Equidae Fossil Analysis

Roxana Nojoumi

Summary

Equidae is a family of odd-toed ungulate mammals of horses and horse-like animals. It is sometimes known as the horse family. In this study I used the Paleobiology database to obtain the fossil occurrences for Equidae family to find the extinction rate through time using PyRate. To find a possible cause for the change in extinction rate of the horse family I again used the Paleobiology database for Equidae and Plantae family to analyze the relationship between the plants diversity and horse diversity through time. The results showed that there is no correlation between the plants and horses' diversities. In addition, I used the teeth measurements for the horse fossils through time to find any possible pattern for the change in the teeth size through time. The result showed an increase of the teeth measurement through time. More studies can be done focusing on diversity of each individual family of plants to find a possible relationship between the diversity and density of a specific plant family and horse extiction through time.

Introduction

55 million years phylogeny of family Equidae shows the macroevolutionary histry of this family. Interpretation of fossil records suggests a diversity in the extinct forms. By 55 million years ago, the first members of the horse family, the dog-sized Hyracotherium, were scampering through the forests that covered North America. For more than half their history, most horses remained small, forest browsers. But changing climate conditions allowed grasslands to expand, and about 20 million years ago, many new species rapidly evolved.(MacFadden 2005)Some but not all became larger and had the familiar hooves and grazing diets that we associate with horses today. Only these species survived to the present, but in the past, small and large species lived side by side.(MacFadden 1986) In this study, I used the data set provided by *paleobiology* to obtain the Equidae's fossil occurrence, Equidae's teeth measurement, Equidae's location and also plantae's occurrence through time. I used the program PyRate to find the the extinction and speciation rate of the Equidae family. The extiction rate showed a jump in extinction rate towards the present time. To find any possible explanation for this pattern I compared the diversity of the plants and horses through time to find any possible correlation between them. In addition, to see the pattern for the change in morphology of the horses in time, I plotted the teeth measurement for Equidae in time. I predict that the change in teeth measurement should correspond to the change in types of plant existed in a specific time.

Methods

Data description

There are four main data files, which I have used. All the files were downloaded from Paliobiology website. (https://paleobiodb.org) First one is for Equidae fossils which contains the specimen number and teeth measurements in centimeters. Second one is for Equidae fossils which contains the species names, their rank and the minimum and maximum age of the fossils in Ma. The third one is for Plantae fossils which contains the species names, their rank and the minimum and maximum of the age of the fossils. Fourth one is for Equidae fossils which contains the species name and the location of the fossil found. A link to the github repository containing this file is shown below.

https://github.com/rnojoumi/eeb-174-final-project

Developing the comparative database

I used online data base Paliobiology to download the data for the teeth measurement of the Equidae fossils. The file downloaded contained the specimen number for the species and the length measurements. With the following code, I extracted the length measurement and speciment number and wrote it in anohter file.

```
# extract measurements
#open the file with the measurements for the equidae family
#then extract only the measurements for their length and write
#the length with its speciment number in a new file
my_file5 = open("equidae-measurement.csv", encoding = "ISO-8859-15")
my_line5 =my_file5.readlines()[18:]
output= open("eq_length.csv" ,"w")

from collections import defaultdict
length = defaultdict(list)

#from collections import defaultdict
#width = dedfaultdict(list)
for line in my_line5:
    line= line.replace('"', '')
    line= line.replace("'", '')

    sp_no= line.split(",")[1]
    len_or_wid= line.split(",")[5]
    meas = line.split(",")[6]

    if len_or_wid == "length":
```

```
        length[sp_no].append(meas)
        output.write(sp_no + ","+ meas + "\n")
```

The output file does not have the species name. In order to find that, I downloaded another file that contained species names and their corresponding speciment number. I used dictionaries to combine these two files and write another file which has the species names and their length measurements. The python code I used is as follows.

```
#here the output is a file with the species name and their length
#measurement open file eq_sp_occ_no.csv which its first column is
#the specimen number, its 20th column is the accepted name for the
#family equidae. I will use these columns to make a dictionary
#with speciment number as its keys and accepted name as its values.
#open file eq_length.csv which its first column is the speciment number
#and its second column is the length
#measurement for the equidae family from the file
#equidae-measurement.csv downloaded.make a second dictionary
#with specimen number as its keys and length measurements as its
#values.so i can compare these two dicts to find the name of the
#species for the length measurements i found.

file= open("eq_sp_occ_no.csv", encoding = "ISO-8859-15")
my_line6= file.readlines()[18:]

file2 =open("eq_length.csv", encoding = "ISO-8859-15")
my_line7= file2.readlines()

output = open("sp_name_length.csv", "w")

mydict ={}
mydict2 ={}
for line in my_line6:#for lines in first file
    line= line.replace('"', '')
    line= line.replace("'", '')

    sp_no = line.split(",")[0] #specimen number
    acpt_name = line.split(",")[19]#accepted name


    mydict[sp_no]= acpt_name
    #first dictionary with specimen number and species names

for line in my_line7:#for lines in second file
    line= line.replace('"', '')
    line= line.replace("'", '')
```

```
    line= line.replace("\n", "")

    sp_number= line.split(",")[0] #specimen number
    length= line.split(",")[1] #length measurements

    mydict2[sp_number]= length

for each in mydict.keys():
    for keys in mydict2.keys():
    #for each key value in the first dict,
    #if it was equal to the value in second
    #dict(comparing specimen numbers)
        if each == keys:
                output.write(mydict[each] + "," + mydict2[each]+ "\n")
```

