

Belly Button Microbiome Sequence Data Analysis


Gabrielle McCabe, Ethan Cohen, Kevin Saroya, Aimee Watts, and Jenna Kim

BIOL 432
April 13th, 2020

Introduction:

Referenced Study

A Jungle in There: Bacteria in Belly Buttons are Highly Diverse, but Predictable

Jiri Hulcr, Andrew M. Latimer, Jessica B. Henley, Nina R. Rountree, Noah Fierer, Andrea Lucky, Margaret D. Lowman, Robert R. Dunn 

Published: November 7, 2012 • <https://doi.org/10.1371/journal.pone.0047712>

- The researchers of this study set out to better understand the diversity of microbes which preside within the belly buttons of humans sampled within a nation-wide citizen science project.
 - *They wanted to know if the microbiome of the human belly button is dominated by numerous phylotypes or just a few phylotypes.*
- The researchers of the study were also interested in witnessing if the “Oligarchy Hypothesis” was at play
 - *The researchers wanted to know if the frequency of phylotypes within one human population sample can predict the frequency of the same phylotypes of a second independent sample.*

Introduction:

Data Used in Analysis

Addressing Question 1.

- `sample_info.csv`: This file contains two columns, one with the sample ID and one with the population from which the sample came from.
- `OTU_file.txt`: This file is the OTU table which contains the number of sequences of each Operational Taxonomic Unit (OTU) in each sample (individual belly button). The row names are the OTUs and the column names are the sample IDs; however, the first column is the lowest taxonomic level.

Addressing Question 2.

- `raw_seqs_BB.fna`: This file contains the raw sequences of each sequence found in all the samples. Each sequence is linked to the sample they were found in and contain a unique identifier.



Biological Questions

1. Do the belly button microbiomes differ among the two populations of people sampled?
 - I.e. Can the diversity within one population predict another?
1. What is the evolutionary relationship among the phyla found in one sample (one belly button)?
 - Is the microbiome dominated by a few phylotypes only, or are the phylotypes very diverse and distantly related?

Question 1

Do the belly button microbiomes differ among the two populations sampled?

Outline of Steps:

- Import and setup the .csv with sample information and the OTU table
- Calculate the binary distance and create a neighbour-joining tree
- Calculate the euclidean distance and create a neighbour-joining tree
- Calculate the Bray-Curtis Dissimilarity and create a neighbour-joining tree
- Plot the Non-Metric Multidimensional Scaling (NMDS) algorithm

Sample Information & OTU Table

- First the sample information was imported. This linked each sample with the population it came from and will be used to colour code the cluster analyses.
- Next, the OTU table was imported. In order to properly work with the data, however, the column containing the taxonomy was removed and the total sequences in the OTU table was calculated.

```
OTU_table <- read.delim("data/OTU_file.txt", header = T, row.names = "X.OTU.ID")
```

```
OTU_data <- OTU_table[, -c(1, ncol(OTU_table))]
```

```
x <- rowSums(OTU_data)
```

```
sum(x)
```

- With 24,000 reads, there is a likelihood of contaminated data. To remove these, we removed any OTUs that did not have more than one sequence read in more than one sample.

```
drop <- rowSums(OTU_data) < 2
```

```
sum(drop) # The number of sequences being removed.
```

```
OTU_red <- OTU_data[!drop, ]
```

- Next, the table was formatted so that the species were laid across the top and samples were along the side.

```
OTU_red[1:3, 1:3]
```

```
OTUs <- as.data.frame(t(OTU_red))
```

```
OTUs[1:3, 1:3]
```

