Introduction to Quantitative Genetics

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-03-23

Contents

1 Introduction

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 height, obesity or longevity vary greatly among individuals. Quantitative trait phenotypes are continuously distributed and do not show simple Mendelian inheritance (i.e., phenotypes that are distributed in discrete categories determined by one or a few genes). The quantitative genetics framework can also be used to analyze categorical traits like number of children given birth to (which consist of discrete counts like 0, 1, 2, 3, ...) or binary traits like survival to adulthood (which consist of 0 or 1, 'dead' or 'alive', etc.) or multifactorial diseases as diabetes, provided 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 predicting genetic predisposition in humans as well as selective breeding methods in animal and plant populations.

This section introduces basic concepts used in Quantitative Genetics such as:

- Genetic value and variance for a quantitative trait
- Genetic parameters (genetic variance, heritability, and correlation)
- Single locus model, multiple locus model and infinitesimal model
- Linkage disequilibrium (correlation among markers and QTLs)
- Genetic relationship inferred from pedigree or genetic marker data (correlation among individuals)

These concepts are relevant for a range of genetic and statistical analyses of human complex traits and diseases including:

- Estimating the effect of single locus (or marker) for gene discovery
- Estimating the effect of multiple loci (or markers) for genomic prediction
- Estimating the heritability of a trait (the part of its variability due to genetics)
- Estimating genetic predisposition by pedigree or genomic information

2 Genetic Models

In this section we will be introducing the *single-locus model* for a quantitative trait. Although quantitative traits are most likely influenced by many loci, it helps to first consider the case of only one causal locus in the single-locus model. The single-locus model provides the theoretical basis for more complex models including genomic models (statistical models describing the effects of marker loci) and the infinitesimal model.

2.1 Single-locus model with additive and dominance effect:

In the single-locus model, we consider a biallelic locus with allele A1 and A2 in frequencies p and 1-p. Under random mating and Hardy-Weinberg equilibrium, the expected genotype frequencies are $(1-p)^2$, $2p(1-p)$ and p^2, for A2A2, A1A2 and A1A1 respectively. We arbitrarily assign genotypic values (i.e., trait means per genotype class) $-a$, d and a to the three genotypes, d representing the dominance effect (within locus interaction, no interaction when $d = 0$) and $2a$ the difference between the two homozygotes.

Population mean

Under the single-locus model, the population mean is:

$$\mu = (2p-1)a + 2p(1-p)d \tag{1}$$

Average effect of gene (allele) substitution (also called additive effect in the literature).

The transmission of value from parents to offspring occurs through their genes (alleles) and not their genotypes. The average effect of gene substitution (α) is defined as the average effect on the trait when substituting alleles at this locus in the population. It can also be defined as the mean value of genotypes produced by different gametes:

$$\alpha = a + (1-2p)d \tag{2}$$

Importantly, α is also the slope of the linear regression of the genotype means, weighted by their frequency, on the A1 allele dosage (0, 1 or 2).

When performing a standard single marker association analysis, individual phenotypes y are regressed on the number x ($x = 0, 1, 2$) of reference alleles at a given locus, , i.e., the allelic "dosage", where the reference allele for this dosage count is arbitrarily the major or minor allele (but this arbitrary choice is reflected in the sign of the regression coefficient β:

$$y = \mu + x\beta + e \tag{3}$$

Where the residuals e include both the non-additive genetic effects at the locus, the genetic effects (additive and non-additive) at other loci and an environmental and/or chance (non-genetic) effect. The quantity of interest is the slope β of the model (the effect size of the locus), which is the average effect of allele substitution, hence $\beta = \alpha$.

