Genet. Sel. Evol. 32 (2000) 395–402

Original article

The hierarchical island model revisited

Yves VIGOUROUX[a,*], Denis COUVET[b]

[a] Unité de malherbologie et d'agronomie, Institut national de la recherche agronomique, BV 1540, 17 rue Sully, 21000 Dijon, France
[b] Muséum national d'histoire naturelle, CRBPO, 55 rue Buffon, 75005 Paris, France

(Received 7 October 1999; accepted 25 April 2000)

Abstract – Formulae were derived for the genetic differentiation between populations within a metapopulation (F_{SM}), and between metapopulations (F_{MT}) as functions of migration and mutation rates, size and number of populations and metapopulations. We show that $F_{MT} = 1/(1 + 4N_e m)$, where N_e is the effective size of a metapopulation, and where the migration rate between metapopulations is m. The formulae for F_{MT} and F_{SM} were more general than previously proposed since we have relaxed some previously made hypotheses and we included the effect of the mutation rate. Using our formula, some unexpected result of estimation of gene flow, previously obtained, can be explained readily

subdivided populations / effective size / gene flow / mutation rate / *F* statistics

Résumé – Le modèle hiérarchisé en île revisité. Les formules de la différentiation génétique entre populations d'une même métapopulation (F_{SM}), et entre les métapopulations sont calculées en fonction du taux de migration, de mutation, la taille et le nombre de populations et de métapopulations. Nous montrons que $F_{MT} = 1/(1+4N_e m)$, où N_e est l'effectif efficace d'une métapopulation, et où le taux de migration entre métapopulations est m. Les formules proposées sont plus générales que celles proposées précédemment car nous considérons des hypothèses plus générales et notamment nous avons inclus l'effet du taux de mutation. En utilisant nos formules, nous montrons que des résultats inattendus utilisant ce modèle hiérarchique, peuvent être expliqués.

populations subdivisées / taille efficace / flux de gène / taux de mutation / *F* statistiques

1. INTRODUCTION

The analysis of the distribution of genetic diversity allows us to understand the spread and evolution of different species. One important parameter is the amount of gene flow between populations. The statistic F_{ST} was constructed to obtain this parameter from genetic data. Wright [9] has shown that in an island

* Correspondence and reprints
E-mail: yves.vigouroux@epoisses inra.fr

model, the expected value of F_{ST} is $1/(1+4Nm)$, where Nm is the number of migrants per generation, with N representing the size of a population and m, the migration rate between populations. However, gene flow can occur at different levels: for example within and between landscapes (*i.e.* the habitat of a metapopulation). The question which remains, is how does gene flow, at these different levels, relate to corresponding genetic differentiation? To answer this question, Slatkin and Voelm [5] developed the "hierarchical island model", where there are two kinds of migration, (i) between neighbouring populations, *i.e.* populations with in the same metapopulation, and (ii) between distant populations, *i.e.* populations from different metapopulations. However using a coalescent approach, Slatkin and Voelm [5] derived approximate values for F-statistics, where the influence of mutation rate is not considered.

Using an "inbreeding approach", we report the influence of mutation rate on F-statistics at different levels. In addition, a further aim was to relate the differentiation between metapopulations and the effective size of a metapopulation as defined at the mutation-drift equilibrium [6,7]

2. THE MODEL

This hierarchical island model was composed of n metapopulations and each metapopulation was comprised of d populations of N individuals. Migration occurred at two levels: between populations inside a metapopulation and between populations of different metapopulations. Each individual of the population was assumed to be monoecious diploid. There were discrete non-overlapping generations, and panmixia. Migration occurred at the gamete stage, and each gamete had a probability, m_1, of originating from an other population of the same metapopulation and, m_2, from another metapopulation ($m_2 < m_1$).

Our model to study allelic diversity was a k-allele model where k represents the number of possible allelic states and u, the total rate of mutation.

