Lecture 1

Introduction to Population and Quantitative Genetics

Bruce Walsh. July 2005. Asian Institute on Statistical Genetics

OVERVIEW

As background for the rest of the lectures in this course, 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).

We start with what (at first) may seem somewhat of a digression, namely an overview of two of the most important papers in biology, those of Darwin and Mendel, which roughly appeared at the same time. Both revolutionized biology, but Mendel's work took much longer to be accepted. Further, Darwin was concerned with traits that adapt an organism to its environment. These are usually continuous and (as we now know) result from the interaction of a number of genes coupled with the environment. In contrast, Mendelian genetics (in its initial form) was concerned with single genes that have very obvious effects on traits. The modern theory of evolution required R. A. Fisher's classic 1918 paper showing how Mendelian genetics underpins the genetics of complex traits. Fisher's work also introduced several important concepts in modern statistics, and it is not surprising that the analysis of complex traits (quantitative genetics) is a field rich in statistics.

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.

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, especially of continuous traits. He fell back on the standard model of his day, **blending inheritance**. This theory assumes that both parents contribute fluids to the offspring, and these fluids contain the genetic material, which are 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

$$z_o = (z_m + z_f)/2$$

Fleming Jenkin (in 1867) pointed out a serious problem with blending inheritance. Consider the variation in trait value in the offspring,

$$\mathrm{Var}(z_o) = \mathrm{Var}[(z_m + z_f)/2] = \frac{1}{2}\,\mathrm{Var(parents)}$$

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 Naturforschedenden 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. Another idea was that Mendel's original suggestion was perhaps too mathematical for 19th century biologists. While this may indeed 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.

Probability and Genetics

Mendel's key insight was that *genes are discrete particles*, with a (diploid) parent passing one of its two copies of each gene at random to their offspring. Hence, probability plays a key role in the understanding and the analysis of genetics and we start by reviewing a couple of central concepts.

Let $\Pr(A)$ denote the probability that event A occurs. Probabilities are positive and lie between zero and one, so that

$$0 \leq \Pr(A) \leq 1 \tag{1.1a}$$

If $\Pr(A) = 0$, then A never occurs, while if $\Pr(A) = 1$, then A always occurs. If the events $A_1, A_2, \cdots A_n$ are all the possible outcomes, then

$$\sum_{i=1}^{n} \Pr(A_i) = 1 \tag{1.1b}$$

Namely, *probabilities sum to one*. This is an extremely useful result. Suppose we are interested in the probability that any event *except* A_1 occurs. We could compute this as $\sum_{i=2}^{n} \Pr(A_i)$. However, we can often compute this much easier by noting that

$$\Pr(\text{not } A_1) = 1 - \Pr(A_1) \tag{1.1c}$$

Example 1.1 Suppose we cross two Qq parents. What is the probability of getting any genotype *except* qq?

$$\Pr(\text{not } qq) = 1 - \Pr(qq) = 1 - 1/4 = 3/4$$

Now consider two events, A and B. Suppose that A and B are **independent**, namely knowing that B has occurred tells us nothing about A. The probability that both the events A and B occur is

$$\Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B) \tag{1.2a}$$

This is often called the **AND Rule**. If the events are independent, the Probability of A and B and C is just $\Pr(A) \cdot \Pr(B) \cdot \Pr(C)$, so that *and = multiply probabilities*.

Now suppose that events A and B are **mutually exclusive** (they do not contain any overlapping events). For example, if A = roll an even on dice and B = role a 6, these are overlapping events, while if B = roll a 5 then the events A and B are indeed mutually exclusive. If A and B are mutually exclusive, then the probability of A OR B is just their sum,

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) \tag{1.2b}$$

This is often know as the **OR Rule**, with *or = add probabilities*. Note that for Equation 1.1b to hold, we require that the A_i are mutally exclusive events.

