Effects of migration patterns on estimates of dispersal rates inferred from genotypic data: a simulation-based study focused on the software STRUCTURE

Maia H. Malonzo[1,2], Rachel G. Ravago-Gotanco[1], Arturo O. Lluisma[1,*]

[1]Marine Science Institute, University of the Philippines, 1101 Quezon City
[2]Current Address: Department of Information and Computer Science, Aalto University School of Science, Helsinki, Finland

Recent developments in the analysis of genetic data now make possible the direct measurement of migration rates through individual-based assignment methods at ecological time frames relevant for resource management. While several software implementing these assignment methods have been examined for accuracy under various conditions of spatial patterns and rates of gene flow and population size, previous analyses have not examined the effects of temporal variations in dispersal rate on assignment accuracy. In this study, we evaluated the assignment accuracy of the widely used software, STRUCTURE, using simulated genetic datasets generated to reflect two patterns of temporal variation in dispersal rate: a normal distribution and a negative binomial distribution, the latter reflecting a pattern of migration commonly observed in natural populations in which the movement of a large number of migrants into the sink population is a rare event. We also evaluated the accuracy of different assignment models and varying sample sizes. The results of the simulations suggest that at the mean migration rate of 5 individuals per generation, STRUCTURE exhibits greater assignment accuracy from a negative binomial distribution relative to a normal distribution at smaller sample sizes of 20-50 individuals. This however is attributed to greater population structure among populations in which migration followed a negative binomial distribution, and its effect on recovering more accurate assignments. At sample sizes of 100 to 200 individuals, assignment accuracy was similar for the two distributions. Increasing the sample size generally resulted in reduced specificity in classification. At the larger sample sizes, increasing numbers of false positives were recovered for both normal and negative binomial distribution patterns, likely due to the proportionally increased probability of sampling individuals with recent migrant ancestry. Incorporating prior population information into models of migrant inference resulted in higher levels of accuracy in detecting actual migrants (true positive assignments), and reducing false positive assignments.

*Corresponding author
Email Address: aolluisma@gmail.com
Received: September 2, 2012
Revised: December 27, 2012
Accepted: December 27, 2012
Published: February 21, 2013
Editor-in-charge: Eduardo A. Padlan
Reviewer: Jingky P. Lozano-Kühne

KEYWORDS

assignment methods, migration, accuracy, genetic data, simulation

INTRODUCTION

The task of managing and conserving marine resources requires a clear understanding of population connectivity, or the extent to which populations are linked by the exchange of individuals (Sale et al 2005). The rates, spatial scales, and patterns of dispersal and connectivity among local populations have important implications to a wide range of ecological and evolutionary processes, from the dynamics, maintenance and persistence of populations to genetic diversity, local adaptation and speciation (reviewed in Strathmann et al 2002, Kinlan et al 2005), and is of critical importance in the context of resource management and conservation. Such an understanding is crucial considering that natural populations, particularly in the marine realm, generally consist not of closed, independent populations, but rather of multiple populations interconnected by various population dynamic processes (metapopulations, Kritzer and Sale 2004). Demographic exchange and gene flow facilitated by larval dispersal and adult migration, represent important mechanisms influencing the coupling of population dynamics among neighboring populations. Consequently, an estimation of levels and patterns of connectivity is critical to the design and implementation of management interventions, particularly for spatially-explicit management schemes (Sale et al 2005).

Molecular genetic markers have found widespread application as a rich source of information on the evolutionary and ecological history of populations (Avise 1994). Successful migrants leave behind a genetic trail of their movements, offering a means of estimating population connectivity via gene flow (Hellberg et al 2002). However, equilibrium-based methods for estimating gene flow rely on indirect estimators of genetic differentiation in the form of standardized variance in allele frequencies which are based on theoretical models of population structure (F_{ST}, Wright 1931). Such estimators are thought to reflect migration rates averaged over evolutionary time, and not over ecological timescales relevant for guiding management initiatives (Hedgecock et al 2007, Jones et al 2009). Moreover, the assumptions underlying theoretical models of population structure are often violated in natural populations, and represent significant drawbacks to estimating migration rates from F_{ST} (Bossart and Prowell 1998, Whitlock and McCauley 1999).

Recent developments in the generation and analysis of genetic data now make possible the measurement of gene flow over ecological time frames (reviewed in Manel et al 2005). These direct methods focus on individual-based methods for identifying populations of origin (assignment methods) or putative parents (parentage analysis). These approaches are most promising towards yielding contemporary estimates of connectivity and dispersal over temporal scales spanning single to several generations (Waser and Strobeck 1998, Berry et al 2004, Pearse and Crandall 2004), of relevance to fisheries management.

The algorithms for assignment methods, and the variety of software implementing them were developed fairly recently (Waples and Gaggioti 2006). Although a number of studies have demonstrated the potential of these programs, the accuracy of the estimates of various population parameters generated by the algorithms remain insufficiently characterized statistically. Thus, the statistical capabilities of these algorithms can constrain the accuracy of estimates derived from their use. Our inference of connectivity can only be as accurate as these algorithms allow. In addition, the range of parameter values within which the algorithms are expected to generate reliable results is not well-defined. Because the use of these algorithms is practically inevitable, and because the estimates that these algorithms yield may be directly used as basis for actual resource management decisions, the accuracy of the estimates must be well-understood to be reliable and useful.

Perhaps the most popular and widely-used program used for inferring population structure based on assignment methods is STRUCTURE (Pritchard et al 2000). Implementing a Bayesian approach, the software compares various models of population structure (where the number of subpopulations K is fixed) using an optimization criterion based on minimal linkage disequilibrium within subpopulations, i.e. finding groups of individuals that are as far as possible from disequilibrium. Consequently, the number of subpopulations can be inferred, and individuals assigned probabilistically to these populations. Migrants whose population of origin (source) is different from the population where it was sampled are also identified.

