Linkage Disequilibrium

Biostatistics 666

Last Lecture

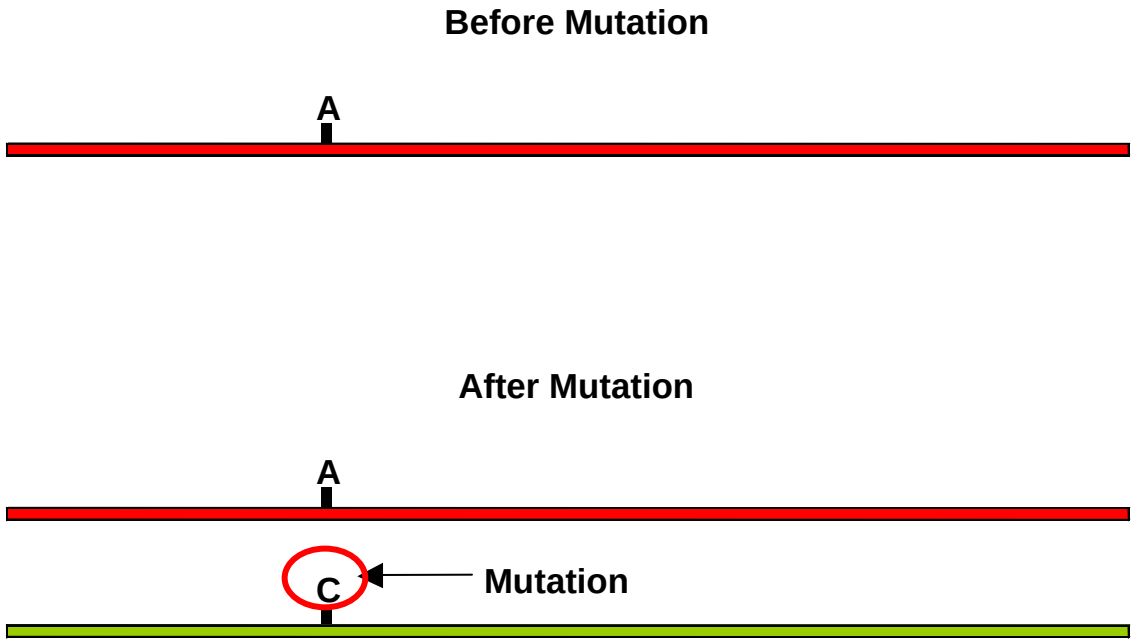
- Basic properties of a locus
  - Allele Frequencies
  - Genotype Frequencies
- Hardy-Weinberg Equilibrium
  - Relationship between allele and genotype frequencies that holds for most genetic markers
- Exact Tests for Hardy-Weinberg Equilibrium

Today …

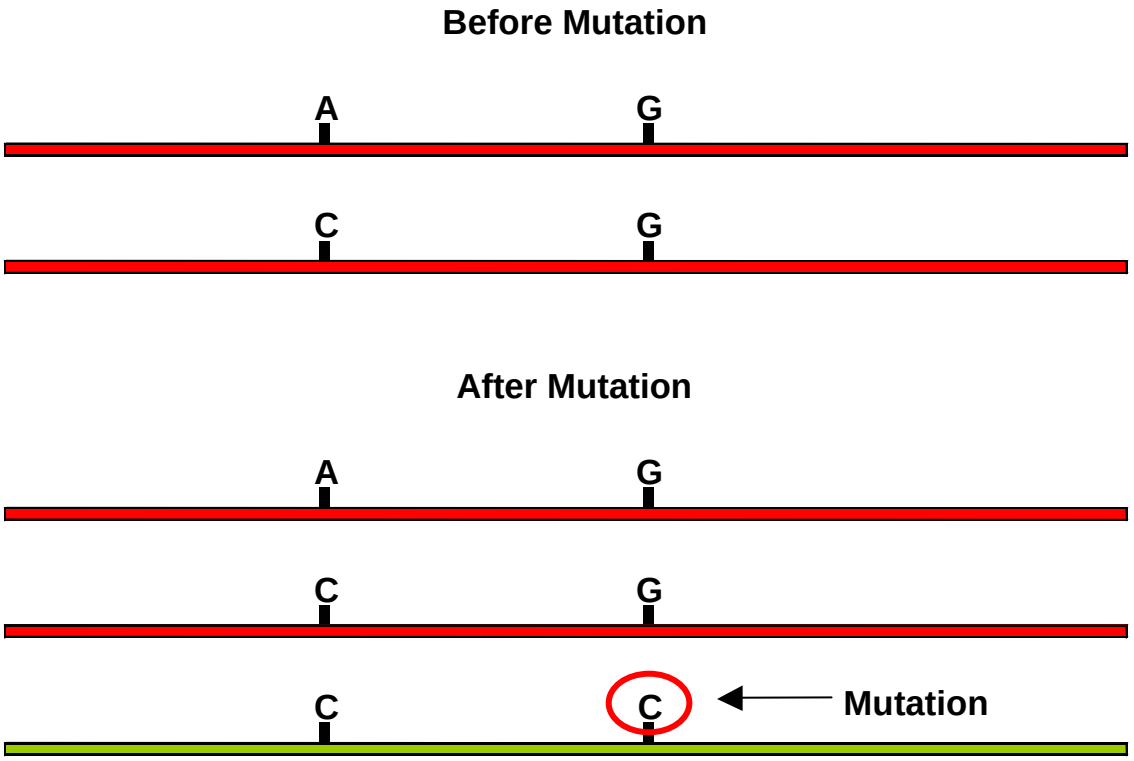
- We'll consider properties of pairs of alleles
- Haplotype frequencies
- Linkage equilibrium
- Linkage disequilibrium

Let's consider the history of two neighboring alleles…

Alleles that exist today arose through ancient mutation events...



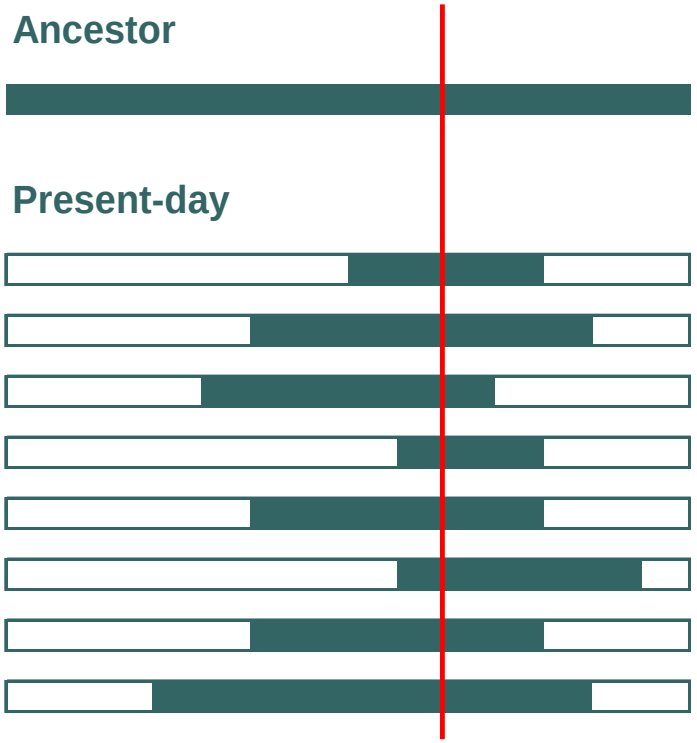
One allele arose first, and then the other...



Recombination generates new arrangements for ancestral alleles

Linkage Disequilibrium

- Chromosomes are mosaics
- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection
- Combinations of alleles at very close markers reflect ancestral haplotypes



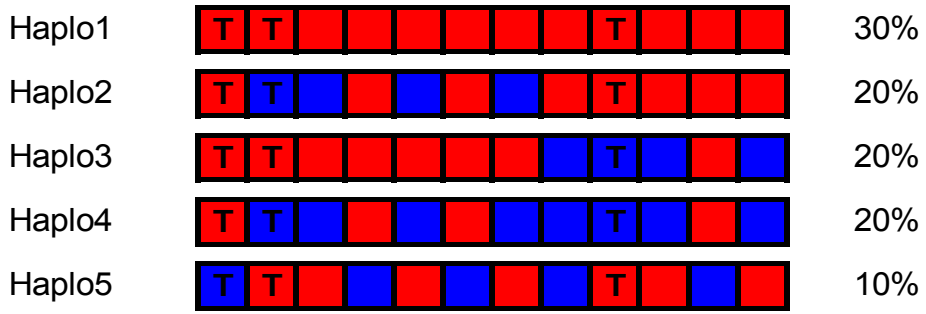
Why is linkage disequilibrium important for gene mapping?

