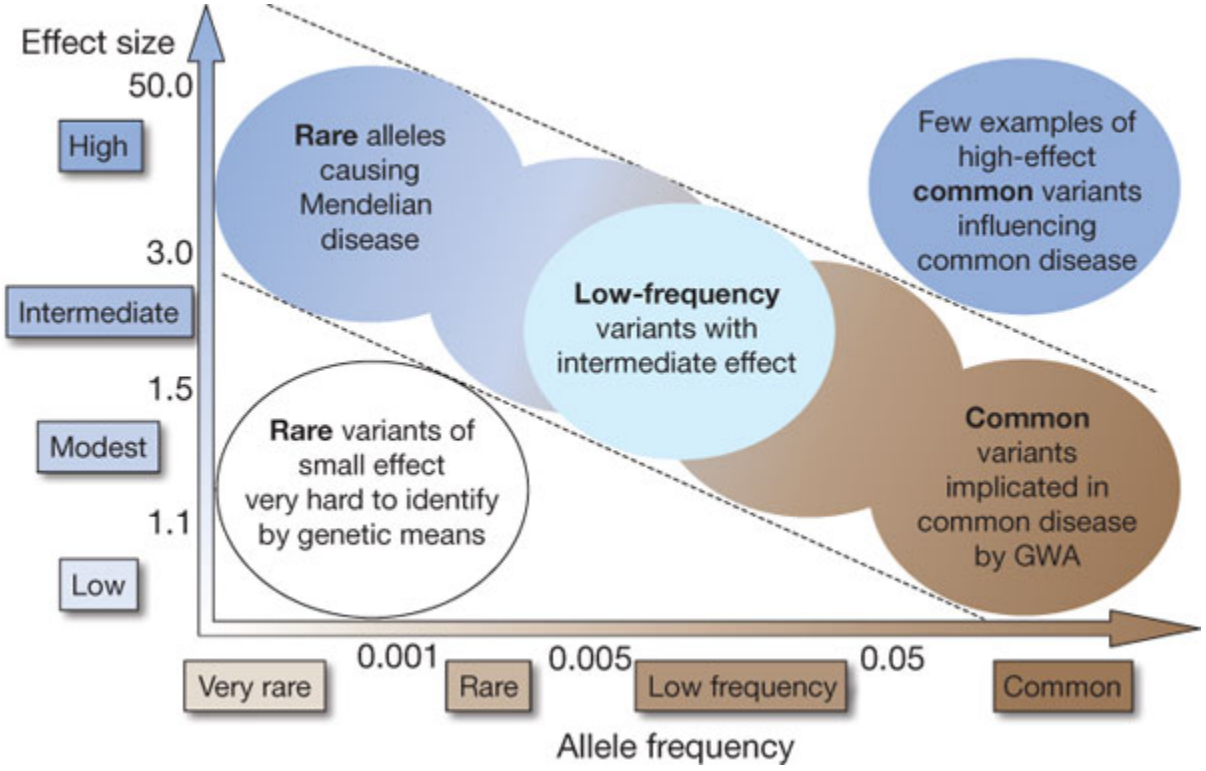
TGACTGAGGAGTTTACGGGAGCAAAGCGGCGTCATTGCTATTCGTATCTGTTTAG

01010111000100101000010101010101001101111001100011000101000100101

Human Population Genomics

Heritability & Environment

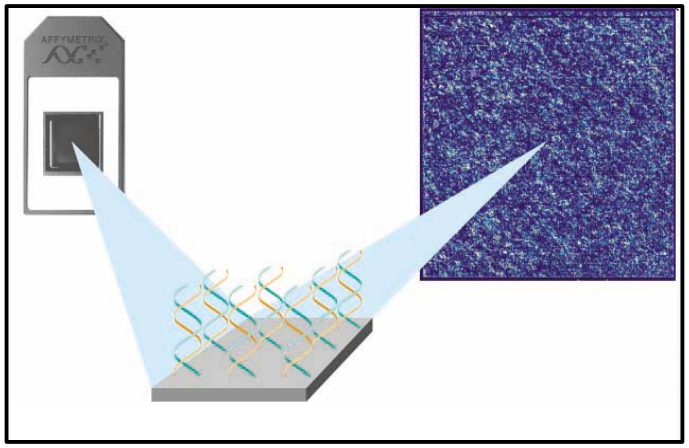
Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).

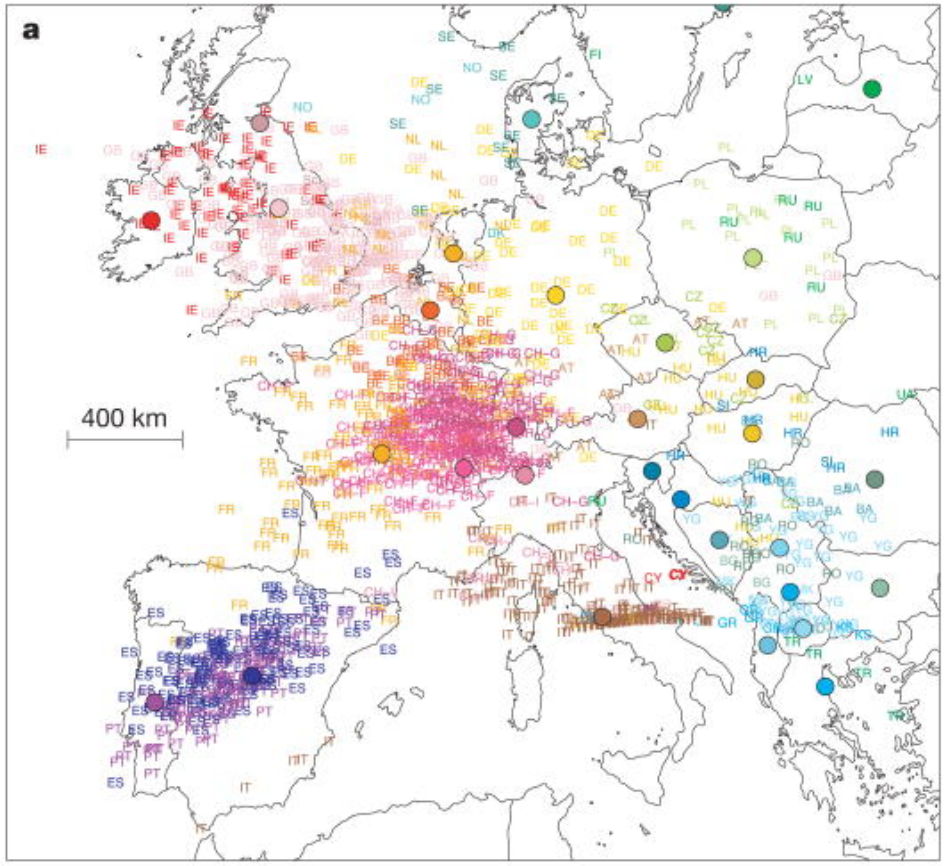


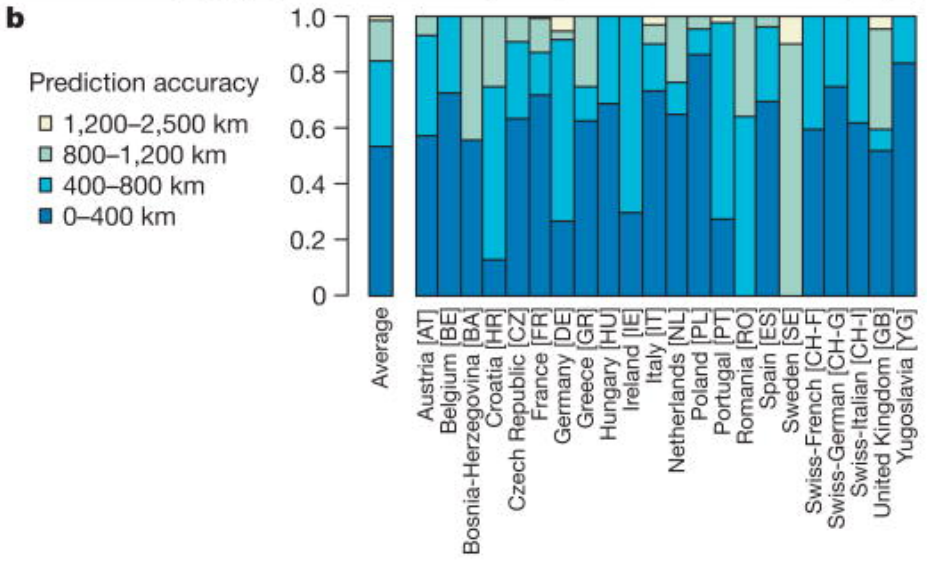
TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

nature

Global Ancestry Inferen







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

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

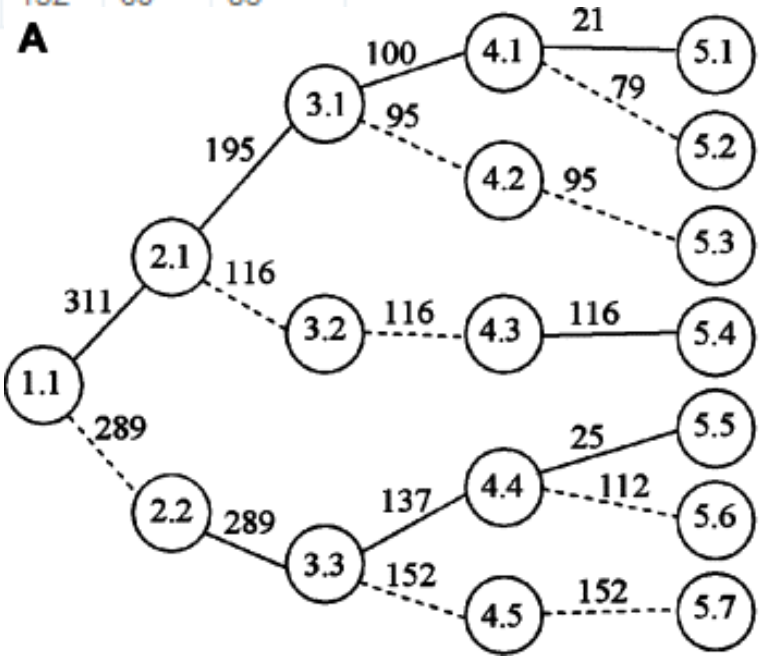
Modeling population haplotypes – VLMC

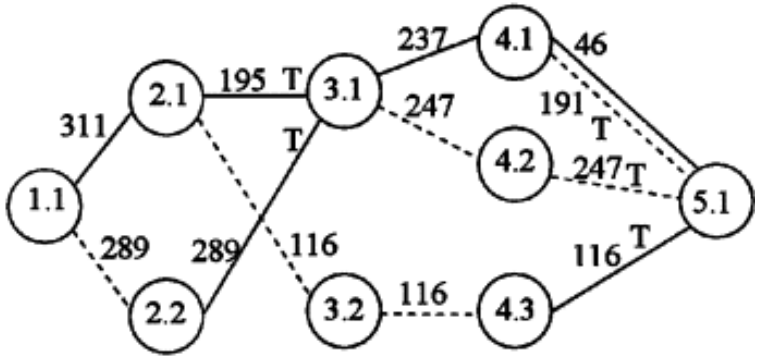
Haplotype	Count		
	Total	Case	Control
0000	21	12	9
0001	79	43	36
0011	95	43	52
0110	116	59	57
1000	25	14	11
1001	112	60	52
1011	152	69	83

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

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

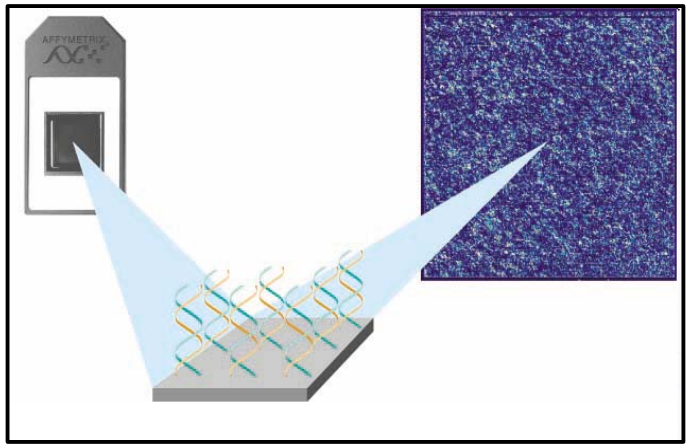
$H_i = h_{ij1} \ldots h_{ijn}$; $h_{ijk} \in \{0, 1\}$





