# Phylogenetic Signal of Sub-Arctic Beetle Communities

Samantha E. Majoros and Sarah J. Adamowicz
samjoros@uguelph.ca, sadamowi@uoguelph.ca
https://orcid.org/0000-0001-8788-5523
University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1

Post-glacial dispersal and colonization processes have shaped community patterns in sub-Arctic regions such as Churchill, Manitoba, Canada. Important questions remain about the species that colonized this area, in particular the role of glacial history and biological traits in governing colonization patterns from refugial and southerly geographic regions. This study quantifies sub-Arctic beetle phylogenetic community structure using the net relatedness index (NRI) and nearest taxon index (NTI); calculated using publicly available data from BOLD; compares patterns across families with different traits (habitat, diet) using standard statistical analysis (ANOVA) as well as phylogenetic generalized least squares (PGLS) using a higher-level beetle phylogeny; and compares phylogenetic community structure in Churchill with a region in southern Canada (Guelph, Ontario). The dominant pattern detected in our study was that aquatic families were much better represented in Churchill compared to terrestrial families, when compared against richness sampled from across Canada and Alaska. Individually, most families showed significant phylogenetic clustering in Churchill. Closely related species were likely found together due to the strong environmental filtering present in Arctic environments. There was no significant difference in phylogenetic structure between Churchill and Guelph, although the trend was towards stronger clustering in the North. Similarly, there was no difference in phylogenetic structure metrics calculated for aquatic vs. terrestrial beetle families, again with a trend towards stronger clustering in water beetles. By contrast, there was a

significant relationship between traits and community structure. Predators showed significantly stronger clustering in Churchill compared to other feeding modes, perhaps due to phylogenetic conservatism of their overwintering ability or generalist diet of some clades within families. This study contributes to our understanding of the traits and processes structuring insect biodiversity and macroecological trends in the sub-Arctic.

Keywords: Phylogenetic community structure, Arctic, Entomology, Biogeography, Environmental filtering, Macroecology

Declarations

Authors acknowledge financial support for this project from the Ted Morwick Research Assistantship in Aquatic Biology to SEM and by a Discovery Grant to SJA from the Natural Sciences and Engineering Research Council of Canada (NSERC). Thank you to Matthew Orton, Jacqueline May and Cameron Nugent for assistance with sections of the R code. Thank you to Alex Smith and Kamil Chatila-Amos for helpful project input and discussions. This project was designed by Samantha Majoros and Sarah Adamowicz. The code design, analysis and writing were done by Samantha Majoros, with edits and input from Sarah Adamowicz. We are grateful to the dozens of researchers who made their data available on the Barcode of Life Data Systems (BOLD).

# Phylogenetic Signal of Sub-Arctic Beetle Communities

## Introduction

The Arctic is a land of change (Pielou 1995). Glaciation changed, or largely eliminated, the communities inhabiting sub-Arctic areas such as Churchill, Manitoba, Canada (Pielou 1995). Ever since the last glacial maximum, post-glacial colonization has been ongoing in Arctic North America, with species coming from both the south and from the Beringian glacial refuge (Pielou 1995, Woodcock *et al*. 2013). While diversity in general tends to decrease with latitude, Arctic environments still provide a diverse range of habitats and niches in which organisms exist (Danks 1992, Woodcock *et al.* 2013). As the climate shifts, these communities and habitats are experiencing rapid changes; this may be due to increasing temperature, melting sea ice, increased greenery, changing nutrient levels, or invading species (Walseng *et al.* 2018). Important questions remain about Arctic biodiversity, such as what species and traits make up Arctic communities, where did they colonize from, what patterns exist in their community structure, and how will these patterns shift in the future? With ongoing climate change, it is important to understand the traits of Arctic and sub-Arctic species, as well as to predict how their geographic ranges and community structure may shift in the future.

Investigating evolutionary community structure can help us understand the relationships among species in Arctic communities and their distribution patterns. Phylogenetic community structure metrics are used to quantify the relatedness among cohabiting species against patterns in a broader source community (Webb 2000, Webb *et al*. 2002, Kraft *et al.* 2007, Mayfield & Levine 2010, Emerson *et al.* 2011, Smith *et al*. 2014, Boyle & Adamowicz 2015). Are the species found in a local community more closely related than those in a broader community? What does this tell us about the mechanisms underlying their relationships and distributions?

In order to identify and understand the phylogenetic relationships between species, it is beneficial to analyze DNA sequence data, which is a rich source of data for inferring relationships (Hebert *et al*. 2003, Hebert & Gregory 2005). DNA barcodes are standardized DNA sequences that are used for specimen identification and species discovery (Hebert *et al*. 2003, Hebert & Gregory 2005). The barcode most commonly used for animals is an approximately 658 base pair region of cytochrome c oxidase subunit I (COI), a mitochondrial gene (Wilson 2010, Wilson 2011, Smith *et al*. 2014, Boyle & Adamowicz 2015). DNA barcoding allows for data to be readily available to other scientists through data banks like the Barcode of Life Data Systems (BOLD), which contains a large collection of geo-referenced specimens from locations around the world (Ratnasingham & Hebert 2007). This study leverages publicly available, geo-referenced sequence data for beetles from BOLD, combined with a published multi-gene backbone phylogeny (Hunt *et al*. 2007), to combine the merits of both approaches for community phylogenetics (Boyle & Adamowicz, 2015).

Various patterns can occur in phylogenetic community structure, including patterns of clustering, overdispersal, or random (Webb 2000, Webb *et al*. 2002). A clustered pattern occurs when closely related species are found together more often than expected by chance, often caused by environmental filtering (Fig.1a) (Kraft *et al*. 2007, Emerson *et al*. 2011, Weiher *et al*. 2011, Smith *et al*. 2014, Boyle & Adamowicz 2015). In this case, cohabiting species typically share the traits needed to survive in a given environment and are therefore found in the same region, while more distantly related species that lack these traits are excluded. Overdispersion occurs when closely related species cohabit in the same local community less than is expected (Fig.1b) (Kraft *et al*. 2007, Mayfield & Levine 2010, Emerson *et al.* 2011, Weiher *et al.* 2011, Boyle & Adamowicz 2015). This is often interpreted as evidence for competitive exclusion,

whereby closely related species compete for the same resource, and this results in one species being forced out of the environment or into a different niche (Kraft *et al.* 2007, Emerson *et al.* 2011, Weiher *et al.* 2011, Boyle & Adamowicz 2015). However, it is difficult to draw conclusion about mechanisms and the causes of these patterns based on the phylogenetic patterns alone. Mayfield & Levine (2010) suggest that competitive exclusion can also cause clustering. If competitive ability is phylogenetically clustered and is more important for surviving in the environment than niche differences, we can expect competitive exclusion to cause clustering rather than overdispersion (Mayfield & Levine 2010). In order to draw conclusions about mechanisms, it may be beneficial to examine traits rather than community phylogenetic patterns alone.

There are various environmental and biotic factors that may influence the phylogenetic structure of communities, and these may change with latitude. Factors such as the strength of competition and environmental filtering change across latitude with Danks (1993), suggesting that the relative importance of competition for resources decreases as latitude increases. In northern environments, the climate and environmental factors are more important than biotic interactions when determining the survival of populations in the environment (Ernst & Buddle 2015), which would be expected to result in a more clustered phylogenetic community structure. The traits of the species within a community, such as diet or lifestyle, can also affect the phylogenetic structure (Mayfield & Levine 2010). For example, Poulin *et al*. (2011) found that closely related parasitic species are found together in local communities more than expected, likely due to closely related species having similar hosts. If these hosts are clustered geographically, we can expect the same of the parasites (Poulin *et al*. 2011, Eagalle & Smith 2017). Similarly, Vamosi & Vamosi (2007) discussed the effects of an aquatic lifestyle on

community structure with dytiscid beetle communities in the lakes of Alberta and showing phylogenetic clustering. This may have been caused by a decrease in the importance of competition and an increase in environmental filtering in aquatic systems relative to terrestrial (Vamosi & Vamosi 2007). In order to survive in aquatic environments, species need to have a certain set of physiological tolerances, and environmental factors such as salinity and pH influence the diversity (Heino *et al*. 2016) and composition of species found in the environment (Vamosi & Vamosi 2007). However, different processes interact to determine species survival and co-existence, and it may be difficult to pinpoint one cause or mechanism (Peres-Neto *et al*. 2012). Across these varied examples, the lifestyles and characteristics of the species influence the community structure.

While prior studies have investigated clustering patterns and community structure within specific taxa and locations, few have compared these patterns across taxa or investigated how community structure is related to traits (Kraft *et al*. 2007, Vamosi & Vamosi 2007, Poulin *et al*. 2011, Weiher *et al*. 2011). In this study, we investigate the patterns that occur in phylogenetic community structure at a species level across taxa and traits and investigate the phylogenetic relatedness of species inhabiting the sub-Arctic site of Churchill, Manitoba using northern North America as the regional species pool. This study allows us to investigate what traits, such as feeding modes and habitat preferences, are relatively more prevalent in Arctic communities and whether families with these traits tend to exhibit phylogenetic clustering. By understanding the current traits and community structure, and how these relate to environmental factors, we can better prepare for the changes likely to occur in the future.

The focal organisms for this study are sub-Arctic Coleoptera. Beetles are understudied in previous community structure research yet are hyper-diverse, with species occupying a variety of

niches and habitats and exhibiting substantial variability in traits (Woodcock *et al.* 2013). There are also 322,157 public records available on the BOLD database as of June 25, 2019. Particularly, we will be focusing on the Churchill region as there has been a concerted effort to barcode fauna in northern communities, particularly Churchill (e.g. Woodcock *et al*. 2013, Zhou *et al*. 2009, 2010). In the BOLD database, there are 306 recorded species of Coleoptera in Churchill as of June 25, 2019 (Ratnasingham & Hebert 2007).

We hypothesize that environmental filtering will impact community structure of sub-Arctic communities due to the harsh environmental conditions present at higher latitudes. Specifically, we predict that the species in Churchill will present a significantly clustered pattern when compared against the broader North America species phylogeny. When comparing other regions within North America, we expect the regions found at higher latitudes to show a more significant clustered pattern. Secondly, we hypothesize that the traits and characteristics of the species will influence the community structure. We predict that taxonomic groups with traits that expose them to more environmental filtering, such as being aquatic, or relying on a host species, such as being a parasite or parasitoid, will have a more clustered pattern than their terrestrial and free-feeding counterparts.

## Methods

### Data and Taxa

Using BOLD's application programming interface (API), all data for this study were pulled from the BOLD database [June 19$^{th}$ 2019] directly into the R environment (Supplementary Material Appendix 1). All coding was done in R version 3.5.0 (R Core Team, 2013). Data for both Canada and Alaska were used as the regional species pool and compared to

the data from Churchill, which will be defined as the local community for this study. Coleoptera families were retained for analysis if they were represented by three or more BINs (Barcode Index Number; Ratnasingham and Hebert 2013), in Churchill.

Filtering Data and Defining Churchill

Once the sequences and metadata had been pulled from BOLD, the data were filtered. DNA sequences without a BIN or GPS coordinates were removed. Sequences were also removed if they were not from the COI-5P marker, if they had internal missing data ('N' nucleotides) or gap content greater than 1% of the sequence length, or were less than 500 base pairs. The decision to use COI is explained in Supplementary Material Appendix 2. The sequences were aligned within each BIN in order to choose a representative sequence for each BIN, defined as the sequence with the minimum average distance to all others in its BIN (as in Orton *et al.,* 2018). Alignments were preformed using the muscle algorithm (Edgar 2004) with the following parameters: maxiters equaled 3, diags equaled true, and gapopen equaled -3000. These parameters were chosen in order to limit the number of iterations for optimization to allow for an alignment to be quickly generated. Then, the selected centroids (one per BIN) were aligned within each family. A preliminary alignment was performed with the above parameters in order to trim the sequences and to screen for outliers. These sequences were then aligned using a reference sequence. A reference BIN that met the following criteria was selected from the public data on BOLD; it contained at least 10 CO1-5P sequences, it had at least one specimen photograph and did not have taxonomic conflicts at order level or above. The reference sequence was chosen from this BIN and had to be 658 base pairs long, have 2 trace file chromatograms and no missing information or stop codons. The final alignment was performed using the package muscle (Edgar 2004) with the same settings as the previous alignments, but with the

default maxiters parameter (maxiter = 8 in R implementation using muscle package) (Edgar 2004). The gap opening penalty is based on preliminary analyses performed by Orton *et al.* (2018) on taxonomic groups that contained gap regions (amino acid insertions or deletions in the COI barcode region). This gap opening penalty provided biologically realistic alignments that preserved amino acid alignment homology across taxonomic groups (Orton *et al*. 2018). The centroid, alignment, and filtering code were adapted from publicly available code by May (2017) and Orton *et al*. (2018).

