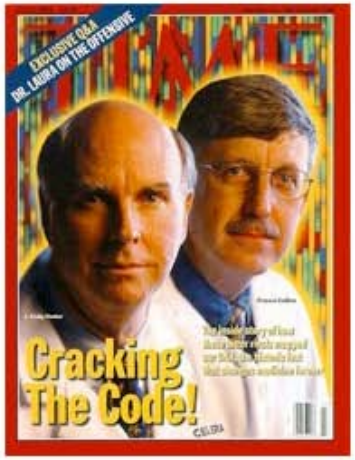
Human Genetic Diversity

2000

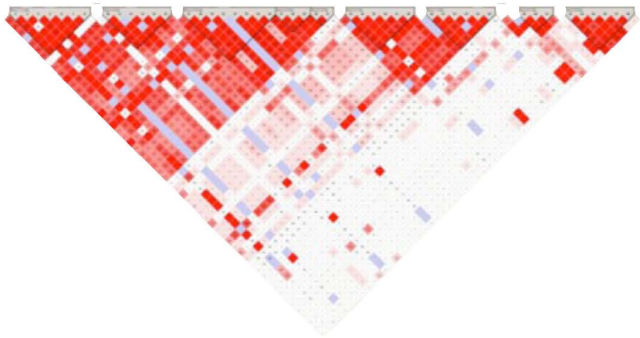


0.1% x 3.3 billion
= 3,300,000 bp of differences

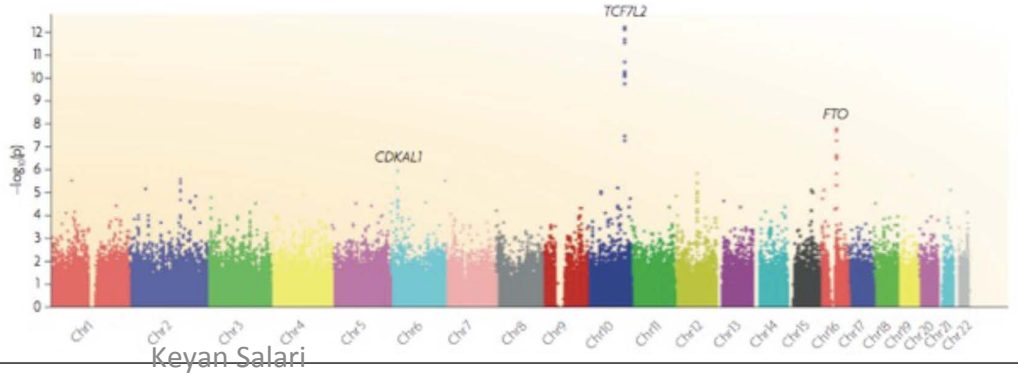
"I believe one of the great truths to emerge from this triumphant expedition inside the human genome is that in genetic terms, all human beings, regardless of race, are more than 99.9 percent the same."
President Bill Clinton, June 26, 2000, The White House East Room

