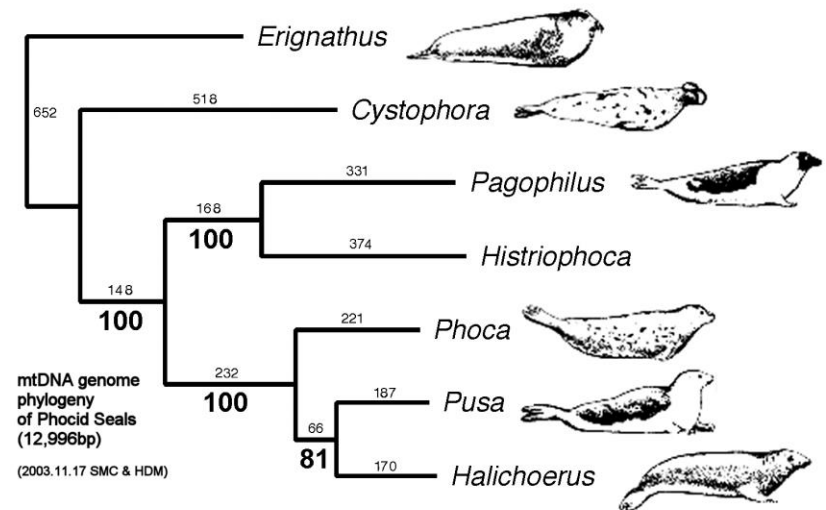
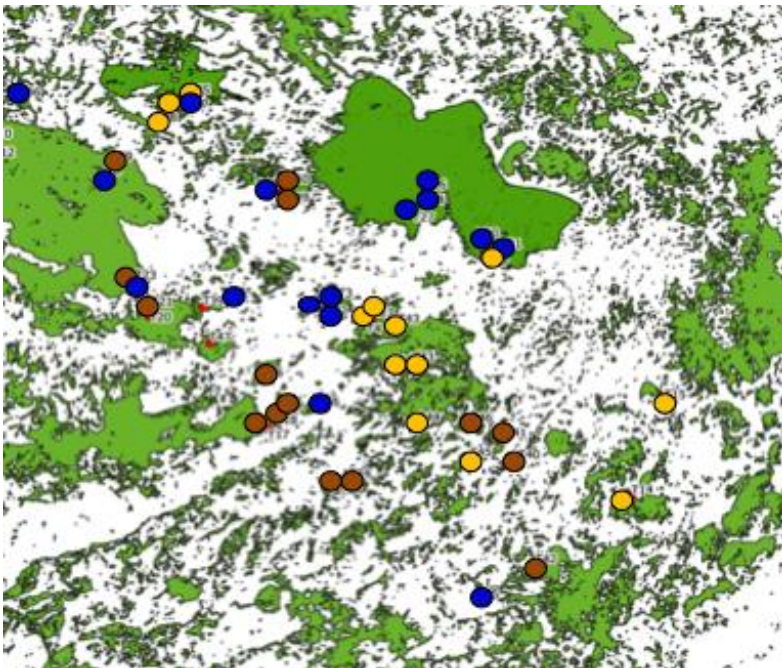


# Lesson 9:

## Phylogenetic regression

# Dealing with auto-correlation

- Data points that are close together in space, time or phylogeny are not independent.



# IID

## Independently Identically Distributed

Statistical models assume that generated errors,  $\varepsilon$ , are IID.

- We assume that all errors are sampled from the same distribution (identically distributed)
- We assume that that they are independent of one another (If we know  $\varepsilon_k$  we cannot predict  $\varepsilon_{k+3}$ )

In general, it is hard to get errors fully IID, but we try to do so.

