MMG1001 Genomics
Week 2 Tutorial

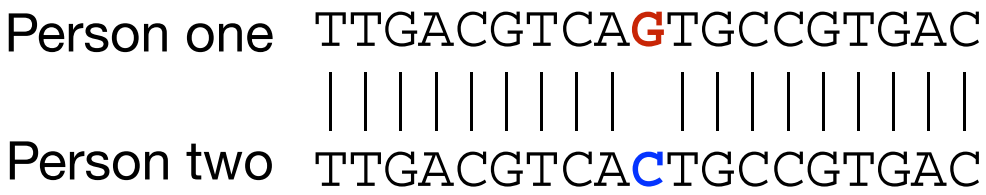
The Structure of Haplotype Blocks in the Human Genome

TA: Heather Gibling

*Many slides borrowed and/or modified from Kaitlin Laverty

Single Nucleotide Polymorphisms (SNPs)

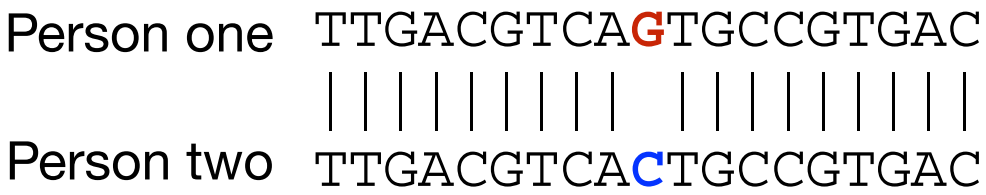
A SNP is SNV (single nucleotide variant) that occurs with greater frequency in the population)



Most are probably silent, but many impact gene expression, splice variation, or protein properties.

Single Nucleotide Polymorphisms (SNPs)

A SNP is SNV (single nucleotide variant) that occurs with greater frequency in the population)

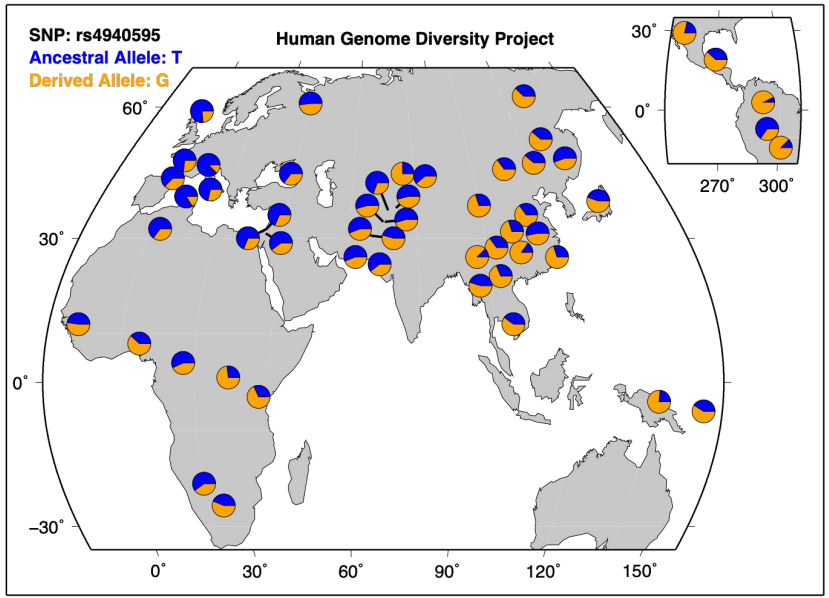


Most are probably silent, but many impact gene expression, splice variation, or protein properties.

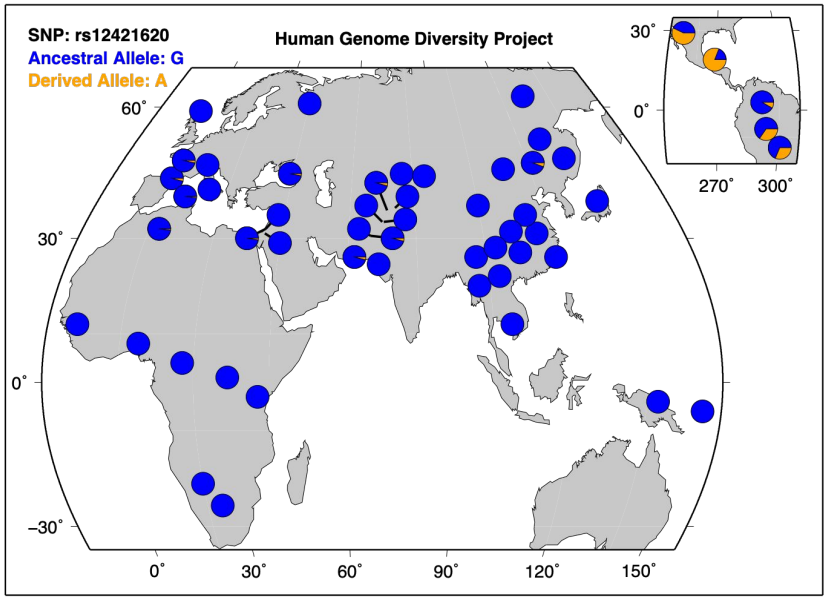
What is a minor allele? Minor allele frequency (MAF)?

Allele frequencies are different between populations

Both alleles occur in all populations, but with variation in the frequency

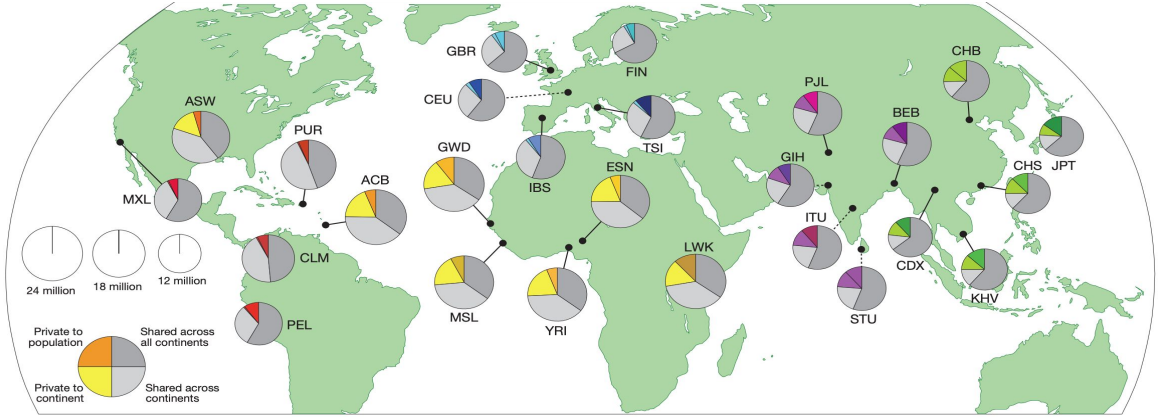


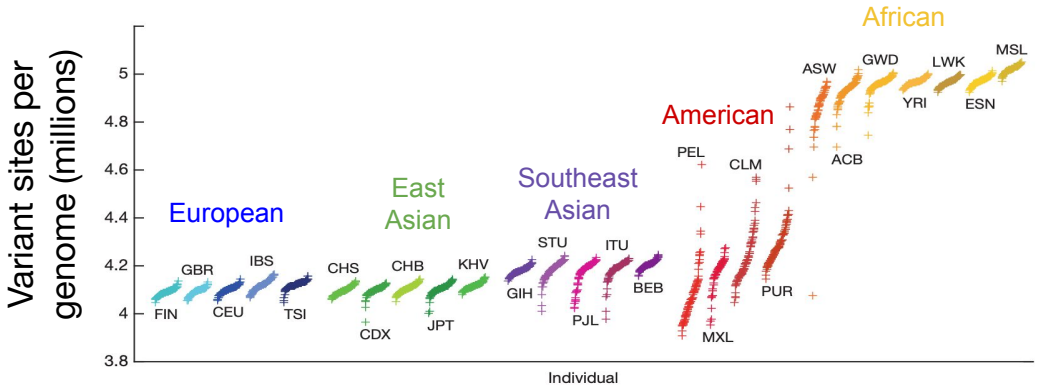
Derived allele is only seen in some populations



Human Genome Diversity Project Browser Data

Genetic diversity within populations



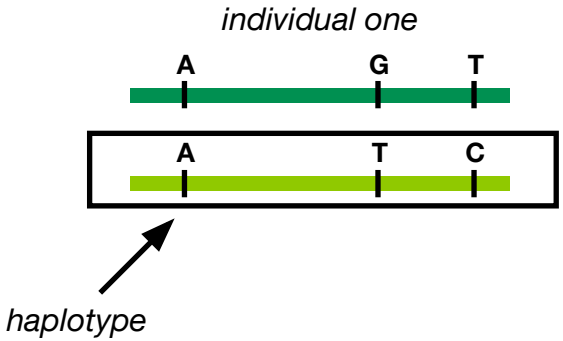


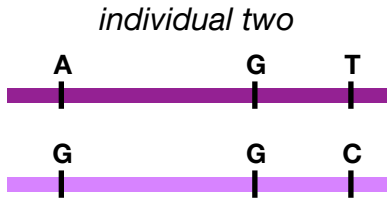
African populations are genetically more diverse than any other population.

