Estimation of migration rates and population sizes in geographically structured populations

Peter Beerli
Department of Genetics, University of Washington, Seattle WA 98195-7360, USA

The estimation of population parameters from genetic data can help reveal past migration patterns or past population sizes. The transformation from raw genetic data to population parameters needs a model, which should reflect the true relationships between subpopulations. Often the models are overly simplified and do not allow, for example, for differences in population sizes and differences in migration rates. I stress here the point that it is important to consider possible asymmetries in migration rates and differences in population sizes. Very recently several estimators based on the direct use of allele frequencies and based on coalescence theory have been developed. All these outperform migration rate estimators based on F_{ST}.

1. Introduction

The estimation of population parameters such as population size and migration rates between subpopulations of a species is crucial for many ecological studies. Two very different approaches to estimating population parameters are in use: (1) direct methods using direct observations or radio-telemetry data of migrating individuals, and (2) indirect methods using genetic data from samples of individuals in several subpopulations for the inference of migration rates. Direct methods can help to determine the migration pattern of individuals during the study, and can deliver information about very recent history. Under the assumption that the few tracked individuals are picked at random and that their movements are not artefacts of the study, these data can give interesting insights into the migration pattern of a specific population. Limits are also evident, however: small migration rates are undetectable and the accuracy of the parameter estimates is small when the study is based only on a few individuals. If the study is too short and not repeated we cannot know if the migration pattern we observed was accidental or is general. Current progress in establishing the relationship between individuals using DNA finger-printing may help to generate accurate information about very recent migration patterns. Using these methods, it is simple to increase the number of individuals studied and to make estimates that are less dependent on the length of the study, because one uses the shared genetic history of parents and offspring. In the future DNA finger-printing will certainly provide a valuable tool for the detection of very recent migration pattern between small populations Bossart & Prowell 1998).

I will concentrate on indirect methods that also are based on genetic history. They do not assume that we can find parent-offspring pairs, but instead use probabilistic models. The array of possible markers is large and can include allozymes, restriction length polymorphism, microsatellites, or protein or DNA sequences. These genetic data are then the basis for our inference of migration patterns. This chapter focus on models for discrete populations and their assumptions, interested readers may find more information about the influence of different data types on the analysis in Neigel (1997). For some of these methods we first estimate certain meta-quantities from which we then infer population parameters, such as population sizes and migration rates. These methods all have some advantages compared to the direct approach. One can investigate large sample sizes or many loci and therefore detect small amounts of migration. There is no need to track individuals over time: the estimates are averages over evolutionary time and reveal general rather than individual migration patterns. We can use the indirect methods for any organism. There are also disadvantages. We need to assume that the markers are selectively neutral, so that similarity between different subpopulations is a result of migration rather than similar selection pressures. The population parameters are also confounded with the mutation rate. The markers need to show enough variability, so that we can see differences between subpopulations. A marker with a very slow mutation rate will not reveal recent migration events, but may still have some information about migration events far in the past, when compared with geographically more distant populations.

Several groups of approaches using genetic data for the inference of migration rates are recognized: (1) estimators based on allele frequencies and Wright (1951)'s F-statistic (reviewed in Michalakis & Excoffier 1996); (2) maximum likelihood estimators based on allele frequencies (Rannala & Hartigan 1996; Tufto *et al.* 1996); and (3) estimators based on genealogies of the sampled individuals (coalescent theory: Kingman 1982b) with migration rates estimated using procedures of Wakeley (1998), Bahlo & Griffiths (1998), or Beerli & Felsenstein (1998). Some estimators are mixtures of these groups. Slatkin & Maddison (1989), for example, developed a method that uses results from coalescent theory and then presents an interpolation table produced by simulating the coalescence with migration, in which the minimal number of migration events found on the best genealogy is related to a migration rate, $4N_em$. Most of these estimators were developed under simplifying assumptions; for example all current "all-purpose" migration rate estimators assume that the population is in migration-mutation equilibrium; in other words they assume that the migration and the mutation rates are constant through time. Additionally, almost all methods use a constant sized population. There is currently one method, developed by Bahlo & Griffiths (1998), that can allow for subpopulations that are growing. The availability of methods for estimating population parameters under non-constant condition will increase, but development of programs and new method is often a slow task. Additionally, the more parameters we want to estimate the more data we need; this makes it perhaps impractical to allow for all possibilities. Methods allowing for non-constant conditions have first to be fully developed and then have to show that they can deliver accurate estimates.

Most migration models are based on the Wright-Fisher population. The Wright-Fisher population model with migration is rather simple (Fig. 1) and has properties that makes its mathematical treatment easy. A subpopulation consists of a constant number of individuals, either haploid or diploid (I will describe the diploid case). In each generation each individual is producing a large number of gametes. These gametes either stay in a subpopulation or migrate into another subpopulation. New individuals are formed by randomly choosing two gametes in a subpopulation, and these individuals replace their parents.

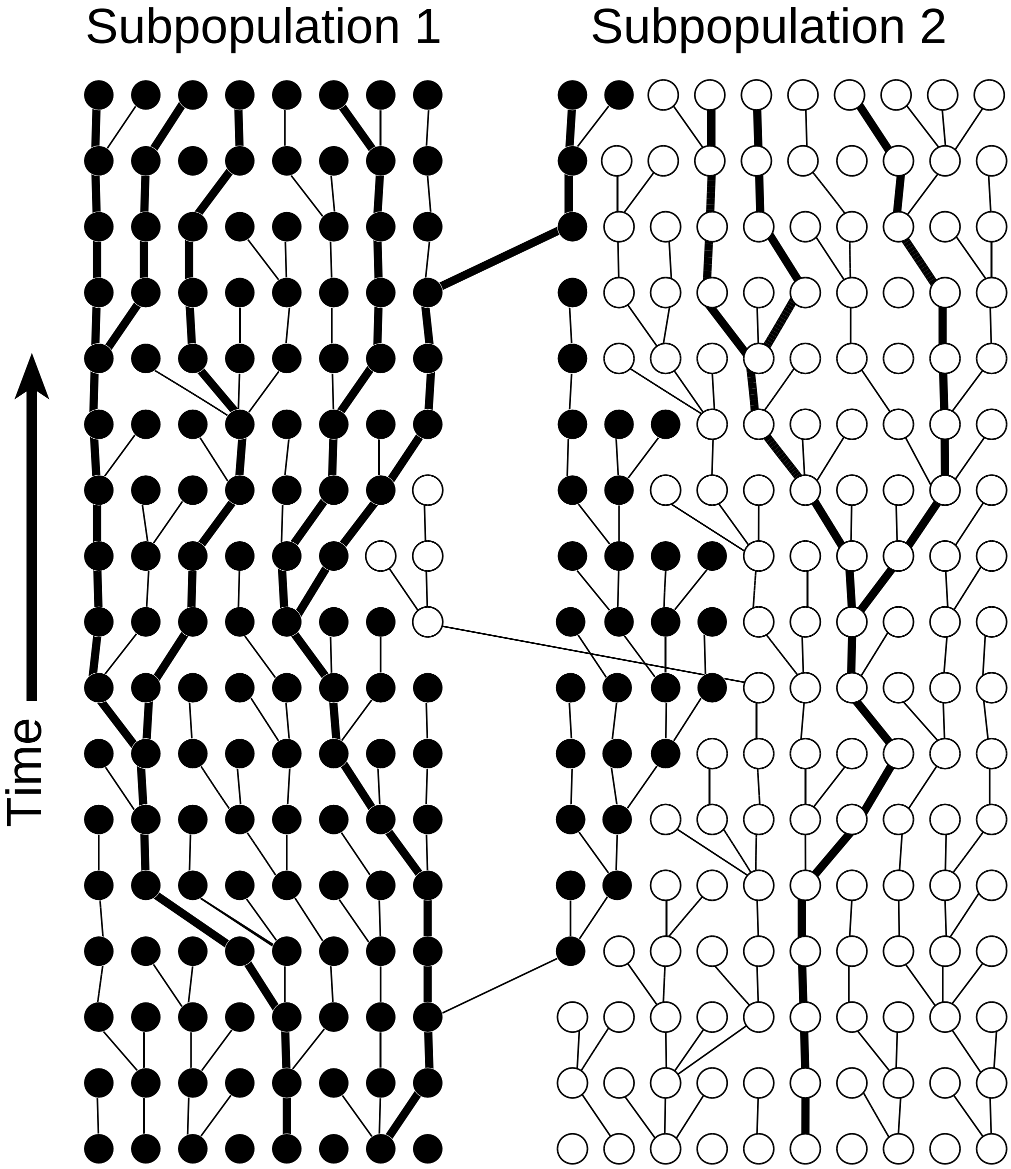

Fig. 1: Gene genealogy of a Wright-Fisher population with migration between two subpopulations. Thick lines show the coalescence of lineages from a sample of 4 individuals in each population.

I will focus on some of these estimators and discuss their properties and limitations when subpopulations may exchange migrants at different rates or have different sizes, and in which the mutation rate may vary among loci. Additionally, I will develop an F-statistic framework introduced by Maynard Smith (1970), Maruyama (1970), and Nei & Feldman (1972), so that population sizes and migration rates can be jointly estimated. Finally, I will compare these F_{ST}-based approaches with an approach based on the coalescence theory.

2. How to compare different population parameter estimators

For practitioners choosing a method for the estimation of population parameters is a difficult task, because many methods are available (see the almost certainly incomplete list of available programs in this book). Each of these different methods has its own set of assumptions; sometimes these are incompatible with the study. Results obtained with real data provide rather unreliable guidance for picking a method, because we normally do not know the underlying true parameter values. With computer generated data sets, however, it is easy to create many replicates with the same population model and the same population parameters. For these arbitrary data generating parameters I will use terms like the "true value" or the "truth". The chosen population parameters should then be recovered from each data set with some error from randomness in the data generation (Hudson 1983) and errors in the analysis. The averages of the parameter estimates from many data sets should converge to the true values when we increase the number of replicates and if the method is unbiased (does not, for example, always yield estimates that are too high). Additionally, a good method should have a small variance and therefore produce small confidence intervals. In short, superior methods are unbiased or have a small bias and a small variance.

3. Migration rate estimators based on F-statistics

Wright (1951) described a framework that uses his earlier inbreeding coefficient F for a subdivided population with three coefficients: F_{IS}, F_{IT}, and F_{ST}. For the infinite allele model F can be understood as a probability that two randomly chosen alleles are identical by descent, and the F_{IJ}'s are ratios between the F in an individual I, a subpopulation S, and all subpopulations T (Total).

I will focus on F_{ST}, which is the correlation between the probability that two randomly chosen gene copies within a subpopulation share an ancestor in the last generation relative to the probability that two gene copies picked from the total population share an ancestor in the last generation. In other words, this index uses the partitioning of total genetic variability into variability within and between subpopulations. Using insights from Slatkin (1991) and from Michalakis & Excoffier (1996), a highly generalized overview for the relationship of F_{ST} is

$$F_{ST} = \frac{g(\sigma_W) - g(\sigma_B)}{g(\sigma_B)} \qquad (1)$$

where F_{ST} can be replaced by different specific estimators such as θ (Weir 1996), N_{ST} (Lynch & Crease 1990), $<F_{ST}>$ (Hudson *et al.* 1992), Φ_{ST} (Excoffier & Smouse 1994), ρ_{ST} (Rousset 1996), G_{ST} (Nei 1973), and R_{ST} (Slatkin 1993). The g(x) are correction functions used for the different F_{ST}-estimators scaling the variance within a population (σ_W) or between populations (σ_B). These variances are proportional to the mean coalescence times in a subpopulation and the whole population. This summary statistic, F_{ST}, is interpreted as a measure of the differentiation between subpopulations, where values close to zero indicate that the population is not structured.

F_{ST} is commonly transformed into a more direct measure for migration. Wright (1951) showed that for an n-island population model (Fig. 2) with an infinite number of subpopulations and no mutation, we can use

$$F_{ST} = \frac{1}{1 + 4N_e m} \qquad (2)$$

where N_e is the population size of a Wright-Fisher population (Fig. 1), and m is the migration rate per generation. I use the term effective population size N_e to mark the fact that even when the population is not exactly behaving like a Wright-Fisher population, we still can use N_e to make comparisons with other populations. Relaxing the rather strong assumption of having an infinite number of subpopulations is simple and has been described several times, for example by Li (1976):

$$F_{ST} = \frac{1}{1 + 4N_e m \frac{d^2}{(d-1)^2}} \quad (3)$$

where d is the number of subpopulations, d is most certainly different from the number of sampled populations. Charlesworth (1998) pointed out that results for $N_e m$ can be very different depending on which version of F_{ST} is used.

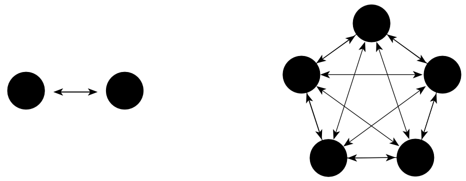

Fig. 2: Island population model. Examples with 2 and 5 islands. The relevant population parameters are an overall population size N_e and an immigration rate m that is the same for all subpopulations.

The n-island population model uses only two parameters: the effective population size N_e and the immigration rate per generation m. It is assumed that the subpopulation sizes are the same and that the migration rate is the same between all the subpopulations. These assumptions are often violated in studies of natural populations, for which we know neither the true migration patterns nor the population sizes.

I created simulated data sets using a technique first used by Hudson (1983). 100 different data containing sequence data (500 bp) for 20 individuals in each of 2 subpopulations were created using specific population sizes Θ_1 and Θ_2, and migration rates M_1 and M_2, where Θ is $4N_e$ μ μ (Fig. 3). From these data sets γ (γ=$4N_e m$=ΘM), was estimated using Wright's formula (2,3) with $<F_{ST}>$ (Hudson *et al.* 1992). For this simple two-population situation the estimates for γ are appropriate if the subpopulation sizes are the same and the migration rates are symmetric; as soon as the assumptions of symmetry of migration rates or of equal population sizes is violated, however, the estimates are wrong (Table 1). Relethford (1996) revealed similar problems with approaches that assume that the population sizes are equal and develops an alternative approach, that allows for different sizes but not different rates, so that $m_{ij} = m_{ji}$. Laurent Excoffier (personal communication 1998) and coworkers have done extensive simulations with different population sizes and have shown that the assumption of equal population sizes is critical to the analysis.

