
0101011000100101000010101010011011100110001100101000100101

Human Population Genomics

Heritability & Environment

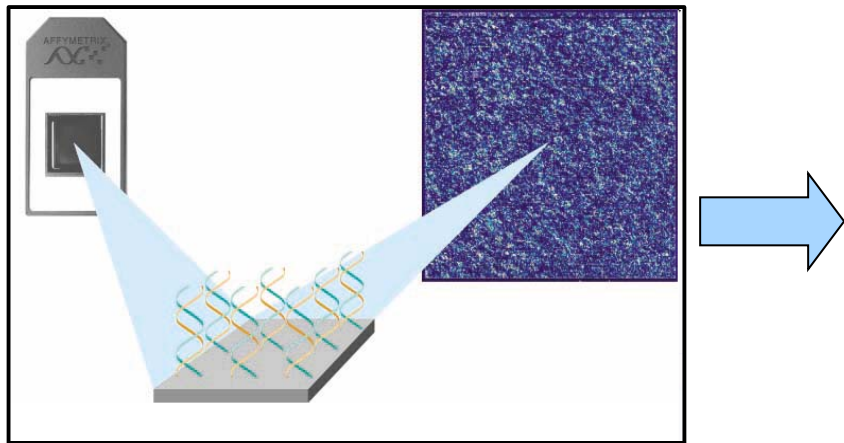


Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).

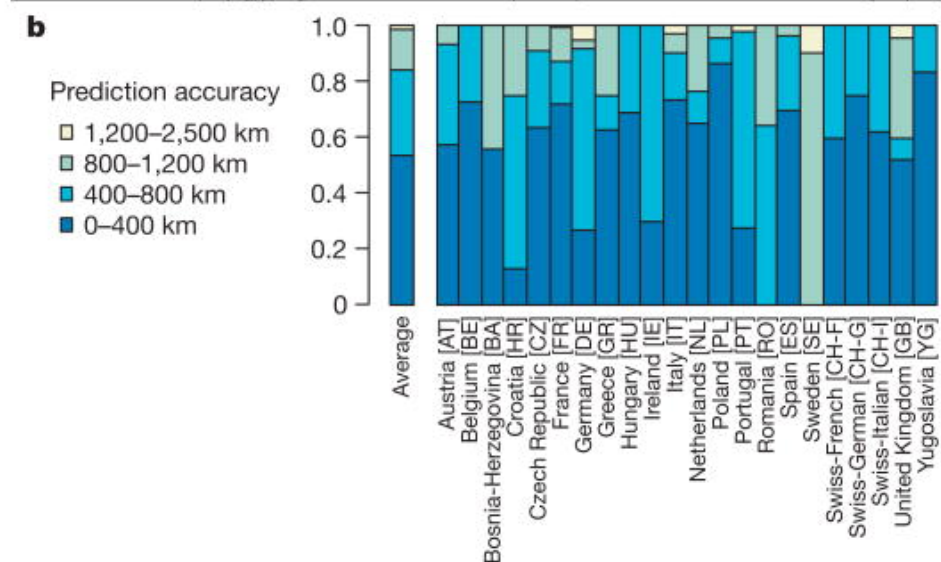
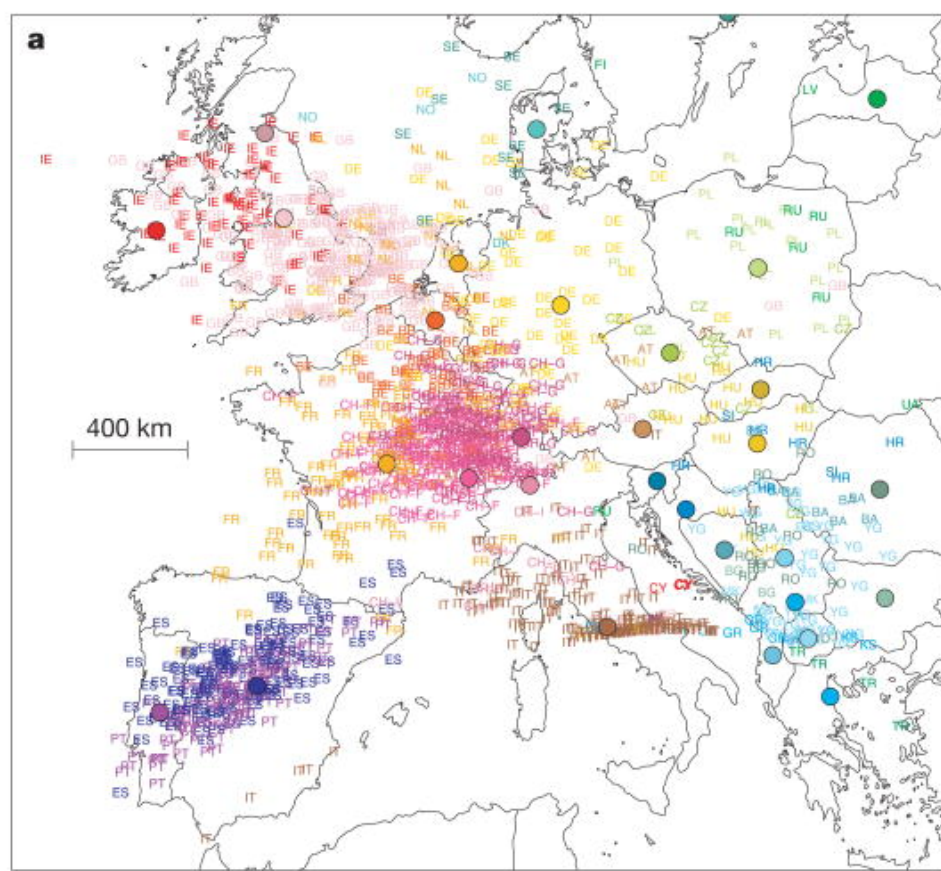


TA Manolio *et al.* *Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Global Ancestry Inference



$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$





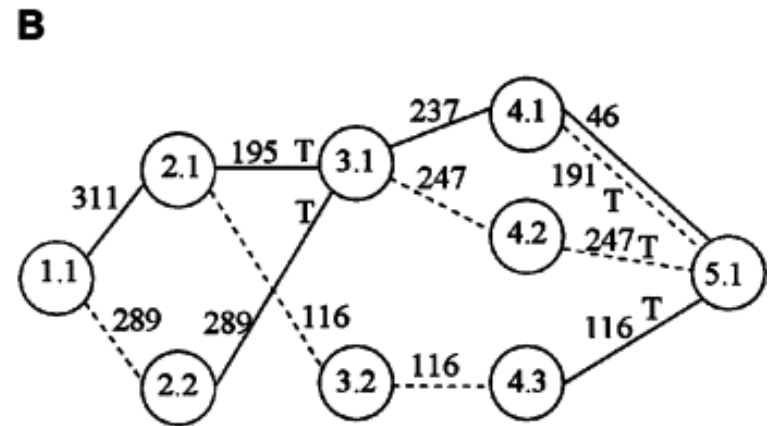
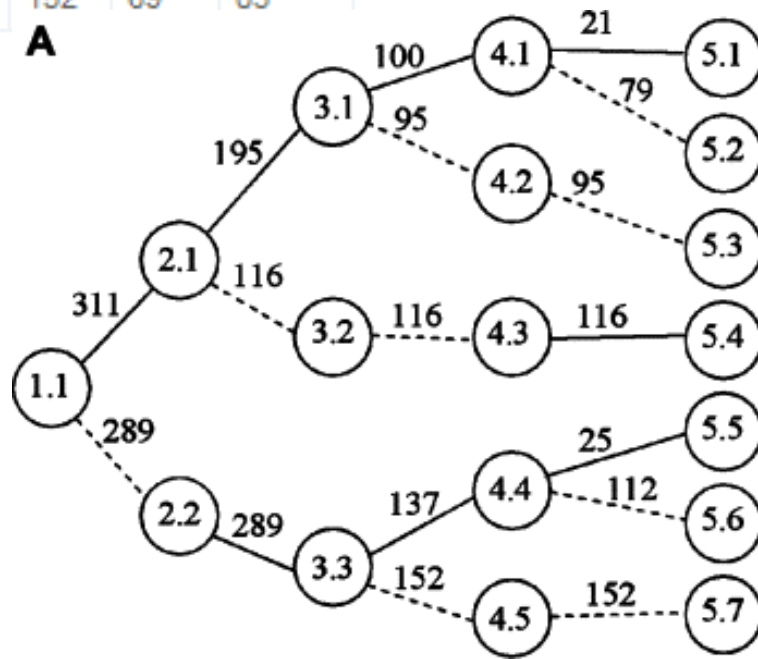
Modeling population haplotypes – VLMC

HAPLOTYPE	COUNT		
	Total	Case	Control
0000	21	12	9
0001	79	43	36
0011	95	43	52
0110	116	59	57
1000	25	14	11
1001	112	60	52
1011	152	69	83

$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

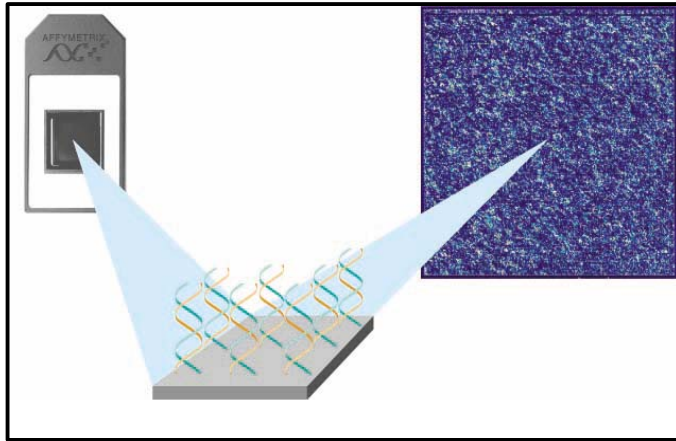
$$G_i = H_{i1} + H_{i2}, \text{ where,}$$

$$H_i = h_{ij1} \dots h_{ijn}; \quad h_{ijk} \in \{0, 1\}$$





Phasing



Haplotype Phasing

Haplotypes

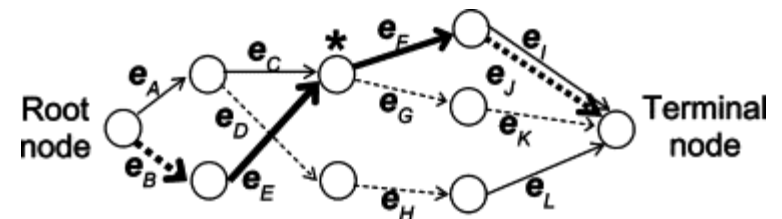
ATCCGA
AGACGC

Genotype

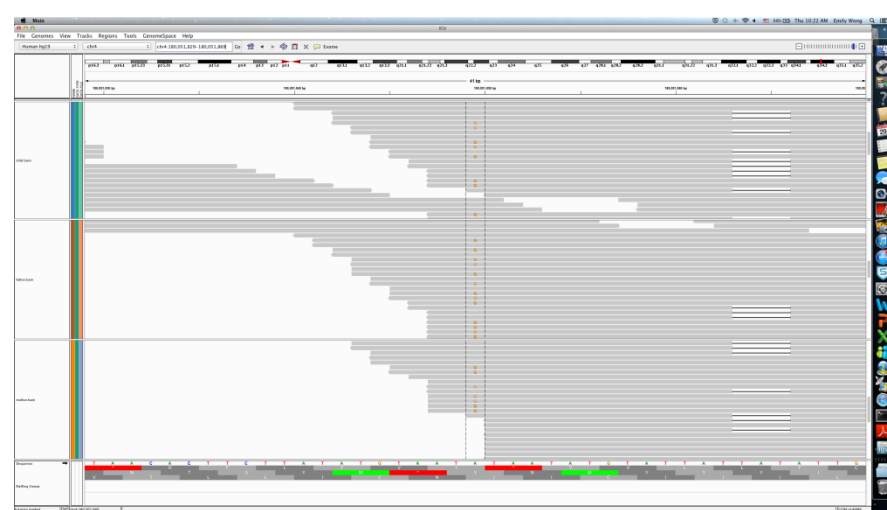
$$A \begin{Bmatrix} T \\ G \end{Bmatrix} \begin{Bmatrix} C \\ A \end{Bmatrix} C G \begin{Bmatrix} C \\ A \end{Bmatrix}$$

