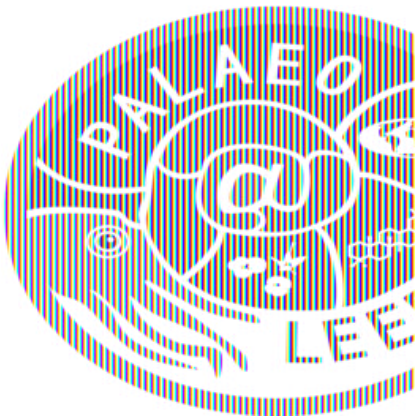
# Sampling bias in the fossil record

Bethany J. Allen | Alex M. Dunhill | Graeme T. Lloyd



# Workshop structure

- Alex talk (video | slides)
- This talk (video | slides)
- Scripts (hands on)
- Ask questions (best to do other bits first!)
  - Drop-in Zoom, Friday 12th June 2pm-3:30pm BST, check your emails for the link
  - Use the ‘#workshop’ channel on Discord
  - Email the organisers

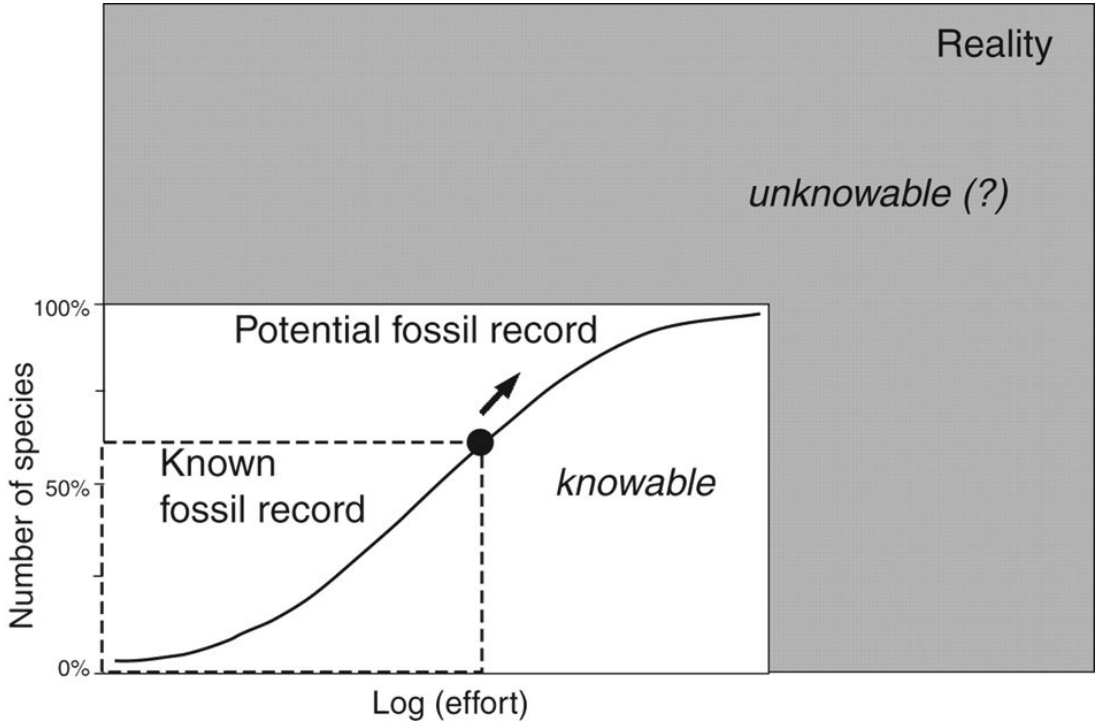
# Outline for this talk

Core concepts

Big data

Methods

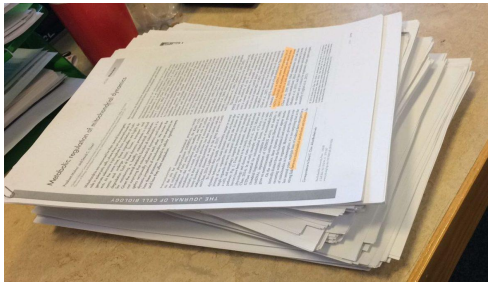
# The fossil record is an incomplete record

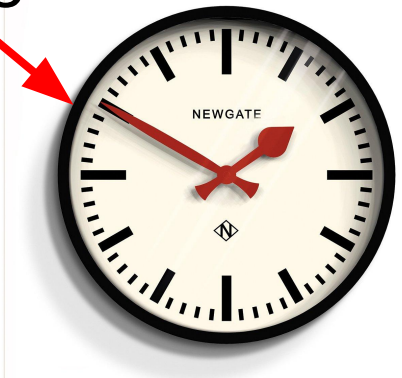



Benton *et al.* 2011; In *Comparing the Geological and Fossil Records: Implications for Biodiversity*

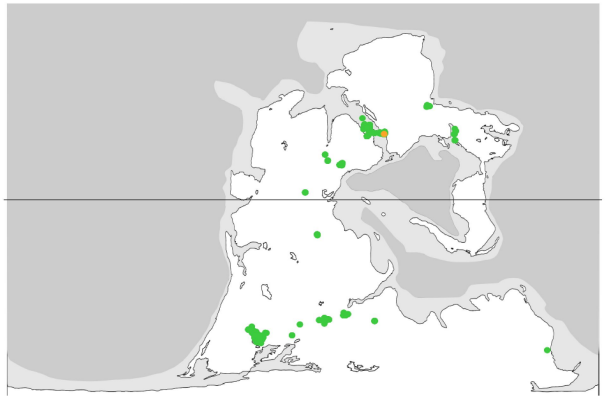
# Sampling has many facets

Sampling

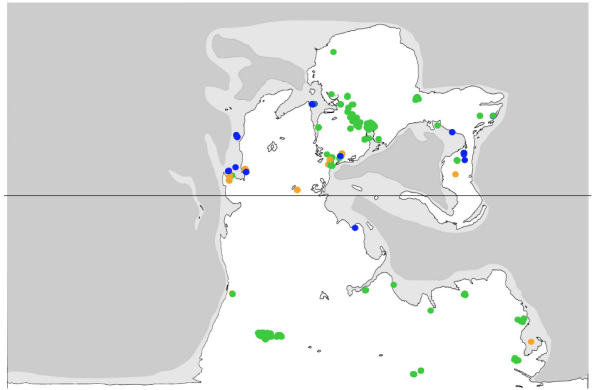




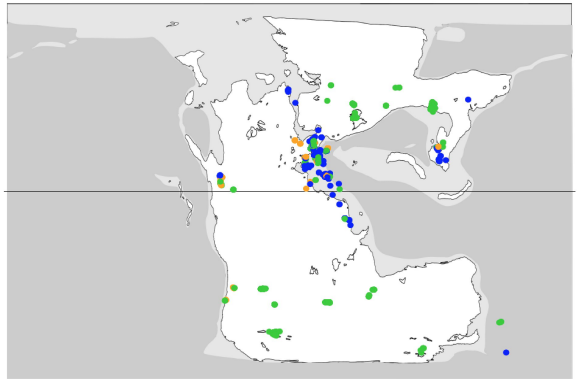
# Sampling varies spatially and temporally



Late Permian



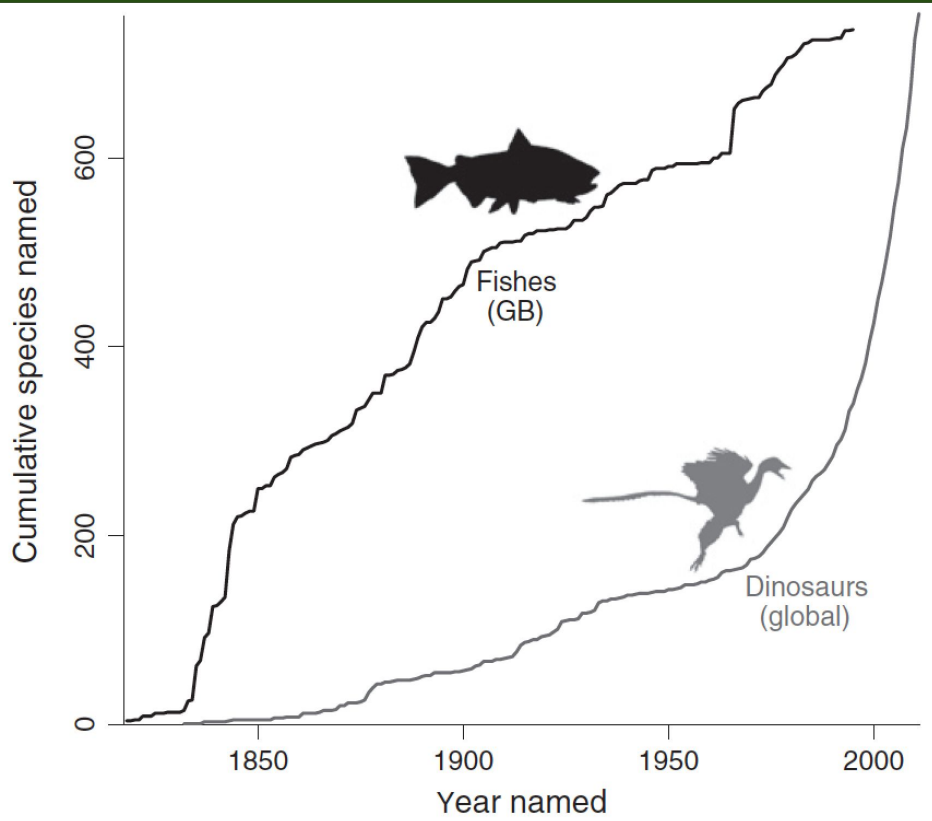
Early Triassic



Middle Triassic

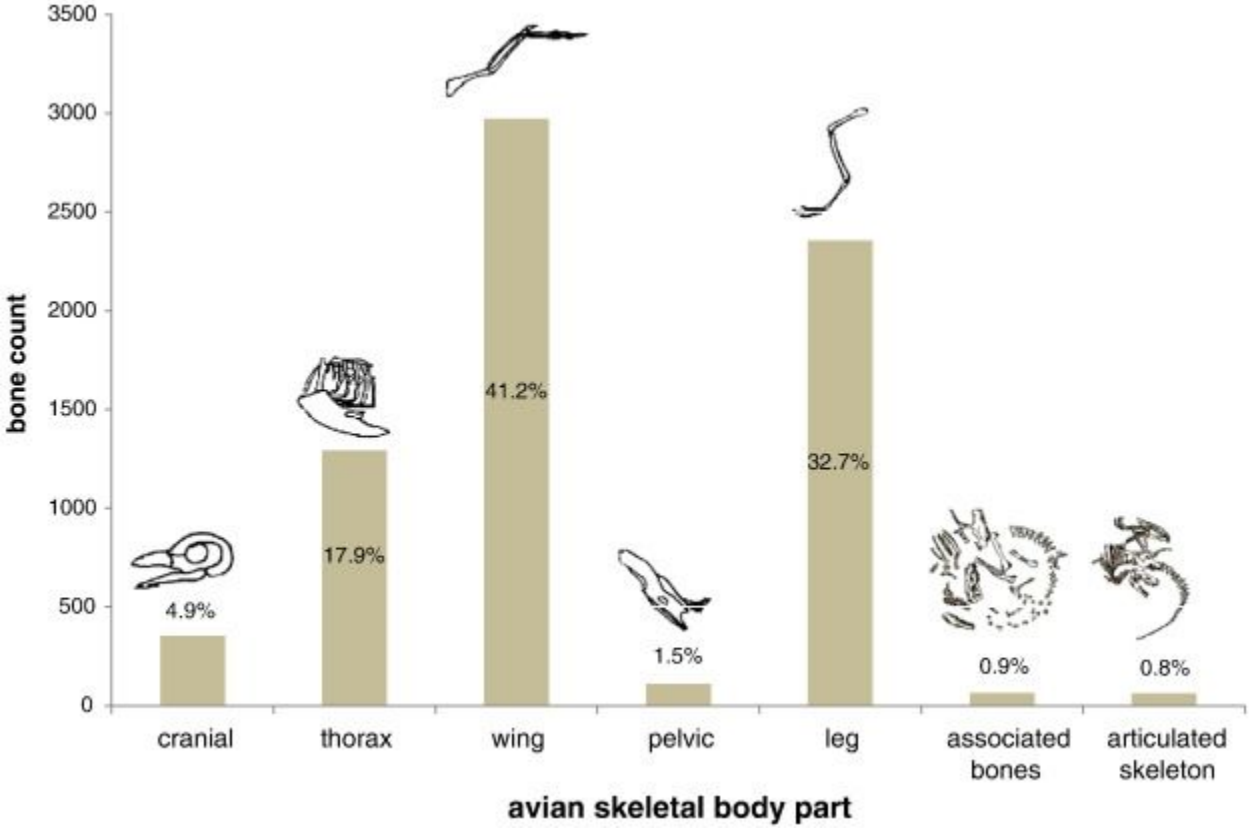
Allen *et al.* in press; *Proc. Roy. Soc. B*

