A WORLDWIDE SURVEY OF HAPLOTYPE VARIATION AND LINKAGE DISEQUILIBRIUM IN THE HUMAN GENOME

QUESTIONS:

1. How useful are current SNP datasets for studying haplotype variation in diverse human populations?

2. To what extent are patterns of haplotype variation similar?

3. What does those patterns imply about human history and recombination events?

4. To what extent do the HapMap populations predict patterns of haplotype diversity found in the worldwide set of populations?

DATA

- ~1000 distinct individuals for 3,024 SNPs

- Samples divided according to major geographic regions: Africa, Europe, Middle East, Central and South Asia, East Asia, Oceania, Americas representing 52 populations

- Smallest sample size of region is 27 individuals from Oceania

- 36 geomic regions (16 across all autosomes, 16 from chr21, 4 from non-pseudoautosomal chrX) across 12Mb

- Each region covered ~330kb, using 84 SNPs

- Genotyped with Illumina BeadLab 1000 platform

GENOMIC LOCATIONS OF 36 REGIONS
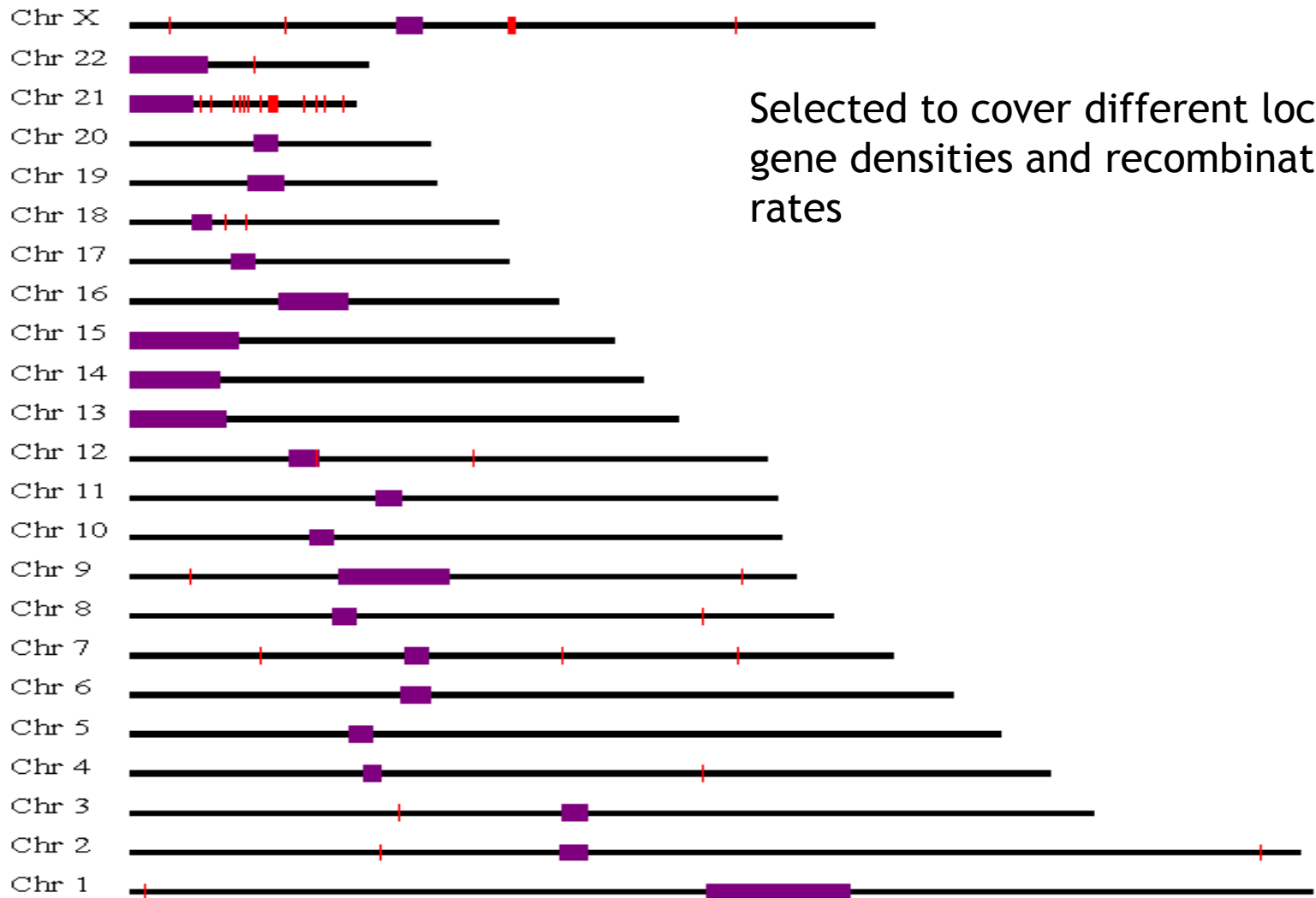
Selected to cover different local gene densities and recombination rates
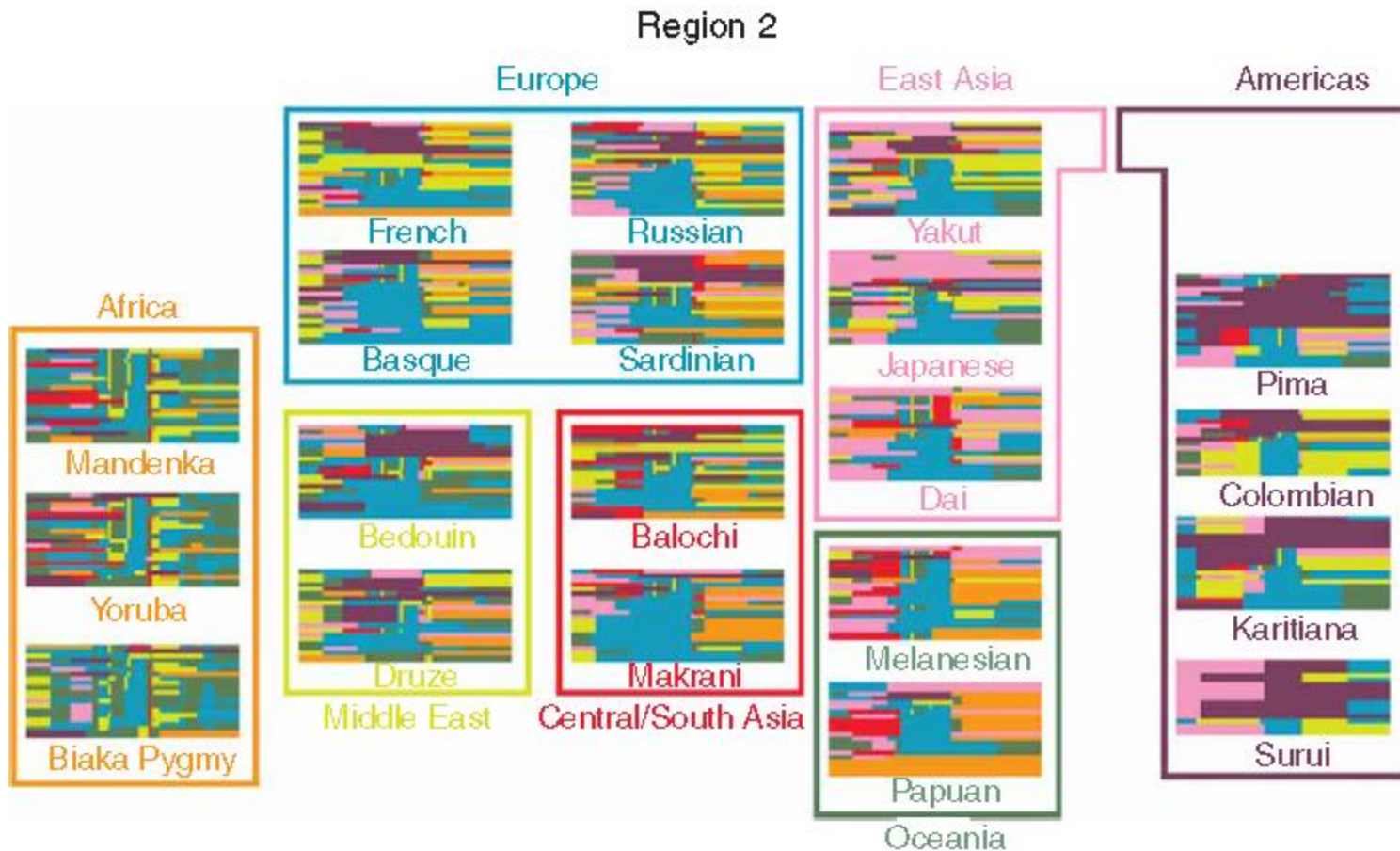
PHASED HAPLOTYPES

