SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

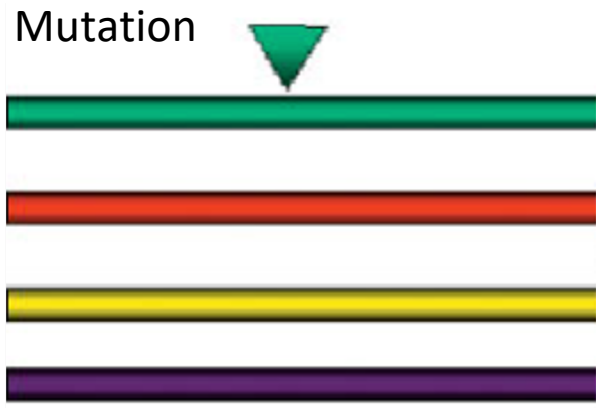
Session 4:
Linkage Disequilibrium (LD) and Hardy-Weinberg Equilibrium (HWE)

Linkage Disequilibrium (LD)

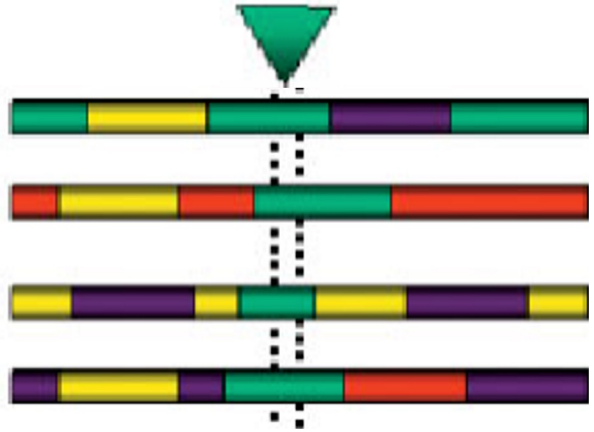
- LD is the non-random association between alleles at two or more loci

- Why do we care about LD?
  - Probably the most recurring term in this course (and in genetic epidemiology)

  - Population genetics, association mapping, imputation etc. are all based on LD.

