First association

gg

19 July, 2016

To run this file: Rscript -e “rmarkdown::render(‘first_association.Rmd’)”

Correlation

Correlation is particularly problematic for compositional data because the abundance of one part affects the abundance of every other. We can identify positively associated taxa by finding those taxa where the ratio between two or more taxa is approximately constant in all samples (Lovell et al. 2015). Such groups of taxa can be said to be associated (or positively correlated). Negative associations are very problematic because they can arise from underlying biology (which we want to detect) or from the negative correlation bias (when one thing goes up, one or more others must go down!).

Identifying correlated taxa is a very active and open research problem. Two methods have been proposed for datasets with particular sparsity profiles (Friedman and Alm 2012; Kurtz et al. 2015), and one method has been proposed for general datasets (Lovell et al. 2015). We will demonstrate ϕ, the general method, and give one or two examples of how this can be useful when analyzing microbiome datasets. The ϕ metric has been modified to allow us to include taxa with 0 values by using the Monte-Carlo replicates of the data that were generated by ALDEx2.

It is important to note that ϕ is a strength of association measure, and not a p value measure. In that way, it is similar to an effect size. You have some intuition about how ϕ behaves, and you choose a cutoff you are comfortable with. Guidance suggests that values less than 0.1 are appropriate for transcriptome datasets where the data are highly reproducible, and that 0.2-0.3 are more appropriate for the noisier microbiome datasets.

This plots the connectivity graph for $\phi <= 0.3$, and includes additional ways to display the information from ϕ. Displaying correlation graphs in a useful and intuitive way is a research topic by itself. Interested participants are encouraged to look at the igraph R package.

There is quite a bit of setup, since this is not in any easy-to-use R package as of yet.

```
# set a list of defined colours
colours <- c("indianred1", "steelblue3",  "skyblue1", "mediumorchid","royalblue4", "olivedrab3",
   "pink", "#FFED6F", "mediumorchid3", "ivory2", "tan1", "aquamarine3", "#C0C0C0",
    "mediumvioletred", "#999933", "#666699", "#CC9933", "#006666", "#3399FF",
   "#993300", "#CCCC99", "#666666", "#FFCC66", "#6699CC", "#663366", "#9999CC", "#CCCCCC",
   "#669999", "#CCCC66", "#CC6600", "#9999FF", "#0066CC", "#99CCCC", "#999999", "#FFCC00",
   "#009999", "#FF9900", "#999966", "#66CCCC", "#339966", "#CCCC33", "#EDEDED"
)

# read in the dataset and associated taxonomy file
d.subset <- read.table("data/ak_vs_op.txt", sep="\t", header=T, row.names=1)
taxon.1 <- read.table("data/ak_vs_op_taxon.txt", sep="\t", header=T, row.names=1)

# ALDEx2: generate Monte Carlo replicates of our clr data
# use 16 samples of the data to keep this short(ish)
x <- aldex.clr(d.subset, mc.samples=16, verbose=FALSE, useMC=TRUE)
```

```
## [1] "multicore environment is is OK -- using the BiocParallel package"
```

```r
# propr: calculate the phi statistic.
d.sma.df <- propr.aldex.phi(x)

# choose a cutoff. In practice this can be very low for transcriptomes and higher
# for microbiomes. Somewhere between 0.2 and 0.3 is a good place to start
phi.cutoff <- 0.30

# get the subset of OTUs that are joined by one or more low phi connections
d.sma.lo.phi <- subset(d.sma.df, phi < phi.cutoff)

# igraph: convert the connections into a graphical object
# igraph is a full-featured graph analysis and display tool
# the full use of which is well beyond what we can demonstrate here
g <- graph.data.frame(d.sma.lo.phi, directed=FALSE)

# igraph: group by clusters
g.clust <- clusters(g)

# make a table to examine the cluster membership by hand
g.df <- data.frame(Systematic.name=V(g)$name, cluster=g.clust$membership,
    cluster.size=g.clust$csize[g.clust$membership])

# generate a set of clusters larger than some size
# minimum is 2 (obviously)
big <- g.df[which(g.df$cluster.size >= 2),]
colnames(big) <- colnames(g.df)

# igraph: rename the cluster members by their genus name
V(g)$name <- as.vector(taxon.1[names(V(g)),"genus"])

# igraph:
plot(g, vertex.size=5, vertex.color=rgb(0,0,0,0.2),
   vertex.frame.color="white")
```

We next what to display the results on top of the compositional biplot to get some sense of how good the projection is, and to see if any of the low-ϕ clusters are distributed asymmetrically in the dataset. This can be done with a bit of fiddling.

