*Preliminary Summary of Methods and Results for the Phylogenetic Biodiversity Analysis of University Greenhouses*

Caroline Dolt, Hilary Callahan

A. Receiving and Resolving Original Lists

Emails were sent out to various department faculty and staff, usually the greenhouse manager or a biology professor with botanical research. We received many replies and a total of 29 lists from institutions across the nation. Many of these lists contained old names or best guesses, so it was necessary to go through each of them by hand. The first task was to resolve the names of each plant on the individual lists. This was done going through them by hand and using Google, The Plant List, IPNI, GBIF, and ITIS to determine the most updated names for the specimen. Many hybrids, predominately orchids and succulents, were eliminated from the lists since the programs used to analyze the lists deal best with species binomials. The total came to over 16,000 specimens with 7,366 unique species. Since there is a total of over 350,000 plant species total, this is a small slice of the phylogenetic pie. All the lists were cut into their binomials so they could be used in the RStudio package, taxize, which removed some information concerning subspecies or variation.

B. Finding Higher Level Taxonomic Classifications for all Specimens on Resolved Lists

To get the higher level classification of each specimen, lists were put into RStudio and run through the taxize package. Here is a sample of the coding used to read and organize the list in three chunks:

```{r}

wsulist<-read.csv("WSU\_Binomials.csv")

wsulist

wash2<-classification(wsulist$Taxon, db = 'gbif', verbose=FALSE, row=1, return\_id = FALSE)

wash2

washu<-cbind(wash2)

washu

dim(washu)

class(washu)

```

Each binomial list was put through this until there were a total of 29 profiles, going from the species level up to the kingdom classification. The ‘row=1’ function was added to the command because the database chosen, gbif, often had multiple possibilities for each name. This command automatically chooses the first row of options, which is usually correct. Some plants, such as *Eugenia uniflora* and *Ficus carica*, were consistently catalogued as animals, so manual correction and checking of each profile was required. Once the profiles were corrected, new lists with orders, families, and genera were cut to analyze individually. There was both a total and a unique list made for each taxonomic level.

C. Taxonomic analysis using VEGAN with total numbers of orders, families and genera

Using the package, VEGAN, the cut lists made from the individual profiles could be bound and analyzed for similarities and differences.

The orders were all checked against the APG4 list released in 2016, which combined many existing orders. The lists included mostly angiosperms, but mosses and ferns are kept in some greenhouses. This is an example of how the orders lists were imported and renamed:

```{r}

centre<-read.csv("CentreOrders.csv")

Centre30=count(centre)

Centre30

Centre30<-data.frame("Centre", Centre30[,1:2])

names(Centre30)<-c("ListID", "OrderID", "Freq")

Centre30

```

After all the lists were brought in, they were bound to create one long list that had columns classifying the college, listing the order, and listing the frequency. There is a total of 84 orders across all specimens. Then, this bound list was put through ‘matrify’, which created a large frequency matrix.

From the frequency matrix, a jaccard index was made, which assigns a number between 0 and 1 to a pairwise comparison between two collections. A heatmap was from the jaccard matrix to visualize similarities and differences (Fig 1).

The same process was followed for both the family and genus analyses. There were a total of 326 families and 2091 genera. A frequency and jaccard matrix were generated for both, and these are the resulting heatmaps(Fig 2, Fig 3).

The complete visualization of this data was put into a bar graph to show the complete tally of all orders, families, and genera per each collection(Fig 7).

D. Taxonomic analysis using VEGAN for unique numbers of orders, families and genera

The original cut lists were duplicated and then the duplicates within the list were deleted so that only a single occurrence for each taxon remained. These were loaded into RStudio so that the same process as above would be followed but studying only the presence/absence of a taxon would make some analysis easier.

It was found that there are 8 orders that occur in every greenhouse and 8 orders that occurred only in one greenhouse.

|  |  |
| --- | --- |
| Universal Orders | Endemic Orders |
| Alismatales | Berberidopsidales(UConn) |
| Asparagales | Blasiales(UConn) |
| Caryophyllales | Bruniales(UW Seattle) |
| Gentianales | Hymenophyllales(UW Seattle) |
| Lamiales | Isobryales(UW Seattle) |
| Malpigphiales | Notothyladales(UConn) |
| Poales | Pallaviciniales(UConn) |
| Polypodiales | Polytrichales(UWis Madison) |

For the families, there were 63 that occurred in only one greenhouse and 8 that were in all 28 greenhouses.

