Co-phylogenetic analyses

Background

Zoonoses, pathogens transmitted between humans and animal taxa, are a rising threat to public health and wildlife conservation. With the increase in emergence of zoonotic infectious diseases, research to predict and survey potentially zoonotic host-pathogen interactions is critical (Cleavland et al. 2001; Taylor et al. 2001; Han et al. 2016; Olival et al. 2017). Recent work indicates that hosts infected by phylogenetically diverse pathogens are at a higher risk for disease spillover (Olival et al. 2017). Additionally, host shifts are more likely to occur between closely related hosts (Davies & Pederson 2008; Ramsden et al. 2009; Cooper et al. 2012; Huang et al. 2013; Longdon et al. 2014), suggesting that close relatives of humans infected by phylogenetically diverse pathogens are most likely to cause a spillover event. Understanding which ecological and evolutionary components of host-pathogen interactions lead to spillover events remains challenging but necessary to more accurately surrey potential emerging infectious diseases.

The ecological and evolutionary components of host-pathogen interactions can be explained by four hypotheses. The first is that host-pathogen associations could be random with respect to both the host and pathogen phylogenies, in which host-pathogen interactions are likely driven by abiotic or biotic factors rather than evolution. For instance, associations between nematodes and their stick insect hosts are likely determined by geographic separation (Larose & Schwander 2016). In contrast with the first hypothesis, the second suggests that host-pathogen associations are driven by codivergence, where closely related hosts are infected by closely related pathogens. This pattern has been observed in pocket gophers with chewing lice (Demastes et al. 2012; CITE). The last hypotheses assume that the pattern of host-pathogen association is largely driven by either host or pathogen evolution. For instance, if the pattern was driven by host evolution, closely related hosts would be infected by similar pathogens, regardless of the relatedness of the pathogens (shown in McCoy et al. 2001; Streicker et al. 2010). Similarly, if the pattern was driven by pathogen evolution, closely related pathogens would infect similar hosts, regardless of the relatedness of those hosts (indicated in . . .).

Determining which of these hypotheses provides the best fit with data can be used to identify potential zoonotic threats. For example, if neither host or pathogen evolution shapes host-pathogen interactions, then determining the ecological factors, abiotic and biotic, that shape these interactions is the first step in identifying potential sources of emerging infectious diseases (host geographic range or parasite life history. . . CITES). In contrast, if host-pathogen associations are driven by codivergence, then the pathogens most likely to spill into humans will be pathogens that both infect close relatives, i.e. the great apes, and are closely related to pathogens already infecting humans. If host evolution alone plays the dominant role, then surveillance should shift attention to the pathogens of humans' close relatives, regardless of whether they are similar to pathogens that already infect humans. Surveying primates for potentially emerging pathogens is a relatively straightforward task and is one that researchers are already pursuing (see. . . CITES. . . Nunn or Goldberg?). On the other hand, if parasite evolution alone plays the dominant role, surveying potential zoonoses is more difficult because it would require identifying the close relatives of the pathogens that already infect humans and surveying their current hosts.

We use host-pathogen databases containing associations between helminth taxa and free-range mammals (GMPD, Stephens et al. 2017; Nearctic dataset, Botero-Cañola 2020) in combination with reconstructed host and pathogen molecular phylogenies (Faurby et al. 2018; Pfenning-Butterworth et al. ???) to assess which hypothesis, stated above, best fit helminth-mammal associations. We conducted co-phylogenetic analyses using multiple methods (CITATIONS). Briefly, we first used ParaFit to evaluate the evidence for codivergence in the data (Legendre et al. 2012). Then, we tested for phylogenetic signal in the pattern of helminth-mammal associations following Ives and Godfray (2006). Next, we .... (Hommla et al. 2009). Lastly, we estimated the modularity of helminth-mammal associations following Krasnov et al. (2012).

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

The data for this analysis is comprised of three things: (1) an incidence matrix of host-parasite interactions; (2) a host phylogeny; (3) a parasite phylogeny. To determine whether there are any general patterns in the effect of host and parasite (co)evolution on host-parasite interactions, we work with two different datasets of host-parasite interaction: the first comes from the Global Mammal Parasite Database, and the second from a database of museum-verified Nearctic host-parasite associations. Additionally, we analyze these patterns using only the data on Primate-parasite associations in the GMPD. With these data, we can carry out several different co-phylogenetic analyses using different methods that have been proposed in the literature.

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 computes a so-called "fourth corner" statistic (Legendre) based on the product of matrices describing (A) the presence/absence of each host-parasite association; (B) the parasite phylogenetic tree; and (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.)

The second 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. We compute this correlation in our three datasets, and then compare the observed correlation against randomly permuted data, using the same set of permutations described above. Like the method of Legendre et al., this test can provide evidence for phylogenies affecting host-parasite associations.

The third and fourth methods, in contrast to the first two, provide additional information that can help us to determine whether the pattern of host-parasite association is more strongly structured by host evolutionary history, parasite evolutionary history, or coevolution. The third method is that of Ives and Godfray (2006); 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. This is in contrast to the related method of Legendre et al., which holds the branch lengths constant. Statistically speaking, the Legendre et al. method assumes that the covariance between species is determined by the branch lengths under a Brownian motion evolutionary process. In contrast, the Ives and Godfray method adjusts the covariance between any two tips in either tree to maximize the fit of the evolutionary model. This covariance is specified by an Ornstein-Uhlenbeck process with a parameter d that determines the strength of the phylogenetic signal; this 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 covariance between tips, which can be descrbied as a star phylogeny, whereas $d = 1$ implies no selection and a covariance that is described by Brownian motion. A value of $0 < d < 1$ implies some amount of stabilizing selection, and $d > 1$ suggests disruptive selection. Since there are two phylogenies, 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. 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.

We use the method in full for the Nearctic and Primate datasets, first fitting simplified versions of the model that either fix $d_h = d_p = 0$ (star phylogeny) or $d_h = d_p = 1$ (Brownian motion), then fitting the full model that fits d_h and d_p to the data. Unfortunately, the GMPD dataset is too large for the full method to work: both the code provided in the supplementary material of Ives and Godfray (2006) and in the R package picante (ref) that implements the method generate matrices that are so large that they consume all available computer memory.

However, Hadfield et al. (2014) provide code for computing a statistic that is proportional to mean square error under the assumption that $d_h = d_p = 1$ (Brownian motion evolution); their code does not generate large matrices, and is thus possible to apply all three datasets. This statistic is very similar to that of Legendre et al. (2002). It is also possible to compute the mean square error under the assumption that $d_h = d_p = 0$ (star

phylogeny) for all three models. We therefore compute these statistics for all three datasets, and use the same set of permutations described above to evaluate whether these two statistics are different from a null distribution based on randomly permuted host-parasite assocations. This provides an alternative metric to evaluate the evidence for phylogenetic signal in the pattern of host-parasite association.

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). 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. It does this by calculating the correlation between comembership in a module and phylogenetic distance, that is, it asks whether, across all hosts (and parasites), is the pairwise phylogenetic distance between hosts (or parasites) within a single module less than the pairwise phylogenetic distance between hosts in separate modules. If so, then there is evidence for phylogeny structuring host-parasite associations. Again, we use permutations of the host-parasite association data to determine whether the observed correlations between comembership and phylogenetic distance are different from what you would expect for randomly constructed host-parasite association networks.

Note for Jonathan: we also attempted to run the full method of Hadfield et al. (2014), which is an extension of the approach of Ives and Godfray that uses a Bayesian approach to estimate the contribution of many different terms to the covariance observed in the host-parasite association data. In particular, Hadfield et al. estimate (1) whether the number of parasites infecting each host is explained by host phylogeny (the host evolutionary effect); (2) whether host range of each parasite is explained by parasite phylogeny (the parasite evolutionary effect); (3) whether related hosts are infected by similar parasites (the host evolutionary interaction effect); (4) whether related parasites infect similar hosts (the parasite evolutionary interaction); and (5) whether related parasites infect related hosts (the coevolutionary interaction). We were able to obtain preliminary results from a shorter run (100,000 iterations), but that had not yet converged. Runs of the model take several weeks and during a longer run, my computer crashed (perhaps due to the run itself). It's unclear to us whether it makes sense to stub in these results now, with the goal of having a more complete run to analyze by the time of revisions, or whether to just leave them out entirely.

Results

Legendre method Using the Legendre et al. (2002) method, we can see good evidence for coevolution, host and parasite evolutionary interactions, and for phylogenetic signal in parasite species richness and host range in all three datasets. The observed value of the test statistic (which is meaningless in and of itself) is more extreme than the values observed for almost all of the bootstrap permutations, using any method of permutation. Thus, we conclude that there appears to be strong evidence for phylogenetic signal in the pattern of host-parasite associations, though we are not really able to determine whether that is more due to the host, or more due to the parasite, using this method.

```
res <- readRDS("legendre_results.RDS")
res
```

```
##      dataset    value              L1-pval L2-pval H-pval
## [1,] "GMPD"     "637.819049188227" "0"     "0"     "0"
## [2,] "Primates" "6.69776365341545" "0.004" "0"     "0.001"
## [3,] "Nearctic" "12.3106371966957" "0"     "0"     "0"
```

Hommola method The results indicate that there is fairly low correlation between host and parasite shared branch lengths observed in the GMPD data ($r < 0.001$). Moreover, this correlation is fairly likely under random perturbations of the host-parasite association data, with bootstrap p-values ranging from 0.33 to 0.37. Similarly, there is essentially no correlation between shared branch lengths in the Nearctic dataset

either. However, there is a significant positive correlation in the Primate dataset, suggesting that there is a detectable phylogenetic signal in the pattern of host-parasite associations among Primates and their parasites.

```
res <- readRDS("hommola_results.RDS")
res
```

```
##      dataset    value                    L1-pval L2-pval H-pval
## [1,] "GMPD"     "0.000976480200552879"  "0.353" "0.365" "0.334"
## [2,] "Primates" "0.0623094454501342"    "0.001" "0.115" "0.071"
## [3,] "Nearctic" "-0.000468175281842404" "0.365" "0.383" "0.336"
```

Ives & Godfray method For the Nearctic dataset, the best-fitting evolutionary model was a star phylogeny ($d_h = d_p = 0$). This can be seen both in the mean square error estimates of the different models and from the estimates of d_h and d_p in the full method. For the Primate dataset, however, the best-fitting evolutionary model was the full model that fit both d_h and d_p. This is despite the fact that the values of d_h and d_p are both very small. However, the value of d_h is significantly larger than that of d_p, suggesting a larger role for the host phylogeny in shaping host-parasite associations among Primates and their parasites.

```
ig_method_nearctic <- readRDS("IvesGodfray_full_nearctic.RDS")
## MSE estimates for each each model fitted to the Nearctic dataset
ig_method_nearctic$MSE
```

```
##     MSETotal   MSEFull    MSEStar  MSEBase
## 1 0.05731509 0.0589009 0.05731509 30.49405
```

```
ig_method_primate <- readRDS("IvesGodfray_full_primate.RDS")
## MSE estimates for each each model fitted to the Priamte dataset
ig_method_primate$MSE
```

```
##     MSETotal    MSEFull    MSEStar  MSEBase
## 1 0.05893056 0.05601232 0.05893056 1.921781
```

```
nearctic_boots <- readRDS("IvesGodfray_full_nearctic_bootstraps.RDS")
primate_boots <- readRDS("IvesGodfray_full_primate_bootstraps.RDS")

signal_estimates <- array(NA, dim=c(2,3))
colnames(signal_estimates) <- c("dataset", "d_h", "d_p")
c("Nearctic",
  paste0(signif(ig_method_nearctic$signal.strength$estimate[1],3), " (",
         signif((lapply(nearctic_boots, function(b) b[[1]]) %>% unlist %>% sort)[3],3), ", ",
         signif((lapply(nearctic_boots, function(b) b[[1]]) %>% unlist %>% sort)[97],3),")"),
  paste0(signif(ig_method_nearctic$signal.strength$estimate[2],3), " (",
         signif((lapply(nearctic_boots, function(b) b[[2]]) %>% unlist %>% sort)[3],3), ", ",
         signif((lapply(nearctic_boots, function(b) b[[2]]) %>% unlist %>% sort)[97],3),")")) -> signal_
c("Primates",
  paste0(signif(ig_method_primate$signal.strength$estimate[1],3), " (",
         signif((lapply(primate_boots, function(b) b[[1]]) %>% unlist %>% sort)[3],3), ", ",
         signif((lapply(primate_boots, function(b) b[[1]]) %>% unlist %>% sort)[97],3),")"),
  paste0(signif(ig_method_primate$signal.strength$estimate[2],3), " (",
         signif((lapply(primate_boots, function(b) b[[2]]) %>% unlist %>% sort)[3],3), ", ",
         signif((lapply(primate_boots, function(b) b[[2]]) %>% unlist %>% sort)[97],3),")")) -> signal_
```

Interestingly, if we look at the permutation test results, we see that the fit of *both* the star phylogeny and a Brownian motion model to the real data in all three datasets is significantly different from random. This is difficult to interpret, as the star phylogeny and Brownian motion are often set in opposition to one another in phylogenetic comparative analysis: one suggests a complete lack of phylogenetic signal, whereas the other

suggests a strong phylogenetic signal. Here, we interpret these results merely as indicating that there is significant structure in the pattern of host-parasite association that may reflect both strong independent adaptation (as suggested by the star phylogeny) and by some phylogenetic constraint.

```
res <- readRDS("IvesGodfray_simplified_results.RDS")
res
```

```
##      dataset    model   value                L1-pval L2-pval H-pval
## [1,] "GMPD"     "brown" "339667.560349322"   "0.009" "0.002" "0.361"
## [2,] "GMPD"     "star"  "0.019480127589224"  "0"     "0"     "0"
## [3,] "Primates" "brown" "3453.44944604846"   "0.006" "0"     "0.047"
## [4,] "Primates" "star"  "0.058930562184038"  "0"     "0"     "0"
## [5,] "Nearctic" "brown" "189337.530222331"   "0.016" "0.1"   "0.158"
## [6,] "Nearctic" "star"  "0.0573150916712507" "0"     "0"     "0"
```

Krasnov method For the GMPD dataset, we detected 38 distinct host-parasite clusters, with an the overall modularity score of 0.63 (Fig. @ref(fig:gmpd-module-graph). 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.

For the Nearctic dataset, we detected 17 distinct host-parasite clusters, with an the overall modularity score of 0.5 (Fig. @ref(fig:nearctic-module-graph)). Similar to the GMPD dataset, there are several large modules in the Nearctic dataset, although the overall pattern reveals more connections between modules than was evident in the GMPD data. Some of these modules appear to be structured by parasite. For example, the largest network contains *all* of the Artiodactyla in the dataset. The next largest network contains 46% of the Carnivores in the dataset, although Carnviores are also common in the third-largest network (which also contains marsupials and rodents). The overall higher interconnectedness of the network, and lower modularity, appears to be driven by the inclusion of rodents in this dataset, as rodents appear in 12 of the 17 modules.

For the Primate dataset, we detected 20 distinct host-parasite clusters, with an the overall modularity score of 0.53. There are many very small modules, comprising a single host-parasite pair. The largest module (of 23 species), contains 18 species of Great Apes (both *Pan* species), lesser apes, and Old and New World monkeys.

Looking across the datasets, there was always a very strong and significant negative correlation between comembership in a module and phylogenetic distance for hosts, indicating that more closely related hosts are more likely to end up in the same module. However, parasite relatedness was never significantly related to membership in a module.

```
res <- readRDS("Krasnov_results.RDS")
res
```

```
##      dataset    host-corr            L1-pval L2-pval H-pval
## [1,] "GMPD"     "-0.487084121706913" "0"     "0"     "0"
## [2,] "Primates" "-0.215131659453853" "0"     "0"     "0"
## [3,] "Nearctic" "-0.269710261492701" "0"     "0"     "0"
##      parasite-corr         L1-pval L2-pval H-pval
## [1,] "0.0280484528908893"  "0.792" "0.828" "0.857"
## [2,] "0.0171651827625029"  "0.502" "0.558" "0.59"
## [3,] "-0.0437206964798904" "0.165" "0.143" "0.092"
```

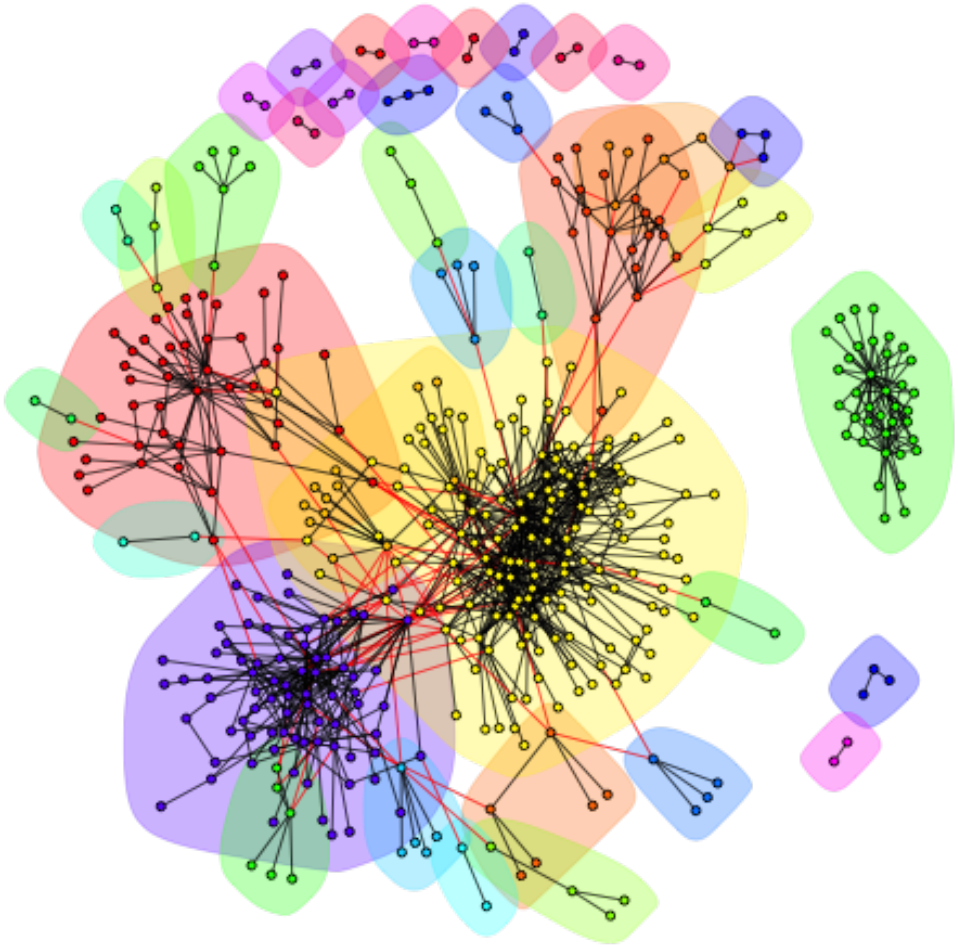

Figure 1: Host-parasite interaction network within the GMPD data, broken into 38 distinct modules.

