An introduction to the `treebase` Package

Carl Boettiger[a,*], Duncan Temple Lang[b]

[a] *Center for Population Biology, University of California, Davis, United States*
[b] *Department of Statistics, University of California, Davis, United States*

Abstract

1. TreeBASE is an important and rapidly growing repository of phylogenetic data. The R statistical environment has become a primary tool for the applied phylogenetic analyses that use this kind of data for across a range of questions, from comparative evolution to community ecology to conservation planning.
2. We have developed `treebase`, an open-source package (freely available from http://cran.r-project.org/web/packages/treebase) for the R environment, providing simplified, programmatic and interactive access to phylogenetic data.
3. We illustrate how this package creates a bridge between the repository and the rapidly growing ecosystem of R packages for phylogenetics that can reduce barriers to discovery and integration across phylogenetic research.
4. We provide several examples, including one that replicates and extends the results of an analysis on a single phylogeny, and one which includes an analysis of changes in diversification rate across over 1000 phylogenies.

Keywords: R, software, API, TreeBASE, tools, e-science

1. 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 (Drummond and Rambaut, 2007; Huelsenbeck and Ronquist, 2001; Stamatakis, 2006). Just as this goldmine of available sequence data has led to the subsequent explosion of phylogenetic methods, and many other avenues of research, this rapid expanse of phylogenetic data now primes new innovations across ecology and evolution. Once again the product of one field has become the raw data of the next. But 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 technology, its counterpart in evolutionary informatics remains "scattered, poorly documented, and in formats that impede discovery and integration" (Parr et al., 2011). The goal of this paper is to address this gap by illustrating how programmatic and interactive access between the repositories that store this phylogenetic data and the software tools most commonly used to analyze them can breach some of these barriers to discovery and integration.

While the task of inferring phylogenies from sequence data remains dominated by dedicated compiled software such as MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond and Rambaut, 2007), RAXML (Stamatakis, 2006), the research methods that make use of the phylogenies they infer are largely based in R. To distinguish between the processes of inferring the phylogeny and these applications, We will refer collectively to this area of research using phylogenetic relationships as input data as applied phylogenetics.

The R statistical environment (R Development Core Team, 2012) has become a dominant platform for researchers using this phylogenetic data to address a rapidly expanding set of questions in ecological and evolutionary processes. These methods include tasks such as ancestral state reconstruction (Butler and King, 2004;

*Corresponding author.
Email address: `cboettig@ucdavis.edu` (Carl Boettiger)

Paradis, 2004), diversification analysis (Fitzjohn, 2010; FitzJohn et al., 2009; Goldberg et al., 2011; Harmon et al., 2008; Paradis, 2004; Rabosky, 2006; Stadler, 2011a), quantifying the rate and tempo of trait evolution (Butler and King, 2004; Eastman et al., 2011; Harmon et al., 2008; Hipp and Escudero, 2010; Paradis, 2004; Revell et al., 2011), identifying evolutionary influences and proxies for community ecology Kembel et al. (2010); Webb et al. (2008), performing phyloclimatic modelling (Evans et al., 2009; Warren et al., 2008), and simulation of speciation and character evolution (Boettiger et al., 2012; Harmon et al., 2008; Stadler, 2011b), as well as various manipulation and visualization of phylogenetic data Jombart et al. (2010); Paradis (2004); Revell et al. (2011); Schliep (2010). 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. Several programs exist outside the R language for applied phylogenetic methods, incuding Java (Maddison and Maddison, 2011), MATLAB (Blomberg et al., 2003) and Python (Sukumaran and Holder, 2010) and online interfaces (Martins, 2004).

This paper describes the purpose and use of the `treebase` software package for the R language, which provides a programmatic link between these programs and TreeBASE. 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 (Morell, 1996; Sanderson et al., 1994). The database can be searched through an online interface which 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. Our `treebase` package uses this API to create a direct link between this data and the R language, making it easier for users and developers to take advantage of this data.

We provide three kinds of examples to illustrate what such programmatic and interactive access could mean for applied phylogenetics research. The first illustrates the process of exploration and discovery of available data to characterize the kind of data available in TreeBASE and illustrate its rapid increase in both the number and size of phylogenies provided. The second example focuses on reproducible research using a more detailed analysis of a single phylogeny, in which we replicate some results from an existing paper and extend the analysis to more recently developed methods. In the third example we take all of TreeBASE as our scope for a meta-analysis on diversification rates, illustrating full potential of interactive and programmatic access to the repository.

1.1. Basic Queries

The basic functions of the TreeBASE API allow search queries through two separate interfaces. The `OAI-MPH` interface provides the metadata associated with the publications from which the phylogenies have been taken, whiel the `Phylo-WS` interface provides information and access to the phylogenetic data itself. These interfaces are well-documented on the TreeBASE website. The `treebase` package allows these queries to be made directly from R, just as a user would make them from the browser. Because the queries can be implemented programmatically in R, a user can construct more complicated filters than permitted by the web interface, and can maintain a record of the queries they used to collect their data as an R script. The ability to script this data-gathering step of research can go a long way to reducing errors and ensuring that an analysis can be replicated later, by the author or other groups (Peng et al., 2011).