Example 1.2 Let's revisit Example 1.1. We can write Pr(not qq) = Pr(QQ or Qq). From the OR Rule,

$$\Pr(QQ \text{ or } Qq) = \Pr(QQ) + Pr(Qq) = 1/4 + 1/2 = 3/4$$

How do we know that $\Pr(QQ) = 1/4$? This follows from the AND rule, as to get a QQ offspring, the father must contribute a Q *AND* the mother must contribute a Q. Hence

$$\Pr(QQ) = \Pr(Q \text{ from father}) \cdot \Pr(Q \text{ from mother}) = (1/2) * (1/2) = 1/4$$

To see both the AND and OR rules in action, consider $\Pr(Qq)$. This can occur two different (mutually exclusive) ways, as

$$\Pr(Qq) = \Pr(Q \text{ from father AND } q \text{ from mother OR } q \text{ from father AND } Q \text{ from mother})$$

$$\Pr(Qq) = \Pr(Q \text{ from father AND } q \text{ from mother }) + \Pr(q \text{ from father AND } Q \text{ from mother})$$

$$\Pr(Qq) = \Pr(Q \text{ from father}) \cdot \Pr(q \text{ from mother }) + \Pr(q \text{ from father}) \cdot \Pr(Q \text{ from mother})$$

$$= (1/2)(1/2) + (1/2)(1/2) = 1/4$$

Finally, if Q is a dominant allele, we are often interested in the probability of a genotype that contains at least one Q, namely

$$\Pr(Q-) = \Pr(QQ) + \Pr(Qq) = 3/4$$

What happens if A and B are **dependent**, namely that event A contains information about B? In this case, we use conditional probability, and define $\Pr(A \mid B)$ is the *Probability of A given B*, or the **conditional probability** of A given that we know B. We can compute $\Pr(A \mid B)$

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)} \tag{1.3a}$$

where $\Pr(A, B)$ is the **joint probability** that both A and B occur. We can rearrange this to give

$$\Pr(A, B) = \Pr(A \mid B) \cdot \Pr(B) \tag{1.3b}$$

If A and B are independent, then $\Pr(A \mid B) = \Pr(A)$ and we recover the AND rule (Equation 1.2a)

Example 1.3 Suppose individuals that have at least one Q are purple, while qq are green. If we cross two Qq parents, what is the probability that a purple offspring is really QQ? Using the definition of conditional probability (Equation 1.3a),

$$\Pr(QQ \,|\, \text{Purple}) = \frac{\Pr(QQ, \text{Purple})}{\Pr(\text{Purple})} = \frac{\Pr(QQ)}{\Pr(Q-)} = \frac{1/4}{3/4} = 1/3$$

which follows in that all QQ are purple, hence $\Pr(QQ, \text{Purple}) = \Pr(QQ)$

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 yellow pea line to a pure-breeding green pea line. 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?

Pr(YY) = Pr(Allele Y from dad)*Pr(Allele Y from mom) = $(1/2)*(1/2) = 1/4$
Pr(gg) = Pr(Allele g from dad)*Pr(Allele g from mom) = $(1/2)*(1/2) = 1/4$
Pr(Yg) = 1-Pr(YY) - Pr(gg) = $1/2$

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

Pr(Yg) = Pr(dad = Y)*Pr(mom = g) + Pr(dad = g)*Pr(mom = Y) = $(1/4)(1/4) + (1/4)(1/4) = 1/2$

Hence, Prob(Yellow phenotype) = Pr(YY) + 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 α_i is the average trait value of allele A_i and α_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_iA_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 despot 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, epsistais occurs when the two-locus genotypic value $G_{ijkl} \neq G_{ij} + G_{kl}$, the sum of the two single-lcous values.