Browning, 2006

Phasing



Haplotype Phasing

Haplotypes

ATCCGA

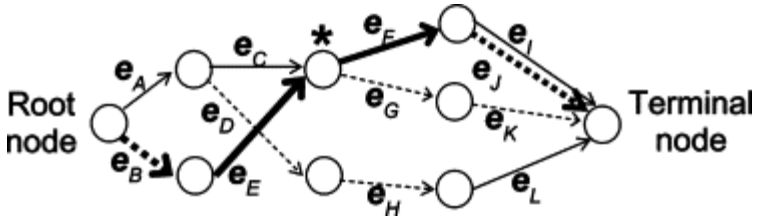
AGACGC

Genotype

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

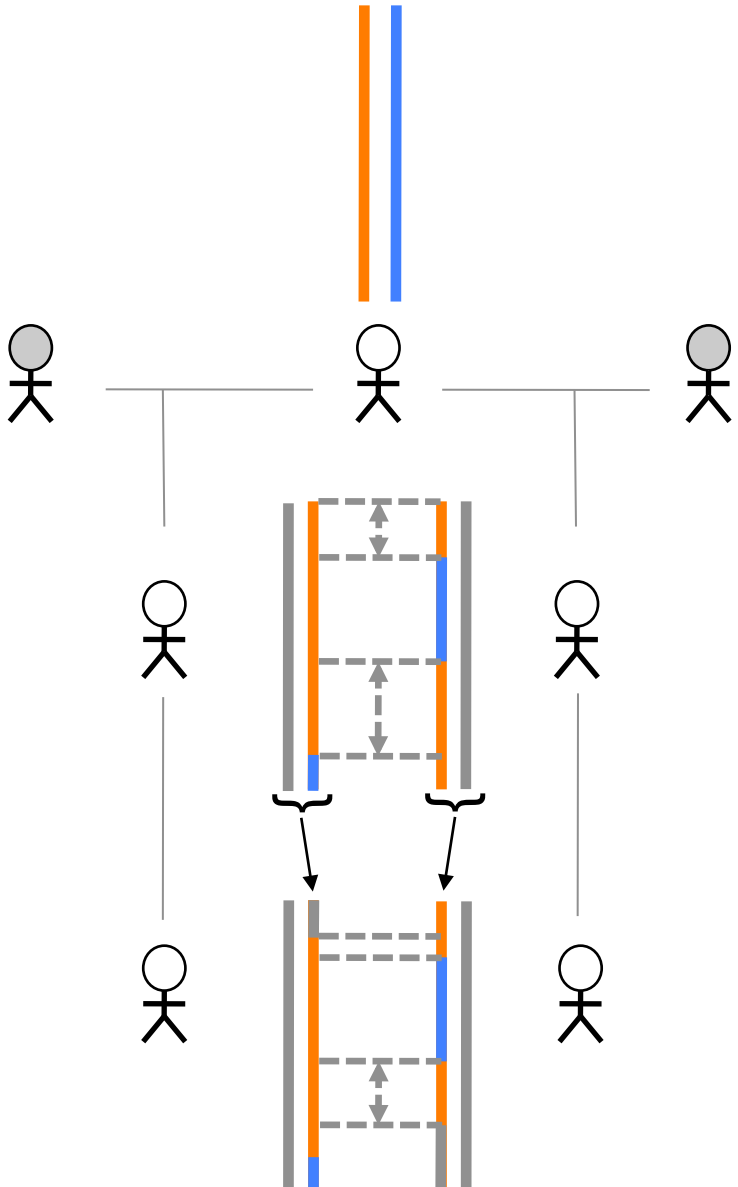
- High throughput cost effective sequencing technology gives genotypes and not haplotypes.

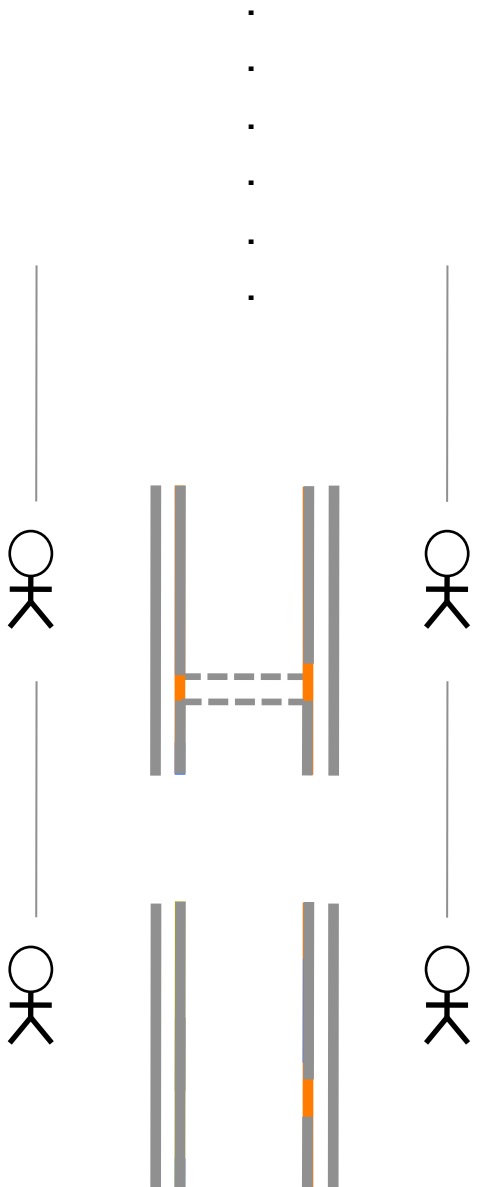
Possible phases: ATACGA AGACGA AGCCGC ATCCGC ….



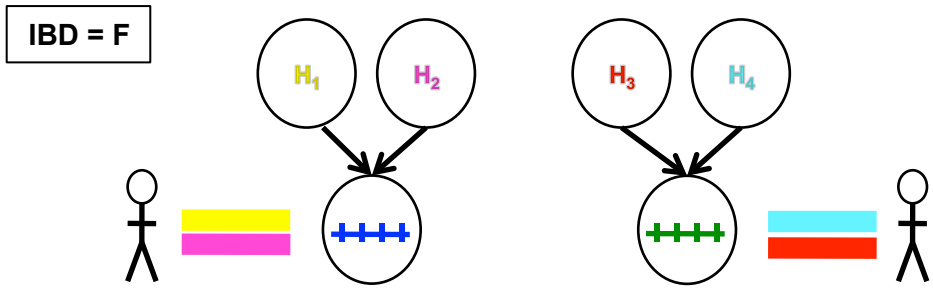
[Browning & Browning, 2007]

