A General Quantitative Genetic Model for Haplotyping a Complex Trait in Humans

Song Wu, Jie Yang, Chenguang Wang and Rongling Wu*

Department of Statistics, University of Florida, Gainesville, FL 32611, USA

Abstract: Uncertainty about linkage phases of multiple single nucleotide polymorphisms (SNPs) in heterozygous diploids challenges the identification of specific DNA sequence variants that encode a complex trait. A statistical technique implemented with the EM algorithm has been developed to infer the effects of SNP haplotypes from genotypic data by assuming that one haplotype (called the risk haplotype) performs differently from the rest (called the non-risk haplotype). This assumption simplifies the definition and estimation of genotypic values of diplotypes for a complex trait, but will reduce the power to detect the risk haplotype when non-risk haplotypes contain substantial diversity. In this article, we incorporate general quantitative genetic theory to specify the differentiation of different haplotypes in terms of their genetic control of a complex trait. A model selection procedure is deployed to test the best number and combination of risk haplotypes, thus providing a precise and powerful test of genetic determination in association studies. Our method is derived on the maximum likelihood theory and has been shown through simulation studies to be powerful for the characterization of the genetic architecture of complex quantitative traits.

Received on: July 12, 2006 - Revised on: August 17, 2007 - Accepted on: September 2, 2007

Key Words: Complex trait, diplotype, haplotype, quantitative genetics, quantitative trait nucleotides.

INTRODUCTION

The high-throughout technology of single nucleotide polymorphisms (SNPs) provides a powerful tool for studying the detailed genetic and developmental architecture of complex traits, such as human diseases, because SNPs residing within a coding sequence can alter the biological function of a protein that forms a phenotype [1-2]. However, current experimental techniques have still not achieved a point at which multiple SNPs can be easily observed at their diplotype level [3]. Such a technological limitation makes it difficult to associate the phenotypes of a trait with specific DNA sequence variants (known as haplotypes) constructed by a set of SNPs, although recent genetic studies suggest that a gene may determine a complex trait through its haplotype rather than genotype [4-8]. More recently, a statistical model has been derived to estimate and test haplotype effects on trait variation with a random sample drawn from a natural population [9-11]. This model implements the population genetic properties of gene segregation into a unifying mixture-model framework for haplotype discovery. It assumes that one haplotype composed of alleles at multiple SNPs is different from the remaining haplotypes in terms of genetic effects on a trait. The former is called the reference or risk haplotype [9], whereas the latter is collectively called the non-reference or non-risk haplotype. This simplified assumption allows the direct use of a traditional biallelic quantitative genetic model [12] and facilitates the definition and estimation of genetic effects triggered by different haplotypes, but it is limited in practical use when there is substantial variation among the non-risk haplotypes.

The motivation of this work is to expand Liu *et al.*'s [9] original idea to model all possible effects of individual haplotypes by constructing a multi-allelic quantitative genetic model within the mixture model framework. The multi-allelic model deals with genetic effects triggered by multiple alleles at a single gene and is thought to be important for explaining genetic variation in a natural population. We use the multi-allelic model to define various additive and dominance effects due to multiple risk haplotypes. Conventional model selection criteria are incorporated to choose the optimal number and combination of risk haplotypes responsible for quantitative variation of a trait. We derived closed forms for the EM algorithm to estimate a variety of genetic parameters including haplotype frequencies and haplotype effects. Simulation studies are used to test the statistical behavior of the model and validate its usefulness and utilization.

METHOD

Population and Quantitative Genetic Models

Suppose there are genetically associated SNPs each with two alleles designated as 1 and 0. Let p and q be the 1-allele frequencies for the first and second SNP, respectively. Thus, the 0-allele frequencies at different SNPs will be $1\text{-}p$ and $1\text{-}q$. These two SNPs segregating in a natural population form four haplotypes, 11, 10, 01 and 00, whose frequencies are constructed by allele frequencies and linkage disequilibrium (D) between the two SNPs, i.e., $p_{11} = pq + D$, $p_{10} = \mathrm{p}(1 - q) - D$, $p_{01} = (1 - p)q - D$, and $p_{00} = (1 - p)(1 - q) - D$. We use $\Theta_p = (p_{11}, p_{10}, p_{01}, p_{00})$ to denote the haplotype frequency

[*] Address correspondence to this author at the Department of Statistics, University of Florida, Gainesville, FL 32611, USA; Tel: (352)392-3806; Fax: (352)392-8555; E-mail: rwu@stat.ufl.edu

1389-2029/07 $50.00+.00

vector. These haplotypes unite randomly to generate 10 distinct diplotypes and 9 distinct genotypes. If the population is at Hardy-Weinberg equilibrium, the frequency of a diplotype is expressed as the product of the frequencies of the two haplotypes that constitute the diplotype (Table **1**). The frequency of the double zygotic genotype is the summation of the frequencies of its two possible diplotypes.

We are interested in the detection of risk haplotype(s) constructed by the alleles of the two SNPs which encodes a quantitative trait. Below given are different genetic models used to identify risk haplotypes.

