

Neutral Evolution at Two Loci

Alan R. Rogers

March 24, 2021

Linkage disequilibrium (LD)

Gamete	Locus	
	1	2
1	A	B
2	A	B
3	A	B
4	A	B
5	A	B
6	A	b
7	a	B
8	a	B
9	a	b
10	a	b

	A	a	
B	5	2	7
b	1	2	3
	6	4	10

- B is more common among A -gametes than a -gametes.
- A is more common among B -gametes than b -gametes.
- This is LD.

1 / 28

2 / 28

Linkage equilibrium (LE)

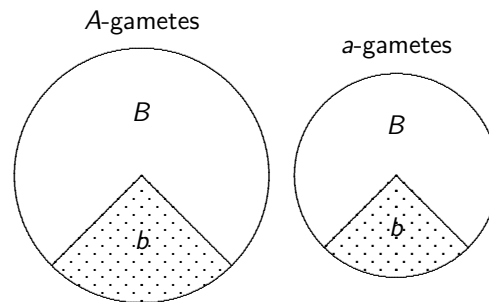
Gamete	Locus	
	1	2
1	A	B
2	A	B
3	A	B
4	A	B
5	A	b
6	A	b
7	a	B
8	a	B
9	a	b

	A	a	
B	4	2	6
b	2	1	3
	6	3	9

- B is equally common among A -gametes and a -gametes.
- A is equally common among B -gametes and b -gametes.
- This is LE.

3 / 28

Linkage Equilibrium \iff shaded fractions equal



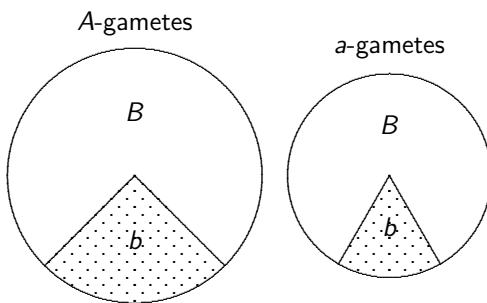
LE: Knowledge about one locus tells nothing about other.

Here, b is equally common among a -gametes and A -gametes.

(Also, a is equally common among b -gametes and B -gametes.)

4 / 28

LD \iff shaded fractions unequal



LD: Knowledge about one locus helps predict the other.

Here, b is more common among A -gametes than among a -gametes.

(Also, a is more common among B -gametes than among b -gametes.)

5 / 28

People sometimes confuse LD with epistasis

Epistasis

In a comparison of diploid genotypes, the *phenotypic effect* of one locus depends on another locus.

Linkage disequilibrium

In a comparison of haploid chromosomes, the *allele frequency* at one locus depends on another locus.

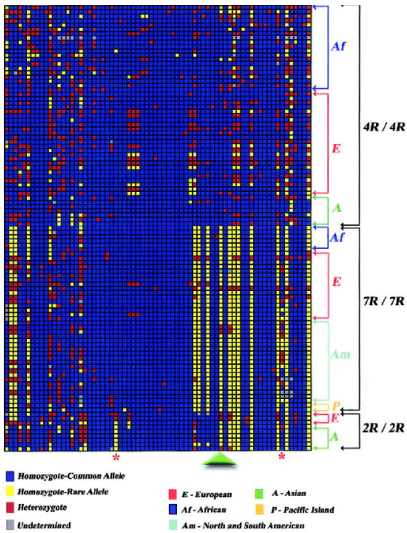
6 / 28

You can see LD in sequence data

		Nucleotide position															
		3	8	1	1	1	1	1	1	2	2	2	2	2	2	2	2
		1	9	4	6	4	4	3	3	3	5	2	6	0			
		2	7	7	1	3	4	4	9	3	1	0	3	4			
Orang		T	G	C	A	T	G	T	A	A	C	G	C	T			
Chimp		T	G	C	A	T	G	T	A	A	T	G	C	T			
A		A	G	A	A	.			
B		A	G	A	A	.			
C		T	.	.	G	C			
D		C	G	G	.	.	.	C			
E		C	C	G	G	.	.	.	C			
F		C	C	G	.	.	.	C			
G		C	.	T	C	G	.	.	.	C			
H		C	.	T	G	.	.	.	C	G	.	.	.	C			

(GARRIGAN ET AL 2004)

- ▶ Dots: identical to chimp sequence.
- ▶ Sites not independent.
- ▶ A at site 1343 predicts G at 1951
- ▶ This is linkage disequilibrium (LD).

