Journal of Genetics, Vol. 96, No. 5, November 2017, pp. 795–800
https://doi.org/10.1007/s12041-017-0837-0

HALDANE AT 125

Haldane, Waddington and recombinant inbred lines: extension of their work to any number of genes

AREEJIT SAMAL[1]* and OLIVIER C. MARTIN[2]*

[1] *The Institute of Mathematical Sciences, Homi Bhabha National Institute, Chennai 600 113, India*
[2] *GQE-Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France*
*E-mail: Areejit Samal, asamal@imsc.res.in; Olivier C. Martin, olivier.c.martin@inra.fr.

Published online 24 November 2017

Abstract. In the early 1930s, J. B. S. Haldane and C. H. Waddington collaborated on the consequences of genetic linkage and inbreeding. One elegant mathematical genetics problem solved by them concerns recombinant inbred lines (RILs) produced via repeated self or brother–sister mating. In this classic contribution, Haldane and Waddington derived an analytical formula for the probabilities of 2-locus and 3-locus RIL genotypes. Specifically, the Haldane–Waddington formula gives the recombination rate R in such lines as a simple function of the per generation recombination rate r. Interestingly, for more than 80 years, an extension of this result to four or more loci remained elusive. In 2015, we generalized the Haldane–Waddington self-mating result to any number of loci. Our solution used self-consistent equations of the multi-locus probabilities 'for an infinite number of generations' and solved these by simple algebraic operations. In practice, our approach provides a quantum leap in the systems that can be handled: the cases of up to six loci can be solved by hand while a computer program implementing our mathematical formalism tackles up to 20 loci on standard desktop computers.

Historical overview

To explain the mathematical genetics problem posed in 1931 by Haldane and Waddington, it is necessary to first define recombinant inbred lines (RILs). In genetics, inbred lines (or inbred strains) are animals or plants which have undergone inbreeding to the point where their genomes are fixed, i.e., within each pair of homologous chromosomes, the two chromosomes are identical. As a result, an inbred line is 'homozygous' for each gene and will produce gametes that have no genetic variation, a major advantage for genetic research (Crow 2007). RILs are typically produced by starting with two distinct inbred lines (the parents), crossing them, and then generating new lines by inbreeding over many generations. Figure 1 illustrates such a process in plants that are able to self-pollinate so that a single individual can play the role of both male and female, and thus produce offspring on its own. Note that each individual is 'diploid', i.e., its chromosomes come in pairs (figure 1). The chromosomes in the resulting lines are mosaics of those of the two parents. Because of the associated genetic shuffling, RILs provide powerful ways to relate genetic variation (presence of one of the parental alleles versus the other) to phenotypes such as resistance to pathogens, fertility, height, etc. (Lynch and Walsh 1998).

The realization of mosaic chromosomes in RILs depends on the formation of meiotic crossovers during successive generations. Specifically, consider two loci, i.e., two positions on a chromosome. Let parent P (respectively P') have the allele A_1 (respectively a_1) at the first locus and the allele A_2 (respectively a_2) at the second locus. Because these 'diploid' parents are homozygous, they produce 'haploid' gametes that have no genetic variation. A male and a female gamete need to fuse to produce the children of the next generation. In our case here, each child will have a pair of homologous chromosomes, the first carrying alleles A_1A_2 and the second carrying alleles a_1a_2, and the 2-locus 'genotype' is denoted by A_1A_2/a_1a_2. Note that all children in this F_1 generation are genetically identical. However, when going to the next generation F_2, the chromosome carried by a gamete can be of a parental type, i.e. A_1A_2 or a_1a_2, or it can be 'recombinant', A_1a_2 or a_1A_2 because of the presence of a crossover between the two loci during meiosis (Griffiths *et al.* 1999). The probability of recombination between two loci during meiosis is generally denoted by r. The problem formulated by Haldane and Waddington consists in determining the *RIL* recombina-

