Genet. Sel. Evol. 37 (2005) 129–150

DOI: 10.1051/gse:2004041

Original article

Statistical power for detecting epistasis QTL effects under the F-2 design

Yongcai Mao, Yang Da*

Department of Animal Science, University of Minnesota, Saint Paul, MN 55108, USA

(Received 16 April 2004; accepted 9 September 2004)

Abstract – Epistasis refers to gene interaction effect involving two or more genes. Statistical methods for mapping quantitative trait loci (QTL) with epistasis effects have become available recently. However, little is known about the statistical power and sample size requirements for mapping epistatic QTL using genetic markers. In this study, we developed analytical formulae to calculate the statistical power and sample requirement for detecting each epistasis effect under the F-2 design based on crossing inbred lines. Assuming two unlinked interactive QTL and the same absolute value for all epistasis effects, the heritability of additive × additive (a × a) effect is twice as large as that of additive × dominance (a × d) or dominance × additive (d × a) effect, and is four times as large as that of dominance × dominance (d × d) effect. Consequently, among the four types of epistasis effects involving two loci, 'a × a' effect is the easiest to detect whereas 'd × d' effect is the most difficult to detect. The statistical power for detecting 'a × a' effect is similar to that for detecting dominance effect of a single QTL. The sample size requirements for detecting 'a × d', 'd × a' and 'd × d' are highly sensitive to increased distance between the markers and the interacting QTLs. Therefore, using dense marker coverage is critical to detecting those effects.

epistasis / QTL / statistical power / sample size / F-2

1. INTRODUCTION

Epistasis refers to gene interaction effect involving two or more genes. Evidence from studies in several species, including cattle [18, 21], dogs [1], rat [23], drosophila [8] and humans [10, 20], indicates that epistasis can play a significant role in both quantitative and qualitative characters. Epistasis effects of quantitative trait loci (QTL) have been found in soybean [17], maize [9] and tomato [11]. Among the many models to study epistasis effects, the linear partition of genotypic values into additive, dominance, and epistasis effects by Fisher [12] is considered a classical model for epistasis [6]. Cockerham [5]

* Corresponding author: yda@umn.edu

and Kempthorne [15] further partitioned Fisher's epistasis effects into four components, additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistasis effects with the genetic interpretation of allele × allele, allele × genotype, genotype × allele, and genotype × genotype interactions respectively. The partition of Fisher's epistasis effect by Cockerham and Kempthorne provides a necessary tool to understand the precise nature of gene interactions. The genetic modeling of epistasis by Fisher, Cockerham and Kempthorne can be applied for testing epistasis effects of candidate genes and for mapping interactive QTL using genetic markers. Several statistical methods for mapping interactive QTL have been reported, *e.g.*, the ANOVA method [22], the randomization test [3], the mixture model likelihood analysis based on Cockerham's orthogonal contrast [14], and the Bayesian approach for an outbred population [24]. However, little is known about the statistical power and sample size requirement for detecting each epistasis effect. The purpose of this article is to derive and analyze the statistical power and sample size requirements of the F-2 design for detecting epistasis effects. Formulae for statistical power take into account major factors affecting statistical power and sample size, including separate testing and estimation of additive × additive, additive × dominance, dominance × additive, and dominance × dominance effects, various levels of epistasis effects, marker-QTL distances, type-I error and sample size; formulae for sample size requirements take into account various levels of epistasis effects, marker-QTL distances, type-I and type-II errors.

2. MATERIALS AND METHODS

2.1. Assumptions

Throughout this paper, two unlinked quantitative trait loci (QTL) on different chromosomes, QTL 1 and QTL 2, are assumed. Let A and B denote codominant marker loci linked to QTL 1 and QTL 2, respectively, and let θ_1, and θ_2 denote the recombination frequencies between marker A and QTL 1, and between marker B and QTL 2, respectively. Then, the marker-QTL orders are A-θ_1-QTL1 and B-θ_2-QTL2. Two inbred lines with gene fixation for the markers and QTL are assumed for the convenience of analytical derivations. We assume Line 1 has $AAQ_1Q_1BBQ_2Q_2$ genotype and Line 2 has $aaq_1q_1bbq_2q_2$ genotypes, where A, a, B and b are marker alleles, and Q_1, q_1, Q_2 and q_2 are QTL alleles. The cross between these two lines yields the F-1 generation with $AQ_1/aq_1BQ_2/bq_2$ individuals. The F-2 design is a result of matings

among the F-1 individuals. The least squares partitioning of genotypic values and variances [12, 15] will be used to derive the common genetic modeling for QTL values, assuming Hardy-Weinberg equilibrium. The statistical testing of epistasis effects using genetic markers will use a least squares model, because analytical solutions are available from this method. From this least squares model, elements required for the calculation of statistical power and sample size requirements will be derived, including the marker contrast for testing each epistasis effect and the variance of the contrast. Theoretical results for experimental designs will be compared with simulation studies assuming various levels of epistasis effects, and marker-QTL distance.

2.2. Genetic modeling and marker contrasts

The purpose of genetic modeling of the QTL genotypic values and individual phenotypic values is to establish a theoretical foundation for defining marker contrasts for testing epistasis effects. The least squares partitioning of genotypic values by Kempthrone [15] will be used for the genetic modeling. Let g_{ijkl} = genotypic value of individuals with genotype ij at locus 1 and kl at locus 2, ($i = Q_1$ and $j = q_1$ of locus 1, $k = Q_2$ and $l = q_2$ of locus 2). Then, using Kempthrone's partitioning of genotypic values for the case of two unlinked loci [15, 19], the genotypic value can be modeled as:

$$\begin{aligned} g_{ijkl} &= \mu + (\alpha_i + \alpha_j) + (\alpha_k + \alpha_l) + \delta_{ij} + \delta_{kl} + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\ &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl}) + (\delta\alpha_{ijk} + \delta\alpha_{ijl}) + \delta\delta_{ijkl} \\ &= \mu + \alpha_{ij} + \alpha_{kl} + \delta_{ij} + \delta_{kl} + \alpha\alpha_{ijkl} + \alpha\delta_{ijkl} + \delta\alpha_{ijkl} + \delta\delta_{ijkl} \end{aligned} \tag{1}$$

where μ is the population mean of QTL genotypic values, α_i, α_j, α_k, α_l are the additive effects of QTL allele Q_1, q_1, Q_2, q_2, respectively; δ_{ij}, δ_{kl} are the dominance effects of locus 1 and locus 2, respectively; $\alpha\alpha_{ik}$, $\alpha\alpha_{il}$, $\alpha\alpha_{jk}$, $\alpha\alpha_{jl}$ are the additive × additive effects accounting for the dependency of the effect of an allelic substitution at one locus on the allele present at a second locus; $\alpha\delta_{ikl}$, $\alpha\delta_{jkl}$ are the additive × dominance effects accounting for the interaction of single alleles at locus 1 with the genotype at locus 2; $\delta\alpha_{ijk}$, $\delta\alpha_{ijl}$ are the dominance × additive effects representing the interaction of the genotype at locus 1 with single alleles at locus 2; and $\delta\delta_{ijkl}$ is the dominance × dominance effect representing the interaction between the genotype at locus 1 and the genotype at locus 2. In equation (1), $\alpha_{ij} = \alpha_i + \alpha_j$, $\alpha_{kl} = \alpha_k + \alpha_l$, $\alpha\alpha_{ijkl} = \alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}$, $\alpha\delta_{ijkl} = \alpha\delta_{ikl} + \alpha\delta_{jkl}$, $\delta\alpha_{ijkl} = \delta\alpha_{ijk} + \delta\alpha_{ijl}$. For an F-2 population with equal allele frequencies, it can be shown that the genetic

effects in equation (1) have the following symmetry property:

$$
\begin{aligned}
a_1 &= \alpha_i = -\alpha_j \\
a_2 &= \alpha_k = -\alpha_l \\
d_1 &= \delta_{ii} = -\delta_{ij} = \delta_{jj} \\
d_2 &= \delta_{kk} = -\delta_{kl} = \delta_{ll} \\
i_{aa} &= \alpha\alpha_{ik} = -\alpha\alpha_{il} = -\alpha\alpha_{jk} = \alpha\alpha_{jl} \\
i_{ad} &= \alpha\delta_{ikk} = -\alpha\delta_{ikl} = \alpha\delta_{ill} = -\alpha\delta_{jkk} = \alpha\delta_{jkl} = -\alpha\delta_{jll} \\
i_{da} &= \delta\alpha_{iik} = -\delta\delta_{ijk} = \delta\alpha_{jjk} = -\delta\alpha_{iil} = \delta\alpha_{ijl} = -\delta\alpha_{jjl} \\
i_{dd} &= \delta\delta_{iikk} = -\delta\delta_{iikl} = \delta\delta_{iill} = -\delta\delta_{ijkk} \\
&\quad = \delta\delta_{ijkl} = -\delta\delta_{ijll} = \delta\delta_{jjkk} = -\delta\delta_{jjkl} = \delta\delta_{jjll}.
\end{aligned}
$$

