Genet. Res., Camb. (1980), **35**, *pp.* 131–155
With 10 *text-figures*
Printed in Great Britain

The distribution of the fraction of the genome identical by descent in finite random mating populations

BY P. STAM
Department of Genetics, Agricultural University, 53 Generaal Foulkesweg, 6703 BM, Wageningen, The Netherlands

(*Received* 30 *August* 1979)

SUMMARY

The probability distribution of the heterogenic (non-identical by descent) fraction of the genome in a finite monoecious random mating population has been derived. It was assumed that in any generation the length of both heterogenic and homogenic segments are exponentially distributed. An explicit expression is given for the expected number of 'external junctions' (sites that mark the end of a heterogenic segment) per unit map length in any generation. The latter necessitates the introduction of two higher-order identity relations between three genes, and their recurrence relations. Theoretical results were compared with the outcome of a series of simulation runs (showing a very good fit), as well as with the results predicted by Fisher's 'theory of junctions'. In contrast to Fisher's approach, which only applies when the average heterogeneity is relatively small, the present model applies to any generation.

1. INTRODUCTION

The inbreeding coefficient of an individual is defined as the probability that the homologous genes at a randomly drawn locus are identical by descent (Malécot, 1948). For most breeding systems, including random mating populations of finite size, this probability (F) can be calculated exactly, given the initial conditions, i.e. the probabilities of identity by descent for genes in the base population. The coefficient F represents the expected proportion of the individual's genome which is identical by descent. Since blocks of linked genes rather than individual genes are transmitted from one generation to the next, blocks of linked genes rather than single genes become identical by descent as inbreeding proceeds. The length of transmitted gene blocks (or chromosome segments) is variable; therefore, the proportion of the genome identical by descent in a given generation will vary between individuals, although the expected proportion may be the same for all individuals of that generation. This paper deals with the probability distribution of the fraction of the genome that is identical by descent in any generation of a finite, monoecious random mating population with no selfing.

A problem rather similar to the one indicated above has been treated by Fisher (1949, 1954, 1959) and Bennet (1953, 1954) for full sib mating and parent–offspring mating by means of the 'theory of junctions'. In their terminology a population

0016-6723/80/2828-7780 $01.00 © 1980 Cambridge University Press

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

is *homogenic* at a given locus if *all* the genes of the population at that locus are identical by descent; otherwise the population is *heterogenic*. Since they mainly dealt with populations of size 2, the terms homogenic and heterogenic applied to four homologous genes or chromosome segments. The theory of junctions deals with the fraction of the germ plasm (i.e. the total genetic content of the population) that is heterogenic in a given generation. The analogy between the problem treated in this paper and Fisher's problem is obvious: here we will consider one homologous pair of chromosomes at a time, whereas Fisher considered all chromosome pairs of the population simultaneously.

The basic approach to the problem used in this paper is the same as the one introduced by Fisher (1949). I will therefore briefly outline his method. In the rest of this paper the terms homogenic and heterogenic are used to indicate identity and non-identity by descent of homologous *pairs* of genes or chromosome segments.

At any stage of inbreeding the genome of an individual will consist of alternating heterogenic and homogenic segments ('tracts'). The sites that mark the ends of a heterogenic tract are called *external junctions*, a term that will become clear later on. The basic idea developed by Fisher is as follows. Let the mean number of external junctions per unit map length in a given generation be Z_t, and let H_t be the expected heterogenic proportion of the genome ($H_t = 1-F_t$). Then, for a diploid organism with total genetic map length L, distributed over n chromosome pairs, the mean number of ends of heterogenic tracts equals LZ_t+2nH_t; the second term here represents the $2n$ chromosome termini, of which a fraction H_t is heterogenic. Thus the mean number of heterogenic tracts (m) is

$$m = \tfrac{1}{2}Z_t\,.\,L+nH_t. \tag{1}$$

If the distribution (p_k) of the number of heterogenic tracts is known, and if further the distribution of the lengths of heterogenic tracts is known, then the distribution of the sum of all heterogenic tracts can be obtained. Fisher (1949) and Bennet (1954) made the following basic assumptions about these distributions.

(i) The number of heterogenic tracts is Poisson-distributed, i.e.

$$p_k = e^{-m}m^k/k!. \tag{2}$$

(ii) The length of heterogenic tracts is distributed exponentially, i.e. the probability density function (p.d.f.) is

$$f(t) = ae^{-at}, \quad t > 0, \quad a > 0.$$

The sum of k-independent heterogenic tracts then follows a gamma distribution with p.d.f.

$$f_k(t) = \frac{a^k t^{k-1}e^{-at}}{(k-1)!}.$$

By randomization of the latter distribution with respect to k, according to (2) one obtains

$$\begin{aligned} f(t) &= \sum_{k=1}^{\infty} a^k t^{k-1}e^{-at}e^{-m}m^k/(k-1)!\,k! \\ &= \exp\{-(m+at)\}\sqrt{(ma/t)}\,I_1(2\sqrt{(mat)}), \end{aligned} \tag{3}$$

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

where $I_1(z)$ is the modified Bessel function of the first kind of order one (Bennet, 1954). The distribution has a probability condensation of e^{-m} at $t = 0$, corresponding to the probability of complete homogeneity. The mean of this distribution is m/a, the expected length being heterogenic. Thus

$$m/a = LH_t,$$

or, combined with (1),

$$a = \frac{Z_t}{2H_t} + \frac{n}{L}. \tag{4}$$

The relations (1) and (4) thus completely specify the probability distribution (3), once H_t and Z_t are known. In order to obtain H_t and Z_t (in the sense of Fisher) for full sib mating and parent–offspring mating, Fisher applied the elaborate method of generation matrices.

It will be clear that the calculations leading to the p.d.f. (3) tacitly assume a genome of infinite length. When the average length of heterogenic tracts is short, relative to the total map length, this introduces no serious error. However, in early generations, when heterogenic tracts are still relatively long, the p.d.f. (3) cannot be applied.

This paper presents an extension of Fisher's theory. I will introduce two modifications. First, by using identity relations between genes, instead of a generation matrix, a generalization to any finite population size is obtained. Second, I will assume that at any stage of inbreeding the lengths of both heterogenic and homogenic tracts are distributed exponentially, each with its own parameter. The external junctions can then be regarded as events in two alternating Poisson processes. The probability distribution of the sum of heterogenic segments *in any finite interval* can then be obtained.

The first modification necessitates the introduction of two higher-order identity relations between genes in order to obtain the mean number of external junctions per unit map length. The second modification enables application of the theory to early generations, because the average length of heterogenic tracts, relative to the map length considered, is irrelevant. The crucial assumption of my approach is that both heterogenic and homogenic tract lengths are distributed exponentially. The validity of this assumption has been verified by means of a series of Monte Carlo simulations.

A different approach to the distribution of heterogeneity has been indicated by Franklin (1977). This approach uses the concept of joint identity by descent for pairs of loci, developed by Cockerham and Weir (Cockerham & Weir, 1968; Weir & Cockerham, 1974). As pointed out by Franklin (1977), summation of this probability over all possible pairs of loci on a chromosome gives the second moment of the total number of homogenic loci on that chromosome. Some of Franklin's results will be discussed further on.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

2. ANALYSIS

(i) *Junctions*

An exchange between two unlike strands leads to a *junction*. The formation of a junction is a unique event, since no two crossovers will occur at exactly the same site. A junction can therefore be regarded as a unique point mutation, and the fate of a junction is the same as that of a unique mutation: ultimately it will either be lost or become fixed. Considering the state at the site of a junction in a diploid we can distinguish two types, i.e. *internal* and *external* ones. Let j denote a junction between unlike strands (a and b). We will denote a strand which is, at the site of the junction, identical to either a or b by x; a strand unlike both a and b is denoted by y. Then, at the site of a junction the possible genotypes are xx, xj, xy, jj, jy and yy (see Fig. 1). In the genotypes jx (i.e. ja and jb) the genome switches from heterogenic to homogenic at the site of the junction. Here, the junction is external; it marks the end of a heterogenic tract. In the other genotypes there is no switch from heterogenic to homogenic. In jj the genome is homogenic at either side of the junction; in jy it is heterogenic at either side. Therefore the junction is said to be internal in the genotypes jj and jy. Notice that an internal junction may become external in a later generation, and vice versa. The concept of the state of a junction in a diploid (i.e. internal or external) also applies to randomly sampled gametes from distinct individuals. Fig. 1 further illustrates the concept of internal and external junctions.