We first generate a biplot using scale=0) so that the axes values correspond to the loadings. We color the sample names in light grey, and do not plot the loadings values. Then we overplot the clusters and give each cluster a unique color from the list of colours. For this plot, we will include only clusters with sizes greater than 5 to plot to keep it simple.

```r
big <- g.df[which(g.df$cluster.size >= 5),]

# zCompositions: make a pcx object as we did for the biplot
d.n0 <- cmultRepl(t(d.subset), label=0, method="CZM")
```

```
## No. corrected values:  891009
```

```r
d.clr <- t(apply(d.n0, 1, function(x) log(x) - mean(log(x)) ) )
pcx.d <- prcomp(d.clr)
```

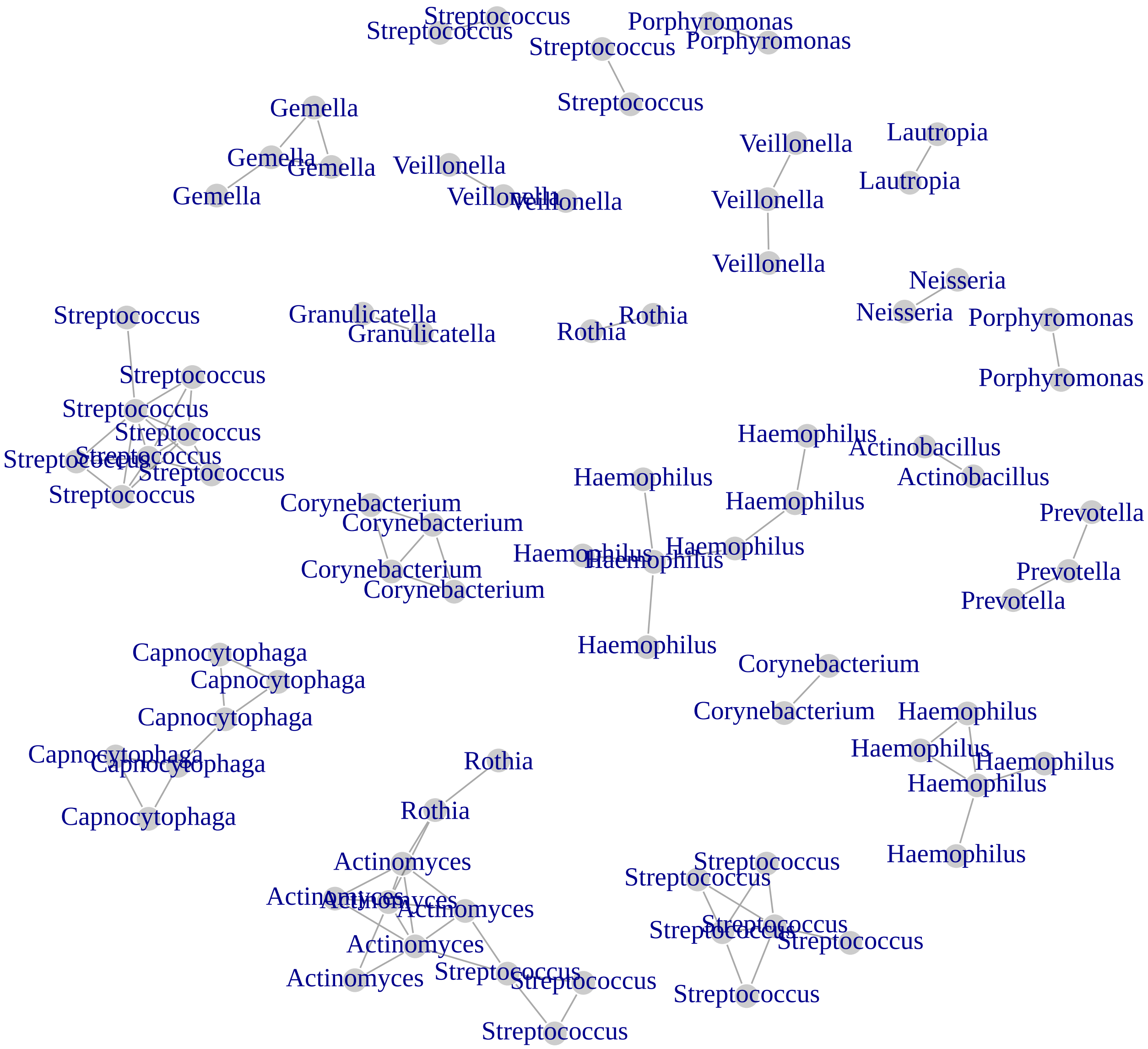

Figure 1: The connectivity graph for phi <=0.29. In this dataset, for the most part, we have small clusters composed of taxa from the same genus. This could be because of inefficient clustering, or because of similar organisms in the same genus exploiting exactly the same resources. There are a few clusters composed of different genera, and these are most likely members of different genera exploiting resources similarly.

```r
biplot(pcx.d, var.axes=F, cex=c(0.4,0.01), scale=0,
  col=c(rgb(0,0,0,0.2), "red"),
  xlab=paste("PC1: ", round(sum(pcx.d$sdev[1]^2)/sum(pcx.d$sdev^2), 3), sep=""),
  ylab=paste("PC2: ", round(sum(pcx.d$sdev[2]^2)/sum(pcx.d$sdev^2),3),sep="")
  )

#abline(v=0, lty=2, col="grey")
#abline(h=0, lty=2, col="grey")

lev <- factor(big$cluster)
for(i in as.numeric(levels(lev))){
nms <- rownames(big)[big$cluster==i]

text(pcx.d$rotation[nms,][,1], pcx.d$rotation[nms,][,2],
    labels = taxon.1[rownames(big)[big$cluster==i],"genus"],col=colours[i], cex=1)
}
```

