Relationship Between Functional And Statistical Effects Of Multiple Alleles

R.-C. Yang[*,†] and J. M. Álvarez-Castro[‡]

Introduction

Genome-wide association studies have recently been conducted in human and domestic animals to locate and identify chromosomal regions or genes for complex traits (known as QTLs) as an aid for selection of animals with superior performance and quality. These studies are possible because cheap and abundant single nucleotide polymorphisms (SNPs) are increasingly available in many animal species. The increased marker density would increase linkage disequilibrium (LD) between SNP markers and QTLs so that the marker effects may serve as the reliable surrogates of QTL effects (Meuwissen et al. 2001). Consequently, while the effort to locate and estimate the effect of a particular QTL remains important, a growing focus is now on modeling actions and interactions of detected genome-wide QTL effects that contribute collectively to overall genetic merits of animals (see Yang and Alvarez-Castro 2008 for literature review).

A QTL effect can be defined at both genotypic (functional) and gene (statistical) levels. Earlier studies focus on the properties and relationships between functional and statistical genetic effects defined for two alleles at a locus. This paper will examine and clarify how functional and statistical effects are related to each other for multiple alleles [see Yang and Alvarez-Castro (2008) for background and technical details]. Such a relationship for multiple alleles is not always obvious. For example, C.C. Li in Kempthorne (1955) questioned why, in a three-allele case, statistical dominance effects are present when apparently there is lack of functional dominance for some genotype pairs. The question remains unanswered since then. Since many major genes have multiple alleles [e.g., multiple double muscling mutations of myostatin gene in cattle (Bellinge et al. 2005)], it is reasonable to assume that at least some genes affecting complex traits have multiple alleles as well. However, this assumption is not readily tested because typical sizes of mapping populations are not sufficiently large to detect small effects of multiple QTL alleles (Goddard and Hayes 2009). In addition, there is a definite need to model genetic effects for mapping populations derived from multi-way crosses between more than two inbreds or for QTL mapping using multi-allelic molecular markers (e.g., microsatellites).

[*] Agriculture Research Division, Alberta Agriculture and Rural Development, Edmonton, Alberta, Canada T6H 5T6;

[†] Department of Agricultural, Food and Nutritional Science University of Alberta, Edmonton, Alberta, Canada T6G 2P5

[‡] Department of Genetics, University of Santiago de Compostela, 27002 Lugo, Spain

Methods and results

Relationship between functional effects and genotypic values. A vector of genotypic values at a given locus, say locus A ($\mathbf{G}_A$), is transformed to obtain a vector of genetic effects ($\mathbf{E}_{X.A}$) through a genetic-effect design matrix ($\mathbf{S}_{X.A}$), $\mathbf{G}_A = \mathbf{S}_{X.A}\mathbf{E}_{X.A}$, where subscript X represents a transformation operator. For r ($r \geq 1$) alleles at locus A, the number of possible genotypes at locus A is $q = r(r+1)/2$ and the dimensions of $\mathbf{G}_A$, $\mathbf{E}_{X.A}$ and $\mathbf{S}_{X.A}$ are $q \times 1$, $q \times 1$ and $q \times q$, respectively. The special case of $r = 1$ and $q = 1$ (all genotypes are one type of homozygotes) is trivial and thus $\mathbf{G}_A = \mathbf{E}_{X.A}$. In the case of two alleles ($r = 2$ and $q = 3$), $\mathbf{G}_A = [G_{11}\ G_{12}\ G_{22}]'$ where the prime ($'$) denotes matrix or vector transposition and G_{ij} is the value of the genotype carrying alleles A_i and A_j. The genetic-effect vector is $\mathbf{E}_{X.A} = [R_{X.A}\ a_A\ d_A]'$, where $R_{X.A}$ is the reference point for a given transformation operator, a_A is the additive effect measured and d_A is the dominance effect. Since the functional additive and dominance effects are independent of the transformation operator, the reference point can be an arbitrary value including a mean of all genotypic values, a single arbitrarily chosen genotypic value or any numerical number. The only constraint to the choice of the reference point is that matrix $\mathbf{S}_{X.A}$ must be of full rank to ensure two-way transformations: $\mathbf{G}_A = \mathbf{S}_{X.A}\mathbf{E}_{X.A}$ and $\mathbf{E}_{X.A} = \mathbf{S}_{X.A}^{-1}\mathbf{G}_A$ with $\mathbf{S}_{X.A}^{-1}$ being the inverse of $\mathbf{S}_{X.A}$.