Why? What is the founder effect?

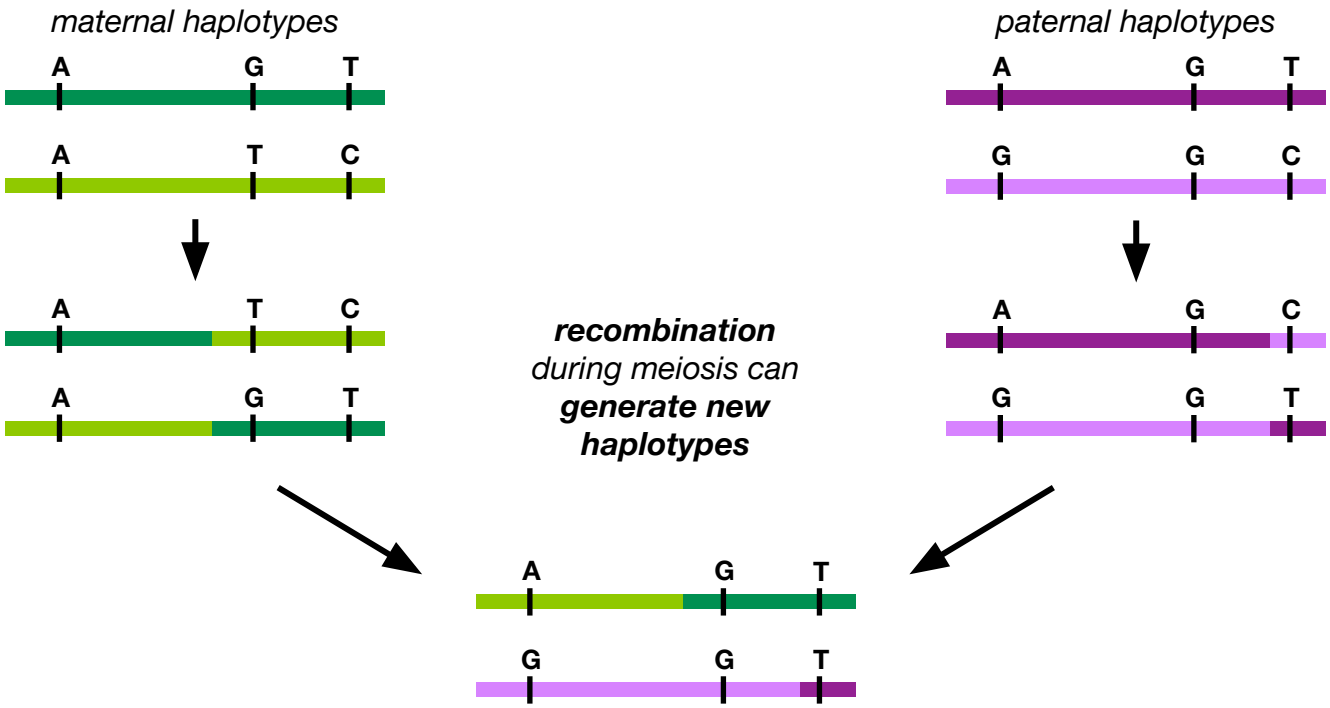
1000 Genomes Project Consortium (2015) *Nature* 526:68–74

Haplotypes are sets of alleles





Haplotypes are sets of alleles

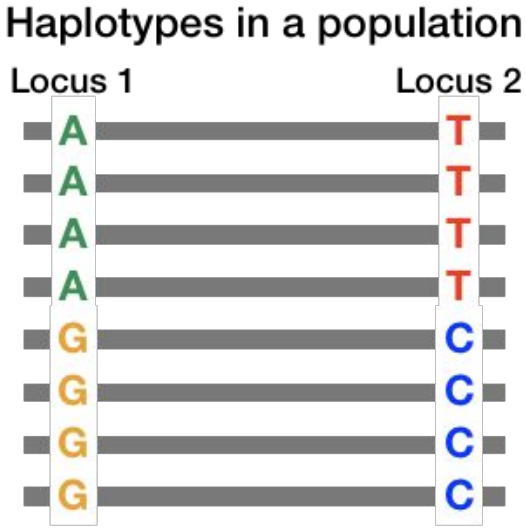


*offspring **inherit one haplotype** from **each parent***

Linkage disequilibrium (LD)

Linkage disequilibrium (LD): “non-random association of alleles at two or more loci in a general population”*

If you have two SNP loci with two alleles each:
4 possible haplotypes: **AC**, **AT**, **GC** and **GT**



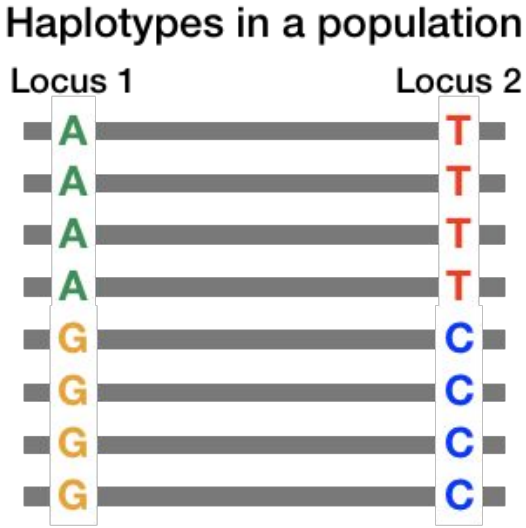
Are the alleles **G** and **C** in LD in this population?

*Goode E.L. (2011) Linkage Disequilibrium. In: Schwab M. (eds) Encyclopedia of Cancer

Linkage disequilibrium (LD)

Linkage disequilibrium (LD): “non-random association of alleles at two or more loci in a general population”*

If you have two SNP loci with two alleles each:
4 possible haplotypes: **AC**, **AT**, **GC** and **GT**



Are the alleles **G** and **C** in LD in this population?

D = pAB - pApB = 0.5 - 0.5*0.5 = **0.25**

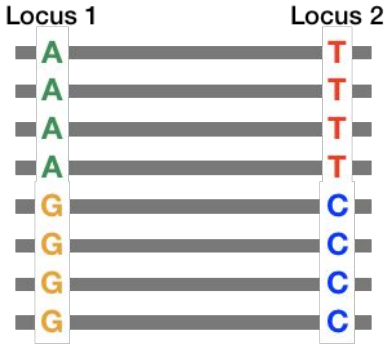
pAB - frequency of the haplotype (50% of haps are **GC**)
pA - frequency of the first allele (**G** occurs 50% of the time)
pB - frequency of the second allele (**C** occurs 50% of the time)

If D = 0 alleles are in equilibrium
If D ≠ 0 alleles are in **LD**

*Goode E.L. (2011) Linkage Disequilibrium. In: Schwab M. (eds) Encyclopedia of Cancer

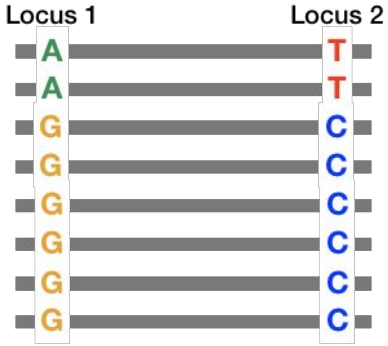
Linkage disequilibrium (LD)

Haplotypes in a population



D = pAB - pApB
= 0.5 - 0.5*0.5
= **0.25**

Haplotypes in a population

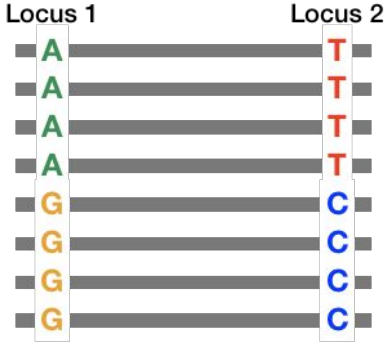


D = pAB - pApB
= 0.75 - 0.75*0.75
= **0.1875**

The value of D depends on the haplotype frequencies.
G and C are in LD in both of these populations, but the values of D are very different!

Linkage disequilibrium (LD)

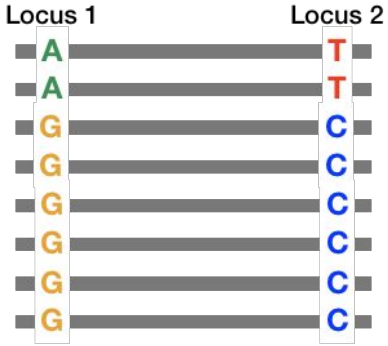
Haplotypes in a population



D = pAB - pApB
= 0.5 - 0.5*0.5
= **0.25**

D' = D / min(pApb, papB)
= 0.25 / min((0.5 * 0.5), (0.5 * 0.5))
= 0.25 / min((0.25), (0.25))
= 0.25 / 0.25
= **1**

Haplotypes in a population



D = pAB - pApB
= 0.75 - 0.75*0.75
= **0.1875**

D' = D / min(pApb, papB)
= 0.1875 / min((0.75 * 0.25), (0.25 * 0.75))
= 0.1875 / min((0.1875), (0.1875))
= 0.1875 / 0.1875
= **1**

The value of D depends on the haplotype frequencies.
G and **C** are in LD in both of these populations, but the values of D are very different!

