

# Introduction to Genetic Analysis: Basic Concepts in Mendelian, Population and Quantitative Genetics

Bruce Walsh (jbwalsh@u.arizona.edu). July 2009.

2nd Annual NSF short course on Statistical Genetics, Honolulu

Note that there is additional material presented here not covered in my talk. These sections are denoted by \*

## OVERVIEW

As background for the rest of the talks in this workshop, our goal is to introduce some basic concepts from Mendelian genetics (the rules of gene transmission), population genetics (the rules of how genes behave in population), and quantitative genetics (the rules of transmission of complex traits, those with both a genetic and environmental basis).

### \* A Tale of Two Papers: Darwin vs. Mendel

The two most influential biologists in history, Darwin and Mendel, were contemporaries and yet the initial acceptance of their ideas suffered very different fates. Darwin was concerned with the evolution of complex traits (and hence concepts from population and quantitative genetics), while Mendel was concerned with the transmission of traits that had a simple genetic basis (often a single gene). Modern genetics and evolutionary theory was dependent on a successful fusion of their two key ideas (Mendel's that genes are discrete particles, Darwin's of evolution by natural selection). Against this background, it's interesting to consider the initial fates of both of their original papers.

In 1859, Darwin published his *Origin of Species*. It was an instant classic, with the initial printing selling out within a day of its publication. His work had an immediate impact that restructured biology. However, Darwin's theory of evolution by natural selection, as he originally presented it, was not without problems. In particular, Darwin had great difficulty dealing with the issue of inheritance. He fell back on the standard model of his day, **blending inheritance**. Essentially, both parents contribute fluids to the offspring, and these fluids contain the genetic material, which is blended to generate the new offspring. Mathematically, if  $z$  denotes the phenotypic value of an individual, with subscripts for father ( $f$ ), mother ( $m$ ) and offspring ( $o$ ), then blending inheritance implies

$$z_o = (z_m + z_f)/2 \quad (1a)$$

In 1867, in what was the first population genetics paper, the Scottish engineer Fleming Jenkin pointed out a serious problem with blending inheritance. Consider the variation in trait value in the offspring,

$$\text{Var}(z_o) = \text{Var}[(z_m + z_f)/2] = \frac{1}{2} \text{Var}(\text{parents}) \quad (1b)$$

Hence, under blending inheritance, half the variation is removed each generation and this must somehow be replenished by mutation. This simple statistical observation posed a very serious problem for Darwin, as (under blending inheritance) the genetic variation required for natural selection to work would be exhausted very quickly.

The solution to this problem was in the literature at the time of Jenkin's critique. In 1865, Gregor Mendel gave two lectures (delivered in German) on February 8 and March 8, 1865, to the Naturforschenden Vereins (the Natural History Society) of Brünn (now Brno, in the Czech Republic). The Society had been in existence only since 1861, and Mendel had been among its

founding members. Mendel turned these lectures into a (long) paper, "Versuche über Pflanzen-Hybriden" (Experiments in Plant Hybridization) published in the 1866 issue of the *Verhandlungen des naturforschenden Vereins*, (the *Proceedings of the Natural History Society in Brünn*). You can read the paper on-line (in English or German) at <http://www.mendelweb.org/Mendel.html>. Mendel's key idea: **Genes are discrete particles passed on intact from parent to offspring.**

Just over 100 copies of the journal are known to have been distributed, and one even found its way into the library of Darwin. Darwin did not read Mendel's paper (the pages were uncut at the time of Darwin's death), though he apparently did read other articles in that issue of the *Verhandlungen*. In contrast to Darwin, Mendel's work had no impact and was completely ignored until 1900 when three botanists (Hugo DeVries, Carl Correns, and Erich von Tschermak) independently made observations similar to Mendel and subsequently discovered his 1866 paper.

Why was Mendel's work ignored? One obvious suggestion is the very low impact journal in which the work was published, and his complete obscurity at the time of publication (in contrast, Darwin was already an extremely influential biologist before his publication of *Origins*). However, this is certainly not the whole story. One additional factor was that Mendel's original suggestion was perhaps too mathematical for 19th century biologists. While this may be correct, the irony is that the founders of statistics (the biometricians such as Pearson and Galton) were strong supporters of Darwin, and felt that early Mendelian views of evolution (which proceeds only by new mutations) were fundamentally flawed.

## BASIC MENDELIAN GENETICS

### Mendel's View of Inheritance: Single Locus

To understand the genesis of Mendel's view, consider his experiments which followed seven traits of the common garden pea (as we will see, seven was a very lucky number indeed). In one experiment, Mendel crossed a pure-breeding line with yellow peas to a pure-breeding line with green peas. Let  $P_1$  and  $P_2$  denote these two parental populations. The cross  $P_1 \times P_2$  is called the **first filial**, or  $F_1$ , population. In the  $F_1$ , Mendel observed that all of the peas were yellow. Crossing members of the  $F_1$ , i.e.  $F_1 \times F_1$  gives the **second filial** or  $F_2$  population. The results from the  $F_2$  were shocking – 1/4 of the plants had green peas, 3/4 had yellow peas. This **outbreak of variation**, recovering both green and yellow from yellow parents, blows the theory of blending inheritance right out of the water. Further, Mendel observed that  $P_1$ ,  $F_1$  and  $F_2$  yellow plants behaved very differently when crossed to the  $P_2$  (pure breeding green). With  $P_1$  yellows, all the seeds are yellow. Using  $F_1$  yellows, 1/2 the plants had yellow peas, half had green peas. When  $F_2$  yellows are used, 2/3 of the plants have yellow peas, 1/3 have green peas. Summarizing all these crosses,

Cross	Offspring
$P_1$	Yellow Peas
$P_2$	Green Peas
$F_1 = P_1 \times P_2$	Yellow Peas
$F_2 = F_1 \times F_1$	3/4 Yellow Peas, 1/4 green Peas
$P_1$ yellow $\times P_2$	Yellow Peas
$F_1$ yellow $\times P_2$	1/2 Yellow Peas, 1/2 green Peas
$F_2$ yellow $\times P_2$	2/3 Yellow Peas, 1/3 green Peas

What was Mendel's explanation of these rather complex looking results? **Genes are discrete particles, with each parent passing one copy to its offspring.**

Let an **allele** be a particular copy of a gene. In **diploids**, each parent carries two alleles for each gene (one from each parent). Pure Yellow parents have two  $Y$  (or yellow) alleles, and thus we can write their **genotype** as  $YY$ . Likewise, pure green parents have two  $g$  (or green) alleles, and a genotype of  $gg$ . Both  $YY$  and  $gg$  are examples of **homozygous** genotypes, where both alleles are the same. Each parent contributes one of its two alleles (at random) to its offspring, so that the

homozygous  $YY$  parent always contributes a  $Y$  allele, and the homozygous  $gg$  parent always a  $g$  allele. In the  $F_1$ , all offspring are thus  $Yg$  **heterozygotes** (both alleles differing). The **phenotype** denotes the trait value we observed, while the **genotype** denotes the (unobserved) genetic state. Since the  $F_1$  are all yellow, it is clear that both the  $YY$  and  $Yg$  genotypes map to the yellow pea phenotype. Likewise, the  $gg$  genotype maps to the green pea phenotype. Since the  $Yg$  heterozygote has the same phenotype as the  $YY$  homozygote, we say (equivalently) that the  $Y$  allele is **dominant** to  $g$  or that  $g$  is **recessive** to  $Y$ .

With this model of inheritance in hand, we can now revisit the above crosses. Consider the results in the  $F_2$  cross. Here, both parents are  $Yg$  heterozygotes. What are the probabilities of the three possible genotypes in their offspring?

$$\text{Prob}(YY) = \text{Pr}(Y \text{ from dad}) * \text{Pr}(Y \text{ from mom}) = (1/2)*(1/2) = 1/4$$

$$\text{Prob}(gg) = \text{Pr}(g \text{ from dad}) * \text{Pr}(g \text{ from mom}) = (1/2)*(1/2) = 1/4$$

$$\text{Prob}(Yg) = 1 - \text{Pr}(YY) - \text{Pr}(gg) = 1/2$$

Note that we can also compute the probability of a  $Yg$  heterozygote in the  $F_2$  as follows:

$$\begin{aligned} \text{Prob}(Yg) &= \text{Pr}(Y \text{ from dad}) * \text{Pr}(g \text{ from mom}) + \text{Pr}(g \text{ from dad}) * \text{Pr}(Y \text{ from mom}) \\ &= (1/4)(1/4) + (1/4)(1/4) = 1/2 \end{aligned}$$

Hence,  $\text{Prob}(\text{Yellow phenotype}) = \text{Pr}(YY) + \text{Pr}(Yg) = 3/4$ , as Mendel observed. This same logic can be used to explain the other crosses. (For fun, explain the  $F_2$  yellow  $\times$   $P_2$  results).