In this study, we investigated the effects of several parameters on the accuracy of STRUCTURE to identify migrant individuals, in particular: (1) the temporal variation in dispersal rate among sampled populations; (2) implementation of ancestry models with or without prior information on geographic sampling of individuals; and (3) the number of individuals sampled for analysis (sample size). Existing population genetic models have generally ignored the effects of temporal variation in dispersal rate, assuming instead normal distributions to approximate complex phenomena. Ecological phenomena however, such as spatial and temporal patterns of distribution and abundance among natural populations, can be better characterized by a negative binomial distribution (e.g., Pielou 1969, Taylor et al 1979) where the variance can be greater than the mean. Meanwhile, the effects of sample size on accuracy of population structure delineation and assignment probabilities is analyzed to determine sample size thresholds at which reliable inferences can be made (e.g., Paetkau et al 2004).

The accuracy of the analytical tools such as STRUCTURE is typically assessed using simulated data as input so that the parameters whose effects are under investigation can be controlled with precision (which is not possible with actual environmental samples). Here we used the program Nemo (Guillaume and Rougement 2006) to simulate genotypic data of populations subject to two patterns of temporal migration rate. In both scenarios, individuals migrate unidirectionally (source to sink population only) at the same migration rate (averaged over several generations). However, the rate varies between generations following a normal and negative normal binomial distribution, respectively. Data from 20 to 200 randomly selected individuals was used as input to examine the effect of varying sample sizes, and ancestry models, on the accuracy of STRUCTURE in identifying migrants.

MATERIALS AND METHODS

Generation of genotype data

Genotypic data was generated for each model of temporal dispersal pattern (normal or negative binomial distribution) using Nemo as the simulation platform. The following parameters were implemented for each model: each model consisted of two patches (populations), with 4000 individuals per patch, consisting of equal numbers of males and females. Genotype data was generated for 20 neutral markers, with 30 alleles per locus, a mutation rate of 5×10^{-4} characteristic of hypervariable markers (Ellegren 2000), a recombination rate of 0.5 corresponding to unlinked loci, and maximum values of initial variance. Dispersal rates between the 2 patches were assigned using a dispersal matrix, where unidirectional source to sink migration rate per generation varied according to either a normal or negative binomial distribution. To ensure that the pattern of number of migrants per generation varied according to a fixed mean (i.e. 5 individuals), a list of 4000 integers were drawn from a normal and negative binomial distribution with mean = 5. For the latter distribution, the maximum number of migrants in a generation was restricted to a third of the population size. Mating within a population was random, and all offspring had an equal probability of survival, i.e. no fitness or selection differentials. Population regulation of adult individuals was also random. Genotype data was generated for each dispersal model by running the simulations for 4000 generations, where the dispersal probability for each generation corresponded with a value from the discrete distribution divided by the population size. For each model, 20 replicates were generated to simulate different population scenarios with the same mean migration rate. To ensure that the temporal dispersal rates indeed followed the specified pattern, the number of migrants per generation were obtained for each replicate. The histograms indicate that the two sets of replicates follow the desired probability distributions (data not shown).

To simulate the effect of sampling different collections of individuals, datasets were generated with varying sample sizes for each model. For each model (normal and negative binomal distribution, respectively), 10 datasets from each of the 20 replicates were created. Each dataset consists of 6 different sample sizes: 20, 30, 50, 100, 150, and 200 randomly selected individuals. A total of 200 different sample size datasets were generated for each of the two models.

Assignment of individuals and inferring migrants using STRUCTURE

The STRUCTURE software uses a Bayesian approach to infer population structure by clustering individuals probabilistically into subpopulations (K) where linkage disequilibrium is minimized (Pritchard et al 2000). For each dataset, population structure was examined for concordance with the expected number of subpopulations (K=2). This was done by perfoming four replicate STRUCTURE runs for each dataset at each of four hypothesized number of subpopulations (K = 1, 2, 3 and 4). The optimum K was identified based on the highest estimated log probability of the data (ln K values). Replicates with an optimal number of subpopulations other than the expected K=2, and replicates where the probability of membership in each of the two expected clusters was ambiguous ($p = 0.5$, approaching homogeneity) were noted.

Migrants and migrant ancestry were inferred based on estimates of the probability (p), of observing the individual genotype in each subpopulation or cluster (K). For this analysis, two ancestry models were used and compared for their accuracy in inferring migrant individuals. In the first approach, a model of admixture without prior population information was used. An individual was considered a migrant if the probability (p) of its ancestry in a cluster other than where it was sampled was lower than a defined threshold. Two threshold values were used (0.85 and 0.90) to test the effect of stringency on classifying migrants. The lambda parameter (which affects the prior values of the allele frequencies) was estimated (instead of being fixed it to 1), since a smaller value could yield better results when most marker frequencies are either very high or very low (Pritchard et al 2000).

In the second approach, we used admixture with prior population information in the form of the location of sampling (cluster) from which an individual was sampled as the prior. Model parameters used to infer the number of clusters were retained, namely the assumption of independent allele frequencies and estimation of lambda. Two values for the probability that an individual is a migrant or has migrant ancestry were used ('MIGPRIOR', v): 0.01 and 0.05. This model probabilistically determines if an individual is a migrant or whether it has migrant ancestry in its predefined group. The model was set to detect ancestry for up to two generations, i.e. if an individual is a migrant, has a migrant parent, or a migrant grandparent.

For each dataset, inferred migrants were counted from the replicate with the highest estimated log probability (ln K), and the mean counts for the datasets for each replicate were obtained.

Comparison of actual and inferred number of migrants

For each dataset, the actual ancestry of each individual going back one generation (parents) was determined by tracing its record in the original Nemo-simulated genotype data file. An individual was identified as a migrant if the population of origin is different from the population where it was 'sampled'. Conversely, an individual had a migrant parent if its record indicated that its parents came from different patches, since in this case there are only two patches for each simulated model.

