# Supplemental File 2: Assessment of vaginal microbiota composition with clustering and ordination analysis

*Roxana J. Hickey*

## Contents

## Description

This is a supplement to the paper “Vaginal microbiota of adolescent girls resemble those of reproductive-age women prior to the onset of menarche” by Hickey et al. Please refer to the paper for complete information about the objectives and study design.

The embedded R code works through the first set of analyses and generation of figures related to the assessment of vaginal microbiota composition in girls and mothers. The analyses can be run directly in RStudio from the R Markdown file “adolescent-supp-02.Rmd” located at https://github.com/roxanahickey/adolescent. It should be run after “adolescent-supp-01.Rmd”, which prepares the data and color palettes used in this script.

### Objective

The first major objective of this study is to characterize the composition of vaginal microbiota in girls both premenarche and postmenarche as well as mothers in this study. To do this we will first perform hierarchical clustering analysis of the vaginal microbiota from girls and mothers to determine what the major types of communities are and how they are distributed across the sample groups we are interested in. Then we will perform principal coordinates analysis to get a different perspective of the similarity among samples in relation to other variables of interest.

---

## Initial setup

After running adolescent-supp-01.Rmd, two RData files named "01-data-prep-[date].RData" and "01-data-prep-last-run.RData" are saved in the data-postproc directory (two files are written so as to preserve older versions if necessary). Load either file to get all of the data necessary to run the analyses and generate figures below.

*Note: If you run the R Markdown script 'as is' from the same directory containing it and the 'data-input' and 'data-postproc' subdirectories, all figures will be printed inside the resulting PDF or HTML output. If you want to save the figures as individual files, uncomment the lines below starting with 'dir.create()' as well as any lines throughout the script starting with 'ggsave()' or 'pdf()'. I made note of each of these within the chunk code.*

```
## Clear current workspace
rm(list=ls())

## Load the RData file created from adolescent-supp-01.Rmd
load("data-postproc/01-data-prep-last-run.RData")

## Load packages
library(ape)
library(cluster)
library(gclus)
library(ggplot2)
library(gplots)
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(grid)
library(plyr)
library(reshape)
```

```
##
```

```
## Attaching package: 'reshape'
##
## The following objects are masked from 'package:plyr':
##
##     rename, round_any
```

```
library(scatterplot3d)
library(vegan)
```

```
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.0-10
```

```
## Uncomment the next two lines to create directories for figure output
# dir.create("hclust-analysis")
# dir.create("pcoa-analysis")
```

## Subset data

The dataframes from the first step contain both vagina and vulva samples. Since we are only interested in the vagina samples at this step, we first subset the data for ease of use. Below we make new phylotype/taxon abundance and proportion tables from the 'abund.red' and 'prop.red' tables, a new metadata table from the 'meta' dataframe, and a reduced color palette table from 'col.meta'.

```
## Subset abund and prop tables to include only vagina samples
spe.abund.vag <- t(abund.red[,meta$site=="vag"])
spe.prop.vag <- t(prop.red[,meta$site=="vag"])

## Subset the metadata table
meta.vag <- subset(meta, site=="vag")

## Subset the metadata color palette
col.meta.vag <- col.meta[meta$site=="vag",]
```

## Hellinger standardization of taxon abundance data

Next, we standardize the taxon abundances using the Hellinger method. This is a recommended approach when the "species" are sparsely populated at some sites, resulting in many zeros in the species abundance matrix. We then compute the Bray-Curtis dissimilarity matrix from the Hellinger-standardized abundance matrix. This will be used in subsequent clustering and ordination analyses. It does not matter whether you apply the Hellinger transformation to the count (spe.abund.vag) or proportion (spe.prop.vag) table; they will produce the same result.

```
## Perform the Hellinger transformation
spe.vag.hel <- decostand(spe.abund.vag, method="hellinger")

## Compute the Bray-Curtis dissimilarity matrix
spe.vag.hel.bc <- vegdist(spe.vag.hel, method="bray")
```

# Part I: Perform hierarchical clustering

## Hierarchical clustering analysis

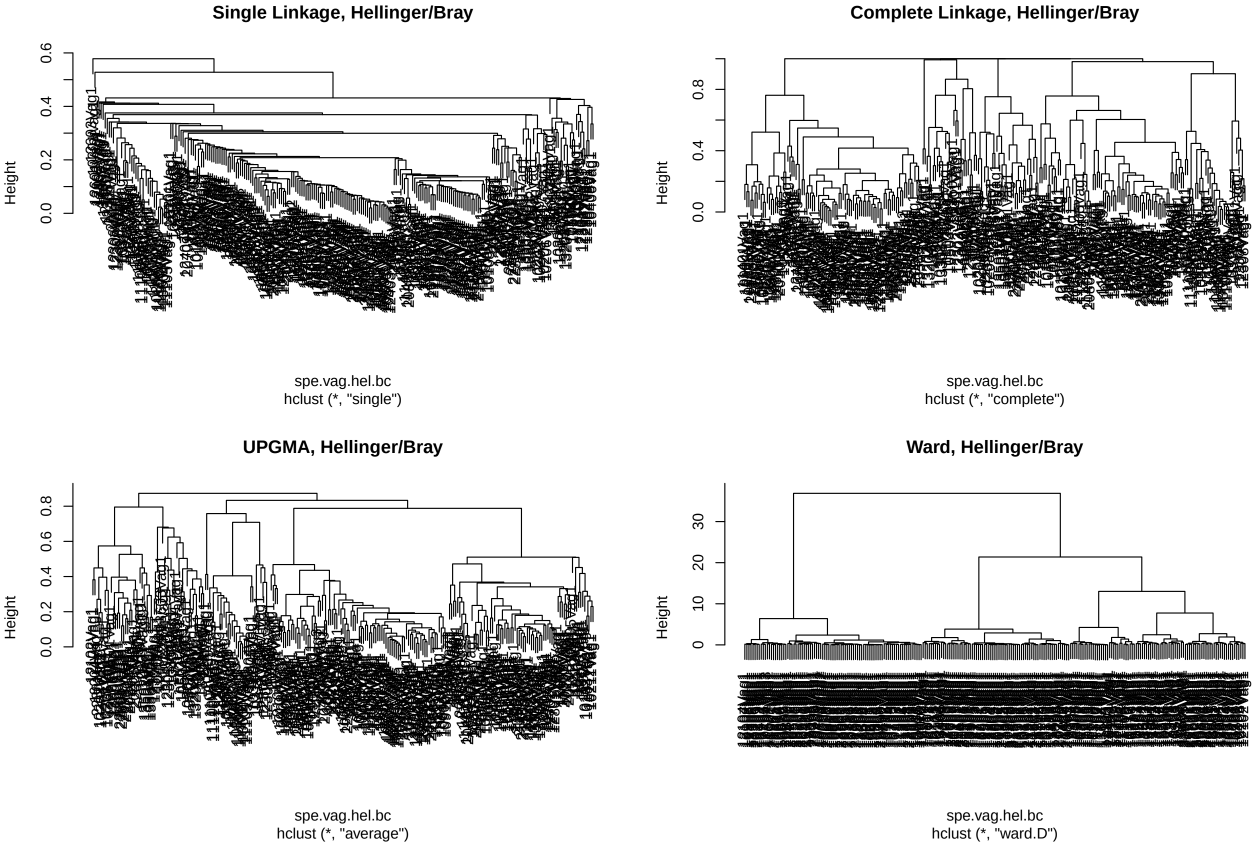
The first set of analyses involves clustering the samples based on community composition and selecting the optimal clustering model and number of clusters. The approaches are based on those outlined in the following texts:

- Legendre P, Legendre L. (2012). *Cluster analysis.* 3rd ed. Elsevier.
- Borcard D, Gillet F, Legendre P. (2011). *Numerical ecology with R.* Springer.

*Note: vegan and other ecological software commonly refer to "sites" and "species". In our case, "sites" refers to the individual samples (vaginal and vulvar swabs) while "species" refers to the taxa present in each sample (our "species" are actually taxa at multiple taxonomic levels, mostly genus).*

We start by performing hierarchical clustering using multiple linkage strategies (single, complete, average/UPGMA, Ward) and select the optimal one using Gower's distance. After that, we select the optimal number of clusters using the maximum silhouette width. First, compute the clusters and look at the dendrograms:

```r
## Compute hierarchical clustering using four linkage methods
spe.vag.hb.single <- hclust(spe.vag.hel.bc, method="single")
spe.vag.hb.complete <- hclust(spe.vag.hel.bc, method="complete")
spe.vag.hb.upgma <- hclust(spe.vag.hel.bc, method="average")
spe.vag.hb.ward <- hclust(spe.vag.hel.bc, method="ward.D")

## Plot to compare (these are ugly because the labels obscure each another,
## but it's useful to take a look at the shape of the dendrograms)
par(mfrow=c(2,2))
plot(spe.vag.hb.single, main="Single Linkage, Hellinger/Bray")
plot(spe.vag.hb.complete, main="Complete Linkage, Hellinger/Bray")
plot(spe.vag.hb.upgma, main="UPGMA, Hellinger/Bray")
plot(spe.vag.hb.ward, main="Ward, Hellinger/Bray")
```



```r
dev.off()
```

```
## null device
##           1
```

## Selection of optimal clustering model

Now we select the best clustering method by determining the cophenetic distance of each hierarchical clustering, followed by calculation of the Gower distance (Gower 1983), which is the sum of squared differences between the original and cophenetic distances. The method with the lowest Gower distance is considered the optimal clustering model for the distance matrix used. Below, this method identifies average/UPGMA is the best clustering model.

```r
## Calculate cophenetic distance for each hclust object
spe.vag.hb.single.coph <- cophenetic(spe.vag.hb.single)
spe.vag.hb.complete.coph <- cophenetic(spe.vag.hb.complete)
spe.vag.hb.upgma.coph <- cophenetic(spe.vag.hb.upgma)
spe.vag.hb.ward.coph <- cophenetic(spe.vag.hb.ward)

## Calculate the Gower distance
gow.vag.dist.single <- sum((spe.vag.hel.bc-spe.vag.hb.single.coph)^2)
gow.vag.dist.complete <- sum((spe.vag.hel.bc-spe.vag.hb.complete.coph)^2)
gow.vag.dist.upgma <- sum((spe.vag.hel.bc-spe.vag.hb.upgma.coph)^2)
gow.vag.dist.ward <- sum((spe.vag.hel.bc-spe.vag.hb.ward.coph)^2)
```

```
## Compare Gower distances and identify the lowest
gow.vag.dist.single
```

```
## [1] 6384
```

```
gow.vag.dist.complete
```

```
## [1] 1355
```

```
gow.vag.dist.upgma
```

```
## [1] 427.5
```

