FRONTIERS IN GENETICS

ORIGINAL RESEARCH ARTICLE
published: 25 September 2014
doi: 10.3389/fgene.2014.00328

# A revised Fisher model on analysis of quantitative trait loci with multiple alleles

***Tao Wang****

*Division of Biostatistics, Institute for Health and Society, Medical College of Wisconsin, Milwaukee, WI, USA*

***Edited by:***
*Qizhai Li, Chinese Academy of Sciences, China*

***Reviewed by:***
*Guohua Zou, Chinese Academy of Sciences, China*
*Tian-Qing Zheng, Chinese Academy of Agricultural Sciences, China*

***Correspondence:**
*Tao Wang, Division of Biostatistics, Institute for Health and Society, Medical College of Wisconsin, 8701 Watertown Plank Road, PO Box 26509, Milwaukee, WI 53226, USA*
*e-mail: taowang@mcw.edu*

Zeng et al. (2005) proposed a general two-allele (G2A) model to model bi-allelic quantitative trait loci (QTL). Comparing with the classical Fisher model, the G2A model can avoid using redundant parameters and be fitted directly using standard least square (LS) approach. In this study, we further extend the G2A model to general multi-allele (GMA) model. First, we propose a one-locus GMA model for phase known genotypes based on modeling the inheritance of paternal and maternal alleles. Next, we develop a one-locus GMA model for phase unknown genotypes by treating it as a special case of the phase known one-locus GMA model. Thirdly, we extend the one-locus GMA models to multiple loci. We discuss how the genetic variance components can be analyzed using these GMA models in equilibrium as well as disequilibrium populations. Finally, we apply the GMA model to a published experimental data set.

**Keywords: Fisher's genetic model, genetic variance components, general two-allele model, general multi-allele model, orthogonality, least square approach, average allelic effects and interactions**

## 1. INTRODUCTION

Currently there are two types of statistical genetic models that are commonly used in genetic analysis of quantitative traits. One is the $F_\infty$ type models that concentrate on direct modeling of the expected genotypic values at targeted quantitative trait loci (QTL) or genetic markers and association testing for various allelic effects and interactions (Fisher, 1918; Cheverud, 2000; Hansen and Wagner, 2001; Wang, 2011). Another is Fisher's analysis of variance (ANOVA) models that focus on assessing variations contributed by some genetic components (i.e., grouped allelic effects or allelic interactions) at targeted QTL or genetic markers (Fisher, 1918; Cockerham, 1954, 1963; Kempthorne, 1969; Weir and Cockerham, 1977; Wang and Zeng, 2006). A considerable amount of discussion has been made about the distinction between these two different types of genetic models (Zeng et al., 2005; Álvarez-Castro and Carlborg, 2007; Yang and Álvarez-Castro, 2008; Wang and Zeng, 2009). More recently, a comprehensive review of various genetic models was also given in Álvarez-Castro (2012).

In genetic association studies, we are often interested in direct comparison of the expected genotypic values at certain QTL or marker loci. The $F_\infty$ models are appealing in this setting due to their simplicity in interpretation of their model parameters, which are often referred as the fixed genetic effects such as the additive and dominance effects or the allelic effects and allelic interactions in terms of the expected genotypic values. By applying the $F_\infty$ models, we can compare the expected genotypic values via hypothesis tests on various fixed genetic effects. However, as pointed out in Wang and Zeng (2009), the $p$-value based association tests on these fixed genetic effects could highly depend upon the regression model, the distribution assumption and the available sample size. Besides, a statistically significant genetic effect with a small enough $p$-value may not necessarily imply a clinically important finding. On the other hand, the Fisher type models allow us to assess the genetic variations contributed by certain genetic components to the overall variation of a quantitative trait. By definition, these genetic variance components do not depend on the sample size, and they can provide additional information on better understanding the genetic etiology and assessing for clinical importance. Nonetheless, both the $F_\infty$ and the Fisher type models form basis in the analysis of quantitative traits. They provide different perspectives in assessing the genetic effects of QTL or genetic markers.

The basic genetic model on assessing the genetic variance components was first proposed by Fisher (1918). Cockerham (1954, 1963) extended Fisher's one-locus model to two bi-allelic QTL with a particular focus on epistatic variance components. Kempthorne (1954, 1957) further extended the two-locus model to multiple alleles. Wang and Zeng (2006) also explored the Fisher type multi-allele two-locus model on partition of the genotypic variance in the presence of linkage disequilibrium (LD). With their model parameters referred as the average effect of the gene substitution (see Falconer and Mackay, 1996) or the average allelic effects and interactions (Wang and Zeng, 2009), these classical Fisher type models often contain constraints on their model parameters due to an over-parameterization of the expected genotypic values. These constraints could make it difficult to fit this type of models using standard least square (LS) regression approach. To avoid this problem, under a regression model framework, Zeng et al. (2005) proposed a general two-allele (G2A) model for variance component analysis on bi-allelic QTL. The G2A model retains the nice property of the classical Fisher's model on orthogonal partition of the genotypic variance in equilibrium populations. Meanwhile, without using redundant parameters, the G2A model can be fitted directly using the standard LS approach. Wang and Zeng (2009) further explored the origin of the G2A model and clarified its theoretical basis

on pertaining the orthogonal partition of the genotypic variance in equilibrium populations. Álvarez-Castro and Carlborg (2007) also proposed a different way for re-parameterization of the expected genotypic values, which the authors referred as the natural and orthogonal interactions (NOIA) model. It appears that the NOIA model first defines the genetic effects in terms of the expected genotypic values, and then derives the design matrix for the genetic effects via an inverse of the matrix. Though the matrix formulation is helpful for facilitating the transformation between different parameterizations, making inverse of the matrix to construct the design matrix is difficult to implement analytically. For multiple QTL, how to define various genetic effects in terms of the expected genotypic values could also be a challenge especially when locus-by-locus interactions are involved. Besides, the NOIA model assumes that the design matrix is of full rank, which makes it unsuitable for reduced re-parameterization of the genotypic values. Currently, no NOIA model has been proposed for reduced re-parameterization of the expected genotypic values, or for multiple QTL in the presence of LD.

In this study, we further extend the G2A model to QTL with multiple alleles and multiple loci. In bi-allelic case, only one additive effect and one dominance effect are needed at each locus, and the locus-by-locus interactions can be easily included for constructing a full re-parameterization of the genotypic values. For one QTL with multiple alleles, how to define the dominance effects for various allelic interactions is not straightforward especially when phases of its genotypes are unknown. The extension to multiple loci is also cumbersome by the much more complex structure of locus-by-locus interactions. How to present the model and define various genetic variance components are not trivial tasks. To construct one-locus general multi-allele (GMA) model, we overcome the phase problem by appropriately merging the paternal and maternal allelic effects and allelic interactions in the phase-known situation. Typically, with phase unknown genotypes at a locus, we may have to assume that the paternal and maternal alleles have the same frequencies and contribute the same genetic effects so that we could merge them without distinguishing their parental origins. With phase-known genotypes, we can further break down the additive variance component into paternal and maternal variance components. For multiple QTL with multiple alleles, we develop concise expressions for constructing multi-locus GMA models and defining various genetic components. Explicit formulas for calculating various genetic variance components in equilibrium population are also derived.

The structure of this manuscript is organized as the following. First, we consider one multi-allele QTL with phase known genotypes. Following the same strategy as adopted in Wang and Zeng (2009), we start by introducing indicator variables that describe the inheritance of paternal and maternal alleles. Then we make mean corrections on these indicator variables and build one-locus GMA model based on the mean-corrected index variables. Next, we consider one QTL with phase unknown genotypes. We construct a one-locus GMA model for unphased genotypes by appropriately combining the paternal and maternal allelic effects in the phase known one-locus GMA model. We derive formulas on partitioning the genotypic variance into the additive and dominance variance components under Hardy-Weinberg equilibrium (HWE) as well as in Hardy-Weinberg disequilibrium (HWD) for both the phase-known and phase-unknown one-locus GMA models. Thirdly, we extend the one-locus GMA models to multiple loci with either phase known or phase unknown genotypes. Based on these multi-locus GMA models, we describe how the various genetic components can be defined. An orthogonal partition on the genotypic variance in an equilibrium population is also presented in both the phase known and unknown cases. In addition, we discuss how to construct the reduced multi-locus GMA models in practice. The difference in using $F_\infty$ models and GMA models to build reduced models for the expected genotypic values is explored. Finally, we apply the GMA model to a published experimental data set.

## 2. METHODS AND RESULTS

The variation of a quantitative trait $Y$ is usually assumed to be contributed by both genetic and environmental effects. Let $G$ denote the true (unobservable) genotypic value from the joint contribution of all the genetic factors to the quantitative trait $Y$. Given genotypes at targeted QTL or genetic marker loci, we focus on assessing the variations (i.e., variances) contributed by the allelic effects and interactions of the QTL or marker loci to the total genotypic variance $V_G$.

### 2.1. ONE-LOCUS MODEL FOR PHASE KNOWN

First, let us consider a single QTL with multiple alleles $A_1, \ldots, A_m$ ($m \geq 2$). Suppose we know the parental origins of the alleles for observed genotypes. Then we can distinguish the paternal and maternal allelic effects separately. With $m$ alleles, there are in total $m^2$ possible phased genotypes: $(A_i, A_j)$, $i, j = 1, \ldots, m$, where $A_i$ and $A_j$ represent the paternal and maternal allele, respectively. Let $G_j^i = E[G|g = (A_i, A_j)]$ denote the expected genotypic value for a phased genotype $(A_i, A_j)$ in a study population. Then there are totally $m^2$ possible expected genotypic values $G_j^i$, $i, j = 1, \ldots, m$. To model these expected genotypic values, Fisher's classical one-locus model is given by

$$G_j^i = \mu + \alpha^i + \alpha_j + \delta_j^i, \ i, j = 1, \ldots, m, \tag{1}$$

where $\alpha^i$ (or $\alpha_j$) is the so-called *average* additive or main allelic effect of a paternal (or maternal) allele $A_i$ (or $A_j$), and $\delta_j^i$ is the *average* allelic interaction between a paternal allele $A_i$ and a maternal allele $A_j$. The above model is a typical two-way ANOVA model with the paternal and maternal gametes being treated as two independent risk factors. Although the paternal and maternal gametes often share (but do not have to) the same set of alleles $A_1, \ldots, A_m$ at the QTL, it allows the paternal and maternal gametes to have different allele frequencies and allelic effects at the QTL.

