Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

R programming in phylogenetics and evolution

Christoph Heibl

August 2014

Gutachter

1. Gutachterin: Prof. Dr. Susanne S. Renner

2. Gutachter: Prof. Dr. Dirk Metzler

Tag der Einreichung: 13. August 2014

Tag der mündlichen Prüfung: 11. December 2014

Meinen Eltern
Meiner Susi
Meinem Sohn Valentin

Preface

Statutory Declaration

Erklärung

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dr. Susanne S. Renner betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt wurde.

Christoph Heibl

Miesbach, 13. August 2014

Note

In this thesis, I present the results from my doctoral research, carried out in Munich from January 2006 to December 2009 under the guidance of Prof. Dr. Susanne S. Renner. My thesis resulted in three publications. I also gave the presentations and developed the software listed below. For the study of cucurbit historical biogeography, I estimated divergence times including sensitivity analyses on the impact of different sets of fossil time constraints. The study on niche evolution in *Hordeum* was done in collaboration with Dr. Sabine Jakob and Prof. Dr. Frank Blattner (Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany) and Dr. Dennis Rödder (Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany) and my contribution was the estimation of climatic ancestral tolerances and the evolution of disparity in climatic niche dimensions. For the study on *Oxalis*, I generated all data and conducted all analyses myself. Writing and discussion involved collaboration with Prof. Susanne Renner.

List of publications

Peer-reviewed journal articles

Heibl, C. & S.S. Renner. 2012. Distribution models and a dated phylogeny for Chilean *Oxalis* species reveal occupation of new habitats by different lineages, not rapid adaptive radiation. *Systematic Biology* **61**: 823–834.

Jakob, S., C. Heibl, D. Rödder & F. Blattner. 2010. Population demography influences climatic niche evolution: evidence from South American *Hordeum* species (Poaceae). *Molecular Ecology.* **19**: 1423–1438.

Schaefer, H., C. Heibl, & S.S. Renner. 2009. Gourds afloat: A dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc. Roy. Soc. B.* **276**: 843–851.

Posters

Heibl, C. & S.S. Renner 2007. Growth form evolution in Mediterranean *Oxalis* - Maximum likelihood estimates of ancestral states. *Origin and evolution of biota in Mediterranean climate zones.* University of Zurich, Switzerland, July 14–15, 2007.

Heibl, C. & S.S. Renner. 2009. *Reconstruction of past climatic niches and growth form evolution in south-central Andean Oxalis. Niche evolution - a unifying concept for systematics, ecology, paleontology, and conservation biology.* University of Zurich, Switzerland, July 3–4, 2009.

Oral presentations

Evolution of *Oxalis* in arid southern South America under changing climates. *Institutskolloquium Biologie und Umweltwissenschaften.* Carl von Ossietzky University, Oldenburg, Germany, November 12, 2012.

Evolutionary history of *Oxalis* section *Giganteae* in Chile. *Seminars on Evolution and Systematics of Plants and Fungi.* Ludwig-Maximilians University, Munich, Germany, December 3, 2008.

Evolution der Gattung *Oxalis* im südwestlichen Südamerika. *Seminars on Evolution and Systematics of Plants and Fungi.* Ludwig-Maximilians University, Munich, Germany, December 20, 2006.

Rain shadow and fog desert: Orogenic influences on the evolution of *Oxalis* in the coastal desert of Atacama, Chile. *Annual meeting of the Gesellschaft für Systematische Biologie.* Museum for Natural History, Basel, Switzerland, September 15, 2005.

Herbaria visited

- Botanische Staatssammlung München (M), Munich, Germany, 2006–2010
- Museo Nacional de Historia Natural (SGO), Santiago, Chile, October 2005, November/October 2006
- Universidad de Concepción (CONC), Concepción, Chile, December 2004, October 2005, November 2006, November 2007, January/February 2008
- Universidad de La Serena (ULS), La Serena, Chile, October 2006
- Universidad de Magallanes (HIP), Punta Arenas, Chile, November 2007
- Universidad Nacional Mayor de San Marcos (USM), Lima, Peru, November 2007

- Royal Botanic Gardens (K), London, United Kingdom, April 2010
- Royal Botanic Garden Edinburgh (E), Edinburgh, United Kingdom, April 2010

Fieldwork

- Atacama Desert and Altiplano, Chile, October 27–December 15, 2004
- Atacama Desert and Altiplano, Chile, October 2–25, 2005
- Atacama Desert, Chile, September 22–October 27, 2006
- Atacama Desert, Chile, October 18–November 3, 2007
- Última Esperanza and Tierra del Fuego, Chile, November 5–11, 2007
- Coastal Peruvian Desert, Peru, November 20–26, 2007
- Southern to Central Andes, Chile, January 23–March 7, 2008

Funding

- Fieldwork in Chile 2005 by the "Freunde des Botanischen Gartens e.V.", München
- The three-year research period (2007–2009) was funded by the German Research Council (DFG project RE 603/5-1)
- Visits to the herbaria of the Royal Botanic Gardens Edinburgh and Kew by the "Freunde des Botanischen Gartens e.V.", München
- Travel to conferences in Zurich 2007 and 2009 by EES$^{\mathrm{LMU}}$ Travel Grants

Contents

Summary

This dissertation addresses the application of the statistical computing language R in the study of evolution and diversification of plants. The topics included range from the worldwide historical biogeography of the cucurbit family and the phylogenetic composition of the Mediterranean *Oxalis* flora in central Chile to the interplay between population genetics and climatic niche evolution in four *Hordeum* clades in the Americas. In these studies, I drew on existing methods in R and on java and C programs that could be easily integrated with R. Whenever necessary, I created additional software available in four new R packages. R's features, e.g., intersystem-interfaces, extensibility, reproducibility and advanced graphical capability, proved well suited for evolutionary and phylogenetic research.

My coauthors and I addressed the history of Cucurbitaceae, one of the most economically important families of plants, using a multi-gene phylogeny for 114 of the 115 genera and 25 per cent of the 960 species. Worldwide sampling was achieved by using specimens from 30 herbaria. Results reveal an Asian origin of Cucurbitaceae in the Late Cretaceous, followed by the repeated spread of lineages into the African, American and Australian continents via transoceanic long-distance dispersal (LDD). North American cucurbits stem from at least seven range expansions of Central and South American lineages; Madagascar was colonized 13 times, always from Africa; Australia was reached 12 times, apparently always from Southeast Asia. Overall, Cucurbitaceae underwent at least 43 successful LDD events over the past 60 Myr, which would translate into an average of seven LDDs every 10 Myr. These and similar findings from other angiosperms stress the need for an increased tapping of museum collections to achieve extensive

geographical sampling in plant phylogenetics.

