# First biplot

*gg*

*04 April, 2018*

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

We will use as an example a transcriptome dataset (Schurch et al. 2016; Gierliński et al. 2015) containing 96 samples, 48 each from wt and SNF2 knockout strain These data have been filtered to include only those features 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 dataset. It shows, in one plot, the essences of your results. Do my samples separate into groups? features are driving this separation? what features 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.agg <- read.table("data/barton_agg.tsv", sep="\t", header=T, row.names=1)

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

# it is important to first filter to remove rows that are exclusively 0 values
d.filt <- codaSeq.filter(d.agg, min.count=1, min.prop=0, samples.by.row=FALSE)

# we are using the Count Zero Multiplicative approach
d.n0 <- cmultRepl(t(d.filt), method="CZM", label=0)

# generate the centered log-ratio transformed data
# samples by row
d.clr <- apply(d.n0, 2, function(x) log(x) - mean(log(x)))

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

# 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)
```

```r
plot(pcx, main="hist")
```

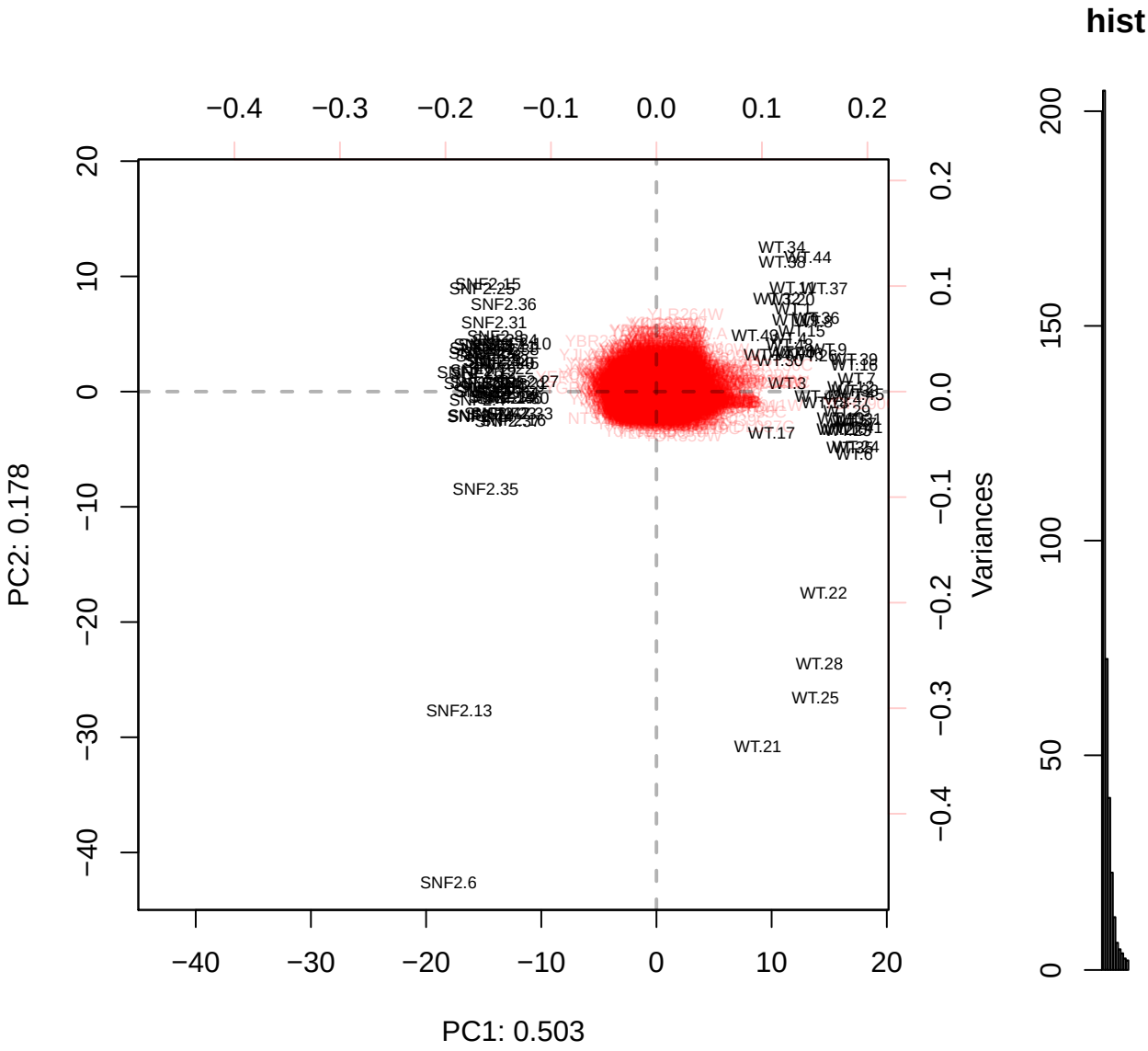
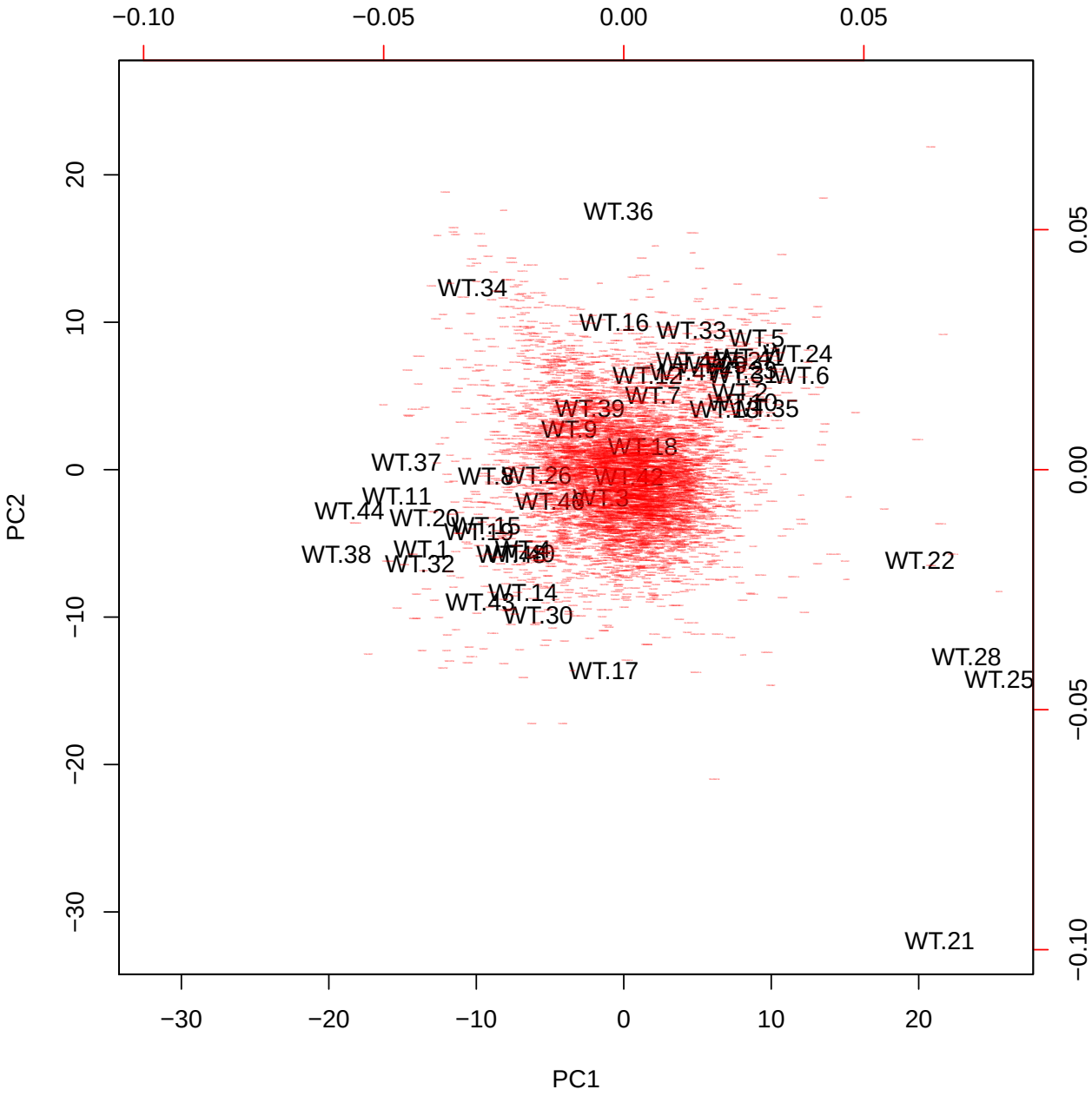


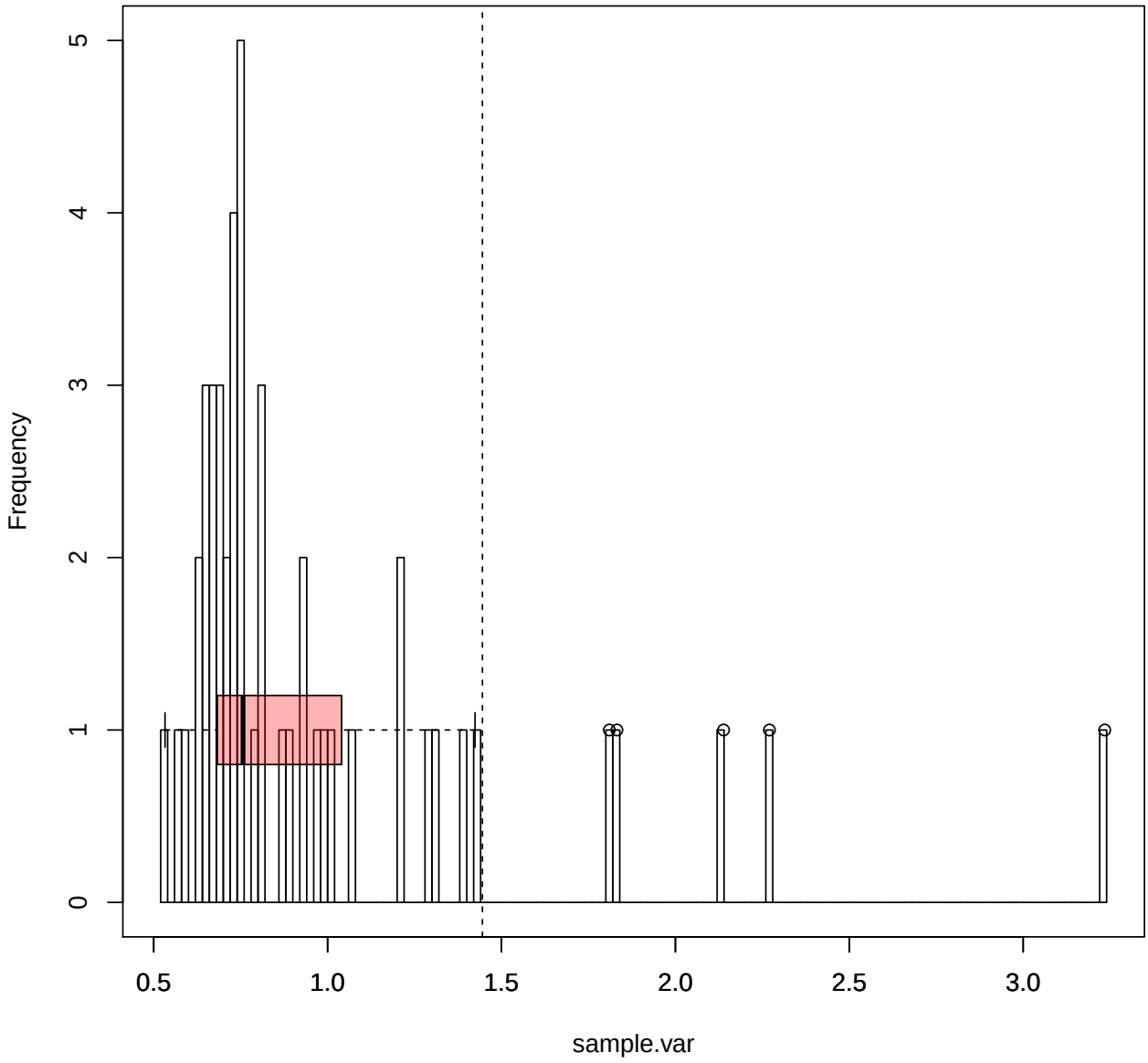

Figure 1: The compositional biplot is the workhorse tool for CoDa. This plot summarizes the entire analysis in a qualitative manner. We can see that the op and ak samples separate very well, although the proportion of variance explained on component 1 is small. Furthermore, we can see the genus names of some of the features 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.

