Mar 10, 2016

Script R6: Virome Beta Diversity

F1000Research

In 1 collection

DOI
https://dx.doi.org/10.17504/protocols.io.einbcde

HANNIGAN GD[1], GRICE EA[1], ET AL.[1]

[1]DEPARTMENT OF DERMATOLOGY UNIVERSITY OF PENNSYLVANIA

VERVE Net | Club Grice

Geoffrey D Hannigan
Merck Research Labs

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.einbcde

External link: http://mbio.asm.org/content/6/5/e01578-15.full

Protocol Citation: HANNIGAN GD, GRICE EA, ET AL. 2016. Script R6: Virome Beta Diversity. **protocols.io** https://dx.doi.org/10.17504/protocols.io.einbcde

Manuscript citation:
Kindler L, Stoliartchouk A, Teytelman L, Hurwitz BL, Method-centered digital communities on protocols.io for fast-paced scientific innovation. F1000Research doi: 10.12688/f1000research.9453.2

Protocol status: Working

Created: February 11, 2016

Last Modified: November 09, 2017

Protocol Integer ID: 2350

Keywords: virome beta diversity, stranded dna virome, virome beta diversity this protocol, clustering of the virome sample, virome sample, virome, dynamic associations with the host microbiome, host microbiome, genetic enrichment, skin environment, human skin, dna

Abstract

This protocol outlines the Bray-Curtis dissimilarity NMDS ordination and significance analysis of our manuscript. Here we will look at the clustering of the virome samples using our reference independent OTU table based on contig hits. We will compare skin environments (sebaceous, etc), occlusion status, and time. We also look at the difference between the background controls and the rest of the samples. Based on the methods from the following publication:

Hannigan, Geoffrey D., et al. "The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome." *mBio* 6.5 (2015): e01578-15.

Guidelines

sessionInfo()

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.4 (Yosemite)
## ## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5    formatR_1.2    tools_3.2.0    htmltools_0.2.6
## [5] yaml_2.1.13    stringi_0.4-1    rmarkdown_0.7    knitr_1.10.5
## [9] stringr_1.0.0    digest_0.6.8    evaluate_0.7
```

Before start

Supplemental information available at:

https://figshare.com/articles/The_Human_Skin_dsDNA_Virome_Topographical_and_Temporal_Diversity_Genetic_Enrichment_and_Dynamic_Associations_with_the_Host_Microbiome/1281248

1 Load the required R packages.

Command

```
library(vegan)
packageVersion(
```

Expected result

```
## [1] '2.3.0'
```

Expected result

```
## [1] '1.0.1'
```

Expected result

```
## [1] '0.3.35'
```

Expected result

```
## [1] '1.4.1'
```

Expected result

```
## [1] '1.8.2'
```

Expected result

```
## [1] '1.6.0'
```

2 Import OTU table.

Command

```
INPUT <- read.delim(
```

3 Remove last column because it contains NAs (due to transposing Python script upstream).

Command

```
INPUT_NO_FINAL <- INPUT[ ,c(1:74361)]
```

4 Import mapping file.

Command

```
INPUT_MAP <- read.delim(
```

5 Here we need to reformat the mapping files. This means only looking at the two time points for which we have a complete data set (we have only partial data for time point 1), as well as excluding the sites and subjects for which we only have partial sampling (as mentioned in previous notebooks).

6 We will be generating NMDS plots to visualize the distances between virome samples based on site environment, occlusion status, and time point. We also calculate the significance using the adonis test which was stratified across subjects.

7 Generate subset of mapping file for only the specific anatomic sites and all time pints (2 and 3). Remove rows in mapping file with ID of NA (meaning the mapping row belongs only to the metagenome dataset).

Command

```
SUBSET_MAP <- INPUT_MAP[-which(INPUT_MAP$NexteraXT_Virome_SampleID
%in% NA), ]
SUBSET_MAP <- SUBSET_MAP[which(SUBSET_MAP$TimePoint %in% c(2,3)), ]
SUBSET_MAP <- SUBSET_MAP[-which(SUBSET_MAP$Site_Symbol %in% c(
```

8 Get only the samples described in the map subset.

Command