Let $p^i$ be the frequency of allele $A_i$ on paternal gametes $\left(\sum_{i=1}^m p^i = 1\right)$, and $p_j$ be the frequency of allele $A_j$ on the maternal gametes $\left(\sum_{j=1}^m p_j = 1\right)$. One nice feature of the Fisher model above is that it can assess the additive and dominance variance components $V_A$ and $V_D$, which are defined as variations contributed by the additive allelic effects from the paternal and maternal alleles and the allelic interactions between the paternal or maternal alleles, respectively. For example, under HWE,

it is well known that the total genotypic variance $V_G = \sum_{i,j} p^i p_j (G^i_j - \mu)^2$ has an orthogonal partition $V_G = V_A + V_D$, where

$$V_A = \sum_i p^i (\alpha^i)^2 + \sum_j p_j (\alpha_j)^2 \,, \quad V_D = \sum_{i,j} p^i p_j (\delta^i_j)^2.$$

Note that there are in total $m^2 + 2m + 1 = (m+1)^2$ parameters being involved in Fisher model (1) including the intercept $\mu$, which is more than the total number $m^2$ of the expected genotypic values $G^i_j$, $i, j = 1, \ldots, m$. As a result, not all the model parameters are estimable. To avoid this problem, some constraints on the model parameters are often required. It is usually assumed that all the genetic effects are averaged to zero over any index; i.e.,

$$\sum_i p^i \alpha^i = 0, \quad \sum_j p_j \alpha_j = 0, \quad \sum_i p^i \delta^i_j = 0, \quad \sum_j p_j \delta^i_j = 0. \tag{2}$$

With these constraints, Fisher (1918) showed that the least square estimates (LSE) of the model parameters are given by

$$\mu = E(G),\ \alpha_i = G^i_{\cdot} - G^{\cdot}_{\cdot}\,,\ \alpha_j = G^{\cdot}_j - G^{\cdot}_{\cdot}\,,\ \delta^i_j = G^i_j - G^i_{\cdot} - G^{\cdot}_j + G^{\cdot}_{\cdot}$$

where $G^{\cdot}_{\cdot} = E(G)$, $G^i_{\cdot} = E[G|g = (A_i, -)]$ and $G^{\cdot}_j = E[G|g = (-, A_j)]$. In simple cases, we could estimate the model parameters using these formulas. In general, however, those "irregular" constraints in (2) make it difficult to fit model (1) using the standard LS approach via commonly used software such as SAS (SAS Institute INC, Raleigh, NC), especially when we need to adjust for certain environmental covariates.

Here we propose a way to get rid of the redundant parameters. Let us first introduce the following indicator variables that describe the transmission inheritance of the paternal and maternal alleles.

$$z_{P_i}(g) = \begin{cases} 1, & \text{the paternal allele is } A_i \\ 0, & \text{the paternal allele is not } A_i \end{cases},$$

$$z_{M_j}(g) = \begin{cases} 1, & \text{the maternal allele is } A_j \\ 0, & \text{the maternal allele is not } A_j \end{cases},$$

for $i, j = 1, 2, \ldots, m$ at the QTL. Next, using the same strategy as adopted in Wang and Zeng (2006), we further make mean corrections on these indicator variables $z_{P_i}$, $z_{M_j}$ and define the following mean-corrected index variables

$$x_{P_i}(g) = z_{P_i}(g) - E[z_{P_i}(g)] = \begin{cases} 1 - p^i, & \text{the paternal allele is } A_i \\ -p^i, & \text{the paternal allele is not } A_i \end{cases},$$

$$x_{M_j}(g) = z_{M_j}(g) - E[z_{M_j}(g)] = \begin{cases} 1 - p_j, & \text{the maternal allele is } A_j \\ -p_j, & \text{the maternal allele is not } A_j \end{cases}.$$

Then we can re-write the Fisher model (1) as

$$E(G|g) = \mu + \sum_{i=1}^{m} \alpha^i x_{P_i}(g) + \sum_{j=1}^{m} \alpha_j x_{M_j}(g) + \sum_{i,j} \delta^i_j x_{P_i}(g) x_{M_j}(g), \tag{3}$$

where $E(G|g)$ denotes the expected genotypic value given a genotype $g$ at the targeted QTL. As shown in Wang and Zeng (2006), the above model is simply a different representation of the Fisher model (1) with the same constraints being applied on the model parameters. However, as the HWD measurements can be quantified as covariances between those mean-corrected index variables $x_{P_i}$'s and $x_{M_j}$'s, this model expression can facilitate derivation on examining the genetic variance components under HWD.

Now, we further exclude the redundant parameters in model (3). For a diploid subject such as human being, his or her genotype at a locus on a pair of homologous chromosomes consists of two alleles with one from the father and the other one from the mother. Therefore, we always have $\sum_{i=1}^{m} z_{P_i}(g) \equiv 1$ and $\sum_{j=1}^{m} z_{M_j}(g) \equiv 1$. Thus, $x_{P_m} = -\sum_{i=1}^{m-1} x_{P_i}$ and $x_{M_m} = -\sum_{j=1}^{m-1} x_{M_j}$. So we can simply replace $x_{P_m}$ by $(-\sum_{i=1}^{m-1} x_{P_i})$, and $x_{M_m}$ by $(-\sum_{j=1}^{m-1} x_{M_j})$ in model (3), which leads to the following revised Fisher model

$$E(G|g) = \mu + \sum_{i=1}^{m-1} \alpha^{*i} x_{P_i}(g) + \sum_{j=1}^{m-1} \alpha_{*j} x_{M_j}(g) + \sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \delta^{*i}_{*j} x_{P_i}(g) x_{M_j}(g). \tag{4}$$

Model (4) provides a full re-parameterization of the $m^2$ expected genotypic values $G^i_j$, $i, j = 1, \ldots, m$, without using redundant parameters. We refer it as a revised one-locus Fisher model or a general multi-allele (GMA) model. We also refer its model parameters $\alpha^{*i}$ (or $\alpha_{*j}$) as the average (additive) allelic effect of the paternal (or maternal) allele $A_i$ (or $A_j$), and $\delta^{*i}_{*j}$ the average allelic interaction between the paternal allele $A_i$ and maternal allele $A_j$, with respect to the baseline allele $A_m$. Here, we choose allele $A_m$ as the baseline allele but it could be any other allele as well. It is easy to show that, in terms of the parameters in the original Fisher model (1), we have $\alpha^{*i} = \alpha^i - \alpha^m$, $\alpha_{*j} = \alpha_j - \alpha_m$ and $\delta^{*i}_{*j} = \delta^i_j - \delta^m_j - \delta^i_m + \delta^m_m$, for $i, j = 1, \ldots, m$.

Model (4) retains most of the nice features of the original Fisher model (1) on partition of the genotypic variance. Based on this model, we have the genetic additive variance components $V_{A_P} = \mathrm{Var}(\sum_{i=1}^{m-1} \alpha^{*i} x_{P_i})$ and $V_{A_M} = \mathrm{Var}(\sum_{j=1}^{m-1} \alpha_{*j} x_{M_j})$, which denote variations contributed by the additive effects of the paternal and maternal alleles, respectively. The genetic dominant variance component $V_D = \mathrm{Var}(\sum_{i=1}^{m-1} \sum_{j=1}^{m-1} \delta^{*i}_{*j} x_{P_i} x_{M_j})$, which accounts for a variation contributed by the interactions between paternal and maternal alleles in addition to the additive variance components. Under HWE, the paternal index variables $\{x_{P_i}, i = 1, \ldots, m\}$ are independent of the maternal index variables $\{x_{M_j}, j = 1, \ldots, m\}$.

Therefore, we have $\mu = E(G)$ and an orthogonal partition on the variance of the expected genotypic values: $V(E(G|g)) = V_{A_P} + V_{A_M} + V_D$, where

$$V_{A_P} = \sum_{i=1}^{m-1} p^i(\alpha^{*i})^2 - \left(\sum_{i=1}^{m-1} p^i\alpha^{*i}\right)^2,$$

$$V_{A_M} = \sum_{j=1}^{m-1} p_j(\alpha_{*j})^2 - \left(\sum_{j=1}^{m-1} p_j\alpha_{*j}\right)^2$$

$$V_D = \sum_{i,j=1}^{m-1} p^i p_j(\delta_{*j}^{*i})^2 - \sum_{i=1}^{m-1} p^i\left(\sum_{j=1}^{m-1} p_j\delta_{*j}^{*i}\right)^2 - \sum_{j=1}^{m-1} p_j\left(\sum_{i=1}^{m-1} p^i\delta_{*j}^{*i}\right)^2 + \left(\sum_{i,j=1}^{m-1} p^i p_j\delta_{*j}^{*i}\right)^2.$$

In HWD, the disequilibrium measurements can be captured by the covariances between the index variables $x_{P_i}$'s and $x_{M_j}$'s. In this case, we no longer have an orthogonal partition on the variance of the genotypic values for the additive and dominance variance components (see **Appendix A**). It should be pointed out that although the model parameters defined in model (4) depend upon the choice of the reference allele $A_m$, the additive and dominant variance components $V_{A_P}$, $V_{A_M}$, $V_D$ as well as their covariance components do not depend on such a choice.