Binary Method

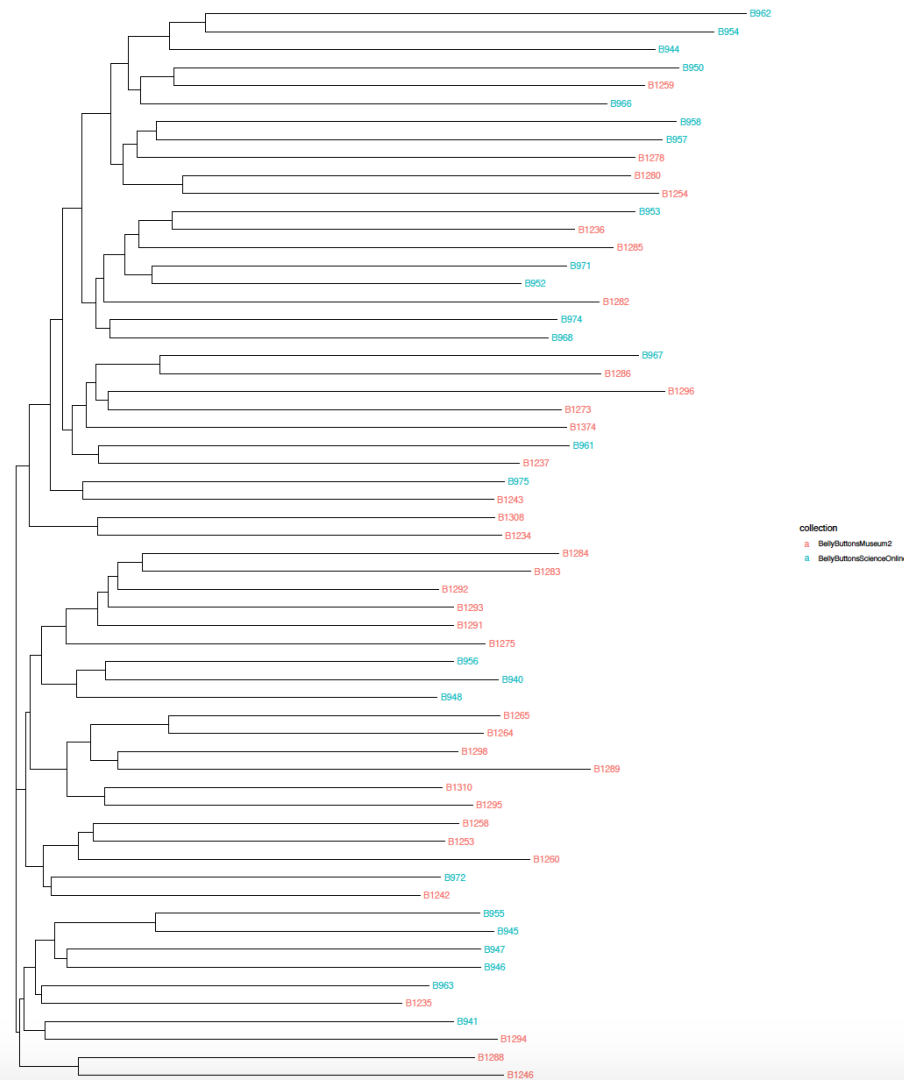
- Next, a pairwise distance of the binary matrix was calculated

```
OTU_bin_dist <- dist(OTU_bin, method = "binary")
```

- The pairwise distance of the binary matrix was utilized to create a binary neighbour-joining tree which was annotated using sample information
- It was then outputted as a pdf

```
OTU_bin_tree <- nj(OTU_bin_dist)
ggtree(OTU_bin_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
```

```
pdf(width = 16, height = 20, "Binary_Tree.pdf")
ggtree(OTU_bin_tree, layout = "rectangular") %<+% Samples +
  geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")
dev.off()
```