Biallelic Model

Liu *et al*. [9] assumed that all haplotypes are sorted into two groups, risk and non-risk, and defined the combination of risk and non-risk haplotypes as a composite diplotype. Let A_1 and A_0 be the risk and non-risk haplotypes, respectively, which are equivalent to two alternative alleles if the two associated SNPs considered are viewed as a "locus". Thus, for such a "biallelic locus", we have three possible composite diplotypes whose genotypic values are specified as

Composite Diplotype	Genotypic Value
$A_1 A_1$	$\mu_1 = \mu + a$
$A_1 A_0$	$\mu_2 = \mu + d$
$A_0 A_0$	$\mu_3 = \mu - a$

(1)

where μ is the overall mean, a is the additive effect due to the substitution of the risk haplotype by the non-risk haplotype, and d is the dominance effect due to the interaction between the risk and non-risk haplotypes. These parameters are arrayed in $\Theta_{qB} = (\mu, a, d)$.

There are a total of seven options to choose the risk haplotype. First, because any one haplotype from 11, 10, 01 and 00 can be risk, there are four choices for determining the risk haplotype. Second, any two haplotypes can be different from the rest, which includes three possibilities for combining the risk vs. non-risk haplotypes. All these options can be tabulated as follows:

No.	Risk Haplotype	Non-risk Haplotype
B_1	11	10,01,00
B_2	10	11,01,00
B_3	01	11,10,00
B_4	00	11,10,01
B_5	11,10	01,00
B_6	11,01	10,00
B_7	11,00	10,01

(2)

The optimal choice of a risk haplotype for the biallelic model is based on the maximum of the likelihoods calculated for each of the seven options described above.

Triallelic Model

It is possible that there are two distinct risk haplotypes which are each different from non-risk haplotypes. This case

Table 1. Diplotypes and their Frequencies for each of Nine Genotypes at Two SNPs, Haplotype Composition Frequencies for Each Genotype, and Composite Diplotypes under Biallelic, Triallelic and Quadriallelic Models

Genotype	Diplotype		Relative Diplotype Frequency	Composite Diplotype		
	Configuration	Frequency		Biallelic	Triallelic	Quadriallelic
11/11	[11][11]	p_{11}^2	1	A_1A_1	A_1A_1	A_1A_1
11/10	[11][10]	$2p_{11}p_{10}$	1	A_1A_0	A_1A_2	A_1A_2
11/00	[10][10]	p_{10}^2	1	A_0A_0	A_2A_2	A_2A_2
10/11	[11][01]	$2p_{11}p_{01}$	1	A_1A_0	A_1A_0	A_1A_3
10/10	[11][00] [10][01]	$2p_{11}p_{00}$ $2p_{10}p_{01}$	ϕ $1-\phi$	A_1A_0 A_0A_0	A_1A_0 A_2A_0	A_1A_0 A_2A_3
10/00	[10][00]	$2p_{10}p_{00}$	1	A_0A_0	A_2A_0	A_2A_0
00/11	[01][01]	p_{01}^2	1	A_0A_0	A_0A_0	A_3A_3
00/10	[01][00]	$2p_{01}p_{00}$	1	A_0A_0	A_0A_0	A_3A_0
00/00	[00][00]	p_{00}^2	1	A_0A_0	A_0A_0	A_0A_0

Two alleles for each of the two SNPs are denoted as 1 and 0, respectively. Genotypes at different SNPs are separated by a slash. Diplotypes are the combination of two bracketed maternally and paternally derived haplotypes. Risk haplotype(s) is assumed as [11] for the biallelic model, [11] and [10] for the triallelic model, and [11], [10] and [01] for the quadriallelic model.

is regarded as a "triallelic locus". Let A_1 and A_2 be the first and second risk haplotypes, and A_0 be the non-risk haplotype, which form six composite diplotypes with genotypic values expressed as

Composite Diplotype	Genotypic Value
$A_1 A_1$	$\mu_1 = \mu + a_1$
$A_2 A_2$	$\mu_2 = \mu + a_2$
$A_0 A_0$	$\mu_3 = \mu - a_1 - a_2$
$A_1 A_2$	$\mu_4 = \mu + \frac{1}{2}(a_1 + a_2) + d_{12}$
$A_1 A_0$	$\mu_5 = \mu - \frac{1}{2} a_1 + d_{10}$
$A_2 A_0$	$\mu_6 = \mu - \frac{1}{2} a_1 + d_{20}$

(3)

where μ is the overall mean, a_1 and a_2 are the additive effects due to the substitution of the first and second risk haplotype by the non-risk haplotype, and d_{12}, d_{10} and d_{20} are the dominance effects due to the interaction between the first and second risk haplotype, between the first risk haplotype and the non-risk haplotype and between the second risk haplotype and non-risk haplotype, respectively. These parameters are arrayed in $\Theta_{qT} = (\mu, a_2, a_2, d_{12}, d_{10}, d_{20})$.

The triallelic model may include a total of six haplotype combinations, which are

No.	Risk Haplotype		Non-risk Haplotype
	1	2	
T_1	11	10	01,00
T_2	11	01	10,00
T_3	11	00	10,01
T_4	10	01	11,00
T_5	10	00	11,01
T_6	01	00	11,10

(4)

The optimal combination of risk haplotypes for the triallelic model corresponds to the maximum of the likelihoods calculated for each of the six possibilities.