Association Studies and Linkage Disequilibrium

- If all polymorphisms were independent at the population level, association studies would have to examine every one of them…
- Linkage disequilibrium makes tightly linked variants strongly correlated producing cost savings for association studies

Tagging SNPs

- In a typical short chromosome segment, there are only a few distinct haplotypes
- Carefully selected SNPs can determine status of other SNPs

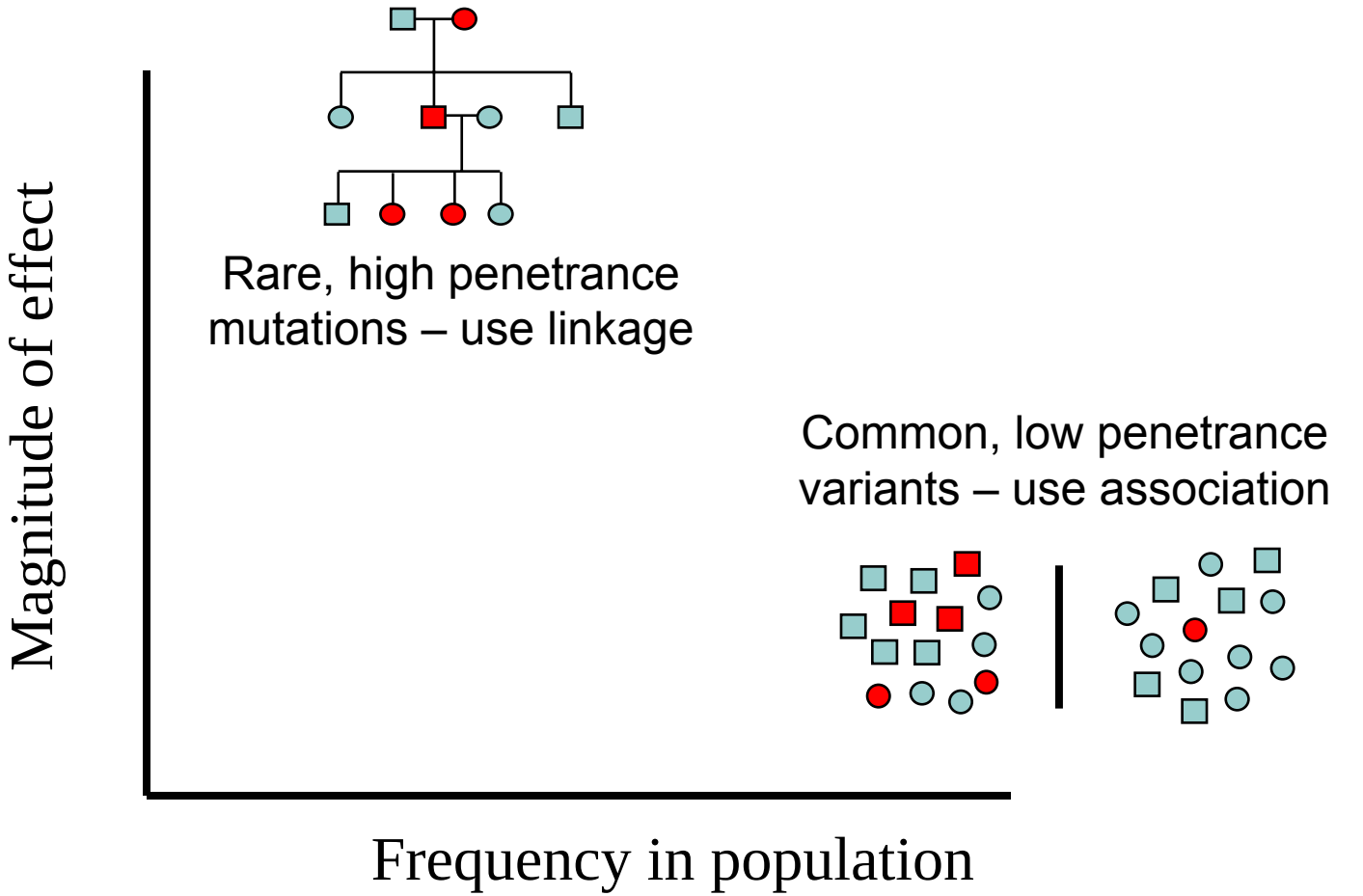


Linkage Disequilibrium Enables Genetic Association Studies

- In contrast to linkage studies, association studies can identify variants with relatively small individual contributions to disease risk
- However, they require detailed measurement of genetic variation and there are >10,000,000 catalogued genetic variants
- Until recently, studies limited to candidate genes or regions
  - A hit-and-miss approach…
- Because assay costs are decreasing and a modest number of variants can represent all others, genome-wide association studies are now possible.

The Allelic Architecture of Disease

What is it and how do we discover it?



Basic Descriptors of Linkage Disequilibrium

Commonly Used Descriptors

- Haplotype Frequencies
  - The frequency of each type of chromosome
  - Contain all the information provided by other summary measures
- Commonly used summaries
  - D
  - D'
  - r^2 or Δ^2

Haplotype Frequencies

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

Linkage Equilibrium
Expected for Distant Loci

$$p_{AB} = p_A p_B$$

$$p_{Ab} = p_A p_b = p_A(1 - p_B)$$

$$p_{aB} = p_a p_B = (1 - p_A) p_B$$

$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B)$$

Linkage Disequilibrium Expected for Nearby Loci

$$p_{AB} \neq p_A p_B$$

$$p_{Ab} \neq p_A p_b = p_A(1 - p_B)$$

$$p_{aB} \neq p_a p_B = (1 - p_A) p_B$$

$$p_{ab} \neq p_a p_b = (1 - p_A)(1 - p_B)$$

Disequilibrium Coefficient D_{AB}

$$D_{AB} = p_{AB} - p_A p_B$$

$$p_{AB} = p_A p_B + D_{AB}$$

$$p_{Ab} = p_A p_b - D_{AB}$$

$$p_{aB} = p_a p_B - D_{AB}$$

$$p_{ab} = p_a p_b + D_{AB}$$

D_{AB} is hard to interpret

- Sign is arbitrary …
  - A common convention is to set A, B to be the common allele and a, b to be the rare allele
- Range depends on allele frequencies
  - Hard to compare between markers

What is the range of D_{AB}?

- What are the maximum and minimum possible values of D_{AB} when
  - $p_A = 0.3$ and $p_B = 0.3$
  - $p_A = 0.2$ and $p_B = 0.1$
- Can you derive a general formula for this range?

D' – A scaled version of D

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

- Ranges between –1 and +1
  - More likely to take extreme values when allele frequencies are small
  - ± 1 implies at least one of the observed haplotypes was not observed

More on D'

- Pluses:
  - D' = 1 or D' = -1 means no evidence for recombination between the markers
  - If allele frequencies are similar, high D' means the markers are good surrogates for each other
- Minuses:
  - D' estimates inflated in small samples
  - D' estimates inflated when one allele is rare

Δ^2 (also called r^2)

$$\Delta^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$$

$$= \frac{\chi^2}{2n}$$

- Ranges between 0 and 1
  - 1 when the two markers provide identical information
  - 0 when they are in perfect equilibrium
- Expected value is $1/2n$

More on r^2

- $r^2 = 1$ implies the markers provide exactly the same information
- The measure preferred by population geneticists
- Measures loss in efficiency when marker A is replaced with marker B in an association study
  - With some simplifying assumptions (e.g. see Pritchard and Przeworski, 2001)

When does
linkage equilibrium hold?

Equilibrium or Disequilibrium?