In order to find the fossil ages for each of these length measurements, I used another file which contained the species names and their min and max age. I used dictionaries to combine this file with the previous file made to write another file which has the species name, length measurement,min age, max age, range(max-min) and average of min and max.

```
#here i want to combine two file into one with the output
#containing these columns 1)species name 2)species length measured
#3)min age 4)max age 5)range(max-min) 6)average of min and max
import csv
file1= open("sp_uniq_range.csv", encoding = "ISO-8859-15")
#contains 1)name 2)min 3)max 4)range(max-min)
file2= open("sp_name_length.csv",encoding = "ISO-8859-15")
#contains 1)name 2)length


line1= file1.readlines()#read through lines
line2= file2.readlines()
#define 2 dictionaries
dict_min={}
dict_max={}
dict_range={}
dict_len={}
dict_ave={}#also want to calculate the everage of the
#min and max to be able to plot it vs length
for line in line1:
    line= line.replace('"', '')
    line= line.replace("'", '')
    line= line.replace("\n", "")
```

```
    sp_name= line.split(",")[0] #species names
    sp_min = line.split(",")[1] #minimum number is the second column
    sp_max = line.split(",")[2] #maximum number
    sp_range = line.split(",")[3] #range
    sp_ave = (float(sp_min) + float(sp_max))/2
    #find the average to use for graphing later

    #define dicts
    dict_min[sp_name]= sp_min
    dict_max[sp_name]= sp_max
    dict_range[sp_name]=sp_range
    dict_ave[sp_name]=str(sp_ave)
output5= open("sp_len_min_max_range_ave.csv", "w")

for line in line2:
    line= line.replace('"', '')
    line= line.replace("'", '')
    line= line.replace("\n", "")

    sp_name2 = line.split(",")[0] #species names
    sp_length = line.split(",")[1]#length measurements


    dict_len[sp_name2]= sp_length


#combine the dictionaries and write in the output file
for each in dict_len.keys():
#for all the dicts with the same species name:
    for keys in dict_min.keys():
        for g in dict_max.keys():
            for h in dict_range.keys():
                if each == keys==g==h:
                    output5.write(each +","+ dict_len[each] +
                    ","+dict_min[each] +","+ dict_max[each] +
                    ","+ dict_range[each] + "," + dict_ave[each]+ "\n")


file1.close()
file2.close()
```

I used ggplot in R to graph the teeth measurements through time from the file created above for Equidae family. The R code used to make this graph is as follows:

```
meas <- read.csv("sp_len_min_max_range_ave.csv", header = FALSE)
ggplot(data = meas, aes(x = V6, y = log(V2))) +
  geom_point()  + geom_smooth(method="lm") +
  ggtitle("Fossil occurence vs. length") + xlab("occurrence (Ma)")+
  ylab("length measurement")
```

For visualization of the location of the fossil occurrences, I downloaded a file which contains the location of the fossils. I used R and a library named *rworldmap* to graph the fossil locations in world map.

```
airports <- read.csv("./eq_new_col.csv", header = TRUE)
#install.packages("rworldmap")
head(airports)

library(rworldmap)
newmap <- getMap(resolution = "low")
plot(newmap, xlim = c(-180,150 ), ylim = c(-40, 55), asp = 1)
points(airports$lng, airports$lat, col = "green", cex = .4)
```

I downloaded a file for the fossil occurrences of Equidae from PBDB. First I used the file I downloaded to write another file which excludes the headers and contains data for only those who have been identified to family level. The python code is as follows:

```
#to clean up the file from headers, we read its line
#and give it a range which is after line 18 and rewrtie
#it to another file
my_file= open("equidae.csv" , encoding = "ISO-8859-15")
my_line = my_file.readlines()[18:]
output = open("plants_new.csv", "w")
for line in my_line:
    output.write(line)
```

The output however contains multiple sets of data for each family. In order to have only one set of data for the min and max fossil age of each family, I used defaultdict in python which combines the keys that are the same and so the output would have unique set of family without repetition.

```
def sp_range(filename):
    file = open(filename,'r')
    my_line = file.readlines()   #read through lines
    #from the filename
   #define dictionary to write
    from collections import defaultdict
    species_ranges = defaultdict(list)
    for line in my_line:
```

```
        line= line.replace('"', '')
        line= line.replace("'", '')
        species_name= line.split(",")[5]
        sp_or_genus= line.split(",")[6]
        max_int = line.split(",")[14]
        min_int = line.split(",")[15]
        if sp_or_genus=="family": # for only when the
        #genus rank has been identified append the min
        #and max to the species name in dict
            species_ranges[species_name].append(max_int)
            species_ranges[species_name].append(min_int)
         #sort the dictionary
        sorted(species_ranges.keys())

    file.close()
    return (species_ranges)
sp_range("equidae_new.csv")
```

