*Proc. Natl. Acad. Sci. USA*
Vol. 82, pp. 8611–8613, December 1985
Genetics

# Homozygosity, effective number of alleles, and interdeme differentiation in subdivided populations

**(geographical variation/migration/genetic drift/population genetics)**

THOMAS NAGYLAKI

Department of Molecular Genetics and Cell Biology, The University of Chicago, Chicago, IL 60637

*Communicated by James F. Crow, August 23, 1985*

ABSTRACT **The amount and pattern of genetic variability in a geographically structured population at equilibrium under the joint action of migration, mutation, and random genetic drift is studied. The monoecious, diploid population is subdivided into panmictic colonies that exchange migrants. Self-fertilization does not occur; generations are discrete and nonoverlapping; the analysis is restricted to a single locus in the absence of selection; every allele mutates to new alleles at the same rate. It is shown that if the number of demes is finite and migration does not alter the deme sizes, then population subdivision produces interdeme differentiation and the mean homozygosity and the effective number of alleles exceed their panmictic values. A simple relation between the mean probability of identity and the mean homozygosity is established. The results apply to a dioecious population if the migration pattern and mutation rate are sex independent.**

Although most studies of neutral models of geographical variation have involved the detailed investigation of particular migration patterns (see refs. 1 and 2 for refs.), several general properties of subdivided populations have also been established. In ref. 2, the strong- and weak-migration limits, properties invariant under population subdivision, and the approximation of diploid migration by gametic dispersion are reviewed. Here, we examine the mean homozygosity, effective number of alleles, and interdeme differentiation in a model with diploid migration and no self-fertilization. Our results also hold in a simpler, less realistic model with gametic dispersion and selfing in each deme at a rate equal to the reciprocal of the number of individuals in that deme (2).

## Formulation

We assume that a monoecious, diploid population is subdivided into a finite number of panmictic colonies that exchange migrants in a fixed pattern; colony $i$ contains $N_i$ adults. Self-fertilization does not occur; generations are discrete and non-overlapping; the analysis is restricted to a single locus in the absence of selection; every allele mutates to new alleles at the same rate $u$ ($0 < u < 1$). We measure time, $t$ ($=0, 1, 2, \ldots$), in generations. Random genetic drift operates through population regulation.

To begin the life cycle, the adults in each colony mate at random and produce without fertility differences a very large number of offspring. Migration and mutation follow, and finally population regulation returns the number of individuals in deme $i$ to $N_i$. Let $I_{ij}(t)$ represent the probability that two genes chosen at random from distinct adults just before reproduction in generation $t$, one from deme $i$ and one from deme $j$, are the same allele. We designate by $J_i(t)$ the probability that the two genes of an adult chosen at random from deme $i$ just before reproduction in generation $t$ are the same allele. Thus, $J_i$ is the expected homozygosity in colony $i$. Our formal scheme is displayed below:

$$\begin{array}{ccccc} \text{adults} & \xrightarrow[\text{reproduction}]{} & \text{zygotes} & \xrightarrow[\text{migration}]{} & \text{zygotes} \\ N_i, I_{ij}, J_i & & \infty, -, - & & \infty, -, - \end{array}$$

$$\begin{array}{ccccc} \xrightarrow[\text{mutation}]{} & \text{zygotes} & \xrightarrow[\text{regulation}]{} & \text{adults} \\ & \infty, -, - & & N_i, I'_{ij}, J'_i \end{array}$$

We use the backward migration matrix, $M$, to describe the pattern of dispersion: $m_{ij}$ denotes the probability that an individual in colony $i$ comes from colony $j$. The probabilities of identity in state satisfy (3, 4)

$$I'_{ij} = v\left[\sum_{kl} m_{ik}m_{jl}I_{kl} + \sum_{k} m_{ik}m_{jk}(2N_k)^{-1}(1 + J_k - 2I_{kk})\right], \qquad \textbf{[1a]}$$

$$J'_i = v \sum_k m_{ik}I_{kk}, \qquad \textbf{[1b]}$$

where $v = (1 - u)^2$ and the prime signifies the next generation. We place mutation after migration only for definiteness; actually, **[1]** holds if mutation occurs at any time between reproduction and regulation. Population regulation during this period would have no effect if it were sufficiently weak to leave very large numbers of zygotes. It is easy to see that **[1]** holds after one generation for a dioecious population if the migration pattern and mutation rate are sex independent and we take

$$N_i = 4N_i^{(1)}N_i^{(2)}/(N_i^{(1)} + N_i^{(2)}),$$

where $N_i^{(1)}$ and $N_i^{(2)}$ denote the numbers of males and females in deme $i$ (3, 4).

