Classical Genetics

Pak Sham

16th August 2017

Outline

- Classical genetics: definition and relevance
- Segregation analysis
- Kinship coefficients
- Family, twin and adoption studies
- Estimating heritability
- Estimating genetic covariance
- (Linkage and association: excluded)

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.
- Includes
  - Segregation analysis
  - Heritability analysis
  - Classical linkage analysis
  - Classical association analysis

Relevance to modern research

- Genotypes no longer need to be latent!
  - Modern molecular technologies can perform genotyping / sequencing very efficiently
  - Genomes of many species have been sequenced
- Nevertheless the principles of classical genetics are still useful for:
  - Assessing whether a phenotype is a Mendelian trait (SEGREGATION)
  - Estimating the overall genetic contribution to a phenotype (HERITABILITY)
  - Estimating the overall shared genetic contributions to two correlated phenotypes (GENETIC COVARIANCE)
  - Identifying the genes that influence the values of a phenotype (MAPPING: LINKAGE, ASSOCIATION)

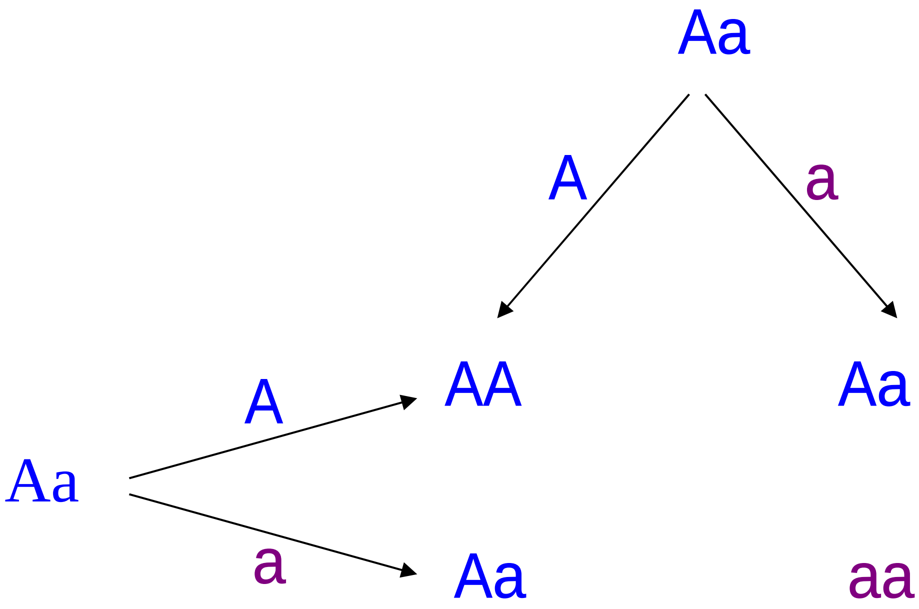
Mendelian traits

- What is a Mendelian trait?
  - Determined by the genotype at a single locus
- Example:
  - Locus with 2 alleles: 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”)
- Typical research questions
  - Where / Which is the gene ? (Mapping)
  - How does the mutation lead to the disease ? (Functional analysis)

Segregation analysis

- How do we find out if a trait is inherited as a Mendelian trait, before trying to map it?
- 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

