HW52018

Jasmine Nakayama

3/10/2018

Link to repository: https://github.com/jynakay/Assignments (https://github.com/jynakay/Assignments)

Download the dataset from the Canvas folder. Then let's load everything in:

```
# Load all packages
library(phyloseq)
packageVersion("phyloseq")
```

```
## [1] '1.22.3'
```

```
library(ggplot2)
packageVersion("ggplot2")
```

```
## [1] '2.2.1'
```

```
library(RColorBrewer)
packageVersion("RColorBrewer")
```

```
## [1] '1.1.2'
```

```
#Load HMP data
load("HMPv35.RData")
HMPv35
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 45336 taxa and 4743 samples ]
## sample_data() Sample Data:       [ 4743 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 45336 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 45336 tips and 45099 internal nodes ]
## refseq()      DNAStringSet:      [ 45336 reference sequences ]
```

Problem 1

Subset the HMPv35 object to obtain only the samples from Tongue_dorsum. Call this new object HMPv35sub2

```
# Code to subset with new object HMPv35sub2
sub <- get_variable(HMPv35, "HMPbodysubsite") %in% c("Tongue_dorsum")
sample_data(HMPv35)$sub <- factor(sub)
HMPv35sub2 <- prune_samples(sample_data(HMPv35)$sub == TRUE, HMPv35)
summary(sample_data(HMPv35sub2))
```

```
##    X.SampleID             RSID              visitno          sex
##  Min.   :700014409   Min.   :132902142   Min.   :1.000   female:132
##  1st Qu.:700033504   1st Qu.:159586626   1st Qu.:1.000   male  :184
##  Median :700097802   Median :161250552   Median :1.000
##  Mean   :700074079   Mean   :389803522   Mean   :1.415
##  3rd Qu.:700106136   3rd Qu.:763638144   3rd Qu.:2.000
##  Max.   :700114709   Max.   :970836795   Max.   :3.000
##
##     RUNCENTER         HMPbodysubsite Mislabeled      Contaminated
##  WUGC    :103   Tongue_dorsum:316    Mode :logical   Mode :logical
##  BI      : 68                        FALSE:245       FALSE:245
##  JCVI    : 64                        NA's :71        NA's :71
##  BCM     : 44
##  BCM,BI  : 11
##  BCM,JCVI:  7
##  (Other) : 19
##                                                                  Description
##  HMP_Human_metagenome_sample_700014409_from_subject_158398106__sex_male_  :  1
##  HMP_Human_metagenome_sample_700014515_from_subject_158418336__sex_male_  :  1
##  HMP_Human_metagenome_sample_700014609_from_subject_158438567__sex_male_  :  1
##  HMP_Human_metagenome_sample_700014731_from_subject_158458797__sex_female_:  1
##  HMP_Human_metagenome_sample_700014785_from_subject_158479027__sex_male_  :  1
##  HMP_Human_metagenome_sample_700014911_from_subject_158499257__sex_male_  :  1
##  (Other)                                                                  :310
##    sub
##  TRUE:316
##
##
##
##
##
##
```

```
HMPv35sub2
```

