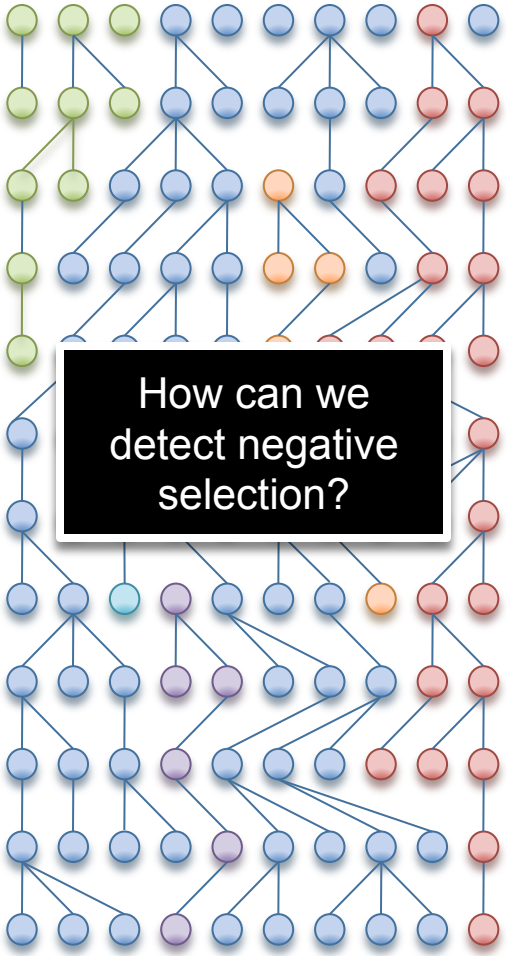
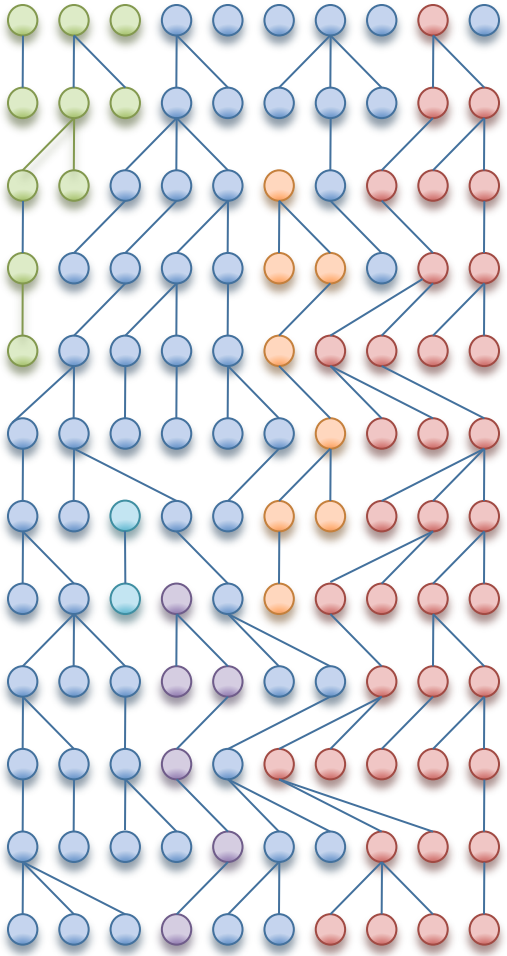
TGACTGAGGAGTTTACGGGAGCAAAGCGGCGTCATTGCTATTCGTATCTGTTTAG
0101011000010010100001010101010100110111001100011000101000100101

Human Population Genomics

Fixation, Positive & Negative Selection



Negative Selection



Neutral Drift

Positive Selection

How can we detect positive selection?

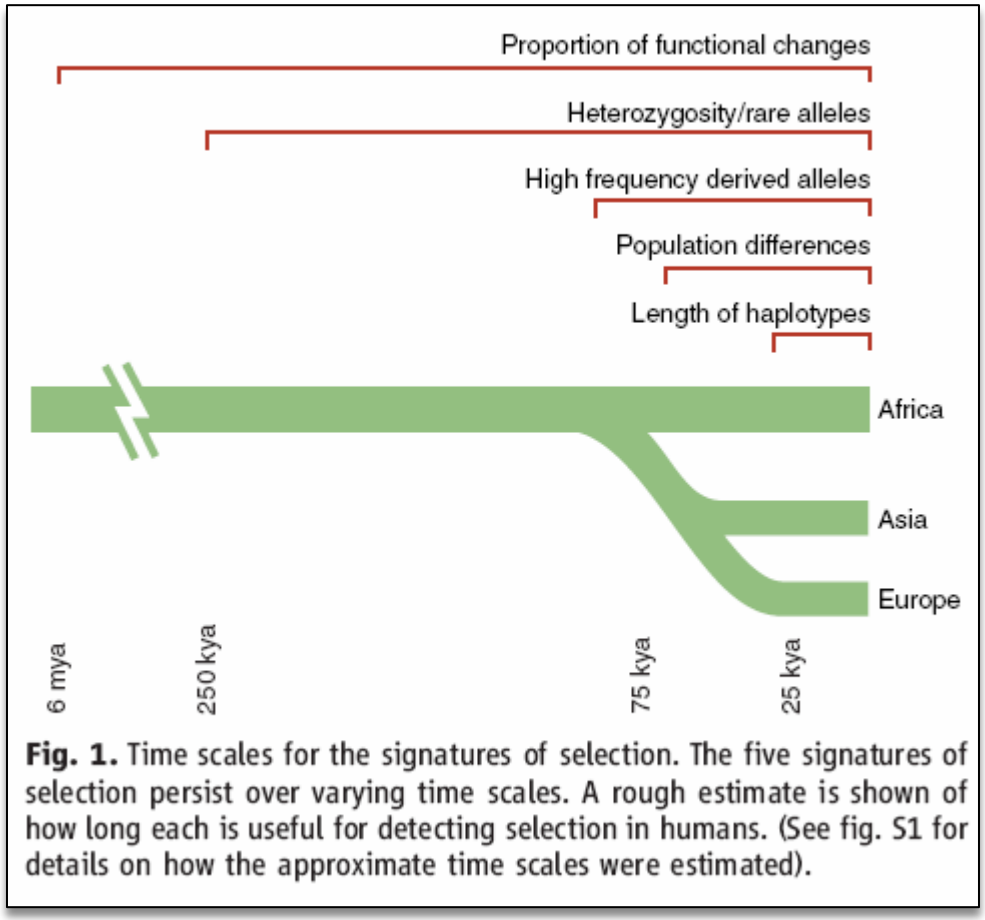


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans. (See fig. S1 for details on how the approximate time scales were estimated).

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

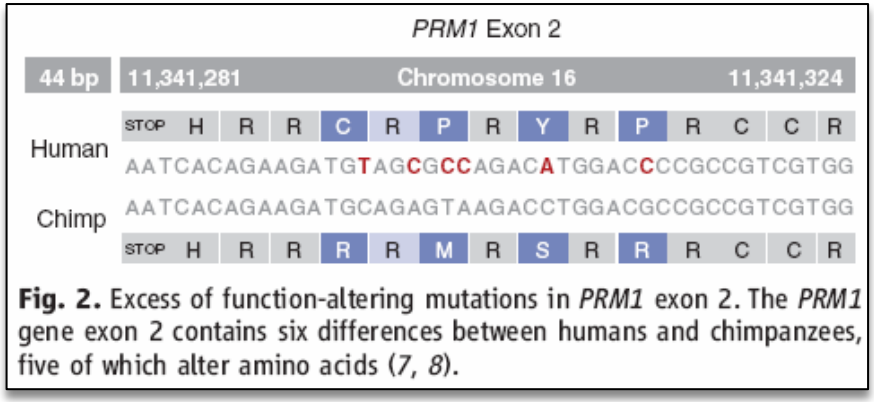


Fig. 2. Excess of function-altering mutations in *PRM1* exon 2. The *PRM1* gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (*7*, *8*).

How can we detect positive selection?

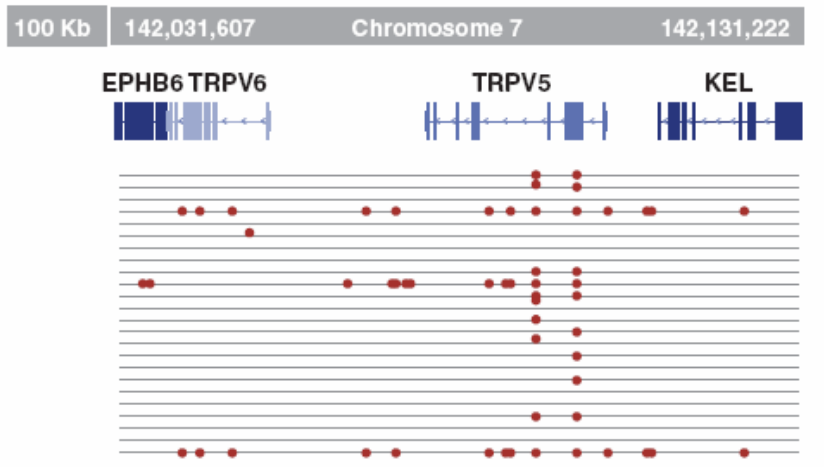


Fig. 3. Low diversity and many rare alleles at the Kell blood antigen cluster. On the basis of three different statistical tests, the 115-kb region (containing four genes) shows evidence of a selective sweep in Europeans (*28*).

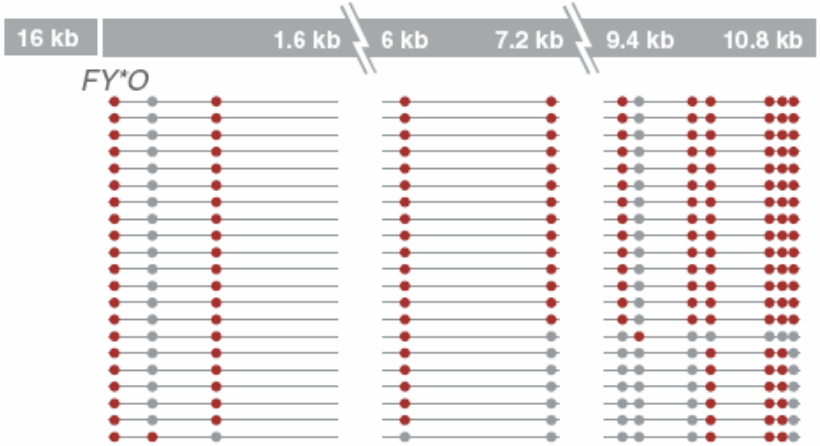


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (*FY*) gene (*34*). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

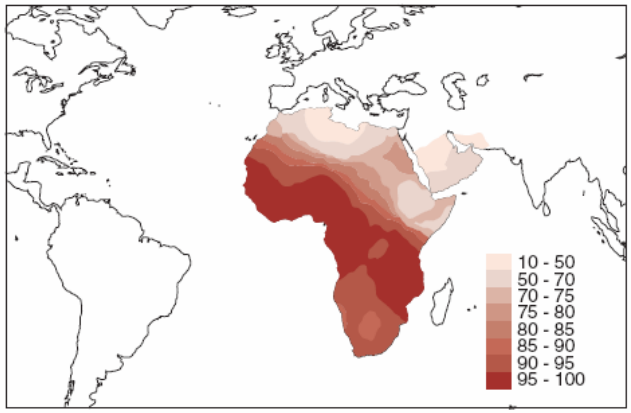


Fig. 5. Extreme population differences in *FY*O* allele frequency. The *FY*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (*38*).

