BMC Genetics

METHODOLOGY ARTICLE Open Access

On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits

Tao Wang

Abstract

Background: In genetic association study of quantitative traits using F_∞ models, how to code the marker genotypes and interpret the model parameters appropriately is important for constructing hypothesis tests and making statistical inferences. Currently, the coding of marker genotypes in building F_∞ models has mainly focused on the biallelic case. A thorough work on the coding of marker genotypes and interpretation of model parameters for F_∞ models is needed especially for genetic markers with multiple alleles.

Results: In this study, we will formulate F_∞ genetic models under various regression model frameworks and introduce three genotype coding schemes for genetic markers with multiple alleles. Starting from an allele-based modeling strategy, we first describe a regression framework to model the expected genotypic values at given markers. Then, as extension from the biallelic case, we introduce three coding schemes for constructing fully parameterized one-locus F_∞ models and discuss the relationships between the model parameters and the expected genotypic values. Next, under a simplified modeling framework for the expected genotypic values, we consider several reduced one-locus F_∞ models from the three coding schemes on the estimability and interpretation of their model parameters. Finally, we explore some extensions of the one-locus F_∞ models to two loci. Several fully parameterized as well as reduced two-locus F_∞ models are addressed.

Conclusions: The genotype coding schemes provide different ways to construct F_∞ models for association testing of multi-allele genetic markers with quantitative traits. Which coding scheme should be applied depends on how convenient it can provide the statistical inferences on the parameters of our research interests. Based on these F_∞ models, the standard regression model fitting tools can be used to estimate and test for various genetic effects through statistical contrasts with the adjustment for environmental factors.

Background

Genetic markers with multiple alleles are common phenomena in genetic studies. It is well known that the ABO blood types in human are determined by three alleles at a genetic locus on chromosome 9. Molecular markers such as microsatellites often have multiple alleles. The major histocompatibility complex (MHC), a highly polymorphic genome region that resides on the human chromosome 6, encompasses multiple genes that encode for many human leukocyte antigens (HLA) and play an important role in regulation of the immune responses. Depending on the resolution level of allele typing, each of the HLA-A, B, C, DR, DQ and DP gene loci could contain tens to hundreds of allele types. In addition, in the haplotype analysis of single-nucleotide polymorphisms (SNPs), various haplotypes from a set of SNPs can also be treated as different alleles from a 'super' marker locus that consists of the set of SNPs.

Presently, there are mainly three types of genetic models that are commonly used in the genetic analysis of quantitative traits. One is Fisher's analysis of variance (ANOVA) models that focus on a decomposition of the genotypic variance into genetic variance components contributed by various genetic effects at quantitative trait loci (QTL) [1-6]. Another is the F_∞ models that concentrate on direct statistical modeling of the expected genotypic values at target genetic markers or

Correspondence: taowang@mcw.edu
Division of Biostatistics, Institute for Health and Society, Medical College of Wisconsin, Milwaukee, WI 53226, USA

QTL and the association testing of various genetic effects. The other one is the so-called functional genetic models that emphasize on modeling the functional effects of genes [7]. Both Fisher's and F_∞ models can be referred to as statistical models, while the functional genetic models have fundamentally different objectives and estimation methods from the statistical models. A considerable amount of discussion has been made about the distinction between these different types of genetic models [8-11].

The F_∞ models have been widely used in genetic association studies of quantitative traits. In building F_∞ models, how to code genotypes at a marker (or QTL) and interpret the model parameters are fundamental issues for constructing appropriate testing hypotheses and making correct statistical inferences. While the Fisher's ANOVA models can be directly applicable to genetic markers with multiple alleles, the F_∞ models by contrast have been mainly discussed in the biallelic case [1,9,12]. For haplotype analysis, Zaykin *et al*. in [13] proposed a simple coding which included only the additive effects of haplotypes but ignored their interactions. More recently, Yang *et al*. in [11] explored an extension of the biallelic F_∞ models to multi-allele models with a focus on the definition of various genetic effects and their relationships with the average genetic effects defined in the Fisher's models. A thorough work on coding of marker genotypes and interpretation of model parameters for F_∞ models has not been done in the past especially for genetic markers with multiple alleles.

In general, there are two different strategies in coding the marker or QTL genotypes. One is to treat each marker or QTL as a potential risk factor with its genotypes as the risk units. Then, similar to the strategy in handling categorical covariates in classical regression models, at each locus we can create one dummy variable per genotype and then include all but one (as the reference) of these dummy variables into a model. But this genotype coding is often limited by the available sample sizes especially when the number of alleles at the marker locus is large. Alternatively, as alleles are often supposed to be the basic genetic risk units that may contribute to disease phenotypes in genetic studies, we may want to treat alleles at each marker or QTL as the risk units and examine the effects of alleles. However, genetic data has some specialty that needs to be taken into account in order to build the allele-based models. In the genome of diploid species such as human being, alleles normally appear in pairs to form a genotype at each marker locus or QTL with one from the father and one from the mother, except for the sex chromosomes in males. That is, at each locus we have two within-locus risk factors that reside on a homologous pair of chromosomes. Unlike the classical two-way ANOVA model in which the two risk factors own different risk units, the paternal and maternal risk factors at a locus often share the same set of alleles. Besides, the parental origins (i.e., the phase) of the two alleles at each locus are quite often unknown. These features could sometimes complicate the allele-based coding of marker genotypes and generate confusion in interpretation of the model parameters.

In this study, we introduce three allele-based coding schemes for building F_∞ models, namely allele, F_∞ and allele-count codings. First, we formulate F_∞ models under a general regression framework to model the expected genotypic values at given markers or QTL. Then, under a standard ANOVA model setting, we present several fully parameterized one-locus models using the three allele-based coding schemes. Some potential collinearity relationships among the coding variables of the marker genotypes are clarified. Strategies to avoid the redundant model parameters are also proposed. After that, we examine the definition of model parameters under a reduced one-locus model framework. The impact of a linear relationship among the coding variables of marker genotypes on the estimability of the model parameters is fully explored based on the linear model theory. Finally, we consider extension of the one-locus models to two-locus situation. Several fully parameterized as well as reduced two-locus models are addressed. A focus of this study is to establish the relationships between the model parameters and the expected genotypic values at given marker loci or QTL for various F_∞ models from these three coding schemes under various different model frameworks, and explain how to estimate and test for various genetic effects through statistical contrasts. Relationships among different coding schemes and models are also illustrated through simulation.

Results

Fully parameterized one-locus models

In genetic studies, a quantitative trait Y is typically considered as a combination of a genetic component G and an environmental component E with perhaps the genetic by environmental interactions $G \times E$, where G is the true genotypic value from a joint (unobservable) contribution of all the genetic factors to the quantitative trait Y. In practice, given a random sample of N individuals from a study population, let g_i be the observed genotypes at certain target marker loci or QTL and z_i be a vector of some environmental covariates that may contribute to the variation of the quantitative trait for individuals $i = 1, \ldots, N$. By ignoring the genetic by environmental interactions and assuming that the genotypic value G and environmental component E do not depend on the environmental covariates z_i and g_i, respectively, then the observed quantitative trait y_i of an

individual i can be expressed through a regression model as

$$y_i = G(g_i) + z_i\beta + e_i, i = 1, \ldots, N \quad (1)$$

where $G(g_i) = E(G|g_i)$ is the expected genotypic value of G given the marker (or QTL) genotypes g_i, β denotes the effects of the environmental covariates, and e_i is the residual error of the model with $E(e_i) = 0$. Similar to introducing dummy variables for the covariates z_i which allow us to assess various environmental effects β in the model, it is convenient to further represent $G(g_i)$ as $G(g_i) = x(g_i)\alpha$ so that we can fit the regression model and assess the genetic effects α of the markers or QTL, where $x(g_i)$ is a coding function of the marker genotypes. When the marker locus is not associated with the phenotype, then $G(g_i) = E(G)$ is a constant which does not depend on g_i. In the rest of the paper, we will focus on the interpretation of the marker effects α in terms of the expected genotypic values $G(g) = E(G|g)$ according to different coding schemes. When certain genetic by environmental interactions are included in the model, the interpretation of α could be modified accordingly. It has to be pointed out that QTL are generally assumed to be unknown genomic regions that may contribute to the variation of the quantitative traits with their genotypes unobserved. But the results (i.e., the coding schemes and the relationships between the model parameters and the expected genotypic values) are held for QTL as well, although the expected genotypic values at a target QTL can no longer be directly estimated via fitting the regression models.

Now, consider one target marker locus with multiple alleles A_1, ..., A_m, $m \geq 2$. In general, there are m possible homozygous genotypes A_jA_j, $j = 1$ ..., m, and $m(m-1)/2$ possible heterozygous genotypes A_jA_k, $j \neq k$. Let $G_{jk} = E(G|g = A_jA_k)$ be the expected genotypic values, given the marker genotypes A_jA_k in a study population. Without knowing the parental origins of the alleles, we assume as usual that the parental origin of the alleles does not make a difference (i.e., no imprinting). We have then $G_{jk} = G_{kj}$ for $j, k = 1, ..., m$, and there are totally $m(m+1)/2$ possible distinctive expected genotypic values G_{jk}, $j, k = 1, ..., m$, which could be estimated through the means in the genotypic subgroups after adjustment for the environmental covariates. Here we assume no missing genotypes for the sampled individuals, and the random sample has its individuals carrying all possible genotypes. How to handle missing genotypes will be discussed in the discussion. To fully re-parameterize these expected genotypic values through a linear model, we then need totally $m(m+1)/2$ parameters including the intercept in the model. By treating the paternal and maternal alleles as two independent risk factors and following the classical two-way ANOVA notation, we can represent the genotypic values G_{jk} as

$$G_{jk} = \mu^* + \alpha_j^* + \alpha_k^* + \delta_{jk}^*, j, k = 1, \ldots, m \quad (2)$$

where α_j^* and δ_{jk}^* are the realized (but unobservable) additive effects of allele A_j and the allelic interaction between the two alleles A_j and A_k, respectively. The above model is different from the classical two-way ANOVA model in that here both the paternal and the maternal risk factors share the same set of alleles A_1, ..., A_m. As usual, with the unknown paternal origins of alleles at the locus, we assume the paternal and maternal alleles have the same genetic effect. More precisely, the paternal allele A_j and maternal allele A_j have the same additive allelic effects α_j^* for $j = 1, ..., m$. Besides, the allelic interaction between a paternal allele A_j and a maternal allele A_k is the same as that between the paternal allele A_k and the maternal allele A_j; i.e., $\delta_{jk}^* = \delta_{kj}^*$, for $j, k = 1, ..., m$. Still, with m additive allelic effects and $m(m+1)/2$ allelic interactions plus the intercept, it is clear that model (2) is over-parameterized on modeling the $m(m+1)/2$ expected genotypic values G_{jk} for $j, k = 1, ..., m$. As a result, the parameters μ^*, α_j^* and δ_{jk}^* in model (2) are not all estimable in terms of the expected genotypic values G_{jk} (see [14,15]).

In order to avoid the inestimability issue, one way is to add constraints on the model parameters. However, those constraints, together with the symmetry property of δ_{jk}^*, could make it difficult to fit the model using the standard software package such as SAS. Alternatively, we consider dropping certain redundant parameters in the model. Similar to the biallelic case [10], let us first introduce the following indicator variables to describe the transmission of alleles from parents to their offspring

$$z_{1j} = \begin{cases} 1, & \text{inherited } A_j \text{ on paternal gamete,} \\ 0, & \text{inherited other alleles on paternal gamete} \end{cases}$$

and

$$z_{2j} = \begin{cases} 1, & \text{inherited } A_j \text{ on maternal gamete,} \\ 0, & \text{inherited other alleles on maternal gamete} \end{cases}$$

for each allele type A_j, $j = 1, ..., m$. Then we define the following coding variables of the marker genotypes

$$w_j(g) = z_{1j} + z_{2j} = \begin{cases} 2, & \text{if } g = A_jA_j \\ 1, & \text{if } g = A_jA_j^c \\ 0, & \text{if } g = A_j^cA_j^c \end{cases}$$

$$v_{jk}(g) = z_{1j}z_{2k} = \begin{cases} 1, & \text{if } g = A_jA_k \\ 0, & \text{otherwise} \end{cases}$$

for $j, k = 1, ..., m$, where A_j^c denotes any other allele type except A_j. Note that z_{1j}, $z_{2j}$are not observable

because we do not know exactly which allele is inherited from paternal or maternal gamete for the sampled individuals without their parental information. But this unknown phase problem does not affect the definitions of w_j, v_{jk} since w_j only counts the number of allele A_j in the genotypes and the value of v_{jk} is 1 when the genotype is A_jA_k and 0 otherwise regardless of where the two alleles come from. We refer to the above coding of marker genotypes as an allele coding scheme. Model (2) can then be re-written in a linear model form as

$$G(g_i) = \mu^* + \sum_{j=1}^{m} \alpha_j^* w_j(g_i) + \sum_{j=1}^{m}\sum_{k=j}^{m} \delta_{jk}^* v_{jk}(g_i) \quad (3)$$

for $i = 1, ..., N$. As each individual always carries two alleles at a marker locus with one from the father and the other from the mother, we have $\sum_{j=1}^{m} z_{1j}(g_i) = \sum_{k=1}^{m} z_{2k}(g_i) = 1$, for any $i = 1, ..., N$. Therefore, given a particular j, $w_{jk} = 2 - \sum_{k\neq j} w_k$, which is a linear combination of the rest of $\{w_k, k \neq j\}$. For v_{jk}, we also have $\sum_{j=1}^{m} v_{jk} = z_{2k}$, or $v_{jk} = w_k/2 - \sum_{l\neq j} v_{lk}$. Hence, each of the v_{jk}, $k = 1, ..., m$, is also a linear combination of the coding variables $\{w_k, k \neq j\}$ and $\{v_{lk}, l, k \neq j\}$. To avoid the redundancy of parameters due to these collinearity relationships among the coding variables in model (3), without losing generality, we consider dropping w_m and $\{v_{km}, k = 1, ..., m\}$ in (3). Then

$$G(g_i) = \mu + \sum_{j=1}^{m-1} \alpha_j w_j(g_i) + \sum_{j=1}^{m-1}\sum_{k=j}^{m-1} \delta_{jk} v_{jk}(g_i) \quad (4)$$

for $i = 1, ..., N$. Model (4) now provides a full re-parameterization of the $m(m + 1)/2$ expected genotypic values G_{jk} for $j, k = 1, ..., m$ with its parameters α_j can be referred to as the additive allelic effects and δ_{jk} the allelic interactions with respect to the reference allele A_m. Given a random sample, we can then incorporate model (4) into (1) and fit the regression model (1) using the standard least-square approach. In terms of the expected genotypic values, it is easy to show that $\mu = G_{mm}$, $\alpha_j = G_{jm} - G_{mm}$ and $\delta_{jk} = (G_{jk} - G_{km}) - (G_{jm} - G_{mm})$, for $j = 1, ..., m - 1$ and $k = j, ..., m - 1$. Therefore, the additive allelic effect α_j can be interpreted as the substitution effect of replacing allele A_m by A_j when paired with another allele A_m to form the genotypes. Meanwhile, the allelic interaction δ_{jk} is the difference between the substitution effect of replacing allele A_m by A_j (or A_k) when paired with allele A_k (or A_j) and that when paired with allele A_m. Or, in other words, δ_{jk} is the difference between the substitution effects of replacing allele A_m by A_j (or A_k) with paired alleles A_k (or A_j) and A_m. Note that dropping w_j and $\{v_{kj}, k = 1, ..., m\}$ for a particular $j \neq m$ instead of w_m and $\{v_{km}, k = 1, ..., m\}$ can lead to similar interpretations of the model parameters with A_j being the reference allele. Using model (4), we can also estimate and test for various other genetic effects. For example, the so-called functional 'additive effects' $a_{jk}^* = (G_{jj} - G_{kk})/2$ and the 'dominance effects' $d_{jk}^* = G_{jk} - (G_{jj} + G_{kk})/2$, $j \neq k$ defined in [11] can be expressed as $a_{jk}^* = (\alpha_j - \alpha_k) + (\delta_{jj} - \delta_{kk})/2$ and $d_{jk}^* = \delta_{jk} - (\delta_{jj} + \delta_{kk})/2 - 2\mu$, $j \neq k$, respectively, in terms of the above model parameters. So we can estimate a_{jk}^*, d_{jk}^* using the fitted model parameters or test for the hypothesis of $H_0 : a_{jk}^* = 0$ or $H_0 : d_{jk}^* = 0$ through the general linear contrasts [15] using the standard software such as SAS. To test whether a particular allele A_j has an overall effect, the null hypothesis is $H_0 : \alpha_j = \delta_{jk} = 0$ for $k = 1, ..., m - 1$, which can be performed through either a general linear contrast (or likelihood ratio test) with the degrees of freedom being m for the test statistic. The association test for overall effects of the locus corresponds to the null hypothesis of $H_0 : \alpha_j = \delta_{jk} = 0$ for any $j, k = 1, ..., m - 1$, which has its degrees of freedom being $m(m + 1)/2 - 1$ for the test statistic. Currently, the so-called F_∞ model has been widely used in genetic association studies. In the simple biallelic case with two alleles A and α, an F_∞ model gives [16-19].