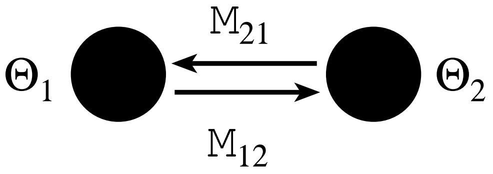

Fig. 3: Two population model: Θ_1 is $4N_e^{(1)}\mu$, Θ_2 is $4N_e^{(2)}\mu$, M_{21} is m_{21}/μ, and M_{12} is m_{12}/μ, where N_e is the effective population size, μ is the mutation rate per generation, and m_{ji} is the migration rate per generation from population j to population i.

Table 1: Estimates of migration rates γ = $4N_em$ based on Wright's formula (Formulas 2, 3) in a two population system (Fig. 3). Averages ± standard deviations of 100 simulated data sets generated with the same population parameters are shown. For each data set 2 populations with 20 sampled individuals with 500bp of sequence data were created according to Hudson (1983). T: the parameter values under which the data sets were created, A: Wright's relation of F_{ST} and migration rate γ without correction for finite and known number of subpopulations (Formula 2), B: Wright's relation with the correction for two populations (Formula 3).

	Population 1		Population 2	
	Θ	γ	Θ	γ
T	0.01	1.00	0.01	1.00
A	-	4.56 ± 3.08	-	4.56 ± 3.08
B	-	1.14 ± 0.77	-	1.14 ± 0.77
T	0.01	10.00	0.01	1.00
A	-	31.20 ± 88.78	-	31.20 ± 88.78
B	-	7.80 ± 22.20	-	7.80 ± 22.20
T	0.05	10.00	0.005	1.00
A	-	45.86 ± 74.15	-	45.86 ± 74.15
B	-	11.46 ± 18.54	-	11.46 ± 18.54

The incorporation of asymmetric migration rates in a two-population model seems simple using the framework developed by Maynard Smith (1970), Maruyama (1970), and Nei & Feldman (1972). Interestingly, the formulas outlined by Nei & Feldman (1972) could have been used to estimate the effective population size and the migration rate jointly using the F-statistic, but were to my knowledge never used in that context. Their work considered N_e, the mutation rate μ, and the migration rate m, but did not show how to translate these parameters into a more practical estimator, given that the mutation rate is usually unknown. With two populations (Fig. 3) we have three quantities: the probability that two randomly chosen gene copies in subpopulation 1 share the same ancestor in the past generation ($F_W^{(1)}$), a similar probability for subpopulation 2 ($F_W^{(2)}$), and the probability that two copies from different subpopulations have the same ancestor in the past generation (F_B). These statistics can be simply estimated from data using heterozygosity in the subpopulation and an overall heterozygosity; Slatkin & Hudson (1991) outlined a procedure for sequence data. By relating F_B and $F_W^{(i)}$ to population sizes, migration rate and mutation rate, we can use the following recurrence formulas, which are adapted from Nei & Feldman (1972), for two populations with different population sizes and migration rates. The exact formulas are simplified by removing quadratic terms like μ^2, m^2, μm and divisions by number of individuals in a population (e.g. m/N). For two populations in equilibrium we get the equation system

$$\begin{aligned} F_W^{(1)} &= \frac{1}{2N_1} + \left(1 - 2\mu - 2m_1 - \frac{1}{2N_1}\right) F_W^{(1)} + 2m_1 F_B\,, \\ F_W^{(2)} &= \frac{1}{2N_2} + \left(1 - 2\mu - 2m_2 - \frac{1}{2N_2}\right) F_W^{(2)} + 2m_2 F_B\,, \\ F_B &= F_B\left(1 - 2\mu - m_1 - m_2\right) + m_1 F_W^{(1)} + m_2 F_W^{(2)} \end{aligned} \tag{4}$$

where μ is the mutation rate, m_i is the immigration rate into population i, and N_i is the subpopulation size. Because we do not know the mutation rate μ, I follow a common practice in coalescence theory and use a compound parameter Θ which is $4N_e\mu$, and define $M = m/\mu$. Multiplying the equation system by $1/(2\mu)$ we get

$$\begin{aligned}
F_W^{(1)} &= \frac{1}{\Theta_1} - \left(M_1 + \frac{1}{\Theta_1} \right) F_W^{(1)} + M_1 F_B \, , \\
F_W^{(2)} &= \frac{1}{\Theta_2} - \left(M_2 + \frac{1}{\Theta_2} \right) F_W^{(2)} + M_2 F_B \, , \\
2F_B &= F_B \left(- M_1 - M_2 \right) + M_1 F_W^{(1)} + M_2 F_W^{(2)} .
\end{aligned} \tag{5}$$

This system can be solved only for three quantities and not for the four quantities we would need to describe the two-population system (Fig. 3). We must either require the population sizes to be the same, but allow different migration rates, or require the migration rates to be the same, but allow different population sizes. For a model with $\Theta = \Theta_1 = \Theta_2$ and two variable migration rates M_1 and M_2 we get

$$\begin{gathered}
\Theta = \frac{2 - F_W^{(1)} - F_W^{(2)}}{2F_B + F_W^{(1)} + F_W^{(2)}} \, , \\
M_1 = \frac{2F_B F_W^{(1)} + F_W^{(1)} - F_W^{(2)} - 2F_B}{\left(F_W^{(1)} - F_B\right)\left(F_W^{(1)} + F_W^{(2)} - 2\right)} \, , \qquad M_2 = \frac{2F_B F_W^{(2)} + F_W^{(2)} - F_W^{(1)} - 2F_B}{\left(F_W^{(2)} - F_B\right)\left(F_W^{(1)} + F_W^{(2)} - 2\right)} \, ,
\end{gathered} \tag{6}$$

and for a model with two variable Θ_1 and Θ_2, and $M = M_1 = M_2$ we get

$$\begin{gathered}
\Theta_1 = \frac{\left(1 - F_W^{(1)}\right)\left(F_W^{(1)} + F_W^{(2)} - 2F_B\right)}{\left(F_W^{(1)}\right)^2 + F_W^{(1)} F_W^{(2)} - \left(2F_B\right)^2} \, , \qquad \Theta_2 = \frac{\left(1 - F_W^{(2)}\right)\left(F_W^{(1)} + F_W^{(2)} - 2F_B\right)}{\left(F_W^{(2)}\right)^2 + F_W^{(1)} F_W^{(2)} - \left(2F_B\right)^2} \, , \\
M = \frac{2F_B}{F_W^{(1)} + F_W^{(2)} - 2F_B} .
\end{gathered} \tag{7}$$

These estimators will fail when $F_W^{(i)} \leq F_B$. This can happen more often with more subpopulations and is dependent on the asymmetry of migration rates and on the population sizes (Table 3).

The equation system (5) can be rewritten for more than two populations. One needs, however, to decide whether to base the F_B values on pairwise differences among subpopulations or on an average difference among all pairs of subpopulations. If we want to solve the full model with n population sizes and $n(n-1)$ migration rates we need n^2 quantities. Table 2 shows that we cannot estimate all parameter with one locus. Adding a second locus enables us to solve for all parameters, but complicates the analysis even more. Additionally, we need to assume that the mutation rate is the same for all loci, which is certainly not true for all type of data, for example microsatellites.

Table 2: Variance quantities needed to estimate asymmetric migration rates and population sizes jointly. F_W is the "homozygosity" in a population, F_B is the "homozygosity" between pairs of subpopulations or the averages among all subpopulations.

Populations	Parameters	Quantities			Missing dimension	
		$F_B^{(all)}$	$F_B^{(pairs)}$	F_W	over all	pairwise
2	4	1	1	2	1	1
3	9	1	3	3	5	3
4	16	1	6	4	11	6
.	.	.	.	.	.	.
n	*n*^{2}	1	*n(n-1)/2*	*n*	*n(n-1)-1*	*n(n-1)/2*

Table 3: Estimates of migration rates $\gamma = 4N_em$ based on formula (5) in a two population system (Fig. 3). Averages ± standard deviations of 100 simulated data sets with the same population parameters are shown. For each data set 2 populations with 20 sampled individuals with 500bp sequence data were created according to Hudson (1983). T: the parameter values under which the data sets were created, C: Θ is the same for both subpopulations and M can be different for each population (Formula 7), D: Θ of the two subpopulations can be different, the migration rate M is the same for both subpopulations (Formula 6). N is the number of simulation runs used for calculating the averages and standard deviation. C_1, D_1: cases with illegal population parameters were discarded. C_2, D_2: illegal results were set to zero.

	Population 1		Population 2		N
	Θ	γ	Θ	γ	
T	0.01	1.00	0.01	1.00	-
C_1	0.0096 ± 0.0056	3.28 ± 6.35	0.0096 ± 0.0056	3.21 ± 8.43	66
C_2	0.0097 ± 0.0059	2.32 ± 5.34	0.0097 ± 0.0059	2.52 ± 7.05	100
D_1	0.0160 ± 0.0214	3.61 ± 6.09	0.0157 ± 0.0271	3.88 ± 12.33	95
D_2	0.0153 ± 0.0211	4.08 ± 7.49	0.0150 ± 0.0266	3.69 ± 12.05	100
T	0.01	10.00	0.01	1.00	-
C_1	0.0063 ± 0.0025	21.66 ± 59.37	0.0063 ± 0.0025	53.43 ± 162.01	34
C_2	0.0064 ± 0.0026	7.79 ± 35.74	0.0064 ± 0.0026	19.75 ± 96.75	100
D_1	0.0349 ± 0.1393	31.48 ± 75.02	0.0186 ± 0.0665	26.64 ± 107.75	46
D_2	0.0166 ± 0.0954	15.89 ± 52.86	0.0106 ± 0.0465	14.83 ± 73.86	100
T	0.05	10.00	0.005	1.00	-
C_1	0.0133 ± 0.0069	22.89 ± 41.42	0.0133 ± 0.0069	9.10 ± 35.20	34
C_2	0.0116 ± 0.0058	7.83 ± 26.27	0.0116 ± 0.0058	3.59 ± 20.73	100
D_1	0.1874 ± 0.7685	58.02 ± 258.86	0.0071 ± 0.0064	2.06 ± 3.29	39
D_2	0.0732 ± 0.4849	22.63 ± 162.88	0.0040 ± 0.0048	2.54 ± 4.35	100

Table 3 gives values for this more complex estimation procedure. The results are not really reassuring. Several of the 100 runs had to be discarded or parameters had to be set to zero because the values for the migration rates were negative for one population. Additionally, the estimates for the migration rates are biased upwards (cf. Beerli & Felsenstein 1998).

4. Maximum likelihood estimators based on allele frequencies

Rannala & Hartigan (1996) and Tufto *et al.* (1996) developed methods based on maximum likelihood to use the allele frequency data of n subpopulations to estimate migration rates directly. Both methods assume a specific probability distribution for the allele frequencies and use this distribution for their likelihood functions. Rannala & Hartigan (1996) used a simpler estimator to calculate the allele frequency estimates and therefore reduced the number of parameters for this approximate likelihood analysis to one, $4N_em$. This shortcut makes this method very fast and it is a better estimator of $4N_em$ than estimators based on F_{ST}. Rannala & Hartigans's implementation (`http://mw511.biol.berkeley.edu/homepage.html`) has, however, the same limitation as all symmetric estimators and will not to deliver correct estimates when the migration rates are asymmetric, but it may be possible to expand it to estimate a full migration matrix to handle asymmetric migration rates. The likelihood method of Tufto *et al.* (1996) is capable of estimating any migration scenario for a finite number of subpopulations. This likelihood method does not make the same assumptions about the allele frequency distribution as the method of Rannala & Hartigan (1996), but needs to estimate the most likely population allele frequencies given the sampled allele frequencies and a migration matrix. It seems that the approach of Tufto *et al.* (1996) would work well for allele frequency data, and Tufto has recently made the method available in form of S-PLUS functions [S-PLUS is a computer statistics package] (`http://www.math.ntnu.no/jarlet/migration/`).

5. Estimators using the coalescent

The introduction of coalescence theory (Kingman 1982a, b) changed the field of theoretical population genetics considerably. The coalescence theory is based on an approximation of a sampling process in a Wright-Fisher population (Fig. 1). Looking backwards in time, one can construct a genealogy of the sampled individuals. In a single population this process is only dependent on the effective population size. Kingman showed that the probability of a coalescence of two randomly chosen gene copies from a sample of size k in time interval u which is measured in generations scaled by mutation rate μ, is

$$\text{Prob}(u \mid N_e, \mu) = \exp\left(-u\frac{k(k-1)}{4N_e}\right)\frac{2}{4N_e}. \tag{8}$$

We can now calculate the probability Prob(g| N_e, μ) of a whole genealogy g by starting with k sampled alleles or sequences and, going back in time, multiplying the probabilities for each time interval u between nodes on this genealogy. We could now examine all possible genealogies and find the genealogy or a group of genealogies for which the probability given the population parameters is highest.

This framework can be expanded and now, for the first time it seems possible to include all possible population parameters into a single consistent framework (Hudson 1990), Kingman's original framework can be easily expanded by incorporating other population parameters (Hudson 1990) such as population growth (Griffiths & Tavaré 1994; Kuhner *et al.* 1998), migration rates (Nath & Griffiths 1996; Bahlo & Griffiths 1998; Beerli & Felsenstein 1998), recombination rates (Griffiths & Marjoram 1996), and selection (Krone & Neuhauser 1997; Neuhauser & Krone 1997). Including migration with a two population system (Fig. 4) changes the formula to

$$\text{Prob}(u \mid N_e^{(1)}, N_e^{(2)}, m_{21,} m_{12}, \mu) = \exp\left(-u\left[\frac{k_1(k_1-1)}{4N_e^{(1)}} + \frac{k_2(k_2-1)}{4N_e^{(2)}} + k_1 m_1 + k_2 m_2\right]\right)\beta \tag{10}$$

where β is the actual probability of the event, either a coalescent in population i, with probability $2/4N_e^{(i)}$ or a migration event from population j to i with probability m_{ji}.

