Haplotyping a quantitative trait with a high-density map in experimental crosses

Wei Hou  
*University of Florida*

John Stephen F. Yap  
*University of Florida*

Song Wu  
*University of Florida*

Tian Liu  
*University of Florida*

James M. Cheverud  
*Washington University School of Medicine in St. Louis*

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Part of the [Medicine and Health Sciences Commons](https://digitalcommons.wustl.edu/open_access_pubs)

**Recommended Citation**
[https://digitalcommons.wustl.edu/open_access_pubs/858](https://digitalcommons.wustl.edu/open_access_pubs/858)
Authors
Wei Hou, John Stephen F. Yap, Song Wu, Tian Liu, James M. Cheverud, and Rongling Wu
Haplotyping a Quantitative Trait with a High-Density Map in Experimental Crosses

Wei Hou¹, John Stephen F. Yap², Song Wu², Tian Liu², James M. Cheverud³, Rongling Wu²*

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 School of Medicine, 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 (≤ 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 F₂ 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.

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 statistic 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, Arabidopsis, 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

Copyright: © 2007 Hou et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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
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 F2, 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 F2 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 the 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 F2 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 diplotype effects on a quantitative trait. In the F2, 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 F2 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 r1,…,rL, 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[r1,…,rL], is decomposed into the following components:

\[
p[r_1, \ldots, r_L] = p_1 p_2 \cdots p_L
\]

\[
+ (-1)^{r_1} p_2 \cdots p_L D_{123} + \cdots
\]

\[
+ (-1)^{r_1} p_2 \cdots p_L D_{12L} + \cdots + (-1)^{r_2} p_1 \cdots p_{L-1} D_{123} + \cdots
\]

\[
+ \cdots + (-1)^{r_L} p_{1-L} D_{123} + \cdots
\]

\[
+ \cdots + (-1)^{r_1} p_2 \cdots p_L D_{1-L} + \cdots
\]

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

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

Figure 2. Diplopy configuration of a genotype for two hypothesized SNPs. doi:10.1371/journal.pone.0000732.g002
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 \([r^m_1 \ldots r^m_L] \) \([r^p_1' \ldots r^p_L']\) \((r^m_1, r^p_1' \ldots , r^m_L, r^p_L' = 1, 0)\) and \(3^L\) genotypes expressed as \(r^m_1' \ldots r^m_L' r^p_1' \ldots r^p_L'\) \((r^m_1' \ldots r^m_L' \geq r^p_1' \ldots r^p_L' = 1, 0)\). 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 \(P_{r^m_1 \ldots r^m_L} \) and \(P_{r^p_1' \ldots r^p_L'}\) to denote the diploype and genotype frequencies, respectively, and \(n_{r^m_1' \ldots r^m_L' r^p_1' \ldots r^p_L'}\) to denote genotype observation.

A special case: Two-point linkage disequilibrium: For two given SNPs (\( S_1 \) and \( S_2 \)), 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

\[
P_{11} = \frac{1}{4} - D, \quad P_{10} = \frac{1}{4} - D, \quad P_{01} = \frac{1}{4} + D, \quad P_{00} = \frac{1}{4} + D.
\]

(2)

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

\[
D = \frac{1}{4}(1 - 2r),
\]

or

\[
r = \frac{1}{2}(1 - 4D).
\]

A special case: Three-point linkage disequilibrium: For three given SNPs (\( S_1, S_2 \), and \( S_3 \)), 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

\[
p_{111} = \frac{1}{8} D_{32} + \frac{1}{2} D_{23} + \frac{1}{2} D_{12} + D_{123},
\]

\[
p_{110} = \frac{1}{8} D_{32} - \frac{1}{2} D_{23} + \frac{1}{2} D_{12} - D_{123},
\]

\[
p_{101} = \frac{1}{8} D_{32} + \frac{1}{2} D_{23} - \frac{1}{2} D_{12} - D_{123},
\]

\[
p_{100} = \frac{1}{8} D_{32} - \frac{1}{2} D_{23} - \frac{1}{2} D_{12} + D_{123},
\]

\[
p_{011} = \frac{1}{8} D_{32} + \frac{1}{2} D_{23} + \frac{1}{2} D_{12} - D_{123},
\]

\[
p_{010} = \frac{1}{8} D_{32} - \frac{1}{2} D_{23} - \frac{1}{2} D_{12} + D_{123},
\]

\[
p_{001} = \frac{1}{8} D_{32} - \frac{1}{2} D_{23} + \frac{1}{2} D_{12} + D_{123},
\]

\[
p_{000} = \frac{1}{8} D_{32} + \frac{1}{2} D_{23} - \frac{1}{2} D_{12} - D_{123}
\]