$$G_{AA} = \tau + a, \quad G_{Aa} = \tau + d, \quad G_{aa} = \tau - a$$

where $G_{AA} = E(G|AA)$, $G_{Aa} = E(G|Aa)$ and $G_{aa} = E(G|aa)$ are the three possible expected genotypic values at the marker. The parameters a, d are often referred to as the additive and dominance effects of the allele A over a, and in terms of the expected genotypic values we have $a = (G_{AA} - G_{aa})/2$ and $d = G_{Aa} - (G_{AA} + G_{aa})/2$. This F_∞ model can also be written in a linear model form as [10]

$$G(g_i) = \tau + af(g_i) + dh(g_i), i = 1, \ldots, N$$

where f, h are two coding variables of the marker genotypes that are defined as

$$f(g) = \begin{cases} 1, \text{ if } g = AA \\ 0, \text{ if } g = Aa \\ -1, \text{ if } g = aa \end{cases}$$

$$h(g) = \begin{cases} 1, & \text{if } g = Aa \\ 0, \text{ otherwise} \end{cases}$$

We refer to the above coding of the marker genotypes as the F_∞ coding. As a straightforward extension of the F_∞ coding scheme to multiple alleles, we can define the following coding variables

$$f_j(g) = \begin{cases} 1, \text{ if } g = A_jA_j \\ 0, \text{ if } g = A_jA_j^c \\ -1, \text{ if } g = A_j^cA_j^c \end{cases}$$

$$h_j(g) = \begin{cases} 1, \text{ if } g = A_jA_j^c \\ 0, \text{ otherwise} \end{cases}$$

for each $j = 1, ..., m$. It is easy to see that f_j, h_j and the previous w_j, v_{jk}, $j, k = 1, ..., m$ have the relationships: $f_j(g) = w_j(g) - 1$, $h_j(g) = w_j(g) - 2v_{jj}(g)$, and $v_{jk}(g) = h_j(g)h_k(g)$ as $j \neq k$. Thus, for the same reason to avoid collinearity, we can exclude some redundant coding variables and write a fully parameterized one-locus model using the F_∞ coding as

$$G(g_i) = \tau + \sum_{j=1}^{m-1} a_j f_j(g_i) + \sum_{j=1}^{m-1} d_{jj} h_j(g_i) + \sum_{j=1}^{m-1}\sum_{k=j+1}^{m-1} d_{jk} h_j(g_i) h_k(g_i) \quad (5)$$

for $i = 1, ..., N$. By having model (5) equivalent to (4), we can first build the relationships between the two model parameters and then establish the relationships between the parameters of model (5) and the expected genotypic values as following

$$\begin{cases} \tau = \mu + \sum_{j=1}^{m} \left(\alpha_j + \frac{\delta_{jj}}{2}\right) \\ \quad = G_{mm} + \frac{1}{2}\sum_{j=1}^{m-1}\left(G_{jj} - G_{mm}\right) \\ a_j = \alpha_j + \frac{\delta_{jj}}{2} = \frac{G_{jj} - G_{mm}}{2}, j = 1, \ldots, m-1 \\ d_{jj} = -\frac{\delta_{jj}}{2} = G_{jm} - \frac{G_{jj} + G_{mm}}{2}, j = 1, \ldots, m-1 \\ d_{jk} = \delta_{jk} = \left(G_{jk} - G_{jm}\right) - \left(G_{km} - G_{mm}\right), j \neq k \end{cases}$$

Therefore, a_j can be interpreted as a half of the difference between the two expected homozygous genotypic values G_{jj} and G_{mm}, which is the same as the additive effect a^*_{jm} defined in [11]. Besides, d_{jj} is the difference between the expected heterozygous genotypic value G_{jm} and the averaged expected homozygous genotypic value $(G_{jj} + G_{mm})/2$, which is the same as the dominance effect d^*_{jm} defined in [11]. It is interesting to see that d_{jk}, $j \neq k$, has the same interpretation as δ_{jk} in model (4), which is the difference between the substitution effects of replacing allele A_m by A_j when paired with alleles A_k and A_m. Note that d_{jj} can also be interpreted as the allelic interaction - the difference between the substitution effects of replacing allele A_j by A_m when paired with another A_j and A_m. In addition, based on model (5), the additive effects a^*_{jk} and the dominance effects d^*_{jk} proposed in [11] have the relationship with the model parameters: $a^*_{jk} = a_j - a_k$, $d^*_{jk} = d_{jk} + (d_{jj} + d_{kk})$, $j \neq k$. The overall effect of a particular allele A_j can be tested through the composite hypothesis of $H_0 : a_j = d_{jk} = 0$ for $k = 1, ..., m - 1$, and the overall effects of the locus can be tested via the null hypothesis of $H_0 : a_j = d_{jk} = 0$ for any $j, k = 1, ..., m - 1$.

In addition to the allele and F_∞ codings, another way of coding the marker genotypes which occasionally appears in practice is to count the number of alleles in marker genotypes for each specific allele A_j. As each individual can have 0, 1 or 2 copies of an allele A_j, by taking the genotypic group with 0 copy of allele A_j as the baseline, we can introduce the following two indicator (or dummy) variables for the genotypic groups with 1 and 2 copies of the allele A_j, respectively.

$$h_{1j}(g) = \begin{cases} 1, & \text{if } g = A_j A_j^c \\ 0, & \text{otherwise} \end{cases}$$

$$h_{2j}(g) = \begin{cases} 1, & \text{if } g = A_j A_j \\ 0, & \text{otherwise} \end{cases}$$

for each $j = 1, ..., m - 1$. These coding variables of marker genotypes have relationships $h_{1j}(g) = h_j(g) = w_j(g) - 2v_{jj}(g)$ and $h_{2j}(g) = v_{jj}(g)$ with previous ones. We refer to this coding of marker genotypes as the allele-count coding. Similar to models (4) and (5), by excluding some redundant coding variables, the allele-count coding leads to another fully parameterized one-locus model as

$$G(g_i) = \pi_0 + \sum_{j=1}^{m-1} \pi_j h_{1j}(g_i) + \sum_{j=1}^{m-1} \eta_{jj} h_{2j}(g_i) + \sum_{j=1}^{m-1}\sum_{k=j+1}^{m-1} \eta_{jk} h_{1j}(g_i) h_{1k}(g_i) \quad (6)$$

for $i = 1, ..., N$. Similarly, by having model (6) equivalent to (4), we can establish the following relationships

$$\begin{cases} \pi_0 = \mu = G_{mm} \\ \pi_j = \alpha_j = G_{jm} - G_{mm}, j = 1, \ldots, m-1 \\ \eta_{jj} = 2\alpha_j + \delta_{jj} = G_{jj} - G_{mm}, j = 1, \ldots, m-1 \\ \eta_{jk} = \delta_{jk} = \left(G_{jk} - G_{jm}\right) - \left(G_{km} - G_{mm}\right), j \neq k \end{cases}$$

Therefore, π_j in model (6) can still be interpreted as the substitution effect of replacing allele A_m by A_j when paired with allele A_m, or the difference between the genotypic values of the genotype group A_jA_m with one copy of A_j versus the genotype group A_mA_m (baseline). η_{jj} is the difference between the expected genotypic value G_{jj} in the homozygous genotypic group A_jA_j with two copies of A_j and G_{mm} in the baseline group A_mA_m. Besides, η_{jk} in model (6) has the same interpretation as δ_{jk} (or d_{jk}) before. From model (6), the general additive effects $a^*_{jk} = (\eta_{jj} - \eta_{kk})/2$ and the dominance effects $d^*_{jk} = \eta_{jk} - (\eta_{jj} + \eta_{kk})/2 - 2\pi_0$, $j \neq k$, which can be tested either separately or jointly. The overall effect of a particular allele A_j can be tested through the composite hypothesis of $H_0 : \pi_j = \eta_{jk} = 0$ for $k = 1, ..., m - 1$. The overall effects of the locus can also be tested via the null hypothesis of $H_0 : \pi_j = \eta_{jk} = 0$ for any $j, k = 1, ..., m - 1$.

Each of the three models (4), (5) and (6) provides a full re-parameterization of the $m(m + 1)/2$ expected genotypic values under the same model framework (3). The relationships between their model parameters and the expected genotypic values are summarized in Table 1. It is interesting to see from Table 1 that the null hypothesis of $\alpha_j = \delta_{jj} = 0$ is equivalent to either $a_j = d_{jj} = 0$ or $\pi_j = \eta_{jj} = 0$, which implies $G_{jj} = G_{jm} = G_{mm}$. So the three models above should provide the same test statistics for testing $\alpha_j = \delta_{jj} = 0$, $a_j = d_{jj} = 0$ or $\pi_j = \eta_{jj} = 0$.

For a biallelic locus with alleles A (or A_1) and a (or A_2), we have $m = 2$ with three possible genotypic values $G_{AA} = E(G|AA)$, $G_{Aa} = E(G|Aa)$ and $G_{aa} = E(G|aa)$. If we adopt the allele coding, then $w_2(g) = 2 - w_1(g)$, $\nu_{12}(g) = w_1(g) - \nu_{11}(g)$, and $\nu_{22}(g) = 1 - w_1(g) + \nu_{11}(g)$. For the F_∞ coding, we have $f_2(g) = -f_1(g)$ and $h_2(g) = h_1(g)$. So we can further drop d_2 in model (5). For the allele-count coding, we have $h_{12}(g) = h_{11}(g)$ and $h_{22}(g) = 1 - h_{11}(g) - h_{21}(g)$. The interpretation of model parameters for these three biallelic QTL models are summarized in Table 2, which is a special case of Table 1.

For a locus with three alleles A_1, A_2 (i.e., $m = 3$), we have six possibly distinctive expected genotypic values G_{11}, G_{22}, G_{33}, G_{12}, G_{13} and G_{23}. Each of the three fully parameterized models (4), (5) and (6) can provide a full re-parameterization of the six expected genotypic values. In a matrix form, from the allele coding model (4), we have

$$\begin{bmatrix} G_{11} \\ G_{22} \\ G_{33} \\ G_{12} \\ G_{13} \\ G_{23} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \delta_{11} \\ \delta_{22} \\ \delta_{12} \end{bmatrix}$$

Table 2 Parameterization of one-locus models (4), (5), (6) when $m = 2$.

Codings	Models	Relationships
Allele	$G_{AA} = \mu + 2\alpha_1 + \delta_{11}$	$\mu = G_{aa}$
	$G_{Aa} = \mu + \alpha_1$	$\alpha_1 = G_{Aa} - G_{aa}$
	$G_{aa} = \mu$	$\delta_{11} = G_{AA} + G_{aa} - 2G_{Aa}$
F_∞	$G_{AA} = \tau + a_1$	$\tau = \frac{G_{AA}+G_{aa}}{2}$
	$G_{Aa} = \tau + d_{11}$	$a_1 = \frac{G_{AA}-G_{aa}}{2}$
	$G_{aa} = \tau - a_1$	$d_{11} = G_{Aa} - \frac{G_{AA}+G_{aa}}{2}$
Allele-count	$G_{AA} = \pi_0 + \eta_{11}$	$\pi_0 = G_{aa}$
	$G_{Aa} = \pi_0 + \pi_1$	$\pi_1 = G_{Aa} - G_{aa}$
	$G_{aa} = \pi_0$	$\eta_{11} = G_{AA} - G_{aa}$

From the F_∞ coding model (5), we have

$$\begin{bmatrix} G_{11} \\ G_{22} \\ G_{33} \\ G_{12} \\ G_{13} \\ G_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau \\ a_1 \\ a_2 \\ d_{11} \\ d_{22} \\ d_{12} \end{bmatrix}$$

And the allele-count coding model (6) gives

$$\begin{bmatrix} G_{11} \\ G_{22} \\ G_{33} \\ G_{12} \\ G_{13} \\ G_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \eta_{11} \\ \eta_{22} \\ \eta_{12} \end{bmatrix}$$

By multiplying the design matrices on the left side of the equations, we can show that the model parameters

Table 1 Parameterization of fully parameterized one-locus models (4), (5), (6).

Codings	Relationships
Allele	$\mu = G_{mm}, \alpha_j = G_{jm} - G_{mm}$
	$\delta_{jj} = G_{jj} + G_{mm} - 2G_{jm}, j = 1, \ldots, m-1$
	$\delta_{jk} = (G_{jk} - G_{jm}) - (G_{km} - G_{mm}), j, k = 1, \ldots, m-1; j < k$
F_∞	$\tau = G_{mm} + \frac{1}{2}\sum_{j=1}^{m-1}(G_{jj} - G_{mm})$
	$a_j = \frac{G_{jj}-G_{mm}}{2}, d_{jj} = G_{jm} - \frac{G_{jj}+G_{mm}}{2}, j = 1, \ldots, m-1$
	$d_{jk} = (G_{jk} - G_{jm}) - (G_{km} - G_{mm}), j, k = 1, \ldots, m-1; j < k$
Allele-count	$\pi_0 = G_{mm}, \pi_j = G_{jm} - G_{mm}$
	$\eta_{jj} = G_{jj} - G_{mm}, j = 1, \ldots, m-1$
	$\eta_{jk} = (G_{jk} - G_{jm}) - (G_{km} - G_{mm}), j, k = 1, \ldots, m-1; j < k$

Table 3 Parameterization of one-locus models (4), (5), (6) when $m = 3$.

Codings	Relationships
Allele	$\mu = G_{33}$
	$\alpha_1 = G_{13} - G_{33}, \alpha_2 = G_{23} - G_{33}$
	$\delta_{11} = G_{11} + G_{33} - 2G_{13}$
	$\delta_{22} = G_{22} + G_{33} - 2G_{23}$
	$\delta_{12} = G_{12} + G_{33} - G_{13} - G_{23}$
F_∞	$\tau = \frac{G_{11}+G_{22}}{2}$
	$a_1 = \frac{G_{11}-G_{33}}{2}, a_2 = \frac{G_{22}-G_{33}}{2}$
	$d_{11} = G_{13} - \frac{G_{11}+G_{33}}{2}$
	$d_{22} = G_{23} - \frac{G_{22}+G_{33}}{2}$
	$d_{12} = G_{12} + G_{33} - G_{13} - G_{23}$
Allele-count	$\pi_0 = G_{33}$
	$\pi_1 = G_{13} - G_{33}, \pi_2 = G_{23} - G_{33}$
	$\eta_{11} = G_{11} - G_{33}$
	$\eta_{22} = G_{22} - G_{33}$
	$\eta_{12} = G_{12} + G_{33} - G_{13} - G_{23}$

and the expected genotypic values have the relationships as summarized in Table 3, which is consistent with that in Table 1.

