MIGRATION AND MUTATION

Contents

This section is based on two papers [5] [3] that my colleague Barrie Latter and I wrote together. It started off when both of us independently found that the results of computer simulations we were doing didn't quite agree with what was expected from Sewall Wright's theory.

1. THE ORIGINAL MIGRATION MODEL

As is often the case with population genetics theory, there are two ways of looking at the effects of migration. The first follows gene frequencies and variances of gene frequencies in populations, giving the most direct picture of what is expected to happen in a population. However inbreeding probability methods can give the same results more easily, even though their applicability is sometimes harder to see. In the following I have attempted to give both approaches, since this is what we did in our papers.

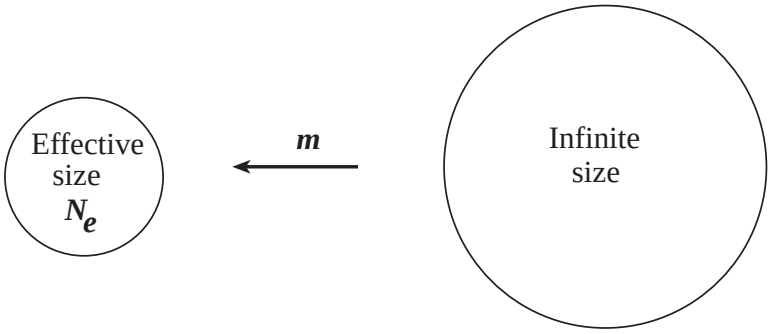

FIGURE 1. Model of a small (island) population receiving migrants from a large (mainland) population

1.1. **Migration and drift - frequency approach.**
Consider a single gene, whose frequency in the population is p. It is convenient to think of this population as one of a number of similar populations, say the kth population. The overall gene frequency, ie the frequency in the large population providing the migrants, is P. The variance of gene frequencies between populations at time t is σ_t^2. The mean can be assumed to be P.

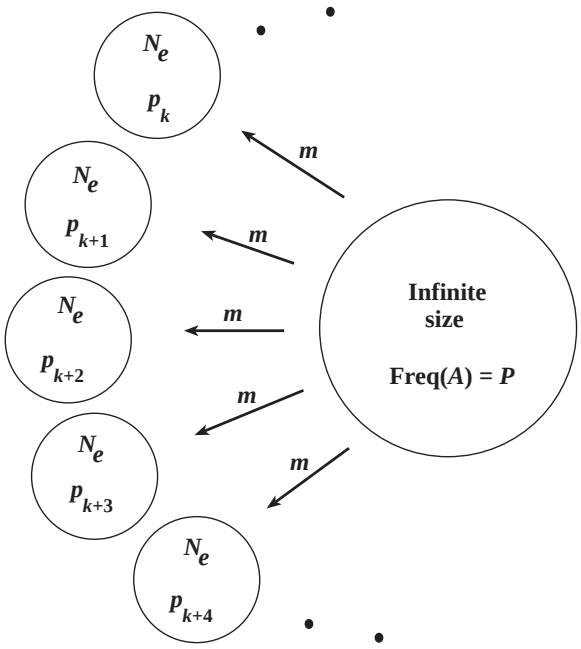

FIGURE 2. Replicate populations

Drift increases the variance in frequency of the gene. The variance before migration is taken into account, σ^{o2}, is:

$$\sigma^{o2}_{t+1} = \frac{P(1-P)}{2N_e} + (1 - \frac{1}{2N_e})\sigma^2_t \tag{1}$$

This calculation comes from our paper [5], but I believe it is basically Wright's argument. It assumes the Wright-Fisher model of population replacement, where each gene is chosen independently from the previous generation.