In this hierarchical island model, three different F-statistics were defined, which correspond to comparisons at different levels of the hierarchy: F_{ST} represents the correlation between two randomly chosen alleles in a population relative to alleles chosen in different metapopulations; F_{SM}, the correlation between two randomly chosen alleles in a population, relative to the alleles chosen in different populations of the same metapopulation, and F_{MT}, the correlation between two randomly chosen alleles in different populations of a metapopulation, relative to the alleles chosen in different metapopulations. These three coefficients can be written in terms of probability of identities by descent, following Cockerham and Weir [3]:

$$F_{ST} = \frac{f_p - f_t}{1 - f_t} \tag{1}$$

$$F_{SM} = \frac{f_p - f_m}{1 - f_m} \tag{2}$$

$$F_{MT} = \frac{f_m - f_t}{1 - f_t} \tag{3}$$

with the three probabilities of identity by descent f_p, f_m and f_t defined as follows:

- f_p, probability of identity of two alleles drawn at random from the same population;
- f_m probability of identity of two alleles drawn at random from the same metapopulation in two different populations;
- f_t probability of identity of two alleles drawn at random from two different metapopulations.

3. RESULTS

3.1. Analytical resolution

To find the expected value at equilibrium of the three probabilities of identity by descent f_p, f_m and f_t, we derived a recurrent relationship following Crow and Aoki [4]. At time t, f'_p, f'_m and f'_t are related to f_p, f_m and f_t, at time $t-1$:

$$f'_p = v[\alpha(a+bf_p)+\beta f_m+\gamma f_t]+w[b\alpha(1-f_p)+\beta(1-f_m)+\gamma(1-f_t)] \quad (4)$$

$$f'_m = v[t_0\beta(a+bf_p)+(\alpha+\beta-t_0\beta)f_m+\gamma f_t] + w[bt_0\beta(1-f_p)+(\alpha+\beta-t_0\beta)(1-f_m)+\gamma(1-f_t)] \quad (5)$$

$$f'_t = v[s\gamma t_1(a+bf_p)+s\gamma(1-t_1)f_m+(1-s\gamma)f_t] + w[bs\gamma t_1(1-f_p)+s\gamma(1-t_1)(1-f_m)+(1-f_t)(1-s\gamma)] \quad (6)$$

with

$$\alpha=(1-m_1-m_2)^2+\frac{m_1^2}{d-1}+\frac{m_2^2}{d(n-1)},$$

$$\beta=2m_1(1-m_1-m_2)+\frac{m_1^2(d-2)}{d-1}+\frac{m_2^2(d-1)}{d(n-1)},$$

$$\gamma=2m_2(1-m_2)+\frac{m_2^2(n-2)}{n-1},$$

$$a=\frac{1}{2N},\quad b=1-a,\quad v=(1-u)^2,\quad w=\frac{2u(1-u)}{k-1},$$

$$t_0=\frac{1}{d-1},\quad t_1=\frac{1}{d}\quad\text{and}\quad s=\frac{1}{n-1}\cdot$$

The following resolution was obtained for the equilibrium value, when $f_p=f'_p$, $f_m=f'_m$ and $f_t=f'_t$.

Following Cockerham and Weir [3], taking away equation (4) from equation (5) leads to the elimination of terms which are functions of f_t. The equation obtained is solved after isolation of the two quantities f_p-f_m and $1-f_m$

$$\frac{f_p-f_m}{1-f_m}=\frac{a(v-w)(\alpha-t_0\beta)}{1-b(v-w)(\alpha-t_0\beta)}\cdot \quad (7)$$

Equation (2) gives a relationship between f_p and f_m: $f_p = F_{SM}(1 - f_m) + f_m$. Using this formula, f_p is substituted in the equations (5) and (6). Then, taking away equation (5) from equation (6), and after the isolation of the two quantities $f_t - f_m$ and $1 - f_t$, F_{MT} is deduced:

$$\frac{f_m - f_t}{1 - f_t} = \frac{(v-w)(t_0\beta - st_1\gamma)(a + bF_{SM})}{1 + (v-w)\left[(t_0\beta - st_1\gamma)(a + bF_{SM}) - (1 - \gamma - s\gamma)\right]}. \quad (8)$$

3.2. Approximated value for F_{SM}, F_{MT} and F_{ST}

Assuming that the two different migration rates were sufficiently small so that, at most, one gene migrates per generation, the approximated values of α, β and γ are $\alpha \approx 1 - 2m_1 - 2m_2$, $\beta \approx 2m_1$, $\gamma \approx 2m_2$. These approximations were used further in the formulae given for different F-statistics parameters.