Note that the partition of the total genotypic variance $V_G$ based on the Fisher model (1) assumes that the genotypic values are completely determined by the QTL. In general, beyond the targeted QTL, the genotypic value $G$ could also be affected by other unobserved QTL. In this case, the total genotypic variance $V_G = V(E(G|g)) + E(V(G|g))$, where $V(G|g)$ is a variation that cannot be explained by the targeted QTL. So the Fisher model (1), or its revised model (4), actually provides a partition on the variance $V(E(G|g))$ of the expected genotypic values at the targeted QTL instead of the total genotypic variance $V_G$. In practice, given a random sample from the study population, if we ignore the genetic by environmental interaction, a quantitative trait can typically be expressed in a linear regression model form as

$$y_k = \beta z_k + E(G|g_k) + \epsilon_k,\ k = 1, \ldots, n,$$

where, for $k = 1, \ldots, n$, $g_k$ is the observed QTL genotype of subject $k$, $z_k$ is a vector of subject $k$ from some adjusted environmental covariates with fixed effects $\beta$, and $\epsilon_k$ is a model residual contributed by other environmental and genetic factors that cannot be captured by $z_k$ and $g_k$ at the targeted QTL. Assuming that $\epsilon_k$, $k = 1, \ldots, n$, are independent and identically distributed (i.i.d) with a variance $\sigma_\epsilon^2$, we have $V_y = V(E(G|g)) + \sigma_\epsilon^2$. To further partition $V(E(G|g))$ based on the QTL genotypes, we can first calculate the allele frequencies and compute the values of the index variables $\{x_{P_i}(g_k), x_{M_j}(g_k)\}$ for each subject $k = 1, \ldots, n$, in the sample. Then, by treating these index variables as fixed covariates, we can incorporate the GMA model (4) into the regression model above and fit the model using the standard LS approach. Based on the LSE $\hat{\alpha}^{*i}$, $\hat{\alpha}_{*j}$ and $\hat{\delta}_{*j}^{*i}$ of the model parameters, we can compute the additive and dominance genetic components $A_P(k) = \sum_{i=1}^{m-1}\hat{\alpha}^{*i}x_{P_i}(g_k)$, $A_M(k) = \sum_{j=1}^{m-1}\hat{\alpha}_{*j}x_{M_j}(g_k)$ and $D(k) = \sum_{i=1}^{m-1}\sum_{j=1}^{m-1}\hat{\delta}_{*j}^{*i}x_{P_i}(g_k)x_{M_j}(g_k)$, for $k = 1, \ldots, n$. Finally, the genetic variance components $V_{A_P}$, $V_{A_M}$ and $V_D$ can be estimated as the sample variances of $\{A_P(k), k = 1, \ldots, n\}$, $\{A_M(k), k = 1, \ldots, n\}$, and $\{D(k), k = 1, \ldots, n\}$, respectively. Meanwhile, the genetic covariance components $\mathrm{Cov}(A_P, A_M)$, $\mathrm{Cov}(A_P, D)$, and $\mathrm{Cov}(A_M, D)$ can be estimated through the sample covariances

$$\widehat{\mathrm{Cov}(A_P, A_M)} = \frac{1}{n}\sum_{k=1}^{n}(A_P(k) - \bar{A}_P)(A_M(k) - \bar{A}_M),$$

$$\widehat{\mathrm{Cov}(A_P, D)} = \frac{1}{n}\sum_{k=1}^{n}(A_P(k) - \bar{A}_P)(D(k) - \bar{D}),$$

$$\widehat{\mathrm{Cov}(A_M, D)} = \frac{1}{n}\sum_{k=1}^{n}(A_M(k) - \bar{A}_M)(D(k) - \bar{D}),$$

respectively, where $\bar{A}_P = \sum_{k=1}^{n}A_P(k)/n$, $\bar{A}_M = \sum_{k=1}^{n}A_M(k)/n$ and $\bar{D} = \sum_{k=1}^{n}D(k)/n$.

### 2.2. ONE-LOCUS MODEL FOR PHASE UNKNOWN

In this subsection, we consider modeling a multi-allele QTL with phase unknown genotypes—a more common situation in practice. As we cannot distinguish the parental origins of alleles in QTL genotypes, as usual, we assume that the paternal and maternal gametes share the same set of alleles with the same allele frequencies. Let $A_1, \ldots, A_m$ ($m \geq 2$) denote the alleles at a target QTL or marker locus with allele frequencies $p_i$, $i = 1, \ldots, m$. Ignoring the phases, there are $m$ possible homozygous genotypes $A_iA_i$, $i = 1 \ldots, m$, and $m(m-1)/2$ possible heterozygous genotypes $A_iA_j$, $i \neq j$. We also assume that these alleles contribute the same genetic effects regardless of their parental origins, which implies that the expected genotypic values $G_{ij} = E(G|g = A_iA_j)$ satisfy the symmetric property: $G_{ij} = G_{ji}$, for $i \neq j$. So there are totally $m(m+1)/2$ possible distinctive expected genotypic values $G_{ij}$, $i, j = 1, \ldots, m$. In this case, by treating the paternal and maternal gametes as two independent risk factors, the Fisher's ANOVA model for the expected genotypic values $G_{ij}$ can be written as

$$G_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij},\ i, j = 1, \ldots, m, \tag{5}$$

where $\alpha_i$ is the average (additive) allelic effect of the paternal or maternal allele $A_i$ ($i = 1, \ldots, m$), and $\delta_{ij}$ is the average allelic interaction between two alleles $A_i$ and $A_j$ ($i, j = 1, \ldots, m$). As pointed out in Wang (2011), the above model is different from the classical two-way ANOVA model in that the paternal and the maternal gametes share the same set of alleles and have the same allelic effects at the locus. To avoid the inestimability of model parameters in model (5) due to over-parameterization, the following constraints are often added on the model parameters

$$\sum_i p_i\alpha_i = 0, \quad \sum_i p_i\delta_{ij} = 0.$$

From the symmetric property of $G_{ij}$, we also assume $\delta_{jk}$'s are symmetric. Similarly, the above constraints together with the symmetry property of $\delta_{jk}$ make it difficult to fit model (5) using the standard LS approach.

Note that model (5) can be treated as a special case of model (1) or (3) with $\alpha^i = \alpha_i$ and $\delta^i_j = \delta^j_i$ for $i, j = 1, \ldots, m$. To construct a similar GMA model for model (5), we can combine the term $\alpha^{*i}x_{P_i}$ with $\alpha_{*i}x_{M_i}$, and $\delta^{*i}_{*j}x_{P_i}x_{M_j}$ with $\delta^{*j}_{*i}x_{P_j}x_{M_i}$ in model (4). By denoting $\alpha^{*i} = \alpha_{*i}$ as $\alpha^*_i$ for $i = 1, \ldots, m-1$, and $\delta^{*j}_{*i} = \delta^{*i}_{*j}$ as $\delta^*_{ij}$ for $i \le j$, we obtain the following GMA model

$$E(G|g) = \mu + \sum_{i=1}^{m-1}\alpha^*_i w_i(g) + \sum_{i=1}^{m-1}\delta^*_{ii}v_{ii}(g) + \sum_{j=2}^{m-1}\sum_{i<j}\delta^*_{ij}v_{ij}(g), \quad (6)$$

where, for $i = 1, \ldots, m-1$,

$$w_i(g) = x_{P_i} + x_{M_i} = \begin{cases} 2(1-p_i), & \text{if } g = A_iA_i \\ 1-2p_i, & \text{if } g = A_iA^c_i \\ -2p_i, & \text{if } g = A^c_iA^c_i \end{cases},$$

$$v_{ii}(g) = x_{P_i}x_{M_i} = \begin{cases} (1-p_i)^2, & \text{if } g = A_iA_i \\ -p_i(1-p_i), & \text{if } g = A_iA^c_i \\ p_i^2, & \text{if } g = A^c_iA^c_i \end{cases},$$

and for $i < j$

$$v_{ij}(g) = x_{P_i}x_{M_j} + x_{P_j}x_{M_i} = \begin{cases} (1-p_i)(1-p_j) + p_ip_j, & \text{if } g = A_iA_j \\ -2p_j(1-p_i), & \text{if } g = A_iA_i \\ -p_j(1-2p_i), & \text{if } g = A_iA^c_{ij} \\ -2p_i(1-p_j), & \text{if } g = A_jA_j \\ -p_i(1-2p_j), & \text{if } g = A_jA^c_{ij} \\ 2p_ip_j, & \text{if } g = A^c_{ij}A^c_{ij} \end{cases}.$$

Here $A^c_{ij}$ denotes an allele which is different from $A_i$ and $A_j$. Note that the combined index variables $w_i(g)$, $v_{ii}(g)$, and $v_{ij}(g)$ above are well defined on unphased genotypes, although $x_{P_i}$, $x_{M_j}$ are not. We refer them as genotype coding variables, and the model parameter $\alpha^*_i$ as the average allelic effect of allele $A_i$ $(i = 1, \ldots, m)$, and $\delta^*_{ij}$ as the average allelic interaction between two alleles $A_i$ and $A_j$ $(i, j, = 1, \ldots, m)$, with respect to the baseline allele $A_m$. In terms of the parameters in the original model (5), we can show that $\alpha^*_i = \alpha_i - \alpha_m$, for $i = 1, \ldots, m-1$; and $\delta^*_{ij} = \delta_{ij} - \delta_{im} - \delta_{jm} + \delta_{mm}$, for $i \le j$, $i, j = 1, \ldots, m-1$.

Model (6) is an extension of the one-locus G2A model proposed in Zeng et al. (2005) to one QTL with multiple alleles. Note that no $v_{ij}(g)$'s for $i < j$ are needed in the bi-allelic case. The combined index variables $v_{ii}(g)$, $i = 1, \ldots, m-1$, are also slightly different from the ones defined in Zeng et al. (2005) by a scalar of $(-2)$ folds. Here we drop the scalar $(-2)$ so that the coefficient $\delta^*_{ii}$ can keep the same interpretation as $\delta^{*i}_{*i}$ in model (4). In addition, the combined index variables $v_{ij}$ $(i < j)$ defined above are not exactly the same as the ones we suggested in the discussion section of Wang (2011). Based on the previously defined inheritance indicator variables, we define $w^*_i = z_{P_i} + z_{M_i}$ and $v^*_{ij}(g) = z_{P_i}z_{M_j} + z_{P_j}z_{M_i}$ for $i < j$. Then here we have $v_{ij} = v^*_{ij} - (p_iw^*_j + p_jw^*_i) + 2p_ip_j$, for $i < j$.

Still, model (6) retains the nice feature of the classical Fisher's model on partition of the genotypic variance. The additive variance component $V_A = V\left(\sum_{i=1}^{m-1}\alpha^*_i w_i(g)\right)$, which is contributed by the additive effects of both paternal and maternal alleles. The dominant variance component $V_D = V\left(\sum_{i=1}^{m-1}\delta^*_{ii}v_{ii}(g) + \sum_{i<j}\delta^*_{ij}v_{ij}(g)\right)$, which is contributed by all the interactions between paternal and maternal alleles. Under HWE, we have $\mu = E(G)$ and an orthogonal partition on the variance of the expected genotypic values $V(E(G|g)) = V_A + V_D$, where

$$V_A = 2\sum_{i=1}^{m-1}p_i(\alpha^*_i)^2 - 2\left(\sum_{i=1}^{m-1}p_i\alpha^*_i\right)^2,$$

$$V_D = \sum_{i,j=1}^{m-1}p_ip_j(\delta^*_{ij})^2 - 2\sum_{j=1}^{m-1}p_j\left(\sum_{i=1}^{m-1}p_i\delta^*_{ij}\right)^2 + \left(\sum_{i,j=1}^{m-1}p_ip_j\delta^*_{ij}\right)^2.$$

Here we define $\delta^*_{ji} = \delta^*_{ij}$, for $i < j$. In HWD, the desired orthogonal partition on the variance of the expected genotypic values is no longer held. But the model can still allow us to capture the disequilibria via covariances between those index variables $x_{P_i}$'s and $x_{M_j}$'s (see **Appendix B**).

