Phylogenetic Biogeography with the R package *BioGeoBEARS*: testing for trait-dependent dispersal in southern conifers

The goal of phylogenetic biogeography is to understand how and why the geographic distributions of lineages have changed over millions of years.

Modern methods attempt to do this by using explicit, probabilistic models of the process of geographic range evolution, and applying them to datasets consisting of (1) a time-scaled phylogeny, plus (2) the geographic ranges of the species at the tips of the phylogeny.

By fitting different models to a single dataset, we hope to be able to infer which models fit best, as well as infer the relative influence of different processes. Once a model has been selected, we may then use to estimate the probability of ancestral ranges at any point on the phylogeny.

The method we use to fit models is called "Maximum Likelihood," (ML) and the method used to statistically compare models is called Akaike Information Criterion (AIC). If you have taken BIOSCI 220: Quantitative Biology, these methods should be familiar to you. If not, it has been reviewed in lecture, and in any case, instructions are given here.

Software: *BioGeoBEARS* is a freely available R package. To install it on a University of Auckland computer, open *RStudio*, and copy/paste these commands:

```
library(ape)
library(optimx)
library(GenSA)
library(FD)
library(rexpokit)
library(cladoRcpp)
library(devtools)
devtools::install_github(repo="nmatzke/BioGeoBEARS", INSTALL_opts="--byte-
compile")
```

The "library" command loads R packages that have already been installed. UoA computers already all of the R packages on CRAN (the Comprehensive R Archive Network) installed, so all you should need is the "library" command.

If you find that a package is missing, causing an error in the *library* command, or you are doing the lab on a laptop or other computer, you will have to install any missing packages before doing the *library* command. This can be done with the "install.packages" command, e.g.: `install.packages("ape")`

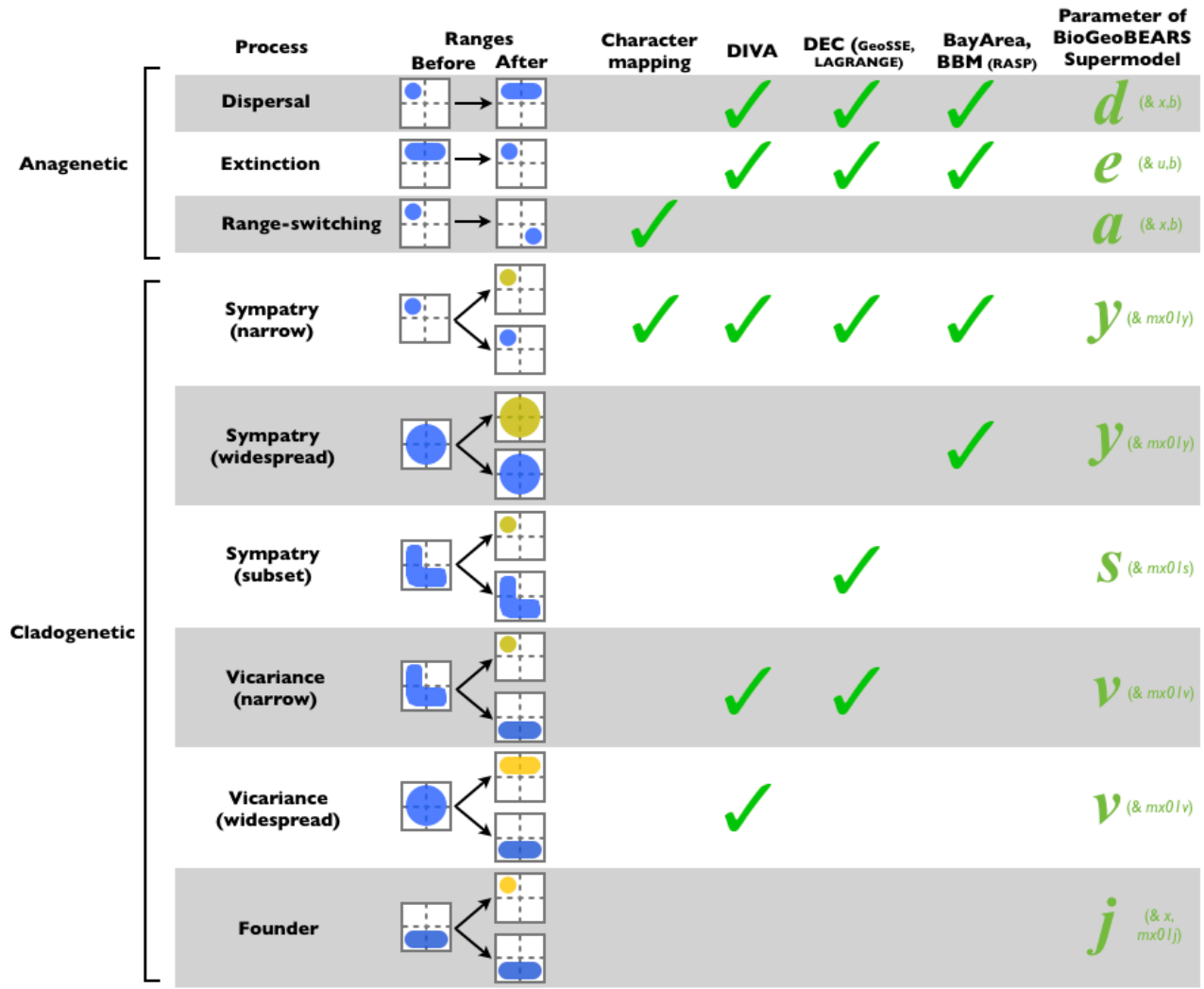
This only has to be done once per package on a personal computer; once a package has been installed, it is saved on your hard drive.