(5)

where \( D_{12}, D_{23} \), and \( D_{13} \) are the linkage disequilibria between SNP \( S_1 \) and \( S_2 \) between \( S_1 \) and \( S_3 \) and between \( S_2 \) and \( S_3 \), respectively, and \( D_{123} \) is the linkage disequilibrium among the three SNPs. The four disequilibrium coefficients can be estimated, by solving equation (5), as

\[
D_{12} = \frac{1}{4}(p_{111} + p_{110} + p_{001} + p_{000}) - (p_{011} + p_{100} + p_{011} + p_{010})
\]

\[
D_{23} = \frac{1}{4}(p_{111} + p_{101} + p_{100} + p_{010}) - (p_{110} + p_{000} + p_{011} + p_{010})
\]

(6)

\[
D_{13} = \frac{1}{4}(p_{111} + p_{101} + p_{100} + p_{010}) - (p_{110} + p_{000} + p_{011} + p_{010})
\]

\[
D_{123} = \frac{1}{4}(p_{111} + p_{100} + p_{101} + p_{010}) - (p_{110} + p_{001} + p_{011} + p_{010})
\]

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 chromated 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 diplotypes 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 \( \theta = (p_{11}, p_{10}, p_{01}, p_{00}) \). 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 \( F_2 \) were classified into 9 genotypes for the two SNPs, each genotype with observation generally expressed as \( n_{r^m_1' \ldots r^m_L' r^p_1' \ldots r^p_L'} ~ (r^m_1' \leq r^m_2' \leq \ldots \leq r^m_L' \geq r^p_1' \leq \ldots \leq r^p_L' = 1, 0) \). The frequency of each genotype can be expressed in terms of haplotype frequencies (Table 1). Considering a quantitative trait controlled by diploype (rather than genotype) diversity, the phenotypic value...
haplotype frequencies (
loss of generality, let \([11]\) be such a distinct haplotype, called risk haplotypes form three 
[01] and [00], are collectively expressed as 
haplotype likelihoods, one for haplotype frequencies (composite diplotypes and residual variance are arrayed by 
diplotypes, respectively (Table 1). The means for different 

\[ j \]

considered is compatible with subject 
\( u \)

where

\[ z \]

Assume that this QTN triggers an effect on the trait because at 
\( 10 \)

= 10 log (2 
\( r \)

is the indicator variable defined as 1 if a diplotype 
\( r \)

is expressed by a linear model, i.e.,

\[ P \]

f

\[ (\phi) \]

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

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

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 
\( L(\theta_j | M) \), whereas the estimates of diplotype genotypic means 
and residual variance are based on the log-likelihood function 
\( L(\theta_j | y, M) \). 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 (\( \phi \)) 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

\[ p_{11} = \frac{1}{2n}(2n_{11/11} + \phi n_{10/10} + n_{11/10} + n_{10/11}) \]

\[ p_{10} = \frac{1}{2n}(2n_{10/00} + n_{11/10} + (1 - \phi)n_{10/10} + n_{10/00}) \]

\[ p_{01} = \frac{1}{2n}(2n_{00/11} + n_{00/10} + (1 - \phi)n_{00/10} + n_{00/00}) \]

\[ p_{00} = \frac{1}{2n}(2n_{00/10} + \phi n_{00/10} + n_{00/00} + n_{00/00}) \]

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

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 \( \phi \) and 1-\( \phi \), expressed as

\[ \phi = \frac{p_{11}p_{00}}{p_{11}p_{00} + p_{10}p_{01}}, \] (9)

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

\[ y_i = \sum_{j=1}^{1} \sum_{k=0}^{1} \sum_{l=0}^{1} \zeta_{ij} \mu_{ikj} + \epsilon_i, \] (7)

where \( \zeta_{ij} \) is the indicator variable defined as 1 if a diplotype 
considered is compatible with subject \( i \) and 0 otherwise, \n\( \mu_{ikj} \) is the genotypic value for diplotype 
\( [j] [k] \), and \( \epsilon_i \) is the residual error distributed as \( N(0, \sigma^2) \).