After the data were filtered, a Churchill subset was defined using coordinates: a latitude between 58.6 and 58.7 degrees and a longitude between -94.2 and -93.8 degrees. These coordinates were found using Google Earth (Google, 2018) and based on a map provided in Boyle (2012) that showed the accessible areas in the vicinity of Churchill, MB, included in prior DNA barcoding research. This map is compatible with maps in other Churchill-related DNA barcoding papers (e.g. Zhou et al. 2009, 2010; Woodcock et al. 2013).

Community Phylogenetic Metrics

In order to test for phylogenetic clustering and overdispersion, we calculated net relatedness index (NRI) and nearest taxon index (NTI); the calculation of these metrics requires a phylogeny as one of the inputs. First, we generated a maximum likelihood tree for each Coleoptera family using one sequence per BIN for all BINs present in Canada and Alaska. The family level was chosen because beetle families often share important traits, such as feeding mode (Hunt *et al*. 2007). Before reconstructing the phylogenies, we first estimated the best-fit model of nucleotide evolution using the R package phangorn version 2.4.0 (Schliep 2011). The model with the lowest Bayesian Information Criterion (BIC) score was chosen, and the proportion of invariant sites was determined based on the fitted model. The number of intervals

of discrete gamma distribution (the k value) was set to 4. A neighbour joining tree, generated using the function NJ from phangorn version 2.4.0 (Schliep 2011), was used as the guide tree. The maximum likelihood trees were generated using the function optim.pml from phangorn version 2.4.0 (Schliep 2011) and optNni, optGamma and optInv were set to true. These trees were used in the NRI and NTI analysis. NRI and NTI calculate the pairwise distance between two species and use this to estimate the community relatedness (Webb 2000) (Supplementary Material Appendix 3). These calculations were preformed using the R package picante version 1.7 (Kembel *et al*. 2010) and the null model "taxa.labels", which indicates that random draws of the same species richness as the Churchill community were made from each family phylogeny; and NRI and NTI are re-calculated with each randomization. The analysis was repeated 1000 times. The observed NRI and NTI values were then compared against the null distribution to obtain a p-value. These tests determined whether species inhabiting the Churchill region are more significantly phylogenetically clustered than expected by chance, when compared against the phylogeny of DNA barcoded beetles of northern North America.

Kraft *et al*. (2007) state that the power for the NRI and NTI analysis is highest when local species richness is 30-60% of regional species richness. For the Coleoptera of Churchill, all families are below this range except for Dytiscidae, Gyrinidae, and Haliplidae. To determine the effects of this, a sensitivity analysis was performed (Supplementary Material Appendix 4). A Holm-Bonferroni correction was done for the p-values in order to account for the test being performed 16 times.

Community Phylogenetic Metrics for a Temperate Region

In order to compare the phylogenetic community structure patterns in Churchill to a temperate location, the analysis above was repeated for the Guelph region. Guelph was selected

due to its temperate climate and the abundance of data available on the BOLD database (Ratnasingham & Hebert 2007). A Guelph subset was defined using coordinates: a latitude between 43.4 and 43.6 degrees and a longitude between -80.3 and -80.1 degrees, selecting after consulting. These coordinates were found using Google Earth (Google, 2018). In order to determine if the community structure of the Churchill and Guelph subsets were significantly different, a t-test was performed to compare mean NTI and NRI values for beetle families between these sites.

Trait Analysis

For the trait analyses, we investigated whether families with different traits have different phylogenetic community structure, by comparing the NRI/NTI values across trait categories using an ANOVA. First, we created a character matrix for each family. Characters/traits were found for each family based on the literature (Le Conte 1862, Marshall 2007, Slipinski *et al.* 2011). The traits that describe the majority of members of a given family were used; this included habitat (terrestrial or aquatic) and feeding mode (predator, herbivore, or scavenger). We then used a one-way ANOVA to compare the average phylogenetic structure (NRI or NTI metric) of families across trait categories, treating each family as an independent unit (as supported by the results of Pyle 2018). We conducted a second analysis considering phylogenetic relationships among families. We created a family-level phylogenetic tree, i.e. treating each family as one tip, using the phylogenetic hypothesis provided in Hunt *et al*. (2007) based upon three gene regions, and assigned branch lengths of 1, before fitting a phylogenetic generalized least squares (PGLS) model using picante version 1.7 (Kembel *et al*. 2010). This allowed us to determine whether families with particular traits have different clustering patterns while taking

into account the phylogeny of the entire order. The PGLS analysis used Brownian motion as the model of trait evolution and the log-likelihood was maximized for the method.

## Results

### Phylogenetic Clustering Metrics

Sixteen families of Coleoptera were analyzed for the study, following the data filtering steps described above, of which eight showed significant phylogenetic clustering (min p-value = 0.0009, max p-value = 0.045) (Fig.2a). Five showed a non-significant trend toward clustering (min p-value = 0.05, max p-value = 0.44) and three showed a trend toward overdispersion but insignificant (min p-value = 0.58, max p-value = 0.83) (Table 1). For 11 of 16 families, the results for NRI and NTI showed the same trend and significance. For Hydrophilidae, Cryptophagidae and Staphylinidae, NTI suggested significant clustering while NRI was not significantly different from zero. For Gyrinidae and Haliplidae, NRI suggested significant clustering while NTI was insignificant. Though NRI and NTI conflict in significance, both show a trend toward clustering.

After applying the Holm-Bonferroni correction, Cantharidae (original p-value = 0.0009, corrected p-value = 0.0144) was the only family below 0.05. This suggests, that while the results for Cantharidae are very significant, there could be some false positives in the other families who did not meet this threshold.

The same analysis was completed for the Guelph subset. Thirty-two families were analyzed, five of which showed significant phylogenetic clustering (Fig.2b). Nine families showed a trend toward phylogenetic clustering in both NRI and NTI, but this was insignificant. Twelve showed a trend toward overdispersion in both NRI and NTI, but this was also

insignificant. The remaining six families had conflicting trends in NRI and NTI, with NRI showing a trend toward clustering and NTI showing a trend toward overdispersion in five of the families. The remaining family, Nitidulidae, showed clustering in NTI and overdispersion in NRI. The results for NRI and NTI agreed on significance for all but 3 families. Cryptophagidae was significant only in NTI and Scirtidae and Throscidae were significant only in NRI. Overall, Guelph appears to be more overdispersed than Churchill. However, the NRI values (t-statistic = -0.9, p-value = 0.38 and NTI values (t-statistic = -1.53, p-value = 0.14) of the two subsets were not significantly different.

Trait Analysis

Within the families studied in Churchill, only 5 were aquatic while 11 were terrestrial. However, aquatic families have a larger percent of their total BINs found in Churchill (Fig. 3). A Chi Square test was used to determine if there were relatively more BINs in a particular habitat than expected. This test indicated that the distribution of BINs is not independent of habitat ($X$-squared = 23.51, df = 1, p-value = $1.243 \times 10^{-6}$), with aquatic families better represented in the sub-Arctic than terrestrial families. A similar result was shown for feeding mode, with the number of sequences being significantly related to the feeding mode ($X$-squared = 244.59, df =2, p-value = $2.2 \times 10^{-16}$) and predaceous BINs occurring relatively more frequently in the sub-Arctic than other feeding modes. Six families were herbivores, six were predators and four were scavengers. The ANOVA showed no significant relationship between the community structure metrics and the traits of the families (Table 2a). There was no relationship between structure and habitat (F-value=2.02, p-value=0.18, Habitat df =1, Residuals df = 14) and no relationship between structure and feeding mode (F-value=0.71, p-value=0.51, Diet df=2, Residuals df = 13). These results were consistent with both the NRI and NTI values. The results of the PGLS

differed from that of the ANOVA. While community structure was not significantly related to habitat, there was a significant relationship to feeding mode (Table 2b, Fig.4). Predators were significantly more clustered than other feeding modes in both NRI (t-value=2.12, p-value=0.05) and NTI (t-value = 2.57, p-value = 0.02). There are also appeared to be a trend toward increased clustering in aquatic families.

## Discussion

Overall, we found significant clustering in eight of the Coleoptera families studied, and a trend toward clustering in five. This provides support for the hypothesis that due to the harsh conditions present at high latitudes, environmental filtering would be strong in sub-Arctic communities. The species present in the Churchill region possessed the traits needed to survive in this environment, while more distantly related species likely did not. This was not true for all families studied, as three families showed a trend toward overdispersal, though this was insignificant. All families were widely sampled across Canada, though less sampling was done in Northern Canada than Southern. The overdispersed families could potentially still have been experiencing clustering, just at a larger scale, with closely related species being clustered in Canada.

When comparing the Guelph and Churchill subsets, Guelph appeared to be more overdispersed than Churchill, though this difference was not significant. It is possible that Guelph, at 43.5 degrees north, and Canada in general, is still far enough north to experience significant clustering. Guelph was less clustered than Churchill (58.7 degrees north) and if compared to more regions at lower latitudes, it is possible there will be a more pronounced difference. This supports the hypothesis that sub artic communities are experiencing greater environmental filtering. Temperate areas are under less extreme environmental pressures, likely

resulting in stronger competition and more phylogenetically dispersed communities compared to polar regions (Danks 1993).

For some families in both Churchill and Guelph, the NRI and NTI differed in their estimates of significance, or, in some cases, even their predicted trend. This likely relates to the different ways NRI and NTI measure distance between nodes. If NRI suggests clustering, this is due to clustering occurring deeper within the phylogeny (Webb 2000). For NTI, the clustering is occurring within the clades and at the tips of the phylogeny (Webb 2000). For this study, it is beneficial to use both in order to detect clustering patterns at all levels.

Our results were similar to those found in other studies. Ernst & Buddle (2015) found that assemblage structure was correlated with latitude and that climate was more important than biotic factors for determining community structure in northern communities of beetles when species were placed in functional groupings. Similarly, Shibuya *et al*. (2011) found that in the beetle family Carabidae in Japan, the environmental conditions were more important for determining community patterns than competition, and there was actually very little interaction between the beetle species. Carabidae was significantly clustered in our study, in accordance with the findings of Shibuya *et al*. (2011). While Dytisicidae was not significantly clustered in this study, it still showed a trend toward clustering, similar to Vamosi & Vamosi (2007). Not all families exhibited this pattern. The importance of competition, as well as the strength of environmental filtering, likely differs between species, and this results in different community structure patterns, even under harsh environmental conditions. There were some families that exhibited a trend toward overdispersion. Ulrich & Fattorni (2013) found a similar pattern in Tenebrionidae and suggested that this could be due to colonization patterns. Differences in the past colonization patterns of the families could also influence the community structure.

In contrast to our predictions, there was no significant relationship between phylogenetic community structure and habitat in both the ANOVA and PGLS analyses at the family level, though a trend toward increased clustering in aquatic families was shown. Competition is often less important in aquatic communities due to the strong influence of environmental factors (Vamosi & Vamosi 2007, Heino *et al*. 2016). Therefore, we expected aquatic families to be significantly more clustered than terrestrial. Both terrestrial and aquatic families exhibited similar clustering patterns, but only terrestrial families showed any trend toward overdispersion. This could be due to stronger clustering in aquatic habitats. This pattern could also be influenced by the low richness of some of the terrestrial families, as well as low plant species richness in the sub-Arctic. Though there was no significant evidence for a relationship between the phylogenetic clustering and the habitat at the level investigated in this study, there may be evidence for the influence of habitat at a higher taxonomic level. Getting a true representation of terrestrial versus aquatic families was difficult due to the variability within families and the limited families located in Churchill. Only five of the sixteen families studied were aquatic. However, these aquatic families had a larger percent of their total species found in Churchill than terrestrial families, and habitat was shown to be strongly related to the representation of Northern North American BINs that have been found in Churchill. This suggests that it may be easier for aquatic species to colonize the Arctic than terrestrial. In order to better understand this pattern, other locations and taxonomic levels should be investigated. Habitat is likely not the trait determining community structure within families in this study.

