The analysis of DNA polymorphisms is a powerful approach to understand the evolutionary process and to establish the functional analyses using a friendly graphical user interface (GUI) (Rozas et al., 2007; Excoffier and Heckel, 2006), and more of the data (massive amounts of data, missing data, genotypes, an increasing rate, but need to be adapted to the particularities of 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 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 data. 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 heterogeneous 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 D, 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., 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 regions, the population mutation parameter, the number of nucleotide substitutions per site, etc., and neutrality tests (such as Tajima’s D, Fu and Li’s and Fu’s tests).

3 IMPLEMENTATION
DnaSP version 5 has been developed in Microsoft Visual Basic v.6.0, 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 2005SGR0166).

Conflict of Interest: none declared.

REFERENCES

1452