D' is a standardized version of D
Essentially, D is the difference between the observed and expected frequencies. If you divide this by the maximum possible difference you could get, you get a standardized measure between 0 and 1.

Both examples have **D' = 1**

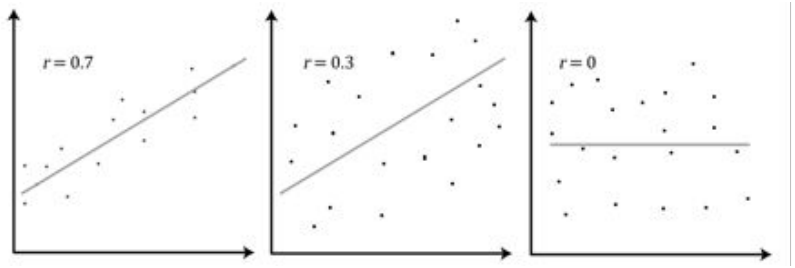
$$D' = D/D_{max}$$
$$If\ D > 0, \quad D_{max} = \min(p_A p_b, p_a p_B)$$
$$If\ D < 0, \quad D_{max} = -\max(p_A p_B, p_a p_b)$$

Measures of Linkage Disequilibrium (LD)

D: Difference between the number of times you OBSERVE two alleles together and the number of times you EXPECT to see them together

D': D standardized by the maximum possible D you could see for the two alleles you are examining

Measures of Linkage Disequilibrium (LD)



$$r = \frac{D}{\sqrt{p_A p_a p_B p_b}}$$

r²: Pearson correlation coefficient squared to give a value between 0 and 1

D'

- Ranges between 0 and 1
- 1 implies at least one of the possible haplotypes was not observed
- D' estimates inflated in small samples or when one allele is rare

r²

- Ranges between 0 and 1
- 1 when the two markers provide identical information
- Prefered measure for population geneticists

https://statsmethods.wordpress.com/2013/05/10/pearson-correlation-coefficient-r

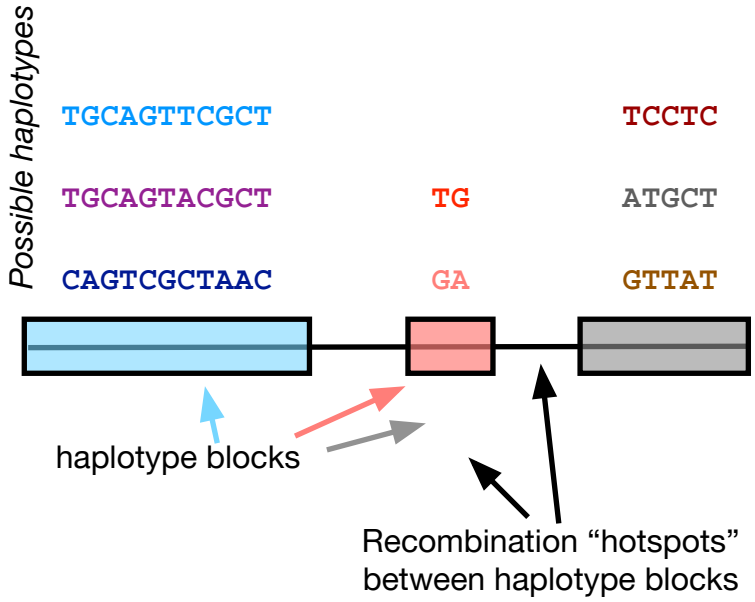
Haplotype blocks contain many haplotypes

Haplotype blocks are stretches of DNA that contain a number of different haplotypes.

The haplotypes are generally in strong LD.

They blocks are separated by sites of recombination.

“We defined a haplotype block as a region over which a very small proportion (5%) of comparisons among informative SNP pairs show strong evidence of historical recombination.”



REPORTS

The Structure of Haplotype Blocks in the Human Genome

Stacey B. Gabriel,[1] Stephen F. Schaffner,[1] Huy Nguyen,[1] Jamie M. Moore,[1] Jessica Roy,[1] Brendan Blumenstiel,[1] John Higgins,[1] Matthew DeFelice,[1] Amy Lochner,[1] Maura Faggart,[1] Shau Neen Liu-Cordero,[1,2] Charles Rotimi,[3] Adebowale Adeyemo,[4] Richard Cooper,[5] Ryk Ward,[6] Eric S. Lander,[1,2] Mark J. Daly,[1] David Altshuler[1,7]*

What are the main goals of the paper?

What are the main goals of the paper?

To try to answer the following open questions:

1. How much of the human genome exists in such blocks, and what are the size and diversity of haplotypes within blocks?

2. To what extent do these characteristics vary across population samples?

3. Can haplotype patterns be parsed using only common SNPs sampled from the population, or will the pattern only emerge after complete resequencing?

4. How completely does such a haplotype framework capture common sequence variation within each region?

How did they do it?

How did they do it?

Genotyping! ...the old-school way (primer extension of multiplex products with detection by **MALDI-TOF** mass spec)*

* I have no idea what this is

54 autosomal regions (~0.4% human genome)

- average size: 250kb
- contain ~1 SNP per 2kb
- 4532 SNPs in total

275 individuals

- Yoruban - 90 individuals (30 trios)
- European - 93 individuals (12 pedigrees)
- Japanese and Chinese - 42 individuals (unrelated)
- African American - 50 individuals (unrelated)

What were the major findings?

Figure 1 A

Majority of SNPs assayed were polymorphic

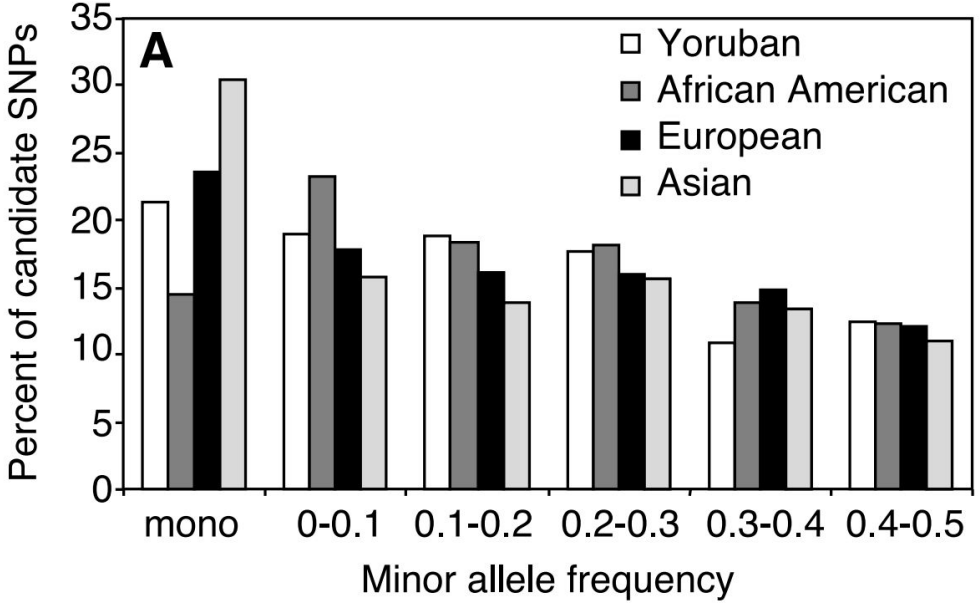
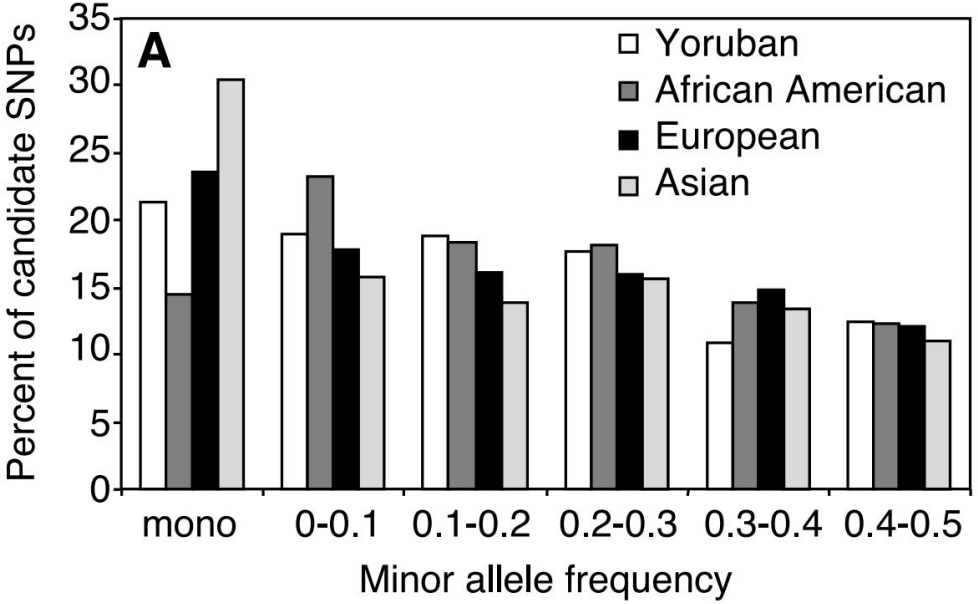


Figure 1 A

Majority of SNPs assayed were polymorphic



Percent of SNPs that were polymorphic:

- 70% (Asian)
- 86% (African American)

What does 'mono' mean?