This symmetrical property leads to simplified modeling of equation (1), as shown in Table I. More importantly, this symmetry property will greatly simplify the marker contrasts for testing epistasis effects, allowing simple analytical solutions for evaluating statistical power and sample size requirement, as to be shown later. By combining the nine equations in Table I and solving for μ, a_1, a_2, d_1, d_2, i_{aa}, i_{ad}, i_{da}, and i_{dd}, the unique solutions of the effect parameters in terms of the genotypic values are:

$$\mu = \tfrac{1}{16}(g_{iikk} + 2g_{iikl} + g_{iill} + 2g_{ijkk} + 4g_{ijkl} + 2g_{ijll} + g_{jjkk} + 2g_{jjkl} + g_{jjll}) \tag{2}$$

$$a_1 = \tfrac{1}{16}[(g_{iikk} + 2g_{iikl} + g_{iill}) - (g_{jjkk} + 2g_{jjkl} + g_{jjll})] \tag{3}$$

$$a_2 = \tfrac{1}{16}[(g_{iikk} + 2g_{ijkk} + g_{jjkk}) - (g_{iill} + 2g_{ijll} + g_{jjll})] \tag{4}$$

$$d_1 = \tfrac{1}{16}[(g_{iikk} + 2g_{iikl} + g_{iill}) - 2(g_{ijkk} + 2g_{ijkl} + g_{ijll}) + (g_{jjkk} + 2g_{jjkl} + g_{jjll})] \tag{5}$$

$$d_2 = \tfrac{1}{16}[(g_{iikk} + 2g_{ijkk} + g_{jjkk}) - 2(g_{iikl} + 2g_{ijkl} + g_{jjkl}) + (g_{iill} + 2g_{ijll} + g_{jjll})] \tag{6}$$

$$i_{aa} = \tfrac{1}{16}[(g_{iikk} - g_{jjkk}) - (g_{iill} - g_{jjll})] \tag{7}$$

$$i_{ad} = \tfrac{1}{16}(g_{iikk} - 2g_{iikl} + g_{iill} - g_{jjkk} + 2g_{jjkl} - g_{jjll}) \tag{8}$$

$$i_{da} = \tfrac{1}{16}(g_{iikk} - 2g_{ijkk} + g_{jjkk} - g_{iill} + 2g_{ijll} - g_{jjll}) \tag{9}$$

$$i_{dd} = \tfrac{1}{16}(g_{iikk} - 2g_{iikl} + g_{iill} - 2g_{ijkk} + 4g_{ijkl} - 2g_{ijll} + g_{jjkk} - 2g_{jjkl} + g_{jjll}). \tag{10}$$

In equations (2–10), a_1 = additive effect of QTL 1, d_1 = dominance effect of QTL 1, a_2 = additive effect of QTL 2, d_2 = dominance effect of QTL 2, i_{aa} = additive × additive epistasis effect, i_{ad} = additive × dominance epistasis

Table I. Representation of genotypic values in terms of additive, dominance, and epistasis contributions for the case of two loci with two equally frequent alleles.

Genotype	Frequency	Genotypic Value
$Q_1Q_1Q_2Q_2$	1/16	$g_{iikk} = \mu + 2a_1 + 2a_2 + d_1 + d_2 + 4i_{aa} + 2i_{ad} + 2i_{da} + i_{dd}$
$Q_1Q_1Q_2q_2$	1/8	$g_{iikl} = \mu + 2a_1 + d_1 - d_2 - 2i_{ad} - i_{dd}$
$Q_1Q_1q_2q_2$	1/16	$g_{iill} = \mu + 2a_1 - 2a_2 + d_1 + d_2 - 4i_{aa} + 2i_{ad} - 2i_{da} + i_{dd}$
$Q_1q_1Q_2Q_2$	1/8	$g_{ijkk} = \mu + 2a_2 - d_1 + d_2 - 2i_{da} - i_{dd}$
$Q_1q_1Q_2q_2$	1/4	$g_{ijkl} = \mu - d_1 - d_2 + i_{dd}$
$Q_1q_1q_2q_2$	1/8	$g_{ijll} = \mu - 2a_2 - d_1 + d_2 + 2i_{da} - i_{dd}$
$q_1q_1Q_2Q_2$	1/16	$g_{jjkk} = \mu - 2a_1 + 2a_2 + d_1 + d_2 - 4i_{aa} - 2i_{ad} + 2i_{da} + i_{dd}$
$q_1q_1Q_2q_2$	1/8	$g_{jjkl} = \mu - 2a_1 + d_1 - d_2 + 2i_{ad} - i_{dd}$
$q_1q_1q_2q_2$	1/16	$g_{jjll} = \mu - 2a_1 - 2a_2 + d_1 + d_2 + 4i_{aa} - 2i_{ad} - 2i_{da} + i_{dd}$

effect, i_{da} = dominance × additive epistasis effect, and i_{dd} = dominance × dominance epistasis effect between QTL 1 and QTL 2. Equations (2–10) are foundations for marker contrasts for QTL detection under the F-2 design, and can be used for testing candidate genes in an F-2 design.

When a QTL genotypic value is to be predicted by linked markers, as is the case in QTL detection, the QTL genotypic value can be modeled as:

$$g_{ijkl} = m_{ijkl} + r_{ijkl} \tag{11}$$

where m_{ijkl} is the effect of markers, and r_{ijkl} is the genotypic residual value due to recombination between the markers and QTL. Note that the common genetic mean (μ) term in equation (11) is dropped for convenience of derivations. The two marker models with or without the μ term are equivalent models [13, 19] that achieve the same result for statistical testing. In matrix notation, the QTL genotypic value modeled by genetic markers can be expressed as:

$$\mathbf{g} = \mathbf{Xm} + \mathbf{r} \tag{12}$$

where $\mathbf{g}$ is the column vector of QTL genotypic values, $\mathbf{X}$ is the design matrix for the marker effects, $\mathbf{r}$ is the recombination residual of the QTL value not explained by the common mean and the markers, and $\mathbf{m}$ is the vector of marker effects, *i.e.*,

$$\mathbf{m} = (m_{iikk}, m_{iikl}, m_{iill}, m_{ijkk}, m_{ijkl}, m_{ijll}, m_{jjkk}, m_{jjkl}, m_{jjll})'.$$

The normal equations for equation (12) in matrix notation are $\mathbf{X'Xm} = \mathbf{X'g}$, and the solution to this normal equation is $\mathbf{m} = (\mathbf{X'X})^{-1}\mathbf{X'g}$, where $(\mathbf{X'X})^{-1}$ is the inverse of $\mathbf{X'X}$, and $\mathbf{X'}$ is the transpose of $\mathbf{X}$.

A phenotypic value is modeled as the summation of a QTL genotypic value (g_{ijkl}) and a random residual (e) with N(0,σ_e^2) distribution, *i.e.*,

$$y_{ijkl} = g_{ijkl} + e_{ijkl} = m_{ijkl} + (r_{ijkl} + e_{ijkl}) = m_{ijkl} + \varepsilon_{ijkl} \tag{13}$$

with ε_{ijkl} = phenotypic residual value not explained by the marker effects due to the recombination and random residuals. Using matrix notation, equation (13) can be expressed as:

$$\mathbf{y} = \mathbf{Xm} + (\mathbf{r} + \mathbf{e}) = \mathbf{Xm} + \boldsymbol{\varepsilon}. \tag{14}$$

