Conservation Genetics (2006) 7:295–302
DOI 10.1007/s10592-005-9098-1


# Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation

Emily K. Latch[1,*], Guha Dharmarajan[1], Jeffrey C. Glaubitz[2] & Olin E. Rhodes Jr.[1]
[1]*Department of Forestry and Natural Resources, Purdue University, West Lafayette, Indiana, 47907, USA;*
[2]*Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin, 53706, USA*
*(*Corresponding author: Phone: +1-765-496-6868; Fax: +1-765-496-2422; E-mail: latche@purdue.edu)*

Received 13 May 2005; accepted 9 November 2005

*Key words:* assignment, Bayesian, $F_{ST}$, microsatellite, population structure

**Abstract**

Traditional methods for characterizing genetic differentiation among populations rely on *a priori* grouping of individuals. Bayesian clustering methods avoid this limitation by using linkage and Hardy–Weinberg disequilibrium to decompose a sample of individuals into genetically distinct groups. There are several software programs available for Bayesian clustering analyses, all of which describe a decrease in the ability to detect distinct clusters as levels of genetic differentiation among populations decrease. However, no study has yet compared the performance of such methods at low levels of population differentiation, which may be common in species where populations have experienced recent separation or high levels of gene flow. We used simulated data to evaluate the performance of three Bayesian clustering software programs, PARTITION, STRUCTURE, and BAPS, at levels of population differentiation below $F_{ST} = 0.1$. PARTITION was unable to correctly identify the number of subpopulations until levels of $F_{ST}$ reached around 0.09. Both STRUCTURE and BAPS performed very well at low levels of population differentiation, and were able to correctly identify the number of subpopulations at $F_{ST}$ around 0.03. The average proportion of an individual's genome assigned to its true population of origin increased with increasing $F_{ST}$ for both programs, reaching over 92% at an $F_{ST}$ of 0.05. The average number of misassignments (assignments to the incorrect subpopulation) continued to decrease as $F_{ST}$ increased, and when $F_{ST}$ was 0.05, fewer than 3% of individuals were misassigned using either program. Both STRUCTURE and BAPS worked extremely well for inferring the number of clusters when clusters were not well-differentiated ($F_{ST} = 0.02$–$0.03$), but our results suggest that $F_{ST}$ must be at least 0.05 to reach an assignment accuracy of greater than 97%.

## Introduction

Traditional population genetic analyses, such as *F*-statistics or genetic distances, remain the most common approach for characterizing population differentiation. However, inappropriate *a priori* grouping of individuals into populations may diminish the power of such analyses to elucidate biological processes, potentially leading to unsuitable conservation or management strategies. Clustering methods that aim to decompose a sample of individuals into genetically distinct groups without *a priori* characterization of these groups have been recognized as a powerful means by which populations can be defined prior to undertaking traditional population genetic analyses. For example, Bowcock et al. (1994) used tree-based hierarchical clustering of individuals to define clusters of genetically similar human populations for inference of evolutionary relationships.

The types of questions that can be addressed with clustering methods are central to conservation and management of wildlife, and include delineation of population boundaries and detection of cryptic population structure (Paetkau et al. 1995; Kyle and Strobeck 2001), estimation of dispersal rates and patterns (Castric and Bernatchez 2004; Berry et al. 2005), characterization of hybrid individuals (Beaumont et al. 2001; Randi et al. 2001; Latch et al. 2006), and determination of the origin of unknown individuals (Primmer et al. 2000; Manel et al. 2002).

