BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1187–1189
doi:10.1093/bioinformatics/btt763

Genetics and population analysis

Advance Access publication January 2, 2014

DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data

Jean-Marie Cornuet[1], Pierre Pudlo[1,2,3], Julien Veyssier[1,3,4], Alexandre Dehne-Garcia[1,3], Mathieu Gautier[1,3], Raphaël Leblois[1,3], Jean-Michel Marin[2,3] and Arnaud Estoup[1,3,*]

[1]Inra, UMR1062 cbgp, Montpellier, France, [2]Université Montpellier 2, UMR CNRS 5149, I3M, Montpellier, France, [3]Institut de Biologie Computationnelle (IBC), 34095 Montpellier, France and [4]CNRS-UM2, Institut de Biologie Computationnelle, LIRMM, Montpellier, France

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: DIYABC is a software package for a comprehensive analysis of population history using approximate Bayesian computation on DNA polymorphism data. Version 2.0 implements a number of new features and analytical methods. It allows (i) the analysis of single nucleotide polymorphism data at large number of loci, apart from microsatellite and DNA sequence data, (ii) efficient Bayesian model choice using linear discriminant analysis on summary statistics and (iii) the serial launching of multiple post-processing analyses. DIYABC v2.0 also includes a user-friendly graphical interface with various new options. It can be run on three operating systems: GNU/Linux, Microsoft Windows and Apple Os X.

Availability: Freely available with a detailed notice document and example projects to academic users at http://www1.montpellier.inra.fr/CBGP/diyabc

Contact: estoup@supagro.inra.fr

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

Received on July 31, 2013; revised on November 26, 2013; accepted on December 25, 2013

1 INTRODUCTION

One prospect of current biology is that molecular data will help us to reveal the complex demographic processes that have acted on natural populations. The extensive availability of various molecular markers and increased computer power has promoted the development of inferential methods. Among these novel methods, approximate Bayesian computation (ABC) method (Beaumont *et al.*, 2002) is increasingly used to make inferences from large datasets for complex models in various research fields, including population and evolutionary biology.

General statistical features, practical aspects and applications of ABC in evolutionary biology have been reviewed in three recent papers (Beaumont, 2010; Bertorelle *et al.*, 2010; Csilléry *et al.*, 2010). Briefly, ABC constitutes a recent approach to carrying out model-based inference in a Bayesian setting in which model likelihoods are difficult to calculate (due to the complexity of the models considered) and must be estimated by massive simulations. In ABC, the posterior probabilities of different models and/or the posterior distributions of the demographic parameters under a given model are determined by measuring the similarity between the observed dataset (i.e. the target) and a large number of simulated datasets; all raw datasets (i.e. multilocus genotypes or individual sequences) are summarized by statistics, such as mean number of alleles or F_{st}.

Several ABC programs have been proposed to provide solutions to non-specialist biologists (Table 1 in Bertorelle *et al.*, 2010; see also Supplementary Appendix S1). Cornuet *et al.* (2008, 2010) developed the (coalescent based) software DIYABC in which a user-friendly interface helps non-expert users to perform historical inferences using ABC. DIYABC allows considering complex population histories including any combination of population divergence events, admixture events and changes in past population size (with population samples potentially collected at different times). DIYABC can be used to compare competing evolutionary scenarios and quantify their relative support and estimate parameters for one or more scenarios. Eventually, it provides a way to evaluate the amount of confidence that can be put into the various estimations and to achieve model checking computation.

In this article, we present DIYABC v2.0, a completely rewritten version of the software DIYABC (Cornuet *et al.*, 2008, 2010). Version 2.0 implements a number of new features and analytical methods allowing extensive analyses of large molecular datasets, including single nucleotide polymorphism (SNP) data.

2 NEW FEATURES

2.1 Analysis of SNP data

DIYABC v2.0 allows analyzing statistically independent SNP markers, apart from microsatellite and DNA sequence data. Compared with other types of markers, SNP loci have low mutation rates, so that polymorphism at such loci results from a single mutation during the whole population gene tree, and genotypes are bi-allelic. To generate a simulated polymorphic dataset at a given SNP locus, we proceeded following the algorithm proposed by Hudson (2002) (cf—*s 1* option in the program *ms* associated to Hudson, 2002). Briefly, the genealogy at a given

*To whom correspondence should be addressed.

locus of all genes sampled in all populations of the studied dataset is simulated until the most recent common ancestor according to coalescence theory. Then a single mutation event is put at random on one branch of the genealogy (the branch being chosen with a probability proportional to its length relatively to the total gene tree length). This algorithm provides the simulation efficiency and speed necessary in the context of ABC, where large numbers of simulated datasets including numerous SNP loci have to be generated (see Supplementary Appendix S1 for additional comments on Hudson's algorithm).