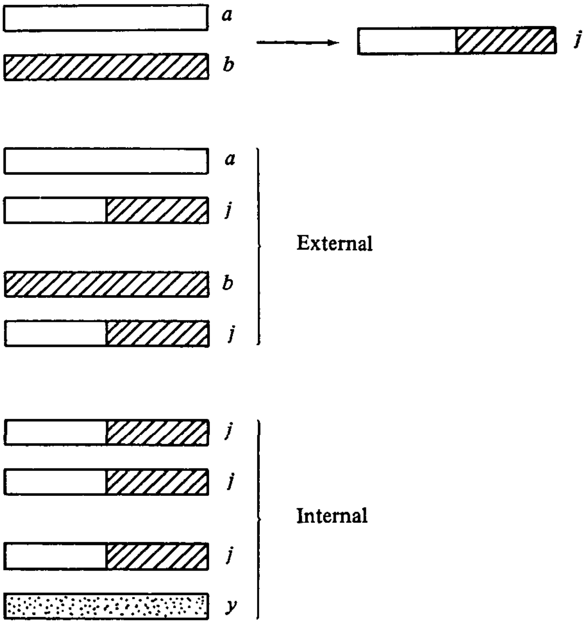


Fig. 1. The concept of external and internal junctions: j is a junction resulting from a crossover between two unlike strands (here denoted by a and b). The genotypic notation, aj, bj, etc., refers to the status at the site of the junction: y denotes a strand unlike both a and b. For further explanation see text.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

(ii) *Identity relations*

We will develop the theory for a finite population of constant size N that reproduces by random union of gametes, while selfing is excluded (i.e. monoecy). Generations are assumed to be non-overlapping. Besides the usual identity relations between two genes, we introduce two higher-order identity relations between three genes (S and R, below). In the following we will use the term 'unlike' in the sense of non-identity by descent; so when three genes, α, β and γ, are unlike, this means $\alpha \not\equiv \beta$, $\alpha \not\equiv \gamma$, $\beta \not\equiv \gamma$. We will use the following probabilities (indices refer to generations):

H_t: the probability that the homologous genes of an individual at a randomly chosen locus are unlike ($1 - H_t = F_t$, the inbreeding coefficient).

K_t: the probability that two homologous genes sampled from distinct individuals are unlike.

S_t: the probability that three genes, α, β and γ are unlike, when two of the genes constitute a homologous pair of a single individual and the third is sampled from a distinct individual.

R_t: the probability that three homologous genes, α, β and γ are unlike when these genes are drawn from three distinct individuals. These relations are shown in Fig. 2.

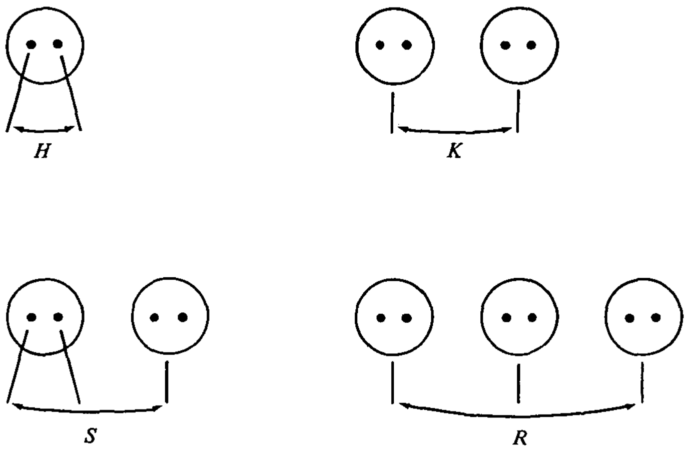


Fig. 2. Identity relations between two (H and K) and three (S and R) genes.

Higher-order identity relations have been discussed by, among others, Harris (1964) and Cockerham (1971). In Cockerham's (1971) notation, R and S read as follows

$$R = 1 - \theta_{xy} - \theta_{xz} - \theta_{yz} + 2\gamma_{xyz},$$

$$S = 1 - 2\theta_{xy} - F_{\ddot{x}} + 2\gamma_{\ddot{x}y}.$$

For H and K we have the well-known recurrence relations

$$H_{t+1} = K_t,$$

$$K_{t+1} = \frac{1}{2N} H_t + \left(1 - \frac{1}{N}\right) K_t. \qquad (5)$$

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

General rules for the computation of higher-order identity relations are given by Cockerham (1971). We will here derive the recurrence relations for S and R.

Considering S_{t+1} we see that with probability $2/N$ the three genes derive from two distinct parents; in that case the probability of being unlike equals $\frac{1}{2}S_t$. With probability $(1-2/N)$ the genes derive from three distinct parents, in which case the probability of being unlike equals R_t. Thus,

$$S_{t+1} = \frac{1}{N}S_t + \left(1-\frac{2}{N}\right)R_t. \tag{6}$$

Next consider R_{t+1}. With probability $1/N^2$ the three genes derive from a single parent; the probability of being unlike then is zero. With probability $3(N-1)/N^2$ the genes derive from two distinct parents; in that case the probability of being unlike is $\frac{1}{2}S_t$. With probability $(N-1)(N-2)/N^2$ the genes derive from three distinct parents, in which case the probability of being unlike is R_t. Thus,

$$R_{t+1} = \frac{3(N-1)}{2N^2}S_t + \frac{(N-1)(N-2)}{N^2}R_t. \tag{7}$$

Equations (5)–(7) are the basic recurrence relations needed for the calculation of the mean number of external junctions per unit map length in any generation.

(iii) *The mean number of external junctions per unit map length*

New junctions are formed at the rate H_t; i.e. the expected number of new junctions per 100 centimorgans formed in generation t equals H_t. We now define the quantity P_t as the mean number of external junctions (per 100 centimorgans) considering one of the two homologous chromosomes, that is the mean number of junctions in external state on one of the two homologous chromosomes of an individual. Then the mean number of external junctions per individual (Z_t) equals $2P_t$. Analogously we define Q_t as the mean number of external junctions on one chromosome of a homologous pair, drawn at random from two distinct individuals of generation t. In the following, j indicates a junction between strands a and b, and c is the homologous site of this junction in either the same individual or in a distinct individual.

We will now set up recurrence relations for P and Q using the identity relations defined in the previous section. It will be clear that the contributions to both P_{t+1} and Q_{t+1} come from distinct sources: we can distinguish a part that comes from junctions that already existed in generation t and a part coming from junctions that were first formed in generation t.

First consider P_{t+1} (cf. Fig. 3a). The contribution to P_{t+1} from existing junctions simply equals Q_t. The expected number of new junctions equals H_t. From this we must subtract those that lead to the internal condition in the next generation. A junction formed in generation t will be internal in the next generation if in the

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

gamete from the other parent the corresponding site is unlike both a and b. The latter occurs with probability S_t. Thus,

$$P_{t+1} = Q_t + H_t - S_t. \tag{8}$$

Next consider Q_{t+1} (cf. Fig. 3*b*). The contribution from existing junctions to Q_{t+1} equals

$$\frac{P_t}{2N} + \left(1 - \frac{1}{N}\right) Q_t,$$

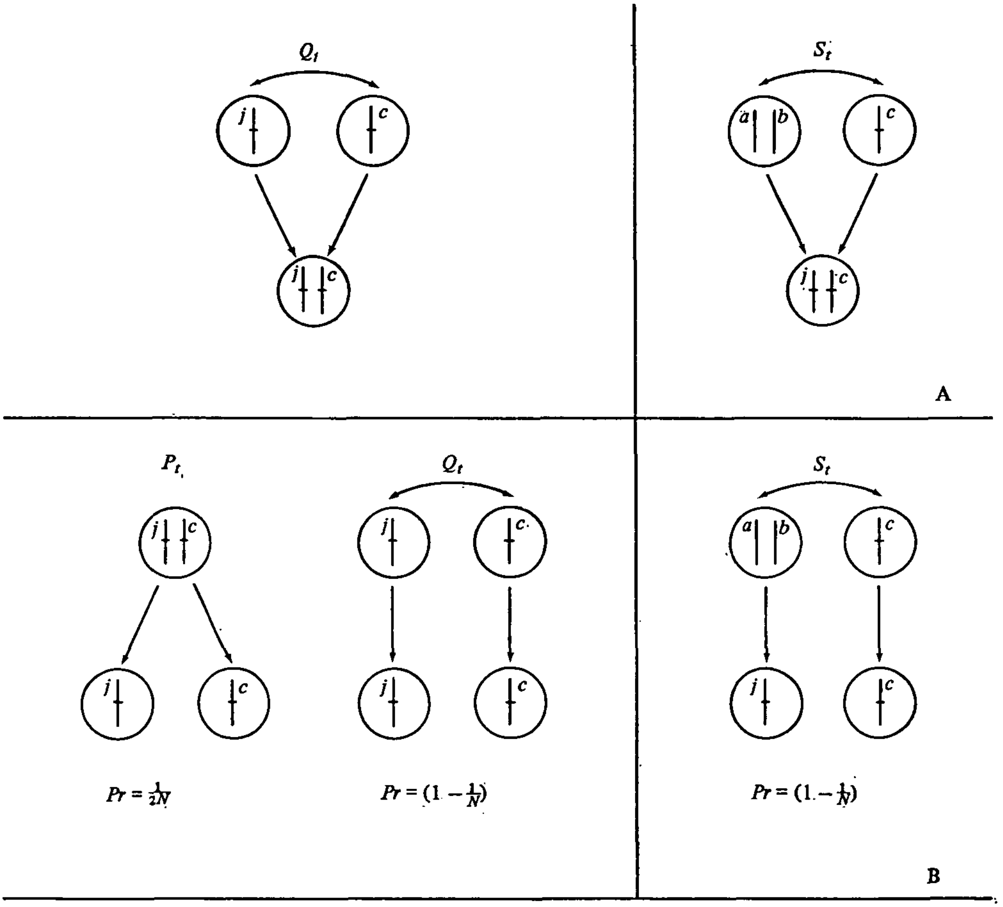


Fig. 3. Contributions to P_{t+1} (A) and Q_{t+1} (B). Left: from junctions formed before generation t; right: from junctions formed after generation t. j denotes a junction between unlike strands a and b; c denotes a gene at the site of the junction in the homologous chromosome (A) or in a chromosome from a distinct individual (B).