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

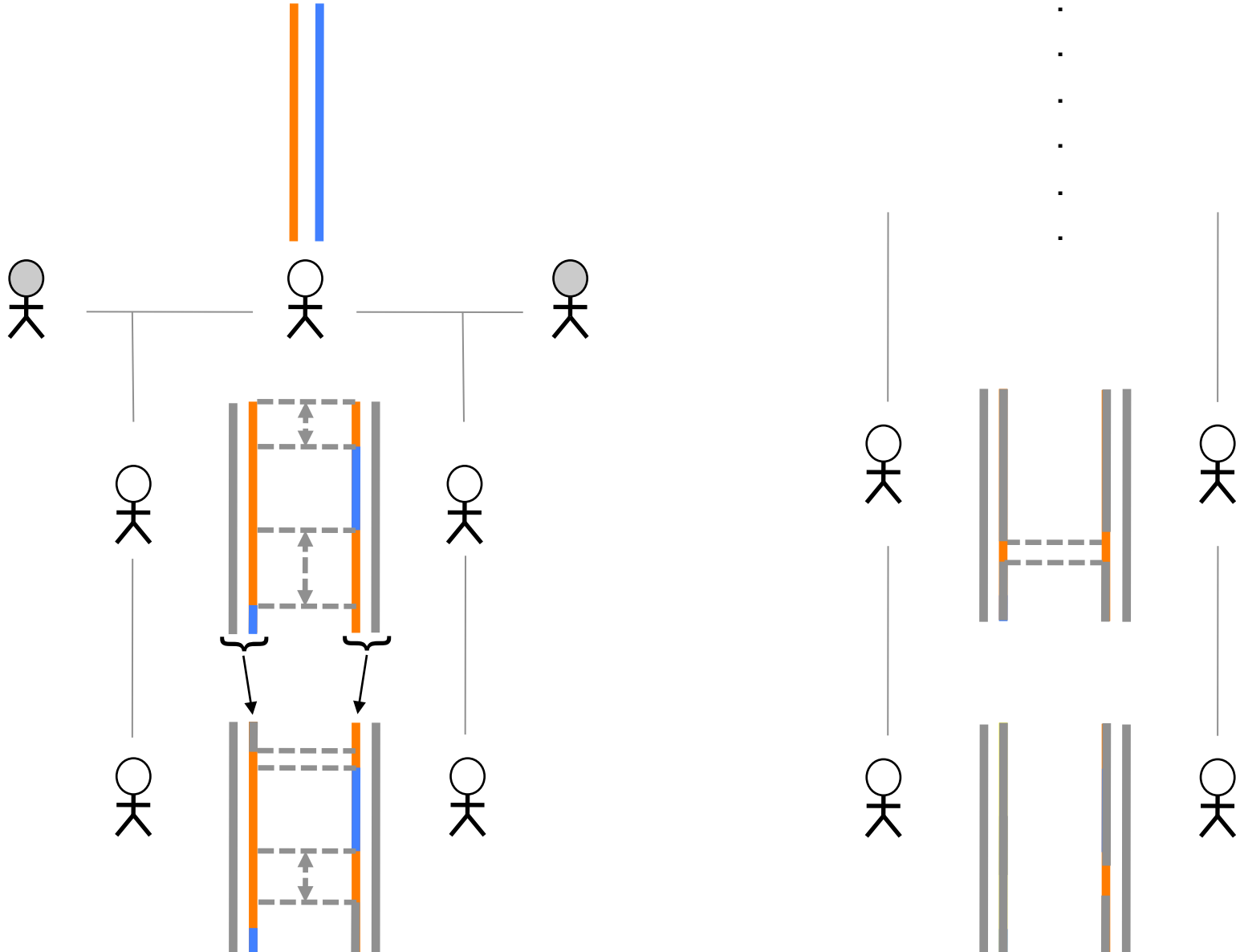
Possible phases: ATACGA AGACGA
AGCCGC ATCCGC

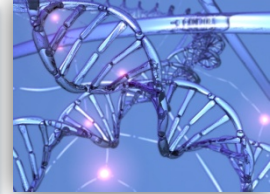


[Browning & Browning, 2007](#)

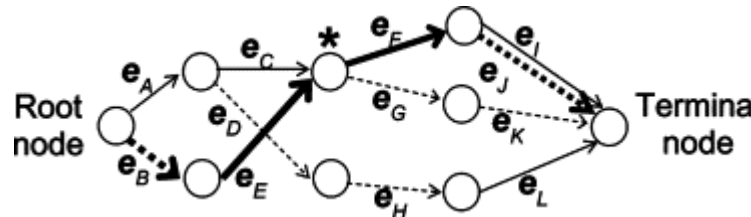
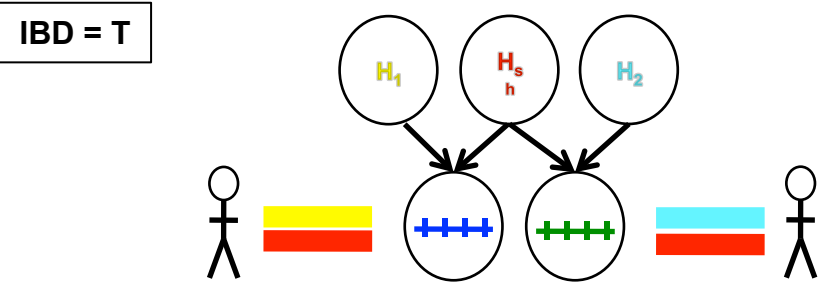
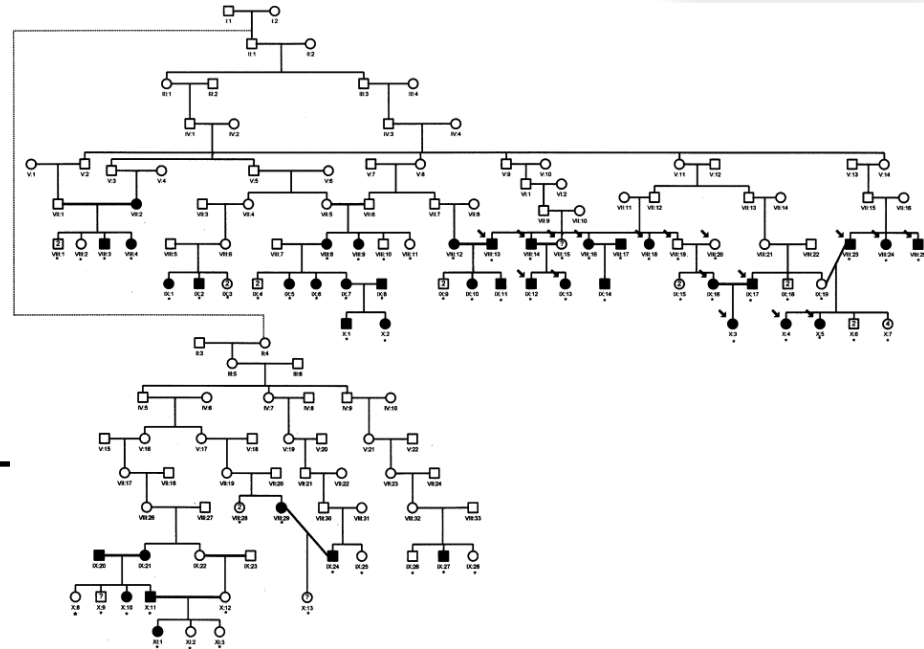
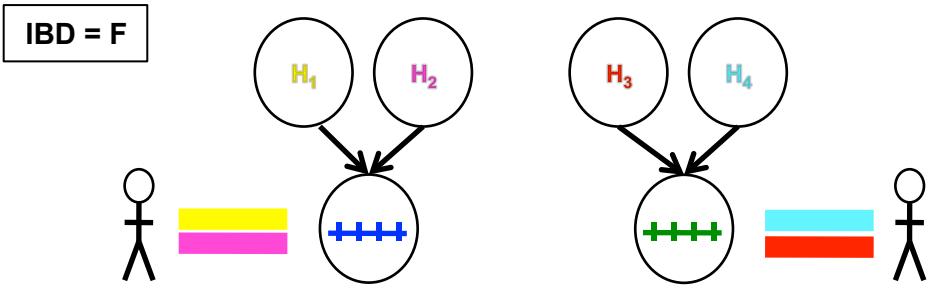


Identity By Descent





IBD detection



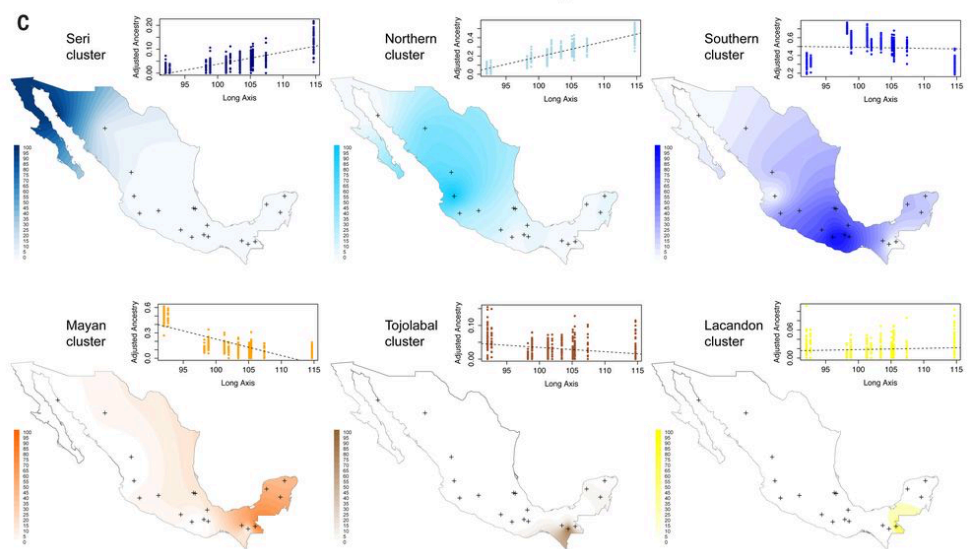
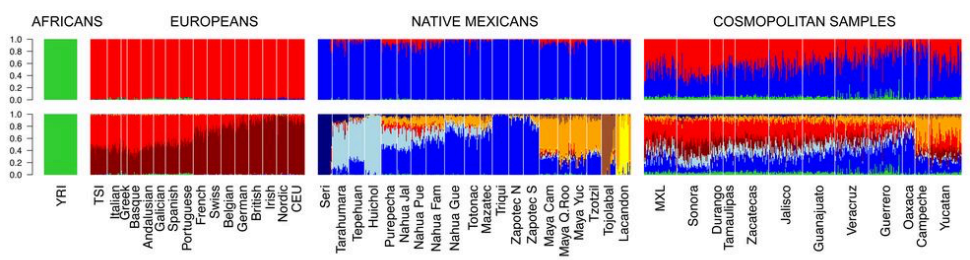
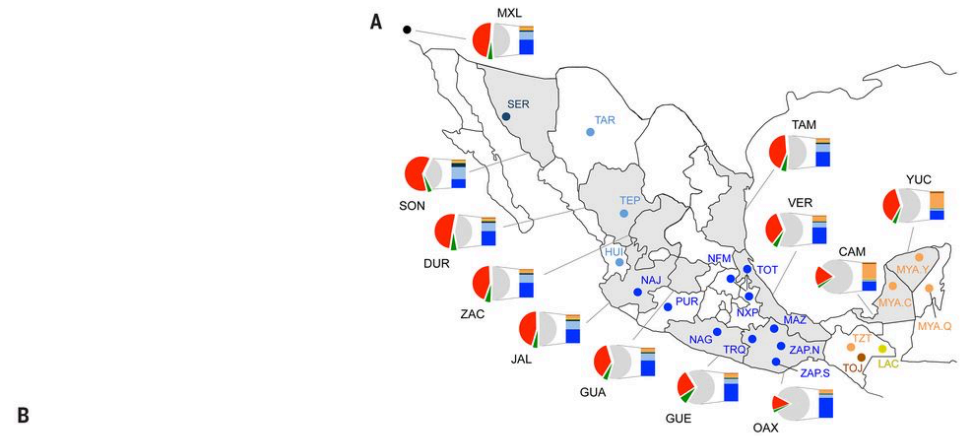
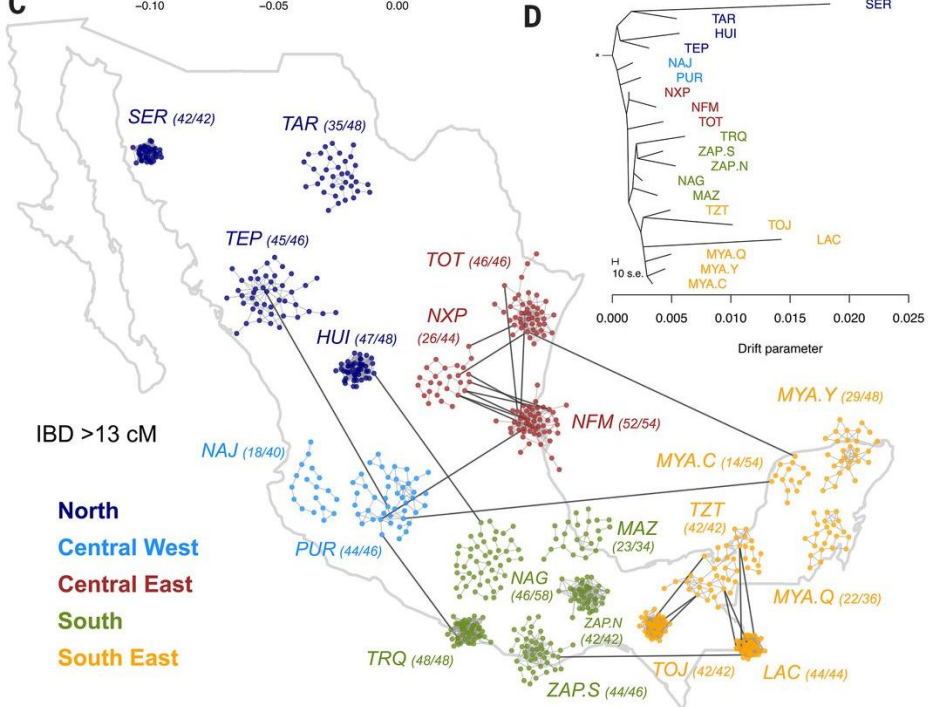
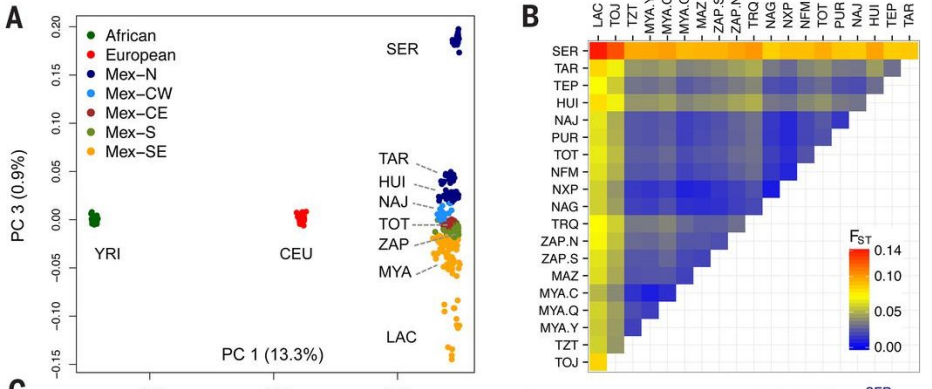
Parente