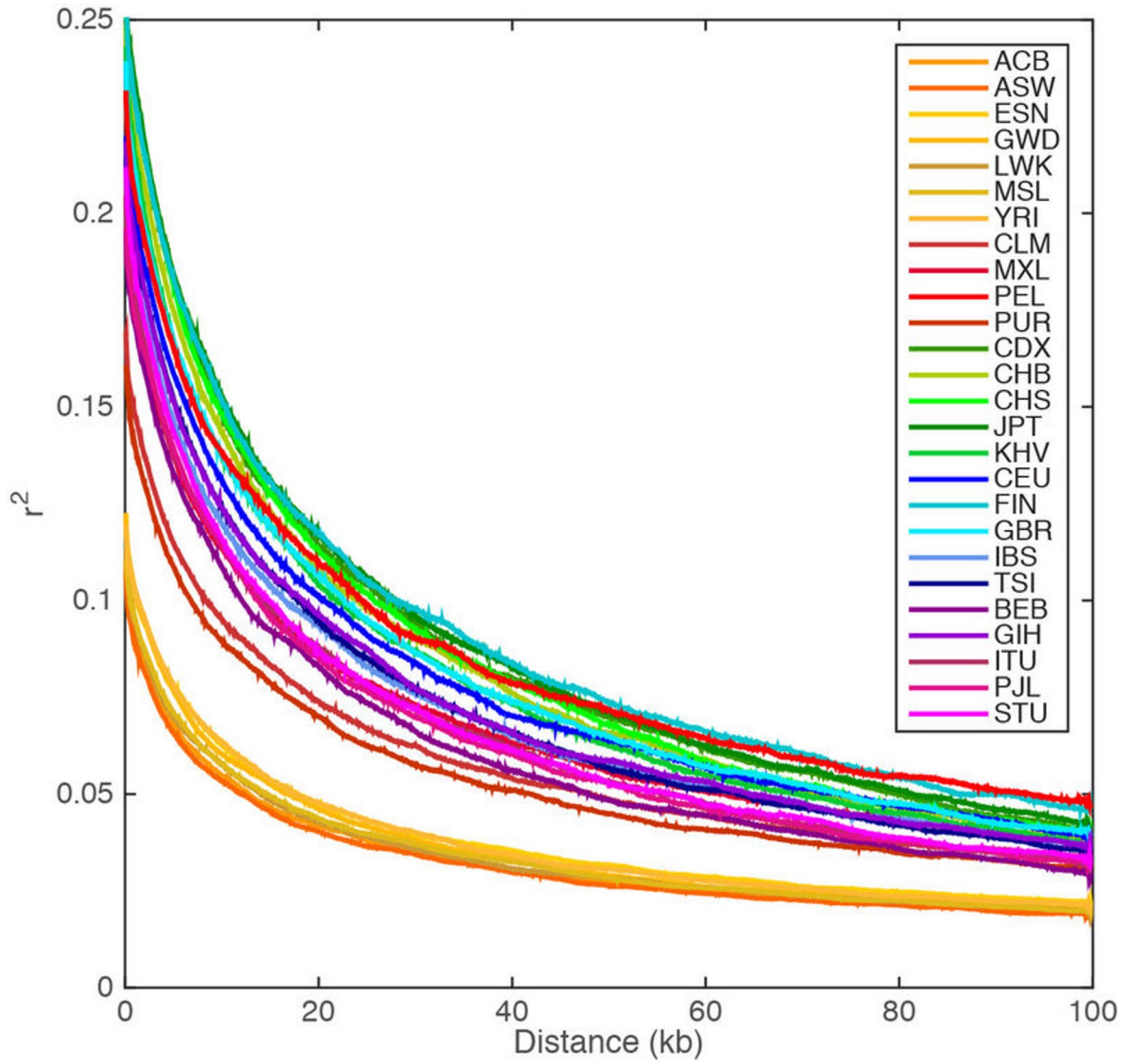


Ancestral Haplotypes



Present-day Haplotypes

SNPs physically closer to each other tend to be in stronger LD



1000 Genomes Project, Nature 2015

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Factors that influence LD

- New mutations

Factors that influence LD

- New mutations
- Genetic drift
  - Have larger effect on smaller populations

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
  - The faster the population grows, the less LD.

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
  - The larger difference between the two populations, the more impact on LD

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
  - Causes long-range LD

Factors that influence LD

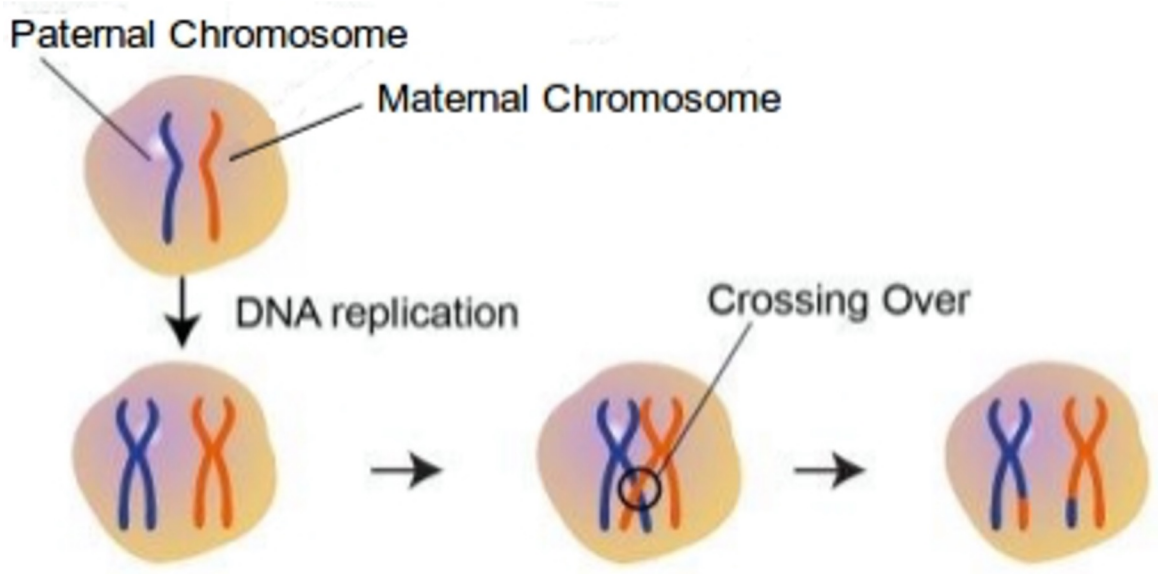
- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency
- Recombination (recombination hotspots)

Recombination

- Alleles on the same chromosome are inherited together unless *recombination (crossing over)* occurs
- The probability of recombination between two alleles increases with the distance between them
- The parameter θ estimates the probability of observing a recombinant gamete (the *recombination fraction*)



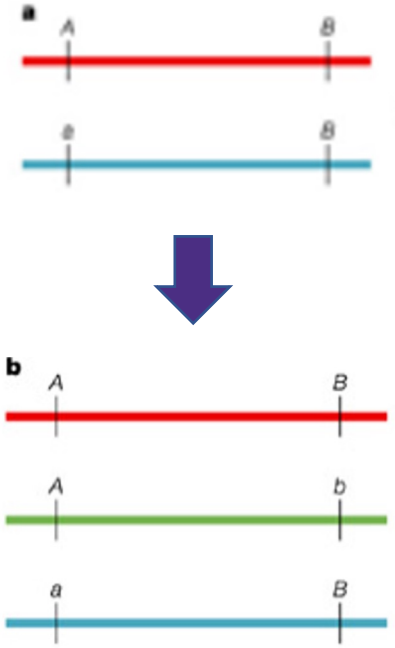
Recombination breaks up LD

Start with a polymorphic locus with alleles *A* and *a*.