BioGeoBEARS itself is not on CRAN, but instead on GitHub, at https://github.com/nmatzke/BioGeoBEARS . This is why the special `devtools::install_github` command is used for *BioGeoBEARS*.

For a summary of the capabilities and rationale behind *BioGeoBEARS*, see:

Matzke, Nicholas J. (2013). "Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing." *Frontiers of Biogeography*, 5(4), 242-248. http://escholarship.org/uc/item/44j7n141

Figure 1 gives a summary of the biogeographical processes allowed under different *BioGeoBEARS* models:



Dataset: The data and models we will be using are a near-complete dated phylogeny of southern conifers. Famous representatives include kauri (in family Araucariaceae) and totara and rimu (family Podocarpaceae).

This dataset was assembled in this publication: Klaus, Kristina; Matzke, Nicholas J. (2020). Statistical Comparison of Trait-dependent Biogeographical Models indicates that Podocarpaceae Dispersal is influenced by both Seed Cone Traits and Geographical Distance. *Systematic Biology*, 69(1), 61-75.

Some of these species have fleshy cones that are bird-dispersed, and others do not:

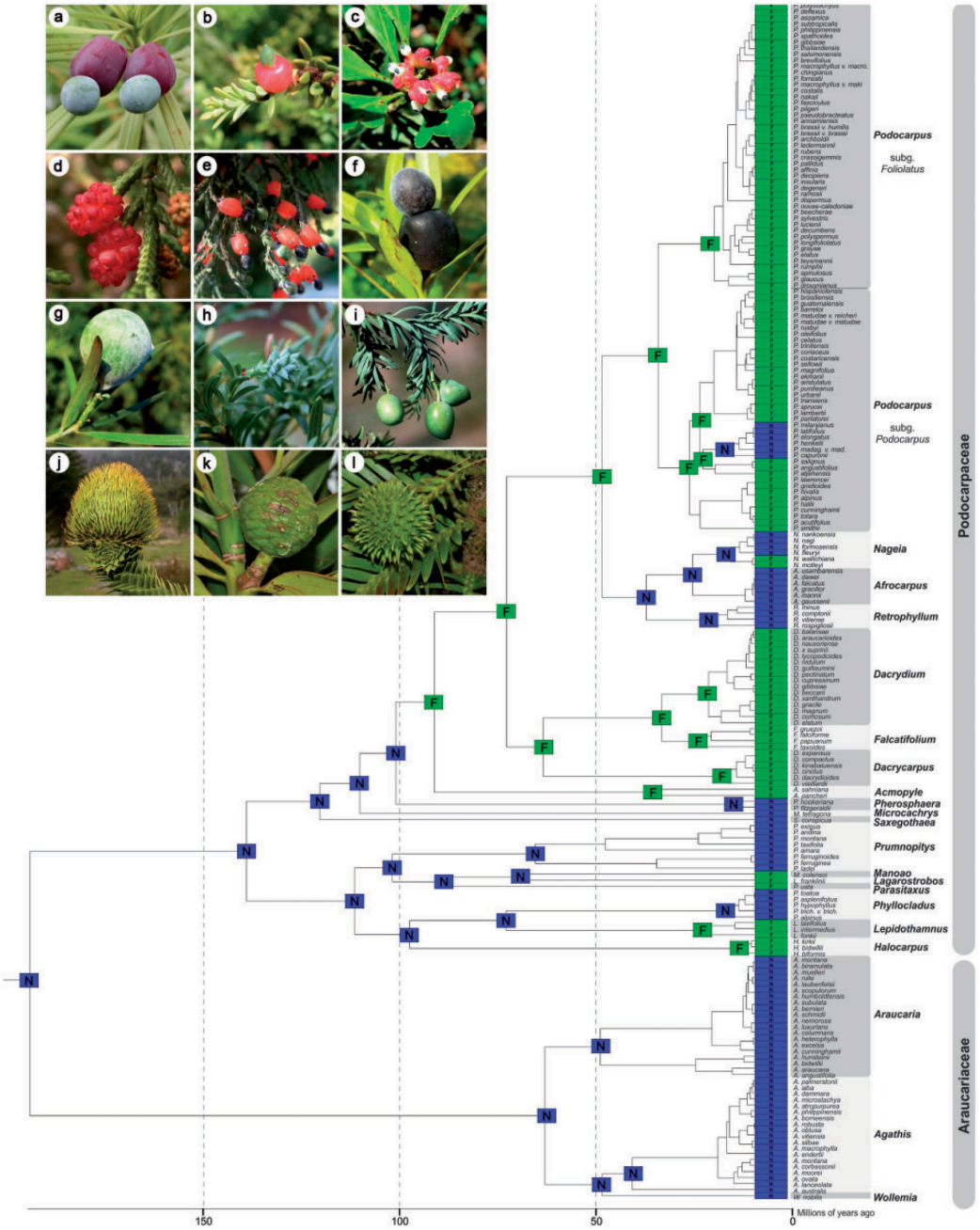


FIGURE 1. Seed cones of the Podocarpaceae and Araucariaceae. a) *Podocarpus macrophyllus*, b) *Podocarpus nivalis*, c) *Phyllocladus asplenifolius*, d) *Microcachrys tetragona*, e) *Dacrycarpus dacrydioides*, f) *Podocarpus elatus*, g) *Afrocarpus falcatus*, h) *Saxegothaea conspicua*, i) *Prumnopitys amara*, j) *Araucaria araucana*, k) *Agathis robusta*, and l) *Wollemia nobilis*. Depicted at each node: the most-probable ancestral state for the seed cone fleshiness trait in Podocarpaceae and Araucariaceae under the best-fitting model (DIVALIKE+j+x+t_{12}+t_{21}+m_2). F (light nodes) = fleshy seed cone structures; N (dark nodes) = nonfleshy seed cone structures.