# Sampling varies by clade




Lloyd and Friedman 2013; *Palaeo3*

# Sampling varies anatomically




Gardner et al. 2016; *Palaeo3*

# Sampling as explanatory variable: patient zero

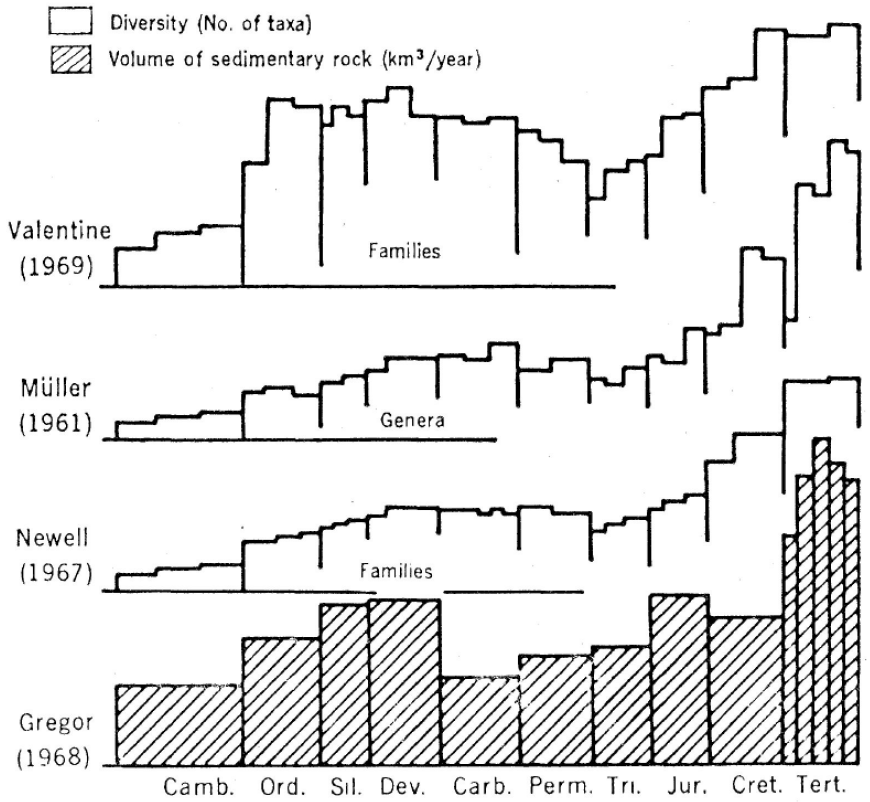
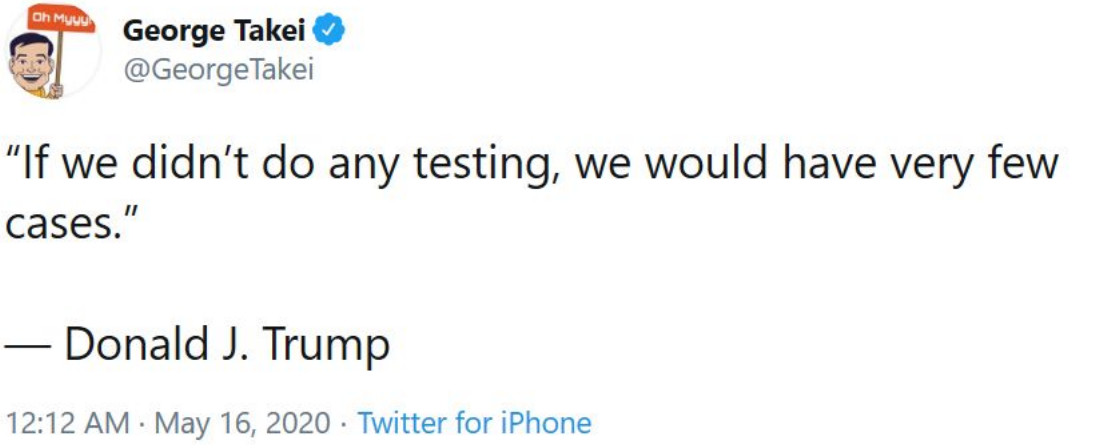


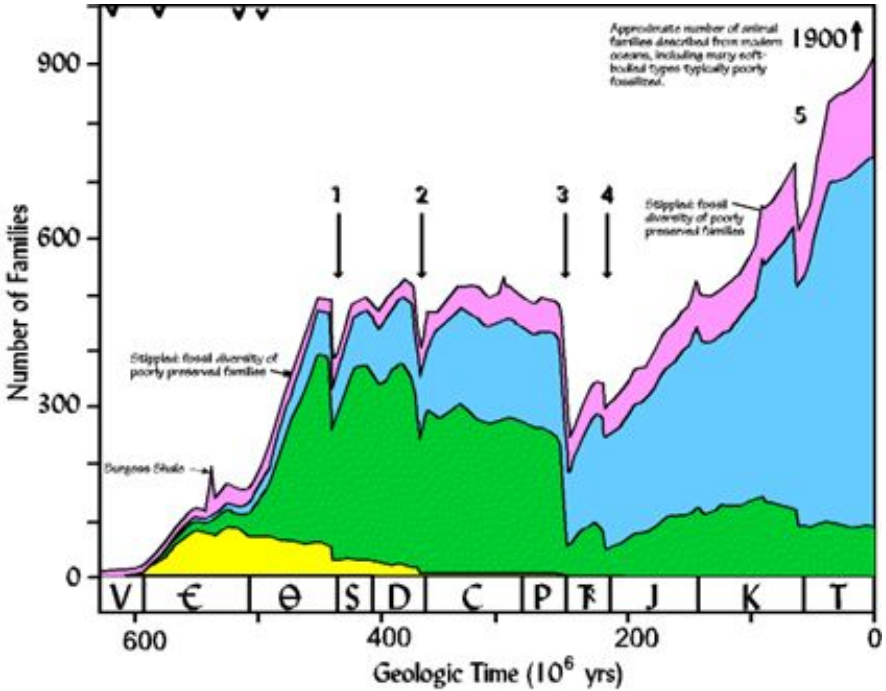

Fig. 1. Comparison of the number of taxa and the volume of sedimentary rock during the Phanerozoic. The diversity data are based mainly on well-skeletonized marine invertebrates (*1*, *8*, *9*, *12*).

