MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2012) **12**, 562–565 doi: 10.1111/j.1755-0998.2011.03108.x

Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding

SAMUEL D. J. BROWN,[1] RUPERT A. COLLINS,[1] STEPHANE BOYER,[2] MARIE-CAROLINE LEFORT,[1] JAGOBA MALUMBRES-OLARTE,[2] COR J. VINK[3,4] and ROBERT H. CRUICKSHANK[2]

[1]*Bio-Protection Research Centre, PO Box 84, Lincoln University 7647, Canterbury, New Zealand,* [2]*Department of Ecology, Faculty of Agriculture and Life Sciences, PO Box 84, Lincoln University 7647, Canterbury, New Zealand,* [3]*Biosecurity Group, AgResearch, Private Bag 4749, Christchurch 8140, New Zealand,* [4]*Entomology Research Museum, PO Box 84, Lincoln University, Lincoln 7647, New Zealand*

Abstract

Spider: SPecies IDentity and Evolution in R is a new R package implementing a number of useful analyses for DNA barcoding studies and associated research into species delimitation and speciation. Included are functions essential for generating important summary statistics from DNA barcode data, assessing specimen identification efficacy, and for testing and optimizing divergence threshold limits. In terms of investigating evolutionary and taxonomic questions, techniques for assessing diagnostic nucleotides and probability of reciprocal monophyly are also provided. Additionally, a sliding window function offers opportunities to analyse information across a gene, essential for marker design in degraded DNA studies. Spider capitalizes on R's extensible ethos and offers an integrated platform ideal for the analysis of both nucleotide and morphological data. The program can be obtained from the comprehensive R archive network (CRAN, http://cran.r-project.org) and from the R-Forge package development site (http://spider.r-forge.r-project.org/).

Keywords: barcoding gap, characteristic attributes, integrative taxonomy, mini-barcodes, neighbour-joining, threshold optimization

Received 10 September 2011; revision received 11 November 2011; accepted 27 November 2011

Introduction

Increased interest in biodiversity assessment, specimen identification and species delimitation have spurred the development of numerous methods to investigate both pattern and process in the evolution of populations and species (e.g. Sites & Marshall 2003; Padial *et al*. 2010). Up until now, these methods have been implemented in a multitude of standalone programs, requiring users to analyse their data in a piecemeal fashion, and in different programs from those used to plot and present figures. Spider: SPecies IDentity and Evolution in R is a new R package, providing a suite of functions for exploring and testing both molecular (DNA sequence) and morphological (discrete character) species-level data. It is of direct use to all practitioners of DNA barcoding techniques, as well as taxonomists and evolutionary biologists interested in integrative approaches to systematic biology (Padial *et al*. 2010).

The statistical programming environment R (R Development Core Team, 2011a) has been widely used as a platform for analysis in the fields of phylogenetics and genomics research, with the development of packages including ape (Paradis *et al*. 2004), phangorn (Schliep 2011) and the bioconductor range (Hahne *et al*. 2008). However, its acceptance in the fields of population genetics and speciation has lagged. Because of its advantages for complex data manipulation, R is an ideal environment to conduct these analyses and it provides flexible analytical tools coupled with powerful graphical capabilities.

Correspondence: Samuel D. J. Brown, Fax: +64 3 325 3844;
E-mail: sam.brown@lincoln.ac.nz

Data input

Spider uses the *DNAbin* class for DNA sequence manipulation, which is implemented in ape (Paradis *et al*. 2004). In addition to existing functions in ape for downloading sequences from GenBank, spider also implements functions for searching and downloading publicly available sequences from BOLD, the international Barcode of Life Data System (http://www.barcodinglife.com). Local DNA alignments can be loaded using ape's *read.dna* or *read.nexus.data* functions, while morphological data can be loaded using the *read.table* family of functions in the base R installation.

The size of the data set that can be handled by R, and therefore SPIDER, is dependant on the computing platform on which it is installed and the patience of the user. All R objects are stored in memory, meaning that the upper limit in most circumstances is a data set of a few hundred megabytes (R Development Core Team 2011b). In the context of DNA sequence manipulation, however, the data structures provided by APE are able to handle decently sized alignments and trees. As an example, calculating a distance matrix and neighbour-joining tree from an alignment 850 bp in length with 3000 specimens takes around 220 s (measured on an Intel Pentium 1.66 GHz, 1024 kb CPU Cache, 2 Gb RAM, running Ubuntu 10.04). Processing time increases linearly with sequence length, but exponentially with sequence number (Supplementary Material Data S3).