Figure 1 B

Evidence for historical recombination rises rapidly with respect to distance

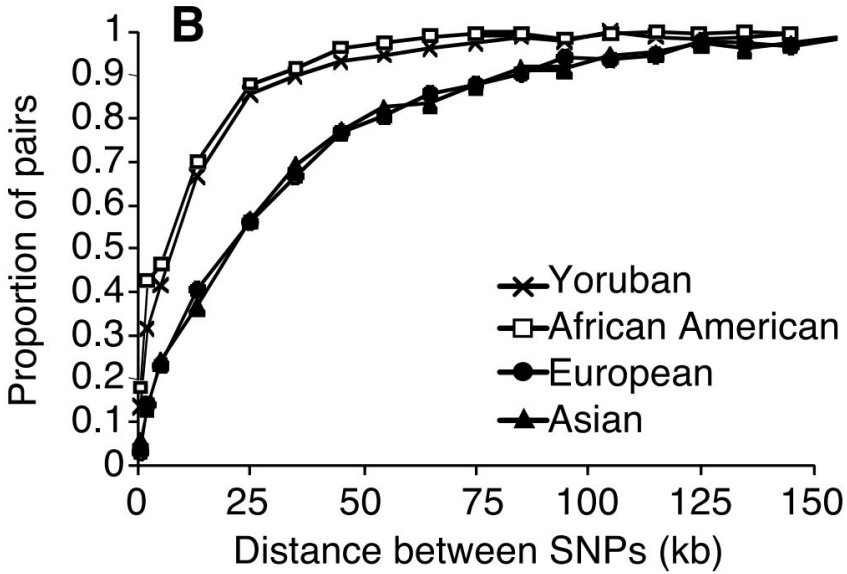
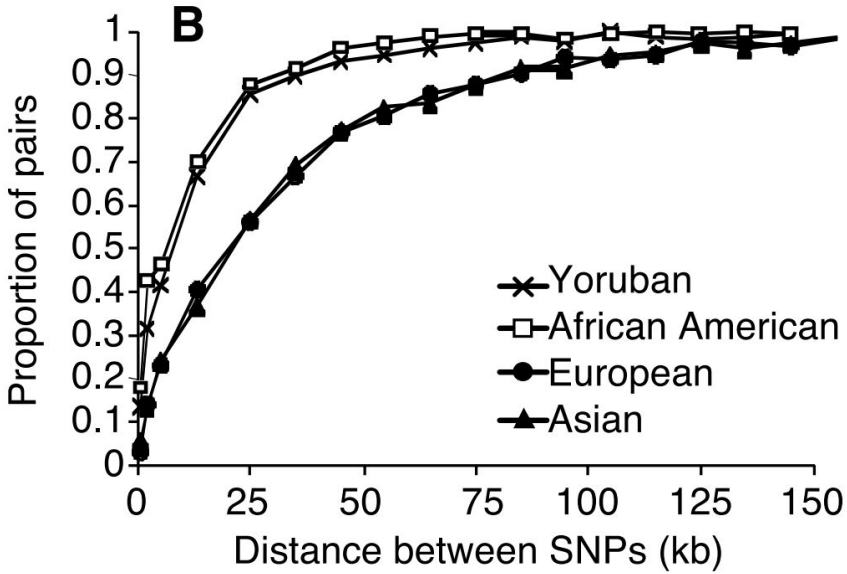


Figure 1 B

Evidence for historical recombination rises rapidly with respect to distance



Categorized informative SNPs as:

- **Strong LD**:
  D' confidence bounds [>0.7, >0.98]
- **Strong evidence for historical recombination**:
  D' confidence bounds [0, < 0.9]

50% of SNP pairs were informative for recombination at:

- 8kb apart (Yoruban, African American)
- 22kb apart (European, Asian)

What does this tell us about African haplotype blocks?