Mendel's View of Inheritance: Independent Assortment at 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 seed traits: shape (round vs. wrinkled) and color (green vs. yellow). 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

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

Assuming genotypes at the different loci are independently inherited, the probability of seeing a round, yellow F_2 individual is

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

Likewise,

$$\begin{aligned}
\Pr(\text{yellow, wrinkled}) &= \Pr(rrY-) = \Pr(rr) \cdot \Pr(Y-) = (1/4) * (3/4) = 3/16 \\
\Pr(\text{green, round}) &= \Pr(R-gg) = \Pr(R-) \cdot \Pr(gg) = (3/4) * (1/4) = 3/16 \\
\Pr(\text{green, wrinkled}) &= \Pr(rrgg) = \Pr(rr) \cdot \Pr(gg) = (1/4) * (1/4) = 1/16
\end{aligned}$$

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 $AaBBCcDD \times aaBbCcDd$, 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,

$$\begin{aligned}\Pr(AaBbCc) &= \Pr(Aa) * \Pr(Bb) * \Pr(Cc) \\ &= (1/2) * (1/2) * (1/2) = 1/8\end{aligned}$$

Mendel was Wrong: Linkage

Shortly after the rediscovery of Mendel, Bateson and Punnet looked at a cross in peas involvig 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 is a significant departure from independent assortment, with an excess of PL and pl gametes over Pl and pL, and evidence that the P and L 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 (and most other eukaryotes) also have another type of DNA molecule, 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. The chloroplast found in plants and some unicellular organisms also contain multiple copies of the **chloroplast genome** (or cpDNA). These genomes are also usually (although not always) strictly maternally inherited.

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 configuation as on the parental chromosomes.

Consider the Bateson-Punnett pea data, and let PL/pl denote that 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 chromosomes** 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 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.

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

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

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

Generally, to compute the expected genotype probabilities, need 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 **replusion**.

BASIC POPULATION GENETICS

Mendelian genetics provides the rules of transmission of genes and genotypes 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,

$$p_i = \text{freq}(A_i) = \text{freq}(A_iA_i) + \frac{1}{2}\sum_{i\neq j}\text{freq}(A_iA_j) \tag{1.4}$$

The $1/2$ appears since only half of the alleles in A_iA_j heterozygotes are A_i. Equation 1.4 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,

$$\text{freq}(A_iA_j) = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_ip_j & \text{for } i \neq j \end{cases} \tag{1.5}$$

Equation 1.5 is the first part of the **Hardy-Weinberg thereom**, 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 will 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 1.5) 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,

$$\text{freq}(AABB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \tag{1.6a}$$

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

$$\text{freq}(AaBB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(aB|\text{mother}) + \text{freq}(aB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \tag{1.6b}$$

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 equilibrum**, the frequency of a multi-locus gamete equals the product of the individual allele frequencies. For example, for two and three loci, the linkage equilibrium gamete frequencies are just

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) \quad \text{for 2 loci}, \quad \text{freq}(ABC) = \text{freq}(A) \cdot \text{freq}(B) \cdot \text{freq}(C) \quad \text{for 3 loci}$$

In linkage equilibrium, the alleles are different loci are independent — knowledge that a gamete contains one allele (say A) provides no information on the allele from the second locus. More generally, loci can show **linkage disequilibrium** (LD), which is also called **gametic phase disequilbirum** 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

$$D_{AB} = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) \tag{1.7a}$$

Rearranging Equation 1.7a shows that the gamete frequency is just

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) + D_{AB} \tag{1.7b}$$

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

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 straight-forward 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), the method of 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. This includes (as a special case) the situation where a trait is influenced by variation at only a single gene but that is 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. Examples would include weight, blood pressure, and cholesterol levels. For all of these traits there are both genetic and environmental risk factors. The distinction is sometime made between **metric traits** (thoses that can take on continuous values, such as height of weight) and **meristic traits**, those that take on countable values, such as number of leaves on a tree. Typically, however, we tend to treat meristic traits as being continuous.

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 your sib (or some other relative) has a disease/trait, what are your odds of showing that trait? 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. Finally, quantitative genetics allows both breeders and envolutionary biologists to predict the response to selection and the effects of different mating systems (such as selfing vs. outcrossing) on complex traits.

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 apparent phenotypic simplistic can easily mask a very complex genetic basis.