[Rodriguez et al. 2013](#)

FastIBD: sample haplotypes for each individual, check for IBD

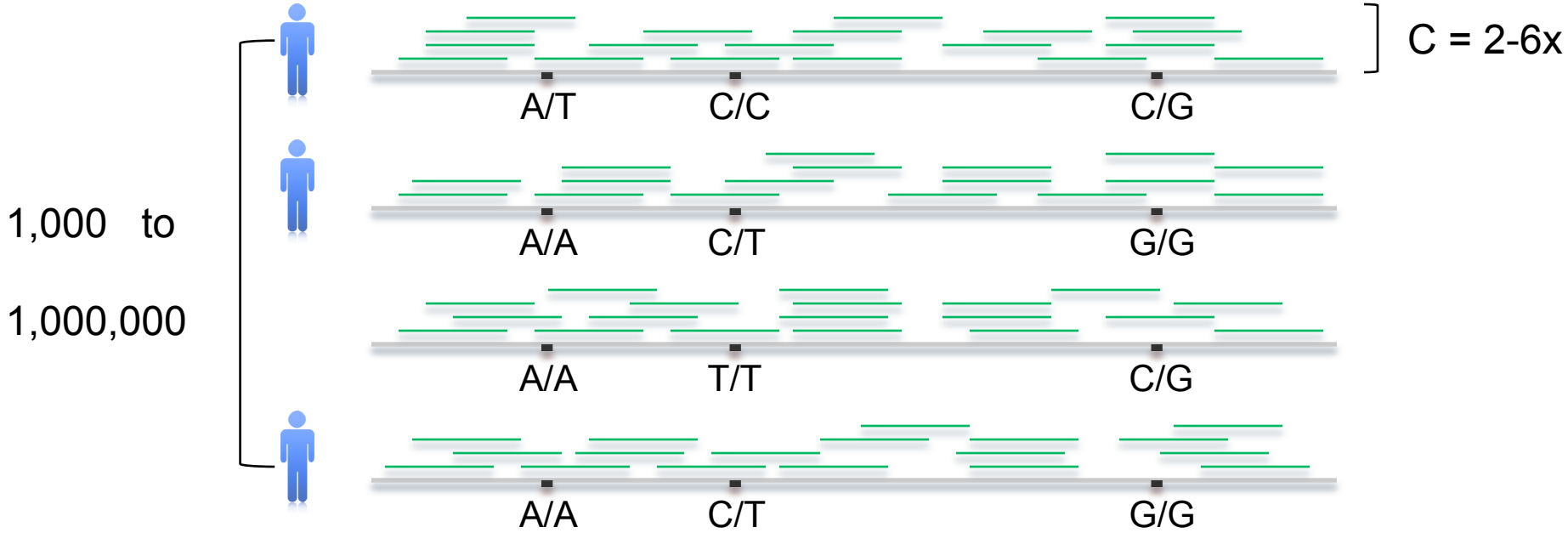
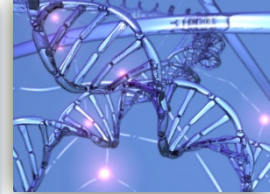
[Browning & Browning 2011](#)

Mexican Ancestry

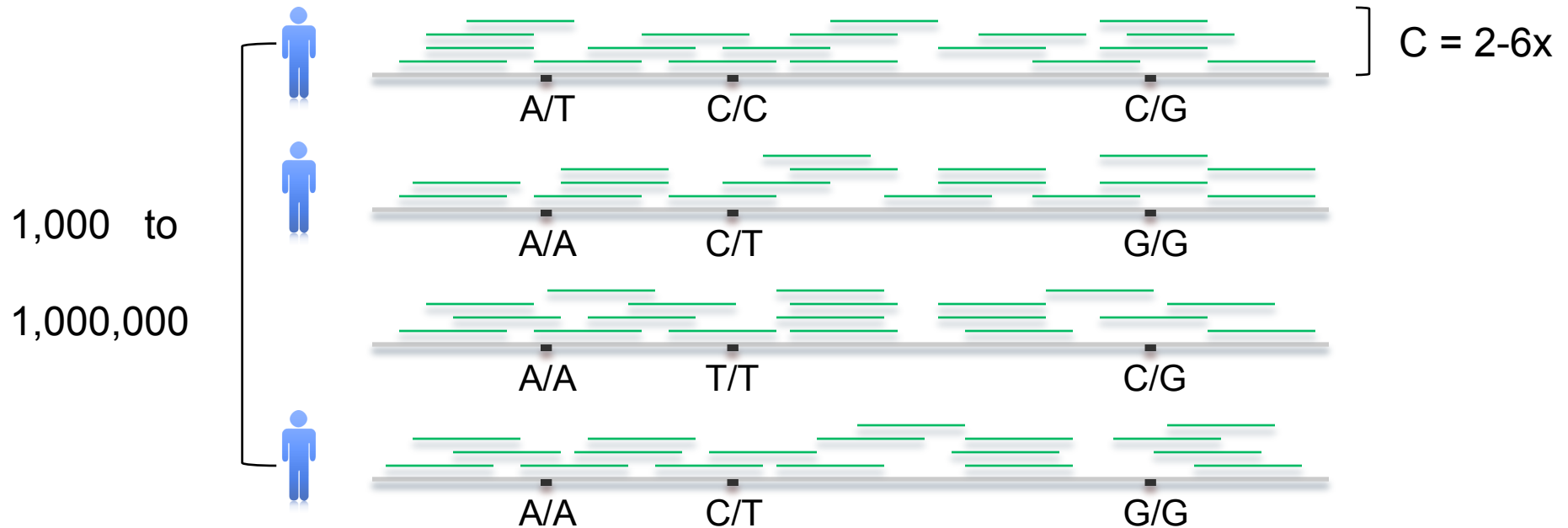


The genetics of Mexico recapitulates Native American substructure and affects biomedical traits, Moreno-Estrada et al. Science, 2014.

Population Sequencing



Population Sequencing



$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$



Population Sequencing

When C is high ($>30x$),

$$\text{Prob}(g_{ij} = k \mid \text{data}) \sim$$

$$\text{Prob}(g_{ij} = k \mid \text{reads mapping on } (i, j))$$

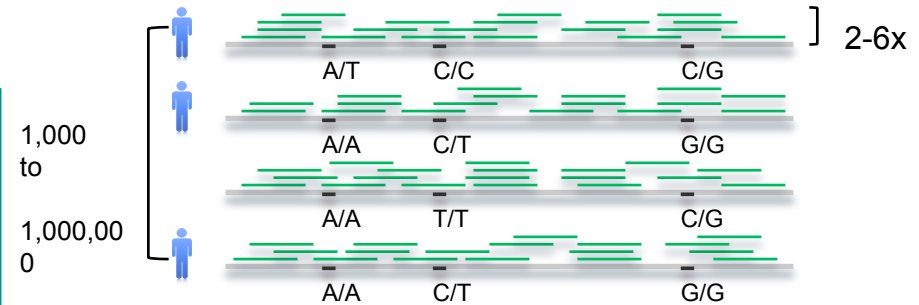
fast & easy

When C is low,

$\text{Prob}(g_{ij} = k \mid \text{data})$ needs to leverage LD:

positions $j' \neq j$ in all individuals

in principle, intractable



$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$



Population Sequencing

Summarization - Maximization:

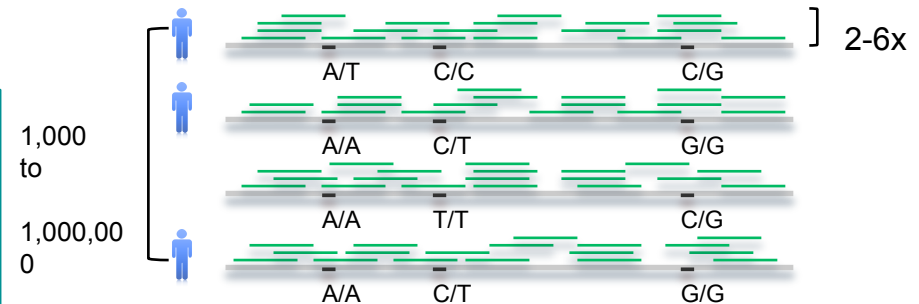
1. Identify candidate polymorphic sites
2. Initialize $G^{(0)}$
3. Summarization:

$$p^{(n+1)}_{ijk} = \text{Prob}(g_{ij} = K \mid G^{(n)}, \text{data})$$

4. Maximization:

$$g^{(n+1)}_{ijk} = \text{argmax}_{ijk} p^{(n+1)}_{ijk}$$

5. Repeat until convergence

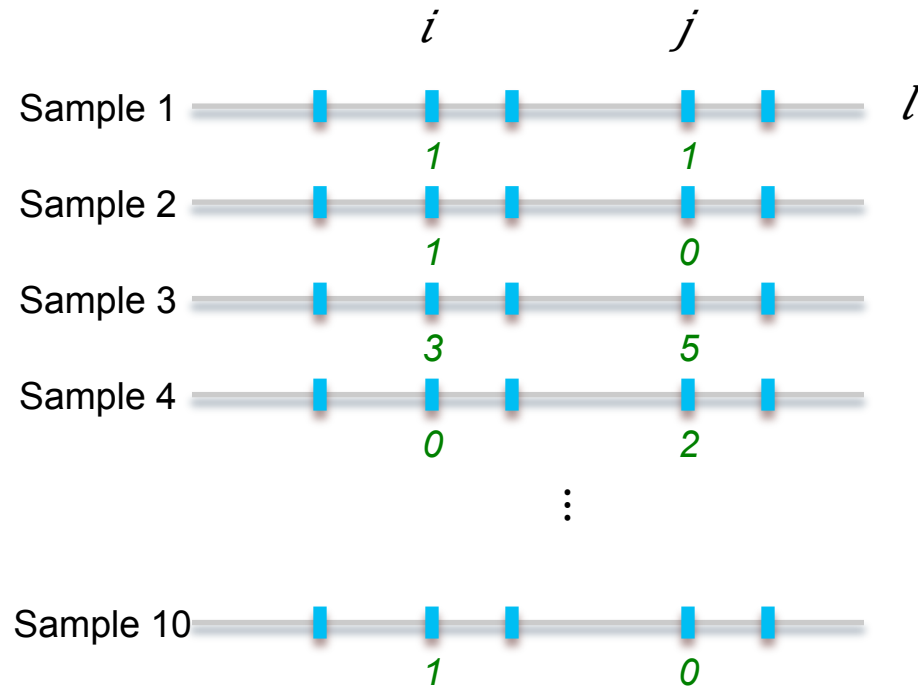


$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$



Modeling LD: Nearest Neighbors



Let

$S_i = \{ \text{samples with } \geq \text{one read covering minor allele} \}$

$S_i = \{1, 2, 3, 10\}$

$S_j = \{1, 3, 4\}$

Then,

$$\text{Sim}_1(i, j) = (S_i \cap S_j) / (S_i \cup S_j) = 2/4$$



Reveel Algorithm

Summarization Maximization:

1. Identify candidate polymorphic sites

2. Initialize $G^{(0)}$

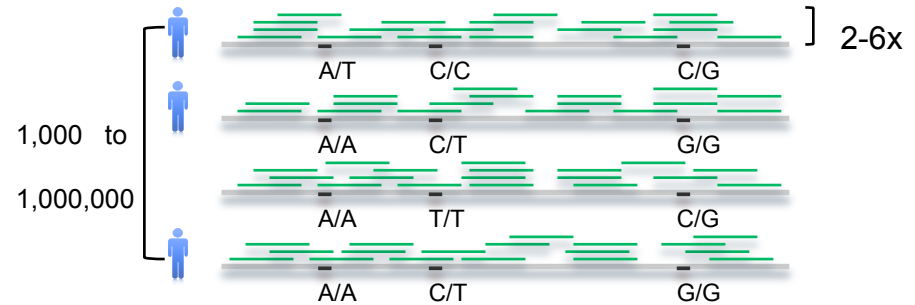
3. Summarization:

$$p^{(n+1)}_{ijk} = \text{Prob}(g_{ij} = K \mid G^{(n)}, \text{data})$$

4. Maximization:

$$g^{(n+1)}_{ijk} = \text{argmax}_{ijk} p^{(n+1)}_{ijk}$$

5. Repeat until convergence



$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$

Candidate Polymorphic site

Essentially, pos'n j where some individuals have at least 2 reads with same minor allele



Reveel Algorithm

Summarization Maximization:

1. Identify candidate polymorphic sites

2. Initialize $G^{(0)}$

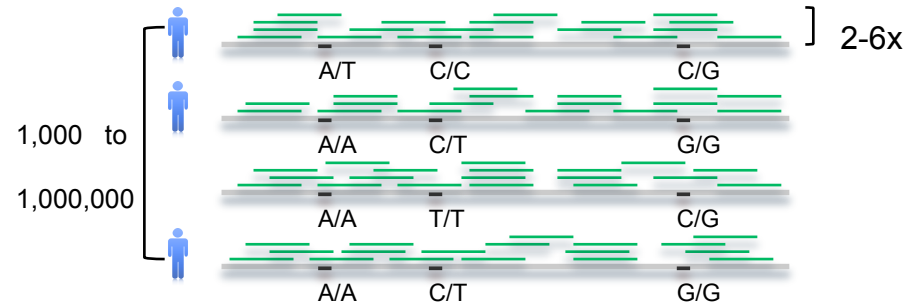
3. Summarization:

$$p^{(n+1)}_{ijk} = \text{Prob}(g_{ij} = K \mid G^{(n)}, \text{data})$$

4. Maximization:

$$g^{(n+1)}_{ijk} = \text{argmax} p^{(n+1)}_{ijk}$$

5. Repeat until convergence



$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$

At each position j ,

Use sum of read counts at j and its nearest neighbors

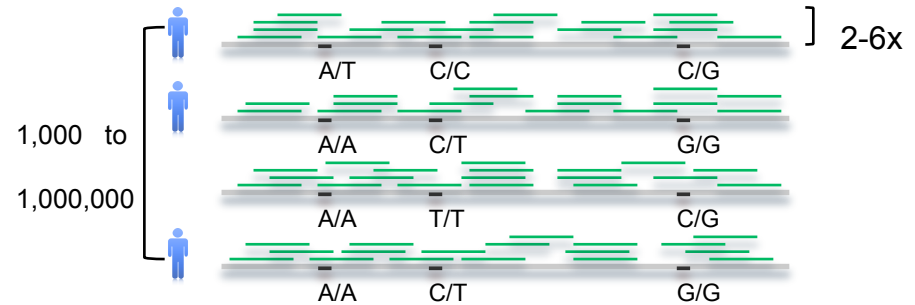


Reveel Algorithm: calculate $P^{(n+1)}$

$$p^{(n+1)}_{ijk} = P(g_{ij} = k \mid G^{(n)}, \text{reads})$$

$$\sim P(g_{ij} = k \mid g_{kNN}, \text{reads})$$

$$= P(\text{reads} \mid g_{ij} = k) P(g_{ij} = k \mid g_{kNN})$$



$G_1, \dots, G_N; G_i = g_{i1} \dots g_{in}; g_{ij} \in \{0, 1, 2\}$

$P_1, \dots, P_N; P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$

$P(\text{reads} \mid g_{ij} = k)$: easy

$$P(g_{ij} = k \mid g_{kNN}) =$$

Let C_0, C_1, C_2
 = # samples matching i in kNN at (n) ,
 with j^{th} genotype pos'n = 0, 1, 2

Then,

$$P(g_{ij} = k \mid g_{kNN}) = C_k / (C_0 + C_1 + C_2)$$



Reveel Algorithm

Summarization Maximization:

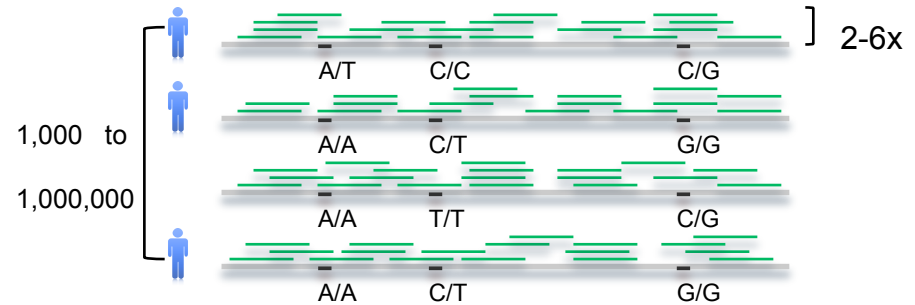
1. Identify candidate polymorphic sites
2. Initialize $G^{(0)}$
3. Summarization:

$$p^{(n+1)}_{ijk} = \text{Prob}(g_{ij} = K \mid G^{(n)}, \text{data})$$

4. Maximization:

$$g^{(n+1)}_{ijk} = \text{argmax}_{ijk} p^{(n+1)}_{ijk}$$

5. Repeat until convergence



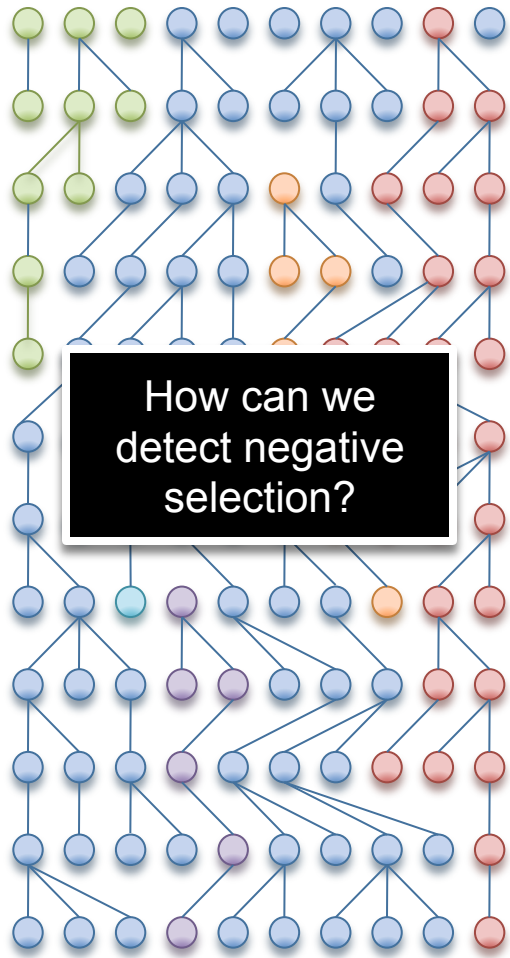
$$G_1, \dots, G_N; \quad G_i = g_{i1} \dots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$$

$$P_1, \dots, P_N; \quad P_i : [p_{ijk} = \text{Prob}(g_{ij} = k \mid \text{data})]$$

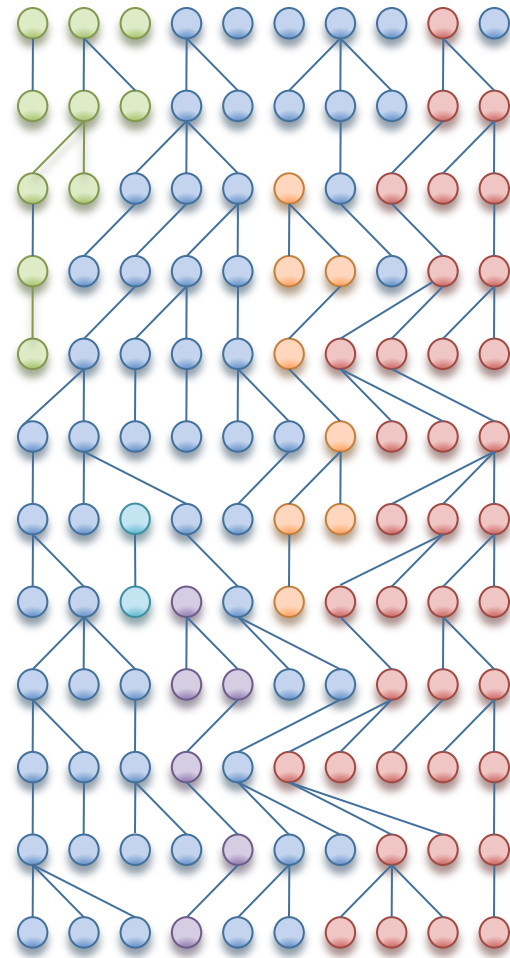
At each position j ,

Use sum of read counts at j and its nearest neighbors

Fixation, Positive & Negative Selection



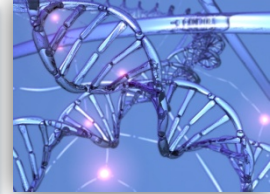
Negative Selection



Neutral Drift



Positive Selection



How can we detect positive selection?

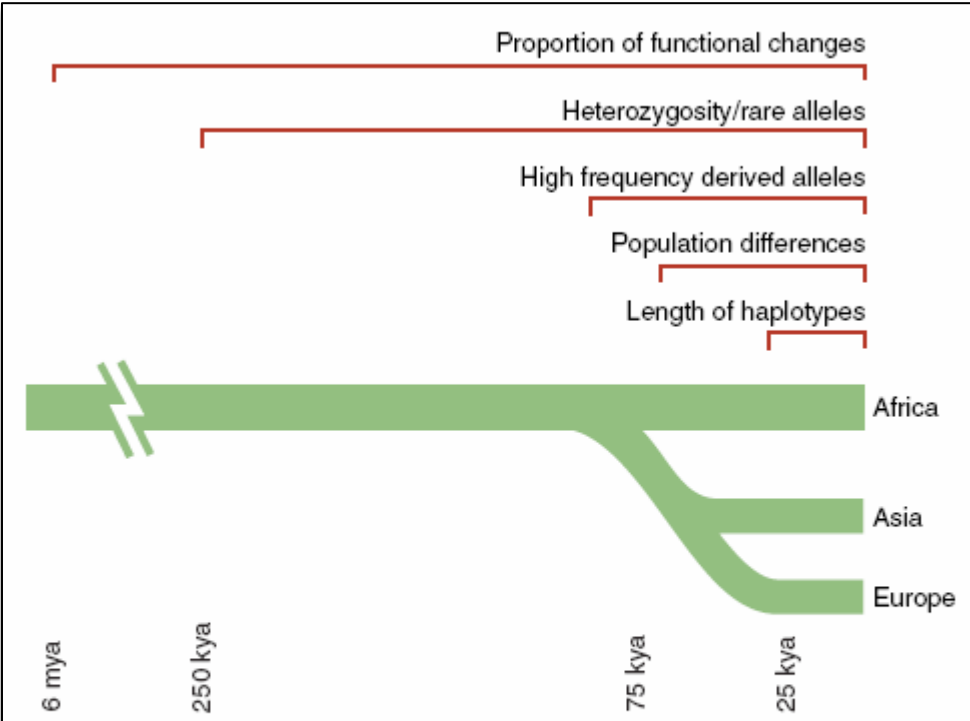


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

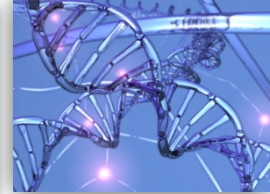
Ka/Ks ratio:
 Ratio of nonsynonymous to synonymous substitutions

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

Examples: immune response, spermatogenesis

		PRM1 Exon 2																													
44 bp		11,341,281	Chromosome 16										11,341,324																		
Human		STOP	H	R	R	C	R	P	R	Y	R	P	R	C	C	R															
		AATCACAGAAGAT	G	T	A	G	C	G	C	C	A	G	A	C	A	T	G	G	A	C	C	C	G	C	C	G	T	C	G	T	G
Chimp		STOP	H	R	R	R	R	M	R	S	R	R	R	C	C	R															
		AATCACAGAAGAT	G	C	A	G	A	G	T	A	A	G	A	C	C	T	G	G	A	C	G	C	C	G	C	C	G	T	C	G	T

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?