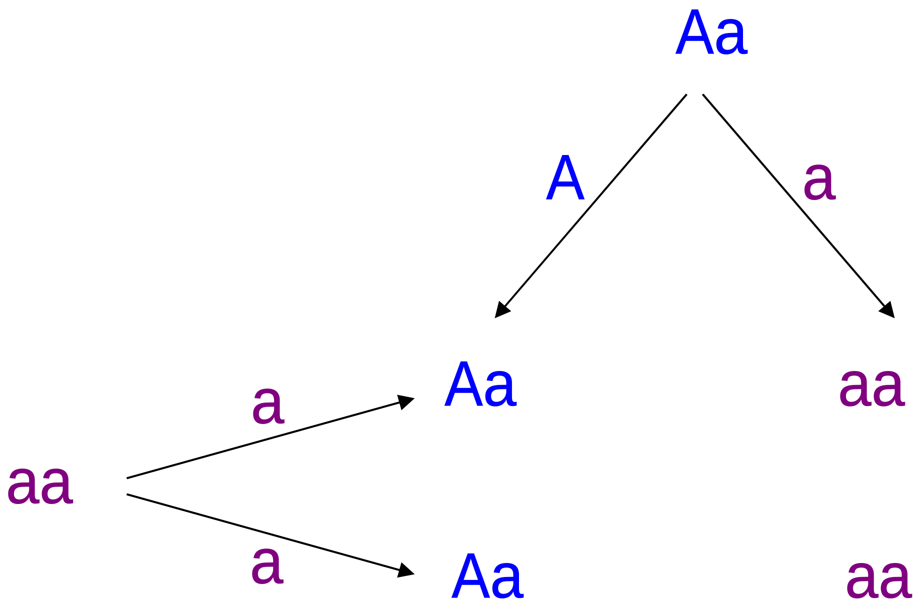


Aa

Aa	A	a
A	AA	Aa
a	Aa	aa

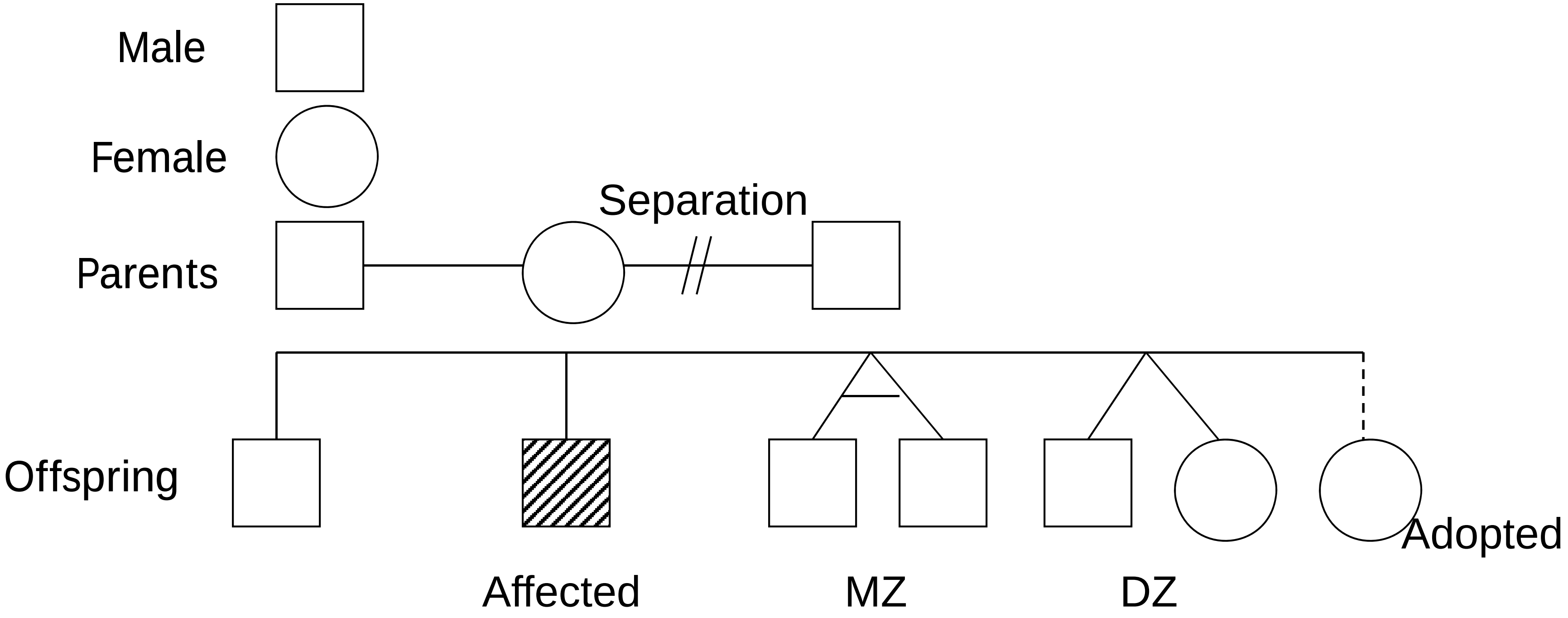
3:1 Segregation ratio

Mendelian backcross



1:1 Segregation ratio

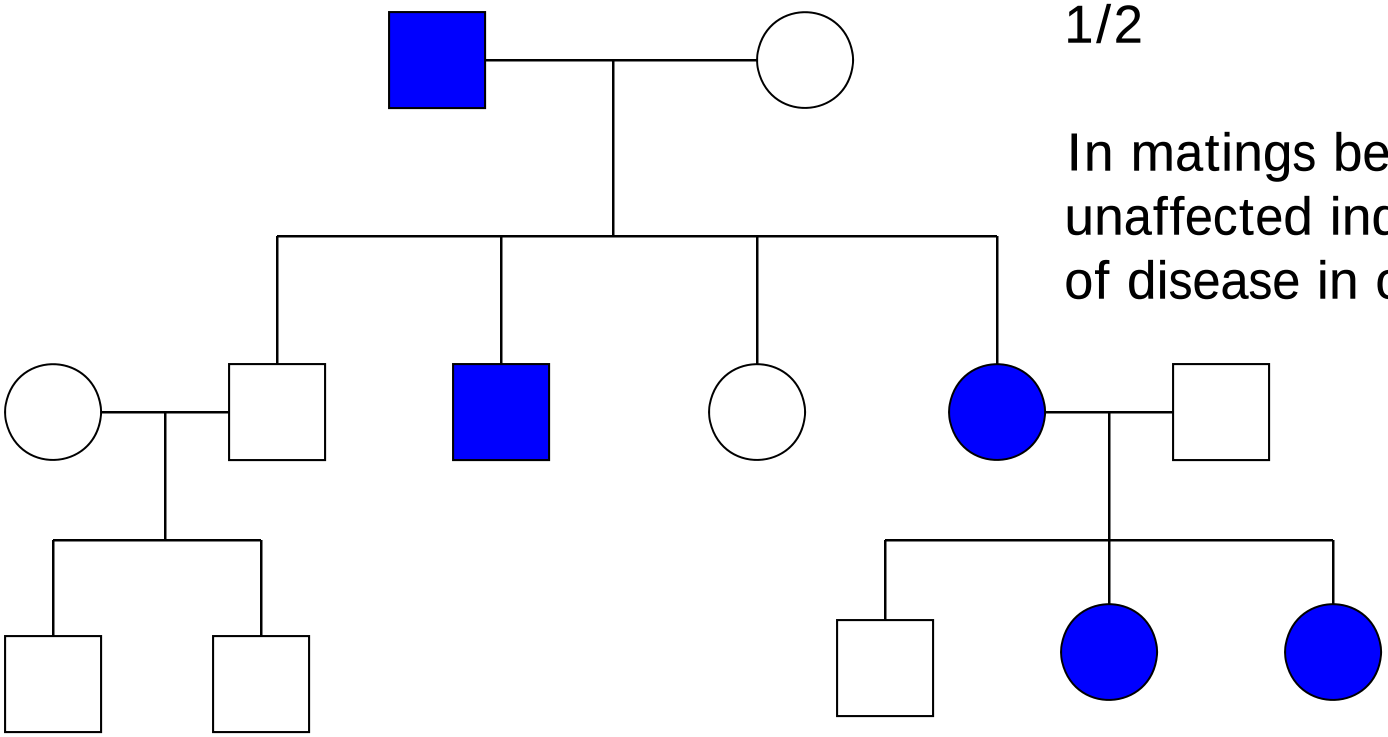
Pedigree Tree Symbols



MZ: Monozygotic twins; developed from same fertilized ovum

DZ: Dizygotic twins; developed from 2 separate fertilized ova

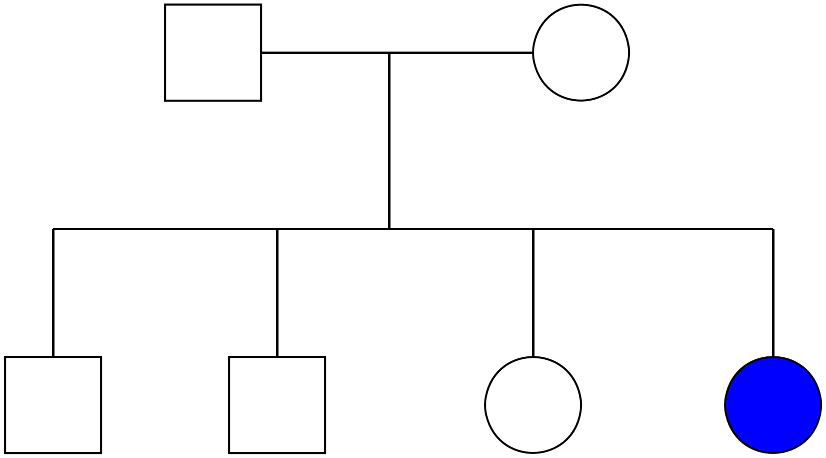
Autosomal Dominant Diseases



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

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.

Single offspring families

- Consider all families where both parents are heterozygous carriers with 1 offspring
- The offspring will be affected in 1/4 of these families.