### The Genotype to Phenotype Mapping: Dominance and Epistasis

For Mendel's simple traits, the genotype to phenotype mapping was very straightforward, with complete dominance. More generally, we will be concerned with metric traits, namely those that we can assign numerical value, such as height, weight, IQ, blood chemistry scores, etc. For such traits, dominance occurs when alleles fail to act in an additive fashion, i.e. if  $\alpha_i$  is the average trait value of allele  $A_i$  and  $\alpha_j$  the average value of allele  $j$ , then dominance occurs when  $G_{ij} \neq \alpha_i + \alpha_j$ , namely that the genotypic value for  $A_i A_j$  does not equal the average value of allele  $i$  plus the average value of allele  $j$ .

In a similar fashion, **epistasis** is the non-additive interaction of genotypes. For example, suppose  $B-$  (i.e., either  $BB$  or  $Bb$ ) gives a brown coat color, while  $bb$  gives a black coat. A second gene,  $D$  is involved in pigment deposition, so that  $D-$  individuals deposit normal amounts of pigment, while  $dd$  individuals deposit no pigment. This is an example of epistasis, in that both  $B-$  and  $bb$  individuals are albino under the  $dd$  genotype. For metric traits, epistasis occurs when the two-locus genotypic value is not simply the sum of the two single-locus values, namely that  $G_{ijkl} \neq G_{ij} + G_{kl}$ .

### Mendel's View of Inheritance: Multiple Loci

For the seven traits that Mendel followed, he observed **independent assortment** of the genetic factors at different loci (genes), with the genotype at one locus being independent of the genotype at the second. Consider the cross involving two traits: round vs. wrinkled seeds and green vs. yellow peas. The genotype to phenotype mapping for these traits is  $RR, Rr$  = round seeds,  $rr$  = wrinkled seeds, and (as above)  $YY, Yg$  = yellow,  $gg$  = green. Consider the cross of a pure round, green ( $RRgg$ ) line  $\times$  a pure wrinkled yellow ( $rrYY$ ) line. In the  $F_1$ , all the offspring are  $RrYg$ , or round and yellow. What happens in the  $F_2$ ?

A quick way to figure this out is to use the notation  $R-$  to denote both the  $RR$  and  $Rr$  genotypes. Hence, round peas have genotype  $R-$ . Likewise, yellow peas have genotype  $Y-$ . In the  $F_2$ , the probability of getting an  $R-$  genotype is just

$$\text{Pr}(R- | F_2) = \text{Pr}(RR|F_2) + \text{Pr}(Rr|F_2) = 1/4 + 1/2 = 3/4$$

Since (under independent assortment) genotypes at the different loci are independently inherited, the probability of seeing a round, yellow  $F_2$  individual is

$$\text{Pr}(R- Y-) = \text{Pr}(R-) \cdot \text{Pr}(Y-) = (3/4) * (3/4) = 9/16$$

Likewise,

$$\Pr(\text{yellow, wrinkled}) = \Pr(rrY-) = \Pr(rr) \cdot \Pr(Y-) = (1/4) * (3/4) = 3/16$$

$$\Pr(\text{green, round}) = \Pr(R- gg) = \Pr(R-) \cdot \Pr(gg) = (3/4) * (1/4) = 3/16$$

$$\Pr(\text{green, wrinkled}) = \Pr(rrgg) = \Pr(rr) \cdot \Pr(gg) = (1/4) * (1/4) = 1/16$$

Hence, the four possible phenotypes are seen in a 9 : 3 : 3 : 1 ratio.

Under the assumption of independent assortment, the probabilities for more complex genotypes are just as easily found. Crossing  $AaBBcDD \times aaBbCcDd$ , what is  $\Pr(aaBBCCDD)$ ?

$$\begin{aligned}\Pr(aaBBCCDD) &= \Pr(aa) * \Pr(BB) * \Pr(CC) * \Pr(DD) \\ &= (1/2 * 1) * (1 * 1/2) * (1/2 * 1/2) * (1 * 1/2) = 1/2^5\end{aligned}$$

Likewise,

$$\Pr(AaBbCc) = \Pr(Aa) * \Pr(Bb) * \Pr(Cc) = (1/2) * (1/2) * (1/2) = 1/8$$

### Mendel was Wrong: Linkage

Shortly after the rediscovery of Mendel, Bateson and Punnett looked at a cross in peas involving a flower color locus (with the purple  $P$  allele dominant over the red  $p$  allele) and a pollen shape locus (with the long allele  $L$  dominant over the round allele  $l$ ). They examined the  $F_2$  from a pure-breeding purple long ( $PPLL$ ) and red round ( $ppll$ ) cross. The resulting genotypes, and their actual and expected numbers under independent assortment, were as follows:

Phenotype	Genotype	Observed	Expected
Purple long	$P-L-$	284	215
Purple round	$P-ll$	21	71
Red long	$ppL-$	21	71
red round	$ppll$	55	24

This was a significant departure from independent assortment, with an excess of  $PL$  and  $pl$  gametes over  $Pl$  and  $pL$ , evidence that the genes are **linked**, physically associated on the same chromosome.

### Interlude: Chromosomal Theory of Inheritance

Early light microscope work on dividing cells revealed small (usually) rod-shaped structures that appear to pair during cell division. These are **chromosomes**. It was soon postulated that Genes are carried on chromosomes, because chromosomes behaved in a fashion that would generate Mendel's laws — each individual contains a pair of chromosomes, one from each parent, and each individual passes along one random chromosome from each pair to its offspring. We now know that each chromosome consists of a single double-stranded DNA molecule (covered with proteins), and it is this DNA that codes for the genes.

Humans have 23 pairs of chromosomes (for a total of 46), consisting of 22 pairs of autosomes (chromosomes 1 to 22) and one pair of sex chromosomes — XX in females, XY in males. Humans also have another type of DNA molecule, namely the mitochondrial DNA genome that exists in tens to thousands of copies in the mitochondria present in all our cells. mtDNA is unusual in that it is strictly maternally inherited — offspring get only their mother's mtDNA.

### Linkage

If genes are located on different chromosomes they (with very few exceptions) show independent assortment. Indeed, peas have only 7 chromosomes, so was Mendel lucky in choosing seven traits at random that happen to all be on different chromosomes? (Hint, the probability of this is rather

small). However, genes on the same chromosome, especially if they are close to each other, tend to be passed onto their offspring in the same configuration as on the parental chromosomes.

Consider the Bateson-Punnett pea data, and let  $PL/pl$  denote that in the parent, one chromosome carries the  $P$  and  $L$  alleles (at the flower color and pollen shape loci, respectively), while the other chromosome carries the  $p$  and  $l$  alleles. Unless there is a **recombination** event, one of the two parental chromosome types ( $PL$  or  $pl$ ) are passed onto the offspring. These are called the **parental gametes**. However, if a recombination event occurs, a  $PL/pl$  parent can generate  $Pl$  and  $pL$  **recombinant gametes** to pass onto its offspring.

Let  $c$  denote the **recombination frequency** — the probability that a randomly-chosen gamete from the parent is of the recombinant type. For a  $PL/pl$  parent, the gamete frequencies are

Gamete Type	Frequency	Expectation under independent assortment
$PL$	$(1 - c)/2$	$1/4$
$pl$	$(1 - c)/2$	$1/4$
$pL$	$c/2$	$1/4$
$Pl$	$c/2$	$1/4$

Parental gametes are in excess, as  $(1 - c)/2 > 1/4$  for  $c < 1/2$ , while recombinant gametes are in deficiency, as  $c/2 < 1/4$  for  $c < 1/2$ . When  $c = 1/2$ , the gamete frequencies match those under independent assortment.

Suppose we cross  $PL/pl \times PL/pl$  parents. What are the expected genotype frequencies in their offspring?

$$\Pr(PPLL) = \Pr(PL|\text{father}) * \Pr(PL|\text{mother}) = [(1 - c)/2] * [(1 - c)/2] = (1 - c)^2/4$$

Likewise,  $\Pr(ppll) = (1 - c)^2/4$ . Recall from the Bateson-Punnett data that  $\text{freq}(ppll) = 55/381 = 0.144$ . Hence,  $(1 - c)^2/4 = 0.144$ , or  $c = 0.24$ .

A (slightly) more complicated case is computing  $\Pr(PpLl)$ . Two situations (linkage configurations) occur, as  $PpLl$  could be  $PL/pl$  or  $Pl/pL$ .

$$\begin{aligned} \Pr(PL/pl) &= \Pr(PL|\text{dad}) * \Pr(pl|\text{mom}) + \Pr(PL|\text{mom}) * \Pr(pl|\text{dad}) \\ &= [(1 - c)/2] * [(1 - c)/2] + [(1 - c)/2] * [(1 - c)/2] = (1 - c)^2/2 \end{aligned}$$

$$\begin{aligned} \Pr(Pl/pL) &= \Pr(Pl|\text{dad}) * \Pr(pL|\text{mom}) + \Pr(Pl|\text{mom}) * \Pr(pL|\text{dad}) \\ &= (c/2) * (c/2) + (c/2) * (c/2) = c^2/2 \end{aligned}$$

Thus,  $\Pr(PpLl) = (1 - c)^2/2 + c^2/2$ .

