Introduction to Genomic Epidemiology

Human Genomic Epidemiology – Asia

Pak Sham

13th June 2022

INTRODUCTION

Learning objectives

- Introduce key concepts and terminologies in genomic epidemiology, from classical genetics to GWAS

Individual differences

- People are remarkably similar to each other
  - Anatomy
  - Physiology
  - Psychology
- At the same time, no two individuals are identical. People differ in:
  - Appearance
  - Personality and temperament
  - Physical fitness
- Importantly, people differ in disease status
  - Understanding why is the goal of genomic epidemiology

Causes of disease

- Inborn (congenital)
  - Genetic inheritance
  - Prenatal environment
- Acquired
  - Upbringing (parental care, education)
  - Physical environment (e.g. toxins, pathogens)
  - Social environment (family, friends, colleagues)
  - Habits and lifestyle (nutrition, exercise, drug use)
  - Accidents

Genetic variation

- The genetic material differ between people, which partly explain their difference in disease occurrence and predisposition
- All variations arise from mutations that occurred in ancestors and subsequently passed from generation to generation down to the current population.
- Genetic variations can range in size from a single nucleotide (single nucleotide variants, SNV), to large segments of DNA (structural variants), up to an entire chromosome (aneuploidies).

Genomic epidemiology

- Genomic epidemiology is the study of genetic variations, and their relationships to disease, on a whole-genome scale, in populations.
- The goal of genomic epidemiology is to understand aetiological factors and biological mechanisms of diseases, that explain the occurrence of disease geographically, temporally, and in particular population groups, in order to formulate effective strategies for the surveillance, prevention, early detection, and treatment of disease.

CLASSICAL GENETICS

Classical genetics

- Based solely on visible results of reproductive acts, i.e. **phenotypes**. Genotypes are inferred, not directly observed.
- Goes back to the breeding experiments on the garden pea by Gregor Mendel.
- Segregation analysis: to assessing whether a phenotype is a Mendelian trait
- Heritability analysis: to estimate the overall genetic contribution to a phenotype

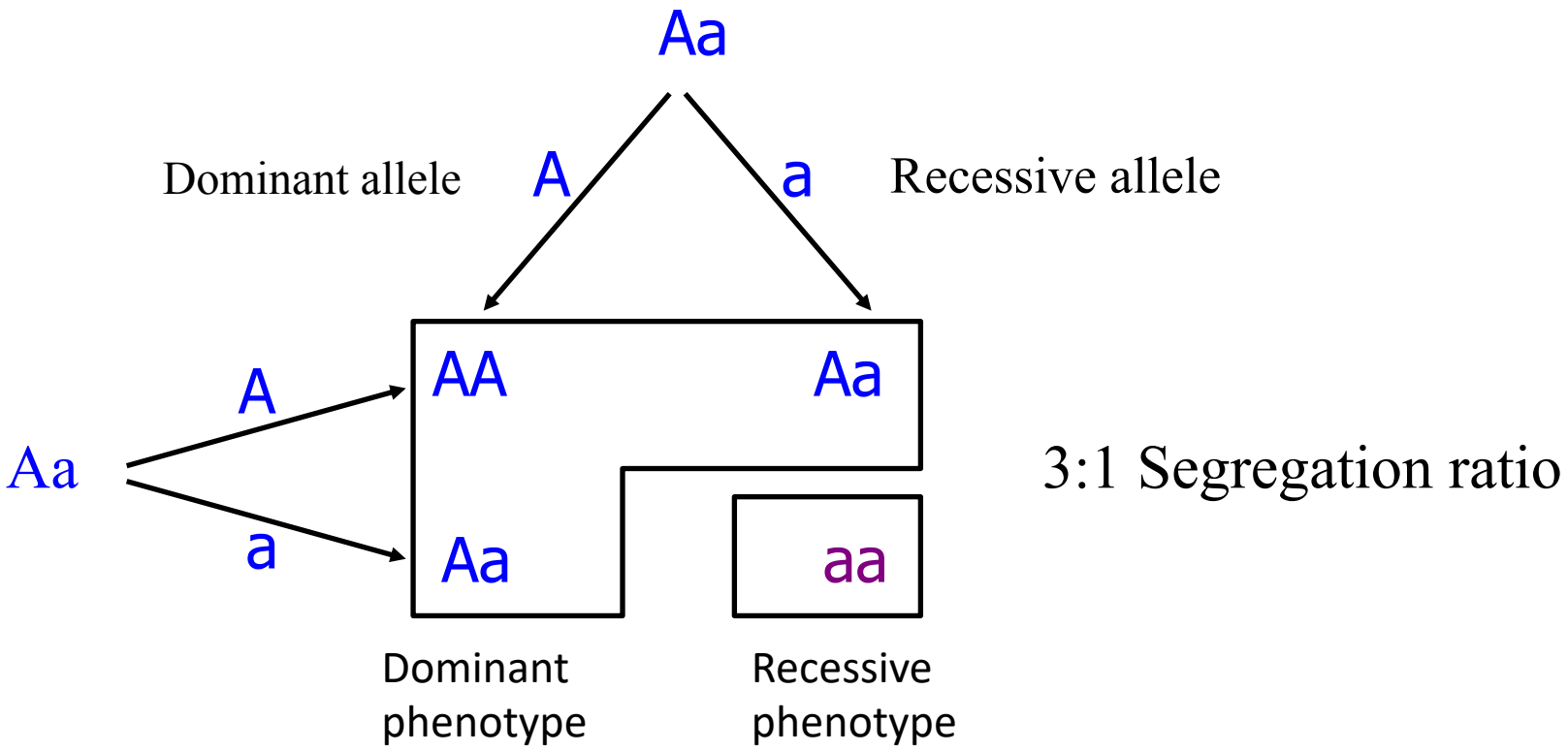
Mendelian traits

- What is a Mendelian trait?
  - Determined by the genotype at a single locus
- Example:
  - Locus with 2 alleles: e.g. A (reference) and G (variant)
  - Three genotypes: AA, GG, AG
  - Genotype phenotype relationship
    - AA -> Normal (“Wild Type”)
    - AG -> Affected (Heterozygous “Mutant”)
    - GG -> Affected (Homozygous “Mutant”)

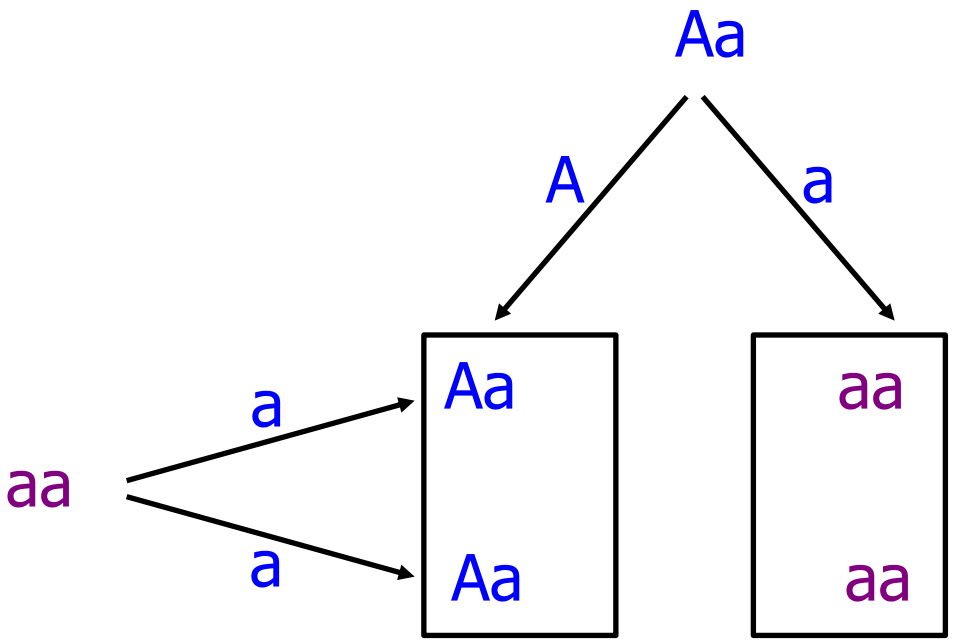
Segregation analysis

- How to determine whether whether a trait is inherited as a Mendelian trait, without genotypings.
- In animals / plants
  - Perform crosses
  - Consistency of offspring trait distribution with Mendelian ratios.
- In humans
  - In population: trait is distinctive and (usually) rare
  - In families: characteristic Mendelian patterns