As $t \to \infty$, $[I_{ij}(t), J_i(t)]$ converges at least as fast as $v^t$ to the unique solution of (3, 4)

$$\hat{I}_{ij} = v\left[\sum_{kl} m_{ik}m_{jl}\hat{I}_{kl} + \sum_k m_{ik}m_{jk}s_k\right], \qquad \textbf{[2a]}$$

$$\hat{J}_i = v \sum_k m_{ik}\hat{I}_{kk}, \qquad \textbf{[2b]}$$

$$s_k = (1 + \hat{J}_k - 2\hat{I}_{kk})/(2N_k). \qquad \textbf{[2c]}$$

Since $u > 0$, some genetic variability is preserved: **[2]** is not satisfied if $\hat{I}_{ij} = 1$ and $\hat{J}_i = 1$ for every $i$ and $j$.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

## Lemma

In our analysis, we shall need the fact that $s_i \geq 0$; i.e.,

$$\frac{1}{2}(1 + \hat{J}_i) \geq \hat{I}_{ii} \qquad [3]$$

in every deme. This inequality is a special case of the combinatorial result stated and proved below and is therefore independent of the evolutionary details of our model.

Consider a population of $N$ monoecious, diploid individuals and focus attention on a single locus with alleles $A_1, \ldots, A_r$. The genotypic distribution is arbitrary. Let $x$ designate the probability that $K$ genes sampled *with replacement* from an individual chosen at random are the same allele. We denote by $y$ the probability that $K$ genes chosen at random from distinct individuals are the same allele. Then

$$x \geq y. \qquad [4]$$

Clearly, the special case of [4] with $K = 2$ implies [3]. This special case can be established by introducing genotypic frequencies. Here, a shorter, more informative proof, due to R. R. Bahadur and S. L. Zabell (personal communication), of the general inequality is presented. The proof has two parts: first, we show that it suffices to prove [4] for $N = K$ and then we establish [4] under this simplification.

Let $C$ represent a set of $K$ distinct individuals chosen at random and denote by $q_C$ the probability of choosing this set. Let $x_C$ designate the probability that $K$ genes sampled with replacement from an individual chosen at random from $C$ are the same allele. Since the random choice of $C$ followed by the random choice of an individual from $C$ produces an individual chosen at random from the entire population, we have

$$x = \sum_C x_C q_C. \qquad [5]$$

Furthermore,

$$y = \sum_C y_C q_C, \qquad [6]$$

where $y_C$ signifies the probability that $K$ genes chosen at random, one from each individual in $C$, are the same allele. Obviously, if $x_C \geq y_C$ for every set $C$, then [5] and [6] imply [4].

Suppose now that $N = K$ and number the individuals in the population. Denote by $p_{ij}$ the probability that a gene chosen at random from individual $i$ is $A_j$. Since the arithmetic mean is no less than the geometric mean, we obtain

$$x = \sum_{j=1}^{r} \frac{1}{K} \sum_{i=1}^{K} p_{ij}^{K} \geq \sum_{j=1}^{r} \prod_{i=1}^{K} p_{ij} = y. \qquad [7]$$

Equality holds in [7] if and only if $p_{ij}$ is independent of $i$ for every $j$, i.e., if and only if all individuals have the same (homozygous or heterozygous) genotype.

Scrutiny of our proof reveals that [4] holds regardless of ploidy; in fact, it is not even necessary for all $N$ individuals to have the same ploidy. Furthermore, [4] is also valid if the $K$ individuals in the definition of $y$ are sampled with replacement. If all individuals have the same ploidy, this is equivalent to sampling with replacement $K$ genes from the population.

