Received: 7 June 2017 | Revised: 1 September 2017 | Accepted: 12 September 2017

DOI: 10.1111/1755-0998.12719

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods

Yu-Long Li[1,2] | Jin-Xian Liu[1,2]

[1]CAS Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, Shandong, China

[2]Laboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

Correspondence
Jin-Xian Liu, CAS Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, Shandong, China.
Email: jinxianliu@gmail.com

Funding information
Qingdao National Laboratory for Marine Science and Technology, Grant/Award Number: 2015ASTP-ES05; NSFC-Shandong Joint Fund for Marine Ecology and Environmental Sciences, Grant/Award Number: U1606404; National Natural Science Foundation of China, Grant/Award Number: 41676137.

Abstract

Inferences of population genetic structure are of great importance to the fields of ecology and evolutionary biology. The program STRUCTURE has been widely used to infer population genetic structure. However, previous studies demonstrated that uneven sampling often leads to wrong inferences on hierarchical structure. The most widely used ΔK method tends to identify the uppermost hierarchy of population structure. Recently, four alternative statistics (MEDMEDK, MEDMEAK, MAXMEDK and MAXMEAK) were proposed, which appear to be more accurate than the previously used methods for both even and uneven sampling data. However, the lack of easy-to-use software limits the use of these appealing new estimators. Here, we developed a web-based user-friendly software STRUCTURESELECTOR to calculate the four appealing alternative statistics together with the commonly used Ln Pr($X|K$) and ΔK statistics. STRUCTURESELECTOR accepts the result files of STRUCTURE, ADMIXTURE or FASTSTRUCTURE as input files. It reports the "best" K for each estimator, and the results are available as HTML or tab separated tables. The program can also generate graphical representations for specific K, which can be easily downloaded from the server. The software is freely available at http://lmme.qdio.ac.cn/StructureSelector/.

KEYWORDS
best K, clustering, population genetic structure, Puechmaille method, visualization

1 | INTRODUCTION

Inferring population genetic structure is essential to the fields of ecology, evolution and conservation biology. The program STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) is the most widely used and cited software to detect population genetic structure. However, this program introduces the problem of choosing the best number of genetic clusters (K). STRUCTURE performed well in recovering the correct number of clusters when the sample size is even, but the program does not reliably recover the correct population structure when sampling is uneven between subpopulations and/or hierarchical levels of population structure exist (Puechmaille, 2016). The commonly used method of identifying the "optimal" number of clusters introduced by Evanno, Regnaut, and Goudet (2005) (also known as ΔK method) is more likely to identify the uppermost level of hierarchical population structure, thus will lead to underestimating of structure. While uneven sampling and hierarchical population structure are common in empirical studies, and misidentification of population structure could result in putting wildlife populations at risk, careful considerations should be taken by researchers when choosing the right number of genetic clusters based on STRUCTURE results.

Puechmaille (2016) proposed four new supervised estimators, MEDMEDK, MEDMEAK, MAXMEDK and MAXMEAK, which are based on the count of the number of clusters that are contained in at least one subpopulations (e.g., sampling location/region). These new estimators were found to be more accurate than the Ln Pr($X|K$) and ΔK method on both evenly and unevenly sampled datasets (Puechmaille, 2016), thus providing appealing alternatives for the selection of optimal

number of genetic clusters. These estimators can also be applied to other genetic structure programs that provide membership coefficients (Q) (Puechmaille, 2016), such as ADMIXTURE (Alexander, Novembre, & Lange, 2009) and FASTSTRUCTURE (Raj, Stephens, & Pritchard, 2014). While these new estimators are appealing alternatives, they have not been tested widely yet. Although the author provided a useful R script along with a manual to calculate these estimators, it might be difficult for some researchers to run the script. With the advance of the next-generation sequencing, rather than STRUCTURE, both ADMIXTURE and FASTSTRUCTURE have been widely adopted in population genomic studies. However, no software is available to calculate the estimators of Puechmaille (2016) for both softwares. So, a user-friendly software, which could significantly stream the process of obtaining these estimators, is urgently needed.

In this study, we developed a web-based software STRUCTURESELECTOR to calculate these new estimators easily and assist in the selection and visualization of the "best" *K*. In addition to MEDMEDK, MEDMEAK, MAXMEDK and MAXMEAK, STRUCTURESELECTOR can also calculate the commonly used Ln Pr(*X*|*K*) (Pritchard et al., 2000) and Δ*K* estimators for standard output results of STRUCTURE. STRUCTURESELECTOR also adopted the choose *K* algorithm (Raj et al., 2014) in FASTSTRUCTURE for choosing model complexity. These measures all combined should help researchers to choose the "best" *K* that fits the data once considering the biological meanings. Furthermore, STRUCTURESELECTOR can generate graphical representations of the results by integrating the CLUMPAK program (Kopelman et al., 2015).

STRUCTURESELECTOR accepts the results of standard STRUCTURE, ADMIXTURE, FASTSTRUCTURE or other STRUCTURE-like program that produces Q-matrices as input file, which should be compressed into zip format. In addition to results of a single dataset, STRUCTURESELECTOR can also accept results of multiple datasets, which can be put either in different subfolders or all together in one folder (recognized by file names specific to each dataset). When calculating estimators of Puechmaille (2016), users can input multiple threshold values separated by";"at once. Different grouping options are also available by uploading different popmap files or different vectors of grouping sizes. Results are available as HTML or tab separated tables and figures, and the plots of each estimator are generated by R (R Core Team 2017), which are suitable for further publications. The program can also generate graphical representations of specific *K* by running CLUMPAK on the selected *K*, which can be easily downloaded from the server. This allows easy data submission, quick visualization and rapid import of graphical plots into scientific works. The software could be accessed at http://lmme.qdio.ac.cn/StructureSelector/. The main programs were written in Perl, and source codes are available from the author upon request.

ACKNOWLEDGEMENTS

We thank for Dr. Sébastien J. Puechmaille and another anonymous reviewer for their useful suggestions on the first release of the program. This study was supported by the AoShan Talents Program supported by Qingdao National Laboratory for Marine Science and Technology (No. 2015ASTP-ES05), the NSFC-Shandong Joint Fund for Marine Ecology and Environmental Sciences (No. U1606404) and a grant from the National Natural Science Foundation of China (No. 41676137).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

The study was designed by L.Y.L. and L.J.X.; program was written by L.Y.L.; manuscript was written by L.Y.L. and L.J.X.

DATA ACCESSIBILITY

The program, instructions and example datasets are available on the website (http://lmme.qdio.ac.cn/StructureSelector/).

ORCID

Jin-Xian Liu http://orcid.org/0000-0002-0756-2984

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664.

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

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*, 1179–1191.

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

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*, 608–627.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FASTSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*, 573–589.

How to cite this article: Li Y-L, Liu J-X. STRUCTURESELECTOR: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol Ecol Resour*. 2018;18:176–177. https://doi.org/10.1111/1755-0998.12719