Chapter 4

Genome-Wide Association Studies (GWAS)

Marcos Deon Vilela de Resende,[a,b] Fabyano Fonseca e Silva,[b] Márcio Fernando R. Resende Júnior,[c] and Camila Ferreira Azevedo[b]
[a]*Brazilian Enterprise for Agricultural Research on Forestry, Colombo, Brazil,* [b]*Federal University of Viçosa, Viçosa, Brazil,* [c]*University of Florida, Gainesville, FL, USA*

INTRODUCTION

Efforts to use genetic markers for genetic improvement research have diverged into two approaches: QTL (*quantitative trait loci*) identification and mapping; and marker use in genetic selection programs, through marker-assisted selection (MAS) and genome-wide selection (GWS) or genomic selection (GS). This chapter and the next cover both approaches with a special emphasis on genetic selection from genomic data.

QTL ANALYSIS AND GENOMIC SELECTION: CONCEPTS

The use of molecular genetic markers for selection and genetic improvement is based on genetic linkage between these markers and a quantitative trait locus (QTL) of interest. Thus, linkage analyses between markers and QTLs and between the proper multiple markers are essential for genetic selection from genomic information. It must be made clear that by definition, a QTL refers only to the statistical association between a genomic region and a trait.

In classical genetics, linkage between genetic factors or genes has been reported since 1906 and means that closely linked genes on a chromosome are inherited together. In other words, these genes do not segregate independently and thus they do not obey Mendel's Second Law or the Law of Independent Assortment. When these genes are close to each other on a chromosome or linkage group, the linkage is complete. When the genes are part of the same linkage group but are distant from each other, there is a partial linkage.

The calculated genetic distance between two genes is a function of the recombination frequency between the genes and forms the basis for linkage map construction. For linkage between loci to be detected and used in selection, there must be linkage disequilibrium (LD) in the studied population or family.

Biotechnology Applied to Plant Breeding. http://dx.doi.org/10.1016/B978-0-12-418672-9.00004-0

LD or gametic phase disequilibrium is a measurement of the allele interdependence at two or more loci. In a group of individuals, if two alleles from distinct loci are found together more often than would be expected, based on the product of their frequencies, it can be inferred that such alleles are in LD. Disequilibrium values near zero suggest equilibrium or independence between alleles from different genes, and values near one indicate disequilibrium or strong linkage.

LD between markers and QTLs is essential for QTL detection, MAS, and GWS. Of particular importance is the extent of this disequilibrium in a chromosome in a selected population. If a marker and a QTL are in equilibrium in the population, this marker will segregate independently from the QTL. Thus, the marker genotype of an individual has no informative value for selection. The persistence in the population of LD among linked loci depends on the distance between the loci; in other words, it depends on the recombination rate between the two loci. For closely linked loci, any LD that has been created will persist for many generations. However, for weakly linked loci (a recombination rate greater than 0.1), the LD will decrease rapidly. Although a marker (locus *m*) linked to a QTL (locus *q*) might be in linkage equilibrium in a population, there is always disequilibrium within families or crosses, even for weakly linked loci. Additionally, this disequilibrium can extend over large distances because it comes from only one recombination performed to produce the descendants of a heterozygous F_1 individual.

For example, take two linked loci, *m* (marker) and *q* (QTL), in four individuals who are heterozygous for the marker and have the following genotypes: MQ//*mq*, M*q*//*m*Q, MQ//*m*Q, and M*q*//*mq*. The families coming from the two first individuals will be in LD (because, for linked loci, parental gametes are more common than recombinant ones) but in opposite directions because the phase of the QTL marker differs in the two parents. The families from the two last individuals will not be in LD because the QTL does not segregate in these families. When combined across families, the four types of disequilibrium will cancel each other out, creating linkage equilibrium in the population. Thus, LD within each family is useful for QTL analysis as long as different linkage phases are considered.

In population genetics, disequilibrium generically refers to the discrepancy between the joint frequency of a combination of alleles and the product of the alleles' individual frequencies. The term normally refers to alleles from different loci in the same gamete, but can also refer to pairs of alleles of the same locus that show a lack of Hardy-Weinberg equilibrium.

QTL mapping, MAS, proposed by Lande and Thompson (1990), and GWS, proposed by Meuwissen et al. (2001), are based on the presence of linkage disequilibrium in the studied population (or cross). In this situation, marker alleles provide information about the existence and effects of loci that control quantitative traits, providing ways to estimate the effects of QTLs and allowing them to be used efficiently in genetic selection. The causes of LD in a population are mutation, migration, selection, and a small effective population size

(genetic drift due to sampling). In other words, all of the factors that affect Hardy-Weinberg equilibrium in a population also affect linkage equilibrium.

Recently, molecular genetic markers that consist of SNPs (single-nucleotide polymorphisms), which are based on the detection of polymorphisms that arise from a single base change in the genome, have been used. Generally, for an SNP to be considered genetically derived, the polymorphism must occur in at least 1% of the population. SNPs are the most common form of genomic DNA variation and are preferred over other genetic markers due to their low mutation rates and ease and low cost of genotyping. Thousands of SNPs can be used to cover the entire genome of an organism with markers that are not more than 1 cM apart from each other. Microsatellite markers can also be used. These markers are efficient because they are codominant, multi-allelic, abundant, and highly transferable between individuals and species. However, the marker density of microsatellites is usually limited to a few hundred markers. This density can compromise the association study especially in populations with lower LD.

SNP markers are most often bi-allelic, as shown below:

Individual 1: TCA*C*CGCG
Individual 2: TCA*T*CGCG

In this example, there is an SNP polymorphism between the two individuals. A single base change in the DNA sequence results in a polymorphism. More than 1.5 million SNPs have been identified in the human genome. These SNPs exist at an average spacing of 2×10^{-3} cM (Hartl and Jones, 2002).

DArT (*Diversity Array Technology*) markers are also bi-allelic and well suited for GWS because they are abundant, like SNPs, and can be determined rapidly and in large numbers. However, these markers are dominant, which may be a disadvantage when compared to the codominant SNPs. GWS or GS (see Chapter 5) was proposed by Meuwissen et al. (2001) as a way to accelerate and increase the effectiveness of genetic improvement programs. GWS emphasizing the simultaneous prediction (without using significance tests for individual markers) of the genetic effects of thousands of DNA genetic markers that are spread throughout the genome of an organism, in order to the capture the effects of all loci (both small and large effects) and explain all genetic variation of a quantitative trait. To accomplish this, population level LD between the marker alleles and the genes that control the trait is essential.

Linkage Analysis (LA) and Linkage Disequilibrium Analysis (LDA)

The quantity of heritable genetic material in an individual is finite and depends on the genome size. In humans, the genome comprises approximately 35 thousand genes (Ewing and Green, 2000). Thus, a finite number of genes should control a given quantitative trait, making it possible to analyze all of the loci linked to the genetic control of the trait.

