This is the peer reviewed version of the following article: Cullingham, C.I., Miller, J.M., Peery, R.M., Dupuis, J.R., Malenfant, R.M., Gorrell, J.C., & Janes, J.K. (2020). Confidently identifying the correct K value using the ΔK method: When does $K=2$? *Molecular Ecology*, 29(5), 862-869, which has been published in final form at <http://doi.org/10.1111/mec.15374>. This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Use of Self-Archived Versions](#).

Confidently identifying the correct K value using the ΔK method: when does $K = 2$?

Cunningham CI¹, Miller JM², Peery RM², Dupuis JR³, Malenfant RM⁴, Gorrell JC⁵, Janes JK⁵

¹Department of Biology, Carleton University, Ottawa, ON, Canada

²Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

³Department of Entomology, University of Kentucky, Lexington, KY, USA

⁴Department of Biology, University of New Brunswick, Fredericton, NB, Canada

⁵Biology Department, Vancouver Island University, Nanaimo, BC, Canada

Corresponding author: Catherine Cunningham, Department of Biology, Carleton University, Ottawa, ON, Canada. email: catherine.cunningham@carleton.ca

Running head: When does $K = 2$?

Abstract

Populations delineated based on genetic data are commonly used for wildlife conservation and management. Many studies use the program STRUCTURE, combined with the ΔK method to identify the most probable number of populations (K). We recently found $K = 2$ was identified more often when studies used ΔK compared to studies that did not. We suggested two reasons for this: hierarchical population structure leads to underestimation, or the ΔK method does not evaluate $K = 1$ causing an overestimation. Our current work aims to develop a better understanding of the limits of the method using one, two, and three population simulations across migration scenarios. From these simulations we identified the “best K ” using model likelihood and ΔK . Our findings show that mean probability plots and ΔK are unable to resolve the correct number of populations once migration rate exceeds 0.005. We also found a strong bias towards selecting $K = 2$ using the ΔK method. We used these data to identify the range of values where the ΔK statistic identifies a value of K that is not well supported. Finally, using the simulations and a review of empirical data, we found that the magnitude of ΔK corresponds to the level of divergence between populations. Based on our findings, we suggest researchers should use the ΔK method cautiously; they need to report all relevant data, including the magnitude of ΔK , and an estimate of connectivity for the research community to assess whether meaningful genetic structure exists within the context of management and conservation.

Keywords: population structure, ΔK , genetic structure, migration rate, F_{ST}

Introduction

A major contribution of population genetics to the management and conservation of wildlife is the identification of populations (Allendorf, Luikart, & Aitken, 2013; Frankham, Ballou, & Briscoe, 2002; Palsbøll, Bérubé, & Allendorf, 2007; Reiss, Hoarau, Dickey-Collas, & Wolff, 2009), where a population can be defined as an independent group of interbreeding individuals. Misidentification of populations can have negative consequences, including over-harvesting, loss of unique genetic diversity, and misallocation of conservation resources. Given the importance of identifying genetic populations, population geneticists have been developing and refining approaches without needing *a priori* information (e.g. Chen, Durand, Forbes, & Francois, 2007; Corander, Waldmann, & Sillanpää, 2003; Dawson & Belkhir, 2001; François, Ancelet, & Guillot, 2006; Guillot, Mortier, & Estoup, 2005; Huelsenbeck, Andolfatto, & Huelsenbeck, 2011; Jombart, Devillard, & Balloux, 2010; Pritchard, Stephens, & Donnelly, 2000). Currently, the most widely used program is STRUCTURE (Pritchard, Stephens, & Donnelly, 2000; cited more than 19 000 times, source ISI Web of Science, 10-31-2019). The popularity of this method is due to its ease of use, ability to identify weakly differentiated populations (Latch, Dharmarajan, Glaubitz, & Rhodes, 2006), and consistency of performance (although see Gilbert et al., 2012). However, one major drawback of this program is that the selection of the most appropriate number of populations (K) is a subjective task. To make the choice of K more objective, Evanno et al. (2005) presented the ΔK method, which uses the rate of change in model likelihoods to identify the optimal K . We recently reviewed the literature regarding the use of STRUCTURE and found that application of the ΔK method more often led researchers to choose $K = 2$ as the optimal solution than if the method was not used (Janes et al., 2017). We hypothesized that an over-representation of $K = 2$ may be the result of either structure not being correctly identified, e.g. underestimation due to hierarchical population structure (Puechmille, 2016), or overestimation due to the inability of the ΔK method to evaluate $K = 1$ (Evanno, Regnaut, & Goudet, 2005; Gilbert et al., 2012).