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

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

These functions loads the matching phylogenies into R, ready for analysis. The package documentation provides many examples of possible queries. The `search_treebase` function is the heart of the `treebase` package. While basic queries such as these seem simple, we present several use-cases of how this programmatic access can be leveraged to allow rapid exploration of phylogenetic data that opens doors to faster and easier verification of results and also new kinds of analysis and new scales of analysis.

To illustrate this potential, we introduce the second core function, `search_metadata`, which provides metadata about the resources available in the TreeBASE repository. Using programmatic access to this metadata and standard analysis tools available in R, we can quickly paint an up-to-the-minute picture of the data currently available TreeBASE. We will then return to our `search_treebase` function to illustrate several ways we can take advantage of the programmatic access to the data.

2. Quantifying TreeBASE

The `treebase` package provides access to the metadata of all publications containing trees deposited in TreeBASE using a separate API for metadata. This can help the user discover phylogenies of interest and also allows the user to perform statistical analyses on the data deposition itself, which could identify trends or biases in the phylogenetics literature.

The following command downloads the metadata for all publications associated with TreeBASE. (An optional argument can restrict searches to a given date range)

```
metadata <- search_metadata()
```

This returns an R list object, in which each element is an entry with bibliographic information corresponding to a published study that has deposited data in TreeBASE. From the length of this list we see that there are currently 3056 published studies in the database.

R provides a rich statistical environment in which we can extract and visualize the data we have just obtained. For instance, we may wish to obtain a list of all the dates of publication & names of the journals (publishers) that have submitted data:

```
dates <- sapply(metadata, `[[`, "date")
pub <- sapply(metadata, `[[`, "publisher")
```

which we organize into a table,

```
pub_table <- sort(table(as.character(pub)), decreasing=TRUE)
```

Many journals have only a few submissions, so we will group them together as "Other":

```
top_contributors <- names(head(pub_table,10))
pub[!(pub %in% top_contributors)] <- "Other"
```

We can then look at the distribution of publication years for phylogenies deposited in TreeBASE, color coding by publisher in Fig 1. It is encouraging to see that no single journal dominates the submissions, and taxa-specific publications and more broad-scope journals share the top ten spots. It will be interesting to watch these trends as more journals extend mandatory archiving requirements over the coming years.

In addition to this information about the publications, we can obtain metadata about the phylogenies themselves, including the study identifier number of where they were published, the number of taxa in the tree, a quality score, (if available), kind of tree (gene tree, species tree, or barcode tree) and whether the phylogeny represents a single or consensus type.

```
tree_metadata <- cache_treebase(only_metadata=TRUE)
```

For instance, we can summarize how the 1.0365×10^4 trees break out by kind or type:

```
kind <- table(sapply(tree_metadata, `[[`, "kind"))
xtable::xtable(kind) #pretty print
```

	V1
Barcode Tree	11
Gene Tree	290
Species Tree	10049

```
meta <- data.frame(publisher = as.character(pub), dates = dates)
require(ggplot2)
ggplot(meta) + geom_bar(aes(dates, fill = publisher))
```

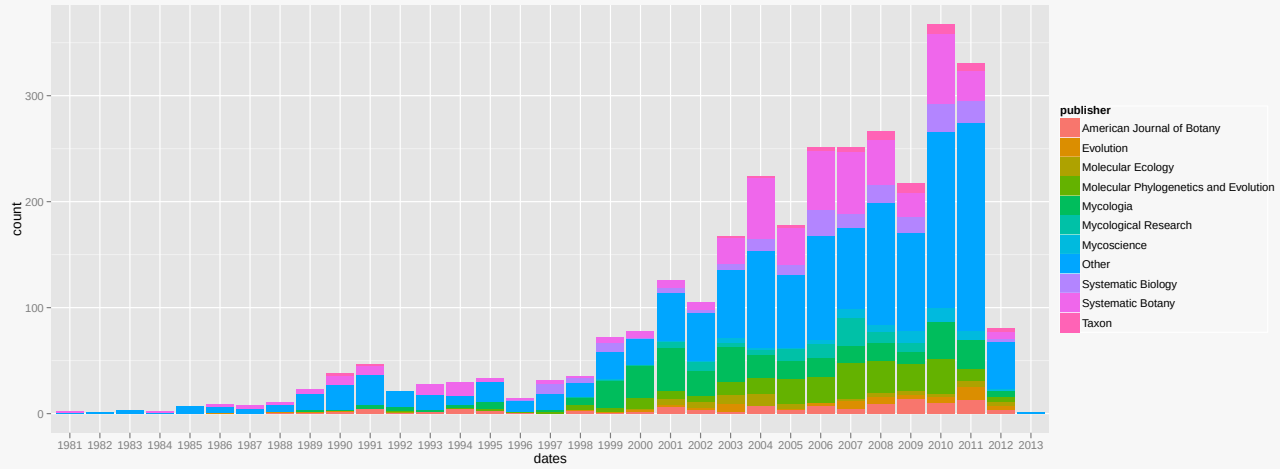

Figure 1: Histogram of publication dates by year, with the code required to generate the figure.

```
type <- table(sapply(tree_metadata, `[[`, "type"))
xtable::xtable(type)
```