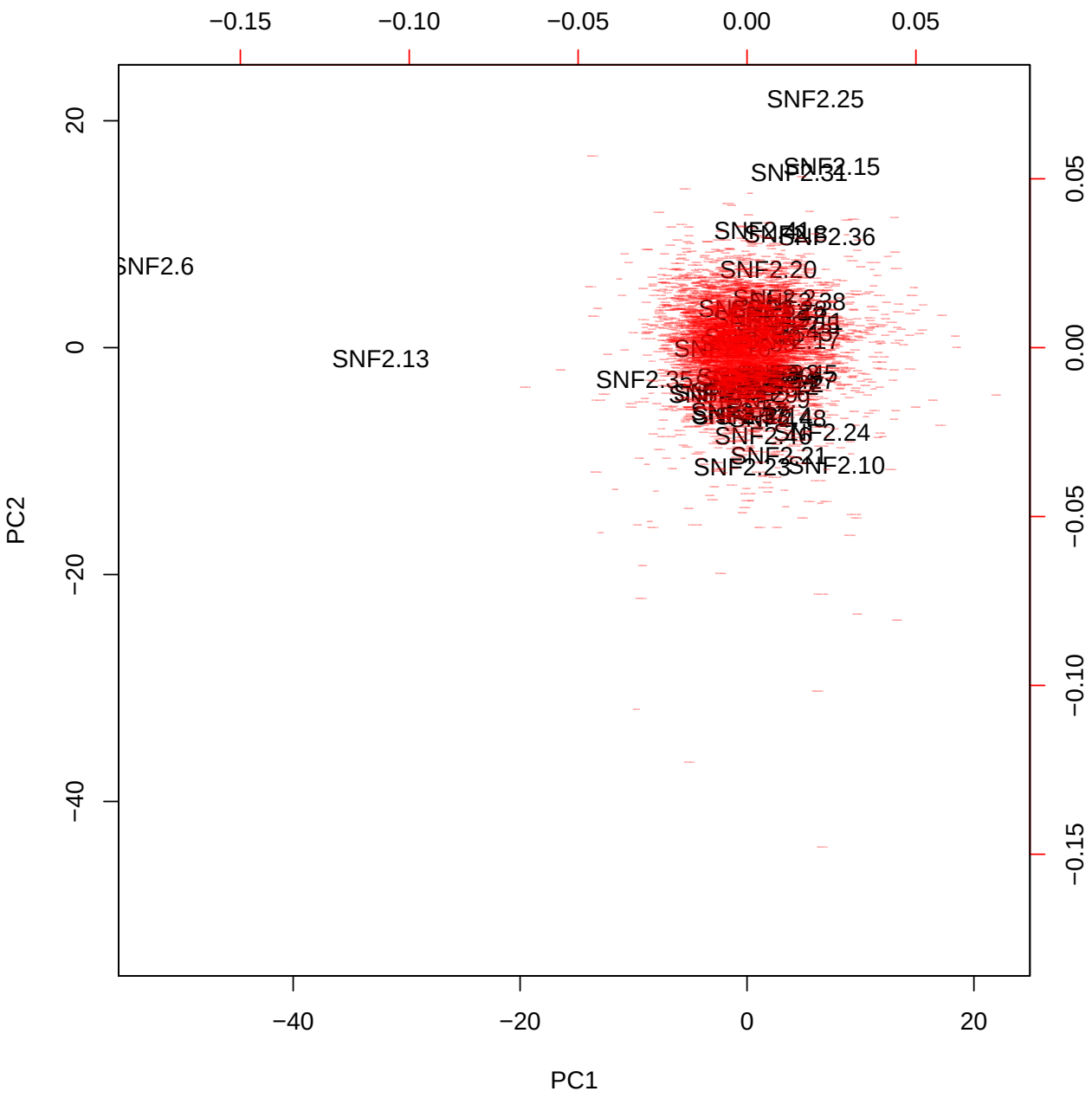
**Rules for interpreting compositional biplots:**

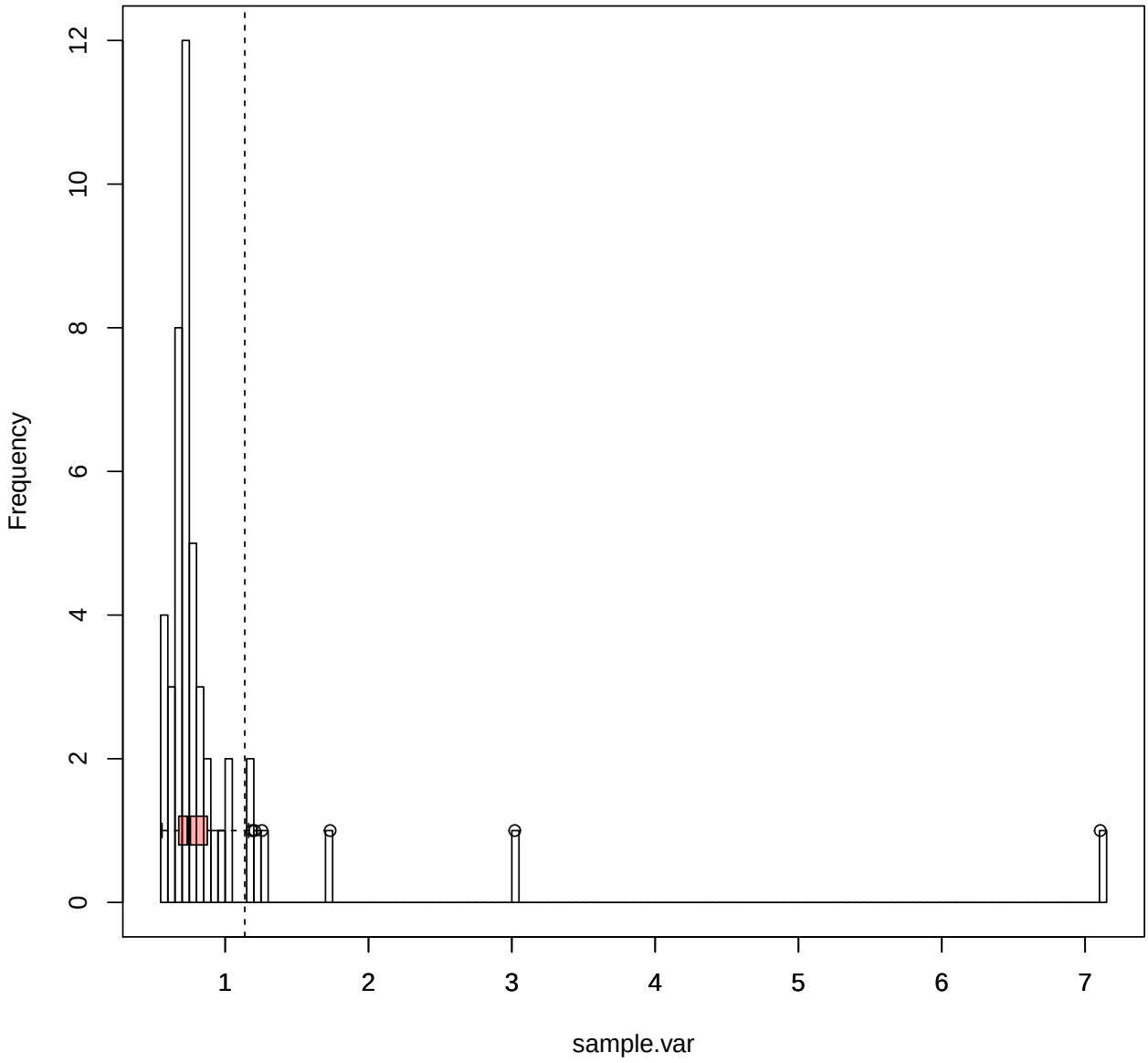
- All interpretations are up to the limit of the variance explained. We can think of this as a shadow of the multidimensional dataset (4545 dimensions!) projected onto two dimensions. If the variance explained is high ( $> 0.8$) then the edges of the shadows are sharp, however, if the variance explained is low, as it is here, then we have little confidence in the exact placement of any individual sample or feature.
- 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 features as ratios. Abundance information is not directly available on these plots.