Mendelian intercross

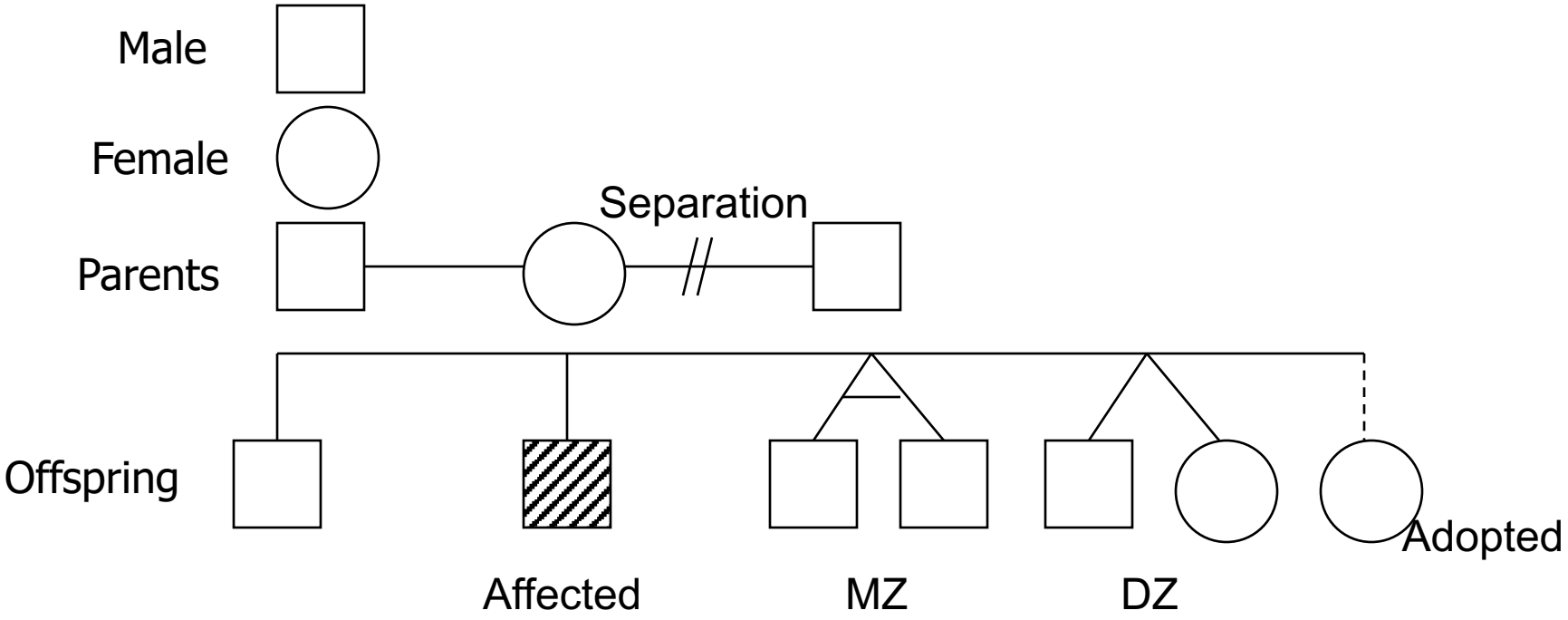


Mendelian backcross



1:1 Segregation ratio

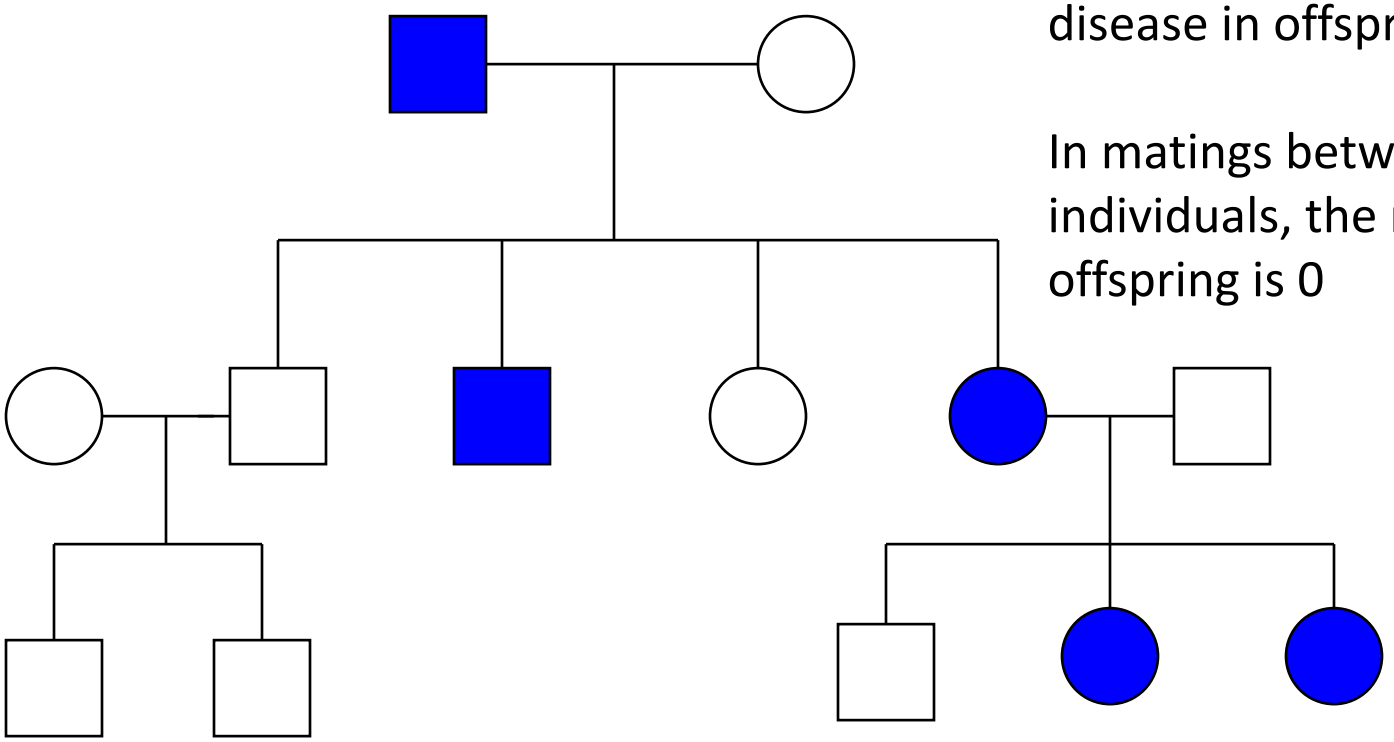
Pedigree Tree Symbols



MZ: Monozygotic twins; developed from same fertilized ovum

DZ: Dizygotic twins; developed from 2 separate fertilized ova

Autosomal Dominant Diseases

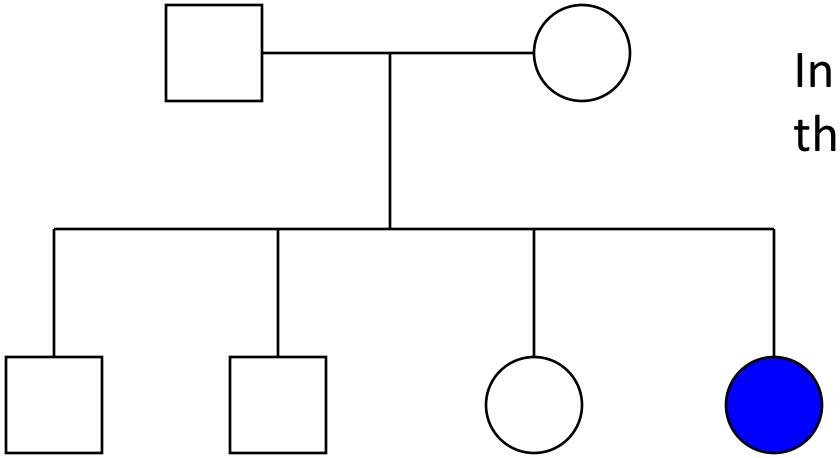


In matings between affected and unaffected individuals, the risk of disease in offspring is 1/2

In matings between two unaffected individuals, the risk of disease in offspring is 0

Complication: some individuals with the mutation may not have the disease because of protective factors (e.g. age)

Autosomal Recessive Diseases

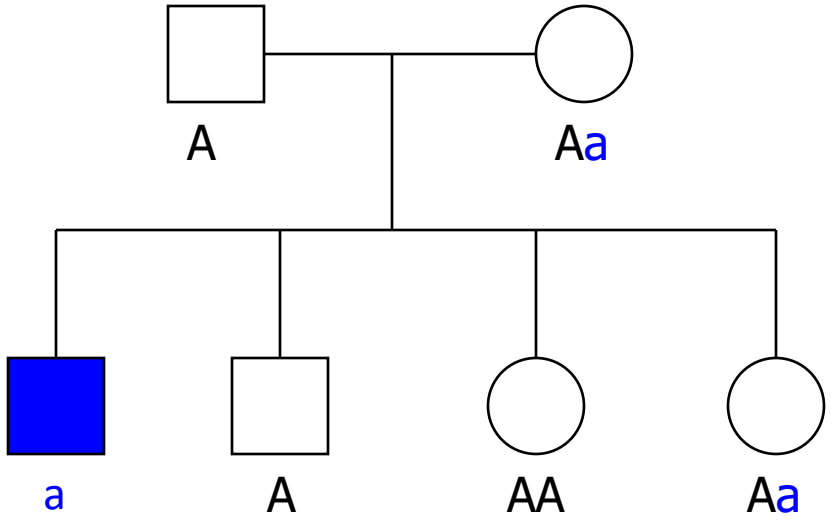


In matings between 2 heterozygous carriers, the risk of disease in offspring is 1/4

Complication: in these families the genotypes of the parents have to be inferred from the presence of an affected offspring.

This leads to the problem of “incomplete selection” which, if ignored, would lead to the over-estimation of the segregation ratio.

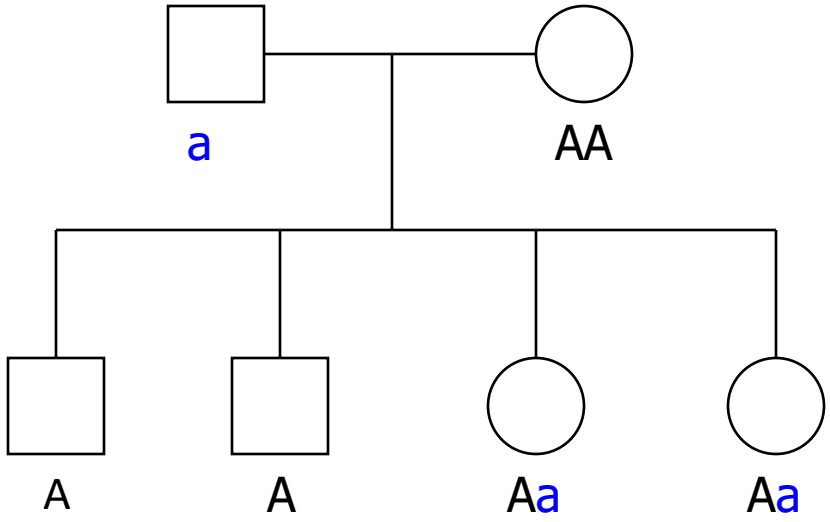
X-linked Recessive Disorders



In matings between normal father and carrier mother, the risk in sons is 1/2, while the risk in daughters is 0

Complication: X inactivation may give rise to mild form of disease of variable severity in heterozygous females

X-linked Recessive Disorders



In matings between affected father and normal mother, the risk in sons is 0, while all daughters are heterozygous carriers and may have mild form of disease

Quiz 1

- You are studying a rare disease. From a clinic you have recruited families where both parents are normal but have at least one affected child.
- Selecting all the families with 2 children, in what proportion of such families will both children be affected, under the hypothesis that the disease is autosomal recessive?

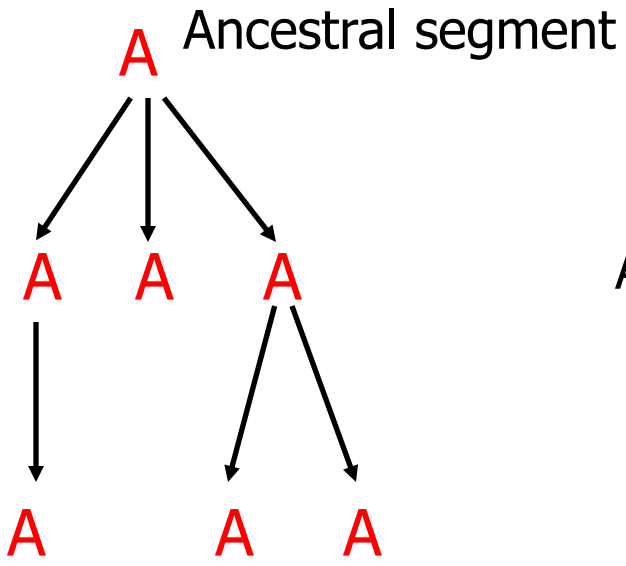
A: 1/4
B: 1/7
C: 1/16

Quiz 1

- Consider all families where both parents are heterozygous carriers with 2 offspring
- In $(3/4)^2=9/16$ of these families, both offspring will be unaffected. Such families will not among families with at least one affected child
- In $(1/4)^2=1/16$ of these families, both offspring will be affected.
- In the remaining 6/16 of these families, 1 offspring will be affected.
- Thus the proportion of families where both children are affected is 1/7

Identity-by-Descent (IBD)

- Two DNA segments (e.g. genes) are identical-by-descent if they are descended from, and there replicates of, a single ancestral DNA segment



All 7 A "alleles" are IBD with each other

Kinship coefficients

The kinship coefficient (*K*) between two individuals is defined as the probability that two alleles, one from each individual, drawn at random at any autosomal locus, will be identical-by-descent (IBD).

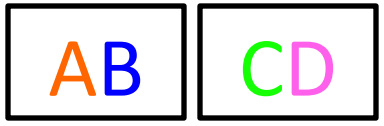
What is the kinship coefficient for two individuals who are

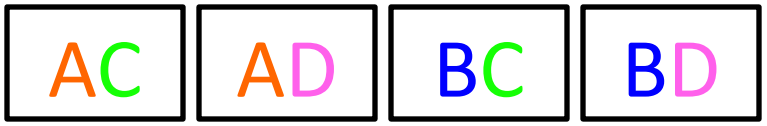
1. Full siblings
2. Monozygotic (MZ) twins

Asssuming that the parents are not inbred and are not genetically related to each other?

Kinship between full sibs

As the two parents are neither inbred nor related to each other, they have 4 distinct DNA segments at any genomic location:





Each child will have 1 of 4 equally likely genotypes:

Considering 2 children jointly, there are 4x4=16 possible combinations:

	AC	AD	BC	BD
AC	½	¼	¼	0
AD	¼	½	0	¼
BC	¼	0	½	¼
BD	0	¼	¼	½

The overall kinship coefficient (*K*) is the average of *K* of these 16 equally probable scenarios, i.e. ¼

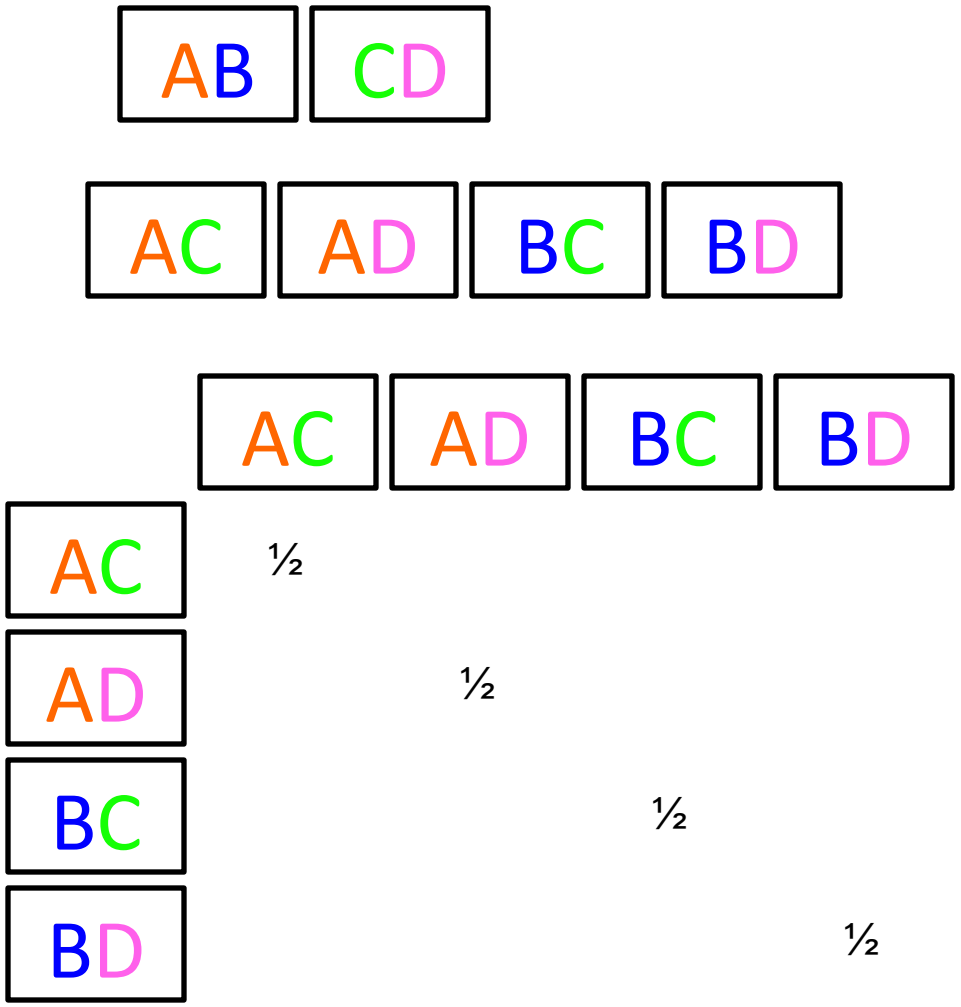
Kinship between MZ twins

As the two parents are neither inbred nor related to each other, they have 4 distinct DNA segments at any genomic location:

Twin 1 will have 1 of 4 equally likely genotypes:

Since Twin 2 is derived from the same zygote, it will always have the same genotype as Twin 1.

The overall kinship coefficient (*K*) is the average of *K* of these 4 equally probable scenarios, i.e. ½



Quantitative traits

- Many traits can be measured numerically, e.g. height, body mass index, blood pressure, blood cholesterol level, general intelligence
- Many such traits have a normal distribution in the population, or can be mathematically transformed to have a normal distribution
- Central limit theorem: the sum of **many small independent influences** will have a normal distribution, regardless of the distribution of the influences
- This suggests that normally distributed traits may be determined by many small independent influences (including genetic differences) – **polygenic multifactorial model**

IFA.com – The Galton Board: Math in Motion
https://www.youtube.com/watch?v=9QuPHf1xi-4&t=78s

Variances, covariances and correlations

- The expectation of a variable *X*, written as *E*(*X*), is the population average (mean) of the variable

$$E(X) = \mu$$

- We can also take the expectations of functions of *X*, e.g.

$$E(X - \mu) = E(X) - \mu = 0 \qquad E\left(X - \mu\right)^2 = \sigma^2$$

- The latter is defined as the variance of *X*, *Var*(*X*). With 2 variables, *X* and *Y*, we define their covariance as

$$E(X - \mu_X)(Y - \mu_Y) = \sigma_{XY}$$

- Obviously, Cov(*X*,*X*)=Var(*X*). A covariance can be rescaled to a correlation by

$$\frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho_{XY}$$

Variance partitioning

- When a trait is determined additively by 2 variables

$$P = X + Y$$

- Its covariance with another variable Q, is given by

$$\begin{aligned} Cov(P,Q) &= Cov(X+Y,Q) \\ &= Cov(X,Q) + Cov(Y,Q) \end{aligned}$$

- Thus

$$\begin{aligned} Var(P) &= Cov(P,P) \\ &= Cov(X+Y,X+Y) \\ &= Cov(X,X) + Cov(Y,Y) + 2Cov(X,Y) \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

Partitioning single locus effects

- Let the two alleles present at a locus in an individual be X_1 and X_2, which are coded 0 (reference) and 1 (variant)
- The expected values of X_1 and X_2 are both p, the frequency of the variant in the population
- The influence of the alleles on the trait consists of both the main (**additive**) effects of the two alleles, and their interaction (**dominance**):

$$Y = b(X_1-p) + b(X_2-p) + g(X_1-p)(X_2-p)$$

Additive Dominance

- The mean centering ensures that the main effects are uncorrelated with the interaction term.