The output of this function contains more than one min and max value for the age of each unique genus. In order to find the minimum and maximum from all the values, I used functions *min* and *max*, and I wrote genus and species name and their minimum and maximum value in a new file.

```
#second function
dictname=sp_range("equidae_new.csv")

def dict_min_max(dictname, outputfile):

    output= open("equidae_min_max.csv", "w")
    for keys, values in dictname.items():
        minage = min(values)
        maxage = max(values)
        genus = keys.split(" ")[0]
        species = keys
        output.write(genus + ","+ species+ ","+ minage+
        ","+ maxage+ "\n")


dict_min_max(dictname, "equidae_min_max.csv")
```

The output of this file contains genus name, species name and the minimum and maximum value. To plot the fossil occurrence of the horses through time and also plot the diversity of the horses through time I used R and ggplot library. The R code for the plots are as follows:

```
library(ggplot2)
equidaes <- read.csv("~/Desktop/eeb-177/
```

```
                    eeb-177-final-project/
                    equidae_min_max.csv", header = F, as.is = T)
names(equidaes) <- c("genus", "species", "minage", "maxage")
library(forcats)
equidaes_occ <- ggplot(equidaes, aes( x = fct_reorder(species,
                                        minage, .desc = T)
                                      , maxage, colour = genus))

#to plot the diversity through time with line range and
#different colors representing different genus.
equidaes_occ <- equidaes_occ + geom_linerange(aes(ymin = minage,
ymax = maxage + 0.5)) + theme(legend.position="none") +
coord_flip() +  theme(axis.text.y = element_text(size=3.5)) +
scale_y_continuous(limits=c(0, 58),
expand = c(0, 0),
breaks=c(0, 10, 20, 30, 40, 50)) +
labs(title = "Equidae Fossil Occurrences", x = "Family", y = "Ma ago")
+ theme(plot.title = element_text(hjust = 0.5, size=22, face = "bold")
, axis.title =element_text(size=20))

equidaes_occ
#installed.packages("dplyr")
library(tidyr)
library(dplyr)
# to plot the diversity
diversity <- equidaes %>% gather(key = type, value = age, minage, maxage)
%>% mutate(count = ifelse(type == "maxage", 1, -1)) %>% group_by(age)
%>% summarise(count = sum(count))  %>% arrange(-age, -count)
%>% mutate(diversity = cumsum(count))

ggplot(diversity, aes(x = age, y = diversity))
+ geom_step() + ggtitle("Horse diversity")
```

I downloaded another file from PBDP for the Plantae fossil occurrences through time. In order to exclude the headers and only get the data for the ones which were identified to the genus level, I used the following python code which uses defaultdict to exclude similar keys in the dictionary.

```
my_file= open("./plant/plant.csv" , encoding = "ISO-8859-15")
#open the file downloaded
my_lines = my_file.readlines()[19:]
#read through lines
output=open("./plant/plant_new2.csv", "w")
for line in my_lines:
  output.write(line)
```

```
my_file.close()

def sp_range(filename):
    file = open(filename,'r')
    my_line = file.readlines()
   #define a new dictionary
    from collections import defaultdict
    species_ranges = defaultdict(list)
    for line in my_line:

        line= line.replace('"', '')
        line= line.replace("'", '')
        species_name= line.split(",")[5]
        sp_or_genus= line.split(",")[6]

        max_int = line.split(",")[10]
        min_int = line.split(",")[11]
        #for only when genus rank is identified
        if sp_or_genus=="genus":

            species_ranges[species_name].append(max_int)
            species_ranges[species_name].append(min_int)

        sorted(species_ranges.keys())

    file.close()
    return (species_ranges)
sp_range("plant_new2.csv")
```

The output of this file contains more than one set of min and max for each genus. I used *min* and *max* function to only have one min and one max for each genus.

```
dictname=sp_range("plant_new2.csv")

def dict_min_max(dictname, outputfile)

    output= open("plant_min_max.csv", "w")
    for keys, values in dictname.items():
        minage = min(values)
        maxage = max(values)
        genus = keys.split(" ")[0]
        species = keys
        output.write(genus + ","+ species+ ","+ minage+
        ","+ maxage+ "\n")
```

```
dict_min_max(dictname, "plant_min_max.csv")
```

The output contains species names and their min and max value which I used in the follwoing R code to plot the diversity of the Plants' fossil through time.

```
library(ggplot2)
plants <- read.csv("~/Desktop/eeb-177/eeb-177-final-project/
                   plant/plant_min_max.csv", header = F, as.is = T)
names(plants) <- c("family", "species","minage", "maxage")

#install.packages("dplyr")
library(tidyr)
library(dplyr)
diversity1 <- plants %>% gather(key = type, value = age, minage, maxage)
%>% mutate(count = ifelse(type == "maxage", 1, -1)) %>% group_by(age)
%>% summarise(count = sum(count))  %>% arrange(-age, -count)
%>% mutate(diversity = cumsum(count))

ggplot(diversity1, aes(x = age, y = diversity)) + geom_step()
+ ggtitle("plant diversity")+ scale_x_continuous(limits=c(0, 60))
```