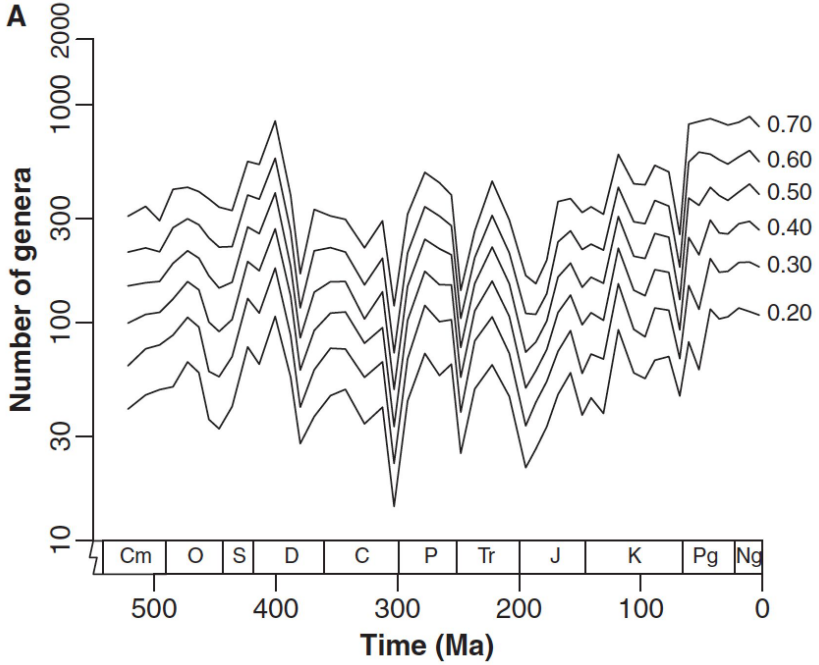
Raup 1972; *Science*

# Sampling can drive pattern



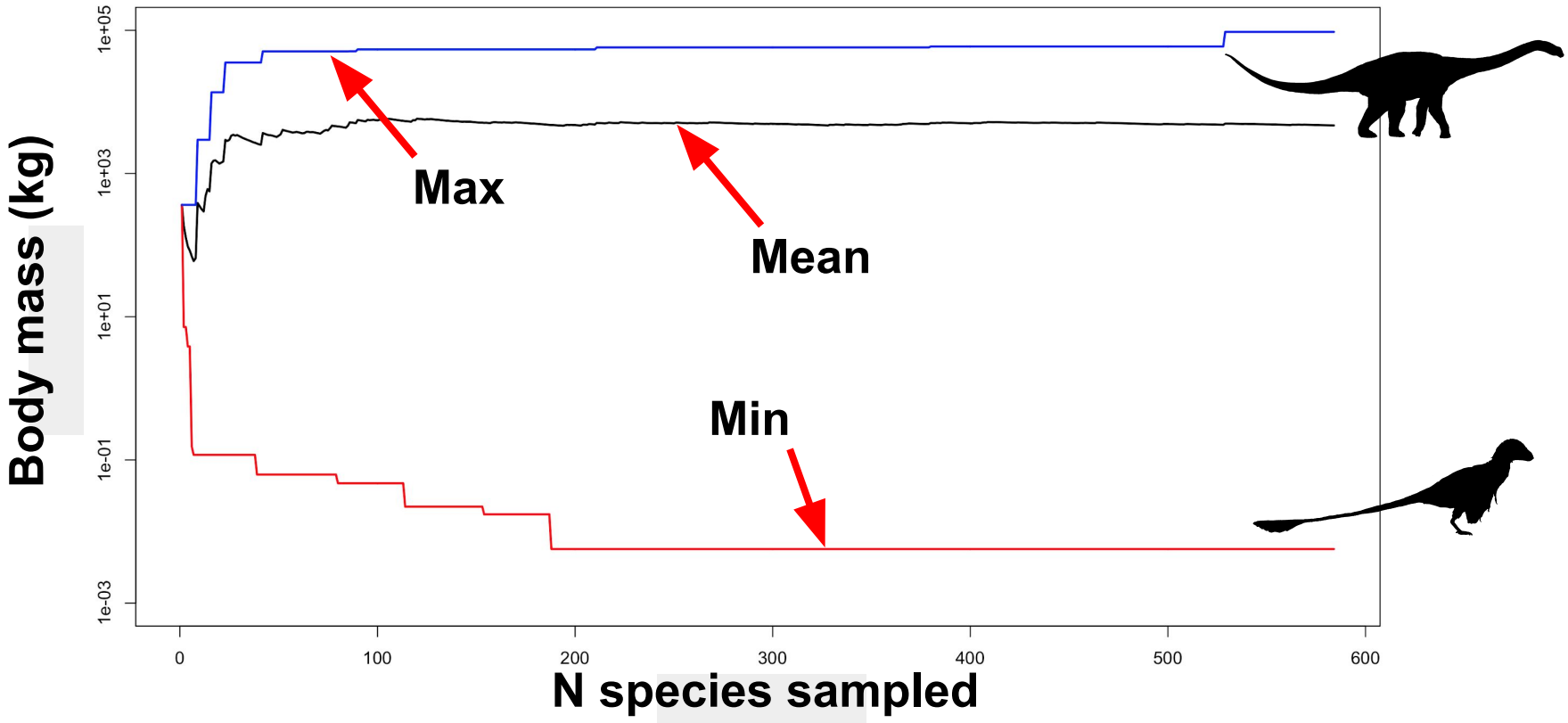
# Removing sampling bias




Sepkoski 1984, Alroy 2010; *Palaeontology*

# The monotony of it all




**Data from:** Benson *et al.* 2018; *Palaeontology*

# The monotony of it all [code]

```
# Code to produce plot on previous plot

# Grab the Benson dinosaur body mass data from GitHub:
DinoBodyMasses <-
read.table("https://raw.githubusercontent.com/bethany-j-allen/sampling_bias_workshop/master/Data/DinoBodyMasses.txt?token=ABZPFHSQJZOGHAKWFKTP73K6ZUEPA", header = TRUE)

# Shuffle the order to randomly resample the data:
x <- sample(DinoBodyMasses[, "Mass_kg"])

# Set random seed (so you will get exactly same plot):
set.seed(9)

# Resample data as minimum, mean and maximum:
Samples <- do.call(rbind, lapply(as.list(1:length(x)), function(y) {z <- x[1:y]; c(min(z), mean(z), max(z))}))

# Add column names to output:
colnames(Samples) <- c("MinMass", "MeanMass", "MaxMass")

# Make plot:
plot(x = 1:length(x), y = Samples[, "MeanMass"], log = "y", xlab = "N species sampled", ylab = "Mass (kg)", type = "l", ylim = c(0.001, 100000), lwd = 2)
points(x = 1:length(x), y = Samples[, "MinMass"], type = "l", col = "red", lwd = 2)
points(x = 1:length(x), y = Samples[, "MaxMass"], type = "l", col = "blue", lwd = 2)
```

# Big data

The Paleobiology Database
revealing the history of life

NEOTOMA PALEOECOLOGY DATABASE

GBIF Global Biodiversity Information Facility

Neptune SANDBOX Berlin

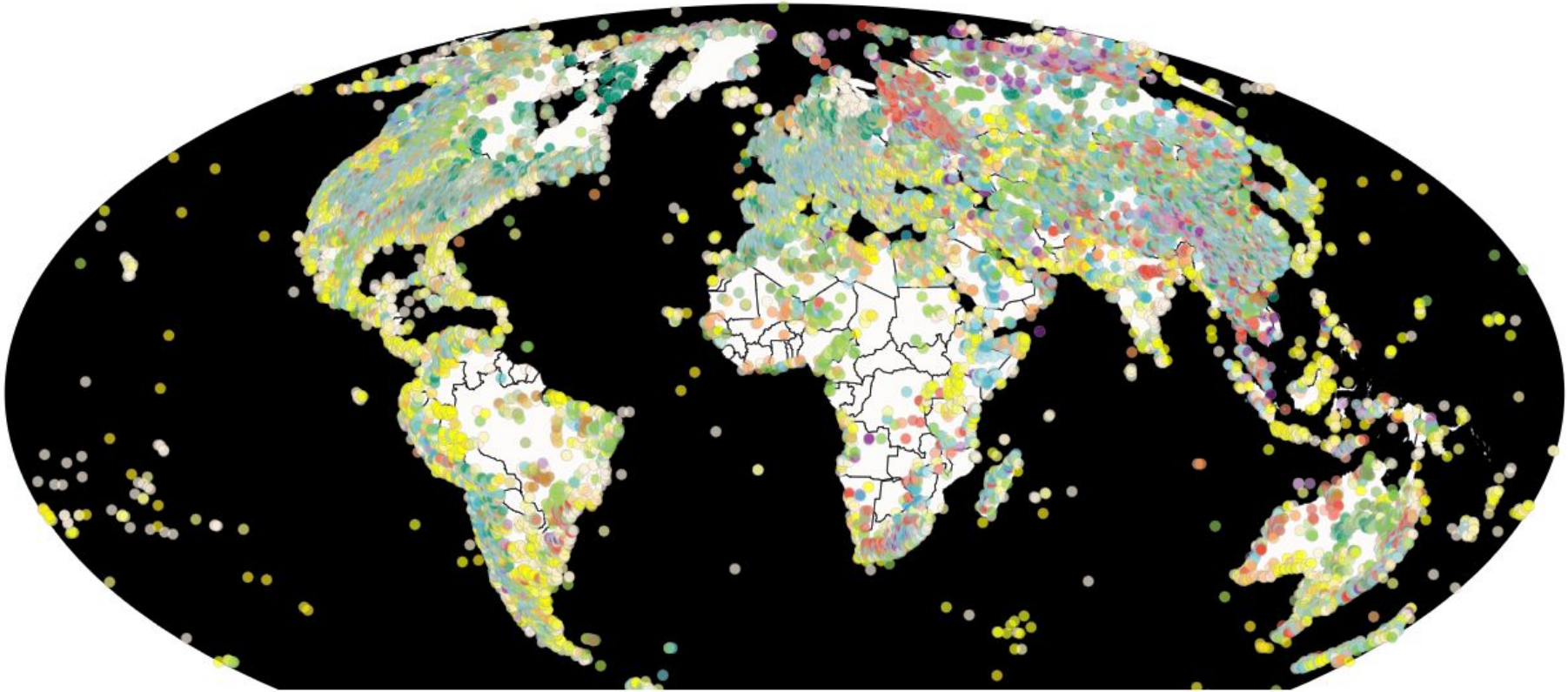
iDigBio
Integrated Digitized Biocollections

GBDB Geobiodiversity Database
Promoting collaborations based on global and regional database

# Big data

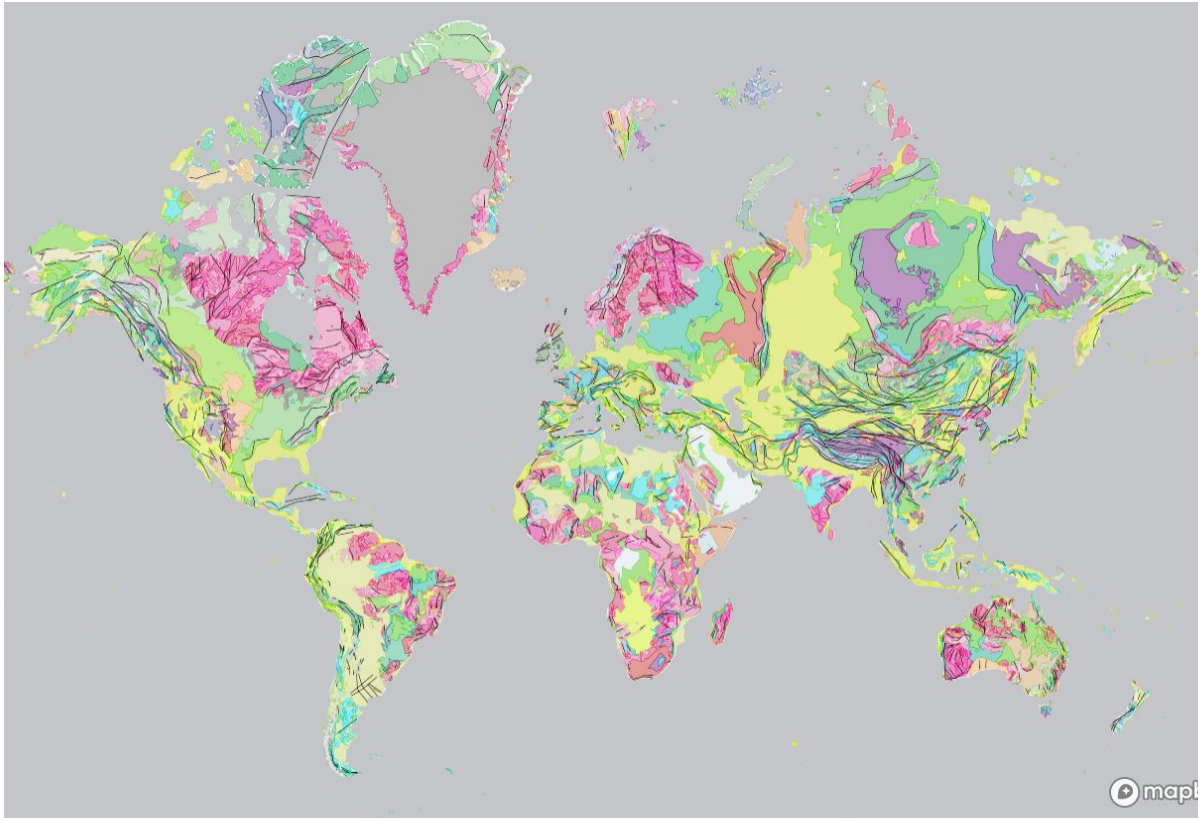

PaleobioGoogling
@PaleoGoogling
what happens when the entire fossil record is in the @PaleoDB?
3:13 PM · Jan 15, 2020 · Twitter Web App
3 Retweets
31 Likes

