Biometrical Models and Introduction to Genetic Analysis

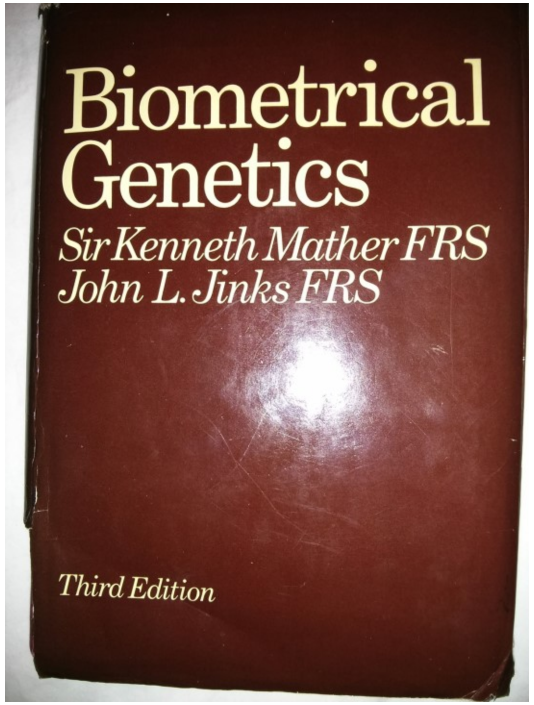
Pak Sham, University of Hong Kong

4th March 2019

The 2019 International Workshop on Statistical Genetics Methods for Human Complex Traits

What is biometrical genetics?

- How do genes contribute to the biometrical (statistical) properties of continuous (quantitative) traits in the populations
- For single trait, biometrical properties include
  - Means and Variances in individuals
  - Covariances between relatives
- For multiple traits, biometrical properties also include
  - Covariances between different traits in the same individual
  - Covariances between different traits in different (related) individuals



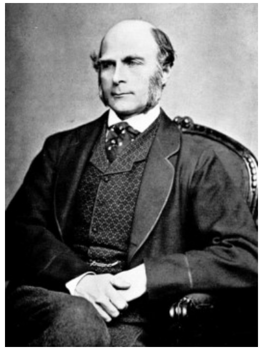
Biometrical Genetics

Sir Kenneth Mather FRS
John L. Jinks FRS

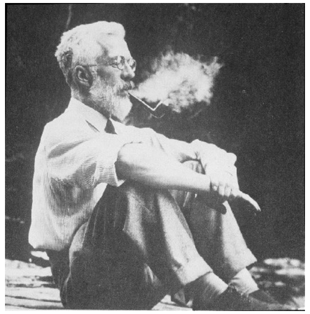
Third Edition

History of biometrical genetics

Mendel
Genes



Galton
Biometrics



Fisher
Correlation between relatives on the supposition of mendelian inheritance

Mather

Jinks

Biometrical genetics

Fulker

Eaves

Statistical modelling in biometrical genetics

Genes are discrete entities

https://gameofthrones.fandom.com/wiki/Dwarfism

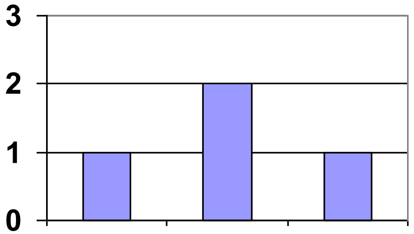
- Mendelian disorders are caused by mutations in a single gene
- Mendelian disorders are also discrete entities
- How can discrete entities produce continuous variation?

Origin of continuous variation

- Continuous (quantitative) variation can be explained by polygenic inheritance
- The sum of independent and approximately equal influences will approach a continuous, normal distribution, as the number of influences increases (central limit theorem)

