Introduction to Basic Concepts in Quantitative Genetics

Guillaume Ramstein & Peter Sørensen

2023-01-11

Contents

Learning objective:

This section introduces the basic concepts in quantitative genetics such as:

- Genetic value and variance for a quantitative trait
- Genetic parameters (heritability, genetic variance and correlation)
- Difference between genotypic values and breeding values
- Infinitesimal model

These concepts are relevant for a range of genetic and statistical analyses of complex traits and diseases in animal and plant populations, including:

- Estimating the heritability of a trait (the part of its variability due to genetics)
- Estimating breeding values by pedigree or genomic information
- Selection of breeding individuals based on estimated breeding values
- Prediction of selection response based on estimated heritability (breeder's equation)

1 Quantitative Genetics

Quantitative genetics, also referred to as the genetics of complex traits, is the study of quantitative traits. Quantitative genetics is based on models in which many genes influence the trait, and in which non-genetic factors may also be important. Quantitative traits such as size, obesity or longevity vary greatly among individuals. Their phenotypes are continuously distributed phenotypes and do not show simple Mendelian inheritance (i.e., their phenotypes are distributed in discrete categories determined by one or a few genes). The quantitative genetics framework can also be used to analyze discrete traits like litter size (which consists of discrete counts like 0, 1, 2, 3, . . .) or binary traits like survival to adulthood (which consists of 0 or 1, 'dead' or 'alive', etc.), provided that they have a polygenic basis (i.e., they are determined by many genes). The quantitative genetics approach has diverse applications: it is fundamental to an understanding of variation and covariation among relatives in natural and managed populations; it is also used as basis for selective breeding methods in animal and plant populations (https://doi.org/10.1098/rstb.2009.0203).

1.1 Infinitesimal model

The infinitesimal model, also known as the polygenic model, is a widely used genetic model in quantitative genetics. Originally developed in 1918 by Ronald Fisher, it is based on the idea that variation in a quantitative trait is influenced by an infinitely large number of genes, each of which makes an infinitely small (infinitesimal) contribution to the phenotype, as well as by environmental (non-genetic) factors. In the most basic model the phenotype (P) is the sum of genetic effects (G), and environmental effects (E):

$$P = G + E \tag{1}$$

The genotypic effect (G) in the model can be split into additive effects (A), dominance effects (D), and epistatic effects (I) such that the expanded infinitesimal model becomes:

$$P = A + D + I + E \tag{2}$$

The genotypic effect may also depend on the environment in which they are expressed (e.g., in plants a drought-tolerance gene may have a favorable effect on grain yield under water-limited conditions, but may be useless under irrigation). Therefore we may consider an extended version of the infinitesimal model where the phenotype (P) is the sum of genotypic effects (G), environmental effect (E), and genotype-environment interaction effects (G×E):

$$P = G + E + G \times E \tag{3}$$

In practice, the genotype-environment interaction effect can be important for the phenotype of individuals, but for the sake of simplicity we will ignore them in the remainder of this section. Therefore, hereafter, we will assume that genotypic effects are not impacted by environmental factors.

1.1.1 Genotypic effects and breeding values

The genotypic effect (G) in the model can include additive effects (A), dominance effects (D), and epistatic effects (I). Additive effects are the summed effects of individual alleles. Dominance effects are interactions between alleles within loci. Epistatic effect are interactions between alleles in different loci and can therefore only occur if two or more loci affect the trait.

Consider an individual that is diploid, like most animals and plants like maize, soybean, barley (i.e., they carry two copies of every genes, except in their sexual chromosomes). Assume that one locus in its genome exists under two possible alleles: A_1 and A_2, with respective allele effects +1 and -1. How do the individual's alleles combine into a genotype? They may combine additively, so that the value of a genotype (the combination of two alleles genotype) is simply the sum of allele effects, but this is only a very special case! If genetic effects are entirely additive, then the value of each possible genotype is the sum of their respective allele effects, i.e., -2 if the individual is A_2A_2, 0 if it is A_2A_1 (or A_1A_2), and +2 if it is A_1A_1.

Generally, the value of each genotype will depend on the combination of alleles within one locus (G = A + D) or across multiple loci (G = A + D + I). For example, in presence of dominance, the value of each possible genotype may be -2 if the individual is A_2A_2, +1 if it is A_2A_1 (or A_1A_2), and +2 if it is A_1A_1.