Allele frequency of SNPs varies across populations

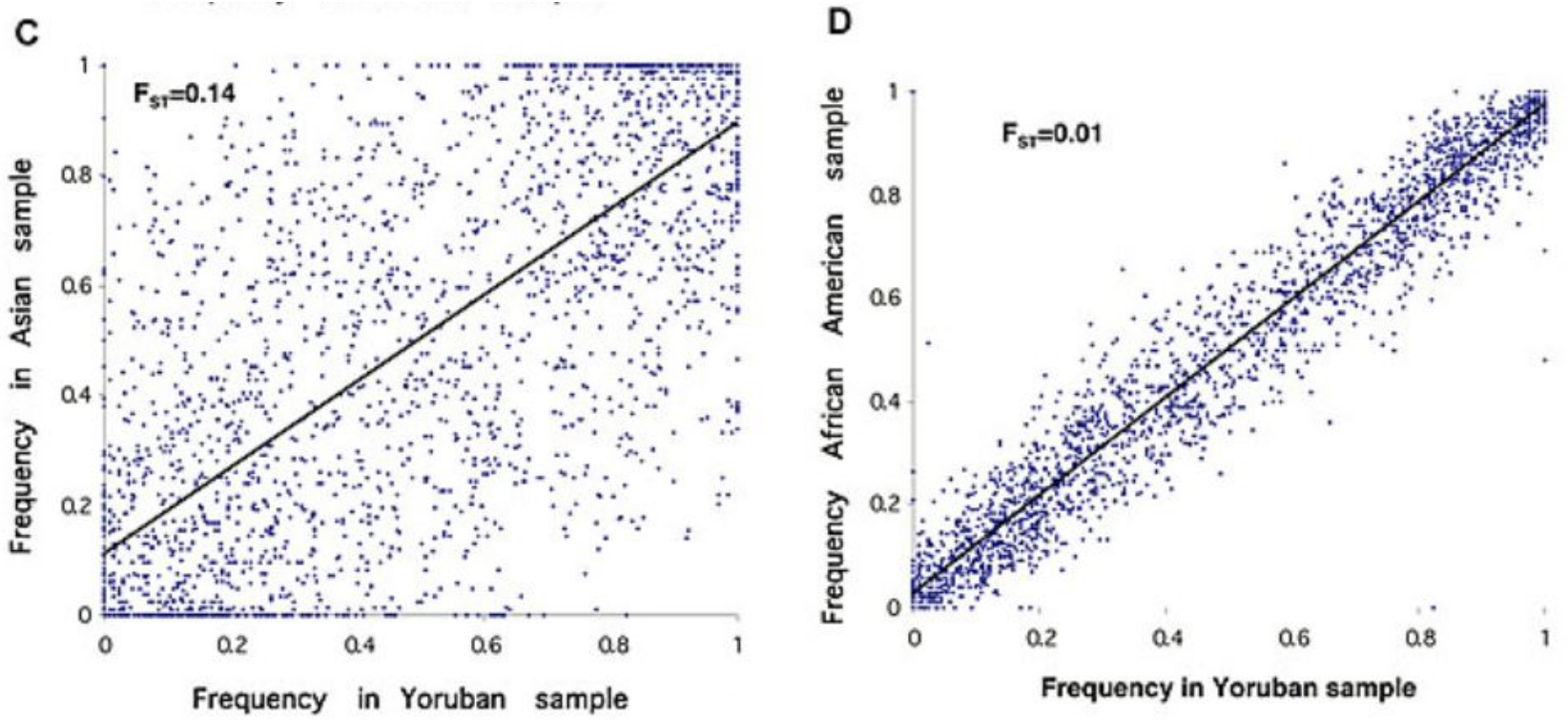


Figure 2 A,B

Only a few SNPs required for informative markers about historical recombination

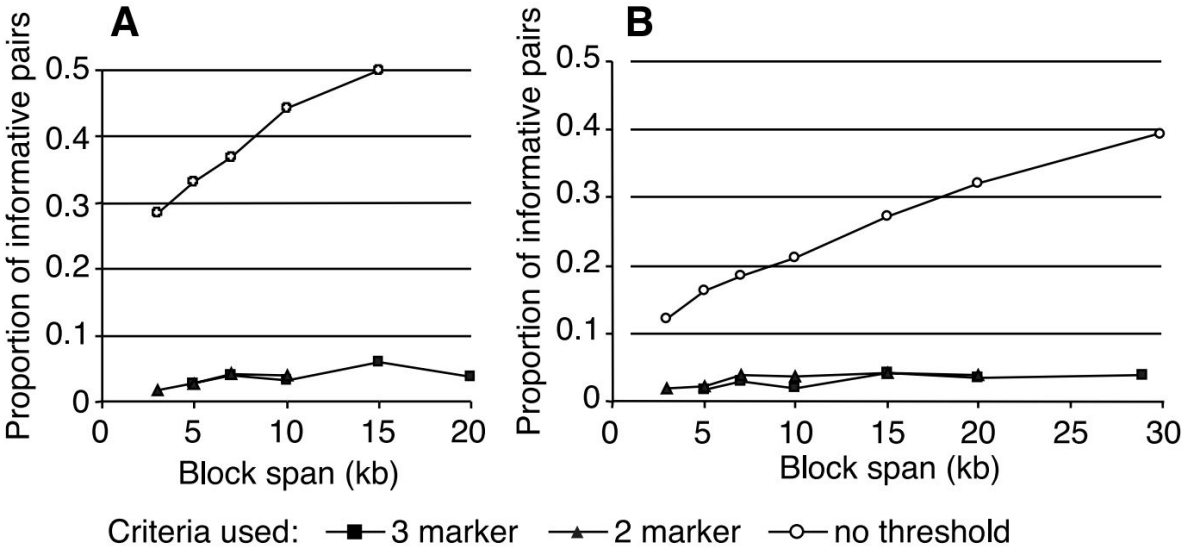
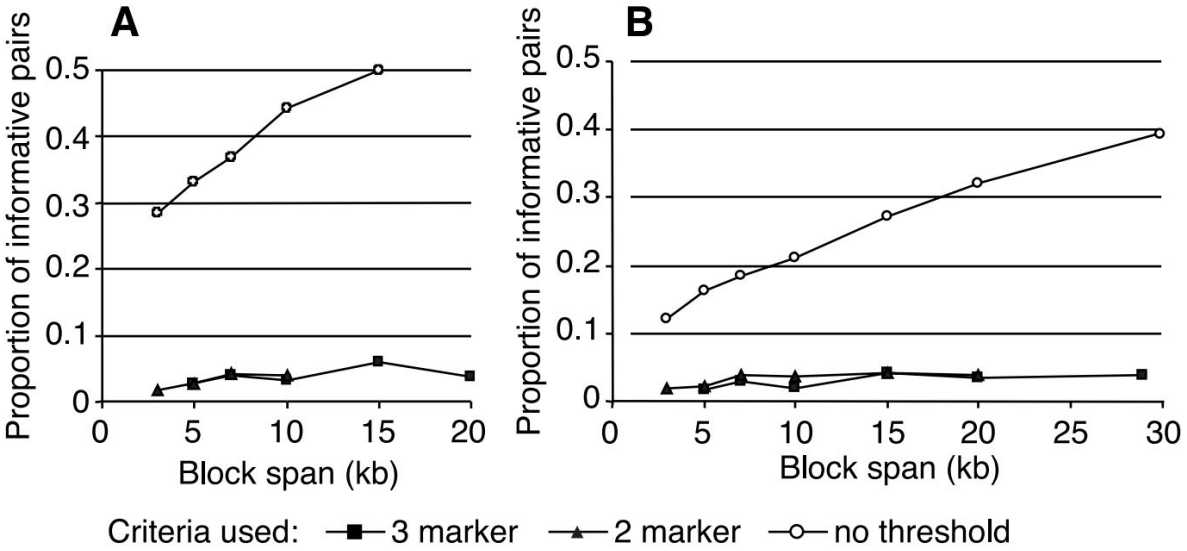


Figure 2 A,B

Only a few SNPs required for informative markers about historical recombination



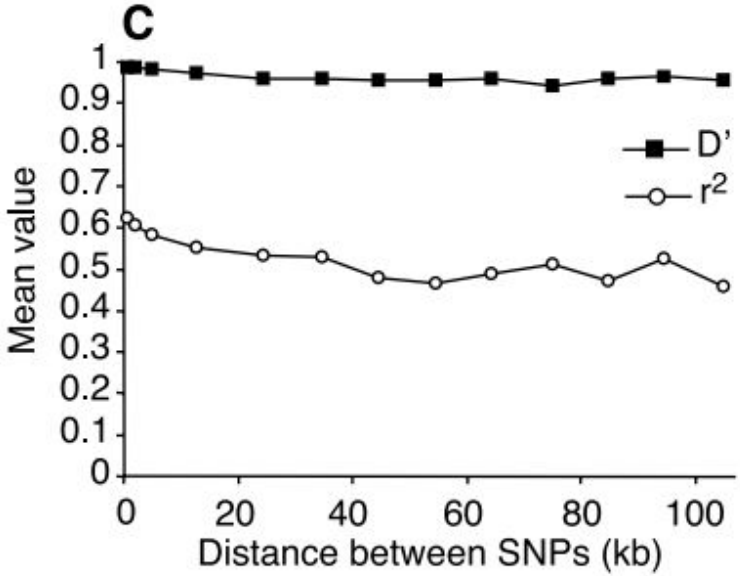
Sampled SNP pairs, found **2-3 markers sufficient** to characterize haplotype blocks (even long blocks).

Can define haplotype blocks where marker coverage is less dense.

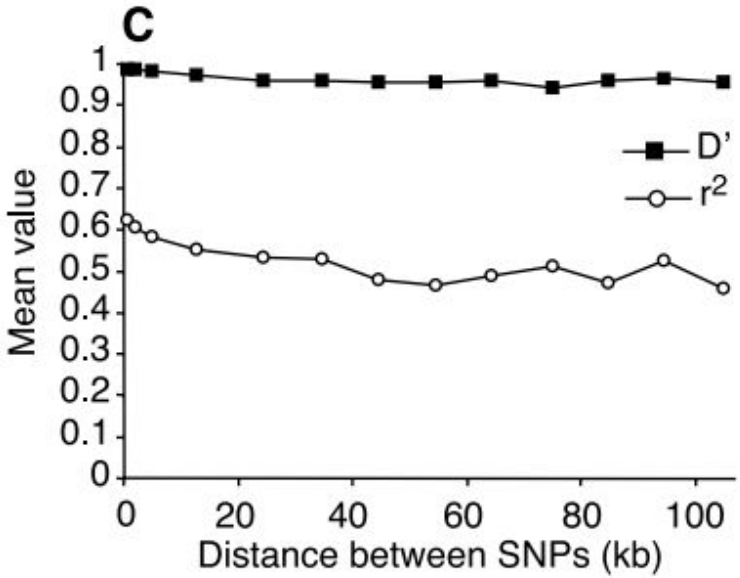
Are marker SNPs uniformly distributed?

Figure 2 C

Measures of LD consistent across the length of defined haplotype blocks



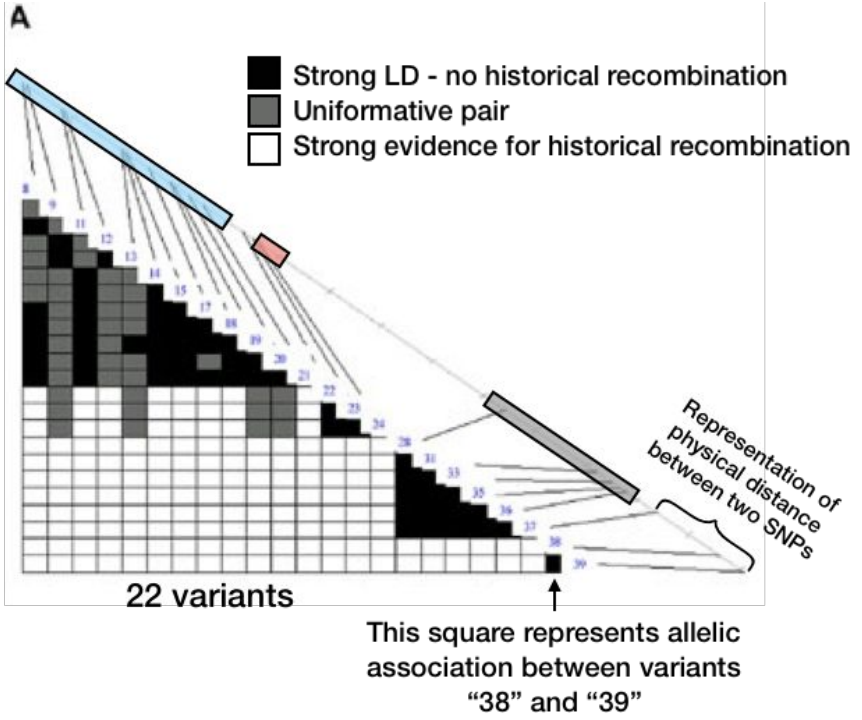
Measures of LD consistent across the length of defined haplotype blocks

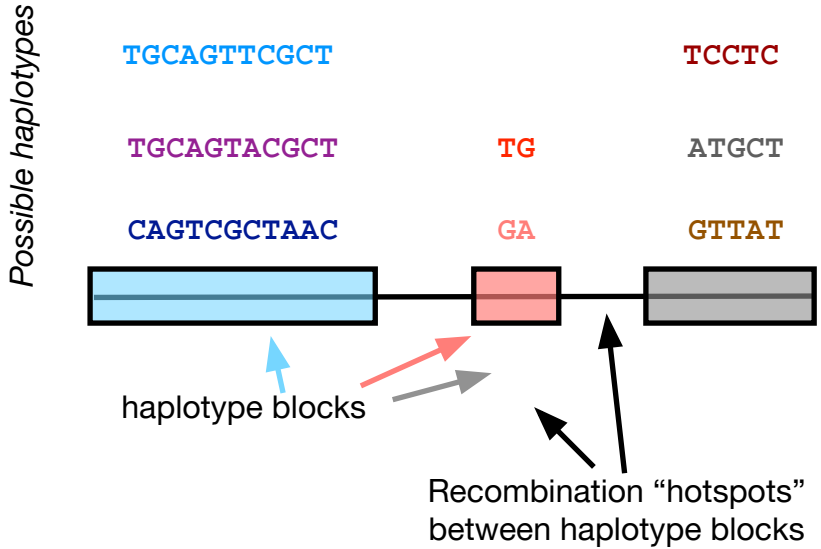


SNP pairs (independent of those used to define haplotype block boundaries) tend to have consistent measures of LD even as distance between them increases.