The normal equations are $\mathbf{X'Xm} = \mathbf{X'y}$, and the estimator of $\mathbf{m}$ is given by

$$\hat{\mathbf{m}} = (\mathbf{X'X})^{-1}\mathbf{X'y}. \tag{15}$$

Let $\hat{\mathbf{m}} = (\hat{m}_{iikk}, \hat{m}_{iikl}, \hat{m}_{iill}, \hat{m}_{ijkk}, \hat{m}_{ijkl}, \hat{m}_{ijll}, \hat{m}_{jjkk}, \hat{m}_{jjkl}, \hat{m}_{jjll})'$ be the least squares estimate of $\mathbf{m} = (m_{iikk}, m_{iikl}, m_{iill}, m_{ijkk}, m_{ijkl}, m_{ijll}, m_{jjkk}, m_{jjkl}, m_{jjll})'$ defined in equation (15), then the four marker contrasts for testing epistasis effects are:

$$L_{aa} = \tfrac{1}{16}\left(\hat{m}_{iikk} - \hat{m}_{iill} - \hat{m}_{jjkk} + \hat{m}_{jjll}\right) \tag{16}$$

$$L_{ad} = \tfrac{1}{16}\left(\hat{m}_{iikk} - 2\hat{m}_{iikl} + \hat{m}_{iill} - \hat{m}_{jjkk} + 2\hat{m}_{jjkl} - \hat{m}_{jjll}\right) \tag{17}$$

$$L_{da} = \tfrac{1}{16}\left(\hat{m}_{iikk} - \hat{m}_{iill} - 2\hat{m}_{ijkk} + 2\hat{m}_{ijll} + \hat{m}_{jjkk} - \hat{m}_{jjll}\right) \tag{18}$$

$$\begin{aligned} L_{dd} = \tfrac{1}{16}\big(&\hat{m}_{iikk} - 2\hat{m}_{iikl} + \hat{m}_{iill} - 2\hat{m}_{ijkk} + 4\hat{m}_{ijkl} - 2\hat{m}_{ijll} + \hat{m}_{jjkk} \\ &-2\hat{m}_{jjkl} + \hat{m}_{jjll}\big) \end{aligned} \tag{19}$$

where L_{aa}, L_{ad}, L_{da} and L_{dd} are the contrasts for testing additive × additive, additive × dominance, dominance × additive, and dominance × dominance effects, respectively.

2.3. Variances of recombination and phenotypic residuals

Following the approach of Bulmer [2] (Eq. (5.1) on page 58), Table I can be expressed more succinctly as:

$$\begin{aligned} g_{ijkl} = \mu &+ 2(z_1 - 1)a_1 + 2(z_2 - 1)a_2 + \left(1 - 4z_1 + 2z_1^2\right)d_1 + \left(1 - 4z_2 + 2z_2^2\right)d_2 \\ &+ 4(z_1 - 1)(z_2 - 1)i_{aa} + 2(z_1 - 1)\left(1 - 4z_2 + 2z_2^2\right)i_{ad} \\ &+ 2\left(1 - 4z_1 + 2z_1^2\right)(z_2 - 1)i_{da} + \left(1 - 4z_1 + 2z_1^2\right)\left(1 - 4z_2 + 2z_2^2\right)i_{dd} \end{aligned} \tag{20}$$

where z_1 is the number of Q_1 alleles in a particular individual at the first putative QTL, and z_2 is the number of Q_2 alleles in a particular individual at the second putative QTL, (z_1, z_2 = 0, 1, or 2 respectively). Equation (20) provides a convenient model for deriving variance and covariance of the QTL genotypic values (App. A). Let σ_g^2 = the total QTL genotypic variance, σ_{A1}^2 = additive variance of QTL 1, σ_{A2}^2 = additive variance of QTL 2, σ_{D1}^2 = dominance variance of QTL 2, σ_{D2}^2 = dominance variance of QTL 2, σ_{AA}^2 = additive × additive variance, σ_{AD}^2 = additive × dominance variance, σ_{DA}^2 = dominance × additive variance, and σ_{DD}^2 = dominance × dominance variance of the two QTLs, then

$$\begin{aligned}\sigma_g^2 &= \sigma_{A1}^2 + \sigma_{A2}^2 + \sigma_{D1}^2 + \sigma_{D2}^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DA}^2 + \sigma_{DD}^2 \\ &= 2a_1^2 + 2a_2^2 + d_1^2 + d_2^2 + 4i_{aa}^2 + 2i_{ad}^2 + 2i_{da}^2 + i_{dd}^2 \end{aligned} \tag{21}$$

with

$$\begin{aligned}&\sigma_{A1}^2 = 2a_1^2, \sigma_{A2}^2 = 2a_2^2, \sigma_{D1}^2 = d_1^2, \sigma_{D2}^2 = d_2^2, \\ &\sigma_{AA}^2 = 4i_{aa}^2, \sigma_{AD}^2 = 2i_{ad}^2, \sigma_{DA}^2 = 2i_{da}^2, \sigma_{DD}^2 = i_{dd}^2. \end{aligned} \tag{22}$$

Derivations for equations (21, 22) are given in Appendix A. The total genetic variance is partitioned into eight independent components, and each variance component is a function of one effect only. This property greatly facilitates the evaluation of the contribution of an effect to the total genetic variance. Note that equivalent partitions can be obtained under alternative models, but they have different meanings in interpreting gene effects, different structures of variance components, and different properties in statistical estimation that may affect the study of QTL [14].

The population variance of recombination residuals for equation (12) is

$$\sigma_r^2 = \tfrac{1}{n}(\mathbf{g}'\mathbf{g} - \mathbf{m}'\mathbf{X}'\mathbf{g}). \tag{23}$$

Applying equation (23) to the F-2 design, and utilizing equation (22), the recombination residual variance of QTL genotypic values is found to be:

$$\begin{aligned}\sigma_r^2 = &\left[1 - (1 - 2\theta_1)^2\right]\sigma_{A1}^2 + \left[1 - (1 - 2\theta_2)^2\right]\sigma_{A2}^2 \\ &+ \left[1 - (1 - 2\theta_1)^4\right]\sigma_{D1}^2 + \left[1 - (1 - 2\theta_2)^4\right]\sigma_{D2}^2 \\ &+ \left[1 - (1 - 2\theta_1)^2(1 - 2\theta_2)^2\right]\sigma_{AA}^2 + \left[1 - (1 - 2\theta_1)^2(1 - 2\theta_2)^4\right]\sigma_{AD}^2 \\ &+ \left[1 - (1 - 2\theta_1)^4(1 - 2\theta_2)^2\right]\sigma_{DA}^2 + \left[1 - (1 - 2\theta_1)^4(1 - 2\theta_2)^4\right]\sigma_{DD}^2. \end{aligned} \tag{24}$$

Derivation of equation (24) is given in Appendix B. The residual variance of phenotypic values for F-2 design under equation (14) is:

$$\sigma_{\varepsilon}^2 = \sigma_r^2 + \sigma_e^2.$$

3. RESULTS

3.1. Mathematical formulae for statistical power and sample size

Statistical power (π) is the probability that an effect is detected when the effect is present, commonly denoted by $\pi = 1 - \beta$, where β is the type II error, *i.e.*, the probability of false 'negatives'. A standardized normal distribution denoted by N(0,1) is assumed for deriving the statistical power. The normal distribution is chosen because the calculation of the exact residual degrees of freedom is unnecessary, providing analytical simplicity. Since the residual degrees of freedom are sufficiently large for the sample sizes discussed in this article ($N \geq 200$), the normal distribution practically yields identical results as the *t*-distribution that is often used in QTL analysis. The general expression for π is:

$$\pi = 1 - \beta = 1 - \Pr(Z < z_i) = 1 - \Phi(z_i) \tag{25}$$

where Z is a N(0,1) normal variable, z_i is the ordinate of the standardized normal curve corresponding to the type II error of β, and Φ is the cumulative distribution function of standard normal random variable. The application of equation (25) to QTL mapping designs requires two key elements: a marker contrast for detecting each epistasis effect, and the variance of the contrast.

