BIOINFORMATICS MODULE

Table 1. Summary and reasoning of the best BLAST hit results for OTUs 301 - 310.

Otu #	Taxonomy	Score(Bits)	E value	Reasoning
Otu00301	*Eubacterium coprostanoligenes*	346	3.3E-95	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain HL)
Otu00302	*Parabacteroides johnsonii* *Parabacteroides merdae*	407	1E-113	The two species of bacteria had the same score and the E value, so only the *Parabacteroides* genus could be determined. Otherwise, these are the best BLAST hits because they had the highest scores and the lowest E values.
Otu00303	*Blastopiellula marina*	233	3E-61	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain DSM 3645)
Otu00304	*****No Hits Found*****	-	-	The genus and the species could not be determined. The bacterial sequences that produced significant alignments could not be found. (Not in the database)
Otu00305	*Prevotella copri*	357	2E-98	This is the best BLAST hit because it had the highest score and the lowest E value. Two strains of the bacteria, JCM 13464 and CB7 shared the same score and the E value. Longer query sequences may be needed to tell the strains apart from each other.
Otu00306	*Ethanoligenens harbinense*	329	3E-90	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain YUAN-3)
Otu00307	*Prevotella buccae*	244	1E-64	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain JCM 12245)
Otu00308	*Bacillus massilioanorexius*	468	7E-132	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain AP8)
Otu00309	*Muribaculum intestinale*	283	3E-76	This is the best BLAST hit because it had the highest score and the lowest E value. (Strain YL27)
Otu00310	*Paraprevotella clara*	278	1E-74	This is the best BLAST hit because it had the highest score and the lowest E value. The strains JCM 14859 and YIT 11840 had equal scores and E values. Longer query sequences may be needed to tell the strains apart from each other.

Codes used to obtain the BLAST hits:

```
> awk ‘NR==601, NR==620’ Module1_OTU.fasta > my_seqs.fasta
> blastn -query my_seqs.fasta -db 16SMicrobial -out my_seqs.blast.txt
```

Table 2. Alpha Diversity(Shannon/Inverse Simpson) and Alpha Richness(Chao1/ACE) values for V2

Shannon	Inverse Simpson	Chao1	ACE
3.296458	10.25484	143.071429	143.826306

Codes:

Loading the Dataset:

```
> library(vegan)
> OTU.table = read.table(file="Module1_OTU.txt", header=TRUE, row.names=1, sep="\t")
> alpha = read.table("alpha_values.txt", header=TRUE, sep="\t")
> DatasetV = alpha[alpha$Dataset_ID == "V",]
```

To obtain alpha diversity values (Shannon and Inverse Simpson):

```
> diversity(OTU.table["V2",], index="shannon")
> diversity(OTU.table["V2",], index="invsimpson")
> estimateR(OTU.table["V2",])
```

```
                V2
S.obs    118.000000
S.chao1 143.071429
se.chao1 12.784686
S.ACE    143.826306
se.ACE     6.051734
```

Figure 1. Boxplot of Alpha Diversity for dataset V. The y-axis shows the Shannon's Diversity Index. The boxplot shows the mean of 3.19094.

