Heredity (2008) 101, 39–47

www.nature.com/hdy

ORIGINAL ARTICLE

Mapping epistatic quantitative trait loci underlying endosperm traits using all markers on the entire genome in a random hybridization design

X-H He and Y-M Zhang
Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement, National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing, China

Triploid endosperm is of great economic importance owing to its nutritious quality. Mapping endosperm trait loci (ETL) can provide an efficient way to genetically improve grain quality. However, most triploid ETL mapping methods do not produce unbiased estimates of the two dominant effects of ETL. A random hybridization design is an alternative method that may be used to overcome this problem. However, epistasis has an important role in the dissection of genetic architecture for complex traits. In this study, therefore, an attempt was made to map epistatic ETL (eETL) under a triploid genetic model of endosperm traits in a random hybridization design. The endosperm trait means of random hybrid lines, together with known marker genotype information from their corresponding parental F_2 plants, were used to estimate, efficiently and without bias, the positions and all of the effects of eETL using a penalized maximum likelihood method. The method proposed in this article was verified by a series of Monte Carlo simulation experiments. Results from the simulated studies show that the proposed method provides accurate estimates of eETL parameters with a low false-positive rate and a relatively short running time. This new method enables us to map triploid eETL in the same way as diploid quantitative traits.
Heredity (2008) **101,** 39–47; doi:10.1038/hdy.2008.23; published online 7 May 2008

Keywords: endosperm trait; epistasis; penalized maximum likelihood method; quantitative trait loci; triploid inheritance

Introduction

Endosperm, a result of double fertilization in flowering plants, is a triploid tissue whose genetic constitution is consequently more complex than that of common diploid tissue. Endosperm traits, such as protein and amino-acid content in wheat, amylose content and gel consistency in rice, sugar content in sweetcorn and starch and gum content in barley, are of great economic importance because they are directly related to grain quality. Mapping endosperm trait loci (ETL) can provide an efficient way to genetically improve grain quality (Hospital and Charcosset, 1997; Moreau *et al*., 1998; Peleman and Voort, 2003; Servin *et al*., 2004). However, quantitative trait loci (QTL) mapping methods are usually designed for traits that are under diploid control (Lander and Botstein, 1989; Haley and Knott, 1992; Martinez and Curnow, 1992; Jansen, 1993; Zeng, 1994; Kao *et al*., 1999; Xu, 2003, 2007; Zhang and Xu, 2005a, b; Zhang, 2006). The development of a new method for mapping ETL is thus warranted.

The key to understanding the genetic architecture of endosperm traits is found in the study of the properties of individual genes and their interactions. However, classical statistical methodologies (Gale, 1976; Mo, 1987; Bogyo *et al*., 1988; Foolad and Jones, 1992; Pooni *et al*., 1992; Zhu and Weir, 1994) generally focus on partitioning the phenotypic variance of an endosperm trait into genetic and nongenetic (environmental) components, and limit the analysis of the genetic variation to the collective properties of genes. With the advent of molecular markers, QTL mapping became popular. Early QTL mapping used diploid methods to analyze endosperm traits (Tan *et al*., 1999; Wang and Larkins, 2001; Wang *et al*., 2001). This simple treatment failed to take into account the triploid nature of endosperm traits.

To overcome this problem, several approaches have been proposed. Wu *et al*. (2002a, b) pointed out that diploid QTL mapping models require modification to encompass the trisomic inheritance of endosperm traits and the generation difference between a maternal plant and its corresponding endosperm. Such a model requires simultaneous use of two successive generations (two-stage hierarchical design). Theoretically, this can lead to an increase in genetic information extraction from both the maternal plant and its offspring embryo genomes, and in resolution for ETL mapping, compared with a single segregation generation (one-stage) design. Xu *et al*. (2003) expressed the mean value of endosperm traits of $F_{2:3}$ seeds as a dependent variable and the expectations of genotypic indicators for additive and dominant effects of a putative ETL as independent variables for iteratively

Correspondence: Dr Y-M Zhang, College of Agriculture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, China.
E-mail: soyzhang@njau.edu.cn
Received 25 September 2007; revised 17 December 2007; accepted 23 December 2007; published online 7 May 2008

reweighted least-squares mapping. Recently, Hu and Xu (2005) postulated that genetic expression of an endosperm trait may be controlled simultaneously by triploid endosperm and diploid maternal genotypes, and proposed a statistical method for ETL mapping that included maternal genetic effects. However, both of these methods are problematic. First, they handle only models with a single ETL. Only the effects of the putative ETL at the current position are included in the model; all other ETL effects are ignored. Thus, this model is biased in estimating the effects and the positions of ETL provided that multiple and epistatic ETL (eETL) control the trait. Wu *et al*. (2002b) proposed a two-ETL genetic model to detect eETL, but theirs is not a true multiple eETL genetic model. Subsequently, Kao (2004) developed a method of triploid multiple interval mapping (MIM) that combined the triploid nature of endosperm with their diploid MIM (Kao *et al*., 1999).

Second, the existing methods do not produce unbiased estimates of the two dominant effects of ETL. If the genotype of a plant is *QQ* (or *qq*), all the endosperms of the seeds on the plant will be *QQQ* (or *qqq*); if the genotype of a plant is *Qq*, all the endosperms will be 0.25 (*QQQ* + *QQq* + *Qqq* + *qqq*). This means that the first and second dominant effects cannot be distinguished individually, only collectively, so the result is equivalent to that obtained from a diploid genetic model (Wen and Wu, 2006). Wen and Wu (2006) put forward a random hybridization design to estimate the two dominant effects of ETL without bias, but their method does not consider epistasis.

Epistasis, the interaction between QTL, plays an important role in the dissection of genetic architecture for complex traits (Phillips, 1998; Carlborg and Haley, 2004). To date, several approaches have been developed, including the MIM method (Kao and Zeng, 1997; Kao *et al*., 1999), the least-squares multiple regression model (Broman and Speed, 1999), the Bayesian shrinkage estimation method (Xu, 2003; Wang *et al*., 2005; Zhang and Xu, 2005b), stochastic search variable selection methodology derived from George and McMulloch (1993) (Oh *et al*., 2003; Yi *et al*., 2003a, b), the unified Bayesian method (Yi, 2004), the penalized maximum likelihood (PML) method (Zhang and Xu, 2005a) and the empirical Bayes method (Xu, 2007; Xu and Jia, 2007). Most of these are feasible methods for identifying epistatic QTL. Although PML is an all-marker analysis method, it has some advantages. It is simple to use, its result is concise, its running time is much shorter than that of the Bayesian analysis method (Zhang and Xu, 2005a) and it has been proved to be very effective (Broman and Speed, 1999; Xu, 2003; Zhang and Xu, 2005a). Because of these advantages, we used the PML method in our study.