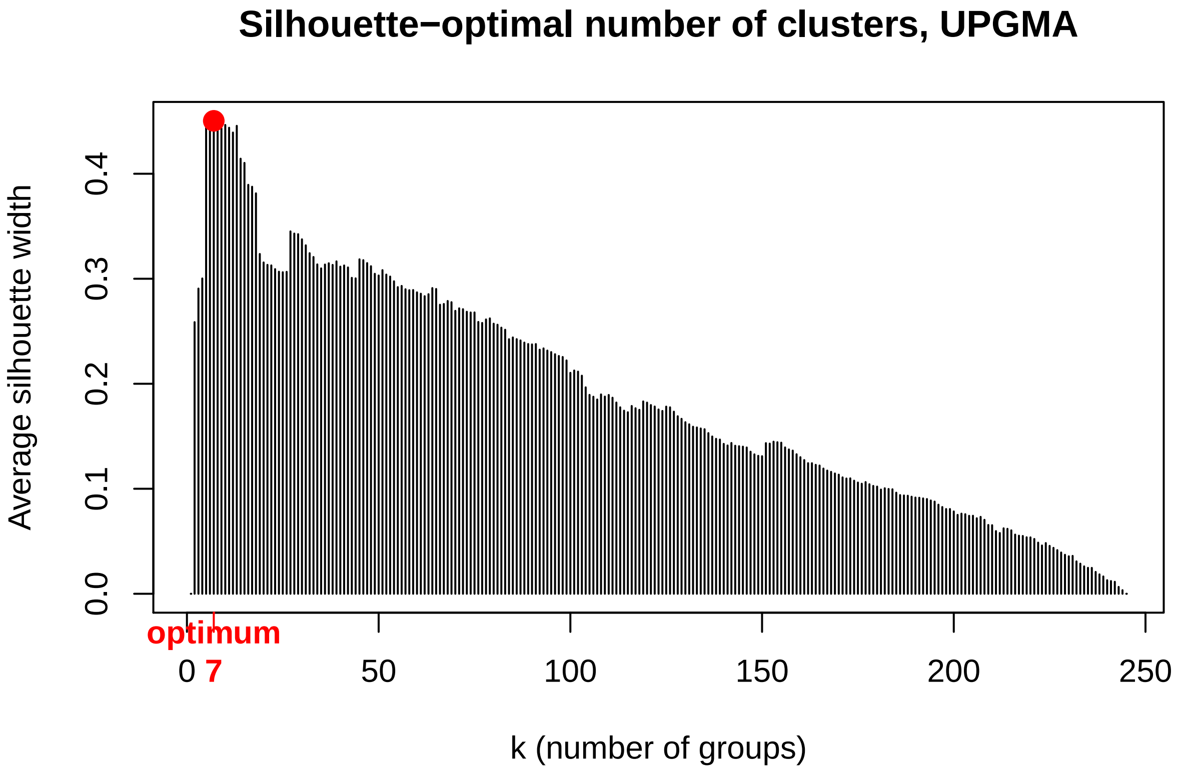
```
gow.vag.dist.ward
```

```
## [1] 20826257
```

### Selection of optimal number of clusters

Now we pick the optimal number of clusters according silhouette widths (Rousseew quality index). To do this we plot the average silhouette widths for all partitions except for the trivial partition in a single group (k=1). Below, we find that seven clusters are identified as optimal.

```
## Create an empty vector for the average silhouette width values
asw <- numeric(nrow(spe.abund.vag))

## This function calculates and plots the silhouette width, indicating the
## optimal number in red (function from Borcard et al. 2011)
for (k in 2:(nrow(spe.abund.vag)-1)) {
  sil <- silhouette(cutree(spe.vag.hb.upgma, k=k), spe.vag.hel.bc)
  asw[k] <- summary(sil)$avg.width
}

k.best <- which.max(asw)

plot(1:nrow(spe.abund.vag), asw, type="h",
     main="Silhouette-optimal number of clusters, UPGMA",
     xlab="k (number of groups)", ylab="Average silhouette width")
axis(1, k.best, paste("optimum",k.best,sep="\n"), col="red", font=2, col.axis="red")
points(k.best, max(asw), pch=16, col="red", cex=1.5)
```

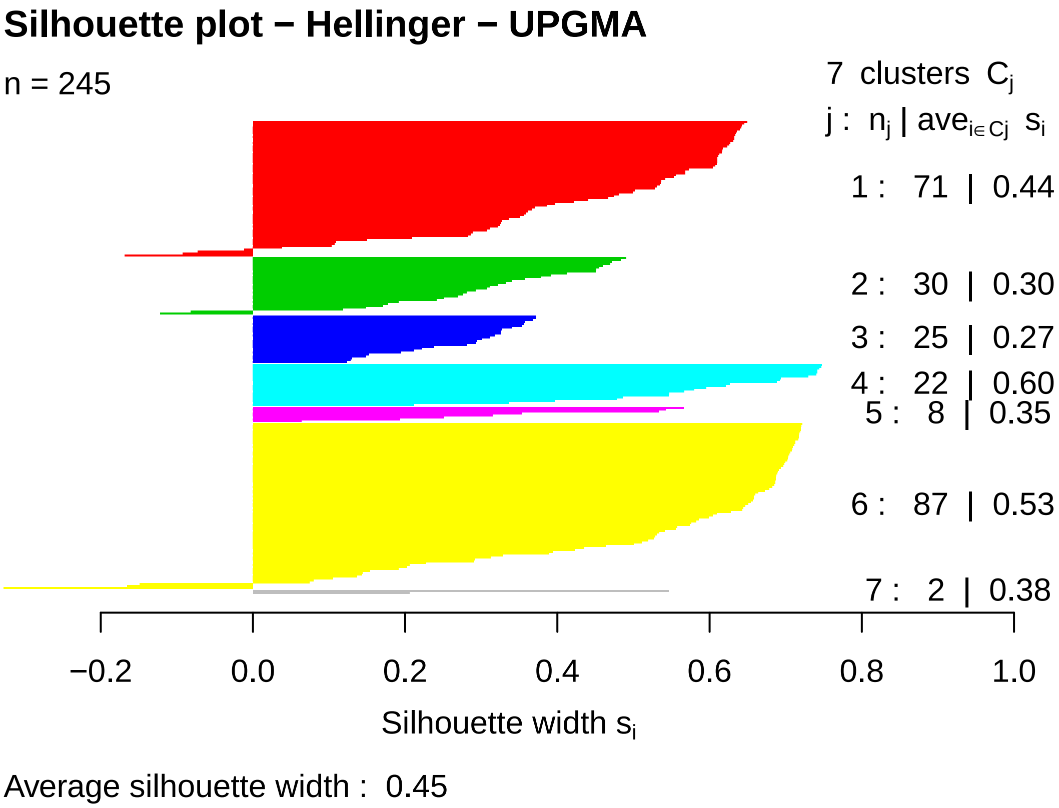


```
cat("", "Silhouette-optimal number of clusters k =", k.best, "\n",
    "with an average silhouette width of", max(asw), "\n")
```

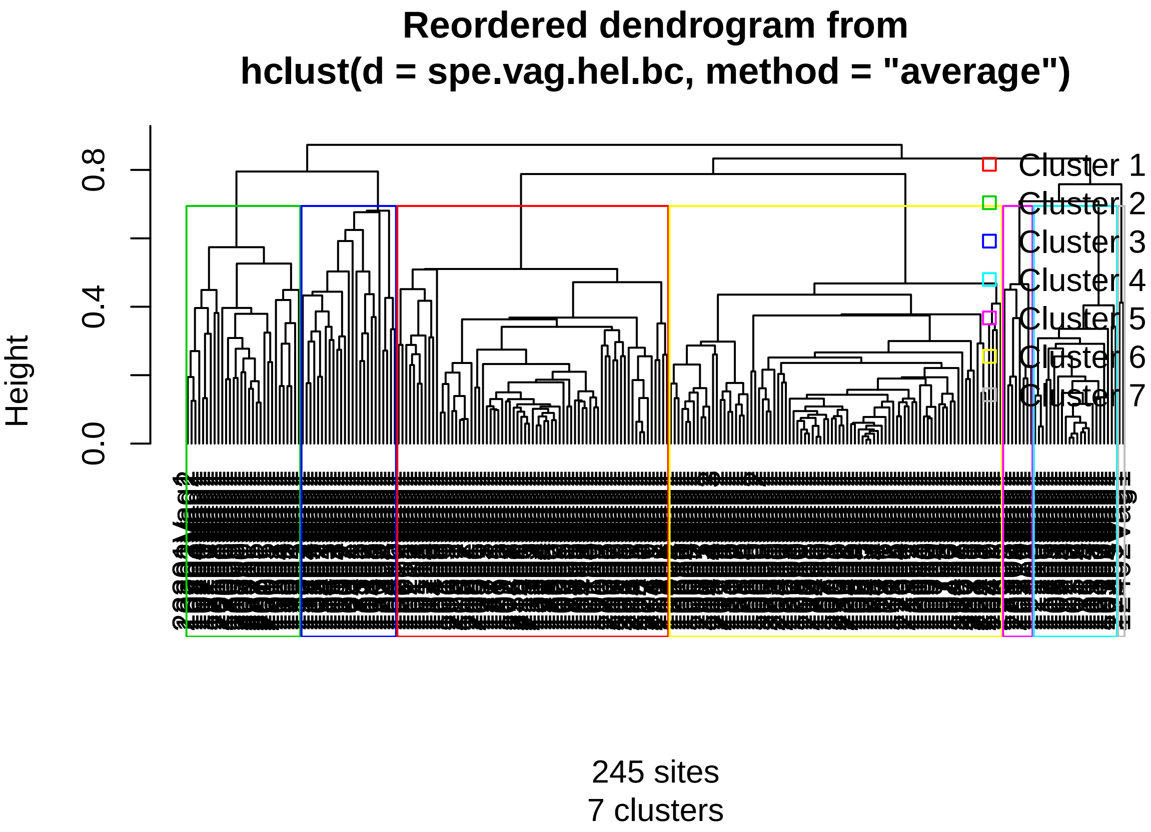
```
##  Silhouette-optimal number of clusters k = 7
##  with an average silhouette width of 0.4504
```

Now we can look at how well the number of clusters agrees with the hierarchical clustering of our samples:

```
## Set the optimal cluster number found above
k <- 7

## Cut the tree and assign samples to each of the seven groups
cutg <- cutree(spe.vag.hb.upgma, k=k)
sil <- silhouette(cutg, spe.vag.hel.bc)
rownames(sil) <- row.names(spe.abund.vag)

## Plot silhouette partition
plot(sil, main="Silhouette plot - Hellinger - UPGMA",
     cex.names=0.8, col=2:(k+1), nmax=100)
```



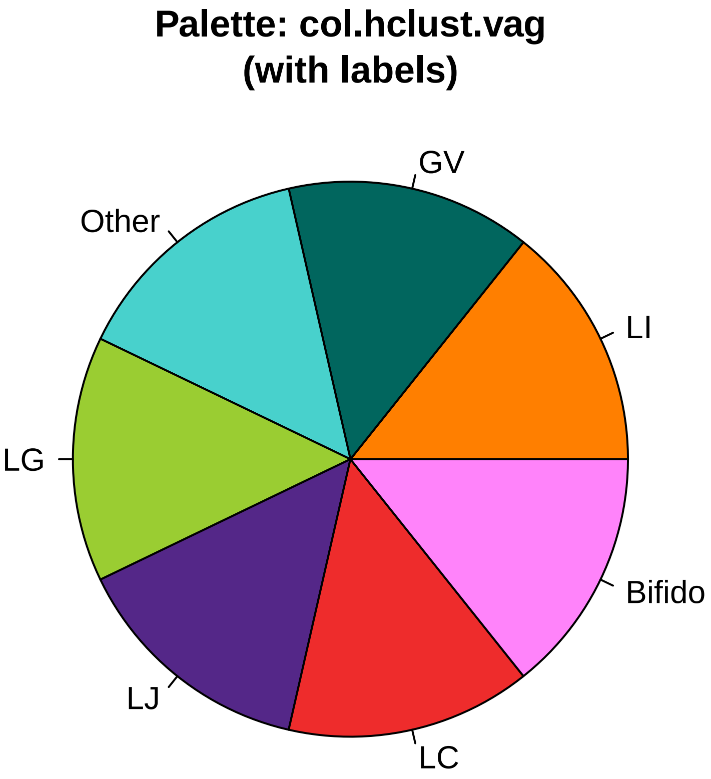
```
## Plot dendrogram with group labels
hcoplot(spe.vag.hb.upgma, spe.vag.hel.bc, k=7)
```