Reduced one-locus models

Due to limited available sample sizes in practice, it may not always be feasible to use the fully parameterized models. Quite often, one may want to check the main effects of alleles first before including all possible allelic interactions. Here we consider the case of including possible interactions between A_j and itself for the homozygous genotypes $A_j A_j$, $j = 1, ..., m$, but ignore other interactions between different alleles A_j and A_k ($j \neq k$). Then we obtain a reduced case of model (2) as below

$$G_{jk} = \mu^* + \alpha_j^* + \alpha_k^* + \delta_j^* 1_{\{j=k\}} \quad (7)$$

for $j, k = 1, ..., m$. Similarly, using the allele coding, we can present this model in a linear model form as

$$G(g_i) = \mu^* + \sum_{j=1}^{m} \alpha_j^* w_j(g_i) + \sum_{j=1}^{m} \delta_j^* v_j(g_i) \quad (8)$$

for $i = 1, ..., N$, where $v_j(g) = v_{jj}(g)$ for $j = 1, ..., m$, with $v_{jj}(g)$ defined as before.

Model (8) contains only one redundant parameter in the α^*'s due to the fact that $\sum_{j=1}^{m} w_j(g_i) = 2$ for $i = 1, ..., N$. In this case, as shown in Appendix A, the parameters $\delta_1^*, \ldots, \delta_m^*$ in model (8) are estimable but the parameters μ^* and $\alpha_1^*, \ldots, \alpha_m^*$ are not estimable. To overcome the redundant parameter problem, we can drop w_m from model (8) and consider

$$G(g_i) = \mu + \sum_{j=1}^{m-1} \alpha_j w_j(g_i) + \sum_{j=1}^{m} \delta_j v_j(g_i) \quad (9)$$

for $i = 1, ..., N$. Note that $v_m = z_{1m} z_{2m} = 1 - \sum_{j=1}^{m-1} w_j + \sum_{j=1}^{m-1} \sum_{k=1}^{m-1} v_{jk}$, which cannot be completely determined by $\{w_j, v_j, j = 1, ..., m - 1\}$. Therefore, dropping $\{\delta_{jk}, j, k = 1, ..., m - 1, j < k\}$ from model (4) does not directly lead to an equivalent model of (9) as the latter contains v_m. In fact, as further dropping v_m in (9), it will lead to a more restricted model structure for the expected genotypic values with the similar interpretation of its model parameters as presented in model (4). It is also interesting to see that the haplotype coding proposed in [13] is a special case of model (9) when we further ignore all the allelic interactions and drop all the $\{v_j, j = 1, ..., m\}$ in the model.

By definition, a reduced model can be derived from its original model by adding certain restrictions on the model parameters. Typically, the model parameters in a reduced model could be interpreted similarly as that in its original model when these restrictions are simple enough (e.g., by setting a subset of them being zero). When the restrictions on the original model parameters are complicated, however, the interpretation of the reduced model parameters could be different from that presented in the original model. For model (9), we can establish the relationship between its model parameters and the expected genotypic values using a classical matrix approach, as shown in Appendix B. An alternative way of building this relationship is to simply treat model (9) as a reduced form of model (8) by adding a restriction $\alpha_m^* = 0$ and taking $\mu = \mu^*$, $\alpha_j = \alpha_j^*$ for $j = 1, ..., m - 1$, and $\delta_j = \delta_j^*$ for $j = 1, ..., m$. Note that adding the restriction $\alpha_m^* = 0$ on (8) does not change the modeling structure of the expected genotypic values because α_m^* is a redundant parameter given the others. Therefore,

$$\begin{cases} \mu = G_{mm} - \delta_m^* = G_{jm} + G_{km} - G_{jk}, \\ \quad j \neq k \neq m \\ \alpha_j = G_{jm} - \mu^* = G_{jk} - G_{km}, \ k \neq j, m, \\ \quad j = 1, \ldots, m-1 \\ \delta_j = G_{jj} - (\mu^* + 2\alpha_j^*) \\ \quad = (G_{jj} - G_{jk}) - (G_{jl} - G_{kl}), j \neq k \neq l, \\ \quad j = 1, \ldots, m \end{cases}$$

Comparing with the parameters in model (4), we can see that the interpretation of the parameters in model (9) have changed slightly. The intercept μ now becomes $(G_{mm} - \delta_m^*)$ instead of G_{mm}, the α_j is the substitution effect of replacing allele A_m by A_j when paired with any allele A_k ($k \neq j, m$) instead of just A_m, while the δ_j is the

difference between the substitution effect of replacing any allele A_k by A_j when paired with A_j itself and that when paired with another allele A_l ($l \neq j, k$). If both α_j and δ_j are zero for a particular $j < m$, then $G_{jj} = G_{jm} = \mu$ and $G_{jk} = G_{km}$ for any $k \neq j, m$.

Under the same model framework (8), the F_∞ coding leads to the following model

$$G(g_i) = \tau + \sum_{j=1}^{m-1} a_j f_j(g_i) + \sum_{j=1}^{m} d_j h_j(g_i) \quad (10)$$

for $i = 1, ..., N$. By applying the relationship $f_j(g) = w_j(g) - 1$ and $h_j(g) = w_j(g) - 2v_j(g)$ for $j = 1, ..., m$, we can show that for models (10) and (8) to be equivalent their model parameters have the relationship

$$\begin{cases} \tau = \mu^* + \sum_{j=1}^{m} (\alpha_j^* + \frac{\delta_j^*}{2}) \\ a_j = \alpha_j^* + \frac{\delta_j^*}{2}, j = 1, \ldots, m-1 \\ d_j = -\frac{\delta_j^*}{2}, j = 1, \ldots, m \\ \alpha_m^* + \frac{\delta_m^*}{2} = 0 \end{cases}$$

In other words, model (10) leads to a restriction $2\alpha_m^* + \delta_m^* = 0$ on the parameters in model (8) which makes $\mu^* = G_{mm} - (2\alpha_m^* + \delta_m^*) = G_{mm}$, $\alpha_m^* = -\delta_m^*/2$ and $\alpha_j^* = G_{jm} - (\mu^* + \alpha_m^*) = G_{jm} - G_{mm} + \delta_m^*/2$, $j = 1, ..., m - 1$. Thus,

$$\begin{cases} \tau = G_{mm} + \frac{1}{2} \sum_{j=1}^{m-1} (G_{jj} - G_{mm}) \\ a_j = \frac{G_{jj} - G_{mm}}{2}, j = 1, \ldots, m-1 \\ d_j = -\frac{(G_{jj} - G_{jk}) - (G_{jl} - G_{kl})}{2}, j \neq k \neq l, \\ \quad j = 1, \ldots, m \end{cases}$$

Now d_j becomes a half of the difference between the substitution effect of replacing any allele A_k by A_j when paired with another A_j and that when paired with an allele A_l ($l \neq j, k$), which can no longer be referred to as a dominance effect.

With the allele-count coding, we can actually construct two equivalent models in this case

$$G(g_i) = \pi_0 + \sum_{j=1}^{m-1} \pi_j h_{1j}(g_i) + \sum_{j=1}^{m} \eta_j h_{2j}(g_i) \quad (11)$$

and

$$G(g_i) = \pi'_0 + \sum_{j=1}^{m} \pi'_j h_{1j}(g_i) + \sum_{j=1}^{m-1} \eta'_j h_{2j}(g_i) \quad (12)$$

for $i = 1, ..., N$. Similarly, we can show that model (11) can be treated as a reduced model by adding the restriction $\alpha_m^* = 0$ on parameters in model (8) with the following relationships

$$\begin{cases} \pi_0 = \mu^* = G_{jm} + G_{km} - G_{jk}, \\ \quad j \neq k \neq m \\ \pi_j = \alpha_j^* = G_{jk} - G_{km}, k \neq j, m, \\ \quad j = 1, \ldots, m-1 \\ \eta_j = 2\alpha_j^* + \delta_j^* = (G_{jj} - G_{jm}) + (G_{jk} - G_{km}), \\ \quad k \neq j, m, j = 1, \ldots, m-1 \\ \eta_m = \delta_m^* = (G_{mm} - G_{jm}) - (G_{km} - G_{jk}), \\ \quad j \neq k \neq m \end{cases}$$

On the other hand, model (12) can be treated as a reduced model by adding the restriction $2\alpha_m^* + \delta^* = 0$ on parameters in model (10) with the following relationships

$$\begin{cases} \pi'_0 = \mu^* = G_{mm} \\ \pi'_j = \alpha_j^* = \frac{(G_{jm} - G_{mm}) + (G_{jk} - G_{km})}{2}, \\ \quad k \neq j, m, j = 1, \ldots, m-1 \\ \pi'_m = -\frac{\delta_m^*}{2} = -\frac{(G_{mm} - G_{jm}) - (G_{km} - G_{jk})}{2}, \\ \quad j \neq k \neq m \\ \eta'_j = 2\alpha_j^* + \delta_j^* = G_{jj} - G_{mm}, \\ \quad j = 1, \ldots, m-1 \end{cases}$$

While the effect η_{jj} in model (6) is the difference between the two expected homozygous genotypic values G_{jj} and G_{mm}, the effect η_j in model (11) becomes the sum of the substitution effects of replacing allele A_m by A_j when paired with A_j itself and when paired with another allele A_k ($k \neq j, m$. It is also interesting to see that the definition of parameters in models (11) and (12) are quite different. A null hypothesis of $H_0 : \pi'_j = \eta'_j = 0$ for a particular $j < m$ in model (12) implies that $G_{jj} = G_{mm}$ and $G_{jm} - G_{mm} = G_{jk} - G_{km}$ for any $k \neq j, m$, while the null hypothesis of $H_0 : \pi_j = \eta_j = 0$ for a $j < m$ in model (11) implies that $G_{jj} = G_{jm}$ and $G_{jk} = G_{km}$ for any $k \neq j, m$, which has nothing to do with G_{mm}.

Under the same model framework (8), each of the above four models (9), (10), (11) and (12) contains $2m$ non-redundant parameters (including the intercept) to model the $m(m + 1)/2$ expected genotypic values. When $m > 3$, we have $m(m + 1)/2 > 2m$. Therefore, the model framework (7) enforces certain constraints on the $m(m + 1)/2$ genotypic values. If $m = 3$, then each of the four models actually provides a full re-parameterization of the six expected genotypic values G_{11}, G_{22}, G_{33}, G_{12}, G_{13} and G_{23}. The relationships between the four model parameters and the expected genotypic values are summarized in Table 4.

Comparing Table 4 with Table 1, we can see that the definition of model parameters depends not only on the coding schemes of marker genotypes but also on the underlying framework for the structure of the expected genotypic values. From Table 4, it is also interesting to see that the null hypothesis of $H_0 : \alpha_j = \delta_j = 0$ ($j < m$) in model (9) is equivalent to $\pi_j = \eta_j = 0$ in model (11),

Table 4 Parameterization of one-locus models (9), (10), (11), (12) when $m \geq 3$.

Codings	Restrictions	Relationships
Allele	$\alpha_m^* = 0$	$\mu = \mu^* = (G_{jm} + G_{km}) - G_{jk}, j \neq k \neq m$
		$\alpha_j = \alpha_j^* = G_{jk} - G_{km}, j = 1, \ldots, m-1; j \neq k, m$
		$\delta_j = \delta_j^* = (G_{jj} - G_{jk}) - (G_{jl} - G_{kl}), j = 1, \ldots, m; k \neq j \neq l$
F_∞	$2\alpha_m^* + \delta_m^* = 0$	$\tau = \mu^* + \frac{1}{2}\sum_{j=1}^{m-1}(2\alpha_j^* + \delta_j^*) = G_{mm} + \frac{1}{2}\sum_{j=1}^{m-1}(G_{jj} - G_{mm})$
		$a_j = \frac{1}{2}(2\alpha_j^* + \delta_j^*) = \frac{G_{jj} - G_{mm}}{2}, j = 1, \ldots, m-1$
		$d_j = -\frac{\delta_j^*}{2} = -\frac{(G_{jj} - G_{jk}) - (G_{jl} - G_{kl})}{2}, j = 1, \ldots, m; j \neq k \neq l$
Allele-count	$\alpha_m^* = 0$	$\pi_0 = \mu^* = (G_{jm} + G_{km}) - G_{jk}, j \neq k \neq m$
		$\pi_j = \alpha_j^* = G_{jk} - G_{km}, j = 1, \ldots, m-1; k \neq j, m$
		$\eta_j = 2\alpha_j^* + \delta_j^* = (G_{jj} - G_{jm}) + (G_{jk} - G_{km}), j = 1, \ldots, m-1; k \neq j, m$
		$\eta_m = \delta_m^* = (G_{mm} - G_{jm}) - (G_{km} - G_{jk}), j \neq k \neq m$
Allele-count	$2\alpha_m^* + \delta_m^* = 0$	$\pi'_0 = \mu^* = G_{mm}$
		$\pi'_j = \alpha_j^* = \frac{(G_{jm} - G_{mm}) + (G_{jk} - G_{km})}{2}, j = 1, \ldots, m-1; k \neq j, m$
		$\pi'_m = -\frac{\delta_m^*}{2} = -\frac{(G_{mm} - G_{jm}) - (G_{km} - G_{jk})}{2}, j \neq k \neq m$
		$\eta'_j = 2\alpha_j^* + \delta_j^* = G_{jj} - G_{mm}, j = 1, \ldots, m-1$

which implies $\alpha_j^* = \delta_j^* = 0$ in model (8) with restriction $\alpha_m^* = 0$, or $G_{jk} = G_{km}$ for any $k = 1, ..., m$. On the other hand, the null hypothesis of $H_0 : a_j = d_j = 0$ ($j < m$) in model (10) is equivalent to $\pi'_j = \eta'_j = 0$ in model (12), which implies $\alpha_j^* = \delta_j^* = 0$ in model (8) with a restriction $2\alpha_m^* + \delta_m^* = 0$, or $G_{jj} = G_{mm}$ and G_{jj} - $G_{jm} = G_{jk}$ - G_{km} for any $k \neq m$. In general, the two null hypotheses of $\alpha_j = \delta_j = 0$ and $a_j = d_j = 0$ may not always be equivalent. For example, when $m = 3$, similar to the three-allele models discussed in the previous section, we can show that the four model parameters and the expected genotypic values have the relationships as shown in Table 5, which is a special case of Table 4. We can see from Table 5 that $\alpha_1 = \delta_1 = 0$ is equivalent to $\pi_1 = \eta_1 = 0$ which implies $G_{12} = G_{23}$ and $G_{11} = G_{13}$; while $a_1 = d_1 = 0$ is equivalent to $\pi'_1 = \eta'_1 = 0$ which implies $G_{11} = G_{33}$ and $G_{12} + G_{13} = G_{11} + G_{23}$. So, depending on the underlying true setting of the expected genotypic values, the null hypotheses of $\alpha_1 = \delta_1 = 0$ in model (9) could be different from that of $a_1 = d_1 = 0$ in model (10).

Table 5 Parameterization of one-locus models (9), (10), (11), (12) when $m = 3$.

