Choosing a denominator

Greg Gloor

14 August, 2017

About this document

This document is an .Rmd document and part of the documentation chain for ALDEx2:

The document is the supplement and companion to the "Microbiome datasets are compositional: and this is not optional." review article. This document contains interspersed markdown and R code that is compiled into a pdf document that supports the figures and assertions in the main article. R code is exposed in the pdf document.

From an R command prompt you can compile this document into PDF if you have LaTeXand pandoc installed:

`rmarkdown::render("high_abund_low_var.Rmd")` or you can open the file in RStudio and compile in that environment.

The first code block loads the data and sets the analysis parameters and the list of genes that will be assumed to be housekeeping functions. These are functions that are assumed invariant in relative abundance, and if we were conducting a qPCR, would be reasonable functions to choose. For this analysis, we assume that glycolytic or ribosomal protein functions are invariant between conditions.

```r
# load the libraries and data
library(ALDEx2)

# dataset: metatranscriptome of vaginal samples
e.min <- read.table("data/twntyfr.txt",
  header=T, row.names=1, check.names=F,
  sep="\t", comment.char="", quote="")

# the vector of conditions
conds <-c("H","H","H","H","B","H","B",
  "B","H","B","H","B","B","B","B",
  "B","B","B","H","B","H","H")

# standard housekeeping genes
# glycolsis or ribosomal protein genes
glycol <- c(2418,1392,1305,1306,2421,1049)
ribo <- c(grep("LSU", rownames(e.min)),
  grep("SSU", rownames(e.min)))
```

The problem with a pre-chosen set of housekeeping functions is that they may or may not be constant in each sample and so we want a non-arbitrary way to choose the relatively constan functions. The qualities of a good housekeeping gene (or function) is that it be somewhat abundant so that it is easy to measure, and that it not vary much between the samples so that it serves as a good internal standard. We chose to identify those functions with variance in the bottom quartile and relative abundance in the top quartile *in all conditions*. The idea here is to find functions that are somewhat invariant across all conditions. Note that we tried this on a global basis with less success.

The second code block is an `R` function, `find_lowvar_high_abund` that identifies these molecular functions and places the offsets into the vector `invariant.set`.

```r
find_lowvar_highabund <- function(data, conds)
{
  invariant.set.list <- vector("list",
  length(unique(conds)))

  # clr transform
  reads.clr <- t(apply(data + 0.5, 2,
  function(x){log2(x) - mean(log2(x))}))

  # per-condition offsets found
  for(i in 1:length(unique(conds))){
    these.rows <- which(conds ==
      unique(conds)[i])

  # find the least variable
    reads.var <- apply(reads.clr[these.rows,],
    2, function(x){var(x)})
    var.set <- which(reads.var
      < quantile(unlist(reads.var))[2])

    # find the most relative abundant
    # top quartile
    reads.abund <- apply(reads.clr[these.rows,],
      2, sum)
    abund.set <- which(reads.abund >
      quantile(reads.abund)[4])

    invariant.set.list[[i]] <-
      intersect(var.set, abund.set)
  }

  # get the intersect of all conditions
  # successive operations on the list elements

  invariant.set <-
    Reduce(intersect, invariant.set.list)
  return(invariant.set)
}

invariant.set <-
  find_lowvar_highabund(e.min, conds)
```

This third code block generates the ALDEx2 output for the non-centred dataset. This is the default method in ALDEx2. We then plot the result; the plotting code is not shown for brevity but is in `R_block_3Fig`.

```
x <- aldex.clr(e.min, conds=conds)
x.e <- aldex.effect(x, conds)
```

