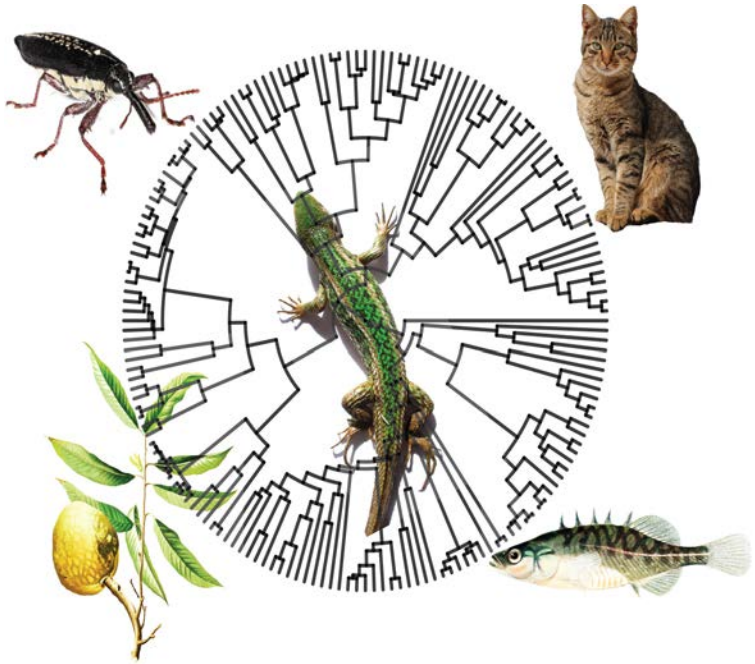
Biology 559R: Introduction to Phylogenetic Comparative Methods

Topics for this week:

- Course general information
- About the course
- Course objectives
- Comparative methods: An overview
- R as language: uses and benefits
- Basic R functions
- Data manipulation for analysis in R



Course general information

- Lectures and Workshops: Tuesday and Thursday afternoon 2:00 – 2:50 p.m.
- Room: 2142 LSB (2nd floor LSB)
- Instructor: Juan C. Santos
  Office: LSB 4136
  Phone: 801-422-4398
  E-mail: infraguttatus@gmail.com
  Office Hours: Appointment by email

- Bring your own laptop if you have one, with latest **R 3.1.2** installed:

  MAC OSX: http://cran.r-project.org/bin/macosx/

  Windows: http://cran.r-project.org/bin/windows/base/

Grading

	Total Grade (%)	Due Date
Participation (showing up in class, asking questions, participate in the discussions)	20	
Project Proposal (5-10 minute presentation)	10	Feb-5
Final Presentation (10-15 minute presentation)	20	Apr 9 and Apr 14
Final Project Report	50	Day of presentation

- There are no exams

Course general information

Date	Topic	Software and R packages
Jan 6	Introduction to R and comparative methods	--
Jan 8	Introduction to R basic functions	--
Jan 13	Getting sequences from GenBank	'ape', 'seqinr'
Jan 15	Alignment: Simultaneous alignment and tree estimation Alignment visualization and manipulation	'Sate-II', 'Mesquite'
Jan 20	Statistical estimation of models of sequence evolution	'jmodeltest'
Jan 22	Implementation of models of sequence evolution and phylogenetic inference	'Garli-2.0', 'RAxML'
Jan 27	Chronogram estimation	'BEAST'
Jan 29	Tree visualization, plotting	'FigTree', 'ape'
Feb 3	Tree retrieval, manipulation and simulations	'ape', 'geiger'
Feb 5	Project Proposal Presentation	--
Feb 17	Ancestral state reconstruction (discrete)	'ape', 'phytools'
Feb 19	Ancestral state reconstruction (continuous)	'ape', 'phytools'
Feb 24	Diversification Analysis	'ape', 'geiger'
Feb 26	Diversification Analysis: BiSSE based models	'diversitytree'
Mar 3	Diversification Analysis: GoeSSE based models	'diversitytree'
Mar 5	Trait evolution: Data manipulation and exploration	--
Mar 10	Trait evolution: Correlation analyses (continuous traits)	'geiger', 'caper'
Mar 12	Trait evolution: Correlation analyses (discrete traits)	'geiger', 'caper'
Mar 17	Trait evolution: Rates of trait evolution	'geiger', 'phytools'
Mar 19	Trait evolution: Trait simulations	'geiger', 'caper'
Mar 24	Trait evolution: Multivariate methods (PPCA)	'phytools', "caper'
Mar 26	Request to practice of specific methodologies	--
Mar 31	Request to practice of specific methodologies	--
Apr 2	Request to practice of specific methodologies	--
Apr 7	Request to practice of specific methodologies	--
Apr 9	Student presentations of final project	--
Apr 14	Student presentations of final project	--