Quadriallelic Model

If there are three distinct risk haplotypes, we need a quadriallelic genetic model to specify haplotype effects. Let A_1, A_2 and A_3 be the first, second and third risk haplotypes, and A_0 be the non-risk haplotype, which form 10 composite diplotypes with genotypic values expressed as

Composite Diplotype	Genotypic Value
A_1A_1	$\mu_1 = \mu + a_1$
A_2A_2	$\mu_2 = \mu + a_2$
A_3A_3	$\mu_3 = \mu + a_3$
A_0A_0	$\mu_4 = \mu - (a_1 + a_2 + a_3)$
A_1A_2	$\mu_5 = \mu + \frac{1}{2}(a_1 + a_2) + d_{12}$
A_1A_3	$\mu_6 = \mu + \frac{1}{2}(a_1 + a_2) + d_{13}$
A_2A_3	$\mu_7 = \mu + \frac{1}{2}(a_2 + a_3) + d_{23}$
A_1A_0	$\mu_8 = \mu - \frac{1}{2}(a_2 + a_3) + d_{10}$
A_2A_0	$\mu_9 = \mu - \frac{1}{2}(a_1 + a_3) + d_{20}$
A_3A_0	$\mu_{10} = \mu - \frac{1}{2}(a_1 + a_2) + d_{30}$

(5)

where μ is the overall mean, a_1, a_2 and a_3 are the additive effects due to the substitution of the first, second and third risk haplotype by the non-risk haplotype, and d_{12}, d_{13}, d_{23}, d_{10}, d_{20}, and d_{30} are the dominance effects due to the interaction between the first and second risk haplotype, between the first and third risk haplotype, between the second and third risk haplotype, between the first risk and non-risk haplotype, between the second risk and non-risk haplotype, and between the third risk and non-risk haplotype, respectively. These parameters are arrayed in $\Theta_{qQ} = (\mu, a_1, a_2, a_3\ d_{12}, d_{13}, d_{23}, d_{10}, d_{20}, d_{30})$.

LIKELIHOOD

Assume that a total of n subjects are sampled from a Hardy-Weinberg equilibrium population and that each subject is genotyped for many SNPs and phenotyped for a quantitative trait. Consider two of the SNPs that form nine genotypes with observed numbers generally expressed as $n_{r_1 r'_1/r_2 r'_2}$ (r_1, r'_1, r_2, r'_2 = 1,0). The phenotypic value of the trait for subject i is expressed in terms of the two-SNP haplotypes as

$$y_i = \sum_{j=1}^{J} \xi_i \mu_J + e_i, \tag{6}$$

where ξ_i is the indicator variable defined as 1 if subject i has a composite diplotype j and 0 otherwise, e_i is the residual error, normally distributed as $N(0, \sigma^2)$, and J is the number of composite diplotypes expressed as

$$J = \begin{cases} 3 & \text{for the biallelic model} \\ 6 & \text{for the triallelic model} \\ 10 & \text{for the quadriallelic model.} \end{cases} \tag{7}$$

The genotypic values of composite diplotypes and variance are arrayed by a quantitative genetic parameter vector $\Theta_q = (\Theta_{qB}, \sigma^2)$ for the biallelic model, (Θ_{qT}, σ^2) for the triallelic model, and (Θ_{qQ}, σ^2) for the quadriallelic model.

The log-likelihood of haplotype frequencies, genotypic values of the diplotypes and residual variance given the phenotypic (**y**) and SNP data (**S**) is factorized into two parts, expressed as

$$\log L(\Theta_p, \Theta_q \mid \mathbf{y}, \mathbf{S}) = \log L(\Theta_p \mid \mathbf{S}) + \log L(\Theta_q \mid \mathbf{y}, \mathbf{S}, \Theta_p) \qquad (8)$$

where

$$\log L(\Theta_p \mid \mathbf{S}) = \text{constant}$$

$$+2n_{11/11} \ln p_{11} + n_{11/10} \ln(2\, p_{11}p_{10}) + 2n_{11/00} \ln p_{10}$$

$$+n_{10/11} \ln(2\, p_{11}p_{01}) + n_{10/10} \ln(2\, p_{11}p_{00} + 2\, p_{10}p_{01}) + n_{10/00} \ln(2\, p_{10}p_{00}) + 2n_{00/11} \ln p_{01} + n_{00/10} \ln(2\, p_{01}p_{00}) + 2n_{00/00} \ln p_{00}, \quad (9)$$

$$\log L(\Theta_{qB} \mid \mathbf{y}, \mathbf{S}, \Theta_p)$$

$$= \sum_{i=1}^{n_{11/11}} \log f_1(y_i) + \sum_{i=1}^{n_{11/10}} \log f_2(y_i) + \sum_{i=1}^{n_{11/00}} \log f_3(y_i)$$

$$+ \sum_{i=1}^{n_{10/11}} \log f_2(y_i) + \sum_{i=1}^{n_{10/10}} \log[\phi f_2(y_i) + (1-\phi) f_3(y_i)] +$$

$$\sum_{i=1}^{n_{10/00}} \log f_3(y_i)$$

$$+ \sum_{i=1}^{n_{00/11}} \log f_3(y_i) + \sum_{i=1}^{n_{00/10}} \log f_3(y_i) + \sum_{i=1}^{n_{00/00}} \log f_3(y_i) \quad (10)$$

