Linkage Disequilibrium

Biostatistics 666

Logistics: Office Hours, Exams

- Office hours on Fridays from 3-4 pm.
- Working to reserve a room in the School of Public Health.
- Aiming for mid-term on October 25.
  - Let me know immediately if you will need accommodations.
    My e-mail is goncalo@umich.edu

Previously …

- 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 …

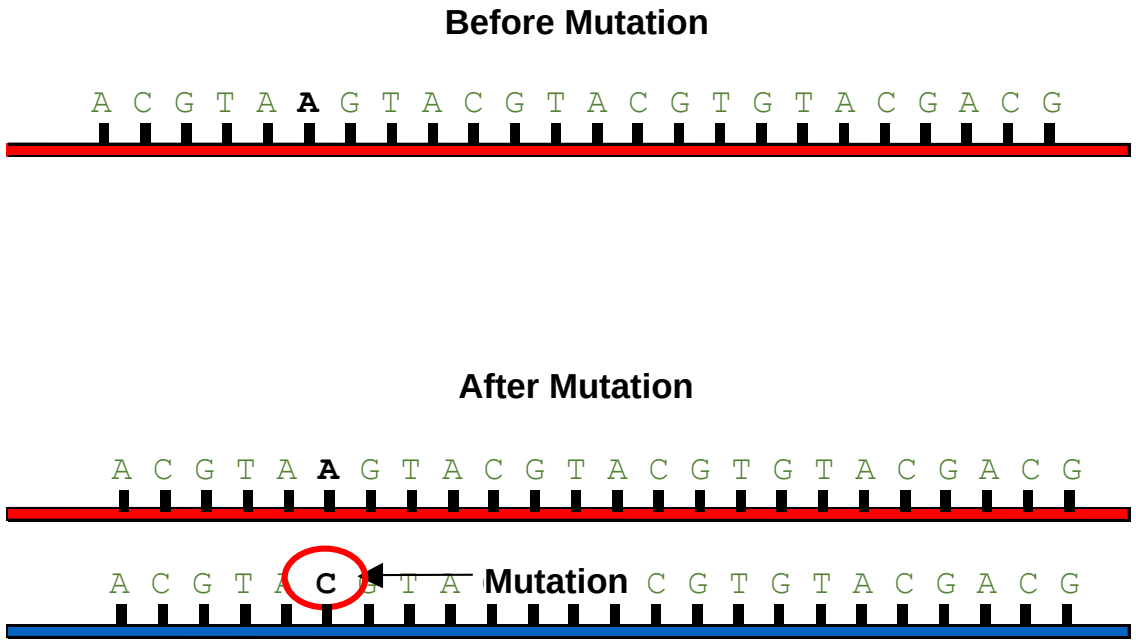
- We’ll consider properties of pairs of alleles
- Haplotype frequencies
- Linkage equilibrium
- Linkage disequilibrium

Intuition

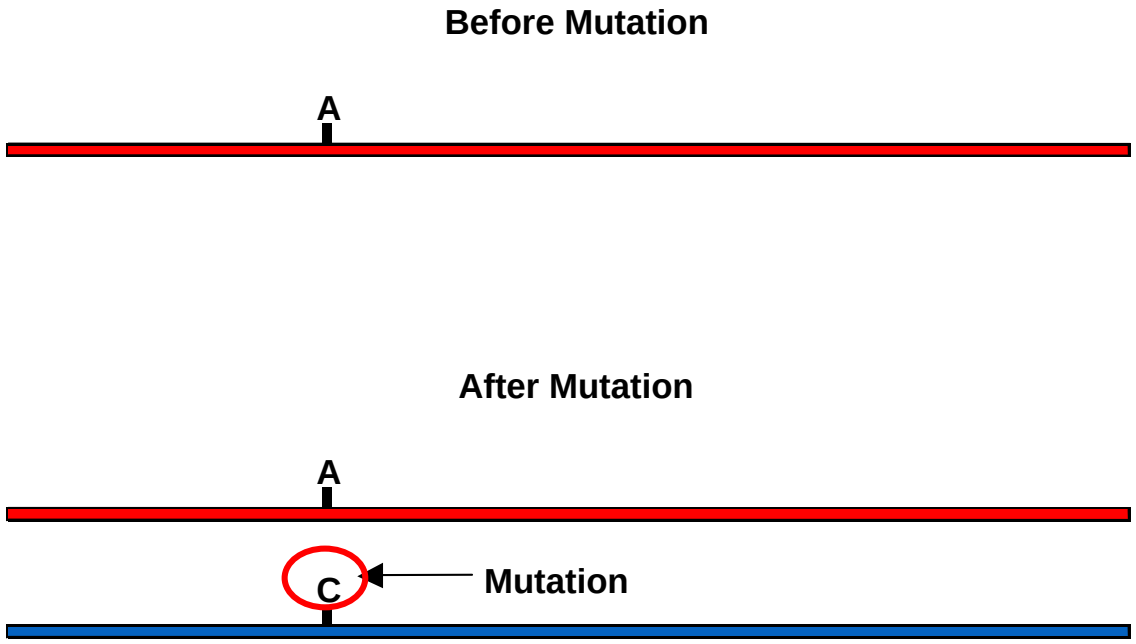
- Genetic variants band together through time and populations …
- … this often results in correlated distributions for nearby variants
- The phenomenon is termed linkage disequilibrium

Let’s consider the history of two neighboring alleles…

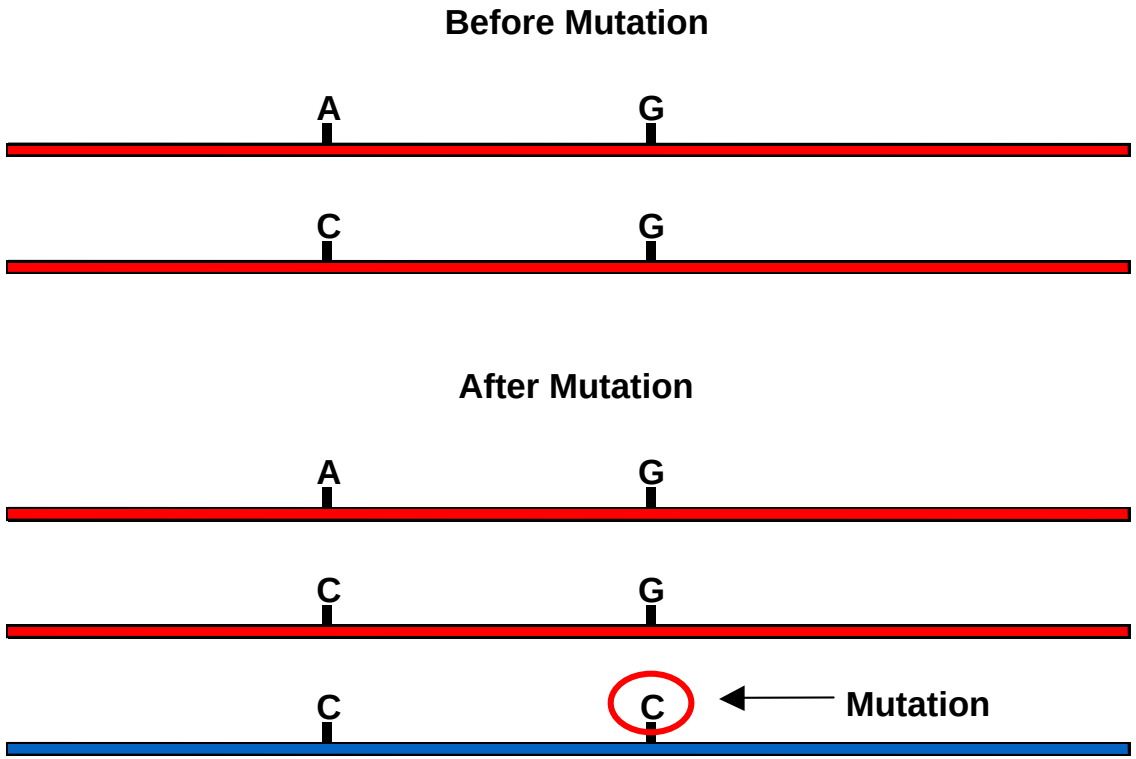
Alleles that exist today arose through ancient mutation events…



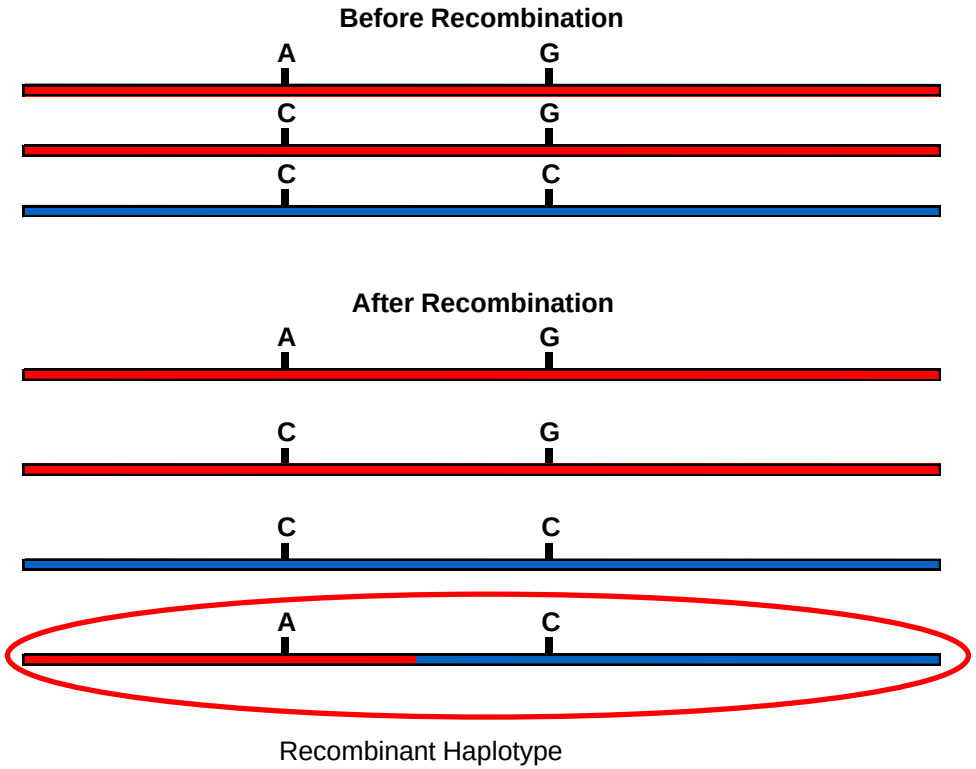
Alleles that exist today arose through ancient mutation events…