Course website:

http://www.jcsantosresearch.org/

- Updated regularly
- Lecture pdfs will be placed before class
- Relevant papers and datasets will be placed there

Reference Textbook (not required):

Emmanuel Paradis (2012) Analysis of Phylogenetics and Evolution with R. Second Edition. Springer. ISBN 978-1-4614-1742-2

*This book is available to download at the Harold B. Lee Library

About the course

• This course is a ‘hands on’ practice and application of comparative methods.

• The course will cover from very basic data manipulation starting from the construction of phylogenetic trees to analyses of multivariate datasets.

• Different methods will be introduced for analyzing character evolution and evolutionary process along phylogenetic trees.

• The theoretical foundations of techniques will be discussed, but the emphasis of this course is on the “practical” implementation of comparative methodologies. Please read the relevant papers posted on the website.

• We will use different R packages and other freely available software that allow us to analyze and retrieve data from online repositories, but students are also encouraged to use their own research data for this purpose.

About the course

The main topics include:

(1) Brief introduction to R scripting and environment (Other software will also be introduced)

(2) Data retrieval from online databases for the construction of phylogenetic trees

(3) Sequence alignment and phylogenetic inference

(4) Tree manipulation and visualization

(5) Ancestral state reconstruction (continuous and discrete traits)

(6) Diversification analyses and trait evolution inferences

About the course

The main topics include:

(7) Tree and data simulations

(8) Multivariate comparative methods

(9) Specific topics or R-packages that you might want to review

See this page for some ideas:

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

Course objectives

- To help you to prepare and organize your data for phylogenetic comparative analyses.

- To provide you with the basis to do data manipulation using different R-packages and other software.

- To provide you a general introduction of current comparative methods. For your research projects, it is expected that you will develop further understanding of some particular methods or R-packages that turn out to be most appropriate.

- The creative combination of functions from different R-packages will also be explored in order to achieve the objective of the topics for each class.

Course objectives

• Learning R has a notoriously steep learning curve and the primary focus of this course is the direct application of methods: learn by doing.

• The independent projects aim that the students (i.e., you) conduct further practice of the comparative methods using their own selected datasets with a final report and presentation.

Comparative methods: An overview

Vol. 125, No. 1 The American Naturalist January 1985

PHYLOGENIES AND THE COMPARATIVE METHOD

JOSEPH FELSENSTEIN

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195

Submitted November 30, 1983; Accepted May 23, 1984

Recent years have seen a growth in numerical studies using the comparative method. The method usually involves a comparison of two phenotypes across a range of species or higher taxa, or a comparison of one phenotype with an environmental variable. Objectives of such studies vary, and include assessing whether one variable is correlated with another and assessing whether the regression of one variable on another differs significantly from some expected value. Notable recent studies using statistical methods of this type include Pilbeam and Gould's (1974) regressions of tooth area on several size measurements in mammals; Sherman's (1979) test of the relation between insect chromosome numbers and social behavior; Damuth's (1981) investigation of population density and body size in mammals; Martin's (1981) regression of brain weight in mammals on body weight; Givnish's (1982) examination of traits associated with dioecy across the families of angiosperms; and Armstrong's (1983) regressions of brain weight on body weight and basal metabolism rate in mammals.

