Signals of recent positive selection in a worldwide sample of human populations

Pickrell, Coop, Novembre,..., Pritchard

Department of Human Genetics
The University of Chicago

Journal Club presentation
Maido Remm
29.09.2009

Questions asked in Pickrell et al.

- Main question: which regions in human genome are under **positive selection in one or another population**.

So far only HapMap and Perlegen datasets have been studied: one European, one African and one or two East Asian populations.

Positive selection at different timescale

Sabeti et al. Science 312, 1614 (2006)

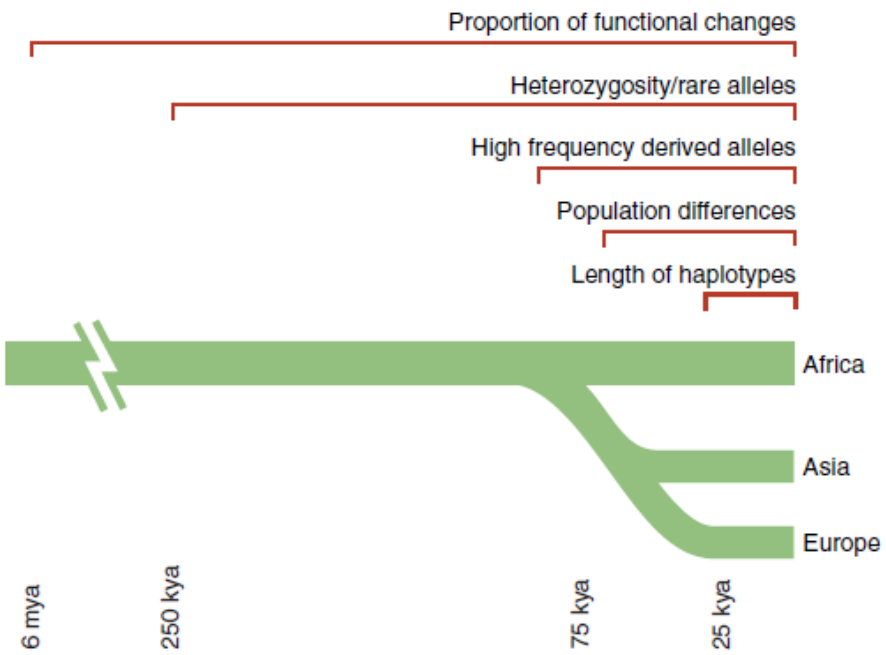


Fig. 1. Time scales for the signatures of selection. The five signatures of selection persist over varying time scales. A rough estimate is shown of how long each is useful for detecting selection in humans.

Example of frequent functional changes:

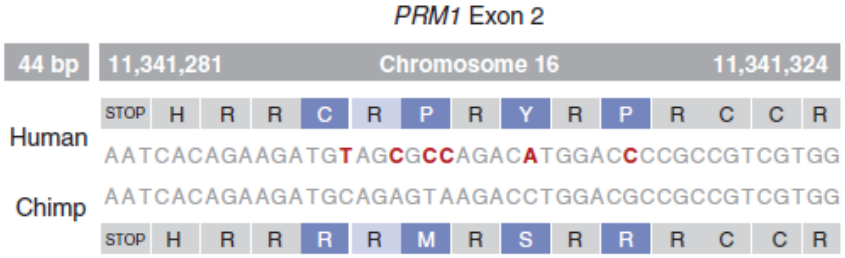


Fig. 2. Excess of function-altering mutations in PRM1 exon 2. The PRM1 gene exon 2 contains six differences between humans and chimpanzees, five of which alter amino acids (7, 8).

Positive selection at different timescale

Sabeti et al. Science 312, 1614 (2006)

Example of high frequency of derived alleles:

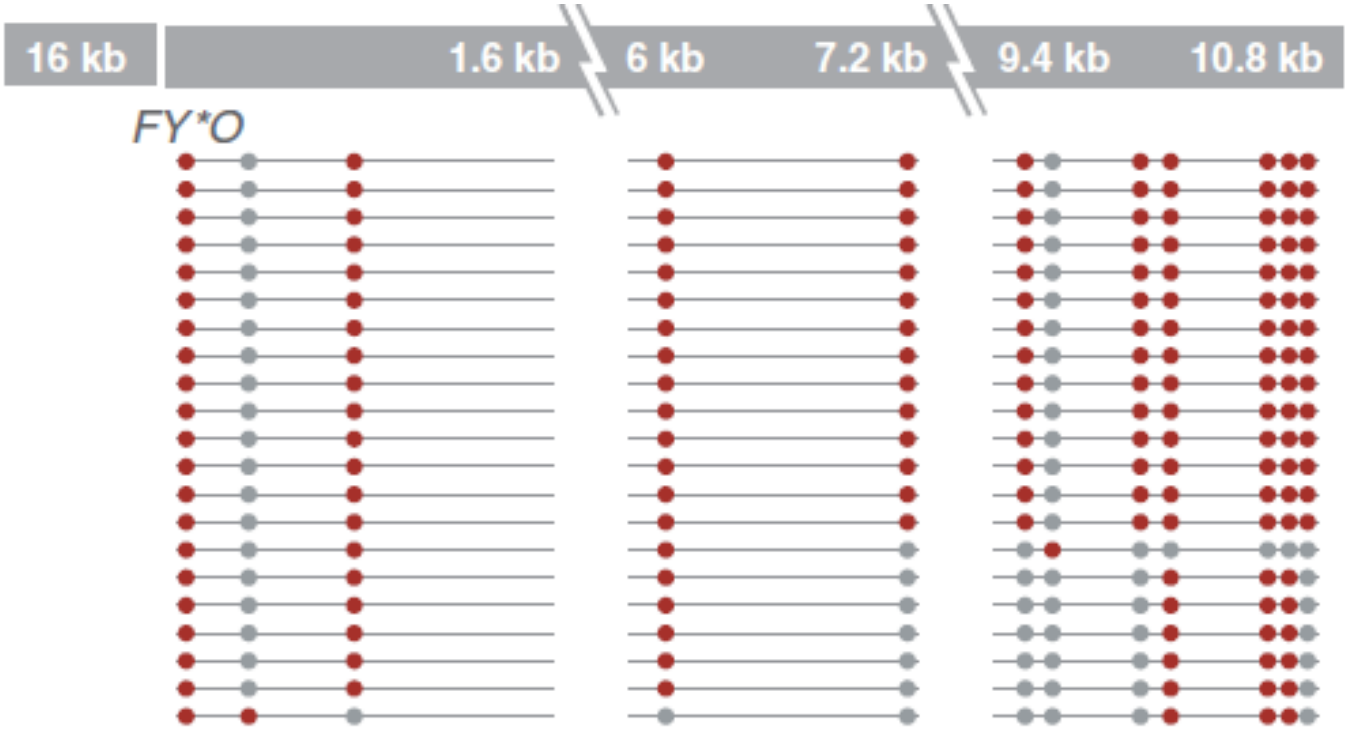


Fig. 4. Excess of high-frequency derived alleles at the Duffy red cell antigen (FY) gene (34). The 10-kb region near the gene has far greater prevalence of derived alleles (represented by red dots) than of ancestral alleles (represented by gray dots).

Positive selection at different timescale

Sabeti et al. Science 312, 1614 (2006)

Example of population differences and haplotype length tests:

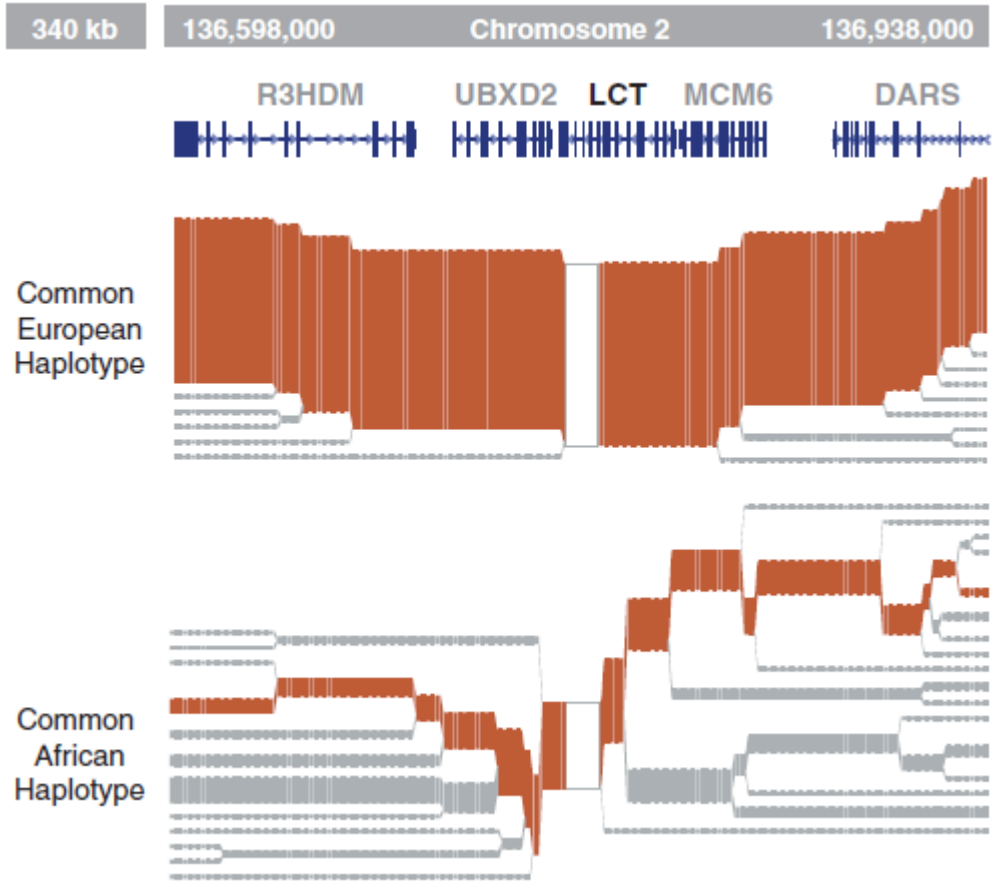


Fig. 6. Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (E77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (6).

