Notes for: Walsh and Lynch. Genetics and Analysis of Quantitative Traits

Thomas Battram

2020-05-28

Contents

Preface

This is a good book, but if I make it through the whole thing I deserve several medals and some cake.

Chapter 1

An overview of quantitative genetics

BORWANG!!!

This chapter just introduces the book and some simple concepts.

Chapter 2

Properties of distributions

ALSO BORWANG!

You can guess what this chapter was on and also how much of a hoot it was...

Chapter 3

Covariance, regression, and correlation

3.1 Covariance

Covariance is a measure of association and the covariance between x and y would be denoted by $\sigma(x, y)$. If x and y are independent then $\sigma(x, y) = 0$, BUT if $\sigma(x, y) = 0$, x and y aren't necessarily independent.

3.1.1 Useful identities for covariance

Covariance of x with itself = variance of x:

$$\sigma(x, x) = \sigma^2(x) \tag{3.1}$$

For constants (here represented by a) see (3.2) below

$$\begin{aligned} \sigma(a, x) &= 0 \\ \sigma(ax, y) &= a\sigma(x, y) \\ \sigma^2(a, x) &= a^2\sigma^2(x) \\ \sigma[(a + x), y] &= \sigma(x, y) \end{aligned} \tag{3.2}$$

The covariance of 2 sums can be written as the sum of covariances, i.e. just multiply out the brackets:

$$\sigma[(x + y), (w + z)] = \sigma(x, w) + \sigma(x, z) + \sigma(y, w) + \sigma(y, z) \tag{3.3}$$

Variance of a sum is sum of variances and covariances:

$$\sigma^2(x+y) = \sigma^2(x) + 2\sigma(x,y) + \sigma^2(y) \tag{3.4}$$

3.2 Least squares linear regression

Linear model:

$$y = \alpha + \beta x + e \tag{3.5}$$

Continuing on, α and β will be the true population values and a and b will be the intercept and slope for the line of best fit derived from observed data. The derivation of a and b using the least-squares model can be found on pages 39-41. Buuut, who cares about that, here are the results:

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ b &= \frac{Cov(x,y)}{Var(x)} \end{aligned} \tag{3.6}$$

3.2.1 Properties of least squares

6 in the book, just writing down important/not obvious ones.

- The mean residual ($\bar{e}$) is 0
- Residual errors are uncorrelated with predictor variable x (see book for why)
  - BUT e and x may not be independent if the relationship between x and y is non-linear. If it is truly non-linear $E(e|x)! = 0$
- Variance of e can vary with x, in this situation the the regression is said to display heteroscedasticity (see Figure 3.4 for great illustration)
- The regression of y on x is different to the regression of x on y!

3.3 Correlation

Correlation coefficient between x and y:

$$r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} \tag{3.7}$$

The correlation coefficient is a dimensionless measure of association and it is symmetrical (i.e. $r(x,y) = r(y,x)$).

Scaling x or y by constants does not change the correlation coefficient, but it does affect variances and covariances.

The correlation coefficient is a standardised regression coefficient -> the regression coefficient resulting from rescaling x and y such that each has unit variance).

r^2 assumes $E(y|x)$ is linear!

3.4 Differential selection (brief intro)

The directional selection differential, S, is the difference between the mean phenotype within that generation before selection (μ_s) and the mean phenotype within that generation after (μ) selection.

$$S = \mu_s - \mu \tag{3.8}$$

If all individuals have equal fertility and viability then selecting individuals won't change anything so $\mu_s = \mu$ and $S = 0$.

If $W(z)$ is the probability that individuals with phenotype z survive to reproduce and $p(z)$ is the density of z (pretty much means distribution) before selection, then the density after selection is:

$$p_s(z) = \frac{W(z)p(z)}{\int W(z)p(z)dz} \tag{3.9}$$

The denominator here is the mean individual fitness ($\bar{W}$). The relative fitness of z is $w(z) = \frac{W(z)}{\bar{W}}$.

After some sweet derivation (see page 46), you finish with:

$$S = \sigma[z, w(z)] \tag{3.10}$$

Therefore the directional selection is equivalent to the covariance of the phenotype and the relative fitness.

If you regress offspring phenotype on the midparent phenotype and that relationship is linear with slope β, a change in mean midparent phenotype induces an expected change in mean phenotype across generations equal to:

$$\begin{aligned} \Delta\mu &= \mu_0 - \mu \\ &= \beta(\mu_s - \mu) \\ &= \beta S \end{aligned} \tag{3.11}$$

This is the breeders' equation!

3.5 Correlation between genotype and phenotype (brief intro)

Only when there is no gene-environment interaction is the variance explained by genetics (broad-sense heritability) the equation below:

$$H^2 = \frac{\sigma_G^2}{\sigma_z^2}, \tag{3.12}$$

where z is the phenotype and G is the sum of the total effects (not just additive) at all loci on the trait.

The slope of a midparent-offspring regression provides an estimate of the proportion of the phenotypic variance that is attributable to additive genetic factors (the narrow-sense heritability).

$$h^2 = \frac{\sigma_A^2}{\sigma_z^2} \tag{3.13}$$

So as h^2 is just the regression of offspring phenotype on midparent phenotype it can actually be used in the breeders' equation!

$$\Delta\mu = h^2 S \tag{3.14}$$

So the narrow-sense heritability can be thought of as the efficiency of the response to selection. If $h^2 = 0$ there can be no evolutionary change regardless of strength of selection. Although this should be obvious because if h^2 is 0 then there is clearly no passing of genetic material onto the next generation that is influencing that trait.

3.6 End of chapter questions

1. True or false, if $\sigma(x, y) = 0$, x and y are independent
2. Finish these equations:

$$\begin{aligned} \sigma[(x+y),(w+z)] &= ... \\ \sigma^2(a,x) &= ... \\ \sigma[(a+x),y] &= ... \\ \sigma^2(x+y) &= ... \end{aligned} \tag{3.15}$$

3. Give 2 properties of residuals from least-squares regression
4. What is heteroscedasticity?

5. True or false, $y \sim x$ will always give the same effect estimate as $x \sim y$
6. Give an assumption of the correlation coefficient
7. How can you work out h^2 from trio data?
8. Give two definitions of the breeders' equation.

Chapter 4

Properties of single loci

4.1 Introduction

too easy

4.2 Allele and genotype frequencies

too easy

4.3 The transmission of genetic information

4.3.1 The Hardy-Weinberg principle

$$p^2 + 2pq + q^2 = 1 \tag{4.1}$$

where p = allele frequency of first allele at a locus and q = allele frequency of the second allele at that same locus.