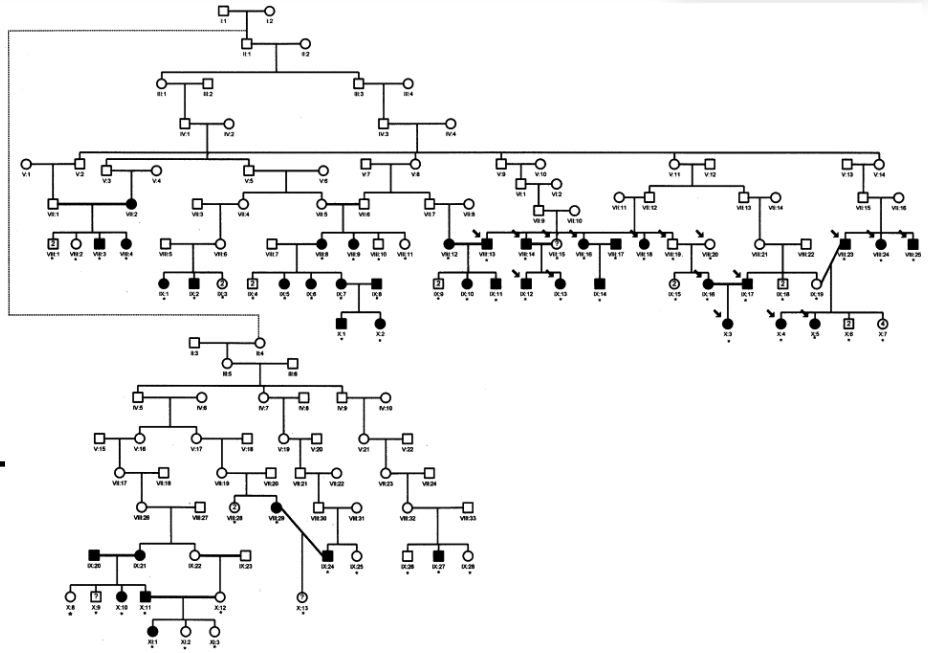
Identity By Descent

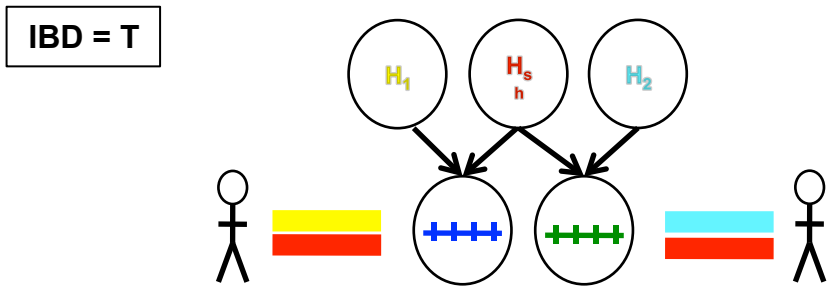


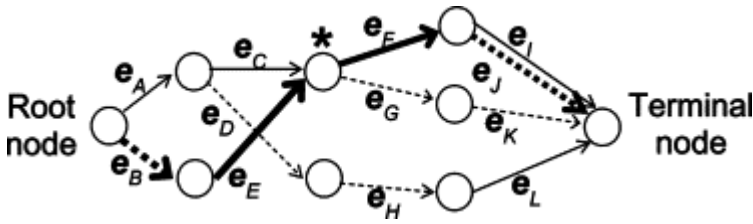


IBD detection









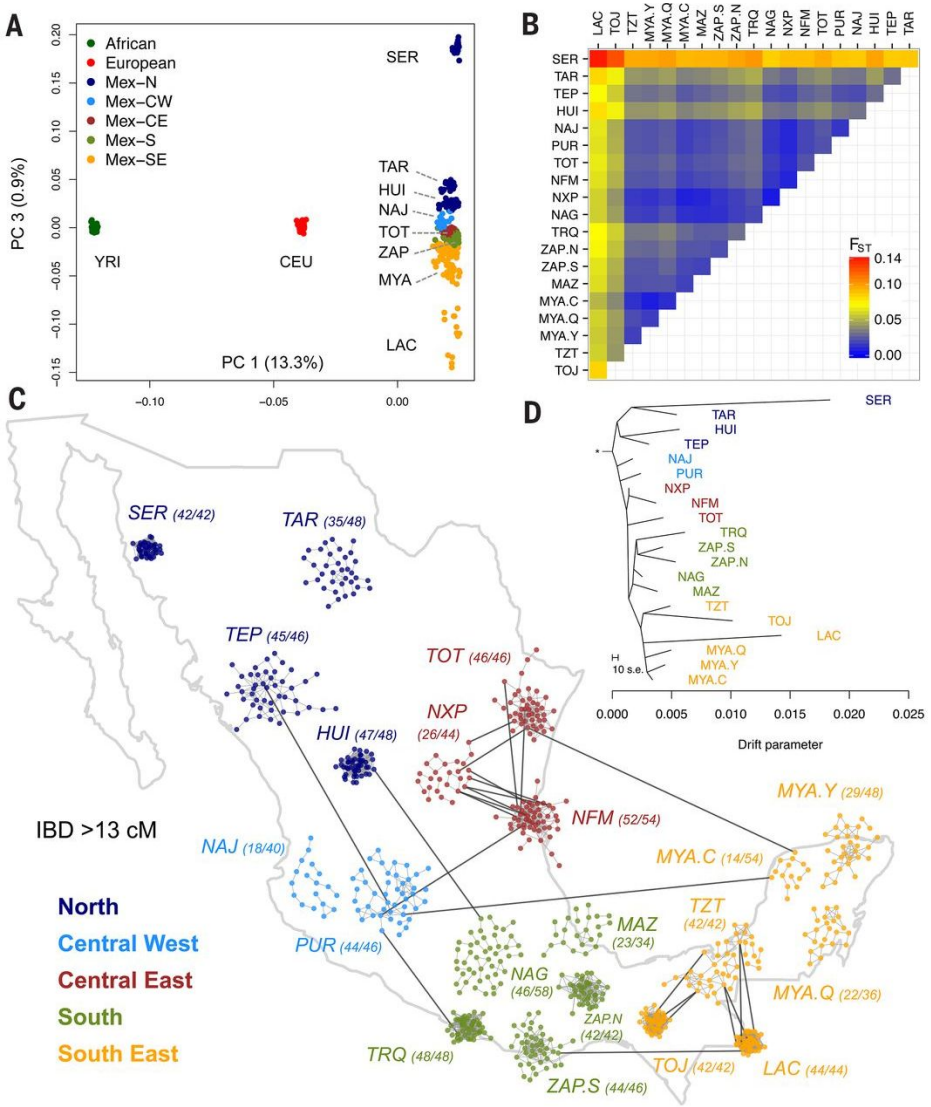
Parente

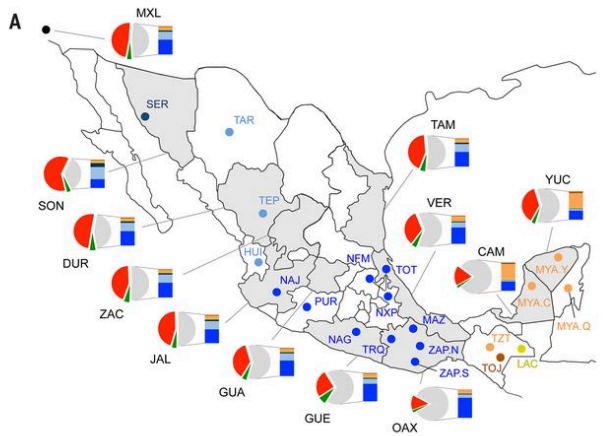
[Rodriguez et al. 2013]

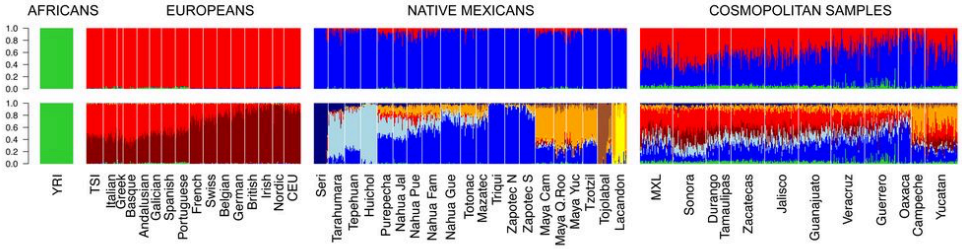
FastIBD: sample haplotypes for each individual, check for IBD

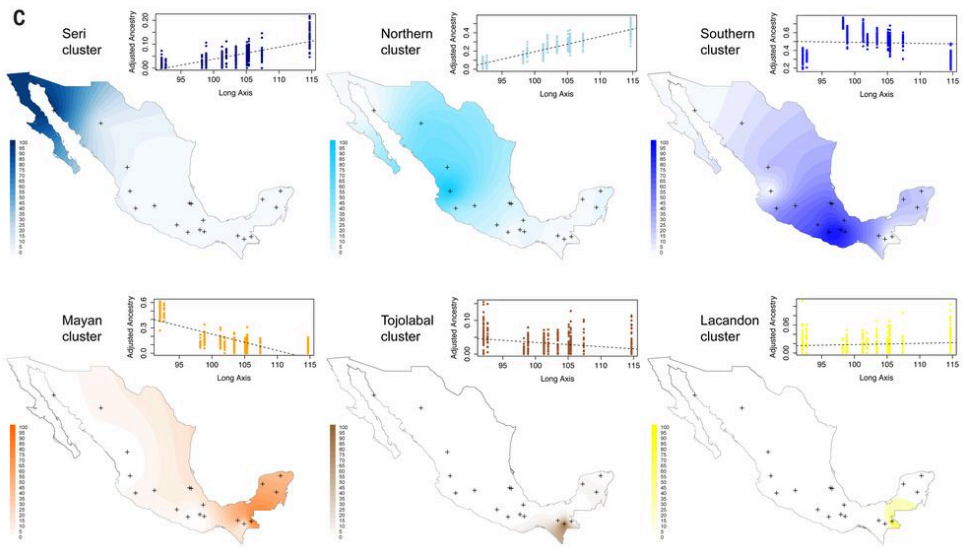
[Browning & Browining 2011]

Mexican Ancestry



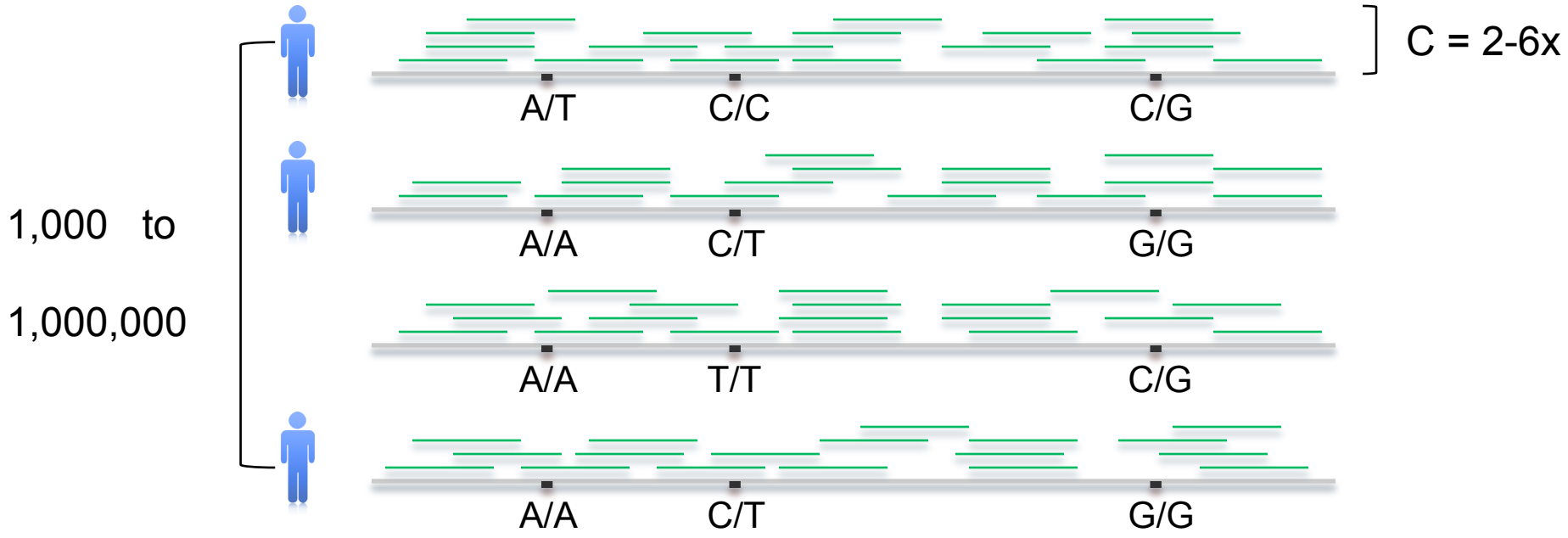




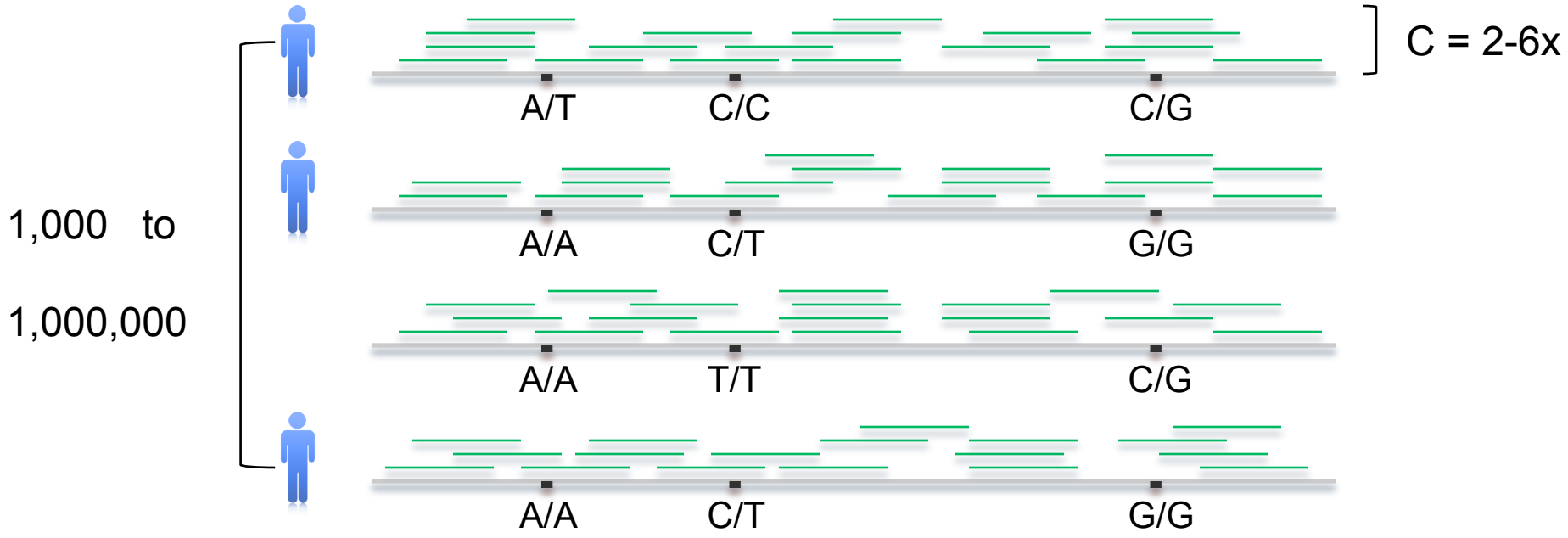


The genetics of Mexico recapitulates Native American substructure and affects biomedical traits, Moreno-Estrada et al. Science, 2014.