Additive (breeding) values and dominance deviations

The breeding values are the expected genotypic values under additivity (the predictions from the linear model). Expressed as deviations from the population mean μ, the breeding values of the 3 genotypes A2A2, A1A2 and A1A1 are $-2p\alpha$, $(1-2p)\alpha$ and $(2-2p)\alpha$. The residuals of the linear regression are the deviations due to the within locus interaction (dominance)

Genetic variance

The total genotypic variance (σ^2_g) is partitioned into additive (σ^2_a) and dominance (σ^2_d) variance.

$$\sigma^2_g = \sigma^2_a + \sigma^2_d \tag{4}$$

Additive variance

The additive variance (σ^2_a) is the variance of additive (breeding) values. When values are expressed in terms of deviation from the population mean, the variance simply become the mean of the squared values. Hence, σ^2_a is obtained by squaring the additive (breeding) values described above, multiplying by the corresponding frequencies and summing over the 3 genotypes, leading to:

$$\sigma^2_a = 2p(1-p)\alpha^2 = 2p(1-p)[a + d(1-2p)]^2 = H\alpha^2, \tag{5}$$

with H the heterozygosity at the locus. Note that σ^2_a is the variance explained by a biallelic marker in a GWAS ($2p(1-p)\alpha^2 = 2p(1-p)\beta^2$) and contain both a term due to additivity (a) and dominance (d) through the average effect α^2.

Dominance variance

Similarly, the dominance variance (σ^2_d) is the variance of dominance deviations:

$$\sigma^2_d = (2p(1-p)d)^2 = H^2d^2 \tag{6}$$

Therefore, the dominance variance disproportionally depends on the locus heterozygosity compared to the additive variance (H^2 versus H).

2.2 Two-locus model with additive and additive-by-additive effect:

We extend the one-locus to a two-locus model with additive and additive-by-additive epistatic interaction only, assuming no within loci dominance effects ($d = 0$ at both loci). We introduce a second (unlinked) locus with alleles B1 and B2 in frequencies q and $1 - q$ respectively. The genotypic values and allele frequencies of the 9 genotypes are:

where a_A (a_B) is the genotypic value for the upper homozygote A1A1 (B1B1) and a_{AB} is the additive-by-additive interaction effect. This is a re-parametrization of the model described by Mäki-Tanila and Hill (2014).

Population mean

In our model, the mean of the genotypic values is:

$$\mu = a_A(2p - 1) + a_B(2q - 1) + a_{AB}(1 - 2(p + q) + 4pq) \tag{7}$$

Note that the expression of μ depends on the arbitrarily assigned genotypic values.

Average effect of gene (allele) substitution In this model, the locus specific average effects are:

$$\alpha_A = a_A + 2qa_{AB} \tag{8}$$
$$\alpha_B = a_B + 2pa_{AB} \tag{9}$$

Genetic variance

The total genotypic variance (σ_g^2) of the model is partitioned into additive (σ_a^2) and additive-by-additive (σ_{aa}^2) variance.

$$\sigma_g^2 = \sigma_a^2 + \sigma_{aa}^2 \tag{10}$$

Additive variance

The additive variance of the model is:

$\sigma_a^2 = \sum_i H_i\alpha_i^2$, with H_i the heterozygosity at locus i (i = A, B) and α_i the average effect of locus i. Hence:

$$\sigma_a^2 = 2p(1 - p)[a_A + 2qa_{AB}] + 2q(1 - q)[a_B + 2pa_{AB}] \tag{11}$$

Note that σ_a^2 contains a term due pairwise additive-by-additive interaction between locus A and B (a_{AB}).

Additive-by-Additive variance

The additive-by-additive variance of the model is:

$\sigma_{aa}^2 = \sum_i \sum_{j>i} H_iH_ja_{ij}^2$, with H_i the heterozygosity at locus i (i = A, B) and a_{ij} the additive-by-additive interaction effect between locus i and j. Hence:

$$\sigma^2_{aa} = 4p(1-p)q(1-q)a_{AB} \tag{12}$$

Therefore, the additive-by-additive variance disproportionally depends on the locus heterozygosity as compared to the additive variance.