To follow up on our review, we compared the performance of the ΔK method versus STRUCTURE likelihoods under scenarios where $K = 1, 2$, or 3 over a variety of biologically relevant parameters. We simulated populations under migration scenarios that ranged from low migration (highly diverged populations) to highly admixed. Using these data, we wanted to explore how the ΔK method is biased towards $K = 2$ in comparison to using STRUCTURE probability plots, and whether the value of the ΔK statistic tells us anything about model support, i.e. can we identify the range of ΔK values where there is strong support for the given solution. Finally, we extracted the optimal ΔK ($\Delta K_{\text{Optimal}}$) and F_{ST} values from empirical studies showing $K = 2$ using the ΔK method within our original review and compared these values with our simulated data. Knowing that there is a bias with the ΔK method, it is important for researchers to recognize when their $K = 2$ results are strongly supported, and when they should be cautious in their interpretation.

Methods

Simulations

We used EASYPOP 1.7 (Balloux, 2001) to simulate one, two, and three randomly mating, diploid populations for four different sampling and population size scenarios across different migration rates (detailed parameter input is included in Table S1). For the two- and three-population scenarios, starting population size for each population was either $N = 100$ or $N = 500$, and from these we sampled either $n = 10$ or $n = 20$ individuals. We modelled both microsatellites and SNPs as markers and chose parameters to reflect typical population genetic study designs, while ensuring reasonable computation time. In our original review, the average number of microsatellites used was 15, and the majority of SNP studies used < 100 loci (Janes et al. 2017, supplemental 2). Therefore, we simulated 15 microsatellite and 50 SNP loci. The mutation model for microsatellites was a combination of the single-step mutation (SSM: 75%), and K -allele model (KAM: 25%) with 20 variable states ($\mu = 0.001$), while for SNPs we used the KAM with two variable states ($\mu = 0.0001$). For migration rates, we first explored seven different migration rates

using two populations ($m = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.5$) for 10 replicates of each parameter set (280 simulations). Based on the results of this preliminary analysis, we chose to narrow the migration rates to those at which ΔK drops sharply (Figure S1), selecting five migration rates ($m = 0.0001, 0.001, 0.005, 0.01, 0.5$). For the three-population scenarios we used an island model of migration and simulated four migration rates ($m = 0.0001, 0.001, 0.005, 0.01$) as the last two migration rates ($m = 0.01, 0.5$) had similar outcomes based on the two-population scenarios (see Results). Aside from the panmictic simulations ($m = 0.5$), the migration rates represent small numbers of effective migrants per generation ($Nm_{500} = 0.05, 0.5, 2.5, 5$; $Nm_{100} = 0.01, 0.1, 0.5, 1$). Fifty replicates were run for each parameter set, resulting in a total of 3600 population simulations (Figure S2). Simulations were run for 20 000 generations starting at minimum diversity; we examined the models for stability to ensure they were run long enough to reach equilibrium (Figure S3a/b). For the one-population simulations we only used a subset of the parameters: we used both markers at the larger sample size ($n = 20$) for $N = 100$ and 500 with 25 simulations each totaling 100 simulations (generation time and marker mutation models were the same as above).

Structure analysis

The output files from EASYPOP were converted to FSTAT and STRUCTURE files for downstream analyses (custom scripts for conversion are available here: https://github.com/rmpeery/deltaK_for_K-2). We ran all simulations in STRUCTURE 2.3.4 (Falush, Stephens, & Pritchard, 2003; Pritchard et al., 2000) with the following parameters: $K = 1-5$, correlated allele frequencies, burn-in = 100 000, sampling = 250 000, and five replicates. We used CLUMPAK 1.1 (Kopelman, Mayzel, Jakobsson, Rosenberg, & Mayrose, 2015) to summarize all of the runs, and to obtain the ΔK statistic for the optimal K and the STRUCTURE model likelihoods. Using FSTAT 2.9.3.2 (J. Goudet, 1995), we obtained F_{ST} estimates for all simulations. As the presence of population structure is inherently associated with Hardy-Weinberg disequilibrium, we calculated overall deviation from Hardy-Weinberg equilibrium (HWE) for each simulation (without using

population flags) using the *adegenet* (Thibaut Jombart, 2008) and *pegas* (Paradis, 2010) packages in R 3.5.3 (R Core Team 2019).