We attempted to detect triploid eETL using a random hybridization design and to estimate, without bias, all effects of eETL, using the PML method.

Method

Experimental design

To form a randomly hybridized population, the parental F_2 population was divided into two groups (maternal and paternal) of equal size. The order of the F_2 plants in each parental group was randomly permuted, and pairs of plants with corresponding order numbers in the two parental groups were crossed. This procedure was repeated until sufficient hybrid lines were obtained. For each hybrid line, the phenotypic value of the endosperm trait and molecular marker information was required. To obtain the phenotypic value of the trait, we measured the mixture of seeds on the maternal plant for each hybrid line to calculate the mean of the line. Molecular marker information was derived from diploid tissues rather than from the triploid endosperm, since the three genotypes *MMM*, *MMm* and *Mmm* could not be distinguished from one another for dominant markers; nor could genotypes *MMm* and *Mmm* be distinguished for co-dominant markers (Wu *et al*., 2002b). Therefore we predicted ETL behavior using marker information from parental F_2 plants. These endosperm trait means of hybrid lines and known marker genotype information from the parental F_2 plants were used to map eETL.

Genetic model for random hybrid line mean of an endosperm trait

Let n be the number of random hybrid (RH) lines and m be the number of markers. We assume that there are no maternal effects affecting endosperm trait expression and that, in the RH population, there is one ETL residing on each marker in the entire genome with two different alleles (Q and q). All pair-wise eETL are considered. The mean of hybrid line j, y_j, for the trait is described by the following genetic model

$$\begin{aligned} y_j = \mu + \sum_{k=1}^{m}(x_{jk}a_k + z_{jk1}d_{k1} + z_{jk2}d_{k2}) + \sum_{r<s}^{m}[(x_{jr}x_{js})i_{a_ra_s} \\ + (x_{jr}z_{js1})i_{a_rd_{s1}} + (x_{jr}z_{js2})i_{a_rd_{s2}} + (z_{jr1}x_{js})i_{d_{r1}a_s} \\ + (z_{jr1}z_{js1})i_{d_{r1}d_{s1}} + (z_{jr1}z_{js2})i_{d_{r1}d_{s2}} + (z_{jr2}x_{js})i_{d_{r2}a_s} \\ + (z_{jr2}z_{js1})i_{d_{r2}d_{s1}} + (z_{jr2}z_{js2})i_{d_{r2}d_{s2}}] + \varepsilon_j \end{aligned} \quad (1)$$

where μ is the population mean; a_k is the additive effect for locus k, which measures the average effect of substituting Q for q; d_{k1} (d_{k2}) is the first (second) dominant effect for locus k, which measures the departure of the substitution effect in QQ (qq) background; $i_{\cdot\cdot}$ is the epistatic effect between two loci (Kao, 2004); ε_j is the residual error with an assumed $N(0, \sigma^2)$ distribution; and x, z_1 and z_2 are dummy variables taking values depending on the genotype combination of the two parental F_2 plants randomly hybridized (Table 1).

We now use l to index the lth genetic effect (the additive, the first and second dominant and epistatic effects) for $l = 1, \ldots, q$. We can rewrite model (1) as

$$y_j = b_0 + \sum_{l=1}^{q} x_{jl}' b_l + \varepsilon_j \quad (2)$$

where $b_0 = \mu$, $q = 1.5m\ (3m-1)$,

$$\begin{aligned} \mathbf{b} = \{b_1, \cdots, b_q\}^{\mathrm{T}} \triangleq \{ & a_1, d_{11}, d_{12}, \cdots, a_m, d_{m1}, d_{m2}, i_{a_1a_2}, i_{a_1d_{21}}, \\ & i_{a_1d_{22}}, i_{d_{11}a_2}, i_{d_{11}d_{21}}, i_{d_{11}d_{22}}, i_{d_{12}a_2}, i_{d_{12}d_{21}}, i_{d_{12}d_{22}}, \\ & \cdots, i_{a_{m-1}a_m}, i_{a_{m-1}d_{m1}}, i_{a_{m-1}d_{m2}}, i_{d_{(m-1)1}a_m}, i_{d_{(m-1)1}d_{m1}}, \\ & i_{d_{(m-1)1}d_{m2}}, i_{d_{(m-1)2}a_m}, i_{d_{(m-1)2}d_{m1}}, i_{d_{(m-1)2}d_{m2}} \}^{\mathrm{T}} \end{aligned}$$

and $\mathbf{x}_l' = \{x_{1l}', \ \ldots, \ x_{nl}'\}^{\mathrm{T}}$ is an $n \times 1$ incidence vector corresponding to the effect b_l ($\forall l = 1, \ldots, q$).

Table 1 Values of dummy variables for x, z_1 and z_2 in random hybridization design of F_2 plants

The kth *marker genotype of* F_2 *plant*		*Genetic constitution for hybrid line for endosperm trait at* kth *locus*	x	z_1	z_2
Maternal	*Paternal*				
M_kM_k	M_kM_k	QQQ	$\frac{3}{2}$	0	0
M_kM_k	M_km_k	$\frac{1}{2}$ $(QQQ+QQq)$	1	$\frac{1}{2}$	0
M_kM_k	m_km_k	QQq	$\frac{1}{2}$	1	0
M_km_k	M_kM_k	$\frac{1}{2}$ $(QQQ+Qqq)$	$\frac{1}{2}$	0	$\frac{1}{2}$
M_km_k	M_km_k	$\frac{1}{4}$ $(QQQ+QQq+Qqq+qqq)$	0	$\frac{1}{4}$	$\frac{1}{4}$
M_km_k	m_km_k	$\frac{1}{2}$ $(QQq+qqq)$	$-\frac{1}{2}$	$\frac{1}{2}$	0
m_km_k	M_kM_k	Qqq	$-\frac{1}{2}$	0	1
m_km_k	M_km_k	$\frac{1}{2}$ $(Qqq+qqq)$	-1	0	$\frac{1}{2}$
m_km_k	m_km_k	qqq	$-\frac{3}{2}$	0	0

Parameter estimation

The PML method (Zhang and Xu, 2005a) was used to estimate the parameters in model (2). The method is briefly described here; for technical detail the reader is referred to the original study (Zhang and Xu, 2005a).

