TECHNICAL NOTE

STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method

Dent A. Earl · Bridgett M. vonHoldt

Received: 19 May 2011 / Accepted: 28 September 2011 / Published online: 13 October 2011

Abstract We present STRUCTURE HARVESTER (available at http://taylor0.biology.ucla.edu/structureHarvester/), a web-based program for collating results generated by the program STRUCTURE. The program provides a fast way to assess and visualize likelihood values across multiple values of *K* and hundreds of iterations for easier detection of the number of genetic groups that best fit the data. In addition, STRUCTURE HARVESTER will reformat data for use in downstream programs, such as CLUMPP.

Keywords Structure · Population structure · Population genetics · Evanno method · Visualization · Clustering

Introduction

The program STRUCTURE (Pritchard et al. 2000) has become one of the most widely used programs by population geneticists to assess the level of genetic stratification in a multi-locus data set. Users can assess the likelihood values of partitioning their data into different numbers of clusters (*K*), with various updates to the model's assumptions previously discussed (e.g. Falush et al. 2003; Falush et al. 2007; Hubisz et al. 2009; François and Durand 2010). Inferring the number of clusters, *K*, that best fit the data is an important challenge that has been previously reviewed with many ad hoc methods available (Pritchard et al. 2000; Evanno et al. 2005; Latch et al. 2006; Hubisz et al. 2009; Schwartz and McKelvey 2009). Moreover, parsing STRUCTURE results can be cumbersome and time-consuming, especially when a larger number of replicates (>10) are analyzed over many successive values of *K*. Additional data parsing and formatting is required if the user wants to subsequently analyze STRUCTURE results with CLUMPP (Jakobsson and Rosenberg 2007) or DISTRUCT (Rosenberg 2004). CLUMPP aligns cluster assignment across replicate analyses, while DISTRUCT aids in the visual representation of the aligned cluster assignments.

There is a clear need for a program to quickly collate results from STRUCTURE, visualize likelihood scores to assist the user in delineating the most likely level of population subdivision, and format the output for use in downstream programs. The program STRUCTURE HARVESTER is a Python program with a web-based front-end for quickly parsing and summarizing output data from STRUCTURE.

STRUCTURE HARVESTER processes STRUCTURE results, generates in-files for use with CLUMPP, and when possible, executes the "Evanno" method (Evanno et al. 2005). It then produces a plot of the mean likelihood values per *K* and a tab-delimited table of the Evanno results for use in downstream programs such as Microsoft Excel. STRUCTURE HARVESTER runs on any computer with either (1) a web browser or (2) Python installed (for the stand-alone version). The web version provides a simple and easy to use graphical user interface for managing in and out-files.

Other parsers exist for STRUCTURE, including CorrSieve (Campana et al. 2011), though we know of no other

D. A. Earl (✉)
Center for Biomolecular Science and Engineering, Biomolecular Engineering Department, University of California, Santa Cruz, Mail Stop: CBSE/ITI; 1156 High Street, Santa Cruz, CA 95064, USA
e-mail: dearl@soe.ucsc.edu

B. M. vonHoldt
Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA 92697, USA

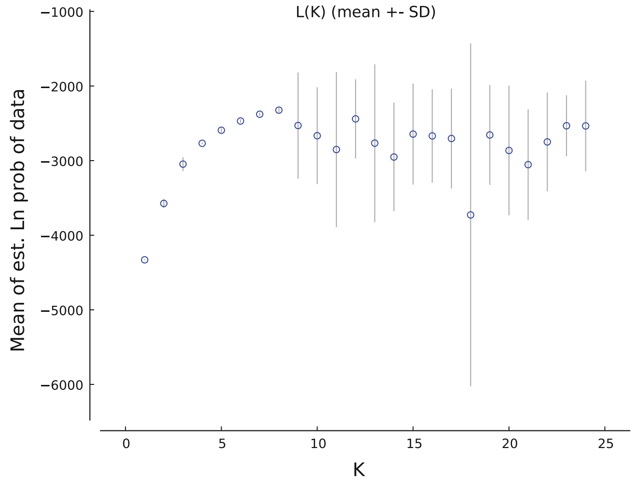


Fig. 1 Plot of mean likelihood L(K) and variance per *K* value from STRUCTURE on a dataset containing 95 individuals genotyped for 19 polymorphic microsatellite loci. The data displayed is the full dataset from Earl et al. (2010)

web-based parsing program. Web-based software solutions allow the user to always use the latest version of the software when visiting the website, incorporating all patches and feature enhancements, without having to bother with complicated installation and dependency issues. For users that need offline access to the parser we provide an offline version that is kept functionally synchronized with the web-based program. The user can obtain the standalone version at http://users.soe.ucsc.edu/~dearl/software/structureHarvester/ under an MIT license.