Universal(n=8): Amaryllidaceae, Apocynaceae, Araceae, Asparagaceae, Bromeliaceae, Euphorbiaceae

Endemic(n=63):

|  |  |
| --- | --- |
| University of Alaska - Fairbanks | Hydrophyllaceae |
| Southern Illinois University | Dipsacaceae |
| Cornell University | Calophyllaceae |
| Rollins College | Platanaceae, Styracaceae |
| University of Washington - Seattle | Bruniaceae, Fontinalaceae, Hymenophyllaceae |
| Smith College | Asteliaceae, Chrysobalanaceae, Kewaceae, Linnaeaceae, Philesiaceae, Stemonaceae |
| University of Wisconsin – Madison | Alzateaceae, Amblystegiaceae, Bryaceae, Hylocomiaceae, Jungermanniaceae, Polytrichaceae, Putranjivaceae |
| University of Connecticut | Berberidopsidaceae, Blasiaceae, Brachytheciaceae, Butomaceae, Byblidaceae, Cabombaceae, Cephaloziaceae, Clethraceae, Coriariaceae, Corynocarpaceae, Corynocarpaceae, Curtisiaceae, Daphniphyllaceae, Datiscaceae, Diervillaceae, Drosophyllaceae, Griseliniaceae, Grubbiaceae, Hernandiaceae, Hypnaceae, Juncaginaceae, Kirkiaceae, Lanariaceae, Lepidoziaceae, Loganiaceae, Martyniaceae, Montiniaceae, Myrothamnaceae, Notothyladaceae, Pallaviciniaceae, Penaeaceae, Philydraceae, Racopilaceae, Resedaceae, Salvadoraceae, Schoepfiaceae, Staphyleaceae, Stilbaceae, Surianaceae, Tamaricaceae, Trichocoleaceae |

Heatmaps were generated at each level of analysis(Fig 4-6). Concerning genera, there are 820 genera that endemic to one greenhouse. No genera occur in all greenhouses, but two appear in 27 of the 29 greenhouses(*Euphorbia* and *Trandescantia*). See Figure 8.

There are no endemic taxa at any level in the collections at the University of Iowa or Wheaton College. The University of Connecticut has the most endemic taxa at every level.

E. Phylogenetic analysis using picante to determine species richness and phylogenetic diversity statistics

A phylogenetic tree with branch lengths was generated for each collection using a list of species binomials placed into phylomatic(<http://phylodiversity.net/phylomatic/>). Using the megatree Zanne2014, newick code was made and copied into Interactive Tree of Life to visualize the trees. Centre College and the University of Connecticut were analyzed only at the genus level. Centre College provided a list missing most species, so to create a usable tree, the genera provided were used. The University of Connecticut provided a comprehensive and long list, which was too much for the program to handle at a species level. All newick codes are missing some species, but the majority are included. This could be from names outdated pre-2014 or missing updated names from the 2016 APG4 list that would not be in the 2014 megatree.

These newick codes were uploaded into RStudio to be analyzed using the ‘picante’ package. The purpose of this package is to analyze the phylogenetic diversity of a plant list and do comparisons between lists. Four statistics were calculated at the order and family level: phylogenetic diversity (PD), species richness (SR), mean nearest taxon distance (MNTD), and mean pairwise distance (MPD).

Phylogenetic diversity is based on the system from Faith 1992, which assigns the number of tree units between taxa. Species richness is the number of unique taxa. Mean nearest taxon distance is the mean distance between each species and its closest relative within a community. The mean pairwise distance is the mean phylogenetic distance between all pairs of species within a community.

i. Results of the Order Phylogenetic Analysis

The PD had the range of 10,035.9, mean of 6,522 and median 6,607.0. The University of Connectucut had the greatest PD at 12,994.0 while Centre College had the lowest at 2,958.1. The SR had a range of 58, mean 39.28 and median 41. The greatest SR was 74 from the University of Connecticut and the lowest was 16 from East Carolina University.

For MNTD and MPD, a positive z value and high p value(p>0.05) indicate greater phylogenetic distance than expected, or overdispersion. A negative z value and a low p value(p<0.05) indicate clustering and low phylogenetic distance.

Universities that had contradictory results, such as a high p value for MNTD and MPD but a negative z value for both statistics, were classified as neither overdispersed or clustered. If three of the four statistics produced results that put the university strongly towards a conclusion, the group was classified as mostly overdispersed or mostly clustered.

For the order analysis, 13 groups were clustered, 7 were mostly clustered, 6 were neither clustered not overdispersed, 1 was mostly overdispersed, and 2 were overdispersed. The University of Connecticut(n=74) and East Carolina University(n=16) were the overdispersed groups. They are on opposite ends of the spectrum for number of taxa, so it is likely that UConn is overdispersed due to almost encompassing all orders while ECU is overdispersed because its orders are phylogenetically diverse and few.

