Bioinformatics, 38(7), 2022, 2070–2071
https://doi.org/10.1093/bioinformatics/btac046
Advance Access Publication Date: 26 January 2022
Applications Note

OXFORD

Genetics and population analysis

Struct-f4: a Rcpp package for ancestry profile and population structure inference from f_4-statistics

Pablo Librado * and Ludovic Orlando

Centre for Anthropobiology and Genomics of Toulouse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, 31000 Toulouse, France

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 16, 2021; revised on January 14, 2022; editorial decision on January 17, 2022

Abstract

Summary: Visualization and inference of population structure is increasingly important for fundamental and applied research. Here, we present Struct-f4, providing automated solutions to characterize and summarize the genetic ancestry profile of individuals, assess their genetic affinities, identify admixture sources and quantify admixture levels.

Availability and implementation: Struct-f4 is written in Rcpp and relies on f_4-statistics and Markov Chain Monte Carlo (MCMC) optimization. It is freely available under GNU General Public License in Bitbucket (https://bitbucket.org/plibradosanz/structf4/).

Contact: plibradosanz@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-generation sequencing has opened for the routine characterization of genome variation at the population scale, including in non-model organisms, which provides invaluable insights into evolution (Nielsen *et al.*, 2017). Enhanced characterization of population structure has also found a range of applications in medicine, forensics, conservation biology and more. Many statistical methods are available to visualize (e.g. Principle Component Analysis; Patterson *et al.*, 2006) and model population structure, e.g. as combinations of K ancestry clusters (e.g. ADMIXTURE; Alexander *et al.*, 2009). However, these methods can be biased by the amount of genetic drift exclusive to single populations (Lawson *et al.*, 2018). Other methods based on shared drift aimed to overcome such limitations and increasingly gained popularity in the last decade (Patterson *et al.*, 2012). For example, qpGraph and qpAdm leverage patterns of allele sharing between population quartets (the so-called f_4-statistics) to model evolutionary histories, including admixture coefficients. This methodology is, however, highly supervised through the specification of homogeneous groups, potentially acting as admixture sources, and requires to assess alternative models individually. This becomes practically challenging as the number of populations and/or admixture events increase, rapidly exceeding the current capacity of automated solutions (Leppälä *et al.*, 2017).

To remediate this situation, we developed Struct-f4, a package leveraging the power of f_4-statistics and automating statistical inference within an MCMC framework. Struct-f4 first estimates the shared drift across pairs of individuals, allowing visualization of population structure through Multi-Dimensional Scaling (MDS). It also models individual genetic profiles as mixtures from K ancestral populations, not assumed to follow Hardy–Weinberg equilibrium, and accommodates both supervised and unsupervised analyses.

2 Materials and methods

Struct-f4 was originally proposed by Fages *et al.* (2019) to visualize the genetic structure within ancient and modern horse populations. The methodology involved maximum likelihood optimization to place individuals within the 3D-Euclidean space that best fits the observed combination of f_4-statistics. Here, we redesigned the underlying statistical model to retrieve direct estimates of the shift in allele frequency that occurred between pairs of individuals, as follows:

$$f_4(\mathrm{H1},\ \mathrm{H2};\ \mathrm{H3},\ \mathrm{H4}) = (p_{H1} - p_{H2})(p_{H3} - p_{H4}) = d_{H1H2}d_{H3H4},$$

where d_{H1H2} the difference in allele frequency between individuals H_1 and H_2. Assuming f_4-statistics follow normal distributions, the likelihood of the d_{ij} parameters can be calculated, allowing for their optimization within an adaptive Metropolis-Hastings MCMC framework (Supplementary Information). This approach can be generalized to model individual profiles as mixtures of K ancestral populations, where K is user-defined

$$f_4(\mathrm{H1},\ \mathrm{H2};\ \mathrm{H3},\ \mathrm{H4}) = \left(\sum_{i=1}^{K} Q_{i\_H1} \sum_{j=1}^{K} Q_{j\_H2} d_{ij}\right)\left(\sum_{i=1}^{K} Q_{i\_H3} \sum_{j=1}^{K} Q_{j\_H4} d_{ij}\right)$$

and d_{ij} now represents the allele frequency shift that occurred between the ancestral components i and j, while Q_{i_H1} the proportion