There are three basic approaches to QTL discovery: (1) candidate gene approaches, (2) mapping by linkage analysis (LA), and (3) mapping by LD

analysis (LDA). The candidate gene strategy assumes that a gene involved in the underlying physiology of the trait contains a mutation that causes trait variation. This gene is then sequenced in various individuals and differences in the DNA sequences are analyzed for their associations with trait phenotype variations (Anderson and Georges, 2004). This approach has the following problems: there are a large number of potential gene candidates and it is possible that a causative mutation exists in a gene that was not chosen *a priori* as a candidate.

Mapping approaches seek to identify chromosomal regions associated with phenotypic variations in the trait of interest and do not assume *a priori* that the genes themselves are known. Such approaches are therefore based on associations between genetic marker alleles and differences in quantitative traits. A molecular DNA marker is an identifiable physical location on the chromosome for which inheritance can be monitored. The approach therefore consists of identifying markers that are associated with the trait. These markers can be the mutation affecting the quantitative trait, or it can be in LD with the cause of that effect.

A marker is considered to be informative when one can accurately determine which parental allele was transmitted to the progeny. Thus, if a genotyped parent is homozygous for the marker, this marker will not be informative in any of the progeny because it will be impossible to determine which parental allele was transmitted. If both the parent and the progeny are heterozygous, the marker might also be uninformative. If only one parent is genotyped and the progeny have the same genotype as the parent, the progeny could have received the allele from either parent. The expected frequency of individuals for whom the allele origin can be determined will be $1-(p+q)/2$, where p and q are the frequencies of the two parental alleles for the marker. Thus, if only two marker alleles are present in the population, half of the offspring will have the same genotype as their parent. For multi-allelic loci such as microsatellites, $(p+q)$ can be much lower than 1 (Weller, 2001).

The LA strategy considers only the LD that exists within a family or cross; this can extend for dozens of cM and is interrupted by recombination after a few generations. This approach uses a limited number of markers per chromosome; therefore, due to recombination between distant markers and the QTL, the association between the markers and the QTL will persist only within a family and for a limited number of generations. This strategy allows QTL mapping of a large interval on the chromosome unless a large number of individuals per family are used. The formula devised by Darvasi and Soller (1997) illustrates this. In a specific genetic map of cattle created by high marker densities, the confidence interval (CI) is given as $CI = 3000/(kns^2)$, where k is the number of informative genitors per individual (1 for families with half siblings and 2 for families with full siblings and F_2 populations), n is the number of genotyped individuals, s is the effect of allelic substitution in connection with the favorable QTL and 3000 cM is the size of the bovine genome (in this species, 1 cM contains approximately eight genes).

Based on this equation and a QTL that segregates with an s equal to a 0.5 residual standard deviation, a family of 1000 half-siblings will have a 95% CI of 12 cM. The following points should be considered for this large CI: (1) if the goal is to use a gene candidate approach within this interval, a large number of genes will need to be sequenced and studied (80 genes, assuming a total of 20,000 genes in a 3000 cM genome) and (2) if the goal is to use MAS, the linkage between the marker and the QTL will not be sufficiently close to guarantee that the association between the marker and the QTL will be consistent across the population. In this latter case, the marker-QTL linkage phase within each family should be established before undergoing MAS (Hayes, 2008). For example, an individual in the population could carry allele *M* of the marker linked to the favorable allele of the QTL, and another individual from the same population but from a different family could carry allele *m* for the marker linked to the same favorable QTL allele.

The LA approach is based on associations between marker alleles and the phenotypic QTL classes and, until recently, was widely used because the numbers of markers identified in various species were low and the cost of genotyping was very high. With the recent advent of SNP markers, which exist in large numbers and can be inexpensively genotyped, the use of high-density markers in the genome became possible and finding markers close to the actual QTL became viable. Thus, adoption of the LDA approach, which is superior to LA, became possible.

The LDA strategy is based on linkage disequilibrium between a marker and a QTL in a population and not just within a family, as is done in LA. For this to occur, the marker and QTL must be linked very closely. When this occurs, the association between them is a property of the entire population and will persist for many generations.

Meuwissen and Goddard (2000) showed that the confidence interval could be reduced to 1 cM with LDA mapping. If a QTL polymorphism is due to a recent mutation or to a recent introduction of another population, the LD between a QTL and closely linked markers can be detected at the population level. The closer the marker to the QTL, the greater will be the LD. The confidence interval can be further reduced by combining the LA and LDA strategies and by multiple trait analysis (Meuwissen and Goddard, 2004).

Association analysis is used for fine mapping and is based on population level LD. Linkage can occur when the gene directly affects a trait and when there is LD between the marker and the gene controlling the trait. In the first case, the effect of the gene is measured directly and the marker is functional. The functional mutations are known as quantitative trait nucleotides (QTN). In the second case, the linkage test requires LD between the marker and the QTL. When a mutation occurs on a given chromosome, it creates a haplotype with adjacent loci on the chromosome. In the subsequent generation, this mutation tends to occur within the same haplotype unless there is recombination. This creates the LD used for association mapping.

GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Coefficients and Measurements of Linkage Disequilibrium

As seen previously, LD is defined as the non random association of alleles at different loci. For example, take a locus with alleles A and a and another locus with alleles B and b. The gametic disequilibrium is given by $D = \text{prob}(AB)\ \text{prob}(ab) - \text{prob}(Ab)\ \text{prob}(aB)$, where prob denotes the probability or frequency of the respective haplotypes. Thus, disequilibrium exists (D is not zero) when the gametes in coupling occur at a different frequency than those in repulsion. Positive values for D show that the gametes in coupling are in excess. Negative D values show that the gametes in repulsion are in excess. After t generations of random crosses, $D_t = D_0(1-r)^t$, and therefore $t = (\log D_t)/[\log D_0(1-r)]$ calculates the number of generations needed to reach equilibrium, where D_0 is the initial disequilibrium and r is the recombination rate.

Take the following allelic frequencies: $\text{p}(A) = p_1$; $\text{p}(a) = p_2$; $\text{p}(B) = q_1$ and $\text{p}(b) = q_2$. These have the following equalities: $D = \text{prob}(AB)\ \text{prob}(ab) - \text{prob}(Ab)\ \text{prob}(aB) = P_{11}\ P_{22} - P_{12}\ P_{21} = p_1q_1\ p_2q_2 - p_1q_2\ p_2q_1 = P_{11} - p_1q_1 = P_{22} - p_2q_2 = p_1q_2 - P_{12} = p_2q_1 - P_{21}$. Thus, the maximum and minimum values for disequilibrium are given by $D\text{max} = \min\ (Ab, aB) = \min\ (p_1q_2, p_2q_1)$ and $D\text{min} = \max\ (AB, ab) = \max\ (-p_1q_1, -p_2q_2)$.

