Chapter 1: The Genome and Phenotypes

Understanding processes that generate variation between animals and how this variation is inherited between generations is essential for understanding the importance and challenges of estimating genetic values. This chapter describes the standard quantitative genetic model where phenotype values are generated from genetic, environmental, and other sources of variation. To this end, DNA molecules, their organisation in genomes, and their variation are described. This is followed by an encoding of DNA variation for quantitative genetic analyses and a functional relationship with phenotype values. While this standing genetic and phenotypic variation is often substantial, the inheritance of DNA between generations shuffles genetic variation in parents and generates new combinations through recombination, segregation and mutation.

Introduction

This book is about methods for analysing variation between animals to estimate their genetic value. For this estimation, we use statistical models, as we will show in the following chapters. Before we delve into these statistical models, it is instructive to overview the biological processes that generate the data we are analysing. Here, we will also describe models, but these are data generation models upon which the theory of quantitative and statistical genetics is built (Falconer and MacKay, 1996; Lynch and Walsh, 1998). Although these data generation models are often similar to the statistical models we use in our data analysis, it is important to note the following three interrelated points. Firstly, all models are an abstraction of complex biology, the true model, that generates variation between animals. Secondly, simple models can often adequately describe complex phenomena with a small number of parameters. Thirdly, although a very good attempt is usually made to match data generation and statistical models, we cannot fully unravel complex biology. This is so because we typically have only a limited amount of data, or the data resolution is too coarse to decipher this complexity. Continued advancements in data recording technologies will allow us to decipher more and more biology in forthcoming years.

The remainder of this chapter is organised into the following five sections. First, we continue this introduction and conceptualise variation between animals, including the definition of traits and underlying genetic, environmental, and other sources. Secondly, we describe the molecule that encodes genetic information, the DNA, and its organisation in the genome. We discuss the major DNA variation sources and how we encode this variation for quantitative genetic analysis. Thirdly, we delve into a model that generates variation between animals from genetic and environmental effects. Fourthly, we describe the inheritance of DNA from parents to offspring and how this process generates variation in a new generation. Understanding the processes that generate variation between animals and how this variation is inherited between generations is essential for understanding what we estimate with pedigree-based and genome-based statistical models described in the following chapters. Fifthly, we point to the different types of traits, multiple traits, genotype-by-environment interactions, and additive and non-additive genetic effects.

It is well known that most, if not all, traits vary between animals and that this variation is due to many effects. But what is a trait? Any characteristic you can see or measure on animals may be called a trait. For example, weight, height, colour, and so on. All the traits

we observe are called phenotypes of that animal. Derived from the Greek *pheno*, meaning "to show", and *type*, meaning, well, "type".

There are many ways to organise traits into various groups. For example, milk yield, number of laid eggs, and body weight are often called production traits. The number of days between two calvings is an example of a reproduction trait, and so on. In this book, we will be most interested in grouping traits by their phenotypic expression or how we record this expression. Without being exhaustive, let us look at continuous, ordinal, and binary traits. For example, milk yield is a continuous trait we usually record in kilograms, such as 7812.4kg per cow's lactation. Looking at the distribution of such continuous traits (Figure 1a), we will generally see a spread of recorded values around the central value with a decaying frequency towards the tails of this distribution. Another group are ordinal traits. Ordinal traits represent traits whose expression we count and hence have distinct categories. The frequency of animals recorded in each category can vary significantly between different systems. For example, the distribution of the number of progenies in a litter in pigs (Figure 1b) or in sheep (Figure 1c) are both ordinal. Binary traits are an extreme example, with only two categories, such as healthy or diseased (Figure 1d). Ordinal and binary traits are often called discrete or categorical traits to distinguish them from continuous traits. Also, sometimes a trait has a continuous phenotype expression, but we record it as a categorical trait. This book focuses on continuous traits because most traits have such distributions. It is also generally recommended that recording is continuous to capture full trait variation. When this is not the case, we can use methods described in chapter 15.

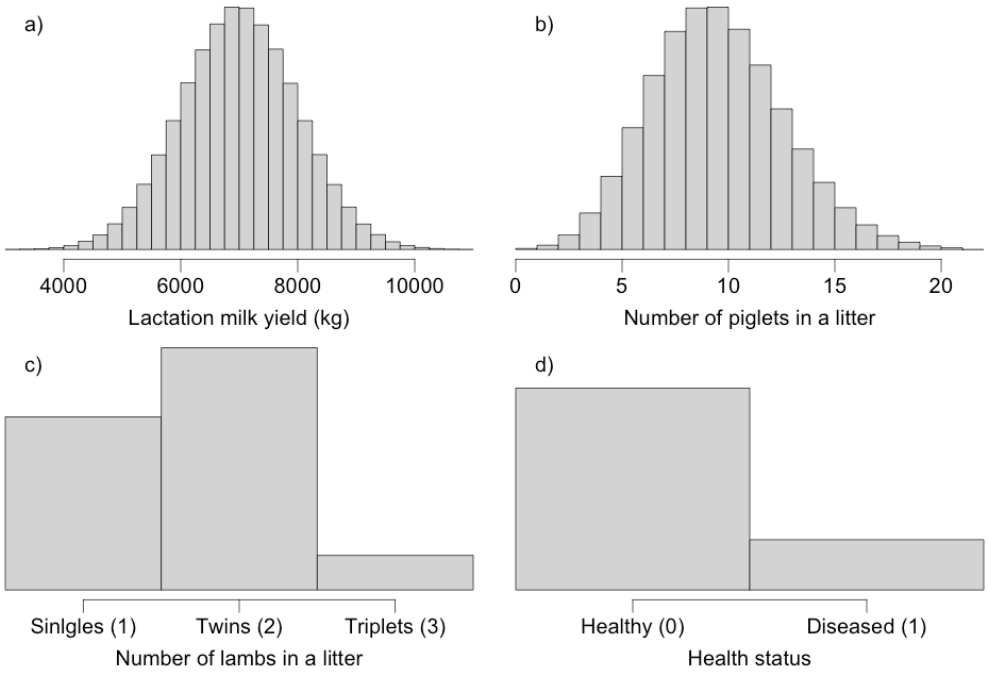

Figure 1: Examples of distributions for a) cow milk yield per lactation (continuous trait with mean of 7,000 and standard deviation of 1,000), b) litter size in pigs (ordinal trait with a mean of 10 and variance of 10), c) litter size in sheep (ordinal trait with mean of 1.7 and standard deviation of 0.6, given the small number of categories we can also report 38% singles, 54% twins, and 8% triplets), and d) health status (binary trait encoded as 0 for healthy and 1 for diseased, with a mean of 0.2, standard deviation of 0.4, and 20% diseased animals) (CC-BY 4.0)

What drives this phenotype variation between animals? You might have heard of the concept "*nature versus nurture"*. An animal's phenotype is a result of a combination of genes inherited from parents, known as the genotype, the environment they live in, and other factors:

$$Phenotype = f(Genotype, Environment, Other\ factors).$$

We will look at the genotype and its effect on phenotype in the next two sections. The environment includes the amount and quality of feed consumed, temperature, humidity, etc. Other factors include sex, the type of recording device, the data recording technician, the farmer's knowledge of animal husbandry, etc. We have loosely mentioned phenotype, genotype, environment, and other factors. More concretely, an animal's recorded phenotype *value* is a function of the *effect* of the animal's genotype, the *effect* of the environment where the animal lives, and the *effect* of other factors. We emphasise the concepts of *values* and *effects* because they enable us to quantify the contribution of different sources to phenotype variation. If we knew the effect of genotype, environment, and other factors and their functional relationships, we would fully understand sources of variation in phenotype values. We never know these effects and their functional relationships. We use collected data and statistical models to estimate these effects and their functional relationships. While the collected data that we feed into the statistical models will vary substantially between animal systems, it will generally include phenotype values, associated descriptors (such as animal identification, sex, farm identification, etc.), pedigree, and, increasingly, genomic data. The following chapters will show examples of such datasets.