The optimal K was determined as the smallest value of K for which there was an increase in mean probability prior to a plateau or decrease in the mean likelihood plot by estimating the differences between the likelihoods for each successive K (e.g. likelihood for $K = 2$ minus likelihood for $K = 1$). For simulations where the difference was less than 50, we examined the plots visually (Figures 1, 2, S4, S5).

One of our study goals was to develop a reference for researchers to use in identifying the range of $\Delta K_{\text{Optimal}}$ values which correctly identify $K = 2$ or 3 when the mean probability plots also indicate $K = 2$ or 3. Therefore, we created logistic regression models for the two- and three-population scenarios to look at the relationship between the probability of selecting the correct K (based on the K estimated using STRUCTURE likelihoods) with the value of $\ln(\Delta K_{\text{Optimal}})$ as the explanatory variable. We used the `glm` function in R, randomly selecting 60% of the data to develop the model, and 40% to test the model accuracy. We used the *ROCR* 1.0-7 package (Sing, Sander, Beerenwinkel, & Lengauer, 2005) to assess explanatory power of each model. For these analyses we did not use the two-population scenarios with $m = 0.5$ because they represented panmixia.

In addition to developing the model, we also examined the relationship between $\Delta K_{\text{Optimal}}$ and estimated global F_{ST} . We estimated the correlation between these values using a Spearman's rank correlation because the relationship between variables was not linear.

Empirical data

Simulations may not reflect natural populations accurately therefore we recorded F_{ST} and $\Delta K_{\text{Optimal}}$ values for all papers reviewed in Janes et al. (2017) that identified two populations using the ΔK method. Where possible, we directly recorded the reported $\Delta K_{\text{Optimal}}$ values, but for studies where values were only presented in a graph we used WEBPLOTDIGITIZER (<https://apps.automeris.io/wpd/>, accessed 15-04-2019)

to estimate $\Delta K_{\text{Optimal}}$ from the plot. When recording F_{ST} values, we only included those where the authors explicitly stated that the F_{ST} was between the two genetic clusters they found, not between sampling localities.

Results

Comparison between ΔK and mean likelihood

After simulating single populations, the mean probabilities correctly identified $K = 1$ in 100% of the runs as expected (Figure S6). While ΔK is not able to identify $K = 1$, $K = 2$ was selected more frequently than expected, assuming all solutions of K were equally probable ($\chi^2 = 14.06$, $\text{df} = 2$, $p < 0.001$; Table 1). For the two- and three-population scenarios, we found that once the migration rate was ≥ 0.005 , the mean probability plots indicated $K = 1$ for most of the runs across the different simulation parameters (no change in mean probability across all K values; microsatellite data: Figures 1 and 2, SNP data: Figures S4 and S5). When we compare these results with the ΔK method for the two-population scenarios, the two methods agree for the first two migration rates ($m = 0.0001$, 0.001), but because the ΔK method cannot evaluate $K = 1$, they disagree for the majority of simulations where $m \geq 0.005$. For simulations without significant deviation from HWE, $\Delta K = 2$ was selected more often than expected assuming all solutions are equally probable ($\chi^2 = 332.72$, $\text{df} = 2$, $p < 0.001$; Table 1). For the three-population scenarios, there was a greater disagreement between the two methods for the low migration rates ($m = 0.0001$, 0.001), where the ΔK method identified $K = 2$ in 12% of the simulations while the mean probability method strongly supported $K = 3$. In instances without significant deviation from HWE, we again found a bias towards $K = 2$ using ΔK ($\chi^2 = 264.05$, $\text{df} = 2$, $p < 0.001$, Table 1).

Modelling the optimal ΔK

Our models for the two and three population scenarios had high explanatory power based on the area under the receiver operating characteristic curve (AUC) (two population: $\text{AUC} = 0.99$, three population:

AUC = 0.97; Figure S7). Plotting the predicted versus the actual data, we identified the area of model uncertainty, where the correct number of clusters may or may not be selected using the ΔK method (Figure 3) for two populations ($6 < \Delta K_{\text{Optimal}} < 53$), and three populations ($2.5 < \Delta K_{\text{Optimal}} < 400$). We recognize that model parameters may affect the relationship between the probability of selecting the correct K with the value of $\ln(\Delta K_{\text{Optimal}})$, therefore we examined each set of simulation parameters separately for the two-population scenarios. We found very similar areas of uncertainty for each of the models (Figure S8).