which is analogous to the equation for K_{t+1} (5). In order to obtain the contribution to Q_{t+1} from new junctions (formed in generation t) we subtract from H_t the part that leads to the internal condition. The internal condition occurs only if the sites j and c derive from distinct parents and if c is unlike both a and b (Fig. 3*b*). This has probability $(1 - 1/N) S_t$. Taking these contributions together we have

$$Q_{t+1} = \frac{1}{2N} P_t + \left(1 - \frac{1}{N}\right) Q_t + H_t - \left(1 - \frac{1}{N}\right) S_t. \tag{9}$$

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

Combining equations (5)–(9) in matrix notation, with

$$\mathbf{X}' = (P, Q, H, K, S, R),$$

we have

$$\mathbf{X}_{t+1} = \mathbf{C}\mathbf{X}_t,$$

where the transition matrix $\mathbf{C}$ is

$$\begin{pmatrix} 0 & 1 & 1 & 0 & -1 & 0 \\ \frac{1}{2N} & \left(1-\frac{1}{N}\right) & 1 & 0 & -\left(1-\frac{1}{N}\right) & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2N} & \left(1-\frac{1}{2N}\right) & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{N} & 1-\frac{2}{N} \\ 0 & 0 & 0 & 0 & \frac{3(N-1)}{2N^2} & \frac{(N-1)(N-2)}{N^2} \end{pmatrix}.$$

It is easy to see that the eigenvalues of the transition matrix are the roots of

$$[\lambda^2-(1-1/N)\lambda-1/2N]^2[\lambda^2-\{1-2(N-1)/N^2\}\lambda-(N-1)(N-2)/2N^3] = 0.$$

This has two double roots, i.e.

$$\lambda_1 = \lambda_2 = (N-1+\surd(N^2+1))/2N,$$

and

$$\lambda_3 = \lambda_4 = (N-1-\surd(N^2+1))/2N,$$

and further

$$\lambda_{5,6} = (\alpha \pm D)/2N,$$

where

$$\alpha = N+(N-1)(N-2),$$

and

$$D = \surd(\alpha^2+2N(N-1)(N-2)).$$

The explicit expression for $Z_t = 2P_t$ is then of the form

$$Z_t = (c_1+tc_2)\lambda_1^t+(c_3+tc_4)\lambda_3^t+c_5\lambda_5^t+c_6\lambda_6^t, \tag{11}$$

the constants $c_1, \ldots, c_6$ depending on the initial conditions. Assuming that initially all genes in the population are unlike, the initial conditions are

$$P_0 = Q_0 = 0; \quad H_0 = K_0 = S_0 = R_0 = 1.$$

The constants c_1–c_6 are given in the Appendix.

For $N = 2$, i.e. full sib mating, the sixth eigenvalue is zero because the relation R between three genes does not exist. It is straightforward to verify that the explicit expression for Z_t is then

$$Z_t = \left\{-\frac{4}{25}(3+38\epsilon)+\frac{4}{5}(1+4\epsilon)t\right\}\epsilon^t$$
$$-\left\{\frac{8}{25}(11-19\epsilon)-\frac{4}{5}(3-4\epsilon)t\right\}(\tfrac{1}{2}-\epsilon)^t+4(\tfrac{1}{2})^t, \tag{13}$$

where $\epsilon = \lambda_1 = (1+\surd 5)/4$.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

Table 1. *The mean number of external junctions* (Z_t), *per 100 centimorgans in full sib mating. t, generation; A, theoretical value (equation 13); B, simulation results*

	Z_t			Z_t	
t	A	B	t	A	B
1	0·0000	0·00	11	3·1000	3·06
2	2·0000	2·01	12	2·7734	2·75
3	3·0000	3·00	13	2·4590	2·38
4	3·7500	3·77	14	2·1636	2·11
5	4·1250	4·11	15	1·8914	1·90
6	4·2500	4·26	16	1·6442	1·64
7	4·1875	4·12	17	1·4225	1·42
8	4·0000	4·03	18	1·2255	1·24
9	3·7344	3·65	19	1·0519	1·08
10	3·4258	3·34	20	0·8999	0·90

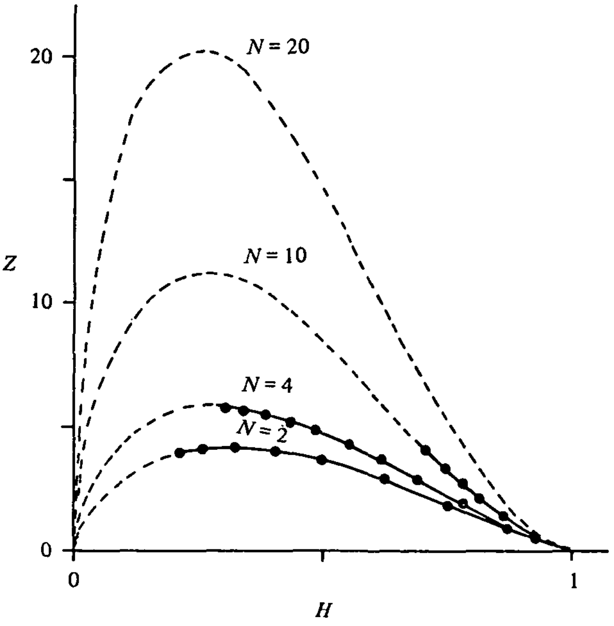


Fig. 4. Trajectories of H (mean heterogeneity) and Z (mean number of external junctions) for different values of N (population size). Dots indicate successive generations (starting at $H = 1$, $t = 1$). In the dashed part of the graphs the generations are not indicated.

Table 1 gives the values of Z for the first 20 generations of full sib mating (equation (13)), together with the mean values obtained in 500 replicate simulation runs (see the section on simulation). In Fig. 4 are plotted corresponding values of Z and H for several population sizes. It is seen that at a given level of H, Z increases as population size increases. This is because it takes longer for a large population to reach a given value of H than for a small one. Let us compare two populations of different size (small and large). The initial rate of increase of Z per generation is

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

smaller for the large population because a larger fraction of its junctions will then be internal ones. However, the quick drop of H in the small population rapidly reduces the opportunity to create new junctions. The net effect is a larger value of Z for the large population at a given level of H.

(iv) *The limiting number of junctions*

When complete identity by descent has been attained, the genome of every individual consists of a number of segments of different origin separated by internal junctions that have become fixed. The mean number of these junctions is found as follows. In any generation the expected number of junctions formed in the total population equals $2NH_t$. Since each junction has probability $1/2N$ of surviving, the expected number still present in the limit equals $\Sigma_{t=0}^{\infty} H_t$. It is easily verified that H_t can be written as

$$H_t = \{(1-\lambda_3)/(\lambda_1-\lambda_3)\}\lambda_1^t - \{(1-\lambda_1)/(\lambda_1-\lambda_3)\}\lambda_3^t, \tag{14}$$

so that

$$\sum_{t=0}^{\infty} H_t = \{(1-\lambda_3)/(1-\lambda_1)-(1-\lambda_1)/(1-\lambda_3)\}/(\lambda_1-\lambda_3) = 2(N+1). \tag{15}$$

For an organism with total map length L, distributed over n chromosome pairs, the expected number of distinct segments of a given origin thus equals

$$2(N+1)L+n.$$

This result was obtained by Bennet (1953) for full sib mating ($N = 2$).

(v) *The distribution of heterogeneity*

In order to obtain the probability distribution of the heterogenic fraction of the genome, we will assume that in any generation the lengths of both homogenic and heterogenic segments are distributed exponentially, each with its own parameter. External junctions can then be regarded as events in two alternating Poisson processes. Let us denote these alternating states by A (heterogenic) and B (homogenic). Let

$$f(x) = \lambda e^{-\lambda x} \tag{16}$$

be the p.d.f. of the length of a heterogenic tract, and let

$$g(x) = \alpha e^{-\alpha x} \tag{17}$$

be the corresponding p.d.f. for homogenic tracts. The mean lengths of heterogenic and homogenic tracts then are $1/\lambda$ and $1/\alpha$, respectively. Since a junction cannot coincide with a chromosome end, the latter (or any other point on the chromosome) can be considered as a random point in the process. In other words, a chromosome (or any other segment of finite length) can be considered as a section of the process in stationary phase. From elementary renewal theory (see e.g. Cox, 1962) it can be shown that the probability, p_A, that at a random point the process is in state A, is

$$p_A = \mu_A/(\mu_A+\mu_B) = \alpha/(\alpha+\lambda).$$

Similarly,

$$p_B = \lambda/(\alpha+\lambda).$$

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

In terms of H_t this is $H_t = \alpha/(\alpha+\lambda), \quad 1-H_t = \lambda/(\lambda+\alpha).$ (18)

It can further be shown (cf. Cox, 1962) that the mean number of events per unit length in the stationary process equals $2\alpha\lambda/(\alpha+\lambda)$. Thus,

$$Z_t = 2\alpha\lambda/(\alpha+\lambda). \tag{19}$$

Combining (18) and (19) allows us to calculate the parameters α and λ for any generation, i.e.

$$\alpha = Z_t/2(1-H_t), \quad \lambda = Z_t/2H_t. \tag{20}$$

The p.d.f. of the heterogenic fraction in a segment of length L, in terms of α, λ and L, reads (see Appendix for a derivation):

$$\begin{aligned}\phi(x) &= \frac{\alpha\lambda}{\alpha+\lambda} L \exp[-L\{\lambda x+\alpha(1-x)\}] \\ &\quad \times \left[\frac{\alpha x+\lambda(1-x)}{\alpha\lambda x(1-x)} I_1\{2L\surd(\alpha\lambda x(1-x))\}+2I_0\{2L\surd(\alpha\lambda x(1-x))\}\right],\end{aligned} \tag{21}$$

where I_0 and I_1 are modified Bessel functions of the first kind of order zero and one, respectively. The probability condensations P_0 at $x = 0$ and P_1 at $x = 1$ are

$$P_0 = \frac{\lambda}{\alpha+\lambda} e^{-\alpha L}, \quad P_1 = \frac{\alpha}{\alpha+\lambda} e^{-\lambda L}. \tag{22}$$

The result (21)–(22) is not particularly useful because the second and higher moments cannot be obtained directly from it.