What does this imply about where recombination occurs?

Anatomy of a haplotype block





SNPs with frequency >20% in the given population

Figure 3 A, B

Most haplotype blocks are small

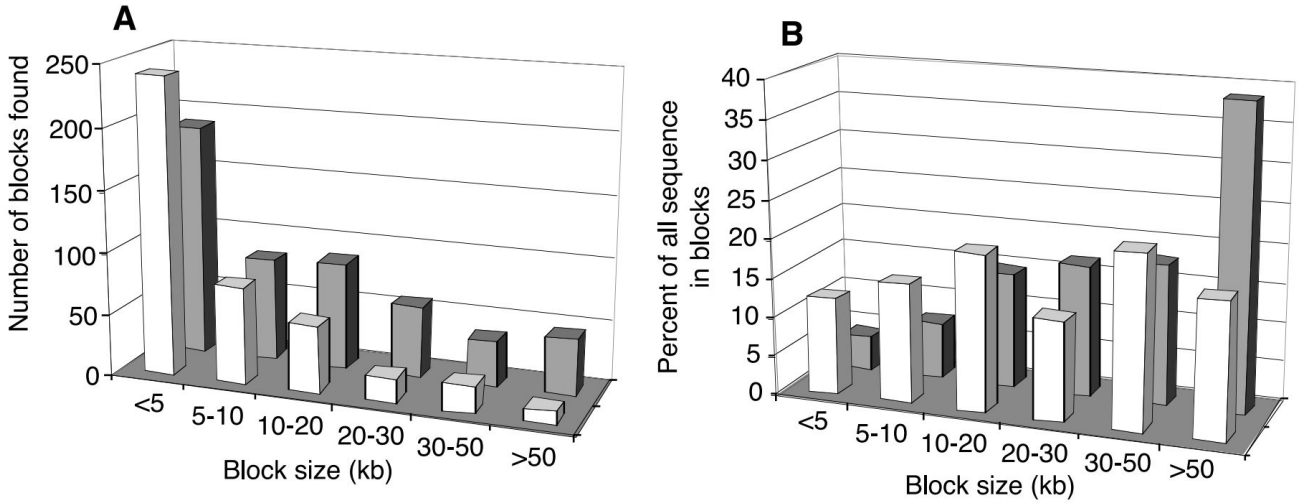
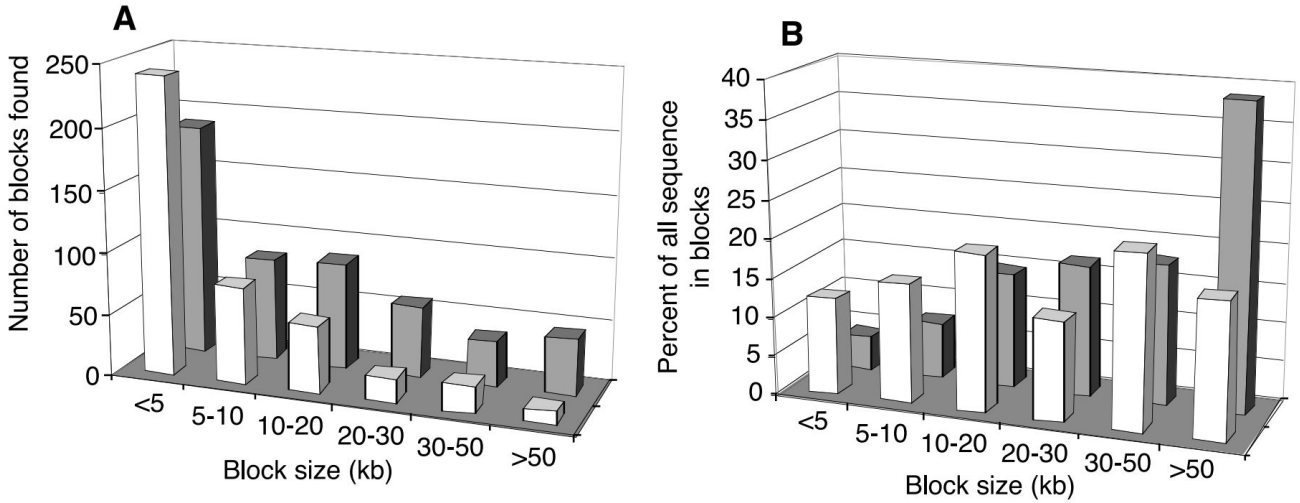


Figure 3 A, B

Most haplotype blocks are small

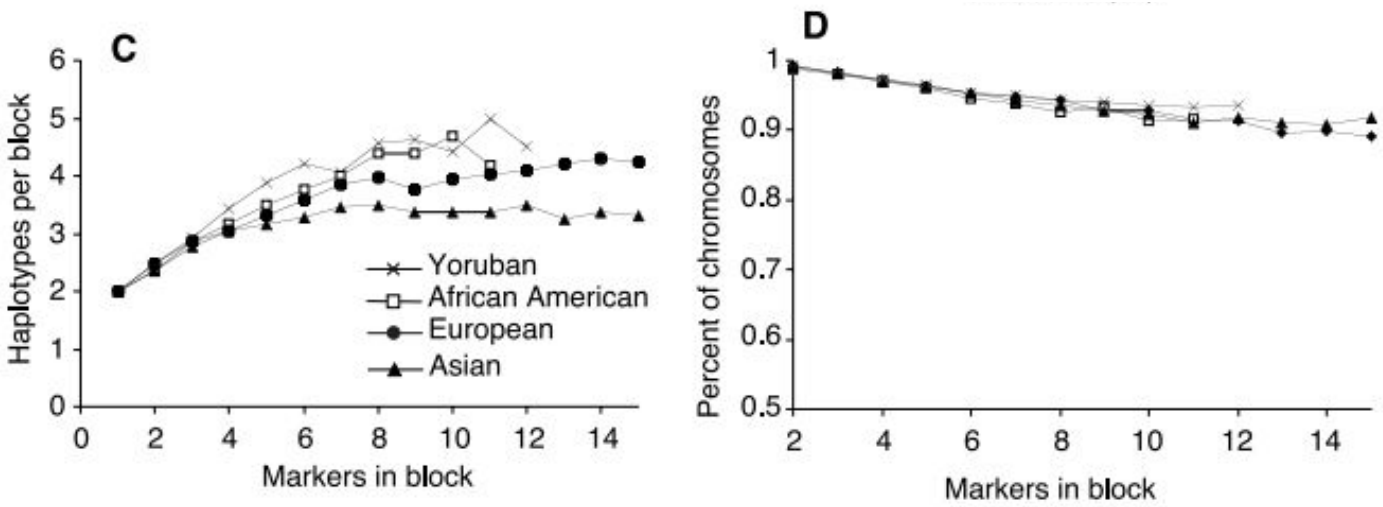


Most haplotype blocks are small, but most sequence is contained in large blocks.