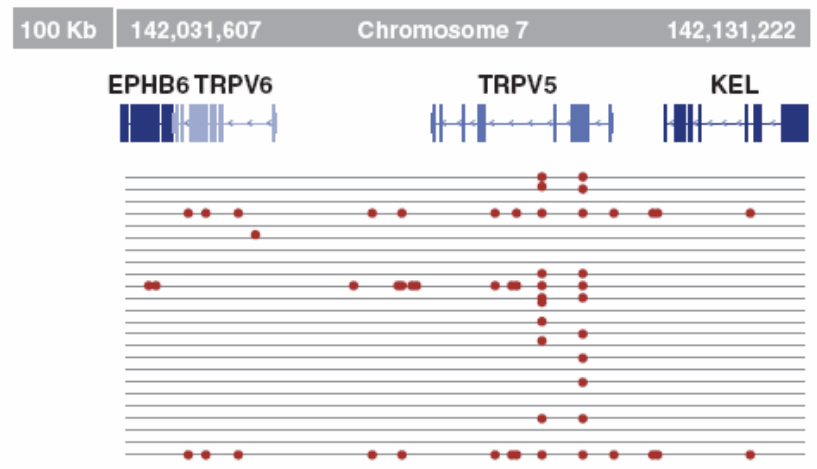


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (28).

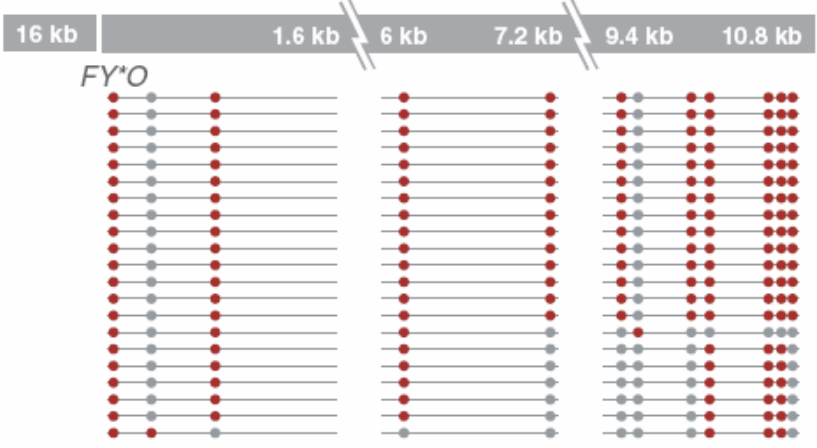


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (34). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

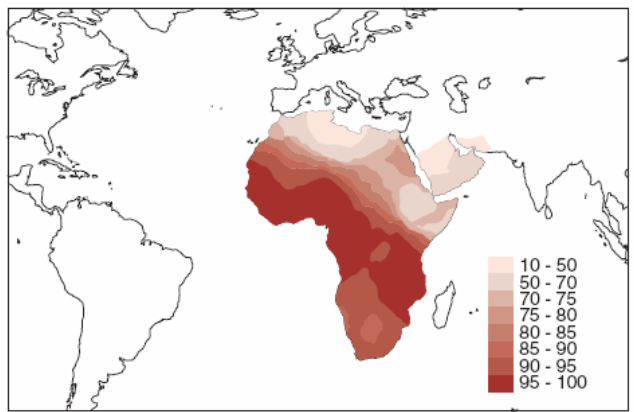


Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (38).

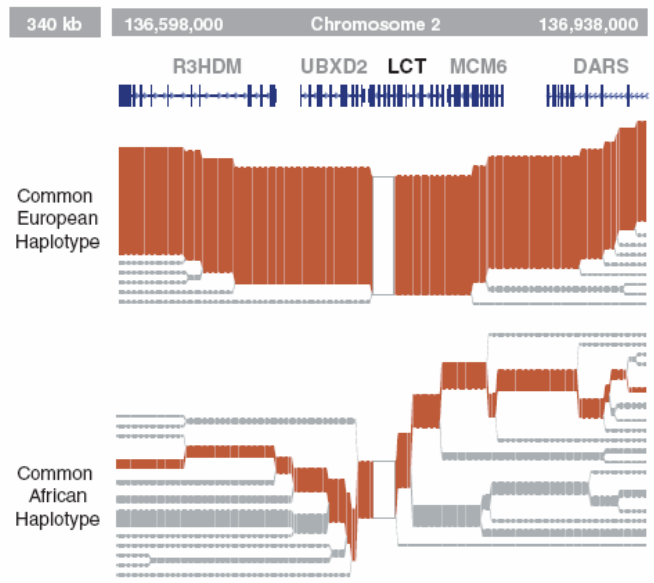
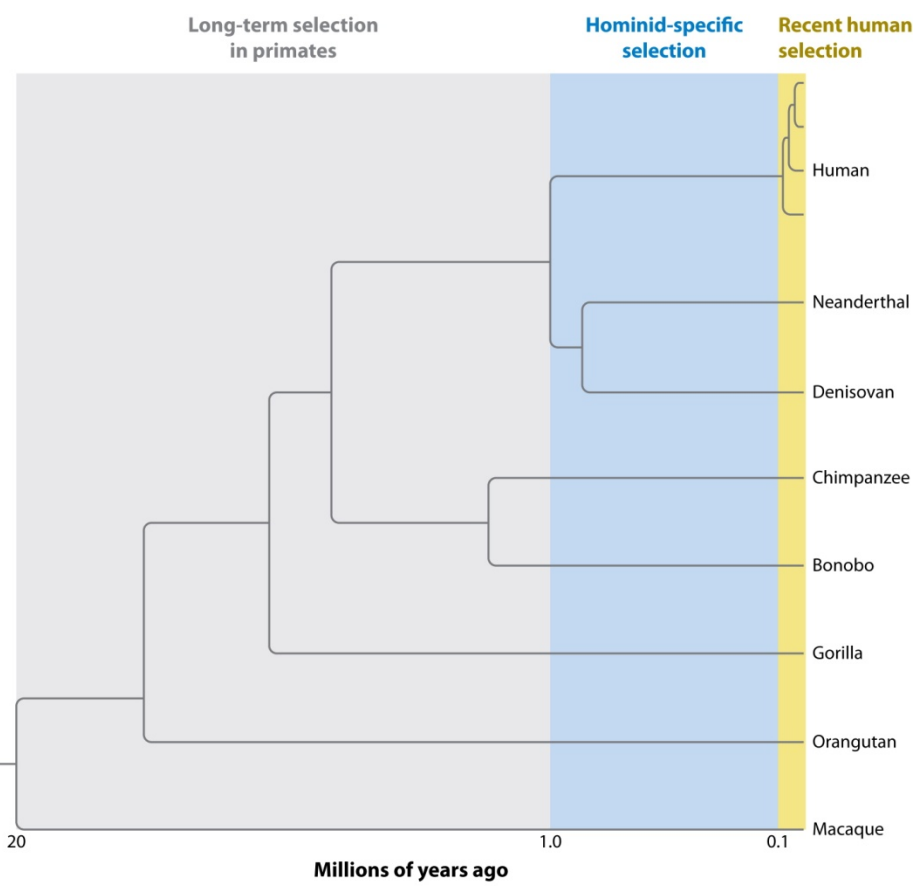


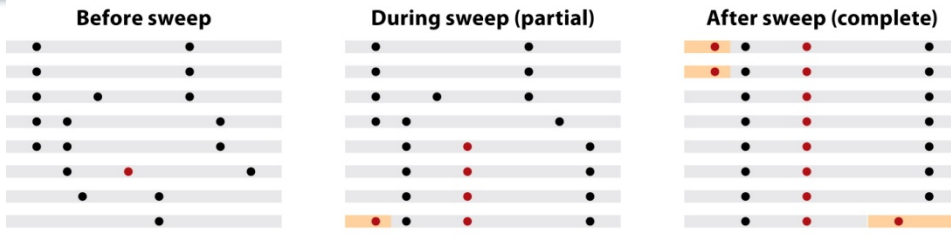
Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (6).



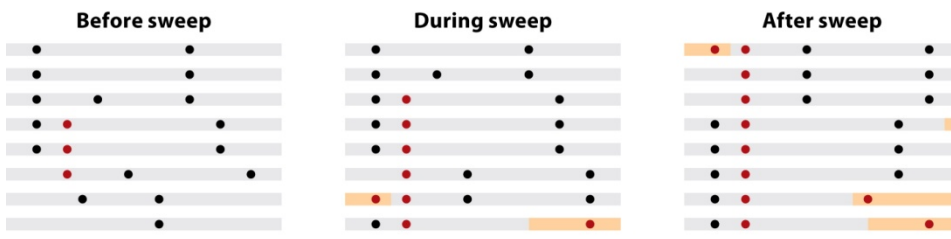
Positive Selection in Human Lineage



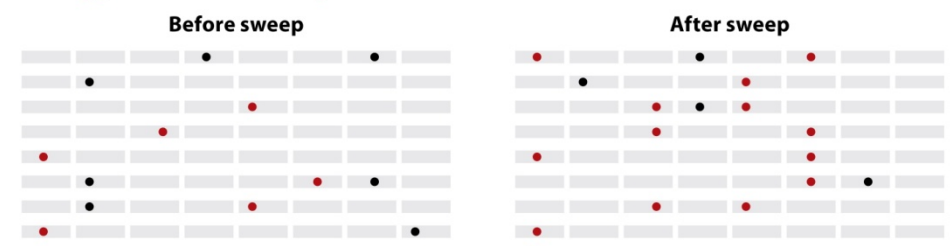
a Hard sweep



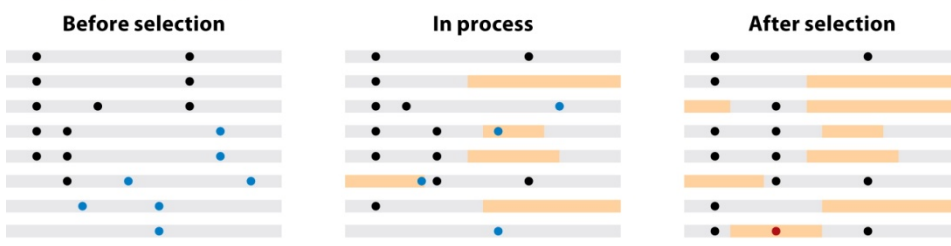
b Positive selection on standing variation





c Polygenic selection (adaptation)