To determine the accuracy of STRUCTURE in inferring migrants, the following categories comparing actual and inferred ancestry were examined: (1) an actual migrant was detected, i.e. actual and inferred ancestry were concordant (true positive); (2) a non-migrant was detected as a migrant (false positive); and (3) an actual migrant was not detected (false negative). Moreover, to see the effect of migrant ancestry on migrant inference, we also noted the instances when an individual with a migrant parent was inferred as a migrant.

RESULTS

Temporal Migration Pattern: Normal and Negative Binomal Distribution

To assess the results generated by STRUCTURE, we compared the frequency (out of 200 runs) of observing a category of result (where each category corresponds to a combination of different values of parameters under study) against those of other categories for each sample size. The parameters included the two temporal migration patterns (normal and negative binomial), two modes of migrant inference (with or without prior information) and the number of migrants in the sample based on the NEMO simulation and on the STRUCTURE results. The frequencies observed for each sample size and category are shown in Table 1.

We first examined whether the two probability distributions generated similar numbers of migrants for comparisons of the classification performance of STRUCTURE, by evaluating the total number of samples with at least one migrant individual. Data simulated from the two probability distributions generated similar numbers of samples with migrants for varying sample sizes, with the exception of the two largest sample sizes (150 and 200), where datasets following a normal distribution generated approximately 10 more samples on average, compared to those following a negative binomial distribution (Table 1). The disparity can be attributed to the fact that as sample size increases, the proportion of samples with and without a migrant more closely approximate the actual proportion in the total population of 4000 individuals, over the 200 replicate datasets. For a negative binomial distribution, the number of populations with or without a migrant would expectedly be a smaller proportion of the total number of samples, but exhibiting greater skew, i.e. more extreme values of number of migrants. Examination of the population genotypic data reveals that of the 20 replicates in the negative binomial populations, only 6 replicates contain migrants in the sampled generation, in contrast to the normal distribution populations where all 20 replicates included at least one migrant. To account for the effect of this difference on subsequent comparisons, analyses were based on the proportion of the classification counts with respect to the total number of samples with or without migrants instead of the raw, absolute counts (Fig. 1, 2).

Majority of the datasets sampled contained no migrants. Datasets following a negative binomial distribution exhibited a greater number of migrants (ranging from 0 to 6) relative to datasets from normal distribution patterns (range = 0 to 2). Only negative binomal distributions generated sample datasets with more than two migrants, and only at larger sample sizes (at least 100 individuals for 3 migrants, and 150 individuals for migrants; Table 1). Overall, the proportion of samples with migrants accurately identified by STRUCTURE was slightly higher for the negative binomial distributions at smaller sample sizes of 20 – 50 individuals compared to the normal distribution (Fig. 1), for both migrant inference approaches (with or without priors). No significant differences were observed in assignment accuracy of STRUCTURE between normal and negative binomal distributions at larger sample sizes (at least 100 individuals). These general observations were true for both models of migrant inference, whether with or without prior population information.

Assignment accuracy with and without prior population information

Majority of the sampled datasets contained zero migrants (Table 1), across all sample sizes. With an increase in sample size, the number of migrant-less samples decreased for both normal and negative binomial distributions, e.g. from 96.5% and 96% (of the 200 datasets), respectively for a sample size of 20 individuals, to 81.5% and 88% respectively, sampling 200 individuals (data not shown). Most of these samples were correctly inferred by STRUCTURE, with or without prior population information, to have no migrant (Table 1). At larger sample sizes (150-200 individuals), including prior population information resulted in an increased number of correct classifications (5-10% more) than the model without prior information. This suggests that improvement in accuracy when using prior population information is more apparent at larger sample sizes. At smaller sample sizes (20-50 individuals) only marginal differences were observed.

In samples with one migrant individual, incorporating prior population information into the assignment model slightly

increased the detection of true migrants at smaller sample sizes (20-100 individuals), for both normal and negative binomial distributions. At larger sample sizes (150-200 individuals), incorporating prior information yielded the same proportion of correct classifications as the model without priors. Results generated from using more relaxed parameter values, e.g. probability of cluster ancestry, p (without prior information) and probability v (with prior information), that an individual is an immigrant, yielded either a greater or equal number of migrants from sets with one true migrant in contrast to sets with 2-6 migrants wherein the number of migrants detected were equal for both parameter values at all sample sizes. This implies that at this migration rate, classification is easier for samples having more than one migrant. In the samples with two migrant individuals, no increase in the proportion of true positives (accurate detection) was observed when using prior population information.

The relative number of false positive classifications was used to measure classification specificity. At smaller sample sizes (20-50 individuals), there were instances where a greater number of false positives were observed for samples with prior population information (Fig. 2), for both normal and negative binomial distributions. At sample sizes above 100 individuals, all classifications using prior population information yielded considerably less numbers of false positives. When using prior population information, only a single individual was misclassified in a sample; without prior information, as many as two individuals were misclassified in a sample (data not shown).

To demonstrate the effect of migrant ancestry on classification specificity, false positive classifications were decomposed into samples with falsely classified non-immigrants with migrant and non-migrant parents. Predictably, since there were very few individuals with migrant parents, there were more misclassified individuals with non-migrant parents. However, the differences in number of misclassified individuals using the more stringent and less stringent parameters values were in general, significantly lower in individuals with migrant parents compared to those with non-migrant parents, indicating that the former tend to have higher values of probability of ancestry in their parents' ancestral cluster.

In both types of misclassification (false positives and false negatives), only samples following a negative binomial migration pattern showed a consistent nondecreasing trend as sample size increased and only when analyzed without prior population information (Fig. 2). When using prior population information, the number of misclassified individuals plateaued at around a sample size of 50 individuals, suggesting that the model's ability to improve specificity takes effect at the specified minimum sample size. Two misclassified individuals and those with both correct and incorrect classifications were found only in samples analyzed without prior population information, except for two instances where the misclassified individuals had migrant parents, which further supports the observation that incorporating prior population information improves assignment accuracy.