The second study focused on the interplay of population demography with the evolution of ecological niches during or after speciation in *Hordeum.* While large populations maintain a high level of standing genetic diversity, gene flow and recombination buffers against fast alterations in ecological adaptation. Small populations harbor lower allele diversity but can more easily shift to new niches if they initially survive under changed conditions. Thus, large populations should be more conservative regarding niche changes in comparison to small populations. My coauthors and I used environmental niche modeling together with phylogenetic, phylogeographic and population genetic analyses to infer the correlation of population demography with changes in ecological niche dimensions in 12 diploid *Hordeum* species from the New World, forming four monophyletic groups. Our analyses found both shifts and conservatism in certain niche dimensions within and among clades. Speciation due to vicariance resulted in three species with no pronounced climate niche differences, while species originating due to long-distance dispersals or otherwise encountering genetic bottlenecks mostly revealed climate niche shifts. Niche convergence among clades indicates a niche-filling pattern during the last 2 Myr in South American *Hordeum.* We provide evidence that species that did not encounter population reductions mainly show ecoclimatic niche conservatism, while major niche shifts have occurred in species that have undergone population bottlenecks. Our analyses allow the conclusion that population demography influences adaptation and niche shifts or conservatism in South American *Hordeum* species.

Finally, I studied the phylogenetic composition of *Oxalis* flora of Mediterranean zone of Chile by asking whether in such a species-rich clade xerophytic adaptations arose in parallel, at different times, or simultaneously. Answering this type of question has been a major concern of evolutionary biology over the past few years, with a growing consensus that lineages tend to be conservative in their vegetative traits and niche requirements. Combined nuclear and chloroplast DNA sequences for 112 species of Oxalidales (4900 aligned nucleotides) were used for a fossil-calibrated phylogeny that includes 43 of the 54 species of Chilean *Oxalis*, and species distribution models (SDMs) incorporating precipitation, temperature, and fog, and the phylogeny were used to reconstruct ancestral habitat preferences, relying on likelihood and Bayesian techniques. Since uneven collecting can reduce

the power of SDMs, we compared 3 strategies to correct for collecting effort. Unexpectedly, the *Oxalis* flora of Chile consists of 7 distant lineages that originated at different times prior to the last Andean uplift pulse; some had features preadapting them to seasonally arid or xeric conditions. Models that incorporated fog and a 'collecting activity surface' performed best and identified the Mediterranean zone as a hotspot of *Oxalis* species as well as lineage diversity because it harbors a mix of ancient and young groups, including insufficiently arid-adapted species. There is no evidence of rapid adaptive radiation.

General Introduction

Data analysis and computation

A seemingly small invention in 1947 marks the beginning of the Digital Revolution: that of the transistor. The massive and ongoing increase of computer power since the 1970s was the main driver of this transformation process (foreseen by Moore, 1965), which also has affected life sciences and environmental sciences. Wet laboratory techniques, such as the polymerase chain reaction (PCR) and Sanger sequencing, have led to the development of molecular phylogenetics and phylogeography, branches of evolutionary biology aiming to reconstruct the Tree of Life, or parts of it, with the estimation of gene genealogies. Similarly, since the launch of Sputnik 1 by the Soviet Union in 1957, satellite-based technologies, such as global positioning and remote sensing, have become standard tools for gathering spatial information about single individuals or whole communities. Common to both sets of techniques is the massive accumulation of data, posing a challenge for their curation and analysis.

Although it is commonly stated that increasing hardware performance has improved our ability to analyze large datasets, a significant fraction of the increase in computer speed is due to improved efficiency of algorithms (Schuster, 2013). It follows that it is the implementation, distribution, and application of such algorithms that helps us meet the challenge of huge amounts of information, a fact highlighted by even the smallest next-generation sequencing projects (McCormack *et al.*, 2013).

Nearly half a century after the invention of the transistor, Gentleman (2004), among others, raised the concern that methods for data analysis will not be produced, distributed, and applied to the degree required by the increasing amounts of data. He proposed three measures to foment the production and application of algorithms for data analysis. First, software environments should be developed that are both easy to use and flexible to extend, being attractive for both users and developers. Having users and developers work in the same software environment would stimulate the distribution and improvement of methods because of bidirectional flow of information: Users benefit from the latest code that developers provide and give feedback that helps developers to debug and generalize the code. Second, publication of results should follow the principles of reproducible research (Gentleman & Temple Lang, 2004; Fomel & Claerbout, 2009; Xie, 2013), i.e., data and source code should be included in the publication to enable the reader to reproduce the results and test their sensitivity to different sets of assumptions or parameterizations. This would not only enhance the trustfulness of the results, but more importantly, stimulate the use and iterative refinement of methods, an important principle in the scientific method. Iterative refinement requires methods not to be programmed every time *de novo*, but instead that the present implementation of the method be scrutinized and extended. This leads naturally to Gentleman's last postulate: The ideal software environment should be capable of inter-system interfacing, i.e, data exchange with a large number of programming languages and database management systems (DBMS) including GIS and different file formats to make efficient use of data sets and methods already available in other systems.

Now, ten years later, the accumulation of data is flourishing, as are statistical applications. In what follows I will review concepts in computer programming that lend themselves to the task of data analysis and demonstrate the suitability of the R software environment for phylogenetic analyses. I will use my empirical studies to illustrate this.

The R language for data analysis

A powerful programming language for statistical data analysis was the major goal when Ihaka & Gentleman (1996) created the R language. In the light of the ideas introduced above, this meant that the new programming language was to be intuitive to use and easy to extend and interface, and that its computations were to be trustworthy and reproducible. To achieve this goal, R's developers relied on several programming concepts that had not been combined hitherto: interpretation, functional programming, and object-oriented programming.