Genetic variance components

- Additive: $V_A = 2p(1-p)b^2$
- Dominance: $V_D = p^2(1-p)^2g^2$
- Note 2p(1-p) is the expected heterozygosity under random mating
- Dominance variance is usually much smaller than additive genetic variance, especially when p is close to 0 or 1

Genetic covariances

- Let the alleles present in 2 individuals be (X_{11},X_{12}) and (X_{21},X_{22}), then the covariances between the genetic effects of the 2 individuals are
- For additive effects

$$Cov[(bX_{11}+bX_{12}),(bX_{21}+bX_{22})] = 2KV_A$$

- For dominance

$$Cov[g(X_{11}-p)(X_{12}-p),g(X_{21}-p)(X_{22}-p)] = P(K=½\ )V_D$$

- **Intuitively**, the additive effects of 2 alleles (1 from each individual) are shared if the 2 alleles are IBD, and this occurs with probability *K*, with variance $V_A/2$. Since there are 4 pairs of alleles, the overall covariance is $4xKxV_A/2 = 2KV_A$
- In contrast, the dominance of the two genotypes are shared only when *K*=½ (meaning that the 2 genotypes must be identical).

Overall variance components

- Let's now redefine V_A and V_D as the **TOTAL** additive genetic and dominance variances, respectively, across all variable sites in the genome.
- If the overall variance of the trait is V_T, then the proportion of variance contributed by additive effects is V_A/V_T (also called **narrow heritability**, h^2), and the proportion of variance contributed by dominance is V_D/V_T (d^2)
- Note, the broad heritability is the proportion of overall variance contributed by all genetic effect, including additive effects, dominance and epistasis (interactions between alleles of different loci)
- One important goal in classical quantitative genetics is to estimate h^2 and d^2

Parent-offspring studies

- An offspring inherits exactly one copy of the parent's two alleles, across the whole-genome.
- Not only is *K*=¼ overall, but *K*=¼ locally across the entire genome, and nowhere is *K*=½
- Therefore, between a parent and an offspring, the additive genetic covariance is $\frac{1}{2}V_A$, and the dominance covariance is 0
- If additive genetic effects are the only source of familial resemblance, then parent-offspring correlation is expected to be $\frac{1}{2}V_A/V_T$.
- Thus the narrow heritability h^2 can be estimated by $2r_{PO}$ (r_{PO} being the empirical parent-offspring correlation)
- However, parent-offspring correlation may arise from similarities in their environment. If this the the case, then $2r_{PO}$ would be an over-estimate of h^2.

Adoption studies

- If an offspring was adopted away from their biological parents at an early age and brought up by unrelated adoptive parents, then the possibility of shared environmental influences with the biological parents is much reduced.
- In an adoption study, heritability can be estimated by $2r_{POA}$, where r_{POA} is the empirical correlation between a biological parent and their adopted away offspring.
- Furthermore, the contribution of dominance to overall variance can be estimated by $4(r_{SA} - r_{POA})$, where r_{SA} is the empirical correlation between biological full sibs who have been brought up separately (**reared apart**).
- The contribution of shared environmental effects for "intact" parent/offspring pairs is estimated by $r_{PO} - r_{POA}$

Twin studies

- Twins provide another way of separating genetic effects from shared environmental influences
- This takes advantage of the existence of two types of twins: monozygotic (MZ) and dizygotic (DZ)
- MZ twins are developed from the same zygote, and therefore $K = ½$
- DZ twins are full siblings, and therefore $K = ¼$ and Prob(K=½) = ¼
- The crucial assumption is that shared environmental influences are equally important for MZ and DZ twins.
- Let r_{MZ} and r_{DZ} denote the empirical MZ and DZ correlations from a study.
- In the absence of dominance, h^2 can be estimated by $2(r_{MZ} – r_{DZ})$, then shared environmental variance is estimated by $r_{MZ} – h^2$
- If $r_{MZ} < h^2$, this indicates the presence of dominance

Quiz 2

- You have collected data on a quantitative trait in a sample of MZ and DZ twins.
- From the data you calculated MZ and DZ correlations to be 0.7 and 0.4 respectively
- What are the estimates of
  - Heritability
  - Proportion of variance explained by shared environmental influences

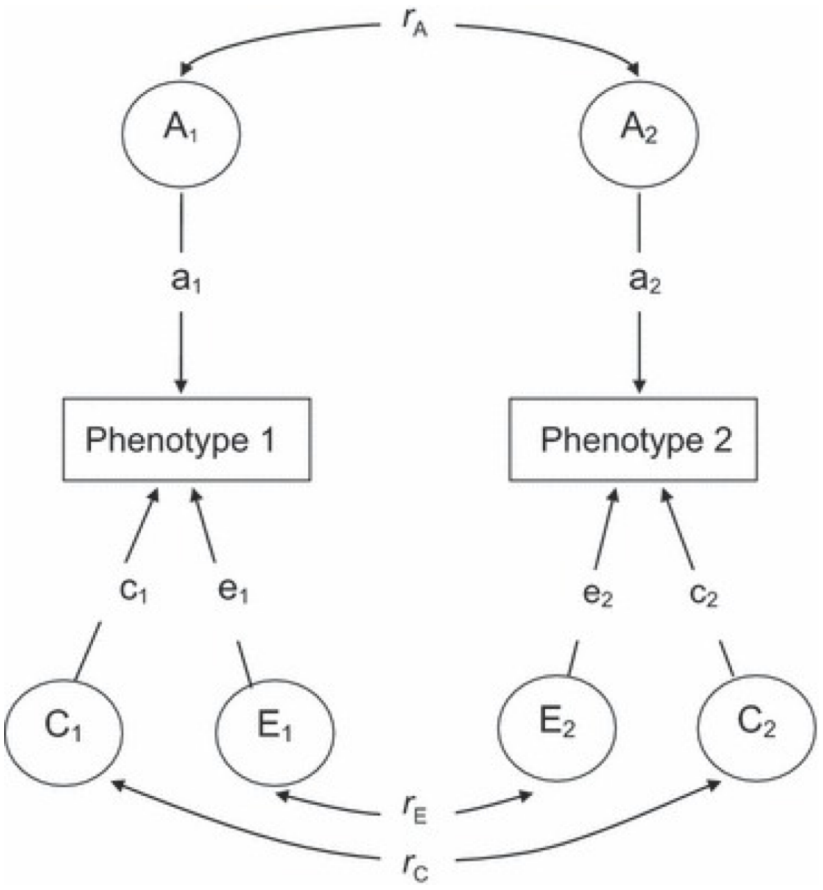
1. 0.7, 0.4
2. 0.6, 0.1
3. 0.4, 0.3

Quiz 2

- Estimate of Heritability = 2(0.7-0.4) = 0.6
- Estimate of proportion of variance explained by shared environmental influences = 0.7-0.6=0.1

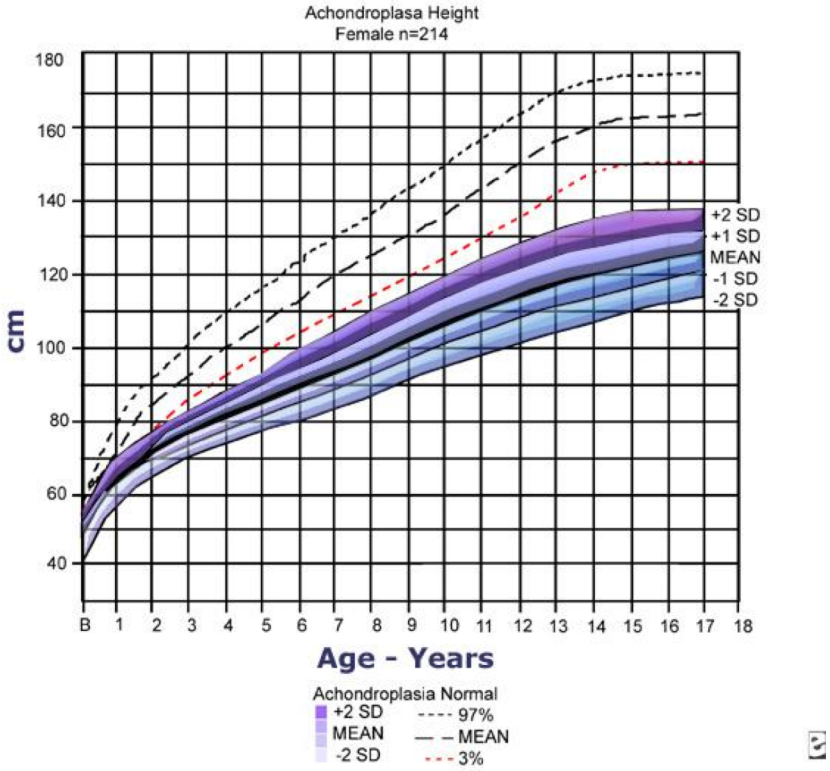
Multivariate twin studies

- Twins can also be used to estimate the genetic covariances between different traits.
- In the absence of dominance, the genetic covariance between 2 different traits (1 and 2) is estimated by $2(c_{12MZ} - c_{12DZ})$ where c_{12MZ} is the empirical cross-trait covariance between MZ twins and c_{12DZ} is the empirical cross-trait covariance between DZ twins.



Major locus effects

- Quantitative traits can be influenced by genetic mutations with very large effects (**major loci**) in addition to multiple genetic variants with small effects (**polygenes**)
- Adult males with achondroplasia have mean height of 52 inches, compared to the population adult male mean of 69 inches. This difference of 17 inches is almost 6 standard deviations of adult male height in the general population.
- Thus even the tallest adults with achondroplasia are seldom taller than the shortest adults without achondroplasia.



Height for females with achondroplasia (mean/standard deviation [SD]) compared to normal standard curves. The graph is based on information from 214 females. Adapted from Horton WA, Rotter JI, Rimoin DL, et al. Standard growth curves for achondroplasia. J Pediatr. 1978 Sep; 93(3): 435-8.

Liability-threshold model

- Liability is a quantitative trait determined by multiple genetic and environmental factors
- When liability exceeds threshold, disease develops
- Under this model, the heritability of the liability for a disease can be estimated from twin and adoption studies.

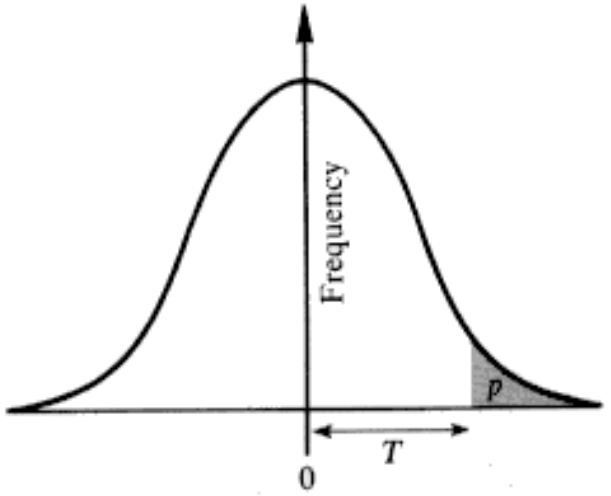


FIGURE 9.5
Threshold model. All individuals with a value of x greater than T are affected. The proportion of affected individuals is the area under the distribution curve beyond T.

POPULATION GENETICS

Genotype frequencies

Genotype frequency - the proportion of individuals in a population with a particular genotype.

For example: In a population of 1000 individuals, the number of individuals with the 3 possible genotypes at a SNP are: AA 500, AB 200, BB 300

The genotype frequencies are calculated as:

$p_{AA} = 500/1000 = 0.5$

$p_{AB} = 200/1000 = 0.2$

$p_{BB} = 300/100 = 0.3$

Note the genotype frequencies should sum to 1

Allele frequencies

Allele frequency – the proportion of alleles at a variable site in a population that is a particular type

Proportion of allele A in genotype AA = 1

Proportion of allele A in genotype AB = 1/2

Proportion of allele A in genotype BB = 0

The frequencies of alleles A and B thus

$p_A = p_{AA} + ½\ p_{AB}$

$p_B = p_{BB} + ½\ p_{AB}$

Note that allele frequencies should sum to 1

Hardy-Weinberg equilibrium

- Hardy-Weinberg equilibrium (HWE) describes SNP where the genotype frequencies are given by the terms of the binomial expansion

$$(p+q)^2 = p^2 + 2pq + q^2$$

- Where p and q are allele frequencies ($p+q=1$) and p^2, $2pq$ and q^2 are genotype frequencies
- This result comes from the "multiplicative law" of probability for independent events:

$$P(AA) = P(A)P(A) = p^2 \text{ for homozygous genotypes}$$

$$P(AB)=P(BA)=P(A)P(B) = pq \text{ for heterozygous genotypes}$$

- Hardy-Weinberg equilibrium holds when there is random mating in a large population, with respect to the SNP, unless the genotype data contain errors.

Quiz 3

An autosomal recessive disease has a frequency of 1 per 10,000 in the population. Assuming that the normal allele and the disease mutation are in Hardy-Weinberg Equilibrium, what is the frequency of heterozygous carriers in the population?

1. 2/10,000
2. 1/100
3. 2/100

Quiz 3

- Let the mutation frequency be p
- Under Hardy-Weinberg Equilibrium
- Homozygous mutant frequency $p^2 = 1/10000$
- Mutant allele frequency $p=\sqrt{(1/10,000)} = 1/100$
- Heterozygous frequency $2p(1-p) = 2 \times 1/100 \times 99/100 = 198 / 10,000 \sim 2\%$

Kinship coefficients

The kinship coefficient (*K*) between two individuals is defined as the probability that two alleles, one from each individual, drawn at random at any autosomal locus, will be identical-by-descent (IBD).