for the biallelic model assuming that haplotype 11 is a risk haplotype,

$$\log L(\Theta_{qT} \mid \mathbf{y}, \mathbf{S}, \Theta_p)$$

$$= \sum_{i=1}^{n_{11/11}} \log f_1(y_i) + \sum_{i=1}^{n_{11/10}} \log f_4(y_i) + \sum_{i=1}^{n_{11/00}} \log f_2(y_i)$$

$$+ \sum_{i=1}^{n_{10/11}} \log f_5(y_i) + \sum_{i=1}^{n_{10/10}} \log[\phi f_5(y_i) + (1-\phi) f_6(y_i)] +$$

$$\sum_{i=1}^{n_{10/00}} \log f_6(y_i)$$

$$+ \sum_{i=1}^{n_{00/11}} \log f_3(y_i) + \sum_{i=1}^{n_{00/10}} \log f_3(y_i) + \sum_{i=1}^{n_{00/00}} \log f_3(y_i) \quad (11)$$

for the triallelic model assuming that haplotypes 11 and 10 are the first and second risk haplotypes, respectively,

$$\log L(\Theta_{qQ} \mid \mathbf{y}, \mathbf{S}, \Theta_p)$$

$$= \sum_{i=1}^{n_{11/11}} \log f_1(y_i) + \sum_{i=1}^{n_{11/10}} \log f_5(y_i) + \sum_{i=1}^{n_{11/00}} \log f_2(y_i)$$

$$+ \sum_{i=1}^{n_{10/11}} \log f_6(y_i) + \sum_{i=1}^{n_{10/10}} \log[\phi f_8(y_i) + (1-\phi) f_7(y_i)] +$$

$$\sum_{i=1}^{n_{10/00}} \log f_9(y_i)$$

$$+ \sum_{i=1}^{n_{00/11}} \log f_3(y_i) + \sum_{i=1}^{n_{00/10}} \log f_{10}(y_i) + \sum_{i=1}^{n_{00/00}} \log f_4(y_i) \quad (12)$$

for the quadriallelic model assuming that haplotypes 11, 10 and 01 are the first, second and third risk haplotypes, respectively, with f_i (y_i) being a normal distribution density function of composite diplotype *j* with mean μ_j and variance σ^2.

We have shown that maximizing $L(\Theta_p, \Theta_q \mid \mathbf{y}, \mathbf{S})$ in equation (8) is equivalent to individually maximizing $\log L(\Theta_p \mid \mathbf{S})$ in equation (9) and $\log L(\Theta_q \mid \mathbf{y}, \mathbf{S}, \Theta_p)$ in equation (10), (11) or (12) (unpublished results).

THE EM ALGORITHM

A closed-form solution for the EM algorithm has been derived to estimate the unknown parameters that maximize the likelihoods. The estimates of haplotype frequencies are based on the log-likelihood function $L(\Theta_p \mid \mathbf{S})$, whereas the estimates of genotypic values of composite diplotypes and the residual variance are based on the log-likelihood function $L(\Theta_q \mid \mathbf{y}, \mathbf{S}, \Theta_p)$. These two different types of parameters can be estimated using a two-stage hierarchical EM algorithm (see [9] for a detailed implementation).

MODEL SELECTION

The formulation of likelihoods (10), (11) and (12) is based on the assumption that one or more haplotypes are risk haplotypes for the biallelic, triallelic and quadriallelic model. However, a real risk haplotype under each of these models is unknown from raw data (**y**, **S**). Also, we are uncertain about the optimal number of risk haplotypes. An additional step for the choice of the most likely risk haplotypes and their number should be implemented. The simplest way to do so is to calculate and compare the likelihood values within the model by assuming that any one or more of the four haplotypes can be a risk haplotype, and AIC or BIC among the models by assuming different numbers of risk haplotypes [13]. Thus, we obtain possible likelihood values and AIC/BIC as follows:

Model	No.	Likelihood	AIC/BIC
Biallelic	B_l	$\log L_{B_l}(\widehat{\Theta}_p, \widehat{\Theta}_{qB} \mid \mathbf{y}, \mathbf{S})$	C_{B_l}
Triallelic	T_l	$\log L_{T_l}(\widehat{\Theta}_p, \widehat{\Theta}_{qT} \mid \mathbf{y}, \mathbf{S})$	C_{T_l}
Quadriallelic	Q	$\log L_Q(\widehat{\Theta}_p, \widehat{\Theta}_{qQ} \mid \mathbf{y}, \mathbf{S})$	C_Q

(13)

The largest likelihood and the smallest AIC/BIC value calculated is thought to correspond to the most likely risk haplotypes and their optimal number.