Positive selection and it's measures

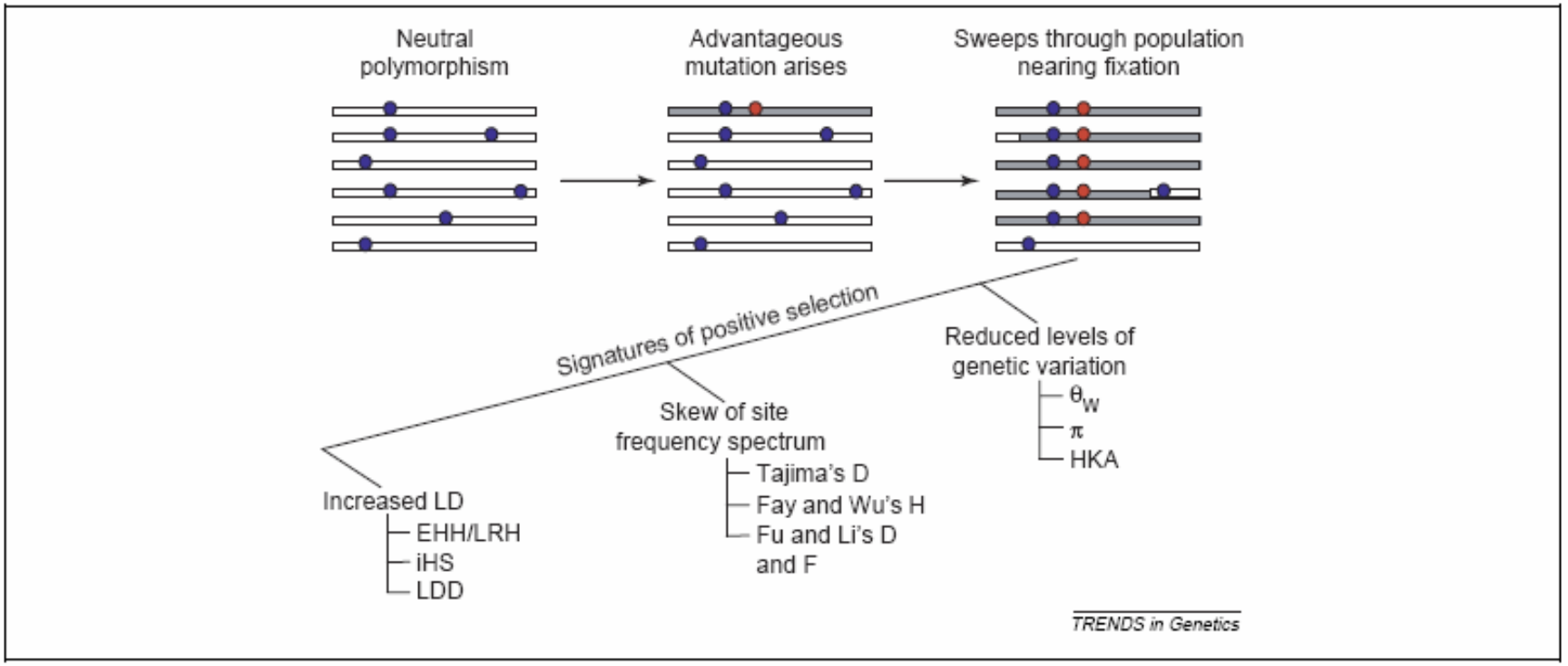


Figure 1. Signatures of positive selection. On the left, patterns of neutral polymorphism (denoted as blue circles) are shown for a sample of six haplotypes. A new advantageous mutation (indicated by the red circle) arises on a specific haplotype (middle panel highlighted in gray). As the advantageous allele increases in frequency it drags along linked neutral polymorphisms. On the right, an incomplete selective sweep is shown such that the advantageous allele has not yet reached fixation. This process perturbs patterns of genetic variation relative to neutral expectations and imparts signatures such as reduced levels of genetic variation, a skew in the site frequency spectrum (also referred to as allele frequency distribution), and increased levels of LD. Recombination between haplotypes carrying and not carrying the advantageous allele delimit the region over which the signature of selection extends. Commonly used summary statistics that have been proposed to test for these signatures are also indicated and described in more detail in Box 1. Note that the relative magnitude of these signatures of positive selection depend on many parameters such as when the advantageous allele arose, the strength of selection, whether the sweep is ongoing or has reached fixation, the amount of time that has elapsed since fixation, and local rates of recombination and mutation.

Biswas and Akey. Trends in Genetics 22:437 (2006)

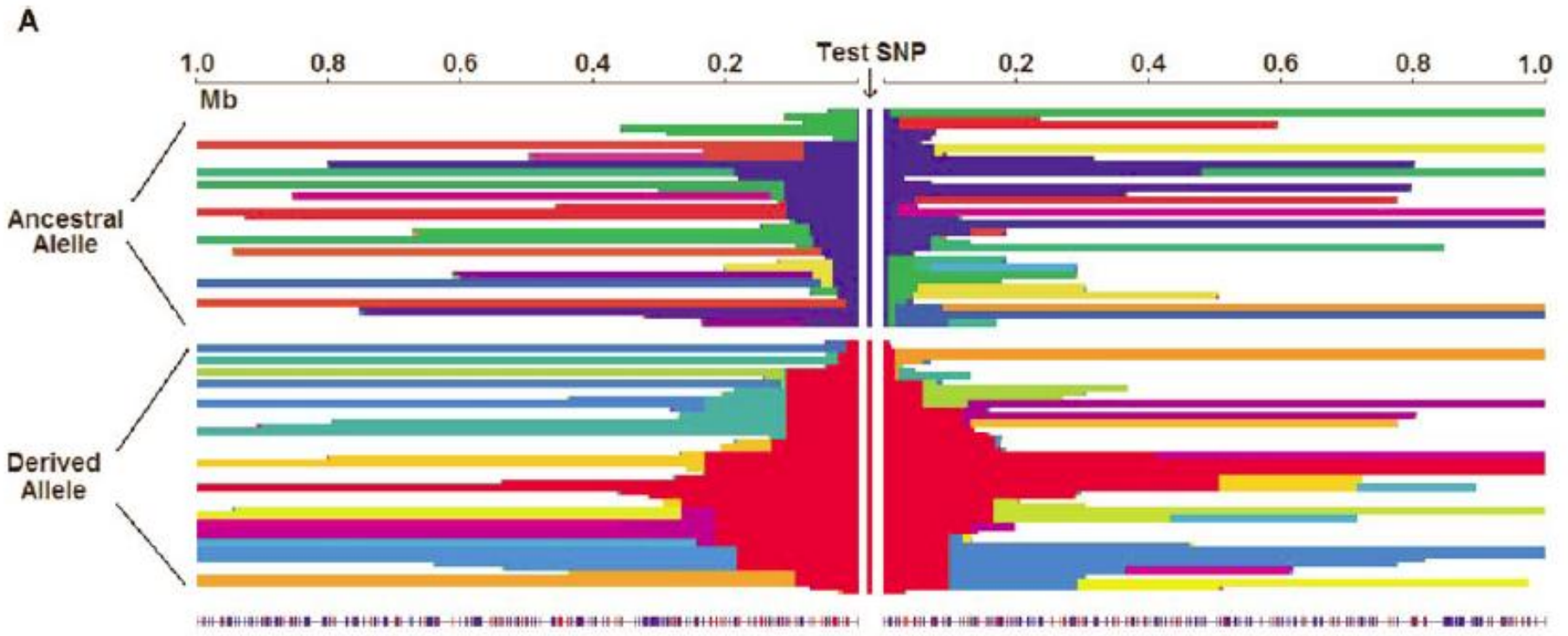
What methods were used in this paper?