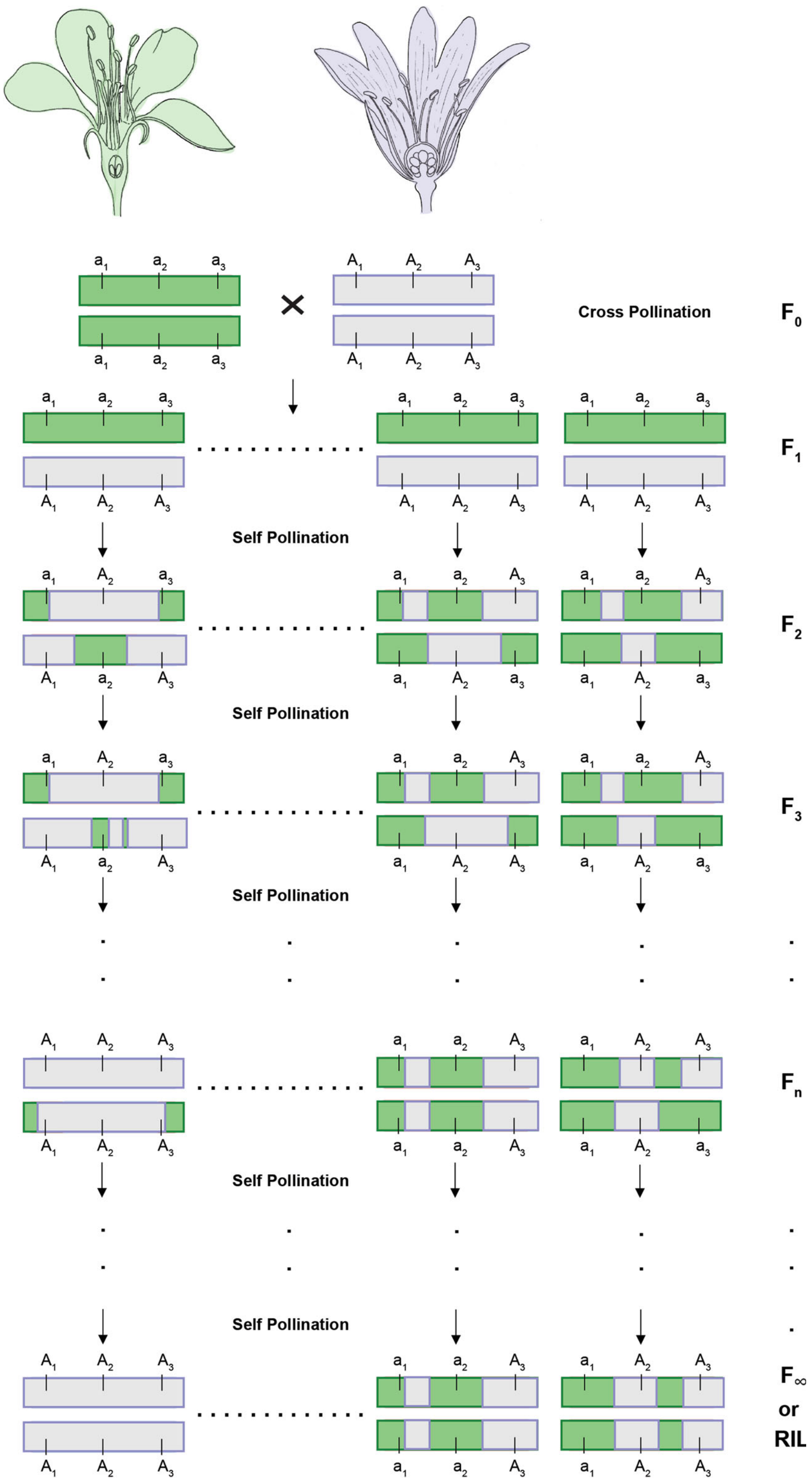

Figure 1. Generation of multiple recombinant inbred lines in self-pollinating plants. In F_0 generation, two homozygous parents are crossed to produce individuals in the next F_1 generation. Subsequently, each individual self mates, producing one gamete via female meiosis and one gamete via male meiosis, and the two gametes from the same individual are fused together to produce the individual of the next generation. Meiotic recombination leads to intrachromosomal shuffling of allelic content and ultimately to fixed mosaic chromosomes.

