Linkage Disequilibrium

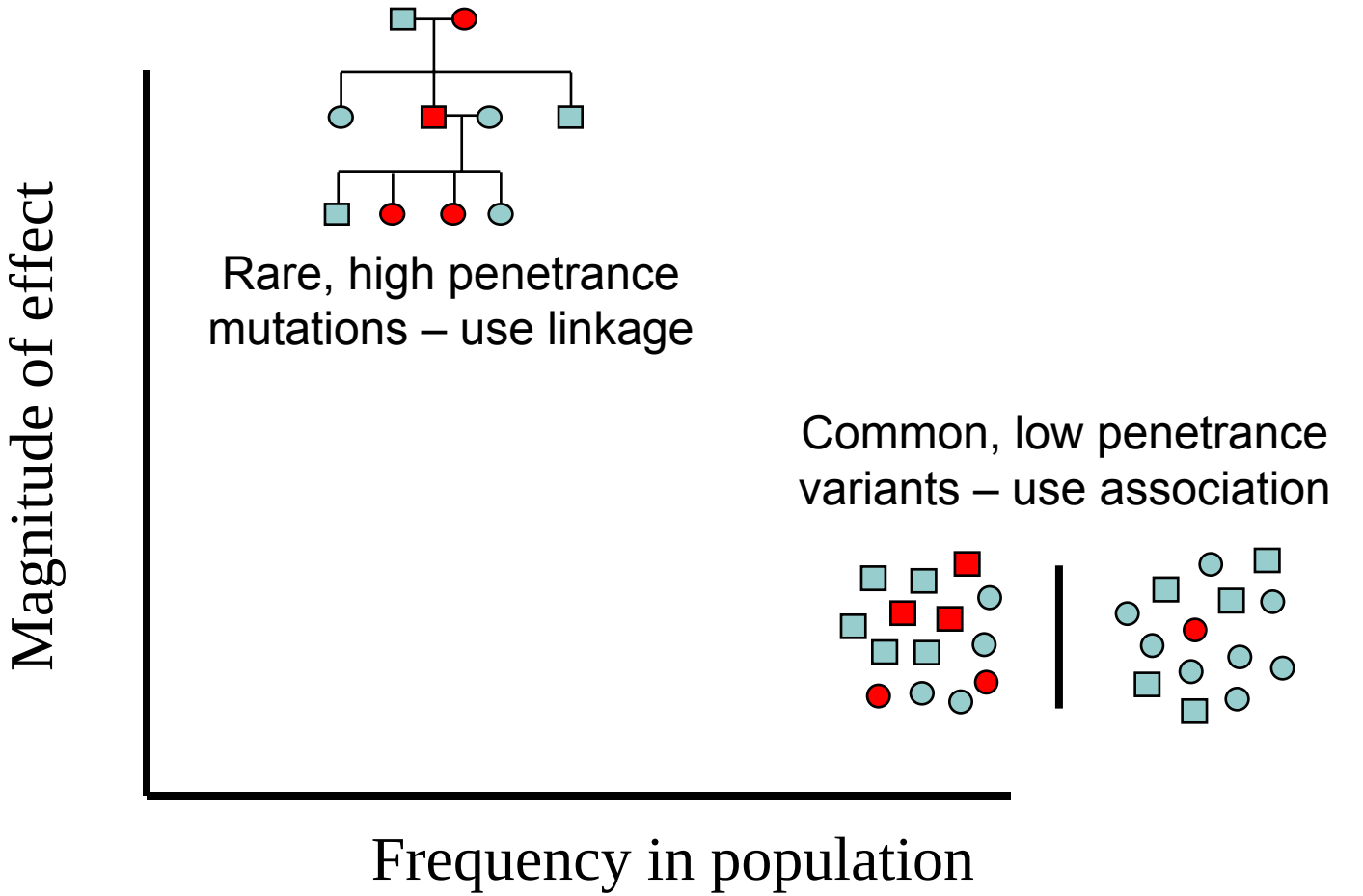
Biostatistics 666

Last Lecture

- Common study designs
- Descriptions of a single locus
  - Allele Frequencies
  - Genotype Frequencies
- Hardy-Weinberg Equilibrium
  - Exact Test for HWE
  - Procedure is analogous to tests for paired data

The Allelic Architecture of Disease

What is it and how do we discover it?



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

Today …

- Descriptors for multiple markers
- Relating allele frequencies at pairs of loci
  - Linkage equilibrium
  - Linkage disequilibrium
- D’ and r^2 statistics

