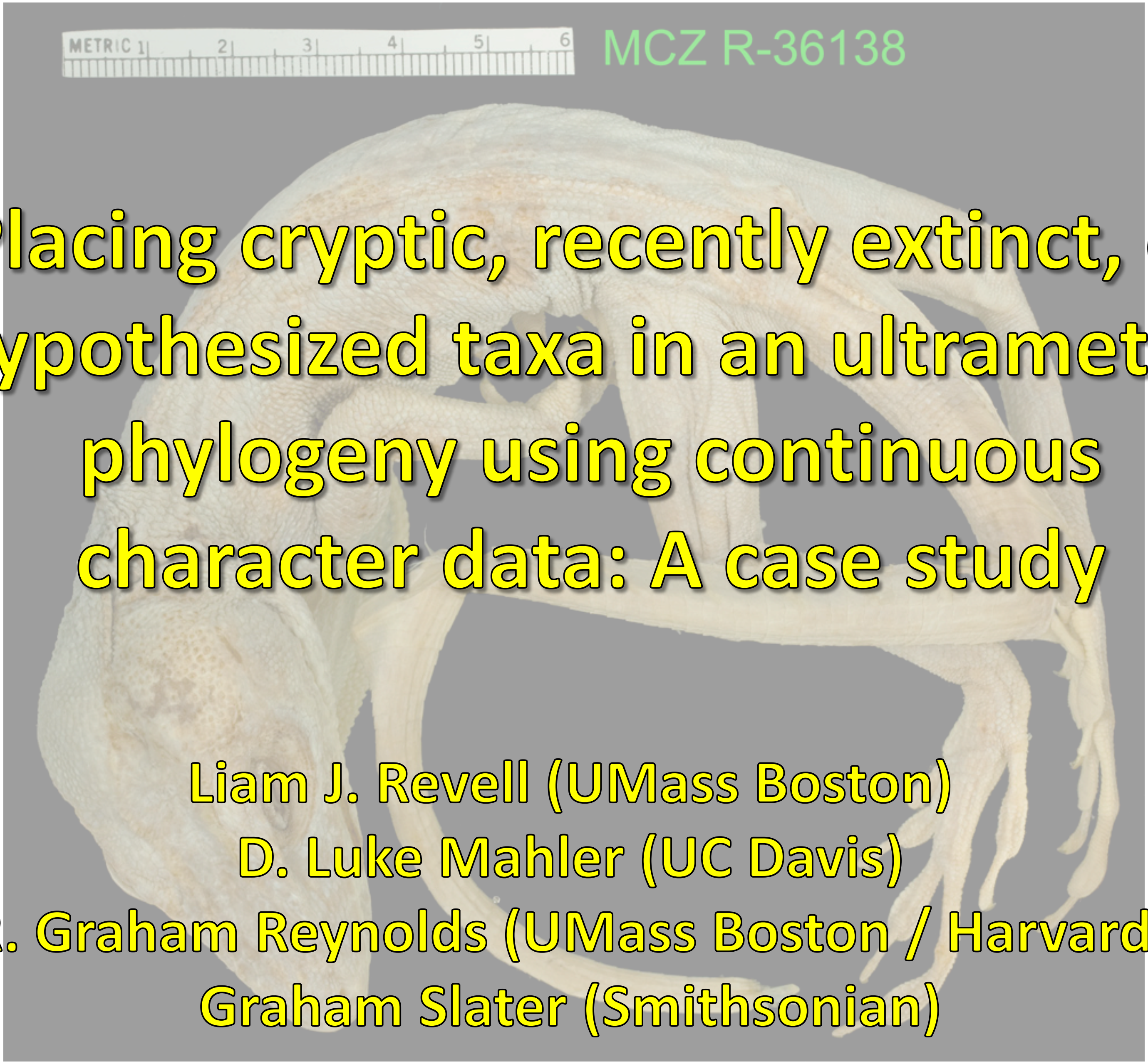


Placing cryptic, recently extinct, or hypothesized taxa in an ultrametric phylogeny using continuous character data: A case study

Liam J. Revell (UMass Boston)
D. Luke Mahler (UC Davis)
R. Graham Reynolds (UMass Boston / Harvard)
Graham Slater (Smithsonian)

The problem

- We have an ultrametric species tree (based on, say, DNA sequence data), and we want to add a single extant or recently extinct taxon to the phylogeny based on multivariable continuous trait data.
- Our missing taxon might be recently extinct (e.g., the thylacine), cryptic, hypothesized, or merely very difficult to obtain (but present in museum collections).

Placing taxon on a tree

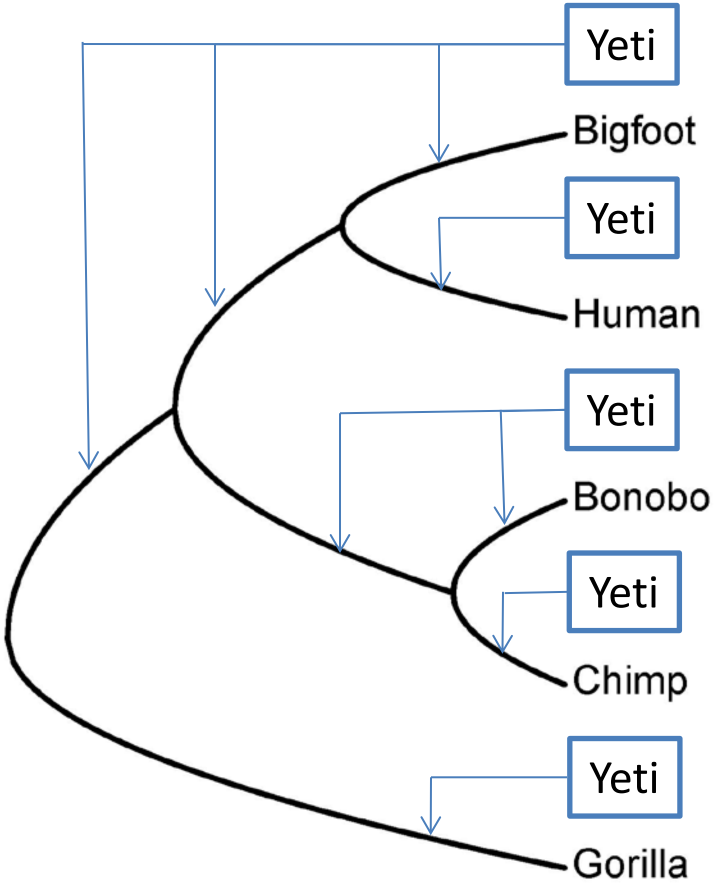
We need:

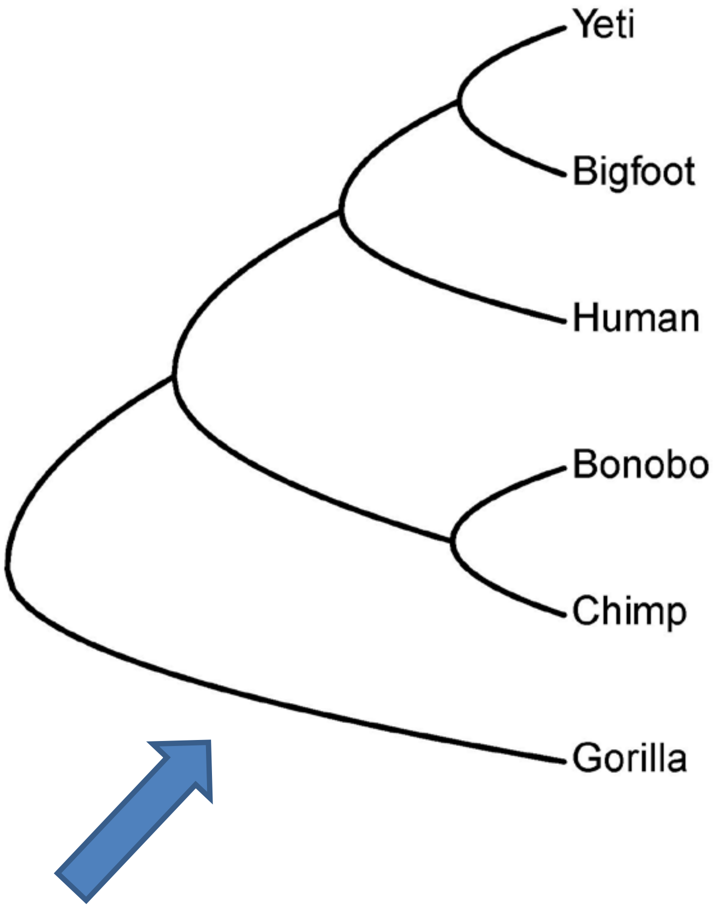
- Multivariable continuous character dataset of *N* species.
- Ultrametric base tree for *N* – 1 taxa.
- Single leaf to be added to the tree.