In the PML method, the objective function to be maximized for parameter estimation is the penalized likelihood function, that is, the product of the likelihood function $L(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{M})$ and the penalty function $P(\boldsymbol{\theta}, \boldsymbol{\xi})$. The former is

$$L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M}) = \prod_{j=1}^{n} \varphi(y_j; \mu_j, \sigma^2) \tag{3}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{M}$ is marker information, and $\varphi(y_j; \mu_j, \sigma^2)$ is a normal probability density function with mean μ_j and variance σ^2; the latter is

$$P(\boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{l=1}^{q} [\varphi(b_l; \mu_l, \sigma_l^2)\varphi(\mu_l; 0, \sigma_l^2/\eta)] \tag{4}$$

where $\boldsymbol{\theta} = (b_0, b_1, \ldots, b_q, \sigma^2)$, $\boldsymbol{\xi} = (\mu_1, \ldots, \mu_q, \sigma_1^2, \ldots, \sigma_q^2)$ is the vector of hyperparameters, and $\eta > 0$ is prior sample size for accessing μ_k. Therefore, the penalized likelihood function is

$$\psi(\boldsymbol{\theta}, \boldsymbol{\xi}) = L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M})P(\boldsymbol{\theta}, \boldsymbol{\xi}) \tag{5}$$

The PML estimates for both model parameters and hyperparameters are

$$b_0 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \sum_{l=1}^{q} x_{jl}' b_l\right) \tag{6}$$

$$b_l = \left(\sum_{j=1}^{n} x_{jl}'^2 + \sigma^2/\sigma_l^2\right)^{-1} \times \left[\sum_{j=1}^{n} x_{jl}'(y_j - b_0 - \sum_{t\neq l}^{q} x_{jt}' b_t) + \mu_l\sigma^2/\sigma_l^2\right] \tag{7}$$

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - b_0 - \sum_{l=1}^{q} x_{jl}' b_l\right)^2 \tag{8}$$

$$\mu_l = b_l/(\eta + 1) \tag{9}$$

$$\sigma_l^2 = \frac{1}{2}[(b_l - \mu_l)^2 + \eta\mu_l^2] \tag{10}$$

The procedures for parameter estimation are the same as those used by Zhang and Xu (2005a).

Statistical test

As noted by Zhang and Xu (2005a), the usual likelihood ratio test (LRT) cannot be performed with the PML method because of overparameterization. We proposed the following two-stage selection process to screen the markers (Zhang and Xu, 2005a). In the first stage, all markers with $|\hat{b}/\hat{\sigma}| > 10^{-6}$ are picked up. In the second stage, the epistatic genetic model is modified so that only effects past the first round of selection are included in the model. Owing to the smaller dimensionality of the modified model, we can use the maximum likelihood method to reanalyze the data and perform the LRT. The procedure for the LRT is as follows.

The overall null hypothesis is no effect of ETL at the locus of interest, denoted by H_0: $a = d_1 = d_2 = 0$ or H_0: $\mathbf{Lu} = 0$, where $\mathbf{L} = \{1\,0\,0;\ 0\,1\,0;\ 0\,0\,1\}$ and $\mathbf{u} = \{a\ d_1\ d_2\}^{\mathrm{T}}$. If we determine the maximum likelihood estimates of the parameters under the restriction of $\mathbf{Lu} = 0$ and calculate the log-likelihood value of the solutions with this restriction, we have $L(\hat{\theta} \mid \mathbf{Lu} = 0)$. At the same time, we can also evaluate the log-likelihood value of the solutions without restriction and obtain $L(\hat{\theta})$. Therefore, the LRT statistic is

$$\mathrm{LR} = -2[L(\hat{\theta}|\mathbf{Lu} = 0) - L(\hat{\theta})] \tag{11}$$

Various other statistical tests can be carried out by redefining the $\mathbf{L}$ matrix. To test the hypothesis of H_1: $a = 0$, for example, we define $\mathbf{L}_1 = \{1\,0\,0\}$. The LRT statistic is $\mathrm{LR}_1 = -2\ [L(\hat{\theta} \mid \mathbf{L_1u} = 0) - L(\hat{\theta})]$.

For eETL, we may define $\mathbf{L} = \mathrm{diag}\ (\{1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\})_{15\times 15}$ and $\mathbf{u} = \mathbf{b}$. In the same way, the significance of epistatic effects can be tested. The significance threshold of log of the odds (LOD) score is set at 3.0 where $\mathrm{LOD} = \mathrm{LR}/4.605$.

Simulation studies

Genetic design

We simulated RH populations, with a sample size of 300 in most cases. Twenty-one equally spaced markers were

simulated on three-chromosome segments 360 cM long. We used three main ETL effects and one pair-wise interaction effect, all of which overlapped with markers. All three ETL effects were located at the center (60 cM) of the chromosome. Their genetic parameters were: $a_1 = 2.0$ (marginal variance 5.00), $d_{11} = 5.2$ (marginal variance 5.07) and $d_{12} = -5.2$ (marginal variance 5.07) for the first ETL; $a_2 = 3.0$ (marginal variance 11.25), $d_{21} = 3.0$ (marginal variance 1.69) and $d_{22} = 0.0$ (marginal variance 0.00) for the second ETL; $a_3 = 1.0$ (marginal variance 1.25) and $d_{31} = d_{32} = 0.0$ (marginal variance 0.00) for the third ETL. The eETL was the additive-by-additive interaction between the second and third ETL ($i_{a_2a_3}$) and its effect was set to be equal to 1.50 (marginal variance 3.52). The marginal genetic variances explained by the three main effect ETL were 23.72, 15.19 and 1.25, respectively (Appendix). The total genetic variance for the endosperm trait (σ_g^2) was 43.67. The environmental variance was calculated by $\sigma_e^2 = (1-h^2)\sigma_g^2/h^2$ with h^2 being a 0.50 heritability for most cases. A mixture of ten seeds from each maternal plant for each hybrid line was simulated for the endosperm trait to obtain the mean of the line. To investigate the performance of the proposed method, different cases were considered. Each case was replicated 200 times. For each simulated ETL, we counted the samples in which the LOD statistic had passed 3. A detected ETL within 20 cM of the simulated ETL was considered as a true ETL. The ratio of the number of such samples to the total number of replicates (200) represented the empirical power for this ETL. The false-positive rate was calculated as the ratio of the number of false-positive effects to the total number of zero effects considered in a multiple-ETL genetic model.