Migration then reduces the variance by the factor $(1-m)^2$, giving

$$\sigma^2_{t+1} = (1-m)^2\,[\frac{P(1-P)}{2N_e} + (1 - \frac{1}{2N_e})\sigma^2_t] \tag{2}$$

Putting $\sigma^2_{t+1} = \sigma^2_t$ gives an equilibrium value of

$$\hat{\sigma^2} = \frac{(1-m)^2 P(1-P)}{1 + (2N_e - 1)(2m - m^2)} \tag{3}$$

1.2. **Inbreeding approach.**
This argument comes from Section 6.6 in Crow and Kimura's 1970 textbook [1], usually my bible for elementary population genetic calculations. What inbreeding does is to increase homozygosity through identity-by-descent (IBD). If f_t is the probability of IBD in generation t, or the autozygosity [1], then this is increased between generations according to the relationship

$$f_{t+1} = \frac{1}{2N_e} + (1 - \frac{1}{2N_e})f_t \tag{4}$$

Another way of writing this is

$$1 - f_{t+1} = (1 - \frac{1}{2N_e})(1 - f_t)$$

from which it can be seen that heterozygosity goes down by a fraction $1 - \frac{1}{2N_e}$ in each generation.

Migration is introduced in the same way as previously, modifying equation (4) as follows (cf equation 6.6.2 in Crow and Kimura)

$$f_t = (1-m)^2\,[\frac{1}{2N_e} + (1 - \frac{1}{2N_e})f_{t-1}] \tag{5}$$

The justification for this is that $(1-m)^2$ is the probability that the two genes in a homozygote are non-immigrant genes. Any immigrant

genes are not IBD with each other nor with any genes within the island population.

An equilibrium is now expected in which autozygosity does not become complete. At equilibrium, putting f_{t+1} equal to f_t, the equilibrium value of f becomes

$$\hat{f} = \frac{(1-m)^2}{1+(2N_e-1)(2m-m^2)} \tag{6}$$

Under the assumption that m is small and N_e large, equation (6) becomes

$$\hat{f} = \frac{1}{1+4N_e m} \tag{7}$$

This is the form of the equilibrium that one normally sees.

1.3. **The relation between the frequency and probability approaches.**

Equations (3) and (6) are basically the same, if we replace $\hat{\sigma^2}$ by $\hat{f}P(1-P)$. But at first glance they seem unrelated. One refers to the variance of gene frequencies between replicate populations and the other to the expected level of autozygosity within a population.

The formal relationship between the two can be seen by considering the expected frequencies of genotypes at a single locus. I've simplified here by putting $Q = 1 - P$.

	AA	Aa	aa
$1-f$	P^2	$2PQ$	Q^2
f	P	-	Q
Sum	P^2+fPQ	$2PQ-2fPQ$	Q^2+fPQ
Wahlund	$P^2+\sigma^2$	$2PQ-2\sigma^2$	$Q^2+\sigma^2$

The line 'Sum' gives the frequencies summed over the cases of non-IBD and IBD. The following line gives the frequencies looked at from the point of view of homozygosity and heterozygosity in individual populations in terms of the overall gene frequencies, a result sometimes known as 'Wahlund's principle'. The two lines are equivalent if σ^2 is equal to fPQ.

2. THE NEW MODEL

It is equation (3), or alternatively (6), that Barrie Latter and I both had trouble in confirming in our computer simulations. Our answers

weren't far off, but there was definitely something that we were doing in the simulations that wasn't taken into account in the formula. We decided that this had to do with the way that migration was put into the computer simulation models.

2.1. **Four interpretations of migration.**

There are at least four ways in which migration can be interpreted:

[1] m can be interpreted as a frequency. So if the frequency of a particular gene before migration is p in the island population and P in the mainland, then after migration the frequency in the island population is $(1-m)p + mP$. This model can be described as 'deterministic migration'

[2] A fixed number of migrant genes comes from the mainland to the island. A migration rate of m then corresponds to a migration of $2N_e m$ individual genes from the mainland to the island. This looks superficially the same as [1] but it isn't the same. Each of these genes is sampled, rather than a fixed gene.

[3] Each gene in the island population has probability m of being an immigrant gene and probability $1-m$ of being a non-immigrant gene. Clearly this interpretation is different to the first two because it involves a variable rather than fixed number of migrant genes.

[4] Class [3] refers to migration of individual genes rather than pairs of genes, as expected if migration involves diploid individuals. Our first paper [5], which I wrote, with migration from the mainland into a single island population, ignores this complication. I have to admit that I entirely forgot about this case, but Barrie included it later in [3] which he wrote, dealing with migration between different islands.