The shortest phylogenetic distance is 211.35 myr between Ericales and Cornales. The longest phylogenetic distances are over 880.58 myr between all species and these outgroups: Blasiales, Lunulariales, and Marchantiales. This makes sense, as all four of these families are liverworts and would be much older than the angiosperms that are the majority of species kept in the greenhouses.

See Figure 9

The rarefaction curve shows how much the diversity of orders increases with each added site(greenhouse). There is a steep increase for the first seven or eight sites, but then the line plateaus. This means that the diversity does not increase by adding more and more greenhouses. Since there is such a finite number of orders, it makes sense that a small number of greenhouses can cover the spectrum of orders fairly well.

See Figure 10

ii. Results of the Family Phylogenetic Analysis

The PD had a range of 23,578.7, a mean of 10,706.8 and a median of 9828.2. The greatest PD was 27,100.2 from the University of Connecticut and the smallest was 3,521.5 from East Carolina. The SR had a range of 236, a mean of 91.2 and a median of 80. University of Connecticut had the highest SR at 256 and East Carolina had the lowest at 20.

For the family analysis, 0 groups were clustered, 4 were mostly clustered, 7 were neither clustered not overdispersed, 13 was mostly overdispersed, and 5 were overdispersed. This is a large change from the results of the order analysis since there are a greater number of possible families and reflects how they are better dispersed across university greenhouses than the orders.

The shortest phylogenetic distance is 112.39 myr between Asteraceae and Goodeniaceae. The longest phylogenetic distances are over 880.58 myr between all species and these outgroups: Blasiaceae, Lunulariaceae, Ricciaceae, and Conocephalaceae. Ricciaceae and Conocephalaceae are families of liverworts belonging to the order Marchantiales.

See Figure 11

The rarefaction curve for families has a slightly higher slope than the one for orders. This means that there is a greater increase in diversity with each added site, although the line does begin to plateau as the number of sites approaches 20.