## Analysis

Defining the matrix $A$ and the vector $\mathbf{b}$ by

$$A = vM \otimes M, \qquad b_{ij} = v \sum_k m_{ik} m_{jk} s_k, \qquad [8]$$

where the notation signifies that $A$ is a Kronecker product, we rewrite [2a] as the vector equation

$$\hat{\mathbf{I}} = A\hat{\mathbf{I}} + \mathbf{b}. \qquad [9]$$

Treating $\mathbf{b}$ as known, we can solve [9] immediately:

$$\hat{\mathbf{I}} = (I - A)^{-1}\mathbf{b} = \sum_{n=0}^{\infty} A^n \mathbf{b}, \qquad [10]$$

in which $I$ denotes the identity matrix. Since $M$ is stochastic, its spectral radius is one. Hence, the spectral radius of $A$ is $v < 1$, which implies convergence of the sum in [10]. Substituting [8] into [10] leads to

$$\hat{I}_{ij} = \sum_{n=0}^{\infty} v^{n+1} \sum_p m_{ip}^{(n+1)} m_{jp}^{(n+1)} s_p, \qquad [11]$$

where $m_{ij}^{(n)} = (M^n)_{ij}$. Malécot (5) has obtained a similar result. Since $s_p \geq 0$ for every $p$, from [11] we infer at once, for *any* migration pattern,

$$\hat{I}_{ij} \leq \sum_{n=0}^{\infty} v^{n+1} \sum_p \frac{1}{2} \{[m_{ip}^{(n+1)}]^2 + [m_{jp}^{(n+1)}]^2\} s_p$$

$$= \frac{1}{2}(\hat{I}_{ii} + \hat{I}_{jj}); \qquad [12]$$

equality holds if and only if $m_{ip}^{(n+1)} = m_{jp}^{(n+1)}$ for every $n$ and $p$.

Let $P_{i\alpha}$ represent the frequency of the allele $A_\alpha$ in deme $i$. Had we sampled with replacement in the definition of $I_{ij}$, then [12] would have followed at once from the trivial inequality

$$\sum_\alpha P_{i\alpha} P_{j\alpha} \leq \frac{1}{2} \sum_\alpha (P_{i\alpha}^2 + P_{j\alpha}^2).$$

To see that [12] is not merely a combinatorial result, focus attention on demes $i$ and $j$, $i \neq j$, and suppose that $N_i = N_j$, no allele occurs more than once in either deme, and the two demes are genetically identical. Then $I_{ij} > 0$ and $I_{ii} = I_{jj} = 0$; [12] fails because the population is not at equilibrium.

We assume now that migration is *conservative* (6); i.e., it does not change the deme sizes. Define the proportion of adults in deme $i$, the total population number, the global and local means of the probability of identity, and the mean homozygosity:

$$\kappa_i = N_i/N_T, \qquad N_T = \sum_i N_i, \qquad [13]$$

$$\hat{\bar{I}} = \sum_{ij} \kappa_i \kappa_j \hat{I}_{ij}, \qquad \hat{\bar{I}}_0 = \sum_i \kappa_i \hat{I}_{ii}, \qquad \hat{\bar{J}} = \sum_i \kappa_i \hat{J}_i. \qquad [14]$$

Averaging [12] with the aid of [14] shows that the mean probability of identity cannot exceed the mean homozygosity:

$$\hat{\bar{I}} \leq \frac{1}{2} \sum_{ij} \kappa_i \kappa_j (\hat{I}_{ii} + \hat{I}_{jj}) = \hat{\bar{I}}_0. \qquad [15]$$

Thus, population subdivision produces interdeme differentiation. Equality occurs in [15] if and only if

$$m_{ij}^{(n+1)} = c_j^{(n+1)} \qquad [16]$$

for every $i, j$, and $n$, for some $c_j^{(n+1)}$. For conservative migration (6),

$$\kappa_j = \sum_i \kappa_i m_{ij}. \qquad [17]$$

On the one hand, if **[16]** holds, we take $n = 0$ and substitute $m_{ij} = c_j$ into **[17]** to conclude that $c_j = \kappa_j$ for every $j$, which means that the population is panmictic. On the other hand, if we posit panmixia, then $m_{ij} = \kappa_j$ for every $i$ and $j$ (7). But if $m_{ij}^{(n)} = \kappa_j$ for every $i$ and $j$, then

$$m_{ij}^{(n+1)} = \sum_k m_{ik}^{(n)} m_{kj} = \kappa_j, \qquad [18]$$

so **[16]** holds. Therefore, equality occurs in **[15]** if and only if the entire population mates at random.

