Exploring the Ecological Database of World's Insect Pathogens

Tad Dallas

This vignette is designed to introduce you to the EDWIP data contained within the `insectDisease` data package. This serves to reproduce the figures presented in the manuscript, as well as to go over some of the data processing steps and idiosyncracies of this data resource.

load required packages for vignette

```
library(insectDisease)
library(plyr)
library(dplyr)
library(corrplot)
knitr::opts_chunk$set(fig.width=6, fig.height=6)
```

Loading the data

Each data product (see the `R` folder, `man` folder, or the package README for more information on each data product) can be loaded into the workspace using the `data()` function. Here, we'll load the four main data products, corresponding to 1) nematode-insect associations, 2) virus-insect associations, 3) non-viral pathogen - insect associations, and 4) all the negative associations (attempted and failed infections).

```
data(nematode)
data(viruses)
data(nvpassoc)
data(negative)
```

We can explore the structure of these data using R functions like `head`, `tail`, `table`, and `summary`.

By example, we can see that `head(nematode)` provides some detail on the nematode association data, including host species (`HostSpecies`), nematode species (`PathogenSpecies`), and other relevant taxonomic and infection information, including the country where the interaction was recorded, the soil type, and the host tissue and stage infected.

```
head(nematode)
```

```
##   ERNnem        HostSpecies       PathogenSpecies PathogenStrain StageInfected
## 1     32 Bibio ferruginatus   Steinernema affinis not maintained         Larva
## 2     33    Bibio hortulana   Steinernema affinis not maintained         Larva
## 3     34  Dilophus vulgaris   Steinernema affinis not maintained         Larva
## 4     35   Helina duplicata   Steinernema affinis not maintained         Larva
## 5     36      Anomala dubia Steinernema arenarium         Riazan         Larva
## 6     37      Anomala dubia Steinernema arenarium        Voronez         Larva
##   TissueInfected FieldOrLab Country         SoilType  AssociatedBacterium
```

```
## 1       Hemocoel      Field Denmark grain and others Xenorhabdus bovienii
## 2       Hemocoel      Field Denmark grain and others Xenorhabdus bovienii
## 3       Hemocoel      Field Denmark grain and others Xenorhabdus bovienii
## 4       Hemocoel      Field Germany   sand / pasture Xenorhabdus bovienii
## 5       Hemocoel      Field  Russia             <NA>      Xenorhabdus sp.
## 6       Hemocoel      Field  Russia             <NA>      Xenorhabdus sp.
##   IntermediateHost CreationDate ModificationDate    Group HostTaxID HostGenus
## 1             none   1997-05-08       2000-09-21 Nematode        NA      <NA>
## 2             none   1997-05-12       2000-02-23 Nematode        NA      <NA>
## 3             none   1997-05-12       1999-02-12 Nematode        NA      <NA>
## 4             none   1997-05-12       1999-02-12 Nematode        NA      <NA>
## 5             none   1997-05-12       1999-02-12 Nematode   1143021   Anomala
## 6             none   1997-05-12       1999-02-12 Nematode   1143021   Anomala
##     HostFamily  HostOrder HostClass PathTaxID   PathGenus       PathFamily
## 1         <NA>       <NA>      <NA>        NA        <NA>             <NA>
## 2         <NA>       <NA>      <NA>        NA        <NA>             <NA>
## 3         <NA>       <NA>      <NA>        NA        <NA>             <NA>
## 4         <NA>       <NA>      <NA>        NA        <NA>             <NA>
## 5 Scarabaeidae Coleoptera   Insecta    172738 Steinernema Steinernematidae
## 6 Scarabaeidae Coleoptera   Insecta    172738 Steinernema Steinernematidae
##    PathOrder   PathClass PathKingdom
## 1       <NA>        <NA>        <NA>
## 2       <NA>        <NA>        <NA>
## 3       <NA>        <NA>        <NA>
## 4       <NA>        <NA>        <NA>
## 5 Rhabditida Chromadorea     Metazoa
## 6 Rhabditida Chromadorea     Metazoa
```

The focus of the EDWIP data is largely on insect pests. This becomes a bit clearer when we look at the insect host species data (available in the `data(hosts)` produce), and at the parasite species richness for given insect hosts.

```
data(hosts)
table(hosts$InsectStatus)
```

```
##
##         Beneficial Endangered Species               Pest          uncertain
##                 67                   1                901                  1
##    unknown species           Wildlife
##                101                348
```

This `hosts` data also has information on insect generation time, habitat, and diet preferences. Looking at nematode parasite richness, and sorting parasite species richness from largest to smallest value, we see that ...

```
sort(table(nematode$HostSpecies), decreasing=TRUE)[1:5]
```

```
##
##        Popillia japonica    Cyclocephala borealis          Cydia pomonella
##                       22                      12                       9
##   Phyllopertha horticola Amphimallon solstitialis
##                        8                       6
```

... the Japanese beetle (*Popillia japonica*) has the most nematode associations in the data currently. This beetle is a documented generalist pest species in North America and Europe. In fact, all of the top five host species in terms of nematode parasite species richness are defined as 'pests'. Some of the associations are repeated observations, as the number of unique nematode species per host is often less than the above sorted table of interactions would suggest.

The identity of these highly parasitized (or well-sampled) insect host species changes as a function of what data resource we consider, as the insect hosts with a large number of interactions with viruses tend to be markedly different (fewer beetles and more loopworms, earworms, and bees).

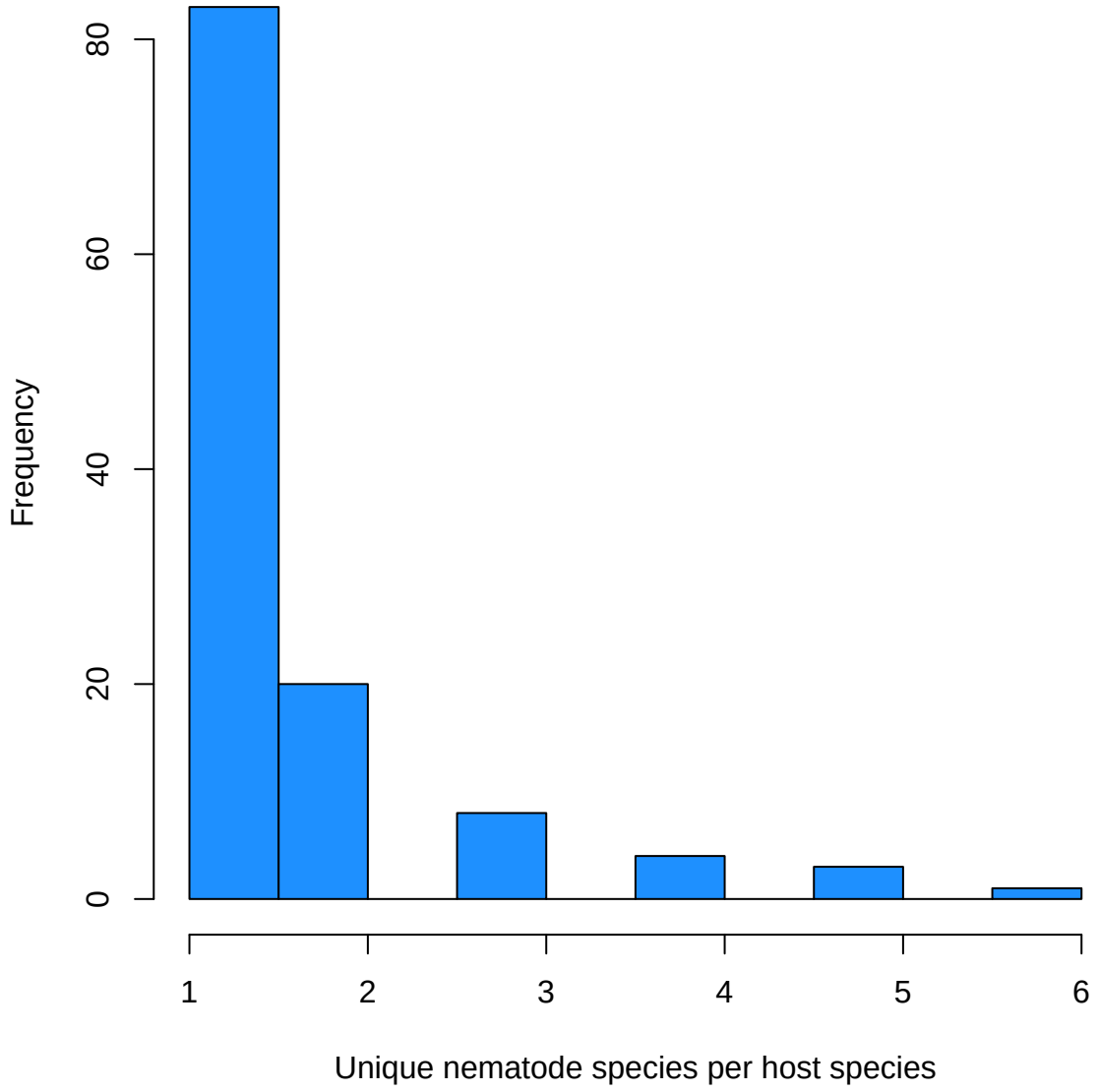
```
sort(table(viruses$HostSpecies), decreasing=TRUE)[1:5]
```

```
##
##      Trichoplusia ni       Apis mellifera      Helicoverpa zea
##                   24                   16                   14
##  Galleria mellonella Helicoverpa armigera
##                   12                   11
```

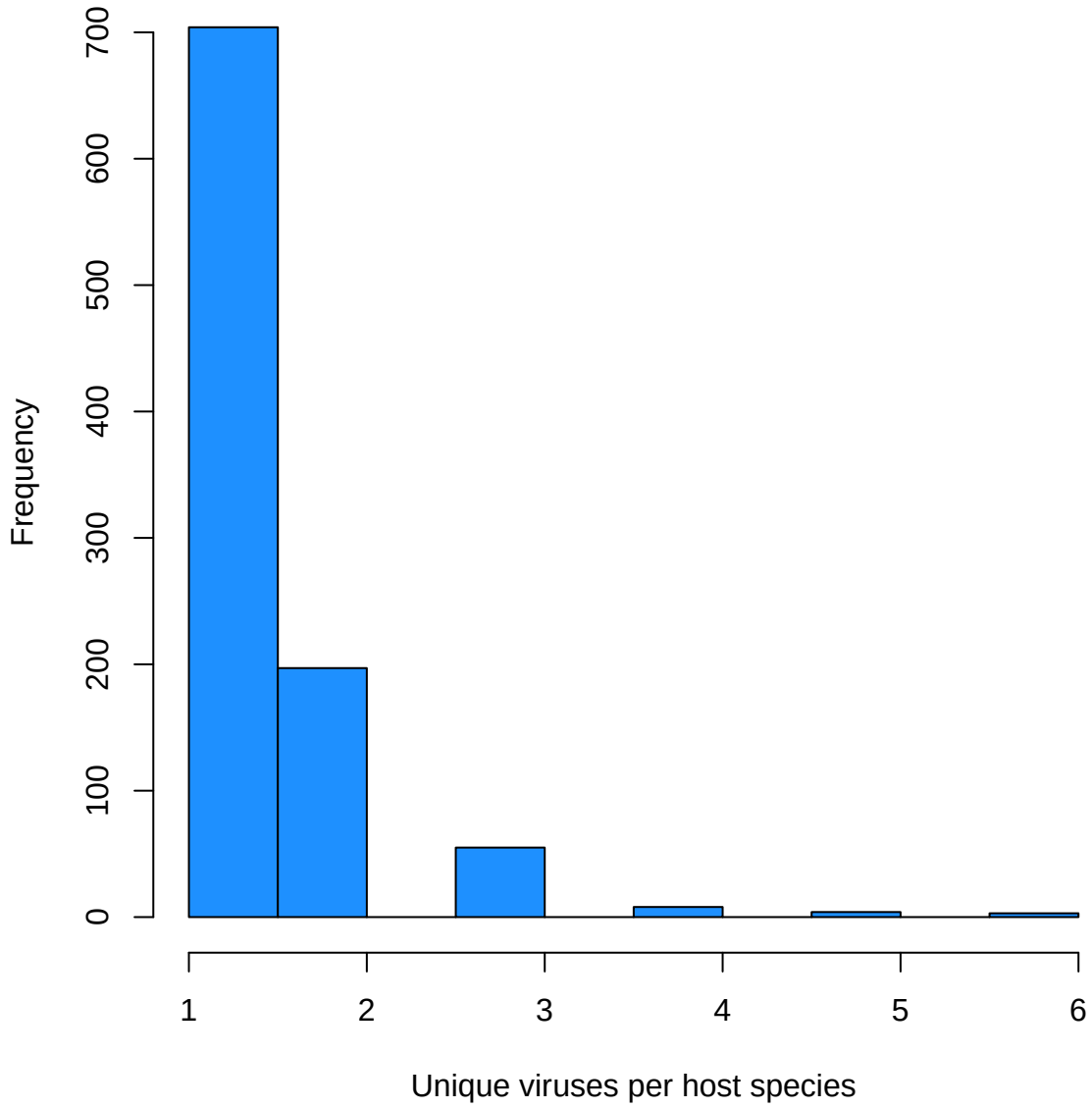
If we wanted to actually get at true parasite species richness, we would calculate the number of unique pathogen species per host species, whereas above we just consider the total number of interactions recorded. That is, a host species could be infected by the same pathogen 20 times in the above calculation, whereas these differences in unique pathogen species would be evident below.

Here, we plot histograms for nematode parasite species richnes and viral richness of insect hosts.

```
nema2 <- nematode %>%
  dplyr::group_by(HostSpecies) %>%
  dplyr::summarise(uniqueNema=length(unique(PathogenSpecies)))

par(mar=c(4,4,0.5,0.5))
hist(nema2$uniqueNema, col='dodgerblue',
  main='', ylab='Frequency',
  xlab='Unique nematode species per host species')
```

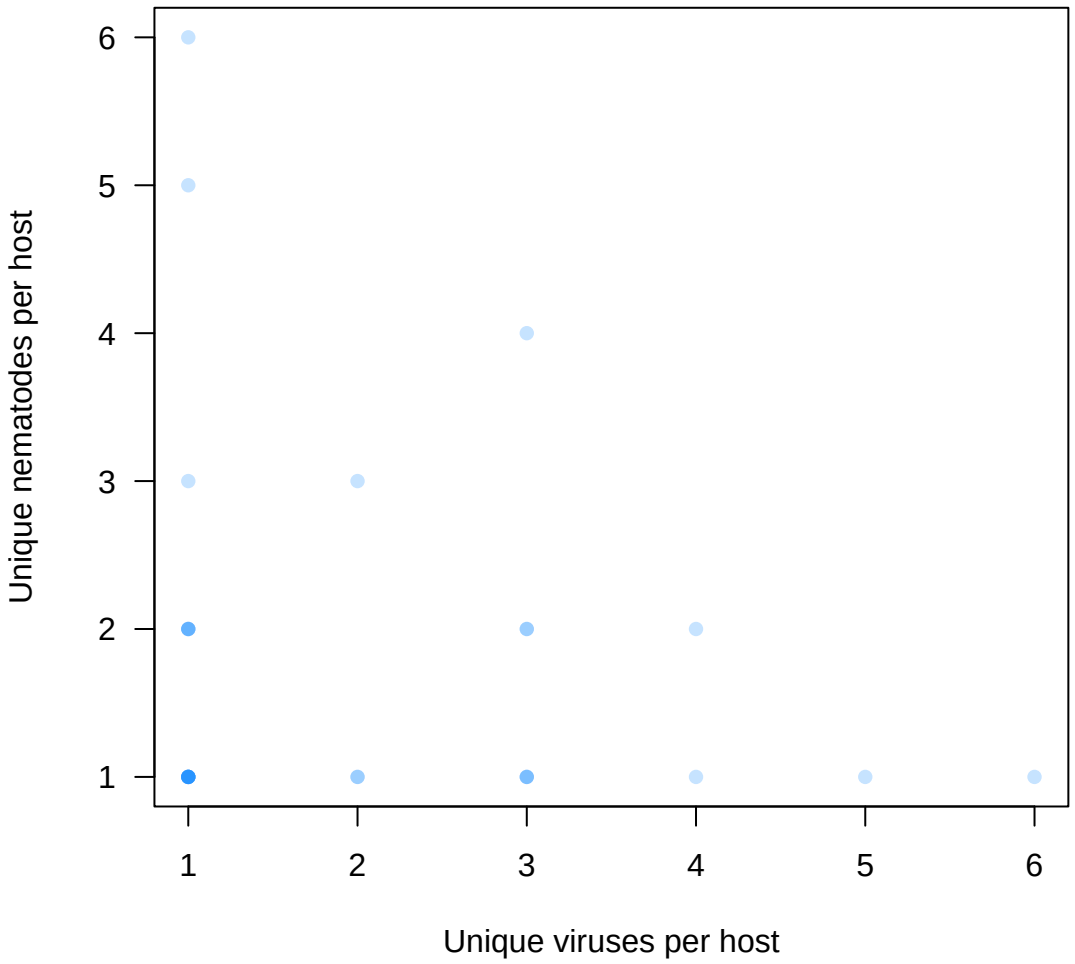



```
viru2 <- viruses %>%
  dplyr::group_by(HostSpecies) %>%
  dplyr::summarise(uniqueViru=length(unique(PathogenSpecies)))

par(mar=c(4,4,0.5,0.5))
hist(viru2$uniqueViru, col='dodgerblue',
  main='', ylab='Frequency',
  xlab='Unique viruses per host species')
```



Then we join the two resulting data.frame objects to allow the comparison of host species by their pathogen species richness. That is, do insect species with more nematode species infecting them also have more recorded viruses which infect them? We do not find very strong support for this idea given the data.

```
nemaViru <- dplyr::full_join(nema2, viru2, by='HostSpecies')

plot(y=nemaViru$uniqueNema,
  x=nemaViru$uniqueViru,
  xlab='Unique viruses per host',
  ylab='Unique nematodes per host',
  pch=16, las=1,
  col=adjustcolor('dodgerblue',0.25)
)
```



NOTE: Some of the columns in the different data sets are different, as the nematode data have information on soil type (`SoilType`), associated bacteria (`AssociatedBacterium`) and intermediate hosts (`IntermediateHost`), while the non-viral pathogen data have information on biogeographic realm (`BiogeographicRegion`), for example.

Visualizing the data

Another way to explore the EDWIP data in the package is through visualization. Here, we reproduce the manuscript figures, highlighting aspects of data size in terms of number of unique species and interactions across the different pathogen types. Below, we write a quick function to subset each data product down to just columns `PathogenSpecies`, `HostSpecies`, and `Group`.

```
#' subset the edwip data objects
#'
#' @param dat the edwip data.frame
#'
#' @return subset data

getSubset <- function(dat){
  nms <- c('PathogenSpecies', 'HostSpecies', 'Group')
  tmp <- dat[,which(colnames(dat) %in% nms)]
  tmp <- tmp[,order(colnames(tmp))]
```

```
  return(tmp)
}
```

We then bind all these rows together into one data.frame called `edwip3`, and create a column called `interaction` which describes the presence or absence of a known interaction between host and pathogen.

```
edwip3 <- rbind(
  getSubset(nematode),
  getSubset(viruses),
  getSubset(nvpassoc)
)
edwip3$interaction <- 1
```

This data.frame is then joined with a subset version of the `negative` data, which describes attempted (and failed) infections of hosts by pathogens. This is incredibly important (and rarely known) information that sets clear barriers on host usage. For these interactions, we set the `interaction` column to 0, as these correspond to known non-interactions.

```
neg <- getSubset(negative)
neg$interaction <- 0
```

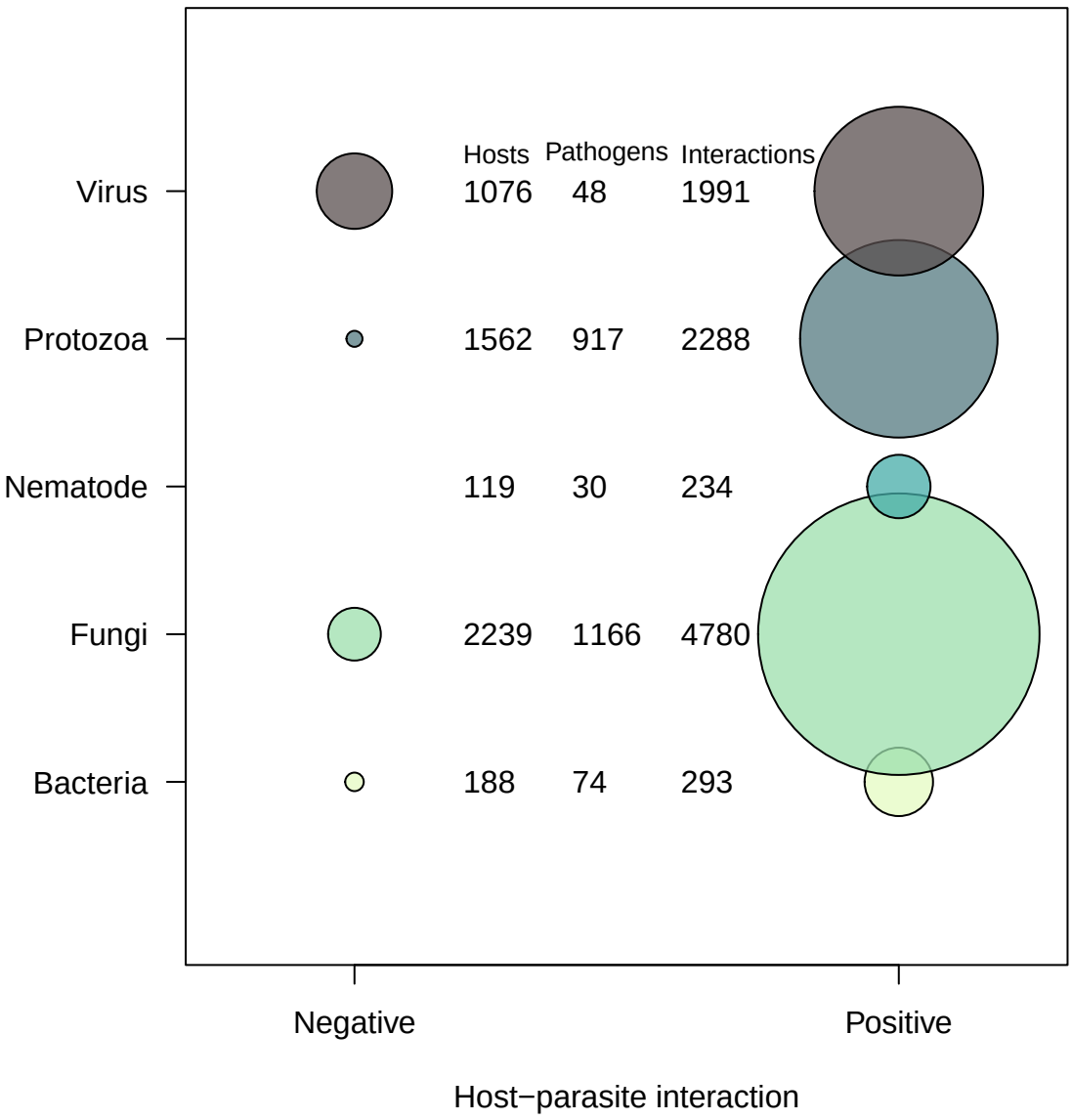
We then bind everything together, and change a couple of instances where `Group` was incorrect or misleading (e.g., `Mollicutes` should be included as `Bacteria`).

```
edwip4 <- rbind(edwip3, neg)
edwip4$Group[which(edwip4$Group == 'Mollicutes')] <- 'Bacteria'
edwip4$Group[which(edwip4$Group == 'Viruses')] <- 'Virus'
```

We can then make the bubble plot, which describes the number of unique hosts, pathogens, and the number of known interactions (both successful infections and failed infections).

```
bubble <- edwip4 %>%
  dplyr::group_by(Group) %>%
  dplyr::summarise(n=length(HostSpecies),
    nHosts=length(unique(HostSpecies)),
    nPaths=length(unique(PathogenSpecies)),
    positives=sum(interaction==1, na.rm=TRUE),
    negatives=sum(interaction==0, na.rm=TRUE)
  )

colorz <- c('#E5FCC2', '#9DE0AD', '#45ADA8', '#547980', '#594F4F')

par(mar=c(4,6,0.5,0.5))
plot(x=1:2,
  y=1:2, type='n', las=1,
  ylim=c(0,6),
  xaxt='n', yaxt='n',
  xlim=c(0.75,2.25),
  xlab='Host-parasite interaction',
  ylab='')

scul <- sqrt(c(unlist(bubble[,6]), unlist(bubble[,5])))
```