Figure 1 Haplotype structure in diverse populations for two genomic regions of size B330 kb. For each population, haplotypes are plotted in rows; horizontal position is proportional to physical position in the sequence. Each haplotype is represented as a mosaic of seven 'template' haplotypes, each of which is common in a different part of the world and is colored using the same color as the text for that region (see Methods). Haplotypes are sorted so that within populations, similar haplotypes are drawn close together. A pair of haplotypes is identical across the entire region if and only if both share the same coloring pattern (rare minor alleles not on any template are dropped from the analysis). The same coloring scheme is used for all populations.
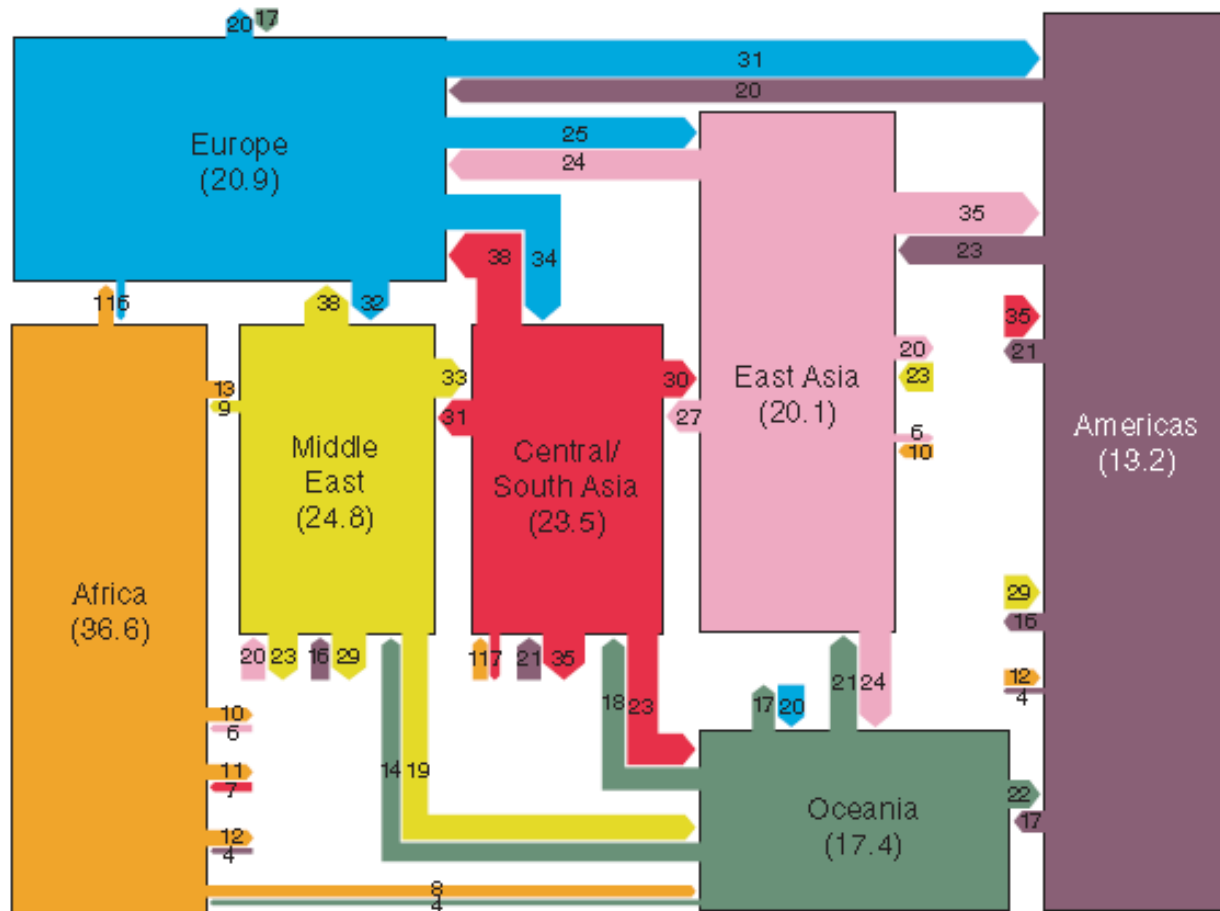
HAPLOTYPE MIGRATION