As an example, consider the following: two loci with two alleles are segregating in the population, and the following information is given: $\text{prob}(AB) = 0.35$, $\text{p}(A) = 0.7$, and $\text{p}(b) = 0.4$. Is this population in gametic equilibrium? Based on the provided information, we have $\text{p}(B) = 1 - 0.4 = 0.6$, and the expected probability of AB is $\text{P}(AB) = \text{p}(A)\ \text{p}(B) = 0.7 \times 0.6 = 0.42$. Thus, $D = \text{prob}(AB) - \text{p}(A)\ \text{p}(B) = P_{11} - p_1q_1 = 0.35 - 0.42 = -0.07$. Therefore, the population is in LD and there is an excess of gametes in repulsion. Assuming that the linked loci have a recombination rate of 2%, the number of generations until the disequilibrium falls to one half ($D_t/D_0 = 0.5$) is given as $D_t/D_0 = (1-r)^t = 0.5$. Thus, $0.5 = (1-r)^t$ and $0.5 = (1-0.02)^t$, and solving for t yields $t = 34.31$ generations.

The LD statistic shown above, $D = \text{prob}(AB)\ \text{prob}(ab) - \text{prob}(Ab)\ \text{prob}(aB)$, is very sensitive to individual allele frequencies and therefore cannot be used to compare LD between multiple pairs of loci that cover various locations across the genome. The r^2 statistic developed by Hill and Robertson (1968), is more adequate because it is less dependent on allelic frequencies. This statistic is given by $r^2 = D^2/[\text{prob}(A)\ \text{prob}(a)\ \text{prob}(B)\ \text{prob}(b)]$. Values for r^2 vary from zero (pairs of loci with no disequilibrium between them) to 1 (pairs of loci in complete LD). In the example above, the following haplotype frequencies are observed: $\text{P(AB)} = 0.35$; $\text{P}(ab) = 0.05$; $\text{P}(aB) = 0.25$; $\text{p}(Ab) = 0.35$. Therefore, $D = \text{P}(AB)\ \text{P}(ab) - \text{P}(Ab)\ \text{P}(aB) = -0.07$ and $D^2 = 0.0049$. The value of r^2 is thus given by $r^2 = D^2/[\text{prob}(A)\ \text{prob}(a)\ \text{prob}(B)\ \text{prob}(b)] = 0.0049/[(0.7)\ (0.3)\ (0.6)\ (0.4)] = 0.0972$. This level of disequilibrium is considered to be low. Moderate r^2 values are on the order of 0.2 or higher (Hayes et al., 2006).

Another measure for LD is the statistic $D' = \text{modulus}\ (D)/D\text{max}$, proposed by Lewontin (1964), which standardizes D by Dmax. The Dmax is given by $D\text{max} = \min(p_1q_2,\ p_2q_1)$ if $D > 0$ and $D\text{max} = \min(p_1q_1,\ p_2q_2)$ if $D < 0$. This measure of LD is not very accurate because it can be inflated when estimated from small samples or in situations with low allelic frequencies (McRae et al., 2002). Another limitation of D' is its inability to predict the necessary marker density to completely cover the genome by LD.

Thus, the r^2 statistic is preferred. The genetic significance of the r^2 between a marker and an unobserved QTL is that the r^2 measures the proportion of the variation caused by the QTL alleles as explained by the marker alleles. Thus, r^2 decreases as the distance increases, therefore indicating the numbers of markers and phenotypes necessary to make accurate predictions in the context of GWS and QTL detection by LD at the population level. Sample sizes should increase at a rate given by $1/r^2$ to detect an unobserved QTL relative to the sample needed to measure the own QTL (Pritchard and Przeworski, 2001).

The measures of disequilibrium presented herein specifically address loci with two alleles or bi-allelic markers. This is sufficient for SNP markers, but these measures can be extended to multi-allelic markers such as microsatellites. Thus, an estimator for multi-allelic linkage disequilibrium, proposed by Zhao et al. (2005), uses the x^{2*} statistic given by $\chi^{2*} = [1/(m-1)]\sum_{i=1}^{k}\sum_{j=1}^{n}\{D_{ij}^2/[p(a_i)p(b_j)]\}$, where $D_{ij}^2 = p(a_ib_j) - p(a_i)p(b_j)$ and $p(a_i)$ and $p(b_j)$ are the frequencies of the alleles i and j of the markers a and b, respectively. Additionally, $p(a_ib_j)$ represents the frequency of the haplotype (a_ib_j). The quantity m represents the minimum of the number of alleles for the markers a and b. The x^{2*} statistic is a generalization of r^2, and $x^{2*} = r^2$ for bi-allelic markers. Simulations performed by Zhao et al. (2005) show that x^{2*} is the best predictor of the proportion of the variance caused by alleles of the QTL explained by the markers.

The r^2 statistic developed by Hill and Robertson (1968), given by $r^2 = D^2/$ [prob(A) prob(a) prob(B) prob(b)], has an expectation or expected value that is shown by the following equation by Sved (1971): $E(r^2) = 1/(4\ Ne\ L + 1)$. This equation can also be expressed as a function of the recombination rate r in Morgans, resulting in $E(r^2) = 1/(4\ Ne\ r + 1)$. Thus, based on the effective population size (Ne) and the recombination rate, the r^2 can be inferred. Inferred r^2 values are important for calculating the accuracy of GWS.

In exogamous domestic species (animals and perennial plants that are preferentially allogamous), the reduced effective population size is the main cause of LD. In such cases, the expected value for disequilibrium in a given chromosomal segment of size L (in Morgans) can be calculated by the expression $E(r^2) = 1/(4\ Ne\ L + 1)$. The Sved equation shows that the LD decreases rapidly as the distance between the genes increases or as the size of the segment in question increases. This reduction becomes even larger as the effective population size increases (Table 4.1).

For the effective population sizes used in perennial plant improvements (30 to 100), sufficiently large LDs (equal to or greater than 0.2) for QTL selection

TABLE 4.1 Expected Values ($E(r^2)$) for the Linkage Disequilibrium between Two Loci as a Function of the Effective Population Size (*Ne*) and Length (*L*) of the Chromosomal Segment That Separates the Two Loci

Ne	L (Morgan)	L (centiMorgan)	$E(r^2)$	*Ne*	*L* (Morgan)	L (centiMorgan)	$E(r^2)$
10	0.005	0.5	0.83	100	0.005	0.5	0.33
10	**0.01**	**1**	**0.71**	**100**	**0.01**	**1**	**0.20**
10	0.02	2	0.56	100	0.02	2	0.11
10	0.03	3	0.45	100	0.03	3	0.08
10	0.04	4	0.38	100	0.04	4	0.06
10	0.05	5	0.33	100	0.05	5	0.05
20	0.005	0.5	0.71	200	0.005	0.5	0.20
20	**0.01**	**1**	**0.56**	**200**	**0.01**	**1**	**0.11**
20	0.02	2	0.38	200	0.02	2	0.06
20	0.03	3	0.29	200	0.03	3	0.04
20	0.04	4	0.24	200	0.04	4	0.03
20	0.05	5	0.20	200	0.05	5	0.02
30	0.005	0.5	0.63	500	0.005	0.5	0.09
30	**0.01**	**1**	**0.45**	**500**	**0.01**	**1**	**0.05**
30	0.02	2	0.29	500	0.02	2	0.02
30	0.03	3	0.22	500	0.03	3	0.02
30	0.04	4	0.17	500	0.04	4	0.01
30	0.05	5	0.14	500	0.05	5	0.01
50	0.005	0.5	0.50	1000	0.005	0.5	0.05
50	**0.01**	**1**	**0.33**	**1000**	**0.01**	**1**	**0.02**
50	0.02	2	0.20	1000	0.02	2	0.01
50	0.03	3	0.14	1000	0.03	3	0.01
50	0.04	4	0.11	1000	0.04	4	0.01
50	0.05	5	0.09	1000	0.05	5	0.00

can be obtained with markers that are spaced 1 to 3 cM apart. The r^2_{mq} or $E(r^2)$ is a weighted average of the r^2 of each marker-QTL pair, being r^2 the square of the correlation (r) between the alleles or genotypes present at the marker locus and the QTL locus (Table 4.2).