See Figure 12

References coming soon

Appendix A

Figure 1. Heatmap of collections for total orders

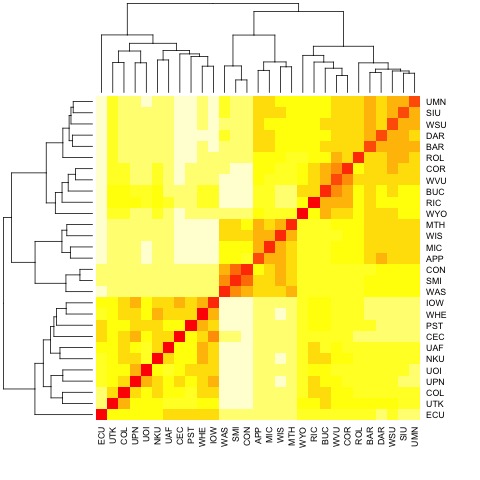


Figure 2. Heatmap of collections for total families

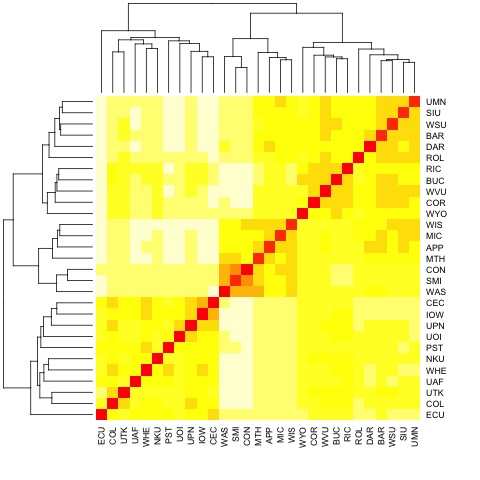


Figure 3. Heatmap of collections for total genera

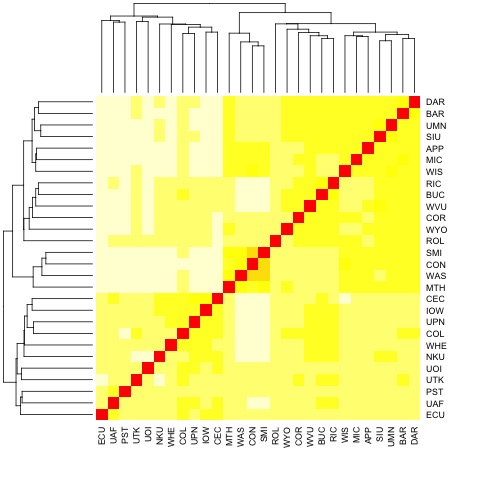


Figure 4. Heatmap of collections for unique orders

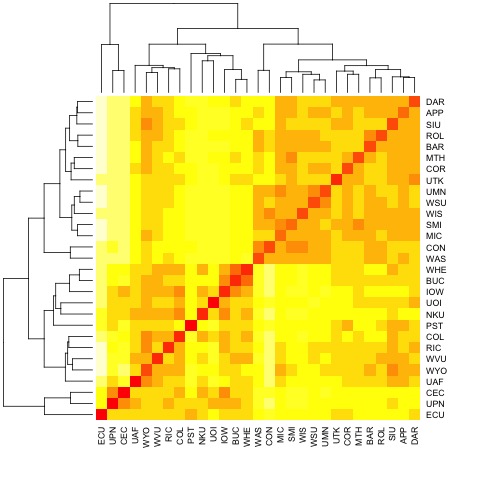


Figure 5. Heatmap of collections for unique families

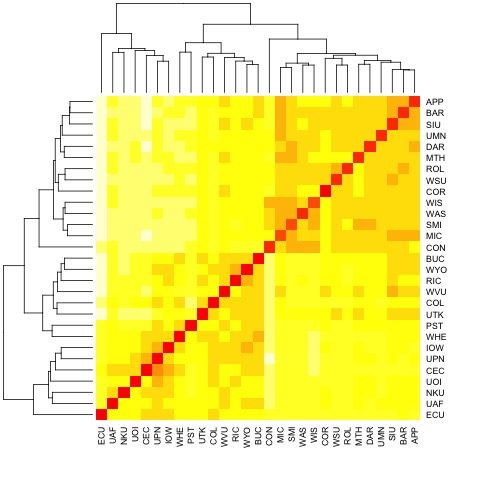


Figure 6. Heatmap of collections for unique genera

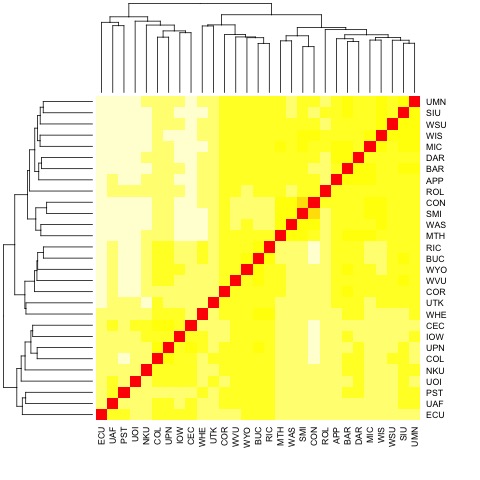


Figure 7. Comparing the Total Number of Orders, Families, and Genera by University

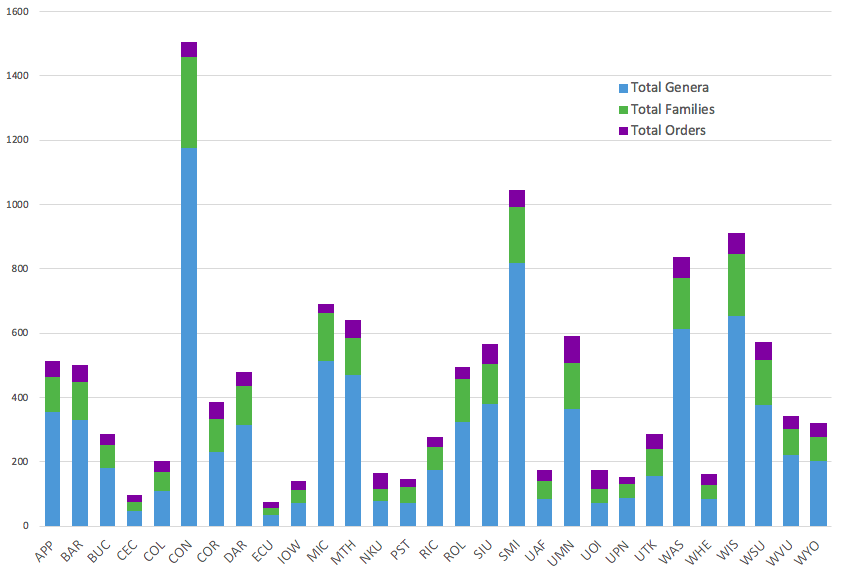


Figure 8. Comparing the Unique Number of Orders, Families, and Genera by University