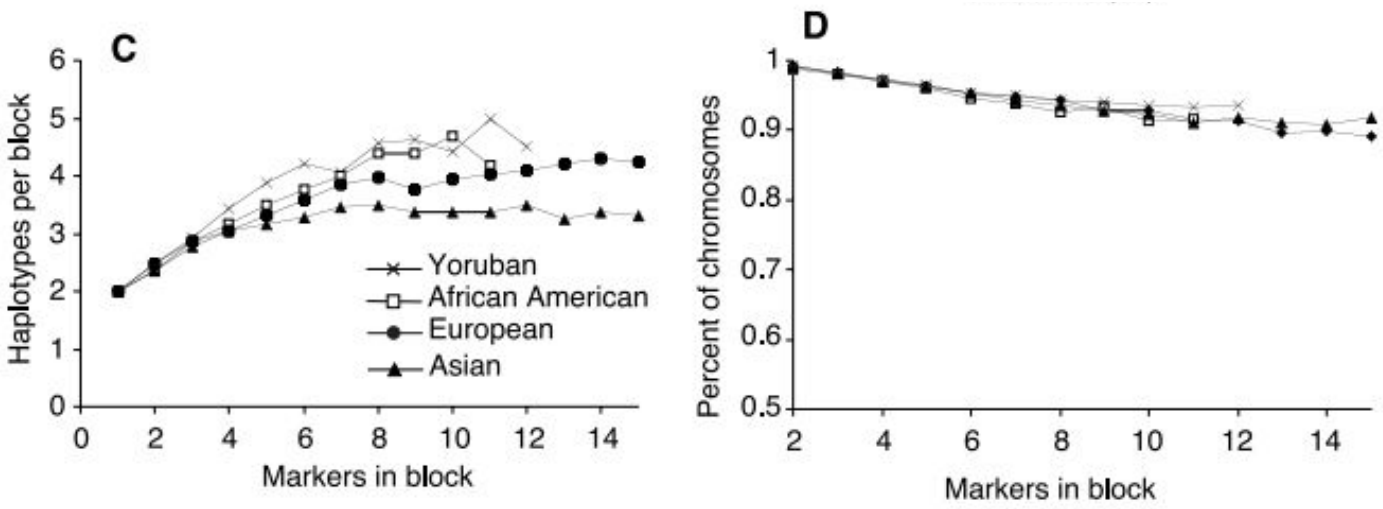
The legend doesn't specify what the colors represent!
Which do you think is Yoruban/African American and which is European/Asian?

Figure 3 C, D

Most haplotype blocks had 3-5 common haplotypes



Most haplotype blocks had 3-5 common haplotypes



Haplotype blocks generally had 3-5 common haplotypes (Africans had more), and most individuals perfectly matched one of the common haplotypes.

Do these figures imply that genotyping more than 14 markers in a block is beneficial?

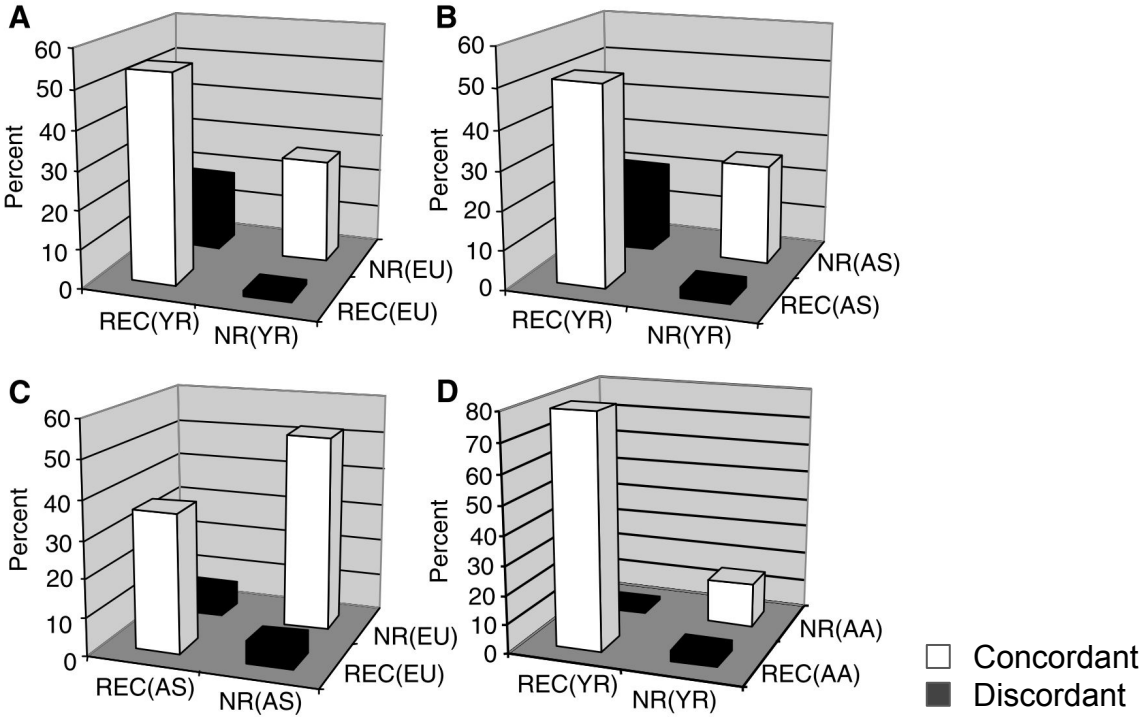
Table 1

Proportion of genome spanned by blocks ≥10kb is 65% (African) to 85% (European/Asian)

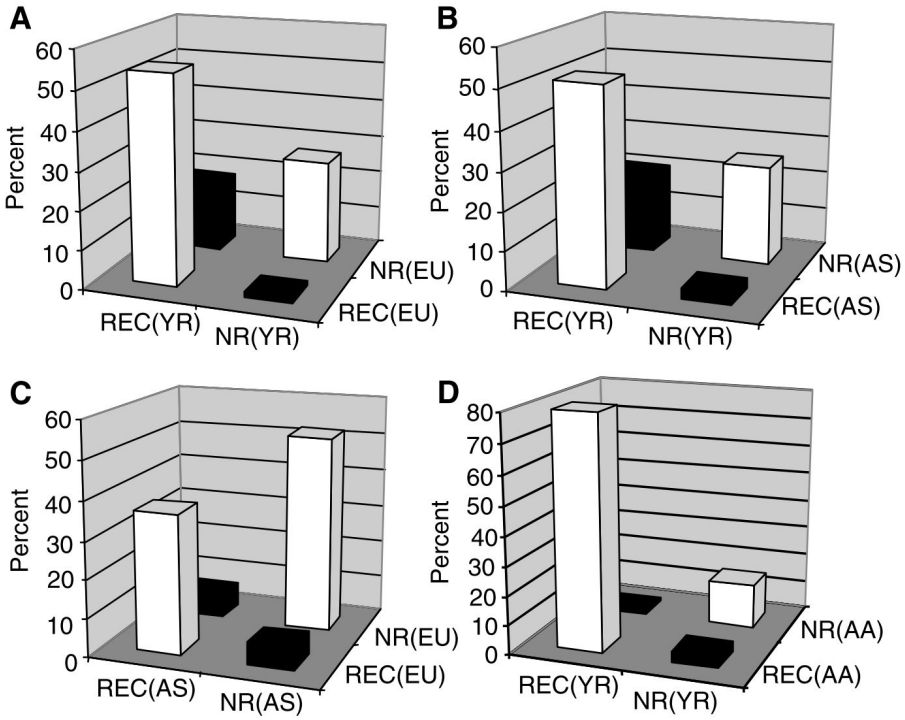
Block size	Yoruban sample		European sample	
	Obs.	Pred.	Obs.	Pred.
0 to 5	12.4	6.3	4.4	1.8
5 to 10	15.3	15.1	7.4	5.2
10 to 20	20	31.5	14.9	15.2
20 to 30	12.8	21.8	16.6	16.6
30 to 50	22.2	19.1	18	26.9
>50	17.4	6.3	38.7	34.2

Number of small blocks likely underestimated (less likely to detect with few SNPs).
Block sizes might also be underestimated (likely extend beyond known boundary SNPs).
Simulations fit observed data with mean block size 11kb (Yoruban) or 22kb (European).

Haplotype block boundaries often shared across populations



Haplotype block boundaries often shared across populations



Examined adjacent SNPs:

- **Concordant**: in same block in two populations
- **Discordant**: in different blocks

77-95% SNP pairs between any two populations are concordant

Why is there more discordance when comparing Africans to Europeans or Asians?

Figure 4 E

Haplotypes often shared across populations

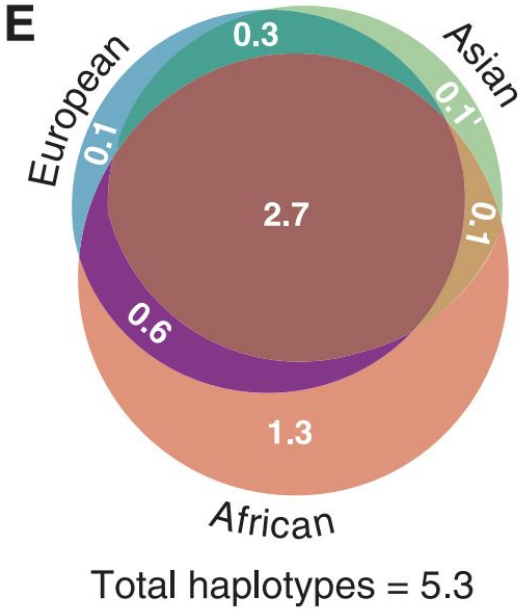
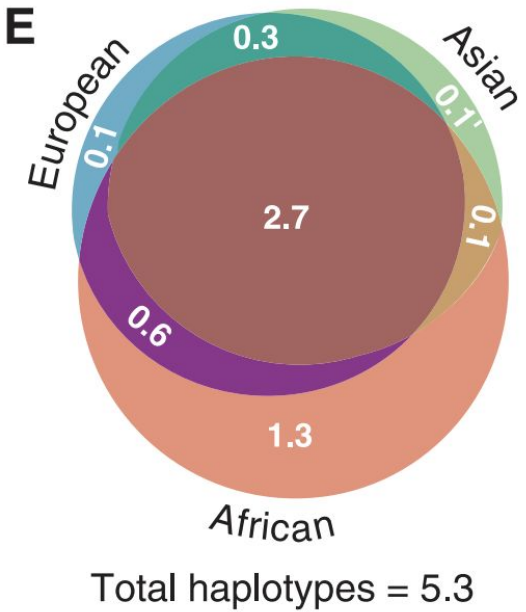