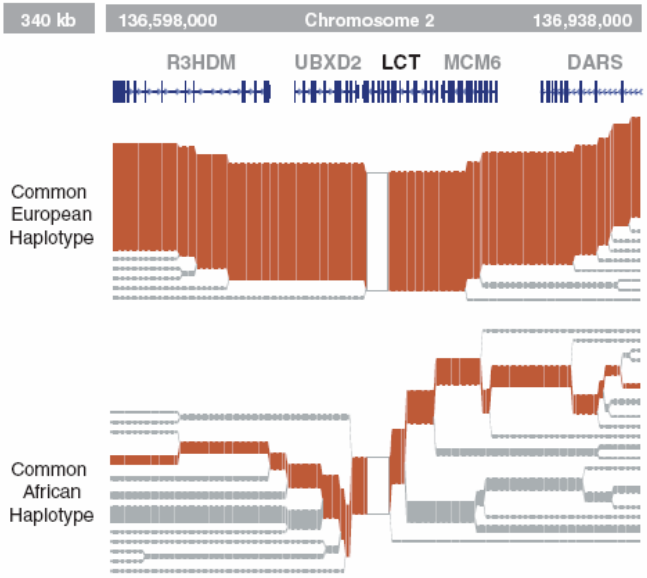
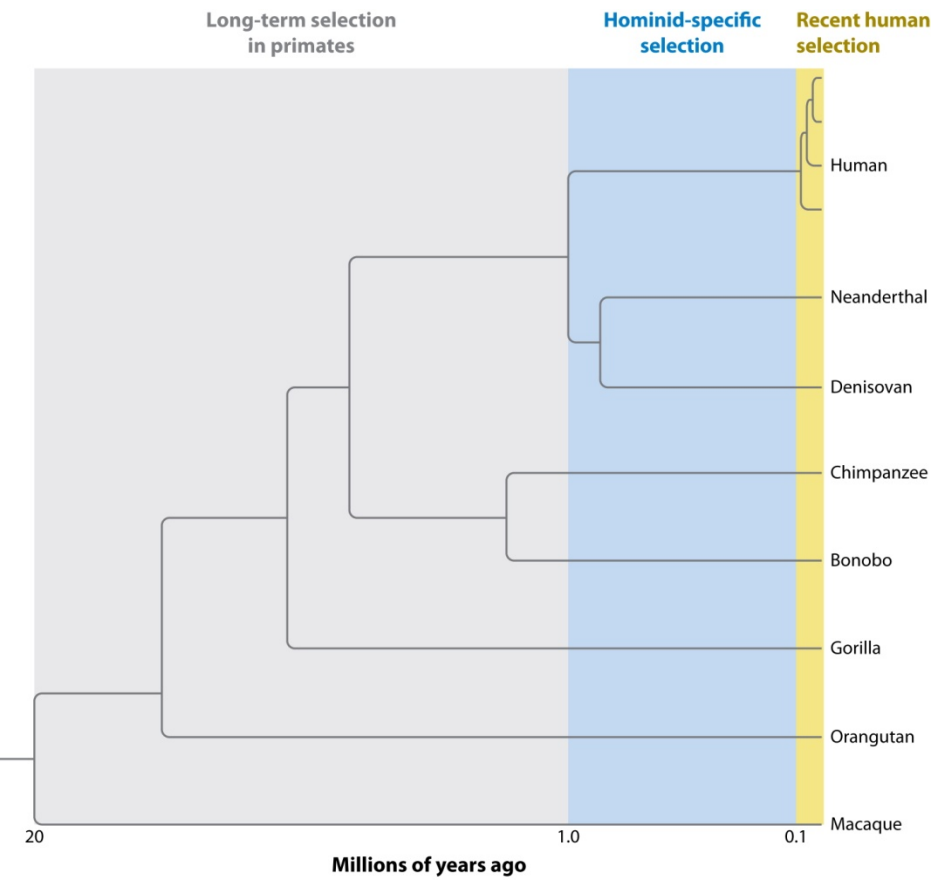
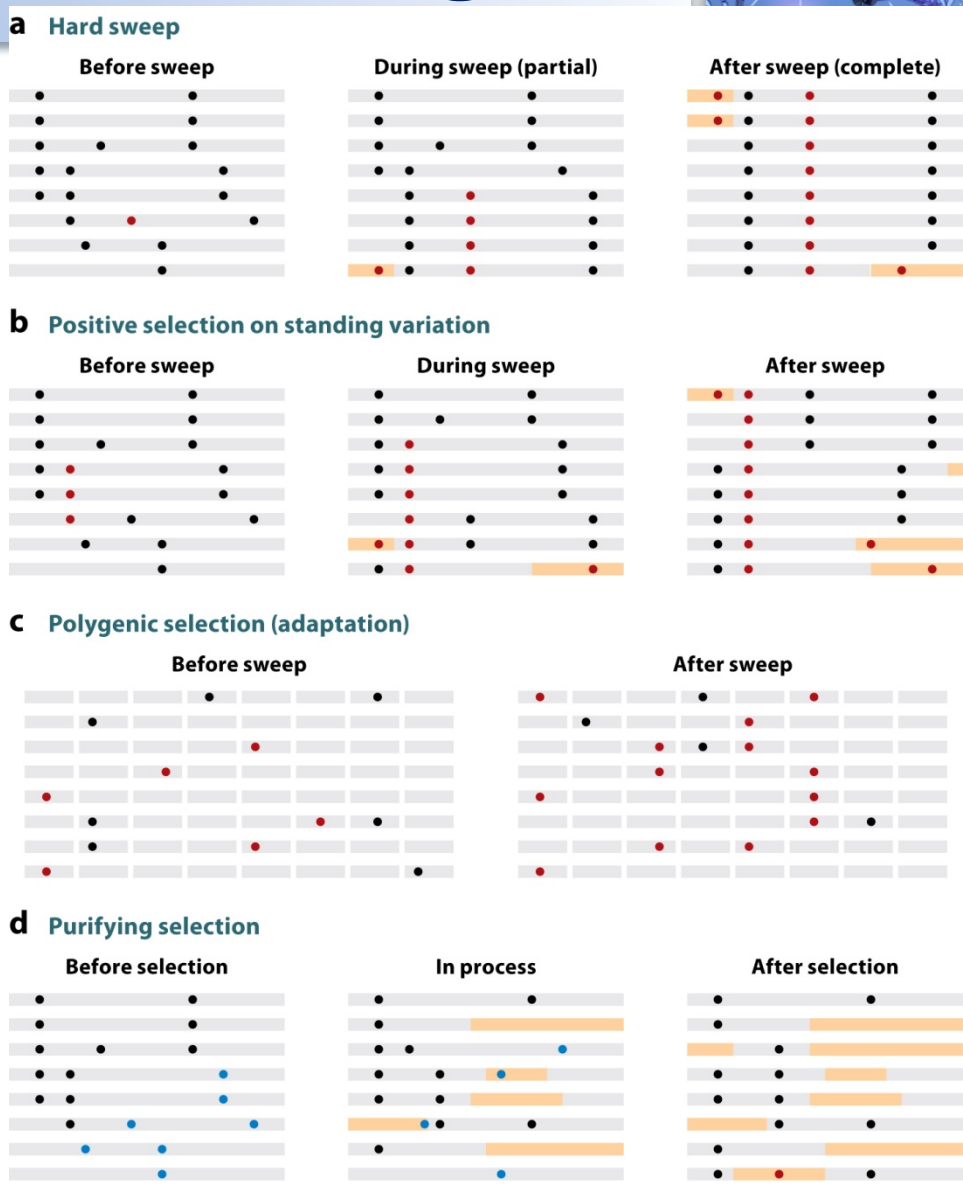


Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (*6*).

Positive Selection in Human Lineage

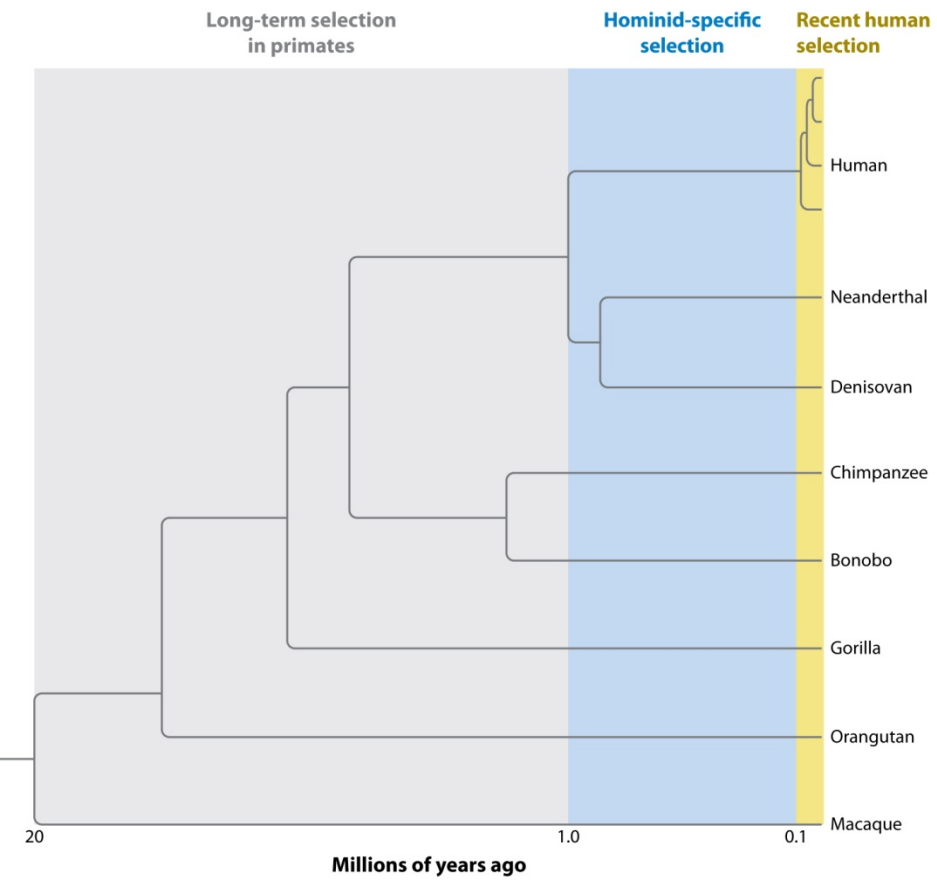


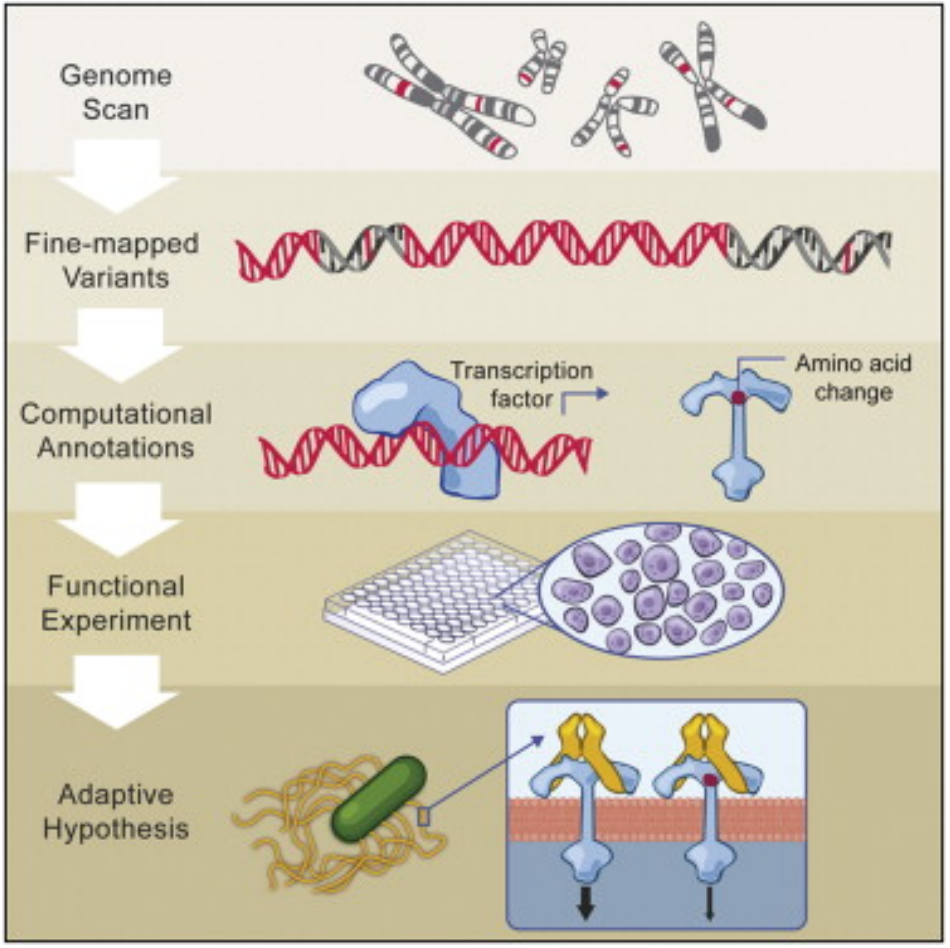
Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89



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

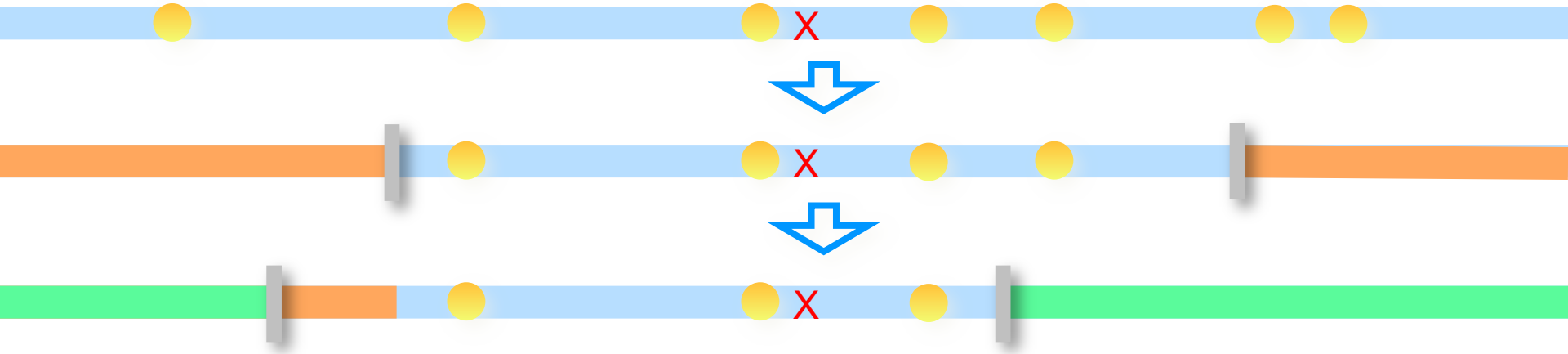
Positive Selection in Human Lineage





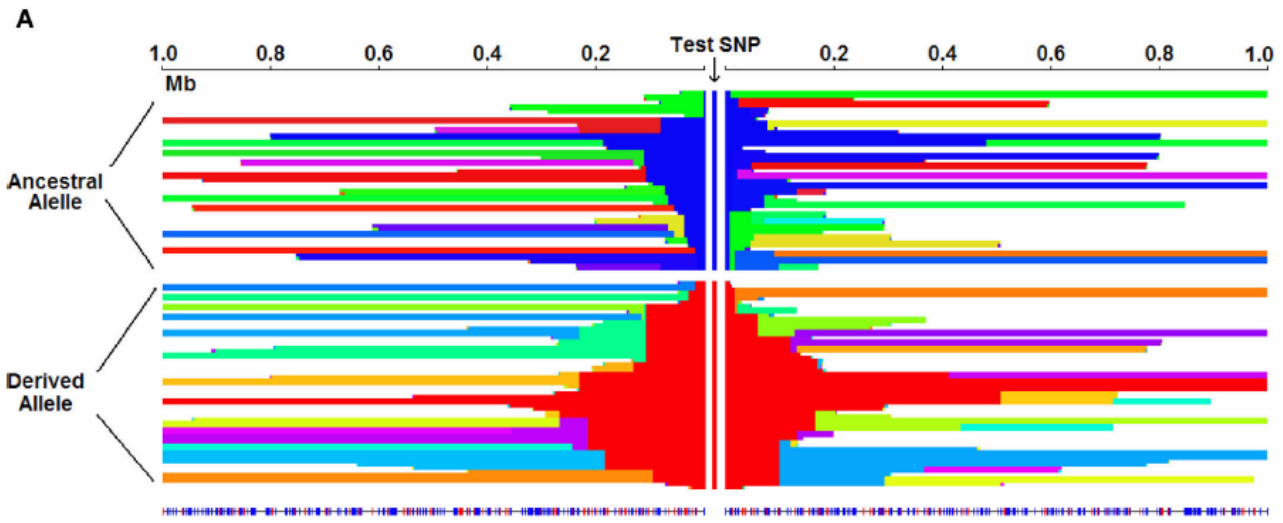
Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

