Haplotyping a Quantitative Trait with a High-Density Map in Experimental Crosses

Wei Hou[1], **John Stephen F. Yap**[2], **Song Wu**[2], **Tian Liu**[2], **James M. Cheverud**[3], **Rongling Wu**[2]*

1 Department of Epidemiology and Health Policy Research, University of Florida, Gainesville, Florida, United States of America, **2** Department of Statistics, University of Florida, Gainesville, Florida, United States of America, **3** Department of Anatomy and Neurobiology, Washington University Medical School, St. Louis, Missouri, United States of America

Background. **The ultimate goal of genetic mapping of quantitative trait loci (QTL) is the positional cloning of genes involved in any agriculturally or medically important phenotype. However, only a small portion ($\leq$ 1%) of the QTL detected have been characterized at the molecular level, despite the report of hundreds of thousands of QTL for different traits and populations.** ***Methods/Results.*** **We develop a statistical model for detecting and characterizing the nucleotide structure and organization of haplotypes that underlie QTL responsible for a quantitative trait in an F_2 pedigree. The discovery of such haplotypes by the new model will facilitate the molecular cloning of a QTL. Our model is founded on population genetic properties of genes that are segregating in a pedigree, constructed with the mixture-based maximum likelihood context and implemented with the EM algorithm. The closed forms have been derived to estimate the linkage and linkage disequilibria among different molecular markers, such as single nucleotide polymorphisms, and quantitative genetic effects of haplotypes constructed by non-alleles of these markers. Results from the analysis of a real example in mouse have validated the usefulness and utilization of the model proposed.** ***Conclusion.*** **The model is flexible to be extended to model a complex network of genetic regulation that includes the interactions between different haplotypes and between haplotypes and environments.**

Citation: Hou W, Yap JSF, Wu S, Liu T, Cheverud JM, et al (2007) Haplotyping a Quantitative Trait with a High-Density Map in Experimental Crosses. PLoS ONE 2(8): e732. doi:10.1371/journal.pone.0000732

INTRODUCTION

The basic principle for quantitative trait locus (QTL) mapping is the cosegregation of the alleles at a QTL with those at one or a set of known polymorphic markers genotyped on a genome in an experimental cross [1,2]. If a QTL is cosegregating with molecular markers, the genetic effects of the QTL on a quantitative trait and its genomic position can be estimated from the marker genotypes and phenotypic values of the trait. This estimation process particularly assumes the QTL to be located within an interval constructed by a pair of flanking markers in which a test statistics calculated under the reduced (there is no QTL) and full model (there is a QTL) is used to test the existence of the QTL and estimate its position. This so-called interval mapping approach and its extensions [3–5] is robust and powerful for the detection of major QTL and presents the most efficient way to utilize marker information when marker maps are sparse [6]. However, interval mapping is limited by its incapacity to infer any information about the sequence structure and organization of the QTL. Partly for this reason, only a few QTL mapped from markers have been successfully cloned [7–9], despite a considerable number of QTL reported in the literature.

Interval QTL mapping also has an unsolved statistical difficulty when it is used with a high-density linkage map. With more markers genotyped, a genetic map for QTL identification has tended to be infinitely dense. For such an infinitely dense map in which markers are located everywhere over the genome, test statistics at nearby intervals are not independent any more. Thus, the critical threshold used to acclaim the existence of a QTL by interval mapping will be difficult to analytically determine. Although an empirical alternative based on permutation tests has been proposed for threshold determination [10], extensive computing may affect the use efficiency of interval mapping.

Despite its unsuitability for interval mapping of QTL, an infinitely dense map provides an important tool for characterizing genetic variants that contribute to quantitative variation via the analysis of haplotypes composed of non-alleles at a set of highly linked markers. Recent genetic studies suggest that a gene may determine a complex trait, such as body weight or drug response, through its haplotype rather than genotype [11,12]. The completion of the genome projects for several important organisms, Arabdopsis, chicken, human, mouse and poplar, has made massive amounts of DNA sequence data available. In particular, single nucleotide polymorphisms (SNPs), being the most common type of variant in the DNA sequence, provide a powerful means for genotyping the whole genome or any part of it. This facilitates the identification of specific SNP-constructed haplotypes which are responsible for quantitative traits. A set of SNPs that cause quantitative differences among individuals are called quantitative trait nucleotides (QTNs). Liu et al. [13] proposed a statistical model for estimating and testing haplotype effects at a QTN in a random sample drawn from a natural population. This model is based on the population genetic properties of gene segregation. Through the implementation of the EM algorithm, population genetic parameters of SNPs, such as haplotype frequencies, allele frequencies and linkage disequilibria, and quantitative genetic parameters, such as haplotype effects of a QTN, are estimated with closed forms.

Academic Editor: Peter Heutink, Vrije Universiteit Medical Centre, Netherlands

Received April 10, 2007; **Accepted** July 13, 2007; **Published** August 15, 2007

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

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: rwu@mail.ifas.ufl.edu

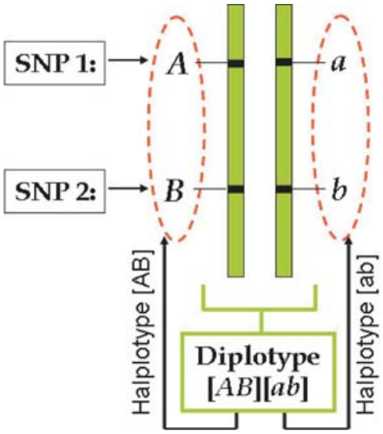

Figure 1. Haplotype configuration of a diplotype for two hypothesized SNPs.
doi:10.1371/journal.pone.0000732.g001

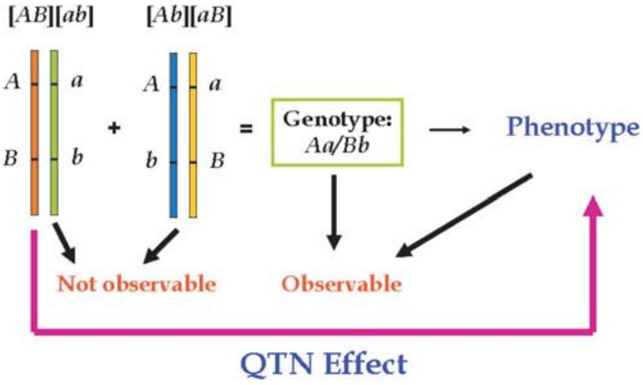

Figure 2. Diplotype configuration of a genotype for two hypothesized SNPs.
doi:10.1371/journal.pone.0000732.g002

The motivation of this work is to derive a statistical model for haplotype discovery responsible for quantitative variation in a mapping population derived an experimental cross. Unlike a natural population in which gene co-segregation analysis is based on linkage disequilibria [14], experimental crosses, such as the backcross or F_2, have usually been analyzed in terms of the linkage between different markers and QTL. In this article, we will frame a general statistical model for estimating the linkage between different SNPs and testing haplotype effects within the context of linkage disequilibrium analysis in an F_2 pedigree. We show that the new model can test for the dependence of SNPs when a multi-point analysis is performed. We have derived closed forms for the EM algorithm to estimate a variety of genetic parameters. A worked example is used to validate the usefulness and utilization of the model.

METHODS

Haplotype and diplotype

A haplotype represents a linear arrangement of nucleotides (alleles) at different SNPs on a single chromosome, or part of a chromosome. The pair of haplotypes is called a diplotype. The observed phenotype of a diplotype is called a genotype. A diplotype is always constructed by two haplotypes, one from the maternal parent and the other from the paternal parent. Suppose there are two different SNPs on the same genomic region, one with two alleles A and a and the other with two alleles B and b, respectively. Allele A from SNP 1 and allele B from SNP 2 are located on the first homologous chromosome, whereas allele a from SNP 1 and allele b from SNP 2 located on the second homologous chromosome. Thus, $[AB]$ is one haplotype and $[ab]$ is a second haplotype, and both constitute a diplotype $[AB][ab]$ (Fig. 1).

In a practical genetic analysis, we can only observe the genotype expressed as Aa/Bb. However, the double heterozygote may be one (and only one) of two possible diplotypes $[AB][ab]$ and $[Ab][aB]$. But these two diplotypes cannot be directly observed and should be inferred from SNP genotype data (Fig. 2). In practice, it is important to estimate haplotype effects on a quantitative trait based on the diplotypes and therefore genotypes. For example, if an animal carries haplotype $[AB]$, it will grow better than other animals that carries any other haplotypes, $[Ab]$, $[aB]$ and $[ab]$. For this reason, the same genotype Aa/Bb may perform differently, depending on what diplotype it carries. If this genotype is diplotype $[AB][ab]$, then it will have a better growth. If the animal is diplotype $[Ab][aB]$, its growth will be poorer. The statistical model being developed will be used to determine which diplotype is associated with better growth in experimental crosses.

