THE APPROACH TO HOMOZYGOSITY IN A DIPLOID SELFING SERIES WITH NO MUTATION OR SELECTION

D. S. Robson

BU-162-M March, 1964

Abstract

In a diploid selfing series the expected rate of loss of heterozygosity is 50 percent per generation, provided that mutations and selection do not occur. If the heterozygous loci are randomly distributed over the n chromosome pairs then deviations from this mean rate must be approximately normally distributed and hence essentially dependent only on the cross-over probabilities between pairs of loci. The variance in this case indicates that a large number of linked loci will effectively behave the same as $1.5\ n\bar{k}$ unlinked loci, where $50\ \bar{k}$ is the average chromosome length measured in cross-over units. For N independently segregating loci the most probable number of generations required to reach complete homozygosity is approximately $\log_2 N$, and there is a 50 percent chance of achieving this state before $-\log_2 \left[1 - 2^{-\frac{1}{N}}\right]$ generations have elapsed. The mathematical analysis of this system is facilitated by its representation as a Markov chain.

Biometrics Unit, Plant Breeding Department, Cornell University

THE APPROACH TO HOMOZYGOSITY IN A DIPLOID SELFING SERIES WITH NO MUTATION OR SELECTION

BU-162-M D. S. Robson March, 1964

Introduction

A single line of descent in a selfing series executed under constant environmental conditions forms a simple Markov chain where the probability distribution of genotypes in any generation depends only on the genotype of the parent of that generation. In the absence of mutation and selection this conditional probability distribution for the offspring of any given parental genotype depends only on the linkage relationships among the segregating loci. Although rates of corssing-over may in general be influenced by genotype, the present analysis will presums these to be constant for all genotypes, as in classical linkage analysis.

The approach to homozygosity, irrespective of the particular genes being fixed and the order in which the different loci reach fixation, must ultimately be analyzed as a stochastic process of transition between states of non-decreasing degrees of heterozygosity. A selfed parent which is heterozygous at exactly N loci can only produce progeny which are heterozygous at N or fewer loci; thus, if S_i is the state defined by

S_i = the collection of genotypes g_i heterozygous at exactly i loci

then parent → offspring transitions of the form $S_i \rightarrow S_j$ are possible only if $i \geq j$. Although the process of transition between specific genotypes $g_i \rightarrow g_j$ heterozygous at i and j loci, respectively, is a simple Markov process, the

process on the combined states S_i will have the simple Markov structure if and only if the parent → offspring transition $g_i \rightarrow S_j$ has the same probability p_{ij} for every genotype g_i in S_i; that is, if and only if

$$\sum_{g_j \mid g_j \in S_j} P\left\{\text{progeny genotype} = g_j \mid \text{parent genotype} = g_i\right\} \equiv_{g_i \in S_i} p_{ij}$$

With linkage operating to determine the transition probabilities between individual genotypes, however, two different parental genotypes g_i and g_i' which are heterozygous at different sets of i loci will have different probabilities of producing offspring which are heterozygous at exactly j loci. Exploitation of Markov structure in computing transition probabilities for the S_i-process therefore requires the introduction of an intermediate Markov process with states S_{A_i} specifying not only the degree i of heterozygosity but also the collection A_i of loci at which heterozygosity occurs. The previously mentioned assumption that the probability distribution of cross-overs is independent of genotype assures that individual genotypes may be combined to this extent,

S_{A_i} = the collection of genotypes g_i heterozygous at the i loci in the set A_i and homozygous elsewhere,

while retaining simple Markov structure.