To compare the diversity for the plants and horses, I used the following R code to plot the two graphs in one page.

```
library(ggplot2)

plants <- read.csv("~/Desktop/eeb-177/eeb-177-final-project/
                   plant/plant_min_max.csv", header = F, as.is = T)
names(plants) <- c("family", "species","minage", "maxage")
#install.packages("dplyr")
library(tidyr)
library(dplyr)
diversity1 <- plants %>% gather(key = type, value = age, minage, maxage)
%>% mutate(count = ifelse(type == "maxage", 1, -1)) %>% group_by(age)
%>% summarise(count = sum(count))  %>% arrange(-age, -count)
%>% mutate(diversity = cumsum(count))
ggplot(diversity1, aes(x = age, y = diversity))
+ geom_step()+ ggtitle("plant diversity")
+ scale_x_continuous(limits=c(0, 60))

equidaes <- read.csv("~/Desktop/eeb-177/eeb-177-final-project/
                     equidae_min_max.csv", header = F, as.is = T)
names(equidaes) <- c("genus", "species", "minage", "maxage")
library(tidyr)
library(dplyr)
```

```
diversity <- equidaes %>% gather(key = type, value = age, minage, maxage)
%>% mutate(count = ifelse(type == "maxage", 1, -1)) %>% group_by(age)
%>% summarise(count = sum(count))  %>% arrange(-age, -count)
%>% mutate(diversity = cumsum(count))
ggplot(diversity, aes(x = age, y = diversity)) + geom_step()
+ ggtitle("Horse diversity")+scale_x_continuous(limits=c(0, 60))

#install.packages("cowplot")
#install.packages("gridExtra")

library(cowplot)
library(gridExtra)

p1<-ggplot(diversity, aes(x = age, y = diversity))
+ geom_step() + ggtitle("Horse diversity")
+scale_x_continuous(limits=c(0, 60))
p2<-ggplot(diversity1, aes(x = age, y = diversity))
+ geom_step()+ ggtitle("plant diversity")
+ scale_x_continuous(limits=c(0, 60))

plot_grid(p1, p2, labels = c("A","B", ncol = 2, nrow = 1 ))
```

In order to find the extinction rate, speciation rate, net diversification rate and longevity of Equidae family, I used the program called PyRate which helps account for uncertainty in estimates of fossil species' origination and extinction dates.(Silvestro *et al.* 2014) For the first step, I cloned the whole program with the following code:

```
git clone https://github.com/dsilvestro/PyRate.git
```

Next, I wrote the following R script to format my data for Equidae.

```
source("~/PyRate/pyrate_utilities.r")

# we need to give the utilities a list of extant species
extant_horses = c("Equus quagga","Equus burchelli","Equus grevyi","Equus zebra","Equus a

# use the extract.ages.pbdb() function in pyrate_utilities to reformat our dataset...
extract.ages.pbdb(file= "eq.csv",extant_species=extant_horses)
```

Next I used the following command to run the pyrate for my data.

```
python ~/PyRate/PyRate.py canid_occ_PyRate.py -n 5000000
```

Results

The extinction rate for Equidae family is almost constant and less than 0.2 from 60 Ma to around 2 Ma. There is a jump in the extinction rate from approximately 0.2 to 1.1 from 2 Ma to present time. The speciation rate is approximately 0.2 from 60 to 22 Ma. The speciation rate decrease to 0.09 after around 22 Ma and stays constant after that time. The net diversification rate is almost constant and positive from 60 to 10 Ma. There is a great decrease in diversification rate after 10 Ma which goes down to approximately -1.0. The longevity decreases from 17 to 7 in 60 to 50 Ma time period. It then goes up to approximately 19 from 30 to 20 Ma. The greatest longevity drop is from 10 Ma to present which decreases to approximately 1.

