Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 232–236 doi: 10.1111/2041-210X.12310

APPLICATION

BAT – Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity

Pedro Cardoso[1,2]*, François Rigal[2] and José C. Carvalho[2,3]

[1]*Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland;* [2]*CE3C - Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group, University of the Azores, Angra do Heroísmo, Portugal; and* [3]*CBMA – Centre of Molecular and Environmental Biology, Department of Biology, University of Minho, Braga, Portugal*

Summary

1. Novel algorithms have been recently developed to estimate alpha and partition beta diversity in all their dimensions (taxon, phylogenetic and functional diversity – TD, PD and FD), whether communities are completely sampled or not.

2. The R package BAT – Biodiversity Assessment Tools – performs a number of analyses based on either species identities (TD) or trees depicting species relationships (PD and FD). Functions include building randomized accumulation curves for alpha and beta diversity, alpha diversity estimation from incomplete samples and the partitioning of beta diversity in its replacement and richness difference components.

3. All functions allow the rarefaction of communities. Estimation methods include curve-fitting and nonparametric algorithms. Beta diversity indices include the Jaccard and Sørensen families of measures and deal with both incidence and abundance data. Two auxiliary functions that allow judging the efficiency of the algorithms are also included.

4. Several examples are shown using the data included in the package, which demonstrate the usefulness of the different methods. The BAT package constitutes an open platform for further development of new biodiversity assessment tools.

Key-words: accumulation curves, biodiversity estimation, diversity partition, extrapolation, nonparametric estimators, sampling bias

Introduction

The most studied biodiversity components are alpha diversity (α) that measures the diversity intrinsic to each community or site and beta diversity (β) that reflects the differences among communities or sites. Biodiversity can also be characterized according to different aspects measured by how species relate. Taxon diversity (TD) is the most common and is quantified based on the number of taxa and, often, on the distribution of abundances. However, TD disregards the fact that communities are composed of species with different evolutionary histories and a diverse array of ecological functions. Thus, the last decade has seen a growing interest in alternative representations of biodiversity, including phylogenetic diversity (PD) and functional diversity (FD) (Stegen & Hurlbert 2011).

Studying biodiversity patterns and the respective driving processes in all these dimensions is invariably challenging. As alpha diversity (TD, PD or FD) increases with the number of individuals or sampling units, observed richness almost invariably underestimates true richness (Coddington *et al.* 2009; Cardoso *et al.* 2014a). Without complete inventories of communities, comparisons of alpha diversity values cannot be reliably made if one does not consider the sampling effort or completeness attained (Gotelli & Colwell 2001). The same problems have been identified for beta diversity, since underestimation of similarity also occurs because of the failure to account for unseen shared species (Cardoso, Borges & Veech 2009). Some measures are more sensitive than others, and the comparison of randomized accumulation curves for both alpha (Gotelli & Colwell 2001) and beta (Cardoso, Borges & Veech 2009) diversity has been advocated as a previous step in many studies for which undersampling is suspected.

Estimating species richness from incomplete samples has long been proposed as a potential way of dealing with undersampling (Colwell & Coddington 1994). Two kinds of data may be used for estimating diversity: (i) abundance data and (ii) incidence data. A number of approaches for estimating true diversity are possible, mainly the fit of asymptotic functions to accumulation curves (Soberón & Llorente 1993) and the use of nonparametric estimators based on the abundance or incidence of rare species. Among the latter are the jackknife (Heltshe & Forrester 1983) and Anne Chao's formulas (Chao

*Correspondence author. E-mail: pedro.cardoso@helsinki.fi

1984, 1987). All these approaches were first developed for TD and only recently adapted for other dimensions of biodiversity, namely PD and FD (Cardoso *et al*. 2014a).

The term beta diversity has been used to refer to a variety of phenomena, although all of these encompass compositional heterogeneity between communities. Two distinct processes shape communities and their differences: species replacement and species loss (or gain) (Williams, de Klerk & Crowe 1999; Lennon *et al*. 2001). How to partition these different components has been a field of much recent development (Baselga 2010, 2012; Podani & Schmera 2011; Carvalho, Cardoso & Gomes 2012; Carvalho *et al*. 2013; Legendre 2014), including when phylogenetic or functional relations between species are taken into account (Cardoso *et al*. 2014c).

