Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution doi: 10.1111/j.2041-210X.2012.00247.x

APPLICATION

Treebase: an R package for discovery, access and manipulation of online phylogenies

Carl Boettiger[1]* and Duncan Temple Lang[2]

[1]*Center for Population Biology, University of California, Davis, CA,95616,USA; and* [2]*Department of Statistics, University of California, Davis, CA,95616,USA*

Summary

1. The TreeBASE portal is an important and rapidly growing repository of phylogenetic data. The R statistical environment has also become a primary tool for applied phylogenetic analyses across a range of questions, from comparative evolution to community ecology to conservation planning.

2. We have developed `treebase`, an open-source software package (freely available from http://cran.r-project.org/web/packages/treebase) for the R programming environment, providing simplified, *programmatic* and interactive access to phylogenetic data in the TreeBASE repository.

3. We illustrate how this package creates a bridge between the TreeBASE repository and the rapidly growing collection of R packages for phylogenetics that can reduce barriers to discovery and integration across phylogenetic research.

4. We show how the `treebase` package can be used to facilitate replication of previous studies and testing of methods and hypotheses across a large sample of phylogenies, which may help make such important reproducibility practices more common.

Key-words: application programming interface, database, programmatic, R, software, TreeBASE, workflow

Introduction

Applications that use phylogenetic information as part of their analyses are becoming increasingly central to both evolutionary and ecological research. The exponential growth in genetic sequence data available for all forms of life has driven rapid advances in the methods that can infer the phylogenetic relationships and divergence times across different taxa (Huelsenbeck & Ronquist 2001; Stamatakis 2006; Drummond & Rambaut 2007). The product of one field has become the raw data of the next. Unfortunately, while the discipline of bioinformatics has emerged to help harness and curate the wealth of genetic data with cutting-edge computer science, statistics and Internet technologies, its counterpart in evolutionary informatics remains 'scattered, poorly documented, and in formats that impede discovery and integration' (Parr *et al.* 2011). Our goal in developing the `treebase` package is to provide steps to reduce these challenges through programmatic and interactive access between the repositories that store this data and the software tools commonly used to analyse them.

The R statistical environment (R Development Core Team 2012) has become a dominant platform for researchers using phylogenetic data to address a rapidly expanding set of questions in ecological and evolutionary processes. These methods include, but are not limited to, ancestral state reconstruction (Paradis 2004; Butler & King 2004), diversification analysis (Paradis 2004; Rabosky 2006; Harmon *et al.* 2008), identifying trait-dependent speciation and extinction rates, (Fitzjohn 2010; Goldberg, Lancaster & Ree 2011; Stadler 2011b), quantifying the rate and tempo of trait evolution (Butler & King 2004; Harmon *et al.* 2008; Eastman *et al.* 2011), identifying evolutionary influences and proxies for community ecology (Webb, Ackerly & Kembel 2008; Kembel *et al.* 2010), connecting phylogeny data to climate patterns (Warren, Glor & Turelli 2008; Evans *et al.* 2009), and simulation of speciation and character evolution (Harmon *et al.* 2008; Stadler 2011a; Boettiger, Coop & Ralph 2012), as well as various manipulations and visualizations of phylogenetic data (Paradis 2004; Schliep 2010; Jombart, Balloux & Dray 2010; Revell *et al.* 2011). A more comprehensive list of R packages by analysis type is available on the phylogenetics taskview, http://cran.r-project.org/web/views/Phylogenetics.html. Libraries and packages are developed for use in other general purpose programming environments and languages, including Java (Maddison & Maddison 2011), MATLAB (Blomberg, Thedore Garland & Ives 2003) and Python (Sukumaran & Holder 2010) and online interfaces (Martins 2004). In particular, the Bio::Phylo toolkit (Vos *et al.* 2011) not only provides a PERL implementation of some of the common phylogenetic simulation and visualization tools found in these R libraries, but can already provide

*Correspondence author. E-mail: cboettig@ucdavis.edu

programmatic access to TreeBASE. Our goal is to bring similar functionality to the larger suite of applied phylogenetics methods and user in the R community.