tion rate R, i.e., its value at the level of the RIL descendants (which formally arises after an infinite number of generations of inbreeding) in terms of the rate per meiosis r. In the context of figure 1 where each individual is both male and female, the celebrated 'Haldane-Waddington' formula (Haldane and Waddington 1931), is simply

$$R = \frac{2r}{1+2r}. \tag{1}$$

In their 1931 paper, Haldane and Waddington wrote the linear equations giving the frequencies of all the 2-locus genotypes at one generation given the frequencies at the previous generation. Naively, since each chromosome of the considered homologous pair can have any of four possible combinations of alleles, there are 16 different frequencies to consider which by symmetry can be reduced to 5. Then the recursion from one generation to the next corresponds to applying a 5 × 5 matrix to the reduced vector of 5 frequencies. Thus, the frequencies of all 2-locus RIL genotypes are obtained by applying the 5 × 5 matrix an infinite number of times to an initial vector. Haldane and Waddington obtained these limiting frequencies by clever algebraic manipulations. In a more general context, one may instead put such matrices into a canonical form by a change of basis. In the canonical form, it is simple to raise the matrix to any power and thereby extract the limit of an infinite number of generations. What is more striking is that Haldane and Waddington were able to use their algebraic manipulations to treat also the case of RILs in which one mates a brother to a sister at each generation (Haldane and Waddington 1931). In this case of brother–sister mating, the number of 2-locus genotypes to consider is significantly higher: naively there are 256 genotypes but using symmetries the number can be reduced to 22. Nevertheless, the task of exploiting these recursions is arduous! The reader may be amused by what Crow (2007) says about Waddington after this work:

> I wonder if experience with the exhausting and tedious algebra in this article sensitized him forever against any further work in this field.

In contrast, Haldane very characteristically enjoyed delving into such complex mathematics. It is worth mentioning two points of potential interest to historians here. First, as stressed by James F. Crow (Crow 2007), the term 'recombinant' inbred line—an extension of inbred lines—was only coined two decades after the work of of Haldane and Waddington (Haldane and Waddington 1931). This could justify why Fisher in his later book on the mathematical treatment of inbreeding (Fisher 1949) did not survey the work of previous contributors to this field, not even that of Wright or Haldane (Bartlett 1950; Lush 1950). Second, given that in his 1949 book Fisher provides only three references for which he is not an author, it seems quite inevitable that he would not cite the Haldane and Waddington paper of 1931.

Extension of the challenge to multi-locus genotypes

Computational complexity of the Haldane–Waddington approach and short cut for 3 loci

The framework used by Haldane and Waddington generalizes to any number of loci. Hereafter we focus on the case of RILs produced by selfing as is possible with plants. If one considers L loci, there are 4^L different diploid genotypes and thus that many frequencies to follow in a recursion. For the case of $L = 2$, Haldane and Waddington used symmetries to reduce the number of unknowns from 16 to 5, but for larger L exploiting such symmetries becomes unmanageable. Thus, in practice one has to consider the recursion equations for all 4^L unknowns and the problem can be solved by bringing the associated $4^L \times 4^L$ recursion matrix into a canonical form. But to transform the $4^L \times 4^L$ recursion matrix into that canonical form requires on the order of 4^{3L} operations. For $L = 4$ this transformation corresponds to over 16 million operations and for $L = 6$ to over 68 billion (68×10^9) operations, and thus producing a 'formula' is out of the question.

In spite of the pessimism emerging from such massive numbers, a formula was nevertheless provided by Haldane and Waddington in 1931 for the case of $L = 3$ (Haldane and Waddington 1931). The formula for $L = 3$ follows quite simply by noticing that the $2^L = 8$ different fixed (RIL) genotypes come in pairs of equal frequencies, and thus there are only four unknown frequencies. Further, since these add up to 1, one really has just 3 independent unknowns. These can be simply determined in terms of the 3 RIL recombination rates R_{ij} between loci i and j (running from 1 to 3). For instance, R_{13} is the sum of the frequencies of genotypes in which just one of the two adjacent intervals (1, 2) and (2, 3) are recombinant. Building on this optimistic note, some authors considered that the trick for $L = 3$ extended to $L = 4$ and beyond, but this is not true, a piece of information is missing as justified below (Samal and Martin 2015).