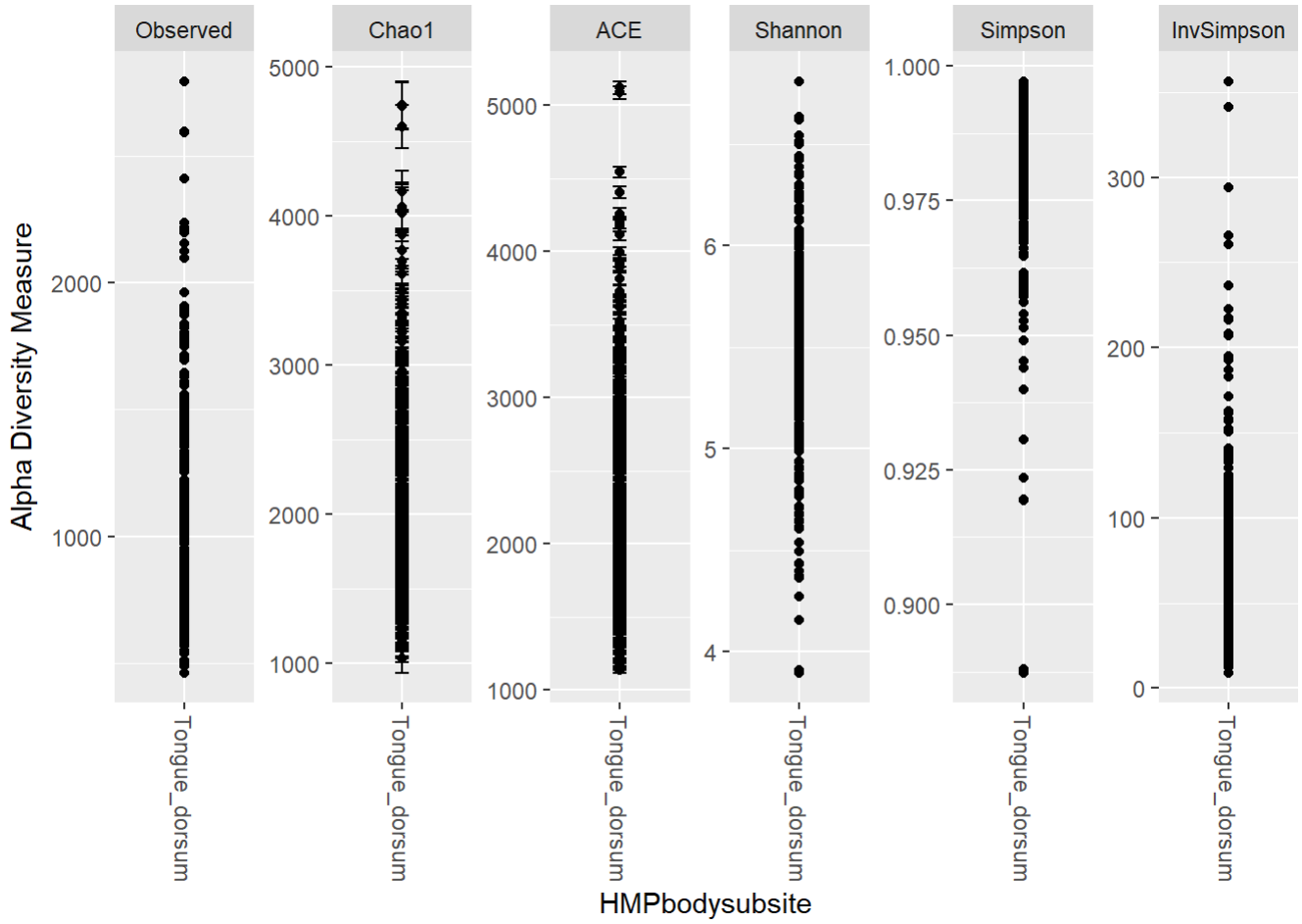
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 45336 taxa and 316 samples ]
## sample_data() Sample Data:       [ 316 samples by 10 sample variables ]
## tax_table()   Taxonomy Table:    [ 45336 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 45336 tips and 45099 internal nodes ]
## refseq()      DNAStringSet:      [ 45336 reference sequences ]
```

