AMOVA-Based Clustering of Population Genetic Data

PATRICK G. MEIRMANS

Department of Ecology and Evolution, Biophore, University of Lausanne, Switzerland; Laboratoire d’Ecologie Alpine, CNRS UMR 5553; Université Joseph Fourier, 38041 Grenoble, France; and Institute of Biodiversity and Ecosystem Dynamics, Universiteit van Amsterdam, The Netherlands.

Address correspondence to Patrick Meirmans at the address above, or e-mail: p.g.meirmans@uva.nl

Abstract

Determining the genetic structure of populations is becoming an increasingly important aspect of genetic studies. One of the most frequently used methods is the calculation of $F$-statistics using an Analysis of Molecular Variance (AMOVA). However, this has the drawback that the population hierarchy has to be known a priori. Therefore, the population structure is often based on the results of a clustering analysis. Here I show how these two steps, clustering and calculation of $F$-statistics, can be combined in a single analysis. I do this by showing how the AMOVA framework is theoretically related to the widely used method of $K$-means clustering and can be used for the clustering of populations into groups. Simulations were used to show that the method performed very well both under random mating and under nonrandom mating. However, when the migration rates were high, the results were better under random mating than under predominant selfing or clonal reproduction. Two summary statistics were tested for estimating the number of clusters. Overall, pseudo-$F$ showed the better performance, but BIC is better for detecting whether any significant structure is present. The results show that the AMOVA-based $K$-means clustering is useful for clustering population genetic data. Programs to perform the clustering can be downloaded from www.patrickmeirmans.com/software.

Key words: AMOVA, clonality, genetic structure, $K$-means clustering, self-fertilization, simulated annealing

Many evolutionary processes, such as selection, local adaptation, and genetic drift strongly depend on a species’ past and present population structure. Therefore, our understanding of evolutionary processes often depends on detailed knowledge about a species’ population structure. Assessment of population structure also has practical importance in conservation biology and the study of invasive species. Not surprisingly, determining the population structure of a species comprises a major part of population genetics. The strength of the genetic population structure of species is usually described in terms of $F$-statistics that relate the within-population diversity to the total genetic diversity (e.g., Weir and Cockerham 1984).

One popular method of calculating $F$-statistics is via an Analysis of Molecular Variance (AMOVA), which makes it easy to test for the presence of hierarchical population structure (Excoffier et al. 1992). However, this technique actually requires that the hierarchical structure of the population is known in advance. Therefore, a clustering analysis (e.g., STRUCTURE, Pritchard et al. 2000) may be performed on the data, the results of which are then used as a basis for the AMOVA. This two-step process, clustering and calculation of $F$-statistics, could be greatly simplified if the clustering were included in the AMOVA itself. This would also remove any ambiguity in how the clustering results, which are mostly individual based and may contain admixture, are converted into a hierarchical population structure. One method that actually does such a combined analysis is the Spatial Analysis of Molecular Variance (Dupanloup et al. 2002), that combines an AMOVA (Excoffier et al. 1992) method with Voronoi polygons for the spatially restricted clustering of populations. However, this method is somewhat restricted in applicability, since it requires the spatial coordinates of the populations as input and assumes that the geographical distances are indicative of the “ecological” distances between populations. Most importantly, the spatial AMOVA also assumes that the correct number of clusters is known a priori, as it does not provide a way to estimate the number of clusters from the data.

In connection with these issues, it is interesting to notice that there is currently an increased interest in the use of multivariate statistics in population genetics, triggered by the availability of large genetic data sets. On the one hand, this includes the use of standard multivariate statistics to discover
patterns in genetic data. For example, Jombart et al. (2010) showed how K-means clustering, which is one of the most widely used general-purpose clustering techniques, can be combined with Discriminant Analysis to visualize population structure. On the other hand, a direct link has now been made between multivariate statistics and population genetic theory (Goudet 1999; McVean 2009; Engelhardt and Stephens 2010). For example, McVean (2009) showed that in a standard Principal Components Analysis the axis scores can be described in terms of coalescent times and the variance explained can be described in terms of $F_{ST}$.