```
KEEP_SAMPLES <- as.vector(SUBSET_MAP$NexteraXT_Virome_SampleID)
INPUT_SUBSET <- INPUT_NO_FINAL[which(INPUT_NO_FINAL$ContigID %in%
c(KEEP_SAMPLES)), ]
row.names(INPUT_SUBSET) <- INPUT_SUBSET[,1]
INPUT_SUB_FORMAT <- INPUT_SUBSET[,-1]
head(INPUT_SUB_FORMAT)[,c(1:5)]
```

Expected result

	##		X1		X2		X4		X5
	## MG100195		0.0000		0		1.343610		2.71473
	## MG100198		0.0000		0		0.363769		0.00000
	## MG100199		0.0000		0		0.955394		0.00000
	## MG100200		25.9488		0		0.457572		2.31129
	## MG100201		0.0000		0		0.155531		0.00000
	## MG100202		50.3225		0		1.361750		5.15885

9 Generate distance matrix using Bray Curtis for all time point (2 and 3).

Command

```
INPUT_SUBSET_DIST_MATRIX <- vegdist(INPUT_SUB_FORMAT, method =
```

10 Visualize the distance matrix using NMDS.

Command

```
BRAY_ORD_NMDS <- metaMDS(INPUT_SUBSET_DIST_MATRIX,k=3)
```

Expected result

```
## Run 0 stress 0.1579314
## Run 1 stress 0.1611112
## Run 2 stress 0.1626692
## Run 3 stress 0.1582604
## ... procrustes: rmse 0.03032907 max resid 0.2032916
## Run 4 stress 0.1601818
## Run 5 stress 0.1595255
## Run 6 stress 0.1567941
## ... New best solution
## ... procrustes: rmse 0.04472259 max resid 0.3011937
## Run 7 stress 0.157843
## Run 8 stress 0.1617624
## Run 9 stress 0.165309
## Run 10 stress 0.1575912
## Run 11 stress 0.1580084
## Run 12 stress 0.1601537
## Run 13 stress 0.1579342
## Run 14 stress 0.1634542
## Run 15 stress 0.1602448
## Run 16 stress 0.1593178
## Run 17 stress 0.157845
## Run 18 stress 0.1563824
## ... New best solution
## ... procrustes: rmse 0.02635918 max resid 0.1612987
## Run 19 stress 0.1583989
## Run 20 stress 0.1600246
```

11 Record the stress value.

Command

```
BRAY_ORD_FIT = data.frame(MDS1 = BRAY_ORD_NMDS$points[,1], MDS2 =
BRAY_ORD_NMDS$points[,2], MDS3 = BRAY_ORD_NMDS$points[,3])
BRAY_ORD_NMDS$stress
```

Expected result

```
## [1] 0.1563824
```

12 Generate and visualize merged NMDS and MAP data.

Command

```
BRAY_ORD_FIT$SampleID <- rownames(BRAY_ORD_FIT)
NMDS_AND_MAP <- merge(BRAY_ORD_FIT, SUBSET_MAP, by.x=
```

Expected result

	##		SampleID		MDS1		MDS2		MDS3		NexteraXT_SampleID
	## 1		MG100195		-0.039764304		0.10493926		-0.27817131		MG100171
	## 2		MG100198		-0.164123841		0.04221565		0.02443569		MG100174
	## 3		MG100199		-0.034656783		0.06907268		-0.09141366		MG100175
	## 4		MG100200		-0.037125834		0.08835104		-0.01446685		MG100176
	## 5		MG100201		0.0102459480		0.09353651		0.11426126		MG100177
	## 6		MG100202		0.002223516		0.10249228		0.02265875		MG100178

13 Start plotting the figures.

Note

The legends will show up on top of the figures here, but running the first plot without the second legend line will generate the plot alone (this is what was used in publication)

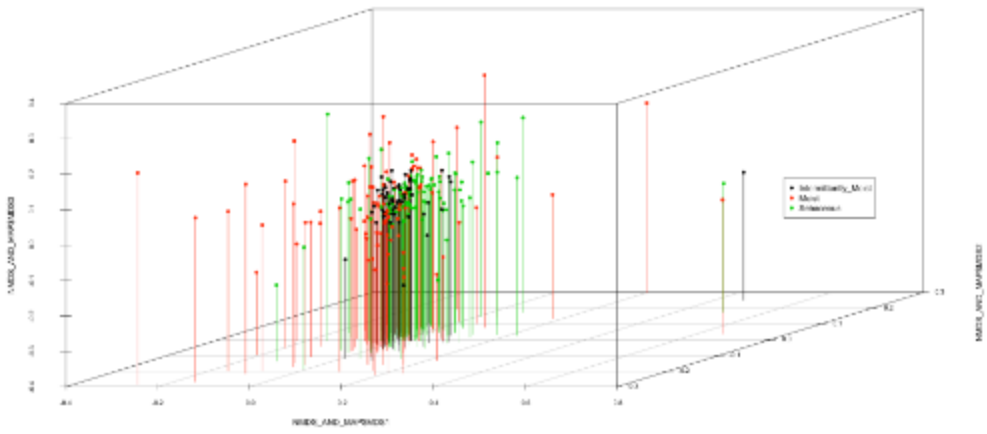
Note

The legends will show up on top of the figures here, but running the first plot without the second legend line will generate the plot alone (this is what was used in publication)

Command