Problem 2

Produce the geometric box plot of diversity measures for your object, HMPv35sub2

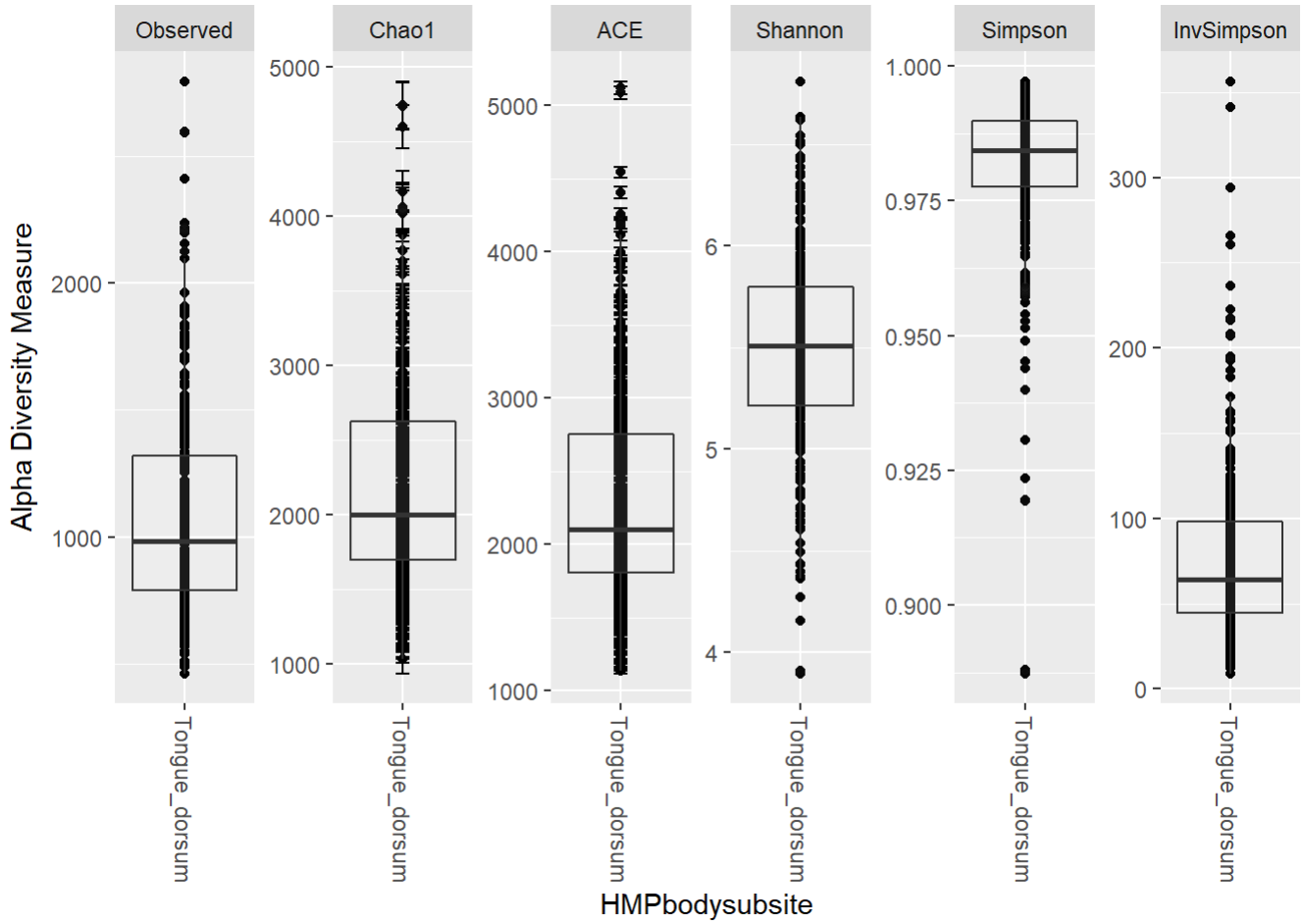
```
#box plot for diversity here
alpha_meas = c("Observed", "Chao1", "ACE", "Shannon", "Simpson", "InvSimpson")
(p <- plot_richness(HMPv35sub2, "HMPbodysubsite", measures=alpha_meas))
```