Linkage disequilibrium in the F_2 intercross

A general model: Haplotype analysis in the backcross is straightforward because the diplotype is determined for all the backcross genotype. Simple analysis of variance can be used to detect haplotype effects on a quantitative trait. In the F_2, this is not a case in which the double heterozygote is a mixture of two possible diplotypes.

Suppose many SNPs are genotyped each of which is segregating in a 1:2:1 Mendelian ratio in the F_2 population. As seen in the human genome [15], these SNPs are divided into different haplotype blocks. 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 [16–18]. Consider a QTN that contains L htSNPs among which there exist linkage disequilibria of different orders. The two alleles, 1 and 0, at each of these SNPs are symbolized by $r_1,\ldots,r_L$, respectively. For a cross initiated with two inbred parents, the allele frequencies for each of these htSNPs should be $1/2$. A haplotype frequency, denoted as $p_{r_1r_2\cdots r_L}$, is decomposed into the following components:

$$\begin{aligned}
&p_{r_1r_2\ldots r_L} && \\
&=p_{r_1}p_{r_2}\ldots p_{r_L} && \text{No LD}\\
&+(-1)^{r_{L-1}+r_L}p_{r_1}\ldots p_{r_{L-2}}D_{(L-1)L}+\ldots && \\
&+(-1)^{r_1+r_2}p_{r_3}\ldots p_{r_L}D_{12} && \text{Digenic LD}\\
&+(-1)^{r_{L-2}+r_{L-1}+r_L}p_{r_1}\ldots p_{r_{L-3}}D_{(L-2)(L-1)L} && \\
&+\ldots+(-1)^{r_1+r_2+r_3}p_{r_4}\ldots p_{r_L}D_{123} && \text{Trigenic LD}\\
&+\ldots && \\
&+(-1)^L(-1)^{r_1+\ldots+r_L}D_{1\ldots L} && \text{L}-\text{genic LD}
\end{aligned} \tag{1}$$

where D's are the linkage disequilibria of different orders among particular SNPs.

Totally, L SNPs form 2^L haplotypes expressed as $[r_1\ldots r_L]$, $2^{L-1}(2^L+1)$ diplotypes, i.e., a pair of maternally- (m) and paternally-derived haplotypes (p), expressed as $[r_1{}^{\mathrm{m}}\ldots r_L{}^{\mathrm{m}}][r_1{}^{\mathrm{p}}\ldots r_L{}^{\mathrm{p}}]$ $(r_1{}^{\mathrm{m}}, r_1{}^{\mathrm{p}},\ldots;r_L{}^{\mathrm{m}},r_L{}^{\mathrm{p}}=1,0)$ and 3^L genotypes expressed as $r_1r'_1/\ldots/r_Lr'_L$ $(r_1\geq r'_1,\ldots,r_L\geq r'_L=1,0)$. Only genotypes can be observed. The number of diplotypes is smaller than the number of genotypes because the genotypes that are heterozygous at two or more SNPs contain multiple different diplotypes. Diplotype (and therefore genotype) frequencies can be expressed in terms of haplotype frequencies. We use $P_{[r_1^{\mathrm{m}}\ldots r_L^{\mathrm{m}}][r_1^{\mathrm{p}}\ldots r_L^{\mathrm{p}}]}$ and $P_{r_1\,r'_1\,/\ldots/r_L\,r'_L}$ to denote the diplotype and genotype frequencies, respectively, and $n_{r_1\,r'_1\,/\ldots/r_L\,r'_L}$ to denote genotype observation.

A special case: Two-point linkage disequilibrium: For two given SNPs ($\mathbf{S}_1$ and $\mathbf{S}_2$), there are four different haplotypes in a cross population. According to the definition given above, these four haplotypes are denoted as [11], [10], [01] and [00], whose frequencies in a cross population are, respectively, expressed as

$$\begin{aligned}p_{11}&=\frac{1}{4}+D,\\ p_{10}&=\frac{1}{4}-D,\\ p_{01}&=\frac{1}{4}-D,\\ p_{00}&=\frac{1}{4}+D.\end{aligned}\tag{2}$$

Assume that the two SNPs are linked with a recombination fraction r. The haplotype frequencies can be expressed in terms of r, i.e., $p_{11}=\frac{1}{2}(1-r)$, $p_{10}=\frac{1}{2}r$, $p_{01}=\frac{1}{2}r$ and $p_{00}=\frac{1}{2}(1-r)$. Combining equation (2), this establishes the relation between the linkage disequilibrium and recombination fraction as

$$D=\frac{1}{4}(1-2r),\tag{3}$$

or

$$r=\frac{1}{2}(1-4D).\tag{4}$$

A special case: Three-point linkage disequilibrium: For three given SNPs ($\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$), there are eight different haplotypes, i.e., [111], [110], [101], [100], [011], [010], [001], and [000]. The haplotype frequencies in a cross population are, respectively, expressed as

$$\begin{aligned}p_{111}&=\frac{1}{8}+\frac{1}{2}D_{23}+\frac{1}{2}D_{13}+\frac{1}{2}D_{12}+D_{123}\\ p_{110}&=\frac{1}{8}-\frac{1}{2}D_{23}-\frac{1}{2}D_{13}+\frac{1}{2}D_{12}-D_{123}\\ p_{101}&=\frac{1}{8}-\frac{1}{2}D_{23}+\frac{1}{2}D_{13}-\frac{1}{2}D_{12}-D_{123}\\ p_{100}&=\frac{1}{8}+\frac{1}{2}D_{23}-\frac{1}{2}D_{13}-\frac{1}{2}D_{12}+D_{123}\\ p_{011}&=\frac{1}{8}+\frac{1}{2}D_{23}-\frac{1}{2}D_{13}-\frac{1}{2}D_{12}-D_{123}\\ p_{010}&=\frac{1}{8}-\frac{1}{2}D_{23}+\frac{1}{2}D_{13}-\frac{1}{2}D_{12}+D_{123}\\ p_{001}&=\frac{1}{8}-\frac{1}{2}D_{23}-\frac{1}{2}D_{13}+\frac{1}{2}D_{12}+D_{123}\\ p_{000}&=\frac{1}{8}+\frac{1}{2}D_{23}+\frac{1}{2}D_{13}+\frac{1}{2}D_{12}-D_{123}\end{aligned}\tag{5}$$

where D_{12}, D_{23} and D_{13} are the linkage disequilibria between SNP $\mathbf{S}_1$ and $\mathbf{S}_2$, between $\mathbf{S}_2$ and $\mathbf{S}_3$ and between $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, and D_{123} is the linkage disequilibrium among the three SNPs. The four disequilibrium coefficients can be estimated, by solving equation (5), as

$$\begin{aligned}D_{12}&=\frac{1}{4}[(p_{111}+p_{110}+p_{001}+p_{000})-(p_{101}+p_{100}+p_{011}+p_{010})]\\ D_{23}&=\frac{1}{4}[(p_{111}+p_{011}+p_{100}+p_{000})-(p_{110}+p_{010}+p_{101}+p_{001})]\\ D_{13}&=\frac{1}{4}[(p_{111}+p_{101}+p_{010}+p_{000})-(p_{110}+p_{100}+p_{011}+p_{001})]\\ D_{123}&=\frac{1}{8}[(p_{111}+p_{100}+p_{010}+p_{001})-(p_{110}+p_{101}+p_{011}+p_{000})]\end{aligned}\tag{6}$$

The first three first-order linkage disequilibria can be used to describe the linkage between different SNPs and crossover interference, whereas the last second-order linkage disequilibrium is thought to be associated with chromatid interference.

Haplotyping a trait with two SNPs

Our interest is to search for the haplotype diversity that can explain phenotypic variation in a complex trait. The association between haplotype diversity and phenotypic variation has been detected in several studies of drug responses [11,12]. This allows us to assume that a particular haplotype is different from other haplotypes for a given trait. Here, our focus will be on modelling haplotype effects in experimental crosses. Although haplotypes (comprising diplotypes) can be directly observed in the backcross, this is not possible for the F_2 because their heterozygous genotypes are not concordant with diplotypes or haplotypes. For the F_2 population, the effects of different haplotypes on the phenotype need be postulated from observed zygotic genotypes. The inference of diplotypes for a particular genotype is statistically a missing data problem that can be formulated by a finite mixture model.