	V1
Consensus	3011
Single	7354

```
quality <- table(sapply(tree_metadata, `[[`, "quality"))
xtable::xtable(quality)
```

	V1
Alternative Tree	76
Preferred Tree	270
Suboptimal Tree	15
Unrated	9989

It is possible to use the `only_metadata` option with `search_treebase` as well. While this information is always returned by the query, this is useful for a quick look at the data matching any search, such as

```
genetrees <- search_treebase( "'Gene Tree'", by='kind.tree', only_metadata=TRUE )
```

which returns just the metadata for the 296 gene trees in the database, from which we can look up the corresponding publication information, etc. For certain applications a user may wish to download all the available phylogenies from TreeBASE. Using the `cache_treebase` function allows a user to download a local copy of all trees. Because direct database dumps are not available, this function has intentional delays to avoid overtaxing the TreeBASE servers, and should be allowed a full day to run.

```
studyid <- sapply(tree_metadata,'[[', 'S.id')
sid <- sapply(metadata, '[[', 'identifier')
sid <- gsub(".*TB2:S(\\d*)", "\\1", sid)
matches <- sapply(sid, match, studyid)
Ntaxa <- sapply(matches, function(i)  tree_metadata[[i]]$ntax)
Ntaxa[sapply(Ntaxa, is.null)] <- NA
taxa <- data.frame(Ntaxa=as.numeric(unlist(Ntaxa)), meta)
ggplot(taxa, aes(dates, Ntaxa)) +
  geom_point(position = 'jitter', alpha = .8) +
  scale_y_log10() + stat_smooth(aes(group = 1))
```

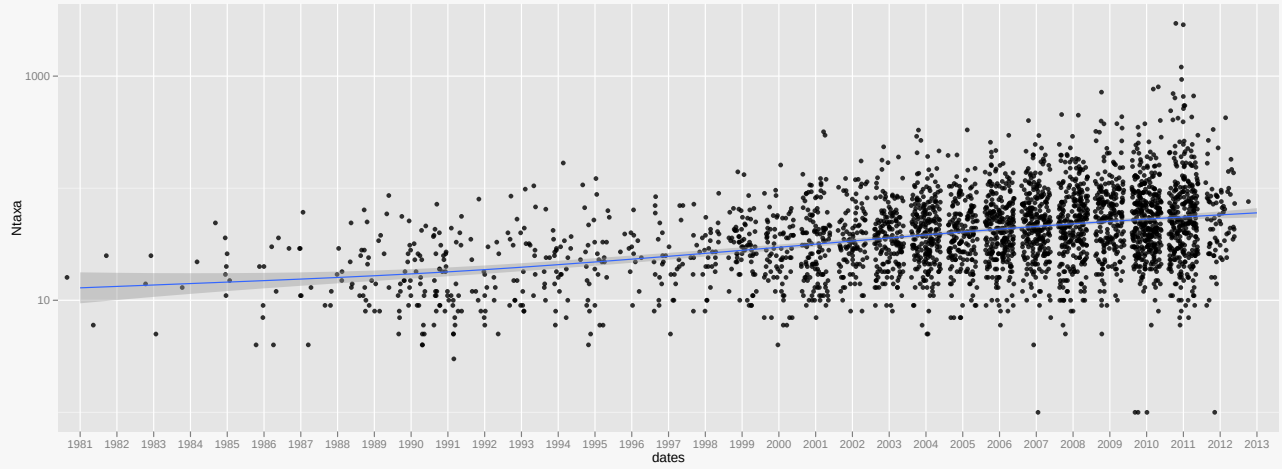

Figure 2: Combining the metadata available from publications and from phylogenies themselves, we can visualize the growth in taxa on published phylogenies. Note that the maximum size tree deposited each year is growing far faster than the average number.

