Genet. Res., Camb. (1978), **31**, *pp.* 255–264
With 2 text-figures
Printed in Great Britain

The time of detection of recessive visible genes in small populations

By ALAN ROBERTSON

*Department of Animal Husbandry, University of Sydney, Sydney, N.S.W., Australia**

(*Received* 15 *November* 1977)

SUMMARY

Homozygotes for recessive visible genes have often been discovered in lines under artificial selection, sometimes many generations from the start. As a help in the interpretation of this phenomenon, the distribution of the time to first detection as a homozygote of a recessive gene occurring only once in the initial generation has been obtained. Alternatively the results may be considered as referring to the time of first appearance as a homozygote of a new mutation occurring in a finite population. For a monoecious random mating population of size N with selfing permitted, the mean time to detection is very close to $2N^{\frac{1}{2}}$ over a range of N from 1 to 500 with a coefficient of variation of roughly 2/3 and a 95% upper limit about 2·5 times the mean. If selfing is prohibited, the mean time is increased by a little over 1 generation. The treatment is extended to cover the effects of artificial selection in favour of the heterozygote, of the frequency of occurrence in the initial generation and of the examination of more individuals each generation than are used as parents.

1. INTRODUCTION

It is not unusual for recessive visible genes to be discovered in populations under artificial selection, sometimes many generations from the start. Sometimes, but not always, these can be shown to have increased in frequency because of selection acting on the heterozygote. The question then arises as to whether they have been present since the start or whether they have been produced during selection by some event either of mutation or crossing over. There appears to be no treatment in the literature of the distribution of times of detection of initially rare alleles as homozygotes in small populations. This paper is intended to fill that gap. The results of course also refer to the distribution of times of first occurrence as a homozygote of a new mutant in a finite population.

* Permanent Address: Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN, Scotland

2. METHODS

(i) *Using transition matrices*

Defining E_i as the state of a monoecious random mating population with selfing in which there are i copies of the recessive gene, the process can be described by a matrix of transition probabilities p_{ij}, the probability that a population in state E_i in one generation will be in E_j in the next. In the much studied case of a population of N individuals with no selection, p_{ij} is given by

$$p_{ij} = {}^{2N}C_j \left(\frac{i}{2N}\right)^j \left(1-\frac{i}{2N}\right)^{2N-j}. \tag{1}$$

There are then two 'absorbing' states from which there can be no exit, with $i = 0$ and $2N$, representing loss and fixation of the allele respectively. If $V(0)$ is the initial row vector of frequencies of states and P is a $2N+1 \times 2N+1$ matrix whose elements equal p_{ij}, then the equivalent vector at generation t is given by

$$V(t) = V(0).P^t.$$

To deal with the present problem with this technique we introduce a new absorbing state, H, containing all populations in which recessive homozygotes are found. This involves subdividing all E_i's into a subset of D_i's in which there are no recessive homozygotes amongst the N parents. Let k_i be the proportion of states E_i in which there are no recessive homozygotes. This is given by the product of $i-1$ terms, i.e.

$$k_i = \frac{2N-i}{2N-1}\,\frac{2N-i-1}{2N-3} \cdots \frac{2N-2i+2}{2N-2i+3}.$$

The first term is the probability that a first recessive will be paired with a dominant, the second the probability that the second recessive will be likewise paired and so on. Obviously, if $i > N$, $k_i = 0$ and also $k_0 = k_1 = 1$. It can be shown that if i/N is small,

$$k_i = \exp\left(-i(i-1)/(4N)\right).$$

The matrix which describes this process is then derived from P by writing $k_j p_{ij}$ for p_{ij} and substituting H for E_{2N} with $p_{iH} = 1-\Sigma k_j p_{ij}$. I am indebted to Dr Peter Avery for pointing out that k_i can be expressed as

$$k_i = \frac{2^i N!\,(2N-i)!}{(2N)!\,(N-i)!}.$$

This leads to the expressions

$$k_j p_{ij} = \frac{N!}{j!\,(N-j)!}\left(\frac{i}{N}\left(1-\frac{i}{2N}\right)\right)^j \left(1-\frac{i}{2N}\right)^{2N-2j}$$

and

$$p_{iH} = 1-\left(1-\frac{i^2}{4N^2}\right)^N.$$

These will be recognized as, firstly, the probability of drawing j heterozygotes and $N-j$ dominant homozygotes and, secondly, the probability of drawing at

https://doi.org/10.1017/S0016672300018036 Published online by Cambridge University Press

least one recessive homozygote in a sample of N individuals from a population with a recessive gene frequency of $i/(2N)$ at Hardy–Weinberg equilibrium.