- 3/4 of the families, which have an unaffected offspring, are indistinguishable from other families in the population and cannot be recruited for study
- In the families recruited for study, the segregation ratio appears to be 1:0
- However, because recruitment requires the offspring to be affected, these families actually do not contain any information on the segregation ratio.

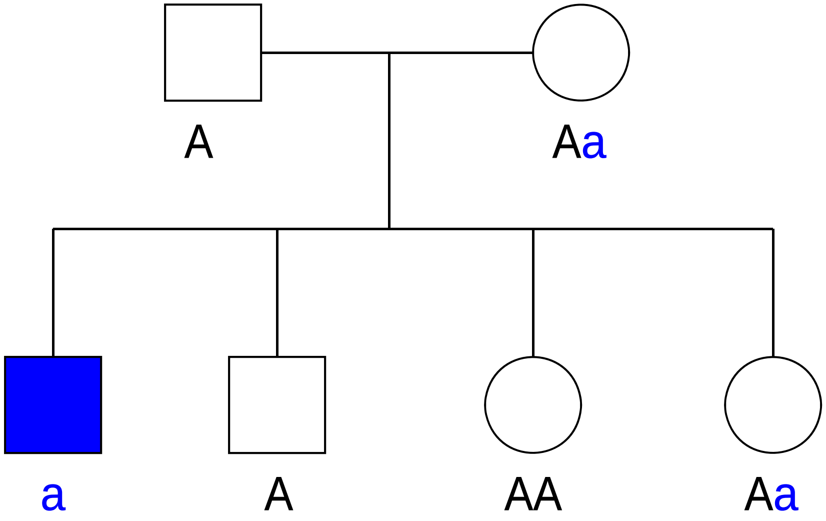
Two offspring families

- 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 be included in the study
- 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 ratio of the number of families with 2 affected offspring, to the number of families with 1 affected offspring, is 1:6
- Such a ratio would be consistent with autosomal recessive inheritance