- ***iHS*** (integrated Haplotype Score)
- **XP-EHH** (Cross-population Extended Haplotype Homozygosity)
- **Heterozygosity -** fraction of heterozygote individuals in population. Low values may be sign of recent selection.
- **F_{st}** - measures proportion of variance between <u>two populations</u>

F_{st} = (average_number_of_diff_**between**_pop - average_number_of_diff_**within**_pop) / average_number_of_diff_**between**_pop

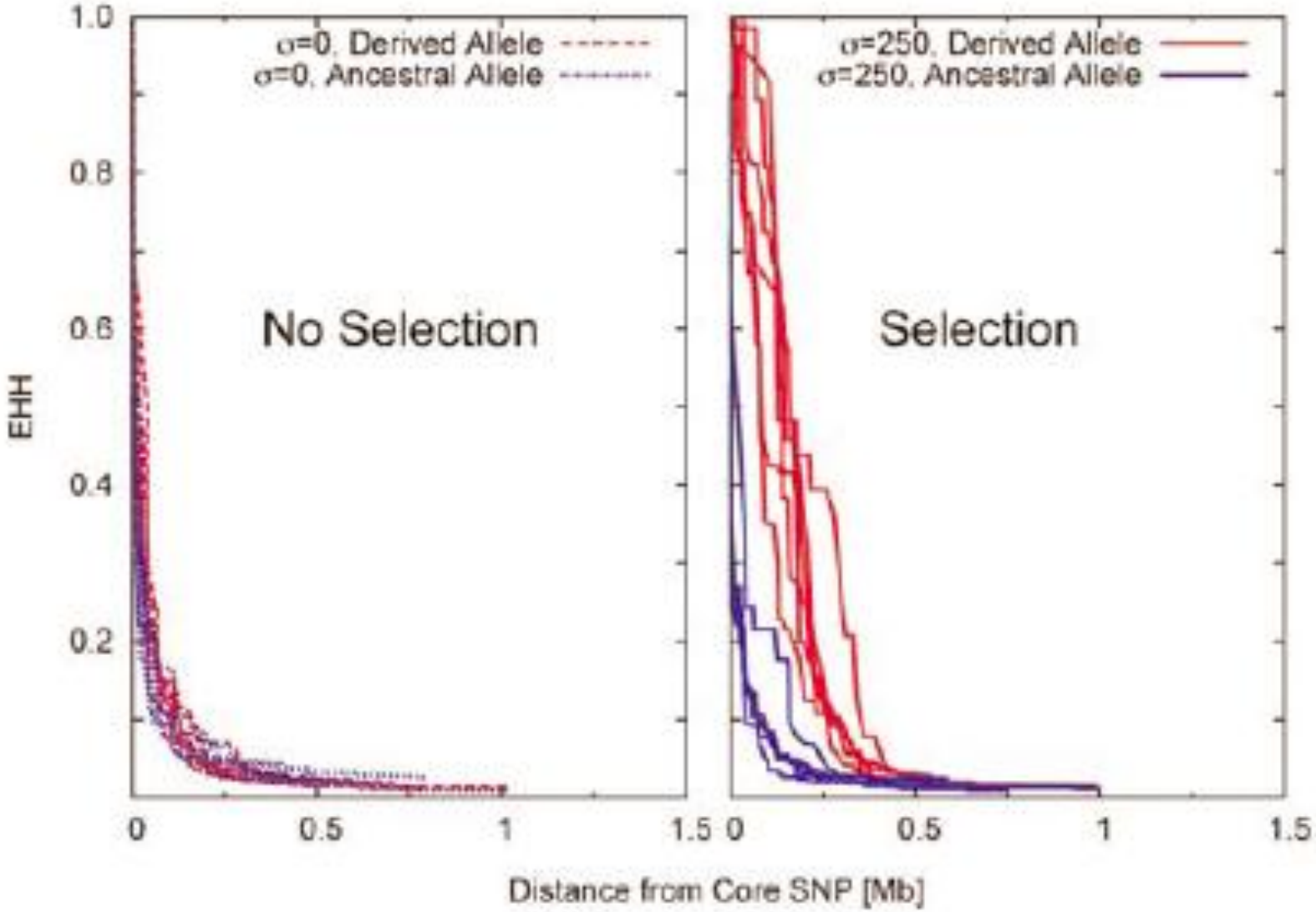
F_{st} 0 means populations are very similar, 1 means populations are very different.

Assumption:
selected haplotypes are longer than expected



Voight et al. PLos Biology 4:e72 (2006)

Simulated example



Voight et al. PLos Biology 4:e72 (2006)

What has been done before?

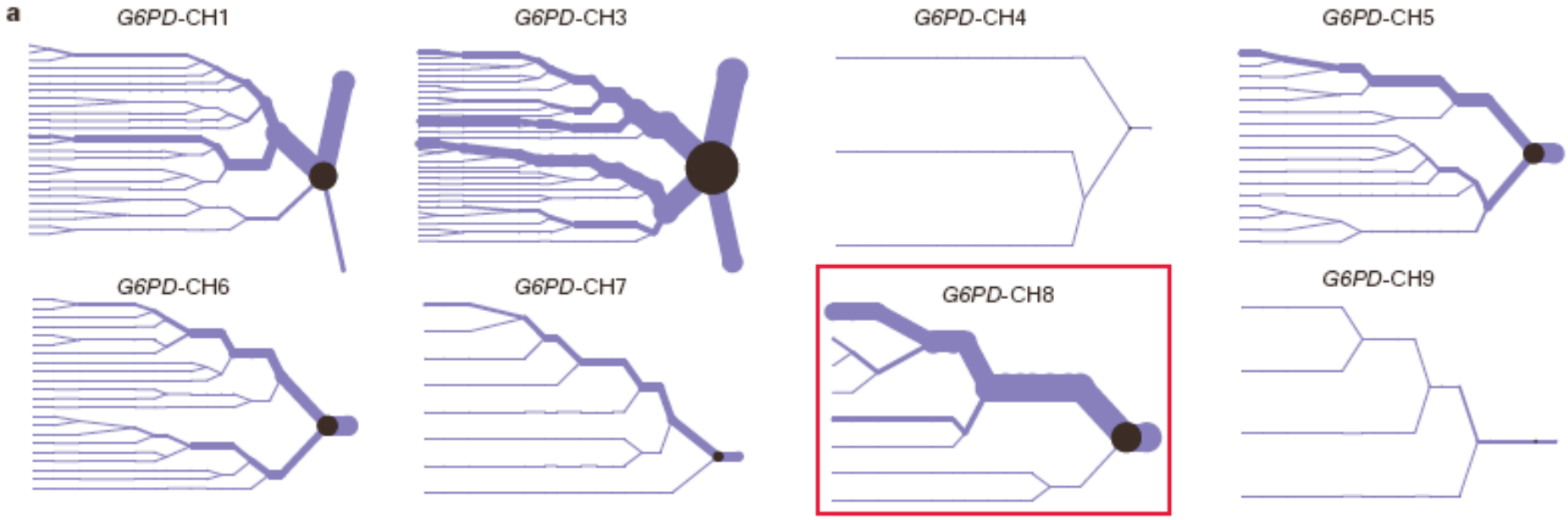
- **Sabeti 2002, Science:**
- **Used EHH method: Extended haplotype heterozygosity**
- The key characteristic of positive selection is that it causes an unusually rapid rise in allele frequency, occurring over a short enough time that recombination does not substantially break down the haplotype on which the selected mutation occurs. A signature of positive natural selection is thus an allele having unusually long-range LD given its population frequency.

EHH example

Example of EHH:

GP6D – confers resistance to malaria

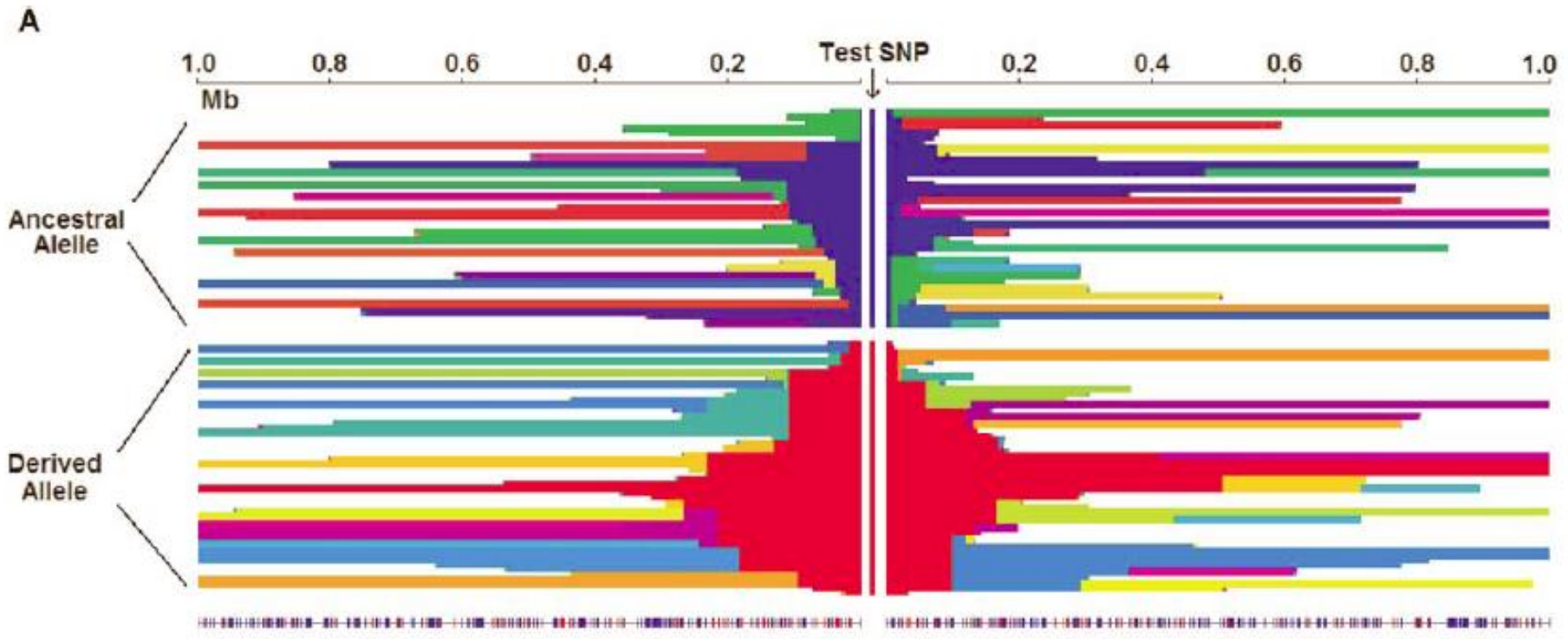
If one haplotype is much more frequent (thicker) than others
=> positive selection.



