Lecture 1

Introduction to Mendelian and Molecular Genetics

Bruce Walsh. Aug 2004. Royal Veterinary and Agricultural University, Denmark

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. 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

$$z_o = (z_m + z_f)/2 \tag{1.1}$$

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

$$\text{Var}(z_o) = \text{Var}[(z_m + z_f)/2] = \frac{1}{2}\,\text{Var(parents)} \tag{1.2}$$

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. One suggestion was that Mendel's original suggestion was perhaps too mathematical for 19th century biologists. While his 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.

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?

Prob(YY) = yellow(dad)*yellow(mom) = (1/2)*(1/2) = 1/4
Prob(gg) = green(dad)*green(mom) = (1/2)*(1/2) = 1/4
Prob(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:

Prob(Yg) = yellow(dad)*green(mom) + green(dad)*yellow(mom) = (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 i plus the average value of 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: 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

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

Since 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,

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

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

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

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 $AaBBCcDD \times aaBbCcDd$, 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 Mendels 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 usual in that it is strictly maternally inherited — offspring get only their mothers 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 configuation 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 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, $\Pr(PpLl) = (1-c)^2/2 + c^2/2$.

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**.

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 α), 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 α and β and fraction of null hypotheses, π_0. To see this, apply Bayes' theorem,

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

Consider the numerator first. Let π_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 α, 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 $1-\beta$ (β 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

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

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 α = 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 · 0.05 = 47.5 of these to show a false-positive. Conversely, only 50 are expected to be linked, and we would declare 50 · 0.80 = 40 of these to be significant, so that 47.5/87.5 of the significant results are due to false-positives.

Introduction to Molecular Genetics: Structure of DNA

Finally, a few brief comments on molecular genetics are in order. **Deoxyribonucleic Acid** (DNA for short) is the biological molecule that stores the information that is the genes. As mentioned, each chromosome consists of a single DNA molecule, usually extensively coated with proteins. The DNA molecular has four **bases** or **nucleotides** as its fundamental building blocks: Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). A DNA molecular is a very long double-stranded polymer constructed from these four bases. The molecule is a double helix, with each strand being complementary to the other (i.e., if we know only one strand, we know the exact sequence of the complementary strand. This occurs because A on one strand is paired with T on the other, and likewise G with C. DNA is an ideal information-storage molecule because of this redundancy and also because it is a very stable molecule, i.e., it generally does not readily interact with other molecules.

In contrast, RNA, or **Ribonucleic Acid**, is a very active molecule and can start a number of chemical and biological reactions. In the cell, the stable information for a gene on the DNA is turned into a particular **protein** by first **transcribing** the DNA sequence into an RNA sequence (in particular, a messenger, or mRNA sequence), which is then **translated** into the proteins that make up the cell on special structures called **ribosomes**, which are a mixture of both RNA (rRNA in the case) and proteins.

Regulation of the expression of a gene (i.e., the timing, amount, and location of the final product, be it a protein or RNA) can occur through control of its transcription, through control at translation, or through post-translation control. One popular tool are **microarrays**, which allow one to examine the mRNA levels for all the genes in the cell at a particular time-point. However, while quite powerful, this is only examining one level of the control of gene expression.

Introduction to Molecular Genetics: Molecular Markers

DNA from natural populations is highly **polymorphic**, in that if we looked at the DNA sequence for a random collection of the same chromosome (say the X chromosome), no two sequences would be the same (except for identical twins). In Humans, one polymorphism occurs roughly every 100 to 1000 bases. Any two random humans differ by over 20,000,000 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 proposes, we will distinguish between the two most widely used types of markers, **SNPs**, or **single nucleotide polymorphism** and **STRs** or **simple tandem repeats**. SNPs result from the change in a single base, for example AAG**G**AA to AAG**T**AA. 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, –ACACACAC— vs. –ACACACACAC—. 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 typically a large number of alleles in the population. As a result, STR sites are usually very highly polymorphic, making them idea for certain types of mapping. SNPs on the other hand, also have advantages, as we will see later on in the course

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. What value of α should we choose, assuming 80% power (i.e., $1 - \beta = 0.80$), to give a PER value of 0.05 in a human linkage analysis study?

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?

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

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

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

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

Hence, a 20% 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?

$$\Pr(sss) = \Pr(s|\text{parent 1}) * \Pr(s|\text{parent 2}) * \Pr(s|\text{parent 3}) = (1/3) * (2/3) * (1) = 4/9$$

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. What value of α should we choose, assuming 80% power (i.e., $1 - \beta = 0.80$), to give a PER value of 0.05 in a human linkage analysis study?
Here $\pi_0 = 0.95$.

$$0.05 = \frac{\alpha \cdot 0.95}{\alpha \cdot 0.95 + 0.8 \cdot 0.05}$$

or

$$0.05 * [\alpha \cdot 0.95 + 0.04] = 0.0475 \cdot \alpha + 0.002 = \alpha \cdot 0.95$$

giving $\alpha(0.95 - 0.0475) = 0.002$, or $\alpha = 0.00222$