Genet. Res., Camb. (1971), **18**, *pp.* 81–84
Printed in Great Britain

An invariant property of a structured population*

BY TAKEO MARUYAMA
National Institute of Genetics, Mishima, Japan

(*Received* 26 *March* 1971)

SUMMARY

Considering a single mutant gene in a geographically subdivided finite population, it is shown that the average number of heterozygotes, due to this gene, that appear in the population before fixation or loss of this gene is equal to twice the total population number. This is invariant under the geographical structure of the population.

The genetic variability and the number of heterozygotes maintained in a finite population are of considerable interest in population genetics. If a population has geographic structure and is not a panmictic unit as a whole, these quantities may depend on its structure. For example, suppose that the frequency of an allele is $\bar{x}$ in the whole population, but the occurrence of the allele is localized. Then we should expect more than $\bar{x}^2$ homozygotes for this allele.

Consider a locus, and assume that there is an allele present only once in the whole population. We may ask how many heterozygotes due to this allele will appear in the population before the allele becomes fixed in the population or lost from it, provided that no mutation occurs at this locus before fixation or loss of the allele and that no complete splitting of the population occurs. In this report, I shall show that this number is equal to twice the population number and that it is invariant under the population structure. The only assumptions necessary for this property to hold are (1) that the total population number (N_T) is constant with time and (2) that the allele is selectively neutral. Under these assumptions, the average number (H_T) of heterozygotes that appear in the population is

$$H_T = 2N_T. \tag{1}$$

This was first obtained by Kimura & Crow (1963) for a situation where every individual contributes exactly two gametes to the next generation. In that case, however, $H_T = 4N_T$. Their argument may be generalized to a situation where every individual has an equal expectation of the variance of offspring number throughout the habitat and in all generations. However, taking a different approach, I shall show that formula (1) holds for the following model in which the only biologically essential assumption is the neutrality of alleles under natural selection.

Model. Generation time is discrete. At the end of each generation the population

* Contribution number 827 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan. Aided in part by a grant-in-aid from the Ministry of Education, Japan.

is divided into nesting colonies where each colony produces exactly the same number of offspring as parents by random mating within the colony. Therefore the total population number (N_T) is constant over time. However, the number of colonies and the number of individuals in a colony may vary arbitrarily over time. Thus the expectation of the variance of offspring number may vary, depending upon the position of individual, and may also vary with time. Let $N_i^{(t)}$ be the number of individuals in colony i in some generation t, and let $m_{ij}^{(t)}$ be the probability that one born in colony i in generation t will participate in reproduction in colony j in the next generation, $t+1$. We assume that

$$\sum_j m_{ij}^{(t)} = 1 \quad \text{for all } i.$$

This assumes that the migration does not change the genetic composition in the entire population, though it will often change the local composition. Let $f_{ij}^{(t)}$ be the probability that two randomly chosen homologous genes, one each from colonies i and j, are the same allele (with $i = j$ the two genes are chosen without replacement). Define

$$F(t) \equiv \frac{1}{N_T^2} \sum_{ij} f_{ij}^{(t)} N_i^{(t)} N_j^{(t)}$$

and

$$F_0(t) \equiv \frac{1}{N_T} \sum_i f_{ii}^{(t)} N_i^{(t)}.$$

Then $F(t)$ is the probability that two homologous genes randomly chosen from the entire population are the same allele in some generation t, and $F_0(t)$ is the frequency of homozygotes. We assume no mutation during the time considered. The fraction of colony i which comes from colony k is

$$\frac{m_{ki}^{(t)} N_k^{(t)}}{N_i^{(t+1)}},$$

and

$$\frac{m_{ki}^{(t)} N_k^{(t)}}{N_i^{(t+1)}} \frac{m_{lj}^{(t)} N_l^{(t)}}{N_j^{(t+1)}} f_{kl}^{(t)}$$

is the probability that two homologous genes, one each from colonies i and j, are the same allele and they come from colonies k and l, provided $k \neq l$. With $k = l$, two homologous genes coming from a single colony are an identical gene in the previous generation with probability $1/2N_k^{(t)}$, and they are two different genes with probability $(1 - 1/2N_k^{(t)})$. Thus the probability that two homologous genes in colonies i and j are the same allele and both come from colony k is

$$\frac{m_{ki}^{(t)} N_k^{(t)}}{N_i^{(t+1)}} \frac{m_{kj}^{(t)} N_k^{(t)}}{N_j^{(t+1)}} \left\{ (1 - \tfrac{1}{2} N_k^{(t)}) f_{kk} + \tfrac{1}{2} N_k^{(t)} \right\}.$$