We use:

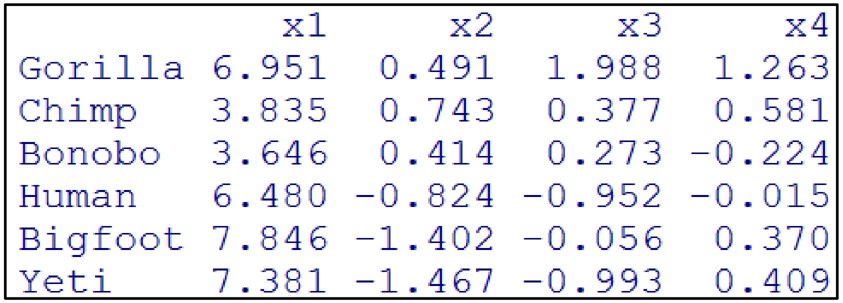
- Formulae of Felsenstein (1973, 1981) to compute the likelihood of hypothesized placements on the tree.
- Function to maximize the likelihood is called `locate.yeti`.

Graham Slater





L(model & tree|



```
             x1      x2      x3      x4
Gorilla  6.951   0.491   1.988   1.263
Chimp    3.835   0.743   0.377   0.581
Bonobo   3.646   0.414   0.273  -0.224
Human    6.480  -0.824  -0.952  -0.015
Bigfoot  7.846  -1.402  -0.056   0.370
Yeti     7.381  -1.467  -0.993   0.409
```

)*

* Felsenstein (1973, 1981)

What about…

… the covariances between characters?

- The problem with covariances among traits when the tree is unknown is that for each hypothesized topology & branch lengths we have a potentially different among-trait evolutionary covariance structure.
- To compute the likelihood of our tree & model from these data, we need to each time invert a covariance matrix of dimension N x m.
- A (barely) approximate solution when a *single* leaf is to be added is to use the principal components from phylogenetic PCA (or any evolutionary orthogonalization) using the $N – 1$ species in the base tree to rotate all the data before analysis; then just add the log(L) for each trait.

What about…

… finding the maximum likelihood placement?

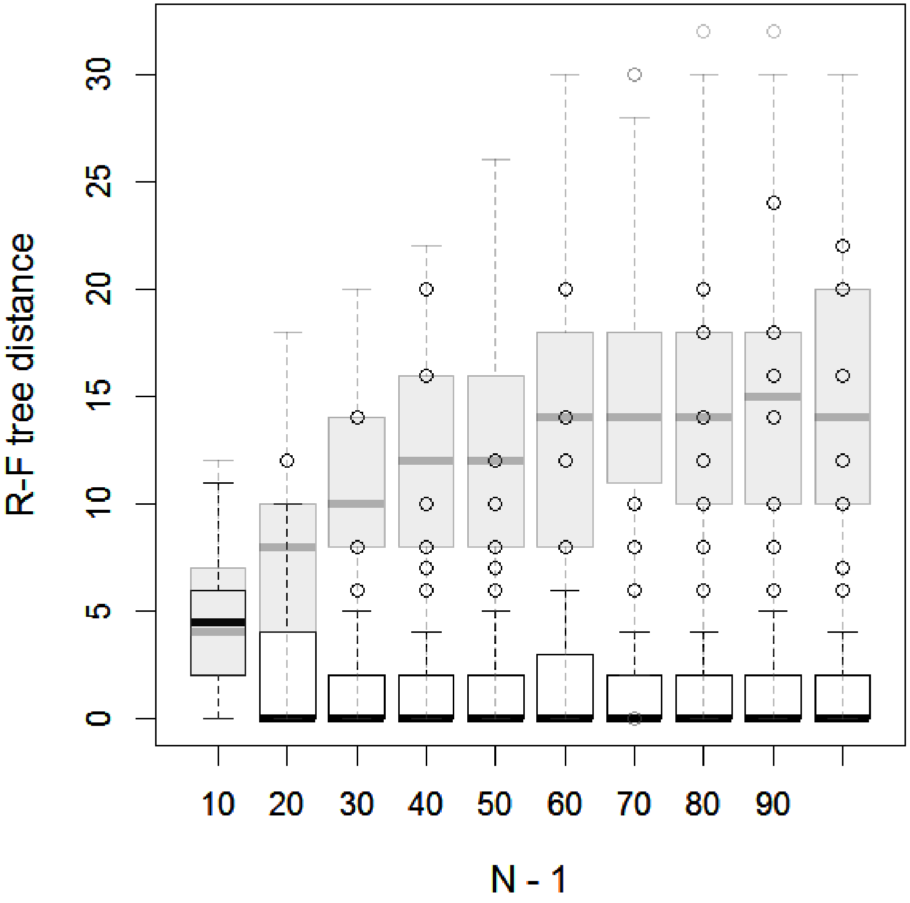
- We can use a *heuristic* approach in which we start by attaching the taxon to be added to each of the $2(N - 2)$ edges of the tree; and the optimize the location of the taxon on the edges with the highest likelihoods.
- However even for trees of 100 or more taxa an *exhaustive* search for the ML position of the tip taxon is often possible.
- Since our tree is ultrametric, the length of the new terminal edge is fully determined based on its location of attachment.

It works….

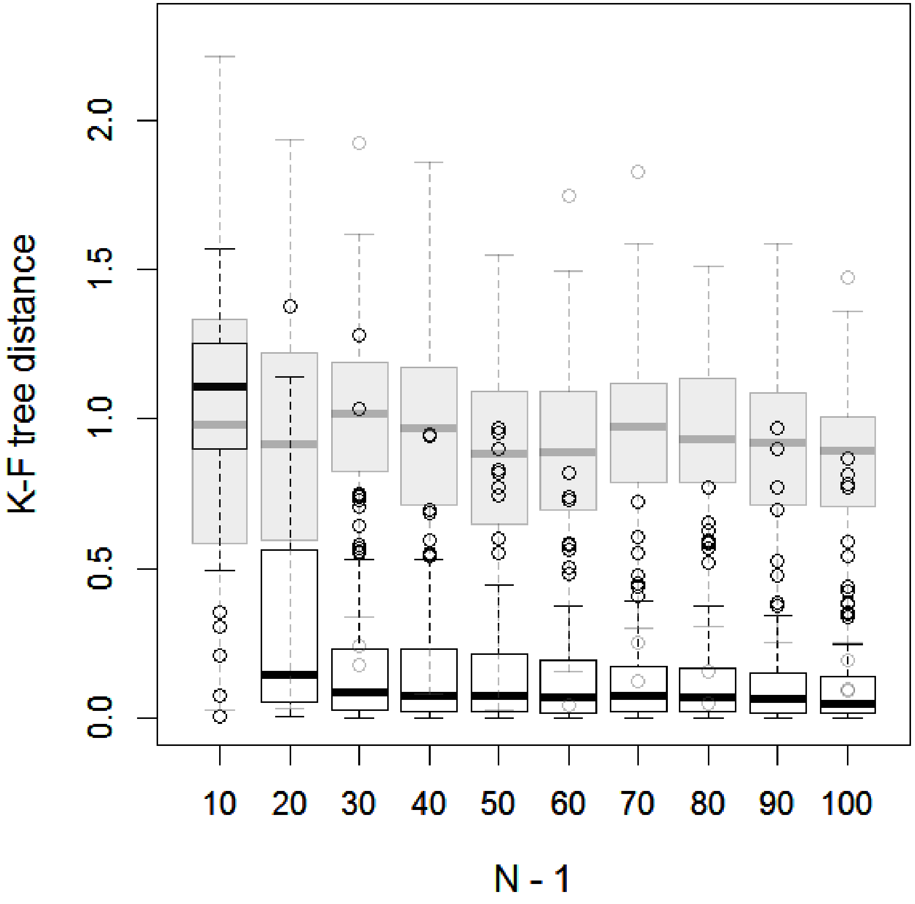
1) Simulate trees with *N* taxa (for various *N*).
2) Simulate *correlated* character evolution for *m* = 10 traits.
3) Trim one leaf at random.
4) Estimate the location of the leaf.
5) Compare to original tree using Robinson-Foulds distance (Robinson & Foulds 1981) and Kuhner & Felsenstein branch score (Kuhner & Felsenstein 1994).