Next, we average **[2a]** and **[2b]**, appealing to **[2c]**, **[13]**, **[14]**, and **[17]**:

$$\hat{\bar{I}} = v[\hat{\bar{I}} + (2N_T)^{-1}(1 + \hat{\bar{J}} - 2\hat{\bar{I}}_0)], \qquad [19a]$$

$$\hat{\bar{J}} = v\hat{\bar{I}}_0, \qquad [19b]$$

whence

$$\hat{\bar{I}} = \frac{v[1 - (2 - v)\hat{\bar{I}}_0]}{2N_T(1 - v)}. \qquad [20]$$

Eqs. **19b** and **20** enable us to relate the mean probability of identity to the mean homozygosity:

$$\hat{\bar{I}} = \frac{v - (2 - v)\hat{\bar{J}}}{2N_T(1 - v)}. \qquad [21]$$

There is a simpler, analogous result for gametic dispersion (8, 9).

Combining **[15]** and **[20]** yields

$$\hat{\bar{I}}_0 \geq v/[2N_T(1 - v) + v(2 - v)] = I_r, \qquad [22a]$$

where $I_r$ denotes the probability of identity between distinct individuals in a panmictic population (5, 10). From **[19b]** and **[22a]** we obtain

$$\hat{\bar{J}} \geq v^2/[2N_T(1 - v) + v(2 - v)] = J_r, \qquad [22b]$$

in which $J_r$ represents the expected homozygosity in a panmictic population (5, 10). Thus, the expected homozygosity is at least as great as for panmixia; equality holds in **[22]** if and only if the entire population mates at random. Inserting **[22a]** into **[20]** gives

$$\hat{\bar{I}} \leq I_r; \qquad [23]$$

i.e., the mean probability of identity is decreased by population subdivision. For the effective number of alleles (11, 12), we find

$$n_e = 1/\hat{\bar{I}} \geq 1/I_r = n_e^r, \qquad [24]$$

where $n_e^r$ designates the panmictic value; equality holds in **[24]** if and only if the population mates at random. Eq. **24** proves that population subdivision raises at least one index of genetic diversity.

Thus, in addition to the invariance result **[21]**, we have established the inequalities

$$\hat{\bar{I}} \leq I_r \leq \hat{\bar{I}}_0, \qquad J_r \leq \hat{\bar{J}}, \qquad [25]$$

in which equality holds if and only if the entire population mates at random.

The inequalities **[23]** and **[24]** can fail for *nonconservative* migration. In the strong-migration limit, an effective population number $N_e < N_T$ appears in the theory, and we fix $M$ and let $u \to 0$ and $N_e \to \infty$ so that $N_e u$ remains fixed. One finds (4)

$$\hat{I}_{ij} \to \phi \qquad \hat{J}_i \to \phi, \qquad [26a]$$

$$\phi = 1/(1 + 4N_e u) > 1/(1 + 4N_T u). \qquad [26b]$$

Since the right side of **[26b]** is the limit of $I_r$ and $J_r$ as $u \to 0$ and $N_T \to \infty$ with $N_T u$ fixed, we conclude that **[23]** and **[24]** must be reversed in this case.

I am very grateful to S. Sawyer for helpful critical comments and to R. Bahadur and S. Zabell for the proof of the lemma. This work was supported by National Science Foundation Grant DEB81-03530.

1. Nagylaki, T. (1984) in *Proc. Symp. Appl. Math.*, Vol. 30, ed. Levin, S. A. (Am. Math. Soc., Providence, RI), pp. 19–36.
2. Nagylaki, T. (1985) in *Stochastic Spatial Processes*, ed. Tautu, P. (Springer, Berlin), in press.
3. Sawyer, S. (1976) *Ann. Probab.* **4**, 699–728.
4. Nagylaki, T. (1983) *Theor. Popul. Biol.* **24**, 268–294.
5. Malécot, G. (1948) *Les mathématiques de l'hérédité* (Masson, Paris) [Extended translation: Malécot, G. (1969) *The Mathematics of Heredity* (Freeman, San Francisco)].
6. Nagylaki, T. (1980) *J. Math. Biol.* **9**, 101–114.
7. Nagylaki, T. (1977) *Selection in One- and Two-Locus Systems* (Springer, Berlin), p. 142.
8. Crow, J. F. & Maruyama, T. (1971) *Theor. Popul. Biol.* **2**, 437–453.
9. Nagylaki, T. (1982) *J. Theor. Biol.* **99**, 159–172.
10. Malécot, G. (1946) *C. R. Acad. Sci.* **222**, 841–843.
11. Kimura, M. & Crow, J. F. (1964) *Genetics* **49**, 725–738.
12. Maruyama, T. (1970) *Theor. Popul. Biol.* **1**, 273–306.