Psycholoy 5135:
Class Notes on Genetics

James J. Lee

March 2, 2019

Contents

1 Fundamentals of Mendelian Inheritance 2
- 1.1 Genes, Alleles, Sites, and Genomes 2
- 1.2 Meiosis, Mating, and Mendel's Laws 6
- 1.3 Exceptions to Mendel's Laws 7
- 1.4 Asexual Reproduction 10

2 Quantitative Genetics 11
- 2.1 An Even Simpler Representation: "Plus" and "Minus" Alleles 11
- 2.2 The Average Effect of Gene Substitution 12
- 2.3 Heritability and the Correlations Between Relatives: A First Pass 18
- *2.4 Heritability and the Correlations Between Relatives: A Second Look 23

3 Modern Genetic Research with DNA-Level Data 27
- 3.1 Detection of Individual Causal Sites 27
- 3.2 Estimating Heritability Using Unrelated Individuals 29
- 3.3 Estimating Genetic Correlations Between Traits 30

4 Population Genetics 31

5 Summary 35

Genetics is important to differential psychology because it is usually the case that a large fraction of a psychological trait's variance is caused by genetic differences among individuals. Given the advent of modern genotyping and sequencing technology, the importance of genetics will only grow as we become increasingly able to identify the specific parts of the genome responsible for contributing to individual differences.

These notes present a "cartoon" version of genetics. Many important facts are omitted, and others are greatly simplified. This very streamlined account, however, will suffice for this course. The optional footnotes try to give some idea of the underlying complexity, and you can get an even more complete picture from PSY5137 (Behavioral Genetics) or the many excellent textbooks on the subject (e.g., Lynch & Walsh, 1998; Gillespie, 2004; Pierce, 2010; Watson et al., 2014). You might also find the following links helpful:

- 23andMe Genetics Primer I: What Are Genes?
- 23andMe Genetics Primer II: What Are SNPs?
- 23andMe Genetics Primer III: Where Do Genes Come From?
- 23andMe Genetics Primer IV: What Are Phenotypes?
- DNA Structure
- Meiosis
- Gene Expression

Unfortunately, special terms are not always used consistently in genetics; the exact meaning of a word can depend on the context. I will try to keep my usage consistent, which means that occasionally the same special term might be used slightly differently in these notes than in the links above.

These notes are admittedly quite challenging. It might be a good idea to just skim these notes before the lecture on genetics and then to study the material more carefully in small chunks distributed evenly across the time between the first and second midterms. If by then you understand all of the figures (with the exception of Figure 8) and can follow the summary in Section 5, you can rest assured that you understand the genetics necessary to do well in this course.

I strongly recommend that you print out a hard copy of these notes. You will often need to go back to one of the figures, and this is easier to do if you have a hard copy to flip through.

1 Fundamentals of Mendelian Inheritance

1.1 Genes, Alleles, Sites, and Genomes

The **genome** of an organism is the totality of its genetic material (DNA). Each cell in your body contains its own copy of your genome, although only certain parts of the genome tend to be active in any given cell. For example, parts of the genome that are particularly important for brain development or function will tend to more active in your brain cells

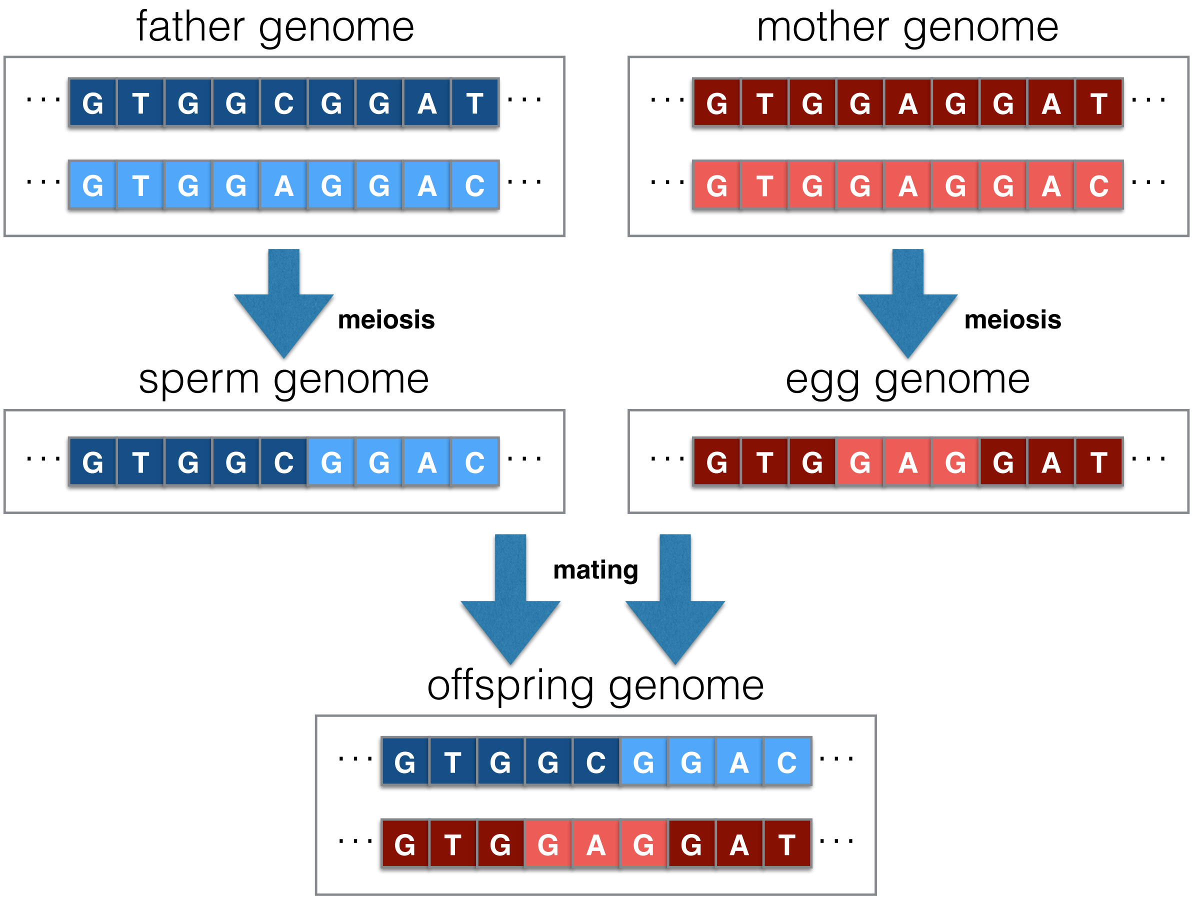

father genome
··· G T G G C G G A T ···
··· G T G G A G G A C ···
mother genome
··· G T G G A G G A T ···
··· G T G G A G G A C ···
meiosis
meiosis
sperm genome
··· G T G G C G G A C ···
egg genome
··· G T G G A G G A T ···
mating
offspring genome
··· G T G G C G G A C ···
··· G T G G A G G A T ···

Figure 1: A simplified picture of Mendelian inheritance. Each individual's genome consists of two "duplicate" versions, one from each parent. For example, we can imagine that the dark-blue contribution to the father's genome came from the paternal grandfather and that the light-blue contribution came from the paternal grandmother.

than parts that are less important. How each cell "knows" which parts of the genome to activate is a very interesting question, but we do not consider it too deeply in this course.

An "active" region of the genome is one that is being transcribed into **messenger RNA**, which is then translated into a **protein**. Proteins are folded chains of **amino acids** and make up the basic building blocks of your body. For example, the receptors used by your brain cells to bind neurotransmitters are proteins. By specifying the proteins (basic ingredients) and the timing and manner of their production, your genome plays a fundamental role in constructing your body and mind.

The fundamental laws of inheritance were discovered by the botanist Gregor Mendel in the 19th century. Figure 1 depicts the essence of Mendel's discoveries, in the light of our modern molecular understanding.

We can think of the genome as equivalent to a string of letters. Any one of four possible letters (bases)—`G`, `C`, `T`, or `A`—occupies each position in this string. We will call a position in this string a **site**; the token letter occupying the site is called a **gene**. Only a nine-site stretch of the genome is explicitly depicted in Figure 1; the ellipses preceding and following each genome mean that many more sites are not shown. In fact, the human genome contains over 3 *billion* sites.[1]

Each individual genome actually consists of two parts, each contributed by one parent. Figure 1 captures this fact by representing the genome of each mature individual as *two* strings. Let us concentrate on the genome of the father in this hypothetical pedigree. We can imagine that the dark-blue "half genome" was contributed by the *paternal grandfather* (the father's own father) and that the light-blue half genome was contributed by the *paternal grandmother* (the father's mother).

Similarly, we can imagine that the dark-red string within the mother's genome was contributed by the *maternal grandfather* and that the light-red string was contributed by the *maternal grandmother*.

At roughly 99.9 percent of all sites, nearly all human beings carry the same letter. In fact, *all humans and chimpanzees* carry the same letter at roughly 99 percent of all sites. Only about one percent of the genome is responsible for the differences *between* ourselves and our closest cousin species; only about 0.1 percent of the genome is responsible for individual differences *within* our own species. However, this small percentage still leaves plenty of room for individual differences, because it amounts to roughly 10 *million* sites where it is common for two people to carry different letters (or for the same person to inherit different letters from mother and father).

[1] The word **gene** has another meaning that is far more commonly used in modern genetics: a contiguous region of the type genome (e.g., the "human genome"), spanning many sites, whose expression or transcription leads to a particular biological product—typically a messenger RNA (mRNA) transcript, the translation of which leads to the production of a protein. In this course, however, we will frequently need to use the term to mean a token of a hereditary material (a base pair) that can fill an atomic site in the genome. For this reason I will try to consistently use the term **protein-coding region** to refer to a "gene" in the modern sense.

Let us examine the fifth site in Figure 1. An individual's **genotype** refers to the allelic states of the genes carried by the individual at a particular site. We can see that the genotype of the father at the fifth site is `CA`, whereas the genotype of the mother at the same site is `AA`. Thus, the fifth site is one of the millions in the human genome where not all individuals carry the same genotype. Such sites are called **polymorphic** (which is derived, I believe, from the Greek for "possessing many forms"). In Figure 1 the ninth site is also polymorphic. The father and mother happen to carry the same genotype at this site, `TC`, but it is possible that other individuals carry the genotypes `TT` and `CC` at this site. A polymorphic site of this kind is called a **single-nucleotide polymorphism** (SNP, pronounced "snip," for short).[2]

At polymorphic sites, where more than one letter can possibly occur, each distinct class of genes is called an **allele**. For example, at the fifth site in Figure 1, there are at least two alleles in this population, `C` and `A`; at the ninth site, the two alleles are `T` and `C`.[3]

A genotype composed of two different alleles is called **heterozgyous**, while a genotype composed of one allele represented twice is called **homozygous**. In Figure 1 the father and offspring are both heterozygous at sites 5 and 9; the mother is heterozygous at only site 9.

A quantity of great importance in population genetics is the **allele frequency**. This is simply the total number of genes falling within a specified allelic class divided by the total number of genes at the site. To get a better grasp on this concept, focus on the father and mother in the pedigree of Figure 1. In this sample of two individuals, the allele frequency of `C` at the fifth site is 0.25; one gene is `C`, and there are four total genes. Conversely, the allele frequency of `A` at the fifth site is 0.75. At the ninth site, both `T` and `C` have allele frequencies of 0.50. Of course, since that we are examining a sample of only two individuals, these numbers are probably not good estimates of the allele frequencies in the overall population to which the father and mother belong. Notice that allele frequencies can only take on values between 0 and 1.

There are many potential mechanisms by which a polymorphic site might contribute to individual differences. If the site lies within a genomic region that is ultimately translated into a protein, the two different alleles might specify different amino acids, thereby producing two variants of the protein with different properties. An example might be a polymorphic site in a region encoding a nicotinic acetylcholine receptor; the "smoking" allele might lead to a receptor that binds nicotine more readily, making it more likely that individuals carrying that allele will become smokers. Polymorphic sites outside of protein-coding regions can still contribute to individual differences by affecting the quantity of the protein encoded by a nearby region or the timing of the protein's production.

[2] More precisely, a SNP is a site where the identities of the letters may differ. There are other kinds of polymorphic sites. For example, at some sites, distinct genomes may differ by the insertion or deletion of at least one letter. SNPs, however, are the most abundant type of polymorphic site.

[3] A typical distance between polymorphic sites is roughly 300 bases.