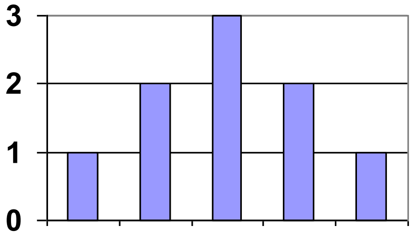
1 Gene

→ 3 Genotypes
→ 3 Phenotypes



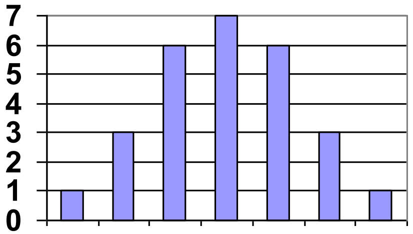
2 Genes

→ 9 Genotypes
→ 5 Phenotypes



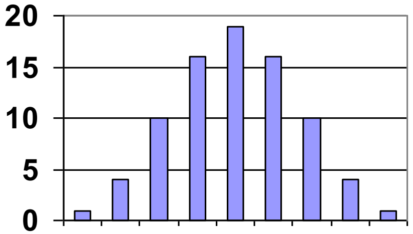
3 Genes

→ 27 Genotypes
→ 7 Phenotypes



4 Genes

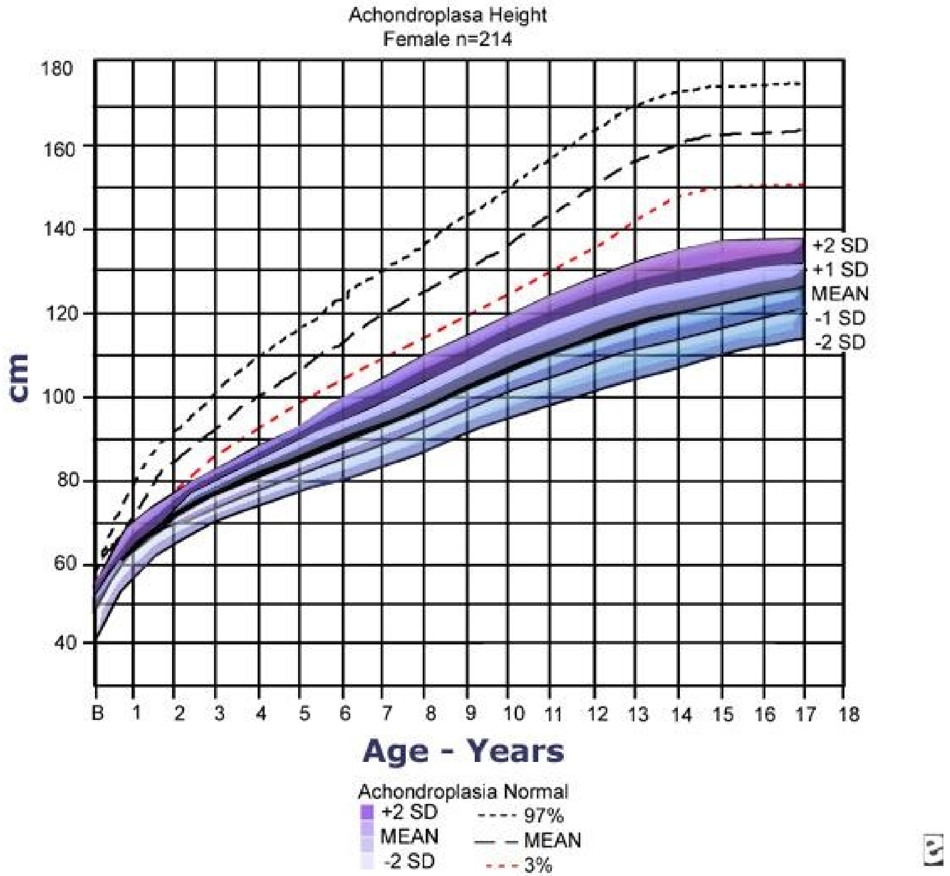
→ 81 Genotypes
→ 9 Phenotypes



https://www.youtube.com/watch?v=kDkmSI39sWQ

Major loci and polygenes

- 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.

Genetic polymorphisms, alleles, genotypes

- A genetic polymorphism is a variable site in the genome (e.g. single nucleotide polymorphism, SNP)
- The alternative sequences at a locus are called alleles, often denoted as capital and small letters (e.g. A, a)
- The alleles present at the polymorphic site (locus) of an individual is called his or her genotype (e.g. AA, aa, Aa)

Analysis of variance

- Fisher developed Analysis of Variance (ANOVA) for “factorial designs”, where the factors have discrete levels (e.g. binary).
- The overall variance of a trait is decomposed into components due to the main effects of the factors, two-way interactions, 3-way interactions, etc., in a hierarchical fashion

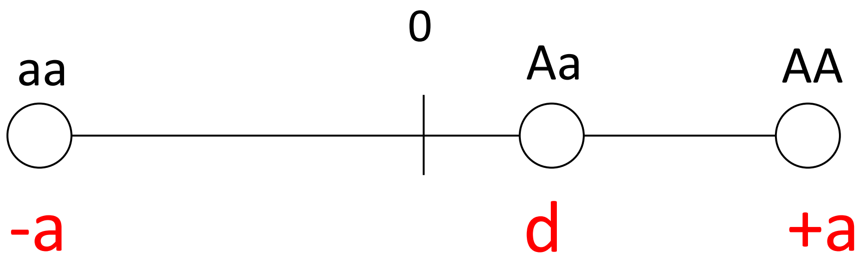
	0	**1**
0		
1		

Biometrical model for single locus

- Consider the effects of a single locus on a quantitative trait
- All other influences are considered as “error” or “residual”, which are assumed to be uncorrelated and have no interaction with the locus being considered
- In Fisher’s convention, effects are measured from the “midpoint” of two homozygous genotypes

Model: $Y = c + X + R$

X:



R: Residual influences

Note: we do not distinguish paternal from maternal transmitted alleles, implicitly assuming that their effects are the same

	Genotype means
AA	c + a
Aa	c + d
aa	c – a

Population genotype frequencies

- We also need to specify the frequencies of the 3 genotypes in the population
- In a large population under random mating, the frequencies of genotypes AA, Aa and aa follow the binomial proportions s $\mathbf{p^2}$:**2pq**:$\mathbf{q^2}$, where p and q (=1-p) are the frequencies of alleles A and a
- Genotypes in such proportions are said to be in Hardy-Weinberg equilibrium; deviation from such proportions is called Hardy-Weinberg Disequilibrium (HWD)

Derivation of Hardy-Weinberg proportions

Parental frequencies – not necessarily in Hardy-Weinberg proportions

Genotype	Frequency
AA	P
Aa	Q
aa	R

Allele	Frequency
A	P+Q/2
a	R+Q/2

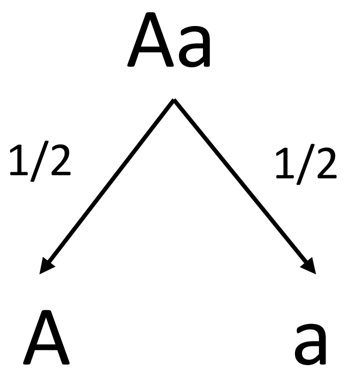
Random mating

Under random mating, the mating type frequencies are

	AA	**Aa**	**aa**
AA	P^2	PQ	PR
Aa	PQ	Q^2	QR
aa	PR	QR	R^2

Mendelian segregation

- Mendel's law of segregation: when a parent has heterozygous genotype Aa, there is equal probability for the two alleles (A and a) to be transmitted to an offspring)



Segregation ratios

According to Mendel’s law of segregation, the offspring genotype frequencies for the mating types are:

	AA	**Aa**	**aa**
AA	AA	AA:Aa 0.5:0.5	Aa
Aa	AA:Aa 0.5:0.5	AA:Aa:aa 0.25:0.5:0.25	Aa:aa 0.5:0.5
aa	Aa	Aa:aa 0.5:0.5	aa

Offspring genotype frequencies

Averaging over the mating types, the offspring genotype frequencies are

Genotype	**Frequency**
AA	$P^2+PQ+Q^2/4 = (P+Q/2)^2$
Aa	$2PR+PQ+QR+Q^2/2 = 2(P+Q/2)(R+Q/2)$
aa	$R2+QR+Q^2/4 = (R+Q/2)^2$

Offspring allele frequencies

Averaging over the genotypes, the offspring allele frequencies are

Allele	**Frequency**
A	$(P+Q/2)^2 + (P+Q/2)(R+Q/2) = P+Q/2$
a	$(R+Q/2)^2 + (P+Q/2)(R+Q/2) = R+Q/2$

Hardy-Weinberg equilibrium

The genotype can be thought of as consisting of 2 independent factors, one from each parent (as in a 2-way factorial design)

