Script R5: Virome Alpha Diversity

HANNIGAN GD, GRICE EA, ET AL.

Abstract

This protocol outlines our alpha diversity analyses of the virome (from PHACCS) and whole metagenome (from MetaPhlan OTU table). We start by comparing the virome and whole metagenome alpha diversity values, and then look at the differences in virome and whole metagenome diversity between skin sites. 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.

Citation: HANNIGAN GD, GRICE EA, ET AL. Script R5: Virome Alpha Diversity. **protocols.io**
dx.doi.org/10.17504/protocols.io.eimbcc6
Published: 10 Mar 2016

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

Protocol

Step 1.

Load the required R packages.

cmd COMMAND

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

library(ggplot2)
packageVersion("ggplot2")

library(pgirmess)
packageVersion("pgirmess")

library(plyr)
packageVersion("plyr")

library(Hmisc)

packageVersion("Hmisc")
```

EXPECTED RESULTS

```
## [1] '2.3.0'



## [1] '1.0.1'



## [1] '1.6.0'



## [1] '1.8.2'


## Loading required package: grid
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:plyr':
##
## is.discrete, summarize
##
## The following objects are masked from 'package:base':
##
## format.pval, round.POSIXt, trunc.POSIXt, units
## [1] '3.16.0'
```

Step 2.

Import the whole microbiome OTU table.

cmd COMMAND

```
INPUT_WM <-
 read.delim("../../IntermediateOutput/Alpha_diversity/skinmet_metaphlan_formatted.tsv", sep
="\t", header=TRUE)
```

Step 3.

Check out a summary of the file to see what it looks like.

cmd COMMAND

```
head(INPUT_WM)[,c(1:6)]
```

EXPECTED RESULTS

```
##    OTU_ID  MG100128  MG100129  MG100130  MG100131  MG100132
## 1 1        0         0         0         0         0
## 2 2        0         0         0         0         0
## 3 3        0         0         0         0         0
## 4 4        0         0         0         0         10535.1
## 5 5        0         0         0         0         0
## 6 6        0         0         0         0         0
```

Step 4.

Import the virome OTU table.

cmd COMMAND

```
INPUT_PHACCS <-
 read.delim("../../IntermediateOutput/Alpha_diversity/PHACCS_results_all.txt", header=TRUE,
 sep="\t")
head(INPUT_PHACCS)
```

EXPECTED RESULTS

```
##    SampleID   SW_Index
## 1 MG100098   11.48738
## 2 MG100100   11.86418
## 3 MG100101   11.29502
## 4 MG100102   12.33014
## 5 MG100104   12.23747
## 6 MG100107   12.16315
```

Step 5.

Import mapping file for whole metagenome and virome.

cmd COMMAND

```
MAP <-
 read.delim("../../IntermediateOutput/Mapping_files/SkinMet_and_Virome_001_metadata.tsv", s
ep="\t", header=TRUE)
head(MAP)[,c(1:6)]
```

EXPECTED RESULTS

```
##     NexteraXT_SampleID          NexteraXT_RunName  NexteraXT_Virome_SampleID
## 1   MG100151                    NexteraXT_007      MG100102
## 2   MG100150                    NexteraXT_007      MG100101
## 3   MG100149                    NexteraXT_007      <NA>
## 4   MG100146                  NexteraXT_007     MG100098
## 5   MG100157                  NexteraXT_007     MG100107
## 6   MG100153                  NexteraXT_007      MG100104
##     NexteraXT_Virome_RunName  SubjectID          TimePoint
## 1   NexteraXT_005             1                  1
## 2   NexteraXT_005             1                 1
## 3   <NA>                      1                 1
## 4   NexteraXT_005             1                 1
## 5   NexteraXT_005             1                 1
## 6   NexteraXT_005                               1
```

Step 6.

While the virome PHACCS diversity is included in the output, MetaPhlan only provides OTU relative abundance information, which means the Shannon diversity must be calculated here with Vegan. We also calculate the median diversity and other information required for graphing. This will all be used for the data visualization.

Step 7.

Here we also need to reformat the mapping files. This means only looking at the two time points for which we have a complete data set, as well as excluding the sites and subjects for which we only have partial sampling.

Step 8.

Transpose the whole microbiome matrix.

cmd COMMAND