1. Natural variation in the human genome

2. Genetic Association & Linkage Disequilibrium



3. Genome-wide association studies



Steve Quake

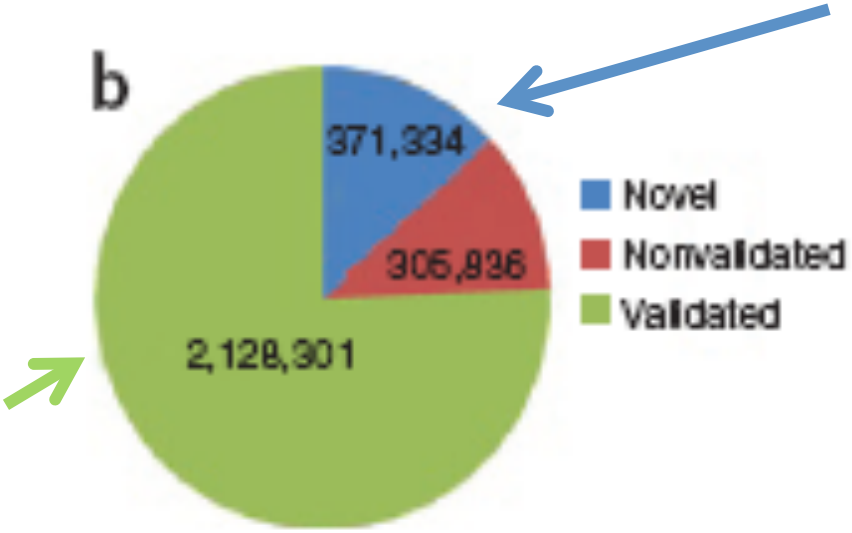
nature biotechnology

LETTERS

Single-molecule sequencing of an individual human genome

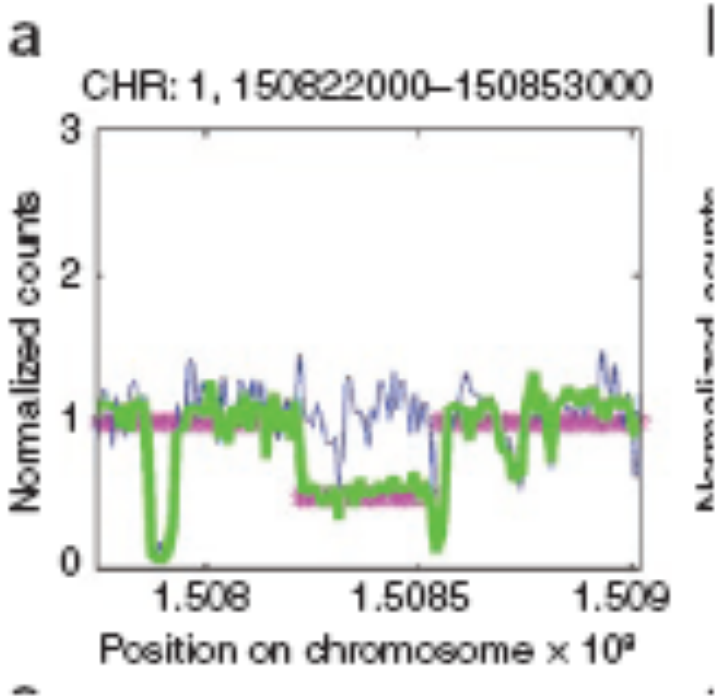
Dmitry Pushkarev[1,2], Norma F Neff[1,2] & Stephen R Quake[1]

Nature Biotech 27, 847, 2009



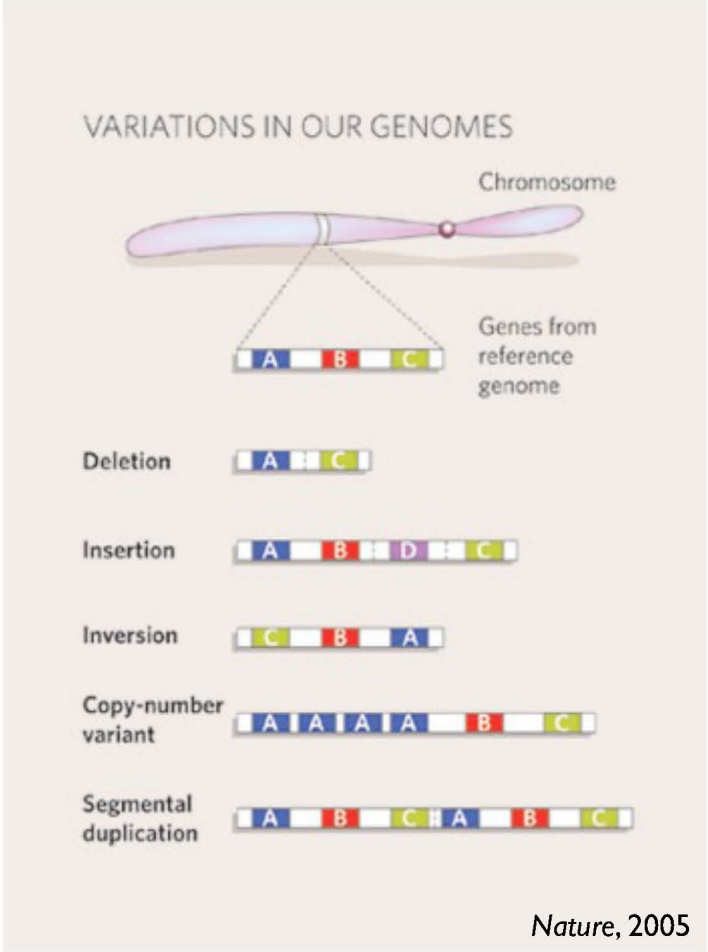
- 2.8 Million SNPs
  - 371 Thousand SNPs (13%) are novel

Copy number variation



- 752 copy number variations
- 16 Mb total

Human Genetic Variation



Nature, 2005

Structural variation

- 12% of our genome
- thousands of genes, disease loci, functional elements
- likely role in phenotypic variation and human disease

