POPHELPER: an R package and web app to analyse and visualize population structure

R. M. FRANCIS
Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

Abstract

The POPHELPER R package and web app are software tools to aid in population structure analyses. They can be used for the analyses and visualization of output generated from population assignment programs such as ADMIXTURE, STRUCTURE and TESS. Some of the functions include parsing output run files to tabulate data, estimating K using the Evanno method, generating files for CLUMPP and functionality to create barplots. These functions can be streamlined into standard R analysis workflows. The latest version of the package is available on GITHUB (https://github.com/royfrancis/pophelper). An interactive web version of the POPHELPER package is available which covers the same functionalities as the R package version with features such as interactive plots, cluster alignment during plotting, sorting individuals and ordering of population groups. The interactive version is available at http://pophelper.com/.

Keywords: ADMIXTURE, CLUMPP, DISTRUCT, population structure, STRUCTURE, TESS

Received 19 October 2015; revision received 12 January 2016; accepted 16 January 2016

Introduction

The use of multilocus molecular markers to differentiate populations and to infer their genetic structure and composition is a powerful tool in ecology, conservation and landscape genetics (Sunnucks 2000; Falush et al. 2003). One of the most successful approaches employed is the assignment of individuals (Manel et al. 2005) which includes distance-based methods (Piry et al. 2004), frequency-based methods (Paetkau et al. 1995), Bayesian methods (Rannala & Mountain 1997; Baudouin & Lebrun 2000) and those involving Markov chain Monte Carlo methods (Cornuet et al. 1999; Pritchard et al. 2000; Paetkau et al. 2004). Several programs have been published in the recent years with various features and differences to investigate the population structure. Some of the softwares discussed in this article includes ADMIXTURE (Alexander et al. 2009), FASTSTRUCTURE (Raj et al. 2014), STRUCTURE (Pritchard et al. 2000) and TESS (François & Durand 2010). Along with assignment of individuals, it is often useful to cluster individuals into K populations. Estimation of the number of populations (K) is often a tricky problem, and no standard method exists that can be applied in all situations. The approach by Evanno et al. (2005) is one method to estimate the value of K and has been widely cited (Morgan et al. 2007; Blackburn & Maddison 2014; Diez et al. 2015).

With STRUCTURE, the input formats have been made convenient for users using data conversion tools such as GENALEX (Peakall & Smouse 2012), but the handling of output files have not been implemented in a convenient manner. Both of these programs also lack functionalities to generate high-quality graphics. Two common downstream approaches are to align assignment clusters across replicate runs using CLUMPP (Jakobsson & Rosenberg 2007) and to visualize the output using DISTRUCT (Rosenberg 2004). The use of these programs is often difficult and cumbersome. Therefore, a program to easily and quickly accept run files and generate outputs that can be directly used in downstream programs was a necessity.

The most popular of these 'helper' programs was STRUCTURE HARVESTER (Earl 2012). STRUCTURE HARVESTER is a web-based utility with a graphical user interface that can accept STRUCTURE runs to generate Evanno plots and input files for CLUMPP, but does not produce plots or allow to work with CLUMPP output. A more recent development was STRUCTURE PLOT (Ramasamy et al. 2014), which uses R shiny web framework (Chang et al. 2015) to interactively work in the browser. Some of the limitations of STRUCTURE PLOT include the lack of parsing run files, data tabulation, cluster alignment, CLUMPP export or Evanno method calculation. These programs also do not handle input from multiple softwares.

The statistical programming language R (R Development Core Team 2013) is widely used in data analysis
and its use in ecology and population analysis has risen over the years. The program \textsc{pophelper} presented in this article is provided as an \textit{R} package for command-line use and as a web app for interactive use. Many of the limitations of previously discussed applications have been rectified in this package which include parsing run files from various softwares, tabulation of runs, plotting bar-plots with population labels, sorting of individuals and ordering of populations, alignment of clusters and merging of repeats using \textsc{clumpp}, etc.

Features

The latest version of \textsc{pophelper} \textit{R} package can be installed from the Github repository (https://github.com/royfrancis/pophelper) with installation instructions provided. A brief introduction on functions, usage and workflow are provided on the github page. A detailed demonstration is provided as an \texttt{html} vignette along with the package accessible from the github page. The \textsc{pophelper} package employs the \textsc{ggplot} (Wickham 2009) graphing package to create high-quality graphics. The \textsc{pophelper} package currently accepts \textsc{admixture}, \textsc{faststructure}, \textsc{structure} and \textsc{teSS} runs as input. Files from other sources or modified data can be used as input as long as they are all numeric, tabular data delimited by white space, tab or comma. The files must be free of headers and decimal must be defined by dot.