Mutations and LD

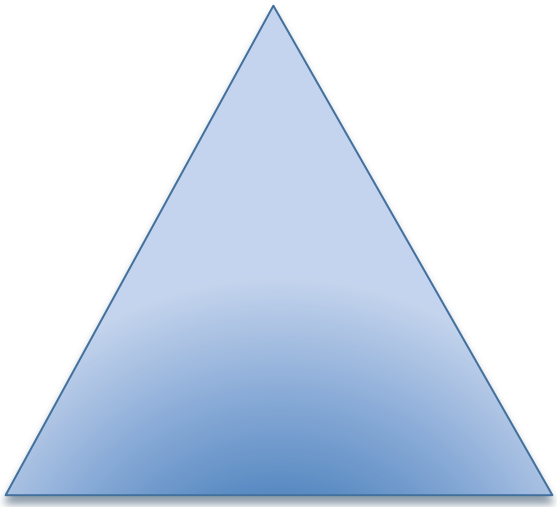


Slide Credits:
Marc Schaub

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

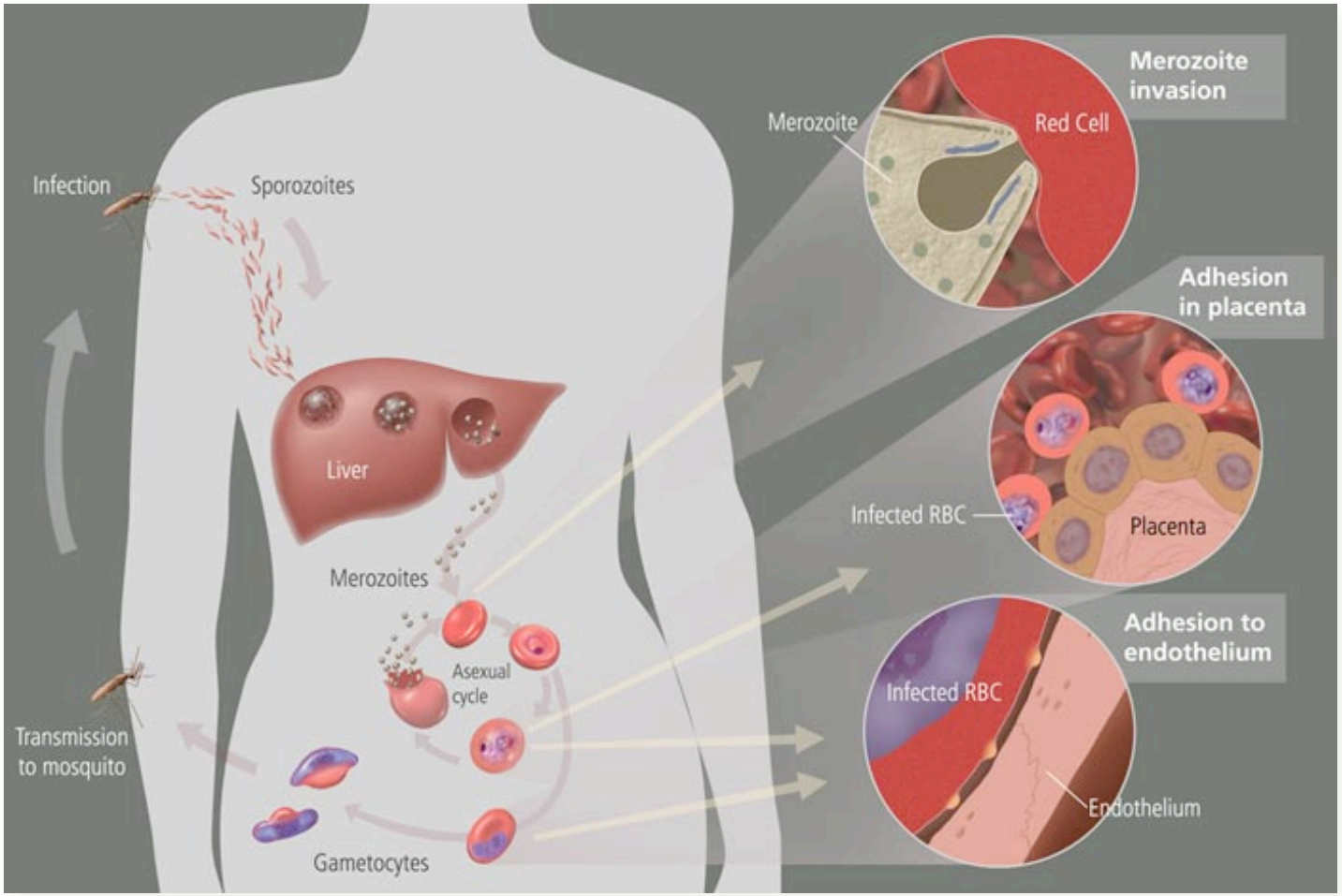


Image source:
NIH - http://history.nih.gov/exhibits/bowman/images/malariacycleBig.jpg

Slide Credits:
Marc Schaub

Application: Malaria

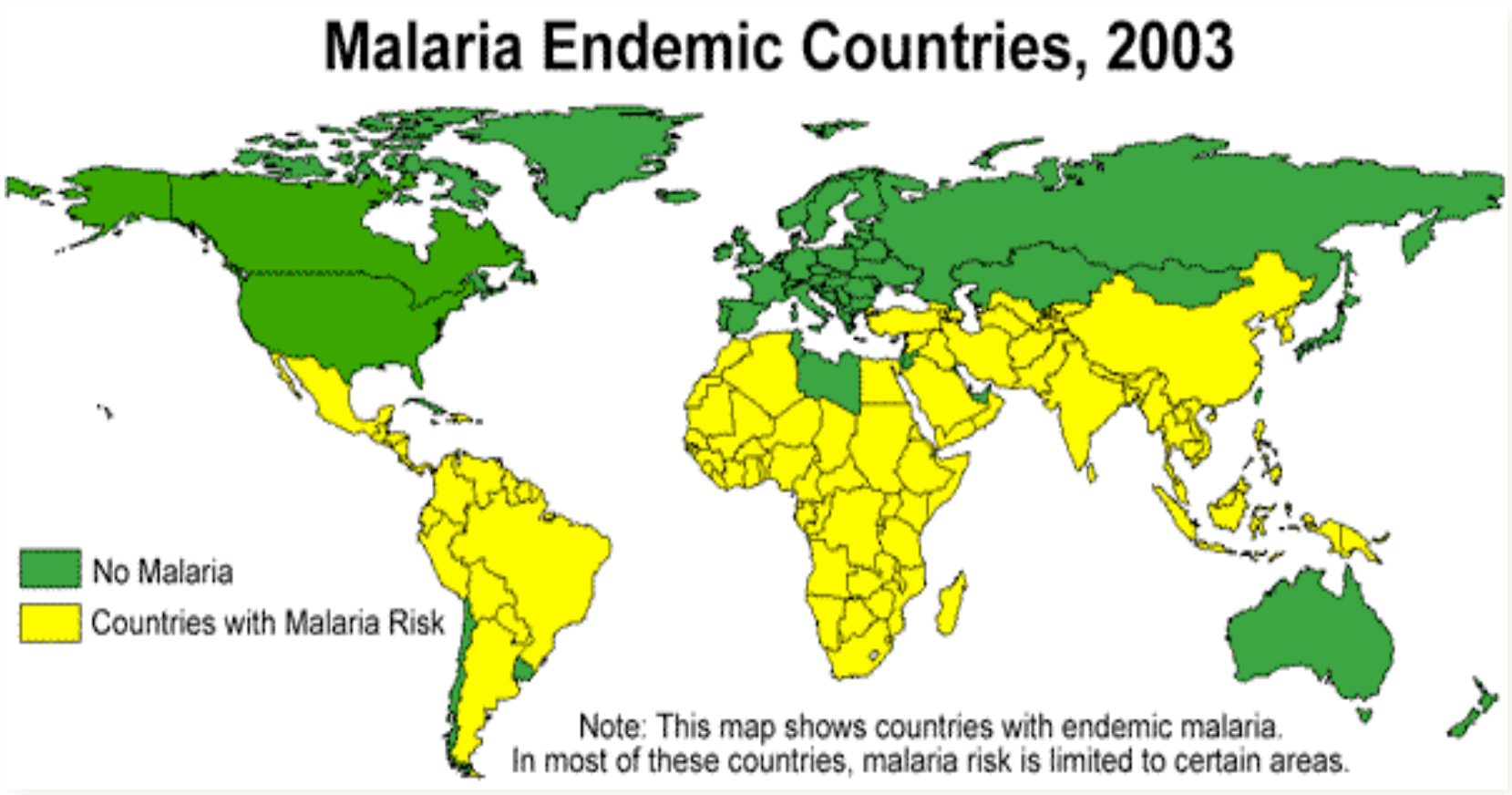
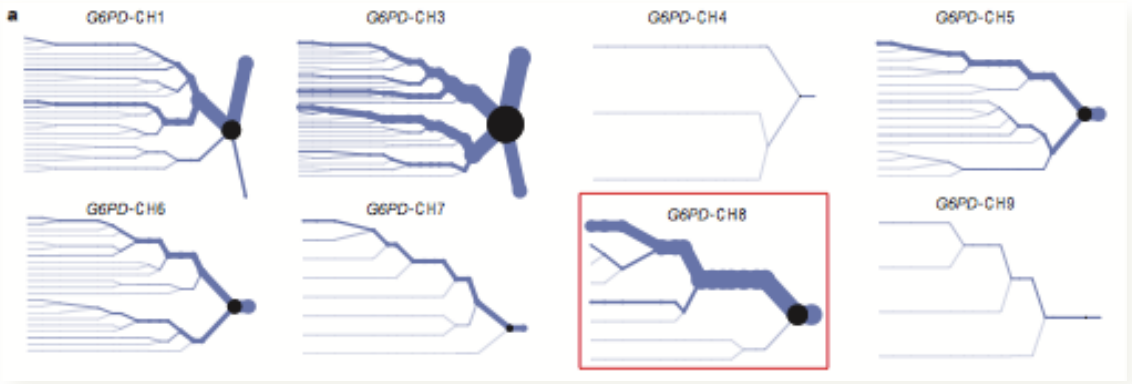
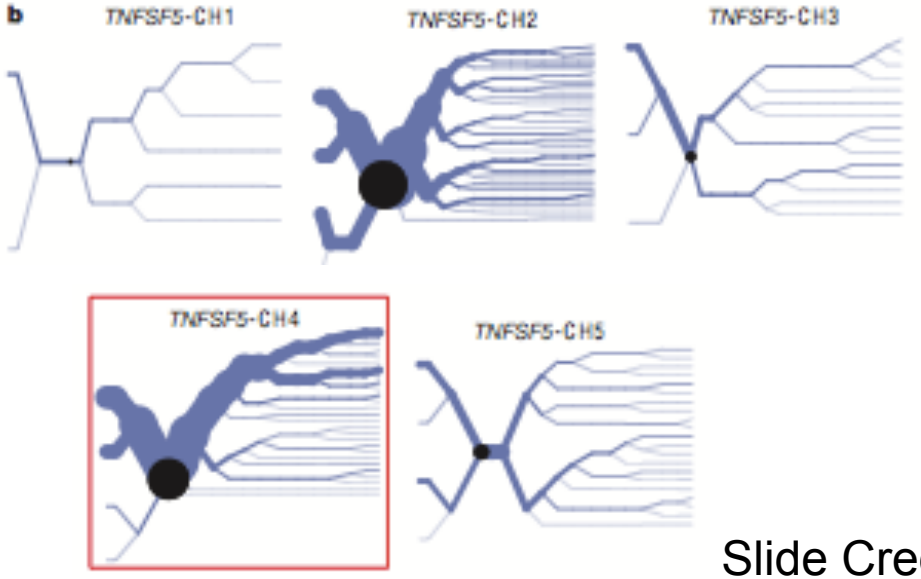


Image source: CDC - http://www.dpd.cdc.gov/dpdx/images/ParasiteImages/M-R/Malaria/malaria_risk_2003.gif

Slide Credits:
Marc Schaub

Results: G6PD, TNFSF5





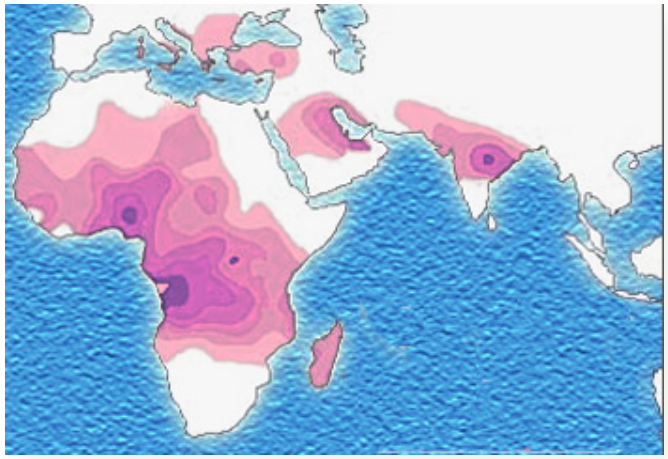
Source: Sabeti *et al.* Nature 2002.