Assign the group IDs as a new variable in the metadata and define colors/names for each group:

```
## Add the group assignments as a new variable to the metadata
meta.vag$hclust <- cutg

## Define new colors and names for hclust groups (I determined these by
## looking at the heatmap in the next step then coming back to name the
## clusters and set colors that match the dominant taxon, if any)
col.hclust.vag <- c("1"=col.taxa["Lactobacillus_iners"],
                    "2"=col.taxa["Gardnerella_vaginalis"],
                    "3"="mediumturquoise",
                    "4"=col.taxa["Lactobacillus_gasseri"],
                    "5"=col.taxa["Lactobacillus_jensenii"],
                    "6"=col.taxa["Lactobacillus_crispatus"],
                    "7"=col.taxa["Bifidobacterium"])
names(col.hclust.vag) <- c("LI", "GV", "Other", "LG", "LJ", "LC", "Bifido")

## Replace cluster numbers with new names in metadata
meta.vag$hclust <- gsub("1", "LI", meta.vag$hclust)
meta.vag$hclust <- gsub("2", "GV", meta.vag$hclust)
meta.vag$hclust <- gsub("3", "Other", meta.vag$hclust)
meta.vag$hclust <- gsub("4", "LG", meta.vag$hclust)
meta.vag$hclust <- gsub("5", "LJ", meta.vag$hclust)
meta.vag$hclust <- gsub("6", "LC", meta.vag$hclust)
meta.vag$hclust <- gsub("7", "Bifido", meta.vag$hclust)

## Add these colors as a new variable to col.meta.vag
col.meta.vag$hclust <- col.hclust.vag[meta.vag$hclust]
```

```
## Print pie charts to show color palette
par(mar=c(1,1,2,1))
pie(rep(1,7), col=col.hclust.vag,
    main="Palette: col.hclust.vag\n(with labels)",
    labels=names(col.hclust.vag))
```



```
dev.off()
```

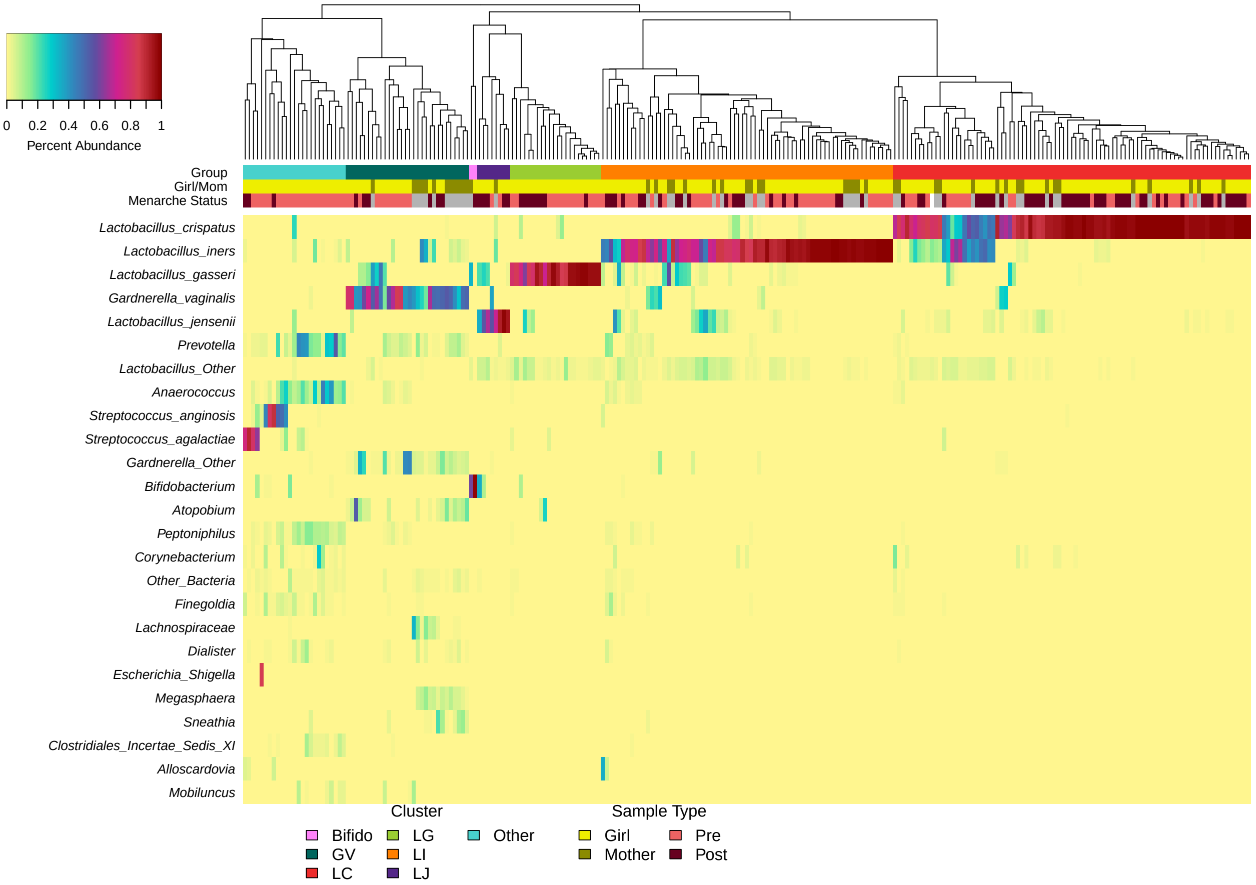
```
## null device
##           1
```

Now we visualize community composition as a heatmap along with the UPGMA dendrogram and cluster assignments just determined.

### Figure 1. Heatmap of the proportions of bacterial taxa in the vaginal microbiota of 31 adolescent girls and 24 mothers sampled longitudinally.

Each column in the graph represents the vaginal microbiota sampled from a single individual at a single point in time. In total 198 samples from girls and 47 samples from mothers are represented. The dendrogram above the graph represents the average linkage (UPGMA) hierarchical clustering of samples based on the Bray-Curtis dissimilarity matrix computed from Hellinger standardized taxon abundance data. The colored bars below the dendrogram represent hierarchical cluster assignments (top row) and sample types (second and third rows). Clusters are named to signify the most abundant taxon, when applicable: LC (*Lactobacillus crispatus* dominant, n=87), LI (*L. iners*, n=71), LG (*L. gasseri*, n=22), LJ (*L. jensenii*, n=8), GV (*Gardnerella vaginalis*, n=30), ‘Other’ (n=25), and ‘Bifido’ (*Bifidobacterium*, n=2). The heatmap shows the proportions (prior to Hellinger standardization) of the 25 overall most abundant taxa within each community as indicated by the color scale at top right. Sample type categories include girl/mother and premenarche/postmenarche (girls only; no menarche status is indicated for mother samples).

```r
## Sort taxa by abundance (column-wise, by taxon)
csum <- colSums(spe.prop.vag)

## Pick the top 25 taxa across all samples for the heatmap
pick <- order(csum,decreasing=TRUE)[1:25]

## Plot the dendrogram and heatmap with selected metadata
## Uncomment the next line to save as a PDF
# pdf("hclust-analysis/fig-1-hclust-heatmap.pdf", width=12, height=8, pointsize=10)
fcol <- cbind(col.meta.vag$men.stat, col.meta.vag$type, col.meta.vag$hclust)
colnames(fcol) <- c("Menarche Status", "Girl/Mom", "Group")
par(oma=c(2,2,2,2), lwd=0.75)
heatmap.3(t(spe.prop.vag[,pick]),
          col=col.heatmap,
          distfun=vegdist,
          hclustfun=function (z) hclust(spe.vag.hel.bc, method="average"),
          labCol="",
          ColSideColors=fcol,
          cexRow=1,
          cexCol=0.5,
          mar=c(5,5),
          dendrogram="column",
          Rowv=F,
          Colv=T,
          keysize=1,
          trace="none",
          key=T,
          density.info="none",family="sans")
par(oma=c(1,2,1,2), new=TRUE, xpd=TRUE)
plot(0:1, 0:1, type = "n", axes = F, xlab="", ylab="")
legend(0.15, -0.085, legend=sort(names(col.hclust.vag)), title="Cluster",
       fill=col.hclust.vag[order(names(col.hclust.vag))], bty="n",
       bg="#ffffff55", inset=0, cex=0.8, ncol=3)
legend(0.4, -0.085, legend=c("Girl", "Mother", "Pre", "Post"),
       title="Sample Type", fill=c(col.type[c("girl","mom")],
       col.men.stat[c("pre","post")]), bty="n",
       bg="#ffffff55", inset=0, cex=0.8, ncol=2)
```

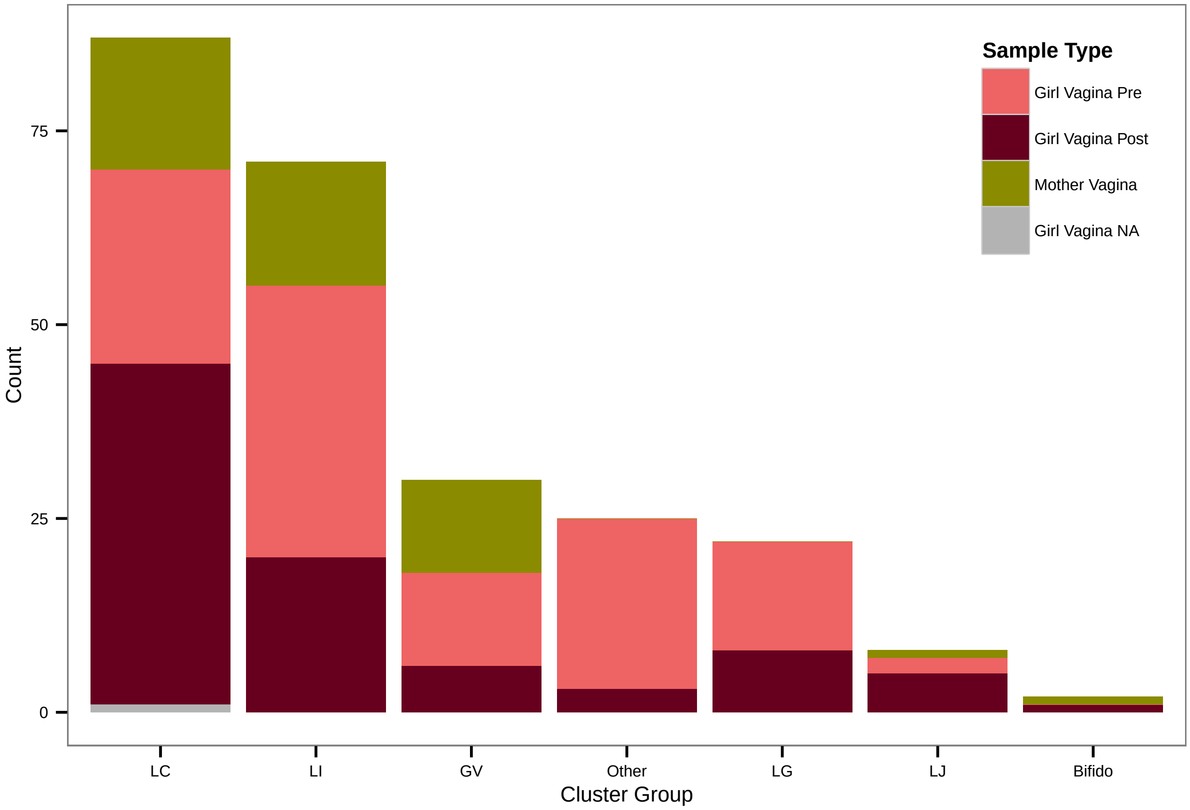


```
dev.off()
```

```
## null device
##           1
```