- The distance and direction of an feature from the origin is the standard deviation of the ratio of that feature to the geometric mean of all features.
- The line between any set of features is called a link. Links that pass through more than one feature are permitted and do not change the interpretation.
- Short links indicate a constant or near constant ratio between the two (or more) linked features in the dataset. This dataset is too complex to identify links easily
- Long links indicate a non-constant ratio between the joined features, and define a ratio relationship that can be inverse or random. There is no principled method to determine which is the case.

We can see that there are a number of samples that appear to be outlier samples. Should we include SNF2.6 in the analysis or not? One of the messages of the Barton papers (Schurch et al. 2016; Gierliński et al. 2015) was that about 10% of samples, even carefully prepared samples can be outliers for unknown methodological reasons. We approach outliers by finding those samples that contribute more variance than expected to the variance of the group. Outliers are defined as those samples that contribute greater than the median plus twice the interquartile range of the sample variance to the total variance of the group.

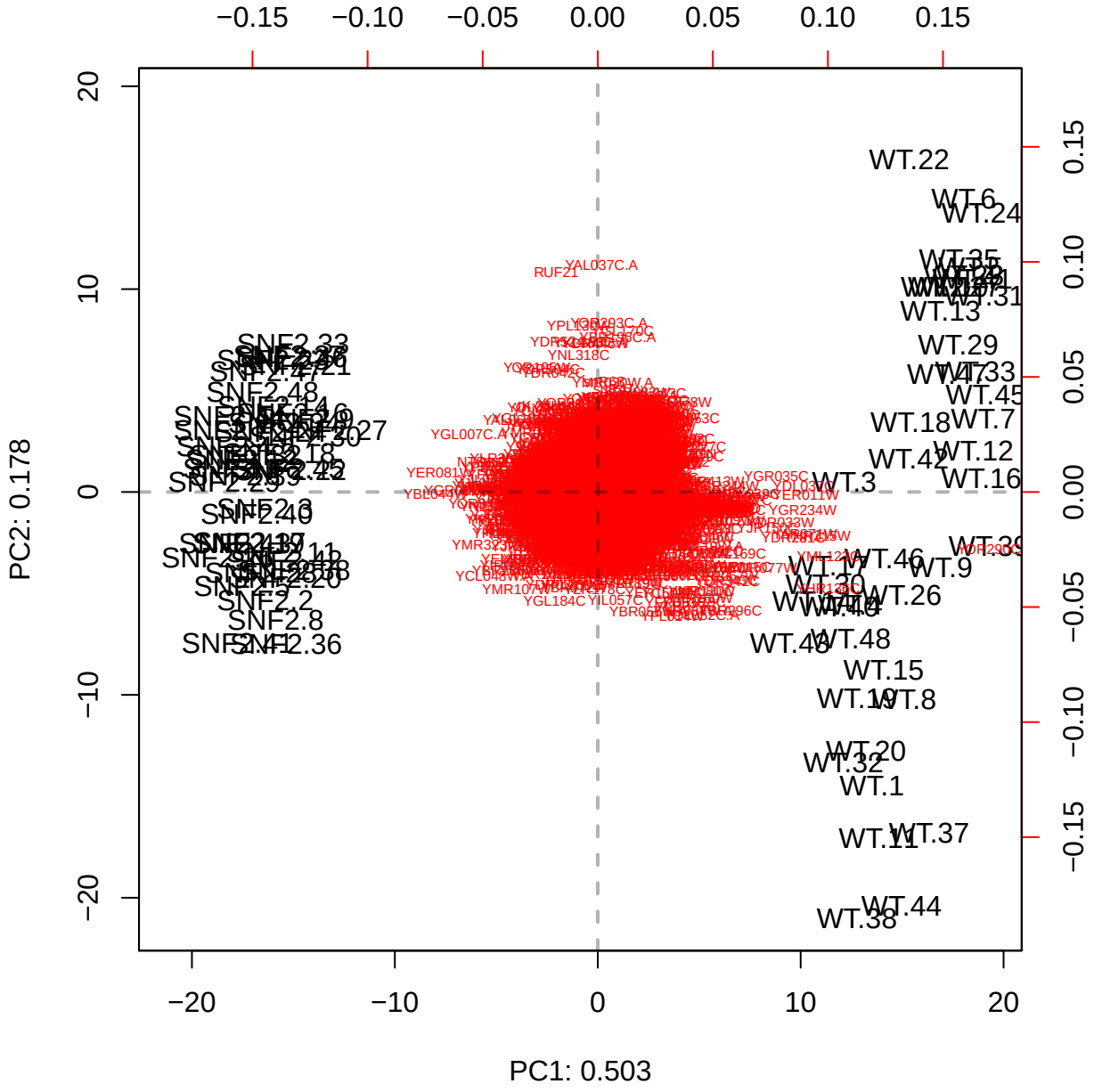



Histogram of sample.var
Frequency
0
1
2
3
4
5
0.5
1.0
1.5
2.0
2.5
3.0
sample.var

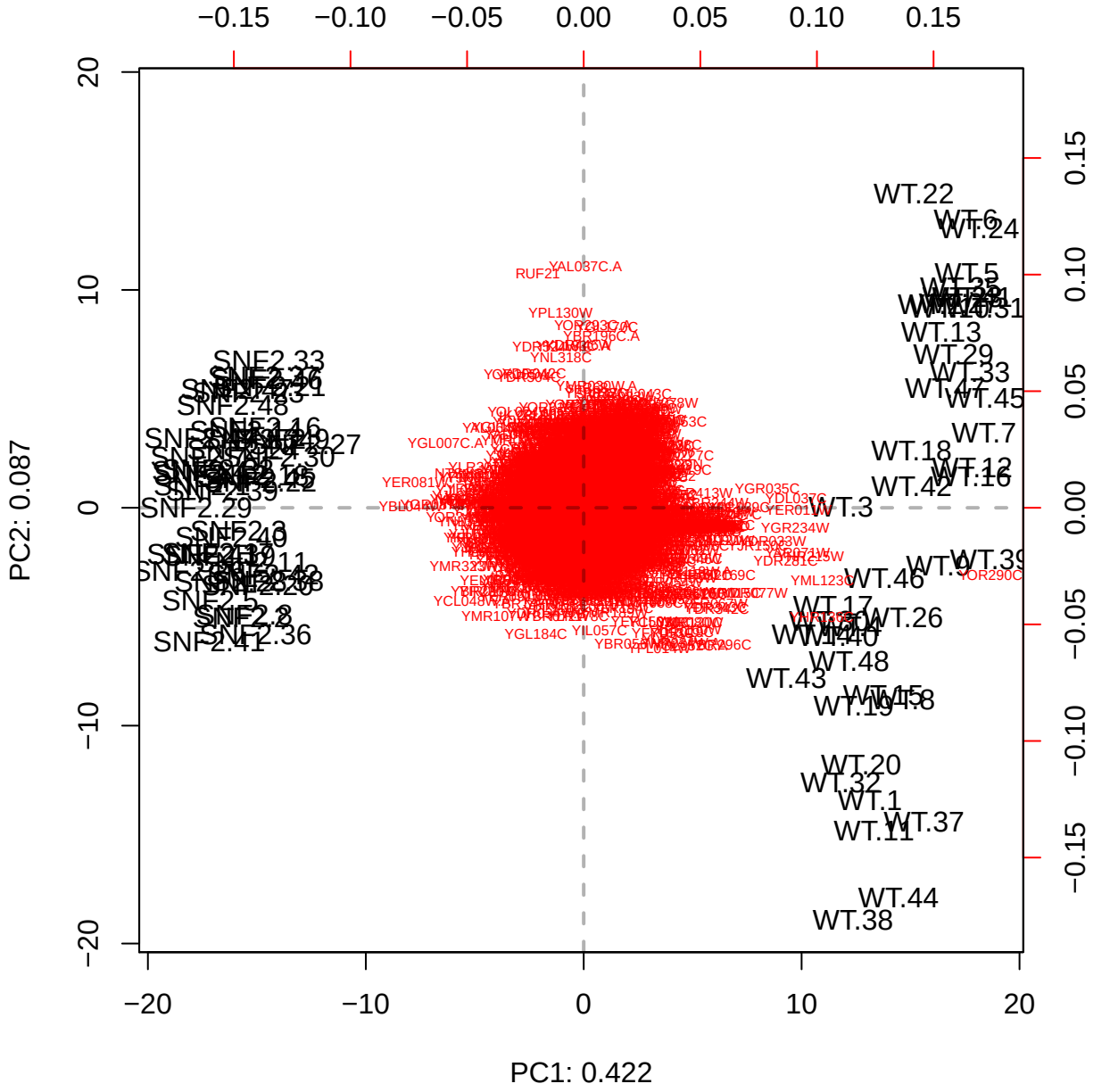