	A	a	
A	p^2	pq	p
a	pq	q^2	q
	p	q	

Biometrical model: mean

The mean effect of genotype under Hardy-Weinberg Equilibrium is thus

Genotype	AA	Aa	aa
Frequency	p^2	2pq	q^2
Effect	a	d	-a

Mean $m = p^2(a) + 2pq(d) + q^2(-a)$

$= (p-q)a + 2pqd$

$$\mu = E(X) = \sum_i x_i f(x_i)$$

Biometrical model: variance

The variance of the genotypic effect is therefore

Genotype	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
$(X\text{-}m)^2$	$(a-m)^2$	$(d-m)^2$	$(-a-m)^2$

Variance $= (a-m)^2p^2 + (d-m)^2 2pq + (-a-m)^2q^2$

$= 2pq[a+(q-p)d]^2 + (2pqd)^2$

(intermediate steps not shown)

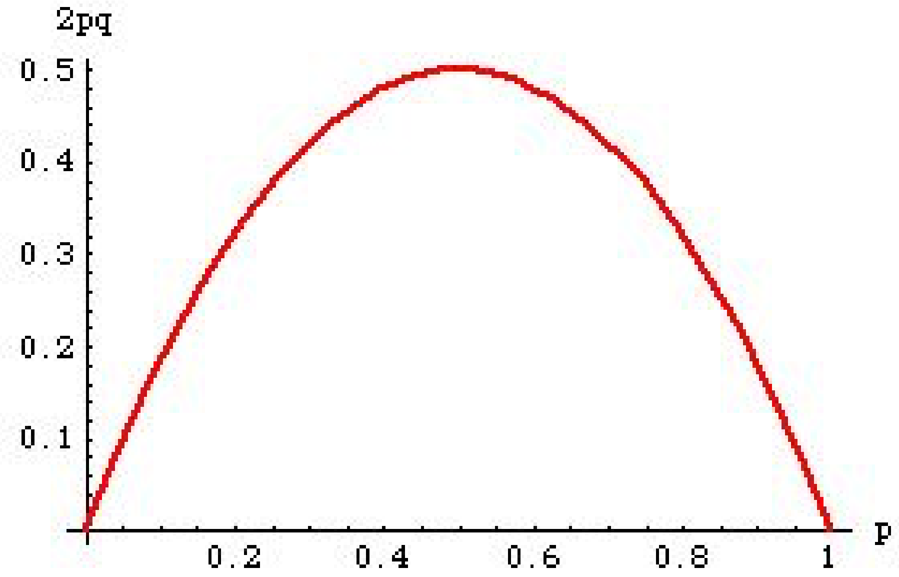
$$Var(X) = E(X-\mu)^2$$
$$= \sum_i (x_i - \mu)^2 f(x_i)$$

Average allele effect, additive variance

- The first variance component is due the additive effects of the two alleles of the genotype:
- The presence of dominance (i.e. when $d \neq 0$) means that the effect of an allele depends on the other allele in the genotype:
  - When the other allele is A, the effect of allele A is a-d (i.e. effect of AA – effect of Aa)
  - When the other allele is a, the effect of allele A is a+d (i.e. effect of Aa – effect of aa)
- Therefore the average effect of allele A = $p(a-d)+q(a+d)$ = $a+(q-p)d$
- If genotype is coded additively as G= number allele A in the genotype (i.e. G = 0, 1 or 2), then the regression coefficient is the trait on G is $a + (q-p)d$
- Thus the additive genetic variance is $(a+(q-p)d)^2$ Var(G) = $2pq(a+(q-p)d)^2$
- The second component of the variance, $(2pqd)^2$, is therefore attributed to the dominance deviation (2nd order interaction between the 2 alleles at the genotype at the same locus)

Variance components and heterozygosity

- 2pq is the expected heterozygosity of a biallelic locus under Hardy-Weinberg equilibrium
- When p=q=1/2, the expected heterozygosity takes its highest value of 1/2. As allele frequency approaches 0 or 1, heterozygosity approaches 0
- Additive genetic variance is proportional to the expected heterozygosity
- Dominance genetic variance is proportional to the square of the expected heterozygosity
- Dominance genetic variance declines much more rapidly than additive genetic variance, as allele frequency approaches 0 or 1. (Why is this intuitively obvious?)



Covariance between pairs of relatives

	AA	**Aa**	**aa**
AA	$(a-m)^2$		
Aa	$(a-m)(d-m)$	$(d-m)^2$	
aa	$(a-m)(-a-m)$	$(-a-m)(d-m)$	$(-a-m)^2$

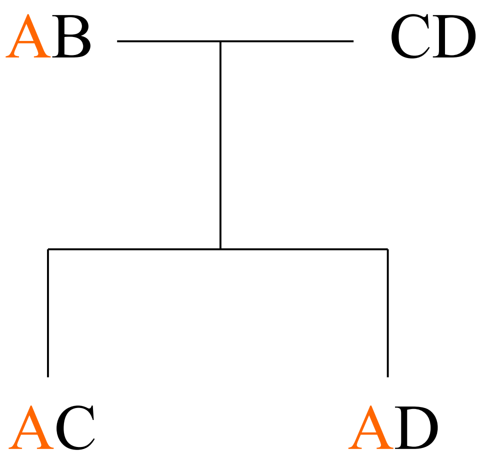
The matrix is symmetrical, therefore upper triangular elements are not shown,

The covariance between relatives of a certain class is the weighted average of these cross-products, where each cross-product is weighted by its frequency in that class.

$$
\begin{aligned}
Cov(X,Y) &= E\left(X-\mu_X\right)\left(Y-\mu_Y\right)\\
&\sum_i \left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)f\left(x_i,y_i\right)
\end{aligned}
$$

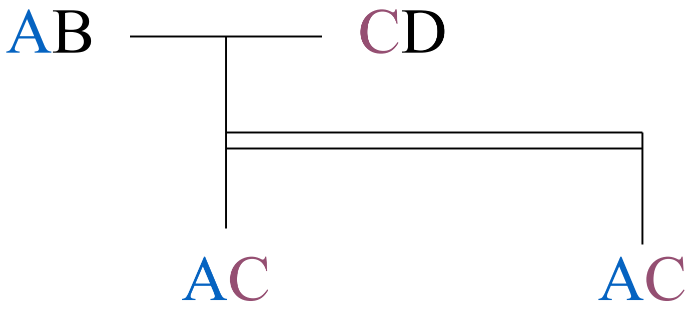
Genetic identity-by-descent (IBD)