Selection can easily be introduced into (1) by replacing $i/(2N)$ with $i/(2N)+s(i/(2N))(1-i/N)$. The second term can be shown to be the expected change in gene frequency in D_i, a state in which there are i heterozygotes, when the latter have selective advantage s.

The initial vector $V(0)$ depends on the initial gene frequency assumed. For most of the calculations I assume that the recessive occurs initially only once so that $V(0)$ is then 0100.... The entry in $V(t)$ corresponding to H gives the proportion of populations in which a homozygote has been detected up to and including generation t.

(ii) *Simulation*

The matrix method has the advantage that it gives directly and without error the complete distribution of times of detection. As it assumes that genes are sampled from parents with replacement, it implicitly allows selfing and homozygotes may appear in the first progeny generation. To be more realistic, I decided to look at the effect of a division of the population into two sexes by simulation.

The simulation will be described first when selfing is allowed. Any gene in the population is identified by a four-digit integer. The first two digits refer to the generation in which it first occurred and the last two identify it precisely. For example, the genes present at the start are identified as 0001, 0002, etc. The first generation is formed by sampling at random from these with replacement. The genes are then compared in pairs – the first with the second, the third with the fourth, and so on. If a pair are identical, their 'age' is noted and all occurrences of the allele concerned are replaced by new alleles, each occurring once. For instance, if an allele occurs in a pair in the first generation (and three times in all) it is replaced by three new alleles 0101, 0102 and 0103. The majority of new alleles never appear as 'homozygotes' – they are lost by chance before this happens.

The output of the programme is a list of alleles lost by homozygosity and of new alleles inserted (as these occur) and then, at every 20th generation, the distribution of age of alleles at detection as homozygotes is printed out. The results should be the same as with the matrix method, though they are now subject to sampling error. The method is easily modified to include two sexes. Instead of sampling alleles at random from the whole $2N$, the first of each pair is sampled from the first N (the genes in males) and the next from the second N (those in females). This assumes that equal numbers of males and females are used as parents.

3. RESULTS

(i) *Matrix method*

In developing the programme, $2N$ was taken as 10 and subsequent runs were done with values of 20, 40, 100, 200 and 1000. Unless specially indicated, it is

assumed that the alleles being followed occurred only once in the initial sample. The distribution of times of first detection as a homozygote is shown in Fig. 1 for $2N = 40$. It has a mode at 3 generations and a long tail to higher values. It is to be expected that at high values the distribution of times will take the form $k(1-\lambda)^t$, a geometric distribution. In such situations the distribution of transient state frequencies tends to a constant form with all values decreasing by the same proportion, λ, each generation. In consequence, the number of populations entering H for the first time at any generation will decline by the same proportion.

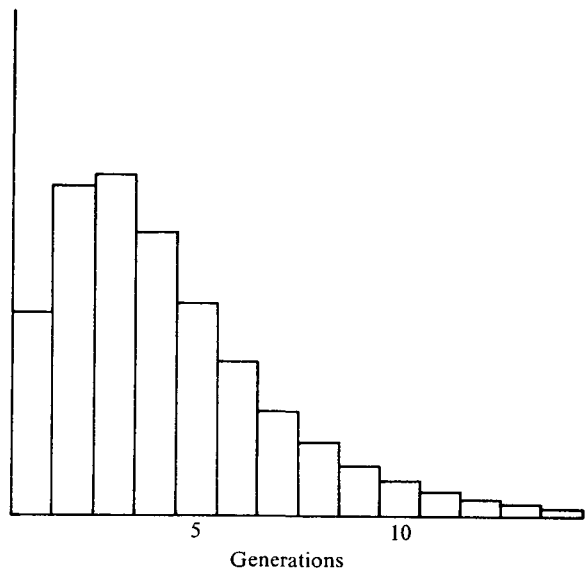

Fig. 1. The distribution of time to detection of a recessive initially occurring once in a population with 20 parents, with 20 individuals examined each generation.