Mixture model: The statistical method for the genomewide scan of QTN is formulated on the basis of a finite mixture model. The mixture model assumes that each observation comes from one of an assumed set of distributions. The mixture model derived to detect haplotype effects on a quantitative trait based on SNP genotype data contains three major parts: (1) the mixture proportions of each distribution, denoted as the relative frequencies of different diplotypes for the same SNP genotype, (2) the mean for each diplotype in the density function, and (3) the residual variance common to all diplotypes.

For simplicity, we consider a QTN that is composed of only two SNPs each with two alleles designated as 1 and 0. These two SNPs segregating in the F_2 population form four haplotypes whose frequencies are arrayed in vector $\Theta\mathrm{p}=(p_{11}, p_{10}, p_{01}, p_{00})$. All the genotypes are consistent with diplotypes, except for the double heterozygote, 10/10, that contains two different diplotypes [11][00] with a frequency of $2\,p_{11}p_{00}$ and [10][01] with a frequency of $2\ p_{10}p_{01}$ (Table 1). The relative frequencies of different diplotypes for the double heterozygote are a function of haplotype frequencies.

A total of n individuals in the F_2 are classified into 9 genotypes for the two SNPs, each genotype with observation generally expressed as $n_{r_1\,r'_1\,/r_2\,r'_2}$ $(r_1\geq r'_1,r_2\geq r'_2,r_3\geq r'_3=1,0)$. The frequency of each genotype can be expressed in terms of haplotype frequencies (Table 1). Considering a quantitative trait controlled by diplotype (rather than genotype) diversity, the phenotypic value

Table 1. Diplotypes and their frequencies for each of nine genotypes at two SNPs within a QTN, haplotype composition frequencies for each genotype, and composite diplotypes for four possible risk haplotypes.

Genotype	Diplotype		Relative diplotype frequency	Risk haplotype			
	Configuration	Frequency		[11]	[10]	[01]	[00]
11/11	[11][11]	$p^2{}_{11}$	1	AA	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$
11/10	[11][10]	$2p_{11}p_{10}$	1	$A\bar{A}$	$A\bar{A}$	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$
11/00	[10][10]	$p^2{}_{10}$	1	$\bar{A}\bar{A}$	AA	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$
10/11	[11][01]	$2p_{11}p_{01}$	1	$A\bar{A}$	$\bar{A}\bar{A}$	$A\bar{A}$	$\bar{A}\bar{A}$
10/10	$\begin{cases}[11][00]\\ [10][01]\end{cases}$	$\begin{cases}2p_{11}p_{00}\\ 2p_{10}p_{01}\end{cases}$	$\begin{cases}\phi\\ 1-\phi\end{cases}$	$\begin{cases}AA\\ AA\end{cases}$	$\begin{cases}AA\\ AA\end{cases}$	$\begin{cases}AA\\ AA\end{cases}$	$\begin{cases}AA\\ AA\end{cases}$
10/00	[10][00]	$2p_{10}p_{00}$	1	$\bar{A}\bar{A}$	$A\bar{A}$	$\bar{A}\bar{A}$	$A\bar{A}$
00/11	[01][01]	$p^2{}_{01}$	1	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$	AA	$\bar{A}\bar{A}$
00/10	[01][00]	$2p_{01}p_{00}$	1	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$	$A\bar{A}$	$A\bar{A}$
00/00	[00][00]	$p^2{}_{00}$	1	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$	$\bar{A}\bar{A}$	AA

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. By assuming different haplotypes as a risk haplotype, composite diplotypes are accordingly defined and their genotypic values are given.

doi:10.1371/journal.pone.0000732.t001

of the trait (y_i) for individual i is expressed by a linear model, i.e.,

$$y_i=\sum_{r_1^{\mathrm{m}}=0}^{1}\sum_{r_1^{\mathrm{p}}=0}^{1}\sum_{r_2^{\mathrm{m}}=0}^{1}\sum_{r_2^{\mathrm{p}}=0}^{1}\xi_i u_{[r_1^{\mathrm{m}}r_2^{\mathrm{m}}][r_1^{\mathrm{p}}r_2^{\mathrm{p}}]}+e_i, \tag{7}$$

where ξ_i is the indicator variable defined as 1 if a diplotype considered is compatible with subject i and as 0 otherwise, $u_{[r_1^{\mathrm{m}}r_2^{\mathrm{m}}][r_1^{\mathrm{p}}r_2^{\mathrm{p}}]}=u_{[r_1^{\mathrm{p}}r_2^{\mathrm{p}}][r_1^{\mathrm{m}}r_2^{\mathrm{m}}]}$ is the genotypic value for diplotype $[r_1^{\mathrm{m}}r_2^{\mathrm{m}}][r_1^{\mathrm{p}}r_2^{\mathrm{p}}]$, and e_i is the residual error distributed as $\mathcal{N}(0,\sigma^2)$.

Assume that this QTN triggers an effect on the trait because at least one haplotype is different from the remaining seven. Without loss of generality, let [11] be such a distinct haplotype, called *risk haplotype*, designated as A. All the other non-risk haplotypes, [10], [01] and [00], are collectively expressed as $\bar{A}$. The risk and non-risk haplotypes form three *composite diplotypes* AA (**2**), $A\bar{A}$ (**1**) and $\bar{A}\bar{A}$ (**0**). Let μ_2, μ_1 and μ_0 be the genotypic value of the three composite diplotypes, respectively (Table 1). The means for different composite diplotypes and residual variance are arrayed by a quantitative genetic parameter vector $\Theta_q=(\mu_{2,}\ \mu_{1,}\ \mu_{0,}\ \sigma^2)$.

Likelihoods: With the above notation, we construct two likelihoods, one for haplotype frequencies (Θ_p) based on SNP data (**S**) and the other for quantitative genetic parameters (Θ_q) based on haplotype frequencies (Θ_p), phenotypic (y) and SNP data (**S**). They are, respectively, expressed as

$$\begin{array}{rr}
log L(\Theta_p|\boldsymbol{S})= & \log L(\Theta_p|\Theta_q,y,\boldsymbol{S})= \\
+2n_{11/11}\log p_{11} & \sum_{i=1}^{n_{11/11}}\log f_2(y_i) \\
+n_{11/10}\log(2p_{11}p_{10}) & +\sum_{i=1}^{n_{11/10}}\log f1(y_i) \\
+2n_{11/00}\log p_{10} & +\sum_{i=1}^{n_{11/00}}\log f_0(y_i) \\
+n_{10/11}\log(2p_{11}p_{01}) & +\sum_{i=1}^{n_{10/11}}\log f_1(y_i) \\
+n_{10/10}\log(2p_{11}p_{00}+2p_{10}p_{01}) & +\sum_{i=1}^{n_{10/10}}\log[\phi f_1(y_i)+(1-\phi)f_0(y_i)] \\
+n_{10/00}\log(2p_{10}p_{00}) & +\sum_{i=1}^{n_{10/00}}\log f_0(y_i) \\
+2n_{00/11}\log p_{01} & +\sum_{i=1}^{n_{00/11}}\log f_0(y_i) \\
+n_{00/10}\log(2p_{01}p_{00}) & +\sum_{i=1}^{n_{00/10}}\log f_0(y_i) \\
+2n_{00/00}\log p_{00} & +\sum_{i=1}^{n_{00/00}}\log f_0(y_i)
\end{array} \tag{8}$$

where f_j (y_i) is a normal distribution density function of composite diplotype j ($j=2,1,0$), i.e.,

$$f_j(y_i)=\frac{1}{\sqrt{2\pi}\sigma}exp\left[-\frac{(y_i-\mu_j)^2}{2\sigma^2}\right].$$

It can be seen from the above likelihood functions that, although most zygote genotypes contain a single component (diplotype), the double heterozygote is the mixture of two possible diplotypes weighted by ϕ and 1-ϕ, expressed as

$$\phi=\frac{p_{11}p_{00}}{p_{11}p_{00}+p_{10}p_{01}}, \tag{9}$$

which represents the relative frequency of diplotype [11][00] for the double heterozygote.