For two-locus genotype frequencies of two relatives, the concept of genetic identity by descent is helpful



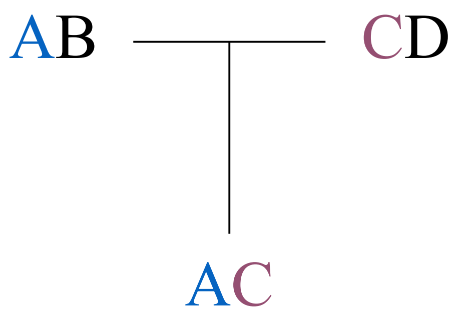
- DNA segments (e.g. genes) are identical-by-descent if they are descended from, and therefore replicates of, a single ancestral DNA segment.
- The IBD genetic segments should have identical genetic sequence (unless new mutation has occurred)
- At any autosomal location, two individuals can share 0, 1 or 2 alleles
- There are 3 genetic relationships where the IBD sharing is the same throughput the genome (What are these?)

IBD for MZ twins



MZ twins share 2 alleles IBD for all loci

IBD for parent-offspring (PO)



When the parents are unrelated to each other, PO pairs share 1 allele IBD at all loci

IBD for unrelated individuals

- Two unrelated individuals share 0 alleles IBD at all loci

Covariance of MZ twins

	AA	**Aa**	**aa**
AA	p^2		
Aa	0	2pq	
aa	0	0	q^2

$$\begin{aligned}\text{Covariance} &= (a\text{-}m)^2p^2 + (d\text{-}m)^2 2pq + (\text{-}a\text{-}m)^2q^2 \\ &= 2pq[a+(q\text{-}p)d]^2 + (2pqd)^2 \\ &= V_A + V_D\end{aligned}$$

$$Cov(X,Y) = E\left(X-\mu_X\right)\left(Y-\mu_Y\right)$$

$$\sum_i \left(x_i-\mu_X\right)\left(y_i-\mu_Y\right)f\left(x_i,y_i\right)$$

Covariance for parent-offspring (P-O)

	AA	**Aa**	**aa**
AA	p^3		
Aa	p^2q	pq	
aa	0	pq^2	q^3

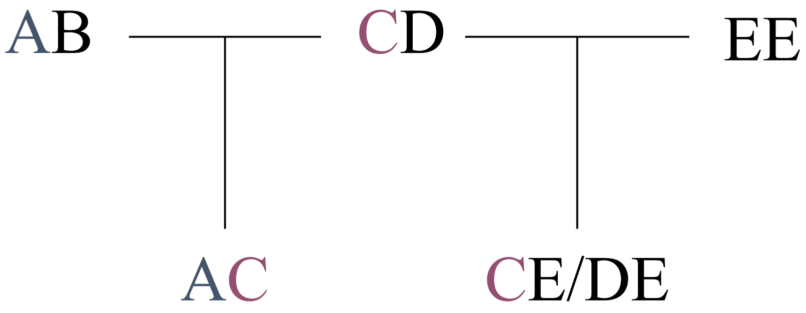
$$
\begin{aligned}
\text{Covariance} &= (a-m)^2p^3 + (d-m)^2pq + (-a-m)^2q^3 \\
&\quad + (a-m)(d-m)2p^2q + (-a-m)(d-m)2pq^2 \\
&= pq[a+(q-p)d]^2 \\
&= V_A / 2
\end{aligned}
$$

Covariance for unrelated pairs (U)

	AA	**Aa**	**aa**
AA	p^4		
Aa	$2p^3q$	$4p^2q^2$	
aa	p^2q^2	$2pq^3$	q^4

$$\begin{aligned}\text{Covariance} &= (a\text{-}m)^2p^4 + (d\text{-}m)^2 4p^2q^2 + (\text{-}a\text{-}m)^2q^4 \\ &\quad + (a\text{-}m)(d\text{-}m)4p^3q + (\text{-}a\text{-}m)(d\text{-}m)4pq^3 \\ &\quad + (a\text{-}m)(\text{-}a\text{-}m)2p^2q^2 \\ &= 0\end{aligned}$$

IBD: half sibs



IBD Sharing	Probability
0	½
1	½

Average IBD sharing = 0(1/2) + 1(1/2) = 1/2

In terms of IBD sharing, half siblings are similar to
- Parent-offspring for ½ of the genome
- Unrelated individuals for ½ of the genome

Covariance: half sibs

Genotype frequencies are weighted averages:

½ Parent-offspring (when IBD=1)

½ Unrelated (when IBD=0)

Covariance $= \frac{1}{2}(V_A/2) + \frac{1}{2}(0)$

$= \frac{1}{2}V_A$

IBD: full sibs

IBD paternal alleles

		0	**1**
IBD maternal alleles	0	0	1
	1	1	2

IBD Sharing	Probability
0	1/4
1	1/2
2	1/4

Average IBD sharing = 0(1/4) + 1(1/2) + 2(1/4) = 1

In terms of IBD sharing, full siblings are similar to

MZ twins for ¼ of the genome
Parent-offspring for ½ of the genome
Unrelated individuals for ¼ of the genome

Covariance: full sibs

Genotype frequencies are weighted averages:

¼ MZ twins (when IBD=2)
½ Parent-offspring (when IBD=1)
¼ Unrelated (when IBD=0)

$$\text{Covariance} = \tfrac{1}{4}(V_A+V_D) + \tfrac{1}{2}(V_A/2) + \tfrac{1}{4}(0)$$

$$= \tfrac{1}{2}V_A + \tfrac{1}{4}V_D$$

Generalization: proportion of alleles IBD (π)

- IBD can be expressed as a proportion π (= number IBD / 2), thus $\pi = 0, 1/2$ or 1
- The probability distribution π is Prob(π=0), Prob(π=1/2), Prob(π=1)
- $E(\pi) = \text{Prob}(\pi=1) + (1/2)\,\text{Prob}(\pi=1/2)$
- $\text{Var}(\pi) = \text{Prob}(\pi=1) + (1/4)\,\text{Prob}(\pi=1/2) - (E(\pi))^2$

Relationship	**E(π)**	**Var(π)**	**Prob(π=1)**
MZ	1	0	1
Parent-Offspring	0.5	0	0
Unrelated	0	0	0
Half sibs	0.25	0.0625	0
Full sibs	0.5	0.125	0.25