TreeBASE (http://treebase.org) is an online repository of phylogenetic data (e.g. trees of species, populations or genes) that have been published in a peer-reviewed academic journal, book, thesis or conference proceedings (Sanderson *et al*. 1994; Morell 1996). The database can be searched through an online interface that allows users to find a phylogenetic tree from a particular publication, author or taxa of interest. TreeBASE provides an application programming interface (API) that lets computer applications make queries to the database, known as PhyloWS (Vos *et al*. 2010). Our `treebase` package uses this API to create a direct link between this data and the R environment. This has several immediate and important benefits:

1. *Data discovery:* Users can leverage the rich, higher-level programming environment provided by the R environment to better identify data sets appropriate for their research by iteratively constructing queries for data sets that match appropriate metadata requirements.

2. *Programmatic data access:* Many tasks that are theoretically made possible by the creation of the Web-based interface to the TreeBASE repository are not pursued because they would be too laborious for an exploratory analysis. The ability to use programmatic access across data sets to automatically download and perform a reproducible and systematic analysis using the rich set of tools available in R opens up new avenues for research.

3. *Automatic updating:* The TreeBASE repository is expanding rapidly. The scriptable nature of analyses run with our `treebase` package means that a study can be rerun on the latest version of the repository without additional effort, but with potential new information.

PROGRAMMATIC WEB ACCESS

The TreeBASE repository makes data accessible via Web queries through a RESTful (REpresentational State Transfer) interface, which supplies search conditions in the address URL. The repository returns the requested data in XML (extensible markup language) format, conforming to the RSS1.0 standard. Because the RSS1.0 format allows the search results to also be viewed in a human-readable format in common browsers such as Safari and Firefox, the `treebase` package echoes this URL to the console, so that the user can explore the results in the browser as well. The `treebase` package uses the `RCurl` package (Temple Lang 2012a) to make queries over the Web to the repository, and the `XML` package (Temple Lang 2012b) to parse the Web page returned by the repository into meaningful R data objects. While these querying and parsing functions comprise most of the code provided in the `treebase` package, they are hidden from the end-user who can interact with these rich data retrieval and manipulation tools to access data from these remote repositories in much the same way as data locally available on the users hard-disk.

While the TreeBASE repository provides phylogenies in both the traditional Nexus file format and the more data-rich NeXML format (Vos *et al*. 2012), none of the R packages currently available for phylogenetic research are positioned to read these NeXML files. The next version of the `treebase` package will provide the extraction of metadata information from the NeXML through XML parsing.

BASIC QUERIES

The `treebase` package allows these queries to be made directly from R. Programmatic access also allows a user to go beyond the utilities of the Web interface, constructing more complicated queries and allowing the user to maintain a record of the commands used to collect their data as an R script. Scripting the data-gathering process helps reduce errors and assists in replicating the analysis later, either by the authors or by other researchers (Peng 2011).

The `search_treebase` function forms the base of the `treebase` package. Table 1 lists each of the types of queries available through the `search_treebase` function. This list can also be found in the function documentation through the R command `?search_treebase`.

Any of the queries available on the Web interface can now be made directly from R, including downloading and importing a phylogeny into the R session. For instance, one can search for phylogenies containing dolphin taxa, 'Delphinus,' or all phylogenies submitted by a given author, 'Huelsenbeck' using the R commands

```
search_treebase('Delphinus', by='taxon')
search_treebase('Huelsenbeck', by='author')
```

The `search_treebase` function returns the matching phylogenies as an R object, ready for analysis. The package documentation provides many examples of possible queries.

ACCESSING ALL PHYLOGENIES

For certain applications, a user may wish to download all the available phylogenies from TreeBASE. Using the

Table 1. Queries available in `search_treebase`. The first argument is the keyword used in the query such as an author's name, and the second argument indicates the type of query (i.e. 'author')

Search 'by='	Purpose
abstract	search terms in the publication abstract
author	match authors in the publication
subject	Matches in the subject terms
doi	The unique object identifier for the publication
ncbi	NCBI identifier number for the taxon
kind.tree	Kind of tree (Gene tree, species tree, barcode tree)
type.tree	Type of tree (Consensus or Single)
ntax	Number of taxa in the matrix
quality	A quality score for the tree, if it has been rated.
study	Match words in the title of the study or publication
taxon	Taxon scientific name
id.study	TreeBASE study ID
id.tree	TreeBASE's unique tree identifier (Tr.id)
id.taxon	Taxon identifier number from TreeBase
tree	The title for the tree