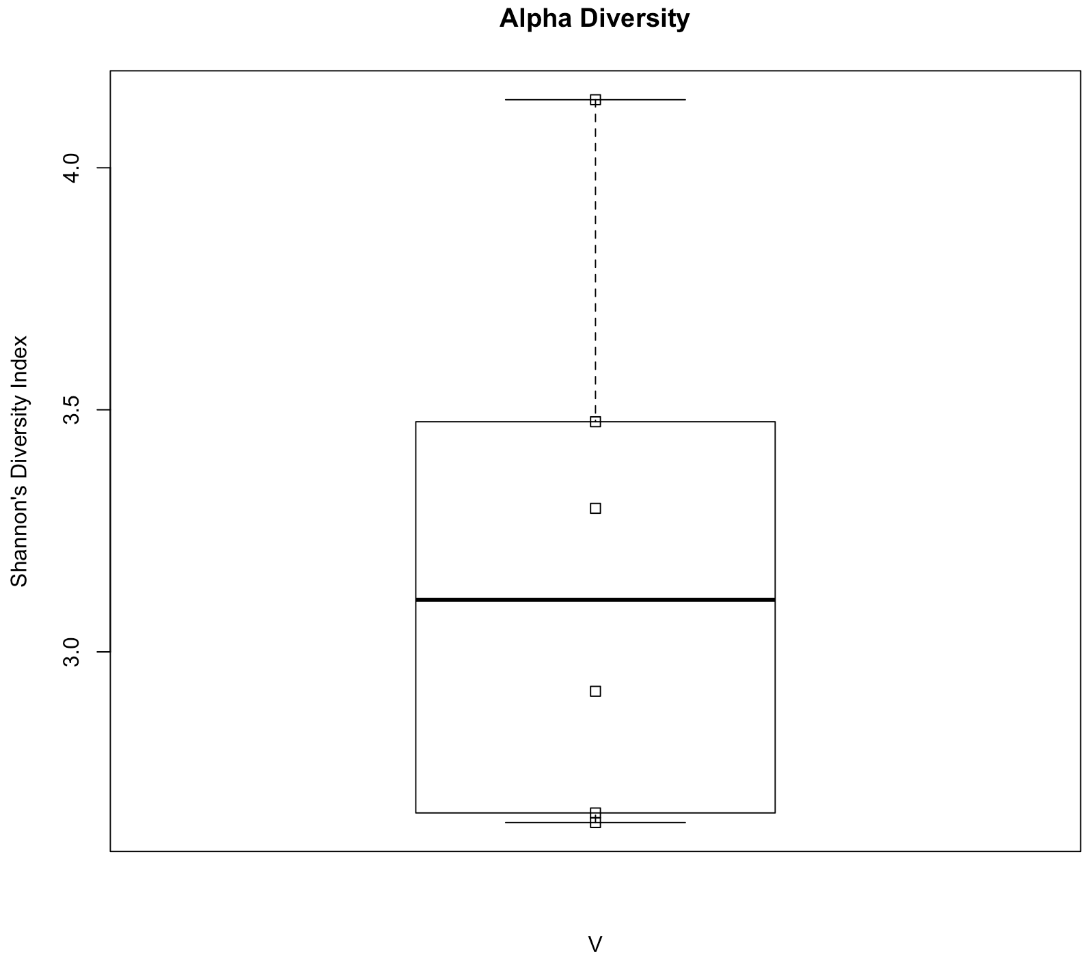
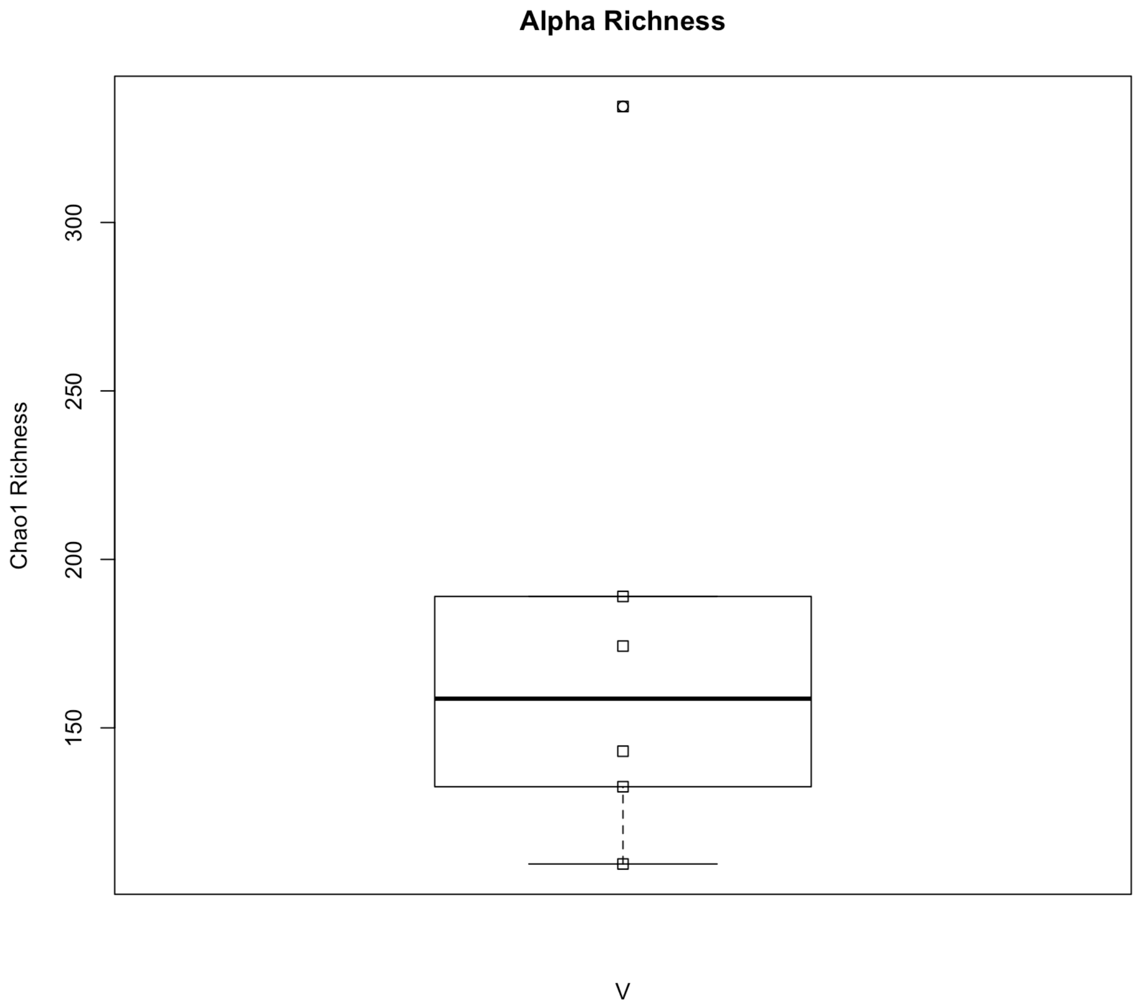