It should be noted that $L(\Theta_p,\ \Theta_q\ |\ y,\ \mathbf{S})$ relies on the haplotype frequencies defined in $L(\Theta_p|\mathbf{S})$ and, thus, the latter is thought to be nested within the former. The estimates of parameters that maximize $L(\Theta_p|\mathbf{S})$ can also maximize the $L(\Theta_p,\ \Theta_q\ |\ y,\ \mathbf{S})$.

The EM algorithm: A closed-form solution for the EM algorithm has been derived to estimate the unknown parameters that maximize the two likelihoods of (26) [13]. The estimates of haplotype frequencies are based on the log-likelihood function $L(\Theta_p|\mathbf{M})$, whereas the estimates of diplotype genotypic means and residual variance are based on the log-likelihood function $L(\Theta_p,\ \Theta_q\ |\ y,\ \mathbf{M})$. These two different types of parameters can be estimated using a two-stage hierarchical EM algorithm.

At a higher hierarchy of the EM algorithm, the E step is aimed to calculate the relative frequency (ϕ) of diplotype [11][00] in the double heterozygote is calculated by equation (9). The M step is aimed to estimate the haplotype frequencies based on the probabilities calculated in the previous iteration using

$$\begin{aligned}
\hat{p}_{11}&=\frac{1}{2n}(2n_{11/11}+\phi n_{10/10}+n_{11/10}+n_{10/11})\\
\hat{p}_{10}&=\frac{1}{2n}[2n_{11/00}+n_{11/10}+(1-\phi)n_{10/10}+n_{10/00}]\\
\hat{p}_{01}&=\frac{1}{2n}[2n_{00/11}+n_{10/11}+(1-\phi)n_{10/10}+n_{00/10}]\\
\hat{p}_{00}&=\frac{1}{2n}(2n_{00/00}+\phi n_{10/10}+n_{01/00}+n_{10/00})
\end{aligned} \tag{10}$$

At a lower hierarchy of the EM algorithm, the E step is derived to calculate the posterior probability ($\Omega[_{11}]_{[00]i}$) of individual i with the double heterozygous genotype to be diplotype [11][00] by

$$\Omega_{[11][00]i}=\frac{\phi f_{[11][00]}(y_i)}{\phi f_{[11][00]}(y_i)+(1-\phi)f_{[10][01]}(y_i)}.$$

Note that for all the other genotypes, such posterior probabilities do not exist.

By assuming that [11] is a risk haplotype, the M step is derived to estimate the genotypic values (μ_j) for each composite diplotype and the residual variance based on the calculated posterior probabilities by

$$\begin{aligned}\hat{\mu}_2 &= \frac{\sum_{i=1}^{n_{11/11}} y_i}{n_{11/11}},\\ \hat{\mu}_1 &= \frac{\sum_{i=1}^{\dot{n}} y_i+\sum_{i=1}^{n_{10/10}} \Omega_{[11][00]i}y_i}{\dot{n}+\sum_{i=1}^{n_{10/10}} \Omega_{[11][00]i}},\\ \hat{\mu}_0 &= \frac{\sum_{i=1}^{\ddot{n}} y_i+\sum_{i=1}^{n_{10/10}} (1-\Omega_{[11][00]i})y_i}{\ddot{n}+\sum_{i=1}^{n_{10/10}} (1-\Omega_{[11][00]i})},\end{aligned} \tag{12}$$

$$\begin{aligned}\hat{\sigma}^2 &= \tfrac{1}{n}\{\sum_{i=1}^{n_{11/11}} (y_i-\hat{\mu}_2)^2+\sum_{i=1}^{\dot{n}} (y_i-\hat{\mu}_1)^2+\sum_{i=1}^{\ddot{n}} (y_i-\hat{\mu}_0)^2\\ &+ \sum_{i=1}^{n_{10/10}} \left[\Omega_{[10/10]i}(y_i-\hat{\mu}_1)^2+(1-\Omega_{[10/10]i})(y_i-\hat{\mu}_0)^2\right]\},\end{aligned} \tag{13}$$

where

$$\begin{aligned}\dot{n}&=n_{11/10}+n_{10/11},\\ \ddot{n}&=n_{11/00}+n_{10/00}+n_{01/01}+n_{01/00}+n_{00/00}.\end{aligned}$$

Iterations including the E and M steps are repeated at the higher hierarchy between equations (9) and (10) and at the lower hierarchy among equations (12) and (13) until the estimates of the parameters converge to stable values. The sampling errors of these parameters can be estimated by calculating Louis' [19] observed information matrix.

Haplotype frequencies can be expressed as a function of allelic frequencies and linkage disequilibrium. Based on equation (2), we solve the linkage disequilibrium between two SNPs by

$$\begin{aligned}\hat{D}&=\hat{p}_{11}-\frac{1}{4}\\ &=\frac{1}{4}-\hat{p}_{10}.\end{aligned} \tag{14}$$

With the genotypic means of composite diplotypes, we can estimate the overall mean (μ) and additive (a) and dominant genetic effects (d) due to the QTN detected, respectively, by

$$\begin{aligned}\hat{\mu} &= \frac{1}{2}(\hat{\mu}_2+\hat{\mu}_0)\\ \hat{a} &= \frac{1}{2}(\hat{\mu}_2-\hat{\mu}_0)\\ \hat{d} &= \hat{\mu}_1-\frac{1}{2}(\hat{\mu}_2+\hat{\mu}_0)\end{aligned}$$

Model selection: The likelihood $L(\Theta_p, \Theta_q \mid y, \mathbf{S})$ is formulated by assuming that haplotype [11][11] is a risk haplotype. However, a real risk haplotype is unknown from raw data (y, $\mathbf{S}$). An additional step for the choice of the most likely risk haplotype should be implemented. The simplest way to do so is to calculate the likelihood values by assuming that any one of the four haplotypes can be a risk haplotype (Table 1). Thus, we obtain four possible likelihood values under different risk haplotypes; that is, (1) $L_1(\hat{\Theta}_p, \hat{\Theta}_{1q} | y,\boldsymbol{S})$ for [11], (2) $L_2(\hat{\Theta}_p, \hat{\Theta}_{2q} | y,\boldsymbol{S})$ for [10], (3) $L_3(\hat{\Theta}_p, \hat{\Theta}_{3q} | y,\boldsymbol{S})$ for [01], and (4) $L_4(\hat{\Theta}_p, \hat{\Theta}_{4q} | y,\boldsymbol{S})$ for [00]. Under each possible risk haplotype, we estimate the quantitative genetic parameters $\hat{\Theta}_{kq}$ ($k=1,\ldots,4$). The largest likelihood value calculated is thought to correspond to the most likely risk haplotype.

In practice, it is also possible that there exist more than one risk haplotypes for a QTN. Relative to the bi-"allelic" QTN with one risk haplotype, such a QTN is called a multi-"allelic" QTN. If there are two risk haplotypes, we will have six composite diplotypes. Assuming that [11] (denoted by A_1) and [10] (denoted by A_2) are risk haplotypes and the remaining haplotypes [10] and [01] are non-risk haplotypes (denoted by A_3), then six composite diplotypes, expressed as A_1A_1, A_1A_2, A_1A_3, A_2A_2, A_2A_3 and A_3A_3, can be specified according to the diplotype distribution as shown in Table 1. Totally, there are six such haplotype combinations for a two-SNP QTL, each combination corresponding to a likelihood value. Based on the calculated likelihoods, we can determine a most likely risk and non-risk haplotype combination. If there are three risk haplotypes, we will have 10 different composite diplotypes. The optimal risk and non-risk haplotype combination will be selected from three combinations based on the likelihoods.

The likelihood can be used as a criterion to select the optimal risk and non-risk haplotype combination when the number of risk haplotype is the same. However, when the number of risk haplotype is different, an AIC- or BIC-based model selection strategy [20] should be used because of different numbers of parameters being estimated in this case.

Hypothesis tests: We can test two major hypotheses in the following sequence: (1) the association between two SNPs by testing their linkage disequilibrium, and (2) the difference of a given haplotype from the remaining haplotypes by testing the significance of haplotype additive and dominant effects on the trait. The linkage disequilibrium between two given SNPs can be tested using two alternative hypotheses:

$$H_0 : D=0 \text{ vs. } H_1 : D\neq 0 \tag{15}$$

The log-likelihood ratio test statistic for the significance of LD is calculated by comparing the likelihood values under the H_1 (full model) and H_0 (reduced model) using

$$LR_1=-2[logL(p_{11}=p_{10}=p_{01}=p_{00}=\frac{1}{4}|\boldsymbol{S})-logL(\hat{\Theta}_p|\boldsymbol{S})], \tag{16}$$

The LR_1 is considered to asymptotically follow a χ^2 distribution with one degree of freedom.