Extensions to cases of more than two alleles are straightforward. For example, with three alleles ($r = 3$), there are six possible genotypes: A_1A_1, A_1A_2, A_2A_2, A_1A_3, A_2A_3 and A_3A_3. Thus, vector $\mathbf{G}_A$ is expanded from $\mathbf{G}_A = [G_{11}\ G_{12}\ G_{22}]'$ for $r = 2$ to $\mathbf{G}_A = [G_{11}\ G_{12}\ G_{22}\ G_{13}\ G_{23}\ G_{33}]'$ for $r = 3$. Correspondingly, $\mathbf{E}_{X.A}$ is expanded from $[R_{X.A}\ a_{12}\ d_{12}]'$ to $[R_{X.A}\ a_{12}\ d_{12}\ a_{13}\ d_{13}\ d_{23}]'$, where additive and dominance effects are simply comparisons between pairs of homozygotes and deviations of a given heterozygote from the average of the two corresponding homozygotes, $a_{ij} = (G_{ii} - G_{jj})/2$ and $d_{ij} = G_{ij} - (G_{ii} + G_{jj})/2$. It should noted that a_{23} is missing from the $\mathbf{E}_{X.A}$ vector but it can be recovered from the relation, $a_{23} = a_{12} - a_{13}$. This example illustrates that when there are more than two alleles, the additive effects derived from individual comparisons between pairs of homozygotes are linearly dependent of each other. In general, for $r > 2$, there are r homozygotes and (r-1) basic additive effects are defined as the differences between the values of the reference homozygote (G_{11}) and the remaining (r-1) homozygotes. The remaining (r-1)(r-2)/2 additive effects can be recovered from the basic additive effects as $a_{ji} = a_{1j} - a_{1i}$, $i, j = 2,\ldots,r$.

Matrix $\mathbf{S}_{X.A}$ can be expanded similarly. Let P_{uv} be the frequency of the genotype carrying alleles A_u and A_v with $\Sigma_{u,v\geq u}^{r} P_{uv} = 1$ and let p_u be the frequency of allele A_u with $p_u = \Sigma_{v=1}^{r} P_{uv}$. If the reference point, $R_{X.A}$, is arbitrarily defined as a general function (GF) of all genotypic values, i.e., $GF = \Sigma_{u,v\geq u}^{r} P_{uv} G_{uv}$, then the matrix, $\mathbf{S}_{GF.A}$, can be readily written out.

It is also possible to directly calculate functional genetic effects without the need of finding the design matrix, $\mathbf{S}_{X.A}$. Once again, of $r(r+1)/2$ possible genotypes with r alleles at locus A, there are r homozygotes (A_1A_1, A_2A_2, …, and A_rA_r), and $r(r-1)/2$ heterozygotes (A_1A_2, A_1A_3, …, and $A_{r-1}A_r$). There are numerous ways to define comparisons among these genotypes. One of such definitions includes comparisons that correspond to the following two sets of

meaningful hypotheses: (i) all r homozygotes are functionally equivalent (i.e., all homozygotes have the same genotypic values, $G_{11} = G_{22} = \ldots = G_{rr}$) and (ii) a heterozygote is functionally equivalent to the average of the two corresponding homozygotes [i.e., $G_{uv} = (G_{uu} + G_{vv})/2$]. Testing for hypotheses (i) embodies (r-1) comparisons between a base homozygote value (say G_{11}) and each of the remaining (r-1) homozygote values (i.e., $G_{11} = G_{22}$; $G_{11} = G_{33}$; …; $G_{11} = G_{rr}$); testing for hypotheses (ii) requires $r(r\text{-}1)/2$ comparisons, each being between a heterozygote and the average of the two corresponding homozygotes.