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 
ahaplotype, 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 \( AA \), \( A \bar{A} \) 
and \( \bar{A}A \). Let \( \mu_2, \mu_1 \) and \( \mu_0 \) 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 \( \theta_j = (\mu_2, \mu_1, \mu_0, \sigma^2) \).

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

\[ \log L(\theta_j | S) = \sum_{i=1}^{n} \log f(y_i) \]

\[ + 2n_{11/11} \log p_{11} + n_{11/10} \log (2p_{11}p_{10}) + 2n_{10/11} \log p_{10} + n_{10/10} \log (2p_{10}p_{10}) + n_{11/11} \log (2p_{11}p_{11}) + \sum_{i=1}^{n} \log f(y_i) \] (8)

\[ \log L(\theta_j | y, S) = \sum_{i=1}^{n} \log f(y_i) \]

\[ + n_{11/11} \log p_{11} + n_{11/10} \log (2p_{11}p_{10}) + 2n_{10/11} \log p_{10} + n_{10/10} \log (2p_{10}p_{10}) + n_{11/11} \log (2p_{11}p_{11}) + \sum_{i=1}^{n} \log f(y_i) \] (8)

where \( f_j(y) \) is a normal distribution density function of composite 
diplototype \( j (j = 2, 1, 0) \), i.e.,
At a lower hierarchy of the EM algorithm, the E step is derived to calculate the posterior probability \( \Omega_{[1]}[00] \) of individual \( i \) with the double heterozygous genotype to be diplotype \([11][00]\) by

\[
\Omega_{[1]}[00] = \frac{\phi_{(11)}[00](y_i)}{\phi_{[1]}[00](y_i)+(1-\phi_{[10]}[00])}.
\]

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 \( \tilde{\mu}_d \) for each composite diplotype and the residual variance based on the calculated posterior probabilities by

\[
\tilde{\mu}_d = \sum_{i=1}^{n_{[11][1]}} y_i, \\
\tilde{\mu}_1 = \sum_{i=1}^{n_{[1]}} y_i + \sum_{i=1}^{n_{[10][0]}} \Omega_{[10][0]}(y_i), \\
\tilde{\mu}_0 = \sum_{i=1}^{n_{[0][0]}} y_i + \sum_{i=1}^{n_{[10][0]}} (1-\Omega_{[10][0]}(y_i)).
\]

\[
\tilde{a}_2 = \frac{1}{n} \left\{ \sum_{i=1}^{n_{[11][1]}} (y_i - \tilde{\mu}_2)^2 + \sum_{i=1}^{n_{[1]}} (y_i - \tilde{\mu}_1)^2 + \sum_{i=1}^{n_{[0][0]}} (y_i - \tilde{\mu}_0)^2 \right. \\
+ \left. \sum_{i=1}^{n_{[1][0]}} \Omega_{[10][0]}(y_i - \tilde{\mu}_2)^2 + (1-\Omega_{[10][0]})(y_i - \tilde{\mu}_0)^2 \right\},
\]

where

\[
\tilde{\mu} = \frac{1}{2} (\tilde{\mu}_2 + \tilde{\mu}_0), \\
\tilde{a} = \frac{1}{2} (\tilde{\mu}_2 - \tilde{\mu}_0), \\
\tilde{d} = \tilde{\mu}_1 + \frac{1}{2} (\tilde{\mu}_2 + \tilde{\mu}_0).
\]

**Model selection:** The likelihood \( L(\Theta_{p}, \Theta_{q} | y, S) \) is formulated by assuming that haplotype \([11][11]\) is a risk haplotype. However, a real risk haplotype is unknown from raw data \([y, S]\). 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) \( L_A(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | y, S) \) for \([11]\), (2) \( L_A(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | y, S) \) for \([10]\), (3) \( L_A(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | y, S) \) for \([01]\), and (4) \( L_A(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | y, S) \) for \([00]\). Under each possible risk haplotype, we estimate the individual genetic parameters \( \tilde{\Theta}_{q} \) for each composite diplotype. The optimal risk and non-risk haplotype combination will be selected from these combinations based on the likelihoods.

**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 four 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:

\[
H_0 : D = 0 \text{ vs. } H_1 : D \neq 0
\]

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

\[
LR_1 = -2[logL(p_{11} = p_{10} = p_{01} = p_{00} = \frac{1}{4}(S) - logL(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | S)],
\]

The LR test is considered to asymptotically follow a \( \chi^2 \) 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

\[
H_0 : \mu_i \equiv \mu, \mu \text{ vs. } H_1 : \text{at least one equality in } H_0 \text{ does not hold}, j = 2, 1.0
\]

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

\[
LR_2 = -2[logL(\tilde{\Theta}_{p}, \tilde{\mu} | y, S) - logL(\tilde{\Theta}_{p}, \tilde{\Theta}_{q} | y, S)],
\]
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 LE2 may asymptotically follow a $\chi^2$ distribution with two degrees of freedom, so that the threshold can be obtained from the $\chi^2$ distribution table.

**Haplotyping a trait with multiple SNPs**

**Haplotype structure:** The statistical method for QTN mapping is exemplified by a set of three SNPs, $S_1, S_2, S_3$, 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 Table 2. Some genotypes are consistent with diploypes, whereas the others that are heterozygous at two or more SNPs are not. Each double heterozygote contains two different diploypes. One triple heterozygote, i.e., 10/10/10, by assuming [111] as a risk haplotype (labelled by $A_{111}$), the relative frequencies of different diploypes for this triple heterozygote are a function of haplotype frequencies (Table 2).

In the $F_2$ population, there are 27 genotypes for the three SNPs. Let $N_{i1}, r_1, r_2, r_1, r_2, r_3 = 1, 0, 1, 0, 1, 0$ 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 genetic value of the trait for individual $i$ is expressed, at the diplotype level, as

$$y_i = \sum_{r_1} \sum_{r_2} \sum_{r_3} \sum_{f_1} \sum_{f_2} \sum_{f_3} \epsilon_i \hat{H}_{r_1 r_2 r_3 f_1 f_2 f_3} + \epsilon_i, \quad (19)$$

where $\epsilon_i$ is the indicator variable defined as 1 if a diploptye considered is compatible with subject $i$ and as 0 otherwise, $\hat{H}_{r_1 r_2 r_3 f_1 f_2 f_3}$ is the genetic value for diploype $[r_1 r_2 r_3 f_1 f_2 f_3]$, and $\epsilon_i$ is the residual error distributed as $N(0, \sigma^2)$. Note that $m_1$ and $p_1$ stand for the maternally and paternally derived alleles, respectively.

By assuming [111] as a risk haplotype labelled by $A_1$ and all the others as non-risk haplotypes labelled by $A_2$. Table 2 provides the formulation of genetic values for three composite diploypes, $\mu_k$ for $A_1A_1$, $\mu_k$ for $A_1A_2$ and $\mu_k$ for $A_2A_2$. The haplotype effect parameters and residual covariance matrix are arrayed by a quantitative genetic parameter vector $\Theta = [\mu_{11}, \mu_{12}, \mu_{22}, \sigma^2]$.

**Likelihoods and algorithms:** With the above notation, we construct two likelihoods, one for haplotype frequencies ($\Theta_1$) based on SNP data ($S$) and the other for quantitative genetic parameters ($\Theta_2$) based on haplotype frequencies ($\Theta_2$), phenotypic ($y$) and SNP data ($S$). They are, respectively, expressed as

$$\log L(\Theta_1|S) = \text{constant} + \sum_{i=1}^{n} \sum_{r_1} \sum_{r_2} \sum_{r_3} \sum_{f_1} \sum_{f_2} \sum_{f_3} \log f(y_i | H_{r_1 r_2 r_3 f_1 f_2 f_3})$$

where $\phi = (1 - \psi)$ are defined below, and $f_j$ ($y_j = 2, 1, 0$) is a normal distribution density function of composite diploype $j$.

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 diploype within double or triple heterogeneous genotypes (E step) by

$$\phi_1 = \frac{P_{11}P_{10} + P_{10}P_{00}}{P_{11}P_{10} + P_{10}P_{00}}, \quad \text{for genotype 11/10/10}$$

$$\phi_2 = \frac{P_{11}P_{10} + P_{10}P_{00}}{P_{11}P_{10} + P_{00}P_{10}}, \quad \text{for genotype 11/10/10}$$

$$\phi_3 = \frac{P_{11}P_{10} + P_{10}P_{00}}{P_{11}P_{10} + P_{00}P_{10}}, \quad \text{for genotype 10/10/10}$$

$$\phi_4 = \frac{P_{11}P_{10} + P_{10}P_{00} + P_{10}P_{00} + P_{00}P_{10}}{P_{11}P_{10} + P_{10}P_{00} + P_{00}P_{10} + P_{00}P_{00}}, \quad \text{for genotype 11/10/10}$$

$$\phi_5 = \frac{P_{11}P_{10} + P_{00}P_{10}}{P_{11}P_{10} + P_{00}P_{10}}, \quad \text{for genotype 10/10/10}$$

$$\phi_6 = \frac{P_{11}P_{10} + P_{10}P_{00} + P_{10}P_{00} + P_{00}P_{10}}{P_{11}P_{10} + P_{10}P_{00} + P_{00}P_{10} + P_{00}P_{00}}, \quad \text{for genotype 10/10/00}$$

$$\phi_7 = \frac{P_{11}P_{10} + P_{00}P_{10}}{P_{11}P_{10} + P_{00}P_{10}}, \quad \text{for genotype 00/10/10}$$

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 (A) are the non-risk haplotype).

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Configuration</th>
<th>Frequency</th>
<th>Relative frequency</th>
<th>Composite diplotype</th>
<th>Symbol</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/11/11</td>
<td>[111][111]</td>
<td>2p111</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>11/11/10</td>
<td>[111][110]</td>
<td>2p111p110</td>
<td>1</td>
<td>AÅ</td>
<td>µ2</td>
<td></td>
</tr>
<tr>
<td>11/11/00</td>
<td>[110][110]</td>
<td>2p110</td>
<td>1</td>
<td>AÅ</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>11/10/11</td>
<td>[111][101]</td>
<td>2p111p101</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>11/10/00</td>
<td>[110][100]</td>
<td>2p110</td>
<td>1</td>
<td>AÅ</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/00/11</td>
<td>[101][101]</td>
<td>2p101</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/00/10</td>
<td>[101][100]</td>
<td>2p101p100</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/00/00</td>
<td>[100][100]</td>
<td>2p100</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>10/11/11</td>
<td>[111][011]</td>
<td>2p111p011</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>10/11/00</td>
<td>[110][010]</td>
<td>2p110p010</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>10/10/11</td>
<td>[111][001]</td>
<td>2p111p001</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>10/10/00</td>
<td>[110][000]</td>
<td>2p110p000</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>10/00/11</td>
<td>[101][001]</td>
<td>2p101p001</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>10/00/10</td>
<td>[101][000]</td>
<td>2p101p000</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>10/00/00</td>
<td>[100][000]</td>
<td>2p100p000</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/11/10</td>
<td>[111][100]</td>
<td>2p111p100</td>
<td>1</td>
<td>AA</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>11/11/00</td>
<td>[110][100]</td>
<td>2p110p100</td>
<td>1</td>
<td>AÅ</td>
<td>µ1</td>
<td></td>
</tr>
<tr>
<td>11/00/11</td>
<td>[101][100]</td>
<td>2p101p100</td>
<td>1</td>
<td>AÅ</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/00/10</td>
<td>[101][100]</td>
<td>2p101p100</td>
<td>1</td>
<td>AÅ</td>
<td>µ0</td>
<td></td>
</tr>
<tr>
<td>11/00/00</td>
<td>[100][100]</td>
<td>2p100p100</td>
<td>1</td>
<td>AA</td>
<td>µ0</td>
<td></td>
</tr>
</tbody>
</table>

Calculation of frequencies for each genotype (n = 2)

\[
\begin{align*}
\hat{p}_{111} &= \frac{1}{2n} (2n_{111/111} + 2n_{111/110} + 2n_{111/101} + 2n_{111/100}) \\
\hat{p}_{110} &= \frac{1}{2n} (2n_{111/110} + 2n_{111/100} + 2n_{111/100} + 2n_{111/101}) \\
\hat{p}_{101} &= \frac{1}{2n} (2n_{111/101} + 2n_{111/100} + 2n_{111/100} + 2n_{111/101}) \\
\hat{p}_{100} &= \frac{1}{2n} (2n_{111/100} + 2n_{111/100} + 2n_{111/100} + 2n_{111/100}) \\
\hat{p}_{011} &= \frac{1}{2n} (2n_{110/110} + 2n_{110/101} + 2n_{110/100} + 2n_{110/100}) \\
\hat{p}_{010} &= \frac{1}{2n} (2n_{110/101} + 2n_{110/100} + 2n_{110/100} + 2n_{110/100}) \\
\hat{p}_{001} &= \frac{1}{2n} (2n_{110/100} + 2n_{110/100} + 2n_{110/100} + 2n_{110/100}) \\
\hat{p}_{000} &= \frac{1}{2n} (2n_{110/100} + 2n_{110/100} + 2n_{110/100} + 2n_{110/100})
\end{align*}
\]