`cache_treebase` function allows a user to download a local copy of all trees. Because direct database dumps are not currently available from treebase.org, this function has intentional delays to avoid overtaxing the TreeBASE servers and should be allowed a full day to run.

```
treebase <- cache_treebase()
```

Users should still be mindful that these servers are a shared community resource and not place many queries at once. Users running large jobs should consider joining the TreeBASE mailing list (http://sourceforge.net/mailarchive/forum.php?forum_name=treebase-users) to discuss such queries ahead of time.

Once run, the cache is saved compactly in memory where it can be easily and quickly restored. For convenience, the `treebase` package comes with a copy already cached, which can be loaded into memory.

```
data(treebase)
```

The cache included in the package will be updated during major package revisions. The timestamp of the cache provided can be viewed in the help file for the data object using the command `?treebase` (Current cache is May 14, 2012). All queries from `metadata()` and `search_treebase()` are run against the current online version of the database.

All of the examples shown in this manuscript are run as shown using the `knitr` package for authoring dynamic documents (Xie 2012), which helps ensure the results shown are reproducible. These examples can be updated by copying and pasting the code shown into the R terminal, or by recompiling the entire manuscript from the source files found on the development Web page for the TreeBASE package, github.com/ropensci/treebase. Data were accessed to produce the examples shown on Thursday 9 August 2012 at 10:51:51 Pacific Time.

Data discovery in TreeBASE

Data discovery involves searching for existing data that meet certain desired characteristics. Such searches take advantage of metadata – summary information describing the data entries provided in the repository. The Web repository uses separate interfaces (APIs) to access metadata describing the publications associated with the data entered, such as the publisher, year of publication, and a different interface to describe the metadata associated with an individual phylogeny, such as the number of taxa or the kind of tree (e.g. gene tree vs. species tree). The `treebase` package can query these individual sources of metadata separately, but this information is most powerful when used in concert – allowing the construction of complicated searches that cannot be automated through the Web interface. The `metadata` function updates a list of all available metadata from both APIs and returns this information as an R `data.frame`.

```
meta <- metadata()
```

From the number of rows of the metadata list, we see that there are currently 3164 published studies in the database. The field columns provided by `metadata` are listed in Table 2.

Metadata can also be used to reveal trends in the data deposition which may be useful in identifying patterns or biases in research or emerging potential types of data. As a simple example, we look at trends in the submission patterns of publishers over time:

```
date <- meta[['date']]
pub <- meta[['publisher']]
```

Many journals have only a few submissions, so we will label any not in the top 10 contributing journals as 'Other':

```
topten <- sort(table(pub), decreasing=TRUE)[1:10]
meta[['publisher']] <- as.character(meta[['publisher']])
meta[['publisher']][!(pub %in% names(topten))] <- 'other'
meta[['publisher']] <- as.factor(meta[['publisher']])
```

We plot the distribution of publication years for phylogenies deposited in TreeBASE, colour coding by publisher in Fig. 1.

```
library(ggplot2)
library(reshape2)
df <- acast(meta, date ~ publisher, value.var='publisher', length)
df <- melt(df, varnames=c('date', 'publisher'))
ggplot(df) + geom_area(aes(x=date,y=value, fill=publisher))
```

Typically, we are interested in the metadata describing the phylogenies themselves rather than just in the publications in which they appeared. Phylogenetic metadata include features such as the number of taxa in the tree, a quality score (if available), kind of tree (gene tree, species tree or barcode tree) or whether the phylogeny represents a consensus tree from a distribution or just a single estimate.