Covariance: general relative pair

The covariance is a weighted average of the covariances for MZ twins, parent-offspring and unrelated individuals

$$\begin{aligned} \text{Covariance} &= \text{Prob}(\pi=1)(V_A+V_D) + \text{Prob}(\pi=1/2)(V_A/2) + \text{Prob}(\pi=0)(0) \\ &= (\text{Prob}(\pi=1)+\text{Prob}(\pi=1/2)/2)V_A + \text{Prob}(\pi=1)V_D \\ &= E(\pi)V_A + \text{Prob}(\pi=1)V_D \end{aligned}$$

Kinship coefficient

- The kinship coefficient (*K*) between two individuals is defined as the probability that two alleles, one from each individual, drawn at random at an autosomal locus, will be identical-by-descent (IBD)
- Let the paternal and maternal alleles of individuals 1 and 2 be denoted G_{1P}, G_{1M}, G_{2P}, G_{2M}. The genotypes of the 2 individuals, additively coded (0,1,2), would be $G_1=G_{1P},+G_{1M}$ and $G_2=G_{2P},+G_{2M}$
- The covariance between the two genotypes is

$Cov(G_1, G_2) = Cov(G_{1P},G_{2P})+Cov(G_{1P},G_{2M})+Cov(G_{1M}, G_{2P})+Cov(G_{1M},G_{2M})$

- In the absence of inbreeding, $Var(G_1) = Var(G_2) = 2pq$, and each covariance term is either pq when the alleles are IBD or 0 when they are not. Also, each allele of one person can be IBD with at most 1 allele of the other person. In this scenario $E(\pi)$ is equivalent to $2K$ and represents the correlation between G_1 and G_2
- *K* is of wider applicability than $E(\pi)$ when there is inbreeding

“Attenuation” of kinship

- 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?
- 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*.
- Applying this result recursively, we can show that the Kinship coefficient between two individuals sharing one common ancestor is equal to $(½)^{g+1}$, where g is the number of meioses separating the 2 individuals

Inbreeding coefficient

- The inbreeding coefficient of an individual , I, is the probability that the 2 alleles at any locus are IBD. It is equal to the kinship coefficient of his or her parents, since in meiosis an allele is randomly drawn from the genotype of a parent.
- Inbreeding inflates the variance of a additively coded genotype:

$$\mathrm{Var}(G) = \mathrm{Var}(G_P)+\mathrm{Var}(G_M)+2\mathrm{Cov}(G_P,G_M)$$

- Inbreeding also inflates the covariance between the additively coded genotypes of 2 individuals, since now it is possible for an allele in one person to be IBD with both alleles of the other person.

Two-locus biometrical genetic model

- Generalize biometrical model to 2 loci
- This is necessary only when there is either correlation or interaction between the 2 loci; otherwise the loci can be considered separately
- Two-locus Interactions include
  - second-order inter-loci interactions involving 1 allele each locus, additive-additive (AA)
  - third-order interactions involving both alleles from one locus and 1 allele from the other, additive-dominance (AD)
  - fourth-order interactions involving both alleles from both loci, dominance-dominance (DD)
- For 2 loci, there are 3x3=9 genotypic groups (assuming no parent-of-origin effect). In principle, if we can write down the trait means and population frequencies of these 9 genotypic groups, then we can proceed with variance partitioning using a hierarchical ANOVA, when the two loci are not correlated. This is straightforward by computer program but tedious by hand - see Sham (1997) Statistics in Human Genetics, Chapter 5.

Covariance of epistatic components

- The AA interaction between 2 alleles are shared by 2 individuals when the 2 alleles are both IBD, and not shared when at least one of them is not IBD. When the 2 loci are independent, the probability of sharing is the product of proportion of IBD sharing of the 2 loci, $\pi_1\pi_2$. For a particular class of relative pairs, the expected covariance is $E(\pi_1\pi_2)=[E(\pi)]^2$
- Similarly, the expected covariance of the AD interactions for a class of relative pairs is $E(\pi)Prob(\pi=1)$
- Finally, the expected covariance of the DD interactions for a class of relative pairs is $[Prob(\pi=1)]^2$

Covariance: general relative pair

Including 2 loci interactions, the covariance for 2 relatives of a given class is:

Covariance = $E(\pi)V_A + P(\pi=1)V_D + [E(\pi)]^2V_{AA} + E(\pi)P(\pi=1)V_{AD} + [P(\pi=1)]^2V_{DD}$

This can be further extended to epistasis involving more than 2 loci

Genetic linkage - two-locus transmission

- The correlation between 2 loci depends on “linkage”
- 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?