Sabeti et al. Nature 419, 832 (2002)

XP-EHH

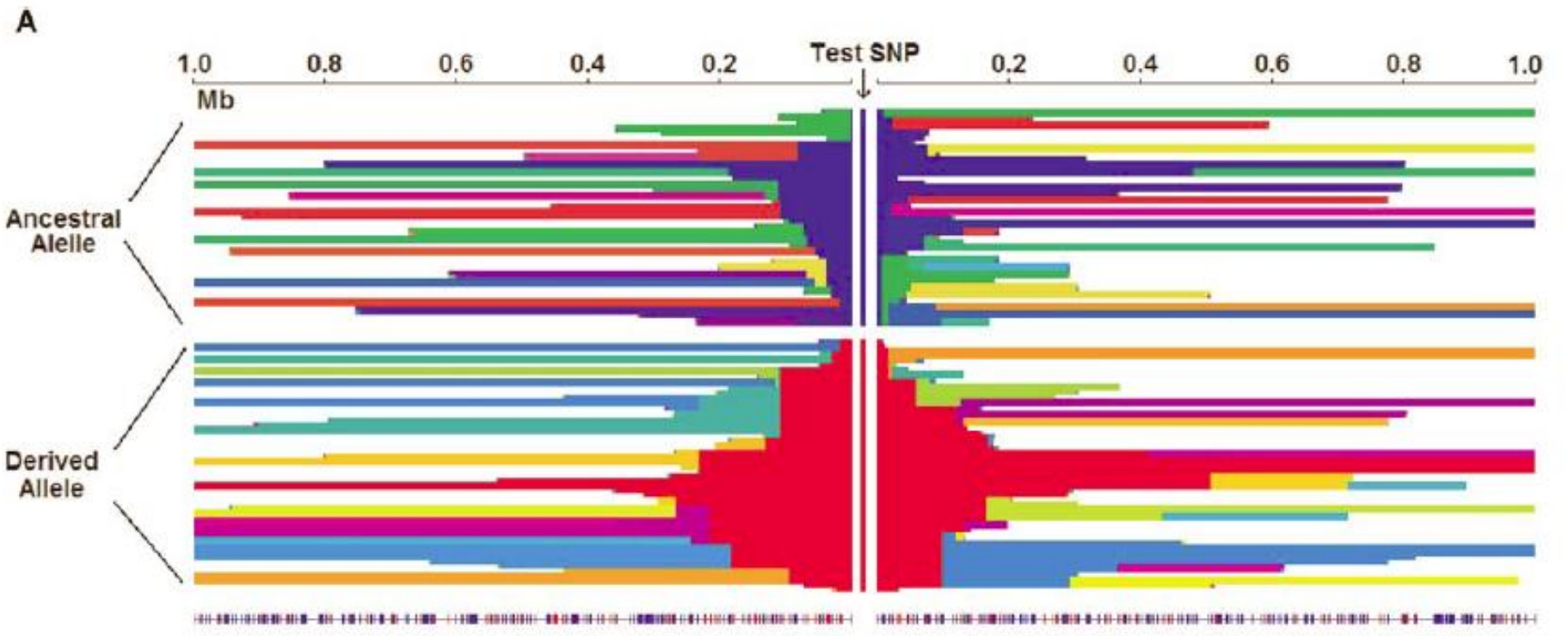
- XP-EHH: Expected means expected for same haplotypes in **another population** at the same locus. XP-EHH was first described in Sabeti et al. Nature 449: 913 (2007).



Picture from Voight et al. PLos Biology 4:e72 (2006)

iHS

- iHS: Expected means **other haplotypes** in the **same population** at the same locus



Voight et al. PLos Biology 4:e72 (2006)

iHS

- iHS: Expected means **other haplotypes** in the **same population** at the same locus

$$\text{unstandardized } iHS = \ln\left(iHH_A \,/\, iHH_D\right)$$

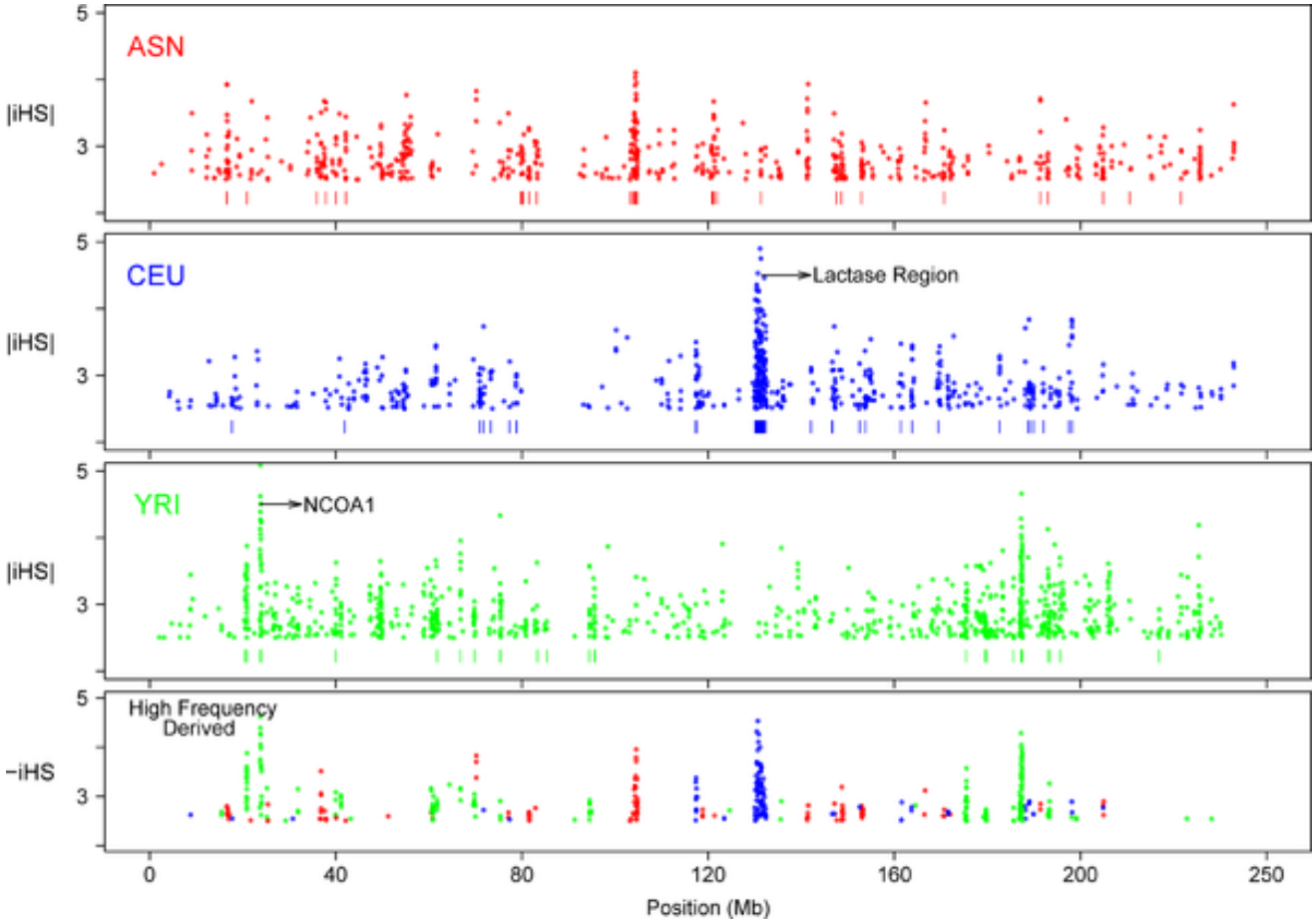
where
iHH_A is integrated Haplotype Homozygosity for ancestral allele
iHH_D is integrated Haplotype Homozygosity for derived allele

Integration is done over both directions from SNP

Low frequency alleles tend to have longer haplotypes. To avoid over-representation of low-freq alleles, the iHS is further **standardized** w.r.t. mean and SD of all alleles in the genome with similar allele frequency.

iHS example

- Extreme values of iHS correlate with known regions of positive selection



Large negative values indicate unusually long haplotypes carrying the derived allele.

