A General Probabilty Model for the Inheritance of Binary Traits

by

Katy Simonsen
Purdue University

Technical Report #03-04

Department of Statistics
Purdue University

2003

A general probability model for the inheritance of binary traits

Katy L. Simonsen

Department of Statistics, Purdue University, 150 N. University Ave., West Lafayette, IN 47907-2068; Ph. 765-494-6036, Fax 765-494-0558; simonsen@purdue.edu

Keywords: transmission genetics, penetrance, QTL, BTL

Abstract

In the era of genomics, large scale trait mapping studies are conducted. These experiments involve hundreds of markers and increasingly frequently they focus on complex traits. Complex traits are interesting because of the inherent multigenic nature of the traits, and the possible interaction between genes on the phenotype. Binary traits, such as disease resistance, have a simple phenotype but may also be multigenic. Likelihood and regression methods for modeling genotype and phenotypic relationships rely on underlying classical transmission genetics and the ability to implement complicated models. In principle, the expansion from one to two, three and more loci is understood as is the transmission probability from an F_2 to an F_3. What is missing is a comprehensive, flexible notation that allows for an arbitrary number of loci and the transmission for multiple generations. In effect, the integration of theory involving mapping functions with the theory of quantitative inheritance. We used classical genetics theory to construct a probability model for the joint distribution of marker genotypes, trait genotypes and trait values for any number of marker and trait loci. This framework is described in detail for a binary trait where the trait probabilities are parameterized in terms of a penetrance function. The transmission portion of the probability model is completely general and will apply regardless of the model employed for the trait values. We provide Maple code that generates the marginal distributions as well as the joint distribution. The explicit mathematical formulation described here will assist in implementing different models for the behavior of the complex trait and provides a common set of useful notation that can be used regardless of the trait model employed.

Introduction

The development of map functions (Haldane 1919, Kosambi 1944) and the importance of correctly modelling recombinational events have led to much interesting work on theories involving the recombination model itself (e.g. Housworth and Stahl, 2003). Map functions provide an important understanding about how adjacent maker loci behave and have been used as a jumping off point for the creation of a rich literature in the mapping of quantitative traits extending recently to sophisticated theory involving multiple loci, epistatic interactions, polypolids and binary traits. Applications in human genetics from full pedigree analyses to simple transmission disequilibrium tests have contributed to broad understanding in the inheritance and expression of phenotypeic data. With more than 62,000 articles published in the last 5 years on applications of QTL mapping methodology, this mature field is a rich resource for many scientists seeking to understand underlying genetic contributions to complex traits.

The proliferation of the methodology in this arena is at times dizzying. Notation abounds and the experimentalist seeking to make his or her way through the morass of options available is confronted with a plethora of terminology, and mathematical notation. In addition many of the examples are given for two or three loci with no general expansion. The mechanics of the theory upholding this vast literature are often invisible and for that reason it becomes difficult to compare methods.

Previous models have been constructed around the effects of the QTL leaving a morass of models and tests to navigate through. We propose a modular probability model that allows for separate discussion of the transmission genetics and the genetic models for effects. This modular approach allows for a clean separation of each of the underlying components and allows for direct comparisons between methodologies by directly comparing which element of the probability model is being changed.

The goal of this paper is to provide a general set of vocabulary that can be used to unify the current thinking about the probability models underlying much of the current statistical theory for detection of QTL. We show in detail how to construct a probability model and give an easy to implement algorithm for constructing a joint probability distribution for trait values, marker genotypes, and trait genotypes in the context of experimental crosses, for an arbitrary number of loci. This probability distribution is a necessary first step in the construction of tests for association between markers and phenotypic traits.

There has been a substantial amount of work done on the ordering of loci to make a genetic map. (Lander and Botstein 1989, Lander and Green 1987) For this reason we do not address the formation of the genetic map, but assume that the map has already been created. Once we have a map, then basic transmission genetic theory tells us how markers are transmitted from one generation to the next and this has been applied to develop a general matrix algebra application for understanding multi-generational transmissions. This recalls the landmark theory of Thompson (1982) and provides a direct link between human and experimental populations.

Our goal is not to defend any one of our particular assumptions but rather to point out exactly at what point specific assumptions enter and how in our modular approach it is easy to modify assumptions and explore the impact of that modification on the final joint probability. We focus on binary traits but also point how extensions to quantitative traits are easily achieved.

Model Framework

Our goal is to construct a general probability model for markers, trait genes, and trait values. In particular, we seek the joint probability distribution for these three quantities. The joint probability is the most general function; from this, any related quantities such as the joint probability of markers and trait values or the conditional probability of trait genes given markers are easily calculated. These probabilities can subsequently be used in inference procedures such as maximum likelihood calculations or Bayesian methods.

Assumptions

We construct the probability distributions with the following assumptions:

1. The population of interest is ultimately descended from an ancestral population (set of parents) with known genotype.
2. The mating system is known (e.g. backcross, F_2, F_3, ...)
3. Random gamete formation (No gametic selection or segregation distortion)
4. Mating is random within designated crosses (random union of gametes, no selection)
5. Markers are biallelic.
6. No interference in recombination.
7. Phase is known.
8. No mutation in markers or genes for the duration of the controlled crosses

In a typical application the original parents will be homozygous (RI lines), but this is not assumed, only that their marker types are known (with phase). We have developed the notation and the model with the understanding that it will be highly desirable to relax many of these assumptions. The model is derived in a modular way so that the point at which each assumption is used is clear. We illustrate the removal of an assumption in the section Relaxing Model Assumptions, where we show how to change the model when phase is unknown, eliminating assumption 7 above.

Notation and Preliminaries

The word diplotype will be used to indicate a double haplotype, that is, a genotype which consists of an ordered pair of haplotypes. A diplotype includes phase information and makes a distinction between the two possible heterozygotes (0/1 and 1/0) at a locus. It is easiest to construct the probability model when phase information is included. When phase information is not obtained in the data, the probabilities can be collapsed by summing over the unknown phase, as will be shown later. To avoid confusion, symbols which refer to phase-unknown genotypes will be designated with a prime ($'$), while diplotype and haplotype symbols will not be primed. For example, the random variable M refers to a marker diplotype, and M' refers to a marker genotype with phase unknown.

Denote by k the number of marker loci. Recombination rates between adjacent markers are denoted θ, and between markers and genes are denoted r. All matrices and vectors are indexed from 0 up to one less than their length, as is done in C/C++. While initially confusing to some, this greatly simplifies the binary index representation.

Genes and traits

The main random variables of interest are M, the marker diplotype, G, the trait gene (or "gene" diplotype), and Y, the trait value. One goal is to produce a joint probability distribution for these three random variables. This probability distribution will be denoted $\Pr(Y, M, G)$. Probability distributions for other combinations will be denoted similarly, for example, we use $\Pr(M)$ for the marginal distribution of M, and $\Pr(M, G)$ for the joint distribution of M and G. For expository purposes we will initially suppose Y is a binary trait, but will show later how to apply the technique to a quantitative trait. Additional models for the behaviour of Y can be easily incorporated into the proposed framework. Following the usual convention in statistics, we use upper case letters indicate random variables, while lower case letters indicate actual realizations or outcomes of these random variables, with subscripts on the lower case letters used to enumerate the different possibilities. For example, the probability that the random variable M takes on a specific value m_i is denoted $\Pr(M = m_i)$, where $i = 0 \ldots K\text{-}1$.

The probability distribution for a discrete random variable is a specification of the probability of every possible outcome. For our purposes, it will be convenient to use matrices or vectors to represent the probability distributions of our random variables. If the random variable M has a sample space of size K, that is, there are K possible outcomes, then $\Pr(M)$, the probability distribution for M, is represented as a vector of length K, where the i^{th} entry in the vector $\Pr(M)$ is the probability that M takes on the i^{th} possible outcome, that is, $\Pr(M = m_i)$. For this notation to be unambiguous it is necessary to define a fixed, canonical ordering for all the possible outcomes of each random variable. We define such an ordering below. The joint probability distribution of two random variables can similarly be described with a matrix. For example, the joint probability distribution of M and G, $\Pr(M, G)$, is represented with a matrix of size $K \times K$, whose i, j^{th} entry is the value $\Pr(M = m_i, G = g_j)$. Although markers and genes are in reality interspersed along the genome, we consider them as separate random variables; this is convenient since markers are observable and genes are not.

Several other random variables will be considered in the derivation. These are intermediate steps, and are used because their consideration simplifies the derivation of probabilities of M and G. These intermediate variables include N, the marker type of a haploid gamete, and H, the trait gene type of a haploid gamete. Specific values of these haplotypes will be denoted n_i and h_j, $(i, j = 0 \ldots L\text{-}1)$ respectively, according to the convention previously described. The relationship between the diplotypes M and G, and the haplotypes N and H will be clarified below. The other intermediate random variables that will be considered are called "transmission indicators", and will be defined in the next section. These are denoted T and S for markers and genes respectively, with particular outcomes t_u and s_v $(u, v = 0 \ldots L\text{-}1)$. As with M and G, we will need to define a canonical ordering for the possible outcomes for N, H, T, and S, so that we are able to describe their probability distributions with vectors and matrices. To assist the reader, a complete list of symbols and their meanings is provided in Box 1.

Transmission Indicators

A gamete haplotype is a random combination of its two parental haplotypes (parental diplotype). We label the two parental haplotypes "a" and "b", which could represent maternal and paternal origin. At each locus ℓ, it is possible that either the first ("*a*") or second ("*b*") parental haplotype transmits its genetic material to the gamete. We can therefore denote the transmission (or parental origin) for a single gamete, *T*, as a string of *k* 0's and 1's, in locus order, where a 0 in position ℓ indicates that the first parental haplotype ("*a*") was transmitted to the gamete at that locus, and 1 indicates that the second parental haplotype ("*b*") was transmitted to the gamete at that locus. For example, with $k = 3$ loci, a transmission indicator 011 indicates that the first locus in the gamete came from the parent's first haplotype and the second and third loci from the parent's second haplotype, so a recombination occurred between the first and second loci. A transmission of all 0's, i.e. $\underline{00\ldots00}$, indicates that the offspring gamete received an intact copy of the parent's first haplotype, and that no recombination occurred, whereas a transmission of all 1's indicates that the offspring gamete received an intact copy of the parent's second haplotype, again with no recombination. A transmission indicator that alternates 0's and 1's, $\underline{0101\ldots01}$ or $\underline{1010\ldots10}$ indicates that recombination occurred between every adjacent pair of loci. In general whenever the indicator at locus ℓ does not match the indicator at locus $\ell + 1$, then recombination occurred between those loci, whereas whenever they do match, recombination did not occur. The probability of any transmission indicator can be calculated in terms of recombination rates. A general formula for the probability of these events is given in a following section. As indicated previously, in order to write down probability distributions in terms of a vector, it is necessary to define a canonical ordering of all the possible outcomes. The transmission indicator is a string of *k* 0's and 1's and as such can be considered as an integer in base 2. This provides a natural ordering of the possible outcomes. There are $L = 2^k$ possible outcomes for the transmission indicator for *k* loci, so the transmission indicator *T* can take on the possible values $t_0, \ldots t_{L-1}$. The subscript on the *t* refers to its position in this ordering, and is equal to the decimal representation of the binary number. For example, with k = 3, the transmission indicator $\underline{011} = t_3$ since $3 = 0\times2^2 + 1\times2^1 + 1\times2^0$. More detail on this representation and ordering is given in the next section.

It is important to stress that the transmission indicator does not refer to the value of the actual allele of a gamete or its parents, only to the *parental origin* of the gamete at that locus, i.e. which parental chromosome transmitted its information to the offspring. Of course, the transmission indicator cannot unambiguously be determined at loci where parents are homozygous. However, inferring transmission indicators has been addressed in other literature and those methods could be easily incorporated in the framework we describe. For this work we consider all possible transmission indicators. When transmission is known then recombination probabilities are easy to calculate. When transmission is ambiguous at certain loci we simply sum over all the possibilities.

Binary index representation and ordering

The random variables *M* (marker diplotype) and *G* (trait gene diplotype) are discrete, each with $4^k = K$ possible outcomes (diplotypes), since at each locus there are 4 possibilities for the two chromosomal alleles *a*/*b*: 0/0, 0/1, 1/0, and 1/1. Since a diplotype is made up of two haplotypes, we can write a diplotype random variable (*M* or *G*) as an ordered pair of two haplotype random variables (*N* or *H*). Thus, we write $M \equiv [N_a, N_b]$ and $G \equiv [H_a, H_b]$ where the "*a*" subscript refers to the haplotype consisting of the chromosomes inherited from one parent (parent "*a*") and the "*b*" subscript refers to the haplotype consisting of the chromosomes inherited from the other parent (parent "*b*").

We define a canonical ordering of the possible diplotypes so that the probability distribution for each of these random variables can be unambiguously stated as a vector of length K, and the joint distribution of the two variables can be stated as a $K \times K$ matrix. Similarly, the random variables N (marker haplotype) and H (trait gene haplotype) are multinomial with L possible outcomes, and require a related ordering. We first define the ordering for N and H, and then use that ordering to define the ordering for M and G. It is the same as that used for transmission indicators. For example: with k biallelic loci (0 or 1), haplotypes (i.e. $N = n_i$ or $H = h_j$) can be written as a string of 0's and 1's, with the order corresponding to the order of the loci. These strings have a natural ordering given by counting in base 2. This ordering is $n_0 = \underline{00\cdots00} < n_1 = \underline{00\cdots01} < n_2 = \underline{00\cdots10} < \ldots < n_{L-2} = \underline{11\cdots10} < n_{L-1} = \underline{11\cdots11}$. By ordering in this way, the subscript is equal to the value of the diplotype when considered as a k-digit number in base 2. The outcomes $h_0 \ldots h_{L-1}$ appear identical to $n_0 \ldots n_{L-1}$, but they refer to the gene haplotypes rather than the marker haplotypes. The haplotype outcomes look identical to transmission indicators, but haplotypes refer to the actual allelic values of the gamete offspring, whereas transmission indicators refer to their parental origin.

The ordering of the diplotypes M and G are derived from the ordering of their components N and H respectively. A diplotype is made up of two haplotypes, one contributed from each parent. These are designated as $M = [N_a, N_b]$ and $G = [H_a, H_b]$. So each possible outcome m_i (or g_i) is made up of two components: $m_i = \left[n_{i_a}, n_{i_b}\right]$ $\left(g_j = \left[h_{j_a}, h_{j_b}\right]\right)$. The ordering of the diplotypes can thus be derived from the ordering of the haplotypes, if we define $i = i_a \times L + i_b$ (and $j = j_a \times L + j_b$). Thus $m_0 = [n_0, n_0] < m_1 = [n_0, n_1] < m_2 = [n_0, n_2] < \ldots < m_{L-1} = [n_0, n_{L-1}] < m_L = [n_1, n_0] < m_{L+1} = [n_1, n_1] < \ldots < m_{K-2} = [n_{L-1}, n_{L-2}] < m_{K-1} = [n_{L-1}, n_{L-1}]$, and similarly for g_i, $i = 0 \ldots K$-1. The subscript is equal to the value of the genotype when considered as a $2k$-digit number in base 2, and the "a" haplotype is placed to the left of the "b" haplotype. The transformation from i to i_a, i_b is unique (a bijection), since with $0 \leq i_a, i_b < L$, i_a and i_b are just the quotient and remainder, respectively, when i is divided by L.

For example, if $k = 3$, so that $L = 8$ and $K = 64$, the usual recombinant inbred parents would be P_0 with $M = m_0 = [n_0, n_0] = \underline{[000,000]}$ and $G = g_0 = \underline{[000,000]}$, and P_1 with $M = m_{63} = [n_7, n_7] = \underline{[111,111]}$ and $G = g_{63} = \underline{[111,111]}$. (Note that 63 = 7×8 + 7). The F_1 offspring of the cross $P_0 \times P_1$ would all be of the type $M = m_7 = [n_0, n_7] = \underline{[000,111]}$ (and $G = g_7$), while the F_1 offspring of the cross $P_1 \times P_0$ would all have $M = m_{56} = [n_7, n_0] = \underline{[111,000]}$ (and $G = g_{56}$). If both these crosses were done in equal numbers, then for the F_1 generation the probability distribution of M (or G) would be represented as a vector with ½ in positions 7 and 56, and 0 in all other positions from 0 to 63.

This particular choice of ordering facilitates the construction of diplotype from haplotype probabilities (shown later). While other orderings are possible and reasonable, the most important thing is to keep the orderings consistent throughout so that the appropriate products are formed when assembling the various desired distributions.

Model Derivation

We initially assume that the marker types of the original parents are known, with phase, and that their trait gene types can be labeled. The most typical application of this will be crosses of inbred lines, where one originating parent can be labeled "0" at all loci and the other originating parent can be labeled "1" at all loci. The progeny examined will typically be the result of a backcross, F_2, or subsequent design. The reader may wish to keep such an example in mind as the general derivation proceeds.

Road Map

We begin by calculating Pr(N) for a single gamete conditional on its parents. This will involve the probability distribution of the transmission indicator T, and the probability distribution of haplotypes N conditional on T. Then we combine the probabilities of its two component gametes N_a and N_b to form Pr(M). Next, we extend the method to derive the joint probability distribution of M and G. First, we calculate Pr(N,H) for a single gamete, conditional on its parents. This will involve the joint probability distribution of the two transmission indicators T (for markers) and S (for genes), and the probability distributions of haplotypes N conditional on T, and H conditional on S. Then we combine the probabilities of its two component gametes N_a,H_a and N_a,H_b to form Pr(M, G). Once Pr(M, G) is available, Pr(Y, M, G) is easy to obtain, from which we can easily derive any desired marginal or conditional probabilities such as Pr(Y, M) and Pr($G|M$). Then we will show how to apply the calculation of Pr(M, G) recursively over generations, so that arbitrary designs can be used.

Probability of Marker Genotypes: Pr(M)

We first consider the probabilities of marker haplotypes, and then combine them to construct diplotypes.

Marker Haplotypes

Suppose parent "a" in generation t has marker diplotype m_i. We wish to determine the probability that an offspring gamete in generation t+1 has marker haplotype n_j. Of interest is

$$\Pr\left(N\left(t+1\right)=n_j \mid M\left(t\right)=m_i\right),$$

where $M(t)$ is the parental marker diplotype and $N(t+1)$ is the gametic marker haplotype. (Generation indicators may be omitted when they are clear from context.)

Denote by T the transmission indicator random variable and by t_0, ... t_{L-1} its possible values. We drop the generation indicators for brevity, understanding that the M on which we condition is parental. Then

$$\begin{aligned}\Pr\left(N\left(t+1\right)=n_j \mid M\left(t\right)=m_i\right)&=\sum_{u=0}^{L-1}\Pr\left(N=n_j,T=t_u \mid M=m_i\right)\\&=\sum_{u=0}^{L-1}\Pr\left(N=n_j \mid M=m_i,T=t_u\right)\Pr\left(T=t_u \mid M=m_i\right)\\&=\sum_{u=0}^{L-1}\Pr\left(N=n_j \mid M=m_i,T=t_u\right)\Pr\left(T=t_u\right)\end{aligned} \tag{2.1}$$

The last line holds under the assumption (3 that markers are equally likely to be transmitted to gametes regardless of their actual allele values, that is, $\Pr\left(T=t_u \mid M=m_i\right)=\Pr\left(T=t_u\right)\forall m_i$ (no selection). We can easily construct an L-vector Θ whose entries are $\Pr(T = t_u)$ as follows. Let t_u^ℓ refer to the value of the transmission indicator t_u at locus ℓ, that is, the ℓ^{th} binary digit in the string representation of t_u, and let $I(u,\ell) = \begin{cases} 0 & t_u^\ell \neq t_u^{\ell+1} \\ 1 & t_u^\ell = t_u^{\ell+1} \end{cases}$ indicate whether (0) or not (1) recombination occurred between marker loci ℓ and $\ell + 1$ in transmission indicator t_u. Then

$$\Theta_u = \Pr(T = t_u) = \frac{1}{2}\prod_{\ell=1}^{k}(1-\theta_\ell)^{I(u,\ell)}\,\theta_\ell^{\,1-I(u,\ell)} \tag{2.2}$$

Note that the vector Θ does not depend on any genotypic values or the experimental design.

The transmission indicator t_u dictates exactly which allele is passed from m_i to n_j at each locus. Therefore, the expression $\Pr\left(N(t+1) = n_j \mid M(t) = m_i, T = t_u\right)$ is either 0 or 1 depending on whether n_j is the haplotype that results when m_i transmits its alleles according to t_u. We construct a matrix Δ whose j,u^{th} entry is that probability. Let $i = i_a L + i_b$ and let

$$\delta_{ju}^{\ell} = \begin{cases} 0, & t_u = 0, n_{i_a}^{\ell} \neq n_j^{\ell} \\ 0, & t_u = 1, n_{i_b}^{\ell} \neq n_j^{\ell} \\ 1, & t_u = 0, n_{i_a}^{\ell} = n_j^{\ell} \\ 1, & t_u = 1, n_{i_b}^{\ell} = n_j^{\ell} \end{cases} \tag{2.3}$$

denote whether (1) or not (0) the gamete allele at locus ℓ $\left(n_j^\ell\right)$ matches the transmitted parental allele at locus ℓ. Then let

$$\Delta_{ju}(m_i) = \prod_{\ell=1}^{k}\delta_{ju}^{\ell} \tag{2.4}$$

In other words, $\Delta_{ju}(m_i) = 1$ if at every locus, the gamete haplotype j matches the parental haplotype indicated by t_u, and $\Delta_{ju}(m_i) = 0$ if there is at least one locus where they do not match. Notice that for any given m_i, the matrix Δ has exactly one non-zero entry in each column, placing a 1 in the row corresponding to the gamete transmitted from m_i through the transmission represented by that column.

Then, according to these definitions, the value $\Pr\left(N(t+1) = n_j \mid M(t) = m_i\right)$ is the j^{th} entry of the L-vector $\Gamma(m_i) = \Delta(m_i) \times \Theta$, using ordinary matrix multiplication. This is true because

$$\begin{aligned}\Gamma(m_i)_j &= (\Delta(m_i)\times\Theta)_j = \sum_{u=0}^{L-1}\Delta(m_i)_{ju}\Theta_u \\ &= \sum_{u=0}^{L-1}\Pr(N=n_j \mid T=t_u, M=m_i)\Pr(T=t_u) \\ &= \Pr(N=n_j \mid M=m_i)\end{aligned} \tag{2.5}$$

For short, we write Pr(*N*) = Γ(*m_i*).

Marker Diplotypes

From the possible gametes n_j we can construct the marker diplotypes in generation *t*+1, by random union of gametes. The probability of a marker diplotype is simply the product of the probability of the two component gametes. Let $m_j = \left[n_{j_1}, n_{j_2}\right]$ where $j = j_1L + j_2$. Then

$$\begin{aligned}&\Pr\left(M(t+1)=m_j \mid M_a(t)=m_{i_a}, M_b(t)=m_{i_b}\right) \\ &= \Pr\left(N_a(t+1)=n_{j_1} \mid M_a(t)=m_{i_a}\right)\cdot\Pr\left(N_b(t+1)=n_{j_2} \mid M_b(t)=m_{i_b}\right) \\ &= \Gamma_{j_i}\left(m_{i_a}\right)\cdot\Gamma_{j_2}\left(m_{i_b}\right)\end{aligned} \tag{2.6}$$

Kronecker Product: The Kronecker product C of two matrices, A and B, where A is *p*×*q* and B is *r*×*s*, is a matrix of size *pr*×*qs* defined as follows: $C = A\otimes B$ has entries

$$\begin{aligned}&C_{ij} = A_{i_1j_1}B_{i_2j_2}, \text{ where } i = i_1r+i_2, \text{ and } j = j_1s+j_2 \\ &i_1 = 0\ldots p-1, i_2 = 0\ldots r-1, j_1 = 0\ldots q-1, j_2 = 0\ldots s-1. \\ &i = 0\ldots pr-1, j = 0\ldots qs-1\end{aligned} \tag{2.7}$$

In terms of block matrices, this is

$$A\otimes B = \begin{bmatrix} A_{00}B & \cdots & A_{0,q-1}B \\ \vdots & \ddots & \vdots \\ A_{p-1,0}B & \cdots & A_{p-1,q-1}B\end{bmatrix} \tag{2.8}$$

For the special case when A and B are both column vectors of length *L* ($q = s = 1, p = r = L$), $C = A\otimes B$ becomes a vector of length $L^2 = K$ with entries

$$\begin{aligned}&C_j = A_{j_1}B_{j_2}, \text{ where } j = j_1L + j_2 \\ &j_1, j_2 = 0\ldots L-1, j = 0\ldots K-1\end{aligned} \qquad (2.9),$$

or more explicitly,

$$A\otimes B = \left[A_0B_0, A_0B_1, \cdots A_0B_{L-1}, A_1B_0, \cdots A_1B_{L-1}, \cdots A_{L-1}B_0, \cdots, A_{L-1}B_{L-1}\right] \tag{2.10}$$

Using this definition we see that Pr(*M*) is constructed from the two *L*-vectors $\Gamma\left(m_{i_a}\right)$ and $\Gamma\left(m_{i_a}\right)$ via a Kronecker product, so that $\Pr\left(M\right)=\Gamma\left(m_{i_a}\right)\otimes\Gamma\left(m_{i_b}\right)$. Pr(*M*) is a *K*-vector valued function which takes the parental marker diplotypes m_{i_a} and m_{i_b} as input.

In summary, the *K*-vector Pr(*M*) gives the probabilities of marker diplotypes for offspring, given that the marker diplotypes for parents are m_{i_a} and m_{i_b}, according to the formula

$$\Pr\left(M\right)=\left(\Delta\left(m_{i_a}\right)\times\Theta\right)\otimes\left(\Delta\left(m_{i_b}\right)\times\Theta\right) \qquad (2.11)$$

The parental genotypes m_{i_a} and m_{i_b} in the previous generation are assumed known and fixed in the previous derivation. If the parental diplotypes are not fixed but rather are given by a probability distribution, then Pr(*M*) for the offspring is obtained by a weighted average of the values for all possible parental diplotypes.

Joint Probability of genotypes and markers: Pr(*M*, *G*)

For simplicity of notation, it is assumed that there is one potential gene to the right of each marker. The probability of recombination between marker ℓ and gene ℓ is r_ℓ, and the probability of recombination between marker ℓ and marker $\ell + 1$ is θ_ℓ. Since we assume no interference in recombination, the recombination probability between gene ℓ and marker $\ell + 1$ is therefore $q_\ell=\dfrac{\theta_\ell-r_\ell}{1-2r_\ell}$, where $r_\ell<\theta_\ell$ $\left(\text{since } 1-\theta=\left(1-r\right)\left(1-q\right)+rq\Rightarrow q=\dfrac{\theta-r}{1-2r}\right)$.

Of interest is Pr(*M*, *G*): the joint probability of a particular set of markers and a particular set of genes. This probability can be represented as a $4^k \times 4^k$ matrix where rows refer to markers and columns refer to genes. The i,j^{th} entry of the Pr(*M*, *G*) matrix is the joint probability that the marker type is m_i, and the gene type is g_j. It is also implicitly conditional on the parental marker and gene diplotypes, and the experimental design.

Haplotypes

For every marker locus ℓ there is a gene locus to its right separated from it by recombination rate $r_\ell < \theta_\ell$. A parent *a*, with marker diplotype $M_a\left(t\right)=\left[N_{a_1}\left(t\right),N_{a_2}\left(t\right)\right]$ and gene diplotype $G_a\left(t\right)=\left[H_{a_1}\left(t\right),H_{a_2}\left(t\right)\right]$ produces gametes with markers $N_a\left(t+1\right)$ and genes $H_a\left(t+1\right)$. (The derivation for parent *b* is identical.) We now consider the joint probability distribution for the random variables $N_a\left(t+1\right),H_a\left(t+1\right)\mid M_a\left(t\right),G_a\left(t\right)$. As in the derivation for Pr(*M*), we consider possible transmission indicators. In this case, however, we need to consider transmission for both markers and genes. Let *T* denote the transmission for markers as before, and *S* denote the transmission for genes (these are not independent). We drop the generation notation and the "*a*" subscript for brevity.

$$
\begin{aligned}
&\Pr\left(N_a(t+1)=n_i, H_a(t+1)=h_j \mid M_a(t)=m_{ia}, G_a(t)=g_{ja}\right)\\
&=\sum_{u=0}^{L-1}\sum_{v=0}^{L-1}\Pr\left(N=n_i, H=h_j, T=t_u, S=s_v \mid M=m_{ia}, G=g_{ja}\right)\\
&=\sum_{u=0}^{L-1}\sum_{v=0}^{L-1}\Pr\left(N=n_i, H=h_j \mid T=t_u, S=s_v, M=m_{ia}, G_a=g_{ja}\right)\Pr\left(T=t_u, S=s_v \mid M_a=m_{ia}, G_a=g_{ja}\right)\\
&=\sum_{u=0}^{L-1}\sum_{v=0}^{L-1}\Pr\left(N=n_i \mid T=t_u, M=m_{ia}\right)\Pr\left(H=h_j \mid S=s_v, G=g_{ja}\right)\Pr\left(T=t_u, S=s_v\right)
\end{aligned}
\tag{2.12}
$$

The last line holds because conditional on transmission indicators, markers and genes are independent, and transmission is independent of the parents' alleles. The components $\Pr\left(H_a=h_j \mid S=s_v, G_a=g_{j_a}\right)$ and $\Pr\left(N_a=n_i \mid T=t_u, M_a=m_{i_a}\right)$ are simply $\Delta\left(g_{j_a}\right)_{jv}$ and $\Delta\left(m_{i_a}\right)_{iu}$ respectively, as defined in the derivation of Pr(*M*). The component $\Pr\left(T=t_u, S=s_v\right)$ depends only on the recombination rates, and is derived next.

Let $R(u,v,\ell)=\begin{cases}0, & s_v^\ell \neq t_u^\ell\\ 1, & s_v^\ell = t_u^\ell\end{cases}$ and $Q(u,v,\ell)=\begin{cases}0, & s_v^\ell \neq t_u^{\ell+1}\\ 1, & s_v^\ell = t_u^{\ell+1}\end{cases}$, indicate whether (0) or not (1) recombination occurred between marker ℓ and gene ℓ, or gene ℓ and marker $\ell+1$, respectively, according to the transmission indicators t_u (markers) and s_v (genes). Then let

$$
\begin{aligned}
\Omega_{uv} &= \Pr\left(T=t_u, S=s_v\right)\\
&=\frac{1}{2}\left(\prod_{\ell=1}^{k}\left(1-r_\ell\right)^{R(u,v,\ell)} r_\ell^{\left(1-R(u,v,\ell)\right)}\right)\left(\prod_{\ell=1}^{k-1}\left(1-q_\ell\right)^{Q(u,v,\ell)} q_\ell^{\left(1-Q(u,v,\ell)\right)}\right),
\end{aligned}
\tag{2.13}
$$

where $q_\ell = \dfrac{\theta_\ell - r_\ell}{1-2r_\ell}$

define a matrix with the necessary components; that is, the joint probability that markers have transmission indicator $T = t_u$ and genes have transmission indicator $S = s_v$. (Ω is independent of all genotypes and the experimental design, depending only on the recombination parameters r and θ. Notice also that the sum over columns of Ω is Θ, i.e. Θ is the marginal.)

Thus, we can define

$$
\begin{aligned}
\Lambda\left(m_{i_a}, g_{j_a}\right) &= \Pr\left(N, H \mid M=m_{i_a}, G=g_{j_a}\right)\\
&=\Delta\left(m_{i_a}\right)\times\Omega\times\left(\Delta\left(g_{j_a}\right)\right)^{\dagger}
\end{aligned}
\tag{2.14}
$$

since the i,jth entry of that matrix is

$$\begin{aligned}\Lambda\left(m_{i_a},g_{j_a}\right)_{ij} &= \left(\Delta\left(m_{i_a}\right)\times\Omega\times\left(\Delta\left(g_{j_a}\right)\right)^{\dagger}\right)_{ij}\\ &= \sum_{u=0}^{L-1}\Delta\left(m_{i_a}\right)_{iu}\left(\sum_{v=0}^{L-1}\Omega_{uv}\Delta\left(g_{j_a}\right)_{jv}\right)\\ &= \sum_{u=0}^{L-1}\sum_{v=0}^{L-1}\Delta\left(m_{i_a}\right)_{iu}\Omega_{uv}\Delta\left(g_{j_a}\right)_{jv}\\ &= \sum_{u=0}^{L-1}\sum_{v=0}^{L-1}\Pr\left(N=n_i \mid T=t_u, M=m_{i_a}\right)\Pr\left(T=t_u, S=s_v\right)\Pr\left(H=h_j \mid S=s_v, G=g_{j_a}\right)\\ &= \Pr\left(N_a=n_i, H_a=h_j \mid M_a=m_{i_a}, G_a=g_{j_a}\right)\end{aligned}$$

(2.15)

For short, we write $\Pr(N,H)=\Lambda\left(m_{i_a},g_{i_a}\right)$.

Diplotypes

To obtain the probability for diplotypes, we multiply the probabilities for the corresponding haplotypes, again assuming random union of gametes.

$$\begin{aligned}&\Pr\left(M(t+1)=m_i, G(t+1)=g_j \mid G_a(t)=g_{j_a}, M_a(t)=m_{i_a}, G_b(t)=g_{j_b}, M_b(t)=m_{i_b}\right)\\ &=\Pr\left(N_a=n_{i_1}, H_a=h_{j_1} \mid G_a=g_{j_a}, M_a=m_{i_a}\right)\times\Pr\left(N_b=n_{i_2}, H_b=h_{j_2} \mid G_b(t)=g_{j_b}, M_b(t)=m_{i_b}\right) \quad (2.16)\\ &=\Lambda\left(m_{i_a},n_{j_a}\right)_{i_1 j_1}\times\Lambda\left(m_{i_b},n_{j_b}\right)_{i_2 j_2}\text{, where } \mathrm{i}=i_1L+i_2 \text{ and } j=j_1L+j_2\end{aligned}$$

Therefore, the $K \times K$ matrix Pr(M, G) is the Kronecker product of the haplotype matrices from the two parents:

$$\begin{aligned}\Pr(M,G) &= \Pr\left(N_a,H_a\right)\otimes\Pr\left(N_b,H_b\right)\\ &=\Lambda\left(m_{i_a},g_{j_a}\right)\otimes\Lambda\left(m_{i_b},g_{j_b}\right) \qquad (2.17)\\ &=\left(\Delta\left(m_{ia}\right)\times\Omega\times\left(\Delta\left(g_{ja}\right)\right)^{\dagger}\right)\otimes\left(\Delta\left(m_{i_b}\right)\times\Omega\times\left(\Delta\left(g_{j_b}\right)\right)^{\dagger}\right)\end{aligned}$$

Penetrance: p

A binary trait Y is a Bernoulli random variable that can take on only two possible values Y, denoted 0 and 1. The penetrance parameter is the probability that $Y = 1$ for each trait gene diplotype. Since there are K possible genotypes, the penetrance parameter **p** is a K-vector of probabilities, where the j^{th} element of **p** is $\Pr\left(Y=1 \mid G=g_j\right)$. The most general genetic model allows a different penetrance parameter for each genotype. Constraints among the elements of **p** indicate particular genetic models. In particular, consider four genotypes $g_{j_1}\ldots g_{j_4}$ that are identical except at a specific locus ℓ, and which at locus ℓ have the four possible types 0/0, 0/1, 1/0, and 1/1. If the elements of **p** are equal for these four genotypes, so that $\pi_{j_1} = \pi_{j_2} = \pi_{j_3} = \pi_{j_4}$ depends only on the genotypes at the other loci, then that locus has no effect on the trait.

If the penetrance does not depend on which parent contributed the genotype, then the genotypes with 0/1 or 1/0 at a particular locus will always have the same penetrance parameter. Such a constraint reduces the number of free parameters in **p** from 4^k to 3^k. Since this case is expected to be particularly common, we define a constrained penetrance $\mathbf{p}'$, a 3^k vector whose entries are penetrances for phase-unknown genotypes, in the order described in the section on phase. Further assumptions about the genetic model (*e.g.* dominance, additivity) can further reduce the dimension of the penetrance, and thus the number of free parameters to be estimated.

Joint Probability: Pr(*Y*, *M*, *G*) and Pr(Y, M)

We now have the building blocks by which to construct the joint probability distribution of *Y*, *G*, and *M*. We seek the probability

$$\begin{aligned}
&\Pr\left(Y=1, M=m_i, G=g_j\right) \\
&=\Pr\left(Y=1 \mid M=m_i, G=g_j\right)\Pr\left(M=m_i, G=g_j\right) \\
&=\Pr\left(Y=1 \mid G=g_j\right)\Pr\left(M=m_i, G=g_j\right) \\
&=\left[\Pr(M,G)\right]_{ij} p_j
\end{aligned} \qquad (2.18)$$

This can be represented as a *K*×*K* matrix whose *i,j*th entry is the above quantity:

$$\Pr(Y=1, M, G) = \Pr(M,G) \times \operatorname{diag}(\mathbf{p}) \qquad (2.19)$$

Summing over all possible values of *G* gives the joint probability of the two observable random variables *Y* and *M*:

$$\begin{aligned}
\Pr(Y=1, M) &= \sum_{j=0}^{K-1} \Pr\left(Y=1, M, G=g_j\right) \\
&= \Pr(M,G) \times \mathbf{p}
\end{aligned} \qquad (2.20)$$

where **p** is a column vector, and the matrix multiplication sums over the possible (unknown) genotypes. This probability could be used in a likelihood calculation.

Chaining Generations Together

The derivation above assumes that the genotypes and marker types in the previous generation were fixed and known. However, the results easily generalize to the case where the genotypes and marker types are described by a probability distribution. In this way, it is possible to "chain" the generations together to create the distribution for specific experimental designs. If the parental genotypes in generation t are described by the probability matrices $A = \Pr\left(M_a(t), G_a(t)\right)$ and $B = \Pr\left(M_b(t), G_b(t)\right)$ whose elements are

$$\begin{aligned} A_{i_a,j_a} &= \Pr\left(M_a(t) = m_{i_a}, G_a(t) = g_{j_a}\right) \\ B_{i_b,j_b} &= \Pr\left(M_b(t) = m_{i_b}, G_b(t) = g_{j_b}\right) \end{aligned} \qquad (2.21)$$

then probabilities in generation $t + 1$ are given by

$$\begin{aligned} &\Pr\left(M(t+1) = m_i, G(t+1) = g_j\right) \\ &= \sum_{i_a,j_a=0}^{K-1} \sum_{i_b,j_b=0}^{K-1} A_{i_a j_a} B_{i_b j_b} \Pr\left(M(t+1) = m_i, G(t+1) = g_j \middle| \begin{matrix} M_a(t) = m_{i_a}, G_a(t) = g_{j_a}, \\ M_b(t) = m_{i_b}, G_b(t) = g_{j_b} \end{matrix} \right) (2.22) \\ &\sum_{i_a,j_a=0}^{K-1} \sum_{i_b,j_b=0}^{K-1} A_{i_a j_a} B_{i_b j_b} \left(\Lambda\left(m_{i_a}, g_{j_a}\right) \otimes \Lambda\left(m_{i_b}, g_{j_b}\right)\right)_{i,j} \end{aligned}$$

Since the matrices A and B are the result of calls to the formula for Pr(M,G) based on known parents, for complex designs it is possible to start with known parents and repeatedly calculate Pr(M,G), combining results with the above formula. We give examples of such calculations in a following section.

Relaxing Model Assumptions

We made several assumptions in constructing our model. In this section we consider how the model makes use of these assumptions, pointing out at which point each assumption is used. We briefly consider the consequences to the model of modifying each assumption, and, where appropriate, indicate

Phase Unknown

In this section we show how the probabilities calculated for diplotypes with phase known can be combined to form probabilities for genotypes with phase unknown. Our initial derivation distinguished between genotypes of different phase, that is, considered separately the two possible heterozygotes 0/1 and 1/0 at each locus. Since this derivation is about probability distributions and not statistical inference, it does not actually require that which parent contributed each gamete can be determined unambiguously, (or that the phase is known) for any particular dataset. Here's an analogy to clarify the situation: consider repeated Bernoulli trials. In order to discover the probability distribution (binomial) for the number of successes, it is necessary to distinguish, and thereby count, all possible different orderings of successes and failures. If we ignore the fact that SSF is conceptually distinct from SFS and FSS, we will miscount the possible outcomes, and come up with the wrong probability distribution. However, once the binomial distribution has been derived, we can collapse these cases together as ``two successes". If we have an actual data set then we are able to estimate the parameter just from the total number of successes, even if for some reason it is not possible to observe the ordering. In a similar way, our derivation of the probability distribution has treated the case where 0 is maternally inherited and 1 paternally inherited as distinct from the reverse. However, since in practice it is often not possible to observe phase in actual data, in preparation for a future statistical inference step we show how to reduce the phase-known case to the phase-unknown case by the use of a phase matrix *C*, a 3^k x 4^k matrix. This involves grouping together indistinguishable genotypes and adding their probabilities.

Phase Matrix

Sometimes certain diplotypes are indistinguishable in practice (heterozygotes 0/1 vs. 1/0 at same locus). This reduces the sample space from a 4^k space of diplotypes to a 3^k space of phase-unknown genotypes. Before showing how to convert between these spaces, we must define an ordering on this 3^k space.

The genotype at a single locus can take one of three different values, which we label by the total number of "1" alleles at that locus. The diplotype 0/0 is labeled 0, the diplotypes 0/1 and 1/0 (the two heterozygotes) are both labeled 1, and the diplotype 1/1 is labeled 2. The phase-unknown genotype for all loci is thus a string of *k* 0's, 1's, and 2's. This corresponds to a *k*-digit number in base 3, which has 3^k possible values. We use the natural ordering of numbers in base 3 to define a canonical ordering of phase-unknown genotypes, just as was done with haplotypes in base 2. Thus the marker genotypes are ordered as $m_0' = \underline{0\cdots00} < m_1' = \underline{0\cdots01} < m_2' = \underline{0\cdots02} < m_3' = \underline{0\cdots10} < \ldots < m_{3^k-1}' = \underline{2\ldots22}$

Now we need to map the existing phase-known probabilities based on diplotypes into the phase-unknown setting. This is accomplished by the left and right-multiplication of an appropriate matrix, which combines the appropriate rows and columns of a probability matrix. This phase matrix, denoted C_k, indicates which rows and columns are to be combined. It is a $3^k \times 4^k$ matrix and operates on 4^k column vectors whose rows are indexed by phase-known genotypes. Therefore, if G' and M' indicate the phase-unknown versions of *G* and *M*, we have

$$\Pr(M',G') = C_k \times \Pr(M,G) \times C_k^\dagger \qquad (3.1)$$

C_k can similarly be used to collapse a probability column vector with a single multiplication.

$$\Pr(M') = C_k \Pr(M) \qquad (3.2)$$

A conditional probability matrix (Pr(*G*|*M*)) is row normalized (entries in each row sum to 1). In order to maintain the row-normalization, the left multiplication must in this case be by a row-normalized version of C_k, which we denote C_k^*. (Note that $C_k^* C^\dagger = I$) Thus

$$\Pr(G' \mid M') = C_k^* \times \Pr(G \mid M) \times C_k^\dagger \qquad (3.3)$$

Examples:

For $k = 1$:

$$C_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad C_1^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

This corresponds to the two heterozygotes being indistinguishable.

For $k = 2$:

$$C_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The outlined row 4 shows, for example, that the four different diplotypes [$\underline{00}$,$\underline{11}$], [$\underline{01}$,$\underline{10}$], [$\underline{10}$,$\underline{01}$], [$\underline{11}$,$\underline{00}$] (columns 3, 6, 9, and 12) all correspond to the same phase-unknown genotype, namely the base three number $\underline{11}$ (row 4 since in base 3, 11 = 1×3 + 1 = 4), that is, heterozygous at both loci.

General Formula for C_k:

The i, j^{th} entry of the matrix C_k is given by

$$\left(C_k\right)_{ij} = \begin{cases} 1, & j = \sum_{\ell=1}^{k} \left(i_1^\ell + i_2^\ell\right) \times 3^{k-\ell} \\ 0, & \text{otherwise} \end{cases}, \text{ where } i = i_1 L + i_2 \text{ and } i_a^\ell \text{ is the allele at locus } \ell \qquad (3.4)$$

That is, each column i (representing a diplotype) has a single non-zero entry in the row corresponding to its phase-unknown genotype. Correspondingly, each column j (representing a genotype) has one or more non-zero entries corresponding to its possible diplotypes.

Probabilities when phase is unknown

Summary of how to do it.

Other Assumptions

Assumption (1) of known ancestors can only be relaxed to the extent that a probability distribution for the ancestral generation can be used instead of the exact ancestral genotypes. If no ancestral information is known, the probabilities cannot be chained forward. Likewise assumption (2) of known mating system is necessary, although a wide variety of models can be used (e.g. selfing).

The assumption (3) of no gametic selection is used to calculate the probabilities of gamete formation. In particular, in (2.12), it is used to equate $\Pr\left(T=t_u, S=s_v \mid M_a=m_{ia}, G_a=g_{ja}\right)$ and $\Pr\left(T=t_u, S=s_v\right)$. If selection causes some alleles to be more frequently transmitted, this equation would have to be modified. The assumption (4) of random union of gametes is used in (2.17), where the probability of a diplotype was the product of its haplotype probabilities. This equation would have to be modified if that assumption were to be relaxed.

The assumption of biallelic markers (5) is central to our notation and ordering for markers and diplotypes, where the number of alleles dictates the size of the arrays. This assumption is justified in any of the standard crosses of inbred lines. If the parents are instead known to be multiallelic, this would mean working in that base, rather than in binary. In principle, this is a straightforward extension of the present work, although one could imagine that if the number of alleles varied from marker to marker the bookkeeping would be difficult.

Interference models of recombination eliminate the assumption (6) that recombination events at linked loci are independent. This assumption was used in the formula (2.13) for $\Omega_{uv}=\Pr\left(T=t_u, S=s_v\right)$, in which transmission probabilities are calculated as a product across loci. This formula would change in an interference model; however the resulting Ω could be used in the same way.

The assumption of no mutation (8) is used to define $\Delta_{ju}=\Pr\left(N(t+1)=n_j \mid M(t)=m_i, T=t_u\right)$ in (2.3) and (2.4) as either 0 or 1 depending on whether the parent's alleles transmitted via transmission indicator t_u give the gamete's alleles. If mutation were allowed then in this matrix each 0 entry would be replaced by the probability of achieving that gamete through mutation, while each 1 entry would be replaced by the probability of no mutation. For example, a simple mutation model could consider a probability of mutation p at each locus. Then the 1's would be replaced by $(1-p)^k$ and the 0's replaced by p raised to the power of the number of mutations needed to achieve the corresponding gamete. An even simpler model (an approximation for small p) would only permit those gametes achievable by a single mutation, so that the k gametes "one step" away would have probability p while the "no mutation" gamete would have probability $1-kp$.

Specific Crosses of Inbred Lines: Examples

In the traditional cross of inbred lines, Parent$_0$ is defined (at $t=0$) to have $G(0)=g_0$ and $M(0)=m_0$, while Parent$_1$ is defined to have $G(0)=g_{K-1}$ and $M(0)=m_{K-1}$. The F_1 offspring of the cross where parent a = Parent$_0$ and parent b = Parent$_1$ has $G(1)=\left[h_0, h_{L-1}\right]=g_{L-1}$, and $M(1)=\left[n_0, n_{L-1}\right]=m_{L-1}$. (The reciprocal cross with a = Parent$_1$ and b = Parent$_0$ would produce F_1 offspring of type $G(1)=\left[h_{L-1}, h_0\right]=g_{K-L}$, and $M(1)=\left[n_{L-1}, n_0\right]=m_{K-L}$).

For the backcross to parent a = Parent$_0$ and the F_1 as above, probabilities are obtained using the formulas for Pr(Y, $G(2)$, $M(2)$) with the a and b parents as $G_a(1)=g_0$, $M_a(1)=m_0$, $G_b(1)=g_L$, $M_b(1)=m_{L-1}$.

Probabilities for the F_2 population are obtained similarly, with both a and b parents (F_1) as either $G(1)=[h_0,h_{L-1}]=g_{L-1}$, and $M(1)=[n_0,n_{L-1}]=m_{L-1}$ or $G(1)=[h_{L-1},h_0]=g_{K-L}$, and $M(1)=[n_{L-1},n_0]=m_{K-L}$, as appropriate.

Next we give some examples of the probability matricies for small values of k.

1 gene 1 marker ($k = 1$)

The following are Pr(M, G) for several common breeding schemes, as derived by the above formulas.

$$\text{Parent}_0:\Pr(G,M)=\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix},\quad \text{Parent}_1:\Pr(G,M)=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix} \tag{3.5}$$

$$F_1:\Pr(G,M)=\begin{bmatrix}0&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\text{ or }\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}\text{or}\begin{bmatrix}0&0&0&0\\0&\frac{1}{2}&0&0\\0&0&\frac{1}{2}&0\\0&0&0&0\end{bmatrix} \tag{3.6}$$

Backcross Parent$_0$ x F$_1$:

$$\Pr(G,M)=\begin{bmatrix}\frac{1}{2}(1-r_1)&\frac{1}{2}r&0&0\\\frac{1}{2}r_1&\frac{1}{2}(1-r_1)&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix} \tag{3.7}$$

Next we give the matrix Pr(M, G) for F$_2$, …, F$_6$. In each case the matrix has the same overall structure, but with different entries as shown.

$$\Pr(M,G)=\begin{bmatrix}p_0&p_1&p_1&p_2\\p_1&p_0&p_2&p_1\\p_1&p_2&p_0&p_1\\p_2&p_1&p_1&p_0\end{bmatrix} \tag{3.8}$$

$$F_2:\Pr(G,M)=\begin{bmatrix}\frac{1}{4}(1-r_1)^2 & \frac{1}{4}r_1(1-r_1) & \frac{1}{4}r_1(1-r_1) & \frac{1}{4}r_1^2\\ \frac{1}{4}r_1(1-r_1) & \frac{1}{4}(1-r_1)^2 & \frac{1}{4}r_1^2 & \frac{1}{4}r_1(1-r_1)\\ \frac{1}{4}r_1(1-r_1) & \frac{1}{4}r_1^2 & \frac{1}{4}(1-r_1)^2 & \frac{1}{4}r_1(1-r_1)\\ \frac{1}{4}r_1^2 & \frac{1}{4}r_1(1-r_1) & \frac{1}{4}r_1(1-r_1) & \frac{1}{4}(1-r_1)^2\end{bmatrix}\text{, i.e. }\begin{array}{l}p_0=\frac{1}{4}(1-r_1)^2\\ p_1=\frac{1}{4}r_1(1-r_1)\\ p_2=\frac{1}{4}r_1^2\end{array}\quad(3.9)$$

$$F_3:\begin{array}{l}p_0=\tfrac{1}{16}\left(2r_1^2-3r_1+2\right)^2\\ p_1=\tfrac{1}{16}r_1\left(3-2r_1\right)\left(2r_1^2-3r_1+2\right)\\ p_2=\tfrac{1}{16}r_1^2\left(3-2r_1\right)^2\end{array}$$

$$F_4:\begin{array}{l}p_0=\tfrac{1}{16}\left(2-4r_1+5r_1^2-2r_1^3\right)^2\\ p_1=\tfrac{1}{16}r_1\left(2r_1^2-5r_1+4\right)\left(2-4r_1+5r_1^2-2r_1^3\right)\\ p_2=\tfrac{1}{16}r_1^2\left(2r_1^2-5r_1+4\right)^2\end{array}\quad(3.10)$$

$$F_5:\begin{array}{l}p_0=\tfrac{1}{16}\left(2r_1^4-7r_1^3+9r_1^2-5r_1+2\right)^2\\ p_1=\tfrac{1}{16}r_1\left(5-9r_1+7r_1^2-2r_1^3\right)\left(2r_1^4-7r_1^3+9r_1^2-5r_1+2\right)\\ p_2=\tfrac{1}{16}r_1^2\left(5-9r_1+7r_1^2-2r_1^3\right)^2\end{array}$$

(3.11)

$$F_6:\begin{array}{l}p_0=\tfrac{1}{16}\left(2-6r_1+14r_1^2-16r_1^3+9r_1^4-2r_1^5\right)^2\\ p_1=\tfrac{1}{16}r_1\left(2r_1^4-9r_1^3+16r_1^2-14r_1+6\right)\left(2-6r_1+14r_1^2-16r_1^3+9r_1^4-2r_1^5\right)\\ p_2=\tfrac{1}{16}r_1^2\left(2r_1^4-9r_1^3+16r_1^2-14r_1+6\right)^2\end{array}\quad(3.12)$$

Considering phase unknown combines the two heterozygotes, which sums the two middle rows and columns, so that in each of the above cases $\Pr(M',G')=\begin{bmatrix}p_0 & 2p_1 & p_2\\ 2p_1 & 2(p_0+p_2) & 2p_1\\ p_2 & 2p_1 & p_0\end{bmatrix}$. For example,

$$F_2:\Pr\left(M',G'\right)=\begin{bmatrix}\frac{1}{4}\left(1-r_1\right)^2 & \frac{1}{2}r_1\left(1-r_1\right) & \frac{1}{4}r_1^2\\ \frac{1}{2}r_1\left(1-r_1\right) & \frac{1}{2}r_1^2+\frac{1}{2}\left(1-r_1\right)^2 & \frac{1}{2}r_1\left(1-r_1\right)\\ \frac{1}{4}r_1^2 & \frac{1}{2}r_1\left(1-r_1\right) & \frac{1}{4}\left(1-r_1\right)^2\end{bmatrix}$$

2 genes 2 markers ($k = 2$)

Parent$_0$ has $(M, G) = (0,0)$ and Parent$_1$ = (15, 15). The F_1 is (3,3) if a is Parent$_0$ and b is Parent$_1$, or is (12, 12) if a is Parent$_1$ and b is Parent$_0$. The 16×16 matrix Pr(M, G) is in general too large to typeset easily.

Backcross

$$\Pr\left(M',G'\right)=\frac{1}{2\left(1-2r_1\right)}\left[\begin{matrix}\left(1-\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & \left(1-\theta_1-r_1\right)\left(1-r_1\right)r_2 & 0 & \left(\theta_1-r_1\right)r_1\left(1-r_2\right) & \left(\theta_1\right.\\ \left(\theta_1-r_1\right)\left(1-r_1\right)r_2 & \left(\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & 0 & \left(1-\theta_1-r_1\right)r_1r_2 & \left(1-\theta_1\right.\\ 0 & & 0 & 0 & \\ \left(1-\theta_1-r_1\right)r_1\left(1-r_2\right) & \left(1-\theta_1-r_1\right)r_1r_2 & 0 & \left(\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & \left(\theta_1-\right.\\ \left(\theta_1-r_1\right)r_1r_2 & \left(\theta_1-r_1\right)r_1\left(1-r_2\right) & 0 & \left(1-\theta_1-r_1\right)\left(1-r_1\right)r & \left(1-\theta_1-r\right.\\ 0 & & 0 & 0 & \\ 0 & & 0 & 0 & \\ 0 & & 0 & 0 & \\ 0 & & 0 & 0 & \end{matrix}\right.$$

(3.13)

$$BC_0:\Pr\left(M,G\right)=\frac{1}{2\left(1-2r_1\right)}\left[\begin{matrix}\left(1-\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & \left(1-\theta_1-r_1\right)\left(1-r_1\right)r_2 & \left(\theta_1-r_1\right)r_1\left(1-r_2\right) & \left(\theta_1-r_1\right)r_1r_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ \left(\theta_1-r_1\right)\left(1-r_1\right)r_2 & \left(\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & \left(1-\theta_1-r_1\right)r_1r_2 & \left(1-\theta_1-r_1\right)r_1\left(1-r_2\right) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ \left(1-\theta_1-r_1\right)r_1\left(1-r_2\right) & \left(1-\theta_1-r_1\right)r_1r_2 & \left(\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & \left(\theta_1-r_1\right)\left(1-r_1\right)r_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ \left(\theta_1-r_1\right)r_1r_2 & \left(\theta_1-r_1\right)r_1\left(1-r_2\right) & \left(1-\theta_1-r_1\right)\left(1-r_1\right)r_2 & \left(1-\theta_1-r_1\right)\left(1-r_1\right)\left(1-r_2\right) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{matrix}\right.$$

(3.14)

F2

$$\Pr\left(M',G'\right)=\frac{1}{4\left(1-2r_1\right)^2}\begin{bmatrix} p_0 & 2p_1 & p_2 & 2p_3 & 4p_4 & 2p_5 & p_6 & 2p_7 & p_8 \\ 2p_9 & & 2p_9 & 2p_{24} & & 2p_{24} & 2p_{10} & & 2p_{10} \\ p_{11} & 2p_{12} & p_{13} & 2p_5 & 4p_4 & 2p_3 & p_{14} & 2p_{15} & p_{16} \\ 2p_{17} & 4p_{18} & & & & & 2p_{20} & 4p_{21} & 2p_{22} \\ 4p_4 & & 4p_4 & & & & 4p_4 & & 4p_4 \\ 2p_{22} & 4p_{21} & 2p_{20} & & & & 2p_{19} & 4p_{18} & 2p_{17} \\ p_{16} & 2p_{15} & p_{14} & 2p_3 & 4p_4 & 2p_5 & p_{13} & 2p_{12} & p_{11} \\ 2p_{10} & & 2p_{10} & 2p_{24} & & 2p_{24} & 2p_9 & & 2p_9 \\ p_8 & 2p_7 & p_6 & 2p_5 & 4p_4 & 2p_3 & p_2 & 2p_1 & p_0 \end{bmatrix}$$

$$p_0=\left(1-r_1\right)^2\left(1-r_2\right)^2\left(1-r_1-\theta_1\right)^2$$
$$p_1=\left(1-r_1\right)^2 r_2\left(1-r_2\right)\left(1-r_1-\theta_1\right)^2$$
$$p_2=\left(1-r_1\right)^2 r_2^2\left(1-r_1-\theta_1\right)^2$$
$$p_3=r_1\left(1-r_1\right)\left(1-r_2\right)^2\left(\theta_1-r_1\right)\left(1-r_1-\theta_1\right)$$
$$p_4=r_1\left(1-r_1\right)r_2\left(1-r_2\right)\left(\theta_1-r_1\right)\left(1-r_1-\theta_1\right) \quad p_{24}=r_1\left(1-r_1\right)r_2\left(1-r_2\right)\left(1+2\theta_1^2-2\theta_1+2r_1^2-2r_1\right)$$
$$p_5=r_1\left(1-r_1\right)r_2^2\left(\theta_1-r_1\right)\left(1-r_1-\theta_1\right)$$
$$p_6=r_1^2\left(1-r_2\right)^2\left(\theta_1-r_1\right)^2$$
$$p_7=r_1^2 r_2\left(1-r_2\right)\left(\theta_1-r_1\right)^2$$
$$p_8=r_1^2 r_2^2\left(\theta_1-r_1\right)^2$$

Statistical Inference

Likelihood

Likelihood methods require the specification of a probability model. In this section we illustrate how the probability model developed above can be used to construct a likelihood. We anticipate that statistical inference will frequently be performed in the context of phase unknown marker types, M', and so these are used in this section; however, diplotypes could also be used.

For a single observation with $Y=y$ (0 or 1) and $M'=m_i'$, the contribution to the likelihood is $\Pr\left(Y=y,M'=m_i'\mid \boldsymbol{r},\boldsymbol{\theta},\boldsymbol{p}'\right)$. If $y=1$, this is the i^{th} element of the vector $\Omega_1=\Pr\left(Y=1,M'\right)=C_k\times\Pr\left(M,G\right)\times\mathbf{p}=\Pr\left(\mathbf{M}',\mathbf{G}'\right)\times\mathbf{p}'$, whereas if $y=0$, this is the i^{th} element of

$\Omega_0 = \Pr(Y=0, M) = \Pr(\mathbf{M'}, \mathbf{G'}) \times (\mathbf{1}\text{-}\mathbf{p'})$. The Ω's are functions of the parameters **r**, θ, and **p**. For a given sample, let z_i be the observed number of individuals of marker type m_i' who exhibit Y=1, and let n_i be the total number of individuals of that marker type. Thus the number of individuals of type m_i' who exhibit Y = 0 is $n_i - z_i$, and the total sample size is $\sum_{i=0}^{K-1} n_i$The observed variables ***n*** and ***z*** are equivalent to knowing Y and M for all individuals. The likelihood for this sample is $L(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{p}) = \Pr(\boldsymbol{n}, \boldsymbol{z} \mid \boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{p}) = \prod_{i=0}^{K-1} (\Omega_{1i})^{z_i} (\Omega_{0i})^{(n_i - z_i)}$, and the log likelihood is

$\log L(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=0}^{K-1} z_i \log(\Omega_{1i}) + (n_i - z_i)\log(\Omega_{0i})$.

The observed mean of marker class *i* is simply $\hat{\pi}_i = \frac{z_i}{n_i}$. We denote the vector of expected means for all marker classes by *π*. The expected mean of marker class *i* is

$$\begin{aligned} \pi_i &= E(Y \mid M' = m_i') = \Pr(Y = 1 \mid M' = m_i') \\ &= \frac{\Pr(Y = 1, M' = m')}{\Pr(M' = m')} = \frac{\Omega_{1i}}{\Pr(M' = m')} \end{aligned}$$

The vector π can be computed via an elementwise division of the two vectors $\boldsymbol{\Omega_1} = \Pr(Y, M')$ and $\Pr(M')$, i.e. $\boldsymbol{\pi} = \boldsymbol{\Omega_1} \div \Pr(M')$, where we specify $\pi_i = 0$ if $\Pr(M' = m_i') = 0$ for the design in question. We also note that $\boldsymbol{1} - \boldsymbol{\pi} = \Pr(Y = 0 \mid M') = \boldsymbol{\Omega_0} \div \Pr(M')$ where division is again done elementwise. (It follows that $\boldsymbol{\Omega_0} = \Pr(M') - \boldsymbol{\Omega_1}$.) Since $\Omega_{1i} = \pi_i \times \Pr(M' = m_i')$ and $\Omega_{0i} = (1 - \pi_i) \times \Pr(M' = m_i')$, the log likelihood can be written in terms of marker class means as

$$\begin{aligned} \log L(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{p}) &= \sum_{i=0}^{K-1} z_i \left(\log(\pi_i) + \log\left(\Pr(M' = m_i')\right)\right) + (n_i - z_i)\left(\log(1 - \pi_i) + \log\left(\Pr(M' = m_i')\right)\right) \\ &= \sum_{i=0}^{K-1} z_i \log(\pi_i) + (n_i - z_i)\log(1 - \pi_i) + n_i \log\left(\Pr(M' = m_i')\right) \end{aligned}$$

The first two terms depend on the parameters **r**, θ, and **p** via π, while the last term depends only on the parameter θ. The maximum likelihood estimators for the π_i are easily obtained to be the observed marker class means:

$$0 = \frac{\partial}{\partial \pi_i} \log L = \frac{z_i}{\pi_i} - \frac{n_i - z_i}{1 - \pi_i} = \frac{z_i(1 - \pi_i) + (z_i - n_i)\pi_i}{\pi_i(1 - \pi_i)} \Rightarrow \hat{\pi}_i = \frac{z_i}{n_i} \qquad (3.15)$$

This result is natural since the z_i are simply binomial random variables.

Since $\boldsymbol{\pi} = \Pr(Y, M') \div \Pr(M') = (\Pr(M', G) \times \boldsymbol{p}) \div \Pr(M')$ (elementwise division), we have $\boldsymbol{\pi} \circ \Pr(M') = \Pr(M', G) \times \boldsymbol{p}$, and so $\boldsymbol{p} = [\Pr(M', G)]^{-1} (\boldsymbol{\pi} \circ \Pr(M'))$, where $\circ$ indicates elementwise multiplication. Recall that $\Pr(M')$ depends only on θ, while $\Pr(M', G')$ depends on **r** and θ.. In many realistic experiments, the marker map θ will be known (or accurately estimated) from previous experiments, so that $\Pr(M', G')$ can be considered a function of **r**. In such a case, the invariance property of MLE's can be used to obtain the MLE $\hat{\boldsymbol{p}}$ for **p** in terms of the estimates $\hat{\boldsymbol{\pi}}$ and the unknown parameters ***r***, whenever the matrix $\Pr(M', G')$ is invertible, that is, $\hat{\boldsymbol{p}} = [\Pr(M', G)]^{-1} (\hat{\boldsymbol{\pi}} \circ \Pr(M'))$. In a design such as a backcross in which certain genotypes *cannot* be observed, those rows and columns of zeros must be removed from all matrices and vectors, or the matrix will not be invertible. In addition, certain regularity conditions (such as $r_i < ½$, $r_i < \theta_i$) are required, which makes sense since the penetrance of unlinked genes cannot be estimated.

(Section to be added here on simultaneous estimation of r and p)

Alternative Trait Models

While our derivation focuses on binary traits using a single penetrance parameter, the above models could be applied to Xu's liability model by replacing the penetrance vector with the appropriate vector of liability functions, and most of the derivation can also be applied to a quantitative trait.

Quantitative Traits

Suppose that all genotypes have the same known variance σ and that genotype g_j has trait mean μ_j, that is, $\mu_j = E(Y \mid G = g_j)$. These means can be arranged in a vector μ with the same ordering as the genotypes g_j. Then the contribution to the likelihood from an observation (y, m_i) is $f(y, m_i)$, calculated as follows. For any gene diplotype g_j,

$$f(y, m_i, g_j) = f(y \mid m_i, g_j) \Pr(M = m_i, G = g_j) = f(y \mid g_j) (\Pr(M, G))_{ij}$$

$$f(y, m_i) = \sum_{j=0}^{K-1} f(y, m_i, g_j) = \sum_{j=0}^{K-1} f(y \mid g_j) (\Pr(M, G))_{ij} \qquad (3.16)$$

$$= \sum_{j=0}^{K-1} \phi_{\mu_j}(y) (\Pr(M, G))_{ij} = (\Pr(M, G) \times \boldsymbol{\Phi}(y))_i$$

where ϕ_{μ_j} is the normal pdf with mean μ_j and variance σ, and $\boldsymbol{\Phi}$ is a vector of those pdf's.

The mean of marker class m_i is

$$\lambda_i = E\left(Y \mid M = m_i\right) = \frac{\sum_{j=0}^{K=1} \mu_j \Pr\left(M = m_i, G = g_j\right)}{\Pr\left(M = m_i\right)} = \frac{\sum_{j=0}^{K=1} \mu_j \Pr\left(M,G\right)_{i,j}}{\Pr\left(M\right)_i} = \frac{\left(\Pr\left(M,G\right)\times\boldsymbol{\mu}\right)_i}{\Pr\left(M\right)_i} \quad (3.17)$$

$$= \left(\Pr\left(G \mid M\right)\times\boldsymbol{\mu}\right)_i$$

Thus, the means for all marker classes can be represented as a vector $\boldsymbol{\lambda} = \Pr\left(G \mid M\right)\times\boldsymbol{\mu}$, where the rows of μ and the columns of Pr(G|M) refer to gene diplotypes g_j, and the rows of λ and Pr(G|M) refer to marker diplotypes m_i.

The major difference between the present matrix formulation and those previously found in the literature, is that previous attempts have focused on specific cases of additivity, dominance and epistasis. (See, for example, Falconer and MacKay, 1996.) However, in this formulation, no explicit genetic model is assumed. Conventional models are a subset of the present model, where the free parameters are reduced in number according to the constraints of the assumed genetic model. For example, under certain assumptions trait means are often represented in terms of additive and dominance factors; that is, assuming no epistasis, trait locus ℓ is often assumed to contribute additively to the trait mean, by a factor of $+a_\ell$ if the genotype is 0, by $+d_\ell$ if the genotype is 1, and by $-a_\ell$ if the genotype is 2. How does this correspond to our notation? Let g_j^ℓ represent the allele (0, 1, or 2) of genotype g_j at locus ℓ (the ℓ^{th} digit in the base 3 representation of j). Then $\mu_j = \mu + \sum_{\ell=1}^{k} a_\ell \delta\left(g_j^\ell = 0\right) + \sum_{\ell=1}^{k} d_\ell \delta\left(g_j^\ell = 1\right) - \sum_{\ell=1}^{k} a_\ell \delta\left(g_j^\ell = 2\right)$, where μ is the base trait mean and δ is an indicator function. The 3^k free parameters in an unconstrained μ are thus reduced to $2k$+1 parameters by these assumptions. A less constrained μ permits the introduction of epistasis into the model but does not force epistasis to be present. Thus, the full set of parameters can be estimated, and equivalences tested to determine the genetic model.

Discussion

We use classical transmission genetics to derive a general mathematical framework for the transmission of an arbitrary number of marker and trait loci. This notation is particularly useful in the study of complex traits, where the underlying factors are assumed to be multigenic. We explicitly assume no interference and random union of gametes in this particular formation. Also omitted are mutation and selection effects. However, the mathematics will hold together even when these assumptions are relaxed.

We present the solution for an arbitrary number of generations in a direct line, that is, F2, F3, F4 assuming random mating between the individuals in the parental generation. This model is easily modified to allow for selfing down the generations simply by restricting the probability of the transmissions and to allow for other mating designs by constructing appropriate parental matrices and forming the correct product.

We have coded the model into Maple for ease of use and to allow easy access for researchers seeking to implement this model. This code is available on request from the author (simonsen@stat.purdue.edu). Maple allows symbolic representation and is an easy to use package that integrates well with other programming languages.

End Matter

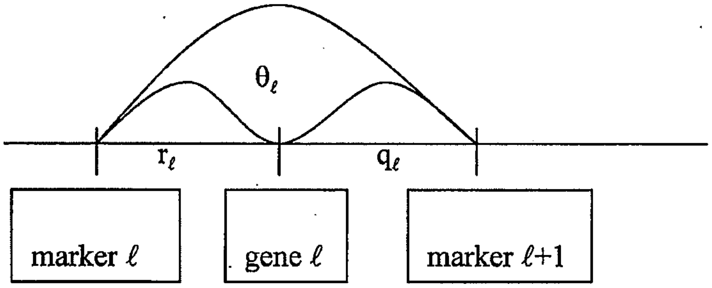

Figure 1: Map

References

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics, 4th ed.* Addison Wesley Longman, Harlow, England.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics* 8: 299–309.

Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Ann Eugen* 12: 172–175.

Lander, E.S. and Botstein D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121 (1): 185–199.

Lander, E. S. and Green P. (1987). Construction of multilocus genetic-linkage maps in humans. *P Natl Acad Sci USA* 84 (8): 2363–2367

Xu, S. Z. and Atchley, W. R. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* 143 (3): 1417-1424.

Box 1: Notation

Random Variables

M: marker diplotype (random variable); possible values $m_0 \ldots m_{K-1}$
N: marker haplotype (random variable) possible values $n_0 \ldots n_{L-1}$
G: trait gene diplotype (random variable); possible values $g_0 \ldots g_{K-1}$
H: trait gene haplotype (random variable); possible values $h_0 \ldots h_{L-1}$
Y: binary trait value random variable; possible values $y = 0$ or 1
T: transmission indicators for marker haplotypes, possible values $t_0, \ldots t_{L-1}$
S: transmission indicators for gene haplotypes, possible values $s_0, \ldots, s_{L-1}$

Parameters

k: number of marker loci = number of potential trait genes
$L = 2^k$: number of possible haplotypes
$K = 4^k = L^2$: number of possible diplotypes
θ_ℓ: recombination probability between marker ℓ and marker $\ell+1$, $\ell = 1, \ldots, k-1$
r_ℓ: recombination probability between marker ℓ and gene ℓ , $\ell = 1, \ldots, k$
q_ℓ: recombination probability between gene ℓ and marker $\ell+1$, $\ell = 1, \ldots, k-1$
$\mathrm{p}_i = \Pr(Y = 1 | G = g_i)$: penetrances, $i = 0, \ldots, K-1$
ℓ: index for loci
C_k: phase matrix

Operators

$\otimes$: Kronecker product of matrices or vectors
$\times$: matrix multiplication
$(\ )'$: phase-unknown
$(\ )^\dagger$: matrix (or vector) transpose

Probability distributions

Θ: transmission probabilities for marker haplotypes (4^k)
Ω: joint transmission probabilities for marker and gene haplotypes ($2^k \times 2^k$)
Δ: probability of haplotypes conditional on transmission indicators (marker or gene) ($2^k \times 2^k$)
Λ: joint probabilities of marker and gene haplotypes
Γ: probabilities of marker haplotypes
$\Pr(M)$: vector of probabilities; i^{th} entry is $\Pr(M = m_i)$
$\Pr(M, G)$: matrix of probabilities; i,j^{th} entry is $\Pr(M=m_i, G = g_j)$