Even simple queries can illustrate the advantage of interacting with TreeBASE data through an R interface has over the Web interface. A Web interface can only perform the tasks built in by design. For instance, rather than performing six separate searches to determine the number of consensus vs. single phylogenies available for each kind of tree, we can construct a 2 by 2 table with a single line of code:

```
table(meta[['kind']],meta[['type']])
```

Table 2. Columns of metadata available from the `metadata` function

metadata field	description
Study.id	TreeBASE study ID
Tree.id	TreeBASE's unique tree identifier
kind	Kind of tree (Gene tree, species tree, barcode tree)
type	Type of tree (Consensus or Single)
quality	A quality score for the tree, if it has been rated.
ntaxa	Number of taxa in the matrix
date	Year the study was published
author	First author in the publication
title	The title of the publication

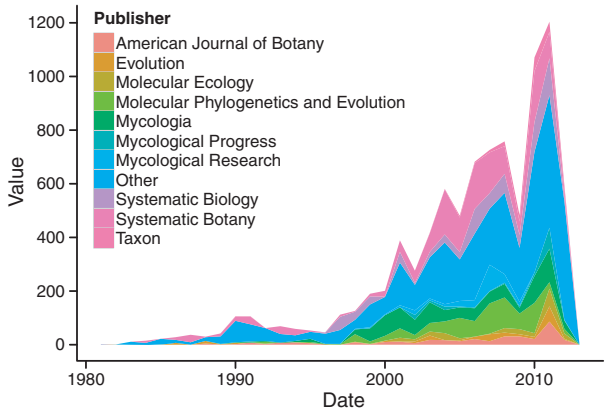

Fig. 1. Histogram of publication dates by year, with the code required to generate the figure.

Reproducible computations

Reproducible research has become a topic of increasing interest in recent years, and facilitating access to data and using scripts that can replicate analyses can help lower barriers to the replication of statistical and computational results (Schwab, Karrenbach & Claerbout 2000; Gentleman & Temple Lang 2004; Peng 2011). The `treebase` package facilitates this process, as we illustrate in a simple example.

Consider the shifts in speciation rate identified by Derryberry *et al.* (2011) on a phylogeny of ovenbirds and treecreepers. We will seek to not only replicate the results the authors obtained by fitting the models provided in the R package `laser` (Rabosky 2006), but also compare them against methods presented in Stadler (2011b) and implemented in the package `TreePar`, which permits speciation models that were not available to Derryberry *et al.* (2011) at the time of their study.

OBTAINING THE TREE

The most expedient way to identify the data uses the digital object identifier (doi) at the top of most articles, which we use in a call to the `search_treebase` function, such as

```
results <- search_treebase('10.1111/j.15585646.2011.01374.x','doi')
```

The search returns a list, because some publications can contain many trees. In this case our phylogeny is in the only element of the list.

Having imported the phylogenetic tree corresponding to this study, we can quickly replicate their analysis of which diversification process best fits the data. These steps can be easily implemented using the phylogenetics packages we have just mentioned.

For instance, we can calculate the branching times of each node on the phylogeny,

```
bt <- branching.times(results[[1]])
```

and then begin to fit each model the authors have tested, such as the pure birth model,

```
yule <- pureBirth(bt)
```

or the birth–death model,

```
birth_death <- bd(bt)
```

The estimated models are now available in the active R session where we can further explore them as we go along. The appendix shows the estimation and comparison of all the models originally considered by Derryberry *et al.* (2011).

In this fast-moving field, new methods often become available between the time of submission and the time of publication of a manuscript. For instance, the more sophisticated models introduced in the study by Stadler (2011b) were not used in the original study, but have since been made available in the recent package, `TreePar`. These richer models permit a shift in the speciation or extinction rate to occur multiple times throughout the course of the phylogeny.

We load the new method and format the phylogeny appropriately using the R commands:

```
library(TreePar)
x <- sort(getx(results[[1]]), decreasing = TRUE)
```

As a comparison of speciation models is not the focus of this paper, the complete code and explanation for these steps are provided in an appendix. Happily, this analysis confirms the original author's conclusions, even when the more general models of Stadler (2011b) are considered.

