BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 12 2011, pages 1717–1718
doi:10.1093/bioinformatics/btr253

Genetics and population analysis

Advance Access publication April 19, 2011

Mcheza: a workbench to detect selection using dominant markers

Tiago Antao[1,*,†] and Mark A. Beaumont[2,†]

[1]Molecular and Biochemical Parasitology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool and [2]Department of Mathematics, University of Bristol, Bristol, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Dominant markers (DArTs and AFLPs) are commonly used for genetic analysis in the fields of evolutionary genetics, ecology and conservation of genetic resources. The recent prominence of these markers has coincided with renewed interest in detecting the effects of local selection and adaptation at the level of the genome.

Results: We present Mcheza, an application for detecting loci under selection based on a well-evaluated F_{ST}-outlier method. The application allows robust estimates to be made of model parameters (e.g. genome-wide average, neutral F_{ST}), provides data import and export functions, iterative contour smoothing and generation of graphics in an easy to use graphical user interface with a computation engine that supports multicore processors for enhanced performance. Mcheza also provides functionality to mitigate common analytical errors when scanning for loci under selection.

Availability: Mcheza is freely available under GPL version 3 from http://popgen.eu/soft/mcheza.

Contact: tra@popgen.eu

Received on January 14, 2011; revised on April 11, 2011; accepted on April 12, 2011

1 INTRODUCTION

Non-specific amplification methods, such as Diversity Arrays Technology (DArT) markers and amplified fragment length polymorphism (AFLP), are commonly used for analysis of within-species variation because they allow the rapid acquisition of substantial genetic information, at relatively low cost. Although other alternative sequencing techniques have since been developed, DArTs and AFLPs are still widely used in the fields of evolutionary genetics, ecology and conservation. One of the most important applications of these dominant markers is in detecting the effects of selection and local adaptation at the level of the genome, in areas ranging from parasitology to conservation genetics.

There are two current approaches to detect selection: 'classical' F_{ST}-outlier approaches [reviewed in Storz (2005)], based on the distribution of summary statistics; and those based on likelihood (such as Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). The original F_{ST}-outlier methods do not account for dominant markers such DArTs or AFLPs. These markers have two phenotypes detectable at each locus: one allele (the plus-allele) is amplifiable, whereas the other (the null-allele) is not. Heterozygous genotypes cannot be directly distinguished from homozygotes making the estimation of allele frequencies non-trivial.

A widely used F_{ST}-outlier method (Pérez-Figueroa *et al*., 2010) for detecting selection with dominant markers is implemented in the package DFDIST. DFDIST is a modification of the FDIST program (Beaumont and Nichols, 1996) to allow for dominant markers, and implements the method of Zhivotovsky (1999) to estimate allele frequencies. Briefly, coalescent simulations are used to generate a null sampling distribution of estimates of F_{ST} based upon neutral expectations. The performance of the method has been examined, using data simulated with known levels of selection, in Caballero *et al*. (2008) and Pérez-Figueroa *et al*. (2010). DFDIST has a complicated text-based interface that makes it difficult to use it correctly. For example, the tuning of parameters for the coalescent simulations is non-trivial, and the input dataset has to be formatted in a non-standard way. We describe Mcheza, a new application based on DFDIST, with a graphical user interface allowing easier configuration of some non-trivial parameters.

2 SOFTWARE IMPLEMENTATION

The Mcheza architecture is composed of two parts: the front-end implemented in Jython and the DFDIST back-end implemented in C. The front-end provides an interface similar to LOSITAN (Antao *et al*., 2008) (A selection workbench based on the analogous method for co-dominant markers). The interface provides the following functionality on top of DFDIST:

(1) Estimation of the mean neutral F_{ST}, while taking into account loci that might be under selection. While DFDIST requires an estimate of the neutral F_{ST}, an empirical dataset will probably include loci under selection. Mcheza provides a mechanism similar to that in LOSITAN for estimating the neutral F_{ST} based on the removal of loci that are potentially under selection.

(2) An improved method for ensuring that the simulated distribution of F_{ST} has a mean that is close to the required value. DFDIST is only capable of providing a reliable approximation when close to theoretical conditions (i.e. when simulating a large number of populations). The Mcheza interface provides a correction that accurately approximates F_{ST} even when the number of demes is very low.