However, there was a significant relationship between clustering and feeding mode, with predator families showing significantly more clustering compared to other feeding modes. This pattern could possibly be due to the predator's reliance on their prey species. If the prey is

clustered, so is the predator. However, out of the six predatory families studied, five were generalist predators (Marshall 2006). The sixth family, Coccinellidae, consumes mostly aphids (Marshall 2006). This is also the only predator family that showed a trend toward overdispersion. Therefore, the general diet of most of the families suggests that the clustering pattern observed is not dependent on their prey and that having a more general diet is beneficial for surviving in the Arctic. It is also possible that there are limits to the vegetation available in Arctic climates, therefore limiting the survival and diversity of herbivores. Another possible explanation is that these predators are able to survive in these northern habitats due to their cold tolerance and overwintering abilities. Predacious families such as Coccinellidae and Carabidae have overwintering strategies that allow for survival in cold temperatures (Knapp & Saska 2011, Hamedi *et al*. 2013). This includes occupying microhabitats that buffer the effects of the climate, lowering temperature thresholds for activity or increasing cryoprotectant concentrations (Knapp & Saska 2011, Hamedi *et al*. 2013). If these traits are phylogenetically conserved, this would result in clustering. Diet has been shown to be important in other studies, such as Poulin *et al*. (2011), who found clustering in parasitic families due to their reliance on specific host species. Like habitat, feeding mode was shown to influence the number of individuals found in Churchill, with predacious families having a larger percent of their total species found in Churchill compared to other feeding modes. This suggest that predators may be more likely to colonize Arctic environments and that feeding mode is an important part of determining community structure in Churchill. Overall, there was support for our hypothesis that traits would impact phylogenetic community structure. The idea that traits are important for determining community structure is also supported by the literature; Mayfield & Levine (2010) suggest that phylogeny

alone cannot determine community structure, and studies such as Vamosi & Vamosi (2007) have found traits such as body size to be related to community structure.

One of the limitations of this study was limited richness of BINs within some habitats and traits among the families present in Churchill, reflecting primarily the biological patterns as sampling has been extensive (Woodcock et al. 2013; Pyle 2018). There were more terrestrial species than aquatic and more herbivores and predators than scavengers; this makes it hard to accurately compare the observed clustering patterns in relation to the trait states. There were only 16 families studied, providing limited data for the analysis of variance assessing the phylogenetic community structure of the families and traits. Future studies may expand on these results by conducting this analysis for other taxa, other geographic regions, and other traits. While this study did look at one temperate region, including more regions along a latitudinal gradient would allow us to better understand the effects of latitude and environmental conditions on communities. There was a connection between community structure and traits, but only two traits were investigated. By including more traits, we can discover what other traits are being filtered for in Arctic communities and how these traits are affecting phylogenetic community structure. It may also be beneficial to investigate the effect of traits at a lower taxonomic level, as families are diverse, and one trait state may not adequately describe the ecology of every species. There is also the issue that some families are understudied and under sampled (Brunke 2019). While sampling in general is extensive for Churchill beetles (Woodcock et al. 2013; Pyle 2018), families such as Scirtidae and Latridiidae are poorly understood and appear to be more diverse in Canada than the number of recorded species suggests (Brunke 2019). By continuing to study Canada's insect communities, including intensive sampling at focal sites, we can future explore diversity, traits, and community structure based upon more complete sampling.

During post-glacial colonization, species came from the south and from the Beringian glacial refugium (Pielou 1995, Woodcock *et al.* 2013). Was this colonization random? The results of this study suggest that it wasn't. Closely related species, sharing similar traits were found in sub-Arctic communities due to the environmental filtering occurring in this area. Arctic communities are particularly vulnerable to climate change and increasing temperatures (Danks 1992, Walseng 2017). If Arctic conditions change, it is possible that some of their extreme environmental pressures will decrease or shift, and the environmental filtering occurring in these environments will likely also change. By understanding the current community structure and the factors and traits influencing this, we can better predict how these communities are likely to change in the future. If temperate locations show less clustering than those in northern regions, as shown by the comparison of Guelph and Churchill in this study, we can expect communities to become less phylogenetically clustered as species move northward.

## Data Accessibility Statement

DNA Sequences and related data are publicly available on the Barcode of Life Database http://boldsystems.org/. An excel file containing the process ids of sequences used in this analysis can be found in Supplementary material.

The character matrices and phylogenetic tree used in the analysis are available on Data Dryad.

485 References
486

487 Boyle, E.E. & Adamowicz, S.J. 2015. Community phylogenetics: assessing tree reconstruction
488 methods and the utility of DNA barcodes. – PLoS ONE. 10: p.e0126662. doi:
489 10.1371/journal.pone.0126662
490
491 Brunke, A.J., Bouchard, P., Douglas, H.B. & Pentinsaari, M. 2019. Coleoptera of Canada. –
492 ZooKeys. 819: 361-376. doi: 10.3897/zookeys.819.24724
493
494 Danks, H.V. 1992. Arctic insects as indicators of environmental change. – Arctic. 45: 159-166.
495 doi: http://dx.doi.org/10.14430/arctic1389
496 Danks, H.V. 1993. Patterns of diversity in the Canadian insect fauna. – Mem Entomol Soc. 165:
497 51-74. doi: https://doi.org/10.4039/entm125165051-1
498
499 Eagalle, T. & Smith, M.A. 2017. Diversity of parasitoid and parasitic wasps across a latitudinal
500 gradient: using public DNA records to work within a taxonomic impediment. – Facets. 2:
501 937-954. doi: 10.7717/peerj.4642
502
503 Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
504 throughput. - Nucleic Acids Res. 32: 1792-1797. doi: 10.1093/nar/gkh340.
505
506 Emerson, B.C., Cicconardi, F., Franciulli, P.P. & Shaw, P.J.A. 2011. Phylogeny, phytogeography,
507 phylobetadiversity and the molecular analysis of biological communities. – Philos T Roy
508 Soc B. 366: 2391-402. doi: 10.1098/rstb.2011.0057
509
510 Ernst, C.M. & Buddle, C.M. 2015. Drivers and patterns of ground-dwelling beetle biodiversity
511 across Northern Canada. – PloS ONE. 10: p.e0122163. doi:
512 10.1371/journal.pone.0122163
513
514 Google. 2018. Google Earth. https://earth.google.com/web/
515
516 Hamedi, N., Moharramipour, S. & Barzeger, M. 2013. Temperature-dependent chemical
517 components accumulation in *Hippodamia variegata* (Coleoptera: Coccinellidae) during
518 overwintering. – Environ Entomol. 42: 375-380. doi:10.1603/EN11084
519
520 Hebert, P.D.N. & Gregory, T.R. 2005. The promise of DNA barcoding for taxonomy. –
521 Systematic Biol. 54: 852-859. doi: 10.1080/10635150500354886
522
523 Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. 2003. Biological identifications
524 through DNA barcodes. – P Roy Soc B-Biol Sci. 270: 313– 21. doi:
525 10.1098/rspb.2002.2218
526
527
528

529 Heino, J., Soininen, J., Alahuhta, J., Lappalainen, J. &Virtanen, R. 2017. Metacommunity
530 ecology meets biogeography: effects of geographical region, spatial dynamics and
531 environmental filtering on community structure in aquatic organisms. – Oecologia. 183:
532 121-137. doi: 10.1007/s00442-016-3750-y
533
534 Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O.S., Wild, R., Hammond, P.M.,
535 Ahrens, D., Balke, M., Caterino, M.S., Gomez-Zuirta, J., Ribera, I., Barraclough, T.G.,
536 Bocakova, M., Bocak, L. & Alfred, P. 2007. A comprehensive phylogeny of beetles
537 reveals the evolutionary origins of a superradiation. – Science. 318: 1913-6. doi:
538 10.1126/science.1146954
539
540 Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D.,
541 Blomberg, S.P. and Webb. C.O. 2010. Picante: R tools for integrating phylogenies and
542 ecology. – Bioinformatics. 26:1463-1464. Doi: 10.1093/bioinformatics/btq166
543
544 Knapp, M. & Saska, P. 2011. The effects of habitat, density, gender and duration on
545 overwintering success in *Bembidion lampros* (Coleoptera: Carabidae). – J Appl Entomol.
546 136: 225-233. doi: 10.1111/j.1439-0418.2011.01643.x
547
548 Kraft, N.J.B., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. 2007. Trait evolution, community
549 assembly, and the phylogenetic structure of ecological communities. – Am Nat. 170:
550 271–83. doi: 10.1086/519400
551
552 Le Conte, J.L. 1862. Classification of the Coleoptera of North America. – Smithsonian
553 Institution.
554
555 Marshall, S. 2006. Insects: their natural history and diversity: with a photographic guide to
556 insects of eastern North America. – Firefly Books.
557
558 May, J.A. 2017. A new bioinformatics pipeline to reveal the correlates of molecular evolutionary
559 rates in ray-finned fishes. – Msc Thesis, University of Guelph.
560
561 Mayfield, M.M. & Levine, J.M. 2010. Opposing effects of competitive exclusion on the
562 phylogenetic structure of communities. – Ecol Lett. 13: 1085-1093. doi: 10.1111/j.1461-
563 0248.2010.01509.x
564
565 Orton, M.G., May, J.A., Ly, W., Lee, D.J. & Adamowicz, S.J. 2018. Is molecular evolution
566 faster in the tropics? – Heredity. 122: 513-524. doi: 10.1038/s41437-018-0141-7
567
568 Peres-Neto, P.R., Leibold, M.A. & Dray, S. 2012. Assessing the effects of spatial contingency
569 environmental filtering on metacommunity phylogenetics. – Ecology. 93: S14-S30. doi:
570 10.1890/11-0494.1
571
572 Pielou, E.C. 1995. A Naturalist Guide to the Arctic. – University of Chicago Press.
573

574 Poulin, R., Krasnov, B.R., Mouillot, D. & Thieltges, D.W. 2011. The comparative ecology and
575 biogeography of parasites. – Philos T Roy Soc B. 366: 237-2390. doi:
576 10.1098/rstb.2011.0048
577
578 Pyle, M.N. 2018. Insights into post-glacial colonization of sub-Arctic environments: using beetle
579 phylogeny to determine the role of feeding strategy. – Msc Thesis, University of Guelph.
580
581 R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for
582 Statisitcal Computing, Vienna, Austria. URL http://www.R-project.org/.
583
584 Ratnasingham, S. & Hebert, P.D.N. 2007. Bold: the barcode of life data system. – Mol Ecol
585 Notes. 7: 355-364. doi: 10.1111/j.1471-8286.2007.01678.x
586
587 Schliep K.P. 2011. f: phylogenetic analysis in R. – Bioinformatics. 27: 592-593. doi:
588 10.1093/bioinformatics/btq706
589
590 Shibuya, S., Kubota, K., Ohsawa, M. & Kikvidze, Z. 2011. Assembly rules for ground beetle
591 communities: what determines community structure, environmental factors or
592 competition? – Eur J Entomol. 108: 453-459. doi:
593 http://dx.doi.org/10.14411/eje.2011.058
594
595 Slipinski, S. A., Leschen, R.A.B., Beutel, R.G. & Lawrence, J.F. 2011. Coleoptera beetles.
596 Volume 2, Morphology and systematics (Elateroidae, Bostrichiformia, Cucujiformia
597 partim) – Walter de Gruyter
598
599 Smith, M.A., Hallwachs, W. & Janzen, D.H. 2014. Diversity and phylogenetic community
600 structure of ants along a Costa Rican elevational gradient. – Ecography. 37: 720-731. doi:
601 10.1111/j.1600-0587.2013.00631.x

602 Suplementary Material (Appendix EXXXXX at www.oikosoffice.lu.se/appendix). Appendix 1-4.

603
604 Ulrich, W. & Fattorini, S. 2013. Longitudinal gradients in the phylogenetic community structure
605 of European Tenebrionidae (Coleoptera) do not coincide with the major routes of
606 postglacial colonization. – Ecography. 36: 1106-1116. doi: 10.1111/j.1600-
607 0587.2013.00188.x
608
609 Vamosi, J.C. & Vamosi, S.M. 2007. Body size, rarity, and phylogenetic community structure:
610 insights from diving beetle assemblages of Alberta. – Divers Distrib. 13: 1-10. doi:
611 10.1111/j.1472-4642.2006.00299.x
612
613 Walseng, B., Jensen, T., Dimante-Deimantovica, I., Christoffersen, K.S., Chertoprud, M.,
614 Chertoprud, E., Novichkova, A. & Hessen, D.O. 2017. Freshwater diversity in Svalbard:
615 providing baseline data for ecosystems in change. – Polar Biol. 1-11. doi:
616 10.1007/s00300-018-2340-3
617