1. If two individuals (A and B) have kinship coefficient *K*, what is the kinship coefficient between A and the offspring of B, assuming that the other parent of this offspring is unrelated to A?

“Propagation” of kinship

At any genomic location, the offspring of B will have inherited 1 of the 2 DNA segments of B.

When a DNA segment is drawn at random from the offspring of B, there is a probability ½ that this is inherited from B, and probability ½ that this is inherited from the other parent.

If the segment is inherited from B, then there is probability *K* that it is IBD with a segment drawn from the corresponding genomic location from A.

If the segment is inherited from the other parent, then the probability is 0 because the other parent is unrelated to A.

Therefore the kinship coefficient between A and the offspring of B is ½ *K*.

Kinship: general formulae

- Applying the result for the propagation of kinship recursively, we see that each meiosis decreases kinship by a factor of ½.
- Since the kinship of a non-inbred individual with itself is ½, the Kinship coefficient between two individuals sharing one common ancestor is equal to $(½)^{g+1}$, where is the total number of meioses separating the 2 individuals.
- For example, half siblings share one parent and are separated by 2 meioses, therefore their kinship is $(1/2)^3=1/8$
- If 2 individuals are descended from two individuals who are full siblings, then since full siblings have kinship ¼, the kinship of the 2 individuals is $(½)^{g+2}$, where g is the total number of meioses separating the 2 individuals, downwards from the 2 siblings.
- For example, first cousins are children of full siblings, therefore their kinship is $(1/2)^4=1/16$

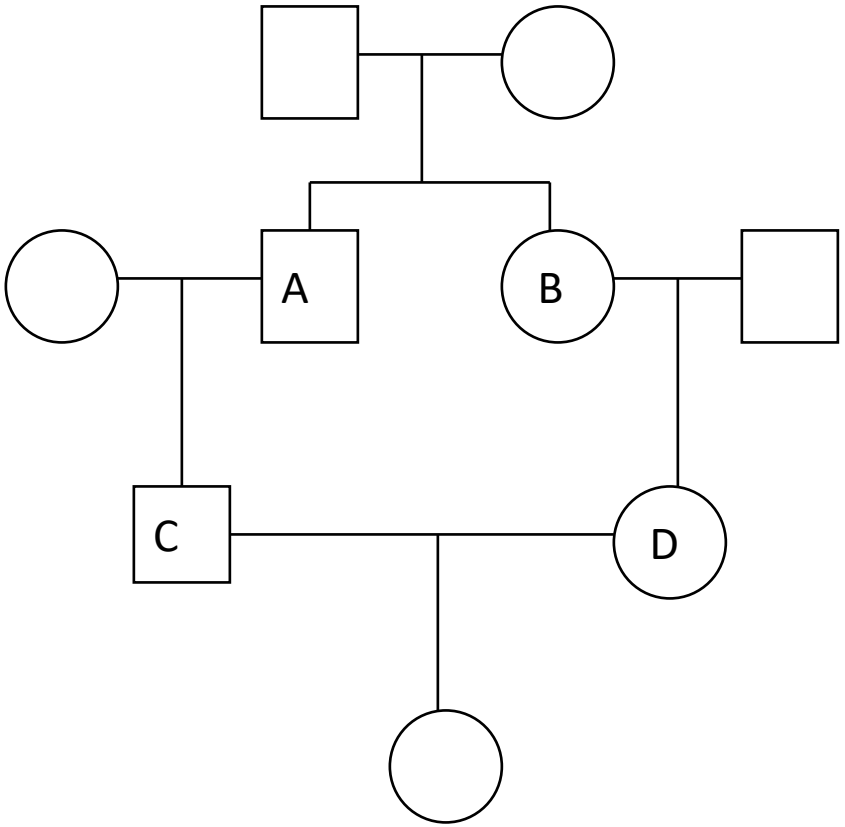
Inbreeding and recessive disease

Consanguineous marriages (usually between cousins) are encouraged in some cultures (mostly in Asia and Africa). These populations have an increased risk of recessive diseases.

Example

A newly-wed couple, who are first cousins, are worried that their future child will have a very high risk of cystic fibrosis (CF). Although the general incidence of CF in the population is only 1/1800, and they have no knowledge of whether there is CF or not in their family history, they have heard that the children of consanguineous marriages have increased risk of CF. What is your estimate of the risk of CF if they were to have a child?

Kinship coefficients



K(A,B) = 1/4

K(A,D) = 1/8

K(C,D) = 1/16

Overall disease risk

At any genomic location, the genotype of an offspring is the result of drawing a DNA segment at random from each of the 2 parents.

The probability that the 2 DNA segments in the offspring are IBD with each other is therefore equal to the kinship coefficient of the 2 parents. (This is also called as the **inbreeding coefficient** of the offspring).

The kinship coefficient between 2 first cousins is 1/16. Therefore there is a probability 1/16 that the 2 DNA segments containing the CF gene in the offspring will be IBD with each other.

In this case, the probability that the offspring will be homozygous for CF mutation will be the frequency of the CF mutation in the population (approximately 1/42, the square root of 1/1800).

If this is not the case, the probability that the offspring will be homozygous for CF mutation will be the population frequency of CF, i.e. 1/1800.

Therefore the total probability of CF in the offspring is
$1/42 \times 1/16 + 1/1800 \times 15/16 \sim 1/500$

Population differentiation

- The extent to which a population is homogeneous or diverse subpopulations with different allele frequencies is measured by the **fixation index** F_{ST}.
- If the i'th (of r) subpopulation accounts for a fraction f_i of the overall population and has allele frequency p_i at a single SNP, and the population as a whole has allele will have allele frequency p, then the fixation index is

$$F_{ST} = \frac{\sum_i f_i (p_i - p)^2}{p(1-p)}$$

- An overall F_{ST} index is obtained by taking the average of F_{ST} of all available SNPs
- When population differentiation is absent, $F_{ST}=0$. The larger the F_{ST}, the more diverse are the subpopulations.

Quiz 4

- A population is made up of 2 subpopulations with frequencies 0.4 and 0.6. A SNP has frequencies 0.1 and 0.2 in these 2 populations, respectively. What is the F_{ST} of the population at this SNP?

1. 0.0016
2. 0.0024
3. 0.1504

Quiz 4

- Overall allele frequency in population p = 0.4 x 0.1 + 0.6 x 0.2 = 0.16
- Numerator = 0.4 x $(0.1-0.16)^2$ + 0.6 x $(0.2-0.16)^2$ = 0.0024
- Denominator = 0.16 * (1-0.16) = 0.1504
- F_{ST} = 0.0024 / 0.1504 ~ 0.016

Natural selection on mutations

- Mutations that cause serious derangements in cellular functions, and result in non-viable embryo, early death or reduced fertility are less likely to be passed on to the next generation than other, less damaging mutations.
- Such mutations are thus under **negative selection** preventing them from become common in the population. Indeed, many will survive only a small number of generations becoming extinct. Extremely damaging mutations often become extinct in just one generation, so that when such mutations are detected, they are often ***de novo***.
- Mutations in genomic regions containing functionally important genes are likely to be damaging. Negative selection causes many such mutations to become extinct, reducing the genetic diversity of these regions. Negative selection thus also known as "**purifying selection**", and genes with low levels of diversity are inferred to be **mutation-intolerant**.

Selection on quantitative traits

- If a certain range of values of a quantitative trait is optimal for survival and reproduction under the prevailing environment and most individuals have trait values below this optimum, then there would be positive selection for variants that increase the trait, and negative selection for variants that decrease the trait.
- If the prevailing environment does not change over many generations, then the allele frequencies will evolve such that most individuals have trait values in the optimal range. At this stage, small effects of the trait, whether increasing or decreasing, will have little or no average effect on fitness.
- However, variants with very large effects in either direction may push the phenotype to outside its optimal range, and therefore be subjected to negative selection. This is known as **stabilizing selection**.
- The opposite of stabilizing selection is **disruptive selection**, which favours individuals with extreme phenotypes.

MOLECULAR GENETICS

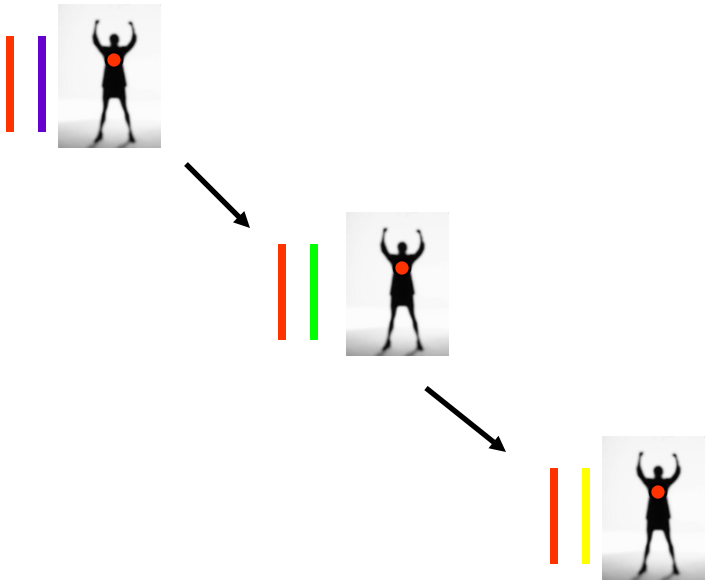
Measured genotypes

- Advances in molecular genetics made it possible for genetic variations to be discovered and measured throughout the genome (e.g. restriction fragment length polymorphisms, RFLPs)
- This meant that genotypes no longer need to be latent variables, but can be measured in individual subjects.
- Genotyping revolutionized genetics in the 1980, when numerous mutations responsible for Mendelian diseases were mapped to their chromosomal position, and the affected genes were subsequently identified.

Mapping Strategies

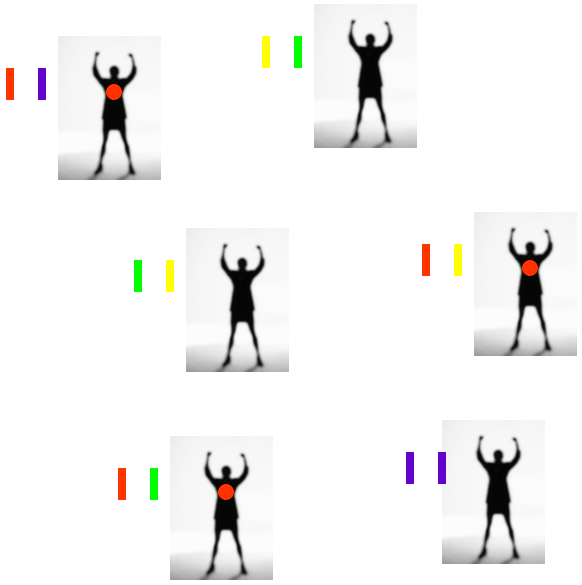
Linkage

- Co-segregation of a DNA segment with phenotype
- Long range



Association

- Correlation of a DNA sequence variant with phenotype
- Short range



Genetic linkage - two-locus transmission

- Given a heterozygous genotype Aa, the 2 possible haplotypes (A and a) are equally likely to be transmitted to an offspring (Mendel's first law)
- How about an individual heterozygous for two loci, AaBb, what are the probabilities of transmitted each of the 4 haplotypes AB, Ab, aB, ab?
- If segregation at the 2 loci are independent, then transmission probability of each haplotype is ½ x ½ = ¼.
- This is true when the loci are on different chromosomes (Mendel's second law), but not when they are on the same chromosome.
- Which two types will be more likely to be transmitted?