One allele arose first, and then the other…

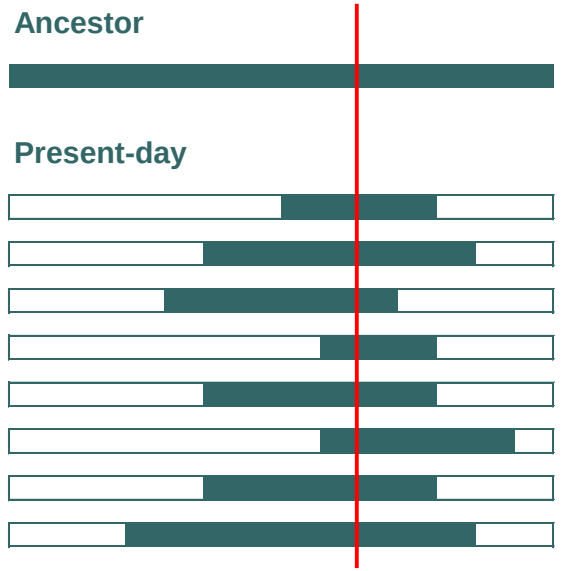


Recombination generates new arrangements for ancestral alleles



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



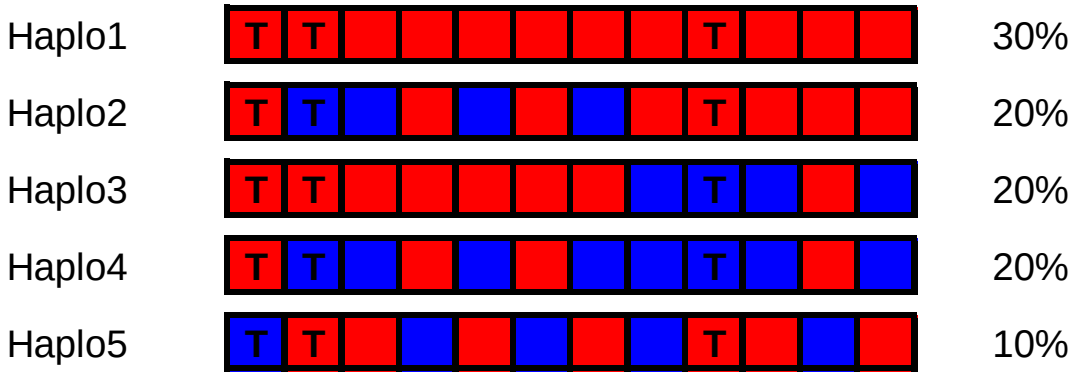
Why is linkage disequilibrium important for genetic studies?

Benefits …

Challenges …

Tagging SNPs

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



Basic Descriptors of Linkage Disequilibrium

Commonly Used Descriptors

- Haplotype Frequencies
  - How often do we see each allele combination along a 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* as the common alleles
    - *a, b* as the rare allele
- Range depends on allele frequencies
  - Hard to compare between markers
- Can you see why the range of D_{AB} depends on allele frequencies?

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

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

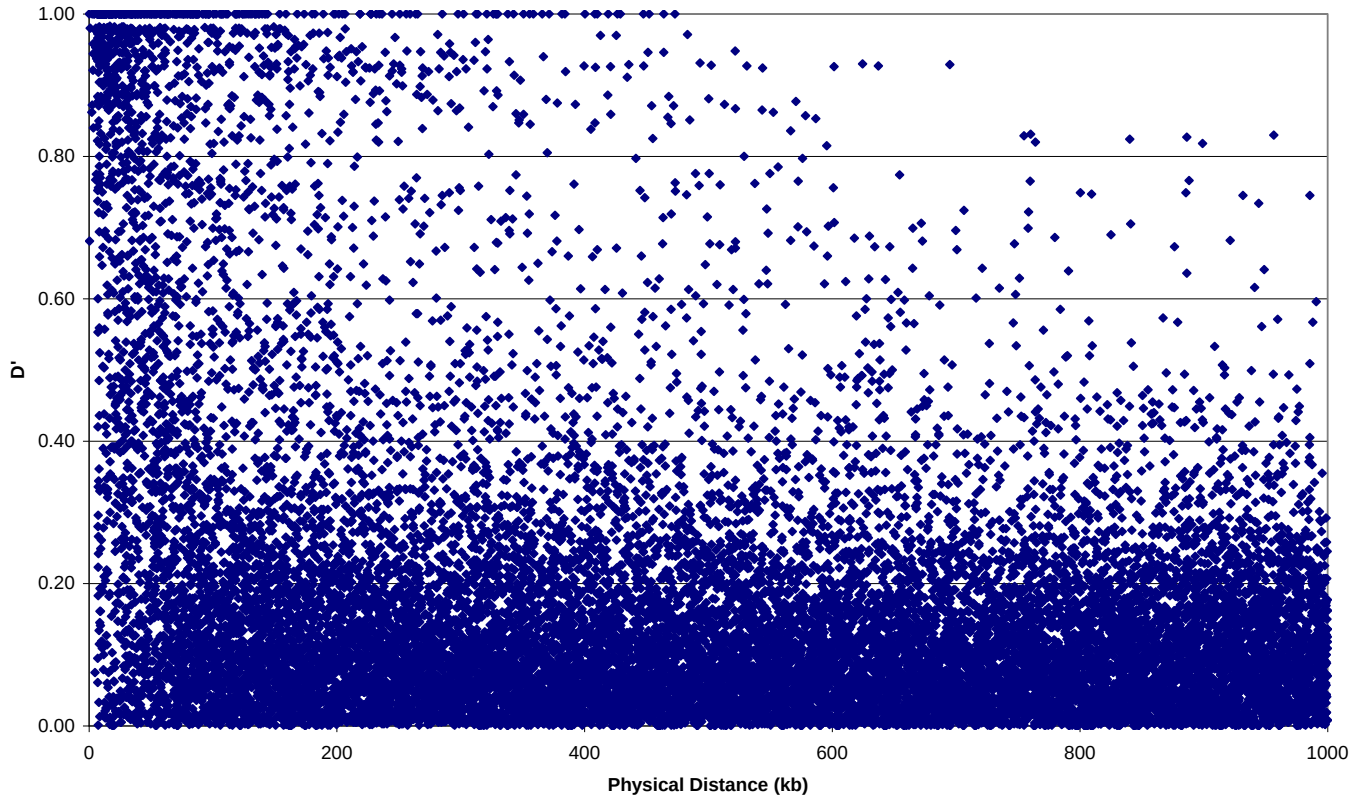
- Ranges between –1 and +1
  - More likely to take extreme values when allele frequencies are small
  - ± 1 implies at least one of the observed haplotypes was not observed

More on D’

- Pluses:
  - D’ = 1 or D’ = -1 means no evidence for recombination between the markers
  - If allele frequencies are similar, high D’ implies markers are good surrogates for each other

- Minuses:
  - D’ estimates inflated in small samples
  - D’ estimates inflated when one allele is rare

Raw |D’| data from Chr22



Dawson et al, *Nature,* 2002

Δ^2 (also called r^2)

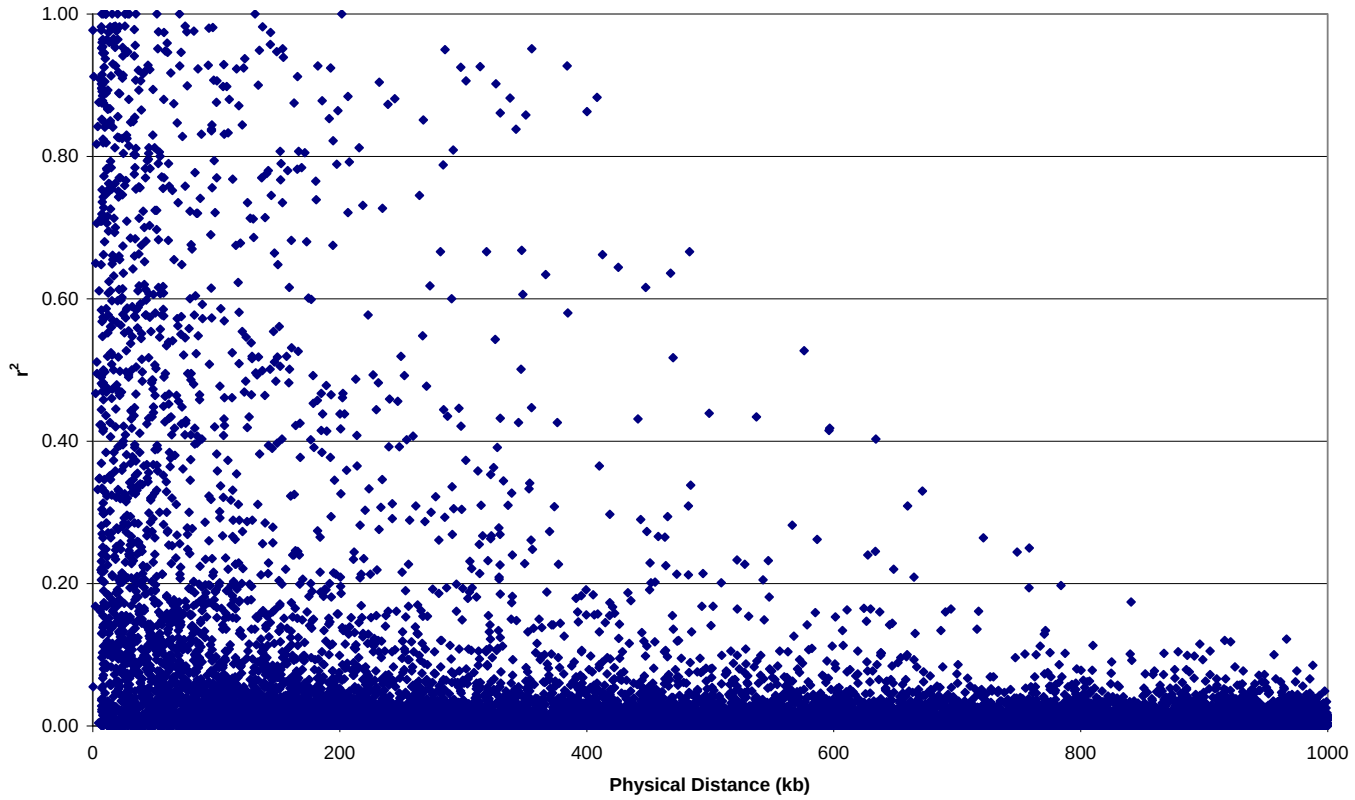
$$\Delta^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}$$
$$= \frac{\chi^2}{2n}$$