Here, I show that the AMOVA framework is theoretically related to K-means clustering. Therefore the AMOVA itself can be used for clustering populations (or sample collections) into groups, combining the clustering and calculation of $F$-statistics into a single analysis. Such an unsupervised AMOVA will be useful in detecting large-scale patterns in population structure. Furthermore, it will allow a comparison of how well hypotheses of population structure, defined as an a priori hierarchical structure, fit the structure present in the data. In addition, I describe how two different summary statistics can be used to infer the number of population clusters. I then use simulations to test which of these summary statistics is best suited for determining the number of clusters under three different modes of reproduction: random mating, predominant clonal reproduction, and predominant selfing.

**Materials and Methods**

**Linking AMOVA and K-Means Clustering**

Under the class of clustering methods that are collectively known as K-means clustering, the optimal clustering is the one with the smallest amount of variation within clusters, which is calculated using the within-clusters sum of squares (for an introduction, see Legendre and Legendre 1998 [Chapter 8.8]; for applications in genetics, see Liu and Zhao 2006; Lee et al. 2009 Rodriguez-Ramilo et al. 2009; Jombart et al. 2010). AMOVA (Excoffier et al. 1992; Dupanloup et al. 2002) is similarly based on the calculation of sums of squares between and within groups of individuals. Therefore, these sums of squares can be used to find the clustering of populations into groups that have the lowest amount of genetic variation within groups. The minimization of the within-groups sum of squares that is used in K-means clustering is, in the context of a hierarchical AMOVA, equivalent to minimizing the among-populations-within-groups sum of squares ($\text{SSD}_{AP/WG}$) (Excoffier et al. 1992, Equation 8b). The hierarchical population structure in the AMOVA then consists of different hierarchical levels: individuals, populations, and clusters of populations. Different $F$-statistics can be calculated based on the variance components for the different hierarchical levels. In terms of $F$-statistics, the minimization of $\text{SSD}_{AP/WG}$ comes down to a maximization of $F_{CT}$, the variance among clusters ($\phi$) relative to the total variance ($\phi_T$).

In an AMOVA, the within and between-groups sums of squares are calculated from a matrix of squared Euclidean distances between all pairs of individuals (Li 1976; Excoffier et al. 1992). For codominant markers this is usually done using a locus-by-locus approach, where a separate distance matrix is used for every locus (Michalakis and Excoffier 1996). However, K-means clustering uses a single distance matrix since this greatly simplifies the clustering algorithm. Smouse and Peakall (1999) developed a method to calculate a single distance matrix of squared Euclidean distances from one or more loci that can readily be used for an AMOVA (Maguire et al. 2002). However, this distance metric disregards the variation within individuals and as a result yields higher values for the $F$-statistics than the locus-by-locus approach (comparable to values for haploids). For diploids, this distance measure is simply calculated as twice the squared Euclidean distance among the vectors of within-individual allele frequencies of the two individuals $i$ and $j$:

$$d_{ij}^2 = 2 \sum_{l=1}^{L} \sum_{a=1}^{A_l} (p_{ia} - p_{ja})^2$$

where $L$ is the number of loci, $A_l$ is the number of alleles at locus $l$, and $p_{ia}$ is the frequency of allele $a$ at locus $l$ within individual $i$ (which for diploids can take the value 0, 0.5, or 1). For a hierarchical population structure that includes individuals, populations, and groups of population, the Euclidean distances can then be used to calculate the different sums of squares. Using equations 8a-c from Excoffier et al. (1992), we obtain the within-populations sum of squares ($\text{SSD}_{WP}$), the among-populations-within-groups sum of squares ($\text{SSD}_{AP/WG}$), and the among-groups sum of squares ($\text{SSD}_{MG}$). If there are many missing values in a data set, these should be taken into account before the analysis. For example, individuals or loci with a lot of missing data may be removed from the analysis, or the missing values can be replaced with randomly drawn alleles.