Numerical evaluation of (21) however is quite feasible. The variance can best be obtained by a direct argument (see Appendix); it reads

$$\operatorname{var}(x) = \frac{\{2H(1-H)\}^2}{ZL^2}\left[L - \frac{2H(1-H)}{Z}\left\{1-\exp\left(\frac{ZL}{2H(1-H)}\right)\right\}\right]. \tag{23}$$

The distribution (21)–(22) applies to a single chromosome or chromosome segment of length L. In order to obtain the corresponding distribution for n independently segregating chromosomes, each of length L, the n-fold convolution should be taken, together with a transformation of scale. However, in most cases we can, without serious error, treat n chromosomes of length L as if they were a single one of length n. The reason for this is that the approach to stationarity of the alternating process, given the state at $t = 0$, is rather fast. In fact,

$$\left.\begin{aligned} p_A(t) &= \frac{\alpha}{\alpha+\lambda}[1-\exp\{-(\alpha+\lambda)t\}]+p_A(0)\exp\{-(\alpha+\lambda)t\}, \\ p_B(t) &= \frac{\lambda}{\alpha+\lambda}[1-\exp\{-(\alpha+\lambda)t\}]+p_B(0)\exp\{-(\alpha+\lambda)t\}, \end{aligned}\right\} \tag{24}$$

where $p_A(t)$ are the probabilities of being in state A and B, respectively, at a distance t from the origin (cf. Cox & Smith, 1961). It is seen from (24) that for $(\alpha+\lambda)t \gg 1$, the state at $t = 0$ hardly influences the values of $p_A(t)$ and $p_B(t)$. This means that a long segment may, for our purposes, be treated as a number of independent smaller

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

ones, provided the latter are not too short. Since the reverse is also true, we may treat independent segments as if they constituted a single one. In Fig. 5 values of $\alpha+\lambda\,(=Z/2H(1-H))$ are plotted against the corresponding value of H for some population sizes. It is seen that in general $(\alpha+\lambda)$ is much larger than unity. The lower boundary of $(\alpha+\lambda)$ can be shown to be 4. This justifies treating independent chromosomes as if they constituted a single one.

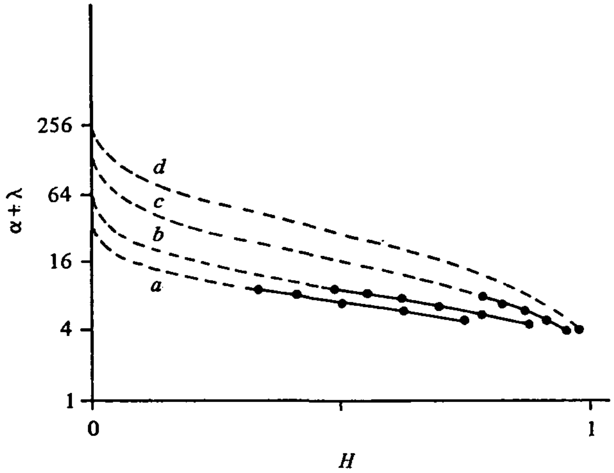


Fig. 5. Trajectories of H and $\alpha+\lambda$ for different population sizes (N). *a*: $N = 2$; *b*: $N = 4$; *c*: $N = 10$; *d*: $N = 20$. Trajectories start at the right at generation 2. Further legend as Fig. 4.

As an example we compare the variance for a single chromosome of length nL, var $(x; nL)$, with that for n independent chromosomes of length L, i.e. var $(x; L)/n$. These are given in Table 2 for $N = 4$ and $N = 16$ in generations 4 and 16 for $nL = 20$. From Table 2 it is seen that, unless $L \leqslant 0\cdot5$, there is no substantial discrepancy between var $(x; nL)$ and var $(x; L)/n$.

Several numerical examples of the distribution (21)–(22) are given in the next section.

Table 2. *Variance of the heterogenic fraction for a total genome length of 20 morgans with different numbers* (n) *of chromosomes of equal length* (L). *N, population size; t, generation*

		$t = 4$		$t = 16$	
n	L	$N = 4$	$N = 16$	$N = 4$	$N = 16$
1	20	0·0031	0·0013	0·00076	0·0013
2	10	0·0031	0·0013	0·00076	0·0013
4	5	0·0029	0·0013	0·00075	0·0013
8	2·5	0·0029	0·0012	0·00074	0·0013
10	2	0·0029	0·0012	0·00074	0·0013
20	1	0·0027	0·0011	0·00072	0·0012
40	0·5	0·0022	0·0009	0·00068	0·0012
80	0·25	0·0016	0·0006	0·00060	0·0010

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

3. SIMULATION

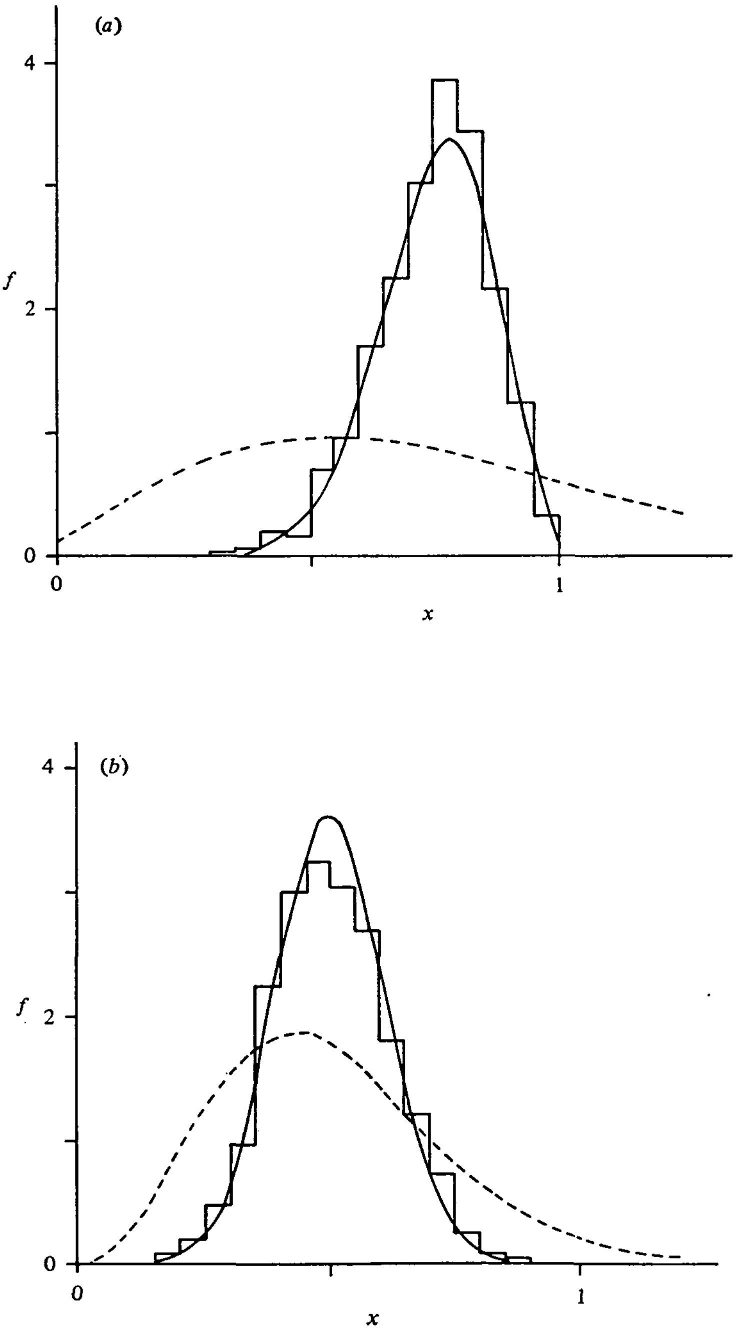
The crucial assumption of the model concerns the probability distribution of homogenic and heterogenic tract lengths. Both distributions are assumed to be exponential. In order to test the validity of this assumption a series of Monte Carlo simulations was run. In the simulation each chromosome was described by the following characteristics: (1) the number of junctions (n), (2) the sites of the junctions: $s_1, \ldots, s_n$ and (3) the 'origin' of the chromosome between any two junctions, being one of the numbers $1, 2, \ldots, 2N$ (corresponding to the $2N$ possible origins). From these characteristics the number of external junctions and the heterogenic fraction can be obtained for a pair of homologous chromosomes. Random mating with the exclusion of selfing was simulated by repeated sampling of two distinct individuals as the parents of a new zygote (this implies monoecy). Crossover sites were generated by successive sampling from an exponential distribution (i.e. no interference was assumed) until a site fell outside the pre-assigned chromosome length. After recovering the gamete (a rather complicated procedure because for each gamete the sites of the junctions and the origins must be determined from the parental homologues and the crossover sites), this was stored until $2N$ gametes had been generated. The parental population was then replaced by the offspring. At pre-assigned generation intervals output was gathered and stored; this was used after all replicates had been run. From the results of replicate runs a frequency diagram, the overall mean and overall variance of the heterogenic fraction were calculated. The histograms with 20 classes of width 0·05, plus the classes 0 and 1 (complete homogeneity and complete heterogeneity) can be compared with the theoretical distribution (21–22). The mean and variance of the number of external junctions per 100 centimorgans were also recorded. The number of independently segregating chromosomes could be varied, as well as the chromosome lengths.