As an interpreted language, R's source code, i.e., a list of sequential commands, is parsed line-by-line and evaluated directly. This is in contrast to compiled languages where the source code has to be compiled into machine code prior to the execution of the program. The machine code is then executed directly by the central processing unit of a computer, which is much faster and more efficient than interpreting code at runtime. Nevertheless, interpreted languages confer some essential advantages if our goal is data analysis and statistical modeling, in particular in view of the principles of reproducible research (Gentleman, 2004). Because programs (or 'scripts') are evaluated line by line, R programming is easy to learn, easy to debug, easy to share and easy to reproduce.

R is also a functional programming language, i.e., any computation is treated as the evaluation of mathematical functions and, as a consequence, programs in R consist entirely of functions (Chambers, 2008). Like mathematical functions, functions in R depend only on their input, which is given as arguments, and not on the program state. This behavior, called referential transparency (Søndergaard & Sestoft, 1990), makes it much easier to understand and predict the behavior of a program. More important from a user's perspective, however, might be that functional programming leads to a syntax that is similar to mathematical notation and is thus well suited for statistical modeling (Ihaka & Gentleman, 1996).

Another concept that has influenced R language is object-oriented programming (Chambers, 1998). In object-oriented programming, entities of interest (e.g., molecular sequences, phylogenetic trees) are represented as objects. Objects have data fields called attributes that describe the object and associated functions, called methods. The exact definition of attributes and methods are tied to a class definition. Any object of a particular class is called an instance of that class: its structure and behavior is completely dependent on the class definition. The encapsulation of attributes and methods in the same object adds to the consistency of in object-orientated programming and created a system of modular objects that are easy to maintain and to extend. Extensibility is further enhanced by the concept of inheritance: New classes can build upon preexisting classes, i.e., they inherit their attributes and methods, supplemented with additional attributes and methods. R is not an object-orientated language in the strict sense. In its old S3 class system neither attributes nor methods of objects were formally defined. Instead, S3 objects could have classes as attributes and inheritance was emulated by using vectors of classes. Generic functions (such as `print`, `plot`, `summary`) recognized class attributes and dispatched the object to the corresponding function. Improvement came from the formally defined S4 class system designed by Chambers (1998) and implemented in the **methods** package, which has been part of the R distribution since version 1.7.0.

Evolutionary data analysis

Evolutionary data analysis became available when Paradis *et al.* (2004) presented **ape**, an R package for analysis of phylogenetics and evolution. This package was seminal in many ways, but its perhaps most important achievement was the definition of the `phylo` class. This class derived its simple structure from graph theory and allowed objects in R to represent phylogenetic trees, the pivotal data structure in macroevolutionary analysis. Supplemented with functions for import, export, plotting, and manipulation of tree topology and branch length, the `phylo` class in **ape** triggered a massive development of methods and packages (see the Taskview Phylogenetics[a] maintained by Brian O'Meara for an updated overview).

The evolutionary data analyses presented in this thesis make extensive use of R in general and of **ape** and its companion packages in particular. At the same time, I developed many additional functions and interfaces to complement the existing software. Below, I introduce three examples of macroevolutionary data analyses, one from each publications of this dissertation, chosen to indicate the challenges that I encountered and their solution in R.

Divergence times

Schaefer *et al.* (2009) explore the biogeographic history of the cucumber family

[a] http://cran.r-project.org/web/views/Phylogenetics.html

(Cucurbitaceae) based on molecular sequences from five chloroplast loci covering 114 of 115 extant cucurbit genera and five independent time calibration points. Techniques to estimate ancestral ranges, whether explicitly model-based (Ree *et al.*, 2005; Ree & Smith, 2008; Lemey *et al.*, 2009) or not (Ronquist, 1997), rely on a time-calibrated, phylogenetic tree, today often called time-tree or chronogram. The basic problem in the estimation of time-trees is that molecular divergence has two components, molecular rate and time, which can only be disentangled by making assumptions about the evolution of molecular rates, so-called molecular clock models. Two types of molecular clocks are possible: the 'strict clock', which does not allow for any evolution of rates, meaning that the molecular rate found at the root remains constant throughout the tree (Zuckerkandl & Pauling, 1965), and 'relaxed clocks', which allow molecular rates to change according to some model (e.g., Thorne *et al.*, 1998; Sanderson, 2002; Drummond *et al.*, 2006; Drummond & Suchard, 2010. Relaxed clock models are well suited for biogeographic analysis of larger clades with several time calibration points (Renner, 2005), because with the increasing number of lineages, the assumption of rate constancy is likely to be violated, and multiple calibrations points can rarely be reconciled with a single molecular rate. Because the size of Cucurbitaceae (∼960 species) and the availability of at least five time calibration points, we chose a relaxed clock method with auto-correlated molecular rates implemented in the multidivtime approach (Thorne *et al.*, 1998) with multi-locus data (Thorne & Kishino, 2002) for divergence time estimation based on five chloroplast loci under varying selective constraints (*rbc*L and *mat*K genes, the *trn*L intron and the *trn*L-F and *rpl*20-*rps*12 intergenic spacers).

The multidivtime approach connects five C language programs in a pipeline: (1) BASEML estimates the parameters of a F84 model of sequence evolution given a set of DNA sequences and a corresponding tree topology; (2) PAML2MODELINF parses the output file produced by BASEML and writes the parameters of interest to an input file for the following program; (3) ESTBRANCHES estimates branch lengths and their variance based on the tree topology and parameters estimated by BASEML; (4) INSEED generates random number seeds for setting up a Markov chain Monte Carlo (MCMC) procedure; (5) MULTIDIVTIME uses a MCMC to estimate the posterior distribution of divergence times based on the branch length parameters estimated by ESTBRANCHES. BASEML is part of the PAML package

(Yang, 2007) and the remaining programs have been developed by Thorne *et al.* (1998). Neither of the programs is particularly user-friendly. They are run from command-line and upon violation of their input format they can get stuck or crash without any diagnostic message. This behavior and the fact that the pipeline has to be interrupted several times to manually adjust control files led me to encapsulate the whole pipeline in R language to check the input data for consistency, produce syntactically correct input and control files, automatically set priors and time constraints, and control the workflow in the case of multi-locus data.