**Clustering Algorithm**

The above-described method of calculating the sums of squares can be used with most K-means clustering algorithms, such as the classical algorithm by MacQueen (1967). However, this algorithm has the drawback that it may get stuck in local maxima, especially for large data sets. Here, I use the technique of simulated annealing (Kirkpatrick et al. 1983), which is the method used by Dupanloup et al. (2002) for their spatial clustering technique. In every step of the simulated annealing algorithm, one randomly picked population (with all the individuals in it) is placed in a randomly picked cluster. The new clustering is then either accepted or rejected depending on the change in sum of squares, and a parameter called the “temperature” of the chain (see Supplementary Material online). Simulated annealing has the ability to move away from a local maximum; nevertheless, it is not guaranteed to yield the global maximum. Therefore, the chain is usually run a number of times and the best result is chosen from among these runs. For the simulated data sets described below, the simulated annealing algorithm consistently outperformed the classical algorithm (results not shown).
Determining K Using a Summary Statistic

Finding the best way to distribute the population samples over \( k \) clusters is often only a part of the problem, since in many cases the true value of \( K \) is not known \( \text{a priori} \) (note, I here use uppercase \( K \) to refer to the true value and lowercase \( k \) for the estimate). A common method is to calculate a summary statistic for every value of \( k \) and choose the optimum value of \( k \) based on these values (Legendre and Legendre 1998, Chapter 8.8). Many different summary statistics are available that differ in their performance and their applicability (Milligan and Cooper 1985). Here, I test the performance of two different summary statistics that have a track record of having been proved for clustering and for model selection: pseudo-\( F \) (Caliński and Harabasz 1974) and Bayesian Information Criterion (Schwarz 1978).

Pseudo-\( F \) (Caliński and Harabasz 1974) relates \( r^2 \), the fraction of the total variance that is explained by the clustering, to the number of clusters \( k \) and the number of populations \( n \):

\[
F_k = \frac{r^2}{(1-r^2)(n-k)}.
\]

Where \( r^2 = (\text{SSD}_T - \text{SSD}_{AP/WG})/(\text{SSD}_T - \text{SSD}_{WP}) \). The clustering with the highest value for pseudo-\( F \) is regarded to provide the best fit. One drawback of pseudo-\( F \) is that it is not defined for \( k=1 \). This means that pseudo-\( F \) cannot be used to determine whether there actually is any significant clustering in the data (\( k \geq 1 \)) or not (\( k=1 \)).

The Bayesian Information Criterion (BIC) was developed as a summary statistic for model selection using a Bayesian framework (Schwarz 1978). However, the criterion is also applicable outside of the Bayesian context and has been used in a number of fields, including population genetic (Jombart et al. 2010):

\[
\text{BIC}_k = n \cdot \ln(\text{SSF}) + k \cdot \ln(n)
\]

Unlike pseudo-\( F \), BIC can be calculated for \( k=1 \).

Simulations

The program EASYPOP (Balloux 2001) was used to generate genotypic data with different numbers of population clusters and with different modes of reproduction. Migration followed a hierarchical island model with 20 populations, divided into 1, 2, 4, or 5 archipelagos. Migration between populations within archipelagos occurred at a rate of 0.1, migration between archipelagos occurred at a rate of 0.01, 0.02, or 0.03. Each population consisted of \( N=100 \) diploid hermaphroditic individuals. Each metapopulation model was run with three different modes of reproduction: complete random mating, 95% clonal reproduction with 5% random mating, and 95% selfing with 5% random mating. In all cases, the random mating included a probability of self-fertilization of 1/N. Mutation followed a K-alleles model with 50 possible states and a mutation rate of \( 10^{-4}\), this resulted in an average of 6.3, 6.3, and 3.9 alleles per locus for the three modes of reproduction respectively, and corresponding expected heterozygosities (\( H_s \)) of 0.44, 0.44 and 0.29 (calculated using HIERFSTAT; Goudet 2005). A total of 100 loci were simulated, but I also tested the performance of the clustering when only 20 loci were used. Simulations were started with maximum diversity at all loci and were then run for 10 000 generations to allow for equilibrium between mutation, drift, and migration to be reached. For all combinations of parameters 100 independent replicates were run. The program GENODIVE (Meirmans and Van Tienderen 2004) was used to calculate the among-cluster \( F_{CT} \) statistics based on the known correct clustering. This was done using the method from Smouse and Peakall (1999) to calculate the matrix of squared Euclidean distances (Table 1) and using the locus-by-locus approach (see Supplementary Table S1 online).