# Getting to IID

In order to get errors close to IID, we try to include information into our model of response  $Y$  that account for sources of covariance among  $y_i$

e.g. in lmm, we included random blocking effects to account for grouped data pts

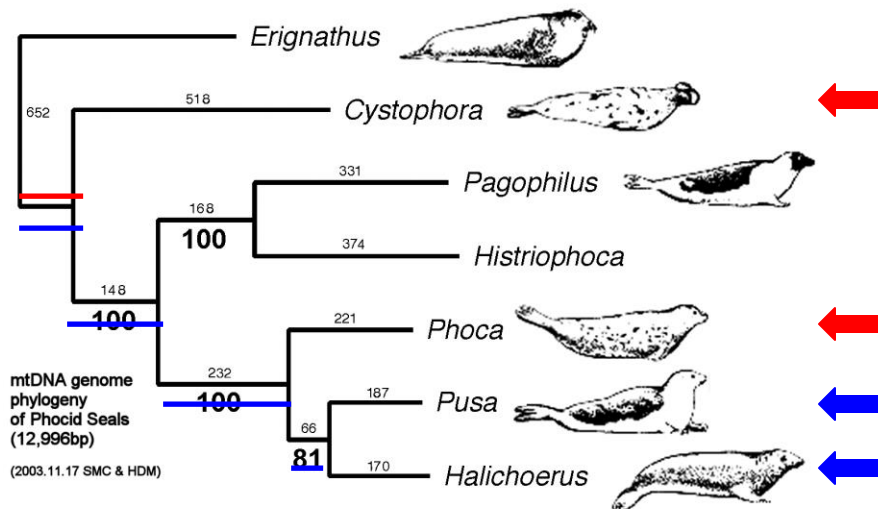
(pts from the same group may be more covariant with one another than pts outside the group)

e.g. in gls, we include information about distance between points (pts closer in distance may be more covariant than pts further apart)

# Covariance in phylogenies

Covariance in phylogenies is fairly intuitive

- > based on shared evolutionary history
- > measured using common branch length in phylogeny
- > always relative, being based on the phylogeny provided



Species that diverged recently have longer common evolutionary history and hence high covariance

Species that diverged a long time ago have shorter common evolutionary history and low covariance

Shared path length = measure of covariance

# Phylogenies & The Comparative Method

Many of us are involved in trying to understand variation among species (morphology, physiology, functional traits, niche space), by relating differences between them to environmental drivers. This is known as the Comparative method.

It is intuitive that two closely related species could share much greater variation because of their common evolutionary history than distantly related species.

When trying to understand variation differences between species in response to present-day environmental pressure, it is essential that we account for this shared evolutionary history.

Code 9.1

Evolution in a phylogeny

# Evolution or environment?

## Phylogenetic regression

Phylogenetic regression allows us to incorporate both sources of variation, treating the evolutionary history as random variation and concentrating on the environment components as fixed effects.



# The linear model with error covariance

- Recall that:  $\underline{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$  where  $\underline{\varepsilon} \sim N(0, \mathbf{V})$

Where  $\mathbf{V} = \text{Cov}[\underline{\varepsilon} | \mathbf{X}]$ , the covariance matrix of errors. Ideally all off-diagonal elements of  $\mathbf{V}$  are zero (i.e. no covariance)

$$\mathbf{Y} = \mathbf{X} * \underline{\beta} + \underline{\varepsilon} \quad \text{where} \quad \underline{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}) \quad \checkmark$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix} \quad \text{where}$$

$$\begin{bmatrix} \varepsilon_{1,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \varepsilon_{2,2} & 0 & \dots & 0 & 0 \\ 0 & 0 & \varepsilon_{3,3} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \varepsilon_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & \varepsilon_{n,n} \end{bmatrix}$$

# The linear model with error covariance

- In reality  $\underline{\epsilon} \sim N(0, \mathbf{V})$ , where off-diagonal elements are non-zero. How large this off-diagonal component is affects how reliable our model inferences are.

$$\begin{array}{c}
 \text{Y} \\
 \left[ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{array} \right]
 \end{array}
 =
 \begin{array}{c}
 \text{X} \\
 \left[ \begin{array}{ccc} \text{Int} & \text{X1} & \text{X2} \\ 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ & \vdots & \vdots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{array} \right]
 \end{array}
 *
 \begin{array}{c}
 \beta \\
 \left[ \begin{array}{c} \beta_0 \\ \beta_1 \\ \beta_2 \end{array} \right]
 \end{array}
 +
 \begin{array}{c}
 \epsilon \\
 \left[ \begin{array}{c} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{n-1} \\ \epsilon_n \end{array} \right]
 \end{array}
 \quad \text{where}$$