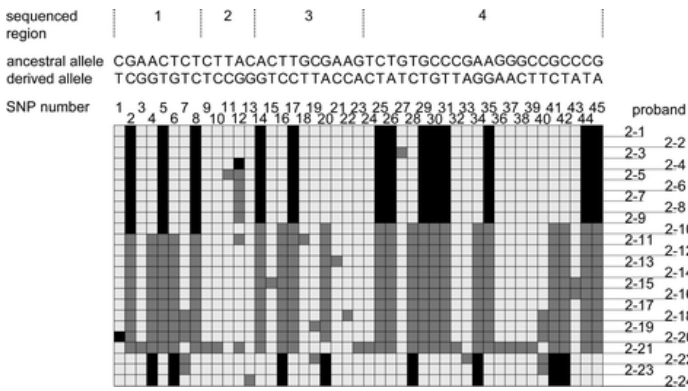


- ▶ Columns are SNPs
- ▶ Rows are diploid genotypes
- ▶ Blue: common homozygote
- ▶ Yellow: rare homozygote
- ▶ Red: heterozygote
- ▶ Note LD w/i 7R genotypes

7 / 28

8 / 28

LD at the NF1 locus (Schmegner et al 2005)



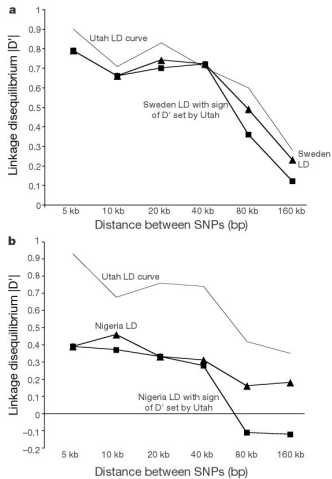
9 / 28

DNA sequences from region of human lactase gene

```
cgcttcaggcattcctatctaacaagacacgaAggggtacaatgctaaccagagctttcaactct
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
```

10 / 28

More LD in Europe than Africa

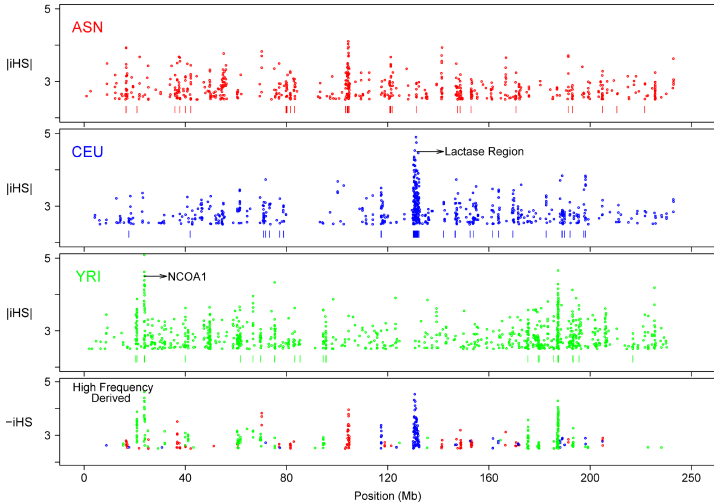


- ▶ LD declines with distance along chromosome
- ▶ More LD in Europe than Africa
- ▶ Why?

(REICH ET AL 2001)

11 / 28

LD unevenly distributed within genome



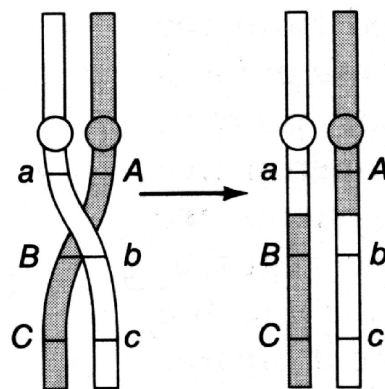
12 / 28

Facts we need to understand

- ▶ LD decays w/ distance along chromosome.
- ▶ Populations differ.
- ▶ Unevenly distributed w/i genome