Assumptions of H-W:

- No selection
- No mutation
- Random mating
- No differential migration
- No random drift

Even though these assumptions will never be met completely in the real world, for the majority of the time the H-W prinicple holds regardless.

Assuming assumptions are met, 2 important points from H-W:

1. It takes no more than a single generation to equilibriate and stabilize gene frequencies in the two sexes.
2. Only one additional generation is required for the stabilisation of the genotype frequencies into the predictible Hardy-Weinberg proportions.

4.3.2 Sex-linked loci

Alleles on sex chromosomes in diploid organisms are obviously different. Sons can only receive an X chromosome from their mother so the frequency of X linked loci in the sons is equal to that of their mothers. Daughters receive both an X chromosome from Mum + from Dad.

Overall this means allele frequencies oscilate around an equilibrium state, but continually get closer to that state over the generations (see Figure 4.2 and page 56 for equation).

4.3.3 Polyploidy

Skipped over this section because it's not relevant to human quant gen. Buuut, essentially it just details how to derive allele frequencies under a certain case of polyploidy. Also, it should be noted that of course H-W does not hold under polyploidy!

4.3.4 Age structure

Age structure also complicates our idealised model of H-W. In populations composed of several age classes, the generations overlap, and this causes the approach of genotype frequencies towards the H-W expectations to be gradual (rather than just by 1 or 2 generations), even in the case of an autosomal locus. Doesn't explain this very much, but it's covered elsewhere. Importantly, when newly founded populations have significant age structure, fluctuations in both gene and genotoype frequencies may occur for a substantial period of time even in the abscence of selection!

4.3.5 Testing for Hardy-Weinberg proportions

Says in the book that LRT can be used to test for departures from HWE and it can, but another common method is the chi-squared test and in PLINK they used Haldane's exact test which is apparently analogus to Fisher's exact test (papier on it) (can also use Fisher's exact test if the sample size is tiny and the allele is rare.). Essentially, in a population, at a specific locus, you can calculate the allele frequencies (and from that expected genotype frequencies) from the observed genotype frequencies then test if there is a difference between the observed and expected values. LRT equation for it given on page 60. See code for some comparisons.

Should remember (as pointed out above), that just because some assumptions are violated, doesn't mean you'd get a departure from HWE!

4.4 Characterising the influence of a locus on the phenotype

If a trait is entirely influenced by a single locus then the genetic effect on that trait can be characterised pretty easily and the dominance and additive effects of the alleles can be calculated. So if a locus has genotypes B_1B_1, B_1B_2, B_2B_2, then the values given to these genotypes can be said to be: $-a$, $(1+k)a$ and $+a$. Now if you have genotype data at that locus and data on the trait you can work out the effect of the B_2 allele by taking the mean phenotypic value of individuals with B_2B_2 and subtracting the mean phenotypic value of individuals with B_1B_1 and dividing by 2 i.e.

$$B_{2eff} = \frac{p_{B2} - p_{B1}}{2} \tag{4.2}$$

where B_{2eff} is the effect of allele B_2, p_{B2} is the mean phenotypic value of individuals with B_2B_2 and p_{B1} is the mean phneotypic value of individuals with B_1B_1.

As $B_{2eff} = a$ you can then substitute this in to $(1+k)a$ to get the dominance coefficient k. Of course if $k = 0$ then there is no dominance (in reality you would calculate probability of dominance).

4.5 The basis of dominance

Confusing part... Don't really get the enzyme activity bit...

Main point (I think) is that new deleterious mutations are very likely to be recessive and new mutations with a slight deleterious effect interact in an almost entirely additive fashion (no dominance!).

4.6 Fisher's decomposition of the genotypic value

Recalling that the phenotypic value can be partitioned like so:

$$z = G + E \tag{4.3}$$

where z is the phenotype, G is the genotypic value and E is the environmental value.

The genotypic value of a specific locus can be partitioned into it's "expected" values based on there being only additive effects ($\hat{G}$) and the deviations from the expected values or dominance effects (δ). So for genotype B_iB_j:

$$G_{ij} = \hat{G_{ij}} + \delta_{ij} \tag{4.4}$$

This can be formalised (whatever the fuck that means) by regressing the genotypic values on the number of B_1 and B_2 alleles in the genotype (N_1 and N_2):

$$G_{ij} = \hat{G_{ij}} + \delta_{ij} = \mu_G + \alpha_1 N_1 + \alpha_2 N_2 \tag{4.5}$$

μ_G = the mean genotypic value in the population, α_1 and α_2 are the slopes of the regression, N_1 and N_2 are the number of B_1 and B_2 alleles. So the regression is:

$$G_{ij}\ N_2 + N_1 \tag{4.6}$$

By noting that for any individual, $N_1 = 2 - N_2$ you can reduce the multiple regression model into an easier to work with univariate model. Give it a go (use equation (4.5)):

$$\begin{aligned} G_{ij} &= \mu_G + \alpha_1(2 - N_2) + \alpha 2N_2 + \delta_{ij} \\ &= l + (\alpha_2 - \alpha_1)N_2 + \delta_{ij} \end{aligned} \tag{4.7}$$

where $l = \mu_G + 2\alpha_1$ is the intercept and the slope is $\alpha = \alpha_2 - \alpha_1$

If you plotted the genotypic value (G) against gene content (N_2 or number of B_2 alleles) and calculated residuals these residuals would be δ, the dominance deviation (see Figure 4.6).

The rest of the chapter uses this regression and what we know about genotype frequencies to derive a formula for the average effect of allelic substitution:

$$\alpha = a[1 + k(p_1 - p_2)] \tag{4.8}$$

where a = genotypic value of B_2 (see above), k is the dominance coefficient and p_1 and p_2 are the frequencies of B_1 and B_2. This value α represents the average change in genotypic value that results when a B_2 allele is randomly substituted for a B_1 allele. If no dominance ($k = 0$) then $\alpha = a$. Except in the case of additivity, the average effect of allelic substitution is not simply a function of the inherent physiological properties of the allele. It can only be defined in the context of the population!

4.7 Partioning the genetic variance.

Deriving variance of G:

$$\begin{aligned} G &= \hat{G} + \delta \\ \sigma^2_G &= \sigma^2(\hat{G} + \delta) \\ &= \sigma^2(\hat{G}) + 2\sigma(\hat{G} + \delta) + \sigma^2(\delta) \end{aligned}$$

The top equation is just like a regression, with δ being the residual error and we know that for least-squares there is no correlation between the residual error and the predictor. So there is no correlation between $\hat{G}$ and δ. Therefore:

$$\sigma^2_G = \sigma^2(\hat{G}) + \sigma^2(\delta)$$

OR more commonly

$$\sigma^2_G = \sigma^2_A + \sigma^2_D \tag{4.9}$$

σ^2_A is the variance of G explained by regression on N_2 (or N_1), and σ^2_D is the residual variance of that regression. The variance of the additive and dominance effects!