As an example, let us consider a QTL with 3 alleles $A_1$, $A_2$, and $A_3$. By taking $A_3$ as the baseline allele, model (6) leads to

$$E(G|g) = \mu + \alpha^*_1w_1(g) + \alpha^*_2w_2(g) + \delta^*_{11}v_{11}(g) + \delta^*_{22}v_{22}(g) + \delta^*_{12}v_{12}(g). \quad (7)$$

Or, in a matrix form, we have

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \\ G_{13} \\ G_{23} \\ G_{33} \end{pmatrix} = \begin{pmatrix} 1 & 2(1-p_1) & -2p_2 & (1-p_1)^2 & p_2^2 & -2p_2(1-p_1) \\ 1 & 1-2p_1 & 1-2p_2 & -p_1(1-p_1) & -p_2(1-p_2) & 1-p_1-p_2+2p_1p_2 \\ 1 & -2p_1 & 2(1-p_2) & p_1^2 & (1-p_2)^2 & -2p_1(1-p_2) \\ 1 & 1-2p_1 & -2p_2 & -p_1(1-p_1) & p_2^2 & -p_2(1-2p_1) \\ 1 & -2p_1 & 1-2p_2 & p_1^2 & -p_2(1-p_2) & -p_1(1-2p_2) \\ 1 & -2p_1 & -2p_2 & p_1^2 & p_2^2 & 2p_1p_2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha^*_1 \\ \alpha^*_2 \\ \delta^*_{11} \\ \delta^*_{22} \\ \delta^*_{12} \end{pmatrix}$$

If we choose $A_1$ (or $A_2$) instead of $A_3$ as the baseline allele, we can obtain different re-parameterizations of the six expected genotypic values. But they all give the same partition on the variance of the expected genotypic values. Álvarez-Castro and Yang (2011) also presented a similar re-parameterization of the six

expected genotypic values from their NOIA model. It appears that their one-locus 3-allele NOIA model and the GMA model (7) above are likely equivalent on partitioning the variance of the six expected genotypic values, as we will see later in the Example Subsection 2.4.

Similar to the phase-known case, we can estimate the additive, dominance variance components and the covariance Cov($A$, $D$) from a random sample when the QTL genotypes are available. First, we can compute the values of the combined index variables $w_i(g_k)$ and $v_{ij}(g_k)$ at the target QTL for each subject $k = 1, \ldots, n$. Next, we can incorporate model (6) into a regression model with other possible adjusted covariates. By treating those combined index variables as regular fixed covariates, we can fit the regression model using the standard LS approach. Then, based on the fitted model, we compute the additive and dominance components $A(k) = \sum_{i=1}^{m-1} \hat{\alpha}_i^* w_i(g_k)$ and $D(k) = \sum_{i=1}^{m-1} \hat{\delta}_{ii}^* v_{ii}(g_k) + \sum_{i<j} \hat{\delta}_{ij}^* v_{ij}(g_k)$ for each subject $k = 1, \ldots, n$, where $\hat{\alpha}_i^*$, $\hat{\delta}_{ii}^*$ and $\hat{\delta}_{ij}^*$ are LSE of the model parameters. Finally, we can estimate $V_A$ and $V_D$ as the sample variances of $\{A(k), k = 1, \ldots, n\}$ and $\{D(k), k = 1, \ldots, n\}$, respectively. Meanwhile, an estimate of Cov($A$, $D$) is given by the sample covariance between $\{A(k), k = 1, \ldots, n\}$ and $\{D(k), k = 1, \ldots, n\}$.

### 2.3. MULTI-LOCUS MODELS

The one-locus GMA models can be extended to multiple loci. Typically, for each locus $k$, we can create the mean-corrected index variables $x_{P_i}^{(k)}$, $x_{M_j}^{(k)}$ (or combined index variables $w_{k,i}$, $v_{k,ij}$ for phase-unknown genotypes) in the same way as in the one-locus case. Then we can build a multi-locus GMA model by including the within locus additive and dominance effects as well as the locus-by-locus interactions (i.e., epistases). The proposed multi-locus GMA model allows LD among multiple loci even though the LD may affect the partition on the variance of the expected genotypic values.

Consider $L$ loci with alleles $A_{k1}, \ldots, A_{km_k}$ at the $k$-th locus for $k = 1, \ldots, L$. With phase known, there are totally $m_1^2 \cdots m_L^2$ possible expected genotypic values: $G_{t_1 \cdots t_L}^{s_1 \cdots s_L} = \mathrm{E}[G|g = (A_{1s_1} \cdots A_{Ls_L}, A_{1t_1} \cdots A_{Lt_L})]$, where the joint genotype $(A_{1s_1} \cdots A_{Ls_L}, A_{1t_1} \cdots A_{Lt_L})$ is formed by the union of a paternal gamete $A_{1s_1} \cdots A_{Ls_L}$ and a maternal gamete $A_{1t_1} \cdots A_{Lt_L}$ with $s_k, t_k \in \{1, \ldots, m_k\}$ for $k = 1, \ldots, L$. Let $p_{ks_k}^P$ and $p_{kt_k}^M$ be the frequencies of the paternal allele $A_{ks_k}$ and maternal allele $A_{kt_k}$, respectively. At each locus $k = 1, \ldots, L$, we define the mean-corrected index variables $x_{Ps_k}^{(k)}$ for paternal alleles $A_{ks_k}$ ($s_k = 1, \ldots, m_k$) and $x_{Mt_k}^{(k)}$ for maternal alleles $A_{kt_k}$ ($t_k = 1, \ldots, m_k$) in the same way as in the one-locus case. Then, when we choose $A_{1m_1}, \ldots, A_{Lm_L}$ as the baseline alleles, a fully parameterized $L$-locus GMA model can be expressed as

$$E(G|g) = \sum_{i_1=0}^{1} \sum_{j_1=0}^{1} \cdots \sum_{i_L=0}^{1} \sum_{j_L=0}^{1} \sum_{\{s_1, \text{if } i_1=1\}} \sum_{\{t_1, \text{if } j_1=1\}} \cdots \sum_{\{s_L, \text{if } i_L=1\}} \sum_{\{t_L, \text{if } j_L=1\}} \alpha_{t_1^* \cdots t_L^*}^{s_1^* \cdots s_L^*} \cdot \left(x_{Ps_1}^{(1)}\right)^{i_1} \left(x_{Mt_1}^{(1)}\right)^{j_1} \cdots \left(x_{Ps_L}^{(L)}\right)^{i_L} \left(x_{Mt_L}^{(L)}\right)^{j_L} \quad (8)$$

where the summation of $s_k$ (or $t_k$) is from 1 up to $(m_k - 1)$ for $k = 1, \ldots, L$; $s_k$ (or $t_k$) refers to a particular paternal allele $A_{ks_k}$ (or maternal allele $A_{kt_k}$) at the $k$-th locus; and $i_k$ (or $j_k$) is an indicator variable for the presence or absence of a paternal (or maternal) allele at locus $k$ which is involved in a term. The coefficient $\alpha_{t_1^* \cdots t_L^*}^{s_1^* \cdots s_L^*}$ in each term represents an average allelic effect of a single paternal or maternal allele, or an allelic interaction from a set of paternal and maternal alleles that are involved in this term with respect to the baseline alleles $A_{1m_1}, \ldots, A_{Lm_L}$. The superscripts (or subscripts) in $\alpha_{t_1^* \cdots t_L^*}^{s_1^* \cdots s_L^*}$ are defined as $s_k^* = i_k \cdot s_k$ (or $t_k^* = j_k \cdot t_k$), indicating which paternal (or maternal) allele at locus $k$ is involved in this term. If the term does not have any paternal (or maternal) allele at locus $k$ being involved, then we have $s_k^* = 0$ (or $t_k^* = 0$). Note that $na_P = \sum_{k=1}^{L} i_k$ (or $na_M = \sum_{k=1}^{L} j_k$) specifies the total number of paternal (or maternal) alleles being involved in a term. We refer the total number of both paternal and maternal alleles that are involved in a term $na = na_P + na_M$ as the order of this term.

The multi-locus GMA model (8) provides a full re-parameterization of the $m_1^2 \cdots m_L^2$ expected genotypic values $G_{t_1 \cdots t_L}^{s_1 \cdots s_L}$ with phase-known genotypes without using redundant parameters. Note that $E(G|g)$ can also be partitioned into a sum of the following genetic components

$$C_{j_1 \cdots j_L}^{i_1 \cdots i_L} = \sum_{\{s_1, \text{if } i_1=1\}} \sum_{\{t_1, \text{if } j_1=1\}} \cdots \sum_{\{s_L, \text{if } i_L=1\}} \sum_{\{t_L, \text{if } j_L=1\}} \alpha_{t_1^* \cdots t_L^*}^{s_1^* \cdots s_L^*} \cdot \left(x_{Ps_1}^{(1)}\right)^{i_1} \left(x_{Mt_1}^{(1)}\right)^{j_1} \cdots \left(x_{Ps_L}^{(L)}\right)^{i_L} \left(x_{Mt_L}^{(L)}\right)^{j_L},$$

where $i_1, j_1, \ldots, i_L, j_L = 0, 1$ are indicator variables, which specify the parental origins of the contributing alleles. In other words, each genetic component consists of the terms with the same order and having their alleles all coming from the same subset of loci with the same parental origins but allowing varied allelic types. The classical genetic variance components can then be defined as variances of these genetic components. Note that the component $C_{0 \cdots 0}^{0 \cdots 0}$ is a constant, which corresponds to an intercept without any alleles being involved. Therefore, for a $L$-locus GMA model with phase known genotypes, there are in total $(2^{2L} - 1)$ genetic variance components $V\left(C_{j_1 \cdots j_L}^{i_1 \cdots i_L}\right)$, where $i_1, j_1, \ldots, i_L, j_L = 0, 1$ but not all being zeros. Among them, there are $L$ paternal and $L$ maternal variance components of order 1 with each having a single paternal or maternal allele being involved, $L$ dominance variance components of order 2 with each having two alleles within the same locus being involved, $L(L-1)/2$ paternal by paternal (or maternal by maternal) variance components of order 2 with two paternal (or maternal) alleles coming from different loci, and $L(L-1)$ paternal by maternal variance components of order 2 with one paternal and one maternal allele coming from two different loci. For examples, the within-locus paternal, maternal and dominance variance components at a locus $k$ can be written as