Ardlie, Nat Rev Genetics 2002

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Recombination breaks up LD

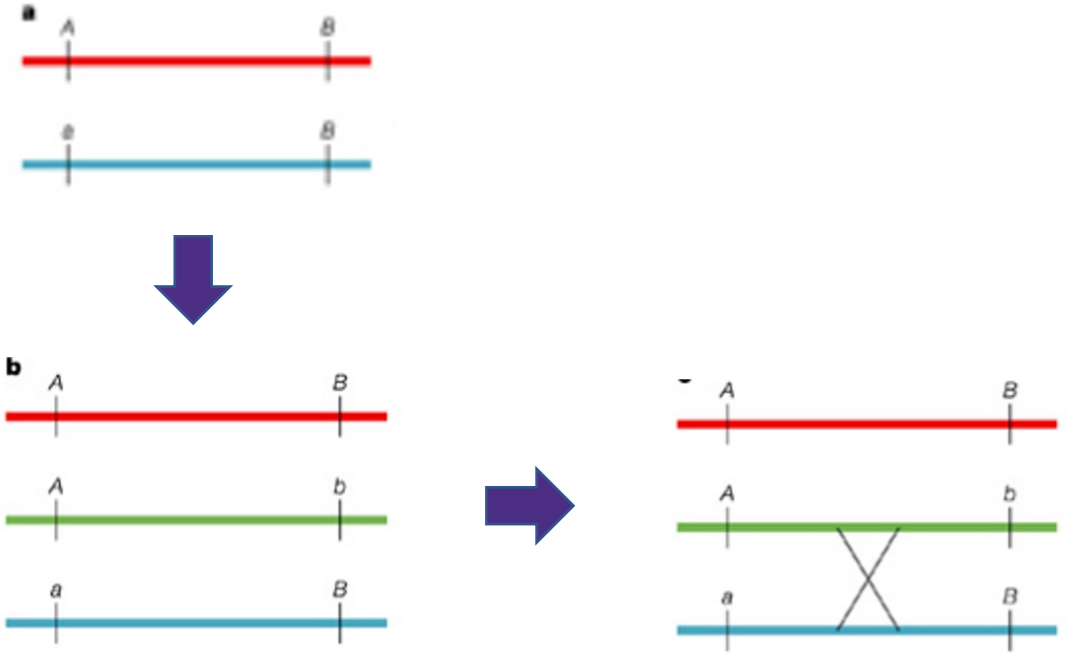


When a mutation occurs at a nearby locus (B->b), this occurs on a single chromosome bearing either allele *A* or *a* at the first locus (*A* in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The *b* allele will always be found on a chromosome with the *A* allele.