Diplotype or haplotype effects on the trait, i.e., the existence of a QTN, can be tested using the following hypotheses expressed as

$$H_0 : \mu_j \equiv \mu \text{ vs. } H_1 : \text{at least one equality in } H_0 \text{ does not hold}, j=2,1,0 \tag{17}$$

The log-likelihood ratio test statistic (LR_2) under these two hypotheses can be similarly calculated,

$$LR_2=-2[logL(\tilde{\Theta}_p,\tilde{\mu}|y,\boldsymbol{S})-logL(\hat{\Theta}_p,\hat{\Theta}_q|y,\boldsymbol{S})], \tag{18}$$

where the tildes and hats denote the MLEs of parameters under the null and alternative hypotheses of (17), respectively. Although the critical threshold for determining the existence of a QTN can be based on empirical permutation tests, the LR_2 may asymptotically follow a χ^2 distribution with two degrees of freedom, so that the threshold can be obtained from the χ^2 distribution table.

Haplotyping a trait with multiple SNPs

Haplotype structure: The statistical method for QTN mapping is exemplified by a set of three SNPs, $\mathbf{S}_1$–$\mathbf{S}_3$, for a QTN. Two alleles 1 and 0 at each SNP are symbolized by r_1, r_2 and r_3, respectively. Eight haplotypes, [111], [110], [101], [100], [011], [010], [001] and [000], formed by these three SNPs, have the frequencies arrayed in $\Theta_p = (p_{111,}\ p_{110,}\ p_{101,}\ p_{100,}\ p_{011,}\ p_{010,}\ p_{001,}\ p_{000})$. Some genotypes are consistent with diplotypes, whereas the others that are heterozygous at two or more SNPs are not. Each double heterozygote contains two different diplotypes. One triple heterozygote, i.e., 10/10/10, contains four different diplotypes, [111][000] (in a probability of $2p_{111}p_{000}$), [110][001] (in a probability of $2p_{110}p_{001}$), [101][010] (in a probability of $2p_{101}p_{010}$) and [100][011] (in a probability of $2p_{100}p_{011}$). The relative frequencies of different diplotypes for this double or triple heterozygote are a function of haplotype frequencies (Table 2).

In the F_2 population, there are 27 genotypes for the three SNPs. Let $n_{r_1 r'_1/r_2 r'_2/r_3 r'_3}$ ($r_1 \geq r'_1, r_2 \geq r'_2, r_3 \geq r'_3 = 1,0$) be the number of offspring for a genotype. As seen in Table 2, the frequency of each genotype is expressed in terms of haplotype frequencies. Similar to equation (25), the phenotypic value of the trait for individual i is expressed, at the diplotype level, as

$$y_i = \sum_{r_1^m=0}^{1} \sum_{r_1^p=0}^{1} \sum_{r_2^m=0}^{1} \sum_{r_2^p=0}^{1} \sum_{r_3^m=0}^{1} \sum_{r_3^p=0}^{1} \xi_i u_{[r_1^{\mathrm{m}} r_2^{\mathrm{m}} r_3^{m}][r_1^{\mathrm{p}} r_2^{\mathrm{p}} r_3^{p}]} + e_i, \qquad (19)$$

where ξ_i is the indicator variable defined as 1 if a diplotype considered is compatible with subject i and as 0 otherwise, $u_{[r_1^{\mathrm{m}} r_2^{\mathrm{m}} r_3^{m}][r_1^{\mathrm{p}} r_2^{\mathrm{p}} r_3^{p}]} = u_{[r_1^{\mathrm{p}} r_2^{\mathrm{p}} r_3^{p}][r_1^{\mathrm{m}} r_2^{\mathrm{m}} r_3^{m}]}$ is the genotypic value for diplotype $[{r_1}^{\mathrm{m}}{r_2}^{\mathrm{m}}{r_3}^{\mathrm{m}}][{r_1}^{\mathrm{p}}{r_2}^{\mathrm{p}}{r_3}^{\mathrm{p}}]$, and e_i is the residual error distributed as $\mathcal{N}(0,\sigma^2)$. Note that **m** and **p** stand for the maternally and paternally derived alleles, respectively.

By assuming [111] as a risk haplotype (labelled by A) and all the others as non-risk haplotypes (labelled by $\bar{A}$), Table 2 provides the formulation of genotypic values for three composite diplotypes, μ_2 for AA, μ_1 for $A\bar{A}$ and μ_0 for $\bar{A}\bar{A}$. The haplotype effect parameters and residual covariance matrix are arrayed by a quantitative genetic parameter vector $\Theta_q = (\mu_2,\ \mu_1,\ \mu_0, \sigma^2)$.

Likelihoods and algorithms: With the above notation, we construct two likelihoods, one for haplotype frequencies (Θ_p) based on SNP data ($\mathbf{S}$) and the other for quantitative genetic parameters (Θ_q) based on haplotype frequencies (Θ_p), phenotypic (y) and SNP data ($\mathbf{S}$). They are, respectively, expressed as

$$\begin{array}{ll}
logL(\Theta_p|\mathbf{S}) = \text{constant} & logL(\Theta_q|\Theta_p, y, \mathbf{S}) = \\
+2n_{11/11/11} logp_{111} & \sum_{i=1}^{n_{11/11/11}} logf_2(y_i) \\
+n_{11/11/10} log(2p_{111}p_{110}) & + \sum_{i=1}^{n_{11/11/10}} logf_1(y_i) \\
+2n_{11/11/00} logp_{110} & + \sum_{i=1}^{n_{11/11/00}} logf_0(y_i) \\
+n_{11/10/11} log(2p_{111}p_{101}) & + \sum_{i=1}^{n_{11/10/11}} logf_1(y_i) \\
+n_{11/10/10} log(2p_{111}p_{100} + 2p_{110}p_{101}) & + \sum_{i=1}^{n_{11/10/10}} log[\phi_1 f_1(y_i) + \bar{\phi}_1 f_0(y_i)] \\
+n_{11/10/00} log(2p_{110}p_{100}) & + \sum_{i=1}^{n_{11/10/00}} logf_0(y_i) \\
+2n_{11/00/11} logp_{101} & + \sum_{i=1}^{n_{11/00/11}} logf_0(y_i) \\
+n_{11/00/10} log(2p_{101}p_{100}) & + \sum_{i=1}^{n_{11/00/10}} logf_0(y_i) \\
+2n_{11/00/00} logp_{100} & + \sum_{i=1}^{n_{11/00/00}} logf_0(y_i) \\
+n_{10/11/11} log(2p_{111}p_{011}) & + \sum_{i=1}^{n_{10/11/11}} logf_1(y_i) \\
+n_{10/11/10} log(2p_{111}p_{010} + 2p_{110}p_{011}) & + \sum_{i=1}^{n_{10/11/10}} log[\phi_2 f_1(y_i) + \bar{\phi}_2 f_0(y_i)] \\
+n_{10/11/00} log(2p_{110}p_{010}) & + \sum_{i=1}^{n_{10/11/00}} logf_0(y_i) \\
+n_{10/10/11} log(2p_{111}p_{001} + 2p_{101}p_{011}) & + \sum_{i=1}^{n_{10/10/11}} log[\phi_3 f_1(y_i) + \bar{\phi}_3 f_0(y_i)] \\
+n_{10/10/10} log(2p_{111}p_{000} + 2p_{101}p_{010} & \\
\quad +2p_{110}p_{001} + 2p_{100}p_{011}) & + \sum_{i=1}^{n_{10/10/10}} log[\phi_4 f_1(y_i) + \bar{\phi}_4 f_0(y_i)] \\
+n_{10/10/00} log(2p_{110}p_{000} + 2p_{100}p_{010}) & + \sum_{i=1}^{n_{10/10/00}} log[\phi_5 f_0(y_i) + \bar{\phi}_5 f_0(y_i)] \\
+n_{10/00/11} log(2p_{101}p_{001}) & + \sum_{i=1}^{n_{10/00/11}} logf_0(y_i) \\
+n_{10/00/10} log(2p_{101}p_{000} + 2p_{100}p_{001}) & + \sum_{i=1}^{n_{10/00/10}} log[\phi_6 f_0(y_i) + \bar{\phi}_6 f_0(y_i)] \\
+n_{10/00/00} log(2p_{100}p_{000}) & + \sum_{i=1}^{n_{10/00/00}} logf_0(y_i) \\
+2n_{00/11/11} logp_{011} & + \sum_{i=1}^{n_{00/11/11}} logf_0(y_i) \\
+n_{00/11/10} log(2p_{011}p_{010}) & + \sum_{i=1}^{n_{00/11/10}} logf_0(y_i) \\
+2n_{00/11/00} logp_{010} & + \sum_{i=1}^{n_{00/11/00}} logf_0(y_i) \\
+n_{00/10/11} log(2p_{011}p_{001}) & + \sum_{i=1}^{n_{00/10/11}} logf_0(y_i) \\
+n_{00/10/10} log(2p_{011}p_{000} + 2p_{010}p_{001}) & + \sum_{i=1}^{n_{00/10/10}} log[\phi_7 f_0(y_i) + \bar{\phi}_7 f_0(y_i)] \\
+n_{00/10/00} log(2p_{010}p_{000}) & + \sum_{i=1}^{n_{00/10/00}} logf_0(y_i) \\
+2n_{00/00/11} logp_{001} & + \sum_{i=1}^{n_{00/00/11}} logf_0(y_i) \\
+n_{00/00/10} log(2p_{001}p_{000}) & + \sum_{i=1}^{n_{00/00/10}} logf_0(y_i) \\
+2n_{00/00/00} logp_{000} & + \sum_{i=1}^{n_{00/00/00}} logf_0(y_i)
\end{array} \qquad (20)$$