- We will present simple argument for why linkage equilibrium holds for most loci

- Balance of factors
  - Genetic drift (a function of population size)
  - Random mating
  - Distance between markers
  - …

Why Equilibrium is Reached…

- Eventually, random mating and recombination should ensure that mutations spread from original haplotype to all haplotypes in the population…

- Simple argument:
  - Assume fixed allele frequencies over time

Generation t, Initial Configuration

	B	b	
A	$p_A p_B + D_{AB}$	$p_A p_b - D_{AB}$	p_A
a	$p_a p_B - D_{AB}$	$p_a p_b + D_{AB}$	p_a
	p_B	p_b	

Assume arbitrary values for the allele frequencies and disequilibrium coefficient

Generation t+1, Without Recombination

	B	b	
A	$p_A p_B + D_{AB}$	$p_A p_b - D_{AB}$	p_A
a	$p_a p_B - D_{AB}$	$p_a p_b + D_{AB}$	p_a
	p_B	p_b	

Haplotype Frequencies Remain Stable Over Time
Outcome has probability 1- r

Generation t+1, With Recombination

	B	b	
A	$p_A p_B$	$p_A p_b$	p_A
a	$p_A p_b$	$p_a p_b$	p_a
	p_B	p_b	

Haplotype Frequencies Are Function of Allele Frequencies
Outcome has probability r

Generation t+1, Overall

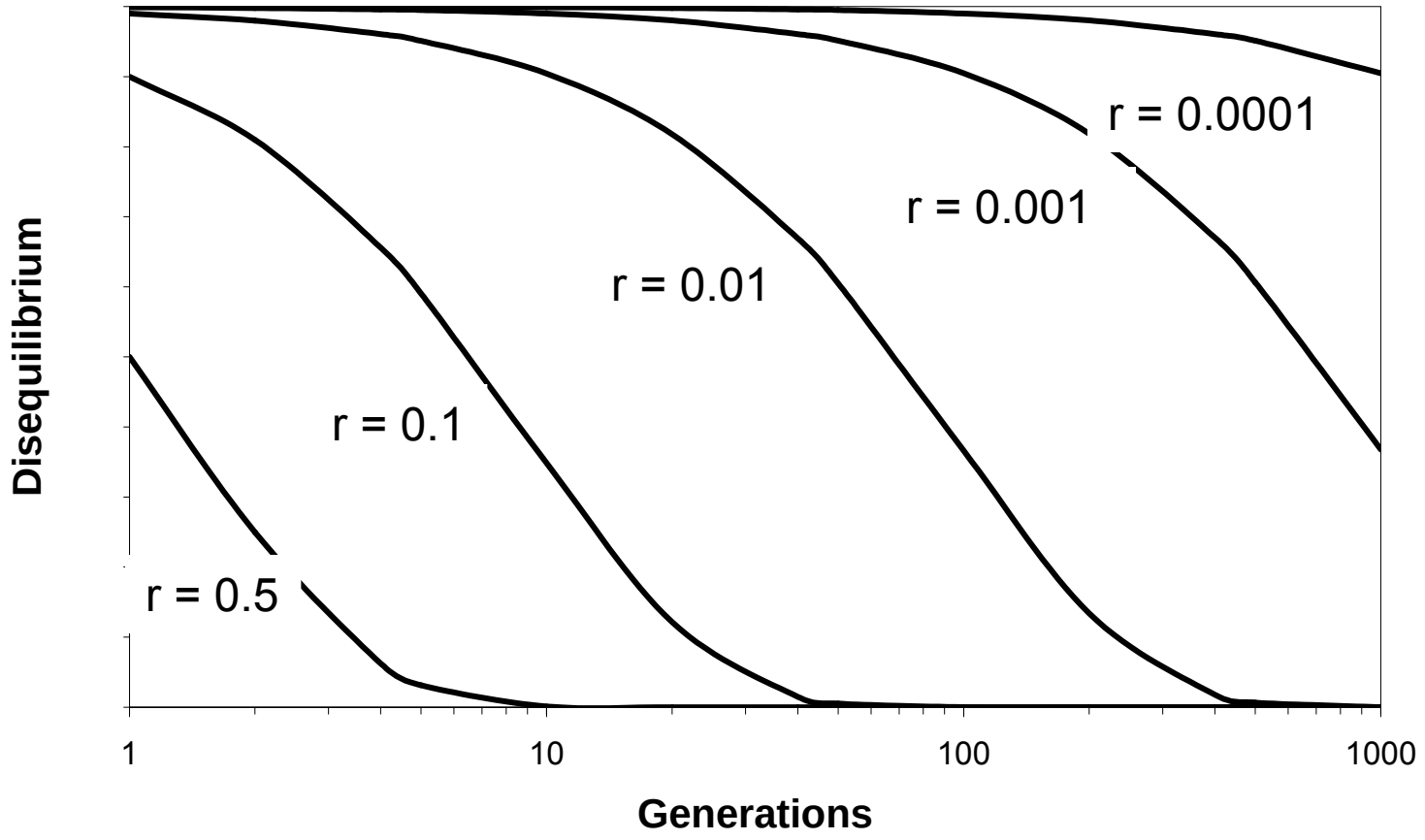
	B	b	
A	$p_A p_B + (1-\theta) D_{AB}$	$p_A p_b - (1-\theta) D_{AB}$	p_A
a	$p_A p_b - (1-\theta) D_{AB}$	$p_a p_b + (1-\theta) D_{AB}$	p_a
	p_B	p_b	

Disequilibrium Decreases...

Recombination Rate (r)

- Probability of an odd number of crossovers between two loci
- Proportion of time alleles from two different grand-parents occur in the same gamete
- Increases with physical (base-pair) distance, but rate of increase varies across genome

Decay of D with Time



Predictions

- Disequilibrium will decay each generation
  - In a large population
- After t generations…
  - $D^t_{AB} = (1-\theta)^t D^0_{AB}$
- A better model should allow for changes in allele frequencies over time…

Linkage Equilibrium

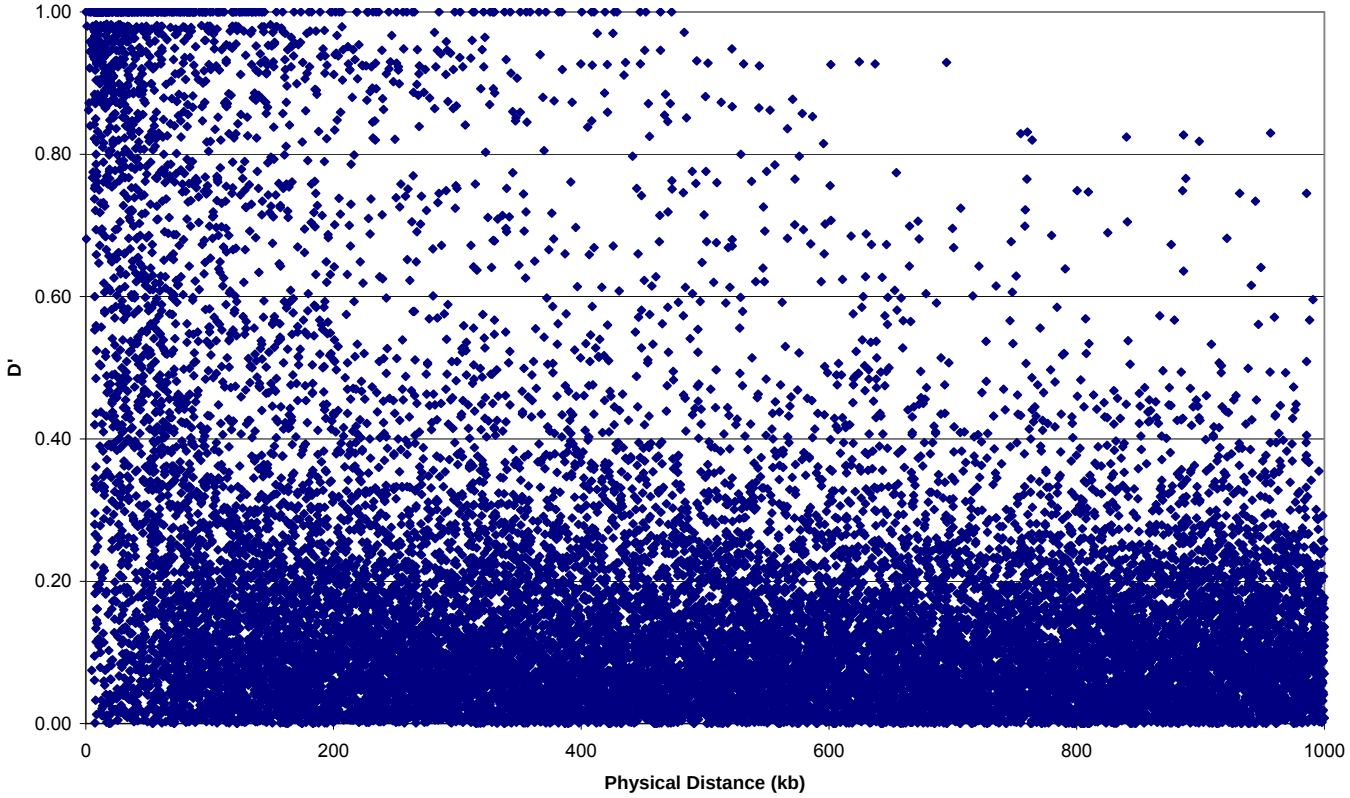
- In a large random mating population haplotype frequencies converge to a simple function of allele frequencies