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

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)
```

Having access to both the metadata from the studies and from the phylogenies in R lets us quickly combine these data sources in interesting ways. For instance, with a few commands we can visualize how the number of taxa on submitted phylogenies has increasing over time, Figure 2.

The promise of this exponential growth in the sizes of available phylogenies, with some trees representing 2957 taxa motivates the more and more ambitious inference methods being developed which require large trees to have adequate signal (Boettiger et al., 2012; FitzJohn et al., 2009). It will be interesting to see how long into the future this trend is maintained. These visualizations help identify research trends and can also help identify potential data sets for analyses. In this next section we highlight a few ways in which programmatic access can be leveraged for various research objectives.

3. Reproducible research

Reproducible research has become a topic of increasing concern in recent years (Gentleman and Temple Lang, 2004; Peng, 2011; Schwab et al., 2000). Access to data and executable scripts that reproduce the results presented are two central elements of this process which are addressed by the `treebase` package.

Imagine reading the recent phylogenetics paper, Derryberry et al. (2011). The paper analyzes speciation models on a phylogeny of bird taxa to identify if the rate of speciation shows a substantial shift among any of the groups, using the R package `laser` (Rabosky, 2006). We recall that more recent methods for identifying these rate shifts were presented in Stadler (2011a), and would like to see if the results presented hold up under the newer approach. The `treebase` package can help us both verify the results presented and test the data against the newer method with minimal effort. Further, because the process can be entirely scripted in R, from accessing the data to performing the analyses, it can be easily replicated & extended to additional datasets or methods.

Obtaining the tree

By drawing on the rich data manipulation tools available in R (and thus familiar to the large R phylogenetics community), the `treebase` package allows us to construct richer queries than are possible through TreeBASE alone. We begin our search by asking for a phylogenies by one of the paper's authors:

```
derryberry_results <- search_treebase("Derryberry", "author")
```

This shows several results. We would like the phylogeny appearing in *Evolution* in 2011. Each phylogeny includes a TreeBASE study id number, stored in the "S.id" element, which we use to look up the metadata for each paper.

```
ids <- lapply(derryberry_results, `[[`, "S.id")
meta <- lapply(ids, metadata)
```

We can then look through the metadata to find the study matching our description.

```
i <- which( sapply(meta, function(x) x$publisher == "Evolution" && x$date=="2011") )
derryberry <- derryberry_results[[i]]
```

This is simply one possible path to identify the correct study, certainly this query could be constructed in other ways, including direct access by the study identifier.

Having successfully imported the phylogenetic tree corresponding to this study, we can quickly replicate their analysis of which diversification process best fits the data. Different diversification models make different assumptions about the rate of speciation, extinction, and how these rates may be changing over time. The authors consider eight different models, implemented in the `laser` package (Rabosky, 2006). This code fits each of the eight models to that data:

```
require(laser)
tt <- branching.times(derryberry)
models <- list(              yule = pureBirth(tt),
                      birth_death = bd(tt),
                      yule.2.rate = yule2rate(tt),
      linear.diversity.dependent = DDL(tt),
 exponential.diversity.dependent = DDX(tt),
         varying.speciation_rate = fitSPVAR(tt),
         varying.extinction_rate = fitEXVAR(tt),
                    varying_both = fitBOTHVAR(tt)
               )
```

Each of the model estimate includes an AIC score indicating the goodness of fit, penalized by model complexity (lower scores indicate better fits) We ask R to tell us which model has the lowest AIC score,

```
aics <- sapply(models, `[[`, 'aic')
best_fit <- names(models[which.min(aics)])
```

and confirm the result presented in Derryberry et al. (2011), that the `yule.2.rate` model is the best fit to the data.

In this fast-moving field, new methods often become available within the time-frame that another manuscript is submitted by its authors and the time at which if first appears in print. For instance, the more sophisticated methods available in the more recent package, `TreePar`, introduced in Stadler (2011a) were not used in this study.

We load the new method and format the data as its manual instructs us

```
require(TreePar)
x<-sort(getx(derryberry), decreasing=TRUE)
```

The best-fit model in the laser analysis was a yule (net diversification rate) models with two separate rates. We can ask `TreePar` to see if a model with more rate shifts is favored over this single shift, a question that was not possible to address using the tools provided in `laser`. The previous analysis also considers a birth-death model that allowed speciation and extinction rates to be estimated separately, but did not allow for a shift in the rate of such a model. Here we consider models that have up to 4 different rates in Yule models,[1]

```
yule_models <- bd.shifts.optim(x, sampling = c(1,1,1,1), grid = 5, start = 0, end = 60, yule
= TRUE)[[2]]
```

We also want to compare the performance of models which allow up to four shifts and also estimate extinction and speciation separately:

```
birth_death_models <- bd.shifts.optim(x, sampling = c(1,1,1,1), grid = 5, start = 0, end =
60, yule = FALSE)[[2]]
```

The models output by these functions are ordered by increasing number of shifts. We can select the best-fitting model by AIC score,