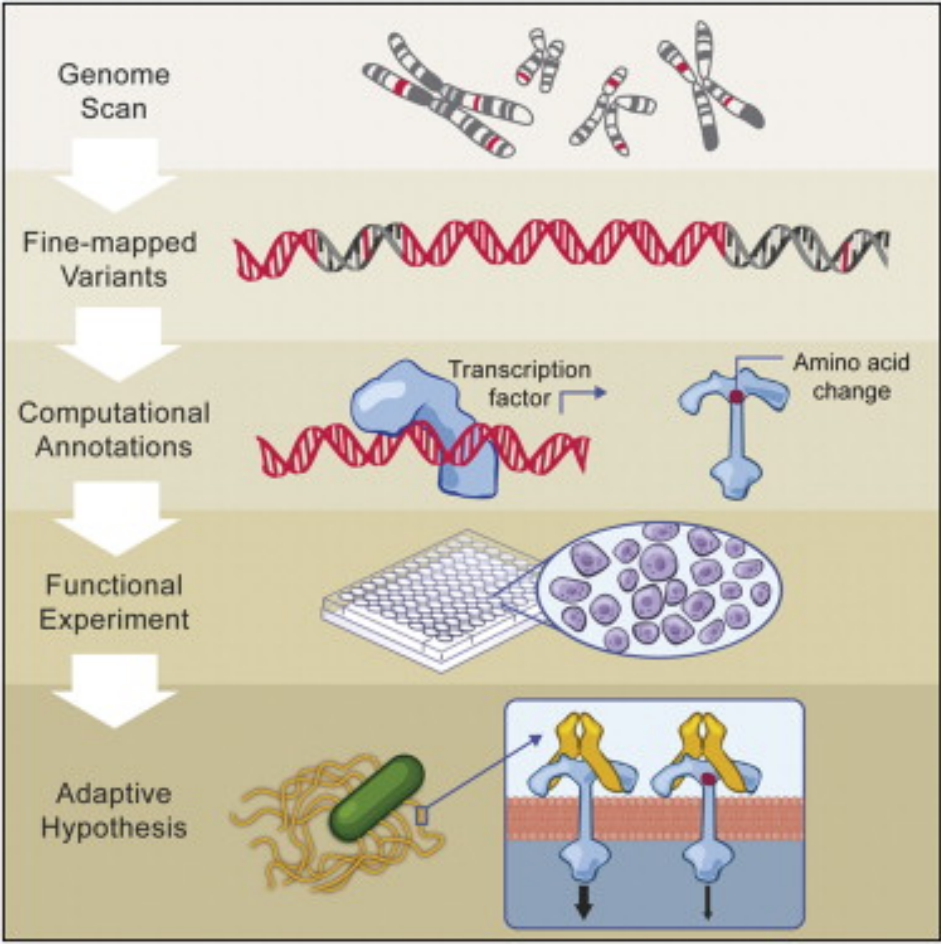
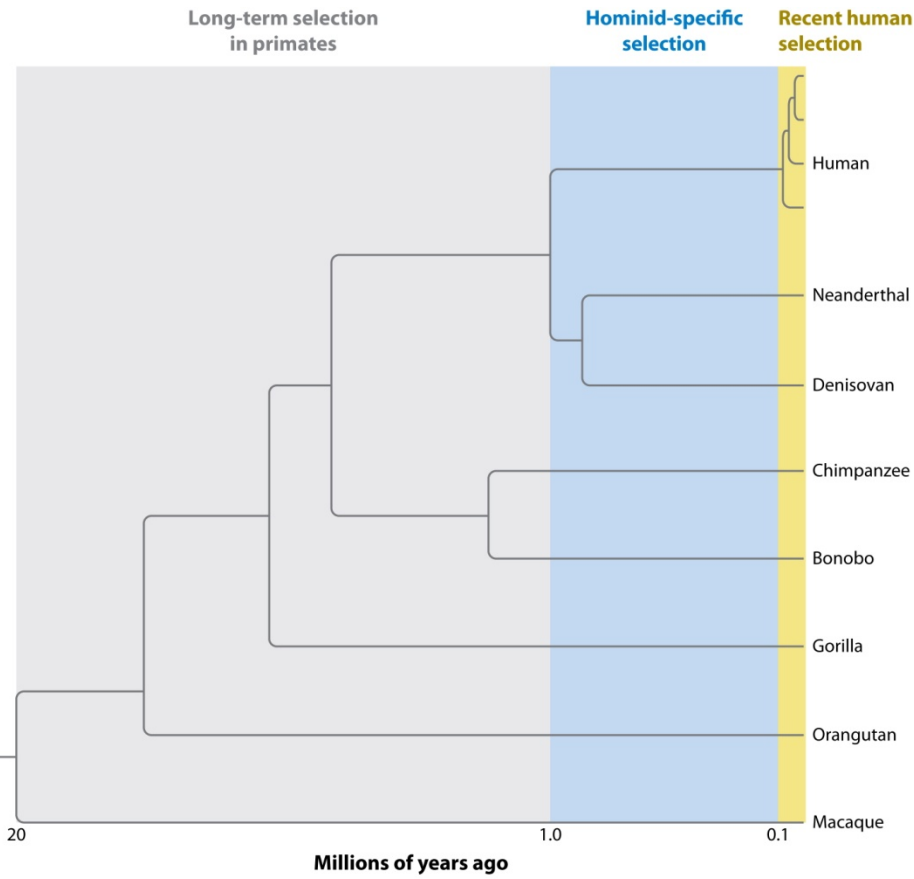
d Purifying selection




 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

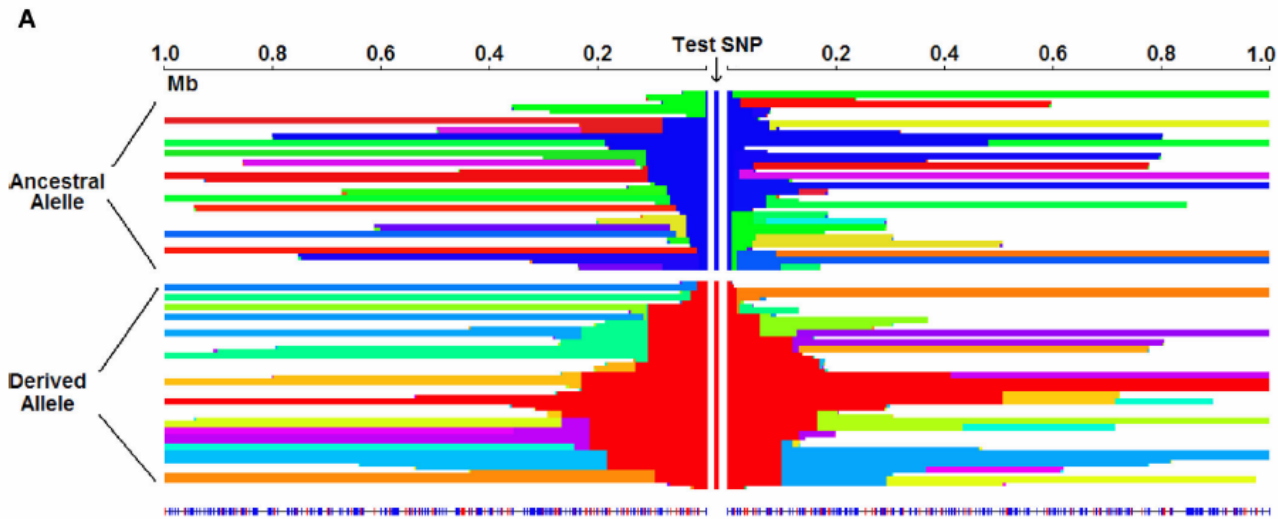
Positive Selection in Human Lineage



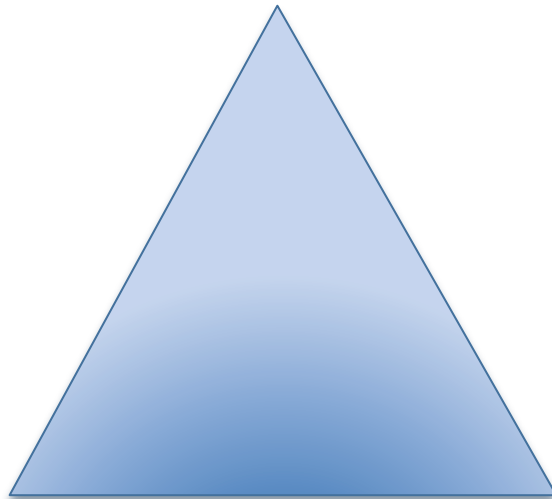
 Fu W, Akey JM. 2013. Annu. Rev. Genomics Hum. Genet. 14:467–89



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



Application: Malaria

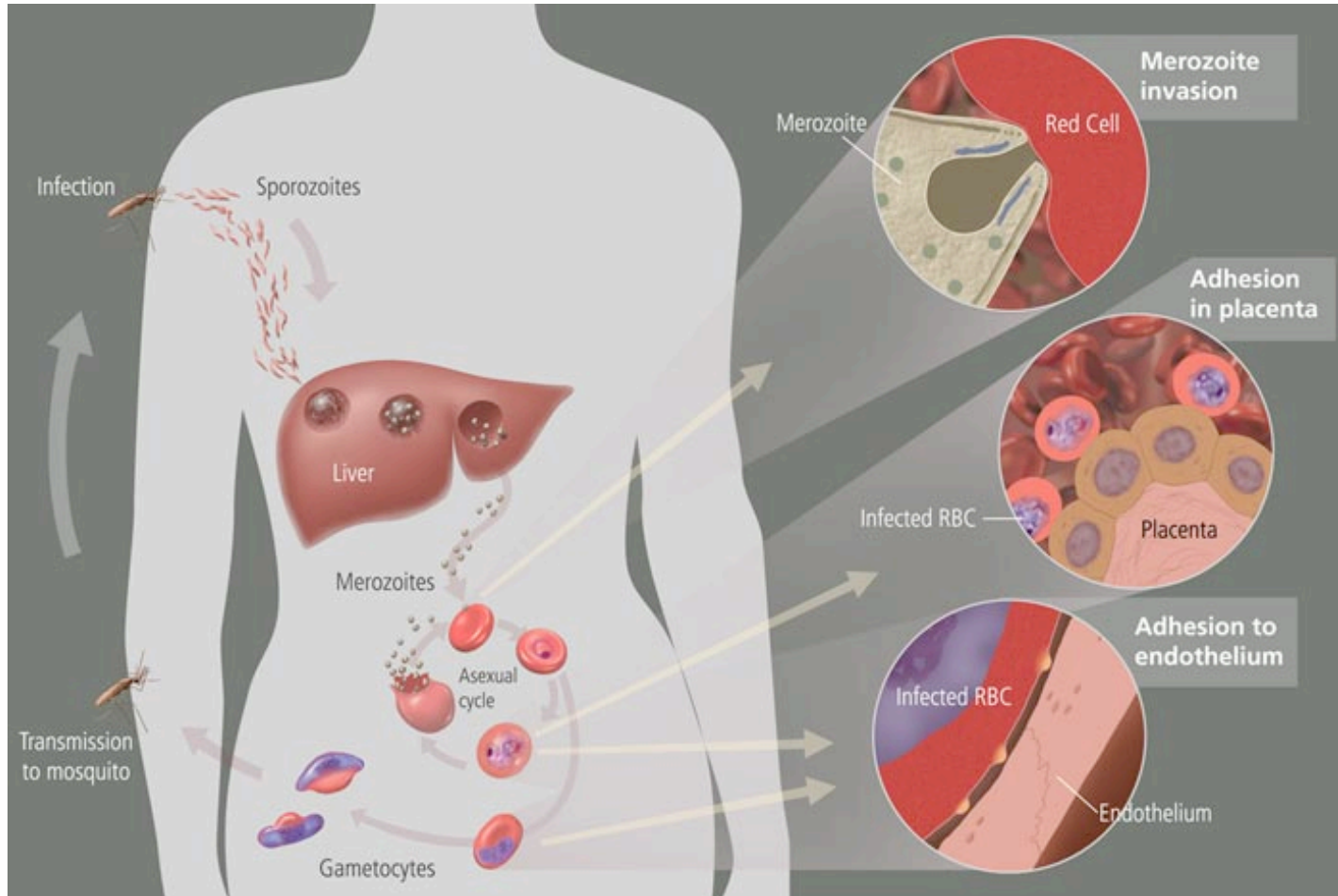


Image source:

NIH - <http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg>

Slide Credits:
Marc Schaub

Application: Malaria



Malaria Endemic Countries, 2003

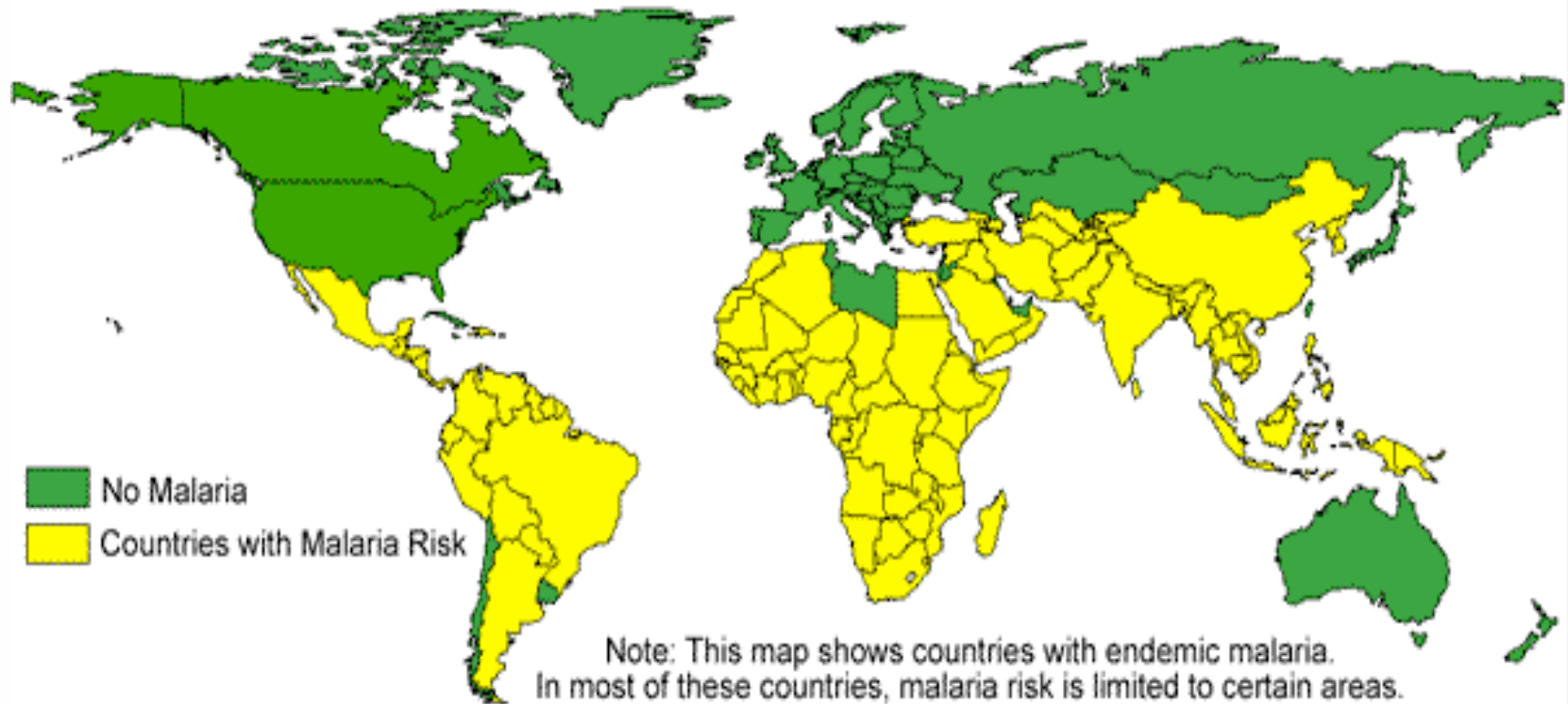
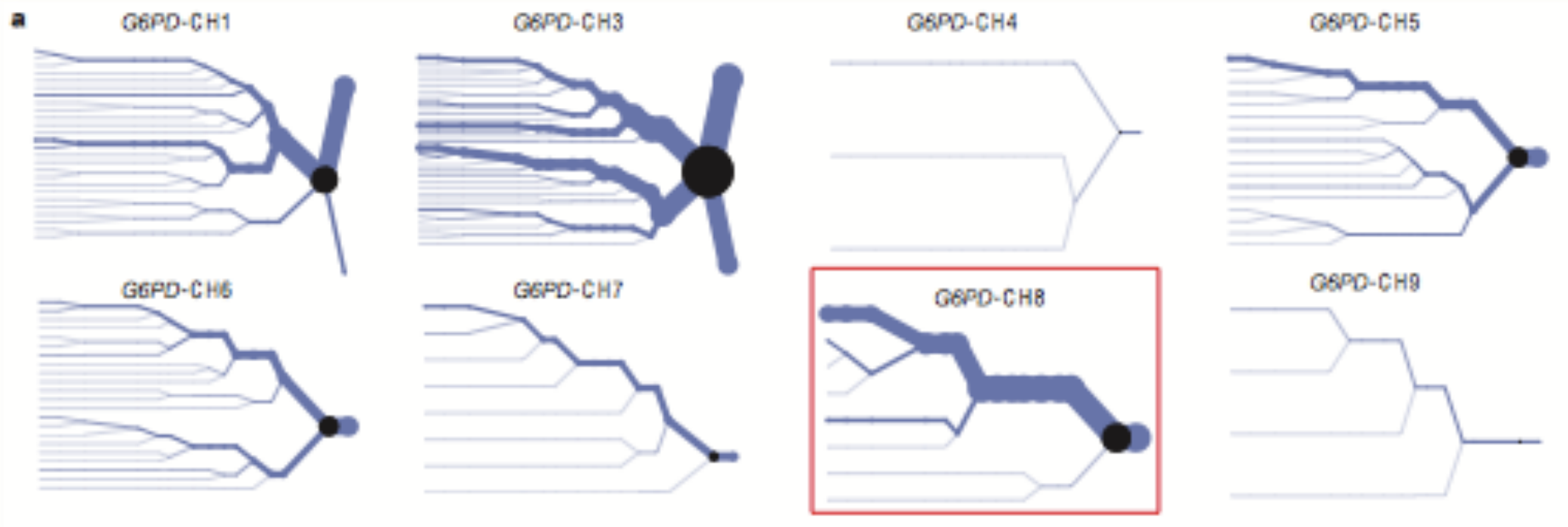


Image source: CDC -

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

Slide Credits:
Marc Schaub

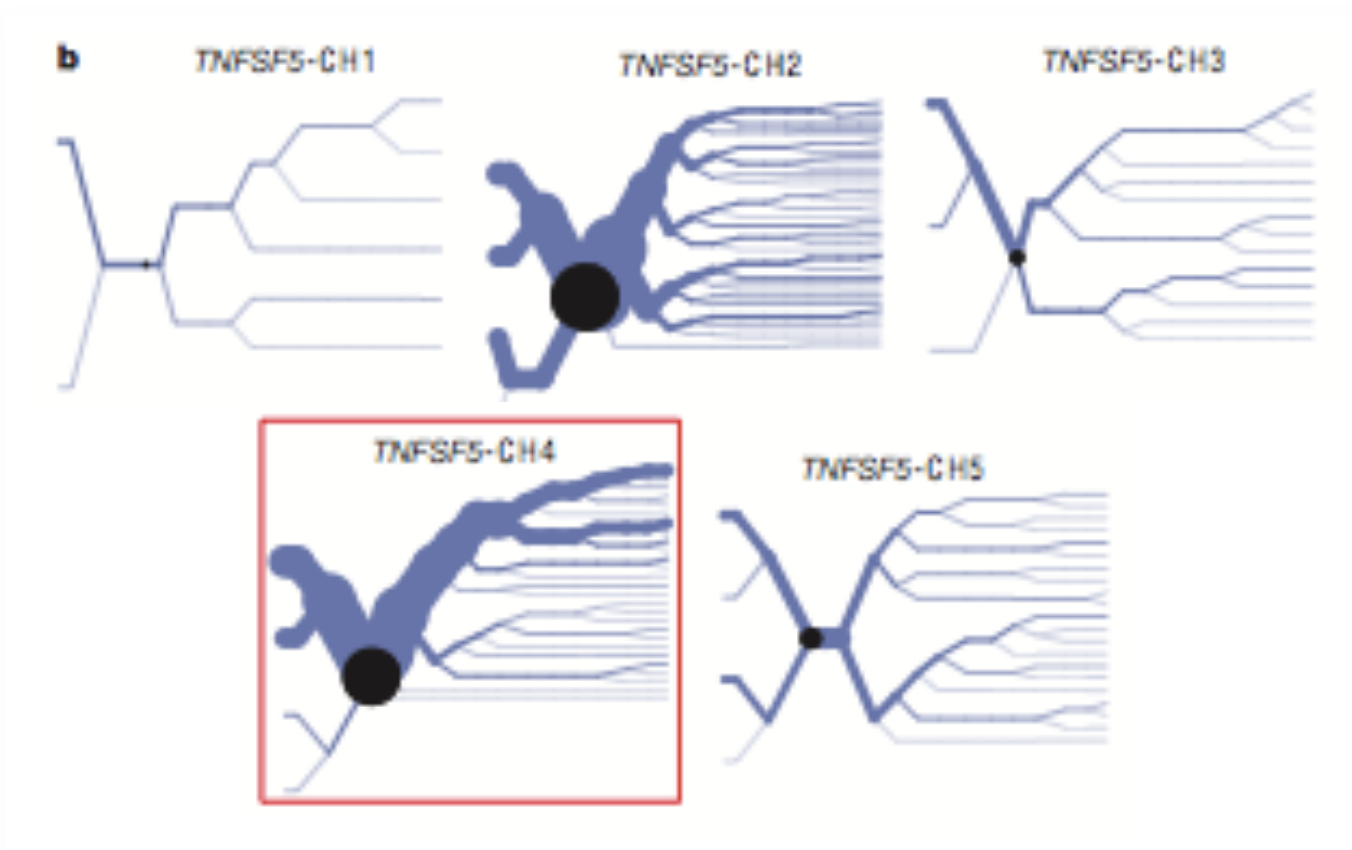
Results: G6PD



Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

Results: TNFSF5



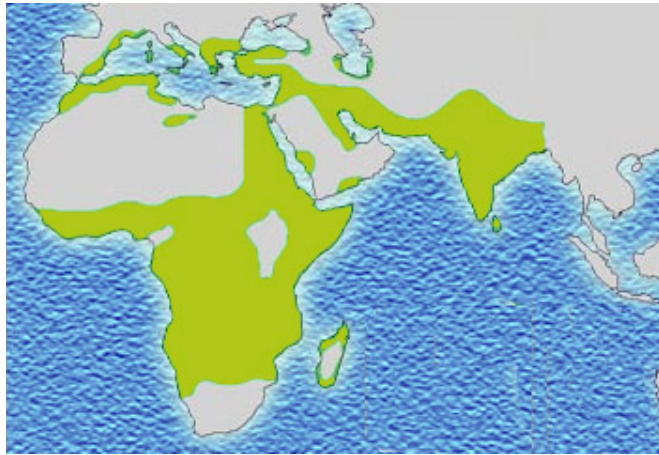
Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

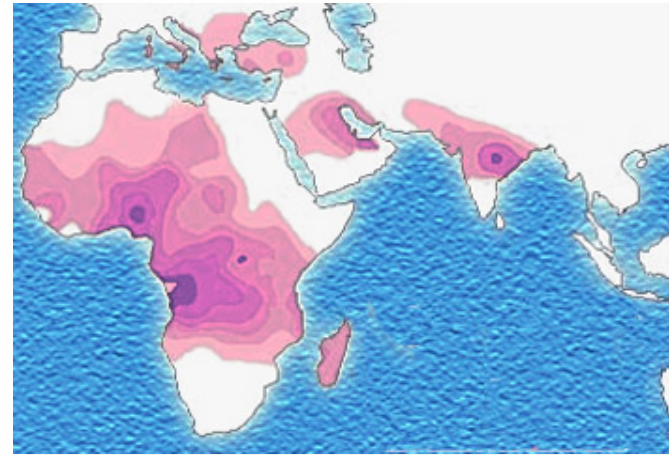
Malaria and Sickle-cell Anemia



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



Distribution of malaria

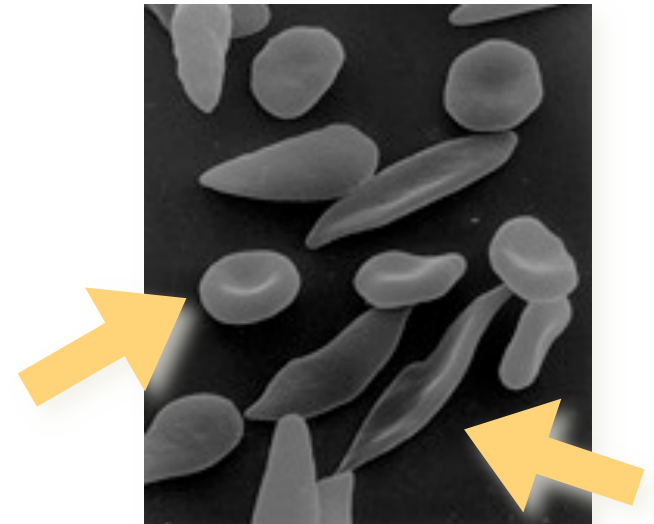


Distribution of sickle-cell anemia

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.