The results for different values of N are given in Table 1. As N increases the proportion of genes detected declines and the average time to detection increases. Note that the coefficient of variation of time of detection is almost independent of N. This is a consequence of the 'geometric' tail of the distribution. If the entire distribution were geometric, the coefficient of variation would be $(1-\lambda)^{\frac{1}{2}}$. The fifth column gives the time at which 95% of all detected alleles have been found, which is close to $2\frac{1}{2}$ times the mean. For a geometric distribution, the 95% value is approximately 3 times the mean. Note that the mean time is approximately $1{\cdot}5/\lambda$. The relationship of λ to N will be discussed later.

The matrix method was used to examine two further points. As mentioned earlier, the effect of selection was looked at by modifying the gene frequency in the binomial expansions which lead to the fundamental matrix. Values of $2N$ equal to 20 and 100 were used. The results, given in Table 2, show that, although the probability of detection is much increased, the time to detection is changed little, even by an s value as high as 0·4. With no selection the number of

generations to detection is small and presumably the expected change of gene frequency due to selection in this time is not large enough to have much effect on the process.

On the other hand, the effect of the initial numbers of the recessive proved to be large. The results are summarized in Table 3. Examination of the results with $2N = 20$ and 40 suggested a generalization which was further examined with $2N = 80$.

It appears that the mean time to detection is mainly determined by the initial gene frequency, as exemplified by the rearrangement of the results in Table 4.

Table 1. *The time to detection as homozygotes of recessives initially present once only – matrix results, selfing permitted, λ is the limiting rate of decline of the gene frequency distribution*

		Time to detection (generations)			
$2N$	Proportion detected	Mean	S.D.	95% upper limit	$1/\lambda$
2	0·50	2·0	1·4	4·2	2·00
10	0·26	3·4	2·3	8·6	2·63
20	0·23	4·4	2·9	10·5	3·16
40	0·18	5·5	3·6	13·0	3·81
100	0·14	7·5	4·9	17·6	4·94
200	0·11	9·5	6·2	21·0	6·06
1000	0·07	16·3	10·5	37·0	9·92

Table 2. *The effect of selection – s is the selective advantage of the heterozygote relative to the wild-type homozygote: matrix results*

			Time to detection		
	s	Proportion detected	Mean	S.D.	95% upper limit
$2N = 20$	0	0·23	4·4	2·9	10·5
	0·1	0·31	4·4	2·8	10·2
	0·2	0·39	4·3	2·7	9·9
	0·4	0·53	3·9	2·3	8·6
$2N = 100$	0	0·14	7·5	4·8	17·6
	0·1	0·24	7·6	4·7	17·0
	0·2	0·34	6·3	4·2	15·5
	0·4	0·50	6·1	3·1	12·5

(ii) *Simulation*

Simulation was used partly as a check on the matrix method but also to allow the removal of the effect of selfing inherent in the latter. The results are summarized in Table 5 for $2N = 20$ and 100. The agreement between the two methods is satisfactory. The effect of excluding selfing is to increase the mean time to detection by about one generation, a result not altogether unexpected, as the allele now cannot be detected in the first generation. When $N = 2$, for instance,

the expected mean time with selfing is 2·54 generations and without it (full sib mating) it is 4·00.

All models so far used differ from an artificial selection experiment since I have assumed that only the individuals used as parents are examined each generation. Selection of course entails the observation of more animals than are

Table 3. *The effect of initial gene numbers: matrix results*

	Initial count	Proportion detected	Time to detection	
			Mean	S.D.
$2N = 20$	1	0·23	4·4	2·9
	2	0·44	3·6	2·7
	3	0·61	3·0	2·5
	4	0·75	2·5	2·2
$2N = 40$	1	0·18	5·5	3·6
	2	0·37	4·7	3·5
	3	0·55	4·0	3·3
	4	0·70	3·5	3·0
$2N = 80$	1	0·15	7·0	4·6
	2	0·30	6·1	4·4
	4	0·54	4·8	4·0
	6	0·73	3·7	3·5
	8	0·85	2·9	2·9

Table 4. *The data of Table 3 rearranged in terms of initial gene* frequency

Initial frequency	$2N$	Mean time to detection
0·025	40	5·5
	80	6·1
0·05	20	4·4
	40	4·7
	80	4·8
0·075	40	4·0
	80	3·7
0·100	20	3·6
	40	3·5
	80	2·9

used as parents. The simulation with two sexes was therefore modified so that the effect of this could be examined for selection intensities of 1 in 2 and 1 in 5 with $2N = 20$ and $2N = 100$. The results are given in Table 6. As would be expected the increased number of individuals examined increases the chance of detection and decreases the mean time.