Redon *et al. Nature*, 2006

Genetic variation for a simple trait

Chr12: ALDH2 - SNP rs671

```
...GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
...GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
... G  L  Q  A  Y  T  E  V  K  T  V  S  V
```

Genotype: G/G

Protein: functional

Phenotype: none

Chr12: ALDH2 - SNP rs671

```
... GGGCTGCAGGCATACACTGAAGTGAAAACTGTGAGTGTG
... GGGCTGCAGGCATACACTAAAGTGAAAACTGTGAGTGTG
... G  L  Q  A  Y  T E/K V  K  T  V  S  V
```

Genotype: A/G

Protein: 1/2 functional

Phenotype: alcohol flush reaction

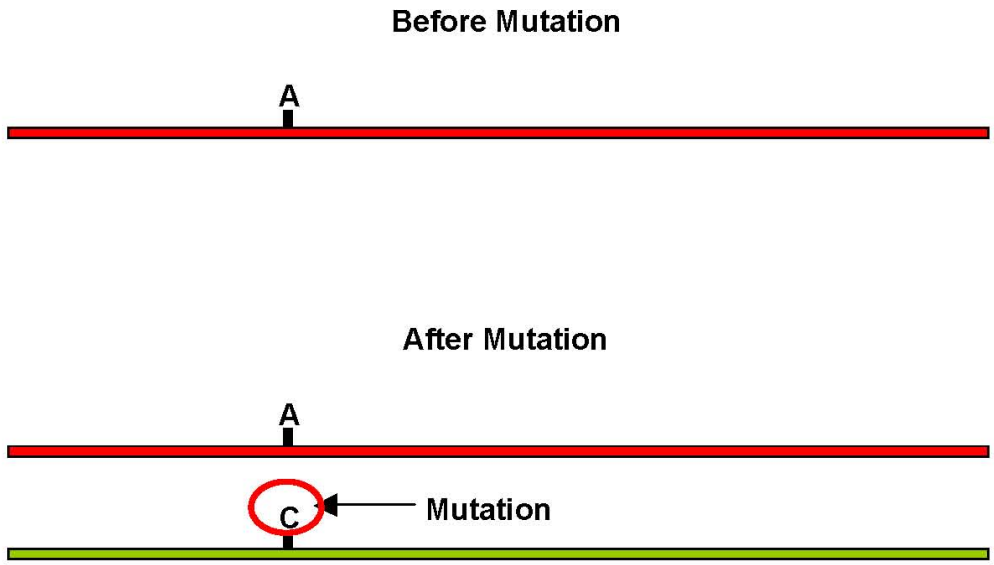
G allele functional
A allele missense (null)