One generation transition probabilities

A parent which is heterozygous at the particular set A_i of i loci and homozygous at all other loci may, for present purposes, be regarded as a pair of homologous chromosomes, each bearing i genes but of unlike allelic form, and with one of the "chromosomes" representing the male gametic contribution of the grandparent, the other representing the female gamete of the grandparent. This second classification of alleles according to gametic origin is redundant since the two unlike allelic forms at each locus necessarily originate from unlike gametes; for any homozygous locus outside of A_i, however, classification according to allelic form would fail to distinguish the two genes at that locus. With one such homozygous locus A_1 appended to the "chromosome" in question, the transition from the state S_{A_i} of the parent to the state S_{A_j} of the offspring partitions into two events:

$S_{A_i} \rightarrow S_{A_j}$ and the offspring receives a gene at A_1 from each grandparental gamete

or

$S_{A_i} \rightarrow S_{A_j}$ and the offspring receives both genes at A_1 from the same grandparental gamete

Because cross-over probabilities are assumed to be independent of allelic forms, the probabilities of these two events would be unaltered if the two alleles at the A_1 locus were unlike; hence, transition probabilities in this Markov process are related by the recursion

$$p_{A_i,A_j} = P\left\{ \text{offspring in } S_{A_j} \mid \text{parent in } S_{A_i} \right\}$$

$$= P\left\{ \text{offspring in } S_{A_j+A_1} \mid \text{parent in } S_{A_i+A_1} \right\}$$

$$+ P\left\{ \text{offspring in } S_{A_j} \mid \text{parent in } S_{A_i+A_1} \right\}$$

$$= p_{A_i+A_1,A_j+A_1} + p_{A_i+A_1,A_j}$$

When rearranged in the form

$$(1) \qquad p_{A_i,A_j} = p_{A_i-A_1,A_j} - p_{A_i,A_j+A_1}$$

where A_1 is now an element of A_i-A_j, repeated application of this recursion to the terms on the right hand side of the equation produces the relation

$$(2) \qquad p_{A_i,A_j} = \sum_{k=0}^{i-j} (-1)^k \sum_{A_k \subset A_i-A_j} p_{A_j+A_k,A_j+A_k}$$

With relation (2) established, one-generation transition probabilities p_{A_i,A_j} need be defined in terms of cross-over probabilities only for the case $A_i=A_j$. The transition $S_{A_i} \rightarrow S_{A_i}$ will occur if, for any set of loci $A_j \subset A_i$, the male gamete of the parent carries paternal grandparental genes at the loci in A_j and maternal grandparental genes at the loci in A_i-A_j while the female

gamete of the parent carries the reverse configuration of paternal grandparental genes at the loci in $A_i - A_j$ and genes from the maternal grandparental gamete at the loci in A_j. Since these two kinds of gametes have the same probability of occurrence, say C_{A_i,A_j}, then the probability of their pairing is $C^2_{A_i,A_j}$, and since $S_{A_i} \rightarrow S_{A_i}$ is the union of all such mutually exclusive events for which $A_j \subset A_i$ then

$$(3) \qquad p_{A_i,A_i} = \sum_{j=0}^{i} \sum_{A_j \subset A_i} C^2_{A_i,A_j}$$

where

$$C_{A_i,A_j} = C_{A_i,A_i - A_j}$$

and

$$\sum_{j=0}^{i} \sum_{A_j | A_j \subset A_i} C_{A_i,A_j} = 1 \ .$$

Transition probabilities for the reduced, non-Markovian process may now be expressed in terms of the transition probabilities p_{A_i,A_i}, but only for transitions out of a maximal state S_N, which is equivalent to S_{A_N}, and which denotes the genotype of the initial parent of the selfing series. Thus,

$$\pi_{N,i} = P\left\{ \text{offspring heterozygous at exactly i loci} \mid \text{parent heterozygous at N loci} \right\}$$

$$= \sum_{A_i | A_i \subset A_N} p_{A_N,A_i}$$

$$= \sum_{j=0}^{N-i} (-1)^j \sum_{A_i \subset A_N} \sum_{A_j \subset A_N - A_i} p_{A_i+A_j, A_i+A_j}$$

or

$$(4) \quad \Pi_{N,i} = \sum_{j=1}^{N} (-1)^{j-i} \binom{j}{i} \sum_{A_j \subset A_N} p_{A_j, A_j} \quad .$$

t-generation transition probabilities

The probability $p^{(t)}_{A_i, A_j}$ that a line which has achieved the state S_{A_i} will be in state S_{A_j} after t more generations may be readily computed from the preceding results. The basic relation (1) among the one-generation transition probabilities applies, by the same argument, to the t-generation transition probabilities, and the corresponding relation (2) then expresses $p^{(t)}_{A_i, A_j}$ in terms of $p^{(t)}_{A_k, A_k}$. Since transitions $S_{A_k} \xrightarrow{(t)} S_{A_k}$ occur if and only if the system remains in state S_{A_k} for t generations then, by the Markov property,