Figure 2 Schematic world map of haplotype diversity. Colored boxes represent regions of the world, positioned geographically. The average number of haplotypes per genomic core region in a sample size of 54 chromosomes is written for each geographic region. Links entering a geographic region indicate the percentages of distinct haplotypes from the geographic region found in other regions (and are drawn proportionately in width). For example, on average 11% of haplotypes observed in Europe in a given part of the genome are found in Africa, whereas 6% of African haplotypes are found in Europe. The links can be viewed as a description of haplotype 'flow': for example, 11% gives a measurement of the proportion of distinct European haplotypes that could have come from Africa (without mutation or recombination), and 6% gives the proportion of African haplotypes that could have come from Europe.

ASCERTAINMENT BIAS
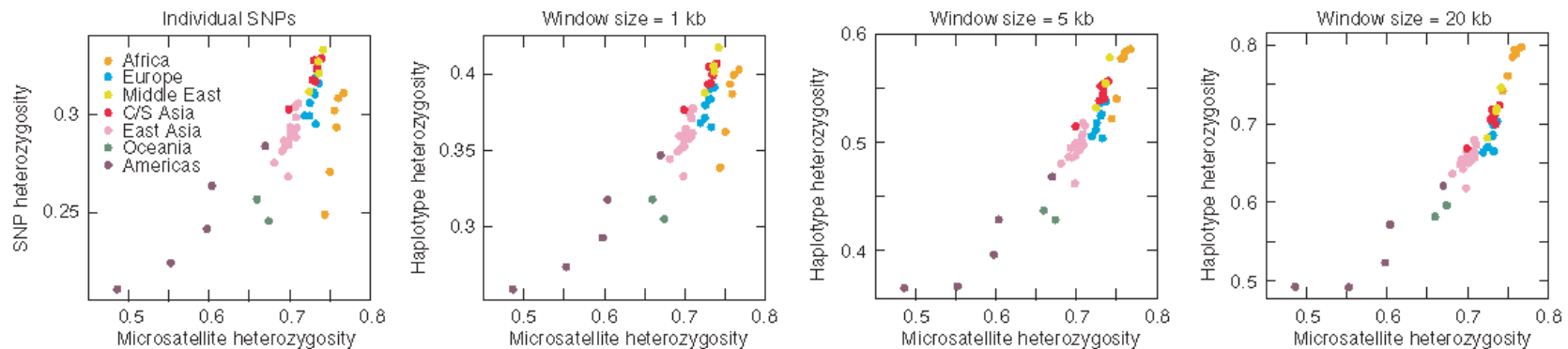
Figure 3 Effect of ascertainment bias on haplotype diversity. The plots show haplotype heterozygosity computed in windows of four sizes, plotted against heterozygosity of microsatellite loci in the same populations[29,41].
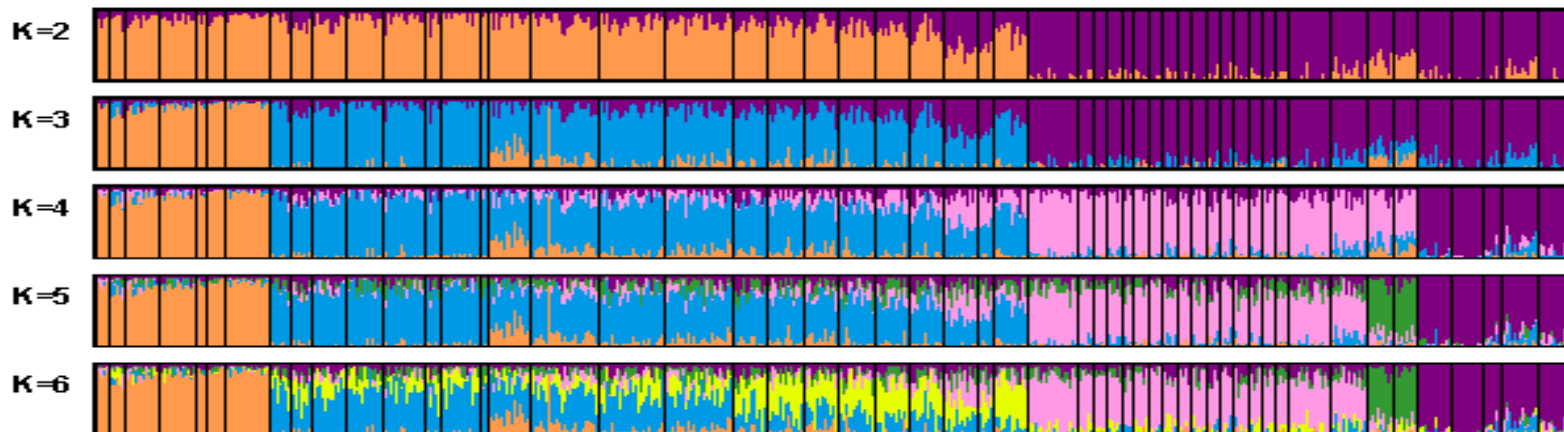
Ascertainment bias occurs when false results are produced by non-random sampling. (wikipedia)