Our objective was to develop code for the measurement and the estimation of alpha and beta diversities in their multiple facets (taxon, phylogenetic and functional) along with the partitioning of beta diversity.

The BAT package

The R package BAT – Biodiversity Assessment Tools (Cardoso, Rigal & Carvalho 2014b) allows performing a number of analyses based on either species identities (TD) or ultrametric trees depicting species relationships (PD and FD). Functions include building randomized accumulation curves for alpha (*alpha*, *alpha.accum*) and beta (*beta.accum*) diversity, alpha diversity estimation from incomplete samples (*alpha.estimate*) and the partitioning of beta diversity in its replacement and richness differences components (*beta*), including for multiple sites simultaneously (*beta.multi*). Beta diversity indices include the Jaccard and Sørensen families of measures and deal with both incidence and abundance data. The package also implements two auxiliary functions that allow judging the efficiency of the algorithms (*accuracy*, *slope*).

The BAT package is written in R and can be installed from the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/BAT/). The raw data accepted by most functions are (1) a sites/sampling units (rows) by species (columns) matrix, with either abundance or incidence data and (2) a 'phylo' object as implemented in the R package *ape* (Paradis, Claude & Strimmer 2004) or a 'hclust' object (both used only for PD or FD). All functions allow the rarefaction of communities. The functions provided by the package may be grouped into three categories, according to their scope, as follows:

ALPHA DIVERSITY

1. *alpha (comm, tree, raref, runs)* calculates the observed alpha diversity of multiple sites with possible rarefaction. The argument *comm* is a sites × species matrix, with either abundance or incidence data. As in all functions below, *tree* is a 'phylo' or 'hclust' object (used only for PD or FD). Calculations of PD and FD are based on Faith (1992) and Petchey & Gaston (2002), respectively, which measure PD and FD of a community as the total branch length of a tree linking all species represented in such community. As in other functions below, *raref* specifies the number of individuals to be used for individual-based rarefaction. *runs* is the number of rarefaction procedures to execute. The function returns a matrix of sites × diversity values.

2. *alpha.accum (comm, tree, func, runs)* estimates alpha diversity of a single site with accumulation of sampling units. The argument *comm* is a sampling units × species matrix, with either abundance or incidence data. *func* is the class of estimators to be used, either curve-fitting, nonparametric or, for PD and FD, based on the completeness of TD (see Cardoso *et al*. 2014a). *runs* is the number of random permutations to be made to the sampling units order. The function returns a matrix of sampling units × diversity values (individuals, observed and estimated diversity).

3. *alpha.estimate (comm, tree, func)* estimates alpha diversity of multiple sites simultaneously. The argument *comm* is a sites × species matrix, with either abundance or number of incidences (sampling units where each species occurs) data. As above, *func* is the class of estimators to be used, but only nonparametric (jackknife and Chao for both incidence and abundance data) or, for PD and FD, based on the completeness of TD, are available (as no sampling unit data are given for curve fitting). The function returns a matrix of sites × diversity values (individuals, observed and estimated diversity).

BETA DIVERSITY

4. *beta (comm, tree, abund, func, raref, runs)* calculates beta diversity and its partitioning for multiple sites simultaneously with possible rarefaction. *comm* is a sites × species matrix, with either abundance or incidence data. *abund* is a Boolean indicating whether abundance data should be used or converted to incidence before analysis. As for all beta diversity functions in the package, *func* indicates whether the Jaccard or Sørensen family of beta diversity measures should be used. The function returns three distance matrices with beta diversity measured between all pairs of sites (*dist* object), one per each of the three beta diversity measures (β_{total} – overall beta diversity, β_{repl} – the replacement component and β_{rich} – the richness differences component).

5. *beta.accum (comm1, comm2, tree, abund, func, runs)* computes an accumulation curve for the beta diversity between two communities (*comm1* and *comm2*). Both *comm1* and *comm2* are sampling units × species matrices, with either abundance or incidence data. The function returns three matrices of sampling units × diversity values, one per each of the three beta diversity measures (individuals and observed diversity).

6. *beta.multi (comm, tree, abund, func, raref, runs)* calculates beta diversity among multiple sites with possible rarefaction. The multiple sites measure is calculated as the average or variance of all pairwise values. The argument *comm* is a sites × species matrix, with either abundance or incidence data. The function returns a matrix of beta measures × diversity values (average and variance).