Generally, to compute the expected genotype probabilities, one needs to consider the frequencies of gametes produced by both parents. Suppose dad =  $Pl/pL$ , mom =  $PL/pl$ .

$$\Pr(PPLL) = \Pr(PL|\text{dad})\Pr(PL|\text{mom}) = [c/2] * [(1 - c)/2]$$

*Notation:* when the allele configurations on the two chromosomes are  $PL/pl$ , we say that alleles  $P$  and  $L$  are in **coupling**, while for  $Pl/pL$ , we say that  $P$  and  $L$  are in **repulsion**.

### Map Distances are Obtained from Recombination Frequencies via Mapping Functions

Construction of a **genetic map** involves both the ordering of loci and the measurement of distance between them. Ideally, distances should be additive so that when new loci are added to the map, previously obtained distances do not need to be radically adjusted. Unfortunately, recombination frequencies are not additive and hence are inappropriate as distance measures. To illustrate, suppose that three loci are arranged in the order  $A, B$ , and  $C$  with recombination frequencies  $c_{AB}$ ,  $c_{AC}$ , and

$c_{BC}$ . Each recombination frequency is the probability that an odd number of crossovers occurs between the markers, while  $1 - c$  is the probability of an even number (including zero). There are two different ways to get an odd number of crossovers in the interval  $A-C$ : an odd number in  $A-B$  and an even number in  $B-C$ , or an even number in  $A-B$  and an odd number in  $B-C$ . If there is no **interference**, so that the presence of a crossover in one region has no effect on the frequency of crossovers in adjacent regions, these probabilities can be related as

$$c_{AC} = c_{AB}(1 - c_{BC}) + (1 - c_{AB})c_{BC} = c_{AB} + c_{BC} - 2c_{AB}c_{BC}$$

This is **Trow's formula**. More generally, if the presence of a crossover in one region depresses the probability of a crossover in an adjacent region,

$$c_{AC} = c_{AB} + c_{BC} - 2(1 - \delta)c_{AB}c_{BC}$$

where the **interference parameter**  $\delta$  ranges from zero if crossovers are independent (no interference) to one if the presence of a crossover in one region completely suppresses crossovers in adjacent regions (complete interference).

Thus, in the absence of very strong interference, recombination frequencies can only be considered to be additive if they are small enough that the product  $2c_{AB}c_{BC}$  can be ignored. This is not surprising given that the recombination frequency measures only a part of all recombinant events (those that result in an odd number of crossovers). A **map distance**  $m$ , on the other hand, attempts to measure the total number of crossovers (both odd and even) between two markers. This is a naturally additive measure, as the number of crossovers between  $A$  and  $C$  equals the number of crossovers between  $A$  and  $B$  plus the number of crossovers between  $B$  and  $C$ .

A number of **mapping functions** attempt to estimate the number of cross-overs ( $m$ ) from the observed recombination frequency ( $c$ ). The simplest, derived by Haldane (1919), assumes that crossovers occur randomly and independently over the entire chromosome, i.e., no interference. Let  $p(m, k)$  be the probability of  $k$  crossovers between two loci  $m$  map units apart. Under the assumptions of this model, Haldane showed that  $p(m, k)$  follows a Poisson distribution, so that the observed fraction of gametes containing an odd number of crossovers is

$$c = \sum_{k=0}^{\infty} p(m, 2k + 1) = e^{-m} \sum_{k=0}^{\infty} \frac{m^{2k+1}}{(2k + 1)!} = \frac{1 - e^{-2m}}{2} \quad (2a)$$

where  $m$  is the expected number of crossovers. Rearranging, we obtain **Haldane's mapping function**, which yields the (Haldane) map distance  $m$  as a function of the observed recombination frequency  $c$ ,

$$m = -\frac{\ln(1 - 2c)}{2} \quad (2b)$$

For small  $c$ ,  $m \simeq c$ , while for large  $m$ ,  $c$  approaches  $1/2$ . Map distance is usually reported in units of **Morgans** (after T. H. Morgan, who first postulated a chromosomal basis for the existence of linkage groups) or as **centiMorgans** (cM), where  $100 \text{ cM} = 1 \text{ Morgan}$ . For example, a Haldane map distance of  $10 \text{ cM}$  ( $m = 0.1$ ) corresponds to a recombination frequency of  $c = (1 - e^{-0.2})/2 \simeq 0.16$ .

Although Haldane's mapping function is frequently used, several other functions allow for the possibility of crossover interference in adjacent sites. For example, human geneticists often use **Kosambi's mapping function** (1944), which allows for modest interference,

$$m = \frac{1}{4} \ln \left( \frac{1 + 2c}{1 - 2c} \right) \quad (2c)$$

### \*The Prior Probability of Linkage and Morton's Posterior Error Rate

Time for an interesting statistical aside motivated by linkage analysis. Morton in 1955 introduced the concept of a **Posterior Error Rate (PER)**, in the context of linkage analysis in humans. Morton's PER is simply the probability that a single significant test is a false positive. Framing tests in terms of the PER highlights the **screening paradox**, namely that "type I error control may not lead to a suitably low PER". For example, we might choose  $\alpha = 0.05$ , but the PER may be much, much higher, so that a test declared significant may have a much larger probability than 5% of being a false-positive. The key is that since we are *conditioning on the test being significant* (as opposed to conditioning on *the hypothesis being a null*, as occurs with  $\alpha$ ), this could include either false positives or true positives, and the relative fractions of each (and hence the probability of a false positive) is a function of the single test parameters  $\alpha$  (the Type I error) and  $\beta$  (the Type II error), and fraction of null hypotheses,  $\pi_0$ . To see this, apply Bayes' theorem,

$$\Pr(\text{false positive} \mid \text{significant test}) = \frac{\Pr(\text{false positive} \mid \text{null true}) \cdot \Pr(\text{null})}{\Pr(\text{significant test})}$$

Consider the numerator first. Let  $\pi_0$  be the fraction of all hypotheses that are truly null. The probability that a null is called significant is just the type I error  $\alpha$ , giving

$$\Pr(\text{false positive} \mid \text{null true}) \cdot \Pr(\text{null}) = \alpha \cdot \pi_0$$

Now, what is the probability that a single (randomly-chosen) test is declared significant? This event can occur because we pick a null hypothesis and have a type I error or because we pick an alternative hypothesis and avoid a type II error. Writing the power as  $1 - \beta$  ( $\beta$  being the type II error, the failure to reject an alternative hypothesis), the resulting probability that a single (randomly-draw) test is significant is just

$$\Pr(\text{significant test}) = \alpha\pi_0 + (1 - \beta)(1 - \pi_0)$$

Thus

$$PER = \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + (1 - \beta) \cdot (1 - \pi_0)} = \left( 1 + \frac{(1 - \beta) \cdot (1 - \pi_0)}{\alpha \cdot \pi_0} \right)^{-1}$$

In Morton's original application, since there are 23 pairs of human chromosomes, he argued that two randomly-chosen genes had a  $1/23 \simeq 0.05$  *prior probability of linkage*, i.e.,  $1 - \pi_0 = 0.05$  and  $\pi_0 = 0.95$ . Assuming a type I error of  $\alpha = 0.05$  and 80% power to detect linkage ( $\beta = 0.20$ ), this would give a PER of

$$\frac{0.05 \cdot 0.95}{0.05 \cdot 0.95 + 0.80 \cdot 0.05} = 0.54$$

Hence with a type-one error control of  $\alpha = 0.05\%$ , a random test showing a significant result ( $p \leq 0.05$ ) has a 54% chance of being a false-positives. This is because most of the hypotheses are expected to null — if we draw 1000 random pairs of loci, 950 are expected to be unlinked, and we expect  $950 \cdot 0.05 = 47.5$  of these to show a false-positive. Conversely, only 50 are expected to be linked, and we would declare  $50 \cdot 0.80 = 40$  of these to be significant, so that  $47.5/87.5$  of the significant results are due to false-positives.

### Molecular Markers

DNA from natural populations is highly **polymorphic**, in that if we looked at the DNA sequences of a particular region for a random sample from the population, no two sequences would be the same (except for identical twins). In humans, one polymorphism occurs roughly every 100 to 1000 bases, with any two random humans differing by over 20 million DNA differences. This natural variation in DNA provides us with a richly abundance set of **genetic** (or **molecular**) **markers** for gene mapping.

A variety of molecular tools have been used to detect these differences. For our purposes, we just consider the two most widely used types of markers, **SNPs (single nucleotide polymorphisms)** and **STRs (simple tandem repeats)**. SNPs result from the change in a single base, for example AAGGAA to AAGTAA. As a result, there are typically only two alleles in any population and the level of polymorphism between individuals can be modest. In contrast, STRs (also called **microsatellites**) are variations in the lengths of short repeated regions. For example, –ACACAC– vs. –ACACACAC– (e.g., AC<sub>3</sub> vs AC<sub>4</sub>). Such differences are easily scored with a variety of DNA sequencing technologies.