Variation in DNA

Variation in the composition of an animal's genetic material, their genotype, is determined by the DNA inherited from their parents. This DNA instruct biological functions, such as the growth and reproduction of an animal, in all the trillions of cells (~10^{12+}). Inside each cell is a nucleus with a complete copy of the inherited DNA. DNA is a long molecule that looks like a twisted ladder. The rungs of this ladder are smaller molecules called nucleotides or bases. There are four bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) (Figure 2). These bases bind in pairs forming the twisted ladder, the double helix. Adenine (A) binds with Thymine (T), while Cytosine (C) binds with Guanine (G) (Figure 2).

The complete collection of DNA molecules in a cell is called a genome. The structure and size of the genome vary between species. For example, in cattle, the genome is organised into 30 chromosomes. Cattle are diploid, meaning it has two copies of each chromosome, in total 60 DNA molecules. Each of the copies is inherited from one parent. We call each chromosome copy a haplotype, and the combination of two chromosome copies a genotype. The total length of the cattle genome is about 3 billion base pairs (~$3x10^9$). This is the length of one copy of 30 chromosomes. Hence, each cell in cattle has about 6 billion base pairs. Some genomes are much smaller. For example, the honeybee genome has only about 250 million base pairs organised in 16 chromosomes.

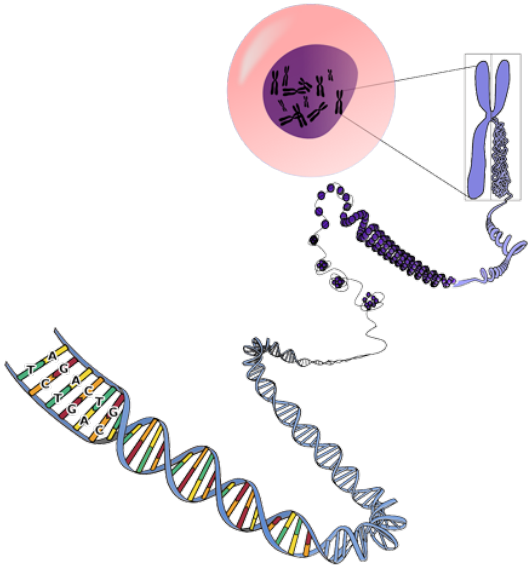

Figure 2: Diagram zooming in from the cell's nucleus to an animal chromosome and to the unwinding of DNA double helical molecule with its bases © OpenClipart-Vectors (2013) CC0

Most of the genome is the same across all chromosome copies in a population. We are, however, most interested in the parts of the genome that differ between chromosome copies. This variation can be present both within one animal and between animals. These variable parts of the genome are called segregating/polymorphic sites or loci. If a locus is segregating, it means that there is variation in DNA at that position within and between families. In other words, DNA variation (polymorphism) exists at that position. For example, some chromosome copies in a population have the A-T base pair at that locus (say, a single base pair A-T), while other chromosome copies have the G-C base pair. These loci show variation because, at some point in the past, one of the chromosomes has been copied with an error, with a mutation. If a mutation occurs in germline (reproductive) cells, it can be passed from parents to their progeny.

We refer to the different base pair sequences at a locus as alleles. A mutation is usually called a derived allele, while the original allele is called the ancestral allele. In the context of reference genomes, the genome that other genomes are compared to, we often use the term reference allele, which denotes the allele present in the reference genome. Alleles that differ from the reference allele are usually called alternative alleles. Variation at the single base pair mentioned above is called a Single Nucleotide Polymorphism (SNP). There are additional types of DNA polymorphism, such as deletions, insertions, repetitions, and inversions, at a small-scale involving few base pairs or at a large-scale involving chromosome regions or whole chromosomes. Because the DNA molecule has a direction (that is, it is read from the 3' end towards the 5' end), we can observe four possible SNP alleles: A-T, T-A, G-C, and C-G (Figure 3). Of the billions of DNA base pairs, most studies have found tens to hundreds of millions of SNPs (~10-100 x 10^6 = ~10^7-10^8) and other types of DNA polymorphisms (Hayes and Daetwyler, 2019; Halldorsson et al., 2022; Ros-Freixedes et al., 2022). This suggests that every 100th to 10th base pair in a genome could show polymorphism in a population. Many of these loci will have very low frequencies of mutated alleles.

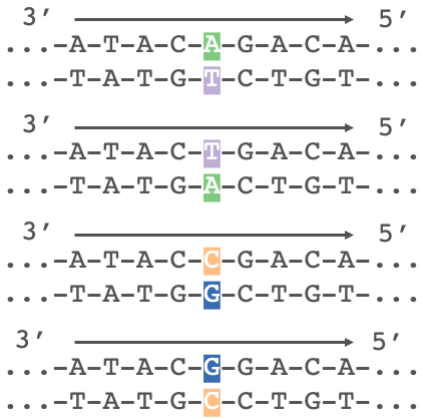

Figure 3: Four possible SNP alleles at one base pair in four DNA fragments taking DNA read direction into account (CC-BY 4.0)

The two most important technologies for generating genomic data are sequencing and SNP arrays. Sequencing simply means reading the DNA. There are two phases in using sequencing. Initially, we must de-novo sequence the genome of one animal and build the so-called reference genome. Then, further animal genomes are re-sequenced against the reference genome. Most modern sequencing techniques involve high-molecular-weight DNA isolation, cutting the genome into smaller fragments, repeatedly sequencing these fragments, aligning the sequence reads to the reference genome, and finally, calling the alleles and genotypes of an animal. The accuracy of the resulting data depends on the quality of all the steps. For example, good quality DNA isolation is critical, as is repeated sequencing of DNA fragments to capture variation at both chromosome copies and to distinguish sequencing errors from real DNA variation. The advantage of sequencing is that it captures most of the genomic variation, all the millions of SNP loci and some structural variants. Note that some reference genomes do not contain all variation within a species, so re-sequencing against such a reference misses that variation. SNP arrays (also called SNP chips) are conceptually doing the same as sequencing but are using previously designed array probes to capture variation at a selected set of SNP markers. Most “standard” SNP arrays have a density of ~60 thousand (60K) markers. In cattle, this gives a marker every ~50 thousand (50K) base pairs, of which about 500 to 1000 are expected to be polymorphic yet represented by a single marker. The selection of SNP markers aims for a uniform spread along chromosomes and allele frequency spectrum, as well as reliable genotype calling across batches of animals. The cost of SNP array genotyping is generally lower than sequencing and has a lower DNA isolation quality requirement, but it captures less DNA variation.