Slide Credits:
Marc Schaub

Malaria and Sickle-cell Anemia

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

Distribution of malaria



Distribution of sickle-cell anemia

Image source: wikipedia.org

Slide Credits:
Marc Schaub

Lactose Intolerance

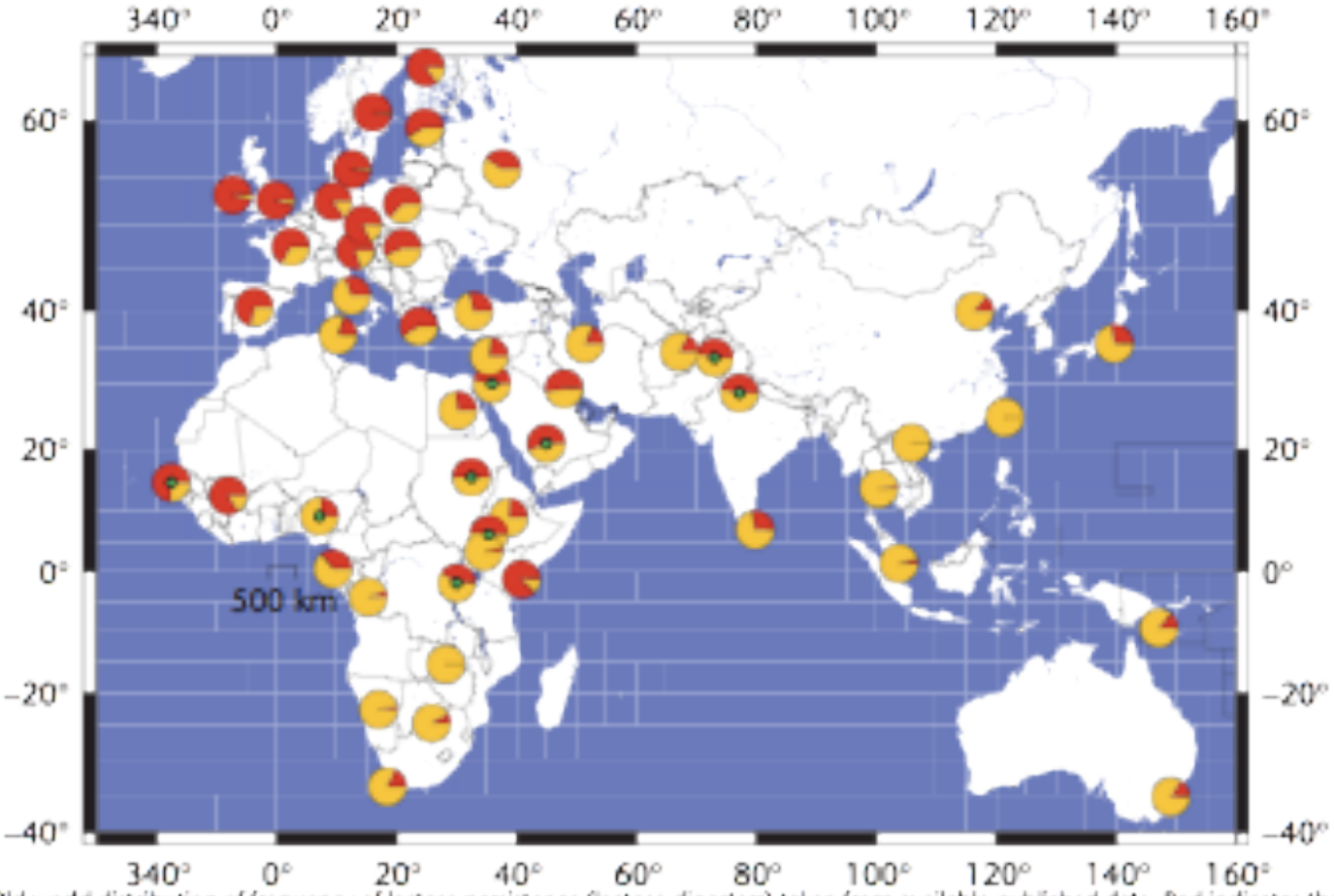
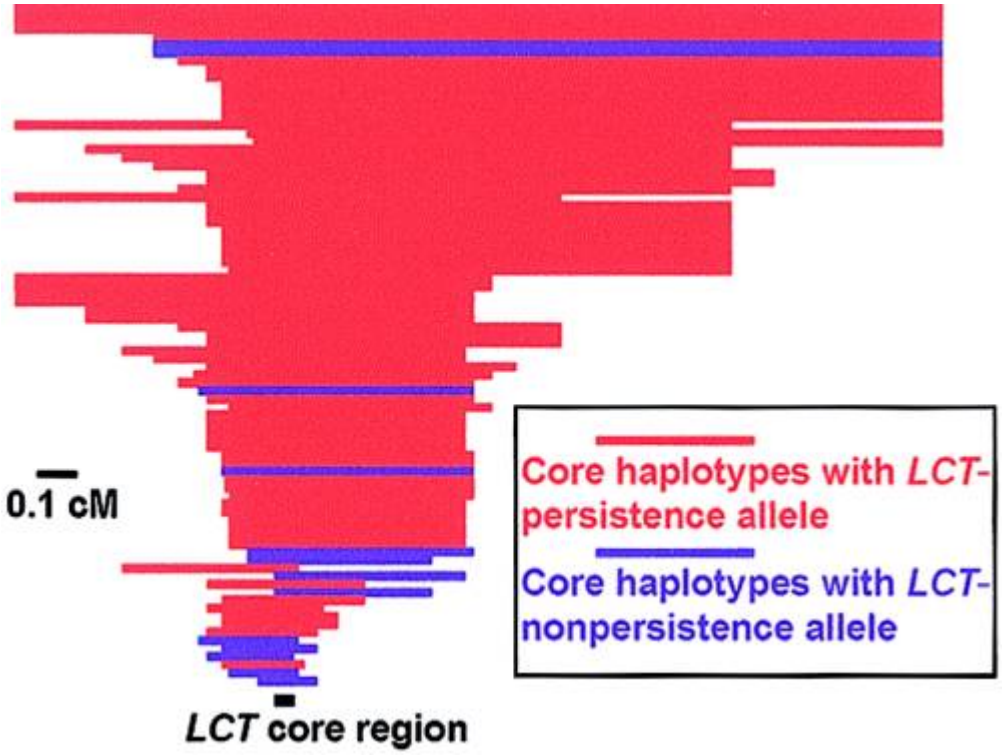


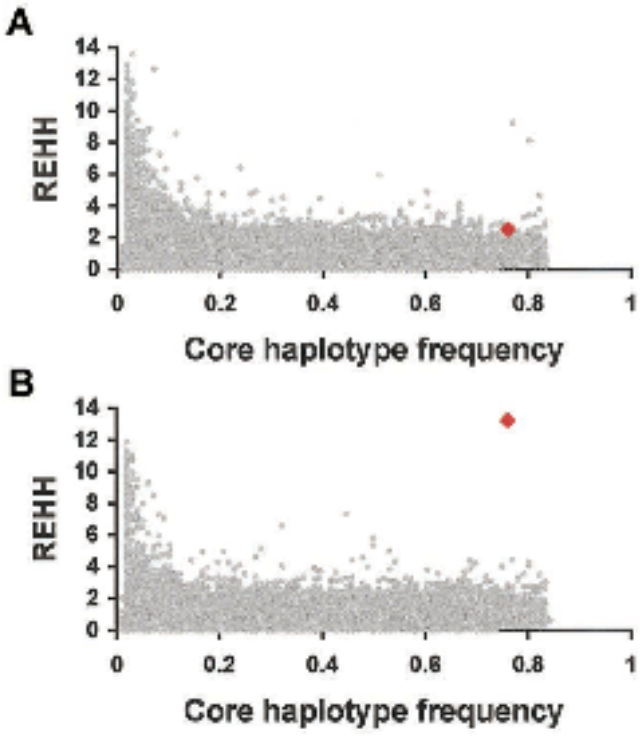
Figure 1 Old world distribution of frequency of lactase persistence (lactose digesters) taken from available published data. Red indicates the proportion of lactose digesters in a given population with yellow representing maldigesters. Charts with a green central circle indicate that the overall published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

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

Slide Credits:
Marc Schaub

Lactose Intolerance





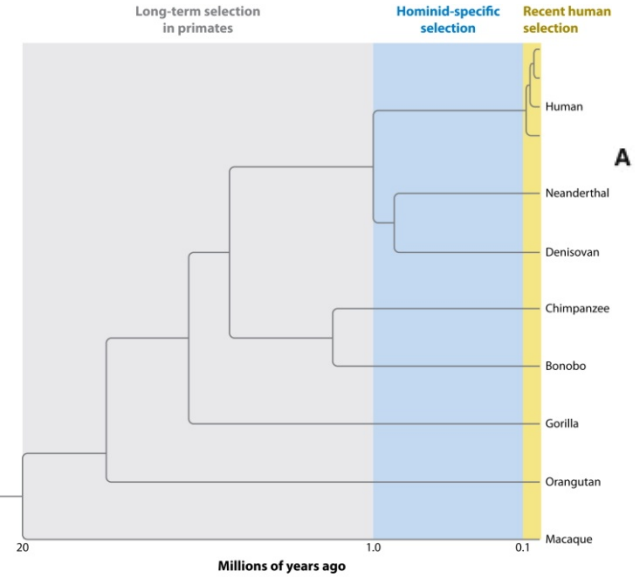
LCT, 5'

LCT, 3'

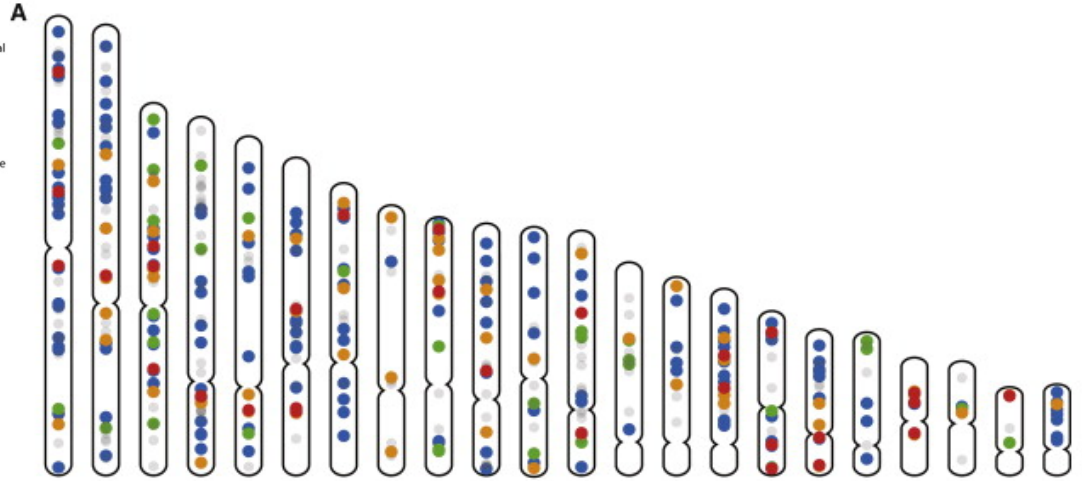
Source: Bersaglieri *et al.* Am. J. Hum. Genet. 2004.