One advantage of STRs is that they have very high mutation rates (typically on the order of 1/500 vs. the 1/billions for SNPs) and hence there are usually a large number of alleles segregating in the population. As a result, STR sites are generally very polymorphic, making them ideal for certain types of mapping, such as within a family or extended pedigree. SNPs, on the other hand, have very low mutation rates, and since there is usually (at most) two alleles, the amount of polymorphism for a SNP is much less than a typical STR. Thus, they tend to be much less informative in pedigree studies. However, the low mutation rate means that the SNP alleles tend to be quite stable over long periods of time, making them ideal for population-level association studies where allele identities must remain unchanged over long periods of evolutionary time.

## BASIC POPULATION GENETICS

Mendelian genetics provides the rules of transmission from parents to offspring, and hence (by extension) the rules (and probabilities) for the transmissions of genotypes within a pedigree. More generally, when we sample a population we are not looking at a single pedigree, but rather a complex collection of pedigrees. What are the rules of transmission (for the population) in this case? For example, what happens to the frequencies of alleles from one generation to the next? What about the frequency of genotypes? The machinery of population genetics provides these answers, extending the mendelian rules of transmission within a pedigree to rules for the behavior of genes in a population.

### \* Allele and Genotype Frequencies

The frequency  $p_i$  for allele  $A_i$  is just the frequency of  $A_iA_i$  homozygotes plus half the frequency of all heterozygotes involving  $A_i$ ,

$$p_i = \text{freq}(A_i) = \text{freq}(A_iA_i) + \frac{1}{2} \sum_{i \neq j} \text{freq}(A_iA_j) \quad (3)$$

The 1/2 appears since only half of the alleles in heterozygotes are  $A_i$ . Equation 3 allows us to compute *allele* frequencies from *genotypic* frequencies. Conversely, since for  $n$  alleles there are  $n(n + 1)/2$  genotypes, the same set of allele frequencies can give rise to very different genotypic frequencies. To compute genotypic frequencies solely from allele frequencies, we need to make the (often reasonable) assumption of random mating. In this case,

$$\text{freq}(A_iA_j) = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_i p_j & \text{for } i \neq j \end{cases} \quad (4)$$

Equation 4 is the first part of the **Hardy-Weinberg theorem**, which allows us (assuming random mating) to predict genotypic frequencies from allele frequencies. The second part of the Hardy-Weinberg theorem is that allele frequencies remain unchanged from one generation to the next, *provided*: (1) infinite population size (i.e., no genetic drift), (2) no mutation, (3) no selection, and (4) no migration. Further, for an autosomal locus, a single generation of random mating gives genotypic frequencies in **Hardy-Weinberg proportions** (i.e., Equation 4) and the genotype frequencies forever remain in these proportions.



## Gamete Frequencies, Linkage, and Linkage Disequilibrium

Random mating is the same as gametes combining at random. For example, the probability of an  $AABB$  offspring is the chance that an  $AB$  gamete from the father and an  $AB$  gamete from the mother combine. Under random mating,

$$\text{freq}(AABB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (5a)$$

For heterozygotes, there may be more than one combination of gametes that gives rise to the same genotype,

$$\text{freq}(AaBB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(aB|\text{mother}) + \text{freq}(aB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (5b)$$

If we are only working with a single locus, then the gamete frequency is just the allele frequency, and under Hardy-Weinberg conditions, these do not change over the generations. However, when the gametes we consider involve two (or more) loci, recombination can cause gamete frequencies to change over time, even under Hardy-Weinberg conditions. At **linkage equilibrium**, the frequency of a multi-locus gamete is just the product of the individual allele frequencies. For example, for two and three loci,

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B), \quad \text{freq}(ABC) = \text{freq}(A) \cdot \text{freq}(B) \cdot \text{freq}(C)$$

In linkage equilibrium, the alleles at different loci are independent — knowledge that a gamete contains one allele (say  $A$ ) provides no information on the allele at the second locus in that gamete. More generally, loci can show **linkage disequilibrium** (LD), which is also called **gametic phase disequilibrium** as it can occur between unlinked loci. When LD is present,

$$\text{freq}(AB) \neq \text{freq}(A) \cdot \text{freq}(B)$$

Indeed, the disequilibrium  $D_{AB}$  for gamete  $AB$  is defined as

$$D_{AB} = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) \quad (6a)$$

Rearranging Equation 6a shows that the gamete frequency is just

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) + D_{AB} \quad (6b)$$

$D_{AB} > 0$  implies  $AB$  gametes are more frequent than expected by chance, while  $D_{AB} < 0$  implies they are less frequent.

We can also express the disequilibrium as a covariance. Code allele  $A$  as having value one, other alleles at this locus having value zero. Likewise, at the other locus, code allele  $B$  with value one and all others with value zero. The covariance between  $A$  and  $B$  thus becomes

$$\text{Cov}(AB) = E[AB] - E[A] \cdot E[B] = 1 \cdot \text{freq}(AB) - (1 \cdot \text{freq}(A)) \cdot (1 \cdot \text{freq}(B)) = D_{AB} \quad (7)$$

If the recombination frequency between the two loci is  $c$ , then the disequilibrium after  $t$  generations of recombination is simply

$$D(t) = D(0)(1 - c)^t \quad (8)$$

Hence, with loose linkage ( $c$  near  $1/2$ )  $D$  decays very quickly and gametes quickly approach their linkage equilibrium values. With tight linkage, disequilibrium can persist for many generations. As we will see in numerous instances throughout this course, it is the presence of linkage disequilibrium that allows us to map genes.

### \* The Effects of Population Structure

Many natural populations are *structured*, consisting of a mixture of several subpopulations. Even if each of the subpopulations are in Hardy-Weinberg proportions, samples from the entire population need not be. Suppose our sample population consists of  $n$  subpopulations, each in HW equilibrium. Let  $p_{ik}$  denote the frequency of allele  $A_i$  in population  $k$ , and let  $w_k$  be the frequency that a randomly-drawn individual is from subpopulation  $k$ . The expected frequency of an  $A_iA_i$  homozygote becomes

$$\text{freq}(A_iA_i) = \sum_{k=1}^n w_k \cdot p_{ik}^2 \quad (9a)$$

while the overall frequency of allele  $A_i$  in the population is

$$p_i = \sum_{k=1}^n w_k \cdot p_{ik} \quad (9b)$$

We can rearrange this as

$$\text{freq}(A_iA_i) = p_i^2 - \left( p_i^2 - \sum_{k=1}^n w_k \cdot p_{ik}^2 \right) = p_i^2 + \text{Var}(p_i) \quad (9c)$$

Hence, Hardy-Weinberg proportions hold only if  $\text{Var}(p_i) = 0$ , which means that all the subpopulations have the same allele frequency. Otherwise, the frequency of homozygotes is *larger* than we expect from Hardy-Weinberg (based on using the average allele frequency over all subpopulations), as

$$\text{freq}(A_iA_i) \geq p_i^2$$

While homozygotes are always over-represented, there is no clear-cut rule for heterozygotes. Following the same logic as above yields

$$\text{freq}(A_iA_j) = 2p_i p_j + \text{Cov}(p_i, p_j) \quad (10)$$

Here, the covariance can be either positive or negative.

Population structure can also introduce linkage disequilibrium (even among unlinked alleles). Consider an  $A_iB_j$  gamete and assume that linkage-equilibrium occurs in all subpopulations, then

$$\text{Freq}(A_iB_j) = \sum_{k=1}^n w_k \cdot p_{A_{ik}} \cdot p_{B_{jk}}$$

The expected disequilibrium is given by

$$\begin{aligned} D_{ij} &= \text{Freq}(A_iB_j) - \text{Freq}(A_i) \cdot \text{Freq}(B_j) \\ &= \sum_{k=1}^n w_k \cdot p_{A_{ik}} \cdot p_{B_{jk}} - \left( \sum_{k=1}^n w_k \cdot p_{A_{ik}} \right) \left( \sum_{k=1}^n w_k \cdot p_{B_{jk}} \right) \end{aligned} \quad (11)$$

Consider the simplest case of two populations, where the allele frequencies for  $A_i$  differ by  $\delta_i$  and by  $\delta_j$  for  $B_j$ . In this case, Equation 11 simplifies to

$$D_{ij} = \delta_i \cdot \delta_j \cdot [w_1(1 - w_1)] \quad (12)$$

Hence, in order to generate disequilibrium, the subpopulations must differ in allele frequencies at both loci. Further, the amount of disequilibrium is maximal when both subpopulations contribute equally ( $w_1 = 0.5$ ).