Tackling the 4-locus case: a strategy based on self-consistency

In the 4-locus case there are $2^4 = 16$ different RIL genotypes. Due to underlying symmetry, these 16 different RIL genotypes come in pairs of equal frequencies, reducing the number of unknown frequencies to 8. We need 8 independent equations to determine these unknowns. One equation follows from the fact that the frequencies sum to 1. Further, by using the 2-locus RIL recombination rates R_{ij} between pairs of loci, one obtains $4\,(4-1)/2 = 6$ additional equations. Thus, we need one more equation. Uncovering this missing information was the stumbling block to extending the Haldane–Waddington result to four or more loci. In a recent contribution (Samal and Martin 2015), we have solved this challenge, deriving the exact probabilities of RIL genotypes with 'any' number of loci, using two concepts borrowed from theoretical physics: Glauber's formula (Glauber 1963) and Schwinger–Dyson equations (Dyson 1949; Schwinger 1951) that are self-consistent relations between moments of multivariate distributions.

Our framework based on moments and Glauber's formula

In their 1931 derivation of the analytical formulas for the probabilities of 2-locus and 3-locus RIL genotypes, Haldane and Waddington used the exact inbreeding recursion equations that are then iterated an infinite number of times. However such an approach becomes too cumbersome even for the 4-locus case. In contrast, our method solves self-consistent equations to obtain the multi-locus RIL probabilities 'directly at the infinite generation limit'. We use a notation inspired by Slatkin (1972) to represent a L-locus RIL genotype that is homozygous at every locus via a vector S of 'spin' variables S_i, $i = 1, 2, \ldots, L$ where $S_i = 1$ if locus i is a_i/a_i and $S_i = -1$ if locus i is A_i/A_i. This notation is very convenient for expressing the probability of any RIL genotype S in terms of expectation of spin products as follows. If we consider a single locus i, the probability that the spin S_i has value s_i is given by, $P(S_i = s_i) = E[(1 + s_i S_i)/2]$, where the average is taken over the distribution of the random variable S_i. More importantly, one can use the generalization due to Glauber (Glauber 1963) to determine the probability of a L-locus RIL genotype via:

$$P(S_1 = s_1, S_2 = s_2, \ldots, S_L = s_L) = E\left[\left(\frac{1+s_1S_1}{2}\right)\left(\frac{1+s_2S_2}{2}\right)\cdots\left(\frac{1+s_LS_L}{2}\right)\right], \quad (2)$$

where

$$E\left[\left(\frac{1+s_1S_1}{2}\right)\left(\frac{1+s_2S_2}{2}\right)\cdots\left(\frac{1+s_LS_L}{2}\right)\right]$$

is the average over all possible RIL genotypes (Samal and Martin 2015). We remark that Glauber's formula is exact, the spins S_i at different loci need not be independent.

From equation 2 it follows that the challenge of obtaining the probabilities of L-locus RIL genotypes is solved if one can determine the expectation values of all spin products. Specifically, for the case of 4 loci ($L = 4$), equation 2 becomes:

$$P(S_1 = s_1, S_2 = s_2, S_3 = s_3, S_4 = s_4) = \frac{1}{16}\left(1 + \sum_{i<j} s_i s_j E\left[s_i s_j\right] + s_1 s_2 s_3 s_4 E\left[s_1 s_2 s_3 s_4\right]\right), \quad (3)$$

where the right-hand side is based on expectations of 2-spin and 4-spin products. Note that in Glauber's formula for the 4-locus case, we have used the property that the expectation of a product with an odd number of spins vanishes due to the global invariance $P(S) = P(-S)$ which corresponds to exchange of a's and A's in RIL genotypes.

Self-consistency based on Schwinger–Dyson equations

As an illustrative example, consider calculating the expectation of a 4-locus product. The Schwinger–Dyson formalism equates the value of that multivariate moment to a sum of terms, each being a weight times the expectation of a k-locus product (k ranging from 0 to 4). Under the hypothesis that meiotic crossovers are noninterfering (Haldane 1919), these weights can be written 'directly' in terms of the 2-locus recombination rates (Samal and Martin 2015). The resulting self-consistent 'Schwinger–Dyson' equation (Dyson 1949; Schwinger 1951) for the expectation of the 4-spin product is:

$$\begin{aligned} E\left[S_1S_2S_3S_4\right] = {} & \frac{(1-2r_1)\left((1-r_2)^2+r_2^2\right)(1-2r_3)}{2} \\ & \cdot \left(E\left[S_1S_2S_3S_4\right]+1\right) \\ & + \frac{(1-2r_1)\left(2\left(1-r_2\right)r_2\right)(1-2r_3)}{2} \\ & \cdot \left(E\left[S_1S_2\right]+E\left[S_3S_4\right]\right), \end{aligned} \quad (4)$$