Watterson (1975) used the number of segregating sites in a sample of sequences to infer population size. Coalescence theory facilitates finding expectations and variances for population parameters based on the segregating sites method (Wakeley 1998), so that there are two main streams of inference using the coalescent: (1) methods using segregating sites, (2) methods using maximum likelihood analysis based on the coalescence of the total sample. There are additional approximations of the coalescent process (for example Fu 1994), but they have not been extended to incorporate migration.

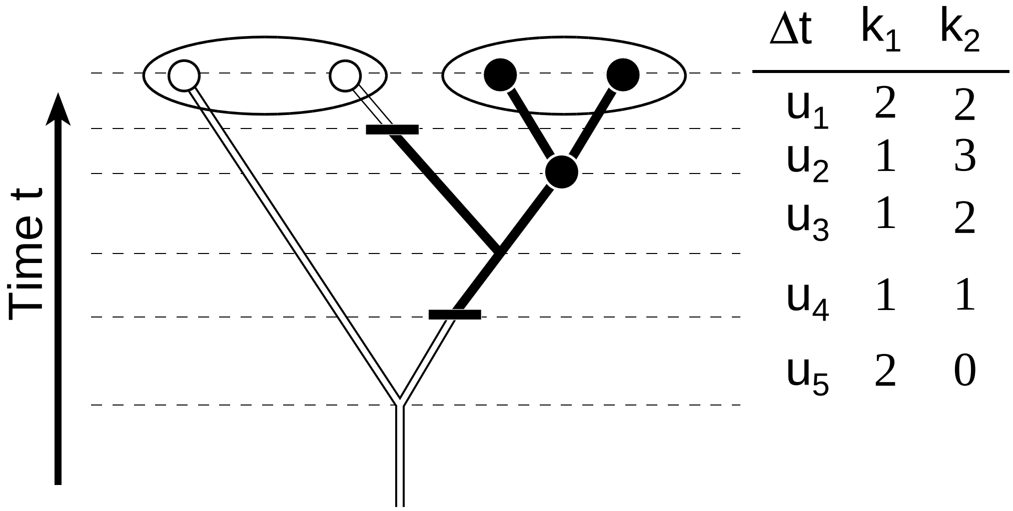

Fig. 4: Coalescent genealogy of 4 sampled individuals with migration. Samples from population 1 are shown in white, samples from population 2 in black. The migration events are shown as black bars. Δt is the time interval, k_1 and k_2 are the number of lineages in population 1 or 2, respectively, during a time interval u_i. Formula (8) shows the probability for each given time interval u_i on the genealogy.

5.1 Estimators based on analysis of segregating sites

Wakeley (1998) developed an approximation for the length of a genealogy when there are an infinite number of subpopulations that uses the number of segregating sites of a sample of genes from one to several subpopulations. This approximation to the length of the genealogy is then used to find a new estimator $M_S = 2Nm$ that is dependent on the number of segregating sites within subpopulations and the average number of segregating sites among subpopulations. This estimator is better than those based on F_{ST}, because it has a smaller standard deviation, but shares the same problems: (1) it estimates a symmetric migration rate; and (2) some data sets cannot be analyzed because the estimation of M_S fails when the number of segregating sites in a subpopulation is too large relative to the number of segregating sites over all subpopulations.

5.2 Maximum likelihood estimators using the coalescent

Kingman's probability calculation can be used to construct a maximum likelihood method for the estimation of population parameters. One can weight the probability of a given genealogy *g* with the likelihood of *g* which is the probability of the data given the genealogy. This is a quantity well known in phylogenetic studies. Because we do not know the true genealogy of our sample, we use the sum over all possible genealogies and then maximize this function to find the population parameters P,

$$L(\mathrm{P}) = \sum_{g\in G} \mathrm{Prob}(g \mid \mathrm{P})\,\mathrm{Prob}(D \mid g) \tag{10}$$

where *D* is the sampled data, and P are the population parameters we want to estimate. Taking into account the uncertainty of the genealogy should deliver more accurate parameter values than methods (Slatkin & Maddison 1989) that assume that the topology and the branch length of the genealogy is known.

Griffiths & Tavaré (1994) were the first o use this type of inference of population parameters. There is, however, a problem, that one cannot sample all genealogies, because there are too many. Our group (Joseph Felsenstein, Mary K. Kuhner, Jon Yamato, and PB) uses a Markov chain Monte Carlo Metropolis-Hastings importance sampling scheme to generate an approximation of the likelihood (10) where we integrate over a large sample of genealogies *G* with different topologies and different branch lengths (Kuhner *et al.* 1995 1998). The approach chosen by Griffiths & Tavaré (1994) and Bahlo & Griffiths (1998) appears rather different from ours (Kuhner *et al.* 1995; Kuhner *et al.* 1998; Beerli & Felsenstein 1998), but estimates the same quantitites. Felsenstein (unpubl.) showed that one can explain the method of Griffiths & Tavaré (1994) in terms used by our group and that it then is very similar to our approach. Both groups have released programs to estimate population parameters. For migration patterns these are `genetree` at the Internet site `http://www.maths.monash.edu.au/~mbahlo/mpg/gtree.html`, and `migrate` at `http://evolution.genetics.washington.edu/lamarc.html`. Both programs use a Markov chain Monte Carlo approach and sample genealogies that are then used to find the maximum likelihood estimate of a full migration matrix with population sizes. *Migrate* estimates the parameters

$$\mathrm{P} = \begin{pmatrix} \Theta_1 & M_{21} & \ldots & M_{n1} \\ M_{12} & \Theta_2 & \ldots & M_{n2} \\ \ldots & M_{ji} & \Theta_i & \ldots \\ M_{1n} & \ldots & \ldots & \Theta_n \end{pmatrix}$$

$$\text{with} \quad \Theta_{\mathrm{i}} = 4N_e\mu \quad \text{and} \quad M_{ji} = \frac{m_{ji}}{\mu}.$$

For potential users of these methods, differences in the respective underlying models of evolutionary change are perhaps more important than the similarities between the approaches. Griffiths & Tavaré (1994) and Bahlo & Griffiths (1998) use an infinite sites model that is inappropriate for highly variable sequence data, because it does not allow multiple substitutions at the same site and the researcher needs to discard such sites from the data. This is unfortunate, because discarding variable sites from the data set biases the population parameters; the population size estimates, for example, are too small. Our group uses a more generalized, two parameter mutation model developed by Felsenstein in 1984 (PHYLIP 3.2) (Swofford *et al.* 1996) that is an extension of Kimura's (1980) two-parameter model. Additionally a stepwise mutation model for microsatellites and a model for electrophoretic data are available (Beerli & Felsenstein 1998).

Simulated data sets were analyzed with `migrate`. As Table 4 shows, `migrate` delivers less biased estimates and smaller standard deviations than the other methods I have presented (cf. Tables 1, 3) when the population parameters are unequal.

Table 4: Simulation with unequal population parameters of 100 two-locus data sets with 25 individuals in each population and 500 base pairs (bp) per locus. T: Parameter values used to generate the data sets; Migrate: maximum likelihood estimator using the coalescence theory; B: Wright's relation between F_{ST} and γ with the correction for two populations; C_2: Θ is the same for both subpopulations and M can be different for each population (Formula 7); D_2: Θ of the two subpopulations can be different, the migration rate M is the same for both subpopulations (Formula 6); illegal results in C_2 and D_2 were set to zero.

	Population 1		Population 2	
	Θ	γ	Θ	γ
T	0.05	10.00	0.005	1.00
`Migrate`	0.0476 ± 0.0052	8.35 ± 1.09	0.0048 ± 0.0005	1.21 ± 0.15
B	-	11.46 ± 18.54	-	11.46 ± 18.54
C_2	0.0116 ± 0.0058	7.83 ± 26.27	0.0116 ± 0.0058	3.59 ± 20.73
D_2	0.0732 ± 0.4849	22.63 ± 162.88	0.0040 ± 0.0048	2.54 ± 4.35

5. Variable mutation rate

When we assume that loci are independent and neutral, then each locus delivers an independent estimate of the population parameters N_e and m. It is difficult, however, to exclude the mutation rate from these estimates, so that we normally estimate $\Theta = 4N_e\mu$ and $M = m/\mu$. In principle, this allows us to estimate $\gamma = 4N_e\ m$, which is independent of the mutation rate. For real data, this is only partly true: if there are only a few variable sites, or very few alleles in the data M will probably be high, but there is much uncertainty because we do not know whether these high values are caused by a high migration rate m or by a very small mutation rate μ. The estimate of γ therefore has large confidence intervals.

Using more than one locus improves the parameter estimates because, when the loci are unlinked, we have independent replicates from which we can calculate means and variances. We still need, however, to assume that the mutation rate is the same for all loci. If one is willing to assume that the mutation rate follows a Gamma-distribution with shape parameter α (Fig. 5), it is then easy to incorporate variable mutation rates into a maximum likelihood framework by integration over all possible mutation rates.

A comparison of `migrate` with the F_{ST} approach (Table 5) shows that the estimates for Θ are less biased when we take variation of the mutation rate into account. The estimates for the F_{ST}-based migration parameters γ are remarkably good, suggesting that the mutation rates really cancel in γ and that for the parameter values used, F_{ST} is a good estimator for symmetrical migration rates, even when the mutation rate is exponentially distributed.

Table 5: Estimates of population parameters from data with mutation rate variation between loci. The true mutation rate variation follows an exponential distribution ($\alpha = 1$, see Fig. 5). The values are estimates from one single data set for two populations with 30 electrophoretic marker loci. T: the parameter values used to generate the data sets; D: calculation based on formula (6); Migrate: maximum likelihood estimator based on the coalescence theory. θ is $4N_e\mu$, μ θ can be different for the two subpopulations and the scaled migration rate ($M = m/\mu$) is symmetrical. γ is θM .

	Population 1		Population 2		Shape α
	θ	γ	θ	γ	
T	1.00	1.00	1.00	1.00	1.0
`Migrate`	1.02	1.03	1.26	1.48	1.7
D	0.70	1.10	0.68	1.07	∞

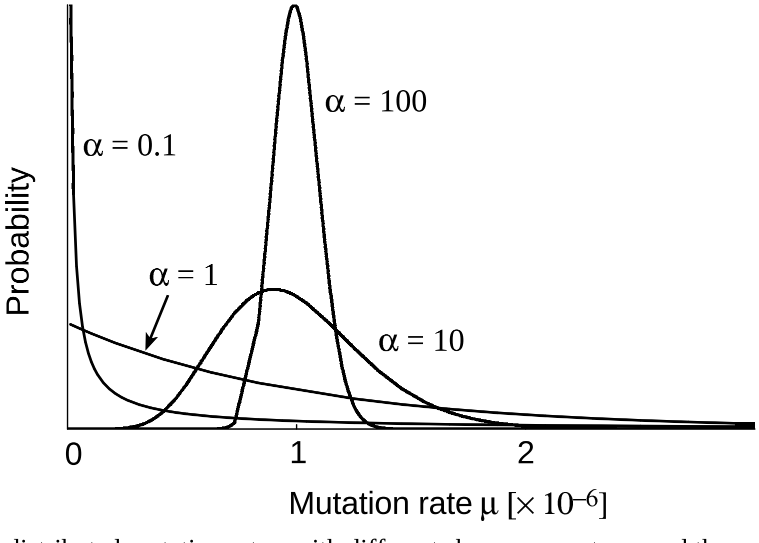

Fig. 5: Gamma distributed mutation rates, with different shape parameter α and the same mean μ for all curves. With α = 1 the Gamma distribution is an exponential distribution.

6. Testing whether migration rates are symmetric

Given the difficulties of developing a general framework to estimate asymmetric migration rates using F_{ST} , it seems rather cumbersome to develop a test for symmetry of migration rates. A way to solve the problem would be to generate simulated data sets using the estimated migration rates and population sizes (Rousset & Raymond 1997). These simulated data sets could then be used to compare the variances between the different migration rates with an ANOVA.

In the maximum likelihood framework, one would use a log-likelihood ratio test. The standard procedure for testing uses the assumption that in the limit, when we have an infinite amount of data the log-likelihood curves can be approximated by a normal distribution, so that we can use a χ^2-distribution for the test statistic. This approach may encounter difficulties because the current maximum likelihood estimators (Beerli & Felsenstein 1998; Bahlo & Griffiths 1998) approximate the likelihood using a Markov chain Monte Carlo approach. These methods are known to deliver good point estimates, but these approximate likelihood curves are exact only close to the point estimated; we (Mary K. Kuhner, Jon Yamato, & PB) are currently investigating this issue. The current speed of computers makes it too tedious to generate a data dependent test distribution for these maximum likelihood based estimators.

7. Discussion

The simulation studies for the F_{ST} based estimators (Tables 1, 3) show rather clearly that, when we violate the assumption that exchange of migrants between subpopulations is symmetrical the estimates of the migration rates are biased, if not completely wrong. Only those methods allowing the estimation of asymmetric migration rates have a chance of recovering the possible migration pattern in natural populations. This comes at a price: we need to estimate many more parameters. At least for estimation of growth rate and population sizes, coalescence theory suggests that we need to increase the quantity of data not by sampling more individuals, but by sampling more loci (Kuhner *et al.* 1998). This is probably true even for estimation of migration rates (cf. Wakeley 1998; PB unpubl.). Adding more individuals will mainly add lineages in the very recent past, so little additional information is gained about historical events at the bottom of the genealogy.