## BASIC QUANTITATIVE GENETICS

When there is a simple genetic basis to a trait (i.e., phenotype is highly informative as to genotype), the machinery of Mendelian genetics is straightforward to apply. Unfortunately, for many (indeed most) traits, the observed variation is a complex function of genetic variation at a number of genes plus environmental variation, so that phenotype is highly *uninformative* as to the underlying genotype. Developed by R. A. Fisher in 1918 (in a classic and completely unreadable paper that also introduced the term variance and the statistical method of analysis of variance), quantitative genetics allows one to make certain statistical inferences about the genetic basis of a trait given only information on the phenotypic covariances between sets of known relatives.

The machinery of quantitative genetics thus allows for the analysis of traits whose variation is determined by both a number of genes and environmental factors. Examples are traits influenced by variation at only a single gene that are also strongly influenced by environmental factors. More generally, a standard complex trait is one whose variation results from a number of genes of equal (or differing) effect coupled with environmental factors. Classic examples of complex traits include weight, blood pressure, and cholesterol levels. For all of these there are both genetic and environmental risk factors. Likewise, in the genomics age, complex traits can include molecular traits, such as the amount of mRNA for a particular gene on a microarray, or the amount of protein on a 2-D gel.

The goals of quantitative genetics are first to partition total trait variation into **genetic** (*nature*) vs. **environmental** (*nurture*) components. This information (expressed in terms of variance components) allows us to predict resemblance between relatives. For example, if a sib has a disease/trait, what are your odds? Recently, molecular markers have offered the hope of localizing the underlying loci contributing to genetic variation, namely the search for **QTL (quantitative trait loci)**. The ultimate goal of quantitative genetics in this post-genomic era is the prediction of phenotype from genotype, namely the deduction of the molecular basis for genetic trait variation. Likewise, in the genomics era we often speak of **eQTLs (expression QTLs)**, loci whose variation influences gene expression (typically the amount of mRNA for a gene on a microarray). As we will see, operationally QTLs involve a genomic region, often of considerable length. Thus the goal is to find **QTNs, for quantitative trait nucleotides**, the specific nucleotides (as opposed to genomic regions) underlying the trait variation.

### Dichotomous (Binary) Traits

While much of the focus of quantitative genetics is on continuous traits (height, weight, blood pressure), the machinery also applies to dichotomous traits, such as disease presence/absence. This apparently phenotypic simplicity can easily mask a very complex genetic basis.

Loci harboring alleles that increase disease risk are often called **disease susceptiblely** (or DS) loci. Consider such a DS locus underlying a disease, with alleles  $D$  and  $d$ , where allele  $D$  significantly increases disease risk. In particular, suppose  $\Pr(\text{disease} | DD) = 0.5$ , so that the **penetrance** of genotype  $DD$  is 50%. Likewise, suppose for the other genotypes that  $\Pr(\text{disease} | Dd) = 0.2$ ,  $\Pr(\text{disease} | dd) = 0.05$ . Hence, the presence of a  $D$  allele significantly increases your disease risk, but  $dd$  individuals can rarely display the disease, largely because of exposure to adverse environmental conditions. Such  $dd$  individuals showing the disease are called **phenocopies**, as the presence of the disease does not result from them carrying a high-risk allele. If the  $D$  allele is rare, most of the observed disease cases are environmental (from  $dd$ ) rather than genetically (from  $D-$ ) caused. For example, suppose  $\text{freq}(d) = 0.9$ , what is  $\text{Prob}(DD | \text{show disease})$ ? First, the **population prevalence**  $K$  (the frequency) of the disease is

$$\begin{aligned} K &= \text{freq}(\text{disease}) \\ &= \Pr(DD) * \Pr(\text{disease}|DD) + \Pr(Dd) * \Pr(\text{disease}|Dd) + \Pr(dd) * \Pr(\text{disease}|dd) \\ &= 0.12 * 0.5 + 2 * 0.1 * 0.9 * 0.2 + 0.92 * 0.05 = 0.0815 \end{aligned}$$

Hence, roughly 8% of the population shows the disease. **Bayes' theorem** states that

$$\Pr(b|A) = \frac{\Pr(A|b) * \Pr(b)}{\Pr(A)} \quad (13)$$

Applying Bayes' theorem (with  $A$  = disease,  $b$  = genotype),

$$\Pr(DD|\text{disease}) = \frac{\Pr(\text{disease}|DD) * \Pr(DD)}{\Pr(\text{disease})} = \frac{0.5 * 0.12}{0.0815} = 0.06$$

Hence, if we pick a random individual showing the disease, there is only a 6% chance that they have the high-risk ( $DD$ ) genotype. Likewise,  $\Pr(Dd | \text{disease}) = 0.442$ ,  $\Pr(dd | \text{disease}) = 0.497$ .

### Contribution of a Locus to the Phenotypic Value of a Trait

The basic model for quantitative genetics is that the **phenotypic value**  $P$  of a trait is the sum of a **genetic value**  $G$  plus an **environmental value**  $E$ ,

$$P = G + E \quad (14)$$

The genetic value  $G$  represents the average phenotypic value for that particular genotype if we were able to replicate it over the distribution (or **universe**) of environmental values that the population is expected to experience.

The genotypic value  $G$  is usually the result of a number of loci that influence the trait. However, we will start by first considering the contribution of a single locus, whose alleles are alleles  $Q_1$  and  $Q_2$ . We need a parameterization to assign genotypic values to each of the three genotypes, and there are three slightly different notations used in the literature:

	Genotypes		
	$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
Average Trait Value:	$C$	$C + a(1 + k)$	$C + 2a$
	$C - a$	$C + a + d$	$C + 2a$
		$C + d$	$C + a$

Here  $C$  is some background value, which we usually set equal to zero. What matters here is the difference  $2a$  between the two homozygotes,  $a = [G(Q_2Q_2) - G(Q_1Q_1)]/2$ , and the relative position of the heterozygotes compared to the average of the homozygotes. If it is exactly intermediate,  $d = k = 0$  and the alleles are said to be additive. If  $d = a$  (or equivalently  $k = 1$ ), then allele  $Q_2$  is completely dominant to  $Q_1$  (i.e.,  $Q_1$  is completely recessive). Conversely, if  $d = -a$  ( $k = -1$ ) then  $Q_1$  is dominant to  $Q_2$ . Finally if  $d > a$  ( $k > 1$ ) the locus shows **overdominance** with the heterozygote having a larger value than either homozygote. Thus  $d$  (and equivalently  $k$ ) measure the amount of dominance at this locus. Note that  $d$  and  $k$  are related by

$$ak = d, \quad \text{or} \quad k = \frac{d}{a} \quad (15)$$

The reason for using both  $d$  and  $k$  is that some expressions are simpler using one parameterization over another.

### Example: Apolipoprotein E and Alzheimer's age of onset

One particular allele at the apolipoprotein E locus (we will call it  $e$ , and all other alleles by  $E$ ) is associated with early age of onset for Alzheimer's. The mean age of onset for  $ee$ ,  $Ee$ , and  $EE$  genotypes are 68.4, 75.5, and 84.3, respectively. Taking these to be estimates of the genotypic values ( $G_{ee}$ ,  $G_{Ee}$ , and  $G_{EE}$ ), the homozygous effect of the  $E$  allele is estimated by  $a = (84.3 - 68.4)/2 = 7.95$ . The dominance coefficient is estimated by  $ak = d = G_{Ee} - [G_{ee} + G_{EE}]/2 = -0.85$ . Likewise,  $k = d/a = 0.10$ .

## Fisher's Decomposition of the Genotypic Value

Quantitative genetics as a field dates back to R. A. Fisher's brilliant (and essentially unreadable) 1918 paper, in which he not only laid out the field of quantitative genetics, but also introduced the term variance and developed the analysis of variance (ANOVA). Not surprisingly, his paper was initially rejected.

Fisher had two fundamental insights. First, that *parents do not pass on their entire genotypic value to their offspring, but rather pass along only one of the two possible alleles at each locus*. Hence, only part of  $G$  is passed on and thus we decompose  $G$  into component that can be passed along and those that cannot. Fisher's second great insight was that *phenotypic correlations among known relatives can be used to estimate the variances of the components of  $G$* .

Fisher suggested that the genotypic value  $G_{ij}$  associated with an individual carrying a  $Q_iQ_j$  genotype can be written in terms of the **average effects**  $\alpha$  for each allele and a **dominance deviation**  $\delta$  giving the deviation of the actual value for this genotype from the value predicted by the average contribution of each of the single alleles,

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij} \quad (16)$$

The predicted genotypic value is  $\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$ , where  $\mu_G$  is simply the average genotypic value,

$$\mu_G = \sum G_{ij} \cdot \text{freq}(Q_iQ_j)$$

Note that since we assumed the environmental values have mean zero,  $\mu_G = \mu_P$ , the mean phenotypic value. Likewise  $G_{ij} - \hat{G}_{ij} = \delta_{ij}$ , so that  $\delta$  is the residual error, the difference between the actual value and that predicted from the regression. Since  $\alpha$  and  $\delta$  represent deviations from the overall mean, they have expected values of zero.

