First comparison

gg

05 April, 2018

To run this file: Rscript -e "rmarkdown::render('ALDEx_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 features:

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

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

We want to answer the question: Have the abundances of the features 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 feature as a reference, say feature 4, and determine a ratio, i.e.:

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

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

Here we can see that if we assume one feature is constant (feature 4), then the last two are seen to be very similar in abundance. Now we can infer that the majority of change is in the first feature. We cannot compare the last feature because it is assumed to be constant, that is, the assumed change in the last feature 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 feature, if any, is constant, we can assume that a large number of the features exhibit only random change. Then rather than using one feature as a reference we can use the geometric mean abundance of all features. Note: this approach works poorly if there are only a small number of features (less than about 50) or if the features 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)
```

```
## Loading required package: methods
```

```
# read the dataset
d <- read.table("data/barton_agg.tsv", 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("SNF", length(grep("SNF", rownames(d))) ),
    rep("WT", length(grep("WT", rownames(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
# this workflow can take several minutes

# note that the latest version of ALDEx2 requires conditions explicitly
d.x <- aldex.clr(t(d), conds=d.conds, mc.samples=128)
```

```
## [1] "operating in serial mode"
## [1] "computing center with all features"
```

```
# calculate effect sizes for each mc instance, report the expected value
d.eff <- aldex.effect(d.x, d.conds, include.sample.summary=TRUE)
```

```
## [1] "operating in serial mode"
## [1] "sanity check complete"
## [1] "rab.all  complete"
## [1] "rab.win  complete"
## [1] "rab of samples complete"
## [1] "within sample difference calculated"
## [1] "between group difference calculated"
## [1] "group summaries calculated"
## [1] "effect size calculated"
## [1] "summarizing output"
```

```
# 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)
```

```
## [1] "running tests for each MC instance:"
## |-----------(25%)----------(50%)----------(75%)----------|
```