Figure 4 E

Haplotypes often shared across populations



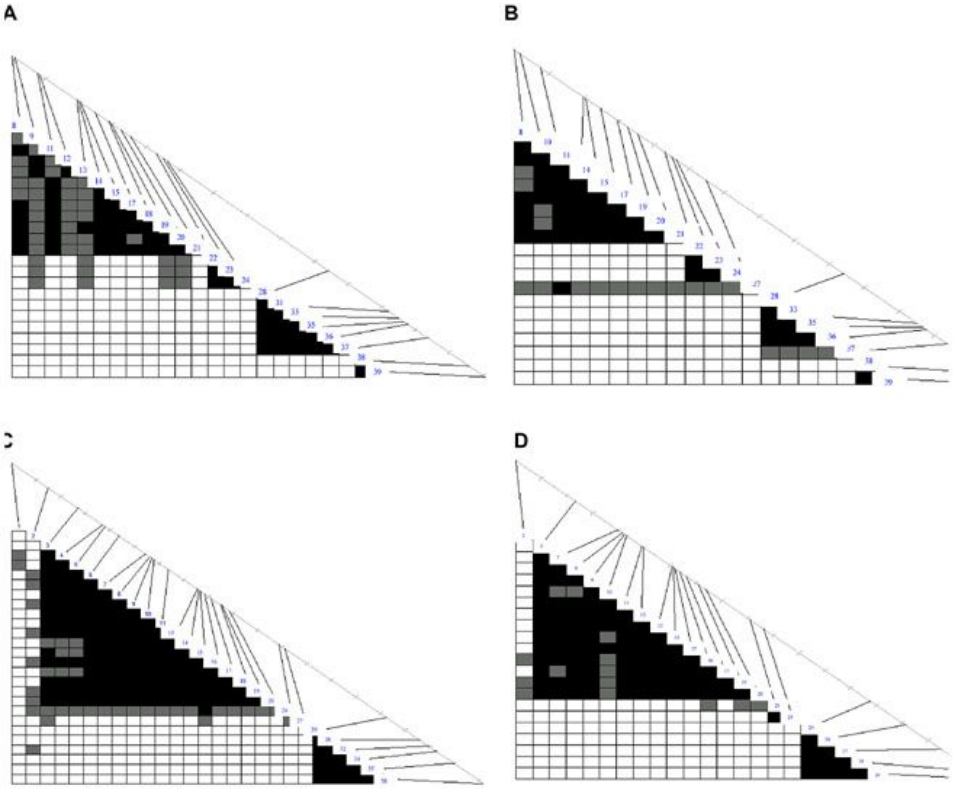
*Haplotypes with a frequency >5%

Looked at haplotypes from blocks with ≥ 6 markers

- Each individual population had an average of 3.1-4.9 haplotypes
- Union of sets had only 5.3 average haplotypes (remarkable similarity across populations)
- 51% of haplotypes observed in all 3 populations
- Average of 0.1 haplotypes unique to either Europeans or Asians, 1.3 to Africans

90% of haplotypes that were only found in one population were in the Yoruban sample.
How does this relate to the Out of Africa theory?

Haplotype blocks across four different populations

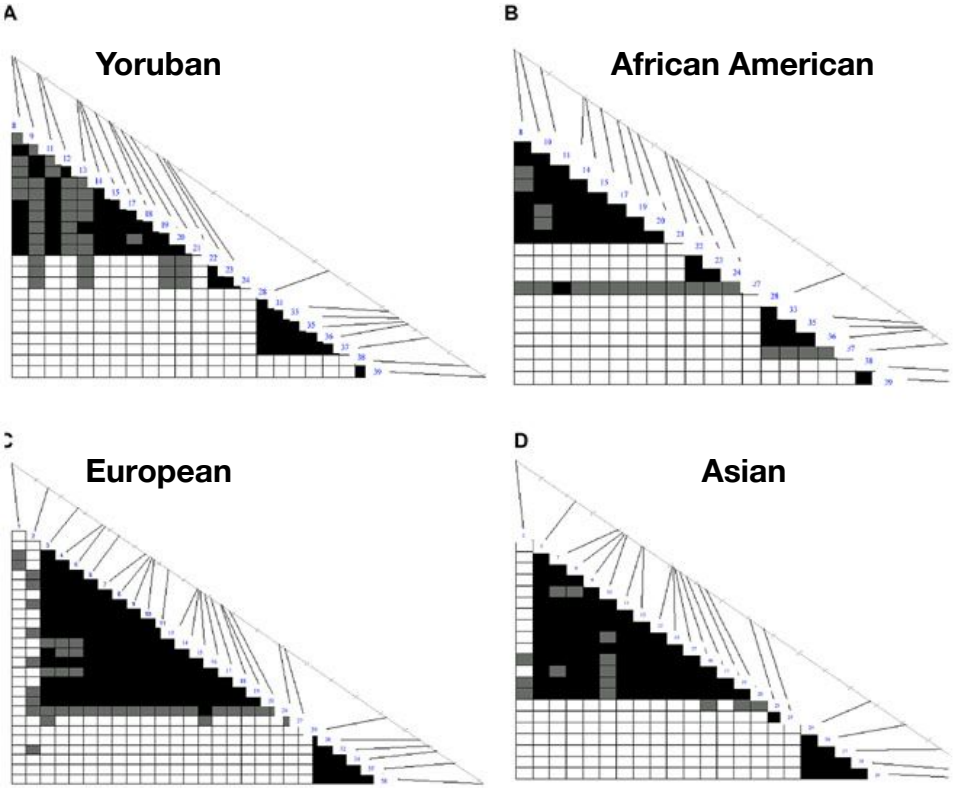


Examine:

- The number of haplotype blocks in each sample
- The size of the blocks
- The concordance of block boundaries

Which images correspond with the Yoruban/African American populations? With the European/Asian populations?

Haplotype blocks across four different populations



Examine:

- The number of haplotype blocks in each sample
- The size of the blocks
- The concordance of block boundaries

Yoruban/African American graphs have more blocks and the blocks are smaller.

Take home messages

The average size of a haplotype block is 11-22 kb

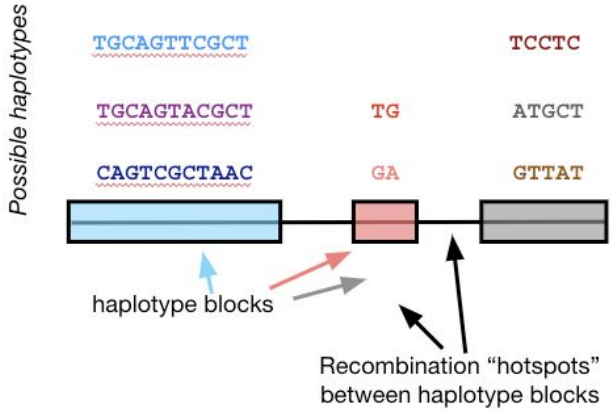
Only a few informative SNPs are needed to define blocks

Higher genetic diversity is found among African populations

- smaller haplotype blocks with more possible haplotypes within the blocks

Variation is shared across populations

- 51% of haplotypes are found in all populations, 72% are found in 2 of 3 populations



Why does it matter?

Why does it matter?

Described a lot of valuable information regarding haplotype diversity and haplotype block sizes

Helped provide a framework for larger haplotype block projects (HapMap Project).

If we know all the haplotypes, we don’t have to genotype every SNP in haplotype block, and can instead **infer** SNPs by genotyping a set of tag SNPs and **impute** additional SNPs as needed.

Assignment: Investigating linkage disequilibrium in different populations

Two parts:

- Group worksheet (link in solo worksheet)
- Solo worksheet

Useful Terminology
Please put your initials next to one term and fill out the description with a sentence or two.

Initials	Term	Description
	allele	
	biallelic SNP	
	D'	
	genetic marker	
	genotyping	
	haplotype	
	haplotype block	
	linkage disequilibrium (LD)	
	minor allele frequency (MAF)	
	out of Africa	
	population bottleneck	
	r^2	

SNP Information
From the following list of SNPs (same as those in the txt files provided), please select one and search the dbSNP database to fill out the appropriate information.

Initials	Block	SNP ID	Chromosome & Position	Alleles	MAF (TOPMED)	Gene : Consequence
	1	rs1325809				
		rs1998676				
		rs111679233				
		rs79888670				
		rs79192154				
		rs142261558				
	2	rs76370881				
		rs143161038				
		rs11571743				
		rs9567609				
		rs61946969				
		rs9534318				