You might notice that Equation 16 looks like a regression. Indeed it is. Suppose we have only two alleles,  $Q_1$  and  $Q_2$ . Notice that we can re-express Equation 16 as

$$G_{ij} = \mu_G + 2\alpha_1 + (\alpha_2 - \alpha_1)N + \delta_{ij} \quad (17)$$

where  $N$  is the number of copies of allele  $Q_2$ , so that

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g., } Q_1Q_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g., } Q_1Q_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g., } Q_2Q_2 \end{cases} \quad (18)$$

Thus we have a regression, where  $N$  (the number of copies of allele  $Q_2$ ) is the predictor variable, the genotypic value  $G$  the response variable,  $(\alpha_2 - \alpha_1)$  is the regression slope, and  $\delta_{ij}$  the residuals of the actual values from the predicted values. Recall from the standard theory of least-squares regression that the correlation between the predicted value of a regression ( $\mu_G + \alpha_i + \alpha_j$ ) and the residual error ( $\delta_{ij}$ ) is zero, so that  $\sigma(\alpha_i, \delta_j) = \sigma(\alpha_k, \delta_j) = 0$ .

To obtain the  $\alpha$ ,  $\mu_G$  and  $\delta$  values, we use the notation of

Genotypes:	$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
Average Trait Value:	0	$a(1+k)$	$2a$
frequency (HW):	$p_1^2$	$2p_1p_2$	$p_2^2$

A little algebra gives

$$\mu_G = 2p_1p_2a(1+k) + 2p_2^2a = 2p_2a(1+p_1k) \quad (19a)$$

Recall that the slope of a regression is simply the covariance divided by the variance of the predictor variable, giving

$$\alpha_2 - \alpha_1 = \frac{\sigma(G, N_2)}{\sigma^2(N_2)} = a[1+k(p_1-p_2)] \quad (19b)$$

See Lynch and Walsh, Chapter 4 for the algebraic details leading to Equation 19b. Since we have chosen the  $\alpha$  to have mean value zero, it follows that

$$p_i\alpha_1 - p_2\alpha_2 = 0$$

When coupled with Equation 19b this implies (again, see L & W Chapter 4)

$$\alpha_2 = p_1 a [1 + k (p_1 - p_2)] \quad (19c)$$

$$\alpha_1 = -p_2 a [1 + k (p_1 - p_2)] \quad (19d)$$

Finally, the dominance deviations follow since

$$\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j \quad (19e)$$

Note the important point that both  $\alpha$  and  $\delta$  are functions of allele frequency and hence change as the allele (and/or genotype) frequencies change. While the  $G_{ij}$  values remain constant, their weights are functions of the genotype (and hence allele) frequencies. As these change, the regression coefficients change.

### Average Effects and Additive Genetic Values

The  $\alpha_i$  value is the **average effect** of allele  $Q_i$ . Animal breeders are concerned (indeed obsessed) with the **breeding values** (BV) of individuals, which are related to average effects. The BV is also called the **additive genetic value**,  $A$ , and since our focus here is human genetics (not animal/plant breeding) we will stick with  $A$ . The additive genetic value associated with genotype  $G_{ij}$  is just

$$A(G_{ij}) = \alpha_i + \alpha_j \quad (20a)$$

Likewise, for  $n$  loci underlying the trait, the BV is just

$$A = \sum_{k=1}^n (\alpha_i^{(k)} + \alpha_k^{(k)}) \quad (20b)$$

namely, the sum of all of the average effects of the individual's alleles. Note that since the additive genetic values are functions of the allelic effects, they change as the allele frequencies in the population change.

So, why all the fuss over breeding/additive-genetic values? If the additive genetic value of one parent is  $A_1$ , and the other parent is chosen at random, then the average value of their offspring  $\mu_o$  is

$$\mu_o = \mu + \frac{A_1}{2}$$

where  $\mu$  is the population mean. Similarly, the expected value of the offspring given the additive of both parents is just their average,

$$\mu_o = \mu_G + \frac{A_1 + A_2}{2} \quad (21)$$

The focus on additive genetic values thus arises because they predict offspring means.

## Genetic Variances

Recall that the genotypic value is expressed as

$$G_{ij} = \mu_g + (\alpha_i + \alpha_j) + \delta_{ij}$$

The term  $\mu_g + (\alpha_i + \alpha_j)$  corresponds to the regression (best linear) estimate of  $G$ , while  $\delta$  corresponds to a residual. Recall from regression theory that the estimated value and its residual are uncorrelated, and hence  $\alpha$  and  $\delta$  are uncorrelated. Since  $\mu_G$  is a constant (and hence contributes nothing to the variance) and  $\alpha$  and  $\delta$  are uncorrelated,

$$\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij}) \quad (22)$$

Equation 22 is the contribution from a single locus. Assuming linkage equilibrium, we can sum over loci,

$$\sigma^2(G) = \sum_{k=1}^n \sigma^2(\alpha_i^{(k)} + \alpha_j^{(k)}) + \sum_{k=1}^n \sigma^2(\delta_{ij}^{(k)})$$

This is usually written more compactly as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \quad (23)$$

where  $\sigma_A^2$  is the **additive genetic variance** and represents the variance in breeding values in the population, while  $\sigma_D^2$  denotes the **dominance genetic variance** and is the variance in dominance deviations.

Suppose the locus of concern has  $m$  alleles. Since (by construction) the average values of  $\alpha$  and  $\delta$  for a given locus have expected values of zero, the contribution from that locus to the additive and dominance variances is just

$$\sigma_A^2 = E[\alpha_i^2 + \alpha_j^2] = 2E[\alpha^2] = 2 \sum_{i=1}^m \alpha_i^2 p_i, \quad \text{and} \quad \sigma_D^2 = E[\delta^2] = \sum_{i=1}^m \sum_{j=1}^m \delta_{ij}^2 p_i p_j \quad (24)$$

For one locus with two alleles, these become

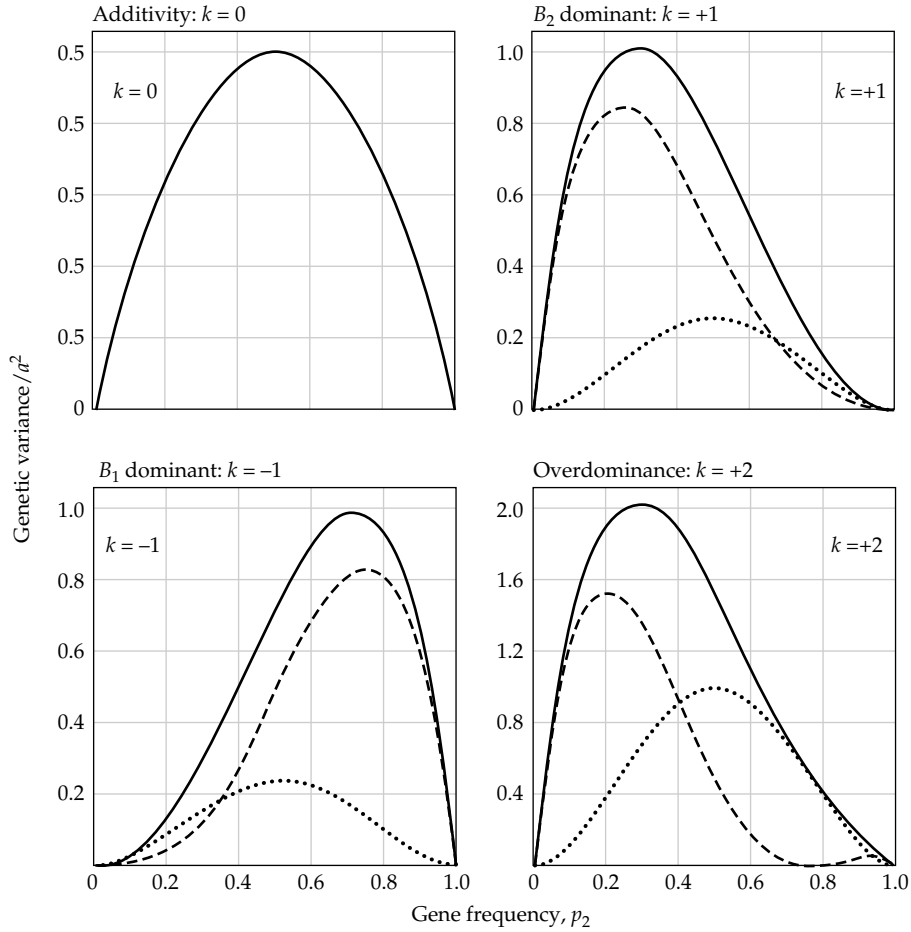
$$\sigma_A^2 = 2p_1 p_2 a^2 [1 + k(p_1 - p_2)]^2 \quad (25a)$$

and

$$\sigma_D^2 = (2p_1 p_2 ak)^2 \quad (25b)$$

The additive (dashed line), dominance (dotted line) and total ( $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$ , solid line) variance are plotted below for several different dominance relationships.