Even when by some external knowledge, for example using direct methods, we know that the subpopulations are approximately equal in size and the migration rates are symmetric, the F_{ST}-based approaches are still superseded by using the pseudo-maximum likelihood approach of Rannala & Hartigan (1996) or using the segregating sites approach of Wakeley (1998). Alternatively, the computationally more demanding but more accurate maximum likelihood methods that sample over all genealogies (Bahlo & Griffiths 1998; Beerli & Felsenstein 1998) can be employed.

8. Conclusion

During the last 60 years many researchers have used and continue to use F-statistics or genetic distances to make inferences about migration patterns. With the advent of newer methods, such as maximum likelihood using allele frequencies and their possible extensions or methods based on coalescence theory, tools now exist that allow us to estimate migration patterns without the unrealistic assumption of symmetry of migration rates or equal population sizes.

9. References

Bahlo M, Griffiths RC (1998) `Genetree`. Program and documentation distributed by the authors. Department of Mathematics, Monash University, Sydney, Australia.

Beerli P, Felsenstein J (1998) `Migrate` *- Maximum likelihood estimation of migration rates and population numbers*. Program and documentation distributed by the authors. Department of Genetics, University of Washington, Seattle, Washington.

Bosshart JL, Prowell DP (1998) Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology and Evolution* 13:202-206.

Charlesworth B (1998) Measures of Divergence Between Populations and the Effect of Forces that Reduce Variability. *Molecular Biology and Evolution* 15: 538-543.

Excoffier L, Smouse P (1994) Using Allele Frequencies and Geographic Subdivision to Reconstruct Gene Trees Within Species: Molecular Variance Parsimony. *Genetics* 136: 343-359.

Fu YX (1994) A phylogenetic estimator of effective population size or mutation rate. *Genetics*136:685-692.

Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society London Series B: Biological Sciences* 344: 403-410.

Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* 3: 479-502.

Hudson RR (1983) Properties of a natural allele model with intragenic recombination. *Theoretical Population Biology* 23: 183-210.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7:1-44.

Hudson RR, Slatkin M, Maddison WP (1994)

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Kingman J (1982a) The coalescent. *Stochastic Processes and their Applications* 13: 235-248.

Kingman J (1982b) On the genealogy of large populations. In *Essays in Statistical Science* (eds. Gani J, Hannan E), pp. 27-43. Applied Probability Trust, London.

Krone SM, Neuhauser C (1997) Ancestral Processes with Selection. *Theoretical Population Biology* 51: 210-237.

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140: 1421-1430.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149:429-434.

Li W-H (1976) Effect of migration on genetic distance. *American Naturalist* 110: 841-847.

Lynch M, Crease T (1990) The analysis of population survey data on DNA sequence variation. *Molecular Biology and Evolution* 7: 377-394.

Maruyama T (1970) Effective Number of Alleles in a Subdivided Population. *Theoretical Population Biology* 1: 273-306.

Maynard Smith J (1970) Population size, polymorphism, and the rate of non-Darwinian evolution. *American Naturalist* 104: 231-237.

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142: 1061-1064.

Nath H, Griffiths RC (1996) Estimation in an island model using simulation. *Theoretical Population Biology* 50: 227-253.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* USA 70: 3321-3323.

Nei M, Feldman MW (1972) Identity of genes by descent within and between populations under mutation and migration pressures. *Theoretical Population Biology* 3: 460-465.

Neigel JE (1997) A comparison of alternative strategies for estimating gene flow from genetic markers. *Annual Review of Ecology and Systematics* 28: 105-128.

Neuhauser C, Krone SM (1997) The genealogy of samples in models with selection. *Genetics* 145: 519-34.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research* 67: 147-158.

Relethford JH (1996) Genetic drift can obscure population history: problem and solution. *Human Biology 68*: 29-44.

Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* 142: 1357-1362.

Rousset F, Raymond M (1997) Statistical analyses of population genetic data: new tools, old concepts. *Trends in Ecology and Evolution* 12:313-317.

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research* 58: 167-75.

Slatkin M (1993) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.

Slatkin M, Hudson R (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555-562.

Slatkin M, Maddison W (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* 123: 603-613.

Swofford D, Olsen G, Waddell P, Hillis D (1996) Phylogenetic Inference. In *Molecular Systematics* (eds. Hillis D, Moritz C, Mable B), pp. 407-514. Sinauer Associates, Sunderland, Massachusetts.

Tufto J, Engen S, Hindar K (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* 144: 1911-1921.

Wakeley J (1998) Segregating sites in Wright's island model. *Theoretical Population Biology* 53:166-174.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7: 256-276.

Weir SB (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics* 15: 323-354.

A guide to software packages for data analysis in molecular ecology

Andrew SCHNABEL[1], Peter BEERLI[2], Arnaud ESTOUP[3], David HILLIS[4]

[1] *Department of Biological Sciences, Indiana University South Bend, South Bend, Indiana 46615 USA*

[2] *Department of Genetics, University of Washington, Seattle, Washington 98195 USA*

[3] *Laboratoire de Modélisation et de Biologie Evolutive, 488 rue croix Lavit, URBL-INRA, 34090 Montpellier, France*

[4] *Department of Zoology and Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712 USA*

Abstract. We briefly discuss software packages for the analysis of molecular ecological data, focusing on three levels of analysis: parentage and relatedness, population genetic structure, and phylogeny reconstruction. For the first two levels of analysis, we have gathered lists of some of the packages that we consider to be the most useful and user-friendly. For each package, we provide information on names of authors, date of latest update, compatible operating systems, types of data handled and analyses supported, availability, and literature citations. For software packages dealing with phylogeny reconstruction, we refer the reader to specific literature and website sources where this information has already been compiled.

1. Introduction

Molecular ecologists use protein or DNA markers to address questions about interactions between organisms and their biotic and abiotic environments. These studies often result in the generation of large and complex molecular data sets, and one of the challenges facing many workers is how to analyze those data properly. In this chapter, we present summary information on several of the numerous computer software packages for the analysis of genetic relationships among individuals, populations, and species. We do not claim that this information is complete, because or perfectly up-to-date, because new programs and updates of older programs are appearing almost monthly. Although some overlap will inevitably exist between different levels of analysis, we have chosen to divide the summary into three areas: parentage and relatedness, population genetic structure and gene flow, and phylogeny reconstruction.

2. Relationships among individuals: parentage and relatedness

Our understanding of social systems, mating behaviours, correlates of reproductive success, and dispersal patterns in natural populations depends on the possibility of genetically differentiating individuals, assigning both male and female parentage to individual progeny, and estimating with sufficiently high precision the genetic relatedness between groups or pairs of interacting individuals (Queller & Goodnight 1989; Cruzan 1998; Parker *et al.* 1998; Estoup, this volume). Studies of plant populations often use parentage analyses to address questions of outcrossing rates, genetic relatedness per se (Schnabel, this volume). On the other hand, in animal studies, relatedness and parentage are linked through studies of altruistic behavior, social and genetic mating systems, and kin selection (Hughes 1998; Rico, this volume). Although polymorphic genetic markers have been used for a long time in cases when pedigree information must be ascertained, as in animal breeding selection programs or in human paternity analysis, the advent of molecular markers with high levels of polymorphism has opened new perspectives for studies of parentage and relatedness in natural populations (Queller *et al.* 1993; Avise 1994; Estoup *et al.* 1994; Morin *et al.* 1994; Blouin *et al.* 1996; Taylor *et al.* 1997; Aldrich and Hamrick 1998; Hughes 1998; Parker *et al.* 1998; Prodöhl *et al.* 1998).

Compared with the number of software packages available for higher levels of analyses (see below), very few programs are available for the analysis of parentage or for the estimation of genetic relatedness (Appendix 1). Written specifically for plants, the set of programs by Ritland (1990) is the most widely used package for the analysis of outcrossing rates. More detailed parentage analyses are possible with PollenFlow (JD Nason, unpublished), which combines paternity exclusion analyses with the fractional paternity model of Devlin *et al.* (1988) and the maximum-likelihood models of Roeder *et al.* (1989) and Devlin & Ellstrand (1990), such that the user is able to obtain estimates both of pollen gene flow into the study population and relative fertilities of all possible male parents (Schnabel, this volume). A similar approach to parentage inference is taken in CERVUS (Marshall *et al.* 1998), which implements the likelihood models of Thompson (1975, 1976) and Meagher (1986). A very simple approach is taken by Danzmann (1997) in PROBMAX, which calculates probabilities that individuals are the offspring of specific parental pairs. The probability values indicate the number of loci sampled for each progeny that conform to Mendelian expectations for the pair of parents being tested. Finally, two well-developed programs are available for the estimation of relatedness based on the models of Queller and Goodnight (1989). The package Kinship tests hypotheses of pedigree relationships between pairs of individuals, and Relatedness uses a regression technique to measure relatedness between groups of individuals.

3. Relationships among populations: population genetic structure and gene flow

Many questions asked by molecular ecologists require that they conduct a survey of genetic diversity across several populations of a species. Such surveys estimate how much genetic variation a particular species maintains within its populations for a particular set of molecular markers (e.g., allozymes) and how that variation is partitioned among populations. Based on these data, inferences can be made about effective population sizes, natural selection, patterns of mating and dispersal, gene flow, and biogeographical history of the populations (e.g., Gentile & Sbordoni 1998; Godt & Hamrick 1998; Gonzalez *et al.* 1998; Xu *et al*. 1998). Hundreds of surveys of population genetic structure can be found in the literature, most of which prior to 1990 used either allozymes or mitochondrial DNA restriction sites (Hamrick & Godt 1989; Avise 1994). In the past decade, however, a large and rapidly growing number of studies have used a wider variety of molecular markers, such as DNA sequences, RAPDs, AFLPs, and microsatellites (e.g., Arden & Lambert 1997; Fischer & Matthies 1998; Paetkau *et al.* 1998; Winfield *et al.* 1998).

Given this great abundance of studies, it is not surprising that the number of available software packages for the analysis of population genetic structure is also large. Most of the early programs were written with allozymes in mind, and several of the currently available packages are still limited in the types of data that can be handled. In contrast to software for phylogenetic reconstruction (see below), there is no single source either in print or as a website that brings all of this information together. In collecting this diverse array of programs, we found that many of the more user-friendly programs (e.g., analyses they perform (Appendix 2). First, several packages calculate basic statistics of genetic variation, such as the proportion of polymorphic loci, the average number of alleles per locus, and heterozygosity. Those programs that handle a wider variety of data types also calculate statistics such as nucleotide diversity. Second, many packages will conduct tests for Hardy-Weinberg equilibrium. Third, most of the programs we report will estimate patterns of genetic stucturing using the hierarchical approach of Wright and/or Cockerham and Weir. A smaller proportion of the programs also include methods for analyzing microsatellite data using R_{ST} (e.g., Arlequin, Fstat, GENEPOP, RSTCALC) or Analysis of Molecular Variance (AMOVA in Arlequin). Fourth, a several of the programs will calculate one or more pairwise genetic distance measures (e.g., Nei's distance, Rogers distance), and will analyze those distances using some sort of clustering algorithm (e.g., UPGMA or neighbor joining). Last, several of the programs will estimate the level of linkage disequilibrium between loci. Although a number of home-grown programs for Macintosh computers must certainly exist, the large majority of packages we found were written for either a DOS or Windows platform. On a final note, the best available program for analyzing genetic structure within hybrid zones appears to be Analyse by Barton & Baird (1998).

All of the programs in Appendix 2 work within a traditional Wrightian framework of geographic structuring and estimation of gene diversity statistics based on allele frequencies. The introduction of coalesence theory by Kingman (1982) created new methods for analyzing population data (Hudson 1990; Beerli, this volume). Coalescence theory focuses on the sampled gene copies and looks backward in time to calulate the probability that two randomly chosen gene copies in the sample have a common ancestor *t* time units in the past. This process is driven only by the effective population size, N_e, and mutation rate. Kingman (1982) showed that the time when all lineages coalesced for a sample of 2, 4, and infinite gene copies is $2N_e$, $3N_e$, and $4N_e$, respectively. This Kingman coalescence process can be easily extended to incorporate other population parameters like population size, growth rate, recombination rate, and migration rates (Hudson 1990). Beerli (this volume) and Wakeley (1998) have shown that approaches based on coalescence theory are superior to approaches based on allele frequencies.

Two main groups of programs exist for coalescence analysis (Appendix 3), those that use segregating sites in the sample and those that integrate over all possible genealogies. In the sofware package SITES (Hey & Wakeley 1997), the coalescent is used to generate expectations for the number of segregating sites in a sample of sequences, and these expecetations are subsequently used to estimate population parameters. Most other programs listed in Appendix 3 integrate over all possible genealogies (e.g., MIGRATE). These programs are very computer intensive, but they use all possible information in the data, such as the history of mutation events. They also can be applied to several different types of molecular data other than sequence data. The general approach is to find the maximum likelihood of the population parameters, where the likelihood function is defined as the sum of probabilities over all possible genealogies. For each of these genealogies, one calculates the probability given the parameters and given the sampled data (Beerli, this volume).

4. Relationships among species: phylogeny reconstruction

The field of molecular systematics has become an increasingly important part of ecological studies during the past two decades. During that time, the number of computer programs for data preparation (e.g., entering and aligning DNA sequences), phylogenetic inference, tree comparisons, and other associated analyses has mushroomed to the point of being beyond the scope of any paper or website. For those readers considering phylogenetic analysis for the first time, we recommend reading Hillis (this volume) and the volume, *Molecular Systematics* (Hillis *et al.* 1996), within analysis of phylogenetic and population genetic data. Because that publication is now nearly 3 years old, some of the information may be out of date, but nonetheless it represents a good starting point. Alternatively, we advise visiting the website of the J. Felsenstein lab at the University of Washington (http://evolution.genetics.washington.edu). At that website, one can find descriptions of approximately 120 phylogeny packages that are arranged by (i) method of phylogenetic inference; (ii) computer systems on which they work; (iii) most recent listings; and (iv) those most recently updated.