Functionality description

On the homepage, the user can upload a compressed file of the STRUCTURE output files (filename_f files, found in the STRUCTURE “Results” directory) and STRUCTURE HARVESTER will perform the analyses and direct the user to a dynamically generated results page. All generated content is compressed into a single archive (tar gzip) for downloading and a link is provided at the top of the result page. Results are stored on the server for one week. The first plot on a results page is the mean likelihoods per *K* value, including standard deviation bars to display likelihood variance (Fig. 1). All images are produced in PNG, PDF, and EPS formats, allowing easy generation of publication-quality figures for inclusion in reports or manuscripts. Additionally, CLUMPP in-files (both indfiles and popfiles) for each *K* value are produced as links for the user to download.

The algorithm employed by STRUCTURE HARVESTER to determine if the Evanno method can be performed requires that at least three values of sequential

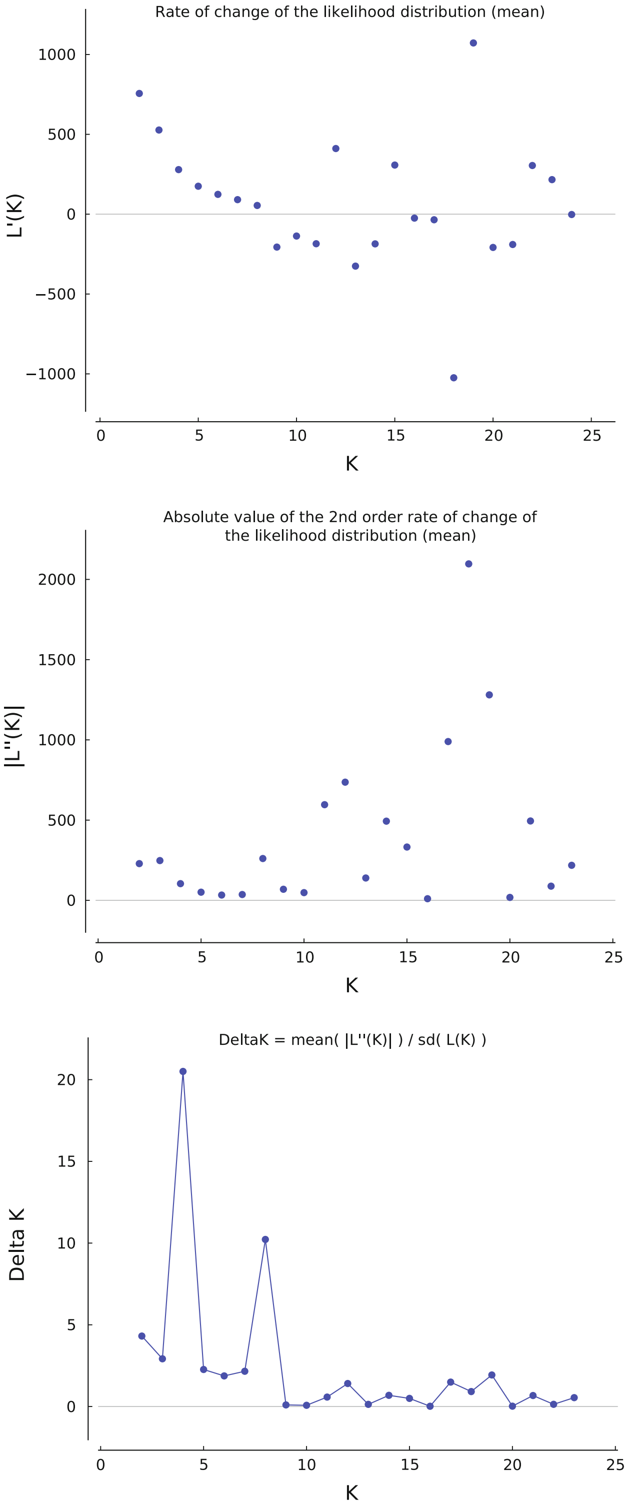
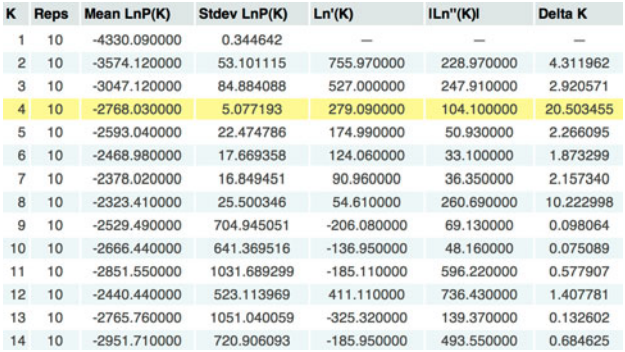