Note (from both the figures and from Equation 25) that there is plenty of additive variance even in the face of complete dominance. Indeed, dominance (in the form of the dominance coefficient  $k$ ) enters the expression for the additive variance. This is not surprising as the  $\alpha$  arise from the best-fitting line, which will incorporate some of the departures from additivity. Conversely, note that the dominance variance is zero if there is no dominance ( $\sigma_D^2 = 0$  if  $k = 0$ ). Further note that  $\sigma_D^2$  is symmetric in allele frequency, as  $p_1 p_2 = p_1(1 - p_1)$  is symmetric about 1/2 over (0,1).



## Epistasis

Epistasis, nonadditive interactions between alleles at different loci, occurs when the single-locus genotypic values do not add to give two (or higher) locus genotypic values. For example, suppose that the average value of a  $AA$  genotype is 5, while an  $BB$  genotype is 9. Unless the value of the  $AABB$  genotype is  $5 + 4 = 9$ , epistasis is present in that the single-locus genotypes do not predict the genotypic values for two (or more) loci. Note that we can have strong dominance within each locus and no epistasis between loci. Likewise we can have no dominance within each locus but strong epistasis between loci.

The decomposition of the genotype when epistasis is present is a straightforward extension of the no-epistasis version. For two loci, the genotypic value is decomposed as

$$\begin{aligned}
 G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kl}) \\
 &\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\
 &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\
 &\quad + (\delta\delta_{ijkl}) \\
 &= \mu_G + A + D + AA + AD + DD
 \end{aligned} \tag{26}$$

Here the breeding value  $A$  is the average effects of single alleles averaged over genotypes, the dominance deviation  $D$  the interaction between alleles at the same locus (the deviation of the single locus genotypes from the average values of their two alleles), while  $AA$ ,  $AD$  and  $DD$  represent the



(two-locus) epistatic terms.  $AA$  is the **additive-by-additive** interaction, and represents interactions between a single allele at one locus with a single allele at another.  $AD$  is the **additive-by-dominance** interaction, representing the interaction of single alleles at one locus with the genotype at the other locus (e.g.  $A_i$  and  $B_j B_k$ ), and the **dominance-by-dominance** interaction  $DD$  is any residual interaction between the genotype at one locus with the genotype at another. As might be expected, the terms in Equation 26 are uncorrelated, so that we can write the genetic variance as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 \quad (27)$$

More generally, with  $k$  loci, we can include terms up to (and including)  $k$ -way interactions. These have the general form of  $A^n D^m$  which (for  $n + m \leq k$ ) is the interaction between the  $\alpha$  effects at  $n$  individual loci with the dominance interaction ( $\delta$ ) at  $m$  other loci. For example, with three loci, the potential epistatic terms are

$$\sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + \sigma_{AAA}^2 + \sigma_{AAD}^2 + \sigma_{ADD}^2 + \sigma_{DDD}^2$$

### Resemblance Between Relatives: Basic Concepts

The **heritability** of a trait, a central concept in quantitative genetics, is the proportion of variation among individuals in a population that is due to variation in the additive genetic (i.e., breeding) values of individuals:

$$h^2 = \frac{V_A}{V_P} = \frac{\text{Variance of breeding values}}{\text{Phenotypic Variance}} \quad (28)$$

Since an individual's phenotype can be directly scored, the phenotypic variance  $V_P$  can be estimated from measurements made directly on the population.

In contrast, an individual's breeding value cannot be observed directly, but rather must be inferred from the mean value of its offspring (or more generally using the phenotypic values of other known relatives). Thus estimates of  $V_A$  require known collections of relatives. The most common situations (which we focus on here) are comparisons between parents and their offspring or comparisons among sibs.

**Key observation:** *The amount of phenotypic resemblance among relatives for the trait provides an indication of the amount of genetic variation for the trait. If trait variation has a significant genetic basis, the closer the relatives, the more similar their appearance.*

### Fisher's Theory of Phenotypic Resemblance Between Relatives

One of the key ideas in Fisher's 1918 paper was that phenotypic resemblance between relatives provides information on their genetic resemblance, and in particular allows us to estimate genetic variance components for the trait. Writing the phenotypic value of relatives  $x$  and  $y$  as  $P_x = G_x + E_x$  and  $P_y = G_y + E_y$ , the phenotypic covariance (resemblance) between relatives can be written as

$$\text{Cov}(P_x, P_y) = \text{Cov}(G_x + E_x, G_y + E_y) = \text{Cov}(G_x, G_y) + \text{Cov}(E_x, E_y)$$

showing that relatives resemble each other for quantitative traits more than they do unrelated members of the population for two potential reasons:

- Relatives share genes ( $\text{Cov}(G_x, G_y) > 0$ ). The closer the relationship, the higher the proportion of shared genes. This genetic covariance between relatives is a function of their degree of relatedness and the genetic variance components.
- Relatives share similar environments ( $\text{Cov}(E_x, E_y) > 0$ ).

## The Genetic Covariance Between Relatives

The Genetic Covariance  $Cov(G_x, G_y)$  = covariance of the genotypic values ( $G_x, G_y$ ) of the related individuals  $x$  and  $y$ .

We will first show how the genetic covariances between parent and offspring, full sibs, and half sibs depend on the genetic variances  $V_A$  and  $V_D$ . We will then discuss how the covariances are estimated in practice.

Genetic covariances arise because two related individuals are more likely to share alleles than are two unrelated individuals. Sharing alleles means having alleles that are **identical by descent** (IBD): namely that both copies of an allele can be traced back to a single copy in a recent common ancestor. Alleles can also be **identical in state** but not identical by descent. For example, both alleles in an  $A_1A_1$  individual are the same type (identical in state), but they are only identical by descent if both copies trace back to (descend from) a single copy in a recent ancestor.

As a further example, consider the offspring of two parents and label the four allelic copies in the parents by 1 - 4, independent of whether or not any are identical in state.

$$\text{Parents: } A_1A_2 \times A_3A_4$$

$$\text{Offspring: } o_1 = A_1A_3 \quad o_2 = A_1A_4 \quad o_3 = A_2A_3 \quad o_4 = A_2A_4$$

Here,  $o_1$  and  $o_2$  share one allele IBD,  $o_1$  and  $o_3$  share two alleles IBD,  $o_1$  and  $o_4$  share no alleles IBD.

### 1. Offspring and one parent.

What is the covariance of genotypic values of an offspring ( $G_o$ ) and its parent ( $G_p$ )? Denoting the two parental alleles at a given locus by  $A_1A_2$ , since a parent and its offspring share *exactly* one allele, one allele in the offspring came from the parent (say  $A_1$ ), while the other offspring allele (denoted  $A_3$ ) came from the other parent. To consider the genetic contributions from a parent to its offspring, write the genotypic value of the parent as  $G_p = A + D$ . We can further decompose this by considering the contribution from each parental allele to the overall breeding value, with  $A = \alpha_1 + \alpha_2$ , and we can write the genotypic value of the parent as  $G_p = \alpha_1 + \alpha_2 + D_{12}$  where  $D_{12}$  denotes the dominance deviation for an  $A_1A_2$  genotype. Likewise, the genotypic value of its offspring is  $G_o = \alpha_1 + \alpha_3 + D_{13}$ , giving

$$Cov(G_o, G_p) = Cov(\alpha_1 + \alpha_2 + D_{12}, \alpha_1 + \alpha_3 + D_{13})$$

We can use the rules of covariances to expand this into nine covariance terms,

$$\begin{aligned} Cov(G_o, G_p) = & Cov(\alpha_1, \alpha_1) + Cov(\alpha_1, \alpha_3) + Cov(\alpha_1, D_{13}) \\ & + Cov(\alpha_2, \alpha_1) + Cov(\alpha_2, \alpha_3) + Cov(\alpha_2, D_{13}) \\ & + Cov(D_{12}, \alpha_1) + Cov(D_{12}, \alpha_3) + Cov(D_{12}, D_{13}) \end{aligned}$$

By the way have (intentionally) constructed  $\alpha$  and  $D$ , they are uncorrelated. Further,

$$Cov(\alpha_x, \alpha_y) = \begin{cases} 0 & \text{if } x \neq y, \text{ i.e., not IBD} \\ Var(A)/2 & \text{if } x = y, \text{ i.e., IBD} \end{cases} \quad (29a)$$

The last identity follows since  $Var(A) = Var(\alpha_1 + \alpha_2) = 2Var(\alpha_1)$ , so that

$$Var(\alpha_1) = Cov(\alpha_1, \alpha_1) = Var(A)/2$$

Hence, when individuals share one allele IBD, they share half the additive genetic variance. Likewise,

$$Cov(D_{xy}, D_{wz}) = \begin{cases} 0 & \text{if } xy \neq wz, \text{ i.e., both alleles are not IBD} \\ Var(D) & \text{if } xy = wz, \text{ both alleles are IBD} \end{cases} \quad (29b)$$

Two individuals only share the dominance variance when they share both alleles. Using the above identities (29a and 29b), eight of the above nine covariances are zero, leaving

$$Cov(G_o, G_p) = Cov(\alpha_1, \alpha_1) = Var(A)/2 \quad (30)$$

## 2. Half-sibs.

Here, one parent is shared, the other is drawn at random from the population. Let  $o_1$  and  $o_2$  denote the two sibs. Again consider a single locus. First note that  $o_1$  and  $o_2$  share either one allele IBD (from the father) or no alleles IBD (since the mothers are assumed unrelated, these sibs cannot share both alleles IBD as they share no maternal alleles IBD). The probability that  $o_1$  and  $o_2$  both receive the same allele from the male is one-half (because whichever allele the male passes to  $o_1$ , the probability that he passes the same allele to  $o_2$  is one-half). In this case, the two offspring have one allele IBD, and the contribution to the genetic covariance when this occurs is  $Cov(\alpha_1, \alpha_1) = Var(A)/2$ . When  $o_1$  and  $o_2$  share no alleles IBD, they have no genetic covariance.