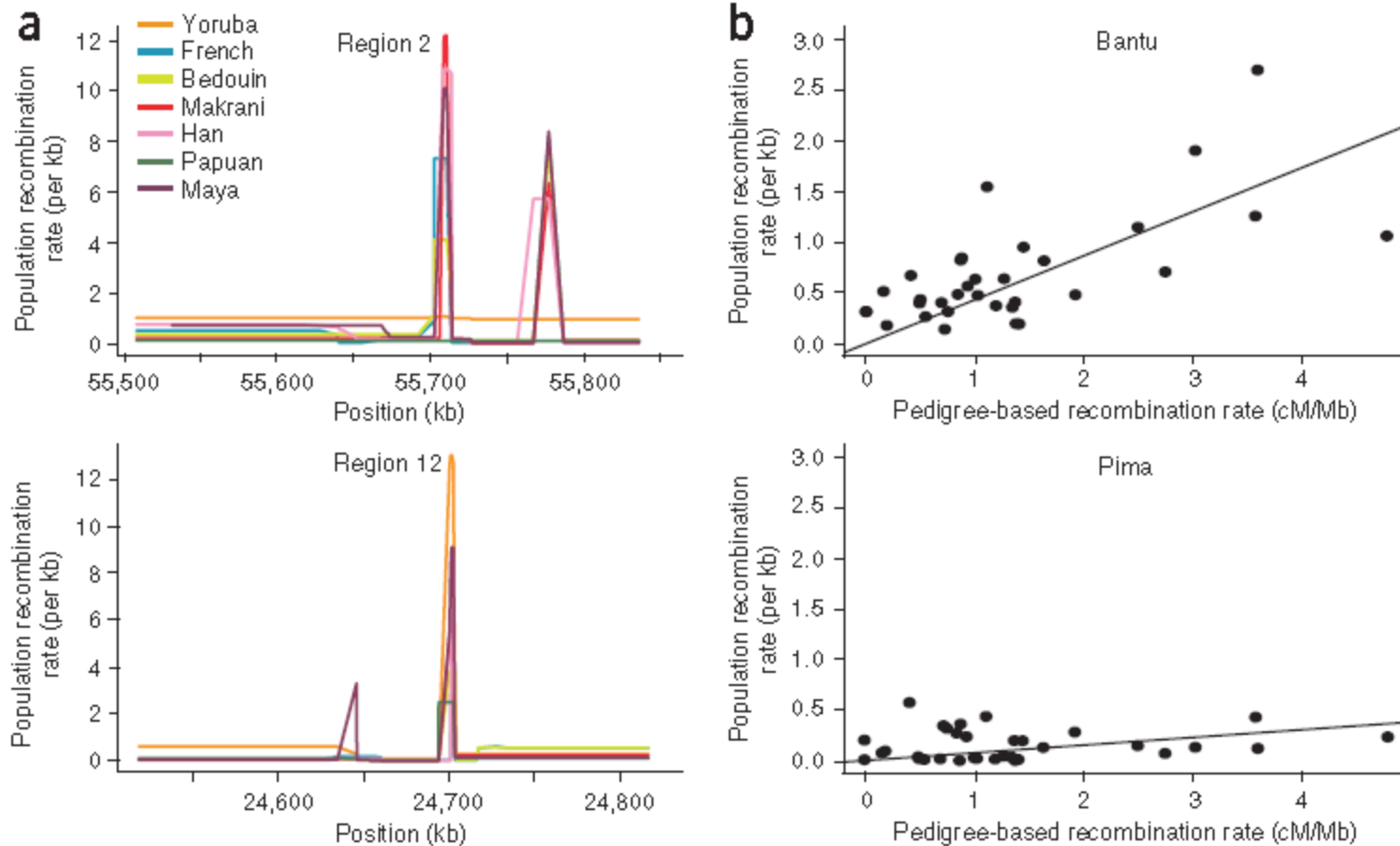# F$_{ST}$ OF SNPS VS. MICROSATELLITES

A. Microsatellites
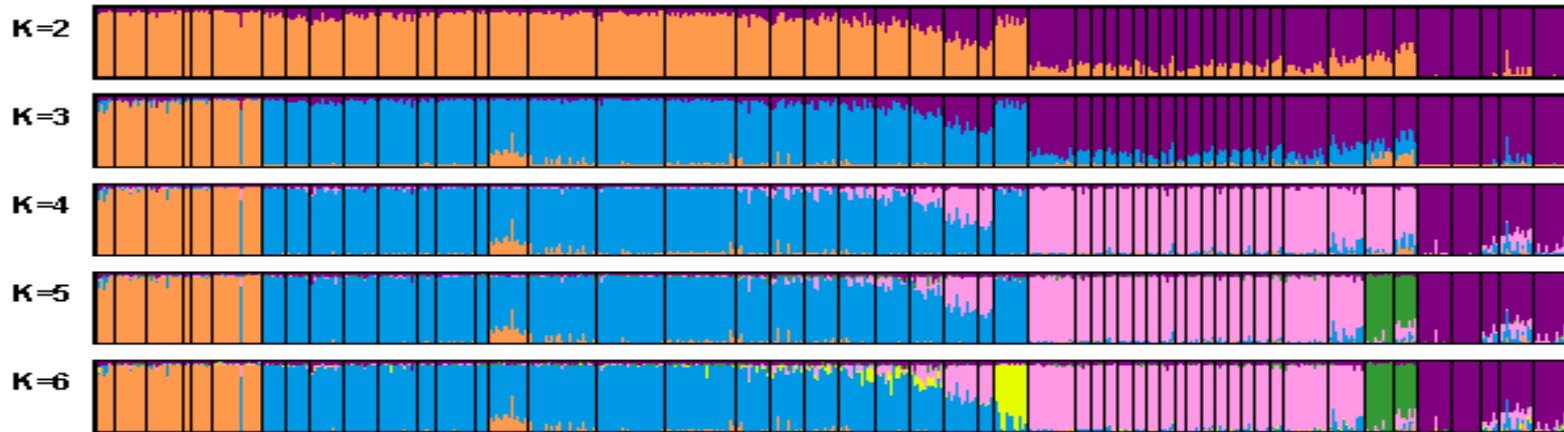
B. SNPs

RECOMBINATION RATES


Figure 4 Population recombination rates (ρ) across genomic regions. **(a)** Fine-scale estimates of variation in ρ across the same two genomic regions shown in **Figure 1**. **(b)** Scatter plot of the relationship between ρ and the meiotic recombination rates for each autosomal region, as estimated from Icelandic pedigrees[27] (data shown for autosomal regions only).