Population Sequencing



Population Sequencing



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

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

Population Sequencing

When C is high (>30x),

Prob(g_{ij} = k | data) ~

Prob(g_{ij} = k | reads mapping on (i, j))

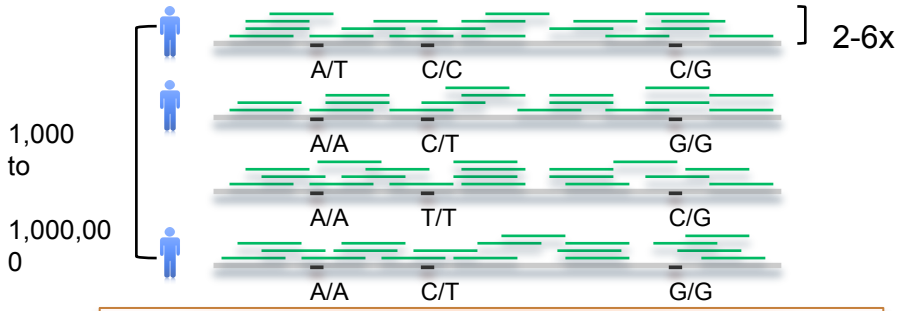
fast & easy

When C is low,

Prob(g_{ij} = k | data) needs to leverage LD:

positions j' ≠ j in all individuals

in principle, intractable



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

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

Population Sequencing

Summarization - Maximization:

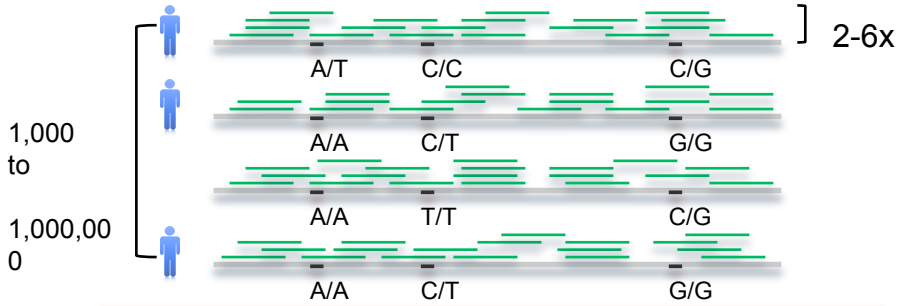
1. Identify candidate polymorphic sites
2. Initialize $G^{(0)}$
3. Summarization:

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

4. Maximization:

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

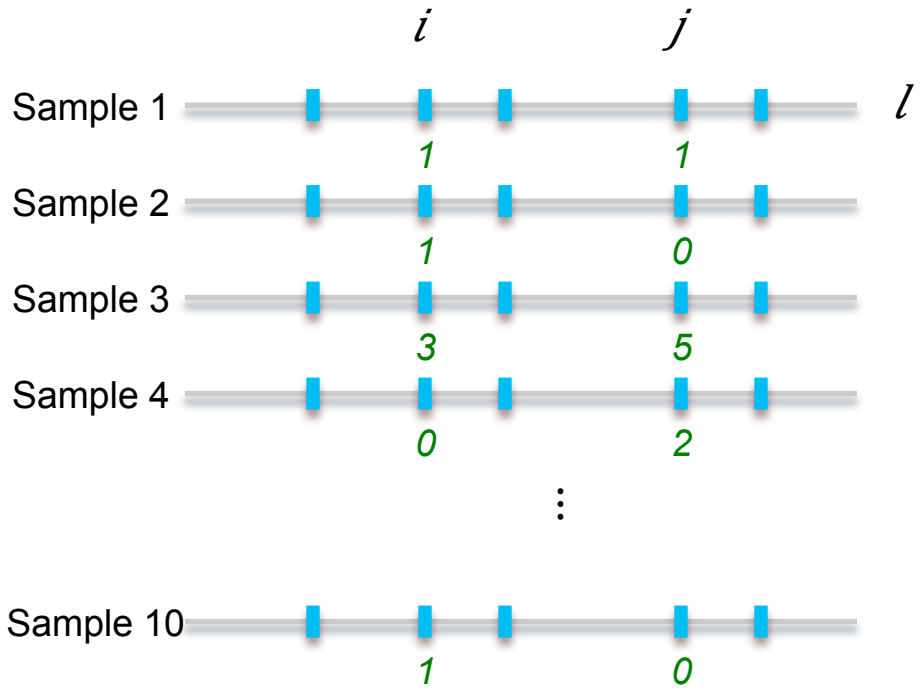
5. Repeat until convergence



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

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

Modeling LD: Nearest Neighbors



Let

S_i = { samples with >= one read covering minor allele }

S_i = {1, 2, 3, 10}
S_j = {1, 3, 4}

Then,

$Sim_1(i, j) = (S_i \quad S_j) / (S_i \cup S_j) = 2/4$

Reveel Algorithm

Summarization Maximization:

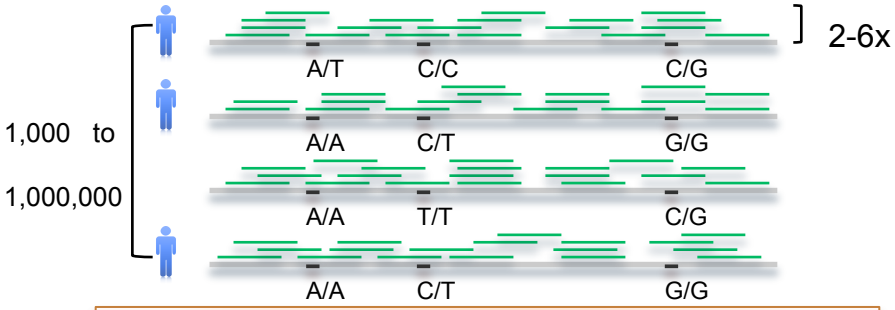
1. **Identify candidate polymorphic sites**
2. Initialize $G^{(0)}$
3. Summarization:

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

4. Maximization:

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

5. Repeat until convergence



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

$P_1, \ldots, P_N$; 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:

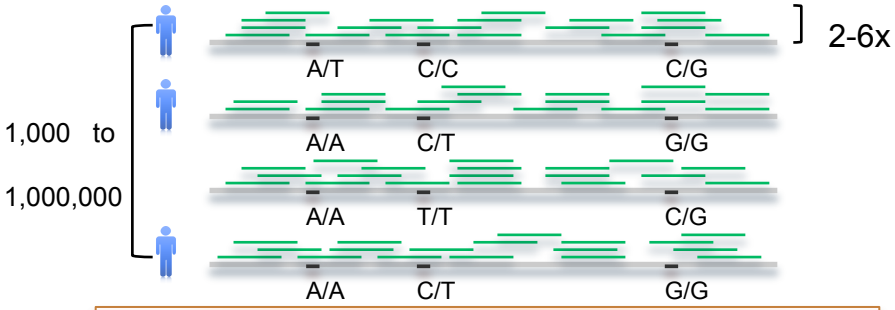
1. Identify candidate polymorphic sites
2. **Initialize $G^{(0)}$**
3. Summarization:

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

4. Maximization:

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

5. Repeat until convergence



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

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

At each position j,

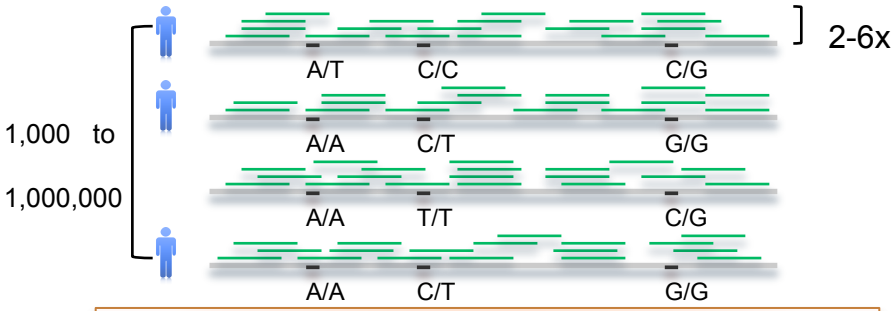
Use sum of read counts at j and its nearest neighbors

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

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

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

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



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

$P_1, \ldots, P_N$; 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:

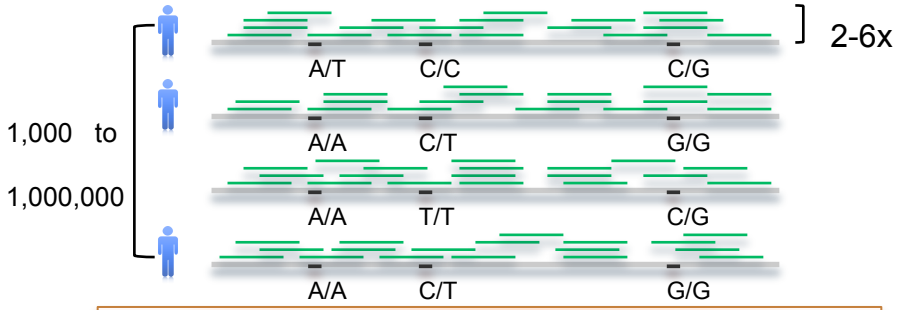
1. Identify candidate polymorphic sites
2. Initialize $G^{(0)}$
3. Summarization:

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

4. **Maximization:**

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

5. Repeat until convergence



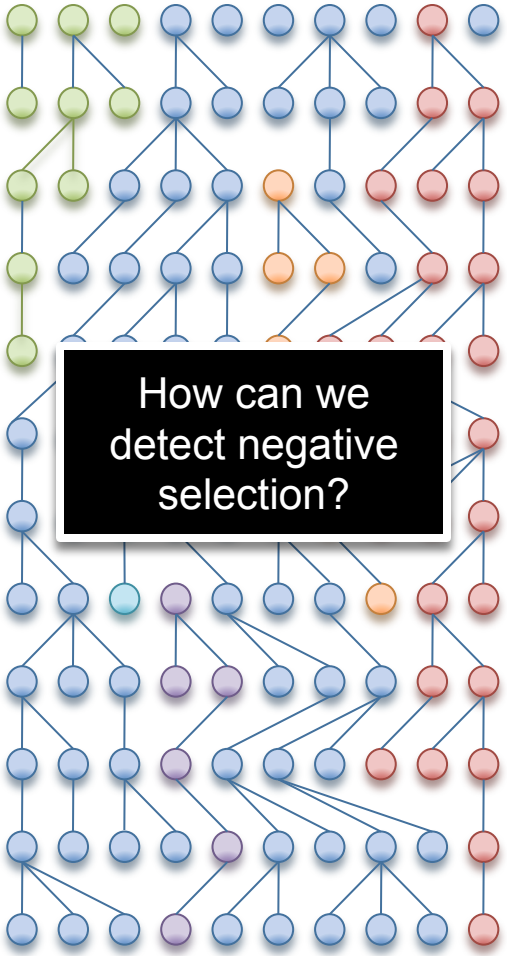
$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

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

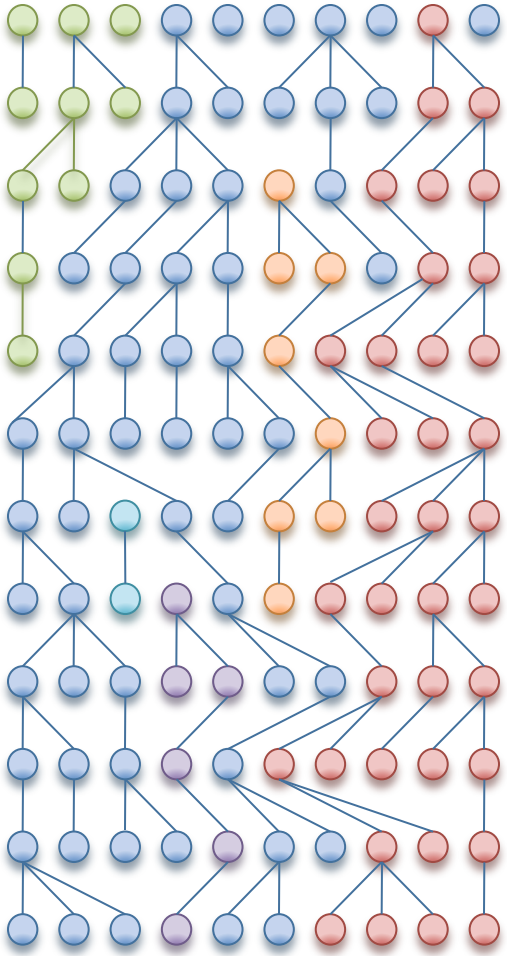
At each position j,

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?