Sample Size

Increasing sample size resulted in an increase in the proportion of samples (out of the total number of 200 simulated datasets) where migrants were identified, for both true or false positive classifications (Fig. 1 & 2, respectively). This was true for both sets with normal and negative binomial temporal migration pattern.

For false positive classifications, the proportion of samples identified without incorporating prior information increased with sample size, and was more pronounced for negative binomial distributions (Fig. 2). Incorporating prior information considerably reduced classification errors, i.e. the number of false positives at larger population sizes.

For true positive classifications, the number of samples increased with sample size as well (Fig. 1). Classification approximated the actual proportion (approached 100%) at a minimum population size of 100 individuals, for both normal and negative binomial migration patterns.

DISCUSSION

The mean migration rates and other simulation parameters used in the study generated populations that are highly differentiated (F_{ST} = 0.01 to 0.04, $p < 0.05$). At the migration rate used in this study (mean of 5 migrants per generation), results indicate that (1) the temporal pattern of migration could affect the performance of STRUCTURE in detecting migrants; (2) increasing sampling effort (sampling size or number of individuals), while increasing true positive classifications, also leads to reduced accuracy with increased false positive identifications; and (3) accuracy in migrant identification is significantly improved by incorporating prior population information into assignment methods.

Many studies have previously analyzed the accuracy of STRUCTURE in assigning individuals into the correct cluster using both empirical and simulated data under varying conditions. These studies have focused on examining the effects of various factors on recovering population structure (K) and assignment accuracy, such as the nature of markers, number and alleles, sample size, extent of genetic differentiation (e.g., Bernatchez and Duchesne 2000, Evanno et al 2005, Berry et al 2004, Hauser et al 2006, Latch et al 2006), according to various parameters optimal for handling such data (e.g., confidence thresholds, prior immigrant probability values, baseline data, among others). In this study, we simulated population genetic data with varying migration among generations, following either a normal or negative binomial distribution, over 4000 generations. Both distributions however have a single mean overall migration rate,

to examine whether a population's migration history has any effect on the accuracy of STRUCTURE in identifying population migrants, i.e. its ability to infer the sampled generation's dispersal rate. Moreover, we looked at the effect of sample size as well as the use of baseline data and levels of stringency thresholds on STRUCTURE's accuracy in assigning individuals to the correct source population.

To facilitate comparisons, we classified STRUCTURE results into categories wherein migrants were identified (true positives,

Table 1. Frequency (and percent) of runs observed for each sample size and for each result category, where each category represents a combination of values of the different migrant counts in the sample: actual number of migrants in the drawn sample of size N based on the NEMO simulation (NM), number of individuals in the sample inferred as migrant based on analysis of STRUCTURE outputs (ST), number of true migrants in ST (i.e., true positives) (TM), number of individuals in ST which are not real migrants but with a parent that is a migrant (MP), number of individuals in ST that are wrongly inferred a migrants (NP), and number of individuals in the sample that are migrants but were not inferred as migrants (i.e., false negatives) (ND). A total of 200 runs were run per sample size. For rows where STRUCTURE accurately estimated the actual number of migrants, the values are shown in bold font.

A. Temporal pattern of migration rate follows a normal distribution.

	Number of migrants						Sample Size, *N*											
Method	NM	ST	TM	MP	NP	ND	20		30		50		100		150		200	
method1, p = 0.85	**0**	**0**	**0**	**0**	**0**	**0**	**185**	**(92.5)**	**188**	**(94.0)**	**185**	**(92.5)**	**167**	**(83.5)**	**151**	**(75.5)**	**146**	**(73.0)**
	0	1	0	0	1	0	7	(3.5)	3	(1.5)	2	(1.0)	6	(3.0)	11	(5.5)	10	(5.0)
	0	1	0	1	0	0	0	(0.0)	0	(0.0)	1	(0.5)	6	(3.0)	7	(3.5)	7	(3.5)
	0	2	0	0	2	0	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)
	0	2	0	1	1	0	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)
	1	0	0	0	0	1	4	(2.0)	4	(2.0)	3	(1.5)	0	(0.0)	0	(0.0)	0	(0.0)
	1	**1**	**1**	**0**	**0**	**0**	**3**	**(1.5)**	**4**	**(2.0)**	**8**	**(4.0)**	**15**	**(7.5)**	**22**	**(11.0)**	**27**	**(13.5)**
	1	2	1	0	1	0	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	0	(0.0)	0	(0.0)
	1	2	1	1	0	0	0	(0.0)	0	(0.0)	0	(0.0)	2	(1.0)	3	(1.5)	3	(1.5)
	2	1	1	0	0	1	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	2	**2**	**2**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**2**	**(1.0)**	**5**	**(2.5)**	**7**	**(3.5)**
method1, p = 0.90	**0**	**0**	**0**	**0**	**0**	**0**	**192**	**(96.0)**	**191**	**(95.5)**	**188**	**(94.0)**	**174**	**(87.0)**	**159**	**(79.5)**	**152**	**(76.0)**
	0	1	0	0	1	0	1	(0.5)	0	(0.0)	0	(0.0)	3	(1.5)	6	(3.0)	7	(3.5)
	0	1	0	1	0	0	0	(0.0)	0	(0.0)	0	(0.0)	2	(1.0)	5	(2.5)	4	(2.0)
	0	2	0	1	1	0	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	0	(0.0)	0	(0.0)
	1	0	0	0	0	1	5	(2.5)	5	(2.5)	4	(2.0)	1	(0.5)	0	(0.0)	0	(0.0)
	1	**1**	**1**	**0**	**0**	**0**	**2**	**(1.0)**	**3**	**(1.5)**	**7**	**(3.5)**	**14**	**(7.0)**	**23**	**(11.5)**	**27**	**(13.5)**
	1	2	1	0	1	0	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	0	(0.0)	0	(0.0)
	1	2	1	1	0	0	0	(0.0)	0	(0.0)	0	(0.0)	2	(1.0)	2	(1.0)	3	(1.5)
	2	1	1	0	0	1	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	2	**2**	**2**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**2**	**(1.0)**	**5**	**(2.5)**	**7**	**(3.5)**
method2, prior = 0.01	**0**	**0**	**0**	**0**	**0**	**0**	**190**	**(95.0)**	**190**	**(95.0)**	**187**	**(93.5)**	**180**	**(90.0)**	**168**	**(84.0)**	**162**	**(81.0)**
	0	1	0	0	1	0	3	(1.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)
	0	1	0	1	0	0	0	(0.0)	0	(0.0)	1	(0.5)	0	(0.0)	2	(1.0)	1	(0.5)
	1	0	0	0	0	1	3	(1.5)	3	(1.5)	2	(1.0)	2	(1.0)	1	(0.5)	4	(2.0)
	1	1	0	1	0	1	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	0	(0.0)
	1	**1**	**1**	**0**	**0**	**0**	**4**	**(2.0)**	**5**	**(2.5)**	**9**	**(4.5)**	**16**	**(8.0)**	**23**	**(11.5)**	**26**	**(13.0)**
	2	1	1	0	0	1	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	2	**2**	**2**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**2**	**(1.0)**	**5**	**(2.5)**	**7**	**(3.5)**
method2, prior = 0.05	**0**	**0**	**0**	**0**	**0**	**0**	**183**	**(91.5)**	**184**	**(92.0)**	**185**	**(92.5)**	**179**	**(89.5)**	**168**	**(84.0)**	**159**	**(79.5)**
	0	1	0	0	1	0	9	(4.5)	6	(3.0)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	0	1	0	1	0	0	1	(0.5)	1	(0.5)	2	(1.0)	1	(0.5)	2	(1.0)	4	(2.0)
	1	0	0	0	0	1	3	(1.5)	3	(1.5)	2	(1.0)	1	(0.5)	1	(0.5)	2	(1.0)
	1	1	0	1	0	1	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)
	1	**1**	**1**	**0**	**0**	**0**	**4**	**(2.0)**	**5**	**(2.5)**	**9**	**(4.5)**	**16**	**(8.0)**	**23**	**(11.5)**	**27**	**(13.5)**
	1	2	1	1	0	0	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)
	2	1	1	0	0	1	0	(0.0)	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	2	**2**	**2**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**2**	**(1.0)**	**5**	**(2.5)**	**7**	**(3.5)**