We often focus on biallelic SNPs, those with two alleles. The reason for this focus is that transition mutations between Adenine (A) and Guanine (G) and between Cytosine (C) and Thymine (T) are much more common than transversion mutations between Adenine (A) and Cytosine (C) or Thymine (T) and between Guanine (G) and Cytosine (C) or Thymine (T). This is driven by the molecular structure of the bases. Hence, if we have the A-T base pair as the ancestral allele, the frequency of the alternative C-G and T-A base pairs will be lower than that of the G-C base pair (considering the DNA orientation). Ultimately, the frequency of each mutation will depend on their spread between generations. We often focus on biallelic

SNPs to avoid a mix-up between potential data recording errors and rare mutations. The accumulation of vast genomic data in recent years will likely broaden this focus. When calculating with biallelic SNPs, we numerically encode the two alleles in a computer with numbers: 0 represents the reference (ancestral) allele, and 1 represents the mutated (derived or alternative) allele. In diploid species, we can observe three possible genotypes at a biallelic SNP: homozygote for allele 0 (genotype 0/0), heterozygote for allele 0 and 1 (genotypes 0/1 or 1/0), and homozygote for allele 1 (genotype 1/1) (Figure 4). Following the numerical encoding, the homozygote 0/0 is encoded as 0+0=0, heterozygotes 0/1 or 1/0 are encoded as 0+1=1+0=1, and homozygote 1/1 is encoded as 1+1=2 (Figure 4). The numerical codes 0, 1, or 2 for the three genotypes mean that an animal has respectively 0, 1, or 2 alternative alleles. These codes are usually called allele dosage.

Figure 4: Three possible genotype combinations at a biallelic SNP with the corresponding allele dosage encoding of the alleles and genotypes (CC-BY 4.0)

We can now write a sequence of biallelic SNPs along a chromosome as a series of zeroes (0) and ones (1). We will write such sequences of SNP alleles from one chromosome or chromosome region (haplotypes) in rows. Figure 5 shows two haplotypes of an animal across six SNPs and the corresponding genotype as a sum of the two haplotypes. The top haplotype has 3 alternative alleles in total. The bottom haplotype has 4 alternative alleles in total. Hence, the genotype has 7 alternative alleles in total.

Haplotype 1	0	1	1	0	0	1
Haplotype 2	1	1	1	1	0	0
Genotype	1	2	2	1	0	1

Figure 5: Example of allele dosage encoding for two haplotypes across six SNPs of an animal and the corresponding genotype (CC-BY 4.0)

There are many ways to summarise DNA variation across animals and across loci. The simplest way is to calculate the frequency of alleles at a locus. Assume we have a matrix of genotype allele dosages for biallelic SNPs where animals are represented in rows and loci in columns. Then locus allele frequencies are calculated as column means divided by 2 – each column will give allele frequency p_l for the corresponding locus l. Next, we can calculate the frequency of genotypes at a locus by tabulating the frequency of three allele dosages: 0, 1, and 2. There are many other ways to summarise variation in DNA. Statistics used in chapters 11 and 12 are the expected allele dosage at a locus, the variance of allele dosages at a locus, and the correlation between allele dosages at two loci. We can estimate the expected allele dosage at a locus by multiplying allele frequency at a locus by 2: $2*p_l$. If allele frequency p_l is 0.2, we would expect an average genotype allele dosage of 2*0.2=0.4 (see SNP2 in Table 1). This means that the frequency of genotypes 0, 1, and 2 will be such that their average will be 0.4. We can estimate the variance of allele dosages by calculating the variance of observed allele dosages in our dataset. It is common to compare the observed and expected genotype variation according to the Hardy-Weinberg equilibrium. This equilibrium is achieved primarily by random mating of parents and avoiding selection between their progeny. Under such conditions, we expect that the frequency of alleles in parents and progeny will be the same, and the frequency of genotypes 2, 1, and 0 will be respectively p_l, $2*p_l*q_l$, and q_l, where p_l is the frequency of the alternative allele 1 and q_l=1-p_l is the frequency of the reference allele 1. Following the binomial sampling of alleles under such conditions (this is a mathematical way of describing random mating), the variance of allele dosages is $2*p_l*q_l$. This quantity is often referred to as heterozygosity, the proportion of heterozygotes, as well as genic variance (expected variance of allele dosages at one locus, that is, one gene, hence the term genic) under Hardy-Weinberg equilibrium. Finally, we can calculate the correlation between allele dosages at two loci to study covariation between different genome regions. This quantity is referred to as linkage-disequilibrium because a non-zero correlation suggests that alleles appear together more often than expected by chance. This can happen when loci are physically linked (placed on the same chromosome) or influenced by selection, population stratification, or admixture processes.

Table 1: Genotype allele dosages at two loci in five animals and corresponding summary

Animal	SNP1	SNP2
1	0	0
2	2	1
3	2	0
4	1	1
5	0	0
Mean	1.00	0.40
Standard deviation	1.00	0.55
Variance	1.00	0.30
Allele frequency	0.50	0.20
Genic variance	0.50	0.32
Correlation	0.46	

Variation in phenotype values

How is the DNA variation related to variation in phenotype values between animals? We generally do not know which DNA loci affect traits. We know that DNA gene regions are translated into RNA molecules, which are further transcribed into proteins that perform biological functions. Polymorphic loci in or around the genes drive genetic differences between animals' phenotype values. We call such loci causal loci or Quantitative Trait Loci (QTL). Some traits are affected only by the genotype of an animal. We call such traits as Mendelian traits. Mendelian traits are commonly affected by only a few DNA loci. When a trait is affected by one gene (or one locus), we call it a monogenic trait. When a trait is affected by several genes (or several loci), we call it an oligogenic trait. Most traits are affected by many DNA loci. We call such traits as polygenic traits. Some traits, especially polygenic traits, are also affected by the environment in which animals live. We call such traits as complex traits. This complex situation is the basis for the "*nature versus nurture"* concept and recognition that an animal's observed phenotype value is a function of the animal's genotype, the *effect* of the environment where the animal lives, and the *effect* of other factors.

How many DNA loci affect polygenic traits? We do not know. But we can make an educated estimate. There are about 20,000 genes in the genomes of many species. Let's assume that each gene affects a polygenic trait and has at least one SNP. In this case, the number of causal SNP loci will be about 20,000. This might seem like a large number. However, note that traits related to biological processes such as growth or lactation are incredibly complex and involve many, if not most, body functions and hence many proteins and their upstream genes in one way or another.