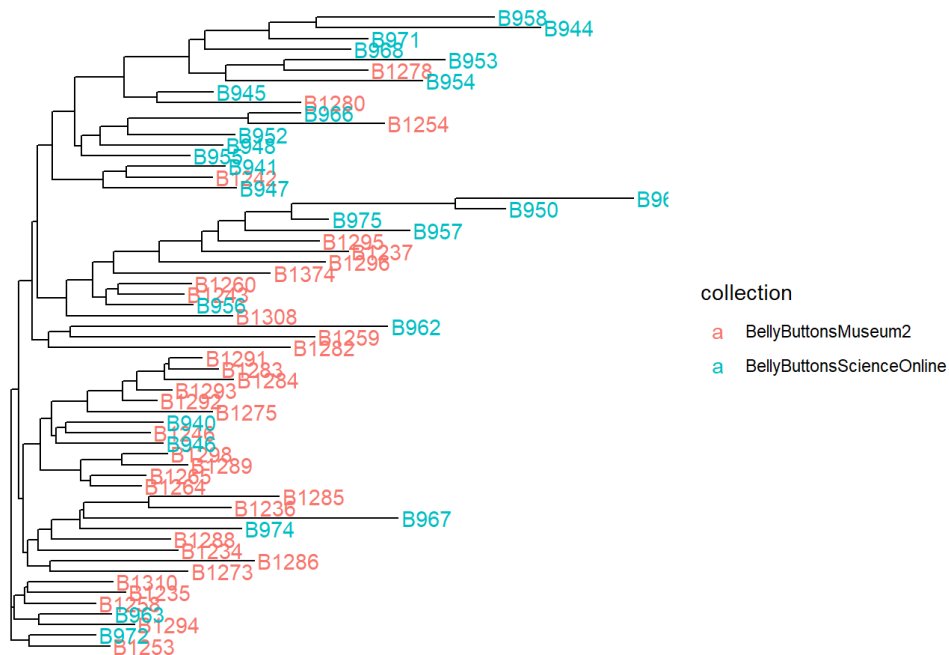


Euclidean Method

- Here we calculated the euclidean distance which was used to build the euclidean neighbour-joining tree.
- It was annotated using the sample information and was saved as a pdf

```
OTU_euc_dist <- dist(OTUs, method = "euclidean")
OTU_euc_tree <- nj(OTU_euc_dist)
ggtree(OTU_euc_tree, layout = "rectangular") %<+%
Samples + geom_tiplab(aes(colour = collection)) +
theme(legend.position = "right")

pdf(width = 16, height = 20, "Euclidean_Tree.pdf")
ggtree(OTU_euc_tree, layout = "rectangular") %<+%
Samples + geom_tiplab(aes(colour = collection)) +
theme(legend.position = "right") dev.off()
```