For a diallelic locus we can do some rearranging of equations in Table 4.1 of the book and get these equations:

$$\sigma^2_A = 2p_1p_2\alpha^2 \tag{4.10}$$

$$\sigma^2_D = (2p_1p_2ak)^2 \tag{4.11}$$

From these we can clearly see that both components depend on allele frequencies, the dominance coefficient and the homozygous effect (remember α is just the slope of the G N_2 regression!).

By plotting how genetic variance changes with gene frequency under different scenarios (see 4.1). You see some interesting patterns. Firstly, at a single diallelic locus, you see that σ^2_A reaches it's peak when $p_1 = p_2 = 0.5$. Secondly, it's clear that, even in the case of overdominance (which is rare!), additive genetic variance will almost always be much higher than genetic variance from dominance effects, even when the frequency of the dominant allele is high.

4.8 Additive effects, average excesses and breeding values

The dominance deviation of a parent, which is a function of the interaction between the two parental alleles, is eliminated when gametes are produced. Thus,

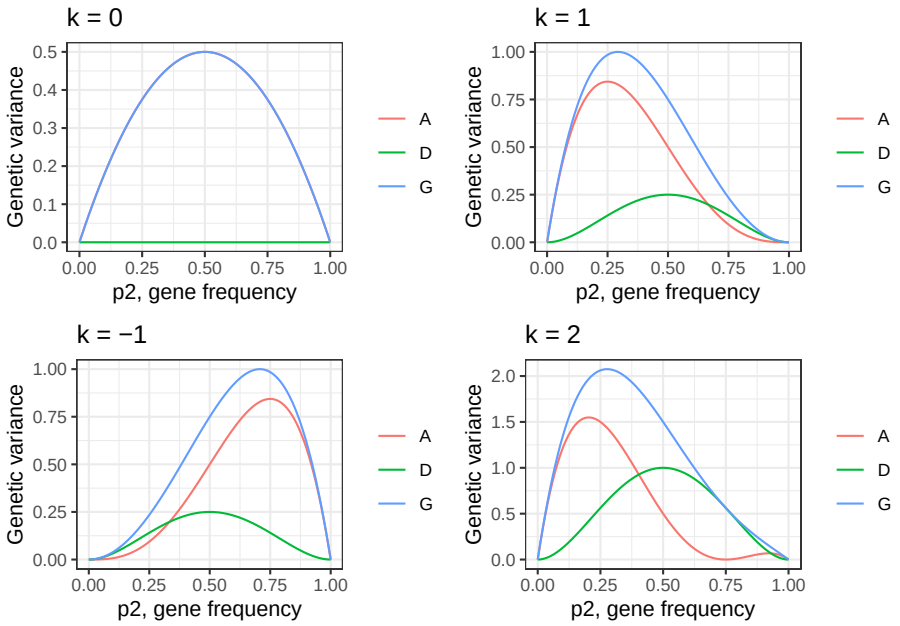

Figure 4.1: The dependence of components of genetic variance at a locus on the frequency of the B_2 allele. a is set to be one, which scales the vertical axes so that for any particular case, the actual variances are obtainable by multiplying by a^2.

one can think of $\hat{G}$ and δ as the heritable and nonheritable components of an individual's genotypic value.

Fisher proposed two different measures of the effect of an allele: one being the additive effect (α_i) and then the average excess α_i^x. The average excess α_2^x of allele B_2 is the difference between the mean genotypic value of individuals carrying at least one copy of B_2 and the mean genotypic value of a random individual from the entire population:

$$\alpha_2^x = (G_{12}P_{12|2} + G_{22}P_{22|2}) - \mu_G \tag{4.12}$$

where P_{ij} is the conditional probability of a B_iB_j genotype given that one allele is B_i. Under random mating $P_{ij|i} = p_j$ (p_j = frequency of allele B_j). THINK ABOUT HARDY-WEINBERG AND IT MAKES SENSE!

So under random mating,

$$\alpha_2^x = G_{12}p_1 + G_{22}p_2 - \mu_G \tag{4.13}$$

$G_{12} = a(1+k)$ and $G_{22} = 2a$. By substituting these into the equation above for α_1^x and α_2^x and then calculating α_1 and α_2 (additive effects) by the method previously mentioned (regressing genotypic value G on the number of B_2 alleles, N_2), we will see they're equivalent (shown on page 72):

$$\begin{aligned} \alpha_2 &= p_1\alpha \\ \alpha_1 &= -p_2\alpha \end{aligned} \tag{4.14}$$

The breeding value of an individual (A) is the sum of the additive effects of its genes. So the breeding value of a B_1B_1 homozygote is just $2\alpha_1$. In randomly mating populations the breeding value of a genotype is equivalent to twice the expected deviation of its offspring mean phenotype from the population mean. Soooo, no genotype information is needed to calculate the breeding value. All we have to do is mate an individual to many randomly chosen individuals from the population and taking twice the deviation of its offspring mean from the population mean. EASY IN HUMANS!!!

In Chapter 13 this will be discussed wrt candidate gene studies.

4.9 Extensions for multiple alleles and non random mating

So this section seems mostly unrelevant as we're unlikely to deal with situations with more than 2 alleles. Non-random mating could be encountered if we're

interested in some phenotypes (e.g. alcohol intake). Buuuut, it's still good to note some of the generalised equations for what we've been discussing so far in the chapter.

4.9.1 Average excess

When n alleles are present, the average excess, α_i^x, for any allele B_i is given by

$$\alpha_i^x = \sum_{j=1}^{n} P_{ij|i} G_{ij} - \mu_G \tag{4.15}$$

Remember, under random mating $P_{ij|i} == p_j$

4.9.2 Additive effects

The genotypic value can also be obtained using regression as before, but in it's generalised form is a multivariate regression. For n alleles

$$G = \mu_G + \sum_{i=1}^{n} \alpha_i N_i + \delta \tag{4.16}$$

After some re-arranging can derive the regression coefficients and finally end with

$$\alpha_i = \sum_{j=1}^{n} p_j G_{ij} - \mu_G \tag{4.17}$$

i.e. under random mating, the average effects (α_i) are equal to the conditional mean deviations from the mean genotypic value of the population (μ_G).