false positives) or were not identified (false negatives). Due to the low mean migration rate (5 out of 4000 individuals), most samples of the total 200 datasets for each migration model contained no migrants (81.5% to 96.5% of the total datasets; Table 1). The number of zero-migrant samples decreased with increasing sampling effort (number of individuals). Assignment accuracy without incorporating prior information generally decreased with increasing sampling size, i.e. more individuals were identified as migrants (false positives) as more individuals are sampled, for both temporal migration patterns (Fig. 2). Incorporation of prior population information results in more conservative identification of migrants, reducing the proportion of false positives at larger sample sizes relative to inferences without priors.

STRUCTURE's accuracy in classifying migrants was generally similar for samples following normal and negative binomial distribution patterns, with the exception of slightly greater sensitivity observed from the latter sets at lower sample sizes (Fig. 1). Most of the correctly classified samples from the

Table 1. (continued.)

B. Temporal pattern of migration rate follows a negative binomial distribution.

Method	Number of migrants NM	ST	TM	MP	NP	ND	Sample Size 20		30		50		100		150		200	
method1, p = 0.85	0	0	0	0	0	0	188	(94.0)	188	(94.0)	182	(91.0)	168	(84.0)	162	(81.0)	154	(77.0)
	0	1	0	0	1	0	3	(1.5)	1	(0.5)	5	(2.5)	9	(4.5)	11	(5.5)	15	(7.5)
	0	1	0	1	0	0	0	(0.0)	1	(0.5)	2	(1.0)	3	(1.5)	4	(2.0)	5	(2.5)
	0	2	0	0	2	0	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)
	0	2	0	2	0	0	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)
	1	0	0	0	0	1	1	(0.5)	1	(0.5)	1	(0.5)	0	(0.0)	0	(0.0)	0	(0.0)
	1	**1**	**1**	**0**	**0**	**0**	**6**	**(3.0)**	**7**	**(3.5)**	**8**	**(4.0)**	**12**	**(6.0)**	**12**	**(6.0)**	**12**	**(6.0)**
	2	**2**	**2**	**0**	**0**	**0**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**6**	**(3.0)**	**8**	**(4.0)**	**7**	**(3.5)**
	3	**3**	**3**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**
	4	**4**	**4**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**2**	**(1.0)**
	6	**6**	**6**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**2**	**(1.0)**
method1, p = 0.90	**0**	**0**	**0**	**0**	**0**	**0**	**192**	**(96.0)**	**188**	**(94.0)**	**185**	**(92.5)**	**173**	**(86.5)**	**166**	**(83.0)**	**160**	**(80.0)**
	0	1	0	0	1	0	0	(0.0)	2	(1.0)	2	(1.0)	5	(2.5)	8	(4.0)	11	(5.5)
	0	1	0	1	0	0	0	(0.0)	1	(0.5)	2	(1.0)	2	(1.0)	3	(1.5)	5	(2.5)
	1	0	0	0	0	1	3	(1.5)	2	(1.0)	1	(0.5)	0	(0.0)	0	(0.0)	1	(0.5)
	1	**1**	**1**	**0**	**0**	**0**	**4**	**(2.0)**	**6**	**(3.0)**	**8**	**(4.0)**	**12**	**(6.0)**	**12**	**(6.0)**	**11**	**(5.5)**
	2	**2**	**2**	**0**	**0**	**0**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**6**	**(3.0)**	**8**	**(4.0)**	**7**	**(3.5)**
	3	**3**	**3**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**
	4	**4**	**4**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**2**	**(1.0)**
	6	**6**	**6**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**2**	**(1.0)**
method2, prior = 0.01	**0**	**0**	**0**	**0**	**0**	**0**	**190**	**(95.0)**	**189**	**(94.5)**	**186**	**(93.0)**	**177**	**(88.5)**	**173**	**(86.5)**	**173**	**(86.5)**
	0	1	0	0	1	0	2	(1.0)	1	(0.5)	1	(0.5)	2	(1.0)	3	(1.5)	2	(1.0)
	0	1	0	1	0	0	0	(0.0)	1	(0.5)	2	(1.0)	1	(0.5)	1	(0.5)	1	(0.5)
	1	0	0	0	0	1	2	(1.0)	2	(1.0)	2	(1.0)	2	(1.0)	1	(0.5)	1	(0.5)
	1	**1**	**1**	**0**	**0**	**0**	**5**	**(2.5)**	**6**	**(3.0)**	**7**	**(3.5)**	**10**	**(5.0)**	**11**	**(5.5)**	**11**	**(5.5)**
	2	1	1	0	0	1	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	1	(0.5)	1	(0.5)
	2	**2**	**2**	**0**	**0**	**0**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**5**	**(2.5)**	**7**	**(3.5)**	**6**	**(3.0)**
	3	**3**	**3**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**
	4	**4**	**4**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**2**	**(1.0)**
	6	**6**	**6**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**2**	**(1.0)**
method2, prior = 0.05	**0**	**0**	**0**	**0**	**0**	**0**	**188**	**(94.0)**	**186**	**(93.0)**	**181**	**(90.5)**	**173**	**(86.5)**	**167**	**(83.5)**	**169**	**(84.5)**
