Final Report in Data Science

Sander Boisen 20114143

Sunday, October 11, 2015

Plant pollinator interactions in the Rocky Mountains

Exploring the effect of systematics on interactions

Background on the data and why I choose it:

Last year I was contacted by a good friend of mine who had met with an american Ph.D. student named Paul. Paul was going to spend some time in Denmark and wanted to start a collaboration with some of the students at AU. Specifically he wanted to work with someone interested in systematics and phylogeny which led to me. For some time we worked with prepping his data for analysis - i.e. making the interaction matrix, gathering the insect phylogeny, extracting the plant phylogeny from a molecular super tree, etc. The insect phylogeny in perticular has caused some trouble, as I build it from scratch. At the order level it was based on published molecular phylogenies based on published molecular phylogenetic studies. Then the family level was added on the background of several other studies. Finally the genus level was seperated based on taxonomi and the distance between genera found from the time tree of life project. Species was added as 0-length polytomies at genus level. All branchlengths were found using the time tree of life projects homepage. Concretely the tree was made by hand in the newick format with a basic text editor. Then it was opened in the Mega6 programs tree viewer, exported as newick and then opened in R. This was necessary as it solves problems with 0-length polytomies and singletons, which was R's main problem with reading the phylogenies. The goal of this project was to explore the effect of phylogeny on in a plant-pollinator community. Another goal was to try to do the same analysis on different taxonomic levels - i.e. species level and family level. In addition I wanted to explore several different meassures quantifying the phylogenetic component of the interactions. This was done to see if there is a distinct effect of the analysis applied, and to familiarize myself with the meassures for future use. I choose to work with this data as I mostly have been spending my spare time working with it, and I saw this course as a possibility to start an actual analysis. My hope was to get the Mixed Linear Phylogenetic Modelling function in the Picante package to work, but it turns out that this meassure is not compatible with our data, as it uses a log10 transformation and the data contains a bunch of zeroes which in turn gives negative infinity when log10 transformed. On the other hand I got to explore the effect of looking at the interactions on different taxonomic levels. Overall no perticular pattern of phylogenetic signal emerged, which is interesting as pollination interactions are one of the classical examples of coevolution and in term a strong phylogenetic signal.

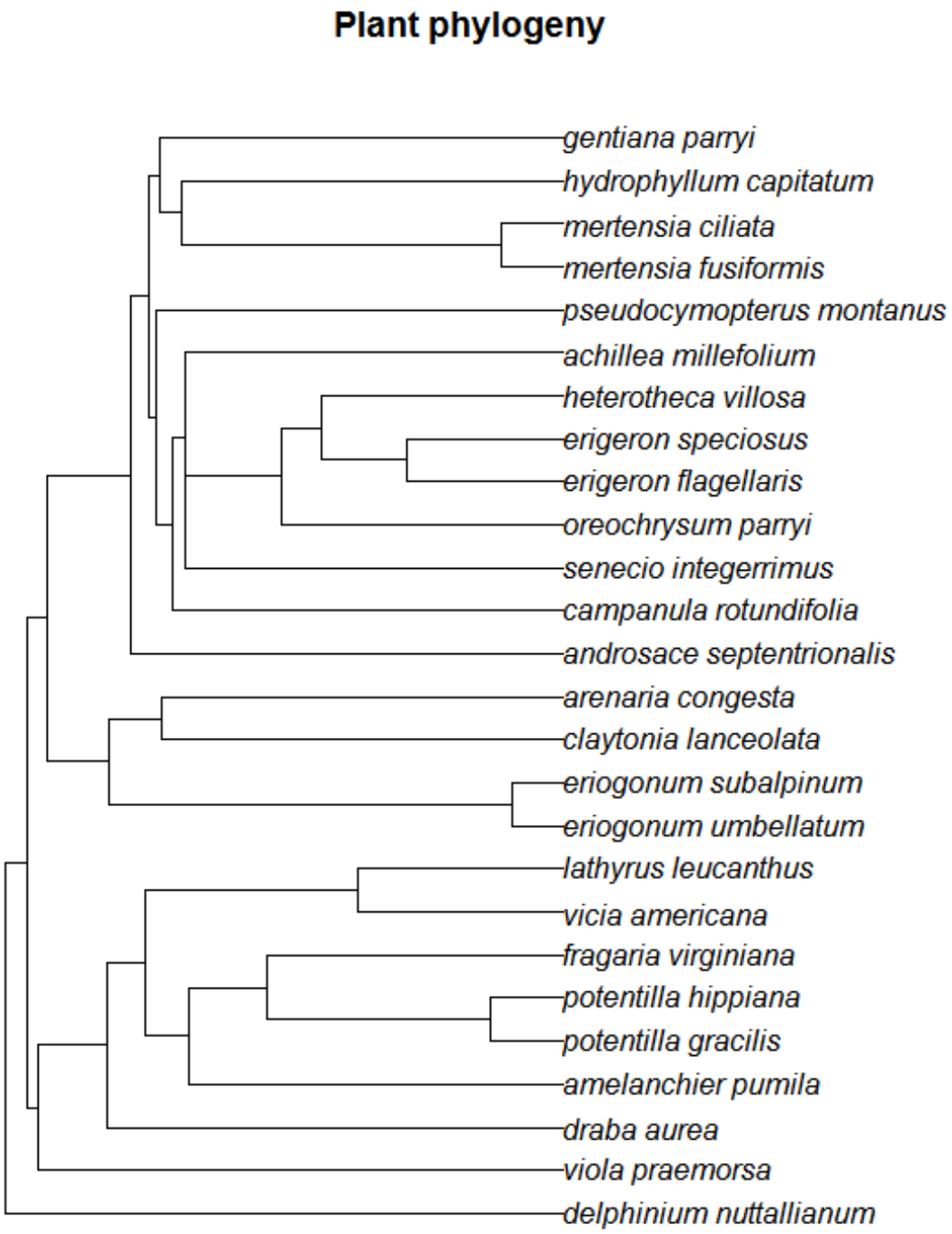
About the site and the plant-pollinator community:

When the snow melts in Gothic, Colorado there is a brief period where conditions are suitable for plant growth and flowering. This ephemeral plant community is very vulnerable to flash frost, as this generally leads to destruction of the flowers. Insect pollinators are also very dependent temperature and flash frost can also mean that some species are not active during the entire season in any given year.

About the manner of data collection:

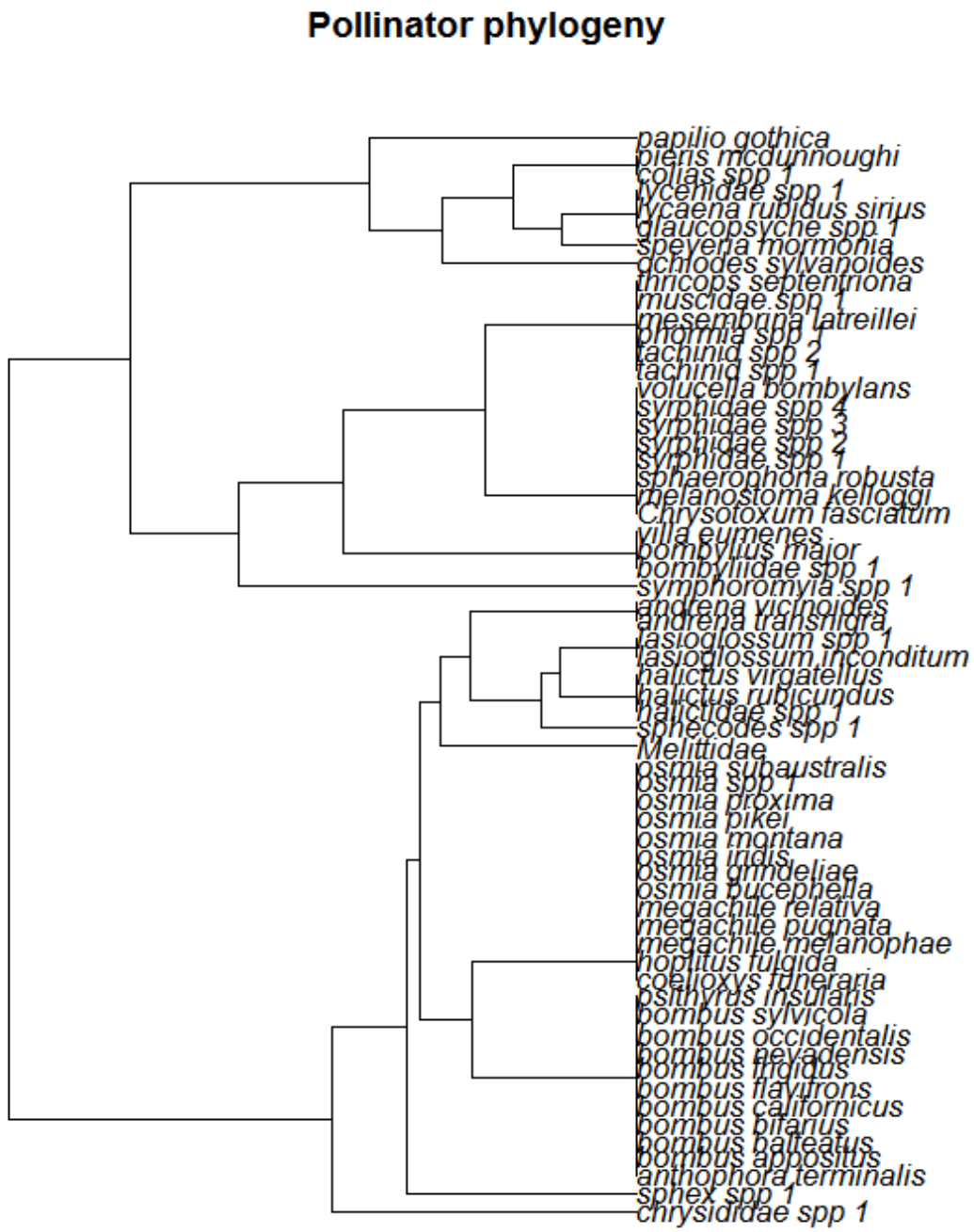
The data was collected by transections. Each transect lasted for one hour and interactions between plant and insect species where noted as they were observed. Insects species may be hard to identify by first glance, and all researchers were therefore trained in the identification of the species known to occur as pollinators in the area. Plant species likewise, and confidence in correct field identification is high. The data collected during the entire season was compiled into an interaction matrix for the entire season.

```
#Plotting in a seperate box because knitr was complaiing for some reason

plot(plant_tree)
title(main = "Plant phylogeny")
```



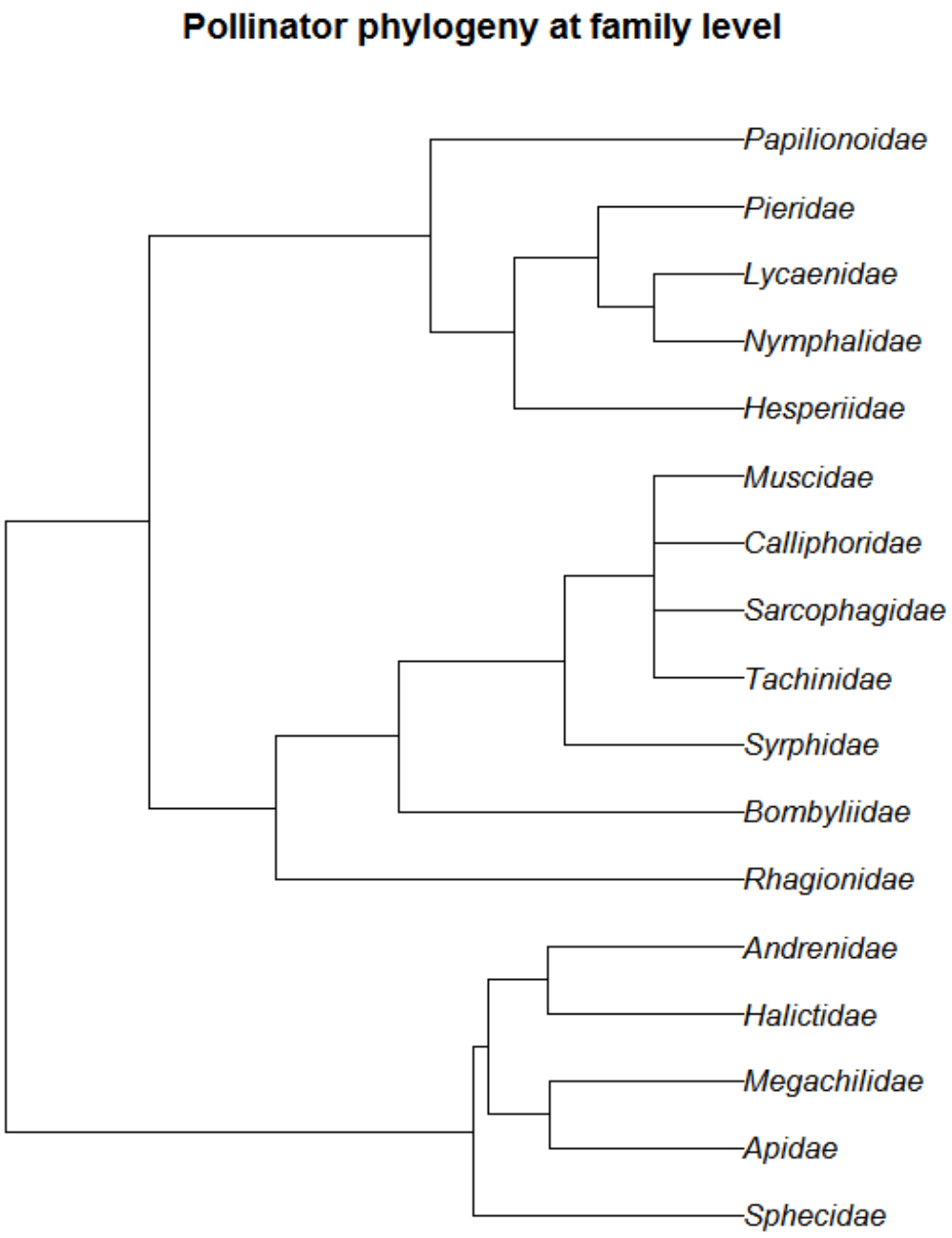
Figur 1 shows a phylogeny of the plants occuring in the study

```
plot(pol_tree)
title(main = "Pollinator phylogeny")
```



Figur 2 shows a phylogeny of all the pollinators occuring in the study

```
plot(pol_fam_tree)
title(main = "Pollinator phylogeny at family level")
```

