

QStrap: bootstrapping confidence intervals around admixture estimates from STRUCTURE output

Version 3.1

January 4, 2008

Kevin B. Rogers, Aquatic Research, Colorado Division of Wildlife, PO Box 775777, Steamboat Springs, CO 80477, kevin.rogers@state.co.us

Overview

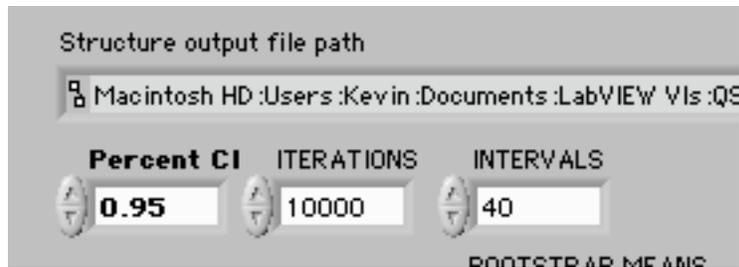
Evaluating taxonomic relationships with microsatellite or AFLP non-diagnostic markers has generated a need for more complex analysis software to evaluate population structure with genotype data. One program that has seen increasing popularity for making these determinations is Program STRUCTURE (Falush et al. 2007; Pritchard et al. 2007). While the program easily generates average admixture coefficients (q) to characterize levels of admixture, it does not provide confidence intervals around those values to yield insight into the precision of the estimated q . QStrap uses the estimated q values from STRUCTURE runs to bootstrap confidence intervals around admixture values.

Installation

To install off the application CD, simply double click on the “setup.exe” icon and follow the instructions. To download QStrap off the web, visit <http://wildlife.state.co.us/Research/Aquatic> and click on the software tab in the left dropdown. Scroll down to the QStrap description and click on the appropriate link.

Running QStrap

QStrap can be invoked by selecting the program from the START menu (PC), or double clicking on the application (MacOSX). Once open, the percent confidence interval, number of bootstrap iterations, and bin intervals used for plotting should be entered.



The default unknown reference population is set at position six, since that is the most common position for our cutthroat trout taxonomy work (five reference taxa and one unknown). If the program is used for other applications, the following dropdown should reflect the population number of the unknown as indicated in the STRUCTURE output below:

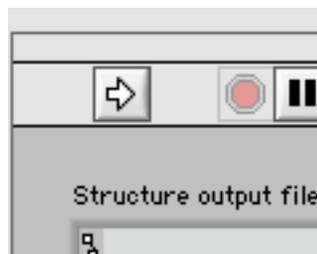


```

Inferred ancestry of individuals:
Probability of being from assumed population | prob of other pops
Label (%Miss) Pop
1 BBC-62326 (0) 1 : 1.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 | Pop 5: 0.000 |
2 BBC-62327 (0) 1 : 1.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 | Pop 5: 0.000 |
3 BBC-62328 (0) 1 : 1.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 | Pop 5: 0.000 |
4 BRC_62329 (0) 1 : 1.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 | Pop 5: 0.000 |

243 CRR-46xxx (0) 5 : 1.000 | Pop 1: 0.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 |
244 CRR-46xxx (0) 5 : 1.000 | Pop 1: 0.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 |
245 CRR-46xxx (0) 5 : 1.000 | Pop 1: 0.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 |
246 CRR-46xxx (0) 5 : 1.000 | Pop 1: 0.000 | Pop 2: 0.000 | Pop 3: 0.000 | Pop 4: 0.000 |
247 WF9-66959 (0) 6 : 0.997 0.000 0.001 0.002 0.000
248 WF9-66960 (0) 6 : 0.998 0.001 0.001 0.001 0.000
249 WF9-66961 (0) 6 : 0.997 0.001 0.000 0.001 0.000
250 WF9-66962 (0) 6 : 0.994 0.001 0.001 0.003 0.000
251 WF9-66963 (0) 6 : 0.998 0.001 0.001 0.001 0.000
  
```

QStrap is then run by clicking on the right facing arrow on the upper left portion of the application window.



Selecting appropriate STRUCTURE output

Once the run arrow is selected, QStrap will prompt the user for the file to analyze. The current Pisces protocol calls for generating bootstrapped confidence intervals around q-values generated in the STRUCTURE run with the highest log likelihood (best of ten runs).