13 / 28

Cross-overs shuffle DNA



- ▶ occur during reproduction.
- ▶ shuffle parental chromosomes.
- ▶ sites far apart more likely to recombine
- ▶ result: “recombinant” chromosomes

14 / 28

Why loci are independent on recombinants

```
|
| .....A.....b.....      A recombinant chromosome.
| \_____/ \_____/
|  from dad  from mom
|
|
|                               Gamete from Dad carried A.
|                               Gamete from Mom carried b.
|
|
| Probability of this?       $p_A p_b$  under random mating
```

15 / 28

Ingredients of a model

x_1 = frequency of AB -gametes among parents
 p_A = frequency of A -gametes among parents
 p_B = frequency of B -gametes among parents
 c = prob of recombination between the two loci

In any generation, there are two kinds of AB gamete:

1. non-recombinants: these were AB s in the last generation
Frequency: $(1 - c)x_1$
2. recombinants: formed from an A gamete and a B gamete, drawn at random. Frequency: cp_Ap_B

Next step: sum these contributions.

16 / 28

Model with random mating, no selection

x_1 = frequency of AB -gametes among parents
 p_A = frequency of A -gametes among parents
 p_B = frequency of B -gametes among parents
 c = prob of recombination between the two loci

Change in frequency of AB -gametes during one generation:

$$\begin{aligned} x_1' &= \overbrace{(1-c)x_1}^{\text{nonrecombinants}} + \overbrace{cp_{APB}}^{\text{recombinants}} \\ &= x_1 - c(x_1 - p_{APB}) \\ &= x_1 - cD \end{aligned}$$

17 / 28

Several equivalent definitions of D

The previous slide defined D , a measure of LD:

Gamete	Definition
AB	$D = x_1 - p_A p_B$
Ab	$-D = x_2 - p_A p_b$
aB	$-D = x_3 - p_a p_B$
ab	$D = x_4 - p_a p_b$

If the association between A and B is positive, then that between A and b must be negative. A more convenient formula:

$$D = x_1x_4 - x_2x_3$$

They all give the same answer.

18 / 28

Calculating *D*

Gamete	Locus		AB	Ab	aB	ab
	1	2	x_1	x_2	x_3	x_4
1	A	B				
2	A	B				
3	A	B				
4	A	B				
5	A	B				
6	A	b				
7	a	B				
8	a	B				
9	a	b				
10	a	b				

A

a

5

2

B

b

1

2

6

4

7

3

10

$$D = \frac{x_1 x_4 - x_2 x_3}{\frac{5}{10} \cdot \frac{2}{10} - \frac{1}{10} \cdot \frac{2}{10}}$$
$$= \frac{2}{25}$$

All four gametes, still no selection

Gamete	Recurrence
AB	$x'_1 = x_1 - cD$
Ab	$x'_2 = x_2 + cD$
aB	$x'_3 = x_3 + cD$
ab	$x'_4 = x_4 - cD$

How recombination affects *D*

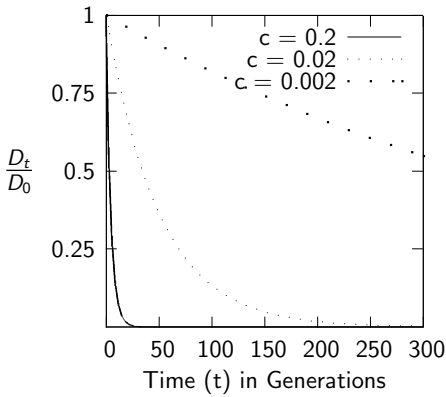
After one generation,

$$D' = x'_1 x'_4 - x'_2 x'_3$$
$$= (x_1 - cD)(x_4 - cD) - (x_2 + cD)(x_3 + cD)$$
$$= (1 - c)D$$