Ardlie, Nat Rev Genetics 2002

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

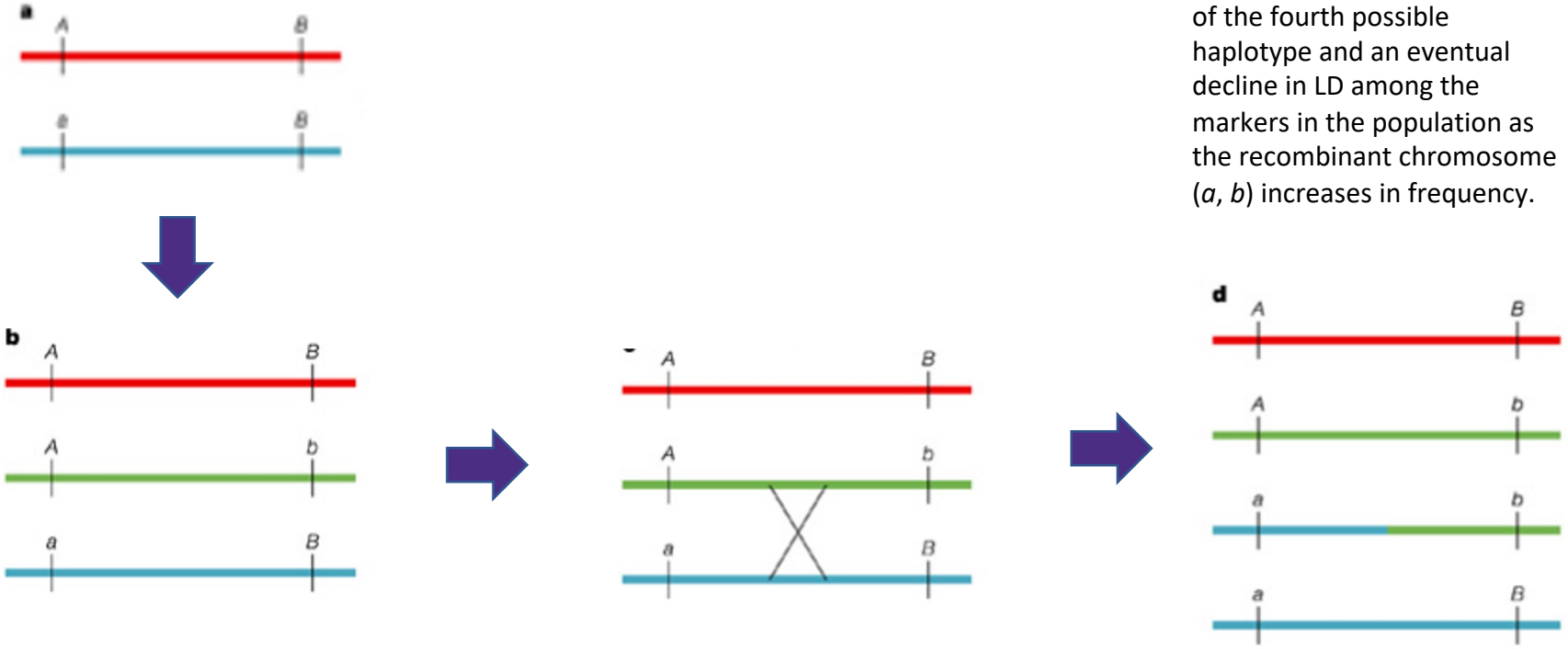
Recombination breaks up LD



With time, a recombination event will take place and the association between alleles at the two loci will gradually be disrupted

Ardlie, Nat Rev Genetics 2002

Recombination breaks up LD



This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (*a*, *b*) increases in frequency.

Ardlie, Nat Rev Genetics 2002

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Calculation of LD

There are 4 possible haplotypes for SNP1 (Aa) and SNP2 (Bb)

	SNP2 (Bb)		
SNP1 (Aa)	AB	Ab	p_A
	aB	ab	p_a
	p_B	p_b	1

Calculation of LD

Haplotype frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB} = p_A p_B$	Ab $p_{Ab} = p_A p_b$	p_A
	aB $p_{aB} = p_a p_B$	ab $p_{ab} = p_a p_b$	p_a
	p_B	p_b	1

Calculation of LD

Haplotype frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other
(This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB}= p_A p_B$	Ab $p_{Ab}= p_A p_b$	p_A
	aB $p_{aB}= p_a p_B$	ab $p_{ab}= p_a p_b$	p_a
	p_B	p_b	1

	SNP2 (Bb)		
SNP1 (Aa)	$p_{AB}=$ $0.6*0.8=0.48$	$p_{Ab}=$ $0.6*0.2=0.12$	$p_A=0.6$
	$p_{aB}=$ $0.4*0.8=0.32$	$p_{ab}=$ $0.4*0.2=0.08$	$P_a=0.4$
	$p_B=0.8$	$P_b=0.2$	1

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Calculation of LD

$$D=p_{AB}p_{ab}-p_{Ab}p_{aB}$$

Haplotype frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB}= p_Ap_B$	Ab $p_{Ab}= p_Ap_b$	p_A
	aB $p_{aB}= p_ap_B$	ab $p_{ab}= p_ap_b$	p_a
	p_B	p_b	1

We can infer LD as the deviation of observed haplotype frequencies from what is expected if SNP1 and SNP2 are independent of each other

	SNP2 (Bb)		
SNP1 (Aa)	AB p_Ap_B+D	Ab p_Ap_b-D	p_A
	aB p_ap_B-D	ab p_ap_b+D	p_a
	p_B	p_b	1

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Instead of D, we often express LD in terms of D' (normalized D) or r^2 (correlation coefficient)

$$D' = \frac{D}{D_{max}},$$

$$D_{max} = \begin{cases} min\{p_A p_B, (1-p_A)(1-p_B)\}, when\ D < 0 \\ min\{p_A(1-p_B), (1-p_A)p_B\}, when\ D > 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

r^2

- ranges from 0 [no LD] to 1 [perfect LD]
- is sensitive to allele frequencies

D' and r^2

- If 2 SNPs are independent (not inherited together), D´=0 and r^2=0 regardless of allele frequencies.
- For 2 SNPs, there are 4 possible haplotypes. If not all 4 haplotypes are observed, D'=1. D'<1 indicates a recombination event between the SNPs.
- If 2 SNPs have allele frequency 50% and are always inherited together, both D'=1 and r^2=1.
- If SNP1 has frequencies p_A=0.5, p_a=0.5 and SNP2 has p_B=0.99, p_b=0.01, and b is always inherited with A => D'=1 BUT $r^2 < 1$
  - there cannot be an 'ab' haplotype => D'=1
  - Given SNP2=B, you cannot say whether SNP1 will be A or a. => $r^2 < 1$. Thus, r^2 is sensitive to allele frequencies.

How does LD influence our study power?

- If a genotyped SNP M and a causal SNP G are in LD with each other with some r^2, then a study with N cases and controls which measures M (but not G) will have the same power to detect an association between M and disease as a study with r^2N cases and controls that directly measured G.