# Paleobiology Database



paleobiodb.org

# Macrostrat

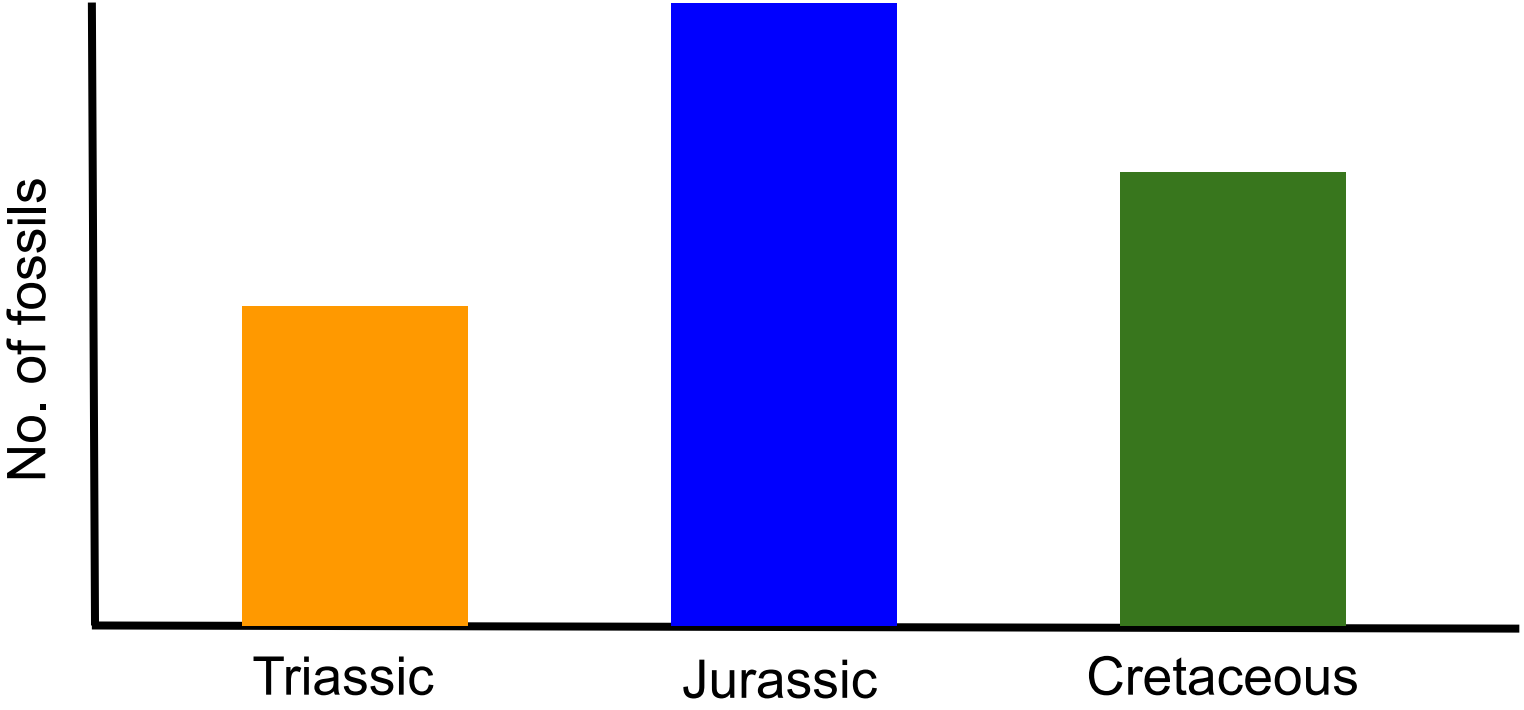


macrostrat.org

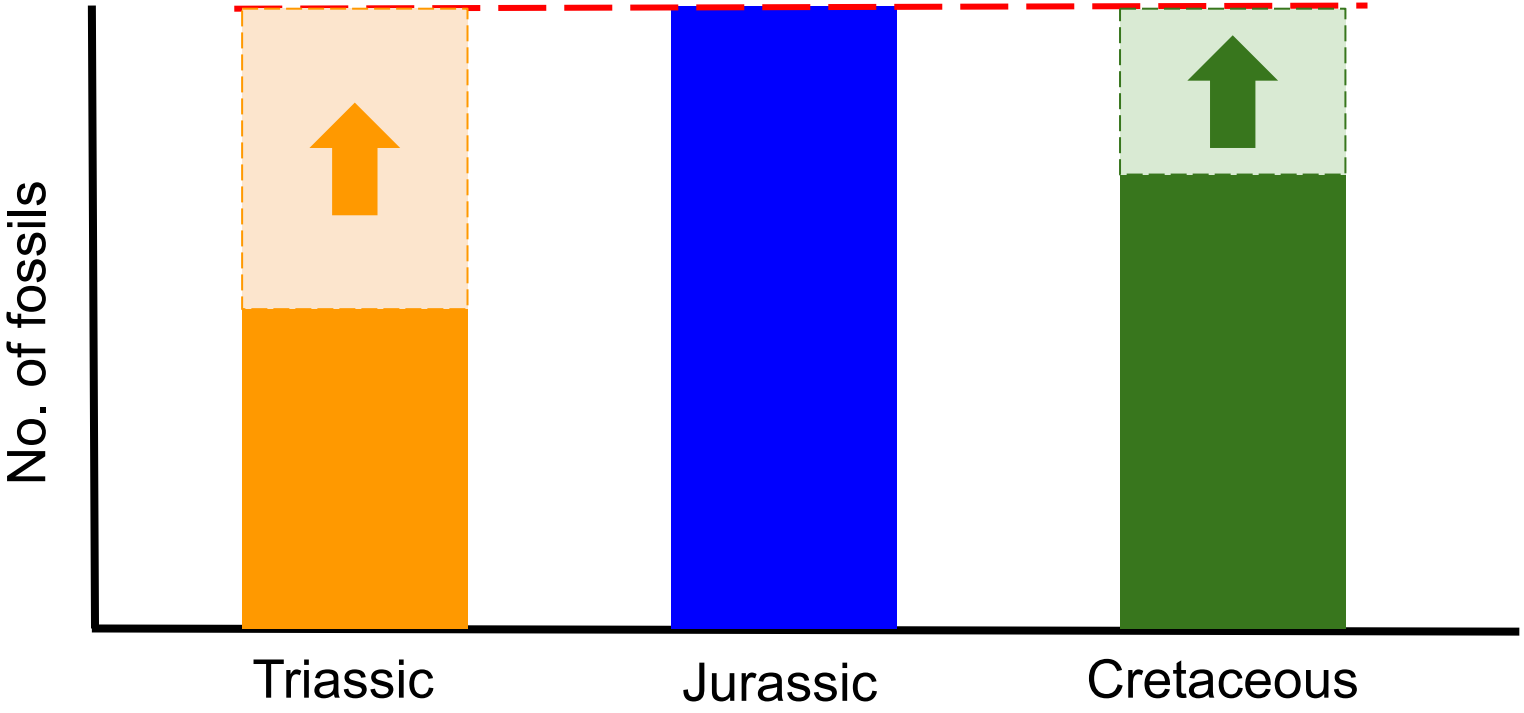
# Help! My data are biased

# Methods

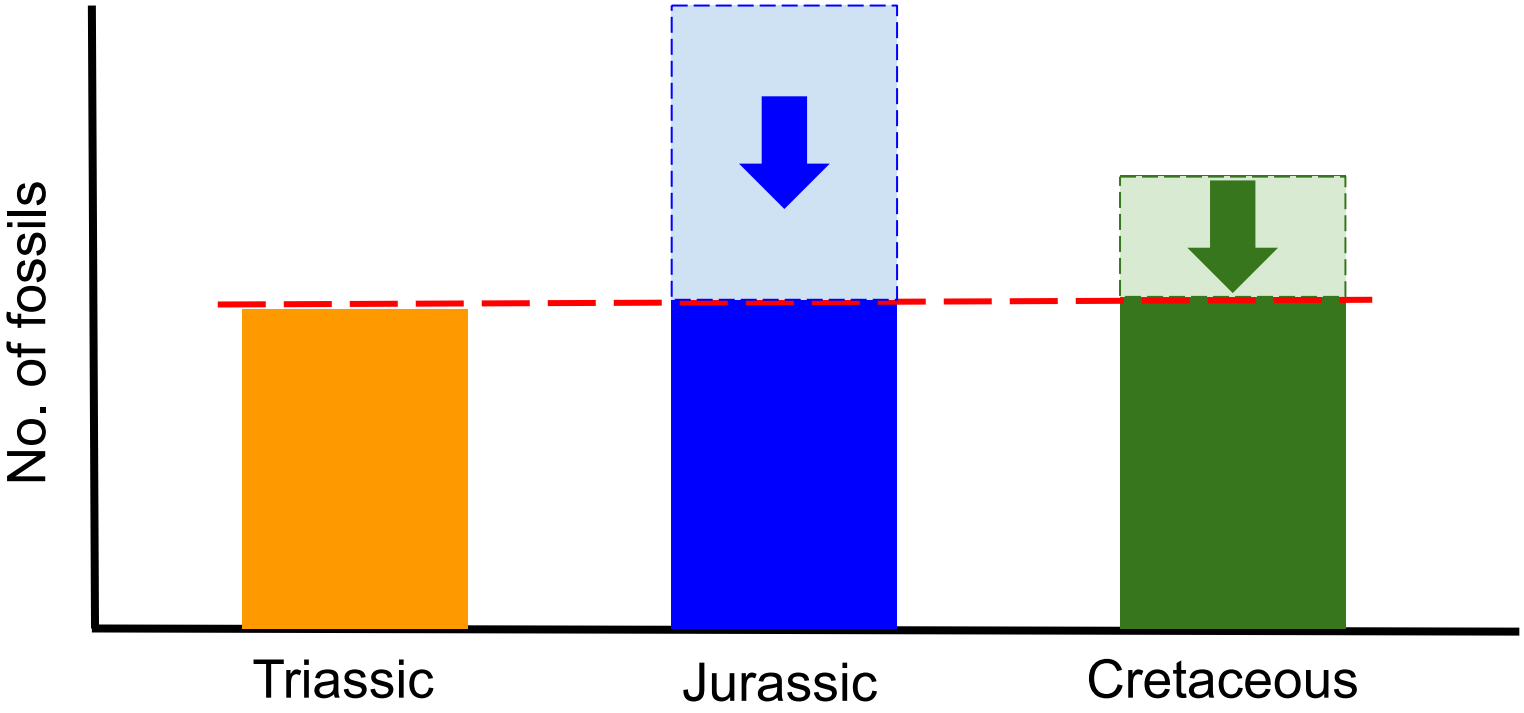
# Keeping it simple?



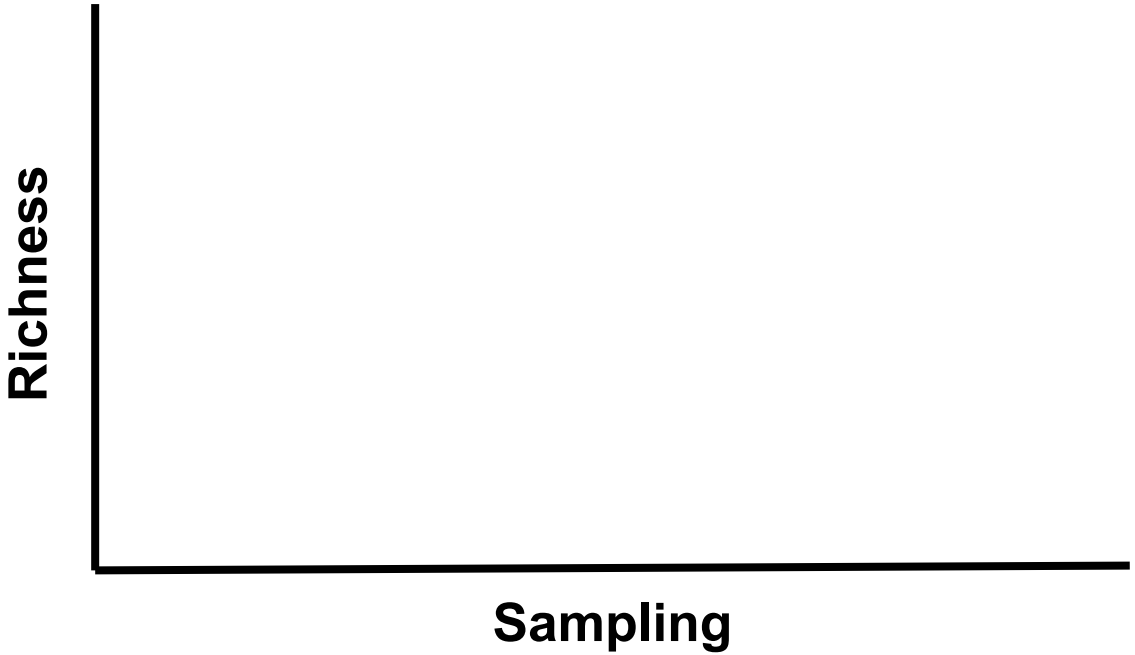
# Extrapolation



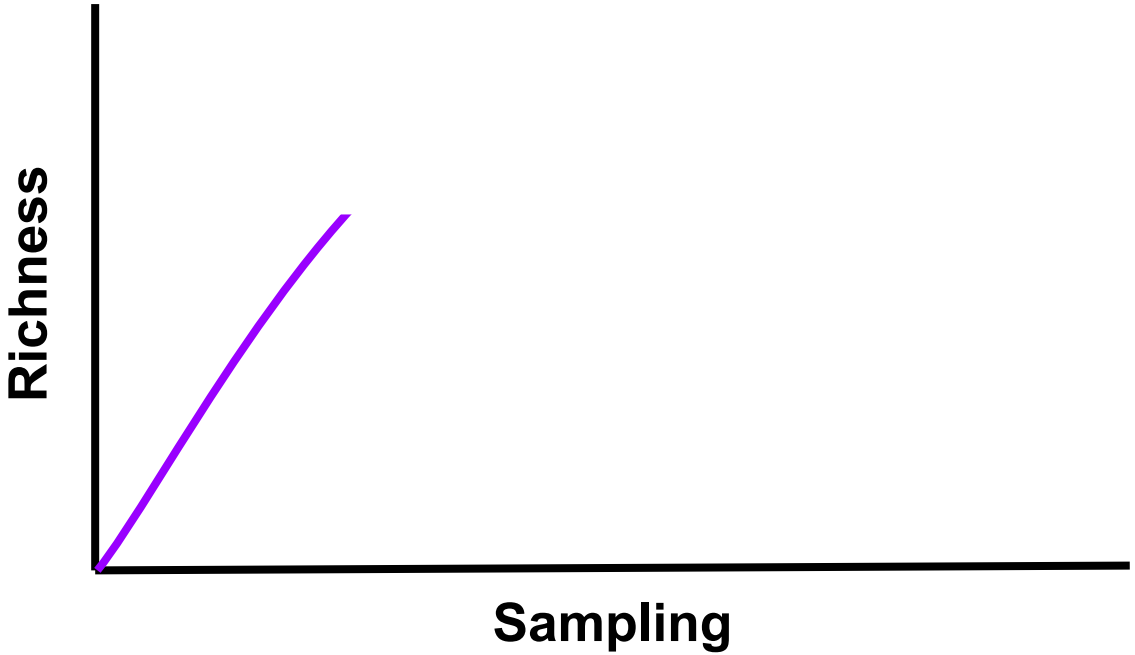
# Subsampling: rarefaction and bootstrapping