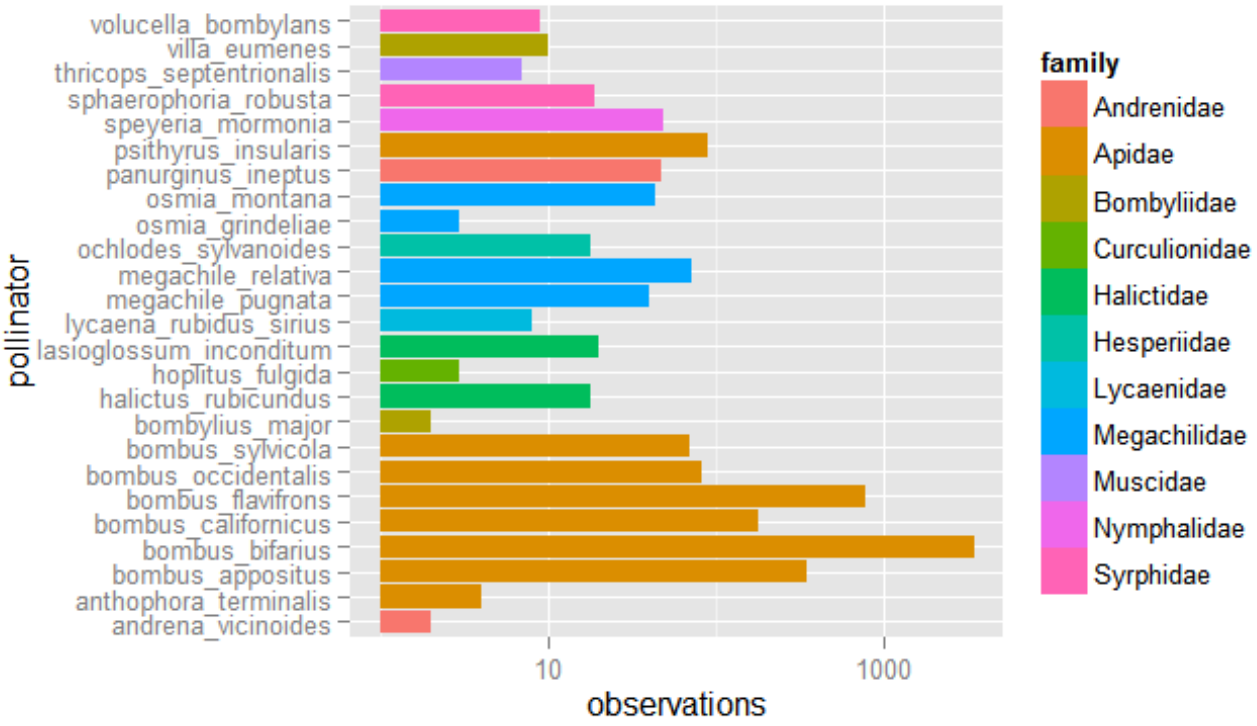


Figur 3 shows a family level phylogeny of the pollinators occuring in the study

About the data set:

The data consisted of 228 observations of plant-pollinator pairs. The data contained all plant-pollinator pairs wether or not an interaction was found. In the study there were 36 species of flowering plants distributed across 17 families. For the pollinators there were 39 species distributed across 14 families and 5 orders. One weaknes is that I had not been supplied with abundance data, so I have no oppurtunity to correct for the effect of abbundance on interaction strength. The entry for the interaction that was observed most frequently looks like so: . One interesting note is that this interaction in between an Asteracean and an Apid, both families that are known to contain a large number of generalists. To illustrate this point consider the following plot for all the interactions of *Heterotheca villosa*:

```
counts%>%
  filter(flower=='heterotheca_villosa')%>%
  filter(observations>0)%>%
  ggplot()+ geom_bar(aes(x=pollinator, y = observations, fill=family,
order=family),
                     stat = 'identity')+
  coord_flip() + scale_y_log10()
```

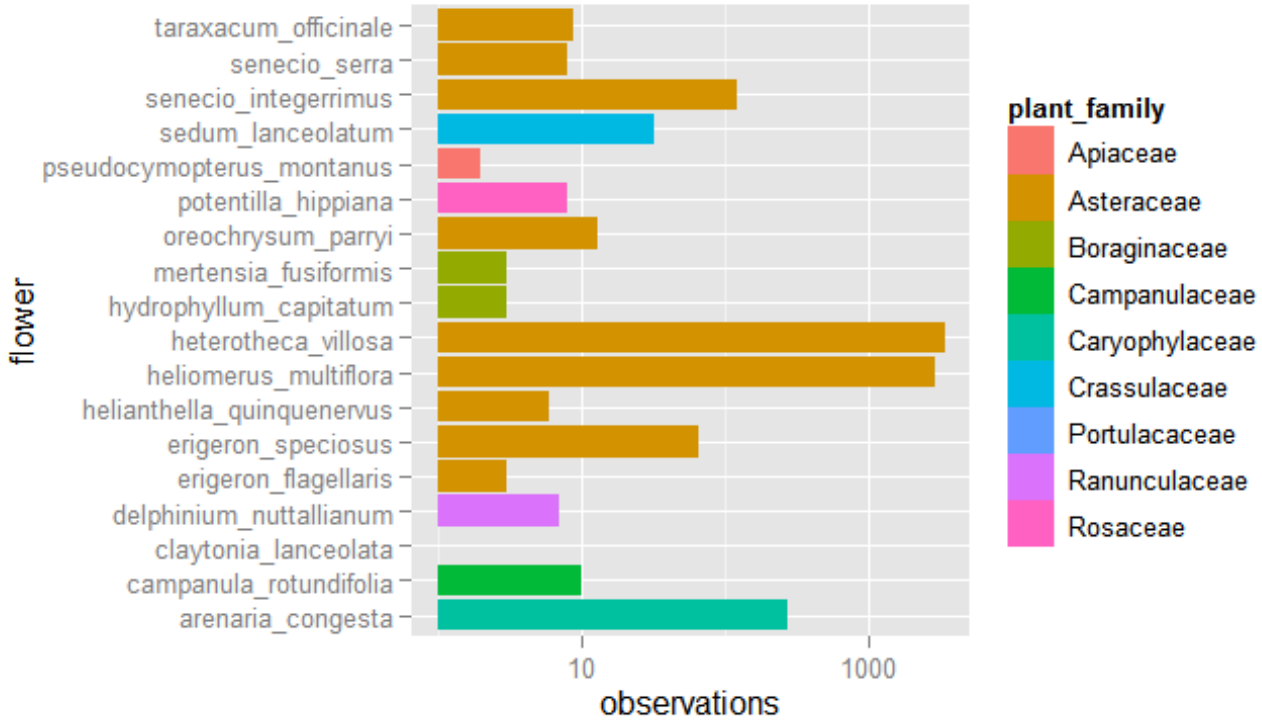


Figur 4 shows the interactions of Heterotheca villosa

As can be seen from the plot one species seem to have a lot more interactions with *Heterotheca villosa* than all the others(Note: the x-axis i scaled log 10), namely *Bombus bifarius*. The following plot shows all the interactions for this species:

```
counts %>%
  filter(pollinator=='bombus_bifarius')%>%
```

```
  filter(observations>0)%>%
  ggplot() + geom_bar(aes(x=flower, y = observations, fill=plant_family),stat =
'identity')+
  coord_flip() + scale_y_log10()
```

