First biplot

gg

19 July, 2016

To run this file: Rscript -e "rmarkdown::render('first_biplot.Rmd')"

We will use as an example the HMP (Proctor 2011) oral microbiome dataset and compare the attached keratinized gingiva and the supra-gingival plaque. For comparison, the sub and supra-gingival samples are compared as well. These are adjacent, in the mouth but the former pair have separable microbiome compositions, and the latter pair has few if any real differences. These data have been filtered to include only those OTUs that are present with a mean count of at least 0.1 across all samples.

The compositional biplot is the first exploratory data analysis tool that should be used whenever exploring a microbiome dataset. It shows, in one plot, the essences of your results. Do my samples separate into groups? what taxa or groups of taxa are driving this separation? what taxa are irrelevant to the analysis?

Compositional biplots appear to be complex and intimidating, but with a little patience and practice they are easily interpretable (Aitchison and Greenacre 2002). They are based on the variance of the ratios of the parts, and are substantially more informative that the commonly used PCoA plots that are driven largely by abundance (Gorvitovskaia, Holmes, and Huse 2016).

```
# 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)

# load the library zCompositions to perform 0 replacement
library(zCompositions)

# it is important to first filter to remove rows that are exclusively 0 values
# we are using the Count Zero Multiplicative approach
d.n0 <- cmultRepl(t(d.subset), method="CZM", label=0)
```

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

```
# generate the centered log-ratio transformed data
d.clr <- t(apply(d.n0, 1, function(x) log(x) - mean(log(x))))

# apply a singular value decomposition to the dataset
# do not use princomp function in R!!
pcx <- prcomp(d.clr)

# rename the OTUs by their genus
rownames(pcx$rotation) <- taxon.1[rownames(pcx$rotation), "genus"]

# get the labels for the first two components
PC1 <- paste("PC1: ", round(pcx$sdev[1]^2/sum(pcx$sdev^2),3), sep="")
PC2 <- paste("PC2: ", round(pcx$sdev[2]^2/sum(pcx$sdev^2),3), sep="")

par(fig=c(0,1,0,1), new=TRUE)
# generate a scree plot
par(fig=c(0,0.8,0,1), new=TRUE)
biplot(pcx, cex=c(0.6,0.6), col=c("black", rgb(1,0,0,0.2)), var.axes=F, scale=0,
```

```
    xlab=PC1, ylab=PC2)
abline(h=0, lty=2, lwd=2, col=rgb(0,0,0,0.3))
abline(v=0, lty=2, lwd=2, col=rgb(0,0,0,0.3))

par(fig=c(0.8,1,0,1), new=TRUE)
plot(pcx, main="hist")
```