1.1.1.1   Additive Effects   Additive effects are the summed effects of average allele effects. Quite confusingly, additive effects depend on the population, because average allele effects depend on the frequency of genotypes in the population! For example, assume that genotypes have values -2 (A_2A_2), +1 (A_2A_1) and +2 (A_1A_1). In a population consisting of 25% A_2A_2, 50% A_2A_1 and 25% A_1A_1, you would expect the A_1 allele in a A_1A_2 genotype 2/3 of the time, and you would expect the A_1 allele in a A_1A_1 genotype 1/3 of the time. In another population consisting of 90% A_2A_2, 18% A_2A_1 and 2% A_1A_1, you would expect the A_1 allele in a A_1A_2 genotype about 95% of the time, and in a A_1A_1 genotype only about 5% of the time. As a result, the effect of the A_1 allele, averaged over genotypes, will not be the same, from one population to another. The concept of additive genetic effects and average allele effects is fundamental to quantitative genetics. However, it is one of is most confusing, precisely because of the dependance of allele effects on genotype frequencies.

1.1.1.2   Dominance Effects   Dominance genetic effects are the interactions among alleles at a given locus. This is an effect that is extra to the sum of the additive allele effects. Each genotype has its own dominance effect, denoted by δ_{ij}, for the specific combination of alleles i and j, (e.g., $\delta_{A_1A_2}$), and each of them are non-zero quantities.

1.1.1.3   Epistatic Genetic Effects   Epistatic genetic effects encompass all possible interactions among the loci impacting the trait, whenever there is more than one such loci. This includes all two-way interactions (e.g., interactions between loci A and B, A and C), three-way interactions (e.g., joint interaction among A, B and C), etc. Epistasis can be decomposed, so it includes interactions between additive effects at different loci, interactions between additive effects at one locus with dominance effects at a second locus, and interactions between dominance effects at different loci.

1.1.1.4   Genotypic value versus Breeding value   For selective breeding purposes additive genetic effects are of primary interest. This is because additive effects generally make most of the genotypic effects, and the allelic effects are passed directly to offspring while the other genetic effects are not transmitted to the progeny, and are generally smaller in magnitude. The sum of the additive effects of all loci on a quantitative trait is known as the true breeding value.

- Breeding value = the value of genes to progeny (additive effects only)

- Genotypic value = the value of genes to self (which includes additive, dominance and epistatic effects)

The difference between genotypic value and breeding value is largely dominance deviation. This is because an individual can express dominance deviation (e.g. an A_1A_2 heterozygote). However, an individual cannot pass on dominance deviation to its progeny as it only transmits one allele (e.g., an A_1A_2 heterozygote will either transmit a A_1 gamete or an A_2 gamete to one of its progeny, but not both!) With fully inbred lines, offspring have the same genotype as their parent, and hence the entire parental genotypic value G is passed along. Hence, favorable interactions between alleles are not lost by randomization under random mating but rather passed along. When offspring are generated by crossing (or random mating), each parent contributes a single allele at each locus to its offspring, and hence only passes along a part of its genotypic value. This part is determined by the average effect of the allele. However, any favorable interaction between alleles is not passed along to their offspring.

1.1.2   Infinite number of loci each with small effect on the phenotype

Quantitative traits do not behave according to simple Mendelian inheritance laws. More specifically, their inheritance cannot be explained by the genetic segregation of one or a few genes. Even though Mendelian inheritance laws accurately depict the segregation of genotypes in a population, they are not tractable with the large number of genes which typically affect quantitative traits. To better understand the infinitesimal

model assume Mendelian inheritance to occur at every locus in the genome. Let's say there are 30,000 gene loci in the genome. The number of alleles at each locus varies from 2 to 30 or more. If we assume that there are only two alleles (3 possible genotypes) per locus, and gene loci segregate independently, then the number of possible genotypes (considering all loci simultaneously) would be 3^{30000} which is large enough to give the illusion of an infinite number of loci. Furthermore each of these loci could contribute additive and dominance effects in addition to interaction effects.

1.1.2.1 Distribution of genotypic and phenotypic values in single locus model First we will consider how to model the genetic basis of a quantitative trait when a single locus affects the trait of interest. We call this a single-locus model. The distribution of the genotypic values for a set of individuals will be discrete. The frequency of the genotypic values depend on genotype frequencies, which in turn depend on allele frequencies of A_1 and A_2. The phenotype is however also influenced by the environment. If we assume that the environmental effects are normally distributed (e.g. $\mathcal{N}(0, \sigma^2 = 1)$) then we can observe that the phenotype distribution is infact normally distributed.