RELIABILITY

7. *accuracy (accum, target)* calculates the scaled mean squared error of accumulation curves compared with a known true diversity value (*target*). The argument *accum* is a matrix resulting from the *alpha.accum* or *beta.accum* functions (sampling units × diversity values). The function returns a vector with accuracy values for all observed and estimated curves.

8. *slope (accum)* calculates the slope between adjacent points along the accumulation curves produced by either *alpha.accum* or *beta.accum* (the single argument *accum*). This value reflects the expected gain in diversity when sampling a new individual. The function returns a matrix of sampling units × slope values.

DATA SETS

The BAT package includes five data sets. Three of them depict the abundance of 338 species of spiders (Araneae) in each of either 192 or 320 sampling units collected in three sites in Portugal (*arrabida*, *geres* and *guadiana*, see details in Cardoso 2009 and references therein). The data set *phylotree* represents an approximation to the phylogenetic tree for the 338 species. The tree is based on the Linnean hierarchy, with different suborders separated by 1 unit, families by 0·75, genera by 0·5 and species by 0·25. Finally, the data set *functree* represents the functional tree for these same species. Several traits were recorded for each species: average size, type of web, type of hunting, stenophagy, vertical stratification in vegetation and circadian activity (Cardoso *et al.* 2011). The species × traits matrix was subjected to a hierarchical clustering procedure using UPGMA with Gower distances to produce a final dendrogram.

EXAMPLES

Here, we provide an example of a typical R session using the BAT package. All examples can be reproduced using the included data sets. We start by loading the package and all data.

```
library(BAT)
data(arrabida)
data(geres)
data(guadiana)
data(phylotree)
data(functree)
```

We create a new matrix with species abundance per site.

```
comm < - rbind(colSums(arrabida), colSums(geres), colSums(guadiana))
sites < - c('Arrabida', 'Geres', 'Guadiana')
row.names(comm) < - sites
```

and calculate alpha diversity (TD, PD and FD) for all sites.

```
alpha(comm)
alpha(comm, phylotree)
alpha(comm, functree)
```

Although Gerês has more species and phylogenetic diversity, it is less diverse functionally than Arrábida. But are these results determined by the sampled abundance? We may rarefy to 1000 individuals and plot the results.

```
raref.td < - alpha(comm, raref = 1000)
raref.pd < - alpha(comm, phylotree, raref = 1000)
raref.fd < - alpha(comm, functree, raref = 1000)
par(mfrow=c(1,3))
boxplot(t(raref.td), names = sites, main='TD', pars = list (boxwex = 0·8, staplewex = 0, outwex = 0·5), range=1·5, lwd=0·8, lty=1, col = c('grey30', 'grey60', 'grey90'), ylab = expression(alpha))
boxplot(t(raref.pd), names = sites, main='PD', pars = list (boxwex = 0·8, staplewex = 0, outwex = 0·5), range=1·5, lwd=0·8, lty=1, col = c('grey30', 'grey60', 'grey90'), ylab = expression(alpha))
boxplot(t(raref.fd), names = sites, main='FD', pars = list (boxwex = 0·8, staplewex = 0, outwex = 0·5), range=1·5, lwd=0·8, lty=1, col = c('grey30', 'grey60', 'grey90'), ylab = expression(alpha))
```