The proband method

- In a study where families are recruited through an initial list of affected individuals (e.g. patients who attended a particular hospital), the affected individuals in the list are called "index cases" or "probands"
- Estimates the segregation ratio by the proportion of the probands' siblings who are affected
- When ALL 2 offspring families with at least 1 affected offspring are included, this indicates that all affected offspring are probands (**complete ascertainment**)
- Thus each 2-affected-offspring family contains 2 probands each with 1 affected sibling, while each 1-affected-offspring family contains 1 proband with 1 unaffected sibling
- The expected segregation ratio estimate is (1x2x1+6x1x0)/(1x2+6x1) = 1/4
- This method still works under **incomplete ascertainment**.

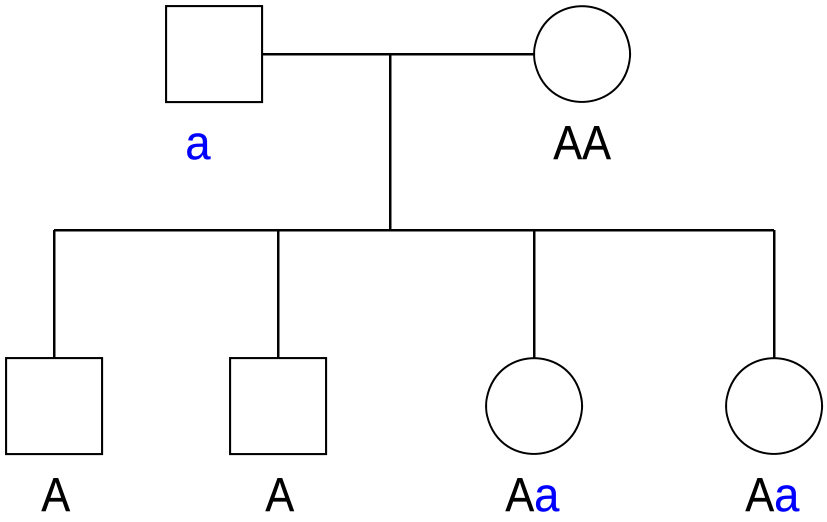
X-linked Recessive Disorders



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

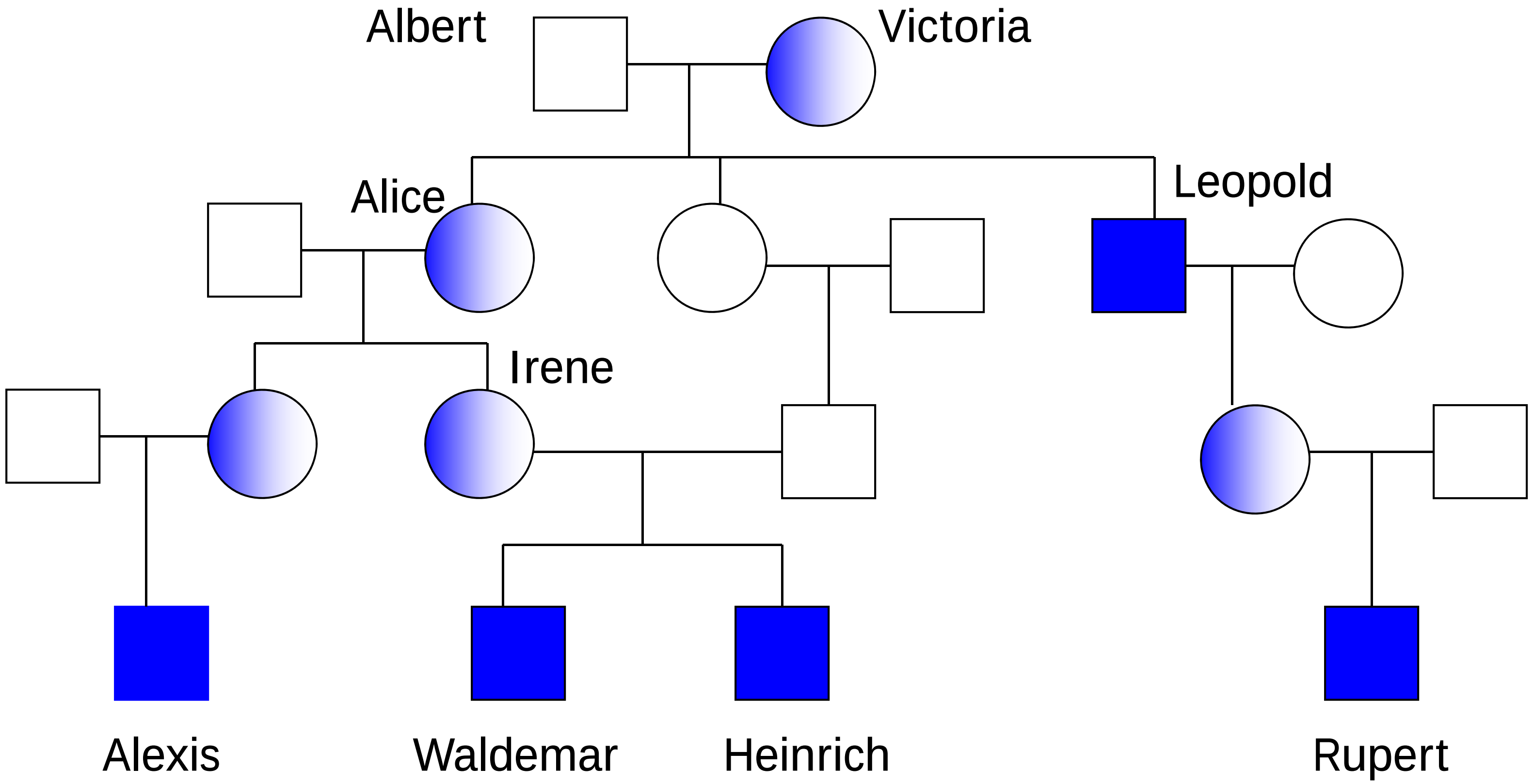
Complication: X inactivation may give rise to mild form of disease of variable severity in heterozygous females