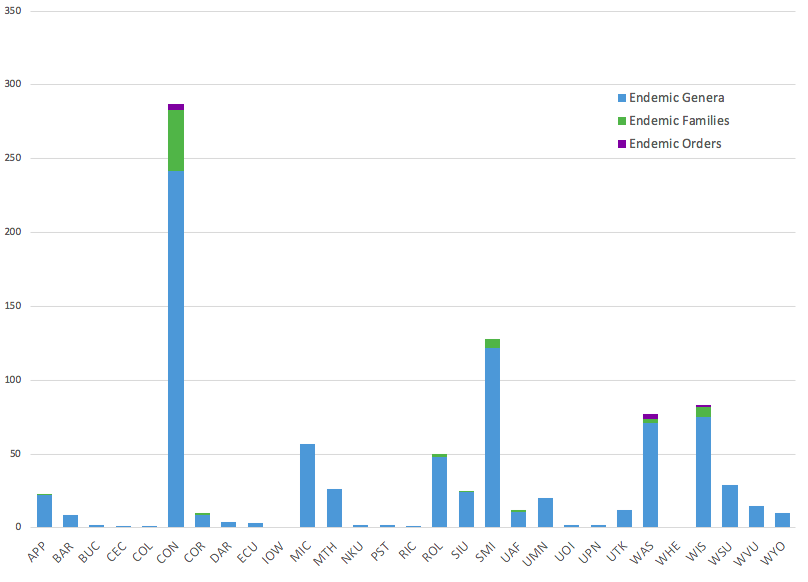


Figure 9. Hierarchical Clustering of Order Diversity

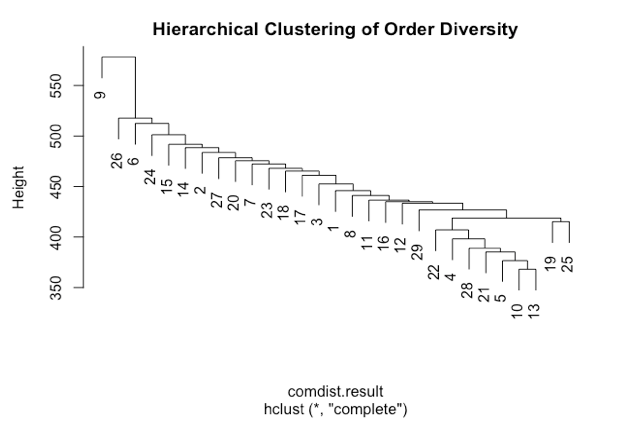


Figure 10. Rarefaction curve for species richness and phylogenetic species richness of orders

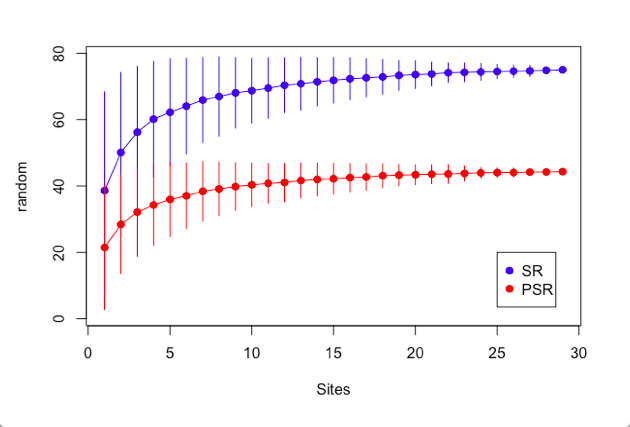


Figure 11. Hierarchical Clustering of Family Diversity

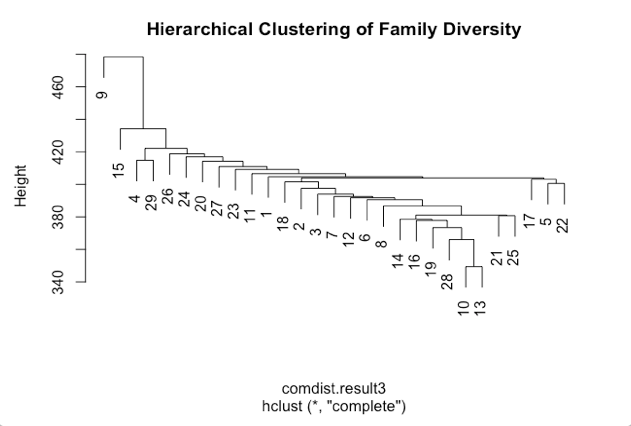


Figure 12. Rarefaction curve for species richness and phylogenetic species richness of families 