618 Webb, O.C., 2000. Exploring the phylogenetic structure of ecological communities: an example
619 for rain forest trees. – Am Nat. 156: 145-155. doi: 10.1086/303378
620
621 Webb, O.C., Ackerly, D.D., McPeek, M.A. and Donoghue, M.J. 2002. Phylogenies and
622 community ecology. – Annu Rev Ecol Syst. 33: 475-505. doi:
623 10.1146/annurev.ecolsys.33.010802.150448
624
625 Weiher, E., Freund, D., Bunton, T., Stefanski, A., Lee, T. & Bentivenga, S. 2011. Advances,
626 challenges and a developing synthesis of ecological community assembly theory. – Philos
627 T Roy Soc B. 366: 2403-2413. doi: 10.1098/rstb.2011.0056
628
629 Wilson, J. 2010. Assessing the value of DNA barcodes and other priority gene regions for
630 molecular phylogenetics of Lepidoptera. – PloS ONE. 5: p.e10525. doi:
631 10.1371/journal.pone.0010525
632
633 Wilson, J. 2011. Assessing the value of DNA barcode for molecular phylogenetics: effect of
634 increased taxon sampling in Lepidoptera. – PloS ONE. 6: p.e24769. doi:
635 10.1371/journal.pone.0024769
636
637 Woodcock, T.S., Boyle, E.E., Roughley, R.E., Kevan, P.G, Labbee, R.N., Smith, A.B.T., Goulet,
638 H., Steinke, D. & Adamowicz, S.J. 2013. The diversity and biogeography of the
639 Coleoptera of Churchill: insights from DNA barcoding. – BMC Ecol. 13: 258. doi:
640 10.1186/1472-6785-13-40
641
642 Zhou, X., Adamowicz, S.J., Jacobus, L.M., DeWalt, R.E. & Hebert, P.D.N. 2009. Towards a
643 comprehensive barcode library for Arctic life-Ephemeroptera, Plecoptera, and
644 Trichoptera of Churchill, Manitoba, Canada. – Front Zool. 6: 30. doi: 10.1186/1742-
645 9994-6-30
646
647 Zhou, X. Jacobus, L.M., DeWalt, R.E., Adamowicz, S.J. & Hebert, P.D.N. 2010. Ephemeroptera,
648 Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into
649 biodiversity patterns from DNA barcoding. – J N Am Benthol Soc. 29: 814-837. doi:
650 10.1899/09-121.1

651 Tables & Figures

652 Table 1: A table showing the community phylogenetic and other metrics for each of the Coleoptera families. Significant values are in
653 bold. Significant values under the Holm-Bonferroni threshold are in italics.

| Family | Clustering Value NRI | p-value NRI | Clustering Value NTI | p-Value NTI | Number of BINs in Canada and Alaska | Number of BINs in Churchill | % of Total Found in Churchill | Number of Sequences in Canada and Alaska | Number of Sequences in Churchill | Habitat | Feeding Mode |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Buprestidae | **2.02** | **0.04** | **2.2** | **0.05** | 84 | 3 | 4% | 470 | 3 | Terrestrial | Scavenger |
| Cantharidae | **2.78** | ***0.0009*** | **2.52** | **0.005** | 95 | 6 | 6% | 5043 | 23 | Terrestrial | Predator |
| Carabidae | **1.94** | **0.02** | **2.29** | **0.01** | 398 | 20 | 5% | 3642 | 90 | Terrestrial | Predator |
| Chrysomelidae | -0.72 | 0.77 | -0.47 | 0.67 | 259 | 5 | 2% | 3805 | 71 | Terrestrial | Herbivore |
| Coccinellidae | -0.19 | 0.58 | -0.62 | 0.74 | 104 | 5 | 5% | 2481 | 9 | Terrestrial | Predator |
| Cryptophgidae | 1.38 | 0.07 | **1.88** | **0.04** | 65 | 3 | 5% | 428 | 5 | Terrestrial | Herbivore |
| Curculionidae | 0.21 | 0.44 | 0.19 | 0.39 | 356 | 8 | 2% | 7453 | 11 | Terrestrial | Herbivore |
| Dytiscidae | 1.34 | 0.08 | 1.52 | 0.07 | 84 | 36 | 43% | 1531 | 140 | Aquatic | Predator |
| Elateridae | 0.39 | 0.34 | 0.38 | 0.43 | 246 | 5 | 2% | 3035 | 20 | Terrestrial | Herbivore |
| Gyrinidae | **2** | **0.045** | 1.17 | 0.13 | 18 | 7 | 39% | 215 | 22 | Aquatic | Predator |
| Haliplidae | **2.17** | **0.03** | 0.99 | 0.17 | 9 | 6 | 67% | 75 | 6 | Aquatic | Herbivore |
| Hydrophilidae | 1.14 | 0.12 | **2.7** | **0.01** | 56 | 6 | 11% | 265 | 13 | Aquatic | Scavenger |
| Latridiidae | -0.74 | 0.76 | -0.89 | 0.83 | 80 | 3 | 4% | 4216 | 11 | Terrestrial | Scavenger |
| Leiodidae | 0.68 | 0.24 | 0.25 | 0.38 | 124 | 5 | 4% | 593 | 19 | Terrestrial | Scavenger |
| Scirtidae | 1.03 | 0.17 | 0.67 | 0.2 | 43 | 3 | 7% | 2881 | 9 | Aquatic | Herbivore |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stapylinidae | 0.61 | 0.29 | **2.48** | **0.009** | 951 | 21 | 2% | 7187 | 35 | Terrestrial | Predator |

654

Table 2: The results from the a) ANOVA, i) for the NRI values and ii) for the NTI values. There is no significant relationship between the phylogenetic community structure within families and the habitat or feeding mode of the families. These results differed from the b) PGLS analysis comparing community structure within families to feeding mode and comparing community structure to habitat for both i) NRI values and ii) NTI values, taking into account the family-level phylogeny of beetles. There was no significant relationship between community structure and habitat but there was a significant relationship with feeding mode. Predators were significantly more clustered.

a) i) ANOVA: NRI Values

| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
|---|---|---|---|---|---|
| Habitat | 1 | 2.07 | 2.07 | 2.023 | 0.18 |
| Residuals | 14 | 14.32 | 1.02 | | |

| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
|---|---|---|---|---|---|
| Adult Diet | 2 | 1.62 | 0.81 | 0.71 | 0.51 |
| Residuals | 13 | 14.77 | 1.14 | | |

ii) ANOVA: NTI Values

| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
|---|---|---|---|---|---|
| Habitat | 1 | 0.89 | 0.89 | 0.59 | 0.46 |
| Residuals | 14 | 21.15 | 1.51 | | |

| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
|---|---|---|---|---|---|
| Adult Diet | 2 | 3.02 | 1.51 | 1.03 | 0.384 |
| Residuals | 13 | 19.02 | 1.46 | | |

677 b) i) PGLS: NRI Values

| | Value | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Herbivore | 0.42 | 0.68 | 0.62 | 0.54 |
| Predator | 1.14 | 0.54 | 2.12 | **0.05** |
| Scavenger | 0.72 | 0.66 | 1.09 | 0.3 |

678

| | Value | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Aquatic | 1.31 | 0.79 | 1.66 | 0.12 |
| Terrestrial | -0.39 | 0.74 | -0.53 | 0.61 |

679

680 ii) PGLS: NTI Values

681

| | Value | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Herbivore | 0.14 | 0.72 | 0.2 | 0.84 |
| Predator | 0.47 | 0.57 | 2.57 | **0.02** |
| Scavenger | 1 | 0.7 | 1.42 | 0.18 |

682

| | Value | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Aquatic | 1.23 | 0.89 | 1.38 | 0.19 |
| Terrestrial | -0.39 | 0.83 | -1.47 | 0.65 |

683

684

685

686

687

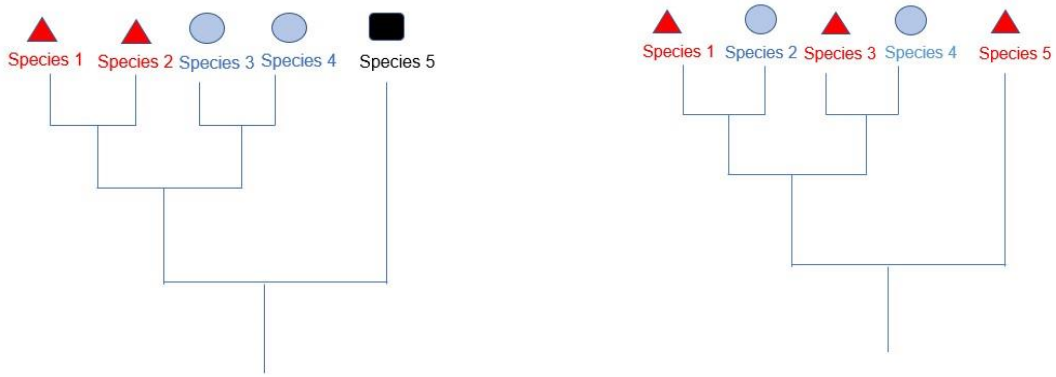
688

689

690

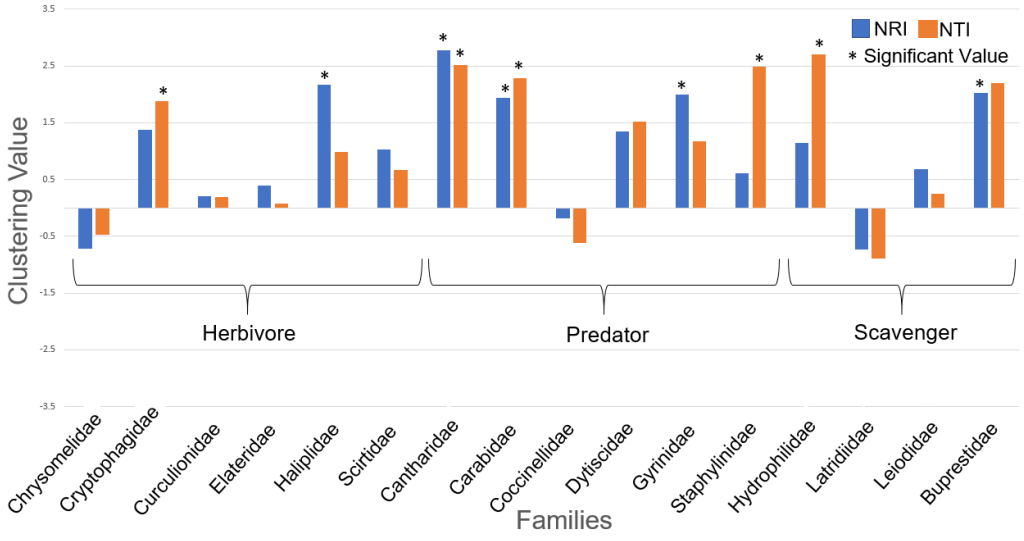
691

692

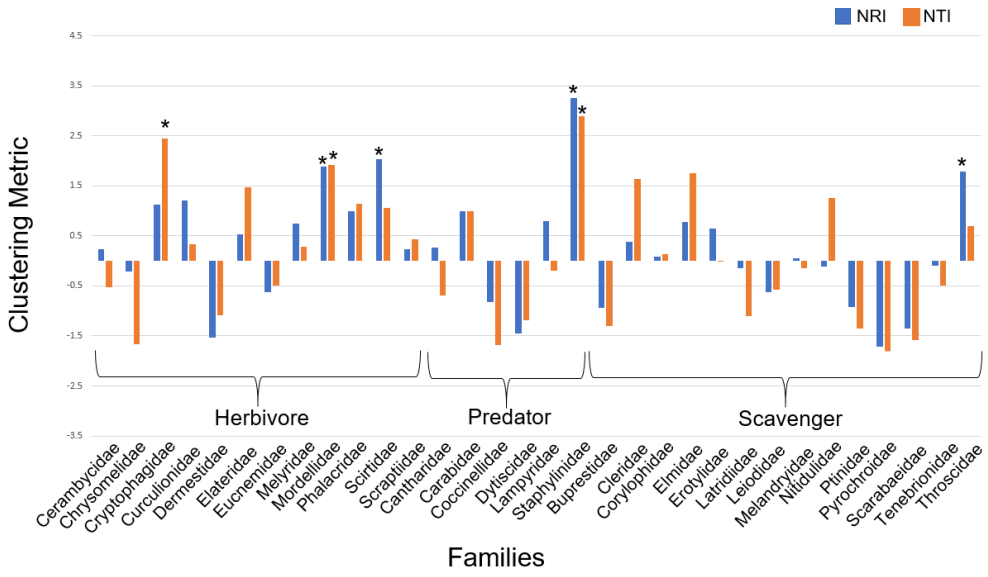



a) b)

Figure 1: Phylogenetic trees demonstrating phylogenetic community structure patterns. Each habitat or geographic region is shown by a different colour and shape. a) Pattern A shows a clustering pattern, where closely related species share the same region. b) Pattern B shows an overdispersed pattern, where closely related species inhabit different regions or environments.




705 a)




706 b)

707 Figure 2: a) Graph showing the phylogenetic community metrics for *Coleoptera* families in
708 Churchill, MB. A positive value indicates a clustered pattern, and a negative value marks an
709 overdispersed pattern. Families exhibiting significant (p-value < 0.05) clustering are marked by
710 an asterisk. The majority of families tend towards a clustering pattern. b) Graph showing the
711 clustering values for Coleoptera families in Guelph, ON. The phylogenetic community structure
712 is generally random, without a clear trend toward overdisperion or clustering. Families are more
713 overdispersed in this region than Churchill.

