Supplemental Information: VEGAN based analysis of community composition

Simpson, Bettauer et al.

May 2020

```
> options(warn = -1)
> # Load the tree data.frame and genome size fits.
> root <- rprojroot::find_root(".git/index");  source(file.path(root, "src/init.R"))
> source(file.path(root, "experiments/exp5-vegan/local.R"))
> figurefile <- file.path(mainfigurefile, "vegan")
> #install.packages("extraDistr")
```

Rarefy for the whole tree

```
> library(vegan)
> genera <- which(tree$rank == "genus")
> species <- which(tree$rank == "species")
> counts <- t(tree[species, c("br_bel", "br_may")])
> colnames(counts) <- tree[species, "name"]
> rownames(counts) <- c("Bellairs", "Maycocks")
> adj <- rarefy( counts, min(rowSums(counts)))
> prestondistr( counts[1,]) %>% veiledspec
```

```
Extrapolated     Observed       Veiled
9.089028e+03 9.089000e+03 2.779913e-02
```

```
> prestondistr( counts[2,]) %>% veiledspec
```

```
Extrapolated     Observed       Veiled
9.461097e+03 9.461000e+03 9.671192e-02
```

So no change.

```
> df1 <- data.frame(counts = counts[1,], site = 'Bellairs')
> df2 <- data.frame(counts = counts[2,], site = 'Maycocks')
> df <- rbind(df1, df2)
> g <- ggplot(df, aes(x=counts, fill=site, color=site)) +
+  # scale_x_log10() +
+   scale_x_continuous(trans='log2') +
+   geom_histogram(aes(y=..density..), position="identity", alpha=0.1)+
+   geom_density(alpha=0.6)+
+
+   scale_color_manual(values=c("red", "blue", "#56B4E9"))+
+   scale_fill_manual(values=c("darkred", "lightblue", "#56B4E9"))+
+   labs(title="Number of reads at each site across all species",x="Number Reads (log2)", y = "Density")+
+   theme_classic()
> g
> g <- ggplot(df, aes(x=counts, fill=site, color=site)) +
+  # scale_x_log10() +
+   scale_x_continuous(trans='log2') +
+   geom_histogram(alpha=0.1, binwidth=(1/(2^3)))+
+ #  geom_density(alpha=0.6)+
```

```
+   scale_color_manual(values=c("red", "blue", "#56B4E9"))+
+   scale_fill_manual(values=c("darkred", "lightblue", "#56B4E9"))+
+   labs(title="Number of reads at each site across all species",x="Number Reads (log2)", y = "Counts")+
+   theme_classic()
> g
```

Our next goal is to undersample from the observed counts to see the effect on richness at both sites.

```
> library(extraDistr)
> br_tot <- sum(tree$bell_orig_est_reads[species]); may_tot <-  sum(tree$may_orig_est_reads[species])
> br <- tree$bell_orig_est_reads[species]
> may <- tree$may_orig_est_reads[species]
> res <- data.frame(); idx <- 1
> for (i in seq(from=0.01, to = 1, by= 0.01 )) {
+
+   tmp_br <- rmvhyper( nn = 10, n = br,   k = floor(br_tot * i))
+   num_taxa <- apply(tmp_br, 1, FUN=function(x) {return(length(which(x > 0)))})
+   res <- rbind(res,  data.frame(site = "Bellairs", percent = i,
+                              tot_reads = floor(i * br_tot),
+                              num_taxa  = num_taxa ))
+
+   tmp_may <- rmvhyper( nn = 10, n = may,   k = floor(may_tot * i))
+   num_taxa <- apply(tmp_may, 1, FUN=function(x) {return(length(which(x > 0)))})
+   res <- rbind(res,  data.frame(site = "Maycocks", percent = i,
+                            tot_reads = floor(i * may_tot),
+                            num_taxa  = num_taxa ))
+
+ }
> to_remove <- intersect(which(res$site == "Maycocks"), which(res$percent  > .5))
> res <- res[-to_remove, ]
> bmax <- res[ which(res$site == "Bellairs"), ]; bbmax <- res[which.max(bmax$num_taxa), "num_taxa"]
> mmax <- res[ which(res$site == "Maycocks"), ]; mmmax <- res[which.max(mmax$num_taxa), "num_taxa"]
> ggplot(res, aes(x=tot_reads, y=num_taxa, color=site)) +
+   geom_point(shape=3) +    # Use hollow circles
+   labs(title="Number of taxa identified as a function of read coverage",x="Number of Reads: downsampling
+   geom_hline(yintercept = 9500, color = "blue") +
+   geom_text(aes(y=9500, label="\n9461 taxa", x=110000),
+             colour="blue", angle=0, size=4) +
+   geom_hline(yintercept = 9120, color = "blue") +
+   geom_text(aes(y=9120, label="\n9089 taxa", x=110000),
+             colour="blue", angle=0, size=4) +
+   geom_vline(xintercept=4486003, color = "blue") +
+   geom_text(aes(x=4486003, label="\ntruncated at 4.48M reads", y=8000), colour="blue", angle=90, size=5)
+   theme_tufte()
```

Check the genera level just in case.

```
> library(extraDistr)
> br_tot <- sum(tree$br_bel[genera]); may_tot <-  sum(tree$br_may[genera])
> br <- tree$br_bel[genera]
> may <- tree$br_may[genera]
> res <- data.frame(); idx <- 1
> for (i in seq(from=0.01, to = 1, by= 0.01 )) {
+
+   tmp_br <- rmvhyper( nn = 10, n = br,   k = floor(br_tot * i))
+   num_taxa <- apply(tmp_br, 1, FUN=function(x) {return(length(which(x > 0)))})
+   res <- rbind(res,  data.frame(site = "Bellairs", percent = i,
+                              tot_reads = floor(i * br_tot),
```

```
+                            num_taxa  = num_taxa ))
+
+   tmp_may <- rmvhyper( nn = 10, n = may,   k = floor(may_tot * i))
+   num_taxa <- apply(tmp_may, 1, FUN=function(x) {return(length(which(x > 0)))})
+   res <- rbind(res,  data.frame(site = "Maycocks", percent = i,
+                            tot_reads = floor(i * may_tot),
+                            num_taxa  = num_taxa ))
+ }
> to_remove <- intersect(which(res$site == "Maycocks"), which(res$percent  > .5))
> res <- res[-to_remove, ]
> bmax <- res[ which(res$site == "Bellairs"), ]; bbmax <- res[which.max(bmax$num_taxa), "num_taxa"]
> mmax <- res[ which(res$site == "Maycocks"), ]; mmmax <- res[which.max(mmax$num_taxa), "num_taxa"]
> ggplot(res, aes(x=tot_reads, y=num_taxa, color=site)) +
+   geom_point(shape=3) +    # Use hollow circles
+   labs(title="Number of genera identified as a function of read coverage",x="Number of Reads: downsampli
+   geom_hline(yintercept = 2520, color = "blue") +
+   geom_text(aes(y=2520, label="\n2479 taxa", x=110000),
+             colour="blue", angle=0, size=4) +
+   geom_hline(yintercept = 2430, color = "blue") +
+   geom_text(aes(y=2440, label="\n2415taxa", x=110000),
+             colour="blue", angle=0, size=4) +
+   geom_vline(xintercept=3400000, color = "blue") +
+   geom_text(aes(x=3400000, label="\ntruncated at 3.40M reads", y=2200), colour="blue", angle=90, size=5)
+   theme_tufte()
```