With G_{11} being assigned as the reference point ($R_{G11.A} = G_{11}$), we can freely define (r-1) additive effects, $a_{1v} = (G_{vv} - G_{11})/2$ for $v = 2, 3, \ldots, r$, and we can obtain the remaining $(r\text{-}1)(r\text{-}2)/2$ a's from the relation, $a_{uv} = a_{1u} - a_{1v}$. The $r(r\text{-}1)/2$ dominance effects (d's) are defined as follows, $d_{uv} = G_{uv} - (G_{uu} + G_{vv})/2$ for $u<v$, $u = 1, 2, \ldots, r\text{-}1$. Collecting and writing out all $r(r+1)/2$ equations from $R_{G11.A}$, (r-1) a's and $r(r\text{-}1)/2$ d's in matrix form, functional genetic effects are automatically expressed in terms of the inverse of **S**-matrix multiplied by the vector of genotypic values, $\mathbf{E}_{G11.A} = \mathbf{S}_{G11.A}^{-1}\mathbf{G}_A$. Since the choice of the reference point has no impact on the additive and dominance effects, a convenient choice is the population mean ($R_{\mu.A} = \mu$), giving $\mathbf{E}_{\mu.A} = \mathbf{S}_{\mu.A}^{-1}\mathbf{G}_A$. Furthermore, if deviations of genotypic values from μ are used, then the reference point is zero (i.e., $R_{0.A} = 0$) and $\mathbf{E}_{0.A} = \mathbf{S}_{\mu.A}^{-1}(\mathbf{G}_A - \mathbf{1}\mu)$.

Relationship between statistical effects and genotypic values. The statistical additive and dominance effects for multiple alleles in a Hardy-Weinberg equilibrium (HWE) population are well known (e.g., Lynch and Walsh 1998), $G_A = 1\,\mu + N\,\boldsymbol{\alpha} + \boldsymbol{\delta}$, where N is a matrix of gene content with the elements in the ith column having values of 2, 1 and 0 presenting two, one and zero copies of the ith allele, respectively; $\boldsymbol{\alpha}$ is the vector of statistical additive effects; and $\boldsymbol{\delta}$ is the vector of statistical dominance deviations. The weighted least squares solution of $\boldsymbol{\alpha}$ is given by $\boldsymbol{\alpha} = (\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}\mathbf{N})^{-1}\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}(\mathbf{G}_A - \mathbf{1}\mu)$, where weighting matrix $P_{HWE} = \mathrm{diag}\{ p_1^2, 2p_1p_2, p_2^2, \cdots, p_r^2 \}$ is the diagonal matrix whose diagonal elements are the HWE genotypic frequencies. The statistical dominance deviations are given by, $\boldsymbol{\delta} = \mathbf{G}_A - \mathbf{1}\mu - \mathbf{N}\boldsymbol{\alpha}$.

Relationship between functional effects and statistical effects. The statistical additive effects may be obtained from knowing the functional effects. This relationship can be established by finding a matrix, $\mathbf{T}_\alpha$, such that $\boldsymbol{\alpha} = \mathbf{T}_\alpha\mathbf{E}_{0.A}$. Since $\mathbf{E}_{0.A} = \mathbf{S}_{\mu.A}^{-1}(\mathbf{G}_A - \mathbf{1}\mu)$, it is easy to see that $\mathbf{T}_\alpha = (\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}\mathbf{N})^{-1}\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}\mathbf{S}_{\mu.A}$. Similarly, we can obtain the statistical dominance effects by finding another matrix, $\mathbf{T}_\delta$, $\mathbf{T}_\delta = [\mathbf{I} - \mathbf{N}(\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}\mathbf{N})^{-1}\mathbf{N}'\mathbf{P}_{\mathrm{HWE}}]\mathbf{S}_{\mu.A}$, such that $\boldsymbol{\delta} = \mathbf{T}_\delta\mathbf{E}_{0.A}$. It should be noted that since the first component of $\mathbf{E}_{0.A}$ is zero, the first column of $\mathbf{T}_\alpha$ and $\mathbf{T}_\delta$ can be chosen arbitrarily. For three alleles, we have,