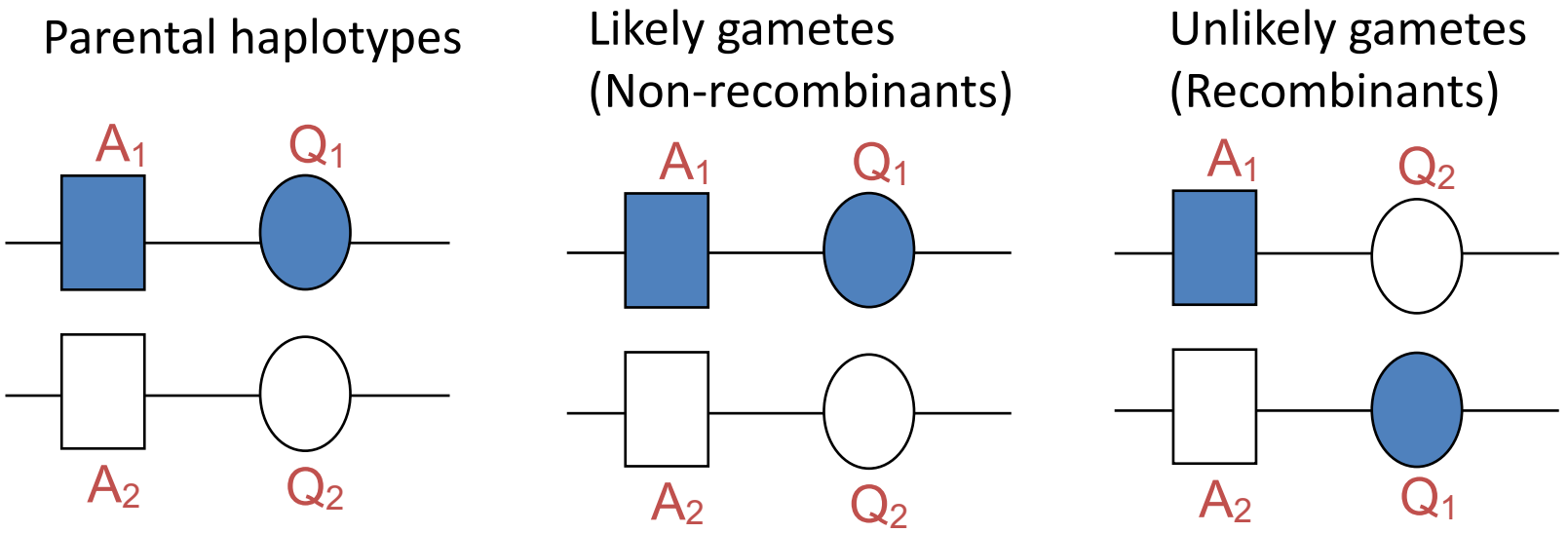
AaBb

Parent

AB Ab ab aB

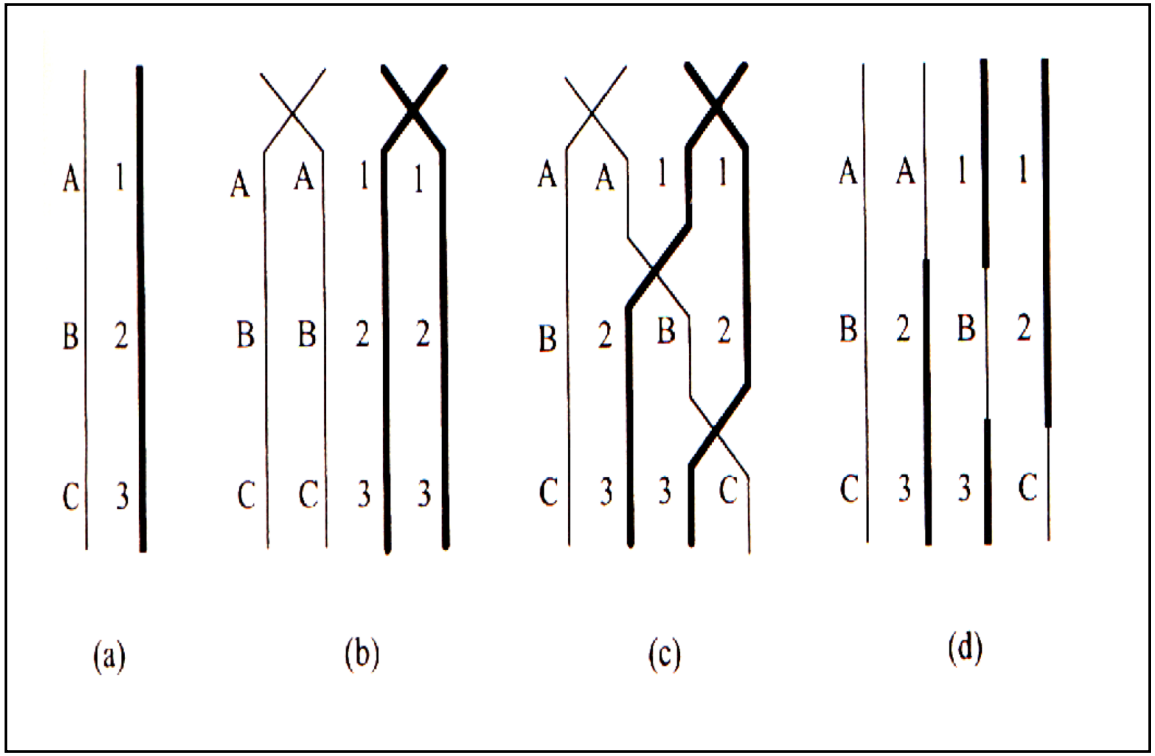
Gametes

Haplotypes and recombination



- Haplotype = set of alleles inherited from the same parent
- Alleles that were inherited together from the previous generation are more likely to be transmitted together to the next generation, if the loci are on the same chromosome
- Alleles which have different parental origins but are transmitted together in the same gamete are called "recombinant"
- The proportion of gametes of 2 loci that are recombinant is called the recombination fraction
- Two loci are "linked" if their recombination fraction is less than 1/2

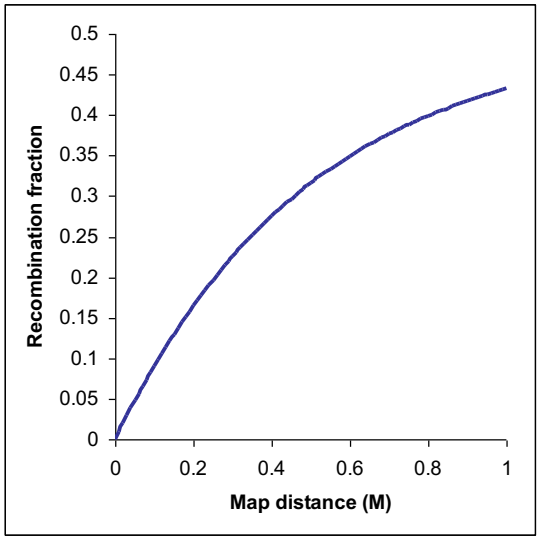
Crossovers during meiosis



- A chromosome inherited from a parent is usually not transmitted intact to a offspring
- Instead, crossovers between chromatids occur during meiosis, resulting in each transmitted chromosome being a hybrid of alternating segments of the paternal and maternal chromosomes

Map function

The genetic distance (Morgans) between two loci is the average number of cross-over events that occur between them per meiosis



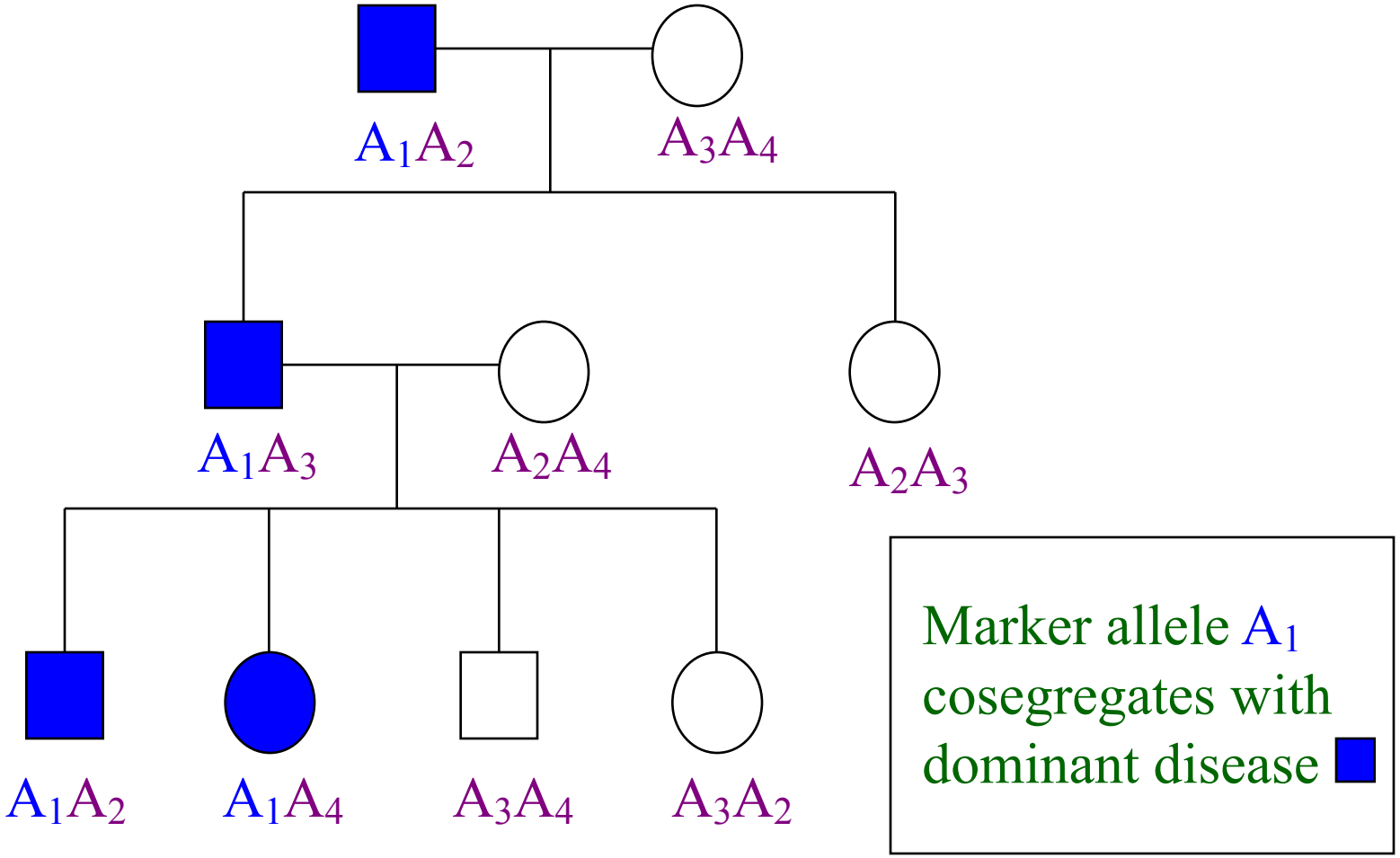
Haldane map function

$$\theta = \frac{1 - e^{-2m}}{2}$$

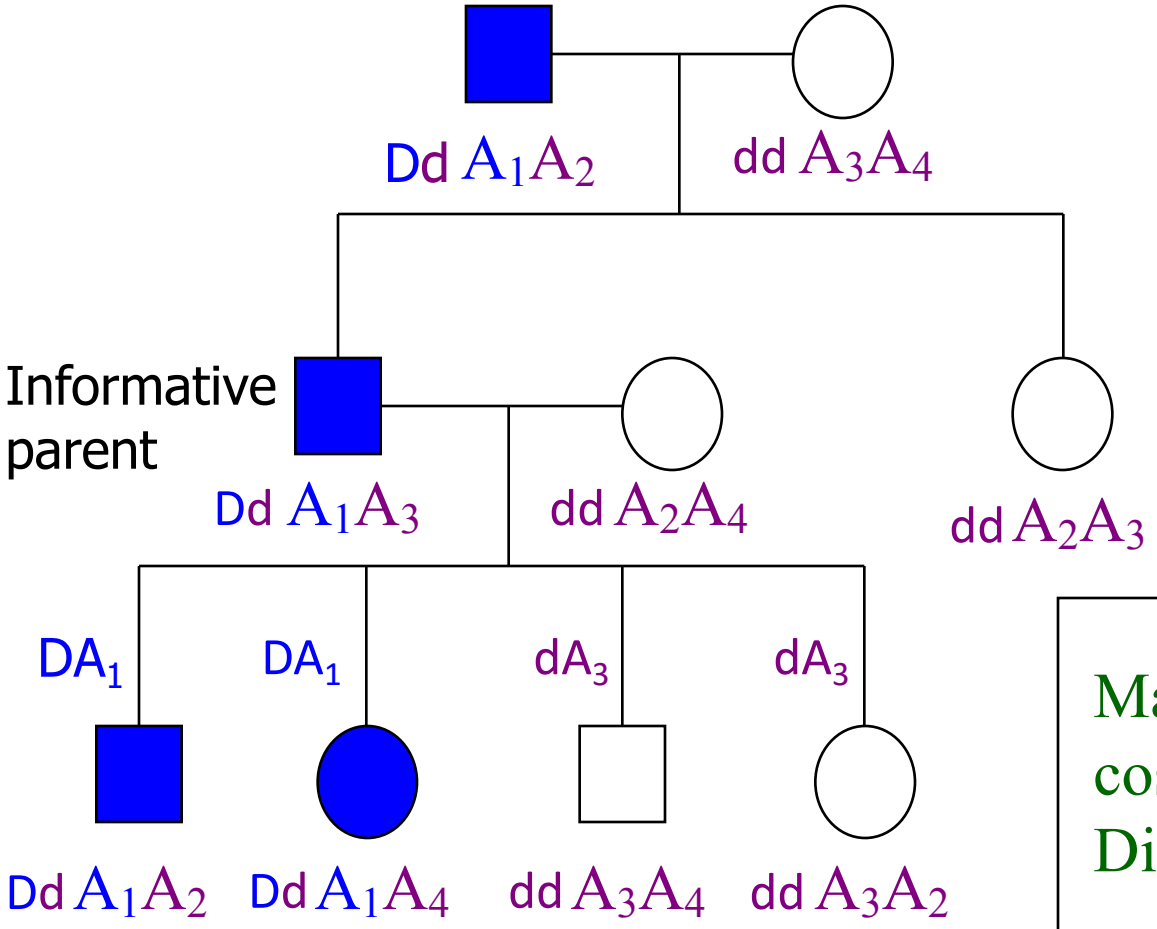
Assumes independence of cross-over points

Implication: genetic loci can be mapped relative to each other by examining the recombination fractions between them

Linkage Mapping



Linkage Mapping



Marker allele A_1 cosegregates with Disease mutation D

Recombination fraction: 0/4

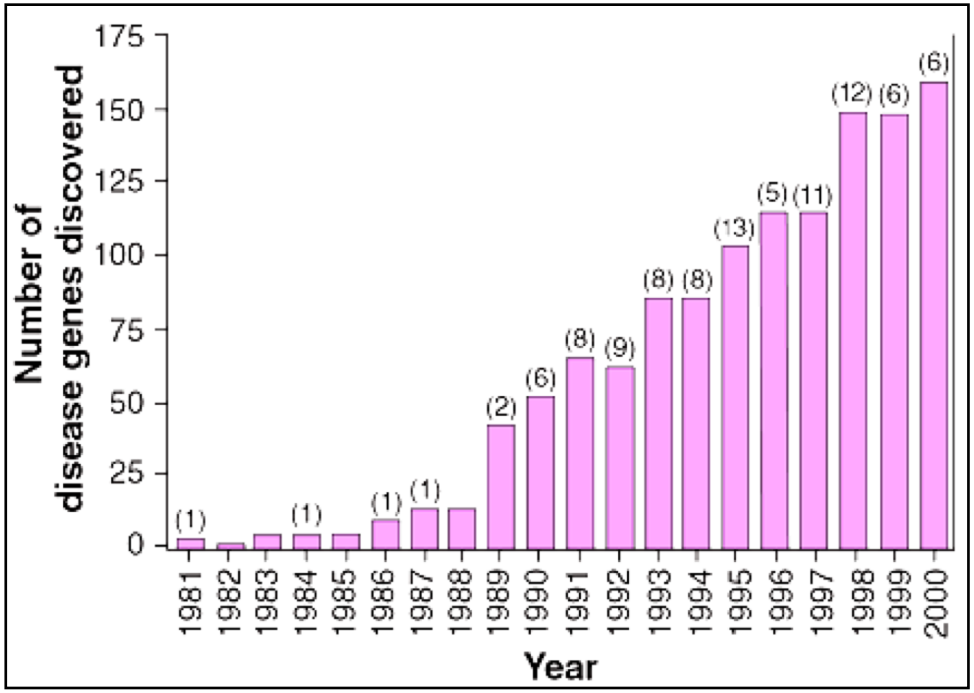
Parametric linkage analysis

- When recombinant and non-recombinants gametes can be inferred, linkage analysis involves simply counting the 2 types of gametes and calculating the fraction that are recombinant.
- In human data, it is often not possible to infer the recombination status of gametes with certainty. Thus, it is necessary to adopt likelihood inference, where the probabilities (called likelihoods) of the observed data under different values of the recombination fraction (θ) are calculated.
- The value of θ that maximizes the likelihood provides an estimate of the true recombination fraction.
- Whether θ differs significantly from 0.5 is evaluated by a lod score, defined as

$$lod(\theta = t) = \log_{10}\left(\frac{L(\theta = t)}{L(\theta = 0.5)}\right)$$

