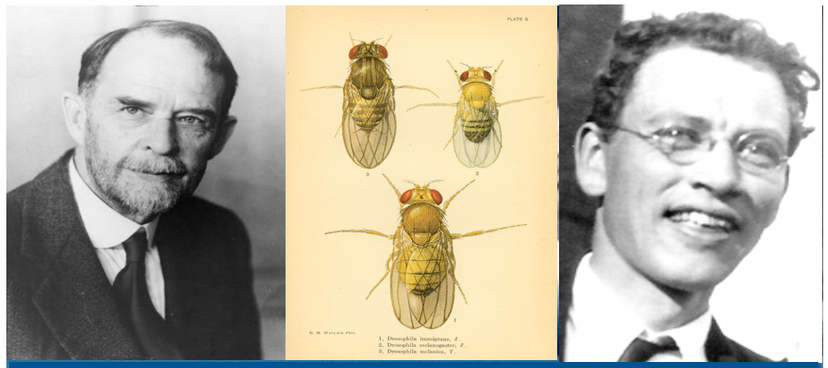
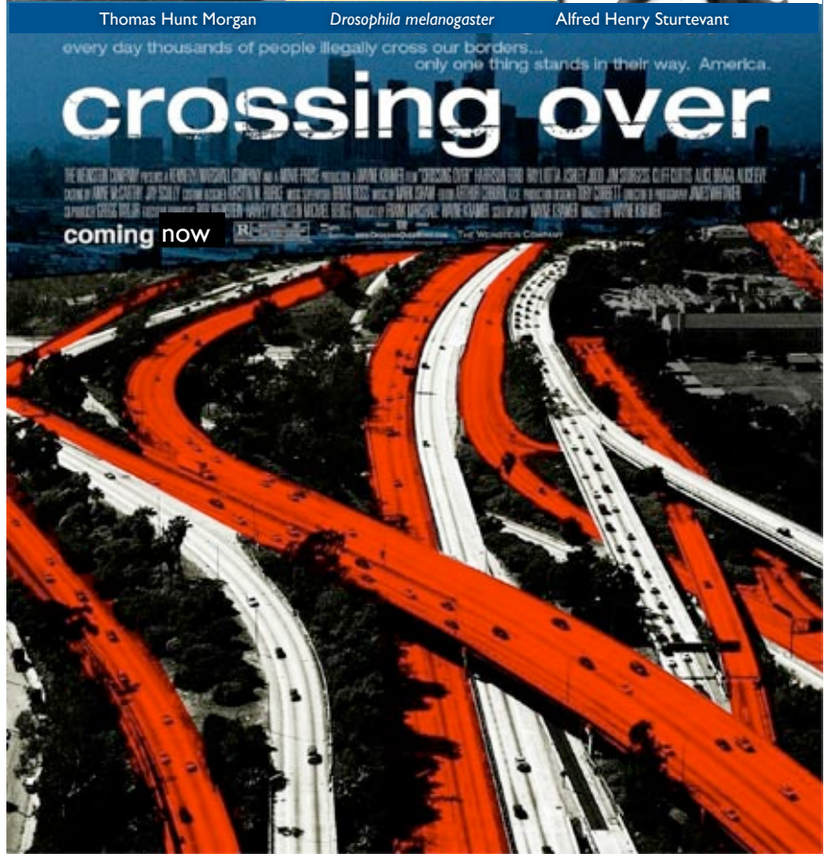
Evolutionary Genetics

LV 25600-01 | Lecture with exercises | 4KP

Linkage Disequilibrium



Thomas Hunt Morgan *Drosophila melanogaster* Alfred Henry Sturtevant



In our treatment of population genetics up to this point, we have often assumed that (a) **the transmission of alleles at a given locus across generations is independent of alleles at a second locus** and (b) **the fitness of genotypes at one locus is not affected by genotypes at another locus**. Both assumptions are likely to be violated for a reasonable number of loci.

When genetic variation at two or more loci is considered simultaneously, allele frequencies are not sufficient to describe their dynamics in natural populations. It is therefore necessary to consider the extent of non-random association of alleles at different loci. The non-random association of alleles at different loci is called gametic phase disequilibrium or, more simply, linkage disequilibrium.

Linkage equilibrium is a term used in genetics to describe the absence of association or linkage between alleles at different loci on a chromosome. In other words, it refers to a situation where the occurrence of a particular **allele at one genetic locus is independent of the alleles present at another locus**. It is a crucial concept in population genetics because deviations from linkage equilibrium can indicate genetic linkage between loci, meaning that certain alleles at different loci are inherited together more often than expected by chance.

Deviations from linkage equilibrium (**disequilibrium**) can indicate genetic linkage, genetic association, or the presence of evolutionary forces that affect allele frequencies within a population. Understanding patterns of linkage equilibrium and deviations from it helps geneticists study patterns of inheritance, genetic disease and evolutionary processes.

In a population at linkage equilibrium:

Independent segregation - Alleles at different loci segregate independently during meiosis. This means that the inheritance of one allele does not affect the inheritance of alleles at other loci.

Random association - Alleles of different genes assort randomly into gametes without any preferential association or linkage between them.

Hardy-Weinberg equilibrium - In the absence of factors such as mutation, migration, selection or genetic drift, a population that satisfies the conditions of HWE will exhibit linkage equilibrium among unlinked loci.

Linkage equilibrium occurs when the genotype present at one locus is **independent** of the genotype at a second locus.

D = 0

Linkage disequilibrium occurs when genotypes at the two loci are **not independent** of another.

D ≠ 0

Stumbling blocks:

a) non-random associations of alleles at two loci can occur even if the two genes are unlinked

b) just because two loci are linked does not mean that they will be in linkage disequilibrium

Locus A	Locus B
A_1	B_1
A_2	B_2

$$D_{min} = 0 \rightarrow N(gt) = ?$$

Locus A	Locus B
A_1	B_1
A_2	B_2

$$D_{min} = 0 \rightarrow N(gt) = 4$$

$$A_1B_1, A_1B_2, A_2B_1, A_2B_2$$

Locus A	Locus B
A_1	B_1
A_2	B_2

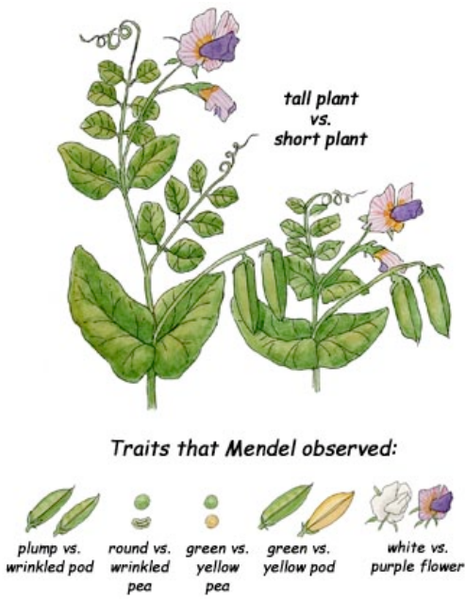
$D_{max} \rightarrow N(gt) = ?$

Locus A	Locus B
A_1	B_1
A_2	B_2

$D_{max} \rightarrow N(gt) = \mathbf{2}$

A_1B_1 / A_2B_2 or A_1B_2 / A_2B_1

Gregor Mendel (1823-1884)



Why did Mendel not encounter linkage?

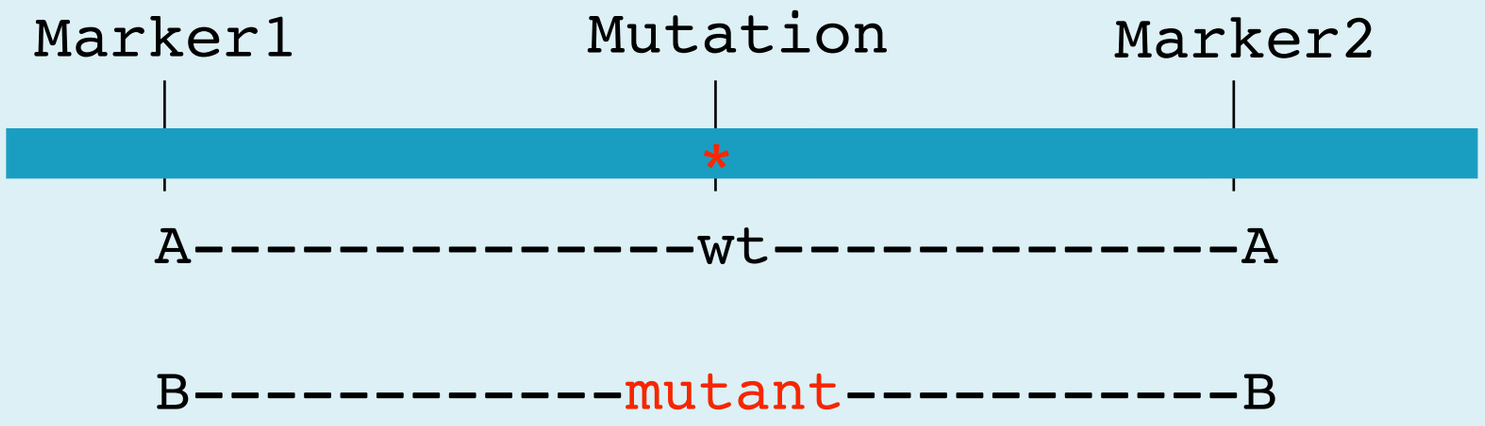
Mendel had no complications of linkage in his experiments. His (published) results showed that the seven genes (traits) he studied segregated freely. This is possible because the pea genome is organised into seven chromosomes. But this was not the case. Mendel worked with **three genes on chromosome 4**, **two genes on chromosome 1** and **one gene each on chromosomes 5 and 7**. At first glance, it seems that of the 21 dihybrid combinations that Mendel could theoretically have studied, no fewer than four should have resulted in linkage. However, as has been found in hundreds of crosses and as shown by the genetic map of the pea, the loci on chromosome 1 are so far apart on the chromosome that no linkage is normally detected. The same is true for two of the loci on chromosome 4. This leaves one locus that should have shown linkage. Mendel wrote that all 128 crosses with the seven genes were made, but not all of them were thoroughly analysed, including the one that might have shown linkage.