```
s3d <-
scatterplot3d(NMDS_AND_MAP$MDS1,NMDS_AND_MAP$MDS2,NMDS_AND_MAP$MDS3,
pch=16, color=as.integer(factor(NMDS_AND_MAP$Site_Categories)), type=
```

Expected result



Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$Site_Categories)	2		2.562		1.28109		5.0118
	Residuals	249		63.649		0.25562		
	Total	251		66.211				

Residuals
Total

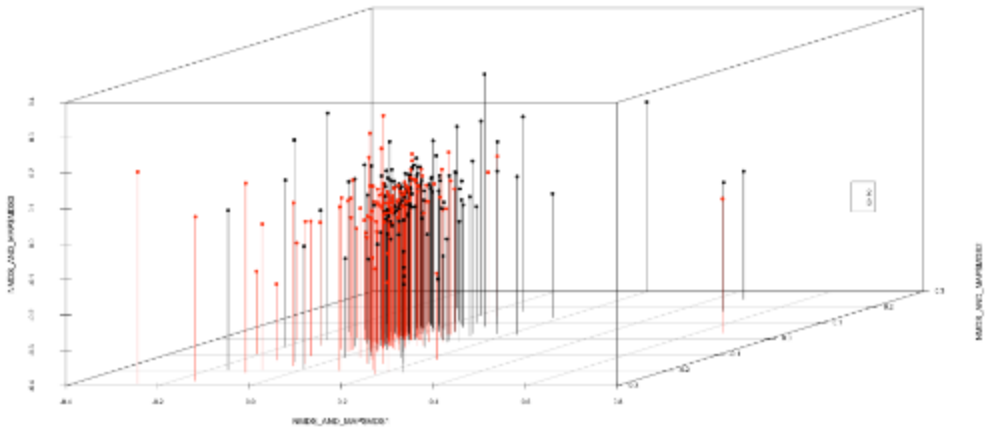
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

14 Plot time point.

Command

```
s3d <-
scatterplot3d(NMDS_AND_MAP$MDS1,NMDS_AND_MAP$MDS2,NMDS_AND_MAP$MDS3,
pch=16, color=as.integer(factor(NMDS_AND_MAP$TimePoint)), type=
```

Expected result



Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$TimePoint)	1		2.279		2.27899		8.9118
	Residuals	250		63.932		0.25573		
	Total	251		66.211				

Residuals
Total

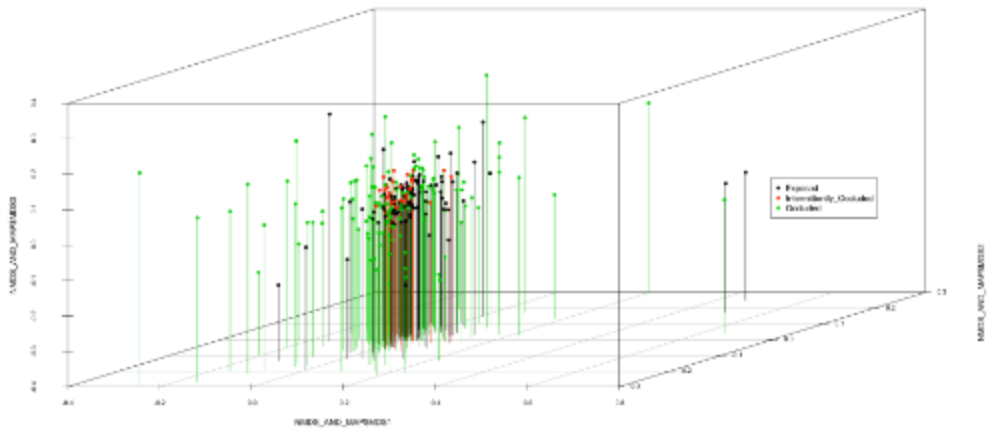
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

15 Plot ordination for occlusion site status.

Command