Let $\mathbf{c}$ be a contrast vector that defines an estimable function of $\mathbf{m}$, then $\mathrm{E}(\mathbf{c}'\hat{\mathbf{m}}) = \mathbf{c}'\mathbf{m}$. Based on this result and the $\mathbf{m}$ vector in Appendix B, the mathematical expectation of each contrast given by equations (16–19), denoted by $\mathrm{E}(L_i)$, $i = aa, ad, da, dd$, are functions of markers-QTL recombination frequencies and the QTL effects being tested, *i.e.*,

$$\mathrm{E}(L_{aa}) = (1 - 2\theta_1)(1 - 2\theta_2)\, i_{aa} \tag{26}$$

$$\mathrm{E}(L_{ad}) = (1 - 2\theta_1)(1 - 2\theta_2)^2\, i_{ad} \tag{27}$$

$$\mathrm{E}(L_{da}) = (1 - 2\theta_1)^2(1 - 2\theta_2)\, i_{da} \tag{28}$$

$$\mathrm{E}(L_{dd}) = (1 - 2\theta_1)^2(1 - 2\theta_2)^2\, i_{dd}. \tag{29}$$

Using $\mathrm{E}(L_i)$ in place of L_i for $i = aa, ad, da, dd$ as defined by equations (16–19), the z_i value in equation (25) can be expressed as:

$$z_i = z_{\alpha/2} - \frac{\mathrm{E}(L_i)}{\sqrt{\mathrm{var}(L_i)}} = z_{\alpha/2} - \frac{\sqrt{N_i}\mathrm{E}(L_i)}{\sqrt{V_i}}$$

where N_i is the sample size and $V_i = N_i \text{ var}(L_i)$, for $i = aa, ad, da, dd$. For convenience, V_i will be referred to as the 'kernel' of the contrast variance, meaning that V_i differs from $\text{var}(L_i)$ only by a constant of N_i. Let

$$\begin{aligned} h_\varepsilon^2 &= \frac{\sigma_r^2 + \sigma_e^2}{\sigma_y^2} \\ &= 1 - (1-2\theta_1)^2 h_{a_1}^2 - (1-2\theta_2)^2 h_{a_2}^2 - (1-2\theta_1)^4 h_{d_1}^2 - (1-2\theta_2)^4 h_{d_2}^2 \\ &\quad - (1-2\theta_1)^2(1-2\theta_2)^2 h_{aa}^2 - (1-2\theta_1)^2(1-2\theta_2)^4 h_{ad}^2 \\ &\quad - (1-2\theta_1)^4(1-2\theta_2)^2 h_{da}^2 - (1-2\theta_1)^4(1-2\theta_2)^4 h_{dd}^2 \end{aligned} \tag{30}$$

where $h_{a_1}^2 = \sigma_{A1}^2/\sigma_y^2$, $h_{a_2}^2 = \sigma_{A2}^2/\sigma_y^2$, $h_{d_1}^2 = \sigma_{D1}^2/\sigma_y^2$, $h_{d_2}^2 = \sigma_{D2}^2/\sigma_y^2$, $h_{aa}^2 = \sigma_{AA}^2/\sigma_y^2$, $h_{ad}^2 = \sigma_{AD}^2/\sigma_y^2$, $h_{da}^2 = \sigma_{DA}^2/\sigma_y^2$, and $h_{dd}^2 = \sigma_{DD}^2/\sigma_y^2$. For convenience, we will refer to $h_{a_1}^2$ and $h_{a_2}^2$ as the additive heritabilities, $h_{d_1}^2$ and $h_{d_2}^2$ as the dominance heritabilities, and h_{aa}^2, h_{ad}^2, h_{da}^2 and h_{dd}^2 as the additive × additive, additive × dominance, dominance × additive and dominance × dominance heritabilities, respectively. Then, the expressions of V_i in terms of QTL parameters are given as follows:

$$V_{aa} = \tfrac{1}{4}\sigma_y^2 h_\varepsilon^2 \tag{31}$$

$$V_{ad} = \tfrac{1}{2}\sigma_y^2 h_\varepsilon^2 \tag{32}$$

$$V_{da} = \tfrac{1}{2}\sigma_y^2 h_\varepsilon^2 \tag{33}$$

$$V_{dd} = \sigma_y^2 h_\varepsilon^2. \tag{34}$$

The derivations of equations (31–34) are given in Appendix B.

Letting $\lambda_i = \text{E}(L_i)/\sqrt{V_i}$, then z_i in equation (25) can be expressed in terms of QTL parameters as:

$$z_i = z_{\alpha/2} - \sqrt{N_i}\lambda_i \tag{35}$$

where

$$\lambda_{aa} = \frac{(1-2\theta_1)(1-2\theta_2)h_{aa}}{h_\varepsilon} \tag{36}$$

$$\lambda_{ad} = \frac{(1-2\theta_1)(1-2\theta_2)^2 h_{ad}}{h_\varepsilon} \tag{37}$$

$$\lambda_{da} = \frac{(1-2\theta_1)^2(1-2\theta_2)h_{da}}{h_\varepsilon} \tag{38}$$

$$\lambda_{dd} = \frac{(1-2\theta_1)^2(1-2\theta_2)^2 h_{dd}}{h_\varepsilon}. \tag{39}$$

In equations (36–39), θ_1, θ_2 are the marker-QTL recombination frequencies. Theoretical predictions of statistical power for various parameters using equation (25) and equations (35–39) are shown in Figures 1–4. The implications of these results will be discussed along with the results of simulation studies.

Using the above results, the minimum sample size required for given levels of type I and type II errors can be expressed as:

$$N_i = \frac{V_i(Z_{\alpha/2} + Z_\beta)^2}{E^2(L_i)} = \frac{(Z_{\alpha/2} + Z_\beta)^2}{\lambda_i^2} \tag{40}$$

where $Z_{\alpha/2}$ and Z_β are the ordinate of the standardized normal curve corresponding to the probabilities of $\alpha/2$ and β. The sample size given by equation (40) is an increasing function of marker-QTL recombination frequencies, as well as type-I and type-II errors, and a decreasing function of heritability. Sample size requirements obtained from equation (40) for two type-I errors corresponding to the "suggestive" and "significant" linkages proposed by Lander and Kruglyak [16], and different levels of epistasis heritabilities and marker-QTL recombination frequencies are given in Table II, assuming a 95% statistical power.

3.2. Simulation studies on statistical power

Simulation studies were conducted using the Monte Carlo method to evaluate the theoretical results on statistical power for detecting epistasis effects under the F-2 designs. Markers and QTL genotypes were generated such that the true recombination frequencies and each QTL effect used to generate these genotypes can be obtained reversely from the data. A total of 100 sets of marker-QTL genotypes were generated. The phenotypic value of each individual is obtained as the summation of the individual QTL genotypic value and a random residual with N(0,1) distribution. For each set of genotypic data, 10 000 replicates were generated for Figures 1–4. Two interactive QTL without linkage and all epistasis effects have the same absolute value are assumed, thus the heritability of additive × additive (a × a) effect is twice as large as that of additive × dominance (a × d) or dominance × additive (d × a) effect, and is four times as large as that of dominance × dominance (d × d) effect. Heritabilities of additive × additive effect are used in the range of 0.025 to 0.25. Sample sizes of 200–2000 individuals resulting from crossing between inbred lines were generated. The significant levels (type I errors) used were those corresponding to "suggestive linkage" and "significant linkage" proposed by Lander and Kruglyak [16] with type-I errors of 0.0034 and 0.00072 respectively.

Table II. Sample size required to achieve 95% power with a type I error of 5% for the F-2 design.

Heritability ($a_1 = a_2 = d_1 = d_2 = i_{aa} = i_{ad} = i_{da} = i_{dd}$)[a]			$\theta_1 = \theta_2$	a × a	a × d = d × a	d × d
h^2_{aa}	$h^2_{ad} = h^2_{da}$	h^2_{dd}		N	N	N
0.1000	0.0500	0.0250	0	81	162	325
			0.05	150	370	914
			0.1	268	837	2615
			0.15	488	1992	8131
			0.2	942	5233	29074
0.1500	0.0750	0.0375	0	38	76	152
			0.05	84	207	511
			0.1	162	506	1583
			0.15	308	1256	5125
			0.2	608	3376	18758
0.2000	0.1000	0.0500	0	16	32	65
			0.05	51	126	310
			0.1	109	341	1066
			0.15	218	888	3623
			0.2	441	2448	13600

[a] '$a_1 = a_2 = d_1 = d_2 = i_{aa} = i_{ad} = i_{da} = i_{dd}$' indicates that all the eight effects are assumed to be of the same size in defining each heritability. For the F-2 design, $h^2_{aa} = 2h^2_{ad} = 2h^2_{da} = 4h^2_{dd}$ when $a_1 = a_2 = d_1 = d_2 = i_{aa} = i_{ad} = i_{da} = i_{dd}$.

Since the powers for a × d effect and d × a effect as expected identical in the theoretical derivations and almost identical in the simulations, we only show the results for a × d effect, and the results for d × a effect are not included. Table III shows the observed statistical power for epistasis effects for different sample sizes and heritability levels. Statistical powers observed from the simulated data agreed well with the predicted powers as shown in Figures 1–4. Among the four types of epistasis effects involving two loci, 'a × a' effect is the easiest to detect whereas 'd × d' effect is the most difficult to detect. The statistical power for detecting 'a × a' effect is similar to that for detecting dominance effect of a single QTL. The power is poor for detecting 'a × d' or 'd × a' effect and is extremely poor for detecting 'd × d' effect. This trend is consistent across a range of sample sizes (Fig. 1), heritabilities (Fig. 2), and

Table III. Statistical power for detecting epistasis effects based on simulated data ($\theta_1 = \theta_2 = 0.10$).

