First comparison

gg

06 February, 2018

To run this file: Rscript -e "rmarkdown::render('first_comparison.Rmd')" ## Types of data

Comparison of 'differential abundance' is problematic for compositional data (Fernandes et al. 2013, 2014). Since the apparent abundance of every value depends on the apparent abundance of every other value, we can get into real difficulties if we are not careful. Take the simple example where we have two samples. The samples contain the following counts for five taxa:

$A = [1000, 100, 50, 250]$

and $B = [10, 500, 250, 1250]$.

We want to answer the question: Have the abundances of the taxa changed?

We sequence, and have a total count of about 100 (it is a first generation machine!)

So we get: $A_s = [71, 7, 4, 18], B_s = [1, 25, 12, 62]$

Note that these values appear to be very different between the groups. However, if we take one taxon as a reference, say taxon 4, and determine a ratio, i.d.:

$A_r = [74/18, 7/18, 4/18] = [4.1, 0.39, 0.22]$

$B_r = [1/62, 25/62, 12/62] = [0.02, 0.40, 0.20]$

Here we can see that if we assume one taxon is constant (taxon 4), then the last two are seen to be very similar in abundancd. Now we can infer that the majority of change is in the first taxon. We cannot compare the last taxon because it is assumed to be constant, that is, the assumed change in the last taxon is 0. This approach is the one used by ANCOM, a recently developed tool to assess change in microbiome datasets (Mandal et al. 2015).

Since we cannot know which taxon, if any, is constant, we can assume that a large number of the taxa exhibit only random changd. Then rather than using one taxon as a reference we can use the geometric mean abundance of all taxa. Note: this approach works poorly if there are only a small number of taxa (less than about 50) or if the taxa are very asymmetrically distributed between groups. This approach is the one used by ALDEx2 (Fernandes et al. 2013, 2014), and is the method that we will use.

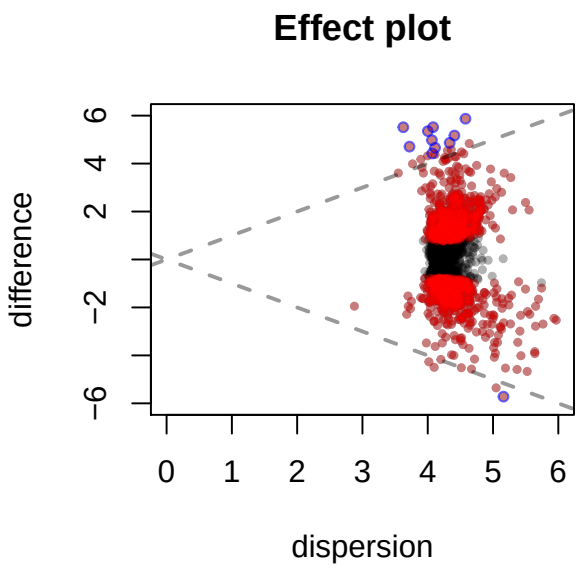
One complication is that a genometric mean cannot be determined if any of the values have a count of 0. For pairwise comparisons

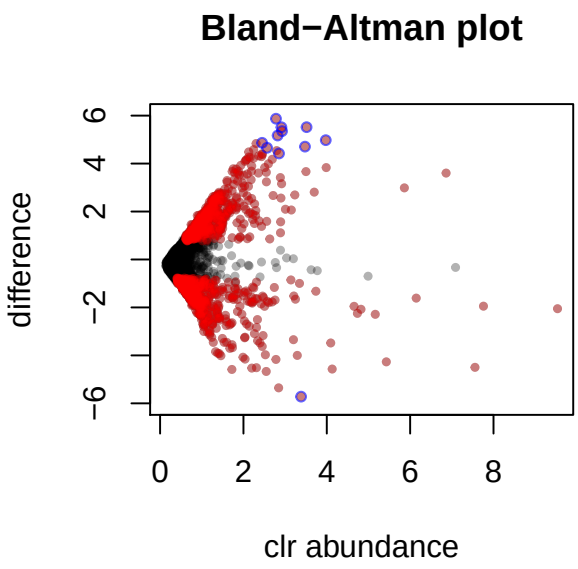
```
library(ALDEx2)

# read the dataset
d.subset <- read.table("data/ak_vs_op.txt", row.names=1, header=T)
# make a vector containing the two names of the conditions
# in the same order as in the column names
d.conds <- c(rep("ak", length(grep("ak", colnames(d.subset))) ), rep("op", length(grep("op", colnames(d
# generate Monte-Carlo instances of the probability of observing each count
# given the actual read count and the observed read count.
# use a prior of 0.5, corresponding to maximal uncertainty about the read count
# this returns a set of clr values, one for each mc instance
# note that the latest version of ALDEx2 requires conditions explicitly
d.x <- aldex.clr(d.subset, conds=d.conds, mc.samples=128)
# calculate effect sizes for each mc instance, report the expected value
```