PopGen ▷ Linkage Disequilibrium

Marker1
A / B

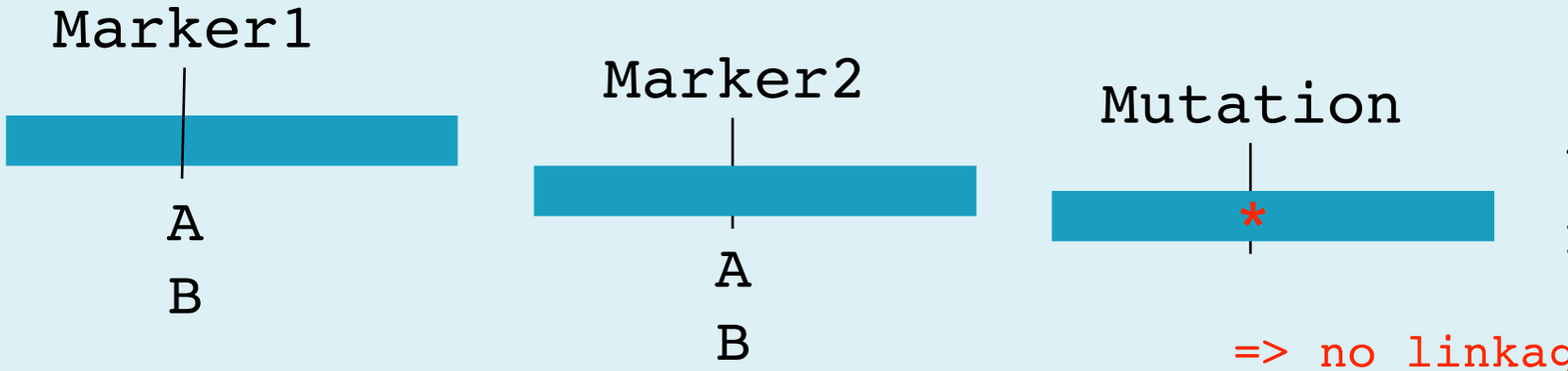
Phenotype
wt / mutant

Marker2
A / B



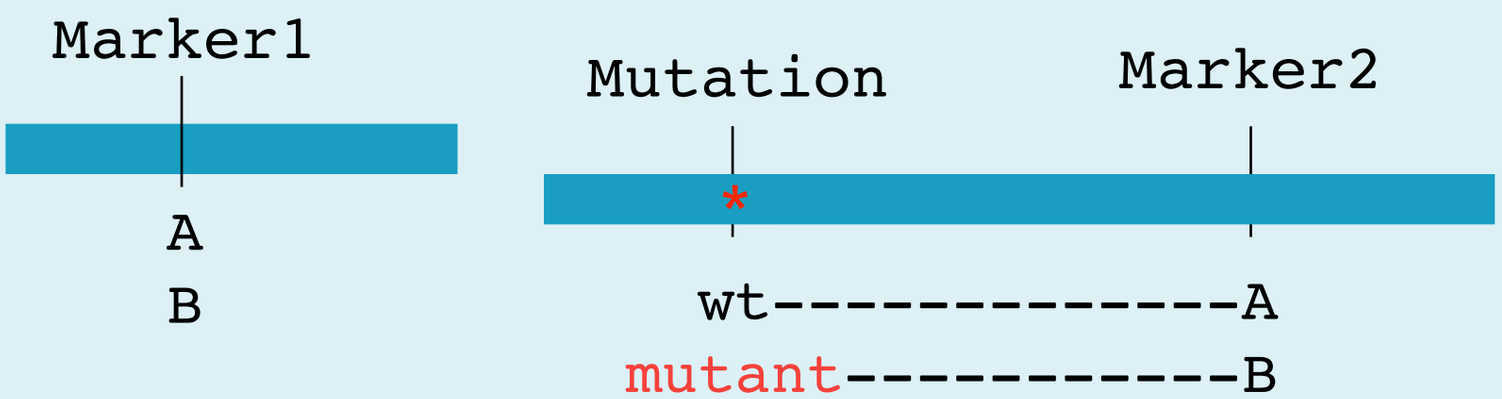
wt => AA
mutant => BB

=> linkage for Markers and Phenotyp



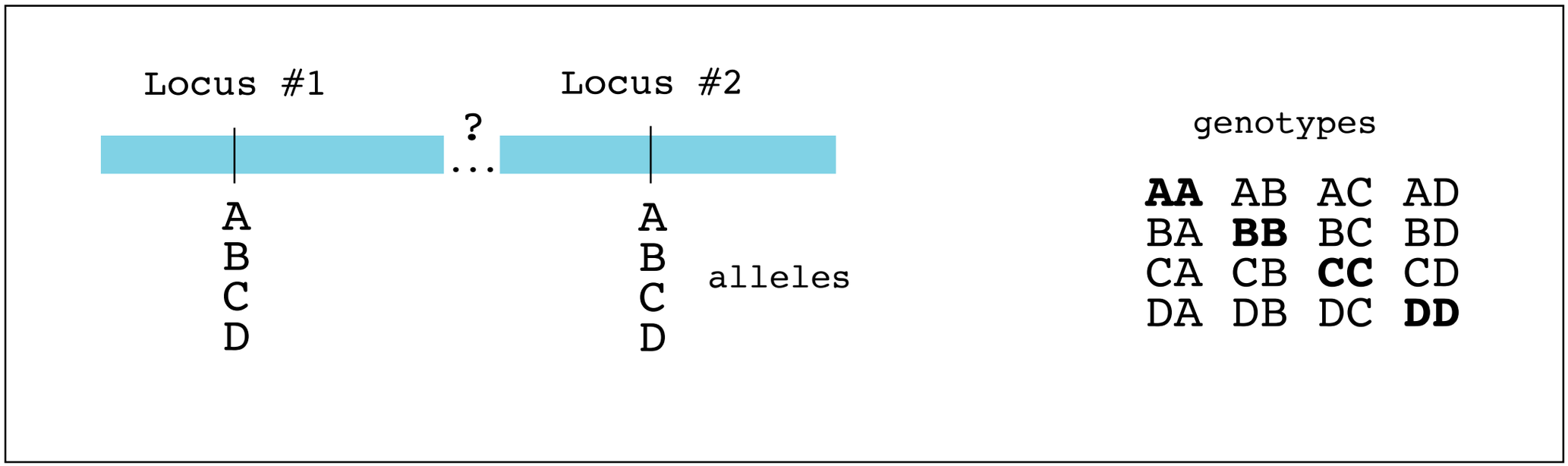
wt => AA,AB,BA,BB
mutant => AA,AB,BA,BB

=> no linkage - independent segregation



wt => A**A**,B**A**
mutant => A**B**,B**B**

=> linkage for Marker2 and Phenotyp



healthy

1	2	3	4	5	6	7	8	9	10
f_{emale}	f	m_{ale}	f	m	m	f	m	m	f
AC	AD	BB	AB	AC	CD	AC	BC	AA	AD

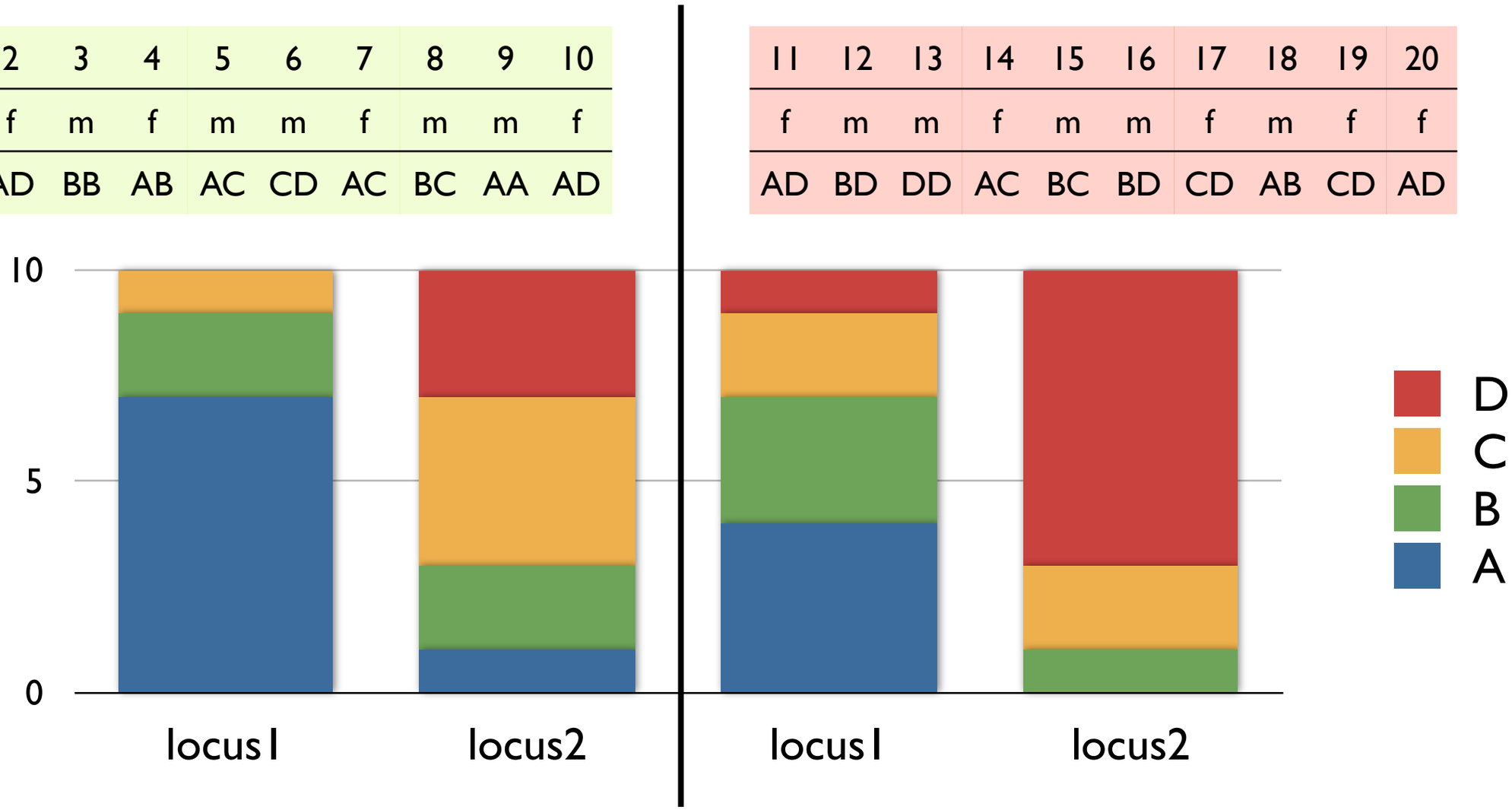
affected

11	12	13	14	15	16	17	18	19	20
f	m	m	f	m	m	f	m	f	f
AD	BD	DD	AC	BC	BD	CD	AB	CD	AD

Example: Association of diseases and markers

1	2	3	4	5	6	7	8	9	10
f	f	m	f	m	m	f	m	m	f
AC	AD	BB	AB	AC	CD	AC	BC	AA	AD

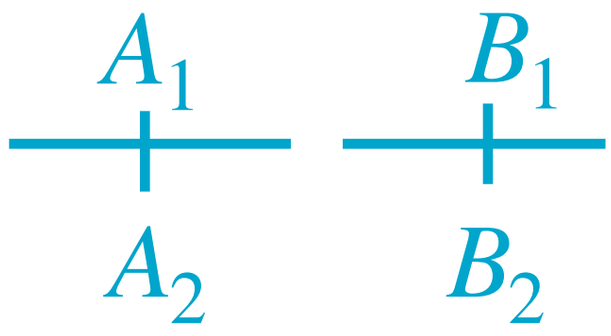
11	12	13	14	15	16	17	18	19	20
f	m	m	f	m	m	f	m	f	f
AD	BD	DD	AC	BC	BD	CD	AB	CD	AD



☛ We can say that a marker allele is associated with a disease if the allele is found more frequently in cases than in the background population, or in a group of unaffected controls.

The parameter **D is the coefficient of linkage disequilibrium** and was first proposed by Lewontin and Kojima (1960).

Linkage equilibrium	Linkage disequilibrium
$D = 0$	$D \neq 0$



$\rightarrow (A_1B_1) = x_{11}$ $\nearrow (A_2B_1) = x_{21}$

$\rightarrow (A_2B_2) = x_{22}$ $\searrow (A_1B_2) = x_{12}$

$$D = \underbrace{X_{11}X_{22}}_{\text{part that shows linkage}} - \underbrace{X_{12}X_{21}}_{\text{part that shows recombination}}$$

Biologically speaking: if D is not equal to 0 then the frequency of A_1B_1_A_2B_2 is not equal the frequency of A_1B_2_A_2B_1, i.e. the two double heterozygotes occur in unequal frequencies.

If the alleles at the two loci are not randomly associated, there is a deviation in the expected frequencies:

Frequency of A_1B_1 or A_2B_2

$x_{11} = p_1q_1$

$x_{22} = p_2q_2$

linkage **increases** the frequency of association

$X_{11} = p_1q_1 + D$

$X_{22} = p_2q_2 + D$

Frequency of A_1B_2 or A_2B_1

$x_{12} = p_1q_2$

$x_{21} = p_2q_1$

linkage **reduces** the frequency of recombination events

$X_{12} = p_1q_2 - D$

$X_{21} = p_2q_1 - D$

“expected” frequency (no linkage disequilibrium)		“observed” frequency (linkage disequilibrium)
$x_{11}=p_1q_1$	➤	$\mathbf{X_{11}=p_1q_1+D}$
$x_{12}=p_1q_2$	➤	$\mathbf{X_{12}=p_1q_2-D}$
$x_{21}=p_2q_1$	➤	$\mathbf{X_{21}=p_2q_1-D}$
$x_{22}=p_2q_2$	➤	$\mathbf{X_{22}=p_2q_2+D}$

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

↓

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

Testing for significance of linkage disequilibrium

$$\chi^2 = \rho^2 n$$

$$\rho = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}}$$

Hill and Robertson (1968)

ρ (roh) has the convenient feature that the χ^2 value for goodness of fit to the hypothesis $D = 0$ can be calculated, where *n* is the number of chromosomes (haplotypes) in the sample. The test has 1 degree of freedom (thus, values larger 4 are significant at $p < 0.05$).

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

$$\rho = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}} \rightarrow \frac{0.0699}{\sqrt{0.543 \cdot 0.457 \cdot 0.308 \cdot 0.692}} = 0.3040$$

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

$$\rho = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}} \rightarrow \frac{0.0699}{\sqrt{0.543 \cdot 0.457 \cdot 0.308 \cdot 0.692}} = 0.3040$$

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

$$\chi^2 = \rho^2 n \rightarrow 0.3040^2 \cdot 2000 = 184.7 \Rightarrow p < 0.01$$

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

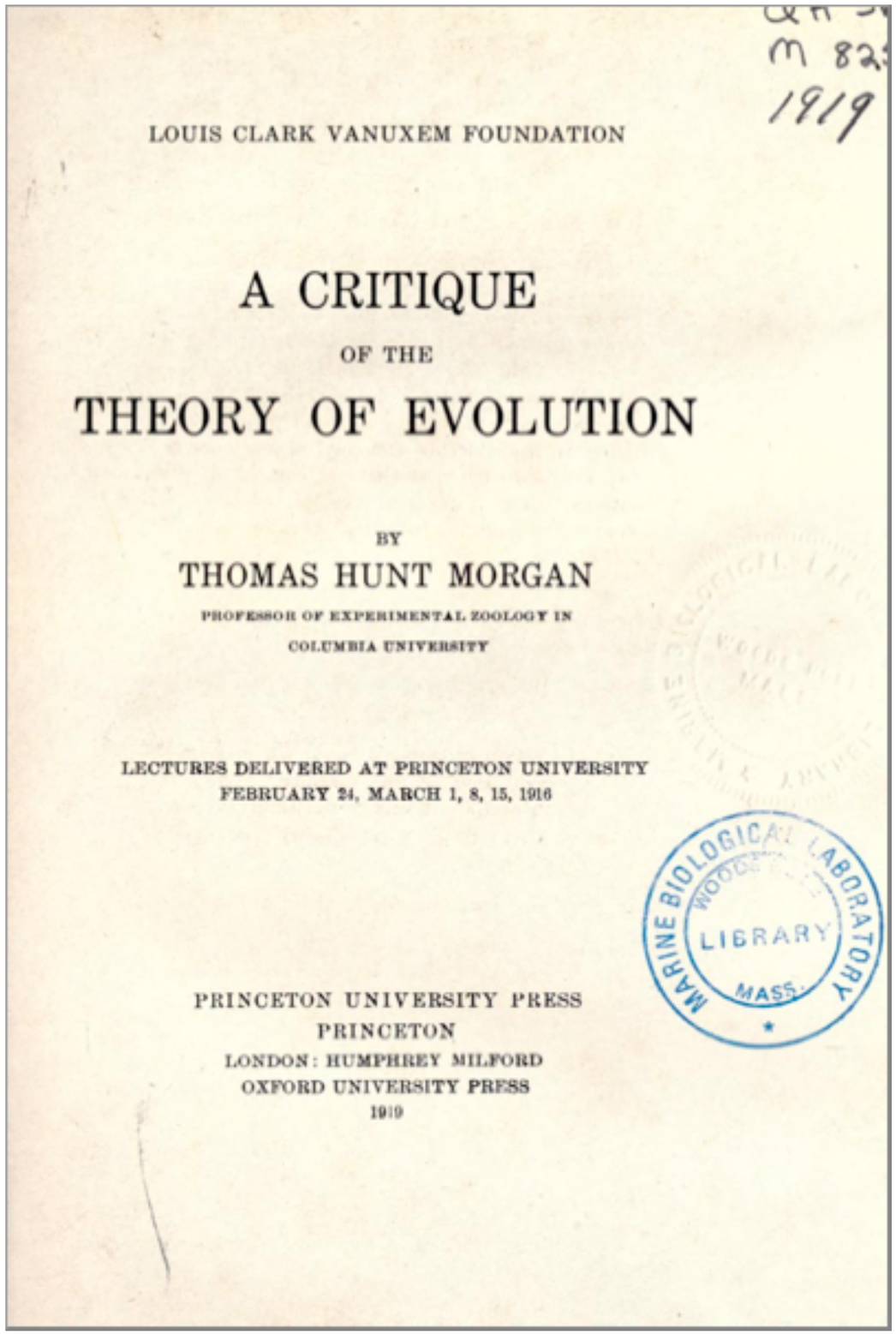
$$\rho = \frac{D}{\sqrt{p_1 q_1 p_2 q_2}} \rightarrow \frac{0.0699}{\sqrt{0.543 \cdot 0.457 \cdot 0.308 \cdot 0.692}} = 0.3040$$

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

$$\chi^2 = \rho^2 n \rightarrow 0.3040^2 \cdot 2000 = 184.7 \Rightarrow p < 0.01$$

⇨ D is not significantly different from zero.