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

$$p_{AB} = p_A p_B$$

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

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

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

Linkage Disequilibrium

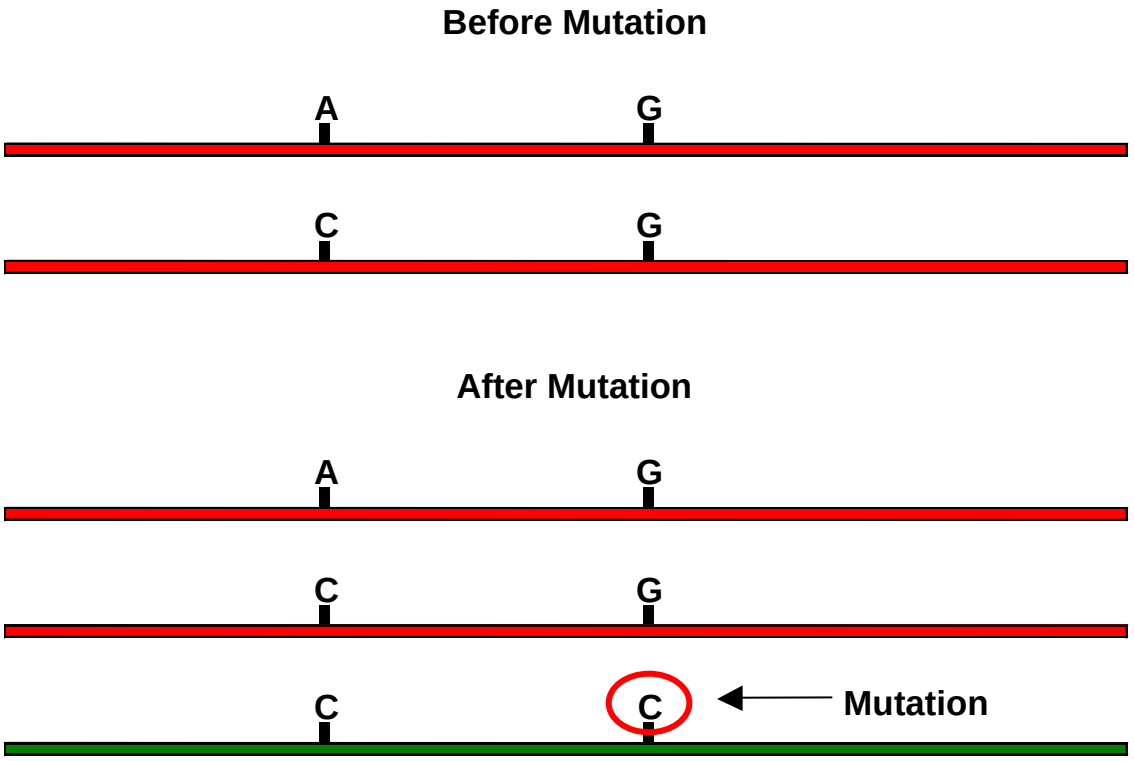
$$p_{AB} \neq p_A p_B$$

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

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

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

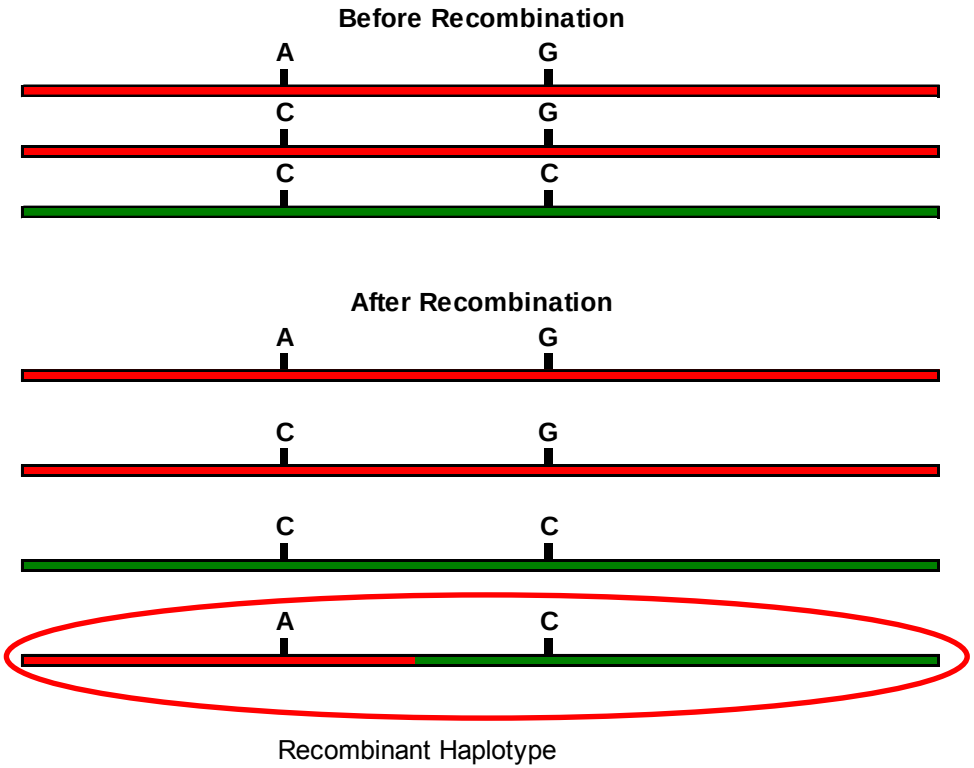
A new mutation...



For a new mutation...

- One haplotype frequency is zero
- Linkage equilibrium does not hold
  - In contrast, **linkage disequilibrium**
- How do haplotype frequencies change over time?

Recombination



Recombinant Haplotype

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
  - …

Disequilibrium Coefficient D_{AB}

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

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

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

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

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

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

Recombination Rate (θ)