### Table 1

<table>
<thead>
<tr>
<th>Migration among archipelagos</th>
<th>Number of archipelagos</th>
<th>Random mating</th>
<th>Clonality</th>
<th>Selfing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2</td>
<td>0.020 (0.003)</td>
<td>0.019 (0.005)</td>
<td>0.019 (0.005)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.062 (0.005)</td>
<td>0.060 (0.01)</td>
<td>0.061 (0.012)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.081 (0.005)</td>
<td>0.081 (0.015)</td>
<td>0.080 (0.012)</td>
</tr>
<tr>
<td>0.02</td>
<td>2</td>
<td>0.008 (0.002)</td>
<td>0.009 (0.005)</td>
<td>0.008 (0.003)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.029 (0.003)</td>
<td>0.029 (0.005)</td>
<td>0.029 (0.006)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.039 (0.004)</td>
<td>0.038 (0.007)</td>
<td>0.039 (0.007)</td>
</tr>
<tr>
<td>0.03</td>
<td>2</td>
<td>0.004 (0.001)</td>
<td>0.005 (0.002)</td>
<td>0.005 (0.002)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.018 (0.002)</td>
<td>0.018 (0.004)</td>
<td>0.018 (0.004)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.024 (0.002)</td>
<td>0.024 (0.005)</td>
<td>0.025 (0.004)</td>
</tr>
</tbody>
</table>

The values represent the averages over 100 simulated data sets, with the standard deviations given within parentheses. The \( F \)-statistics cannot be computed when all populations form a single cluster.
Clustering was performed using a command-line program specifically written for this purpose that was compiled and run on both Macintosh and Windows computers. A graphical user interface for performing the clustering has been implemented in the program GENODIVE Version 2.0 (Meirmans and Van Tienderen 2004), which is available for computers running Mac OS X. The programs and the source-code for the command line utility can be downloaded from http://www.patrickmeirmans.com/software.

Results
Quality of the Clustering with Known K
When the correct value of K was known a priori, the AMOVA-based clustering of populations into groups yielded good results for most scenarios (Table 2). Especially under random mating, the assignment of populations to clusters was close to 100% correct. When there was clonality or selfing, the results of the population clustering were not as good as under random mating, but still dependable, especially for the lowest migration rate (Table 2). The assignment of populations to clusters was more difficult when there were 2 archipelagos than when there were 4 or 5. This is because the total number of individuals was kept fixed at 2,000 for the whole metapopulation. Therefore, at K=2, there are more individuals per cluster (1,000) than for K=4 and K=5 (500 and 400 individuals resp.). As a result, genetic drift within clusters is stronger at higher K, and the impact of migration among clusters is smaller. This leads to a lower degree of differentiation as evidenced by the lower FCT values at K=2 (Table 2).

Estimating the Number of Clusters
It is difficult to tell which summary statistic showed a better overall performance for estimating the number of population clusters. The performance of the two statistics was mostly comparable, though pseudo-F was better at the highest migration rate (Figure 1). One important limitation of pseudo-F is that it cannot be calculated for k=1 and therefore cannot be used to test whether there is any significant structure. For the simulated data sets without any structure, pseudo-F typically returned an optimal k of 2. The mode of reproduction had a strong effect on the performance of both summary statistics. The best results were obtained under random mating where in some cases a 100% correct inference of the number of clusters was observed. The effect of non-random mating could especially be seen at higher rates of gene flow when population structure is weak. In that case, the results of BIC at K=1, and of pseudo-F at K=2 seem very good, but that is simply due to the fact that in the absence of any detectable structure, these statistics return their minimum value. The number of loci used strongly affected the estimation of the number of clusters (see Supplementary Figure S1 online). With 20 loci, the results were still good under random mating and when the rates of gene flow are not very high. However, for the highest rate of gene flow, the results are notably worse than for 100 loci, especially under nonrandom mating.