Most of the recent advances in clustering methodology have been made in a Bayesian statistical framework, to allow simultaneous estimation of many interdependent parameters in complex models; these methods have proven to be very useful for the field of population genetics (Mank and Avise 2004; Pearse and Crandall 2004; Manel et al. 2005). Three of the most widely used software programs available for Bayesian clustering are PARTITION (Dawson and Belkhir 2001), STRUCTURE (Pritchard et al. 2000; Falush et al. 2003; Pritchard and Wen 2003), and BAPS (Corander et al. 2003, 2004, 2005). All three methods operate by minimizing the Hardy–Weinberg and linkage disequilibrium (HWD and LD) that would result if individuals from different, randomly-mating populations were incorrectly grouped into a common population. PARTITION estimates the number of clusters ($K$) in a sample by employing a Markov chain Monte Carlo (MCMC) method to generate an estimate of the posterior distribution of the sample partition (Dawson and Belkhir 2001). Individuals are assumed to be of pure ancestry, which in the context of this program means that each sampled individual is assigned to a single cluster (Dawson and Belkhir 2001). STRUCTURE employs a somewhat *ad hoc* approach for inferring $K$, by deriving the posterior probability distribution of $K$ from separate MCMC chains, each with a different fixed value of $K$ (Pritchard et al. 2000; Falush et al. 2003). The most recent version of BAPS (version 3.1) uses a greedy stochastic optimization algorithm (Fletcher 1987) to search for the most likely $K$ (Corander et al. 2005). Both STRUCTURE and BAPS allow individuals to be of mixed ancestry, proportionally assigning an individual genome to clusters.

As the level of genetic differentiation among the true subpopulations represented in a sample decreases, so will the HWD and LD. As such, the performance of Bayesian clustering methods, which rely on this disequilibrium, will decrease with decreasing genetic differentiation (Pearse and Crandall 2004; Manel et al. 2005). Low values of $F_{ST}$ ($< 0.1$) are commonly observed in wild populations and can reflect high levels of gene flow between populations or recent divergence. Although appropriate conservation and management strategies depend on our ability to correctly delineate genetically distinct populations, no study to date has addressed the question most recently put forth by Manel et al. (2005): 'How well do Bayesian clustering methods perform when genetic differentiation among populations is modest ($F_{ST} < 0.05$)?' In this study, we evaluate the relative utility of three widely used Bayesian clustering methods (PARTITION, STRUCTURE, and BAPS) for identifying population substructure and assigning individuals to populations, using simulated data.

## Methods

Each of our simulated datasets consisted of a population structured into five subpopulations (finite island model), differentiated at one of ten $F_{ST}$ levels ($F_{ST} = 0.01$–$0.10$). We independently generated 5 datasets at each level of $F_{ST}$. In order to generate a single dataset, we initially created an infinite reference population with a uniform allelic frequency distribution across 10 codominant unlinked loci. A set of 5 subpopulations was created by drawing a random set of founder individuals from the reference population, and the $F_{ST}$ among these subpopulations was calculated (Nei 1977). This process was repeated by iteratively modifying the number of founders drawn until the target $F_{ST}$ (within a precision of 0.001) was reached. Once a set of subpopulations was obtained with the target $F_{ST}$ value, 500 multilocus genotypes (100 per subpopulation) were randomly drawn from the subpopulation allele frequency distributions to form a single dataset. This process mimicked an instantaneous expansion of each subpopulation to a very large size with no mutation, such that the allelic frequency distributions of the subpopulations, and thus $F_{ST}$ among subpopulations, remained constant after the initial founding event. We used the software CONVERT (version 1.2;

Glaubitz 2004) to ease the conversion of files from text format to those used by PARTITION, STRUCTURE, and BAPS. Not all levels of $F_{ST}$ were tested using all software programs: the lower limit of $F_{ST}$ used for a given software was that at which $K$ could no longer be correctly inferred as five; the upper limit of $F_{ST}$ was that at which 95% of all individuals were correctly assigned to their subpopulation of origin. At each value of $F_{ST}$, we analyzed five independently-generated, replicate datasets to evaluate the consistency of the results.

We performed our analysis in PARTITION using version 2 of the software (Dawson and Belkhir 2001). We set the maximum number of possible source populations at 10 to provide a sufficiently broad prior. We set the priors theta (prior distribution of allelic diversity within the source population) and $\mu$ (prior distribution of $K$) to 1.0. The estimates of posterior probabilities were made on the basis of 50,000 iterations of the Markov chain, yielding (after thinning) 10,000 observations. We implemented a burn-in of 1000 observations and computed posterior co-assignment probabilities for pairs of individuals (dimension = 2) based only on observations from the Markov chain where $K$ was five. Each dataset required approximately 30 h to complete (all times provided are appropriate for a computer with a 2.2 GHz Celeron processor and 512 MB of RAM). We also ran longer MCMC chains (25,000 observations) for five datasets with $F_{ST} = 0.08$ to see if this led to improved results. The mode of the posterior distribution of $K$ was taken as a point estimate of $K$, and the Bayes factor was used to evaluate the evidence for $K = 1$ against the alternative of $K > 1$. We defined group membership for each individual based upon visual inspection of the co-assignment probability dendrogram.