Nevertheless the absolute value of *iHS* is frequently used!

Voight et al. PLos Biology 4:e72 (2006)

What are these tests able to detect?

- iHS detects recent events of positive selection where one haplotype have recently increased but ancestral alleles are still present. Does NOT need reference population.

- XP-EHH detects well the regions that are nearly fixed, however it needs reference population

- Heterozygosity: reduced heterozygosity indicates reduced variation and thus can detect long-term selection, but there might be other reasons for reduced variation.

- F_{st} measures proportion variance. Needs a properly chosen reference population.

The analysis (1):

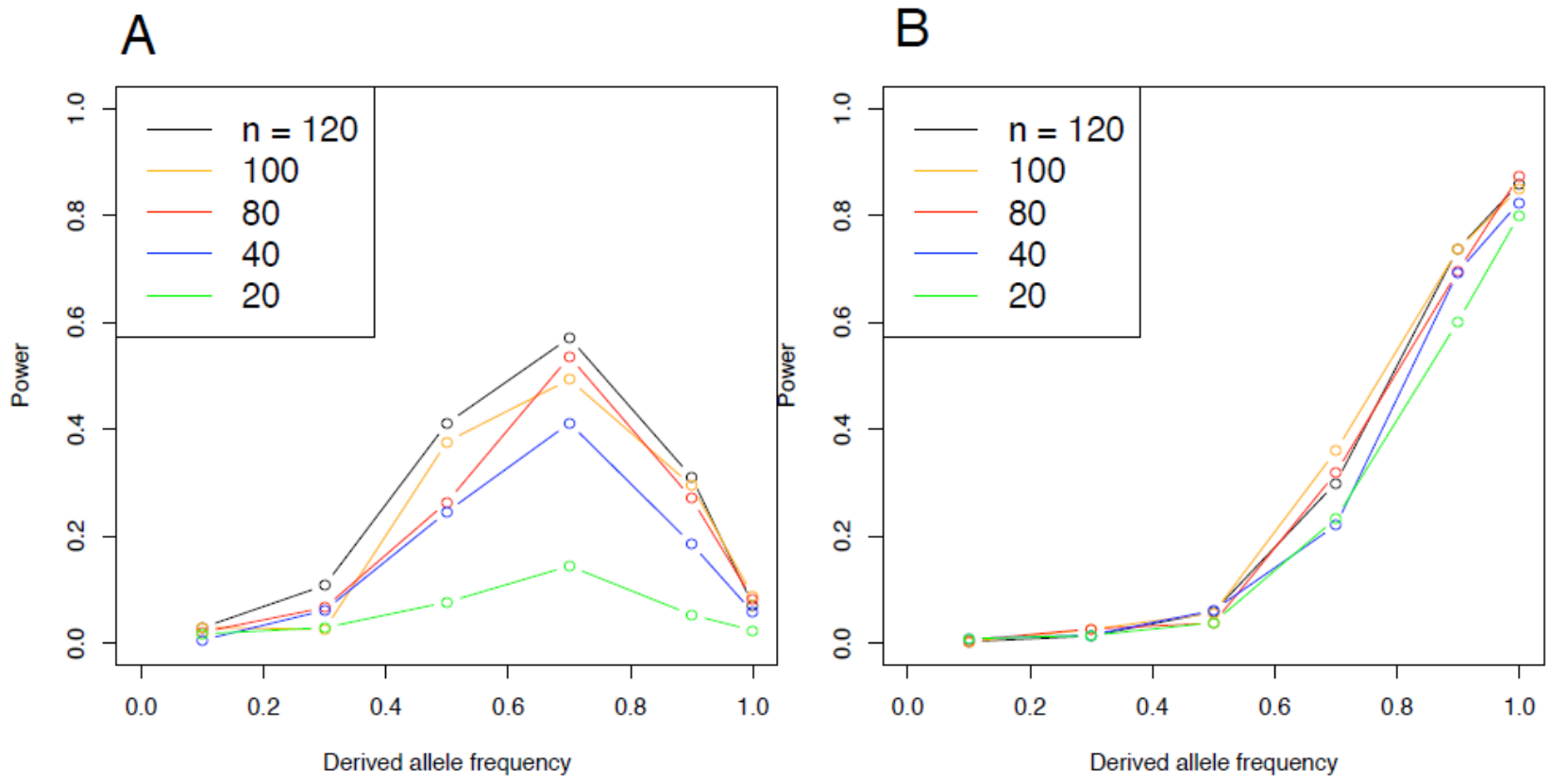


Figure 3: **Power of iHS (A) and XP-EHH (B) in the YRI demography for different sample sizes (in number of chromosomes).** Selected alleles were introduced at a random time with a selection coefficient of 1% and simulated forwards in time. Simulations were binned into six bins according the the final frequency of the selected allele. The type I error was set to 1%, with the threshold determined by neutral simulations.

The analysis (2):

iHS and XP-EHH were calculated for all SNPs.

maximum XP-EHH and the **fraction of extreme** (|iHS| > 2) **iHS scores** was recorded in 200kb non-overlapping windows

For $\mathbf{F_{st}}$, the **maximum value of single SNP** in 100 kb window around previously known SNP was used.

The importance of these values was estimated by genome-wide distribution of scores.

Phasing was done using fastPHASE, with the settings that allow variation in the switch rate between subpopulations.

The analysis (3):

Analysis was done on HGDP populations.
After quality control and removal of related individuals, the HGDP data consist of
657 143 SNPs typed on
938 individuals in 53 populations.

Populations were grouped into 8 groups to achieve larger sample size.

- Bantu-speaking populations,
- Biaka Pygmies,
- Europeans,
- Middle Easterners,
- South Asians,
- East Asians,
- Oceanians and
- Native Americans.

The Mbuti Pygmies and San were dropped from these groups because their large divergence from other African populations.

The analysis (4):

Top-ten genomic regions for each population were calculated and are shown in Fig.1

Some known candidate genes are analyzed in detail:
- pigmentation genes
- genes revealed by recent GWA on common diseases

Some top signals are analyzed in detail:
- C21orf34, a locus of unknown function on chromosome 21
- selection signals in the NRG–ERBB4 pathway