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

A Royal Pedigree Tree



What is the likely mode of inheritance of this disease?

Hardy-Weinberg Law

In a large population under random mating:

- Allele frequencies in the offspring, denoted as **p** and **q**, are the same as those in the parental generation.
- Genotype frequencies in the offspring will follow the ratios **p^2**:**$2pq$**:**q^2**, regardless of the genotype frequencies in the parents.

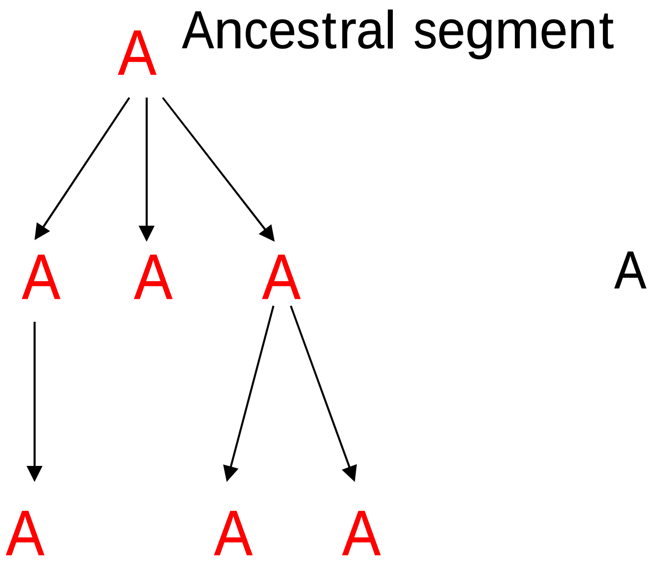
Exercise

An autosomal recessive disorder has a frequency of 1 in 10,000 births in a large, random mating population

1. What would be the predicted frequency of the disease allele in the population?
2. What would be the predicted frequency of carriers of the disease in the population?

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).

1. What is the kinship coefficient for two individuals who are full siblings, assuming that the parents are not inbred and are not genetically related to each other?

2. 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?

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:

AB CD

Each child will have 1 of 4 equally likely genotypes:

AC AD BC BD

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

The probability of *K* = ½ is also important. For full sibs this probability is ¼

Important: Kinship has a **local** as well as a **genome-wide** meaning

“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.