It works….

Robinson-Foulds distance:



Branch-score distance:

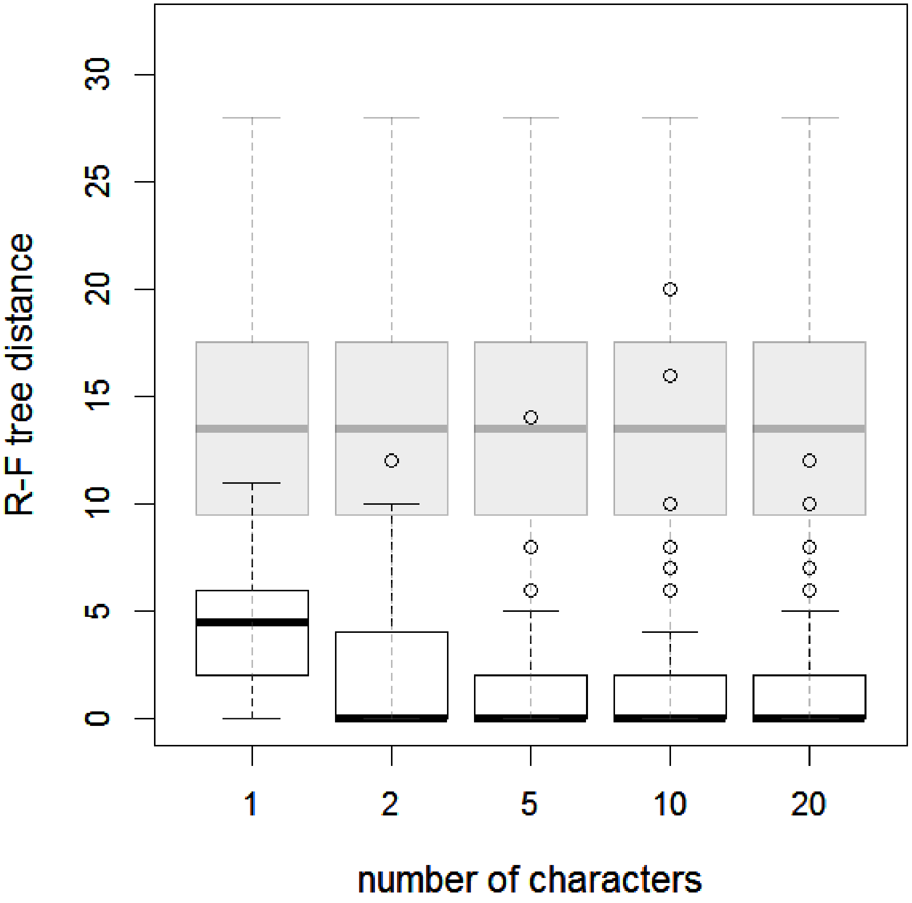


It works….

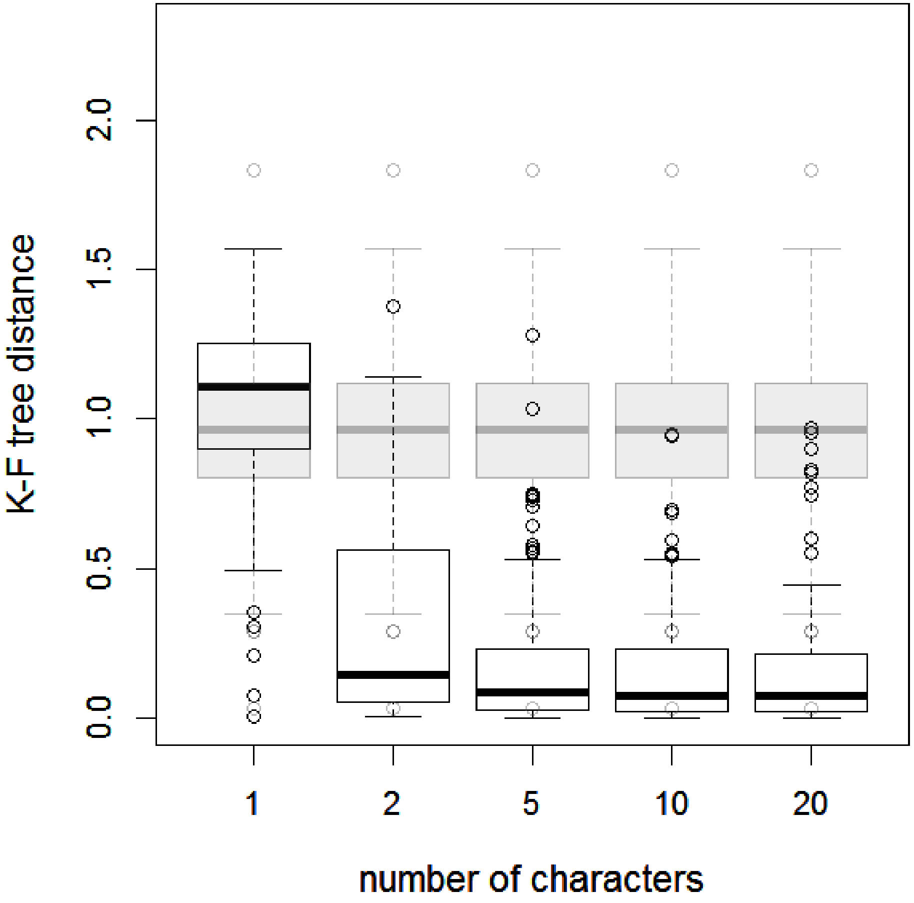
1) Simulate trees with *N* = 65 taxa.
2) Simulate *correlated* character evolution for *m* = 1, 2, 5, 10, & 20 traits.
3) Trim one leaf at random.
4) Estimate the location of the leaf.
5) Compare to original tree using Robinson-Foulds distance (Robinson & Foulds 1981) and Kuhner & Felsenstein branch score (Kuhner & Felsenstein 1994).

It works….