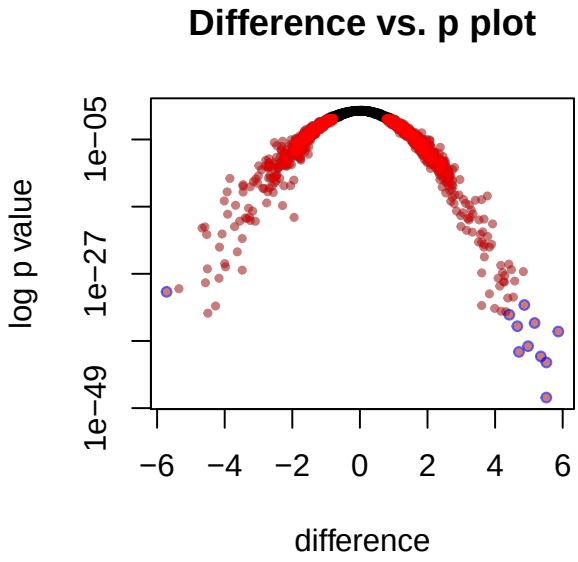
```
d.eff <- aldex.effect(d.x, d.conds)
# perform parametric or non-parametric tests for difference
# report the expected value of the raw and BH-corrected P value
d.tt <- aldex.ttest(d.x, d.conds)
# concatenate everything into one file
d.all <- data.frame(d.eff,d.tt)

#### this can be slow, so I have pre-computed and saved the file
```