Effect of ETL heritability on results of ETL mapping

In the first simulation experiment, we studied the effect of ETL heritability on the results of ETL mapping. The parameters simulated in this experiment, with the exception of ETL heritability, were described in the section on genetic design. By changing the size of residual variance, the total heritability for an endosperm trait was set at four levels: 0.20, 0.40, 0.60 and 0.80. The true and estimated values for the effects and the positions of ETL along with the empirical powers in the detection of ETL are listed in Table 2. As expected, the precision of the estimates of the effects and positions of ETL and the empirical power increase as the heritability increases. Note that the estimates for most of the effects and positions of ETL are unbiased; all coefficients of variance (CV) are below 30%; and the CV falls below ~10%, whereas the marginal variance of a genetic effect accounts for >5% of the total phenotypic variance. We also noted that, in the case of 0.20 heritability, the powers in the detection of d_{21}, a_3 and $i_{a_2a_3}$ are relatively low owing to low genetic variances and explained by their corresponding effects (0.78, 0.57 and 0.69%). In addition, the false-positive rate is low.

Effect of sample size on ETL mapping

In the second experiment, we evaluated the effect of sample size on the results of ETL mapping. By changing the number of RH lines, sample size was set at five levels: 100, 200, 400, 600 and 1000. The results from the simulated experiments are listed in Table 3. They show

Table 2 Effect of ETL heritability on results of ETL mapping in random hybridization design of F_2 plants (200 replicates)

Heritability	*Statistic*	b_0	σ^2	*ETL_1*				*ETL_2*			*ETL_3*		*$ETL_2 \times ETL_3$*		*False-positive rate (%)*				
				a_1	d_{11}	d_{12}	*Posi.*	a_2	d_{21}	*Posi.*	a_3	*Posi.*	$i_{a_2a_3}$	*Posi.*	a	d	i_{aa}	i_{ad}	i_{dd}
True values		100.000	—	2.000	5.200	−5.200	60.00	3.000	3.000	60.00	1.000	60.00	1.500	60.00	60.00					
0.20	Mean	100.710	20.061	1.944	4.891	−5.238	59.95	3.168	3.938	59.81	1.048	60.83	1.534	58.81	58.61	0.14	0.04	0.12	0.00	0.00
	s.d.	0.802	1.867	0.393	1.376	1.366	1.04	0.437	0.862	1.67	0.276	11.52	0.374	10.13	8.13					
	Power			1.000	0.975	0.985		1.000	0.170		0.485		0.505							
0.40	Mean	100.329	8.801	1.970	5.126	−5.037	59.99	3.109	2.966	60.04	0.958	59.51	1.437	59.86	59.63	0.17	0.05	0.12	0.00	0.00
	s.d.	0.500	0.910	0.231	0.833	0.890	0.25	0.266	0.585	1.77	0.243	7.36	0.291	6.34	5.16					
	Power			1.000	1.000	1.000		1.000	0.570		0.865		0.900							
0.60	Mean	100.117	5.099	1.970	5.128	−5.142	59.99	3.047	2.851	59.73	0.944	59.15	1.428	59.27	59.83	0.08	0.04	0.08	0.00	0.00
	s.d.	0.355	0.456	0.166	0.577	0.601	0.06	0.197	0.509	1.87	0.194	4.99	0.232	4.13	2.90					
	Power			1.000	1.000	1.000		1.000	0.885		0.975		0.975							
0.80	Mean	100.053	3.274	2.006	5.169	−5.166	60.00	3.004	2.885	59.95	0.964	59.51	1.452	59.77	59.75	0.03	0.04	0.06	0.00	0.00
	s.d.	0.283	0.339	0.134	0.456	0.443	0.08	0.132	0.395	0.69	0.142	2.96	0.176	2.14	2.36					
	Power			1.000	1.000	1.000		1.000	0.970		0.995		1.000							

Abbreviations: *a*, additive effect; *d*, dominant effect; ETL, endosperm trait locus; *i*, interaction effect; Posi., ETL position (cM); s.d., standard deviation.

Table 3 Effect of sample size on results of ETL mapping in random hybridization design of F_2 plants (200 replicates)

Sample size	*Statistic*	b_0	σ^2	*ETL_1*				*ETL_2*			*ETL_3*		*$ETL_2 \times ETL_3$*			*False-positive rate (%)*				
				a_1	d_{11}	d_{12}	*Posi.*	a_2	d_{21}	*Posi.*	a_3	*Posi.*	$i_{a_2a_3}$	*Posi.*	*Posi.*	a	d	i_{aa}	i_{ad}	i_{dd}
True value		100.000	—	2.000	5.200	−5.200	60.00	3.000	3.000	60.00	1.000	60.00	1.500	60.00	60.00					
100	Mean	100.728	8.250	2.070	5.875	−5.820	59.84	3.151	4.073	59.89	1.283	57.69	1.746	57.54	57.94	0.03	0.00	0.04	0.00	0.00
	s.d.	1.297	1.765	0.536	1.681	1.722	1.84	0.418	.	1.57	0.238	9.42	0.529	12.63	8.36					
	Power			0.985	0.730	0.720		1.000	0.005		0.260		0.140							
200	Mean	100.474	6.945	1.961	5.086	−5.039	59.97	3.102	3.218	59.94	0.952	58.75	1.387	58.94	58.84	0.11	0.04	0.07	0.00	0.00
	s.d.	0.611	0.901	0.273	0.978	0.936	0.61	0.309	0.599	1.88	0.212	9.23	0.328	8.55	8.06					
	Power			1.000	0.995	0.995		1.000	0.345		0.720		0.760							
400	Mean	100.095	6.450	2.010	5.123	−5.187	59.99	2.998	2.938	59.95	0.952	59.57	1.431	59.52	59.77	0.11	0.05	0.10	0.01	0.00
	s.d.	0.326	0.480	0.197	0.532	0.519	0.22	0.186	0.448	0.44	0.175	3.63	0.216	3.13	1.78					
	Power			1.000	1.000	1.000		1.000	0.940		0.985		1.000							
600	Mean	100.058	6.510	1.988	5.141	−5.204	60.00	3.005	2.866	60.00	0.973	59.93	1.475	59.96	60.07	0.11	0.09	0.09	0.02	0.00
	s.d.	0.244	0.368	0.133	0.431	0.426	0.04	0.147	0.401	0.03	0.142	1.41	0.159	0.48	0.55					
	Power			1.000	1.000	1.000		1.000	1.000		1.000		1.000							
1000	Mean	100.012	6.570	2.004	5.172	−5.201	60.00	3.003	2.956	60.00	0.993	59.99	1.476	60.02	59.96	0.17	0.10	0.13	0.03	0.00
	s.d.	0.186	0.311	0.108	0.353	0.326	0.02	0.109	0.302	0.00	0.118	0.63	0.131	0.25	0.33					
	Power			1.000	1.000	1.000		1.000	1.000		1.000		1.000							

Abbreviations: *a*, additive effect; *d*, dominant effect; *i*, interaction effect; s.d., standard deviation; ETL, endosperm trait locus; Posi., ETL position (cM).

