Task 03: Utilizing Scientific Databases

I. General

Type up all of your work in a text editor. Basically, you should NEVER type things directly into the `R` terminal. Type them into a text editor, then either run them or copy/paste them into `R`.

When you're **done** with this assignment, turn it in by (1) saving your text document, (2) opening your Terminal or GitBash (3) navigating to the appropriate folder on your computer using `cd`, and (4) typing:

`git add -A` (enter)
`git commit -m "<your name> Task 03"` (enter)
`git push -u origin master` (enter)

II. Downloading Data

This week, we're going to use real data downloaded from a curated central repository. The Paleobiology Database (paleodb.org) is a website that seeks to document and make accessible the occurrence of every fossil ever found. They're far from doing so, but for some groups they do have every single scientifically-documented fossil (including some that are in museums but have not been published on). And their database is extensive enough that, despite several analyses, it itself does not seem biased within major groups any more than the fossil record itself is biased within those groups.

That is to say, the database is biased in that there aren't a lot of annelid fossils recorded in it. But that's because there aren't a lot of annelid fossils! However, although it is missing lots of, say, bivalve occurrences, it isn't missing those bivalves in a biased manner. So for most major fossil-preserving groups, the database is reliable.

Thus, we're going to use `R` to analyze fossil occurrences downloaded from this database. You can go to their website and download the data using a clicky-format, or you can download it by directly accessing their website via `R`. We'll do the latter. In `R`, type `install.packages(‘‘paleobioDB’’, dep = T)` and hit enter. Then, choose your mirror (any US based one) and open your coding file for the week.

Set your working directory as per last week, and type `library(paleobioDB)` to activate the functions stored in that package. Then add:

```
# download data for a specific taxon. Here, dinosauria! Must use a
   scientific term for the taxon
Taxon <- "Dinosauria"

# the min_ma and max_ma arguments control the time-window that you are
   pulling fossils from in millions of years.
```

```
MinMA <- 66
MaxMA <- 252
fossils <- pbdb_occurrences(base_name = Taxon, show = c("phylo", "
   coords", "ident"), min_ma=MinMA, max_ma=MaxMA)
```

Now you have chosen a taxon (dinosaurs) and downloaded all of their occurrences between 252 and 66 million years ago, and stored those occurrences in an object that's called “fossils”. That's your data! Let's analyze it.

III. Analyzing the Data: Through Time

66 million years separate us from the last (nonavian) dinosaur, but *186* million years separated the first and last (nonavian) dinosaur from each other. That's a lot of time for species to evolve! So let's look at how the number of dinosaur species changed during their reign, and also at how fast and slow they made new species.

```
# how many species are known from each time period?
# We'll define a time period Resolution (Res) of 5Ma
Res <- 5
nspeciesOverTime <- pbdb_richness(fossils, rank = "genus", temporal_
   extent = c(MaxMA,MinMA), res=Res)

# I don't like the default plot. Here's an alternative.
par(mar=c(4,5,2,1), las=1, tck=-0.01, mgp=c(2.5,0.5,0))
plot(seq(to=MaxMA, from=MinMA, length.out=nrow(nspeciesOverTime)),
   nspeciesOverTime[,2], xlim=c(MaxMA, MinMA), type="l", xlab="age (
   millions of years ago)", ylab="num. of species", main = Taxon)
```

That plot shows how the raw number of known species of dinosaurs changes over time. We can also look at the rates of change. That is, if the number of species changes, it means that either a new species appeared in the record, or an old species disappeared. So we can plot, over time, the number of first appearances and last appearances to see what is driving the fluctuations.

```
# get the appearance data
newspeciesOverTime <- pbdb_orig_ext(fossils, res=5, rank="species",
   temporal_extent=c(MinMA, MaxMA))

# set up the plot
par(mar=c(4,5,2,1), las=1, tck=-0.01, mgp=c(2.5,0.5,0))

# plot the first appearances
plot(seq(to=MaxMA, from=MinMA, length.out=nrow(newspeciesOverTime)),
   newspeciesOverTime[,1], xlim=c(MaxMA, MinMA), type="l", xlab="age (
   millions of years ago)", ylab="num. of species", main = Taxon)

# add a line for the last appearances
lines(seq(to=MaxMA, from=MinMA, length.out=nrow(newspeciesOverTime)),
   newspeciesOverTime[,2], col='red')

# add a legend
```

```r
legend("topleft", legend=c("first appear", "go extinct"), col=c('black', 'red'), lty=1, bty="n")
```

What do you notice about these two lines? Is there anything that surprises you? When, if we assume these numbers accurately reflect the rate new species were evolving, were dinosaurs evolving the fastest?

IV. Analyzing the Data: Through Space

The Paleobiology Database records everything we know about each fossil occurrence. And if someone finds a fossil, then they know *where* they found the fossil! So let's plot our fossil occurrences on a map.

```r
# Let's map these data so that instead of looking through time, we look across space
# We'll set a color for the oceans and the land on our map
OceanCol <- "light blue"
LandCol <- "black"

# we'll also set some colors for the fossil occurrences. I like to choose colors from http://colorbrewer2.org/
# there is a package (RColorBrewer) that lets you generate those palettes in R, and there are many other ways to choose nice colors
# but for now, let's just use some from that website. Feel free to change these to whatever you want. You just need 2 - 5 colors here.
Cols <- c('#fee5d9','#fcae91','#fb6a4a','#de2d26','#a50f15')

# Now, let's make a map!
par(las=0)
pbdb_map_richness(fossils, col.ocean=OceanCol, col.int = LandCol, col.rich=Cols)
```