TABLE 4.2 Calculations of the Linkage Disequilibrium between a Marker and QTL

Individual	Num. alleles at marker locus (X_a)	Num. alleles at QTL locus (X_b)
1	0	0
2	2	1
3	1	1
4	1	0
5	2	1
Correlation (r)	$r = 0.76$	$r^2 = 0.58$

The correlation coefficient between two variables or alleles at loci a and b is given by

$$\begin{aligned} r &= \frac{Cov(a,b)}{[Var(a)Var(b)]^{1/2}} = \frac{\sum ab - \sum a \sum b}{[Var(a)]^{1/2}[Var(b)]^{1/2}} \\ &= \frac{\Pr ob(ab) - \Pr ob(a) \Pr ob(b)}{[pq]^{1/2}[rs]^{1/2}} = \frac{D}{[pq\ rs]^{1/2}} \end{aligned}$$

The square of this quantity equals $r^2 = \frac{D^2}{[pq\ rs]}$, which is the standard measurement for LD. Using the incidence matrix X for the markers, the value of r can be expressed as

$$r_{(a,b)} = \frac{Cov(X_{ia}, X_{ib})}{[Var(X_{ia})]^{1/2}[Var(X_{ib})]^{1/2}}$$

D is defined by $D = \text{Prob}(ab) - \text{Prob}(a)\ \text{Prob}(b)$, where $\text{Prob}(a)$ is the frequency of allele a and $\text{Prob}(ab)$ is the frequency of genotype ab. Generically, p, q, r, and s are the frequencies of the alleles A, a, B, and b, respectively. The equation $\text{Var}(a) = pq$ assumes a Bernoulli distribution for the presence of an allele.

The relationship between the genetic effects of a marker and a QTL can be better understood with the following models: genetic effect of the QTL on the phenotype $(g_{QTL}), y = u + g_{QTL} + e$ and genetic effect of the marker on the phenotype $(g_m), y = u + g_{QTL} + e = u + Xg_m + e$. The variable g_m is a regression coefficient given by

$$\begin{aligned} g_m &= Cov(y,X)/Var(X) = Cov(g_{QTL},X)/Var(X) \\ &= r[Var(g_{QTL})/Var(X)]^{1/2} = r\{Var(g_{QTL})/[2p(1-p)]\}^{1/2} \end{aligned}$$

The amount of variation in the QTL explained by the marker is given by

$$Var(Xg_m) = g_m^2 Var(X) = r^2[Var(g_{QTL})/Var(X)]Var(X) = r^2 Var(g_{QTL})$$

Thus, we find that r^2 is the proportion of variance in the QTL explained by the marker.

The extent of the LD depends on recent and historical recombinations as well as the current and past *Ne*. Domesticated plant and animal populations have lower current *Ne* than past *Ne*. In humans, the contrary is true due to the current rapid increase in the population. Hayes et al. (2003) stated that LD in short chromosomal segments (short distances) depends on the historical effective population size many generations ago. However, disequilibrium across long distances depends on the recent history of the population. As linear changes in populations occur, it must follow that the measure r^2 reflects the *Ne* associated to $1/(2L)$ generations back. Therefore, the expected r^2 when *Ne* changes over time is given by $E(r^2) = 1/(4\ Ne_t\ L + 1)$, where $t = 1/(2L)$. In humans, the *Ne* equals approximately 10,000 (Kruglyak, 1999). In domesticated animals and perennial plants, the *Ne* can be lower (in the order of 100). Therefore, the LD should be lower in humans. However, in the past, the *Ne* of the human population was low. Thus, for long distances between markers, the r^2 values are lower in humans than in domesticated plant and animal species. And the r^2 values for short distances between markers are more similar in humans and domesticated animal species. Moderate LDs (r^2 greater than or equal to 0.2) in humans extend for fewer than 5 kb or 0.005 cM. In cattle, moderate LDs extend up to 100 kb. However, high LD values (r^2 greater than or equal to 0.8) extend for only very short distances in both humans and cattle (Tenesa et al., 2007).

Dutch and Australian dairy cattle populations show a similar decline in LD because these populations are related by origin and have similar histories and *Ne* values. Norwegian red cattle (*Ne* equal to 400) have a faster decline in LD than Dutch dairy cattle (global *Ne* of 150). The different *Ne* values justify the different LDs in both populations (Zenger et al., 2007).

Methods for QTL Analysis via LDA

For a long time, mapping studies were based on linkage analyses associated with pedigree data. More recently, methods based on LD in unrelated individuals have been recommended as powerful tools with which to produce precise estimates of gene locations. These methods are based on the premises listed below. When a novel allele is introduced into the population either by mutation or migration, it exists in the population with a group of marker alleles. The length of this haplotype is reduced over several generations due to recombination events and, after many generations, only the markers in the immediate neighborhood of the new allele locus are likely to remain on the same chromosome segment.

Due to the low LD present in these populations, the associations of markers with the trait are harder to find and extremely high marker density is required to increase the probability that a marker will be in LD with the causal loci. Despite this disadvantage, whenever markers are identified to be associated with an allele that influences a given trait, a strong correlation between the trait and the marker should indicate that the coding locus for the trait is located very near the marker or the identified marker could in fact be QTN.

LDA mapping seeks to increase the precision of QTL position estimates because, in some situations, the meiosis number in the genotyped pedigree is not sufficient for the LA to be precise. LDA methods provide fine mapping based on quantified LD in the gametic phase which is persistent across families in an allogamous population. In this case, the linkage phase does not change between families or between generations. This method is based on the fact that in a small population, the founders have a small number of distinct haplotypes and, at highly linked loci, there is not sufficient time for recombination to break the associations between markers and the mutations that affect the QTL (Perez-Enciso et al., 2003). This type of mapping is also known as association mapping, which became possible with the advent of high-density SNP and DArT markers. The strategy of population-level trait-marker association depends on small blocks of genes in disequilibrium, and therefore the resolution is very high (small distances between genes). Although the resolution is higher, QTL detection and precision mapping require a large number of markers. Association mapping uses the general population rather than a specific mapping population. The association between a marker and a QTL depends on the recombination frequency between them. To find a marker reasonably close to a QTL, there must be a low recombination rate. The larger the LD, the closer the marker is to the gene, and this LD or association will be valid even for genetically distant individuals.