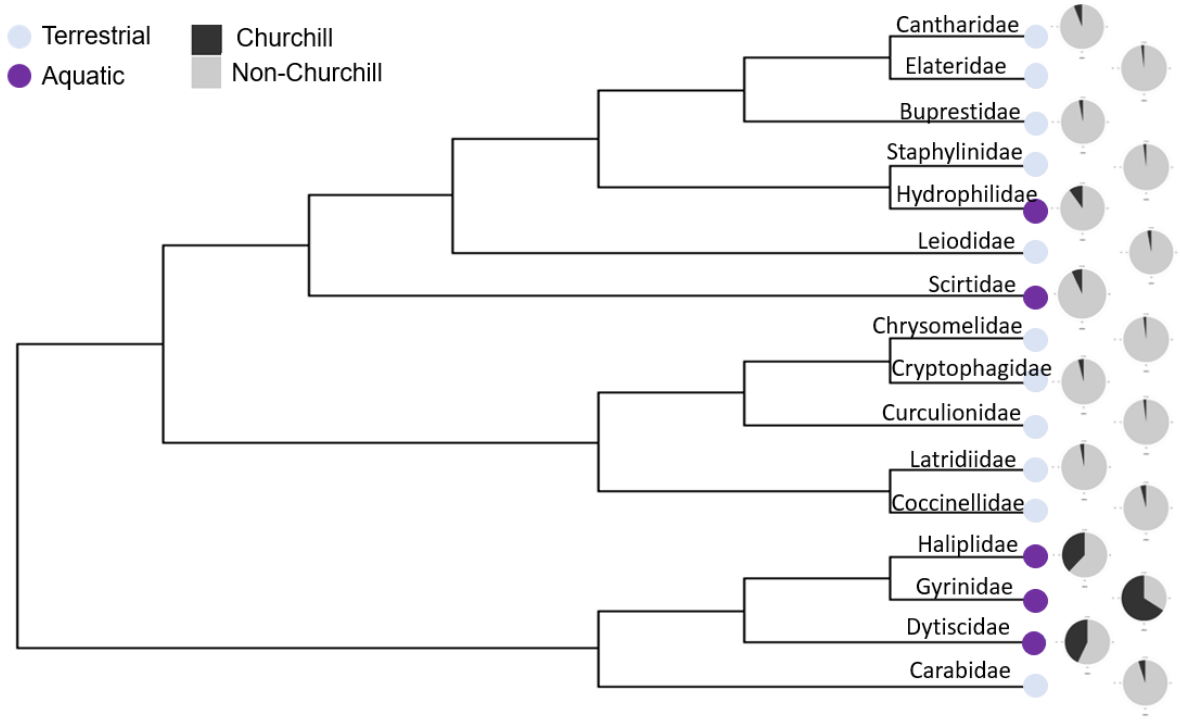



Figure 3: Phylogenetic tree showing the terrestrial and aquatic families present in Churchill. The pie graphs show the percent of the total BINs from Canada and Alaska that have been found in Churchill. Aquatic families have a larger percent of their total BINs found in Churchill than terrestrial families.

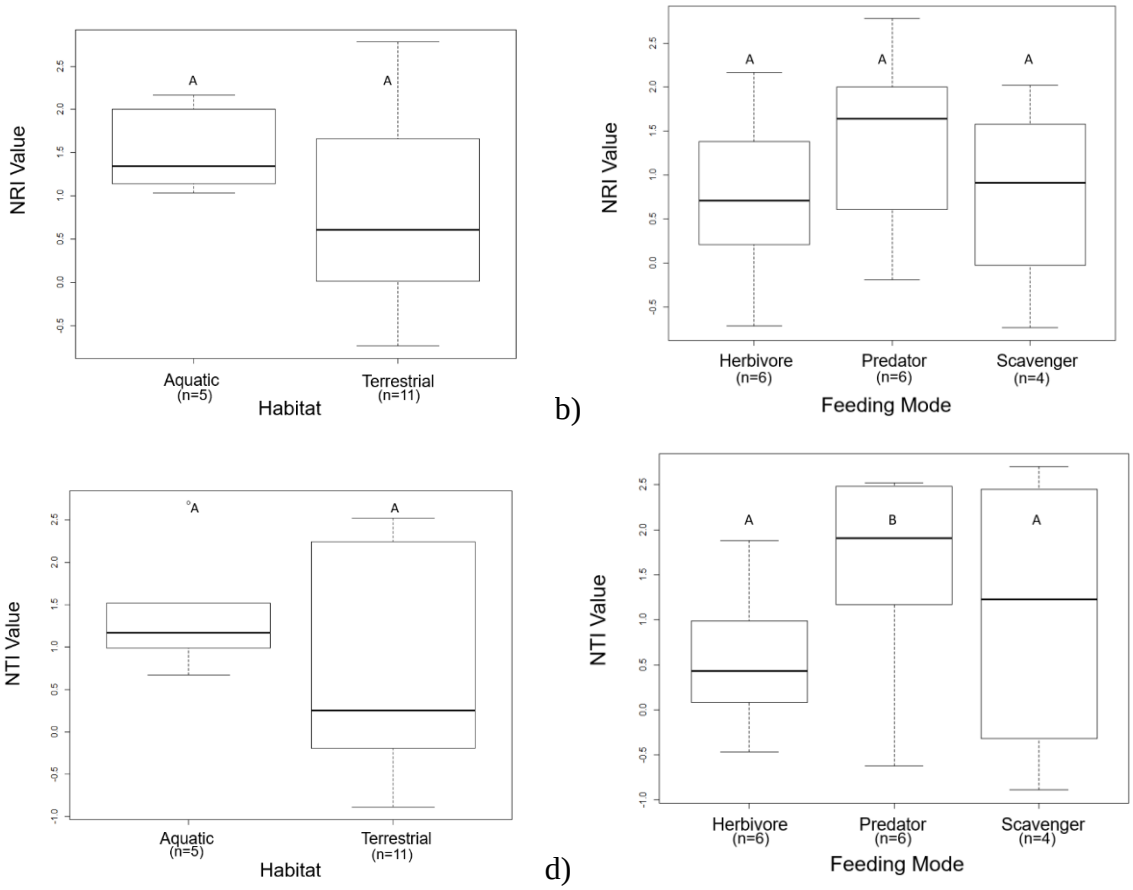



Figure 4: Boxplots showing the results of the PGLS for the clustering values of Coleoptera families inhabiting species habitats using a) NRI and b) NTI and the clustering values of Coleoptera families exhibiting different feeding modes using c) NRI and d) NTI. The same letter above bars denotes groups that do not differ significantly, while different letters denote a significant difference; predators are significantly clustered using the NTI.