While we do not know the form of this phenotype generation function, nor its effects, we will use Fisher's quantitative genetics framework to reason about the effects and later estimate them (Fisher, 1919). Fisher assumed that the observed phenotype value of an individual (y_i) can be partitioned into the effect of various factors that capture deviation of the phenotype value from the baseline of a population (μ), most importantly, the genetic value of the individual (g_i) and environmental effect (e_i), plus possibly interaction between the genotype and environment ($g_i \times e_i$):

$$y_i = \mu + g_i + e_i + g_i \times e_i.$$

An important simplification here is that this phenotype generation function is assumed to be linear, where we add up the effects of different factors. Here we ignore the highly non-linear and interconnecting biochemical pathways, metabolic processes, etc. All this biological complexity is swept under the "model carpet". We simply associate changes in phenotype values with changes in the genetic composition of individuals, which we quantify with genetic values, while the remainder is assumed to be due to environmental effects. Such linear models can be seen as a first-order (local) approximation of the highly complex biological system. This is the simplest possible, yet informative, approximation. Following the same linear framework, we can further assume that the total genetic value of an individual (g_i) is a sum of the genetic values of that individual across causal loci ($g_{i,1}$, $g_{i,2}$, …,

$g_{i,l}$, …, $g_{i,k}$, where *k* is the number of causal loci and $g_{i,l}$ takes as many values as there are genotypes observed at the locus *l*) and possibly their epistatic interactions ($g_{i,1} \times g_{i,2}$ + …):

$$g_i = g_{i,1} + g_{i,2} + \cdots + g_{i,k} + g_{i,1} \times g_{i,2} + \cdots.$$

Fisher also assumed possible interactions between alleles within a locus, which further decomposes genetic values into the additive genetic value ($a_{i,l}$) and the dominance genetic value ($d_{i,l}$) of an individual at each causal locus as well as across causal loci. We leave the topic of additive and non-additive (dominance and epistasis) genetic effects for Chapter 13. From this point onwards, we will assume additive allele effects only; hence, genetic values g_i will be additive genetic values, often called breeding values.

The following example demonstrates the decomposition of genetic value across loci. Assume that the baseline value of a population is 10 units and that there is a single causal locus *l* with an additive effect, a_l. The effect is such that substituting the reference allele 0 with the alternative allele 1 increases the phenotype value for 1 unit. Hence, substituting two reference alleles will increase the phenotype value for 2 units. With the three possible genotypes at a biallelic SNP (encoded as $x_{i,l}$ = 0, 1, and 2), we respectively expect the following three phenotype values:

$E(y_i|x_{i,l} = 0, a_l = 1) = \mu + x_{i,l}a_l = 10 + 0 \times 1 = 10,$
$E(y_i|x_{i,l} = 1, a_l = 1) = \mu + x_{i,l}a_l = 10 + 1 \times 1 = 11,$ and
$E(y_i|x_{i,l} = 2, a_l = 1) = \mu + x_{i,l}a_l = 10 + 2 \times 1 = 12.$

Observed phenotype values will deviate from these expectations due to environmental effects. Assuming that environmental effects come from a normal distribution with mean zero and standard deviation (σ_e) of 0.5 unit, we can expect variation in phenotype values as shown in Figure 6.

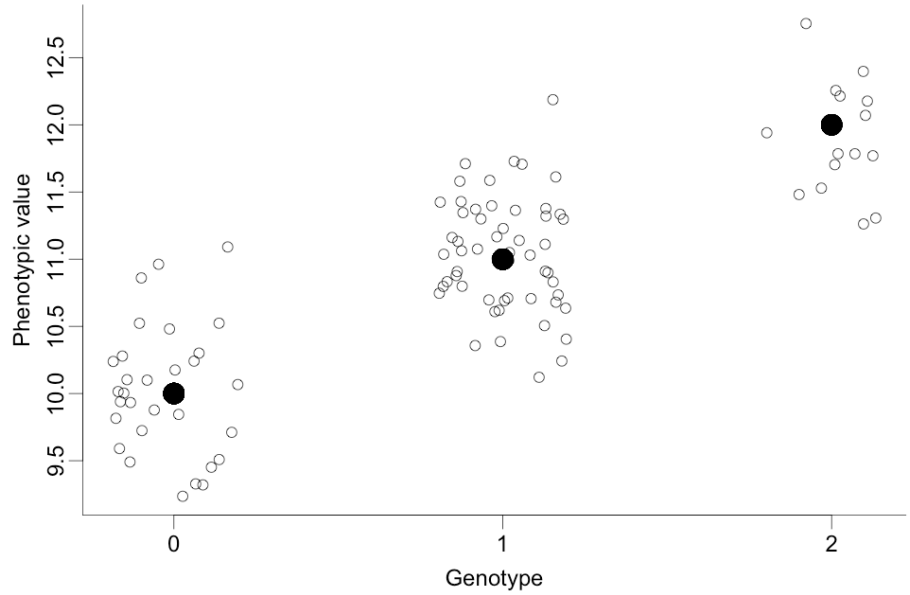

Figure 6: Example of expected (large full circle) and observed (small empty circle) phenotype values as a function of three genotypes at a causal biallelic SNP locus (jittered to improve

the display of points) with an allele substitution effect of 1 unit and normal environmental effects with a standard deviation of 0.5 units

Figure 6 shows two sources of variation in phenotype values – genetic differences and environmental differences between animals. For every animal, we can write the following phenotype generation model:

$$y_i = \mu + g_i + e_i.$$

Assuming that environmental effects are normally distributed and that we know the genotype values (under the data generation model), we can write the model in a probabilistic form as:

$$y_i|g_i \sim N(\mu + g_i, \sigma_e^2),$$

where we see that the expectation of this data generation process is $E(y_i|g_i) = \mu + g_i$ and the variance of this process is $Var(y_i|g_i) = \sigma_e^2$, the environmental variance. Note that if we do not know the genotypes and their effect, that is, we are looking at phenotype variation across all the genotypes together, the probabilistic form changes to:

$$y \sim N\left(\mu, \sigma_g^2 + \sigma_e^2\right),$$

where σ_g^2 is genetic variance, that is, the variance between genetic values of individuals $Var(g_i) = \sigma_g^2$; and $\sigma_g^2 + \sigma_e^2 = \sigma_y^2$ is the phenotypic variance, that is, the variance between phenotypic values of individuals $Var(y_i) = \sigma_y^2$, which is driven by genetic and environmental variation.

Until now, we have omitted a description of allele substitution effects' size, sign, and distribution. While there is a growing body of literature on this topic, the field is still grappling with the challenge of identifying causal loci among all the loci. Namely, there are tens to hundreds of millions of SNPs and additional types of polymorphisms. We expect that only a fraction of these loci is causal (perhaps on the order of hundreds, thousands, or tens of thousands). Whatever the distribution of allele substitution effects, once we add up these effects across loci, the resulting distribution of whole-genome genetic values will tend towards a normal distribution due to the central limit theorem. We demonstrate this by showing the distribution of a sample of the population baseline plus genetic values (expected phenotype value) in Figure 7 for the trait affected by one, two, three, or ten biallelic SNPs – all having the allele frequency of 0.5 and being on different chromosomes. We assumed that the baseline value of the population is 10 units and that each alternative allele has an effect of *1/k* units, where *k* is the number of causal SNPs (this ensures that the scale of genetic values is comparable between the four examples, but note that this also scales genetic variance).