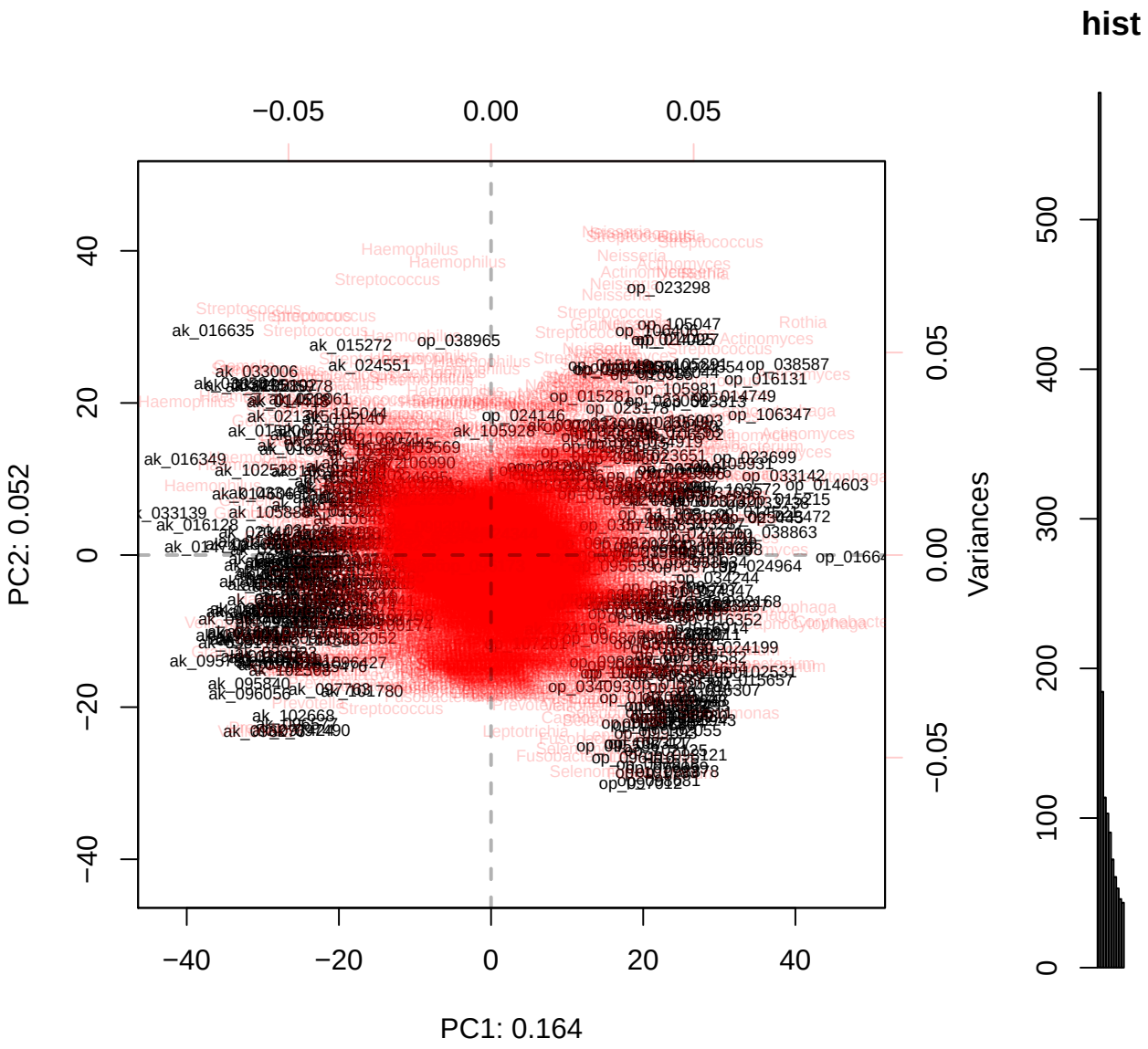

Figure 1: The compositional biplot is the workhorse tool for CoDa. This plot summarizes the entire analysis in a qualitative manner. We can see that the op and ak samples separate very well, although the proportion of variance explained on component 1 is small. Furthermore, we can see the genus names of some of the OTUs that are driving this divide. Finally, component 1 has substantially more variance than does componet 2, and we can explain this experiment as a simple two part comparison with the largest variance along the axis of the comparison.

Rules for interpreting compositional biplots:

- All interpretations are up to the limit of the variance explained. We can think of this as a shadow of the multidimensional dataset (4545 dimensions!) projected onto two dimensions. If the variance explained is high (> 0.8) then the edges of the shadows are sharp, however, if the variance explained is low, as it is here, then we have little confidence in the exact placement of any individual sample or OTU.
- The distance between samples is related to their multivariate similarity of the parts as ratios. If all components are relatively the same (ie, the ratios between all parts are identical), then two samples are in the same location.
- We must interpret the OTUs as ratios. Abundance information is not directly available on these plots.
- The distance and direction of an OTU from the origin is the standard deviation of the ratio of that OTU to the geometric mean of all OTUs. So Haemophilus at (-40,20) is highly variable, and more abundant in the ak than in the op samples
- The line between any set of OTUs is called a link. Links that pass through more than one OTU are permitted and do not change the interpretation.
- Short links indicate a constant or near constant ratio between the two (or more) linked OTUs in the dataset. So the two Haemophilus OTUs at (-5,40) will likely have a relatively constant ratio across all samples since their link is short.
- Long links indicate a non-constant ratio between the joined OTUs, and define a ratio relationship that can be inverse or random. There is no principled method to determine which is the case. However, when we have a clear separation, we can see if the link between Haemophilus (-40,20) and Corynebacterium (45,-10) defines a reciprocal abundance relationship that separates the two groups. However, note that this relationship will be very noisy because of the low variance explained, driven in part by the high variation within groups.