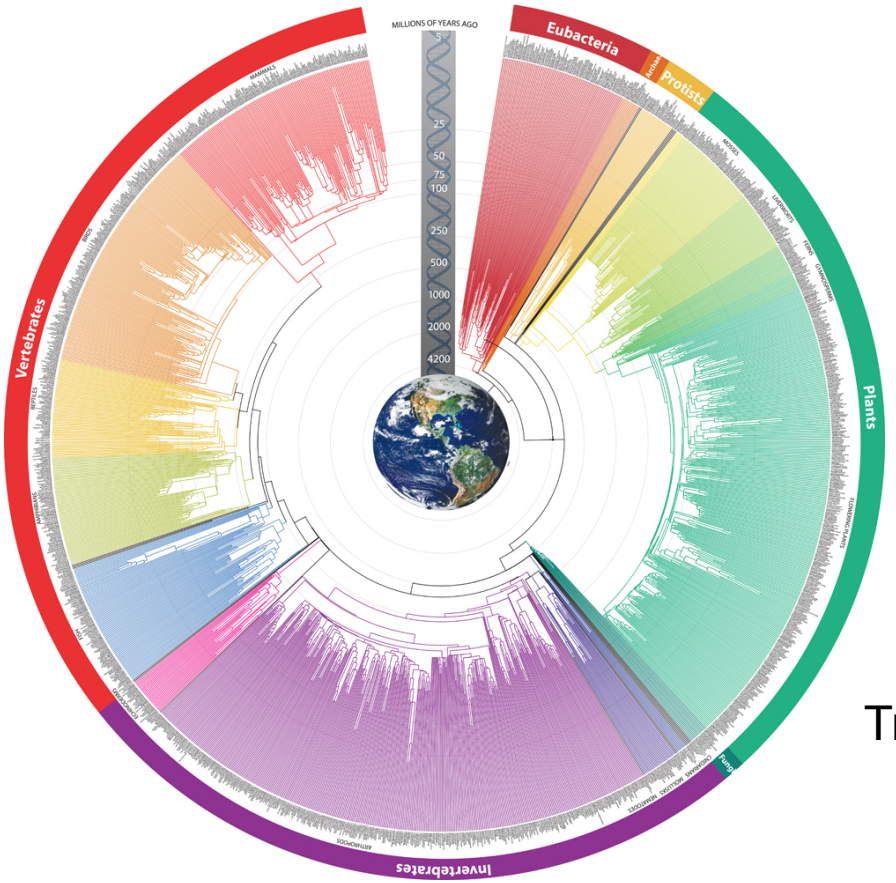
My intention is to point out a serious statistical problem with this approach, a problem that affects all of these studies. It arises from the fact that species are part of a hierarchically structured phylogeny, and thus cannot be regarded for statistical purposes as if drawn independently from the same distribution. This problem has been noticed before, and previous suggestions of ways of coping with it are briefly discussed. The nonindependence can be circumvented in principle if adequate information on the phylogeny is available. The information needed to do so and the limitations on its use will be discussed. The problem will be discussed and illustrated with reference to continuous variables, but the same statistical issues arise when one or both of the variables are discrete, in which case the statistical methods involve contingency tables rather than regressions and correlations.