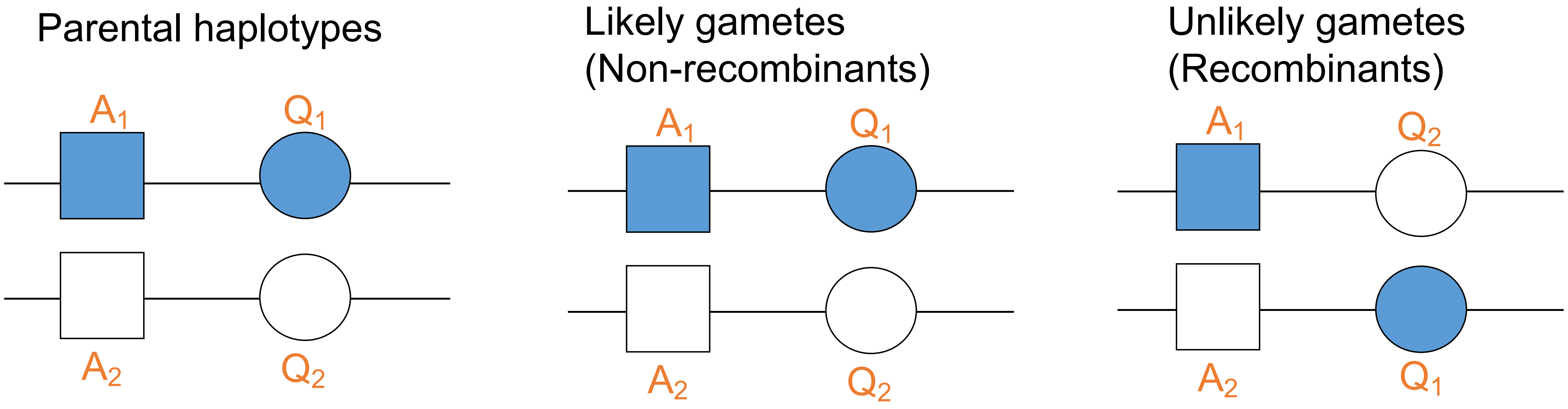
AaBb

Parent

AB Ab ab aB

Gametes

Haplotypes and recombination



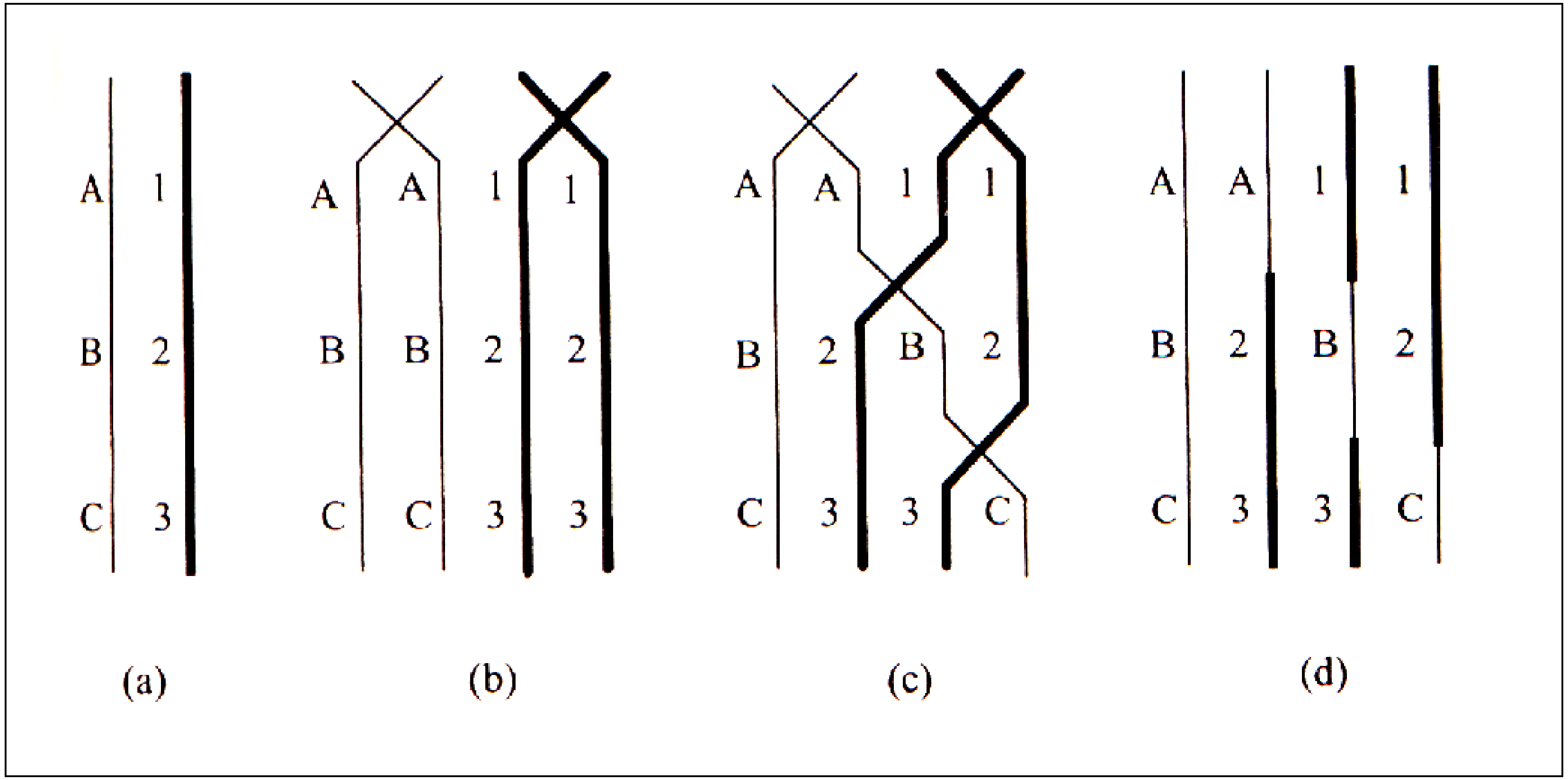
Parental haplotypes

Likely gametes (Non-recombinants)

Unlikely gametes (Recombinants)

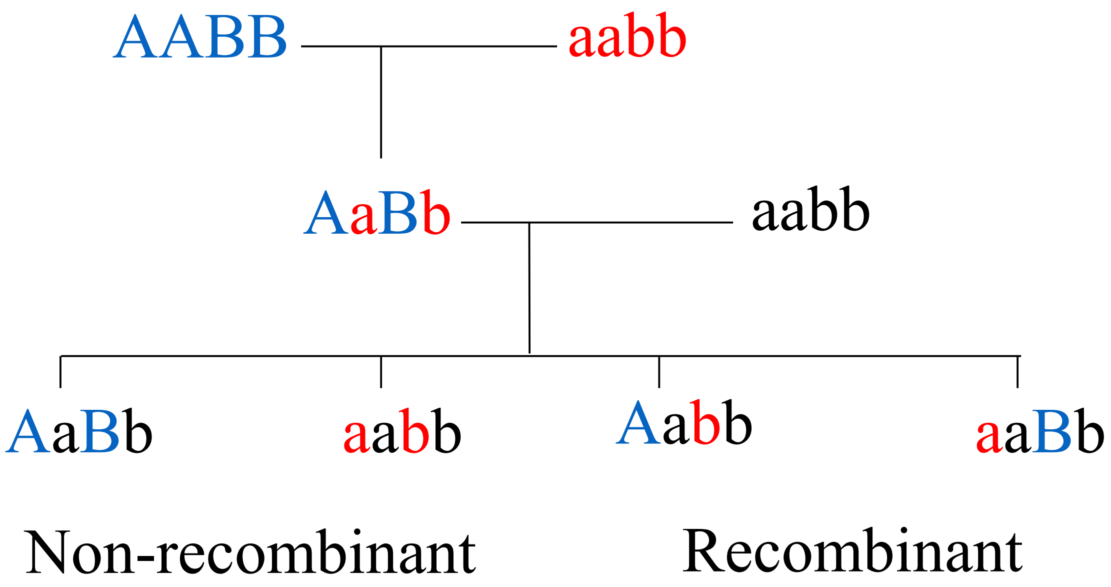
- Haplotype = set of alleles inherited from the same parent
- Alleles that were inherited together from the previous generation are more likely to be transmitted together to the next generation, if the loci are on the same chromosome
- Alleles which have different parental origins but are transmitted 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