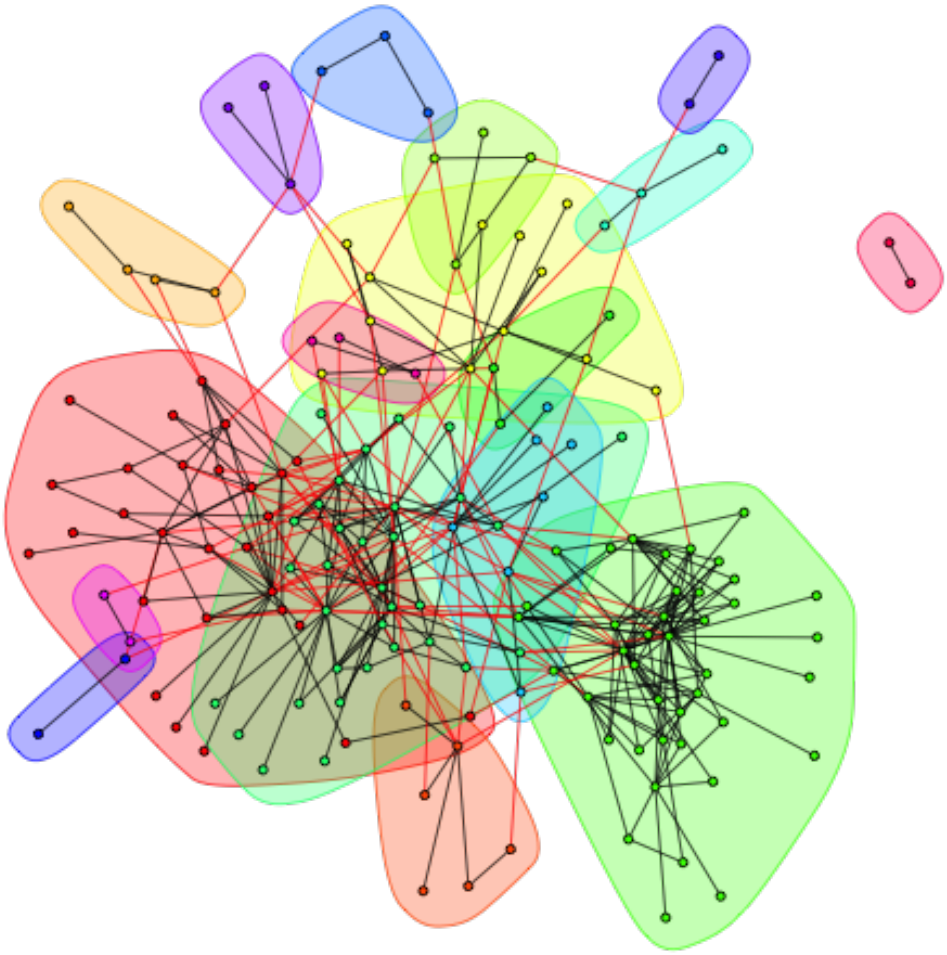

Figure 2: Host-parasite interaction network within the Nearctic dataset, broken into 38 distinct modules.

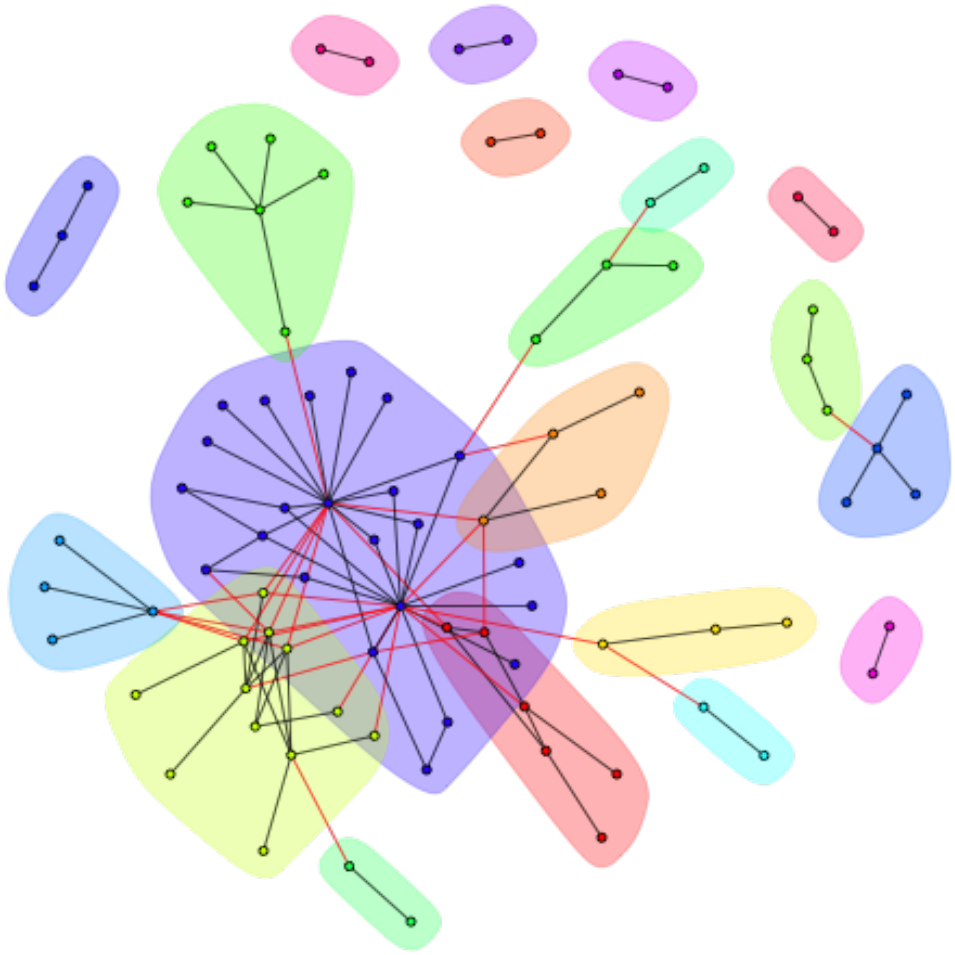

Figure 3: Host-parasite interaction network within the Primate dataset, broken into 20 distinct modules.

All of the results so far seem to point towards a strong role for the host phylogeny in structuring host-parasite associations.