	0	1	0	0	1	0	4	(2.0)	4	(2.0)	6	(3.0)	6	(3.0)	7	(3.5)	5	(2.5)
	0	1	0	1	0	0	0	(0.0)	1	(0.5)	2	(1.0)	1	(0.5)	3	(1.5)	2	(1.0)
	1	0	0	0	0	1	2	(1.0)	1	(0.5)	1	(0.5)	0	(0.0)	1	(0.5)	1	(0.5)
	1	**1**	**1**	**0**	**0**	**0**	**5**	**(2.5)**	**7**	**(3.5)**	**8**	**(4.0)**	**12**	**(6.0)**	**11**	**(5.5)**	**11**	**(5.5)**
	2	1	1	0	0	1	0	(0.0)	0	(0.0)	0	(0.0)	1	(0.5)	1	(0.5)	1	(0.5)
	2	**2**	**2**	**0**	**0**	**0**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**5**	**(2.5)**	**7**	**(3.5)**	**6**	**(3.0)**
	3	**3**	**3**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**	**1**	**(0.5)**
	4	**4**	**4**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**1**	**(0.5)**	**2**	**(1.0)**
	6	**6**	**6**	**0**	**0**	**0**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**0**	**(0.0)**	**1**	**(0.5)**	**2**	**(1.0)**

negative binomial set were drawn from a single replicate population with an unusually high number of migrants (68 individuals) in the original population. Despite the high number of migrants in these samples (1-2 migrants among 20 individuals), patches exhibited high levels of differentiation based on F_{ST} values (F_{ST} = 0.04 to 0.05, relative to F_{ST} = 0.02 to 0.03 for other samples). Greater genetic differentiation in negative binomial distributions expectedly increases assignment accuracy as demonstrated in earlier studies (Berry et al 2004, Latch et al 2006). At higher sample sizes, migrants were detected in most samples from both distributions, consistent with findings by Yang et al (2005) where assignment accuracy using STRUCTURE increased from sample size 10 to 100 individuals, where accuracy approached 100% for 100 replicates. However, the proportion of false positive classifications likewise generally increased with sample size, indicating that sensitivity to migrant detection (identifying the number of migrants) increases with sample size at the expense of specificity (identifying the correct migrant individuals, true positives only), possibly because at greater sample sizes the probability of sampling individuals with recent migrant ancestry proportionally increases.

We compared STRUCTURE's accuracy when using the model with and without prior population information. The former model incorporates prior data by assigning all or some individuals membership to a designated population, such as the location at time of sampling. The model postulates that a proportion of an individual's genome may have originated from a population different from the group from which it was sampled either through dispersal (whole genome) or through a migrant ancestor (partial genome), under the assumption that the probability of migrant ancestry is proportion to $2^{-t}v$, where t is the generation and v is the probability of migration (user defined). By setting v at a higher or lower value, one can control the stringency at which an individual is classified as a migrant. Using prior population information generally resulted in higher accuracy in classifying migrants, both in terms of sensitivity and specificity (Fig. 1 and 2). In terms of sensitivity, using prior information detected significantly more migrants at lower sample sizes (20 to 50 individuals), even when results using the lower migration probability value (v = 0.01) were compared with the results using the less stringent threshold probability (p = 0.85), when analyzed without prior population information, ranging from 4 to 10% (Fig. 1). However, from size 100 to 200 individuals, the model with prior population information becomes more conservative at classifying migrants and missed 1-4 true migrants, whereas the model without priors detected all migrants. Incorporating prior population information however, yields reduced false positive classifications. These results give a more optimistic assessment of STRUCTURE's assignment sensitivity without baseline data, compared to Hauser and colleagues' (2006) study wherein STRUCTURE classification between wild and hatchery-bred populations without baseline data was unable to assign any sample to the latter with a p = 0.95 threshold probability. These results however, may be due to the less stringent threshold used in this study, at p of 0.85 and 0.90, respectively.