Summing over all possible combinations of k and l we have

$$f_{ij}^{(t+1)} = \sum_{k,l} \frac{m_{ki}^{(t)} N_k^{(t)}}{N_i^{(t+1)}} \frac{m_{lj}^{(t)} N_l^{(t)}}{N_j^{(t+1)}} f_{kl}^{(t)} + \sum_k \frac{m_{ki}^{(t)} N_k^{(t)}}{N_i^{(t+1)}} \frac{m_{kj}^{(t)} N_k^{(t)}}{N_j^{(t+1)}} \frac{1 - f_{kk}^{(t)}}{2N_k^{(t)}}. \tag{2}$$

https://doi.org/10.1017/S001667230001243X Published online by Cambridge University Press

Multiplying the left side and the right side of (2) by $N_i^{(t+1)}N_j^{(t+1)}/N_T^2$ and summing over i and j we have

$$F(t+1) = \frac{1}{N_T^2}\sum_{k,l} N_k^{(t)}N_l^{(t)}f_{kl}^{(t)}\sum_{i,j} m_{ki}m_{lj} + \frac{1}{2N_T^2}\sum_k (1-f_{kk}^{(t)})N_k^{(t)}\sum_{i,j} m_{ki}m_{lj}$$

$$\simeq F(t) + \frac{1}{2N_T}(1-F_0(t)).$$

Therefore

$$1-\frac{1-F(t+1)}{1-F(t)} = \frac{1}{2N_T}\frac{1-F_0(t)}{1-F(t)}. \tag{3}$$

This formula was first given by Robertson (1964) for the general model in which the number of offspring is assumed to be exactly two for every individual. Equation (3) can be written as

$$2N_T\{F(t+1)-F(t)\} = 1-F_0(t).$$

The right side of the above equation is the frequency of heterozygotes, and thus

$$H_T = \sum_{t=0}^{\infty} N_T\{1-F_0(t)\} = 2N_T^2\{F(\infty)-F(0)\}. \tag{4}$$

If there is initially one mutant gene in the entire population, $F(0) = 1-(1/N_T)$, and if complete splitting of the whole population does not occur, $F(\infty) = 1$. Thus, substituting, these quantities into (4), we have $H_T = 2N_T$.

We next state a formula analogous to (3) for a continuous generation model:

$$\frac{1}{1-F(t)}\frac{dF(t)}{dt} = \frac{1}{2N_T}\frac{1-F_0(t)}{1-F(t)}.$$

Consider many such loci. Assume that every mutation occurs at a locus which is homallelic in the entire population, and let U be the rate of occurrence of a mutation in the entire population. Then the average probability that a locus in an individual is heterozygous is

$$U \times 2N_T \times \frac{1}{N_T} = 2U.$$

Therefore, if the mutation rate per gene per generation is sufficiently low, the average number of heterozygous loci is invariant under the population structure.

In order to demonstrate the validity of formula (1), I have performed several computer simulations using the following two different models. Model I: individuals are distributed uniformly on a linear habitat with unit density; at the end of each discrete generation, each individual is replaced by a union of two gametes chosen randomly from its neighbourhood according to a normal distribution with variance σ^2. Model II: the population is divided into ten colonies and an individual moves from its native colony to another colony with probability m and it stays in its native colony with probability $1-m$. At the end of each generation, each colony produces exactly the same number of individuals as

https://doi.org/10.1017/S001667230001243X Published online by Cambridge University Press

there were parents by random mating within the colony. Thus colony sizes vary with time. In both models a mutant gene is introduced when and only when the whole population becomes homallelic. The total number of heterozygotes due to each mutant gene is counted, and in order to see the effect of the population structure, the average number of generations required for fixation of a mutant gene is recorded. All the probabilistic events are simulated by drawing pseudo-random numbers. The results of the simulations are presented in Table 1.

Table 1. *Results of simulations on the number of heterozygotes due to a single mutant gene*

Case	N_T	m or σ^2	H_T	T	Model
1	100	10	194·1	539·3	I
2	100	1	202·4	1622·9	I
3	100	0·5	212·7	4049·1	I
4	100	0·1	195·8	533·4	II
5	200	0·05	420·2	947·8	II
6	200	0·005	404·6	3425·5	II

(Note N_T = the total population number, σ^2 = the dispersion variance in model I, m = the migration rate in model II, H_T = the average number of heterozygotes that appeared, T = the average fixation time. The number of repetition for each case is $1000 \sim 5000$.)

The results on H_T agree well with the theoretical expectation $2N_T$ in all cases. It is remarkable to note that, despite the large differences in the fixation time, the number of heterozygotes are approximately the same in cases 2 and 3 and in cases 5 and 6.

I would like to thank Dr Motoo Kimura for the encouragement and help he has offered at all times. I would also like to thank Dr Joseph Felsenstein who corrected the English and offered useful criticisms.

REFERENCES

KIMURA, M. & J. F. CROW (1963). On the maximum avoidance of inbreeding. *Genetical Research* **4**, 399–415.

ROBERTSON, A. (1964). The effect of non-random mating within inbred lines of the rate of inbreeding. *Genetical Research* **5**, 164–167.