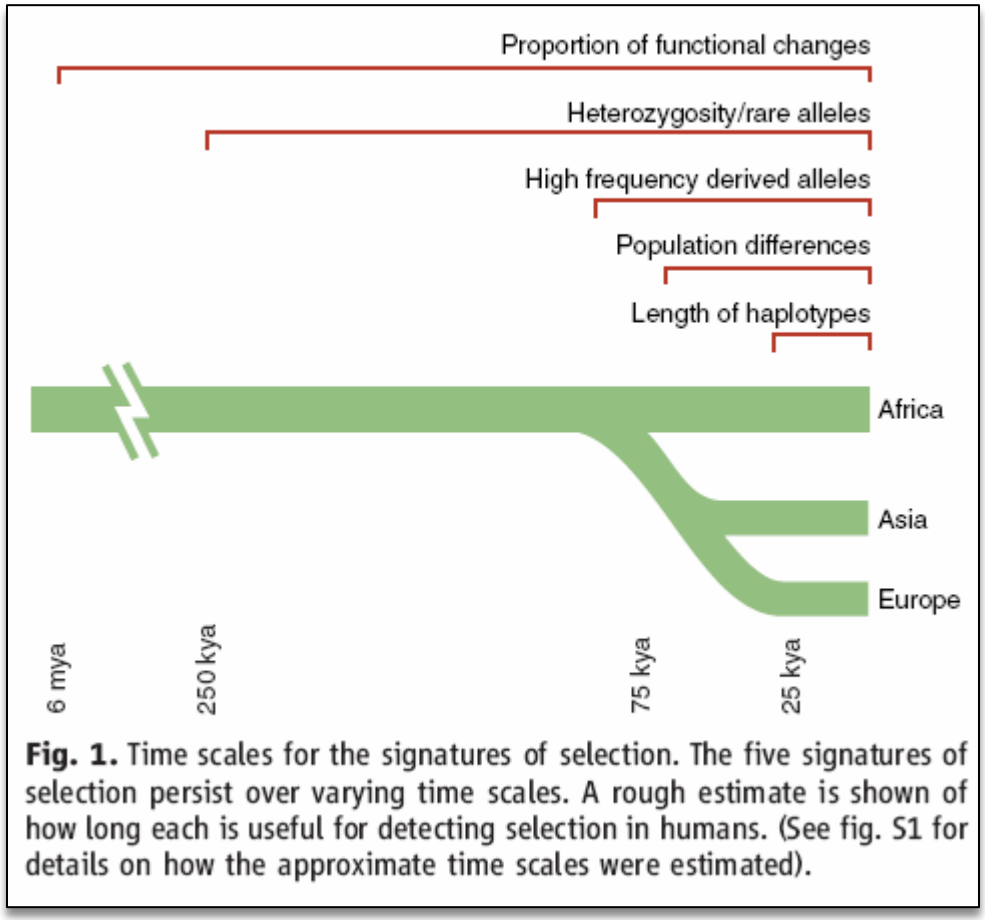


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

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

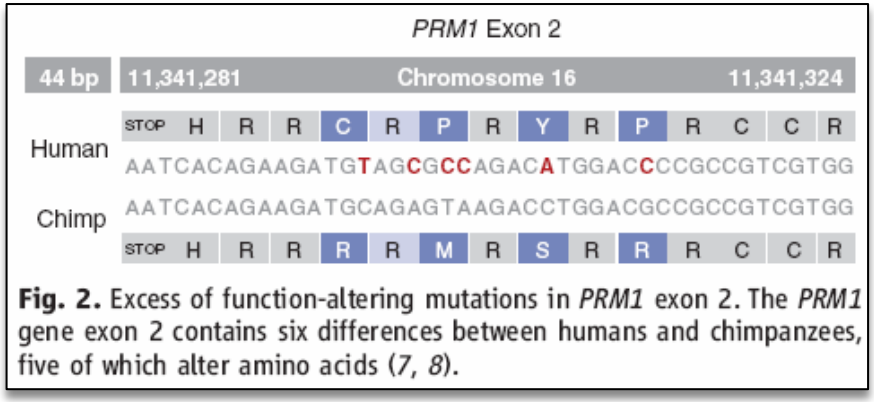


Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

How can we detect positive selection?

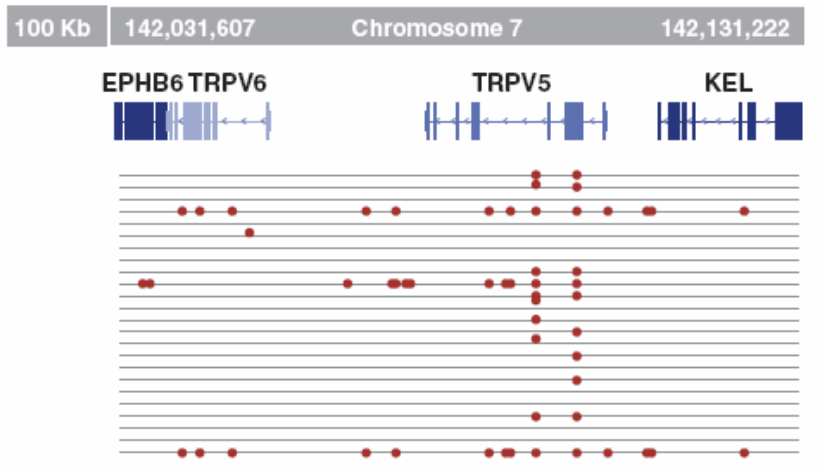


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

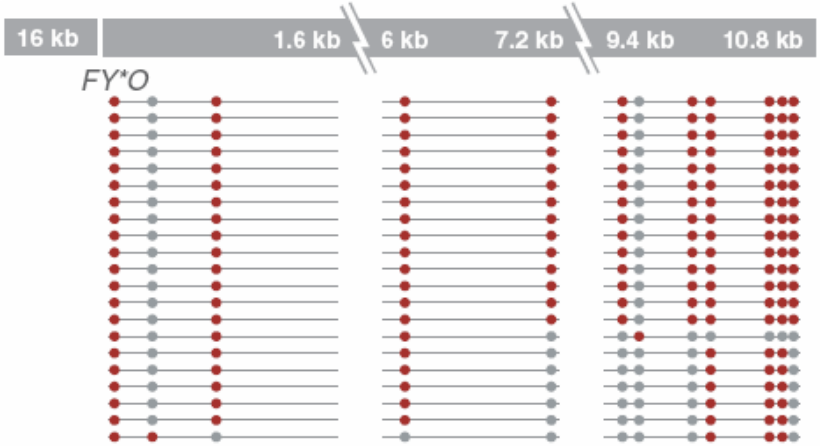


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (*34*). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

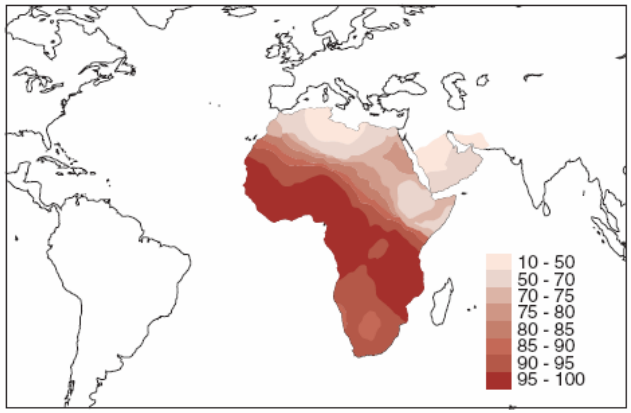


Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (*38*).

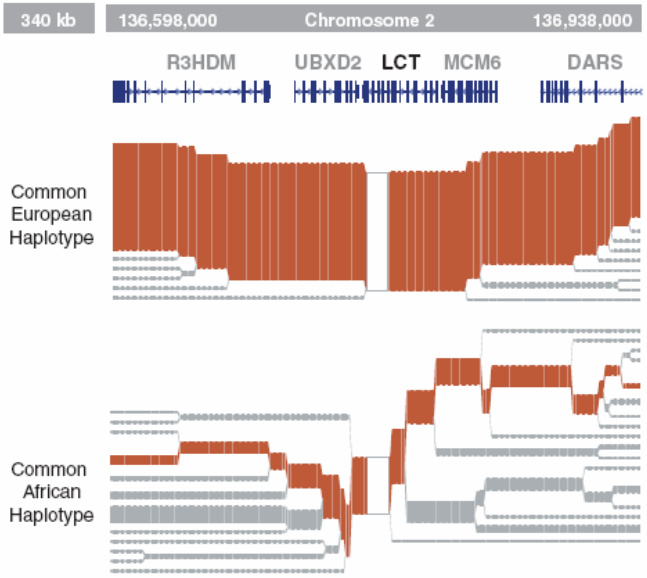
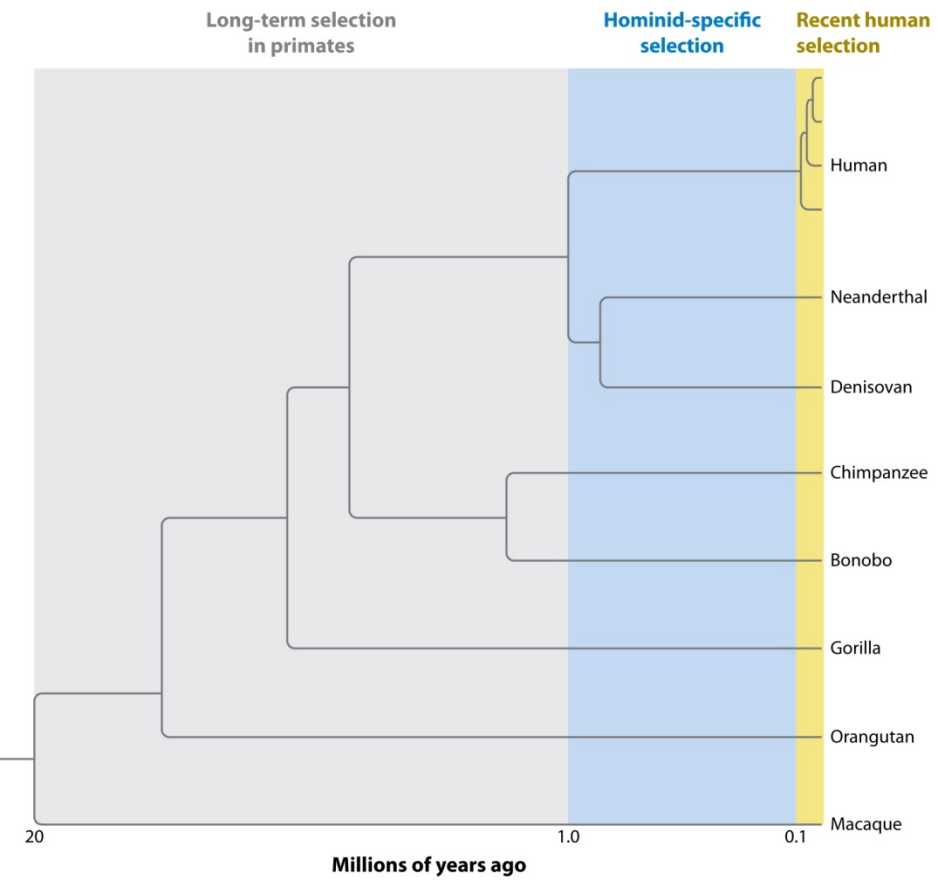
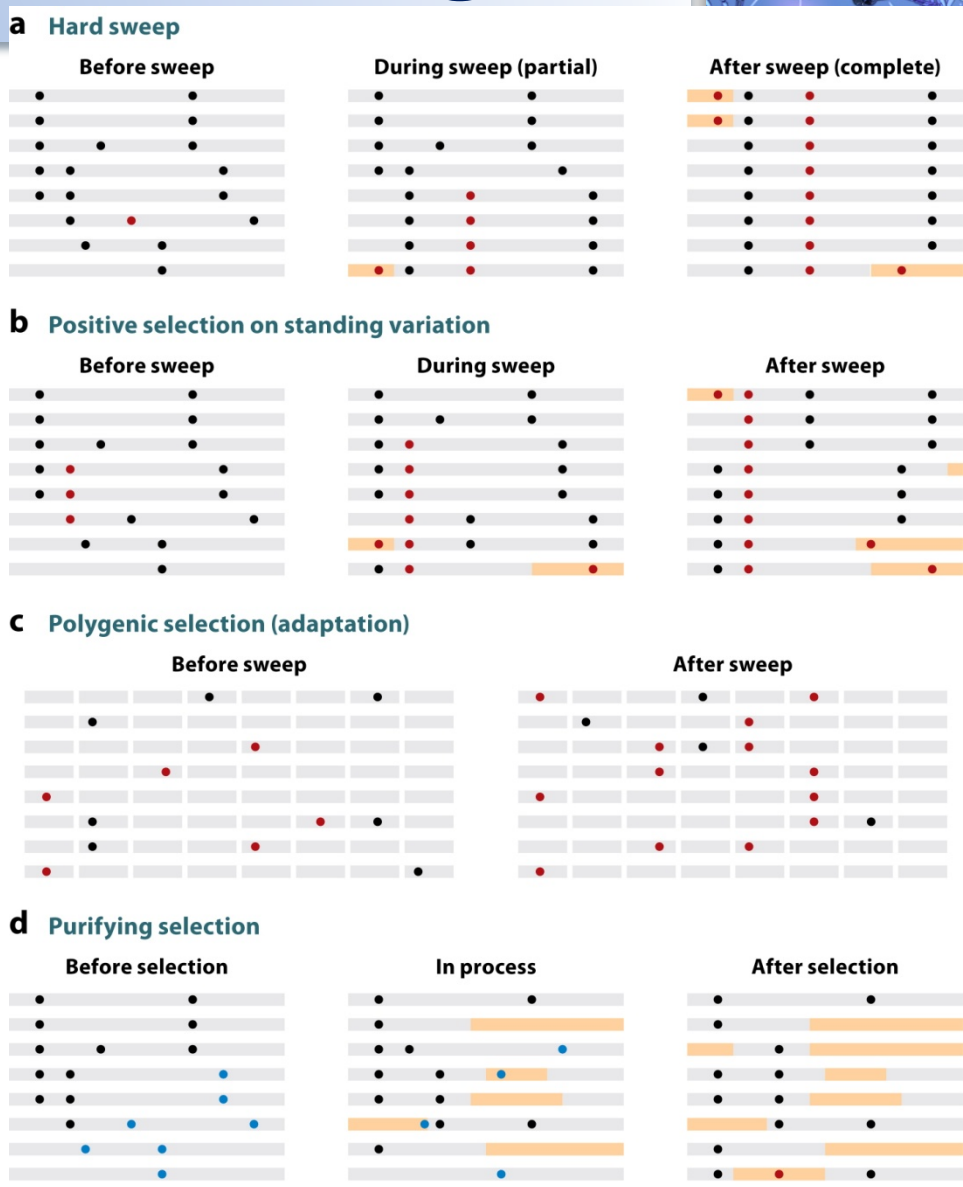


Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (*6*).

Positive Selection in Human Lineage

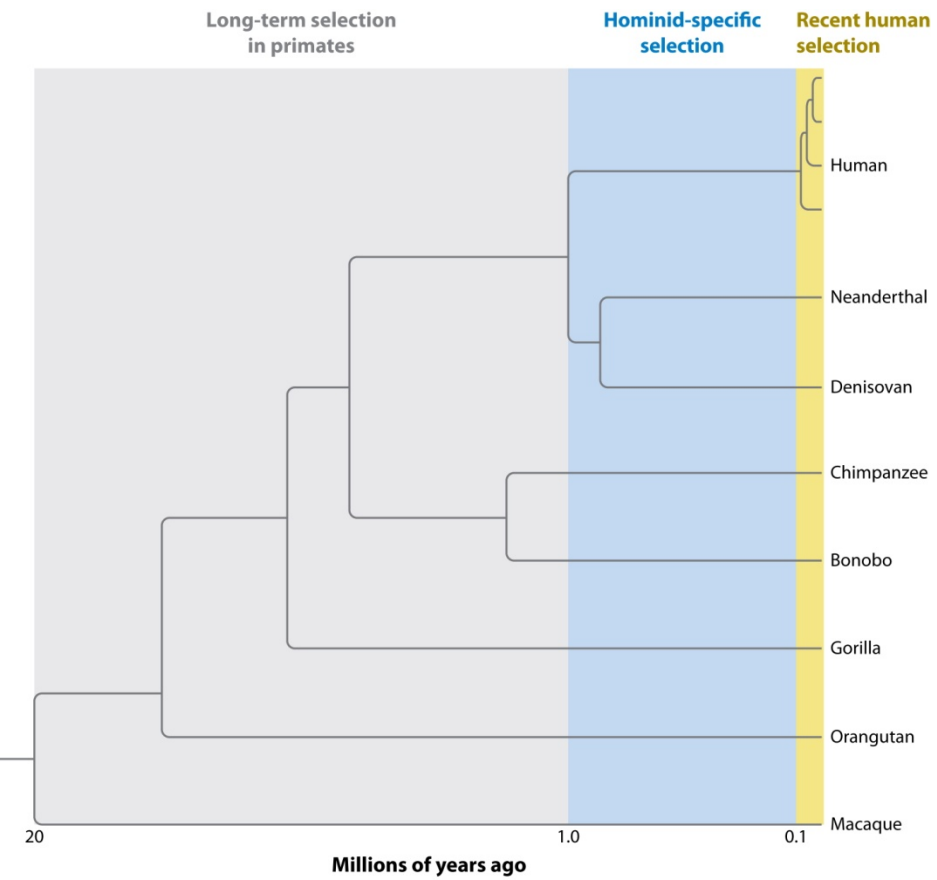


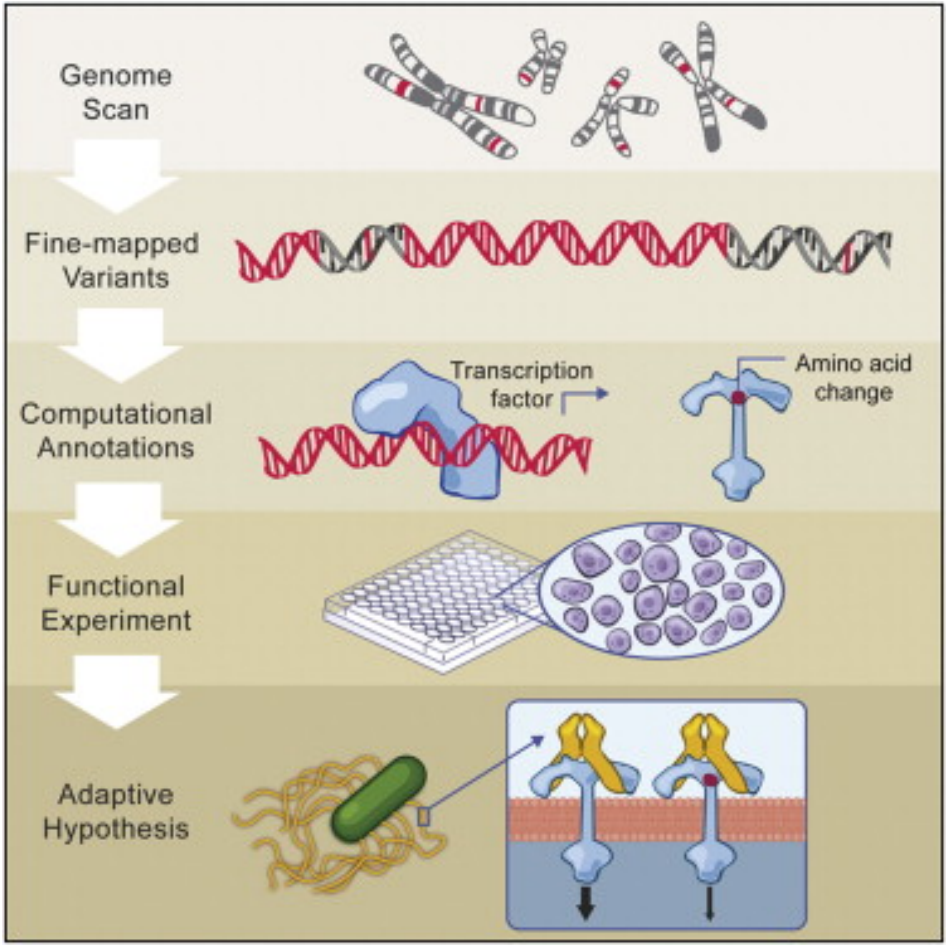
Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89