```
INPUT_WM_NO_FIRST_COL
```

Step 9.

Calculate alpha diversity for the whole metagenome samples.

cmd COMMAND

```
SHANNON_WM
```

Step 10.

Merge the mapping file info with the alpha diversity information.

cmd COMMAND

```
MERGE_WM
```

Step 11.

Calculate median diversity for each individual anatomical location. For error bar calculation, see the boxplot notching formula implemented in ggplot2:

LINK:

http://www.inside-r.org/packages/cran/ggplot2/docs/geom_boxplot

cmd COMMAND

```
SUMRY_WM_MEDIAN
```

EXPECTED RESULTS

```
##    Site_Symbol_WM  mean_WM    IQR_WM     N_WM  se_WM      mean_plus_WM
## 1 Ac               0.7163576  0.8142767  13    0.3568268  1.0731844
## 2 Ax               0.7493853  0.6449232  13    0.2826138  1.0319992
## 3 Fh               0.2219632  0.4138806  13    0.1813679  0.4033311
## 4 Oc               0.3387557  0.4252576  13    0.1863535  0.5251092
## 5 Pa               1.0617160  1.1723905  13    0.5137569  1.5754730
## 6 Ra                0.3428010   0.3872246   13       0.1696869   0.5124879
```

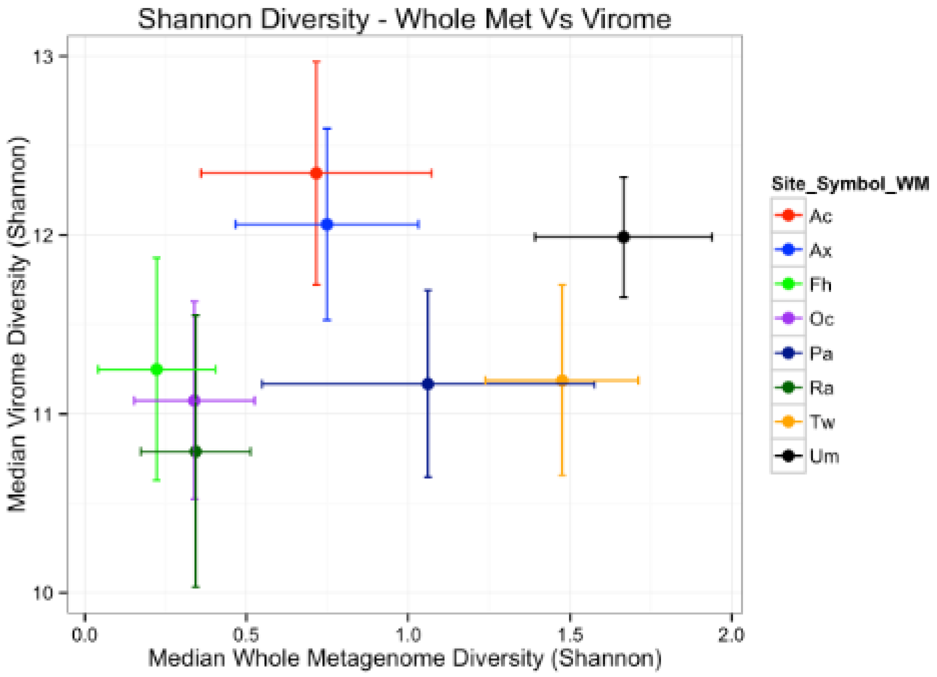
Step 12.

Plot the diversity values as a scatter plot with notch deviation bars.

cmd COMMAND