1.2 Meiosis, Mating, and Mendel's Laws

Each cell in your body contain a copy of your genome, but **germline cells** have a special status because of their role in reproduction.

The normal duplicate status of the genome inside a typical cell is called **diploidy**. During a process known as **meiosis**, germline cells end up with only a single copy of the genome. The status of carrying just one "half genome" is called **haploidy**. The haploid copy of the genome present in each product of meiosis inside your sex organs is a *mixture* of the genetic material that you inherited from both of your parents. The technical term for the mixing process within meiosis is **recombination**.

Figure 1 provides a schematic overview of what meiosis does. Each sperm cell produced by the father contains one string of DNA rather than two; that is, meiosis has reduced an initially diploid germline cell to a haploid product. The haploid string of DNA inside the depicted sperm cell is a hybrid of the contributions made by the paternal grandfather (dark blue) and paternal grandmother (light blue); the genes at sites 1 through 5 come from the paternal grandfather, while the genes at sites 6 through 9 come from the paternal grandmother. During the process of meiosis itself, there was a recombination event between sites 5 and 6 that spliced together the distinct contributions of the two paternal grandparents.

Inside the depicted egg cell produced by the mother, we can see that the nine-site stretch of the haploid genome is the outcome of two recombination events: one between sites 3 and 4, and another between sites 6 and 7.[4] The result is that this particular egg cell contains the genetic material from the maternal grandfather at sites 1 through 3 and again at sites 7 through 9; at sites 4 through 6, the genetic material comes from the maternal grandmother.

The process of **mating** brings together the haploid sperm and egg cells to constitute a diploid offspring of the two parents.[5] A newly fertilized diploid organism is called a **zygote**; later it is called a fetus, a baby, a toddler, etc. The blue portion of the offspring genome in Figure 1 was contributed by the father, whereas the red portion was contributed by the mother. Notice that the coloring convention allows us to see that the respective paternal and maternal contributions to the offspring genome are themselves complex hybrids of the grandparental genomes.

This system of inheritance naturally raises some questions. Consider the father's genotype at the fifth site, which happens to be `CA`. In Figure 1, `C` was the gene that happened to be passed on to the sperm and eventually to the offspring. But it conceivably could have been `A`. Can we predict whether it will be `C` or `A` that ends up in the next sperm cell? More generally, is there any external variable that will provide some information about which of the two parental genes at a given site will enter the final product of meiosis?

[4] In reality, it would be rare for two recombination events to occur so close together in the genome.

[5] Hey, I told you that these notes were overly simplified.

Mendel gave the answer to this question: *No.*

Mendel's Law 1 *Each diploid individual possesses a pair of genes at any particular site. Upon becoming a parent, the individual transmits a* randomly *selected member of the pair to the offspring.*

What this law means is that whether, for example, `C` or `A` fills the fifth site in the sperm cell is perfectly analogous to a coin flip. In both cases the sheer complexity of the low-level physical events leads to a high-level fifty-fifty symmetry.

Here is another way to describe Mendel's First Law. The father depicted in Figure 1 will produce many millions of sperm cells over his lifetime. Half of these cells will carry `C` at the fifth site, while the other half will carry `A`. But before the conception of any *particular* offspring, it is not possible to tell whether the sperm cell that fertilizes the egg will be carrying `C` or `A`.

A human female produces only about 500 mature egg cells in her lifetime, but the same principle applies. Roughly half of the eggs borne by the mother in Figure 1 will carry `T` at the ninth site; the other half will carry `C`; and whether a particular egg cell is of one kind or the other cannot be predicted in advance.

Mendel's First Law is of such importance that it is depicted more clearly in Figure 2, which shows the sequence of events at the ninth site without the visual clutter of the other eight sites. Both parents carry the genotype `TC` at site 9. When it comes time to produce their offspring, Nature essentially flips a coin when she decides whether it is `C` or `T` that passes into each final product of meiosis.

In Figures 1 and 2, the newly conceived offspring happens to carry the genotype `CT` at site 9, but the genotype could equally well have been `TC`, `TT`, or `CC`. Since in most situations the sex of the parent who transmitted a particular gene does not matter, we can treat `CT` and `TC` as equivalent.

One might argue that Mendel's Second Law can be deduced from the First Law, but in any event we state it here for completeness.

Mendel's Law 2 *Genes at distinct sites are transmitted independently of one another. That is, the identity of the gene transmitted by parent to offspring at one site provides no information about the identity of the gene transmitted at another site.*

What this law means is that transmission of genes at two different sites, for instance, is perfectly analogous to two *independent* coin flips. Whether one coin lands heads or tails has no bearing on the outcome of the second coin flip.

1.3 Exceptions to Mendel's Laws

In the century and a half since Mendel formulated his two laws of inheritance, geneticists have learned that they are subject to exceptions.

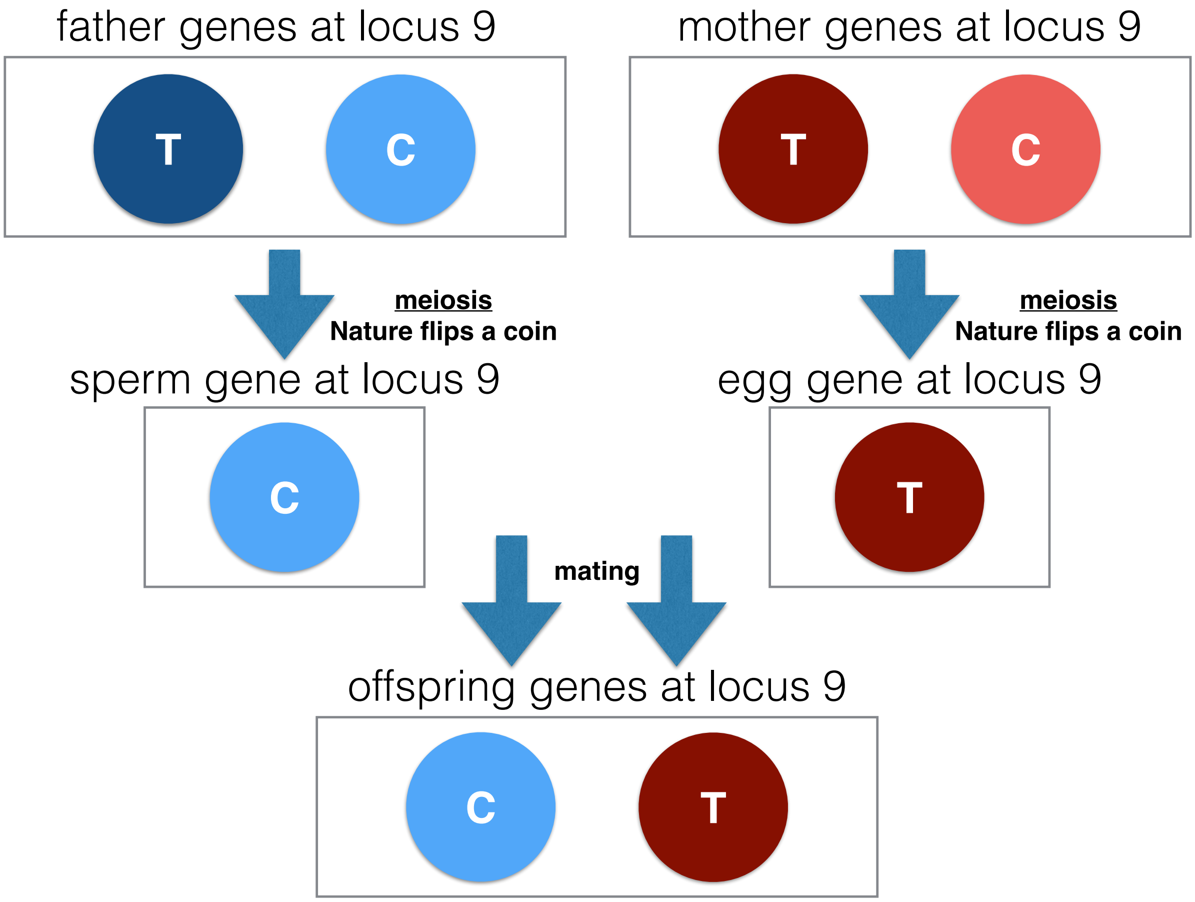


Figure 2: A simplified picture of Mendelian inheritance at one site. We can think of this figure as another representation of what occurred at the ninth depicted site in Figure 1. Each individual inherits two genes at the site, one from each parent. For example, we can imagine that the dark-blue gene in the father's genotype came from the paternal grandfather and that the light-blue gene came from the paternal grandmother.

Figure 1 gives the misleading impression that each haploid genome consists of a single long string (DNA molecule). In reality, the genetic material is arranged into several distinct strings, called **chromosomes**. The genomes of different species are characterized by different number of chromosomes; the genome of our own species, *Homo sapiens*, contains 23 chromosomes. You can think of Figure 1 as depicting just one pair of chromosomes out of the 23 pairs making up the genome of each individual. The **sex chromosome** differs from the other 22 in that it comes in two distinct types, X and Y; a zygote receiving XX from its parents becomes female, whereas one receiving XY becomes male. Because the mother can only transmit X in its own gametes, whether the offspring is male or female depends entirely on whether the father transmits its X or Y.

Sites located on the same chromosome will *not* follow the Second Law if they are so close together that recombination events are not likely to occur between them. This means, for example, that if the father in Figure 1 passes on `A` at the fifth site, he is also likely to pass on `C` at the ninth site (although in our example it happens that there was a recombination event between these sites in the meiosis producing the sperm cell). The Second Law is nevertheless a good approximation because the human genome is so large. If we select any two polymorphic sites at random, they will almost certainly be far enough apart to follow the Second Law very closely. In fact, they will probably be on different chromosomes.

One consequence of the Second Law's inexactness is that polymorphic sites close together in the genome often show strong correlations. That is, it may be that if we observe `A` at the fifth site in a token chromosome, we also observe `C` at the ninth site more often than expected from its allele frequency. (Conversely, we may observe token chromosomes carrying `C` and `T` at sites five and nine respectively more often than expected from their allele frequencies.) Such a correlation between polymorphic sites is called **linkage disequilibrium** (**LD**).[6] If the Second Law applied everywhere in the genome, any LD between nearby sites would quickly decay with the passage of generations. But since the Second Law does not so apply, LD between nearby sites often persists for some time.

One type of exception to the First Law occurs at sites where one allele produces a poison that kills sperm cells carrying the other allele. Such exceptions, however, are very exotic. It is almost always safe to assume the validity of the First Law: each of the two genes at a parental site has a fair chance at representation in the genome of a newly fertilized offspring.

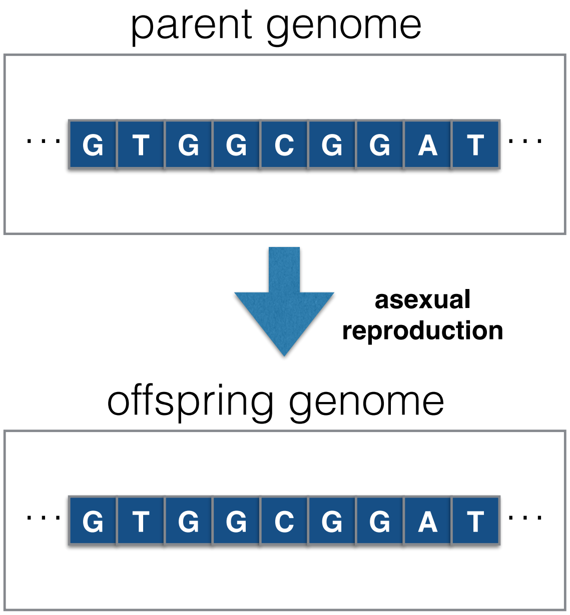


Figure 3: A simplified picture of asexual reproduction in a haploid organism.

1.4 Asexual Reproduction

The Mendelian scheme of inheritance is applicable only to sexually reproducing diploids. Many organisms do not reproduce sexually, and the genomes of such organisms usually remain in the haploid state throughout the life cycle. Since a parent in such a species must always transmit the single gene present at a given site, Mendel's laws describing the transmission of genes within pairs are clearly irrelevant here.

A consequence of asexual reproduction is that the genomes of parent and offspring are identical (Figure 3). In contrast, consider the ninth site in Figure 1. It is possible for the offspring to inherit a genotype at this site (`TT` or `CC`) that was not present in either father or mother. This difference between asexual and sexual reproduction will become important when we study why animals such as *Homo sapiens* bother with sexual reproduction in the first place.