2.3 General two locus model:

Lastly, we use a generalized two-locus model where the user can provide all the genotypic values in an interactive table and choose the allele frequencies at the two loci (p and q). The genotypic values as well as the linear regressions are plotted as a function of the A1 allelic dosage for the different genotypes at locus B, as well as the linear regression of the genotypic values weighted by their frequency on the A1 allele dosage. The total genotypic variance (VG) of this model is then partitioned in five components:

$$\sigma^2_g = \sigma^2_a + \sigma^2_d + \sigma^2_{aa} + \sigma^2_{ad} + \sigma^2_{dd} \tag{13}$$

Where σ^2_{ad} is the additive-by-dominance variance and σ^2_d the dominance-by-dominance variance.

2.4 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):

$$P = G + E \tag{14}$$

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

$$P = A + D + I + E \tag{15}$$

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 ($G \times E$):

$$P = G + E + G \times E \tag{16}$$

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

2.4.1 Genotypic effects

The genotypic effect (G) in the model can include additive effects (A), dominance effects (D), and epistatic effects (I). Consider an individual that is diploid, like humans (i.e., they carry two copies of every gene, 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.

Additive Effects 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. In general additive effects are the summed effects of average allele effects.

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.

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 L_A and L_B, L_A and L_C), three-way interactions (e.g., joint interaction among L_A, L_B, and L_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.

2.4.2 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.

2.4.2.1 Distribution of genotypic and phenotypic values in single locus model When a single locus affects a quantitative trait, a single-locus model is used to model the genetic basis of the trait. 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 in fact normally distributed (or a mixture of normal distributions).

2.4.2.2 Distribution of genotypic and phenotypic values in multiple loci model When multiple loci affect a quantitative trait, a polygenic- or an infinitesimal model is applied to model the genetic basis of the trait. When several loci are causal (i.e., they have an effect on a certain trait), it is referred to as a **polygenic model**. When the number of causal loci tend to infinity, it is referred to as an **infinitesimal model**. From a statistical point of view, the genetic values in an infinitesimal model are considered random with a known distribution. According to the central limit theorem, the distribution of any sum of a large number of very small effects converges to a normal distribution. Therefore, the genetic values in an infinitesimal model tends to a normal distribution, because of the infinitely large number of causal loci.

2.5 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.

2.5.1 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 σ^2_A (variance of additive genetic effects, or additive genetic values), dominance variance σ^2_D (variance of interaction effects between alleles in the same locus), and epistatic variance σ^2_{AA}, σ^2_{AD}, σ^2_{DD} (variance of interaction effects – additive and/or dominance effects – among loci) (Falconer & Mackay 1996; Lynch & Walsh 1998).

Thus the overall phenotypic variance (σ^2_P) can be partitioned:

$$\begin{aligned}\sigma^2_P &= \sigma^2_G + \sigma^2_E \\ &= \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + ... + \sigma^2_E\end{aligned} \tag{17}$$

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.

The variance–covariance matrix of phenotypic values (V_P) of a group of individuals for a single trait can be partitioned in a similar way

σ^2_A

$$\begin{aligned}V_P &= V_G + V_E \\ &= V_A + V_D + V_{AA} + V_{AD} + V_E\end{aligned} \tag{18}$$

$$\sigma^2_P = A\sigma^2_A + D\sigma^2_D + A \circ A\sigma^2_{AA} + A \circ D\sigma^2_{AD} + I\sigma^2_E \tag{19}$$

where A is the additive genetic relation matrix, and D is the dominance relationship matrix. For the epistatic terms, $\circ$ denotes element-by-element multiplication, but applies only for unlinked loci.

Many more terms may be included, such as maternal genetic effects, and genotype by environment interactions. The advantage of this model is that it is all-accomodating, whereas a disadvantage is 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 [? small enough for what?])