For non-random mating we need the inbreeding coefficient, f to define our genotype frequencies:

$$\begin{aligned} P_{ii} &= (1-f)p_i^2 + f p_i \\ P_{ij} &= 2(1-f)p_i p_j \end{aligned} \tag{4.18}$$

Unsure of why, but this means

$$\alpha_i = \frac{\alpha_i^x}{1+f} \tag{4.19}$$

so f is the fractional reduction of heterozygote frequencies relative to those expected under random mating. This means you can kind of do a test for

random mating by checking heterozygote and homozygote frequencies in a population!

4.9.3 Additive genetic variance

The additive genetic variance across n alleles is

$$\sigma_A^2 = 2\sum_{i=1}^{n} p_i \alpha_i \alpha_i^x \tag{4.20}$$

In general inbreeding inflates the additive genetic variance by causing correlations among the effects of alleles within the same individuals.

The broad sense heritability, even under scenarios of non-random mating can be given by

$$\sigma_G^2 = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij}) \tag{4.21}$$

although it should be noted that the definitions of α_i and $\delta_i j$ change with the degree of inbreeding! Random mating means α_i and α_j are uncorrelated so we get back to the good old equation

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

Importantly, under random mating, σ_A^2 is equivalent to the variance of breeding values of individuals in the population.

Summarising some key terms

The homozygous effect, a, and the dominance coefficient, k, are intrinsic properties of allelic products. They are not functions of allele frequencies, but may vary with genetic background

The additive effect, α_i, and the average excess, α_i^x, are properties of alleles in a particular population. They are functions of a, k and genotype frequencies (p_i).

The breeding value, A, is a property of a particular individual in reference to a particular population. It's equivalent to the sum of the additive effects of an individual's alleles.

The additive genetic variance, σ_A^2 is a property of a particular population. It is equivalent to the variance of the breeding values of individuals within the population.

4.10 End of chapter questions

1. What is the Hardy-Weinberg principle and what are it's assumptions?
2. What does the H-W principle mean for gene and genotype frequencies across generations?
3. What is age structure and how does it affect HWE?
4. How can you test for HWE?
5. Are deliterious mutations likely to be dominant or recessive?
6. Assuming a trait was entirely influenced by a single locus, how could you calculate dominance and additive effects knowing the genotypes and phenotypes of the individuals in the sample?
7. What is the formula for the average effect of allelic substitution?
8. For a diallelic locus, what does the additive genetic variance and dominance genetic variance depend on?
9. How does the contribution of additive genetic variance to total genetic variance change when k varies?
10. What is the breeding value of an individual?
11. Define the additive genetic variance in the presence of n alleles
12. Learn the definitions of the key terms!

Chapter 5

Sources of genetic variation for multilocus traits

5.1 Epistasis

Epistasis describes the nonadditivity of effects between loci, i.e. the alleles of one loci influence the effects of another loci.

The genotypic value, G_{ijkl}, needs to take into account all the interaction terms that can arrive between loci, for two loci it's additive x additive effects ($\alpha\alpha$), additive x dominance effects ($\alpha\delta$), and dominance x dominance effects ($\delta\delta$). As the number of loci increases the number of interaction terms increase steadily e.g. $\alpha\alpha\alpha$ will be there for three loci.

5.2 A general least-squares model for genetic effects

This is just an extension of the one-locus linear model introduced in Chapter 4.

For this section, imagine we are interested in measuring the genetic effects of two loci, G_{ijkl}, which can easily be extended to more. The additive effect of an allele on a phenotype is just the phenotypic value in people with that allele minus the mean phenotypic value of the population. When considering epistatic effects we can define it in the same way.

$$\alpha_i = G_{i\ldots} - \mu_G \tag{5.1}$$

$G_{i\ldots}$ is just the conditional mean phenotype of individuals with allele i at the first locus without regard to the other allele at that locus or to the genotype at

the second locus. The other additive terms (for α_j, α_k, α_l) are defined in the same way. Within each locus, the mean value of average effects (weighted by allele frequency) = 0.

Dominance effects can be defined in a similar way, complete these equations by recalling (4.16):

$$\delta_{ij} = G_{ij..} - ... \tag{5.2}$$
$$\delta_{lk} = G_{..lk} - ... \tag{5.3}$$

Like with the additive effects, the mean dominance deviation at each locus is equal to zero.

Epistatic effect terms proceed in a similar fashion. Letting $G_{i.k.}$ be the mean phenotype of individuals with gene i at locus 1 and k at locus 2, without regard to the other two genes, the ikth additive x additive effect is:

$$(\alpha\alpha)_{ik} = G_{i.k.} - \mu_G - \alpha_i - \alpha_k \tag{5.4}$$

So $(\alpha\alpha)_{ik}$ is the deviation of the conditional mean $G_{i.k.}$ from the expectation based on the population mean μ_G and the additive effects α_i and α_k. An additive x dominance effect measures the interaction between an allele at one locus with a genotype of another locus (see equation 5.5 in book) and the dominance x dominance effect involves an interaction between the genotypes at each locus (see equation 5.6 in book).

The complete genotypic value, $G_{ijkl...}$ can be found in equation 5.7 in the book. These parameters depend on genotype frequencies in the population, but the mean value of each type of effect is always equal to zero.

The genotypic value of an individual is often impossible to quantify because of variation in the phenotype due to the environment, but the genotypic value for an individual equation can be extended to populations. Providing mating is random and segregation of loci is independent, there is no statistical relationship between the genes found within or among loci. So the total genetic variance is just the sum of the variance of the individual effects, simplified this is:

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD} + ... \tag{5.5}$$

... here and in other cases just symbolises more terms can be added if more than two loci are used.

Epistatic effects are expected to be common throughout the genome and Wright thought they were the rule, rather than exception. See example two in the book for calculations of epistatic effects and how much variance they contribute to

the overall genetic variance component. Overall, it is clear that even with large epistatic effects, additive genetic variance, σ^2_A will pretty much always (if not always) contribute to the vast majority of overall genetic variance σ^2_G. This is important for two reasons:

1. Variance components provide limited insight into the physiological mode of gene action, i.e. just because genetic variance is explained by additive effects (which means you essentially count each gene separately), it does not mean the interaction between genes is not important in terms of their function!
2. When interested in the variance of a trait that is explained by genetics, you can expect the vast majority of that variance to be explained by additive genetic effects, which makes things like estimating heritability far easier.

5.2.1 Extension to haploids and polyploids

Skipped this section as not relevant to humans.