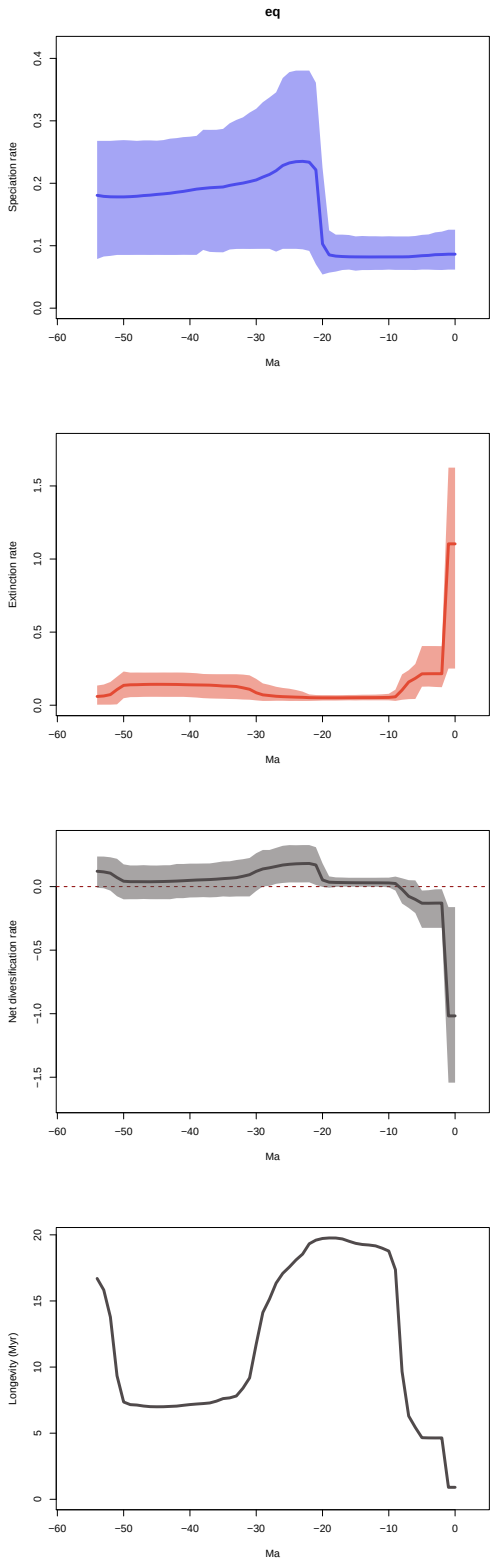

Figure 1: Extiction, speciation, diversification rates and longevity for Equidae through time

The teeth measurement of the Equidae fossils is increasing through time. The measurements increases from approximately 1 to 6 centimeter in length. There is a higher density of fossil collection after 18 Ma to present. Recent fossils show a greater variability of this trait, showing two two major groups of data, approximately 3 and 6 cm.

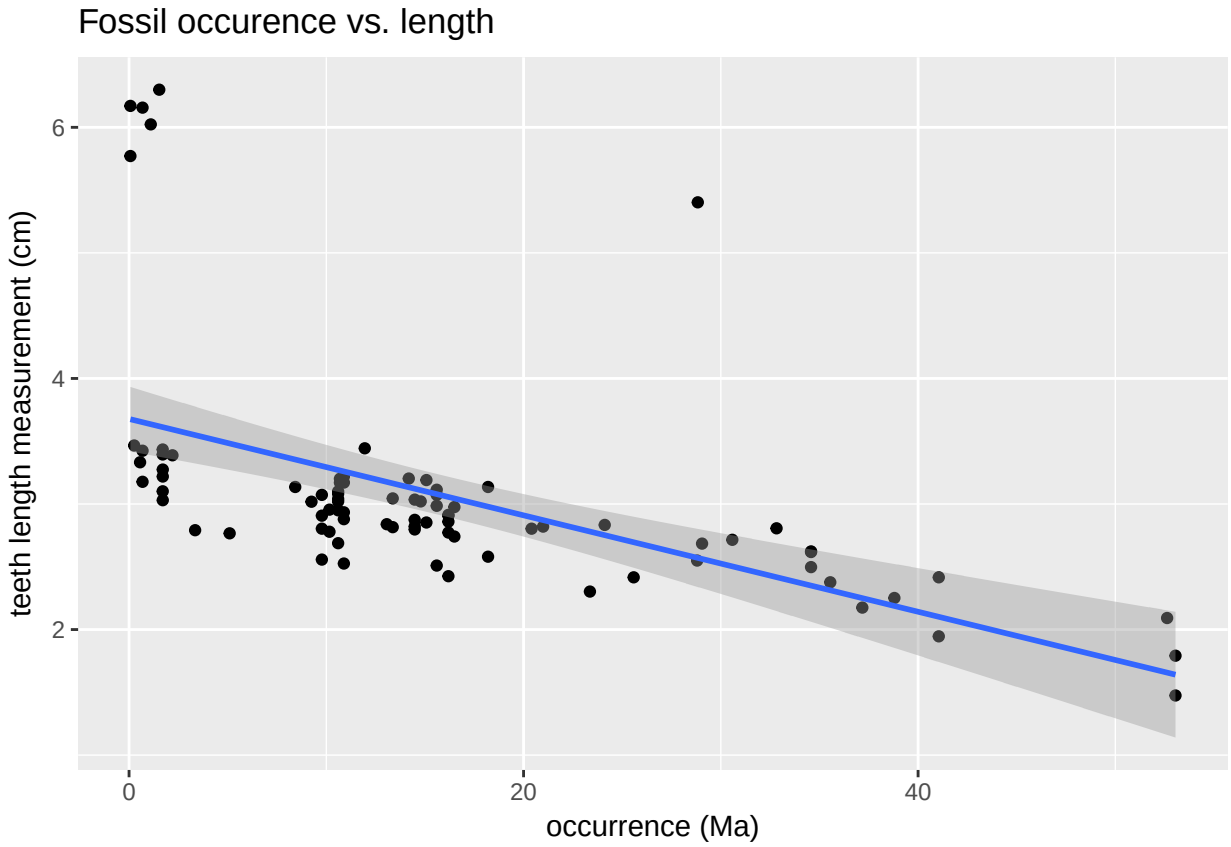

Figure 2-a: Teeth measurement of the Horse Fossils in time. The y axis represents the teeth measurements for the fossils in centimeters.