```
## Warning: Removed 1264 rows containing missing values (geom_errorbar).
```




```
p + geom_boxplot(data=p$data, aes(x=HMPbodysubsite, y=value, color=NULL), alpha=0.1)
```

```
## Warning: Removed 1264 rows containing missing values (geom_errorbar).
```



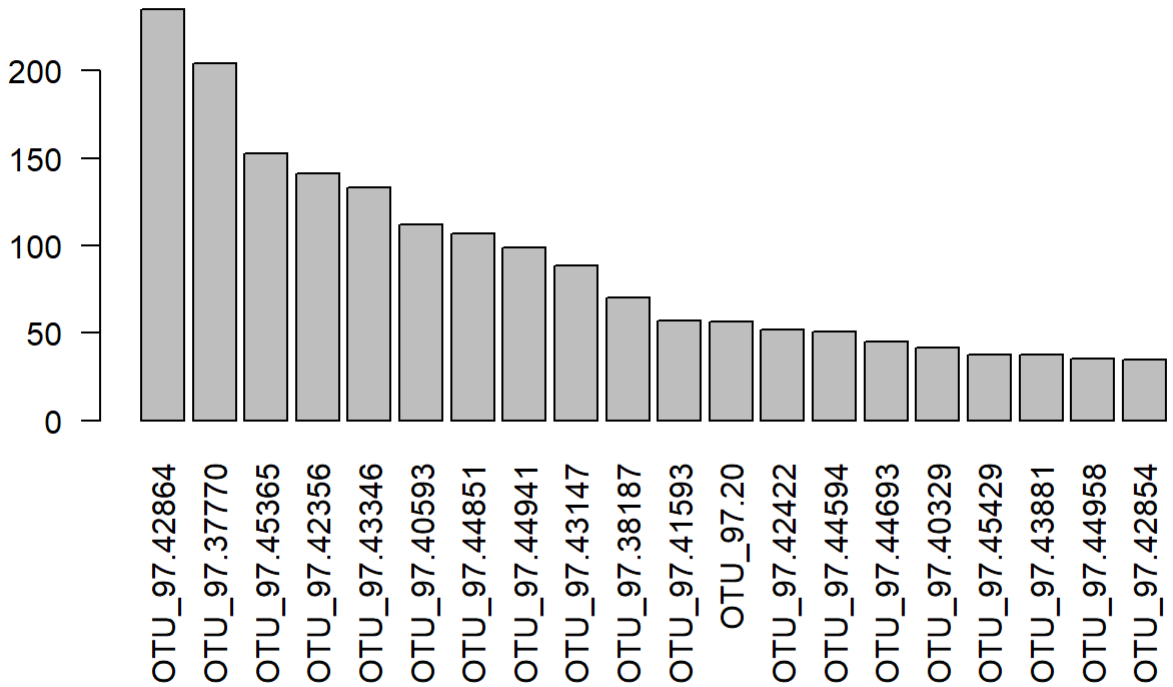
Problem 3

You see what taxa are most prevalent in your subset, HMPv35sub2

```
#Calculate number of taxa in object
ntaxa(HMPv35sub2)
```

```
## [1] 45336
```

```
par(mar = c(10, 4, 4, 2) + 0.1) # make more room on bottom margin
N <- 20
barplot(sort(taxa_sums(HMPv35sub2), TRUE)[1:N]/nsamples(HMPv35sub2), las=2)
```



Problem 4

Using your HMPv35sub2 object, throw the rare taxa out of that object, then reduce to only taxa in the phylum Bacteroidetes. Call this new object HMPv35sub2frbac

```
#throw out the rare taxa from the HMPv35sub2 object with new object HMPv35sub2frbac

#The next step filters out taxa with low occurrence throughout all samples
HMPv35subsub = filter_taxa(HMPv35sub2, function(x) sum(x > 3) > (0.2*length(x)), TRUE)

# The next step filters out all taxa that occur in less than .01% of samples
HMPv35subr  <- transform_sample_counts(HMPv35subsub, function(x) x / sum(x) )
HMPv35subfr <- filter_taxa(HMPv35subr, function(x) mean(x) > 1e-5, TRUE)
HMPv35subfr
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 307 taxa and 316 samples ]
## sample_data() Sample Data:       [ 316 samples by 10 sample variables ]
## tax_table()   Taxonomy Table:    [ 307 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 307 tips and 304 internal nodes ]
## refseq()      DNAStringSet:      [ 307 reference sequences ]
```