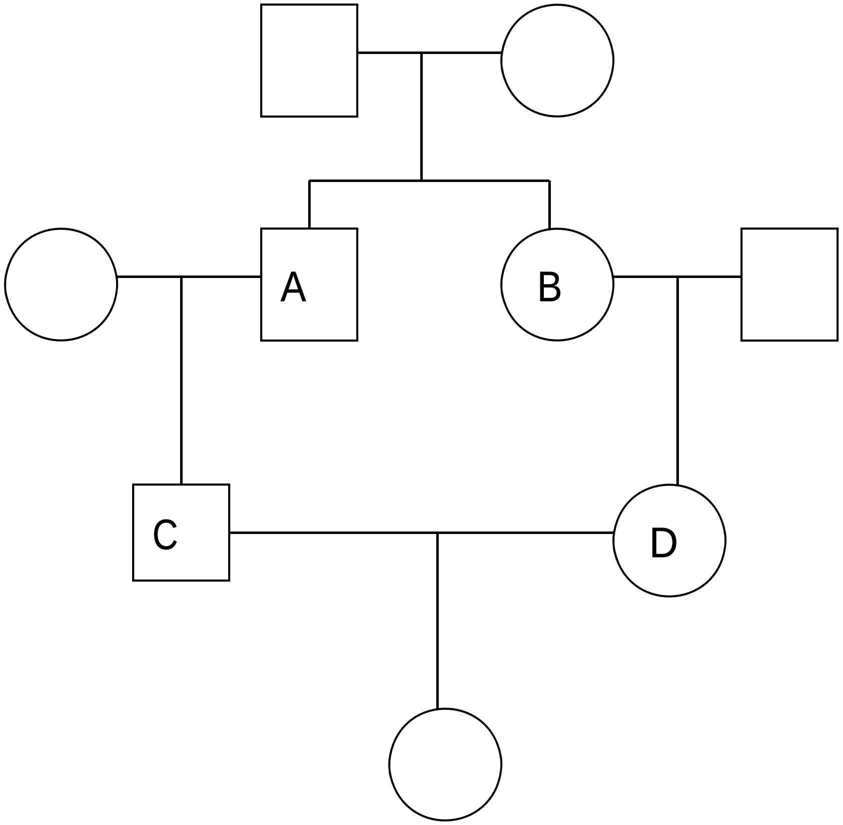
Therefore the kinship coefficient between A and the offspring of B is ½ *K*.

Inbreeding and recessive disease

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?

Pedigree diagram

What is the kinship coefficient between individuals C and D?



K(A,B) = 1/4

K(A,D) = 1/8

K(C,D) = 1/16

Solution

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
$1/42 \times 1/16 + 1/1800 \times 15/16 \sim 1/500$

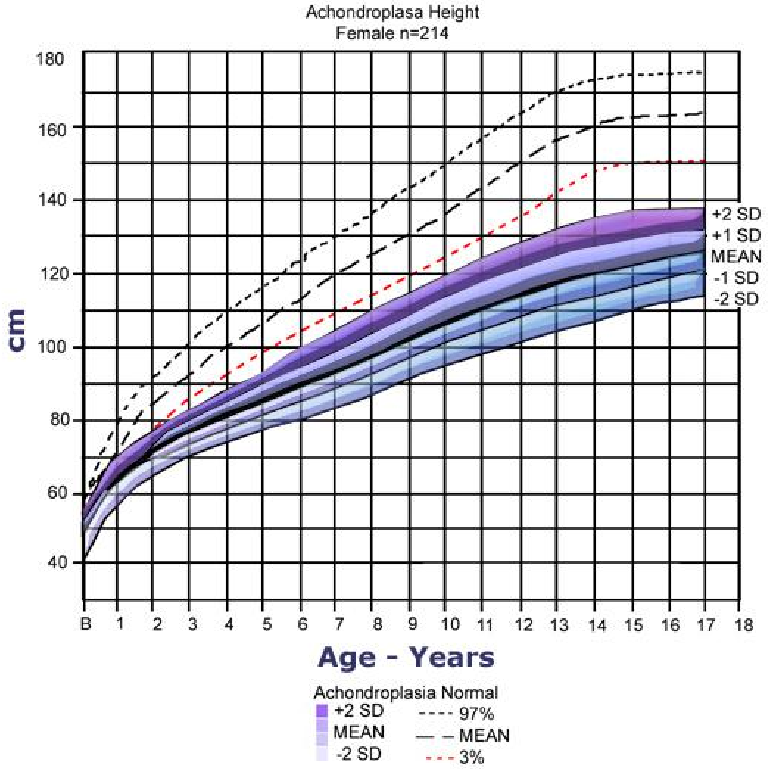
Quantitative traits

- 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**

The dance of chance – a new Galton machine
https://www.youtube.com/watch?v=epq-dpMJIxs

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.

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**):

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

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

Checking zero correlation

X_1	X_2	Frequency	X_1-p	$(X_1-p)(X_2-p)$
0	0	$(1-p)^2$	-p	p^2
0	1	p(1-p)	-p	-p(1-p)
1	0	p(1-p)	1-p	-p(1-p)
1	1	p^2	1-p	$(1-p)^2$

$E(X_1-p) = 0$

$E[(X_1-p)(X_2-p)] = p^2(1-p)^2 – 2p^2(1-p)^2 + p^2(1-p)^2 = 0$

$Cov[(X_1-p),(x_1-p)(X_2-p)] = – p^3(1-p)^2 + p^3(1-p)^2 – p^2(1-p)^3 + p^2(1-p)^3 = 0$