Whereas siblings in an asexual species are genetically identical, the siblings in a sexual species are not (unless they are monozygotic twins) and in fact can be quite different from one another. For instance, at the fifth site in Figure 1, a future offspring of these same parents might carry the genotype `AA` rather than `CA`.

[6] **Linkage disequilibrium** is often confused with **linkage**. If there is linkage between two sites, this simply means that they do not follow the Second Law. If there is linkage disequilibrium between two sites, however, this means that the allelic content at the two sites shows a population correlation. It is very commonly the case, for instance, that two sites are linked but not in linkage disequilibrium. If the distinction is still not clear, you might want to keep thinking about it.

2 Quantitative Genetics

In the context of genetics, a trait such as height or IQ is often called a **phenotype**.

Quantitative genetics is the branch of genetics concerned with the analysis of phenotypes affected by many sites. It is turning out that nearly all defined phenotypes are affected more than one site, but the hypothetical phenotype considered by classical quantitative-genetic models is affected by thousands of sites, no single one of which accounts for a large percentage of the phenotypic variance.

One important application of quantitative genetics is to estimate the **heritability** of a given phenotype, which is the ratio of two variances; the numerator is a "genetic variance" in some sense, and the denominator is the total phenotypic variance. We can think of heritability as one means of quantifying the so-called "nature-nurture" debate. Higher values of heritability correspond to a greater fraction of individual differences being the result of specifically *genetic* differences. There are two senses of heritability that are important in differential psychology, and these will be set out below.

It turns out that the correlations between relatives are functions of the heritability and other parameters, allowing us to work backward and infer plausible values of the heritability from empirically observable correlation. For example, let us examine Figure 7. If certain assumptions hold, then the regression coefficient in biological families is equal to the heritability. The left panel of Figure 7 therefore suggests that the heritability of IQ might be something like 0.60. Regardless of whether that number is accurate, the general idea should make some sense to you; as more of the phenotypic variance is caused by genetic differences, the phenotypic resemblance between biological relatives owed to their shared genetic material should increase.

2.1 An Even Simpler Representation: "Plus" and "Minus" Alleles

Figure 1 might suggest that we are looking at a more-or-less randomly chosen nine-site region of the human genome. In reality it is very likely that none of the nine sites in such a region will be polymorphic. Even if there is a polymorphic site, it probably will not have any effect on our phenotype of interest.

The concept of heritability concerns the magnitude of the causal relation between genotype and phenotype. To portray this concept more vividly, Figure 4 essentially discards many features of Figure 1. We can imagine that the only sites represented in Figure 4 are both polymorphic *and* causal with respect to the phenotype.

A **causal site** with respect to a given phenotype satisfies the following:

1. if we can somehow experimentally "zap" the genotype carried by a new fertilized zygote at the site, so that the zygote now carries a *different* genotype,

2. the organism will grow up and obtain a phenotypic value *differing* from what it would have obtained if its genotype at the site had *not* been zapped in this way.

The allelic types of the genes in Figure 4 are no longer represented by the standard G, C, T, and A, but rather by the symbols $+$ and $-$. The notion is that replacing a $-$ gene with a $+$ gene will increase the phenotypic value obtained by the zapped individual, while the opposite substitution will decrease the phenotype value.[7] Because sites that are either monomorphic or causally irrelevant to phenotype are omitted from Figure 4, the nine depicted sites are not neighbors in the genome.

It is perhaps more evident in Figure 4 that only sites that are heterozygous in a parent can lead to different alleles in different offspring. For example, the mother happens to be homozygous at all sites from 5 through 9. All of her egg cells will be $++-+-$ at these sites. Thus it is only at heterozygous sites where Mendel's First Law comes into play; if a coin has heads on both sides, then we do not care so much about whether it is fair.

Figure 4 also emphasizes how members of the same family can be genetically different from each other. The father carries eleven $+$ genes at the nine depicted sites, and the mother carries ten. The offspring carries twelve—more than either parent. You should verify that the fewest possible $+$ genes carried by an offspring at these nine sites is seven and that the most possible is fourteen. If you ever encounter two siblings where one is much taller than the other, you should keep in mind the possibility that the Mendelian lottery gave the two siblings very different numbers of height-promoting genes.

Throughout the course, I will often refer to "plus" and "minus" genes. You might want to refer back to this section of the notes if you ever get confused by this usage.

2.2 The Average Effect of Gene Substitution

Another term that I will repeatedly use is the "average effect." Here is what I mean by this.

In general, the effect of transforming the allelic type of a gene in our hypothetical experiment will depend on the identity of the other gene at the zapped site, the identities of the genes at other sites, and the environmental experiences of the individual. For example, someone who has already inherited many $+$ genes might not benefit too much from an additional one. Or individuals who grow up in poor environments might obtain low phenotypic values no matter what genes they inherit. If the effect of some change depends on the context in this way, scientists say that the system is **nonlinear** or **nonadditive**. Here is an explanation of these terms:

- We can associate the term *nonlinear* with the graph of the genotype-phenotype relationship. In Figure 5 the solid blue dots represent the phenotypic averages of

[7] This collapsing of four letters to two arithmetic symbols will obviously not work for causal sites where there are more than two alleles. It is usually the case, however, that at multi-allelic sites only two of the alleles are reasonably common.

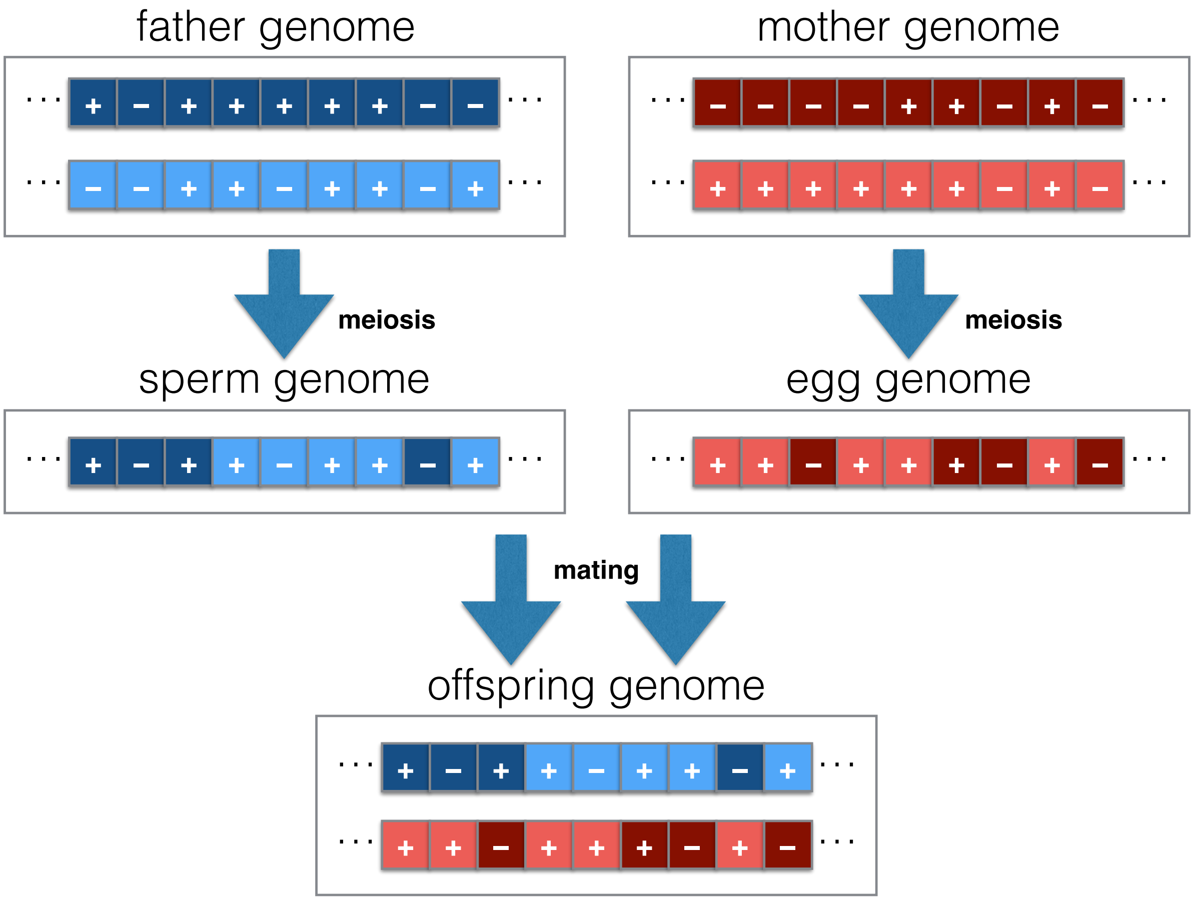
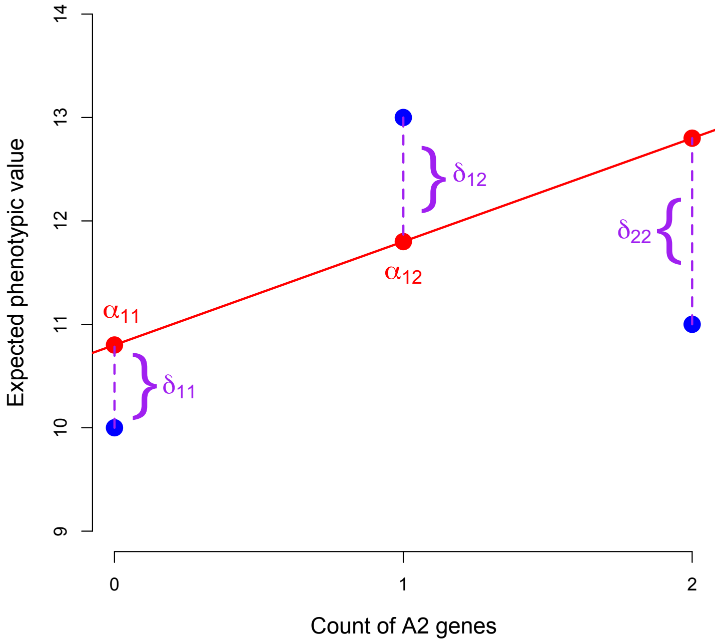


Figure 4: An even more drastically simplified picture of Mendelian inheritance. To create this figure, I used a coin flip at each site to determine whether the dark- or light-colored gene passes into the haploid germline cell.

Figure 5: Breeding (additive genetic) values and dominance (nonlinear) deviations at a single causal site. We assume that the causal site is uncorrelated with all other causes of the phenotype. The allele frequency of + is 0.4. The causal effect of changing $--$ to $+-$ is positive 3, while the effect of changing $+-$ to $++$ is negative 2. The phenotypic mean of each genotype is equal to the sum of its breeding value (α_{ij}) and dominance deviation (δ_{ij}). Each breeding value is the simply the phenotypic value predicted for the corresponding genotype by the least-squares regression line best fitting the genotype-phenotype relationship; each dominance deviation is the difference between the actual phenotypic mean of the corresponding genotype and its breeding value. The phenotypic means are represented by filled blue circles, the corresponding breeding values by open red circles.



the three possible genotypes in this example. If it were true that the effect of changing − to + is always the same, then the three solid blue dots would lie on a straight line. We can see that they do not lie on such a line in this example—hence, nonlinearity. Whereas changing − to + has a *positive* effect on the phenotype when there are no + genes already occupying the site, the same change has a *negative* effect when one + gene is already present. In fact, the only reason to call the counted allele "+" is because the regression line attempting to fit the genotype-phenotype relationship has a positive slope.

- We can associate the term *non-additive* with the fact that incrementing the number of + genes by one might not add as much to the phenotype as the previous increment. Again, in the example of Figure 5, going from no + genes to one adds *positive* 3 units to the phenotype, while the effect of going from one + gene to two adds *negative* 2—hence, non-additivity.

Epistasis is the special term for nonlinearity adopted by geneticists.

In the presence of nonlinearity, one can no longer refer to *the* effect of a gene substitution at a given site; the effect now depends on the context in which the gene substitution is performed. For this reason we define the **average effect of gene substitution** as the *average* of the causal effects across all possible genetic backgrounds and environments.[8]

Now consider the equation

$$Y = \alpha_0 + X_1\alpha_1 + X_2\alpha_2 + X_3\alpha_3 + \cdots + X_L\alpha_L + E = \alpha_0 + \sum_{j=1}^{L} X_j\alpha_j + E, \tag{1}$$

where Y is the individual's phenotype, α_0 is a constant, X_j is the number of site-j + genes (0, 1, or 2) carried by the individual, α_j is the average effect of gene substitution at site j, and E is a "residual" or "disturbance" attributable to nonlinearity, measurement error, and environmental causes of the trait. In fact, many geneticists looking at Equation 1 might think of E as standing for "environment," but the interpretation of this term is somewhat subtler in the presence of nonlinearity.