EFFECTIVE POPULATION SIZE VS. MICROSATELLITE HETEROZYGOSITY
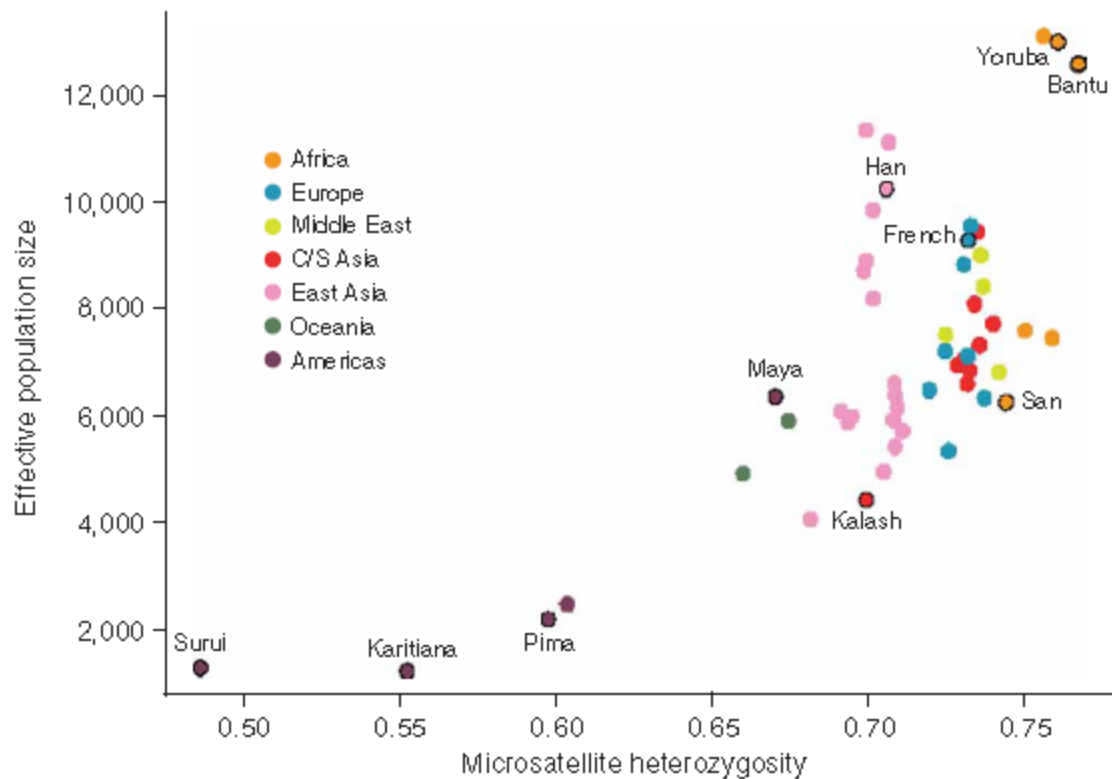
Figure 5 Relationship between $\hat{N}_e(\rho)$ and microsatellite heterozygosity for individual populations.