Some Examples of Linkage Disequilibrium Data

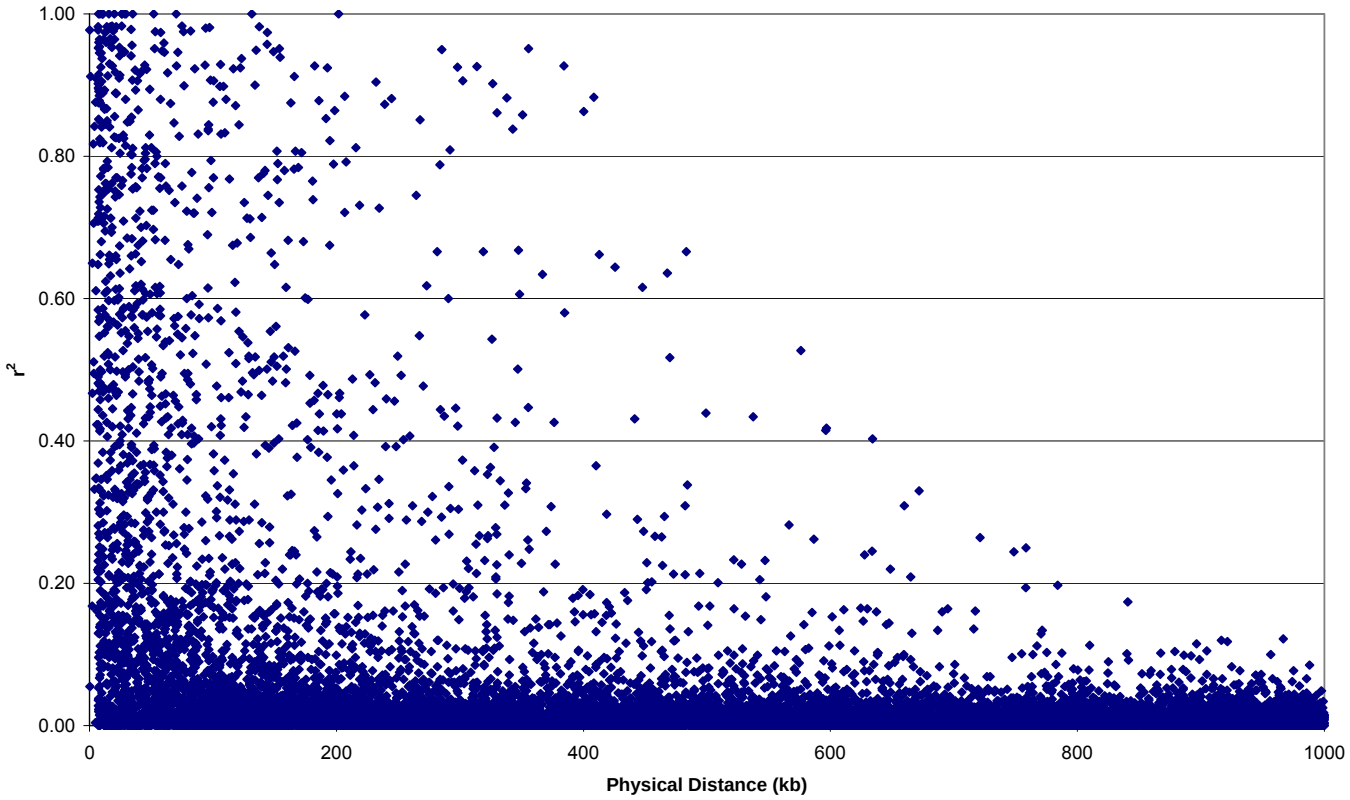
Summary of Disequilibrium in the Genome

- How much disequilibrium is there?
- What are good predictors of disequilibrium?
- What are good predictors of variation in disequilibrium?