Supplementary Material

Appendix 1: Phylogenetic Community Structure Analysis R  Pipeline

```
#pipeline for phylogenetic community structure analysis
#Part 1: Inputting and Filtering Data ----

#Packages needed for this analysis
#If you do not already have these packages, uncomment the code and install.
#install.packages("readr")
library(readr)
#install.packages("plyr")
library(plyr)
#install.packages("dplyr")
library(dplyr)
#install.packages("foreach")
library(foreach)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("stringr")
library(stringr)
#install.packages("stringi")
library(stringi)
#install.packages("ape")
library(ape)
#source("https://bioconductor.org/biocLite.R")
#biocLite("Biostrings")
library(Biostrings)
#source("https://bioconductor.org/biocLite.R")
```

```
#biocLite("muscle")
library(muscle)
#install.packages("phangorn")
library(phangorn)
#install.packages("picante")
library(picante)
#install.packages("data.table")
library(data.table)
#install.packages("phytools")
library(phytools)

#Upload order data into R
#Uncomment the following code to download data directly from BOLD, specifying the required
geographical locations
#dfOrder <-
read_tsv("http://www.boldsystems.org/index.php/API_Public/combined?taxon=Coleoptera&geo
=Alaska|Canada&format=tsv")
#Write file to hard disk
#write_tsv(dfOrder, "Coleoptera_download_June19")
#Read in saved order data
dfOrder <- read_tsv("Coleoptera_download_June19")

#Filtering the data
dfOrder <- dfOrder %>%
  #Filter out those without bin_uri
  filter(str_detect(bin_uri, ":")) %>%
  #Filter out those without a sequence
  filter(str_detect(nucleotides, "[ACTG]")) %>%
  #Filter for COI-5P
```

```
  filter(markercode == "COI-5P") %>%
  #Filter out sequences with fewer than 500 base pairs
  filter(nchar(gsub("-", "", nucleotides)) > 499) %>%
  #Filter out records without a family name
  filter(!is.na(family_name))

#Filter out high gap/N content. A threshold of 1% was chosen because species often differ by
more than 2% divergence. By filtering out records with > 1% N and gap content, we are likely to
get a high-quality data set, given typical patterns of variability in COI in animals.
startNGap <- sapply(regmatches(dfOrder$nucleotides, gregexpr("^[-N]", dfOrder$nucleotides)),
length)
startNGap <- foreach(i=1:nrow(dfOrder)) %do%
  if (startNGap[[i]]>0) {
    split <- strsplit(dfOrder$nucleotides[i], "^[-N]+")
    dfOrder$nucleotides[i] <- split[[1]][2]
  }
endNGap <- sapply(regmatches(dfOrder$nucleotides, gregexpr("[-N]$", dfOrder$nucleotides)),
length)
endNGap <- foreach(i=1:nrow(dfOrder)) %do%
  if (endNGap[[i]]>0) {
    split <- strsplit(dfOrder$nucleotides[i], "[-N]+$")
    dfOrder$nucleotides[i] <- split[[1]][1]
  }
internalNGap <- sapply(regmatches(dfOrder$nucleotides, gregexpr("[-N]",
dfOrder$nucleotides)), length)
internalNGap <- foreach(i=1:nrow(dfOrder)) %do%
  which((internalNGap[[i]]/nchar(dfOrder$nucleotides[i]) > 0.01))
nGapCheck <- sapply(internalNGap, function(x)length(x))
nGapCheck <- which(nGapCheck>0)
dfOrder <- dfOrder[-nGapCheck, ]
```

```
#Remove redundant "BOLD" section from BIN column
dfOrder$bin_uri <- substr(dfOrder$bin_uri, 6, 13)
#Filter out sequences without coordinates
containLatLon <- grep ("[0-9]", dfOrder$lat)
dfOrder <- dfOrder[containLatLon, ]

#Create subset filter using coordinates
#Filter for Churchill, Manitoba
SubsetFilter_Churchill <- which(dfOrder$lat > 58.6 &
                        dfOrder$lon > -94.2 & dfOrder$lat < 58.7 &
                        dfOrder$lon < -93.8)
#Apply filter
dfOrder_Churchill <- dfOrder[SubsetFilter_Churchill, ]

#Find total number of BINs per family in the regional subset of the order
#First convert to datatable
dfOrder_Churchill <- as.data.table(dfOrder_Churchill)
#create datatable showing number of sequences per family
total_species_number <- dfOrder_Churchill[ , .(.N),by=.(family_name)]
#create datatable showing the number of BINs per family
number_of_unique_species <- dfOrder_Churchill[ ,
.(number_of_species=length(unique(bin_uri))), by=family_name]
#convert to dataframe
number_of_unique_species <- as.data.frame(number_of_unique_species)
#Filter down to families with more than 3 or more species
number_of_unique_species <- filter(number_of_unique_species,
number_of_unique_species$number_of_species > 2)

#Create filter to filter down the order to families with three or more species in the subset
```

```
dfOrder_filter <- which(dfOrder$family_name %in% number_of_unique_species$family_name)
#Apply filter
dfOrder <- dfOrder[dfOrder_filter, ]

#Remove unneeded variables
rm(number_of_unique_species, total_species_number, SubsetFilter_Churchill, containLatLon,
endNGap, internalNGap, nGapCheck, split, startNGap, dfOrder_filter, i)

#Part 2: Choosing a Centroid----

#In this section we find a centroid sequence for each BIN present in the order (Not the subset)
#Create a smaller dataframe with needed info
dfBinList <- (dfOrder[, c("processid", "bin_uri", "nucleotides")])
#Create groupings by BIN, each with different bin_uri
binList <- lapply(unique(dfOrder$bin_uri), function(x) dfOrder[dfOrder$bin_uri==x, ])
#Find the number of processids in each bin
binSize <- sapply(binList, function(x)length(x$processid))
#Create new data frame with bin_uri and bin size
dfOrder_bins <- data.frame(binSize)
dfOrder_bins$bin_uri <- c(unique(dfOrder$bin_uri))
#Merge dfBinList and dfOrder_bins
dfBinList <- merge(dfBinList, dfOrder_bins, by.x="bin_uri", by.y="bin_uri")
#Reorder dfFamily_bins by bin_uri
dfOrder_bins <- dfOrder_bins[order(dfOrder_bins$bin_uri), ]

#Find BINs with more than one member
largeBin <- which(dfBinList$binSize > 1)
#Create dataframe with only BINs with more than one member
if (length(largeBin) > 0) {
```

```
  dfCentroid <- dfBinList[largeBin, ]
}

#Subset dfOrder_bins down to number of BINs in dfOrder
dfOrder_bins <- subset(dfOrder_bins, dfOrder_bins$bin_uri %in% dfCentroid$bin_uri)

#Find number of unique BINs in dfCentroid
binNumberCentroid <- unique(dfCentroid$bin_uri)
binNumberCentroid <- length(binNumberCentroid)

#Create dataframe with BINs with only one sequence
dfNonCentroid <- dfBinList[-largeBin, ]

#Create list from dfCentroid
largeBinList <- lapply(unique(dfCentroid$bin_uri), function(x) dfCentroid[dfCentroid$bin_uri
== x, ])
#Extract process Id from each bin
largeBinProcessid <- sapply(largeBinList, function(x) (x$processid))

#Convert sequences to dnaStringSet
dnaStringSet1 <- sapply(largeBinList, function(x) DNAStringSet(x$nucleotides))
#Name dnaStringSet with processids
for(i in seq(from=1, to=binNumberCentroid, by=1)) {
  names(dnaStringSet1[[i]]) <- largeBinProcessid[[i]]
}

#Run multiple sequence alignment for sequences in each BIN in dnaStringSet1
alignment1 <- foreach(i=1:binNumberCentroid) %do%
  muscle::muscle(dnaStringSet1[[i]], maxiters=3, diags=TRUE, gapopen=-3000)
```

```
911
912 #Convert to DNAbin format
913 dnaBINCentroid <- foreach(i=1:binNumberCentroid) %do% as.DNAbin(alignment1[[i]])
914
915 #Calculate a pairwise distance matrix for each BIN
916 geneticDistanceCentroid <- foreach(i=1:binNumberCentroid) %do%
917   dist.dna(dnaBINCentroid[[i]], model="TN93", as.matrix = TRUE,
918           pairwise.deletion = TRUE)
919
920 #Determine centroid sequence; The sequence with the minimum average distance to all other
921 sequences in the BIN.
922 centroidSeq <- foreach(i=1:binNumberCentroid) %do%
923 which.min(rowSums(geneticDistanceCentroid[[i]]))
924 centroidSeq <- centroidSeq %>%
925   unlist() %>%
926   names()
927
928 #Subset dfCentroid by the processid on the list
929 dfCentroid <- subset(dfCentroid, processid %in% centroidSeq)
930
931 #Merge with dfNonCentroid
932 dfAllSeq <- rbind(dfCentroid, dfNonCentroid)
933 #Merge with the original data set
934 dfAllSeq <- merge(dfAllSeq, dfOrder, by.x="processid", by.y="processid")
935 #Reorganize and clean up
936 dfAllSeq <- (dfAllSeq[, c("bin_uri.x", "binSize", "processid", "family_taxID", "family_name",
937 "species_taxID", "species_name", "nucleotides.x", "lat", "lon", "subfamily_name",
938 "order_name")])
939 colnames(dfAllSeq)[1] <- "bin_uri"
940 colnames(dfAllSeq)[8] <- "nucleotides"
```

```
#Delete any possible duplicate entries
dfAllSeq <- (by(dfAllSeq, dfAllSeq["bin_uri"], head, n=1))
dfAllSeq <- Reduce(rbind, dfAllSeq)
#Add an index column
dfAllSeq$ind <- row.names(dfAllSeq)

#Remove unneeded dataframes and variables
rm(alignment1, binList, dfBinList, dfCentroid, dfOrder_bins, dfNonCentroid, dnaBINCentroid,
dnaStringSet1, geneticDistanceCentroid, largeBinList, largeBinProcessid, binNumberCentroid,
binSize,  centroidSeq, i, largeBin)

#Part 3: Alignment----

#Create a function to trim the sequences
RefSeqTrim <- function(x) {
  #Create data frame for reference sequence
  #This reference sequence was taken from BOLD for Coleoptera. Process id: AEDNA549-12.
Species: Colymbetes dolabratus.
  dfRefSeq <- data.frame(taxa=c("Coleoptera"),
nucleotides=c("TAACTTTATATTTTATTTTTGGTGCATGGGCTGGAATGGTAGGAACAT
CTTTAAGTATGTTGATTCGAGCCGAATTAGGAAATCCTGGTTCTCTGATTGGAGATG
ATCAAATTTATAATGTTATTGTAACAGCACATGCTTTTGTAATAATTTTTTTCATAGT
AATACCTATTATAATTGGGGGATTTGGAAATTGATTAGTTCCATTAATATTGGGGGC
CCCAGATATAGCTTTTCCCCGAATAAATAATATAAGTTTTTGACTTCTTCCGCCTTCT
TTAACTCTTCTATTAATAAGAAGAATAGTTGAAAGTGGGGCCGGGACAGGATGAAC
AGTTTACCCCCCTCTATCTTCAGGAATTGCACACGGAGGAGCTTCAGTTGATCTAGC
AATTTTTAGTCTTCATTTAGCTGGAATTTCATCTATTTTAGGGGCTGTAAATTTCATT
ACAACTATTATTAATATACGATCAGTGGGAATAACATTCGACCGAATGCCTCTATTT
GTATGATCCGTAGGAATTACAGCTTTATTACTATTATTATCTTTACCTGTATTAGCGG
GAGCTATTACTATATTATTAACTGATCGTAATCTAAACACCTCATTCTTCGACCCGGC
AGGAGGGGGAGATCCAATTTTATATCAACATTTATT"))
  colnames(dfRefSeq)[2] <- "nucleotides"
  #Convert to datatable
```

```
dfRefSeq <- setDT(dfRefSeq)
dfRefSeq[, "nucleotides":=as.character(nucleotides)]
#Trim sequences to 620bp
dfRefSeq[, nucleotides:=substr(nucleotides, 20, nchar(nucleotides)-19)]
#Check sequence length
dfRefSeq[, seqLength:=nchar(nucleotides)]
#Ensure sequences are of character type
alignmentSeqs <- as.character(x$nucleotides)
#Name according to bin_uri
names(alignmentSeqs) <- x$bin_uri
alignmentref <- as.character(dfRefSeq$nucleotides[1])
#Name reference sequence
names(alignmentref) <- "Reference"
#Put sequences together
alignmentSeqsPlusRef <- append(alignmentref, alignmentSeqs)
#Convert to DNAStringSet
DNAStringSet2 <- DNAStringSet(alignmentSeqsPlusRef)
#Run alignment
alignment2 <- muscle::muscle(DNAStringSet2, diags=TRUE, gapopen=-3000)
#Check alignment
classFileNames <- foreach(i=1:nrow(dfRefSeq)) %do%
  paste("alignmentUntrimmed", dfRefSeq$taxa[i], ".fas", sep="")
alignmentUntrimmed <- DNAStringSet(alignment2)
writeXStringSet(alignmentUntrimmed, file=classFileNames[[1]],
                format = "fasta", width=1500)
#Find stop and start positions in reference
refSeqPos <- which(alignment2@unmasked@ranges@NAMES=="Reference")
refSeqPos <- alignment2@unmasked[refSeqPos]
```

```
  refSeqPosStart <- regexpr("[ACTG]", refSeqPos)
  refSeqPosStart <- as.numeric(refSeqPosStart)
  refSeqPosEnd <- nchar(dfRefSeq$nucleotides[1]) + refSeqPosStart
  refSeqPosEnd <- as.numeric(refSeqPosEnd)
  #Trim sequence
  alignment2Trimmed <- substr(alignment2, refSeqPosStart, refSeqPosEnd)
  #Convert to DNAStringSet
  DNAStringSet3 <- DNAStringSet(alignment2Trimmed)
  #Check alignment
  classFileNames <- foreach(i=1:nrow(dfRefSeq)) %do%
    paste("alignmentTrimmed", dfRefSeq$taxa[i], ".fas", sep="")
  writeXStringSet(DNAStringSet3, file=classFileNames[[1]],
                  format = "fasta", width=1500)
  #Remove reference sequence
  refSeqRm <- which(DNAStringSet3@ranges@NAMES=="Reference")
  dnaStringSet3 <- subset(DNAStringSet3[-refSeqRm])
  alignmentOrder <- DNAStringSet3@ranges@NAMES
  #Reorder based on alignment
  x <- x[match(alignmentOrder, x$bin_uri), ]
  #Replace old sequences with new ones
  trimmedSeqs <- as.character(DNAStringSet3)
  x$nucleotides <- trimmedSeqs
  #Return datafrmae with new sequences
  return(x)
}

#Trim centroid sequences to reference sequence
dfAllSeq2 <- RefSeqTrim(dfAllSeq)
```

```
#Convert sequences to DNAbin format
DNABin <- DNAStringSet(dfAllSeq2$nucleotides)
names(DNABin) <- dfAllSeq2$bin_uri
DNABin <- as.DNAbin(DNABin)
#Construct a distance matrix
distanceMatrix <- dist.dna(DNABin, model="TN93", as.matrix = TRUE, pairwise.deletion =
TRUE)
#Visualize the values in the distance matrix using a histogram
hist(distanceMatrix)

#Using upper threshold of IQR to detect outliers
lowerQuantile <- quantile(distanceMatrix)[2]
upperQuantile <- quantile(distanceMatrix)[4]
iqr <-upperQuantile - lowerQuantile
#Set threshold to 1.5. In order to only remove extreme outliers this can be change to 3.
upperThreshold <- (iqr*1.5) + upperQuantile
#Remove 0 values
distanceMatrix[distanceMatrix==0] <- NA
#Convert to data table
dfOutliers <- as.data.table(distanceMatrix, keep.rownames = T)
#Change the "rn" column to bin_uri
setnames(dfOutliers, "rn", "bin_uri")
#Identify divergent BINs
dfOutliers <- dfOutliers[, outlier := apply(.SD, 1, function(x)all(x>upperThreshold,
na.rm=T))][outlier==TRUE]

#Create remove sequences function
RemoveSequences<-function(x, y){
```

```
  if(length(y)==0){
    print("There are no sequences to remove!")
  }
  else if(length(y)>0){
    x <- x[!x$bin_uri%in%y]
  }
  return(x)
}

#Remove outliers
#Outliers should be blasted prior to removal
dfAllSeq <- RemoveSequences(dfAllSeq, dfOutliers$bin_uri)

#Create final alignment of sequences
#Create RefSeq data frame
#Sequence was taken from BOLD and manually put in
dfRefSeq <- data.frame(taxa=c("Coleoptera"),
nucleotides=c("TAACTTTATATTTTATTTTTGGTGCATGGGCTGGAATGGTAGGAACAT
CTTTAAGTATGTTGATTCGAGCCGAATTAGGAAATCCTGGTTCTCTGATTGGAGATG
ATCAAATTTATAATGTTATTGTAACAGCACATGCTTTTGTAATAATTTTTTTCATAGT
AATACCTATTATAATTGGGGGATTTGGAAATTGATTAGTTCCATTAATATTGGGGGC
CCCAGATATAGCTTTTCCCCGAATAAATAATATAAGTTTTTGACTTCTTCCGCCTTCT
TTAACTCTTCTATTAATAAGAAGAATAGTTGAAAGTGGGGCCGGGACAGGATGAAC
AGTTTACCCCCCTCTATCTTCAGGAATTGCACACGGAGGAGCTTCAGTTGATCTAGC
AATTTTTAGTCTTCATTTAGCTGGAATTTCATCTATTTTAGGGGCTGTAAATTTCATT
ACAACTATTATTAATATACGATCAGTGGGAATAACATTCGACCGAATGCCTCTATTT
GTATGATCCGTAGGAATTACAGCTTTATTACTATTATTATCTTTACCTGTATTAGCGG
GAGCTATTACTATATTATTAACTGATCGTAATCTAAACACCTCATTCTTCGACCCGGC
AGGAGGGGGAGATCCAATTTTATATCAACATTTATT"))

#name nucleotide column and set as character
colnames(dfRefSeq)[2] <- "nucleotides"
```

```
dfRefSeq$nucleotides <- as.character(dfRefSeq$nucleotides)
#Trim references to standard 620
dfRefSeq$nucleotides <- substr(dfRefSeq$nucleotides, 20, nchar(dfRefSeq$nucleotides)-19)
#Check sequence length
dfRefSeq$seqLength <- nchar(dfRefSeq$nucleotides)
#Subset centroid sequences by those found in reference sequence dataframe
dfAllSeq <- subset(dfAllSeq, dfAllSeq$order_name %in% dfRefSeq$taxa)
#Break down dataframe into families
taxalistcomplete <- lapply(unique(dfAllSeq$family_taxID), function(x)
dfAllSeq[dfAllSeq$family_taxID==x, ])

#Extract sequences and bin_uri
familyBin <- foreach(i=1:length(taxalistcomplete)) %do% taxalistcomplete[[i]]$bin_uri
familySequences <- foreach(i=1:length(taxalistcomplete)) %do%
taxalistcomplete[[i]]$nucleotides
familySequenceNames <- familyBin

#Take reference sequences
alignmentref <- as.character(dfRefSeq$nucleotides)
dfRefSeq$reference <- "reference"
#Name reference as a reference
alignmentRefNames <- dfRefSeq$reference
#Merge reference with other sequences
alignmentSequencesPlusRef <- foreach(i=1:length(taxalistcomplete)) %do%
  append(familySequences[[i]], alignmentref[[1]])

#Merge names together
alignmentNames <- foreach(i=1:length(taxalistcomplete)) %do%
  append(familySequenceNames[[i]], alignmentRefNames[[1]])
```

```
#Convert sequences to DNAStringSet format
dnaStringSet3 <- foreach(i=1:length(alignmentSequencesPlusRef)) %do%
  DNAStringSet(alignmentSequencesPlusRef[[i]])

#Name each sequence
for(i in 1:16){
  names(dnaStringSet3[[i]]) <- alignmentNames[[i]]
}

#Multiple sequence alignment
alignmentFinal <- foreach(i=1:length(dnaStringSet3)) %do%
  muscle(dnaStringSet3[[i]], diags=TRUE, gapopen=-3000)
#Check Alignment
familyFileNames2 <- foreach(i=1:length(alignmentFinal)) %do%
  paste("alignmentFinal", dfRefSeq$taxa[i], ".fas", sep="")
alignmentFinalFasta <- foreach(i=1:length(alignmentFinal)) %do%
  DNAStringSet(alignmentFinal[[i]])
foreach(i=1:length(alignmentFinal)) %do%
  writeXStringSet(alignmentFinalFasta[[i]], file=familyFileNames2[[i]], format="fasta",
width=1500)

#Convert to dnaStringSet format
dnaStringSet4 <- foreach(i=1:length(alignmentFinal)) %do%
  DNAStringSet(alignmentFinal[[i]])

#Remove unneeded info
rm(alignmentFinal, alignmentNames, alignmentSequencesPlusRef, dnaStringSet3, familyBin,
alignmentref, alignmentRefNames, i, dfAllSeq2, dfOutliers, distanceMatrix, DNABin,
```

```
familySequences, iqr, lowerQuantile, upperQuantile, upperThreshold, dfRefSeq,
familySequenceNames, alignmentFinal, familyFileNames2)

#Part 4: Create Maximum Likelihood tree----

#Create function to convert DNAStringSets to dataframes
dna_string_to_df = function(dna_string_set){
  out_df = as.data.frame(dna_string_set[[1]])
  for(i in 2:length(dna_string_set)){
    new_df = as.data.frame(dna_string_set[[i]])
    out_df = rbind(out_df, new_df)
  }
  return(out_df)
}
#convert stringsets to dataframes
FamilyDNA = dna_string_to_df(dnaStringSet4)

#Add the bin_uri
FamilyDNA$bin_uri <- row.names(FamilyDNA)
#Merge with the information for dfAllSeq
dfFamilyDNA <- merge(FamilyDNA, dfAllSeq, by.x = "bin_uri", by.y = "bin_uri", all.x =
TRUE)
#Rename the column with your aligned sequences
colnames(dfFamilyDNA)[2] <- "FinalSequences"

#create function to get reference names
get_reference_names <- function(top_ref_num = 16){
  ref_names <- c("reference")
  prefix <- "reference"
```

```
  for(i in 1:top_ref_num){
    new_str <- paste(prefix, as.character(i), sep='')
    ref_names <- c(ref_names, new_str)
  }
  return(ref_names)
}
#create list of reference names
reference_names = get_reference_names()
#remove reference from dataframe
dfFamilyDNA <- dfFamilyDNA[!dfFamilyDNA$bin_uri %in% reference_names , ]

#Pull names from dataframe
familyList <- lapply(unique(dfFamilyDNA$family_name),
                     function(x) dfFamilyDNA[dfFamilyDNA$family_name == x, ])
#Create new dnaStringSet
dnaStringSet5 <- sapply(familyList, function(x) DNAStringSet(x$FinalSequences))
#Pull BIN names from list
binNames <- sapply(familyList, function(x)(x$bin_uri))
#Name the stringsets
for(i in seq(from = 1, to = length(dnaStringSet5), by = 1)) {
  names(dnaStringSet5[[i]]) <- binNames[[i]]
}

#Save family as a fasta file
#For file names make sure to list each family name
familyFileNames <- list("Carabidae", "Curculionidae", "Dytiscidae", "Coccinellidae",
"Leiodidae", "Chrysomelidae", "Staphylinidae", "Buprestidae", "Hydrophilidae", "Haliplidae",
"Cantharidae", "Gyrinidae", "Elateridae", "Cryptophagidae", "Scirtidae", "Latridiidae")