LOUIS CLARK VANUXEM FOUNDATION

A CRITIQUE
OF THE
THEORY OF EVOLUTION

BY

THOMAS HUNT MORGAN
PROFESSOR OF EXPERIMENTAL ZOOLOGY IN
COLUMBIA UNIVERSITY

LECTURES DELIVERED AT PRINCETON UNIVERSITY
FEBRUARY 24, MARCH 1, 8, 15, 1916

PRINCETON UNIVERSITY PRESS
PRINCETON
LONDON: HUMPHREY MILFORD
OXFORD UNIVERSITY PRESS
1919

Thomas Hunt Morgan
(1866-1945)

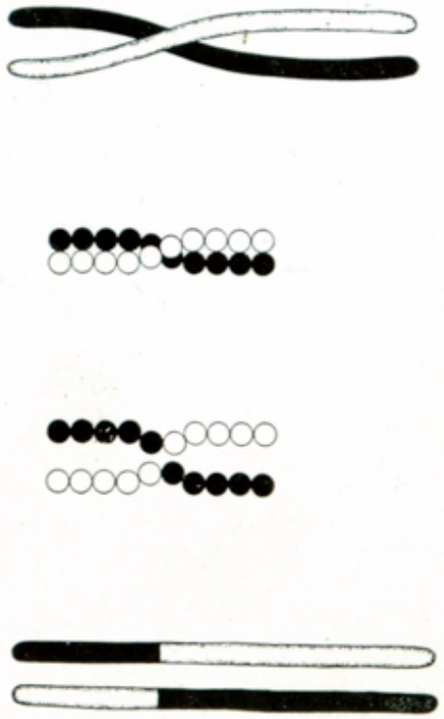


FIG. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Sturtevant, A. H. 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14: 43-59.

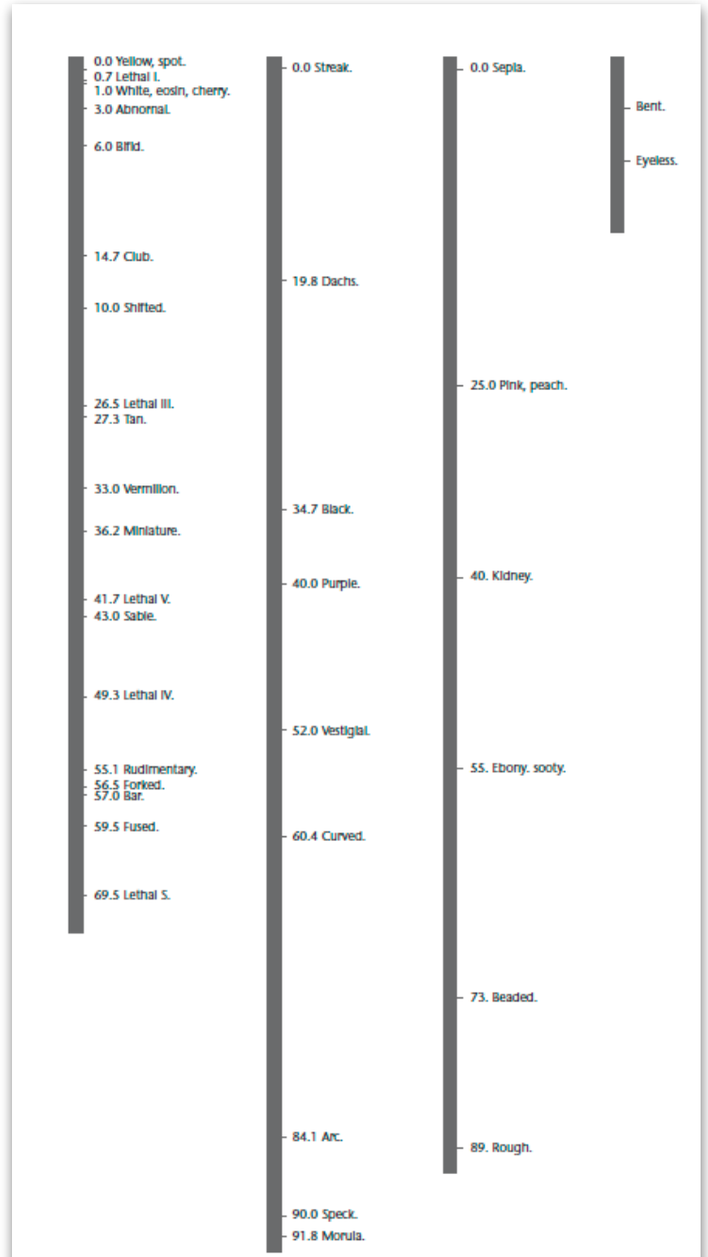
THE LINEAR ARRANGEMENT OF
SIX SEX-LINKED FACTORS IN DROSOPHILA,
AS SHOWN BY THEIR MODE OF ASSOCIATION

A. H. STURTEVANT

A. H. Sturtevant (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology, vol. 14, 43-59.

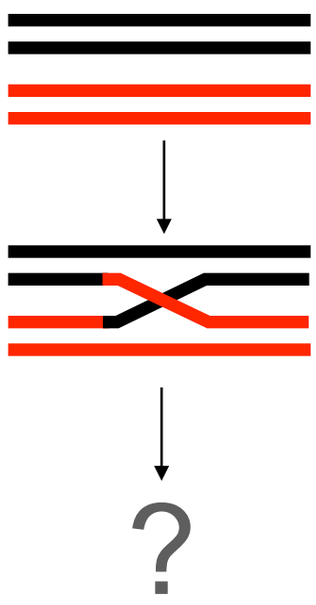
J. B. S. Haldane (1919) The combination of linkage values, and the calculation of distances between linked factors. Journal of Genetics, vol. 8, 299-309.

D. D. Kosambi (1944) The estimation of map distance from recombination values. Annals of Eugenics, vol. 12, 172-175.



A genetic linkage map of the four chromosomes of Drosophila. Reproduced from Morgan TH, Sturtevant AH, Muller HJ and Bridges CB (1915) The mechanism of Mendelian heredity.

There are two types of **homologous recombination**



during meiosis, homologous chromosomes pair and recombine and generate genetic diversity

?

➲ **crossing over** (reciprocal recombination)

> even exchange of homologous sequences between homologous chromosomes

➲ **gene conversion** (nonreciprocal recombination)

> uneven replacement of one sequence by another and loss of one of the variants

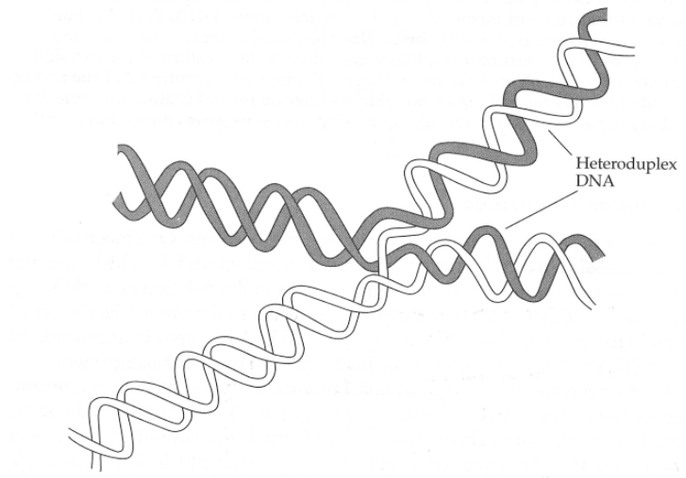


FIGURE 1.14 The Holliday structure. Note the heteroduplex DNA composed of mismatched chromatid strands (gray and white strands).

Robin Holliday (1932–2014) was a British molecular biologist.

Possible outcomes of the resolution of a Holliday junction and the subsequent excision and mismatch repair of heteroduplex DNA.

P

[A]-[B]-[C]-[D]-[E]-[F] × [a]-[b]-[c]-[d]-[e]-[f]
[A]-[B]-[C]-[D]-[E]-[F] [a]-[b]-[c]-[d]-[e]-[f]

F1

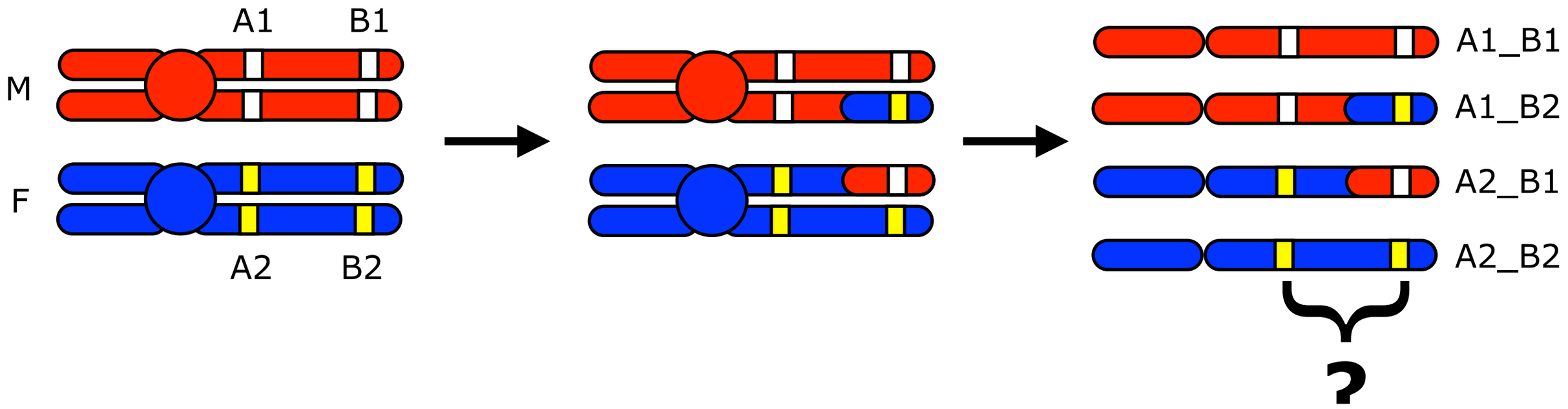
[A]-[B]-[C]-[D]-[E]-[F]
[a]-[b]-[c]-[d]-[e]-[f]

F2

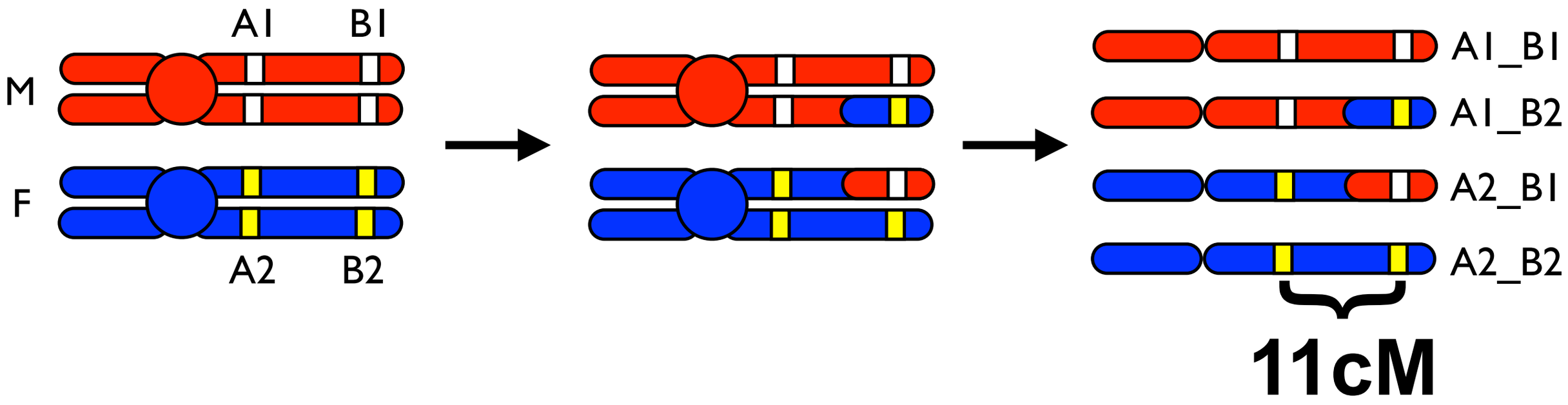
[A]-[B]-[C]-[D]-[E]-[F] × [a]-[b]-[c]-[d]-[e]-[f]
[a]-[b]-[c]-[d]-[e]-[f] [A]-[B]-[C]-[D]-[E]-[F]

[A]-[B]-[C]-[D]-[E]-[F]
[a]-[b]-[c]-[d]-[e]-[f]
[A]-[B]-[C]-d]-[e]-[f]
[A]-[B]-c]-[d]-[e]-[f]
[A]-b]-[C]-[D]-[e]-[F]
[a]-[b]-[c]-[D]-[e]-[f]

...



mother type	A1/B1	165
father type	A2/B2	191
recombinant type	A1/B2	23
recombinant type	A2/B1	21
		400



mother type	A1_B1	165
father type	A2_B2	191
recombinant type	A1_B2	23
recombinant type	A2_B1	21

☛ 44 of the 400 offspring tested are recombinants. The distance between the two loci is 44/400 ⇒ **11cM** (=11 map units).