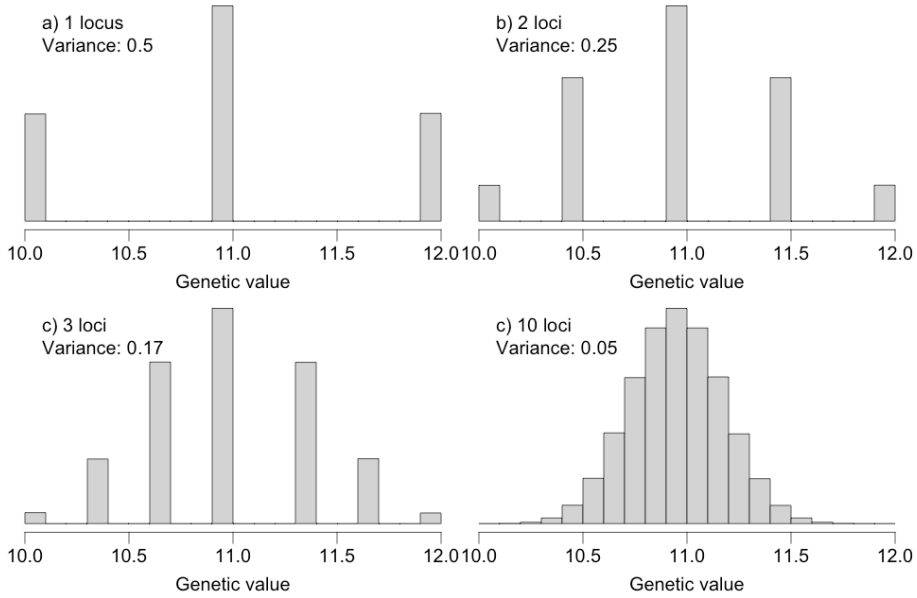

Figure 7: Distribution of a sample of a population baseline plus genetic values for a trait that is affected by (a) one, (b) two, (c) three, or (d) ten biallelic SNP – each sub-plot reports the corresponding variance of genetic values between individuals, the genetic variance $\left(\sigma_g^2\right)$ (CC BY 4.0)

As seen in Figure 7, the number of distinct genetic values is growing quickly with the number of causal loci, and the distribution is rapidly approaching a continuous normal-like distribution. This is not surprising since the total possible number of genotype combinations across *k* biallelic SNP is 3^k: $3^1 = 3$ for one SNP, $3^2 = 9$ for two SNPs, $3^3 = 27$ for three SNPs, and $3^{10} = 59{,}049$ for ten SNPs. With five hundred SNPs, the number of genotype combinations grows to a whooping ~10^{238}, which is more than the number of atoms in the universe (~10^{80}). This is one of the reasons early quantitative genetics work used the term *infinitesimal*, as in the infinitesimal model, to indicate that the contribution of one locus to total genetic variance is infinitely small.

To demonstrate how we generate genetic and phenotypic values for traits that are affected by multiple causal SNPs, let's take the example from Figure 5, assuming a trait has a population baseline of 10 units and is affected by the six biallelic SNPs with additive allele effects. At these SNPs, substituting the reference allele 0 with the alternative allele 1 changes the genetic value for +1 unit at the first SNP, +2 units at the second SNP, -1 unit at the third SNP, +1 unit at the fourth SNP, +1 unit at the fifth SNP, and -2 units at the sixth SNP. In Figure 8 we show how we generate genetic value from an animal's haplotype and genotype allele dosages. At the top are allele dosages for the animal's two haplotypes and the corresponding genotype. At the bottom are values of alleles and corresponding genotypes alongside the six SNPs and their sums on the right. We obtain these allele and genotype values by multiplying the allele dosages with the effects and then summing the values along the SNPs. This animal has one haplotype with value -1 unit, another haplotype with value +3 units, which gives the genetic value of +2 units. If we now add the population baseline of 10 units and assume that the animal experienced a positive environment with an effect of +2 units and that there was no genotype-by-environment effect or non-additive genetic effects, then the phenotype value of this animal would be 10+2+2=14 units.

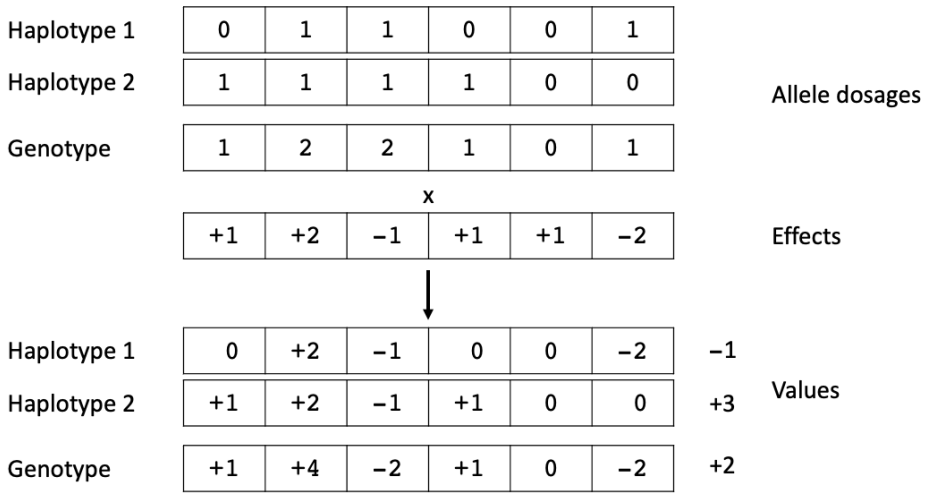

Figure 8: Example of generating genetic value of one individual where the trait is affected by six biallelic SNPs (CC BY 4.0)

For the environmental effects, it is also reasonable to assume that many sources affect complex traits. We don't know all these sources and their effects, but if there are many, the distribution of their total effect will also tend towards a normal distribution due to the central limit theorem. The variance of these total environmental effects is the environmental variance (σ_e^2). Figure 9 repeats the distribution of a sample of genetic values from Figure 7 for the trait affected by ten SNPs with the addition of environmental effects so that the heritability of phenotype values is 0.3, that is, $h^2 = \sigma_g^2/\sigma_y^2 = 0.3$, where $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$ is the phenotypic variance.

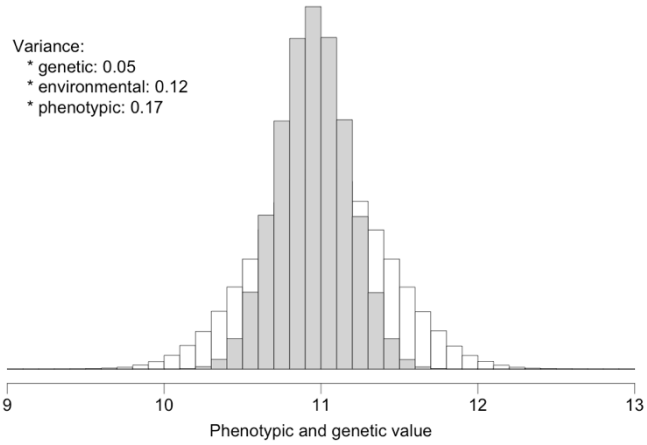

Figure 9: Distribution of a sample of genetic values (dark bars) and phenotypic values (light bars) for the trait that is affected by ten biallelic SNP and environmental effects such that heritability is 0.3 – the plot reports corresponding genetic variance (σ_g^2), environmental variance (σ_e^2), and phenotypic variance (σ_y^2) (CC BY 4.0)

DNA lottery

We will now look at the randomness of DNA inheritance between parents and progeny (the DNA lottery), and how this process drives variation and resemblance between genomes of relatives and their genetic and phenotypic values. This variation comes from mitosis, which

involves mutations, and from meiosis, which involves mutation, recombination, and segregation.

