Exploratory analysis of RNA-seq dataset

gg

2015-11-18

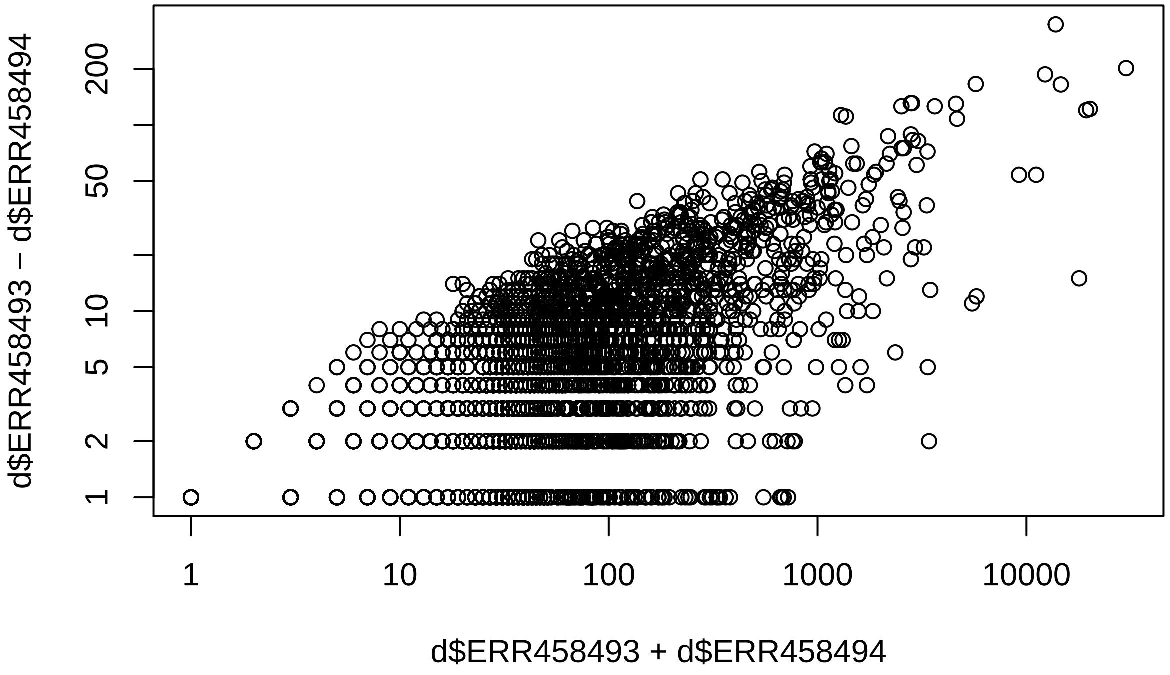
To run this file: Rscript -e "rmarkdown::render('biplots_4.Rmd')"

Exploratory data analysis to determine if technical replication is tight in the Barton dataset

try a couple of scatter plots, and then a biplot

```
# load the compositional framework
library(compositions)
library(zCompositions)
# this is the metadata
meta <- read.table("ERP004763_sample_mapping.tsv", row.names=1,
    header=T)

# this is the table of counts
# should have 672 samples and 6349 genes
d <- read.table("countfinal2.file", header=T, row.names=1)

# plot the difference vs. sum on an absolute scale
# log axes
plot(d$ERR458493+d$ERR458494,d$ERR458493-d$ERR458494,log="xy")
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 523 x values <= 0 omitted
## from logarithmic plot

## Warning in xy.coords(x, y, xlabel, ylabel, log): 3471 y values <= 0 omitted
## from logarithmic plot
```



The correlation between the first two technical replicates is very good being 0.9994563.

However this is tedious, and I'm bored.