$$p^{(t)}_{A_k, A_k} = p^{t}_{A_k, A_k}$$

and

$$(5) \quad p^{(t)}_{A_i, A_j} = \sum_{k=0}^{i-j} (-1)^k \sum_{A_k \subset A_i - A_j} p^{t}_{A_j+A_k, A_j+A_k}$$

so that

$$(6) \quad \Pi^{(t)}_{N,i} = \sum_{j=i}^{N} (-1)^{j-i} \binom{j}{i} \sum_{A_j \subset A_N} p^{t}_{A_j, A_j} \quad .$$

Rates of approach to fixation

The expected 50 percent reduction in heterozygosity per generation is a well known characteristic of the diploid selfing series; if a parent is heterozygous at N loci then the average number of heterozygous loci per offspring is N/2. This result is seen to hold regardless of the linkage relationships or even of any relationship between genotype and cross-over probabilities, since the average loss in heterozygosity at N loci is the sum of the average losses at each of the individual loci, and each individual locus has probability 1/2 of becoming homozygous in one generation. Likewise, the expected number of loci remaining heterozygous after t generations of selfing is $N/2^t$, irrespective of the linkage pattern. A formal, if awkward demonstration of this fact based on the preceding result (6) is given by

$$\sum_{i=0}^{N} i\, \Pi_{N,i}^{(t)} = \sum_{j=0}^{N} (-1)^j \sum_{A_j \subset A_N} p_{A_j,A_j}^{t} \sum_{i=0}^{j} (-1)^i\, i \binom{j}{i}$$

where

$$\sum_{i=0}^{j} (-1)^i\, i \binom{j}{i} = \begin{cases} -1 \text{ for } j = 1 \\ 0 \text{ for } j \neq 1 \end{cases}$$

so

$$(7) \qquad \sum_{i=0}^{N} i\, \Pi_{N,i}^{(t)} = \sum_{A_1 \subset A_N} p_{A_1,A_1}^{t} = N/2^t$$

since any single locus A_1 has probability 1/2 of remaining heterozygous in one generation of selfing.

The variation about an average 50 percent loss of heterozygosity per generation is, of course, profoundly affected by linkage; with no linkage the variance in the proportion lost is $1/4N$ while complete linkage would result in a corresponding variance of 1/4. In general, the variance of the number Q_N of loci becoming homozygous in one generation of selfing will exceed $1/4N$ by an amount equal to twice the sum of the covariances among the N loci. Thus, if c_{ij} denotes the probability that effectively one cross-over occurs between the i^{th} and j^{th} locus ($c_{ij} = 1/2$ if the i^{th} and j^{th} loci are on different chromosomes) then the variance of Q_N will be

$$\sigma^2_{Q_N} = \frac{N}{4} + N(N-1)\left[\frac{1}{4} - \overline{c(1-c)}\right]$$

where $\overline{c(1-c)}$ is the average variance of the number of effective cross-overs (0 or 1) between pairs of loci,

$$\overline{c(1-c)} = \frac{1}{\binom{N}{2}} \sum_{i<j} c_{ij}(1-c_{ij})$$

and could also be expressed as

$$\overline{c(1-c)} = \bar{c}(1-\bar{c}) - \sigma^2_c \quad .$$

The variance of the proportion $H = (N-Q_N)/N$ of heterozygotes remaining after one generation is therefore

$$\sigma_H^2 = \frac{1}{4N} + \left(1 - \frac{1}{N}\right)\left(\frac{1}{4} - \overline{c(1-c)}\right) ,$$

a result which may be obtained from (4) in the same manner that (7) was derived.