Do not proceed further in these notes until you have a solid grasp of what Equation 1 is saying. If a newly fertilized zygote with zero or one + genes at site 1 is "zapped" so that it now has an additional + gene, then on average its phenotype at the time of measurement will be α_1 units greater than what it would have been if this intervention had not been performed. Likewise, if site L is the one that is zapped, on average the phenotype will increase by α_L units.

You should be able to see that the term $\sum_{j=1}^{L} X_j\alpha_j$ can be reasonably regarded as the "genetic part" of the individual's phenotype (trait). Actually, there is a genetic component

[8] To be more accurate, the average effect of gene substitution is a *weighted* average of the causal effects; the different causal effects do not necessarily make the same contribution to the average. The weighting scheme can become somewhat complicated, and the reader is referred to Lee and Chow (2013) for details.

to the residual E if the different genetic sites interact nonlinearly, and we will address this component shortly.

At any given site, different people might carry different numbers of + genes. For example, in Figure 4, the depicted offspring carries two + genes at site 1, but the number carried by a future sibling might be only zero or one, depending on how the Mendelian lottery turns out. Since $\sum_{j=1}^{L} X_j \alpha_j$ therefore varies across individuals, it is a random variable, called the **breeding value** or **additive genetic value**. It is often symbolized by the letter A. The reason for the term "breeding value" is that a parent's value of A often predicts her child's phenotype better than her own phenotype ($Y = A + E$).

In the GWAS literature, an empirical estimate of an individual's $A = \sum_{j=1}^{L} X_j \alpha_j$ is often called that individual's **polygenic score**. It is only an estimate because finite sample size in the GWAS used to estimate SNP j's weight means that the weight is perturbed by a sampling error. For this reason we might use the term **true polygenic score** as a synonym or breeding or additive genetic value.

The variance of breeding values, $\mathrm{Var}(A) = \mathrm{Var}\left(\sum_{j=1}^{L} X_j \alpha_j\right)$, is a quantity of fundamental importance in quantitative genetics and therefore has a special name: the **additive genetic variance**. In the case that the causal sites are all uncorrelated, the additive genetic variance can be broken down into the separate variances contributed by the individual causal sites.[9] The ratio of the additive genetic to the total phenotypic variance,

$$\frac{\mathrm{Var}(A)}{\mathrm{Var}(Y)} = h^2, \tag{2}$$

is called the **heritability in the narrow sense**.

If breeding values and residuals are uncorrelated, then the narrow-sense heritability is the fraction of the total trait variance caused by individual differences in breeding values. Loosely speaking, a near-zero value means that most of the individual differences in the trait are caused by variables *other than* how many + genes are carried at all causal sites. A value near one means the opposite: most of the individual differences in the trait *are* caused by differences in the total number of + genes across the genome.

Referring back to our class notes on statistics, we can see that the narrow-sense heritability is simply the total R^2 in the multiple regression of the phenotype on all causal sites, in the case that breeding values and residuals are uncorrelated.

We now define a new random variable, the **genotypic value** or **total genetic value**, G. The meaning of G is best explained by reference to Figure 5. The breeding value (A) for individuals with no + genes is 10.8. But in a sense this is an approximation; it is the best prediction of an individual's phenotype that we can get from knowing the average effect at this site and the fact that the individual has no + genes. The truth is that the average phenotypic value of individuals with no + genes is really 10. We can thus regard

[9] Even if there are correlations among causal sites, Fisher (1999) derived an expression for the additive genetic variance that is still a sum of terms each corresponding to a causal site.

the true phenotypic value of individuals with a given genotype (G) as the sum of the breeding value (10.8 in this case) and a discrepancy due to the nonlinear interaction of genes at different locations (-0.8 in this case). The respective values of G for individuals carrying zero, one, and two + genes are thus 10, 13, and 11.

This notion carries over to the realistic case of multiple causal sites. In the presence of nonlinear genetic causation, an individual's breeding value $A = \sum_{j=1}^{L} X_j \alpha_j$ differs from the total genetic value (the phenotypic value yielded by the configuration of alleles across all L sites).[10]

The variance of the total genetic values, $\mathrm{Var}(G)$, is called the **genotypic** or **total genetic variance**. Naturally enough, the difference between the total and additive genetic variance, $\mathrm{Var}(G) - \mathrm{Var}(A)$, is called the **non-additive genetic variance**. The ratio of the total genetic to the phenotypic variance,

$$\frac{\mathrm{Var}(G)}{\mathrm{Var}(Y)} = H^2, \tag{3}$$

is called the **heritability in the broad sense**.[11] H^2 is always larger than h^2 because of the principle that "variance is like a pie"; the total genetic variance contains the additive genetic variance and an additional "slice" representing the non-additive genetic variance.

Figure 5 suggests one way to think about the non-additive genetic variance. In the unrealistic case of just one causal site, the nonlinear deviations are equivalent to the vertical discrepancies from the regression line (δ's). The non-additive genetic variance is then the variance of the δ's in our population of interest.

Here is one way to interpret the broad-sense heritability. Suppose that total genetic values and environmental disturbances combine additively (without interacting) and also that genotypes and environments are uncorrelated. Then if we could zap everyone's genotypes at all causal sites so as to be identical—in essence turning the population into a large brood of clones—then the phenotypic variance would shrink by the factor H^2. For example, suppose that the phenotypic variance is 100, the broad-sense heritability is 0.75, and the environment stays constant across time. Zapping the next generation of zygotes in the manner described would result in a cohort exhibiting a variance equal to a mere 25 at the time of measurement—a sizable reduction in individual differences. *If total genetic values and environments are uncorrelated, then the broad-sense heritability is the fraction of the trait variance caused by individual differences in total genetic values.*

[10] The breeding values and nonlinear deviations can still be visualized in the case of two causal sites. If the genetic causation is completely linear, then the total genetic values lie in a plane and coincide with the breeding values. In the nonlinear case, the breeding values still define a plane, but now the total genetic values deviate vertically from this plane.

[11] If genotypes and environments are correlated in such a way that "good" genes tend to be carried by individuals in "bad" environments, then the heritability in either sense might exceed one. This situation, however, is extraordinarily unlikely. It should always be safe to assume that the heritability of a trait is a true proportion.

A total genetic value includes both the breeding value and a deviation due to nonlinear interactions among genes at different locations.

When the term "heritability" is used without specifying the narrow or broad sense, usually the narrow sense is intended, although this usage is not universal.

Much of this discussion has been admittedly quite abstract. To make some of these ideas more concrete, take a look at Figure 6. In lecture I will talk about genetic studies of educational achievement. rs9320913 is one of the sites (SNPs) that have been found to be correlated with years of education (Rietveld et al., 2013). Now suppose that rs9320913 is a true causal site affecting education. Perhaps having more + genes at this site causes you to stay in school longer because it makes you smarter and thus better at school. This SNP has also been found to be correlated with IQ (Rietveld et al., 2014), so this is a plausible story. The two alleles at this site are `C` and `A`; since the latter is the allele associated with more education, it is the + allele in this example.

The average effect of gene substitution on education is estimated to be *one month*. In other words, if we can perform our hypothetical gene-zapping experiment at this site and thereby give someone with either zero or one + genes an additional such gene, then *on average* we expect the person to stay in school for one more month.

Actually, most of the time people quit school at the end of an academic term or year rather than in the middle. So probably what is going on is that people with more + genes are *slightly* less likely to drop out at the end of any given term or year and therefore tend *ever* so slightly to finish with more years of education—by an amount that comes to one month per + gene after averaging across many people staying in school for different amounts of time.

Now it may be that a gene zapping that replaces a − gene with a + gene at rs9320913 will lead to more time in school for one person, but have no effect at all on a second person. Maybe the second person does become smarter, but for whatever reason is not interested in going to college. Or maybe the second person grows up in such a poor environment that his enhanced natural ability is not enough to overcome barriers to pursuing higher education. But if the average effect of gene substitution is truly equal to one month, what this means is that the average hypothetical gain across many different people—some of whom might benefit more than others—is precisely one month.

2.3 Heritability and the Correlations Between Relatives: A First Pass

How can the heritability of a trait be estimated from empirical data? The basic idea is that more highly heritable traits result in higher correlations between the phenotypes of biological relatives. The correlations should also increase with the closeness of the relationship, since individuals who are more closely related share more genetic material. Therefore, if we observe high correlations between biological relatives that increase with the degree of relatedness, we can infer that the trait has a high narrow-sense heritability.

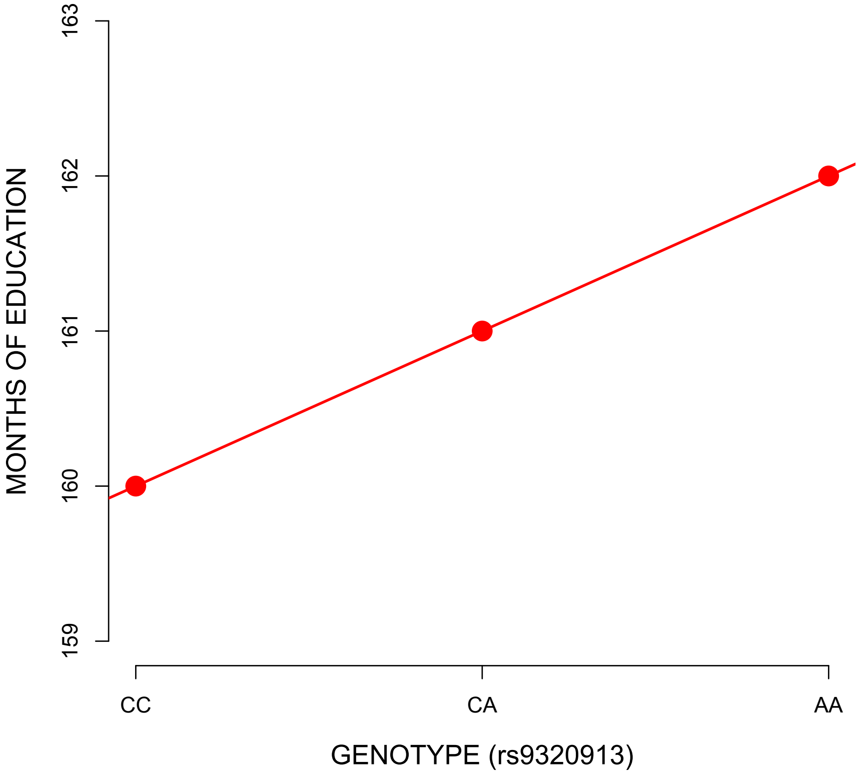


Figure 6: The association between the SNP rs9320913 and education. The x-axis represents the number of `A` base pairs carried by given individual at this site, and the y-axis represents educational attainment (in months). The straight line that best fits the genotype-education relationship is shown in red. The slope of this line is equal to the average effect of gene substitution (how much additional education is obtained, on average, as a result of bearing an additional copy of `A`).

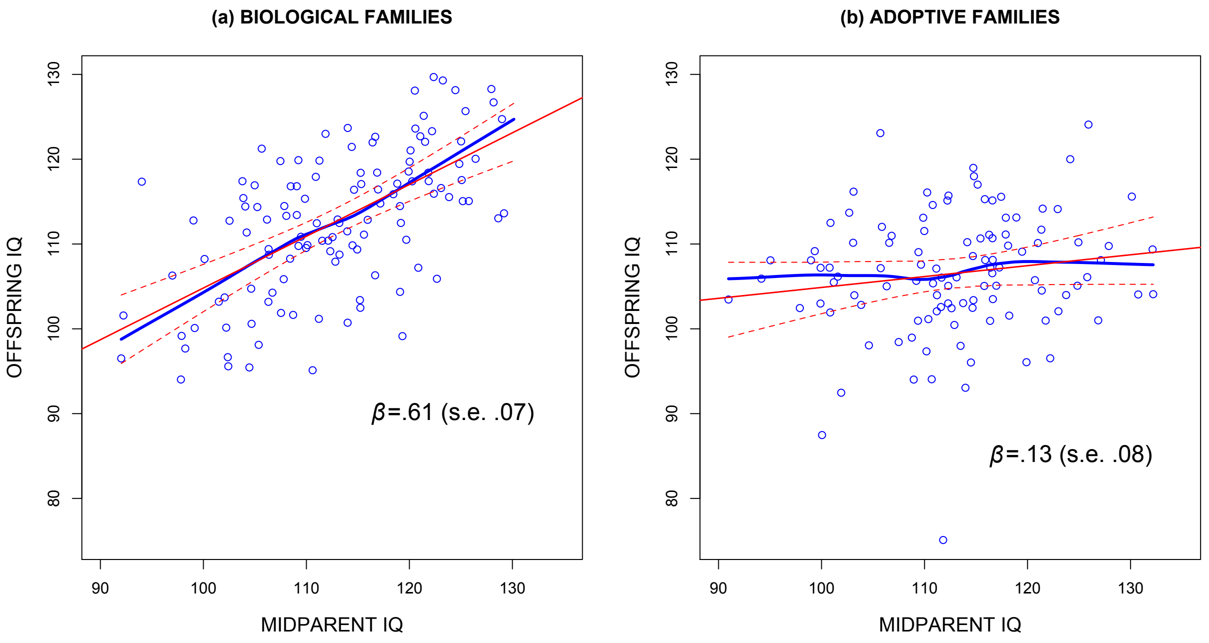


Figure 7: Midparent-offspring regressions for IQ scores, in both (a) biological and (b) adoptive families (Scarr & Weinberg, 1978; Scarr, 1997). The solid red lines are the standard least-squares regression lines; the slopes of these lines are the corresponding regression coefficients (β).