Fig. 2 Evanno et al. (2005) plots for detecting the number of *K* groups that best fit the data. Data as in Fig. 1

K were analyzed with at least 3 replicates and that the sample standard deviation of the log likelihood values across all *K* values is non-zero. The last requirement can occasionally be violated when small numbers of replicates



K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	ILn''(K)I	Delta K
1	10	-4330.090000	0.344642	–	–	–
2	10	-3574.120000	53.101115	755.970000	228.970000	4.311962
3	10	-3047.120000	84.884088	527.000000	247.910000	2.920571
4	10	-2768.030000	5.077193	279.090000	104.100000	20.503455
5	10	-2593.040000	22.474786	174.990000	50.930000	2.266095
6	10	-2468.980000	17.669358	124.060000	33.100000	1.873299
7	10	-2378.020000	16.849451	90.960000	36.350000	2.157340
8	10	-2323.410000	25.500346	54.610000	260.690000	10.222998
9	10	-2529.490000	704.945051	-206.080000	69.130000	0.098064
10	10	-2666.440000	641.369516	-136.950000	48.160000	0.075089
11	10	-2851.550000	1031.689299	-185.110000	596.220000	0.577907
12	10	-2440.440000	523.113969	411.110000	736.430000	1.407781
13	10	-2765.760000	1051.040059	-325.320000	139.370000	0.132602
14	10	-2951.710000	720.906093	-185.950000	493.550000	0.684625

Fig. 3 Table output of the Evanno method results. Yellow highlight is performed dynamically on the website and shows the largest value in the Delta *K* column. Data as in Fig. 1

are used and if, by chance, all runs generate the same estimated log likelihood values. If the conditions are met, then the program will perform the Evanno method for detecting the number of *K* groups that best fit the dataset and produce three plots equivalent to that of Fig. 2 in Evanno et al. (2005) (Fig. 2). Data for the three Evanno plots are provided in a tab-delimited table for the user to download (Fig. 3). Additionally, there is a section on the website to address frequently asked questions regarding the implementation of the Evanno method, specifically highlighting the exact Python code that performs the analysis.

Acknowledgments We thank J Goudet for improving our implementation of the Evanno method (personal communications). We thank C Taylor at University of California at Los Angeles for hosting the website. This work was supported by the National Cancer Institute grants 1U24CA143858-01 and 5R21CA135937.

References

Campana MG, Hunt HV, Jones H, White J (2011) CorrSieve: software for summarizing and evaluating Structure output. Mol Ecol Res 11:349–352

Earl DA, Louie KD, Bardeleben C, Swift CC, Jacobs DK (2010) Rangewide microsatellite phylogeography of the endangered tidewater goby, *Eucyclogobius newberryi* (Teleostei: Gobiidae), a genetically subdivided coastal fish with limited marine dispersal. Conserv Genet 11:103–114

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

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578

François O, Durand E (2010) Spatially explicit Bayesian clustering models in population genetics. Mol Ecol Res 10:773–784

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Res 9:1322–1332

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet 7:295–302

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

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol 4:137–138

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conserv Genet 10:441–452