5.3 Linkage

Genes of the same chromosome tend to be inherited as a group, a tendency that declines with increasing distance between the loci. Crossing-over during meiosis is responsible for this decline.

Difference between linkage and linkage disequilibrium

Loci are linked if they tend to be inherited together. If loci are correlated for any reason (don't need to be inherited together), they are in linkage disequilibrium. The census units for measuring linkage are gamete frequencies, so you can use an individual to estimate this. LD is measured across a population.

Can get linkage without LD, can get linkage and LD, and can get two correlated loci (LD) that aren't linked.

Even though you can get correlated loci for reasons other than linkage, the LD between linked loci are more likely to persist over time as seen in (5.8).

Under linkage equilibrium, the frequency of gametes is the product of allele frequencies, so for loci A and B,

$$Freq(AB) = Freq(A) * Freq(B) \tag{5.6}$$

So A and B are independent of each other.

Measure of disequilibrium is just the departure from this:

$$D_{AB} = Freq(AB) - Freq(A) * Freq(B) \tag{5.7}$$

D_{AB} can be positive or negative depending on whether A and B are in coupling (AB gametes are overrepresented) or repulsion (AB gametes are underrepresented) disequilibrium. D is often referred to as the coefficient of linkage disequilibrium (although can be non-zero without linkage!).

Selection, migration, mutation and drift can help maintain LD. Even without these forces, once LD is established it can be maintained for many generations (especially if loci are more tightly linked!).

Expected LD changes over time depend on the recombination fraction between loci, c. This value ranges from 0 to 0.5, where 0 essentially means the loci are inherited together and 0.5 is free recombination between loci.

If recombination frequency between the A and B loci is c, the disequilibrium in generation t is given by:

$$D(t) = (1 - c)^t D(0) \tag{5.8}$$

This equation is graphically displayed in 5.1.

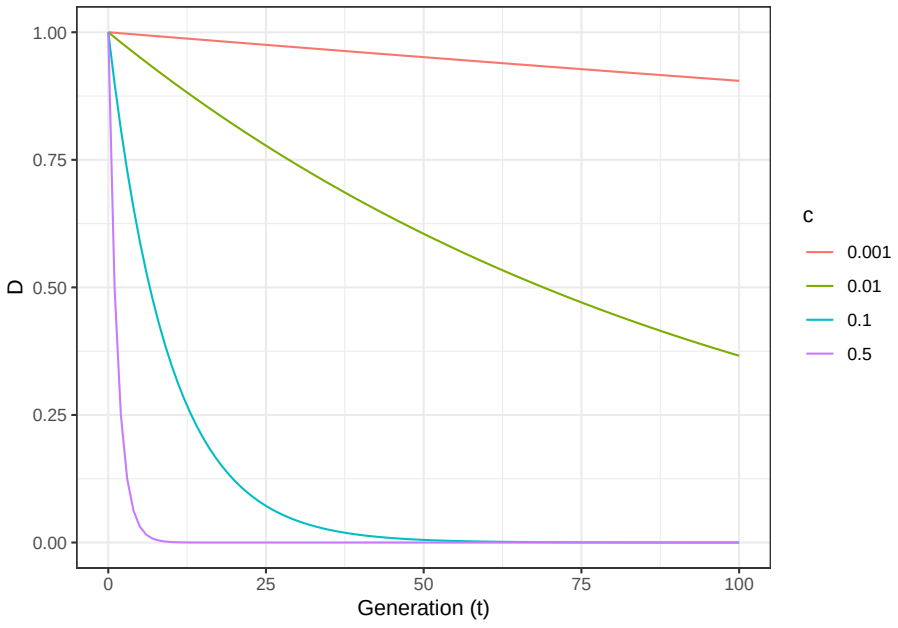

Figure 5.1: The decline, under random mating, of linkage disequilibrium when the initial value [D(0)] is set to 1 as a function of the recombination frequency, c.

To estimate D you can directly count gamete frequencies of an individual. This is not possible for most organisms though and usually you have to make-do with measured multilocus genotypes across a population. You can then workout gametes used to produce the genotypes (e.g. someone with an $AABb$ genotype would be formed from a AB gamete and aB gamete). However, with double heterozygotes you can't be sure if an $ABab$ individual was formed from AB and ab gametes or from Ab and aB gametes. Under random mating, it's not necessary to distinguish between coupling and repulsion heterozygotes so this doesn't really matter. In this case an estimate of D is given by:

$$\hat{D_{AB}} = \frac{N}{N-1}[\frac{4N_{AABB} + 2(N_{AABb} + N_{AaBB}) + N_{AbBb}}{2N} - 2\hat{p_A}\hat{p_B}] \qquad (5.9)$$

The equation for the sampling variance of D is shown on page 99.

Ideally you'd have 1000s of samples to achieve reasonable statistical power when estimating D using multilocus genotype frequencies.

5.4 Effect of disequilibrium of the genetic variance

The aggregate effects of gametic phase disequilibrium might be extensive for quantitative traits whose expression is based on large numbers of loci, even if the average level of disequilibrium between pairs of loci is relatively small. If genes with a positive influence on a character tend to be associated on some chromosomes, and those with a negative influence on others (coupling disequilbrium), the observed genetic variation will be inflated relative to the expectation under random assortment. The opposite will occur if "plus" alleles at one locus tend to be associated with "minus" alleles at another (repulsion disequilibrium). This is illustrated nicely by figure 5.6 in the book. Think of it this way, you're studying gene A, gene B and phenotype X. If upregulation of A leads to an increase in X and an upregulation of B leads to an increase in X by the same proportions, then if genetic variation always occurs so that whenever A is upregulated, B is downregulated by the same amount, then what will be observed at both of the loci is that variation at them is not associated with variation in X. This situation described is complete repulsion linkage (again, see figure 5.6 in book). This is assuming additive effects across loci. When there is no disequilibrium between loci, the variance at each locus is just $2pqa^2$, see (4.10).

As always, dominance effects muddy the waters, but here are the formalized multilocus analogs of (4.10) and (4.11):

$$\sigma_A^2 = 2\sum_{i=1}^{n} \alpha(i)^2 p_i q_i + 2\sum_{i=1}^{n}\sum_{j\neq 1}^{n} \alpha(i)\alpha(j) D_{ij} \tag{5.10}$$

$$\sigma_D^2 = 4\sum_{i=1}^{n} (a_i k_i p_i q_i)^2 + 4\sum_{i=1}^{n}\sum_{j\neq 1}^{n} a_i a_j k_i k_j D_{ij}^2 \tag{5.11}$$

where $\alpha(i)$ is the average effect of allelic substitution at the ith locus (defined in equation (4.8)).