Table 4 Effect of the number of seeds per maternal plant on results of ETL mapping in random hybridization design of F_2 plants (200 replicates)

No. of seeds per plant	*Statistic*	b_0	σ^2	*ETL_1*				*ETL_2*			*ETL_3*		*$ETL_2 \times ETL_3$*			*False-positive rate (%)*				
				a_1	d_{11}	d_{12}	*Posi.*	a_2	d_{21}	*Posi.*	a_3	*Posi.*	$i_{a_2a_3}$	*Posi.*	*Posi.*	a	d	i_{aa}	i_{ad}	i_{dd}
True value		100.000	—	2.000	5.200	−5.200	60.00	3.000	3.000	60.00	1.000	60.00	1.500	60.00	60.00					
1	Mean	100.740	67.579	2.292	7.604	−7.397	59.51	3.049	6.486	59.14	1.551	56.52	2.300	60.77	60.77	0.11	0.00	0.12	0.00	0.00
	s.d.	1.745	6.736	0.708	2.111	1.927	4.06	0.707	1.441	5.41	0.310	11.52	0.446	19.98	19.17					
	Power			0.785	0.455	0.490		0.980	0.010		0.115		0.130							
3	Mean	100.712	22.778	1.908	5.227	−5.510	59.87	3.149	3.922	59.91	1.106	58.65	1.558	58.22	59.04	0.33	0.00	0.11	0.00	0.00
	s.d.	0.937	2.118	0.405	1.481	1.404	1.10	0.357	0.693	0.64	0.327	13.78	0.398	11.37	11.22					
	Power			0.995	0.925	0.935		1.000	0.115		0.370		0.415							
5	Mean	100.410	13.429	1.943	5.185	−4.919	59.95	3.141	3.335	59.83	0.991	57.27	1.423	58.43	58.32	0.11	0.05	0.11	0.00	0.00
	s.d.	0.666	1.194	0.313	1.059	1.055	0.50	0.306	0.569	1.23	0.279	10.71	0.325	9.04	7.64					
	Power			1.000	0.995	0.990		1.000	0.325		0.660		0.765							
10	Mean	100.273	6.655	1.951	5.094	−5.176	60.00	3.057	2.965	59.87	0.950	59.44	1.416	59.14	59.25	0.14	0.05	0.10	0.00	0.00
	s.d.	0.480	0.611	0.208	0.735	0.762	0.17	0.259	0.532	1.04	0.214	5.89	0.267	5.80	5.49					
	Power			1.000	1.000	1.000		1.000	0.710		0.950		0.975							
20	Mean	100.075	3.244	1.991	5.149	−5.157	60.00	3.010	2.843	59.74	0.949	59.67	1.459	59.87	59.81	0.08	0.06	0.08	0.00	0.00
	s.d.	0.299	0.330	0.144	0.447	0.431	0.12	0.147	0.442	1.86	0.151	2.47	0.178	1.75	2.96					
	Power			1.000	1.000	1.000		1.000	0.955		0.995		0.995							

Abbreviations: *a*, additive effect; *d*, dominant effect; ETL, endosperm trait locus; *i*, interaction effect; Posi., ETL position (cM); s.d., standard deviation.

Table 5 Effect of sampling strategy on results of ETL mapping in random hybridization design of F_2 plants (200 replicates)

Sampling strategy	*Statistic*	b_0	σ^2	*ETL_1*				*ETL_2*			*ETL_3*		*$ETL_2 \times ETL_3$*		*False-positive rate (%)*				
				a_1	d_{11}	d_{12}	*Posi.*	a_2	d_{21}	*Posi.*	a_3	*Posi.*	$i_{a_2a_3}$	*Posi.*	a	d	i_{aa}	i_{ad}	i_{dd}
True value		100.000	—	2.000	5.200	−5.200	60.00	3.000	3.000	60.00	1.000	60.00	1.500	60.00					
600×5^a	Mean	100.092	13.245	1.992	5.137	−5.064	60.00	3.042	2.874	59.58	0.931	59.88	1.443	59.81	0.11	0.06	0.09	0.01	0.00
	s.d.	0.413	0.846	0.202	0.676	0.664	0.17	0.231	0.562	2.22	0.199	5.79	0.258	3.07					
	Power			1.000	1.000	1.000		1.000	0.915		0.955		0.985						
300×10	Mean	100.211	6.575	1.970	5.172	−5.080	59.99	3.052	2.928	59.60	0.925	58.71	1.437	58.78	0.06	0.10	0.09	0.00	0.00
	s.d.	0.454	0.598	0.224	0.703	0.705	0.29	0.237	0.508	2.42	0.200	6.81	0.306	6.24					
	Power			1.000	1.000	1.000		1.000	0.730		0.940		0.920						
200×15	Mean	100.350	4.561	1.979	5.055	−5.061	59.96	3.115	2.967	59.60	0.917	58.91	1.416	59.01	0.06	0.03	0.08	0.00	0.00
	s.d.	0.482	0.591	0.209	0.711	0.715	0.28	0.251	0.540	2.43	0.211	6.51	0.271	6.17					
	Power			1.000	1.000	1.000		1.000	0.575		0.870		0.910						
150×20	Mean	100.474	3.775	1.991	5.073	−5.031	59.98	3.140	3.099	59.85	0.967	58.43	1.390	59.03	0.00	0.04	0.04	0.00	0.00
	s.d.	0.543	0.645	0.244	0.807	0.817	0.18	0.267	0.563	1.22	0.239	7.87	0.301	8.09					
	Power			1.000	1.000	1.000		1.000	0.350		0.760		0.700						
100×30	Mean	100.734	3.348	1.960	5.041	−5.094	60.02	3.189	3.296	59.99	1.018	58.81	1.401	59.16	0.03	0.03	0.03	0.00	0.00
	s.d.	0.644	0.873	0.260	0.946	1.020	0.50	0.305	0.736	0.22	0.264	7.91	0.351	10.95					
	Power			1.000	0.960	0.975		1.000	0.060		0.505		0.390						

Abbreviations: *a*, additive effect; *d*, dominant effect; ETL, endosperm trait locus; *i*, interaction effect; Posi., ETL position (cM); s.d., standard deviation.
[a]5 seeds are sampled from each of 600 F_2 maternal plants.

the general behavior of QTL mapping: as sample size increases, the result improves (as judged by the decrease in the standard deviation and the increase in empirical power). When sample size is above 400, accurate estimates and high power can be achieved, even for small genetic effects d_{21}, a_3 and $i_{a_2a_3}$ (marginal heritabilities are 1.95, 1.44 and 1.73%, respectively).