Ancestral climatic tolerances

Jakob *et al.* (2010) focuse on the influence of demographic history on niche evolution versus niche conservatism, which is the tendency of species to retain ancestral ecological tolerances (Harvey & Pagel, 1991). After Peterson *et al.* (1999) had found niches of pairs of sister species of mammals, birds, and butterflies to be conserved over a timescale of several million years, many subsequent studies (reviewed in Wiens & Graham, 2005; Wiens *et al.*, 2010) set out to collect empirical evidence for niche conservatism in diverse groups. Perhaps unsurprisingly, they found different patterns in different clades, causing Wiens *et al.* (2010) to conclude that instead of simply asking whether a species' niche is conserved or not, studies should explicitly focus on the causes and consequences of niche conservatism. Four population-genetic conditions might impede or delay the adaptation of populations/species to divergent niches (Wiens & Graham, 2005; Wiens *et al.*, 2010): (1) A population simply might lack genetic variation in niche-relevant traits that natural selection could act upon. If genetic variation in niche-relevant traits is present, niche evolution might nevertheless be deterred by (2) stabilizing selection in a stable environment, by (3) gene flow in a meta-population swamping out adaptive alleles in a subpopulation, or by (4) pleiotropic effects, with beneficial mutations in niche-relevant traits being linked to another genetic locus, incurring fitness disadvantage.

The 12 species of the South American clade of barley (*Hordeum*, Poaceae) have evolved during the Pleistocene (crown age ∼2 Myr; Blattner, 2006; Jakob *et al.*, 2010). Repeated climate fluctuations between glacial and interglacial periods are likely to have influenced the species' demography. In this scenario changes in climate and population size constantly change the relative importance of selection, drift, and gene flow during the course of lineage evolution, sometimes leading to niche conservatism and sometimes not.

To investigate the influence of demography on niche conservatism in South American *Hordeum*, we combined population genetic methods (see Jakob *et al.* (2010) for details) with a phyloclimatic modeling approach (Yesson & Culham, 2006) devised by Evans *et al.* (2009). Based on a time-tree and environmental niche models (ENM), their approach consists of four steps: (1) creation of predicted niche occupancy (PNO) profiles for each variable in the ENMs; (2) pairwise niche comparisons of whole ENMs or single PNOs using randomization tests (Warren *et al.*, 2008); (3) calculation of (niche) disparity-through-time (DTT) (Harmon *et al.*, 2003); and (4) estimation of ancestral climatic tolerances *sensu* Evans *et al.* (2009). Steps (2) and (3) relied on available software, in particular the JAVA program ENMTools (Warren *et al.*, 2008) and R package **geiger** (Harmon *et al.*, 2003), respectively, and will not be further discussed. For steps (1) and (4), I developed a set of R functions to prepare, analyze and display the data.

Species richness, lineage richness

Heibl & Renner (2012) use Chilean species of *Oxalis* (Oxalidaceae), wood sorrels and allies, to investigate the historical accumulation of biological diversity in the Mediterranean-type climate (MTC) zone of Chile. Mediterranean-type climates are equitable (Cowling *et al.*, 1996) with a low annual amplitude of temperature cycling and little (if any) occurrence of frost. The greatest seasonality is found in the distribution of rainfall, which peaks in winter with a distinctive dry season in summer. In contrast, biomes adjacent to the MTC zone in Chile are considerably harsher with unpredictable precipitation or low winter temperature shortening the growing season (Luebert & Pliscoff, 2006). But these non-MTC biomes are not only less equitable, they have also been less stable during the past, having been exposed to periods of drought (Thompson *et al.*, 1998; Ammann *et al.*, 2001; Stott *et al.*, 2002) and glaciation (Hulton *et al.*, 2002; Kaplan *et al.*, 2008) during the Pleistocene. Whereas Pleistocene climate cycles may have stimulated speciation in the adjacent areas, my hypothesis was that the MTC region might have acted as a longterm sink for members of many lineages and might thus have been accumulating species richness over time.

To test if the biogeographic history of *Oxalis* fits the sink hypothesis of MTC diversity, we inferred the spatial distribution of 'lineage richness', the number of lineages endemic to the MTC region in Chile and adjacent areas (Atacama Desert, Andes range, Patagonian forest and steppes), expecting the highest values in the MTC zone. The mapping of lineage richness involved the following steps: (1) presence records for each species of *Oxalis* were combined with climatic predictor variables to fit environmental niche models (ENMs) mapped onto geographical space as a continuous surface of environmental suitability (Phillips *et al.*, 2006; Phillips & Dudík, 2008; Phillips *et al.*, 2009). (2) geographic ranges of each species were approximated by converting its suitability surfaces by means of a threshold (Jiménez-Valverde & Lobo, 2007; Liu *et al.*, 2013) into binary (presence-absence) range maps (Graham & Hijmans, 2006); (3) we obtained the geographic range of each lineage as the union of the geographic ranges of clade members; (4) finally, we calculated the lineage range richness by summation of all lineage range maps.

Results

Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events

Hanno Schaefer, Christoph Heibl, & Susanne S. Renner. 2009. *Proceedings of the Royal Society B: Biological Sciences* **276**: 843–851.

http://dx.doi.org/10.1098/rspb.2008.1447

Population demography influences climatic niche evolution: evidence from diploid American *Hordeum* species (Poaceae)

Sabine S. Jakob, Christoph Heibl, Dennis Rödder & Frank R. Blattner. 2010. *Molecular Ecology* **19**: 1423–1438.

http://dx.doi.org/10.1111/j.1365-294X.2010.04582.x

Distribution models and a dated phylogeny for Chilean *Oxalis* species reveal occupation of new habitats by different lineages, not rapid adaptive radiation

Christoph Heibl & Susanne S. Renner. 2012. *Systematic Biology* **61**: 823–834.

http://dx.doi.org/10.1093/sysbio/sys034

General Discussion

Analyses of evolution with R

The work presented in this doctoral thesis required, and relied on, statistical modeling in R language. To evaluate R's suitability for evolutionary analyses, I now return to the three examples presented in the General Introduction: divergence times (Schaefer *et al.*, 2009), ancestral climatic tolerances (Jakob *et al.*, 2010), and lineage richness (Heibl & Renner, 2012). In each case, I combined existing software with R's programmability when necessary. The programs (or functions) I developed are mostly included in R packages that are available to the scientific community[a] and guarantee the analyses' reproducibility. Each of the three topic's implementation strategies will be discussed in the remainder of this section, before the following sections derive common features of evolutionary analyses in R and identify achievements and future challenges.