We can illustrate this as follows:

```
# get the centered log-ratio values of the two most extreme OTUs
cor <- d.clr[,"38849"]
haem <- d.clr[,"38193"]

# get the sample names sorted by PC1 location
nms <- sort(pcx$x[,1])

ak <- grep("ak", names(nms))
op <- grep("op", names(nms))
plot(nms,exp(cor[names(nms)] - haem[names(nms)]), log="y",
    ylab="Corynebacterium : Haemophilus ratio", xlab="PC1 location", pch="")
points(nms[op],exp(cor[names(nms[op])] - haem[names(nms[op])]), col="red", pch="O")
points(nms[ak],exp(cor[names(nms[ak])] - haem[names(nms[ak])]), col="blue", pch="A")

abline(v=0, lty=2, lwd=2, col=rgb(0,0,0,0.3))
```

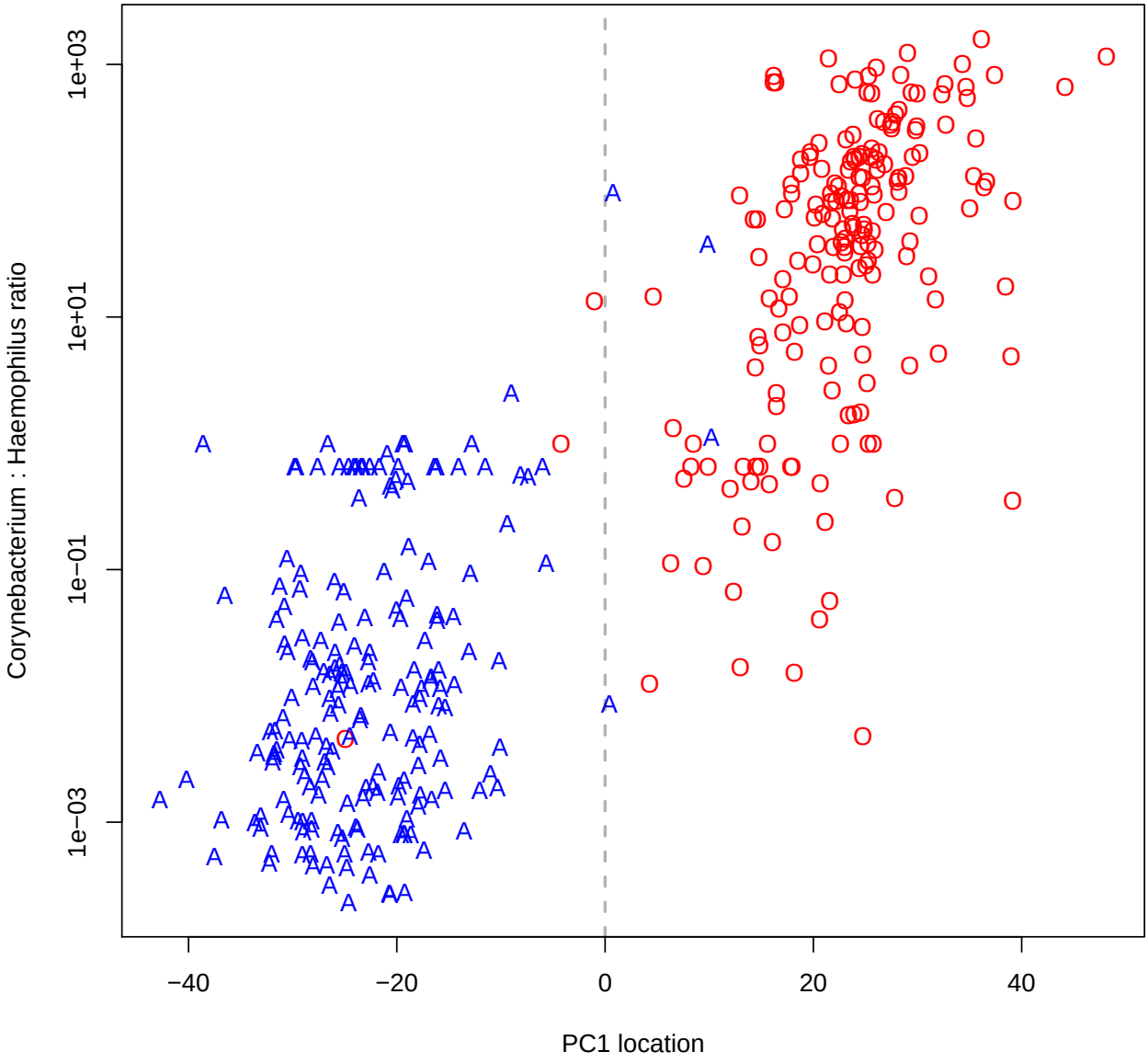

Figure 2: Taxa at extreme ends of a principle component will be seen to have reciprocal abundance relationships. This is true for any component, not just the first. However the relationship will, in general, be strongest along the first component. Note that the ratios between these two OTUs is not really a linear trend so much as we can see that each OTU it is two more or less exclusively more abundant in one group or the other. O is the OP group, and A is the AK group. Identifying taxa that differentiate the two groups is the first step to determining which taxa will be most useful for linear discriminate analysis or other modelling steps.