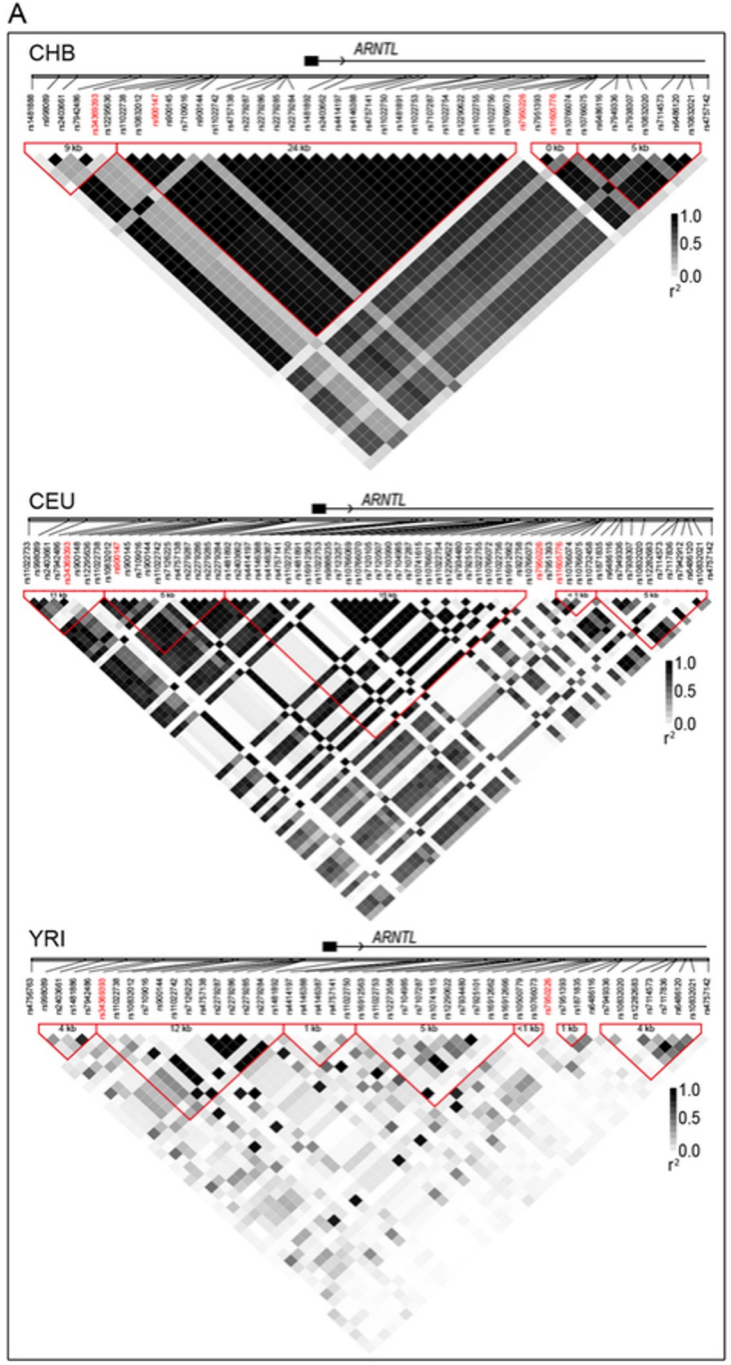
- r^2 N is the "effective sample size"
  - If the r^2 between your genotyped SNP M and causal SNP G is r^2 =0.5 you need to double your sample size to obtain the same power as if you had measured (genotyped) G directly.

Pritchard and Przeworski (2001)

Sometimes just a few SNPs are enough to explain the genetic variation in a region. These SNPs are called ‘tag’ SNPs

Caveat 1: Tag SNPs are not particularly efficient for rare SNPs (remember that r^2 depends on allele frequency)

Caveat 2: LD is population-specific



Pair-wise correlations (r^2): Black means r^2~1

Region associated with Parkinson’s Disease in Han Chinese

Gu, Sci Rep 2015

BREAKOUT ACTIVITY

- We will explore LD using the NCI LDLink online tools (ldlink.nci.nih.gov). There are many different tools to check out, but we will use the LDpop tab.
  - Compare LD for two SNPs that are involved in drug metabolism (rs776746 and rs2740574). Type these into the two SNP boxes (variant RSID) and select "(ALL) all populations", "R^2", then "calculate." After a few seconds, you should see a map of the world with tear drops showing different populations that have been studied.

Q1: Compare R^2 among Colombians from Medellin, Colombia (CLM), British in England and Scotland (GBR) and Luhya in Webue, Kenya (LWK)? You can find the details for each population in the table, or by clicking on the corresponding tear drop.

Q2: Why might these LD values be so different between these populations?

SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

Hardy-Weinberg Equilibrium

Why do we care about HWE?