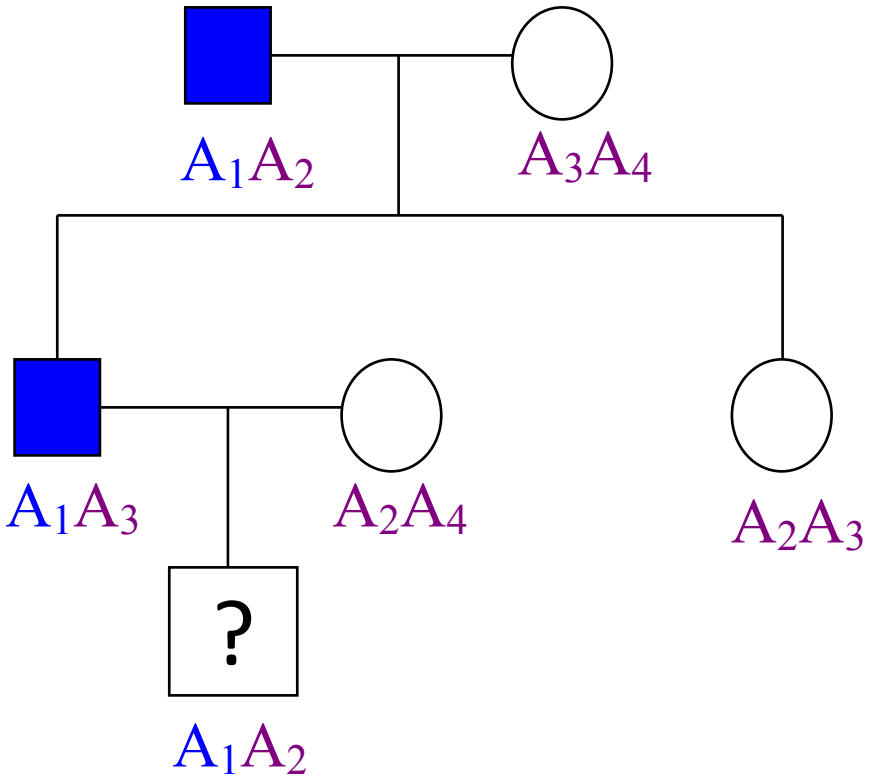
- Linkage is considered significant when lod > 3 for some value of θ.

Disease Gene Discovery: 1981 to 2000



- **1112 disease genes discovered**
- Numbers in parentheses indicate disease-related genes that are polymorphisms ("susceptibility genes").

Quiz 5



Autosomal dominant disorder:

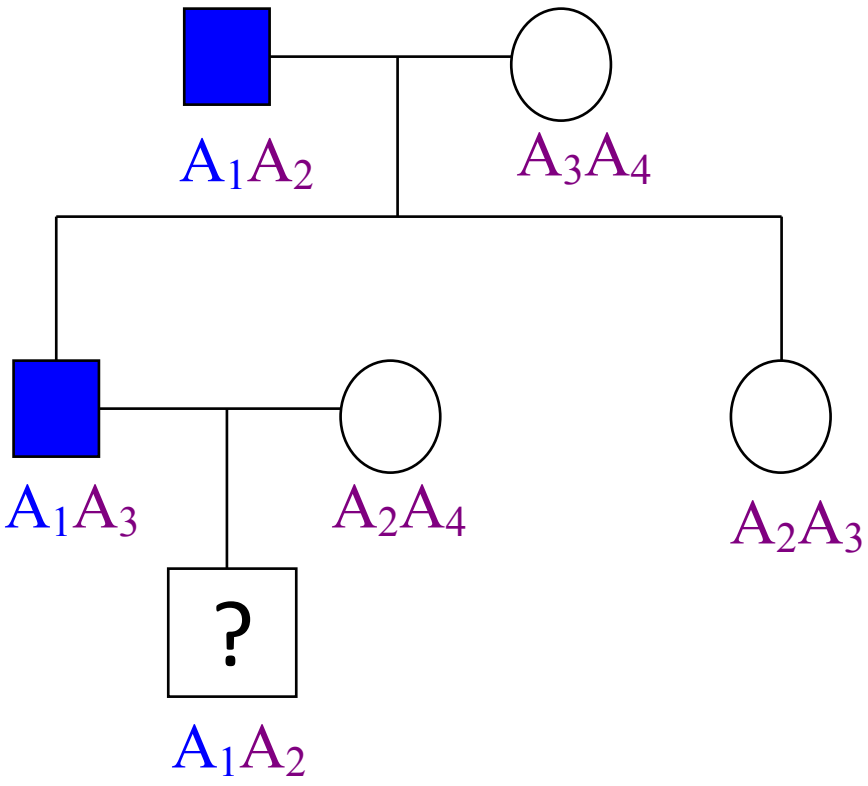
(1) the disease locus has been mapped to be linked to a marker locus with recombination fraction 0.01

(2) the genotypes at that linked locus of family members are as shown in the pedigree diagram

What is the risk of the queried individual?

1. 1.00
2. 0.99
3. 0.01

Quiz 5



(1) The gamete transmitted from the grandfather to the father has haplotype DA_1. Therefore the D mutation and the A_1 allele are on the same chromosome in the father

(2) Since the A1 allele is transmitted from the father to the son, the D mutation is also transmitted to the son unless there is a recombination between the A_1 and D, with probability 0.01

(3) The risk of disease in the son is therefore 1-0.01 = 0.99

Association

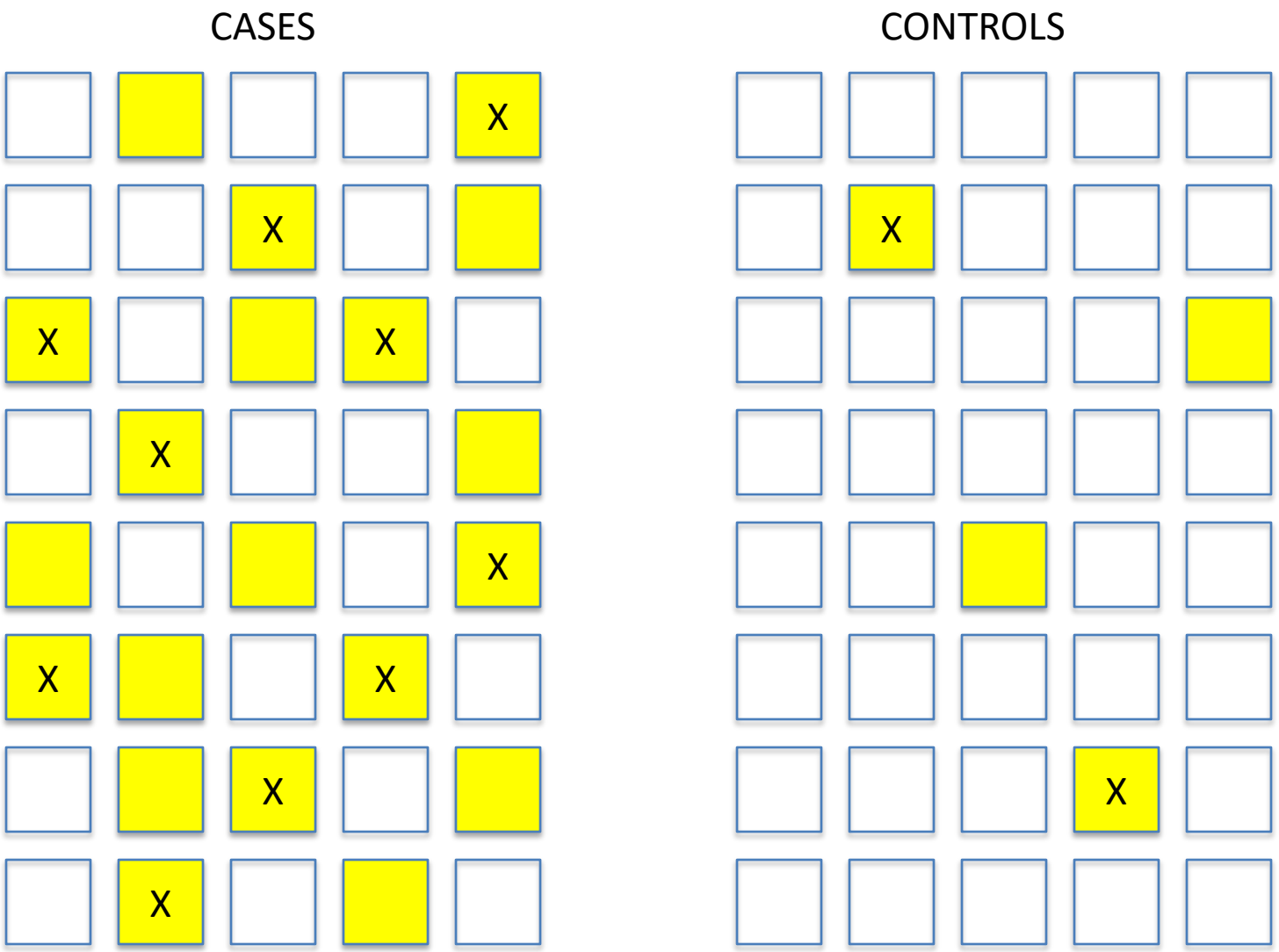
- Association refers to a correlation between genotype and phenotype, that individuals with different genotypes tend to differ in phenotype
- For continuous traits, different genotypes are associated with different phenotypic means (and possibly also different phenotypic variances)
- For disease traits, different genotypes are associated with different disease risks.
- The phenotype effect of a genotype, relative to a reference genotype, is usually measured by a risk ratio.
- An alternative measure of genotypic effect is the odds ratio, where odds = 1-risk

Genotype	Risk	Risk Ratio	Odds	Odds Ratio
BB	0.09	9	0.0989	9.791
AB	0.03	3	0.0309	3.062
AA	0.01	(1)	0.0101	(1)

Possible explanations for association

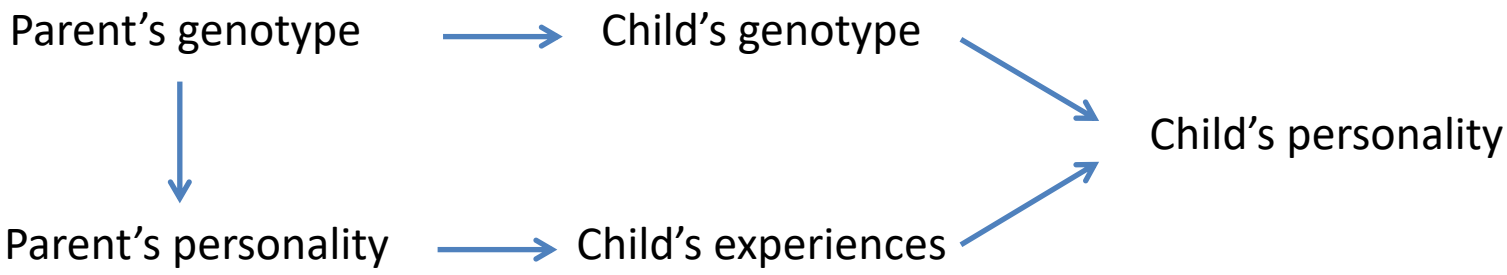
- Causation: Genotype -> Phenotype
- Reverse causation: Phenotype -> Genotype, can be excluded for inherited genotypes (but not somatic mutations)
- Confounding
  - Ancestry: a subpopulation may have higher frequency for both the genotype and the phenotype than the overall population
  - Correlation of the genotype with another genotype (i.e. linkage disequilibrium) which is causally associated with the phenotype
  - Correlation of the genotype with an environmental influence on the phenotype (e.g. parental characteristics)
- Collider bias
  - Genotype and Phenotype both influenced the probability of a individual to be included in the study