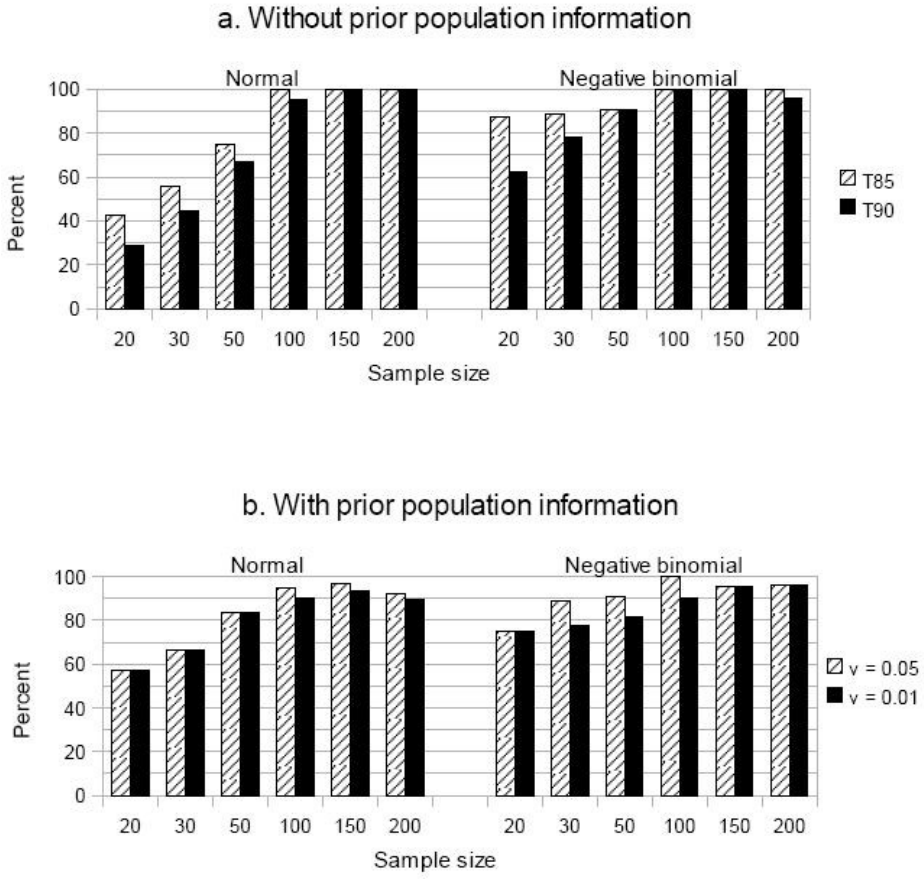

Figure 1. Percent of samples correctly classified to have a migrant/s by STRUCTURE (over the total number of samples with actual migrants), without (a) and with (b) prior population information and following a normal and negative binomial distribution and using a two threshold confidence values, 0.85 and 0.90 (a) and two probability of immigration values, 0.01 and 0.05 (b).

The difference in classification accuracy, when using the model with prior population information, between a higher and lower assumed probability of migration, emphasizes the utility of testing a range of migration probability values. Threshold levels control classification sensitivity and specificity, and increasing one generally leads to a reduction in the other. This was the case in most of the samples wherein a lower classification threshold (lower probability of migration and also, with the model without prior population information, lower threshold confidence), resulted in higher classification sensitivity but at the same time, a proportionally lower specificity (Fig. 1, 2). However, in the set of samples following a normal distribution, a higher migration probability resulted in none or little increase in the number of true migrants detected but generated significantly more false positives (Fig. 1b, 2b). Consequently, although increasing the assumed probability of migration increased the number of individuals classified as migrants, most if not all of these were false detections, even as a significant proportion of the real migrants remain undetected. The *v* values used here (0.01 and 0.05) were based on the suggested range of 0.001 to 0.1 (Pritchard et al 2000) and the known mean migration rate of 5 migrants per 4000 individuals. These results show that at low sample sizes (20-30), even setting the migration probability at a much higher value than expected could still significantly underestimate actual migrant numbers by as much as half.

The initial analyses presented here shows that while extreme deviations in migration history may affect the accuracy of migrant inference, STRUCTURE generally exhibits similar levels of accuracy in migrant classification from populations with different temporal dispersal patterns. Analysis with additional data at varying levels of population genetic differentiation may reveal a more informative picture of how temporal variations affect the estimation of migrants, and consequently migration rates. Since this study was limited to a single mean migration rate, future analysis employing a greater range of migration rates, particularly higher migration rates characteristic of high gene flow scenarios, would allow more in-depth examination of assignment accuracy across varying sample sizes and assignment method models. Nonetheless, general inferences from this initial study can be used to guide aspects of sampling design (sample size thresholds), and genetic analysis (parameters for STRUCTURE analysis), and the probable effects of such on the reliability and accuracy of results generated from STRUCTURE analyses. Estimates of the levels of accuracy, and likewise potential error rates, are particularly important for population genetic studies aimed at inferring patterns of connectivity and dispersal. These are applicable to wild populations in general, and are particularly relevant for marine populations, where the fluid environment drives complex dispersal patterns, and typically large populations presents special challenges to