The common scenario with most users is that they have a directory full of run files. \textsc{pophelper} provides

Fig. 1 A screenshot of the \textsc{pophelper} GUI for Evanno method using \textsc{structure} runs. The table output is shown on the top, and the plot is shown below. The plot shows a typical Evanno analysis with (A) estimated log probability of data of runs over increasing values of $K$, (B) first derivative, (C) second derivative and (D) $\Delta K$ over values of $K$. The same plot is also produced in the \textit{R} package version.
functions to tabulate all runs listing the files and various parameters such as filename, $K$ value, number of individuals and to summarize the runs by number of repeats. TESS files generated in separate directories can be easily collated into a single directory for further processing. ADMIXTURE, FASTSTRUCTURE, STRUCTURE, TESS or any tabular run files can be converted to R data frames for further analysis. The Evanno method to detect $K$ has been implemented for STRUCTURE runs (Fig. 1). This method requires at least three sequential values of $K$ with minimum of two repeats. Informative feedback is provided in cases where the data does not fulfil the requirements. The results are available as plots as well as text files. Input files for CLUMPP (combined file with repeats and the paramfile) are easily generated into separate directories. The CLUMPP executable can be used to generate the aligned and merged files in each $K$ directory. The aligned and merged files can be collated into a single directory using POPHELPER. CLUMPP files can be generated from any of the input run files.

Barplots (Fig. 2) can be created from any of the input run files, combined files, aligned files or
Fig. 3 Barplots created from structure runs. Three runs ($K = 2$, $K = 3$ and $K = 4$) are joined together as a single plot with common grouped labels. The filenames and $K$ values are shown on the right-side panel. Left: shows the input order of individuals. Right: shows individuals sorted by all clusters (similar to the ‘Sort by Q’ option in structure software).

Fig. 4 Barplots created from structure runs. Three repeats for $K = 3$ are joined together as a single plot with common grouped labels. Top left: shows the input in the original order. Top right: shows clusters aligned using clumpp. Bottom left: shows individuals sorted by ‘all’ clusters (similar to the ‘Sort by Q’ option in structure software) within population groups after aligning using clumpp. Bottom right: shows reordered population groups (Pop B before Pop A) after aligning using clumpp.
merged files. The plots can be created separately or as joined plots, with or without population labels (Fig. 3). The combined, aligned and merged files from CLUMPP can also be plotted (Fig. 4). Individuals in the barplot can be sorted by any one of clusters or a clustering that takes all clusters into account (similar to the ‘Sort by Q’ option in STRUCTURE software). When using population labels, population groups can be subsetted or reordered (Fig. 4). When using population labels along with sorting, individuals are sorted within the population groups. Multiline plots (similar to that implemented in STRUCTURE software) can be created from any of the previously mentioned file types to identify individuals. Multiline plots (Fig. 5) produce an A4 size image with a certain number of rows and certain number of individuals per row depending on the number of individuals. A reasonable default is calculated, but this can be adjusted. Multiline plots can be sorted by one or all clusters. The plot functions have several arguments to tweak and customize the figure as required. Export formats include JPEG (lossy), PNG (lossless) or PDF (vector). The package uses simple commands and basic R functions enabling access to beginner R users.

The web implementation of POPHELPER is available at https://pophelper.com. The web app was created in R using shiny (Chang et al. 2015) web framework and is suitable for users who wish to work interactively. The app makes use of several R packages such as fields (Nychka et al. 2015), reshape2 (Wickham 2011), markdown (Allaire et al. 2015), tibble (Wickham 2012), gridExtra (Auguie 2015), plyr (Wickham 2011), shinyace (LLC. Ta 2013), rCharts (Vaidyanathan 2013) and shinyjs (Attali 2015). The web app has similar features to the R package. Interactive graphics are available for data exploration, and standard graphics (Fig. 2) are available for print and publication. The web app has an implementation of the Evanno method (Fig. 1) for determining K as well as cluster alignment during plotting using CLUMPP. Population labels can be uploaded as a text file or copy-pasted to mark populations under the barplot. Individuals can be sorted by one or more clusters. Population groups can be subsetted or reordered. As the cluster alignment is performed during plotting, CLUMPP and related files such as combined and aligned files are not required to be handled by the user. Population clusters can be coloured using a wide range of colour palettes including colour-blind friendly palettes. Options are provided to adjust the size, colour and position of most chart elements as well as population labels.