2.2. **Census before or after migration?**

One more complication concerns the time at which the population is censused to calculate gene frequencies or homozygosity. This can be done either before or after migration has occurred. An alternative way of looking at this is that the order of migration and population replacement can be reversed. Either is legitimate.

The following discussion is simplified by assuming that population replacement occurs first, followed by migration, and that frequency calculations are made after migration.

2.3. **The essence of the difference between migration models.** In model [1], the frequency of A genes coming into the population is exactly P, coming in at a rate of exactly m. In adding migration to the gene frequency model, ie going from equation (1) to (2), it is essentially these assumptions that are made. Any variance in either of these parameters would contribute to σ^2.

There is one circumstance where the deterministic model might be realistic. In fish populations, or others where there is a spraying of a large number of gametes and/or eggs, migration might be modeled in this way. But in terms of adult migration the model seems unrealistic.

The remaining models all have the effect of adding extra components to the variance, due either to variation in the frequency of P (models [2] - [4]), or variation in m (models [3] and [4]). Note that we're not talking about large effects here, particularly if m is small. This is really an exercise in precision, or in being overly fastidious,

2.4. **The effect of variation in P.** The extra component in σ^2 introduced by variation in P is given the symbol Δ_m, so that equation (2) becomes:

$$\sigma^2_{t+1} = (1-m)^2\,[\frac{P(1-P)}{2N_e} + (1-\frac{1}{2N_e})\sigma^2_t] + \Delta_m.$$

After some algebra, an expression can be obtained for Δ_m. When substituted back into the above equation it gives:

$$\sigma^2_{t+1} = \frac{P(1-P)}{2N_e} + [(1-\frac{1}{2N_e})(1-m)^2 - \frac{m(1-m)}{2N}]\sigma^2_t \qquad (8)$$

The $(1-m)^2$ multiplier has disappeared from the first term, and there is an extra component to the σ^2 term.

2.5. **The effect of variation in m.** Rather than m being a constant, it is now assumed to be variable with mean m and variance $V(m)$. This changes equation (8) into

$$\sigma^2_{t+1} = \frac{P(1-P)}{2N_e} + [(1-\frac{1}{2N_e})(1-m)^2 + V(m) - \frac{m(1-m)}{2N}]\sigma^2_t \quad (9)$$

Under model [3], in which each gene has an independent probability m of being an immigrant gene,

$$V(m) = \frac{m(1-m)}{2N}$$

The last two terms in equation (9) cancel out, giving

$$\sigma^2_{t+1} = \frac{P(1-P)}{2N_e} + [(1 - \frac{1}{2N_e})(1-m)^2]\sigma^2_t \tag{10}$$

Equation (10) is seen to be identical to equation (2) except for the missing $(1-m)^2$ factor in the first term. The equilibrium in this case is

$$\hat{\sigma^2} = \frac{P(1-P)}{1 + (2N_e - 1)(2m - m^2)} \tag{11}$$

again just lacking the $(1-m)^2$ term.

Under model [4], the more realistic in terms of adult migration,

$$V(m) = \frac{m(1-m)}{N}$$

leading to a somewhat more complex equation.

2.6. **The inbreeding approach - sampling with or without replacement.**

Equation (5) gave the same result as equation (2). This seems contradictory. As argued above, equation (2) essentially assumed deterministic migration (model [1]), whereas the arguments in connection with equation (5) essentially assume that each gene is considered independently (model [3]).

The resolution of this quandary appears to lie in the definition of what pair of genes are considered for IBD. Wright's original definition of inbreeding was in terms of identity of uniting gametes, or equivalently the two genes contained by an individual. In the Wright-Fisher model generation model, in which each gene is a randomly selected gene from the previous generation, there is no distinction between the two genes possessed by an individual, and any two genes sampled from the population. However this implies sampling *without* replacement.