$\underline{\epsilon} \sim N(0, \mathbf{V})$



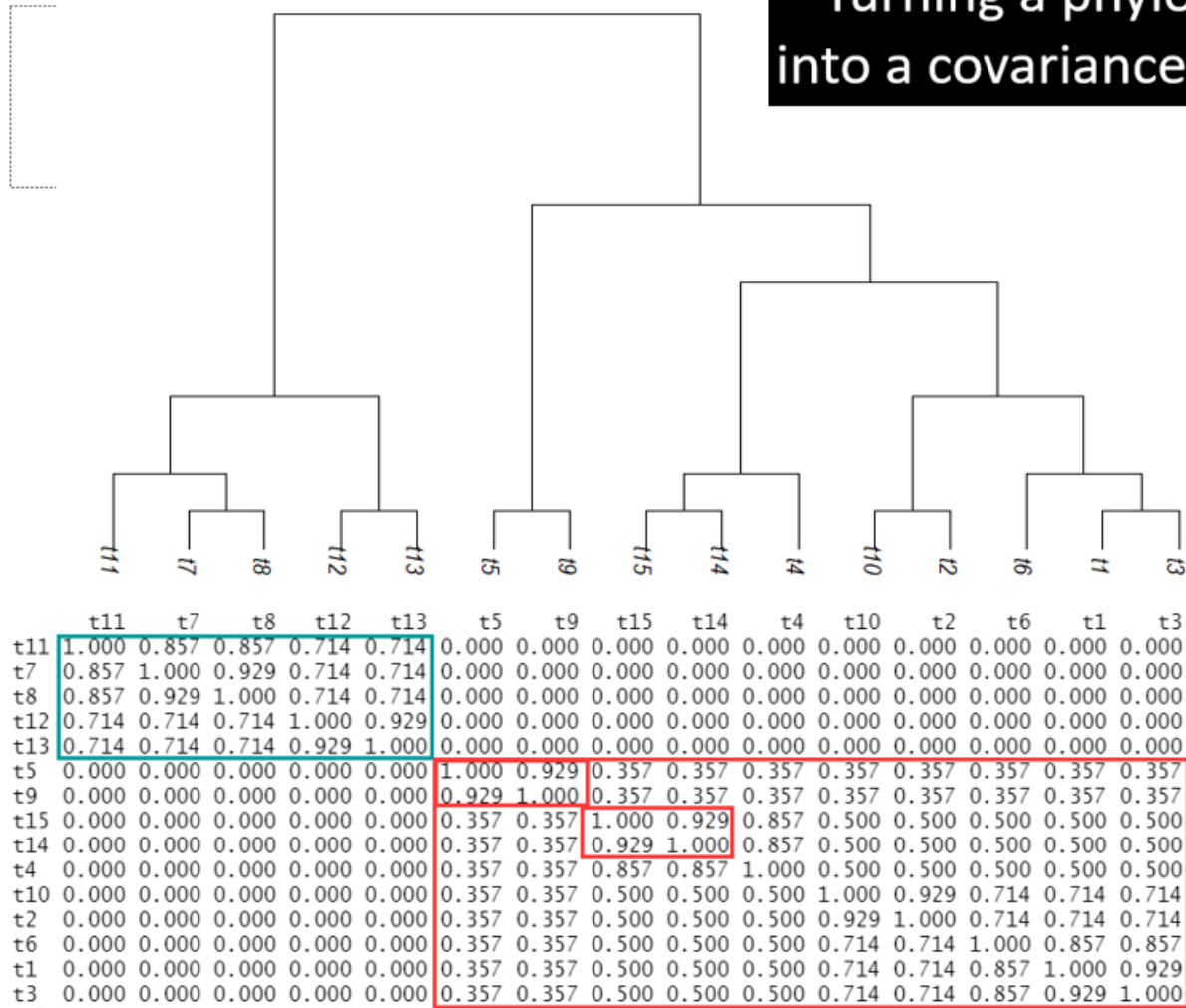
$$\left[ \begin{array}{cccccc}
 \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} & \dots & \epsilon_{1,n-1} & \epsilon_{1,n} \\
 \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} & \dots & \epsilon_{2,n-1} & \epsilon_{2,n} \\
 \epsilon_{3,1} & \epsilon_{3,2} & \epsilon_{3,3} & \dots & \epsilon_{3,n-1} & \epsilon_{3,n} \\
 \vdots & \vdots & \vdots & & \vdots & \vdots \\
 \vdots & \vdots & \vdots & & \vdots & \vdots \\
 \epsilon_{n-1,1} & \epsilon_{n-1,2} & \epsilon_{n-1,3} & \dots & \epsilon_{n-1,n-1} & \epsilon_{n-1,n} \\
 \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} & \dots & \epsilon_{n,n-1} & \epsilon_{n,n}
 \end{array} \right]$$