−0.15
−0.10
−0.05
0.00
0.05
SNF2.25
SNF2.15
SNF2.31
SNF2.36
SNF2.6
SNF2.20
SNF2.13
SNF2.24
SNF2.10
20
0
−20
−40
PC2
0.05
0.00
−0.05
−0.10
−0.15
−40
−20
0
20
PC1


Histogram of sample.var
Frequency
0
2
4
6
8
10
12
1
2
3
4
5
6
7
sample.var

Now we can make a biplot of only those samples that are non-outliers. We see that the SNF2 KO group is more homogeneous than is the WT group. Almost certainly since the SNF2 group is clonal, and the WT group is likely grown from a frozen culture.



We can do additional filtering. Examining the features, most contribute little, if anything, to the separation. These can be removed by filtering out low variance features. Note that we lose some resolution, but that we recapitulate the dataset with only half the features. We could do this iteratively.




## 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.

Gierliński, Marek, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2015. "Statistical Models for Rna-Seq Data Derived from a Two-Condition 48-Replicate Experiment." *Bioinformatics* 31 (22):3625–30. https://doi.org/10.1093/bioinformatics/btv425.

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

Schurch, Nicholas J, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2016. "How Many Biological Replicates Are Needed in an Rna-Seq Experiment and Which

Differential Expression Tool Should You Use?" *RNA* 22 (6):839–51. https://doi.org/10.1261/rna.053959.115.