$$\mathbf{T}_\alpha = \begin{bmatrix} \frac{1}{2} & -p_2 & (1-2p_1)p_2 & -p_3 & (1-2p_1)p_3 & -2p_2p_3 \\ \frac{1}{2} & 1-p_2 & p_1(1-2p_2) & -p_3 & -2p_1p_3 & (1-2p_2)p_3 \\ \frac{1}{2} & -p_2 & -2p_1p_2 & 1-p_3 & p_1(1-2p_3) & p_2(1-2p_3) \end{bmatrix}$$

and

$$\mathbf{T}_{\delta} = \begin{bmatrix} 0 & 0 & -2(1-p_1)p_2 & 0 & -2(1-p_1)p_3 & 2p_2p_3 \\ 0 & 0 & 2p_1p_2+p_3 & 0 & 2p_1p_3-p_3 & 2p_2p_3-p_3 \\ 0 & 0 & -2p_1(1-p_2) & 0 & 2p_1p_3 & -2(1-p_2)p_3 \\ 0 & 0 & 2p_1p_2-p_2 & 0 & 2p_1p_3+p_2 & 2p_2p_3-p_2 \\ 0 & 0 & 2p_1p_2-p_1 & 0 & 2p_1p_3-p_1 & 2p_2p_3+p_1 \\ 0 & 0 & 2p_1p_2 & 0 & -2p_1(1-p_3) & -2p_2(1-p_3) \end{bmatrix}$$

Just like in the case of two alleles, the statistical additive effects depend on both functional additive and dominance effects and the statistical dominance deviations depend only on the functional dominance effects. However, it is clear from transformation matrix $\mathbf{T}_{\delta}$ that any nonzero functional dominance effect between a particular pair of alleles is sufficient to cause nonzero values of all statistical dominance deviations.

Conclusion

Much is discussed about the need to distinguish functional effects from statistical effects of alleles at one or more loci, but such discussion is often limited to the case of two alleles. Here, we extend the discussion to cases of multiple alleles. Additionally, we establish a relationship between functional and statistical effects of multiple alleles (r>2). Our extension reveals three new features that do not exist in the diallelic case: (i) there are $r(r-1)/2$ functional additive effects and $r(r-1)/2$ functional dominance effects, but only $(r-1)$ functional additive effects are freely specified and retrieving the remaining $(r-1)(r-2)/2$ ones is possible only for r>2; (ii) the presence of functional dominance effect at only one allele pair is sufficient to cause the presence of statistical dominance deviations for all the genotypes; (iii) the equality of gene frequencies is no longer a sufficient condition for any direct relationship between physiological and statistical genetic effects in the multi-allelic case. Thus, our new multi-allelic models will have a wider range of applications to genome-wide association studies in domestic animals and other organisms.

References

Bellinge, R. H., Liberles, D. A., Iaschi, S. P. *et al*. (2005) *Anim. Genet.* **36**:1-6.

Goddard, M.E, and Hayes, B.J. (2009) *Nat. Review Genetics* 10:381-391.

Kempthorne, O. (1955) *Cold Spring Harbor Symp. Quant. Biol*. 20: 60-78.

Lynch, M., and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001) *Genetics* 157: 1819-1829.

Yang, R.-C., and Álvarez-Castro, J..M. (2008) *Current Topics in Genetics* 3:49-62.