- 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

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

$P=1-\theta$

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

$P=\theta$

Overall Change

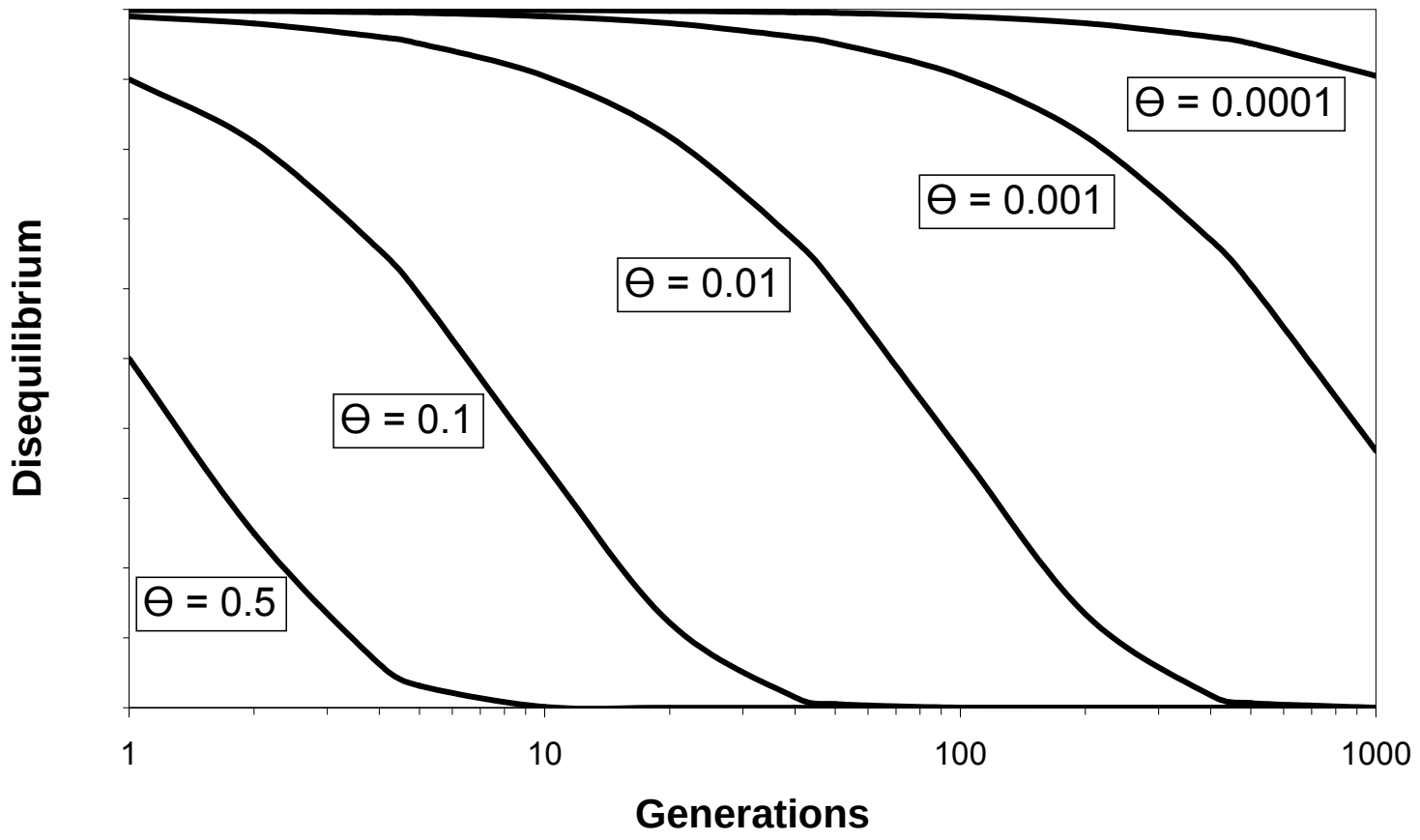
	B	b	
A	$p_A p_B + (1-\theta) D_{AB}$	$p_A p_b - (1-\theta) D_{AB}$	p_A
a	$p_A p_b - (1-\theta) D_{AB}$	$p_a p_b + (1-\theta) D_{AB}$	p_a
	p_B	p_b	

Disequilibrium Decreases...

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…

Decay of D with Time



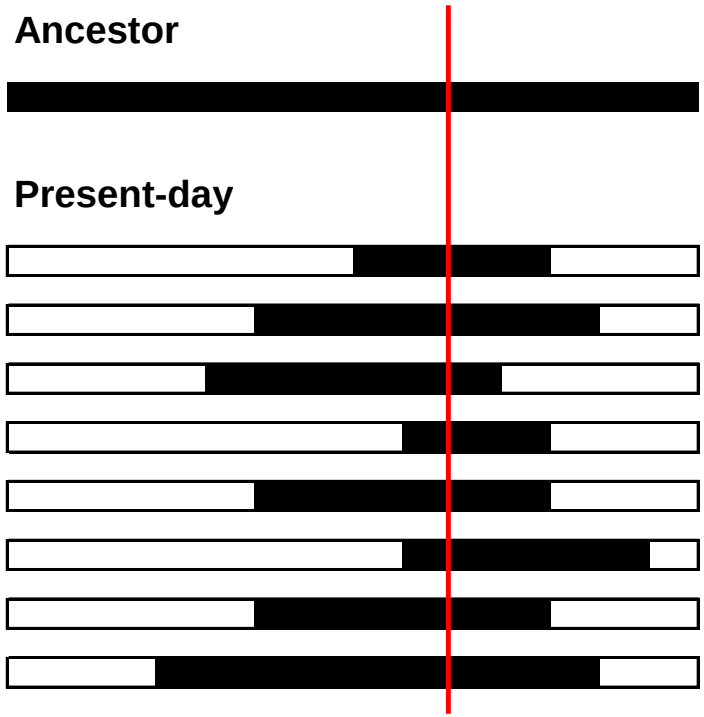
Linkage Equilibrium

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

- (The reverse is always true!)

Linkage Disequilibrium

- Chromosomes are mosaics
- Tightly linked markers
  - Alleles not randomly associated
  - Reflect ancestral haplotypes
- Recombination, Mutation, Drift



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

Range of D_{AB}

- Must be greater than
  - $\text{Max}(-p_A p_B, -p_a p_b)$
- Must be smaller than
  - $\text{Min}(p_A - p_A p_B, p_B - p_A p_B)$