$$V_{A_P^{(k)}} = V\left(\sum_{s_k=1}^{m_k-1} \alpha_{0 \cdots s_k \cdots 0}^{0 \cdots s_k \cdots 0} \cdot x_{Ps_k}^{(k)}\right) = V\left(C_{0 \cdots 0 \cdots 0}^{0 \cdots 1 \cdots 0}\right),$$

$$V_{A_M^{(k)}} = V\left(\sum_{t_k=1}^{m_k-1} \alpha_{0\cdots t_k\cdots 0}^{0\cdots 0\cdots 0} \cdot x_{Mt_k}^{(k)}\right) = V\left(C_{0\cdots 1\cdots 0}^{0\cdots 0\cdots 0}\right),$$

$$V_{D^{(k)}} = V\left(\sum_{s_k=1}^{m_k-1}\sum_{t_k=1}^{m_k-1} \alpha_{0\cdots t_k\cdots 0}^{0\cdots s_k\cdots 0} \cdot x_{Ps_k}^{(k)} x_{Mt_k}^{(k)}\right) = V\left(C_{0\cdots 1\cdots 0}^{0\cdots 1\cdots 0}\right),$$

respectively ($k = 1, \ldots, L$). For a pair of loci $j \neq k$, the two-locus paternal by paternal variance component is

$$V_{A_P^{(j)}\times A_P^{(k)}} = V\left(\sum_{s_j=1}^{m_j-1}\sum_{s_k=1}^{m_k-1} \alpha_{0\cdots 0\cdots 0\cdots 0}^{0\cdots s_j\cdots s_k\cdots 0} \cdot x_{Ps_j}^{(j)} x_{Ps_k}^{(k)}\right) = V\left(C_{0\cdots 0\cdots 0\cdots 0}^{0\cdots 1\cdots 1\cdots 0}\right),$$

and the two-locus paternal by maternal variance component is

$$V_{A_P^{(j)}\times A_M^{(k)}} = V\left(\sum_{s_j=1}^{m_j-1}\sum_{t_k=1}^{m_k-1} \alpha_{0\cdots 0\cdots t_k\cdots 0}^{0\cdots s_j\cdots 0\cdots 0} \cdot x_{Ps_j}^{(j)} x_{Mt_k}^{(k)}\right) = V\left(C_{0\cdots 0\cdots 1\cdots 0}^{0\cdots 1\cdots 0\cdots 0}\right).$$

The variance component of epistases with the highest order is

$$V_{D^{(1)}\times\cdots\times D^{(L)}} = V\left(\sum_{s_1=1}^{m_1-1}\sum_{t_1=1}^{m_1-1}\cdots\sum_{s_L=1}^{m_L-1}\sum_{t_L=1}^{m_L-1} \alpha_{t_1\cdots t_L}^{s_1\cdots s_L} x_{Ps_1}^{(1)} x_{Mt_1}^{(1)} \cdots x_{Ps_L}^{(L)} x_{Mt_L}^{(L)}\right) = V(C_{1\cdots 1}^{1\cdots 1}),$$

which has $2L$ alleles being involved with two alleles per locus.

Based on these genetic components, we can partition the variance of the expected genotypic values into the variances and covariances of the genetic components. Still, the coefficients $\alpha_{t_1^*\cdots t_L^*}^{s_1^*\cdots s_L^*}$ are defined based on the baseline alleles $A_{1m_1}, \ldots, A_{Lm_L}$. But the variances and covariances of these genetic components do not depend on the choice of the baseline alleles. Under HWE, the within-locus paternal index variables $\{x_{Ps_k}^{(k)}, s_k = 1, \ldots, m_k\}$ are independent of the maternal index variables $\{x_{Mt_k}^{(k)}, t_k = 1, \ldots, m_k\}$ for each $k = 1, \ldots, L$. With LE, the index variables $\{x_{Ps_j}^{(j)}, x_{Mt_j}^{(j)}, s_j, t_j = 1, \ldots, m_j\}$ at a locus $j$ are also independent of the index variables $\{x_{Ps_k}^{(k)}, x_{Mt_k}^{(k)}, s_k, t_k = 1, \ldots, m_k\}$ at a different locus $k$ ($k \neq j$). To achieve an orthogonal partition on the variance of the expected genotypic values, we need to further assume that all the paternal index variables $\{x_{Ps_k}^{(k)}, s_k = 1, \ldots, m_k, k = 1, \ldots, L\}$ are independent of the maternal index variables $\{x_{Mt_k}^{(k)}, t_k = 1, \ldots, m_k, k = 1, \ldots, L\}$ across all the loci; i.e., the so-called gametic equilibrium (see Wang and Zeng, 2006). Under both the gametic and linkage equilibria, each genetic component has its mean $E\left(C_{j_1\cdots j_L}^{i_1\cdots i_L}\right) = 0$, and the covariances between different genetic components are zeros. Thus, we have $E(G) = \alpha_{0\cdots 0}^{0\cdots 0}$ and an orthogonal partition on the variance of the expected genotypic values is given by

$$V(E(G|g)) = \sum_{i_1=0}^{1}\sum_{j_1=0}^{1}\cdots\sum_{i_L=0}^{1}\sum_{j_L=0}^{1} V\left(C_{j_1\cdots j_L}^{i_1\cdots i_L}\right), \quad (9)$$

where

$$V\left(C_{j_1\cdots j_L}^{i_1\cdots i_L}\right) = \sum_{\{s_1, s_1', \text{if } i_1=1\}} (p_{1s_1}^P 1_{\{s_1=s_1'\}} - p_{1s_1}^P p_{1s_1'}^P)^{i_1} \sum_{\{t_1, t_1', \text{if } j_1=1\}} (p_{1t_1}^M 1_{\{t_1=t_1'\}} - p_{1t_1}^M p_{1t_1'}^M)^{j_1} \cdots \sum_{\{s_L, s_L', \text{if } i_L=1\}} (p_{Ls_L}^P 1_{\{s_L=s_L'\}} - p_{Ls_L}^P p_{Ls_L'}^P)^{i_L} \sum_{\{t_L, t_L', \text{if } j_L=1\}} (p_{Lt_L}^M 1_{\{t_L=t_L'\}} - p_{Lt_L}^M p_{Lt_L'}^M)^{j_L} \cdot \alpha_{t_1^*\cdots t_L^*}^{s_1^*\cdots s_L^*} \alpha_{t_1'^*\cdots t_L'^*}^{s_1'^*\cdots s_L'^*}.$$

Similarly, we can construct multi-locus GMA models for QTL with phase unknown genotypes. Without distinguishing the parental origin of the alleles, there are totally $\prod_{k=1}^{L} m_k(m_k+1)/2^L$ possible expected genotypic values: $G_{s_1t_1\cdots s_Lt_L}$=E$[G(g)|g = (A_{1s_1}A_{1t_1}, \ldots, A_{Ls_L}A_{Lt_L})]$, for $s_k, t_k = 1, \ldots, m_k$ and $k = 1, \ldots, L$. We assume that the paternal and maternal allele frequencies are the same and denoted by $p_{ks_k}$. We define the combined index variables $w_{k,i}(g)$ and $v_{k,ij}$ at each locus $k$ in the same way as the one-locus case for $k = 1, \ldots, L$. Then, when we choose $A_{1m_1}, \ldots, A_{Lm_L}$ as the baseline alleles, a fully parameterized $L$-locus GMA model for QTL with phase unknown genotypes can be expressed as

$$E(G|g) = \sum_{i_1=0}^{2}\cdots\sum_{i_L=0}^{2}\sum_{\{s_1,\text{if } i_1=1\}}\cdots \sum_{\{s_L,\text{if } i_L=1\}}\sum_{\{s_1\le t_1,\text{if } i_1=2\}}\cdots\sum_{\{s_L\le t_L,\text{if } i_L=2\}} \alpha_{s_1^*\cdots s_L^*} \cdot \left(w_{1,s_1}^{1_{\{i_1=1\}}} \cdot v_{1,s_1t_1}^{1_{\{i_1=2\}}}\right)\cdots\left(w_{L,s_L}^{1_{\{i_L=1\}}} \cdot v_{L,s_Lt_L}^{1_{\{i_L=2\}}}\right), \quad (10)$$

where $1_{\{i_k=j\}}$ is the Kronecker function which equals 1 when $i_k = j$ and 0 otherwise, for $k = 1, \ldots, L$ and $j = 1$ or 2; the summation of $s_k$ (or $t_k$) is from 1 up to $m_k - 1$ with $s_k$ (or $t_k$) referring to allele $A_{ks_k}$ (or $A_{kt_k}$) at a locus $k$ ($k = 1, \ldots, L$); and $i_k$ specifies how many alleles at a locus $k$ are involved in this term. If $i_k = 1$, this term has only one allele $A_{ks_k}$ (either paternal or maternal) being involved via $w_{k,s_k}$ and we set $s_k^* = s_k$. If $i_k = 2$, both the paternal and maternal alleles $A_{ks_k}$ and $A_{kt_k}$ are involved via $v_{k,s_kt_k}$ in this term and we set $s_k^* = s_kt_k$ regardless of the order of $s_k$ and $t_k$. When $i_k = 0$, this term does not have any alleles at locus $k$ being involved and we have $\left(w_{k,s_k}^{1_{\{i_k=1\}}} \cdot v_{k,s_kt_k}^{1_{\{i_k=2\}}}\right) = 1$. The coefficient $\alpha_{s_1^*\ldots s_L^*}$ represents the average allelic effect of a single allele, or an allelic interaction from all the alleles that are involved in this term, with respect to the baseline alleles $A_{1m_1}, \ldots, A_{Lm_L}$. We still

refer the total number of alleles being involved in each term; i.e., $na = \sum_{k=1}^{L} i_k$, as the order of this term.

