DnaSP, DNA polymorphism analyses by the coalescent and other methods.

Author affiliation:

Julio Rozas[1,*], Juan C. Sánchez-DelBarrio[2,3], Xavier Messeguer[2] and Ricardo Rozas[1]

[1] Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain

[2] Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain

[3] Present Address: Departament de Tecnología, Universitat Pompeu Fabra, Barcelona, Spain

Name and address for correspondence:

Julio Rozas

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain.

Tel.: 34 93 402 1495

Fax: 34 93 411 0969

E-mail: **jrozas@ub.edu**

Running head:

DNA polymorphism analysis

[*] To whom correspondence should be addressed

ABSTRACT

Summary: DnaSP is a software package for the analysis of DNA polymorphism data. Present version introduces several new modules and features which, among other options allows, 1) handling big data sets (~5 Mbp per sequence); 2) conducting a large number of coalescent-based tests by Monte Carlo computer simulations; 3) extensive analyses of the genetic differentiation and gene flow among populations; 4) analysing the evolutionary pattern of preferred and unpreferred codons; 5) generating graphical outputs for an easy visualization of results.

Availability: The software package, including complete documentation and examples, is freely available to academic users from: http://www.ub.es/dnasp

Contact: jrozas@ub.edu

INTRODUCTION

Recent advances in DNA sequencing and polymorphism detection methodologies are generating huge datasets of DNA sequence variation and of single nucleotide polymorphisms (SNPs). Analysis of such DNA polymorphism data will definitively enhance our understanding of both the evolutionary significance of DNA polymorphisms and of the evolutionary history of populations and species (Nordborg and Innan 2002). Additionally, DNA polymorphism information has a wide range of applications, including pharmacogenomics, animal and plant breeding, conservation genetics, epidemiology genetics, medicine and forensics.

Current massive datasets are stimulating the development of numerous methods to interpret DNA polymorphism data. These methods capture different features of the data (SNP frequency, association among variants, haplotype structure, synonymous and nonsynonymous changes, recombinational events, codon usage, etc.) (Rosenberg and Nordborg 2002; Bamshad and Wooding 2003). In this context, the coalescent theory (see Hudson, 1990; Rosenberg and Nordborg 2002) has become the primary framework to analyse the data. Indeed, coalescent-based methods are critical for detecting the signature of positive natural selection, in the identification of haplotype blocks across the genome, or for inferring the effect of intragenic recombination. Here, we describe version 4 of the DnaSP software package (Rozas and Rozas 1999). Present version largely extends the capabilities of the software allowing extensive DNA polymorphism analyses on a user-friendly interface.

SYSTEM AND METHODS

DnaSP version 4 is written in Microsoft Visual Basic v. 6.0 and runs on ix86 compatible processors under Microsoft Windows®. DnaSP can also run on Apple Macintosh, Linux and Unix-based platforms using Windows emulator software with one of the required Microsoft Windows® versions.

MAIN NEW FEATURES

DnaSP provides a user-friendly Microsoft Windows graphic interface and can read (and export) five multiple-aligned nucleotide sequence file formats: FASTA, MEGA, NBRF/PIR, NEXUS and PHYLIP. DnaSP allows the analysis of polymorphism, divergence, genetic differentiation, gene flow, gene conversion, linkage disequilibrium, recombination, codon usage and also conducts a number of neutrality tests. The analyses can be performed in a subset of sites (including synonymous, nonsynonymous, non coding, i-fold degenerate sites) or in a subset of DNA sequences. Coding region analysis can be performed using a number of predefined genetic codes and codon usage tables.

Coalescent-based methods

DnaSP has extensively increased the capabilities of the coalescent-based analyses. Present DnaSP version allows conducting most of the developed neutrality tests (with and without outgroup) and linkage disequilibrium statistics, including –among others- (1) Tajima's, Fu's and Fu and Li's tests (Tajima 1989; Fu and Li 1993; Fu 1997); (2) Depaulis and Veuille's haplotype-based tests (Depaulis and Veuille 1998); (3) *B* and *Q* tests (Wall 1999); (4) *H* test (Fay and Wu 2000); (5) Z_{nS}, *ZZ* and Z_A linkage disequilibrium based-statistics (Kelly 1997; Rozas *et al.,* 2001). DnaSP also computes a number of statistical tests

for detecting population growth including the recently developed R_2 test (Ramos-Onsins and Rozas 2002). The Monte Carlo computer simulation module allows generating the empirical distribution for a very large number of test statistics. Simulations can be conducted for different recombination rates.