where $\phi_.$'s ($\bar{\phi}. = 1 - \phi$) are defined below, and $f_j\ (y_j)$ ($j = 2,\ 1,\ 0$) is a normal distribution density function of composite diplotype j.

A two-stage hierarchical EM algorithm is derived to estimate haplotype frequencies and quantitative genetic parameters. At the higher hierarchy of the EM framework, we calculate the proportions of a particular diplotype within double or triple heterozygous genotypes (E step) by

$$\begin{array}{lcll}
\phi_1 & = & \frac{p_{111}p_{100}}{p_{111}p_{100} + p_{101}p_{110}}, & \text{for genotype } 11/10/10 \\
\phi_2 & = & \frac{p_{111}p_{010}}{p_{111}p_{010} + p_{011}p_{110}}, & \text{for genotype } 10/11/10 \\
\phi_3 & = & \frac{p_{111}p_{001}}{p_{111}p_{001} + p_{101}p_{011}}, & \text{for genotype } 10/10/11 \\
\phi_4 & = & \frac{p_{111}p_{000}}{p_{111}p_{000} + p_{101}p_{010} + p_{110}p_{001} + p_{100}p_{011}}, & \text{for genotype } 10/10/10 \\
\phi_4^{'} & = & \frac{p_{101}p_{010}}{p_{111}p_{000} + p_{101}p_{010} + p_{110}p_{001} + p_{100}p_{011}}, & \text{for genotype } 10/10/10 \\
\phi_4^{''} & = & \frac{p_{110}p_{001}}{p_{111}p_{000} + p_{101}p_{010} + p_{110}p_{001} + p_{100}p_{011}}, & \text{for genotype } 10/10/10 \\
\phi_4^{'''} & = & \frac{p_{100}p_{011}}{p_{111}p_{000} + p_{101}p_{010} + p_{110}p_{001} + p_{100}p_{011}}, & \text{for genotype } 10/10/10 \\
\phi_5 & = & \frac{p_{110}p_{000}}{p_{110}p_{000} + p_{100}p_{010}}, & \text{for genotype } 10/10/00 \\
\phi_6 & = & \frac{p_{101}p_{000}}{p_{101}p_{000} + p_{001}p_{100}}, & \text{for genotype } 10/00/10 \\
\phi_7 & = & \frac{p_{011}p_{000}}{p_{011}p_{000} + p_{001}p_{010}}, & \text{for genotype } 00/10/10
\end{array} \qquad (21)$$

The calculated relative proportions by equation (21) were used to estimate the haplotype frequencies with

Table 2. Possible diplotypes and their frequencies for each of 27 genotypes at three SNPs within a QTN, and genotypic value vectors of composite diplotypes (assuming that [111] (A) is the risk haplotype and the others ($\bar{A}$) are the non-risk haplotype).

Genotype	Diplotype			Composite diplotype	
	Configuration	Frequency	Relative frequency	Symbol	Mean
11/11/11	[111][111]	$P^2{}_{111}$	1	AA	μ_2
11/11/10	[111][110]	$2p_{111}p_{110}$	1	$A\bar{A}$	μ_1
11/11/00	[110][110]	$P^2{}_{110}$	1	$\bar{A}\bar{A}$	μ_0
11/10/11	[111][101]	$2p_{111}p_{101}$	1	$A\bar{A}$	μ_1
11/10/10	[111][100] [110][101]	$2p_{111}p_{100}$ $2p_{110}p_{101}$	ϕ_1 ϕ_1	AA AA	μ_1 μ_0
11/10/00	[110][100]	$2p_{110}p_{100}$	1	$\bar{A}\bar{A}$	μ_0
11/00/11	[101][101]	$P^2{}_{101}$	1	$\bar{A}\bar{A}$	μ_0
11/00/10	[101][100]	$2p^2{}_{101}p_{100}$	1	$\bar{A}\bar{A}$	μ_0
11/00/00	[100][100]	$P^2{}_{100}$	1	$\bar{A}\bar{A}$	μ_0
10/11/11	[111][011]	$2p_{111}p_{011}$	1	$A\bar{A}$	μ_1
10/11/10	[111][010] [110][011]	$2p_{111}p_{010}$ $2p_{110}p_{011}$	ϕ_2 ϕ_2	AA AA	μ_1 μ_0
10/11/00	[110][010]	$2p_{110}p_{010}$	1	$\bar{A}\bar{A}$	μ_0
10/10/11	[111][001] [101][011]	$2p_{111}p_{001}$ $2p_{101}p_{011}$	ϕ_3 ϕ_3	AA AA	μ_1 μ_0
10/10/10	[111][000] [110][001] [100][011] [101][010]	$2p_{111}p_{000}$ $2p_{110}p_{001}$ $2p_{100}p_{011}$ $2p_{101}p_{010}$	ϕ_4 ϕ_4' ϕ_4'' ϕ_4'''	AA AA AA AA	μ_1 μ_0 μ_0 μ_0
10/10/00	[110][000] [100][010]	$2p_{110}p_{000}$ $2p_{100}p_{010}$	ϕ_5 ϕ_5	AA AA	μ_0 μ_0
10/00/11	[101][001]	$2p_{101}p_{001}$	1	$\bar{A}\bar{A}$	μ_0
10/00/10	[101][000] [100][001]	$2p_{101}p_{000}$ $2p_{100}p_{001}$	ϕ_6 ϕ_6	AA AA	μ_0 μ_0
10/00/00	[100][000]	$2p_{100}p_{000}$	1	$\bar{A}\bar{A}$	μ_0
00/11/11	[011][011]	$P^2{}_{011}$	1	$\bar{A}\bar{A}$	μ_0
00/11/10	[011][010]	$2p_{011}p_{010}$	1	$\bar{A}\bar{A}$	μ_0
00/11/00	[010][010]	$P^2{}_{010}$	1	$\bar{A}\bar{A}$	μ_0
00/10/11	[011][001]	$2p_{011}p_{001}$	1	$\bar{A}\bar{A}$	μ_0
00/10/10	[011][000] [010][001]	$2p_{011}p_{000}$ $2p_{010}p_{001}$	ϕ_7 ϕ_7	AA AA	μ_0 μ_0
00/10/00	[010][000]	$2p_{010}p_{000}$	1	$\bar{A}\bar{A}$	μ_0
00/00/11	[001][001]	$P^2{}_{001}$	1	$\bar{A}\bar{A}$	μ_0
00/00/10	[001][000]	$2p_{001}p_{000}$	1	$\bar{A}\bar{A}$	μ_0
00/00/00	[000][000]	$P^2{}_{000}$	1	$\bar{A}\bar{A}$	μ_0