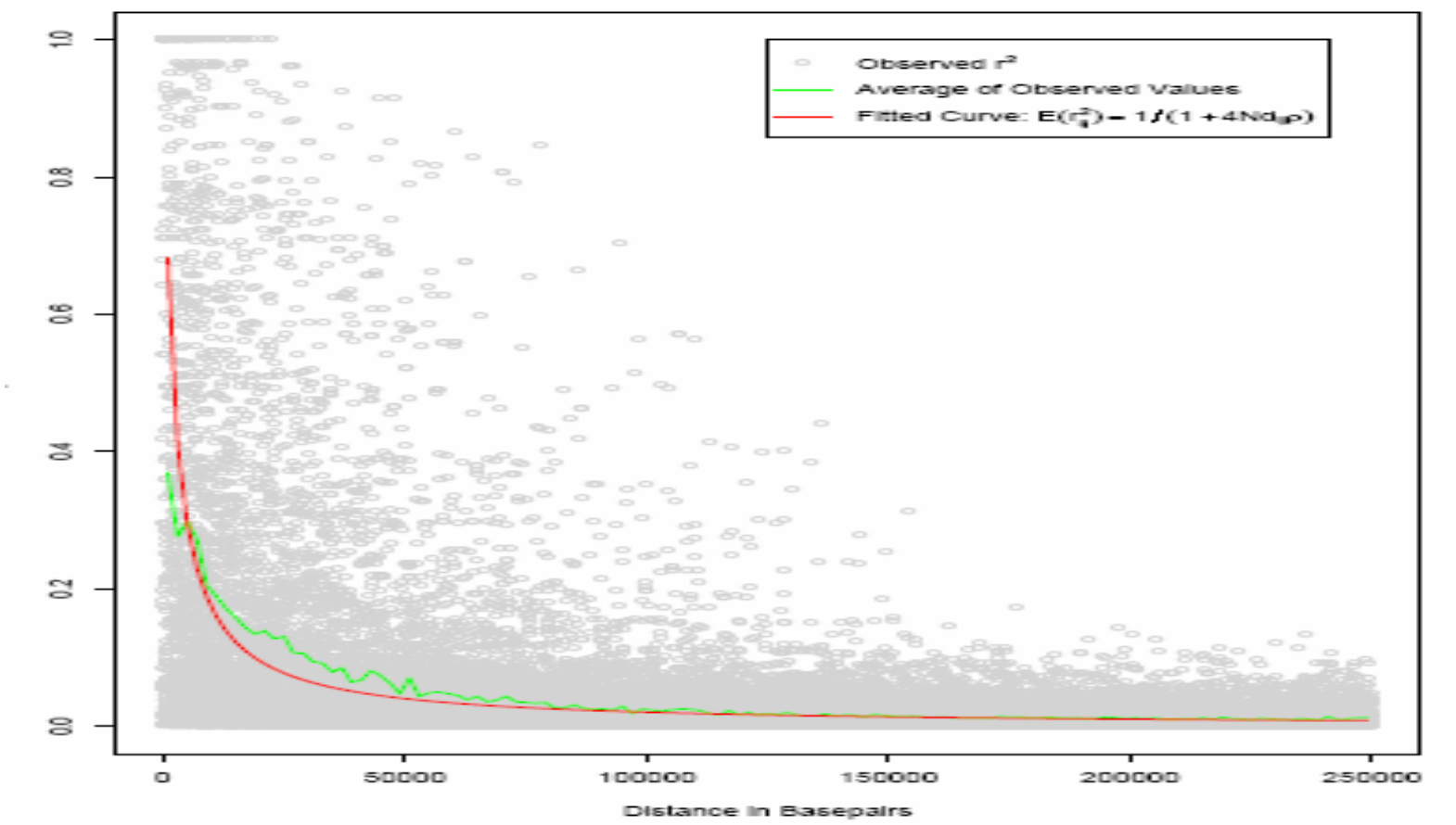
Raw |D'| data from Chr22



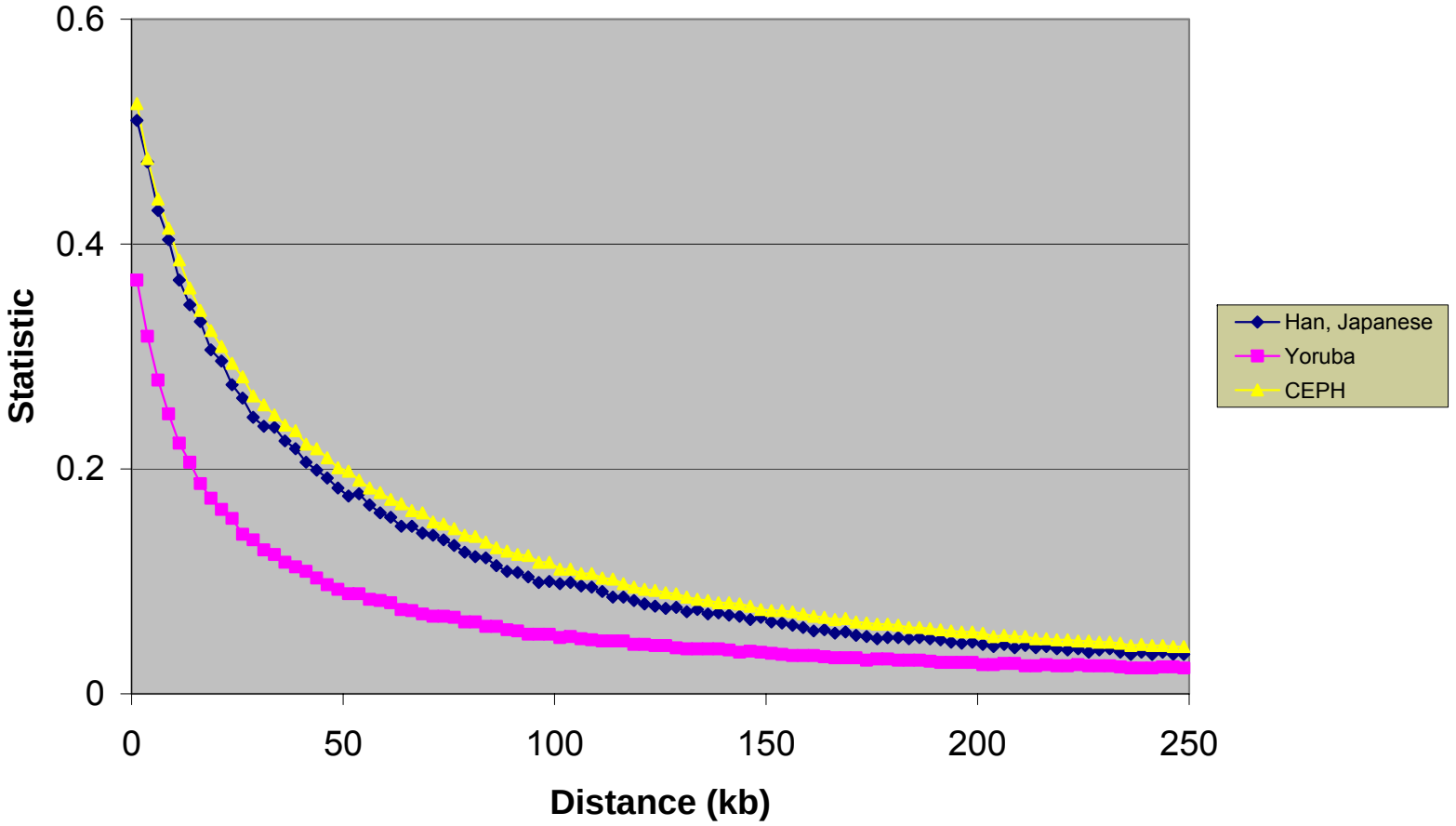
Raw Δ^2 data from Chr22



Summarizing Disequilibrium



Comparing Populations ...