Discussion
The Analysis of Molecular Variance (AMOVA, Excoffier 1992) provides one of the most widely used frameworks for analyzing population genetic data. Part of its appeal is probably the ease with which it can be used to analyze a hierarchical population structure, where individuals are grouped into populations (or sampling locations) and populations are grouped into higher-level clusters. However, the use of an AMOVA in this way requires a priori knowledge of this hierarchy, which may not be available. Commonly, nongenetic criteria, such as geographical, ecological, or linguistic data, are used to infer the hierarchical structure to be tested, with the risk of being incorrect. Alternatively, different possible hierarchies are analyzed to see which are of genetical importance. In this article, I showed that the AMOVA framework is theoretically related to the general method of K-means clustering and that the AMOVA itself can therefore be used to infer the population structure, that is without the need to define the population clusters a priori. This means that the returned clustering is the one with the strongest among-cluster differentiation as measured by FCT. Because an AMOVA can be calculated at multiple hierarchical levels, it is in principle also possible to use it to perform a clustering of individuals into putative populations (i.e., maximizing FST). However, I found the

Table 2  Percentage of populations that was incorrectly assigned to their population clusters when the correct number of archipelagos was set a priori. Results are for three different modes of reproduction: complete random mating, 95% clonal reproduction with 5% random mating, and 95% self-fertilization with 5% random mating

<table>
<thead>
<tr>
<th>Migration among archipelagos</th>
<th>Number of archipelagos</th>
<th>Random mating</th>
<th>Clonality</th>
<th>Selfing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2</td>
<td>0.0 (0.0)</td>
<td>1.4 (2.8)</td>
<td>1.9 (4.3)</td>
</tr>
<tr>
<td>0.02</td>
<td>2</td>
<td>0.0 (0.0)</td>
<td>0.1 (0.5)</td>
<td>0.1 (0.7)</td>
</tr>
<tr>
<td>0.03</td>
<td>2</td>
<td>0.9 (2.3)</td>
<td>10.6 (11.2)</td>
<td>13 (10.4)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.0 (0.0)</td>
<td>1.3 (2.9)</td>
<td>2.7 (4.2)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.0 (0.0)</td>
<td>1.5 (3.5)</td>
<td>1.5 (3.2)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.8 (8.1)</td>
<td>19.6 (11.5)</td>
<td>22.5 (11.3)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.3 (1.1)</td>
<td>6.7 (7.1)</td>
<td>11.3 (8.7)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.1 (0.5)</td>
<td>4.5 (5.0)</td>
<td>6.5 (6.1)</td>
</tr>
</tbody>
</table>

The values represent the averages over 100 simulated data sets, with the standard deviations given within parentheses.
that for such use the method performs poorly, especially under nonrandom mating (results not shown), and in such cases the method does not provide an alternative to model-based clustering methods such as STRUCTURE (Pritchard et al. 2000). Furthermore, STRUCTURE has the ability to detect admixed individuals by partly assigning them to multiple populations. In contrast, in the AMOVA-based method described here the assignment of populations to clusters is "hard": a population cannot be partly assigned to multiple clusters. This means that it is not possible to detect admixed populations using this method.