Next we want to count up the number of samples in each group (girl pre, girl post, mom) assigned to the different clusters.

```
## First we need to do a bit of reshaping to get counts and proportions
df <- meta.vag[,c("type.site.men.stat", "hclust")]

## Convert to "long" format
vag.hclust.count.lg <- ddply(df, .(type.site.men.stat, hclust),
                             summarize, count=length(hclust))

## Make "wide" format of counts of each sample type x cluster
vag.hclust.count.wd <- reshape(vag.hclust.count.lg,
                               idvar="type.site.men.stat",
                               timevar="hclust",
                               direction="wide")
vag.hclust.count.wd[is.na(vag.hclust.count.wd)] <- 0 # replace NA's with 0
colnames(vag.hclust.count.wd) <- gsub("count.", "", colnames(vag.hclust.count.wd))

## Reorder cluster ID columns alphabetically
ord <- c("type.site.men.stat",
         sort(colnames(vag.hclust.count.wd)[-1]))
vag.hclust.count.wd <- vag.hclust.count.wd[,ord]

## Convert wide back to long (now with 0's instead of NA's)
vag.hclust.count.lg <- melt(vag.hclust.count.wd,
                            id.vars="type.site.men.stat")
colnames(vag.hclust.count.lg) <- c("type.site.men.stat", "hclust", "count")

## Plot sample by cluster assignment
gg.hclust.clust <- ggplot(vag.hclust.count.lg,
                          aes(x=hclust, y=count, fill=type.site.men.stat)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=c("gray70", "#67001F", "#EE6363", "#8B8B00"),
                    breaks=c("girl.vag.pre", "girl.vag.post",
                             "mom.vag.mom.post", "girl.vag.NA"),
                    labels=c("Girl Vagina Pre", "Girl Vagina Post",
                             "Mother Vagina", "Girl Vagina NA"),
                    name="Sample Type") +
  scale_x_discrete(limits=c("LC","LI","GV","Other","LG","LJ","Bifido")) +
  xlab("Cluster Group") +
  ylab("Count") +
  theme_cust_nogrid +
  theme(legend.justification=c(1,1), legend.position=c(1,1),
        axis.title=element_text(size=8),
        axis.text=element_text(size=6),
        legend.title=element_text(face="bold", size=8),
        legend.text=element_text(size=6))

print(gg.hclust.clust)
```



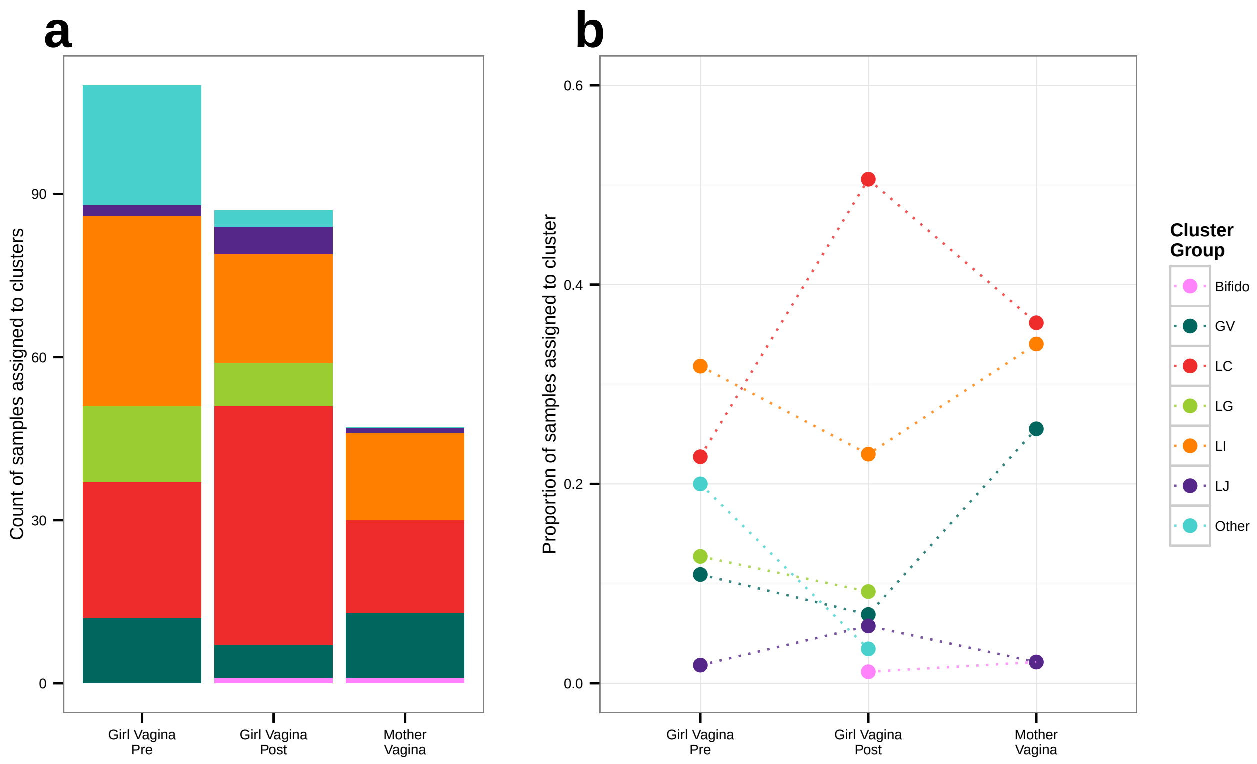
```
## Plot cluster assignment by sample type
gg.hclust.type <- ggplot(vag.hclust.count.lg,
                         aes(x=type.site.men.stat, y=count, fill=hclust)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=col.hclust.vag[vag.hclust.count.lg$hclust],
                    name="Cluster\nGroup") +
  scale_x_discrete(limits=c("girl.vag.pre", "girl.vag.post", "mom.vag.mom.post"),
                   labels=c("Girl Vagina\nPre",
                            "Girl Vagina\nPost",
                            "Mother\nVagina"),
                   name="") +
  ylab("Count of samples assigned to clusters") +
  theme_cust_nogrid +
  theme(axis.title=element_text(size=8),
        axis.text=element_text(size=6),
        legend.title=element_text(face="bold", size=8),
        legend.text=element_text(size=6))

## Determine proportions
vag.hclust.prop <- vag.hclust.count.wd[,-1]/rowSums(vag.hclust.count.wd[,-1])
vag.hclust.prop$type.site.men.stat <- vag.hclust.count.wd$type.site.men.stat

## Convert to long format
vag.hclust.prop.lg <- melt(vag.hclust.prop, id.vars="type.site.men.stat")
colnames(vag.hclust.prop.lg) <- c("type", "hclust","prop")

gg.hclust.prop <- ggplot(subset(vag.hclust.prop.lg, prop>0),
                         aes(x=type, y=prop, group=hclust, color=hclust)) +
  geom_point(size=3) +
  geom_line(lty=3, alpha=0.8) +
  scale_color_manual(values=col.hclust.vag[vag.hclust.count.lg$hclust],
                     name="Cluster\nGroup") +
  scale_x_discrete(limits=c("girl.vag.pre", "girl.vag.post", "mom.vag.mom.post"),
                   labels=c("Girl Vagina\nPre", "Girl Vagina\nPost", "Mother\nVagina"),
                   name="") +
  ylim(c(0,0.6))  +
  ylab("Proportion of samples assigned to cluster") +
  theme_cust +
  theme(axis.title=element_text(size=8),
        axis.text=element_text(size=6),
        legend.title=element_text(face="bold", size=8),
        legend.text=element_text(size=6))
```

### Figure 2. Hierarchical cluster assignment by sample type.

198 vaginal microbiota from girls and 47 vaginal microbiota from mothers were grouped into seven clusters. (a) Count of girl premenarcheal (n=110), girl postmenarcheal (n=87), and mother vaginal microbiota (n=47) assigned to each cluster group, indicated by the legend on far right (cluster names same as in Figure 1). (b) Proportion of samples in each sample type group listed in (a) assigned to each cluster group. The dotted lines serve to highlight differences between sample types and do not necessarily represent changes in cluster group prevalence over time.

```
## Uncomment the next line to save as a PDF
# pdf("hclust-analysis/fig-2-hclust-count-prop-comparison.pdf", width=8, height=5)
multiplot(gg.hclust.type +
            ggtitle("a") +
            theme(legend.position="none",
                  plot.title=element_text(size=22, hjust=-0.05)),
          gg.hclust.prop +
            ggtitle("b") +
            theme(plot.title=element_text(size=22, hjust=-0.05)),
          layout=matrix(c(1,1,2,2,2), nrow=1))
```

```
## Warning: Removed 7 rows containing missing values (position_stack).
## Warning: Removed 1 rows containing missing values (geom_point).
## Warning: Removed 1 rows containing missing values (geom_path).
```



```
dev.off()
```