- Constraints ensure that all haplotype frequencies are greater than zero

Alternative Measures

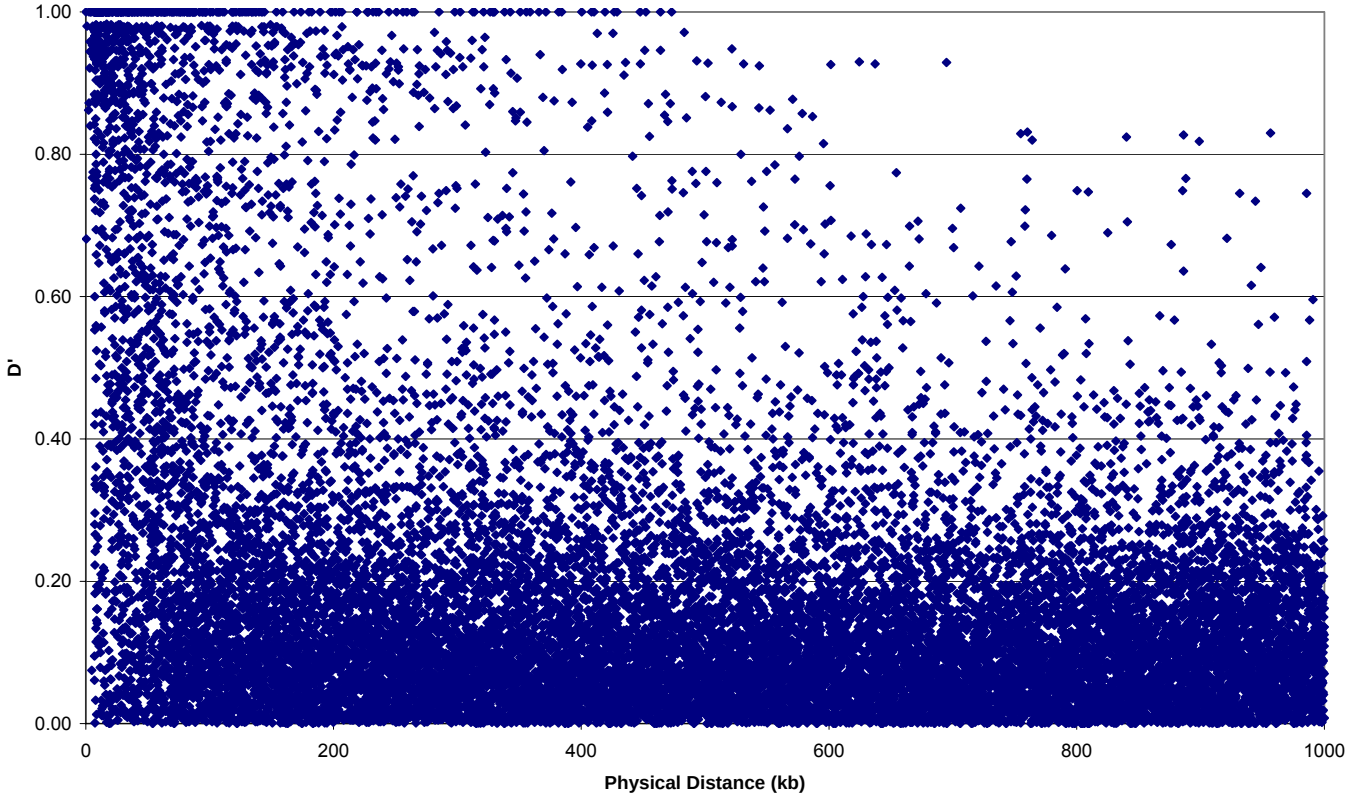
- The most popular measures
  - D’ and |D’|
  - Δ^2 or r^2
- Other common measures
  - Chi-squared
  - P-value

D’

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

- 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

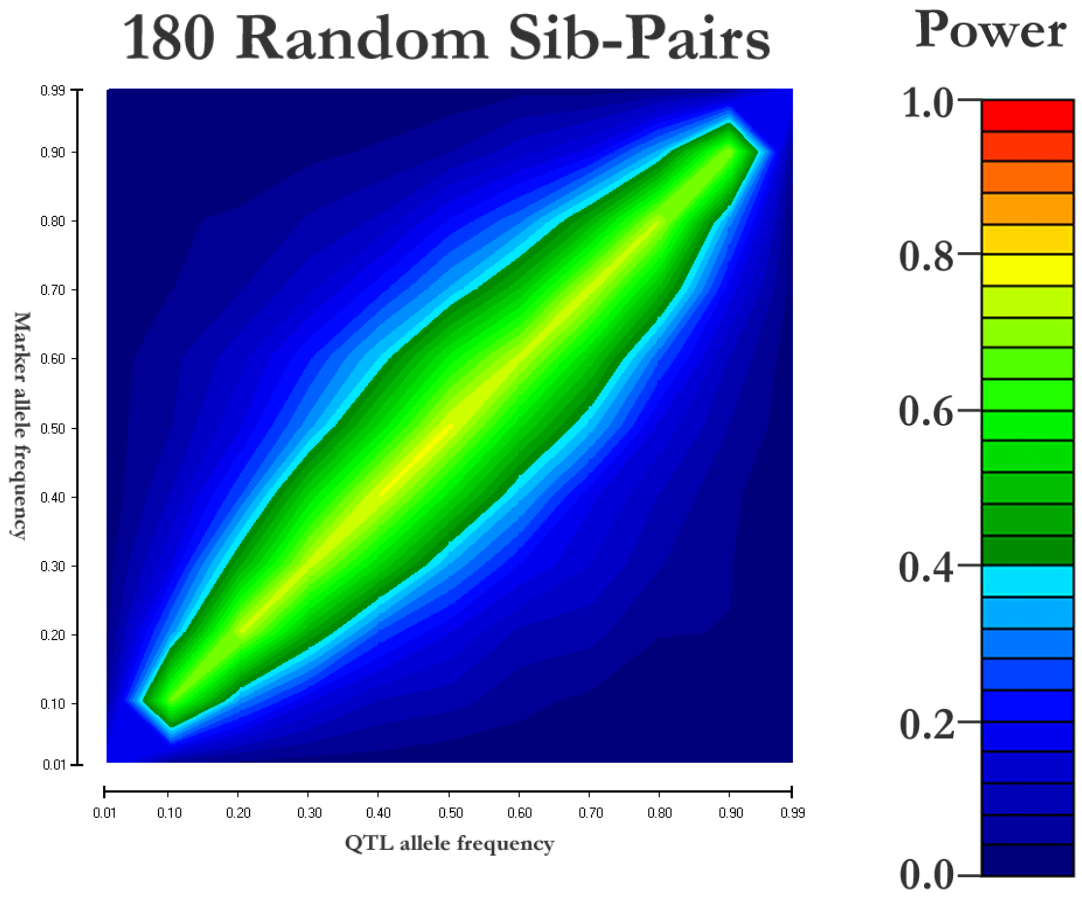
Raw |D'| data from Chr22



More on D'