CEU 100% G
YRI 100% G
CHB/JPT 76-84% G

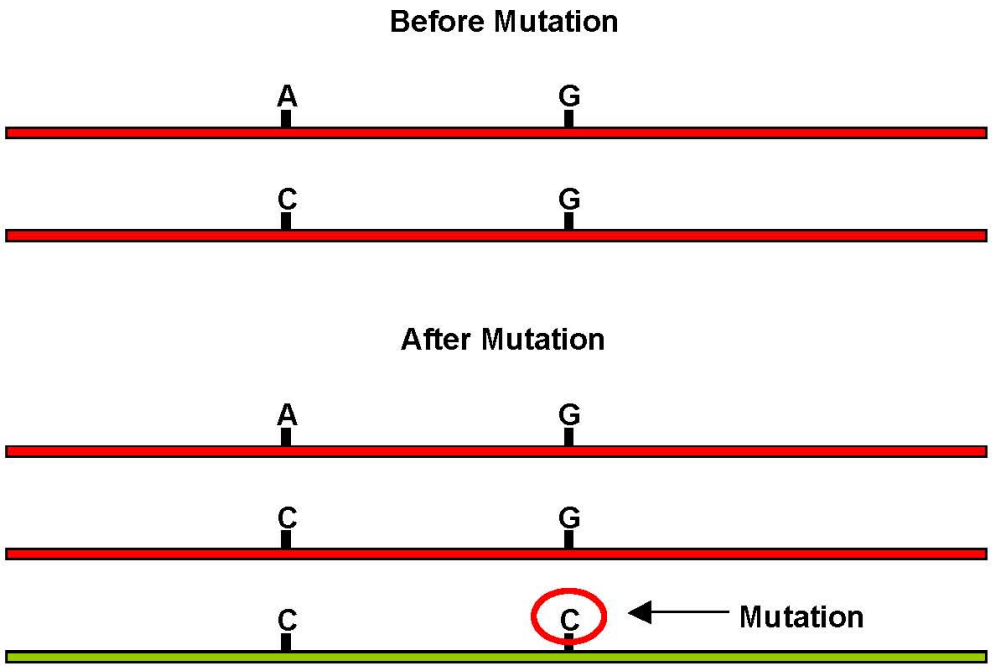
Today …

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

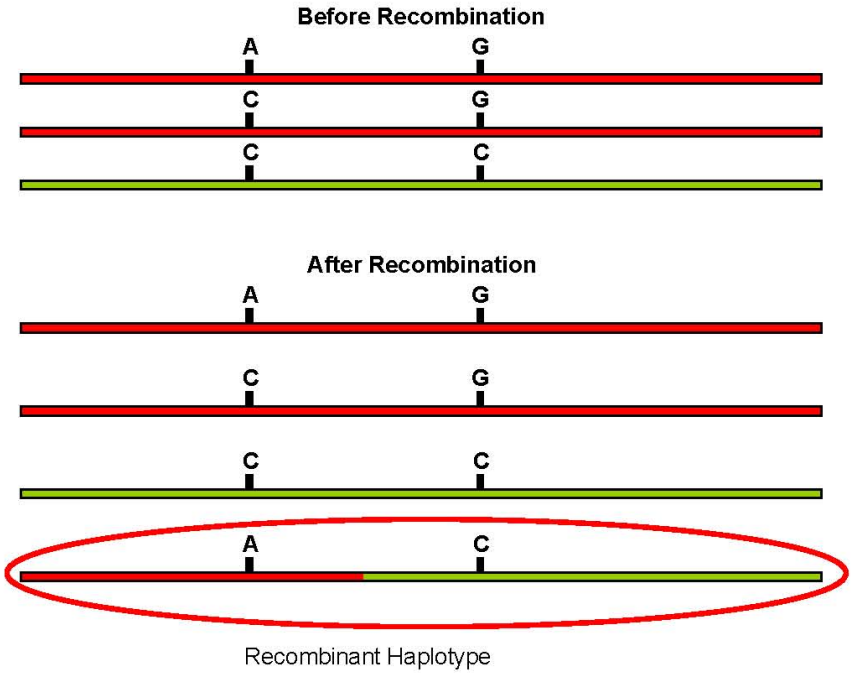
Alleles that exist today arose through ancient mutation events…



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

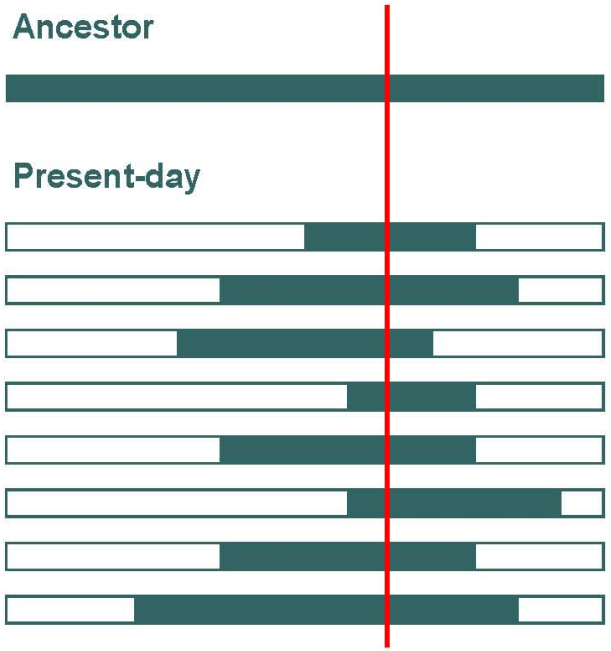


Recombination generates new arrangements for ancestral alleles



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



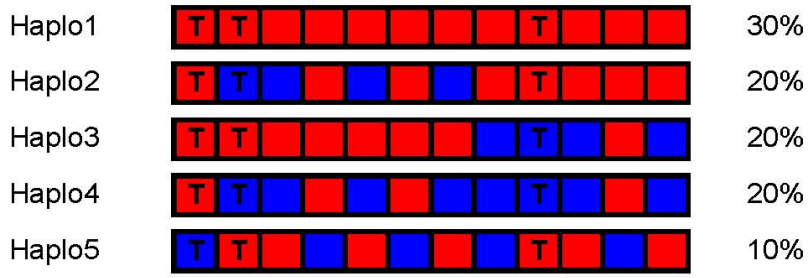
Why is linkage disequilibrium important for gene mapping?