2.2 Computation of scenario probability

Estoup et *al.* (2012) recently proposed a methodological innovation to deal with the discrimination among a large set of complex scenarios through efficient ABC probability computation. It is based on a linear discriminant analysis on summary statistics before the logistic regression analysis (introduced by Fagundes *et al.*, 2007). A major practical advantage is that it substantially decreases the dimension of explanatory variables making computation of scenario probability (~100 times) faster. We have implemented this methodological innovation in DIYABC v2.0 for the analysis of both the real datasets and the simulated pseudo-observed datasets used to evaluate the amount of confidence that can be put into the discrimination of a given set of scenarios.

2.3 New graphical interface and random number generator

DIYABC v2.0 has a new user-friendly graphical interface structured into two main parts: (i) one part including the definition of scenarios, prior distributions, summary statistics and the production of simulated datasets drawing parameter values into priors and (ii) other part including all types of post-processing computations typical of ABC analyses. Among the new options proposed, part (i) allows the definition of different groups of markers characterized by different mutation models and summary statistics and part (ii) allows launching serially multiple post-processing analyses (Supplementary Appendix S2).

Random number generators (RNG) are an important issue especially when several processors are used simultaneously for parallel computing. In DIYABC v2.0, we used RNG of *Mersenne Twister* types. In the multithreaded sections of the codes, which require random draws, each thread uses its own random generator. We initiate the different RNG with the algorithm proposed by Matsumoto and Nishimura (2000) to produce independent random streams.

2.4 Implementation

DIYABC v2.0 is a multithreaded program that runs on three operating systems: GNU/Linux, Microsoft Windows and Apple Os X. Computational procedures are written in C++, and the graphical user interface is based on PyQt, a Python binding of the Qt framework.

3 DISCUSSION

One of the main innovations of DIYABC v2.0 is that it can analyze SNP data, using an efficient simulation algorithm, therefore allowing the treatment of multi-population datasets with large number of loci (e.g. several thousands to ten thousands of loci within a few hours to a few days). The analyzed SNP data are assumed to correspond to independent selectively neutral loci, without any ascertainment bias (AB, i.e. the deviations from expected theoretical results due to the SNP discovery process in which a small number of individuals from selected populations are used as discovery panel). AB may distort measures of diversity and possibly change conclusions drawn from these measures in unexpected ways (e.g. Albrechtsen *et al.*, 2010). AB is mainly a concern when using SNP data obtained from chip-based high-throughput genotyping. It should impact to a much lower extent SNP data obtained from recent next-generation sequencing technologies, such as shotgun sequencing or restriction-site associated DNA sequencing techniques that are increasingly popular, including in population genetics studies of non-model species (Davey *et al.*, 2012). See Supplementary Appendix S1 for additional comments on AB. Another advantage of DIYABC v2.0 is that it provides the posterior distributions of demographic parameters scaled either by the mutation rate or by the effective population size, in parallel to those of original parameters. Scaled parameters are sometimes if not often the only type of parameters that can be robustly inferred under many evolutionary scenarios (e.g. Wakeley, 2005). Owing to the compilation optimization of C++ code and the multithreading of additional computation sections of the program, DIYABC v2.0 is also running faster than the previous version of the program (Supplementary Appendix S3). Finally, the new interface includes an automatic procedure to produce the different files to easily launch simulations on a computer cluster, hence obtaining access to larger computational resources.

ACKNOWLEDGEMENTS

The authors thank the 'beta-users' (Eric Lombaert, Michael Fontaine, Carine Brouat, Thomas Guillemaud, Christophe Plantamp, Johan Michaux and Marie-Pierre Chapuis) who tested the software with their data.

Funding: French Agence Nationale de la Recherche (ANR-09-BLAN-0145-01), Inra–Jeune Equipe IGGiPop, CBGP HPC computational platform and NUMEV Labex.

Conflict of Interest: none declared.

REFERENCES

Albrechtsen,A. *et al.* (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.*, **27**, 2534–2547.

Beaumont,M.A. (2010) Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.*, **41**, 379–406.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Bertorelle,G. *et al.* (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.*, **19**, 2609–2625.

Cornuet,J.M. *et al.* (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Cornuet,J.M. *et al.* (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**, 401.

Csilléry,K. *et al.* (2010) Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.*, **25**, 410–418.

Davey,J.W. *et al.* (2012) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, **12**, 499–510.

Estoup,A. *et al.* (2012) Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics. *Mol. Ecol. Res.*, **12**, 846–855.

Fagundes,N.J.R. *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proc. Natl Acad. Sci. USA*, **104**, 17614–17619.

Hudson,R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Matsumoto,M. and Nishimura,T. (2000) Dynamic Creation of Pseudorandom Number Generators. In: Fang,F. *et al.* (eds) *Monte Carlo and Quasi-Monte Carlo Methods 1998*. Springer-Verlag, New York, pp. 56–69.

Wakeley,J. (2005) The limits of theoretical population genetics. *Genetics*, **169**, 1–7.