- 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)

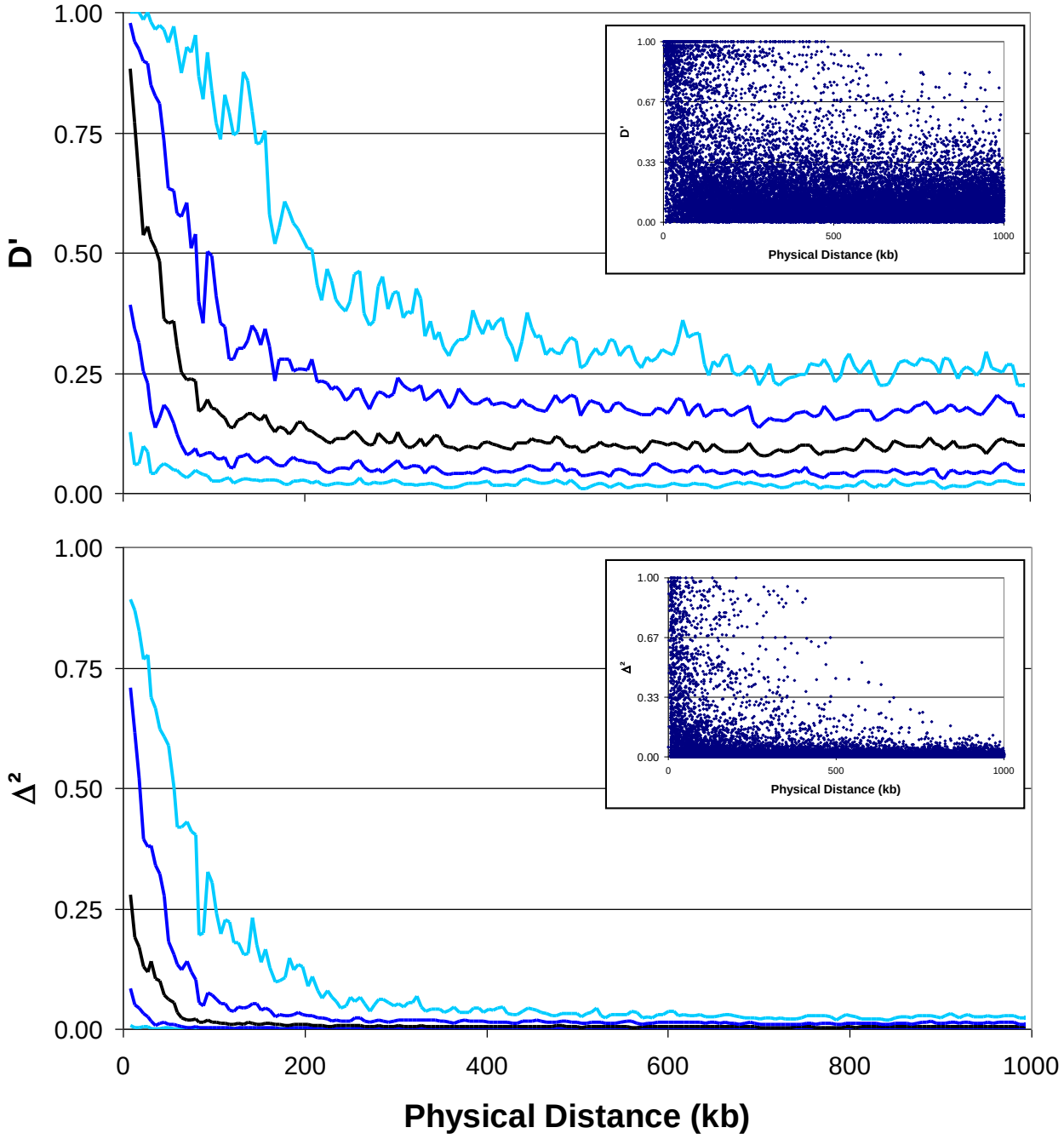
Raw Δ^2 data from Chr22



Dawson et al, *Nature,* 2002

Variability Of Pair-Wise LD

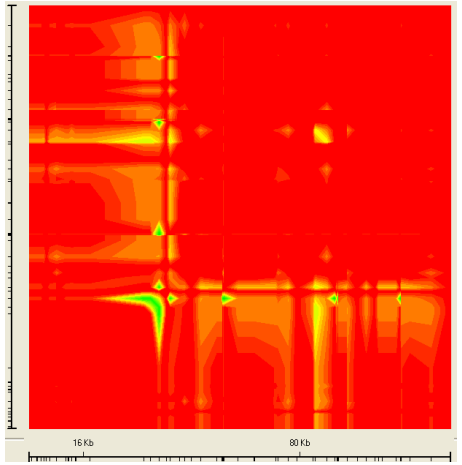
Median
Quartiles
Deciles



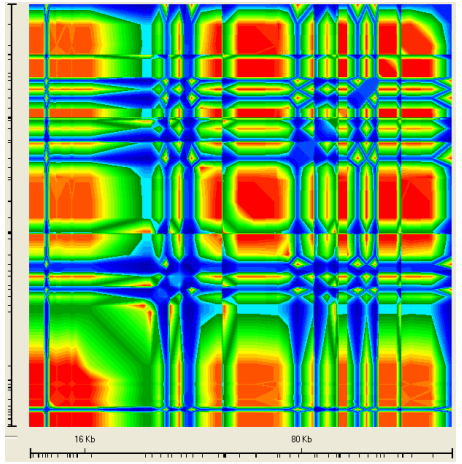
Dense Region 2

- Chromosome 21
  - 57 markers / 130 kb
  - 37.37 – 37.50 Mb
  - High LD region

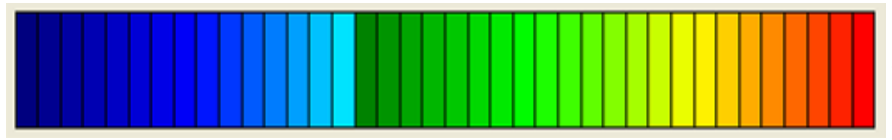
- SNP picking (8/57 = 14%)
  - 5 unique SNPs
  - 3 tagging SNPs
  - Others, average $r^2 = 0.94$



D’



R^2

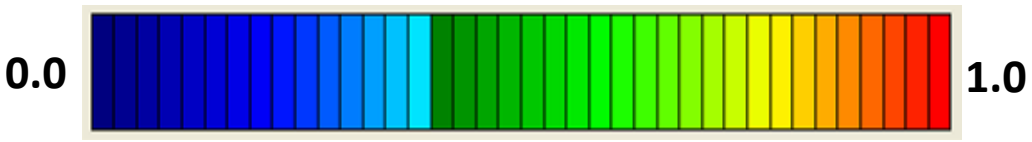


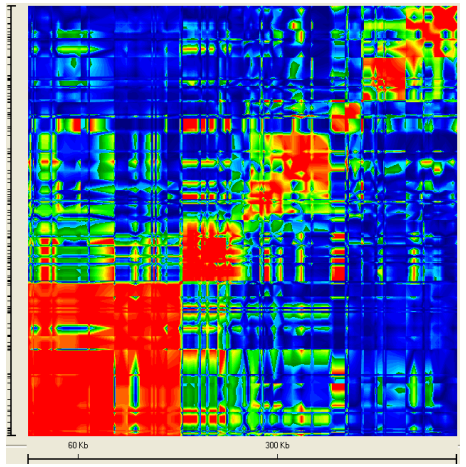
International HapMap Consortium Data

Dense Region 1

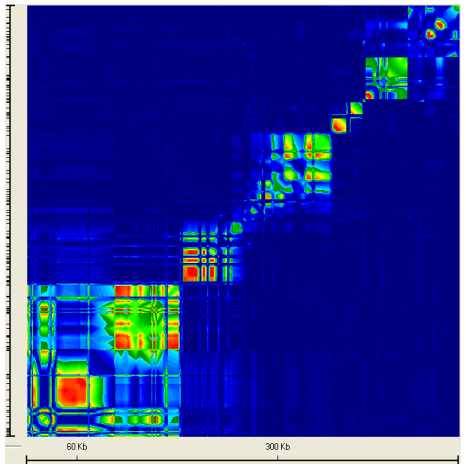
- Chromosome 7
  - 157 markers / 520 kb
  - 27.0 – 27.5 Mb
  - Average LD region

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





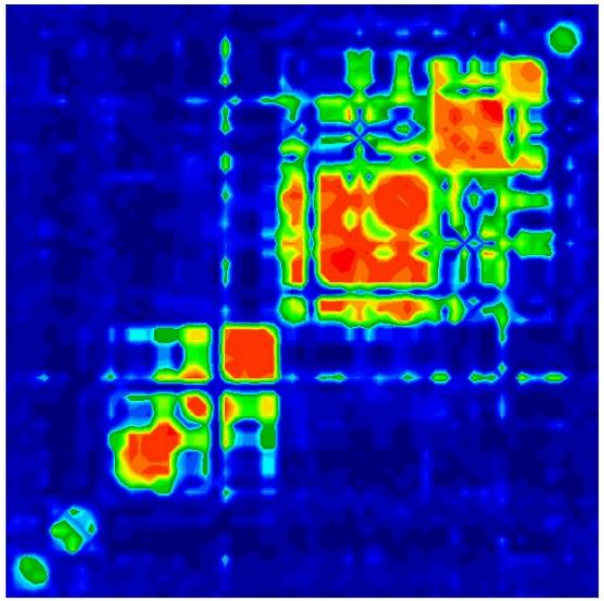
D’