Divergence times

For the divergence time estimation in the Cucurbitaceae (Schaefer *et al.*, 2009), I implemented the multidivtime approach (Thorne *et al.*, 1998; Kishino *et al.*, 2001; Thorne & Kishino, 2002) in the R package **LAGOPUS**[b], with contributions by Natalie Cusimano, who provided functions for parsing and displaying the output. The programs that constitute the multidivtime 'pipeline' (BASEML,

[a] http://www.christophheibl.de/Rpackages.html

[b] http://www.christophheibl.de/mdt/mdtinr.html

paml2modelinf, estbranches, inseed, multidivtime) are written in C language and even their most efficient reimplementation in R would not would not achieve the same computation rate than the C version. In addition, every new implementation bears the danger of introducing errors. As a consequence, I chose to use the original programs, but to connect them in a data handling pipeline to increase the ease of use and repeatability. R proved suitable for this task because the inter-system interfaces needed were already available: writing and reading sequence data and phylogenetic trees (packages **ape**, **ips**) and calling external programs via the `system` command (**base** package). Other available functions make it easy to check the input data for possible errors: Are sequences aligned? Is the tree rooted? Does the tree contain branch lengths? I designed S4 classes to store the input data and program settings ("control files") to secure data consistency all along the pipeline from input to output. Because the setting of rate and age priors could be automated, the whole procedure could be wrapped in one single function (`multidivtime`), which we used successfully in batch scripts exploring different parameter settings on a computer cluster. The size of the data set (146 species, 6 markers) exceeded what could be processed on a desktop computer at the time of analysis (2009), the implementation of the multidivtime approach in R was a crucial step in the study of cucurbit historical biogeography.

Ancestral climatic tolerances

The environmental niche modeling approach in Jakob *et al.* (2010) makes use of a different strength of R: extensibility. Comparative analysis methods in R are particularly well developed (Paradis *et al.*, 2004; Harmon *et al.*, 2008; Kembel *et al.*, 2010; Revell, 2012; Orme *et al.*, 2013). Based on these tools, the estimation of ancestral climatic tolerances (Evans *et al.*, 2009) in *Hordeum* was straightforward to implement. It is a non-parametric approach on top of the parametric estimation of ancestral traits (Schluter *et al.*, 1997). Existing parametric methods for continuous traits (such as climatic variables) whether based on random (Felsenstein, 1985) or directional/stabilizing evolution (Martins & Hansen, 1997) estimate the mean and standard deviation of a single metric, i.e, optima in niche space, but cannot estimate ranges or "tolerances". The approach of Evans *et al.* (2009) circumvents this problem of estimating ancestral trait values by sampling the predicted niche occupancy (PNO) profile, i.e., the ENM's response curve in a single dimension of environmental space, a sufficient number of times (e.g. 1000). The result is a distribution of ancestral trait estimates that approximates the ancestral PNO. I implemented this approach in the function `anc.clim` (package **phyloclim**), which is basically a wrapper extending the `ace` function from the **ape** package by the resampling strategy described above and, in addition, by the option of iterating the whole procedure over a sample of phylogenetic trees drawn from a posterior distribution of a Bayesian analysis. Given n samples from the PNO and m samples of trees, `anc.clim` uses $n \times m$ calls to `ace` to estimate ancestral climatic tolerances, or niche width, taking phylogenetic uncertainty into account. Although computationally costly, the programmability of R allows non-parametric approaches based on randomization where parametric methods are not yet available. A similar example are the tests of niche identity and niche similarity (Warren *et al.*, 2008), first implemented in PERL (ENM-tools) and then transferred to R (**phyloclim**, this thesis; **dismo**, Hijmans *et al.*, 2013).

Species richness, lineage richness

Models of species richness and lineage richness for *Oxalis* of the west-Andean southern cone of South America (Heibl & Renner, 2012) developed in R show both the power of using R as a geographic information system (GIS) and its limitations. In principle, tools exist that allow to conduct the whole modeling process in pure R (packages **sp**, **raster**, **dismo**). This integration is desirable considering the complexity of the data: occurrence data of 42 species, 21 environmental variables, 4 background sampling schemes and numerous types of types of response curves combine to a plethora of models that can be explored. In particular, the **raster** package with its S4 classes `rasterStack` and `rasterBrick` (arrays of raster layers with identical origin, extent, and resolution, that can be processed with array programming techniques) and the ability to analyze large raster datasets (by keeping only parts of the data in memory) is ideally suited for species distribution modeling (SDM). Nevertheless, the extent ($x \times y$) and resolution (30 degree seconds) of the study area in this case posed a considerable computational burden, which I avoided by integration of R and GRASS GIS (GRASS Development Team, 2009). GRASS GIS supports raster and vector data models and consists of several sets of modules that are programmed mostly in C. As a result, GRASS outperforms R in terms of efficiency but is much less flexible despite its modular architecture, which allows extensions of either compiled or interpreted code (Neteler & Mitasova, 2008). My modeling approach took advantage of both systems using rapid GIS operations in GRASS GIS and data handling and management in R. A convenient and powerful interface is provided with the package **spgrass6**, which allows to start R during a GRASS GIS session and import and export data from GRASS GIS to R's current environment and vice versa. This means data can easily be handed back and forth between the two systems, thereby being processed in the system of choice for a specific task. The SDMs themselves were calibrated and predicted with MaxEnt (Phillips *et al.*, 2006), which was embedded in R code via the `system` interface. Alternatively, there is the possibility of calling MaxEnt's JAVA code directly from R (package **dismo**, Hijmans *et al.*, 2013), but this option was not available on Macintosh due to interference with Apple's JAVA support system. (This issue is now solved; see the `maxent` documentation in **dismo**).

Strength of R in evolutionary analysis

Data analysis was the was main focus of Ihaka & Gentleman (1996) when they developed the R language, and while R Core Team members have changed there has never been a change in this priority objective. One therefore expects to find useful features of R that lend power to analyses of phylogenetics and evolution. In the following I discuss the four features I consider most important: inter-system interfaces, extensibility, reproducibility, and graphics.