```
# concatenate everything into one file
x.all <- data.frame(d.eff,d.tt)
```

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.

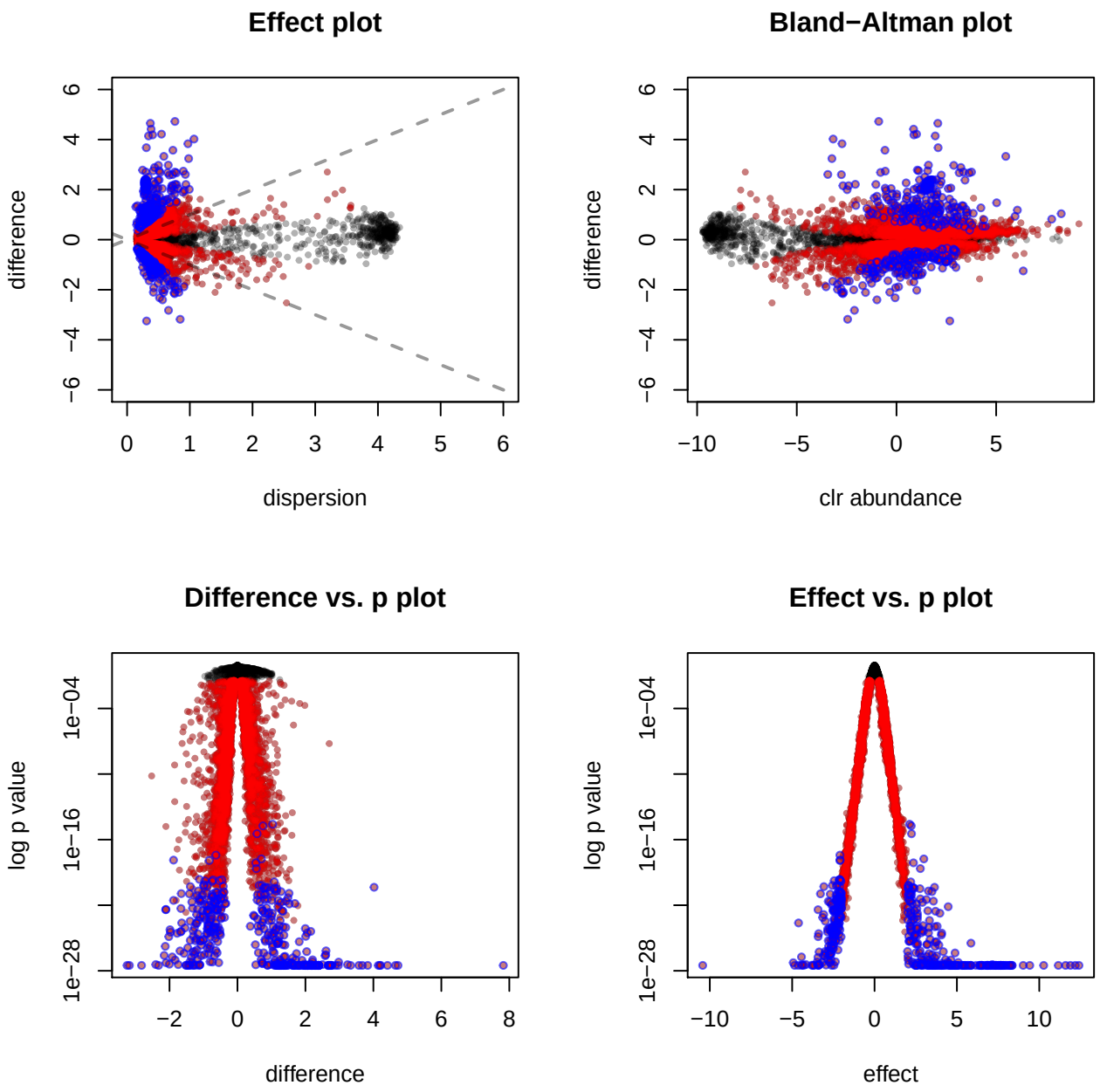

Figure 1: Plotted here are features with no difference between groups (grey), a statistically difference between groups (red), and with an effect larger than 2 (blue circles). These are plotted using different plots (described clockwise from top left). The effect plot (Gloor, Macklaim, and Fernandes 2016) illustrates the difference between groups vs. the dispersion (variance) within groups. If the effect is greater than one (outside the grey lines), then, on average the features are obviosly separable by eye when plotted; roughly, they would be seen to have a greater difference between groups than the pooled standard deviation. Effect is a more robust measure of difference than are P values, since the latter depend on sample size; large sample sizes will always give low P values (Halsey et al. 2015). We can see this here where the large sample size means that even highly variable OTUs are significantly different. The Bland-Altman plot (Altman and Bland 1983) compares difference and abundance, and is often seen in RNA-Seq data. The Volcano plot (Cui and Churchill 2003) shows the association between difference and P value, and the final plot shows the association between effect and P value.