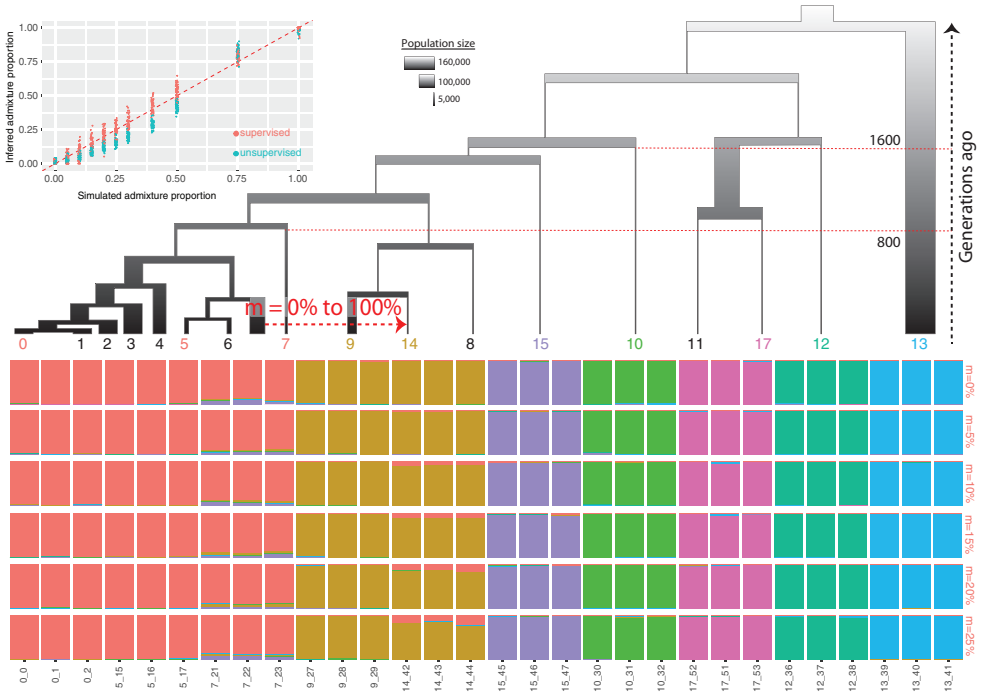

Fig. 1. (**a**) Simulated tree including 18 populations, and three sampled individuals per population. One admixture was simulated (*m*) into population 14 [0,0.25]. (**b**) Admixture proportions estimated by Struct-f4. For consistency with Harney et al. (2021), we simulated the full population history but restricted inference to 10 populations only. The name of each sample is composed of its corresponding population of origin, concatenated to a unique sample identifier. Samples from population 14 show an increasing ancestry from populations 0, 5 and 7 (colored in red) for greater simulated introgression proportions, as expected, and despite the true donor population remained unsampled

of the i ancestry inherited by individual H_1 (i.e. its mixture coefficient).

Struct-f4 is implemented in Rcpp for computational efficiency and requires a matrix of f_4-statistics as input. We also provide the Calc-f4 C program, which was parallelized and optimized for fast computation of f_4-statistics. This reduces the running time to calculate 82 215 f_4-statistics on a single 2700 MHz core to 48′49″, versus 1011′11″ for qpDstat. Struct-f4 outputs posterior mean values and credible intervals for each estimated parameter, together with the full MCMC sample and the corresponding probability used to assess convergence. It also provides (i) an MDS plot of genetic affinities between individuals (Supplementary Fig. S1), (ii) an unsupervised clustering based on the allele frequency shifts that occurred across pairs of individuals and/or K ancestral populations (Supplementary Fig. S2) and (iii) a barplot representation of ancestry profiles (Fig. 1).

3 Results

We evaluated Struct-f4 using the same simulation framework as that implemented by (Harney *et al.*, 2021) for assessing qpAdm performance. Individuals simulated as belonging to populations either closely related or increasingly connected by gene-flow appeared next to each other in the MDS space. Each individual was found to cluster according to the phylogenetically closest cladal group in the simulated model and showed genetic profiles consistent with the intensity of admixture (Fig. 1). Slight underestimates of the admixture proportions were returned for unsupervised analyses, if sampling only three haploid individuals per population. Supervised inference, nevertheless, completely fixed this bias (Fig. 1). Therefore, Struct-f4 can be used even when sampling efforts are limited, advantageously expanding the analytical toolkit in statistical genomics by providing a robust, flexible and user-friendly platform to automatically characterize population genetic structure. We successfully applied Struct-f4 to characterize the population structure underlying 284 ancient and modern horse genomes from over 11 million permutations of f_4-statistics (Librado *et al.*, 2021).

Funding

This work was received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [Grant 681605].

Conflict of Interest: none declared.

References

Alexander,D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.

Fages,A. *et al.* (2019) Tracking five millennia of horse management with extensive ancient genome time series. *Cell*, **177**, 1419–1435.e31.

Harney,É. *et al.* (2021) Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, **217**, iyaa045.

Lawson,D.J. *et al.* (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.*, **9**, 3258.

Leppälä,K. *et al.* (2017) admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics*, **33**, 1738–1740.

Librado,P. *et al.* (2021) The origins and spread of domestic horses from the Western Eurasian steppes. *Nature*, **598**, 634–640.

Nielsen,R. *et al.* (2017) Tracing the peopling of the world through genomics. *Nature*, **541**, 302–310.

Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Patterson,N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.