Relationship with F_{ST}

Estimates of $\Delta K_{\text{Optimal}}$ and F_{ST} were highly correlated across all model parameters for both the two-population scenarios (microsatellites: $r_s = 0.81$, $p < 0.001$, SNPs: $r_s = 0.84$, $p < 0.001$; Figure 4), and three-population scenarios (microsatellites: $r_s = 0.87$, $p < 0.001$, SNPs: $r_s = 0.89$, $p < 0.001$; Figure 4); indicating that the magnitude of $\Delta K_{\text{Optimal}}$ is directly related to the level of differentiation between the clusters.

Empirical data

We found 329 analyses that identified $K = 2$ using the ΔK method and presented this as their conclusion (references and data included in Table S2). Of the 329 analyses, 56 (17%) reported both $\Delta K_{\text{Optimal}}$ and F_{ST} values, 156 (48%) presented only $\Delta K_{\text{Optimal}}$, 8 (2%) presented only F_{ST} , and 109 (33%) studies did not report either value. Data from the 56 analyses that reported both values demonstrated that empirical $\Delta K_{\text{Optimal}}$ and F_{ST} values show a similarly strong correlation as our simulated data ($r_s = 0.69$, $p < 0.001$; Figure 4). Four studies reported $\Delta K_{\text{Optimal}}$ values < 6 .

Discussion

In investigating the behaviour of the ΔK statistic, our analyses highlight three key findings. First, the ability to identify populations using mean probability plots and the ΔK method breaks down at relatively low migration rates ($m \geq 0.005$), where biologically meaningful genetic differentiation will still exist. Second, when $K = 1$, the ΔK method is more likely to identify $K = 2$, than alternative K -values. Finally,

171 there is a strong relationship between the magnitude of $\Delta K_{\text{Optimal}}$ and F_{ST} , which was robust regardless of
172 marker type, sample size, or proportion of sample relative to the underlying population size.

173 Most of the migration rates we selected would result in few migrants per generation ($0.05 - 5$),
174 with panmixia being the exception. Previous simulation-based studies have shown that the mean
175 probability plots from STRUCTURE resolve populations relatively well at low F_{ST} (~ 0.03 ; Latch et al.,
176 2006), and when there are few migrants per generation ($Nm < 5$), but the ΔK method does not perform as
177 well (Waples & Gaggiotti, 2006). We found both mean probability plots and ΔK were unable to resolve
178 the true number of populations once $Nm \geq 2.5$ ($N = 500$) and $Nm \geq 0.5$ ($N = 100$), with the exception being
179 the set of simulations where $N = 100$ and $n = 20$ (Figure 1). These are very low migration rates that will
180 still result in biologically significant levels of genetic differentiation that make important differences for
181 management, e.g. understanding wildlife disease spread risk. Of course, we recognize that including
182 additional markers, and/or samples will potentially increase resolution above what we have observed here.

183 In our initial work we examined several complex population scenarios using ΔK , but we were
184 unable to determine the source of the $K = 2$ bias (Janes et al., 2017). Using our current data, we found that
185 $K = 2$ is selected more often by the ΔK method than $K = 3$ or 4 when there is only a single genetic
186 population. We also found that $K = 2$ was selected more often by the ΔK method when populations
187 appeared to be in HWE. This indicates that some of the bias likely occurs when population structure is
188 weak, and the most optimal solution is $K = 1$. Additionally, there was a portion of simulations where $K = 3$
189 was highly supported by the mean probability plots, and the ΔK method selected $K = 2$. To identify the
190 values of ΔK where the solution is well supported, we created a logistic regression model. For the two-
191 population scenarios, we found that values of $\Delta K_{\text{Optimal}} > 53$ indicate good support for the solution.
192 However, for the three-population scenarios, this value increased almost ten-fold (~ 400). Based on this,
193 we suggest that researchers examine the magnitude of $\Delta K_{\text{Optimal}}$, the mean probability plots, and use
194 additional information to identify whether population structure is significant (e.g. HWE, exact tests;

Goudet, Raymond, de Meeüs, & Rousset, 1996; Waples, 2015). While F_{ST} is known to be an imperfect statistic (Whitlock & McCauley, 1999), it gives a quantitative measure of connectivity once populations have been identified. As well, using the information contained in the assignment probabilities can be informative in helping to resolve the correct K (Lawson, van Dorp, & Falush, 2018; Puechmaille, 2016). Many of the systems that are studied will likely not have the low levels of migration needed to have large values of ΔK ; therefore, understanding the performance of clustering methods at low levels of differentiation will be important. This is especially true for species with high dispersal capacity (e.g. wind-pollinated plants, marine broadcast spawners, large mammals).

