BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1451–1452
doi:10.1093/bioinformatics/btp187

Genetics and population analysis

DnaSP v5: a software for comprehensive analysis of DNA polymorphism data

P. Librado[1,2] and J. Rozas[1,2,*]

[1]Departament de Genètica, Facultat de Biologia and [2]Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Received on February 10, 2009; revised and accepted on April 2, 2009

Advance Access publication April 3, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: DnaSP is a software package for a comprehensive analysis of DNA polymorphism data. Version 5 implements a number of new features and analytical methods allowing extensive DNA polymorphism analyses on large datasets. Among other features, the newly implemented methods allow for: (i) analyses on multiple data files; (ii) haplotype phasing; (iii) analyses on insertion/deletion polymorphism data; (iv) visualizing sliding window results integrated with available genome annotations in the UCSC browser.

Availability: Freely available to academic users from: http://www.ub.edu/dnasp

Contact: jrozas@ub.edu

1 INTRODUCTION

The analysis of DNA polymorphisms is a powerful approach to understand the evolutionary process and to establish the functional significance of particular genomic regions (Begun *et al.*, 2007; Nielsen, 2005; Rosenberg and Nordborg, 2002). In this context, estimating the impact of natural selection (both positive and negative) is of major interest. Furthermore, DNA polymorphisms are relevant as a tool for a broad range of life science disciplines. Consequently, many high-throughput sequencing, genotyping and polymorphism detection systems have been developed and are currently publicly available (Shendure and Ji, 2008). These new technologies are generating massive amounts of data that need to be processed, analyzed and transformed effectively into knowledge.

These technological advances have largely stimulated the development of both analytical methods and computer applications. Population genetic methods, and particularly those based on coalescent theory (Hudson, 1990; Wakeley, 2009), are used at an increasing rate, but need to be adapted to the particularities of the data (massive amounts of data, missing data, genotypes, insertion/deletion (indels) polymorphisms, etc.). Furthermore, new computer applications and algorithms need to be developed for processing massive datasets (Excoffier and Heckel, 2006), and more specifically computer visualization tools for the representation of DNA variation patterns. DnaSP (DNA Sequence Polymorphism) is a software package that allows for extensive DNA polymorphism analyses using a friendly graphical user interface (GUI) (Rozas *et al.*, 2003). Version 5 extends the capabilities of the software, allowing comprehensive DNA polymorphism analyses on multiple data files and on large datasets. Altogether, the present version of DnaSP has the appropriate features for exhaustive exploratory analyses using high-throughput DNA polymorphism data.

*To whom correspondence should be addressed.

2 FEATURES

DnaSP v5 incorporates major improvements. The new version currently allows for the handling and analysis of multiple data files in batch, and implements new algorithms and methods; among other things (see below) includes a new module to identify conserved DNA regions, this feature might be useful for phylogenetic footprinting-based analyses (Vingron *et al.,* 2009). DnaSP provides a convenient GUI facilitating all data management and analytical tasks; the results can be visualized graphically as well as in a text report. DnaSP accepts multiple DNA sequence alignment file formats (Rozas *et al.*, 2003), including NEXUS (Maddison *et al.*, 1997), and HapMap3 files with phased haplotypes (The International HapMap Consortium, 2003). The software allows exhaustive DNA polymorphism analyses, including those based on coalescent theory (Rozas *et al.*, 2003; Wakeley, 2009).

2.1 Haplotype reconstruction

Haplotype reconstruction aims at resolving haplotype phase given genotypic information. DnaSP implements statistical methods to infer haplotype phase, and prepares adequately the phased data for subsequent analyses. The input data (unphased genotype data) are required in FASTA format using IUPAC nucleotide ambiguity codes to represent heterozygous sites. DnaSP reconstructs the phase by applying various algorithms (PHASE v2.1, fastPHASE v1.1 and HAPAR) differing in the underlying population genetic assumptions. PHASE (Stephens and Donnelly, 2003; Stephens *et al.*, 2001) assumes Hardy–Weinberg equilibrium and uses a coalescent-based Bayesian method to infer haplotypes. fastPHASE (Scheet and Stephens, 2006) implements a modification of the PHASE algorithm taking into account the patterns of linkage disequilibrium and its gradual decline with physical distance. This algorithm is faster and allows for the handling of larger datasets than PHASE, while being slightly less accurate. HAPAR (Wang and Xu, 2003) infers haplotype phase by maximum parsimony, i.e. attempts to find the minimum number of haplotypes explaining the genotype sample.