Slide Credits:
Marc Schaub
Image source: wikipedia.org

Lactose Intolerance

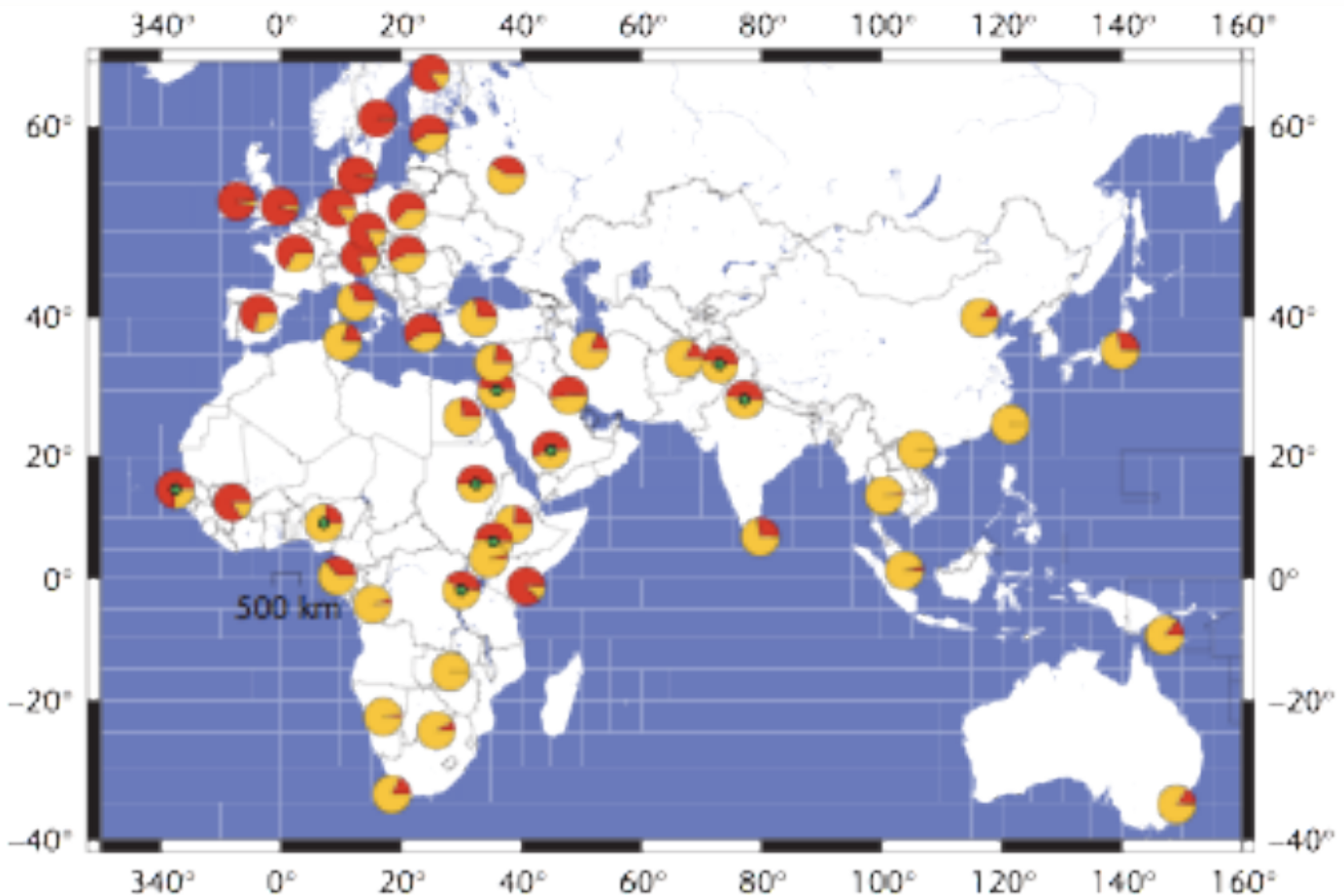
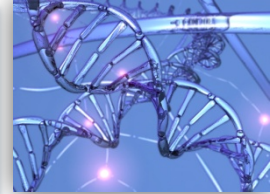
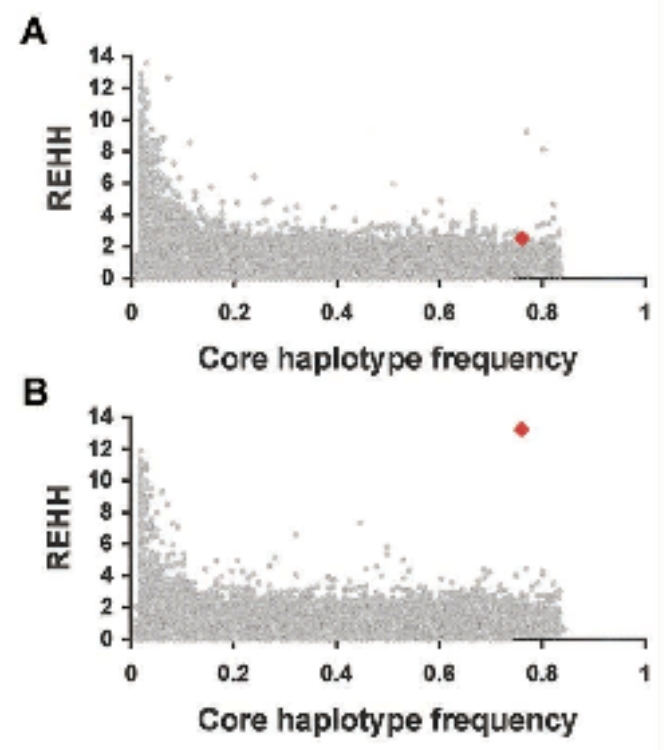
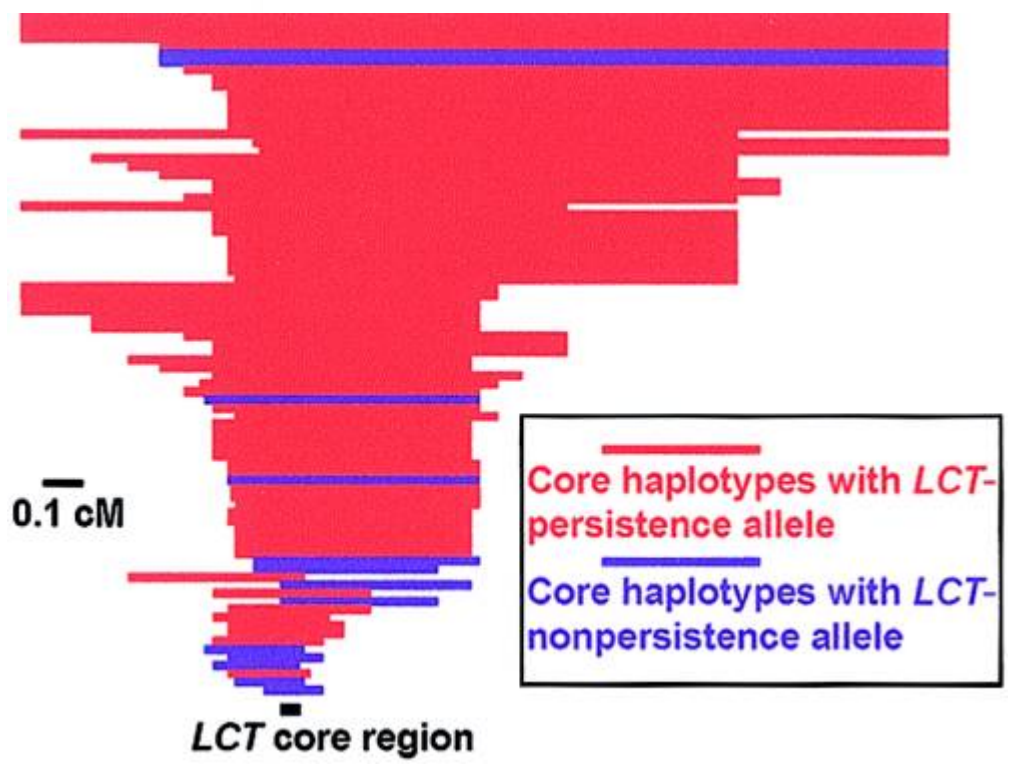
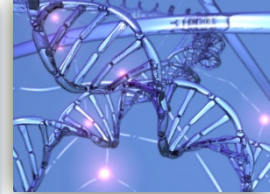


Figure 1 Old world distribution of frequency of lactase persistence (lactose digesters) taken from available published data. Red indicates the proportion of lactose digesters in a given population with yellow representing maldigesters. Charts with a green central circle indicate that the overall published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

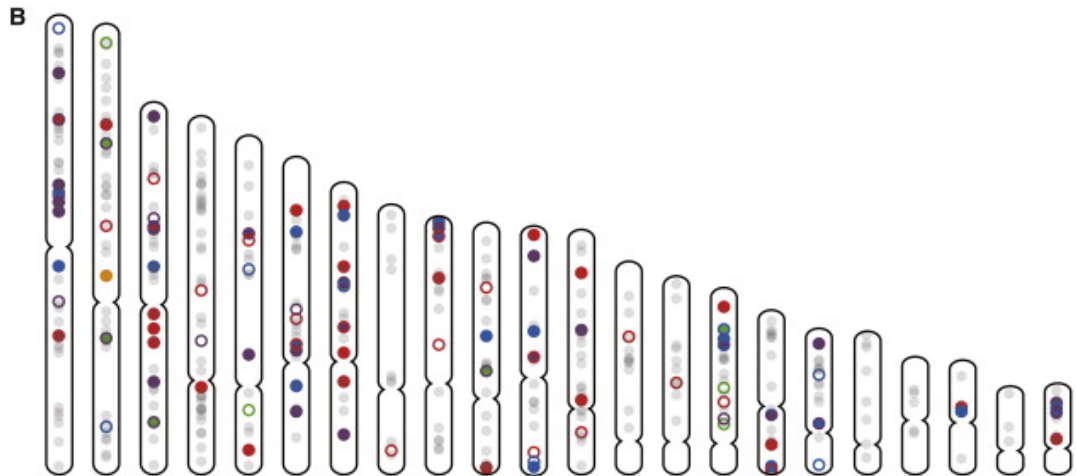
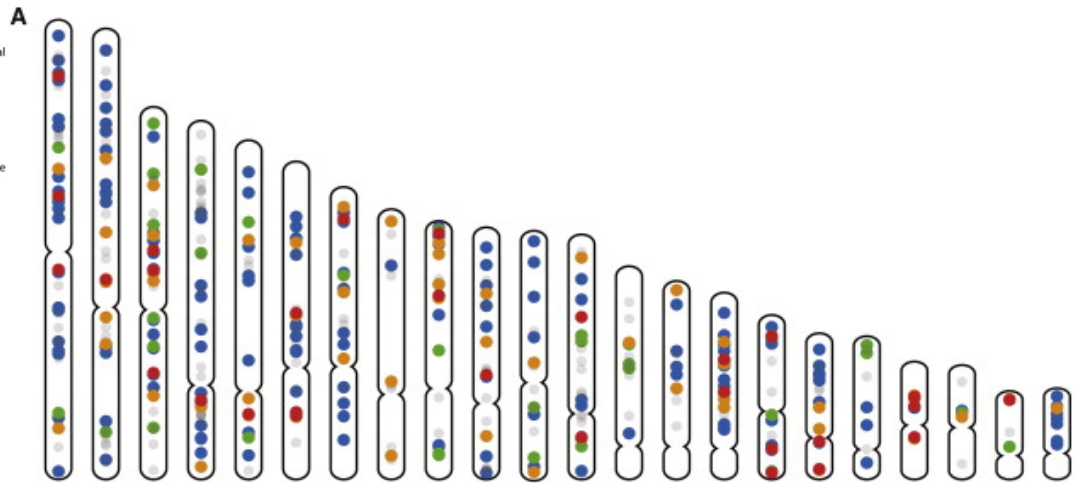
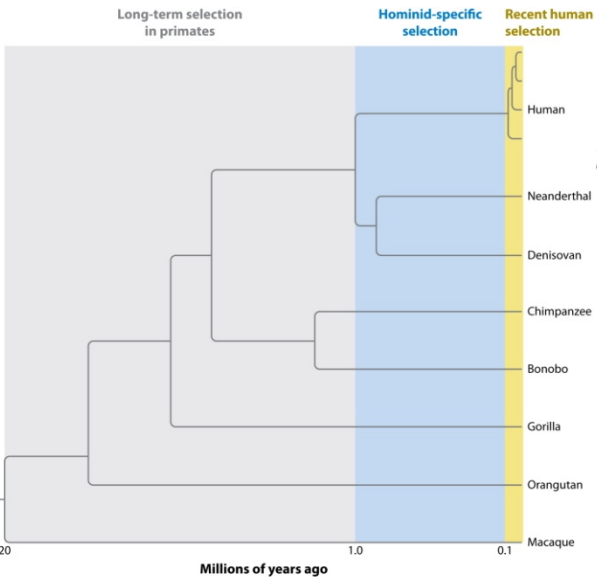
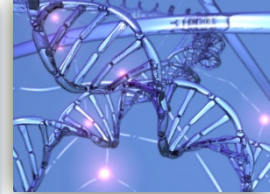
Source: Ingram and Swallow. Population Genetics of Encyclopedia of Life Sciences. 2007.

Slide Credits:
Marc Schaub

Lactose Intolerance



Positive Selection in Human Lineage



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