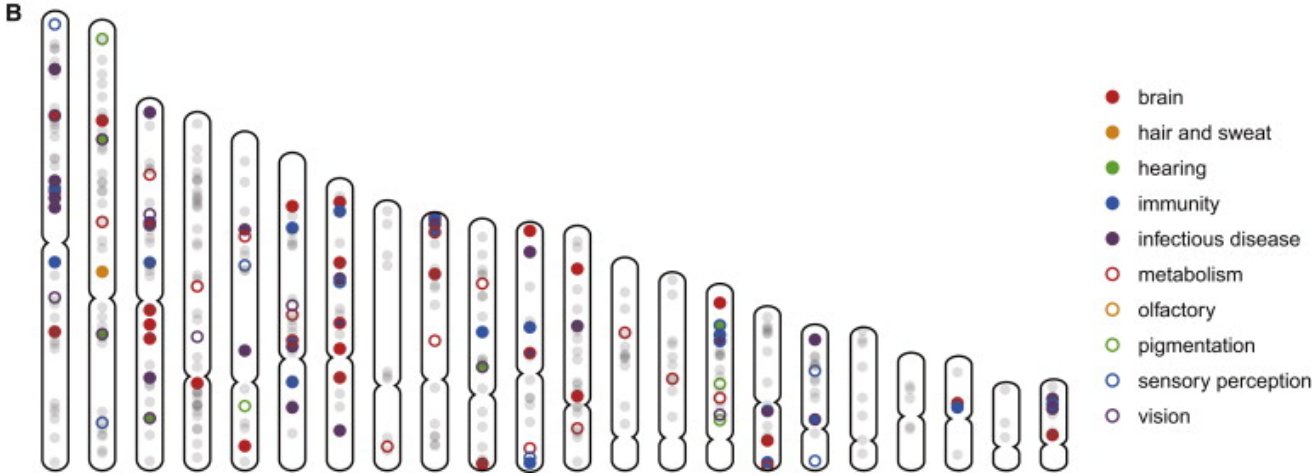
Slide Credits:
Marc Schaub

Positive Selection in Human Lineage

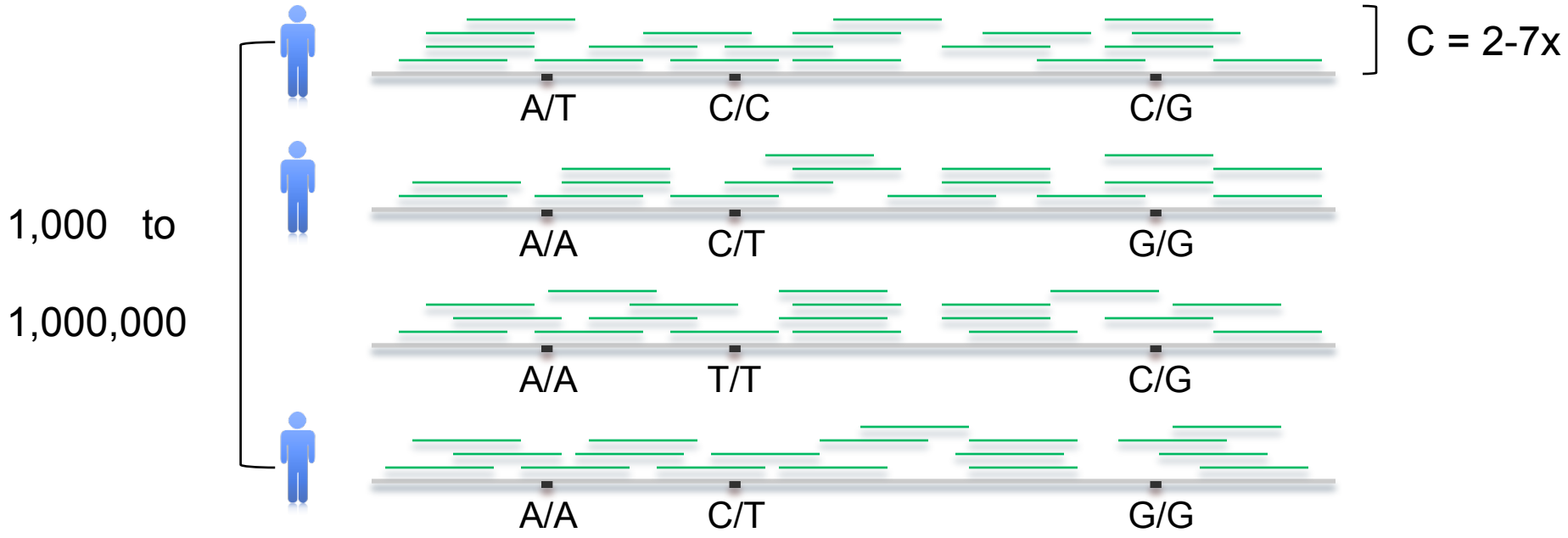


Fu W, Akey JM. 2013.
Annu. Rev. Genomics Hum. Genet. 14:467–89

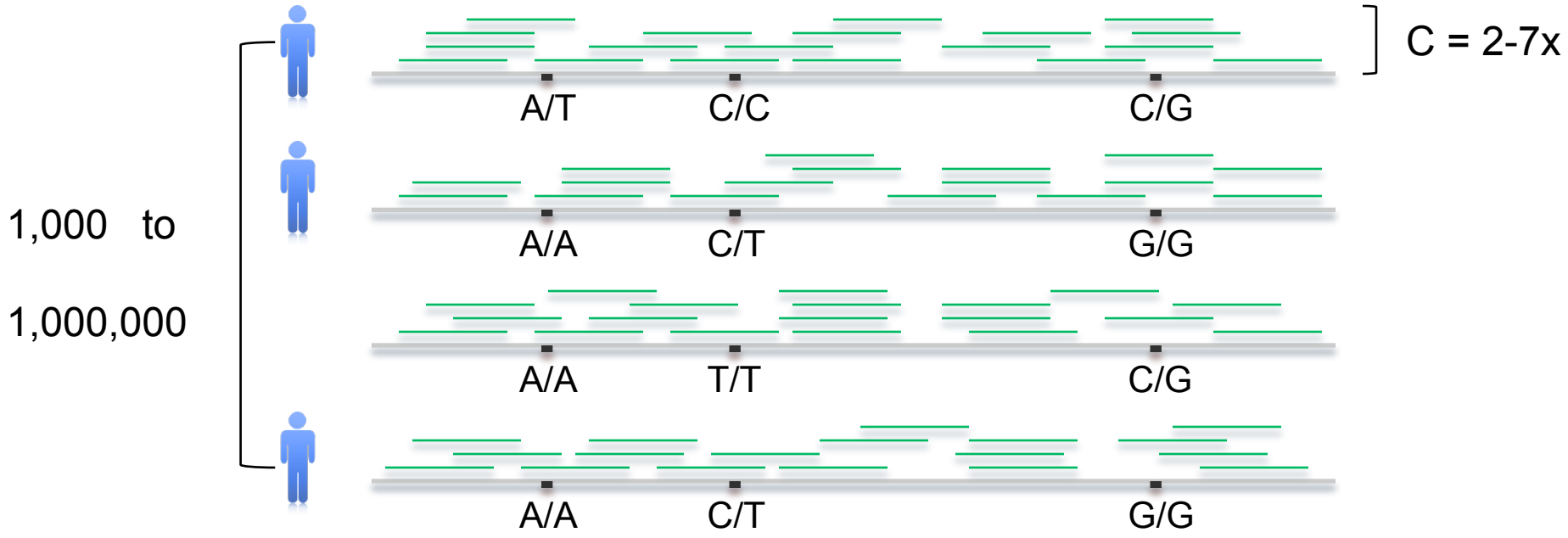




Population Sequencing



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 : [$p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{data})$]

Population Sequencing

When C is high (>30x),

Prob(g_{ij} = *g* | data) ~

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

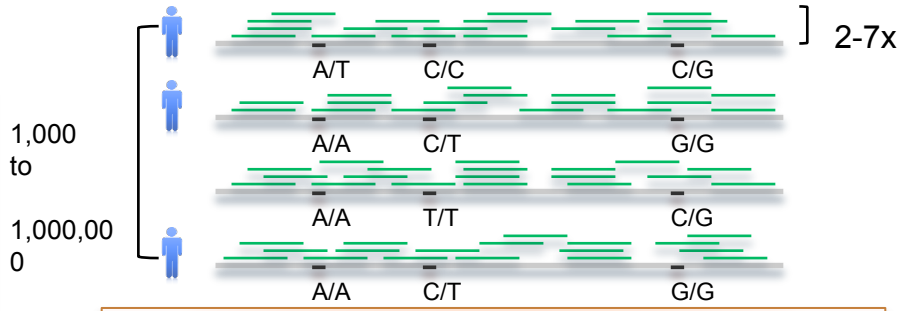
fast & easy

When C is low,

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

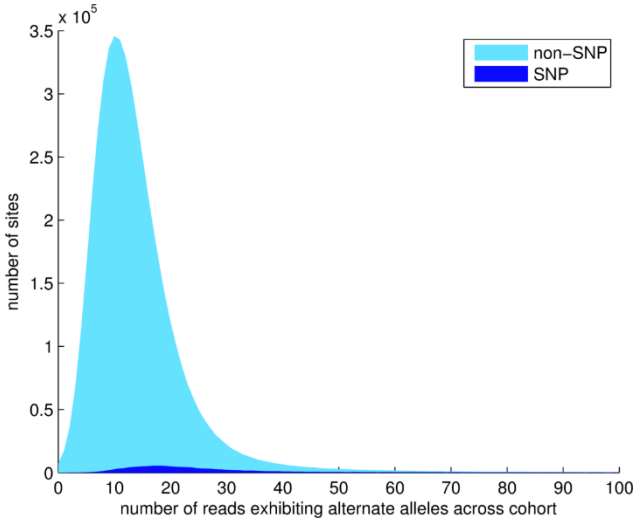
positions j' ≠ j in all individuals

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_{ijg} = Prob(g_{ij} = *g* | data)]

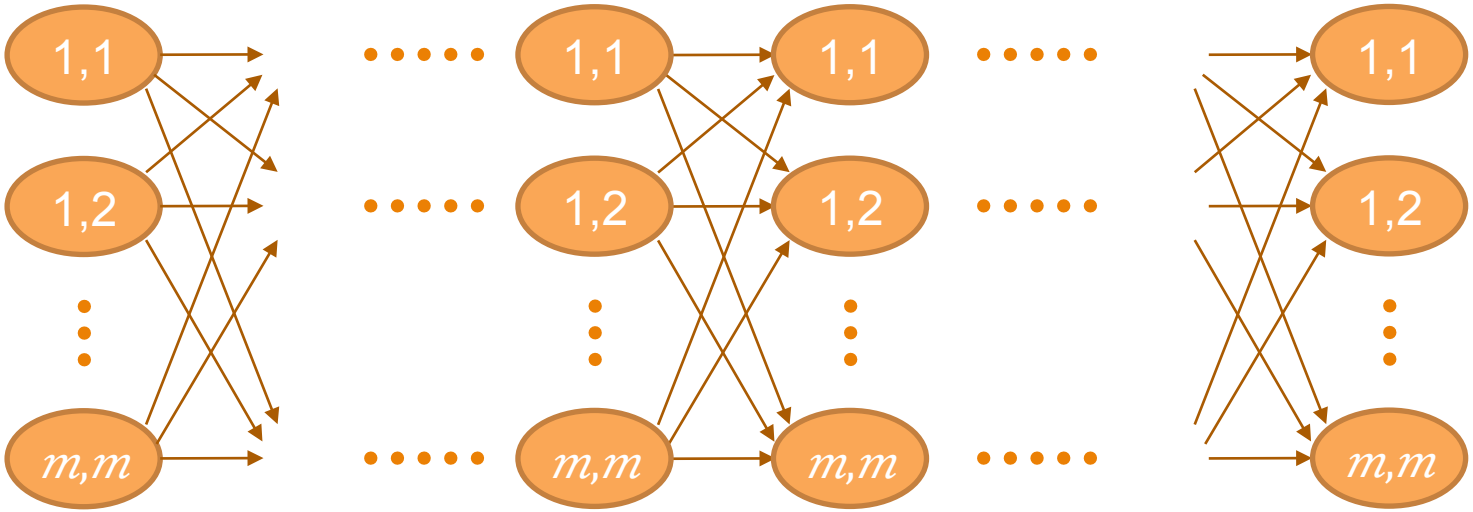


1000 Genomes Project, 2535 individuals, 7x sequencing

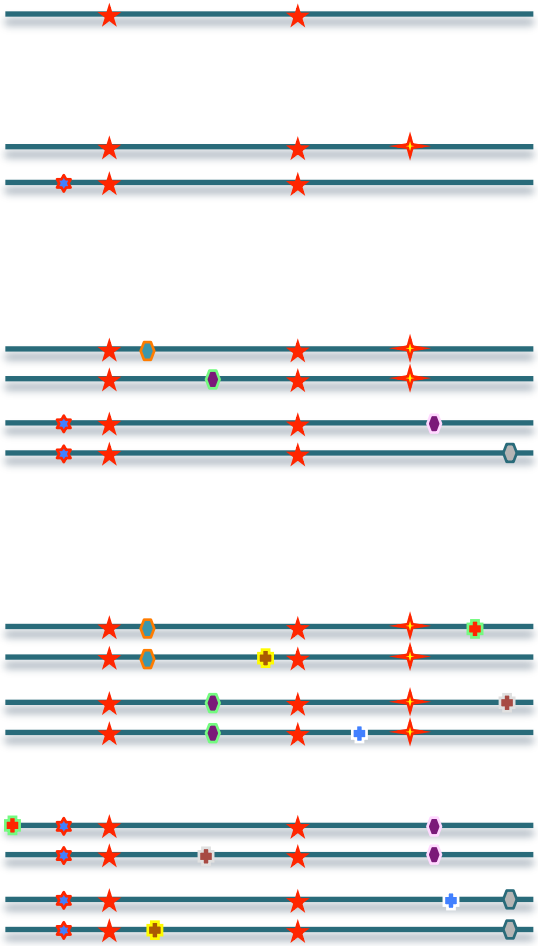
HMM-based models

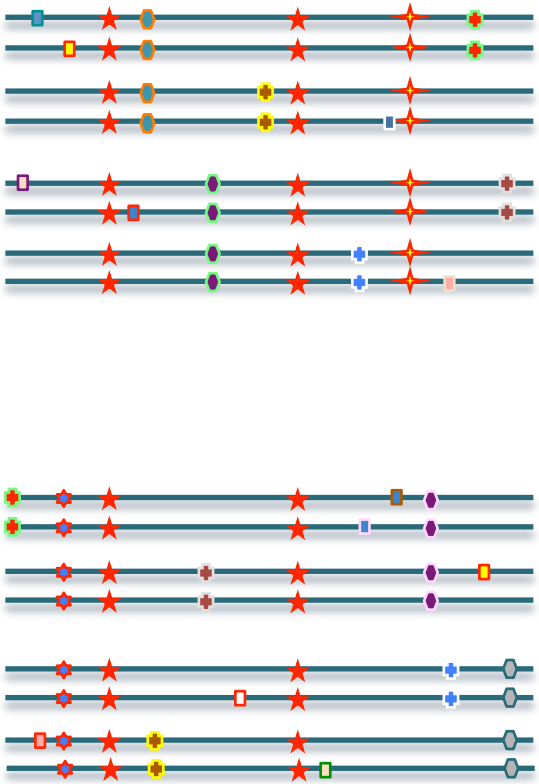
- Li and Stephens 2003

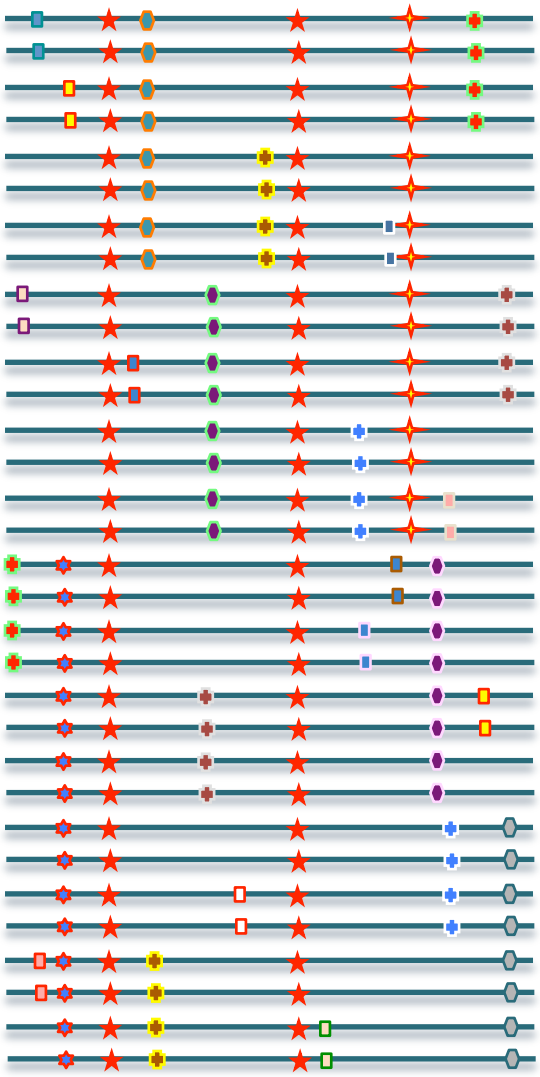
Given m reference haplotypes, and a target sample,
Find the most likely path of haplotype pairs
m^2 states, m^4 transitions per position