HYPOTHESIS TESTS

The genetic architecture of a quantitative trait is characterized by quantitative genetic parameters (including haplotype effects and the mode of their inheritance). The model proposed provides a meaningful way for estimating the genetic architecture of a trait. The estimated genotypic values for the composite diplotypes can be used to estimate additive and dominance genetic effects of haplotypes by

	Additive	Dominace
Biallelic	$a = (\mu_1 - \mu_3)/2$	$d = \mu_2 - (\mu_1 + \mu_3)/2$
Triallelic	$a_1 = [2\mu_2 - (\mu_1 + \mu_3)]/3$	$d_{12} = \mu_4 - (\mu_1 + \mu_2)/2$
	$a_2 = [2\mu_1 - (\mu_2 + \mu_3)]/3$	$d_{10} = \mu_5 - (\mu_1 + \mu_3)/2$
		$d_{20} = \mu_6 - (\mu_2 + \mu_3)/2$
Quadri-allelic	$a_1 = [3\mu_1 - (\mu_2 + \mu_3 + \mu_4)]/4$	$d_{12} = \mu_5 - (\mu_1 + \mu_2)/2$
	$a_2 = [3\mu_2 - (\mu_1 + \mu_3 + \mu_4)]/4$	$d_{13} = \mu_6 - (\mu_1 + \mu_3)/2$
	$a_3 = [3\mu_3 - (\mu_1 + \mu_2 + \mu_4)]/4$	$d_{23} = \mu_7 - (\mu_2 + \mu_3)/2$
		$d_{10} = \mu_8 - (\mu_1 + \mu_4)/2$
		$d_{20} = \mu_9 - (\mu_2 + \mu_4)/2$
		$d_{30} = \mu_{10} - (\mu_3 + \mu_4)/2$

(14)

The additive and dominance effects under different models can be tested by formulating the null hypothesis that the effect being tested is equal. The estimates of the parameters under the null hypotheses can be obtained with the same EM algorithm derived for the alternative hypotheses but with a constraint of the tested effect equal to zero. The log-likelihood ratio test statistics for each hypothesis is thought to asymptotically follow a χ^2-distributed with the degree of freedom equal to the difference of the numbers of the parameters being tested under the null and alternative hypotheses.

HAPLOTYPING WITH THREE SNPS

Li *et al*. [11] constructed a conceptual framework and statistical algorithm for haplotyping a quantitative trait with three SNPs. For a set of three SNPs, there are eight different haplotypes, among which it is possible to have one to seven risk haplotypes. The biallelic model specifies one risk haplotype which may be composed of one (8 cases), two (24 cases), three (56) or four haplotypes (170). The triallelic, quadrialleli, pentaallelic, hexaallelic, septemallelic and octoallelic models contains 28, 56, 170, 56, 24 and 8 cases, respectively. It can be seen that the model selection procedure to determine the optimal number and combination of risk haplotypes will become exponentially more complicated when the number of SNPs increases.

MONTE CARLO SIMULATION

The statistical properties of the model are investigated through simulation studies. Given a certain sample size of subjects (n = 100, 400 or 1000), two SNPs (each with two alleles 1 and 0) were simulated by assuming that 10 diplotypes follow a multinomial distribution with the frequencies determined by allele frequencies $p = 0.6$ and $q = 0.6$ and linkage disequilibrium $D = 0.05$. By hypothesizing risk haplotypes under biallelic, triallelic and quadriallelic models, composite diplotypes can be determined for each double-SNP genotype. The phenotypic values of a quantitative trait were simulated as a normal distribution with mean depending on composite diplotypes and variance determined under different heritability levels (H^2 = 0.1, 0.2 and 0.4).

For a practical data set, the number and combination of risk haplotypes that govern a phenotypic trait is unknown. Thus, the simulation performed here will elucidate the procedure and power to determine risk haplotypes by the new model. The data sets simulated with given risk haplotypes under each quantitative genetic model were analyzed by biallelic, triallelic and quadriallelic models, respectively. For each analysis, the likelihood and model selection criteria, AIC and BIC, are calculated with display (13). The power to correctly identify risk haplotypes was calculated from 1000 simulation replicates. Fig. (**1**) illustrates such power under different heritabilities, sample sizes and genetic models. For the data simulated under the biallelic model, a correct risk haplotype can well be determined with a sample size of 200 even when the heritability of the trait is modest (0.1). In this case in which a small number of genetic parameters are included, the BIC performs better than the AIC. For a data set simulated under the triallelic model, the power of haplotype detection reduces considerably, compared with the data set simulated by the biallelic model. If the heritability of a trait is as low as 0.1, about 1000 subjects are needed to achieve the power of 0.8. With the heritability increasing to 0.2 or 0.4, the same power needs about 600 or 300 subjects, respectively. It is interesting to note that the AIC performs better than the BIC when the heritability is low (0.1 or 0.2), whereas the two criteria perform similarly when the heritability is high (0.4).

The data set simulated under the quadriallelic model contains a very large number of genetic parameters to be estimated. As expected, the power of haplotype detection in this case will be reduced (Fig. **1**). When the heritability is as low as 0.1, a sample size of 1000 can only achieve a power of 0.2. But with an increasing heritability, the power will increase dramatically. For example, a power of $>$ 0.9 can be achieved with 600 subjects when the heritability is 0.4. For the quadriallelic model-simulated data, the AIC always performs better than the BIC because the latter poses too heavy penalty in this case.