5. Acknowledgments

We thank all those people who sent us information about their programs and all those who maintain websites with information about their own and other people's software.

References

Aldrich PR, Hamrick JL (1998) Reproductive dominance of pasture trees in a fragmented tropical forest mosaic. *Science*, **281**, 103-105.

Ardern SL, Lambert DM (1997) Is the black robin in genetic peril? *Molecular Ecology*, **6**, 21-28.

Avise, JC (1994) *Molecular Markers, Natural History and Evolution*, Chapman & Hall, London UK.

Bahlo M, Griffiths RC (1998) Inference from gene trees in a subdivided population. *Theoretical Population Biology*.

Barton NH, Baird SJE (1998) Analyse 2.0. Edinburgh: http://helios.bto.ed.ac.uk/ evolgen/index.html.

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F, (1998) GENETIX, logiciel sous Windows^{TM} pour la génétique des populations. Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier (France)

Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, **5**, 393-401.

Cornuet JM, Luikart G (1997) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001-2014.

Cruzan MB (1998) Genetic markers in plant evolutionary ecology. *Ecology*, **79**, 400-412.

Danzmann RG (1997) PROBMAX: A computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity*, **88**, 333.

Devlin B, Ellstrand NC (1990) The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. *Evolution*, **44**, 248-259.

Devlin B, Roeder K, Ellstrand NC (1988) Fractional paternity assignment, theoretical development and comparison to other methods. *Theoretical and Applied Genetics*, **76**, 369-380.

Estoup A, Solignac M, Cornuet JM (1994). Precise assessment of the number of patrilines and of genetic relatedness in honey bee colonies. *Proceedings of the Royal Society of London: Biological Sciences*, **258**, 1-7.

Fischer M, Matthies D (1998) RAPD variation in relation to population size and plant fitness in the rare *Gentianella germanica* (Gentianaceae). *American Journal of Botany*, **85**, 811-820.

Garnier-Gere P, Dillmann C (1992) A computer program for testing pairwise linkage disequilibrium in subdivided populations. *Journal of Heredity*, **83**, 239.

Gentile G, Sbordoni V (1998) Indirect methods to estimate gene flow in cave and surface populations of *Androniscus dentiger* (Isopoda: Oniscidea). *Evolution*, **52**, 432-442.

Godt MJW, Hamrick JL (1998) Allozyme diversity in the endangered pitcher plant *Sarracenia rubra* ssp. *alabamensis* (Sarraceniaceae) and its close relative *S. rubra* ssp. *rubra*. *American Journal of Botany*, **85**, 802-810.

Gonzalez S, Maldonado JE, Leonard JA, Vila C, Barbanti Duarte JM, Merino M, Brum-Zorrilla N, Wayne RK (1998) Conservation genetics of the endangered Pampas deer (*Ozotoceros bezoarticus*). *Molecular Ecology*, **7**, 47-56.

Goodman SJ (1997) Rst Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology*, **6**, 881-885.

Goudet J (1995) Fstat version 1.2: a computer program to calculate Fstatistics. *Journal of Heredity*, *86*, 485-486.

Griffiths RC, Tavaré S (1996) Computational methods for the coalescent. In: *Progress in Population Genetics and Human Evolution* (eds. Donnelly P, Tavaré S). IMA Volumes in Mathematics and its Applications. Springer Verlag, Berlin.

Hamrick JL, Godt MJW (1989) Allozyme diversity in plant species. In: *Plant PopulationGenetics, Breeding and Genetic Resources* (eds. Brown AHD, Clegg MT, Kahler AL, Weir BS), pp. 43-63. Sinauer, Sunderland, MA.

Hey J, Wakeley J (1997) A coalescent estimator of the population recombination rate. *Genetics*, **145**, 833-846.

Hillis DM, Moritz C, Mable BK (1996) *Molecular Systematics*. Sinauer, Sunderland, MA.

Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1-44.

Hughes C (1998) Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology*, **79**, 383-399.

Kingman J (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235-248.

Kuhner MK, Yamato J, Felsenstein, J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* , **140**, 421-430.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-439.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations**. *Molecular Ecology,* 7,** 639-655**.**

Meagher, TR (1986) Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents**. *The American Naturalist*, 128,** 199-215**.**

Morin PA, Wallis J, Moore JJ, Woodruff DS (1994) Paternity exclusion in a community of wild chimpanzees using hypervariable simple sequence repeats. *Molecular Ecology*, **5**, 469-478.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Vyse E, Ward R, Strobeck C (1998) Variation in genetic diversity across the range of North American brown bears. *Conservation Biology*, **12**, 418-429.

Parker PG, Snow AA, Schug MD, Booton GC, Fuerst PA (1998) What molecules can tell us about populations: choosing and using a molecular marker. *Ecology*, **79**, 361-382.

Prodöhl P A, Loughry WJ, McDonough CM, Nelson WS, Thompson EA, Avise JC (1998) Genetic maternity and paternity in a local population of armadillo assessed by microsatellite DNA markers and field data. *The American Naturalist*, **151**, 7-19.

Queller CR, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258-259.

Queller CR, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in Evolution and Ecology*, **8**, 285-288.

Rannala B, Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research*, **67**, 147-158.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, **94**, 9197-9201.

Raymond M, Rousset F (1995a) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Raymond M, Rousset F (1995b) An exact test for population differentiation. *Evolution*, **49**, 1280-1283.

Ritland K (1990) A series of FORTRAN computer programs for estimating plant mating systems. *Journal of Heredity*, **81**, 235-237.

Roeder K, Devlin B, Lindsay BG (1989) Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics*, **45**, 363-379.

Rozas J, Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Computer Applic. Biosci*, **11**, 621-625.

Rozas J, Rozas R (1997) DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. *Comput. Applic. Biosci.*, **13**, 307-311.

Schneider S, Kueffer JM, Roessli D, Excoffier L (1997) Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Sork VL, Campbell D, Dyer R, Fernandez J, Nason J, Petit R, Smouse P, Steinberg E (1998) Proceedings from a Workshop on Gene Flow in Fragmented, Managed, and Continuous Populations. National Center for Ecological Analysis and Synthesis, Santa Barbara, California. Research Paper No. 3. Available at http://www.nceas.ucsb.edu/papers/geneflow/

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: *Molecular Systematics* (eds. Hillis DM, Moritz C, Mable BK), pp.407-514. Sinauer, Sunderland, MA.

Taylor AC, Horsup A, Johnson CN, Sunnucks P, Sherwin B (1997) Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology*, **6**, 9-19.

Thompson, EA (1975) The estimation of pairwise relationships**. *Annals of Human Genetics*, 39,** 173-188**.**

Thompson, EA (1976) Inference of genealogical structure. ***Social Science Information*, 15,** 477-526**.**

Tufto J,Engen S, Hindar K (1996) Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics*, **144**, 1911-1921.

Yeh FC, Yang RC, Boyle TJB, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belgian Journal of Botany*, **129**, 157.

Wakeley J (1998) Segregating sites in Wright's island model. *Journal of Theoretical Population Biology*, **53**, 166-174.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847-855.

Weir, B (1996) *Genetic Data Analysis*. Sinauer, Sunderland, MA.

Winfield MO, Arnold GM, Cooper F, Le Ray M, White J, Karp A, Edwards KJ (1998) A study of genetic diversity in *Populus nigra* ssp. *betulifolia* in the Upper Severn Area of the UK using AFLP markers. *Molecular Ecology*, **7**, 3-11.

Xu J, Kerrigan RW, Sonnenberg AS, Callac P, Horgen PA, Anderson JB (1998) Mitochondrial DNA variation in natural populations of the mushroom *Agaricus bisporus*. *Molecular Ecology*, **7**, 19-34.

Appendix 1: List of software packages for the study of parentage and relatedness using molecular markers

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
CERVUS v. 1.0, 17/6/98 (T Marshall)	Windows95	Diploid, codominant markers	Uses a most-likely approach to parentage inference and estimates confidence in parentage of most likely parents. Can be used to calculate allele frequencies, run simulations to determine critical values of likelihood ratios and analyse parentage in populations of animals and plants. A simulation system can estimate the resolving power of a series of single-locus marker systems for parentage inference.	Freeware from http://helios.bto.ed.ac.uk/evolgen/index.html Marshall *et al.* (1998)
Kinship v. 1.2, (KF Goodnight, DC Queller, T Posnansky)	MacOS (PowerPC and 68K)	Diploid, codominant markers	Performs maximum likelihood tests of pedigree relationships between pairs of individuals in a population. The user enters two hypothetical pedigree relationships, a primary hypothesis and a null hypothesis, and the program calculates likelihood ratios comparing the two hypotheses for all possible pairs in the data set. Includes a simulation procedure to determine the statistical significance of results. Also calculates pairwise relatedness statistics.	Freeware from http://www.bioc.rice.edu/~kfg/GSoft.html Queller & Goodnight (1989)
MLT (K Ritland)	DOS	Diploid or tetraploid, codominant markers	A set of programs that finds maximum-likelihood estimates of outcrossing rates for plant populations. Also estimates parental gene frequencies and inbreeding coefficients. Special programs within the package can handle autotetraploids and ferns.	Freeware by contacting the author at ritland@unixg.ubc.ca Ritland (1990)

Appendix 1: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
PollenFlow v. 1.0, 26/3/98 (JD Nason)	MacOS (PowerPC and 68K)	Diploid, codominant markers	Implements two different models. First, a paternity exclusion-based model estimates total rate of pollen immigration from a single external source into a defined local population. Second, a likelihood-based model estimates relative male fertility within a population as well as pollen immigration from one or more external sources. Male fertility estimates are adjusted to eliminate biases due to cryptic gene flow.	Freeware by contacting the author at john-nason@uiowa.edu Sork *et al.* (1998)
PROBMAX , 17/11/97 (RG Danzmann)	DOS	Codominant and dominant/ recessive diploid markers	Ascertains the parentage of individuals when genotypic data on both parents and progeny are available. Also includes PROBMAXG, which generates possible progeny genotypes from the parental mixtures to test whether a given set of genetic markers will be able to discriminate all progeny back to parental sets, and PROBMAXN, which allows testing of possible parent/progeny assignments if null alleles segregating at some markers are suspected.	Freeware by anonymous ftp to 131.104.50.2 (password = danzmann) or contact the author at rdanzman@ uoguelph.ca Danzmann (1997)
Relatedness v. 5.0.4, 29/6/1998 (KF Goodnight, DC Queller	MacOS (PowerPC and 68K	Diploid, codominant markers	Estimates genetic relatedness between demographically-defined groups of individuals using a regression measure of relatedness. Calculates symmetrical and asymmetrical relatedness and jackknife standard errors. Allows up to 32 demographic variables in defining those individuals to be used in calculating the relatedness	Freeware from http://www.bioc.rice.edu/~kfg/GSoft.html Queller & Goodnight (1989)

Appendix 2: List of software packages that will analyze geographically structured populations using traditional estimators based on gene frequencies.

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Analyse v. 2.0, 5/98 (SJE Baird, NH Barton)	MacOS (PowerPC)	Diploid and haploid genetic markers, quantitative trait values, spatial coordinates (1 and 2 dimensions), environmental variables	Likelihood analysis of data from hybrid zones. Performs three types of analyses: general data handling (e.g., selecting subsets of the data satisfying particular criteria), analysis of random fluctuations in genotype frequency (e.g., estimating Fst, Fis, and standardized linkage disequilibrium), and analysis of a set of multilocus clines (e.g., estimating variation between clines).	Freeware from http://helios.bto.ed.ac.uk/evolgen/index.html Barton & Baird (1998)
Arlequin v. 1.1, 17/12/97 (S Schneider, JM Kueffer, D Roessli, L Excoffier)	Windows 3.1 or later	RFLPs, microsatellites, allozymes, RAPDs, AFLPs, allele frequencies, DNA sequences	Calculates gene and nucleotide diversity, mismatch distribution, haplotype frequencies, linkage disequilibrium, tests of Hardy-Weinberg equilibrium, neutrality tests, pairwise genetic distances, analyses of molecular variance (AMOVA).	Freeware from http://anthropologie.unige.ch/ arlequin Schneider *et al.* (1997)
DnaSP v. 2.52, 9/97 (J Rozas, R Rozas); v. 2.9 is available as a beta version	Windows 3.1 or later	DNA sequences	Estimates several measures of DNA sequence variation within and between populations (in noncoding, synonymous or nonsynonymous sites), and also linkage disequilibrium, recombination, gene flow, and gene conversion parameters. Also can conduct several tests of neutrality.	Freeware from http://www.bio .ub.es/~julio/DnaSP.html Rozas & Rozas (1995, 1997)
Fstat v. 1.2, 12/95; Fstat for windows v. 2.3, is available as beta upon request (J Goudet)	DOS; new version will be Windows compatible	Allozymes, microsatellites, mtDNA RFLPs	Calculates gene diversity statistics of Weir and Cockerham (Weir, 1996). Computes jackknife and bootstrap confidence intervals of the statistics or can test gene diversity statistics using a permutation algorithm.	Freeware by writing to J. Goudet at jerome.goudet@ izea.unil.ch Goudet (1995)