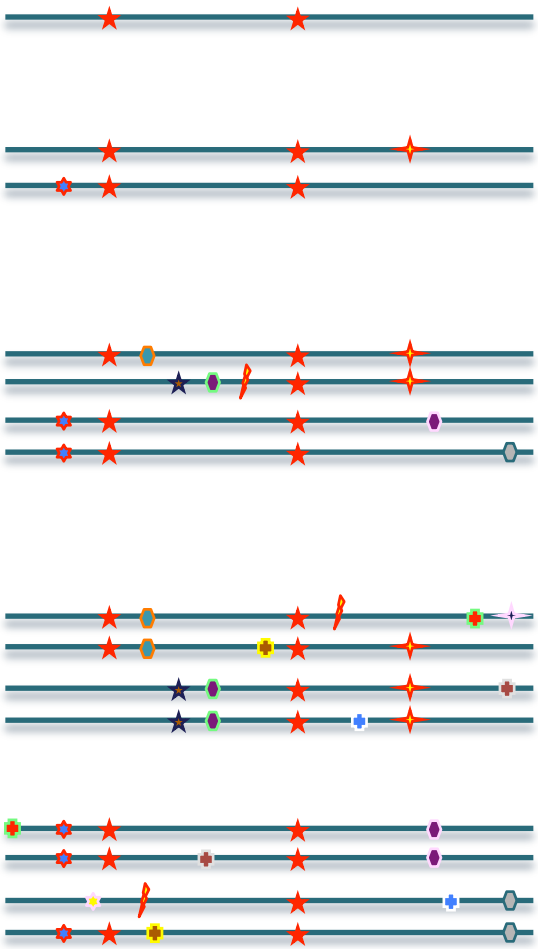
Evolution of a local haplotype

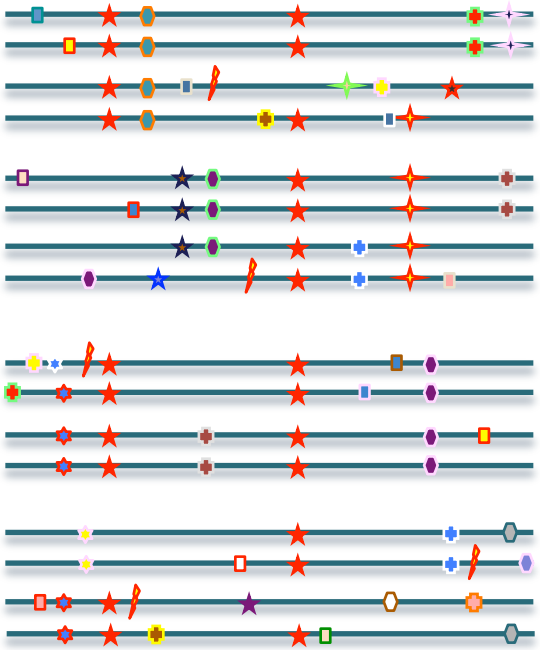


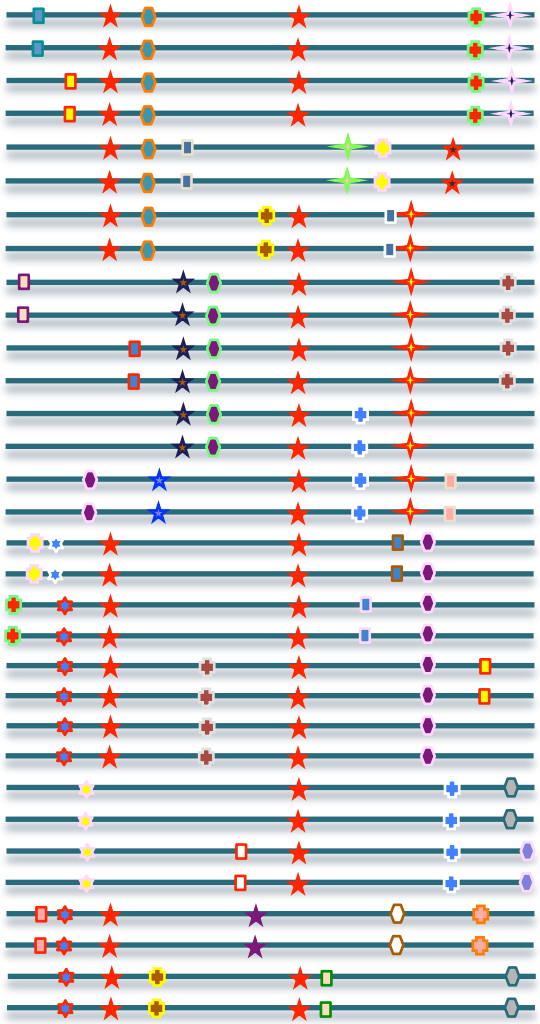




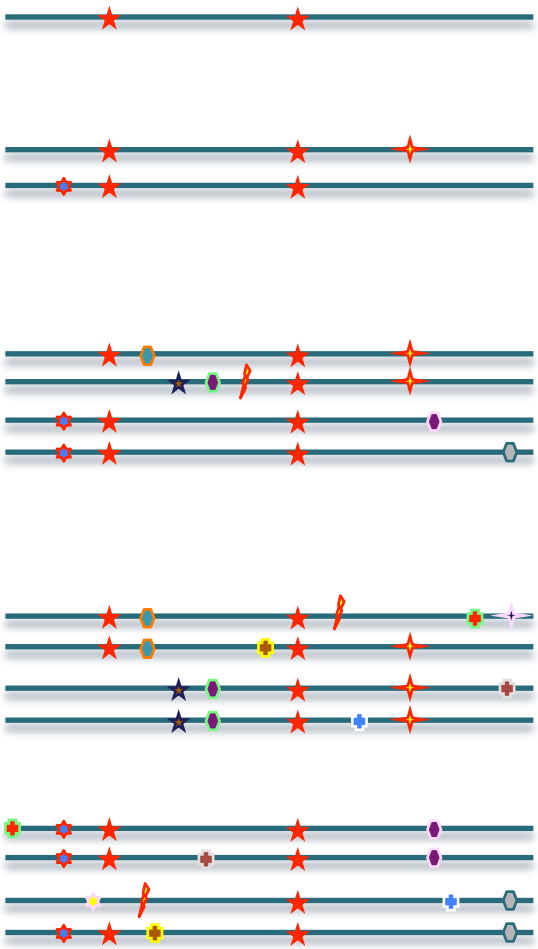
Evolution of haplotypes, recombination

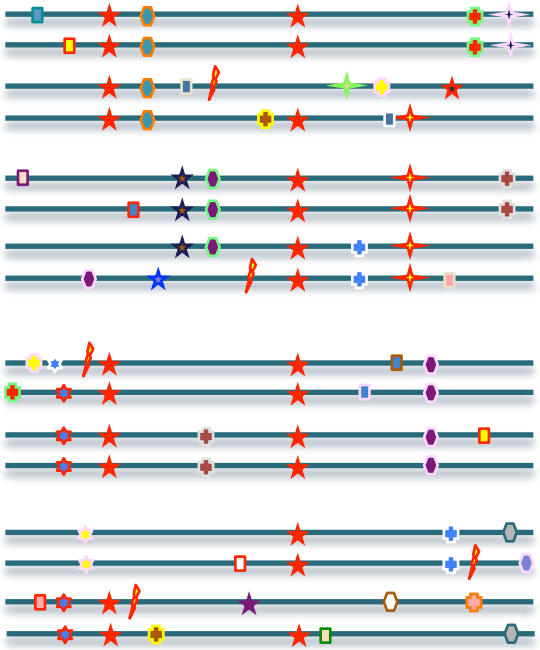


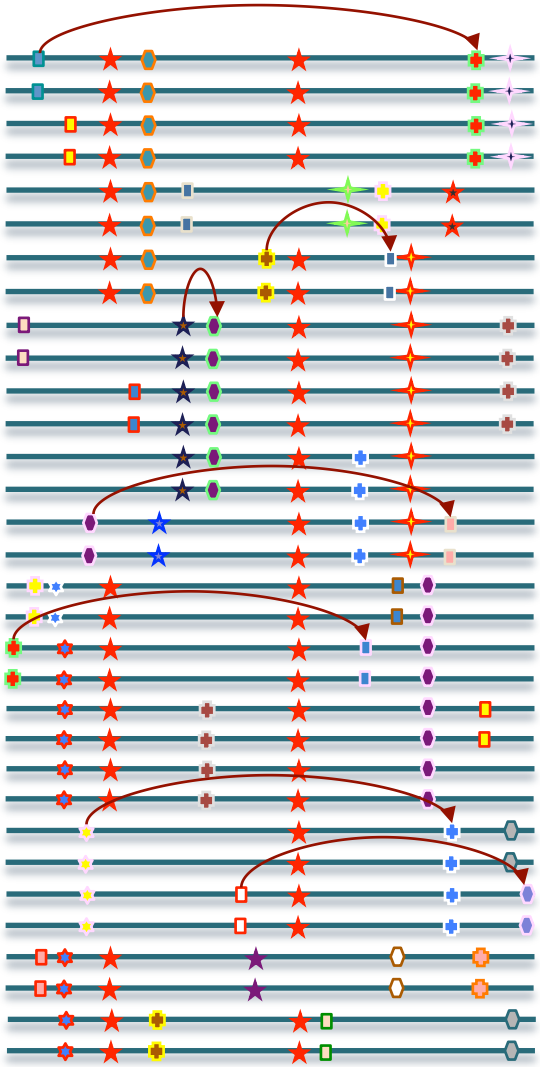




Evolution of haplotypes, recombination





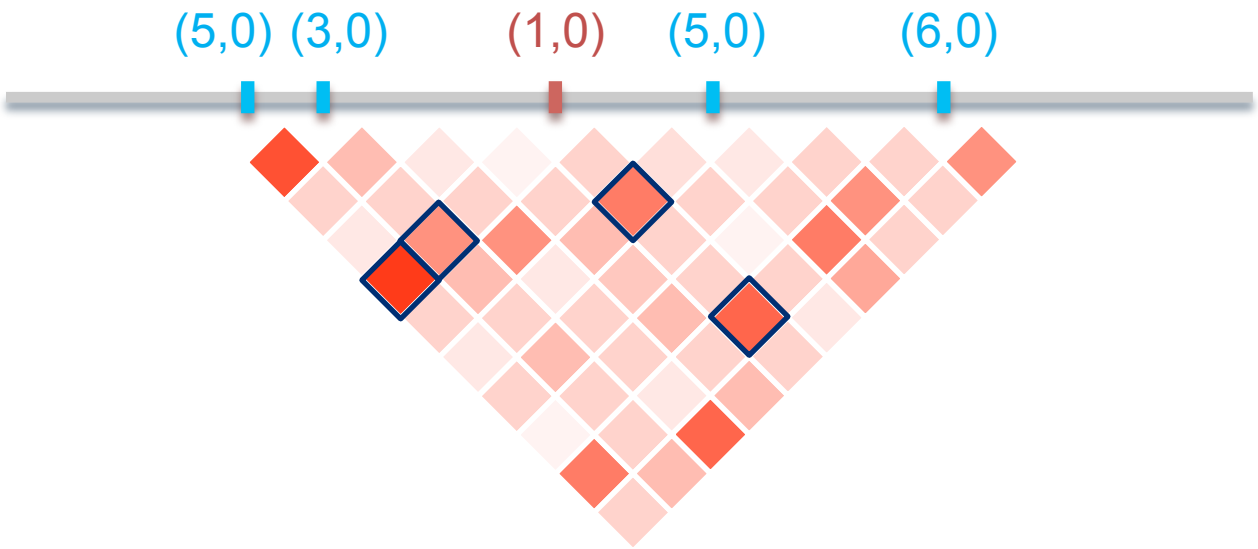


Informative Neighbors

target SNP

k-"nearest" neighbors

in terms of linkage disequilibrium



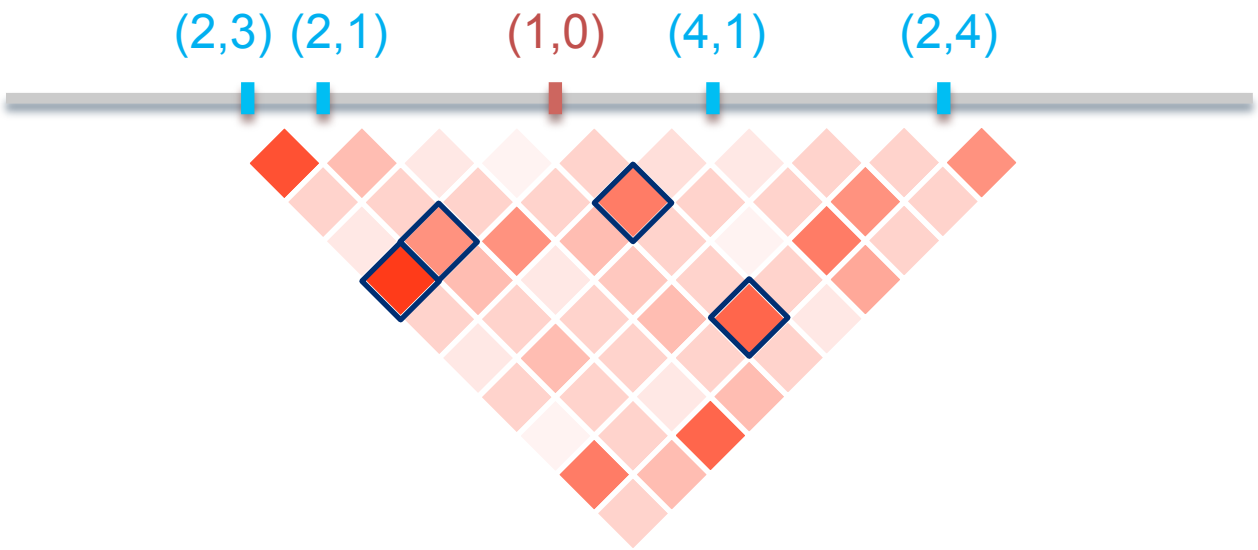
$(R_{ref}, R_{alt}) = \Sigma_{\{target,\ nbrs\}}(r_{ref}, r_{alt}) = (20, 0)$