Epistatic interactions make things crazy complicated!

In summary, the componenets of expressed genetic variance for quantitative traits can be partitioned into expected values under gametic phase equilibrium and deviations from these caused by disequilibrium. From the book: "When the disequilibrium covariance is negative, we refer to it as hidden genetic variance because it is subject to conversion to expressed genetic variance via the breakdown of gametic phase disequilibria". What I think this means: Negative disequilibrium covariance is just when correlated loci covary in a way so that when one increases a traits value, the other decreases it, this is what happens in repulsion linkage! The reason it's not just called repulsion linkage is because you can have covarying loci without them being linked. Sooo, what the book is saying here, is that if linkage disequilibrium (correlation) between the two loci that negatively covary is reduced, the amount of genetic variance they explain in the trait will increase!

5.4.1 The evidence

Hidden genetic variation is expected to be a natural consequence of stabilising selection, which favours linkage groups for their composite properties without regard to the alleles at individual loci. Theoretical work has suggested that stabilising selection encourages the development of substantial hidden genetic variance, potentially depressing the level of expressed genetic variance to 50% or less than its equilibrium expectation.

Of course selection doesn't always favour an increase in hidden genetic variance. Sometimes coupling selection is favoured, so that expressed genetic variance exceeds equilibrium expectations. In this case the disequilibrium covariance is positive, and recombination would be expected to result in a reduction in the expressed genetic variance.

Here we've just considered one trait, but of course the same thinking applies to selection upon multiple traits simultaneously. For example, in populations of insects (LIKE BEES) that exploit multiple host plants, one might expect a genetic correlation to evolve such that individuals prefer to feed on the plant species upon which they perform best. Such correlations could result from LD between a set of genes influencing preference and another influencing performance.

5.5 End of chapter questions

1. Define epistasis
2. Describe the terms that will be needed to define the genotypic value G_{ijkl}
3. Complete equations (5.2) and (5.3)
4. Give two important inferences from the fact the total genetic variance will mostly be attributable to the additive genetic variance, even if there are large dominance and epistatic effects
5. What is linkage and linkage disequilibrium?
6. What is the coefficient of linkage disequilibrium?
7. What can influence maintenance of LD? (5 things)
8. What is the relationship between linkage and LD over time?
9. Give two methods of estimating D
10. What are coupling disequlibrium and repulsion disequilibrium and how do they effect genetic variance?
11. Explain equations (5.10) and (5.11)
12. What is hidden genetic variance?
13. How is stabilising selection thought to influence hidden genetic variance? Why?

Chapter 6

Sources of Environmental Variation

This book divides environmental effects up into 2 different classes:

- General environmental effects: influential factors that are shared by groups of individuals (they include maternal effects in this)
- Special environmental effects: residual deviations from the phenotype expected based on genotype and general environmental effects

6.1 Extension of the linear model to phenotypes

Here we let E and e denote the contributions of general and specifc environmental effects and I denote GxE. Phenotype for kth individual of the ith genotype exposed to the jth general environmental effect can then be described as a linear function of 4 components:

$$z_{ijk} = G_i + I_{ij} + E_j + e_{ijk} \tag{6.1}$$

NOTE to fit an interaction term in R just put the terms into the model and multiply them, e.g. `lm(y ~ x*i, data=df)`.

Explaining some terms:

- I_{ij}, E_j, e_{ijk} are defined in a least-squares sense as deviations from lower-order expectations and so have mean values equal to zero
- $\mu_G = \overline{z}_{ijk}$ is the mean phenotype of all genotypes in the population
- G_i is the expected phenotype of the particular genotype i averaged over all possible environmental conditions

- $\mu_G + E_j$ is the mean phenotypic value expected if all genotypes were assayed in the jth macroenvironment
- $G_i + I_{ij} + E_j$ is the expected phenotype of individuals with genotype i in the jth macroenvironment
- e_{ijk} is the deviation from that expected phenotype so, as per least-squares rules, it isn't correlated with G_i, I_{ij} or E_j

I and e are uncorrelated with other variables (by construction). Remembering that the variance of a sum of uncorrelated variables is just the sum of the variances of each variable (and using equation (3.4)), we can define the phenotypic variance:

$$\sigma_P^2 = \sigma_G^2 + \sigma_I^2 + 2\sigma_{G.E} + \sigma_E^2 + \sigma^2 e \tag{6.2}$$

σ_I^2 is the GxE variance and $\sigma_{G.E}$ is the genotype-environment covariance. These terms are quite different. GxE is concerned variation in phenotypic response of specific genotypes within specific environments. Genotype-environment covariance is simply a measure of association between particular environments and genotypes. So, if individuals were randomly distributed across all environments, $\sigma_{G.E} = 0$, but σ_I^2 will be non-zero if genotypic and environmental effects are non-additive.

Maternal or paternal effects can cause genotype-environment covariance if there is correlation between parental genotype and ability to provision the young.

Genotype-environment covariance is often hard to estimate so and often contributes and unknown amount to estimates of genetic variance.

6.2 Special environmental effects

Two sources: internal developmental noise and external microenvironmental heterogeneity.

6.2.1 Within-individual variation

Can gain some information on within-individual variation by measuring the right and left components of a bilaterally symetrical individual. Pretty difficult to rule out external environmental contributions here though. Total variance of special environmental effects can be written as the sum of within-individual and among-individual environmental components

$$\sigma_e^2 = \sigma_{ew}^2 + \sigma_{ea}^2 \tag{6.3}$$

3 types of asymmetry:

- Directional - consistent bias in one direction (e.g. heart being more to the left)

- Antisymmetry - asymmetry is the rule rather than the exception, but it is nondirectional
- Fluctuating asymmetry - the difference between left and right measures is symmetrically distributed around a mean and mode of 0

Unbiased estimate of the within-individual variance for a trait:

$$\sigma^2_{ew} = \sum_{i=1}^{N} \frac{(r_i - l_i)^2}{2N} - \sigma^2_{em} \tag{6.4}$$

N is the number of individuals sampled, r_i and l_i are the right and left measures for the ith individual, and $\sigma^2 em$ is variance due to measurement error.

The effects of environmental stress on fluctuating asymmetry are fairly predictable - σ^2_{ew} tends to increase in extreme or novel environments. A study suggested humans suffering from malnutrition show increases in fluctuating asymmetry.