The fossil occurrences of the Equidae family shows the extensive distribution of the fossils in the world. There is a higher density of the fossils in North America and Europe.

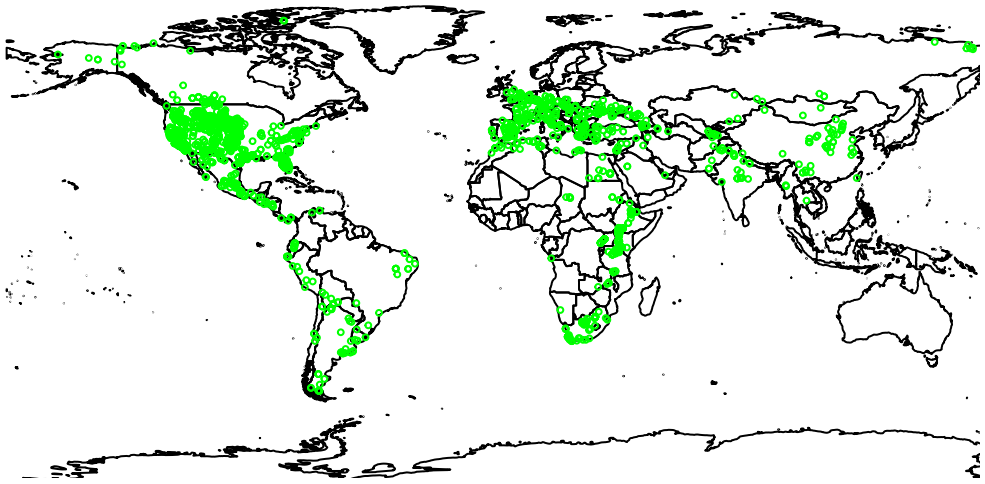

Figure 2: Equidae Fossil Occurrence Locations. Each green spot shows the location of the fossils found in the world

The fossil occurrences for Equidae are extended from 55 Ma to present.

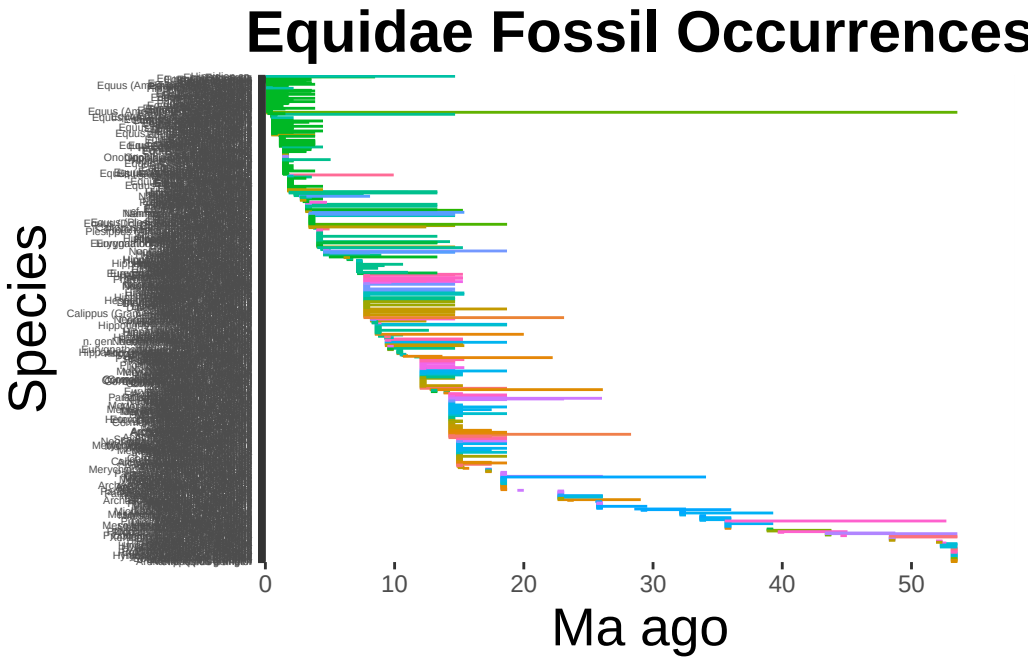

Figure 3: Horse fossil occurrence in time. Each color represents a genus from Equidae. The lines show the start and end time and also length of the presence of each specific species for the horse family.