Informative Neighbors

target SNP

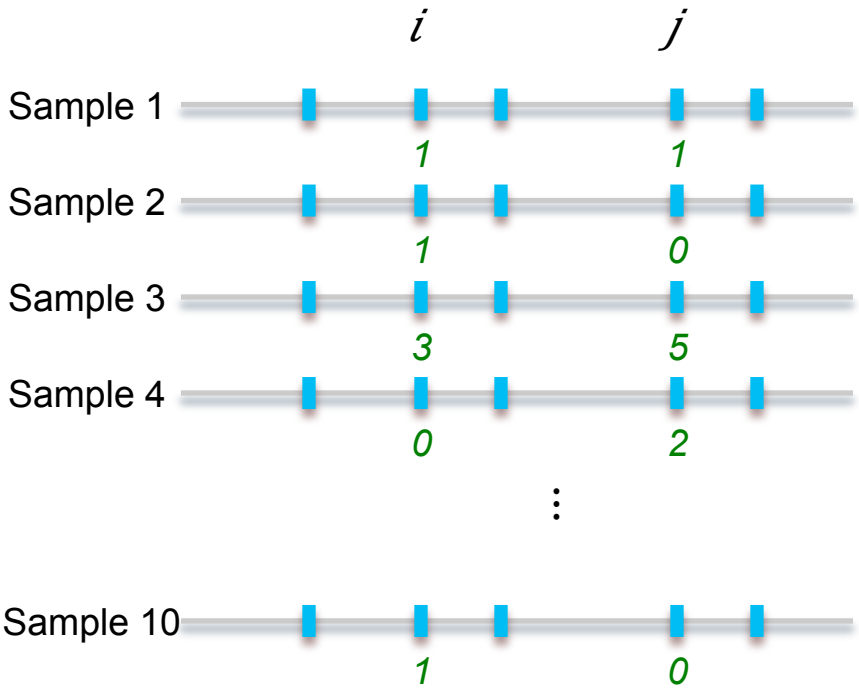
k-"nearest" neighbors

in terms of linkage disequilibrium



$(R_{ref}, R_{alt}) = \Sigma_{\{target,\ nbrs\}}(r_{ref}, r_{alt}) = (11, 9)$

How to pick k nearest neighbors fast



Correlation Coefficient:

$r^2 = (p_{AB} - p_A p_B)^2 / p_A p_B p_a p_b$

Caveat: need **genotyping**, **phasing**

Let

S_i = { samples covering minor allele }

S_i' = { read counts of minor allele }

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

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

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

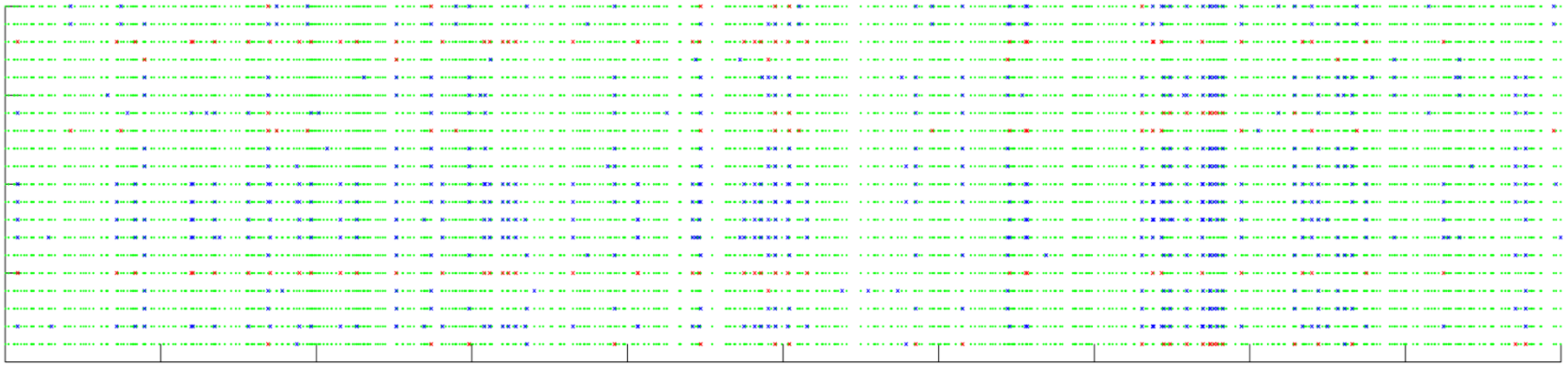
$S_j' = \{1, 3, 3, 3, 3, 3, 4, 4\}$

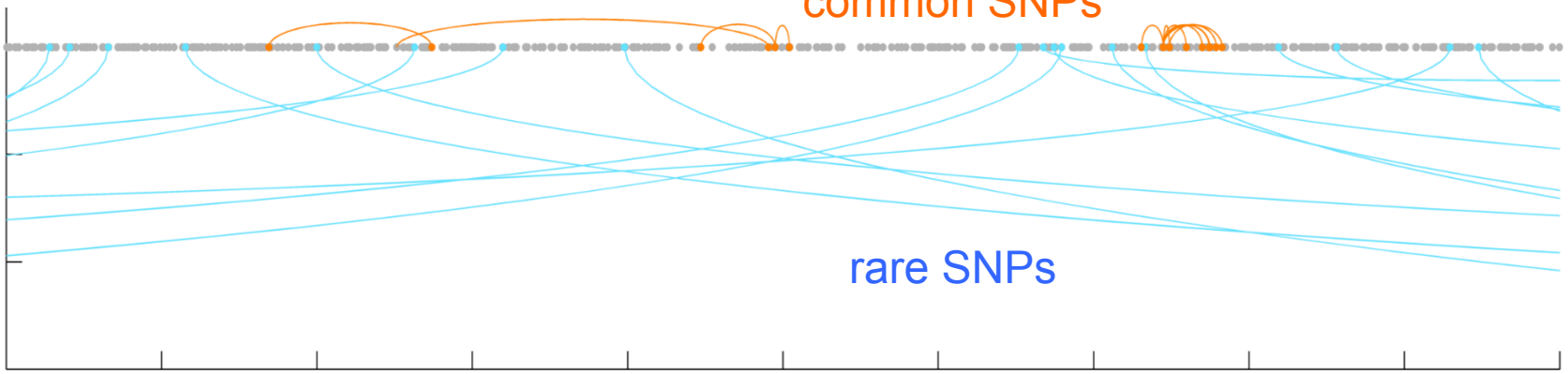
$Sim_1(i, j) = (S_i \cap S_j) / (S_i \cup S_j)$

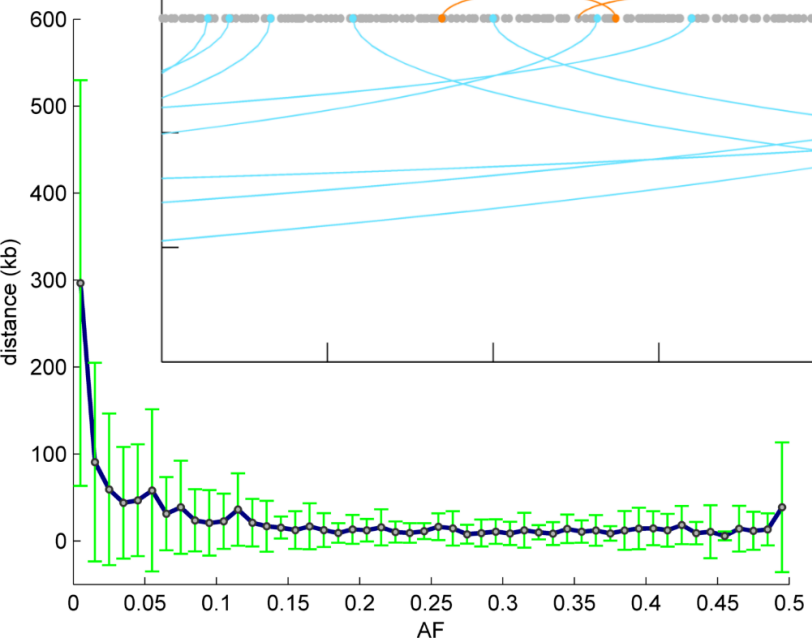
$Sim_2(i, j) = (S_i' \cap S_j') / (S_i' \cup S_j')$

$Sim_3(i, j) = ((S_i' \cap S_j') / (S_i' \cup S_j'))^2$

Genetic distance between NNs



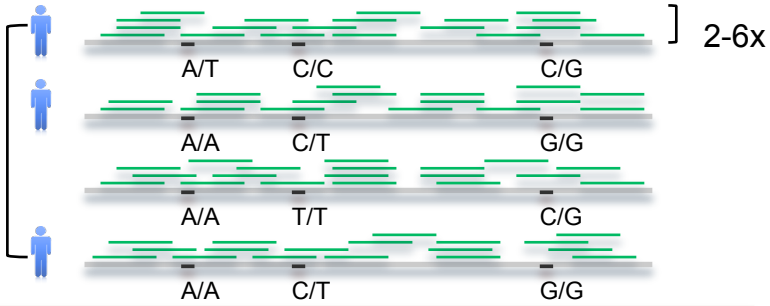




Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



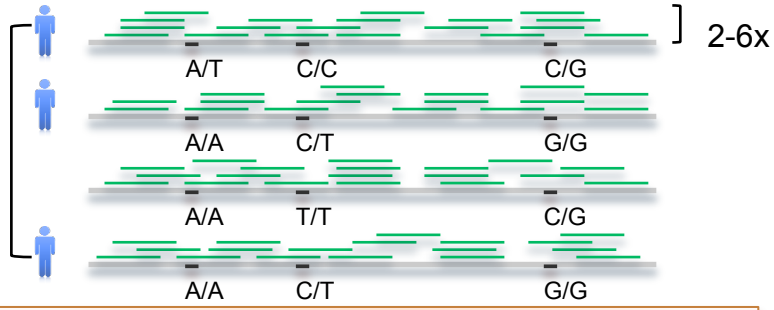
$G_1, \ldots, G_N$; $G_i = g_{i1} \ldots g_{in}$; $g_{ij} \in \{0, 1, 2\}$

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

Overview of Reveel

Reveel:

1. **Identify candidate polymorphic sites**
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



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

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

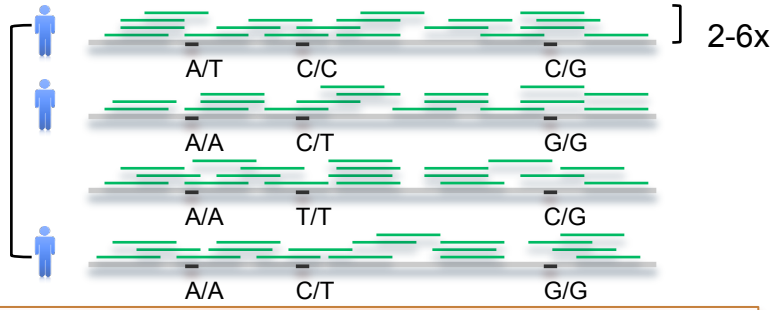
Candidate Polymorphic site

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

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. **Initialize $G^{(0)}$**
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



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

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

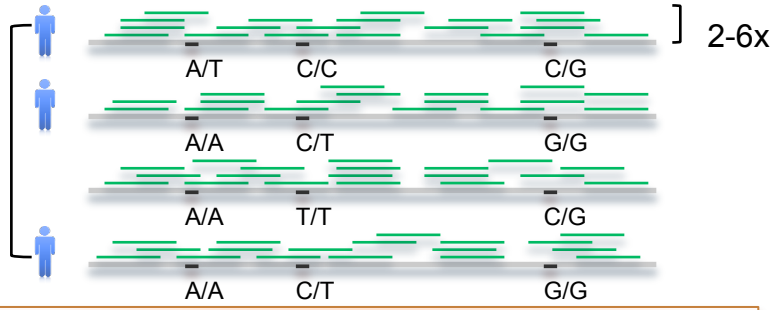
At each position j,

Use sum of read counts at j and its nearest neighbors

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. **Summarization/Maximization**
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



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

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

$p^{(n+1)}_{ijg} = P(g_{ij} = g \mid G^{(n)}, reads)$
$\sim P(g_{ij} = g \mid g_{kNN}, reads)$
$= P(reads \mid g_{ij} = g)\ P(g_{ij} = g \mid g_{kNN})$

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

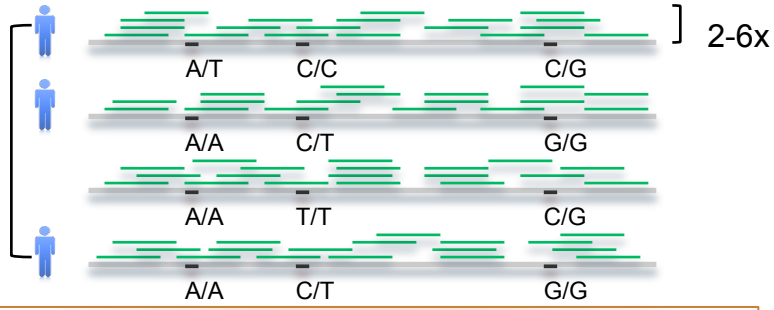
Let C_0, C_1, C_2 = # samples matching i in kNN, with j^{th} genotype pos'n = 0, 1, 2