Inter-system interfaces

The "integration [of different analyses] under a single user interface" is one of four principles that Paradis (2012a) proposed to be indispensable for a data analysis environment suitable for evolutionary biologists. This principle stems from the observation that data exchange between systems is error-prone and sometimes impossible if file converters for a specific data format are not available. These problems can be reduced by embedding different analyses in a single system. We should also make use of software existing outside of R, if we are not to disobey Gentleman (2004)'s demand to improve algorithms by *iterative refinement*. The solution allowing integration of different programs, languages, and data storing facilities is *inter-system interfacing* (Ihaka & Gentleman, 1996; Gentleman, 2004; Paradis, 2012a). My programming made extensive use of inter-system interfaces. All taxonomic and geographic data were stored in postgreSQL[a] and accessed by R with the packages **DBI** and **RPostgreSQL**. This is particularly advantageous if R is used together with GRASS GIS as both systems can share databases of vector data (e.g., locality information with attribute data). For raster data, which cannot easily be stored in a relational database, there is another package, **spgrass6** that offers a convenient interface. For molecular and tree-like data (for instances, multiple sequence alignments and phylogenies), the package **ape** provides well tested functions for import and export.

In general the suite of data exchange interfaces in R can be considered as sufficiently complete for evolutionary data. If new data formats come into use, normally any lacuna in tools can be closed rapidly by extending existing resources. For example, the 'Bayesian Evolutionary Analysis by Sampling Trees' (BEAST) platform (Drummond & Rambaut, 2007; Bouckaert *et al.*, 2014) requires complex evolutionary models to be encapsulated in eXtensible Markup Language (XML) language. Building on R's support for XML (Temple Lang, 2013), it is straight forward to create BEAST input files; see `rbeauti` in package **ips**. Likewise, BEAST's heavily annotated NEXUS tree files can be imported with the function `read.beast` from the same package, a simple extension of `read.nexus` (**ape**), which cannot parse more than one parameter per internal node.

Equally important as data transfer are inter-system interfaces to programs written in other programming languages. The `system` command is powerful tool to embed calls to programs written in compiled languages. Two examples from this work are found in divergence time and environmental niche modeling. In the multidivtime approach, five programs written in C language are wrapped in R code (function `multidivtime` in **LAGOPUS**) resulting in a smoothly running molecular dating software. Similarly, the JAVA application MAXENT was embedded in an R script to model species richness and lineage richness. Both applications show that the `system` interface allows to combine the speed of compiled languages with

[a] http://www.postgresql.org

the programmability and flexibility of R.

Trustfulness, according to Chambers (2008) the "prime directive" of data analysis, is another reason to use inter-system interfaces. Many programs are written by leading scientists in their field and have been abundantly tested by the user community. Particularly in phylogenetics, R developers have worked to integrate many of these programs. Examples include multiple sequence alignment (Clustal, Muscle, T-Coffee, mafft, prank), alignment masking (Gblocks), phylogenetic inference (RAxML, MrBayes) and ancestral character estimation (BayesTraits) in the packages **ape** (Paradis *et al.*, 2004) and **ips** (this work). The PHYLIP software package (Felsenstein, 2013) will by available through another package currently developed by Liam Revell (**Rphylip**)[a].

Extensibility

Extensibility, the ability to adapt classes (including their methods) to new data types or algorithms, comes from R's functional approach to object-oriented programming (Chambers, 2008). Although S4 classes have been developed for phylogenetic data (R Hackathon, 2013), they are not yet widely used in phylogenetics and evolution in spite of their reliability. The old S3 classes, e. g., class `phylo` for phylogenetic trees (Paradis *et al.*, 2004), require the user to take care not to mix up tree topology and trait data. As long as this is kept in mind, S3 classes are not difficult to extend; e. g., the function `read.beast` (see previous section) builds on `read.nexus` and returns an enhanced `phylo` object with node statistics summarized from the posterior distribution of a Bayesian analysis. The predictive behavior of R functions due to the functional programming paradigm in R, facilitates this "informal" S3 class extensibility.

Another informal extension is the use of existing functions for the development of new modeling methods. Among phylogenetic and evolutionary analysis in R, tools for studying trait evolution are particularly well developed (Paradis *et al.*, 2004; Harmon *et al.*, 2008; Kembel *et al.*, 2010; Revell, 2012; Orme *et al.*, 2013). These tools can be used as building blocks for specialized analysis pipelines. Cases in point are the non-parametric estimation of ancestral climatic tolerances based on environmental niche models (Evans *et al.*, 2009, chapter , package **phyloclim**)

[a] Available on GitHub: https://github.com/liamrevell/Rphylip

and the randomization tests for niche identity and niche similarity (Warren *et al.*, 2008; packages **phyloclim**, **dismo**).

Reproducibility

Reproducibility is a key feature of trustworthy data analysis and of scientific research in general (Fomel & Claerbout, 2009). While reproducibility is often discussed with regard to peers and society, it is the researcher that first benefits from reproducible data analysis. The programmability of R as an interpreted language together with a clear syntax that builds on a functional programming concept (Becker *et al.*, 1988; Chambers, 2008) enables the user to set up their analyses as 'scripts'. A script guarantees that the whole computation with all parameter settings can be exactly reproduced and, if properly commented, also understood by other users or even the author herself after some time. This concept has been formalized in 'literate programming' (Knuth, 1984), where the source code of a script could be either rendered as program interpretable by a computer ('tangling') or as a document for human readers ('weaving'). Literate programming was introduced as 'dynamic documents' to R with **Sweave** (Leisch, 2002) and has been further developed in the **knitr** package (Xie, 2013), which was used in the writing of this thesis.

In the divergence time estimation of Cucurbitaceae, we relied on dated fossils to obtain absolute age estimates (Schaefer *et al.*, 2009). Setting up the whole analysis in a script was a good way to explore different combinations of fossils for putting time constraints on certain nodes in the tree without unintentional changes in other parameters. In the same way, I compared combinations of environmental layers and approaches to account for sampling bias in species distribution modeling in *Oxalis* (Heibl & Renner, 2012). The code for SDM was hierarchically organized: a master script sequentially called a number of subscripts. Changes of parameter setting were restricted to the master script leaving the majority of the code untouched. These two examples show how programming of R scripts, as long as certain standards are followed (Paradis, 2012b; Ergül, 2013), leads to reproducible analyses that can be easily shared and published.

Graphics

“R is a language and environment for statistical computing and graphics”[a]; the official description explicitly mentions graphics and R indeed offers powerful graphical tools. Box-plot, scatterplot, histogram and many more classical exploratory techniques are available as ‘high-level’ plotting functions, i.e., functions that produce complete plots with axes, grids, and labels. Whereas these functions are simple to use and sufficient for most data types that can be represented by a table, or `data.frame`, the real power comes from ‘low-level’ plotting function such as `points`, `lines`, and `text` to name just a few.