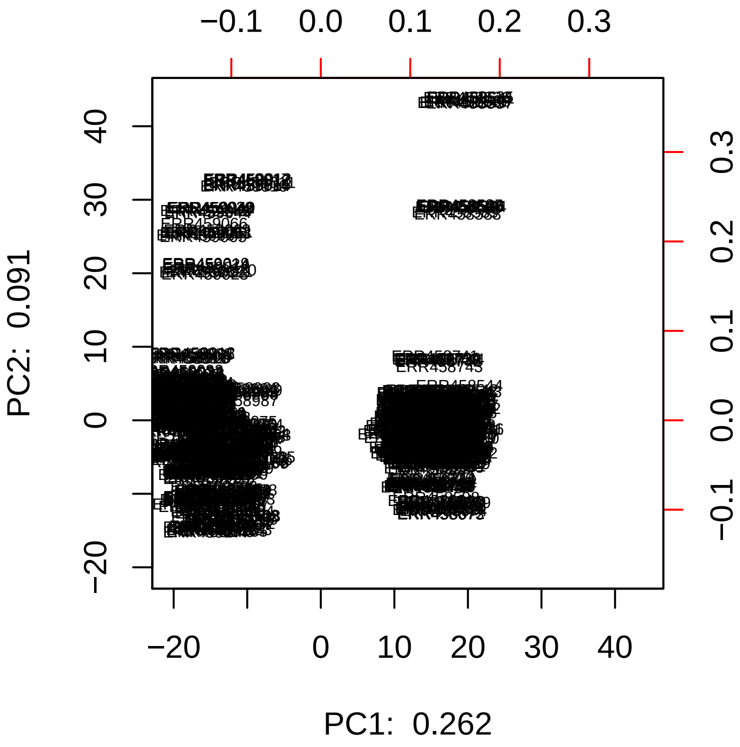
Let's plot all associations between all samples technical and biological on a biplot

For this we need to use a compositional paradigm, since the data have a constant, arbitrary sum.

```
# do we have any throwaway genes
# i.e., structural 0's - 0 in every sample

# apply is determining the row sums
# which is testing the rows for a greater than 0
# we are subsetting the original data to keep rows with sum >0
d.n0 <- d[which(apply(d,1,sum) > 0),]
# this removes 13 rows

# we need to replace 0 values with a best estimate
# use zCompositions CZM by default
# but samples must be by row, so use t()

d.n0.CZM <- cmultRepl(t(d.n0), label=0, method="CZM",
    output="counts")
```

```
## No. corrected values:  2167
```

```
# turn this into a centered log-ratio transform
# samples are by row
# remember that apply by row rotates the data
# R is terrible, so we need to use t() again
d.clr <- t(apply(d.n0.CZM, 1, function(x){log(x)-mean(log(x))}))

# now we can generate the PCA object
pcx <- prcomp(d.clr)
mvar.clr <- mvar(d.clr)

biplot(pcx, scale=0, cex=c(0.5,0.001), var.axes=FALSE,
 xlab=paste("PC1: ", round(sum(pcx$sdev[1]^2)/mvar.clr, 3)),
ylab=paste("PC2: ", round(sum(pcx$sdev[2]^2)/mvar.clr, 3)),
)
```



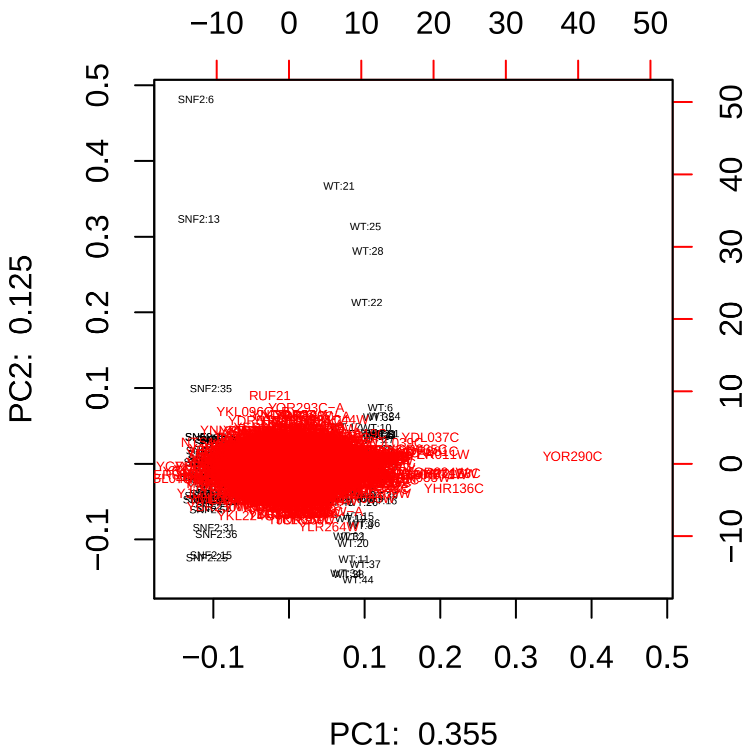
Do we have a good technical replication? we can explore this by observing if the clusters of 7 found in the biplot are technical, biological replicates. So for the cluster of 7 with a value greater than 40 in the biplot, these are 43, 44, 45, 46, 47, 48, 49, and are simply technical replicates. In a like manner you could explore to see if all clusters of 7 are close in biplot space.