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

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. **Recalculate k nearest neighbors**
   - **Approximate Correlation Coefficient (Schaid 2004)**
6. **Summarization/Maximization**
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization



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

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

Correlation Coefficient:
$r^2 = (p_{AB} - p_A p_B)^2 / p_A p_B p_a p_b$

Caveat: need **genotyping**, **phasing**

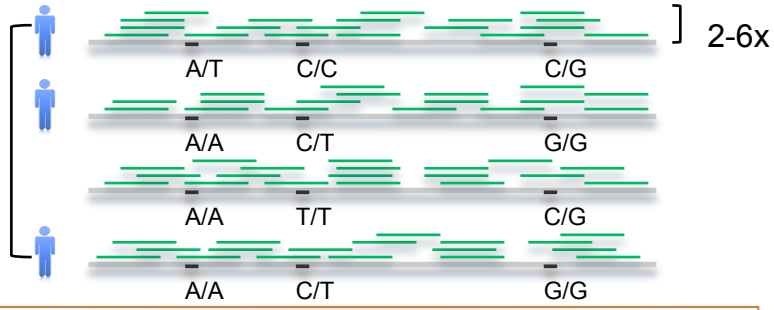
Schaid 2004:
$D = 1/N\ (2N_{AABB} + N_{AABb} + N_{AaBB} + ½N_{AaBb}) - 2p_A p_B$

A faster alternative:
$D = ½\ Sim_1(i, j) + p_A p_B (p_A + p_B - ½\ p_A p_B - 2)$

Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. **Recalculate k nearest neighbors**
   - **Approximate CC, Entropy**
8. **Summarization/Maximization**



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

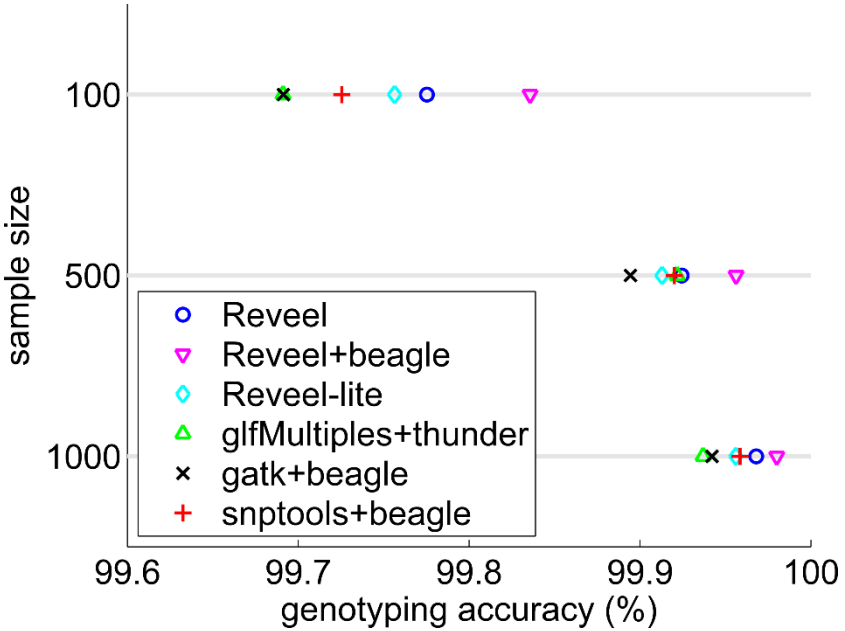
$P_1, \ldots, P_N$; P_i : [$p_{ijg} = Prob(g_{ij} = g \mid data)$]

Identify the sites at which decision trees formed by kNNs have high entropy

Reduce entropy by replacing one or more kNNs

Simulations

- Simulated data set: ~ low-coverage 1KGP
  - 2,535 samples, 1-Mbp region: Chr 20 (43-44 Mbp) of GRCh37
  - Using COSI to simulate variations
  - Reads same positions, lengths, qualities as reads in 1KGP
  - Injecting sequencing base errors, mapping with BWA



Sample size	100	500	1000	2535
Reveel	1.8	14.6	47.4	273
Reveel+Beagle	3.1	25.3	71.2	526
Reveel-lite	1.5	7.8	21.7	145
SNPTools+Beagle	8.2	217	1089	>5days
GATK+Beagle	13.4	388	1806	>5days
glfMultiples+Thunder	307	2736	6120	~15 days

Genotyping Accuracy

Computation Time (mins)

Performance on 1000 Genomes Data

population	# of samples in 1KGP	# of samples in HapMap3	population	# of samples in 1KGP	# of samples in HapMap3
ACB	96	0	**ASW**	66	**50**
BEB	86	0	CDX	99	0
CEU	99	**90**	**CHB**	103	**94**
CHS	108	0	CLM	94	0
ESN	99	0	FIN	99	0
GBR	92	0	**GIH**	106	**93**
GWD	113	0	IBS	107	0
ITU	103	0	**JPT**	104	**97**
KHV	101	0	**LWK**	101	**90**
MSL	85	0	**MXL**	67	**56**
PEL	86	0	PJL	96	0
PUR	105	0	STU	103	0
TSI	108	**96**	**YRI**	109	**103**

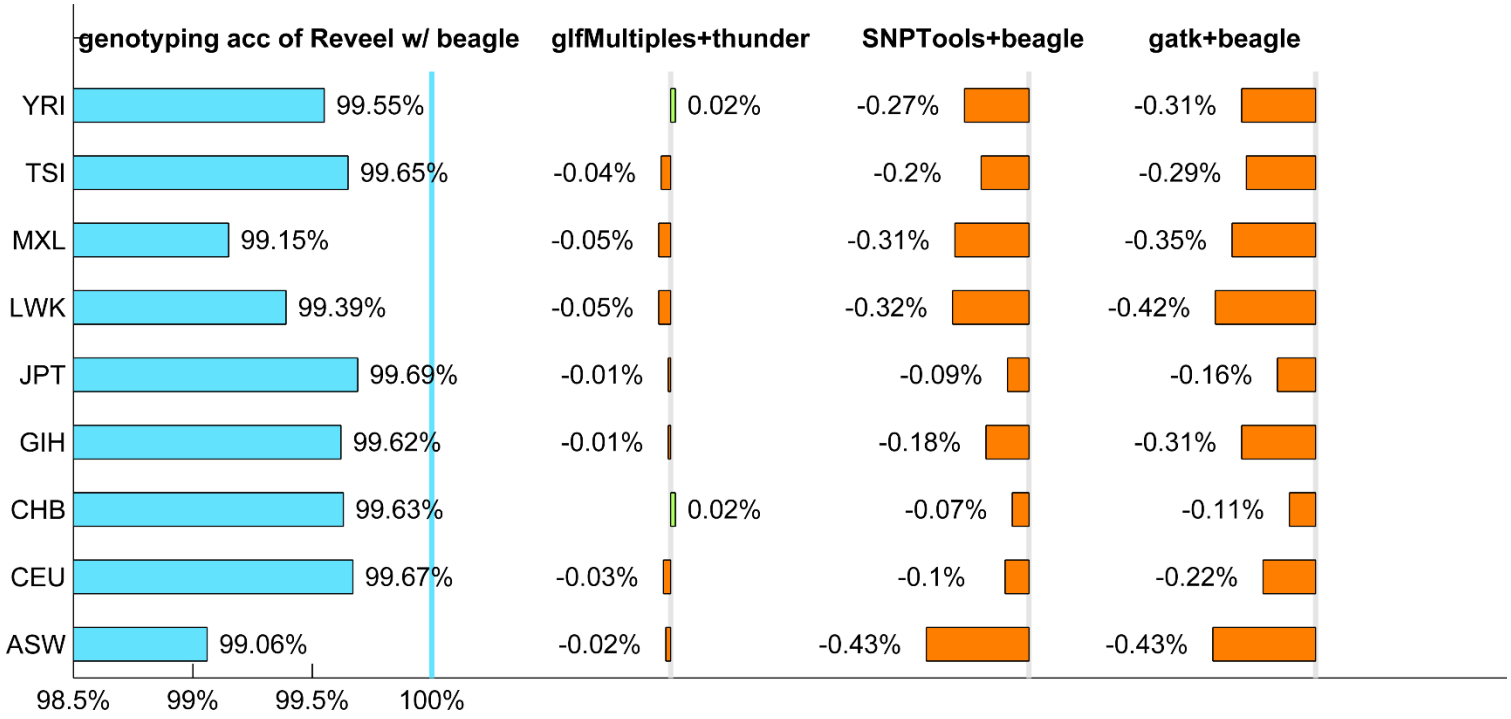
Compared on a 5-Mbp region on chromosome 20 (43-48 Mbp)

Performance on 1000 Genomes Data

HapMap 3 benchmark

769 invididuals from 9 populations

AF ≥ 5%,	2686 SNPs
5% > AF ≥ 1%,	368 SNPs
AF < 1%,	32 SNPs



Performance on 1000 Genomes Data

- SNP discovery
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: Complete Genomics data

Method	false positive rate	sensitivity
Reveel	0.021%	95.80%
Reveel+Beagle	0.020%	95.92%
Reveel-lite	0.021%	95.80%
GATK+Beagle	0.035%	95.62%
glfMultiples+Thunder	0.037%	95.80%
SNPTools+Beagle	0.035%	95.66%
GotCloud (w/ filters)	0.007%	91.29%
Integrated (w/ filters)	0.011%	95.89%

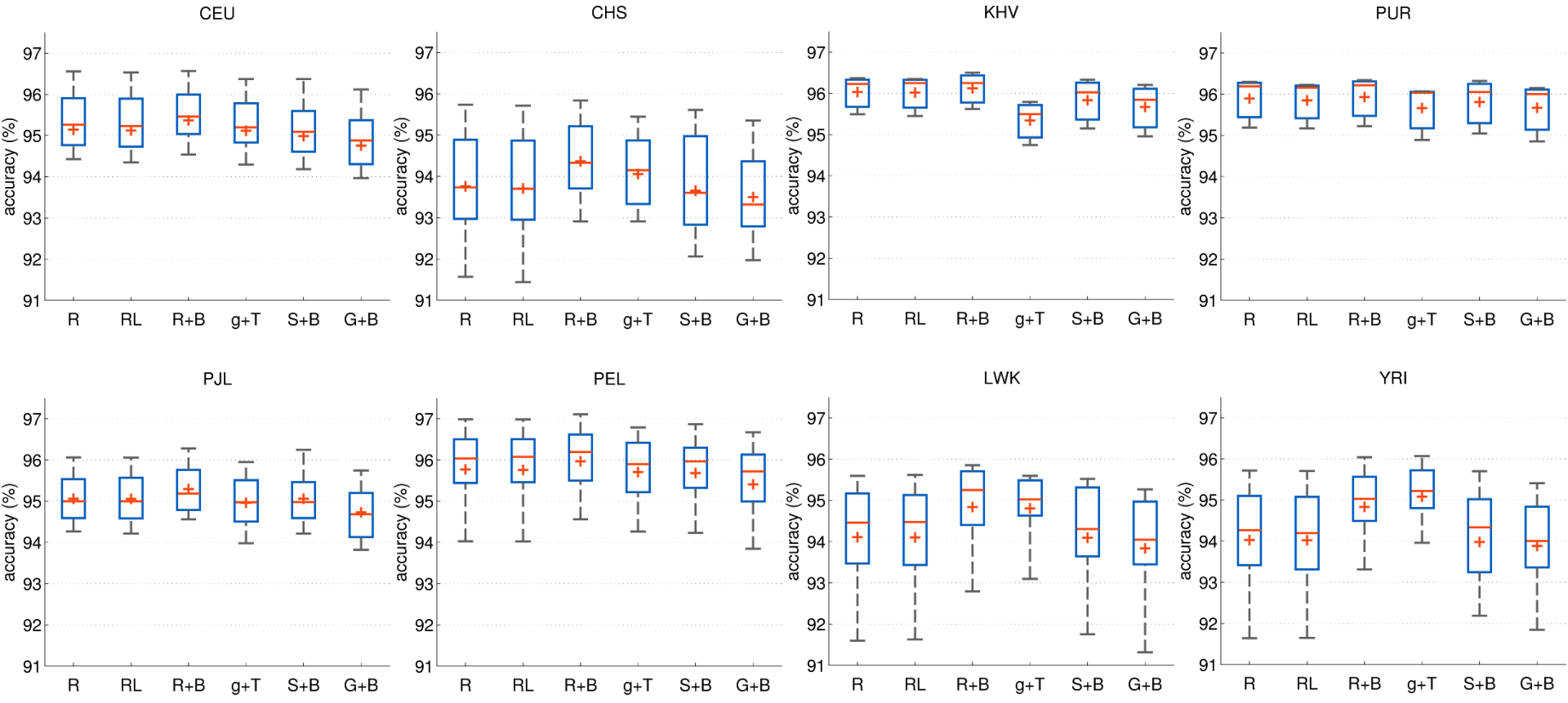
Performance on 1KGP Trios

- SNP discovery
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: 1KGP Pilot2 Trios

Method	false positive rate	sensitivity
Reveel	0.031%	97.06%
Reveel+Beagle	0.031%	97.53%
Reveel-lite	0.031%	97.06%
GATK+Beagle	0.040%	97.38%
glfMultiples+Thunder	0.048%	98.17%
SNPTools+Beagle	0.044%	96.81%
GotCloud (w/ filters)	0.011%	91.46%
Integrated (w/ filters)	0.023%	98.32%

Performance on 1000 Genomes Data

- Genotyping
  - Benchmark: Complete Genomics samples



R: Reveel; RL: Reveel-lite; R+B: Reveel+Beagle; g+T: glfMultiples+Thunder; S+B: SNPTools+Beagle; G+B: GATK+Beagle