```r
# some initial characterization of the whole dataset

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

# it is important to first filter to remove rows that are exclusively 0 values
# we are using the Count Zero Multiplicative approach
e.n0 <- cmultRepl(t(e.subset), method="CZM", label=0)
```

```
## No. corrected values:  927041
```

```r
# this has given the number of corrected values (number of 0s replaced)

# generate the centered log-ratio transformed data
e.clr <- t(apply(e.n0, 1, function(x) log(x) - mean(log(x))))

# apply a singular value decomposition to the dataset
# do not use princomp function in R!!
pcx.e <- prcomp(e.clr)

# rename the OTUs by their genus
rownames(pcx.e$rotation) <- taxon.e[rownames(pcx.e$rotation), "genus"]

# get the labels for the first two components
PC1 <- paste("PC1: ", round(pcx.e$sdev[1]^2/sum(pcx.e$sdev^2),3), sep="")
PC2 <- paste("PC2: ", round(pcx.e$sdev[2]^2/sum(pcx.e$sdev^2),3), sep="")

par(fig=c(0,1,0,1), new=TRUE)
# generate the biplot
par(fig=c(0,0.8,0,1), new=TRUE)
biplot(pcx.e, cex=c(0.6,0.6), col=c("black", rgb(1,0,0,0.2)), var.axes=F, scale=0,
    xlab=PC1, ylab=PC2)
abline(h=0, lty=2, lwd=2, col=rgb(0,0,0,0.3))
abline(v=0, lty=2, lwd=2, col=rgb(0,0,0,0.3))

par(fig=c(0.8,1,0,1), new=TRUE)
plot(pcx.e, main="hist")
```