D declines each generation by a factor of $1 - c$.
After *t* generations,

$$D_t = D_0(1 - c)^t$$

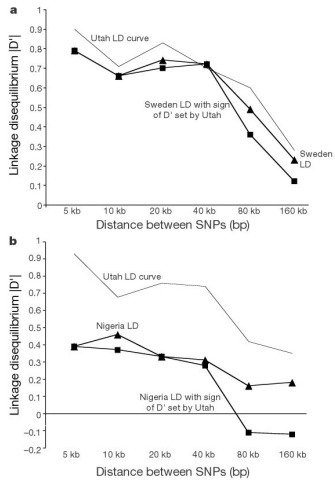
D declines gradually toward zero



- $c = 0.2$
 - ▶ Loci far apart
 - ▶ Loose linkage
 - ▶ LD declines rapidly
- $c = 0.02$
 - ▶ Loci closer
 - ▶ slower decline
- $c = 0.002$
 - ▶ Loci closer still
 - ▶ even slower decline

Is this theory enough to explain the data?

More LD in Europe than Africa



- ▶ *c* increases w/ distance along chromosome.
- ▶ Therefore LD should decline.
- ▶ But why more LD in Europe?

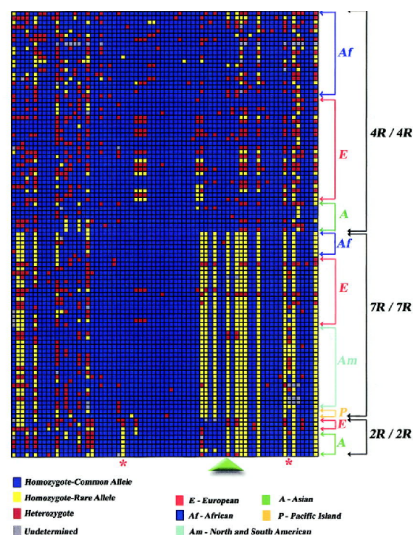
(REICH ET AL 2001)

		Nucleotide position															
		3		8	2		3	3	6	6	7	9	0	2	3	2	
		1	9	4	6	4	4	3	3	3	5	2	6	0			
		2	7	7	1	3	4	4	9	3	1	0	3	4			
Orang		T	G	C	A	T	G	T	A	A	C	G	C	T			
Chimp		T	G	C	A	T	G	T	A	A	T	G	C	T			
A		A	G	A	A	.	.	
B		A	G	A	.	.	.	
C		
D		C	.	C	G	G	C	.	
E		C	C	G	G	C	.	
F		C	C	G	C	.	
G		C	.	T	C	G	C	.	
H		C	.	T	G	.	.	.	C	G	C	.	

(GARRIGAN ET AL 2004)

- ▶ Short DNA sequence.
- ▶ tight linkage: c is small
- ▶ LD decays very slowly
- ▶ But why is it not zero?

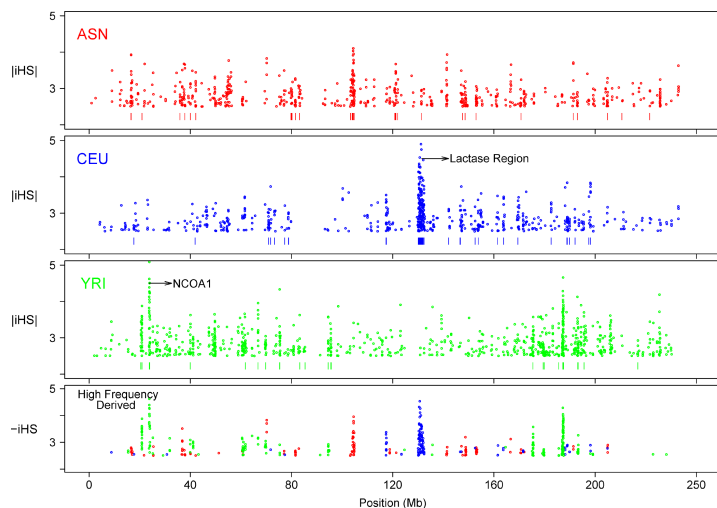
25 / 28



- ▶ Also a short sequence
- ▶ But why is there any LD?

26 / 28

Why is LD unevenly distributed?



27 / 28

Summary

- ▶ Our theory explains why D declines w/ distance along chromosome.
- ▶ If loci are far apart on chromosome, c is high and D declines rapidly.
- ▶ It tells us nothing about the forces that generate LD.
- ▶ Nothing about population differences.
- ▶ Nothing about variation across the genome.

28 / 28