Sample size	400	400	400	1000	1000	1000
$a_1 = a_2 = d_1 = d_2 = i_{aa} = i_{ad} = i_{da} = i_{dd}$	0.2000	0.2928	0.4472	0.2000	0.2928	0.4472
h^2_{aa}	0.1000	0.1500	0.2000	0.1000	0.1500	0.2000
$h^2_{ad} = h^2_{da}$	0.0500	0.0750	0.1000	0.0500	0.0750	0.1000
h^2_{dd}	0.0250	0.0375	0.0500	0.0250	0.0375	0.0500
$\alpha = 0.0034$ (Suggestive linkage)						
a × a	0.8758	0.9891	0.9965	0.9997	1.0000	1.0000
a × d	0.3172	0.5650	0.7325	0.8186	0.9602	0.9881
d × d	0.0564	0.1179	0.2519	0.2417	0.3764	0.7043
$\alpha = 0.00072$ (Significant linkage)						
a × a	0.7725	0.9664	0.9867	0.9988	1.0000	1.0000
a × d	0.1858	0.3964	0.6332	0.6812	0.9064	0.9768
d × d	0.0220	0.0555	0.1339	0.1309	0.2240	0.5435

recombination frequencies (Fig. 3). The general relationship between power and type-I error is shown in Figure 4 for a wide range of type-I errors. The sample size requirements for detecting 'a × d', 'd × a' and 'd × d' are highly sensitive to increased distance between the markers and the interacting QTLs. Therefore, using dense marker coverage is critical to detecting those effects.

4. DISCUSSION

Epistasis, a potentially important genetic component underlying complex traits, has not been extensively explored in QTL analysis [4]. The results obtained in this study provide some guidelines regarding the statistical power and sample size requirement for detecting epistasis effects under the F-2 design. In general, detecting epistasis effect is more difficult than detecting single QTL effect, except for additive × additive effect, which has about the same power as dominance effect. Detecting epistasis effects involving dominance effect is considerably more challenging than detecting single QTL effect. This difficulty could be reduced to some extent by decreasing marker spacing, because the statistical power for detecting epistasis effects involving dominance is highly sensitive to increased marker spacing (Tabs. I and II, Fig. 3). The statistical power and sample size requirements in this study assume the use of a single marker for detecting epistasis effects. For statistical methods using flanking

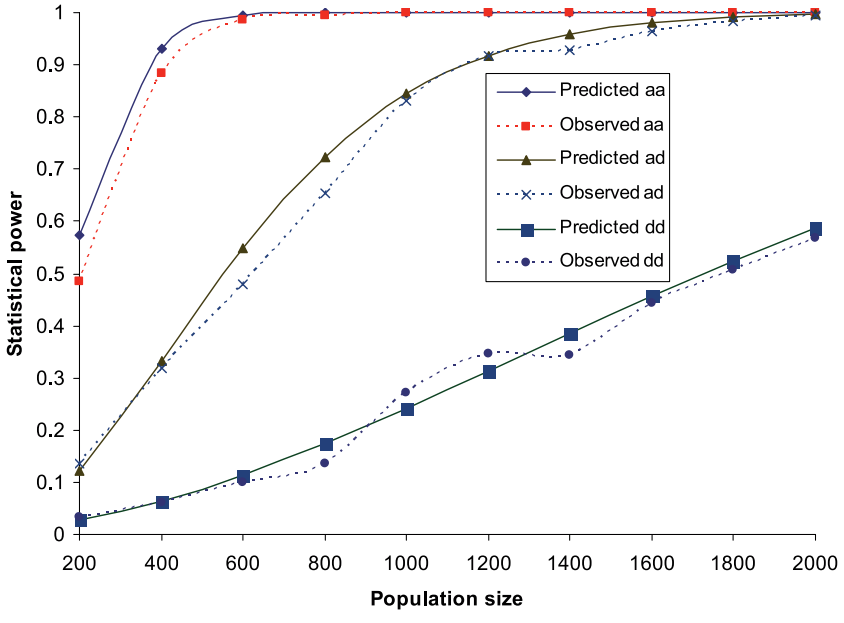

Figure 1. Observed (dotted lines) and predicted (solid lines) statistical power as a function of the population size. (α = 0.0034, $\theta_1 = \theta_2$ = 0.10, h^2_{aa} = 0.1000, $h^2_{ad} = h^2_{da}$ = 0.0500, h^2_{dd} = 0.0250).

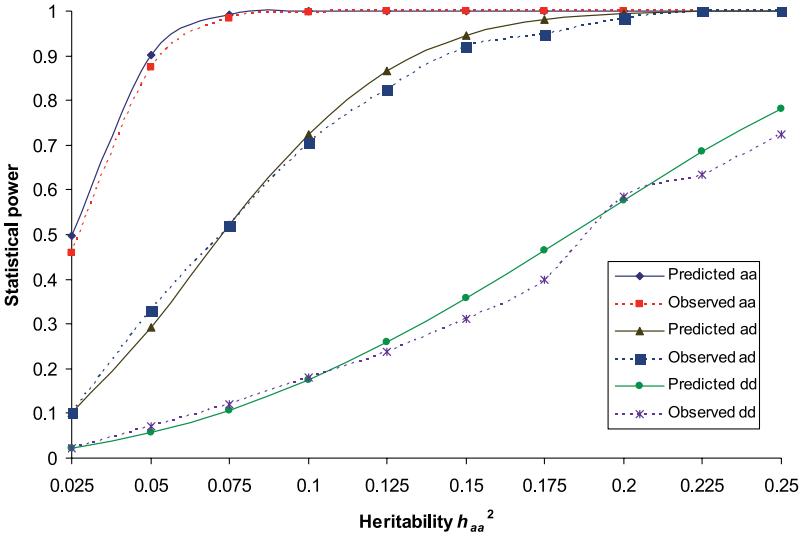

Figure 2. Observed (dotted lines) and predicted (solid lines) statistical power as a function of heritability levels h^2_{aa} with 800 observations (α = 0.0034, $\theta_1 = \theta_2$ = 0.10).

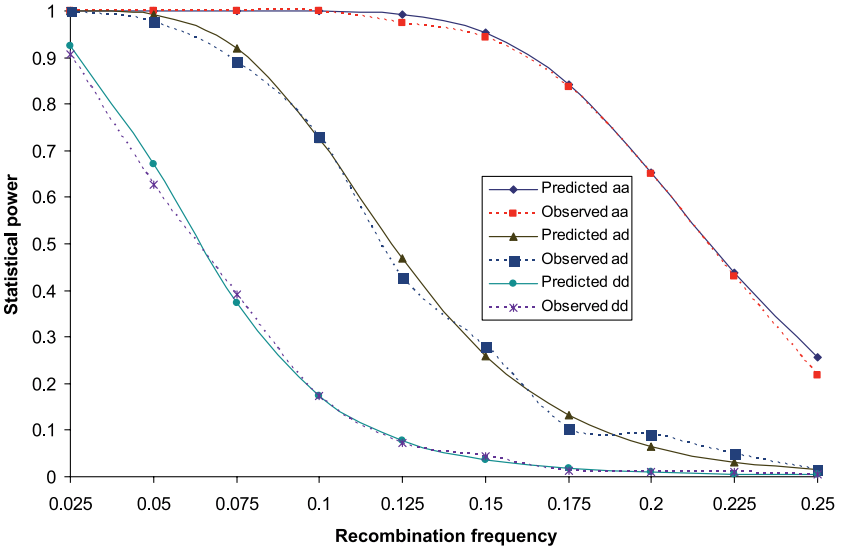

Figure 3. Observed (dotted lines) and predicted (solid lines) statistical power as a function of marker-QTL recombination frequencies $\theta_1 = \theta_2$ with 800 observations ($\alpha = 0.0034$, $h^2_{aa} = 0.1000$, $h^2_{ad} = h^2_{da} = 0.0500$, $h^2_{dd} = 0.0250$).