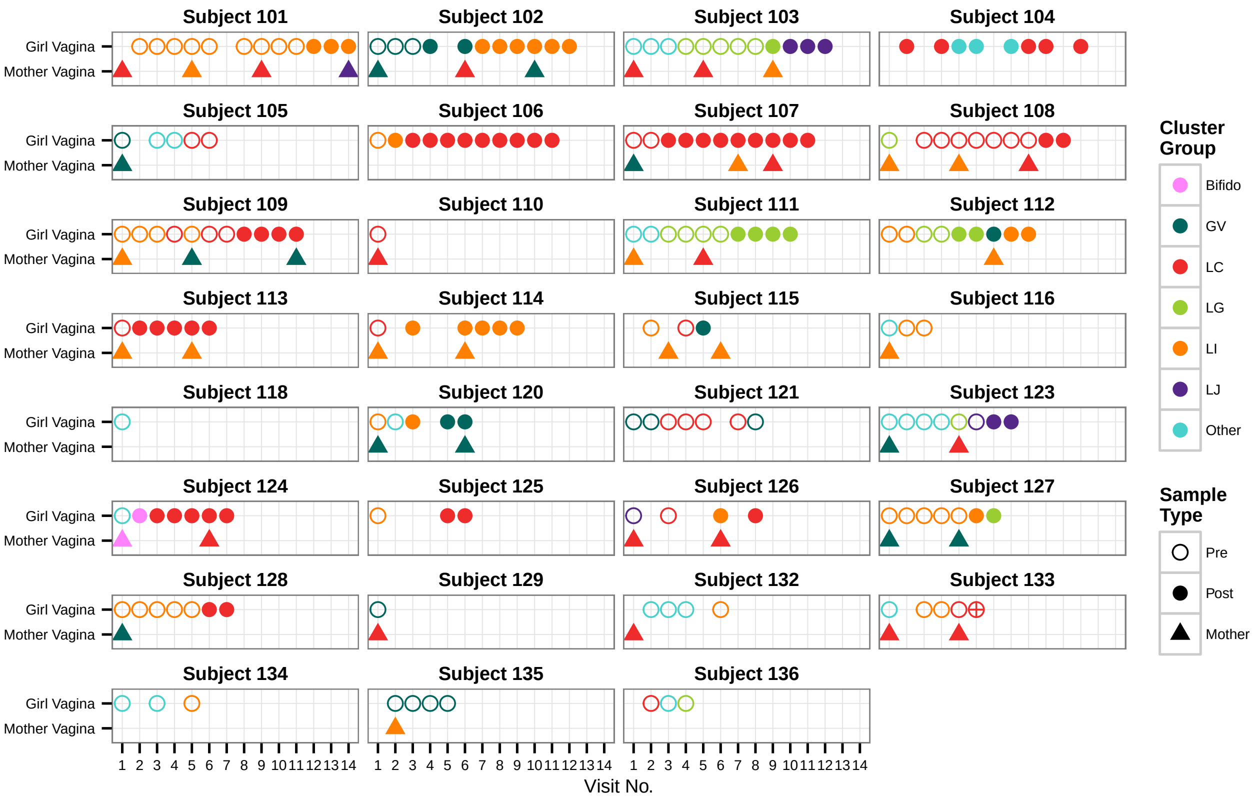
```
## null device
##           1
```

Finally, we look at the cluster assignments for all samples from each subject at each visit.

### Figure 3. Hierarchical cluster assignment over time within individual participants.

Each subplot shows the hierarchical cluster group assignment (same as in Figure 1) of each vaginal microbiota sample from an individual participant (circles) as well as each sample from her mother (triangles), when applicable. The x-axis indicates the clinical visit at which each sample was collected (visits occurred approximately every three months). Open circles signify premenarcheal status, and filled circles signify postmenarcheal status in girls. The menarcheal status was unknown for subject 133 at visit 6, indicated by an open circle with crosshatch.

```
gg.hclust.vag.samp <- ggplot(meta.vag,
                             aes(y=type.site, x=visit, color=factor(hclust), shape=men.stat)) +
  geom_point(size=3) +
  facet_wrap( ~ gm.pair.fullID, ncol=4) +
  scale_shape_manual(values=c(17, 16, 1),
                     breaks=c("pre", "post", "mom.post"),
                     labels=c("Pre", "Post", "Mother"),
                     name="Sample\nType", na.value=10) +
  scale_color_manual(values=col.hclust.vag[vag.hclust.count.lg$hclust],
                     name="Cluster\nGroup") +
  scale_y_discrete(limits=c("mom.vag", "girl.vag"),
                   labels=c("Mother Vagina", "Girl Vagina"),
                   name="") +
  scale_x_discrete(1:14, name="Visit No.") +
  theme_cust +
  theme(axis.title=element_text(size=8),
        axis.text=element_text(size=6),
        legend.title=element_text(face="bold", size=8),
        legend.text=element_text(size=6))

print(gg.hclust.vag.samp)
```



```
## Uncomment the next line to save as a PDF
# ggsave("hclust-analysis/fig-3-hclust-time.pdf", width=8, height=5, units="in")
```

### Cleanup

```
rm(spe.vag.hb.single, spe.vag.hb.complete, spe.vag.hb.ward,
   spe.vag.hb.single.coph, spe.vag.hb.complete.coph, spe.vag.hb.upgma.coph,
   spe.vag.hb.ward.coph, gow.vag.dist.single, gow.vag.dist.complete,
   gow.vag.dist.upgma, gow.vag.dist.ward, asw, sil, k.best, k, cutg,
   csum, pick, fcol, df, ord)
```

# Part II: Perform principal coordinates analysis (PCoA)

Now we perform PCoA to obtain a more nuanced picture of the similarities and differences among vaginal samples. PCoA is an ordination technique that, like principal components analysis (PCA), projects the distance or dissimilarity among objects (i.e., samples) onto a reduced set of orthogonal axes that maximize the variance explained by their multivariate descriptors (i.e., taxon abundances). Unlike PCA, it can be performed on any distance or dissimilarity matrix and need not conform to the Euclidean distance requirement of PCA. Therefore it is ideal to use with our Bray-Curtis dissimilarity matrix computed from Hellinger-standardized taxon abundance data. Again, the approaches below are based on those outlined in the following texts:

- Legendre P, Legendre L. (2012). *Cluster analysis.* 3rd ed. Elsevier.

- Borcard D, Gillet F, Legendre P. (2011). *Numerical ecology with R.* Springer.

## Setup PCoA

*Note: the PCoA method below produces negative eigenvalues unless corrected, which can be problematic for interpreting the R^2-like ratio (essentially variance explained by an eigenvalue in PCA). See Legendre & Legendre Numerical Ecology Ch 9 for more discussion of this (p. 505 in 3rd edition 2012). However, as long as the largest-value negative eigenvalue is smaller in absolute value than any of the positive eigenvalues of interest (typically the first two), the interpretation is still meaningful. A correction was suggested by Cailliez & Pagès to adjust the* $R^2$*-like ratio when negative eigenvalues are present – see Legendre & Legendre p. 506, eq. 9.48.*

```
## Calculate PCoA on Bray-Curtis dissimilarity matrix
spe.vag.hb.pcoa <- cmdscale(vegdist(spe.vag.hel), eig=TRUE, k=nrow(spe.vag.hel)-1)
```

```
## Warning: only 100 of the first 244 eigenvalues are > 0
```

```
## Calculate species scores
spe.vag.hb.wa <- wascores(spe.vag.hb.pcoa$points, spe.vag.hel)

## Apply Cailliez correction using ape::pcoa to obtain R^2-like ratios
spe.vag.hb.c.pcoa <- pcoa(vegdist(spe.vag.hel), correction="cailliez")

## R^2-like ratio for first three axes
spe.vag.hb.c.pcoa$values$Rel_corr_eig[1:3]
```

```
## [1] 0.16794 0.10954 0.07619
```

```
## ~ Total variance explained by first three axes
sum(spe.vag.hb.c.pcoa$values$Rel_corr_eig[1:3])
```

```
## [1] 0.3537
```

```
## Note: we will still use the PCoA computed from cmdscale for plotting
## since it is compatible with vegan's ordiplot functions (next section).
## I had difficulty applying the Cailliez correction to the cmdscale PCoA
## object which is why I used the above ape::pcoa function. If you run it
## without the Cailliez correction (uncomment spe.hb.pcoa.2 above first),
## you get the same eigenvalues as from the cmdscale method. See below:

## PCoA from ape::pcoa without the Cailliez correction
spe.vag.hb.pcoa.2 <- pcoa(vegdist(spe.vag.hel))

## Compare to PCoA from stats::cmdscale
head(spe.vag.hb.pcoa$eig)
```

```
## [1] 24.316 16.741 11.682  6.851  3.897  2.055
```

```
head(spe.vag.hb.pcoa.2$values$Eigenvalues)
```

```
## [1] 24.316 16.741 11.682  6.851  3.897  2.055
```

Now we can plot the PCoA and overlay different variables with point shapes and colors. We'll keep it simple by using only the first 2-3 axes, which as we saw above account for ~35% of the variance after applying the Cailliez correction. We'll look at the same plot colored according to different metadata variables to look for any interesting patterns in the data:

- Color-coding by Tanner breast score (Figure 4)
- Color-coding by hierarchical cluster group (Figure S2)

## Figure 4. PCoA of vaginal microbiota from girls and mothers. (color-coded by Tanner breast stage)

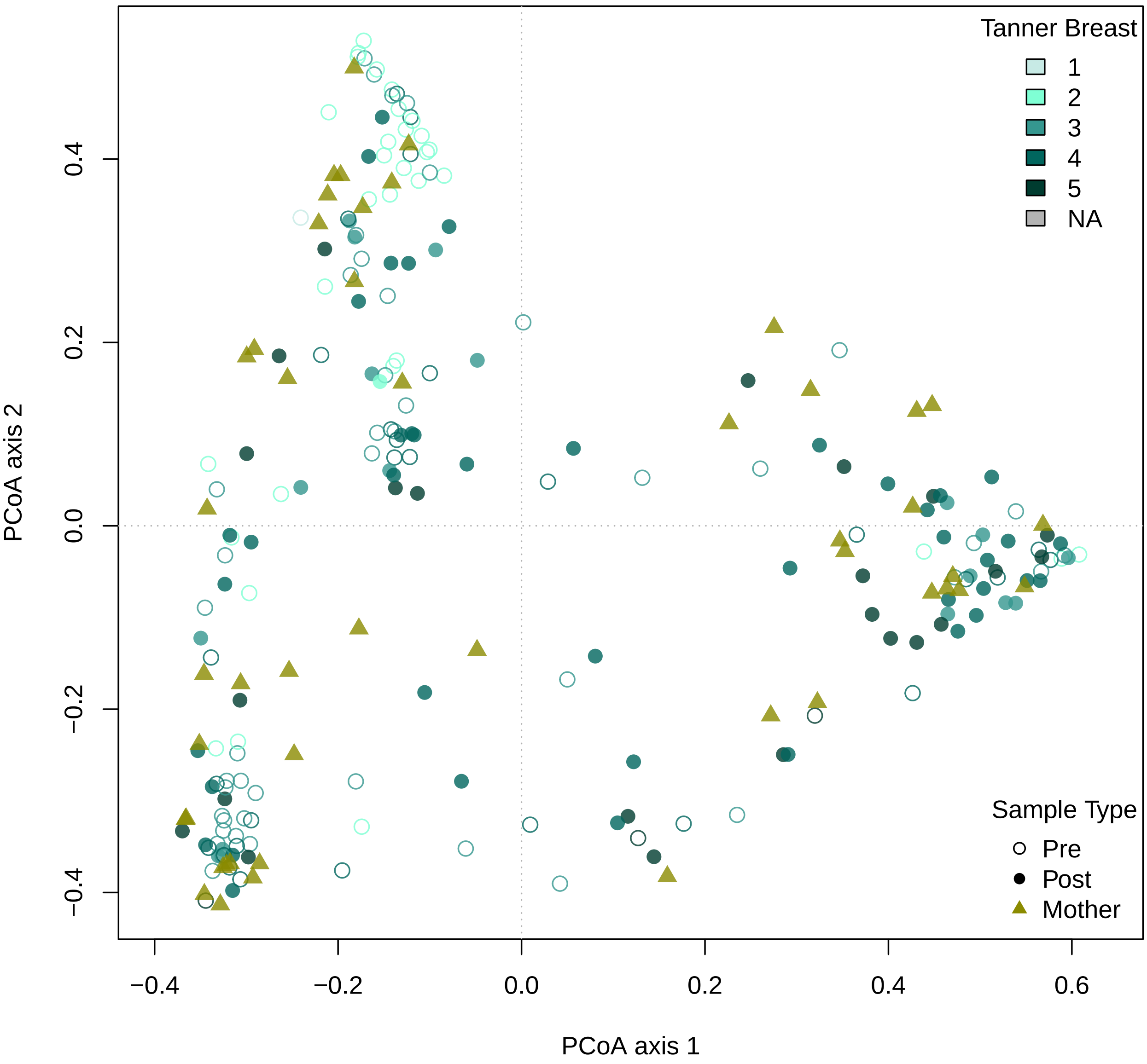
Principal Coordinates analysis (PCoA) was performed on the Bray-Curtis dissimilarity matrix computed from Hellinger-standardized taxon abundance data. Each point represents the vaginal microbiota sampled from a single individual at a single point in time (198 samples from girls and 47 from mothers), color-coded according to Tanner breast stage as indicated by the legend at top right, except for mother samples which are colored green. After applying a Cailliez correction to adjust for negative eigenvalues, the corrected $R^2$-like ratios (essentially percent variance explained) for the first and second PCoA axes are 0.168 and 0.110 (16.8% and 11.0%), respectively.

```
## Rework some of the colors for plotting purposes
col.meta.vag$tan.br.dr[meta.vag$type=="mom"] <- "#8B8B00" # Set mom colors to green
col.tb.vag.alpha15 <- makeTransparent(col.meta.vag$tan.br.dr, alpha=0.15)
col.tb.vag.alpha80 <- makeTransparent(col.meta.vag$tan.br.dr, alpha=0.80)

## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/fig-4-pcoa-tanner-br-2D.pdf", width=6, height=6, pointsize=10)
ordiplot(scores(spe.vag.hb.pcoa), type="n", xlab="PCoA axis 1", ylab="PCoA axis 2")
```

```
## Warning: Species scores not available
```

```
abline(h=0, lty=3, col="gray70")
abline(v=0, lty=3, col="gray70")
points(scores(spe.vag.hb.pcoa),
       col=col.tb.vag.alpha80,
       pch=col.meta.vag$pch.sample.gp,
       cex=1.3)
```

```
legend("topright", bty="n", bg="white", legend=c("1","2","3","4","5","NA"),
       title="Tanner Breast", fill=c(col.tanner,"gray50"))
legend("bottomright", bty="n", title="Sample Type",
       legend=c("Pre", "Post", "Mother"),
       pch=c(1,16,17), col=c(rep("black",2), col.type["mom"]))
```