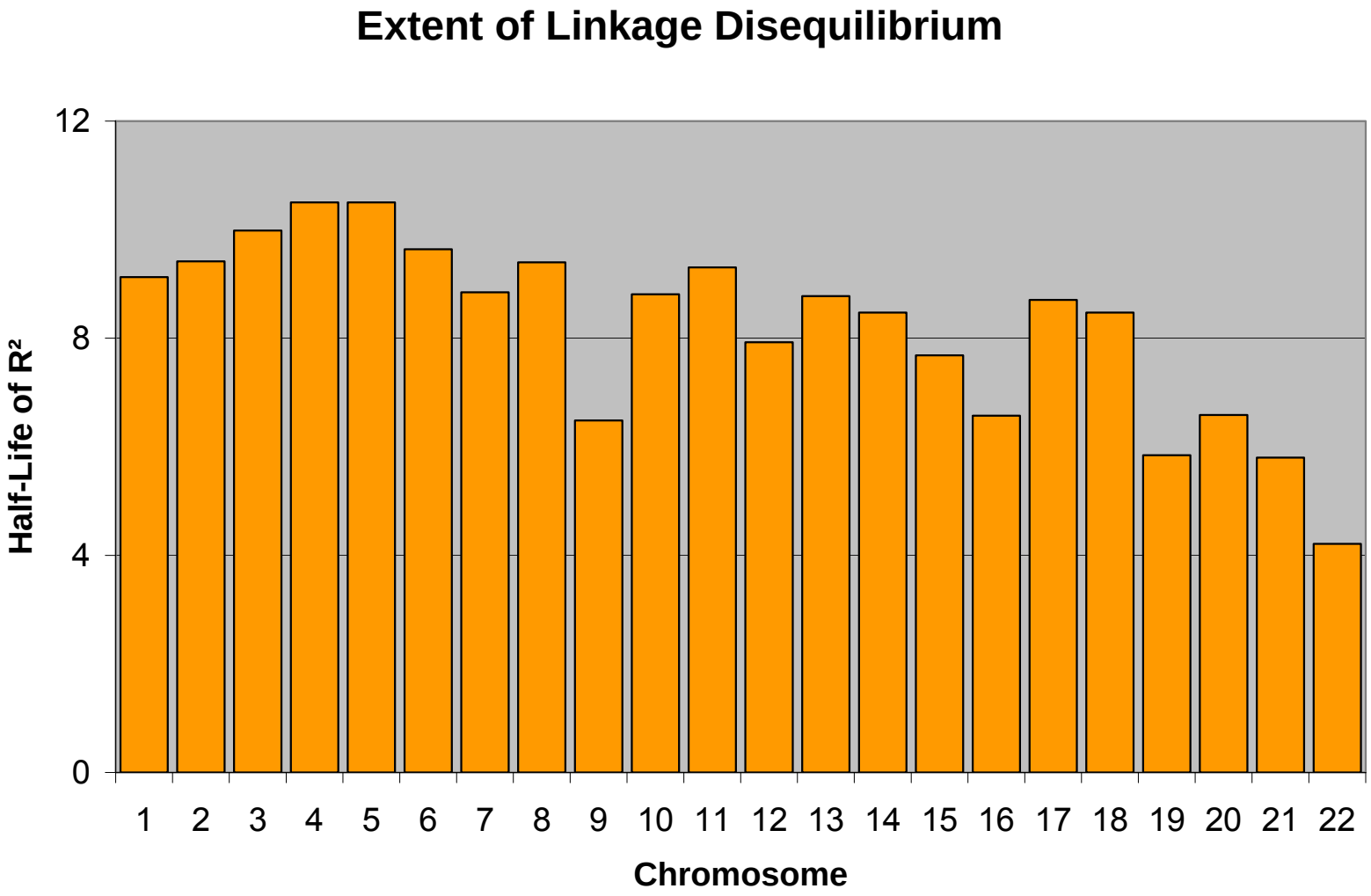


LD extends further in CEPH and the Han/Japanese than in the Yoruba

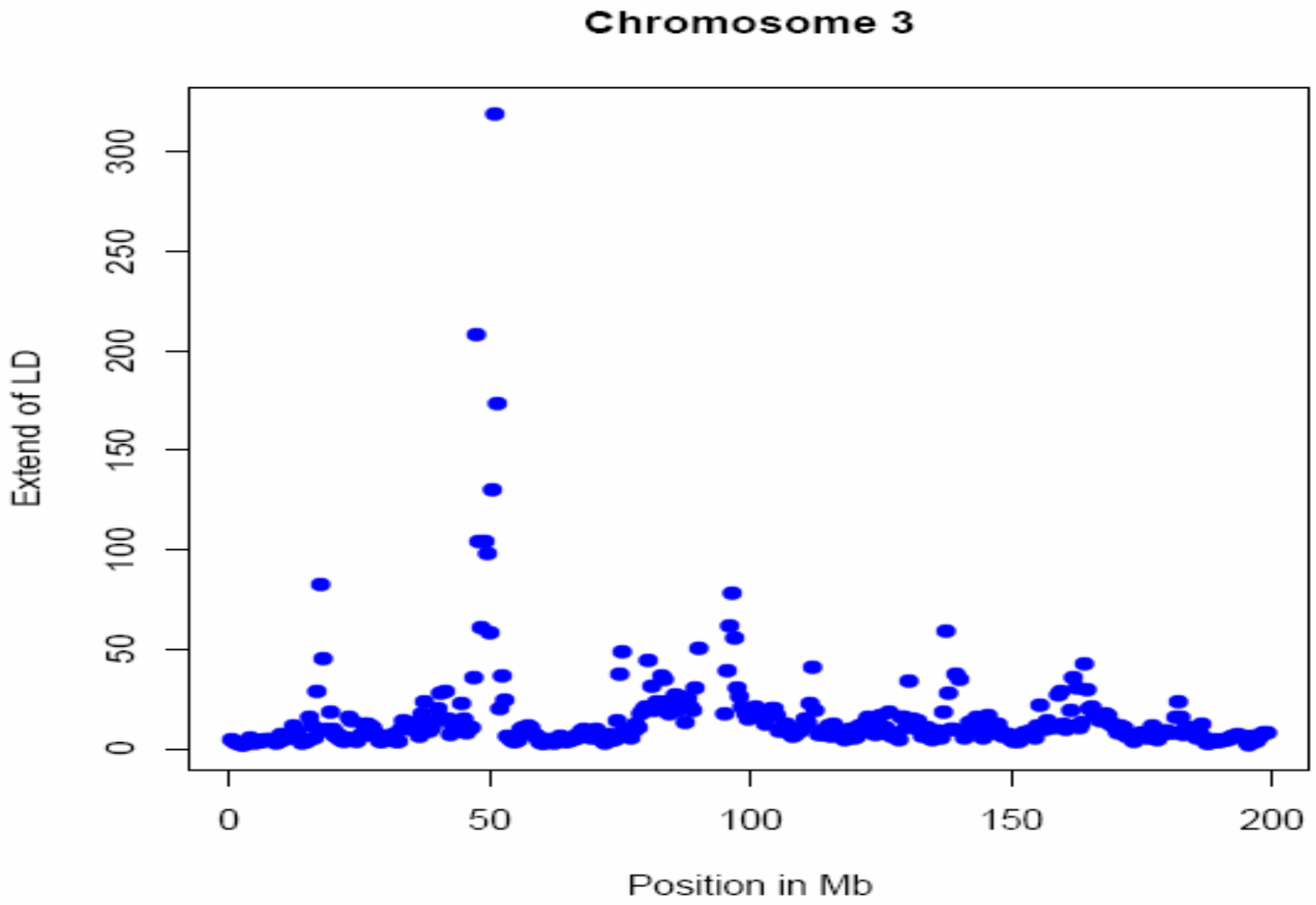
Comparing Genomic Regions ...

- Rather than compare curves directly, it is convenient to a pick a summary for the decay curves

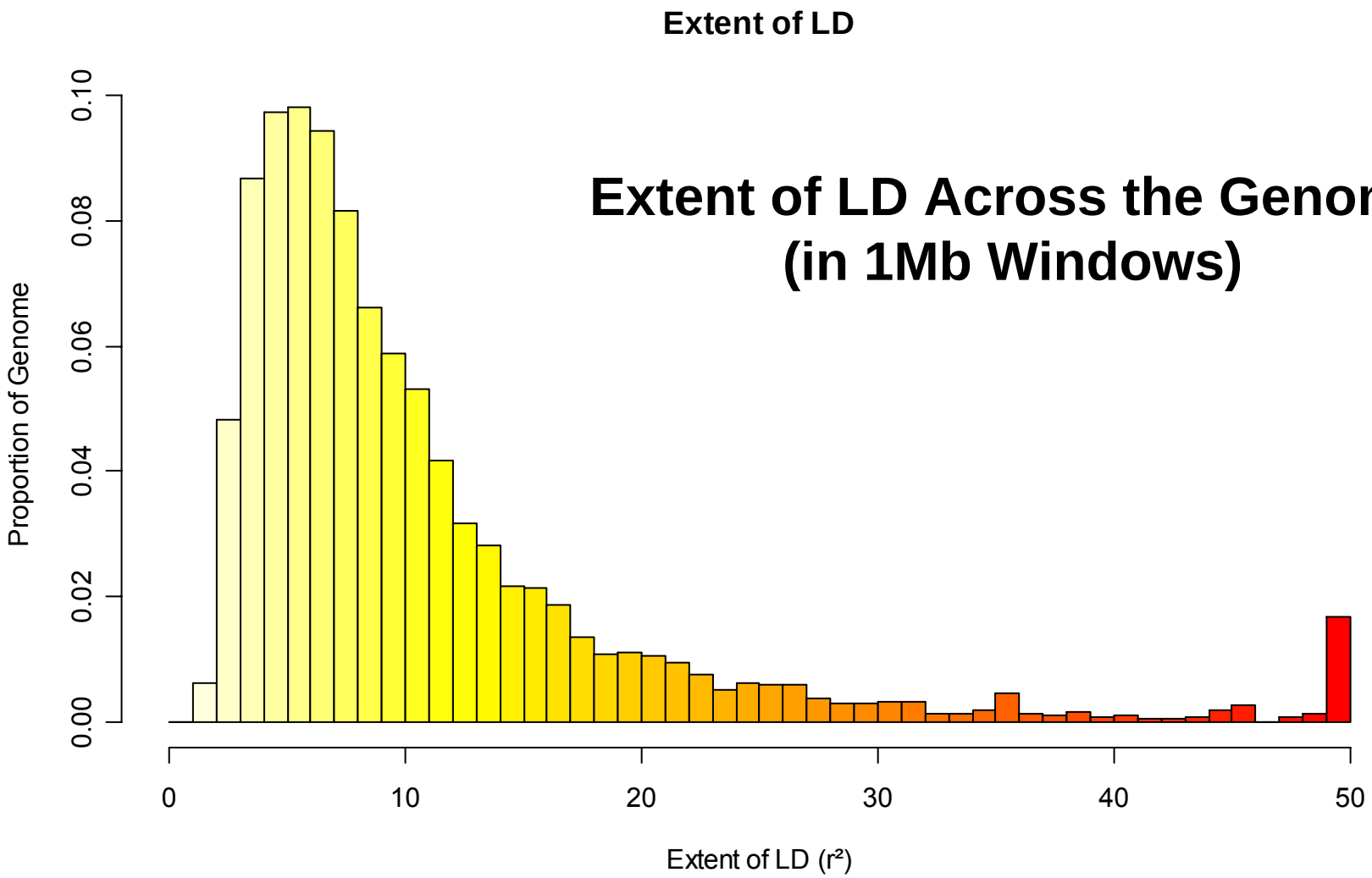
- One common summary is the distance at which the curve crosses a threshold of interest (say 0.50)