Population stratification



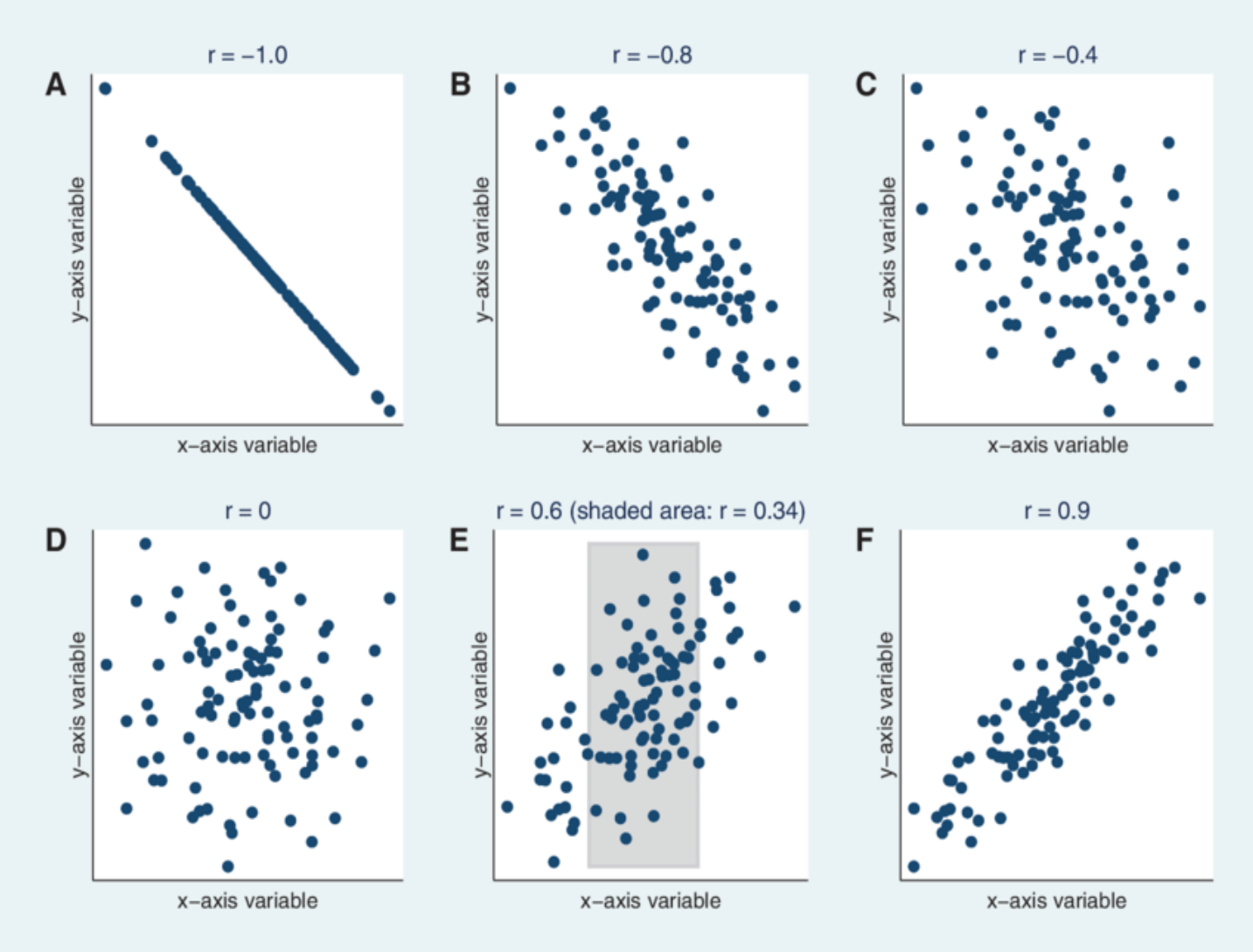
Gene-environment correlation

- A parent can influence a child's phenotype by
  - Transmitting a genetic variant that affects the child's phenotype
  - Altering an environment factor that affects the child's phenotype

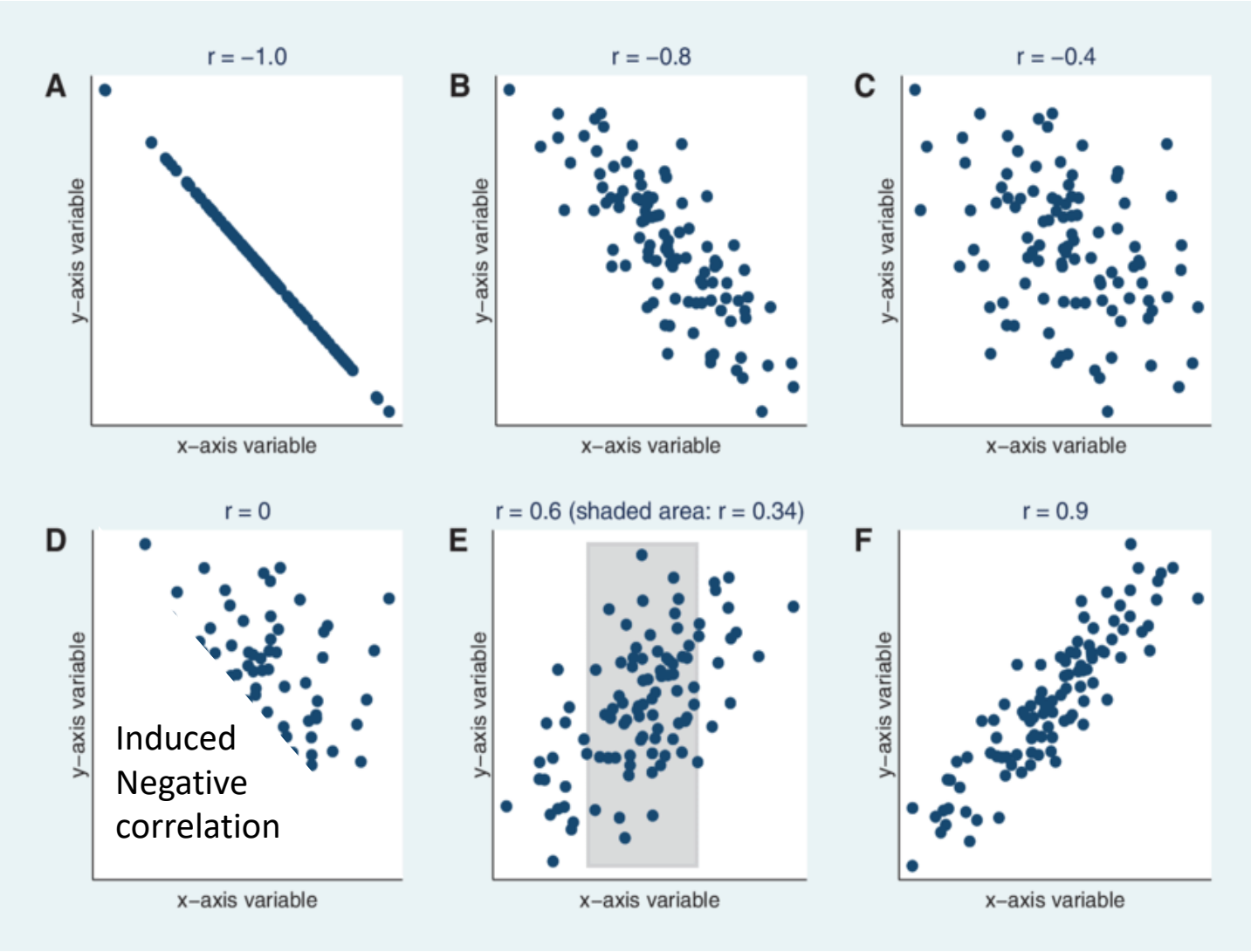


- Even when the child's genotype does not directly influence the child's personality, there can be an association between the two because of the correlation between the child's genotype and experiences

Selection bias from one variable



Collider bias

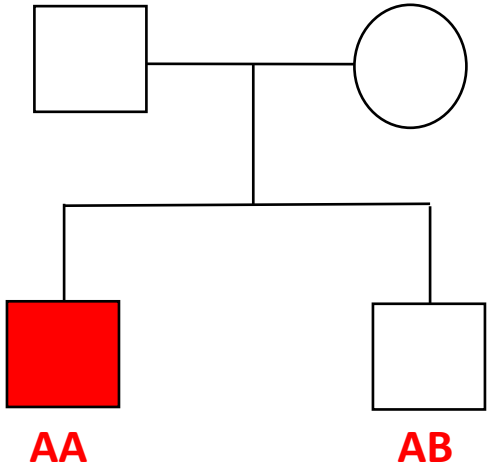


Study designs for association

- Prospective / Cohort studies
  - The sample is recruited from the population regardless of the phenotype.
  - Data on genotype(s), phenotype(s) and confounder(s) are collected
  - Large sample is needed if the phenotype is a rare condition
  - Efficient for studying multiple phenotypes and for gene-environment correlations / interactions
- Retrospective / Case-control studies
  - The sample is recruited according to phenotype
    - Affected individuals are recruited as cases
    - Unaffected individual are recruited as controls
  - Data on genotype(s) and confounder(s) are collected
  - Efficient if the phenotype is a rare condition

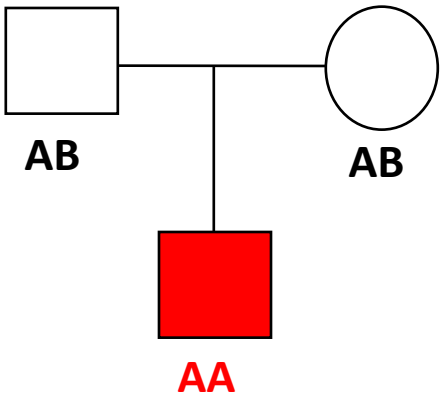
Family-based association studies

- Instead of correlating genotype and phenotype in apparently unrelated individuals, association can also be revealed correlating genotype and phenotype of individuals within a family.
- **Discordant sib-pair design**
  - Recruit sibling pairs where only 1 sib is affected
  - Genotype both siblings
  - Perform association test between disease and genotype for matched-samples



Case-parent trios design

- Recruit affected patients and their 2 parents (trios)
- Genotype all 3 subjects in each family
- Identify the un-transmitted alleles of the parents and regard these as a "**pseudo-control**"
- Perform association test between cases and pseudo-controls
- Alternative test: consider the other 3 possible offspring genotypes of the patents as "**counter-factual genotypes**"



Pseudo-control: **BB**

Counter-factuals: **AB**, **BA**, **BB**

Trios designs: pros and cons

Advantages

- Robust to false positive association from population stratification
- Parents may be easier to recruit than unrelated controls, for childhood diseases
- Able to examine
  - *de novo* mutations: variants present in an individual but absent in the two parents
  - parent-of origin effects: differences in the effect of a variant depending on whether it is inherited from the father or the mothertal
  - genomic nurture: parental genotype affecting offspring phenotype through the offspring's environment

Disadvantages

- Parents may be difficult to recruit for late-onset diseases
- Less statistical power for same number of subjects genotyped (3 subjects in TDT is roughly equivalent to 2 subjects in balanced Case-Control study)
- More prone to false positive association from genotyping errors.

LINKAGE DISEQUILIBRIUM

Population haplotype frequencies

	B	b	
A	pr	ps	p
a	qr	qs	q
	r	s	

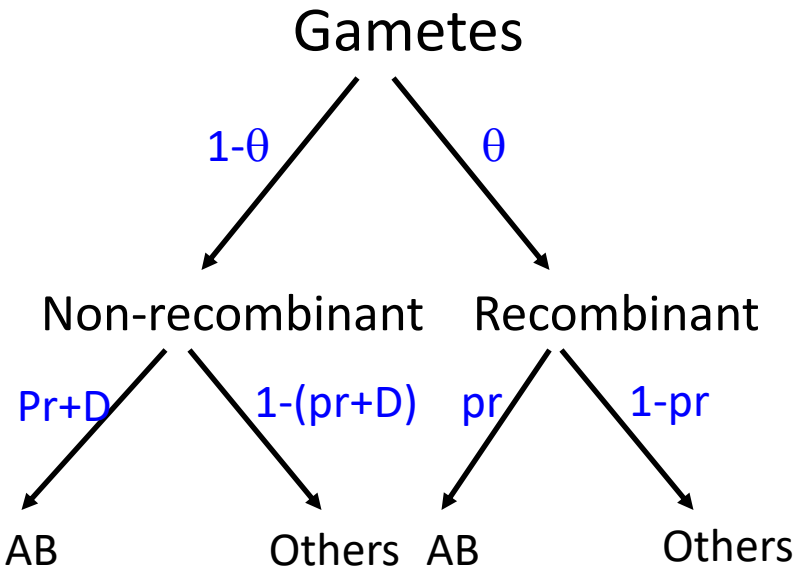
- If there is no association between alleles of the two loci, then the frequency of each haplotype is equal to the product of the frequencies of its constituent alleles
- Two loci with such haplotype frequencies are said to be in **linkage equilibrium**

Linkage Disequilibrium

	B	b	
A	pr+D	ps-D	p
a	qr-D	qs+D	q
	r	s	

- Deviation of haplotype frequencies from the product of constituent allele frequencies is called **linkage disequilibrium** (LD)
- The deviation D is a measure of linkage disequilibrium

Decay of D through recombination



- If haplotype AB has frequency pr+D in the present generation, then its frequency in the next generation is $(1-\theta)(pr+d)+\theta pr = pr+(1-\theta)D$
- Thus, D decays by a factor of $(1-\theta)$ per generation towards 0
- For tightly linked loci, LD will be maintained for many degenerations.

Linkage Disequilibrium: D’

	B	b	
A	h	p-h	p
a	r-h	1+h-p-r	q
	r	s	

- D can be normalized to D’, such that it is bounded between 0 and 1
- $D = D/D_m$
- The upper bound of h is determined by the constraints: p-h≥0 and r-h≥0; giving $h \leqslant \text{Min}(p,r)$
- The lower bound of h is determined by the constraints: 1+h-p-r≥0 and h≥0, giving $h \geqslant \text{Max}(0,p+r-1)$
- When D>0, $D_m = \text{Min}(p,r)-pr$
- When D<0, $D_m = \text{Max}(0,p+r-1)-pr$
- D’=1 when at least 1 of the 4 haplotypes is absent

Linkage Disequilibrium, r^2

	B	b	
A	h	p-h	p
a	r-h	1+h-p-r	q
	r	s	

- D can also be normalized to r^2
- $r = D/\sqrt{(pqrs)}$
- $r^2 = D^2/pqrs$
- r^2 represents the squared corelation between the two haplotypes coded numerically.
- An r^2 of 1 requires that 2 of the 4 haplotypes are absent, implying that the 2 loci have equal allele frequencies.

Quiz 6

	B	b	
A	0.1	0	0.1
a	0.3	0.6	0.9
	0.4	0.6	

For the haplotype frequencies in the table, what are the values of D' and r^2?

1. -1, 0.17
2. 1, 0.17
3. 1, 1

Quiz 6

	B	b	
A	0.1	0	0.1
a	0.3	0.6	0.9
	0.4	0.6	

From haplotype AB

D = 0.1-0.1x0.4 = 0.06

D_m = Min(0.1,0.4)-0.1x0.4 = 0.06

D' = 0.06/0.06 = 1

r^2 = 0.06^2/(0.1x0.9x0.6x0.4) ~ 0.17

Working from haplotype Ab

D = 0-0.1x0.6 = -0.06

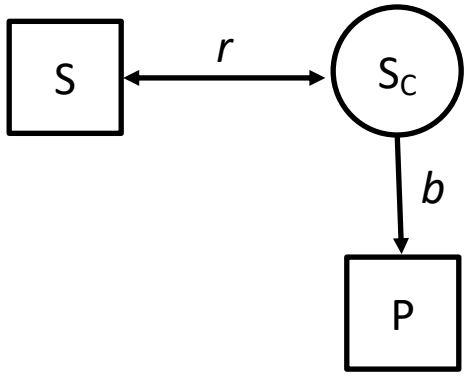
D_m = Max(0,0.1+0.6-1)-0.1x0.6 = -0.06

D' = -0.06/-0.06 = 1

r^2 = (-0.06)^2/(0.1x0.9x0.6x0.4) ~ 0.17

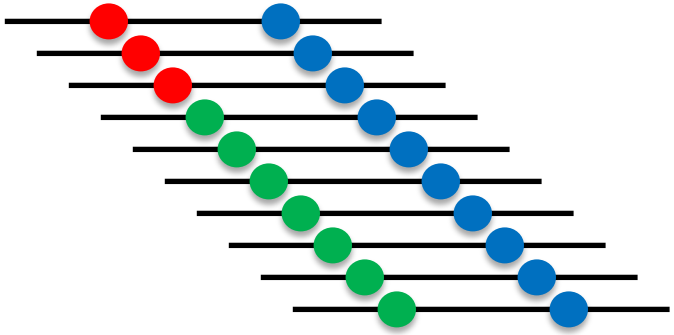
LD and indirect association

- A SNP (S) has no direct causal effect on a phenotype (P), but is in LD with a SNP (S_C) that has a direct causal effect on the phenotype.