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

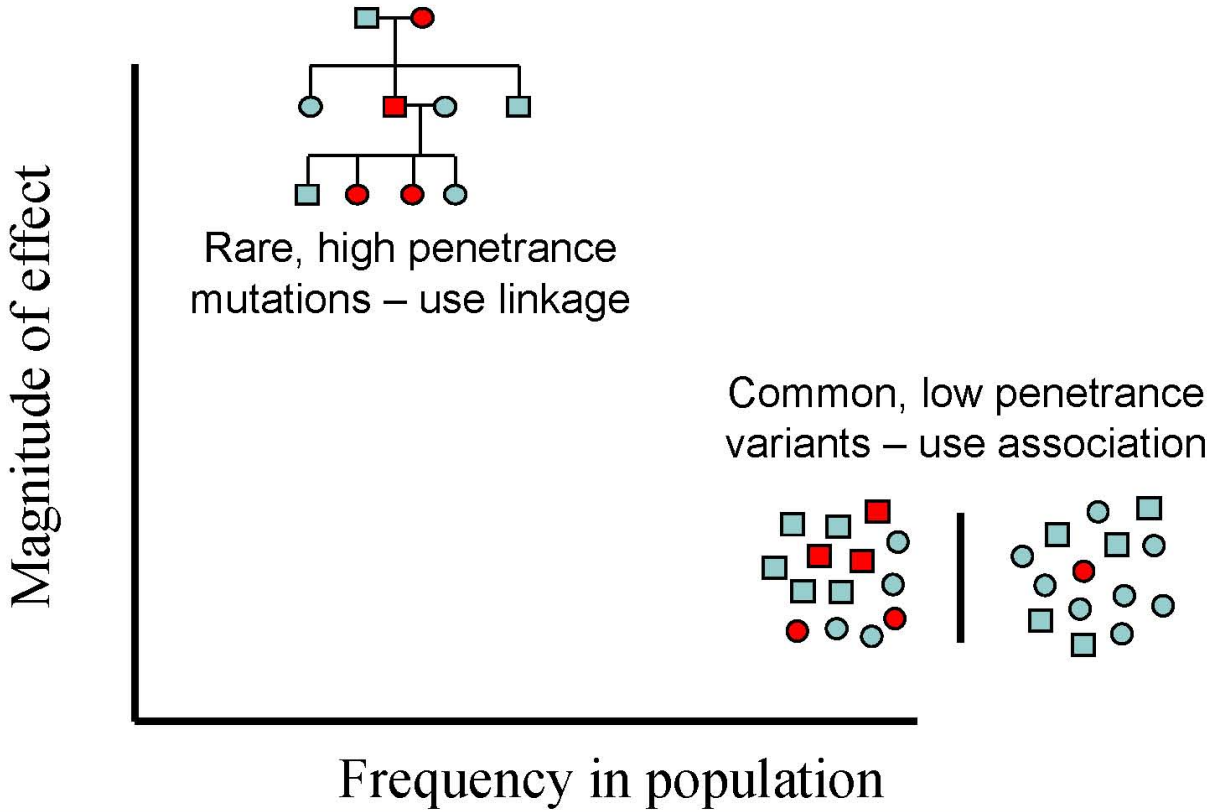


Linkage Disequilibrium Enables Genetic Association Studies

- 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?



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

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)