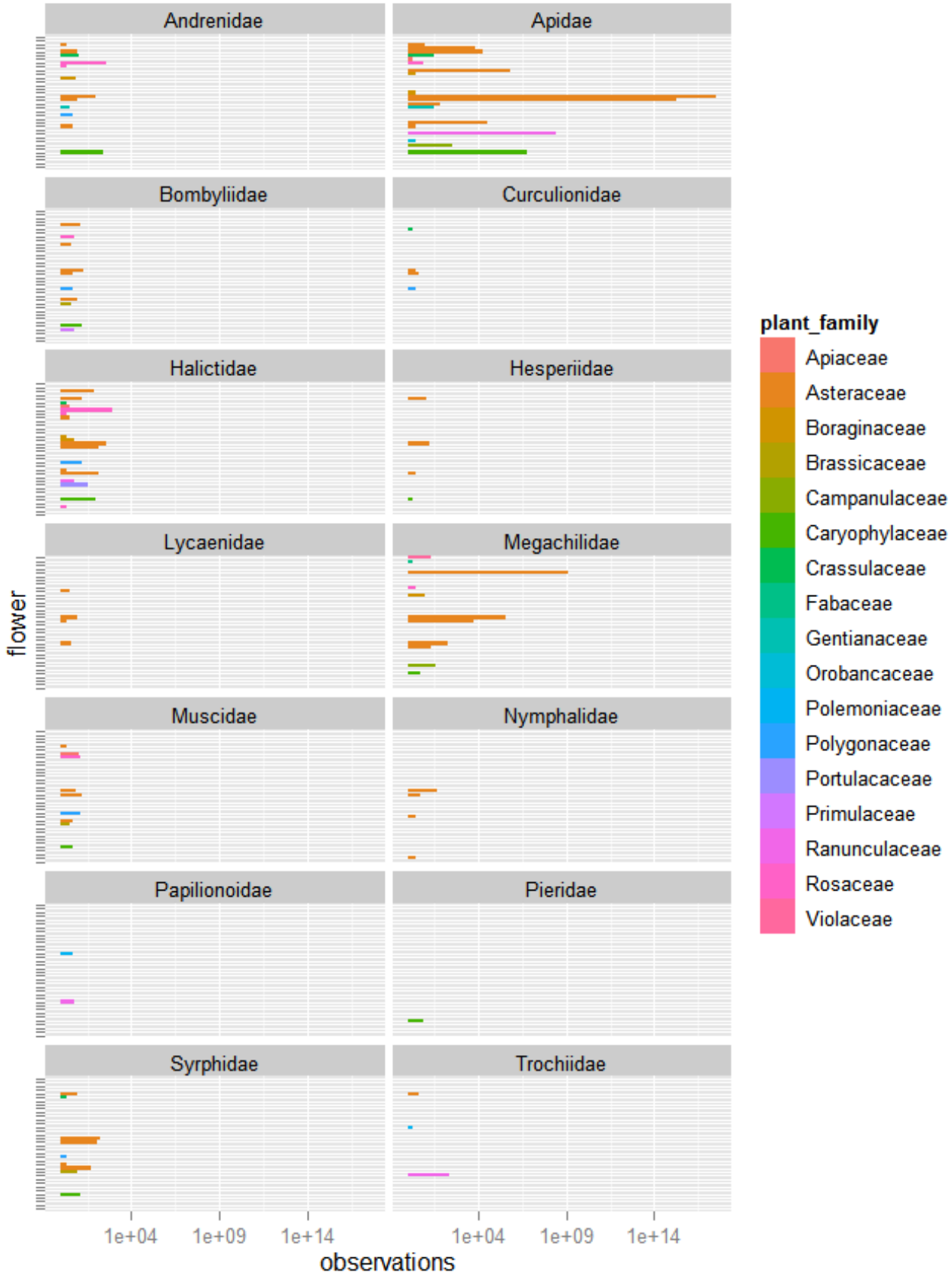


Figur 5 shows the interactions of Bombus bifarius

Fra this plot it is quite obvious the *Bombus bifarius* is a generalist, since it has a very very large number of observed interactions(6961 observations in fact, 0.5953644 of all observations! Further more these interactions are spread across a wide range of plant families.

The following plot is of all interactions distributed across all pollinator families. The colors of the bars represent the plant families as in the plot above.

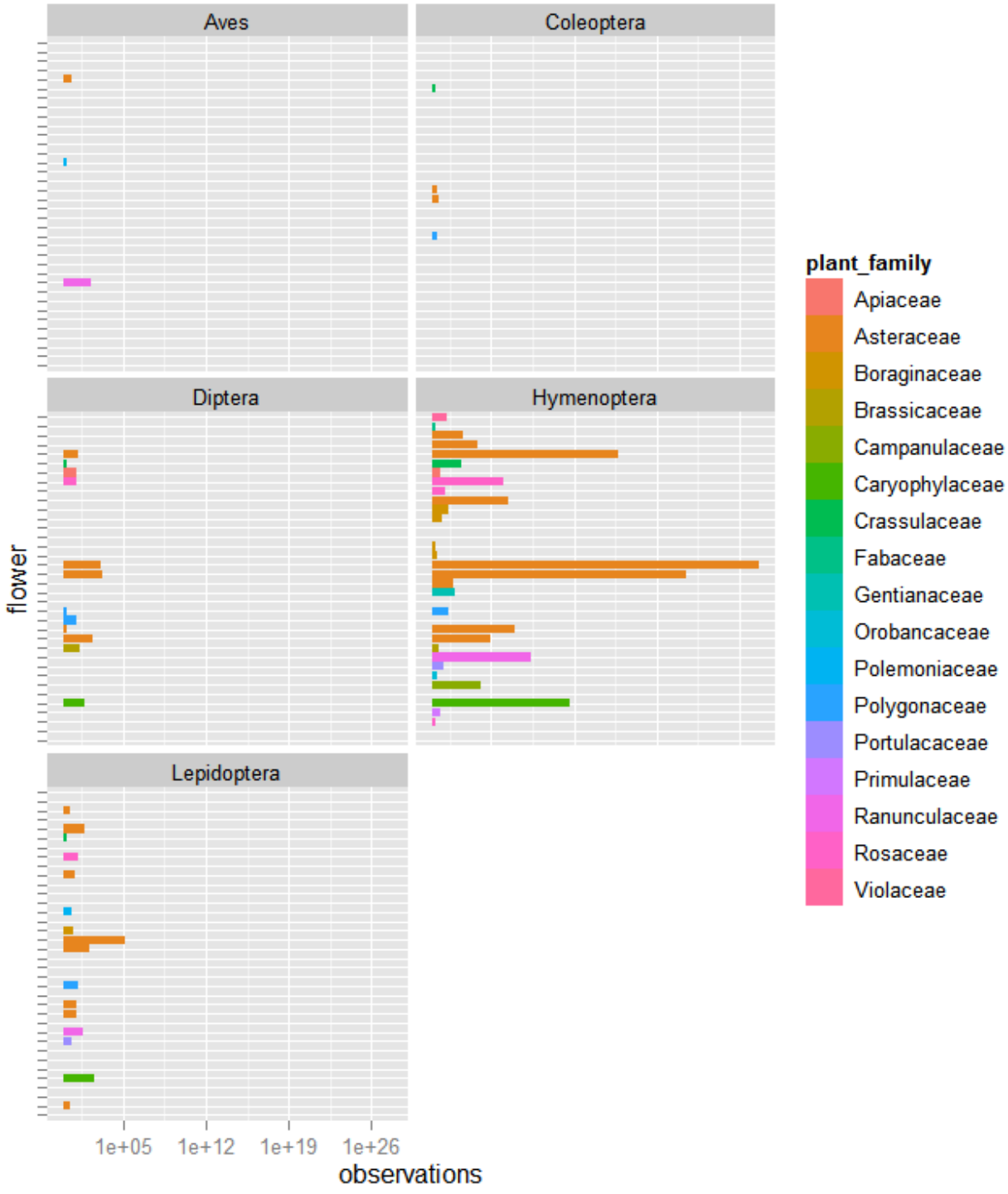
```
counts %>%
  ggplot() + geom_bar(aes(x = flower, y = observations, fill = plant_family ),
stat = 'identity')+
  coord_flip() + scale_y_log10() + theme(axis.text.y = element_blank()) +
facet_wrap(~family, ncol = 2)
```