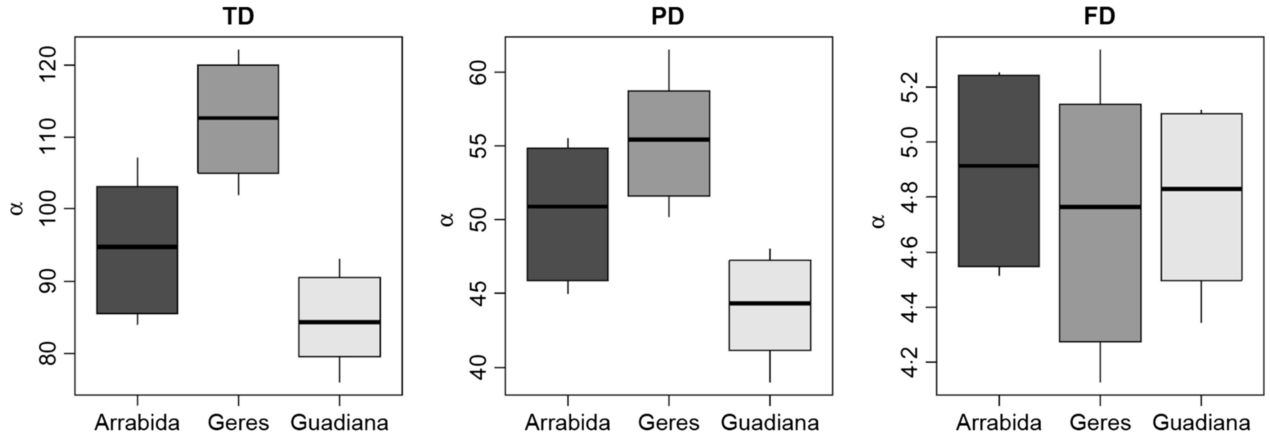

Fig. 1. Results of the rarefaction procedure, for 1000 individuals and 100 runs, for alpha taxon (TD), phylogenetic (PD) and functional (FD) diversity performed with the function *alpha* of the R package BAT. Arrabida, Geres and Guadiana are three sites exhaustively sampled for spiders in Portugal and whose data are included in the package (see text for details).

The three sites present similar FD values when data are rarified (Fig. 1). But results may be better understood if observed diversity is corrected through estimates (Chao1P is used throughout for illustration purposes only).

alpha.estimate(comm)

alpha.estimate(comm, phylotree)

alpha.estimate(comm, functree)

Estimates indicate that Guadiana may be more functionally diverse than the other sites. The reliability of any estimate can be judged based on the accumulation curves.

acc.arrabida < - alpha.accum(arrabida)

par(mfrow = c(1,2))

plot(acc.arrabida[,2], acc.arrabida[,17], xlab = 'Individuals', ylab = 'Chao 1P')

The Chao1P estimator values seem to have reached the asymptote (Fig. 2), which can be tested with

plot(acc.arrabida[,2], slope(acc.arrabida)[,17], xlab = 'Individuals', ylab = 'Slope')

The slope in the last part of the curve is null or very close to it (Fig. 2), an indication of reliability of the estimate (although other indices should be taken into account). If we knew the true diversity of the site (e.g. species richness = 170), the accuracy of the different estimators could be tested as

accuracy(acc.arrabida, target = 170)

In this case, Chao2P has the smallest value (0·017), being considered the most accurate.

Beta diversity between each pair of sites can be calculated as:

beta(comm)

If *func* is not specified, the Jaccard dissimilarity is used by default. Total beta diversity is larger between Gerês and Guadiana ($\beta_{\mathrm{total}} = 0{\cdot}901$), and although dissimilarity between sites is mostly explained by replacement of species ($\beta_{\mathrm{repl}} > 0{\cdot}632$ in all cases), differences in richness are larger between those two sites ($\beta_{\mathrm{rich}} = 0{\cdot}268$). Multiple phylogenetic beta with rarefaction made to the same abundance as the site with least individuals sampled can be calculated as

beta.multi(comm, phylotree, raref = 1)

Again, β_{repl} is clearly dominant overall (averages of $\beta_{\mathrm{total}} = 0{\cdot}731$, $\beta_{\mathrm{repl}} = 0{\cdot}656$, $\beta_{\mathrm{rich}} = 0{\cdot}075$). Finally, we may want to check the reliability of beta diversity values according to sampling intensity, taking abundances into account

acc.beta < - beta.accum(arrabida, geres, abund = TRUE)

par(mfrow = c(1,3))

plot(acc.beta[,2], xlab = 'Sampling units', ylab = expression (beta[total]))

plot(acc.beta[,3], xlab = 'Sampling units', ylab = expression (beta[repl]))

plot(acc.beta[,4], xlab = 'Sampling units', ylab = expression (beta[rich]))

In all cases, the diversity values seem to stabilize and be reliable early in the accumulation process (Fig. 3).