- Pluses:
  - D' = 1 or D' = -1 means no mandatory recombinants between markers
  - Generally, higher D' means two markers are better surrogates for each other
- Minuses:
  - D' estimates inflated in small samples
  - D' estimates inflated when one allele is rare

Effect of Allele Frequencies



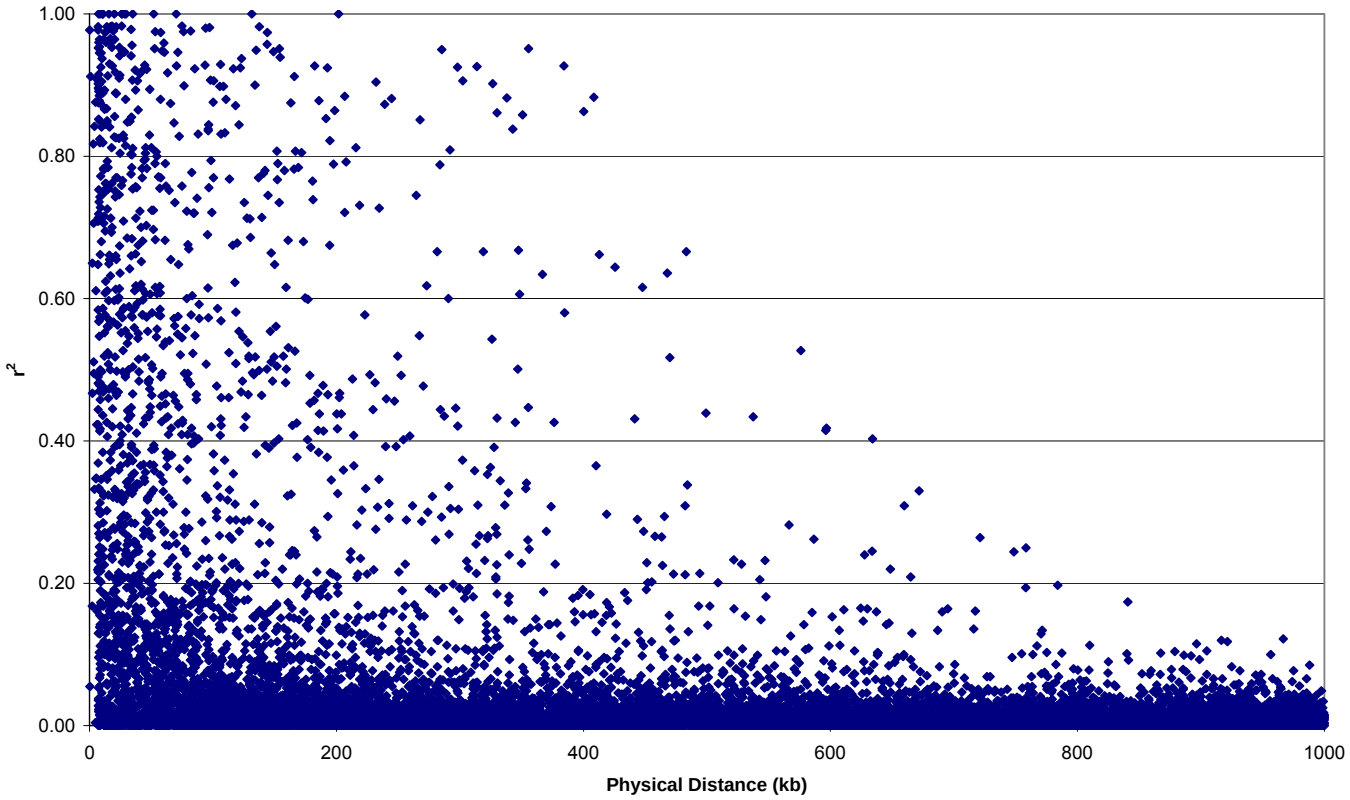
Δ^2 (also called r^2)