The estimates of population (including allele frequencies and linkage disequilibrium) and quantitative genetic parameters (including additive and dominance effects) for each simulated data set were evaluated by calculating their sampling errors. Previous work suggested that the estimates of

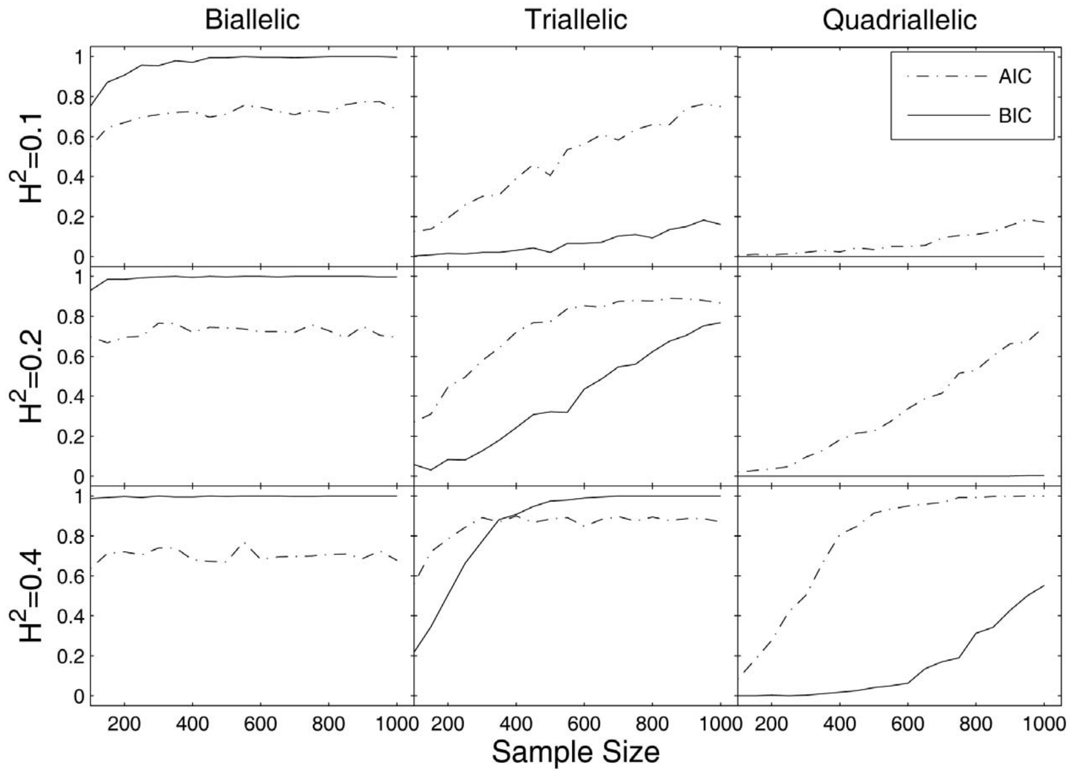

Fig. (1). Power to detect correct risk haplotypes from the data simulated by a biallelic, triallelic and quadriallelic model, respectively, under different heritabilities and sample sizes. Model selection criteria are based on AIC and BIC.

population genetic parameters display great precision even for a sample size of 100 [9]. Here, our simulation studies will focus on the assessment of the precision of quantitative genetic parameter estimates under different heritabilities and sample sizes. For the data set simulated with the biallelic model, the additive and dominance effects can be precisely estimated even with a heritability of 0.1 and a sample size of 100 (Table **2**). Increasing heritabilities and sample sizes increase estimation precision dramatically.

The data set simulated under the triallelic model contains two additive effects and three dominance effects. A sample size of 100 is adequate for precise estimates of the additive effects even for a low heritability (0.1), but the reasonable estimates of the dominance effects need increasing sample size (400 or more) if the heritability is 0.1 (Table **3**). For a high heritability (0.4), a small sample size (100) can provide relatively precise estimates of the dominance effects. For the data set simulated with the quadriallelic model, three additive effects and six dominance effects are included. Still, a low sample size (100) can provide very good estimates of the additive effects even for a low heritability. To reasonably estimate the dominance effects, we need a large sample size (1000) for the heritability of 0.1 or a moderately large sample size (400) for the heritability of 0.4 (Table **4**).

DISCUSSION

Single nucleotide polymorphisms (SNPs) are powerful markers that can explain interindividual differences in disease risk and drug responsiveness in humans. For genes containing multiple SNPs, haplotype structure (i.e., the linear arrangement of different SNP alleles on each of the two homologous chromosomes) is thought to be the principal de-

Table 2. The MLEs of the Additive and Dominance Effects Triggered by a Risk Haplotype and the Square Roots of the Mean Square Errors of the Estimates (in Parentheses) by a Biallelic Model Under Different Heritabilities and Sample Sizes

Genetic Parameter	True Value	H^2 = 0.1			H^2 = 0.4		
		n = 100	*n* = 400	*n* = 1000	*n* = 100	*n* = 400	*n* = 1000
a	10	10.04(0.175)	9.86(0.091)	10.04(0.055)	10.05(0.07)	10.05(0.036)	9.94(0.022)
d	3	2.63(0.244)	3.06(0.123)	3.11(0.08)	2.96(0.102)	2.95(0.051)	3.02(0.031)
σ	22.42	21.9(0.084)	22.27(0.039)	22.39(0.026)			
	9.15				9.02(0.034)	9.08(0.017)	9.13(0.011)

Table 3. The MLEs of the Additive and Dominance Effects Triggered by Two Risk Haplotypes and the Square Roots of the Mean Square Errors of the Estimates (in Parentheses) by a Triallelic Model Under Different Heritabilities and Sample Sizes