6.2.2 Developmental homeostasis and homozygosity

Lerner endorsed the idea that the degree of developmental stability is positively correlated with the overall level of individual heterozygosity. The usual mechanistic explanation is that heterozygosity acts as a buffer against environmental variation. Rest of this section discusses evidence for this hypothesis. It might be a useful exercise to think through how you'd do experiments to test the hypothesis based on the different components of variance that need to be considered. For now, here is the conclusion: "The acceptance of a general causal relationship between heterozygosity and developmental stability should be postponed until additional adequately designed experiments have been performed."

6.2.3 Repeatability

Variance among repeated measures on the same individual can only be due to environmental causes (or measurement errors), so information on the within-individual component of variance can provide some insight into the possible magnitude of the environmental variance for a trait. Time complicates things (phenotypes can vary within individuals at one time and across time), but that aside, the upper-bound estimate of the genetic variance of a trait is provided by:

$$\sigma^2_{G(max)} = \sigma^2_z - \sigma^2_{ew} \tag{6.5}$$

σ^2_z is an estimate for the total phenotypic variance for the trait.

Measurement error always inflates estimates of within-individual variance. As it contributes to total phenotypic variance, this cancels out in the equation above, but it's a pain because often we want to know the contribution of genetic variance to the total phenotypic variance. Repeated measures can help correct

for measurement error where the measure won't change over time - e.g. adult limb length. This is less tractable for measures that vary over time as you can't distinguish variation due natural organismal changes over time and those due to measurement error.

Expected value of $\sigma^2_{G(max)}$ is greater than the total genetic variance for the trait because it includes the among-individual component of variance due to the special environmental effects (σ^2_{ea}) and variance due to general environmental effects (σ^2_E). Letting var(e) denote the variance associated with measurement error, the repeatablitity is:

$$r = \frac{\sigma^2_z - \sigma^2_{ew}}{\sigma^2_z - \sigma^2_{em}} \tag{6.6}$$

and it provides an upper-bound estimate of the broad-sense heritability of a trait (H^2). The degree to which r exceeds H^2 depends on the magnitude of $\sigma^2_{ea} + \sigma^2_E$ relative to σ^2_{ew}. If all environmental variance is just within-individual variance and no measurement error is present, then r gives an unbiased estimate of H^2. Nice thing is that it gives an upper-bound regardless, so if r is low you can say that the environmental components must dominate. Unfortunately, repeatability is often computed as the correlation between two repeated measures (z_1 and z_2) on the same individuals:

$$r_F = \frac{\sigma(z_1, z_2)}{\sigma(z_1)\sigma(z_2)} \tag{6.7}$$

and as measurement error is contained in the denominator, it downwardly biases r_F. So we are no longer necessarily measuring the upper-bound of H^2 :(

6.3 General environmental effects of maternal origin

Before thinking of maternal effects on offspring, remember there is little evidence for intrauterine effects on complex traits in humans, quote from GDS's twitter: "virtually all disease mother-offspring and father-offspring risk concordance the same, except maternal small excess for epilepsy (intrauterine valproate?) and type 2 diabetes" (Link to tweet). Example MR paper.

Unless one runs an experiment where the environment of the past generation is the same as the current generation, one runs the risk that observed phenotypes are largely due to past generation. Similar thinking applies to estimates of heritability when there is assortative mating!

There are some striking examples of maternal effects in the wild and there are plenty of associations that have been drawn between maternal age and various

human traits too, for example the chances of Down's syndrome increases with maternal age.

Lack of data for multigenerational transmission of environmental effects – still a lack of data in 2020 in humans!

6.4 Genotype x environment interaction

This part gives examples of experiments done to detect GxE. In the examples, it was possible to make some inference as to the existence of GxE because members of the same genetic groups were evaluated under well-defined treatments. Of course, for natural populations, assigning individuals to discrete environmental groups is often impossible, so GxE becomes unmeasurable because any GxE will be confounded with the environmental source of variance. Interestingly, I think for Wes's GxE MR paper, they're able to apply the method in cases without clear discrete environments.

Chapter 7

Resemblance between relatives

If you ignore GxE you can express the phenotypic values of individuals x and y (recall equation (6.1)) simply as $Z_x = G_x + E_x + e_x$ and $Z_y = G_y + E_y + e_y$. This chapter is interested in the resemblence between relatives, so using these equations we can specify what the covariance between phenotypic values will be:

$$\begin{aligned}\sigma_z(x,y) &= \sigma[(G_x + E_x + e_x), (G_y + E_y + e_y)] \\ &= \sigma_G(x,y) + \sigma_{G.E}(x,y) + \sigma_{G.E}(y,x) + \sigma_E(x,y)\end{aligned} \tag{7.1}$$

Remember, e (special environmental effects) are derived from random residual deviations so are uncorrelated between individuals (think within-individual variation). You can design experiments so all terms with E in them have expected values of 0 and here we're going to assume that one individual's genotypic effects are not covarying with the others general environmental effects, i.e. $\sigma_{G.E}(x,y) = \sigma_{G.E}(y,x) = 0$. This boils everything nicely down to this simple equation:

$$\sigma_z(x,y) = \sigma_G(x,y) + \sigma_E(x,y) \tag{7.2}$$

$\sigma_G(x,y)$ will be the focus of things to come! Like genetic variance, the covariance can be split into components attributable to additive, dominance, and epistatic effects. Each term is simply one of the terms used to describe genetic variance (e.g. equation (5.5)), weighted by a coefficient that describes the joint distribution of effects in pairs of relatives.

Complications of estimating these coefficients include, non-random mating, LD, assortative mating, sex-linkage, maternal genetic effects and inbreeding.

7.1 Measures of relatedness

Relatedness can only be defined with respect to a specified frame of reference as all individuals are related (DUH). From here on the reference population is the base of the observed pedigree. So if the observed data is just trios, then the base population is the parents in those trios. If, data on grandparents is observed then they're the base population and so-on. Members of the base population are assumed to be unrelated. Also when discussing relatedness we refer to identity by descent (IBD), not identity by state (IBS).

Identity by descent and identity by state

Genes that are identical by descent are those that have been passed down by a common ancestor. The same gene from two individuals may share the same genetic sequence, making them identical by state, but if they do not derive from the same common ancestor they are not identical by descent.

So, genes that are identical by descent must, except for mutations, be identical by state, BUT genes that are identical by state might not be identical by descent.