Low-level plotting functions allow putting together new high-level plotting functions for arbitrarily complex vector graphics, e.g., phylogenetic trees and networks (Paradis *et al.*, 2004; Paradis, 2010). Estimation of ancestral climatic tolerances in *Hordeum* (Jakob *et al.*, 2010) shows another example: plotting phylogenetic relationships in niche space (their figures 2.4, S2.5) with `plotAncClim` from the **phyloclim** package. Also developed during this thesis, the **viper** package offers a suite of low-level functions for phylogenetic trees to highlight clades and add clade support values, confidence intervals, (geological) times scales or trait data descriptions (Schaefer *et al.* 2009: Fig. 2; Jakob *et al.* 2010: Figs 1, S1; Heibl & Renner 2012: Figs 2, 3, 4). Together these examples show that the graphical capabilities can be extended to phylogenetic data structures and in most cases there is no need to recur to graphical software outside R.

Data analysis and computation: Outlook

Ten years after his “thoughts on statistical computing”, many of the ideas that Gentleman (2004) envisioned have been realized in the R language. The R Special-Interest-Group Phylogenetics Mailinglist [b], established in 2007, bears witness to the lively interchange between users and developers of phylogenetic methods. The range of available methods is already compelling (reviewed in Paradis, 2012a), and there is no sign of decreasing productivity of package authors. There

[a] http://www.r-project.org/about.html

[b] https://stat.ethz.ch/mailman/listinfo/r-sig-phylo

is almost no open-source software system that R cannot be interfaced with and long-term compatibility with other systems (e.g., the BEAST platform, Bouckaert *et al.*, 2014) is assured at least in principle. Reproducibility has been developed up to the point that setting up a dynamic document with **knitr** (Xie, 2013) is as easy and convenient as working with the best editors available. More and more journals are calling for fully reproducible research, and R users are certainly well equipped to meet this challenge.

The rapid development of R may continue. In order to maneuver through this 'jungle of methods' and make good use of it, users will need a thorough understanding of probability theory and programming techniques. Maybe we should regard R not only as tool for software analysis, but also for teaching. Playing with probability distributions and simulated data can add something concrete to ideas that many feel are abstract and intangible (Bolker, 2008). As an example, Paradis (2012a, p.146–159) explains how to calculate the likelihood of a phylogenetic tree given a set of aligned DNA sequences by developing the code step by step over several pages. Jim Albert has written two packages (**LernEDA**, **LerningBayes**) that focus explicitly on helping to understand exploratory data analyses and Bayesian inference. Similar packages for topics in phylogenetics and evolution could be valuable tools for teaching evolutionary data analysis beyond simple step-by-step protocols.

General Conclusions

Today there is probably no software package or programming environment that allows the user to access more diverse methods for data analysis under a single user interface than R. The strategic choices made in the development of R now seem to pay off, with a lively community of developers and users widening the range of possible applications ever more. The studies presented in this thesis show that most standard procedures in macroevolutionary analyses can be conducted in R. One example from each study also shows that it is relatively easy to develop new techniques or wrap existing software from outside of R.

Taking advantage of R programming, my research has helped to gain insight in evolutionary processes at different taxonomic, geographical, and temporal scales. Divergence time estimates for nearly all genera of Cucurbitaceae combined with ancestral range estimation suggest an Asian origin of Cucurbitaceae in the Late Cretaceous, followed by the repeated spread of lineages into the African, American and Australian continents via at least 43 transoceanic long-distance dispersal events over the last 60 Myr. Environmental niche modeling together with phylogenetic, phylogeographic, and population genetic analyses of four clades of American *Hordeum* indicate that cases of ecoclimatic niche shift were coupled with genetic bottleneck events, whereas lineages that did not encounter population reductions mainly showed ecoclimatic niche conservatism. Finally, a combination of phylogenetic and species distribution modeling showed that vegetative traits in *Oxalis* are conserved and species richness of *Oxalis* in Mediterranean Chile is not a result of adaptive radiation, but of the invasion of seven different lineages with low to moderate diversification.

References

Ammann, C., Jenny, B., Kammer, K. & Messerli, B. (2001) Late Quaternary glacier response to humidity changes in the arid Andes of Chile. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **172**, 313–326.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988) The new S language. *Pacific Grove, Ca.: Wadsworth & Brooks.*

Blattner, F. R. (2006) Multiple intercontinental dispersals shaped the distribution area of *Hordeum* (Poaceae). *New Phytologist*, **169**, 603–614. doi:10.1111/j.1469-8137.2005.01610.x.

Bolker, B. M. (2008) *Ecological models and data in R.* Princeton University Press.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.

Chambers, J. (2008) *Software for data analysis: programming with R.* Springer.

Chambers, J. M. (1998) *Programming with data: A guide to the S language.* Springer.

Cowling, R. M., Rundel, P. W., Lamont, B. B., Arroyo, M. K. & Arianoutsou, M. (1996) Plant diversity in Mediterranean-climate regions. *TRENDS in Ecology and Evolution*, **11**, 362–366.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88.

Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 240.

Drummond, A. J. & Suchard, M. A. (2010) Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, **8**, 114.

Ergül, Ö. (2013) *Guide to Programming and Algorithms Using R.* Springer.

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*). *American Naturalist*, **173**, 225–240. doi:10.1086/595757.

Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.

Felsenstein, J. (2013) PHYLIP (Phylogeny Inference Package) version 3.695. http://evolution.genetics.washington.edu/phylip.html.

Fomel, S. & Claerbout, J. F. (2009) Reproducible research. *Computing in Science & Engineering*, **11**, 5–7.

Gentleman, R. (2004) Some perspectives on statistical computing. *Canadian Journal of Statistics*, **32**, 209–226.

Gentleman, R. & Temple Lang, D. (2004) Statistical analyses and reproducible research. *Bioconductor Project Working Papers*, **2**, 1–36.

Graham, C. H. & Hijmans, R. J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587. doi:10.1111/j.1466-822x.2006.00257.x.

GRASS Development Team (2009) *Geographic Resources Analysis Support System (GRASS GIS) Software.* Open Source Geospatial Foundation, USA.

Harmon, L. J., Schulte, J. A., Larson, A. & Losos, J. B. (2003) Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, **301**, 961–964.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131. doi:10.1093/bioinformatics/btm538.