Figure 7 gives a rough idea of how the value of heritability can be inferred in this way. The panel on the left shows that there is a strong correlation between parent and offspring IQ in biological families. But without any other information, it is not certain that this correlation points toward high heritability; it is conceivable that the correlation arises from the stimulating environment that smarter parents provide for their children. However, when we look at the panel on the right, we see virtually no relationship between parent and offspring IQ in adoptive families (where the children probably share no more genetic material with their parents than expected by chance). This makes it highly likely that the parent-offspring IQ resemblance in biological families is due primarily to genetic inheritance. Therefore we can reasonably conclude that IQ does indeed have a high heritability.

Furthermore, because of the randomness inherent in Mendel's laws, ordinary siblings (including dizygotic twins) and parent-offspring pairs are not as genetically similar as monozygotic twins (who in fact are genetically identical). It turns out that monozygotic twins show an even higher IQ correlation than parent-offspring pairs. The fact that family resemblance increases with genetic relationship affirms that IQ has a high heritability.

Our textbook on pp. 126–132 gives an explanation of how to use the correlations

Table 1: The correlations between relatives due to the additive genetic variance in a randomly mating population

relationship	correlation
monozygotic twins	h^2
parent-offspring	$\frac{1}{2}h^2$
grandparent-grandchild	$\frac{1}{4}h^2$
full siblings (including dizygotic twins)	$\frac{1}{2}h^2$
half siblings	$\frac{1}{4}h^2$
uncle-nephew (aunt-niece)	$\frac{1}{4}h^2$
first cousins	$\frac{1}{8}h^2$

between relatives to estimate heritability in a more quantitative way. Please be aware that this explanation, although acceptable for an introductory text, contains some errors. Giving a correct explanation of the relationship between heritability and the correlations between relatives using the textbook's approach of "gene sharing" is actually quite difficult, and I will not attempt to do so here. The next subsection explains heritability estimation using an entirely different approach, but this material is entirely optional.

The second column of Table 1 expresses each correlation as a function of the narrow-sense heritability. For example, if the narrow-sense heritability is 0.60, then the correlation between between full siblings due to the narrow-sense heritability is $(1/2)(0.60) = 0.30$.

Remember that this 0.30 is *only* the part of the total sibling correlation due to the narrow-sense heritability. The total sibling correlation itself may in principle be larger than this. For example, siblings growing up in the same home might resemble each other for environmental reasons as well. This is why simply equating an observed correlation between relatives to the theoretical one in Table 1 and solving for h^2 is rarely a good idea. In practice, we use multiple kinships, including adoptive relationships where the individuals are not biologically related, in an attempt to converge on the heritability.

You should study the pattern in Table 1 until it makes sense to you: as relatives become biologically closer, the correlation due to the narrow-sense heritability becomes larger.

Here is an example of how the results in Table 1 can be used in a slightly more sophisticated way to estimate heritability. Suppose that the correlation between monozygotic (MZ) twins is 0.85, the correlation between dizygotic (DZ) twins is 0.50, the correlation between mates is zero, and all of the genetic variance is additive. We assume that any resemblance between twins caused by sharing the same rearing environment increases both the MZ and DZ correlations by the same amount. In this case it is reasonable to think that subtracting the DZ from the MZ correlation might eliminate the environmental contribution and leave only a genetic component.

It is common to use the symbol c^2 denote any contribution to the correlation between twins from shared environment. We find that

$$\begin{aligned}
&2\,[\mathrm{Corr}(\text{MZ twins}) - \mathrm{Corr}(\text{DZ twins})] \\
&= 2\left[\left(h^2 + c^2\right) - \left(\frac{1}{2}h^2 + c^2\right)\right] \\
&= h^2.
\end{aligned}$$

In words, twice the difference between the MZ and DZ correlations leads to an estimate of the narrow-sense heritability. In this numerical example, the narrow-sense heritability is thus estimated to be $2(0.85 - 0.50) = 0.70$.

There will be additional genetic contributions to the correlations between relatives—on top of those coming from the narrow-sense heritability in Table 1—if the genetic sites interact nonlinearly to produce non-additive genetic variance. In this course the most important feature of these additional genetic contributions to family resemblance is that they contribute far more to the MZ correlation than to the others. One way to understand this is to examine Figure 5 and realize that MZ twins, by virtue of their identical genotypes at all sites, share both their breeding values *and* nonlinear deviations.

Another way to put it is that, because of the randomizing tendency inherent in Mendel's laws, any nonlinear deviation of the total genetic value from the breeding value that depends on a precise configuration of alleles at multiple sites is very unlikely to be shared by any relative other than an MZ twin. For example, suppose that a very peculiar phenotype of mine (e.g., always drinking a beer can without touching it with my pinky) depends on the exact allelic states of 100 genes, all at different sites. The probability that all 100 genes are passed on to my child is $(1/2)^{100} \approx 7.9 \times 10^{-31}$, an incredibly small number. In contrast, my monozygotic twin is guaranteed to have inherited all 100 genes from our parents and so is very likely to exhibit this odd phenotype. According to Lykken, McGue, Tellegen, and Bouchard (1992), this principle explains why MZ twins reared apart show so many jaw-dropping resemblances of a kind that are hardly ever observed in DZ twins reared apart.

A very rapid decay of resemblance with decreasing relatedness is evidence of strong nonlinear genetic interactions across multiple sites. For example, if the MZ correlation is 0.70 but the DZ and parent-offspring correlations are only about 0.10, then this is consistent with most of the total genetic variance being non-additive in nature. Among other things, such strong nonlinearity will invalidate the simple subtraction method of heritability estimation using MZ and DZ twins reared together.

We close this subsection with two points:

1. Tables 1 and 2 do not give the contribution to the correlations between relatives from non-additive genetic variance. It turns out that the non-additive genetic variance

contributes fully to the MZ correlation. Therefore the correlation between MZ twins separated at birth and reared apart in uncorrelated environments provides an estimate of the *broad-sense* heritability.

2. Suppose that we perform a regression analysis where the x-axis variable is the average of the two parental phenotypes and the y-axis variable is the average of the offspring phenotypes. Figure 7 provides an example. Then if the only contribution to the parent-offspring resemblance is the average effects of their shared genes, the slope of the regression line is equal to the *narrow-sense* heritability. This result is proved in the next subsection. Even if there are other contributions to parent-offspring resemblance, such as non-additive genetic variance, this regression method usually provides a better estimate of the narrow-sense heritability than methods employing twins.

   The right panel suggests of Figure 7 suggests, perhaps surprisingly, that any non-genetic contributions to parent-offspring IQ resemblance are quite weak. Therefore we may reasonably suppose that the narrow-sense heritability of IQ lies somewhere between 0.45 and 0.60.

*2.4 Heritability and the Correlations Between Relatives: A Second Look

This subsection is entirely optional. Students pursuing advanced research in behavioral genetics may profit from it; the correlations between relatives are rarely explained in this way, despite the fact that I find this style of explanation very intuitive and appealing. The method is owed largely to Wright (1968, 1969).

We assume additive gene action and the linearity of all bivariate relationships. We no longer assume random mating; that is, we now allow the phenotypes of mates to be correlated. We do assume, however, a very specific model of how the mate correlation arises: the probability that a given couple "hits it off" and mates successfully is affected by their phenotypic values. Such assortative mating will change the genetic composition of the population, and therefore we also assume that the population has reached an equilibrium. The same results can hold under different and often less restrictive assumptions (Fisher, 1918; Wilson, 1973; Gimelfarb, 1981; Nagylaki, 1982).

The figure shows a path diagram of the causal system determining the phenotypes of a nuclear family with two offspring. It is convenient to standardize all breeding values and phenotypes. Each directed edge in the diagram represents a causal influence of the head variable on the tail variable. Some of the directed edges are labeled by a "path coefficient," which corresponds to the expected change in the tail variable under an experimental manipulation changing the head variable by one unit and holding constant all other non-tail variables.

Table 2: The correlations between relatives due to the additive genetic variance in an assortatively mating population

relationship	correlation
monozygotic twins	h^2
parent-offspring	$\frac{1}{2}(1+\rho)h^2$
grandparent-grandchild	$\frac{1}{4}(1+\rho)\left(1+\rho h^2\right)h^2$
full siblings (including dizygotic twins)	$\frac{1}{2}\left(1+\rho h^2\right)h^2$
half siblings	$\frac{1}{4}\left(1+2\rho h^2+\rho^2 h^2\right)h^2$
uncle-nephew (aunt-niece)	$\frac{1}{4}\left(1+\rho h^2\right)h^2$
first cousins	$\frac{1}{8}\left(1+\rho h^2\right)^3 h^2$

h^2 is the narrow-sense heritability, and ρ is the phenotypic correlation between mates.

The correlation between any two variables is equal to the sum over connecting paths of coefficient products. A connecting path is any path (sequence of edges passing through any variable no more than once) between the two focal variables satisfying one of the following conditions:

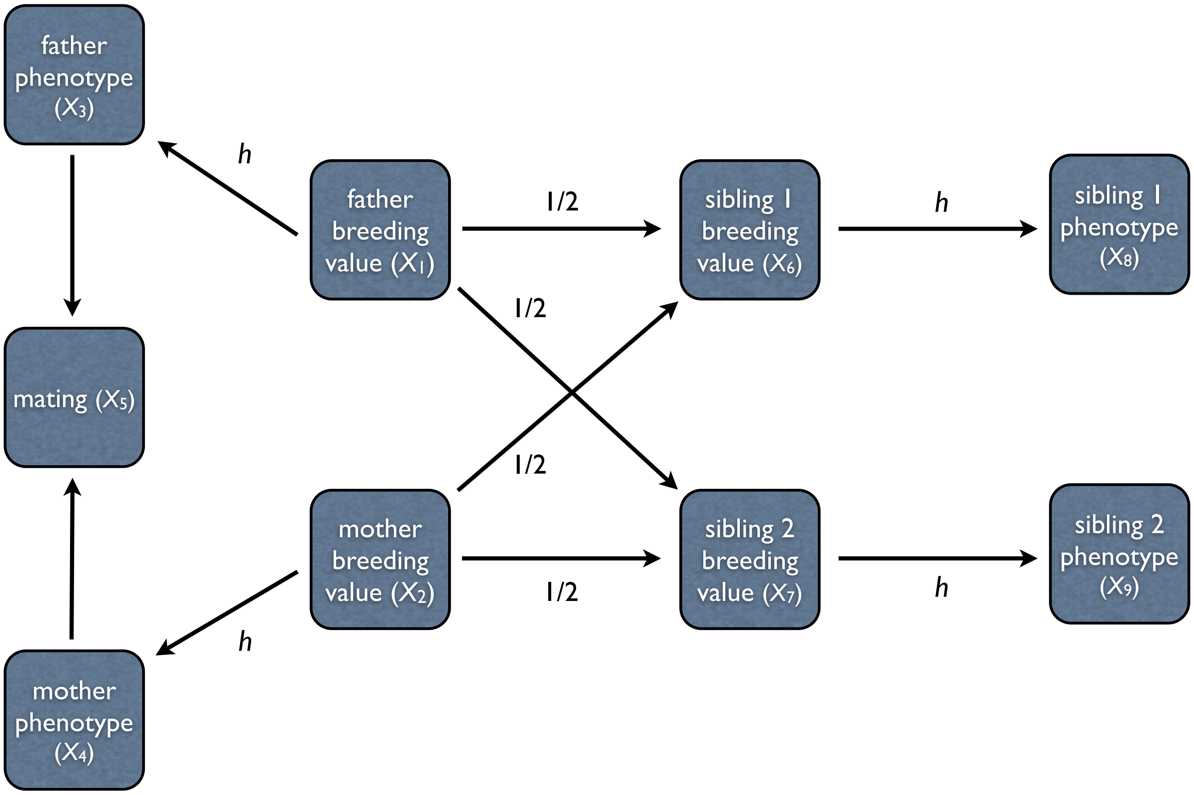
1. The path consists of directed edges all pointing in the same direction.