doi:10.1371/journal.pone.0000732.t002

$$
\begin{array}{rclrcl}
\hat{p}_{111} & = & \frac{1}{2n}(2n_{11/11/11}+n_{11/11/10}+n_{11/10/11}+n_{10/11/11} & \hat{p}_{011} & = & \frac{1}{2n}(2n_{00/11/11}+n_{10/11/11}+n_{00/10/11}+n_{00/11/10} \\
 & + & \phi_1 n_{11/10/10}+\phi_2 n_{10/11/10}+\phi_3 n_{10/10/11}+\phi_4 n_{10/10/10}) & & + & \bar{\phi}_2\, n_{10/11/10}+\bar{\phi}_3\, n_{10/10/11}+\phi'''{}_4\, n_{10/10/10}+\phi_7 n_{00/10/10}) \\
\hat{p}_{110} & = & \frac{1}{2n}(2n_{11/11/00}+n_{11/11/10}+n_{11/10/00}+n_{10/11/00} & \hat{p}_{010} & = & \frac{1}{2n}(2n_{00/11/00}+n_{10/11/00}+n_{00/11/10}+n_{00/10/00} \\
 & + & \bar{\phi}_1\, n_{11/10/10}+\bar{\phi}_2\, n_{10/11/10}+\phi_4'' n_{10/10/10}+\phi_5 n_{10/10/00}) & & + & \phi_2 n_{10/11/10}+\phi'{}_4\, n_{10/10/10}+\bar{\phi}_5\, n_{10/10/00}+\bar{\phi}_7\, n_{00/10/10}) \\
\hat{p}_{101} & = & \frac{1}{2n}(2n_{11/00/11}+n_{11/10/11}+n_{11/00/10}+n_{10/00/11} & \hat{p}_{001} & = & \frac{1}{2n}(2n_{00/00/11}+n_{10/00/11}+n_{00/10/11}+n_{00/00/10} \\
 & + & \bar{\phi}_1\, n_{11/10/10}+\bar{\phi}_3\, n_{10/10/11}+\phi'{}_4\, n_{10/10/10}+\phi_6 n_{10/00/10}) & & + & \phi_3 n_{10/10/11}+\phi''{}_4\, n_{10/10/10}+\bar{\phi}_6\, n_{10/00/10}+\bar{\phi}_7\, n_{00/10/10}) \\
\hat{p}_{100} & = & \frac{1}{2n}(2n_{11/00/00}+n_{11/10/00}+n_{11/00/10}+n_{10/00/00} & \hat{p}_{000} & = & \frac{1}{2n}(2n_{00/00/00}+n_{00/00/10}+n_{00/10/00}+n_{10/00/00} \\
 & + & \phi_1 n_{11/10/10}+\phi'''{}_4\, n_{10/10/10}+\bar{\phi}_5\, n_{10/10/00}+\bar{\phi}_6\, n_{10/00/10}) & & + & \phi_5 n_{00/10/10}+\phi_6 n_{10/00/10}+\phi_7 n_{10/10/00}+\phi_4 n_{10/10/10}).
\end{array}
\qquad (22)
$$

At the lower hierarchy of the EM framework, we calculate the posterior probabilities of a double or triple heterozygous individual i to be a particular diplotype ($A\bar{A}$) (E step), for which where [111] is assumed as the risk haplotype, expressed as

$$\begin{array}{lllll}\Omega_{11|i} & = & \frac{\phi_1 f_1(y_i)}{\phi_1 f_1(y_i)+\bar{\phi}_1 f_0(y_i)}, & \overline{\Omega}_{11|i}=1-\Omega_{11|i} & \text{for genotype } 11/10/10 \\ \Omega_{21|i} & = & \frac{\phi_2 f_1(y_i)}{\phi_2 f_1(y_i)+\bar{\phi}_2 f_0(y_i)}, & \overline{\Omega}_{21|i}=1-\Omega_{12|i} & \text{for genotype } 10/11/10 \\ \Omega_{31|i} & = & \frac{\phi_3 f_1(y_i)}{\phi_3 f_1(y_i)+\bar{\phi}_3 f_0(y_i)}, & \overline{\Omega}_{31|i}=1-\Omega_{13|i} & \text{for genotype } 10/10/11 \\ \Omega_{41|i} & = & \frac{\phi_4 f_1(y_i)}{\phi_4 f_1(y_i)+\bar{\phi}_4 f_0(y_i)}, & \overline{\Omega}_{41|i}=1-\Omega_{14|i} & \text{for genotype } 10/10/10 \end{array} \quad (23)$$

With the calculated posterior probabilities by the above equation (23), we then estimate the quantitative genetic parameters, Θ_q, based on the log-likelihood equations. These equations have similar, but more complicated, forms like equations (12) and (13).

Hypothesis tests can be made for linkage disequilibria among three SNPs and haplotype effects. Four different linkage disequilibria, D_{12}, D_{13}, D_{23} and D_{123}, that describe the linkage among three SNPs can each be tested using the null hypotheses described by equation (21). The log-likelihood ratios for each hypothesis are thought to follow a χ^2 distribution.

***R*-SNP model:** The idea for haplotyping a quantitative trait is described for two- and three-SNP models. It is possible that these models are too simple to characterize genetic variants for quantitative variation. With the analytical line for the two- and three-SNP sequencing model, a model can be developed to include an arbitrary number of SNPs whose sequences are associated with the phenotypic variation. A key issue for the multi-SNP sequencing model is how to distinguish among 2^{r-1} different diplotypes for the same genotype heterozygous at r loci. The relative frequencies of these diplotypes can be expressed in terms of haplotype frequencies. The integrative EM algorithm can be employed to estimate the MLEs of haplotype frequencies. A general formula for estimating haplotype frequencies can be derived.

RESULTS

The statistical model described above can be used to map and identify QTNs for a quantitative trait in an F_2 population. Because the marker data we have for mouse are microsatellites rather than SNPs, we use these microsatellite markers as a surrogate of SNPs for the purpose to demonstrate the utility of the model. Our marker data were from Vaughn et al.'s [21] study in which a linkage map composed of 19 chromosomes was constructed with 96 microsatellite markers for 502 F_2 mice (259 males and 243 females) derived from two strains, the Large (LG/J) and Small (SM/J). This map has a total map distance of ~1780 cM (in Haldane's units) and an average interval length of ~23 cM. The F_2 progeny was measured for their body mass at 10 weekly intervals starting at age 7 days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth, parity and sex [21]. Here, only adult body weights at week 10 are used for "QTN" analysis.

For each F_2 mouse, the parental origin of alleles at each marker can be discerned in molecular studies. Let L and S be the alleles inherited from the Large (LG/J) and Small (SM/J) strains, respectively. For any pair of markers, there are four different haplotypes, LL, LS, SL and SS, whose frequencies are accordingly denoted as

$$p_{LL}=p_{SS}=p$$

and

$$p_{LS}=p_{SL}=\frac{1}{2}-p.$$

By assuming all the four haplotypes as a risk haplotype, respectively, the above model allows for the estimates of haplotype frequencies by the EM iteration at the higher hierarchy and of composite genotypic values by the EM iteration at the lower hierarchy. The estimated haplotype frequencies are used to estimate linkage disequilibrium based on equation (14) and the recombination fraction (r) based on equation (4). This estimation process is moved from the first ($\mathbf{M}_1$–$\mathbf{M}_2$) to last pair of markers ($\mathbf{M}_6$–$\mathbf{M}_7$) on chromosome 1 and then from chromosome 1 to 19.

Table 3 tabulates the results of the MLEs of haplotype frequencies and log-likelihoods under the assumptions of different risk haplotypes. A total of 96 markers are sparsely located on 19 mouse chromosomes, with the estimated recombination fractions from the linkage disequilibrium model [8] consistent with those obtained from the linkage model [21]. Significant likelihood ratios for testing haplotype effects were determined by critical values obtained from the χ^2-square distribution with two degrees of freedom with a Bonferroni adjustment to the type I error. The adjusted critical values for the two- and three-marker QTN models are 18.20 and 18.76, respectively, at the 5% significance level. Significant haplotype effects are detected for a total of eight marker pairs (Table 3), which include one pair on chromosome 4, two consecutive pairs on chromosome 6, four consecutive pairs on chromosome 7 and one pair on chromosome 14. For some pairs, multiple significant risk haplotypes were detected. Risk haplotypes purely composed of alleles inherited from the LG/J or SM/J parent exert a positive or negative additive effect on body weight, respectively. Based on the relative values of estimated additive and dominant effects, the significant marker pairs detected display partial dominant effects (Table 3).

The results from the three-marker model are basically consistent with those from the two-marker model (Table 4). The advantage of the three-marker model is that it incorporates the interferences between adjacent marker intervals into the estimation process and, thus, can potentially increase the estimation precision of haplotype effects.