Analyses across many phylogenies

Large-scale comparative analyses that seek to characterize evolutionary patterns across many phylogenies are increasingly common in phylogenetic methods research (e.g. McPeek & Brown 2007; Phillimore & Price 2008; McPeek 2008; Quental & Marshall 2010; Davies *et al.* 2011). Sometimes referred to by their authors as meta-analyses, these approaches have focused on re-analysing phylogenetic trees collected from many different earlier publications. This is a more direct approach than the traditional concept of meta-analysis where statistical results from earlier studies are weighted by their sample size without being able to access the raw data. Because the identical analysis can be repeated on the original data from each study, this approach avoids some of the statistical challenges inherent in traditional meta-analyses summarizing results across heterogeneous approaches.

To date, researchers have gone through heroic efforts simply to assemble these data sets from the literature. As described in McPeek & Brown (2007) (emphasis added).

> One data set was based on 163 published species-level molecular phylogenies of arthropods, chordates, and molluscs. A PDF format file of each article was obtained, and a digital snapshot of the figure was taken in Adobe Acrobat 7.0. This

.

	Consensus	Single
Barcode Tree	1	4
Gene Tree	65	134
Species Tree	2863	5857

image was transferred to a PowerPoint (Microsoft) file and printed on a laser printer. The phylogenies included in this study are listed in the appendix. *All branch lengths were measured by hand from these printed sheets using dial calipers*.

Despite the recent emergence of digital tools that could now facilitate this analysis without mechanical calipers (e.g. treesnatcher, Laubach & von Haeseler 2007), it is easier and less error-prone to pull properly formatted phylogenies from the database for this purpose. Moreover, as the available data increase with subsequent publications, updating earlier meta-analyses can become increasingly tedious. Using `treebase`, a user can apply any analysis they have written for a single phylogeny across the entire collection of suitable phylogenies in TreeBASE, which can help overcome such barriers to discovery and integration at this large scale. Using the functions we introduced above, we provide a simple example that computes the gamma statistic of Pybus & Harvey (2000), which provides a measure of when speciation patterns differ from the popular birth–death model.

TESTS ACROSS MANY PHYLOGENIES

A standard test of the constant rate of diversification is the gamma statistic of Pybus & Harvey (2000) which tests the null hypothesis that the rates of speciation and extinction are constant. Under the null hypothesis, the gamma statistic is normally distributed about 0; values larger than 0 indicate that internal nodes are closer to the tip than expected, while values smaller than 0 indicate nodes farther from the tip than expected. In this section, we collect all phylogenetic trees from TreeBASE and select those with branch length data that we can time-calibrate using tools available in R. We can then calculate the distribution of this statistic for all available trees and compare these results with those from the analyses mentioned above. As we will use all trees in the repository, we use the cached copy of TreeBASE phylogenies described above to reduce load on TreeBASE servers.

We will only be able to use those phylogenies that include branch length data, which we can determine from the `have_branchlength` function in the `treebase` package. We drop those that do not from the data set,

```
have <- have_branchlength(treebase)
branchlengths <- treebase[ have]
```

Like most comparative methods, this analysis will require ultrametric trees (branch lengths proportional to time, rather than to the nucleotide substitution rate). As most of these phylogenies are calibrated with branch length proportional to mutational step, we must time-calibrate each of them first. The following function drops trees that cannot meet the assumptions of the time-calibration function.

```
timetree <- function(tree)
try(chronoMPL(multi2di(tree)), silent=TRUE)
tt <- drop_nontrees(sapply(branchlengths,
timetree))
```

At this point, we have 1396 time-calibrated phylogenies over which we will apply the diversification rate analysis. Computing the gamma test statistic to identify deviations from the constant-rates model takes a single line,

```
gammas <- sapply(tt, gammaStat)
```

and the resulting distribution of the statistic across available trees is shown Fig 2. While researchers have often considered this statistic for individual phylogenies, we are unaware of any study that has visualized the empirical distribution of this statistic across thousands of phylogenies. The overall distribution appears slightly skewed towards positive values. This could indicate increasing rate of speciation or constant extinction rates. While differences in sampling may account for much of the spread observed, the position and identity of outlier phylogenies could suggest new hypotheses and potential directions for further exploration.

```
qplot(gammas)+xlab('gamma statistic')
```

Conclusion

While we have focused on examples that require no additional data beyond the phylogeny, a wide array of methods combine this data with information about the traits, geography or ecological community of the taxa represented. In such cases, we would need programmatic access to the trait data as well as the phylogeny. The Dryad digital repository (http://datadryad.org) is an effort in this direction. While programmatic access to the repository is possible through the `rdryad` package (Chamberlain, Boettiger & Ram 2012), variation in data formatting must first be overcome before similar direct access to the data is possible. Dedicated databases such as FishBASE (http://fishbase.org) may be another alternative, where morphological data can be queried for a list of species using the `rfishbase` package (Boettiger 2012). The development of similar software for programmatic data access will rapidly extend the space and scale of possible analyses.