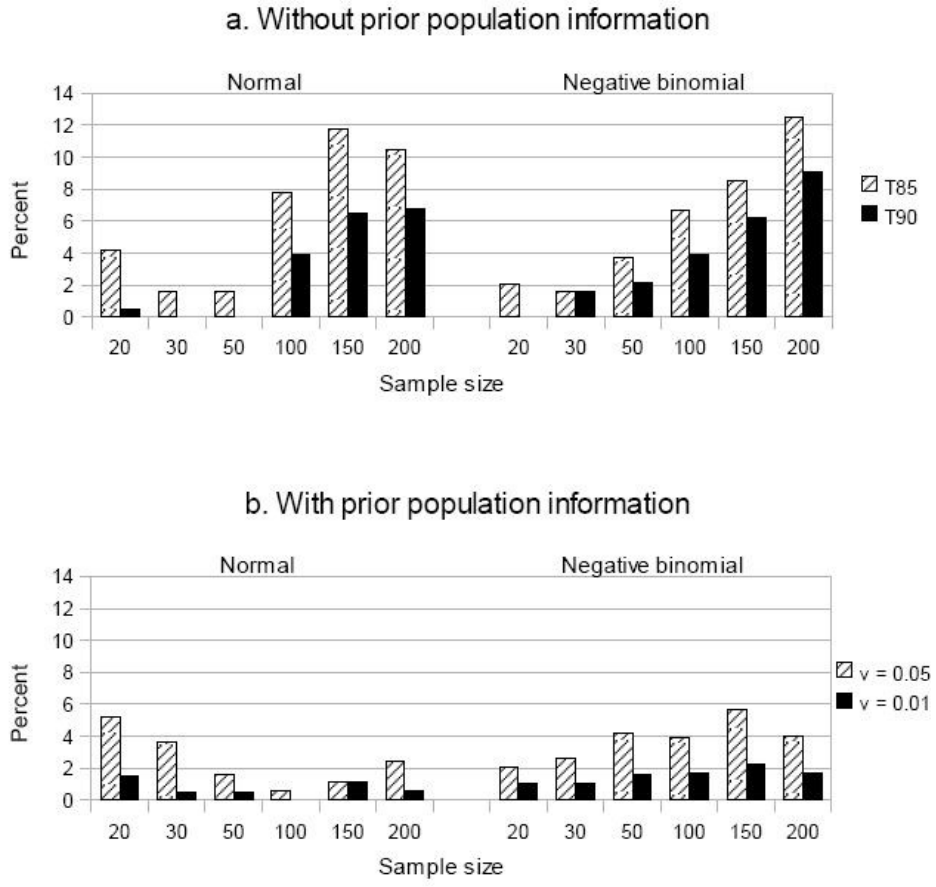

Figure 2. Percent of samples incorrectly classified to have a migrant/s by STRUCTURE (over the total number of samples with inferred migrants), without (a) and with (b) prior population information and following a normal and negative binomial distribution and using two threshold confidence values, 0.85 and 0.90 (a) and two probability of immigration values, 0.01 and 0.05 (b).

representative sampling.

ACKNOWLEDGEMENTS

This study was made possible by a research grant from the University of the Philippines - Office of the Vice President for Academic Affairs to AOL and partial support from the Marine Science Institute.

CONFLICTS OF INTEREST

The authors declare no conlict of interest.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

MHM carried out the simulations and analyses and prepared the draft of the manuscript. RRG performed some of the analyses and refined the discussion, and helped prepare the final draft. AOL conceptualized the project and prototyped the simulations and analyses, wrote some of the Python scripts to generate simulated genotypic data and process the results, and helped in the analysis of data and preparation of the manuscript.

REFERENCE

Avise JC. Molecular markers, natural history, and evolution. New York: Chapman and Hall, 1994.

Bernatchez L, Duchesne P. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? Can J Fish Aquat Sci 2000; 57:1-12.

Berry O, Tocher MD, Sarre SD. Can assignment tests measure dispersal. Mol Ecol 2004; 13:551-561.

Bossart JL, Prowell DP. Genetic estimates of population structure and gene flow: limitations, lessons and new directions. Trends Ecol Evol 1998; 13:202-206.

Ellegren H. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 2000; 16:551-558.

Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 2005; 14:2611-2620.

Guillaume F, Rougemont J. Nemo: an evolutionary and population geneticsprogramming framework. Bioinformatics Applications Note 2006; 22:2556-2557.

Hauser L, Seamons TR, Dauer M, Naish KA, Quinn TP. An empirical verification of population assignment methods by marking and parentage data: hatchery and wild steelhead (*Onchorhynchus mykiss*) in Forks Creek, Washington, USA. Mol Ecol 2006; 15:3157-3173.

Hedgecock D, Barber PH, Edmands S. Genetic approaches to measuring connectivity. Oceanography 2007; 20:70-79.

Hellberg ME, Burton RS, Neigel JE, Palumbi SR. Genetic assessment of connectivity among marine populations. Bulletin of Marine Science 2002; 70:273-290.

Jones GP, Almany GR, Russ GR, Sale PF, Steneck RS, van Oppen MJH, Willis BL. Larval retention and connectivity among populations of corals and reef fishes: history, advances and challenges. Coral Reefs 2009; 28:307-325.

Kinlan BP, Gaines SD, Lester SE. Propagule dispersal and the scales of marine community process. Diversity and Distributions 2005; 11:139-148.

Kritzer JP, Sale PF. Metapopulation ecology in the sea: from Levins' model to marine ecology and fisheries science. Fish and Fisheries 2004; 5:131-140.

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics 2006; 7:302.

Manel S, Gaggioti OE, Waples RS. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol Evol 2005; 20:136-142.

Paetkau D, Slade R, Burden M, Estoup A. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol Ecol 2004; 13:55-65.

Pearse DE, Crandall K. Beyond F_{ST}: Analysis of population genetic data for conservation. Conservation Genetics 2004; 5:585-602.

Pielou EC. An introduction to mathematical ecology. Wiley-Interscience, New York, New York USA. 1969.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000; 155:945-959.

Sale PF, Cowen RK, Danilowicz BS, Jones GP, Kritzer JP, Lindeman KC, Planes S, Polunin NV, Russ GR, Sadovy YJ, Steneck RS. Critical science gaps impede use of no-take fishery reserves. Trends Ecol Evol 2005; 20:74-80.

Strathmann RR, Hughes TP, Kuris AM, Lindeman KC, Morgan SG, Pandolfi JM, Warner RR. Evolution of local recruitment and its consequences for marine populations. Bulletin of Marine Science 2002; 70:377-396.

Taylor LR, Woiwod IP, Perry JN. The negative binomial as a dynamic ecological model for aggregation, and the density dependence of *k*. J Animal Ecol 1979; 48:289-304.

Waples RS, Gaggiotti O. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol 2006; 15:1419-1439.

Waser PM, Strobeck C. Genetic signatures of interpopulation dispersal. Trends Ecol Evol 1998; 13:43-44.

Whitlock MC, McCauley DE. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity 1999; 82:117-125.

Wright S. Evolution in Mendelian populations. Genetics 1931; 16:97-159.

Yang B, Zhao H, Kranzler HR, Gelernter J. Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. Genetic Epidemiology 2005; 28:302-312.