```
s3d <-
scatterplot3d(NMDS_AND_MAP$MDS1,NMDS_AND_MAP$MDS2,NMDS_AND_MAP$MDS3,
pch=16, color=as.integer(factor(NMDS_AND_MAP$Occlusion)), type=
```

Expected result



Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$TimePoint)	1		2.279		2.27899		8.9118
	Residuals	250		63.932		0.25573		
	Total	251		66.211				

Residuals
Total

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

16 We will also perform the negative control ordination to validate we have samples over background, which are significantly different from the rest of the samples.

17 Environmental background contol in the analysis Generate subset of mapping file for only the specific anatomic sites and all time points (2 and 3).

Command

```
SUBSET_MAP <- INPUT_MAP[-which(INPUT_MAP$NexteraXT_Virome_SampleID
%in% NA), ]
SUBSET_MAP <- SUBSET_MAP[which(SUBSET_MAP$TimePoint %in% c(2, 3)), ]
SUBSET_MAP <- SUBSET_MAP[-which(SUBSET_MAP$Site_Symbol %in% c(
```

18 This will define each sample as a control or not:

Command

```
for (i in 1:length(SUBSET_MAP$Site_Categories)) { if
(SUBSET_MAP$Site_Categories[i] ==
```

19 Get only the samples described in the map subset.

Command

```
KEEP_SAMPLES <- as.vector(SUBSET_MAP$NexteraXT_Virome_SampleID)
INPUT_SUBSET <- INPUT_NO_FINAL[which(INPUT_NO_FINAL$ContigID %in%
c(KEEP_SAMPLES)), ]
row.names(INPUT_SUBSET) <- INPUT_SUBSET[, 1]
INPUT_SUB_FORMAT <- INPUT_SUBSET[, -1]
```

20 Generate distance matrix using Bray Curtis for all time points (2 and 3).

Command

```
INPUT_SUBSET_DIST_MATRIX <- vegdist(INPUT_SUB_FORMAT, method =
```

21 Visualize the distance matrix using NMDS.

Command

```
BRAY_ORD_NMDS <- metaMDS(INPUT_SUBSET_DIST_MATRIX, k = 2)
```

Expected result

```
## Run 0 stress 0.2106771
## Run 1 stress 0.2152638
## Run 2 stress 0.2212825
## Run 3 stress 0.2131834
## Run 4 stress 0.2137311
## Run 5 stress 0.2188144
## Run 6 stress 0.2098836
## ... New best solution
## ... procrustes: rmse 0.02215487 max resid 0.2231707
## Run 7 stress 0.2292835
## Run 8 stress 0.218468
## Run 9 stress 0.2248454
## Run 10 stress 0.2147353
## Run 11 stress 0.2143595
## Run 12 stress 0.2191855
## Run 13 stress 0.2148725
## Run 14 stress 0.2248317
## Run 15 stress 0.2357207
## Run 16 stress 0.2133108
## Run 17 stress 0.217102
## Run 18 stress 0.2149535
## Run 19 stress 0.210627
## Run 20 stress 0.2231295
```

22 Record the stress value.

Command