What Barrie Latter and I argued is that the appropriate definition ought to look at IBD of two genes selected from the population *with* replacement. Marc Feldman and I had made the same argument previously for sampling of pairs of gametes in relation to linkage disequilibrium [4]. This at first sight may seem an artificial definition. There is a $1/2N$ chance that the same gene will be selected twice, in which case it must be IBD with itself regardless of any migration or mutation process. How could this be relevant?

The reason that this is relevant, we argued, is that the quantities of interest in the discussion of migration are measures of homozygosity in terms of parameters such as P^2. Although the arguments are applied to finite-size populations, nevertheless the calculation of such homozygosity parameters treats the gene frequencies in the same way as if the populations were infinite. The true level of homozygosity in a small population ought to be a quantity such as:

$$\frac{n_A}{2N} \cdot \frac{n_A - 1}{2N - 1},$$

where n_A is the number of A genes and $2N$ is the total number of genes. Sampling without replacement would be appropriate if homozygosity was calculated in such a way. For conventional homozygosity calculations, however, sampling with replacement is appropriate.

2.7. N_e versus N.

Notice that I have subtly changed the population size from N_e to N. The question of which is appropriate is quite tricky.

All the calculations in this chapter are based on the Wright-Fisher model. I don't think anyone really believes in this model in its pure form. As I understand it, the effective population size N_e is a kind of fudge factor to take account of mating schemes that don't strictly adhere to the Wright-Fisher model. So chance fluctuations in gene frequencies instead of having variance $pq/2N$ can have variance $pq/2N_e$.

But what happens in inbreeding arguments, where one requires probabilities in terms of N or $2N$? Is it legitimate to just substitute N_e for N? Many do this, and I have followed Crow & Kimura [1] in writing equation (4). But when it comes to arguments to do with sampling with and without replacement, I find it difficult to know what to do if the assumptions of the Wright-Fisher model are violated.

2.8. Formulae calculated assuming sampling with replacement.

Consider first model [3], in which each gene has probability m of being a migrant gene. What is the probability of IBD in this case? I will calculate the probabilty of f_{t+1} in terms of f_t, where this f is now assumes sampling with replacement rather than the previous definitions in (4) - (6) which assumed sampling without replacement.

In calculating f_{t+1}, there are two possibilities: (1) The same gene is chosen twice, with probability $1/2N$. In this case IBD is certain regardless of migration. (2) With probability $1 - 1/2N$ two different genes

are chosen. If the two genes are non-immigrant genes, with probability $(1-m^2$, they will be IBD with the relevant IBD probability from the previous generation, f_t. Otherwise, if either is a migrant gene, they won't be IBD. Overall, therefore,

$$f_{t+1} = \frac{1}{2N} + (1 - \frac{1}{2N})(1-m)^2 f_t \tag{12}$$

This equation is now identical with equation (10) if the substitution $\sigma^2 = fP(1-P)$ is made.

It is also possible to use the probability approach to derive the equivalent to equation (2) which assumed deterministic migration. It is only necessary to consider sampling two genes from an infinite gene pool made up of a fraction $1-m$ of non-immigrant genes and m of immigrant genes. This gives

$$f_{t+1} = \frac{1}{2N}(1-m)^2 + (1 - \frac{1}{2N})(1-m)^2 f_t. \tag{13}$$

In general it seems that all results derived using variance - frequency methods can be derived more easily using inbreeding - probability methods,

2.9. **Finite number of islands.**

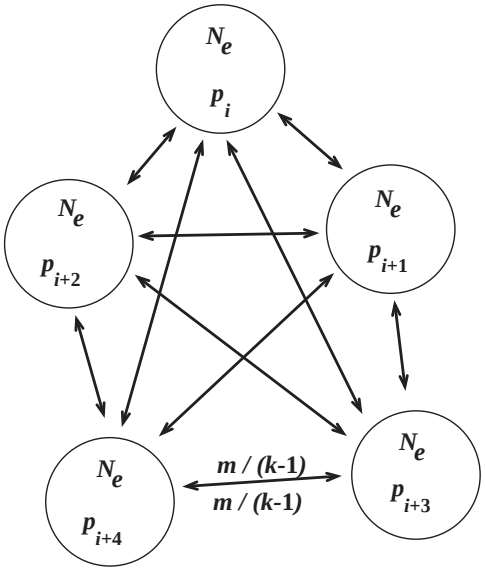

FIGURE 3. Model of k island populations exchanging migration at an equal rate

The second paper in the series, [3], looked at the model where islands exchange migrants at the same rate. The total migration rate into a population is m, implying a rate between populations of $m/(k-1)$. The calculation of frequencies within islands becomes equivalent to the island - mainland model where the number of islands is infinite.