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

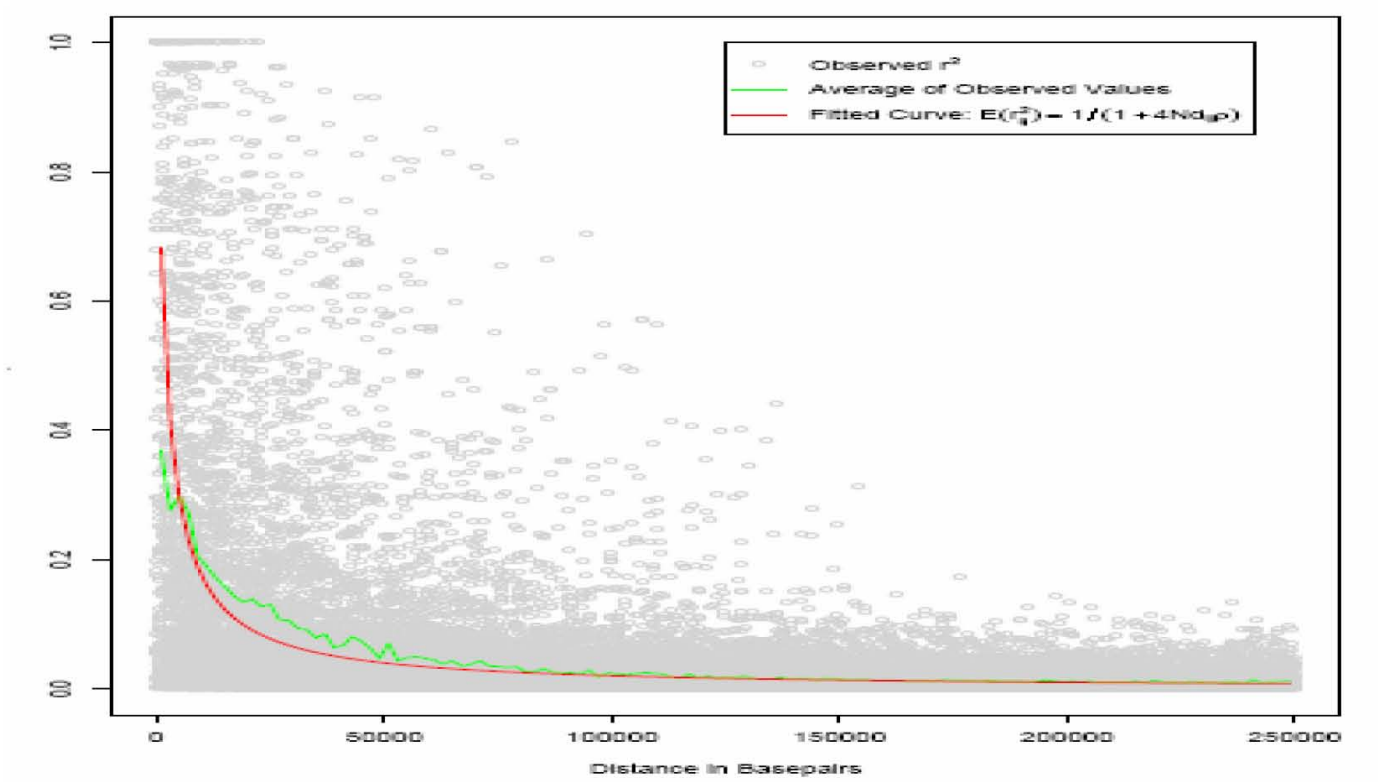
$$= \frac{\chi^2}{2n}$$

- 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

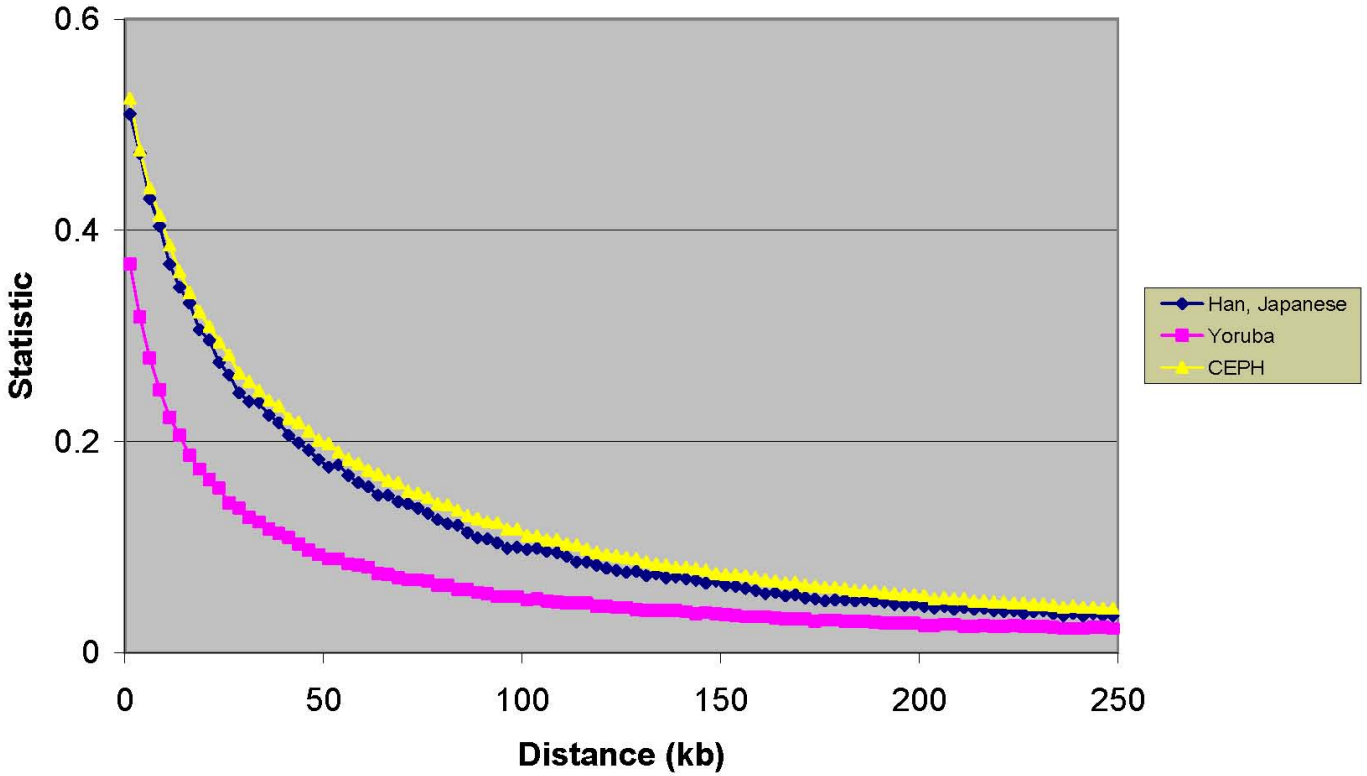
- $r^2 = 1$ implies the markers provide exactly the same information
- The measure preferred by population geneticists
- Measures loss in efficiency when marker A is replaced with marker B in an association study
  - With some simplifying assumptions (e.g. see Pritchard and Przeworski, 2001)

Summarizing Disequilibrium



slide created by Goncarlo Abecasis

Comparing Populations ...



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

slide created by Goncarlo Abecasis

Hum Genet (2002) 111 : 75–87
DOI 10.1007/s00439-002-0763-x

ORIGINAL INVESTIGATION

Stephanie M. Fullerton · Andrew G. Clark
Kenneth M. Weiss · Scott L. Taylor · Jari H. Stengård
Veikko Salomaa · Eric Boerwinkle
Deborah A. Nickerson

Sequence polymorphism at the human apolipoprotein AII gene (*APOA2*): unexpected deficit of variation in an African-American sample

Sequence APOA2 in 72 people