```
yule_aic <- sapply(yule_models, function(pars) 2 * (length(pars) - 1) + 2 * pars[1] )
birth_death_aic <- sapply(birth_death_models, function(pars) 2 * (length(pars) - 1) + 2 *
pars[1] )
best_no_of_rates <- list(Yule = which.min(yule_aic), birth.death =
which.min(birth_death_aic))
best_model <- which.min(c(min(yule_aic), min(birth_death_aic)))
```

which confirms that the Yule 2-rate model is still the best choice based on AIC score. Of the eight models in this second analysis, only three were in the original set considered (Yule 1-rate and 2-rate, and birth-death without a shift), so we could by no means have been sure ahead of time that a birth death with a shift, or a Yule model with a greater number of shifts, would not have fitted better.

This kind of verification of results and validation against alternate methods will not occur regularly as long as the time required to do so is not negligible. While this kind of analysis already enjoys the benefits of scripted software implementations of the methods being employed, access to the actual data has become the rate-limiting step.

Leverging the programmatic access to the phylogenetic data shown in this example, we can both verify the original result and extend the analysis to the new method in a few minutes, without waiting for a drawn-out correspondence to access the data. No additional effort would be required even if many different phylogenies were involved. If verification and validation can be performed in minutes instead of days, they may become routine practices in settings such as journal clubs and peer-review.

[1] The syntax in `TreePar` is slightly cumbersome, the [[2]] indicates where this command happens to store the output models.

4. A self-updating meta-analysis?

Large scale comparative analyses that seek to characterize evolutionary patterns across many phylogenies increasingly common in phylogenetic methods (*e.g.* Davies et al., 2011; McPeek, 2008; McPeek and Brown, 2007; Phillimore and Price, 2008; Quental and Marshall, 2010). Often referred to by their authors as meta-analyses, these approaches have focused on re-analyzing 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 actually repeating the statistical analyses of those papers. 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 and Brown (2007), (emphasis added)

> One data set was based on 163 published species-level molecular phylogenies of arthropods, chordates, and mollusks. [...] A PDF format file of each article was obtained, and a digital snapshot of the figure was taken in Adobe Acrobat 7.0. This 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 appearance of digital tools that could now facilitate this analysis without mechanical calipers, (*e.g.* Laubach and von Haeseler, 2007), it is immensely easier and less error-prone to pull properly formatted phylogenies from the database for this purpose. In this section we describe how `treebase` can help overcome such barriers to discovery and integration at this large scale.

As the available data increases with subsequent publications, updating earlier meta-analyses can become increasingly tedious.

A central question in many studies that look across a large array of phylogenies has been to identify how often these trees show changing rates of speciation and extinction. Understanding these differences in diversification rates in different taxa is fundamental to explaining the patterns of diversity we see today. In this section we illustrate how we can perform a similar meta-analysis to the studies such as Davies et al. (*e.g.* 2011); McPeek (*e.g.* 2008); McPeek and Brown (*e.g.* 2007); Phillimore and Price (*e.g.* 2008); Quental and Marshall (*e.g.* 2010) across a much larger set of phylogenies and with just a few lines of R code. Because the entire analysis, including the access of the data, is scriptable, we could simply recompile this document some time in the future and see how the pattern we find has changed as more data has been added to TreeBASE.

Testing for constant speciation and extinction rates across all of treebase

A standard test of this is the γ statistic of Pybus and Harvey (2000) which tests the null hypothesis that the rates of speciation and extinction are constant. The γ statistic is normally distributed about 0 for a pure birth or birth-death process, values larger than 0 indicate that internal nodes are closer to the tip then expected, while values smaller than 0 indicate nodes farther from the tip then 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.

The `treebase` package provides a compressed cache of the phylogenies available in treebase. This cache can be automatically updated with the `cache_treebase` function,

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

which may require a day or so to complete, and will save a file in the working directory named with "treebase" and the date obtained. For convenience, we can load the cached copy distributed with the `treebase` package:

```
data(treebase)
```

We will only be able to use those phylogenies that include branch length data. We drop those that do not from the data set,

```
qplot(gammas)
```

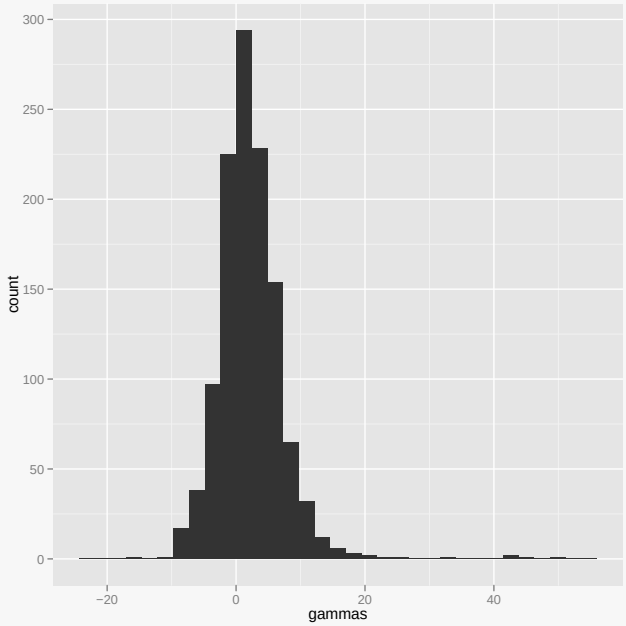

Figure 3: Distribution of the gamma statistic across phylogenies in TreeBASE. Strongly positive values are indicative of an increasing rate of evolution (excess of nodes near the tips), very negative values indicate an early burst of diversification (an excess of nodes near the root). Complete code for this anlysis is presented in the text, and the plotting command is shown above the figure.