2. The path travels against the arrowheads to a variable called a "confounder" and then travels with the arrowheads to the finish at the other focal variable.

3. Normally a path is *not* a connecting path if it contains a "collision" of two arrowheads at an intermediate variable. This corresponds to the commonsensical notion that two causes of the same effect may act independently. If both rain and a sprinkler can cause a pavement to become wet (*rain* → *wet* ← *sprinkler*), it does not follow that rainfall and sprinkler activation are correlated. However, if we condition on the intermediate variable, then the path is "activated" and becomes connecting (correlation inducing). For example, if we observe the pavement only on mornings when it is wet, then the fact that it did not rain during the night means that the sprinkler must have turned on.

To apply these rules, we need to know the relevant path coefficients. Since heritability (h^2) is by definition the proportion of the phenotypic variance caused by differences in breeding values, each path coefficient from breeding value to phenotype has the value h. Now consider how gene substitutions affecting a parent are propagated to its offspring. Each such replacement brings about a change in breeding value equal to the average effect at the targeted site. Recall that a parent only transmits one random gene from each of its

Figure 8: A path diagram of the causal system determining the phenotypes of a nuclear family with two offspring.


father phenotype (X_3)
h
father breeding value (X_1)
1/2
sibling 1 breeding value (X_6)
h
sibling 1 phenotype (X_8)
1/2
mating (X_5)
1/2
mother breeding value (X_2)
1/2
sibling 2 breeding value (X_7)
h
sibling 2 phenotype (X_9)
h
mother phenotype (X_4)

homologous pairs to the offspring. Therefore, at a site where the parent has experienced one gene substitution, the probability of the offspring receiving the altered gene is 1/2. At a site where both parental genes have been altered (thereby increasing the parent's breeding value by twice the average effect), the fact that the offspring receives only a single member of the pair means that its own breeding value is expected to undergo a change equal to half of the change to the parent's breeding value. Thus, regardless of whether one or both parental genes are altered at a given site, half of the resulting change is transmitted to the offspring. Given the constancy of the variance in breeding values across generations, it follows that each path coefficient from parent to offspring breeding value has the value 1/2.

We have implicitly conditioned on a successful mating between phenotypically discriminating individuals of the opposite sex within each family. The path $X_3 \to X_5 \leftarrow X_4$ is thus unblocked, allowing traces to proceed through the arrow-to-arrow collision. In the case that all individuals mate (the only question being with whom), the contribution of the trace through these three variables to the product of coefficients along the embedding path is simply the correlation between spouses.[12]

Taking the father and the first sibling for concreteness, we see that the two parent-offspring connecting paths are

$$X_3 \overset{h}{\leftarrow} X_1 \xrightarrow{1/2} X_6 \overset{h}{\rightarrow} X_8,$$
$$X_3 \overset{\rho}{\rightleftarrows} X_4 \overset{h}{\leftarrow} X_2 \xrightarrow{1/2} X_6 \overset{h}{\rightarrow} X_8.$$

Adding these two products of coefficients, we recover the classical result that the parent-offspring correlation ascribable to the average effects of transmitted genes is

$$\mathrm{Corr}(P, O) = \frac{1}{2}(1+\rho)h^2.$$

It is of interest to find the slope in the regression of offspring phenotype on the mean of the two parental phenotypes:

$$\begin{aligned}\beta &= \frac{\mathrm{Cov}\left(O, \frac{M}{2}+\frac{F}{2}\right)}{\mathrm{Var}\left(\frac{M}{2}+\frac{F}{2}\right)} \\ &= \frac{\frac{1}{2}\mathrm{Cov}\left(O, M\right)+\frac{1}{2}\mathrm{Cov}\left(O, F\right)}{\frac{1}{4}\mathrm{Var}\left(M\right)+\frac{1}{4}\mathrm{Var}(F)+\frac{1}{2}\,\mathrm{Cov}\left(F, M\right)} \\ &= \frac{\frac{1}{2}(1+\rho)h^2}{\frac{1}{2}+\frac{1}{2}\rho} \\ &= h^2.\end{aligned}$$

If there are no other causes of parent-offspring resemblance, the midparent-offspring regression offers perhaps the simplest means of estimating the narrow-sense heritability. Note that the use of this method does not require knowledge of the mating pattern.

The correlation between siblings ascribable to commonly inherited genes can also be obtained from the figure. The four connecting paths are

$$X_8 \overset{h}{\leftarrow} X_6 \xleftarrow{1/2} X_1 \xrightarrow{1/2} X_7 \overset{h}{\rightarrow} X_9,$$
$$X_8 \overset{h}{\leftarrow} X_6 \xleftarrow{1/2} X_2 \xrightarrow{1/2} X_7 \overset{h}{\rightarrow} X_9,$$
$$X_8 \overset{h}{\leftarrow} X_6 \xleftarrow{1/2} X_1 \overset{h}{\rightarrow} X_3 \overset{\rho}{\rightleftarrows} X_4 \overset{h}{\leftarrow} X_2 \xrightarrow{1/2} X_7 \overset{h}{\rightarrow} X_9,$$
$$X_8 \overset{h}{\leftarrow} X_6 \xleftarrow{1/2} X_2 \overset{h}{\rightarrow} X_4 \overset{\rho}{\rightleftarrows} X_3 \overset{h}{\leftarrow} X_1 \xrightarrow{1/2} X_7 \overset{h}{\rightarrow} X_9,$$

from which it follows that

$$\mathrm{Corr(siblings)} = \frac{1}{2}\left(1+\rho h^2\right)h^2.$$

Given a suitable extension of the pedigree in Figure 8, the other correlations in Table 2 can be similarly derived.

[12] The explanation given by Lynch and Walsh (1998) regarding this type of path is incorrect.

3 Modern Genetic Research with DNA-Level Data

Whereas studies of twins, nuclear families, adoptees, and other kinships allow us to make indirect estimates of a trait's heritability, the aim of gene-mapping studies is to identify the actual causal sites in the human genome that are responsible for the heritability. If indirect estimates based on classical quantitative-genetic theory are accurate, then in principle the variance contributed by all causal sites identified in gene-mapping studies will eventually equal the heritability.

3.1 Detection of Individual Causal Sites

If the genetic and environmental causes of a trait are not correlated, then the average effects of gene substitution at the different causal sites are equivalent to the least-squares regression coefficients obtained by performing a multiple linear regression of the phenotype on all causal sites. The sample size of this hypothetical regression study may need to be very large if all regression coefficients are simultaneously estimated.

Multiple regression is more or less how gene-mapping studies estimate the magnitudes of average effects. It is hoped that confounding by past events affecting both the genetic composition of the population and the distribution of the phenotype is mild enough to ignore (or to be controlled by statistical methods), and then the regression coefficient at a given site is estimated. Figure 6 gives an example. In a large sample, we perform a regression of the phenotype on the number of + genes (0, 1, or 2). The resulting regression coefficient (the slope of the red line) is the average effect of gene substitution at this particular site. Of course, in practice we would not know which allele to call "+" until after the regression analysis is performed.

Modern genotyping and sequencing technology allows geneticists to conduct **genome-wide association studies** (GWAS) that look for nonzero regression coefficients at many millions of sites across the genome. A strict threshold of statistical significance ("$p < 5 \times 10^{-8}$" instead of the usual "$p < 0.05$") is applied to pick out a subset of sites with strong evidence for nonzero regression coefficients with respect to the studied phenotype. GWAS conducted so far have found that most phenotypes of interest to geneticists are affected by thousands of sites. In other words, a typical causal site contributes less than 0.1 percent of the total phenotypic variance. Since such small effects on a given trait are difficult to detect given a strict significance threshold, in most cases the majority of the heritability has not yet been pinned down to specific sites.

It should be pointed out that identifying the precise causal variant in a region yielding a GWAS signal is quite difficult because of the tendency for pairs of sites in a relatively small region to show LD. That is, because a person inheriting a certain allele at one site will tend to inherit particular correlated alleles at nearby sites, the presence of one causal site with a true nonzero average effect will lead to spurious correlations with the phenotype at multiple sites in the region. Developing routine procedures for identifying the causal sites responsible for GWAS signals continues to be a challenge for laboratory scientists and bioinformaticians.

How can we be sure that a GWAS signal represents any causal site at all? After all, correlation does not imply causation, and GWAS are not randomized experiments but rather observational studies. There is always the possibility that given site is correlated with a confounding environmental variable that affects the phenotype. One remarkable feature of GWAS, however, is that we can check the results with a design that is nearly equivalent to a randomized experiment (Lee, 2012). Recall that a parent transmits one of its two genes at a particular site *at random* to any given offspring (Figure 2). Gene-mapping studies that take advantage of this feature are called **family-based studies**. There are different kinds of family-based studies, but perhaps the simplest to understand involves the genotyping of two parents (at least one of whom is heterozygous) and two offspring. If there is a significant tendency across many families of this type for the sibling inheriting more + genes from heterozygous parents to exhibit a higher phenotypic value than the sibling inheriting fewer + genes, then we have strong evidence that the genotyped site (or a nearby correlated site) does indeed affect the phenotype. This is because which sibling ends up with more + genes at a given site is determined entirely by Mendelian coin flips that cannot affect or be affected by outside variables.

Family-based studies have tended to confirm GWAS results from studies of unrelated individuals that are in principle subject to confounding. For example, of the first 416 sites reported to be correlated with height in samples of unrelated people, 371 were replicated in family cohorts (Wood et al., 2014). This 89-percent consistency is quite remarkable. A similar result has been obtained in GWAS of schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). For whatever reason, it seems that gene-mapping studies are subject to much less confounding than observational studies in other fields of science. Other evidence for this generalization comes from:

1. the strong tendency of GWAS findings in Europeans to replicate in other ethnicities (including East Asians, South Asians, and Africans) and even other species; and

2. the clustering of GWAS signals for a particular phenotype in and around regions of the genome encoding proteins already known to be biologically related. Often these biological pathways also appear to have a plausible connection with the phenotype. For example, DNA sites correlated with height are overrepresented in regions already implicated in skeletal development, whereas those correlated with psychological phenotypes such as schizophrenia are overrepresented in regions implicated in the central nervous system.

It is tempting to say that the Mendelian mechanism inherently tends to break down correlations between genotypes and potential confounders, but it is difficult to formulate this idea outside of the family setting in a formal way.

3.2 Estimating Heritability Using Unrelated Individuals

Even if the causal sites cannot all be identified because current sample sizes are not large enough to detect very small effects, modern genetic data enables a novel method for estimating narrow-sense heritability that uses *unrelated* individuals.[13] In the literature this method is often called "GCTA" after a well-known software package (Genome-wide Complex Trait Analysis) that researchers can use to apply the method (Yang, Lee, Goddard, & Visscher, 2011).

The basic foundation of this method is that even pairs of unrelated individuals will vary in genetic similarity as a result of sheer chance. In other words, if we randomly sample two other individuals from the population who are not related to you or to each either, it will inevitably turn out to be the case that your genome is *very slightly* more similar to the genome of one person (A) than to the genome of the other (B). Then we can ask: is it also the case that your *phenotypic value* is very slightly closer to the phenotypic value of person A than to the phenotypic value of person B? If the answer is *yes* after considering many pairs of individuals, then we have fairly strong evidence that the phenotype is heritable. The strength of the relationship between directly measured DNA-based genetic similarity and phenotypic similarity tells us the magnitude of the narrow-sense heritability.