Based on the above model, we can define the genetic components as

$$C_{i_1\cdots i_L} = \sum_{\{s_1,\text{if } i_1=1\}} \cdots \sum_{\{s_L,\text{if } i_L=1\}} \sum_{\{s_1\le t_1,\text{if } i_1=2\}} \cdots \sum_{\{s_L\le t_L,\text{if } i_L=2\}} \alpha_{s_1^*\cdots s_L^*} \cdot \left(w_{1,s_1}^{1_{\{i_1=1\}}} \cdot \nu_{1,s_1t_1}^{1_{\{i_1=2\}}}\right) \cdots \left(w_{L,s_L}^{1_{\{i_L=1\}}} \cdot \nu_{L,s_Lt_L}^{1_{\{i_L=2\}}}\right),$$

for $i_1, \ldots, i_L = 0, 1, 2$, with each component being a summation of the terms with their alleles all coming from the same subset of loci and having the same number of alleles from each locus. Excluding the one with $i_1 = \cdots = i_L = 0$, which corresponds to an intercept, there are in total $3^L - 1$ genetic variance components $V(C_{i_1\cdots i_L})$ for $i_1, \ldots, i_L = 0, 1, 2$. Among them, there are $L$ additive variance components with each having a single paternal or maternal allele being involved, $L$ dominance variance components with each having both paternal and maternal alleles within the same locus being involved, $L(L-1)/2$ additive by additive variance components with two alleles coming from different loci, etc. For examples, the within-locus additive and dominance variance components at a locus $k$ can be written as

$$V_{A^{(k)}} = V\left(\sum_{s_k=1}^{m_k-1} \alpha_{0\cdots s_k\cdots 0} \cdot w_{k,s_k}\right) = V(C_{0\cdots 1\cdots 0}),$$

$$V_{D^{(k)}} = V\left(\sum_{s_k\le t_k} \alpha_{0\cdots s_kt_k\cdots 0} \cdot \nu_{k,s_kt_k}\right) = V(C_{0\cdots 2\cdots 0}),$$

respectively, for $k = 1, \ldots, L$. For a pair of loci $j \neq k$, the two-locus additive by additive interaction is

$$V_{A^{(j)}\times A^{(k)}} = V\left(\sum_{s_j=1}^{m_j-1}\sum_{s_k=1}^{m_k-1} \alpha_{0\cdots s_j\cdots s_k\cdots 0} \cdot w_{j,s_j}w_{k,s_k}\right) = V(C_{0\cdots 1\cdots 1\cdots 0}).$$

The variance component of epistases with the highest order of $2L$ is given by

$$V_{D^{(1)}\times\cdots\times D^{(L)}} = V\left(\sum_{s_1\le t_1} \cdots \sum_{s_L\le t_L} \alpha_{s_1t_1,\cdots,s_Lt_L}\nu_{1,s_1t_1}\cdots\nu_{L,s_Lt_L}\right) = V(C_{2\cdots 2}).$$

Under both the gametic and linkage equilibria, we have $E(G) = \alpha_{0\cdots 0}$ and an orthogonal partition on the variance of the expected genotypic values

$$V(E(G|g)) = \sum_{i_1=0}^{2} \cdots \sum_{i_L=0}^{2} V(C_{i_1\cdots i_L}), \tag{11}$$

where

$$V(C_{i_1\cdots i_L}) = \sum_{\{s_1,s_1',\text{if } i_1=1\}} \left\{2\left(p_{1s_1}1_{\{s_1=s_1'\}} - p_{1s_1}p_{1s_1'}\right)\right\}^{1_{\{i_1=1\}}} \cdots \sum_{\{s_L,s_L',\text{if } i_L=1\}} \left\{2\left(p_{Ls_L}1_{\{s_L=s_L'\}} - p_{Ls_L}p_{Ls_L'}\right)\right\}^{1_{\{i_L=1\}}} \sum_{\{s_1,s_1',t_1,t_1',\text{if } i_1=2\}} \left\{\left(p_{1s_1}1_{\{s_1=s_1'\}} - p_{1s_1}p_{1s_1'}\right)\left(p_{1t_1}1_{\{t_1=t_1'\}} - p_{1t_1}p_{1t_1'}\right)\right\}^{1_{\{i_1=2\}}} \cdots \sum_{\{s_L,s_L',t_L,t_L',\text{if } i_L=2\}} \left\{\left(p_{Ls_L}1_{\{s_L=s_L'\}} - p_{Ls_L}p_{Ls_L'}\right)\left(p_{Lt_L}1_{\{t_L=t_L'\}} - p_{Lt_L}p_{Lt_L'}\right)\right\}^{1_{\{i_L=2\}}} \alpha_{s_1^*\cdots s_L^*}\alpha_{s_1^{*\prime}\cdots s_L^{*\prime}}.$$

Here, when $i_k = 2$, we have $s_k^* = s_kt_k$ and $s_k^{*\prime} = s_k't_k'$. We also define $\alpha_{s_1^*\cdots s_L^*}$ (or $\alpha_{s_1^{*\prime}\cdots s_L^{*\prime}}$) to be the same if we switch the order of $s_k$ and $t_k$ in $s_k^*$ (or $s_k'$ and $t_k'$ in $s_k^{*\prime}$) for $k = 1, \ldots, L$. In the presence of disequilibria, the desired orthogonal partition may no longer hold. However, regardless of the equilibrium, the coefficients $\alpha_{s_1^*\cdots s_L^*}$ in model (10) are defined based on the baseline alleles $A_{1m_1}, \ldots, A_{Lm_L}$, while the variances and covariances of the genetic components do not depend on the choice of these baseline alleles.

In practice, we do not have to rely on the derived formula to estimate the genetic variance or covariance components. Similar to the one-locus case, given the observed QTL genotypes for a random sample from a study population, we can always incorporate model (8) or (10) into a regression model with other possible adjusted covariates and fit the model using standard LS approach. Then we can estimate various genetic variance components as well as the covariances among different genetic components based on the fitted model. A good fit of a fully parameterized GMA model often requires that the expected genotypic values for all possible joint genotypes of the QTL are estimable from the study sample. If certain genotypes are not observable or rarely present in subjects from the study sample, a situation which likely happens when the number of alleles or the number of QTL is large with moderate or small sample size, the design matrix for the genetic effects could become singular which implies that some genetic variance components cannot be estimated reliably. But we do not have to use fully parameterized GMA models to model the expected genotypic values. In this case, we may want to build a reduced GMA model that can provide a good approximation to the expected genotypic values overall and meanwhile has a less complicated model structure. The fact that two terms within the same genetic component are unavoidably correlated suggests that we should perhaps treat each genetic component as a whole and keep or drop its terms all at once in building a GMA model. As genetic components of lower orders tend to have bigger impact on the expected genotypic values than the higher order ones, one way to construct a reduced GMA model is perhaps to go through a stepwise forward selection procedure by hierarchically adding

the lowest order genetic component that can achieve a nominal significance level (e.g., 5%) but has not yet been selected in the model into the model one at a time. Here, the classical likelihood ratio statistic can be used to assess each genetic component for entering into or dropping from the model.

It has been known that the model building procedure is often sensitive to potential confounding among the selected variables. A GMA model uses the mean-corrected index variables $x^{(k)}_{Ps_k}, x^{(k)}_{Mt_k}$ (or their combined index variables $w_{k,s_k}, v_{k,s_kt_k}$ in the phase unknown case) as the basic units in constructing all its model terms. At least in an equilibrium population, these mean-corrected index variables can reduce the confounding among different genetic components and make them uncorrelated. This orthogonality implies that each genetic component can be assessed independently regardless of which components are presented in the model. On the other hand, an $F_\infty$ model can be thought of directly using the inheritance indicator variables $z^{(k)}_{Ps_k}$ and $z^{(k)}_{Mt_k}$ (or their merged genotype coding variables $w^*_{k,s_k}(g) = z^{(k)}_{Ps_k} + z^{(k)}_{Ms_k}$ and $v^*_{k,s_kt_k} = z^{(k)}_{Ps_k} z^{(k)}_{Mt_k} + z^{(k)}_{Pt_k} z^{(k)}_{Ms_k}$ in phase unknown case) as the basic units in constructing its terms (Wang and Zeng, 2009; Wang, 2011). Even in an equilibrium population, an $F_\infty$ model could have its low-order terms being highly confounded with other high-order terms when they contain shared alleles (see Zeng et al., 2005). As the result, in building a predictive model based on $F_\infty$ models, the stepwise forward selection procedure could be problematic because failing to include a significant higher order term (or component) in a reduced $F_\infty$ model could make the assessment of some low-order terms (or components) unreliable. On selecting significant QTL from a set of loci without having the locus-by-locus interactions being involved, the choice of using GMA or $F_\infty$ model in building a reduced model for the expected genotypic values should not matter much because using mean-corrected index variables mainly affects the intercept in this case. However, when we consider including epistases for a given set of QTL, the GMA model can appropriately use the orthogonal property among different genetic components to dissect the confounding at least in equilibrium populations, while it appears that the $F_\infty$ models cannot make full use of the equilibrium information. When disequilibria are present, as Hardy-Weinberg equilibria are expected to be held in most of the human genomic regions and LD mainly present for closely linked loci, we would expect that most of the genetic components in a GMA model are likely uncorrelated. Therefore, in most cases, using the GMA model could still be preferable to using $F_\infty$ model in building reduced models for expected genotypic values especially when epistases are involved.

### 2.4. EXAMPLE

As an example, we apply the GMA model to a published experimental data set on the polymorphism at the human acid phosphatase locus (ACP1). The analysis of this data set was first conducted by Greene et al. (2000), and recently re-analyzed as an example in Álvarez-Castro and Yang (2011). The ACP1 gene involves 3 alleles A, B, C. Two phenotypic traits are considered, which measure the ACP1 enzyme activity ($y^{ac}$) and inhibition ($y^{in}$). The estimates of the expected genotypic values and the genotype frequencies are summarized in **Table 1**.

From the genotype frequencies, we first estimate the allele frequencies as $p_A = 0.3534$, $p_B = 0.5818$, and $p_C = 0.0647$. Then, for each trait, we fit a separate GMA model (7) to its expected genotypic values by taking $C$ as the baseline allele and creating the combined index variables $w_1(g)$, $w_2(g)$, $v_{11}(g)$, $v_{22}(g)$, $v_{12}(g)$. For ACP1 enzyme activity, we obtain LSE of the model parameters as $\mu = 167.735$, $\alpha^*_1 = -59.260$, $\alpha^*_2 = -26.254$, $\delta^*_{11} = -4.800$, $\delta^*_{22} = 3.700$ and $\delta^*_{12} = -2.000$. For ACP1 enzyme inhibition, we have $\mu = 39.386$, $\alpha^*_1 = -16.149$, $\alpha^*_2 = -19.714$, $\delta^*_{11} = -0.200$, $\delta^*_{22} = 4.200$, and $\delta^*_{12} = 2.100$.