Mutations may be the source of variation most people are familiar with. Mutations are occasional mistakes made by the DNA replication machinery each time a cell divides. If these mistakes are made in the germline (reproductive) cells, these mutations can be inherited. With billions of DNA bases, it is impressive that mutations only happen at about one mutation per chromosome in the germline (Goriely, 2016). A newborn animal will have about *n* de-novo mutations from each parent, where *n* is the number of chromosomes. In cattle, this could mean about 60 de-novo mutations. These de-novo mutations are in addition to mutations that parents have inherited from their parents (and so on from older ancestors) and have transmitted to their progeny. See the next paragraph on how recombination and segregation affect this transmission. With most chromosomes having about 10^8 base pairs, this number of mutations per chromosome means that the rate of mutations is about $1x10^{-8}$ per DNA base pair per generation in the germline of many animals. The somatic mutation rate seems at least one order of magnitude higher (~$1x10^{-7}$) than the germline mutation rate (Lynch, 2016). With ~100 inherited germline mutations and many somatic cells (~10^{12+}), every animal is expected to carry ~10^{15+} mutations, with most of the genome mutated many times in many cells (Lynch, 2016). While somatic mutations are not inherited, they can affect phenotypes like germline mutations if they occur in key genome regions. Cancer is likely the most prominent example caused by somatic mutations. The effect of mutations can be either negative or positive, and this effect can depend on the environment.

While mutation is the source of new DNA variation, recombination and segregation create new combinations from existing DNA in parents and pass these combinations to offspring. Recombination and segregation happen during meiosis - the process through which germline cells produce gametes, such as sperm and ova. Each gamete contains half of the original set of DNA molecules. Which half a gamete receives is random and referred to as Mendelian sampling. Recombination and segregation can generate many combinations, which enable a continued response to selection from year to year. However, this large variation created in every new generation makes the estimation of the genetic values of newborn animals challenging. Figure 10 shows a diploid cell going through meiosis. For simplicity, we show only one chromosome pair in a cell, colour each chromosome instance differently, assume the chromosome is only six base pairs long, and omit the actual DNA base pairs. While there are several steps in meiosis, we only show four. In the first step, each chromosome copy is doubled into two chromatids. In the second step, crossovers and recombinations occur, where chromatids can exchange DNA. When DNA is exchanged between paternally and maternally derived chromatids, we have recombination. Which chromatids and which parts are exchanged are largely random events. In the third step, the cell divides into two diploid cells. In the fourth step, each diploid cell splits, and we get four haploid gametes. The gametes are now ready for the final act of DNA lottery, segregation. Namely, which of the generated gametes will give rise to an offspring is again down to random events. Note that at DNA replication and recombination steps, there is a chance for germline mutations to occur.

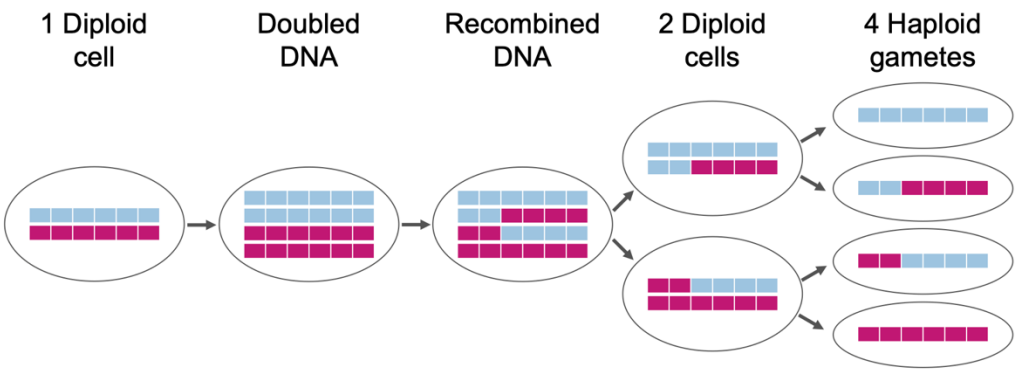

Figure 10: Meiosis process of one diploid cell with one chromosome pair producing four haploid gametes (CC BY 4.0)

To appreciate the power of combining existing DNA variation, let's first look at the number of chromosome combinations we can get in gametes from segregation only, without recombination. With one chromosome pair, we can get two chromosome combinations in gametes - [a light one = 1L] and [a dark one = 1D] (Figure 10). The acronym "1L" refers to the first chromosome and its lightly coloured copy. With two chromosome pairs, we can get four chromosome combinations in gametes - [1L, 2L], [1L, 2D], [1D, 2L], and [1D, 2D]. With three chromosome pairs, we can get eight chromosome combinations in gametes. With *n* chromosomes, we can get 2^n chromosome combinations in gametes. For example, cattle have 30 chromosome pairs, giving 2^{30} = 1,073,741,824 (more than a billion, ~10^9) possible chromosome combinations in gametes. This is the number of possible chromosome combinations in gametes in one parent, assuming no recombination.

As we saw in Figure 10, some chromosomes had no recombinations, and some had one. We usually get about one recombination per generated chromosome of ~10^8 bases, but we can get no recombination or more than one. Hence, the recombination rate is about $1x10^{-8}$ per base pair per generation, like the germline mutation rate. The number and placement of recombinations are random events. While the number of recombinations is not large, random placement adds many possible chromosome combinations in gametes on top of segregation.

The seemingly simple process of meiosis can generate a staggering amount of DNA variation by recombining and segregating the parental chromosomes. This process drives genetic relationships between animals. To put this into the context of relatives, we show a three-generation pedigree in Figure 11. This figure shows two siblings (G and H), their two parents (E and F), and their four grandparents (A, B, C, and D). As before, we show only one chromosome pair with six loci. We have four diploid grandparents, hence eight chromosome instances. To simplify tracking of genetic inheritance within this pedigree, we have coloured the chromosomes and numbered their loci according to grandparental origin. Also, we use a convention that the top chromosome is of paternal origin, and the bottom chromosome is of maternal origin. There is no such ordering in an actual cell. Do not confuse this "descent-based" encoding of alleles (and the related concept of identity-by-descent) with the "state-based" encoding (and the related concept of identity-by-state) that we have used up to now. Behind the numbers 1, 2, ..., 8 are DNA base pairs with the corresponding allele codes 0 and 1, respectively, for reference and alternative alleles.

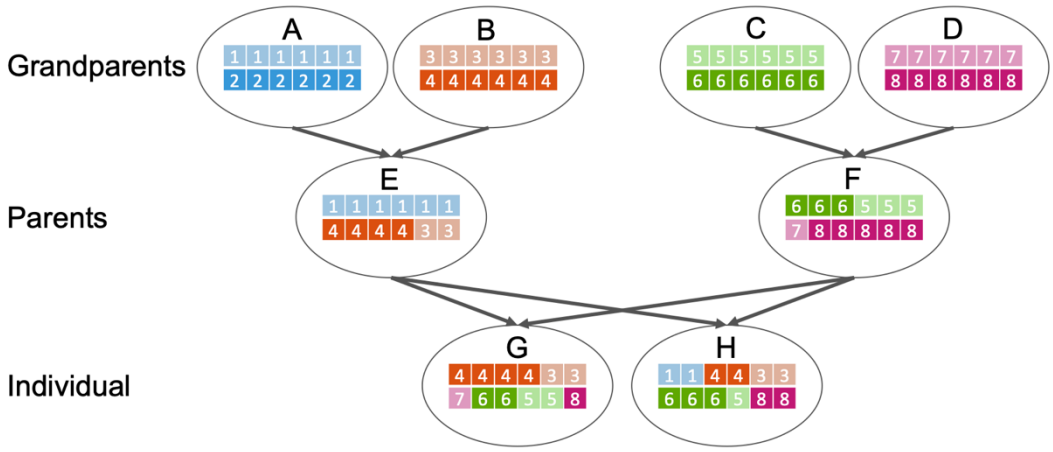

Figure 11: Inheritance of DNA between generations of a pedigree – alleles are represented by descent-based encoding (CC BY 4.0)