1.1.2.2 Distribution of genotypic and phenotypic values in multiple loci model Now we will consider a multiple-locus model. When several loci are causal (i.e., they have an effect on a certain trait), then we talk about a **polygenic model**. Letting the number of causal loci tend to infinity, the resulting model is called an **infinitesimal model**. From a statistical point of view, the breeding values in an infinitesimal model are considered random with a known distribution. Due to the central limit theorem, this distribution tends to a normal distribution, because of the infinitely large number of causal loci. The central limit theorem says that the distribution of any sum of a large number of very small effects converges to a normal distribution. In our case where a given trait of interest is thought to be influenced by a large number of genetic loci each having a small effect, the sum of the breeding values of all loci together can be approximated by a normal distribution. The histograms below show a better approximation to the normal distribution for breeding values (summed allele effects at causal loci), as the number of causal loci increases. In practice, 100 independently segregating causal loci may be large enough, so that the infinitesimal model (and the normal approximation is genomic models) is accurate enough for predictions.

1.1.3 Genetic parameters

Fisher (1918) and Wright (1921) have introduced fundamental statistical methods in quantitative genetics:

- analysis of variance: the partition of phenotypic variation into heritable (A) and non-heritable components (D, I and E).
- resemblance among relatives: the estimations of the proportion of loci shared by relatives under the infinitesimal model.

1.1.3.1 Genetic variance: In the model proposed by Fisher (1918), Cockerham (1954) and Kempthorne (1954), covariance among relatives is described in terms of the additive genetic variance V_A (variance of additive genetic effects, or breeding values), dominance variance V_D (variance of interaction effects between alleles in the same locus), and epistatic variance V_{AA}, V_{AD}, V_{DD}, ... (variance of interaction effects – additive and/or dominance effects – among loci) (Falconer & Mackay 1996; Lynch & Walsh 1998). These partitions are not dependent on numbers of genes or how they interact, but in practice the model is manageable only when the effects are independent from each other, requiring many important assumptions. These include random mating, and hence Hardy-Weinberg equilibrium (i.e. no inbred individuals), linkage equilibrium (independent segregation of loci, which requires many generations to achieve for tightly linked genes) and no selection.

$$\begin{aligned} V_P &= V_G + V_E \\ &= V_A + V_D + V_I + V_E \end{aligned} \tag{4}$$
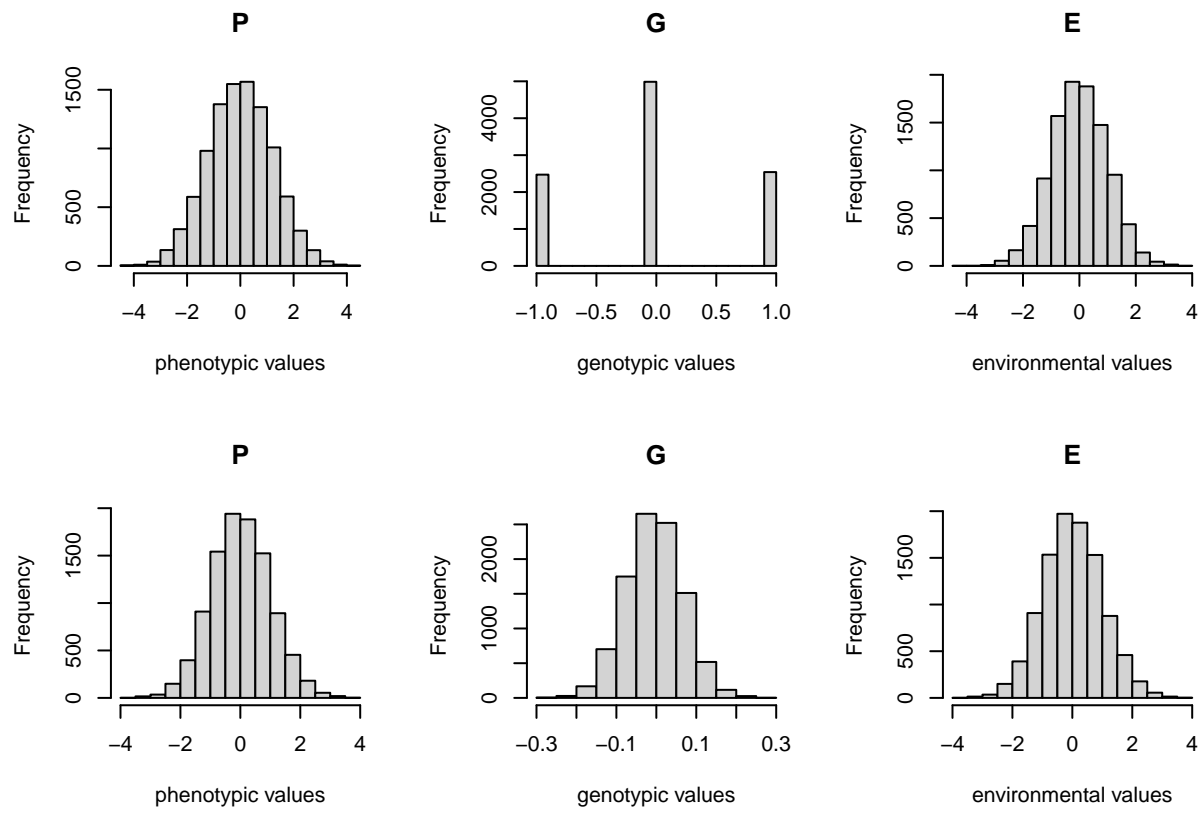
Figure 1: Distribution of genotypic and phenotypic values for a quantitative trait influenced by a single locus model (top panel) or multiple loci (bottom panel)