A key element to many of the functions developed in SPIDER is the 'species vector'; this assignment is predefined by the user and allocates each individual in the data set to a group or taxonomic unit, such as species. This vector can be added manually, but following standard practice of including a species identity element to each unique identifier (i.e. the taxon names given to individuals *a priori*), it is more straightforward and less error prone to extract the species vector from the taxon names. Supplementary Material Data S2 (section: 'Species Vectors') contains examples of methods for extracting this information.

Barcode summary statistics

DNA barcoding is a method of identifying unknown biological material by sequencing a standard region of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene, which is then compared against a reference library of sequences of known origin (Vogler & Monaghan 2007). First proposed by Hebert *et al.* (2003), the method is now widely used in ecology (Jurado-Rivera *et al.* 2009), biodiversity assessment (Janzen *et al.* 2009), taxonomy (Benziger *et al.* 2011), conservation (Francis *et al.* 2010), consumer protection (Lowenstein *et al.* 2010) and biosecurity (Armstrong & Ball 2005). Research on the utility of DNA barcoding for different taxa involves measuring identification accuracy and sequence variation within and between species (Ward 2009).

SPIDER includes indispensable functions for calculating both standard summary statistics and tests of DNA barcode data. Summary statistics include the following: descriptions of the data (number of species, number of individuals, number of haplotypes per species, lengths of sequences, proportion of missing data); calculation of intra- and interspecific distances (both as averages and unsummarized); and assessment of the barcoding gap (maximum intraspecific and minimum interspecific distances).

Tests of barcoding efficacy include the following: species monophyly; 'best close match' (Meier *et al.* 2006); an inclusive threshold analysis similar to the method of specimen identification used by BOLD; and nearest-neighbour identification (cf. *k*-nearest neighbour of Austerlitz *et al.* 2009). It is important to note that these tests of efficacy are not identification tools. All sequences must be identified prior to testing. Each sequence is considered an unknown while the remaining sequences in the data set constitute the DNA barcoding database that is used for identification. If the identification from the test is the same as the preconsidered identification, a 'correct' result is returned.

We also offer a procedure to test the applicability of standard threshold cut-off values (i.e. BOLD's 1%), and optimize custom thresholds based upon error rates directly from the data (Meyer & Paulay 2005). Finally, methods to test taxon sampling protocols are included; a haplotype accumulation curve method offers a randomized examination of the rate of encountering unsampled genetic diversity. Results from all these analyses can be presented as tables or plotted using R's powerful graphical functions.

Sliding window analyses

A major component of SPIDER's capacity are functions for conducting sliding window analyses across DNA sequences. Sliding window analyses partition DNA sequences into shorter fragments, upon which further tests are conducted. These windows can be used to determine the performance of mini-barcodes (Meusnier *et al.* 2008), calculate diversity indices (Roe & Sperling 2007), explore character conflict (Cruickshank 2011) or evaluate genomic data for informative new markers and potential priming sites. SPIDER contains base functions for creating windows of specified width across an alignment at specified intervals. A range of distance matrix- and tree-based analyses can then be performed on these windows to assess their information content. The *slideAnalyses* function is able to create the windows and conduct the analyses in a single function; results can be plotted or presented as a table. Functions are also included to create boxplots of distance matrices and the distribution of pairwise intra- and interspecific distances, revealing a barcoding gap for each window.

Taxonomy and evolution

In addition to the procedures aimed at DNA barcoding described above, SPIDER also includes useful tools for taxonomic and evolutionary investigation. Species monophyly on given trees can be determined, and a bootstrap test allows the uncertainty in the tree be reflected in the result. Rosenberg's probability of reciprocal monophyly (Rosenberg 2007) offers a measure of the suitability of sampling to detect robustly monophyletic clades. Discrete character states using a diagnostic nucleotide approach can be generated—a technique useful for DNA taxonomy (Sarkar *et al.* 2008). Substitution patterns in sequence data can also be explored using the pairwise number of transitions and transversions between sequences, as well as with a visual representation of DNA barcodes. In addition to the nucleotide-sequence-based analyses described above, methods are implemented in SPIDER for species and population level morphological data; currently, the only method of this kind is the population aggregate analysis (Davis & Nixon 1992), which determines traits fixed or polymorphic within populations.