Loci harboring alleles that increase disease risk are often called **disease susceptibly** (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(disease | DD) = 0.5, so that the **penetrance** of genotype DD is 50%. Likewise, suppose for the other genotypes that Pr(disease | Dd) = 0.2, Pr(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 genetic (from $D-$) causes. For example, suppose freq(d) = 0.9, what is Prob (DD | show disease)? First, the **population prevalence** K (the frequency) of the disease is

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

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

$$\Pr(b|A) = \frac{\Pr(A|b) * \Pr(b)}{\Pr(A)} \tag{1.8}$$

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 } disease) = 0.442, Pr(dd } disease) = 0.497

Lecture 1 Problems

1. In the fruit fly *Drosophila*, there is no recombination in males. Suppose we cross a AB/ab male to an Ab/aB female. What is the probability of an $AaBb$ offspring if the recombination frequency between the A and B loci is 0.2?

2. In 2007, NASA will find life on Mars. The discovered life form has three sexes, and in a NASA lab SSs, Sss and sss parents are crossed. What is the probability of an sss offspring? Of an Sss offspring?

3. An application of Bayes' theorem. Suppose there is a genetic disorder that results in all offspring being female and further suppose that the probability a randomly-chosen family has this disorder in 0.05. If we observe a family with 6 girls (and no boys), what is the probability this is a sex-bias family?

Solutions to Lecture 1 Problems

1. In the fruit fly *Drosophila*, there is no recombination in males. Suppose we cross a AB/ab male to an Ab/aB female. What is the probability of an $AaBb$ offspring if the recombination frequency between the A and B loci is 0.2?

$$\Pr(AaBb) = \Pr(Ab/aB) + \Pr(AB/ab)$$

$$\Pr(Ab/aB) = \Pr(Ab|\text{dad}) * \Pr(aB|\text{mom}) + \Pr(Ab|\text{mom}) * \Pr(aB|\text{dad}) = 0 + 0$$

$$\Pr(AB/ab) = \Pr(AB|\text{dad}) * \Pr(ab|\text{mom}) + \Pr(AB|\text{mom}) * \Pr(ab|\text{dad})$$

$$= (1/2) * [0.2/2] + [0.2/2] * (1/2) = 0.1$$

Hence, a 10% probability of an $AaBb$ offspring.

2. In 2007, NASA will find life on Mars. The discovered life form has three sexes, and in a NASA lab SSs, Sss and sss parents are crossed. What is the probability of an sss offspring?

$$\Pr(sss) = \Pr(s|\text{parent 1}) * \Pr(s|\text{parent 2}) * \Pr(s|\text{parent 3}) = (1/3) * (2/3) * (1) = 2/9$$

Of an Sss offspring? $\Pr(Sss)$

$$= \Pr(S|\text{p 1}) * \Pr(s|\text{p 2}) * \Pr(s|\text{p 3}) + \Pr(s|\text{p 1}) * \Pr(S|\text{p 2}) * \Pr(s|\text{p 3}) + \Pr(s|\text{p 1}) * \Pr(s|\text{p 2}) * \Pr(S|\text{p 3})$$

$$= (2/3)(2/3)(1) + (1/3)(1/3)(1) + 0 = 5/9$$

3. First, the probability of having six girls depends on the type of family. With a normal family, this is just $(1/2)^6$, while with a sex-bias family it is one. Hence,

$$\Pr(\text{ 6 girls}) = \Pr(\text{ 6 girls }|\text{normal})\,\Pr(\text{normal}) + \Pr(\text{ 6 girls }|\text{sex-bias})\,\Pr(\text{sex-bias})$$

$$= (1/2)^6 \cdot 0.95 + 1 \cdot 0.05 = 0.0764$$

Using Bayes' theorem,

$$\Pr(\text{sex-bias}|\text{6 girls}) = \frac{\Pr(\text{6 girls}|\text{sex-bias}) * \Pr(\text{sex-bias})}{\Pr(\text{6 girls})} = \frac{1 \cdot 0.05}{0.0764} = 0.65$$