Codings	Restrictions	Relationships
Allele	$\alpha_3^* = 0$	$\mu = G_{13} + G_{23} - G_{12}$
		$\alpha_1 = G_{12} - G_{23}, \alpha_2 = G_{12} - G_{13}$
		$\delta_1 = (G_{11} - G_{13}) - (G_{12} - G_{23})$
		$\delta_2 = (G_{22} - G_{23}) - (G_{12} - G_{13})$
		$\delta_3 = G_{33} + G_{12} - G_{13} - G_{23}$
F_∞	$2\alpha_3^* + \delta_3^* = 0$	$\tau = \frac{G_{11} + G_{22}}{2}$
		$a_1 = \frac{G_{11} - G_{33}}{2}, a_2 = \frac{G_{22} - G_{33}}{2}$
		$d_1 = \frac{(G_{12} + G_{13}) - (G_{23} + G_{11})}{2}$
		$d_2 = \frac{(G_{12} + G_{23}) - (G_{13} + G_{22})}{2}$
		$d_3 = \frac{(G_{13} + G_{23}) - (G_{12} + G_{33})}{2}$
Allele-count	$\alpha_3^* = 0$	$\pi_0 = G_{13} + G_{23} - G_{12}$
		$\pi_1 = G_{12} - G_{23}, \pi_2 = G_{12} - G_{13}$
		$\eta_1 = (G_{11} - G_{13}) + (G_{12} - G_{23})$
		$\eta_2 = (G_{22} - G_{23}) + (G_{12} - G_{13})$
		$\eta_3 = G_{33} + G_{12} - G_{13} - G_{23}$
Allele-count	$2\alpha_3^* + \delta_3^* = 0$	$\pi'_0 = G_{33}$
		$\pi'_1 = \frac{(G_{12} + G_{13}) - (G_{23} + G_{33})}{2}$
		$\pi'_2 = \frac{(G_{12} + G_{23}) - (G_{13} + G_{33})}{2}$
		$\pi'_3 = \frac{(G_{13} + G_{23}) - (G_{12} + G_{33})}{2}$
		$\eta'_1 = G_{11} - G_{33}, \ \eta'_2 = G_{22} - G_{33}$

Extension to two-locus models

In this section, we further explore some extensions of the previous one-locus models to two-locus models. Consider two marker loci with alleles $A_{11}, \ldots, A_{1m_1}$ at locus 1 and alleles $A_{21}, \ldots, A_{2m_2}$ at locus 2, respectively. Without distinguishing the parental origins of the alleles, there are totally $m_1m_2(m_1 + 1)(m_2 + 1)/4$ possible distinctive expected genotypic values: $G_{jkrs} = E(G|$

$A_{1j}A_{1k}A_{2r}A_{2s}$) for $j, k = 1, ..., m_1$, $j \le k$; and $r, s = 1, ..., m_2$, $r \le s$. Using the allele coding, we introduce the following coding variables

$$w_{1j}(g) = \begin{cases} 2, \text{ if } g = A_{1j}A_{1j} \\ 1, \text{ if } g = A_{1j}A_{1j}^c \\ 0, \text{ if } g = A_{1j}^cA_{1j}^c \end{cases}$$

$$v_{1jk}(g) = \begin{cases} 1, \text{ if } g = A_{1j}A_{1k} \\ 0, \text{ otherwise} \end{cases}$$

$j, k = 1, ..., m_1$, for marker genotypes at locus 1 and

$$w_{2r}(g) = \begin{cases} 2, \text{ if } g = A_{2r}A_{2r} \\ 1, \text{ if } g = A_{2r}A_{2r}^c \\ 0, \text{ if } g = A_{2r}^cA_{2r}^c \end{cases}$$

$$v_{2rs}(g) = \begin{cases} 1, \text{ if } g = A_{2r}A_{2s} \\ 0, \text{ otherwise} \end{cases}$$

$r, s = 1, ..., m_2$, for marker genotypes at locus 2, where A_{1j}^c (or A_{2r}^c) denotes any other allele type except A_{1j} (or A_{2r}) at locus 1 (or 2). A fully parameterized two-locus model for G_{jkrs} can then be presented as

$$\begin{aligned} G(g_i) = \mu &+ \sum_{j=1}^{m_1-1} \alpha_{1j}w_{1j} + \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1} \delta_{1jk}v_{1jk} \\ &+ \sum_{r=1}^{m_2-1} \alpha_{2r}w_{2r} + \sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} \delta_{2rs}v_{2rs} \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} (\alpha_{1j}\alpha_{2r})w_{1j}w_{2r} \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (\alpha_{1j}\delta_{2rs})w_{1j}v_{2rs} \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1} (\delta_{1jk}\alpha_{2r})v_{1jk}w_{2r} \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (\delta_{1jk}\delta_{2rs})v_{1jk}v_{2rs} \end{aligned} \quad (13)$$

for $i = 1, ..., N$. Similar to the one-locus models, we can establish the relationship between the model parameters and the expected genotypic values as shown in (C.1) of Appendix C. A nice property of this allele coding model is that a higher order effect is simply the deviation of its corresponding expected genotypic value from an approximation of the other lower order effects. Here the corresponding expected genotypic value of a marker effect is determined by the position of alleles that differ from the two reference alleles A_{1m_1} and A_{2m_2}. So, starting from the lowest order parameter μ, it seems straightforward to build the relationships between the model parameters and the expected genotypic values starting from the low-order effect parameters up to the high-order effect parameters.

For the F_∞ coding, we can define the following coding variables for the genotypes at the two marker loci separately.

$$f_{1j}(g) = \begin{cases} 1, \textit{ if } g = A_{1j}A_{1j} \\ 0, \textit{ if } g = A_{1j}A_{1j}^c \\ -1, \textit{ if } g = A_{1j}^cA_{1j}^c \end{cases}$$

$$h_{1j}(g) = \begin{cases} 1, \textit{ if } g = A_{1j}A_{1j}^c \\ 0, \text{otherwise} \end{cases}$$

for $j = 1, ..., m_1$, and

$$f_{2r}(g) = \begin{cases} 1, \textit{ if } g = A_{2r}A_{2r} \\ 0, \textit{ if } g = A_{2r}A_{2r}^c \\ -1, \textit{ if } g = A_{2r}^cA_{2r}^c \end{cases}$$

$$h_{2r}(g) = \begin{cases} 1, \textit{ if } g = A_{2r}A_{2r} \\ 0, \text{otherwise} \end{cases}$$

for $r = 1, ..., m_2$. A fully parameterized two-locus model using this F_∞ coding is then

$$\begin{aligned} G(g_i) = \tau &+ \sum_{j=1}^{m_1-1} a_{1j}f_{1j}(g_i) + \sum_{r=1}^{m_2-1} a_{2r}f_{2r}(g_i) \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1} d_{1jk}h_{1j}(g_i)h_{1k}(g_i) \\ &+ \sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} d_{2rs}h_{2r}(g_i)h_{2s}(g_i) \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} (a_{1j}a_{2r})f_{1j}f_{2r} \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (a_{1j}d_{2rs})f_{1j}h_{2r}h_{2s} \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1} (d_{1jk}a_{2r})h_{1j}h_{1k}f_{2r} \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r}^{m_2-1} (d_{1jk}d_{2rs})h_{1j}h_{1k}h_{2r}h_{2s} \end{aligned} \quad (14)$$

for $i = 1, ..., N$. Still, using the relationships $w_{1j} = 1 + f_{1j}$, $w_{2r} = 1 + f_{2r}$, $v_{1jj} = (1 + f_{1j} - h_{1j})$, $v_{2rr} = (1 + f_{2r} - h_{2r})$, $v_{1jk} = h_{1j}h_{1k}$ for $j \ne k$, and $v_{2rs} = h_{2r}h_{2s}$ for $r \ne s$ between the F_∞ coding variables and the allele coding variables, we can establish the relationships between the model parameters and the expected genotypic values as shown in (C.2) of Appendix C. We can easily verify that the biallelic two-locus effects $E_{F_\infty \cdot AB}$ in [9] is a special case of our results with $m_1 = m_2 = 2$. It is also interesting to see that the interpretation of model parameters in terms of the expected genotypic values

becomes much more complicated than that in the previous allele coding model. When m_1, $m_2 > 2$, the low-order within-locus main effect $a_{1j}$is a weighted combination of the differences $(G_{jjrr} - G_{m_1m_1rr})$, where $r = 1, ..., m_2$ refer to various homozygous genotypes $A_{2r}A_{2r}$at locus 2. The within-locus effect $d_{1jj}$is a weighted combination of the allelic interactions $(G_{jjrr} - 2G_{jm_1rr} + G_{m_1m_1rr})$, $r = 1, ..., m_2$, at locus 1 with reference $A_{2r}A_{2r}$at locus 2. Even the intercept τ of the model becomes a complex function of various homozygous genotypic values.

Applying the allele-count coding, we can define

$$h_{1j}^{(1)}(g) = \begin{cases} 1, if\ g = A_{1j}A_{1j}^c \\ 0, otherwise \end{cases}$$

$$h_{2j}^{(1)}(g) = \begin{cases} 1, if\ g = A_{1j}A_{1j} \\ 0, \text{otherwise} \end{cases}$$

for $j = 1, ..., m_1$, and

$$h_{1r}^{(2)}(g) = \begin{cases} 1, if\ g = A_{2r}A_{2r}^c \\ 0, \text{otherwise} \end{cases}$$

$$h_{2r}^{(2)}(g) = \begin{cases} 1, if\ g = A_{2r}A_{2r} \\ 0, \text{otherwise} \end{cases}$$

for $r = 1, ..., m_2$. Another fully parameterized two-locus model for G_{jkrs} can be written as

$$\begin{aligned} G(g_i) = \pi_0 &+ \sum_{j=1}^{m_1-1} (\pi_{1j}h_{1j}^{(1)} + \eta_{1jj}h_{2j}^{(1)}) \\ &+ \sum_{r=1}^{m_2-1} (\pi_{2r}h_{1r}^{(2)} + \eta_{2rr}h_{2r}^{(2)}) \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j+1}^{m_1-1} \eta_{1jk}h_{1j}^{(1)}h_{1k}^{(1)} \\ &+ \sum_{r=1}^{m_2-1}\sum_{s=r+1}^{m_2-1} \eta_{2rs}h_{1r}^{(2)}h_{1s}^{(2)} \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} \Big[(\pi_{1j}\pi_{2r})h_{1j}^{(1)}h_{1r}^{(2)} \\ &\quad +(\pi_{1j}\eta_{2rr})h_{1j}^{(1)}h_{2r}^{(2)} + (\eta_{1jj}\pi_{2r})h_{2j}^{(1)}h_{1r}^{(2)} \\ &\quad +(\eta_{1jj}\eta_{2rr})h_{2j}^{(1)}h_{2r}^{(2)}\Big] \\ &+ \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r+1}^{m_2-1} \Big[(\pi_{1j}\eta_{2rs})h_{1j}^{(1)}h_{1r}^{(2)}h_{1s}^{(2)} \\ &\quad +(\eta_{1jj}\eta_{2rs})h_{2j}^{(1)}h_{1r}^{(2)}h_{1s}^{(2)}\Big] \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j+1}^{m_1-1}\sum_{r=1}^{m_2-1} \Big[(\eta_{1jk}\pi_{2r})h_{1j}^{(1)}h_{1k}^{(1)}h_{1r}^{(2)} \\ &\quad +(\eta_{1jk}\eta_{2rr})h_{1j}^{(1)}h_{1k}^{(1)}h_{2r}^{(2)}\Big] \\ &+ \sum_{j=1}^{m_1-1}\sum_{k=j+1}^{m_1-1}\sum_{r=1}^{m_2-1}\sum_{s=r+1}^{m_2-1} (\eta_{1jk}\eta_{2rs}) \\ &\cdot h_{1j}^{(1)}h_{1k}^{(1)}h_{1r}^{(2)}h_{1s}^{(2)} \end{aligned} \tag{15}$$

for $i = 1, ..., N$. In this case, the allele-count coding variables and the allele coding variables have the relationships $w_{1j} = h_{1j}^{(1)} + 2h_{2j}^{(1)}$, $w_{2r} = h_{1r}^{(2)} + 2h_{2r}^{(2)}$, $v_{2rr} = h_{2r}^{(2)}$, $v_{2rr} = h_{2r}^{(2)}$, $v_{1jk} = h_{1j}^{(1)}h_{1k}^{(1)}$ for $j \neq k$, and $v_{2rs} = h_{1r}^{(2)}h_{1s}^{(2)}$ for $r \neq s$. Through the equivalence of the two models (13) and (15), we can also construct relationships between the parameters in model (15) and the expected genotypic values as shown in (C.3) of Appendix C. We can see that the interpretation of parameters in the allele-count coding model (15) are as simple as that in the allele coding model (13) with the same intercept being $G_{m_1m_1m_2m_2}$. Besides, it seems that some parameters such as $(\eta_{1jj}\eta_{2rr})$, $(\eta_{1jk}\eta_{2rs})$ and $(\eta_{1jk}\eta_{2rr})$ have simpler relationships than the corresponding ones in the allele coding model (13).

Finally, let us consider some reduced cases of the two-locus models. By ignoring locus-by-locus interactions (i.e., epistases), we have the following simplified two-locus model framework

$$G_{jkrs} = \mu^* + \alpha_{1j}^* + \alpha_{1k}^* + \delta_{1jk}^* + \alpha_{2r}^* + \alpha_{2s}^* + \delta_{2rs}^* \tag{16}$$

for $j, k = 1, ..., m_1$ and $r, s = 1, ..., m_2$. If we further ignore the within-locus allelic interactions between different alleles, then another reduced two-locus model framework is

$$\begin{aligned} G_{jkrs} = \mu^* &+ \alpha_{1j}^* + \alpha_{1k}^* + \delta_{1j}^* 1_{\{j=k\}} \\ &+ \alpha_{2r}^* + \alpha_{2s}^* + \delta_{2r}^* 1_{\{r=s\}} \end{aligned} \tag{17}$$

Similar to the one-locus models, under each of the two reduced model frameworks we can construct the two-locus models from the three coding schemes. The relationships between the model parameters and the expected genotypic values under framework (14) are summarized in Table 6, which can be treated as an extension of Table 1 to the two-locus case. The relationships between the model parameters and the expected genotypic values under framework (17) are also summarized in Table 7, which is a straightforward extension of Table 4. Further dropping δ_{1j}^* for $j = 1, ..., m_1$ and δ_{2r}^* for $r = 1, ..., m_2$ in (15) will lead to an additive model framework, which has its model parameters interpretable similar to that in Table 6. From Tables 6 and 7, we can see that both the allele and allele-count coding models have their lower-order main effects keep similar interpretation as to that in the previous fully parameterized case with epistases, while the F_∞ coding models have the definition of their lower-order main effects vary depending on whether there are epistases involved in the models.

As pointed out in [9], the genetic effects of a marker may have different interpretation depending upon whether the marker is fitted in a one-locus model or a two-locus model. From the linear model theory, the genetic effects of a marker in a one-locus model are

Table 6 Parameterization of two-locus models under model framework (16).