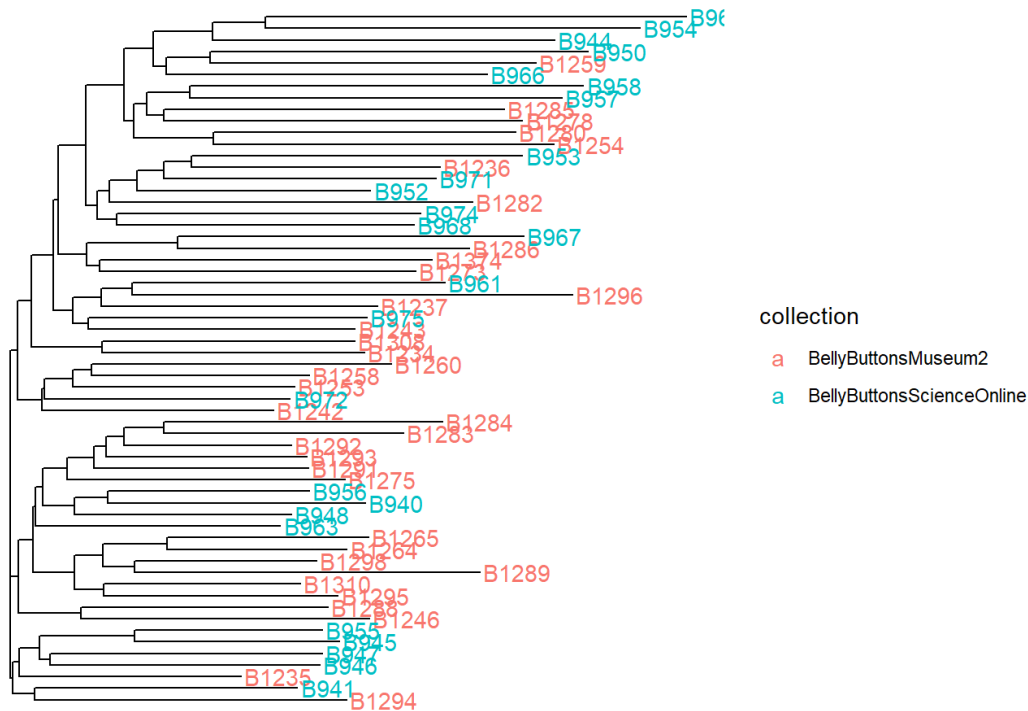


Bray-Curtis Dissimilarity

- The Bray-Curtis dissimilarity was calculated which was used to build a third neighbour-joining tree. This too was annotated using the sample information, and was also saved as a pdf

```
OTU_bc_dist <- vegdist(OTUs, method = "bray", binary = T)
OTU_bc_tree <- nj(OTU_bc_dist)
ggtree(OTU_bc_tree, layout = "rectangular") %<+>%
  Samples + geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")

pdf(width = 16, height = 20, "Bray_Curtis_Tree.pdf")
ggtree(OTU_bc_tree, layout = "rectangular") %<+>%
  Samples + geom_tiplab(aes(colour = collection)) +
  theme(legend.position = "right")dev.off()
```



NMDS

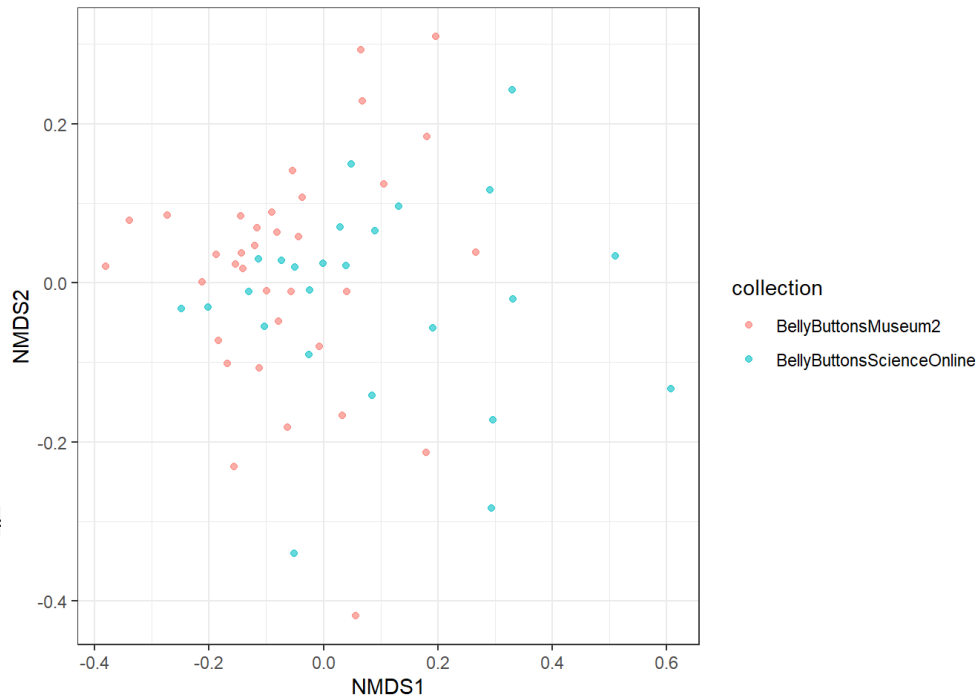
- Finally, we plotted the NMDS to visualize the level of similarity of individual cases of each population

Set up of Model

```
set.seed(13)
NMDSdat <- metaMDS(OTU_bc_dist, k = 2, trymax = 100)
PDat <- data.frame(NMDS1 = NMDSdat$points[, 1], NMDS2 =
  NMDSdat$points[, 2], sample = row.names(OTUs))
PDat <- merge(PDat, Samples, by = "sample", all.x = T, all.y = F)
```

Plotting NMDS

```
qplot(x = NMDS1, NMDS2, colour = collection, alpha = I(0.6), data =
  PDat) + theme_bw()
```



Rationale for Methods (Q1)

Binary Method

The cluster analysis using the binary method shows the similarities/differences between the samples based on the presence or absence of the species found in each sample. Colour coding was then done to compare the similarities/differences between the populations from which the samples were taken from.