One **map unit** (m.u.) is the distance between gene pairs for which one product of meiosis out of hundred is recombinant (= a recombination frequency of 0.01 is defined as 1 map unit). Later one map unit was called 1 **centiMorgan** (cM).

The genetic distance in centimorgans (cM) is numerically equal to the recombination frequency expressed as a percentage.

Plasmodium falciparum (22.8Mb) has an average recombination distance of ~15 kb per centimorgan. Meaning: markers separated by 15 kb of DNA (15,000 nucleotides) have a 1% chance of being separated by crossing over in a single generation. (*Plasmodium falciparum* is a unicellular protozoan parasite of humans, and the deadliest species of Plasmodium that causes malaria in humans.)

One centimorgan corresponds to about 1 million base pairs in humans on average.

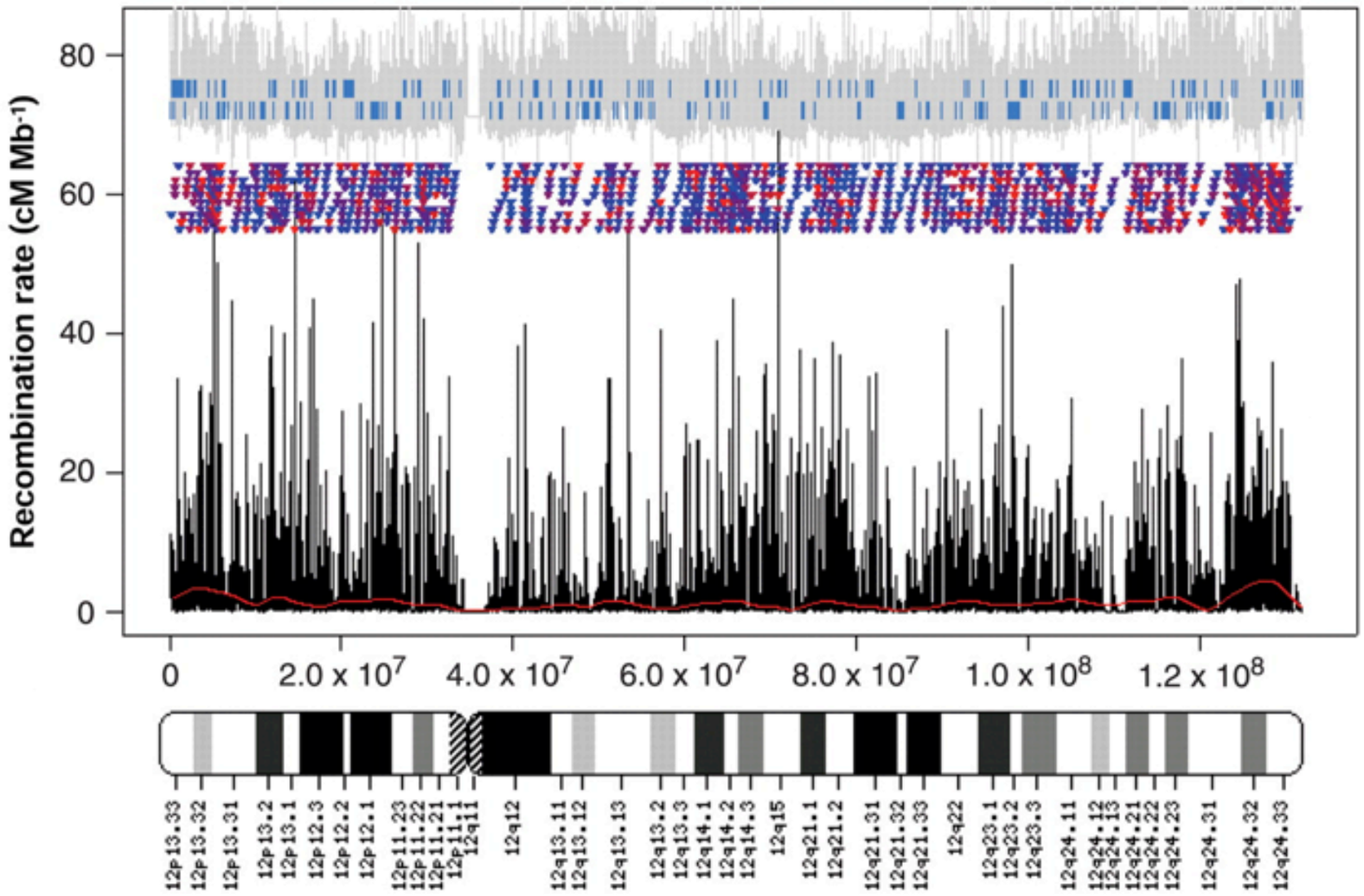


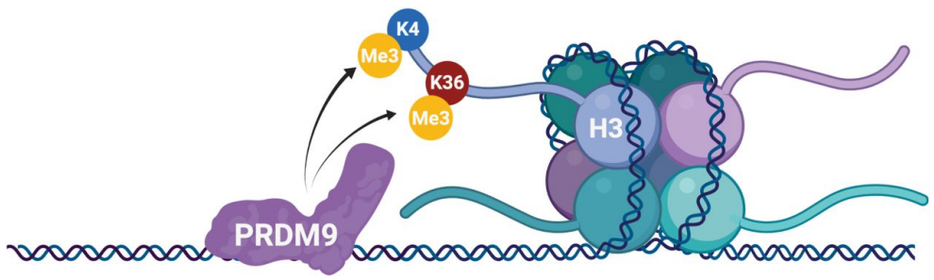
Fig. 1. Recombination rate variation along chromosome 12. Shown are estimated recombination rate (black), locations of statistically significant recombination hotspots [triangles; colors indicate relative amount of recombination from low (blue) to high (red)], and estimated recombination rates from the de-CODE (*6*) genetic map (red curve near bottom). Also shown are the location of ENSEMBL genes on the two strands (blue segments), fluctuations in local GC content (gray lines; averages over 1000-bp windows shown on an arbitrary scale), and an ideogram of chromosome banding.

Myers et al. (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. SCIENCE. 310, 5746: 321-324

PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice

F. Baudat,[1*] J. Buard,[1*] C. Grey,[1*] A. Fledel-Alon,[2] C. Ober,[2] M. Przeworski,[2,3] G. Coop,[4] B. de Massy[1]†

Meiotic recombination events cluster into narrow segments of the genome, defined as hotspots. Here, we demonstrate that a major player for hotspot specification is the *Prdm9* gene. First, two mouse strains that differ in hotspot usage are polymorphic for the zinc finger DNA binding array of PRDM9. Second, the human consensus PRDM9 allele is predicted to recognize the 13-mer motif enriched at human hotspots; this DNA binding specificity is verified by in vitro studies. Third, allelic variants of PRDM9 zinc fingers are significantly associated with variability in genome-wide hotspot usage among humans. Our results provide a molecular basis for the distribution of meiotic recombination in mammals, in which the binding of PRDM9 to specific DNA sequences targets the initiation of recombination at specific locations in the genome.



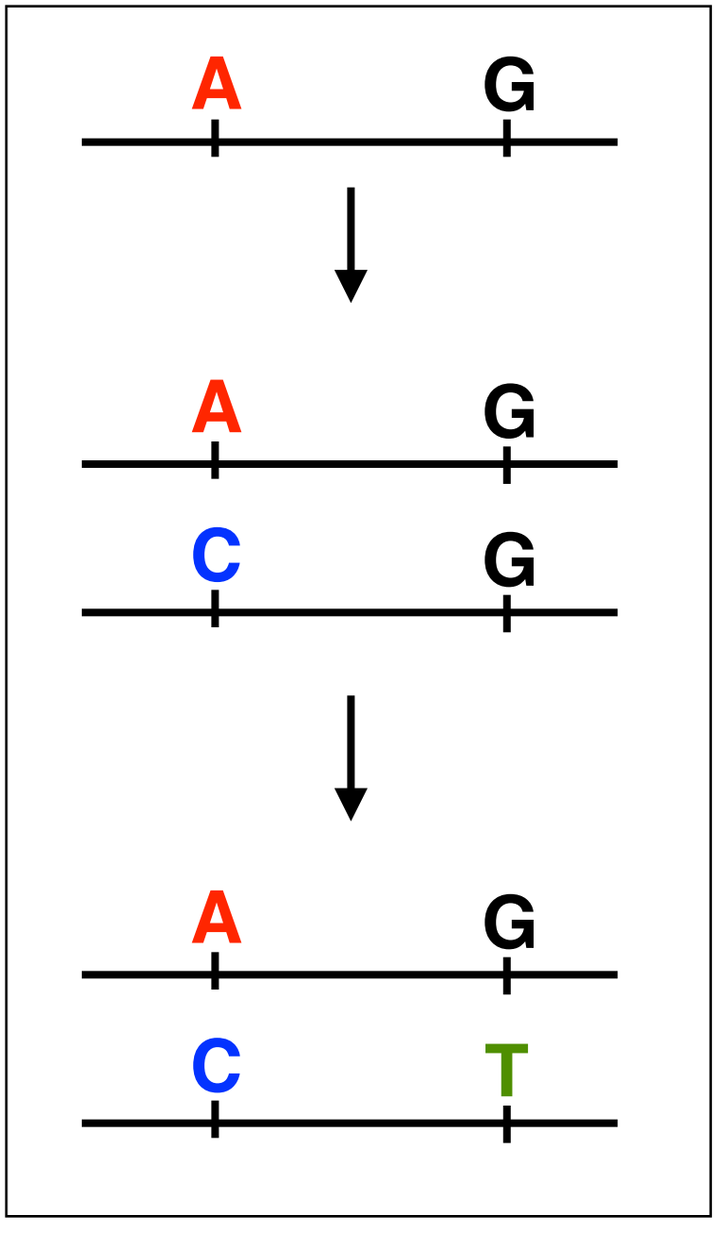
During meiosis, homologous chromosomes pair and recombine, enabling balanced segregation and generating genetic diversity. In many vertebrates, double-strand breaks (DSBs) initiate recombination within hotspots where PRDM9 binds. Source: Wells et al. (2020) ZCWPW1 is recruited to recombination hotspots by PRDM9 and is essential for meiotic double strand break repair. https://doi.org/10.7554/eLife.53392

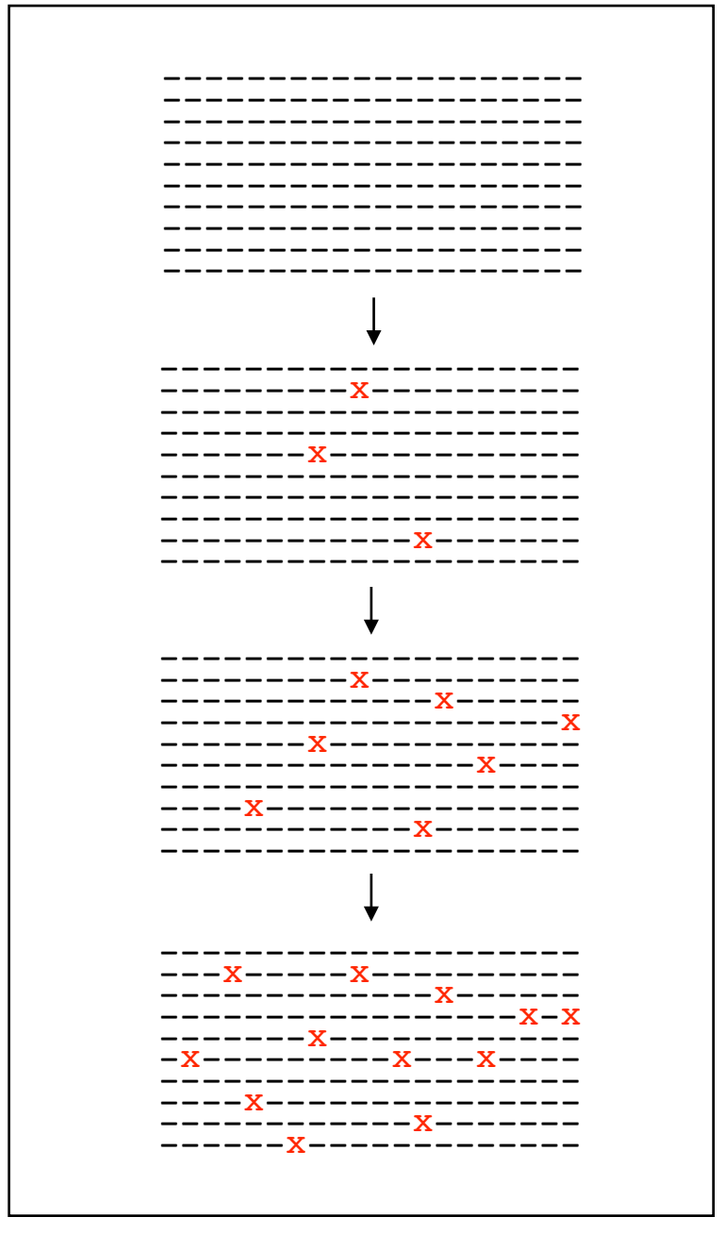
*PRDM9 is a zinc finger protein that binds DNA at specific locations in the genome where it trimethylates histone H3 at lysines 4 and 36 at surrounding nucleosomes