Codings	Relationships
Allele	$\mu = G_{m_1m_1m_2m_2}$
	$\alpha_{1j} = G_{jm_1m_2m_2} - G_{m_1m_1m_2m_2}, j = 1, \ldots, m_1 - 1$
	$\alpha_{2r} = G_{m_1m_1rm_2} - G_{m_1m_1m_2m_2}, r = 1, \ldots, m_2 - 1$
	$\delta_{1jk} = G_{jkm_2m_2} - G_{jm_1m_2m_2} - G_{km_1m_2m_2} + G_{m_1m_1m_2m_2}, j, k = 1, \ldots, m_1 - 1; j \le k$
	$\delta_{2rs} = G_{m_1m_1rs} - G_{m_1m_1rm_2} - G_{m_1m_1sm_2} + G_{m_1m_1m_2m_2}, r, s = 1, \ldots, m_2 - 1; r \le s$
F_∞	$\tau = G_{m_1m_1m_2m_2} + \frac{1}{2}\sum_{j=1}^{m_1-1}(G_{jjm_2m_2} - G_{m_1m_1m_2m_2}) + \frac{1}{2}\sum_{r=1}^{m_2-1}(G_{m_1m_1rr} - G_{m_1m_1m_2m_2})$
	$a_{1j} = \frac{G_{jjm_2m_2} - G_{m_1m_1m_2m_2}}{2}, j = 1, \ldots, m_1 - 1$
	$a_{2r} = \frac{G_{m_1m_1rr} - G_{m_1m_1m_2m_2}}{2}, r = 1, \ldots, m_2 - 1$
	$d_{1jj} = G_{jm_1m_2m_2} - \frac{G_{jjm_2m_2} + G_{m_1m_1m_2m_2}}{2}, j = 1, \ldots, m_1 - 1$
	$d_{1jk} = G_{jkm_2m_2} - G_{jm_1m_2m_2} - G_{km_1m_2m_2} + G_{m_1m_1m_2m_2}, j, k = 1, \ldots, m_1 - 1; j < k$
	$d_{2rr} = G_{m_1m_1rm_2} - \frac{G_{m_1m_1rr} + G_{m_1m_1m_2m_2}}{2}, r = 1, \ldots, m_2 - 1$
	$d_{2rs} = G_{m_1m_1rs} - G_{m_1m_1rm_2} - G_{m_1m_1sm_2} + G_{m_1m_1m_2m_2}, r, s = 1, \ldots, m_2 - 1; r < s$
Allele-count	$\pi_0 = G_{m_1m_1m_2m_2}$
	$\pi_{1j} = G_{jm_1m_2m_2} - G_{m_1m_1m_2m_2}, j = 1, \ldots, m_1 - 1$
	$\pi_{2r} = G_{m_1m_1rm_2} - G_{m_1m_1m_2m_2}, r = 1, \ldots, m_2 - 1$
	$\eta_{1jj} = G_{jjm_2m_2} - G_{m_1m_1m_2m_2}, j = 1, \ldots, m_1 - 1$
	$\eta_{1jk} = G_{jkm_2m_2} - G_{jm_1m_2m_2} - G_{km_1m_2m_2} + G_{m_1m_1m_2m_2}, j, k = 1, \ldots, m_1 - 1; j < k$
	$\eta_{2rr} = G_{m_1m_1rr} - G_{m_1m_1m_2m_2}, r = 1, \ldots, m_2 - 1$
	$\eta_{2rs} = G_{m_1m_1rs} - G_{m_1m_1rm_2} - G_{m_1m_1sm_2} + G_{m_1m_1m_2m_2}, r, s = 1, \ldots, m_2 - 1; r < s$

defined based on the expected genotypic values of certain genotypes at this particular marker locus with genotypes at the other marker loci being averaged out based on the joint genotype distribution. For instance, marker 1 in the two-locus setting above has its effects defined in a one-locus model based on the one-locus genotypic values $E(G_{jk}) = E(G_{jkrs}|A_{1j}A_{1k}) = \sum_{rs} P(A_{2r}A_{2s}|A_{1j}A_{1k})G_{jkrs}$, which could depend on the LDs of alleles between the two loci. When the same marker is fitted in a two-locus model, its effects are usually functions of the expected genotypic values with their joint genotypes taking certain reference alleles or genotypes at the other marker loci. So, in general, even without locus-by-locus interactions, a single marker's effects could be different from the one defined in a multi-locus model when the alleles at different loci are in linkage disequilibrium (LD). Consider a 2-locus haploid model with alleles *A*, *a* at locus 1 and *B*, *b* at locus 2. If we ignore the locus-by-locus interaction, it is easy to show that the additive allelic effects are $\alpha_1 = G_{AB} - G_{aB} = G_{Ab} - G_{ab}$ and $\alpha_2 = G_{AB} -$

Table 7 Parameterization of two-locus models under model framework (17) when $m_1, m_2 \ge 3$.

Codings	Restrictions	Relationships
Allele	$\alpha^*_{1m_1} = \alpha^*_{2m_2} = 0$	$\mu = G_{m_1m_1m_2m_2} - (G_{m_1m_1rs} - G_{jm_1rs} - G_{km_1rs} + G_{jkrs})$ $-(G_{jkm_2m_2} - G_{jkm_2r} - G_{jkm_2s} + G_{jkrs}), j \ne k \ne m_1; r \ne s \ne m_2$
		$\alpha_{1j} = G_{jkrs} - G_{m_1krs}, j = 1, \ldots, m_1 - 1; k \ne j, m_1$
		$\alpha_{2r} = G_{jkrs} - G_{jkm_2s}, r = 1, \ldots, m_2 - 1, r \ne s, m_2$
		$\delta_{1j} = G_{jjrs} - G_{jkrs} - G_{jlrs} + G_{klrs}, j = 1, \ldots, m_1; j \ne k \ne l$
		$\delta_{2r} = G_{jkrr} - G_{jkrs} - G_{jkrt} + G_{jkst}, r = 1, \ldots, m_2; r \ne s \ne t$
F_∞	$2\alpha^*_{1m_1} + \delta^*_{1m_1} = 0$	$\tau = G_{m_1m_1m_2m_2} + \frac{1}{2}\sum_{j=1}^{m_1-1}(G_{jjm_2m_2} - G_{m_1m_1m_2m_2})$
	$2\alpha^*_{2m_2} + \delta^*_{2m_2} = 0$	$+\frac{1}{2}\sum_{r=1}^{m_2-1}(G_{m_1m_1rr} - G_{m_1m_1m_2m_2})$
		$a_{1j} = \frac{G_{jjrs} - G_{m_2m_2rs}}{2}, j = 1, \ldots, m_1 - 1$
		$a_{2r} = \frac{G_{jkrr} - G_{jkm_2m_2}}{2}, r = 1, \ldots, m_2 - 1$
		$d_{1j} = -\frac{G_{jjrs} - G_{jkrs} - G_{jlrs} + G_{klrs}}{2}, j = 1, \ldots, m_1; j \ne k \ne l$
		$d_{2r} = -\frac{G_{jkrr} - G_{jkrs} - G_{jkrt} + G_{jkst}}{2}, r = 1, \ldots, m_2; r \ne s \ne t$
Allele-count	$\alpha^*_{1m_1} = \alpha^*_{2m_2} = 0$	$\pi_0 = G_{m_1m_1m_2m_2} - (G_{m_1m_1rs} - G_{jm_1rs} - G_{m_1krs} + G_{jkrs})$ $-(G_{jkm_2m_2} - G_{jkrm_2} - G_{jkm_2s} + G_{jkrs}), j \ne k \ne m_1; r \ne s \ne m_2$
		$\pi_{1j} = G_{jkrs} - G_{m_1krs}, j = 1, \ldots, m_1 - 1, k \ne j, m_1$
		$\pi_{2r} = G_{jkrs} - G_{jkm_2s}, r = 1, \ldots, m_2 - 1, r \ne s, m_2$
		$\eta_{1j} = G_{jjrs} - G_{jm_1rs} - G_{km_1rs} + G_{jkrs}, j = 1, \ldots, m_1 - 1; k \ne j, m_1$
		$\eta_{1m_1} = G_{m_1m_1rs} - G_{jm_1rs} - G_{km_1rs} + G_{jkrs}, j \ne k \ne m_1$
		$\eta_{2r} = G_{jkrr} - G_{jkrm_2} - G_{jksm_2} + G_{jkrs}, r = 1, \ldots, m_2 - 1; s \ne r, m_2$
		$\eta_{2m_2} = G_{jkm_2m_2} - G_{jkm_2r} - G_{jkm_2s} + G_{jkrs}, r \ne s \ne m_2$
Allele-count	$2\alpha^*_{1m_1} + \delta^*_{1m_1} = 0$	$\pi'_0 = G_{m_1m_1m_2m_2}$
	$2\alpha^*_{2m_2} + \delta^*_{2m_2} = 0$	$\pi'_{1j} = \frac{G_{jm_1rs} - G_{m_1m_1rs} - G_{km_1rs} + G_{jkrs}}{2}, j = 1, \ldots, m_1 - 1; k \ne j, m_1$
		$\pi'_{1m_1} = -\frac{G_{m_1m_1rs} - G_{jm_1rs} - G_{km_1rs} + G_{jkrs}}{2}, j \ne k \ne m_1$
		$\pi'_{2r} = \frac{G_{jkrm_2} - G_{jkm_2m_2} - G_{jksm_2} + G_{jkrs}}{2}, r = 1, \ldots, m_2 - 1; r \ne s, m_2$
		$\pi'_{2m_2} = -\frac{G_{jkm_2m_2} - G_{jkrm_2} - G_{jksm_2} + G_{jkrs}}{2}, r \ne s \ne m_2$
		$\eta'_{1j} = G_{jjrs} - G_{m_1m_1rs}, j = 1, \ldots, m_1 - 1$
		$\eta'_{2r} = G_{jkrr} - G_{jkm_2m_2}, r = 1, \ldots, m_2 - 1$

$G_{Ab} = G_{aB} - G_{ab}$ at locus 1 and 2, respectively. In a one-locus model at locus 1, however, we can show that the locus has its additive allelic effect $\alpha_1^* = \alpha_1 + D\alpha_2/(p_A p_a)$, where $D = P_{AB} - p_A p_B$ is the LD between the two loci.

Simulation Examples

We use some numerical examples to illustrate properties of the models we have discussed. First, we consider the same example discussed in [11] of a three-allele locus with allele frequencies $p_1 = 0.2$ for A_1, $p_2 = 0.3$ for A_2, and $p_3 = 0.5$ for A_3. The six genotypic values are $G_{11} = 10$, $G_{12} = 30$, $G_{22} = 50$, $G_{13} = 36$, $G_{23} = 46$ and $G_{33} = 42$. We adopt a similar strategy to specify the genotype frequencies as: $P_{jj} = p_j^2 - D$ for $j = 1, 2, 3$ and $P_{jk} = 2p_j pk + D$ for $j \neq k$, where D is a measure of departure from Hardy-Weinberg equilibrium (HWE) for the three alleles at the locus and $D^- \leq D \leq D^+$ with

$$D^- = -\min_{j \neq k}\{2p_j p_k\} = -0.12$$

and

$$D^+ = \min_{j=1,2,3}\{p_j^2\} = 0.04$$

We consider two cases: i) D = 0 for HWE, and ii) $D = 0.02$ for Hardy-Weinberg disequilibrium (HWD). The phenotypic value of an individual is simulated as a sum of its true genotypic value and an environmental noise from $N(0, \sigma^2)$, where the σ^2 is chosen to be either 0 or $\sigma^2 = 288$ with the latter one corresponds to a 20% heritability level when D = 0. For each of the four configurations, we simulate 10,000 random samples with 1000 individuals each. For each random sample, we fit the three fully parameterized one-locus models (4), (5) and (6) under model framework (2) using the least square approach and estimate the model parameters as well as the six genotypic values. The means and standard deviations (SD) of the least square estimates (LSE) of the model parameters and the six genotypic values from the 10,000 random samples in fitting these three models are summarized in Table 8.

As each of the three models provides a re-parameterization of the six genotypic values, for each random sample the three models always give exactly the same estimates of the six genotypic values and the residual variance as we expected, even though their model parameters are defined in different ways. As a result, under each configuration, the three models have the same means and SD for the LSE of the six genotypic values and the residual variance. Without environmental variation, each model can accurately estimate its model parameters and the six genotypic values for each random sample regardless of whether there is HWE or HWD. When there is environmental variation on the phenotypes, it is known that the least square estimators of the model parameters are unbiased under either HWE or HWD. However, the HWD may affect the variance of the least square estimators of the model parameters and the six genotypic values. Note that the genotypic frequencies are $P_{11} = 0.04$, $P_{22} = 0.09$, $P_{33} = 0.25$, $P_{12} = 0.12$, $P_{13} = 0.20$ and $P_{23} = 0.30$ under HWE, while with D = 0.02 the genotypic frequencies become $P_{11} = 0.02$, $P_{22} = 0.07$, $P_{33} = 0.23$, $P_{12} = 0.14$, $P_{13} = 0.22$ and $P_{23} = 0.32$. So, under HWD, we tend to have more individuals carrying genotypes A_1A_2, A_1A_3, A_2A_3 but less individuals carrying genotypes A_1A_1, A_2A_2, A_3A_3 in the random samples than that under HWE. Without knowing the accurate genotypic values, more individuals with certain genotypes in a random sample can then provide better estimates of the corresponding genotypic values. This explains why under HWD the estimates of G_{11}, G_{22} and G_{33} have larger SD (or variances) than that under the HWE, and the estimates of G_{12}, G_{13} and G_{23} under HWD have smaller variances than that under the HWE.

As another example, let us consider the statistical modeling of two-locus genotypic values G_{jkrs}, where the first locus have three alleles A_1, A_2, A_3 and the second locus have two alleles B_1, B_2. Assume that the alleles at locus 1 have the same allele frequencies as that in the previous example; i.e., $p_1 = 0.2$ for A_1, $p_2 = 0.3$ for A_2, and $p_3 = 0.5$ for A_3, while the two alleles at locus 2 have frequencies $q_1 = 0.2$ for B_1 and $q_2 = 0.8$ for B_2. The two-locus genotypic values $G_2 = (G_{jkrs})$, $j, k = 1, 2, 3$; $r, s = 1, 2$ are given by

$$G_2 = \begin{bmatrix} G_{1111} & G_{1112} & G_{1122} \\ G_{2211} & G_{2212} & G_{2222} \\ G_{3311} & G_{3312} & G_{3322} \\ G_{1211} & G_{1212} & G_{1222} \\ G_{1311} & G_{1312} & G_{1322} \\ G_{2311} & G_{2312} & G_{2322} \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 10.9 & 9.6 \\ 50 & 50.3 & 49.9 \\ 42 & 42.6 & 41.2 \\ 30 & 30.5 & 29.6 \\ 36 & 36.8 & 35.4 \\ 46 & 46.7 & 45.2 \end{bmatrix}$$

which are modified values from the previous one-locus model in a way that the $G_{jk11} = G_{jk}$, $G_{jk12} = G_{jk} + e_{1jk}$and $G_{jk22} = G_{jk} - e_{2jk}$with $e_{1jk}$and e_{2jk}being some small positive fluctuations according to the genotypes B_1B_2 and B_2B_2 at locus 2. We assume Hardy-Weinberg equilibria at both loci and specify their haplotype frequencies as: $h_{11} = p_1q_1 - D_1$, $h_{12} = p_1q_2 + D_1$, $h_{21} = p_2q_1 - D_2$, $h_{22} = p_2q_2 - D_2$, $h_{31} = p_3q_1 + (D_1 - D_2)$, $h_{32} = p_3q_2 - (D_1 - D_2)$, where D_1 (and D_2) are the linkage disequilibria (LD) between alleles A_1 and B_2 (and A_2 and B_1) at the two loci. We consider two scenarios: i) $D_1 = D_2 = 0$ for linkage equilibrium (LE); and ii) $D_1 = 0$, $D_2 = 0.03$

Table 8 Means (SD) of LSE for three one-locus models (4), (5) and (6) when $m = 3$.