V. Analyzing the Data: Through Space & Time

Not every place in the world preserves rocks from every time period! And certainly not everywhere preserves rocks from every environment in every time period. So where we find fossils from a particular group changes over time. That we don't find fossils of Triassic sauropods in West Virginia doesn't mean that they didn't live here, just that we don't have fossils from both the right time and environment present here now. So let's look at that by making maps of fossil occurrences of dinosaurs...but for different time periods! It's easy.

```r
# Let's use the timescale to look at where dinosaur fossils have been found from different periods!
# This lets us plot both things together
# The Geological Timescale is here: https://www.geosociety.org/documents/gsa/timescale/timescl.pdf

# We'll first get all of the Triassic fossils...
```

```
MinMA <- 201
MaxMA <- 252
triassic_fossils <- pbdb_occurrences(base_name = Taxon, show = c("phylo
   ", "coords", "ident"), min_ma=MinMA, max_ma=MaxMA)

# Then Jurassic fossils...
MinMA <- 145
MaxMA <- 201
jurassic_fossils <- pbdb_occurrences(base_name = Taxon, show = c("phylo
   ", "coords", "ident"), min_ma=MinMA, max_ma=MaxMA)

# Then Cretaceous fossils...
MinMA <- 66
MaxMA <- 145
cretaceous_fossils <- pbdb_occurrences(base_name = Taxon, show = c("
   phylo", "coords", "ident"), min_ma=MinMA, max_ma=MaxMA)

# now let's make a series of maps
dev.new(height = 7.8, width = 13)
pbdb_map_richness(triassic_fossils, col.ocean=OceanCol, col.int =
   LandCol, col.rich=Cols)
mtext(side = 3, ‘‘Triassic (252 - 201Ma)", cex=3, line=-2)

dev.new(height = 7.8, width = 13)
pbdb_map_richness(jurassic_fossils, col.ocean=OceanCol, col.int =
   LandCol, col.rich=Cols)
mtext(side = 3, ‘‘Jurassic (201 - 145Ma)", cex=3, line=-2)

dev.new(height = 7.8, width = 13)
pbdb_map_richness(cretaceous_fossils, col.ocean=OceanCol, col.int =
   LandCol, col.rich=Cols)
mtext(side = 3, ‘‘Cretaceous (145 - 66Ma)’’, cex=3, line=-2)
```

VI. Analyzing the Data: Comparing with Another Group

Looking at dinosaur data is interesting, but as with everything, if we don't have a *basis for comparison* we can't make heads or tails of the numbers. Is 35 species in some place at some time a lot, or a little? How many new species at once is too many, how many is too few? We need to compare the group to another one. So let's compare dinosaurs during the Mesozoic to mammals from the Mesozoic!

```
# We can also compare two groups. Here we'll download data for another
   group of animals during the same time period
Taxon2 <- "Mammalia"
MinMA <- 66
MaxMA <- 252
fossils2 <- pbdb_occurrences(base_name = Taxon2, show = c("phylo", "
   coords", "ident"), min_ma=MinMA, max_ma=MaxMA)
```

```r
nspeciesOverTime2 <- pbdb_richness(fossils2, rank = "genus", temporal_
   extent = c(MaxMA,MinMA), res=Res)

# Now we'll plot both groups together to compare them!
par(mar=c(4,5,2,1), las=1, tck=-0.01, mgp=c(2.5,0.5,0))
Col_dino <- Cols[length(Cols)]
Col_mammal <- Cols[1]
LineWidth <- 2
plot(seq(to=MaxMA, from=MinMA, length.out=nrow(nspeciesOverTime)),
   nspeciesOverTime[,2], xlim=c(MaxMA, MinMA), type="l", xlab="age (
   millions of years ago)", ylab="num. of species", col=Col_dino, lwd=
   LineWidth)
lines(seq(to=MaxMA, from=MinMA, length.out=nrow(nspeciesOverTime2)),
   nspeciesOverTime2[,2], col = Col_mammal, lwd=LineWidth)
legend("topleft", legend=c(Taxon, Taxon2), col=c(Col_dino, Col_mammal),
    bty="n", lwd=LineWidth)
```

VII. Extension

Choose two *different* groups and (almost certainly) a different time period window (i.e., not 252 - 66Ma) and make the above graphs for those two groups. Then, looking at the graphs for your two chosen groups, come up with a *testable hypothesis* to explain any patterns that you see in your data. You don't actually have to test them. Just choose some species, download data, plot the data, look at the plots, and come up with a potential (testable) explanation for the plot(s).

Nota bene: I'd consider including at least one marine group (e.g., echinoderms, brachiopods, bivalves, gastropods...), as it'll make your life a lot easier.

VIII. Extra Credit

Almost all scientific databases have published `R` packages for interfacing with like like the paleobioDB package we used here. Ones that don't are still accessible via `R`, as you can scrape any data you want off the internet (you can look at people who do analyses of tweets using `R`).

Some large, easily-accessed databases are: fishbase (`rfishbase`), iNaturalist (`rinat`, `spocc`), eBird (`auk`), GenBank (`genbankr`), InterMine (`InterMineR`), phylogenetic trees (`phylotastic`), VertNet (`rvertnet`), etc etc etc.

For extra credit, choose one of those databases (I recommend iNaturalist, eBird, or fishbase, but any database on any topic is fine), read up on the interface package, and write some code to download and make a single plot using those data.