The diversity of horse family is less than 10 for 60 to 20 Ma time period. There is an increase in diversity from approximately 19 to 10 Ma.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

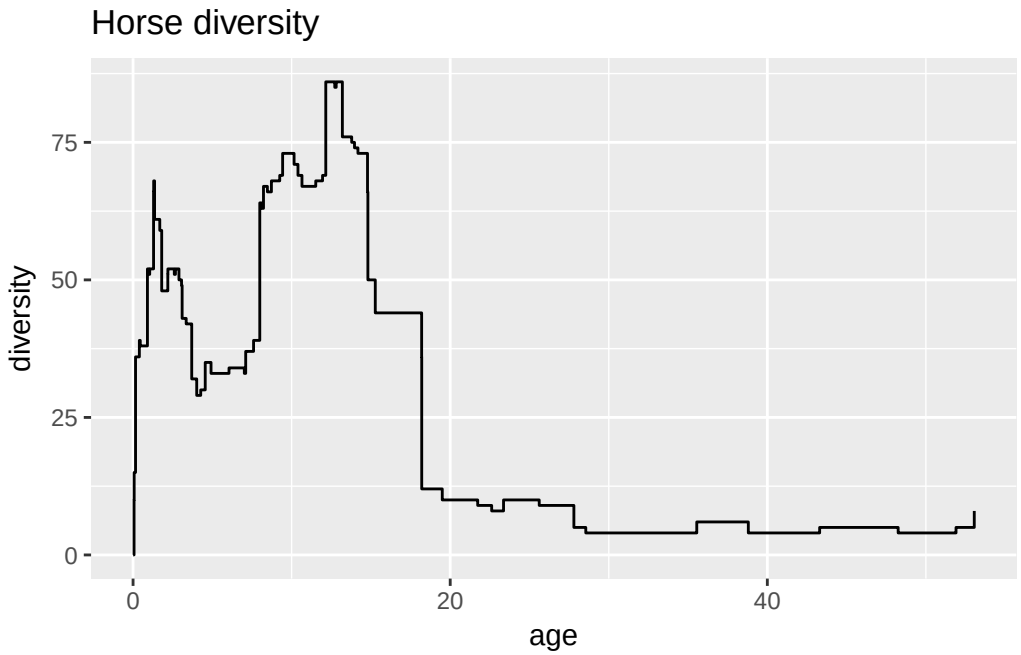

Figure 4: Equidae Diversity in time. The y axis represents the diversity of the species in Equidae through time.

Plantae has its highest diversity from approximately 50 to 40 Ma. The diversity of plants in average is decreasing towards the present time.

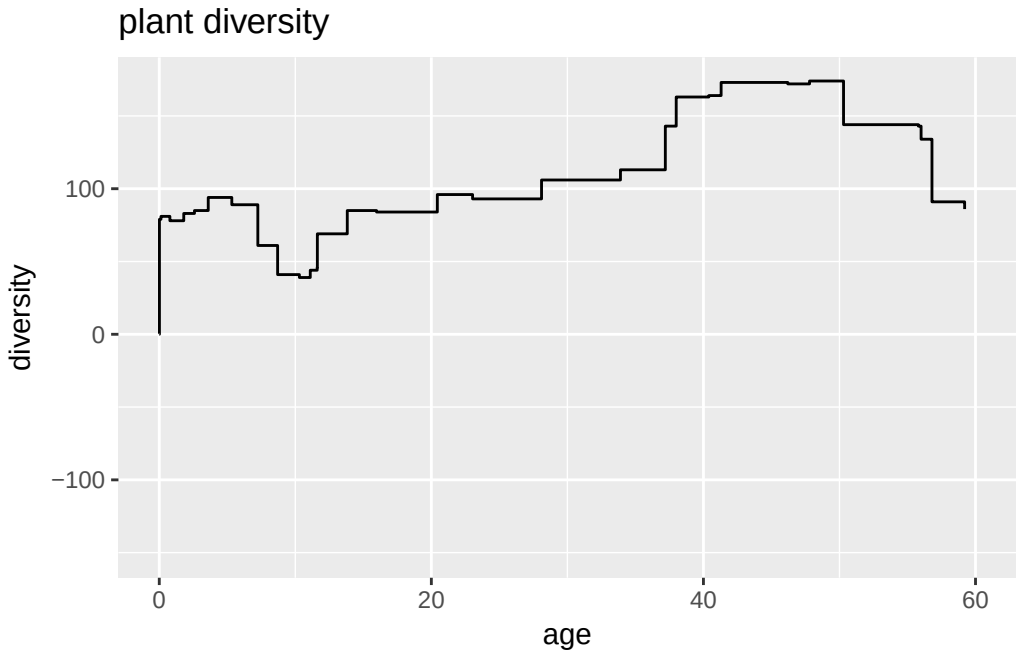

Figure 5: Plantae Diversity in time.

The plant diversity does not show any abvious correlation with the horse diversity.The two plots suggest that there is no correlation between the plant diversity and horse diversity.