Assuming $u \ll 1$, $m_1 \ll 1$ and $m_2 \ll 1$, F_{SM} is equal to:

$$F_{SM} \approx \frac{1}{1 + 4N\left(m_1\frac{d}{d-1} + m_2 + u\frac{k}{k-1}\right)}. \quad (9)$$

Assuming $u \ll 1$, $m_1 \ll 1$ and $m_2 \ll 1$ and d large, F_{MT} is equal to:

$$F_{MT} \approx \frac{1}{1 + 4N_e\left(m_2\frac{n}{n-1} + u\frac{k}{k-1}\right)} \quad (10)$$

with

$$N_e = Nd\left(1 + \frac{1}{4Nm_1}\left(1 + 4Nm_2 + 4Nu\frac{k}{k-1}\right)\right). \quad (11)$$

F_{ST} is related to F_{SM} and F_{MT} using equations (1), (2) and (3):

$$(1 - F_{ST}) = (1 - F_{SM})(1 - F_{MT});$$

so assuming the same hypotheses made previously,

$$F_{ST} \approx \frac{1}{1 + \frac{AB}{1 + A + B}} \quad (12)$$

$$A = 4N\left(m_1\frac{d}{d-1} + m_2 + u\frac{k}{k-1}\right)$$

$$B = 4N_e\left(m_2\frac{n}{n-1} + u\frac{k}{k-1}\right).$$

(See appendix for details of the resolution.)

4. DISCUSSION

The terms migration or migration rate are used for parameters m_1 and m_2, and gene flow for the product Nm_i.

4.1. Comparison of our formula with previous results

Considering only one metapopulation composed of d populations, Cockerham and Weir [3] found that, $F_{ST} = \dfrac{1}{1+4N\left(m\dfrac{d}{d-1}+u\right)}$.

We can compare their formula with our formula for F_{SM} (9). In the case of the infinite allele model *i.e.* $k/(k-1) \approx 1$ and $m_2 =$ zero, the two formulae agree.

For F_{MT}, our formula (10) differed from the one previously proposed by Slatkin and Voelm [5]. In their paper, the term N_e is equal to Nd, due to their approximation, $Nm_1 \gg 1$. If we assume the same approximation, the two formulae agree.

This parameter N_e (11) can be interpreted as the effective number of individuals in a metapopulation. Takahata [6], considering only the equilibrium mutation-drift at the level of the metapopulation (not considering a hierarchical island model) gave a value for this parameter:

$$N_e = N + \left(Nm + \frac{d-1}{4d}\right)\frac{d-1}{e+m}$$

where e is the extinction rate of the population.

The two formulae were almost identical for $m_2 \ll m_1$ and no extinction (*i.e.* $e = 0$). Our formula was also in agreement with the more recent result $N_e = Nd/(1-F_{ST})$ [7]. In our formula, N_e appears proportional to $1/4Nm_1$ and $\left(1+4Nm_2+4Nuk/(k-1)\right)$, and that can be explained as follows. Effective size of a metapopulation is increased by a function of $1/4Nm_1$, which represents the fact that drift is disconnected among populations of a same metapopulation. Input of genetic variability in each population, independently of the other populations of the same metapopulation will further increase this effect. Such input is a function of u and m_2, which both bring variability unrelated to variability found in the other populations of the same metapopulation.

Due to differentiation within a metapopulation, effective size of a metapopulation can be larger than the number of individuals. Although that fact has already been noticed, and also questioned [7], our formula shows that migration between metapopulations might reinforce that effect.

4.2. Influence of mutation rates at different scales

The influence of the mutation rate depends on its value relative to the value of the migration rate at the scale considered (see equations 9 and 10). Intuitively, an obvious result is that genetic differentiation decreases as the mutation rate increases. Then the consequences of F-statistics are that gene

flow will be over-estimated whenever genetic markers with a high mutation rate are used. Such a theoretical result can account for the negative relationship found by Bowcock *et al.* [2] between F_{ST} and locus heterozygosity, assuming that a higher heterozygosity indicates a higher mutation rate. Such a negative relationship will hold if and only if the mutation rate is not negligible relative to the migration rate. Thus, this explanation is not contradictory to simulations showing that median value of F_{ST} is nearly independent of heterozygosity in that case [1].