```
#Finally subset to only bacteria in the phylum Bacteroidetes
HMPv35sub2frbac = subset_taxa(HMPv35subfr, Phylum=="Bacteroidetes")
HMPv35sub2frbac
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 46 taxa and 316 samples ]
## sample_data() Sample Data:       [ 316 samples by 10 sample variables ]
## tax_table()   Taxonomy Table:    [ 46 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 46 tips and 45 internal nodes ]
## refseq()      DNAStringSet:      [ 46 reference sequences ]
```

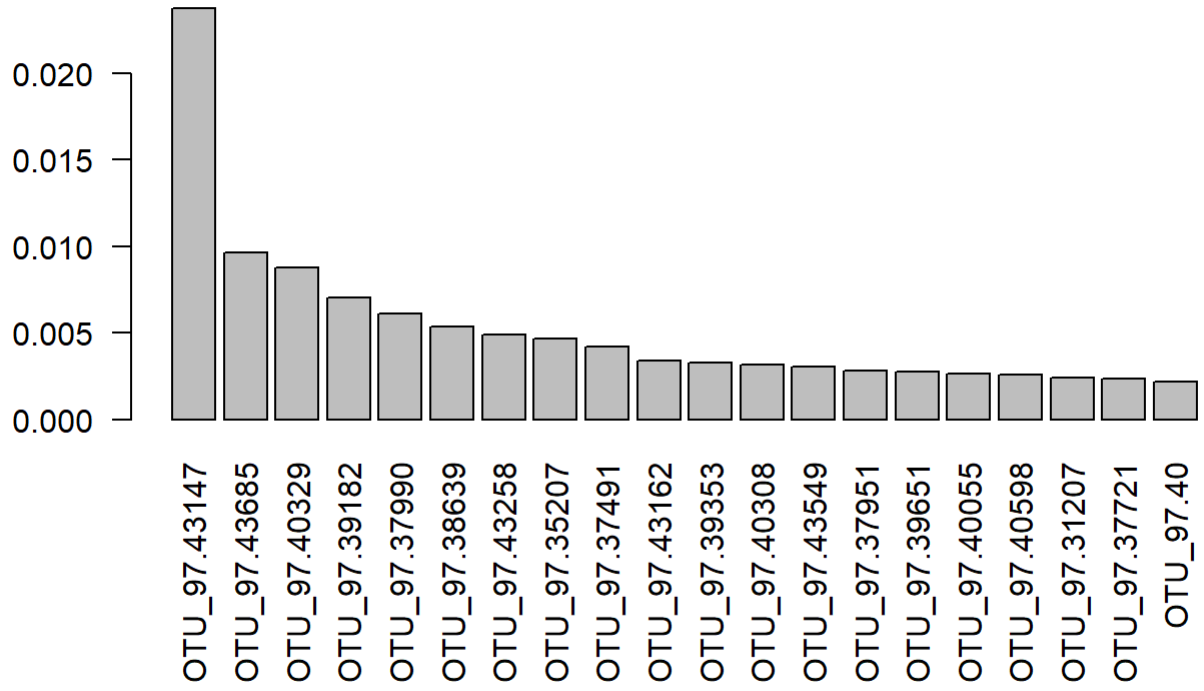
Problem 5

Using your HMPv35sub2frbac object, what is the distribution of the top 20 OTU's?

```
#Calculate number of taxa in object
ntaxa(HMPv35sub2frbac)
```

```
## [1] 46
```

```
par(mar = c(10, 4, 4, 2) + 0.1) # make more room on bottom margin
N <- 20
barplot(sort(taxa_sums(HMPv35sub2frbac), TRUE)[1:N]/nsamples(HMPv35sub2frbac), las=2)
```



Problem 6

Plot sample abundance by body site for your object HMPv35sub2frbac