Euclidean Method

The euclidean method was used to determine if the abundance of different species in the samples had any effect on the results.

Rationale for Methods (Q1) (Con't)

Bray-Curtis Dissimilarity

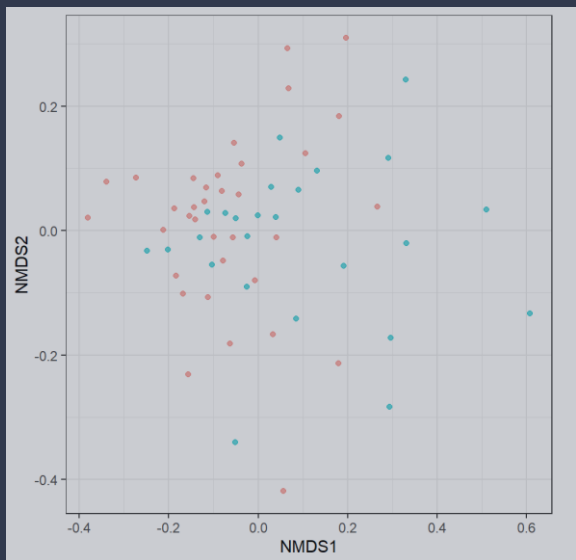
The Bray-Curtis Dissimilarity was used because it takes into consideration the abundance of different species in the samples but it is not as sensitive to abundance as the euclidean method is.

NMDS

This algorithm was used to also determine the similarities/differences between the samples. To visualize the results of this model a bivariate plot was generated where each point represents a sample and points that are closer together are more similar.

Conclusion drawn based on question 1.

Do the belly button microbiomes differ among the two populations sampled?



- After careful observation of all trees constructed as well as the NMDS generated bivariate plot, we were able to conclude that there are no clear distinctions which can be drawn regarding diversity of the microbiomes among the two populations sampled.
- These findings would imply that one population of human belly button microbiome samples is in fact able to predict another.

Question 2

What is the evolutionary relationship among the phyla found in one sample (one belly button)?

Outline of Steps

- Import FASTA file (sample B1285 selected as index)
 - B1285 is a sample selected at random which was used to conduct various analyses to address question 2.
- Align sequence data using MUSCLE
 - A multiple sequence alignment algorithm
- Inspect the alignment in sections
 - Beginning, middle, end of alignment
 - Remove sequences with too many gaps
- Analyze and visualize alignment:
 - Heat map
 - Phylogenetic tree

Import and Align Data

- First, the FASTA file was imported. A single sample was selected (B1285) to observe the diversity of microbes within that sample. A “for loop” was used to compile all sequence reads from B1285 into an object called “sub” since all FASTA reads were originally compiled into a single FASTA file.

```
myFasta <- read.fasta(file = "data/raw_seqs_BB.fna", seqtype = "AA", as.string = TRUE, set.attributes = FALSE)
nam <- names(myFasta)

indexes <- grep("B1285", nam)

sub <- rep(NA, 1170)

for (i in 1:length(indexes)){
  sub[i] <- myFasta[indexes[i]]
}
```

Import and Align Data

- Following that, a new data.frame was created using the index numbers as IDs as well as the sequence data from the myFasta object (which possessed the original raw sequence FASTA data) in the Seq column. The object 'dna' was created using the sapply function to separate each base pair into columns, and the row names were re-named to the indices from the original file into the myFasta file

```
df <- data.frame(ID = as.factor(indexes), Seq = paste(sub), stringsAsFactors = FALSE)
dna <- sapply(df$Seq, strsplit, split = "")
names(dna) <- paste(1:nrow(df), df$ID, sep="_")
dna_bin<-as.DNABin(dna)
```

- The 'dna' object was converted to bin object with as.DNABin so we can use muscle to align the sequence data

```
dna_align <- muscle(dna_bin, quiet=F)
```