From equation 4 one easily extracts the explicit expression for the unknown, and from that, using Glauber's equation, one has the value of all 4-locus RIL genotype probabilities. Such 'by-hand' calculations remain feasible up to $L = 6$ loci. In the supplemental material of Samal and Martin (2015), we similarly give all the weights of the Schwinger–Dyson equation for 6 loci and from there it is easy to obtain all corresponding RIL genotype probabilities using Glauber's formula.

Algorithm and computational aspects

Our formalism shows that one can determine all the moments and thus all the probabilities of RIL genotypes—for any number of loci—by purely algebraic manipulations involving only the four standard operations of addition, subtraction, multiplication and division. Our framework thus extends Haldane and Waddington's 1931 analytic result for RILs generated by selfing. Nevertheless, such operations rapidly become tedious for increasing L. We have thus written a computer program which performs all the calculations numerically (Samal and Martin 2015). The associated algorithm first calculates all the moments recursively: it starts with all $L\,(L-1)/2$ 2-locus moments and then iterates in k to calculate the k-locus moments (k even). For each recursion step, the Schwinger–Dyson equation is used to determine the k-locus moments in terms of previously calculated lower-order moments. Once all moments up to $k = L$ are computed, Glauber's formula is used to obtain all 2^L frequencies of RIL genotypes

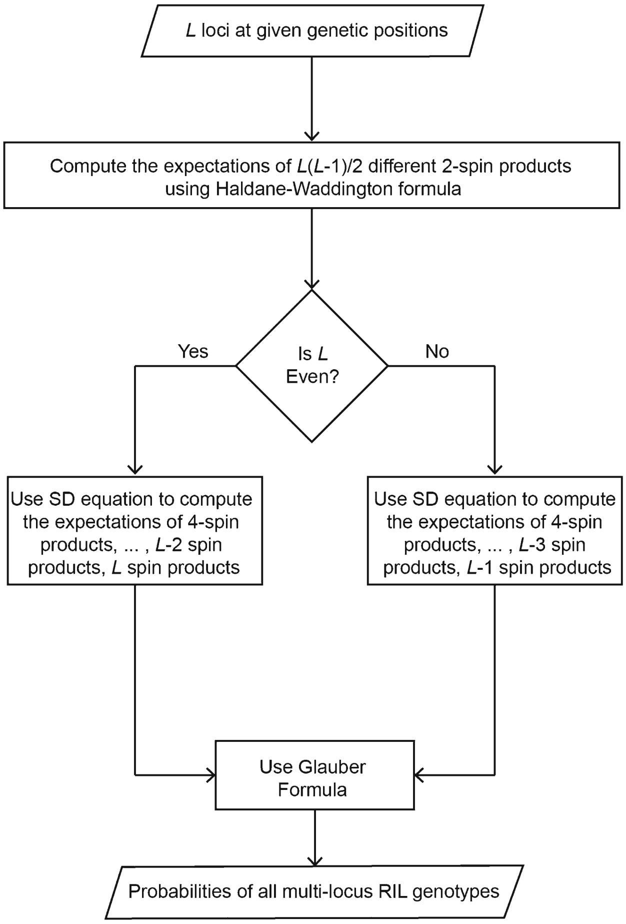
L loci at given genetic positions
Compute the expectations of *L*(*L*-1)/2 different 2-spin products using Haldane-Waddington formula
Is *L* Even?
Yes
No
Use SD equation to compute the expectations of 4-spin products, ... , *L*-2 spin products, *L* spin products
Use SD equation to compute the expectations of 4-spin products, ... , *L*-3 spin products, *L*-1 spin products
Use Glauber Formula
Probabilities of all multi-locus RIL genotypes

Figure 2. Flow chart summarizing our algorithm to compute the probabilities of RIL genotypes with any number of loci.