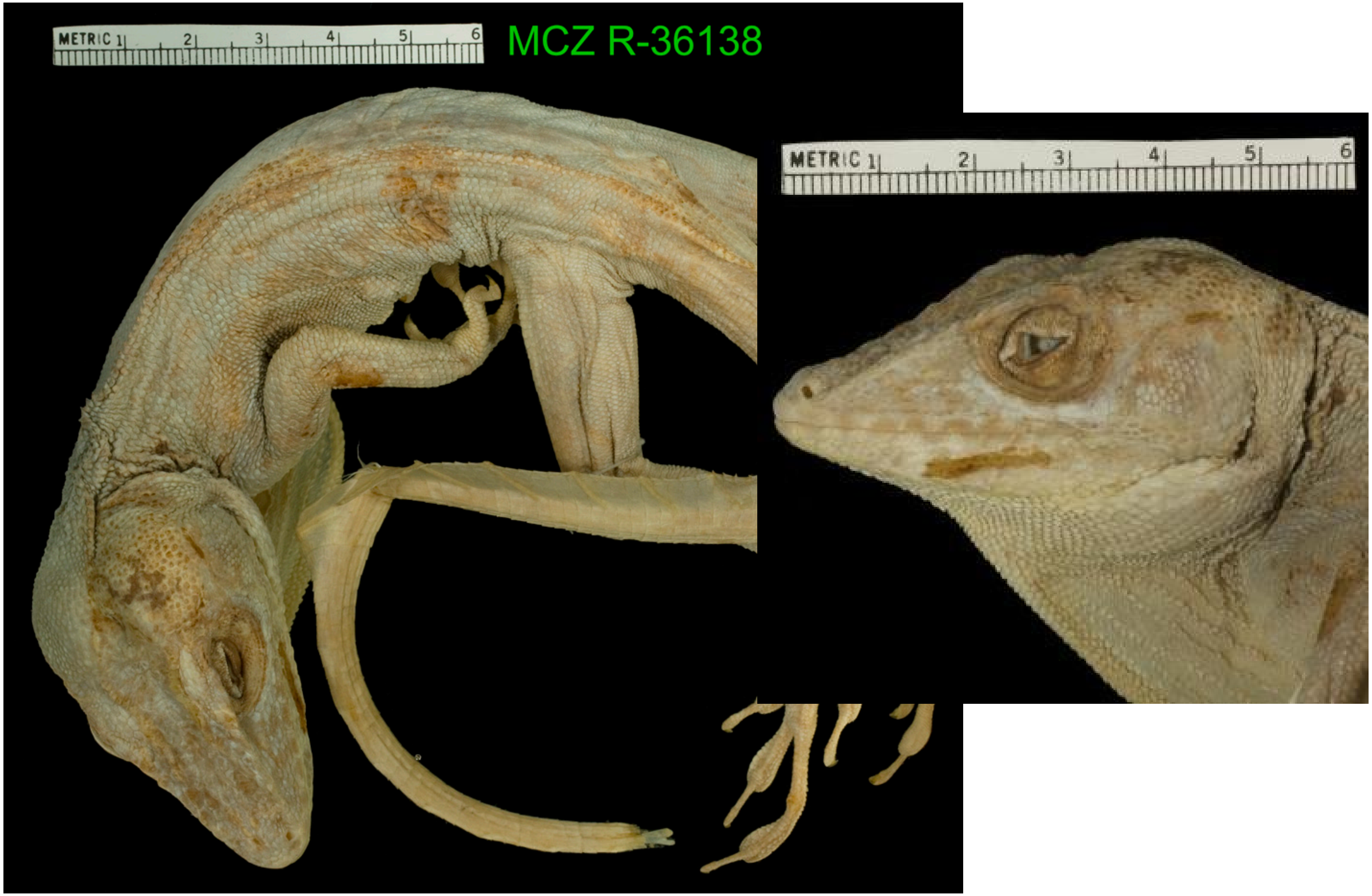
Robinson-Foulds distance:



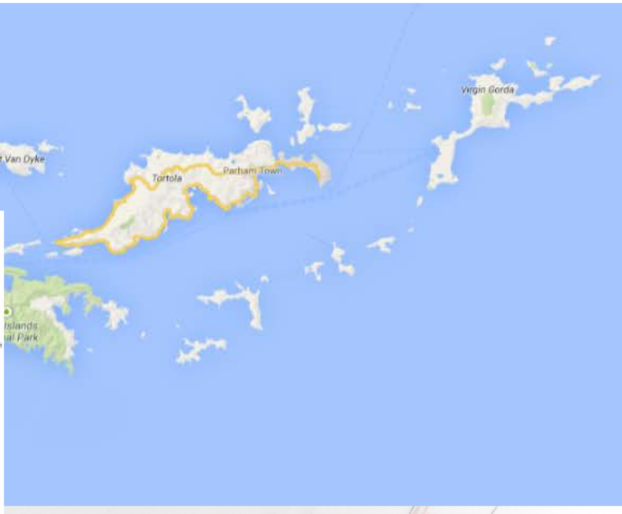
Branch-score distance:

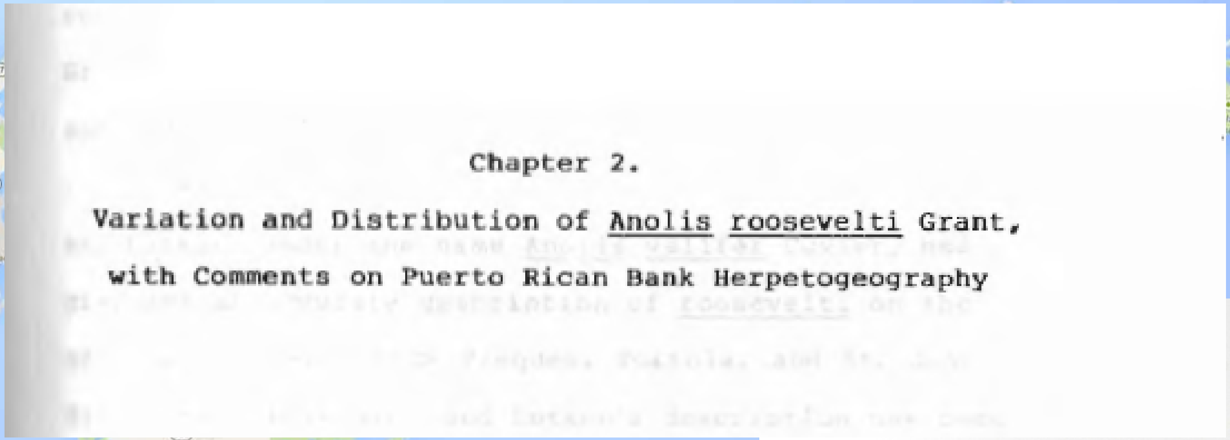


Case study: *Anolis roosevelti*



Anolis roosevelti – the Culebra Giant Anole

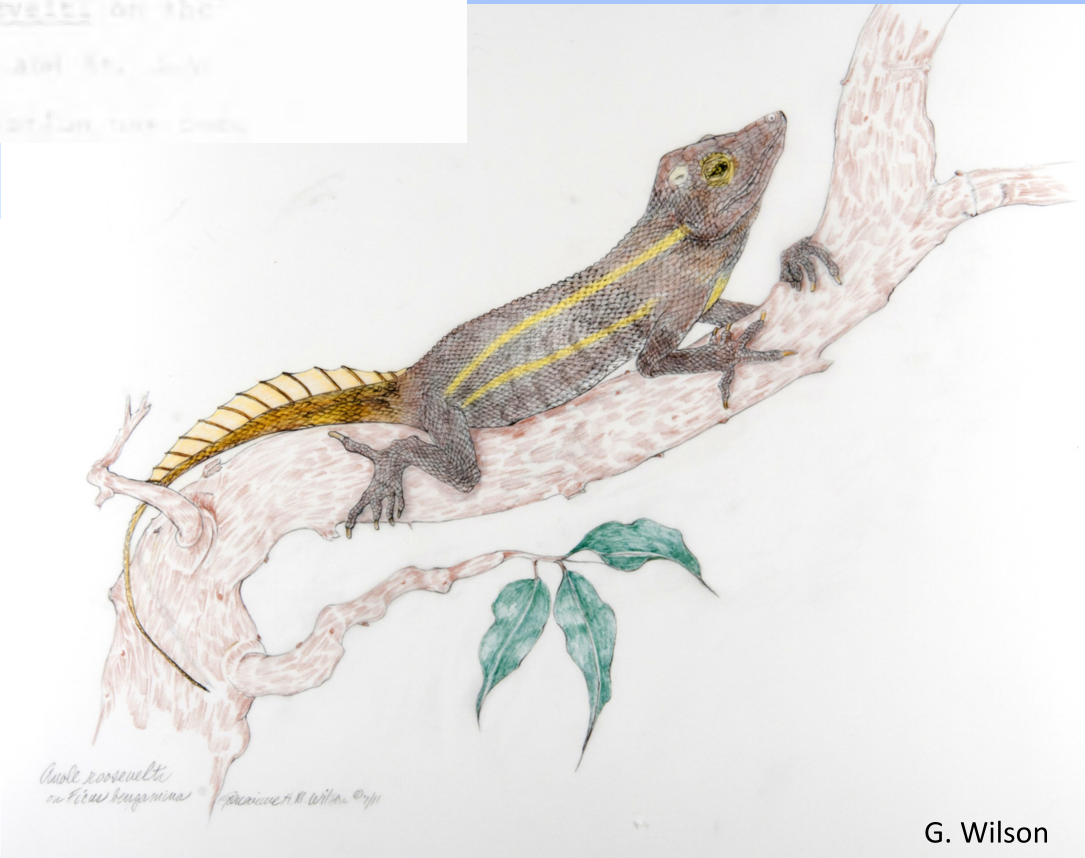




Chapter 2.

Variation and Distribution of Anolis roosevelti Grant, with Comments on Puerto Rican Bank Herpetogeography

- One of very few Caribbean anoles thought to be extinct.
- Called the 'Culebra Giant Anole' but now (Mayer 1989) thought to have been found throughout the PR-bank VIs.
- Probably extinct due to loss of habitat or loss of its preferred tree (ostensibly the 'Gumbo-limbo tree,' *Bursera simaruba*).



G. Wilson

J. Losos

Anolis cuvieri

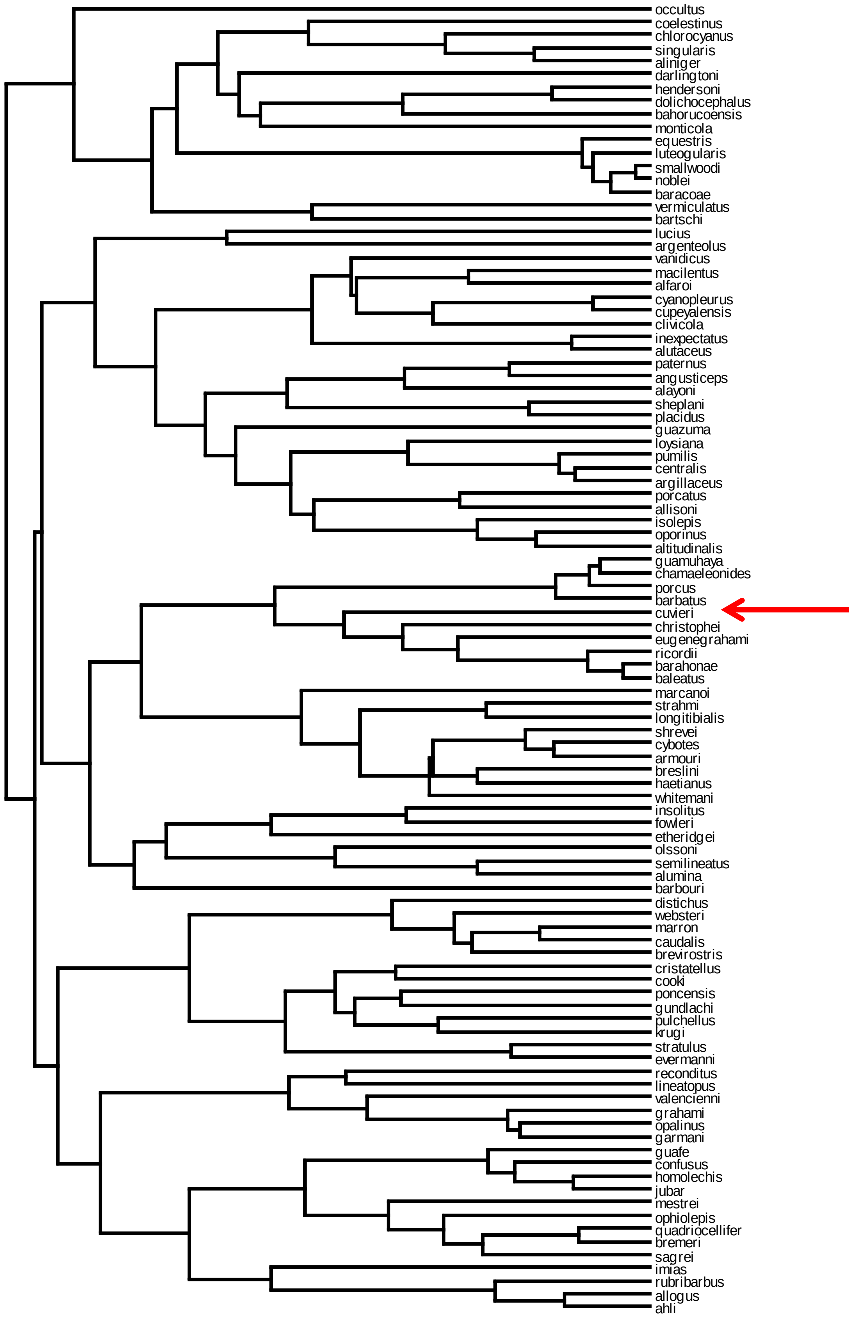
- Long been *assumed* (based on similarity in size, osteological features, and (presumed) ecology to be closely related to the Puerto Rican crown giant – *Anolis cuvieri*.
- However, the reality is we do not know it's phylogenetic relationship – and it's possible that *A. roosevelti* might be closely related to other PR anoles.