Appendix 2: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
GDA, 11/7/97 (PO Lewis, D Zaykin)	Windows 3.1 or later	Allozymes, microsatellites	Calculates standard gene diversity measures, Wright's F-statistics using the method of Weir and Cockerham (Weir, 1996), genetic distance matrices, UPGMA and neighbor-joining dendrograms, exact tests for disequilibrium	Freeware; http://chee.unm.edu/gda Designed to accompany *Genetic Data Analysis* (Weir, 1996).
GENEPOP v. 3.1b, 12/97 (M. Raymond, F. Rousset)	DOS	Allozyme, microsatellites	Calculates exact tests for Hardy-Weinberg equilibrium, population differentiation, and genotypic disequilibrium among pairs of loci. Computes estimates of classical population parameters, such as allele frequencies, Fst, and other correlations. Includes Linkdos (Garnier-Gere and Dillmann, 1992), which is a program for testing pairwise linkage disequilibrium.	Freeware from 3 ftp sites: ftp://ftp.cefe.cnrs-mop.fr/genepop/ ftp://ftp2.cefe.cnrs-mop.fr/pub/pc/ msdos/genepop/ ftp://isem.isem.univ -montp2.fr/pub/ pc/genepop/ Raymond & Rousset (1995a, b)
GENETIX v. 3.3, 14/05/98 (K Belkhir, P Borsa, L Chikhi, J Goudet, F Bonhomme)	Windows95/ NT	Allozymes, microsatellites	Calculates estimates of classical parameters (e.g., genetic distances, variability parameters, Wright's fixation indices, linkage disequilibrium) and tests their departure from null expectations through permutation techniques. The interface is not user-friendly for everyone, because it is currently only in French.	Freeware from http://www.univ-montp2.fr/~genetix/genetix.htm Belkhir *et al.* (1998)
Immanc, 17/10/97 (JL Mountain)	Windows 3.1 or later, MacOS (PowerPC), NeXT HP-RISC, Sun UltraSPARC	Allozymes, microsatellites, RFLPs	Tests whether or not an individual is an immigrant or is of recent immigrant ancestry. The program uses Monte Carlo simulations to determine the power and significance of the test.	Freeware from http://mw511.biol .berkeley.edu/ software.html Rannala & Mountain (1997)

Appendix 2: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	*Analyses supported*	**Availability, literature citations**
Migrlib v. 1.0 (J Tufto)	Unix (available as a collection of S-Plus functions and some C code)	Allele frequencies	Estimates the pattern of migration in a subdivided population from genetic differences generated by local genetic drift. Functions are also provided for carrying out likelihood ratio tests between alternative models such as the island model and the stepping stone model.	Freeware from http://www.math.ntnu.no/~jarlet/migration Tufto *et al.* (1996)
PMLE12 v. 1.2, 4/3/96 (B Rannala)	Windows 3.1 or later, MacOS (PowerPC or 68K), NeXTStep	Allozymes, mtDNA RFLPs	Estimates the gene flow parameter theta for a collection of two or more semi-isolated populations by (pseudo) maximum likelihood. For discrete-generation island model, theta=2Nm. For a continuous-generation island model, theta is the ratio of the immigration rate phi to the individual birth rate lambda.	Freeware from http://mw511.biol.berkeley.edu/bruce/exec.html Rannala & Hartigan (1996)
POPGENE v. 1.21, 22/12/97 (F Yeh, RC Yang, T Boyle)	Windows 3.1 or later	Co-dominant or dominant markers using haploid or diploid data.	Calculates standard genetic diversity measures, tests of Hardy-Weinberg Equilibrium, Wright's F-statistics, genetic distances, UPGMA dendrogram, neutrality tests, linkage disequilibrium	Freeware from http://www.ualberta.ca/~fyeh/index.htm Yeh & Boyle (1997); Yeh *et al.* (1997)
RSTCALC v. 2.2, 6/10/97 (SJ Goodman)	DOS, Windows 3.1 or later	Microsatellites	Performs analyses of population structure, genetic differentiation, and gene flow. Calculates estimates of Rst, tests for significance and calculates 95% CI.	Freeware from http://helios.bto.ed.ac.uk/evolgen Goodman (1997)
TFPGA (Tools for Population Genetic Analyses), 12/5/98 (MP Miller)	Windows 3.1or later	Codominant (allozyme) and dominant (RAPD, AFLP) genotypes	Calculates descriptive statistics, genetic distances, and F-statistics. Performs tests for Hardy-Weinberg equilibrium, exact tests for genetic differentiation, Mantel tests, and UPGMA cluster analyses.	Freeware from http://herb.bio.nau.edu/~miller No citation available

Appendix 3: List of software packages that will analyze geographically structured populations using estimators based on coalescence.

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Bottleneck v. 1.1.03, 27/11/97 (JM Cornuet, G Luikart, S Piry)	Windows95	Allele frequencies	Detects recent reductions in effective population size from allele frequency data. Tests whether a set of loci shows a significant excess of heterozygosity (i.e., the observed heterozygosity is larger than the heterozygosity expected at mutation-drift equilibrium and assuming a given mutation model).	Freeware from http://www.ensam. inra.fr/~piry Cornuet & Luikart (1997)
Fluctuate v. 1.50B, 6/2/98 (M Kuhner, J Yamato)	Windows 95/NT, MacOS (PowerMac); UNIX; available also as C source code	DNA sequences	Estimates the effective population size and an exponential growth rate of a single population using maximum likelihood and Metropolis-Hastings importance sampling of coalescent genealogies.	Freeware from http://evolution. genetics.washington.edu/lamarc.html Kuhner *et al.* (1995, 1998)
Genetree, 9/6/98 (M Bahlo, RC Griffiths)	Windows 95/NT, Dec Alpha; available also as C source code	DNA sequences	Finds maximum likelihood estimates of population sizes, exponential growth rates, migration matrices, and time to the most recent common ancestor.	Freeware from http://www.maths. monash.edu.au/ ~mbahlo/mpg/ gtree.html Griffiths & Tavaré (1996) Bahlo & Griffiths (1998)
Migrate-0.4 v. 0.4.3, 25/5/98 (P Beerli)	Windows 95/98/NT, MacOS (PowerMac), Dec Alpha, LINUX/Intel, NeXTStep; available also as C source code	Allozymes, microsatellites, DNA sequences	Menu driven, character-based program that finds 4+1 maximum-likelihood estimates of population parameters for a two-population model: effective population sizes for subpopulation1 and subpopulation 2, migration rates between the two subpopulations, and for multilocus data, a shape parameter for the distribution of the mutation rate.	Freeware from http://evolution. genetics.washington .edu/lamarc.html Beerli, this volume

Appendix 3: Continued

Package name, lastest update (author)	**Operating system**	*Types of data handled*	**Analyses supported**	**Availability, literature citation**
Migrate-n v. Alpha-3, 25/5/98 (P Beerli)	Windows 95/98/NT, MacOS (PowerMac), Dec Alpha, LINUX/Intel, NeXTStep; available also as C source code	Allozymes, microsatellites, DNA sequences	Menu driven, character-based program that finds n*n maximum-likelihood estimates of population parameters for n-population model: effective population sizes for each subpopulation, migration rates between the n subpopulations, and for multilocus data, a shape parameter for the distribution of the mutation rate.	Freeware from http://evolution. genetics.washington .edu/lamarc.html Beerli, this volume
Recombine v. 1.0, 17/6/98 (MK Kuhner, J Yamato, J Felsenstein)	MacOS (PowerMac), Windows95/NT ; available as C source code that will compile on DEC ULTRIX, DEC alpha, INTEL machines, NeXT, SGI, but needs gcc to compile on Suns	DNA or RNA sequences, single nucleotide polymorphisms	Fits a model which has a single population of constant size with a single recombination rate across all sites. It estimates 4Nu and r, where N is the effective population size, u is the neutral mutation rate per site, and r is the ratio of the per-site recombination rate to the per-site mutation rate.	Freeware from http://evolution. genetics.washington .edu/lamarc.html No citation available
SITES v. 1.1, 21/4/98 (J Hey)	DOS, MacOS; also available as ANSI C source code	DNA sequences	Generates tables of polymorphic sites, indels, codon usage. Computes numbers of synonymous and replacement base positions, pairwise sequence differences, and GC content. Performs group comparisons and polymorphism analyses and estimates historical population parameters. Primarily intended for data sets with multiple closely related sequences.	Freeware from http://heylab.rutgers .edu/index.html #software Hey & Wakeley (1997) Wakeley & Hey (1997)

Analysis of geographically structured populations: (Traditional) estimators based on gene frequencies

Peter Beerli
Department of Genetics, Box 357360,
University of Washington, Seattle WA 98195-7360,
Email: beerli@genetics.washington.edu

This is an introduction and overview of the currently used methods for the analysis of population subdivision and estimation of migration rates. We will discuss theoretical population models such as the group of single migration parameter models with two or *n* islands, stepping stone models, and multi-parameter models such as the migration matrix model. In this lecture I will concentrate on approaches using gene frequencies, and will neglect complicating evolutionary forces such as selection and age structured populations. Sewall Wright introduced 1922 the fixation index F and the term F statistic. This summary statistic is based on the avariability in and between subpopulations. For different data types (e.g. enzyme electrophoretic markers, microsatellite markers, sequence data) different coefficients are in use (e.g. F_{ST}, R_{ST}). These different methods take into account that the variability generating process, mutation, is different for different types of data. Most of these F_{ST} based estimators were developed for symmetrical population models. I will discuss an extension which is able to cope with asymmetrical population models, compare these different methods, and analyze their performance. Confidence limits of $\mathrm{F_{ST}}$ of population parameters can be found using the boostrap over loci, or a maximum likelihood ratio test if we are working in a maximum likelihood framework. Most of these methods will be superseded by either maximum likelihood concepts in the context of gene frequency data, or methods taking the genealogy of the sample into account [second lecture].

Introduction and context

In the early twenties Sewall Wright introduced the notation of the fixation index F to characterize the influence of mating systems on heterozygosity in inbred guinea pig lines. Such an inbred line looks like a "natural" population (Fig. 1) with very few individuals; genes are passed in a random fashion to offspring, who replace their parents. WRIGHT (1973) wrote: "It became evident that the same set of parameters, the F-statistics, which measure relative change of heterozygosis in an array of diverging inbred lines also measures the differentiation of their gene frequencies" and we can apply it to geographically structured populations. F-statistic itself gives us a summary statistic about isolation of subpopulations and their variability, but if we want to understand more clearly the underlying processes we want to know the population parameters such as population size and migration rate and perhaps be able to determine routes of gene flow between populations. A general overview on the problems of estimating effects of migration on gene frequencies can be found in FELSENSTEIN (1982).

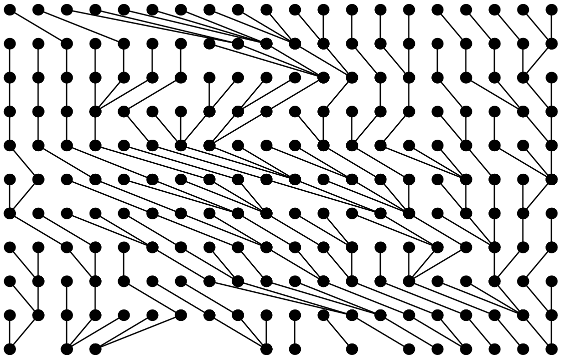

Figure 1: Wright-Fisher population model: idealized population with random mating. The genes are rearranged so that we can see the genealogy. Each line of dots is a generation, the number of individuals is 10 with 20 genes

Models of geographically structured populations

Most of the migration models have several very restrictive assumptions and assume a specific way of replacing individuals from one generation to the other (Fig. 2).

The *n* island model (Figure 4: A,B) **(WRIGHT, 1931):** All subpopulations have the same effective population size, $N_e^{(i)}$. Individuals migrate from one subpopulation to the other with the same rate m. The distances between subpopulations are not taken into account.

Stepping stone model (Figure 4: C) **(MALECOT, 1950; KIMURA, 1953):** All subpopulations have the same effective population size, $N_e^{(i)}$. The migration rate m is constant and defines the rate of exchange from one neighboring population to the other along the possible paths.

Continuum model (WRIGHT, 1940): in which a populaiton is spread out in geographical continuum. Unfortunately, these models have mathematical properties so that they are not able to define stable subpopulations at one location through time, although they come very close to our intuition about real populations.

Migration matrix model (Figure 4: B,D)**(BODMER and CAVALLI-SFORZA, 1968):** All subpopulations have the same effective population size, $N_e^{(i)}$. The migration rates between subpopulations

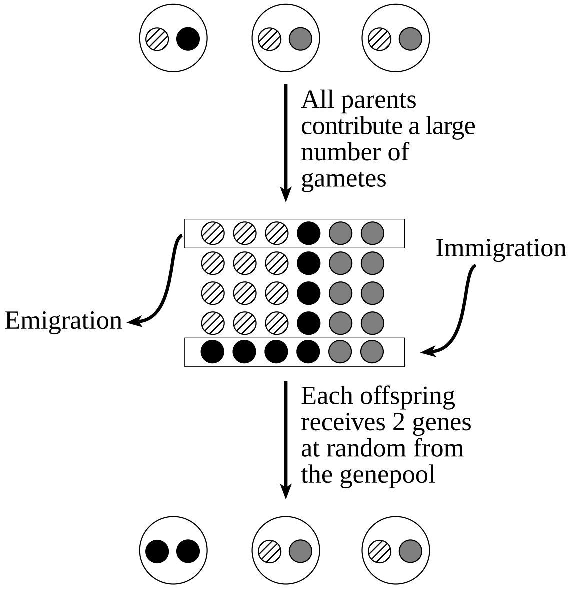
All parents contribute a large number of gametes
Immigration
Emigration
Each offspring receives 2 genes at random from the genepool

Figure 2: Sequence of events in a migration model

can be different and for four populations (Figure 3) one could have for example the following migration matrix (I chose the migration rates to reflect an isolation by distance model).

$$\begin{pmatrix} - & m & \frac{m}{2} & \frac{m}{4} \\ m & - & m & \frac{m}{2} \\ \frac{m}{2} & m & - & m \\ \frac{m}{4} & \frac{m}{2} & m & - \end{pmatrix}$$

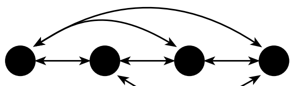

Figure 3: Four populations

In an arbitrary migration model some of the migration path can be disallowed (set to 0.0). A further extension of these models includes variable subpopulation size.