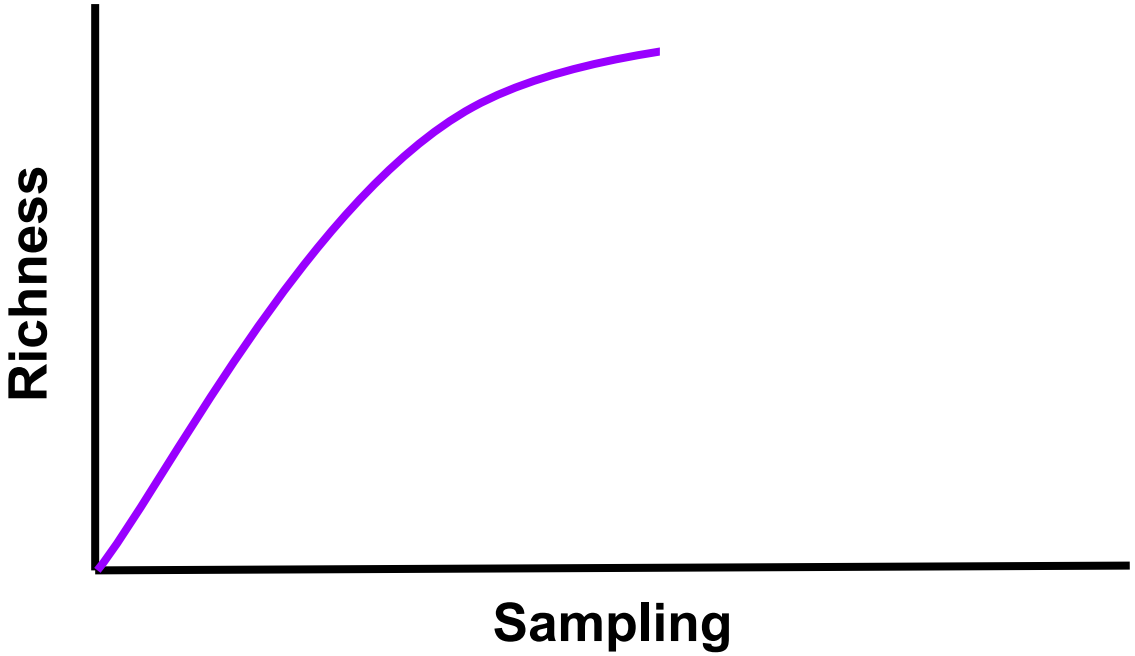
# Collector’s / Species Accumulation Curves



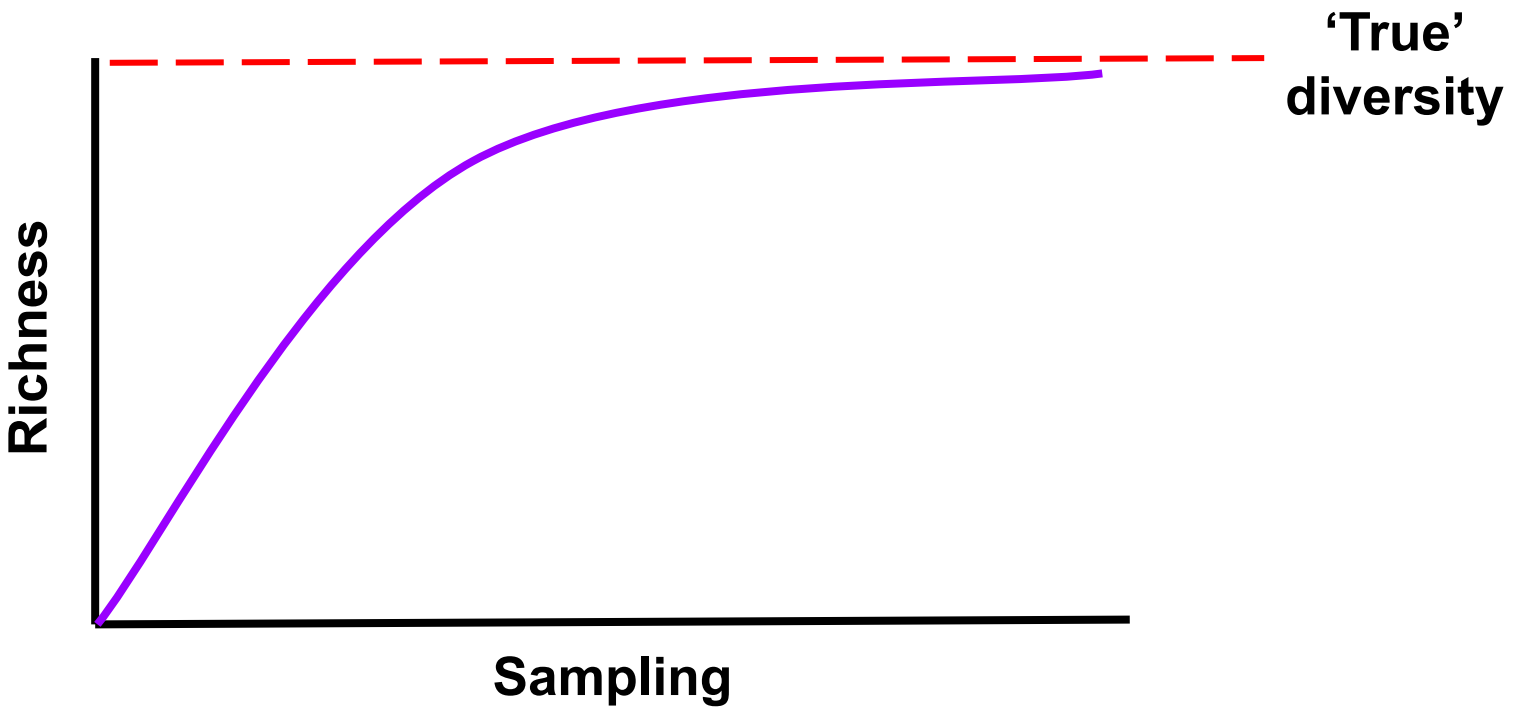
# Collector's / Species Accumulation Curves



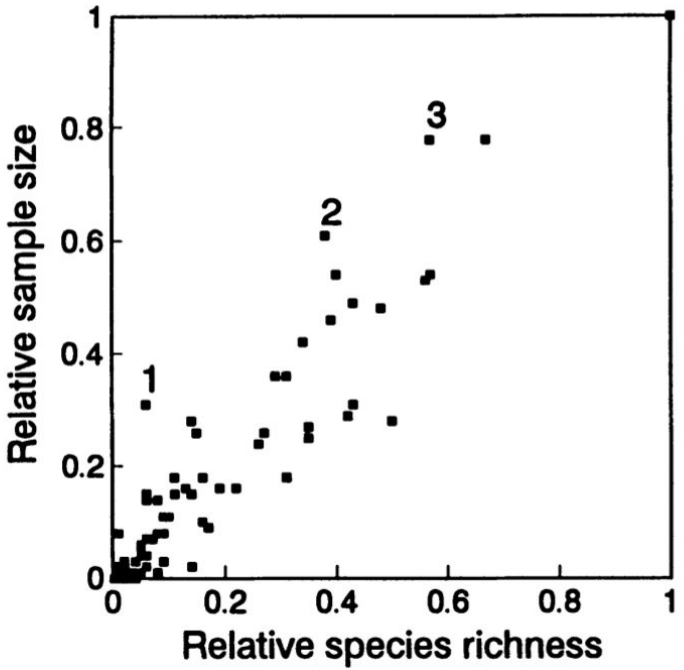
# Collector's / Species Accumulation Curves



# Collector's / Species Accumulation Curves



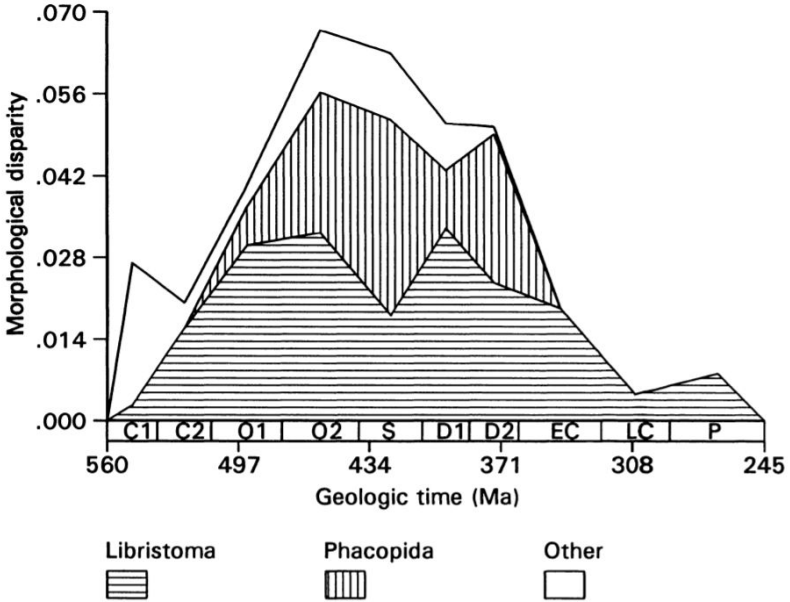
# Bootstrapping example



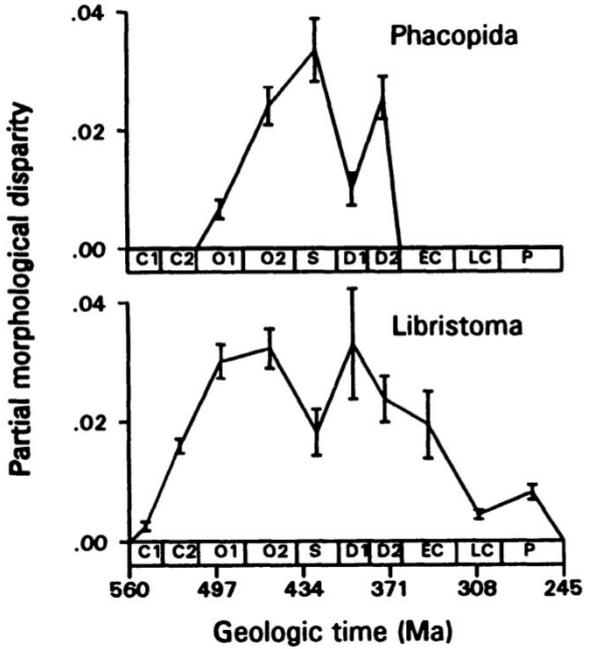

The relationship between sampling and richness in combinations of trilobite subgroups and stratigraphic intervals (Fig. 1)