We also note that ΔK results were under-reported in the set of papers we considered (where $K = 2$). Given the lack of reporting, we wonder if researchers feel the value of this estimate is relative only to their particular study and set of parameters, and not broadly informative. However, we found that the magnitude of $\Delta K_{Optimal}$ indicates whether the two populations are strongly differentiated (Figure 4). In the original evaluation of the method, the authors mentioned that the value of ΔK provided a measure of support for the indicated K , but they did not provide a quantitative treatment of the data (Evanno et al., 2005). Here we demonstrate that $\Delta K_{Optimal}$ is directly related to the underlying biology because the allele frequencies used to generate the solution are also used to estimate F_{ST} .

The objective of this work was not to comment on all of the potential caveats to consider using STRUCTURE and the ΔK method, as these have been reviewed in other work (e.g. Gilbert et al., 2012; Kalinowski, 2011; Meirmans, 2019; Puechmaille, 2016; Waples & Gaggiotti, 2006). We also were not aiming to answer the question of what parameters define a population as a separate unit (Waples & Gaggiotti, 2006). Rather, we wanted to provide information to the field that would be useful in assessing the strength of evidence for their clustering solution when using the ΔK method. We recommend that researchers use the information from HWE tests appropriately (Waples, 2015), and report all relevant data, including F_{ST} (or an analogue of genetic differentiation) between their genetic clusters, the value of ΔK for

all K tested, model likelihoods, and admixture values for individuals (e.g. Gilbert et al., 2012; Puechmaille, 2016). We encourage authors to investigate their data additionally with methods that can evaluate $K = 1$ (e.g. principal component analysis) as most admixture methods do not effectively evaluate this. Using such ordination-based methods provides the added benefit for situations where complex population structure exists (Liu & Zhao, 2006; Novembre & Stephens, 2008), or some source populations are missing (Jombart, Devillard, Dufour, & Pontier, 2008). Therefore, a “total evidence” approach, requiring information from other methods and statistics as well as the biology of the study species, will be needed to support a particular value of K . This total evidence approach will provide readers with information regarding the level of support for the clustering solution presented, and the magnitude of differentiation among the clusters.

Acknowledgements

We wish to thank those who have written on this topic before us and inspired us to expand on this conversation, including R. Andrew, G. Evanno, K. Gilbert, P. Meirmans, J. Pritchard, S. Puechmaille, R. Verity and J. Wang. We would like to acknowledge anonymous reviewers for their comments to help improve this work. Finally, we are thankful for opportunities like this one that keep us working together despite being geographically dispersed!

Data accessibility

The outputs from CLUMPAK are freely available on Dataverse (<https://doi.org/10.7939/DVN/PSDJJG>). Custom scripts used for file conversion are available here: https://github.com/rmpeery/deltaK_for_K-2, and parameter inputs for the EASYPOP simulations are included in Table S1.

References

Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). *Conservation and the genetics of populations* (2nd ed.). John Wiley & Sons. Retrieved from <https://www.wiley.com/en-us/Conservation+and+the+Genetics+of+Populations%2C+2nd+Edition-p-9780470671467>

243 Balloux, F. (2001). EASYPOP (Version 1.7): A Computer Program for Population Genetics Simulations.
244 *Journal of Heredity*, 92(3), 301–302. doi:10.1093/jhered/92.3.301

245 Chen, C., Durand, E., Forbes, F., & Francois, O. (2007). Bayesian clustering algorithms ascertaining
246 spatial population structure: a new computer program and a comparison study. *Molecular Ecology*
247 *Notes*, 7(5), 747–756. doi:10.1111/j.1471-8286.2007.01769.x

248 Corander, J., Waldmann, P., & Sillanpää, M. J. (2003). Bayesian Analysis of Genetic Differentiation
249 Between Populations. *Genetics*, 163(1).