A rough indication of the magnitude of the effect which linkage might be expected to have on the variation of Q_N is illustrated by the case where the initial parent, heterozygous at N loci, is obtained by crossing two "unrelated" homozygotes. In such circumstances the distribution of the N sites of heterozygosity may be expected to be random (uniform) on the scale of cross-over units and randomly allocated among the n chromosome pairs. Thus, if the i^{th} chromosome includes 50 k_i cross-over units it may be expected to include $Nk_i / \sum_1^n k_j$ of the heterozygous loci randomly distributed along its length. The expected value of the cross-over probability c_{ij} for a random pair of such loci would then be

$$\mathcal{E}(c_{ij}) = \frac{1}{2}\left(1 - \frac{1}{k_i}\right)^2 + \frac{1}{2k_i}\left(1 - \frac{2}{3k_i}\right) ,$$

the first term on the right representing the expected contribution from pairs more than 50 cross-over units apart, which occur with probability $\left(1 - \frac{1}{k_i}\right)^2$, and the second representing the contribution of those less than 50 units apart. Similarly,

$$\mathcal{E}\left[c_{ij}(1-c_{ij})\right] = \frac{1}{4} - \frac{4k_i - 1}{24k_i^2}$$

so that expected variance of Q_N becomes

$$\mathcal{E}(\sigma^2_{Q_N}) = \frac{N}{4} + \frac{N(N-1)(4\bar{k}-1)}{24n\bar{k}^2}$$

or

$$\mathcal{E}(\sigma^2_H) = \frac{1}{4N} + \left(1 - \frac{1}{N}\right)\frac{4\bar{k}-1}{24n\bar{k}^2}$$

where $\bar{k} = \Sigma k_i/n$.

Since Q_N is, in fact, a sum of n independent random variables of bounded variation then the distribution of H must be approximately normal with mean 1/2 and the indicated variance, thus further implying that the characteristics of the distribution of H are essentially determined by the linkage relationships simply between pairs of loci. If N is large relative to the number n of chromosome pairs then the expected variance of H reduces to

$$\mathcal{E}(\sigma^2_H) \approx \frac{4\bar{k}-1}{24n\bar{k}^2} = \frac{1}{4}\left(\frac{1}{6n\bar{k}^2/(4\bar{k}-1)}\right)$$

which, when compared to the expression 1/4N arising with N independently segregating loci, implies that the linked system behaves essentially the same as

$$N = \frac{1}{1-4\bar{c}(1-\bar{c})} = \frac{1}{1-4\bar{c}(1-\bar{c})+4\sigma^2_c} \approx \frac{6n\bar{k}^2}{4\bar{k}-1} \approx 1.5\ n\bar{k}$$

unlinked loci.

A more detailed description of the stochastic approach to homozygosity is already contained in expression (6), and particularly in the special case

$$\Pi^{(t)}_{N,0} = \sum_{j=0}^{N} (-1)^j \sum_{A_j \subset A_N} p^t_{A_jA_j}$$

giving the probability that a line will reach complete homozygosity on or before the t^{th} generation. Thus, if $T_{N,0}$ is a random variable denoting the number of generations required for a line to first reach the homozygous state then the cumulative probability distribution of $T_{N,0}$ is

$$P(T_{N,0} \leq t) = \Pi_{N,0}^{(t)}$$

so that

$$P(T_{N,0} = t) = \Pi_{N,0}^{(t)} - \Pi_{N,0}^{(t-1)}$$

$$= \sum_{j=1}^{N} (-1)^{j-1} \sum_{A_j \subset A_N} p_{A_jA_j}^{t-1} (1-p_{A_jA_j})$$

The expected number of generations required to reach fixation is therefore

$$\mathcal{E}(T_{N,0}) = 1 + \sum_{t=1}^{\infty} \left(1-\Pi_{N,0}^{(t)}\right)$$

$$= 1 + \sum_{k=1}^{N} (-1)^{j-1} \sum_{A_j \subset A_N} \frac{p_{A_jA_j}}{1-p_{A_jA_j}} \quad .$$

In the case of independently segregating loci, $p_{A_j,A_j}=(1/2)^j$ and the probability distribution of $T_{N,0}$ is then

$$P(T_{N,0} \leq t) = \Pi_{N,0}^{(t)} = (1-2^{-t})^N$$

with a mean value of

$$\mathcal{E}(T_{N,0}) = 1 + \sum_{k=1}^{N} \binom{N}{k} (-1)^{k-1}(2^k-1)^{-1} \quad .$$