- One of the most fundamental concepts in population genetics
- Most statistical methods assume HWE in their model assumptions
- In practice, a very efficient approach to detect low-quality genotype data

The Hardy-Weinberg principle

- Assume that…
  - Population is large
  - Mating is random
  - No immigration or emigration
  - Natural selection is not occurring (all genotypes have an equal chance of surviving and reproducing)
  - No mutations

- If these assumptions are true, we say that a population is not evolving (allele frequencies stay the same) and in **Hardy-Weinberg Equilibrium**

The Hardy-Weinberg law under the assumption of non-evolving allele frequencies

- The Hardy-Weinberg Law provides two equations allowing us to relate the expected allele and genotype frequencies to each other
- Assume a SNP with alleles A (frequency p) and a (frequency q)
- **p+q=1 (allele frequencies)**
- **$p^2+2qp+q^2=1$ (genotype frequencies)**

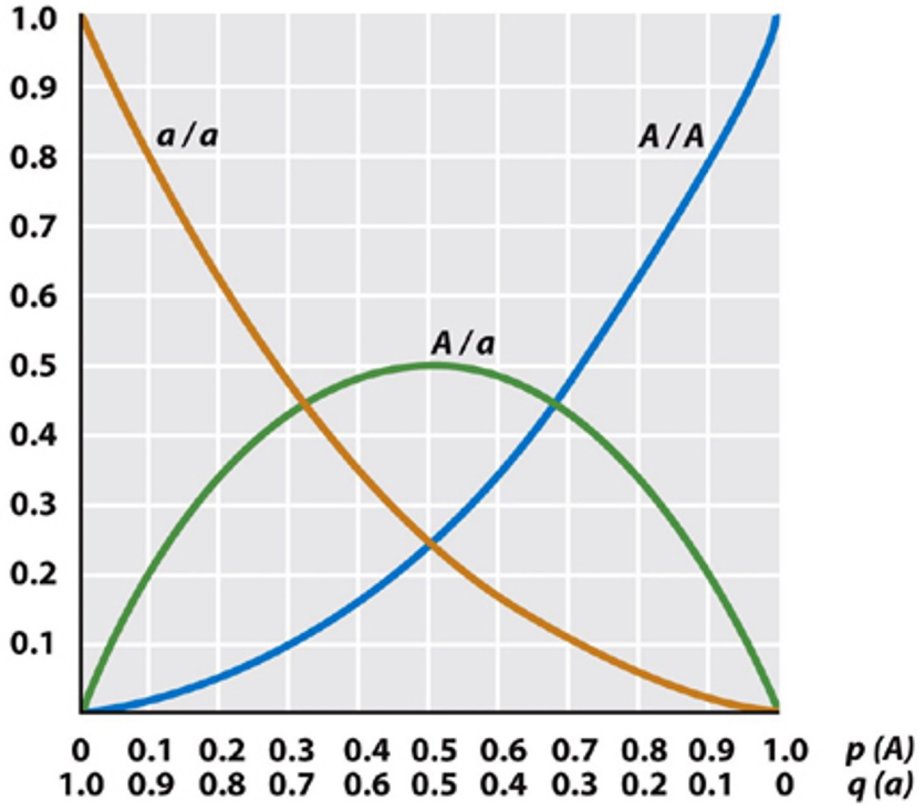
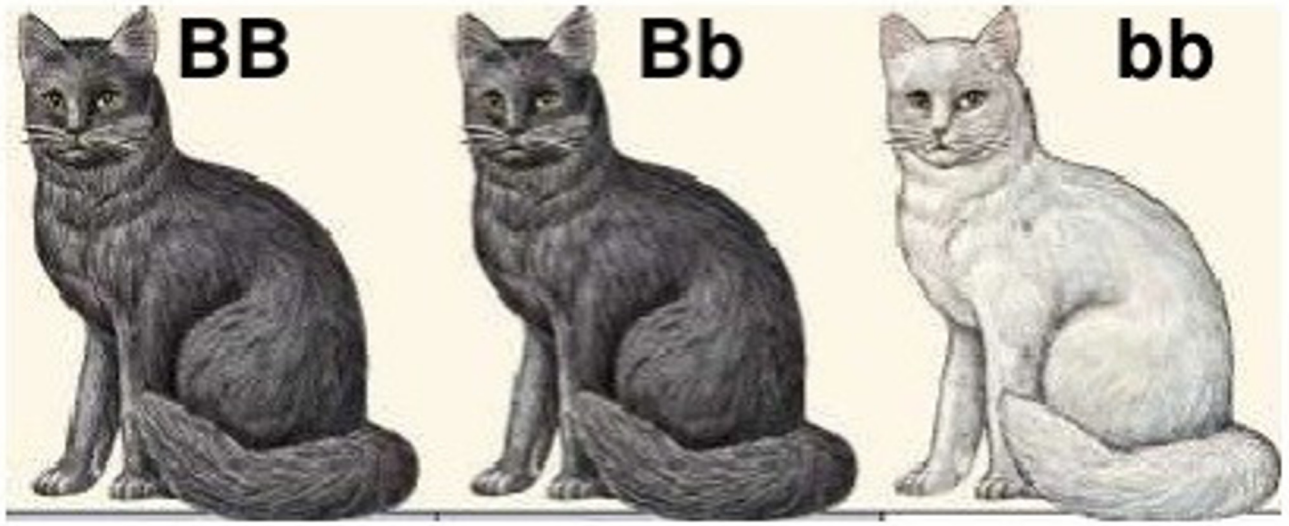