Genetic Parameter	True Value	H^2 = 0.1			H^2 = 0.4		
		n = 100	n = 400	n = 1000	n = 100	n = 400	n = 1000
a_1	4.0	4.15(0.188)	4.15(0.086)	3.89(0.059)	4.04(0.076)	3.95(0.039)	4.05(0.023)
a_2	-1.0	-1.24(0.192)	-1.11(0.092)	-0.84(0.057)	-0.99(0.078)	-0.95(0.039)	-1.03(0.024)
d_{12}	-7.5	-6.91(0.582)	-7.03(0.286)	-7.52(0.169)	-7.05(0.239)	-7.67(0.114)	-7.57(0.072)
d_{10}	-10.5	-11.26(0.409)	-10.22(0.176)	-10.55(0.121)	-10.31(0.146)	-10.47(0.075)	-10.51(0.044)
d_{20}	-14.0	-14.25(0.288)	-13.72(0.144)	-14.1(0.091)	-13.87(0.121)	-14.06(0.058)	-14.03(0.036)
σ	19.11	18.43(0.07)	18.98(0.034)	19.04(0.021)			
	7.80				7.54(0.031)	7.73(0.014)	7.77(0.01)

Table 4. The MLEs of the Additive and Dominance Effects Triggered by Three Risk Haplotypes and the Square Roots of the Mean Square Errors of the Estimates (in Parentheses) by a Quadriallelic Model Under Different Heritabilities and Sample Sizes

Genetic Parameter	True Value	H^2 = 0.1			H^2 = 0.4		
		n = 100	n = 400	n = 1000	n = 100	n = 400	n = 1000
a_1	-19.75	-20.91(0.815)	-19.8(0.405)	-19.88(0.266)	-20(0.346)	-19.73(0.164)	-19.76(0.105)
a_2	-5.75	-2.83(0.784)	-4.48(0.393)	-5.29(0.275)	-6.35(0.344)	-6.18(0.178)	-5.93(0.117)
a_3	-38.25	-37.71(0.819)	-37.97(0.354)	-38.38(0.228)	-37.94(0.307)	-38.31(0.142)	-38.21(0.093)
d_{12}	30.00	32.99(0.47)	30.82(0.262)	29.99(0.193)	29.97(0.197)	29.97(0.118)	29.78(0.078)
d_{13}	18.00	11.9(0.801)	15.84(0.525)	17.03(0.403)	18.55(0.358)	18.39(0.232)	18.56(0.159)
d_{23}	23.00	23.38(0.617)	22.73(0.276)	23.09(0.174)	23.15(0.236)	23.09(0.116)	22.96(0.072)
d_{10}	20.00	19.98(0.98)	19.9(0.433)	20.43(0.261)	19.17(0.387)	20.04(0.17)	20.11(0.113)
d_{20}	16.00	15.91(0.625)	15.84(0.268)	15.98(0.17)	16.04(0.249)	16.03(0.116)	15.96(0.076)
d_{30}	10.00	9.73(0.831)	9.93(0.397)	9.94(0.238)	10.06(0.363)	10.09(0.167)	9.98(0.098)
σ	31.50	29.48(0.12)	30.84(0.057)	31.19(0.039)			
	12.86				12.05(0.052)	12.65(0.026)	12.79(0.016)

terminant of phenotypic traits. While traditional analyses associate phenotypic variability with genotypes, growing evidence shows the important contribution of haplotype diversity to quantitative traits [4-8]. More recently, Liu *et al.* [9] proposed a statistical method for detecting functional (or risk) haplotypes for quantitative traits with a random sample drawn from a natural population. The method allows the characterization of DNA sequence variants that encode the phenotypic value of a trait, thus open a gateway for precisely studying the genetic architecture of quantitativevariation. In this article, we extends Liu *et al.*'s model to estimate the number and combination of multiple functional haplotypes in terms of their genetic effects.

Similar to Liu *et al.*'s work [9], our model was founded on the mixture model-based framework in which the frequencies of haplotype distribution and haplotype effects are estimated with the closed form of the EM algorithm. But our model was incorporated by two important theories from different fields, one regarding the segregation and inheritance of multiple alleles at a single locus in quantitative genetics and the second regarding model selection procedures in statistics. Liu *et al.*'s [9] model was framed on a biallelic model in which one haplotype constructed by a set of associated SNPs was assumed to perform differently from the rest of the haplotypes. Traditional quantitative genetic theory mostly based on biallelic inheritance provide a basis for estimating the additive and dominance effects due to two alternative alleles at a functional gene, but fails to characterize the genetic effects due to all possible combinations between multiple alleles. We have for the first time implemented multiallelic quantitative genetic theory into the estimation process of haplotype effects, in which multiple additive effects and multiple dominance effects due to multiple functional haplotypes can be estimated and tested separately or jointly. The new model expands the idea of haplotyping a complex trait to study the detailed genetic control of the trait in a precise way.