Although the full dataset took many weeks of computation to analyse, we can do a similar analysis on a simplified dataset. We will use this dataset to ask the following questions:

- Which biogeographical "base models" are the best fit to this dataset?
- Does adding geographic distance as a predictor of dispersal probability improve the model fit? How much?
- What are the estimated values of the parameters under the best-fit model?
- How many dispersal events are estimated to have occurred under the best-fit model, and what is the uncertainty of these estimates?
- How much statistical support is there for the role of the cone-fleshiness trait increasing the probability of long-distance dispersal?

As this dataset is simplified from the one published in Klaus & Matzke (2020), you may not get identical answers!

We will also ask you to evaluate your inferences and the models used to make them: all of these biogeographical models must be a great deal simpler than the real-life processes, so how much do we believe the results that we got? What improvements could be made to make the models and resulting inferences more realistic?

Setup instructions

For: BIOSCI 395 lab, room 435-108 (58 Symonds St.).

Room 435-108 is a Macintosh lab. We found that the most feasible way to access RStudio in this lab was via "RStudio Cloud," a service that runs RStudio through a web browser. The same instructions should work on any computer using Rstudio Cloud through a web browser.

You are also free to run the lab on your own laptop. The instructions will be very similar, but you will have to download and install R/RStudio onto your computer (these are free downloads, use google to find them).