Two approaches, genomic scanning and candidate genes, can be used in association genetics or mapping. In the latter approach, only markers in the individual candidate genes are used. For association genetics, the mapping population should be large and have a large degree of LD. LDA mapping uses genomic scanning with high-density markers (1 marker per 0.5 to 2 cM). The success of the method depends on the amount of LD in the population. Because markers can be in incomplete LD with the QTL, both the association between markers and QTLs in the population and the co segregation of markers and QTL within families can be used simultaneously for QTL detection via the LDA-LA method, which combines the properties of LD (in linkage disequilibrium) and LE (in linkage equilibrium) markers, respectively.

LDA-based mapping is performed by calculating the probabilities that the haplotypes shared by individuals are identical by descent from a common ancestor, conditional on marker data. Accurate determination of the linkage phases and QTL genotypes is necessary for fine mapping. Thus, a pure LDA analysis might result in a high number of false positives or false inferences of association in the absence of linkage. Therefore, methods (LA-LDA) that simultaneously

incorporate information about population LD and linkages within families are recommended to mitigate the effects of spurious associations between the markers and QTLs (Meuwissen and Goddard, 2004).

MAS and GWS are increasingly effective as the markers become closer to the QTLs. Given the small distances between the genes on chromosomes, the accurate mapping of QTLs is difficult. On average, a 10-cM chromosomal segment can contain approximately 200 genes. Thus, a high density of genotyped markers increases the QTL mapping resolution. However, if the aim is to find the actual gene that affects a trait, the confidence interval for the QTL remains large, even for a QTL with a large effect and when using a large sample size (Weller, 2001). LDA mapping strategies are described below.

GENOME-WIDE MAPPING VIA SINGLE MARKER REGRESSION

To identify statistically significant effects, genome-wide association studies (GWAS) seek out associations between loci and phenotypic traits in a population by hypothesis testing. The following regression model for single markers can be used to find QTL-marker associations in a panmictic population (Resende, 2008): $y = l\mu + Xm_i + e$, where y is the vector of observed phenotypes, l is a vector with the value of 1, μ is a scalar with the overall mean, m_i is the fixed effect of one of the bi-allelic marker alleles, and e is the vector of random residuals. X is the incidence matrix for m_i. This model assumes that the marker will affect the trait only if it is in LD with the putative QTL. Other fixed and random effects can be incorporated into this model. For example, consider the analysis of 12 individuals for a particular trait and an SNP marker. The genotypic and phenotypic data for the individuals are shown in Table 4.3.

The incidence matrix X connects the number of each SNP allele to the individual phenotypes. Only the effects of one of the alleles need to be included. Thus the matrix X will have only one column for the effects of one of the SNP alleles, such as A. This column contains the number of copies of A possessed by the individuals. Therefore, it has values of 0, 1, or 2 for a diploid individual. The number of rows in the matrix is equal to the number of individuals.

The matrix 1 includes a column for the overall mean. The matrices 1 and X (number of A alleles), shown as transposed matrices, are $1'_{(12x1)} = [1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1]$ and $X'_{(12x1)} = [1\,2\,1\,2\,1\,0\,2\,1\,0\,2\,1\,2]$. The least squares equations for estimating the effects of the overall mean and the SNP is:

$$\begin{bmatrix} 1'1 & 1'X \\ X'1 & X'X \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{m}_i \end{bmatrix} = \begin{bmatrix} 1'y \\ X'y \end{bmatrix},$$

where y is the phenotype vector. Solving this system yields:

$$\begin{bmatrix} \hat{u} \\ \hat{m}_i \end{bmatrix} = \begin{bmatrix} 7.2713 \\ 3.7856 \end{bmatrix}$$

TABLE 4.3 Genotypic and Phenotypic Data for 12 Individuals for a Particular Trait and an SNP Marker

Individual	Phenotype	First SNP1 allele	Second SNP1 allele
1	9.87	*A*	*a*
2	14.48	*A*	*A*
3	8.91	*A*	*a*
4	14.64	*A*	*A*
5	9.55	*A*	*a*
6	7.96	*a*	*a*
7	16.07	*A*	*A*
8	14.01	*A*	*a*
9	7.96	*a*	*a*
10	21.17	*A*	*A*
11	10.19	*A*	*a*
12	9.23	*A*	*A*

The null hypothesis, which states that the marker does not have any effect on the trait, can be evaluated with the *F*-test. The null hypothesis is rejected if $F > F(a,v_1,v_2)$, where F is the *Snedecor* statistic calculated from the data, a is the significance level and v_1 and v_2 are the degrees of freedom for the F distribution. The alternative hypothesis is that the marker affects the trait because the marker and the QTL are in LD. The *F*-statistic value is calculated as

$$F = \frac{QM\ \text{Re}\ gression}{\hat{\sigma}_e^2} = \frac{\hat{m}X'y + \hat{u}1'y - (1/n)(1'y)^2}{(y'y - \hat{m}X'y - \hat{u}1'y)/(n-2)}$$

In this example, the calculated value for F was 9.74. This value can be compared to the *F*-value for a 5% significance level and 1 and 10 degrees of freedom on an *F*-table, which gives 4.96. Thus, the SNP effect is significant. This was expected because the individuals with higher phenotypic values were those with a larger number of *A* alleles for the SNP, as clearly shown in the data table.

STATISTICAL POWER AND SIGNIFICANCE OF ASSOCIATION FOR QTL DETECTION

The power of the marker-QTL linkage test depends on the following factors (Pritchard and Przeworski, 2001; Meuwissen et al., 2002; Hayes et al., 2006; Fernando et al., 2004; Macleod et al., 2010):

1. The r^2 (statistical measurement of the LD) between the marker and QTL. The genetic significance of the r^2 between a marker and an unobserved QTL is that it measures the proportion of the variance caused by alleles of the QTL explained by the marker alleles. The sample size required to detect an unobserved QTL should increase at a rate of $1/r^2$, compared to the sample needed to analyze the QTL itself.
2. The proportion of phenotypic variance (σ_f^2) explained by the QTL; in other words, the coefficient of determination for the q QTL effect ($h_q^2 = \sigma_q^2/\sigma_f^2$).
3. The number of individuals analyzed.
4. The chosen significance level.
5. The frequency, p, of the rare marker allele, which determines the minimum number of observations necessary to estimate the allelic effect. When p is less than 0.1, the power is sensitive to this allelic frequency.

The power of a test is the probability of rejecting H_0 when H_0 is false, or the ability to detect a QTL in the population when it exists. The power of a test to detect a QTL at various r^2 values for a QTL-marker pair can be calculated with a formula derived by Luo (1998). To achieve a power greater than or equal to 80% for detecting a QTL with an h_q^2 equal to 0.05 based on 1000 phenotypic observations, an r^2 of at least 0.2 is required. This result assumed that the frequency of the rare allele is higher than 0.2.

Macleod et al. (2010) reported that the power of QTL detection with an h_q^2 of 5% and 365 genotyped individuals was 37% ($p<0.001$). The authors also found a strong correlation between the F-values of significant SNPs and their r^2 with the QTL. The correlation between the *Snedecor F*-statistic and D' was practically zero.