Inspecting Figure 11, we can see that an animal always receives 50% of DNA from each parent, but there is variation in which half is received due to recombination and segregation. Due to this sampling, an animal might not inherit DNA from one of his grandparents in a particular genome region. This means, that recombination and segregation are sampling different ancestral lineages along the genome for every newborn animal. For example, animal G inherited no DNA from grandparent A. Across multiple chromosome pairs, we expect that a grandchild inherits 25% of DNA from each grandparent. Still, there is variation around this expectation due to recombination and segregation. Similarly, we expect that siblings share 50% of DNA, but there is variation around this expectation due to recombination and segregation.

We can formalise the above observations by relating the genetic value of an individual g_i with the genetic value of its father (sire) $g_{f(i)}$ and its mother (dam) $g_{m(i)}$:

$$g_i = \frac{1}{2}g_{f(i)} + \frac{1}{2}g_{m(i)} + r_i,$$

where $\frac{1}{2}g_{f(i)} + \frac{1}{2}g_{m(i)}$ is the *parent average*, the expected genetic value of an individual given the genetic values of its parents $E\left(g_i|g_{f(i)}, g_{m(i)}\right) = \frac{1}{2}g_{f(i)} + \frac{1}{2}g_{m(i)}$, and r_i is the *Mendelian sampling deviation*, the deviation of individuals' genetic value from the parent average $r_i = g_i - \left(\frac{1}{2}g_{f(i)} + \frac{1}{2}g_{m(i)}\right)$. The above model is sometimes referred to as pedigree regression, where we regress the genetic value of an individual to the genetic values of its parents to get the expected value, while deviations from the regression lines are due to Mendelian sampling. To further connect this formalism with Figure 11, note that the genetic value of an individual is a sum of the genetic values of its two chromosome (genome) copies ($g_{i,1}$ and $g_{i,2}$). Each chromosome (genome) copy originates from a parent, which also has two chromosome (genome) copies and passes a combination of these to its progeny. Hence, we can split parent average and Mendelian sampling terms per parent as:

$$
\begin{aligned}
g_i &= g_{i,1} + g_{i,2}, \\
g_{i,1} &= \tfrac{1}{2} g_{f(i),1} + \tfrac{1}{2} g_{f(i),2} + r_{i,1}, \\
g_{i,2} &= \tfrac{1}{2} g_{m(i),1} + \tfrac{1}{2} g_{m(i),2} + r_{i,2}.
\end{aligned}
$$

Because the genetic values of individuals are a sum of their genetic values across the causal loci, we can show the connection between the genetic values of an individual and its parents along the causal loci as a sum of locus genetic values for each parental chromosome (genome) (giving the parent average) and a sum of locus deviations (giving the Mendelian sampling term):

$$
\begin{aligned}
g_i &= \sum_{l=1}^{k} \left(g_{i,1,l} + g_{i,2,l} \right), \\
g_{i,1} &= \sum_{l=1}^{k} \left(\tfrac{1}{2} g_{f(i),1,l} + \tfrac{1}{2} g_{f(i),2,l} + r_{i,1,l} \right), \\
g_{i,2} &= \sum_{l=1}^{k} \left(\tfrac{1}{2} g_{m(i),1,l} + \tfrac{1}{2} g_{m(i),2,l} + r_{i,2,l} \right)
\end{aligned}
$$

where the summation is across the causal loci *1, 2, …, k*, $g_{i,s,l}$ is the genetic value of individual *i* in genome set *s* at the causal locus *l*, $r_{i,s,l}$ is the corresponding Mendelian sampling deviation, and *k* is the number of causal loci. The above formulation shows how parent average and Mendelian sampling deviation of a genetic value result from DNA inheritance between generations.

Understanding the variation of genetic values between families (that is, between family parent averages) and within families (that is, between Mendelian sampling deviations within families) is important for various breeding operations. Figure 12 demonstrates such variation for two half-sib families originating from crossing parent A with B and parent C with B. Here we assume that such crosses can produce many progenies. If this is not possible, the figure shows the extent of possible variation between potential progenies. There are three notable observations from Figure 12. First, progeny genetic values are distributed around their parent average in line with the abovementioned theory. Second, there is substantial within-family variation due to Mendelian sampling. Third, some progeny genetic values are below or above parental genetic values, again indicating the extent of Mendelian sampling. Importantly, this has been generated from only six causal loci on one chromosome. Many more causal loci across multiple chromosomes will influence many traits, generating even more variation.

To further appreciate the magnitude of between and within-family variation, we can evaluate how much genetic variance is due to variation between and within families. We will address this topic more extensively in Chapter 3, but here we give the standard result by decomposing the genetic variance according to the pedigree regression. In the following, we assume that parents are randomly sampled from a population, not inbred, and unrelated. Under such conditions, genetic variation in a population is 50% due to between-family variation and 50% due to within-family variation. This result is important because it shows the extent of variation we can expect from combining parental genomes (=between-family

or parent average variation) and from recombining and segregating parental genomes (=within=family or Mendelian sampling variation).

$$
\begin{aligned}
Var(g_i) &= Var\left(\tfrac{1}{2}g_{f(i)} + \tfrac{1}{2}g_{m(i)} + r_i\right) \\
\sigma_g^2 &= Var\left(\tfrac{1}{2}g_{f(i)}\right) + Var\left(\tfrac{1}{2}g_{m(i)}\right) + Var(r_i) \\
&= \tfrac{1}{4}Var\left(g_{f(i)}\right) + \tfrac{1}{4}Var\left(g_{m(i)}\right) + Var(r_i) \\
&= \tfrac{1}{2}\sigma_g^2 + \tfrac{1}{2}\sigma_g^2.
\end{aligned}
$$