```
ggplot(SUMRY_MERGE_MEDIAN, aes(x=mean_WM, y=mean_VIR, group=Site_Symbol_WM, colour=Site_Sym
bol_WM, ymax=mean_plus_VIR, ymin=mean_minus_VIR, xmax=mean_plus_WM, xmin=mean_minus_WM)) +
theme_bw() + geom_point(size=3) + geom_errorbar(width=0.025) + geom_errorbarh(height=0.05)
+ scale_colour_manual(values=c("red","blue","green","purple","darkblue","darkgreen","orange
","black")) + ggtitle("Shannon Diversity -
 Whole Met Vs Virome") + xlab("Median Whole Metagenome Diversity (Shannon)") + ylab("Median
 Virome Diversity (Shannon)")
```

EXPECTED RESULTS



Step 13.

We can calculate which samples are significantly different from each other here. This way we can see the statistically significant differences between samples based on bacterial and viral alpha diversity.

cmd COMMAND

```
CUT_LOC_MERGE_VIR
```

EXPECTED RESULTS

```
##      SampleID                 SW_Index    NexteraXT_SampleID   NexteraXT_RunName
## 1    MG100195                 9.134844    MG100171             NexteraXT_008
## 2    MG100199                 8.822409    MG100175             NexteraXT_008
## 3    MG100200                 9.908941    MG100176             NexteraXT_008
## 4    MG100202                 10.327899   MG100178             NexteraXT_008
## 5    MG100204                 10.790570   MG100180             NexteraXT_008
## 6    MG100206                 12.246742    MG100182            NexteraXT_008
##      NexteraXT_Virome_RunName SubjectID
## 1    NexteraXT_009          1
## 2    NexteraXT_009          5
## 3    NexteraXT_009          6
## 4    NexteraXT_009          8
## 5    NexteraXT_009          10
## 6    NexteraXT_009           12
```

Step 14.

Run Kruskal-Wallis on virome dataset.

cmd COMMAND

```
kruskalmc(CUT_LOC_MERGE_VIR$SW_Index, CUT_LOC_MERGE_VIR$Site_Symbol)
```

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
            obs.dif     critical.dif  difference
## Ac-AX    7.210165    52.18821      FALSE
## Ac-Fh    46.603022   52.18821      FALSE
## Ac-Oc    49.900641   54.24178      FALSE
## Ac-Pa    56.049451   52.18821      TRUE
## Ac-Ra    72.081197   52.65148      TRUE
## Ac-Tw    53.299451   52.18821      TRUE
## Ac-Um    2.887960    54.85152      FALSE
## Ax-Fh    39.392857   51.21265      FALSE
## Ax-Oc    42.690476   53.30381      FALSE
## Ax-Pa    48.839286   51.21265      FALSE
## Ax-Ra    64.871032   51.68466       TRUE
## Ax-Tw    46.089286   51.21265       FALSE
## Ax-Um    4.322205    53.92416       FALSE
## Fh-Oc    3.297619    53.30381       FALSE
## Fh-Pa    9.446429    51.21265       FALSE
## Fh-Ra    25.478175   51.68466       FALSE
## Fh-Tw    6.696429    51.21265       FALSE
## Fh-Um    43.715062   53.92416       FALSE
## Oc-Pa    6.148810    53.30381       FALSE
## Oc-Ra    22.180556   53.75747       FALSE
## Oc-Tw    3.398810    53.30381       FALSE
## Oc-Um    47.012681   55.91401       FALSE
## Pa-Ra    16.031746   51.68466       FALSE
## Pa-Tw    2.750000    51.21265       FALSE
## Pa-Um    53.161491   53.92416       FALSE
## Ra-Tw    18.781746   51.68466       FALSE
## Ra-Um    69.193237   54.37264       TRUE
## Tw-Um    50.411491   53.92416       FALSE
```

Step 15.