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$
$$= \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2 \tag{5}$$

Many more terms may be included, such as maternal genetic effects, and genotype $\times$ environment interaction. The model has unlimited opportunities for complexity. This is a strength, in that it is all-accommodating, and a weakness, in that datasets may allow to partition only a few components. In practice, assumptions must be made to reduce the complexity of the resemblance among relatives. Usually, the resemblance among relatives is assumed to depend only on additive genetic variance V_A and dominance variance V_D, so that the following sources of covariation are neglected:

- Epistatic variance (interaction effects among loci are small compared to additive and dominance effects)
- Environmental variance (effects of **shared environments** are assumed to be small enough)

1.1.3.2 Heritability: The models and summary statistics defined by Fisher and Wright have remained at the heart of quantitative genetics, not least because they provide ways to make predictions of important quantities, such as

- Breeding value (A), the expected performance of of an individual's offspring
- Broad-sense heritability, the ratio of total genetic variance V_G to the overall phenotypic variance V_P:

$$H^2 = V_G/V_P$$
$$= (V_A + V_D + V_I)/V_P$$
$$H^2 = \sigma_G^2/\sigma_P^2$$
$$= (\sigma_A^2 + \sigma_D^2 + \sigma_I^2)/\sigma_P^2$$

- Narrow-sense heritability, the ratio of additive genetic variance V_A to the overall phenotypic variance V_P:

$$h^2 = V_A/V_P$$
$$h^2 = \sigma_A^2/\sigma_P^2 \tag{6}$$

- The response to artificial or natural selection, the increase (or decrease) of genotypic values due to selection of individuals, over generations

In view of the assumed complexity of the underlying gene action, involving many loci with unknown effects and interactions, much quantitative genetic analysis has, unashamedly, been at a level of the 'black box'.

1.1.3.3 Genetic correlation: In a general quantitative genetic model, in which, for each individual, two traits (P_1 and P_2) are are each defined as the sum of a genotypic value (G_1 and G_2) and a environmental value (E_1 and E_2):

$$P_1 = G_1 + E_1 \tag{7}$$
$$P_2 = G_2 + E_2 \tag{8}$$

The phenotypic correlation $(\rho_{P_{12}})$ between the traits is defined as:

$$\rho_{P_{12}} = \frac{\sigma_{P_{12}}}{\sqrt{\sigma_{P_1}^2 \sigma_{P_2}^2}}$$

where $\sigma_{P_{12}}$ is the phenotypic covariance and $\sigma_{P_1}^2$ and $\sigma_{P_2}^2$ are the variances of the phenotypic values for the two traits in the population. The genotypic correlation $(\rho_{G_{12}})$ of the traits is defined as:

$$\rho_{G_{12}} = \frac{\sigma_{G_{12}}}{\sqrt{\sigma_{G_1}^2 \sigma_{G_2}^2}}$$

where $\sigma_{G_{12}}$ is the genotypic covariance and $\sigma_{G_1}^2$ and $\sigma_{G_2}^2$ are the variances of the genotypic values for the two traits in the population.

1.1.4 Basic questions remain

On the premise that many genes and environmental factors interact to impact the trait, it will be difficult to determine the action of individual causal genes. Many basic questions remain: What do the genes do; how do they interact; on what traits does natural selection act; why is there so much genetic variation; and can we expect continued genetic improvement in selection programmes? Ultimately, we want to know at the molecular level not just which genes are involved, whether structural or regulatory, but what specific mutation (nucleotide substitution, deletious, copy number variant, etc.) is responsible for genetic effects, and how the causal genes are controlled.