Harvey, P. H. & Pagel, M. (1991) *The Comparative Method in Evolutionary Biology.* Oxford Univ. Press, Oxford.

Heibl, C. & Renner, S. S. (2012) Distribution models and a dated phylogeny for Chilean *Oxalis* species reveal occupation of new habitats by different lineages, not rapid adaptive radiation. *Systematic Biology*, **61**, 823–834. doi:10.1093/sysbio/sys034.

Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. (2013) *dismo: Species distribution modeling.* R package version 0.9-3.

Hulton, N. R. J., Purves, R. S., McCulloch, R. D., Sugden, D. E. & Bentley, M. J. (2002) The Last Glacial Maximum and deglaciation in southern South America. *Quaternary Science Reviews*, **21**, 233–241.

Ihaka, R. & Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, **5**, 299–314.

Jakob, S. S., Heibl, C., Rödder, D. & Blattner, F. R. (2010) Population demography influences climatic niche evolution: evidence from diploid American *Hordeum* species (Poaceae). *Molecular Ecology*, **19**, 1423–1438. doi: 10.1111/j.1365-294X.2010.04582.x.

Jiménez-Valverde, A. & Lobo, J. M. (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, **31**, 361–369.

Kaplan, M. R., Fogwill, C. J., Sugden, D. E., Hulton, N. R. J., Kubik, P. W. & Freeman, S. P. H. T. (2008) Southern Patagonian glacial chronology for the Last Glacial period and implications fot Southern Ocean climate. *Quaternary Science Reviews*, **27**, 284–294.

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.

Kishino, H., Thorne, J. L. & Bruno, W. J. (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.*, **18**, 352–361.

Knuth, D. E. (1984) Literate programming. *The Computer Journal*, **27**, 97–111.

Leisch, F. (2002) Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat*, pp. 575–580. Springer.

Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol*, **5**, e1000520.

Liu, C., White, M. & Newell, G. (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, **40**, 778–789. doi:10.1111/jbi.12058.

Luebert, F. & Pliscoff, P. (2006) *Sinopsis bioclimática y vegetacional de Chile.* Editorial Universitaria, Santiago de Chile.

Martins, E. P. & Hansen, T. F. (1997) Phylogenies and the comparatice method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, **149**, 646–667.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C. & Brumfield, R. T. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Moore, G. E. (1965) Cramming more components onto integrated circuits. *Electronics*, **38**, 82–85.

Neteler, M. & Mitasova, H. (2008) *Open source GIS - a GRASS GIS approach.* Springer.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N. & Pearse, W. (2013) *caper: Comparative Analyses of Phylogenetics and Evolution in R.* R package version 0.5.2.

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

Paradis, E. (2012a) *Analysis of Phylogenetics and Evolution with R.* Springer.

Paradis, E. (2012b) Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. *PLoS Comput Biol.*

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290. doi:10.1093/bioinformatics/btg412.

Peterson, A. T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265–1267.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S. J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

R Hackathon (2013) *phylobase: Base package for phylogenetic structures and comparative data.* R package version 0.6.5.2.

Ree, R. H., Moore, B. R., Webb, C. O. & Donoghue, M. J. (2005) A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, **59**, 2299–2311.

Ree, R. H. & Smith, S. A. (2008) Maximum likelihood inference of geographic evolution by dispersal, local extinction, and cladogenesis. *Syst. Biol.*, **57**, 4–14.

Renner, S. S. (2005) Relaxed molecular clocks for dating historical plant dispersal events. *Trends in plant science*, **10**, 550–558. doi:10.1016/j.tplants.2005.09.010.

Revell, L. J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

Ronquist, F. (1997) Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. *Syst. Biol.*, **46**, 195–203.

Sanderson, M. J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.*, **19**, 101–109.

Schaefer, H., Heibl, C. & Renner, S. S. (2009) Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 843–851. doi:10.1098/rspb.2008.1447.

Schluter, D., Price, T., Mooers, A. Ø. & Ludwig, D. (1997) Likelihood of ancestral states in adaptive radiation. *Evolution*, **51**, 1699–1711.

Schuster, P. (2013) A silent revolution in mathematics. *Complexity*, **18**, 7–10.

Søndergaard, H. & Sestoft, P. (1990) Referential transparency, definiteness and unfoldability. *Acta Informatica*, **27**, 505–517.

Stott, L., Poulson, C., Lund, S. & Thunell, R. (2002) Super ENSO and global climate oscillations at millennial time scales. *Science*, **297**, 222–226.

Temple Lang, D. (2013) *XML: Tools for parsing and generating XML within R and S-Plus.* R package version 3.98-1.1.

Thompson, L. G., Davis, M. E., Mosley-Thompson, E., Sowers, T. A., Henderson, K. A., Zagorodnov, V. S., Lin, P. N., Mikhalenko, V. N., Campen, R. K., Bolzan, J. F., Cole-Dai, J. & Francou, B. (1998) A 25,000-year tropical climate history from Bolivian ice cores. *Science*, **282**, 1858–1864.

Thorne, J. L. & Kishino, H. (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.*, **51**, 689–702.

Thorne, J. L., Kishino, H. & Painter, I. S. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Warren, D. L., Glor, R. E. & Turelli, M. (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, **62**, 2868–2883. doi:10.1111/j.1558-5646.2008.00482.x.

Wiens, J. J., Ackerly, D. D., Allen, A. P., Anacker, B. L., Buckley, L. B., Cornell, H. V., Damschen, E. I., Jonathan Davies, T., Grytnes, J.-A., Harrison, S. P., Hawkins, B. A., Holt, R. D., McCain, C. M. & Stephens, P. R. (2010) Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, **13**, 1310–1324. doi:10.1111/j.1461-0248.2010.01515.x.

Wiens, J. J. & Graham, C. H. (2005) Niche conservatism: intergrating evolution, ecology, and conservation biology. *Annu. Rev. Ecol. Evol. Syst.*, **36**, 519–539.

Xie, Y. (2013) *Dynamic Documents with R and knitr.* CRC Press.

Yang, Z. (2007) Paml 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.

Yesson, C. & Culham, A. (2006) Phyloclimatic modeling: combining phylogenetics and bioclimatic modeling. *Systematic Biology*, **55**, 785–802.

Zuckerkandl, E. & Pauling, L. (1965) *Evolutionary divergence and convergence in proteins.* Academic Press, New York.