Course website:

http://www.jcsantosresearch.org/pdf/week_1/Felsenstein_1985.pdf

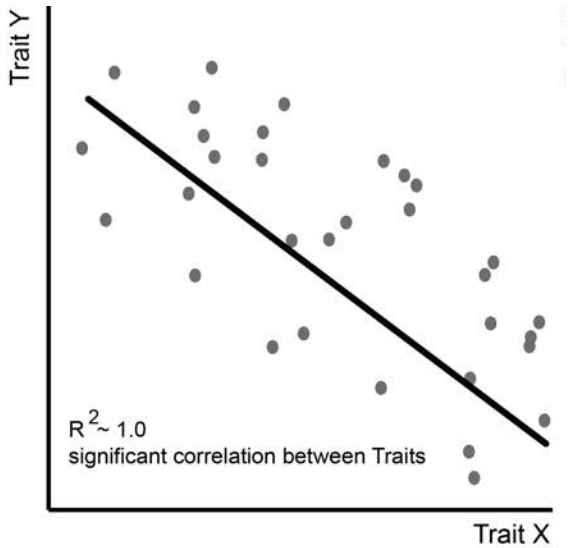
Comparative methods: An overview

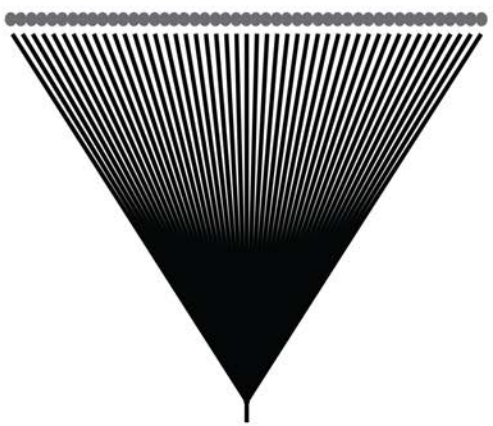
Most statistical methods that compare individuals, populations, species or lineages assume the independence of observations, when in fact most biological groups are differentially related to each other according to their evolutionary history (phylogenetic context)



Tree of Life

Comparative methods: An overview





Assumes each data point is an independent sample

In the figure is visually evident a correlation between Trait A and B.

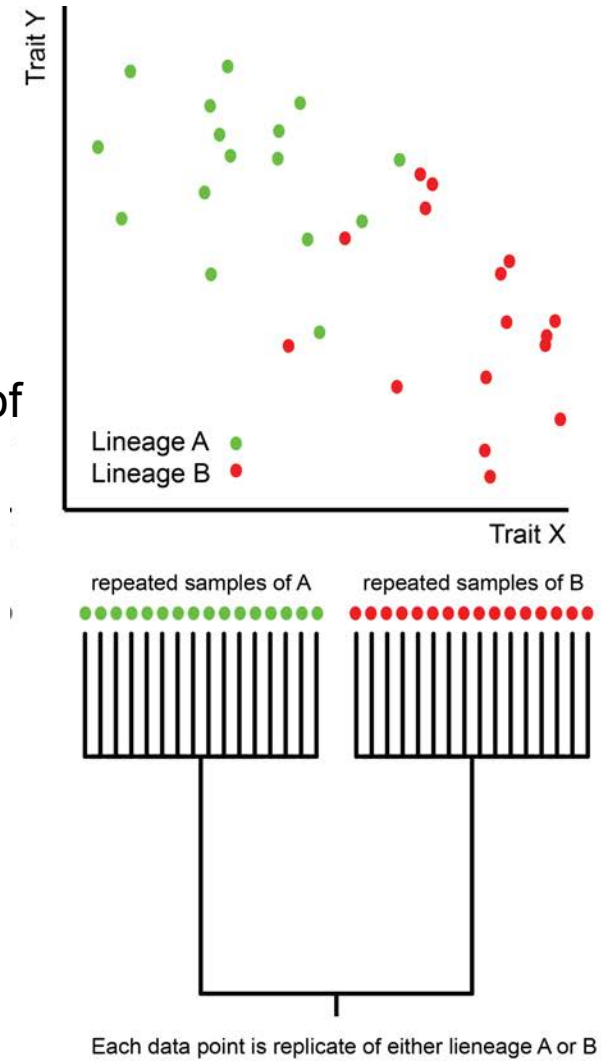
Most statistical methods assume that the measurements from samples are not bias in regard to their origin and find this correlation significant.