Allele	μ	α_1	α_2	δ_{11}	δ_{22}	δ_{12}	σ^2
True	42	-6	4	-20	0	-10	
$D = 0$, $\sigma^2 = 0$	42.0(0.00)	-6.00(0.00)	4.00(0.00)	-20.00(0.00)	0.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0$, $\sigma^2 = 288$	41.99(1.07)	-5.98(1.61)	3.99(1.44)	-20.06(3.80)	0.02(2.85)	-9.98(2.42)	287.84(12.91)
$D = 0.02$, $\sigma^2 = 0$	42.00(0.00)	-6.00(0.00)	4.00(0.00)	-20.00(0.00)	0.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0.02$, $\sigma^2 = 288$	41.98(1.14)	-5.97(1.60)	4.01(1.46)	-20.07(6.21)	0.03(3.09)	-10.04(2.31)	287.81(12.91)
	G_{11}	G_{22}	G_{33}	G_{12}	G_{13}	G_{23}	
	10	50	42	30	36	46	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(2.73)	49.99(1.79)	41.99(1.07)	30.02(1.55)	36.01(1.20)	45.98(0.98)	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(5.66)	50.03(2.21)	41.98(1.14)	29.97(1.38)	36.01(1.12)	45.99(0.93)	
F_∞	τ	a_1	a_2	d_{11}	d_{22}	d_{12}	σ^2
True	30	-16	4	10	0	-10	
$D = 0$, $\sigma^2 = 0$	30.00(0.00)	-16.00(0.00)	4.00(0.00)	10.00(0.00)	0.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0$, $\sigma^2 = 288$	29.98(1.64)	-16.01(1.46)	4.00(1.05)	10.03(1.90)	-0.01(1.42)	-9.98(2.42)	287.84(12.91)
$D = 0.02$, $\sigma^2 = 0$	30.00(0.00)	-16.00(0.00)	4.00(0.00)	10.00(0.00)	0.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0.02$, $\sigma^2 = 288$	29.99(3.05)	-16.01(2.88)	4.03(1.25)	10.04(3.10)	-0.01(1.54)	-10.04(2.31)	287.81(12.91)
	G_{11}	G_{22}	G_{33}	G_{12}	G_{13}	G_{23}	
	10	50	42	30	36	46	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(2.73)	49.99(1.79)	41.99(1.07)	30.02(1.55)	36.01(1.20)	45.98(0.98)	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(5.66)	50.03(2.21)	41.98(1.14)	29.97(1.38)	36.01(1.12)	45.99(0.93)	
Allele-count	π_0	π_1	π_2	η_{11}	η_{22}	η_{12}	σ^2
True	42	-6	4	-32	8	-10	
$D = 0$, $\sigma^2 = 0$	42.00(0.00)	-6.00(0.00)	4.00(0.00)	-32.00(0.00)	8.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0$, $\sigma^2 = 288$	41.99(1.07)	-5.98(1.61)	3.99(1.44)	-32.03(2.92)	8.00(2.09)	-9.98(2.42)	287.84(12.91)
$D = 0.02$, $\sigma^2 = 0$	42.00(0.00)	-6.00(0.00)	4.00(0.00)	-32.00(0.00)	8.00(0.00)	-10.00(0.00)	0.00(0.00)
$D = 0.02$, $\sigma^2 = 288$	41.98(1.14)	-5.97(1.60)	4.01(1.46)	-32.02(5.76)	8.05(2.51)	-10.04(2.31)	287.81(12.91)
	G_{11}	G_{22}	G_{33}	G_{12}	G_{13}	G_{23}	
	10	50	42	30	36	46	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(2.73)	49.99(1.79)	41.99(1.07)	30.02(1.55)	36.01(1.20)	45.98(0.98)	
	10.00(0.00)	50.00(0.00)	42.00(0.00)	30.00(0.00)	36.00(0.00)	46.00(0.00)	
	9.96(5.66)	50.03(2.21)	41.98(1.14)	29.97(1.38)	36.01(1.12)	45.99(0.93)	

for LD. The phenotypic value of an individual is still simulated as a sum of its genotypic value and an environmental noise from $N(0, \sigma^2)$, where the σ^2 was chosen to be either 0 or $\sigma^2 = 286$ with the latter one corresponds to a 20% heritability level when $D_1 = D_2 = 0$. For each of the four configurations, we simulate 10,000 random samples with 1000 individuals each. For each random sample, we consider fitting models under three model frameworks: i) one-locus models (4), (5) and (6) at locus 1 under model framework (2); ii) two-locus models without epistases from the three coding schemes under model framework (14); iii) fully parameterized two-locus models (13), (14) and (15) with epistases. Still, for each random sample, the three allele coding models under the same model framework give exactly the same estimates of the 18 genotypic values as we expected (results not shown here). As the result, under each model framework, the three models have the same means and SD for the LSE of the 18 genotypic values and the residual variance, although the means and SD for the LSE of their model parameters are different. To compare the LSE of model parameters for models from the same coding under different model frameworks, we summarize in Table 9 the means and SD of the LSE of the model parameters from the 10,000 random samples in fitting the three allele-coding models: the one-locus

Table 9 Means (SD) of LSE for three allele-coding models regarding the two-locus genotypic values

One-locus model	μ	α_{11}	α_{12}	δ_{111}	δ_{122}	δ_{112}	σ^2
True	41.68	-5.81	4.03	-20.03	0.29	-10	
$D_1 = D_2 = 0$, $\sigma^2 = 0$	41.68(0.04)	-5.81(0.06)	4.03(0.06)	-20.03(0.14)	0.29(0.09)	-10.00(0.08)	0.37(0.01)
$D_1 = D_2 = 0$, $\sigma^2 = 286$	41.69(1.07)	-5.83(1.61)	4.03(1.44)	-20.00(3.79)	0.27(2.85)	-9.99(2.43)	286.58(12.82)
True	41.55	-5.74	4.21	-20.04	0.09	-10.06	
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 0$	41.55(0.04)	-5.74(0.06)	4.21(0.06)	-20.04(0.14)	0.09(0.09)	-10.06(0.08)	0.36(0.01)
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 286$	41.54(1.07)	-5.74(1.61)	4.23(1.45)	-20.09(3.81)	0.07(2.83)	-10.09(2.43)	286.27(12.94)
Two-locus model - no epistases	μ	α_{11}	α_{12}	δ_{111}	δ_{122}	δ_{112}	α_{21}
True	41.88	-5.81	4.03	-20.03	0.29	-10	0.64
$D_1 = D_2 = 0$, $\sigma^2 = 0$	41.88(0.02)	-5.81(0.01)	4.03(0.01)	-20.03(0.01)	0.29(0.05)	-10.00(0.02)	0.64(0.02)
$D_1 = D_2 = 0$, $\sigma^2 = 286$	41.91(2.88)	-5.82(1.61)	4.03(1.44)	-19.99(3.79)	0.27(2.85)	-9.99(2.43)	0.63(2.90)
	δ_{211}	σ^2					
	-1.92						
	-1.92(0.03)	0.024(0.002)					
	-1.92(3.41)	285.64(12.79)					
True	41.85	-5.80	4.06	-20.04	0.14	-10.04	0.65
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 0$	41.85(0.02)	-5.80(0.01)	4.06(0.01)	-20.04(0.01)	0.14(0.05)	-10.04(0.02)	0.65(0.02)
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 286$	41.87(2.94)	-5.80(1.61)	4.07(1.45)	-20.09(3.81)	0.12(2.83)	-10.07(2.43)	0.62(2.88)
	δ_{211}	σ^2					
	-1.92						
	-1.92(0.03)	0.02(0.00)					
	-1.88(3.38)	285.36(12.94)					
Two-locus model with epistases	μ	α_{11}	α_{12}	δ_{111}	δ_{122}	δ_{112}	α_{21}
True	42	-6	4	-20	0	-10	0.6
$D_1 = D_2 = 0$, $\sigma^2 = 0$	42.00(0.00)	-6.00(0.00)	4.00(0.00)	-20.00(0.00)	0.00(0.00)	-10.00(0.00)	0.60(0.00)
$D_1 = D_2 = 0$, $\sigma^2 = 286$	41.92(5.73)	-5.99(8.65)	4.04(7.67)	-19.79(19.86)	-0.04(15.62)	-9.82(13.54)	0.66(6.04)
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 0$	42.00(0.00)	-6.00(0.00)	4.00(0.00)	-20.00(0.00)	0.00(0.00)	-10.00(0.00)	0.60(0.00)
$D_1 = 0$, $D_2 = 0.03$, $\sigma^2 = 286$	42.24(8.60)	-6.11(11.83)	3.66(10.05)	-20.04(22.95)	0.51(14.85)	-9.77(14.64)	0.38(8.86)
	δ_{211}	$(\alpha_{11}\alpha_{21})$	$(\alpha_{12}\alpha_{21})$	$(\delta_{111}\alpha_{21})$	$(\delta_{122}\alpha_{21})$	$(\delta_{112}\alpha_{21})$	$(\alpha_{11}\delta_{211})$
	-2	0.2	0.1	-0.1	-0.5	-0.4	-0.2
	-2.00(0.00)	0.20(0.00)	0.10(0.00)	-0.10(0.00)	-0.50(0.00)	-0.40(0.00)	-0.20(0.00)
	-2.05(6.99)	0.23(9.12)	0.07(8.08)	-0.35(20.98)	-0.47(16.35)	-0.65(14.30)	-0.29(10.58)
	-2.00(0.00)	0.20(0.00)	0.10(0.00)	-0.10(0.00)	-0.50(0.00)	-0.40(0.00)	-0.20(0.00)
	-1.80(9.71)	0.24(12.24)	0.39(10.44)	-0.03(24.03)	-0.94(15.63)	-0.52(15.27)	-0.15(13.52)
	$(\alpha_{12}\delta_{211})$	$(\delta_{111}\delta_{211})$	$(\delta_{122}\delta_{211})$	$(\delta_{112}\delta_{211})$	σ^2		
	-0.2	0.2	1.7	1			
	-0.20(0.00)	0.20(0.00)	1.70(0.00)	1.00(0.00)	0.00(0.00)		
	-0.18(9.39)	0.55(24.51)	1.70(18.94)	1.35(16.46)	282.45(12.83)		
	-0.20(0.00)	0.20(0.00)	1.70(0.00)	1.00(0.00)	0.00(0.00)		
	-0.44(11.64)	0.07(27.44)	2.11(18.14)	0.98(17.29)	282.81(12.74)		

model (4), the two-locus model under model framework (14), and the two-locus model under model framework (13). Models from the other two coding schemes behave similarly.

As we mentioned before, the one-locus models are actually modeling the expected genotypic values given the genotypes at locus 1. When $D_1 = D_2 = 0$, we can show that the expected genotypic values at locus 1 are $G_{11} = 10.03$, $G_{22} = 50.03$, $G_{33} = 41.68$, $G_{12} = 29.90$, $G_{13} = 35.87$ and $G_{23} = 45.71$, which correspond to $\mu = 41.68$, $\alpha_{11} = -5.81$, $\alpha_{12} = 4.03$, $\delta_{111} = -20.03$, $\delta_{122} = 0.29$ and $\delta_{112} = -10$ as the true parameters in the allele coding one-locus model. When $D_1 = 0$, $D_2 = 0.03$, the expected genotypic values at locus 1 become $G_{11} =$

10.03, G_{22} = 50.08, G_{33} = 41.55, G_{12} = 29.97, G_{13} = 35.81 and G_{23} = 45.77, which correspond to μ = 41.55, α_{11} = -5.74, α_{12} = 4.21, δ_{111} = -20.04, δ_{122} = 0.09 and δ_{112} = -10.06 as the true parameters in the allele coding one-locus model. In both cases, the least square estimators of the one-locus model parameters are unbiased estimators of the true parameters. Note that, unlike the one-locus model in the previous example, the LSE of the model parameters are no longer exactly the same as the true values even when no environmental noises are involved. The reason is that the expected genotypic values at locus 1 depend on not only the genotypic values but also the joint genotype frequencies in the sample, which may change slightly from sample to sample due to the sampling variation.

For the two-locus model without epistases, it cannot provide unbiased estimators for all the genotypic values because of the model mis-specification. However, the LSE of its parameters associated with locus 1 are similar to the ones in the one-locus model at locus 1. In fact, as we know from the linear model theory, the true values of its parameters associated with locus 1 are the same as the ones defined in the one-locus model at locus 1 when the two loci are in LE. Under LD, the least square estimators of its model parameters associated with locus 1 could be biased, and the biasness depends on the LD setting.

The two-locus model with epistases gives a full re-parameterization of the 18 genotypic values. Therefore, when no environmental noises are involved, the LSE of its model parameters are exactly the same as their true values for each random sample regardless of the LD between the two loci. It has to be pointed out that this phenomenon holds only when the random sample contains all the 18 possible genotypes. In our simulation setting, the frequencies for certain genotypes such as $A_1A_1B_1B_1$, $A_1A_3B_1B_1$ and $A_2A_2B_1B_1$ are pretty small. As the result, we occasionally (about 22-23% of the 1000 random samples) may obtain a random sample that has no individuals carrying certain genotypes. In this case, the design matrix in the fully parameterized model becomes singular and the LSE of the model parameters are no longer unique. To keep our illustration of the model properties simple, we excluded those random samples in fitting the two-locus model with epistases (reduced models are less likely to have singular design matrices). Other techniques such as ridge regression could be applied to handle those skewed random samples. In the presence of environmental noises, it is also noted that the LSE for some of its model parameters such as δ_{111}, $(\delta_{111}\alpha_{21})$ and $(\delta_{111}\delta_{211})$ have much larger SD than the LSE of other parameters. This is due to the low frequencies of genotypes $A_1A_1B_1B_1$, $A_1A_3B_1B_1$ and $A_2A_2B_1B_1$. As a random sample has few individuals carrying these genotypes, it has reduced accuracy in estimation of their corresponding true genotypic values to which the model parameters δ_{111}, $(\delta_{111}\alpha_{21})$ and $(\delta_{111}\delta_{211})$ are related.

Discussion

In this study, we introduced three genotype coding schemes to build F_∞ models for multi-allele markers. The relationship between the model parameters and the expected genotypic values were established in some fully parameterized as well as reduced one-locus and two-locus F_∞ models. Our results showed that the relationships between the model parameters and the expected genotypic values could become more intricate in the multi-allele case than that in the biallelic case, even though the extension of the coding schemes from biallelic to multiple alleles appears straightforward. We built the relationships between different model parameters mainly through their coding variables of marker genotypes, which simplified the tedious derivation process comparing with the classical matrix approach. The F_∞ models we proposed can be used directly for association testing of multi-allele markers and their possible interactions with quantitative traits using random unrelated samples. These F_∞ models could also be applied to test for the risk haplotypes and their interactions when incorporated with the likelihood approach (e.g., [20]), or analyze family data by combining them with the likelihood to account for the transmission probability of alleles from parents to their offspring. Although our discussion focused on genetic modeling of quantitative traits, the results can be extended to other phenotypic traits such as binary outcomes in case-control studies using logistic regression models or time-to-event data using the Cox proportional hazard models.

Throughout the paper, we assumed that all the possible genotypes are available from the sampled individuals. If certain genotypes are not observable, then the expected genotypic values on these genotypes will not be estimable by themselves, which could change the interpretation of the model parameters as well. The models we have presented can also be modified to handle the situation when some individuals have missing genotypes at certain marker loci. When the missing genotypes at a marker locus have both alleles missing at the same time, we can simply introduce an indicator variable to code for the missing genotype at the marker. The regression coefficient of this indicator variable for this missing genotype can usually be interpreted as the difference between the expected genotypic value with missing genotype at the marker locus and the intercept

of the model, while the other regression coefficients would keep the same interpretation as before.

It has to be pointed out that the relationships between the model parameters and the expected genotypic values are based on the assumption that the models can correctly specify the structure of the expected genotypic values. When a fully parameterized model is applied, the definition of its model parameters do not depend on the allele frequencies, HWD among alleles within a locus, or LD structure between alleles at different loci. In fitting a reduced model, however, a simplified model may not be totally correct in modeling all the expected genotypic values. In this case, depending on how accurate the simplified model is on approximating the expected genotypic values, the allele frequencies, HWD and LD structure between marker alleles could affect the definition and LSE of its model parameters. In the presence of environmental variation on the phenotypic values, regardless of whether a fully parameterized or reduced model is applied, the allele frequencies, HWD or LD between marker alleles may affect the LSE of the model parameters and the power in detection of the associated marker alleles as shown in our simulation studies.

All the models we have discussed so far are F_∞ models. Statistically, these F_∞ models are fixed-effect models which focus on modeling the expected genotypic values directly. On the other hand, the Fisher's ANOVA models, which target on evaluation of the variations contributed by various allelic effects and interactions, can be treated as random-effect models (see [21]) in which the expected genotypic values come from a discrete random variable $G(g) = E(G|g)$ with its limited genotypes g being randomly sampled from a study population. Both the F_∞ and the Fisher type models form basis in the analysis of quantitative traits and they provide different perspectives in assessing the genetic effects of QTL and markers. For biallelic markers, we proposed in [10] a 'mean corrected' Fisher (mc-Fisher) model for decomposition of the genotypic variances. In the multi-allele marker case, we can also construct similar mc-Fisher models by applying mean corrections on all the indicator variables of the paternal and maternal alleles in the allele coding F_∞ models. For example, based on the allele coding model (4), we can construct its corresponding mc-Fisher model by replacing the coding variables w_j and v_{jk} with $\bar{w}_j = w_j - 2p_j$ and $\bar{v}_{jk} = (z_{1j} - p_j)(z_{2k} - p_k) = v_{jk} - (p_j w_k + p_k w_j)/2 + p_j p_k$, respectively; where p_j is the allele frequency of A_j. Then the genetic additive and dominant variance components V_A and V_D of $G(g)$, which are defined as variations contributed by the additive allelic effect and allelic interactions respectively, can be estimated from $\bar{w}_j$'s and $\bar{v}_{jk}$'s separately. As pointed out in [10], the mc-Fisher model can provide an orthogonal partition of V(G) into the sum of V_A and V_D under Hardy-Weinberg equilibrium, and it can be fitted through the standard least-square regression approach. Similar to the F_∞ models, the definition of the model parameters in such a mc-Fisher model also depend on the choice of the reference allele 'A_m'. But the estimates of the additive and dominant variance components V_A and V_D do not depend on such a choice. In addition, when a fully parameterized model is applied, the mc-Fisher model is equivalent to its original F_∞ model in modeling the expected genotypic values. Therefore, both models have the same residual variance and the F-statistics in testing for the overall effect of the marker locus. When reduced models are applied, the mc-Fisher model could become inequivalent to its original F_∞ model especially when allelic interactions are involved.