We use phylogeny to reduce covariance between error points.

## Code 9.2

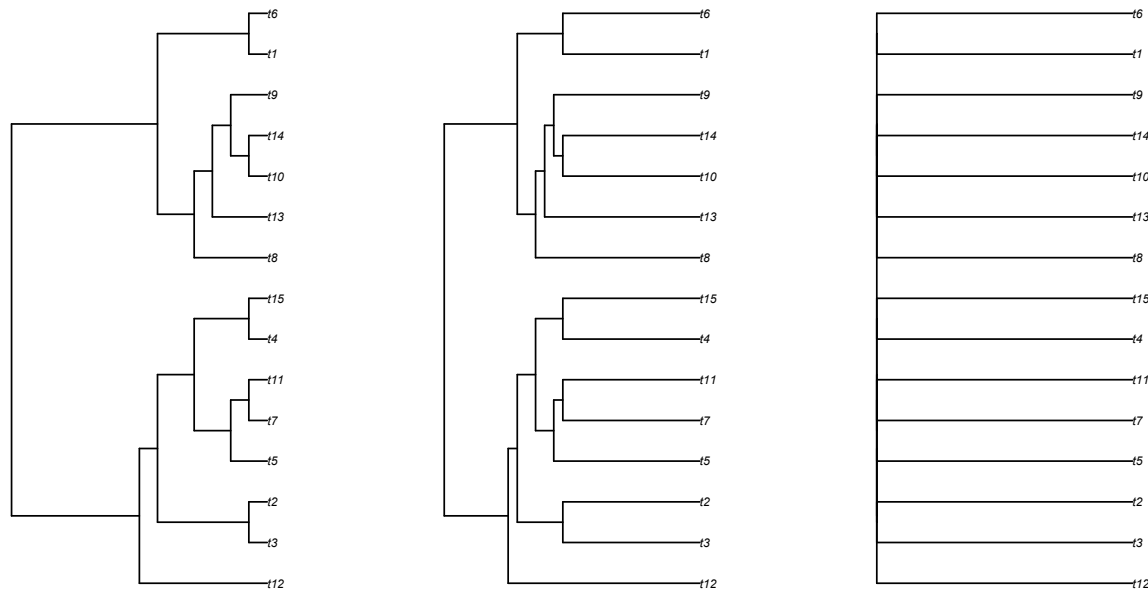
Turning a phylogeny  
into a covariance matrix

# Turning a phylogeny into a covariance matrix



# Modifying phylogenetic signal

OK, so we can model evolutionary covariance between tips (species), but the nature of covariance could change over history...



Strong co-evolutionary signal



Weak co-evolutionary signal

Code 9.3

Modifying phylogenetic signal

# The gls() function for phylogenetic regression

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

```
mod <- gls(y~x, data=dat,  
  correlation=corBrownian(phy= tree.primates))
```

- gls() is set up in the *ape* package to model a covariance matrix that represents the shared evolution between species across the tree, as apparent in the phylogeny
- We can either simply provide the phylogeny OR we can provide the covariance matrix

```
correlation=corSymm(TreeCovar1[lower.tri(TreeCovar1)], fixed=TRUE)
```