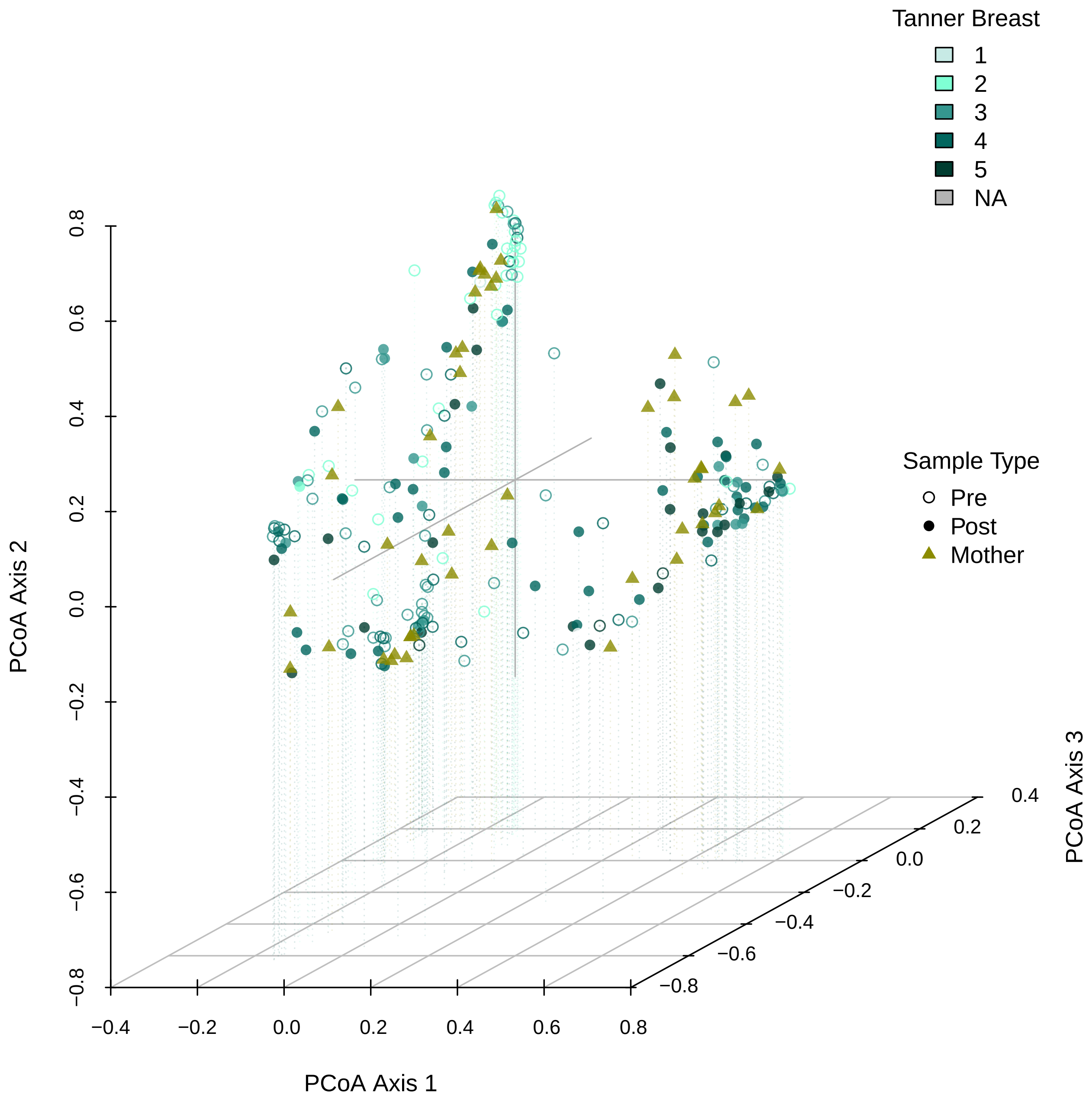
```
dev.off()
```

```
## null device
##           1
```

We get slightly better separation when we include the third axis:

```
## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/vag-pcoa-tanner-br-3D.pdf", width=6, height=6, pointsize=10)
```

```
p1 <- ordiplot3d(spe.vag.hb.pcoa, display="sites", choices=c(1,3,2),
                 xlab="PCoA Axis 1", ylab="PCoA Axis 3", zlab="PCoA Axis 2",
                 type="h", ax.col="gray70", adj=0, box=FALSE, mar=c(4,3,0,3),
                 pch=".", color=col.tb.vag.alpha15, angle=30, lty.hplot=3)
points(p1, "points", col=col.tb.vag.alpha80, pch=col.meta.vag$pch.sample.gp)
legend("right", bty="n", bg="white", title="Sample Type",
       legend=c("Pre", "Post", "Mother"),
       pch=c(1,16,17), , col=c(rep("black",2), col.type["mom"]))
legend("topright", bty="n", bg="white", legend=c("1","2","3","4","5","NA"),
       title="Tanner Breast", fill=c(col.tanner,"gray50"))
```



```
dev.off()
```

```
## null device
##           1
```

We can also color-code the points using different variables, such as the hierarchical cluster assignments determined above.

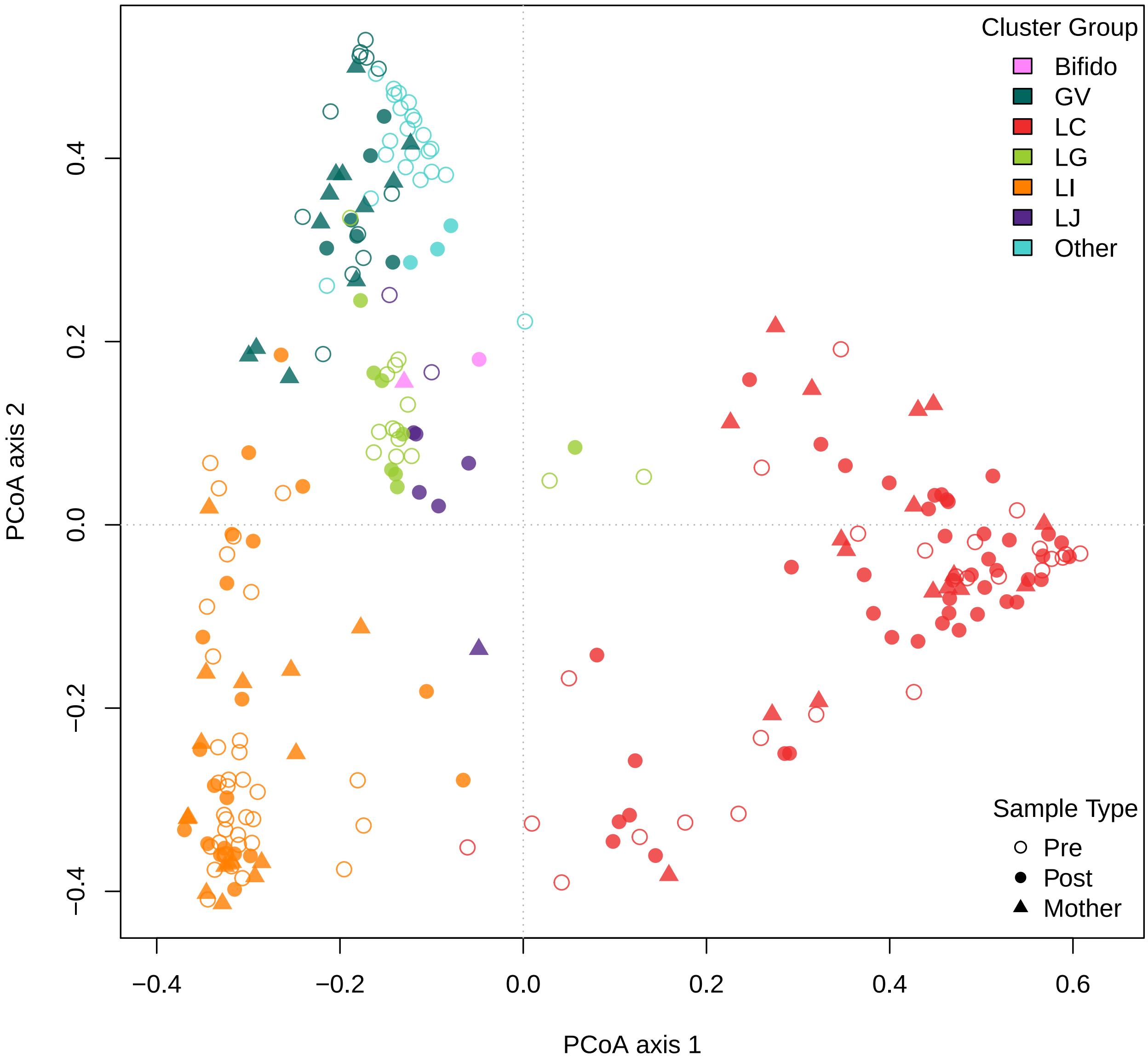
## Figure S2. PCoA of vaginal microbiota from girls and mothers. (color-coded by cluster assignment)