When making a conclusion, a researcher commits a type I error if he/she rejects a true H_0 hypothesis and a type II error if he/she accepts a false H_0 hypothesis. The probability of committing a type I error is called α, and the largest value of α for a true H_0 is called the significance level of a statistical test; in other words, the significance of a test is the maximum probability allowed for making a type I error.

Choosing the significance level to be used in GWAS requires several considerations. This is because thousands of markers will be tested and therefore one confronts the problem of determining the nominal significance levels to be used with multiple tests. Thus, the nominal significance level for each test will not correspond to the level used for the experiment as a whole. At a significance level of 5%, it is expected that 5% of the results will be false positives. If 20,000 markers are tested, 1000 false positives are expected. The Bonferroni correction can mitigate this problem. However, this correction does not account for the fact that tests on the same chromosome are not independent because the markers can be in LD between themselves and also with the QTL.

The permutation test technique was proposed by Churchill and Doerge (1994) to address the challenge of multiple tests in QTL mapping experiments. This technique is used to establish appropriate significance levels. Hoggart et al.

(2008) derived an explicit approximation for the type I error rate, thus avoiding the need for permutation procedures. Another option to avoid false positives is to track this number in comparison with the number of positive results, as done in Fernando et al. (2004). The researcher can thus establish a significance level based on an acceptable proportion of false positives.

In addition to the correction techniques listed above, there is the false discovery rate (FDR), which is defined as the expected proportion of detected false positive QTLs. The FDR can be calculated as FDR = m Pmax/n, where Pmax is the largest QTL P-value that exceeds the significance level, n is the number of QTLs that exceed the significance threshold, and m is the number of tested markers (Weller, 2001). If 10,000 SNPs are tested with a significance level (P-value) of 0.001 and 80 SNPs are considered significant, then FDR = $10{,}000 \times 0.001/80 = 0.125$. This result (12.5%) is considered acceptable for the FDR.

An alternative approach to decreasing the false positive rate is the use of a model that includes a vector of polygenic effects, which includes a relationship matrix and permits correction for the population structure. Macleod et al. (2010) reported an increased number of false positives (type I errors) when polygenic effects were not included in the model. In this case, it is recommended that markers be used to infer the relationship matrix as shown by Hayes et al. (2007). For a given marker locus, the genetic similarity, Sxy, between two individuals x and y is calculated as follows (Hayes et al., 2006, 2007):

1. $Sxy = 1$ when the genotype $x = ii$ (both alleles at the locus are identical) and the genotype $y = ii$ or when $x = ij$ and $y = ij$;
2. $Sxy = 0.5$ when the genotype of $x = ii$ and the genotype of $y = ij$ or vice versa;
3. $Sxy = 0.25$ when the genotype of $x = ij$ and the genotype of $y = ik$;
4. $Sxy = 0$ when the two individuals do not share alleles at the locus.

Similarity arising by chance is represented by

$$Sa = \sum_{i=1}^{g} p_i^2,$$

where p is the allele frequency in the population and g is the number of alleles at the locus. The relatedness between individuals x and y for the locus is thus expressed as $r = (Sxy - \mathrm{Sa})/(1 - \mathrm{Sa})$. The average relatedness between the individuals is calculated as the average r across all of the loci. Thus, when a large number of markers are present, the resulting marker relationship matrix can capture the effects of Mendelian segregation, which are not included in pedigree-based relationship matrices.

Cross-validation methods can be used to estimate confidence intervals in GWAS. For these methods, the data are divided into two halves, and the association analysis is performed three times, once for each half of the data and once with all of the data. The 95% confidence interval for the QTL position is given by the position of the most significant SNP in an analysis of the complete

dataset $\pm 1.96s$, where s is the standard error of the QTL and is calculated from the equation

$$s = \left(\frac{1}{4n} \sum_{i=1}^{n} x_{1i} - x_{2i} \right)^{1/2}$$

for n SNP pairs with significant effects. The components x_{1i} and x_{2i} are the positions of the most significant SNP in each of the half-datasets for the ith most significant QTL position in the complete dataset. This is valid when the analysis of each half-dataset confirms an SNP that was declared to be significant in the complete dataset analysis (Hayes et al., 2006, 2007).

GENOME-WIDE MAPPING WITH HAPLOTYPE MIXED MODELS

Haplotypes are specific combinations of multiple linked markers and can be considered as alleles of a "super-locus." Haplotypes can be used in place of simple markers in GWAS. An advantage is that they can exist in greater disequilibrium with the QTLs. When this occurs, the r^2 is larger and therefore the power of the experiment is increased. The proportion of the QTL variance explained by the markers can be calculated as follows (Hayes et al., 2006): If q_1 and q_2 are the frequencies of the two QTL alleles, the markers can be grouped into n haplotypes with the frequency p_i for the ith haplotype. This can be shown in a contingency table (see Table 4.4).

For the haplotype i in the data, the LD is calculated by $D_i = p_i(q_1) - p_i q_1$, where $p_i(q_1)$ is the proportion of i haplotypes in the dataset that have allele 1 of the QTL (observed in the data), p_i is the proportion of i haplotypes, and q_1 is the frequency of QTL allele 1. The proportion of the variance in the QTL explained by the haplotypes, when corrected for sampling effects, can be calculated by

$$r^2(h,q) = \frac{1}{q_1 q_2} \sum_{i=1}^{n} D_i^2 / p_i$$

TABLE 4.4 Contingency Table

	Haplotypes			
	1	***i***	***n***	**Total**
QTL allele 1	$p_1 q_1 - D_1$	$p_i q_1 - D_i$	$p_n q_1 - D_n$	Q_1
QTL allele 2	$p_1 q_2 + D_1$	$p_i q_2 + D_i$	$P_n q_2 + D_n$	Q_2
Total	p_1	p_i	p_n	1

Thus, the r^2 depends on the LD, the frequency of haplotype h, and the frequencies of the QTL alleles. The r^2 values can be obtained by simulating the different frequencies q_1 and q_2, the genome size, and the haplotypes. The larger the effective population size, the smaller the proportion of genetic variance that will be explained by the haplotypes.

The following mixed linear model is used to estimate the effects of the haplotypes: $y = 1'u + Xh + Za^* + e$, where y is the observed phenotypes vector, u is the scalar of the average (fixed effect), h is the random effects haplotypes vector (intervals), a^* is the polygenic effects vector (random), and e is the random residuals vector. X and Z are the incidence matrices for h and a^*. The effects of the haplotypes should preferably be treated as random because they occur in large numbers and some occur a limited number of times (haplotypes with a small number of observations should be penalized by the shrinkage effect).

The magnitude of h is equal to the number of intervals multiplied by 4 (number of possible haplotypes for each interval). The incidence matrix X contains the values 0, 1, and 2, which represent the number of alleles (of the putative QTL) or type h_i haplotypes in a diploid individual. The algebraic details of this model were previously explained by Resende (2008). The additive variances of the genes ($\sigma^2_{a^*}$) and the haplotypes (σ^2_h) can be estimated by the restricted maximum likelihood applied on phenotypic data or by the variance between haplotypes or the variance of the chromosomal segments of the QTL. The significance of the haplotypic effects is evaluated with the likelihood ratio test. For mapping, the fit of the described model relies on estimations of the variance σ^2_h and LRT to test their significance. There are no specific interests for the best linear unbiased prediction effect for h, which are emphasized and utilized in the MAS.