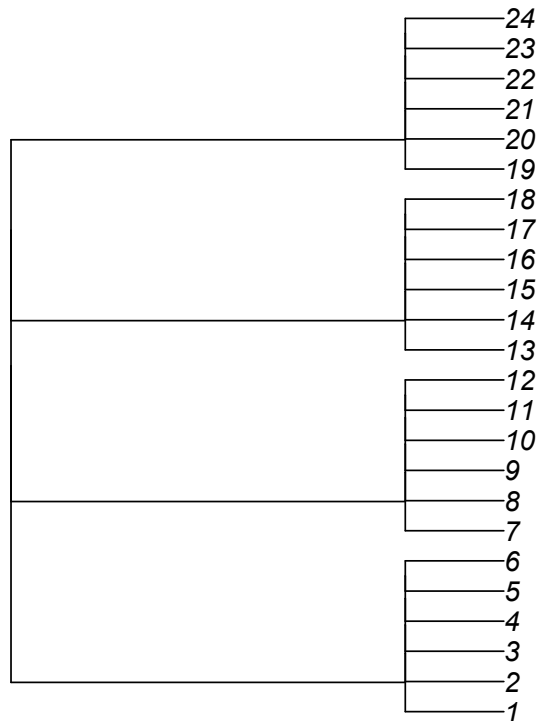
# Code 9.4

## Running gls()



# The link between hierarchical data and covariance: Imm vs gls

It is possible to show that Imm and gls generate similar regression estimates when the data is grouped



The trick is formulating the covariance matrix correctly!

The way to think about this is as some kind of star phylogeny with multiple values per branch

-> grouped values are dependent on one another

Code 9.5

Hierarchical data: Imm vs gls

# The gls() function for phylogenetic regression

```
mod <- gls(y~x, data=dat,  
           correlation=corBrownian(phy= tree.primates))
```

- These corClass objects build the covariance matrix for you.
  - corBrownian
  - corMartins
  - corGrafen
  - corPagel
  - corBlomberg [\(see ?corClasses in R to get more information\)](#)
- These represent different assumptions on trait evolution... which one to choose??

# The model of evolution

There are many ways to view how evolution of traits is happening. We will cover two common ones:

1. **Brownian motion:** trait evolution is purely a random walk process, in which case variance across the phylogeny should increase linearly with time
1. **Stabilized selection (Orstein-Uhlenbeck process):** over time the values across the phylogeny tend towards the long-term mean, so variance is constrained with time more than under BM (traits are constrained by physical limitations)

Code 9.6

Model of evolution: BM vs OU

# Testing for phylogenetic signal

- Here are 4 possible methods (using Pagel's  $\lambda$  as example):
- 1. Likelihood ratio test - compare model with and without phylogeny ( $\lambda=0$ )
- 2. Parametric bootstrap of parameter estimate - Fit the model (estimate  $\lambda$ ), then use the parameter estimate to simulate datasets to get an approximate sampling distribution for the estimator; not good here because the estimator is biased
- 3. Parametric bootstrap of  $H_0$  - Fit the model under  $H_0$  ( $\lambda=0$ ), simulate data based on that model and fit simulations to estimate null distribution. Compare to  $\lambda$  generated by true data.
- 4. Permutation test – Gold standard for correlated data. Traits randomly permuted among species without changing structure of phylogeny, generating a distribution of possible  $\lambda$  values. Compare  $\lambda$  generated by true data to this distribution.

# Bootstrapping versus Permutation

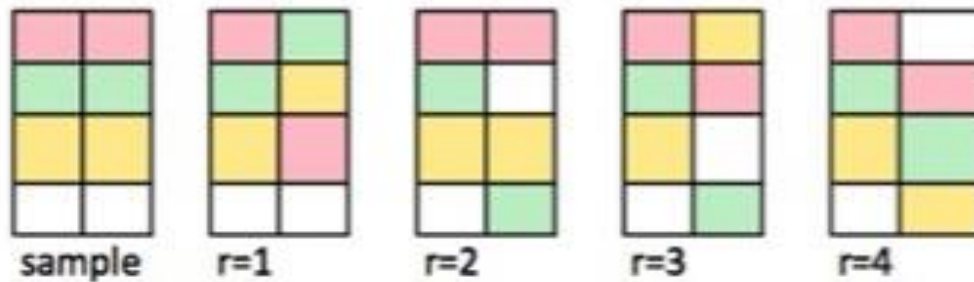
## Comparison of Resampling Techniques

Resampling	Procedure	Applications
<b>Permutation</b>	Samples are drawn at random from original pool without replacement	Tests of hypotheses
<b>Bootstrap<sup>1</sup></b>	Samples are drawn with replacement	Tests of hypotheses <b>AND</b> Standard error, bias, and confidence intervals of estimator
<b>Jack Knife<sup>2</sup></b>	Samples consist of original pool with one at a time withheld	Standard error, bias, and confidence intervals of estimator