Summarizing:

Case	Probability	Contribution
$o_1$ and $o_2$ have 0 alleles IBD	1/2	0
$o_1$ and $o_2$ have 1 allele IBD	1/2	$Var(A)/2$

giving the genetic covariance between half sibs as

$$Cov(G_{o_1}, G_{o_2}) = Var(A)/4 \quad (31)$$

## 3. Full-Sibs.

Here both parents are in common, and three cases are possible as the sibs can share either 0, 1, or 2 alleles IBD. Applying the same approach as for half sibs, if we can compute: 1) the probability of each case; and 2) the contribution to the genetic covariance for each case.

Each full sib receives one paternal and one maternal allele. The probability that each sib receives the same paternal allele is 1/2, which is also the probability each sib receives the same maternal allele. Hence,

$$\Pr(2 \text{ alleles IBD}) = \Pr(\text{paternal allele IBD}) * \Pr(\text{maternal allele IBD}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\Pr(0 \text{ alleles IBD}) = \Pr(\text{paternal allele not IBD}) * \Pr(\text{maternal allele not IBD}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\Pr(1 \text{ allele IBD}) = 1 - \Pr(2 \text{ alleles IBD}) - \Pr(0 \text{ alleles IBD}) = \frac{1}{2}$$

We saw above that when two relatives share one allele IBD, the contribution to the genetic covariance is  $Var(A)/2$ . When two relatives share both alleles IBD, each has the same genotype at the locus being considered, and the contribution is

$$Cov(\alpha_1 + \alpha_2 + D_{12}, \alpha_1 + \alpha_2 + D_{12}) = Var(\alpha_1 + \alpha_2 + D_{12}) = Var(A) + Var(D)$$

Putting these results together gives

Case	Probability	Contribution
$o_1$ and $o_2$ have 0 alleles IBD	1/4	0
$o_1$ and $o_2$ have 1 allele IBD	1/2	$Var(A)/2$
$o_1$ and $o_2$ have 2 allele IBD	1/4	$Var(A) + Var(D)$

This results in a genetic covariance between full sibs of

$$Cov(G_{o_1}, G_{o_2}) = \frac{1}{2} \frac{Var(A)}{2} + \frac{1}{4} (Var(A) + Var(D)) = \frac{Var(A)}{2} + \frac{Var(D)}{4} \quad (32)$$

#### 4. Monozygotic Twins.

Here, both sibs have identical genotypes, giving

$$Cov(G_{o_1}, G_{o_2}) = Var(A) + Var(D) \quad (33)$$

#### 5. General degree of relationship.

The above results for the contribution when relatives share one and two alleles IBD suggests the general expression for the covariance between (noninbred) relatives.

If  $r_{xy} = (1/2) \text{Prob}(\text{relatives } x \text{ and } y \text{ have one allele IBD}) + \text{Prob}(\text{relatives } x \text{ and } y \text{ have both alleles IBD})$ , and  $u_{xy} = \text{Prob}(\text{relatives } x \text{ and } y \text{ have both alleles IBD})$ , then the genetic covariance between  $x$  and  $y$  is given by

$$Cov(G_x, G_y) = r_{xy}V_A + u_{xy}V_D \quad (34a)$$

If epistatic genetic variance is present, this can be generalized to

$$Cov(G_x, G_y) = r_{xy}V_A + u_{xy}V_D + r_{xy}^2V_{AA} + r_{xy}u_{xy}V_{AD} + u_{xy}^2V_{DD} + \dots \quad (34b)$$

#### 6. Maternal and shared environmental effects

The phenotypic covariance  $\sigma(P_x, P_y)$  between two relatives  $x$  and  $y$  is a function of both the genetic covariance (due shared genes) and the environmental covariance (due to shared environmental values),

$$\text{Cov}(P_x, P_y) = \text{Cov}(G_x, G_y) + \text{Cov}(E_x, E_y)$$

It is often assumed that relatives do not share common environmental values (so that  $\text{Cov}(E_x, E_y) = 0$ ), in which case the phenotypic variance equals the genetic covariance. However, in human genetics this is a poor assumption, especially when family members are being compared. First, parents and offspring may share a common environment. A further complication is that environmental material effects can be shared among sibs, with twins potentially sharing more effects (i.e., a higher environmental covariance) than sibs born at different times. Thus, in human studies the default assumption should be for common family effects (due to material effects for sibs sharing the same mother) and other shared environmental effects due to common family up-bringing.

#### Relative Risks for Binary Traits

With binary traits, we are often more interested in the probability of having the trait (i.e., the disease) given that a known relative has the disease. These are closely related to the covariances between relatives. To see this, let  $z_1$  and  $z_2$  denote the trait state (0,1) in two relatives. For relatives of type  $R$  (e.g., full sibs, monozygotic twins, or parent-offspring), the **Recurrence risk**  $K_R = \text{Prob}(z_2 = 1 \mid z_1 = 1)$ , is the conditional probability that you have the disease given that a relative of relationship  $R$  has the disease. This is related to the covariance between relatives via **James' identity**,

$$K_R = K + \frac{Cov(z_1, z_2)}{K} \quad (35a)$$

where  $K = \Pr(z = 1)$ , i.e., the **population prevalence**, the frequency of the trait in the population.

We may also be interested in the **relative risk**,  $\lambda_R = K_R/K$ , which is related to the covariance between relatives via **Risch's identity**,

$$\lambda_R = 1 + \frac{\text{Cov}(z_1, z_2)}{K^2} \quad (35b)$$

### Searching for QTLs

The ability to generate a dense (i.e., closely linked) set of molecular markers for just about any species offers the potential to search for QTLs, loci underlying trait variation. While this will be discussed in great detail over the next several talks, a few brief comments are in order.

Two general approaches have been used. The first makes use of *linkage within a pedigree*, wherein we expect an excess of parental gametes. For example, if the father is  $MQ/mq$ , with molecular marker allele  $M$  tightly linked to a QTL with allele  $Q$  (which increases trait value relative to  $q$ ), the presence of a paternal  $M$  marker allele in the offspring indicates a high probability that these individuals also carry  $Q$ . Conversely, individuals bearing a paternal  $m$  marker likely carry QTL allele  $q$ , so that we expect paternal  $M$ -bearing individuals to (on average) have a larger trait value than paternal  $m$ -bearing individuals. The key here (within a pedigree) is to consider the contrasting of marker alleles from each parent separately. For example, while the father might be  $MQ/mq$ , the mother might have phase  $mQ/Mq$ . Thus, if we simply contrast the trait values in  $M$  vs.  $m$  alleles (without consideration of their parent of origin), we would not observe a difference in trait value between marker allele classes and thus conclude that this particular marker is not linked to a QTL segregating in the pedigree. Conversely, if one contrasts parental  $M$  vs.  $m$ , we would see  $M$  has a larger trait mean, while the  $M$  vs.  $m$  contrast for alleles of maternal origin shows that  $m$  has the larger trait value.

A second approach is to use a collection of individuals from a population, as opposed to a known set of relatives (i.e., a pedigree). The idea is that if we have very tightly-linked markers, then the presence of population-level disequilibrium implies that marker alleles are non-randomly associated with tightly linked QTL alleles. Thus, by contrasting marker alleles across the entire sample, we expect a difference in trait means if the marker is tightly linked to a segregating QTL. This approach avoids the requirement of large pedigrees, but requires a much denser marker map. A second (but very important) complication is that population structure can generate a covariance between unlinked markers, and hence any effects of population structure must first be removed before QTL mapping is attempted.

### A Few Suggested Reference Texts

Ewens, W. J. 2004. *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd Edition. Springer. A very sophisticated mathematical introduction to population and quantitative genetics.

Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*, 4th Edition. Longman. The standard introductory text. Highly recommended as a starter text. Very nice in laying the foundations, but lacking the detail of more advanced texts.

Lynch, M, and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*, Sinauer Associates. The most comprehensive advanced text on quantitative genetics, but then again I'm biased on this one.

Sham, Pak. 1998. *Statistics in Human Genetics*. Arnold, New York. A nice, compact, and often overlooked text.