```
      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 mutational steps). As most of these phylogenies are calibrated with branch length proportional to mutational step, we must time-calibrate each of them first.

```
timetree <- function(tree){
    try( chronoMPL(multi2di(tree)) )
}
tt <- drop_nontrees(sapply(branchlengths, timetree))
```

At this point we have 1217 time-calibrated phylogenies over which we will apply the diversification rate analysis. Computing the γ test statistic to identify devations 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 3.

As the γ statistic is normally distributed under the constant-rates model, we can ask what fraction of trees can reject this model at a given confidence level by calculating the associated p values,

```
p_values <- 2 * (1 - pnorm(abs(gammas)))
non_const <- sum(p_values < 0.025, na.rm=TRUE)/length(gammas)
```

wherein we find that 58% of the trees can reject the constant-rates model at the 95% confidence level. This supports a broad pattern from the above literature that finds deviations from the constant-rates models in smaller phylogentic samples.

```
lambdas <- sapply(tt, function(x) yule(x)$lambda)
ntaxa <- sapply(tt, Ntip)
dat <- data.frame(taxa=ntaxa,lambda=lambdas, gammas=gammas)
ggplot(dat, aes(taxa, lambda)) +
geom_point() + stat_smooth(method=lm, formula=y ~ x) +
scale_x_log10() + scale_y_log10()
```

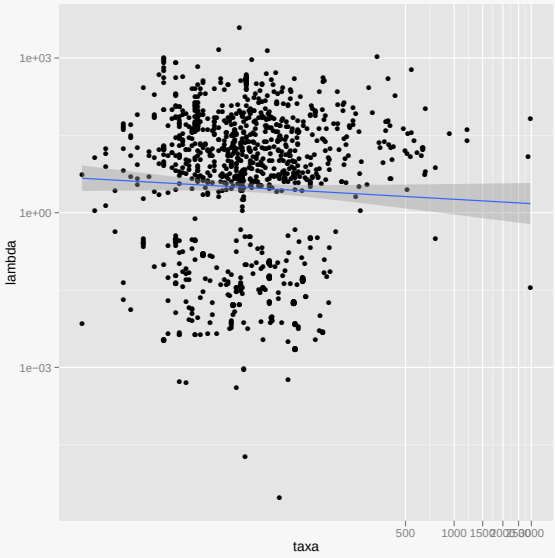

Figure 4: The species richness represented in a phylogeny shows no significant trend with increasing diversification rate.