(figure 2). This is most efficiently done using the multidimensional transform because, like the fast Fourier transform, it requires only $O(N\ln(N))$ operations rather than $O(N^2)$ when transforming a vector of N entries. With this computer program, it is possible to calculate numerically all the 2^L frequencies of RIL genotypes for L up to about 20 on a standard desktop computer.

Discussion and conclusions

In this work, we have focussed on genetic loci, treated as discrete positions on chromosomes in the same manner as Haldane and Waddington (1931). Such a framework may be contrasted with that used in the theory of junctions (Fisher 1949, 1954) which considers 'continuous' positions along chromosomes. That framework has been extensively used in the past two decades to study questions such as identity of descent (Stefanov 2000). Such shared genomic pieces or 'blocks' have applications in human genetics and studies of heritable diseases. In an earlier work (Martin and Hospital 2011), one of the present authors has also used the theory of junctions but to study the statistical properties of the mosaic genomes arising in RILs. In that work, the theory of junctions was used to calculate properties such as number of blocks, size of blocks, and the correlation in size of two successive blocks. However, studying RIL genotypes at particular loci using the framework of junctions involves multidimensional integrals that are very cumbersome (Martin and Hospital 2011) prohibiting the use of such approaches for more than 3 loci. Thus, for questions involving 4 or more loci it is necessary to use a better adapted framework where loci are treated via discrete positions, as provided in the present work.

The 2-locus and 3-locus formulas obtained in Haldane and Waddington (1931) are routinely used in many RIL contexts. As mentioned before, Haldane and Waddington had to introduce clever mathematical tricks to arrive at these concise results, and perhaps this explains why no 4-locus extension was obtained or even attempted for over 80 years. What is surprising though is that the extension as found here in fact requires no new mathematical tools and our formulas only involve the four basic operations of addition, subtraction, multiplication and division. Thus, our extension of the Haldane–Waddington result is based on an elegant solution, albeit the algebra becomes tedious with an increase in the number of loci. Fortunately, the past decades have opened up the possibility of computers performing these operations automatically. Likely, that possibility would have reconciled Waddington with this type of mathematical genetics, but perhaps Haldane would have been sad to see computers supersede human ingenuity.

Acknowledgements

The authors thank M. Karthikeyan for help in preparing the figures and the reviewers for their suggestions to improve the manuscript.

References

Crow J. F. 2007 Haldane, Bailey, Taylor and recombinant-inbred lines. *Genetics* **176**, 729–732.

Bartlett M. S. 1950 Reviews of statistical and economic books. *J. R. Stat. Soc. Ser. A* **113**, 249–250.

Dyson F. J. 1949 The S matrix in quantum electrodynamics. *Phys. Rev.* **75**, 1736–1755.

Fisher R. A. 1949 *The theory of inbreeding*. Oliver & Boyd, Edinburgh and London.

Fisher R. A. 1954 A fuller theory of "junctions" in inbreeding. *Heredity* **8**, 187–197.

Glauber R. J. 1963 Time-dependent statistics of the Ising model. *J. Math. Phys.* **4**, 294–307.

Griffiths A. J. F., Gelbart W. M., Miller J. H. and Lewontin R. C. 1999 *Modern genetic analysis*. W. H. Freeman, New York.

Haldane J. B. S. 1919 The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series. *J. Genet.* **8**, 291–297.

Haldane J. B. S. and Waddington C. H. 1931 Inbreeding and linkage. *Genetics* **16**, 357–374.

Lynch M. and Walsh B. 1998 *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland.

Lush J. L. 1950 The theory of inbreeding (by Ronald A. Fisher). *Am. J. Hum. Genet.* **2**, 97–100.

Martin O. C. and Hospital F. 2011 Distribution of parental genome blocks in recombinant inbred lines. *Genetics* **189**, 645–654.

Schwinger J. 1951 On the Green's functions of quantized fields. I. *Proc. Natl. Acad. Sci.* **37**, 452–455.

Samal A. and Martin O. C. 2015 Statistical physics methods provide the exact solution to a long-standing problem of genetics. *Phys. Rev. Lett.* **114**, 238101.

Slatkin M. 1972 On treating the chromosome as the unit of selection. *Genetics* **72**, 157–168.

Stefanov V. T. 2000 Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. *Genetics* **156**, 1403–1410.