```
BRAY_ORD_FIT = data.frame(MDS1 = BRAY_ORD_NMDS$points[, 1], MDS2 =
BRAY_ORD_NMDS$points[, 2])
BRAY_ORD_NMDS$stress
```

Expected result

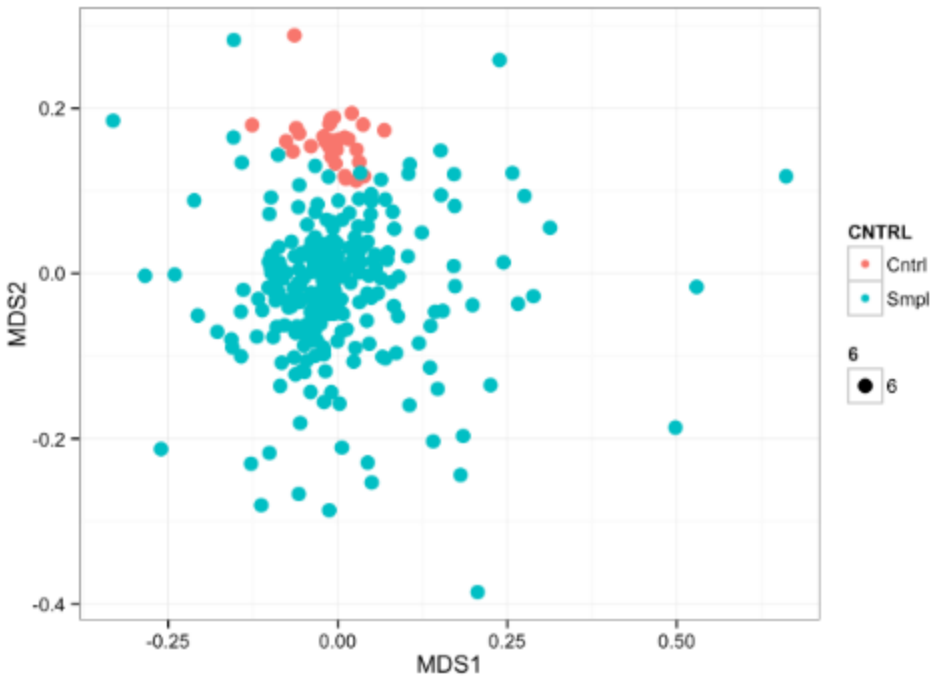
```
## [1] 0.2098836
```

23 Plot MDS1 vs MDS2.

Command

```
RAY_ORD_FIT$SampleID <- rownames(BRAY_ORD_FIT)
NMDS_AND_MAP <- merge(BRAY_ORD_FIT, SUBSET_MAP, by.x =
```

Expected result



24 Calculate the significance of the differences between the environmental background and the rest of the samples.

Command

```
adonis(INPUT_SUBSET_DIST_MATRIX ~ factor(NMDS_AND_MAP$CNTRL), perm =
999, strata = factor(NMDS_AND_MAP$SubjectID))
```

Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$CNTRL)	1		5.439		5.4391		21.422
	Residuals	282		71.601		0.2539		
	Total	283		77.040				

Residuals
Total

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

25 In order to evaluate the utility of reference-independent methods, we also calculated alpha and beta diversity metrics using our reference-dependent taxonomic data.

Command

```
INPUT_SPECIES <- read.delim(
```

26 Also import the mapping file.

Command

```
INPUT_MAP <- read.delim(
```

27 Subset the mapping file.

Command

```
SUBSET_MAP <- INPUT_MAP[which(INPUT_MAP$TimePoint %in% c(2, 3)), ]
SUBSET_MAP <- SUBSET_MAP[-which(SUBSET_MAP$Site_Symbol %in% c(
```

28 Same formatting and parsing as above.

Note

From here we would calculate the mean across the various sites and merge in the mapping file data but we are going to want to use this file here for distance matrix calculations.

Command

```
KEEP_SAMPLES <- as.vector(SUBSET_MAP$NexteraXT_Virome_SampleID)
INPUT_SUBSET <- INPUT_SPECIES[which(INPUT_SPECIES$V3 %in%
c(KEEP_SAMPLES)), ]
INPUT_SUBSET <- INPUT_SUBSET[-which(INPUT_SUBSET$V1 %in% c(
```

29 Convert the file to wide format so that it can be used for distance matrix calculations.

Command