LD extends further in the larger chromosomes, which have lower recombination rates

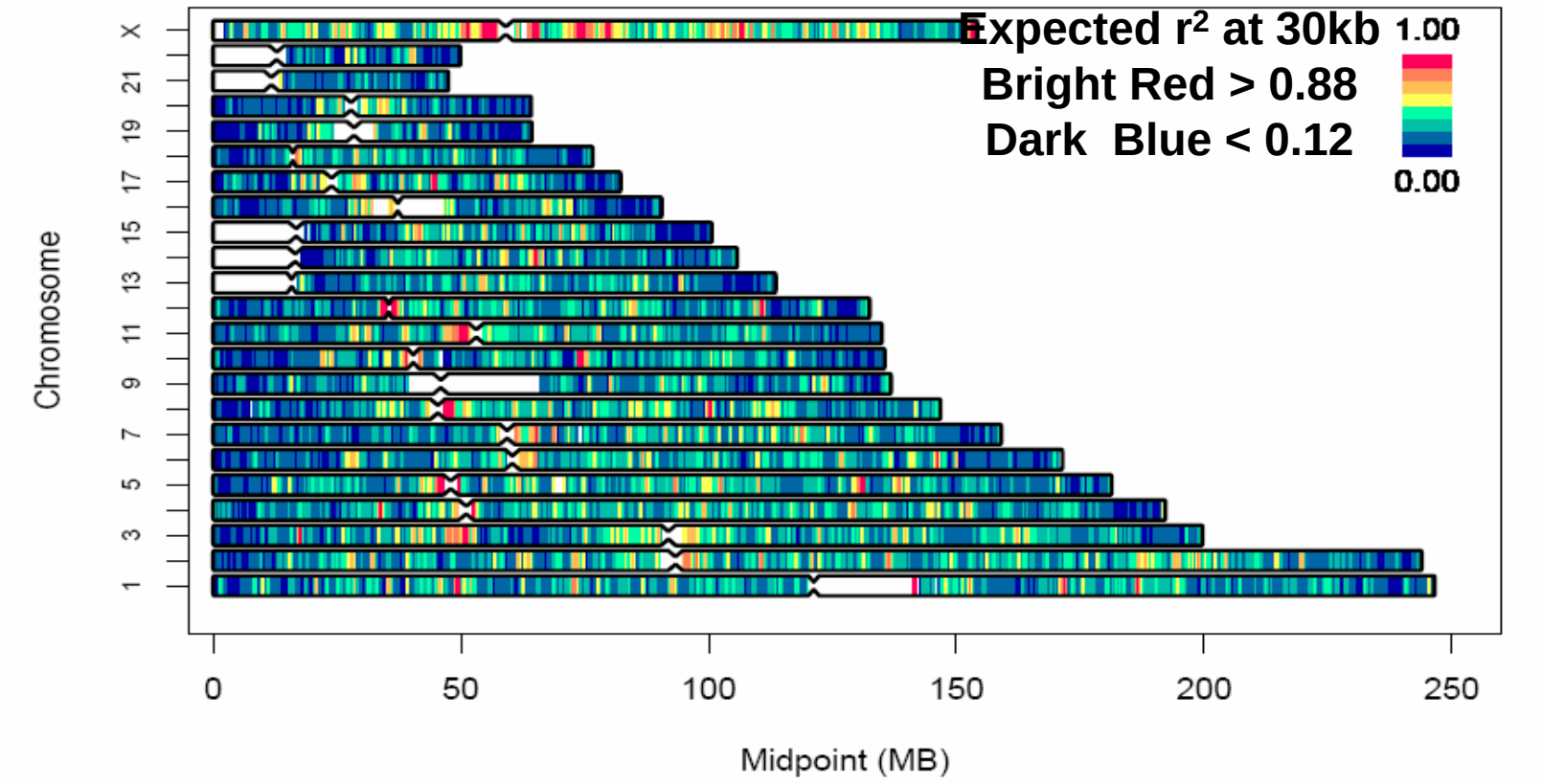


But within each chromosome, there is still huge variability!

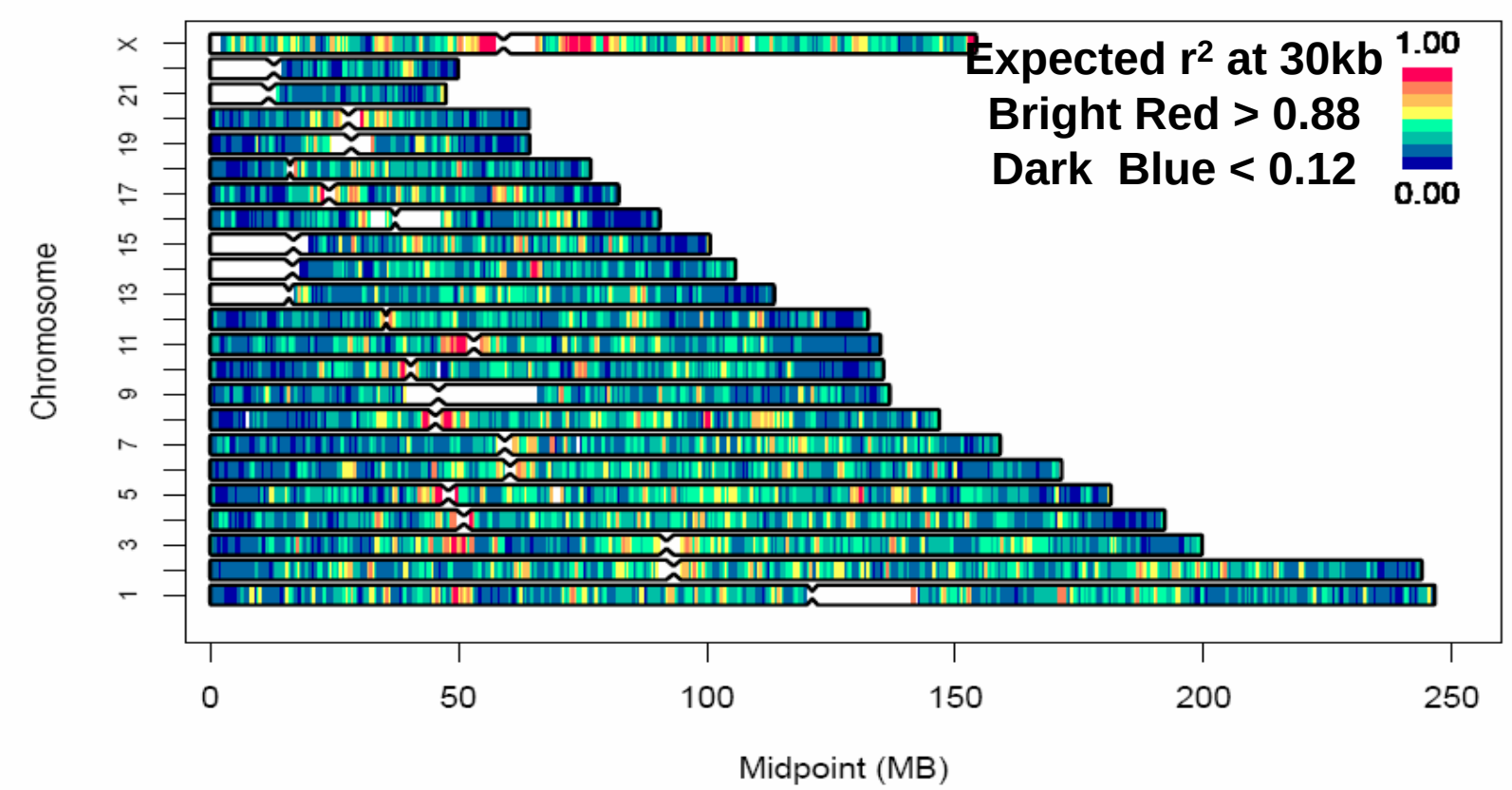


Average Extent:	**11.9 kb**
Median Extent:	**7.8 kb**
10th percentile:	**3.5 kb**
90th percentile:	**20.9 kb**