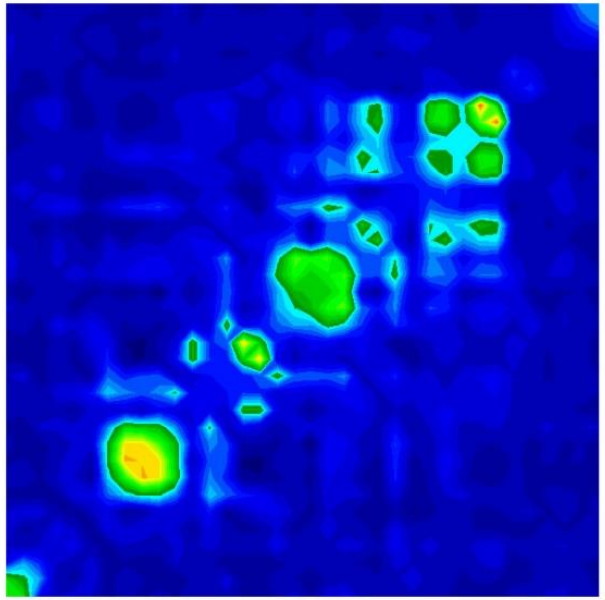
R^2

International HapMap Consortium Data

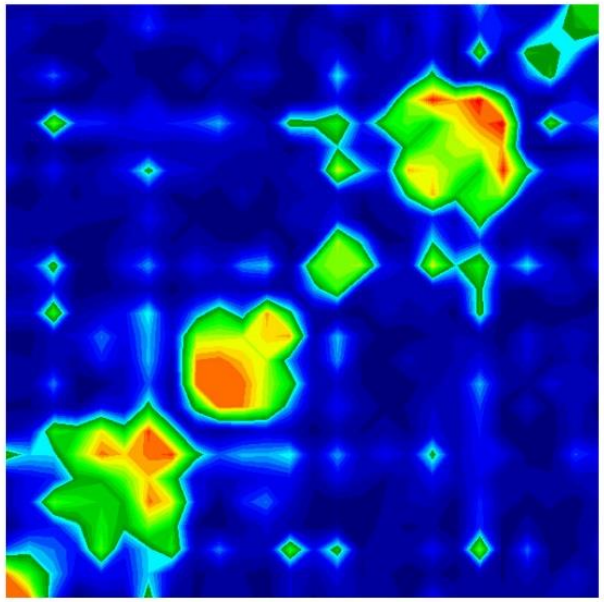
Linkage Disequilibrium in Three Regions



2q13
(63 markers)



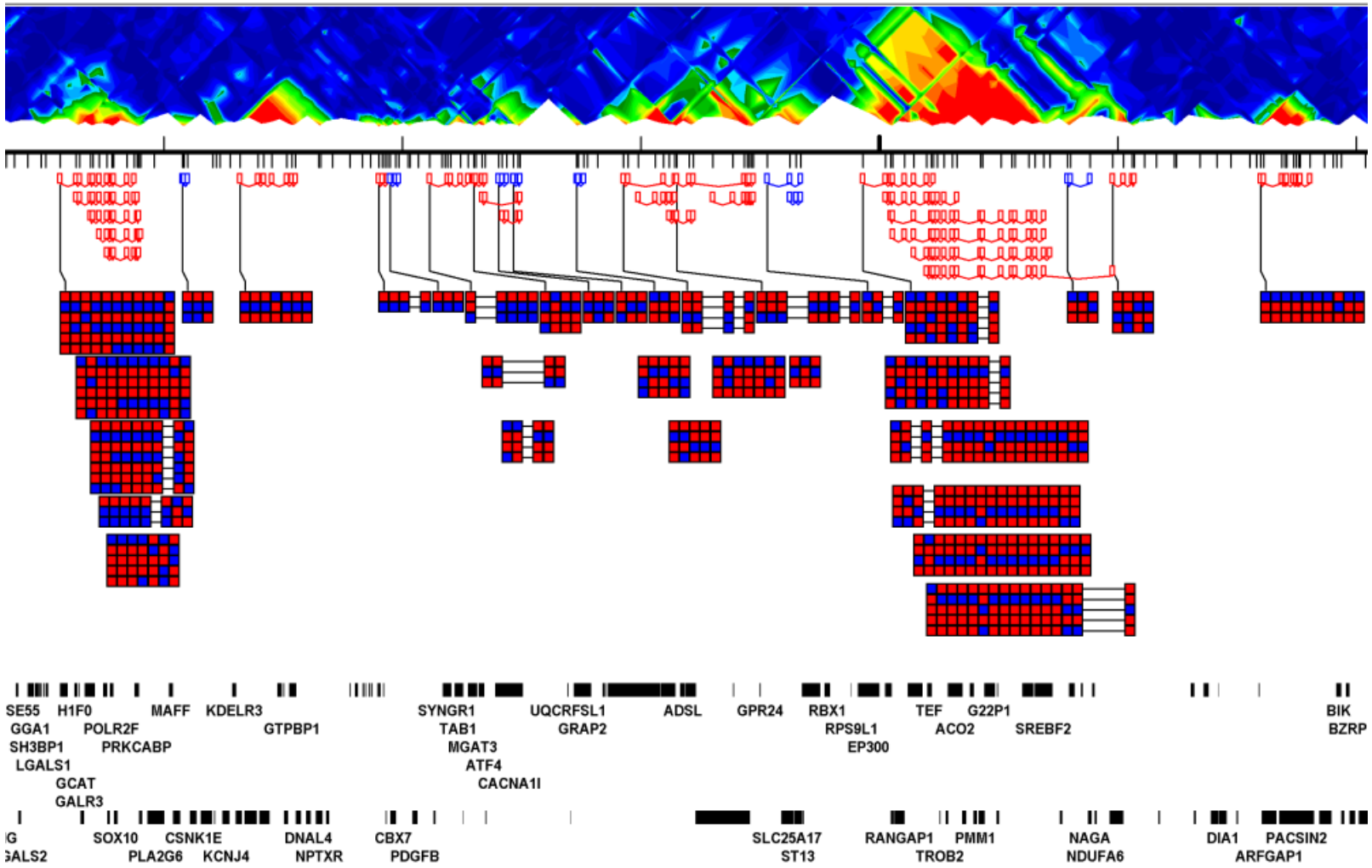
13q13
(38 markers)



14q11
(26 markers)

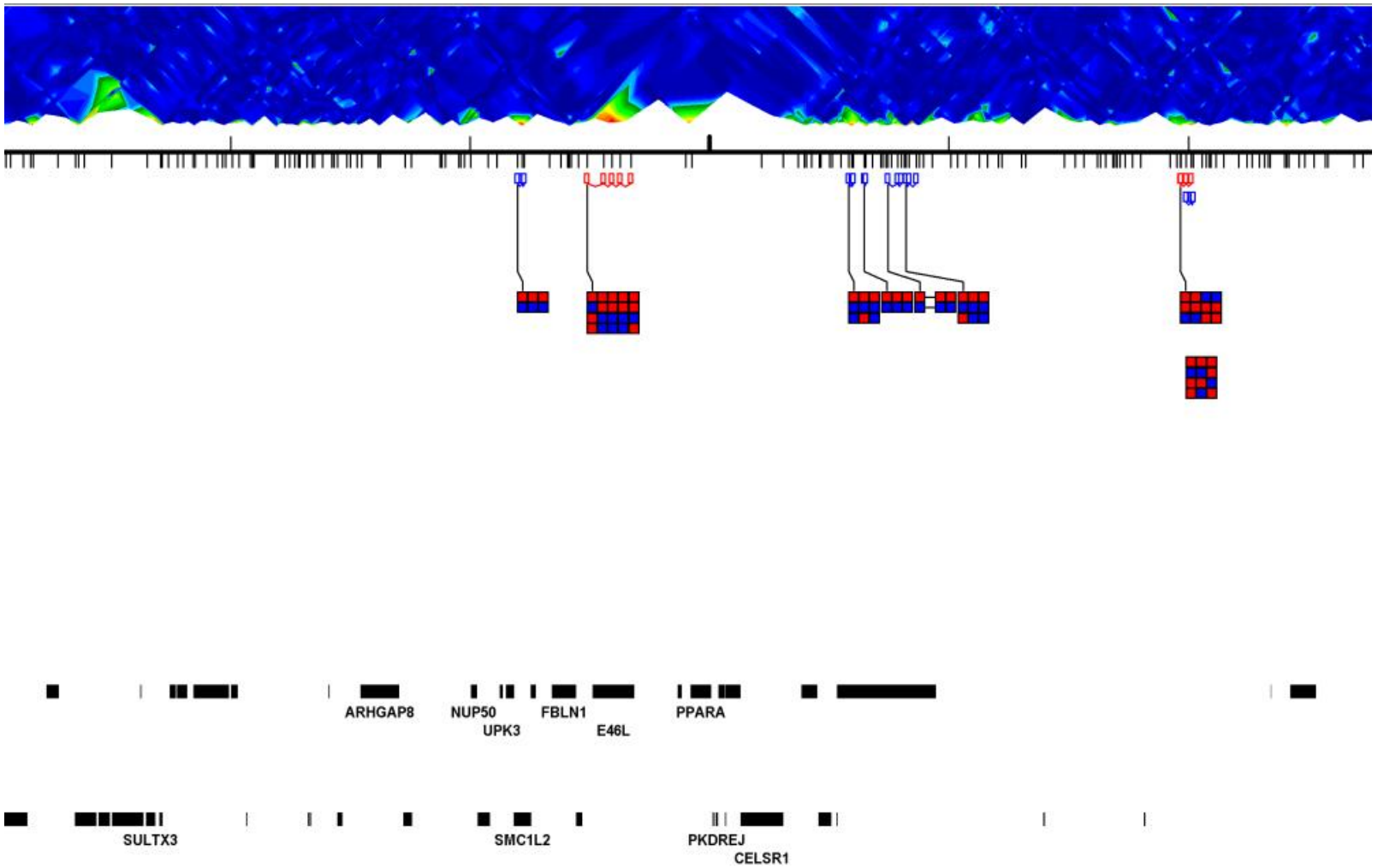
Abecasis et al, *Am J Hum Genet*, 2001

Chr22 High LD: 22-27 Mb



Dawson et al, *Nature*, 2001

Chr22 Low LD: 27-32 Mb



Dawson et al, *Nature*, 2001

When does linkage equilibrium hold?