Following McPeek and Brown (2007) we can investigate if the species richness of a given phylogeny correlates with diversification rate (Nee et al., 1994). Figure 4 shows this analysis, which supports the conclusion that species richness is not explained by increasing diversification rate.

```
xtable::xtable(summary(lm(log(lambda) ~ log(taxa), dat)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8239	0.4672	3.90	0.0001
log(taxa)	-0.1771	0.1144	-1.55	0.1220

Because `treebase` makes it possible to perform this analysis entirely by scripts using the latest treebase data, it is not only easier to perform this analysis but also to update it to reflect the latest data. Note that in this example it is not our objective to provide a thorough analysis of diversification patterns and their possible interpretations, as in McPeek (2008); McPeek and Brown (2007); Phillimore and Price (2008); Pybus and Harvey (2000), but merely to illustrate how the similar calculations to these can be easily applied across the much larger datasets in the repository. This example can be automatically updated to reflect the latest data in TreeBASE simply by rerunning the code we present above.

5. 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 a popular host for such data to support the data archiving requirements mentioned above. While programmatic access to the repository is possible through the `rdryad` package (Chamberlain et al., 2012), variation in data formatting must first be overcome. 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 et al., 2011). 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 phylognetics-based research (*e.g.* Fairbairn, 2010; Piwowar et al., 2011; Whitlock et al., 2010), is a particularly promising development that should continue to fuel 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 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.

6. Acknowledgements

CB wishes to thank the TreeBASE developer team for building and supporting the repository, and all contributers to TreeBASE. CB is supported by a Computational Sciences Graduate Fellowship from the Department of Energy under grant number DE-FG02-97ER25308.

Blomberg, S., Garland, J. T., Ives, A., 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution; international journal of organic evolution 57 (4), 717–745.
URL http://www3.interscience.wiley.com/journal/118867878/abstract

Boettiger, C., Coop, G., Ralph, P., Jan. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution, no–no.
URL http://doi.wiley.com/10.1111/j.1558-5646.2012.01574.x

Boettiger, C., Ralph, P., Coop, G., 2011. Is your phylogeny informative ? Measuring the power of comparative methods.

Butler, M. A., King, A. A., Dec. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. The American Naturalist 164 (6), 683–695.
URL http://www.jstor.org/stable/10.1086/426002

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

Davies, T. J., Allen, A. P., Borda-de Água, L., Regetz, J., Melián, C. J., Jul. 2011. NEUTRAL BIODIVERSITY THEORY CAN EXPLAIN THE IMBALANCE OF PHYLOGENETIC TREES BUT NOT THE TEMPO OF THEIR DIVERSIFICATION. Evolution; international journal of organic evolution 65 (7), 1841–1850.
URL http://doi.wiley.com/10.1111/j.1558-5646.2011.01265.xhttp://www.ncbi.nlm.nih.gov/pubmed/21729042

Derryberry, E. P., Claramunt, S., Derryberry, G., Chesser, R. T., Cracraft, J., Aleixo, A., Pérez-Emán, J., Remsen Jr, J. V., Brumfield, R. T., Jul. 2011. LINEAGE DIVERSIFICATION AND MORPHOLOGICAL EVOLUTION IN A LARGE-SCALE CONTINENTAL RADIATION: THE NEOTROPICAL OVENBIRDS AND WOODCREEPERS (AVES: FURNARIIDAE). Evolution; international journal of organic evolution, no–no.
URL http://doi.wiley.com/10.1111/j.1558-5646.2011.01374.x

Drummond, A. J., Rambaut, A., Jan. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7, 214.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247476&tool=pmcentrez&rendertype=abstract

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., Harmon, L. J., Jul. 2011. A NOVEL COMPARATIVE METHOD FOR IDENTIFYING SHIFTS IN THE RATE OF CHARACTER EVOLUTION ON TREES. Evolution 65 (12), 3578 – 3589.
URL http://doi.wiley.com/10.1111/j.1558-5646.2011.01401.x

Evans, M. E. K., Smith, S. a., Flynn, R. S., Donoghue, M. J., Feb. 2009. Climate, niche evolution, and diversification of the "bird-cage" evening primroses (Oenothera, sections Anogra and Kleinia). The American naturalist 173 (2), 225–40.
URL http://www.ncbi.nlm.nih.gov/pubmed/19072708

Fairbairn, D. J., Nov. 2010. the Advent of Mandatory Data Archiving. Evolution; international journal of organic evolution, no–no.
URL http://doi.wiley.com/10.1111/j.1558-5646.2010.01182.x

Fitzjohn, R. G., Sep. 2010. Quantitative Traits and Diversification. Systematic biology 59 (6), 619–633.
URL http://www.ncbi.nlm.nih.gov/pubmed/20884813

FitzJohn, R. G., Maddison, W. P., Otto, S. P., Dec. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Systematic biology 58 (6), 595–611.
URL http://www.ncbi.nlm.nih.gov/pubmed/20525612

Gentleman, R., Temple Lang, D., 2004. Statistical analyses and reproducible research. Bioconductor Project Working Papers, 2.
URL http://www.bepress.com/cgi/viewcontent.cgi?article=1001&context=bioconductor

Goldberg, E. E., Lancaster, L. T., Ree, R. H., May 2011. Phylogenetic Inference of Reciprocal Effects between Geographic Range Evolution and Diversification. Systematic biology 60 (4), 451–465.
URL http://www.ncbi.nlm.nih.gov/pubmed/21551125

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., Challenger, W., 2008. Geiger: investigating evolutionary radiations. Bioinformatics 24 (1), 129–131.

Hipp, A. L., Escudero, M., 2010. MATICCE: mapping transitions in continuous character evolution. Bioinformatics 26 (1), 132–3.