Inspecting + Visualizing the Alignment

```
checkAlignment(dna_align, what = 1)
```

- It is tough to visualize the entire alignment so, we narrowed it down...

```
checkAlignment(dna_align[1:585, 1:150], what = 1)
```

- We looked at the beginning, middle, and end of the alignment for each half of the sequences

```
checkAlignment(dna_align[586:1170, 1:150], what = 1)
```

- Lastly, we needed to visualize the gaps in order to remove the sequences with an abundance of gaps (right)

```
seqLen <- as.numeric(lapply(dna_bin, length))
```

```
qplot(seqLen) +  
  theme_bw()
```

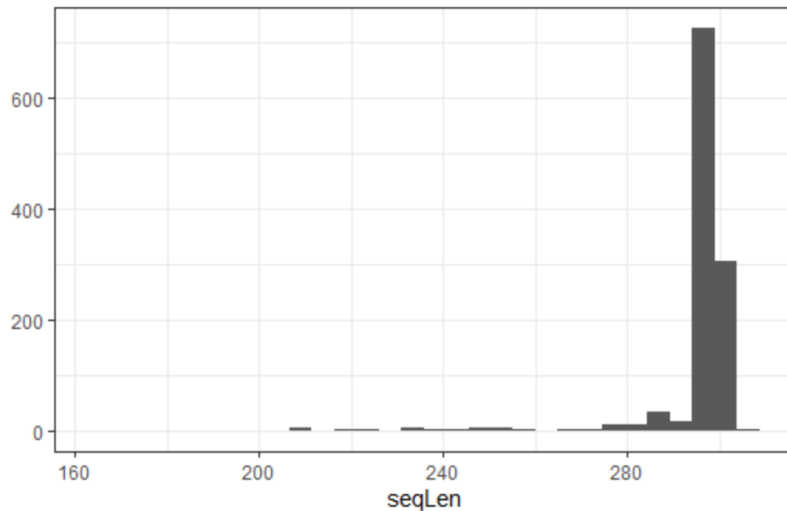
```
keepSeq <- seqLen > 290
```

```
dna_subset <- dna_align[keepSeq, ]
```

```
checkAlignment(dna_subset, what = 1)
```

- From this, we determined 290 as a cutoff to remove those sequences with large gaps. We then checked the alignment again, and determined that enough of the large gaps were removed so we re-aligned the sequences.

```
dna_sub_align <- muscle(dna_subset, quiet = F)
```



Visualization: Building a Heatmap (using ggplot)

- A distance matrix was created using the `dist.dna` function to compute a matrix of pairwise distances from DNA sequences using the `k80` model of DNA evolution
 - This model assumes that in nature, transition mutations are more likely to occur than transversions

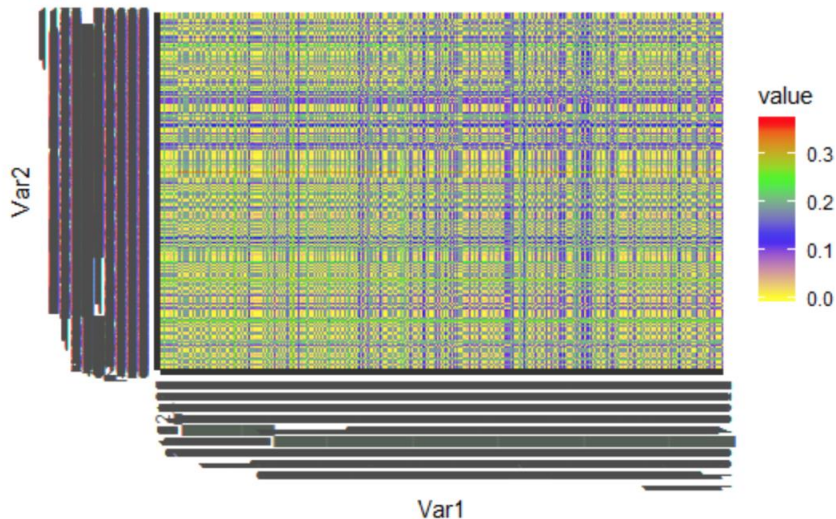
```
dnaDM<-dist.dna(dna_align, model="K80")  
dnaDMmat<-as.matrix(dnaDM)
```