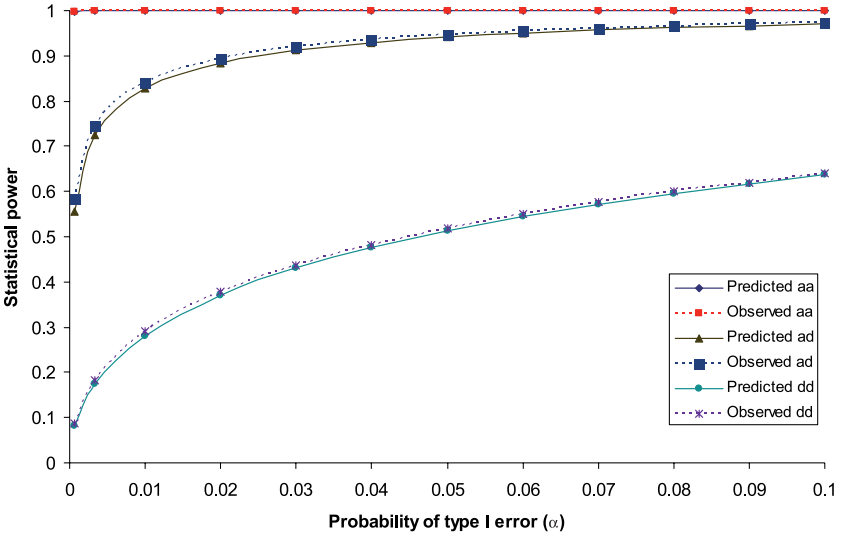

Figure 4. Observed (dotted lines) and predicted (solid lines) statistical power as a function of type I error with 800 observations ($h^2_{aa} = 0.1000$, $h^2_{ad} = h^2_{da} = 0.0500$, $h^2_{dd} = 0.0250$, $\theta_1 = \theta_2 = 0.10$).

markers to detect epistasis effects, results of statistical power could be somewhat overestimates and sample size requirements somewhat underestimates. Extending results in this study to interval mapping is straightforward theoretically but necessarily will require a lengthy development. Since the increase in power of interval mapping over a single marker analysis is only slight [7], results obtained in this study can be considered as close approximation to statistical power and sample size requirements, with statistical power being slightly underestimated and sample size slightly overestimated than those under an interval mapping.

ACKNOWLEDGEMENTS

This research is supported in part by the Agricultural Experiment Station (project MN-16-043) of the University of Minnesota, and by funding from Cargill and the National Research Initiative Competitive Grants Program/United States Department of Agriculture (grant #03275).

REFERENCES

[1] Bourdon R.M., Understanding Animal Breeding, Prentice Hall, 2000, pp. 49–50.
[2] Bulmer M.G., The Mathematical Theory of Quantitative Genetics, Clarendon Press, Oxford, 1980.
[3] Carlborg O., Andersson L., Use of randomization testing to detect multiple epistatic QTLs, Genet. Res. 79 (2002) 175–184.
[4] Carlborg O., Haley C.S., Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. 5 (2004) 618–625.
[5] Cockerham C.C., An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present, Genetics 39 (1954) 859–882.
[6] Cordell H.J., Todd J.A., Hill N.J., Lord C.J., Lyons P.A., Peterson L.B., Wicker L.S., Clayton D.G., Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes, Genetics 158 (2001) 357–367.
[7] Darvasi A., Vinreb A., Minke V., Weller J.I., Soller M., Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map, Genetics 134 (1993) 943–951.
[8] DeSalle R., Slightom J., Zimmer E., The molecular through ecological genetics of abnormal abdomen. II. Ribosomal DNA polymorphism is associated with the abnormal abdomen syndrome in drosophila mercatorum, Genetics 112 (1986) 861–875.
[9] Doebley J., Stec A., Gustus C., *teosinte branched* 1 and the origin of maize: evidence for epistasis and the evolution of dominance, Genetics 141 (1995) 333–346.

[10] El-Hazmi M.A., Warsy A.S., Al-Swailem A.R., Al-Faleh F.Z., Al-Jabbar F.A., Genetic compounds–Hb S, thalassaemias and enzymopathies: spectrum of interactions, J. Trop. Pediatr. 40 (1994) 149–156.
[11] Eshed Y., Zamir D., Less-than-additive epistatic interactions of quantitative trait loci in tomato, Genetics 143 (1996) 1807–1817.
[12] Fisher R.A., The correlation between relatives on the supposition of Mendelian inheritance, Trans. Roy. Soc. Edinburgh 52 (1918) 399–433.
[13] Henderson C.R., Application of linear models in animal breeding, University of Guelph, Ontario, 1984.
[14] Kao C.H., Zeng Z.B., Modeling epistasis of quantitative trait loci using Cockerham's model, Genetics 160 (2002) 1203–1216.
[15] Kempthrone O., The correlation between relatives in a random mating population, Proc. Roy. Soc. London B 143 (1954) 103–113.
[16] Lander E.S., Kruglyak L., Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results, Nature Genet. 11 (1995) 241–247.
[17] Lark K.G., Chase K., Adler F., Mansur L.I., Orf J.H., Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele of another, Proc. Natl. Acad. Sci. USA 92 (1995) 4656–4660.
[18] Long C.R., Gregory K.E., Inheritance of the horned, scurred and polled condition in cattle, J. Hered. 69 (1978) 395–400.
[19] Lynch M., Walsh B., Genetics and Analysis of Quantitative Traits, Sinauer Associates, Sunderland, 1998.
[20] Pedersen J.C., Berg K., Interaction between low density lipoprotein receptor (LDLR) and apolipoprotein E (apoE) alleles contributes to normal variation in lipid level, Clin. Genet. 35 (1989) 331–337.
[21] Potts J.K., Echternkamp S.E., Smith T.P.L., Reecy J.M., Characterization of gene expression in double-muscled and normal-muscled bovine embryos, Anim. Genet. 34 (2003) 438–444.
[22] Routman E.J., Cheverud J.M., Genetic effects on a quantitative trait: two-locus epistatic effects measured at microsatellite markers and at estimated QTL, Evolution 51 (1997) 1654–1662.
[23] Wright S., On the genetics of silvering in the guinea pig with especial reference to interaction and linkage, Genetics 44 (1959) 387–405.
[24] Yi N., Xu S., Allison D.B., Bayesian model choice and search strategies for mapping interacting quantitative trait loci, Genetics 165 (2003) 867–883.

APPENDIX A: PARTITIONING OF GENOTYPIC VALUES AND VARIANCES UNDER THE F-2 DESIGN

Under the F-2 design, the partition of QTL genotypic values is shown in Table I. The QTL alleles are assumed to have equal allele frequency and the two loci are assumed unlinked.

A.1. Analysis of means

Since z_1 is a random variable taking the values 0, 1, and 2 with probabilities 1/4, 1/2, and 1/4. Thus $z_1 - 1$ takes values -1, 0, and 1 with these probabilities, while $1 - 4z_1 + 2z_1^2 = 1 - 2z_1(2 - z_1)$ takes the values 1 and –1 with probabilities 1/2 and 1/2. Hence

$$\mathrm{E}(z_1 - 1) = \mathrm{E}\left(1 - 4z_1 + 2z_1^2\right) = 0.$$

Similarly,

$$\mathrm{E}(z_2 - 1) = \mathrm{E}\left(1 - 4z_2 + 2z_2^2\right) = 0.$$

Also, since the two loci are unlinked,

$$\begin{aligned}
\mathrm{E}[(z_1 - 1)(z_2 - 1)] &= \mathrm{E}(z_1 - 1)\mathrm{E}(z_2 - 1) = 0, \\
\mathrm{E}\left[(z_1 - 1)\left(1 - 4z_2 + 2z_2^2\right)\right] &= \mathrm{E}\left[\left(1 - 4z_1 + 2z_1^2\right)(z_2 - 1)\right] \\
&= \mathrm{E}\left[\left(1 - 4z_1 + 2z_1^2\right)\left(1 - 4z_2 + 2z_2^2\right)\right] = 0.
\end{aligned}$$