The recent advent of mandatory data archiving in many of the major journals publishing phylogenetics-based research (e.g. Fairbairn 2010; Piwowar, Vision & Whitlock 2011; Whitlock *et al*. 2010) is a particularly promising development that should continue to fuel the trend of submissions seen in Fig. 1. Accompanied by faster and more inexpensive techniques of NextGen sequencing, and the rapid expansion in phylogenetic applications, we anticipate this rapid growth in

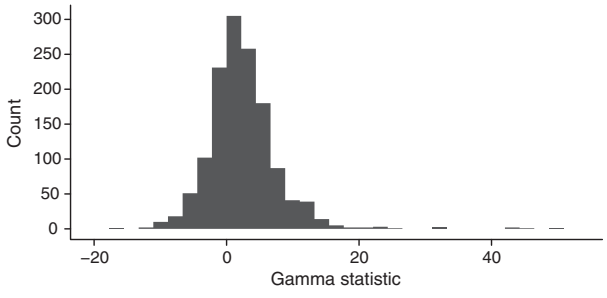

Fig. 2. Distribution of the gamma statistic across phylogenies in TreeBASE. Strongly positive values are indicative of an increasing rate of evolution (excess of nodes near the tips), very negative values indicate an early burst of diversification (an excess of nodes near the root).

available phylogenies will continue. Faced with this flood of data, programmatic access becomes not only increasingly powerful but an increasingly necessary way to ensure we can still see the forest for all the trees.

Acknowledgements

CB wishes to thank S. Price for feedback on the manuscript, the TreeBASE developer team for building and supporting the repository, and all contributors to TreeBASE. CB is supported by a Computational Sciences Graduate Fellowship from the Department of Energy under grant number DE-FG02-97ER25308. The `treebase` package is part of the rOpenSci project (ropensci.org).

References

Blomberg, S.P., Theodore Garland, J.R. & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Boettiger, C., Temple Lang, D. & Wainwright, P.C. (2012) rfishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology* (in press).

Boettiger, C., Coop, G. & Ralph, P. (2012) Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, **66**, 2240–2251.

Butler, M.A. & King, A.A. (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, **164**, 683–695.

Chamberlain, S., Boettiger, C. & Ram, K. (2012) rdryad: Dryad API interface. http://www.github.com/ropensci/rdryad.

Davies, T.J., Allen, A.P., Borda-de-gua, L., Regetz, J. & Melin, C.J. (2011) Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. *Evolution*, **65**, 1841–1850.

Derryberry, E.P., Claramunt, S., Derryberry, G., Terry Chesser, R., Cracraft, J., Aleixo, A., Prez-Emn, J., Remsen Jr, J.V. & Brumfield, R.T. (2011) Lineage diversification and morphological evolution in a large-scale continental radiation: the neotropical ovenbirds and woodcreepers (Aves: Furnariidae). *Evolution*, **65**, 2973–2986.

Drummond, A.J., & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L. & Harmon, L.J. (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, **65**, 3578–3589.

Evans, M.E.K., Smith, S.A., Flynn, R.S. & Donoghue, M.J. (2009) Climate, niche evolution, and diversification of the 'bird-cage' evening primroses (Oenothera, sections Anogra and Kleinia). *The American Naturalist*, **173**, 225–240.

Fairbairn, D.J. (2010) The advent of mandatory data archiving. *Evolution*, **65**, 1–2.

Fitzjohn, R.G. (2010) Quantitative traits and diversification. *Systematic Biology*, **59**, 619–633.

Gentleman, R. & Temple Lang, D. (2004) Statistical analyses and reproducible research. *Bioconductor Project Working Papers*, http://www.bepress.com/bioconductor/paper2.