Run stats on whole metagenome dataset.

cmd COMMAND

```
CUT_LOC_MERGE_WM$Site_Symbol
```

EXPECTED RESULTS

```
##   SampleID                 Shannon_div  NexteraXT_RunName  NexteraXT_Virome_SampleID
## 1 MG100171                 0.0000000    NexteraXT_008      MG100195
## 2 MG100172                 0.4522535    NexteraXT_008      MG100196
## 3 MG100173                 0.4903214    NexteraXT_008      MG100197
## 4 MG100174                 0.2945494    NexteraXT_008      MG100198
## 5 MG100175                 0.1339710    NexteraXT_008      MG100199
## 6 MG100176                 1.3291247    NexteraXT_008      MG100200
##   NexteraXT_Virome_RunName SubjectID
## 1 NexteraXT_009            1
## 2 NexteraXT_009            2
## 3 NexteraXT_009            3
## 4 NexteraXT_009            4
## 5 NexteraXT_009            5
## 6 NexteraXT_009             6
```

Step 16.

Run Kruskal-Wallis on whole metagenome dataset.

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
            obs.dif     critical.dif  difference
## Ac-AX    3.437500    58.72600      FALSE
## Ac-Fh    64.531250   58.72600      TRUE
## Ac-Oc    47.625000   58.72600      FALSE
## Ac-Pa    38.062500   58.72600      FALSE
## Ac-Ra    52.322917   57.07142      FALSE
## Ac-Tw    64.687500   58.72600      TRUE
## Ac-Um   71.687500    58.72600   TRUE
## Ax-Fh   67.968750    58.72600   TRUE
## Ax-Oc   51.062500    58.72600   FALSE
## Ax-Pa   34.625000    58.72600   FALSE
## Ax-Ra   55.760417    57.07142   FALSE
## Ax-Tw   61.250000    58.72600    TRUE
## Ax-Um   68.250000    58.72600    TRUE
## Fh-Oc   16.906250    58.72600    FALSE
## Fh-Pa   102.593750   58.72600    TRUE
## Fh-Ra   12.208333    57.07142    FALSE
## Fh-Tw   129.218750   58.72600    TRUE
## Fh-Um   136.218750   58.72600   TRUE
## Oc-Pa   85.687500    58.72600   TRUE
## Oc-Ra   4.697917     57.07142    FALSE
## Oc-Tw   112.312500   58.72600    TRUE
## Oc-Um   119.312500   58.72600    TRUE
## Pa-Ra   90.385417    57.07142   TRUE
## Pa-Tw   26.625000    58.72600   FALSE
## Pa-Um   33.625000    58.72600    FALSE
## Ra-Tw   117.010417   57.07142    TRUE
## Ra-Um   124.010417   57.07142    TRUE
## Tw-Um   7.000000     58.72600    FALSE
```

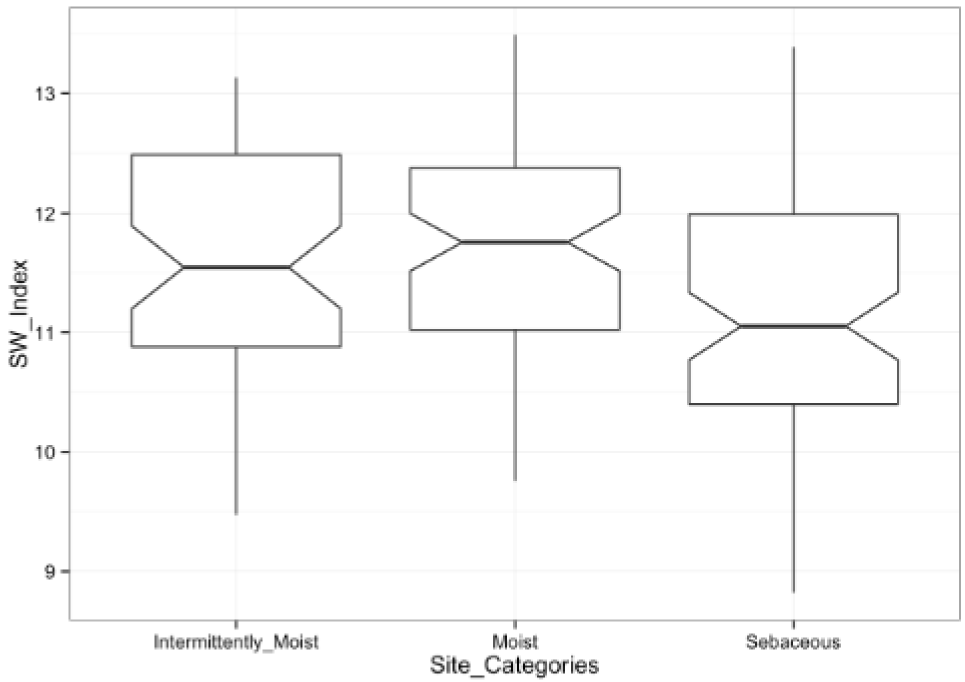
Step 17.

Plot virome diversity by skin location category.

cmd COMMAND