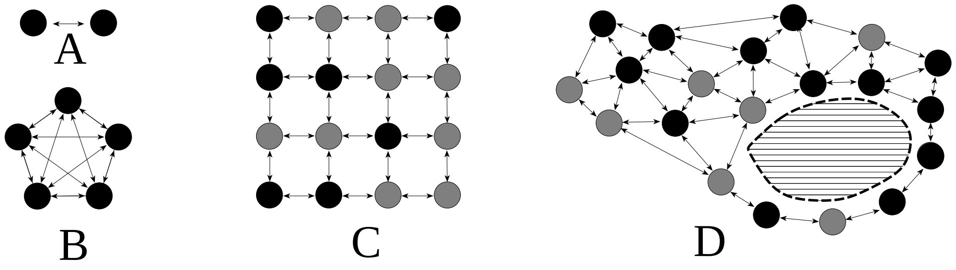

Figure 4: Migration models: A, B: *n*-island model, C: Stepping stone model (2-dimensional), D: arbitrary migration matrix model. Black disks are sampled subpopulations, gray disks are unsampled subpopulations

Transformation of variability into summary statistics

To develop a summary statistic we can use the variability in and between populations, but we need to consider the underlying model of evolution.

$\mathrm{F_{ST}}$[1], $\mathrm{G_{ST}}$, **Infinite allele model:** Weir (1996), Slatkin (1991)

$\mathrm{R_{ST}}$, **Microsatellites:** Slatkin (1993)

$\mathrm{F_{ST}}$, **Sequences:** Hudson *et al.* (1992b), Nei (1982), and Lynch and Crease (1990)

Assessments of confidence limits

Bootstrapping over loci is appropriate to generate confidence limits.

Estimates of migration rate

Wright's formula

$$\mathrm{F_{ST}} = \frac{1}{1+4Nm}$$

to transform $\mathrm{F_{ST}}$ values into migration rates is still most commonly used. It assumes that the mutation rate is 0.0 and the number of subpopulations is very large. Also, we will not gain any information about the population sizes themselves, they are convoluted with the migration rates. Additionally, a mutation rate of 0.0 is perhaps appropriate for enzyme electrophoretic data, but it is not appropriate for microsatellites or intron-sequences. We can incorporate these relaxations of the assumptions. In a two population model (Fig. 5) we can solve the following equation system using the homozygosity within a population F_W and the homozygosity between populations F_B (Nei and Feldman, 1972) by replacing $4N\mu$ with Θ and m/μ with $\mathcal{M}$

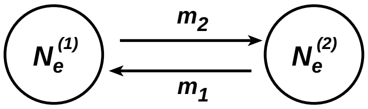

Figure 5: Two population model with population sizes $N_e^{(1)}$, $N_e^{(2)}$, and migration rates m_1, m_2.

$$\begin{aligned}
F_W^{(1)} &= \frac{1}{2N_1} + \left(1 - 2\mu - 2m_1 - \frac{1}{2N_1}\right) F_W^{(1)} + 2m_1 F_B \\
F_W^{(2)} &= \frac{1}{2N_2} + \left(1 - 2\mu - 2m_2 - \frac{1}{2N_2}\right) F_W^{(2)} + 2m_2 F_B \qquad (1) \\
F_B &= F_B\left(1 - \mu - m_1 - m_2\right) + m_1 F_W^{(1)} + m_2 F_W^{(2)}
\end{aligned}$$

With one locus we can only solve for 3 parameters, either a constant $\Theta = 4N\mu$ ($4 \times$ effective population size $N_e \times$ mutation rate μ; because we do not know the mutation rate we include it into the estimate) and two migration rates $\mathcal{M}_1 = m_1/\mu$ and $\mathcal{M}_2 = m_2/\mu$ or for two different Θ_1 and Θ_2 values and one symmetric migration rate $\mathcal{M}$.

[1] Weir (1996) called this θ, but we will use Θ for $4N_e\mu$ in approaches using coalescence theory

Problems with F-statistic approaches:

- Wright's formula is often inappropriate for real world situations.
- Rather complicated estimation procedure, when we consider more than two populations and want to estimate population sizes and migration rates.
- If for some subpopulations the F_W are smaller than the F_B the estimation procedure breaks down.
- Gene frequencies are considered to be the true gene frequencies of the sampled populations. This can produce wrong results with small sample sizes.
- Parameter estimates based on $\mathrm{F_{ST}}$ do not make full usage of the data [see second lecture].

Maximum likelihood estimators

- Estimation using PMLE of RANNALA and HARTIGAN (1996)
- Estimation using the approach of TUFTO *et al.* (1996)

Other approaches

- Distance measures (NEI and FELDMAN, 1972)
- Parsimony related (EXCOFFIER and SMOUSE, 1994)
- Rare allele approach (SLATKIN, 1985)

Summary

- We recognize several different migration models: n-island model, stepping stone model, and migration-matrix model. Their assumptions strongly influence the estimates of population parameters. Complications in computations of estimates can arise by relaxing assumptions such as equal population size or symmetric migrations.
- Quality of transformation of the variability in the data into summary statistics is dependent how well the underlying model for the estimator fits the data.
- Current F-statistic approaches assume symmetry of migrations and often equal population sizes.
- Allowing for unequal population sizes and unequal migration rates complicates migration rate estimation considerably. Also, in a F-statistics framework it is not possible to estimate all four parameters of a two population model with one locus (e.g. mtDNA).

- Maximum likelihood approaches, e.g. work by RANNALA and HARTIGAN (1996) and TUFTO *et al.* (1996), utilizing the distribution of gene frequencies promise to give good results, but some of this work is still in the beginning stages.
- For sequence data the current estimators based on F-statistics are less accurate than coalescence theory based estimators, because they do not not use information about the history of mutations.

Bibliography

BARTON, N. and SLATKIN, M., 1986 A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity (Edinburgh) **56 (Pt 3):** 409–15.

BODMER, W. F. and CAVALLI-SFORZA, L. L., 1968 A migration matrix model for the study of random genetic drift. Genetics **59:** 565–592.

EXCOFFIER, L. and SMOUSE, P., 1994 Using allele frequencies and geographic subdivision to reconstruct gene trees within species: Molecular variance parsimony. Genetics **136:** 343–359.

FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies? Journal of Theoretical Biology **96:** 9–20.

HUDSON, R., BOOS, D., and KAPLAN, N., 1992a A statistical test for detecting geographic subdivision. Molecular Biology and Evolution **9:** 138–151.

HUDSON, R., SLATKIN, M., and MADDISON, W., 1992b Estimation of levels of gene flow from dna sequence data. Genetics **132:** 583–9.

KIMURA, M., 1953 “stepping-stone” model of population. Annual Report of the National Institute of Genetics, Japan **3:** 62–63.

LYNCH, M. and CREASE, T., 1990 The analysis of population survey data on DNA sequence variation. Molecular Biology and Evolution **7:** 377–394.

MALECOT, G., 1950 Some probability schemes for the variability of natural populations (french). Annales de l’Universite de Lyon, Sciences, Section A **13:** 37–60.

NEI, M., 1982 Evolution of human races at the gene level. In *Human Genetics, Part A: The Unfolding Genome*, edited by B. Bohhe-Tamir, P. Cohen, and R. Goodman, pp. 167–181, Alan R. Liss, New York.

NEI, M. and FELDMAN, M. W., 1972 Identity of genes by descent within and between populations under mutation and migration pressures. Theoretical Population Biology **3:** 460–465.

RANNALA, B. and HARTIGAN, J., 1996 Estimating gene flow in island populations. Genetical Research **67:** 147–158.

RANNALA, B. and MOUNTAIN, J., 1997 Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci **94:** 9197–9201.

ROUSSET, F. and RAYMOND, M., 1997 Statistical analyses of population genetic data: new tools, old concepts. Trends in Ecology and Evolution **12:** 313–317.

SLATKIN, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

SLATKIN, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–92.

SLATKIN, M., 1991 Inbreeding coefficients and coalescence times. Genetical Research **58:** 167–75.

SLATKIN, M., 1993 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

SLATKIN, M. and BARTON, N., 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.

SLATKIN, M. and MADDISON, W., 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

SLATKIN, M. and VOELM, L., 1991 Fst in a hierarchical island model. Genetics **127:** 627–629.

TUFTO, J., ENGEN, S., and HINDAR, K., 1996 Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics **144:** 1911–1921.

WEIR, BRUCE, S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16:** 97–159.

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. American Naturalist **74:** 232–248.

WRIGHT, S., 1973 The origin of the f-statistics for describing the genetic aspects of population structure. pp. 3-26 in Genetic Structure of Populations, ed. N. E. Morton. University Press of Hawaii, Honolulu .

Software, with emphasis on methods using gene frequencies

[this list is certainly not complete]

- ANALYSE An ”easy-to-use” MacOS application for the analysis of hybrid zone data. Calculates several statistics: e.g. FST, and isolation by distance.
  Website through `http://helios.bto.ed.ac.uk/evolgen`

- ARLEQUIN is an exploratory population genetics software environment able to handle large samples of molecular data (RFLPs, DNA sequences, microsatellites), while retaining the capacity of analyzing conventional genetic data (standard multi-locus data or mere allele frequency data). A variety of population genetics methods have been implemented either at the intra-population or at the inter-population level.
  Website at `http://anthropologie.unige.ch/arlequin`

- DNASP computes (among lots of other things) different measures of the extent of DNA divergence between populations, and from these measures it computes the average level of gene flow, assuming the island model of population structure. DnaSP estimates the following measures: dST, gST and Nm, NST and Nm, FST and Nm (Rozas, J. and R. Rozas. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. 13: 307-311). Binary for Windows 3.1 and 95.
  Website at `http://www.bio.ub.es/∼julio/DnaSP.html`

- GDA (Genetic Data Analysis) is a Microsoft Windows program for analyzing discrete genetic data based on WEIR (1996).
  Website at `http://chee.unm.edu/gda`

- GENEPOP is a population genetics software package for DOS and can be fetched by anonymous ftp from `ftp.cefe.cnrs-mop.fr` in the directory /PUB/PC/MSDOS/GENEPOP or can be used through a web interface at
  `http://www.curtin.edu.au/curtin/dept/biomed/teach/genepop/web_docs/gene_form.html`

- IMMANC is a program designed to test whether or not an individual is an immigrant or is of recent immigrant ancestry. The method is appropriate for use with allozyme, microsatellite, or restriction fragment length data. Loci are assumed to be in linkage equilibrium. The power of the test depends on the number of loci, the number of individuals sampled, and the extent of genetic differentiation between populations RANNALA and MOUNTAIN (1997). Binaries for Macintosh, Windows, and NEXTSTEP.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- MICROSAT estimates several indices using microsatellite data. C source code and binaries for DOS and Macintosh.
  Website at `http://lotka.stanford.edu/microsat.html`

- PMLE12 estimates the gene flow parameter theta for a collection of two or more semi-isolated populations by (pseudo) maximum likelihood using either allozyme or mtDNA RFLP data RANNALA and HARTIGAN (1996). C source code and binaries for Macintosh, Windows, and NEXTSTEP.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- POPGENE computes both comprehensive genetic statistics (e.g., allele frequency, gene diversity, genetic distance, G-statistics, F-statistics) and complex genetic statistics (e.g., gene flow, neutrality tests, linkage disequilibria, multi-locus structure). Binaries for Windows3.1, Windows95.
  Website at `http://www.ualberta.ca/ fyeh/index.htm`.

- RELATEDNESS 4.2 calculates average genetic relatedness among groups of individuals specified by up to three user-defined demographic variables. It also calculates F-statistics measuring inbreeding and genetic differences among sub-populations. Binary for Macintosh.
  Website at `http://www-bioc.rice.edu/~kfg/GSoft.html`

- RSTCALC is a program for performing analyses of population structure, genetic differentiation and gene flow using microsatellite data. Binary for Windows.
  Website through `http://helios.bto.ed.ac.uk/evolgen`

Analysis of geographically structured populations: Estimators based on coalescence

Peter Beerli
Department of Genetics, Box 357360,
University of Washington, Seattle WA 98195-7360,
Email: beerli@genetics.washington.edu

The rapid increase in the collection of population samples of molecular sequences, plus the great expansion of the use of microsatellite markers, makes it possible to investigate the patterns and rates of migration among geographically subdivided populations with much greater power than was previously possible. The difficulty with methods for analyzing these data has been that they do not allow the researcher to observe the genealogical tree of ancestry of the sampled sequences, but only make an estimate of it which has a great deal of uncertainty. Taking the uncertainty in our estimate of the genealogy into account is the major challenge for a proper statistical analysis of these data. The statistical approach of maximum likelihood is used to infer these rates and patterns, using the Markov Chain Monte Carlo (MCMC) method of computing the likelihoods. This method samples genealogies from the space of possible genealogies, using an acceptance-rejection method to concentrate the sampling in the regions which contribute most to the outcome. Even though the number of possible genealogies is vast, the MCMC sampling can avoid wasting computer time on possibilities that can have made little contribution to the observed outcome. This sampling of different genealogies in computing a likelihood for the parameters correctly accounts for our lack of knowledge of the true gene tree.

It can be shown that these ML-methods are superior to methods based on F_{ST}. Additionally, ML-methods can take into account variability in mutation rate and can estimate all relevant population parameters jointly and also analyze cases with different population sizes and migration rates. Comparison of different data types reveals that number of loci sampled is a key factor in reducing the variability of the parameter estimates.

The coalescent

Most current population genetics analyses are using theoretical findings of Sewall Wright and R. A. Fisher which were made in the early 20th century. Their work is based on a view which uses discrete generations of idealized individuals passing their genes to offspring in the next generation. This "looking forward" strategy implies that calculation of the probability of a given genotype is rather difficult. Kingman (1982a,b) formalized a "looking backward" strategy: the coalescent. Hudson (1990) and Donnelly and Tavaré (1997) give comprehensive reviews on the subject. Coalescence theory takes the relatedness of the sample into account, so it incorporates random genetic drift and mutation. This approach makes it very easy to calculate probabilities of a genealogy of a sample of individuals with a given effective population size, $\mathrm{P}(g|\Theta)$. Hudson (1990) and others showed that we can extend this single population approach to multiple populations and estimate migration rates and also that we can include other forces such as growth, recombination, and selection.