4.3. Application of the model

The formula we propose gave some insight into the unexpected result found using the hierarchical island model to obtain an estimation of gene flow. Wolf and Soltis [8] studied structuration in the *I. aggregata* complex at three levels among populations, subspecies and species. To simplify, we will only discuss the pattern they found between subspecies and species. As they have a direct estimation of the number of subspecies and species, they give, using the Slatkin and Voelm formula [5], an estimation of gene flow $Nm_{\text{subspecies}}$ among subspecies and gene flow Nm_{species} among species. They found an unexpected result: gene flow among species is two to three times higher than that among subspecies. To explain this phenomenon, they proposed the occurrence of hybridisation between species [8].

Using our formula, the estimation of gene flow among species was five to eight times lower than among subspecies. Indeed, since gene flow at the subspecies level is low, on average for all taxa $Nm_{\text{subspecies}} = 0.017$ [8], then the gene flow estimated Nm_{species} (using the Slatkin et Voelm formula) was in fact $Nm_{\text{species}}(1 + 1/4Nm_{\text{subspecies}})$ (see equations 10 and 11). This factor $(1+1/4Nm_{\text{subspecies}})$ inflates the estimation of gene flow at the species level by a factor of fifteen. If we correct the gene flow estimated by Wolf and Soltis [8] by this factor, we find a lower migration rate at the species level compared to the subspecies one.

ACKNOWLEDGEMENTS

Thanks to Valérie LeCorre, Marie-Joséphine Farmer and two anonymous reviewers who improved the previous draft of this article. Yves Vigouroux is supported by a grant from the Burgundy Region Council.

REFERENCES

[1] Beaumont M.A., Nichols R.A., Evaluating loci for use in the genetic analysis of population structure, Proc R Soc. Lond. B 263 (1996) 1619–1626

[2] Bowcock A.M., Ruiz-Linares A., Tomfohrde J , Mınch E., Kidd J.R., Cavalli-Sforza L L., High resolution of human evolutionary trees wıth polymorphic mıcrosatellites, Nature 368 (1994) 455–457.

[3] Cockerham C.C., Weir B.S., Correlations, descent measures: drift with migration and mutation, Proc. Natl. Acad. Sci., USA 84 (1987) 8512–8514.

[4] Crow J.F., Aoki K., Group selection for a polygenic behaviour trait· estimating the degree of population subdivision, Proc Natl. Acad. Sci., USA 81 (1984) 6073–6077.
[5] Slatkin M., Voelm L., Fst in a hierarchical island model, Genetics 127 (1991) 627–629.
[6] Takahata N., Repeated failures that lead to the eventual success in human evolution, Mol. Biol. Evol. 11 (1994) 803–805.
[7] Whitlock M.C., Barton N.H., The effective size of a subdivided population, Genetics 146 (1997) 427–441.
[8] Wolf P.G., Soltis P.S., Estimates of gene flow among populations, geographic races, and species in the *Ipomopsis aggregata* complex, Genetics 130 (1991) 639–647.
[9] Wright S., The genetical structure of populations, Galton lecture at University college, London, 1950, pp. 323–354.

APPENDIX

Resolution of F_{SM}

Taking away equation (4) from equation (5) leads to

$$\begin{aligned} f_p - f_m &= v\left[(\alpha - t_0\beta)(a + bf_m) - (\alpha - t_0\beta)f_m\right] \\ &\quad + w\left[b(\alpha - t_0\beta)(1 - f_p) - (\alpha - t_0\beta)(1 - f_m)\right] \\ \text{then } f_p - f_m &= b(f_p - f_m)(\alpha - t_0\beta)(v - w) + a(1 - f_m)(\alpha - t_0\beta)(v - w) \\ \text{then } \frac{f_p - f_m}{1 - f_m} &= \frac{a(v - w)(\alpha - t_0\beta)}{1 - b(v - w)(\alpha - t_0\beta)}. \end{aligned}$$