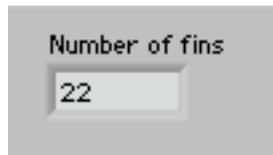
Implementation

QStrap uses a basic bootstrapping approach to derive confidence intervals around mean values of q. Lucid discussions of this approach can be found in Efron and Tibshirani (1986) and Manly (1997). Basically, a sample of n q-values are drawn with replacement from the pool of q-values generated by STRUCTURE for the unknown population of interest based on known reference populations with which admixture is to be evaluated. This exercise is iterated a user specified number of times (default set at 10,000), and the ordered mean values are plotted. The bounds of the middle 95% of values (or whatever other user specified confidence interval is entered) then reflect the upper and lower confidence limits around the mean that is presented.

Interpretation

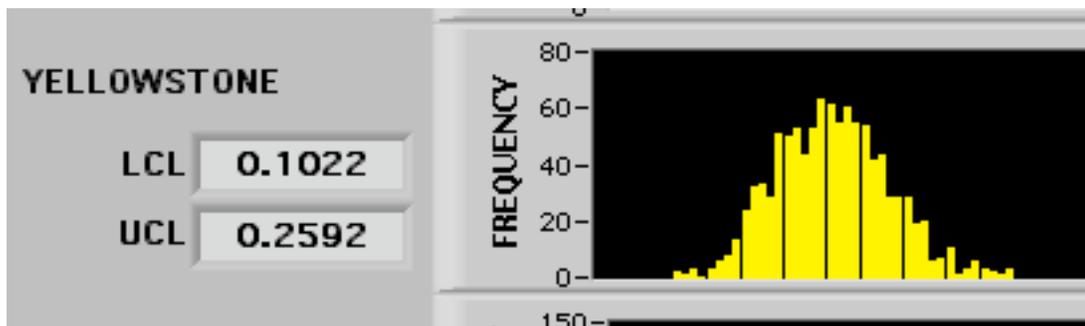
Sample size

The number of individuals (n) used in the analysis is displayed in the upper right portion of the window



Probability plot

The resulting mean q-values from every iteration are plotted for each subspecies so that relative admixture can be assessed. In addition, the confidence intervals that bound the estimated average q-value are presented for each subspecies to the left of the plot



Output table

A table of the individual q-values derived from the STRUCTURE output with the maximum likelihood is displayed to ensure that the program is reading the appropriate information. This table also displays the mean and confidence interval bounds generated from the bootstrapped procedure, and should look something like this:

Output File					
Sample	qCRCT	qGBCT	qRGCT	qYSCT	qCRR
BCL-68275	0.913	0.000	0.000	0.086	0.000
BCL-68276	0.748	0.001	0.001	0.250	0.000
BCL-68277	0.998	0.000	0.000	0.000	0.000
BCL-68278	0.702	0.001	0.001	0.295	0.001
BCL-68279	0.913	0.001	0.001	0.085	0.001
BCL-68280	0.830	0.002	0.004	0.164	0.001
BCL-68281	0.794	0.002	0.002	0.202	0.000
BCL-68282	0.929	0.001	0.001	0.069	0.001
BCL-68283	0.423	0.001	0.001	0.575	0.000
BCL-68284	0.633	0.001	0.001	0.364	0.001
BCL-68285	0.992	0.001	0.001	0.006	0.000
BCL-68286	0.288	0.002	0.001	0.706	0.002
BCL-68287	0.919	0.000	0.001	0.080	0.000
BCL-68288	0.995	0.000	0.001	0.004	0.000
BCL-68289	0.920	0.001	0.001	0.077	0.001
BCL-68290	0.996	0.000	0.000	0.002	0.001
BCL-68291	0.455	0.001	0.003	0.531	0.010
BCL-68292	0.970	0.001	0.001	0.027	0.001
BCL-68293	0.954	0.000	0.001	0.044	0.001
BCL-68294	0.997	0.001	0.001	0.001	0.000
BCL-68295	0.996	0.001	0.002	0.001	0.000
BCL-68296	0.760	0.001	0.005	0.235	0.000
Population Average q	0.82	0.00	0.00	0.17	0.00
Lower 95% CL	0.74	0.00	0.00	0.10	0.00
Upper 95% CL	0.90	0.00	0.00	0.26	0.00

Literature cited

- Falush, D., M. Stephens, and J. K. Pritchard. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*.
- Pritchard, J. K., X. Wen, and D. Falush. 2007. Documentation for structure software: Version 2.2. University of Chicago. Available: <http://pritch.bsd.uchicago.edu/software>. (June 2007).
- Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1:54-77
- Manly, B. J. F. 1997. Randomization, bootstrap, and Monte Carlo methods in biology, 2nd edition. Chapman and Hall, London.