250 Dawson, K. J., & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations
251 and the assignment of individuals. *Genetical Research*, 78(1), 59–77.
252 doi:10.1017/S001667230100502X

253 Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the
254 software structure: a simulation study. *Molecular Ecology*, 14(8), 2611–2620. doi:10.1111/j.1365-
255 294X.2005.02553.x

256 Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus
257 genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567–87. Retrieved
258 from <http://www.ncbi.nlm.nih.gov/pubmed/12930761>

259 François, O., Ancelet, S., & Guillot, G. (2006). Bayesian Clustering Using Hidden Markov Random
260 Fields in Spatial Population Genetics. *Genetics*, 174(2), 805–816. doi:10.1534/genetics.106.059923

261 Frankham, R., Ballou, J. D., & Briscoe, D. A. (2002). *Introduction to conservation genetics*. Cambridge,
262 UK: Cambridge University Press. doi:10.2277/0521639859

263 Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., ... Vines, T. H.
264 (2012). Recommendations for utilizing and reporting population genetic analyses: the reproducibility
265 of genetic clustering using the program structure. *Molecular Ecology*, 21(20), 4925–4930.
266 doi:10.1111/j.1365-294X.2012.05754.x

267 Goudet, J. (1995). FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. *Journal of*
268 *Heredity*, 86(6), 485–486. doi:10.1093/oxfordjournals.jhered.a111627

269 Goudet, Jérôme, Raymond, M., de Meeüs, T., & Rousset, F. (1996). Testing Differentiation in Diploid
270 Populations. *Genetics*, 144(4).

271 Guillot, G., Mortier, F., & Estoup, A. (2005). Geneland: a computer package for landscape genetics.
272 *Molecular Ecology Notes*, 5(3), 712–715. doi:10.1111/j.1471-8286.2005.01031.x

273 Huelsenbeck, J. P., Andolfatto, P., & Huelsenbeck, E. T. (2011). Structurama: Bayesian Inference of
274 Population Structure. *Evolutionary Bioinformatics*, 7, EBO.S6761. doi:10.4137/EBO.S6761

275 Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R.
276 L. (2017). The K = 2 conundrum. *Molecular Ecology*. doi:10.1111/mec.14187

277 Jombart, T., Devillard, S., Dufour, A. B., & Pontier, D. (2008). Revealing cryptic spatial patterns in
278 genetic variability by a new multivariate method. *Heredity*, 101(1), 92–103.
279 doi:10.1038/hdy.2008.34

280 Jombart, Thibaut. (2008). adegenet: a R package for the multivariate analysis of genetic markers.
281 *Bioinformatics*, 24(11), 1403–1405. doi:10.1093/bioinformatics/btn129

282 Jombart, Thibaut, Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a
 283 new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1), 94.
 284 doi:10.1186/1471-2156-11-94

285 Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main
 286 genetic clusters within species: simulations and implications for human population structure.
 287 *Heredity*, 106(4), 625–632. doi:10.1038/hdy.2010.95

288 Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak : a
 289 program for identifying clustering modes and packaging population structure inferences across *K*.
 290 *Molecular Ecology Resources*, 15(5), 1179–1191. doi:10.1111/1755-0998.12387

291 Latch, E. K., Dharmarajan, G., Glaubitz, J. C., & Rhodes, O. E. (2006). Relative performance of Bayesian
 292 clustering software for inferring population substructure and individual assignment at low levels of
 293 population differentiation. *Conservation Genetics*, 7(2), 295–302. doi:10.1007/s10592-005-9098-1

294 Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE
 295 and ADMIXTURE bar plots. *Nature Communications*, 9(1). doi:10.1038/s41467-018-05257-7

296 Liu, N., & Zhao, H. (2006). A non-parametric approach to population structure inference using multilocus
 297 genotypes. *Human Genomics*, 2(6), 353–364. doi:10.1186/1479-7364-2-6-353

298 Meirmans, P. G. (2019). Subsampling reveals that unbalanced sampling affects Structure results in a
 299 multi-species dataset. *Heredity*, 122(3), 276–287. doi:10.1038/s41437-018-0124-8

300 Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population
 301 genetic variation. *Nature Genetics*, 40(5), 646–649. doi:10.1038/ng.139

302 Palsbøll, P. J., Bérubé, M., & Allendorf, F. W. (2007). Identification of management units using
 303 population genetic data. *Trends in Ecology & Evolution*, 22(1), 11–16.
 304 doi:10.1016/J.TREE.2006.09.003

305 Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach.
 306 *Bioinformatics*, 26(3), 419–420. doi:10.1093/bioinformatics/btp696

307 Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus
 308 Genotype Data. *Genetics*, 155(2).