With STRUCTURE (version 2.1; Pritchard et al. 2000; Falush et al. 2003; Pritchard and Wen 2003) we performed 50,000 replicates of the MCMC following a burn-in of 10,000 replicates. We used the admixture model, with a uniform prior on the degree of admixture, alpha (initial value = 1.0, max = 10.0, SD = 0.025). We allowed the allele frequencies to be correlated among subpopulations (prior mean = 0.01, prior SD = 0.05, $\lambda = 1.0$). This configuration, using the admixture model and correlated allele frequencies, has been considered best in the case of subtle population structure (Falush et al. 2003). We ran this parameter set for $K$ between one and ten, for each of the five simulated datasets at each value of $F_{ST}$. Each dataset took approximately 3 h to run. Likelihoods provided for each $K$ were transformed into probabilities (Prichard et al. 2000), and the most probable value of $K$ was used as a point estimate. For many cases, we did observe the phenomenon that once the real $K$ is reached, likelihoods for larger $K$s plateau and the variance among runs increases (Pritchard et al. 2000; Pritchard and Wen 2003). Thus, we also used a $\Delta K$ measure that has been proposed to alleviate this problem and provide a better estimate of the true $K$ (Evanno et al. 2005). Admixture proportions ($q$ values) for each individual, based upon MCMC runs where $K$ was set at (and correctly inferred as) five, were used to define group membership; individuals were considered to be correctly assigned if $q > 0.5$ in their true cluster of origin.

With BAPS (version 3.1; Corander et al. 2005) we clustered groups of individuals, setting the maximum number of clusters at ten. Since the mode of the posterior distribution of $K$ almost always provided an overestimate of $K$ (see Results), we used the number of clusters containing more than 3 individuals as a point estimate of $K$, as recommended by Corander et al. (2005). For runs in which $K$ was correctly estimated, we calculated the average probability of assignment to the 'correct' cluster ('correct' defined as $q > 0.5$ in the correct cluster). Individuals with a likelihood admixture ratio of greater than 3.0 were considered to be significantly admixed. Each run took approximately 30 s to complete.

## Results

PARTITION always detected only a single population when $F_{ST}$ was less than 0.07 (Table 1). At $F_{ST} = 0.07$, PARTITION began to consider multiple clusters; however, significant evidence for substructure (Bayes factor $< 1$) was detected in only one of our five datasets (Table 1). At $F_{ST} = 0.08$, PARTITION found significant evidence for multiple clusters, but still was unable to consistently identify the true value of $K$ (Table 1). When the number of observations was increased to 25,000, the probability of $K = 5$ increased, but it was the most likely solution only two of five times (data not shown). Even at $F_{ST} = 0.09$, PARTITION only

*Table 1.* Posterior probability distributions of the number of subpopulations ($K$) obtained from PARTITION and BAPS software for simulated datasets each with five subpopulations differentiated to varying degrees (true value of $K = 5$)