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

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

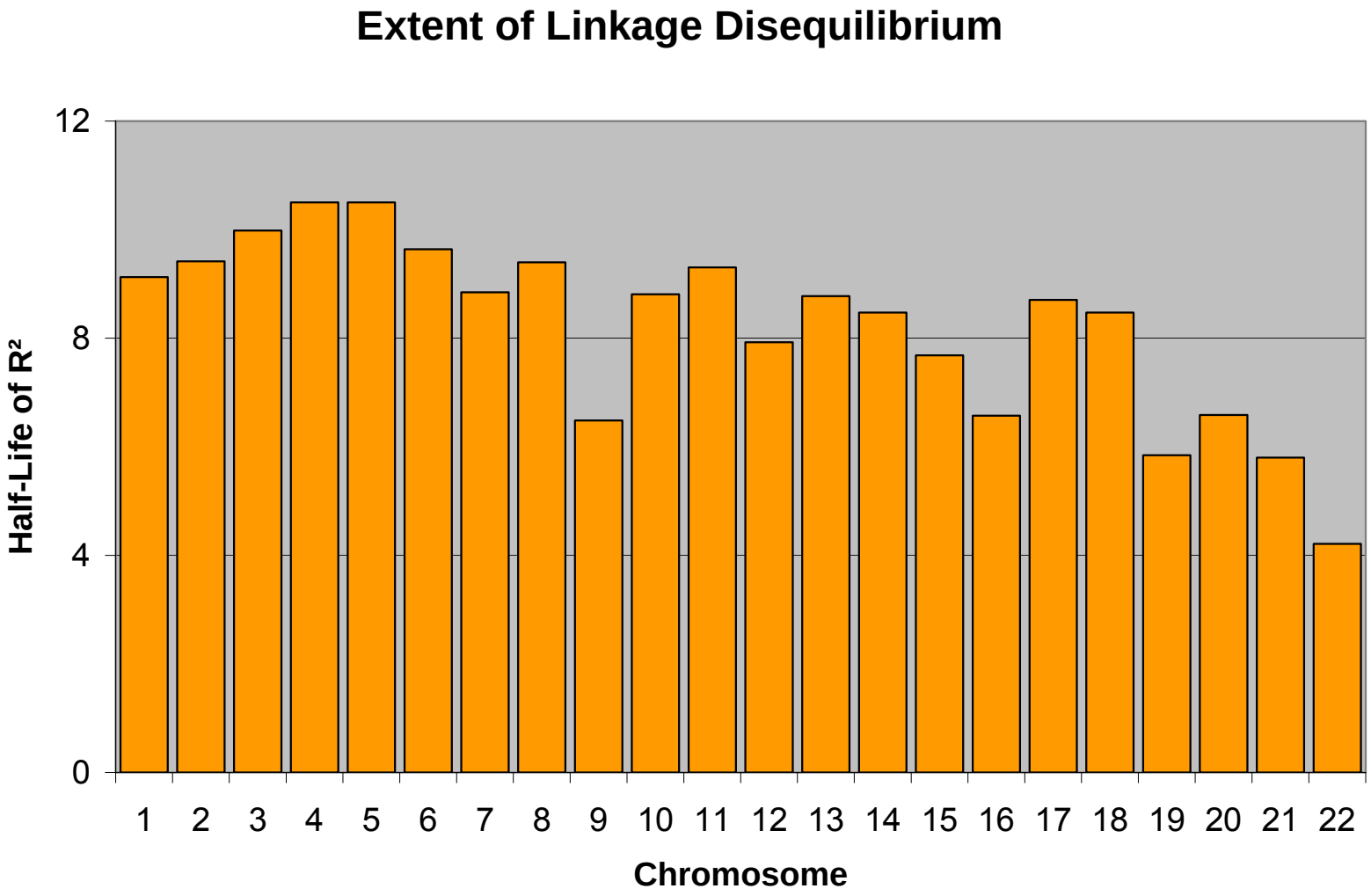
- Ranges between 0 and 1
  - 1 when the two markers provide identical information
  - 0 when they are in perfect equilibrium
- Expected value is $1/2n$