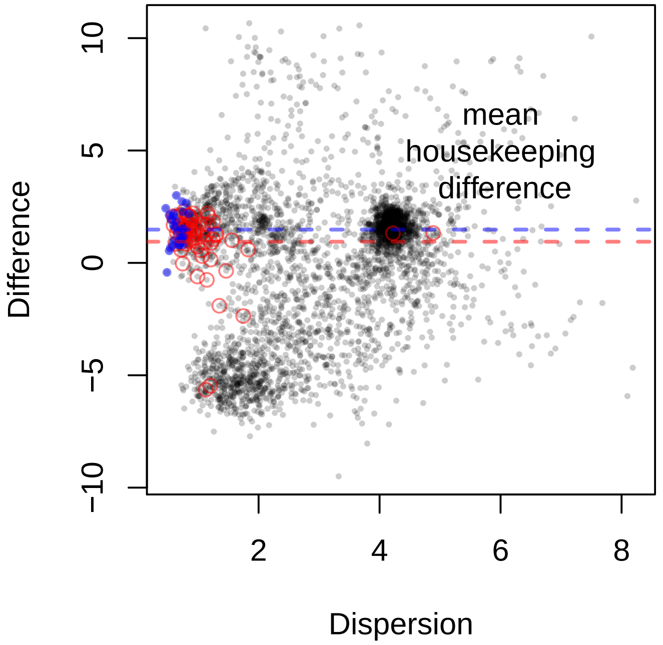

Figure 1: A metatranscriptome asymmetric sparse dataset is not centred properly. The X-axis plots the dispersion of the data (diff.win, akin to the standard deviation) as calculated by ALDEx2, the Y-axis plots the difference between the groups as calculated by ALDEx2 (diff.btw). Each point is an individual SEED function, and are colored in grey if not of interest. Points in red are abundant functions with low variance in group 1, points in brown are abundant functions with low variance in group 2, and points in cyan are abundant functions with low variance in both groups. The pre-selected set of housekeeping functions are circled in black.

We can see in Figure 1 that the mean difference between groups for any of the sets of housekeeping functions are not on the 0 line, and that the per-condition mean values are the most extreme, suggesting that either one would be a poor choice for the other set. However, the joint set in cyan, and the pre-chosen basket of housekeeping functions are closer to the perceived centre of the data, with the joint set seemingly most appropriate.

We now re-calculate the ALDEx2 output using the joint set of invariant functions held in the vector `invariant.set` in code block 4.

```
x.invar <- aldex.clr(e.min, conds=conds,
  denom=invariant.set)
x.e.invar <- aldex.effect(x.invar, conds)
```

After plotting we can see that the housekeeping functions are better centered, and the pre-chosen basket of functions (ribosomal and glycolytic) in black and the automatically chosen set in cyan being closest to the center.

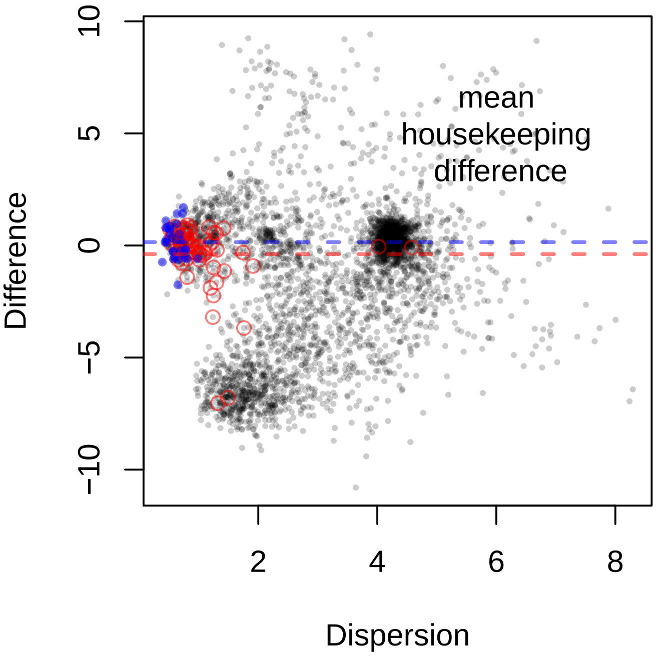

Figure 2: Re-centering the metatranscriptome asymmetric sparse dataset. Colors are as in Figure 1

Another way to look at the whether the mid-point of the set of functions is near 0 is to plot the density of the difference between groups for each function in each set. Figure 3 shows a density plot with the same colors as in Figures 1 and 2. Here we can see that the group 1 functions (in red) are substantially off center, and the group 2 (orange) and the ribosomal and glycolytic functions (black) have a large density centered at approximately 0, but both have a density in the left tail that changes the location of the mean to be less than 0. The density of the automatically chosen set is approximately symmetrical around the 0.

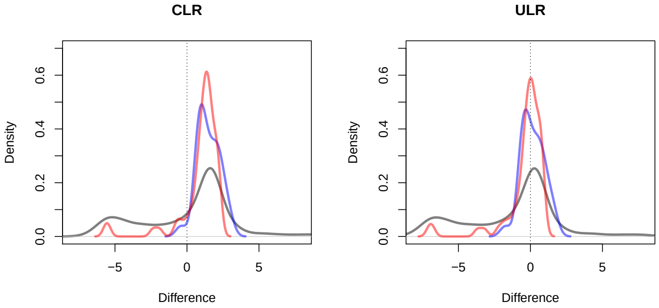

Figure 3: Density plot of difference between for different sets of housekeeping functions.