The modal or most probable value of $T_{N,0}$, found as the solution to the equation

$$\frac{d^2\Pi_{N,0}^{(t)}}{dt^2} = 0$$

is approximately $\log_2 N$ (= log N/log 2) or, more exactly, the solution to the equation

$$\frac{d\Pi_{N,0}^{(t)}}{dt} = \frac{d\Pi_{N,0}^{(t-1)}}{dt}$$

is

$$\text{mode} \approx \log \left\{ \frac{1-.5^{\frac{1}{N-1}}}{1-.5^{\frac{1}{N-1}}} \right\}$$

and the median, found as the solution to

$$\Pi_{N,0}^{(t)} = 0.5$$

is

$$\text{median} = -\log_2 \left(1 - .5^{\frac{1}{N}}\right)$$

The relation median < mode < mean obtains for all N but, as shown by Figure 1, as N gets large these three measures of central tendency become indistinguishable.

It is comforting to note that in accord with the rate of reduction in heterozygosity of 50 percent per generation, the modal number of generations required to first achieve homozygosity at half of the N loci,

$$\frac{d^2\Pi_{N,N/2}^{(t)}}{dt^2} = 0 \quad ,$$

is approximately given by

$$t = - \log_2 \left[\frac{1}{2} + 0\left(\frac{1}{\sqrt{N}}\right) \right] \approx 1 \quad .$$

Appendix

Some of the mathematically convenient properties of Markov chains which have been employed in this analysis of the selfing series are not mentioned in such standard references as Feller's Introduction to Probability Theory, and because these techniques have general applicability in the analysis of mating systems they perhaps deserve special mention here. These properties concern the approach to fixation or, in the terminology of a general Markov chain, the probability of absorption into a closed set of states. In a finite chain consisting of n different states $E_1, E_2, \cdots, E_n$ a subset $C = \{E_{i_1}, \cdots, E_{i_m}\}$ of $m \leq n$ states is said to be closed if transitions from states in C to states outside of C are impossible. The set C is therefore a trap; once the system enters C there is no escape, and the system is then said to have been absorbed in C. An individual state E_i is absorbing if transition out of E_i is impossible or, in other words, if the transition $E_i \rightarrow E_i$ has probability $p_{ii}^{(1)}=1$. Such a state E_i is also classed as recurrent, since recurrent states are defined as those to which ultimate return is certain, and every finite Markov chain includes at least one recurrent (though not necessarily absorbing) state. The remarks here will pertain primarily to the process of absorption into a closed set C taken to include all recurrent states. Asterisks will indicate new - or at least not widely known results.

If $p_{ij}^{(t)}$ denotes the conditional probability that the system will be in state E_j after t steps, given that it starts in state E_i, then for a Markov process the n x n matrix $\{p_{ij}^{(t)}\}$ can be computed as $\{p_{ij}^{(1)}\}^t$, the t^{th} power of the matrix of one-step transition probabilities. The conditional probability,

say $F_i(t)$, that the system will be in the closed set C after t steps, given that E_i is the starting state, is therefore

$$F_i(t) = \sum_{j \mid E_j \epsilon C} p_{ij}^{(t)} \quad ,$$

and the corresponding conditional probability, say $f_i(t)$, of reaching C for the first time on the t^{th} step is

$$(*) \qquad f_i(t) = F_i(t) - F_i(t-1) \approx \frac{dF_i(t)}{dt} \quad .$$

If C contains all of the recurrent states then $F_i(t)$ is the cumulative probability distribution of the random variable T_i representing the number of steps required to first reach C from the state E_i; that is,

$$(*) \qquad P(T_i \leq t) = F_i(t)$$

Such characteristics as the median (t_{med}) and mode (t_{mod}) of this distribution are therefore closely approximated by treating t as a continuous variable and solving for t in the equations

$$(*) \qquad F_i(t_{med}) = \frac{1}{2} \qquad \frac{d^2F_i(t_{mod})}{dt^2} = 0$$