```
CUT_LOC_MERGE_VIR <- MERGE_VIR[-
which(MERGE_VIR$Site_Symbol %in% c("Ba","Ph","Vf","Neg")), ]
CUT_LOC_MERGE_VIR$Type <- "Virome"
ggplot(CUT_LOC_MERGE_VIR, aes(x=Site_Categories, y=SW_Index)) + theme_bw() + geom_boxplot(n
otch=TRUE)
```

EXPECTED RESULTS



Step 18.

Run Kruskal-Wallis on dataset by site category.

cmd COMMAND

```
CUT_LOC_MERGE_VIR$Site_Categories <- factor(CUT_LOC_MERGE_VIR$Site_Categories)
kruskalmc(CUT_LOC_MERGE_VIR$SW_Index, CUT_LOC_MERGE_VIR$Site_Categories)
```

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##                                  obs.dif     critical.dif  difference
## Intermittently_Moist-Moist       6.775434    25.93003      FALSE
## Intermittently_Moist-Sebaceous   27.249883   25.93003      TRUE
## Moist-Sebaceous                  34.025316   23.36625      TRUE
```

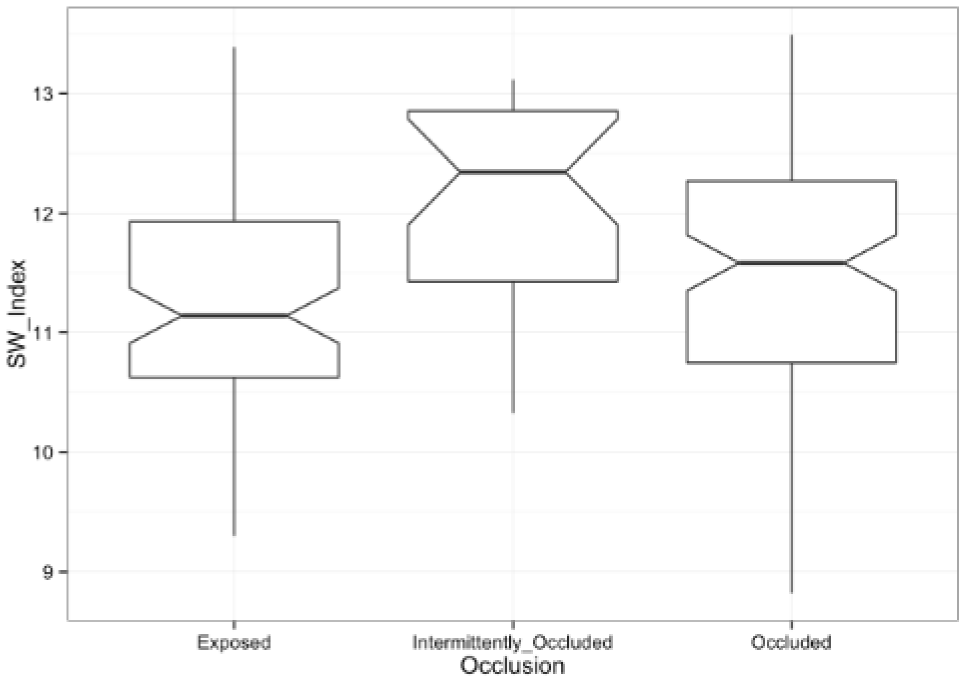
Step 19.