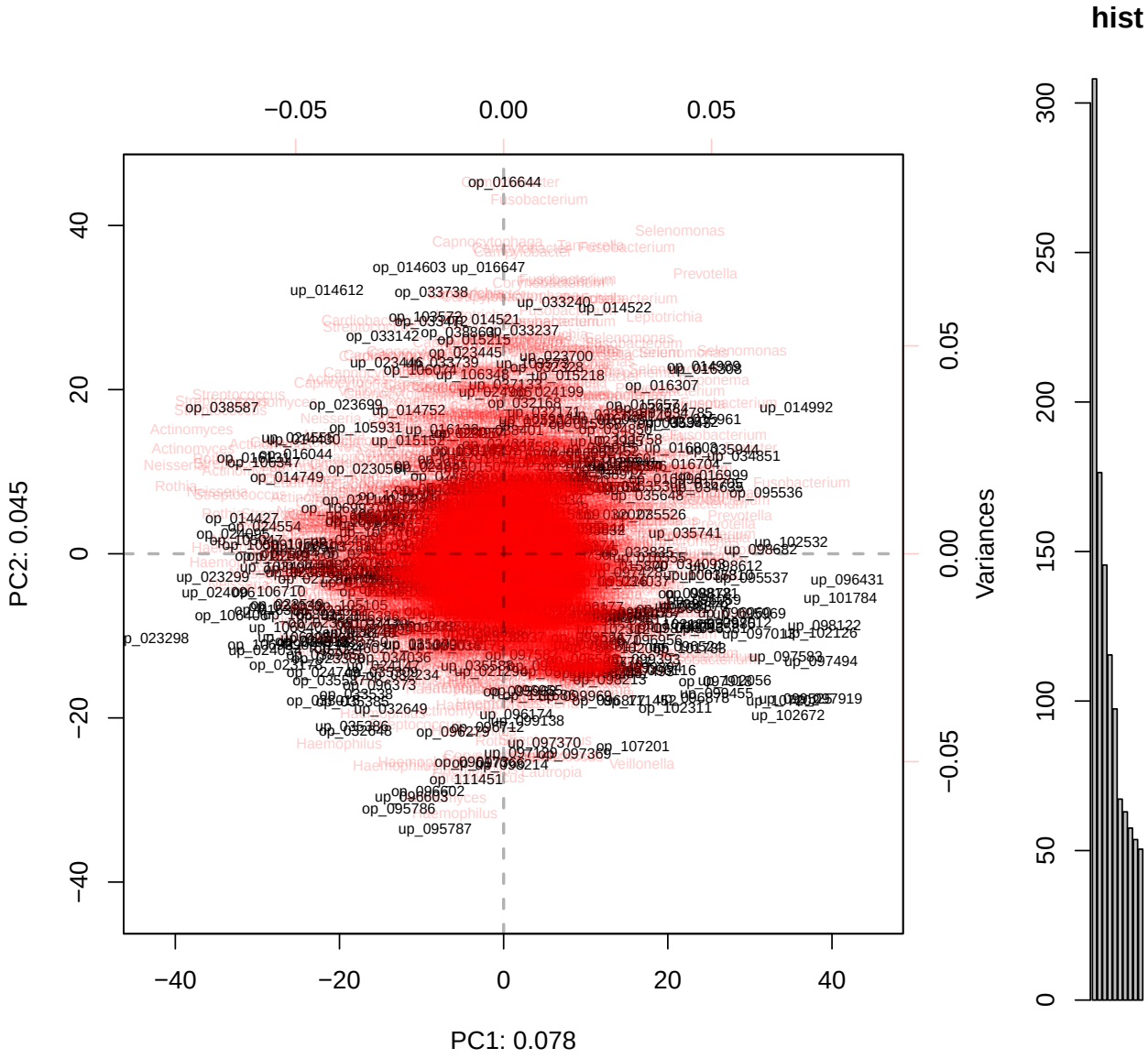

Figure 3: A compositional biplot of sub and supra gingival plaque for comparison. We can see that the op and up samples separate poorly, and the proportion of variance explained on component 1 is far less than for the first dataset. Furthermore, we can see the genus names of some of the OTUs that are driving this divide. Finally, while C component 1 has much more variance than component 2, the reduction in variance appears to be a linear trend until at least component 6 is reached. This biplot thus does not suggest a simple two-part comparison.

References

Aitchison, John, and Michael Greenacre. 2002. "Biplots of Compositional Data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51 (4). Wiley Online Library: 375–92.

Gorvitovskaia, Anastassia, Susan P Holmes, and Susan M Huse. 2016. "Interpreting Prevotella and Bacteroides as Biomarkers of Diet and Lifestyle." *Microbiome* 4: 15. doi:10.1186/s40168-016-0160-7.

Proctor, Lita M. 2011. "The Human Microbiome Project in 2011 and Beyond." *Cell Host Microbe* 10 (4): 287–91. doi:10.1016/j.chom.2011.10.001.