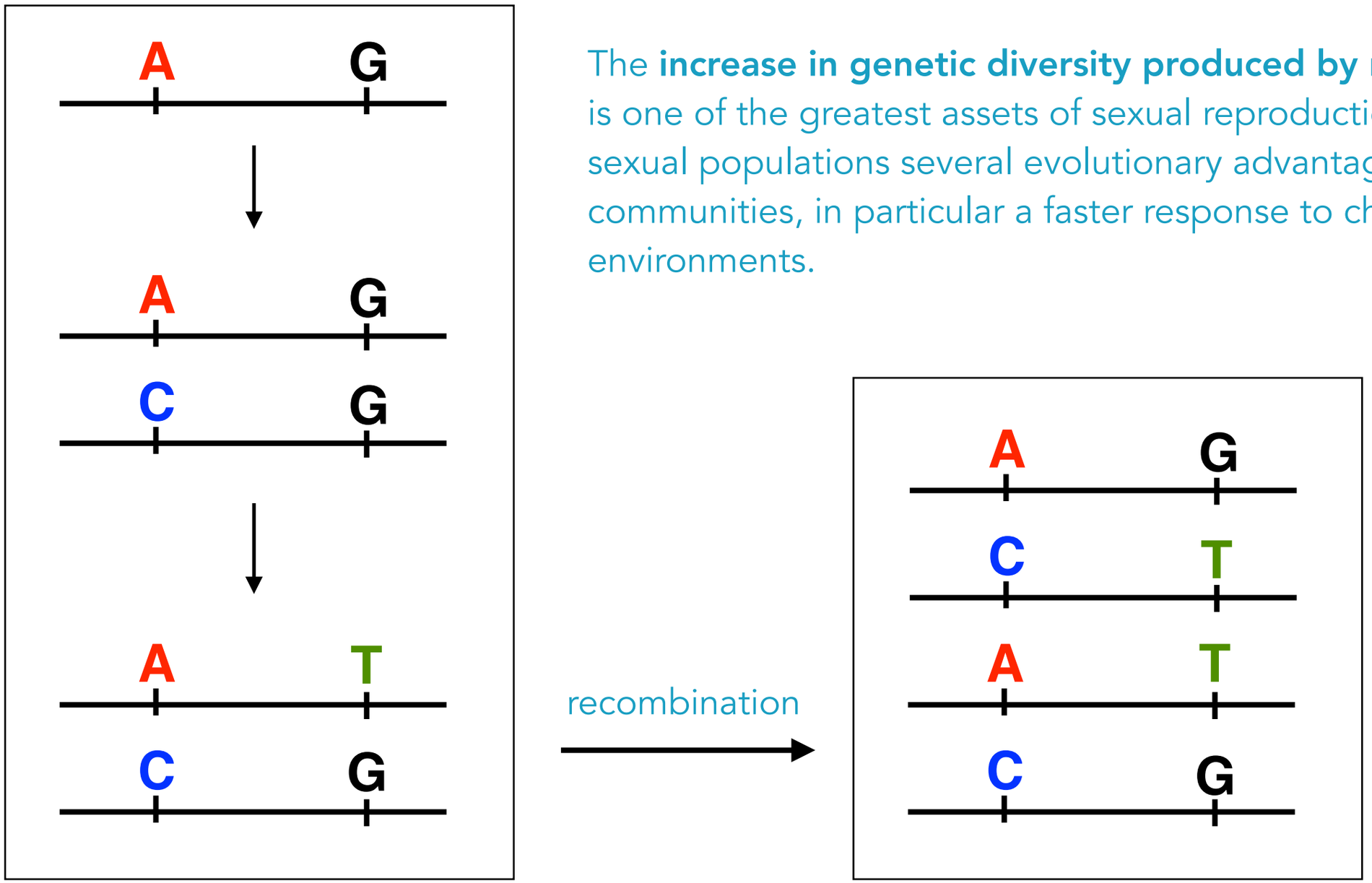
François Jacob [1920-2013]

"It is natural selection that gives direction to changes, orients chance, and slowly, progressively produces more complex structures, new organs, and new species. Novelties come from previously unseen association of old material. **To create is to recombine.**"

Jacob (1977) Evolution and tinkering. Science (196):1161-1166


A
G
A
G
C
G
A
G
C
T

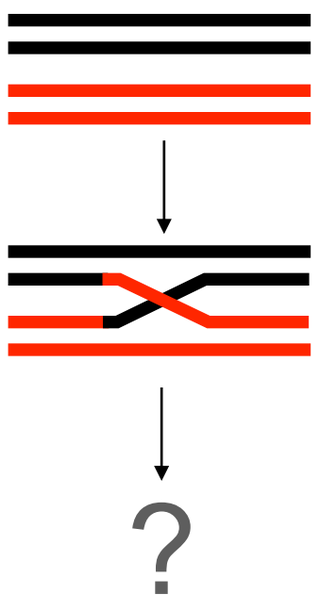




The **increase in genetic diversity produced by recombination** is one of the greatest assets of sexual reproduction and gives sexual populations several evolutionary advantages over asexual communities, in particular a faster response to changing environments.


MORE
THINGS
CONSIDERED
1

There are two types of **homologous recombination**



➲ **crossing over** (reciprocal recombination)

> even exchange of homologous sequences between homologous chromosomes

➲ **gene conversion** (nonreciprocal recombination)

> uneven replacement of one sequence by another and loss of one of the variants

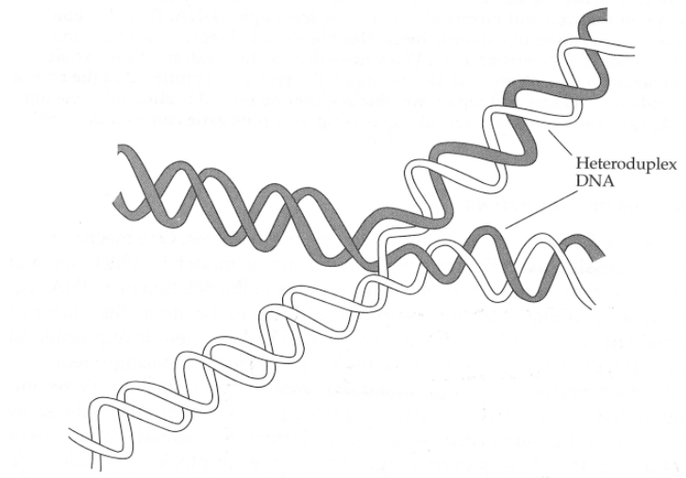
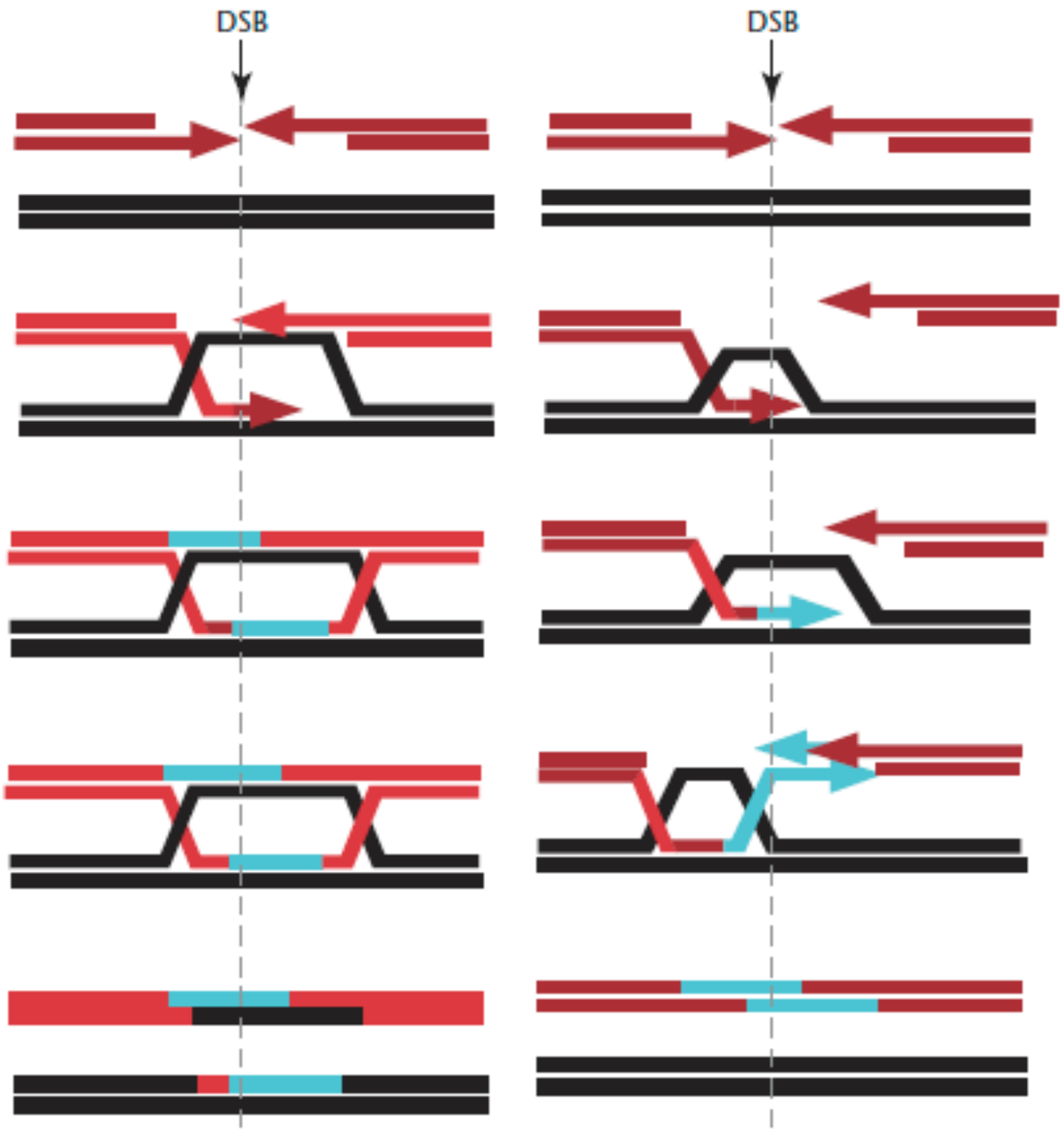


FIGURE 1.14 The Holliday structure. Note the heteroduplex DNA composed of mismatched chromatid strands (gray and white strands).

Robin Holliday (1932–2014) was a British molecular biologist.

Possible outcomes of the resolution of a Holliday junction and the subsequent excision and mismatch repair of heteroduplex DNA.



Two major models of genetic recombination (left) Szostak (1983) DSBR model (right) Allers and Lichten (2001) SDSA model. Modified from Haber et al., 2004. Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model.

Linkage equilibrium

$$D = 0$$

Linkage disequilibrium

$$D \neq 0$$

Linkage Disequilibrium

We consider two loci (A and B), each segregating for two alleles (A_1 / A_2 and B_1 / B_2)

```
--A1---B1--
--A2---B2--   = A1B1_A2B2
```

Allele frequencies: $A_1{:}p_1$ | $A_2{:}p_2$ | $B_1{:}q_1$ | $B_2{:}q_2$ with $p_1+p_2=1$, $q_1+q_2=1$

Linkage Disequilibrium

We consider two loci (A and B), each segregating for two alleles (A_1 / A_2 and B_1 / B_2)

```
--A1---B1--
--A2---B2--   = A1B1_A2B2
```

Allele frequencies: A_1:p_1 | A_2:p_2 | B_1:q_1 | B_2:q_2 with $p_1+p_2=1$, $q_1+q_2=1$

gametes (haplotypes)	frequency
A_1B_1	$x_{11}=p_1q_1$
A_1B_2	$x_{12}=p_1q_2$
A_2B_1	$x_{21}=p_2q_1$
A_2B_2	$x_{22}=p_2q_2$

$\sum x_{ij} = 1.0$

allele	frequency
A_1	$p_1=x_{11}+x_{12}$
A_2	$p_2=x_{21}+x_{22}$
B_1	$q_1=x_{11}+x_{21}$
B_2	$q_2=x_{12}+x_{22}$

↘ $x_{11}+x_{12} = p_1q_1+p_1q_2 = p_1(q_1+q_2) = p_1$

→ In the presented situation, there is no linkage disequilibrium and gamete frequencies can be accurately followed using allele frequencies.

If the alleles at the two loci are not randomly associated, there is a deviation in the expected frequencies:

Frequency of A_1B_1:

$x_{11} = p_1q_1$

linkage **increases** the frequency of association

$X_{11} = p_1q_1 + D$

Frequency of A_1B_2:

$x_{12} = p_1q_2$

linkage **reduces** the frequency of recombination events

$X_{12} = p_1q_2 - D$

“expected” frequency (no linkage disequilibrium)		“observed” frequency (linkage disequilibrium)
$x_{11}=p_1q_1$	➤	$\mathbf{X_{11}=p_1q_1+D}$
$x_{12}=p_1q_2$	➤	$\mathbf{X_{12}=p_1q_2-D}$
$x_{21}=p_2q_1$	➤	$\mathbf{X_{21}=p_2q_1-D}$
$x_{22}=p_2q_2$	➤	$\mathbf{X_{22}=p_2q_2+D}$

The parameter **D is the coefficient of linkage disequilibrium** and was first proposed by Lewontin and Kojima (1960).

$$D = \underbrace{X_{11}X_{22}}_{\text{part that shows linkage}} - \underbrace{X_{12}X_{21}}_{\text{part that shows recombination}}$$

Biologically speaking: if D is not equal to 0 then the frequency of $A_1B_1\_A_2B_2$ is not equal the frequency of $A_1B_2\_A_2B_1$, i.e. the two double heterozygotes occur in unequal frequencies.

After the discovery of the first blood group, ABO, in 1900, Landsteiner and his colleagues continued to experiment with blood to identify other blood groups.

MNS was the second blood group, discovered in 1927, after immunizing rabbits with human RBCs. The M and N antigens were identified first, but it was another 20 years before the S and s antigens were named. Now, more than 40 antigens are known in this blood group, but the M, N, S, and s antigens remain the most common.

Two genes encode the glycophorins that carry the antigens of the MNS blood group: GYPA and GYPB. Both are on the long arm of chromosome 4.

GYPA has two codominant alleles, M and N, which result from three SNPs (59C→T, 71G→A, 72G→T), and the corresponding M and N antigens differ by two amino acids (S1L, G5E). The codominant alleles of GYPB, C and c, result from one SNP (143C→T), and the corresponding S and s antigens differ by a single amino acid (T29M).