Yet for most biological data, this is not necessarily true and some level of relationship exist.

Comparative methods: An overview

For example, this correlation is driven by two clusters of data that correspond to two clades

Each measurement is a ‘pseudoreplicate’ estimate of the same lineage

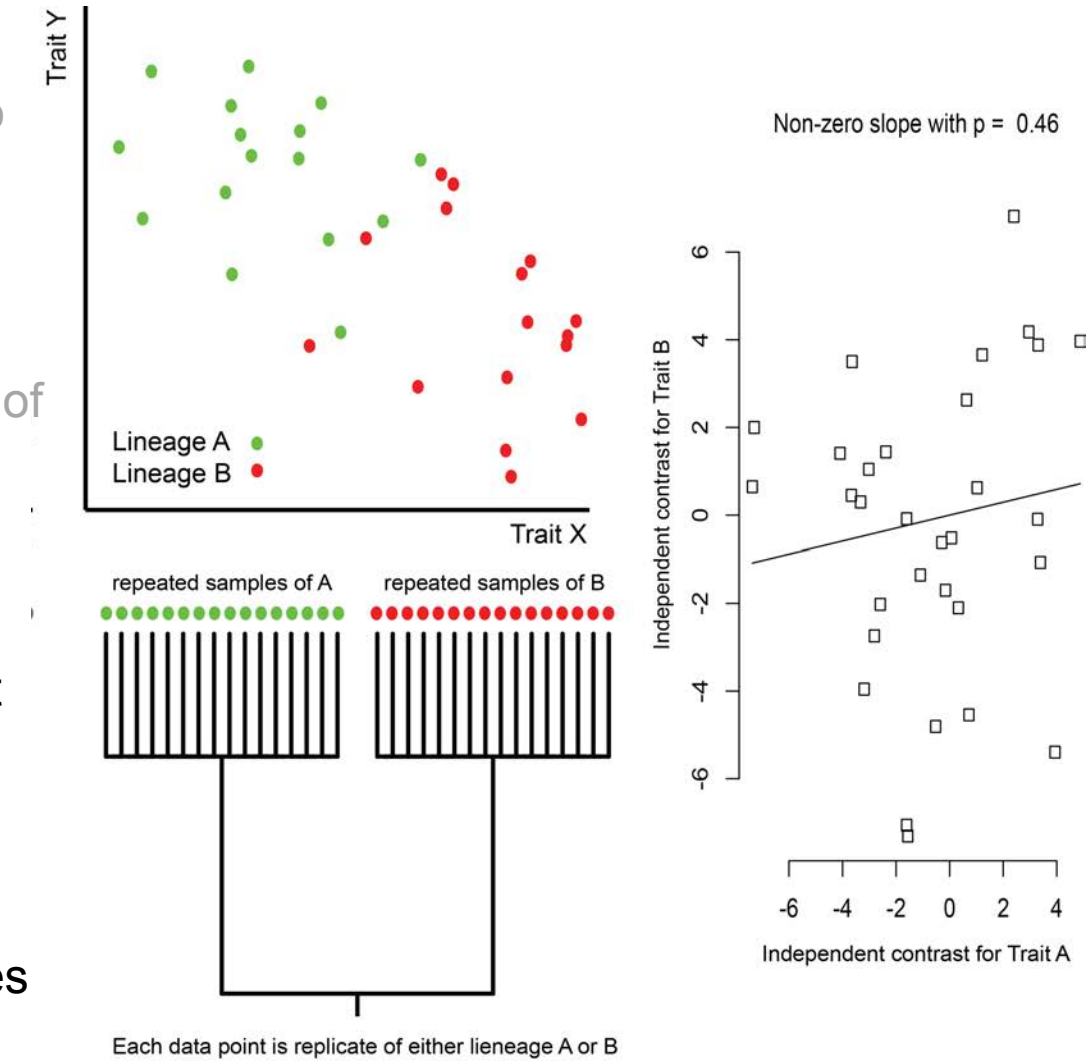


Comparative methods: An overview

For example, this correlation is driven by two clusters of data that correspond to two clades