Effect of the number of seeds per plant on ETL mapping

This simulation experiment aims to evaluate the effect of the number of seeds per maternal plant on the results of ETL mapping. We set the number of seeds per plant at five levels: 1, 3, 5, 10 and 20. The results are given in Table 4. We found that, when the number of seeds per plant was more than 10, all parameters were accurately and precisely estimated. Indeed the power was high, even when there were only three seeds. Therefore, the results are robust.

Effect of sampling strategy on ETL mapping

The effect of sampling strategy on the results of ETL mapping was investigated. We evaluated five schemes of sampling strategy: 600×5 (5 seeds were sampled from each of 600 F_2 maternal plants), 300×10, 200×15, 150×20 and 100×30. The results of 200 replicated simulations are summarized in Table 5. We observed the expected trend of an increase in power as the number of hybrid lines increased; the number of hybrid lines was more important than the number of seeds per maternal plant. The reason for this may be that a larger number of hybrid lines can provide more marker information.

A simulated example of a large genome

Finally, we simulated a large genome 1260 cM long to explore the performance of the proposed method in real data analysis. The genome consisted of 12 chromosomes, each covered by eight evenly spaced markers with a 15 cM per marker interval. The simulated parameters are listed in Table 6 for main effects and in Table 7 for epistatic effects. By changing the size of the residual variance, the total heritability for an endosperm trait was set at 0.60. The total number of ETL effects included in the model was $1.5 \times 96 \times (3 \times 96 - 1) = 41\,328$. We increased the sample size to 600. The number of effects was about 68 times as large as the sample size. Obviously, it was overloaded. At this juncture, a two-stage method was proposed. In the first stage, a full model that included all of the main and pair-wise epistatic effects was divided into many reduced models, each with all of the main effects and proportion of the epistatic effects. It was feasible to estimate the parameters of each reduced model using the PML method. In this way, individual effects apart from zero could be discerned. In the second stage, we modified our epistatic genetic model so that only effects past the first round of the selection were included in the model and we could use the PML method to reanalyze the data. The results are listed in Tables 6 and 7. They show that all ETL are detected with the exception of an eETL with a dominant-by-dominant effect, and that the effects and positions of the detected ETL are close to their corresponding true values. For the undetected eETL, the genetic variance explained by its effect is relatively low. In addition, three false-positive eETL with additive-by-additive epistatic

Table 6 Simulated and estimated ETL positions and effects from a single data set of a large genome

ETL[a]	*Chromosome*	*True parameter*				*Estimate*				*LOD*		
		Posi.	a	d_1	d_2	*Posi.*	a	d_1	d_2	a	d_1	d_2
ETL_1	1	45.00	2.000	0.000	0.000	45.00	1.957	0.000	0.000	133.75	—	—
ETL_2	3	240.00	0.000	2.000	0.000	240.00	0.000	1.698	0.000	—	18.56	—
ETL_3	3	300.00	1.000	2.500	−2.500	300.00	0.980	2.679	−2.441	39.77	24.98	23.04
ETL_4	4	375.00	0.000	0.000	0.000	375.00	0.000	0.000	0.000	—	—	—
ETL_5	5	465.00	0.000	0.000	0.000	465.00	0.000	0.000	0.000	—	—	—
ETL_6	8	765.00	0.000	0.000	2.000	765.00	0.000	0.000	2.135	—	—	24.78
ETL_7	10	1020.00	0.500	−2.500	0.000	1035.00(*a*) 1020.00(d_1)	0.437	−2.556	0.000	10.33	40.04	—
ETL_8	11	1110.00	−2.000	0.000	0.000	1110.00	−1.990	0.000	0.000	143.40	—	—

Abbreviations: *a*, additive effect; *d*, dominant effect; ETL, endosperm trait locus; LOD, log of the odds; Posi., ETL position (cM).
[a]The same is true for Table 7.

Table 7 Simulated and estimated positions and effects of interacting ETL from a single dataset of a large genome

Epistasis	*Type of interaction*	*True parameter*			*Estimate*			*LOD*
		Posi. A	*Posi. B*	*Effect*	*Posi. A*	*Posi. B*	*Effect*	
$ETL_1 \times ETL_2$	Additive-by-additive	45.00	240.00	−1.000	45.00	240.00	−0.986	30.93
$ETL_2 \times ETL_6$	Dominance-by-additive	240.00	765.00	3.000	240.00	765.00	2.988	57.26
$ETL_3 \times ETL_6$	Dominance-by-dominance	300.00	765.00	1.000	300.00	765.00	Missing	—
$ETL_3 \times ETL_7$	Additive-by-additive	300.00	1020.00	1.000	300.00	1020.00	0.908	28.69
$ETL_4 \times ETL_5$	Additive-by-additive	375.00	465.00	1.500	375.00	465.00	1.654	66.23

Abbreviations: LOD, log of the odds; Posi., ETL position (cM).

effects were identified. However, their effects are small, and their LOD values for LRT are about 5 (data not shown)—much less than those for true ETL. Thus, the new method works well.

Discussion

Genetic improvement of grain production and quality is a major aim in plant breeding. Endosperm is a main part of grain seed and many endosperm traits are directly related to grain quality, so endosperm traits are of great importance. To uncover their genetic architecture, several methods of mapping ETL have been proposed (Wu *et al*., 2002a, b; Xu *et al*., 2003; Kao, 2004; Hu and Xu, 2005; Wen and Wu, 2006). These triploid-based methods are all superior to diploid methods for ETL mapping. The method described here, however, offers advantages over triploid-based methods. As in Kao (2004) method, it allows for a model that includes all main and pair-wise epistatic effects, in contrast to other methods in which only a single ETL genetic model is considered (Wu *et al*., 2002a, b; Xu *et al*., 2003; Hu and Xu, 2005; Wen and Wu, 2006). In our new model, biased estimates will not occur if there are linked or eETL. However, our method differs from Kao (2004) method, in which genetic model determination relies on the adoption of a critical statistic whose true distribution is very difficult to determine. The usual technique is the permutation test (Churchill and Doerge, 1994; Kao, 2004), which is very time consuming. In our new method, model selection is unnecessary, and the best model can always be captured (Zhang and Xu, 2005a). Along with Wen and Wu (2006) method, our method can provide unbiased estimates for the first and second dominant effects and corresponding epistatic effects. However, our method differs in that theirs handles only a model with a single ETL. In addition, our method is economical and easy to implement. Although Wu *et al*. (2002b) and Kao (2004) proposed a more advanced two-stage design (with marker information collected from maternal plant and seed embryo), it is difficult to put into practice. The reasons are technical difficulty, imprecise single-seed phenotype measurement, and the high cost of marker assay. In our method, bulked endosperm trait measurement is used for phenotype data, and F_2 plant tissue for marker data.