The GCTA method is very attractive because it does not rely on close biological relatives. Unless biological relatives are separated by adoption, it is always the case that they share both genes *and* environment, making it somewhat difficult to interpret their phenotoypic correlation. But unrelated individuals who by chance happen to be minutely similar at the DNA level are extremely unlikely to share environmental experiences. Furthermore, the GCTA method supplies an internal check on whether genetic similarity is confounded with environmental similarity, although the details of this procedure are beyond the scope of these notes. Another attractive feature of the GCTA method is that its estimates of narrow-sense heritability are not inflated by non-additive genetic variance. Remember that the contribution of non-additive genetic variance is greatest for the correlation between MZ twins and then rapidly decays as the degree of relatedness declines. If we take this decay to the limit of the very small chance levels of genetic similarity exhibited by unrelated individuals, then all that remains is additive genetic variance.

GCTA-based heritability estimates are always smaller than those based on the correlations between close relatives. For example, whereas Figure 7 suggests that the narrow-sense heritability of IQ might be as high as 0.60, GCTA studies have so far returned estimates of ~ 0.35. One reason for the discrepancy may be that estimates from the correlations between close relatives remain slightly inflated even after attempts to account for

[13] Strictly speaking, all individual are genealogically related if we go back far enough in time. In lecture, we learned that we are all related to chimpanzees if we go back roughly 5 million years. When we say "unrelated individuals," we mean that the common ancestry of the two people is far enough back in time that they are no more closely related than, say, second cousins.

environmental sources of similarity and nonlinear interactions. Another possible source of the discrepancy is that current genotyping technology only targets sites where the **minor allele** (the allele with lower frequency) is still reasonably common. Sites with low **minor allele frequencies** (MAFs) are more abundant overall than sites with higher MAFs, and there are reasons to believe that *causal* sites with low MAFs are characterized by larger average effects and thus make a disproportionate contribution to the genetic variance. As DNA studies begin to measure increasingly more sites with low MAFs, GCTA-based heritability estimates should show some convergence with estimates based on the correlations between close relatives.

My opinion is that the original explanation of why GCTA works given by Yang et al. (2010) is not fully satisfactory. Lee and Chow (2014) provide what is hopefully a more transparent account.

3.3 Estimating Genetic Correlations Between Traits

The **genetic correlation** between two phenotypes is simply the correlation between their respective breeding (additive genetic) values. For example, suppose that we are interested in the genetic correlation between schizophrenia (SCZ) and bipolar disorder (BPD). Then write

$$\begin{aligned} Y_{\text{SCZ}} &= \alpha_0 + \underbrace{X_1\alpha_1 + \cdots + X_L\alpha_L}_{\text{SCZ breeding value}} + E_{\text{SCZ}} \\ Y_{\text{BPD}} &= \beta_0 + \underbrace{W_1\beta_1 + \cdots + W_K\beta_K}_{\text{BPD breeding value}} + E_{\text{BPD}}. \end{aligned} \tag{4}$$

If we could somehow lop off the residual ("environmental") terms on the far right and calculate the correlation between the leftover genetic parts, this quantity would be the genetic correlation between schizophrenia and bipolar disorder. Two traits might show a positive and large genetic correlation because many of the L sites affecting one trait overlap with the K sites affecting the other.

How can genetic correlations actually be calculated in practice? This is often done with twin data in an extension of the method sketched in Section 2.3. But nowadays we have GWAS data, and we can use these as well. One popular method for estimating genetic correlations is in fact GCTA, extending its use in the estimation of heritabilities (Lee, Wray, Goddard, & Visscher, 2011). Another method is known as LD Score regression (Bulik-Sullivan et al., 2015); more methods are constantly being introduced.

I would say that the basic idea behind all of these methods is as follows. Suppose it is the case that a SNP with a positive and relatively large effect on liability to schizophrenia also has a positive and relatively large effect on liability to bipolar disorder. If there is a consistent tendency for effects on the two traits to have same sign across the entire genome, then there is indeed a genetic correlation between the traits that can be estimated by the methods just mentioned.

4 Population Genetics

The field of **population genetics** is concerned with how the genetic composition of populations (allele frequencies and related quantities) are affected by various evolutionary forces. Quantitative genetics, the subject of Section 2, is sometimes regarded as a specialized branch of population genetics.

All genetic differences originate by **mutation**, a random change in the allelic state of a gene as it is transmitted from parent to offspring. Figure 9 presents a simplified picture of a mutation event. Note that the father and mother both carry the genotype `GG` at site 3. For all we know, the allele frequency of `G` at this site is equal to one in the parental generation. However, the sperm produced by the father carries the allele `A` at this site, and a gene of this allelic type could not possibly have been inherited from the father. A mutation thus occurred at this site during meiosis. The probability that a mutation will occur at any given site is very small—approximately 0.00000001—but over a long timescale this probability is large enough to produce a substantial amount of genetic variability.

As a result of the mutation, the offspring carries the genotype `AG` at site 3. If it was indeed the case that the allele frequency of `G` was one in the previous generation, then the site has become newly polymorphic. Population geneticists usually use the symbol N to stand for population size. Since each person carries two genes at each site, the allele frequency of `A` in the offspring generation is $1/(2N)$, while the allele frequency of `G` is $1 - 1/(2N)$. Of course, if N is at all large, the frequency $1/(2N)$ is quite small.

In Section 2 we further simplified our model of Mendelian inheritance by considering only causal sites and labeling the allele causing lower phenotypic values "$-$" and the allele causing higher values "$+$" (Figure 4). In this way of looking at things, we have to anticipate the phenotypic consequence of the mutation before labeling the ancestral allele. Suppose that the new mutation will tend to reduce the phenotypic values of its bearers. This is quite plausible in the case of new mutations affecting general intelligence; random changes to a complex machine like the brain are not likely to improve its functioning.[14] Then we can say that the allele frequency of $+$ before the mutation event was one and that the allele frequency of $-$ after the mutation event is $1/(2N)$.

Even in the absence of additional mutations or other evolutionary forces, allele frequencies never stay exactly constant. For example, even if the offspring in Figure 9 grows up and has several children of his own, by chance none of these children might inherit the new `A` allele; the Mendelian coin might come up tails every single time. Then the frequency of `A` will be back at zero in the next generation. Even at site 9, where it seems likely that both alleles `T` and `C` are common, the frequencies are subject to these Mendelian fluctuations. Suppose that the allele frequencies at site 9 are both 0.50 in the parental generation. Then they might become $\{0.49999, 0.50001\}$ in the offspring generation as a

[14] Indeed, we know of many rare mutations that cause mental retardation, but of none that cause their bearers to become intellectually gifted.

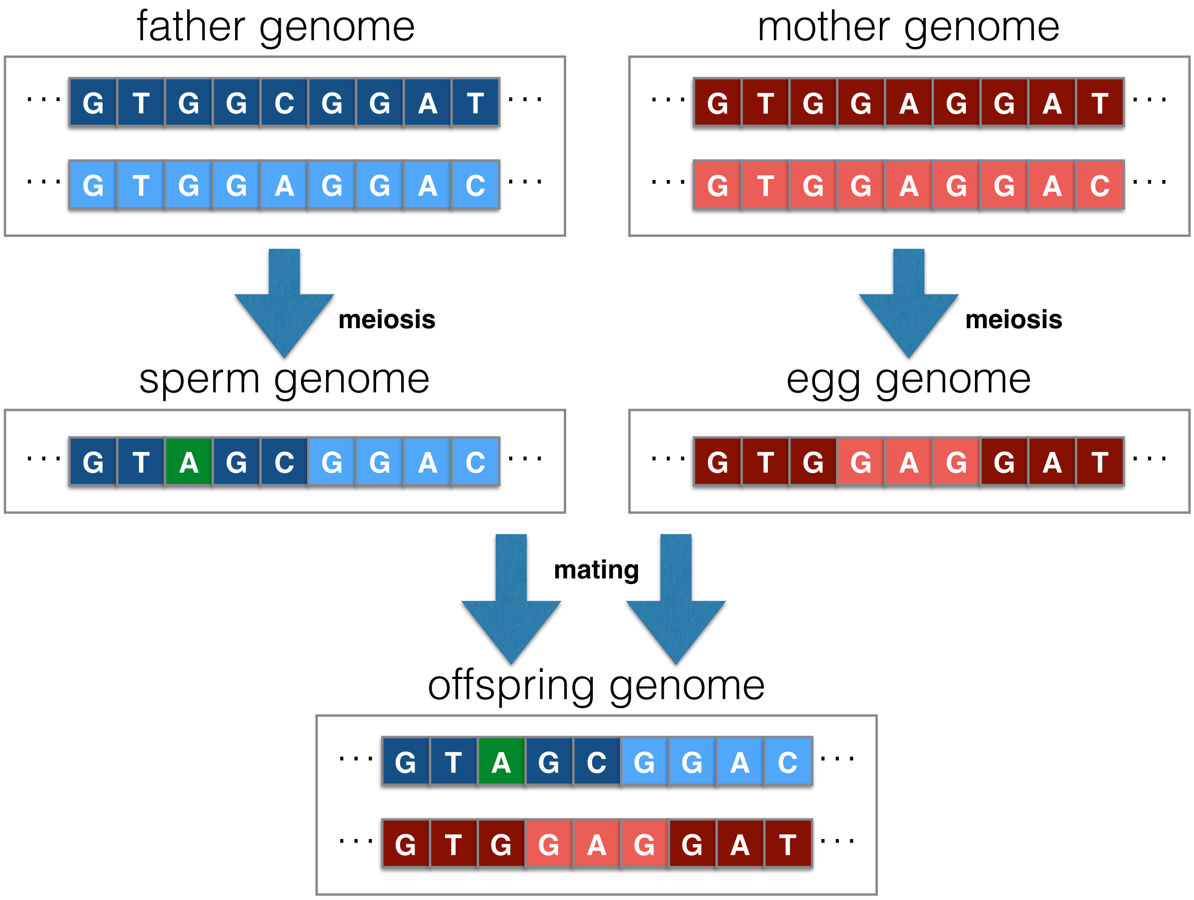

father genome
··· G T G G C G G A T ···
··· G T G G A G G A C ···
mother genome
··· G T G G A G G A T ···
··· G T G G A G G A C ···
meiosis
meiosis
sperm genome
··· G T A G C G G A C ···
egg genome
··· G T G G A G G A T ···
mating
offspring genome
··· G T A G C G G A C ···
··· G T G G A G G A T ···

Figure 9: This is the same sequence of events at the nine-site region in Figure 1 except that a mutation (green) has been added.

result of random transmission. Such changes in allele frequencies over time attributable to random, undirected forces are known as **genetic drift**. Over a sufficiently long time, genetic drift can change allele frequencies at many sites from zero to one, although this fact may not be intuitive at first. Many of the fixed differences between the human and chimpanzee genomes are undoubtedly the result of genetic drift.

The interesting *phenotypic* differences between chimpanzees and humans, however, are probably the result of **natural selection**—heritable variation in fitness leading to the population becoming genetically and phenotypically more like its fittest members in previous generations. Our textbook contains a reasonably good description of natural selection (pp. 154–156), and here we will step through a hypothetical example.

Suppose that in a population of "chumans" (or "humanzees?") living on the African savannah about 5 million years ago, (1) brain volume is heritable and (2) individuals with larger brains are better at surviving and reproducing. This may be because such individuals are better at making tools, hunting, negotiating the social hierarchy, or learning how to do these things from their fellows. Whatever the reasons may be, larger brains are correlated with higher fitness in this environment. The consequence is that the parents of the next generation are not a representative sample of the current generation, but rather a sample biased toward higher frequencies of the + alleles at the polymorphic sites affecting brain size. (The selected parents have these higher frequencies because brain volume is in fact heritable. If you do not understand this point, stop and *think* until you do understand it!) These allele frequencies remain elevated in the next generation; the transmission of genes from parents to offspring does not systematically change any allele frequencies. (They may change as a result of genetic drift, but we ignore this complication.) Figure 10 depicts this process at four of the many polymorphic sites affecting brain size. The blue bars correspond to the initial allele frequencies, and the green bars to the allele frequencies in the selected parents and their offspring.

The magnitudes of the changes in the allele frequencies were computed from standard population-genetic equations that you do not need to know for this course; it was assumed that each site has an average effect of 0.10 standard-deviation units and that the mean phenotypic difference between the selected parents and the total population in the first generation was 2 standard-deviation units.

Let S stand for the just-described phenotypic difference *within* the initial generation, and let R stand for the change in the phenotypic mean *between* generations. If breeding values and residuals are uncorrelated and all variables affecting the phenotype *other* than the allele frequencies maintain the same distributions from one generation to the next, then the relationship between S and R is given by

$$R = h^2 S, \tag{5}$$

which is often called the **breeder's equation**.