Locus GYPA	Locus GYPB
M	S
N	s

$X_{11} \rightarrow$ MS $X_{12} \rightarrow$ Ms

$X_{22} \rightarrow$ Ns $X_{21} \rightarrow$ NS

Locus GYPA	Locus GYPB
M	S
N	s

Observed Genotypes

```
n(MS) = 474
n(Ms) = 611
n(NS) = 142
n(Ns) = 773
```

$X_{11} \rightarrow$ MS $X_{12} \rightarrow$ Ms

$X_{22} \rightarrow$ Ns $X_{21} \rightarrow$ NS

Locus GYPA	Locus GYPB
M	S
N	s

Observed Genotypes

n(MS) = 474
n(Ms) = 611
n(NS) = 142
n(Ns) = 773

$X_{11} \rightarrow$ MS $X_{12} \rightarrow$ Ms

$X_{22} \rightarrow$ Ns $X_{21} \rightarrow$ NS

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

Expected Haplotype-Frequency:
x11=p1q1=0.543*0.308=0.1672
x12=p1q2=0.543*0.692=0.3758
x21=p2q1=0.457*0.308=0.1408
x22=p2q2=0.457*0.692=0.3162 (=1.0)

Expected:
p1q1*2000=334.4
p1q2*2000=751.6
p2q1*2000=281.6
p2q2*2000=632.2

Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

To test for significance of linkage disequilibrium is the calculation of χ^2 from the expected numbers of the four gametic types, which are p_1q_1, p_1q_2, p_2q_1 and p_2q_2.

A problem of interpreting values of D is that its maximum depends on the allele-frequencies. Therefore it is helpful to standardize D, for example as the percentage of its maximum value (if D is positive) or its minimum value (if D is negative).

D_{min} = the larger of $-p1q1$ and $-p2q2$

D_{max} = the smaller of $p1q2$ and $p2q1$

$$D^* = D' = \frac{D}{D_{min} \text{ or } D_{max}}$$

Lewontin (1964)

Another standardized measure of linkage disequilibrium is ρ (rho), which can range from -1 to +1.

$$\rho = D / \text{sqrt}(p_1q_1\, p_2q_2) \quad \rightarrow r^2 = D^2 / [p_1q_1\, p_2q_2]$$

Hill and Robertson (1968)

Testing for significance of linkage disequilibrium

ρ (roh) has the convenient feature that the χ^2 value for goodness of fit to the hypothesis D = 0 can be calculated, where *n* is the number of chromosomes (haplotypes) in the sample. The test has 1 degree of freedom (thus, values larger 4 are significant at $p<0.05$).

$$\chi^2 = \rho^2 n$$

$$\rho = D / \text{sqrt}(p_1 q_1 \, p_2 q_2)$$

Hill and Robertson (1968)

Observed:
MS 474
Ms 611
NS 142
Ns 773 (n=2000)

Allele-Frequency:
M = A1 = p1=(474+611)/2000=0.543 } sum: 1.0
N = A2 = p2=(142+773)/2000=0.457
S = B1 = q1=(474+142)/2000=0.308 } sum: 1.0
s = B2 = q2=(611+773)/2000=0.692

Observed Frequency:
X11=474/2000=0.237
X12=611/2000=0.305
X21=142/2000=0.071
X22=773/2000=0.387 (=1.0)

Expected Haplotype-Frequency:
x11=p1q1=0.543*0.308=0.1672
x12=p1q2=0.543*0.692=0.3758
x21=p2q1=0.457*0.308=0.1408
x22=p2q2=0.457*0.692=0.3162 (=1.0)

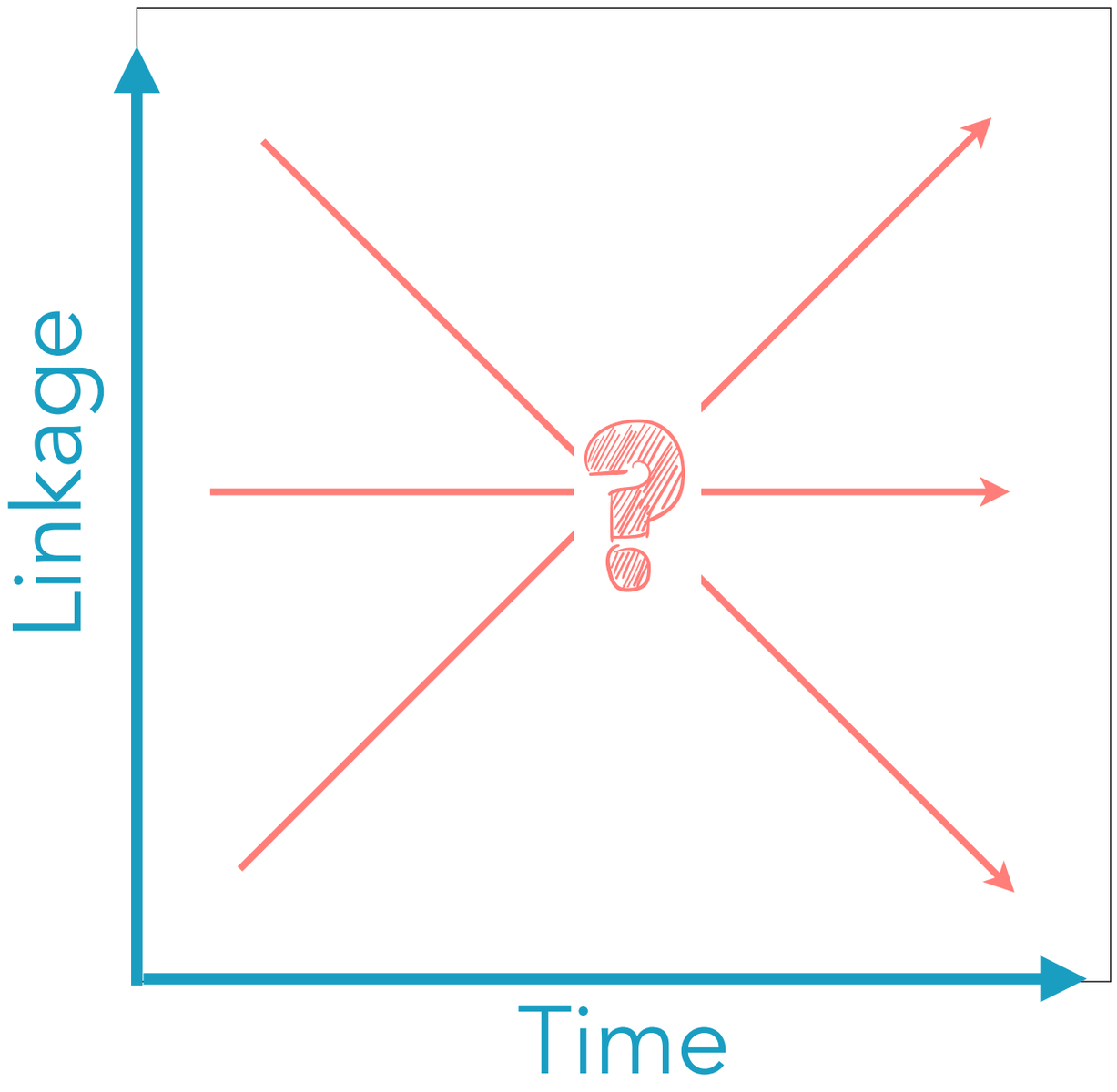
Expected:
p1q1*2000=334.4
p1q2*2000=751.6
p2q1*2000=281.6
p2q2*2000=632.2

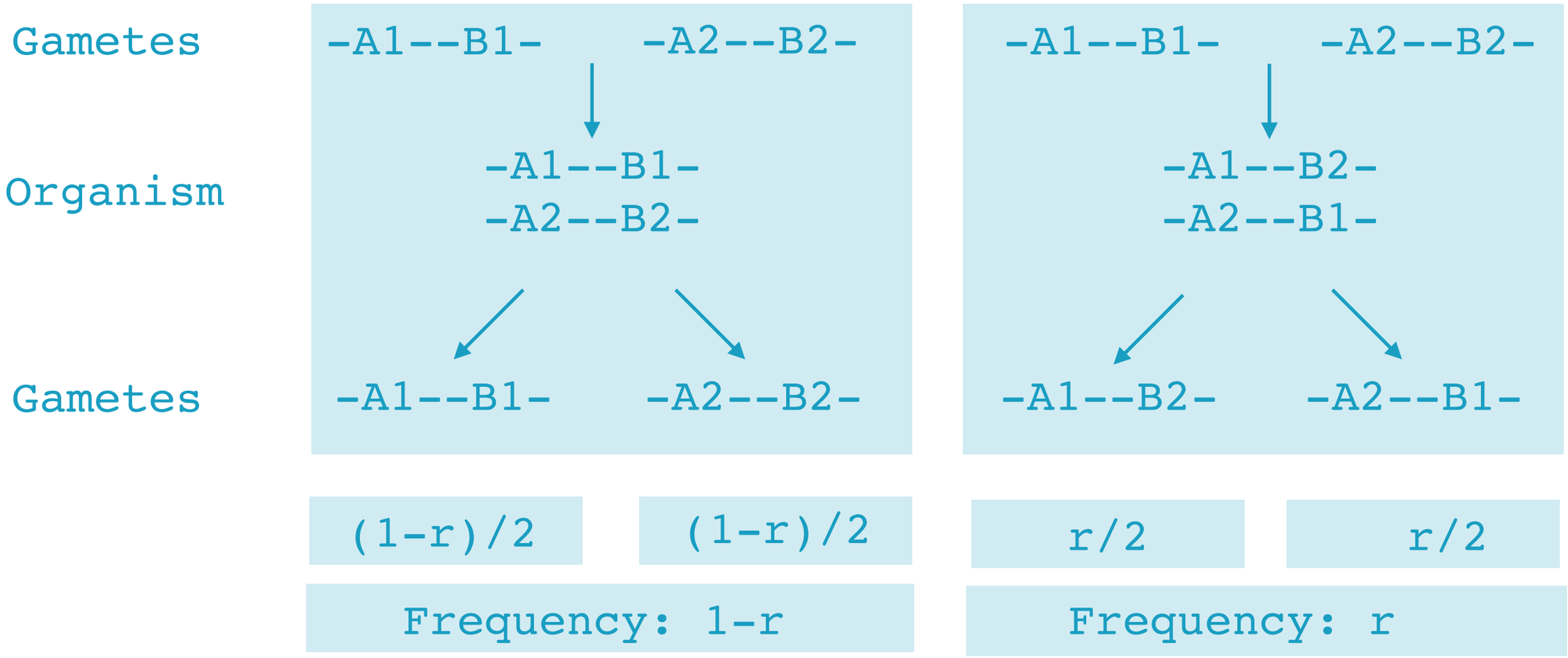
Coefficient of Linkage Disequilibrium:

D = X11*X22 - X12*X21 = 0.0699

$\rho = D/\sqrt{(p_1 q_1 p_2 q_2)} = 0.0699/\sqrt{(0.543*0.457*0.308*0.692)} = \mathbf{0.3040}$

$\chi^2 = \rho^2 n = (0.3040)^2 * 2000 = \mathbf{184.7}$ (df=1) $p < 0.001$ (H_0: D = 0)


Linkage
?
Time



The **frequency of recombination** of 2 loci is defined as r (sometimes c). r is 0.5 if the 2 loci are independent from each other, i.e. if they are on different chromosomes.

$$r:[0 - 0.5]$$

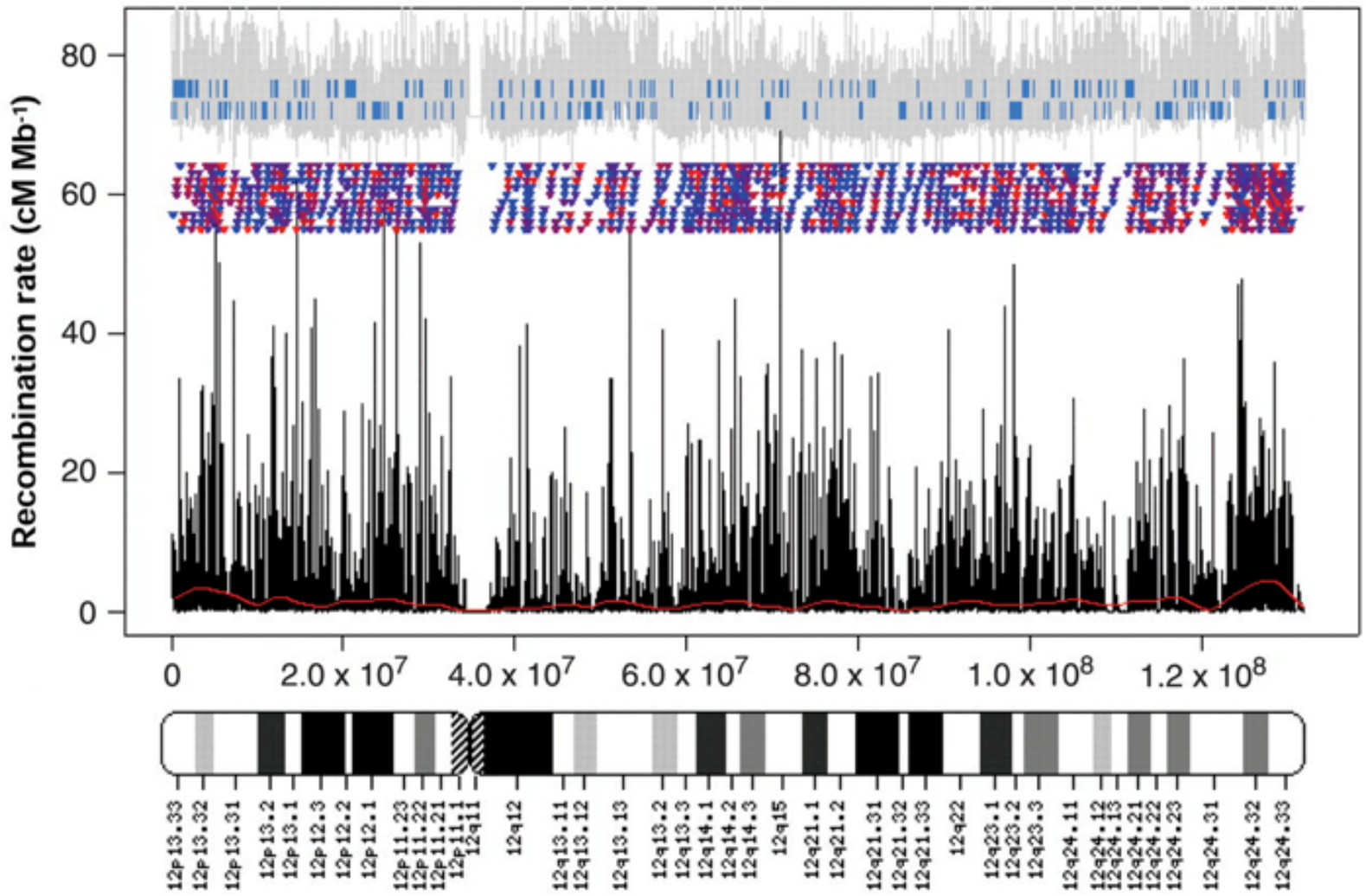


Fig. 1. Recombination rate variation along chromosome 12. Shown are estimated recombination rate (black), locations of statistically significant recombination hotspots [triangles; colors indicate relative amount of recombination from low (blue) to high (red)], and estimated recombination rates from the de-CODE (*6*) genetic map (red curve near bottom). Also shown are the location of ENSEMBL genes on the two strands (blue segments), fluctuations in local GC content (gray lines; averages over 1000-bp windows shown on an arbitrary scale), and an ideogram of chromosome banding.

Myers et al. (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. SCIENCE. 310, 5746: 321-324

Linkage Disequilibrium

Linkage decays each generation at a rate determined by the degree of recombination. The frequencies of the four gametes after one generation are:

$$X'_{11} = X_{11} - rD_0 \qquad X'_{12} = X_{12} + rD_0$$
$$X'_{22} = X_{22} - rD_0 \qquad X'_{21} = X_{21} + rD_0$$

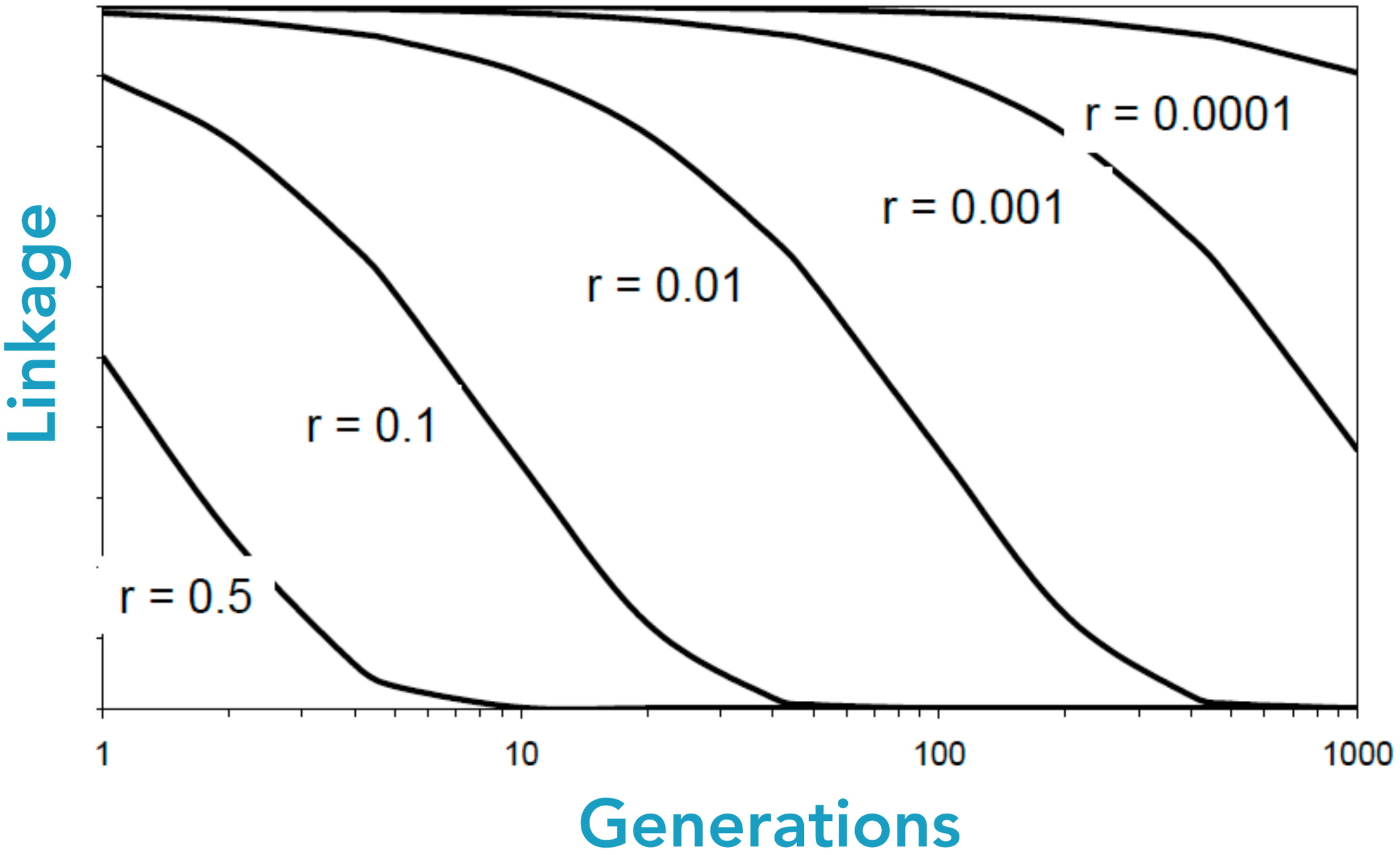
where D_0 is the extent of linkage disequilibrium present in the preceding generation.

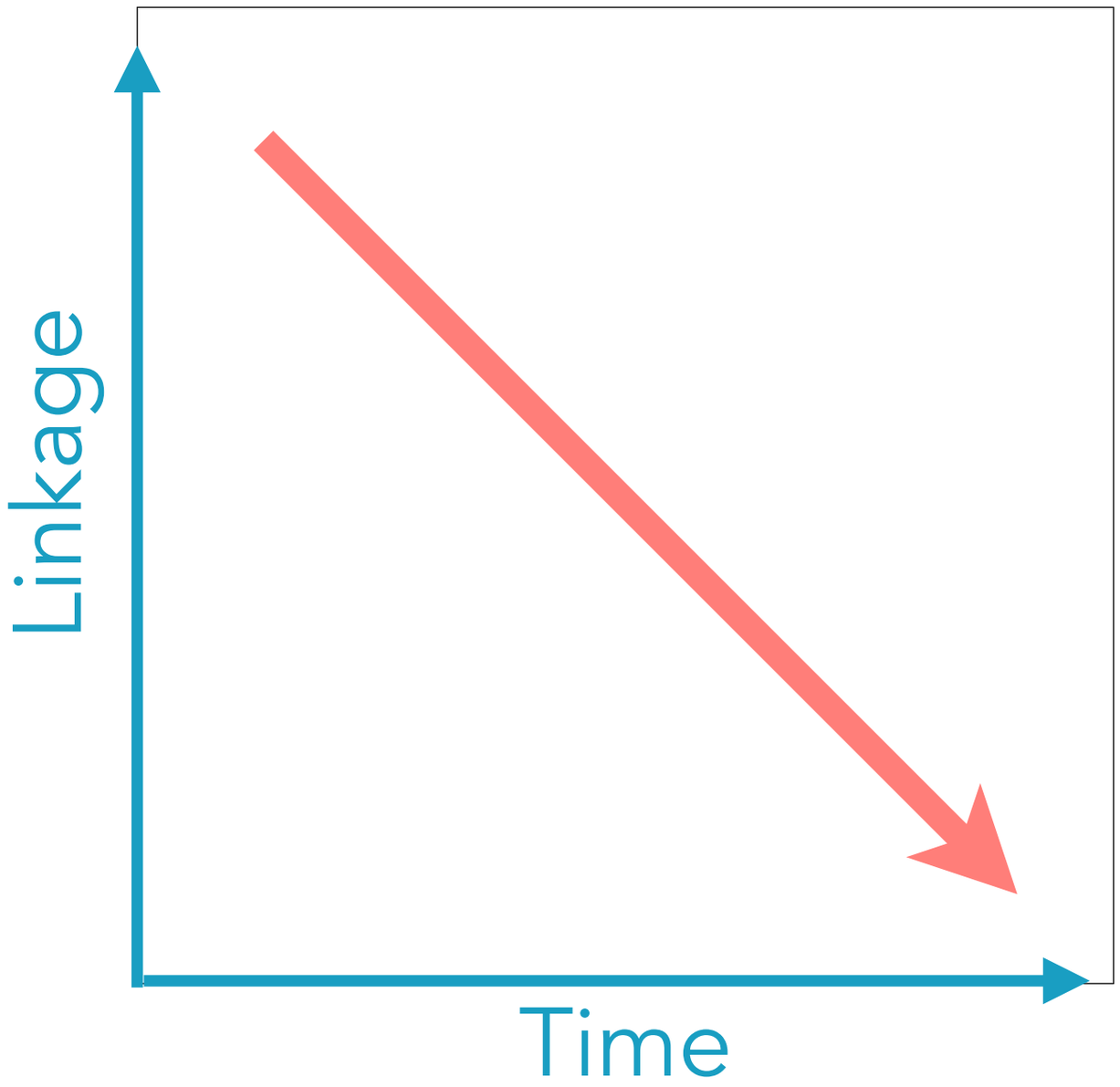
$$D_0 = X_{11}X_{22} - X_{12}X_{21}$$
$$D_1 = X'_{11}X'_{22} - X'_{12}X'_{21}$$
$$= (X_{11} - rD_0)(X_{22} - rD_0) - (X_{12} + rD_0)(X_{21} + rD_0) = (1-r)D_0$$
$$D_2 = (1-r)D_1 = (1-r)^2 D_0$$
$$D_3 = (1-r)D_2 = (1-r)^3 D_0$$
$$\ldots$$

$$\mathbf{D_t = (1-r)D_t = (1-r)^t D_0} \quad \text{(approximated as: } D_t = e^{-ct} D_0\text{)}$$

t: time in generations

frequency of recombination: r [0-0.5]




Linkage
Time

There are various processes that can produce linkage disequilibrium in a population:

- epistatic natural selection (Lewontin 1974, e.g. MHC)
- mutation
- random drift (Ohta and Kimura 1969)
- genetic hitchhiking
- gene flow
- assortative mating (e.g. A_1 mates predominantly with B_1)
- linkage disequilibrium was already in the founding population
- large inversion in chromosome (e.g. Drosophila)
- selfing organisms (inbreeding reduces heterozygosity)

Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait

Ana L. Caicedo[†], John R. Stinchcombe[‡], Kenneth M. Olsen[†], Johanna Schmitt[‡], and Michael D. Purugganan[†§]

15670–15675 | PNAS | November 2, 2004 | vol. 101 | no. 44

Epistatic gene interactions are believed to be a major factor in the genetic architecture of evolutionary diversification. In *Arabidopsis thaliana*, the FRI and FLC genes mechanistically interact to control flowering time, and here we show that this epistatic interaction also contributes to a latitudinal cline in this life history trait within the species. Two major FLC haplogroups (FLC(A) and FLC(B)) are associated with flowering time variation in *A. thaliana* in field conditions, but only in the presence of putatively functional FRI alleles. Significant differences in latitudinal distribution of FLC haplogroups in Eurasia and North Africa also depend on the FRI genotype. **There is significant linkage disequilibrium between FRI and FLC despite their location on separate chromosomes.** Although no nonsynonymous polymorphisms differentiate FLC(A) and FLC(B), vernalization induces the expression of an alternatively spliced FLC transcript that encodes a variant protein with a radical amino acid replacement associated with the two FLC haplogroups. Numerous polymorphisms differentiating the FLC haplogroups also occur in intronic regions implicated in the regulation of FLC expression. The features of the regulatory gene interaction between FRI and FLC in contributing to the latitudinal cline in *A. thaliana* flowering time are consistent with the maintenance of this interaction by epistatic selection. These results suggest that developmental genetic pathways and networks provide the molecular basis for epistasis, contributing to ecologically important phenotypic variation in natural populations and to the process of evolutionary diversification.

The human major histocompatibility complex (MHC) is a cluster of linked loci involved in immune response that is found in all vertebrates. These loci have a major role in fighting pathogens. There are consistent non-random associates of alleles at different loci (linkage disequilibrium) in the MHC.

With the advent of genome wide linkage maps of polymorphic DNA markers it has been possible to examine the influence of genes on the susceptibility to disease in a comprehensive way. Studies of this type have shown that within the human population there are multiple genetic loci which are involved in susceptibility to common autoimmune conditions such as diabetes and rheumatoid arthritis. It is likely that most autoimmune conditions have some genetic component. It is important to distinguish these kind of genetic susceptibilities from traditional inherited disease. No single predisposing allele has to be present for the disease to occur, rather the presence of combinations of susceptibility alleles significantly increases the probability of that individual developing the specific disease.

Haplotype frequencies for HLA-A and HLA-B loci. The color indicates a deficiency (red), or excess (blue) of the haplotype.