Of the two summary statistics tested here, there was no single-best statistic for determining the number of population clusters. Pseudo-$F$ (Caliński and Harabasz 1974) was the better statistic under random mating, while BIC (Schwarz 1978) performed better under nonrandom mating at the intermediate migration rate. Furthermore, BIC also was able to correctly pinpoint when there was no structure ($K=1$). Given this difference in performance between BIC and pseudo-$F$, one has to decide which statistic to use depending on the knowledge about the species at hand. If such knowledge is unavailable or ambiguous, I advise to use both BIC and pseudo-$F$ and see which statistic returns the clustering that makes most biological sense. This is comparable to the way in which often both the original statistic from Pritchard et al. (2000) and the Delta $K$ statistic from Evanno et al. (2005) are used to estimate the number of clusters using STRUCTURE.

Simulations present a good way to judge the usefulness of clustering methods since the actual number of clusters is known a priori. In this article, I have aimed to make the simulations broadly applicable by not only including random mating, but also self-fertilization and asexual reproduction. However, any simulation study is by necessity restricted in its scope and can never capture the breadth of processes that are taking place in nature. The scenarios simulated here all used the Island model. Though this model is the standard in population genetics simulations, it is strictly symmetrical and is not spatially explicit (Meirmans 2012). Furthermore, the degree of genetic variability in any given empirical data set may depart greatly from the values chosen for the simulations presented here. Therefore, the performance of not only this clustering technique, but also that of others, needs to be verified for situations where clusters have different sizes and different shapes.
and should be assessed by users with gene frequencies and number of loci relevant to their own study organisms.

For actual species data there may not even be an objective way to determine the number of clusters since that very much depends on one’s definition of what a population and a group of populations is. There are many different ways to define the concept of a population, but none of them has gained widespread acceptance among biologists (Waples and Gaggiotti 2006). Regarding clustering methods, it may therefore be that different methods give different answers that all have a valid biological meaning. On the other hand, clustering results for a “suboptimal” value of k may be biologically informative. For example, in a study of 27 alpine species, Meirmans et al. (2011) noted that there was no single clustering method that gave a satisfactory answer for all 27 species. On the other hand, a comparison of all values of k from 2 to 5 among the same 27 species revealed meaningful biological patterns (Alvarez et al. 2009). This again shows that though estimating the number of clusters can be useful, it should be taken as a guideline and it is important to consider that in real data sets there can be multiple values of k that represent true biological patterns.

There is a long history of the use of multivariate statistics in population genetics, going back to Menozzi et al.’s (1978) pioneering use of PCA to analyze patterns in allele frequencies. Recently, an effort has started to link these analyses more directly to population genetic theory (Goudet 1999; McVean 2009; Engelhardt and Stephens 2010). For example, McVean (2009) showed that in a PCA, the projection of samples onto the component axes is directly related to the average coalescence times between samples. Furthermore, the proportions of explained variance can be expressed in terms of $F_{ST}$, a point earlier noted by Goudet (1999). Here, I described how the widely used method of K-means clustering can be linked to the AMOVA framework. This connection gives the AMOVA-based K-means clustering a solid basis in population genetic theory. The AMOVA-based K-means clustering also provides an excellent complement to a standard hierarchical AMOVA with a priori defined clusters of population samples. This way, one can test how well the expected population structure matches the structure observed in the data, allowing a more detailed analysis of the biological processes that shape the distribution of genetic variation.

**Supplementary Material**


**Funding**

European Union (FP 6 project nr. 036866); and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (838.06.042).

**Acknowledgments**

I would like to thank Jérôme Goudet for stimulating discussions, the use of a computer cluster for performing the simulations, and comments on an earlier draft of the manuscript. Robin Waples (the associate editor), Thibaut Jombart, and an anonymous reviewer provided comments that have greatly improved the manuscript. As always, Stephanie Meirmans provided discussions and support.

**References**


Goudet J 1999. PCAGEN: Principal Components Analysis on gene frequency data (version 1.2). [http://www2.unil.ch/popgen/softwares/pcagen.htm](http://www2.unil.ch/popgen/softwares/pcagen.htm)


Received August 16, 2011; Revised May 19, 2012; Accepted May 25, 2012

Corresponding Editor: Robin Waples