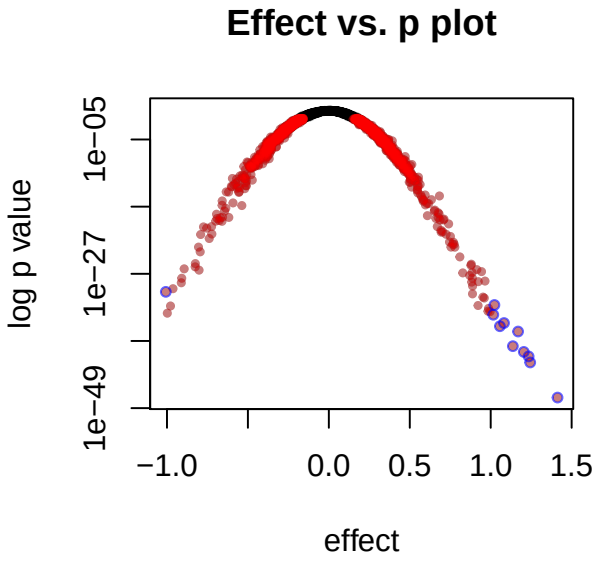
We will display the results using a number of different plots to show how each plot gives a different way of exploring the data. The mainstay that we advocate is the effect plot (Gloor, Macklaim, and Fernandes 2016), that plots the constituents of normalized change, or effect size.

```
x.all <- read.table("data/ak_vs_op_aldex.txt", header=T, row.names=1)

# get 'significant' set
sig <- x.all$wi.eBH < 0.05
eff <- abs(x.all$effect) > 1

# plot all in transparent grey
# low BH-corrected p values as red
# effect sizes > 1 as blue+red
par(fig=c(0,1,0,1), new=TRUE)

par(fig=c(0,0.5,0.5,1), new=TRUE)
plot(x.all$diff.win, x.all$diff.btw, col=rgb(0,0,0,0.3), pch=19,
    cex=0.5, ylim=c(-6,6), xlim=c(0,6), xlab="dispersion", ylab="difference",
    main="Effect plot")
points(x.all$diff.win[sig], x.all$diff.btw[sig], col=rgb(1,0,0,0.3), pch=19, cex=0.5 )
points(x.all$diff.win[eff], x.all$diff.btw[eff], col=rgb(0,0,1,0.6), pch=21, cex=0.7 )
abline(0,1, lty=2, lwd=2, col=rgb(0,0,0,0.4))
abline(0,-1, lty=2, lwd=2, col=rgb(0,0,0,0.4))

par(fig=c(0.5,1,0.5,1), new=TRUE)
plot(x.all$rab.all, x.all$diff.btw, col=rgb(0,0,0,0.3), pch=19,
    cex=0.5, ylim=c(-6,6), xlab="clr abundance", ylab="difference",
    main="Bland-Altman plot")
points(x.all$rab.all[sig], x.all$diff.btw[sig], col=rgb(1,0,0,0.3), pch=19, cex=0.5 )
points(x.all$rab.all[eff], x.all$diff.btw[eff], col=rgb(0,0,1,0.6), pch=21, cex=0.7 )

par(fig=c(0,0.5,0,0.5), new=TRUE)
plot(x.all$diff.btw, x.all$wi.ep, col=rgb(0,0,0,0.3), pch=19,
    cex=0.5, xlab="difference", ylab="log p value",
    main="Difference vs. p plot", log="y")
points(x.all$diff.btw[sig], x.all$wi.ep[sig], col=rgb(1,0,0,0.3), pch=19, cex=0.5 )
points(x.all$diff.btw[eff], x.all$wi.ep[eff], col=rgb(0,0,1,0.6), pch=21, cex=0.7 )

par(fig=c(0.5,1,0,0.5), new=TRUE)
plot(x.all$effect, x.all$wi.ep, col=rgb(0,0,0,0.3), pch=19,
    cex=0.5,  xlab="effect", ylab="log p value",
    main="Effect vs. p plot", log="y")
points(x.all$effect[sig], x.all$wi.ep[sig], col=rgb(1,0,0,0.3), pch=19, cex=0.5 )
points(x.all$effect[eff], x.all$wi.ep[eff], col=rgb(0,0,1,0.6), pch=21, cex=0.7 )
```