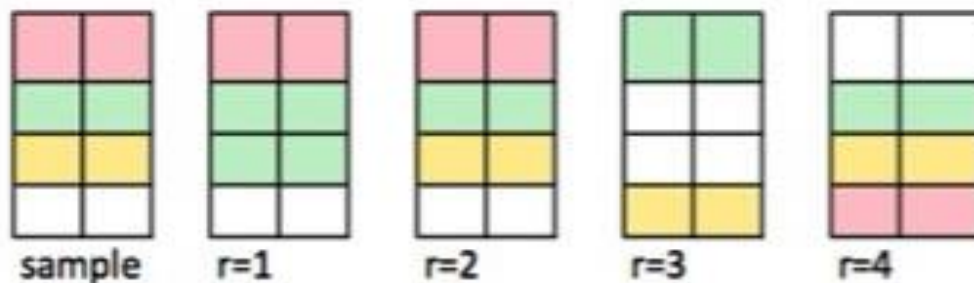
1 Most versatile.

2 Generally outperformed by others.

# Bootstrapping versus Permutation



Permutation  
Randomization test



Bootstrap



# Bootstrapping versus Permutation

## Differences between bootstrap and permutation tests

- Bootstrap
  - estimates confidence interval, bias and standard error
  - Simulates data under the alternative hypothesis
  - Sampling is done with replacement of subjects
  - Many bootstrap samples because of replacement
- Permutation tests
  - estimates p-value and distribution under the null.
  - Simulates data under the null hypothesis
  - Sampling is done without replacement of subjects
  - Finite number of potential permutation samples

Code 9.7

Testing for phylogenetic signal

Phylo signal in the predictor  
( $y = X\beta + \varepsilon$ , where  $\varepsilon \sim N(0, V)$  )

- Caveat: pGLS methods assume phylogenetic structure in the response variable; I am unaware whether there are methods that can deal with phylogenetic signal in both the response AND predictor variables..

Code 9.8

Phylogenetic signal in the independent  
variable

# Running the phylo regression model ( $\underline{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ , where $\underline{\varepsilon} \sim N(0, \mathbf{V})$ )

- Once you have established that there is phylogenetic signal or not, proceed to run analysis and check the fixed model component  $\mathbf{X}\underline{\beta}$
- The fixed model can be evaluated using t-tests, LRTs as before

# Real world problems

Typically for the comparative method you will have the following before you start the analysis in R:

1. A spreadsheet of species with trait & environment data
2. A phylogeny of those species or a larger phylogeny that you can cut down to your set of species

- Steps of analysis:

- A. read in both data files
- B. reduce both data components to the minimum list of common species between dataframe and phylogeny
- C. Check for phylogenetic signal
- D. Run first `gls()` and select best model of fixed effects

# The savanna tree seedling dataset

Question: do traits of seedling trees of humid and semi-arid savannas differ?

Environment: humid savannas are subject to more frequent fire whereas semi-arid savannas are subject to more water stress



# The savanna tree seedling dataset

## Hypotheses:

H1: Tree seedlings in fire-prone systems allocate more biomass root storage ->  $\uparrow$  root mass fractions.

H2: Tree seedlings in water-stressed environments access deeper water ->  $\uparrow$  root extension rate.



## Experiment: Common garden

Tree species from humid and semi-arid savannas on 3 continents.

Grown for 20 weeks.

Many traits measured.






## Code 9.9

The savanna tree seedling dataset

H1: RMF is greater among humid species

# Exercise 9.1

- Test hypothesis 2
- H2: Tree seedlings in water-stressed environments access deeper water ->  root extension rate.  
(the variable is RER in mm day<sup>-1</sup>)