Another major concern is how the PML method deals with a multiple ETL model that potentially can assume one ETL residing on each marker position. A number of questions arise in this regard. First, what are those markers' false-positive rates? The results in Tables 2–5 indicate that if a marker is not associated with a trait, its genetic effect on the locus shrinks to nearly zero. The same result is seen in the simulated experiment with a large genome, and in Zhang and Xu (2005a). Therefore, the false-positive rate is low.

Second, how do we analyze real data? The procedure necessitates pretreatment to deal with dominant and missing markers and marker density. Marker imputation techniques may be used in the case of incomplete information marker data (Xu, 2007). They involve the calculation of the conditional probability of marker genotypes using a multipoint method (Jiang and Zeng, 1997), and the sampling of a complete imputed data set for the marker genotypes. Usually, 10–20 imputed data sets are generated (Sen and Churchill, 2001; Xu, 2007). The reported result is the mean of estimates for each imputed data set. When marker density is too high,

choosing one marker from the cluster of markers avoids a high degree of multicollinearity (Zhang and Xu, 2005a). When the marker is too sparse, a virtual marker (treated as missing data) may be inserted.

Third, is the number of markers that can be applied using the PML method limited? It is preferable to gather more samples or reduce the number of effects considered in the model (Zhang and Xu, 2005a; Hoti and Sillanpää, 2006). If the number of markers is large, however, the number of effects in the model is enormous—more than 40 000 in the simulated experiment with a large genome. In this case, a two-stage method, taking about 22 h, is recommended. The results in Tables 6 and 7 show that this works well, and a further study is under way.

Fourth, how can we fine-map ETL? Although our method, a type of marker analysis, is inadequate for fine-mapping, its strategy has been proved to be very effective (Broman and Speed, 1999; Xu, 2003; Zhang and Xu, 2005a), and we can use the result derived from this method as a starting point for other methods based on a multiple-ETL model, such as Kao (2004) method. Combining the two methods can provide stable model determination and high resolution. Moreover, extension to ETL with epistatic effects, making use of the PML framework, is under way and may be used to fine-map ETL.

It should be noted that in our study an additive-by-additive effect was simulated for most cases. This is because the effect has a relatively high proportion of genetic variance (Appendix) and is easily detected. Larger sample sizes are recommended to explore other kinds of epistatic effects.

Acknowledgements

We thank the subject editor and two anonymous reviewers for their comments on the first version of this article. The work was supported in part by: 973 program (2006CB101708), the National Natural Science Foundation of China (30671333), 863 program (2006AA10Z1E5), Specialized Research Fund for the Doctoral Program of Higher Education (20060307008) and NCET (NCET-05-0489) to YMZ.

References

Bogyo TP, Lance RCM, Chevalier P, Nilan RA (1988). Genetic models for quantitatively inherited endosperm characters. *Heredity* **60**: 61–67.

Broman KW, Speed TP (1999). A review of methods for identifying QTLs in experimental crosses. In: Seillier-Moiseiwitsch F (ed). *Statistics in Molecular Biology and Genetics*. IMS Lecture Notes—Monograph Series, vol. 33(1) pp 114–142.

Carlborg Ö, Haley CS (2004). Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5**: 618–625.

Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 967–971.

Foolad MR, Jones RA (1992). Models to estimate maternally controlled genetic variation in quantitative seed characters. *Theor Appl Genet* **83**: 360–366.

Gale MD (1976). High α-amylase breeding and genetical aspects of the problem. *Cereal Res Commun* **4**: 231–243.

George EI, McMulloch RE (1993). Variable selection via Gibbs sampling. *J Am Stat Assoc* **91**: 883–904.

Haley CS, Knott SA (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Hospital F, Charcosset A (1997). Marker-assisted introgression of quantitative trait loci. *Genetics* **147**: 1469–1485.

Hoti F, Sillanpää MJ (2006). Bayesian mapping of genotype × expression interaction in quantitative and qualitative traits. *Heredity* **97**: 4–18.

Hu Z, Xu C (2005). A new statistical method for mapping QTLs underlying endosperm traits. *Chin Sci Bull* **50**: 1470–1476.

Jansen RC (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.

Jiang CJ, Zeng ZB (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.

Kao CH (2004). Multiple-interval mapping for quantitative trait loci controlling endosperm traits. *Genetics* **167**: 1987–2002.

Kao CH, Zeng ZB (1997). General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**: 359–371.

Kao CH, Zeng ZB, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.

Lander ES, Botstein SD (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Martinez O, Curnow RN (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* **85**: 480–488.

Mo HD (1987). Genetic expression for endosperm traits. In: Weir B, Eisen EJ, Goodmn MM, Namkoong G (eds). *Proceedings of the Second International Conference on Quantitative Genetics*. Sinauer Associates: Sunderland, MA, pp 478–487.

Moreau L, Charcosset A, Hospital F, Gallais A (1998). Marker-assisted selection efficiency in populations of finite size. *Genetics* **148**: 1353–1365.

Oh C, Ye KQ, He QM, Mendell NR (2003). Locating disease genes using Bayesian variable selection with the Haseman–Elston method. *BMC Genet* **4** (Suppl 1): S69.

Peleman JD, Voort JR (2003). Breeding by design. *Trends Plant Sci* **8**: 330–334.

Phillips PC (1998). The language of gene interaction. *Genetics* **149**: 1167–1171.

Pooni HS, Kumar I, Khush GS (1992). A comprehensive model for disomically inherited metrical traits expressed in triploid tissues. *Heredity* **69**: 166–174.

Sen S, Churchill GA (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.

Servin B, Martin OC, Mezard M, Hospital F (2004). Toward a theory of marker-assisted gene pyramiding. *Genetics* **168**: 513–523.

Tan YF, Li JX, Yu SB, Xing YZ, Xu CG, Zhang Q (1999). The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. *Theor Appl Genet* **99**: 642–648.