The effect sizes can be understood as a measure of separability between groups for each feature. Plotted in the figure are features with different effect sizes and the corresponding adjusted p value is included. In RNA-seq data, we have found that an effect size cutoff between 1 and 2 (Macklaim et al. 2013) gives reliable results in meta-transcriptome analyses.

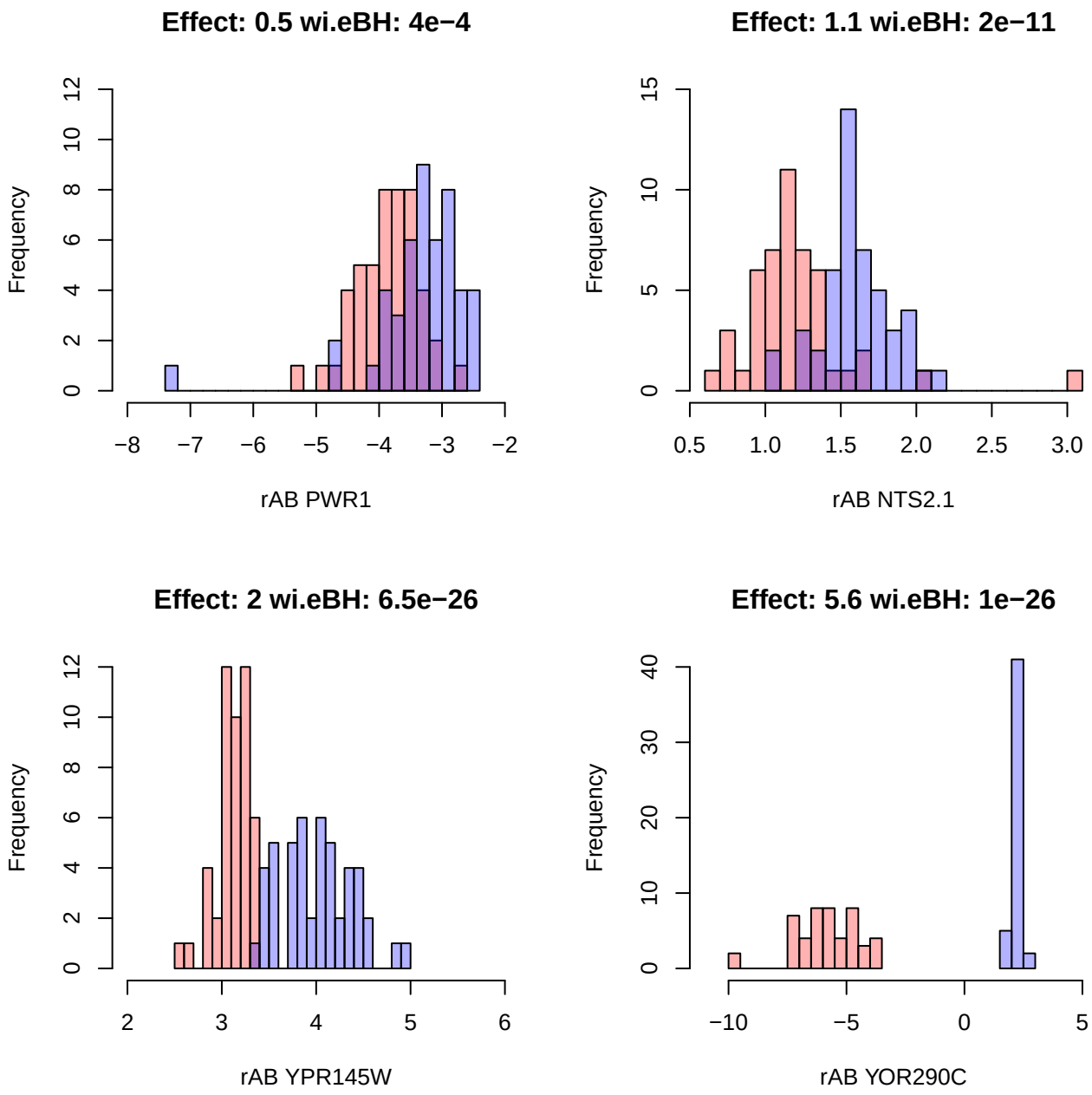

Figure 2: Histograms showing the separation between groups when choosing features withdiffering effect sizes. Features with the largest effect are the most reliably different between groups, and should be chosen over those that are most significantly different whenever possible. Note that an effect size of 0.5 is more than sufficient to give a significant difference but the separation between groups is marginal.

#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.

Macklaim, M Jean, D Andrew Fernandes, M Julia Di Bella, Jo-Anne Hammond, Gregor Reid, and Gregory B Gloor. 2013. "Comparative Meta-RNA-Seq of the Vaginal Microbiota and Differential Expression by *Lactobacillus Iners* in Health and Dysbiosis." *Microbiome* 1:15. https://doi.org/doi: 10.1186/2049-2618-1-12.

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.