Next, for each trait separately, we calculate $A(g_k) = \alpha^*_1 w_1(g_k) + \alpha^*_2 w_2(g_k)$ and $D(g_k) = \delta^*_{11} v_{11}(g_k) + \delta^*_{22} v_{22}(g_k) + \delta^*_{12} v_{12}(g_k)$, $k = 1, \ldots, 6$, for the six ACP1 genotypes. Based on the genotype frequencies, we then calculate the genetic variance components. For ACP1 enzyme activity, we obtain $V_A = 658.868$, $V_D = 0.973$ and the covariance $\mathrm{Cov}(A, D) = -0.356$. The total variance of the expected genotypic values is $V(E(G^{ac}|g)) = 659.129$, and $V_A/V(E(G^{ac}|g)) = 99.96\%$. For ACP1 enzyme inhibition, we have $V_A = 44.920$, $V_D = 0.146$ and the covariance $\mathrm{Cov}(A, D) = -0.217$. The total variance of the expected genotypic values is $V(E(G^{in}|g)) = 44.632$, and $V_A/V(E(G^{in}|g)) = 100.65\%$. Note that the partition $V(E(G|g)) = V_A + V_D + 2\mathrm{Cov}(A, D)$, which is not orthogonal for both traits due to a slight deviation from HWE at ACP1 locus as we have observed from the genotype frequencies in **Table 1**. It is also interesting to see that for ACP1 enzyme inhibition, $V_A/V(E(G^{in}|g))$ is bigger than 100% due to HWD and the fact that $V_D + 2\mathrm{Cov}(A, D) < 0$.

Using the same allele frequencies but assuming HWE, we would have slightly different genotype frequencies. Note that the LSE of the model parameters keep the same and do not depend on the genotype frequencies because they are completely determined by the allele frequencies and the six expected genotypic values when a fully-parameterized GMA is used. But the total variance of the expected genotypic values and its variance components will be different. For ACP1 enzyme activity, we obtain $V_A = 660.588$, $V_D = 0.971$, $\mathrm{Cov}(A, D) = 0.006$, $V(E(G^{ac}|g)) = 661.573$

**Table 1 | Expected genotypic values and observed genotype frequencies.**

| | ACP1 genotypes | | | | | |
|---|---|---|---|---|---|---|
| | **AA** | **AB** | **BB** | **AC** | **BC** | **CC** |
| ACP1 enzyme activity ($G^{ac}$) | 122.4 | 153.9 | 188.3 | 183.6 | 212.3 | 240.0 |
| ACP1 enzyme inhibition ($G^{in}$) | 41.2 | 37.9 | 34.4 | 58.7 | 53.1 | 76.0 |
| Genotype frequency | 0.1242 | 0.4139 | 0.3349 | 0.0445 | 0.0799 | 0.0025 |

and $V_A/V(E(G^{ac}|g)) = 99.85\%$. For ACP1 enzyme inhibition, we have $V_A = 46.422$, $V_D = 0.152$, $\mathrm{Cov}(A, D) = 0.001$, $V(E(G^{in}|g)) = 46.573$ and $V_A/V(E(G^{in}|g)) = 99.68\%$. The minor deviations from orthogonal partitions are likely due to some numerical round-off and the fact that the summation of the three allele frequencies is 0.9999 instead of exactly 1.

For this three-allele example, we also applied the NOIA model using the formulas (10) and (11) provided in Álvarez-Castro and Yang (2011), and in both HWE and HWD cases we obtained exactly the same results as above based on our own calculation, which however appear to be slightly different from what had been reported in Álvarez-Castro and Yang (2011). For example, based on the observed genotype frequencies, an estimate of $V_D = 1.15$ for the trait of ACP1 enzyme inhibition was reported in Álvarez-Castro and Yang (2011), which is noticeably larger than $V_D = 0.146$ that we obtain above.

## 3. DISCUSSION

In the analysis of genetic variance components, a separation of the variations contributed by the additive allelic effects and allelic interactions is complicated by the fact that the observed genotypes are often phase-unknown. In this study, by appropriately merging the paternal and maternal allelic effects and allelic interactions in the phase-known situation, we propose a way to construct one-locus and multi-locus GMA models on analysis of genetic variance components for QTL with multiple alleles. In the same way as building a G2A model, we construct a GMA model by first specifying its design matrix for the genetic effects via some mean-corrected index variables. As these mean-corrected index variables are well defined based on the observed genotypes and allele frequencies, they can be treated as regular covariates for coding QTL genotypes. These one-locus or multi-locus GMA models can then be incorporated into standard regression models with other possible adjusted covariates and fitted using standard LS approach. Based on the fitted models, we can further estimate the genetic variance and covariance components through the sample variances and covariances of various genetic components. As we have pointed out, these GMA models can be applied to equilibrium populations as well as populations in Hardy-Weinberg and/or linkage disequilibria. By using the full set or a low-order subset of the index variables (or genetic components), the GMA model allows us to make either full or reduced re-parameterization of the genotypic values. When some loci have phase known genotypes while other loci have phase unknown genotypes (a possible hypothetical situation), a mixed GMA model could also be constructed by adopting the same modeling strategy.

Sometimes we may want to perform hypothesis tests on the existence of certain genetic variance and covariance components. Note that the GMA models have allele frequencies being involved in their design matrices. As allele frequencies often need to be estimated from the genotype data, they could contribute another source of variation in the LSE ($\hat{\beta}$) of the model parameters as well as the genetic variance components. When the allele frequencies can be accurately estimated, we could simply treat them as fixed constants. When the residuals in a regression model do not depend on the allele frequency estimates, based on the linear model theory (see Ravishanker and Dey, 2002), $\hat{\beta}$ are known to be unbiased with its covariance matrix $\mathrm{cov}(\hat{\beta}) = \sigma^2 \mathrm{E}[(X'X)^{-1}]$, where $X$ is the design matrix of a GMA model. In this case, we can assess the existence of variance components by performing traditional hypothesis tests on $\hat{\beta}$. In general, we could also assess the existence of these genetic variance and covariance components through a bootstrap procedure. By repeatedly drawing random samples of the same size from the observed random sample with replacement, we can estimate the genetic variance and covariance components for each bootstrap sample and meanwhile assess the variances in estimates of these genetic variance and covariance components and test for their existence.

In genetic studies, QTL with missing genotypes is a common phenomenon. GMA model can be used to fit QTL with missing genotypes. Rather than excluding patients with missing QTL genotypes, we could treat "missing" as an allele although this strategy may induce potential bias as we assume that all the missing alleles have the same genetic effect. GMA models could also be applied incorporation with various imputation methods. In recent years, there has been a great deal of interest in developing methodologies for QTL mapping using recombinant intercrosses from multiple inbred lines. In this case, the putative QTL often have their locations and genotypes unknown. But the allele frequencies of QTL could probably be inferred from the study design and the QTL genotypes might be imputable from their neighboring genetic markers. How to apply GMA models to this type of experimental crosses for QTL mapping could be a research topic for further exploration.

In summary, the analysis of genetic variance components for multi-allele QTL has been challenging due to complex allelic interactions and locus-by-locus interactions. In this study, we thoroughly explored the architecture of one-locus and multi-locus GMA models with either phase known or unknown genotypes. Particularly, we described in detail the architecture of the multi-locus GMA model, and how the model terms can be grouped into various genetic components. Under equilibria populations, we also derived formulas for orthogonal partition of the genetic variance components, which could be useful for analytical assessment of the variance components. Comparing to the classical Fisher model, the GMA models can estimate the genetic variance and covariance components more conveniently via standard LS approach for either one or multiple QTL with multiple alleles, in equilibrium as well as disequilibrium populations.

## ACKNOWLEDGMENTS

The author would like to acknowledge Dr. Zhao-Bang Zeng at Bioinformatics Research Center, North Carolina State University, for his thoughtful comments and suggestions on an earlier version of the manuscript.

## REFERENCES

Álvarez-Castro, J. M. (2012). Current applications of models of genetic effects with interactions across the genome. *Curr. Genomics* 13, 163–175. doi: 10.2174/138920212799860689

Álvarez-Castro, J. M., and Carlborg, Ö. (2007). A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176, 1151–1167. doi: 10.1534/genetics.106.067348

Álvarez-Castro, J. M., and Yang, R.-C. (2011). Multiallelic models of genetic effects and variance decomposition in non-equilibrium populations. *Genetica* 139, 1119–1134. doi: 10.1007/s10709-011-9614-9

Cheverud, J. M. (2000). "Detecting epistasis among quantitative trait loci," in *Epistasis and the Evolutionary Process*, eds J. B. Wolf, E. D. Brodie, and M. J. Wade (New York, NY: Oxford University Press), 58–81.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39, 859–882.

Cockerham, C. C. (1963). "Estimation of genetic variances," in *Statistical Genetics and Plant Breeding*, Vol. 982, eds W. D. Henson and H. F. Robinson (Washington, DC: National Academy of Science - National Research Council), 53–94.

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics, 4th Edn*. Harlow: Longman Group Ltd.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi: 10.1017/S0080456800012163

Greene, L. S., Bottini, N., Borgiani, P., and Gloria-Bottini, F. (2000). Acid phosphatase locus 1 (acp1): possible relationship of allelic variation to body size and human population adaptation to thermal stressa theoretical perspective. *Am. J. Hum. Biol.* 12, 688–701. doi: 10.1002/1520-6300(200009/10)12:5<688::AID-AJHB14>3.0.CO;2-C

Hansen, T. F., and Wagner, G. P. (2001). Modeling genetic architecture: a multilinear theory of gene interaction. *Theor. Popul. Biol.* 59, 61–86. doi: 10.1006/tpbi.2000.1508

Kempthorne, O. (1954). The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. B Biol. Sci.* 143, 103–113. doi: 10.1098/rspb.1954.0056

Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. New York, NY: John Wiley and Sons, Inc.

Kempthorne, O. (1969). *An Introduction to Genetic Statistics*. New Haven, CT: Iowa State University Press Ames.