```
inputWide <- dcast(INPUT_SUBSET, V1 ~ V3, value.var =
```

30 I pulled out a couple values to ensure the transformation was correct.

Command

```
inputWideTranspose <- as.data.frame(t(inputWide[, -1]))
INPUT_SUBSET_DIST_MATRIX <- vegdist(inputWideTranspose, method =
```

31 Visualize the distance matrix using NMDS.

Command

```
BRAY_ORD_NMDS <- metaMDS(INPUT_SUBSET_DIST_MATRIX, k = 3)
```

Expected result

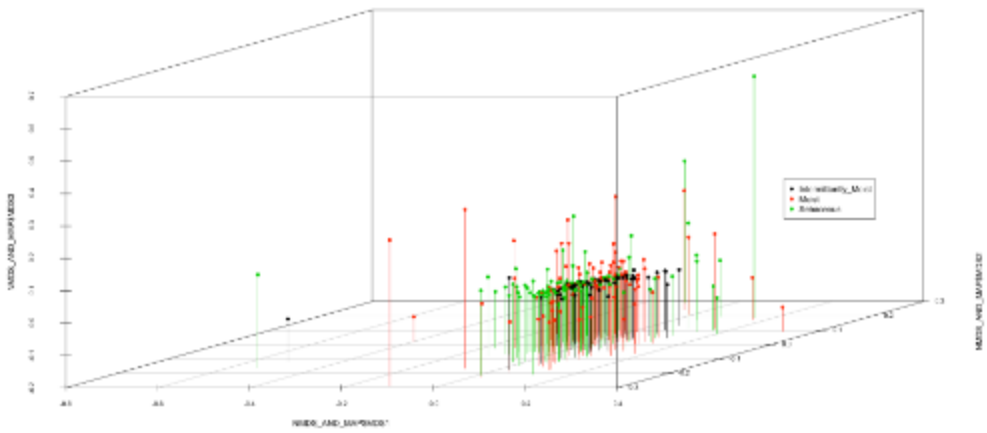
```
## Run 0 stress 0.08479484
## Run 1 stress 0.08752696
## Run 2 stress 0.0945716
## Run 3 stress 0.09044704
## Run 4 stress 0.09281868
## Run 5 stress 0.09634914
## Run 6 stress 0.09194107
## Run 7 stress 0.09115468
## Run 8 stress 0.08718591
## Run 9 stress 0.09208599
## Run 10 stress 0.08650345
## Run 11 stress 0.09094712
## Run 12 stress 0.09453678
## Run 13 stress 0.09349933
## Run 14 stress 0.08866572
## Run 15 stress 0.09333028
## Run 16 stress 0.09083254
## Run 17 stress 0.08825317
## Run 18 stress 0.0877584
## Run 19 stress 0.09264528
## Run 20 stress 0.09402247
```

32 Plot the data.

Command

```
BRAY_ORD_FIT = data.frame(MDS1 = BRAY_ORD_NMDS$points[, 1], MDS2 =
BRAY_ORD_NMDS$points[, 2], MDS3 = BRAY_ORD_NMDS$points[, 3])
BRAY_ORD_NMDS_STRESS <- BRAY_ORD_NMDS$stress
NMDS_AND_MAP <- merge(BRAY_ORD_FIT, SUBSET_MAP, by.x =
```

Expected result



33 Test the significance.

Command

```
adonis(INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), perm = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
```

Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$Site_Categories)	2		0.6099		0.304959		3.8099
	Residuals	249		19.9311		0.080045		
	Total	251		20.5410				

Residuals
Total

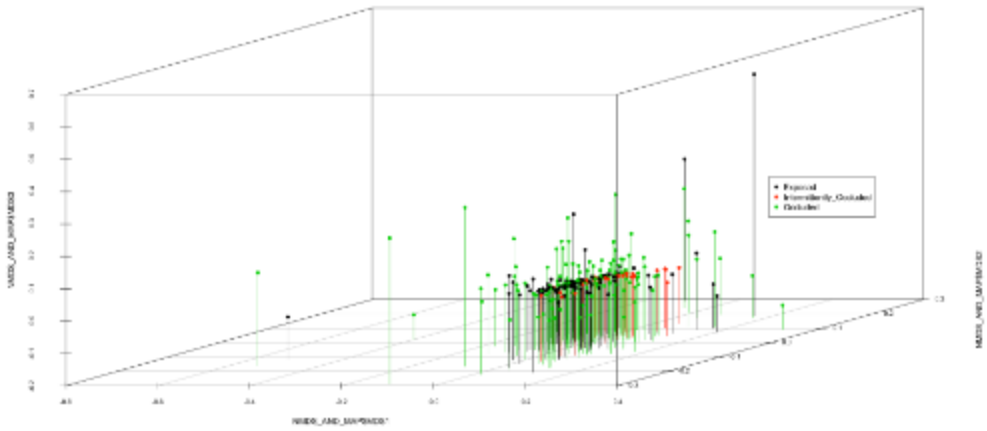
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

34 Also plot occlusion.

Command