With this method the chromosome is 'continuous' rather than 'discontinuous', as it is in the more familiar 'one locus–one bit' simulation technique. With the latter a segment between junctions cannot be smaller than the total map length divided by the number of bits used. This may result in an underestimation of the number of junctions. A similar 'continuous chromosome' procedure was used by Robertson (1977) in simulating artificial selection.

4. SOME NUMERICAL RESULTS

Graphs of the p.d.f. (21)–(22) are shown in Figs. 6–9. In Fig. 6 a sample of simulation results is also shown. In Fig. 6 also the p.d.f. according to Fisher's approach has been plotted (i.e. equation (3) after scale transformation). It is seen from Figs. 6 and 7 that the fit between the theoretical distribution and the simulation results is surprisingly good. Fisher's approximation, which was stated to hold for later generations, shows a poor fit indeed in early generations, the fit becoming increasingly better as H decreases.

Fig. 8 exemplifies the effect of the total map length, L, under consideration: at



See legend for Fig. 6 on opposite page.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

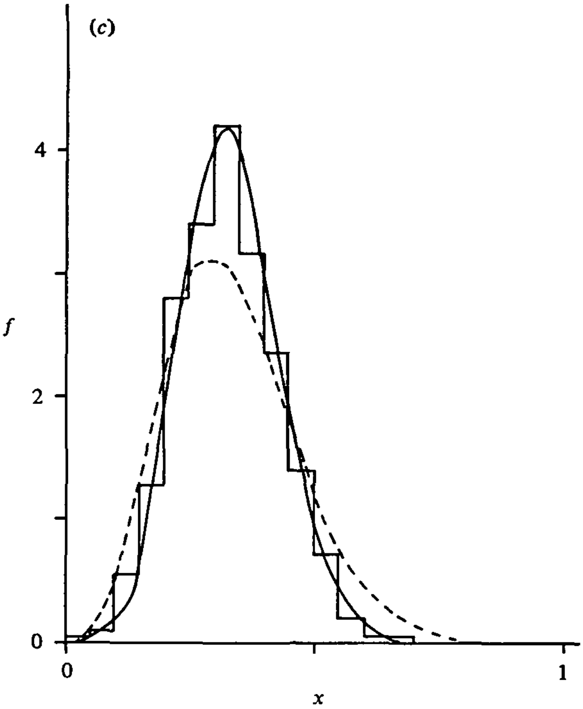


Fig. 6. Probability density function of the heterogenic proportion of the genome. x = heterogenic proportion; f = probability density. Drawn curve: equation (21); dashed curve: Fisher and Bennet's approximation (equation (3)); histogram: simulation results. Arrows indicate the mean of the distribution. N: population size; t: generation; L: genome length (morgans; single chromosome). (*a*) $N = 2$, $t = 2$, $L = 2$; (*b*) $N = 2$, $t = 4$, $L = 2$; (*c*) $N = 2$, $t = 6$, $L = 2$.

a given level of H, the variance of the heterogenic proportion is inversely proportional to L.

Fig. 9 shows the effect of population size on the distribution with given mean. An increase in Z will reduce the variance, because the heterogenic fraction becomes distributed over a larger number of distinct segments. Increasing population size, which corresponds to an increase in Z (cf. Fig. 4), therefore reduces the variance.

Expression (22) makes possible calculation of the probability of 100% homogeneity in any generation. This has been plotted for a few values of N in Fig. 10. It is seen that the variance of the time to 100% homogeneity increases with population size.

5. DISCUSSION

The probability distribution of the heterogenic fraction of the genome in a finite population has been derived under the assumption of selective neutrality. The approach of this paper uses the concept of junctions, developed by Fisher (1949, 1954) and Bennet (1953, 1954). In most of the previous work on inbreeding only expected values were considered. Knowledge of the probability distribution of the

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

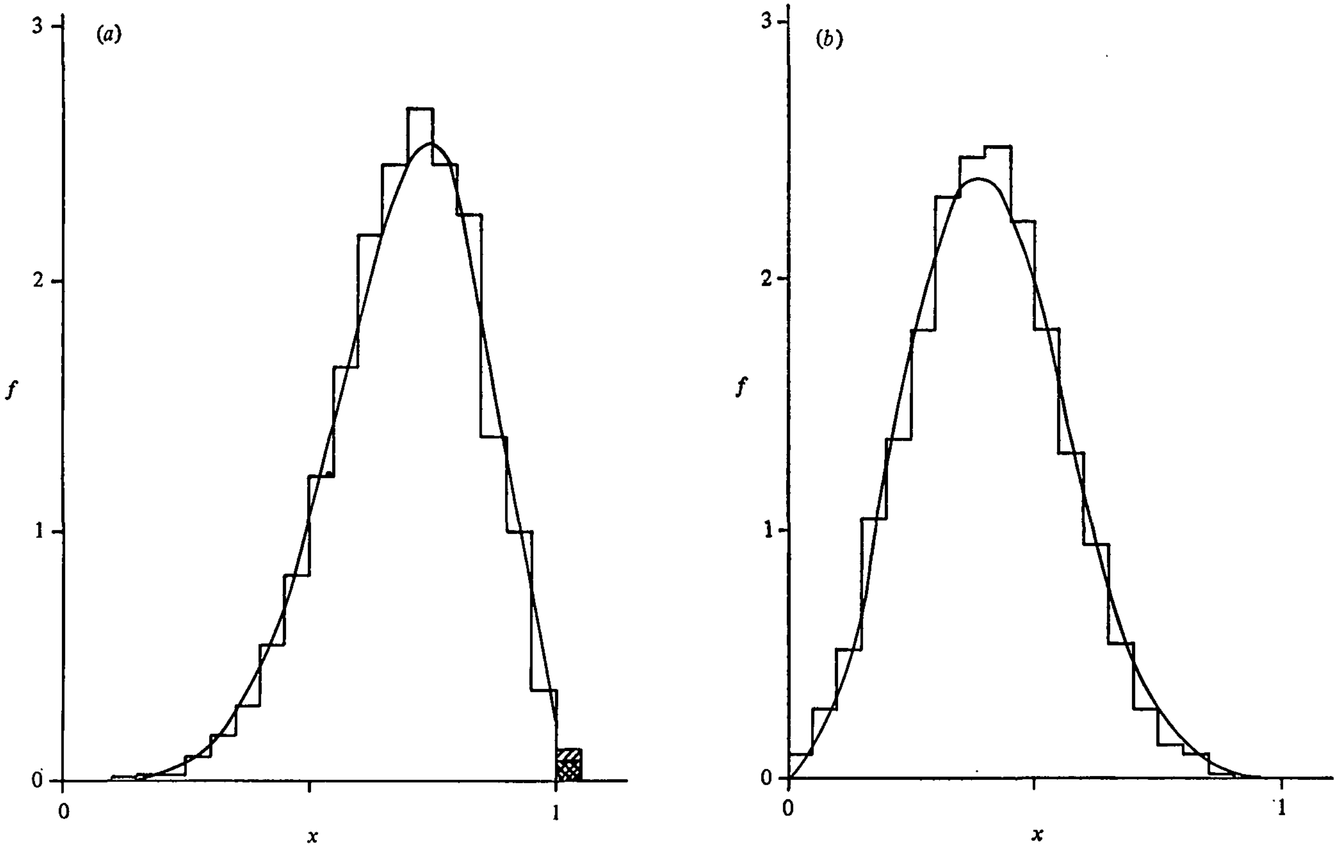


Fig. 7. Same as Fig. 6, without Fisher and Bennet's approximation. (*a*) $N = 4$, $t = 4$, $L = 2{\cdot}5$; (*b*) $N = 8$, $t = 16$, $L = 1$. Hatched and cross-hatched parts indicate, at arbitrary scale, the probability condensations at $x = 1$ for the simulated process and the theoretical distribution, respectively.