Likewise, the moments of the distribution $F_i(t)$ of time to absorption may be computed as

$$(*) \qquad \mathcal{E}(T_i^r) = \sum_{t=1}^{\infty} t^r f_i(t) = \sum_{j \mid E_j \in C} \sum_{t=1}^{\infty} t^r [p_{ij}^{(t)} - p_{ij}^{(t-1)}]$$

or closely approximated by

$$(*) \qquad \mathcal{E}(T_i^r) = \int_1^{\infty} t^r dF_i(t) \quad .$$

The method suggested by Feller for computing the probability $f_i(t)$ has been to solve the recursion relations

$$f_i(t+1) = \sum_{k \mid E_k \in C} p_{ik}^{(1)} f_k(t) \quad ;$$

this method has the advantage of not requiring the calculation of the t^{th} power of the matrix $\{p_{ij}^{(1)}\}$. On the other hand, if $\{p_{ij}^{(1)}\}^t$ is to be computed for other purposes then the relation $f_i(t) = F_i(t) - F_i(t-1)$ may be employed directly without resorting to recursions.

A rather odd consequence of the above recurrence relation is that the moments of the random variables T_i may be computed as linear functions of the one-step transition probabilities $p_{ij}^{(1)}$, without ever computing the probability distributions of the T_i. Multiplying both sides the above equation by t and then summing over t from 1 to ∞ gives

$$\sum_{t=1}^{\infty} t f_i(t+1) = \sum_{k \mid E_k \in C} p_{ik}^{(1)} \sum_{t=1}^{\infty} t f_k(t)$$

where

$$\sum_{t=1}^{\infty} tf_i(t+1) = \sum_{t=1}^{\infty} (t+1)f_i(t+1) - \sum_{t=1}^{\infty} f_i(t+1)$$

$$= \sum_{t=1}^{\infty} tf_i(t) - \sum_{t=1}^{\infty} f_i(t)$$

$$= \mathcal{E}(T_i) - 1 \quad .$$

Thus,

$$(*) \qquad \mathcal{E}(T_i) - 1 = \sum_{k \mid E_k \in C} p_{ik}^{(1)} \, \mathcal{E}(T_k)$$

Similarly, for the second moments,

$$\sum_{t=1}^{\infty} t^2 f_i(t+1) = \sum_{t=1}^{\infty} (t+1)^2 f_i(t+1) - 2\sum_{t=1}^{\infty} (t+1)f_i(t+1) + \sum_{t=1}^{\infty} f_i(t+1)$$

so that

$$(*) \qquad \mathcal{E}(T_i^2) - 2\mathcal{E}(T_i) + 1 = \sum_{k \mid E_k \in C} p_{ik}^{(1)} \, \mathcal{E}(T_k^2)$$

and so on to give, consecutively, as many moments as desired.

Note that if C partitions into k disjoint closed sets of states, $C = \{C_1, C_2, \cdots, C_k\}$, then

$$\mathcal{E}(T_i^r) = \sum_{j=1}^{k} F_{ij}(\infty)\mathcal{E}_j(T_{ij}^r \mid \text{process is ultimately absorbed into } C_j)$$