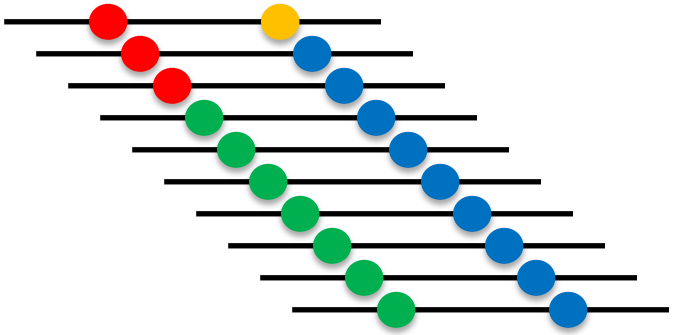
- Since S is correlated with P only through its correlation with S_C, its correlation with P is the correlation between S_C and P, b, attenuated by the correlation between it and S_C, *r*. It is thus *rb*
- The expected value of the chi-squared association test statistic of S is thus the expected chi-squared association test statistic of S_C, attenuated by r^2

Evolution of LD



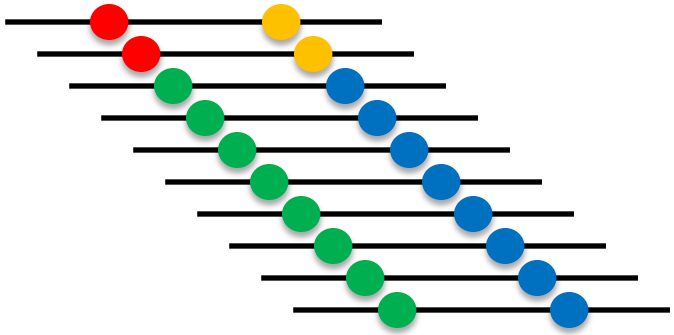
2 haplotypes

Mutation



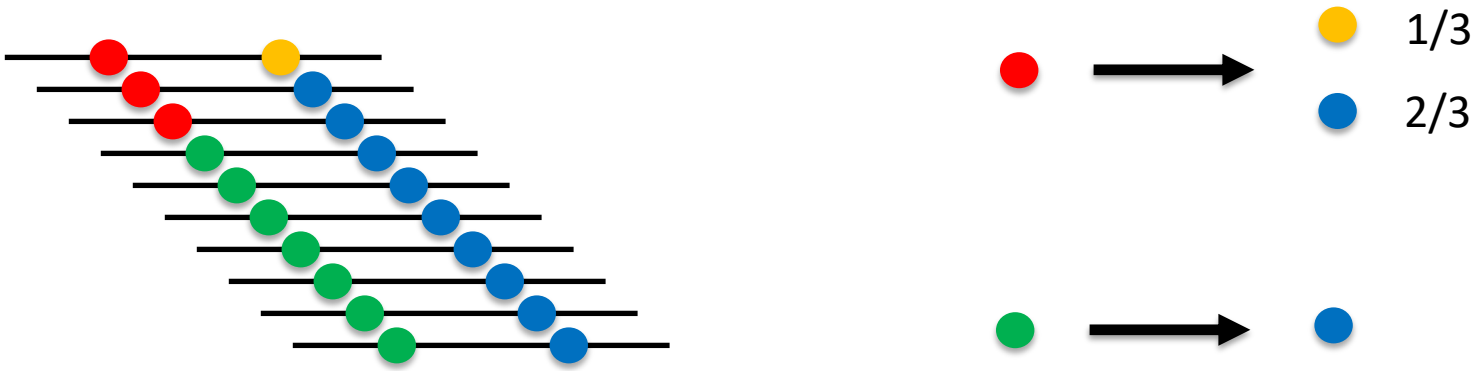
3 haplotypes, D'=1

Drift



2 haplotypes, D'=1, r^2=1

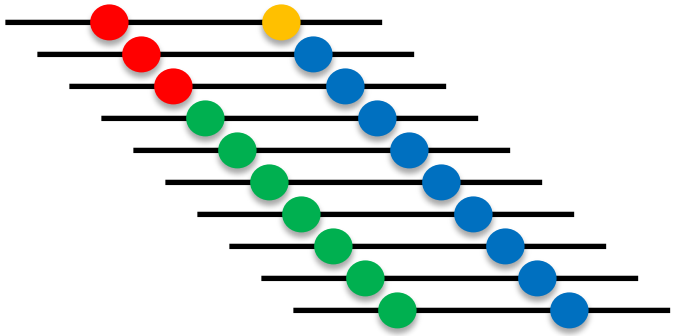
Genotype imputation



3 haplotypes, D'=1

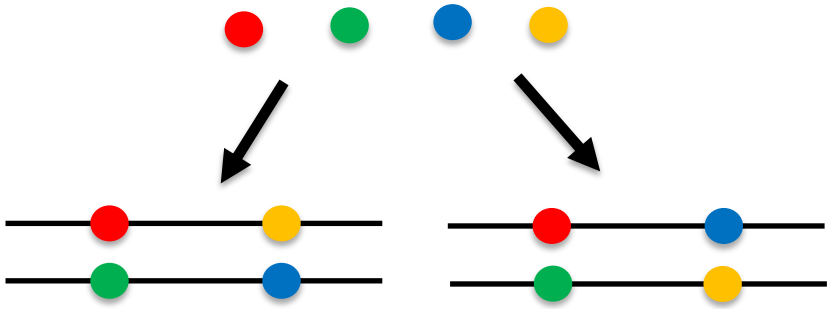
- Knowing the haplotypes present in the population, and their frequencies, allows imputation of untyped loci
- Imputation accuracy is increased when some haplotypes are absent in the population
- With increasing number of loci, more haplotypes are likely to be absent from the population. Thus, considering longer haplotypes should allow more accurate imputation.

Genotype phasing



3 haplotypes, D'=1

Unphased genotype



Excluded

- Phasing means uncovering the 2 haplotypes that make up a genotyping
- Knowing the haplotypes present in the population, and their frequencies, allows phasing of multi-locus genotype
- The absence of possible haplotypes in the population helps phasing
- Having parental genotypes also helps phasing

GENOME-WIDE ASSOCIATION STUDIES

The genomic era

- After the release of the first draft of the human genome sequence by Human Genome Project in 2000, genetic research entered the genomic (or post-genomic) era.
- A SNP Consortium was established in 2001 to catalogue single nucleotide polymorphisms (SNPs) in the human genome.
- Phase 1 of International HapMap Project was completed in 2005, which established the LD structure of over 1 millions SNP throughout the genome in from 3 diverse populations (Caucasians, Africans and East Asians). This HapMap Project was succeeded by the 1000G Project.
- The first SNP genotyping array (HuSNP) was developed by Affymetrix in 1998, with 1,494 SNPs. Subsequent SNP genotyping arrays from Affymetrix and Illumina progressively included more SNPs, from 10,000, to 50,000, 100,000, 500,000 to >1,000,000, designed to maximize SNP coverage using LD data from HapMap or 1000G.

Genome-wide association studies (GWAS)

- The availability of SNP arrays allowed association studies to systematically search for disease-associated SNPs in the entire genome.
- The first successful GWAS was published in 2002, on myocardial infarction. A subsequent GWAS published in 2005 implicated Complement Factor H for age-related macular degeneration.
- Software PLINK was published in 2007, and became a popular tool for the analysis of GWAS data.
- By September 2009, the GWAS Catalogue of NHGRI-EBI contained over 157,000 significant reported associations, from 4,220 publications.

GWAS – general considerations

- GWAS is simply association analysis carried out systematically on an enormous scale
- All the principles for traditional association analysis also apply to GWAS
- However, the very large scale of GWAS raises study design and data analysis issues that require special attention

Data quality control (QC)

- Systematic data **quality control** (QC) became possible and crucial
  - DNA degradation
  - Sample mix-up (sex check)
  - Duplicated samples
  - Cross-contamination
  - Genetic relatedness
  - Population stratification
  - Genetic outliers
  - Batch effects
- No dataset is completely clean. Inadequate data QC leads to misleading results.

SNP-based tests

- Single SNP association tests form the first line of GWAS analysis
  - Simple and **fast** to perform
  - Results can be visualized using Informative **graphical displays**
  - Summary SNP-based statistics suitable for **meta-analysis**
  - Many advanced statistical methods can be modified to use **summary statistics** from single-SNP association analysis instead of raw genotype and phenotype data

Multiple testing adjustment

- The very large number of SNPs tested in GWAS means that by chance alone many SNPs would be significant at traditional p-value threshold such as 0.05
- In order to control the number of false positive, the p-value threshold for significance is commonly the very stringent level of $5x10^{-8}$. This level is known as "**genome-wide significance**".
- Roughly speaking, only one genome scan in 20 is expected to produce a false positive association signal that reaches genome-wide significance.

Sample size

- Large sample size (thousands of cases and controls) is required for GWAS to detect even a small proportion of the risk alleles of common diseases. To capture the majority the risk alleles, the sample size needs to be at least 2 orders of magnitude larger.
  - Stringent significance level
  - Polygenic inheritance (large number of loci each with small effect)
- To maximize sample size, researchers have tended to simplify the criteria for the recruitment of cases and controls, and to not perform detailed clinical assessment of the subjects. This could increase the clinical heterogeneity of the subjects, although clinical heterogeneity does not necessarily reflect aetiological heterogeneity.
- The need for very large sample sizes has motivated the formation of **international consortia** to facilitate data sharing and pooling.

Replication

- When association is tested on many risk alleles, their effect sizes will be over-estimated for some risk alleles and under-estimated for others, purely by chance. For a study will limited sample size such that only a few of the risk alleles reach genome-wide significance, these risk alleles are likely to be the ones whose effect sizes have been most over-estimated. This is known as "**winner's curse**".
- When the same risk alleles are tested in an independent sample of similar size, their effect sizes are unlikely to be as over-estimated as in the original study. Thus they are unlikely to be replicated.
- Many consortia do not try to replicate risk variants using independent samples. Instead, new samples are continuously added to existing samples, and GWAS is repeated when the cumulative sample size has increased substantially. Genome-wide significant risk variants that become even more significant as sample size increases are likely to represent true associations.

Set-based tests

- One of the main purposes of GWAS is the delineate the biological functions that are deranged in disease. Genes are the functional units of the genome, and many biological processes / pathways require the participation of multiple genes.
- If a subset of genetic variants can be assigned to each gene, and a subset of genes assigned to each pathway, then genes and pathways can be tested for association with phenotype using subsets of variants.
- Many **gene-based** and **pathway-based association tests** have been proposed.

Multi-omics integration

- Being a systematic search for association, GWAS is entirely based on statistical evidence and does not require prior knowledge of the disease.
- Nevertheless, incorporating other types omics data to GWAS may enable more powerful and insightful analyses.
  - Incorporating **eQTL** data may enable gene-expression levels to be imputed from SNP genotypes, and to systematically identify genes whose expression level is associated with the phenotype.
  - Incorporating eQTL and **drug-induced gene expression changes** data may identity drugs which may reverse the gene expression changes associated with disease as candidates for drug repurposing
  - Incorporating **tissue-specific epigenomic data** (e.g. ChIPseq data) may enable the identify active regulatory elements in disease-relevant tissues and to prioritize GWAS significant SNPs by whether they could be alter the function of these elements.

Genetic architecture

- GWAS data can also be used to answer questions about the "genetic architecture" of a phenotype:
  - **SNP heritability**: what proportion of phenotypic variation is explained by common variants
  - **Polygenicity**: how many risk variants are involved in the phenotype
  - **Allele frequency spectrum** of risk variants, which may reflect past **selective pressures** on the phenotype
  - **Genetic correlations** with other phenotypes

Prediction

- Polygenicity implies that disease occurrence is a function of multiple risk variants as well as environmental factors.
- Classical and population genetics suggest that the effects of risk variants are largely additive on the liability scale. Thus it is natural to add up the effects of multiple risk factors into a single "**polygenic risk scores**" (PRS).
- GWAS effect size estimates provide the weights for the risk variants for PRS calculation, after pruning SNPs that are unlikely to make an independent contribution to risk (low significance level or high LD with a stronger risk variant)
- The predictive accuracy the of PRS improves with increasing GWAS sample size, but only to the extent that common risk variants determine the overall disease liability (i.e. SNP heritability)

Causal inference

- Causal inference are crucial formulating effective interventions
- Causal inference is most robust when an intervention study (randomized controlled trial) can be performed.
- Mendelian segregation is a random process which allocates alternative variants to offspring.
- A genetic variant that has a strong effect on a putative risk factor for a disease can be can be considered as an “intervention” to infer causation between the risk factor and the disease. Statistically, the genetic variant is used as an “instrumental variable”. A body of methodology called “**Mendelian randomization**” (MR) has been developed for this purpose.
- GWAS data provide multiple candidate instrumental variables for MR studies.
- Other than MR, other methods for causal modelling are being applied to GWAS data.

Limitations of GWAS

- Inability to characterize **rare variants** (although genotype imputation with large reference panels are increasingly able to characterize low frequency variants).
- Limited exploration of **gene-gene** and **gene-environment interactions**
- Insufficiently detailed **phenotypic data** on GWAS samples, e.g. risk factor and symptoms profile, biomarkers, comorbidity, treatment response, complications, course and outcome.
- Not enough studies / samples from **non-European populations**

THANK YOU