Figure 17-5
Introduction to Genetic Analysis, Ninth Edition

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

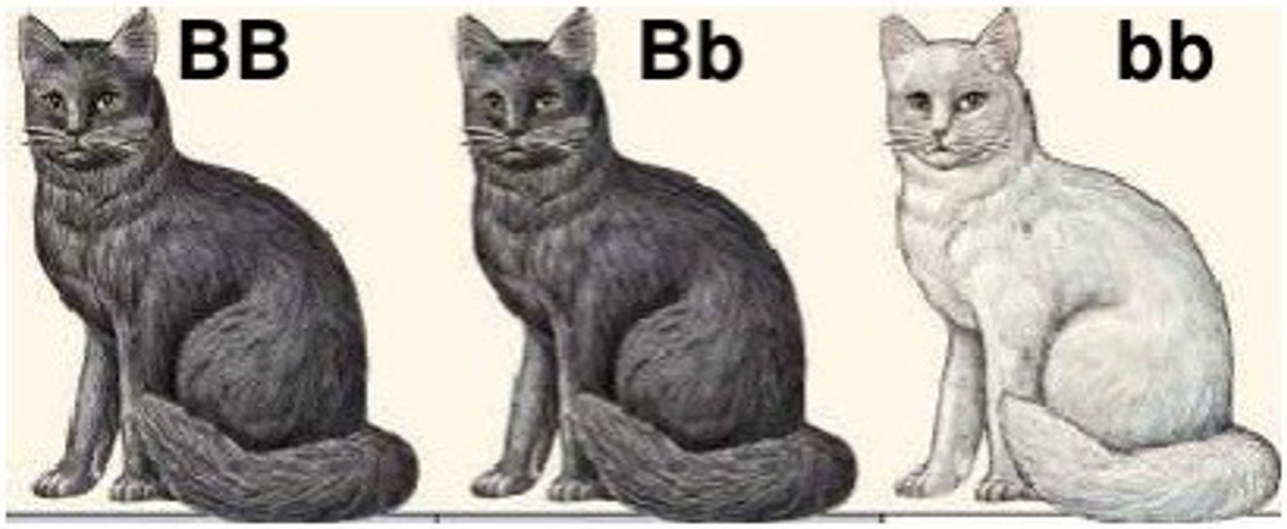
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- Freq(B)=p, Freq(b)=q, Freq(BB)=p^2, Freq(Bb)=2pq, Freq(bb)=q^2
- p+q=1
- $p^2+2qp+q^2=1$



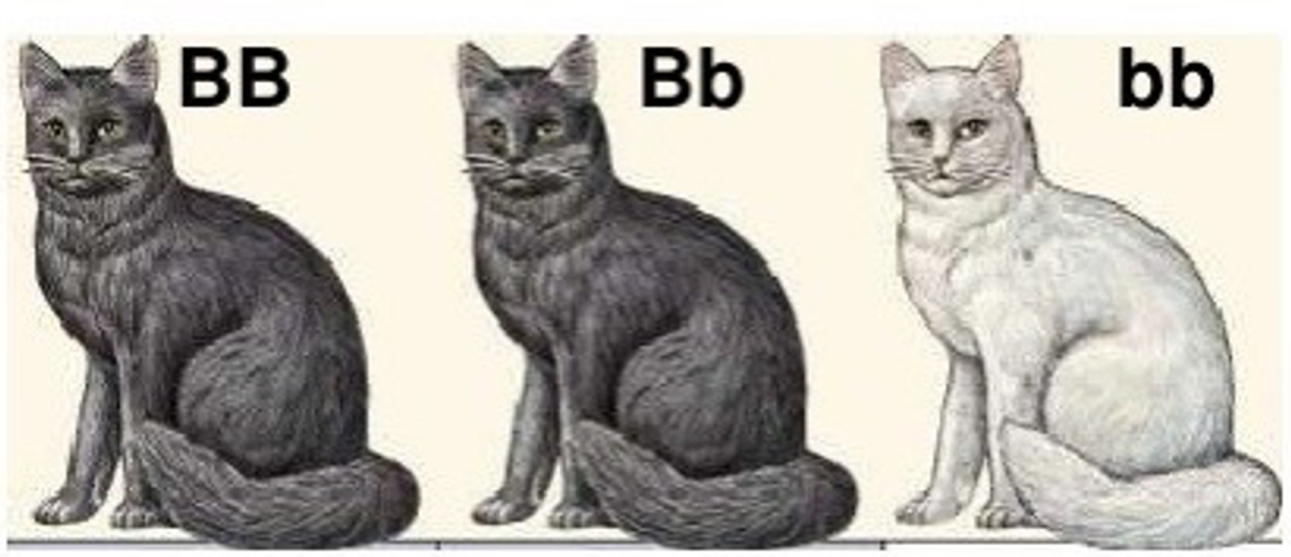
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- Freq(B)=p, Freq(b)=q, Freq(BB)=q^2, Freq(Bb)=2pq, Freq(bb)=q^2
- p+q=1
- $p^2+2qp+q^2=1$
- q^2 =Freq(bb)=16/100=0.16