Top 10 signals

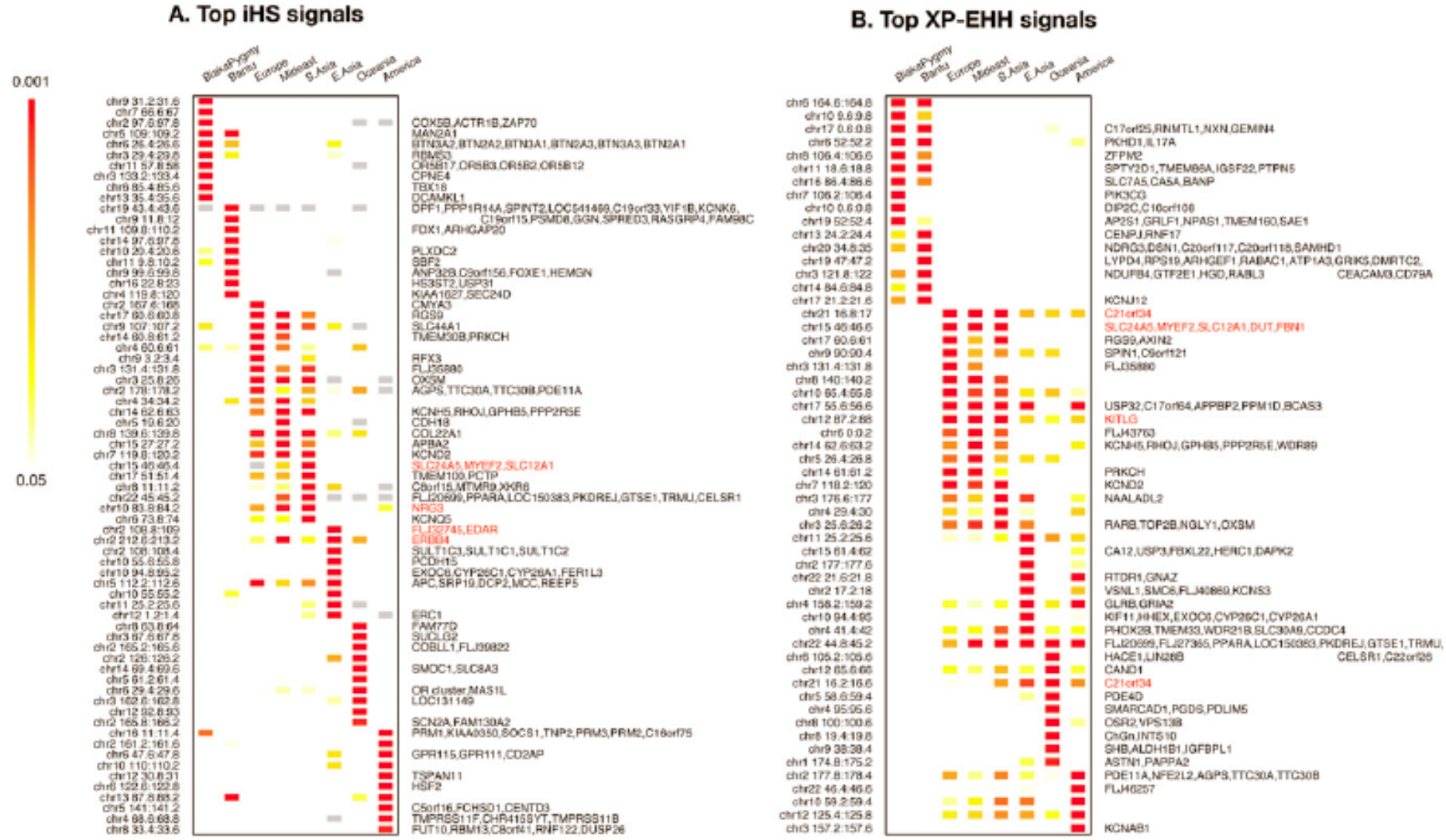


Figure 1. Top 10 iHS (*A*) and XP-EHH (*B*) signals by population cluster. Each row is a 200-kb genomic window, each column is a geographic region, and each cell is colored according to the position of the window in the empirical distribution of scores for that region. Plotted are the most extreme 10 windows for each geographic region. Gray cells in *A* are windows that have fewer than 20 SNPs for which iHS was calculated (see Methods). To the *right* of each row is a list of genes that fall in the window. Windows where the genes are in red are discussed in the text. Note that interpretation of the overlap in XP-EHH signals is complicated by the need for a reference population; see the main text.

Pigmentation genes

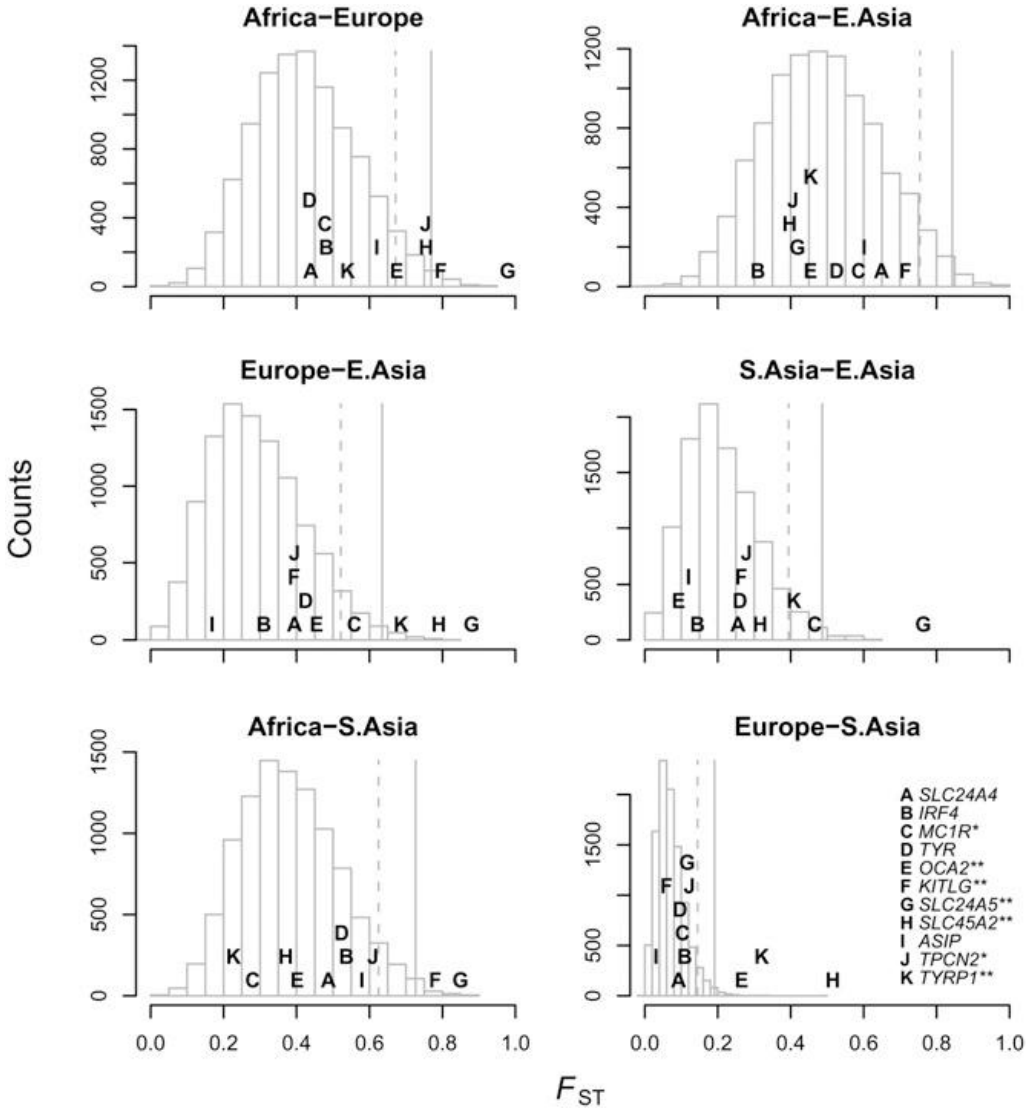


Figure 3. F_{ST} around loci involved in natural variation in pigmentation.

For each SNP found to be associated with pigmentation in a genomewide scan, we plot the maximum pairwise F_{ST} between geographic regions in a 100-kb window surrounding the SNP in the HGDP data, as well as a **histogram of the null distribution calculated by finding the maximum F_{ST} in 100-kb windows surrounding each of 10,000 random SNPs.** The dotted lines shows the position beyond which 5% of the random SNPs fall, and the solid lines the position beyond which 1% of the randomSNPs fall. Gene names that are starred fall in the 5% tail of at least one comparison, and those with two stars fall in the 1% tail of at least one comparison. Letters are positioned along the y-axis to improve readability. The key in the bottom right panel applies to all panels.

Diabetes genes

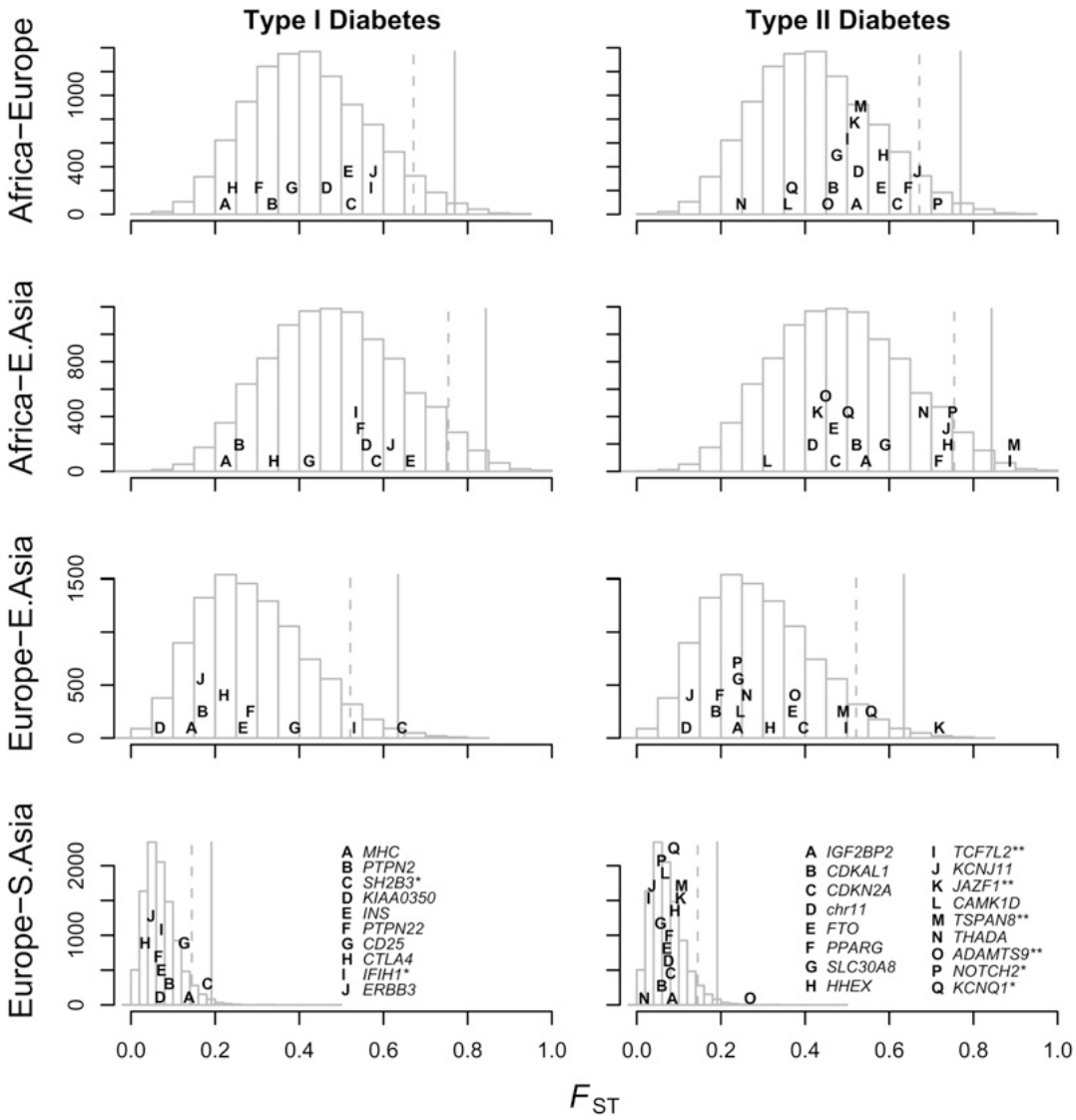


Figure 4. F_{ST} around loci involved in natural variation in diabetes susceptibility.

For each SNP associated with either type I or type II diabetes we plot the maximum pairwise F_{ST} between geographic regions in a 100-kb window surrounding the SNP in the HGDP data, as well as a histogram of the null distribution calculated by finding the maximum F_{ST} in 100-kb windows surrounding each of 10,000 random SNPs. The dotted lines shows the position beyond which 5% of the random SNPs fall, and the solid lines the position beyond which 1% of the random SNPs fall. Gene names that are starred fall in the 5% tail of at least one comparison, and those with two stars fall in the 1% tail of at least one comparison. Letters are positioned along the y-axis to improve readability. The key in the bottom panel of each column applies to the entire column.

C21orf34 region

This region includes terminal three exons of C21orf34, a gene that is expressed in many tissues, as well as three microRNA genes (mir-99a, let-7c, and mir-125-b2).

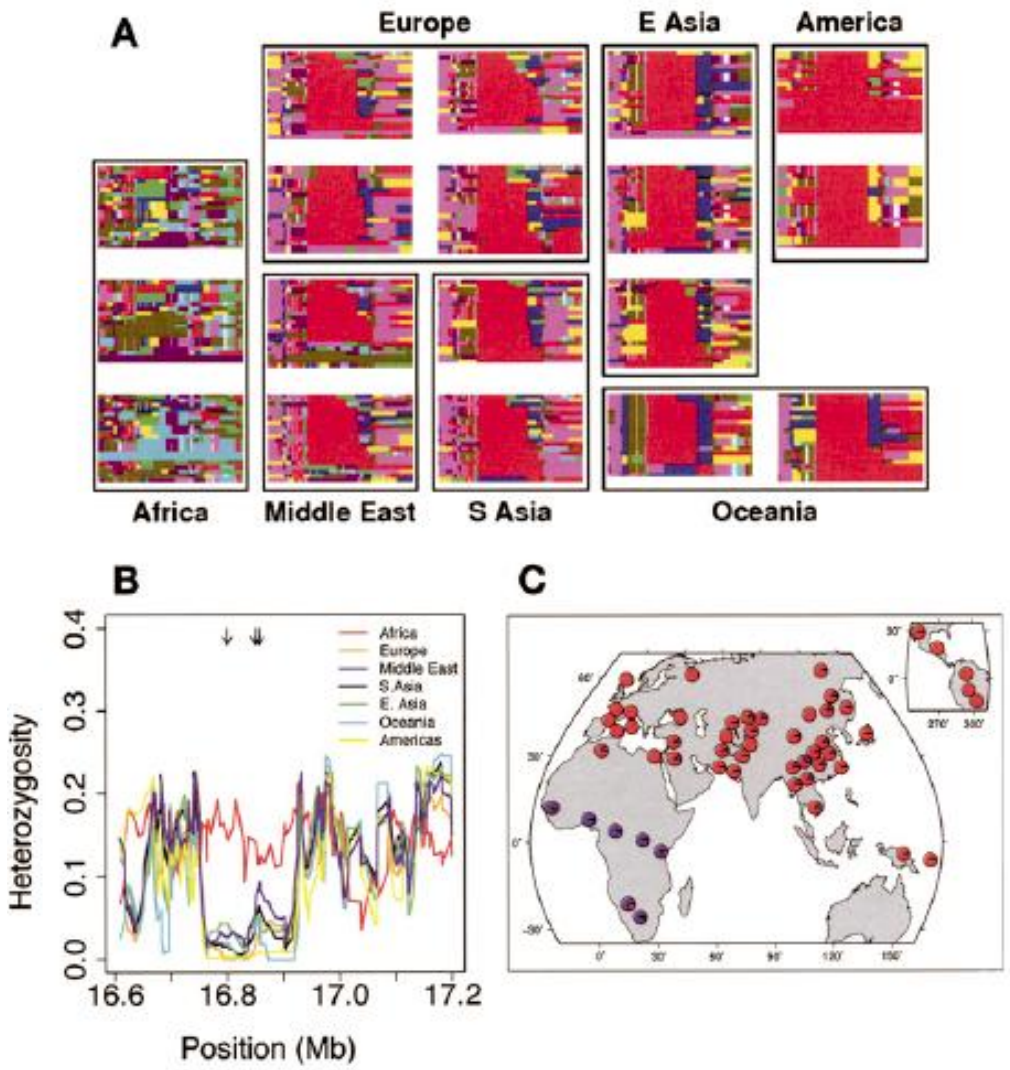


Figure 2. Evidence for selection in a region containing part of the gene C21orf34.

(A) Haplotype plots in a 500-kb region on chromosome 21 surrounding the locus. Each row represents a haplotype, and each column a SNP. Rows are colored the same if and only if the underlying sequence is identical (some low-frequency SNPs are excluded). For full details on the generation of these plots, see Conrad et al. (2006).

(B) Heterozygosity in the same region. Lines show heterozygosity calculated in a sliding window
of three SNPs across the region in different populations. Black arrows at the top of the plot represent
the positions of SNPs with FST > 0.6 (i.e., in the 0.01%tail of worldwide FST).

(C). A pie chart of the worldwide distribution of a SNP that tags the red haplotype in A (rs2823850). (Red) The derived allele frequency; (blue) the ancestral allele frequency.

NRG-ERBB4 pathway

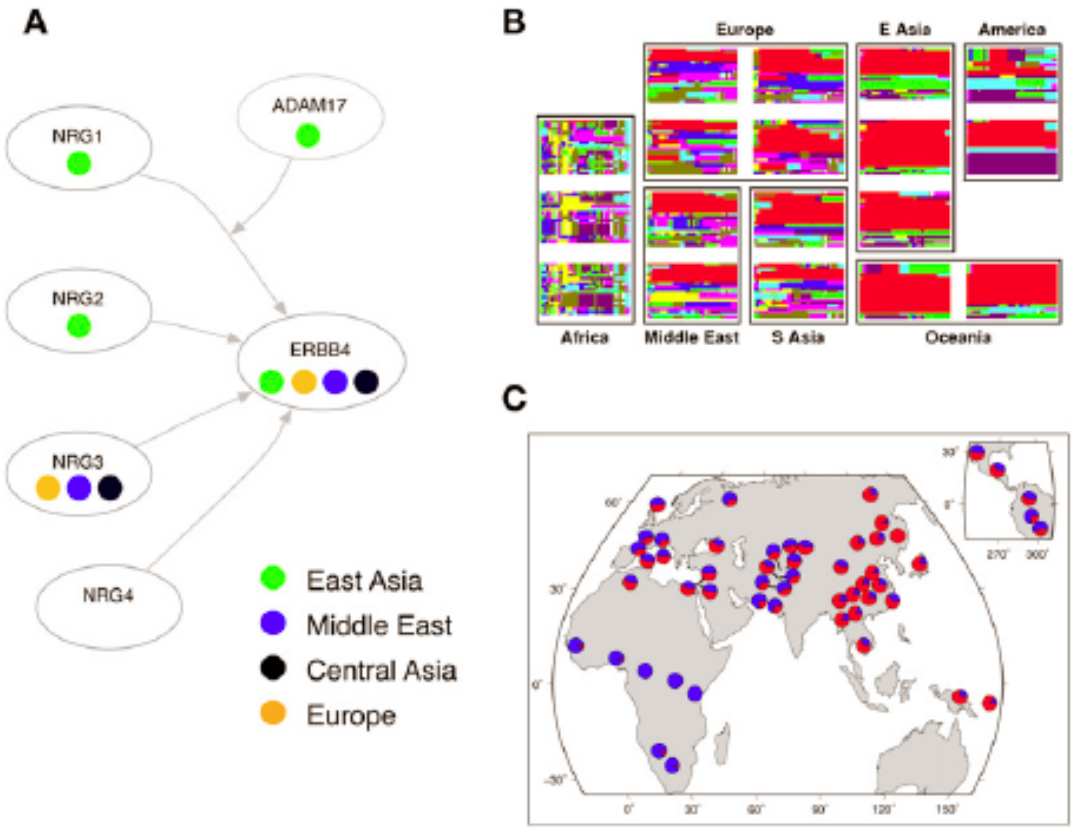


Figure 5. Selection signals in the *NRG–ERBB4* pathway. (*A*) A schematic of the *NRG–ERBB4* pathway, drawn from interactions reported in KEGG (Kanehisa et al. 2008) and Mei and Xiong (2008). Each oval represents a gene, and the colored circles denote the geographic regions that have significant selection signals (empirical scores in the top 5% of the distribution). We excluded Oceania and the Americas from this plot since selection scans are expected to have low power in these regions. For *ADAM17*, the selection statistic is XP-EHH; for the others it is iHS. (*B*) Haplotype plots at the putative selected region in *ERBB4*. (*C*) Worldwide allele frequencies of a SNP that tags the red haplotype in *B* (rs1505353). (Red) The derived allele; (blue) the ancestral allele.

NRG-ERBB4 pathway

Large genes, binding to each other. They both are outliers with respect to the rest of the genome even after a conservative Bonferroni correction for the number of windows (empirical P = 0.001 and P = 0.006 in the Middle East for ERBB4 and NRG3, respectively).

The NRG–ERBB4 signaling pathway is well-studied and known to be involved in the development of a number of tissues, including heart, neural, and mammary tissue.
Variants in genes in this pathway have been associated with risk of schizophrenia and various psychiatric phenotypes.

Other examples of local selection, identified by pairwise F_{st} comparisons

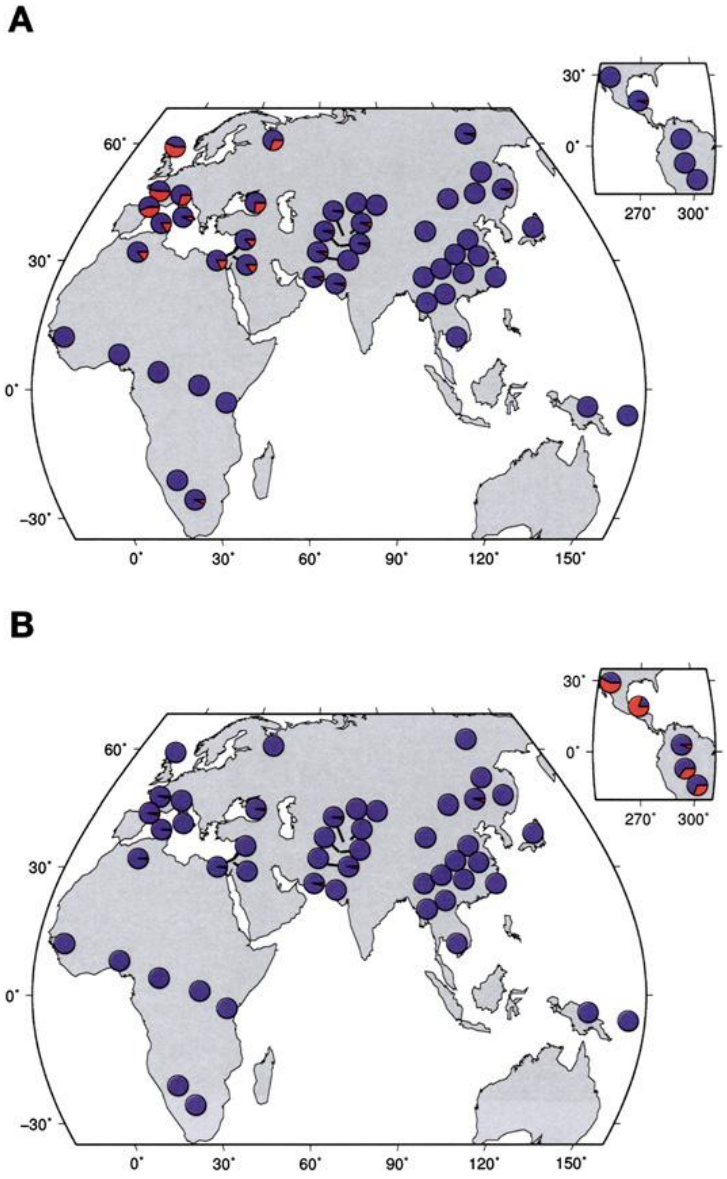


Figure 6. Worldwide allele frequencies of two nonsynonymous SNPs showing evidence of local adaptation.

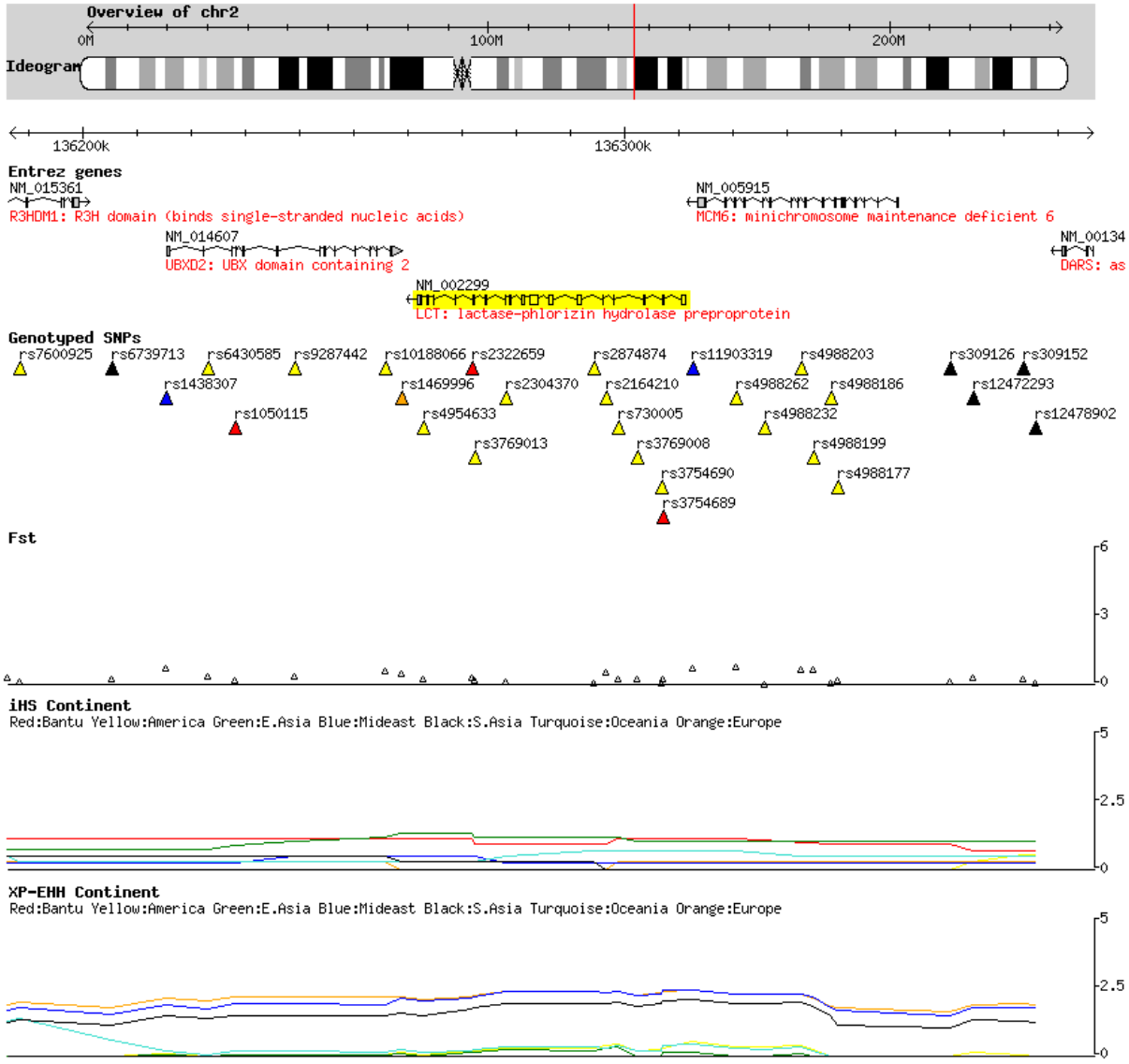
(A) Frequencies of rs5743810 in TLR6 gene;
(B) frequencies of rs12421620 in DPP3 gene.

(Red) The frequency of the derived allele;
(Blue) the frequency of the ancestral allele.

TLR6 is a gene involved in the recognition of bacterial pathogens. DPP3 is highly expressed in lymphoblast - possible link to immunity.

Other attempts:
Pygmy pops (Mbuti and Biaka) vs Bantu did not give any regions that would be clearly associated with body height. However, Biaka vs Yoruba populations indikate differences in PIK3R3 and IGF2R - two genes that might be associated with growth in mice.

Positive selection browser @ Pritchard Lab



Conclusions

Enormous work has been done. More still to do:

- Need more populations with more individuals
- Need to develop methods to detect selection in smaller population groups
- Need better visualization tools
- Need to use rare mutations in addition to SNPs
- More systematic approach to identify the molecular and physiological role of identified genes