| Program | $F_{ST}$ | Dataset | Prob($K$) | | | | | | | | | | Bayes factor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| PARTITION | 0.06 | 1 | 0.47 | 0.29 | 0.17 | 0.04 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 | 7.9 |
| | | 2 | 0.25 | 0.23 | 0.20 | 0.12 | 0.09 | 0.08 | 0.02 | 0.01 | 0 | 0 | 3.1 |
| | | 3 | 0.34 | 0.25 | 0.16 | 0.14 | 0.09 | 0.01 | 0.01 | 0 | 0 | 0 | 4.6 |
| | | 4 | 0.44 | 0.29 | 0.13 | 0.07 | 0.05 | 0.01 | 0.01 | 0 | 0 | 0 | 7.2 |
| | | 5 | 0.41 | 0.35 | 0.11 | 0.07 | 0.03 | 0.03 | 0.01 | 0 | 0 | 0 | 6.2 |
| | 0.07 | 1 | 0.47 | 0.26 | 0.16 | 0.08 | 0.03 | 0.01 | 0 | 0 | 0 | 0 | 7.9 |
| | | 2 | 0.04 | 0.18 | 0.25 | 0.19 | 0.16 | 0.14 | 0.03 | 0.01 | 0.01 | 0.01 | 0.36 |
| | | 3 | 0.33 | 0.32 | 0.18 | 0.13 | 0.04 | 0.01 | 0 | 0 | 0 | 0 | 4.5 |
| | | 4 | 0.20 | 0.27 | 0.23 | 0.15 | 0.10 | 0.05 | 0.01 | 0 | 0 | 0 | 2.3 |
| | | 5 | 0.16 | 0.18 | 0.26 | 0.19 | 0.15 | 0.01 | 0.01 | 0 | 0 | 0 | 1.7 |
| | 0.08 | 1 | 0.25 | 0.30 | 0.26 | 0.13 | 0.06 | 0.02 | 0 | 0 | 0 | 0 | 2.9 |
| | | 2 | 0.03 | 0.11 | 0.1 | 0.18 | 0.28 | 0.24 | 0.04 | 0.01 | 0 | 0 | 0.26 |
| | | 3 | 0.05 | 0.15 | 0.29 | 0.20 | 0.17 | 0.11 | 0.03 | 0 | 0 | 0 | 0.45 |
| | | 4 | 0.09 | 0.14 | 0.19 | 0.22 | 0.17 | 0.15 | 0.04 | 0.01 | 0 | 0 | 0.88 |
| | | 5 | 0.07 | 0.18 | 0.27 | 0.18 | 0.14 | 0.11 | 0.03 | 0.01 | 0 | 0 | 0.66 |
| | 0.09 | 1 | 0.03 | 0.21 | 0.32 | 0.18 | 0.11 | 0.11 | 0.03 | 0.01 | 0 | 0 | 0.31 |
| | | 2 | 0 | 0.02 | 0.09 | 0.16 | 0.42 | 0.24 | 0.05 | 0.01 | 0 | 0 | 0.00 |
| | | 3 | 0.01 | 0.07 | 0.13 | 0.20 | 0.29 | 0.21 | 0.08 | 0.01 | 0 | 0 | 0.10 |
| | | 4 | 0 | 0.12 | 0.11 | 0.16 | 0.28 | 0.23 | 0.08 | 0.02 | 0.01 | 0 | 0.00 |
| | | 5 | 0 | 0 | 0.08 | 0.23 | 0.41 | 0.23 | 0.04 | 0.01 | 0 | 0 | 0.00 |
| BAPS | 0.02 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | 0.03 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.06 | 0.94 | |
| | | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0.96 | |
| | | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.08 | 0.1 | 0.82 | |
| | | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.05 | 0.95 | |
| | 0.04 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.29 | 0.71 | |
| | | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| | | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.11 | 0.89 | 0 | |
| | | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 | 0.11 | 0.87 | 0 | |
| | 0.05 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.17 | 0.83 | |
| | | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| | | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.40 | 0.60 | 0 | |
| | | 4 | 0 | 0 | 0 | 0 | 0.99 | 0.01 | 0 | 0 | 0 | 0 | |
| | | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.64 | 0.36 | 0 | 0 | |

Five replicate data sets were examined at each level of $F_{ST}$. For STRUCTURE, the probability of the data (L($K$)) for the most likely estimate of $K$ was always near one and thus is not included in the table.

estimated $K$ accurately four of five times (Table 1). Visual examination of the co-assignment probability dendrograms showed that, when PARTITION correctly inferred $K$, all individuals were correctly assigned (data not shown).

STRUCTURE always returned a posterior probability of the data (L($K$)) near one at what it considered to be the most likely value of $K$ (which was not necessarily the true value). STRUCTURE could not detect more than one population at an

$F_{ST}$ of 0.01. At 0.02, STRUCTURE identified multiple clusters in four of the five datasets, but could not discern all five subpopulations. At levels of genetic differentiation at or above $F_{ST} = 0.03$, STRUCTURE correctly estimated $K$ as five for all datasets. When we used $\Delta K$ to infer the number of clusters, we found that $K = 5$ was clearly inferred for all $F_{ST} \geq 0.03$. Below $F_{ST} = 0.03$, $\Delta K$ did not exhibit a clear trend or a clearly identifiable mode, suggesting that there was no clear pattern of genetic structure in these data sets. As expected, the ability of STRUCTURE to correctly assign all individuals to their subpopulation of origin (once $K$ was correctly estimated) increased with $F_{ST}$ (Table 2). At an $F_{ST}$ of 0.05, the average $q$ value reached 0.93, and 2.2% of individuals were misassigned (Table 2).