Therefore, all the expectations of effect coefficients in equation (20) are zero, taking the expectation of equation (20) we find that

$$E(g) = \mu. \tag{A.1}$$

A.2. Analysis of variances

Since all the expectations of effect coefficients are zero, and the function of z_1 and function of z_2 are independent, we have

$$\begin{aligned}
&\mathrm{cov}\,(z_1 - 1, z_2 - 1) = 0 \\
&\mathrm{cov}\left(z_1 - 1, 1 - 4z_2 + 2z_2^2\right) = 0 \\
&\mathrm{cov}\left(1 - 4z_1 + 2z_1^2, z_2 - 1\right) = 0 \\
&\mathrm{cov}\left(1 - 4z_1 + 2z_1^2, 1 - 4z_2 + 2z_2^2\right) = 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\operatorname{cov}\left[z_1-1,(z_1-1)(z_2-1)\right] &= E\left[(z_1-1)^2(z_2-1)\right] \\
&= E\left[(z_1-1)^2\right]E(z_2-1)=0 \\
\operatorname{cov}\left[z_1-1,(z_1-1)\left(1-4z_2+2z_2^2\right)\right] &= E\left[(z_1-1)^2\right]E\left(1-4z_2+2z_2^2\right)=0 \\
\operatorname{cov}\left[z_1-1,\left(1-4z_1+2z_1^2\right)(z_2-1)\right] &= E\left[(z_1-1)\left(1-4z_1+2z_1^2\right)\right]E(z_2-1)=0 \\
\operatorname{cov}\left[z_2-1,(z_1-1)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[z_2-1,(z_1-1)\left(1-4z_2+2z_2^2\right)\right] &= 0 \\
\operatorname{cov}\left[z_2-1,\left(1-4z_1+2z_1^2\right)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[z_2-1,\left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right] &= 0 \\
\operatorname{cov}\left[1-4z_1+2z_1^2,(z_1-1)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[1-4z_1+2z_1^2,(z_1-1)\left(1-4z_2+2z_2^2\right)\right] &= 0 \\
\operatorname{cov}\left[1-4z_1+2z_1^2,\left(1-4z_1+2z_1^2\right)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[1-4z_1+2z_1^2,\left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right] &= 0 \\
\operatorname{cov}\left[1-4z_2+2z_2^2,(z_1-1)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[1-4z_2+2z_2^2,(z_1-1)\left(1-4z_2+2z_2^2\right)\right] &= 0 \\
\operatorname{cov}\left[1-4z_2+2z_2^2,\left(1-4z_1+2z_1^2\right)(z_2-1)\right] &= 0 \\
\operatorname{cov}\left[1-4z_2+2z_2^2,\left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right] &= 0.
\end{aligned}$$

Since z_1 is a random variable taking the values 0, 1, and 2 with probabilities 1/4, 1/2, and 1/4. Thus $(z_1-1)(1-4z_1+2z_1^2)$ takes values -1, 0, and 1 with these probabilities, therefore,

$$\begin{aligned}
\operatorname{cov}\left(z_1-1,1-4z_1+2z_1^2\right) &= \mathrm{E}\left[(z_1-1)\left(1-4z_1+2z_1^2\right)\right] \\
&= (-1)\times\tfrac{1}{4}+0\times\tfrac{1}{2}+1\times\tfrac{1}{4}=0.
\end{aligned}$$

Similarly,

$$\operatorname{cov}\left(z_2-1, 1-4z_2+2z_2^2\right)=0$$

$$\operatorname{cov}\left[z_1-1, \left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]$$
$$= E\left[(z_1-1)\left(1-4z_1+2z_1^2\right)\right] E\left(1-4z_2+2z_2^2\right)=0$$

$$\operatorname{cov}\left[(z_1-1)(z_2-1), (z_1-1)\left(1-4z_2+2z_2^2\right)\right]$$
$$= E\left[(z_1-1)^2\right] E\left[(z_2-1)\left(1-4z_2+2z_2^2\right)\right]=0$$

$$\operatorname{cov}\left[(z_1-1)(z_2-1), \left(1-4z_1+2z_1^2\right)(z_2-1)\right]=0$$

$$\operatorname{cov}\left[(z_1-1)(z_2-1), \left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]=0$$

$$\operatorname{cov}\left[(z_1-1)\left(1-4z_2+2z_2^2\right), \left(1-4z_1+2z_1^2\right)(z_2-1)\right]=0$$

$$\operatorname{cov}\left[(z_1-1)\left(1-4z_2+2z_2^2\right), \left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]=0$$

$$\operatorname{cov}\left[\left(1-4z_1+2z_1^2\right)(z_2-1), \left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]=0.$$

Therefore, the covariance of any two different effect coefficients is zero. Also,

$$\operatorname{var}(z_1-1)=E\left[(z_1-1)^2\right]=\tfrac{1}{2}$$

$$\operatorname{var}(z_2-1)=\tfrac{1}{2}$$

$$\operatorname{var}\left(1-4z_1+2z_1^2\right)=E\left[\left(1-4z_1+2z_1^2\right)^2\right]=1$$

$$\operatorname{var}\left(1-4z_2+2z_2^2\right)=1$$

$$\operatorname{var}\left[(z_1-1)(z_2-1)\right]=E\left[(z_1-1)^2(z_2-1)^2\right]=E\left[(z_1-1)^2\right]E\left[(z_2-1)^2\right]=\tfrac{1}{4}$$

$$\operatorname{var}\left[(z_1-1)\left(1-4z_2+2z_2^2\right)\right]=E\left[(z_1-1)^2\left(1-4z_2+2z_2^2\right)^2\right]$$
$$=E\left[(z_1-1)^2\right]E\left[\left(1-4z_2+2z_2^2\right)^2\right]=\tfrac{1}{2}$$

$$\operatorname{var}\left[\left(1-4z_1+2z_1^2\right)(z_2-1)\right]=\tfrac{1}{2}$$

$$\operatorname{var}\left[\left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]=E\left[\left(1-4z_1+2z_1^2\right)^2\left(1-4z_2+2z_2^2\right)^2\right]$$
$$=E\left[\left(1-4z_1+2z_1^2\right)^2 E\left(1-4z_2+2z_2^2\right)^2\right]=1.$$

Therefore,

$$\begin{aligned}\sigma_g^2 &= 4\mathrm{var}\,(z_1-1)\,a_1^2 + 4\mathrm{var}\,(z_2-1)\,a_2^2 + \mathrm{var}\left(1-4z_1+2z_1^2\right)d_1^2\\ &\quad + \mathrm{var}\left(1-4z_2+2z_2^2\right)d_2^2 + 16\mathrm{var}\left[(z_1-1)\,(z_2-1)\right]i_{aa}^2\\ &\quad + 4\mathrm{var}\left[(z_1-1)\left(1-4z_2+2z_2^2\right)\right]i_{ad}^2 + 4\mathrm{var}\left[\left(1-4z_1+2z_1^2\right)(z_2-1)\right]i_{da}^2\\ &\quad + \mathrm{var}\left[\left(1-4z_1+2z_1^2\right)\left(1-4z_2+2z_2^2\right)\right]i_{dd}^2\\ &= 2a_1^2 + 2a_2^2 + d_1^2 + d_2^2 + 4i_{aa}^2 + 2i_{ad}^2 + 2i_{da}^2 + i_{dd}^2. \end{aligned} \tag{A.2}$$

APPENDIX B: PROOFS OF RECOMBINATION RESIDUAL VARIANCES, MEANS AND VARIANCES OF CONTRASTS

The general formula for calculating the residual variance is

$$\sigma_r^2 = \tfrac{1}{n}\left(\mathbf{g}'\mathbf{g} - \mathbf{m}'\mathbf{X}'\mathbf{g}\right)$$

as given by equation (23), where $\mathbf{m} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{g}$. The following results can be established:

$$\begin{aligned}\mathbf{g}'\mathbf{g} &= n\left(\tfrac{1}{16}g_{iikk}^2 + \tfrac{1}{8}g_{iikl}^2 + \tfrac{1}{16}g_{iill}^2 + \tfrac{1}{8}g_{ijkk}^2 + \tfrac{1}{4}g_{ijkl}^2 + \tfrac{1}{8}g_{ijll}^2\right.\\ &\quad \left. + \tfrac{1}{16}g_{jjkk}^2 + \tfrac{1}{8}g_{jjkl}^2 + \tfrac{1}{16}g_{jjll}^2\right)\\ \mathbf{X}'\mathbf{X} &= \tfrac{n}{16}\mathrm{Diag}\left\{1,2,1,2,4,2,1,2,1\right\},\end{aligned}$$

where Diag denotes a diagonal matrix,

$$(\mathbf{X}'\mathbf{X})^{-1} = \tfrac{4}{n}\mathrm{Diag}\{4,2,4,2,1,2,4,2,4\}$$

$$\mathbf{X}'\mathbf{g} = \frac{n}{16}\begin{pmatrix} u_1u_2 & u_1b_2 & u_1c_2 & b_1u_2 & b_1b_2 & b_1c_2 & c_1u_2 & c_1b_2 & c_1c_2\\ u_1b_2 & u_1v_2 & u_1b_2 & b_1b_2 & b_1v_2 & b_1b_2 & c_1b_2 & c_1v_2 & c_1b_2\\ u_1c_2 & u_1b_2 & u_1u_2 & b_1c_2 & b_1b_2 & b_1u_2 & c_1c_2 & c_1b_2 & c_1u_2\\ b_1u_2 & b_1b_2 & b_1c_2 & v_1u_2 & v_1b_2 & v_1c_2 & b_1u_2 & b_1b_2 & b_1c_2\\ b_1b_2 & b_1v_2 & b_1b_2 & v_1b_2 & v_1v_2 & v_1b_2 & b_1b_2 & b_1v_2 & b_1b_2\\ b_1c_2 & b_1b_2 & b_1u_2 & v_1c_2 & v_1b_2 & v_1u_2 & b_1c_2 & b_1b_2 & b_1u_2\\ c_1u_2 & c_1b_2 & c_1c_2 & b_1u_2 & b_1b_2 & b_1c_2 & u_1u_2 & u_1b_2 & u_1c_2\\ c_1b_2 & c_1v_2 & c_1b_2 & b_1b_2 & b_1v_2 & b_1b_2 & u_1b_2 & u_1v_2 & u_1b_2\\ c_1c_2 & c_1b_2 & c_1u_2 & b_1c_2 & b_1b_2 & b_1u_2 & u_1c_2 & u_1b_2 & u_1u_2 \end{pmatrix}\begin{pmatrix} g_{iikk}\\ g_{iikl}\\ g_{iill}\\ g_{ijkk}\\ g_{ijkl}\\ g_{ijll}\\ g_{jjkk}\\ g_{jjkl}\\ g_{jjll} \end{pmatrix}$$