We conclude that the technical replicates are typically very tight, and that we can simply sum across the lanes without introducing bias into the dataset (i.e. there is no strong lane effect)

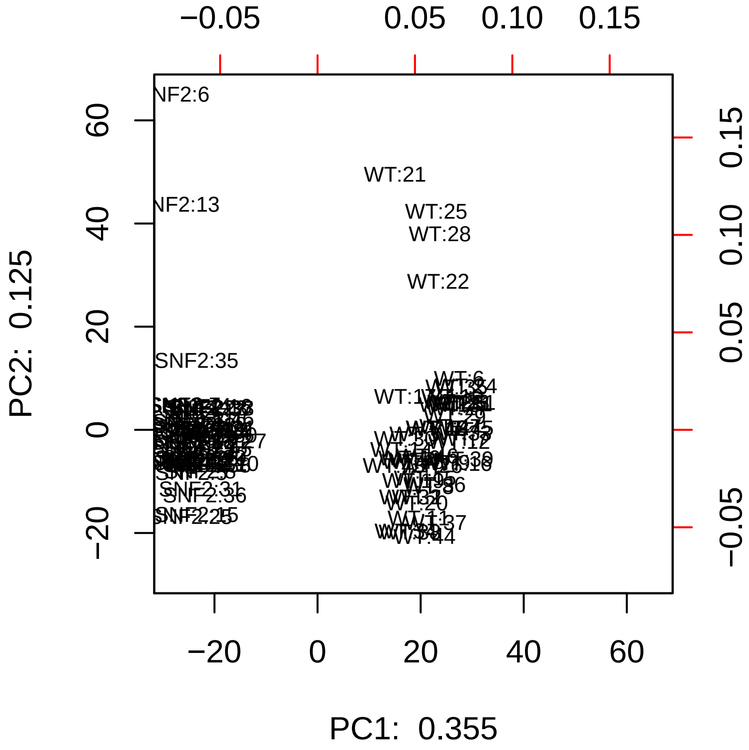
now to aggregate

```
# aggregate all replicates
# need a vector of names that are unique to each sample
# but common across technical replicates
# concatenate the sample and BiolRep values in the
# metadata
nms.agg <- paste(meta[,2], meta[,3], sep=":")

# make an aggregated dataset by sample
# aggregate common elements by some function
# needs samples by row
d.agg <- aggregate(t(d), by=list(nms.agg), FUN=sum)
rownames(d.agg) <- d.agg$Group.1
d.agg$Group.1 <- NULL

# remove columns with 0 counts
# i hate R, since it treats rows and columns differntly in apply
d.agg.gt0 <- d.agg[,apply(d.agg, 2, sum) > 0]

# estimate 0 values (zCompositions)
d.agg.n0 <- cmultRepl(d.agg.gt0, method="CZM", label=0, output="counts")
```

```r
d.agg.n0.clr <- t(apply(d.agg.n0, 1, function(x){log2(x) - mean(log2(x))}))

# SVD and metric variance
pcx.agg <- prcomp(d.agg.n0.clr)

mvar.agg.clr <- mvar(d.agg.n0.clr)

par(mfrow=c(1,1), mar=c(5,4,4,1))
# covariance biplot
# relationships between variables
biplot(pcx.agg, cex=c(0.3,0.4), var.axes=FALSE,
    xlab=paste("PC1: ", round(sum(pcx.agg$sdev[1]^2)/mvar.agg.clr, 3)),
    ylab=paste("PC2: ", round(sum(pcx.agg$sdev[2]^2)/mvar.agg.clr, 3)),
    scale=1)
```



```r
# form biplot
# relationships between samples
biplot(pcx.agg, cex=c(0.7,0.01), var.axes=FALSE,
    xlab=paste("PC1: ", round(sum(pcx.agg$sdev[1]^2)/mvar.agg.clr, 3)),
    ylab=paste("PC2: ", round(sum(pcx.agg$sdev[2]^2)/mvar.agg.clr, 3)),
    scale=0)
```


−0.05
0.05
0.10
0.15
NF2:6
NF2:13
SNF2:35
WT:21
WT:25
WT:28
WT:22
WT:6
SNF2:31
SNF2:36
SNF2:15
SNF2:25
WT:20
WT:37
WT:44
60
40
20
0
−20
0.15
0.10
0.05
−0.05
PC2: 0.125
−20
0
20
40
60
PC1: 0.355