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggplot2':
##
##     ggsave

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

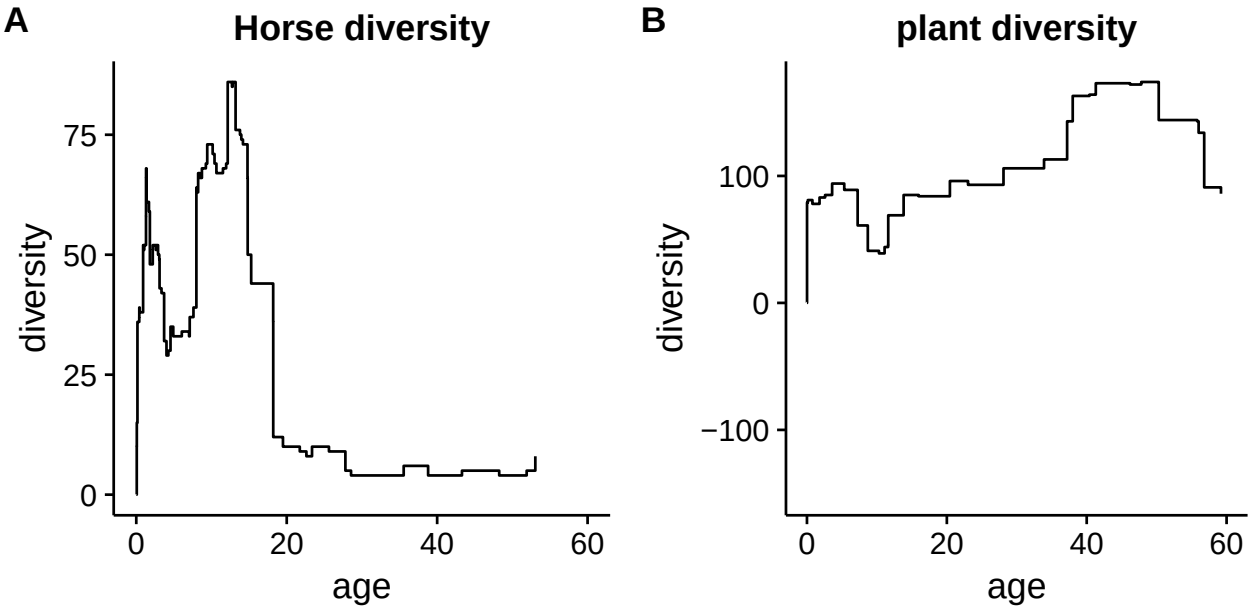

Figure 6: Plantae and Equidae Diversity in time.

Discussion

The diversity of the Horse and plants showed no apparent correlation. The diversity of the horse increases at 19 Ma, however the plant diversity is almost constant at this time. One limitation on this result is that the diversity has been found for all the species existed for plants. However, the food source of horses does not constitudes all of the plant species. As a result, the diversity found for the plants is not a precise comparison for the effect of plant on horse evolution.

Further work needs to be focus on the diversity of different familes of the plants separately. The change in only one or two of the plant families which are the food source of Equidae could show a more precise relation with the horse diversity. For a better understanding, also a study can be done to analyze the horse and plant families specific to a location. For instance, there is a high density of horses in North America. By focusing on only horse fossils found in this location through time and plant fossils specific to this location, a better understanding of the correlation between each specific plant and horse can be gained. One factor for the evolution of Equidae could be their respond to their habitat change, and one important element of this habitat change is the change in plant species.(Strömberg 2006)

In addition, for comparing the diversities, the horse data could be grouped by their traits, for example teeth measurement or body length. The reason is that because different traits evolved at different rates and there is a great variabilty between the traits of organisms in this family. So combining species with all the traits together would reduce the precision of the results.

During the evolutionary history the very first member of this family was *Hyracotherium*, a small dog like forest animal. The food source for this specie was fruits and soft foliage. As we move toward the Oligocene the climate in North America changes and becomes drier and grasses evolve.(Hunt 1995) As the world map of the fossil occurrences(figure 2) shows, there is a high density of these species in North America. This change in food source is one the reasons for evolution of the horse'teeth.(MacFadden 1994) The plot for the teeth measurement(figure 2-a) shows a steady increase of the teeth length as we get closer to present time. The change in teeth morphology helped horses with grinding tougher vegetation. Also the density of the teeth fossils found increases after around 18 Ma. This could be explained by the development of *cement* layer on teeth which increased the preservation rate of these fossils.(Hunt 1995) The jump in the extinction rate from approximately 0.2 to 1.1 from 2 Ma to present time could be explained with the limitations on fossil data. It is harder to find a fossil from a more recent time and this limits our data and effects the results found for this period.

Although no correlation between the horse and plant diversity was found in this study, this does not eliminate the possibility of the effects of the plants on horses. Further study should focus on each horse and plant genera separately and analyze the correlation between their diversities and extinction rates.

Data accessibility:

- All files, figures and functions are available in my github repository: https://github.com/rnojoumi/eeb-174-final-project

References

Hunt, K. (1995). Horse evolution. *TalkOringins Archive*, **304**.

MacFadden, B.J. (1986). Fossil horses from 'eohippus'(Hyracotherium) to equus: Scaling, cope's law, and the evolution of body size. *Paleobiology*, **12**, 355–369.

MacFadden, B.J. (1994). *Fossil horses: Systematics, paleobiology, and evolution of the family equidae*. Cambridge University Press.

MacFadden, B.J. (2005). Fossil horses–evidence for evolution. *Science*, **307**, 1728–1730.

Silvestro, D., Salamin, N. & Schnitzler, J. (2014). PyRate: A new program to estimate speciation and extinction rates from incomplete fossil data. *Methods in Ecology and Evolution*, **5**, 1126–1131.

Strömberg, C.A. (2006). Evolution of hypsodonty in equids: Testing a hypothesis of adaptation. *Paleobiology*, **32**, 236–258.