BAPS consistently overestimated the number of clusters when the mode of the posterior probability distribution of $K$ was used as an estimate of that parameter (Table 1). However, in every case, the vast majority of individuals were assigned to a small number of large clusters. At $F_{ST} = 0.02$, most individuals were grouped into two to four main clusters, indicating an overall inability to correctly delineate subpopulations (data not shown). At levels of differentiation at or above $F_{ST} = 0.03$, as long as the estimate of $K$ was taken as the number of clusters containing more than three individuals, BAPS correctly inferred the true value of $K$ most of the time (Table 2). The average proportion of an individual's genome assigned to its true subpopulation of origin increased with $F_{ST}$, reaching 0.92 at $F_{ST} = 0.05$ (Table 2). Likewise, the average number of misassignments decreased with increasing $F_{ST}$, and when $F_{ST}$ was 0.05, fewer than 3% of individuals were misassigned and less than 4% were considered admixed (Table 2).

## Discussion

PARTITION was unable to correctly identify the number of subpopulations until levels of $F_{ST}$ approached 0.09. Even then, it was not always able to choose the correct $K$. PARTITION's assumption that all individuals are of pure ancestry means that this software may not be appropriate for conservation and management studies where detecting hybridization (admixture) is a concern. However, the dendrogram option provided by PARTITION software offers a unique perspective on assignment of individuals to clusters, and it may sometimes be advantageous to use the dendrogram as a guideline for grouping individuals into clusters rather than relying solely on a single assignment statistic. For example, in a case where it was difficult to assign individuals to discrete populations using traditional statistics, Maingon et al. (2003) found the co-assignment probability dendrogram generated by PARTITION to be a useful alternative. We did not thoroughly examine the co-assignment probability dendrograms for each dataset and thus we may have underestimated the true power of PARTITION to correctly delineate subpopulation structure at low levels of $F_{ST}$.

Although STRUCTURE and BAPS use different methods to search for the most likely number of clusters, both perform well at low levels of population differentiation. The ability of both STRUCTURE and BAPS to differentiate groups whose allele frequency distributions are not extremely different makes them well-suited to many conservation- and management-oriented studies. Each program has unique features that may make it more or less useful for particular applications. For instance, STRUCTURE can incorporate data from linked loci (Falush et al. 2003), while BAPS provides a likelihood ratio as a test statistic to aid the detection of admixed individuals (Corander et al. 2005). The performance of STRUCTURE has recently been evaluated by simulating various dispersal scenarios and seems to perform well with more complex population structure than the finite island model used in this study (hierarchical island model, contact zone model; Evanno et al. 2005).

One potential disadvantage common to STRUCTURE and BAPS exists at the point where the programs begin to break down, around $F_{ST} = 0.02$. At $F_{ST} = 0.02$, neither program correctly identifies the number of subpopulations that exist; however, the probabilities associated with the apparent most likely number of clusters are extremely high. Thus, it seems that these software programs provide false certainty regarding $K$ when $F_{ST}$ is low.

The performance of Bayesian clustering algorithms in detecting population structure depends largely on the properties of the data. As our study was focused on the performance of STRUCTURE, BAPS, and PARTITION relative to one

*Table 2.* Individual assignment data for STRUCTURE and BAPS, obtained for five replicate, simulated datasets at each value of $F_{ST}$