STRENGTHS

The package BAT implements a number of methods that are not available in any other package or software. These include

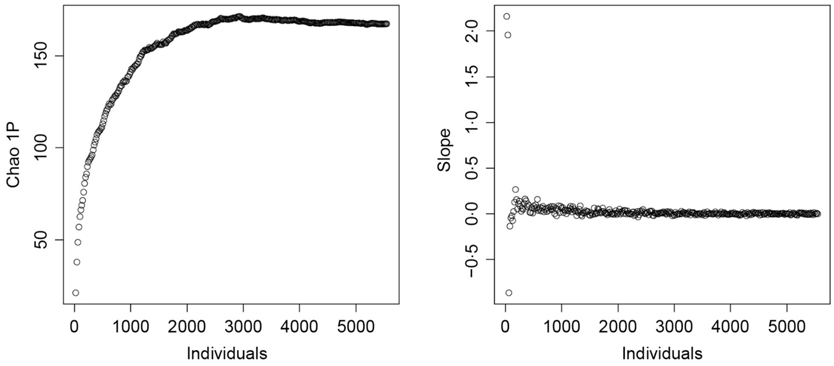

Fig. 2. Accumulation curve for the nonparametric estimator Chao1P provided by the function *alpha.accum* of the R package BAT and the slopes between every consecutive point along the curve, produced by the function *slope*.

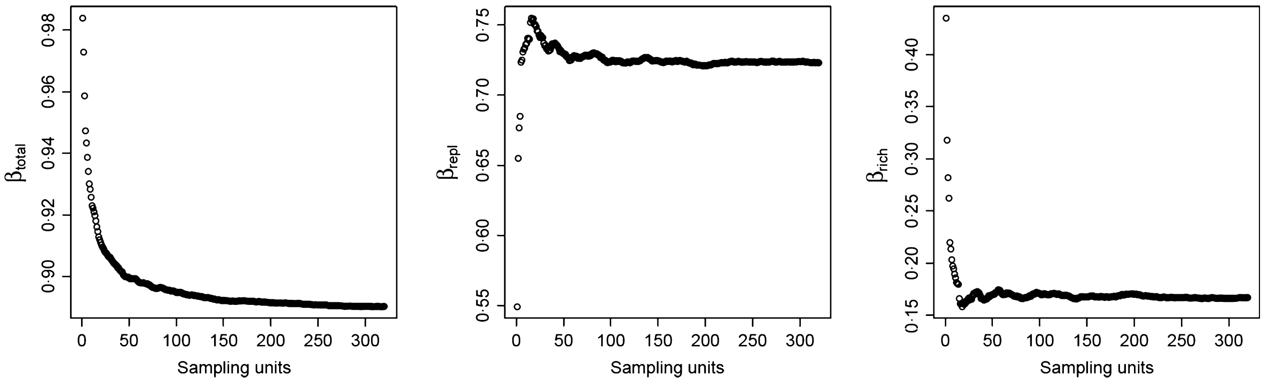

Fig. 3. Accumulation curve for the beta diversity values provided by the function *beta.accum* of the R package BAT.

the estimation of phylogenetic and functional diversity from incomplete sampling, the computation of accumulation curves for beta diversity and the partitioning of beta diversity, including with abundance, phylogenetic or functional data, in its replacement and richness difference components. The output of the functions can be easily used as response variables in statistical models using environmental, spatial or temporal data as explanatory variables.

ALTERNATIVES

Other software packages and R functions exist that calculate parts of the same or alternative methods to those provided by BAT. For species richness estimation, these include the EstimateS software (Colwell 2013), and the estimateR and estaccumR functions of the R package vegan (Oksanen *et al.* 2013). However, none of these is able to deal with phylogenetic or functional data. For beta diversity decomposition, the function *beta.div.comp* found in Legendre (2014) provides similar results to ours for the coefficients computed by both packages. The package betapart (Baselga & Orme 2012) implements alternative methods, although we disagree with the methodological framework developed by these authors (see Carvalho, Cardoso & Gomes 2012; Carvalho *et al.* 2013; Cardoso *et al.* 2014c).

CITATION

Researchers using BAT in a published paper should cite this article and in addition can also cite the BAT package directly. Updated citation information can be obtained by typing: citation ('BAT')

Acknowledgements

We thank all the co-authors of past papers that helped us developing the methods and data sets used here. Pierre Legendre and David Nipperess provided very useful comments on this manuscript.