```
#Pcode for the plot
title = "plot_bar; by site; Bacteroidetes only"
plot_bar(HMPv35sub2frbac, "HMPbodysubsite", "Abundance", title=title)
```

Problem 7

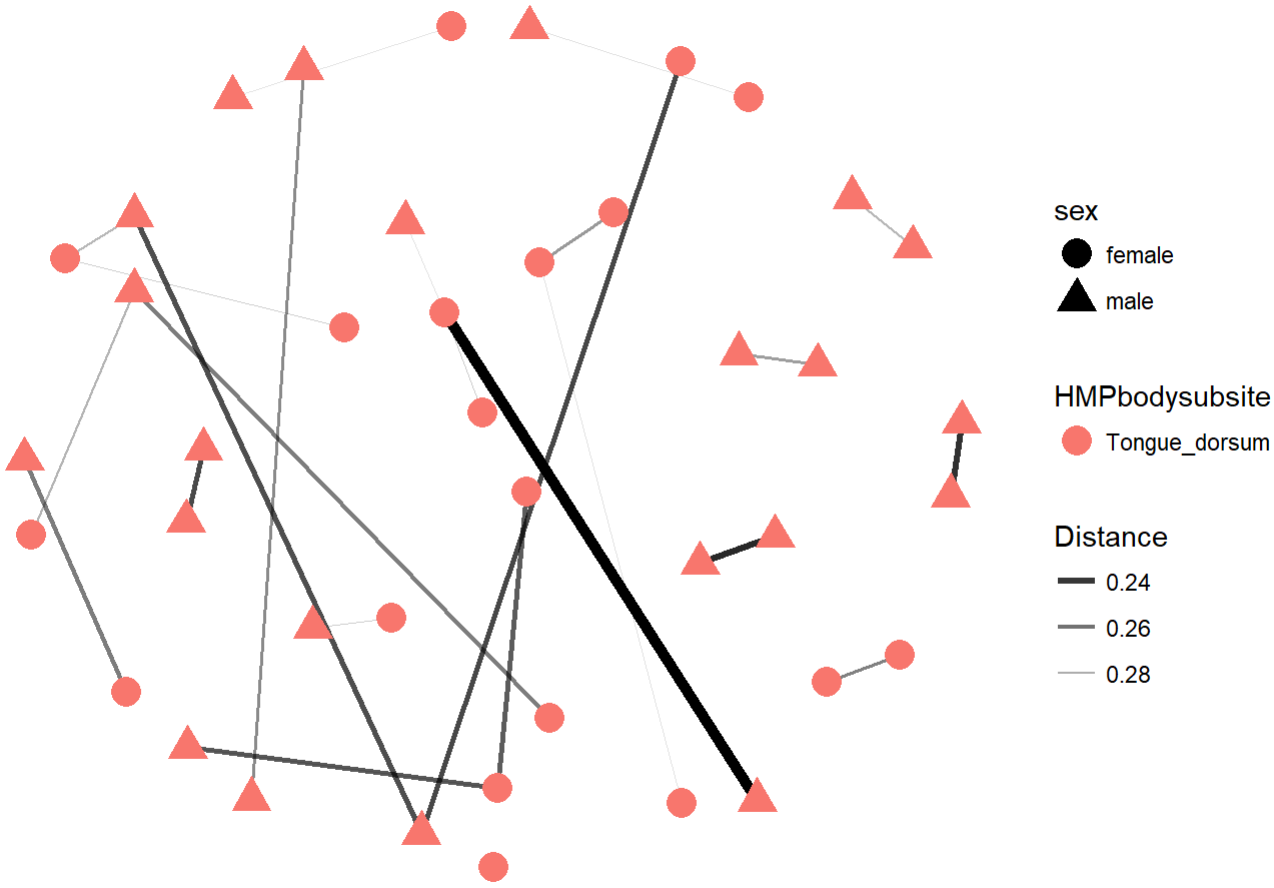
You try it with your HMPv35sub2frbac object