```

```
#Add alignment and .fas to each family name
familyFileNames <- foreach(i=1:length(familyFileNames)) %do%
  paste("Alignment", familyFileNames[[i]], ".fas", sep="")
#Send to your desired working directory
foreach(i=1:length(dnaStringSet5)) %do% writeXStringSet(dnaStringSet5[[i]],
file=familyFileNames[[i]], format="fasta")

#create a list of alignment files
#Calling the alignments in alphabetical order allows for easier analysis during the NRI/NTI step
list_of_files <- c("AlignmentBuprestidae.fas", "AlignmentCantharidae.fas",
"AlignmentCarabidae.fas",
                   "AlignmentChrysomelidae.fas", "AlignmentCoccinellidae.fas",
"AlignmentCryptophagidae.fas",
                   "AlignmentCurculionidae.fas", "AlignmentDytiscidae.fas",
"AlignmentElateridae.fas",
                   "AlignmentGyrinidae.fas","AlignmentHaliplidae.fas",
"AlignmentHydrophilidae.fas",
                   "AlignmentLatridiidae.fas", "AlignmentLeiodidae.fas", "AlignmentScirtidae.fas",
                   "AlignmentStaphylinidae.fas")

#read the alignments into phyDat format
phylo_dat <- lapply(list_of_files, function(x){
  read.phyDat(x, format="fasta", type="DNA")
})

#create distance matrices
dm <- lapply(phylo_dat, function(x){
  dist.ml(x)
})
```

```
#creating NJ tree
tree <- lapply(dm, function(x){
  NJ(x)
})

#run model tests
model_tests <- lapply(phylo_dat, function(x){
  modelTest(x)
})

#create environments
env <- lapply(model_tests, function(x){
  attr(x, "env")
})

#create function to find best model for each family
get_best_model = function(model_df){
  best_model = model_df['Model'][model_df['BIC'] == min(model_df['BIC']) ]
  return(best_model)
}
#create a vector containing the best models
list_of_models = unlist(lapply(model_tests, function(x){
  get_best_model(x)
}))

#get parameters for each model
model_fit <- lapply(env, function(x){
  eval(get(list_of_models, x),x)
```

```
})

#create vector containing inv_values
inv_values <- lapply(1:length(model_fit), function(i){
  model_fit[[i]]$inv
})

#compute likelihood
ml_out = lapply(1:length(tree), function(i){
  pml(tree[[i]], phylo_dat[[i]], k=4, inv = inv_values[[i]])
})

#Drop the suffix from each of the model names
new_list_of_models = unlist(lapply(list_of_models , function(x){unlist(strsplit(x, "\\+"))[[1]]}))

#compute likelihood and optimize parameters
ml_families = lapply(1:length(ml_out), function(i){
  optim.pml(ml_out[[i]], optNni = TRUE, optGamma = TRUE, optInv = TRUE, model =
new_list_of_models[[i]])
})

#Create separate variable for trees
ML_Trees <- lapply(ml_families, function(x){
  x$tree})

#Remove unneeded variables
rm(env, tree, model_tests, model_fit, dm, binNames, familyFileNames, familyList,
dfFamilyDNA, dnaStringSet4, dnaStringSet5, FamilyDNA, ml_families, ml_out,
alignmentFinalFasta, inv_values, i, list_of_files, list_of_models, new_list_of_models,
reference_names, phylo_dat)
```

```
#Part 5: NTI and NRI----

#Create a filter for BINs found in Churchill
ChurchillFilter <- which(dfAllSeq$bin_uri %in% dfOrder_Churchill$bin_uri)
#Create a filter for the BINs not found in Churchill
NotChurchillFilter <- which(!(dfAllSeq$bin_uri %in% dfOrder_Churchill$bin_uri))
#Apply the filters
dfFilter_Churchill <- dfAllSeq[ChurchillFilter, ]
dfFilter_NotChurchill <- dfAllSeq[NotChurchillFilter, ]
#Change to data table and set to 1 if present in Churchill and 0 if not in Churchill
dfFilter_Churchill <- as.data.table(dfFilter_Churchill)
dfFilter_Churchill <- dfFilter_Churchill[, churchill := 1]
dfFilter_NotChurchill <- as.data.table(dfFilter_NotChurchill)
dfFilter_NotChurchill <- dfFilter_NotChurchill[, churchill := 0]
#Bind the new data frames to taxalistcomplete
dfAllSeq <- rbind(dfFilter_Churchill, dfFilter_NotChurchill)

#Create a presence absence matrix for bin_uri in Churchill
#Create new data frame
dfAllSeq2 <- dfAllSeq [, c("bin_uri", "churchill", "family_name")]
#Split into family dataframes
dfAllSeq2 <- split(dfAllSeq2, list(dfAllSeq$family_name))
#Remove the family name column
dfAllSeq2 <- lapply(dfAllSeq2, function(x){
  x[,-3]
})

#Create family matrices
```

```
Family_matrices <- lapply(dfAllSeq2, function(x){
  melt(x, id.var="churchill")
})
Family_matrices <- lapply(Family_matrices, as.data.frame)
Family_matrices <- lapply(Family_matrices, function(x){
  with(x, table(churchill, value))
})
Family_matrices <- lapply(Family_matrices, unclass)

#Calculate net relatedness index (NRI) and nearest taxon index (NTI) using ML Tree
#Ensure ML tree is in correct format
phy.dist <- lapply(ML_Trees, cophenetic)

#Calculate NRI
NRI_Results = lapply(1:length(phy.dist), function(i){
  ses.mpd(Family_matrices[[i]], phy.dist[[i]], null.model = "taxa.labels", abundance.weighted =
FALSE, runs = 1000)
})

#Calculate NTI
NTI_Results = lapply(1:length(phy.dist), function(i){
  ses.mntd(Family_matrices[[i]], phy.dist[[i]], null.model = "taxa.labels", abundance.weighted =
FALSE, runs = 1000)
})

#Remove unneeded variables
rm(Family_phyDat, phy.dist, dfFilter_Churchill, dfFilter_NotChurchill, ChurchillFilter,
NotChurchillFilter, dfAllSeq2, Family_matrices, ML_Trees)
#Part 6: Trait Analysis: ANOVA----
```

```
#Read in character matrix
Coleoptera_Matrix_NRI <-
read_csv(file="C:/Users/sammi/Documents/Coleoptera_Matrix_NRI.csv")
#Run ANOVAs for both traits
Coleoptera_ANOVA_NRI_Feeding <- aov(structure ~ adult_diet, data =
Coleoptera_Matrix_NRI)
Coleoptera_ANOVA_NRI_Habitat <- aov(structure ~ habitat, data = Coleoptera_Matrix_NRI)
#Get ANOVA summary
summary(Coleoptera_ANOVA_NRI_Habitat)
summary(Coleoptera_ANOVA_NRI_Feeding)

#Repeat for NTI values
#Read in matrix
Coleoptera_Matrix_NTI <-
read.csv(file="C:/Users/sammi/Documents/Coleoptera_Matrix_NTI.csv")
#Run ANOVAs for both traits
Coleoptera_ANOVA_NTI_Habitat <- aov(structure ~ habitat, data = Coleoptera_Matrix_NTI)
Coleoptera_ANOVA_NTI_Feeding <- aov(structure ~ adult_diet, data =
Coleoptera_Matrix_NTI)
#Get ANOVA summary
summary(Coleoptera_ANOVA_NTI_Habitat)
summary(Coleoptera_ANOVA_NTI_Feeding)

#Part 7: Trait Analysis: PGLS----

#Read in matrix
PGLSdata_NRI <- read.csv("Coleoptera_Matrix_NRI.csv")
#Read in tree
PGLStree <- read.nexus("Coleoptera_Tree")
```

```
#Set branch lengths to one
PGLStree$edge.length <- replicate((length(PGLStree$edge[, 1])), 1)
PGLStree <- force.ultrametric(PGLStree, method="extend")
#Set the row names to family names
PGLSdata_NRI <- PGLSdata_NRI %>%
  column_to_rownames(var = 'family_name')
#Make sure tree and dataframe are in the same order
PGLSdata_NRI <- PGLSdata_NRI[match(PGLStree$tip.label, rownames(PGLSdata_NRI)), ]
#Run PGLS analysis
pglsModel_NRI1 <- gls(structure ~ habitat, correlation = corBrownian(phy = PGLStree), data =
PGLSdata_NRI, method = "ML")
pglsModel_NRI2 <- gls(structure ~ adult_diet, correlation = corBrownian(phy = PGLStree), data
= PGLSdata_NRI, method = "ML")
#Get PGLS summary
summary(pglsModel_NRI1)
summary(pglsModel_NRI2)

#Create boxplots for traits vs. clustering matrix
plot1 <- boxplot(PGLSdata_NRI$structure ~ PGLSdata_NRI$habitat)
plot2 <- boxplot(PGLSdata_NRI$structure ~ PGLSdata_NRI$adult_diet)

#Repeat for NTI
PGLSdata_NTI <- read.csv("Coleoptera_Matrix_NTI.csv")
#Set row names to family names
PGLSdata_NTI <- PGLSdata_NTI %>%
  column_to_rownames(var = 'family_name')
#Make sure tree and dataframe are in the same order
PGLSdata_NTI <- PGLSdata_NTI[match(PGLStree$tip.label, rownames(PGLSdata_NTI)), ]
#Run PGLS analysis
```

```
pglsModel_NTI1 <- gls(structure ~ habitat, correlation = corBrownian(phy = PGLStree), data =
PGLSdata_NTI, method = "ML")

pglsModel_NTI2 <- gls(structure ~ adult_diet, correlation = corBrownian(phy = PGLStree), data
= PGLSdata_NTI, method = "ML")

#Get PGLS summary

summary(pglsModel_NTI1)

summary(pglsModel_NTI2)


#Create boxplots for traits vs. clustering matrix

plot3 <- boxplot(PGLSdata_NTI$structure ~ PGLSdata_NTI$habitat)

plot4 <- boxplot(PGLSdata_NTI$structure ~ PGLSdata_NTI$adult_diet)
```