Foote 1993; *Paleobiology*

# Bootstrapping example



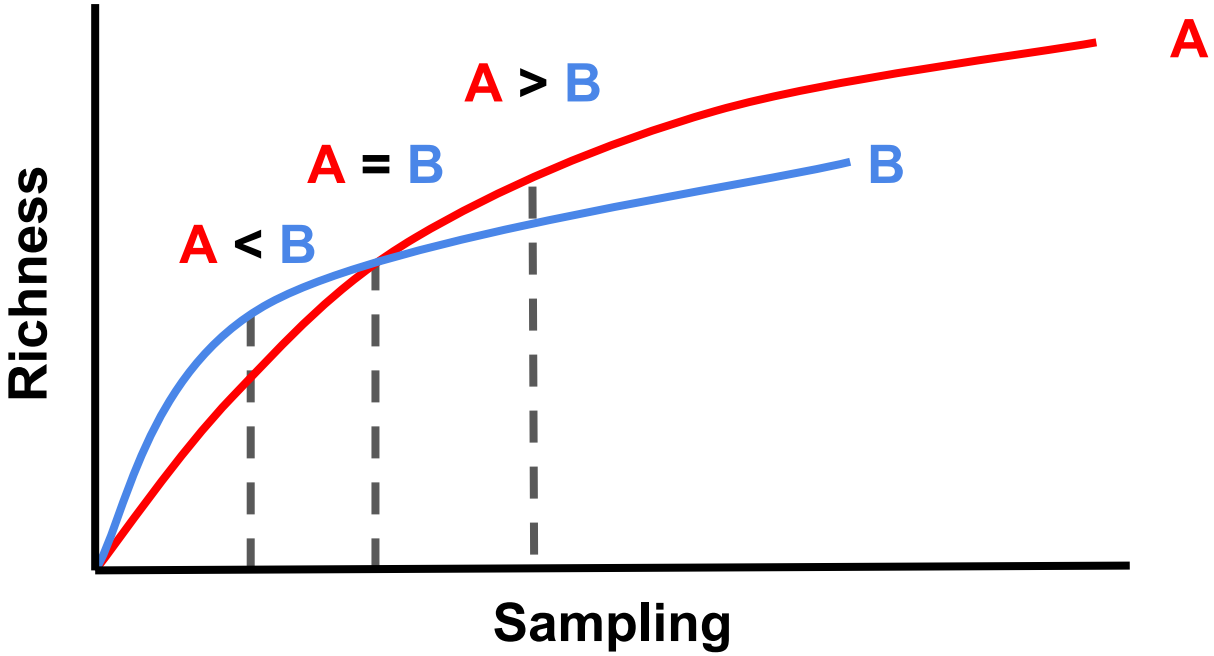

The contribution of different trilobite groups to overall morphological diversity through time (Fig. 2)



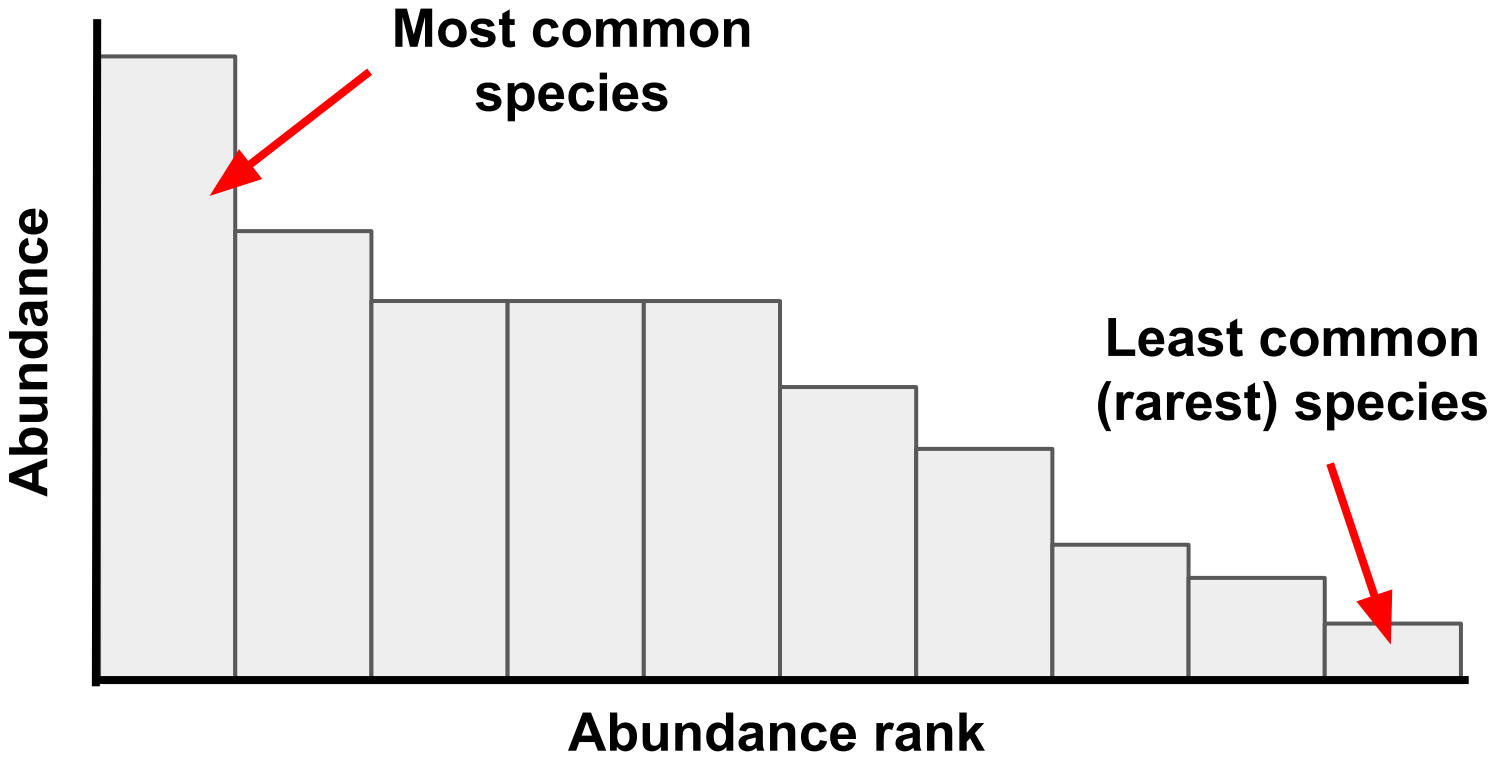

The contributions of Phacopida and Libristoma to overall trilobite disparity, taking into account sampling through a bootstrap regime (Fig. 3)

Foote 1993; *Paleobiology*

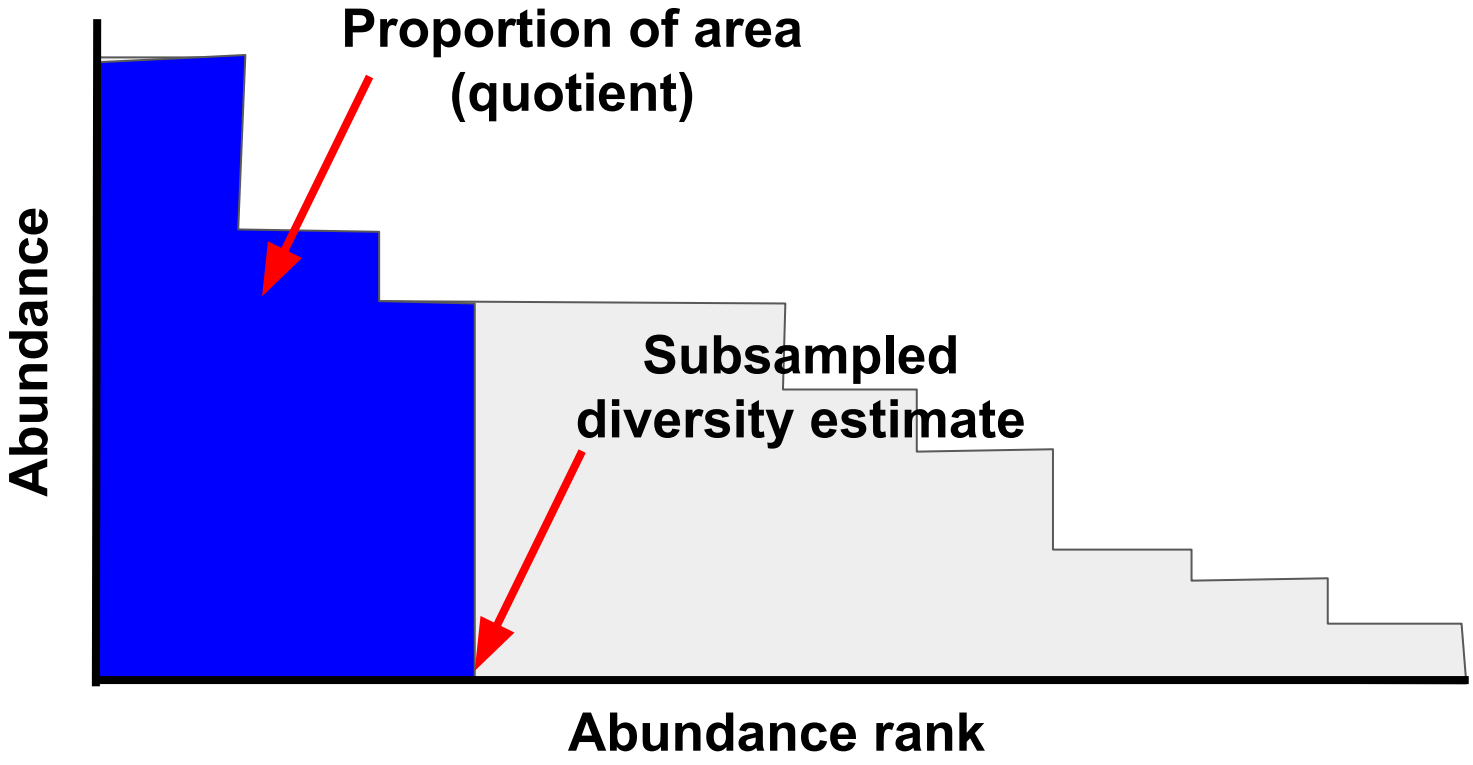
# Subsampling and “The Ghostbuster’s Problem”