Figure 2. Boxplot of Alpha Richness for dataset V. The y-axis shows the Chao1 Richness Index. The individual specimen values are shown as symbols on the stripchart.



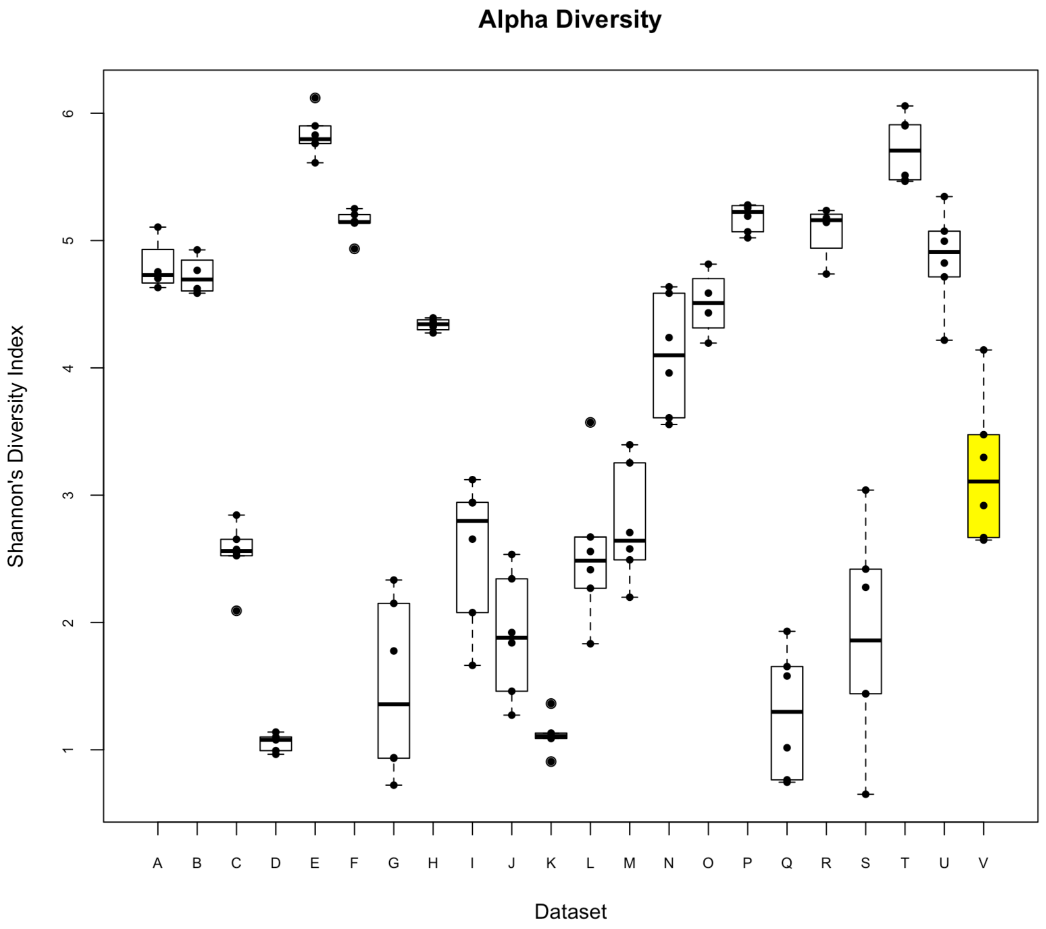
Codes used to obtain Alpha Diversity Boxplot:

```
> boxplot(c(DatasetV$Shannon) ~ c(DatasetV$Dataset_ID), main="Alpha Diversity", xlab="V",
ylab="Shannon's Diversity Index", names=c("V"))
> stripchart(c(DatasetV$Shannon) ~ c(DatasetV$Dataset_ID),vertical=TRUE, add=TRUE)
```