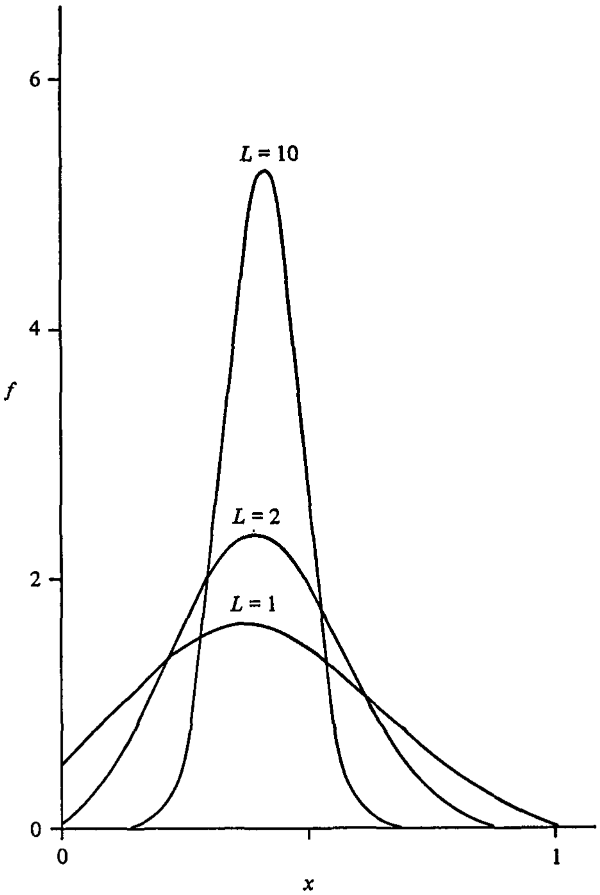


Fig. 8. Same as Fig. 6, exemplifying the effect of L; $N = 2$; $t = 5$.

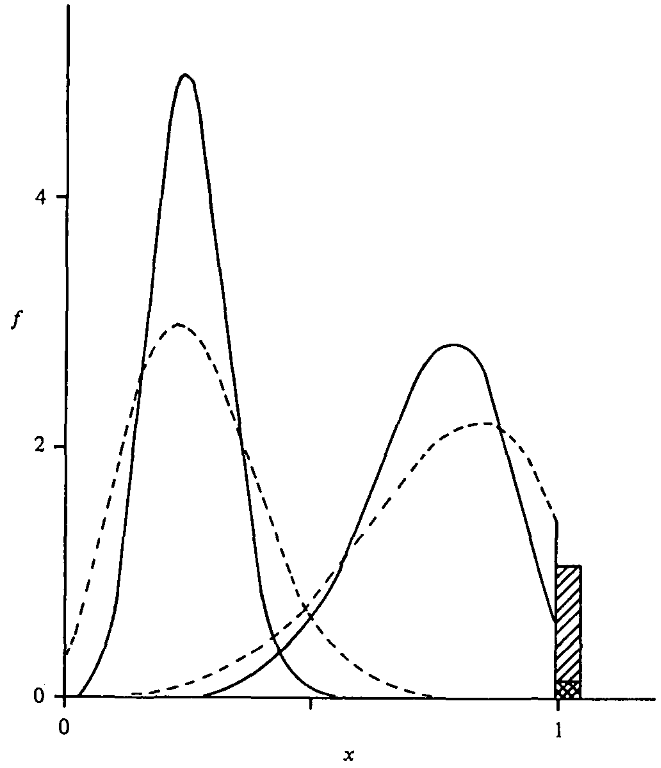


Fig. 9. Same as Fig. 6, exemplifying the effect of population size on the distribution with a given mean. $L = 2$; $H_1 = 0{\cdot}75$; $H_2 = 0{\cdot}25$. Drawn curves: $N = 10$; dashed curves: $N = 2$. Hatching indicates the probability condensations at $x = 1$.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

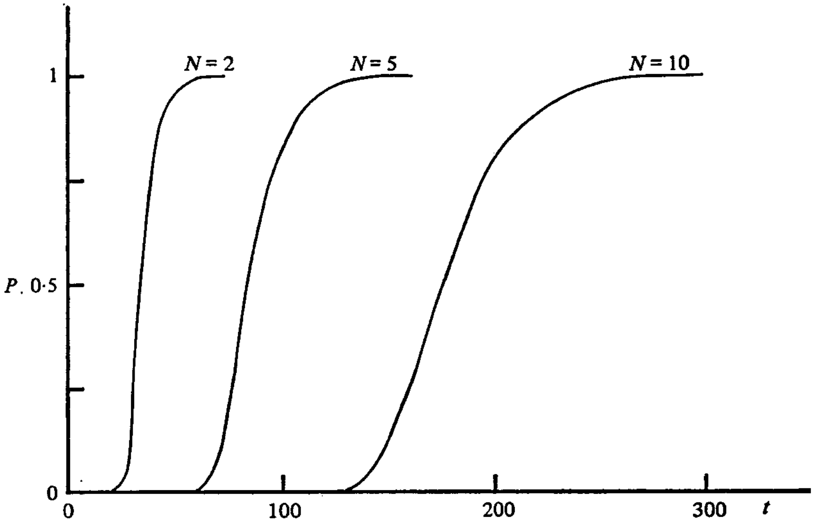


Fig. 10. The probability of complete homogeneity (P) as a function of generation (t), for populations of different size (N). Total genome length is 20 morgans in all cases.

heterogenic proportion of the genome adds a new dimension to the familiar calculation of inbreeding coefficients. It allows us to calculate the probability that a given proportion or more of the genome is homogenic at a certain stage of inbreeding. For practical plant and animal breeding such probabilities are of far more concern than the mean, especially since the variance may be substantial.

Franklin (1977) has presented a method of deriving the variance of the heterogenic fraction using the concept of joint identity by descent at two loci. The basic idea of this method is as follows (see also the Appendix). Let $\theta(r_{ij})$ be the probability that two loci with recombination r_{ij} are both homogenic (recurrence relations for $\theta(r)$ have been given by Cockerham & Weir (1968) for sib mating and by Weir & Cockerham (1974) for monoecious populations of finite size). Then, with n loci, the sum

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\theta(r_{ij})$$

is the second moment of the homogenic fraction. Changing to an infinite number of loci, uniformly distributed over a chromosome of length L, this can be written as

$$\frac{1}{L^2}\int_0^L\int_0^L \theta\{r(x_1,x_2)\}\,dx_1\,dx_2, \tag{25}$$

where $r(x_1, x_2)$ now denotes the recombination fraction between loci at sites x_1 and x_2. A genetic mapping function relating the distance $|x_1-x_2|$ to $r(x_1, x_2)$ is used to evaluate the integral (25). (When interference is present, $r(x_1, x_2)$ may not only depend on the distance $|x_1-x_2|$ but also on x_1 and x_2 themselves; cf. Stam, 1979.) In general, the probability $\theta(r)$ can be written as a polynomial in $(1-2r)$, i.e. as $\Sigma_k a_k(1-2r)^k$. Franklin (1977) has given numerical values of the a_k for the first six generations of full sib mating and parent–offspring mating. For numerical work,

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

$\theta(r)$ can be tabulated using the transition matrix given by Weir & Cockerham (1974); (25) can then be integrated numerically.

The variance of the distribution proposed by Fisher and Bennet, i.e. equation (3), is $2m/a^2L^2$. Table 3 gives a sample of variances calculated in three ways, i.e. (1) Franklin's (1977) method, using his expressions for $\theta(r)$; (2) Fisher and Bennet's approximation, using the values of Z_t and H_t given in this paper, and (3) according to equation (23).

Table 3. *Variance of the heterogenic proportion with full sib mating. L, total genome length; n, number of chromosomes; t, generation. A, Fisher and Bennet's approximation; B, Franklin's method (1977); C, equation (23) of the text, assuming independence of unlinked loci*

t	L	n	A	B	C
3	5	1	0·0962	0·0154	0·0142
		5	0·0735	0·0131	0·0124
	10	1	0·0500	0·0079	0·0072
		10	0·0368	0·0065	0·0062
	20	1	0·0255	0·0040	0·0036
		20	0·0184	0·0033	0·0031
6	5	1	0·0196	0·0109	0·0090
		5	0·0176	0·0095	0·0082
	10	1	0·0100	0·0055	0·0045
		10	0·0089	0·0048	0·0041
	20	1	0·0050	0·0028	0·0023
		20	0·0044	0·0024	0·0020

It is seen from Table 2 that the values according to equation (23) are very close to those obtained by Franklin's method, whereas Fisher and Bennet's approximation largely overestimates the variance in early generations.

In terms of H and Z, the variance of Fisher and Bennet's distribution reads

$$\operatorname{var}(x) = \frac{2H}{L}\left(\frac{Z}{2H}+\frac{n}{L}\right)^{-1}.$$

In later generations, when H becomes very small, and Fisher and Bennet's approximation becomes increasingly better, this can be approximated by

$$\operatorname{var}(x) = 4H^2/LZ,$$

which then hardly differs from (23). Taking the limit of var (x) as $L \to 0$ in (23) we obtain

$$\operatorname{var}(x) = H(1-H),$$

which is the correct expression for the one-locus case.