```
s3d <- scatterplot3d(NMDS_AND_MAP$MDS1, NMDS_AND_MAP$MDS2,
NMDS_AND_MAP$MDS3, pch = 16, color =
as.integer(factor(NMDS_AND_MAP$Occlusion)), type =
```

Expected result



35 Test the significance.

Note

There is a significant difference between the sites, even when using the taxonomic profiles, suggests that even though the annotated portion of the data only represents a small portion of the community, it is actually able to detect the large trends for the overall community.

Command

```
adonis(INPUT_SUBSET_DIST_MATRIX ~ factor(NMDS_AND_MAP$Occlusion),
perm = 999, strata = factor(NMDS_AND_MAP$SubjectID))
```

Expected result

##
Call:
adonis(formula = INPUT_SUBSET_DIST_MATRIX ~
factor(NMDS_AND_MAP$Site_Categories), permutations = 999, strata =
factor(NMDS_AND_MAP$SubjectID))
##
Blocks: strata
Permutation: free ## Number of permutations: 999
##
Terms added sequentially (first to last)
##

		Df		SumsOfSqs		MeansSqs		F.Model
	factor(NMDS_AND_MAP$Occlusion)	2		0.6008		0.300423		3.7515
	Residuals	249		19.9402		0.080081		
	Total	251		20.5410				

Residuals
Total

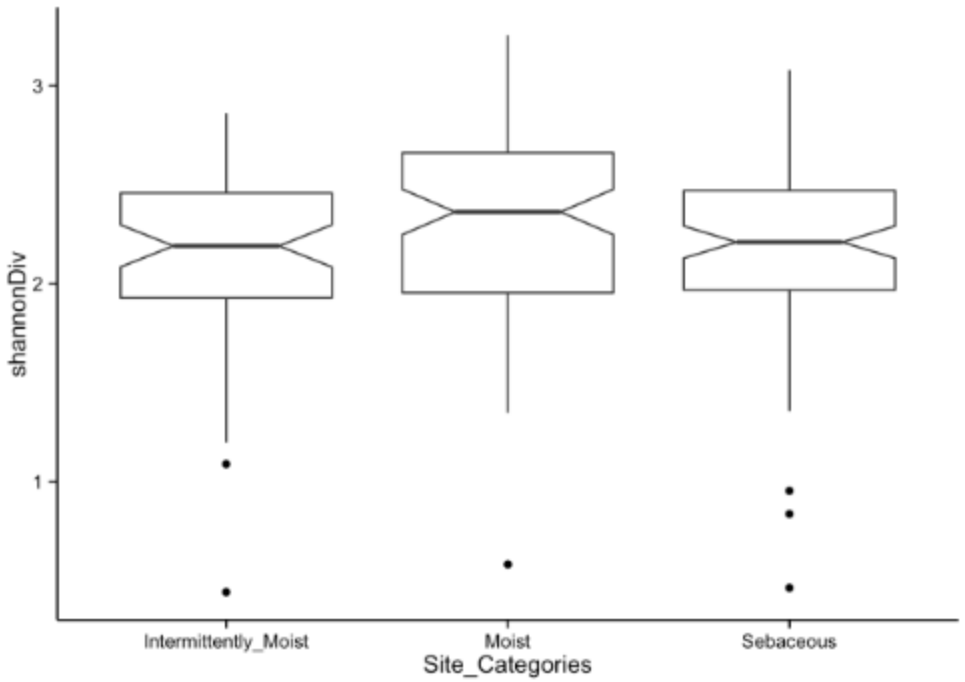
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

36 Also look at the alpha diversity.

Command