Of the three coding schemes that we have discussed, the F_∞ coding is perhaps the most widely used in current genetic association studies of quantitative traits. From what we have shown, the three coding schemes can essentially lead to equivalent models and have the same power in detection of various genetic effects. In practice, just like the various existing coding schemes such as 'Reference', 'GLM' and 'Effect' that are commonly used in the analysis of categorical covariates [22], we usually only need to adopt one specific coding scheme in building the regression models. Which coding scheme should be applied depends on how convenient it can provide the statistical inferences on the parameters of our research interests. In general, the allele coding models can provide direct estimates of certain substitution effects of alleles and allelic interactions and, in the two-locus case, allele coding models are perhaps the easiest among the three codings in building the relationships between their model parameters and the expected genotypic values. Besides, they are generically linked to the genetic variance components as we have shown above. On the other hand, the allele-count coding models are attractive in that it often leads to simple comparisons among the three genotypic groups with 0, 1 or 2 copies of a particular allele. In the two-locus case, the allele-count coding models also have the definition of their model parameters remain as simple as (if not simpler than) that in the allele coding models even in the presence of epistases. Meanwhile, both the allele and allele-count coding show an advantage that their lower-order main effects in the models can keep the same interpretation regardless of whether there are epistases involved in the model or not. In contrast, the F_∞ coding models

may have the definition of their lower-order main effects vary depending on the absence or presence of epistases in the models. Even though the one-locus F_∞ coding model parameters are closely related to the additive and dominance effects, the two-locus F_∞ coding model parameters including the lower-order main effects have more complicated interpretations than that in the allele or allele-count coding models especially when epistases are involved.

The coding of marker genotypes are not limited to the three allele-based coding schemes that we have discussed. Application of a coding scheme could also be subject to the number of individuals available in each genotype group. For example, under the model framework (7), the allele coding scheme typically creates $w_j(g)$, $v_j(g)$ for each allele type A_j, j = 1, ..., m. When the group of a homozygous genotype A_jA_j includes very few individuals for a particular allele A_j, we may want to combine this genotypic group with another genotype such as the one carrying one copy of the allele A_j. Then we can replace the original $w_j(g)$ and $v_j(g)$ by an allele presence-absence coding variable $d_j(g)$ for this specific allele A_j while keeping two coding variables $w_k(g)$, $v_k(g)$ for other alleles A_k, which leads to a mixed use of the allele coding and this allele presence-absence coding variable. In certain situations, the genotype-based coding could also be very useful as it can provide direct tests on pair-wise comparisons of certain genotypic values. Comparing with the genotype-based coding, the allele coding has the advantage of further dissecting the genetic effects into the allelic effects and allelic interactions, which allow us to specify reduced models with varying degrees of interactions among the main allelic effects - a useful tool in the model building procedures. Given a fixed coding, the likelihood ratio test can be applied to compare a full model with its reduced models. Statistical model selection tools such as AIC and BIC criteria, which provide a balance between the goodness of model fitting to the data and the complexity of the models in terms of the number of parameters, could also be used to compare some non-nested reduced models or frameworks. The current study focuses on establishing the theoretical relationships between the model parameters and the expected genotypic values according to different coding schemes under various model frameworks. A power comparison of some reduced models from different coding schemes under various scenarios with respect to the allele frequencies and possible HWD or LDs between marker alleles is beyond the scope of this study and might be worth of further exploration.

Conclusions

In summary, we introduced three allele-based coding schemes to construct F_∞ models for association testing of multi-allele genetic markers with quantitative traits. Depending upon whether certain allelic effects or comparisons between genotypic groups are of the main research interest, investigators may adopt one of the three allele-based codings (i.e., allele, F_∞ or allele-count), or perhaps a genotype-based coding in building an F_∞ model. Based on the F_∞ model from a given coding scheme, standard regression model fitting tools can then be applied to estimate or test for various genetic effects. Understanding the definition of model parameters from different coding schemes under various model frameworks are crucial for constructing appropriate testing hypothesis and making the correct statistical inferences in the genetic association studies.

Appendices

A. Estimability of parameters in model (8)

Let $G = (G(g_1), ...,G(g_N))^T$ denote a vector of the expected genotypic values of all the individuals in the sample, and $\beta^* = (\mu^*, \alpha_1^*, \ldots, \alpha_m^*, \delta_1^*, \ldots, \delta_m^*)$ be a vector of all the model parameters. We can rewrite model (8) in a matrix form as $G = X\beta^* + e$, where $e = (e_1, ..., e_N)$ and the design matrix X is

$$X = [1_N W_1 W_2 \cdots W_m V_1 V_2 \cdots V_m] \tag{18}$$

with $W_j = (w_j(g_1), ..., w_j(g_N))^T$ and $V_j = (v_j(g_1), ..., v_j(g_N))^T$ for j = 1, ..., m. As every individual carries two and only two alleles at the locus, we have $\sum_{j=1}^{m} W_j = 2 \cdot 1_N$, which means that the first (m+1) column vectors 1_N, W_1, W_2, ..., W_m of the design matrix X are linearly dependent. So, $rank(X) \le 2m$; i.e., X is not a full column rank matrix.

From (7), we have $G_{jk} = \mu^* + \alpha_j^* + \alpha_k^*$ for $j \neq k$, and $G_{jj} = \mu^* + 2\alpha_j^* + \delta_j^*$. If we write $G_0 = (G_{12}, G_{13}, ..., G_{1m}, G_{23}, ..., G_{L\text{-}1,\ L}, G_{11}, ..., G_{mm})^T$, then this model gives

$$G_0 = X_0\beta^* = \begin{bmatrix} \mathbf{1}_s & X_A & \mathbf{0}_{s\times m} \\ \mathbf{1}_m & 2\cdot\mathbf{I}_{m\times m} & \mathbf{I}_{m\times m} \end{bmatrix} \beta^*$$

where $s = m(m - 1)/2$, and

$$X_A = \begin{bmatrix} 1\,1\,0\ldots 0 \\ 1\,0\,1\ldots 0 \\ \vdots\ \vdots\ \vdots\ \ddots\ \vdots \\ 1\,0\,0\ldots 1 \\ 0\,1\,1\ldots 0 \\ \vdots\ \vdots\ \vdots\quad\ \vdots \\ 0\,0\,0\ldots 1 \end{bmatrix}$$

Assume that the genotypes of the sampled individuals cover all possible genotypes A_jA_k for $j, k = 1, ..., m$. Then the design matrix X includes all the row vectors of X_0, which implies that $rank(X) \geq rank(X_0)$. It is clear that $rank(X_0) = m + rank([\mathbf{1}_s\ X_A])$, and it can be shown that $rank([\mathbf{1}_s\ X_A]) = rank(X_A) = m$ when $m \geq 3$. Therefore, $rank(X) = 2m$ as $m \geq 3$. Note that when $m = 2$, we have $s = 1$ and $rank(X) = 3$.

From the linear models theory, we know that for a vector $\lambda = (\lambda_0, ..., \lambda_{2m})^T \in R^{2m+1}$, a linear function $\lambda^T\beta^*$ of β^* is estimable if and only if $\lambda \perp \mathcal{N}(X)$, where $\mathcal{N}(X) = \{c \in R^{2m+1} | Xc = 0\}$ is the null space of the design matrix X. It is also known that $\mathcal{N}(X) \oplus \mathcal{R}(X) = R^{2m+1}$, where $\mathcal{R}(X)$ is a linear space generated by the row vectors of X. Hence, we have $rank(\mathcal{N}(X)) = (2m + 1) - rank(\mathcal{R}(X)) = 1$. Note that $c = (2, -1'_m, 0'_m)' \in \mathcal{N}(X)$ due to the linear dependency among the column vectors 1_N, W_1, W_2, ..., W_m in the design matrix X. Therefore, for a vector $\lambda = (\lambda_0, \lambda_1, ..., \lambda_{2m})^T \in R^{2m+1}$, the linear function $\lambda^T\beta^*$ is estimable if and only if $\lambda \perp c$, or equivalently, $2\lambda_0 = \sum_{j=1}^{m} \lambda_j$. As a result, we know that in model (8) the functions of model parameters $G_{jk} = \mu^* + \alpha_j^* + \alpha_k^*$ for $j \neq k$, and $G_{jj} = \mu^* + 2\alpha_j^* + \delta_j^*$ for $j = 1, ..., m$ are estimable, and the parameters $\delta_j^* = G_{jj} - (\mu^* + 2\alpha_j^*) = G_{jj} + G_{kl} - G_{jl} - G_{jk}$ as $j \neq k, l$ and $k \neq l$ (or in abbreviation, $j \neq k, \neq l$) for $j = 1, ..., m$ are also estimable. But the parameters μ^* and $\alpha_1^*, \ldots, \alpha_m^*$ themselves are not estimable.

B. Estimability of parameters in model (9)

For model (9), we have its design matrix

$$W = [1_N W_1 W_2 \cdots W_{m-1} V_1 V_2 \cdots V_m]$$

where $W_j = (w_j(g_1), ..., w_j(g_N))^T$ and $V_j = (v_j(g_1), ..., v_j(g_N))^T$ for $j = 1, ..., m$. It can be shown that the W and the design matrix X defined in (16) for model (8) have the following relationship $W = XT$ or $X = WS^T$, where

$$T = \begin{bmatrix} I_m & 0 \\ 0_{1\times m} & 0_{1\times m} \\ 0 & I_m \end{bmatrix}$$

and

$$S^T = \begin{bmatrix} I_m & d & 0_{m\times m} \\ 0_{1\times m} & 0 & 0_{1\times m} \\ 0_{m\times m} & 0_{m\times 1} & I_m \end{bmatrix}$$

with $d = (2, -1, ..., -1)' \in R^m$. Let $\beta = (\mu, \alpha_1, ..., \alpha_{m-1}, \delta_1, ..., \delta_m)$. Therefore, as (8) and (9) are two equivalent models, we have $G = X\beta^* = WS^T\beta^* = W\beta$, which yields

$$\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{m-1} \\ \delta_1 \\ \vdots \\ \delta_m \end{bmatrix} = S^T\beta^* = \begin{bmatrix} \mu^* + 2\alpha_m^* \\ \alpha_1^* - \alpha_m^* \\ \vdots \\ \alpha_{m-1}^* - \alpha_m^* \\ \delta_1^* \\ \vdots \\ \delta_m^* \end{bmatrix}$$

From this relationship, we have $\delta_j = \delta_j^*$, $j = 1, ..., m$, which are estimable as shown in Appendix A. Besides, the intercept $\mu = \mu^* + 2\alpha_m^* = G_{jm} + G_{km} - G_{jk}$ and $\alpha_j = \alpha_j^* - \alpha_m^* = G_{jk} - G_{km}$, $k \neq j$, $j = 1, ..., m$ - 1, are also estimable.

C. Relationships for fully parameterized two-locus models

(C.1) Relationships between parameters of the fully parameterized two-locus model (13) and the expected genotypic values are

$$\begin{cases}
\mu = G_{m_1m_1m_2m_2} \\
\alpha_{1j} = G_{jm_1m_2m_2} - \mu = G_{jm_1m_2m_2} - G_{m_1m_1m_2m_2} \\
\alpha_{2r} = G_{m_1m_1rm_2} - \mu = G_{m_1m_1rm_2} - G_{m_1m_1m_2m_2} \\
\delta_{1jk} = G_{jkm_2m_2} - \alpha_{1j} - \alpha_{1k} - \mu \\
\quad = G_{jkm_2m_2} - (G_{jm_1m_2m_2} + G_{m_1km_2m_2}) \\
\quad + G_{m_1m_1m_2m_2} \\
\delta_{2rs} = G_{m_1m_1rs} - \alpha_{2r} - \alpha_{2s} - \mu \\
\quad = G_{m_1m_1rs} - (G_{m_1m_1rm_2} + G_{m_1m_1sm_2}) \\
\quad + G_{m_1m_1m_2m_2} \\
(\alpha_{1j}\alpha_{2r}) = G_{jm_1rm_2} - \alpha_{1j} - \alpha_{2r} - \mu \\
\quad = G_{jm_1rm_2} - (G_{jm_1m_2m_2} + G_{m_1m_1rm_2}) \\
\quad + G_{m_1m_1m_2m_2} \\
(\delta_{1jk}\alpha_{2r}) = G_{jkrm_2} - \alpha_{1j} - \alpha_{1k} - \delta_{1jk} \\
\quad - \alpha_{2r} - (\alpha_{1j}\alpha_{2r}) - (\alpha_{1k}\alpha_{2r}) - \mu \\
\quad = G_{jkrm_2} - (G_{jkm_2m_2} + G_{jm_1rm_2} \\
\quad + G_{km_1rm_2}) + (G_{jm_1m_2m_2} + G_{km_1m_2m_2} \\
\quad + G_{m_1m_1rm_2}) - G_{m_1m_1m_2m_2} \\
(\alpha_{1j}\delta_{2rs}) = G_{jm_1rs} - \alpha_{2r} - \alpha_{2s} - \delta_{2rs} \\
\quad - \alpha_{1j} - (\alpha_{1j}\alpha_{2r}) - (\alpha_{1j}\alpha_{2s}) - \mu \\
\quad = G_{jm_1rs} - (G_{m_1m_1rs} + G_{jm_1rm_2} \\
\quad + G_{jm_1sm_2}) + (G_{jm_1m_2m_2} + G_{m_1m_1rm_2} \\
\quad + G_{m_1m_1sm_2}) - G_{m_1m_1m_1m_2} \\
(\delta_{1jk}\delta_{2rs}) = G_{jkrs} - \alpha_{1j} - \alpha_{1k} - \delta_{1jk} - \alpha_{2r} - \alpha_{2s} \\
\quad - \delta_{2rs} - (\alpha_{1j}\alpha_{2r}) - (\alpha_{1j}\alpha_{2s}) - (\alpha_{1k}\alpha_{2r}) \\
\quad - (\alpha_{1k}\alpha_{2s}) - (\alpha_{1j}\delta_{2rs}) - (\alpha_{1k}\delta_{2rs}) \\
\quad - (\delta_{1jk}\alpha_{2r}) - (\delta_{1jk}\alpha_{2s}) - \mu \\
\quad = G_{jkrs} - (G_{jm_1rs} + G_{km_1rs} + G_{jkrm_2} \\
\quad + G_{jksm_2}) + (G_{jkm_2m_2} + G_{jm_1rm_2} \\
\quad + G_{km_1rm_2} + G_{jm_1sm_2} + G_{km_1sm_2} \\
\quad + G_{m_1m_1rs}) - (G_{jm_1m_2m_2} + G_{km_1m_2m_2} \\
\quad + G_{m_1m_1rm_2} + G_{m_1m_1sm_2}) + G_{m_1m_1m_2m_2}
\end{cases}$$

for $j, k = 1, ..., m_1$ - 1; $r, s = 1, ..., m_2$ - 1 and $j \geq k$, $r \leq s$.