Assuming approximations described in the text,

$$\begin{aligned} \frac{f_p - f_m}{1 - f_m} &\approx \frac{\left(1 - 2u\frac{k}{k-1}\right)\left(1 - 2m_1\frac{d}{d-1} - 2m_2\right)}{2N - (2N-1)\left(1 - 2u\frac{k}{k-1}\right)\left(1 - 2m_1\frac{d}{d-1} - 2m_2\right)} \\ \frac{f_p - f_m}{1 - f_m} &\approx \frac{1}{1 + 4Nu\frac{k}{k-1} + 4Nm_1\frac{d}{d-1} + 4Nm_2}. \end{aligned} \tag{A.1}$$

Resolution of F_{MT}

f_p is substituted in the equations (5) and (6) using $f_p = F_{SM}(1 - f_m) + f_m$.

$$\begin{aligned} f_m &= t_0\beta(v - w)(1 - f_m)(a + bF_{SM}) + \gamma(v - w)f_t + (1 - \gamma)(v - w)f_m + w \\ f_t &= st_1\gamma(v - w)(1 - f_m)(a + bF_{SM}) + (1 - s\gamma)(v - w)f_t + s\gamma(v - w)f_m + w \end{aligned}$$

then

$$\begin{aligned} f_m - f_t = (t_0\beta - st_1\gamma)(v - w)\big((1 - f_t) - (f_m - f_t)\big)(a + bF_{SM}) + \\ (1 - \gamma - s\gamma)(v - w)(f_m - f_t) \end{aligned}$$

then

$$\frac{f_m - f_t}{1 - f_t} = \frac{(v-w)(t_0\beta - st_1\gamma)(a + bF_{SM})}{1 + (v-w)\left[(t_0\beta - st_1\gamma)(a + bF_{SM}) - (1 - \gamma - s\gamma)\right]}$$

then

$$\frac{f_m - f_t}{1 - f_t} = \frac{at_0\beta(v-w)\left(1 - \frac{st_1\gamma}{t_0\beta}\right)\left(1 + \frac{b}{a}F_{SM}\right)}{1 + (v-w)\left[at_0\beta\left(1 - \frac{st_1\gamma}{t_0\beta}\right)\left(1 + \frac{b}{a}F_{SM}\right) + (\gamma + s\gamma - 1)\right]}$$

then

$$\frac{f_m - f_t}{1 - f_t} = \frac{(v-w)\left(1 - \frac{st_1\gamma}{t_0\beta}\right)}{\frac{1}{at_0\beta\left(1 + \frac{b}{a}F_{SM}\right)}[1 + (v-w)(\gamma + s\gamma - 1)] + (v-w)\left(1 - \frac{st_1\gamma}{t_0\beta}\right)}. \tag{A.2}$$

Using approximations described in the text,

$$(v-w)\left(1 - \frac{st_1\gamma}{t_0\beta}\right) \approx \left(1 - 2u\frac{k}{k-1}\right)\left(1 - \frac{(d-1)m_2}{d(n-1)m_1}\right) \tag{A.3}$$

$$[1 + (v-w)(\gamma + s\gamma - 1)] \approx \left(2u\frac{k}{k-1} + 2m_2\frac{n}{n-1}\right) \tag{A.4}$$

and, from (A.1)

$$\frac{1}{at_0\beta\left(1 + \frac{b}{a}F_{SM}\right)} \approx \frac{2N(d-1)}{2m_1}\left(\frac{1}{2N} + 2u\frac{k}{k-1} + 2m_1\frac{d}{d-1} + 2m_2\right). \tag{A.5}$$

So, assuming d large, from (A.2), using (A.3, A.4 and A.5)

$$\frac{f_m - f_p}{1 - f_t} \approx \frac{1}{1 + 4N_e\left(u\frac{k}{k-1} + m_2\frac{n}{n-1}\right)}$$

with $N_e = Nd\left(1 + \frac{1}{4Nm_1}\left(1 + 4Nm_2 + 4Nu\frac{k}{k-1}\right)\right)$.