C. Smith

Anolis cristatellus

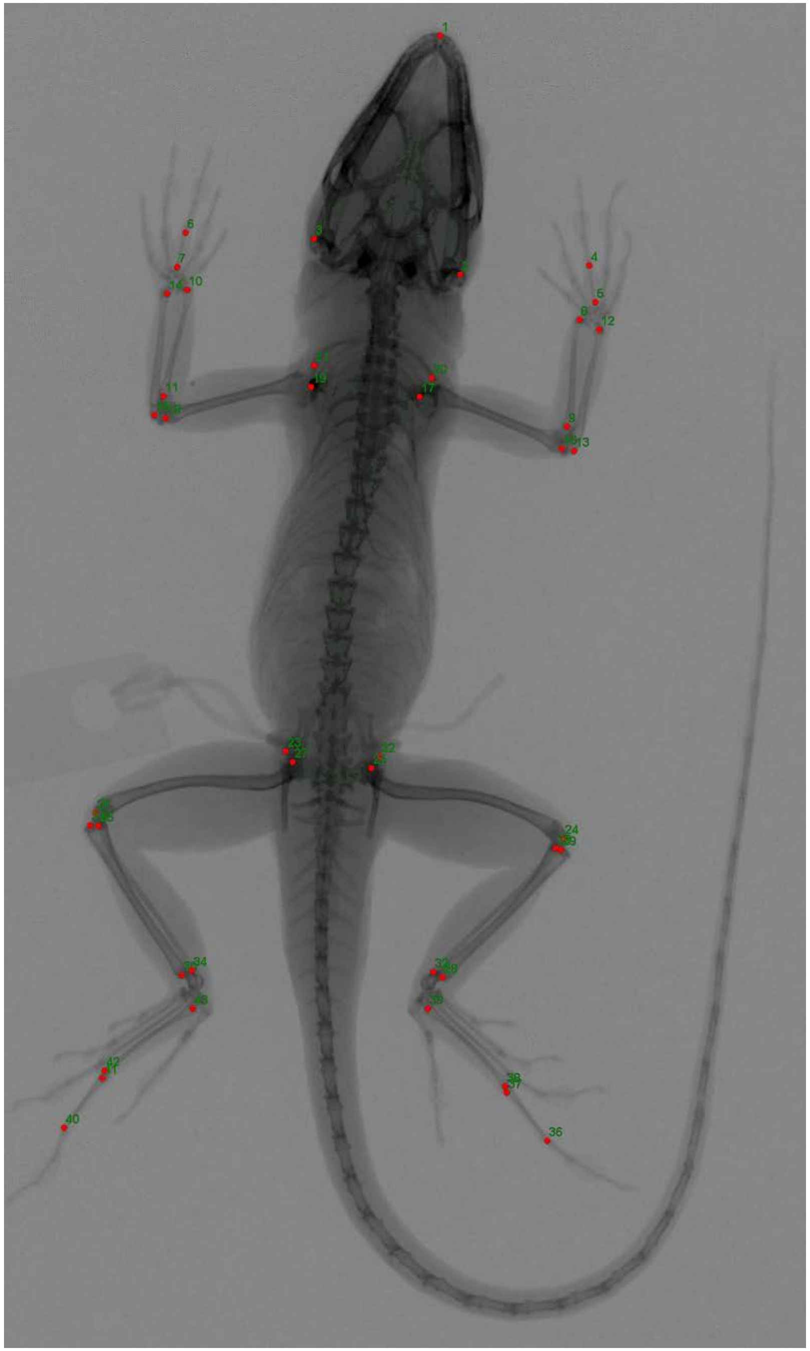
Our tree:

- Near comprehensive Greater Antillean anole phylogeny of Nicholson et al. (2005).
- 100 species, including non-ecomorphs, from all GA islands.



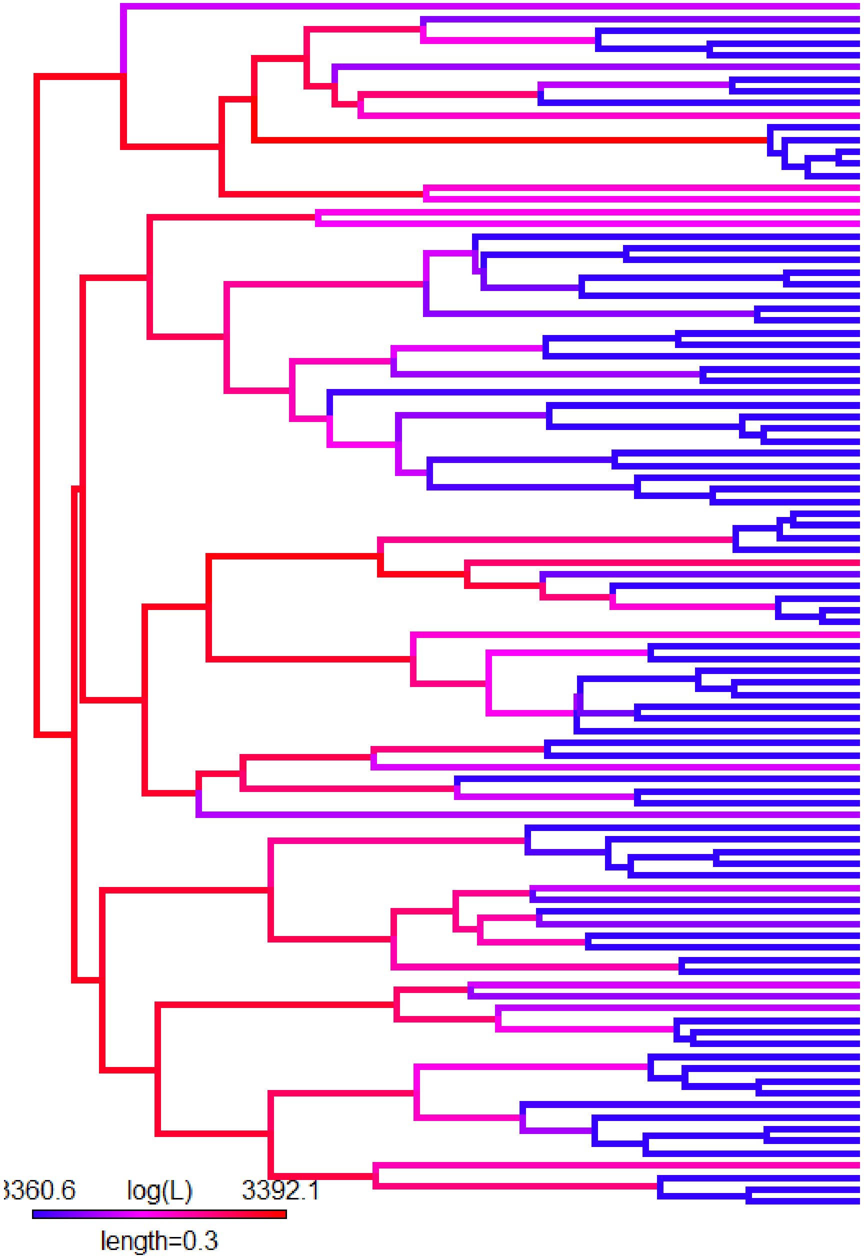
Our data:

- Morphological dataset of Mahler et al. (2010) containing 20 skeletal and external characteristics - averaged by species.
- SVL, head height, head length, head width, jaw length, outlever length, jugal to symphisis length, femur length, tibia length, metatarsal length, …, lamellae numbers, etc.



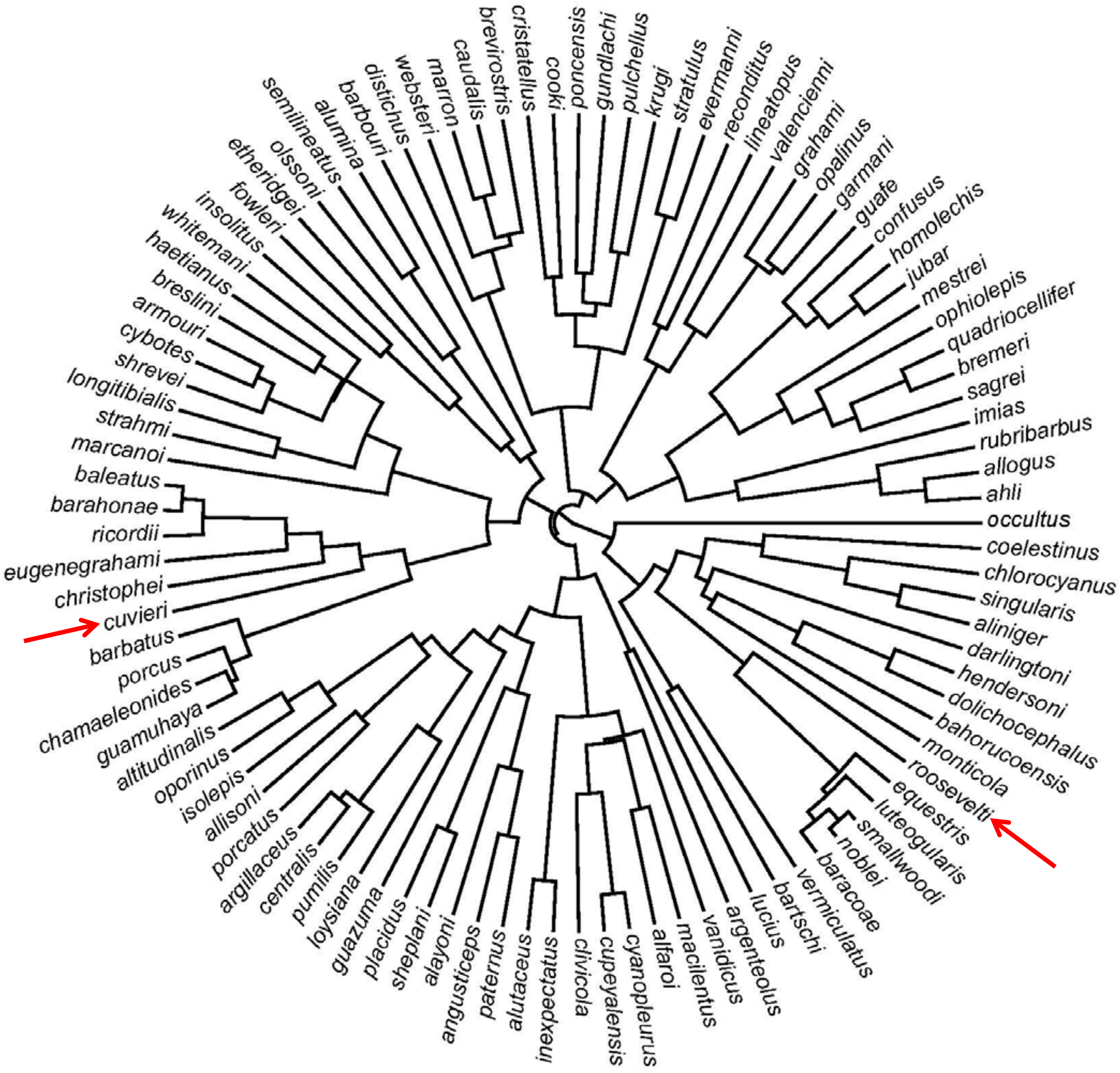
Results:

- Likelihood plot shows that many parts of the tree can be strongly *rejected*; however the surface is very flat towards the root for this dataset.
- This is unsurprising because morphologically similar species have evolved numerous times in different parts of the tree.

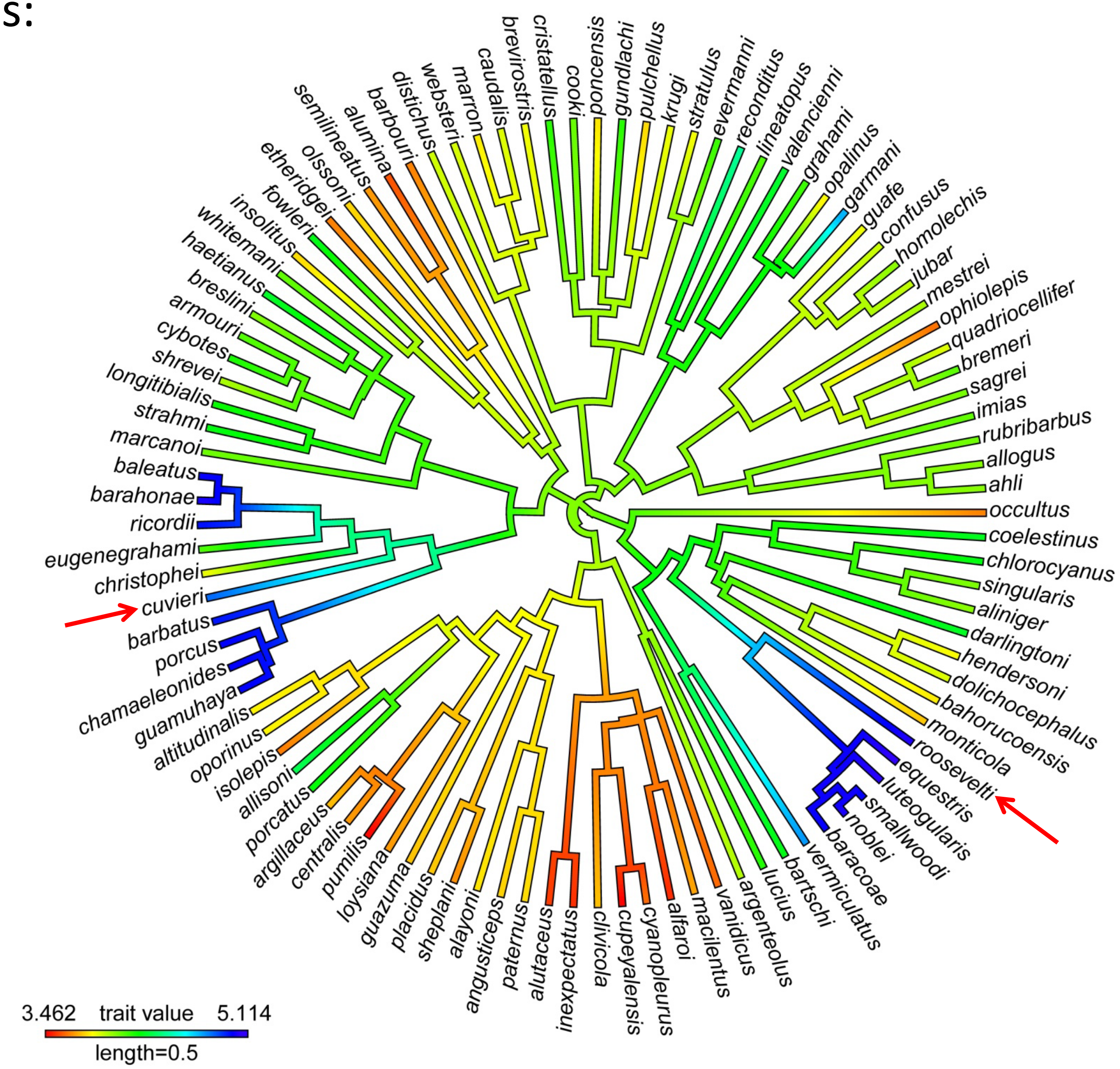


Note: logL <= 3360.63 set to 3360.63 for visualization only

Results:


occultus
coelestinus
chlorocyanus
singularis
aliniger
darlingtoni
hendersoni
dolichocephalus
bahorucoensis
monticola
roosevelti
equestris
luteogularis
smallwoodi
noblei
baracoae
vermiculatus
bartschi
lucius
argenteolus
vanidicus
macilentus
alfaroi
cyanopleurus
cupeyalensis
clivicola
inexpectatus
alutaceus
paternus
angusticeps
alayoni
sheplani
placidus
guazuma
loysiana
pumilis
centralis
argillaceus
porcatus
allisoni
isolepis
oporinus
altitudinalis
guamuhaya
chamaeleonides
porcus
barbatus
cuvieri
christophei
eugenegrahami
ricordii
barahonae
baleatus
marcanoi
strahmi
longitibialis
shrevei
cybotes
armouri
breslini
haetianus
whitemani
insolitus
fowleri
etheridgei
olssoni
semilineatus
alumina
barbouri
distichus
websteri
marron
caudalis
brevirostris
cristatellus
cooki
poncensis
gundlachi
pulchellus
krugi
stratulus
evermanni
reconditus
lineatopus
valencienni
grahami
opalinus
garmani
guafe
confusus
homolechis
jubar
mestrei
ophiolepis
quadriocellifer
bremeri
sagrei
imias
rubribarbus
allogus
ahli

Results:


semilineatus
alumina
barbouri
distichus
websteri
marron
caudalis
brevirostris
cristatellus
cooki
poncensis
gundlachi
pulchellus
krugi
stratulus
evermanni
reconditus
lineatopus
valencienni
grahami
opalinus
garmani
guafe
confusus
homolechis
jubar
mestrei
ophiolepis
quadriocellifer
bremeri
sagrei
imias
rubribarbus
allogus
ahli
occultus
coelestinus
chlorocyanus
singularis
aliniger
darlingtoni
hendersoni
dolichocephalus
bahorucoensis
monticola
roosevelti
equestris
luteogularis
smallwoodi
noblei
baracoae
vermiculatus
bartschi
lucius
argenteolus
vanidicus
macilentus
alfaroi
cyanopleurus
cupeyalensis
clivicola
inexpectatus
alutaceus
paternus
angusticeps
alayoni
sheplani
placidus
guazuma
loysiana
pumilis
centralis
argillaceus
porcatus
allisoni
isolepis
oporinus
altitudinalis
guamuhaya
chamaeleonides
porcus
barbatus
cuvieri
christophei
eugenegrahami
ricordii
barahonae
baleatus
marcanoi
strahmi
longitibialis
shrevei
cybotes
armouri
breslini
haetianus
whitemani
insolitus
fowleri
etheridgei
olssoni
3.462
trait value
5.114
length=0.5

Can we reject alternative phylogenetic positions for *A. roosevelti*?