2.5.2 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

Broad-sense heritability, the ratio of total genetic variance V_G to the overall phenotypic variance V_P:

$$\begin{aligned} H^2 &= V_G/V_P \\ &= (V_A + V_D + V_I)/V_P \\ H^2 &= \sigma_G^2/\sigma_P^2 \\ &= (\sigma_A^2 + \sigma_D^2 + \sigma_I^2)/\sigma_P^2 \end{aligned} \tag{20}$$

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

$$\begin{aligned} h^2 &= V_A/V_P \\ h^2 &= \sigma_A^2/\sigma_P^2 \end{aligned} \tag{21}$$

2.5.3 Genetic correlation:

In a general quantitative genetic model in which, for each individual, two traits (P_1 and P_2) are each defined as the sum of a genetic value (G_1 and G_2) and an environmental value (E_1 and E_2), [OR RATHER USE: In a general quantitative genetic model, where two traits (P_1 and P_2) are each defined as the sum of a genetic value (G_1 and G_2) and an environmental value (E_1 and E_2),]

$$\begin{aligned} P_1 &= G_1 + E_1 \\ P_2 &= G_2 + E_2 \end{aligned}$$

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

$$\rho_{P_{12}} = \frac{\sigma_{P_{12}}}{\sqrt{\sigma_{P_1}^2 \sigma_{P_2}^2}} \tag{22}$$

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,

and the genetic correlation ($\rho_{G_{12}}$) of the traits is defined as:

$$rho_{G_{12}} = \frac{\sigma_{G_{12}}}{\sqrt{\sigma_{G_1}^2 \sigma_{G_2}^2}} \tag{23}$$

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

3 Genomic information

The use of genomic information due to the dramatic development in genotyping technologies has revolutionized the field of quantitive genetic. Today dense genetic maps are available for most of the most important plant and animal species including humans. The genetic maps are based on DNA markers in the form of single nucleotide polymorphisms (SNP) and they enable us to divide the entire genome into thousands of relatively small chromosome segments. Ultimately the entire genome may be sequenced, but this is still very expensive.

3.1 Genetic markers

The different locations in the genome that are considered in genomic analysis are called **markers**. When looking at the complete set of markers making up the genomic information in a population, the so-called **Single Nucleotide Polymorphisms** (SNPs) have been shown to be the most useful types of markers. These SNPs correspond to differences of single bases at a given position in the genome. Based on empirical analyses of very many SNP-loci, almost all SNP just take two different states. Furthermore it is important that these SNPs are more or less evenly spread over the complete genome. Some SNPs may be located in coding regions and some my be placed in regions of unknown function.

3.2 Quantitative Trait Loci and linkage disequilibrium

The loci that are relevant for a quantitative trait are called **Quantitative Trait Loci** (QTL). Any given SNP-Marker can only be informative for a given QTL, if a certain **linkage disequilibrium** between the QTL and the marker locus exists. The idea behind linkage disequilibrium is that a certain positive QTL-allele evolved in a certain genetic neighborhood of a number of SNP loci. As a result of that the positive QTL-allele is very often inherited with the same SNP-allele. Over the generations, recombination between the QTL and the neighboring SNP-loci can happen and thereby weaken the statistical association between the positive QTL-allele and the given SNP-allele. This recombination effect is smaller when the QTL and the SNP-loci are physically closer together on the chromosome. The non-random association between QTL and SNP-markers is called linkage disequilibrium.

The marker locus is called M and the QTL is called Q, then the LD can be measured by

$$D = p(M_1Q_1) * p(M_2Q_2) - p(M_1Q_2) * p(M_2Q_1) \tag{24}$$

where $p(M_xQ_y)$ corresponds to the frequency of the combination of marker allele M_x and QTL allele Q_y. Very often the LD measure shown in (24) is re-scaled to the interval between 0 and 1 which leads to

$$r^2 = \frac{D^2}{p(M_1) * p(M_2) * p(Q_1) * p(Q_2)} \tag{25}$$

In (25) r^2 describes the proportion of the variance at the QTL which is explained by the marker M. Hence the LD must be high such that the marker can explain a large part of the variance at the QTL. A large number of SNP markers (>1M) are required to get a sufficient coverage of the complete human genome.