Avery & Hill (1979) followed an approach similar to Franklin's method (i.e. by deriving the relevant expressions for two loci, and averaging over all pairs of loci) to obtain the various components of variance in heterozygosity in a finite random mating population (including random selfing). The main components of variance are the between-population and within-population variances, although Avery & Hill (1979) considered several other components as well, such as within- and between-half sib families. Unfortunately, the method of the present paper does not allow such a subdivision of the variance in a simple way.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

Sved (1971, see also Sved & Feldman, 1973) has calculated the limiting probability that the genes at two loci, with recombination r, are identical by descent *through the same pathway*. From this probability,

$$Q = \frac{1}{1+4Nr},$$

Sved (1971) derived the p.d.f.

$$\Phi(x) = 4N/(1+4Nx)^2, \tag{26}$$

for the length of an unbroken homogenic segment (through the same pathway) attached to a given homogenic locus. This p.d.f. is length-biased because a randomly sampled locus is more likely to fall in a large segment than in a small one. Denoting the length-unbiased or 'true' p.d.f. of such segments by $f(x)$, the relation between $f(x)$ and $\Phi(x)$ is

$$\Phi(x) = \frac{1-F(x)}{\mu},$$

where

$$F(x) = \int_0^x f(u)\,du$$

and

$$\mu = \int_0^\infty uf(u)\,du.$$

(See e.g. Cox, 1962 for an instructive treatment of this 'waiting time paradox'.) It is easily verified that with (26) the true p.d.f. reads

$$f(x) = \frac{8N}{(1+4Nx)^3},$$

with mean

$$\mu = \frac{1}{4N}. \tag{27}$$

From (15) we see that in the limit the mean length of homogenic segments from a given origin, without the restriction of having descended through the same pathway, equals $1/2(N+1)$, which is approximately twice the mean length of unbroken segments descended through the same pathway (27). Sved (1971) has calculated the expected length of homogenic segments (same pathway) surrounding a given locus, using the p.d.f. (26). This value is then interpreted by Sved as the mean length of homozygous segments (as opposed to heterozygous segments) in an equilibrium population. It is hard to understand the logic behind this interpretation. First, only segments are considered that have remained unbroken (same pathway) since the initial generation; second, the mean length is over-estimated because of the length-bias in the p.d.f. (26), and third, the equilibrium population in the strict sense is completely homozygous, so that it makes little sense to consider homozygous segments as opposed to heterozygous ones.

The model of this paper implies that the identity states at unlinked loci are independent; i.e. the joint probability of non-identity equals H^2 for unlinked loci (this follows directly from (24)). This is correct for full sib mating ($N = 2$), but not

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

for populations of size > 2, in which case the identity states at unlinked loci are (positively) correlated (cf. Weir & Cockerham, 1974). As a consequence, the model slightly underestimates the variance of the heterogenic proportion for long chromosomes. If fitness is related to the homogenic fraction of the genome (as indicated by the phenomenon of inbreeding depression), selection will retard the approach to homogeneity. One might then formulate 'infinite-locus' or 'strand' models of selection, as suggested by Franklin & Lewontin (1970), and applied by Robertson (1977) in the context of artificial selection. With such a model, the variance of the homogenic fraction directly corresponds to variance of fitness. Since variation in fitness is essentially equivalent to selection, the effect of selection on the rate of approach to homogeneity in a finite population is expected to depend on the two factors that determine the variance, i.e. the total genome length L and Z which, for given level of H, in turn depends on population size. Of course, selection will also affect the variance itself; therefore, predictions concerning the effect of selection on the distribution of heterogeneity can only be speculative and at most qualitatively correct. A simulation study of the effect of selection is now under progress.

Franklin & Lewontin (1970) and Lewontin (1974) have raised the question as to whether the analysis of multi-locus selection models can really contribute to our understanding of selection in natural populations. The results obtained by Franklin & Lewontin (1970) indicate that relatively large blocks of linked genes rather than single genes may govern the process of selection (gene frequency changes) at individual loci. They found that the alleles at individual loci are being 'locked up' in semi-permanent super-alleles and that the behaviour of single genes and pairs of genes can hardly be predicted from single- or two-locus theory. Similar results were obtained by Wills, Crenshaw & Vitale (1969) for a rank order selection model. In addition, Avery (1978) has pointed out that the results of deterministic two-locus theory may be of little value as to what may actually happen in a finite population. He showed that genetic drift causes D (measure of linkage disequilibrium) to vary considerably about the values predicted by deterministic theory, so that observed D-values provide no reliable information on the selection regime in a natural population. Although the theory presented in this paper applies to neutral genes only, it is hoped that it contributes to future thinking of 'infinite-locus' models of selection.

The author is indebted to Dr P. van Beek for his suggestions concerning the analysis of alternating renewal processes and to Drs B. Charlesworth and W. G. Hill for their comments on an earlier draft of the paper.

REFERENCES

Avery, P. J. (1978). The effect of finite population size on models of linked overdominant loci. *Genetical Research* **31**, 239–254.

Avery, P. J. & Hill, W. G. (1979). Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. *Genetics* (to appear).

Bennet, J. H. (1953). Junctions in inbreeding. *Genetica* **26**, 392–406.

Bennet, J. H. (1954). The distribution of heterogeneity upon inbreeding. *Journal of the Royal Statistical Society* **16**, 88–99.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

Cockerham, C. C. (1971). Higher order probability functions of identity of alleles by descent. *Genetics* **69**, 235–246.

Cockerham, C. C. & Weir, B. (1968). Sib mating with two linked loci. *Genetics* **60**, 629–640.

Cox, D. R. (1962). *Renewal Theory*. London: Methuen.

Cox, D. R. & Smith, W. L. (1961). *Queues*. London: Chapman & Hall.

Fisher, R. A. (1949). *The Theory of Inbreeding*. London: Oliver & Boyd. (2nd edition, 1963.)

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

Fisher, R. A. (1959). An algebraically exact examination of junction formation and transmission in parent–offspring inbreeding. *Heredity* **13**, 179–186.

Franklin, I. R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology* **11**, 60–80.

Franklin, I. R. & Lewontin, R. C. (1970). Is the gene the unit of selection? *Genetics* **65**, 707–734.

Harris, D. L. (1964). Genotypic covariances between inbred relatives. *Genetics* **50**, 1319–1348.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York and London: Columbia University Press.

Robertson, A. (1977). Artificial selection with a large number of linked loci. In *Proceedings, International Conference on Quantitative Genetics*. Ames, Iowa: Iowa State University Press.

Stam, P. (1979). Interference in genetic crossing over and chromosome mapping. *Genetics* **92**, 573–594.

Sved, J. A. (1977). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.

Sved, J. A. & Feldmann, M. W. (1973). Correlation and probability methods for one and two loci. *Theoretical Population Biology* **4**, 129–132.

Weir, B. & Cockerham, C. C. (1974). Behaviour of pairs of loci in finite monoecious populations. *Theoretical Population Biology* **6**, 323–354.

Wills, C., Crenshaw, J. & Vitale, J. (1969). A computer model allowing maintenance of large amounts of genetic variability in Mendelian populations. I. Assumptions and results for large populations. *Genetics* **64**, 107–123.

APPENDIX

In order to obtain the constants $c_1, \ldots, c_6$ in equation (11), we observe that the relation

$$Q_{t+1} - \lambda P_{t+1} = \alpha\lambda(Q_t - \lambda P_t) + Y_t$$

holds for (i): $\lambda = \lambda_1$, $\alpha = \lambda_3/\lambda_1$, $Y_t = (1-\lambda_1)H_t - \lambda_3 S_t = U_t$

and for (ii): $\lambda = \lambda_3$, $\alpha = \lambda_1/\lambda_3$, $Y_t = (1-\lambda_3)H_t - \lambda_1 S_t = V_t$.

Thus

$$Q_t - \lambda_1 P_t = \lambda_3^t(Q_0 - \lambda_1 P_0) + \sum_{k=0}^{t-1} \lambda_3^{t-1-k} U_k \tag{A 1}$$

and

$$Q_t - \lambda_3 P_t = \lambda_1^t(Q_0 - \lambda_3 P_0) + \sum_{k=0}^{t-1} \lambda_1^{t-1-k} V_k. \tag{A 2}$$

Since $Q_0 = P_0 = 0$, combination of (A 1) and (A 2) yields

$$P_t = \frac{1}{\lambda_1 - \lambda_3}\left\{\sum_{k=0}^{t-1} \lambda_1^{t-k-1} V_k - \sum_{k=0}^{t-1} \lambda_3^{t-k-1} U_k\right\}. \tag{A 3}$$

Evaluation of the sums in (A 3) is straightforward because H_t and S_t can be written as

$$H_t = A\lambda_1^t + B\lambda_3^t, \quad S_t = C\lambda_5^t + D\lambda_6^t,$$

where $A = (1-\lambda_3)/(\lambda_1 - \lambda_3)$; $B = 1 - A$; $C = (\lambda_1 + \lambda_3 - \lambda_6)/(\lambda_5 - \lambda_6)$

and $D = 1 - C$.