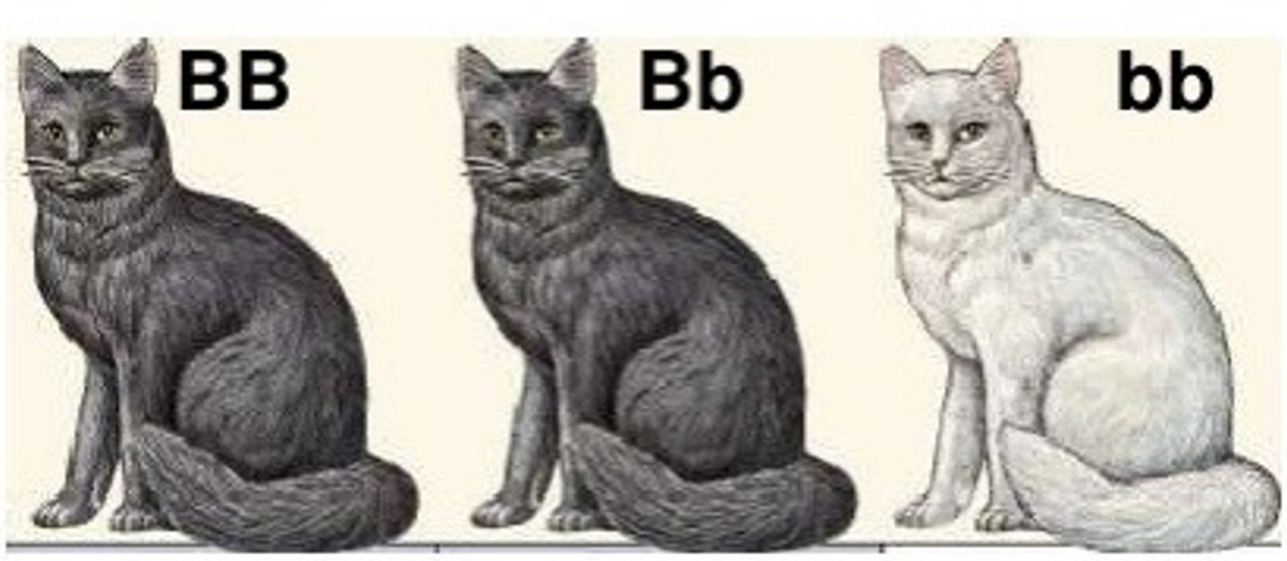
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- Freq(B)=p, Freq(b)=q, Freq(BB)=q^2, Freq(Bb)=2pq, Freq(bb)=q^2
- p+q=1
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- q=0.4, p=0.6
- **0.4*200=80 b**
- **0.6*200=120 B**



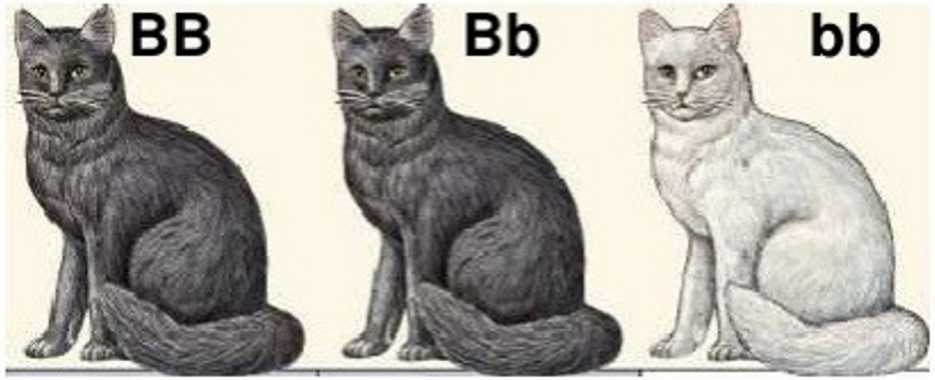
HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- $p+q=1$
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- $q=0.4$, $p=0.6$
- $p^2=0.36$



HWE example

- Assume 100 cats (200 alleles) with alleles B and b. 16 of the cats are white (genotype bb). If you assume HWE, what are the allele (B,b) and genotype (BB, Bb, bb) frequencies?
- Freq(B)=p, Freq(b)=q, Freq(BB)=q^2, Freq(Bb)=2pq, Freq(bb)=q^2
- p+q=1
- $p^2+2qp+q^2=1$
- $q^2=0.16$
- q=0.4, p=0.6
- $p^2=0.36$
- 2pq=2x0.6x0.4=0.48
- Final genotype count: 36 BB, 48 Bb, 16 bb



In practice...

- We use HWE to check the validity of our genotyped data.
- We compare the observed genotype frequencies to the expected genotype frequencies based on the allele frequencies in our data.
- Any deviations from the expected can indicate potential problems
  - Too few heterozygotes (inbreed population)
  - Too many heterozygotes (mixed DNA, low DNA concentration, mixed populations)
- There are many ways to calculate HWE and detect deviations. One common approach is a chi-square test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

LD and HWE are useful tools to detect evolutionary forces acting on a population such as population bottlenecks

deMenocal & Stringer, Nature 2016

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH