frontiers

Probabilities of multilocus genotypes in SIB recombinant inbred lines

Kamel Jebreen [1,2], **Marianyela Petrizzelli** [1] **and Olivier C. Martin** [1,*]

[1] *INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France*

[2] *Department of Mathematics, An-Najah National University, Nablus, Palestine*

Correspondence*:
Olivier C. Martin; the first two authors are "equal contribution"
olivier.c.martin@inra.fr

ABSTRACT

Recombinant Inbred Lines (RILs) are obtained through generations of inbreeding until all alleles are fixed. In 1931 Haldane and Waddington published a landmark paper where they provided the probabilities of achieving any combination of alleles in 2-way RILs for 2 and 3 loci. In the case of SIB RILs where sisters and brothers are crossed at each generation, there has been no progress in treating 4 or more loci, a limitation we overcome here without much increase in complexity. In the general situation of L loci, the task is to determine 2^L probabilities, but we find that it is necessary to first calculate the 4^L "identical by descent" (IBD) probabilities that a RIL inherits at each locus its DNA from one of the four originating chromosomes. We show that these 4^L probabilities satisfy a system of linear equations that follow from self consistency. In the absence of genetic interference – crossovers arising independently –, the associated matrix can be written explicitly in terms of the recombination rates between the different loci. We provide the matrices for L up to 4 and also include a computer program to automatically generate the matrices for higher values of L. Furthermore, our framework can be generalized to recombination rates that are different in female and male meiosis which allows us to show that the Haldane and Waddington 2-locus formula is valid in that more subtle case if the meiotic recombination rate is taken as the average rate across female and male. Once the 4^L IBD probabilities are determined, the 2^L probabilities of RIL genotypes are obtained *via* summations of these quantities. *In fine*, our computer program allows to determine the probabilities of all the multilocus genotypes produced in such sibling-based RILs for $L \leq 10$, a huge leap beyond the $L = 3$ restriction of Haldane and Waddington.

1 INTRODUCTION

There are numerous inference problems in population and quantitative genetics that require comparing experimental frequencies of genotypes to those expected "theoretically". Examples include genetic mapping of genomic markers, localizing causal factors of diseases and

quantitative traits, performing marker assisted selection etc (Lander & Schork (1994); Weir (1996); Walsh & Lynch (2018)). The *expected* frequencies of genotypes, hereafter referred to as probabilities, of interest in such studies often involve multiple loci (Buckler *et al.* (2009)) and are strongly dependent on population structure. In population genetics studies, the structure of *natural* populations is rarely perfectly known. That partly explains why, in both animal and plant genetics, controlled crosses are widely produced to ensure a specific population structure. Arranging the crosses to lead to homozygous lines is greatly advantageous as such lines can be reproduced "identically and indefinitely". The simplest situation satisfying these criteria is that of *recombinant inbred lines* (RILs)(Crow (2007)) founded from two parents as displayed in Fig.1. Given two (generally homozygous) parents that are the founders of the RIL construction (F_0), one first produces the associated hybrids (F_1). Second, starting with these F_1 individuals, one produces a sequence of generations F_2, F_3, etc by iterative inbreeding, crossing male and female siblings until formally at F_∞ one reaches full homozygozity (fixation of the alleles at all loci). As seen in Fig.1, the genomes of the homozygous lines produced by this process are mosaics of the parental genomes.

Consider the allelic content at some set of L genomic markers or loci. There are then 2^L possible RIL genotypes, each having a probability that depends on how meioses generate recombinations between these different loci. In the case of plants that allow for selfing, the same individual is both the mother and the father of its offspring; the RILs are then produced *via* single seed descent (SSD) as opposed to *via* sibling (SIB) mating, this second case being the focus of the present work.

There are numerous generalizations of the RIL construction just given. Instead of using two parents to initiate the inbreeding, the use of 2^k parents leads to 2^k-way RILs (Broman (2005)). 2^k-way RILs start with 2^k parents to form 2^{k-1} offspring that are themselves crossed iteratively following a funnel (specifically a binary tree) pattern. Once the root of this tree is reached, the usual RIL inbreeding process is applied. For instance, the so called "Collaborative Cross" which has been a key community tool for mouse genetics, corresponds to $k = 3$; the choice there of using 8 founding parents at the top of the funnel allows for significantly greater allelic diversity than when using just 2-way RILs. Another generalization is the so called Advanced Intercross RIL (AI-RIL, sometimes referred to as Intermated RIL or IRIL) in which several generations of panmixia are inserted before applying the inbreeding to produce the RILs (Darvasi & Soller (1995); Rockman & Kruglyak (2008); Winkler *et al.* (2003)). Other generalizations include Multi-parent Advanced Generation Inter-Cross (MAGIC) (El-Din El-Assal *et al.* (2001)), *nested association mapping* (NAM) populations (Buckler *et al.* (2009)) etc. All of these population constructions involve some initial generations of allelic shuffling followed by the RIL (inbreeding) construction per se. Those *early* generations produce in effect initial conditions on the genotypes that are at the origin of the RILs and these initial conditions can be computed by direct recurrence from one generation to the next. In contrast, the RIL phase requires crossings that continue until all loci are homozygous and thus – at least mathematically – this phase involves an infinite number of generations. As a result, the computation of the probabilities of multilocus genotypes in RILs does not follow from a simple recursion over a fixed number of generations: either an extrapolation has to be made to deal with the infinite number of generations or some mathematical trick has to be

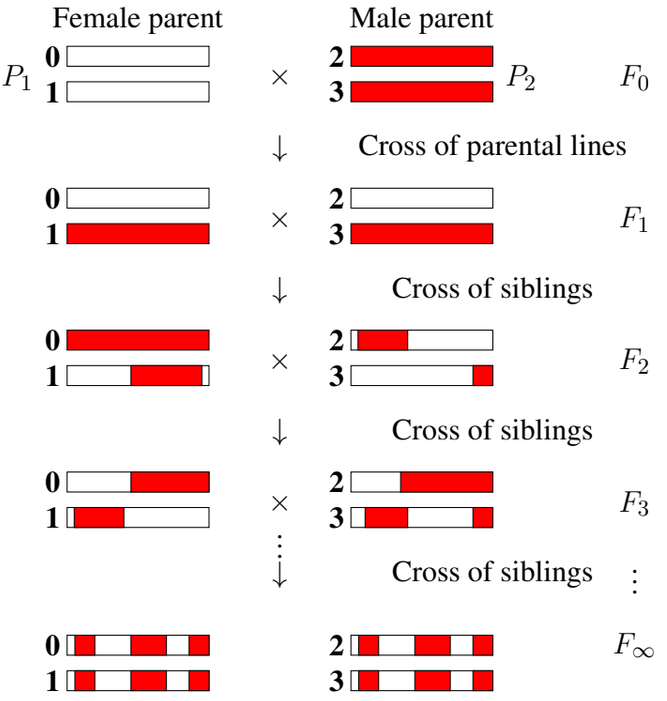

Figure 1. The production of recombinant inbred lines by sibling mating (SIB RILs). The homologous chromosomes (numbered 0, 1, 2, 3) inherit genomic segments from their parents, the boundaries of which identify crossover positions. The allelic content becomes fixed for "enough" generations, and so we introduce the limit of an infinite number of generations with the notation F_∞.

devised to bypass the infinite nature of the process. This fact is at the heart of the difficulty of obtaining exact probabilities of multilocus genotypes in RILs.

The mathematical derivation of such RIL probabilities for two and three loci was provided by Haldane and Waddington (Haldane & Waddington (1931)) for bi-parental RILs in 1931. For two loci, by considering successive generations, they produced recursion equations for the probabilities of the corresponding (fixed or not) SIB genotypes which they then extrapolated to an infinite number of generations. This was quite a feat as they had to solve 22 simultaneous equations, leading *in fine* to their celebrated relation:

$$R = \frac{4r}{1+6r}. \tag{1}$$

In this formula, R is the probability for a RIL two-locus genotype to be recombinant (have the allele of one F_0 parent at one locus and the allele of the other at the other locus) while r is the recombination rate per meiosis between the two loci, assumed identical across male and female meiosis. We will rederive this formula using our framework in Section 3.1 because to our knowledge, the generalization of the Haldane-Waddington formula to situations where male and female recombination rates differ has not been published and our framework allows to deal with this extension.

Given R, it is easy to derive the probabilities of the four different RIL genotypes (each of the two loci can be fixed for either of the two parental alleles). Indeed, the two recombinant genotypes have the same probability and the sum of these two probabilities is precisely R. The probability of each of the two recombinant (respectively non-recombinant) RIL genotypes is then $R/2$ (respectively $(1-R)/2$).

Haldane and Waddington further showed that this two-locus result also determined the three-locus probabilities. A way to see this is to notice that for three loci ($L=3$) there are $2^L=8$ different RIL genotypes (at each locus the homozygous allelic state comes from one of the two parents). These 8 genotypes can be grouped into 4 pairs such that within each pair one genotype is obtained from the other by exchanging the alleles of the parents; for instance if the alleles of the parents are denoted by (A, B, C) and (a, b, c) at the three successive loci, the 4 pairs are $\{(A,B,C),(a,b,c)\}$, $\{(A,B,c),(a,b,C)\}$, $\{(A,b,C),(a,B,c)\}$, and $\{(a,B,C),(A,b,c)\}$. In each pair, the two complementary genotypes have the same probability so in effect it is enough to find the probabilities of each of the 4 pairs. These probabilities add up to one, providing a first equation. Then, labeling the loci as 1, 2, and 3, if the three meiotic recombination rates $r_{1,2}$, $r_{2,3}$ and $r_{1,3}$ are known, the three RIL recombination rates $R_{1,2}$, $R_{2,3}$ and $R_{1,3}$ are also. These quantities provide three further equations relating the four pair probabilities. These four equations uniquely determine the four pair probabilities and thus the probabilities of the 8 RIL genotypes.

Since that 1931 Haldane-Waddington landmark paper, some works have provided generalizations of Eq. 1, for instance in the case of 2^k-way RILs (Broman (2005); Teuscher & Broman (2007)) and in the case of IRILs (Winkler *et al.* (2003); Teuscher & Broman (2007)). However, the problem of dealing with more than three loci seems substantially more difficult. Following the Haldane-Waddington algebraic approach, if there are L loci, there are 16^L possible allelic combinations at each generation and so it is necessary to diagonalize a $16^L \times 16^L$ matrix; that task takes on the order of 16^{3L} operations and thus cannot be done on a standard computer even for $L=4$. To our knowledge, the only work providing closed-form expressions for 4 or more loci is that of Samal and Martin (Samal & Martin (2015)) but their framework for determining exact probabilities of RIL multilocus genotypes applies only to single seed descent RILs, not to SIB RILs. The contribution of the present work is to show that the case of SIB RILs is also to a large extent tractable. In particular, (i) we give the analytic expressions for treating four loci in the absence of crossover interference, and (ii) we show that our framework allows to tackle more loci, though at a computational cost (CPU time and also computer memory) that increases roughly as 16^L. Specifically, our computer scripts, written in R (Ihaka & Gentleman (1996)), can treat $L=8$ loci in approximately 5 minutes when run on a desktop computer while a high-end server allows us to go up to $L=10$ loci.

2 OVERVIEW OF THE METHOD

In the less complex case of single seed descent RILs, it was possible to determine the probabilities of the 2^L RIL multilocus genotypes by writing self-consistent equations directly associated with these unknowns (Samal & Martin (2015)). However, in the case of SIB RILs,

the situation is more subtle because the allele carried by a RIL genotype may come from *either of the two siblings* at the F_1 generation and thus "identical by descent" (IBD) does not reduce to identity by state (having the same allelic content) as can be seen in Fig. 1. As a result, it is necessary to first work with the 4^L probabilities that a RIL inherits IBD at the L loci from any of the four F_1 homologous chromosomes. After introducing in Section 2.1 the 4^L RIL multilocus IBD probabilities, we show in Section 2.2 that each of these unknowns satisfies a self-consistent equation relating it to the others. These equations allow to overcome the technical obstacle of there being an unlimited number of generations in the process of generating RILs. Although these 4^L self-consistent equations constrain the 4^L unknowns, we show in Section 2.3 that one additional equation is necessary to specify the solution. For that last constraint we use the fact that the sum of all probabilities is 1. In Section 2.4 we show how the complexity of the problem can be reduced by working with a subset only of the unknowns. Finally, upon solving the system of equations to determine the IBD quantities, each of the 2^L RIL multilocus genotype probabilities follows by summing the probabilities of all compatible IBDs as will be shown in Section 2.5.

2.1 Probabilities of multilocus IBD inheritances in RILs and the set of non-equivalent Q's

For a given RIL L-locus genotype (specified formally at generation F_∞), the genomic content at any locus ℓ ($\ell \in \{1, \ldots, L\}$) will be IBD with exactly one of the four F_1 homologous chromosomes. (One may note that the allelic fixation can happen before the IBD fixation, but no matter what, after an infinite number of generations both the IBD and the allelic states are fixed, that is they are identical across the four chromosomes of the SIB pair.) We number those four chromosomes 0, 1, 2 and 3 as indicated in Fig. 2 and use the same labeling for the later generations too. The IBD case illustrated is such that the RIL inherits from the F_1 chromosome 2 at the first locus and from the F_1 chromosome 1 at the second locus. (By convention we order the loci from left to right.) More generally, let us introduce the probability $Q(i_1, i_2, \ldots, i_L)$ that a RIL inherits IBD from F_1 chromosome i_ℓ for locus ℓ, $\ell = 1, \ldots, L$ where $i_\ell = 0, 1, 2, 3$. Naturally the sum of these 4^L probabilities (there are four possible values of i_ℓ at each locus ℓ) is equal to 1.

For $L = 1$, there are four IBD probabilities: $Q(0)$, $Q(1)$, $Q(2)$ and $Q(3)$. We shall assume Mendelian segregation with no bias in favor of any particular allele and so in particular the two homologues within each sex are equivalent. Then $Q(i) = 1/4$ for all $i \in \{0, 1, 2, 3\}$. Moving on to $L = 2$ for which there are 16 Q's, the equivalence of homologues leads to the equalities $Q(0, 0) = Q(1, 1)$, $Q(0, 1) = Q(1, 0)$, $Q(2, 2) = Q(3, 3)$, and $Q(2, 3) = Q(3, 2)$ but also to equalities between mixed terms, $Q(0, 2) = Q(0, 3)$, $Q(1, 2) = Q(1, 3)$ etc. Furthermore, if female and male meiosis behave in the same way (so that in particular they have the same recombination rates), we can also conclude that $Q(0, 0) = Q(2, 2)$ etc so that finally there are just *three* probabilities to determine, $Q(0, 0)$, $Q(0, 1)$ and $Q(0, 2)$ instead of the initial 16. More generally, if there are L loci, how many *non-equivalent* Q's are there? We shall assume there is no segregation bias and that female and male meioses have statistically identical behavior. Then it is possible to show (see Supplementary Material for details) that the number

of non-equivalent Q's is exactly

$$N_Q(L) = 2^{L-2}(2^{L-1} + 1). \tag{2}$$

For example $L = 1$ leads to $N_Q(L) = 1$ while $L = 2$ leads to $N_Q(L) = 3$. The number of these non-equivalent Q's grows roughly as $(1/8) \times 4^L$ to be compared with the total number ignoring equivalence of 4^L. The factor (1/8) clearly makes it worth while to use such a reduction in the number of unknowns to simplify the task of writing and solving the equations. The proof of Eq. 2 in the Supplementary Material provides a way to enumerate the Q's to be kept and schematically goes as follows. First, because all four chromosomes play equivalent roles, we can force i_1 to be 0. Second, i_2 can be constrained not to take the value 3 since that value can be replaced by 2, this time by equivalence of chromosomes 2 and 3. If i_2 takes the value 0 or 1, we can again constrain i_3 to be different from 3 by the same reasoning. If instead $i_2 = 2$, then i_3 must be allowed to take all values 0, 1, 2 and 3. We can proceed in this way to define the rules to be applied to the successive i_ℓ. As long as the current list consist of 0s and 1s, the next i can be constrained to not take the value 3 by equivalence between chromosomes 2 and 3, but for all entries after the *first* occurrence of a 2, all values must be allowed (see the Supplementary Material for the final steps required to prove Eq. 2). As an illustration, the reader can check that for $L = 3$ loci, this construction leads to 10 non-equivalent Q's, namely $Q(0,0,0)$, $Q(0,0,1)$, $Q(0,0,2)$, $Q(0,1,0)$, $Q(0,1,1)$, $Q(0,1,2)$, $Q(0,2,0)$, $Q(0,2,1)$, $Q(0,2,2)$, and $Q(0,2,3)$.

2.2 Self-consistent equations for the 4^L IBD probabilities

The IBD inheritance needs an infinite number of generations to become fixed with certainty, at least in principle. Our strategy consist in mapping such an infinite process into a finite one by relying on self-consistency. The probability for F_∞ siblings to inherit IBD the sequence of "indices" $(i_1, i_2, \ldots, i_L)$ from the F_1 chromosomes can be decomposed into trajectories where the inheritance indices at the F_2 level are also made explicit. If we denote these by $(i'_1, i'_2, \ldots, i'_L)$, we can reinterpret $Q(i_1, i_2, \ldots, i_L)$ as a sum of contributions:

$$Q(i_1, i_2, \ldots, i_L) = \sum_{(i'_1, i'_2, \ldots, i'_L)} T\left[(i_1, i_2, \ldots, i_L) \to (i'_1, i'_2, \ldots, i'_L)\right] Q(i'_1, i'_2, \ldots, i'_L) \tag{3}$$

where $T[(\cdot) \to (\cdot)]$ is the transition probability of having the IBD propagate from the first list of indices to the second list of indices when going from the F_1 to the F_2 generation. $T[(\cdot) \to (\cdot)]$ is illustrated graphically in Fig. 2 by considering the case of two loci and having $i_1 = 2$, $i_2 = 1$, $i'_1 = 1$ and $i'_2 = 2$.

Clearly $T[(\cdot) \to (\cdot)]$ depends on the meiotic process, and thus in particular on the recombination rates between loci. To simplify the notation, let us set $u = (i_1, i_2, \ldots, i_L)$ and $v = (i'_1, i'_2, \ldots, i'_L)$. These transition probabilities $T[(\cdot) \to (\cdot)]$ satisfy three properties. First, if $i_k = 0$ or 1, $T[u \to v] = 0$ unless $i'_k = 0$ or 2. Similarly, if $i_k = 2$ or 3, $T[u \to v] = 0$

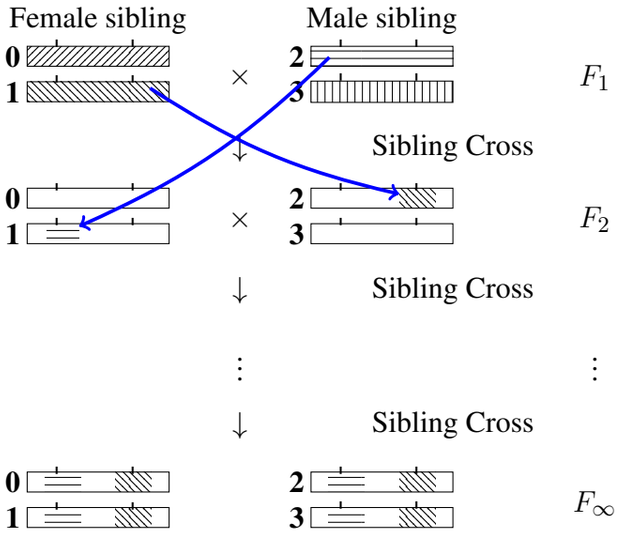

Figure 2. Inheritance during SIB mating and illustration of the construction of a self-consistent equation for any IBD probability. At each generation the homologous chromosomes are labeled 0, 1 (for the female) and 2, 3 (for the male). Note that the chromosomes labeled 0 and 2 are the outcomes of female meiosis while the chromosomes labeled 1 and 3 are the outcomes of male meiosis. The drawing illustrates the transition probability $T\left[(2,1) \to (1,2)\right] Q(1,2)$ entering the self-consistent equation (*cf.* Eq. 3 when the left-hand side is $Q(2,1)$.

unless $i'_k = 1$ or 3. We summarize this via the rules

$$i' \in \begin{cases} \{0,2\} & \text{if } \; i \in \{0,1\} \\ \{1,3\} & \text{if } \; i \in \{2,3\} \end{cases} \tag{4}$$

where i and $i' \in \{0,1,2,3\}$. Second, it turns out that the matrix T is "doubly stochastic" meaning that the sum of its entries in any row or in any column is exactly 1. The result that the sum over elements in a row is 1 follows from the fact that this sum gives the probability of having any of the possible outcomes of inheritances for a given starting point. Analogously, the result that the sum over all elements in a column is 1 corresponds to the fact that a given v is reached by some u and that summing over all possibilities for u again leads to 1. Third, each element of T decomposes into four factors,

$$T\left[u \to v\right] = P_0\left[u \to v\right] P_1\left[u \to v\right] P_2\left[u \to v\right] P_3\left[u \to v\right] \tag{5}$$

where the subscript of each P labels the chromosome of interest (and therefore the meiosis) at the F_2 generation, thus P_j is a probability associated with the meiosis that produces chromosome j when going from F_1 to F_2. Consider for specificity the term P_3. For the computation of this probability, only the entries in v equal to 3 matter. The corresponding *indices* specify which loci are thereafter IBD from chromosome 3 when considering the F_∞ inheritance from the F_2 generation. If those loci numbers are say 2, 5, and $(L-1)$, then $P_3\left[u \to v\right]$ is the probability for the loci 2, 5 and $(L-1)$ to inherit IBD from i_2, i_5 and i_{L-1}

during the meiosis producing chromosome 3 when going from the F_1 generation to the F_2 generation. Note that all the other loci and chromosomes are irrelevant for this factor. The probability of that event is 0.5 (for the probability that the locus 2 will inherit IBD from chromosome i_2) times the probability that the successive intervals 2-5 and 5-$(L-1)$ will be as required – recombinant or not – by the values of i_5 and i_{L-1}. Let us suppose that meioses arise without genetic interference, that is according to the so-called Haldane model (Haldane *et al.* (1919)). (Note that the values of these Ps are the only part of our framework where crossover interference affects our computations; if these single-meiosis probabilities are known, then our framework provides the probabilities of all RIL multilocus genotypes just as in the case of no interference.) For specificity, if there is no interference and both intervals 2-5 and 5-$(L-1)$ are recombinant, the associated (meiotic) probability P is simply 0.5 $\times r_{2,5} \times r_{5,L-1}$. Such a reasoning is easily extended to any situation, leading to the formula

$$P_j\left[u \to v\right] = 0.5 \prod_{<l,l'>} r_{l,l'}^{e_{l,l'}} (1 - r_{l,l'})^{1-e_{l,l'}} \tag{6}$$

where the locus indices l and l' are such that $v_l = v_{l'} = j$, j being the index appearing in the probability P_j. In addition, the $e_{l,l'}$ are defined as

$$e_{l,l'} = \begin{cases} 1 & \text{if the interval is "recombinant"} \\ 0 & \text{if the interval is not "recombinant".} \end{cases} . \tag{7}$$

For Eq. 6, an interval $< l, l' >$ is called "recombinant" if and only if i_l and $i_{l'}$ differ. Lastly, we need to specify the actual pairs of loci l and l' that are to be used in that equation. To do so, we first construct the list of ordered indices that satisfy the constraint $v_l = v_{l'} = j$. The product in Eq. 6 is then over the *successive* pairs of this list. If the list is empty, $P_j = 1$ while if there is only one element in the list, $P_j = 0.5$ .The interpretation of Eq. 6 is then as follows: there is a factor $r_{l,l'}$ if the u list imposes that the interval $< l, l' >$ be recombinant and a factor $1 - r_{l,l'}$ otherwise. Putting together Eqs.3 and 5 specifies the 4^L linear homogeneous equations for the Q's. In our computer software, we determine the matrix elements of T as formal mathematical functions of the $r_{l,l'}$. In these general expressions it is possible to substitute the numerical values of the $r_{l,l'}$ when necessary.

2.3 Adding one linear inhomogeneous equation to uniquely specify all 4^L IBD probabilities

Eq.3 can be rewritten as

$$Q(u) - \sum_v T\left[u \to v\right] Q(v) = 0 \tag{8}$$

for all choices of u, corresponding to a set of 4^L linear homogeneous equations. Given one has as many equations as unknowns, one might hope that this system would determine the Q's but that is not the case because these 4^L equations are not independent. Indeed, consider

the sum of all the equations in the system:

$$\sum_u Q(u) - \sum_u \sum_v T\left[u \to v\right] Q(v) = 0. \tag{9}$$

By interchanging the order of the sums this becomes

$$\sum_u Q(u) - \sum_v \left(\sum_u T\left[u \to v\right] \right) Q(v) = 0 \tag{10}$$

which is automatically satisfied because T is doubly stochastic so that $\sum_u T\left[u \to v\right] = 1$. To overcome the problem coming from this dependence amongst the homogeneous self-consistent equations, we need to include further information. We choose to do that by adding the constraint that the sum of all 4^L IBD probabilities equals 1:

$$\sum_u Q(u) = 1. \tag{11}$$

The inclusion of this (inhomogeneous) linear equation then uniquely specifies the values of all Q's.

2.4 Reducing the system of equations to treat only the $N_Q(L)$ non-equivalent Q's

As mentioned previously, it is advantageous to work with a subset of non-equivalent Q's because this substantially reduces the complexity of the operations to be performed. Specifically, we modify the above approach by considering self-consistent equations only for the reduced list of unknowns – the $N_Q(L)$ non-equivalent Q's chosen in Section 2.1 – so instead of having 4^L homogeneous equations of the type Eq. 8 we have only $N_Q(L)$ of them. In these $N_Q(L)$ equations, we replace each $Q(v)$ by an equivalent $Q(v')$ where $Q(v')$ belongs to our list of $N_Q(L)$ unknowns. This recipe leads to $N_Q(L)$ linear homogeneous equations for our unknowns. Furthermore, we also apply these substitutions to the inhomogeneous equation Eq. 11, with the previously mentioned rule. As a result, by counting the number of Q's arising in each equivalence class defined in Section 2.1, $Q(u)$ occurs with weight 4 if the entries of u are all different from 2 and with weight 8 otherwise.

In practice, to solve this set of equations, it is convenient to have as many equations as unknowns so we remove exactly one of the homogeneous equations. In our computer algorithm we remove the last of these homogeneous equations but any other choice is just as valid. Having obtained as many independent equations as there are unknowns, the direct solution of this linear system (a linear algebra problem) provides the (unique) values of our $N_Q(L)$ non-equivalent Q's.

2.5 Extracting the 2^L probabilities of RIL genotypes

Once the Q's are determined, the probabilities of RIL multilocus genotypes can be computed by summing all IBD probabilities that are *compatible* with the RIL allelic content. Let us

refer to the allelic content of parent 1 as a series of A alleles and that of parent 2 as a series of a alleles. Consider then a RIL multilocus genotype, written as a list $G = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ of L alleles, α_k being A or a. The probability of a genotype G is obtained by summing over all $Q(u)$ for which the u is compatible with the allelic content of G. The compatibility rule can be summarized as follows: if $\alpha_k = A$, then u_k must be 0 or 2, while if $\alpha_k = a$, then u_k must be 1 or 3. This is formalized mathematically by the following equation

$$P(G = (\alpha_1, \alpha_2, \ldots, \alpha_L)) = \sum_u Q(u) \tag{12}$$

where the sum is restricted to the u's satisfying the compatibility rule. Note that the Q's on the right-hand side of Eq. 12 in general will not belong to our list of non-equivalent Q's. As before, just omit all the terms associated with Q's that are not in this list and multiply the other terms by either 8 or 4 depending on whether the associated u has one of its indices u_k equal to 2 or not, again because of the size of the equivalence classes.

3 RESULTS

We illustrate the power of our framework by considering increasing number of loci. The case of two loci is presented both for pedagogical reasons and to give the novel (as far as we know) values of the IBD probabilities when allowing for sex-dependent recombination rates. For three loci we detail the derivation of the coefficients of the self-consistent equations by giving associated graphical representations in the Supplementary Material. For four loci the analytical expression of the 40×40 matrix is also given explicitly. For more loci, the mathematical steps become too cumbersome to be dealt with by hand, but our computer code (in the form of R functions) can be used to first generate the analytic expressions for the linear system of equations, then to solve that system for the Q's, and finally to produce the probabilities of all the RIL multilocus genotypes. The complexity of the computations provided by our framework can be summarized via the dimensionality of the linear system of equations used to compute the Q's. This dimension increases roughly by a factor 4 for each additional locus for the simple reason that the number of unknowns increases in that way (*cf.* Eq. 2).

3.1 Case of two loci: recovering the Haldane-Waddington result and allowing for sex-dependent recombination rates

Haldane and Waddington (Haldane & Waddington (1931)) derived the formula for the probabilities of 2-locus RIL genotypes and Teuscher *et al.* (Teuscher & Broman (2007)) gave an alternative more compact approach. We will derive that Haldane-Waddington result here using our self-consistency approach. Then we show how to extend our framework to the case where female and male recombination rates differ.

Let $r_{l,l'} = r_{1,2}$ denote the recombination fraction between the two loci (this recombination rate is for the moment taken to be the same in female and male as assumed by Haldane and Waddington). Furthermore, let a_l denote the allele at locus l, $l \in \{1, ..., L\}$, on any of the homologous chromosomes in the RIL. By Eq. 2, for $L = 2$ there are 3 unknown Q's. The

indices u for each of these Q's are such that they are not related by the symmetry between chromosomes. Our choice is to use $Q(0,0)$, $Q(0,1)$ and $Q(0,2)$. To build the 3×3 system of equations, begin with the inhomogeneous linear equation

$$4Q(0,0) + 4Q(0,1) + 8Q(0,2) = 1. \tag{13}$$

where the respective factors 8 and 4 follow from whether or not the u list of indices contains a 2. The next step is to write the self-consistent equation for each of the $N_Q(L) - 1$ non-equivalent Q's. For instance for $u = (0,0)$, by Eq. 3 applied to this case and using the rules for the vanishing of the elements of the matrix T, one has

$$\begin{aligned} Q(0,0) = {} & T\left[(0,0) \to (0,0)\right] Q(0,0) \\ & + T\left[(0,0) \to (0,2)\right] Q(0,2) \\ & + T\left[(0,0) \to (2,0)\right] Q(2,0) \\ & + T\left[(0,0) \to (2,2)\right] Q(2,2). \end{aligned}$$

The matrix elements $T\left[u \to v\right]$ are determined by Eqs. 5 and 6. Direct calculation gives $(1 - r_{1,2})/2$, $1/4$, $1/4$, and $(1 - r_{1,2})/2$ respectively. To obtain a self-consistent equation involving only our three non-equivalent Q's, we rewrite Eq. 14 by replacing $Q(2,0)$ by $Q(0,2)$ and $Q(2,2)$ by $Q(0,0)$, leading to

$$r_{1,2}Q(0,0) - Q(0,2)/2 = 0. \tag{14}$$

The self-consistent equation for $Q(0,1)$ is obtained by the same method. Eq. 13 together with Eq. 14 and its analogue for $Q(0,1)$ then lead to the system

$$\begin{bmatrix} 4 & 4 & 8 \\ r_{1,2} & 0 & -\frac{1}{2} \\ r_{1,2} & -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Q(0,0) \\ Q(0,1) \\ Q(0,2) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{15}$$

(Compared to Eq.8, we have changed the signs of each homogeneous equation to obtain a more readable matrix). This system can be solved by hand, leading to

$$Q(0,0) = \frac{1}{4 + 24r_{1,2}}, \quad Q(0,1) = Q(0,2) = \frac{r_{1,2}}{2 + 12r_{1,2}}. \tag{16}$$

Given these three values, we can compute the RIL recombination rate R by summing all the probabilities of IBD events that produce recombinant RILs:

$$R = Q(0,1) + Q(0,3) + Q(2,1) + Q(2,3) + Q(1,0) + Q(1,2) + Q(3,0) + Q(3,2). \tag{17}$$

Using the equivalences ($Q(3,0) = Q(0,2)$ etc), this gives $R = 4Q(0,1) + 4Q(0,2)$; substituting the values from Eq. 16 leads directly to the Haldane-Waddington formula, Eq. 1.

How do these results extend to the case where female and male have different recombination rates, r^f and r^m? The main complication comes from the fact that the symmetries of the system are reduced: one can no longer exchange the roles of female and male SIBs. As a result, there are 6 non-equivalent IBD probabilities. Without loss of generality, we take these to be $Q(0,0)$, $Q(0,1)$, $Q(0,2)$, $Q(2,0)$, $Q(2,2)$, and $Q(2,3)$. The determination of these six unknowns follows the same logic as when $r^f = r^m$. First, use the inhomogeneous equation specifying that the Q's are probabilities that add up to 1:

$$2Q(0,0) + 2Q(0,1) + 4Q(0,2) + 4Q(2,0) + 2Q(2,2) + 2Q(2,3) = 1. \tag{18}$$

Second, determine the homogeneous equations associated with the self-consistency for the first $N_Q(L) - 1$ non-equivalent Q's. This then leads to the following system of equations:

$$\begin{bmatrix} \mathbf{2} & \mathbf{2} & \mathbf{4} & \mathbf{4} & \mathbf{2} & \mathbf{2} \\ (\frac{\mathbf{1}}{\mathbf{2}}\bar{\boldsymbol{r}}^{\boldsymbol{f}} - \mathbf{1}) & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{2}}\bar{\boldsymbol{r}}^{\boldsymbol{f}} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{r}^{\boldsymbol{f}} & -\mathbf{1} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{r}^{\boldsymbol{f}} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{1}}{\mathbf{4}} & -\frac{\mathbf{3}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{4}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{4}} \\ \frac{\mathbf{1}}{\mathbf{2}}\bar{\boldsymbol{r}}^{\boldsymbol{m}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{4}} & (\frac{\mathbf{1}}{\mathbf{2}}\bar{\boldsymbol{r}}^{\boldsymbol{m}} - \mathbf{1}) & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{r}^{\boldsymbol{m}} & \mathbf{0} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{4}} & \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{r}^{\boldsymbol{m}} & -\mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{(\mathbf{0,0})} \\ \boldsymbol{Q}_{(\mathbf{0,1})} \\ \boldsymbol{Q}_{(\mathbf{0,2})} \\ \boldsymbol{Q}_{(\mathbf{2,0})} \\ \boldsymbol{Q}_{(\mathbf{2,2})} \\ \boldsymbol{Q}_{(\mathbf{2,3})} \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{19}$$

For the matrix elements in this system of equations, we have used the notation $\bar{r} = 1 - r$ to designate the complementary value of the recombination rate, such a notation allowing for more compact expressions. The linear system Eq. 19 can be solved by hand, leading to

$$Q(0,0) = \frac{-\frac{1}{2}r^f + \frac{1}{2}r^m + 1}{4(3r^f + 3r^m + 1)} \qquad Q(0,1) = \frac{3r^f + r^m}{8(3r^f + 3r^m + 1)}$$

$$Q(0,2) = Q(2,0) = \frac{r^f + r^m}{4(3r^f + 3r^m + 1)} \qquad Q(2,2) = \frac{\frac{r^f}{2} - \frac{r^m}{2} + 1}{4(3r^f + 3r^m + 1)}$$

$$Q(2,3) = \frac{r^f + 3r^m}{8(3r^f + 3r^m + 1)}.$$

Note that except for $Q(0,2)$ and $Q(2,0)$, all the Q's are *asymmetric* functions of r^f and r^m. Furthermore, the equality $Q(0,2) = Q(2,0)$ follows from the special symmetry of replacing the left-right convention that orients chromosomes by one using the right-left orientation.

Given the non-trivial result of Eq. 20, we can ask what is the consequence for R, the RIL recombination rate. The calculation is straightforward:

$$\begin{aligned} R &= Q(0,1) + Q(0,3) + Q(1,0) + Q(1,2) + Q(2,1) + Q(2,3) + Q(3,0) + Q(3,2) \\ &= 2(Q(0,1) + Q(0,2) + Q(2,0) + Q(2,3)) \\ &= \frac{2(r^f + r^m)}{3(r^f + r^m) + 1}. \end{aligned}$$

Interestingly, this result depends only on the mean of the female and male recombination rates, in spite of the fact that such a property does not hold at the level of the individual Q's. Furthermore, it shows that the Haldane-Waddington relation (Eq. 1) can be used when recombination rates are sex-dependent if in that formula the (sex-independent) recombination rate is replaced by the sex-averaged recombination rate.

Although this example was very simple (it involved only two loci), it should be clear that our framework is generally applicable, for any number of loci, whether the female and male recombination rates are identical or not.

3.2 Case of three loci

Haldane and Waddington showed that the probabilities of two-locus RIL genotypes may be used to derive the probabilities of the three-locus RIL genotypes. Teuscher and Broman also provided this result when they introduced their approach (Teuscher & Broman (2007); Broman (2005)). In the introduction we explained why such a relation holds and so one might expect a similar conclusion to hold for the Q's, but this is not so. Indeed, for this $L = 3$ case, as mentioned in Section 2.1, there are $N_Q(L) = 10$ unknown Q's to determine, corresponding to 9 degrees of freedom, but the information from the $L = 2$ level only provides 6 constraints, two for each pair of loci ($6 = 2 \times 3$).

To determine the values of all the IBD probabilities, we simply apply our framework when using $L = 3$. We begin by specifying the set of non-equivalent Q's that are our unknowns, following the logic of the general case as exposed in Section 2.1. We thus choose $Q(0,0,0)$, $Q(0,0,1)$, $Q(0,0,2)$, $Q(0,1,0)$, $Q(0,1,1)$, $Q(0,1,2)$, $Q(0,2,0)$, $Q(0,2,1)$, $Q(0,2,2)$, and $Q(0,2,3)$. Second, we write the single inhomogeneous equation that sums all Q's (before applying equivalences). Third, we construct the self-consistent equations for the first 9 of our non-equivalent Q's, assuming no genetic interference. The Supplementary Material provides a graphical representation of the $T\,[u \to v]$ entries to be explicit, our R code constructs this matrix automatically. These successive steps lead to the following linear system for our 10

unknowns:

$$\begin{bmatrix} 4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8 \\ \bar{r}_{12}\bar{r}_{23}-1 & 0 & \frac{\bar{r}_{12}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{23}}{2} & 0 \\ r_{23}\bar{r}_{12} & -1 & \frac{\bar{r}_{12}}{2} & 0 & 0 & 0 & \frac{r_{13}}{2} & 0 & \frac{r_{23}}{2} & 0 \\ 0 & \frac{\bar{r}_{12}}{2} & \frac{\bar{r}_{12}-2}{2} & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ r_{12}r_{23} & 0 & \frac{r_{12}}{2} & -1 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{r_{23}}{2} & 0 \\ r_{12}\bar{r}_{23} & 0 & \frac{r_{12}}{2} & 0 & -1 & 0 & \frac{r_{13}}{2} & 0 & \frac{\bar{r}_{23}}{2} & 0 \\ 0 & \frac{r_{12}}{2} & \frac{r_{12}}{2} & 0 & 0 & -1 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{1}{4} & \frac{\bar{r}_{13}-2}{2} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{r_{13}}{2} & 0 & \frac{1}{4} & \frac{r_{13}}{2} & -1 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{2} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{\bar{r}_{23}-2}{2} & 0 \end{bmatrix} \begin{bmatrix} Q(0,0,0) \\ Q(0,0,1) \\ Q(0,0,2) \\ Q(0,1,0) \\ Q(0,1,1) \\ Q(0,1,2) \\ Q(0,2,0) \\ Q(0,2,1) \\ Q(0,2,2) \\ Q(0,2,3) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{20}$$

where as before $\bar{r} = 1 - r$ denotes the complementary value of the recombination rate and $r_{ij} = r_{i,j}$. The solution of Eq. 20 can be obtained either numerically or analytically – that is as an explicit function of the three recombination rates – using e.g., Maple or Mathematica since a treatment by hand would be very tedious.

3.3 Four and more loci

The previous methodology can be extended to more loci but quickly becomes too cumbersome to manage manually. For illustration, in the case $L = 4$, there are 40 Q's to determine (*cf.* Eq. 2). The system of 40 linear inhomogeneous equations determining these unknowns is given in Eq. 21 and barely fits on one page as a figure.

In that display including a 40×40 matrix, we have used the same compact notation as for $L = 3$. Our software produces this system of equations and then can solve for the Q's for any particular values of the r_{ij}. Computing the corresponding probabilities of RIL genotypes is then straightforward and in practice the computer does this very quickly.

It is of course possible to go to larger values of L but then it becomes unweildly to show the corresponding matrix. As expected, the computation time required by our R code grows fast with L, by about a factor 16 for each unit increase of L. The required computer memory also grows in the same way. At $L = 8$ the code takes about 5 minutes to solve the problem, and for still larger values of L it is best to use a server with large memory capacity (we have gone up to $L = 10$).

$$
\left[\begin{array}{cccccccccccccccccccccccccccccccccccc}
4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8 & 4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\
\bar{r}_{12}\bar{r}_{23}\bar{r}_{34}-1 & 0 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{34}\bar{r}_{12}\bar{r}_{23} & -1 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{12}r_{24}}{2} & 0 & \frac{\bar{r}_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{r_{24}\bar{r}_{13}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{r_{34}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & \frac{\bar{r}_{12}\bar{r}_{23}}{2}-1 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
r_{23}r_{34}\bar{r}_{12} & 0 & \frac{r_{23}\bar{r}_{12}}{2} & -1 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{34}\bar{r}_{12}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{23}\bar{r}_{12}\bar{r}_{34} & 0 & \frac{r_{23}\bar{r}_{12}}{2} & 0 & -1 & 0 & \frac{r_{24}\bar{r}_{12}}{2} & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{23}\bar{r}_{12}}{2} & \frac{r_{23}\bar{r}_{12}}{2} & 0 & 0 & -1 & 0 & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}}{4} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}}{4} & 0 & \frac{r_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{\bar{r}_{12}}{4} & \frac{\bar{r}_{12}\bar{r}_{24}}{2}-1 & 0 & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & \frac{r_{24}\bar{r}_{12}}{2} & 0 & \frac{\bar{r}_{12}}{4} & \frac{r_{24}\bar{r}_{12}}{2} & -1 & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & \frac{\bar{r}_{12}\bar{r}_{34}}{2}-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{12}}{2} & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & \frac{r_{34}\bar{r}_{12}}{2} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{r_{34}}{4} \\
r_{23}r_{12}\bar{r}_{34} & 0 & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}\bar{r}_{34}}{2} & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{24}\bar{r}_{13}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{12}r_{23}r_{34} & 0 & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}r_{34}}{2} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{12}r_{23}}{2} & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & 0 & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}}{4} & 0 & \frac{r_{23}}{4} & 0 & 0 & 0 & 0 \\
r_{12}\bar{r}_{23}r_{34} & 0 & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{34}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{12}\bar{r}_{23}\bar{r}_{34} & 0 & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{12}\bar{r}_{23}}{2} & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}}{4} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}}{4} & \frac{r_{12}r_{24}}{2} & 0 & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}}{4} & \frac{r_{12}\bar{r}_{24}}{2} & 0 & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{34}}{2} & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & \frac{r_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & \frac{r_{12}r_{34}}{2} & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & \frac{r_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2}-1 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{34}}{4} & 0 & \frac{r_{34}\bar{r}_{13}}{2} & -1 & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{13}}{4} & \frac{\bar{r}_{13}\bar{r}_{24}}{2}-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{24}\bar{r}_{13}}{2} & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & \frac{r_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{13}}{4} & \frac{r_{24}\bar{r}_{13}}{2} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{34}}{4} & 0 & \frac{r_{13}r_{34}}{2} & 0 & 0 & \frac{r_{13}}{4} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & 0 & \frac{r_{13}}{4} & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{24}}{2} & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{r_{13}}{4} & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{24}}{2} & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & \frac{r_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{r_{13}}{4} & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{\bar{r}_{23}}{4} & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{23}\bar{r}_{14}}{2}-1 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{\bar{r}_{23}}{4} & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{14}\bar{r}_{23}}{2} & -1 & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{24}}{4} & \frac{\bar{r}_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & \frac{\bar{r}_{24}}{4} & 0 & 0 & \frac{\bar{r}_{23}}{4} & \frac{\bar{r}_{23}\bar{r}_{34}}{2}-1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{23}}{2} & \frac{\bar{r}_{23}}{4} & 0 & \frac{r_{24}}{4} & \frac{r_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & \frac{r_{24}}{4} & 0 & 0 & \frac{\bar{r}_{23}}{4} & \frac{r_{34}\bar{r}_{23}}{2} & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}}{4} & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{23}\bar{r}_{14}}{2} & 0 & 0 & \frac{r_{23}}{4} & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}}{4} & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{23}r_{14}}{2} & 0 & 0 & \frac{r_{23}}{4} & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{34}}{2} & \frac{r_{23}}{4} & 0 & \frac{\bar{r}_{24}}{4} & \frac{r_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & \frac{\bar{r}_{24}}{4} & 0 & 0 & \frac{r_{23}}{4} & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & -1 & 0
\end{array}\right]
\begin{bmatrix}
Q(0,0,0,0) \\ Q(0,0,0,1) \\ Q(0,0,0,2) \\ Q(0,0,1,0) \\ Q(0,0,1,1) \\ Q(0,0,1,2) \\ Q(0,0,2,0) \\ Q(0,0,2,1) \\ Q(0,0,2,2) \\ Q(0,0,2,3) \\ Q(0,1,0,0) \\ Q(0,1,0,1) \\ Q(0,1,0,2) \\ Q(0,1,1,0) \\ Q(0,1,1,1) \\ Q(0,1,1,2) \\ Q(0,1,2,0) \\ Q(0,1,2,1) \\ Q(0,1,2,2) \\ Q(0,1,2,3) \\ Q(0,2,0,0) \\ Q(0,2,0,1) \\ Q(0,2,0,2) \\ Q(0,2,0,3) \\ Q(0,2,1,0) \\ Q(0,2,1,1) \\ Q(0,2,1,2) \\ Q(0,2,1,3) \\ Q(0,2,2,0) \\ Q(0,2,2,1) \\ Q(0,2,2,2) \\ Q(0,2,2,3) \\ Q(0,2,3,0) \\ Q(0,2,3,1) \\ Q(0,2,3,2) \\ Q(0,2,3,3)
\end{bmatrix}
=
\begin{bmatrix}
1 \\ 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{bmatrix}
\tag{21}
$$

4 DISCUSSION

The construction of RILs involves successive generations of inbreeding until all alleles are fixed. The probabilities of the multilocus genotypes encountered across successive generations can be followed by recursion equations (Haldane & Waddington (1931); Hospital *et al.* (1996)) which in the case of SIB mating are specified by a dense matrix of size $16^L \times 16^L$ if one has L loci. For the goal of obtaining the probabilities of RIL genotypes (the expected frequency of occurrence when averaging over a large number of repeats), the difficulty is that fixation formally requires an infinite number of generations. Thus, either the recursions must be taken "sufficiently far" to obtain *numerical* convergence or a mathematical trick has to be found. For $L = 2$, Haldane and Waddington succeeded in the second path thanks to much mathematical ingenuity, and interestingly, that $L = 2$ solution automatically determines the probabilities in the $L = 3$ case. However, since that founding work – going back to 1931 – no solution had been proposed to tackle SIB RILs with $L = 4$ or more.

Using a novel method, we have successfully overcome that long-standing challenge here. Our approach provides an algebraic solution, albeit at a computational cost that grows roughly as 16^L for L loci. That exponential growth rate is far less drastic than that of the original proposition of Haldane and Waddington of 1931, so that not only did we break the $L = 4$ barrier but in fact we were able to rather easily treat L's up to 8. We also pointed out that our framework can deal with different female and male recombination rates, a situation that seems to have never been considered before in the context of SIB RILs, even for $L = 2$.

The ability to compute probabilities of RIL multilocus genotypes opens up to a number of applications. For instance, when building genetic maps, the ordering of markers is determined by comparing likelihoods of different orderings. That calculation can now be done using exact rather than approximate multilocus genotype frequencies, putting those mapping algorithms on a more solid footing. Similarly, when RIL genotypes must be inferred because of missing data, determining the most likely value of an allele requires comparing multilocus genotype probabilities. Finally, beyond specific uses in the case of RILs, our framework that exploits self-consistency might be useful in certain population genetics problems involving an infinite number of generations.

AUTHORS CONTRIBUTIONS

OM proposed the project and with MP conceived and implemented a first approach. KJ introduced the analytic formulation and this led to major enhancements to the algorithmic KJ and MP developed the R scripts and all authors wrote, edited and approved the manuscript.

FUNDING

This work has benefited from a French State grant (LabEx Saclay Plant Sciences-SPS, ANR-10-LABX-0040-SPS), managed by the French National Research Agency under an "Investments for the Future" program (ANR-11-IDEX-0003-02) which funded the salary of KJ.

Also, the public Ph.D. grant from the French National Research Agency (ANR) as part of the Investissement d'Avenir program, through the Initiative Doctoral Interdisciplinaire (IDI) 2015 project funded by the Initiative d'Excellence (IDEX) Paris-Saclay, ANR-11-IDEX-0003-02 funded the salary of MP.

ACKNOWLEDGMENT

The authors are grateful to Prof. D. de Vienne and C. Dillmann for insightful comments.

SOFTWARE AVAILABILITY

R code implementing the methodology described in this paper is available online at https://github.com/olivier-c-martin/PMG_SIB_RILs.git

SUPPLEMENTARY MATERIAL

The supplementary material contains two parts: a mathematical proof of Eq.2 from the main text, and the graphical representations of the self-consistent equations for the $L = 3$ case. The Supplementary Material for this article can be found online at: https://github.com/olivier-c-martin/PMG_SIB_RILs.git. The R code is also included as a separate supplementary file.

REFERENCES

Broman, K. W. 2005. The Genomes of Recombinant Inbred Lines. *Genetics*, **169**(2), 1133–1146.

Buckler, Edward S., Holland, James B., Bradbury, Peter J., Acharya, Charlotte B., Brown, Patrick J., Browne, Chris, Ersoz, Elhan, Flint-Garcia, Sherry, Garcia, Arturo, Glaubitz, Jeffrey C., Goodman, Major M., Harjes, Carlos, Guill, Kate, Kroon, Dallas E., Larsson, Sara, Lepak, Nicholas K., Li, Huihui, Mitchell, Sharon E., Pressoir, Gael, Peiffer, Jason A., Rosas, Marco Oropeza, Rocheford, Torbert R., Romay, M. Cinta, Romero, Susan, Salvo, Stella, Villeda, Hector Sanchez, Silva, H. Sofia da, Sun, Qi, Tian, Feng, Upadyayula, Narasimham, Ware, Doreen, Yates, Heather, Yu, Jianming, Zhang, Zhiwu, Kresovich, Stephen, & McMullen, Michael D. 2009. The Genetic Architecture of Maize Flowering Time. *Science*, **325**(5941), 714–718.

Crow, James F. 2007. Haldane, Bailey, Taylor and Recombinant-Inbred Lines. *Genetics*, **176**(2), 729–732.

Darvasi, A., & Soller, M. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, **141**(3), 1199–1207.

El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J., Raz, V., & Koornneef, M. 2001. A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. *Nature Genetics*, **29**(4), 435–440.

Haldane, J. B. S., & Waddington, C. H. 1931. Inbreeding and Linkage. *Genetics*, **16**(4), 357–374.

Haldane, J. S., Meakins, J. C., & Priestley, J. G. 1919. The effects of shallow breathing. *The Journal of Physiology*, **52**(6), 433–453.

Hospital, F., Dillmann, C., & Melchinger, A. E. 1996. A general algorithm to compute multilocus genotype frequencies under various mating systems. *Computer applications in the biosciences: CABIOS*, **12**(6), 455–462.

Ihaka, Ross, & Gentleman, Robert. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Lander, E. S., & Schork, N. J. 1994. Genetic dissection of complex traits. *Science (New York, N.Y.)*, **265**(5181), 2037–2048.

Rockman, Matthew V., & Kruglyak, Leonid. 2008. Breeding designs for recombinant inbred advanced intercross lines. *Genetics*, **179**(2), 1069–1078.

Samal, Areejit, & Martin, Olivier C. 2015. Statistical Physics Methods Provide the Exact Solution to a Long-Standing Problem of Genetics. *Physical Review Letters*, **114**(23), 238101.

Teuscher, Friedrich, & Broman, Karl W. 2007. Haplotype Probabilities for Multiple-Strain Recombinant Inbred Lines. *Genetics*, **175**(3), 1267–1274.

Walsh, Bruce, & Lynch, Michael. 2018. *Evolution and Selection of Quantitative Traits*. Sunderland, Massachusetts, USA: Oxford University Press.

Weir, Bruce S. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. 2 sub edition edn. Sunderland, Mass: Sinauer Associates is an imprint of Oxford University Press.

Winkler, Christopher R, Jensen, Nicole M, Cooper, Mark, Podlich, Dean W, & Smith, Oscar S. 2003. On the determination of recombination rates in intermated recombinant inbred populations. *Genetics*, **164**(2), 741–745.