Evaluation of the geometric series occurring in (A 3) and collecting the coefficients of λ_1^t, λ_3^t, λ_5^t and λ_6^t directly yields the constants $c_1, \ldots, c_6$. These are

$$c_1 = \frac{2}{a}\left\{\frac{2\lambda_1\lambda_3}{a^2} - \frac{\lambda_1}{f}\left(\frac{b}{\lambda_1-\lambda_5} - \frac{d}{\lambda_1-\lambda_6}\right)\right\},$$

$$c_2 = \frac{2}{\lambda_1}\left(\frac{1-\lambda_3}{a}\right)^2,$$

$$c_3 = \frac{2}{a}\left\{\frac{-2\lambda_1\lambda_3}{a^2} + \frac{\lambda_3}{f}\left(\frac{b}{\lambda_3-\lambda_5} - \frac{d}{-\lambda_3-\lambda_6}\right)\right\},$$

$$c_4 = \frac{2}{\lambda_3}\left(\frac{1-\lambda_1}{a}\right)^2,$$

$$c_5 = 2b\{\lambda_1/(\lambda_1-\lambda_5) - \lambda_3/(\lambda_3-\lambda_5)\}/af,$$

$$c_6 = 2d\{\lambda_3/(\lambda_3-\lambda_6) - \lambda_1/(\lambda_1-\lambda_6)\}/af,$$

where $\quad a = \lambda_1 - \lambda_3; \quad b = \lambda_1+\lambda_3-\lambda_6; \quad d = \lambda_1+\lambda_3-\lambda_5; \quad f = \lambda_5-\lambda_6.$

Consider two alternating Poisson processes (A and B) in stationary phase. Let the p.d.f.s of the 'waiting times' be

$$f(t) = \lambda e^{-\lambda t} \text{ (state } A)$$

and

$$g(t) = \alpha e^{-\alpha t} \text{ (state } B).$$

The cumulative distribution functions then are

$$\Pr(x \leqslant t) = F(t) = \int_0^t f(u)\,du = 1 - e^{-\lambda t} \qquad \text{(state } A)$$

and

$$\Pr(y \leqslant t) = G(t) = \int_0^t g(u)\,du = 1 - e^{-\alpha t}. \qquad \text{(state } B)$$

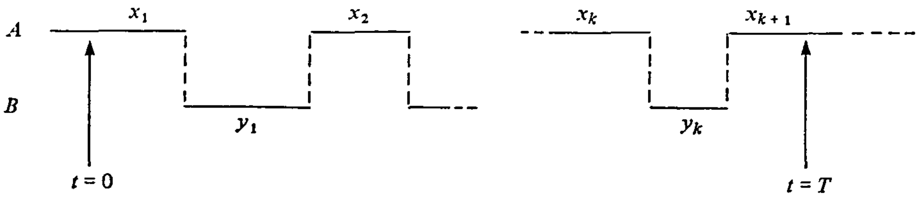
We further define the p.d.f.s of k-independent waiting times:

$$f_k(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!} \text{ (state } A)$$

and

$$g_k(t) = \frac{\alpha^k t^{k-1} e^{-\alpha t}}{(k-1)!} \text{ (state } B).$$

Suppose that at $t = 0$ the process is in state A (this has probability $p_A = \alpha/(\alpha+\lambda)$), and let there be an even number of events ($2k$) in the interval $(0, T)$. See diagram below.



Then at $t = T$, the process is in state A, the number of type A waiting times is $k+1$, and the number of type B waiting times is k. Let the probability of this state

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press

of affairs be p_{2k} (given state A at $t = 0$). Then the conditional p.d.f. of

$$y_1 + y_2 + \ldots + y_k = \Sigma y$$

is
$$\frac{1}{p_{2k}} \int_0^{T-t} g_k(t) f_k(z) \{1 - F(T-z-t)\} dz. \tag{A 4}$$

The factor $g_k(t)$ in the above expression is the k-fold convolution of $g(t)$, corresponding to the k type B waiting times. The factor $f_k(z)$ corresponds to the first k type A waiting times, whereas the last factor expresses the fact that the sum $x_1 + x_2 + \ldots + x_{k+1} + y_1 + \ldots + y_k = \Sigma x + \Sigma y$ is greater than T. The integration is over all possible values of z. Now

$$\int_0^{T-t} f_k(z) \{1 - F(T-z-t)\} dz = \frac{\lambda^k e^{-\lambda(T-t)}}{(k-1)!} \int_0^{T-t} z^{k-1} dz$$
$$= \frac{1}{\lambda} f_{k+1}(T-t).$$

So (A 4) can be written as $\quad g_k(t) f_{k+1}(T-t)/\lambda p_{2k}.$

The conditional p.d.f. of the sum of type A waiting times $(= \Sigma x = T - \Sigma y)$ thus is

$$f_{k+1}(t)\, g_k(T-t)/\lambda p_{2k}.$$

Next consider the case of an odd number $(2k+1)$ of events in $(0, T)$. By the same argument as for an even number of events, we then find that the conditional p.d.f. of the sum of type A waiting time is

$$f_{k+1}(t)\, g_{k+1}(T-t)/\alpha p_{2k+1}.$$

Using a similar argument for the case that at $t = 0$ the process is in state B (by interchanging the roles of f and g and those of Σx and Σy), and adding over all possible number of events (excluding zero events) one obtains

$$\phi(t) = \sum_{k=1}^{\infty} f_k(t) \left[p_A \left\{ \frac{1}{\lambda} g_{k-1}(T-t) + \frac{1}{\alpha} g_k(T-t) \right\} \right.$$
$$\left. + p_B \left\{ \frac{1}{\lambda} g_k(T-t) + \frac{1}{\alpha} g_{k+1}(T-t) \right\} \right],$$

taking $g_k \equiv 0$ for $k < 1$. After applying a change of scale, such that the entire distribution is within the interval $(0, 1)$ (it then applies to the proportion of time spent in state A), we obtain, after some rearrangements

$$\phi(t) = \frac{\alpha\lambda}{\alpha+\lambda} T \exp\left[-T\{\lambda t + \alpha(1-t)\}\right]$$
$$\times \left[\frac{\alpha t + \lambda(1-t)}{\sqrt{(\alpha\lambda t(1-t))}} I_1\{2T\sqrt{(\alpha\lambda t(1-t))}\} + 2I_0\{2T\sqrt{(\alpha\lambda t(1-t))}\} \right], \tag{A 5}$$

where I_0 and I_1 are modified Bessel functions of the first kind of order zero and one, respectively.

Besides the continuous part (A 5), there are probability condensations at $t = 0$ and $t = 1$, corresponding to the case that no events occur in $(0, T)$; the process is then in either of the two states during the complete interval. It is easily seen that the probability condensation at $t = 0$ equals

$$p_0 = p_A e^{-\alpha T} = \frac{\lambda}{\alpha+\lambda}\, e^{-\alpha T}.$$

Similarly, at $t = 1$
$$p_1 = p_B e^{-\lambda T} = \frac{\alpha}{\alpha+\lambda}\, e^{-\lambda T}.$$

In order to obtain the variance of the distribution we use the argument applied by Franklin (1977). Let z_t be a variable defined as follows

$$z_t = \begin{cases} 1 & \text{if the process is in state } A \text{ at } t, \\ 0 & \text{if the process is in state } B \text{ at } t. \end{cases}$$

Then
$$E(z_t) = E(z_t^2) = p_A.$$

Further
$$E(z_{t_0} \cdot z_{t_0+t}) = p_A\{p_A(1-e^{-(\alpha+\lambda)t}) + e^{-(\alpha+\lambda)t}\}$$
$$= p_A^2 + p_A p_B e^{-(\alpha+\lambda)t}. \tag{A 6}$$

(Cf. equation (24).) If we now think of a finite number (n) of points (loci) in the interval $(0, T)$, the proportion of loci in state A equals $1/n \sum_{i=1}^{n} z_i$.

The variance of this proportion is

$$E\left(\frac{1}{n}\sum_i z_i\right)^2 - \left\{E\left(\frac{1}{n}\sum_i z_i\right)\right\}^2$$
$$= \frac{1}{n^2}\left\{E\sum_i z_i^2 + E\sum_{i\neq j}\sum z_i z_j\right\} - p_A^2$$
$$= \frac{p_A}{n} + \frac{1}{n^2} E \sum_{i\neq j}\sum z_i z_j - p_A^2.$$

Changing to an infinite number of loci ($n \to \infty$), uniformly distributed over the interval $(0, T)$, this becomes

$$\int_0^T E(z_{t_0} \cdot z_{t_0+t})\, g(t)\, dt - p_A^2, \tag{A 7}$$

where $g(t)$ is the p.d.f. of the distance between two random points in $(0, T)$, i.e.

$$g(t) = 2(T-t)/T^2. \tag{A 8}$$

Insertion of (A 8) and (A 6) into (A 7) gives

$$\operatorname{var}(x) = p_A p_B \int_0^T \frac{2(T-t)}{T^2}\, e^{-(\alpha+\lambda)t}\, dt$$
$$= \frac{2p_A p_B}{(\alpha+\lambda)T^2}\left\{T - \frac{1}{\alpha+\lambda}(1-e^{-(\alpha+\lambda)T})\right\},$$

which, noting that $p_A = H$ and $\alpha+\lambda = Z/2H(1-H)$, is equation (23) of the text.

https://doi.org/10.1017/S0016672300014002 Published online by Cambridge University Press