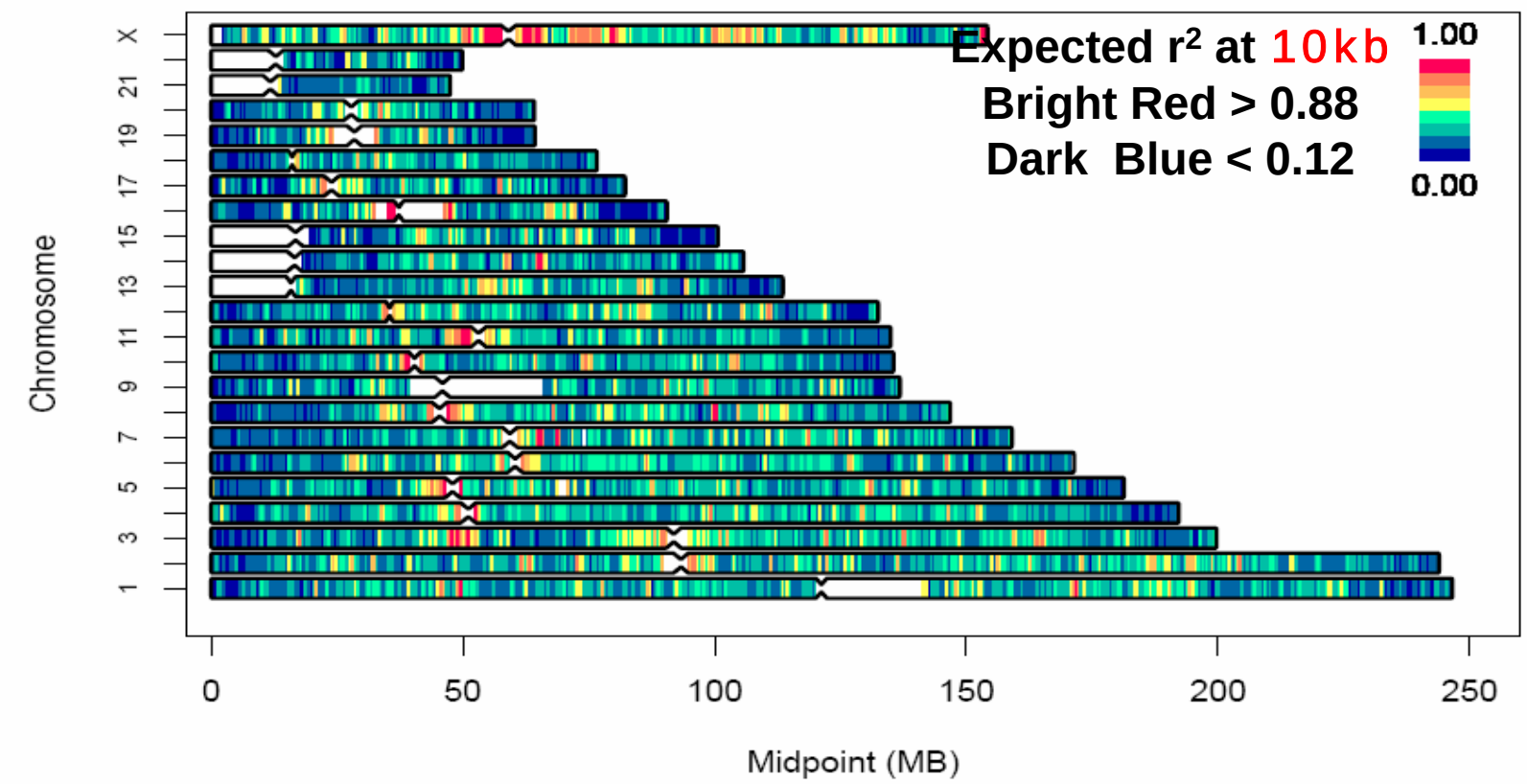
Genomic Variation in LD (CEPH)



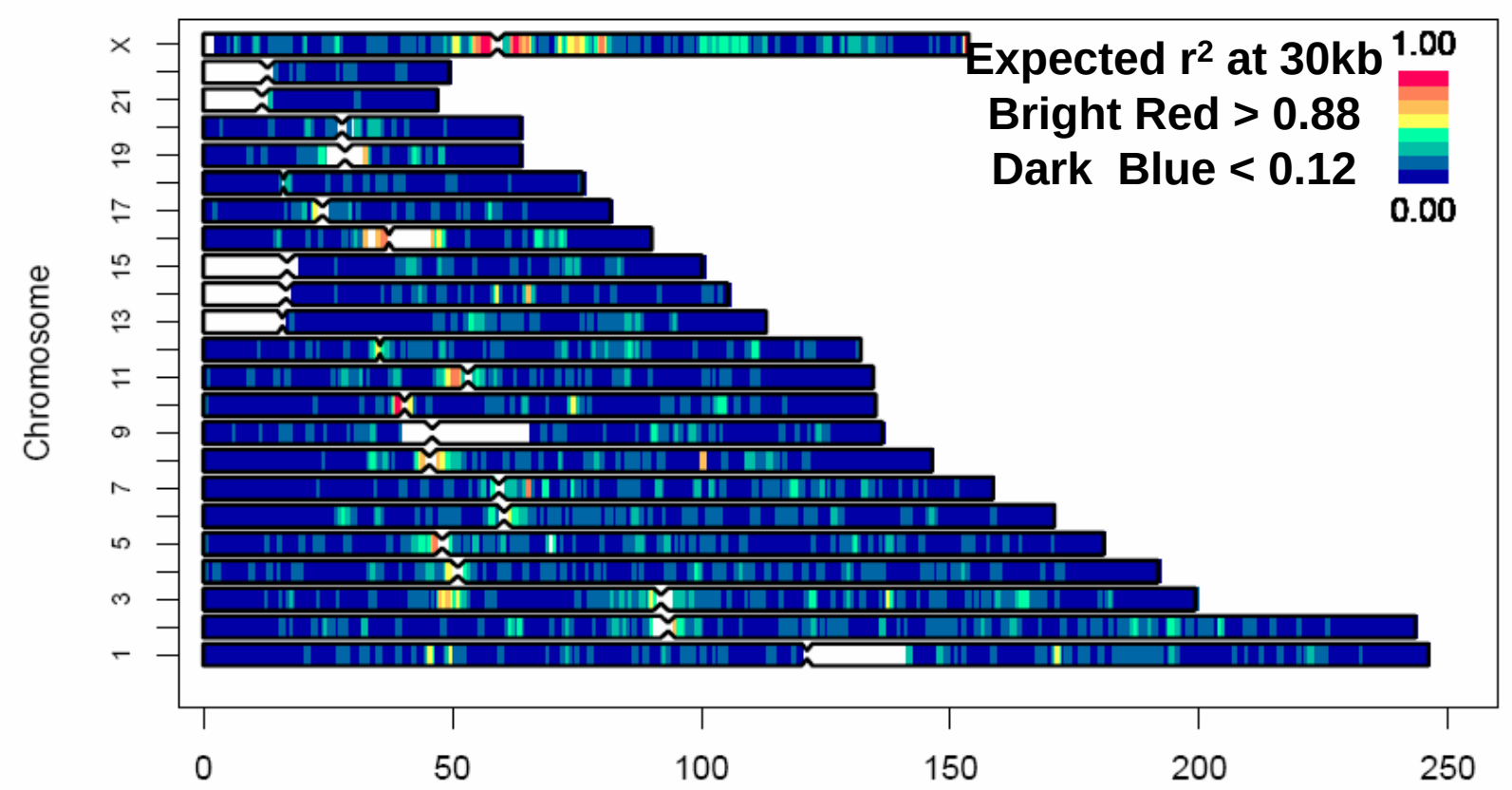
Genomic Variation in LD (JPT + HCB)



Genomic Variation in LD (YRI)



Genomic Distribution of LD (YRI)



What factors might contribute to genomic variation in LD?

Today …

- Basic descriptors of linkage disequilibrium
- Learn when linkage disequilibrium is expected to hold (or not!)