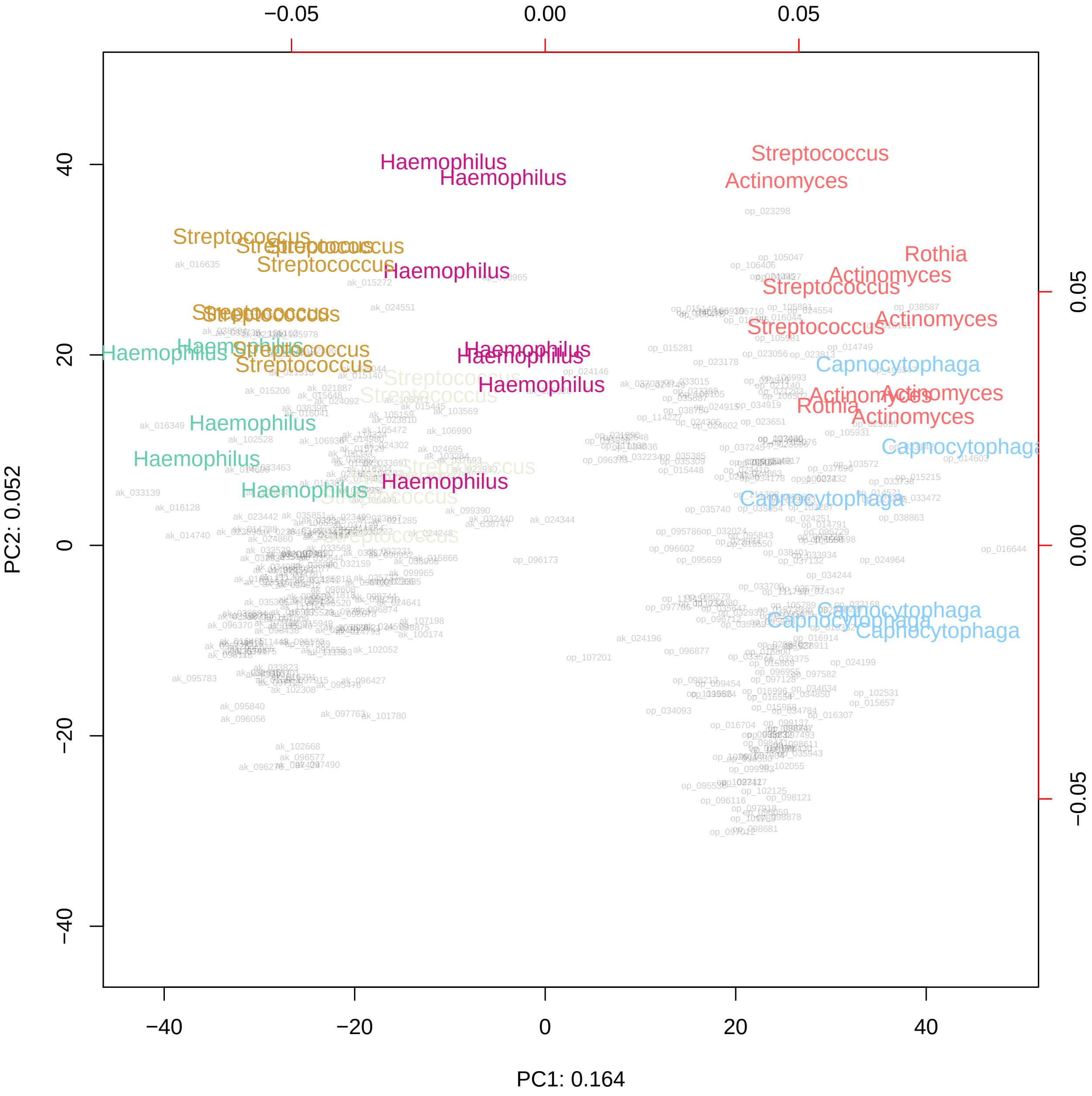

Figure 2: Inspection of the location of the large clusters on the biplot shows that taxa with very similar standard deviations on PC1 are clustered, and that the projection of the data is not particularly good. The Haemophilus cluster, in particular, has a large spread on component 2, but all have a similar distance from the origin on PC1. However, there are two clusters that appear to separate the groups strongly. One, associated with the ak group is composed of Streptococcus, the other associated with the op group contains Streptococus, and Actinomyces. Some instances will have Rothia also associated with the latter group. This is a consequence of examining random instances of the data and indicate that the Rothia exhibit a marginal score.

```r
clop <- c("5425","34862","27937","34896","11687","36010","32985","35898","35936","36038","11","3760")

clopc <- c("26066", "28841","22379", "6107","21779")

clak <- c("38304","39078","39235","39306","39171","39227","37976","38334","38833")

mean.op <- apply(d.clr[,clop], 1, mean)
mean.ak <- apply(d.clr[,clak], 1, mean)
nms <- rownames(d.clr)

hist(mean.op[grep("ak", nms)] - mean.ak[grep("ak", nms)], col=rgb(0,0,1,0.4), xlim=c(-8,8), ylim=c(0,80
hist(mean.op[grep("op", nms)] - mean.ak[grep("op", nms)], col=rgb(1,0,0,0.4), add=T)
```