- Next, in order to compare sequence similarities in a heatmap, the `melt` function was used to turn our file into a linear matrix

```
PDat<-melt(dnaDMmat)  
dim(PDat)  
ggplot(data = PDat, aes(x=Var1, y=Var2, fill=value)) +  
  geom_tile()+scale_fill_gradientn(colours=c("yellow","blue","green","red")) +  
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

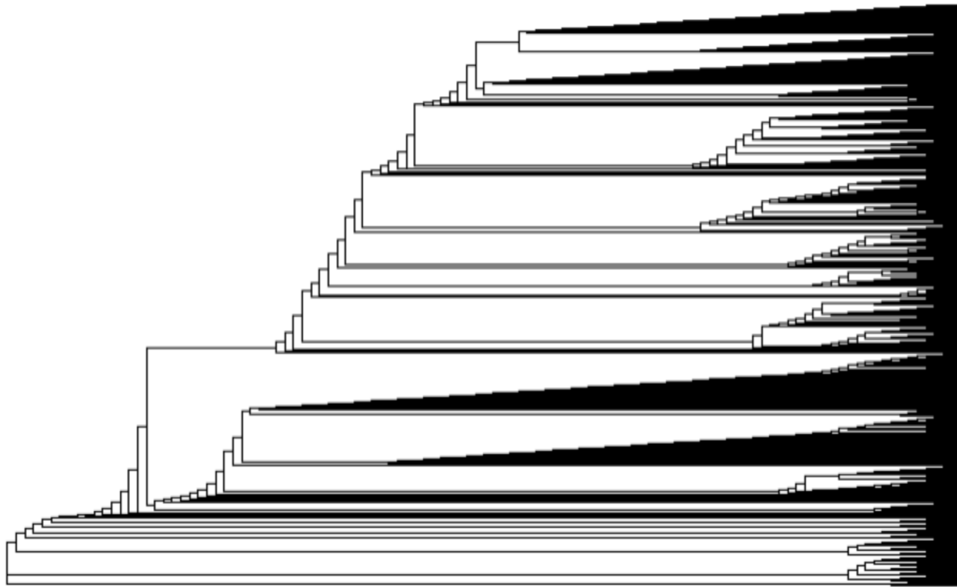
- Finally, the heatmap (right) was downloaded as a pdf

```
pdf(width=20,height=10, "HeatMap_B1285.pdf")  
ggplot(data = PDat, aes(x=Var1, y=Var2, fill=value)) +  
  geom_tile()+scale_fill_gradientn(colours=c("yellow","blue","green","red")) +  
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))  
dev.off()
```



Visualization:

Building a Phylogenetic Tree (using ggtree)



- A phylogenetic tree was created to visualize the evolutionary relationship among samples using the neighbour joining method

```
dnaTree<-nj(dnaDM)
str(dnaTree)
ggtree(dnaTree, branch.length='none')
```
- Each sequence read was less than 300 nucleotides long, the number of taxa was high so minimum evolution method was not optimal
- Branch length was specified as 'none' to observe relationships among taxa rather than evolutionary distance between taxa