2.2 Deletion/insertion polymorphisms

Deletion/insertion polymorphisms (DIPs) analysis can provide insights into the evolutionary forces acting on DNA. This information, however, has been rarely used. One obstacle has been the difficulty of defining clearly homologous states (Young and Healy, 2003). DnaSP incorporates an algorithm for treating indels related to the 'simple indel coding' method of Simmons and Ochoterena (2000). Specifically, only indels with the same 5′ and 3′ termini are considered homologous (resulted from a single event), and indels of different lengths (even in the same position of the alignment) are treated as different events. DnaSP, nevertheless, uses a slightly different method for coding completely overlapping gaps, and allows the user to choose the level of overlap to be coded. Subsequently, DnaSP estimates a number of DIP summary statistics, such as the average indel length, indel diversity, as well as Tajima's *D* (Tajima, 1989) based on indel information. Additionally, it exports the recoded data in the NEXUS format file.

2.3 Analysis of multiple data files

DnaSP can automatically read and analyze multiple data files sequentially (in batch mode). These data files may contain a varying number of sequences (from within one species, or from one species as well as one outgroup), or represent diverse genomic regions. The program estimates the most common DNA polymorphism and divergence summary statistics (such as the nucleotide and haplotype diversity, the population mutation parameter, the number of nucleotide substitutions per site, etc.), and neutrality tests (such as Tajima's, Fu and Li's and Fu's tests).

2.4 Sliding window results visualization

The sliding window technique is a useful tool for exploratory DNA polymorphism data analysis (Hutter *et al.*, 2006; Rozas *et al.*, 2003; Vilella *et al.*, 2005). The current version of DnaSP permits visualizing results of the sliding window (for example, nucleotide diversity or Tajima's *D* values along the DNA sequence) integrating available genome annotations in the UCSC browser (Kent *et al.*, 2002). This feature can greatly facilitate the interpretation of the results; for instance, it is possible to identify the relevant genome annotations (genes, intergenic regions, conserved regions, etc.), which are adjacent to regions with atypical patterns of nucleotide variation.

3 IMPLEMENTATION

DnaSP version 5 has been developed in Microsoft Visual Basic v6.0, C and C++, and it runs under Microsoft Windows operating systems (2000/XP/Vista). With the use of Windows emulators, DnaSP can also run on Apple Macintosh platforms, Linux and Unix-based operating systems. The software has been tested in all three platforms.

ACKNOWLEDGEMENTS

We acknowledge Sergios-Orestis Kolokotronis for helpful comments on the manuscript. Special thanks to the numerous users who tested the software with their data, and particularly to all members of the Molecular Evolutionary Genetics group at the Departament de Genètica, Universitat de Barcelona.

Funding: Spanish Dirección General de Investigación Científica y Técnica (grants BFU2004-02253 and BFU2007-62927); the Catalonian Comissió Interdepartamental de Recerca i Innovació Tecnològica (grant 2005SGR00166).

Conflict of Interest: none declared.

REFERENCES

Begun,D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.*, **6**, e310.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, **7**, 745–758.

Hudson,R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 1–44.

Hutter,S. *et al*. (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

Kent,W.J. *et al*. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Maddison,W.P. *et al.* (1997) NEXUS: an extendible file format for systematic information. *Syst. Biol.*, **46**, 590–621.

Nielsen,R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.*, **39**, 197–218.

Rosenberg,N.A. and Nordborg,M. (2002) Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat. Rev. Genet.*, **3**, 380–390.

Rozas,J. *et al.* (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Simmons,M.P. and Ochoterena, H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, **49**, 369–381.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Stephens,M. *et al.* (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Vilella,A.J. *et al.* (2005) VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, **21**, 2791–2793.

Vingron,M. *et al.* (2009) Integrating sequence,evolution and functional genomics in regulatory genomics. *Genome Biol.*, **10**, 202.

Wang,L. and Xu,Y. (2003) Haplotype inference by maximum parsimony. *Bioinformatics*, **19**, 1773–1780.

Wakeley,J. (2009) *Coalescent Theory. An Introduction*. Roberts and Company Publishers. Greenwood Village.

Young,N.D. and Healy, J. (2003) GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics*, **4**, 6.