Appendix 2: Choice of COI Marker Gene

COI is commonly used for DNA barcoding animals and provides useful phylogenetic signal at low taxonomic levels but has some limitations when used to construct deep phylogenies (Boyle & Adamowicz 2015, Smith *et al*. 2014, Wilson 2010, 2011). This limited phylogenetic signal can be helped by using a constraint tree when constructing phylogenies (Boyle & Adamowicz 2015, Wilson 2011). Despite some limitations, COI can be readily sequenced from a large number of taxa and provides high sequence quality compared to other gene regions (Wilson 2010). Barcode-based trees have also shown similar results when used for community phylogenetics compared to other trees (Boyle & Adamowicz 2015, Smith *et al*. 2014). Because of this, we decided COI was suitable to use for this study. Additionally, this marker had the advantage of large-scale taxonomic and geographic coverage for North American beetles.

## Appendix 3: Net Relatedness Index/ Nearest Taxon Index

NRI and NTI use the pairwise distances among species to quantify the community relatedness (Webb 2000). NRI averages the evolutionary distances between all pairs of tips in the community, while NTI takes only the distances between nearest neighbours (Fig. A1) (Webb 2000). When the NRI/NTI value increases, this indicates increased phylogenetic clustering of the species within the community (Webb 2000). The two tests detect patterns at different levels within the phylogeny; therefore, in order to test for general patterns, both tests should be performed (Kraft *et al.* 2007)

## Appendix 4: Sensitivity Analysis: Size of Regional Species Pool and Taxon Richness of Source Pool

The regional species pool was restricted to the Canadian provinces and territories of Manitoba, Nunavut, Northwest Territories, Saskatchewan, and Ontario. This restriction also helps combat some patterns that may be based on biogeography. For example, the Rocky Mountain Range may act as a barrier to dispersal, and this could create a clustering pattern on its own. By restricting the regional pool, we can remove this effect.

After the regional pool was reduced, Dytiscidae, Haliplidae, and Gyrinidae were still the only families where local species richness was close to 30-60% of regional species richness (Table A1). The results for NRI and NTI did not substantially differ from the original analysis (Fig. A2). This was confirmed with a t-test comparing the NRI and NTI values between the original and restricted source phylogenies (NRI: t-statistic = -0.04, p-value =0.96. NTI: t-statistic = -0.17, p-value=0.87). Significance differed from the original analysis for some families. Staphylinidae exhibited significant evidence of clustering in NRI and Dytiscidae exhibited significant evidence of clustering in both. Carabidae lost its significance in NRI. The trends

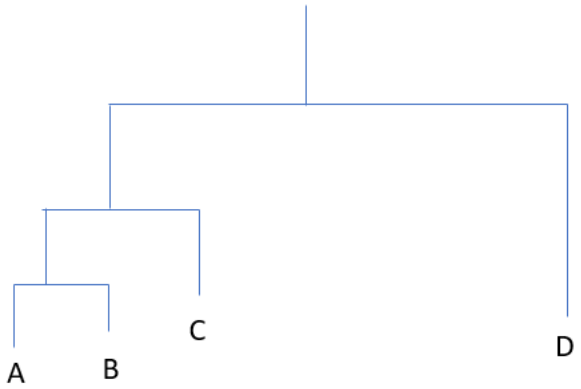
(clustering vs. phylogenetic overdispersion) remained the same for all families except Curculionidae, which became overdispersed, and Latridiidae, which became clustered in NRI only.

Tables

Table A1: Table showing the phylogenetic community structure results for the Coleoptera families after restricting the total BIN source pool. Significant values are bolded.

| Family | Clustering Value NRI | p-value NRI | Clustering Value NTI | p-Value NTI | Number of BINs in Canada and Alaska | Number of BINs in Churchill | % of Total Found in Churchill | Number of Sequences in Regional Species Pool | Number of Sequences in Churchill | Habitat | Feeding Mode |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Buprestidae | **2.33** | **0.02** | **2.46** | **0.02** | 45 | 3 | 7% | 159 | 3 | Terrestrial | Scavenger |
| Cantharidae | **2.18** | **0.001** | **2.1** | **0.007** | 59 | 6 | 10% | 1483 | 23 | Terrestrial | Predator |
| Carabidae | 0.48 | 0.31 | **1.93** | **0.03** | 212 | 20 | 9% | 1702 | 90 | Terrestrial | Predator |
| Chrysomelidae | -0.26 | 0.61 | -0.51 | 0.69 | 199 | 5 | 3% | 2903 | 71 | Terrestrial | Herbivore |
| Coccinellidae | -0.83 | 0.8 | -1.1 | 0.9 | 70 | 4 | 6% | 1113 | 9 | Terrestrial | Predator |
| Cryptophagidae | 1.14 | 0.12 | **1.75** | **0.05** | 34 | 3 | 9% | 203 | 5 | Terrestrial | Herbivore |
| Curculionidae | -0.67 | 0.75 | -0.1 | 0.48 | 204 | 8 | 4% | 3711 | 11 | Terrestrial | Herbivore |
| Dytiscidae | **2.46** | **0.01** | **1.72** | **0.04** | 51 | 36 | 71% | 1368 | 140 | Aquatic | Predator |
| Elateridae | 0.12 | 0.46 | -0.05 | 0.45 | 141 | 5 | 4% | 1092 | 20 | Terrestrial | Herbivore |
| Gyrinidae | **2.89** | **0.01** | 1.34 | 0.09 | 11 | 7 | 64% | 178 | 22 | Aquatic | Predator |
| Haliplidae | **1.91** | **0.04** | 1.19 | 0.15 | 7 | 6 | 86% | 61 | 6 | Aquatic | Herbivore |
| Hydrophilidae | 0.6 | 0.27 | **2.72** | **0.007** | 40 | 6 | 15% | 191 | 13 | Aquatic | Scavenger |
| Lateridiidae | 0.01 | 0.45 | -0.05 | 0.52 | 52 | 3 | 6% | 2252 | 11 | Terrestrial | Scavenger |
| Leiodidae | 1.13 | 0.12 | 1.11 | 0.13 | 68 | 5 | 7% | 293 | 19 | Terrestrial | Scavenger |
| Scirtidae | 0.62 | 0.22 | 0.23 | 0.31 | 32 | 3 | 9% | 1734 | 9 | Aquatic | Herbivore |
| Staphylinidae | **1.65** | **0.04** | **2.9** | **0.006** | 485 | 31 | 6% | 2509 | 35 | Terrestrial | Predator |

Figures




| | A | B | C | D |
|---|---|---|---|---|
| A | | 1 | 2 | 3 |
| B | | | 2 | 3 |
| C | | | | 2 |
| D | | | | |

Figure A1: Example phylogenetic tree with a chart showing nodal distances among members of the community. NRI uses all the distances to find the mean pairwise distance ((1+2+3+2+3+2)/6 =2.16). NTI uses only the distances between nearest neighbors ((1+2+3)/3 =2).

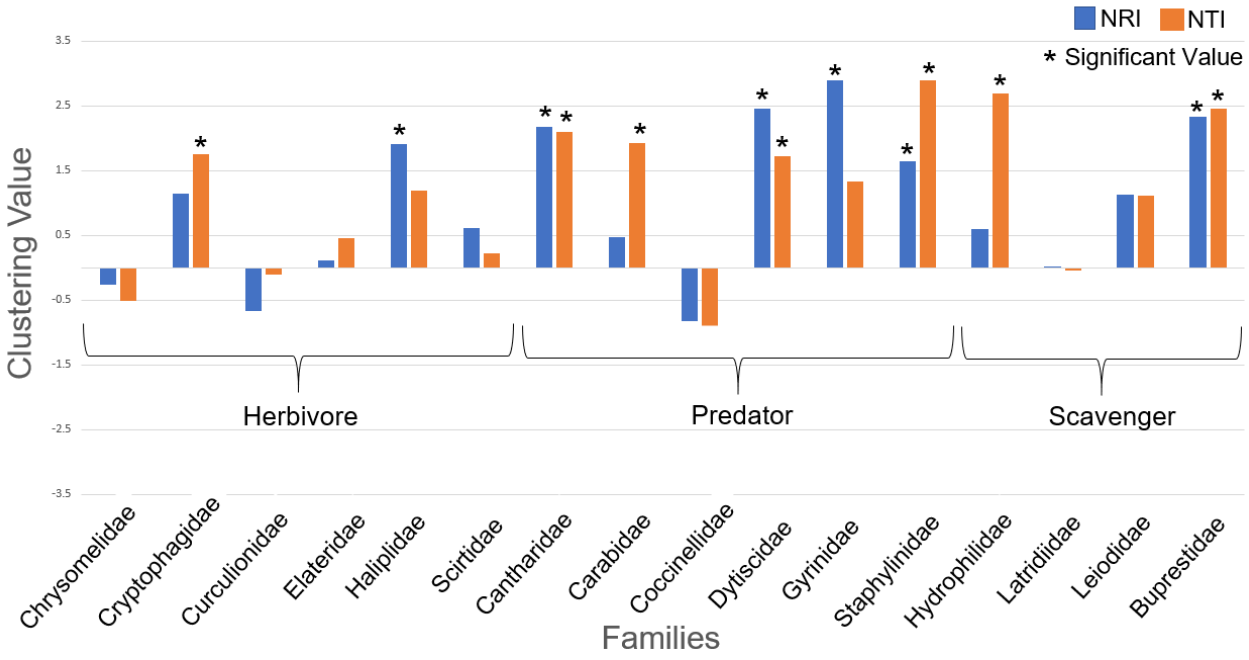



Figure A2: Graph showing the clustering values for Coleoptera families in Churchill, MB, after the sensitivity analysis. The results did not substantially differ from the original analysis, and the main conclusions were supported.

References

Boyle, E.E. & Adamowicz, S.J. 2015. Community phylogenetics: assessing tree reconstruction methods and the utility of DNA barcodes. – PLoS ONE. 10: p.e0126662. doi: 10.1371/journal.pone.0126662

Kraft, N.J.B., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. – Am. Nat. 170: 271–83. doi: 10.1086/519400

Majoros, S.E. and Adamowiz, S.J. Year. Phylogenetic signal of sub-Arctic beetle communities. – Ecography 000:000-000.

Smith, M.A., Hallwachs, W. & Janzen, D.H. 2014. Diversity and phylogenetic community structure of ants along a Costa Rican elevational gradient. – Ecography. 37: 720-731. doi: 10.1111/j.1600-0587.2013.00631.x

Webb, C.O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. – Am Nat. 156: 145-155. doi: 10.1086/303378

Wilson, J. 2010. Assessing the value of DNA barcodes and other priority gene regions for molecular phylogenetics of Lepidoptera. – PloS ONE. 5: p.e10525. doi: 10.1371/journal.pone.0010525

Wilson, J. 2011. Assessing the value of DNA barcode for molecular phylogenetics: effect of increased taxon sampling in Lepidoptera. – PloS ONE. 6: p.e24769. doi: 10.1371/journal.pone.0024769