Each measurement is a ‘pseudoreplicate’ estimate of the same lineage

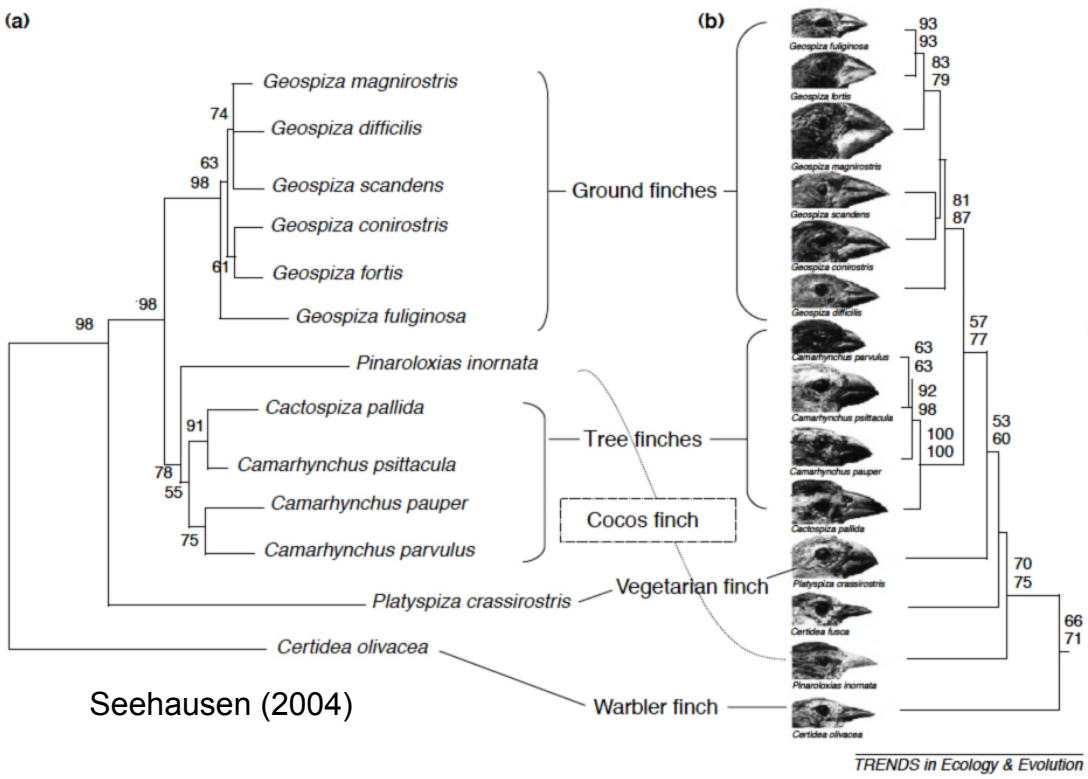
We can correct this bias using phylogenetic independent contrasts that converts the original measurements
into contrasts between pairs of related taxa or (estimated) ancestral nodes in the phylogeny.



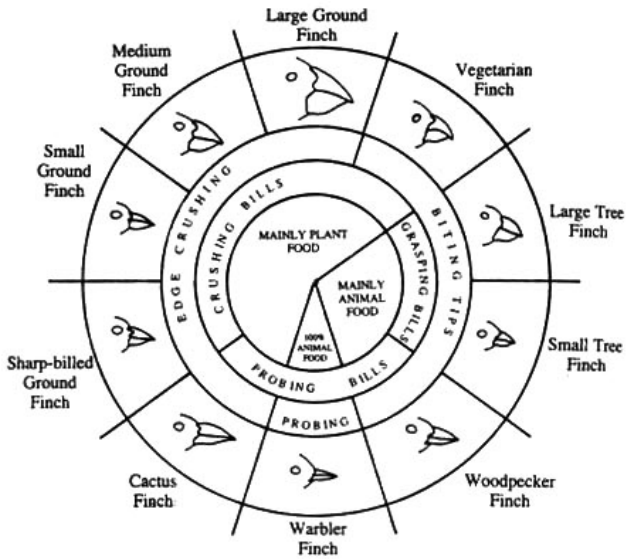
Comparative methods: An overview

• The ‘comparative method’ allows the determine the significance of the differences between populations, species or other higher taxa.

