Co-phylogenetic analyses

A possible pitch for the paper

Given an observed pattern of host-parasite associations and phylogenies for the host and parasite species, we can test four possible explanations for the observed pattern. The pattern could be random with respect to both phylogenies, which might suggest that host-parasite interactions are driven more by biotic or abiotic factors than by evolution. At the opposite end of the spectrum, it could be that the pattern is driven by coevolution between the hosts and parasites, so that closely related hosts tend to be infected by closely related parasites. (Discussion material: the TREE paper we read from Soren Nylin recently that was arguing against cospeciation being an important force). There are also two intermediate perspectives. It is possible that the pattern is largely driven by host or parasite evolution. In the former case, closely related hosts are infected by similar parasites, irrespective of the relatedness of those parasites. In the latter case, closely related parasites infect similar hosts, irrespective of the relatedness of those hosts. Determining which of these explanations provides the best fit with data can help us to identify zoonotic disease threats. For example, if host-parasite associations are driven by coevolution, then the most likely parasites to spill into humans will be those that both infect our close relatives and are closely related to parasites already known to infect us. However, if host evolution plays the dominant role, then we need to focus our attention on all of the parasites of our close relatives, regardless of whether they are similar to parasites that already infect us. This may be a relatively straightforward task. On the other hand, if parasite evolution plays the dominant role, the task is more difficult because it would require us to identify the close relatives of our parasites, whatever may be their current hosts.

Methods and results

The data for this analysis is comprised of three things: (1) an incidence matrix where 1 represents an occurrence of a particular host-parasite interaction in the GMPD database, and a zero indicates ; (2) a host phylogeny; (3) a parasite phylogeny. With these data, we can carry out several different co-phylogenetic analyses using different methods that have been proposed in the literature.

```
library(MCMCglmm)
```

```
## Loading required package: Matrix
## Loading required package: coda
## Loading required package: ape
```

```
library(gdata)
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
```

```
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
library(igraph)

##
## Attaching package: 'igraph'
## The following objects are masked from 'package:MCMCglmm':
##
##     path, sir
## The following objects are masked from 'package:ape':
##
##     edges, mst, ring
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
## The following object is masked from 'package:base':
##
##     union
library(phytools)

## Loading required package: maps
library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------------------------------------------ tidyve
## x dplyr::as_data_frame() masks tibble::as_data_frame(), igraph::as_data_frame()
## x dplyr::combine()       masks gdata::combine()
## x purrr::compose()       masks igraph::compose()
## x tidyr::crossing()      masks igraph::crossing()
## x tidyr::expand()        masks Matrix::expand()
## x dplyr::filter()        masks stats::filter()
## x dplyr::first()         masks gdata::first()
## x dplyr::groups()        masks igraph::groups()
## x purrr::keep()          masks gdata::keep()
## x dplyr::lag()           masks stats::lag()
## x dplyr::last()          masks gdata::last()
## x purrr::map()           masks maps::map()
## x tidyr::pack()          masks Matrix::pack()
## x purrr::simplify()      masks igraph::simplify()
## x tidyr::unpack()        masks Matrix::unpack()
```

```
#library(brms)

################################
##      Load the trees        ##
################################
mam_tree <- read.tree("mammal_tree_clean.tre")
para_tree <- read.tree("helminth_tree_clean.tre")

################################
##  Format the GMPD Data      ##
################################
data <- read.csv("GMPD_clean.csv")

# ## verify that we have concordance between the host phylogeny and the host data in the GMPD
# unique(gsub(" ","_",data$HostCorrectedName))%in%mam_tree$tip.label %>% all
# mam_tree$tip.label%in%unique(gsub(" ","_",data$HostCorrectedName)) %>% all
#
# ## verify that we have concordance between the parasite phylogeny and the parasite data in the GMPD
# unique(gsub(" ","_",data$ParasiteCorrectedName))%in%para_tree$tip.label %>% all
# para_tree$tip.label%in%unique(gsub(" ","_",data$ParasiteCorrectedName)) %>% all

## create a new data.frame with all possible host-parasite combinations
expand.grid(Host.species=(data$HostCorrectedName %>% unique),
            Parasite.species=(data$ParasiteCorrectedName %>% unique)) -> ndata

## ncount = the number of times that each association occurs in data
## presence = incidence data (0/1 if the association occurs or not)
## nhosts.sampled = the number of times each host occurs in the GMPD
## nparas.sampled = the number of times each parasite occurs in the GMPD
mutate(ndata,
       ncount=left_join(ndata, data %>% count(HostCorrectedName,ParasiteCorrectedName),
                        by=c("Host.species"="HostCorrectedName","Parasite.species"="ParasiteCorrectedNa
       presence=ifelse(is.na(ncount),0,1),
       nhosts.sampled=left_join(ndata, data %>% count(HostCorrectedName),
                                by=c("Host.species"="HostCorrectedName"))$n,
       nparas.sampled=left_join(ndata, data %>% count(ParasiteCorrectedName),
                                by=c("Parasite.species"="ParasiteCorrectedName"))$n) -> ndata
```

The first method is ParaFitGlobal (Legendre et al. 2002), which evaluates the evidence for coevolution between parasites and hosts. This method works by testing for congruence between host and parasite phylogenetic trees, that is, it tests whether hosts and their parasites have equivalent positions in their respective trees. Perfect congruence would signal tight codiversification of specialist parasites with their hosts, whereas no congruence would signal that host-parasite associations are formed randomly with respect to the evolutionary history of each species. As such, the null hypothesis that the method is testing is that the evolution of hosts and parasites is independent. This method was one of the first developed that could account for the fact that many parasites can infect more than one host, and that hosts are often infected by many parasites.

The method determines the evidence for coevolution via a permutation test. That is, it computes a statistic based on matrices describing (A) the presence/absence of each host-parasite association; (B) the parasite phylogenetic tree; (C) the host phylogenetic tree. To determine whether this statistic has a value that is different than what you would expect via chance, the presence/absence data is randomly permuted in three ways. Legendre et al. proposed a permutation such that each parasite infects the same number of hosts, but the identity of those hosts is randomly determined. However, an alternative would be a permutation such that each host is infected by the same number of parasites, but the identify of those parasites is randomly

determined. Hommola et al. proposed a third possibility, where only the total number of host-parasite associations is preserved, and those associations are randomly determined. Under all perturbations, the test statistic is computed to produce null distributions of the test statistic that the true value can be compared against. As pointed out by Hadfield et al. (2014), comparing the value of the test statistic against the null distributions generated by different types of permutation provides slightly different information. In particular, the first permutation tests for host-parasite coevolution, for host evolutionary interactions (which occur if related hosts are infected by similar parasites, irrespective of the parasite phylogeny), for parasite evolutionary interactions (which occur if related parasites infect similar hosts, irrespective of the host phylogeny), and for phylogenetic signal in the parasite species richness infecting hosts (because the permutation alters the number of parasites infecting each host). The second permutation tests for coevolution, host and parasite evolutionary interactions, and for phylogenetic signal in the host range of parasites (because the permutation alters the number of hosts each parasite infects). The third permutation tests for coevolution, host and parasite evolutionary interactions, and for phylogenetic signal in both parasite species richness and host range.

(Sidenote: Legendre et al. 2002 also provide a method for testing whether there are host-parasite associations that are particularly important to the overall coevolutionary pattern. I haven't used this method because I wasn't sure why I would, but it could be done pretty easily.)

Based on the results of this analysis, we conclude that there is good evidence for coevolution, host and parasite evolutionary interactions, and for phylogenetic signal in parasite species richness and host range, as the observed value of the test statistic (which is meaningless in and of itself) is more extreme than the values observed for *any* of the bootstrap permutations, using any method of permutation.

```
if(!file.exists("legendre_results.RDS")) {

# matrices for host and parasite based on the branch lengths in the phylogenetic tree
# (Note that differences in tree depth do not matter because of the use of corr=T)
ht<-vcv(mam_tree, corr=T)
pt<-vcv(para_tree, corr=T)

# incidence data
Y<-table(ndata$Host.species, ndata$Parasite.species, ndata$presence)[,,2]>0
## change names to match the phylogeny
rownames(Y) <- gsub(" ", "_", rownames(Y))
colnames(Y) <- gsub(" ", "_", colnames(Y))

## reorder A matrices so match row/columns of Y
hi<-match(rownames(Y), rownames(ht))
pi<-match(colnames(Y), rownames(pt))
ht<-ht[hi,hi]
pt<-pt[pi,pi]

## get principal coordinates of the phylogenetic distance matrix (for the Legendre approach)
htp<-pcoa(1-ht)$vectors
ptp<-pcoa(1-pt)$vectors

## D matrices (see Appendix) for Legendre and Ives methods
lD<-t(htp)%*%Y%*%ptp

## Legendre metric
cl<-sum(diag(t(lD)%*%lD))

## Bootstrap result storage
boot.1 <- 1:1000
```

```
boot.2 <- 1:1000
boot.3 <- 1:1000

###################
## Permutations ##
###################
## Compute the Legendre metric for randomly permuted
## incidence data.
for(j in 1:1000){
  print(j)

  Y2<-apply(Y, 2, sample)                      # Legendre permtations
  lD<-t(htp)%*%Y2%*%ptp                        # Legendre D matrix
  boot.1[j]<-sum(diag(t(lD)%*%lD))               # Legendre metric (ParafitGlobal)

  Y2 <- apply(Y, 1, sample)
  lD<-t(ptp)%*%Y2%*%htp                        # Legendre D matrix
  boot.2[j]<-sum(diag(t(lD)%*%lD))               # Legendre metric (ParafitGlobal)

  Y2<-Y[sample(1:nrow(Y)),sample(1:ncol(Y))]   # Hommola sampling
  lD<-t(htp)%*%Y2%*%ptp
  boot.3[j]<-sum(diag(t(lD)%*%lD))
}

legendre_results <- c(cl, sum(cl < boot.1)/1000, sum(cl < boot.2)/1000, sum(cl < boot.3)/1000)
names(legendre_results) <- c("metric", "L1-pval", "L2-pval", "H-pval")
saveRDS(legendre_results, file="legendre_results.RDS")} else{
  legendre_results <- readRDS("legendre_results.RDS")
  legendre_results
}
```

```
##  metric L1-pval L2-pval  H-pval
## 637.819   0.000   0.000   0.000
```

The second method is that of Ives and Godfray (2006), which provides an alternative, but related, way of testing for phylogenetic signal in the pattern of host-parasite association. As with the Legendre et al. method, Ives and Godfray combine the information contained within the host-parasite association data and the host and parasite phylogenies to estimate so-called "fourth corner" statistics that describe the phylogenetic signal in the pattern of host-parasite associations. However, the method differs from that of Legendre et al. in that it essentially transforms the branch lengths of the host and parasite phylogenies to maximize the fit of the evolutionary model to the observed host-parasite association data, whereas the branch lengths remain fixed by the Legendre et al. method. In other words, during the fitting process, the covariance between any two tips n either tree (which is proportional to the branch lengths separating them) is adjusted to maximize the fit of the evolutionary model. More exactly, the covariance between any two tips is specified by an Ornstein-Uhlenbeck process with a parameter d that determines the strength of the phylogenetic signal, essentially transforming the branch lengths (Blomberg et al. 2003). The Ornstein-Uhlenbeck model is often described as a model for stabilizing selection: a deterministic tendency towards an "optimal" value for a trait evolving along a phylogeny (Hansen 1997). If $d = 0$, there is no phylogenetic correlation (a star phylogeny), whereas $d = 1$ implies no selection (Brownian motion), and a value of $0 < d < 1$ implies some amount of stabilizing selection. The method assumes that there is some value, d_h that best describes the covariance between host species, and a separate parameter, d_p, that describes the covariance between parasite species; specifically, the method estimates the values of d_h and d_p that minimize the mean square error between the model-predicted host-parasite associations and the observed host-parasite associations. Comparing the values of d_h and d_p give a sense of how much phylogenetic signal is due to the host versus the parasite.

Unfortunately, the method in full generality is not applicable to phylogenies this large, because the method generates matrices that are so large that they consume all available computer memory. Instead, we use a simplified version of the method investigated in Hadfield et al. (2014). The simplified method fixes $d_h = d_p = 1$, which fixes the covariance between tips to be strictly proportional to branch lengths, effectively assuming that host-parasite associations evolve according to Brownian motion. Under this assumption, the mean square error can be used as a test statistic. Hadfield et al. (2014) show that there is a close relationship between this test statistic and that of Legendre et al. (2002).

We then use the same set of permutations of the data described above to evaluate the evidence for phylogenetic signal in the pattern of host-parasite association. Again, we see strong evidence that the observed patterns of host-parasite association are non-random. In particular, the p-values of the test comparing against the first and second perturbation (where the number, but not identity, of hosts parasitized by each parasite is preserved and where the number, by not identity, of parasites infecting each host is preserved) are both very small. The smaller value of the p-value for the second perturbation could indicate that there is stronger evidence for phylogenetic signal in host range of parasites than in the parasite species richness of hosts. However, the p-value for the perturbation that randomly reassigns all host-parasite associations is much higher, tempering the conclusions drawn above.

[Sidebar: it might be worth applying the full Ives and Godfray analysis on a subset of the data. For example, we could do the analysis on the big dataset to determine whether there is any signal globally, and then look at only a subset of the data (e.g., just the data from Primates) to determine the relatively influence of host versus parasite evolution. This would have a nice connection to the question of zoonotic risk, as it would let us determine whether humans are most at risk from parasites that are phylogenetically close to ours, or from the parasites that infect our close relatives.]

```
if(!file.exists("IvesGodfray_results.RDS")) {
# matrices for host and parasite based on the branch lengths in the phylogenetic tree
# (Note that differences in tree depth do not matter because of the use of corr=T)
ht<-vcv(mam_tree, corr=T)
pt<-vcv(para_tree, corr=T)

# incidence data
Y<-table(ndata$Host.species, ndata$Parasite.species, ndata$presence)[,,2]>0
## change names to match the phylogeny
rownames(Y) <- gsub(" ", "_", rownames(Y))
colnames(Y) <- gsub(" ", "_", colnames(Y))

## reorder A matrices so match row/columns of Y
hi<-match(rownames(Y), rownames(ht))
pi<-match(colnames(Y), rownames(pt))
ht<-ht[hi,hi]
pt<-pt[pi,pi]

## get unnormalised eigenvectors of phylogenetic distance matrix (for the Ives and Godfray method)
hte<-t(t(eigen(solve(ht))$vectors)*sqrt(eigen(solve(ht))$values))
pte<-t(t(eigen(solve(pt))$vectors)*sqrt(eigen(solve(pt))$values))

## D matrix (see Appendix of Hadfield et al. 2014) for Ives & Godfray methods
iD<-t(hte)%*%(Y-mean(Y))%*%pte

## Metrics of Ives & Godfray (see Appendix of Hadfield et al. 2014)
ci<-sum(diag(t(iD)%*%iD))

boot.1<-1:1000
boot.2<-1:1000
```

```
boot.3<-1:1000
## Consider an alternative Legendre permutation that permutes the rows rather than columns!!

##################
## Permutations ##
##################
for(j in 1:1000){

  Y2<-apply(Y, 2, sample)                       # Legendre permtations
  iD<-t(hte)%*%(Y2-mean(Y2))%*%pte              # Ives D matrix
  boot.1[j]<-sum(diag(t(iD)%*%iD))                # Ives metric (MSEb)

  Y2<-apply(Y, 1, sample)                       # Legendre permtations
  iD<-t(pte)%*%(Y2-mean(Y2))%*%hte              # Ives D matrix
  boot.2[j]<-sum(diag(t(iD)%*%iD))                # Ives metric (MSEb)

  Y2<-Y[sample(1:nrow(Y)),sample(1:ncol(Y))]  # Hommola sampling
  iD<-t(hte)%*%(Y2-mean(Y2))%*%pte
  boot.3[j]<-sum(diag(t(iD)%*%iD))
}
res <- c(ci,sum(ci > boot.1)/1000, sum(ci > boot.2)/1000, sum(ci > boot.3)/1000)
names(res) <- c("metric","L1-pval","L2-pval","H-pval")
saveRDS(res, file="IvesGodfray_results.RDS")} else {
  IvesGodfray_results <- readRDS("IvesGodfray_results.RDS")
  IvesGodfray_results
}
```

```
##     metric    L1-pval    L2-pval     H-pval
## 339667.560      0.016      0.002      0.362
```

The third method is that of Hommola et al. (2009), which tests for a correlation between shared branch lengths. More specifically, the method looks at pairs of host-parasite associations and calculates the phylogenetic distance between the hosts in the pair and between the parasites in the pair: the host distance will be zero for pairs representing two parasites infecting the same host; similarly, the parasite distance will be zero for pairs representing a parasite infecting two hosts. After computing these branch lengths for all pairs, the method then estimates the correlation between the host distances and the parasite distances over all host-parasite association pairs. A high correlation means that, when two hosts are far apart on the tree, the parasites that infect them also tend to be far apart; similarly, when hosts are closely related, their parasites tend to be closely related as well. A low correlation would mean that there is no relationship between the distance between hosts and between parasites.

The results indicate that there is fairly low correlation between host and parasite shared branch lengths observed in the GMPD data ($r = 0.07$). Moreover, this correlation is fairly likely under random perturbations of the host-parasite association data, with bootstrap p-values ranging from 0.33 to 0.38.

```
if(!file.exists("Hommola_results.RDS")) {
  ## make sure you have a compiled Hommola.so available in this directory before running hommola()
source("Hommola.R")

# matrices for host and parasite based on the branch lengths in the phylogenetic tree
# (Note that differences in tree depth do not matter because of the use of corr=T)
ht<-vcv(mam_tree, corr=T)
pt<-vcv(para_tree, corr=T)

# incidence data
```

```
Y<-table(ndata$Host.species, ndata$Parasite.species, ndata$presence)[,,2]>0
## change names to match the phylogeny
rownames(Y) <- gsub(" ", "_", rownames(Y))
colnames(Y) <- gsub(" ", "_", colnames(Y))

## Metrics of Legendre, Ives & Godfray, and Hommola (see Appendix of Hadfield et al. 2014)
ch<-hommola(Y, ht,pt)

boot.1 <-1:1000
boot.2 <-1:1000
boot.3 <-1:1000

##################
## Permutations ##
##################
for(j in 1:1000){

  Y2<-apply(Y, 2, sample)                      # Legendre permtations
  boot.1[j]<-hommola(Y2, ht,pt)                  # Hommola metric

  Y2<-apply(Y, 1, sample)
  boot.2[j]<-hommola(Y2, pt, ht)

  Y2<-Y[sample(1:nrow(Y)),sample(1:ncol(Y))]  # Hommola sampling
  boot.3[j]<-hommola(Y2, ht,pt)
}
Hommola_results <- c(ch, sum(ch < boot.1)/1000, sum(ch < boot.2)/1000, sum(ch < boot.3)/1000)
names(Hommola_results) <- c("metric","L1-pval","L2-pval","H-pval")
saveRDS(Hommola_results, file="Hommola_results.RDS")} else {
  Hommola_results <- readRDS("Hommola_results.RDS")
  Hommola_results
}
```

```
##       metric      L1-pval      L2-pval       H-pval
## 0.0009764802 0.3620000000 0.3730000000 0.3150000000
```

The fourth method comes from Krasnov et al. (2012). It is essentially a two-part algorithm. First, it estimates the modularity of the network formed by host-parasite associations: in a host-parasite network with high modularity, one would find clusters of hosts and parasites that interact mainly with one another, and not with other clusters of hosts and parasites. Modules are computed using the cluster_walktrap function in the R package igraph (Pons and Latapy 2005). Application of this function to the GMPD data produced 38 distinct host-parasite clusters; the overall modularity score was 0.63. You can see the modules in the graph. What is obvious is that there are several highly connected modules comprising dense networks of closely interacting hosts and parasites, surrounded by many small, often disconnected modules. There are five modules with 20 or more interacting hosts and parasites, the largest of which contains 154 hosts and parasites. This module contains 62% of the Carnivore hosts, as well as large fractions of parasites from many of the parasite classes (e.g., 62% of the Platyhelminthes and 59% of the Acanthocephala)). The second largest module (77 species) contains 65% of the Artiodactyla (even-toed ungulates) and their parasites. The third largest module (46 species) contains 57% of the Primates and their parasites. The fourth largest module (37 species) contains all of the Perissodactyla (odd-toed ungulates) and their parasites. In other words, the clustering appears to be highly structured by host phylogeny.

```
#################################################
## Run module on data (from Krasnov et al. 2012) ##
#################################################
```

```r
# A matrices for host and parasite
ht<-vcv(mam_tree, corr=T)
pt<-vcv(para_tree, corr=T)

# incidence data
Y<-table(ndata$Host.species, ndata$Parasite.species, ndata$presence)[,,2]>0
## change names to match the phylogeny
rownames(Y) <- gsub(" ", "_", rownames(Y))
colnames(Y) <- gsub(" ", "_", colnames(Y))

hi<-match(rownames(Y), rownames(ht))
pi<-match(colnames(Y), rownames(pt))

# reorder A matrices so match row/columns of Y
ht<-ht[hi,hi]
pt<-pt[pi,pi]

# assign species to modules
G<-graph_from_incidence_matrix(Y) ## create a bipartite igraph graph from the incidence matrix
MU<-cluster_walktrap(G) ## identify clusters

plot(MU,G,vertex.label="",vertex.size=2)
```

```r
MU<-membership(MU) ## determine which clusters each host and parasite belong to

## sizes of each cluster
sapply(1:38, function(c) sum(MU==c))
```

```
##  [1]  46  77  20   6   7  12 154   5   3   4   3   6  37   6   2   2   2   2   2
## [20]   2   5   4   4   3   3   3   3   2   2   2   2   2   2   2   2   2   2   2
```

```r
## How many of the members of each mammal Order belong in each cluster?
## 65% of Artiodactyla were assigned to cluster 2
## 62% of Carnivores were assigned to cluster 7
## ALL of the Perissodactyla were assigned to cluster 13
## 57% of Primates were assigned to cluster 1
sapply(levels(data$HostOrder),
       function(ord) sapply(1:38, function(c) sum((subset(data,HostOrder==ord)$HostCorrectedName %>% uni
) %>% t
```

```
##                    [,1]      [,2]  [,3]       [,4]   [,5]       [,6]
## Artiodactyla  0.0000000 0.6545455 0.000 0.03636364 0.0000 0.01818182
## Carnivora     0.0000000 0.0000000 0.125 0.01250000 0.0375 0.02500000
## Perissodactyla 0.0000000 0.0000000 0.000 0.00000000 0.0000 0.00000000
## Primates      0.5689655 0.0000000 0.000 0.00000000 0.0000 0.00000000
##                     [,7]   [,8]       [,9]       [,10]      [,11]     [,12]
## Artiodactyla  0.05454545 0.0000 0.00000000 0.01818182 0.00000000 0.0000000
## Carnivora     0.62500000 0.0375 0.00000000 0.00000000 0.00000000 0.0000000
## Perissodactyla 0.00000000 0.0000 0.00000000 0.00000000 0.00000000 0.0000000
## Primates      0.06896552 0.0000 0.03448276 0.00000000 0.03448276 0.0862069
##                [,13]      [,14]  [,15]      [,16]  [,17]      [,18]      [,19]
## Artiodactyla       0 0.09090909 0.0000 0.00000000 0.0000 0.00000000 0.00000000
## Carnivora          0 0.00000000 0.0125 0.00000000 0.0125 0.00000000 0.00000000
## Perissodactyla     1 0.00000000 0.0000 0.00000000 0.0000 0.00000000 0.00000000
```

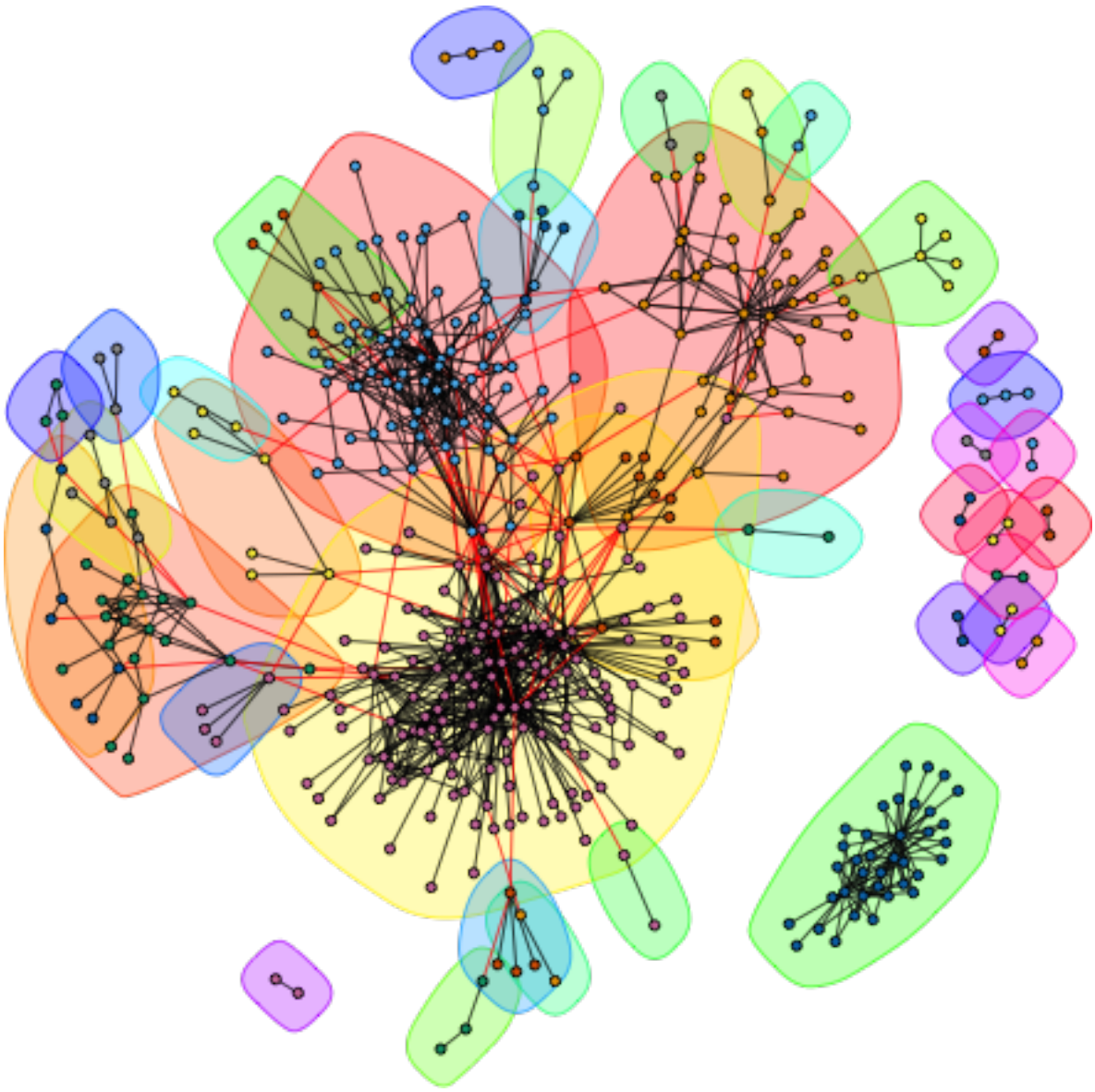

Figure 1: Host-parasite interaction network, broken into 38 distinct modules.

```
## Primates           0 0.00000000 0.0000 0.01724138 0.0000 0.01724138 0.01724138
##                      [,20]      [,21]      [,22]  [,23]  [,24]      [,25]
## Artiodactyla   0.01818182 0.01818182 0.00000000 0.0000 0.0000 0.01818182
## Carnivora      0.00000000 0.00000000 0.00000000 0.0375 0.0125 0.00000000
## Perissodactyla 0.00000000 0.00000000 0.00000000 0.0000 0.0000 0.00000000
## Primates       0.00000000 0.00000000 0.05172414 0.0000 0.0000 0.00000000
##                      [,26] [,27]      [,28]      [,29]      [,30]      [,31]
## Artiodactyla   0.00000000 0.000 0.01818182 0.00000000 0.00000000 0.01818182
## Carnivora      0.00000000 0.025 0.00000000 0.00000000 0.00000000 0.00000000
## Perissodactyla 0.00000000 0.000 0.00000000 0.00000000 0.00000000 0.00000000
## Primates       0.03448276 0.000 0.00000000 0.01724138 0.01724138 0.00000000
##                      [,32]      [,33]  [,34]      [,35]      [,36]  [,37]  [,38]
## Artiodactyla   0.01818182 0.00000000 0.0000 0.01818182 0.00000000 0.0000 0.0000
## Carnivora      0.00000000 0.00000000 0.0125 0.00000000 0.00000000 0.0125 0.0125
## Perissodactyla 0.00000000 0.00000000 0.0000 0.00000000 0.00000000 0.0000 0.0000
## Primates       0.00000000 0.01724138 0.0000 0.00000000 0.01724138 0.0000 0.0000
```

```
## what about if we look structured by parasites?
## Look first by phylum
sapply(levels(data$ParPhylum), function(phy) sapply(1:38, function(c) sum((subset(data,ParPhylum==phy)$
```

```
##                       [,1]      [,2]       [,3]       [,4]       [,5]
## Acanthocephala  0.00000000 0.0000000 0.17647059 0.00000000 0.05882353
## Nematoda        0.07096774 0.2129032 0.02580645 0.01290323 0.01935484
## Platyhelminthes 0.02631579 0.1052632 0.03947368 0.01315789 0.00000000
##                       [,6]      [,7]       [,8]        [,9]      [,10]
## Acanthocephala  0.00000000 0.5882353 0.00000000 0.000000000 0.00000000
## Nematoda        0.05161290 0.2580645 0.01290323 0.006451613 0.01935484
## Platyhelminthes 0.01315789 0.6184211 0.00000000 0.000000000 0.00000000
##                      [,11]       [,12]      [,13]       [,14]      [,15]
## Acanthocephala  0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
## Nematoda        0.006451613 0.006451613 0.20000000 0.006451613 0.00000000
## Platyhelminthes 0.000000000 0.000000000 0.02631579 0.000000000 0.01315789
##                      [,16]      [,17]       [,18]       [,19]       [,20]
## Acanthocephala  0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## Nematoda        0.00000000 0.00000000 0.006451613 0.006451613 0.006451613
## Platyhelminthes 0.01315789 0.01315789 0.000000000 0.000000000 0.000000000
##                      [,21]      [,22]       [,23]     [,24]      [,25]
## Acanthocephala  0.00000000 0.00000000 0.000000000 0.1176471 0.00000000
## Nematoda        0.02580645 0.00000000 0.006451613 0.0000000 0.00000000
## Platyhelminthes 0.00000000 0.01315789 0.000000000 0.0000000 0.02631579
##                      [,26]      [,27]      [,28]       [,29]       [,30]
## Acanthocephala  0.000000000 0.05882353 0.00000000 0.000000000 0.000000000
## Nematoda        0.006451613 0.00000000 0.00000000 0.006451613 0.006451613
## Platyhelminthes 0.000000000 0.00000000 0.01315789 0.000000000 0.000000000
##                      [,31]      [,32]       [,33]       [,34]      [,35]
## Acanthocephala  0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
## Nematoda        0.00000000 0.00000000 0.006451613 0.006451613 0.00000000
## Platyhelminthes 0.01315789 0.01315789 0.000000000 0.000000000 0.01315789
##                      [,36]      [,37]      [,38]
## Acanthocephala  0.000000000 0.00000000 0.00000000
## Nematoda        0.006451613 0.00000000 0.00000000
## Platyhelminthes 0.000000000 0.01315789 0.01315789
```

```
## then by Class
sapply(levels(data$ParClass), function(class) sapply(1:38, function(c) sum((subset(data,ParClass==class)
```

```
##                           [,1]      [,2]       [,3]        [,4]       [,5]
## Archiacanthocephala 0.00000000 0.0000000 0.00000000 0.000000000 0.00000000
## Cestoda             0.00000000 0.1052632 0.07894737 0.026315789 0.00000000
## Enoplea             0.10526316 0.1578947 0.00000000 0.052631579 0.00000000
## Monogenea           0.00000000 0.0000000 0.00000000 0.000000000 0.00000000
## Palaeacanthocephala 0.00000000 0.0000000 0.20000000 0.000000000 0.06666667
## Secernentea         0.06617647 0.2205882 0.02941176 0.007352941 0.02205882
## Trematoda           0.05405405 0.1081081 0.00000000 0.000000000 0.00000000
##                           [,6]      [,7]       [,8]        [,9]      [,10]
## Archiacanthocephala 0.00000000 1.0000000 0.00000000 0.000000000 0.00000000
## Cestoda             0.02631579 0.6052632 0.00000000 0.000000000 0.00000000
## Enoplea             0.15789474 0.4736842 0.00000000 0.000000000 0.00000000
## Monogenea           0.00000000 0.0000000 0.00000000 0.000000000 0.00000000
## Palaeacanthocephala 0.00000000 0.5333333 0.00000000 0.000000000 0.00000000
## Secernentea         0.03676471 0.2279412 0.01470588 0.007352941 0.02205882
## Trematoda           0.00000000 0.6486486 0.00000000 0.000000000 0.00000000
##                          [,11]       [,12]      [,13]       [,14]      [,15]
## Archiacanthocephala 0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
## Cestoda             0.000000000 0.000000000 0.05263158 0.000000000 0.02631579
## Enoplea             0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
## Monogenea           0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
## Palaeacanthocephala 0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
## Secernentea         0.007352941 0.007352941 0.22794118 0.007352941 0.00000000
## Trematoda           0.000000000 0.000000000 0.00000000 0.000000000 0.00000000
##                          [,16]      [,17]       [,18]       [,19]       [,20]
## Archiacanthocephala 0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## Cestoda             0.02631579 0.00000000 0.000000000 0.000000000 0.000000000
## Enoplea             0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## Monogenea           0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## Palaeacanthocephala 0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## Secernentea         0.00000000 0.00000000 0.007352941 0.007352941 0.007352941
## Trematoda           0.00000000 0.02702703 0.000000000 0.000000000 0.000000000
##                          [,21]      [,22]      [,23]     [,24]      [,25]
## Archiacanthocephala 0.00000000 0.00000000 0.00000000 0.0000000 0.00000000
## Cestoda             0.00000000 0.02631579 0.00000000 0.0000000 0.00000000
## Enoplea             0.00000000 0.00000000 0.05263158 0.0000000 0.00000000
## Monogenea           0.00000000 0.00000000 0.00000000 0.0000000 1.00000000
## Palaeacanthocephala 0.00000000 0.00000000 0.00000000 0.1333333 0.00000000
## Secernentea         0.02941176 0.00000000 0.00000000 0.0000000 0.00000000
## Trematoda           0.00000000 0.00000000 0.00000000 0.0000000 0.02702703
##                          [,26]      [,27]      [,28]       [,29]       [,30]
## Archiacanthocephala 0.000000000 0.00000000 0.00000000 0.000000000 0.000000000
## Cestoda             0.000000000 0.00000000 0.00000000 0.000000000 0.000000000
## Enoplea             0.000000000 0.00000000 0.00000000 0.000000000 0.000000000
## Monogenea           0.000000000 0.00000000 0.00000000 0.000000000 0.000000000
## Palaeacanthocephala 0.000000000 0.06666667 0.00000000 0.000000000 0.000000000
## Secernentea         0.007352941 0.00000000 0.00000000 0.007352941 0.007352941
## Trematoda           0.000000000 0.00000000 0.02702703 0.000000000 0.000000000
##                          [,31]      [,32]       [,33]       [,34]      [,35]
## Archiacanthocephala 0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
## Cestoda             0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
```

```
## Enoplea             0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
## Monogenea           0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
## Palaeacanthocephala 0.00000000 0.00000000 0.000000000 0.000000000 0.00000000
## Secernentea         0.00000000 0.00000000 0.007352941 0.007352941 0.00000000
## Trematoda           0.02702703 0.02702703 0.000000000 0.000000000 0.02702703
##                           [,36]      [,37]      [,38]
## Archiacanthocephala 0.000000000 0.00000000 0.00000000
## Cestoda             0.000000000 0.02631579 0.00000000
## Enoplea             0.000000000 0.00000000 0.00000000
## Monogenea           0.000000000 0.00000000 0.00000000
## Palaeacanthocephala 0.000000000 0.00000000 0.00000000
## Secernentea         0.007352941 0.00000000 0.00000000
## Trematoda           0.000000000 0.00000000 0.02702703
```

Second, it estimates whether there is strong phylogenetic signal between modularity and the host or parasite phylogeny: that is, it determines whether the hosts and parasites that belong to modules tend to be closely related. The correlation betwen comembership in a module and phylogenetic distance of hosts was -0.49 in the GMPD data; the correlation between comembership in a module and phylogenetic distance of parasites was 0.03. The significance of these correlations was again assessed using the three permutations. We find that the observed correlation is highly significant for hosts, confirming the results observed in the module graph above, but that the correlation is not significant for parasites.

```
if (!file.exists("Krasnov_results.RDS")) {
lth<-which(lower.tri(ht))

# correlation between comembership and phylogenetic distance of hosts
chk<-cor(outer(MU[1:length(hi)], MU[1:length(hi)], "==")[lth], c(1-ht)[lth])

ltp<-lower.tri(pt)

cpk<-cor(outer(MU[length(hi)+1:length(pi)], MU[length(hi)+1:length(pi)], "==")[ltp], c(1-pt)[ltp])
# correlation between comembership and phylogenetic distance of parasites

## Bootstrapping
# storing host & parasite membership/phylogenetic-distance correlations after the different permutation
spk.1<-1:1000
spk.2<-1:1000
spk.3<-1:1000
shk.1<-1:1000
shk.2<-1:1000
shk.3<-1:1000

for(i in 1:1000){
  Y2<-apply(Y, 2, sample)                    # Legendre sampling
  G2<-graph.incidence(Y2)
  MU2<-walktrap.community(G2)
  MU2<-membership(MU2)
  shk.1[i]<-cor(outer(MU2[1:length(hi)], MU2[1:length(hi)], "==")[lth], c(1-ht)[lth])
  spk.1[i]<-cor(outer(MU2[length(hi)+1:length(pi)], MU2[length(hi)+1:length(pi)], "==")[ltp], c(1-pt)[lt

  Y2<-apply(Y, 1, sample)                    # Legendre sampling
  G2<-graph.incidence(Y2)
  MU2<-walktrap.community(G2)
  MU2<-membership(MU2)
  spk.2[i]<-cor(outer(MU2[1:length(pi)], MU2[1:length(pi)], "==")[ltp], c(1-pt)[ltp])
```

```
  shk.2[i]<-cor(outer(MU2[length(pi)+1:length(hi)], MU2[length(pi)+1:length(hi)], "==")[lth], c(1-ht)[l


  Y2<-Y[sample(1:nrow(Y)),sample(1:ncol(Y))]  # Hommola sampling
  G2<-graph.incidence(Y2)
  MU2<-walktrap.community(G2)
  MU2<-membership(MU2)
  shk.3[i]<-cor(outer(MU2[1:length(hi)], MU2[1:length(hi)], "==")[lth], c(1-ht)[lth])
  spk.3[i]<-cor(outer(MU2[length(hi)+1:length(pi)], MU2[length(hi)+1:length(pi)], "==")[ltp], c(1-pt)[l
  #print(i)
}
k.tests<-cbind(c(chk, cpk),
               c(sum(chk>shk.1)/1000, sum(cpk>spk.1)/1000),
               c(sum(chk>shk.2)/1000, sum(cpk>spk.2)/1000),
               c(sum(chk>shk.3)/1000, sum(cpk>spk.3)/1000))
# store metrics, and the proportion of times the metrics under permutation were greater
rownames(k.tests)<-c("H", "P")
colnames(k.tests)<-c("statistic", "L-pval", "L2-pval", "H-pval")
Krasnov_results <- k.tests
saveRDS(Krasnov_results, file="Krasnov_results.RDS")} else {
  Krasnov_results <- readRDS("Krasnov_results.RDS")
  Krasnov_results
}
```

```
##     statistic L-pval L2-pval H-pval
## H -0.48708412  0.000   0.000  0.000
## P  0.02804845  0.822   0.861  0.859
```

All of the results so far seem to point towards a strong role for the host phylogeny in structuring host-parasite associations. We can get one final piece of evidence for this by employing the method of Hadfield et al. (2014). This is an extension of the method of Ives and Godfray that uses a Bayesian Markov chain Monte Carlo approach to directly estimate the contribution of several different sources of variance to the observed covariance between pairs of interacting hosts and parasites. In particular, it allows us to estimate the contribution of host phylogeny, parasite phylogeny, host evolutionary interactions, parasite evolutionary interactions, and host-parasite coevolution, as well as estimating the contributions of variation that are not due to the phylogeny.

```
res <- readRDS("mI_MCMCb.RDS")
## Another analysis to potentially run: we can compare the host ranges of two parasites
## by computing the phylogenetic distance between the hosts of two parasites. The smaller
## this distance, the more similar the host ranges of the two parasites. To identify
## putative cases of "niche partitioning", we find parasites with very low PD between
## host ranges, but whose hosts don't overlap at all (e.g., the parasitize related, but
## not identical, hosts). You can do a similar analysis, comparing the PD between the
## parasites that infect different hosts
```