Effect plot
difference
dispersion
6
4
2
−2
−6
0
1
2
3
4
5
6


Bland−Altman plot
difference
clr abundance
6
4
2
−2
−6
0
2
4
6
8


Difference vs. p plot
log p value
difference
1e−05
1e−27
1e−49
−6
−4
−2
0
2
4
6


Effect vs. p plot
log p value
effect
1e−05
1e−27
1e−49
−1.0
0.0
0.5
1.0
1.5

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

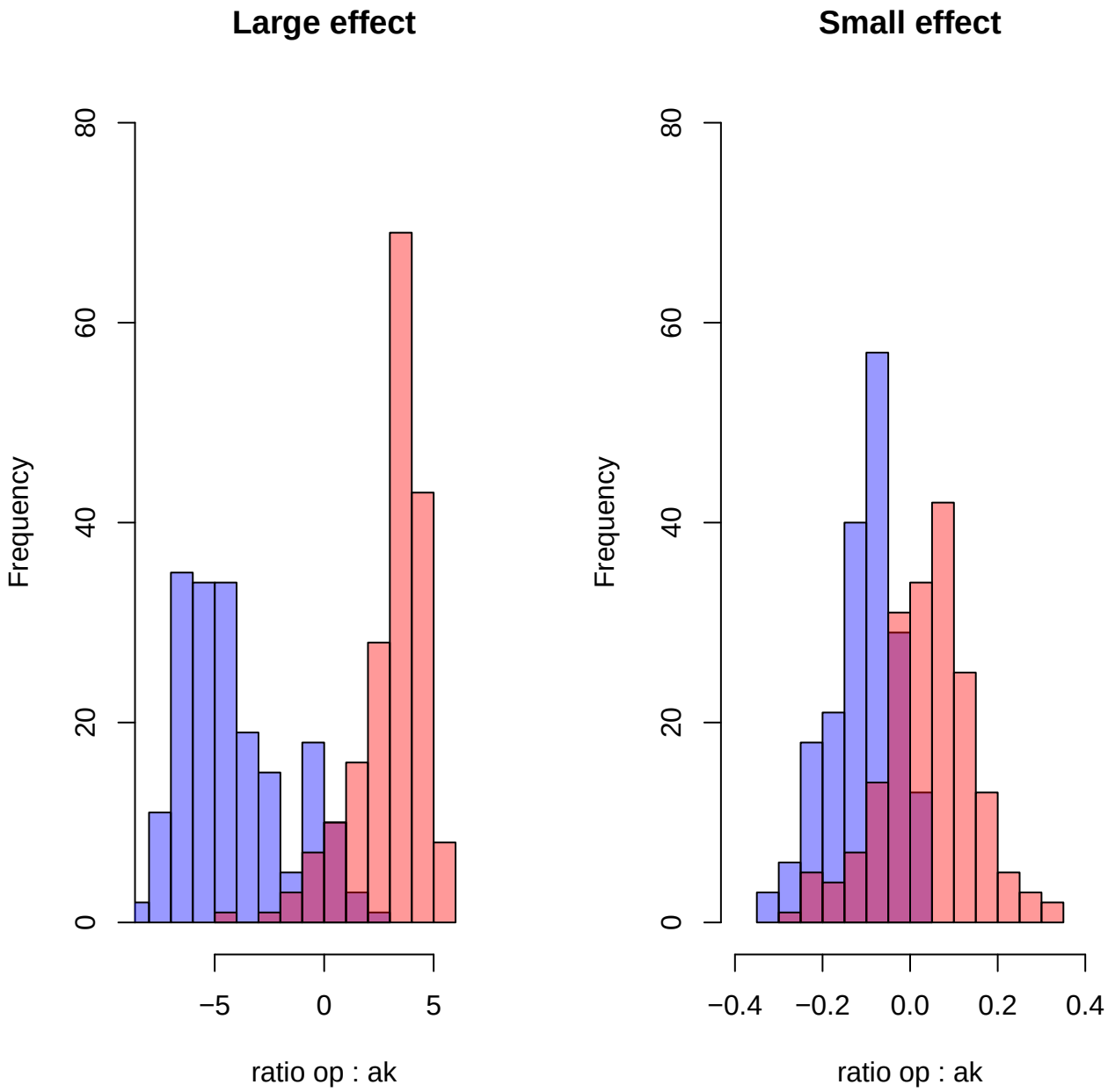


Figure 1: Histograms showing the separation between groups when choosing OTUs with large effect sizes (left), or OTUs with small effect size (right). OTUs with the largest effect are the 'most reliably different' between groups, and should be chosen over those that are 'most significantly different' whenever possible.

#References

Altman, D. G., and J. M. Bland. 1983. "Measurement in Medicine: The Analysis of Method Comparison Studies." *Journal of the Royal Statistical Society. Series D (the Statistician)* 32 (3). Wiley for the Royal Statistical Society:pp. 307–17. http://www.jstor.org/stable/2987937.

Cui, Xiangqin, and Gary A Churchill. 2003. "Statistical Tests for Differential Expression in cDNA Microarray Experiments." *Genome Biol* 4 (4):210.1–210.10.

Fernandes, Andrew D, Jean M Macklaim, Thomas G Linn, Gregor Reid, and Gregory B Gloor. 2013. "ANOVA-Like Differential Expression (Aldex) Analysis for Mixed Population Rna-Seq." *PLoS One* 8 (7):e67019. https://doi.org/10.1371/journal.pone.0067019.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. 2014. "Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S RRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome* 2:15.1–15.13. https://doi.org/10.1186/2049-2618-2-15.

Gloor, Gregory B., Jean M. Macklaim, and Andrew D. Fernandes. 2016. "Displaying Variation in Large Datasets: A Visual Summary of Effect Sizes." *Journal of Computational and Graphical Statistics* 25 (3):971–9.

Halsey, Lewis G, Douglas Curran-Everett, Sarah L Vowler, and Gordon B Drummond. 2015. "The Fickle P Value Generates Irreproducible Results." *Nat Methods* 12 (3):179–85. https://doi.org/10.1038/nmeth.3288.