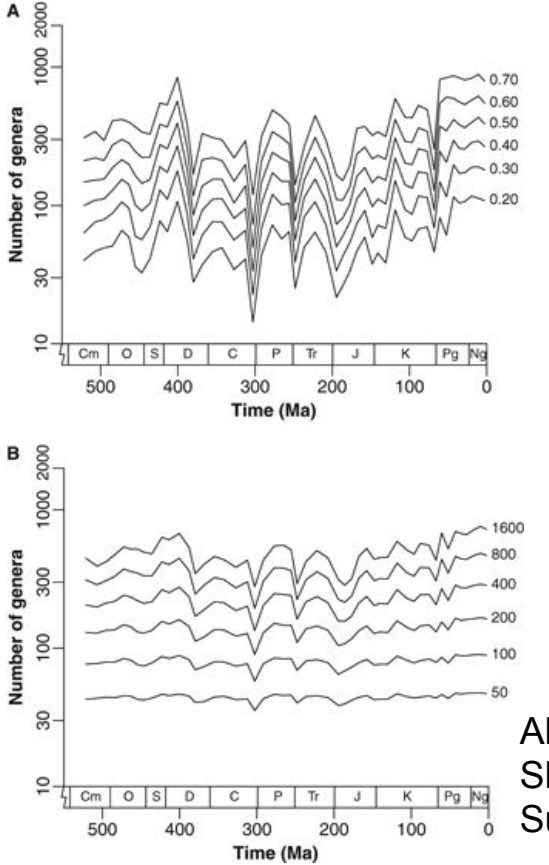
# Rank order abundance



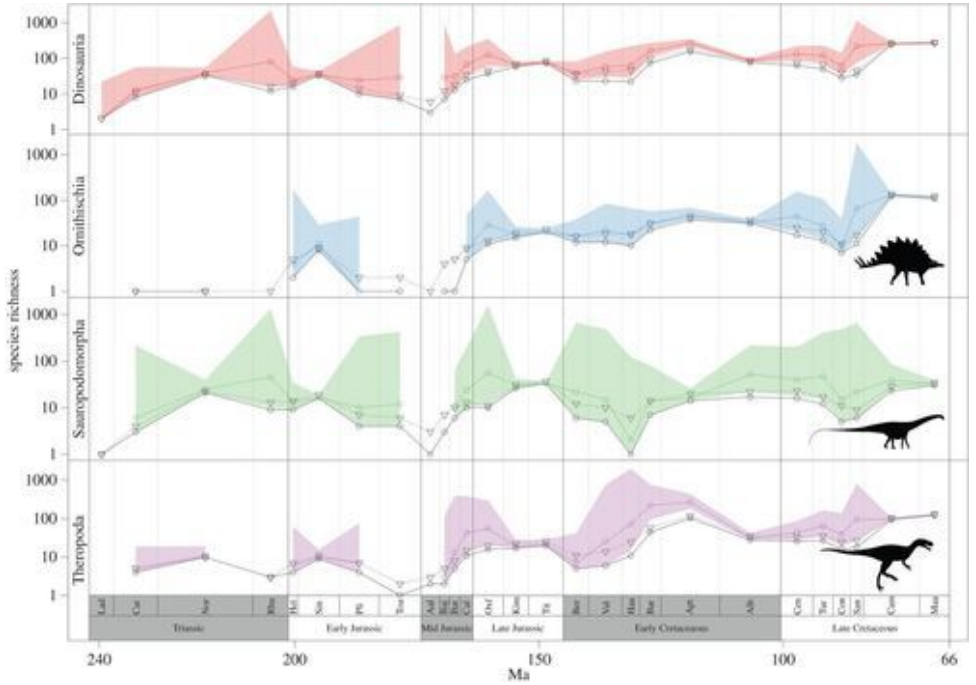
# Coverage-based methods



# Coverage-based methods

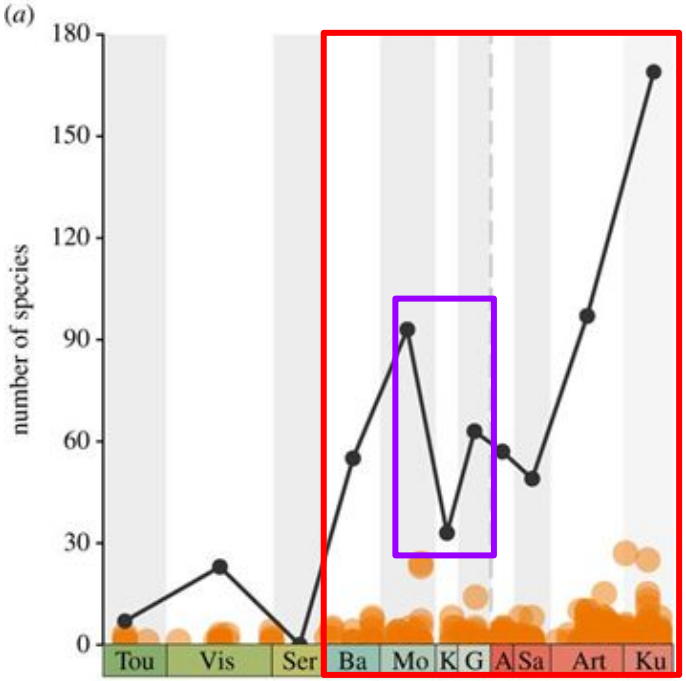



Alroy 2010; *Palaeontology*
Shareholder Quorum Subsampling (SQS)

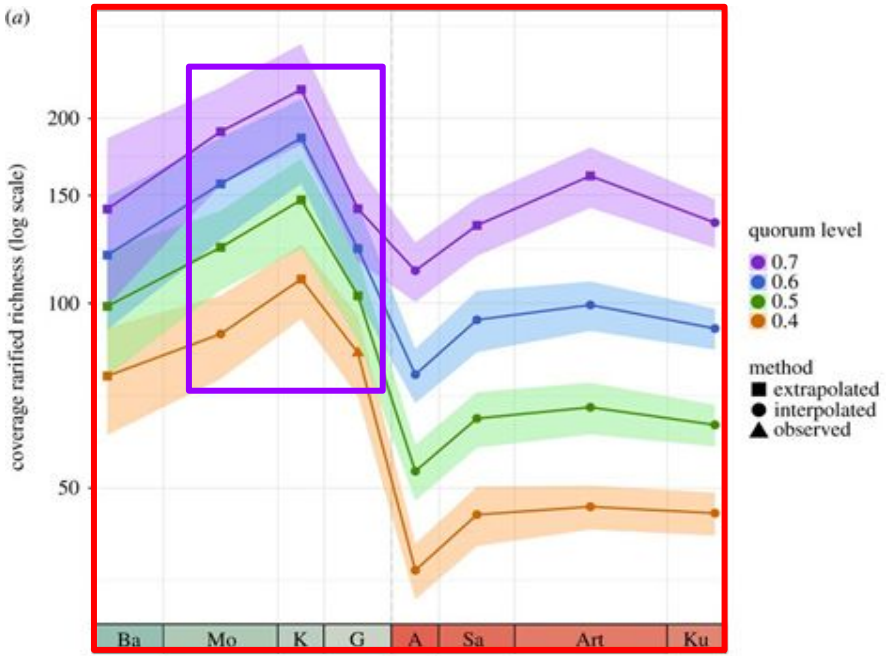



Starrfelt & Liow 2016; *Proc. Roy. Soc. B*
True Richness estimated using a Poisson Sampling model (TriPS)

# Coverage-based examples



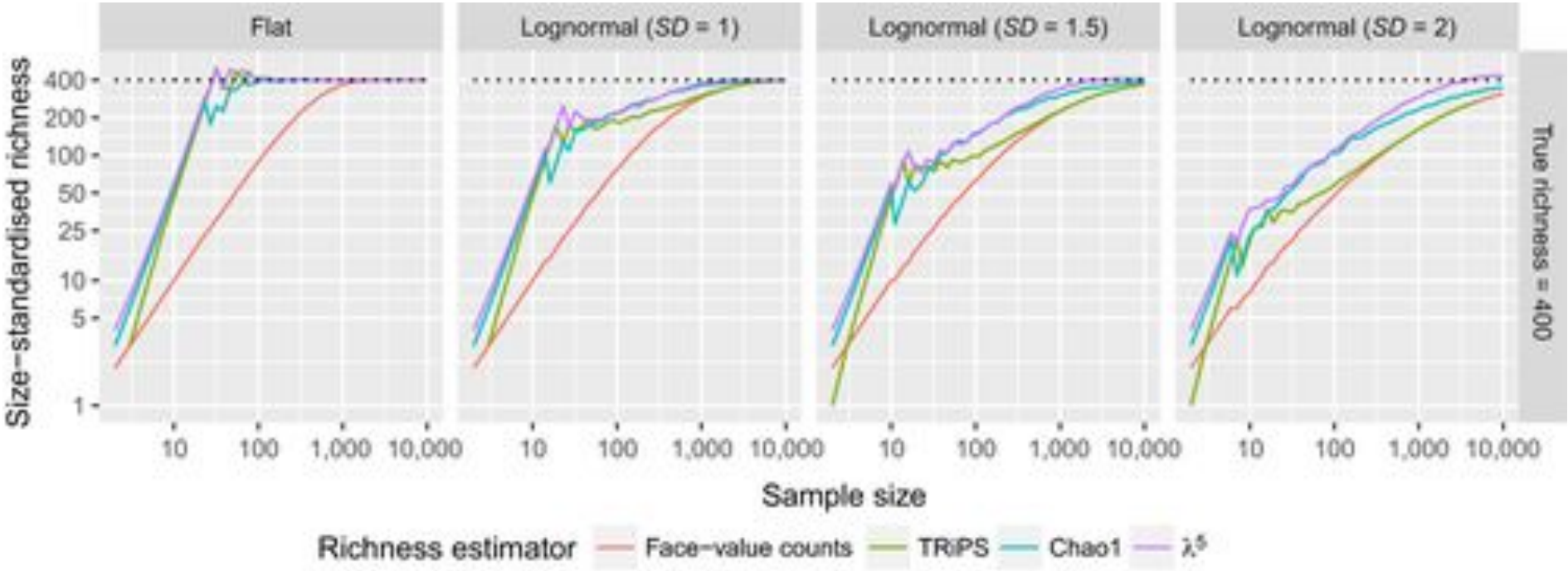

Raw tetrapod diversity curve for the Carboniferous-early Permian (Fig. 1a)




SQS tetrapod diversity curve for the Carboniferous - early Permian (Fig. 3a)

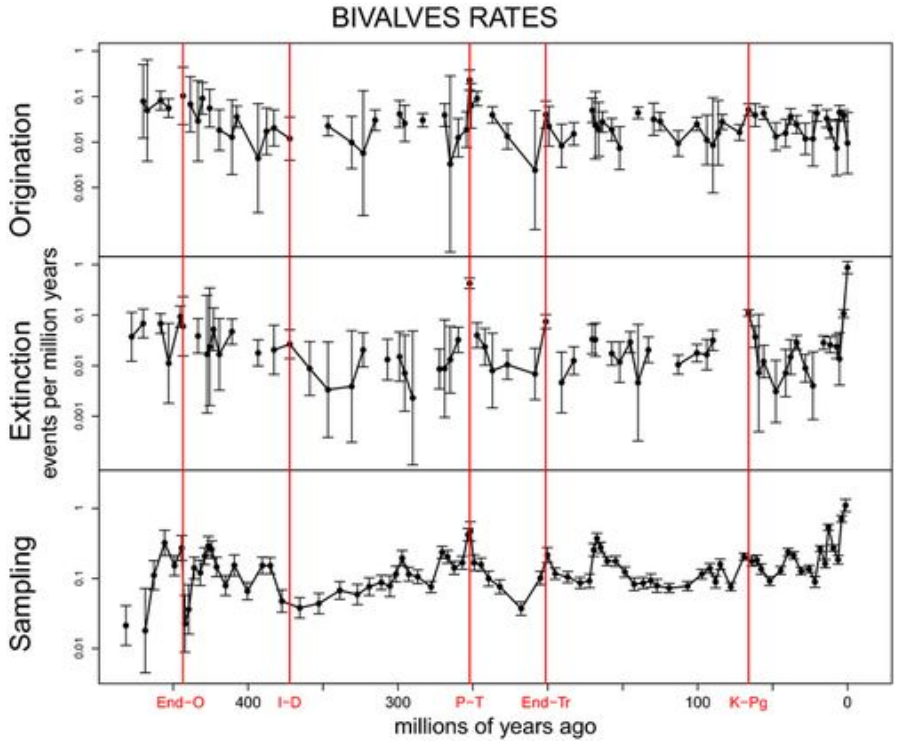
Dunne *et al.* 2018; *Proc. Roy. Soc. B*

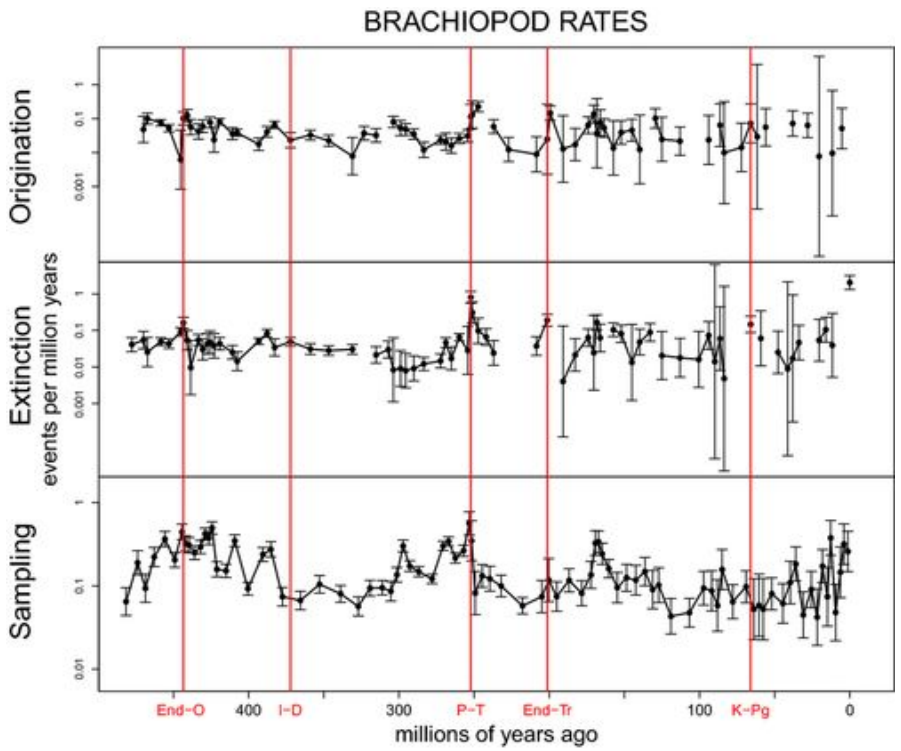
# Which is best?




