Rarefaction and ASV accumulation plots

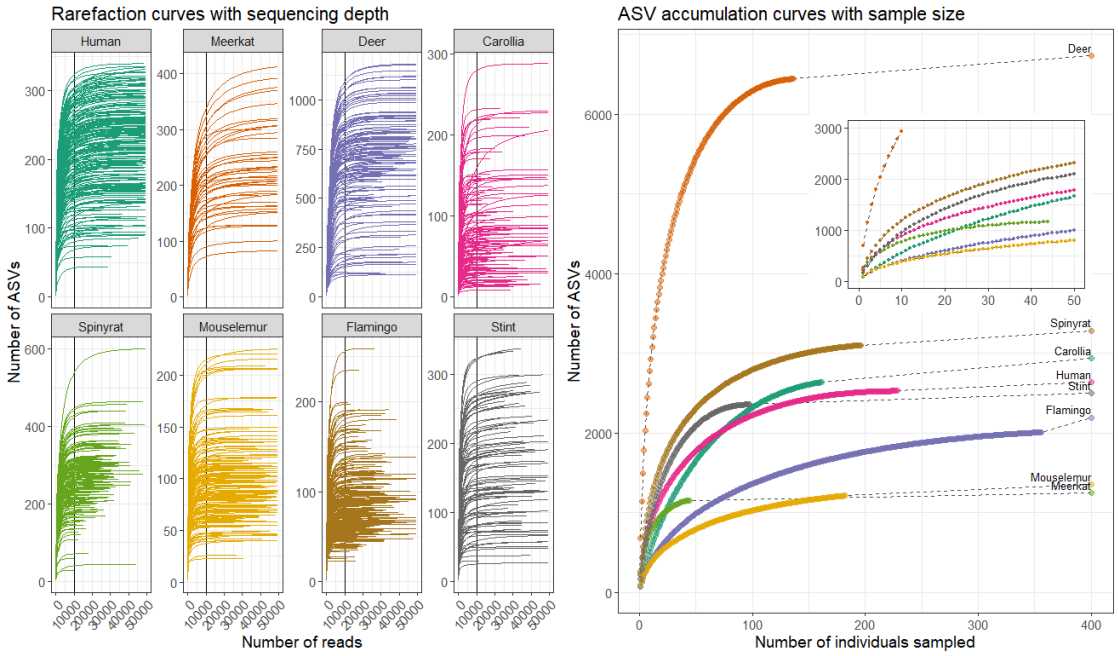
Alice Risely

16/05/2020

Aim of analysis

This code works with a phyloseq object to generate plots that show rarefaction curves (ASV richness with increasing sequencing depth) and ASV accumulation curves (ASV richness with increasing sampling effort). The first shows whether the sequencing depth was sufficient to capture most ASVs in the sample, whilst the second assesses whether the sample size (per group) was sufficient to capture most ASVs within the host population. This is important information and should be explored! In this case, we load a phyloseq object that containts ASV data for eight vertebrate species, and I want to generate raefaction and accumulation curves per species. In other instances you will want to replace species with your group of interest (e.g. site or treatment group).

This is the figure we will generate:



Load data in the form of Phyloseq object

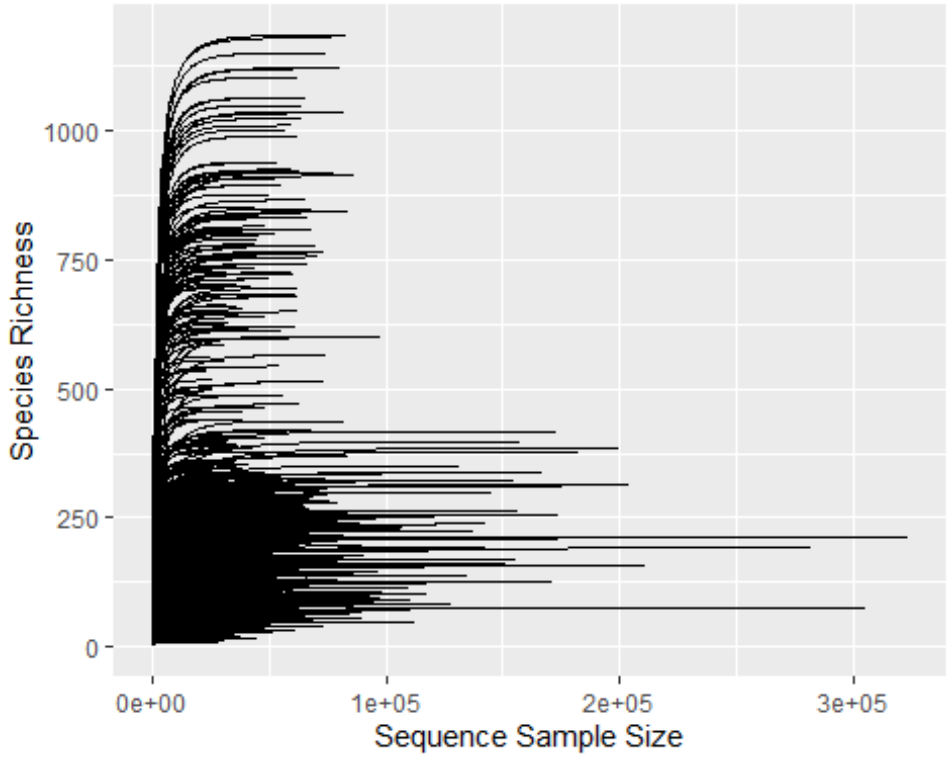
```
merged_8species<-readRDS( "C:\\Users\\risel\\Dropbox\\Sommer postdoc\\PHYLOSEQ OBJECTS\\Rarefaction curves\\merged_8species_unrarefied")

merged_8species
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 19684 taxa and 1400 samples ]
## sample_data() Sample Data:       [ 1400 samples by 17 sample variables ]
## tax_table()   Taxonomy Table:    [ 19684 taxa by 7 taxonomic ranks ]
```

Rarefaction plot

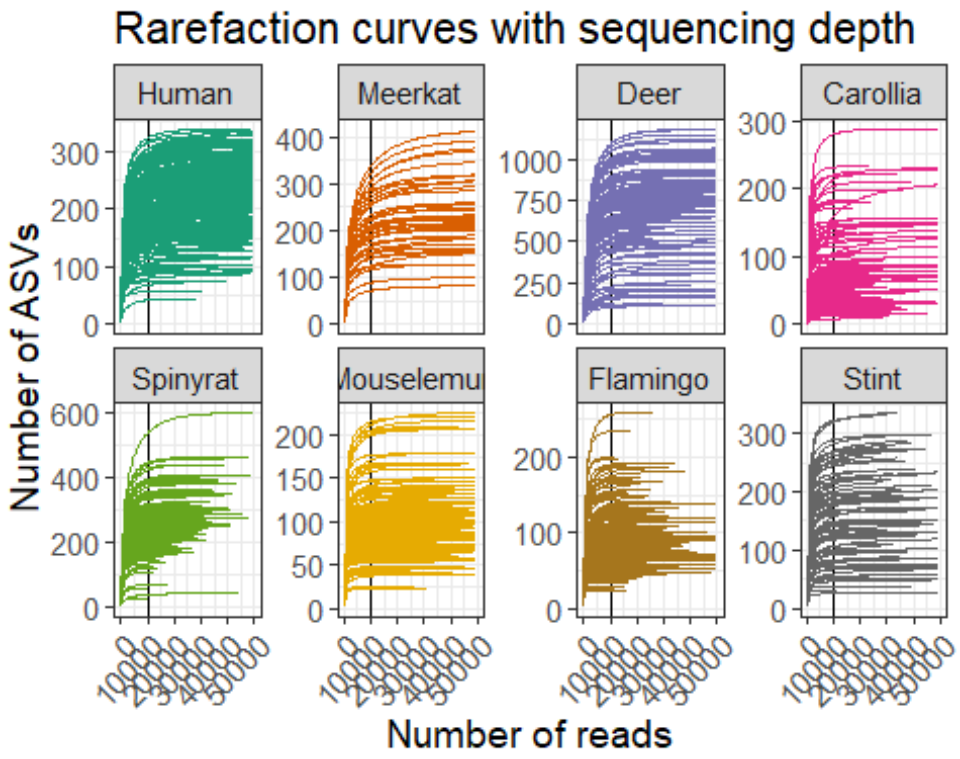
Use the ggrare() function from the ranacapa package

```
rarefaction_fig <- ranacapa::ggrare(merged_8species, step = 500,   se =
FALSE)+
  xlim(c(0,50000))+
  facet_wrap(~Species, scales="free_y", ncol=4)+
  geom_vline(xintercept=10000)+ #I've highlighted 10,000 because that is
where I will rarefy it to later
  theme_bw()+
  theme(legend.position = "none")+
  geom_line(aes(col=Species))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(text=element_text(size=14))+
  ylab("Number of ASVs")+
  xlab("Number of reads")+
  scale_color_brewer(palette = "Dark2")+
  ggtitle("Rarefaction curves with sequencing depth")
```




```
rarefaction_fig
```

```
## Warning: Removed 19506 row(s) containing missing values (geom_path).

## Warning: Removed 19506 row(s) containing missing values (geom_path).
```



Above is the bog standard plot you get with all samples shown together (which is automatically generated with ggrare()), but the object 'rarefaction_fig' is the final rarefaction figure.

ASV accumulation plot

This plot is a bit more complex to code, because you want to treat each species seperately. For the rarefaction figure above, each sample is treated seperately and it doesn't matter which species it belongs to. Because I have eight species I want to repeat this for, it is easier to create a loop and that do them all manually.

The next step codes the for loop, where data for each species is subsetted in turn and an ASV accumulation curves is calculated for each, using vegan's specaccum() function. This randomly picks out samples and calculated how many new ASVs are each added to the total pool each time. We repeat this 500 times so that any biases in sample order are overcome.

In addition, the second part of the loop is estimating the total number of ASVs in the ASV pool; Ie, it extrapolates the curve to calculate how many ASVs we've probably missed had we sampled more. This is important to demonstrate our sample sizes are high enough to capture most of the AVS within the host population.

```
SAClist<-list() #make an enpty list
```

```
uniq <- unique(sample_data(merged_8species)$Species) # make a list of species
to subset sequentially

for (i in 1:length(uniq)){ #for species i

  data_1<-subset_samples(merged_8species, Species == uniq[i]) #subset the
phyloseq object for species i
  data_1<-prune_taxa(taxa_sums(data_1)>0, data_1)  #remove any traces of taxa
that are no longer present in dataset
  data_1_matrix<-data.frame(t(data.frame(otu_table(data_1)))) #transpose the
OTU table
  data_1_specaccum<-vegan::specaccum(data_1_matrix, method="random",
permutations = 500) #apply specaccum()

  ## the output is in list form, so we need to make this into a dataframe

  sac_df<- data_1_specaccum$sites ##sites = samples
  sac_df<-data.frame(sac_df)
  names(sac_df)[1]<-"Site"
  sac_df$Richness <-  data_1_specaccum$richness #import ASV richness to
dataframe
  sac_df$SD <-  data_1_specaccum$sd #import the standard deviation

  ## this next step estimates the TOTAL number of ASVs in the ASV pool.

  sac_total_estimated<-vegan::specpool(data_1_matrix) #estimates total ASV
pool from our otu matrix geenrated above
  sac_df$Total <- sac_total_estimated$boot ##add this to our dataframe
  sac_df$Species <- as.character(sample_data(data_1)$Species[1]) #add species
name, for when we combine dataframes for all species
  SAClist[[i]]<-sac_df #add this dataframe as an element in the empty list
and repeat for the next species

}
```

Now lets look at our final list and combine all those dataframes:

```
names(SAClist)<-uniq #name elements of the list by species
str(SAClist)
```

```
## List of 8
##  $ Carollia  :'data.frame':  161 obs. of  5 variables:
##   ..$ Site    : int [1:161] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:161] 75.6 140.1 201.7 259.2 310.4 ...
##   ..$ SD      : num [1:161] 57.7 71.7 86.6 96.8 105.8 ...
##   ..$ Total   : num [1:161] 2936 2936 2936 2936 2936 ...
##   ..$ Species : chr [1:161] "Carollia" "Carollia" "Carollia" "Carollia"
...
##  $ Flamingo  :'data.frame':  356 obs. of  5 variables:
##   ..$ Site    : int [1:356] 1 2 3 4 5 6 7 8 9 10 ...
```

```
##   ..$ Richness: num [1:356] 85.1 139.8 184 221.6 253.6 ...
##   ..$ SD      : num [1:356] 39.6 52 57.9 61.7 64.4 ...
##   ..$ Total   : num [1:356] 2191 2191 2191 2191 2191 ...
##   ..$ Species : chr [1:356] "Flamingo" "Flamingo" "Flamingo" "Flamingo"
...
##  $ Deer      :'data.frame':  136 obs. of  5 variables:
##   ..$ Site    : int [1:136] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:136] 677 1126 1467 1752 2006 ...
##   ..$ SD      : num [1:136] 237 263 251 248 233 ...
##   ..$ Total   : num [1:136] 6730 6730 6730 6730 6730 ...
##   ..$ Species : chr [1:136] "Deer" "Deer" "Deer" "Deer" ...
##  $ Human     :'data.frame':  228 obs. of  5 variables:
##   ..$ Site    : int [1:228] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:228] 216 349 450 535 610 ...
##   ..$ SD      : num [1:228] 64.2 63 67 64 63.9 ...
##   ..$ Total   : num [1:228] 2637 2637 2637 2637 2637 ...
##   ..$ Species : chr [1:228] "Human" "Human" "Human" "Human" ...
##  $ Meerkat   :'data.frame':  44 obs. of  5 variables:
##   ..$ Site    : int [1:44] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:44] 231 352 442 505 561 ...
##   ..$ SD      : num [1:44] 78.1 83.2 84.3 86.4 82.2 ...
##   ..$ Total   : num [1:44] 1247 1247 1247 1247 1247 ...
##   ..$ Species : chr [1:44] "Meerkat" "Meerkat" "Meerkat" "Meerkat" ...
##  $ Mouselemur:'data.frame':  182 obs. of  5 variables:
##   ..$ Site    : int [1:182] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:182] 96.7 150.6 193.8 226.9 256.4 ...
##   ..$ SD      : num [1:182] 40.4 56.1 64 70.9 75.9 ...
##   ..$ Total   : num [1:182] 1350 1350 1350 1350 1350 ...
##   ..$ Species : chr [1:182] "Mouselemur" "Mouselemur" "Mouselemur"
"Mouselemur" ...
##  $ Spinyrat  :'data.frame':  196 obs. of  5 variables:
##   ..$ Site    : int [1:196] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:196] 261 437 582 700 805 ...
##   ..$ SD      : num [1:196] 81.9 89 89.7 94.1 94.6 ...
##   ..$ Total   : num [1:196] 3284 3284 3284 3284 3284 ...
##   ..$ Species : chr [1:196] "Spinyrat" "Spinyrat" "Spinyrat" "Spinyrat"
...
##  $ Stint     :'data.frame':  97 obs. of  5 variables:
##   ..$ Site    : int [1:97] 1 2 3 4 5 6 7 8 9 10 ...
##   ..$ Richness: num [1:97] 166 297 413 513 600 ...
##   ..$ SD      : num [1:97] 81.3 95.4 96.4 98 99.7 ...
##   ..$ Total   : num [1:97] 2497 2497 2497 2497 2497 ...
##   ..$ Species : chr [1:97] "Stint" "Stint" "Stint" "Stint" ...
```

```
sac_df_all<-do.call(rbind, SAClist) #rbind all our 8 dataframes together

head(sac_df_all) #final dataframe we use to generate the second figure
```

```
##            Site Richness        SD    Total  Species
## Carollia.1    1   75.630  57.71279 2935.775 Carollia
```

```
## Carollia.2    2  140.106  71.74763 2935.775 Carollia
## Carollia.3    3  201.672  86.59919 2935.775 Carollia
## Carollia.4    4  259.152  96.81485 2935.775 Carollia
## Carollia.5    5  310.410 105.84905 2935.775 Carollia
## Carollia.6    6  363.280 108.80895 2935.775 Carollia
```

We could plot the ASV acculumation curves now, but I want to include the the total estimated ASV pool in the figure too, to show that our sample sizes for each species have sufficiently captured almost all ASVs in the host population. TO do this, I just add 8 extra lines to the data frame, withe the species total

```
species_totals<-sac_df_all %>% distinct(Total, .keep_all = T) #subset just
the eight distinct estimated totals per species

species_totals[,1]<-400 #here we put 400 just because we want a number that
is larger than the largest samples size (flamingo)
species_totals[,2]<-species_totals$Total
species_totals[,3]<-NA
species_totals[,4]<-NA

head(species_totals)

##   Site Richness SD Total    Species
## 1  400 2935.775 NA    NA   Carollia
## 2  400 2191.157 NA    NA   Flamingo
## 3  400 6730.058 NA    NA       Deer
## 4  400 2636.761 NA    NA      Human
## 5  400 1247.066 NA    NA    Meerkat
## 6  400 1350.205 NA    NA Mouselemur

sac_df_fig<-rbind(sac_df_all, species_totals) #combine
```

Second figure

Here I've also generated the subplot in the opening figure to show in more the rate of ASV accumulation over the first 50 samples randomly selected. It is optional!

```
#subplot

sub<-ggplot(sac_df_fig, aes(x = Site, y = Richness, group =
Species))+geom_line(alpha=0.7, linetype = "dashed")+
  geom_point( aes(col=Species), size = 1)+theme_bw()+
  xlab("Number of individuals sampled")+
  ylab("Number of ASVs")+
  theme(text=element_text(size=14))+
  theme(axis.title.x=element_blank(),axis.title.y=element_blank())+
  theme(legend.position = "none")+ scale_color_brewer(palette = "Dark2")+
  ylim(c(0,3000))+xlim(c(0,50))


#main plot
```

```
sac_fig<-ggplot(sac_df_fig, aes(x = Site, y = Richness, group = Species))+
  geom_line(alpha=0.7, linetype = "dashed")+
  geom_point( aes(col=Species), size = 2, alpha = 0.5)+theme_bw()+
  xlab("Number of individuals sampled")+
  ylab("Number of ASVs")+
  theme(text=element_text(size=14))+
  theme(legend.position = "none")+
  scale_color_brewer(palette = "Dark2")+
  annotation_custom(ggplotGrob(sub), xmin = 150, xmax=400, ymin = 3500, ymax
= 6000)+
  geom_dl(aes(label = Species), method = list("last.points", cex = 0.8, vjust
= -0.4, hjust = 1))+
  ggtitle("ASV accumulation curves with sample size")

## Warning: Removed 1054 row(s) containing missing values (geom_path).

## Warning: Removed 1054 rows containing missing values (geom_point).

sac_fig
```