Gene Flow and Genetic Differentiation

The *Gene Flow* module has been completely rewritten. Present version allows performing a number of gene flow and genetic differentiation among populations analyses with different options for treating alignment gaps. To detect genetic differentiation among subpopulations DnaSP implements several statistics based both on the number of haplotypes and on the number of nucleotide changes (*i.e.,* sequence-based statistics) (Hudson *et al.,* 1992a; Hudson 2000). DnaSP also estimates several parameters of the standardized measure of the genetic diversity among populations (F_{ST}, and the related statistics G_{ST}, N_{ST}) (see Hudson *et al.,* 1992b). From these F_{ST} based estimators, the migration rates (in terms of *Nm*; where *m* is the migration rate) are obtained. The outcome values can be exported as a distance data file (PHYLIP and MEGA formats) for further phylogenetic analyses. DnaSP incorporates two methods to test for genetic differentiation: 1) the standard χ^2 homogeneity test, and 2) a Monte Carlo permutation (randomization) test (Hudson *et al.,* 1992a).

Analysis of Preferred and Unpreferred codons

Present version implements a number of algorithms and methods to analyse the impact of natural selection and mutational processes on codon usage bias. In addition to the standard codon usage bias estimators (CBI, ENC, Scaled Chi-Square, etc.), DnaSP also implements an algorithm to identify preferred (P) and unpreferred (U) synonymous changes. This information is critical for determining the effect of natural selection (weak selection) on synonymous codons (see Akashi 1999). DnaSP allows estimating the numbers of preferred

and unpreferred changes within species (which requires the availability of one outgroup to polarize the mutations), and also those changes polymorphic within species and fixed between species (which requires the availability of two outgroups). DnaSP also provides several predefined codon usage tables. The user, additionally, can also define his own codon usage table; this user-defined information can be stored on a private block of the NEXUS file format.

ACKNOWLEDGEMENTS

We thank M. Aguadé, A. Blanco-García, H. Quesada, C. Segarra and A. Vilella for critical comments on the manuscript. We also thank the numerous people who tested the program with their data, especially members of the Molecular Evolutionary Genetics group in the Departament de Genètica, Universitat de Barcelona. This work was supported by grant BMC2001-2906 from the Dirección General de Investigación Científica y Técnica, Spain, conferred on M. Aguadé, and by grant TXT98-1802 from the Dirección General de Enseñanza Superior e Investigación Científica, Spain, conferred on J. Rozas.

REFERENCES

Akashi,H., (1999) Detecting the "footprint" of natural selection in within and between species DNA sequence data. *Gene,* **238**, 39-51.

Bamshad,M. and Wooding,S.P. (2003) Signatures of natural selection in the human genome. *Nature Rev. Genetics,* **4,** 99-111.

Depaulis,F. and Veuille, M. (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol. Biol. Evol*., **15,** 1788-1790.

Fay,J.C. and Wu, C.-I. (2000) Hitchhiking under positive Darwinian selection. *Genetics,* **155,** 1405-1413.

Fu,Y.-X. and Li,W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics,* **133,** 693-709.

Fu,Y.-X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics,* **147,** 915-925.

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

Hudson,R.R. (2000) A new statistic for detecting genetic differentiation. *Genetics,* **155,** 2011-2014.

Hudson,R.R., Boos,D.D. and Kaplan,N.L. (1992a) A statistical test for detecting population subdivision. *Mol. Biol. Evol.,* **9,** 138-151.

Hudson,R.R., Slatkin,M. and Maddison,W.P. (1992b) Estimation of levels of gene flow from DNA sequence data. *Genetics,* **132,** 583-589.

Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics,* **146,** 1197-1206.

Nordborg,M. and Innan,H. (2002) Molecular population genetics. *Curr. Op. in Plant Biol.,* **5,** 69-73.

Ramos-Onsins,S.E. and Rozas,J. (2002) Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.,* **19,** 2092-2100.

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

Rozas,J. and Rozas,R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics,* **15,** 174-175.

Rozas,J., Gullaud,M. Blandin,G. and Aguadé.M. (2001) DNA variation at the rp49 gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. *Genetics,* **158,** 1147-1155.

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

Wall,J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet. Res.,* **74,** 65-69.