In domestic animal populations there are usually less male than female parents and I have not examined this situation in detail. There would be some similarity to the results of Table 6 in that now, even in the absence of artificial selection, the effective population size from the point of view of genetic drift is less than the

total number of parents. There is the further complication that the statistics of the detection process will now depend on the sex in which the gene initially occurred.

Table 5. *A comparison of the two methods with one copy initially present – the bottom two lines show the effect of prohibiting selfing*

	Method	Proportion detected	Time to detection: Mean	Time to detection: S.D.
$2N = 20$	Matrix	0·23	4·4	2·9
	Simulation	0·22	4·2 ± 0·2	3·4
$2N = 100$	Matrix	0·14	7·5	4·9
	Simulation	0·14	7·2 ± 0·4	4·9
$2N = 1000$	Matrix	0·07	16·3	10·5
	Simulation	0·07	16·0 ± 0·4	10·0
	Two sexes			
$2N = 20$	Simulation	0·24	5·2 ± 0·2	2·3
$2N = 100$	Simulation	0·13	8·5 ± 0·3	5·0

Table 6. *The effect of increasing the number of individuals observed each generation*

(Simulation with no selfing. One copy initially present of a gene with no selective advantage in the heterozygote.)

	Individuals examined	Proportion detected	Time to detection: Mean	Time to detection: S.D.	Time to detection: 95 % upper limit
$2N = 20$	N	0·24	5·2 ± 0·2	2·3	10·0
	$2N$	0·25	4·6 ± 0·4	2·8	8·5
	$5N$	0·33	3·7 ± 0·3	2·0	7·5
$2N = 100$	N	0·13	8·5 ± 0·3	5·0	17·6
	$2N$	0·18	7·2 ± 0·3	4·3	17·2
	$5N$	0·22	5·9 ± 0·2	3·2	13·0

4. DISCUSSION

The results have interest from two points of view – from their relevance firstly to the experimental results mentioned earlier and, secondly, to the theory of finite populations. From the first aspect, the main points of interest are the small values of the mean times to detection and their low dependence on population size. On the other hand, because the distribution of detection times has a long tail upwards, the standard deviation is high – the 95% upper limit is of the order of 2·5 times the mean, and for most population sizes used in selection is of the order of 15 generations for a gene not subject to selection and occurring only once in the initial sample. Some variants of this basic model (the occurrence of the gene more than once in the initial sample (Table 3) and the examination of more individuals than are used as parents (Table 6)) reduced the times of detection

slightly while artificial selection on the recessive in the heterozygote proved to have surprisingly little effect.

Hollingdale (1971) discussed various aspects of the experimental evidence up to that time in *Drosophila*, though recessive visibles in selection lines are by no means confined to that species. Such results are always subject to the uncertainty that the homozygote may be detected only some generations after it first occurred because no one knew what to look for. The gene scabrous (*sca*) is particularly interesting in having occurred seven times in different selection lines, although only three initial stocks are involved. In one of these stocks, Canberra, the gene was extensively tested for and was not found in over 2000 chromosomes. In most selected lines N was between 10 and 20 and the average time of first detection in these was 12 generations. Our present results would suggest that some structural change during selection is probably involved, perhaps a rare crossover.

The sudden responses in females in 'low' abdominal bristle lines in *Drosophila* observed by Clayton & Robertson (1957) after many generations of selection are another possible example. Recent evidence (Frankham *et al.*, 1978) would suggest that these are caused by variation at the 'bobbed' locus. This is a complex locus, consisting of about 200 repeats of a unit coding for the ribosomal RNA, and alterations in the number of copies on a chromosome are known to occur, by a process which may involve unequal crossing over.

I commented earlier that, in such a stochastic process, the distribution of the frequencies of the transient states slowly approaches a constant form so that thereafter all frequencies decline at the same proportional rate and that this explained some features of the results. In the classical case with no selection and absorbing states E_0 and E_{2N}, the decline is by a fraction of $1/(2N)$ each generation and the limiting form is almost a uniform distribution, the frequencies in the extreme classes being slightly less than the intermediate values. In consequence, a modified time scale with the number of generations divided by N (essentially a measurement of time in terms of the expected inbreeding of the population) gives a greater generality. This is also true for mild selection acting on the individual when the distribution after t generations can be shown to be a function only of q_0 (the initial gene frequency), Ns and t/N (Robertson, 1960). It is known that, when N is large with no selection, a new mutant occurring in the population has a probability of $1/2N$ of eventually being fixed and that the average time to fixation is $4N$ generations.