COMMON HAPLOTYPES WITH HAPMAP
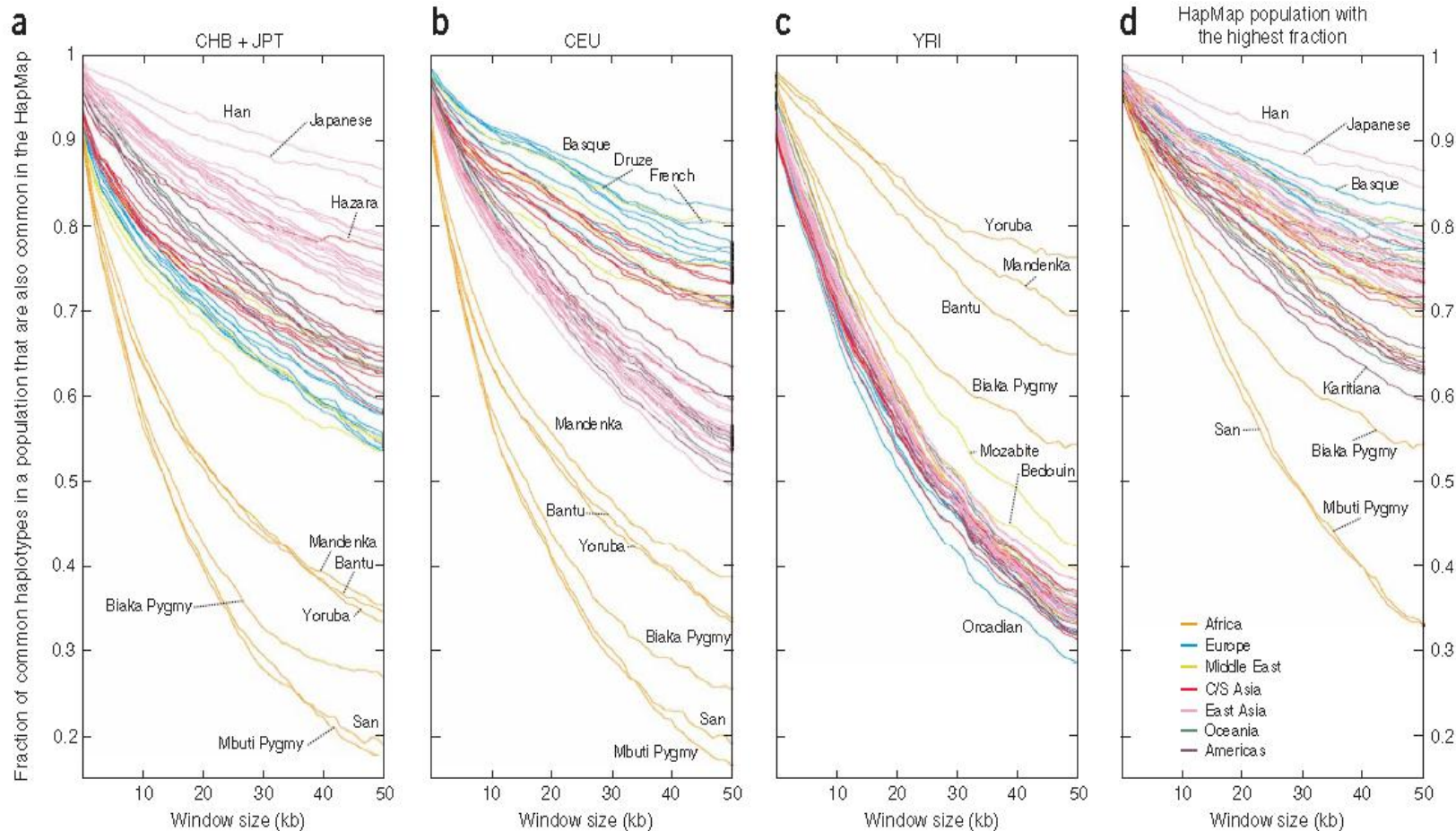
Figure 6 The fraction of common haplotypes in individual populations that are also common in the HapMap. For each plot, we used haplotypes based on the SNPs that overlap between Phase II of the HapMap and our autosomal core regions and averaged over all windows of a given length. Each curve shows the fraction of the common haplotypes of a population (those with >10% frequency) that are also common in a HapMap sample. The HapMap samples are taken as (a) CHB+JPT, (b) CEU, (c) YRI and (d) the maximum for each population of the values in a, b and c, taken pointwise.

PORTABILITY OF HAPMAP'S TAGSNPS

Figure 7 Portability of tag SNPs chosen using the HapMap. **(a)** These columns show for each of 52 populations the proportion of polymorphic non-tag SNPs that have $r^2 > 0.85$ with at least one tag SNP. Tag SNPs were chosen separately from each of the three HapMap groups (CHB+JPT, CEU, YRI; shown separately in each column) to maximize the number of SNPs tagged within those groups, subject to an average tag SNP density of 1 per 7.8 kb (corresponding to a genome-wide screen of about 400,000 SNPs). For each population, the gray bar indicates which tag SNP set is best. Vertical dashed lines indicate 50% tag portability. **(b)** Estimated worldwide population structure based on microsatellite data from the individuals in our study. For each population, the horizontal bar is split into colored segments with lengths proportional to the estimated membership of the population in each of five clusters identified in previous analysis[41] by the program STRUCTURE[47].