• These methods also allow to elucidate past evolutionary process including adaptation as response to natural selection causing differential reproductive success and changes in the frequency of heritable traits.

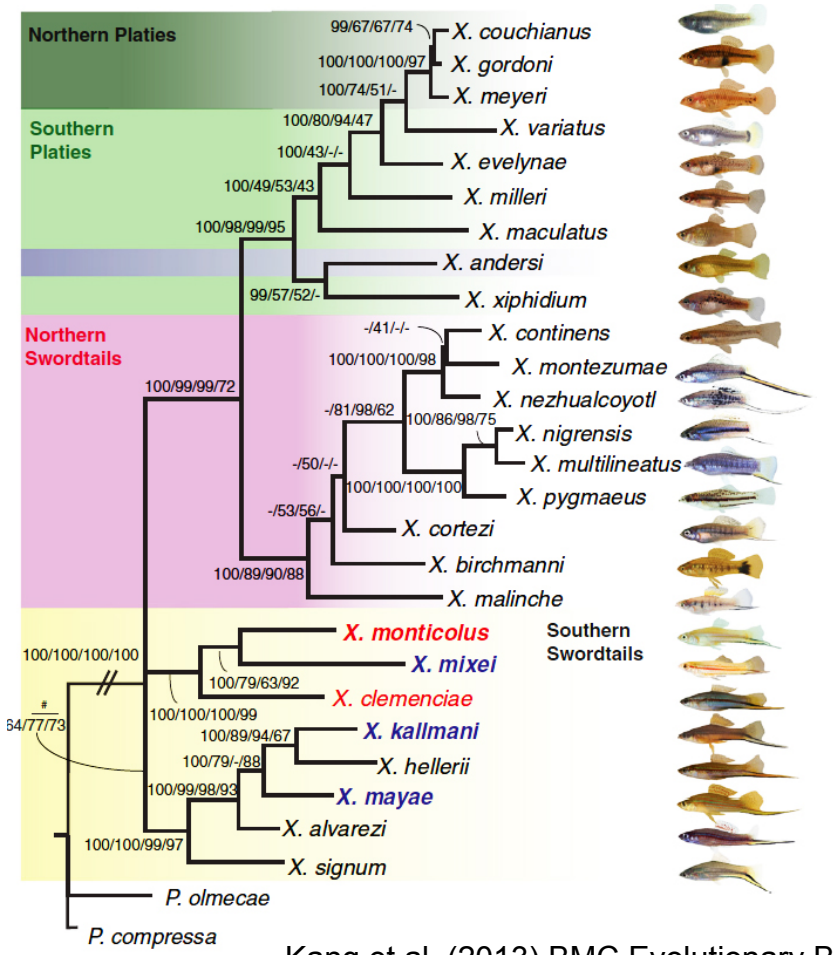


Seehausen (2004)



Comparative methods: An overview

• Other evolutionary process such the effects of sexual selection (which might not be adaptive) can also be studied using comparative methods.



• The evolution of complex courtship displays in swordtails (*Xiphophorus* sp.) with very long tails continues to puzzle the field of sexual selection.

• Comparative methods help to understand the evolutionary basis of female mate choice that drive the evolution of not necessarily adaptive male traits.

Kang et al. (2013) BMC Evolutionary Biology

Comparative methods: An overview

• The development and implementation of comparative phylogenetic methods provided a better understanding of biological evolution (adapted from Garland *et al.* 2005. JEB 208, 3015-3035). Some **key** conceptual advances are:

(1) The results of comparative analyses does not necessarily mean evidence of adaptation.