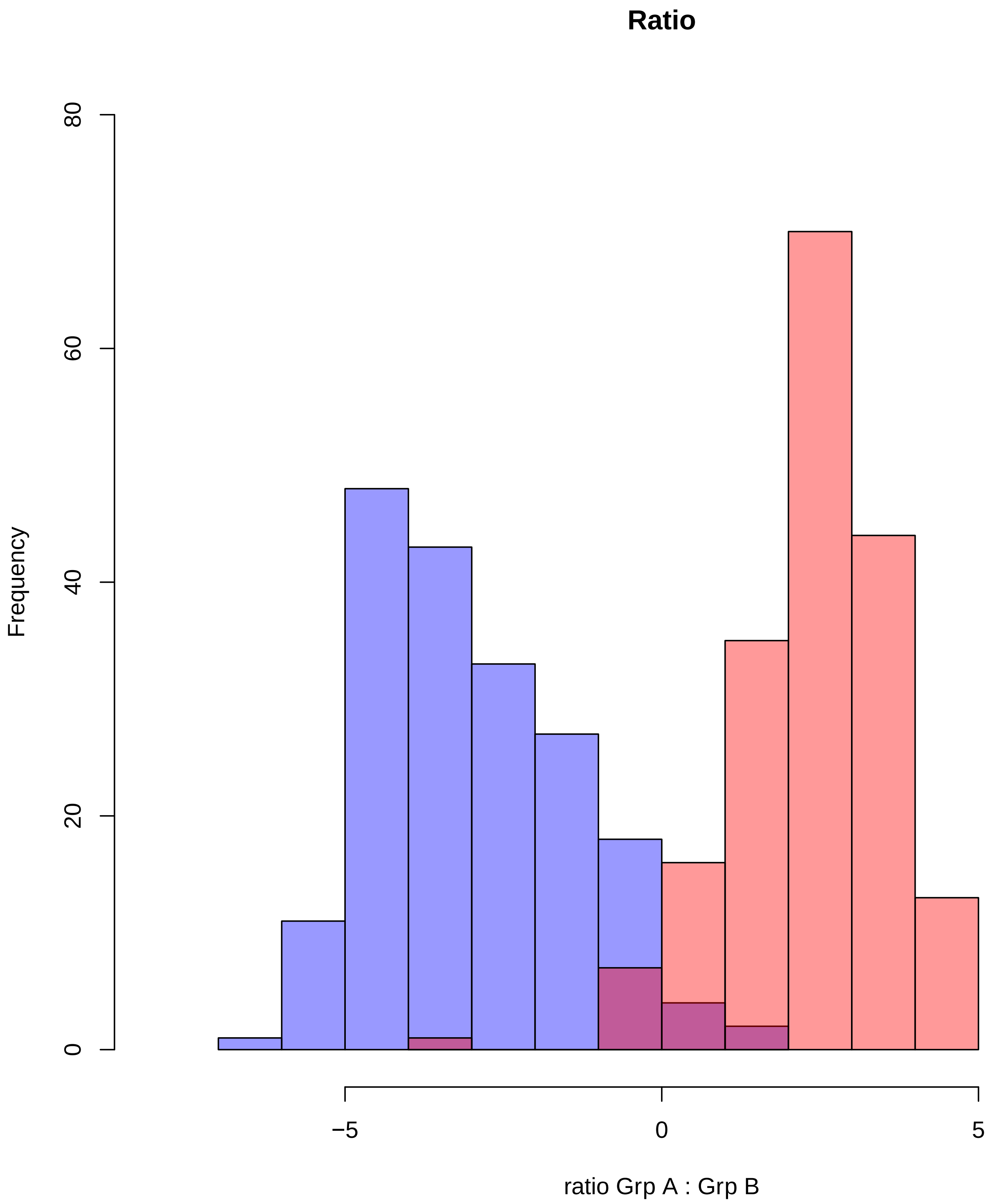

Figure 3: Separation of groups based on ratios of sets of associated taxa.

References

Friedman, Jonathan, and Eric J Alm. 2012. “Inferring Correlation Networks from Genomic Survey Data.” *PLoS Comput Biol* 8 (9): e1002687. doi:10.1371/journal.pcbi.1002687.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. 2015. “Sparse and Compositionally Robust Inference of Microbial Ecological Networks.” *PLoS Comput Biol* 11 (5): e1004226. doi:10.1371/journal.pcbi.1004226.

Lovell, David, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. 2015. “Proportionality: A Valid Alternative to Correlation for Relative Data.” *PLoS Comput Biol* 11 (3): e1004075. doi:10.1371/journal.pcbi.1004075.