1. Google "RStudio Cloud." Get a *free* Rstudio Cloud account and login. (You can use e.g. a gmail account, or set up a different login.)

2. Once you have logged in, click "New Project -> New RStudio Project"

3. Wait a few seconds for the session to set up. Give the project a title like "BGBlab" (for BioGeoBEARS lab).

4. To run the exercise, we will first need to install some R packages.

   Go to: 395_lab_setup_v1.R

   Run these commands:

   (PS: Note that "these quotes" and "these quotes" are different. R only accepts the latter, as they are plain ASCII text.)

   ```
   install.packages("devtools")
   install.packages("Rcpp")
   install.packages("ape")
   ```

```
install.packages("FD")
install.packages("phytools")
install.packages("phangorn")
install.packages("phylobase")
install.packages("optimx")
install.packages("GenSA")
install.packages("gdata")
install.packages("snow")
install.packages("rexpokit")
install.packages("cladoRcpp")
```

…these packages exist on CRAN (the Comprehensive R Archive Network) and are pre-compiled, so they should install easily on all machines.

It *should* be the case that you only ever need to run these install.packages command once per R installation – but they are easy to re-run if needed (new machine, new version of R, etc.).

5. Next, we need to install Matzke's R package, BioGeoBEARS. This exists on GitHub, a website which hosts code that is in active development. The devtools library allows us to install packages from GitHub:

```
library(devtools)
install_github(repo="nmatzke/BioGeoBEARS", upgrade="never")
```

A bunch of warnings may come up, but ignore the warnings. (If you get an actual "Error", then contact an instructor: errors are worse than warnings.)

Check that your installation of BioGeoBEARS worked by typing:

```
library(BioGeoBEARS)
```

If no error messages are returned, then BioGeoBEARS has been loaded into memory.

R tips:
* *install.packages* downloads a package from CRAN onto your machine. This only has to be done once per R installation.
* *library* loads a particular package into the R workspace for use. This must be done for every new R session.

6. Copy/paste these commands into the RStudio console:

```
labdir = paste(extdata_dir, "examples/395lab/", sep="/")
labpt1a = paste(extdata_dir, "examples/395lab/Psychotria_M0_equalRates/",
sep="/")
labpt1b = paste(extdata_dir,
"examples/395lab/Psychotria_M2_oneWayDispersal/", sep="/")
labpt1c = paste(extdata_dir,
"examples/395lab/Psychotria_M4_DistanceDispersal/", sep="/")
labpt2a = paste(extdata_dir, "examples/395lab/conifer_DEC_traits_models/",
sep="/")
labpt2b = paste(extdata_dir,
"examples/395lab/conifer_DEC+x_traits_models/", sep="/")
labpt1a_script = paste(extdata_dir,
```

```
"examples/395lab/Psychotria_M0_equalRates/Psychotria_M0_v1.R",
sep="/")
labpt1b_script = paste(extdata_dir,
"examples/395lab/Psychotria_M2_oneWayDispersal/Psychotria_M2_oneWa
yDispersal_v1.R", sep="/")
labpt1c_script = paste(extdata_dir,
"examples/395lab/Psychotria_M4_DistanceDispersal/Psychotria_M4_Distan
ceDispersal_v1.R/", sep="/")
labpt2a_script = paste(extdata_dir,
"examples/395lab/conifer_DEC_traits_models/conifer_DEC_traits_models_
v1.R", sep="/")
labpt2b_script = paste(extdata_dir,
"examples/395lab/conifer_DEC+x_traits_models/conifer_DEC+x_traits_mod
els_v1.R", sep="/")
```

7. The above set up the locations of the example .R scripts within the BioGeoBEARS GitHub install.

   To open one of the scripts inside RStudio Cloud, type e.g.:

   file.edit(labpt1a_script)

8. To take commands from a .R script, and run them inside the console window, you can:

   * Copy/paste into the console window
   * Highlight the text with the mouse, and press the Run button
   * Highlight the text with the mouse, and press CTRL-ENTER
   * Note that hitting the UP arrow brings back the previous commands you ran

   Note that accidentally skipping part of a script is the most common cause of errors!

9. Once you have loaded BioGeoBEARS, have successfully opened a .R script, and figured out how to get script commands to run in the Console, proceed to the R scripts below.