with

$$\begin{aligned} &u_1 = (1-\theta_1)^2, b_1 = 2\theta_1(1-\theta_1), c_1 = \theta_1^2, v_1 = 2\left[\theta_1^2 + (1-\theta_1)^2\right]\\ &u_2 = (1-\theta_2)^2, b_2 = 2\theta_2(1-\theta_2), c_2 = \theta_2^2, v_2 = 2\left[\theta_2^2 + (1-\theta_2)^2\right]. \end{aligned}$$

Hence,

$$\mathbf{m} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{g}$$

$$= \begin{pmatrix}
u_1u_2 & u_1b_2 & u_1c_2 & b_1u_2 & b_1b_2 & b_1c_2 & c_1u_2 & c_1b_2 & c_1c_2 \\
\frac{1}{2}u_1b_2 & \frac{1}{2}u_1v_2 & \frac{1}{2}u_1b_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1v_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}c_1b_2 & \frac{1}{2}c_1v_2 & \frac{1}{2}c_1b_2 \\
u_1c_2 & u_1b_2 & u_1u_2 & b_1c_2 & b_1b_2 & b_1u_2 & c_1c_2 & c_1b_2 & c_1u_2 \\
\frac{1}{2}b_1u_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1c_2 & \frac{1}{2}v_1u_2 & \frac{1}{2}v_1b_2 & \frac{1}{2}v_1c_2 & \frac{1}{2}b_1u_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1c_2 \\
\frac{1}{4}b_1b_2 & \frac{1}{4}b_1v_2 & \frac{1}{4}b_1b_2 & \frac{1}{4}v_1b_2 & \frac{1}{4}v_1v_2 & \frac{1}{4}v_1b_2 & \frac{1}{4}b_1b_2 & \frac{1}{4}b_1v_2 & \frac{1}{4}b_1b_2 \\
\frac{1}{2}b_1c_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1u_2 & \frac{1}{2}v_1c_2 & \frac{1}{2}v_1b_2 & \frac{1}{2}v_1u_2 & \frac{1}{2}b_1c_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1u_2 \\
c_1u_2 & c_1b_2 & c_1c_2 & b_1u_2 & b_1b_2 & b_1c_2 & u_1u_2 & u_1b_2 & u_1c_2 \\
\frac{1}{2}c_1b_2 & \frac{1}{2}c_1v_2 & \frac{1}{2}c_1b_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}b_1v_2 & \frac{1}{2}b_1b_2 & \frac{1}{2}u_1b_2 & \frac{1}{2}u_1v_2 & \frac{1}{2}u_1b_2 \\
c_1c_2 & c_1b_2 & c_1u_2 & b_1c_2 & b_1b_2 & b_1u_2 & u_1c_2 & u_1b_2 & u_1u_2
\end{pmatrix}
\begin{pmatrix} g_{iikk} \\ g_{iikl} \\ g_{iill} \\ g_{ijkk} \\ g_{ijkl} \\ g_{ijll} \\ g_{jjkk} \\ g_{jjkl} \\ g_{jjll} \end{pmatrix}$$

$$= \begin{pmatrix}
\mu + 2\tau_1a_1 + 2\tau_2a_2 + \tau_1^2d_1 + \tau_2^2d_2 + 4\tau_1\tau_2i_{aa} + 2\tau_1\tau_2^2i_{ad} \\
+2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd} \\
\mu + 2\tau_1a_1 + \tau_1^2d_1 - \tau_2^2d_2 - 2\tau_1\tau_2^2i_{ad} \\
-\tau_1^2\tau_2^2i_{dd} \\
\mu + 2\tau_1a_1 - 2\tau_2a_2 + \tau_1^2d_1 + \tau_2^2d_2 - 4\tau_1\tau_2i_{aa} + 2\tau_1\tau_2^2i_{ad} \\
-2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd} \\
\mu + 2\tau_2a_2 - \tau_1^2d_1 + \tau_2^2d_2 - 2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd} \\
\mu - \tau_1^2d_1 - \tau_2^2d_2 - \tau_1^2\tau_2^2i_{dd} \\
\mu - 2\tau_2a_2 - \tau_1^2d_1 + \tau_2^2d_2 + 2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd} \\
\mu - 2\tau_1a_1 + 2\tau_2a_2 + \tau_1^2d_1 + \tau_2^2d_2 - 4\tau_1\tau_2i_{aa} - 2\tau_1\tau_2^2i_{ad} \\
+2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd} \\
\mu - 2\tau_1a_1 + \tau_1^2d_1 - \tau_2^2d_2 + 2\tau_1\tau_2^2i_{ad} - \tau_1^2\tau_2^2i_{dd} \\
\mu - 2\tau_1a_1 - 2\tau_2a_2 + \tau_1^2d_1 + \tau_2^2d_2 + 4\tau_1\tau_2i_{aa} - 2\tau_1\tau_2^2i_{ad} \\
-2\tau_1^2\tau_2i_{da} + \tau_1^2\tau_2^2i_{dd}
\end{pmatrix},$$

with $\tau_1 = 1 - 2\theta_1$, $\tau_2 = 1 - 2\theta_2$.

Substituting the above results in equation (23) yields equation (24).

Using equations (16–19), we have:

$$\begin{aligned}
E(L_{aa}) &= \tfrac{1}{16}\left(m_{iikk} - m_{iill} - m_{jjkk} + m_{jjll}\right) \\
&= \tfrac{1}{16}\Big[\left(\mu + 2\tau_1 a_1 + 2\tau_2 a_2 + \tau_1^2 d_1 + \tau_2^2 d_2 + 4\tau_1\tau_2 i_{aa} + 2\tau_1\tau_2^2 i_{ad}\right. \\
&\quad \left. + 2\tau_1^2\tau_2 i_{da} + \tau_1^2\tau_2^2 i_{dd}\right) \\
&\quad - \left(\mu + 2\tau_1 a_1 - 2\tau_2 a_2 + \tau_1^2 d_1 + \tau_2^2 d_2 - 4\tau_1\tau_2 i_{aa} + 2\tau_1\tau_2^2 i_{ad}\right. \\
&\quad \left. - 2\tau_1^2\tau_2 i_{da} + \tau_1^2\tau_2^2 i_{dd}\right) \\
&\quad - \left(\mu - 2\tau_1 a_1 + 2\tau_2 a_2 + \tau_1^2 d_1 + \tau_2^2 d_2 - 4\tau_1\tau_2 i_{aa} - 2\tau_1\tau_2^2 i_{ad}\right. \\
&\quad \left. + 2\tau_1^2\tau_2 i_{da} + \tau_1^2\tau_2^2 i_{dd}\right) \\
&\quad + \left(\mu - 2\tau_1 a_1 - 2\tau_2 a_2 + \tau_1^2 d_1 + \tau_2^2 d_2 + 4\tau_1\tau_2 i_{aa} - 2\tau_1\tau_2^2 i_{ad}\right. \\
&\quad \left. - 2\tau_1^2\tau_2 i_{da} + \tau_1^2\tau_2^2 i_{dd}\right)\Big] \\
&= \tau_1\tau_2 i_{aa}.
\end{aligned}$$

This proves equation (26). Equations (27–29) can be proved similarly.

$$\text{Let } \mathbf{k}_{aa} = \tfrac{1}{16}(1 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 1)'.$$

Then,

$$\text{var}(L_{aa}) = \mathbf{k}'_{aa}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{k}_{aa}(\sigma_r^2 + \sigma_e^2) = \frac{1}{4n}\left(\sigma_r^2 + \sigma_e^2\right).$$

This proves equation (31). Equations (32–34) can be proved similarly.