Look at patterns of polymorphisms

Chimp		Site no.[a]														
SNP haplotype no.	Sequence haplotype no.	1 5 5 C	2 0 1 G	1 8 7 T	1 2 1 G	2 6 7 1 ?	2 0 3 8 G	2 0 8 5 C	2 1 1 5 G	2 2 3 3 C	2 8 1 8 C	2 8 6 8 C	3 9 9 4 C	3 0 2 7 T	3 0 9 2 A	2 0 8 G

Find polymorphisms at these positions.

Reference sequence is listed.

Chimp		Site no.[a]															Sample				
SNP haplotype no.	Sequence haplotype no.	1 5 5 C	2 0 1 G	1 8 7 2 T	1 2 1 8 G	2 6 7 1 ?	2 0 3 8 G	2 0 8 5 C	2 1 1 5 G	2 2 3 3 C	2 8 1 8 C	2 8 6 8 C	3 9 9 4 C	3 0 2 7 T	3 0 9 2 A	2 0 8 G	J		N	R	T
Core re-sequenced samples																					
	S9	G		C		20	●		A	●	●	●	●	●	●	●		0	0	1	1

Sequence of the first chromosome.

Circle is same as reference.

Chimp		Site no.[a]														Sample			
SNP haplotype no.	Sequence haplotype no.	15 5 C	20 1 G	18 7 2 T	12 8 G	26 7 1 ?	20 3 8 G	20 8 5 C	21 1 5 G	22 3 3 C	28 1 8 C	28 6 8 C	39 9 4 C	30 2 7 T	30 9 2 A	20 8 G	J	N	R	T
Core re-sequenced samples																			
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1

Sequence of the second chromosome.

Same as the first except at position 2671.