Andrenidae
Apidae
Bombyliidae
Curculionidae
Halictidae
Hesperiidae
Lycaenidae
Megachilidae
Muscidae
Nymphalidae
Papilionoidae
Pieridae
Syrphidae
Trochiidae
flower
1e+04
1e+09
1e+14
observations
plant_family
Apiaceae
Asteraceae
Boraginaceae
Brassicaceae
Campanulaceae
Caryophylaceae
Crassulaceae
Fabaceae
Gentianaceae
Orobancaceae
Polemoniaceae
Polygonaceae
Portulacaceae
Primulaceae
Ranunculaceae
Rosaceae
Violaceae

Figur 6 shows the total interactions of all pollinator families

This plot shows a distinct skew, as the Apidae has the most interactions by far. Further more Halictidae and Andrenidae have a lot of interactions as well and these are in the same order as Apidae, which seems to suggest that this order has the majority of interactions.

```
counts %>%
  ggplot() + geom_bar(aes(x = flower, y = observations, fill = plant_family ),
stat = 'identity')+
  coord_flip() + scale_y_log10() + theme(axis.text.y = element_blank()) +
  facet_wrap(~order, ncol = 2)
```

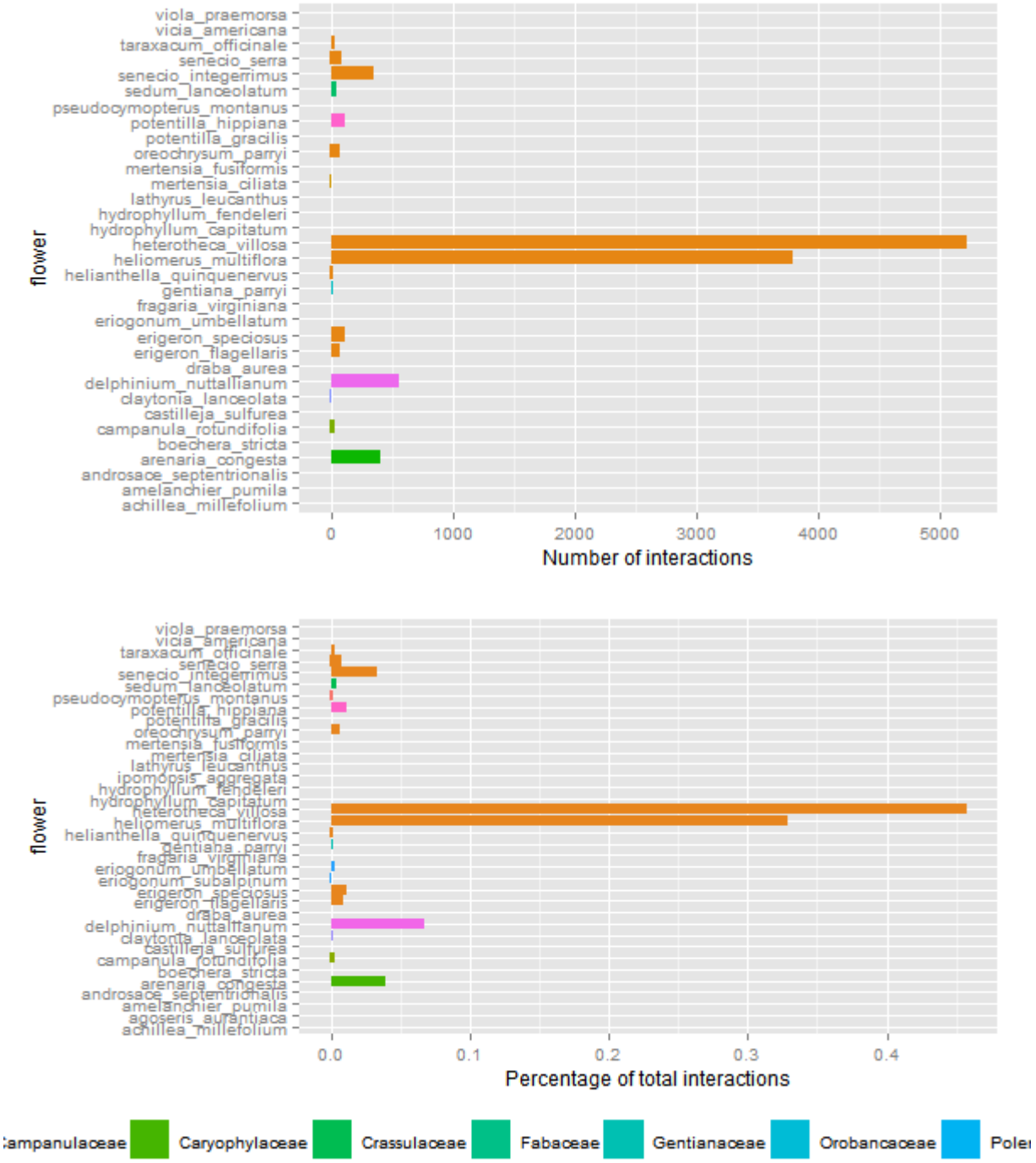


Figur 7 shows the total interactions of all pollinators on the level of order

Again it seems very clear that Hymenoptera has by far the most interactions. This might be an effect of strong interaction with few species.

```
counts %>%
  filter(order == 'Hymenoptera') %>%
  ggplot() +
```

```
  geom_bar(aes(x = flower, y = observations/sum(observations),
               fill = plant_family ),
           stat ='identity', title="Interactions of Hymenoptera") +
coord_flip() +
  scale_y_continuous(name="Number of interactions") +
  theme(legend.position = 'none') -> g1

counts%>%
  ggplot() +
  geom_bar(aes(x = flower,y = observations/sum(observations), fill =
plant_family),
           stat = 'identity') + coord_flip() +
  scale_y_continuous(name="Percentage of total interactions") +
  theme(legend.position = 'bottom',legend.direction='horizontal')->g2
grid.arrange(g1,g2,nrow=2, ncol=1)
```