\textbf{Note:} 10/10/00 = 11/11/11.
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 $\langle j, k \rangle$ (E step), for which where [111] is assumed as the risk haplotype, expressed as

$$
\begin{align*}
\Omega_{1j} &= \frac{g(j)}{g(j) + g(j')} = 1 - \Omega_{1j} \quad \text{for genotype 11/10/10} \\
\Omega_{2j} &= \frac{g(j)}{g(j) + g(j')} = 1 - \Omega_{2j} \quad \text{for genotype 10/11/10} \\
\Omega_{3j} &= \frac{g(j)}{g(j) + g(j')} = 1 - \Omega_{3j} \quad \text{for genotype 10/10/11} \\
\Omega_{4j} &= \frac{g(j)}{g(j) + g(j')} = 1 - \Omega_{4j} \quad \text{for genotype 10/10/10}
\end{align*}
$$

(23)

With the calculated posterior probabilities by the above equation (23), we then estimate the quantitative genetic parameters, $\Theta^{\theta}$, 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 $\chi^2$ 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 $2^n$ 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 F2 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 F2 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 $\sim 1780$ cM (in Haldane’s units) and an average interval length of $\sim 23$ cM. The F2 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 F2 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

$$P_{LS} = P_{SL} = 1 - p.$$
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 F2 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 F2 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 3. The MLEs of haplotype frequencies and significant log-likelihood ratios (LR) by assuming different risk haplotypes in the F2 population of mice.

<table>
<thead>
<tr>
<th>Marker pair</th>
<th>Association</th>
<th>Selection of risk haplotype</th>
<th>Haplotype effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{p}$</td>
<td>$\hat{r}$</td>
<td>Risk haplotype</td>
</tr>
<tr>
<td>D4Mit16-D4mMit13</td>
<td>0.16</td>
<td>0.17</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SS</td>
</tr>
<tr>
<td>D6Mit9-D6Nds5</td>
<td>0.18</td>
<td>0.14</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SL</td>
</tr>
<tr>
<td>D7Mit21–D7Nds1</td>
<td>0.09</td>
<td>0.32</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SL</td>
</tr>
<tr>
<td>D7Nds1–D7Mit17</td>
<td>0.19</td>
<td>0.12</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SL</td>
</tr>
<tr>
<td>D7Mit17–D7Mit9</td>
<td>0.19</td>
<td>0.12</td>
<td>LL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SL</td>
</tr>
<tr>
<td>D7Mit9–D7Nds4</td>
<td>0.12</td>
<td>0.26</td>
<td>LL</td>
</tr>
</tbody>
</table>

Table 4. The MLEs of haplotype frequencies and significant log-likelihood ratios (LR) by assuming different risk haplotypes in the F2 population of mice.

<table>
<thead>
<tr>
<th>Marker pair</th>
<th>Selection of risk haplotype</th>
<th>Haplotype effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk haplotype</td>
<td>Frequency</td>
</tr>
<tr>
<td>D4Mit10–D4Mit16–D4Mit13</td>
<td>LLL</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>LLS</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>LSL</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>LSS</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>SSL</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>SLS</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>SSL</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>SSS</td>
<td>0.40</td>
</tr>
<tr>
<td>D6Mit9–D6Nds5–D6Mit15</td>
<td>SSS</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>SSS</td>
<td>0.30</td>
</tr>
<tr>
<td>D7Mit21–D7Nds1–D7Mit17</td>
<td>LLL</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>SSS</td>
<td>0.30</td>
</tr>
<tr>
<td>D7Nds1–D7Mit17–D7Mit9</td>
<td>LLL</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>SSS</td>
<td>0.40</td>
</tr>
<tr>
<td>D7Mit9–D7Nds4</td>
<td>LLL</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>SSS</td>
<td>0.33</td>
</tr>
<tr>
<td>D14Nds1–D14Mit5–D14Mit7</td>
<td>LLL</td>
<td>0.30</td>
</tr>
</tbody>
</table>

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

The results were obtained by using a three-SNP QTN model.
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 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 F2 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