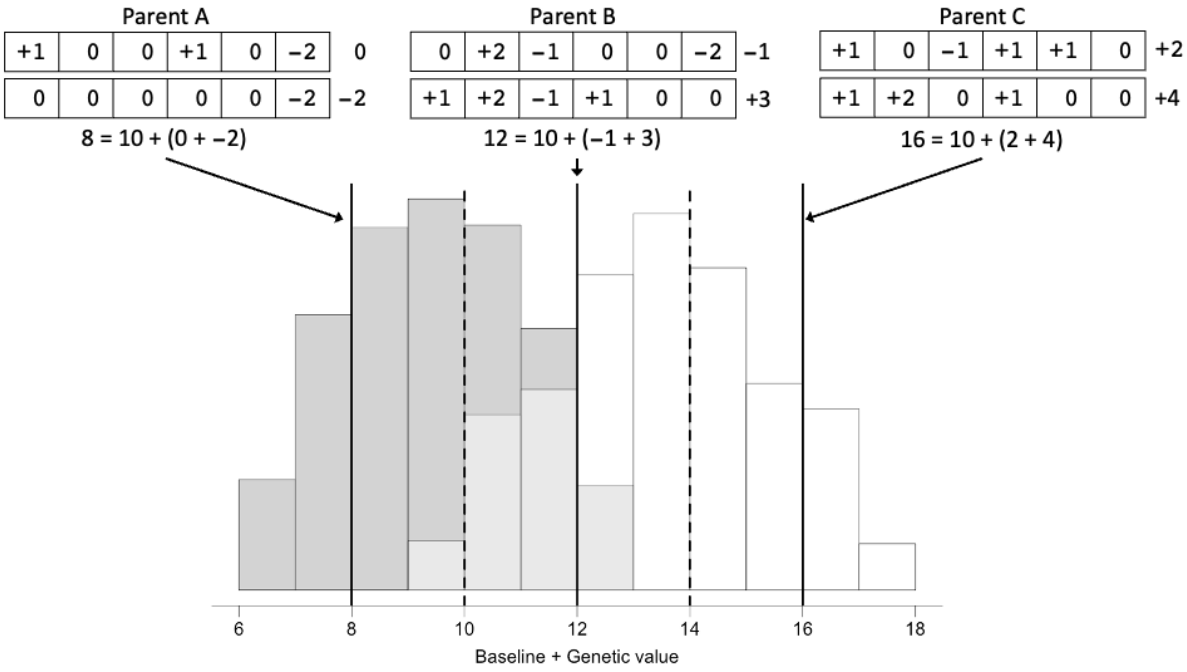

Figure 12: Example of genetic variation between and within two half-sib families due to variation in parental genetic values (parent averages) and Mendelian sampling – at the top are parental haplotypes with allele values and associated haplotype and genotype values, while at the bottom are two distributions of progeny genetic values in the families (dark bars represent progeny from crossing parent A with B and light bars represent progeny from crossing parent C with B) with overlaid vertical lines denoting parental genetic values (full line) and parent averages (dashed line) – all genetic values have the baseline value of 10 added (CC BY 4.0)

Finally, when we combine between and within-family genetic variation with environmental variation (Figure 13), we start to appreciate the challenge of estimating the unknown genetic values of individuals from data. Namely, Figure 13 shows the variation of a sample of phenotypic values based on genetic values from Figure 12. From the genetic values of parents equal to 8, 12, and 16 units, we generated progeny phenotype values with a range between 0 and 25 units. And this phenotype variation does not yet include other factors, such as sex, farm, and other effects, which would increase the phenotypic variance even more. This book describes methods to make the best estimates of genetic values from a combination of collected phenotypic, pedigree, and genomic data.

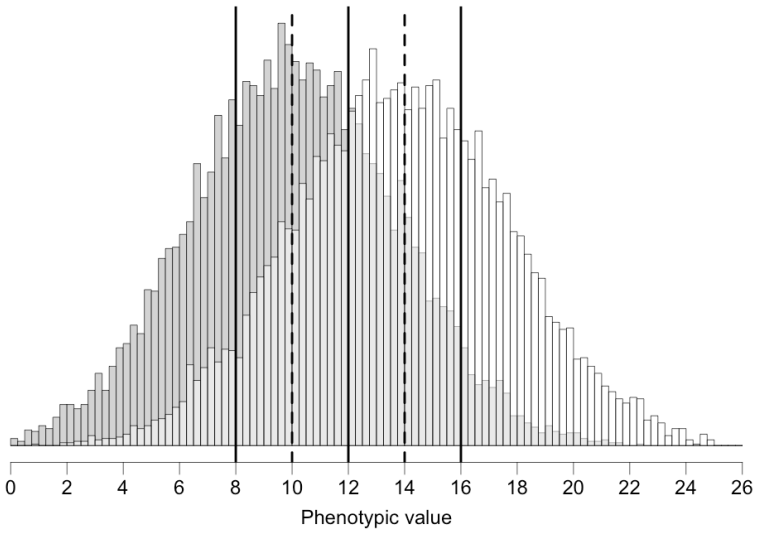

Figure 13: Variation of a sample of phenotypic values with heritability of 0.5 between and within two half-sib families relative to the genetic values of parents (full vertical lines) and corresponding parent average (dashed vertical lines) - see also Figure 12 (CC BY 4.0)

Additional points

To close this chapter, we point to the different types of traits, multiple traits, genotype-by-environment interactions, and biological versus statistical genetic effects. In the introduction of this chapter, we mentioned traits that do not have a continuous distribution. The above-presented genomic and phenotypic data generation processes can also be used for traits with other distributions. When we used Fisher's linear phenotype decomposition, we implicitly used the normal (Gaussian) distribution, assuming a linear link function between the phenotype values and their components. We can relax this assumption with generalized linear models that work with additional distributions and corresponding link functions. An example of such a model is shown in Chapter 15. Further, we have described the data generation model for a single trait only, but the same framework can also be used for multiple traits. The key extension for multiple traits is the addition of genetic and environmental effects for each trait with corresponding variances and covariances among these effects. When we measure the same trait in different environments, we can consider the trait expression in different environments as multiple traits with environment-specific genetic effects representing genotype-by-environment interactions. Finally, in this chapter, we focused solely on additive genetic effects. As described, we do not know the true biological model that generates phenotype values because of the complex underlying biology. Above, we have described a conceptual data generation model following Fisher's linear decomposition of phenotypic values. In these models, the additive genetic effects capture most of the genetic variance (Hill et al., 2008). However, these additive genetic effects are so-called statistical effects – estimated statistically from the data at hand – and are hence data dependent. This means that statistical additive genetic effects likely capture additive genetic variance and a part of non-additive genetic variance. In chapter 13, we describe models for estimating additive and non-additive genetic effects.

Literature

Falconer, D S, MacKay, T F C (1996) Introduction to Quantitative Genetics. 4th edition. Longman, Harlow, UK. ISBN-10: 0582243025.

Fisher, R A (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2):399–433. https://doi.org/10.1017/S0080456800012163

Goriely A (2016) Decoding germline de novo point mutations. Nature Genetics, 48: 823—824. https://doi.org/10.1038/ng.3629

Halldorsson, B V, Eggertsson, H P, Moore, K H S, Hauswedell, H, Eiriksson, O, et al. (2022) The sequences of 150,119 genomes in the UK Biobank. Nature, 607: 732–740. https://doi.org/10.1038/s41586-022-04965-x

Hayes, B J, Daetwyler, H D (2019) 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annual Review of Animal Biosciences, 7(1): 89-102. https://doi.org/10.1146/annurev-animal-020518-115024

Hill W G, Goddard M E, Visscher P M (2008) Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLoS Genetics, 4(2):e1000008. https://doi.org/10.1371/journal.pgen.1000008

Lynch, M (2016) Mutation and Human Exceptionalism: Our Future Genetic Load. Genetics, 202(3):869–875. https://doi.org/10.1534/genetics.115.180471

Lynch, M and Walsh, B (1998) Genetics and Analysis of Quantitative Traits. Sinauer Associates, US. ASIN: B00QAVJKKO.

Ros-Freixedes, R, Valente, B D, Chen, C-Y, Herring, W O, Gorjanc, G, Hickey, J H, Johnsson, M (2022) Rare and population-specific functional variation across pig lines. Genetics Selection Evolution, 54:39. https://doi.org/10.1186/s12711-022-00732-8