Plot virome diversity by occlusion.

cmd COMMAND

```
ggplot(CUT_LOC_MERGE_VIR, aes(x=Occlusion, y=SW_Index)) + theme_bw() + geom_boxplot(notch=T
RUE)
```

EXPECTED RESULTS



Step 20.

Run Kruskal-Wallis by occlusion on virome dataset.

cmd COMMAND

```
CUT_LOC_MERGE_VIR$Occlusion <- factor(CUT_LOC_MERGE_VIR$Occlusion)
kruskalmc(CUT_LOC_MERGE_VIR$SW_Index, CUT_LOC_MERGE_VIR$Occlusion)
```

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##                                     obs.dif   critical.dif  difference
## Exposed-Intermittently_Occluded     50.89856  33.15193      TRUE
## Exposed_Occluded                    15.92795  21.74934      FALSE
## Intermittently_Occluded-Occluded    34.97061  32.13919      TRUE
```

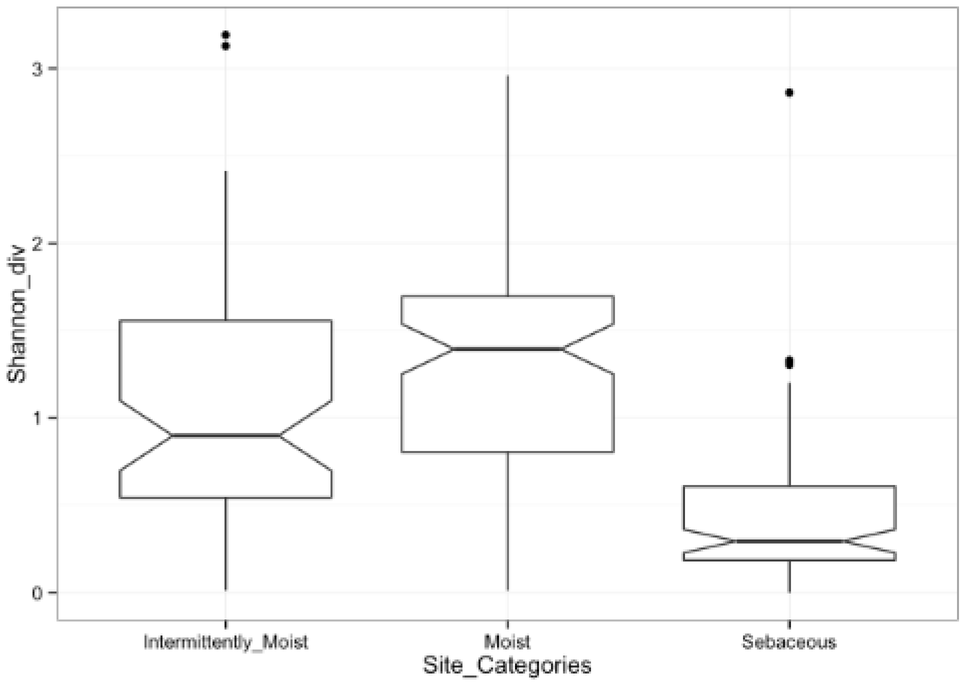
Step 21.

Plot whole metagenome diversity by site category.

cmd COMMAND

```
CUT_LOC_MERGE_WM <- MERGE_WM[-which(MERGE_WM$Site_Symbol %in% c("Ba","Ph","Vf","Neg")), ]
CUT_LOC_MERGE_WM$Type <- "Whole_Metagenome"
CUT_LOC_MERGE_WM$SW_Index <- CUT_LOC_MERGE_WM$Shannon_div
ggplot(CUT_LOC_MERGE_WM, aes(x=Site_Categories, y=Shannon_div)) + theme_bw() + geom_boxplot
(notch=TRUE)
```

EXPECTED RESULTS



Step 22.

Run Kruskal-Wallis on whole metagenome dataset by site category.

cmd COMMAND