309 Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure
 310 when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular*
 311 *Ecology Resources*, 16(3), 608–627. doi:10.1111/1755-0998.12512

312 Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure of marine
 313 fish: mismatch between biological and fisheries management units. *Fish and Fisheries*, 10(4), 361–
 314 395. doi:10.1111/j.1467-2979.2008.00324.x

315 Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCr: visualizing classifier performance
 316 in R. *Bioinformatics*, 21(20), 3940–3941. doi:10.1093/bioinformatics/bti623

317 Waples, R. S. (2015). Testing for Hardy–Weinberg Proportions: Have We Lost the Plot? *Journal of*
 318 *Heredity*, 106(1), 1–19. doi:10.1093/jhered/esu062

319 Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic
 320 methods for identifying the number of gene pools and their degree of connectivity. *Molecular*

321 *Ecology*, 15(6), 1419–1439. doi:10.1111/j.1365-294X.2006.02890.x

322 Whitlock, M. C., & McCauley, D. E. (1999). Indirect measures of gene flow and migration:
323 $F_{ST} \approx 1/(4Nm+1)$. *Heredity*, 82(2), 117–125. doi:10.1046/J.1365-2540.1999.00496.X

324

325 **Author contributions**

326 CIC, JMM, RMP, JRD, RMM, JCG, JKJ helped design the research and write the paper, CIC ran the
327 simulations, CIC, JMM, RMP, JRD, RMM analyzed the data. RMP, JRD, RMM wrote some sweet
328 scripts.

329 Table 1. Optimal cluster solution using mean probability plots (“Mean P(K)”), and the ΔK method for
 330 one- ($K = 1$, $N = 100$), two- ($K = 2$, $N = 2000$), and three-population ($K = 3$, $N = 1600$) simulations. We
 331 categorized simulations into those with and without significant deviation from Hardy-Weinberg
 332 equilibrium (HWE). The columns are the counts of cluster assignments for the simulations using mean
 333 probability, while the rows are the assignments based on ΔK .

		deviation from HWE (p ≤ 0.05)			no deviation from HWE (p > 0.05)			
	ΔK	Mean P(1)	Mean P(2)	Mean P(3)		Mean P(1)	Mean P(2)	Mean P(3)
K = 1	2					51		
	3					25		
	4					24		
K = 2	2	17	612			461	303	
	3	5	1			264	2	
	4	15	1			317	2	
K = 3	2	29	7	113		186	54	54
	3	11	0	798		100	1	145
	4	5	0	2		89	1	5

334

Figure 1. Mean probability plots (left) and ΔK plots (right) from STRUCTURE outputs testing the number of populations ($K = 1$ to 5) across 50 simulations of two populations (indicated by grey bars) created using EASYPOP 1.7. Simulations were generated using 15 microsatellites at four different migration rates (m), two sample sizes (n), and two-population sizes (N) indicated on the right and top (respectively).

Figure 2. Mean probability plots (left) and ΔK plots (right) from STRUCTURE outputs testing the number of populations ($K = 1$ to 5) across 50 simulations of three populations (indicated by grey bars) created using EASYPOP 1.7. Simulations were generated using 15 microsatellites at four different migration rates (m), two sample sizes (n), and two-population sizes (N) indicated on the right and top (respectively).

Figure 3. Predicted probability of the ΔK method selecting the same K -value as the mean probability plots based on the value of $\ln(\Delta K_{\text{Optimal}})$, and our observed data. The model for two-population (top) and three-population simulations (bottom) were run separately. The values of $\ln(\Delta K_{\text{Optimal}})$ where there is model uncertainty are indicated by the vertical blue lines.

Figure 4. Scatterplot of $\ln(\Delta K_{\text{Optimal}})$ versus F_{ST} , for two-population (left) and three-population simulations (right), across simulations of SNPs (orange), microsatellites (blue), the empirical data from the literature was included with the microsatellite data (black), simulated data where $K = 1$ (based on the mean probability plots) is indicated in grey.