Raw Δ^2 data from Chr22

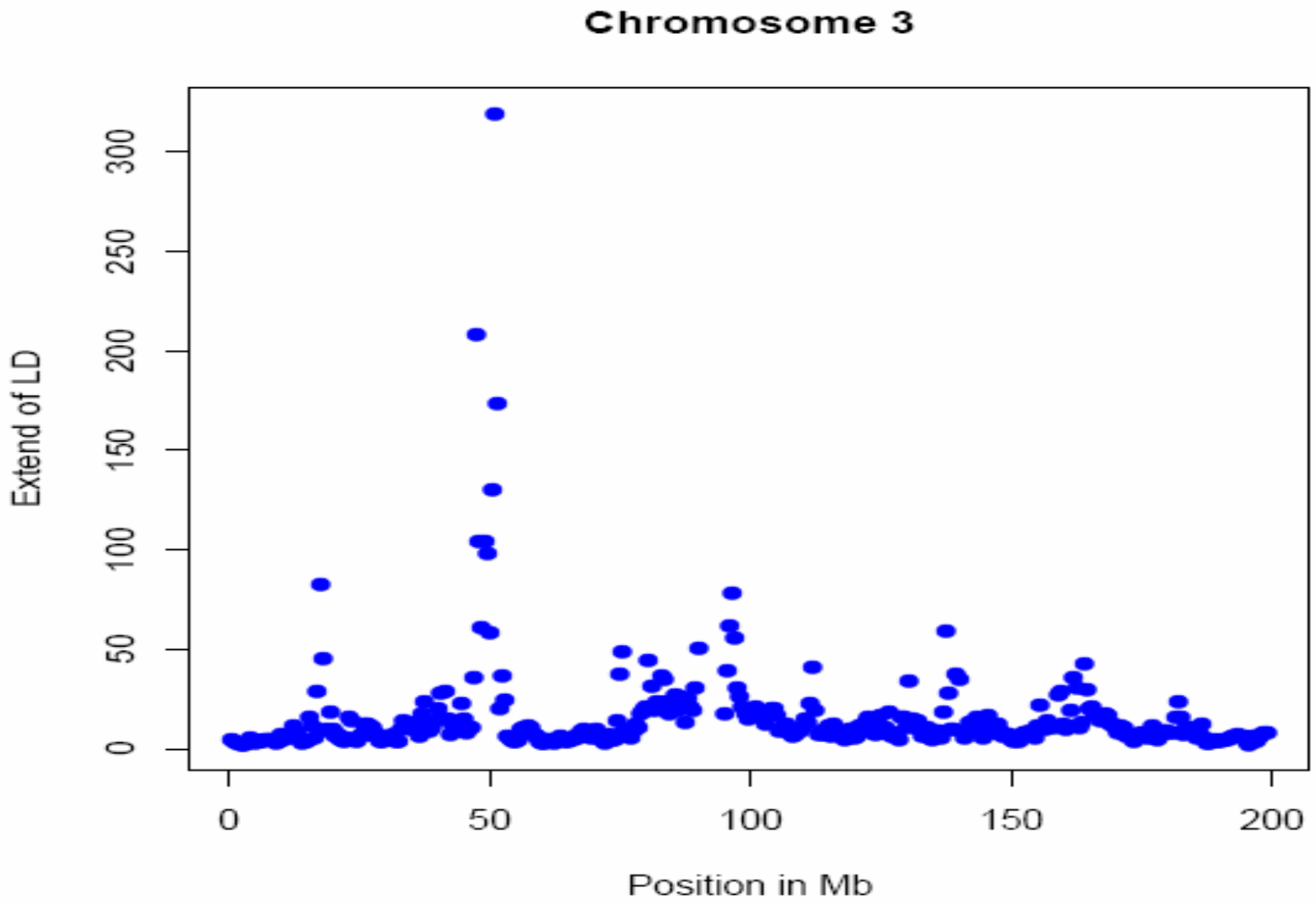


Summary of Disequilibrium in the Genome

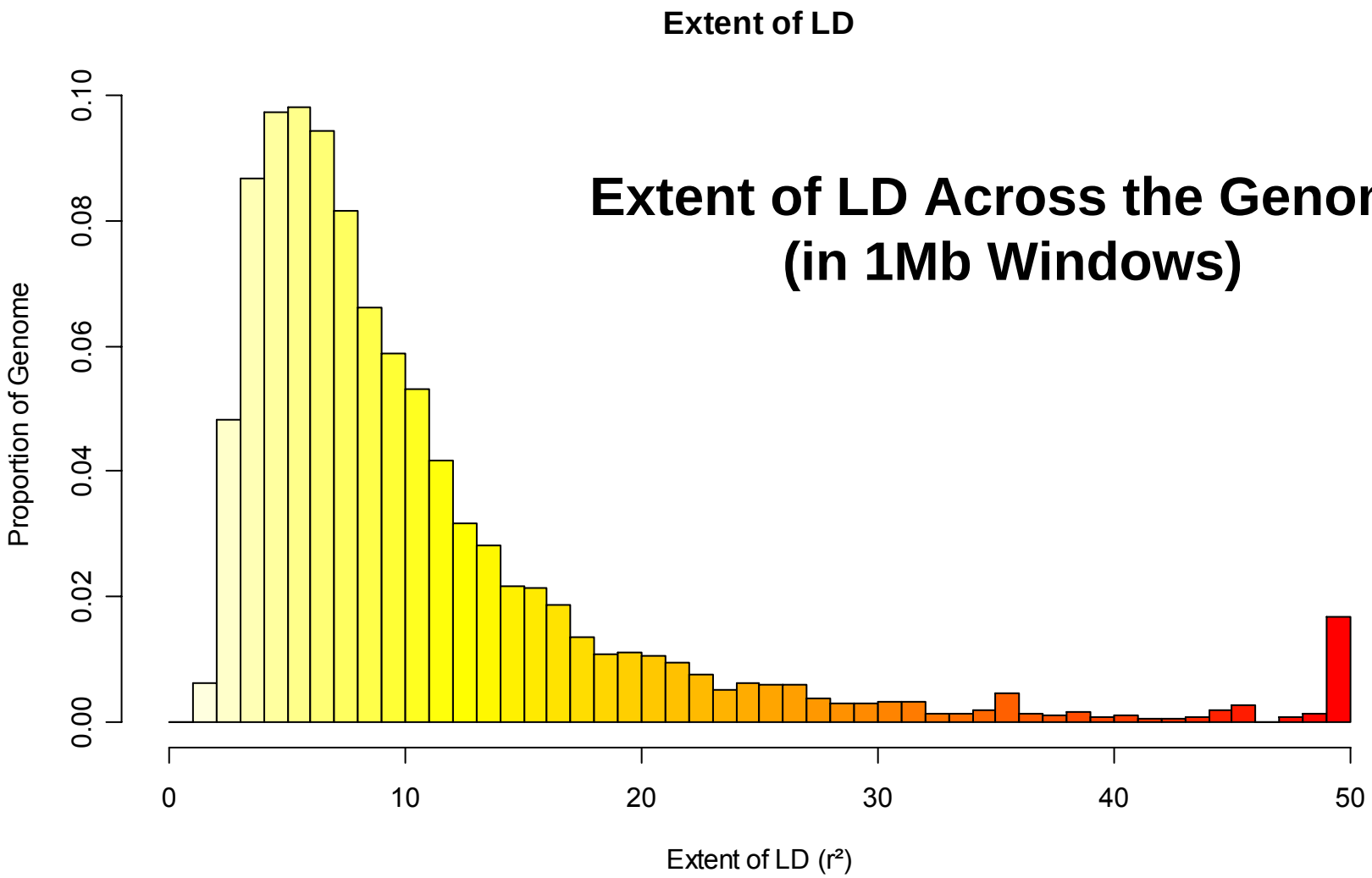
- How much disequilibrium is there?
- What are good predictors of disequilibrium?
- What are good predictors of variation in disequilibrium?