It is clear that the results in Table 1 do not fit into such a framework and that T, the mean time to detection as a homozygote, and $1/\lambda$ show a dependence on N which is much less than the first power. In a related problem, that of selection against recessive lethals (Robertson & Narain, 1971), the rate of decline was found to be proportional to $1/\sqrt{N}$. In that problem, selection is against homozygotes as individuals – in the present, the appearance of a homozygote causes the loss of the whole line in which it occurs. Here a plot of log λ against log N showed that cube-root relationship is better and Fig. 2 gives the values of T and $1/\lambda$ plotted

against $N^{\frac{1}{2}}$. When N equals 1 there is only one transient state which halves in frequency each generation so that both T and $1/\lambda$ equal 2. At higher values of N, T is very close to $2N^{\frac{1}{2}}$ and $1/\lambda$ is about two-thirds of this. I would stress that this surprising relationship is not exact nor does it become more exact when N becomes larger, as is often the case in such studies. In fact, it is exact

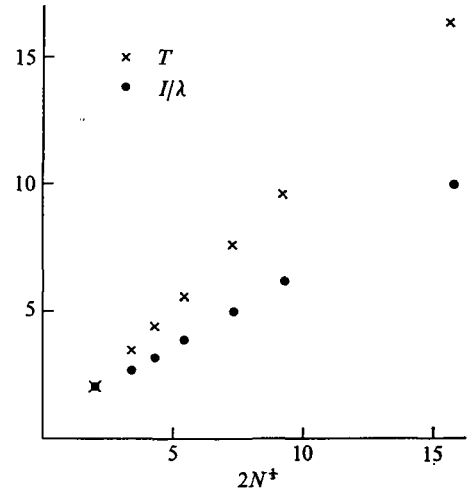

Fig. 2. The mean time to detection, T, and the reciprocal of the limiting rate of decline of the gene frequency distribution, λ, plotted againt $2N^{\frac{1}{2}}$, for a gene occurring once only in the initial generation.

when N equals 1 and the proportional error increases with N though it is only 3% when N equals 500. In the classical case with no selection, the probability of fixation is equal to $2/T$ ($= \lambda$). In the present case the probability of detection as a homozygote is approximately $1/T$ ($\sim 2\lambda/3$). At least we can arrive at the generalization that, over a wide range of N, the mean time to detection, T, of a recessive initially occurring once is approximately $2N^{\frac{1}{2}}$, with a standard deviation of $2T/3$ and a 95% upper limit of $2{\cdot}5T$.

When N equals 1, the distribution of time to detection is a geometric one over the entire range and at higher values of N it is geometric over higher values of t. It was observed that, at higher values of N, the distribution was proportional to t at low values of t. It can be shown that in large populations twice as many genes will be detected in the second generation as are in the first. If there are m copies of a gene in any generation, then if selfing is permitted, the probability that a homozygote will be detected in the next generation is $m^2/4N$. Since, in the first generation, m will be distributed as a Poisson distribution with mean unity if the gene occurred only once in the initial generation, it follows that the chance of detection in the second generation will be

$$\sum_{m=1}^{\infty} m^2/(4Nem!) = 1/2N,$$

twice the chance in the first generation.

https://doi.org/10.1017/S0016672300018036 Published online by Cambridge University Press

REFERENCES

CLAYTON, G. A. & ROBERTSON, A. (1957). An experimental check on quantitative genetic theory. II. The long-term effects of selection. *Journal of Genetics* **55**, 152–170.

FRANKHAM, R., BRISCOE, D. A. & NURTHEN, R. K. (1978). Unequal crossing over at the rRNA locus as a source of quantitive variation. *Nature* **272**, 80.

HOLLINGDALE, B. (1971). Analyses of some genes from abdominal bristle number selection lines in *Drosophila melanogaster*. *Theoretical and Applied Genetics* **41**, 292–301.

ROBERTSON, A. & NARAIN, P. (1971). The survival of recessive lethals in finite populations. *Theoretical Population Biology* **2**, 24–50.

ROBERTSON, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society* B **153**, 234–249.