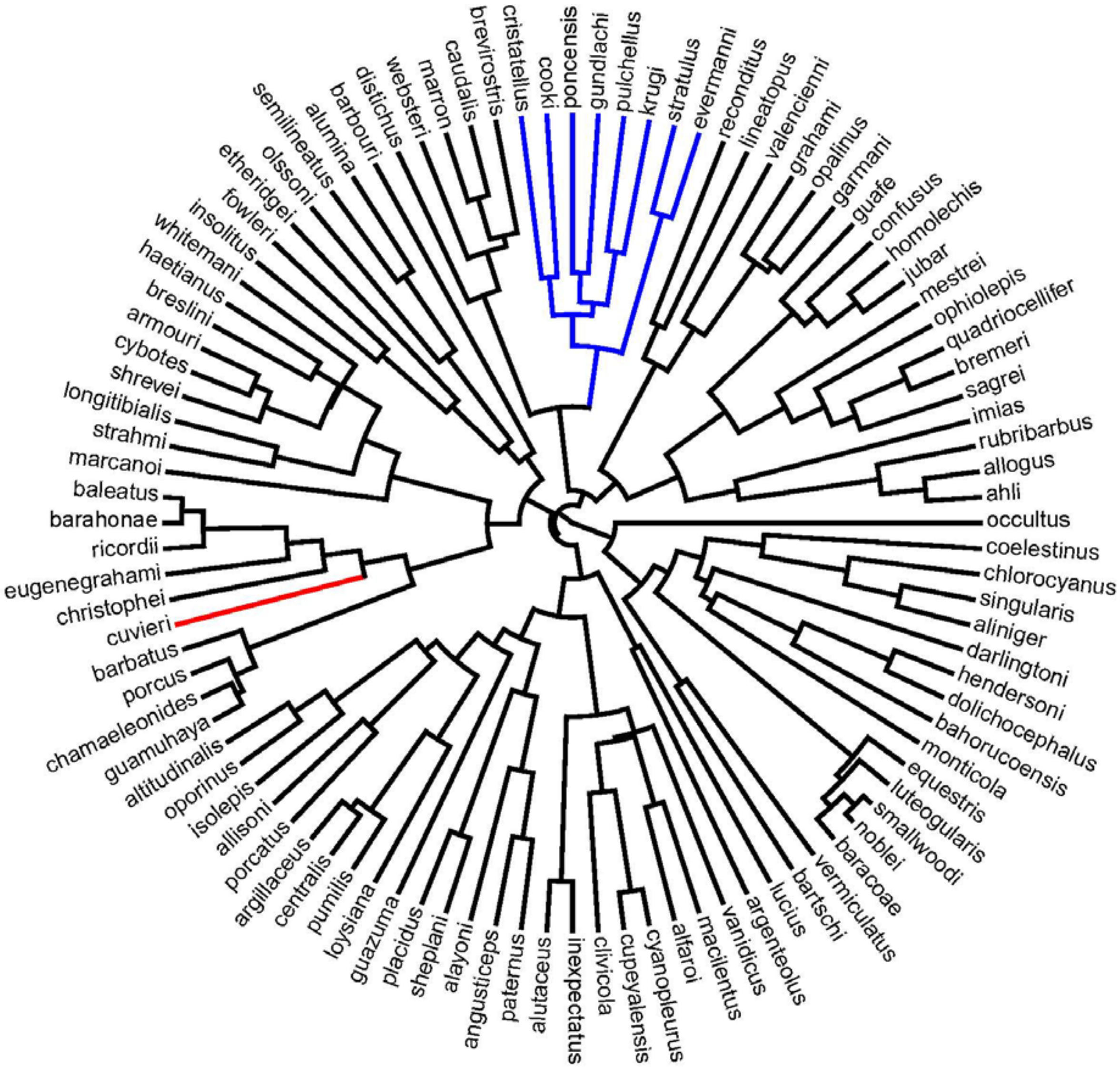
- To answer this, we can simply constrain to alternative hypotheses for the phylogenetic position of our new leaf & compare the likelihood to the unconstrained model.
- For instance, we can constrain *A. roosevelti* to be closely related to different PR species.

A. Sanchez

Anolis evermanni

Hypothesis 1: *Anolis roosevelti* is sister to *A. cuvieri*.

Hypothesis 2: *Anolis roosevelti* is sister or nested within the rest of the PR anoles (excluding *A. occultus*).



What about…

… the number of degrees of freedom consumed by topological constraint?

- If imposing a topological constraint on the tree consumed a specific number of degrees of freedom then we could simply compare our LR to a X^2 with the requisite *d.f*.
- Unfortunately, it is not obvious how many *d.f.* are consumed by topologically constraining the location of the leaf to be added.
- Thus, to obtain a null distribution of the LR, we simulated data on the ML constraint tree, and then estimated both with & without constraint.

Result:

Table. We *cannot* reject placement of *A. roosevelti* as sister to *A. cuvieri* (P = 0.26); however we *can* strongly reject placement with other PR anoles.

Hypothesis	log (L)	LR	P*
H_0: Unconstrained	3392.2	0.000	1.00
H_1: Sister to *A. cuvieri*	3389.6	5.086	0.26
H_2: Sister or nested within other PR anoles.	3387.2	9.955	0.02

*P-value based on simulation

Conclusions

- Method to place recently extinct, cryptic, or hypothesized taxa into an ultrametric tree (`locate.yeti`) works well for simulated trees & data.
- With the *Anolis roosevelti* data the ML placement is *not* with *A. cuvieri* or other PR anole species (where it almost certainly belongs!).
- Method might show better empirical performance in a dataset (or with characteristics) that other studies have not shown to be under such strong selection for convergence!

Future possibilities…

… measurement error / uncertainty in the estimation of species means?

- For the situation in which the uncertainty of the species mean is *known*, this is straightforward to take into account.
- In this case, the variance between species is merely the sum of the evolutionary variance and the variance in the estimates (following Ives et al. 2007).

Future possibilities…

… Bayesian MCMC version of the method?

- This would permit both the use of *prior proabilities* (in lieu of hard constraint) for the phylogenetic position of the unknown leaf…
- … along with an expression of the uncertainty of the placement of a leaf in terms of *posterior probability* that the leaf is connected to each edge in the base tree.

Future possibilities…

… exact REML method (instead of approximate ML method)?

- Finding the REML position (instead of the ML position) of the missing leaf allows us to take advantage of the contrasts algorithm and thus should permit computation of the exact restricted likelihood.
- This is because we don’t have to assume the ML among-trait covariance matrix for the base tree. (We can compute the REML covariance matrix for each proposed leaf placement.)

Future possibilities…

… placing fossils in trees using quantitative characters?

- This approach is closely related to that proposed by Felsenstein (2002).
- In fact, if we merely allow one additional parameter to be optimized (the terminal edge length, with constraint) we have a fossil method.