Close *et al.* 2018; *Methods Ecol. Evol.*, see also: Alroy 2020; *Paleobiology*

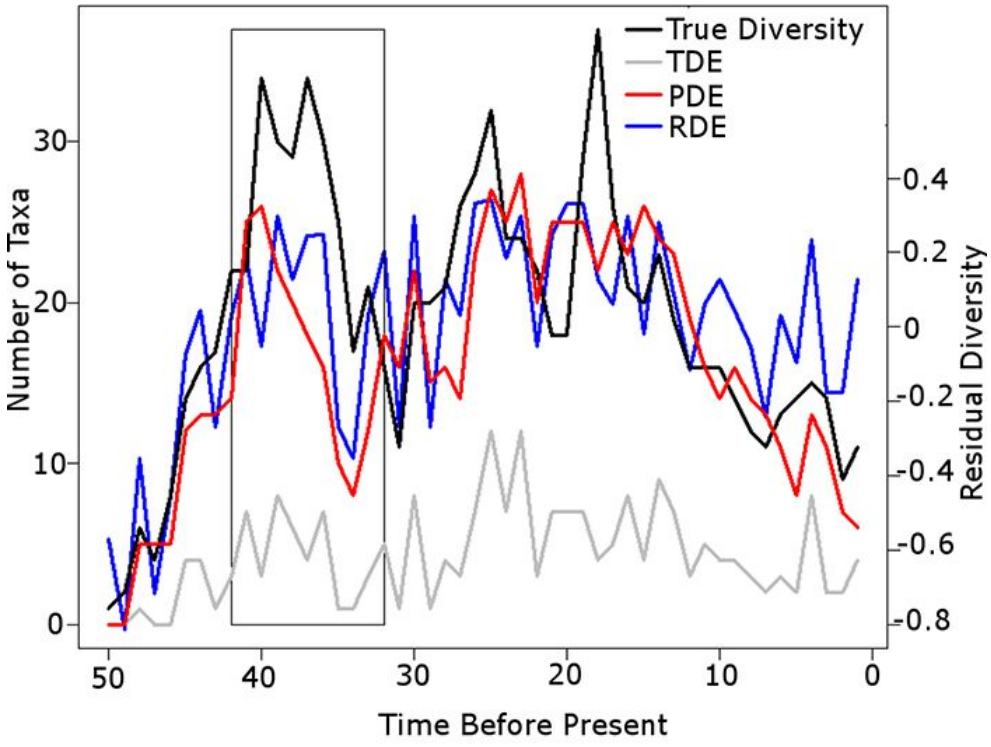
# Capture-mark-recapture

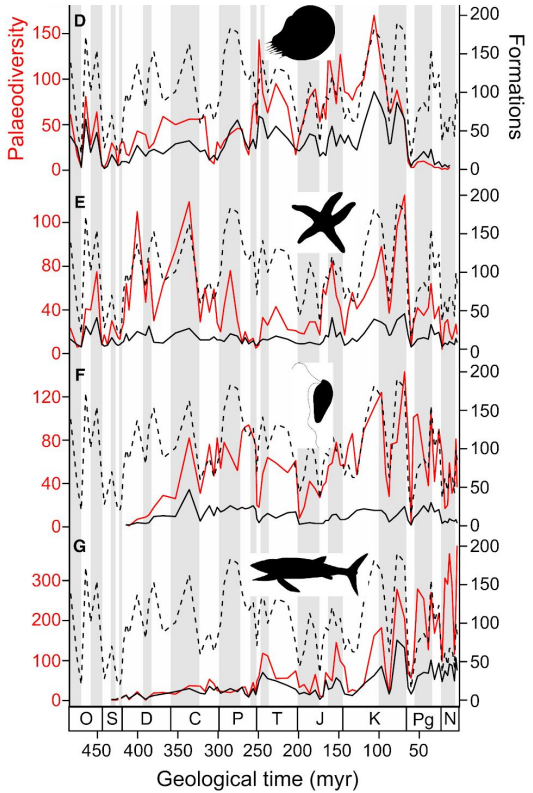



Liow *et al*. 2015; *Ecol. Lett.*

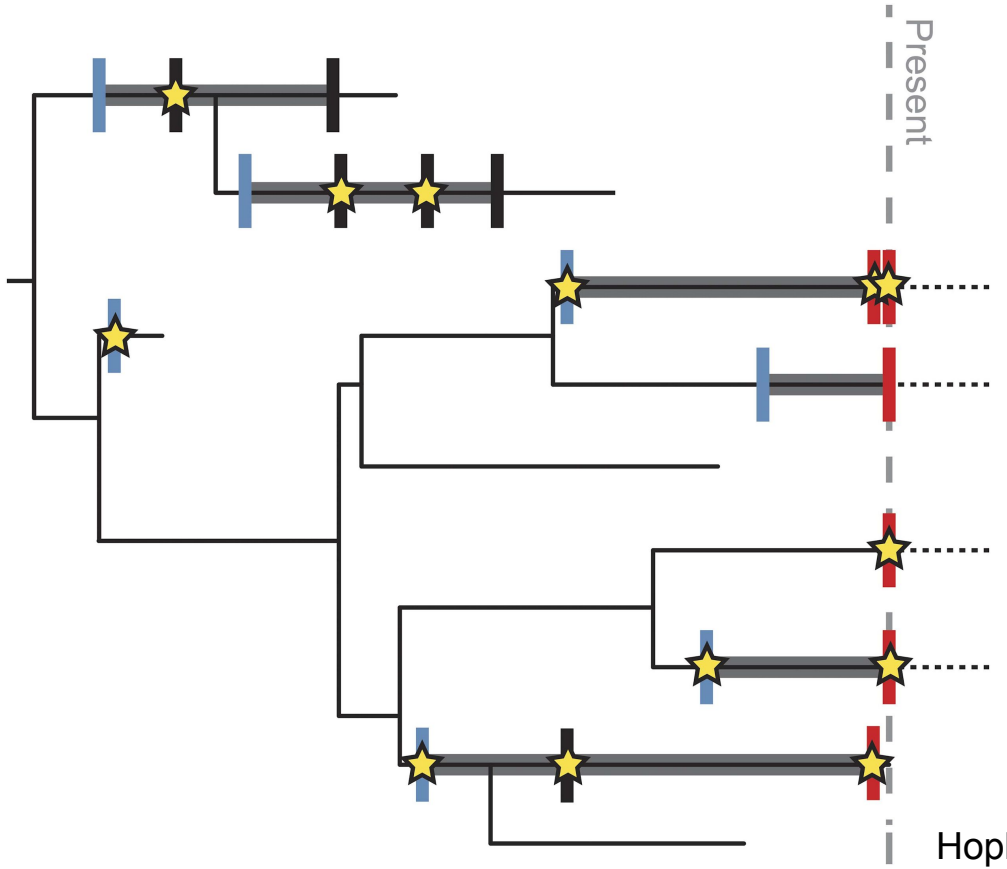
# Sampling proxy correction



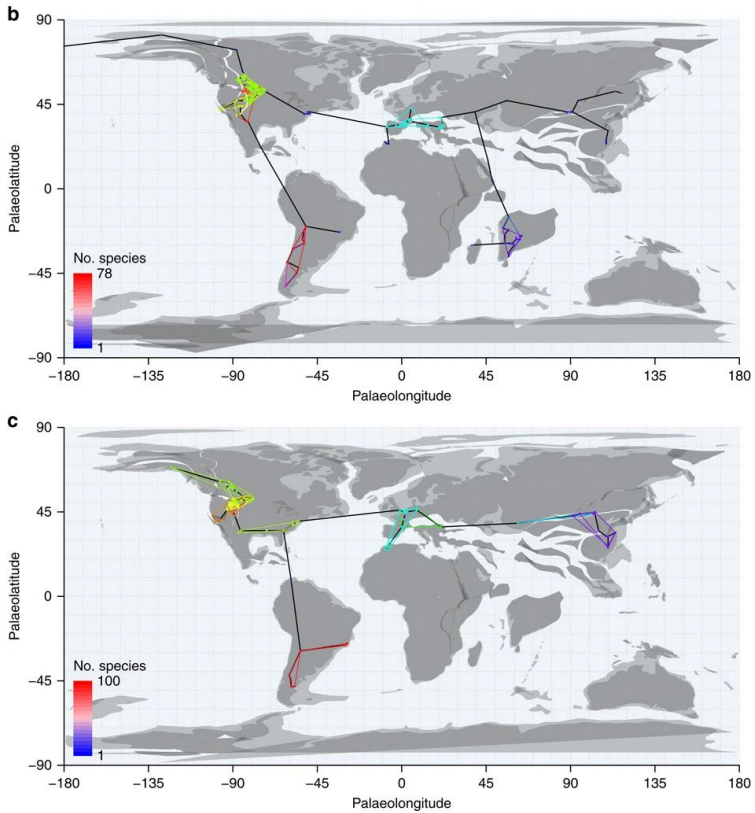

Brocklehurst 2015; *Palaeo. Electronica*, Dunhill *et al.* 2018; *Palaeontology*

# Who needs fossils when you have trees?

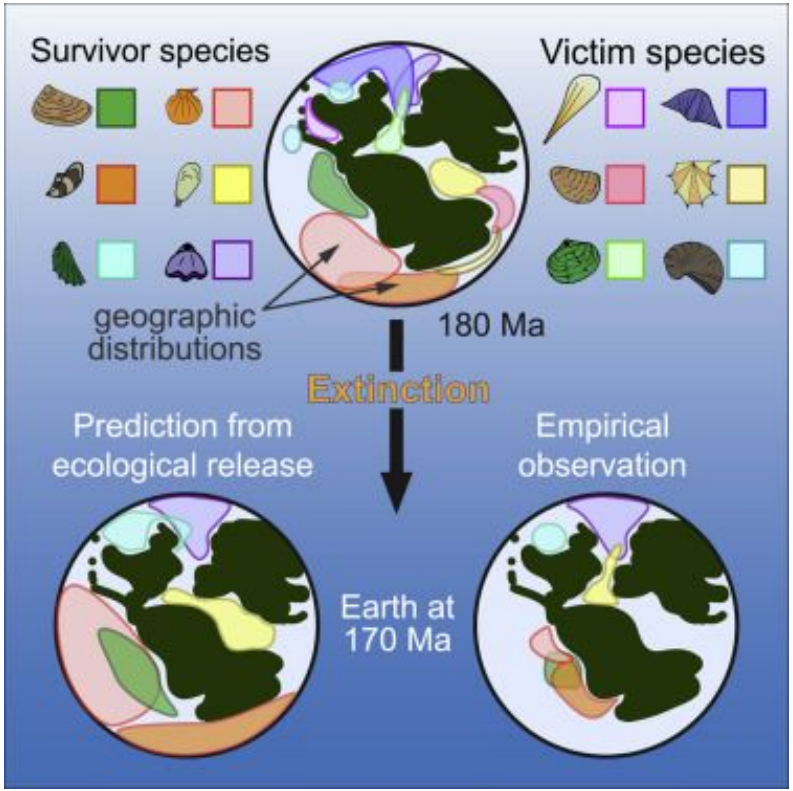



Hopkins *et al*. 2018; *Paleobiology*

# Space vs time




Close *et al.* 2017; *Nature Comms.*




Antell *et al.* 2020; *Curr. Biol.*