4 Genetic relationships among individuals

Estimating heritability and genetic predisposition requires that the phenotypic covariance between related individuals can be expressed by their additive genetic relationship (A) and the additive genetic variance

(σ_a^2). Related individuals share more alleles and thus resemble each other (have correlated phenotypes, to an extent that depends on the genetic relationships). Genetic relationships can be inferred from pedigree og genetic marker data.

4.1 Genetic relationships among individuals estimated from pedigree data

The genetic covariance between individuals depends on the additive genetic relationship. Examples of different types of additive genetic relationships can be found in the table below. The additive genetic relationship (A_{ij}) between the various sources (j) and the individual itself (i) can be seen in the table below.

Table 1: Examples on additive genetic relationship (A_{ij}) between individual i and j.

Type of relative	A_{ij}
Self	1.0
Unrelated	0
Mother	0.5
Father	0.5
Grandparent	0.25
Child	0.5
Full-sib	0.5
Half-sib	0.25
Twins (MZ/DZ)	1/0.5
Cousin	0.0625

The A matrix expresses the additive genetic relationship among individuals in a population, and is called the **numerator relationship matrix** A. The matrix A is symmetric, where the diagonal elements (*i.e.*, A_{ii}) are equal to $1 + F_i$ where F_i is the **coefficient of inbreeding** of individual i. F_i is defined as the probability that two alleles taken at random from individual i are identical by descent (*i.e.*, that the two alleles originate from the same common ancestor). As such, F_i is also the kinship coefficient of its parents (half their genetic relationship).

Each off-diagonal elements (A_{ij}) is the additive genetic relationship between individuals i and j. Multiplying the matrix A by the additive genetic variance σ_a^2 leads to the covariance among the individual genetic values. Thus, if a_i is the genetic value of individual i then,

$$Var(a_i) = A_{ii}\sigma_a^2 = (1 + F_i)\sigma_a^2. \tag{26}$$

The additive genetic relationship matrix A can be computed using a recursive method.

4.2 Genetic relationships among individuals estimated from genetic markers

A large number of genetic marker are required to get an accurate estimate of the genomic relationships among individuals. The elements in genotype matrix M can be encoded in different ways. Genotypes represent the nucleotide configuration that can be found at a given SNP position. For the use in the linear model we have to use a different encoding. Let us assume that at a given SNP-position, the bases G or C are observed and G corresponds to the allele with the positive effect on our trait of interest. Based on the two observed alleles, the possible genotypes are GG, GC or CC. One possible code for this SNP in the matrix M might be the number of G-Alleles which corresponds to 2, 1 and 0. Alternatively, it is also possible to use the codes 1, 0 and -1 instead which corresponds to the factors with which a is multiplied to get the genotypic values in the single locus model.

Multiplying the matrix M with its transpose M^T results in a $n \times n$ square matrix MM^T. On the diagonal of this matrix we get counts of how many alleles in each individual have a positive effect. The off-diagonal

elements count how many individual share the same alleles across all SNP-positions. The additive genomic relationship matrix G is constructed using all genomic markers as follows:

$$G = \frac{WW^T}{\sum_{i=1}^{m} 2p_i(1-p_i)} \tag{27}$$

where W is the centered and scaled genotype matrix, and m is the total number of markers. Each column vector of W was calculated as follows: $w_i = M_i - 2p_i - 0.5$, where p_i is the minor allele frequency of the i'th genomic marker and M_i is the i'th column vector of the allele count matrix, M, which contains the genotypes coded as 0, 1 or 2 counting the number of minor allele. The centering of the allele counts and scaling factor $\sum_{i=1}^{m} 2p_i(1-p_i)$ ensures that the genomic relationship matrix G has similar properties as the numerator relationship matrix A.

The main difference between the two types of genetic relationship matrices (A and G) is that A is based on the concept of identity by descent (sharing of the same alleles, transmitted from common ancestors) whereas G is based on the concept of identity by state (sharing of the same alleles, regardless of their origin).