Utility of the program Structure to recover population clustering on a simulated dataset

Alexa Warwick

Introduction

One of the main goals in biology is to understand differentiation among populations. This understanding is vital to both determining whether two populations are sufficiently different to merit separate species designations, as well as to address within-species dynamics for the purposes of conservation management, i.e. should the populations in question be treated as separate management units. One approach to answer these questions is to collect genetic information for individuals in each population of interest and then probabilistically assign individuals to clusters. By comparing the assignment of individuals for different numbers of clusters (models with one, two, three or more clusters), we can compare the results to determine which model is the most appropriate given our data.

Methods

For this study we used a simulated dataset with known population parameters in order to test the ability of Structure to reconstruct these parameters. This dataset included four ‘sampled’ putative populations for 75 diploid individuals across 10 loci (per population totals listed in Table 1).

Table 1. Number of individuals per population for the simulated dataset

Putative population	Number of individuals
1	40
2	20
3	10
4	5

Structure v2.3.4 (Pritchard and Stephens 2000) was used to assess differentiation among the putative populations. We used mostly default settings, such as the admixture model and the specification that allele frequencies were correlated among populations. We ran 1 million MCMC repetitions after a burn-in period 300,000. The number of clusters, or the “K” value, was tested from K = 1 to K = 6 with five or ten iterations of each model (Table 2). The results from the Structure runs were then used in Structure Harvester (Earl and vonHoldt 2012), which summarized the results of repeated runs and determined the optimal K using the Evanno method (Evanno et al. 2005). We then used CLUMPAK (Kopelman et al. 2015) to summarize the results for comparison, and to generate figures with a sample assignment probability to each K cluster (summarized across runs) using consistent coloring of each cluster.

Results/Discussion

The 75 individuals were assessed for cluster assignment for each value of K, the results of which are shown in Figure 2. The Evanno method suggested two was the optimal K because it had the largest delta K value (Table 2, Figure 2). Structure Harvester and CLUMPAK showed the same results. Under this K = 2 model, putative populations 1 and 2 comprised a single cluster (blue) and populations 3 and 4 comprised the orange cluster (Figure 1). At higher values of K, the blue cluster (populations 1 and 2) had some individuals assigned to new clusters (purple, green, pink, etc.), although these new clusters were not concordant with the putative population origin. In contrast, the orange cluster always contained individuals from putative populations 3 and 4, even at higher values for K. These results suggest that individuals in populations 3 and 4 were likely sampled from a single population (cluster), or from two populations with very high gene flow between them. The putative populations 1 and 2 are clearly more similar to each other than populations 3 and 4; however, 1 and 2 may yet have additional substructure that is not captured via Structure's assignment method, given the mixed assignment of individuals. Using this same dataset to assess structure/differentiation with a different method may be a useful next step.

Table 2. Summary across replicates of K values using Structure Harvester. Using the Evanno method, K = 2 is the optimal value (highlighted in orange).

# K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	10	-2152.26	0.3688	NA	NA	NA
2	10	-1888.25	0.0707	264.01	111.86	1581.93
3	10	-1736.10	0.9809	152.15	78.70	80.23
4	10	-1662.65	0.6654	73.45	27.70	41.62
5	5	-1616.90	1.1023	45.75	9.71	8.80
6	5	-1561.44	0.4980	55.46	NA	NA

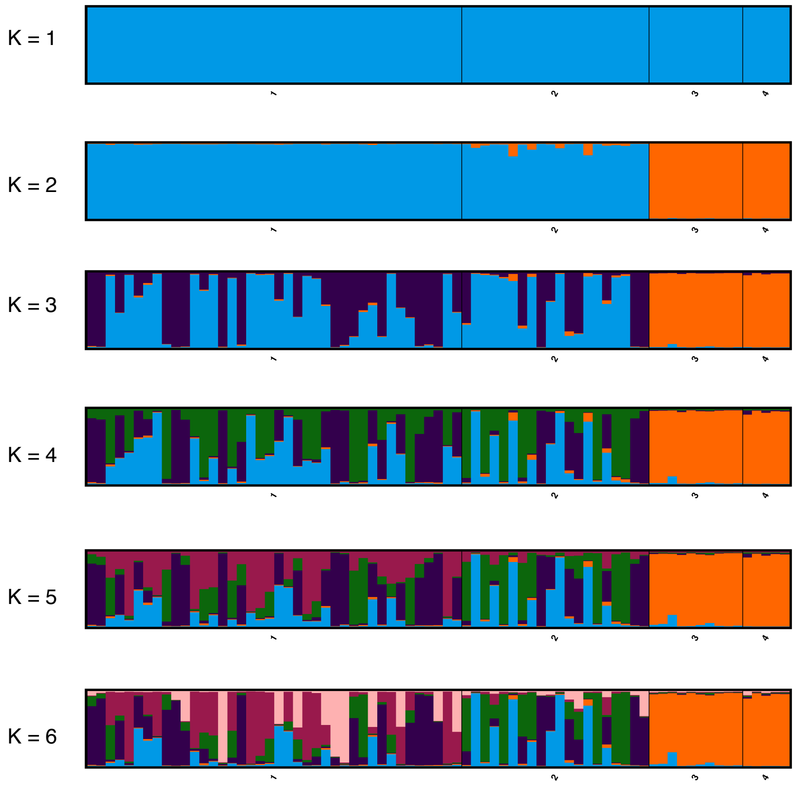

Figure 1. Individual assignment to each cluster for K = 1 to K = 6 (summary across multiple runs for each K value). Putative population numbers from 1 (on the left) to 4. Note: The lines marking the separation between putative populations is present but difficult to see.

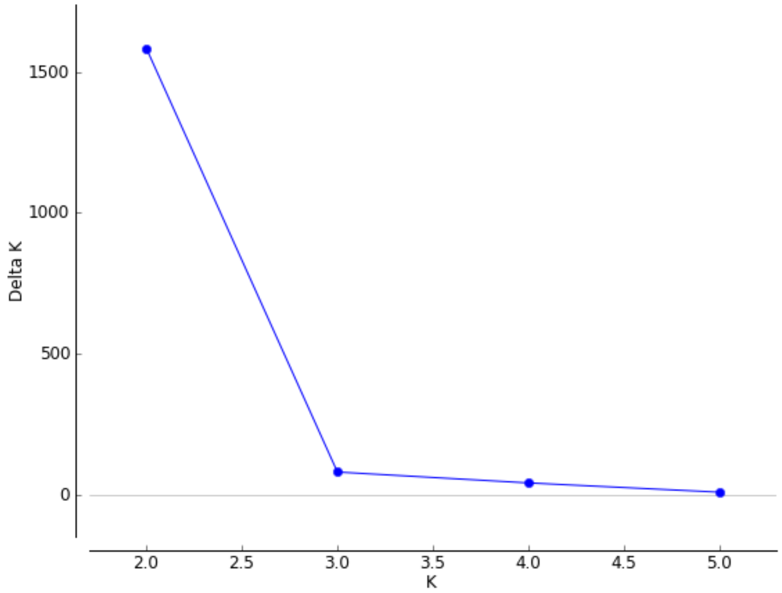

Figure 2. Structure Harvester results for the Evanno method, where K = 2 is the optimal value (largest delta K).

Literature Cited

Earl DA, vonHoldt BM (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetic Resources*, **4**, 359–361.

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

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12387.

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