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

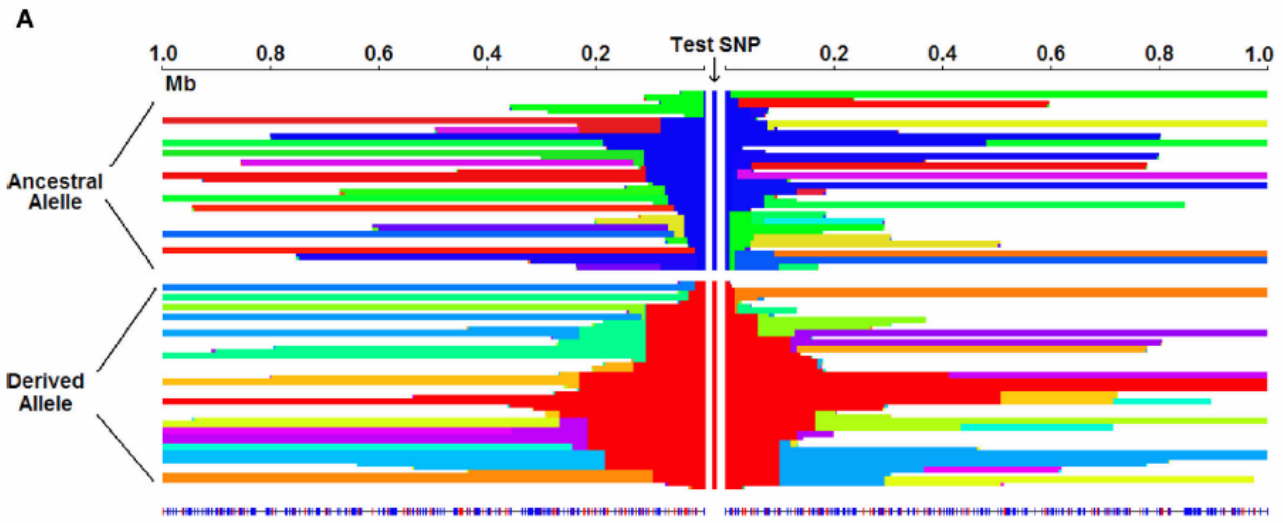
Positive Selection in Human Lineage



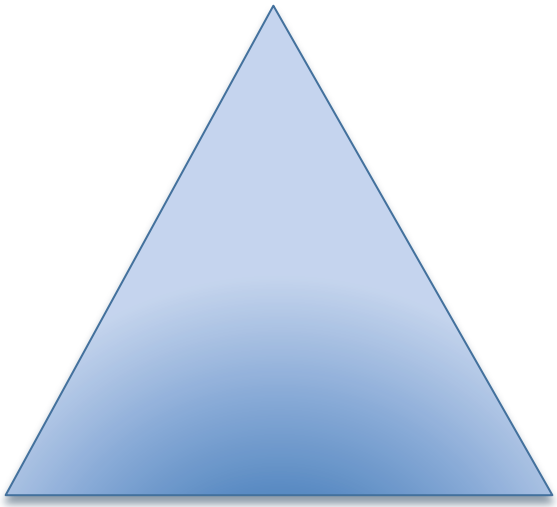


Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

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

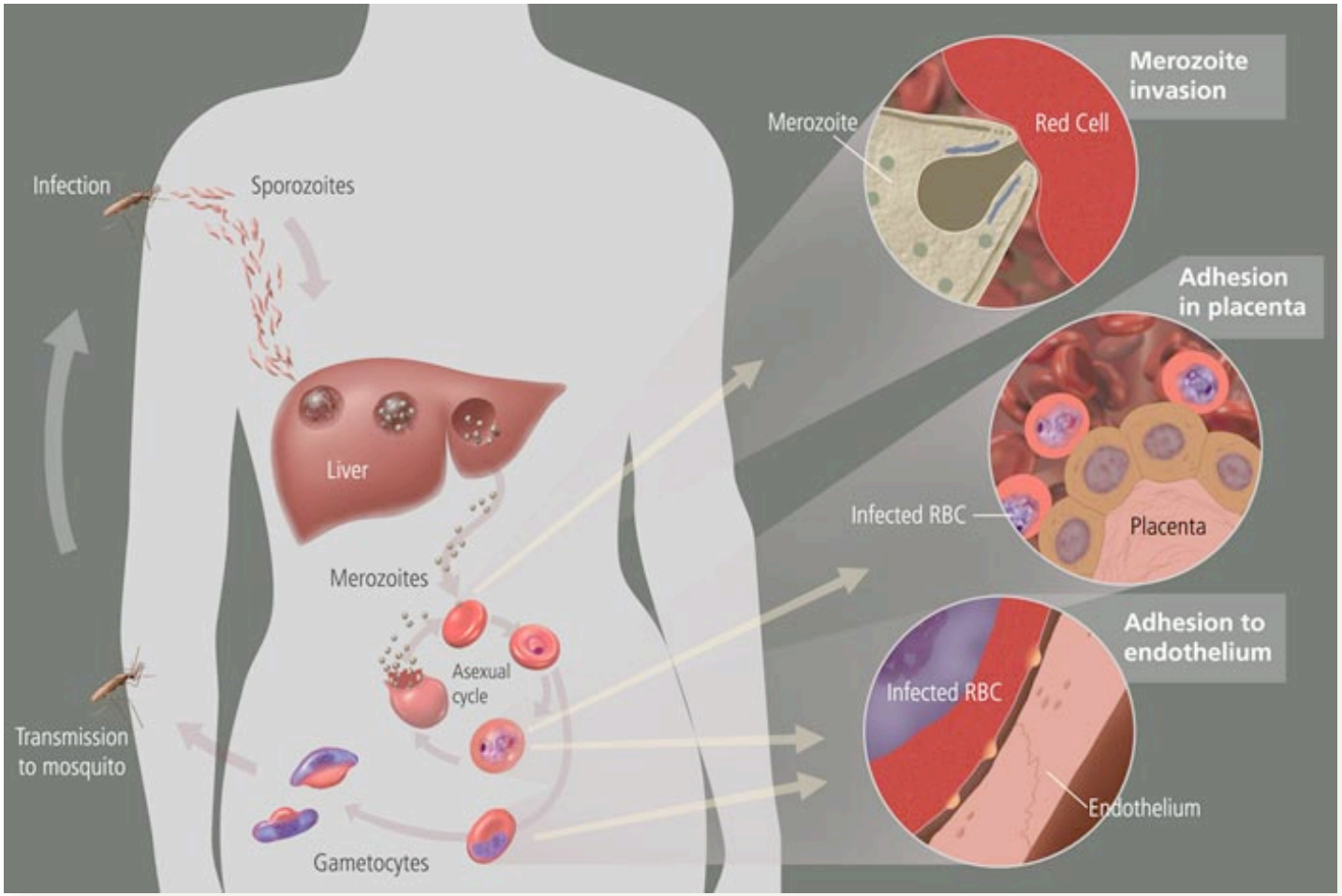


Image source:
NIH - http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg

Slide Credits:
Marc Schaub

Application: Malaria

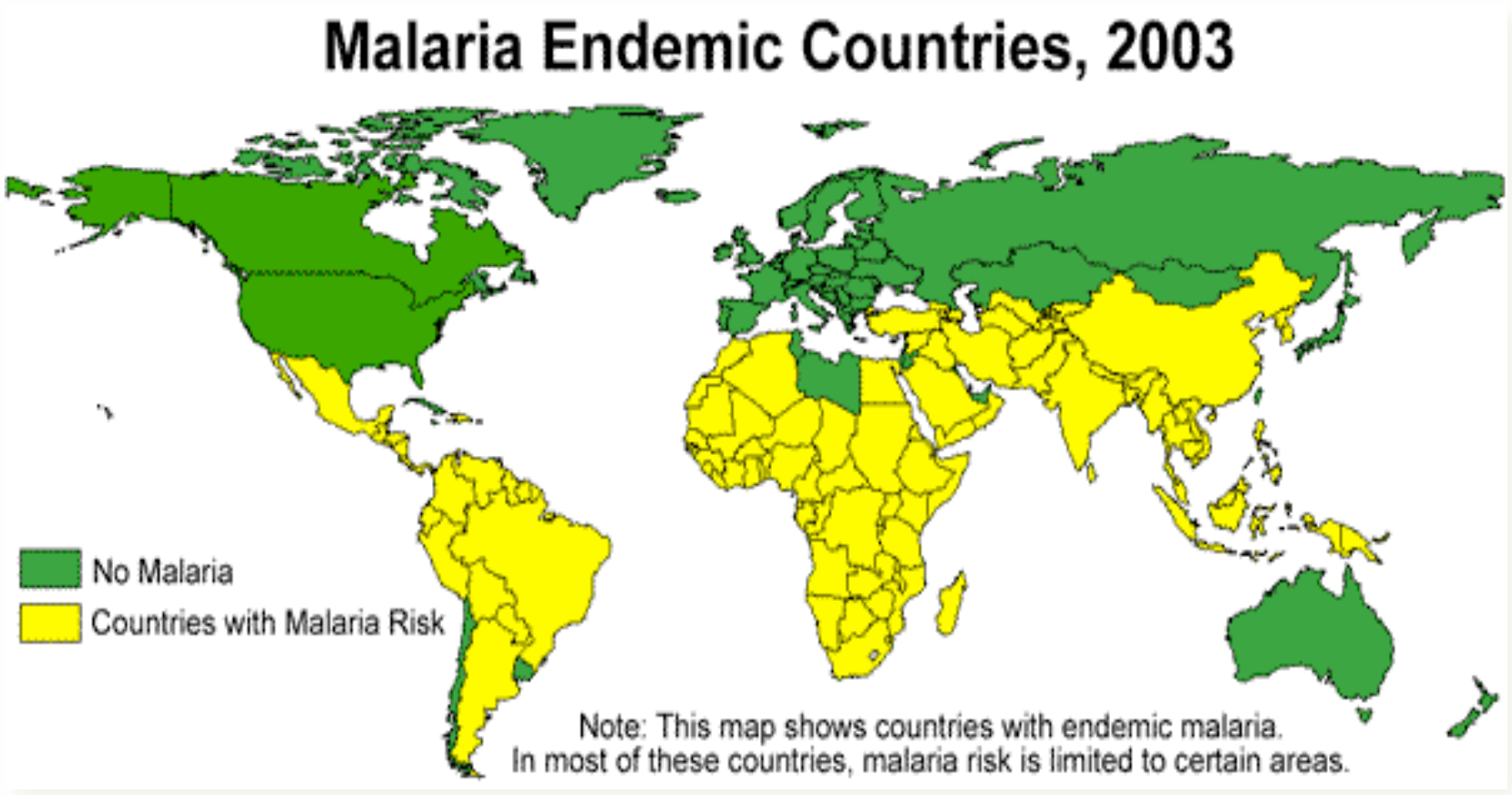
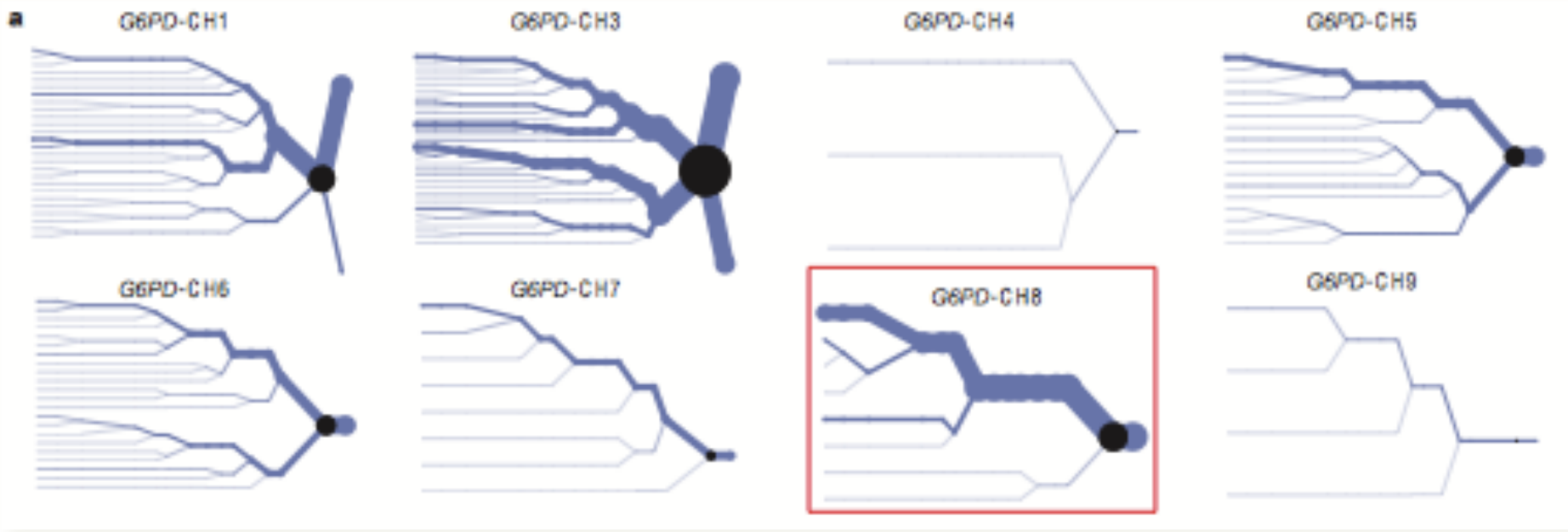


Image source: CDC - http://www.dpd.cdc.gov/dpdx/images/ParasiteImages/M-R/Malaria/malaria_risk_2003.gif

Slide Credits:
Marc Schaub

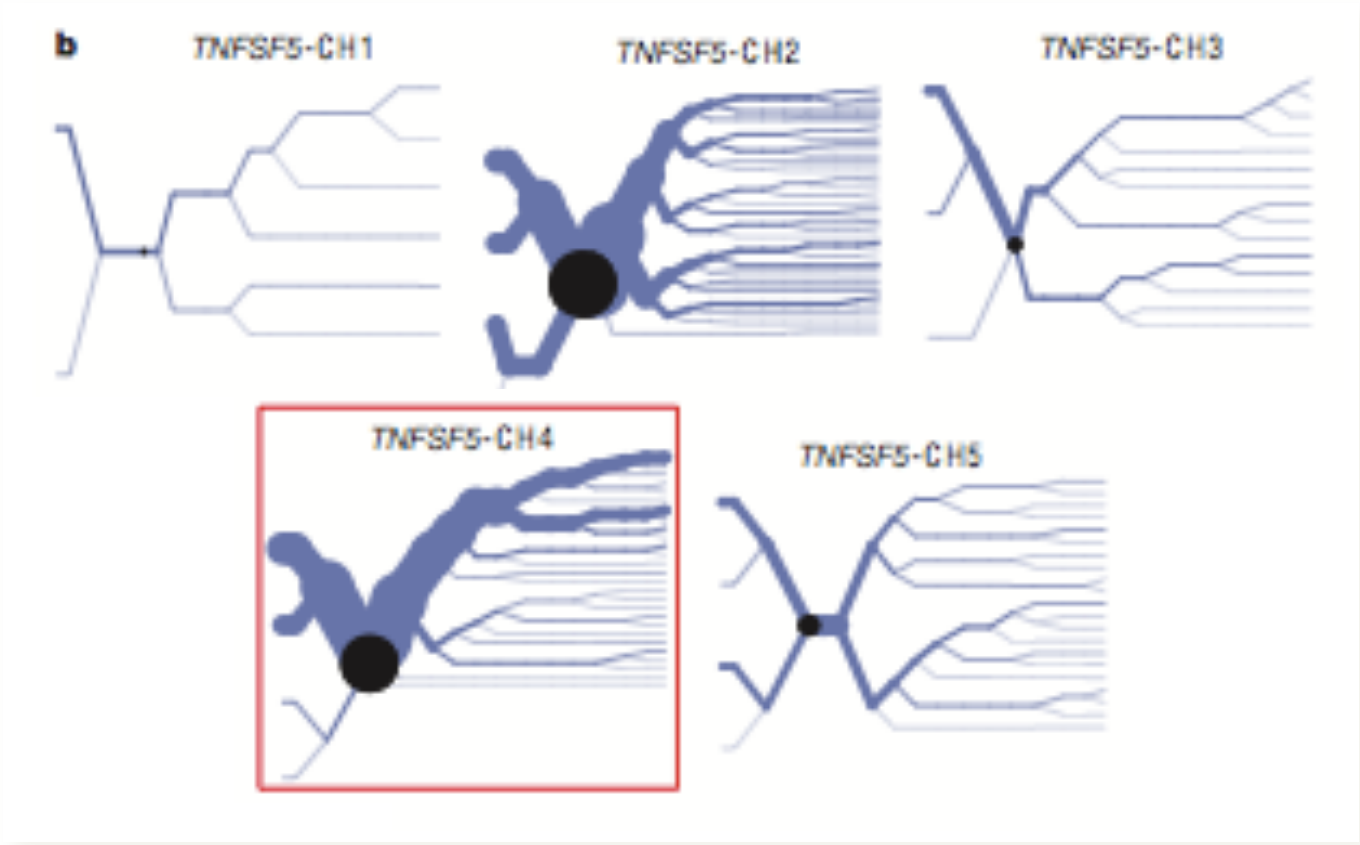
Results: G6PD



Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

Results: TNFSF5



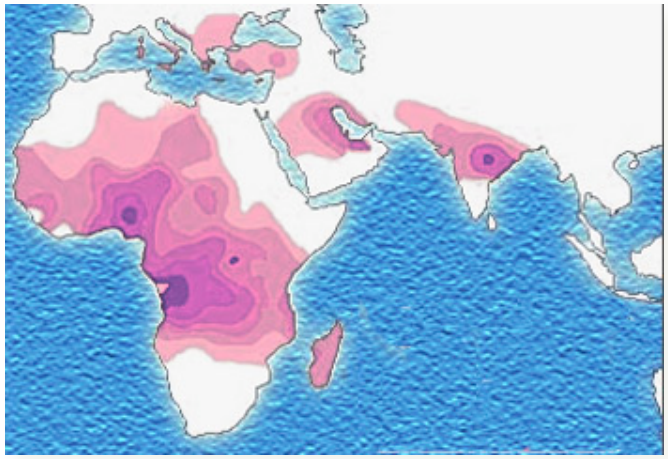
Source: Sabeti *et al.* Nature 2002.