URL http://www.ncbi.nlm.nih.gov/pubmed/19880368

Huelsenbeck, J. P., Ronquist, F., Aug. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics (Oxford, England) 17 (8), 754–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/11524383

Jombart, T., Balloux, F., Dray, S., Aug. 2010. Adephylo: New Tools for Investigating the Phylogenetic Signal in Biological Traits. Bioinformatics (Oxford, England) 26 (15), 1907–9.
URL http://www.ncbi.nlm.nih.gov/pubmed/20525823

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., Webb, C. O., Jun. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics (Oxford, England) 26 (11), 1463–4.
URL http://www.ncbi.nlm.nih.gov/pubmed/20395285

Laubach, T., von Haeseler, A., Dec. 2007. TreeSnatcher: coding trees from images. Bioinformatics (Oxford, England) 23 (24), 3384–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/17893085

Maddison, W. P., Maddison, D., 2011. Mesquite: a modular system for evolutionary analysis.
URL http://mesquiteproject.org

Martins, E. P., 2004. COMPARE, version Computer programs for the statistical analysis of comparative data.
URL http://compare.bio.indiana.edu/

McPeek, M. a., Dec. 2008. The ecological dynamics of clade diversification and community assembly. The American naturalist 172 (6), E270–84.
URL http://www.ncbi.nlm.nih.gov/pubmed/18851684

McPeek, M. a., Brown, J. M., Apr. 2007. Clade age and not diversification rate explains species richness among animal taxa. The American naturalist 169 (4), E97–106.
URL http://www.ncbi.nlm.nih.gov/pubmed/17427118

Morell, V., 1996. TreeBASE: the roots of phylogeny. Science 273 (5275), 569.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.273.5275.569

Nee, S., May, R. M., Harvey, P. H., 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 344 (1309), 305–311.

Paradis, E., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20 (2), 289–290.
URL http://www.bioinformatics.oupjournals.org/cgi/doi/10.1093/bioinformatics/btg412

Parr, C. S., Guralnick, R., Cellinese, N., Page, R. D. M., Dec. 2011. Evolutionary informatics: unifying knowledge about the diversity of life. Trends in ecology & evolution 27 (2), 94–103.
URL http://www.ncbi.nlm.nih.gov/pubmed/22154516

Peng, C., Guiot, J., Wu, H., Jiang, H., Luo, Y., Mar. 2011. Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach. Ecology letters.
URL http://www.ncbi.nlm.nih.gov/pubmed/21366814

Peng, R. D., Dec. 2011. Reproducible Research in Computational Science. Science 334 (6060), 1226–1227.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.1213847

Phillimore, A. B., Price, T. D., Mar. 2008. Density-dependent cladogenesis in birds. PLoS biology 6 (3), e71.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2270327&tool=pmcentrez&rendertype=abstract

Piwowar, H. A., Vision, T. J., Whitlock, M. C., May 2011. Data archiving is a good investment. Nature 473 (7347), 285–285.
URL http://www.nature.com/doifinder/10.1038/473285a

Pybus, O. G., Harvey, P. H., Nov. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings. Biological sciences / The Royal Society 267 (1459), 2267–72.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1690817&tool=pmcentrez&rendertype=abstract

Quental, T. B., Marshall, C. R., Jun. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution, 1–8.
URL http://linkinghub.elsevier.com/retrieve/pii/S0169534710001011

R Development Core Team, T., 2012. R: A language and environment for statistical computing.
URL http://www.r-project.org/

Rabosky, D. L., Jan. 2006. LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evolutionary bioinformatics online 2, 273–6.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2674670&tool=pmcentrez&rendertype=abstract

Revell, L. J., Mahler, D. L., Peres-Neto, P. R., Redelings, B. D., Aug. 2011. a New Phylogenetic Method for Identifying Exceptional Phenotypic Diversification. Evolution; international journal of organic evolution (July), no–no.
URL http://doi.wiley.com/10.1111/j.1558-5646.2011.01435.x

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 (6), 183.

Schliep, K. P., Dec. 2010. phangorn: Phylogenetic analysis in R. Bioinformatics (Oxford, England) 27 (4), 592–593.
URL http://www.ncbi.nlm.nih.gov/pubmed/21169378

Schwab, M., Karrenbach, N., Claerbout, J., 2000. Making scientific computations reproducible. Computing in Science & Engineering 2 (6), 61–67.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=881708

Stadler, T., Mar. 2011a. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences 2011.
URL http://www.pnas.org/cgi/doi/10.1073/pnas.1016876108

Stadler, T., Apr. 2011b. Simulating Trees with a Fixed Number of Extant Species. Systematic biology.
URL http://www.ncbi.nlm.nih.gov/pubmed/21482552

Stamatakis, A., Nov. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics (Oxford, England) 22 (21), 2688–90.
URL http://www.ncbi.nlm.nih.gov/pubmed/16928733

Sukumaran, J., Holder, M. T., Apr. 2010. DendroPy: A Python Library for Phylogenetic Computing. Bioinformatics 26 (12), 1569–1571.
URL http://www.ncbi.nlm.nih.gov/pubmed/20421198

Warren, D. L., Glor, R. E., Turelli, M., Nov. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution; international journal of organic evolution 62 (11), 2868–83.
URL http://www.ncbi.nlm.nih.gov/pubmed/18752605

Webb, C. O., Ackerly, D. D., Kembel, S. W., Sep. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics (Oxford, England) 24 (18), 2098–100.
URL http://www.ncbi.nlm.nih.gov/pubmed/18678590

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., Moore, A. J., Mar. 2010. Data archiving. The American Naturalist 175 (2), 145–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/20073990