Wang XL, Larkins BA (2001). Genetic analysis of amino acid accumulation in opaque-2 maize endosperm. *Plant Physiol* **125**: 1766–1777.

Wang XL, Woo YM, Kim CS, Larkins BA (2001). Quantitative trait locus mapping of loci influencing elongation factor 1α content in maize endosperm. *Plant Physiol* **125**: 1271–1282.

Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ *et al*. (2005). Bayesian shrinkage estimation of QTL parameters. *Genetics* **170**: 465–480.

Wen Y, Wu WR (2006). Methods for mapping QTLs underlying endosperm traits based on random hybridization design. *Chin Sci Bull* **51**: 1976–1981.

Wu RL, Lou XY, Ma CX, Wang XL, Larkins BA, Casella G (2002a). An improved genetic model generates high-resolution mapping of QTL for protein quality in maize endosperm. *Proc Natl Acad Sci USA* **99**: 11281–11286.

Wu RL, Ma CX, Gallo-Meagher M, Littell RC, Casella G (2002b). Statistical methods for dissecting triploid endosperm traits

using molecular markers: an autogamous model. *Genetics* **162**: 875–892.
Xu C, He X, Xu S (2003). Mapping quantitative trait loci underlying triploid endosperm traits. *Heredity* **90**: 228–235.
Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.
Xu S (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**: 513–521.
Xu S, Jia Z (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics* **175**: 1955–1963.
Yi N (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**: 967–975.
Yi N, George V, Allison DB (2003a). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.
Yi N, Xu S, Allison DB (2003b). Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**: 867–883.
Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.
Zhang YM (2006). Advances on methods for mapping QTL in plant. *Chin Sci Bull* **51**: 2809–2818.
Zhang YM, Xu S (2005a). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **95**: 96–104.
Zhang YM, Xu S (2005b). Advanced statistical methods for detecting multiple quantitative trait loci. *Recent Res Dev Genet Breed* **2**: 1–23.
Zhu J, Weir BS (1994). Analysis of cytoplasmic and maternal effects. II. Genetic models for triploid endosperm. *Theor Appl Genet* **89**: 160–166.

Appendix

Assuming that an endosperm trait is controlled by two unlinked QTL, Q_1 and Q_2, the genetic variance in the population of random hybridization lines of F_2 plants is

$$\begin{aligned}
\sigma_g^2 = {} & \frac{5}{4}a_1^2 + \frac{3}{16}d_{11}^2 + \frac{3}{16}d_{12}^2 + \frac{5}{4}a_2^2 + \frac{3}{16}d_{21}^2 + \frac{3}{16}d_{22}^2 + \frac{25}{16}i_{a_1a_2}^2 \\
& + \frac{5}{16}(i_{a_1d_{21}}^2 + i_{a_1d_{22}}^2 + i_{d_{11}a_2}^2 + i_{d_{12}a_2}^2) + \frac{15}{256}(i_{d_{11}d_{21}}^2 + i_{d_{11}d_{22}}^2 + i_{d_{12}d_{21}}^2 + i_{d_{12}d_{22}}^2) \\
& + \frac{1}{4}(a_1d_{11} - a_1d_{12} + a_2d_{21} - a_2d_{22}) + \frac{5}{8}(a_1i_{a_1d_{21}} + a_1i_{a_1d_{22}} + a_2i_{d_{11}a_2} + a_2i_{d_{12}a_2}) \\
& + \frac{1}{16}a_1(i_{d_{11}d_{21}} + i_{d_{11}d_{22}} - i_{d_{12}d_{21}} - i_{d_{12}d_{22}}) + \frac{1}{16}a_2(i_{d_{11}d_{21}} - i_{d_{11}d_{22}} + i_{d_{12}d_{21}} - i_{d_{12}d_{22}}) \\
& + \frac{1}{16}(d_{11} - d_{12})(i_{a_1d_{21}} + i_{a_1d_{22}}) + \frac{1}{16}(d_{21} - d_{22})(i_{d_{11}a_2} + i_{d_{12}a_2}) \\
& + \frac{1}{16}i_{a_1d_{21}}(i_{d_{11}d_{21}} - i_{d_{12}d_{21}}) + \frac{1}{16}i_{a_1d_{22}}(i_{d_{11}d_{22}} - i_{d_{12}d_{22}}) + \frac{1}{16}i_{d_{11}a_2}(i_{d_{11}d_{21}} - i_{d_{11}d_{22}}) \\
& + \frac{1}{16}i_{d_{12}a_2}(i_{d_{12}d_{21}} - i_{d_{12}d_{22}}) + \frac{5}{16}i_{a_1a_2}(i_{a_1d_{21}} - i_{a_1d_{22}} + i_{d_{11}a_2} - i_{d_{12}a_2}) \\
& + \frac{3}{32}[d_{11}(i_{d_{11}d_{21}} + i_{d_{11}d_{22}}) + d_{12}(i_{d_{12}d_{21}} + i_{d_{12}d_{22}}) + d_{21}(i_{d_{11}d_{21}} + i_{d_{12}d_{21}}) \\
& + d_{22}(i_{d_{11}d_{22}} + i_{d_{12}d_{22}})] - \frac{1}{8}(d_{11}d_{12} + d_{21}d_{22}) - \frac{1}{32}[d_{11}(i_{d_{12}d_{21}} + i_{d_{12}d_{22}}) \\
& + d_{12}(i_{d_{11}d_{21}} + i_{d_{11}d_{22}}) + d_{21}(i_{d_{11}d_{22}} + i_{d_{12}d_{22}}) + d_{22}(i_{d_{11}d_{21}} + i_{d_{12}d_{21}})] \\
& + \frac{1}{32}i_{a_1a_2}(i_{d_{11}d_{21}} - i_{d_{11}d_{22}} - i_{d_{12}d_{21}} + i_{d_{12}d_{22}}) + \frac{1}{32}(i_{a_1d_{21}} - i_{a_1d_{22}})(i_{d_{11}a_2} - i_{d_{12}a_2}) \\
& - \frac{1}{128}(i_{d_{11}d_{21}}i_{d_{11}d_{22}} + i_{d_{11}d_{21}}i_{d_{12}d_{21}} + i_{d_{11}d_{21}}i_{d_{12}d_{22}} + i_{d_{11}d_{22}}i_{d_{12}d_{21}} + i_{d_{11}d_{22}}i_{d_{12}d_{22}} + i_{d_{12}d_{21}}i_{d_{12}d_{22}})
\end{aligned}$$