Detailed instructions

Detailed instructions will be provided in a plain-text ".R" file containing the necessary R code to run basic analyses, along with questions to answer along the way.

Week 1. Introduction to BioGeoBEARS: Statistical Model Comparison with the Hawaiian *Psychotria* dataset

As discussed in lecture, Hawaiian *Psychotria* shrubs are a common example dataset for biogeographical models.

Inside the BGB lab directory, you will see 3 Psychotria directories:

Psychotria_M0_equalRates – models with equal dispersal rates between all 4 islands

Psychotria_M2_oneWayDispersal – models where only dispersal to younger islands is allowed

Psychotria_M4_DistanceDispersal – models where dispersal depends on relative distance

Open RStudio. Navigate to each of these directories, open the script (the .R file)

1. Run the script in chunks by copying/pasting major chunks, or highlighting code and pressing CTRL-ENTER.

2. Look at the resulting graphics and try to interpret, with a partner, what each model suggests about the history of this group.

3. Extract the log-likelihood, number of free parameters, and parameter estimates. Calculate AIC, delta AIC, relative likelihood, and AIC weights. Follow the instructions here: http://brianomeara.info/aic.html

   (this is best done in Excel or Google Sheets)

   An example AIC table is here: https://docs.google.com/spreadsheets/d/1ioiALPYMEnwtu6HewUP3iA8ue5k9bZSxPtfm_PT93ow/edit?usp=sharing

   To read more about AIC, see:

   Franklin, Alan B.; Shenk, Tanya M.; Anderson, David R.; Burnham, Kenneth P. (2001). "Statistical Model Selection: An Alternative to Null Hypothesis Testing." *Modeling in Natural Resource Management: Development, Interpretation, and Application*. Edited by T. M. Shenk and A. B. Franklin. Washington, Island Press: 75-90.
   Link: https://books.google.com.au/books?id=Uk7rZ7DCvY4C&dq=burnham+and+anderson&lr=&source=gbs_navlinks_s

   And: http://phylo.wikidot.com/advice-on-statistical-model-comparison-in-biogeobears#refs

4. For Part 1 of the Lab Report, give an approximately 1-2 page summary of your findings (this should include the AIC table, and perhaps screenshot of a graphic if you find it useful to explain something), as if you were giving a summary of a research paper. Include answers to these questions:

   a. Which model is the best fit? Were there any other competitive models?

   b. What does the model-comparison exercise suggest about the process of dispersal in Hawaiian *Psychotria*?

   c. What is the basic conclusion about the biogeographic history of Hawaiian *Psychotria*, from the study you have done?

UPDATE, 16 August 2021: In lab, we discovered that we only had time to work through/understand the *labpt1a* script, i.e. the 6 models in *Psychotria_M0_equalRates*. The week 1 assignment & AIC table/writeup can

therefore be based on just these 6 models. The others (part 1b and 1c) can easily be run, and included if you like, but this is not required.

Part 2. Trait-dependent dispersal in southern conifers

As discussed in lecture, a more complex hypothesis & model is that traits can influence macroevolutionary dispersal raites. As before, run the scripts in each of these directories:

conifer_DEC_traits_models – trait-independent and trait-dependent dispersal models, without geographic distance playing a role

conifer_DEC+x_traits_models – trait-independent and trait-dependent dispersal models, *with* geographic distance playing a role

Note that PDFs and .Rdata files of the results are available in the 395lab.zip zipfile, in case there are difficulties running these models, or displaying the large phylogenies (197 species!).

As before, run the scripts (these will take awhile). Interpret the results and make a model-comparison table.

For Part 2 of the Lab Report, give an approximately 1-2 page summary of your findings (this should include the AIC table, and perhaps screenshot of a graphic if you find it useful to explain something), as if you were giving a summary of a research paper. Include answers to these questions:

a. Which model is the best fit? Were there any other competitive models?

b. What does the model-comparison exercise suggest about the process of dispersal in southern conifers?

c. What are the basic conclusion(s) about the biogeographic and trait history of southern conifers, from the study you have done?

Marking

Your assignment will be marked based on the results tables and your interpretations. It is worth 15% of the grade. It is due (digitally on Canvas, in Word, PDF, or similar) by due 4 pm September 27th.

We will be available to help you in the lab sessions, and the lab should be doable during the lab sessions. Please post questions to Piazza so that the answers can help everyone.