Ongoing development

SPIDER is an actively developing package and aims to provide functions for a range of analyses of species' variation. Future developments within SPIDER's current core capacity will involve expanding the nucleotide diagnostics function to include additional criteria and the inclusion of tools for the identification of unknown DNA sequences using local sequence libraries. SPIDER is also anticipated to implement the following: techniques for exploring conflict in phylogenetic data, randomization tests for statistical analysis of the barcoding gap, tools for studying ancient DNA, and additional methods for morphological and categorical data.

Obtaining SPIDER

SPIDER is a package of the statistical programming environment R, which is available for all computing platforms from the Comprehensive R Archive Network (CRAN, http://cran.r-project.org). A stable version of SPIDER is also available on CRAN and can be downloaded from within R while connected to the internet by entering the following commands at the prompt:

```
> install.packages(''spider'')
> library(spider)
```

In addition to the stable version on CRAN, a development version is available at R-Forge (http://spider.r-forge.r-project.org/). This version can be installed from within R by using the command:

```
> install.packages(''spider'', repos='http://R-Forge.R-project.org')
```

The help manual for version 1.1-0 and a tutorial demonstrating the use of SPIDER are available as Supplementary Material (Data S1 and S2 respectively), or updated versions can be downloaded from http://spider.r-forge.r-project.org/.

SPIDER requires the installation of the packages APE (Paradis *et al.* 2004), PEGAS (Paradis 2010), ADEGENET (Jombart 2008) and MASS (Venables & Ripley 2002), which provide the primary data structures for working with DNA sequences and phylogenetic trees. If these packages are not already on the system, they will automatically be installed when the commands above have been run.

References

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B*, **360**, 1813–1823.

Austerlitz F, David O, Schaeffer B *et al.* (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**, S10.

Benziger A, Philip S, Raghavan R *et al.* (2011) Unraveling a 146 years old taxonomic puzzle: validation of Malabar snakehead, species-status and its relevance for channid systematics and evolution. *PLoS ONE*, **6**, e21272.

Cruickshank RH (2011) Exploring character conflict in molecular data. *Zootaxa*, **2946**, 45–51.

Davis JI, Nixon KC (1992) Populations, genetic variation and the delimitation of phylogenetic species. *Systematic Biology*, **41**, 421–435.

Francis CM, Borisenko AV, Ivanova NV *et al.* (2010) The role of DNA barcodes in understanding and conservation of mammal diversity in Southeast Asia. *PLoS ONE*, **5**, e12575.

Hahne F, Huber W, Gentleman R, Falcon S (2008) *Bioconductor Case Studies*. Use R! Springer, New York.

Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003) Biological identification through DNA barcodes. *Proceedings of the Royal Society of London. B*, **270**, 313–321.

Janzen DH, Hallwachs W, Blandin P *et al.* (2009) Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, **9**, 1–26.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jurado-Rivera JA, Vogler AP, Reid CAM, Petitpierre E, Gómez-Zurita J (2009) DNA barcoding insect-host plant associations. *Proceedings of the Royal Society B*, **276**, 639–648.

Lowenstein JH, Burger J, Jeitner CW, Amato G, Kolokotronis SO, Gochfeld M (2010) DNA barcodes reveal species-specific mercury levels in tuna sushi that pose a health risk to consumers. *Biology Letters*, **6**, 692–695.

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 1–4.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, 2229–2238.

Padial J, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. *Frontiers in Zoology*, **7**, 1–14.

Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419–420.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

R Development Core Team (2011a) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Development Core Team (2011b) *R Data Import/Export*. R Foundation for Statistical Computing, Vienna, Austria.

Roe AD, Sperling FAH (2007) Patterns of evolution of mitochondrial cytochrome *c* oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution*, **44**, 325–345.

Rosenberg NA (2007) Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution*, **61**, 317–323.

Sarkar I, Planet P, DeSalle R (2008) CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*, **8**, 1256–1259.

Schliep K (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Sites JWJ, Marshall JC (2003) Delimiting species: a Renaissance issue in systematic biology. *Trends in Ecology and Evolution*, **18**, 462–470.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*. Springer, New York, 4th edn. ISBN 0-387-95457-0.

Vogler A, Monaghan M (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.

Ward RD (2009) DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources*, **9**, 1077–1085.

Data accessibility

DNA sequences for benchmarking: GenBank sequences GQ337328–GQ337385. Program source code available on CRAN (http://cran.r-project.org).

Supporting Information

Additional supporting information may be found in the online version of this article.

Data S1 SPIDER version 1.1 reference manual.

Data S2 SPIDER version 1.1 tutorial.

Data S3 Results of tests benchmarking the performance of APE's data structures.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.