Variances

X_1	X_2	Frequency	$(X_1\text{-}p)^2$	$[(X_1\text{-}p)(X_2\text{-}p)]^2$
0	0	$(1\text{-}p)^2$	p^2	p^4
0	1	$p(1\text{-}p)$	p^2	$p^2(1\text{-}p)^2$
1	0	$p(1\text{-}p)$	$(1\text{-}p)^2$	$p^2(1\text{-}p)^2$
1	1	p^2	$(1\text{-}p)^2$	$(1\text{-}p)^4$

$\mathrm{Var}(X_1\text{-}p) = p^2(1\text{-}p)^2 + p^3(1\text{-}p) + p(1\text{-}p)^3 + p^2(1\text{-}p)^2 = p(1\text{-}p)$

$\mathrm{Var}[(X_1\text{-}p)(X_2\text{-}p)] = p^4(1\text{-}p)^2 + 2p^3(1\text{-}p)^3 + p^2(1\text{-}p)^4 = p^2(1\text{-}p)^2$

Variance components

- Additive: $V_A = 2p(1\text{-}p)b^2$
- Dominance: $V_D = p^2(1\text{-}p)^2g^2$
- 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

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

- For dominance

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

- **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 $4 \mathrm{x} K \mathrm{x} V_A/2 = 2KV_A$
- 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 dominances variances, respectively, across the whole 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
- Why? How?

Heritability

- The heritability is a property of the population as well as the trait.
- For the same trait, it can vary between different populations, and change over time
- It measures the relative importance of genetic and environmental differences, at a given time in a particular population.
- For traits with very low heritability, studies to identify specific genetic influences may be unwarranted.
- However, a high heritability does not guarantee that specific genetic influences will be easy to identify, as this depends on whether that heritability is made up of a few loci or a very large number of loci.
- A high heritability means that the trait may be somewhat resistant to changes in the environment within the normal range of population.

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 ½V_A, and the dominance covariance is 0
- If additive genetic effects are the only source of parent-offspring correlation is expected to be ½V_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 $r_{PO} – r_{POA}$