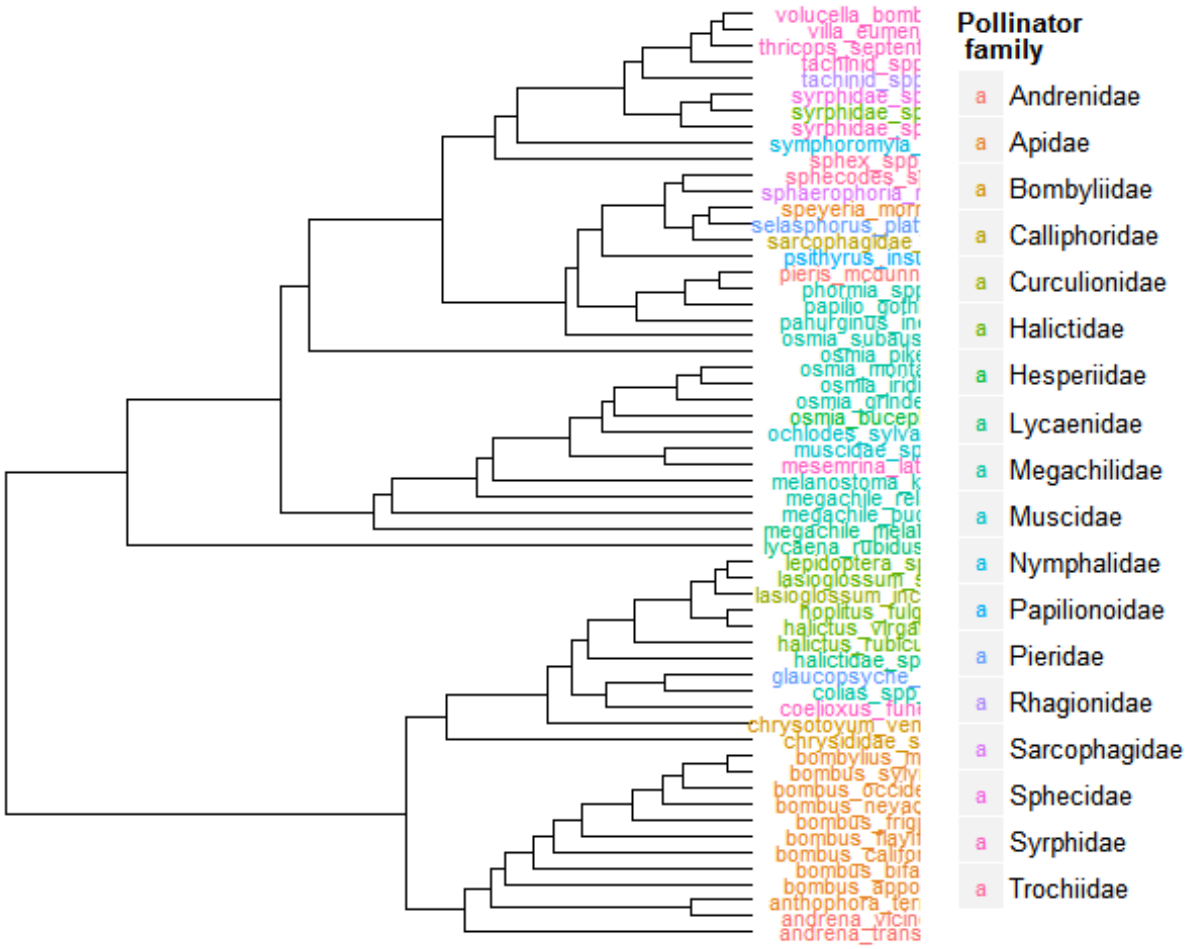


Figur 8 shows the percentage of interactions attributed to Hymenopterans as percentage of the total interactions

As can be seen from this plot *Heterotheca villosa* and *Heliomerus multiflora* are responsible for about 80% of the interactions of the Hymenopterans and these interactions seem to dominate the dataset. One solution for this is to ignore the interaction strength and binarize the interaction matrix. This is rather unfortunate as much of the information in the data will be lost. Another possibility is to remove these two species So what is next? ----------

Now I will try to cluster pollinators using the interactions as a distance meassure. This was not meant to be a result as such, but is another step in exploring the data.

```
# Scale the data but exclude the coloumn with pollinator name
int_scaled <- scale(interactions[,-1])
int_scaled <- cbind(interactions$pollinator, int_scaled)
# Make a distance matrix based on the scaled interaction matrix
int_dist <- vegdist(int_scaled, "euclidean")
#Make a clustering based on the UPGMA
clust <- hclust(int_dist, 'average')
#add the labels as the pollinator names based on the order they appear in the
clustering
clust$labels <- interactions$pollinator[clust$order]

#The raw clustering is not shown, instead a library to draw dendrogram using
ggplot2
#was used to draw the dendrogram. This also means that the branch lengths do
not reflect the
#branch lengths of the original clust, as all tips have the same yend
coordinate

#Extract the data
dhc <- as.dendrogram(clust)
ddata <- dendro_data(dhc, type = "rectangle")

#Make a neat plot
dendro_p <- ggplot(data = segment(ddata))+
  geom_segment( aes(x=x, y=y, xend = xend, yend = yend))+
  coord_flip()+
  geom_text(data = ddata$labels, aes(x = x + 1.75, y = y-5,
            label = interactions$pollinator[label],
            colour = sp_fam$family[label]),
            size = 3, vjust = 2) +
  scale_y_reverse() +
  scale_colour_discrete(name="Pollinator\n family")
dendro_p + theme_dendro()
```



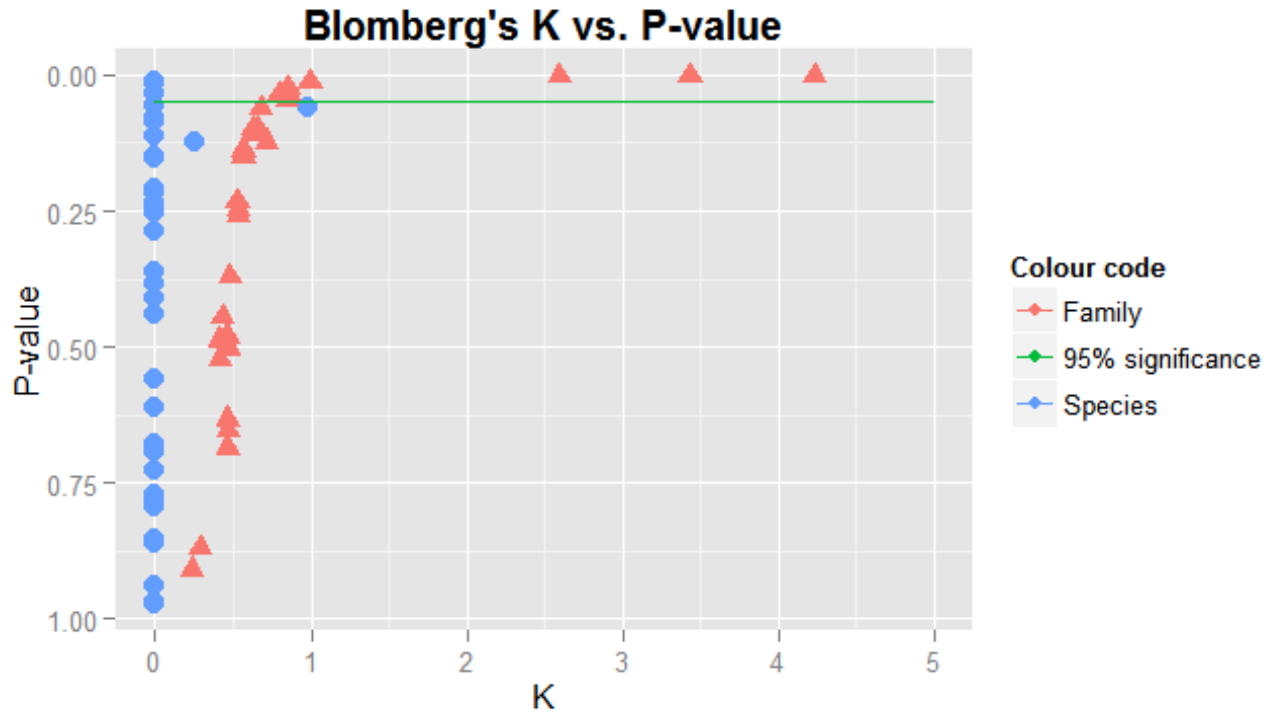
Figur 9 Shows a clustering of all pollinators based on the distance between them based on the interaction matrix