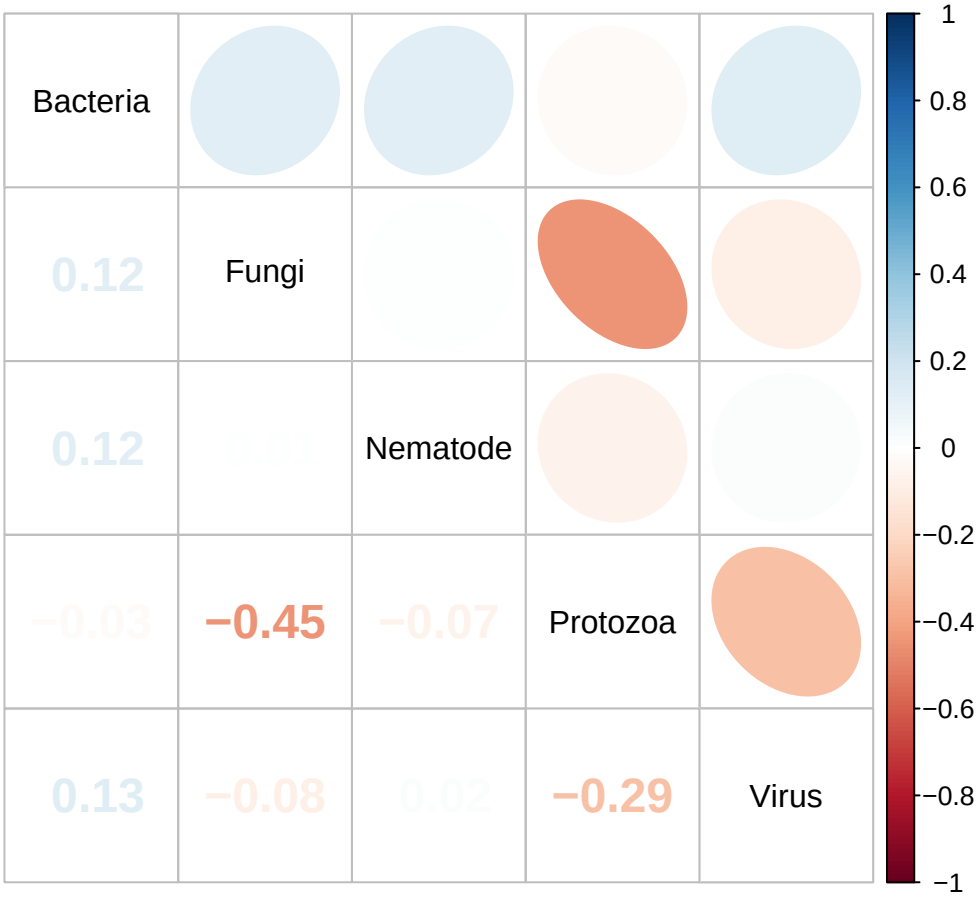
```
scul <- 20*(scul / max(scul))

points(x=sort(rep(c(1,2),5)),
  y=rep(1:5, 2),
  col=1,
  bg=adjustcolor(colorz, 0.75),
  cex=scul,
  pch=21)

text(x=c(1.2,1.35,1.6),
  y=rep(5.25,3), cex=0.85, adj=0,
  c('Hosts', 'Pathogens', 'Interactions'))

text(x=rep(1.2, 5),
  y=1:5, cex=1, adj=0,
  paste(bubble$nHosts))

text(x=rep(1.4, 5),
  y=1:5, cex=1, adj=0,
  paste(bubble$nPaths))

text(x=rep(1.6, 5),
  y=1:5, cex=1, adj=0,
  paste(bubble$n))

axis(1, at = c(1,2),
  labels=c("Negative", "Positive")
)

axis(2, at = 1:5, las=1,
  labels=bubble$Group
)
```



Correlation plot examining covariance among parasite species richness across hosts when parasite species are broken into their respective groups

```
tmp <- as.data.frame.matrix(
  with(edwip4[which(edwip4$interaction==1),],
    table(HostSpecies, Group)
  )
)

par(mar=c(2,2,0.5,0.5))

corrplot::corrplot.mixed(cor(tmp, method='spearman'),
    lower='number', upper='ellipse',
    tl.col=1,
  insig='label_sig',
  number.cex=1.5,
    addgrid.col=grey(0.5,0.5),
    mar=c(0,0,1,0)
```

)


Bacteria
0.12
Fungi
0.12
0.01
Nematode
−0.03
−0.45
−0.07
Protozoa
0.13
−0.08
0.02
−0.29
Virus
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1