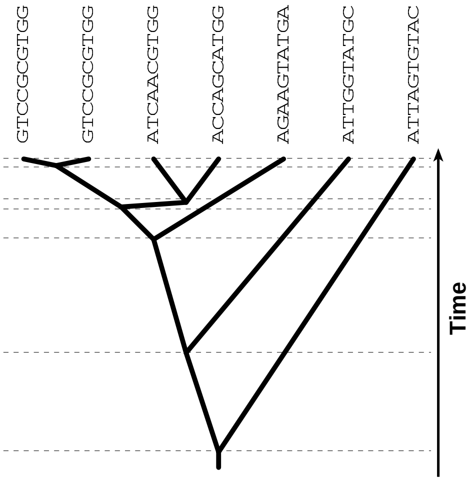

Figure 1: A coalescent tree with sampled sequences

Markov chain Monte Carlo (MCMC) integration

Construction of random genealogies (Simulation studies) is simple with the coalescent approach (e.g. the method of Slatkin and Maddison 1989). Inference of parameters is much harder, especially when we want not to lose any information in the data (Felsenstein 1992). In a likelihood framework we would like to simply integrate over all possible genealogies G and solve for the population parameters Θ at the maximum likelihood

$$\mathrm{L}(\Theta) = \int_{g \in G} \mathrm{P}(g|\Theta)\mathrm{P}(D|g)dg, \tag{1}$$

where $\mathrm{P}(D|g)$ is the likelihood of the genealogy with the sample data. This is not possible; there are too many different topologies with different branch lengths. But we can approximate by using a biased random walk through the genealogy space and then infer the parameters from the sampled genealogies correcting for the biased sampling:

$$\mathrm{L}(\Theta) = \int_{g \sim \mathrm{P}(g|\Theta_0)\mathrm{P}(D|g)} \frac{\mathrm{P}(g|\Theta)}{\mathrm{P}(g|\Theta_0)} dg \tag{2}$$

(MCMC: Hammersley and Handscomb 1964, MCMC and coalescence: Kuhner et al. 1996)

Table 1: Simulation with unequal known parameters of 100 two-locus datasets with 25 individuals in each population and 500 base pairs (bp) per locus. Std. dev. is the standard deviation.

	Population 1		Population 2	
	$4N_e\mu$	$4N_em$	$4N_e\mu$	$4N_em$
Truth	0.0500	10.00	0.0050	1.00
Mean	0.0476	8.35	0.0048	1.21
Std. dev.	0.0052	1.09	0.0005	0.15

Two population exchange migrants

We will explore the details of the MCMC mechanism in a simple two population model with the parameters: $\Theta_1 = 4N_e^{(1)}\mu$, $\Theta_2 = 4N_e^{(2)}\mu$, $\mathcal{M}_1 = m_1/\mu$, $\mathcal{M}_2 = m_2/\mu$ (we need to scale by the unknown mutation rate μ of our data).

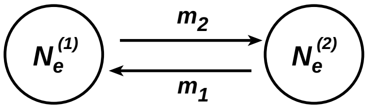

Figure 2: Two population model with population sizes $N_e^{(1)}$, $N_e^{(2)}$, and migration rates m_1, m_2.

- Assumptions: Population have constant size and exist forever, migration rate is constant through time, and the genetic markers are neutral.
- We can jointly estimate migration rates and population sizes
- Example of a simulation study (Table 1), where I generated 100 single locus data sets and then analyzed them with the program MIGRATE (Beerli 1997).
- Problems: perhaps not a natural situation; how long do we need to run the genealogy sampler?

Migration matrix model

- Assumptions: same as with 2 populations
- Simulation studies with (a) 4 sampled populations and (b) with 3 sampled population and one population where we don't have data.

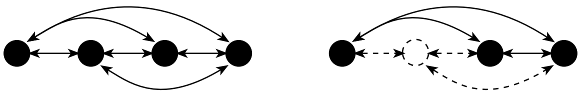

Figure 3: Population structure used in simulations.

- Problems: how many genealogies to sample? Number of parameters increases quadratically.

Comparison with F_{ST}

Simulation studies can show that the ML-estimator delivers better result than F_{ST}, and results are still accurate when population sizes and/or migration rates are unequal (Table 1).

Hypothesis testing using likelihood ratios

The maximum likelihood framework makes it easy to test hypotheses. I expect that these tests will supersede standard test based on F_{ST}. I will show a few examples and hope that I am able to have a version of MIGRATE finished in March so that everybody can experiment with their own data in the “data section”.

$$H_0 : \hat{N}_e = N_e^{(x)}$$

$$\text{Test-statistic:} \qquad -2\log\left(\frac{L(\Theta_x)}{L(\hat{\Theta})}\right) \leq \chi^2_{df,\alpha}$$

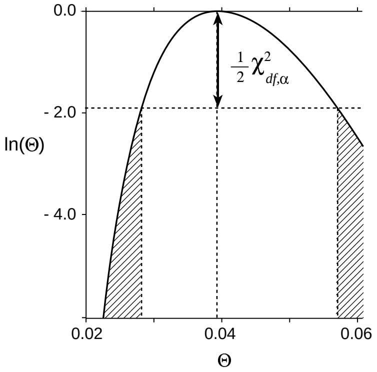

Figure 4: Likelihood ratio test: dashed areas are outside of the 95% confidence limit. Θ is $4N_e\mu$; $df = 1$, $\alpha = 0.05$

Data type and mutation rate

We have mutation models for infinite allele model, microsatellite stepwise mutation model (Valdez and Slatkin 1993, Di Rienzo et al. 1994), and finite sites sequence model (e.g. Swofford et al. 1996).

What’s the effect of the data type to the estimate of migration rates? The data type is not that important, for the quality of the migration rate estimates, but the variance of the estimates is dependent on the number of unlinked loci (Fig. 5) having independent coalescent trees and the variability in the data, the more segregating sites or polymorphic loci are present the better the estimates of the migration rates.

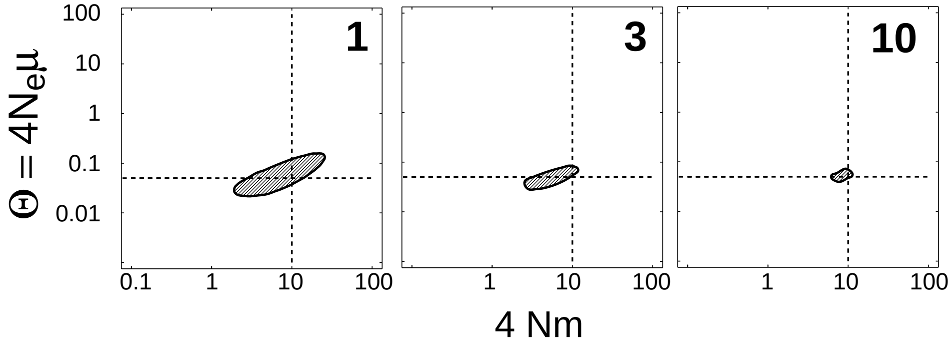

Figure 5: Variance of parameter estimates: the dashed area is the 95% confidence area, the numbers 1, 3, and 10 are the numbers of sampled loci

Mutation rate is not constant: incorporation of the variance of the mutation rate is possible by assuming that it follows a Gamma distribution (Fig. 6) and estimating the shape parameter α of this distribution jointly with the population parameters by integrating over all mutation rates x

$$L(\Theta,\mathcal{M},\alpha)=\prod_{l}\int_{0}^{\infty}\frac{e^{-\alpha x/\Theta_1}x^{\alpha-1}}{\Gamma(\alpha)\left(\frac{\Theta_1}{\alpha}\right)^{\alpha}}L(x,\Theta_l,\mathcal{M}_l)dx,$$

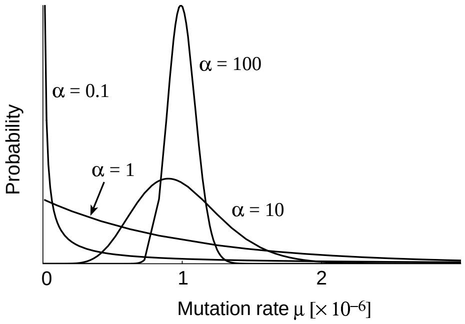

Figure 6: Gamma distributed mutation rates, with different shape parameter α and the same mean

Summary

- Coalescence theory enables us to estimate population parameters by including sample data and taking the possible histories of the populations into account.
- Expansion of the coalescence model to any migration model is possible.
- Maximum likelihood ratio test of arbitrary hypotheses.
- Multi-locus enzyme electrophoretic data and microsatellite markers delivers good migration rate estimates compared to mtDNA sequence data, because the quality of the result is dependent on the number of loci and the variability in the data.
- The assumption that the mutation rate over loci is constant is obviously wrong for electrophoretic markers and microsatellites and taking the variation of the mutation rate into account should improve the estimates of population parameters.

Bibliography

Citations with a ⋆ are recommended to read and/or introductory, citations with a • are rather difficult.

BEERLI, P., 1997, MIGRATE DOCUMENTATION version 0.3. Distributed over the Internet: http://evolution.genetics.washington.edu/lamarc.html.

DI RIENZO A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN, and N. B . FREIMER, 1994, Mutational processes of simple-sequence repeat loci in human populations. Genetics **91** (8): 3166–3170.

⋆ DONNELLY, P. and S. TAVARÉ, 1997, *Progress in population genetics and human evolution*. IMA volumes in mathematics and its applications **87**, Springer, New York.

FELSENSTEIN, J., 1973, Maximum likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics **25**: 471–492.

FELSENSTEIN, J., 1988, Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics **22**: 521–565.

FELSENSTEIN, J., 1992, Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetics Research **59**: 139–147.

• GRIFFITHS, R. and S. TAVARÉ, 1994, Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci **344** (1310): 403–10, Department of Mathematics, Monash University, Clayton, Victoria, Australia.

HAMMERSLEY, J. and D. HANDSCOMB, 1964, *Monte Carlo Methods*. Methuen and Co., London.

⋆ HUDSON, R. R., 1990, Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44.

• KINGMAN, J., 1982a, The coalescent. Stochastic Processes and their Applications **13**: 235–248.

• KINGMAN, J., 1982b, On the genealogy of large populations. In *Essays in Statistical Science*, edited by J. Gani and E. Hannan, pp. 27–43, Applied Probability Trust, London.

⋆ KUHNER, M., J. YAMATO, and J. FELSENSTEIN, 1995, Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140** (4): 1421–30, Department of Genetics, University of Washington, Seattle 98195-7360, USA.

METROPOLIS, N., A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, and E. TELLER, 1953, Equation of state calculations by fast computing machines. Journal of Chemical Physics **21**: 1087–1092.

NATH, H. and R. GRIFFITHS, 1993, The coalescent in two colonies with symmetric migration. Journal of Mathematical Biology **31** (8): 841–51.

NATH, H. and R. GRIFFITHS, 1996, Estimation in an Island Model Using Simulation. Theoretical Population Biology **50**: 227–253.

• NOTOHARA, M., 1990, The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology **29** (1): 59–75.

SLATKIN, M., 1991, Inbreeding coefficients and coalescence times. Genetical Research **58** (2): 167–75, Department of Integrative Biology, University of California, Berkeley 94720.

⋆ SLATKIN, M. and W. MADDISON, 1989, A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123** (3): 603–613, Department of Zoology, University of California, Berkeley 94720.

⋆ SWOFFORD, D., G. OLSEN, P. WADDELL, and D. HILLIS, 1996, Phylogenetic Inference. In *Molecular Systematics*, edited by D. Hillis, C. Moritz, and B. Mable, pp. 407–514, Sinauer Associates, Sunderland, Massachusetts.

• Takahata, N., 1988, The coalescent in two partially isolated diffusion populations. Genetical Research **52** (3): 213–22.

• Takahata, N. and M. Slatkin, 1990, Genealogy of neutral genes in two partially isolated populations. Theoretical Population Biology **38** (3): 331–50, National Institute of Genetics, Mishima, Japan.

Valdez A. M. and M. Slatkin, 1993, Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133** (3): 737–749, Department of Zoology, University of California, Berkeley 94720.

Wakeley, J. and J. Hey, 1997, Estimating ancestral population parameters. Genetics **145** (3): 847–855.

Software, with emphasis on using the coalescent

[this list is certainly not complete]

- LAMARC package [**L**ikelihood **A**nalysis with **M**etropolis **A**lgorithm using **R**andom **C**oalenscence. Three programs are currently available: Coalesce, Fluctuate, and Migrate. C-source code and binaries for Windows, Mac, LINUX, DUNIX, NEXTSTEP.
  Website at `evolution.genetics.washington.edu/lamarc.html`

- MISAT estimates the effective population size of a single population using microsatellite data and can also test if the one-step model or a multi-step model is appropriate. Binaries for Macintosh and Windows.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- Sites is a computer program for the analysis of comparative DNA sequence data (Hey and Wakeley, 1997. A coalescent estimator of the population recombination rate. Genetics 145: 833-846) . C source code and binaries for DOS and Macintosh.
  Website at `http://heylab.rutgers.edu`

- Upblue is a least square estimator for population size (Fu, Y. X., 1994. An phylogenetic estimator of effective population size or mutation rate. Genetics 136:685-692). Fortran program or use the website directly to calculate results
  `http://www.hgc.sph.uth.tmc.edu/fu/`

- Calculation of 4Nm using the method of Slatkin and Maddison (1989), you need to calculate the minimal mumber of migration events on the genealogy either by hand or using MacClade (Maddison and Maddison 1992, Sinauer). Pascal source code.
  Website at `http://mw511.biol.berkeley.edu/software.html`

- Several programs for the estimation of population size, exponential growth, recombination rate, migration rate, time of the last common ancestor. Contact Bob Griffiths (email: ...) for more information.