As can be seen from this plot, there is generally a good consensus between the clusters in the hierachical clustering and between the taxonomic groupings. From this plot it looks like there should be at least some phylogenetic signal in the data. In an attempt quantify the phylogenetic signal in the data, interactions where now considered traits to be able to calculate meassures like Bloomberg's k. The matrix was binarized in an attempt to correct for the large portion of the interactions attributed to the Hymenopterans. While this is not optimal, it is a way to work around the lack of abundance data. First the data will be prepped and a family level interaction matrix is made.

```
#Excludes unnecessary coloumns
fam_counts <- fam_counts[, 1:3]

#Create a new matrix for the plant speciez/pollinator families interaction
matr
matr <- matrix(
```

```
  rep(0, length(levels(fam_counts$family))*length(levels(fam_counts$flower))),
  nrow=length(levels(fam_counts$flower)),
  ncol = length(levels(fam_counts$family)))

#Fill the matrix with the sum of counts for each plant species pollinator
family pair
#For loops are used for simplycity and because the runtime for apply is worse
in this case
for(j in seq(1,length(levels(fam_counts$family)))){
  for(i in seq(1,length(levels(fam_counts$flower)))){
    fam_counts %>% filter(family==levels(family)[j]) %>%
      filter(flower==levels(flower)[i]) -> temp
    sum(temp$observations)->matr[i,j]
  }
}

#Convert the matrix to a data frame
fam_int <- as.data.frame(matr)

#Set coloumn and row names
colnames(fam_int)<- levels(fam_counts$family)
rownames(fam_int)<- levels(fam_counts$flower)
#####

int_bin <- apply(interactions[,-1],c(1,2), FUN = function(x) ifelse(x>0,1,0))
int_bin <- as.data.frame(int_bin)
rownames(int_bin) <- interactions$pollinator
colnames(int_bin) <- names(interactions)[-1]

# Sorting the interactions so that the sequence of coloumns follows the
sequence of tips
# in the phylogeny.
index<-lapply(pol_tree$tip.label, function(x) which(x == rownames(int_bin)))
index <- as.numeric(na.omit(index))
int_bin_sort<-int_bin[index,]
int_bin_sort <- na.omit(int_bin_sort)
pol_tree <- prune.sample(t(int_bin_sort), pol_tree)
pol_tree <- multi2di(pol_tree, random = FALSE)
plant_tree <- prune.sample(interactions[,-1], plant_tree)
# Do the same on a family level

# Binarize the interaction matrix
fam_int_bin <- as.data.frame(apply(fam_int, c(1,2), FUN = function(x)
ifelse(x>0,1,0)))
#Remove groups without interactions
index1 = rep(0, length(colnames(fam_int_bin)))
for(i in seq(1,length(colnames(fam_int_bin)))){
  if(sum(fam_int_bin[,i]) == 0){
    index1[i] = i
  }
}
```

```
index1[which(index1>0)] -> index1
fam_int_bin <- fam_int_bin[,-index1]
#Order to reflect the order of the phylogeny
index2 <- lapply(pol_fam_tree$tip.label, function(x) which(x ==
colnames(fam_int_bin)))
index2<-as.numeric(index2)
index2<-na.omit(index2)
fam_int_bin_sort <- fam_int_bin[,index2]

pol_fam_tree <- prune.sample(fam_int_bin_sort, pol_fam_tree)
pol_fam_tree <- multi2di(pol_fam_tree, random = FALSE)



fam_msig <- multiPhylosignal(t(fam_int_bin_sort), pol_fam_tree)
fam_mpd <- na.omit(mpd(fam_int_bin_sort, cophenetic.phylo(pol_fam_tree)))
fam_ses_mpd <- na.omit(ses.mpd(fam_int_bin_sort, cophenetic(pol_fam_tree),
                               null.model='sample.pool'))

multisig <- multiPhylosignal(int_bin_sort, pol_tree)
# I tried applying these meassures, but the data is not fitting for them see
below.
binPSV <- psv(t(int_bin_sort), pol_tree)
binPSR <- psr(t(int_bin_sort), pol_tree)
binPSE <- pse(t(int_bin_sort), pol_tree)
binPSC <- psc(t(int_bin_sort), pol_tree)
binMPD <- mpd(t(int_bin_sort), cophenetic.phylo(pol_tree))
binNRI <- na.omit(ses.mpd(t(int_bin_sort), cophenetic(pol_tree),
null.model='sample.pool'))

#Plotting values for Blomberg's K
ggplot() + geom_point(data = multisig, aes(x = K, y = PIC.variance.P, shape =
'1', colour = '#FF0000', size = 3)) +
  geom_point(data = fam_msig, aes(x = K, y = PIC.variance.P, shape = '2',
colour= '#000000', size = 3))+
  scale_y_reverse(name = "P-value") +
  geom_line(aes(x=seq(0,5), y=0.05, colour = '#00FF00')) +
  scale_shape_discrete(guide='none') +
  ggtitle("Blomberg's K vs. P-value") +
  scale_colour_discrete(name = "Colour code",labels = c("Family","95%
significance","Species")) +
  scale_size_continuous(guide ='none')+theme(plot.title= element_text(size =
15, face = 'bold'))
```

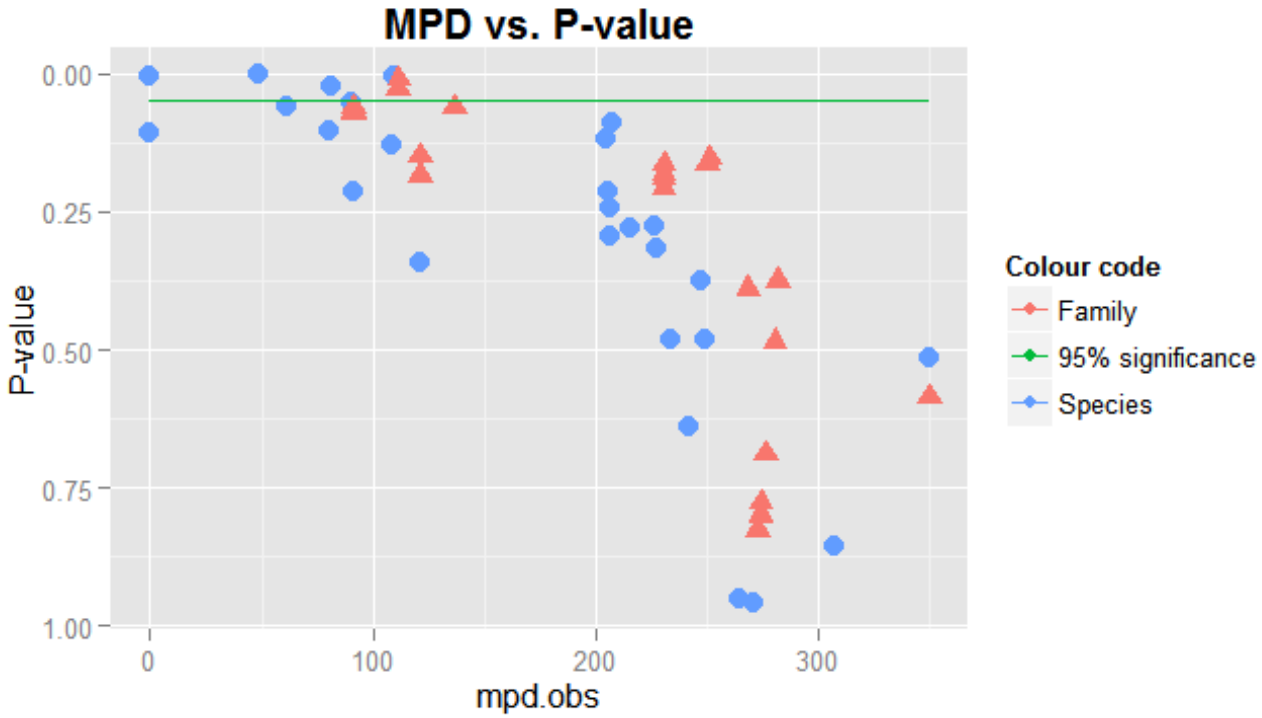


Figur 10 shows Blomberg's K and the P-values of the meassurements.