Equilibrium or Disequilibrium?

- We will outline a simple justification for linkage equilibrium at most loci
  - For now, we will ignore drift in allele frequencies over time.

- In practice, extent of disequilibrium results form balance of factors
  - Distance between markers
  - Genetic drift (a function of population size)
  - Random mating
  - …

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

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

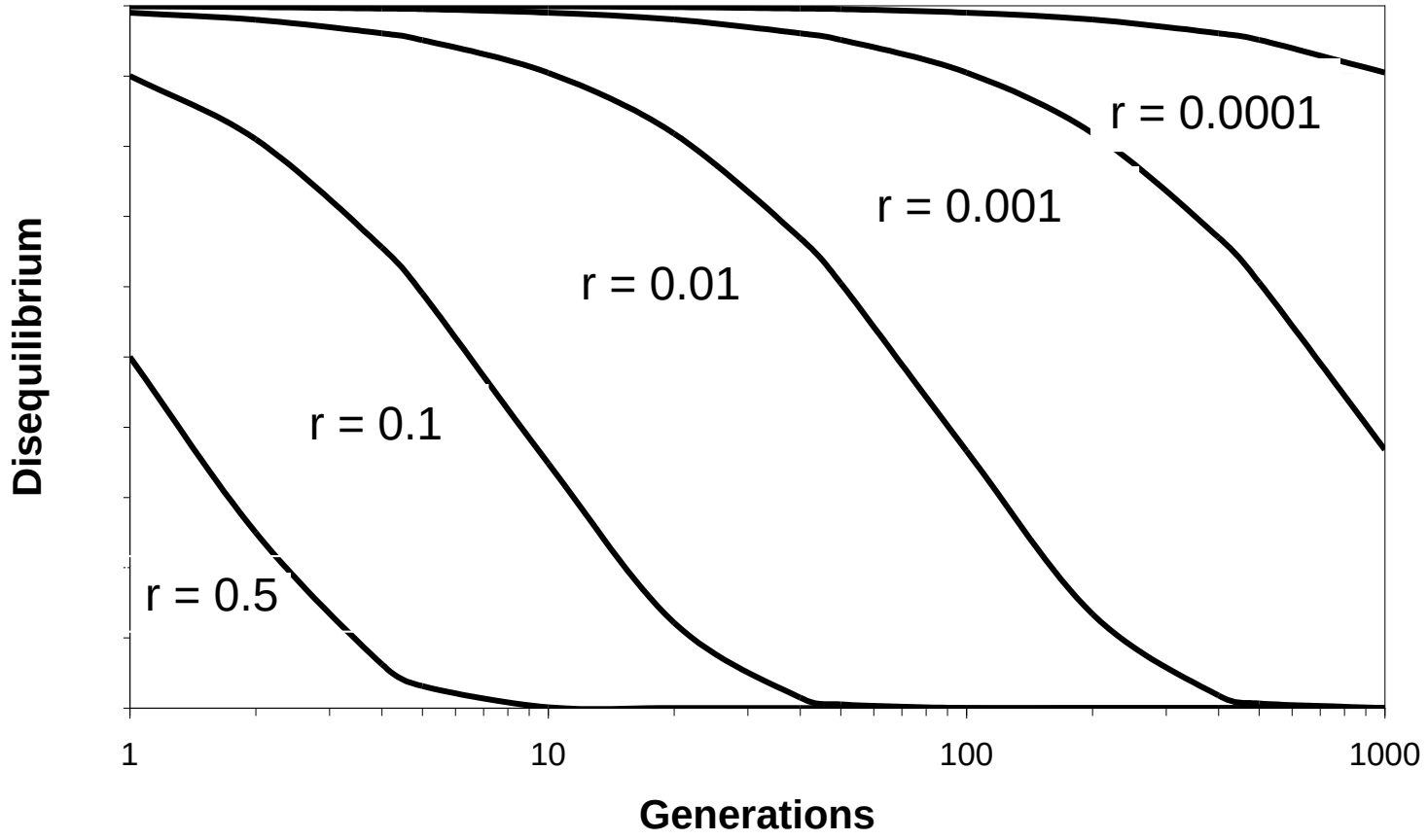
	B	b	
A	$p_A p_B + (1-r) D_{AB}$	$p_A p_b - (1-r) D_{AB}$	p_A
a	$p_A p_b - (1-r) D_{AB}$	$p_a p_b + (1-r) D_{AB}$	p_a
	p_B	p_b	

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…
  - $D^t_{AB} = (1-r)^t D^0_{AB}$
- A better model should allow for changes in allele frequencies over time…

Linkage Equilibrium

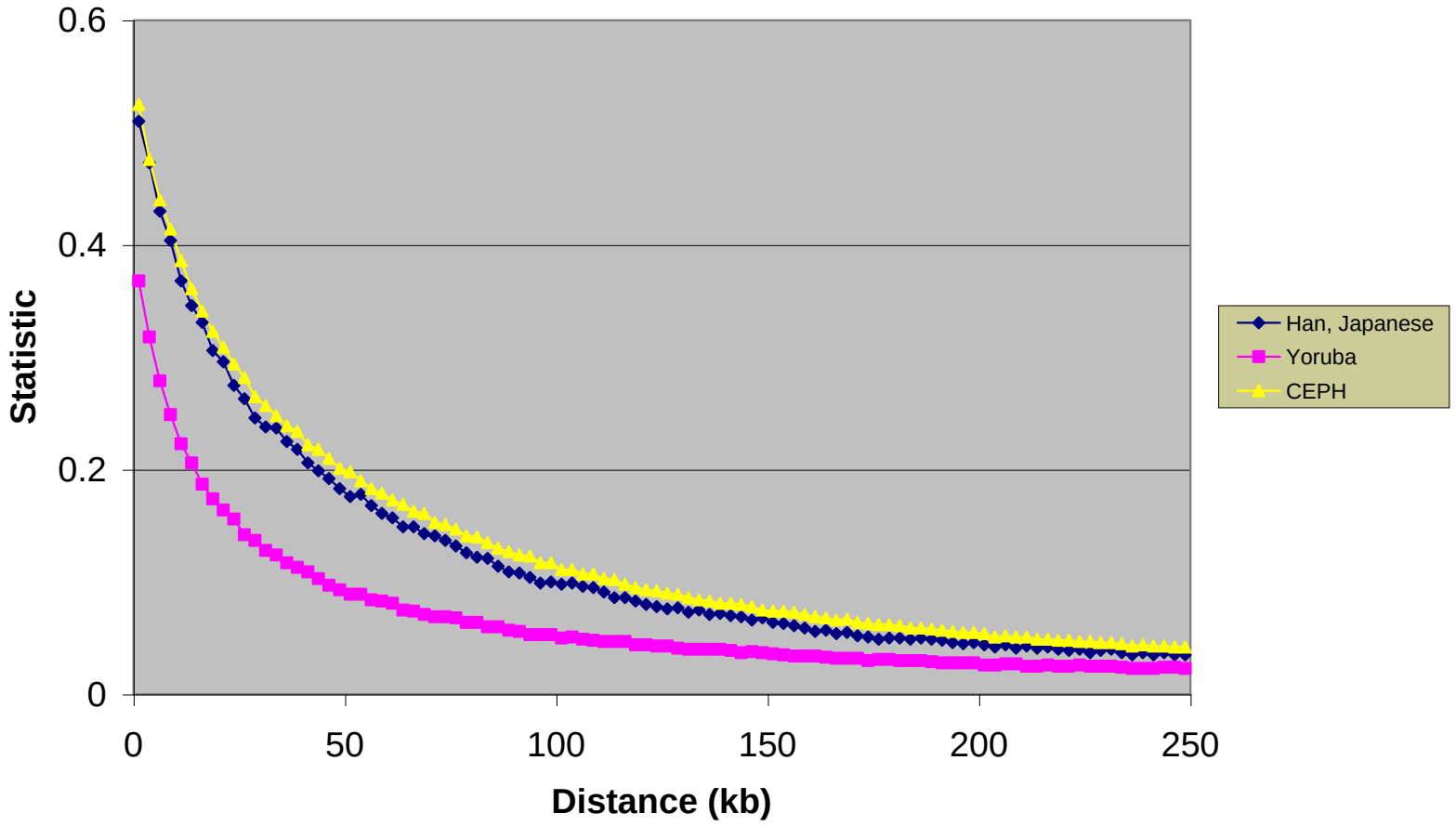
- In a large random mating population haplotype frequencies converge to a simple function of allele frequencies

More Examples of Linkage Disequilibrium Data

How much disequilibrium is there?

What are good predictors of disequilibrium?