DISCUSSION

Quantitative trait locus (QTL) mapping aims to identify narrow chromosomal segments for a quantitative trait by using a statistical method, and has proven its value to study the genetic architecture of the trait in a variety of species [6–8]. The limitations of this technique lie in its inability to characterize the structure and organization of DNA sequences and statistical difficulty in deriving the distribution of test statistics under the null hypothesis of no QTL [22]. At least partly for these reasons, despite thousands of QTL reported for different traits and populations, a very small portion of them have been cloned [9]. With the completion of the genome projects for several important organisms, a new line of thought in the post genomic era has begun to emerge for the identification of specific combinations of nucleotides or haplotypes that contribute to a complex quantitative trait [13,23].

Theory and methods for haplotype discovery have well been established for natural populations [13] in which the non-random association among different single nucleotide polymorphims (SNP), specified by the coefficients of linkage disequilibria, lays a foundation for the mixture model of haplotyping a quantitative trait. In this article, we derived a statistical model for detecting haplotypes and estimating their effects on quantitative variation of

Table 3. The MLEs of haplotype frequencies and significant log-likelihood ratios (LR) by assuming different risk haplotypes in the F_2 population of mice.

Marker pair	Association		Selection of risk haplotype			Haplotype effect	
	$\hat{D}$	$\hat{r}$	Risk haplotype	Frequency	LR_2	$\hat{a}$	$\hat{d}$
D4Mit16-D4mMit13	0.16	0.17	*LL*	0.36	157.59	0.53	0.46
			LS	0.08	152.57	0.60	−0.91
			SL	0.10	155.55	−1.46	0.90
			SS	0.47	153.26	−0.35	0.18
D6Mit9-D6Nds5	0.18	0.14	*LL*	0.42	19.95	1.17	0.24
D6Nds5–D6Mit15	0.14	0.22	*LL*	0.38	25.14	1.25	0.44
			SS	0.41	37.98	−1.69	0.51
D7Mit21–D7Nds1	0.09	0.32	*LL*	0.34	30.84	0.93	1.41
			SS	0.34	36.54	−1.70	−0.07
D7Nds1–D7Mit17	0.19	0.12	*LL*	0.44	46.87	1.66	0.59
			SS	0.45	43.31	−1.75	0.50
D7Mit17–D7Mit9	0.19	0.12	*LL*	0.43	33.41	1.42	0.57
			SS	0.45	34.35	−1.47	0.99
D7Mit9–D7Nds4	0.12	0.26	*SS*	0.38	19.84	−1.15	1.09
D14Mit5–D14Mit7	0.17	0.16	*LL*	0.43	19.35	1.10	0.33

The results were obtained by using a two-SNP QTN model.

doi:10.1371/journal.pone.0000732.t003

a trait in experimental crosses. We used the principle of linkage disequilibrium analysis to characterize the linkage among different markers that is usually described by the recombination fractions in a commonly used F_2 population, initiated with two inbred lines. We established an interchangeable relationship between the linkage and linkage disequilibrium. The merit of this relationship in trait haplotyping includes the incorporation of interferences between adjacent marker intervals into the estimation and test of haplotype effects when multiple markers are modelled simultaneously.

The haplotyping model developed in this article was used to analyze a published F_2 population of mouse [21], but we used microsatellite markers as the surrogate of SNPs so that we can detect the effects of haplotypes constructed by microsatellite

Table 4. The MLEs of haplotype frequencies and significant log-likelihood ratios (LR) by assuming different risk haplotypes in the F_2 population of mice.

Marker pair	Selection of risk haplotype			Haplotype effect	
	Risk haplotype	Frequency	LR_2	$\hat{a}$	$\hat{d}$
D4Mit45–D4Mit16–D4Mit13	*LLL*	0.29	124.70	0.40	0.61
	LLS	0.07	121.34	1.08	−1.00
	LSL	0.01	122.58	−1.88	−2.56
	LSS	0.07	122.08	−0.44	1.40
	SLL	0.07	122.14	0.80	0.12
	SLS	0.01	132.86	-	-
	SSL	0.09	122.33	−1.32	1.09
	SSS	0.40	123.65	−0.55	0.28
D6Mit9–D6Nds5–D6Mit15	*SSS*	0.34	22.65	−1.51	0.39
D7Mit21–D7Nds1–D7Mit17	*LLL*	0.29	38.28	0.81	1.85
	SSS	0.30	33.74	−1.80	0.09
D7Nds1–D7Mit17–D7Mit9	*LLL*	0.38	34.39	1.48	0.47
	SSS	0.40	32.18	−1.61	0.61
D7Mit17–D7Mit9–D7Nds4	*LLL*	0.33	21.74	1.20	0.45
	SSS	0.33	29.41	−1.60	1.36
D14Nds1–D14Mit5–D14Mit7	*LLL*	0.30	19.55	1.44	−0.50

The results were obtained by using a three-SNP QTN model.

doi:10.1371/journal.pone.0000732.t004

alleles. The whole-genome of mouse was scanned for haplotype effects on body weight by a two- and multi-marker model, respectively. Consistent results were observed from the two models, which suggests that four regions in mouse chromosomes 4, 6, 7, and 14 contribute to variation in body weight. These findings are in a good agreement with those from traditional interval QTL mapping [21]. But our haplotype discovery is more informative in terms of the characterization of specific haplotype structure and organization responsible for trait variation.

We have proposed a new model for haplotyping a quantitative trait in the F_2 progeny population. The tenet of the model can be extended to haplotype a complicated trans-generational pedigree, founded with multiple original parents and involving individuals with different relatedness. The model can also be modified to dissect the epistatic effects of different genes [23] and the interaction of genes with environment. For these extensions, haplotype selection aimed to detect the risk haplotypes that are expressed differently from the others present many challenges, but is crucial for the facilitation of the process of detecting the association between haplotype diversity and phenotypic variation.

Our haplotyping model offers a powerful tool for positional cloning of QTL that are important for a complex trait. Flint et al. [9] reviewed the potential of currently available cloning strategies, such as probabilistic ancestral haplotype reconstruction, Yin-Yang crosses and in silico analysis of sequence variants, to identify genes that underlie QTL in rodents. Our model, in conjunction with these strategies, may open a new gateway for the illustration of a detailed picture of the genetic architecture for a complex trait.

ACKNOWLEDGMENTS

Author Contributions

Conceived and designed the experiments: RW. Performed the experiments: JC. Analyzed the data: WH SW TL. Wrote the paper: RW.

REFERENCES

1. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.
2. Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.
3. Jansen RC, Stam P (1994) High resolution mapping of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447–1455.
4. Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.
5. Kao C-H, Zeng Z-B, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152: 1203–1216.
6. Mackay TFC (2001) Quantitative trait loci in *Drosophila*. Nat Rev Genet 2: 11–20.
7. Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, et al. (2000) *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85–88.
8. Li CB, Zhou AL, Sang T (2006) Rice domestication by reducing shattering. Science 311: 1936–1939.
9. Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nat Rev Genet 6: 271–286.
10. Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.
11. Judson R, Stephens JC, Windemuth A (2000) The predictive power of haplotypes in clinical response. Pharmacogenomics 1: 15–26.
12. Bader JS (2001) The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2: 11–24.
13. Liu T, Johnson JA, Casella G, Wu RL (2004) Sequencing complex diseases with HapMap. Genetics 168: 503–511.
14. Lou X-Y, Casella G, Littell RC, Yang MKC, Wu RL (2003) A haplotype-based algorithm for multilocus linkage disequilibrium mapping of quantitative trait loci with epistasis in natural populations. Genetics 163: 1533–1548.
15. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294: 1719–1723.
16. Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci 99: 7335–7339.
17. Sebastiani P, Lazarus SW, Kunkel LM, Kohane IS, Ramoni M (2003) Minimal haplotype tagging. Proc Natl Acad Sci 100: 9900–9905.
18. Eyheramendy S, Marchini J, McVean G, Myers S, Donnelly P (2007) A model-based approach to capture genetic variation for future association studies. Genome Res 17: 88–95.
19. Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J Roy Stat Soc Ser B 44: 226–233.
20. Burnham KP, Andersson DR (1998) Model Selection and Inference. A Practical Information-Theoretic Approach. Springer: New York.
21. Vaughn TT, Pletscher LS, Peripato A, King-Ellison K, Adams E, et al. (1999) Mapping quantitative trait loci for murine growth - A closer look at genetic architecture. Genet Res 74: 313–322.
22. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
23. Lin M, Wu RL (2006) Detecting sequence-sequence interactions for complex diseases. Current Genomics 7: 59–72.