The above plot shows Blomberg's K (Blomberg, Garland, & Ives, 2003) and the corresponding p-values . Blomberg's K is one if there is no phylogenetic signal. If K is less than one there is less phylogenetic signal than expected under a Brownian motion model. From this plot it is fairly obvious that Blomberg's K is much more nuanced on the family level than on species level. On the other hand it is fairly clear that there is a clear pattern towards very low values of K, and a general absence of phylogenetic signal. This is actually interesting and may indicate that in so small and ephemeral community as this there is a general tendency for generalism. This makes perfect sense as pollinators needs to maximize the amount of plants they can interact with to survive. With regard to the plants this also makes sense, as they need to be able to interact with as many pollinators as possible during the very short pollination season. On the family level there are 3 values above onethat achieved significance and a few below one as well. All in all it seems as if using the family level phylogeny gives a more nuanced picture of Blomberg's K than using the species phylogeny.

```
#Plot  MPD values of insects pollinating plants
ggplot() +
  geom_point(data = binNRI, aes(x = mpd.obs, y = mpd.obs.p, shape = "1", colour
= '#FF0000', size = 3))+
  geom_point( data = fam_ses_mpd, aes(x = mpd.obs, y = mpd.obs.p, shape = "2",
colour = '#000000',size =3)) +
 scale_y_reverse(name = "P-value") +
  geom_line(aes(x=seq(0,max(c(binNRI$mpd.obs, fam_ses_mpd$mpd.obs))), y=0.05,
colour = '#00FF00')) +
  scale_shape_discrete(guide='none') +
  ggtitle("MPD vs. P-value") +
  scale_colour_discrete(name = "Colour code",labels = c("Family","95%
```

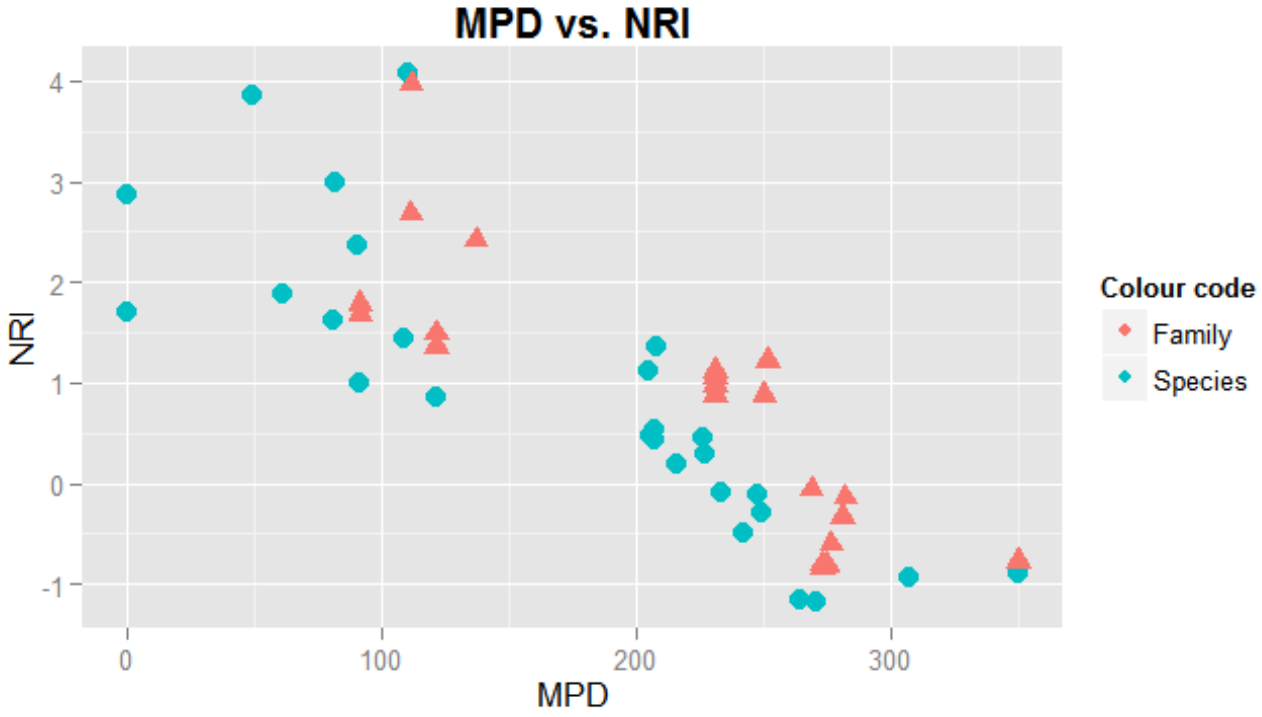
```
significance","Species")) +
  scale_size_continuous(guide ='none')+theme(plot.title= element_text(size =
15, face = 'bold'))
```



Figur 11 shows the MPD and the associated p-values

The Mean Pairwise Distance is defined as exactly that: The mean phylogenetic distance of all pairs of tips in the phylogeny. The smaller the value the higher the average relatedness of the sample. As can be seen the observed MPD increases when looking at the family level compared to the species level. This might be an artifact of the 0-length polytomies in the species level phylogeny, which can result in the the average MPD being 0. As such I believe it to be more robust when applied to family level interaction matrix.

```
ggplot() +
  geom_point(data = binNRI, aes(x = mpd.obs, y = (-1*mpd.obs.z), shape = "1",
colour = '#FF0000', size = 3))+
  geom_point(data = fam_ses_mpd, aes(x = mpd.obs, y = (-1*mpd.obs.z), shape =
"2", colour = '#000000', size = 3))+
 scale_y_continuous(name = "NRI") +
  scale_x_continuous(name = "MPD")  +
  ggtitle("MPD vs. NRI") +
  scale_shape_discrete(guide = 'none') +
  scale_colour_discrete(name = "Colour code",labels = c("Family", "Species")) +
  scale_size_continuous(guide ='none')+theme(plot.title= element_text(size =
15, face = 'bold'))
```



Figur 12 shows NRI vs. MPD

The Net Relatedness Index is defined as the z-standardized MPD multiplied by negative one(Webb, 2000). This is due to the fact that the z-standardized MPD alone would yield higher values with lower relatedness, thus the negative multiplier. As the NRI is directly related to the MPD a general downward trend of NRI is observed when plotted against MPD. An NRI above 1 means that the relatedness of the sample is higher than expected by random. Lower NRI values mean that the sample is less related than expected by chance.

Rounding up:

So while there was little in the way trends or patterns with concern to significance of the results when looking across the collected analysis, the goal of applying several multiple measures at different taxonomic levels was accomplished. It would have been interesting to applying the same measures on the plants using the pollinators as traits. This might give a different result. The binarization of the data set was necessary to account for the large differences in the number of interactions. The data is only part of a larger study that has been running since 2012 with yearly data collection. My data is from the year 2014 only and as the weather varies a lot between years, what is observed in this analysis might not reflect the true nature of the interactions seen across several years. When this is said an analysis of the phenology of flowering plants in the same area showed little or no pattern of phylogenetic signal. Finally most analysis has been done using the Picante package (S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, 2010)which many of the authors cited have contributed to.

References:

Blomberg, S. P., Garland, T., & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution; International Journal of Organic Evolution*, *57*(4), 717–745. doi:10.1111/j.0014-3820.2003.tb00285.x

S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C. O. W. (2010). Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26.

Webb, C. (2000). Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. *The American Naturalist*, *156*(2), 145–155. doi:10.1086/303378