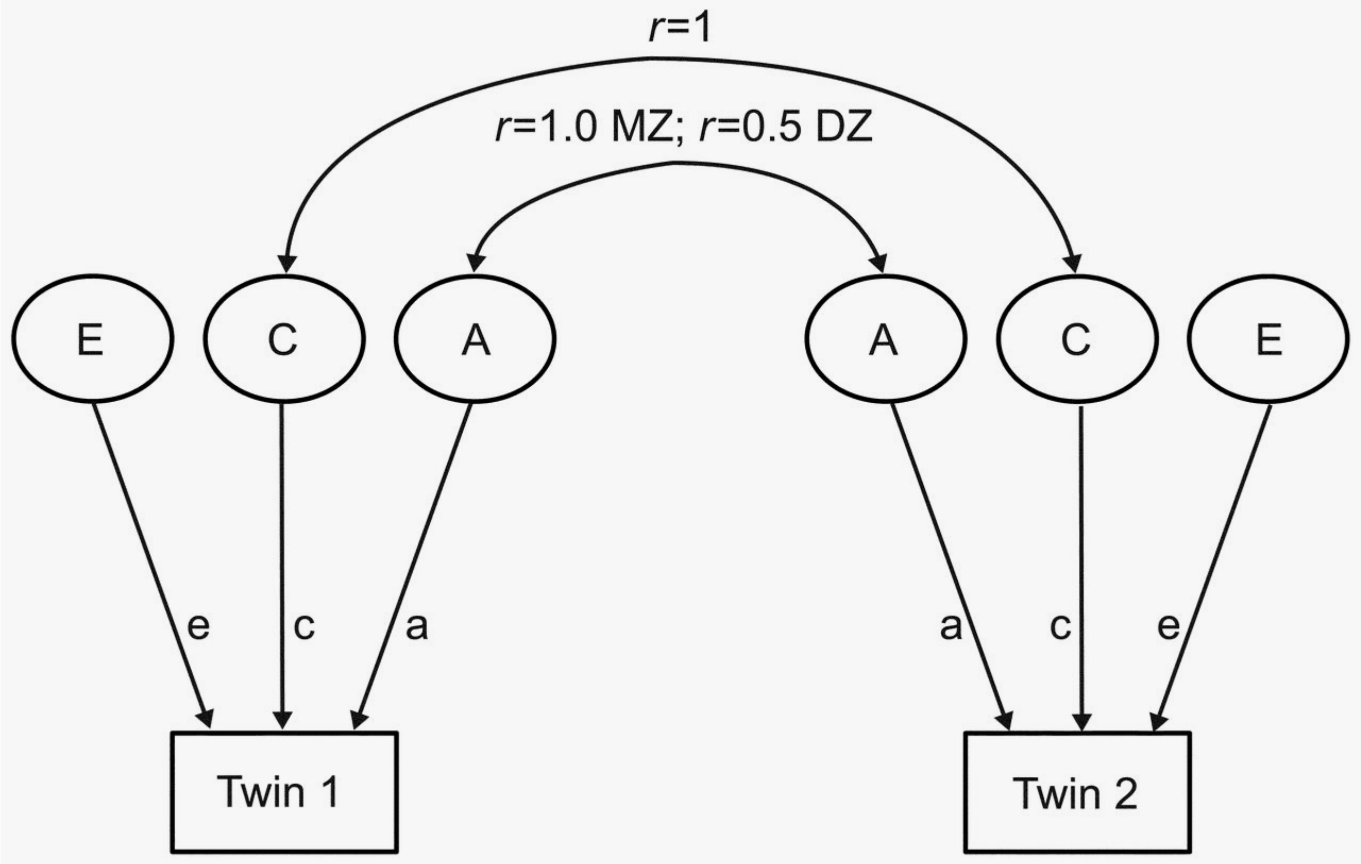
Genetic covariance of 2 traits

- Some loci that influence one trait may also influence another trait.
- Such overlapping genetic loci lead to covariance, and therefore correlation, between the 2 traits in the population.
- In an adoption study, the additive genetic covariance can be estimated by twice the empirical covariance between trait 1 in the biological parent and trait 2 in the adopted away offspring, or twice the empirical covariance between trait 2 in the biological parent and trait 1 in the adopted away offspring, or an average of the two estimates.

Twin studies

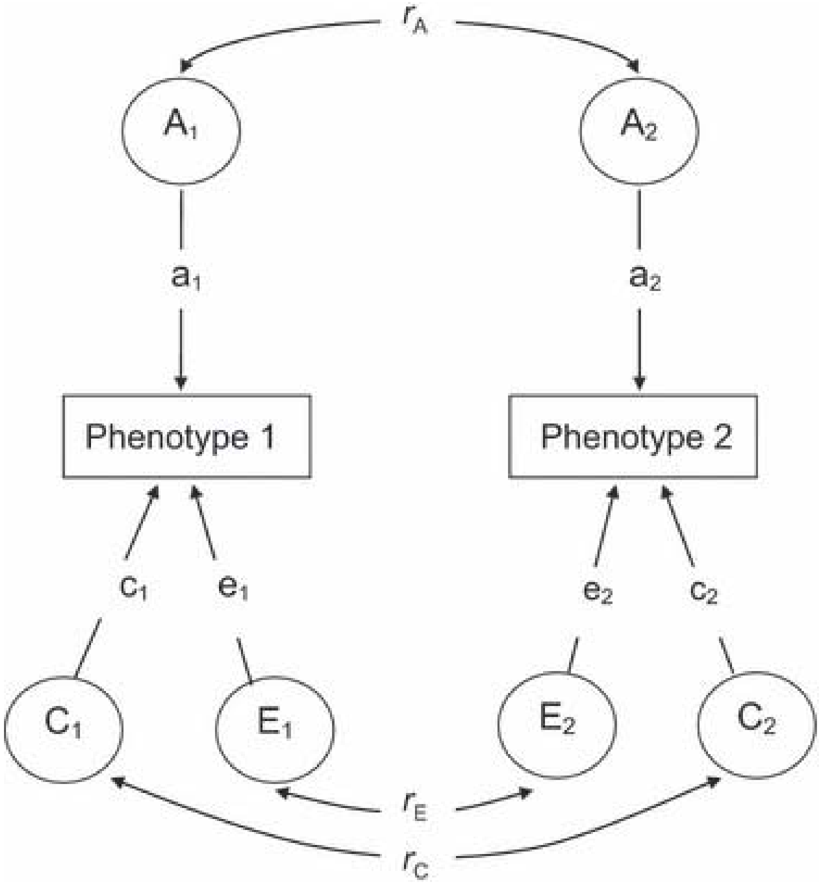
- Twins provide another way of separating genetic effects from shared environmental influences
- This takes advantage of the existence of two types of twins: monozygotic (MZ) and dizygotic (DZ)
- MZ twins are developed from the same zygote, and therefore *K* = ½
- DZ twins are developed from two separate zygotes, 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 $r_{MZ} – h^2$
- What if $r_{MZ} < h^2$?

Path diagrams



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.



Complex twin studies

- The simple approach of calculating empirical covariances or correlations and equating them to functions of variance components is inadequate for more complex studies:
  - Gender differences
  - Shared genetic and environmental effects among multiple phenotypes
  - Change over the lifespan (longitudinal)
  - Causal relationships between phenotypes
  - Additional family members (e.g. parents / siblings / children of twins)
  - Categorical variables (e.g. disease status)
  - Non-random samples (e.g. ascertainment for affected twins)
  - Measured genotype data
- Structural equation models are used in such studies

Summary

- Classical genetics: definition and relevance
- Segregation analysis
- Kinship coefficients
- Family, twin and adoption studies
- Estimating heritability
- Estimating genetic covariance
- (Linkage and association: excluded)

THANK YOU