where

$$F_{ij}(t) = \sum_{h|E_h \in C_j} p_{ih}^{(t)}$$

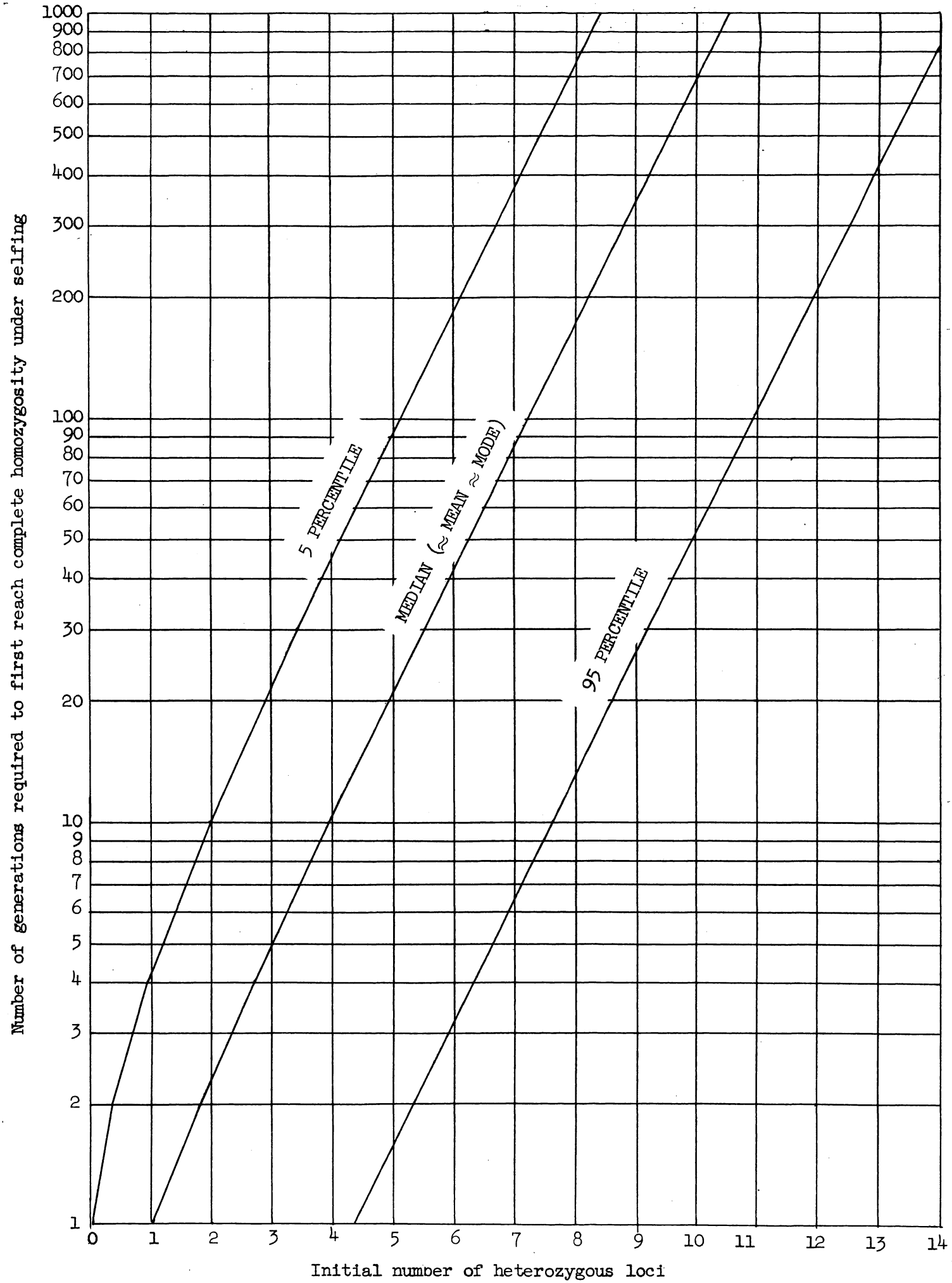
$$= \sum_{h|E_h \notin C} p_{ih}F_{hj}(t-1) + \sum_{h|E_h \in C_j} p_{ih}$$

and

$$F_{ij}(\infty) = \sum_{h|E_h \notin C} p_{ih}F_{hj}(\infty) + \sum_{h|E_h \in C_j} p_{ih} \quad .$$

Conditional moments of the time T_{ij} of absorption into C_j may then be computed as before but using the definitions

$$\sum_{t=1}^{\infty} t^r[F_{ij}(t) - F_{ij}(t-1)] = F_{ij}(\infty)\mathcal{E}_j(T_{ij}^r) \quad .$$


Number of generations required to first reach complete homozygosity under selfing
1000
900
800
700
600
500
400
300
200
100
90
80
70
60
50
40
30
20
10
9
8
7
6
5
4
3
2
1
5 PERCENTILE
MEDIAN (≈ MEAN ≈ MODE)
95 PERCENTILE
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Initial number of heterozygous loci