7.1.1 Coefficients of identity

At a single locus in a diploid individual there are two alleles so with two individuals you have four alleles. Each allele is inherited singularly (a gamete only passes on one copy), so has it's own identity with each of the other three alleles. This means identity within individuals and between individuals can exist. This scenario gives rise to 15 different configurations of identity by descent. Individuals that contain pairs of alleles that are identical by descent are said to be inbred. Ignoring difference between maternally and paternally derived alleles, the number of IBD configurations reduces to nine. These range from a state where all four alleles are identical by descent (two inbred individuals that share a common ancestor) to a state where none of the alleles are identical by descent. In a large population with randomly mating individuals most states don't exist. The probabilities associated with each of the nine states are called the condensed coefficients of identity. Consider the case of a single gene for two non-inbred full sibs. There is a probability of 0.5 that both sibs inherit the same allele from their father and, independently, the probability they inherit the same allele from their mother is 0.5. So there is a probability of 0.25 that both pairs of alleles are identical by descent (i.e. the alleles inherited from the mother were the same and the alleles inherited from the father were the same.), there is a probability of 0.5 that just one pair is identical by descent and a probability of 0.25 that neither pair are identical by descent. All other states have a probability of 0.

7.1.2 Coefficients of coancestry and inbreeding

Suppose single genes (or alleles – side note alleles and genes will probably be used interchangably in this chapter) are drawn randomly from individuals x and y. The probability that these two genes are identical by descent, Θ_{xy}, is the coefficient of coancestry (can be called coefficient of consanguinity or coefficient of kinship). See figure 7.2 in the book for a graphical depiction of the nine IBD state classes and equation 7.2 in the book relates these states to Θ_{xy} –> each state is weighted by the conditional probability that a randomly drawn gene from x is identical by descent with a randomly drawn gene from y.

For an individual, z, their inbreeding coefficient (f_z) is equal to their parents coefficient of coancestry ($f_z = \Theta_{xy}$).

To derive Θ_{xy}, we first need to derive Θ_{xx} (may seem weird, but it ain't). If you took a gene with two alleles, A_1 and A_2 and you could know which parent each came from to distinguish them, if you drew one allele at random then replaced it and drew another you could draw A_1 twice, A_1 then A_2, A_2 then A_1 or A_2 twice. If they're not copies of the same allele (i.e. A_1 doesn't equal A_2) then if A_1 is drawn twice it must be identical by descent and the same goes for A_2. In this scenario, $\Theta_{xx} = (1/4)(1) + (1/4)(1)$. Of course, the individual could be inbred and so the probability that A_1 and A_2 are identical by descent is f_x. A general expression for the coefficient of coancestry of an individual with itself is given below in equation (7.3)

$$\Theta_{xx} = \frac{1}{4}(1 + f_x + f_x + 1) = \frac{1}{2}(1 + f_x) \tag{7.3}$$

Parent (p) and offspring (o) scenario now!! If neither are inbred (p's parents unrelated and they are unrelated to their mate), then that makes things simple. When drawing one of the two alleles from the mother and then one of the two alleles from the offspring, there is only one scenario in which they are the same. As each scenario has an equal probability of occurring, $\Theta_{po} = \frac{1}{4}$. If p is inbred, probability of their alleles being identical by descent is f_p. This is the same as the probability of the offspring allele being identical by descent to the maternal allele the offspring did not inherit. Probability of drawing inherited allele from offspring and allele not passed on from parent is (like the others) 1/4. Therfore, inbreeding inflates the coefficient of coancestry to $\Theta_{po} = \frac{f_p}{4} + \frac{1}{4} = \frac{1+f_p}{4}$. Complete inbreeding means $f_p = 1$ so $\Theta_{po} = \frac{1}{2}$. By thinking about probability of picking paternally derived allele and stuff you can add in f_o (see book page 136) and the general expression is given in equation (7.4) below.

$$\Theta_{po} = \frac{1}{4}(1 + f_p + 2f_o) \tag{7.4}$$

Often in the literature Θ_{po} is considered to simply be 1/4. So no inbreeding is assumed.

Full sibs time! m = mother, f = father, x = kid1 and y = kid2. So if m and f are not inbred or related themselves then there are two situations from which a child could inherit the same allele by descent. Either that allele has come from m or that allele has come from f. Both have the same probability so let's just look at m -> x and m -> y. The probability x and y receive the same maternal allele is 1/2 (i.e. the coefficient of coancestry of the mum with herself, Θ_{mm}). The probability of randomly drawing the maternally inherited allele from x is 1/2 and the same is true for y. Therefore the probability of drawing one allele from x and one from y, that are identical by descent, passed down from m is $\Theta_{mm}/4 = 1/8$. Adding contribution from el dado, we get $\Theta_{xy} = 1/4$ EASY! Appendix 2 contains path analysis, developed by Sewall Wright, that can derive these results. Now we allow for inbreeding of parents so introduce f_m and f_f. There are still only two paths that lead to alleles identical by descent in x and y (they're just more likely with inbreeding because alleles within the father and within the mother are more likely to be identical by descent, i.e. $\Theta_{mm} > 1/2$ and so is Θ_{ff} if there is inbreeding). Including these terms gives:

$$\Theta_{xy} = \frac{1}{4}(\Theta_{mm} + \Theta_{ff}) = \frac{1}{4}(\frac{1+f_m}{2} + \frac{1+f_f}{2}) = \frac{1}{8}(2 + f_m + f_f)$$

By taking into account inbreeding coefficients of kids, we end up with:

$$\Theta_{xy} = \frac{1}{8}(2 + f_m + f_f + 4\Theta_{mf})$$

Under random mating $\Theta_{xy} = 1/4$.

These techniques can be extended to more distant relatives and more complicated schemes of relatedness. The coefficient of coancestry is always the sum of a series of two types of paths between x and y. The first type of path leads from a single common ancestor to the two individuals of interest, while the second type passes through two remote ancestors that are related to each other. Neither type of path is allowed to pass through the same ancestor more than once. This procedure is summarised by equation (7.5) below.

$$\Theta_{xy} = \sum_i \Theta_{ii}(\frac{1}{2})^{n_i - 1} + \sum_j \sum_{j \neq k} \Theta_{jk}(\frac{1}{2})^{n_{jk} - 2} \tag{7.5}$$

where n_i is the number of individuals (including x and y) in the path leading from common ancestor i, and n_{jk} is the number of individuals (including x and y) on the path leading from two different but related ancestors, j and k.

Been assuming autosomal genes until this point! Sex-linked genes means we have to change things a bit. See book for deets (only 1 paragraph).

7.1.3 The coefficient of fraternity

Questions

Have fun answering these Gib!

Chapter 4

1. What the fuck are they talking about with the molecular basis of dominance? - page 63-64

Chapter 5

1. How do they calculate the variance of a phenotype explained by just the dominance effects? - page 91