Comparing Populations …



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

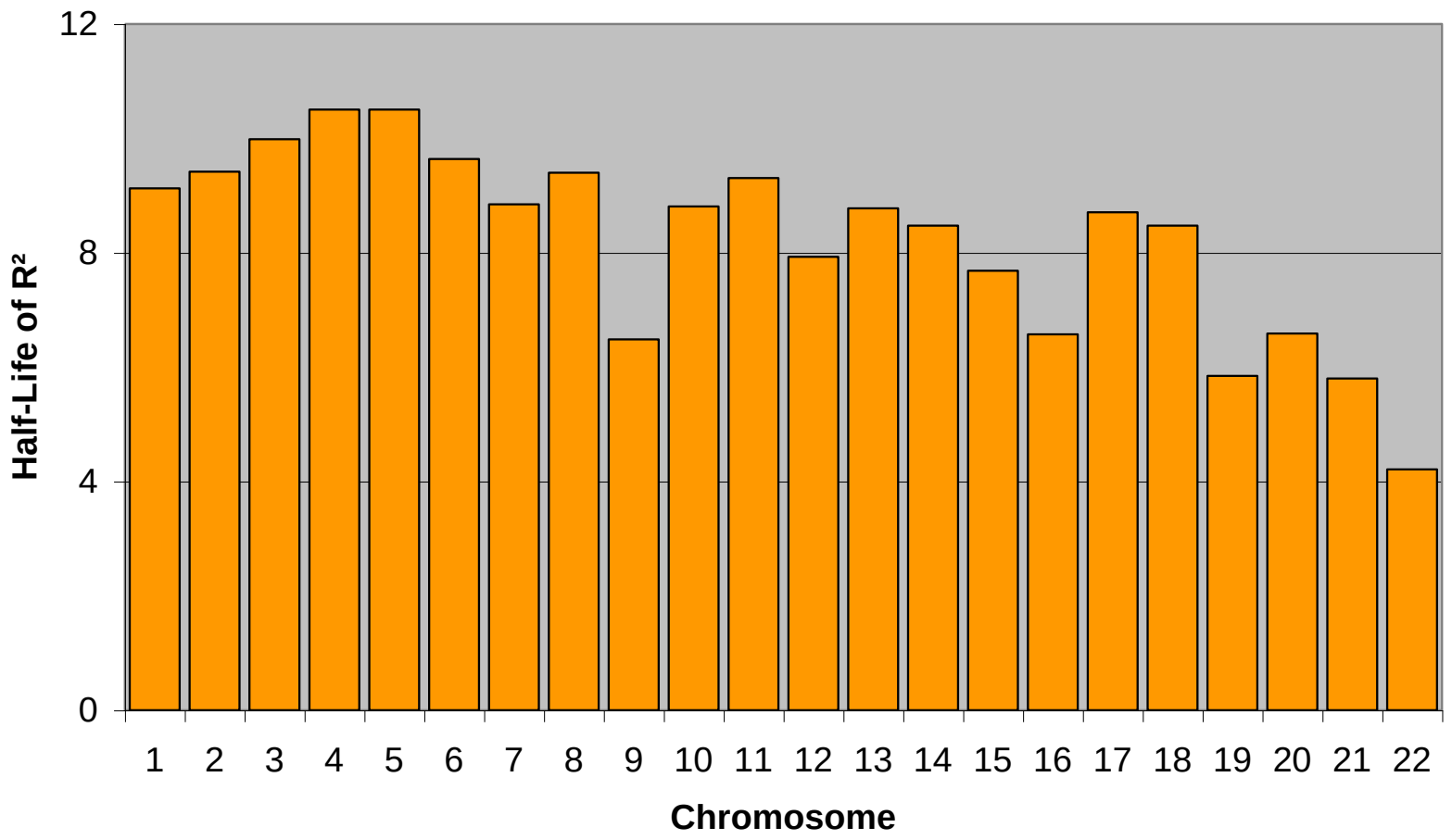
International HapMap Consortium, *Nature,* 2005

Variation in Linkage Disequilibrium Along The Genome

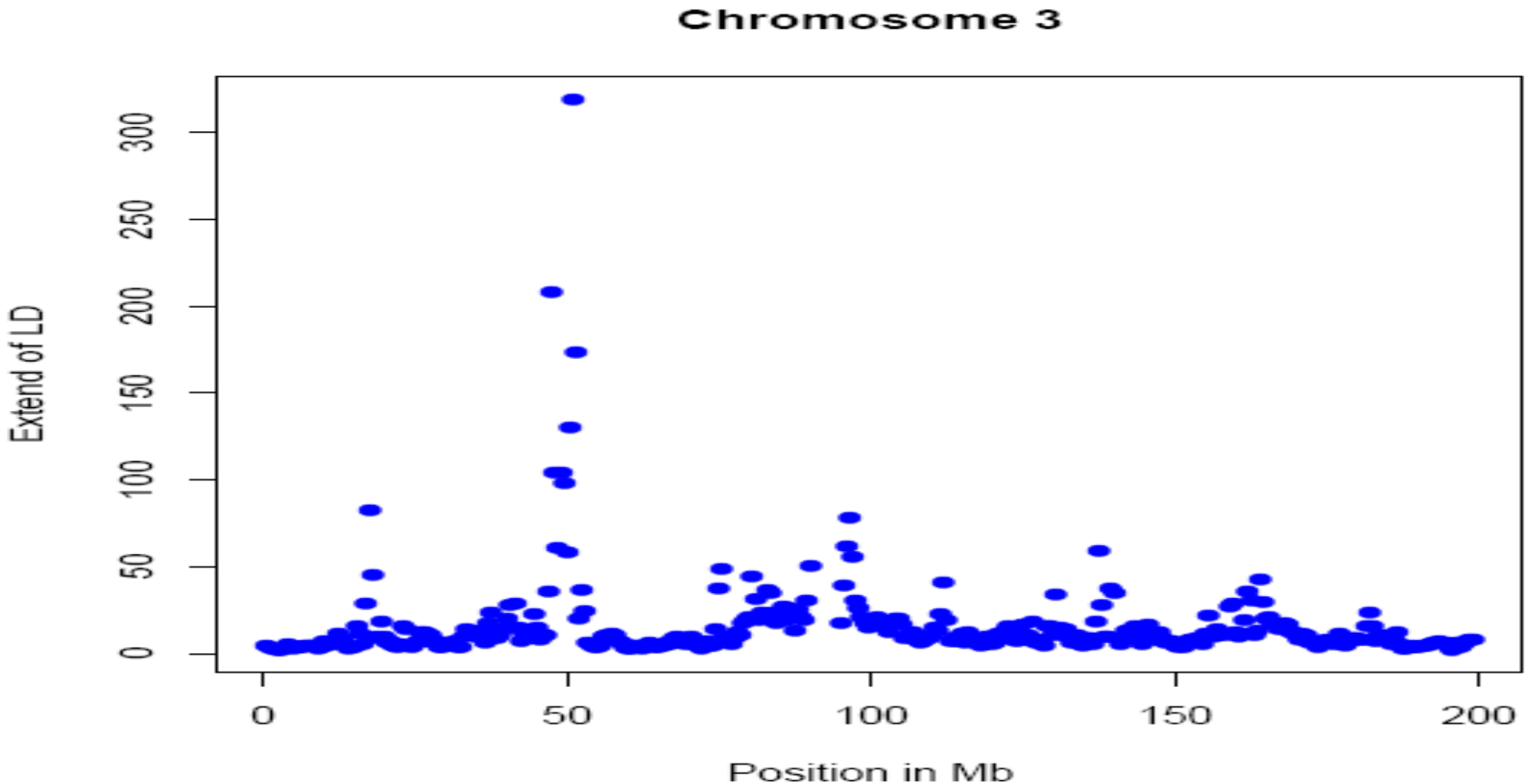
Comparing Genomic Regions …

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

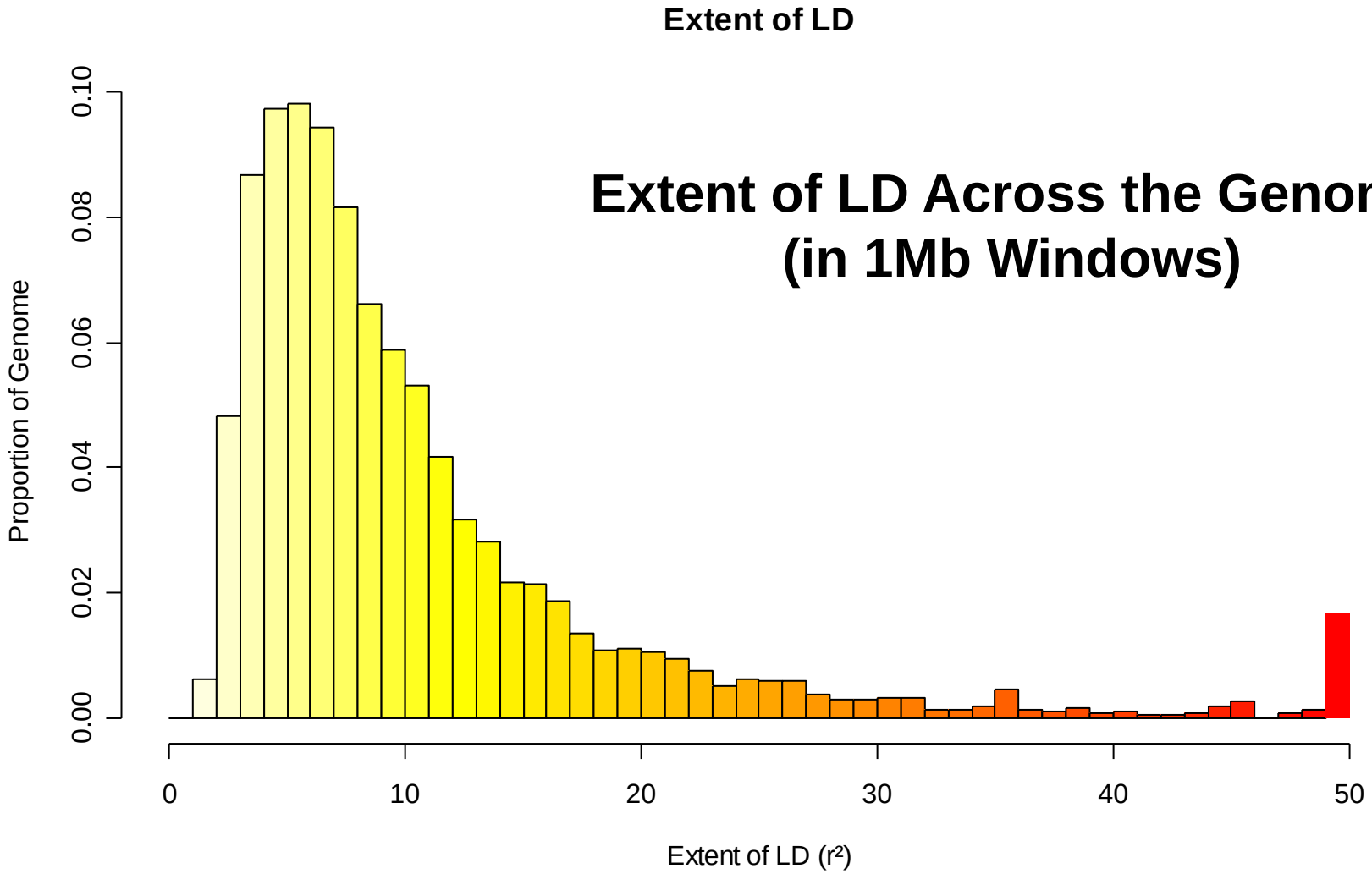
- One common summary is the distance at which the curve crosses a threshold of interest (say 0.50)