```
CUT_LOC_MERGE_WM$Site_Categories <- factor(CUT_LOC_MERGE_WM$Site_Categories)
kruskalmc(CUT_LOC_MERGE_WM$Shannon_div, CUT_LOC_MERGE_WM$Site_Categories)
```

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
 ##                                obs.dif    critical.dif  difference
 ## Intermittently_Moist-Moist     27.57292   29.05167      FALSE
 ## Intermittently_Moist-Sebaceous 73.75750   28.81832      TRUE
 ## Moist-Sebaceous                101.33042  25.72345      TRUE
```

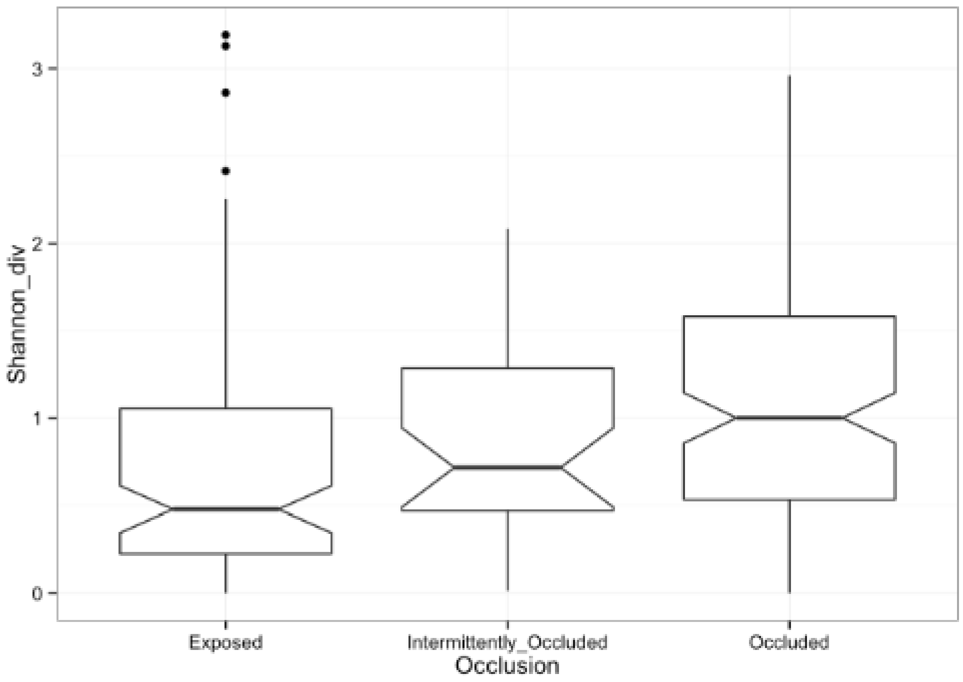
Step 23.

Plot whole metagenome diversity by occlusion.

cmd COMMAND

```
CUT_LOC_MERGE_WM <- MERGE_WM[-which(MERGE_WM$Site_Symbol %in% c("Ba","Ph","Vf","Neg")), ]
CUT_LOC_MERGE_WM$Type <- "Whole_Metagenome"
CUT_LOC_MERGE_WM$SW_Index <- CUT_LOC_MERGE_WM$Shannon_div
ggplot(CUT_LOC_MERGE_WM, aes(x=Occlusion, y=Shannon_div)) + theme_bw() + geom_boxplot(notch
=TRUE)
```

EXPECTED RESULTS



Step 24.

Run Kruskal-Wallis on whole metagenome datset by occlusion.

cmd COMMAND

```
CUT_LOC_MERGE_WM$Occlusion <- factor(CUT_LOC_MERGE_WM$Occlusion)
kruskalmc(CUT_LOC_MERGE_WM$Occlusion, CUT_LOC_MERGE_WM$Occlusion)
```

EXPECTED RESULTS

```
## Multiple comparison test after Kruskal-Wallis
## p.value: 0.05
## Comparisons
##                                     obs.dif  critical.dif  difference
## Exposed-Intermittently_Occluded     64       36.74778      TRUE
## Exposed_Occluded                    146      24.14803      TRUE
## Intermittently_Occluded-Occluded    82       35.47290      TRUE
```