Fully Informative Gametes



- Recombinants and non-recombinants can be inferred in double backcross data.
- The offspring of the double backcross constitute fully informative gametes

Population haplotype frequencies

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

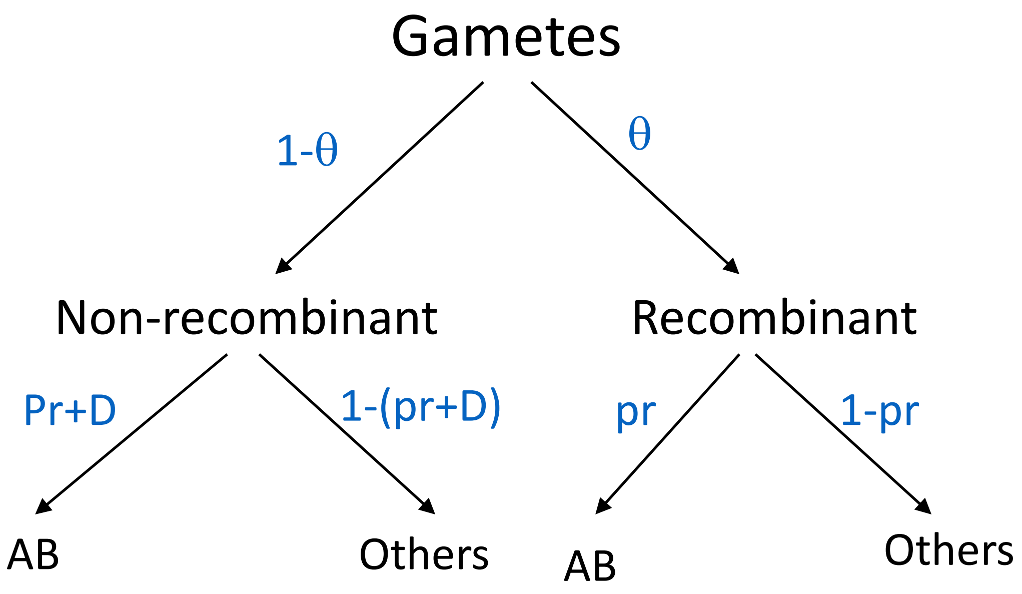
- If there is no association between alleles of the two loci, then the frequency of each haplotype is equal to the product of the frequencies of its constituent alleles
- Two loci with such haplotype frequencies are said to be in linkage equilibrium

Linkage Disequilibrium (LD)

	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
- The deviation D is a measure of linkage disequilibrium
- The normalized D' measure = D/D_{max}. When $D>0$, it cannot exceed the smallest value which causes either $ps-D<0$ or $qr-D<0$. Similar consideration applies when $D<0$.$D'=1$ implies that 1 of the 4 haplotypes is absent.
- The r^2 measure is $D^2/pqrs$ and represents the squared correlation between the two haplotypes coded numerically. An r^2 of 1 implies that 2 of the 4 haplotypes are absent, and that the 2 loci have equal allele frequencies.

Decay of LD through recombination



Gametes

$1-\theta$

θ

Non-recombinant

Recombinant

Pr+D

1-(pr+D)

pr

1-pr

AB

Others

AB

Others

Frequency of AB gametes = $(1-\theta)(pr+d)+\theta pr = pr+(1-\theta)D$

- Thus, the LD measure D decays by a factor of $(1-\theta)$ per generation.
- For unlinked loci, any LD will quickly decay to near 0, whereas for tightly linked loci, any LD will be maintained for many degenerations.
- In any case, once the haplotype frequency decays to pr, it will tend to stay at that frequency (other than random fluctuations), hence “linkage equilibrium”)

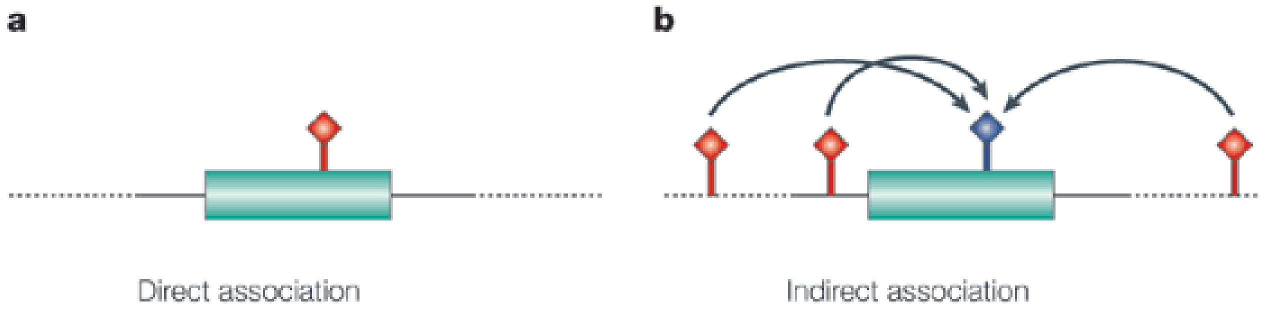
Impact of LD on biometrical model

- Denote the additive genetic effects of loci 1 and 2 by G_1 and G_2, with additive variances V_{1A} and V_{2A} respectively.
- In the absence of LD, the variance of the total additive genetic effects $G=G_1+G_2$ is simply $V_{1A}+V_{2A}$
- However, in the presence of LD, G_1 and G_2 are correlated, and the variance of G becomes $V_{1A}+V_{2A}+2\mathrm{Cov}(G_1,G_2) =V_{1A}+V_{2A}+2r\sqrt{}(V_{1A}V_{2A})$, where r is the correlation between the trait increasing alleles of the 2 loci.
- Denote the the total additive effects of person 1 and person 2 as $(G_{11}+G_{12})$ and $(G_{21}+G_{22})$, respectively, the covariance between these total genetic effects is

$$\mathrm{Cov}(G_{11,}G_{21})+\mathrm{Cov}(G_{12,}G_{22})+\mathrm{Cov}(G_{11,}G_{22})+\mathrm{Cov}(G_{12,}G_{21}) = E(\pi)[(V_{1A}+V_{2A}+2r\sqrt{}(V_{1A}V_{2A})$$

- Thus the correlation between the additive effects remains unchanged at $E(\pi)$
- We do not attempt to address the impact of LD on dominance or epistasis.