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

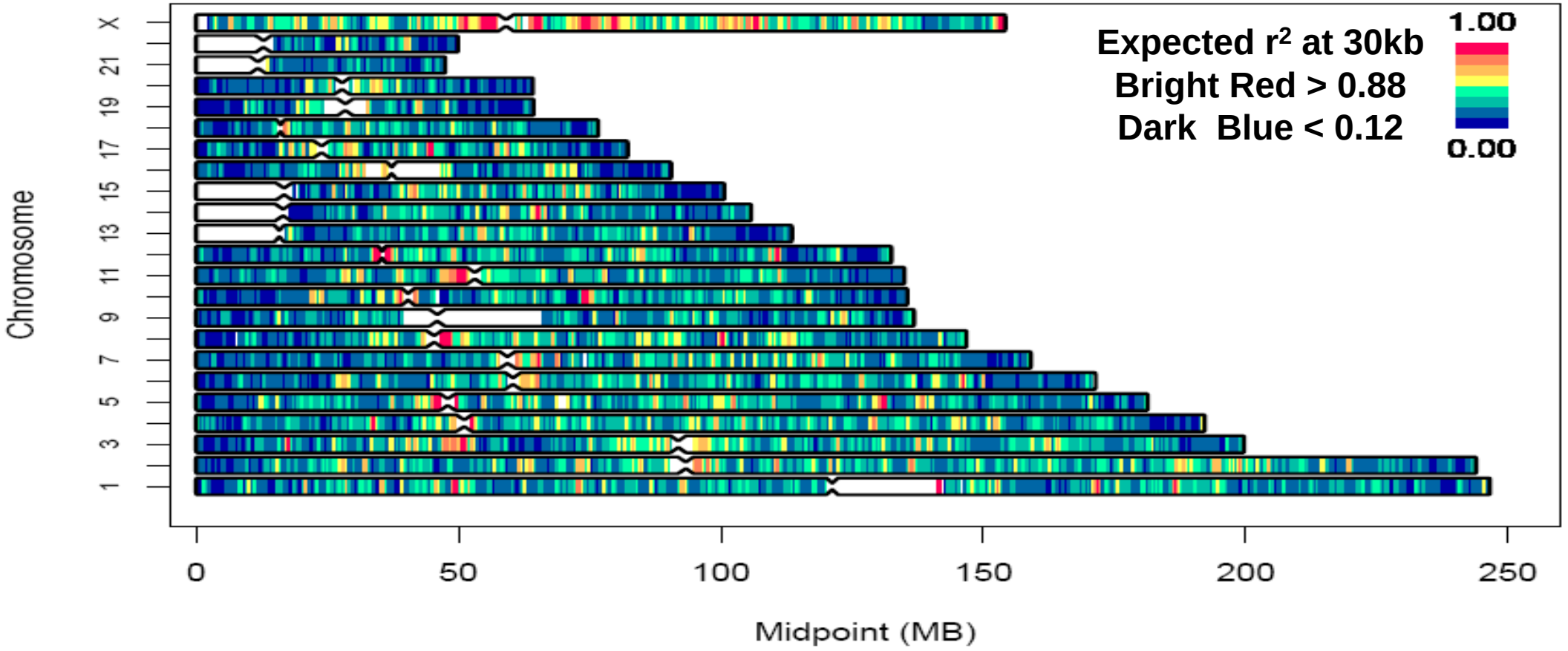


But within each chromosome, there is still huge variability!



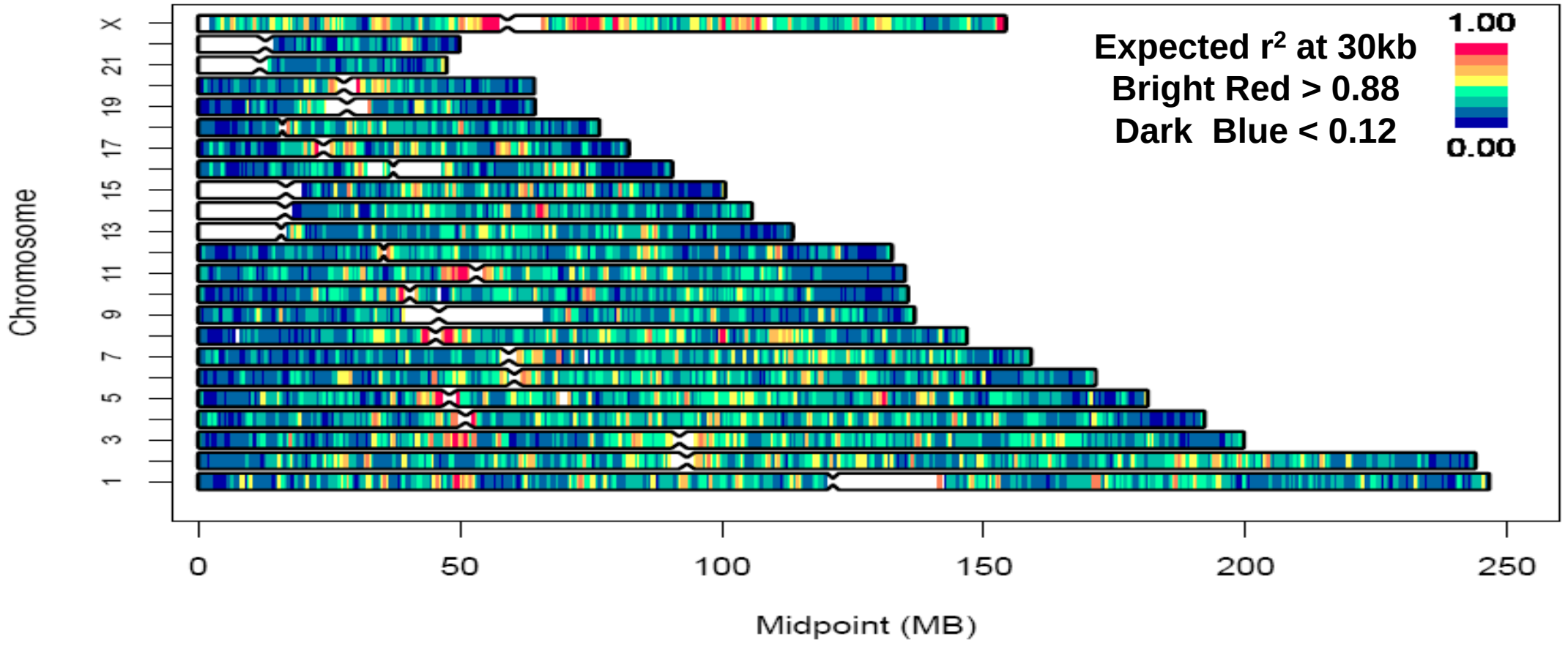
Average Extent:	**11.9 kb**
Median Extent:	**7.8 kb**
10th percentile:	**3.5 kb**
90th percentile:	**20.9 kb**

Genomic Variation in LD (CEPH)



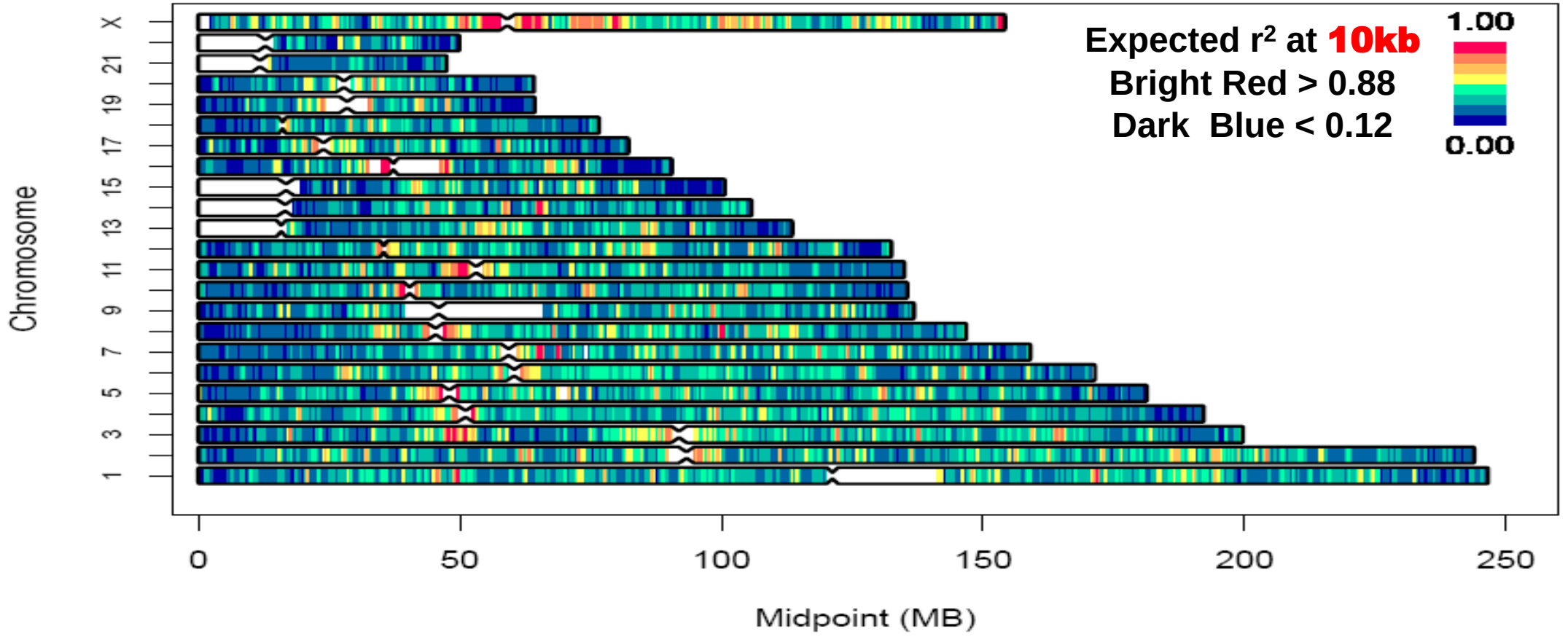
Smith et al, *Genome Research, 2005*

Genomic Variation in LD (JPT + CHB)