Data accessibility

All data used in the manuscript are included in the R package: http://cran.r-project.org/web/packages/BAT/

Funding

F.R. was partly supported by FCT project PTDC/BIA-BIC/119255/2010.

References

Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga, A. (2012) The relationship between species replacement, dissimilarity derived from nestedness and nestedness. *Global Ecology and Biogeography*, **21**, 1223–1232.

Baselga, A. & Orme, C.D.L. (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Cardoso, P. (2009) Standardization and optimization of arthropod inventories – the case of Iberian spiders. *Biodiversity and Conservation*, **18**, 3949–3962.

Cardoso, P., Borges, P.A.V. & Veech, J.A. (2009) Testing the performance of beta diversity measures based on incidence data: the robustness to undersampling. *Diversity and Distributions*, **15**, 1081–1090.

Cardoso, P., Rigal, F. & Carvalho, J.C. (2014b) *BAT: Biodiversity Assessment Tools*. R package version 1.0.1. URL http://cran.r-project.org/package = BAT [accessed 22 October 2014]

Cardoso, P., Pekar, S., Jocque, R. & Coddington, J.A. (2011) Global patterns of guild composition and functional diversity of spiders. *PLoS One*, **6**, e21710.

Cardoso, P., Rigal, F., Borges, P.A.V. & Carvalho, J.C. (2014a) A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. *Methods in Ecology and Evolution*, **5**, 452–461.

Cardoso, P., Rigal, F., Carvalho, J.C., Fortelius, M., Borges, P.A.V., Podani, J. & Schmera, D. (2014c) Partitioning taxon, phylogenetic and functional beta diversity into replacement and richness difference components. *Journal of Biogeography*, **41**, 749–761.

Carvalho, J.C., Cardoso, P. & Gomes, P. (2012) Determining the relative roles of species replacement and species richness differences in generating beta-diversity patterns. *Global Ecology and Biogeography*, **21**, 760–771.

Carvalho, J.C., Cardoso, P., Borges, P.A.V., Schmera, D. & Podani, J. (2013) Measuring fractions of beta diversity and their relationships to nestedness: a theoretical and empirical comparison of novel approaches. *Oikos*, **122**, 825–834.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Coddington, J.A., Agnarsson, I., Miller, J.A., Kuntner, M. & Hormiga, G. (2009) Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. *Journal of Animal Ecology*, **78**, 573–584.

Colwell, R.K. (2013) *EstimateS: Statistical Estimation of Species Richness and Shared Species From Samples. Version 9.1*. URL http://purl.oclc.org/estimates [accessed 01 October 2014]

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London – Biological Sciences*, **345**, 101–118.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Heltshe, J. & Forrester, N.E. (1983) Estimating species richness using the jackknife procedure. *Biometrics*, **39**, 1–11.

Legendre, P. (2014) Interpreting the replacement and richness difference components of beta diversity. *Global Ecology and Biogeography*, **23**, 1324–1334.

Lennon, J.J., Koleff, P., Greenwood, J.J.D. & Gaston, K.J. (2001) The geographical structure of British bird distributions: diversity, spatial turnover and scale. *Journal of Animal Ecology*, **70**, 966–979.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2013) *Vegan: Community Ecology Package*. R package version 2.0-10. URL cran.r-project.org/package = vegan [accessed 01 October 2014].

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

Petchey, O.L. & Gaston, K.J. (2002) Functional diversity (FD), species richness and community composition. *Ecology Letters*, **5**, 402–411.

Podani, J. & Schmera, D. (2011) A new conceptual and methodological framework for exploring and explaining pattern in presence-absence data. *Oikos*, **120**, 1625–1638.

Soberón, M.J. & Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480–488.

Stegen, J.C. & Hurlbert, A.H. (2011) Inferring eco-evolutionary processes from taxonomic, phylogenetic and functional trait β-diversity. *PLoS One*, **6**, e20906.

Williams, P.H., de Klerk, H.M. & Crowe, T.M. (1999) Interpreting biogeographical boundaries among Afro-tropical birds: spatial patterns in richness gradients and species replacement. *Journal of Biogeography*, **26**, 459–474.

Received 21 August 2014; accepted 11 November 2014
Handling Editor: Steven Kembel