(3) Mcheza provides additional features in comparison with LOSITAN by supporting very large datasets: while LOSITAN is only able to support hundreds of loci and hundreds of individuals, Mcheza has been tested using real datasets with

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

25 000 loci. Support for very large datasets is, as expected, computationally more intensive.

(4) Mcheza also introduces support for multitest correction based on false discovery rates (FDRs) (Benjamini and Hochberg, 1995), as implemented in Chiurugwi *et al*. (2011). Without such a correction, there is a danger in overestimating the proportion of loci that are under selection (Beaumont, 2008; Pérez-Figueroa *et al*., 2010).

(5) A multicore aware version of DFDIST with computational performance gains that are near linear with the number of cores.

(6) An easy-to-use interface including the ability to import data in the standard Genepop format (Rousset, 2008), generation of charts, export in standard formats including (R Development Core Team, 2010) and spreadsheets, iterative smoothing of confidence contours, choice of population and loci among other features.

The application, based on the Java Web Start technology, requires only a browser with a modern version of Java installed (on Linux the GNU C compiler is also required). The Java code will detect the operating system and choose the correct DFDIST implementation.

The use of the Jython programming language allows the use of Biopython (Cock *et al*., 2009), which provides a parser for Genepop files. Our Python code to interact with DFDIST is incorporated in the Biopython population genetics module allowing for bioinformatics programmers to directly interface with the DFDIST core using the Python programming language.

3 DISCUSSION

A fundamental consideration in the design of Mcheza is supporting the user by correctly computing important non-trivial parameters that are needed to properly calculate candidate loci for selection. Erroneous usage of population genetics applications can easily produce results that seem correct but are, in effect wrong.

While Mcheza tries to minimize usage errors, the user should be aware of potential limits of the underlying method. Potential users are advised to read Caballero *et al*. (2008), Pérez-Figueroa *et al*. (2010) and Excoffier *et al*. (2009), wherein several scenarios where DFDIST is less applicable are clearly explained. In particular, the Bayesian method of Foll and Gaggiotti (2008), implemented in the program *BayeScan* is an important alternative, with which results should be compared. By improving the estimation of the neutral mean F_{ST} when the number of demes is low, Mcheza addresses situations where DFDIST is known to perform less well (Pérez-Figueroa *et al*., 2010).

In summary, Mcheza tries to provide an intuitive interface, which includes intelligent suggestions to the user with regards to correct usage of software, while enforcing model constraints and providing necessary corrections (e.g. FDR support). It is hoped that this approach will lower barriers to its use, allowing researchers to concentrate more on the biological problems (including the theoretical assumptions and limitations of underlying models) and less on unnecessary software complexity.

Funding: Research grant (SFRH/BD/30834/2006 to T.A.) from Fundação para a Ciência e Tecnologia (FCT), Portugal.

Conflict of Interest: none declared.

REFERENCES

Antao,T. *et al*. (2008) LOSITAN: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinformatics*, **9**, 323.

Beaumont,M.A. (2008) Selection and sticklebacks. *Mol. Ecol.*, **15**. 3425–3427.

Beaumont,M.A. and Balding,D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, **13**. 969–980.

Beaumont,M.A. and Nichols,R.A. (1996) Evaluating loci for use in the genetic analysis of population structure, *Proc. R Soc. B*, **263**, 1619–1626.

Benjamini,Y., Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B*, **57**, 289–300.

Caballero,A. *et al*. (2008) Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. *Genetics*, **179**, 539–554.

Chiurugwi,T. *et al*. (2011) Adaptive divergence and speciation among sexual and pseudoviviparous populations of *Festuca*. *Heredity*, **106**, 854–861.

Cock,P.J.A. *et al*. (2008) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bionformatics*, **25**, 1422–1423.

Excoffier,L. *et al*. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298

Foll,M. and Gaggiotti,O. (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, **180**, 977–993.

Pérez-Figueroa,A. *et al*. (2010) Comparing three different methods to detect selective loci using dominant markers. *J. Evol. Biol.*, **23**, 2267–2276.

Rousset,F. (2008) GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.*, **8**, 103–106.

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Storz,J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence, *Mol. Ecol.*, **14**, 671–688.

Zhivotovsky,L.A. (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Mol. Ecol.*, **8**, 907–913