Codes used to obtain Alpha Richness Boxplot:

```
> boxplot(c(DatasetV$Chao1) ~ c(DatasetV$Dataset_ID), main="Alpha Richness", xlab="V",
ylab="Chao1 Richness", names=c("V"))
> stripchart(c(DatasetV$Chao1) ~ c(DatasetV$Dataset_ID),vertical=TRUE, add=TRUE)
```

Figure 4. Boxplots of Alpha Diversity across all (A – V) datasets. The y-axis shows the Shannon's Diversity Index. Highlighted boxplot shows V dataset.The individual specimens are represented by symbols as stripcharts.



Codes used to obtain Alpha Diversity Boxplot across all datasets (A - V):

```
> DatasetA = alpha[alpha$Dataset_ID == "A",]
> DatasetB = alpha[alpha$Dataset_ID == "B",]
> DatasetC = alpha[alpha$Dataset_ID == "C",]
….. > DatasetU = alpha[alpha$Dataset_ID == "U",]

> boxplot(alpha$Shannon ~ alpha$Dataset_ID, main="Alpha Diversity", xlab="Dataset",
ylab="Shannon's Diversity Index", cex.axis=0.7,
col=c("white","white","white","white","white","white","white","white","white","white","white","
white","white","white","white","white","white","white","white","white","white","yellow"))
> stripchart(alpha$Shannon ~ alpha$Dataset_ID, vertical=TRUE, add=TRUE, pch=20)
```

Table 3. The summary of P-Values for the t-tests and the statistical difference from dataset V.

DataSets	P-Value	P < 0.05
V-A	0.0000178	Yes, Significantly Different
V-B	0.0000555	Yes, Significantly Different
V-C	0.5374749	No
V-D	0.0000000	Yes, Significantly Different
V-E	0.0000000	Yes, Significantly Different
V-F	0.0000000	Yes, Significantly Different
V-G	0.0000001	Yes, Significantly Different
V-H	0.0123062	Yes, Significantly Different
V-I	0.6195657	No
V-J	0.0002030	Yes, Significantly Different
V-K	0.0000000	Yes, Significantly Different
V-L	0.5776164	No
V-M	0.9843585	No
V-N	0.0562718	No
V-O	0.0013597	Yes, Significantly Different
V-P	0.0000000	Yes, Significantly Different
V-Q	0.0000000	Yes, Significantly Different
V-R	0.0000002	Yes, Significantly Different
V-S	0.0001521	Yes, Significantly Different
V-T	0.0000000	Yes, Significantly Different
V-U	0.0000003	Yes, Significantly Different

Codes:

```
> TukeyHSD(aov(alpha$Shannon ~ alpha$Dataset_ID))
> mean(DatasetV$Shannon)
> var(DatasetV$Shannon)
```

I have concluded that my specimen (V2) feasibly originates from the Angel Fish Hindgut. First, the mean diversity of my sample(V) was 3.19042. Looking at Figure 4, the value was in the mid-range when compared to the mean of the other samples. Thus, I suspected that the specimen originates from animals with medium variability in diet. The differences were significant except for a few datasets including C, I, L, M, and N as shown in table 3. Next, the variance of diversity in my sample was 0.3275162, which was relatively high compared to the variance of other samples with similar alpha-diversity with the exception of I and L. This showed that the specimen within my sample had very different diversity values from one another. Moreover, since animals with an enlarged hindgut often shows higher diversity microbiomes, I have concluded that the specimen originates from either Angelfish hindgut or Horse feces. Since I and L had higher variance than V, Angelfish hindgut became a more probable origin because difficult to digest nutrients like lignin is not often found in algae, which contributes to lesser diversity compared to Horse feces. Further confirmation using combinations other measures of diversity such as beta/gamma-diversity may be needed.