Equation 5 is a simple consequence of the equality between the narrow-sense heritability and the slope of the midparent-offspring regression (Section 2). According to the

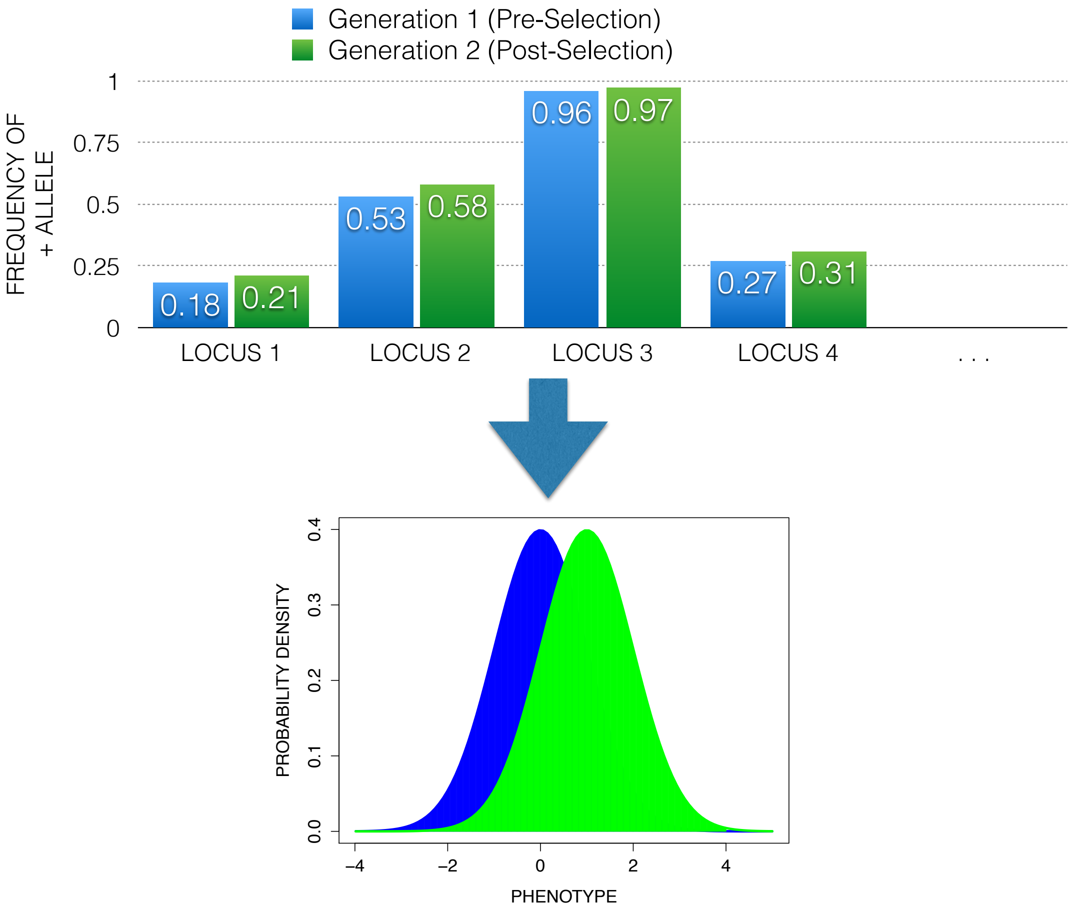


Figure 10: The genetic basis of evolution by natural selection.

interpretation of the regression line given in our materials on statistics, the slope β tells us the mean difference in Y between groups of observations differing by one unit in X. S is the mean phenotypic difference between the selected parents and the total population. Since the Y variable here is offspring phenotype, the mean phenotypic difference between the *actual* offspring and the *hypothetical* offspring that would have been produced in the absence of selection is $\beta S = h^2 S$. The phenotypic mean of the hypothetical offspring would have been the same as the phenotypic mean of the parents (all else being equal), and Equation 5 then follows.[15]

The bottom part of Figure 10 depicts the phenotypic consequence of the underlying changes in allele frequencies. The probability distribution of the phenotype in the offspring generation (green) is shifted to the right; that is, the average brain size of the offspring is larger the average brain size in the previous generation. This is natural selection at work. *Heritable individual differences are the fuel of evolutionary change. The potential response of a trait to natural selection increases with the heritability of the trait.*

In reality, the mean brain size probably never increased in the human lineage by such a large amount in any single generation. Given the amount of time between the chimpanzee-human split and the present day, natural selection only needed to change the mean brain size by an average of 0.00005 modern standard-deviation units each generation to produce the change observed in the fossil record. Another important point is that brain size *per se* was probably not the target of selection. A larger brain imposes many disadvantages, including a difficult birth and high metabolic cost. Natural selection would have thus disfavored larger brains unless they are a cause or effect of some other phenotypic state that improved fitness by a large enough amount to offset the disadvantages. We might reasonably hypothesize that this phenotypic state was a higher level of general intelligence. Confirming such a hypothesis provides a strong motivation for pursuing gene-mapping studies and sequencing DNA from ancient hominin fossils.

5 Summary

- The **genome** is the totality of an organism's genetic material. A genome can be divided into discrete locations that we will call **sites**. Any given individual in a **diploid** species carries two **genes** at any given site, one inherited from each parent. If the population's genes at a site fall into more than one class, the different classes are called **alleles**. A rough analogy to capture the relationships among sites, genes, and alleles is that a site corresponds to a slot in a cash register, genes correspond to coins, and alleles correspond to distinct classes of coins (pennies, nickels, dimes,

[15] The meaning of "all else being equal" can become quite complicated in the presence of nonlinearity (Lee & Chow, 2013). These complications, however, might often fail to matter in practice. For example, Equation 5 will underestimate the magnitude of the evolutionary change between generations if there is a certain kind of non-additive genetic variance, but the surplus phenotypic change will decay in subsequent generations. The *permanent* evolutionary change is thus given by Equation 5.

quarters). One problem with this analogy is that the slot in a cash register usually contains only one type of coin, whereas a person's genes at a given site might represent two different alleles.

- The word **gene** also means a region of the genome, encompassing many sites, that is **transcribed** into **messenger RNA**, which is then in turn **translated** into a **protein** (one of the basic structural and functional components of the body). I will try not to use the word in this sense, but of course you should still understand the fundamentals of gene expression.

- An **allele frequency** is the fraction of the population's genes at a given site that falls within the specific allelic class. We can think of an allele frequency as the mean of a random variable defined over a population of token chromosomes. If there is a correlation between two sites in this population, we say that the sites exhibit **linkage disequilibrium** (**LD**).

- The allelic states of the genes carried by an individual at a particular site constitute that individual's **genotype**. A **heterozygous** genotype consists of two different alleles, while a **homozgyous** genotype consists of one allele represented twice.

- Mendel's First Law states that a diploid parent passes on a randomly selected member from each of its paired genes to the offspring. Mendel's Second Law states that transmission of genes at different sites are uncorrelated.

- Suppose that we could experimentally convert the allelic type of a gene from $-$ to $+$ in a **zygote** immediately after the union of its constituent sperm and egg cells but before the onset of any developmental events. The **average effect of gene substitution** at this site is the average of the resulting phenotypic change.

- An individual's **breeding value** is the weighted sum of how many $+$ genes are carried across causal sites; the weights are the average effects. In other words, an individual obtains a high breeding value by inheriting many $+$ genes, particularly at sites with larger average effects.

- The randomness inherent in Mendel's laws means that offspring of the same parents can possess different genetic values (unless they are monozygotic twins). To put it more simply, siblings can have different numbers of $+$ genes.

- The **additive genetic variance** is the population variance of breeding values. The **narrow-sense heritability** of a **phenotype** (trait) is the ratio of the additive genetic variance to the total phenotypic variance.

- The sites affecting a given phenotype might not combine in a simple additive manner. In this case an individual's **total genetic value** deviates from his breeding value. The variance of such deviations is the **non-additive genetic variance**. The ratio of the total genetic variance to the total phenotypic variance is the **broad-sense heritability**.

- Correlations between relatives (family resemblances) are functions of the heritability and therefore can be used to estimate it. Heritability is best estimated using several different kinds of relatives, including adoptive relatives who are not biologically related and biological relatives who did not grow up together.
- Monozygotic twins separated at birth offer a reasonable way to estimate broad-sense heritability. If other sources of family resemblance are minimal, then the midparent-offspring regression offers a reasonable way to estimate narrow-sense heritability.
- If the genetic and environmental causes of a trait are uncorrelated, then the average effects of gene substitution are equivalent to the partial regression coefficients in the multiple regression of the phenotype on all sites in the genome. Regression is more or less what **genome-wide association studies** (GWAS) use in an effort to identify the precise sites where the average effects are nonzero. So far GWAS have accounted for only a small fraction of phenotypic variance. This is because typical average effects are turning out to be very small, and even larger sample sizes are required to detect them confidently.
- GWAS data have enabled novel methods for estimating the **genetic correlation** between two traits, which is simply the correlation between their respective breeding values.
- A **family-based design** exploits Mendel's First Law to bolster the quality of causal inference in DNA studies. Such studies suggest that GWAS results obtained from unrelated individuals are not unduly contaminated by confounding.
- All genetic variation originates by **mutation**. A current allele frequency is not an unchanging constant of Nature; the frequency of the counted allele was once $1/(2N)$, and it increased to its current level by **genetic drift** or **natural selection**.
- The change in the mean phenotypic value caused by natural selection is proportional to the heritability.

References

Bulik-Sullivan, B. K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., … Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*, 1236–1241. doi:10.1038/ng.3406

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

Fisher, R. A. (1999). *The genetical theory of natural selection: A complete variorum edition*. Oxford, UK: Oxford University Press.

Gillespie, J. (2004). *Population genetics: A concise guide* (2nd ed.). Baltimore, MD: John Hopkins University Press.

Gimelfarb, A. (1981). A general linear model for the genotypic covariance between relatives under assortative mating. *Journal of Mathematical Biology*, *13*, 209–226.

Lee, J. J. (2012). Correlation and causation in the study of personality (with discussion). *European Journal of Personality*, *26*, 372–412. doi:10.1002/per.1863

Lee, J. J. & Chow, C. C. (2013). The causal meaning of Fisher's average effect. *Genetics Research*, *95*, 89–109. doi:10.1017/S0016672313000074

Lee, J. J. & Chow, C. C. (2014). Conditions for the validity of SNP-based heritability estimation. *Human Genetics*, *133*, 1011–1022. doi:10.1007/s00439-014-1441-5

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, *88*, 294–305. doi:10.1016/j.ajhg.2011.02.002

Lykken, D. T., McGue, M., Tellegen, A., & Bouchard, T. J., Jr. (1992). Emergenesis: Genetic traits that may not run in families. *American Psychologist*, *47*, 1565–1577. doi:10.1037/0003-066X.47.12.1565

Lynch, M. & Walsh, B. (1998). *Genetics and the analysis of quantitative traits.* Sunderland, MA: Sinauer.

Nagylaki, T. (1982). Assortative mating for a quantitative character. *Journal of Mathematical Biology*, *16*, 57–74. doi:10.1007/BF00275161

Pierce, B. A. (2010). *Genetics: A conceptual approach* (4th ed.). New York: Freeman.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., … Koellinger, P. D. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences USA*, *111*, 13790–13794. doi:10.1073/pnas.1404623111

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., … Koellinger, P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, *340*, 1467–1471. doi:10.1126/science.1235488

Scarr, S. (1997). Behavior-genetic and socialization theories of intelligence: Truce and reconciliation. In R. J. Sternberg & E. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 3–41). Cambridge, UK: Cambridge University Press.

Scarr, S. & Weinberg, R. A. (1978). The influence of "family background" on intellectual attainment. *American Sociological Review*, *43*, 674–692.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427. doi:10.1038/nature13595

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losick, R. (2014). *Molecular biology of the gene* (7th ed.). Cold Spring Harbor, NY: Pearson/Cold Spring Harbor Laboratory Press.

Wilson, S. R. (1973). The correlation between relatives under the multifactorial model with assortative mating. *Annals of Human Genetics*, *37*, 189–204. doi:10.1111/j.1469-1809.1973.tb01826.x

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*, 1173–1186. doi:10.1038/ng.3097

Wright, S. (1968). *Evolution and the genetics of populations vol. 1: Genetics and biometric foundations*. Chicago: University of Chicago Press.

Wright, S. (1969). *Evolution and the genetics of populations vol. 2: The theory of gene frequencies*. Chicago: University of Chicago Press.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565–569. doi:10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 76–82. doi:10.1016/j.ajhg.2010.11.011