(2) The incorporation of phylogenetic information increases both the quality and even the type of inference from data alone.

(3) All organisms are related to each other at some level, taxa cannot be assumed to be independent of each other for statistical purposes.

Course website:

http://www.jcsantosresearch.org/pdf/week_1/Garland_et_al_2005.pdf

Comparative methods: An overview

• The development and implementation of comparative phylogenetic methods provided a better understanding of biological evolution (adapted from Garland *et al.* 2005. JEB 208, 3015-3035). Some **key** conceptual advances are:

(4) Statistical analyses of that use comparative data must assume or estimate some model of character evolution for effective inference.

(5) Phylogenies help on experimental design by suggesting taxa to be used in comparative analyses in regard to their phylogenetic affinities (e.g., closely versus distantly related taxa).

(6) Phylogenetically based comparisons are purely correlational and inferences of causation can be further tested using enhanced experimental manipulations.

What is R?

• R is a language and environment for gathering and manipulating data, applying statistical and numerical computations, summarizing and visualizing the results.

Website: http://www.r-project.org/

• R provides a wide variety of statistical (linear and nonlinear modeling, classical/Bayesian statistical tests, graphics, phylogenetic methods, genomic analyses, superb data handing).

• The R language provides an Open Source route to participation in that activity (more than 6135 available contributed packages).

List of packages: http://cran.r-project.org/ #see under software (left column) > Packages

• R is available as a Free Software. It runs on most platforms: Linux, Windows and Mac OS.

• You can write your own functions for specific needs

What is R?

• R has an comprehensive documentation

http://cran.r-project.org/doc/manuals/R-intro.html

http://cran.r-project.org/manuals.html

http://cran.r-project.org/other-docs.html

• Other online archives provide visual and problem-based examples:

http://www.bioconductor.org/

http://www.r-phylo.org/wiki/Main_Page

http://rgm3.lab.nig.ac.jp/RGM/R_image_list?page=1436&init=true

https://plot.ly/r/

http://www.r-bloggers.com/

• Many general R-related topics can be easily search using web browsers (e.g., google) or for very specific websites that uses wiki format (allows collaborative modification)

http://stackoverflow.com/

Disadvantages of R

• R uses scripts to run functions and commands rather than displaying menus (‘point and click’ based software). R scripting requires practice and time to learn.

• Some functions in R can take time to master and the causes of errors are ‘non-specific’ and difficult to understand.

• The R-environment evolves fast and some packages become ‘obsolete’ if not maintained. Some package functions also become ‘deprecated’ and they may be removed in the future.

• Data entry is not direct and requires formatting (e.g., enter your data in a standard spreadsheet program and then save in a readable text format.)

Disadvantages of R

• Different packages share the ‘same name’ functions causing that some packages to ‘mask’ others which might prevent correct calculations.

• User provided packages might have ‘errors’ or ‘bugs’ in their functions or calculations which are not obvious to the user unless he/she understands what the software is doing.

Text Editors

• We are going to use at least one text editor for most of the course. Some free options are:

MAC OS: http://www.barebones.com/products/textwrangler/download.html

Windows: http://notepad-plus-plus.org/download/v6.6.9.html

Across platforms: http://jedit.org/

• For TextWrangler (and the others), you might like to add the following preferences:

TextWrangler>Preferences>Appearance: select Line numbers and Tab stops

TextWrangler>Preferences>Editor Defaults: select Show invisible characters and Show spaces

For next class print the reference card

- We are going to use the R-reference card from Matt Baggott: ‘Baggott_refcard_v2.pdf’

*you can get this card from: http://cran.r-project.org/other-docs.html

- Think about your possible projects and read Brian O'Meara's CRAN Task View: Phylogenetics, Especially Comparative Methods for specific packages that you might like to review:

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