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

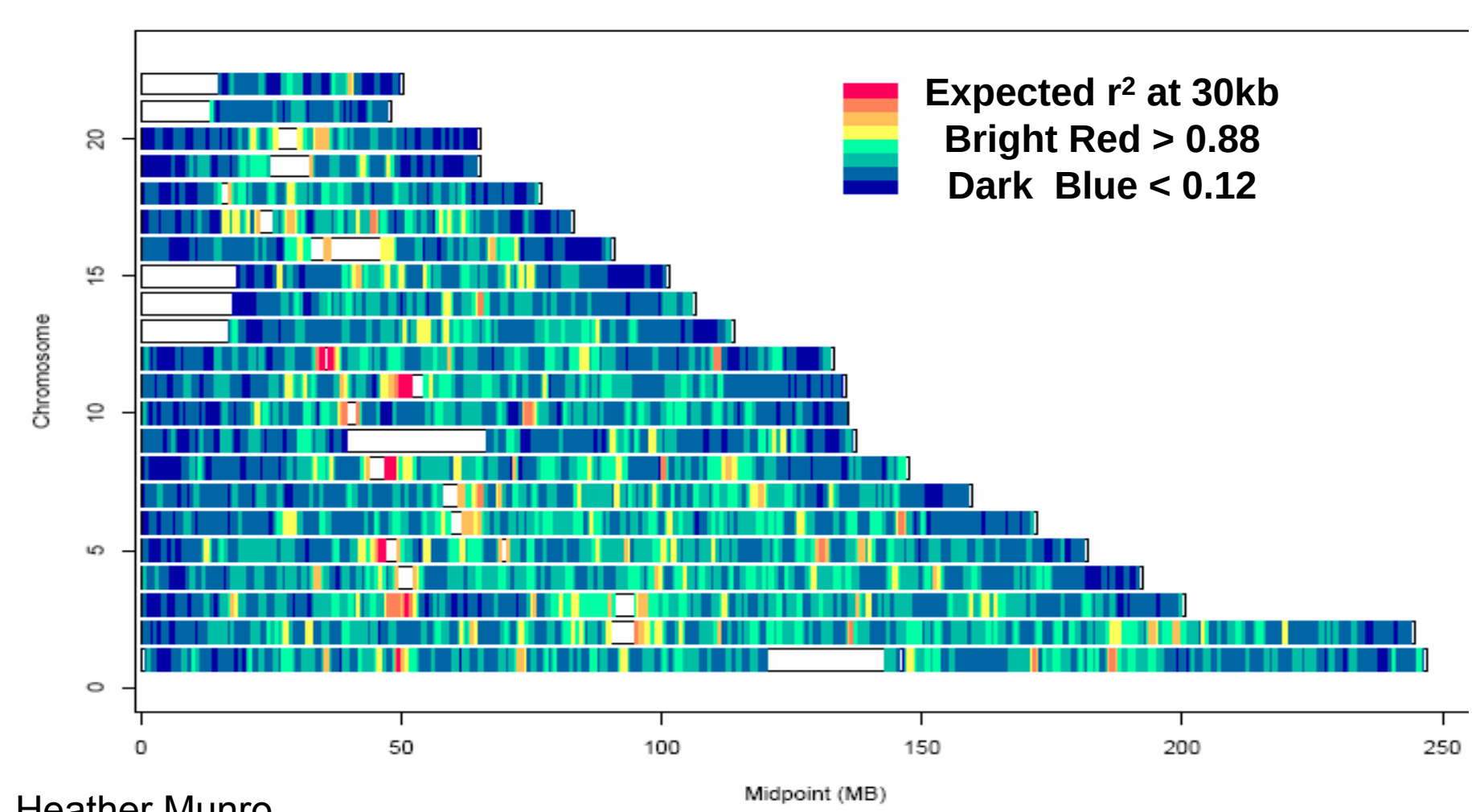


Most chromosomes exhibit regions of extended LD on the megabase scale



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

Genomic Variation in Disequilibrium (CEPH)

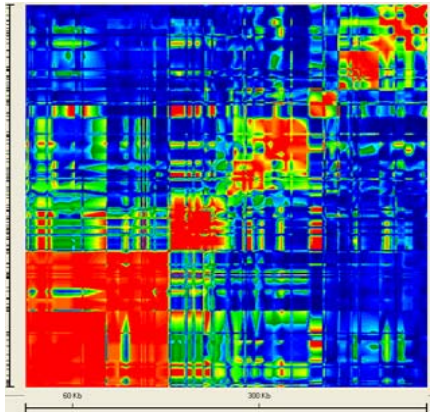


Heather Munro

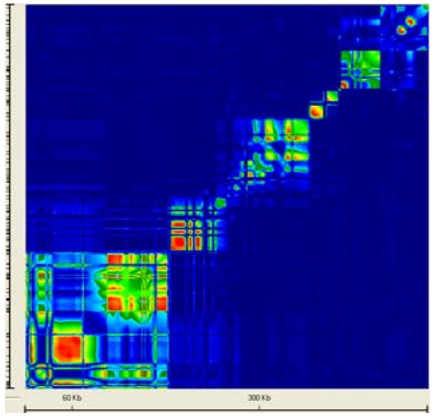
Dense Region 1

- Chromosome 7
  - 157 markers / 520 kb
  - 27.0 – 27.5 Mb
  - Average LD region

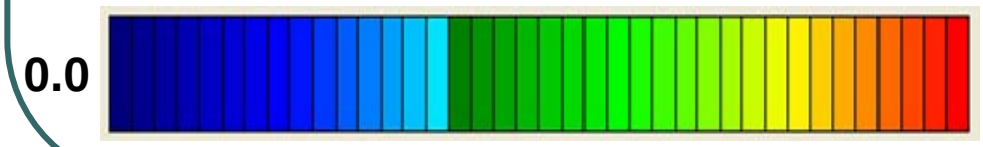
- SNP picking (33/157 = 21%)
  - 12 unique SNPs
  - 21 tagging SNPs
  - Others, average $r^2 = 0.73$



D’



R^2



Predicting Disequilibrium

- How can we predict Δ^2?
  - Depends on allele frequencies
  - Degree of recombination
- How can we predict LD for multiple markers?
- One option: simulation…

Question:

- How would you set up a simulation to predict D’ or r^2 for a pair of markers?
- What parameters might be of interest?

Coalescent Models ...

- A better approach to predicting LD and distribution of genetic variation
- Read book chapter by Richard Hudson