```
shannonDiv <- as.data.frame(diversity(inputWideTranspose, index =
```

Expected result



37 Check for significance of differences between the groups. Check levels of categories.

Command

```
alphaAndMap$Site_Categories <- factor(alphaAndMap$Site_Categories)
levels(alphaAndMap$Site_Categories)
```

Expected result

```
## [1] "Intermittently_Moist" "Moist" "Sebaceous"
```

38 Run Kruskal-Wallis on the dataset.

Command

```
kruskalmc(alphaAndMap$shannonDiv, alphaAndMap$Site_Categories)
```

Expected result

Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons

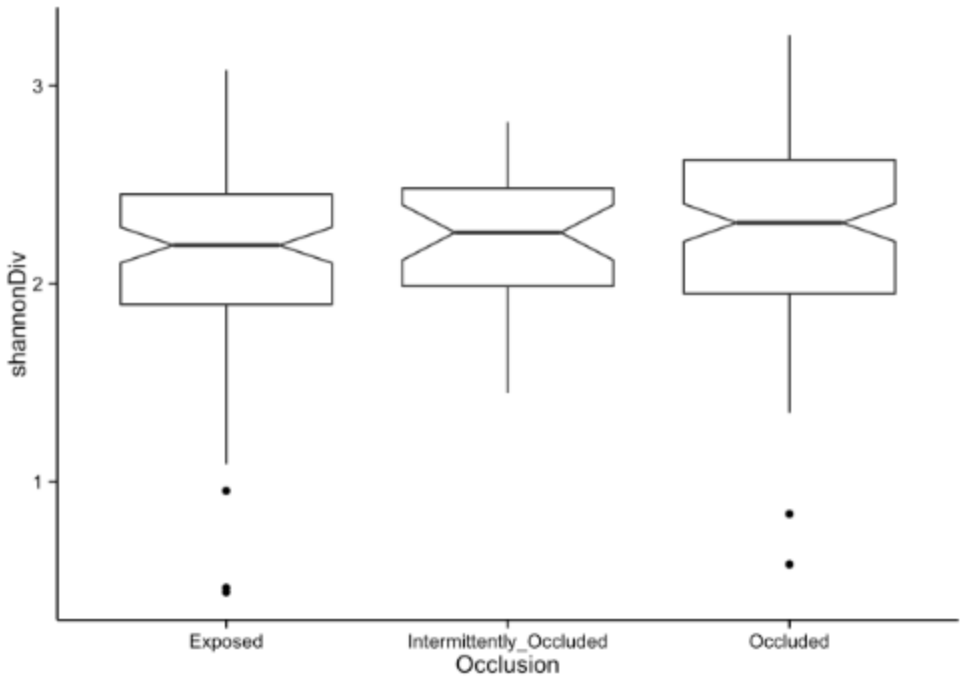
	##		obs.dif		critical.dif		difference
	## Intermittently_Moist-Moist		25.987989		28.54913		FALSE
	Intermittently_Moist-Sebaceous		8.339718		28.43069		FALSE
	Moist-Sebaceous		17.648271		25.32023		FALSE

39 Also look at occlusion status.

Command

```
OcclusionPlot <- ggplot(alphaAndMap, aes(x = Occlusion, y =
shannonDiv)) + theme_classic() + geom_boxplot(notch = TRUE)
OcclusionPlot
```

Expected result



40 Check for significance of differences between groups. Check levels of categories.

Command

```
alphaAndMap$Occlusion <- factor(alphaAndMap$Occlusion)
levels(alphaAndMap$Occlusion)
```

```
## [1] "Exposed" "Intermittently_Occluded" ## [3] "Occluded"
```

41 Run Kruskal-Wallis on the dataset.

Expected result

Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons

	##		obs.dif		critical.dif		difference
	## Exposed-Intermittently_Occluded		13.804754		36.09383		FALSE
	Exposed-Occluded		20.761487		23.71043		FALSE
	Intermittently_Occluded-Occluded		6.959733		34.98437		FALSE