Principal coordinates analysis (PCoA) was performed on the Bray-Curtis dissimilarity matrix computed from Hellinger-standardized taxon abundance data. Each point represents the vaginal microbiota sampled from a single individual at a single point in time (198 samples from girls and 47 from mothers), color-coded according to groups determined by UPGMA hierarchical clustering. Open circles represent girl premenarcheal samples, filled circles represent girl postmenarcheal samples, and triangles represent mother samples. After applying a Cailliez correction to adjust for negative eigenvalues, the corrected $R^2$-like ratios (essentially percent variance explained) for the first and second PCoA axes are 0.168 and 0.110 (16.8% and 11.0%), respectively.

```
## Make semi-transparent colors for better plotting
col.hclust.vag.alpha15 <- makeTransparent(col.meta.vag$hclust, alpha=0.15)
col.hclust.vag.alpha80 <- makeTransparent(col.meta.vag$hclust, alpha=0.80)

## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/fig-s2-pcoa-hclust-2D.pdf", width=6, height=6, pointsize=10)
ordiplot(scores(spe.vag.hb.pcoa), type="n", xlab="PCoA axis 1", ylab="PCoA axis 2")
```

```
## Warning: Species scores not available
```

```
abline(h=0, lty=3, col="gray70")
abline(v=0, lty=3, col="gray70")
points(scores(spe.vag.hb.pcoa),
       col=col.hclust.vag.alpha80,
       pch=col.meta.vag$pch.sample.gp,
       cex=1.3)
legend("topright", bty="n", title="Cluster Group",
       legend=sort(names(col.hclust.vag)),
       fill=col.hclust.vag[order(names(col.hclust.vag))])
legend("bottomright", bty="n", title="Sample Type",
       legend=c("Pre", "Post", "Mother"),
       pch=c(1,16,17))
```

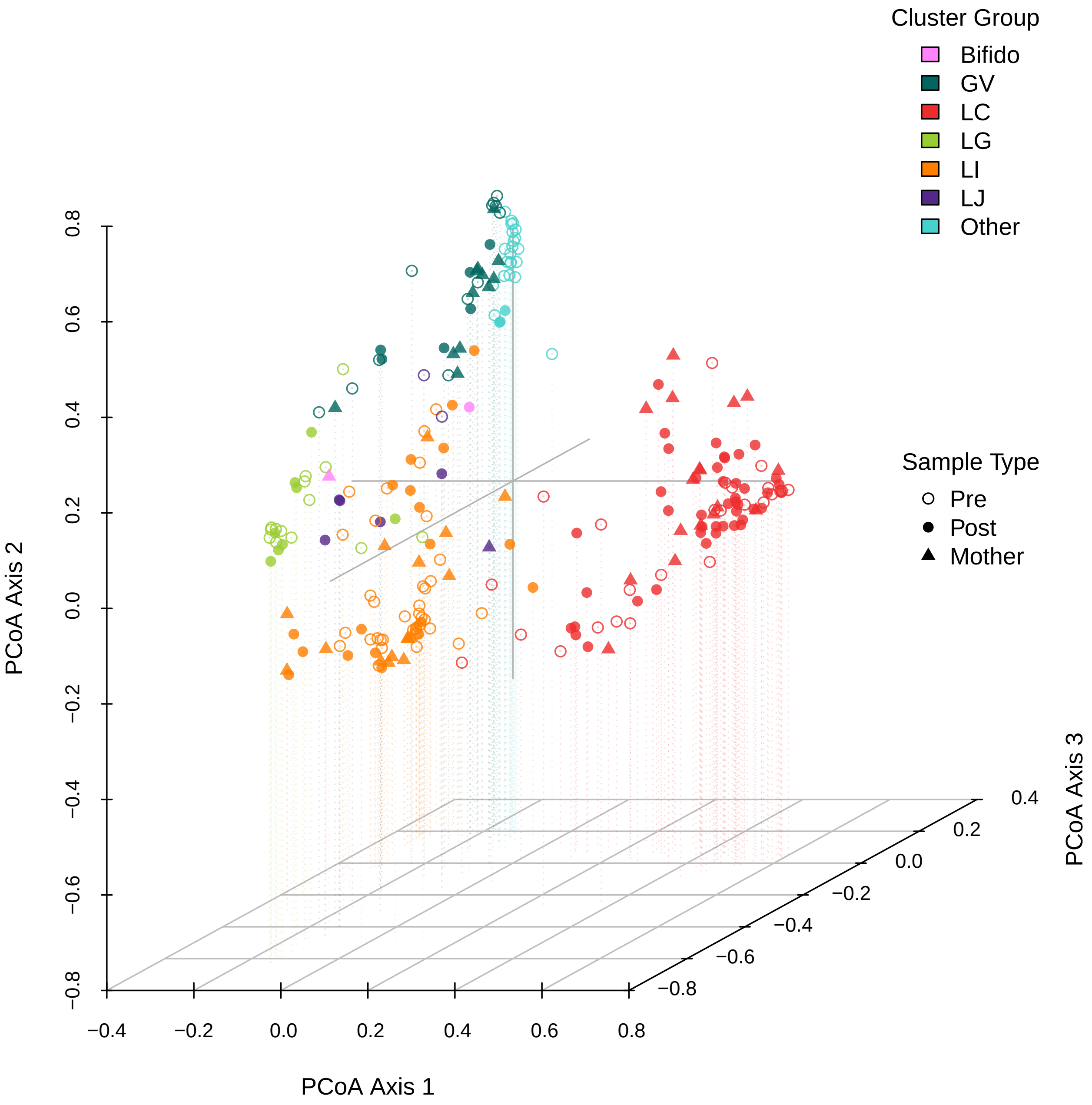


```
dev.off()
```

```
## null device
##           1
```

And again, we can plot in 3D (this allows us to separate out the points near the top, for instance)

```
## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/pcoa-hclust-3D.pdf", width=6, height=6, pointsize=10)
p1 <- ordiplot3d(spe.vag.hb.pcoa, display="sites", choices=c(1,3,2),
                 xlab="PCoA Axis 1", ylab="PCoA Axis 3", zlab="PCoA Axis 2",
                 type="h", ax.col="gray70", adj=0, box=FALSE, mar=c(4,3,0,3),
                 pch=".", color=col.hclust.vag.alpha15, angle=30, lty.hplot=3)
points(p1, "points", col=col.hclust.vag.alpha80, pch=col.meta.vag$pch.sample.gp)
legend("right", bty="n", bg="white", title="Sample Type",
       legend=c("Pre", "Post", "Mother"),
       pch=c(1,16,17))
legend("topright", bty="n", bg="white", legend=sort(names(col.hclust.vag)),
       title="Cluster Group", fill=col.hclust.vag[sort(names(col.hclust.vag))])
```



```
dev.off()
```

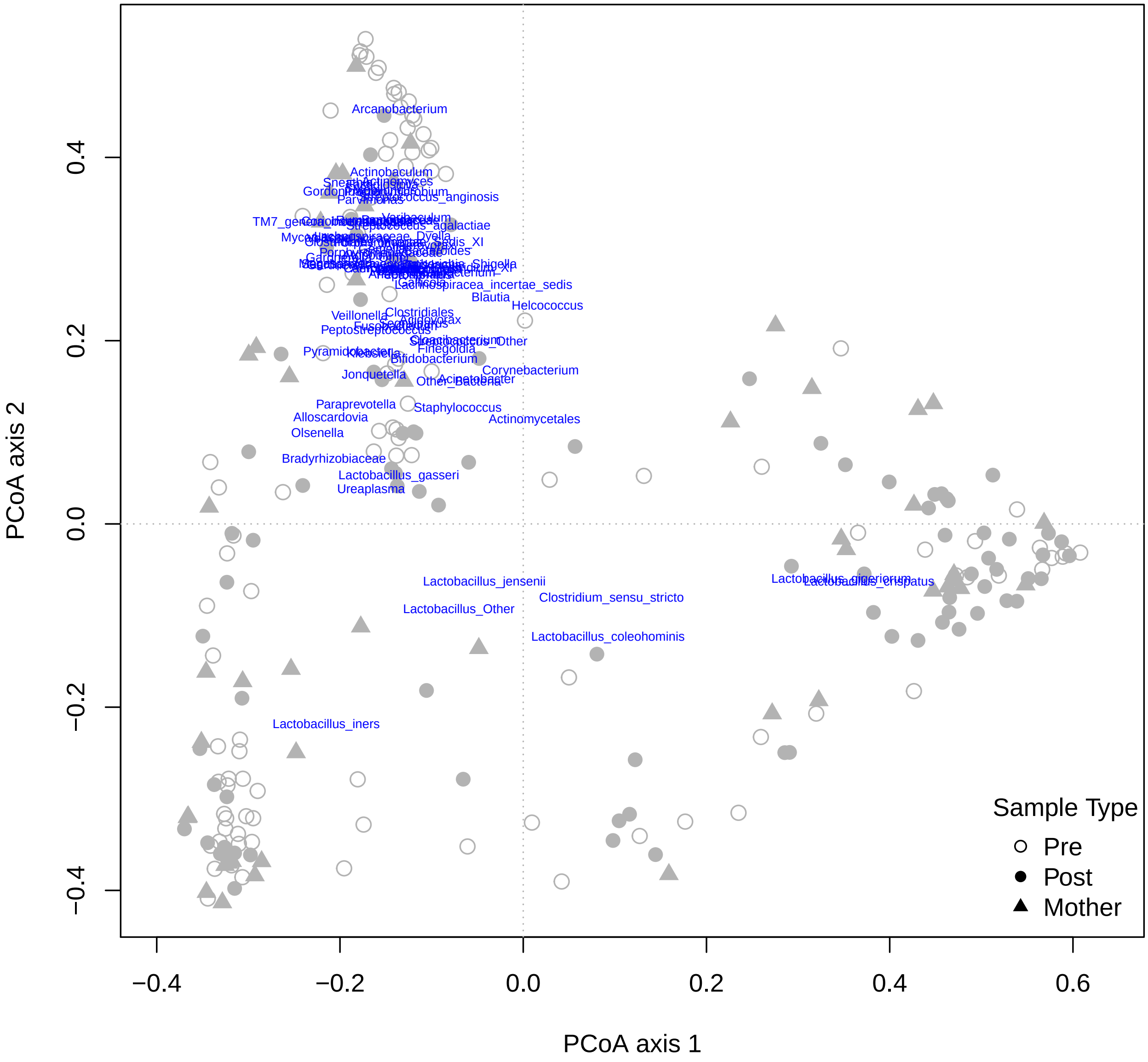
```
## null device
##           1
```

We can also make a "biplot" that shows the eigenvectors (this indicates which variables – bacterial taxa – contribute to separating samples along the axes):

```
## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/pcoa-biplot-2D.pdf", width=6, height=6, pointsize=10)
ordiplot(scores(spe.vag.hb.pcoa), type="n", xlab="PCoA axis 1", ylab="PCoA axis 2")
```

```
## Warning: Species scores not available
```

```
abline(h=0, lty=3, col="gray70")
abline(v=0, lty=3, col="gray70")
points(scores(spe.vag.hb.pcoa),
       col="gray70",
       pch=col.meta.vag$pch.sample.gp,
       cex=1.3)
text(spe.vag.hb.wa, rownames(spe.vag.hb.wa), cex=0.5, col="blue")
legend("bottomright", bty="n", title="Sample Type",
       legend=c("Pre", "Post", "Mother"),
       pch=c(1,16,17))
```



```
dev.off()
```

```
## null device
##           1
```