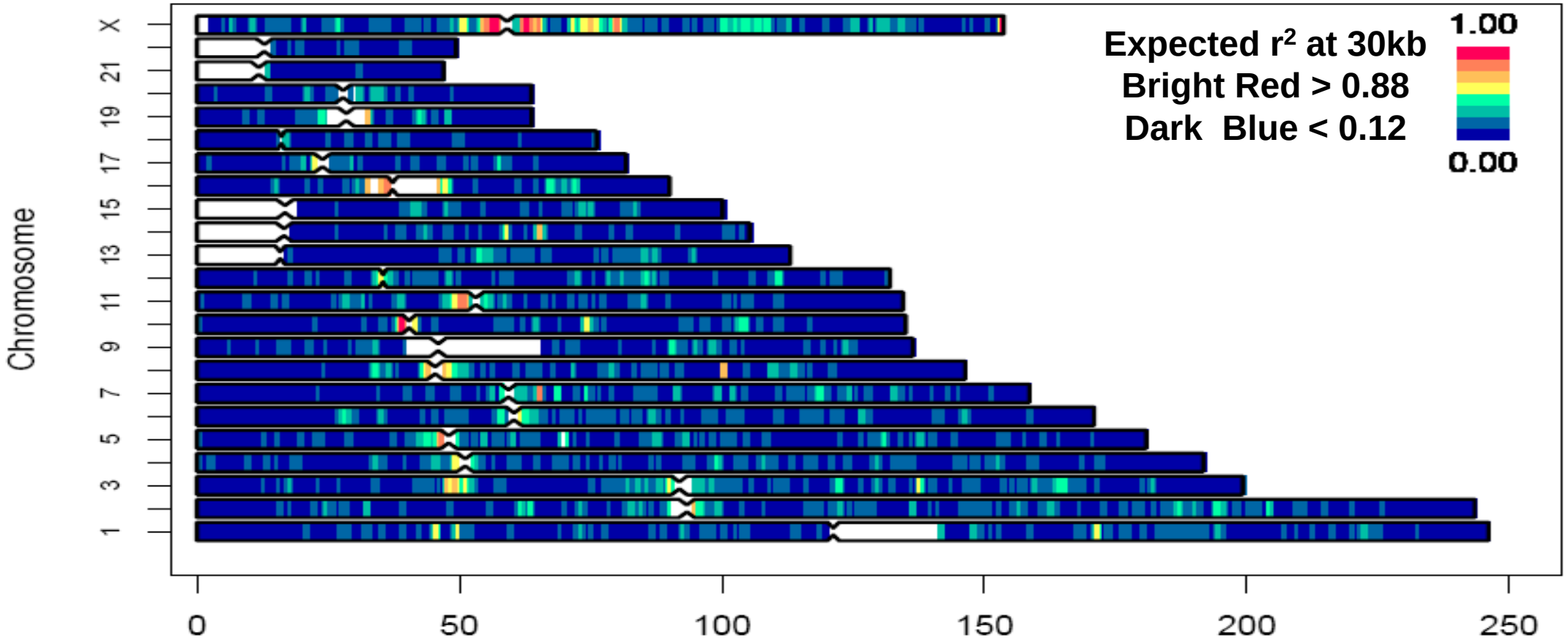
Smith et al, *Genome Research, 2005*

Genomic Variation in LD (YRI)



Smith et al, *Genome Research, 2005*

Genomic Distribution of LD (YRI)



Smith et al, *Genome Research, 2005*

Sequence Composition vs. LD
(some selected comparisons)

		Genome Quartiles, Defined Using LD (Low LD)			(High LD)	
	Genome	**Q1**	**Q2**	**Q3**	**Q4**	**Trend**
Basic Sequence Features						
GC Bases (%)	40.8	43.5	41.0	39.6	39.0	Decreases With LD
Bases in CpG Islands (%)	0.7	0.9	0.7	0.6	0.6	Decreases With LD
Polymorphism ($\Pi * 10{,}000$)	10.1	11.9	10.6	9.6	8.3	Decreases With LD
Genes and Related Features						
Known Genes (per 1000 kb)	*6.4*	*6.6*	*6.1*	*6.2*	*6.7*	U shaped
Genic Bases (Exon, Intron, UTR, %)	38.5	37.6	34.6	36.0	45.9	U shaped
Coding Bases (%)	1.2	1.1	1.0	1.1	1.4	U shaped
Conserved Non-Coding Sequence (%)	1.4	1.6	1.5	1.3	1.1	Decreases with LD
Repeat Content						
Total Bases in Repeats (%)	47.9	44.2	46.4	48.6	52.3	Increases with LD
Bases in LINE repeats (%)	20.9	16.5	19.9	22.4	24.9	Increases with LD
Bases in SINE repeats (%)	13.6	14.7	13.1	12.6	14.0	U shaped

Smith et al, *Genome Research, 2005*

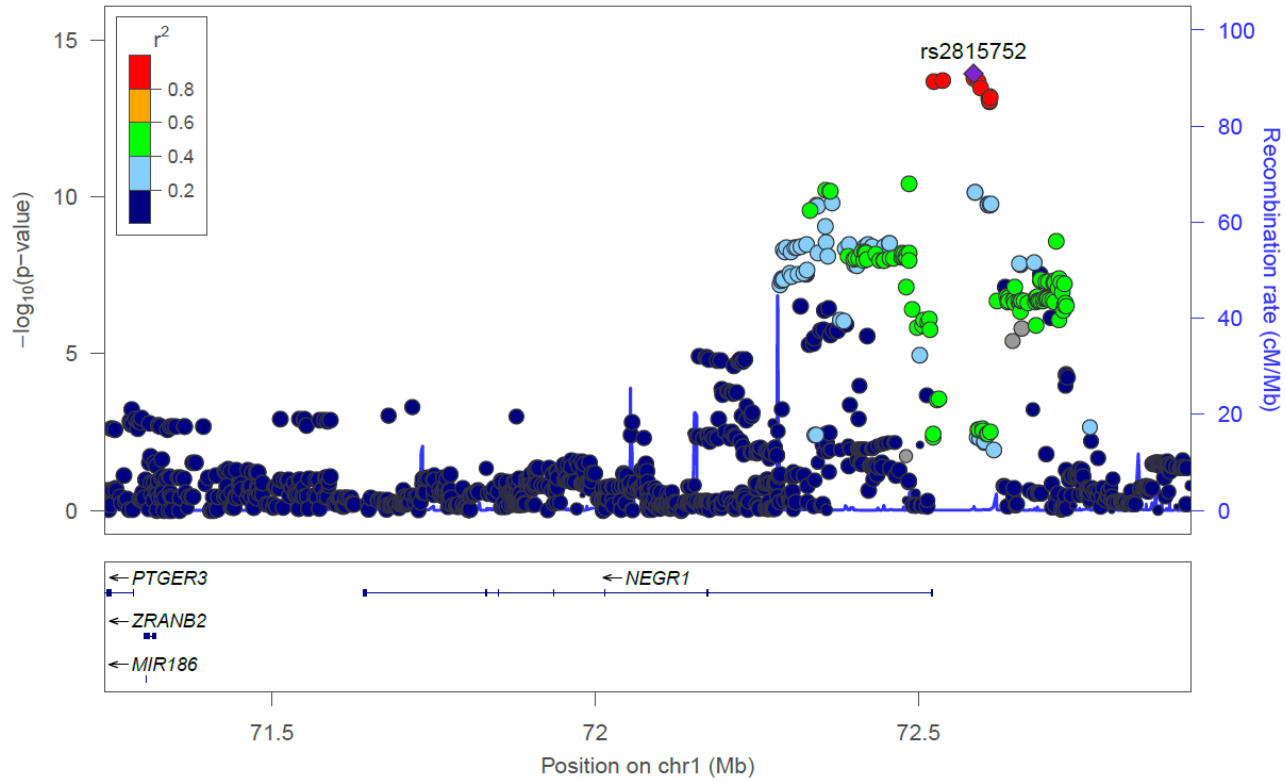
Gene Function in Regions of High and Low Disequilibrium

Gene Function (GO Term)	Annotated Genes	Low LD	High LD	χ^2	P-value
All Swissprot Entries Examined	**7520**	**2305**	**2045**	-	-
DNA metabolism	366	74	139	35.37	<.0001
Immune Response	622	232	94	34.36	<.0001
Cell cycle	493	119	177	26.24	<.0001
Protein Metabolism	1193	318	375	23.33	<.0001
Organelle Organization and Biogenesis	444	107	152	19.65	<.0001
Intracellular Transport	263	56	95	19.61	<.0001
Organogenesis	805	294	162	16.49	0.00005
Cell Organization and Metabolism	545	138	178	16.43	0.00005
RNA Metabollism	208	41	71	15.33	0.00009

Results from a comparison of the distribution of 40 most common gene classifications in the GENE Ontology Database

Implications for Association Studies

Obesity and the *NEGR1* locus



Multiple nearby SNPs show evidence for association with obesity.
The associated alleles appear together often, making it hard to pinpoint the causal variant.
Rapid shifts in association signal often correspond to recombination ‘hotspots’.

Willer et al, *Nature Genetics, 2009*

Linkage Disequilibrium in Association Studies:

Tag SNP Picking

- Many nearby SNPs will typically provide similar evidence for association

- To decrease genotyping costs, most association studies will examine selected “tag SNPs” in each region

- The most common tagging strategy focuses on pairwise r^2 between SNPs

Linkage Disequilibrium in Association Studies:

Pairwise Tagging Algorithm

- Select an r^2 threshold, typically 0.5 or 0.8
  - SNPs with r^2 above threshold can serve as proxies for each other
- For each marker being considered, count the number of SNPs with r^2 above threshold
- Genotype SNP with the largest number of pairwise “proxies”
- Remove SNP and all the SNPs it tags from consideration
- Repeat the previous three steps until there are no more SNPs to pick or genotyping budget is exhausted

Carlson et al, AJHG, 2004

Potential Number of tag SNPs

Table 3 | Number of tag SNPs required to capture common (MAF $\geq$ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

Current tag SNP panels typically examine 500,000 – 1,000,000 SNPs for a cost of $50 - $100 per sample.

The International HapMap Consortium, *Nature*, 2007

Today …

- Basic descriptors of linkage disequilibrium

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

Additional Reading I

- Dawson E et al (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature **418:**544-548

- The International HapMap Consortium. (2005). A haplotype map of the human genome. Nature 437:1299-320

- Carlson CS et al (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106-120

Additional Reading II

- Cardon and Bell (2001) Association study designs for complex diseases. *Nature Reviews Genetics* **2**:91-99

- Surveys important issues in analyzing population data.

- Illustrates shift from focus on linkage to association mapping for complex traits.