This model allows additional information of the covariance in frequencies between populations. All calculations in this case were made using the probability model. Stochastic and deterministic migration were compared, with similar consequences to the single island model.

Some of this paper was devoted to an examination of distance measures. A measure due to Nei that has achieved high levels of use is essentially a measure that measures mutational distance. It has always seemed strange to me that this measure is used in studies where the difference in frequencies is clearly due to genetic drift rather than mutation pressure. Barrie Latter claims, as in previous publications eg [2], that a complete description of genetic distance does require a parameter such as this together with a 'kinship' parameter that is influenced mainly by drift,

3. A MUTATION MODEL

Migration and mutation are both linear forces, and have quite similar effects. We [5] looked at one particular mutation model, the infinite allele model. In this model, each mutation produces a novel allele. The quantity of interest in describing variability under this model is

$$S = \sum_{i=1}^{k} p_i^2, \tag{14}$$

where p_i is the frequency of allele i. The actual measure of variability is 1 - S.

With the infinite allele model, there is a direct relationship between homozygosity as measured by the frequency parameter S, and autozygosity as measured by the probability parameter f. Frequency and probability arguments should therefore lead to the same result.

3.1. **The homozygosity calculation.**

I won't go into the details of this calculation, which by our standards was quite a convoluted one. It utilised the binomial distribution to calculate the probability distribution of existing alleles after mutation.

It then added a component for the newly created mutations. The end result of the calculation was the recurrence relationship:

$$S_{t+1} = \frac{1}{2N} + (1-u)^2(1-\frac{1}{2N})S_t \tag{15}$$

At equilibrium:

$$\hat{S} = \frac{1}{(2N-1)(2u-u^2)} \tag{16}$$

3.2. **The autozygosity calculation.**
The conventional relationship, as given by Crow & Kimura [1] 7.2.1, is

$$f_{t+1} = (1-u)^2\frac{1}{2N} + (1-u)^2(1-\frac{1}{2N})f_t \tag{17}$$

As in the case of migration, there is a $(1-u)^2$ factor in the first term of the relationship which is not present in the frequency calculation. By analogy with the migration calculation, equation (17) would be applicable to a model with a fixed number of mutations, rather than the more realistic model in which mutation occurs independently for each gene.

As with the migration calculation, the probability of autozygosity can be calculated assuming sampling *with* replacement. The justification for using sampling with replacement is clearer in this case, since the quantity S is clearly based on the assumption of an infinitely large population with the frequencies of the finite population. The recurrence relationship for sampling with replacement is

$$f_{t+1} = \frac{1}{2N} + (1-u)^2(1-\frac{1}{2N})f_t \tag{18}$$

At equilibrium:

$$\hat{f} = \frac{1}{(2N-1)(2u-u^2)} \tag{19}$$

These are consistent with the frequency calculations given by equation (15) and (16).

The difference between the equilibrium calculations with and without replacement is the factor $(1-u)^2$. Even more than with the migration example, the difference can scarcely be of practical importance.

References

[1] J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Harper & Row, New York, 1970.

[2] B D Latter. The island model of population differentiation: a general solution. *Genetics*, 73(1):147–157, 1973 Jan.

[3] B. D. Latter and J. A. Sved. Migration and mutation in stochastic models of gene frequency change. ii. stochastic migration with a finite number of islands. *Journal of Mathematical Biology*, 13:95–104, 1981.

[4] J. A. Sved and M. W. Feldman. Correlation and probability methods for one and two loci. *Theor Popul Biol*, 4:129–132, 1973.

[5] J. A. Sved and B. D. Latter. Migration and mutation in stochastic models of gene frequency change. i. the island model. *J Math Biol*, 5(1), 1977.