LD allows indirect association analysis

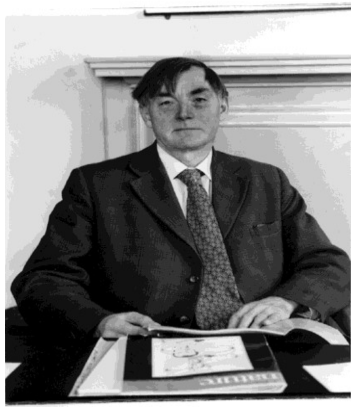


Nature Reviews | Genetics
https://www.nature.com/articles/nrg1521

- If the correlation between a trait and the un-genotyped causal locus is β and the correlation between the causal locus and a genotyped marker locus is r, then the overall correlation the trait and the genotyped marker locus is rβ
- When r is close to 1, testing the marker locus is almost equivalent to testing the causal locus – this makes indirect association feasible
- However, when r is modest this results in substantial reduction in the association signal, such that the sample size needs to be increased by a factor of 1/r^2 to achieve the same statistical power as an direct association analysis of the causal SNP.

Quasi-continuous phenotypes

- Some disease traits are apparently discrete (e.g. myocardial infarction) but reflect an underlying continuous process, e.g. coronary artery narrowing).
- When the underlying continuous process cannot be measured, a latent variable called liability is introduced. When liability exceeds a certain threshold, the diseases occurs
- Liability is assumed to be normally distributed, so that biometrical models developed for continuous traits apply



Douglas Falconer 1965: Inheritance of liability to certain diseases estimated from incidence among relatives

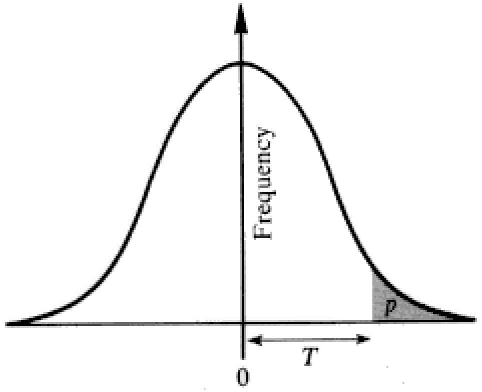


FIGURE 9.5
Threshold model. All individuals with a value of x greater than T are affected. The proportion of affected individuals is the area under the distribution curve beyond T.

Neglected topics

- Environmental influences
- Gene-environment correlation and interaction
- Multivariate (i.e. multi-trait) models
- Assortative mating
- Selection
- Mutation
- Random genetic drift

Data analysis under biometrical models

- Biometrical genetic models provide a coherent understanding of the genetic contributions to the statistical properties of traits in the population.
- Biometrical genetic models also provide a useful framework for the statistical analysis of data

Regression analysis for association

- Regression analysis is appropriate for testing and estimating the fixed effects of one of more genetic loci on a trait
- Linear or logistic regression are commonly used for continuous and binary traits, respectively. Ordinal logistic regression and Cox regression may be appropriate for ordinal or time-to-event data, respectively.
- With appropriate coding of genetic effects, main effects and interactions can be directly tested and estimated
- For example, coding the genotypes additively as aa, Aa and AA as 0, 1 and 2, the regression coefficient of trait on genotype can be directly interpreted as the allelic effect of A.
- Similarly, coding the dominance of aa, Aa and AA as p^2, $-p(1-p)$, and $(1-p)^2$ will directly test and estimate twice the dominance deviation

Variance components model for linkage

- The contribution of the additive genetic variance of a particular locus to the covariance of a trait between relative pairs depends on the proportion of alleles IBD at that locus (π); higher π leads to higher trait covariance
- Other than MZ twins, parent-offspring pairs and unrelated individuals (but see later), there is variation of π across the genome
- The allows the covariance of the an unmeasured additive effect between relative pairs to be specified as a function of π at a specific locus. When π can be estimated from genotyped SNPs at the near the locus, then its effect on trait covariance can be specified as a random effect in a linear mixed model.
- Model fitting (e.g. by restricted maximum likelihood) then provide an estimate the additive genetic variance at the specific locus.
- This analysis is called variance components quantitative trait locus (QTL) linkage analysis, because the estimate would capture of effects of loci that are linked to the specific locus, as well as those of the specific locus itself.
- The use of sib pairs for variance components QTL linkage analysis was a popular approach in the late 1990s. This is seldom used now because of low statistical power when effect size is small.

Variance components model for heritability

- Instead of using the variation of π across the genome for a given class of genetic relationship, we can exploit the differences of E(π) for different relationships to estimate the total additive genetic variance of a trait (which when expressed as a proportion of the total trait variance is the narrow-sense heritability).
- The classic twin design can be analyzed by a variance components model, where the additive genetic effects are specified to have an correlation of 1 for MZ twins and 0.5 for DZ twins (since E(π)=1 for MZ twins, 0.5 for DZ twins).
- A dominance component can be specified by specifying correlation of 1 for MZ twins and 0.25 for DZ twins (since Prob(π=1)=1 for MZ twins and 0.25 for DZ twins).
- Shared environmental influences for both MZ and DZ twins are specified to have correlation 1, whereas non-shared environment influences for both MZ and DZ twins are specified to have correlation 0.

Variance components models for heritability

- Random effects models can be applied to individuals not known to be related, to estimate heritability, since all pairs of individuals share common ancestors if their genealogies are traced back far enough.
- Such remote genetic relationships between pairs of individuals, effectively E(π), can be estimated from numerous SNPs across the genome.
- The estimated remote genetic relationships are then used to specify the covariance in additive genetic effects between pairs of individuals in a random effects model (first proposed in the GCTA method)
- This allows the heritability due to common variation in the genome to be estimated.

More complex structural equation models

- In principle, association, linkage and heritability analyses can all be combined in a single statistical model involving both fixed and random effects and their correlations and interactions, as well as complications such as assortative mating.
- Such an analysis can be formulated in a structural equations model framework, allowing joint estimation of the multiple effects, and hypothesis testing of different components of the model.