(C.2) Relationships between parameters of the fully parameterized two-locus model (14) and the expected genotypic values are

$$
\begin{cases}
\tau = \mu + \sum_{j=1}^{m_1-1}\left(\alpha_{1j} + \frac{\delta_{1jj}}{2}\right) + \sum_{r=1}^{m_2-1}\left(\alpha_{2r} + \frac{\delta_{2rr}}{2}\right) \\
\quad + \sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1}\left[\left(\alpha_{1j}\alpha_{2r}\right) + \frac{(\alpha_{1j}\delta_{2rr})}{2}\right. \\
\quad \left. + \frac{(\delta_{1jj}\alpha_{2r})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4}\right] \\
\quad = \frac{1}{4}\sum_{j=1}^{m_1-1}\sum_{r=1}^{m_2-1} G_{jjrr} + \frac{(3-m_2)}{4}\sum_{j=1}^{m_1-1} G_{jjm_2m_2} \\
\quad + \frac{(3-m_1)}{4}\sum_{r=1}^{m_2-1} G_{m_1m_1rr} + \frac{(3-m_1)(3-m_2)}{4} G_{m_1m_1m_2m_2} \\
a_{1j} = \alpha_{1j} + \frac{\delta_{1jj}}{2} + \sum_{r=1}^{m_2-1}\left[\left(\alpha_{1j}\alpha_{2r}\right) + \frac{(\alpha_{1j}\delta_{2rr})}{2}\right. \\
\quad \left. + \frac{(\delta_{1jj}\alpha_{2r})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4}\right] \\
\quad = \frac{1}{4}\sum_{r=1}^{m_2-1}\left(G_{jjrr} - G_{m_1m_1rr}\right) \\
\quad + \frac{(3-m_2)}{4}\left(G_{jjm_2m_2} - G_{m_1m_1m_2m_2}\right) \\
a_{2r} = \alpha_{2r} + \frac{\delta_{2rr}}{2} + \sum_{j=1}^{m_1-1}\left[\left(\alpha_{1j}\alpha_{2r}\right) + \frac{(\alpha_{1j}\delta_{2rr})}{2}\right. \\
\quad \left. + \frac{(\delta_{1jj}\alpha_{2r})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4}\right] \\
\quad = \frac{1}{4}\sum_{j=1}^{m_1-1}\left(G_{jjrr} - G_{jjm_2m_2}\right) \\
\quad + \frac{(3-m_1)}{4}\left(G_{m_1m_1rr} - G_{m_1m_1m_2m_2}\right) \\
d_{1jj} = -\frac{\delta_{1jj}}{2} - \sum_{r=1}^{m_2-1}\left[\frac{(\delta_{1jj}\alpha_{2r})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4}\right] \\
\quad = -\frac{1}{4}\sum_{r=1}^{m_2-1}\left(G_{jjrr} - 2G_{jm_1rr} + G_{m_1m_1rr}\right) \\
\quad - \frac{(3-m_2)}{4}\left(G_{jjm_2m_2} - 2G_{jm_1m_2m_2} + G_{m_1m_1m_2m_2}\right) \\
d_{2rr} = -\frac{\delta_{2rr}}{2} - \sum_{j=1}^{m_1-1}\left[\frac{(\alpha_{1j}\delta_{2rr})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4}\right] \\
\quad = -\frac{1}{4}\sum_{j=1}^{m_1-1}\left(G_{jjrr} - 2G_{jjrm_1} + G_{jjm_2m_2}\right) \\
\quad - \frac{(3-m_1)}{4}\left(G_{m_1m_1rr} - 2G_{m_1m_1rm_2} + G_{m_1m_1m_2m_2}\right) \\
d_{1jk} = \delta_{1jk} + \sum_{r=1}^{m_2-1}\left[\left(\delta_{1jk}\alpha_{2r}\right) + \frac{(\delta_{1jk}\delta_{2rr})}{2}\right] \\
\quad = \frac{1}{2}\sum_{r=1}^{m_2-1}\left(G_{jkrr} - G_{jm_1rr} - G_{km_1rr} + G_{m_1m_1rr}\right) \\
\quad + \frac{(3-m_2)}{4}\left(G_{jkm_2m_2} - G_{jm_1m_2m_2} - G_{km_1m_2m_2}\right. \\
\quad \left. + G_{m_1m_1m_2m_2}\right), j < k \\
d_{2rs} = \delta_{2rs} + \sum_{j=1}^{m_1-1}\left[\left(\alpha_{1j}\delta_{2rs}\right) + \frac{(\delta_{1jj}\delta_{2rr})}{2}\right] \\
\quad = \frac{1}{2}\sum_{j=1}^{m_1-1}\left(G_{jjrs} - G_{jjrm_2} - G_{jjsm_2} + G_{jjm_2m_2}\right) \\
\quad + \frac{(3-m_1)}{4}\left(G_{m_1m_1rr} - G_{m_1m_1rm_2} - G_{m_1m_1sm_2}\right. \\
\quad \left. + G_{m_1m_1m_2m_2}\right), r < s \\
\left(a_{1j}a_{2r}\right) = \left(\alpha_{1j}\alpha_{2r}\right) + \frac{(\alpha_{1j}\delta_{2rr})}{2} + \frac{(\delta_{1jj}\alpha_{2r})}{2} + \frac{(\delta_{1jj}\delta_{2rr})}{4} \\
\quad = \frac{1}{4}\left(G_{jjrr} - G_{m_1m_1rr} - G_{jjm_2m_2} + G_{m_1m_1m_2m_2}\right) \\
\left(a_{1j}d_{2rr}\right) = -\frac{(\alpha_{1j}\delta_{2rr})}{2} - \frac{(\delta_{1jj}\delta_{2rr})}{4} \\
\quad = -\frac{1}{4}\left(G_{jjrr} - 2G_{jjrm_2} + G_{jjm_2m_2}\right) \\
\quad + \frac{1}{4}\left(G_{m_1m_1rr} - 2G_{m_1m_1rm_2} + G_{m_1m_1m_2m_2}\right) \\
\left(d_{1jj}a_{2r}\right) = -\frac{(\delta_{1jj}\alpha_{2r})}{2} - \frac{(\delta_{1jj}\delta_{2rr})}{4} \\
\quad = -\frac{1}{4}\left(G_{jjrr} - 2G_{jm_1rr} + G_{m_1m_1rr}\right) \\
\quad + \frac{1}{4}\left(G_{jjm_2m_2} - 2G_{jm_1m_2m_2} + G_{m_1m_1m_2m_2}\right) \\
\left(a_{1j}d_{2rs}\right) = \left(\alpha_{1j}\delta_{2rs}\right) + \frac{(\delta_{1jj}\delta_{2rs})}{2} \\
\quad = \frac{1}{2}\left(G_{jjrs} - G_{jjrm_2} - G_{jjsm_2} + G_{jjm_2m_2}\right) \\
\quad - \frac{1}{2}\left(G_{m_1m_1rs} - G_{m_1m_1rm_2} - G_{m_1m_1sm_2}\right. \\
\quad \left. + G_{m_1m_1m_2m_2}\right), r < s \\
\left(d_{1jk}a_{2r}\right) = \left(\delta_{1jk}\alpha_{2r}\right) + \frac{(\delta_{1jj}\delta_{2rr})}{2} \\
\quad = \frac{1}{2}\left(G_{jkrr} - G_{jm_1rr} - G_{km_1rr} + G_{m_1m_1rr}\right) \\
\quad - \frac{1}{2}\left(G_{jkm_2m_2} - G_{jm_1m_2m_2} - G_{km_1m_2m_2}\right. \\
\quad \left. + G_{m_1m_1m_2m_2}\right), j < k \\
\left(d_{1jj}d_{2rs}\right) = -\frac{(\delta_{1jj}\delta_{2rs})}{2}, r < s \\
\left(d_{1jk}d_{2rr}\right) = -\frac{(\delta_{1jk}\delta_{2rr})}{2}, j < k \\
\left(d_{1jj}d_{2rr}\right) = \frac{(\delta_{1jj}\delta_{2rr})}{4} \\
\left(d_{1jk}d_{2rs}\right) = \left(\delta_{1jk}\delta_{2rs}\right), j < k, r < s
\end{cases}
$$

for j, k = 1, ..., m_1 - 1 and r, s = 1, ..., m_2 - 1, where the relationships between the parameters of model (14) and model (13) are built based on the equivalency between the two models. The relationships between the parameters of model (14) and the expected genotypic values can then be derived by replacing the parameters of model (13) with the expected genotypic values from the previous established results in (C.1).

(C.3) Relationships between parameters of the fully parameterized two-locus model (15) and the expected genotypic values are

$$
\begin{cases}
\pi_0 = \mu = G_{m_1m_1m_2m_2} \\
\pi_{1j} = \alpha_{1j} = G_{jm_1m_2m_2} - G_{m_1m_1m_2m_2} \\
\pi_{2r} = \alpha_{2r} = G_{m_1m_1rm_2} - G_{m_1m_1m_2m_2} \\
\eta_{1jj} = 2\alpha_{1j} + \delta_{1jj} = G_{jjm_2m_2} - G_{m_1m_1m_2m_2} \\
\eta_{2rr} = 2\alpha_{2r} + \delta_{2rr} = G_{m_1m_1rr} - G_{m_1m_1m_2m_2} \\
\eta_{1jk} = \delta_{1jk}, j < k;\ \eta_{2rs} = \delta_{2rs}, r < s \\
\left(\pi_{1j}\pi_{2r}\right) = \left(\alpha_{1j}\alpha_{2r}\right) = G_{jm_1rm_2} \\
\qquad - \left(G_{jm_1m_2m_2} + G_{m_1m_1rm_2}\right) + G_{m_1m_1m_2m_2} \\
\left(\pi_{1j}\eta_{2rr}\right) = 2\left(\alpha_{1j}\alpha_{2r}\right) + \left(\alpha_{1j}\delta_{2rr}\right) = G_{jm_1rr} \\
\qquad - \left(G_{jm_1m_2m_2} + G_{m_1m_1rr}\right) + G_{m_1m_1m_2m_2} \\
\left(\pi_{1j}\eta_{2rs}\right) = \left(\alpha_{1j}\delta_{2rs}\right), r < s \\
\left(\eta_{1jj}\pi_{2r}\right) = 2\left(\alpha_{1j}\alpha_{2r}\right) + \left(\delta_{1jj}\alpha_{2r}\right) = G_{jjrm_2} \\
\qquad - \left(G_{jjm_2m_2} + G_{m_1m_1rm_2}\right) + G_{m_1m_1m_2m_2} \\
\left(\eta_{1jk}\pi_{2r}\right) = \left(\delta_{1jk}\alpha_{2r}\right), j < k \\
\left(\eta_{1jj}\eta_{2rr}\right) = 4\left(\alpha_{1j}\alpha_{2r}\right) + 2\left(\alpha_{1j}\delta_{2rr}\right) \\
\qquad + 2\left(\delta_{1jj}\alpha_{2r}\right) + \left(\delta_{1jj}\delta_{2rr}\right) \\
\qquad = G_{jjrr} - \left(G_{jjm_2m_2} + G_{m_1m_1rr}\right) \\
\qquad + G_{m_1m_1m_2m_2} \\
\left(\eta_{1jj}\eta_{2rs}\right) = 2\left(\alpha_{1j}\delta_{2rs}\right) + \left(\delta_{1jj}\delta_{2rs}\right) \\
\qquad = G_{jjrs} - \left(G_{m_1m_1rs} + G_{jjrm_2} + G_{jjsm_2}\right) \\
\qquad + \left(G_{m_1m_1rm_2} + G_{m_1m_1sm_2} + G_{jjm_2m_2}\right) \\
\qquad - G_{m_1m_1m_2m_2}, r < s \\
\left(\eta_{1jk}\eta_{2rr}\right) = 2\left(\delta_{1jk}\alpha_{2r}\right) + \left(\delta_{1jk}\delta_{2rr}\right) \\
\qquad = G_{jkrr} - \left(G_{jkm_2m_2} + G_{jm_1rr} + G_{km_1rr}\right) \\
\qquad + \left(G_{jm_1m_2m_2} + G_{km_1m_2m_2} + G_{m_1m_1rr}\right) \\
\qquad - G_{m_1m_1m_2m_2}, j < k \\
\left(\eta_{1jk}\eta_{2rs}\right) = \left(\delta_{1jk}\delta_{2rs}\right), j < k, r < s
\end{cases}
$$

for j, k = 1, ..., m_1 - 1 and r, s = 1, ..., m_2 - 1, where the relationships between the parameters of model (15) and model (13) are built based on the equivalency between the two models. The relationships between the parameters of model (15) and the expected genotypic values are then derived by replacing the parameters of model (13) with the expected genotypic values from the previous established results in (C.1).

Authors' contributions

TW planned the study, conducted the derivation and wrote the manuscript.

Received: 31 May 2011 Accepted: 21 September 2011
Published: 21 September 2011

References

1. Fisher RA: The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinburgh* 1918, **52**:399-433.

2. Cockerham CC: **An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present.** *Genetics* 1954, **39**:859-882.
3. Cockerham CC: **Estimation of genetic variances.** In *Statistical Genetics and Plant Breeding Natl Acad Sci Natl Res.* Edited by: Henson WD, Robinson HF. Council publ No. 982. Washington, D.C.; 1963:53-94.
4. Weir BS, Cockerham C: **Two-locus theory in quantitative genetics.** In *Proceedings of the international conference on quantitative genetics.* Edited by: Pollack EBT Kempthorne O. Iowa State University Press; 1977:247-269.
5. Kempthorne O: *An introduction to Genetic Statistics* New Haven: Iowa State University Press, Ames; 1969.
6. Wang T, Zeng ZB: **Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium.** *BMC Genetics* 2006, **7**:Article 9.
7. Hansen TF, Wagner GP: **Modeling genetic architecture: a multilinear theory of gene interaction.** *Theor Popul Biol* 2001, **59**:61-86.
8. Alvarez-Castro JM, Carlborg O: **A unified model for functional and statistical epistasis and its application in quantitative trait Loci analysis.** *Genetics* 2007, **176(2)**:1151-1167.
9. Zeng ZB, Wang T, Zou W: **Modeling quantitative trait Loci and interpretation of models.** *Genetics* 2005, **169(3)**:1711-1725.
10. Wang T, Zeng ZB: **Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium.** *BMC Genetics* 2009, **10**:Article 52.
11. Yang RC, Alvarez-Castro JM: **Functional and statistical genetic effects with multiple alleles.** *Current Topics in Genetics* 2008, **3**:49-62.
12. Lynch M, Walsh B: *Genetics and Analysis of Quantitative Traits* Sunderland, MA: Sinauer; 1998.
13. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: **Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals.** *Hum Hered* 2002, **53(2)**:79-91.
14. Searle SR: *Linear Models* John Wiley & Sons Inc., New York, NY; 1971.
15. Ravishanker N, Dey DK: *A First Course in Linear Model Theory* Chapman & Hall, CRC, Boca Raton, Florida; 2002.
16. Van Der Veen JH: **Tests of non-allelic interaction and linkage for quantitative characters in generations derived from two diploid pure lines.** *Genetica* 1959, **30**:201-232.
17. Mather K, Jinks JL: *Biometrical Genetics.* 3 edition. Landon: Chapman and Hall; 1982.
18. Falconer DS, Mackay TFC: *Introduction to Quantitative Genetics.* fourth edition. Harlow, UK: Longman; 1996.
19. Hayman BI, Mather KM: **The description of genetic interactions in continuous variation.** *Biometrics* 1955, **11**:69-82.
20. Liu T, Johnson JA, Casella G, Wu R: **Sequencing complex diseases with HapMap.** *Genetics* 2004, **168**:503-511.
21. Searle SR, Casella G, McCulloch CE: *Variance Components* John Wiley & Sons, NIC, Hoboken, NJ; 1992.
22. Stokes ME, Davis CS, Koch GG: *Categorical Data Analysis using the SAS System.* 2 edition. SAS Institute Inc., Cary, NC; 2001.

doi:10.1186/1471-2156-12-82
Cite this article as: Wang: On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. *BMC Genetics* 2011 **12**:82.