DETERMINANTS OF PORTABILITY OF HAPMAP'S TAGSNPS

Figure 8 The determinants of portability of HapMap tag SNPs. **(a)** The relationship between tag portability and the distance at which the r^2 measure of linkage disequilibrium decays below 0.5. **(b)** The relationship between tag portability and F_{ST} genetic distance to the HapMap population that produces the highest tag portability. For each population, tag portability is computed as the maximum of the three PVT values in **Figure 7a**. Spearman correlation coefficient equals 0.72 between tag portability and r^2-decay distance and equals −0.16 between tag portability and F_{ST}.
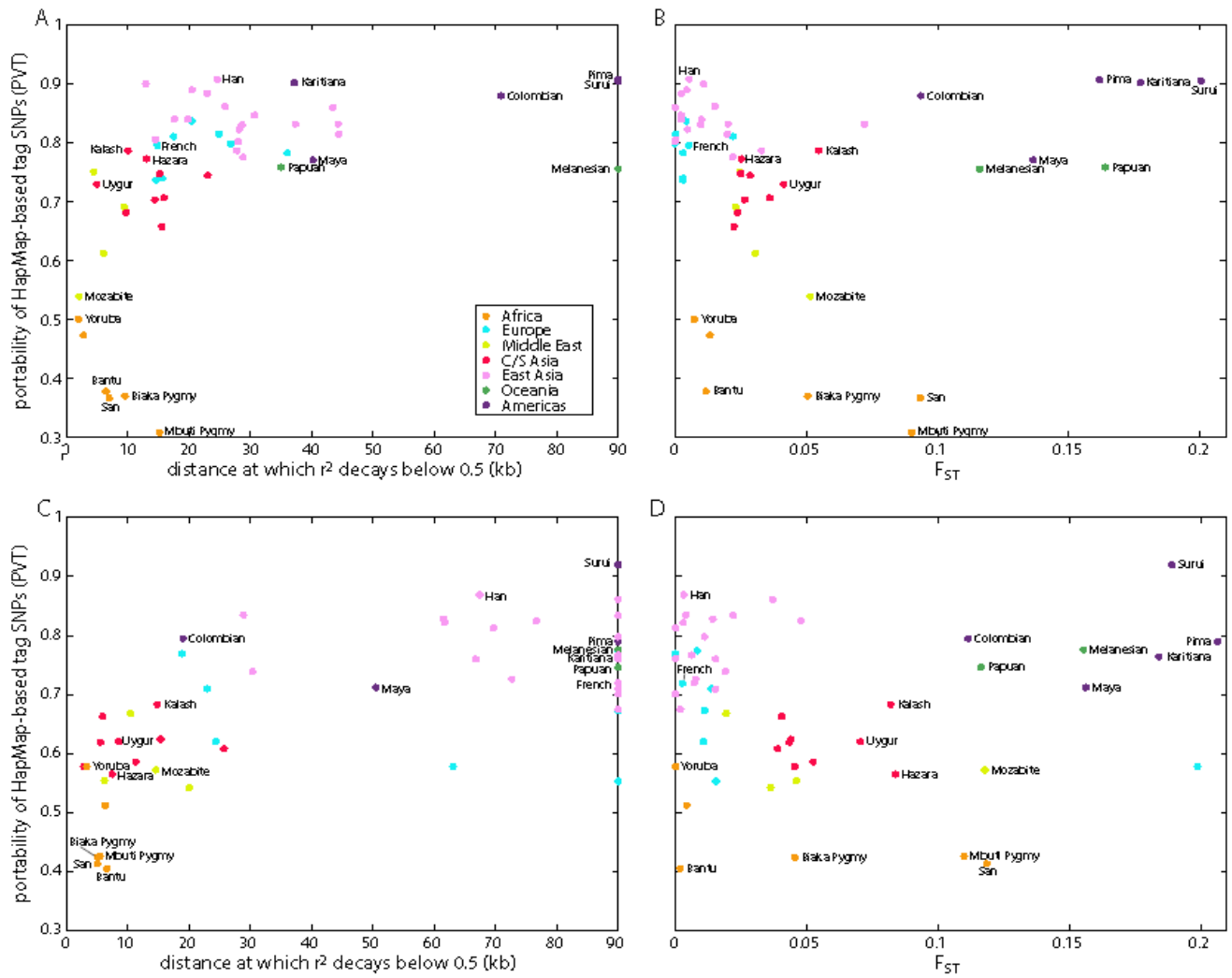
CONCLUSIONS

- 83% of common 20-kb haplotypes in a population are also common in most similar HapMap population

- Although the portability of tag SNPs based on the HapMap is reduced in the low-LD Africans, the HapMap will be helpful for design of genome-wide association mapping studies in nearly all human populations
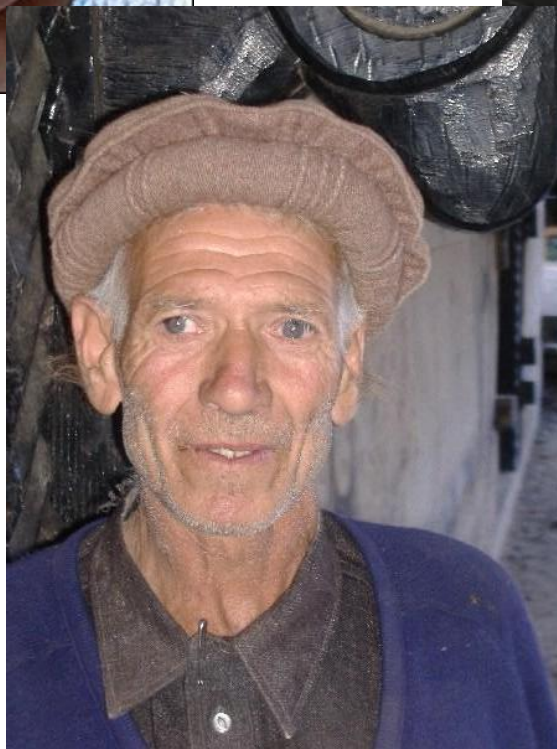
Kalash people

The relationships between [First column] tag portability and the distance at which the r2 measure of linkage disequilibrium decays below 0.5, and between [Second column] tag portability and FST genetic distance to the HapMap population that produces the highest tag portability. For each population, tag portability is computed as the maximum of the three PVT values in Figure 7A. (A),(B): Only SNPs from the Patil regions (chromosome 21) were used. (C), (D): Only SNPs from the non-Patil regions (autosomal non-chromosome 21) were used.