To deal with multiple risk haplotypes, an issue arises naturally about the selection of most likely risk haplotypes from a pool of haplotypes. This will include the optimal number of risk haplotypes and their combination that provide a best fit to the given data. We implemented model selection procedures into the test process of haplotype diversity and effects with two commonly used criteria, AIC and BIC. Extensive simulation studies were performed to investigate the statistical properties of the model and its utilization. Given a real data set, we do not know about the type and number of risk haplotypes. But these can be estimated with model selection by assuming different types of genetic models, biallelic (one risk haplotype), triallelic (two risk haplotypes) and quadriallelic (three risk haplotypes). Simulation studies with two-SNP haplotypes provide a table of model selection approaches (Tables **2–4**) to detect most likely risk haplotypes hidden in a genetic association data set based on a range of sample size and heritability as well as the types of genetic models.

The human genome contains millions of SNPs distributed over 23 pairs of chromosomes [14]. However, these SNPs were observed to locate in different haplotype blocks of the human genome [15-16]. For a given block, there are a particular number of representative SNPs or htSNPs that uniquely identify the common haplotypes in this block or QTN. Several algorithms have been developed to identify a minimal subset of htSNPs that can characterize the most common haplotypes [2, 17-18]. The idea given in this article can be used to find risk haplotypes of these htSNPs by modeling an arbitrary number of SNPs [11], and extended to detect haplotype-haplotype interactions [10], haplotype-environment interactions, parent-of-origin effects of haplotypes in genetic association studies and haplotypes regulating pharmacodynamic reactions of drugs [19]. Although these works will be computationally expensive, it should not be computationally prohibitive if combinatorial mathematics, graphical models, and machine learning are incorporated into closed forms of parameter estimation. With detailed extensions that take account into more realistic biological and genetic problems, our model may provide an efficient solution to the growing need for haplotype data collection and association studies.

ACKNOWLEDGEMENT

The preparation of this manuscript is supported by NSF grant (0540745) to RW.

REFERENCES

[1] Flint, J., Valdar, W., Shifman, S. and Mott, R. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat. Rev. Genet.*, **2005**, *6*: 271-286.

[2] Eyheramendy, S., Marchini, J., McVean, G., Myers, S. and Donnelly, P. A model-based approach to capture genetic variation for future association studies. *Genome Res.*, **2007**, *17*: 88-95.

[3] Konfortov, B. A., Bankier, A. T. and Dear, P. H. An efficient method for multi-locus molecular haplotyping. *Nucleic Acids Res.*, **2007**, *35*: e6.

[4] Judson, R., Stephens, J. C. and Windemuth, A. The predictive power of haplotypes in clinical response. *Pharmacogenomics*, **2000**, *1*: 15-26.

[5] Bader, J. S. The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics,* **2001**, *2*: 11-24.

[6] Winkelmann, B. R., Hoffmann, M. M., Nauck, M., Kumar, A. M., Nandabalan, K., Judson, R. S., Boehm, B. O., Tall, A. R., Ruano, G. and Marz, W. Haplotypes of the cholesteryl ester transfer protein gene predict lipid-modifying response to statin therapy. *Pharmacogenomics J.,* **2003**, *3*: 284-296.

[7] Clark, A. G. The role of haplotypes in candidate gene studies. *Genet. Epidemiol.*, **2004**, *27*: 321-333.

[8] Jin, G. F., Miao, R. F., Deng, Y. M., Hu, Z. B., Zhou, Y., Tan, Y. F., Wang, J. M., Hua, Z. L., Ding, W. L., Wang, L. N., Chen, W. S., Shen, J., Wang, X. R., Xu, Y. C. and Shen, H. B. Variant genotypes and haplotypes of the epidermal growth factor gene promoter are associated with a decreased risk of gastric cancer in a high-risk Chinese population. *Cancer Sci.*, **2007**, *98*: 864-868.

[9] Liu, T., Johnson, J. A., Casella, G. and Wu, R. L. Sequencing complex diseases with HapMap. *Genetics,* **2004**, *168*: 503-511.

[10] Lin, M. and Wu, R. L. Detecting sequence-sequence interactions for complex diseases. *Curr. Genom.,* **2006**, *7*: 59-72.

[11] Li, H. Y., Kim, B. R. and Wu, R. L. Identification of quantitative trait nucleotides that regulate cancer growth: A simulation approach. *J. Theor. Biol.*, **2006**, *242*: 426-439.

[12] Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*. **1998** Sinauer Associates, Sunderland, MA.

[13] Burnham, K. P. and Andersson, D. R. Model Selection and Inference. A Practical Information-Theoretic Approach. **1998** Springer, New York.

[14] Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S. and P. Donnelly, International HapMap Consortium, A haplotype map of the human genome. *Nature,* **2005**, *437*: 1299-1320.

[15] Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, P. D., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A. and Cox, D.R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science,* **2001**, *294*: 1719-1723.

[16] Terwilliger, J. D. and Hiekkalinna, T. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur. J. Hum. Genet.*, **2006**, *14*: 426-437.

[17] Zhang, K., Deng, M., Chen, T., Waterman, M. S. and Sun, F. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci.,* **2002**, *99*: 7335-7339.

[18] Sebastiani, P., Lazarus, S. W., Kunkel, L. M., Kohane, I. S. and Ramoni, M. Minimal haplotype tagging. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*: 9900-9905.

[19] Lin, M., Aquilante, C., Johnson, J. A. and Wu, R. L. Sequencing drug response with HapMap. *Pharmacogenomics J.,* **2005**, *5*: 149-156.