GWAS IN HUMANS

The first studies of human quantitative genetics that sought to understand genetic control of traits were based on estimates of heritability (h^2) by analyzing pairs of twins according to the idea of pedigree-based similarity between relatives (pedigrees: alleles that are identical by descent (IBD)). This approach considers all of the loci or genes, both the common and rare variants (low-frequency genes), that control the trait or the total h^2. The role of individual genes in genetic trait control was later studied according to the Fulker and Cardon (1994) method and by estimating the h^2 of a marker locus in the context of QTL mapping, as described by Resende (2008) and Cruz et al. (2009). This method is based on LA within full sib families and uses two molecular markers at a time.

Visscher et al. (2006) proposed an approach to estimating h^2 by simultaneously using all of the marker loci and also using segregation analysis within full sib families. This genome-wide approach is also based on IBD and takes advantage of the exact relationships. In humans, the estimated h^2 was 0.80 for height. This method includes both common and rare variants (all of the genes or the total h^2) because it also incorporates the pedigree by genotyping the parents

and estimating IBD alleles for all of the loci. Another method for studying trait control at a population level, not only within families, is GWAS. This approach is based on linkage disequilibrium in the population, but only uses one marker locus at a time via a fixed regression analysis of unrelated individuals. It thus seeks to identify significant markers that can be used in the estimation. In humans, the h^2 captured by the significant markers was only 0.10 for height.

When applied to a family with full siblings, GWAS can be described as linkage analysis. In this analysis, markers at some distance from a QTL will exhibit an association with the traits because only one generation of recombination occurred between the progenitors and the siblings. Consequently, a marker allele and a QTL allele on the same chromosome will tend to be inherited together. A more effective procedure (GWAS-SE) for capturing the majority of the heritability of a trait is a population LDA in which all markers are analyzed simultaneously (SE), a method similar to GWS. This method is based on random regression for the prediction of latent QTL effects. It uses unrelated individuals, although all of the individuals of a species are related to some degree because they share common ancestors and thus identical alleles in state (IBS), not necessarily declared as IBD due to a limitation of having recorded a really complete pedigree.

SNP markers capture this ancestral relatedness and therefore estimate genetic relationships by IBS (Powell et al., 2010; Visscher et al., 2010). Population genetics (LA, LDA, and genetic mapping) and quantitative genetics (estimation of heritability) have traditionally been used separately in human genetics. By combining these two areas, GWS allowed an h^2 measurement of 0.45 for height in humans. The remaining portion ($0.80-0.45=0.35$) was not captured due to many low-frequency variants (including loci with large effects). The genetic variance at a locus i is given by $\sigma^2_{ai} = 2\mathrm{p}_i(1-p_i)a_i^2$, ignoring dominance. Thus, a rare allele cannot explain a large part of the genetic variance, even if it has a large effect. A large sample size is needed for these loci to be captured by markers and detected. In GWS experiments, the total additive genetic variance is estimated by $\sigma^2_a = \sum_i 2\mathrm{p}_i(1-p_i)a_i^2$.

Aulchenko et al. (2007) proposed the GRAMMAR method for multiple-stage GWAS, as described below. After fitting the model $y = Xb + Zg + e$, one obtains $\hat{e} = y - X\hat{b} - Z\hat{g}$, where g is a polygenic effect vector. The model $\hat{e} = 1\mathrm{u} + Wm_i + e$ can then be fit to identify significant markers. The model $y = Xb + Wm_i + Zg + e$ is then adjusted by using only the significant SNPs. This reduces the necessary computation time when calculating thousands of markers. The effects of m are fit as the fixed effects because the SNPs do not model the familiar structure in g; in other words, they do not explain the correlation between related individuals by IBD. The method is based on the fact that the effects of the major genes are included in the conditional residual vector after fitting g under an infinitesimal polygenic model (fitting or elimination of family effects or variations between the pedigree or population structure). The final analysis returns to the complete model. This time, the polygenic effect is included to correct the data for the family structure with the relationship matrix A, as $g \sim N(0, A\sigma^2_g)$.

CAPTURING h^2 IN HUMANS WITH IMPERFECT LD BETWEEN SNPS AND CAUSAL VARIANTS

Visscher et al. (2010) addressed the GWAS results for height in humans. The h^2 captured by GWAS in traditional studies was approximately 0.10. This low value was obtained because low frequency variants (MAF<0.10) were not in perfect LD with common markers (MAF>0.10); in such instances, the r^2 is low and variants with small effects are not detected significantly by traditional GWAS, even when they are in LD with common markers. In a study by Yang et al. (2011), the captured h^2 was 0.45. This occurred because variants with small effects were still not detected significantly, but they were captured by GWS when in LD with common markers because GWS does not use significance for marker effects. The maximum possible value for r^2 is largely determined by the allelic frequencies of the two loci. The more the allelic frequencies are different, the lower the r^2 value. Thus, as most genotype SNPs are common, r^2 will be low if the variants are rare and therefore the variance (σ^2_{mi}) associated with the SNP will be substantially less than the variance (σ^2_{ai}) of the QTL (Visscher et al., 2010). The expressions $r^2 = \sigma^2_{mi}/\sigma^2_{ai}$ and $\sigma^2_{mi} = r^2\sigma^2_{ai}$ illustrate this point.

In practice, only the LD between the SNPs can be estimated. This estimate is only useful when the SNP and gene have similar allelic frequencies. A gene can be in LD with several SNPs that can collectively capture the causal variant, even if none of the SNPs are in perfect LD with the gene (Visscher et al., 2010). Thus, an SNP can fail to be significant but, together with other SNPs, can still be important when explaining genetic variance and maximizing the selective accuracy. Therefore, it is not recommended that significance tests be applied prior to GWS. Even when using tens of thousands of markers, the markers will not capture all of the genetic variance if the variants are rare and the markers common. Thus, the efficiency of GWS depends on the genetic architecture of the trait in the population. If the trait is governed by a large number of rare variants that explain a large portion of the genetic variance, GWS will be less successful. In such cases, it is recommended that residual polygenic effects be fitted in the model to capture rare variants.

In summary, the following causes for missing heritability are: (1) low-frequency variants (MAF<0.10) that are not in high LD with common markers (MAF>0.10), resulting in a low r^2; (2) a small number of markers that cause a low r^2; and (3) the use of only significant SNPs in GWAS. Simultaneous estimation is necessary because the SNPs are in LD and are thus dependent and correlated. Simultaneous regression is equivalent to phenotypic regression in all of the principal components derived from the markers because the amount of the experienced shrinkage for each estimated effect is proportional to its quadratic singular value (Campos et al., 2010). This supports the use of GWAS with the simultaneous estimation (GWAS-SE) method, according to Yang et al. (2011).