| Software | $F_{ST}$ | Dataset | Most likely # clusters | Avg proportion of genome belonging to 'correct' subpopulation | Avg % misassigned | Avg % admixed |
|---|---|---|---|---|---|---|
| STRUCTURE | 0.03 | 1 | 5 | 0.71 | 20.6 | |
| | | 2 | 5 | 0.77 | 14.8 | |
| | | 3 | 5 | 0.77 | 13.4 | |
| | | 4 | 5 | 0.77 | 12.8 | |
| | | 5 | 5 | 0.80 | 12.0 | |
| | | Avg | | 0.76 | 14.7 | |
| | 0.04 | 1 | 5 | 0.81 | 10.0 | |
| | | 2 | 5 | 0.87 | 4.8 | |
| | | 3 | 5 | 0.89 | 4.8 | |
| | | 4 | 5 | 0.89 | 5.4 | |
| | | 5 | 5 | 0.88 | 5.0 | |
| | | Avg | | 0.87 | 6.0 | |
| | 0.05 | 1 | 5 | 0.90 | 4.8 | |
| | | 2 | 5 | 0.94 | 1.2 | |
| | | 3 | 5 | 0.94 | 1.6 | |
| | | 4 | 5 | 0.94 | 1.2 | |
| | | 5 | 5 | 0.93 | 2.0 | |
| | | Avg | | 0.93 | 2.2 | |
| BAPS | 0.03 | 1 | 10 (4) | N/A | N/A | N/A |
| | | 2 | 10 (5) | 0.76 | 12.6 | 8.4 |
| | | 3 | 10 (4) | N/A | N/A | N/A |
| | | 4 | 10 (5) | 0.80 | 10.0 | 7.4 |
| | | 5 | 10 (5) | 0.83 | 10.8 | 9.2 |
| | | Avg | | 0.80 | 11.1 | 8.3 |
| | 0.04 | 1 | 10 (5) | 0.82 | 10.0 | 10.2 |
| | | 2 | 7 (5) | 0.87 | 5.2 | 4.4 |
| | | 3 | 9 (5) | 0.88 | 4.4 | 3.6 |
| | | 4 | 10 (5) | 0.87 | 5.8 | 6.6 |
| | | 5 | 9 (5) | 0.88 | 4.6 | 5.6 |
| | | Avg | | 0.86 | 6.0 | 6.1 |
| | 0.05 | 1 | 10 (5) | 0.89 | 5.2 | 4.6 |
| | | 2 | 6 (5) | 0.93 | 1.2 | 2.6 |
| | | 3 | 9 (5) | 0.92 | 3.0 | 2.6 |
| | | 4 | 5 (5) | 0.93 | 1.6 | 2.6 |
| | | 5 | 7 (5) | 0.91 | 2.6 | 4.0 |
| | | Avg | | 0.92 | 2.7 | 3.3 |

Assignment data are provided only for those datasets where the software could correctly infer the number of subpopulations ($K$) as five. With BAPS, $K$ was estimated as the number of clusters containing more than three individuals (numbers in parentheses). Values are not provided in the table for PARTITION, because all individuals were correctly assigned once the correct number of clusters was inferred (at $F_{ST} = 0.09$ or greater).

another, we evaluated their performance against the same sets of data. However, empirical datasets likely will vary in sample size, number of loci, and the variability of loci, all of which may affect the performance of these software programs. Evanno et al. (2005) investigated the impact of sample size (samples of individuals and samples of loci) on the performance of STRUCTURE and noticed a decrease in performance with smaller sample sizes; unfortunately, similar studies are not available for BAPS or PARTITION. Our simulated loci had levels of diversity akin to that of highly variable

microsatellite loci. With highly variable loci, the maximum value of $F_{ST}$ is limited by the amount of within-population heterozygosity (Hedrick 1999). For example, an $F_{ST}$ of 0.05 for highly variable microsatellite loci might indicate a biologically significant level of population differentiation, whereas the same value for less variable allozyme loci may indicate an overall lack of structure. Hedrick (2005) has developed a standardized measure of genetic differentiation that takes within-population heterozygosity into account when estimating $F_{ST}$. For purposes of gauging how markers with different degrees of polymorphism might perform, we provide Hedrick's standardized measure for our simulated markers: $F_{ST}$ values of 0.02 (the point at which STRUCTURE and BAPS cannot correctly identify $K$), 0.03 (the lowest point at which STRUCTURE and BAPS can still correctly identify $K$), and 0.05 (where over 97% of individuals are correctly assigned), become 0.20, 0.28, 0.39 when standardized by Hedrick's (2005) method. Thus, both STRUCTURE and BAPS break down at a standardized $F_{ST}$ between 0.20 and 0.28, and are able to correctly assign over 97% of individuals by the time the standardized $F_{ST}$ reaches 0.39.