		HLA-A alleles			
		A1	A2	A3	overall HLA-B allele frequencies
HLA-B alleles	B7	**0.0074**	0.026	0.0477	0.1143
	B8	0.0672	0.011	0.0019	0.0971
	B35	0.0029	0.0178	0.0257	0.1052
	B44	0.0089	0.0503	0.0068	0.1242
overall HLA-A allele frequencies		0.1439	0.2855	0.1335	

The frequency of A1-B7 haplotype is 0.0074. At linkage equilibrium it would be expected to have a frequency of 0.1439*0.1143=0.0164, so that it shows a deficiency of -0.0090. This deficiency is the likage disequilibrium associated with this halotype and is 55% of the maximum value that D could have for alleles with these frequencies,

The photo shows unpalatable swallowtail model species (left) and palatable mimetic forms of female *Papilio memnon* (right). At bottom is the *Papilio memnon* male. This polymorphic, female-limited Batesian mimicry was first described by Alfred Russel Wallace (1865).

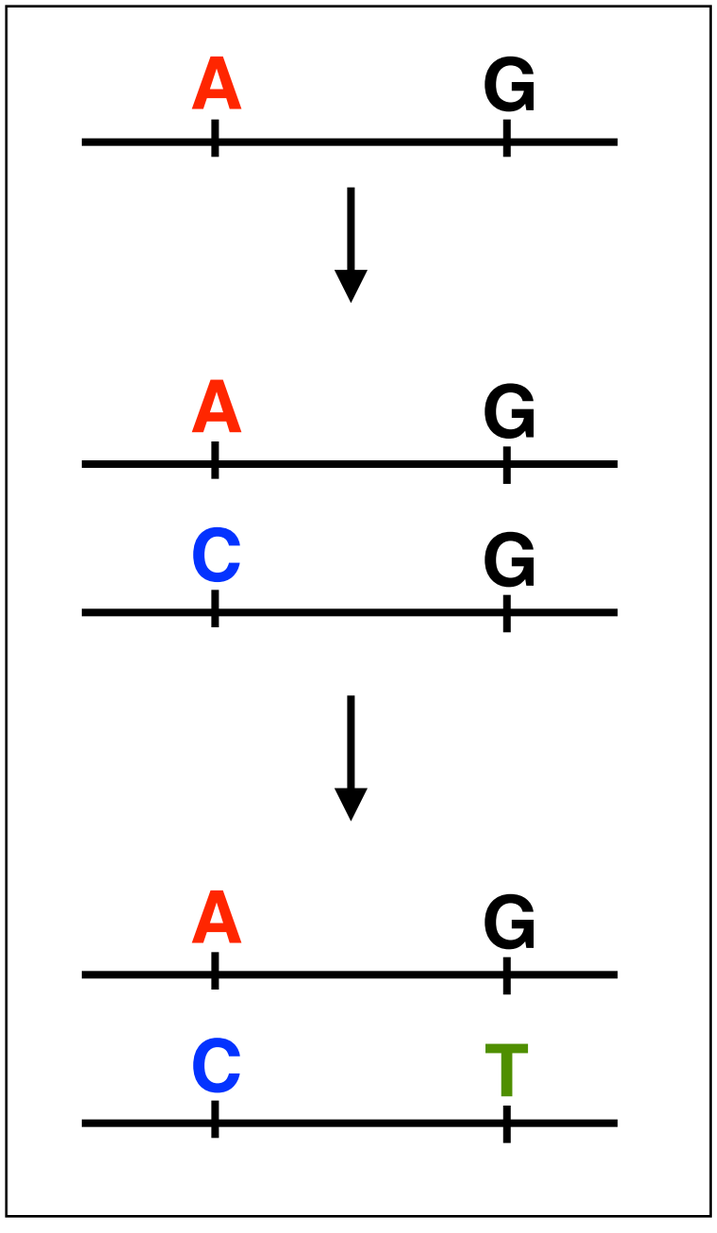
The butterfly *Papilio memnon* provides a good example of linkage disequilibrium. The swallowtail butterfly is living in the Malay Archipelago and Indonesia. The females come in a variety of morphs. The most common morphs mimic natural models which are unpalatable to birds. Each of the different morphs is thought to be controlled by a multi-locus genotype. The loci in question are so tightly linked that recombinants rarely arise. A set of genes so tightly linked that they behave like a single locus has been termed a supergene. However, a sufficiently large number of crosses would be able to break the supergenes into different combinations. In the *P. memnon* there appear to be at least five loci in the mimicry supergene: T, W, F, E, and B. They control respectively presence or absence of tail, hindwing patterns, forewing pattern, epaulette color, and body color. The rare anura morph, for instance, is thought to be a recombinant between the T locus and the other four.

François Jacob [1920-2013]

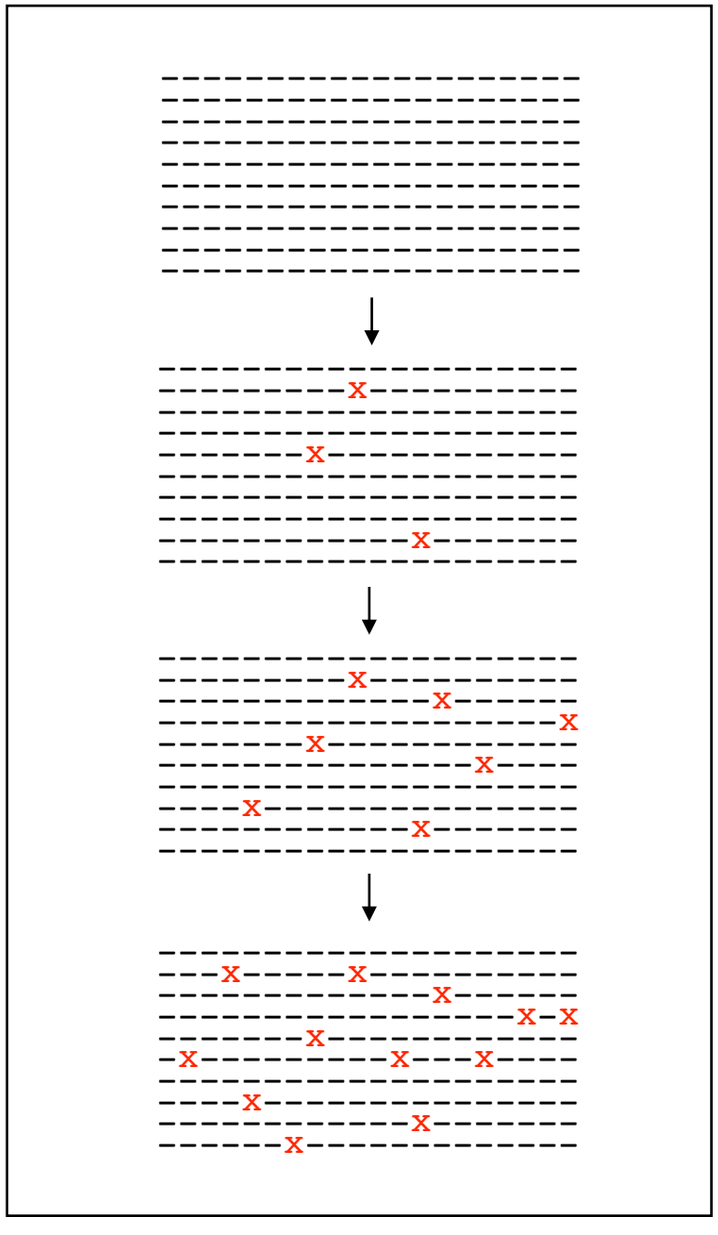
"It is natural selection that gives direction to changes, orients chance, and slowly, progressively produces more complex structures, new organs, and new species. Novelties come from previously unseen association of old material. **To create is to recombine.**"

Jacob (1977) Evolution and tinkering. Science (196):1161-1166

Alleles that exist today arose through ancient mutation events… this process however is slow and limited.



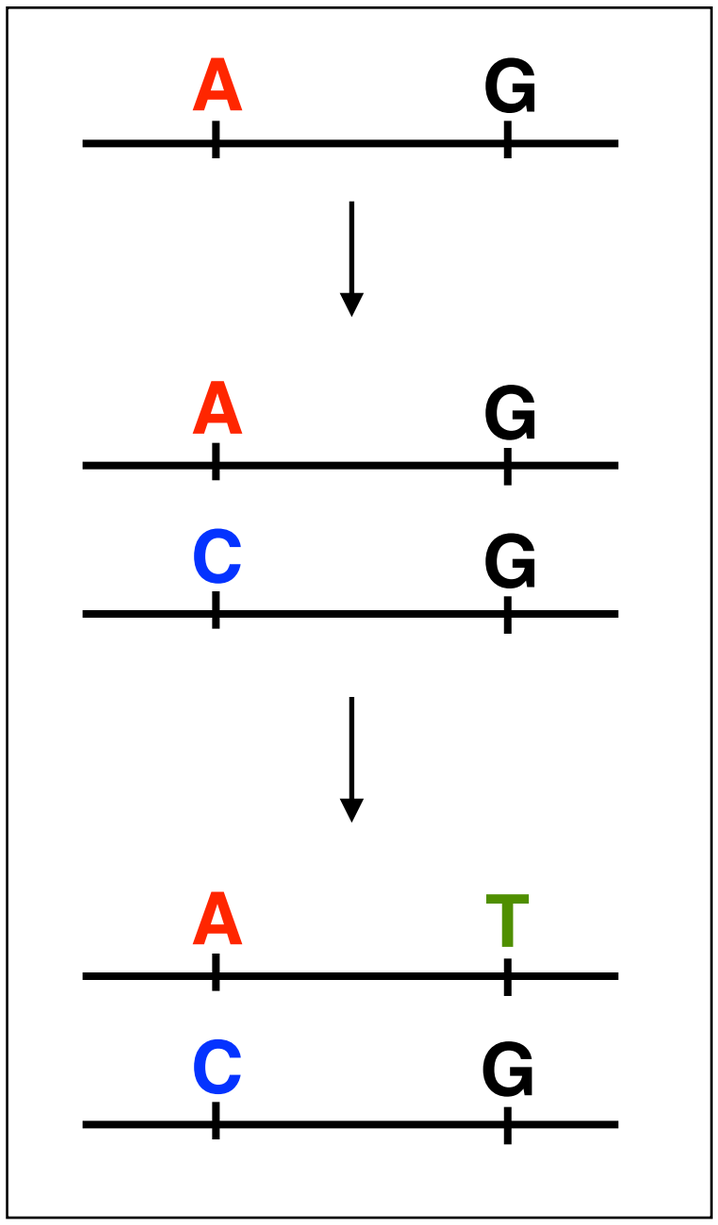
The steady accumulation of (deleterious) mutations may lead to a phenomenon known as Muller's Ratchet.



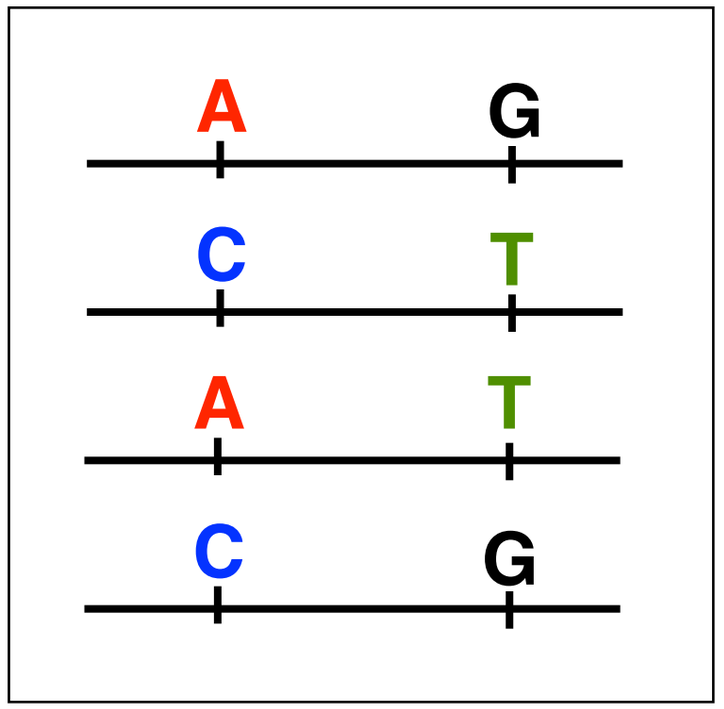
The main effect of Muller's ratchet is the accumulation of slightly deleterious mutations, despite the fact that selection opposes their fixation in the population. Such mutation accumulation can sometimes lead to eventual extinction. This occurs if …

- recombination is absent
- population size is finite
- almost no back mutations occur
- slightly deleterious mutation rates are high
- purifying selection is too weak to remove all new deleterious mutations

Alleles that exist today arose through ancient mutation events…



recombination



The **increase in genetic diversity produced by recombination** is one of the greatest assets of sexual reproduction and gives sexual populations several evolutionary advantages over asexual communities, in particular a faster response to changing environments.

A simple measure of the amount of recombination is the degree of **linkage disequilibrium**.


MORE
THINGS
CONSIDERED
2

A problem of interpreting values of D is that its maximum depends on the allele-frequencies. Therefore it is helpful to standardize D, for example as the percentage of its maximum value (if D is positive) or its minimum value (if D is negative).

D_{min} = the larger of *-p1q1* and *-p2q2*

D_{max} = the smaller of *p1q2* and *p2q1*

$$D^* = D' = \frac{D}{D_{min} \text{ or } D_{max}}$$

Lewontin (1964)

Another standardized measure of linkage disequilibrium is ρ (rho), which can range from -1 to +1.

$\rho = D / \text{sqrt}(p_1q_1\, p_2q_2)$ > $r^2 = D^2 / [p_1q_1\, p_2q_2]$

Hill and Robertson (1968)

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

- 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

- 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

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

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

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

  - $r^2 = 1$ 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)

$$X'_{ij} = (1-r)X_{ij} + rx_{ij}$$

$$X'_{ij} = (1-r)X_{ij} + rx_{ij} = (1-r)X_{ij} + rp_iq_j$$

$$= (1-r)(x_{ij}+D_0)+rx_{ij} = x_{ij}+D_0-\cancel{rx_{ij}}-rD_0+\cancel{rx_{ij}}$$

$$= x_{ij}+D_0-rD_0$$

$$X'_{ij} = x_{ij}+D_0(1-r)$$