```
#code for your plot with family here
plot_bar(HMPv35sub2frbac, "HMPbodysubsite", "Abundance", "Phylum", title=title)
```

We can use the techniques of network science to illustrate how similar or distant samples are.

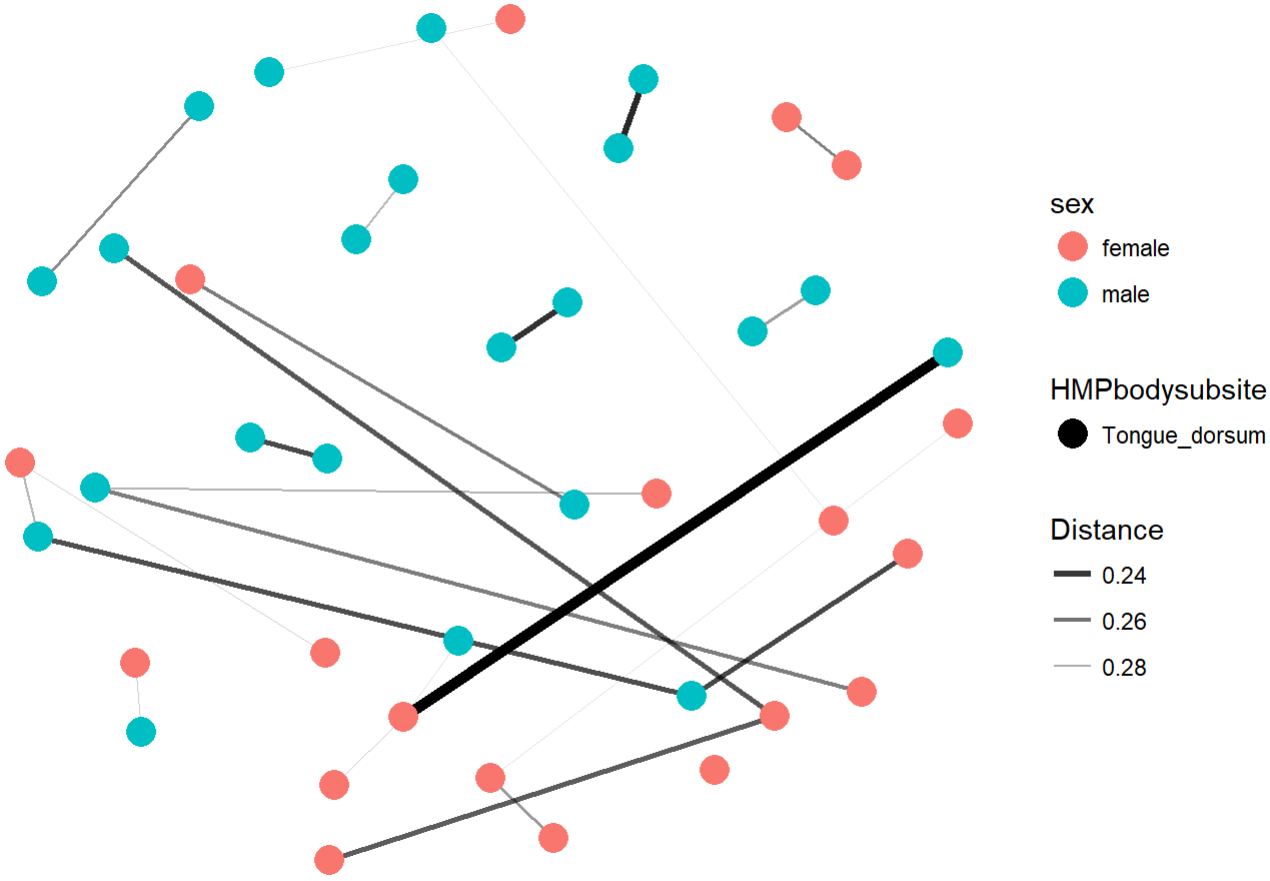
Using our filtered object with all phyla, we use the `plot_net` function to illustrate the “proximity” (or similarity) of samples, while denoting site and the sex of the participant.

```
plot_net(HMPv35subfr, maxdist = 0.3, color = "HMPbodysubsite", shape="sex")
```



Let's redo switching which variable is colored and which variable is differentiated by shape.

```
plot_net(HMPv35subfr, maxdist = 0.3, shape = "HMPbodysubsite", color="sex")
```



####Problem 8

Which method of display do you like best and why?

```
#Place your answer here as another comment.
#I like the second graph more, because it is easier to distinguish sex by color rather than by s
hape. There is higher contrast between the differences.
```