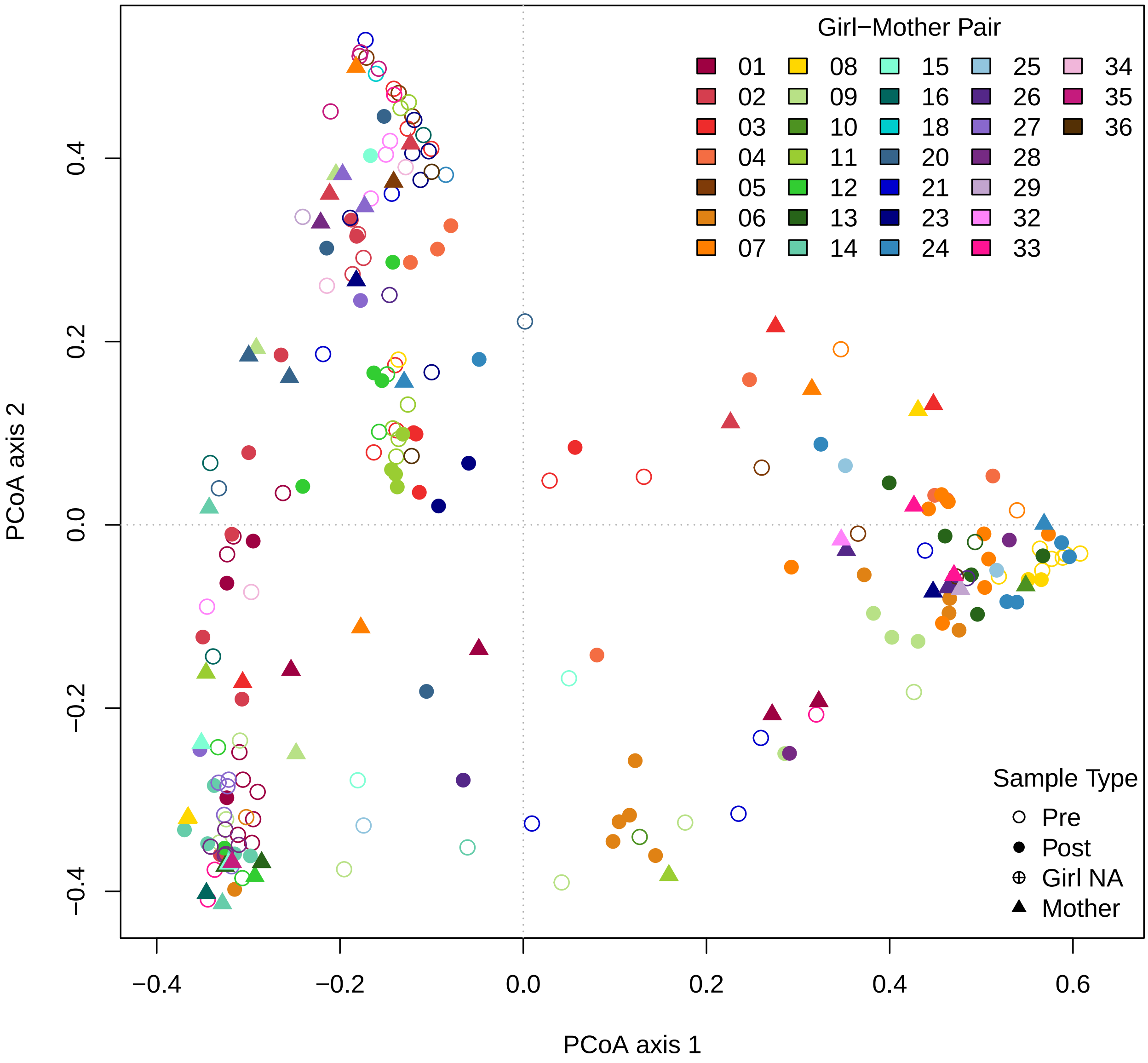
As part of exploratory data analysis, I color-coded by several different variables. These are not reported in the manuscript but may be interesting to look at.

Color-coding by girl-mom pair:

```
## Uncomment the next line to save as a PDF
# pdf("pcoa-analysis/vag-pcoa-girl-mom-2D.pdf", width=6, height=6, pointsize=10)
ordiplot(scores(spe.vag.hb.pcoa), type="n", xlab="PCoA axis 1", ylab="PCoA axis 2")
```

```
## Warning: Species scores not available
```

```
abline(h=0, lty=3, col="gray70")
abline(v=0, lty=3, col="gray70")
points(scores(spe.vag.hb.pcoa),
       col=col.meta.vag$gm.pair,
       pch=col.meta.vag$pch.sample.gp,
       cex=1.3)
legend("topright", bty="n", title="Girl-Mother Pair",
       legend=unique(meta.vag$gm.pair),
       fill=col.gm.pair, ncol=5)
legend("bottomright", bty="n", title="Sample Type",
       legend=c("Pre", "Post", "Girl NA", "Mother"),
       pch=c(1,16,10,17))
```



```
dev.off()
```

```
## null device
##           1
```

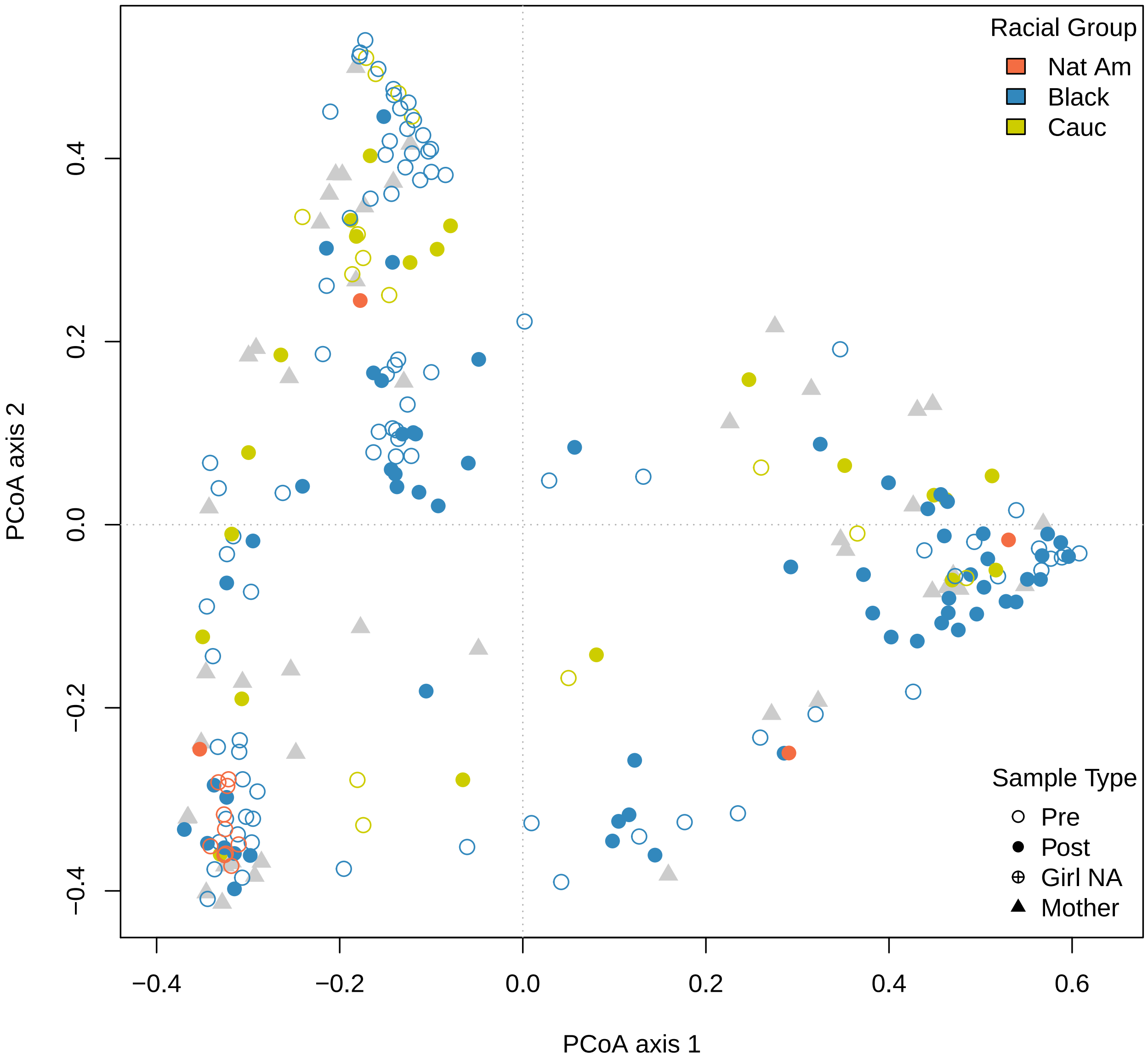
Color-coding by race:

```
## Uncomment next line to save as PDF
# pdf("pcoa-analysis/vag-pcoa-race-2D.pdf", width=6, height=6, pointsize=10)
ordiplot(scores(spe.vag.hb.pcoa), type="n", xlab="PCoA axis 1", ylab="PCoA axis 2")
```

```
## Warning: Species scores not available
```

```
abline(h=0, lty=3, col="gray70")
abline(v=0, lty=3, col="gray70")
```

```
points(scores(spe.vag.hb.pcoa)[meta.vag$type=="mom",1:2],
       col="gray80",
       pch=col.meta.vag$pch.sample.gp[meta.vag$type=="mom"],
       cex=1.3)
points(scores(spe.vag.hb.pcoa),
       col=col.50[c(4,21,39)][unclass(meta.vag$race)],
       pch=col.meta.vag$pch.sample.gp,
       cex=1.3)
legend("topright", bty="n", legend=c("Nat Am", "Black", "Cauc"), title="Racial Group",
       fill=col.50[c(4,21,39)])
legend("bottomright", bty="n", title="Sample Type",
       legend=c("Pre", "Post", "Girl NA", "Mother"),
       pch=c(1,16,10,17))
```



```
dev.off()
```

```
## null device
##           1
```

### Cleanup

```
rm(spe.vag.hb.c.pcoa, spe.vag.hb.pcoa.2, col.tb.vag.alpha15, col.tb.vag.alpha80,
   col.hclust.vag.alpha15, col.hclust.vag.alpha80, p1)
```

---

## Save R workspace

This will save the workspace (data) in two separate images: one named with today's date, in case you ever need to restore that version, and another with a non-dated name that can be easily loaded into subsequent analyses.

```
save.image(paste("data-postproc/02-cluster-pcoa-", Sys.Date(), ".RData", sep=""))
save.image(paste("data-postproc/02-cluster-pcoa-last-run.RData", sep=""))
```

End of Supplemental File 2. Next, move on to *Supplemental File 3. Analysis of vaginal microbiota dynamics in association with pubertal development.*