Detailed description of functionality and tutorials are available with the application (R package as well as the web app). The POPHELPER package is expected to be frequently updated, and new functionalities will be

---

**Fig. 5** Multiline barplots created from a STRUCTURE run file containing 149 individuals and K = 2. Top: shows individuals in the original order. Bottom: shows individuals sorted by ‘all’ clusters.
added as and when possible. It is hoped that the R package will be useful for implementation into analysis workflows and the web app for interactive use.

Conclusion
The pophelper R package and web app are available to assist users working with molecular markers to investigate population structure. The R package is easy to use and possible to fit into workflows and data analysis pipelines. The interactive web app is available for applied users who wish to use a graphical user interface.

Acknowledgement
I wish to thank Per Kryger for the inspiration to embark on this project. I wish to thank the enthusiastic R community for the numerous tools and packages made available.

References
Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19, 1655–1664.
Allaire J, Horner J, Martí V, Porte N (2015) markdown: ‘MarkDown’ Rendering for R. R package version 0.7.7. http://CRAN.R-project.org/package=markdown
Attali D (2015) shinyjs: Perform Common JavaScript Operations in Shiny Apps using Plain R Code. R package version 0.2.0. http://CRAN.R-project.org/package=shinyjs
Auguie B (2015) gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.0.0. http://CRAN.R-project.org/package=gridExtra
Baudouin L, Lebrun P (2000) An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers, 81–93.
Blackburn GS, Maddison WP (2014) Stark sexual display divergence among jumping spider populations in the face of gene flow. Molecular Ecology, 23, 5208–5223.
Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2015) shiny: Web Application Framework for R. http://CRAN.R-project.org/package=shiny
Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics, 153, 1989–2000.
Dey CM, Trujillo I, Martinez-Urdioz N et al. (2015) Olive domestication and diversification in the Mediterranean Basin. New Phytologist, 206, 436–447.
Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4, 359–361.
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.
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.
François O, Durand E (2010) Spatially explicit Bayesian clustering models in population genetics. Molecular Ecology Resources, 10, 773–784.
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.
LLC. TTA (2013) shinyAce: Ace editor bindings for Shiny. R package version 0.1.0. http://CRAN.R-project.org/package=shinyAce
Mandel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology & Evolution, 20, 136–142.
Morgan JA, Vredenburg VT, Rachowicz LJ et al. (2007) Population genetics of the frog-killing fungus Batrachochytrium dendrobatidis. Proceedings of the National Academy of Sciences of the United States of America, 104, 13845–13850.
Nyck A, Furrer R, Sain S (2015) fields: Tools for Spatial Data. R package version 8.2-1. http://CRAN.R-project.org/package=fields
Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology, 4, 347–354.
Paetkau D, Slade R, Burden M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Molecular Ecology, 13, 55–65.
Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics, 28, 2537–2539.
Piry S, Alapetite A, Cornuet JM et al. (2004) GENECLASS2: a software for genetic assignment and first-generation migrant detection. Journal of Heredity, 95, 536–539.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.
R Development Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/
Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 197, 573–589.
Ramassamy RK, Ramasamy S, Bindroo BB, Naik VG (2014) STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus, 3, 431.
Rannala B, Mountain J (1997) Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America, 94, 9197–9201.
Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Molecular Ecology Notes, 4, 137–138.
Summucks P (2000) Efficient genetic markers for population biology. Trends in Ecology & Evolution, 15, 199–203.
Vaidyanathan R (2013) rCharts: Interactive Charts using Javascript Visualization Libraries. R package version 0.4.5.
Wickham H (2007) Reshaping Data with the reshape Package. Journal of Statistical Software, 21, 1–20.
Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer, New York, New York.
Wickham H (2011) The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40, 1–29.
Wickham H (2012) gttable: Arrange grobs in tables. R package version 0.1.2. http://CRAN.R-project.org/package=gttable
Xie Y (2015) DT: A Wrapper of the JavaScript Library ‘DataTables’. R package version 0.1. http://CRAN.R-project.org/package=DT

R.M.F. wrote the code for the program and the manuscript.

Data accessibility
The latest version of the package is available on Github (https://github.com/royfrancis/pophelper).