Chimp		Site no.[a]															Sample			
SNP haplotype no.	Sequence haplotype no.	15 5 C	20 1 G	18 7 2 T	12 1 8 G	26 7 1 ?	20 3 8 G	20 8 5 C	21 1 5 G	22 3 3 C	28 1 8 C	28 6 8 C	39 9 4 C	30 2 7 T	30 9 2 A	20 8 G	J	N	R	T
Core re-sequenced samples																				
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1
	S2	G		C		19	●		●	●	●	●	●	●	●	●	15	10	12	37
	S2a	G		C		20	●		●	●	●	●	●	●	●	●	0	2	3	5
	S2b	G		C		18	●		●	●	●	●	●	●	●	●	0	2	1	3
	S2c	G		C		21	●		●	●	●	●	●	●	●	●	1	0	1	2
	S1d	G		●		19	●		●	●	●	●	●	●	●	●	5	0	0	5
	S1	G		●		16	●		●	●	●	●	●	●	●	●	17	19	14	50
	S1a	G		●		18	●		●	●	●	●	●	●	●	●	5	1	0	6
	S1b	G		●		15	●		●	●	●	●	●	●	●	●	2	0	0	2
	S1c	G		●		17	●		●	●	●	●	●	●	●	●	1	0	0	1
	S6	●		●		16	●		●	●	●	●	●	●	●	●	1	2	0	3
	S5	●		●		14	●		●	T	●	A	●	●	●	●	1	4	2	7
	S3	●		●		14	●		●	T	●	A	●	C	G	A	0	3	6	9
	S7	●		●		13	C		●	●	T	●	●	●	●	●	0	2	0	2
	S8	●		●		13	C		●	●	T	●	●	C	G	●	0	1	1	2
	S4	●		●		13	C		●	●	T	●	T	C	G	●	0	1	6	7
	S4a	?		●		14	C		●	●	T	●	T	C	G	●	0	0	1	1

Calculate D’ and r^2 for 2818 and 3027

Chimp		Site no.[a]														Sample			
SNP haplotype no.	Sequence haplotype no.	15 5 C	20 1 G	18 7 2 T	12 1 8 G	26 7 1 ?	20 3 8 G	20 8 5 C	21 1 5 G	22 3 3 C	28 1 8 C	28 6 8 C	39 9 4 C	30 2 7 T	30 9 2 A	20 8 G	J	N	R	T
Core re-sequenced samples																			
	S9	G		C		20	●		A	●	●	●	●	●	●	●	0	0	1	1
	S9a	G		C		18	●		A	●	●	●	●	●	●	●	0	1	0	1
	S2	G		C		19	●		●	●	●	●	●	●	●	●	15	10	12	37
	S2a	G		C		20	●		●	●	●	●	●	●	●	●	0	2	3	5
	S2b	G		C		18	●		●	●	●	●	●	●	●	●	0	2	1	3
	S2c	G		C		21	●		●	●	●	●	●	●	●	●	1	0	1	2
	S1d	G		●		19	●		●	●	●	●	●	●	●	●	5	0	0	5
	S1	G		●		16	●		●	●	●	●	●	●	●	●	17	19	14	50
	S1a	G		●		18	●		●	●	●	●	●	●	●	●	5	1	0	6
	S1b	G		●		15	●		●	●	●	●	●	●	●	●	2	0	0	2
	S1c	G		●		17	●		●	●	●	●	●	●	●	●	1	0	0	1
	S6	●		●		16	●		●	●	●	●	●	●	●	●	1	2	0	3
	S5	●		●		14	●		●	T	●	A	●	●	●	●	1	4	2	7
	S3	●		●		14	●		●	T	●	A	●	C	G	A	0	3	6	9
	S7	●		●		13	C		●	●	T	●	●	●	●	●	0	2	0	2
	S8	●		●		13	C		●	●	T	●	●	C	G	●	0	1	1	2
	S4	●		●		13	C		●	●	T	●	T	C	G	●	0	1	6	7
	S4a	?		●		14	C		●	●	T	●	T	C	G	●	0	0	1	1

Fill out this table.

X11 is number of times that haplotype is seen.

	2818 C	**2818 T**	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

Fill out this table.
Convert all numbers to frequencies.

	2818 C	**2818 T**	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

Calculate D and D’

	2818 C	2818 T	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

D = x11 – p1q1

D_{max} is given by the smaller of p_1q_2 and p_2q_1

D’ = D/Dmax

Calculate r^2

	2818 C	2818 T	
3027 T	X11	X21	# 3027 T alleles
3027 C	X12	x22	#3027 C alleles
	# 2818 C Allele	# 2818 T allele	

$r^2 = D^2/p1p2q1q2$