```
pdf(width=20,height=10, "PhylogeneticTree_B1285")
ggtree(dnaTree, branch.length='none')
dev.off()
```
- Finally, the tree (left) was downloaded as a pdf

Rationale for Methods (Q2)

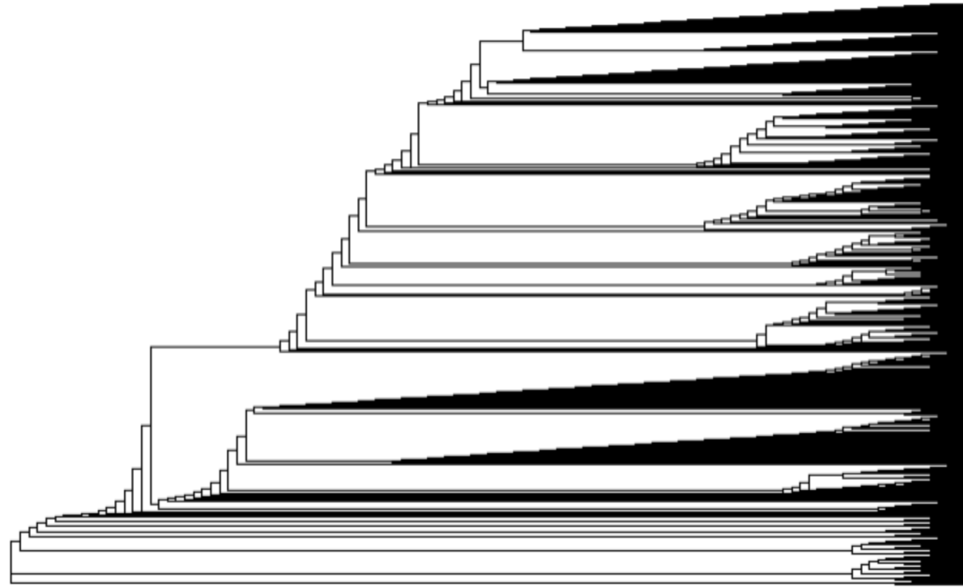
Heat Map

A heatmap was chosen to display this data with ggplot because it is a concise way to visualize three-dimensional data in only two dimensions. The variation among sequences and their similarities are represented by the heatmap. The color patterns in our heatmap may indicate an association between variables, which in this case are the pairwise distances from DNA sequences.

Phylogenetic Tree

In constructing a phylogenetic tree for evolutionary relatedness between all samples taken in the study, the neighbour-joining method was used with ggtree, which is a “bottom up” clustering method that suits sequence data. In particular, with a high volume of taxa but low sequence length (as these are bacterial and archeal genomes), NJ was superior to minimum evolution or hierarchical clustering.

Conclusion drawn
based on question
2.



- Based on the phylogenetic tree, we can interpret that there are a few dominating phylotypes, for example those indicated by the red arrows showing common ancestors that can connect a majority of the species
- This interpretation is in accordance with the Oligarchy Hypothesis presented by the research group

Conclusion

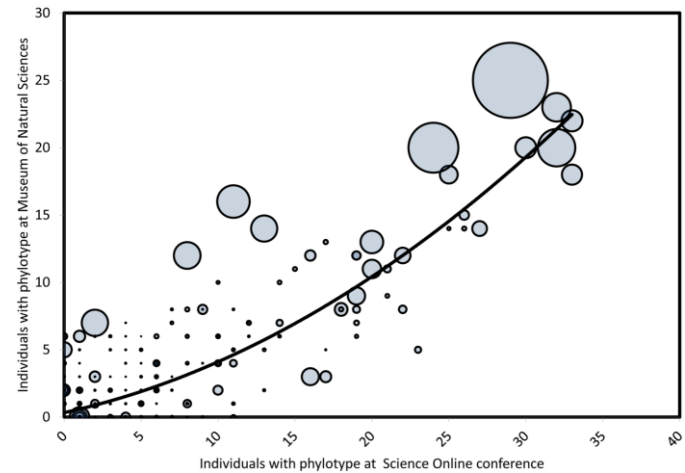


Figure 1. The frequency of bacterial phylotypes (each point = a phylotype) in our first sample of human belly buttons predicts most of the variation in the frequency of the same phylotypes in our second sample

- As seen in the graph displayed above, created by the researchers of the study cited using the same data which we used, the most frequent phylotypes tended to be significantly more abundant than others.
- This graph, along with the neighbour-joining trees generated to address question 1 provides evidence that a subset of phylotypes is both predictably present and abundant.
- The phylogenetic tree built further displays the overall diversity among the sequence reads within one selected sample (as an indicator for the others). While the sample is diverse, each does arise from common ancestors