Mandal, Siddhartha, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. 2015. "Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition." *Microb Ecol Health Dis* 26:27663.

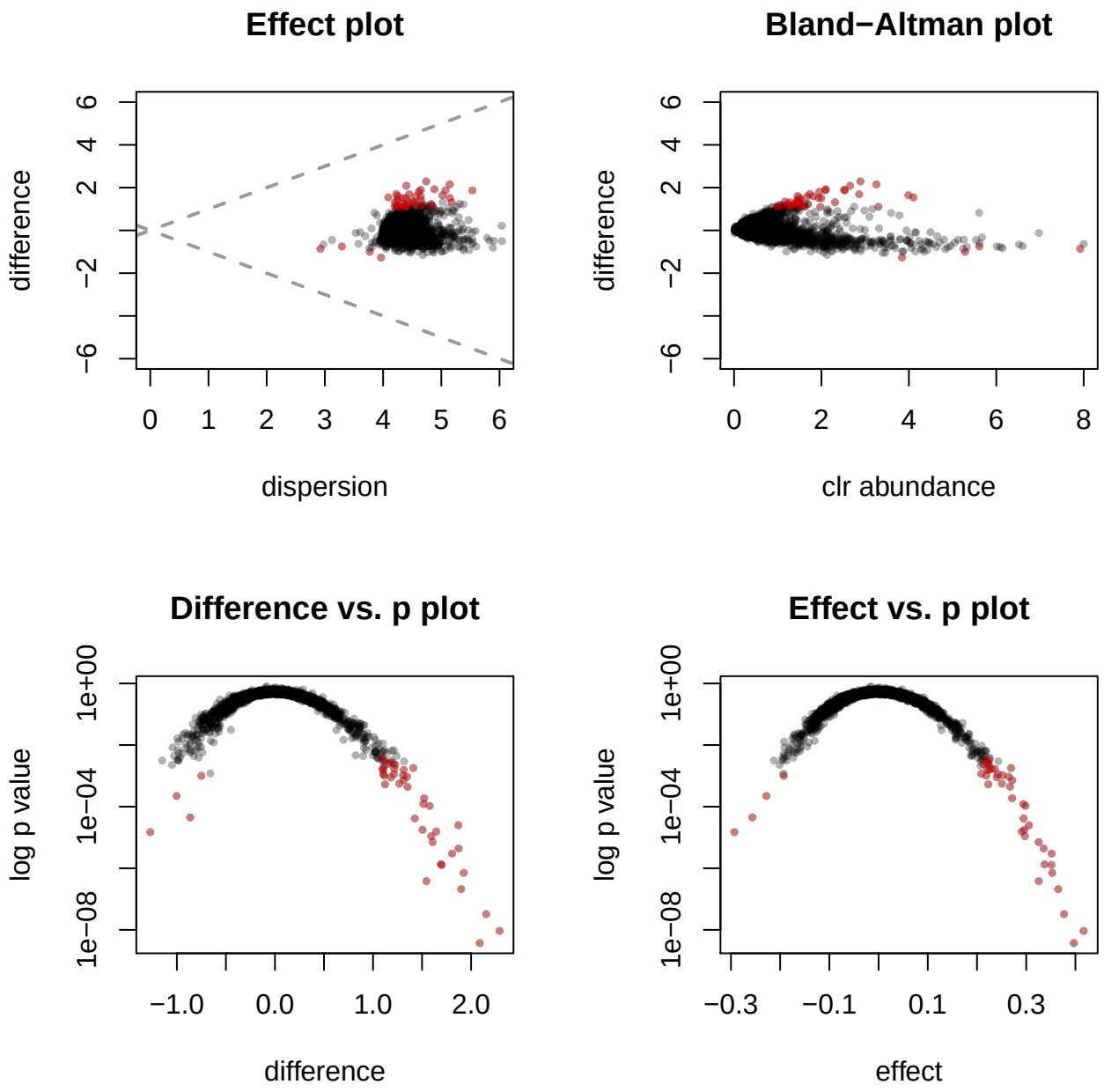


Figure 2: The same plots for the supra and subgingival plaque samples. We see that we have statistical significance, but the biological relevance is difficult to defend because of the very small effect sizes.