GWAS via BayesCπ and BayesDπ

The BayesCπ and BayesDπ methods (described by Habier et al., 2011 and Resende et al., 2011) are advantageous because they provide information on the genetic architecture of the quantitative trait and identify the QTL positions by modeling the frequencies of single nucleotide polymorphisms (SNP) with nonzero effects. They are advantageous over the regression analysis of single markers because they simultaneously account for all markers.

However, care needs to be taken whenever the number of markers is larger than the number of individuals genotyped and phenotyped. Gianola (2013) showed that in these cases, in the Bayesian approaches proposed, such as BayesC and BayesD, the prior is always influential, which could affect the inference to whether or not a marker is associated with the trait.

In the BayesC method, a common variance is specified for all of the loci. The BayesD method maintains specific variances for each locus. Additionally, π is treated as an unknown with a uniform *a priori* distribution (0,1), thus producing the methods BayesCπ and BayesDπ. Modeling π is very interesting in association analysis. The majority of the markers are not in LD with the genes. Thus, a group of markers associated to a trait must be identified. Differently, the BayesB method subjectively determines π. Using the indicator variable δ_i, the BayesCπ and BayesDπ methods model the additive genetic effect of individual j as

$$a_j = \sum_{i=1}^{n} \beta_i x_{ij} \delta_i,$$

where $\delta_i = (0,1)$. The distribution of $\delta = (\delta_{1\ldots}\delta_n)$ is binomial with a probability π. This mixture model is more parsimonious than the BayesB method. According to the model hierarchy, a distribution must be postulated for π and must be a beta distribution, which when appropriately specified becomes a uniform distribution (0,1) (Legarra et al., 2011).

The quantities for x_{ij} are elements of the codominant marker genotype vector and are generally coded as 0, 1, or 2, depending on the number of copies of one of the alleles at marker locus i, and β_i is defined as an element of the vector of the regression coefficients, which includes the marker effects on a phenotypic trait y by means of the LD with the genes that control the trait.

REFERENCES

Anderson, L., Georges, M., 2004. Domestic animal genomes: deciphering the genetics of complex traits. Nature Reviews Genetics 5 (3), 202–212.

Aulchenko, Y.S., Konning, D., Haley, C., 2007. Grammar: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. Genetics, Austin 177, 577–585.

Campos, G.de los, Gianola, D., Allison, D.B., 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Reviews Genetics, London 11, 880–886.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971.

Cruz, C.D., Good God, P.I.V., Bhering, L.L., 2009. Mapeamento de QTLs em populações exogâmicas. In: Borém, A., Caixeta, E.T. (Orgs.), Marcadores Moleculares, vol. 1, second ed. Folha de Viçosa, Viçosa, MG, pp. 443–481.

Darvasi, A., Soller, M., 1997. A simple method to calculate resolving power and confidence interval of QTL map location. Behavior Genetics 27, 125–132.

Ewing, B., Green, P., 2000. Analysis of expressed sequence tags indicates 35,000 human genes. Nature Genetics 25, 232–234.

Fernando, R.L., Nettleton, D., Southey, B.R., Dekkers, J.C.M., Rothschild, M.F., Soller, M., 2004. Controlling the proportion of false positives in multiple dependent tests. Genetics 166, 611–619.

Fulker, D.F., Cardon, L.R., 1994. A sib-pair approach to interval mapping of quantitative trait loci. American Journal of Human Genetics 54, 1092–1103.

Gianola, D., 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics: Early Online, published on May 1, 2013 as 10.1534/genetics.113.151753.

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12, 186.

Hartl, L.D., Jones, E.W., 2002. Essential Genetics: a Genomics Perspective. Jones & Bartlet, Sudbury.

Hayes, B.J., 2008. Course on QTL Mapping, MAS and Genomic Selection. Iowa State University, Ames.

Hayes, B.J., Chamberlain, A.J., Goddard, M.E., 2006. Use of markers in linkage disequilibrium with QTL in breeding programs. In: World Congress of Genetics Applied to Livestock Production, 8, 2006. Proceedings. Belo Horizonte: Ed. da UFMG. 1 CD-ROM.

Hayes, B.J., Chamberlain, A.J., McPartlan, H., Macleod, I., Sethuraman, L., Goddard, M.E., 2007. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. Genetical Research 89, 215–220.

Hayes, B.J., Visscher, P.E., McPartlan, H., Goddard, M.E., 2003. A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. Genome Research 13, 635–643.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226–231.

Hoggart, C.J., Whittaker, J.C., De Iorio, M., Balding, D.J., 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genetics 4 (7), e1000130.

Kruglyak, L., 1999. Prospect for whole genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22, 139–144.

Lande, R., Thompson, R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124, 743–756.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011. Improved Lasso for genomic selection. Genetics Research 93 (1), 77–87.

Lewontin, R.C., 1964. The interaction of selection and linkage. II. Optimal models. Genetics 50, 757–782.

Luo, Z.W., 1998. Linkage disequilibrium in a two-locus model. Heredity 80, 198–208.

Macleod, I.M., Hayes, B.J., Savin, K., Chamberlain, A.J., McPartlan, H., Goddard, M.E., 2010. Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. Journal of Animal Breeding and Genetics 127 (2), 133–142. <http://www.ncbi.nlm.nih.gov/pubmed/20433522>.

McRae, A.F., McEvan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M., Slate, J., 2002. Linkage disequilibrium in domestic sheep. Genetics 160, 1113–1122.

Meuwissen, T.H.E., Goddard, M.E., 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155, 421–430.

Meuwissen, T.H.E., Goddard, M.E., 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genetics Selection Evolution 36, 261–279.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I., Goddard, M.E., 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161, 373–379.

Perez-Enciso, M., Toro, M.A., Tenenhaus, M., Gianola, D., 2003. Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. Genetics 164, 1597–1606.

Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11, 800–805.

Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69, 1–14.

Resende, M.D.V., 2008. Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais. Embrapa Florestas, Colombo. 330p.

Resende, M.D.V., Silva, F.F., Viana, J.M.S., Peternelli, L.A., Resende Junior, M.F.R., et al., 2011. Métodos Estatísticos na Seleção Genômica Ampla. Embrapa Florestas, Colombo. 106p.

Sved, J.A., 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2, 125–141.

Tenesa, A., Navarro, T., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., et al., 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Research 17, 520–526.

Visscher, P.M., Medland, S.E., Ferreira, M.A.R., Morley, K.I., Zhu, G., et al., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genetics 2 (3), e41.

Visscher, P.M., Yang, J., Goddard, M.E., 2010. A commentary on Common SNPs explain a large proportion of the heritability for human height by Yang et al. (2010). Twin Research and Human Genetics 13 (6), 517–524.

Weller, J.I., 2001. Quantitative Trait Loci Analysis in Animals. CABI Publishing, London. 287p.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. Gcta: a tool for genome-wide complex trait analysis. American Journal of Human Genetics 88, 76–82.

Zenger, K.R., Khatkar, M.S., Cavanagh, J.A., Hawken, R.J., Raadsma, H.W., 2007. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian global population variability, including impact of selection. Animal Genetics 38, 7–14.

Zhao, H., Nuttleton, D., Soller, M., Dekkers, J.C.M., 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetical Research 80, 77–97.