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.

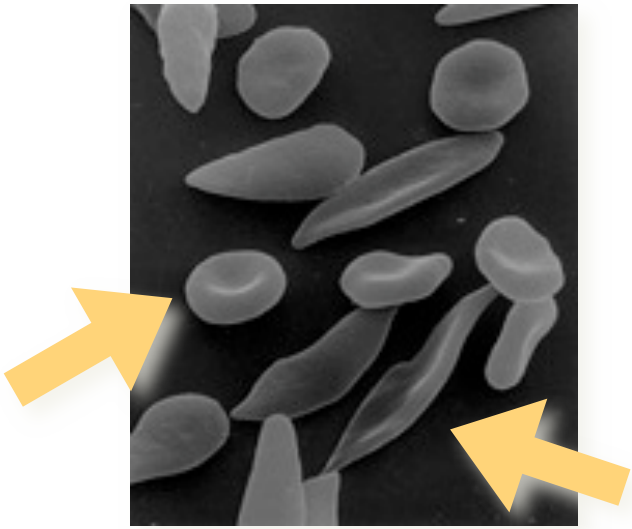


Image source: wikipedia.org

Slide Credits:
Marc Schaub

Lactose Intolerance

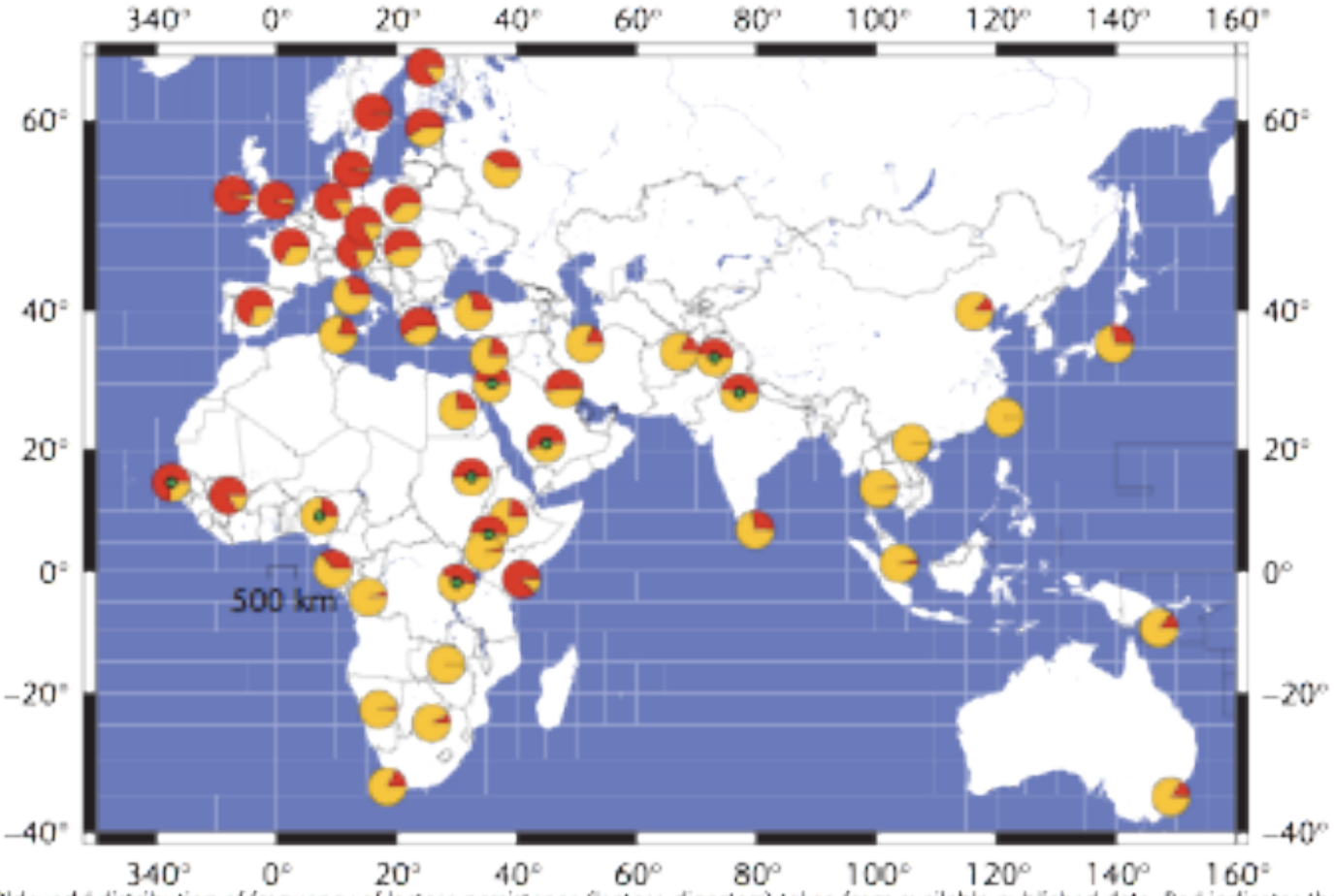
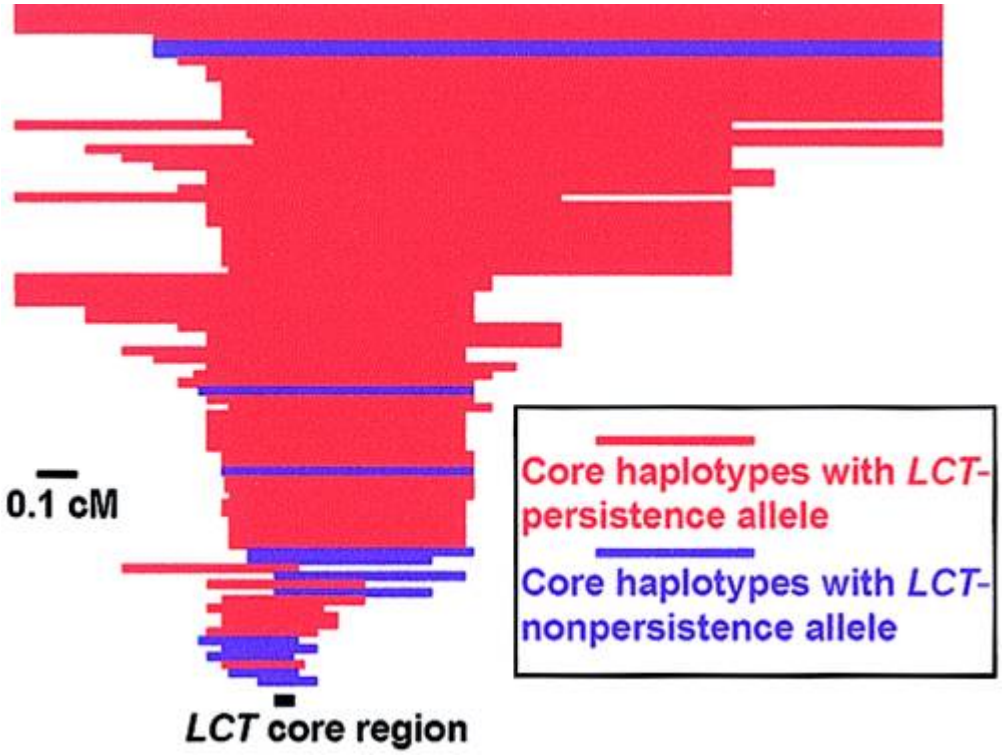


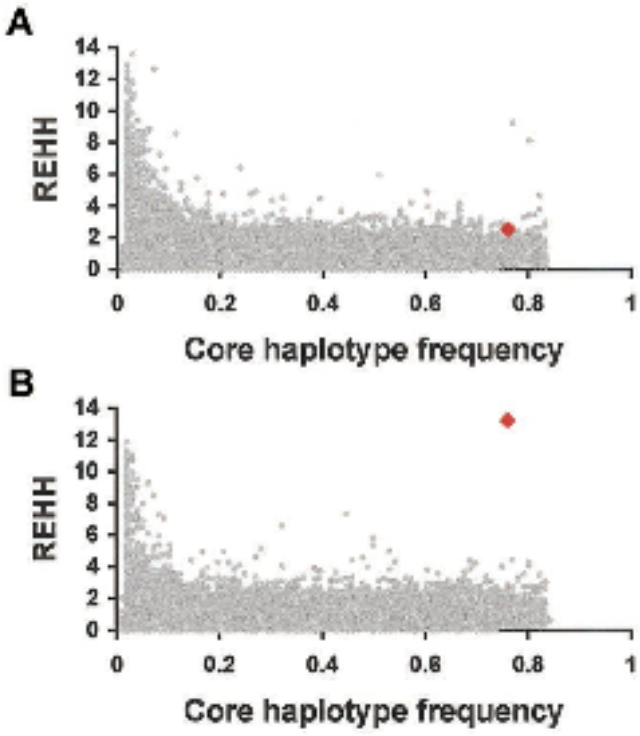
Figure 1 Old world distribution of frequency of lactase persistence (lactose digesters) taken from available published data. Red indicates the proportion of lactose digesters in a given population with yellow representing maldigesters. Charts with a green central circle indicate that the overall published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





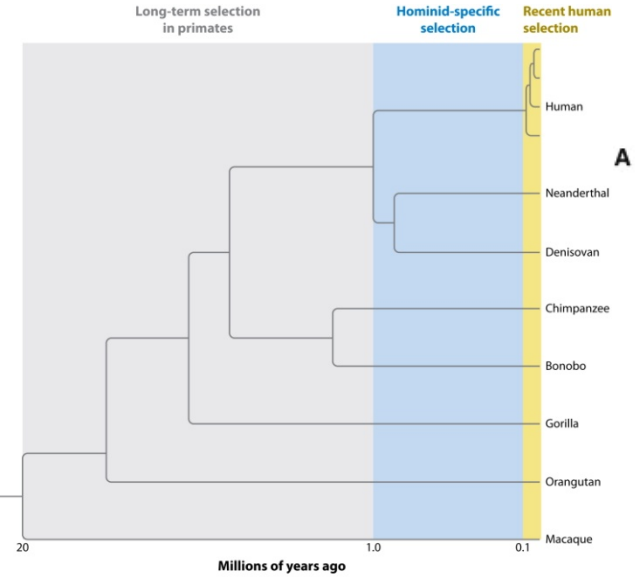
LCT, 5'

LCT, 3'

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

Slide Credits:
Marc Schaub

Positive Selection in Human Lineage



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