Both STRUCTURE and BAPS correctly infer the number of subpopulations in a dataset when genetic differentiation among groups is low. However, levels of genetic differentiation must be considerably higher in order to attain an accuracy rate of over 97% correct assignments. In the context of Hedrick's (2005) standardized measure of genetic differentiation, it seems that the observed proportion of variance partitioned among subpopulations must reach nearly 40% of its maximum possible value in order for over 97% of individuals to be correctly assigned.

The computational efficiency of BAPS 3.1 is impressive; it seems that the use of a direct analytical result to compute the marginal likelihood of a putative partition (equation 3 in Corander et al. 2005), combined with the use of a stochastic greedy optimization algorithm in place of MCMC, has led to a marked improvement in computational efficiency. However, we do acknowledge that the greatest confidence in results is attained when the results are arrived at independently by two different methods, and thus we advocate using both STRUCTURE and BAPS for inferring the number of clusters and assignment of individuals to clusters within an empirical dataset when the level of differentiation among groups is small.

## Acknowledgements

We would like to thank Jukka Corander for providing technical advice and an advance copy of his manuscript for BAPS 3.1, and Khalid Belkhir for offering assistance regarding PARTITION. Funding was provided by Purdue University.

## References

Beaumont M, Barratt EM, Gottelli D, Kitchener AC, Daniels MJ, Pritchard JK, Bruford MW (2001) Genetic diversity and introgression in the Scottish wildcat. *Mol. Ecol.*, **10**, 319–336.

Berry O, Tocher MD, Gleeson DM, Sarre SD (2005) Effects of vegetation matrix on animal dispersal: genetic evidence from a study of endangered skinks. *Conserv. Biol.*, **19**, 855–864.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Castric V, Bernatchez L (2004) Individual assignment test reveals differential restriction to dispersal between two salmonids despite no increase of genetic differences with distance. *Mol. Ecol.*, **13**, 1299–1312.

Corander J, Walmann P, Sillanpaa MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Corander J, Walmann P, Marttinen P, Sillanpaa MJ (2004) BAPS2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Corander J, Marttinen P, Mäntyniemi S (2005) Bayesian identification of stock mixtures from molecular marker data. *Fish. Bull.*, in press.

Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.*, **78**, 59–77.

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

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fletcher R (1987) *Practical methods of optimization*, Wiley, New York.

Glaubitz JC (2004) CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol. Ecol. Notes*, **4**, 309–310.

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313–318.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Kyle CJ, Strobeck C (2001) Genetic structure of North American wolverine (*Gulo gulo*) populations. *Mol. Ecol*., **10**, 337–347.

Latch EK, Harveson LA, King JS, Hobson MD, Rhodes OE (2005) Assessing hybridization in wildlife populations using molecular markers: A case study in wild turkeys. *J. Wildl. Manag*., in press.

Maingon RDC, Ward RD, Hamilton JGC, Noyes HA, Souza N, Kemp SJ, Watts PC (2003) Genetic identification of two sibling species of *Lutzomyia longipalpis* (Diptera: Psychodidae) that produce distinct male sex pheromones in Sobral, Ceara State, Brazil. *Mol. Ecol*., **12**, 1879–1894.

Manel S, Berthier P, Luikart G (2002) Detecting wildlife poaching: Identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conserv. Biol*., **16**, 650–659.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol*., **20**, 136–142.

Mank JE, JC Avise (2004) Individual organisms as units of analysis: Bayesian-clustering alternatives in population genetics. *Genet. Res*., **84**, 135–143.

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet*., **41**, 225–233.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol*., **4**, 347–354.

Pearse DE, KA Crandall (2004) Beyond $F_{ST}$: analysis of population genetic data for conservation. *Conserv. Genet*., **5**, 585–602.

Primmer CR, Koskinen MT, Piironen J (2000) The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. *Proc. R. Soc. Lond. B. Biol. Sci*., **267**, 1699–1704.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard JK, Wen W (2003) *Documentation for STRUCTURE software: Version 2*. Available from http://www.pritch.bsd.uchicago.edu.

Randi E, Pierpaoli M, Beaumont M, Ragni B, Sforzi A (2001) Genetic identification of wild and domestic cats (*Felis silvestris*) and their hybrids using Bayesian clustering methods. *Mol. Biol. Evol*., **18**, 1679–1693.