Goldberg, E.E., Lancaster, L.T. & Ree, R.H. (2011) Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, **60**, 451–465.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. & Challenger, W. (2008) Geiger: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

Huelsenbeck, J.P. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Jombart, T., Balloux, F. & Dray, S. (2010) Adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics*, **26**, 1907–1909.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Laubach, T. & von Haeseler, A. (2007) TreeSnatcher: coding trees from images. *Bioinformatics*, **23**, 3384–3385.

Maddison, W.P. & Maddison, D.R. (2011) Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org.

Martins, E.P. (2004) *COMPARE, version Computer programs for the statistical analysis of comparative data*. Department of Biology, Indiana University, Bloomington, Indiana. http://compare.bio.indiana.edu/.

McPeek, M.A. & Brown, J.M. (2007) Clade age and not diversification rate explains species richness among animal taxa. *The American Naturalist*, **169**, 97.

McPeek, M.A. (2008) The ecological dynamics of clade diversification and community assembly. *The American Naturalist*, **172**, 270.

Morell, V. (1996) TreeBASE: the roots of phylogeny. *Science*, **273**, 569.

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

Parr, C.S., Guralnick, R., Cellinese, N. & Page, R.D.M. (2011) Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology & Evolution*, **27**, 94–103.

Peng, R.D. (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.

Phillimore, A.B. & Price, T.D. (2008) Density-dependent cladogenesis in birds. *PLoS Biology*, **6**, 71.

Piwowar, H.A., Vision, T.J. & Whitlock, M.C. (2011) Data archiving is a good investment. *Nature*, **473**, 285–285.

Pybus, O.G., & Harvey, P.H. (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of The Royal Society B*, **267**, 2267–2272.

Quental, T.B. & Marshall, C.R. (2010) Diversity dynamics: molecular phylogenies need the fossil record. *Trends in Ecology & Evolution*, **25**, 434–441.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.

Rabosky, D.L. (2006) LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics Online*, **2**, 273–276.

Revell, L.J., Luke Mahler, D., Peres-Neto, P.R. & Redelings, B.D. (2011) A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution*, **66**, 135–146.

Sanderson, M.J., Donoghue, M.J., Piel, W. & Eriksson, T. (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany*, **81**, 183.

Schliep, K.P. (2010) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593,

Schwab, M., Karrenbach, N. & Claerbout, J. (2000) Making scientific computations reproducible. *Computing in Science & Engineering*, **2**, 61–67.

Stadler, T. (2011a) Simulating trees with a fixed number of extant species. *Systematic Biology*, **60**, 676–684.

Stadler, T. (2011b) Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, **108**, 6187–6192.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Sukumaran, J. & Holder, M.T. (2010) DendroPy: a Python Library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Temple Lang, D. (2012a) RCurl: General network (HTTP/FTP/...) client interface for R. http://cran.r-project.org/packageR¯Curl.

Temple Lang, D. (2012b) XML: Tools for parsing and generating XML within R and S-Plus. http://cran.r-project.org/packageX¯ML.

Vos, R.A., Lapp, H., Piel, W.H. & Tannen, V. (2010) TreeBase2: rise of the machines. *Nature Precedings*, http://hdl.handle.net/10101/npre.2010.4600.1.

Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E. *et al.* (2012) NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, **61**, 675–689.

Vos, R.A., Caravas, J., Hartmann, K., Jensen, M.A. & Miller, C. (2011) BIO:: Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics*, **12**, 63.

Warren, D.L., Glor, R.E. & Turelli, M. (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, **62**, 2868–2883.

Webb, C.O., Ackerly, D.D. & Kembel, S.W. (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, **24**, 2098–2100.

Whitlock, M.C., McPeek, M.A., Rausher, M.D., Rieseberg, L. & Moore, A.J. (2010) Data archiving. *The American Naturalist*, **175**, 145–146.

Xie, Y. (2012) knitr: A general-purpose package for dynamic report generation in R. http://yihui.name/knitr/.

Received 26 June 2012; accepted 13 August 2012
Handling Editor: Luke Harmon

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Reproducible computation: a diversification rate analysis.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.