Ravishanker, N., and Dey, D. K. (2002). *A First Course in Linear Model Theory*. Boca Raton, FL: Chapman and Hall, CRC.

Wang, T. (2011). On coding genotypes for genetic markers with multiple alleles in genetic association study of quantitative traits. *BMC Genet.* 12:82. doi: 10.1186/1471-2156-12-82

Wang, T., and Zeng, Z. B. (2006). Models and partition of variance for quantitative trait loci with epistasis and linkage disequilibrium. *BMC Genet.* 7:9. doi: 10.1186/1471-2156-7-9

Wang, T., and Zeng, Z. B. (2009). Contribution of genetic effects to genetic variance components with epistasis and linkage disequilibrium. *BMC Genet.* 10:52. doi: 10.1186/1471-2156-10-52

Weir, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates, Inc.

Weir, B. S., and Cockerham, C. C. (1977). "Two-locus theory in quantitative genetics," in *Proceedings of the International Conference on Quantitative Genetics*, eds E. Pollak, O. Kempthorne, and T. B. Bailey Jr. (Ames, IA: Iowa State University Press), 247–269.

Yang, R.-C., and Álvarez-Castro, J. M. (2008). Functional and statistical genetic effects with multiple alleles. *Curr. Top. Genet.* 3, 49–62.

Zeng, Z.-B., Wang, T., and Zou, W. (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* 169, 1711–1725. doi: 10.1534/genetics.104.035857

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 10 July 2014; accepted: 03 September 2014; published online: 25 September 2014.*

*Citation: Wang T (2014) A revised Fisher model on analysis of quantitative trait loci with multiple alleles. Front. Genet.* ***5****:328. doi: 10.3389/fgene.2014.00328*

*This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.*

## APPENDIX A. ONE-LOCUS MODEL FOR PHASE KNOWN – IN HARDY-WEINBERG DISEQUILIBRIUM

In HWD, we can represent the genotype frequencies as $P_{(A_i,A_j)} = p^i p_j + D^i_j$, where $D^i_j$ measures the departure from HWE with constraints $\sum_{i=1}^{m} D^i_j = 0$ for $j = 1, \ldots, m$, and $\sum_{j=1}^{m} D^i_j = 0$ for $i = 1, \ldots, m$. Note that $D^i_j = P_{(A_i,A_j)} - p^i p_j = \mathrm{Cov}(x_{P_i}, x_{M_j})$, we can show that $E(G) = \mu + \sum_{i=1}^{m-1}\sum_{j=1}^{m-1} \delta^{*i}_{*j} D^i_j$, and the variance of the expected genotypic values becomes

$$V(E(G|g)) = V_{A_P} + V_{A_M} + V_D + 2\mathrm{Cov}(A_P, A_M) + 2\mathrm{Cov}(A_P, D) + 2\mathrm{Cov}(A_M, D)$$

where $A_P = \sum_{i=1}^{m-1} \alpha^{*i} x_{P_i}(g)$, $A_M = \sum_{j=1}^{m-1} \alpha_{*j} x_{M_j}(g)$ and $D = \sum_{i=1}^{m-1}\sum_{j=1}^{m-1} \delta^{*i}_{*j} x_{P_i}(g) x_{M_j}(g)$. The formulas on calculating the additive variance components $V_{A_P}$ and $V_{A_M}$ are the same as in HWE case. But the formula for dominance variance component $V_D$ becomes

$$\begin{aligned}
V_D = &\sum_{i,j=1}^{m-1} p^i p_j (\delta^{*i}_{*j})^2 - \sum_{i=1}^{m-1} p^i \left(\sum_{j=1}^{m-1} p_j \delta^{*i}_{*j}\right)^2 - \sum_{j=1}^{m-1} p_j \left(\sum_{i=1}^{m-1} p^i \delta^{*i}_{*j}\right)^2 + \left(\sum_{i,j=1}^{m-1} p^i p_j \delta^{*i}_{*j}\right)^2 \\
&+ \sum_{i,j=1}^{m-1} (\delta^{*i}_{*j})^2 D_{ij} - 2\sum_{i=1}^{m-1}\left(\sum_{j=1}^{m-1} p_j \delta^{*i}_{*j}\right)\left(\sum_{j=1}^{m-1} D^i_j \delta^{*i}_{*j}\right) - 2\sum_{j=1}^{m-1}\left(\sum_{i=1}^{m-1} p^i \delta^{*i}_{*j}\right)\left(\sum_{i=1}^{m-1} D^i_j \delta^{*i}_{*j}\right) \\
&+ 2\left(\sum_{i,j=1}^{m-1} p^i p_j \delta^{*i}_{*j}\right)\left(\sum_{i,j=1}^{m-1} D^i_j \delta^{*i}_{*j}\right) + 2\sum_{i,s=1}^{m-1} p^i \left(\sum_{j=1}^{m-1} p_j \delta^{*s}_{*j}\right)\left(\sum_{j=1}^{m-1} D^s_j \delta^{*i}_{*j}\right) - \left(\sum_{i,j=1}^{m-1} \delta^{*i}_{*j} D^i_j\right)^2
\end{aligned}$$

For the covariances, we have

$$\begin{aligned}
\mathrm{Cov}(A_P, A_M) &= \sum_{i,j=1}^{m-1} D^i_j \alpha^{*i} \alpha_{*j} \\
\mathrm{Cov}(A_P, D) &= \sum_{i,j=1}^{m-1} D^i_j \alpha^{*i} \delta^{*i}_{*j} - \left(\sum_{i=1}^{m-1} p^i \alpha^{*i}\right)\left(\sum_{j,s=1}^{m-1} D^s_j \delta^{*s}_{*j}\right) - \sum_{j=1}^{m-1}\left(\sum_{i=1}^{m-1} \alpha^{*i} D^i_j\right)\left(\sum_{s=1}^{m-1} p^s \delta^{*s}_{*j}\right) \\
\mathrm{Cov}(A_M, D) &= \sum_{i,j=1}^{m-1} D^i_j \alpha_{*j} \delta^{*i}_{*j} - \left(\sum_{j=1}^{m-1} p_j \alpha_{*j}\right)\left(\sum_{i,t=1}^{m-1} D^i_t \delta^{*i}_{*t}\right) - \sum_{i=1}^{m-1}\left(\sum_{j=1}^{m-1} \alpha_{*j} D^i_j\right)\left(\sum_{t=1}^{m-1} p_t \delta^{*i}_{*t}\right)
\end{aligned}$$

As the paternal and maternal alleles are correlated in HWD, we will likely have non-zero covariances among $A_P$, $A_M$ and $D$.

## APPENDIX B. ONE-LOCUS MODEL FOR PHASE UNKNOWN – IN HARDY-WEINBERG DISEQUILIBRIUM

In HWD, we can represent the genotype frequencies as $P_{A_iA_i} = p_i^2 + D_{ii}$ for $j = 1, \ldots, m$, and $P_{A_iA_j} = 2p_ip_j + 2D_{ij}$ for $i \neq j$. Since $p_i = P_{A_iA_i} + \sum_{j\neq i} P_{A_iA_j}/2$, we have $\sum_{i=1}^{m} D_{ij} = 0$ for $j = 1, \ldots, m$; $\sum_{j=1}^{m} D_{ij} = 0$ for $i = 1, \ldots, m$; and $D_{ij} = D_{ji}$. As the disequilibrium measures can be represented as $D_{ij} = \mathrm{Cov}(x_{P_i}, x_{M_j})$ for $i, j = 1, \ldots, m$, we can show that $E(G) = \mu + \sum_{i=1}^{m-1} \delta^*_{ii} D_{ii} + 2\sum_{j=2}^{m-1}\sum_{i<j} \delta^*_{ij} D_{ij}$, and the variance partition of the expected genotypic values $V(E(G|g)) = V_A + V_D + 2\mathrm{Cov}(A, D)$, where

$$\begin{aligned}
V_A &= 2\sum_{i=1}^{m-1} p_i (\alpha^*_i)^2 - 2\left(\sum_{i=1}^{m-1} p_i \alpha^*_i\right)^2 + 2\left(\sum_{i=1}^{m-1} D_{ii}(\alpha^*_i)^2 + 2\sum_{i<j} \alpha^*_i \alpha^*_j D_{ij}\right) \\
V_D &= \sum_{i,j=1}^{m-1} p_i p_j (\delta^*_{ij})^2 - 2\sum_{j=1}^{m-1} p_j \left(\sum_{i=1}^{m-1} p_i \delta^*_{ij}\right)^2 + \left(\sum_{i,j=1}^{m-1} p_i p_j \delta^*_{ij}\right)^2 + \sum_{i,j=1}^{m-1} D_{ij} (\delta^*_{ij})^2 - 4\sum_{j=1}^{m-1}\left(\sum_{i=1}^{m-1} p_i \delta^*_{ij}\right)\left(\sum_{i=1}^{m-1} D_{ij} \delta^*_{ij}\right) \\
&\quad + 2\left(\sum_{i,j=1}^{m-1} p_i p_j \delta^*_{ij}\right)\left(\sum_{i,j=1}^{m-1} D_{ij} \delta^*_{ij}\right) + 2\sum_{i,s=1}^{m-1} p_i \left(\sum_{j=1}^{m-1} p_j \delta^*_{sj}\right)\left(\sum_{j=1}^{m-1} D_{sj} \delta^*_{ij}\right) - \left(\sum_{i,j=1}^{m-1} D_{ij} \delta^*_{ij}\right)^2 \\
\mathrm{Cov}(A, D) &= 2\sum_{i,j=1}^{m-1} D_{ij} \alpha^*_i \delta^*_{ij} - 2\left(\sum_{i=1}^{m-1} p_i \alpha^*_i\right)\left(\sum_{s,j=1}^{m-1} D_{sj} \delta^*_{sj}\right) - 2\sum_{j=1}^{m-1}\left(\sum_{i=1}^{m-1} \alpha^*_i D_{ij}\right)\left(\sum_{s=1}^{m-1} p_s \delta^*_{sj}\right)
\end{aligned}$$

In HWD, the genotype frequencies can also be parameterized as $P_{A_iA_i} = p_i^2 + p_i(1-p_i)f$, and $P_{A_iA_j} = 2p_ip_j(1-f)$ for $i \neq j$ (see Weir, 1996). This is a special case of the above parameterization with $D_{ii} = p_i(1-p_i)f$ and $D_{ij} = -p_ip_jf$, for $i \neq j$.