Exploratory analysis of RNA-seq dataset

gg

2015-11-30

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

Exploration of the 'Statistical Models for RNA-seq' paper from Barton lab, Bioinformatics 2015.

In this paper, they generated a 7 technical, 48 biological replicate dataset generated from a yeast experiment comparing SNF2-KO to WT cells.

The paper has three main messages.

- that technical replicates are essentially Poisson distributed, this is similar to the Marioni paper, and we demonstrated that in the first ALDEx paper (Fernandes PLoS ONE 2013), and in the AJS paper (Gloor, Austrian Journal of Statistics, submitted). Nothing more needed I think, we use the sum of the technical replicates for all work. However, you should always check your lane replicates on a biplot first to ensure that the lane effects are minimal.
- that they have developed a protocol to identify poor biological replicates that involves a linear function including Pearson's correlation (ugh!), outlier fraction, and Chi-squared sequencing depth variance. In some sense, these are all measuring variation.

  thus, this can be addressed using a PCA plot of clr-transformed data very simply
- that the mean-variance relationship is negative-binomial distributed

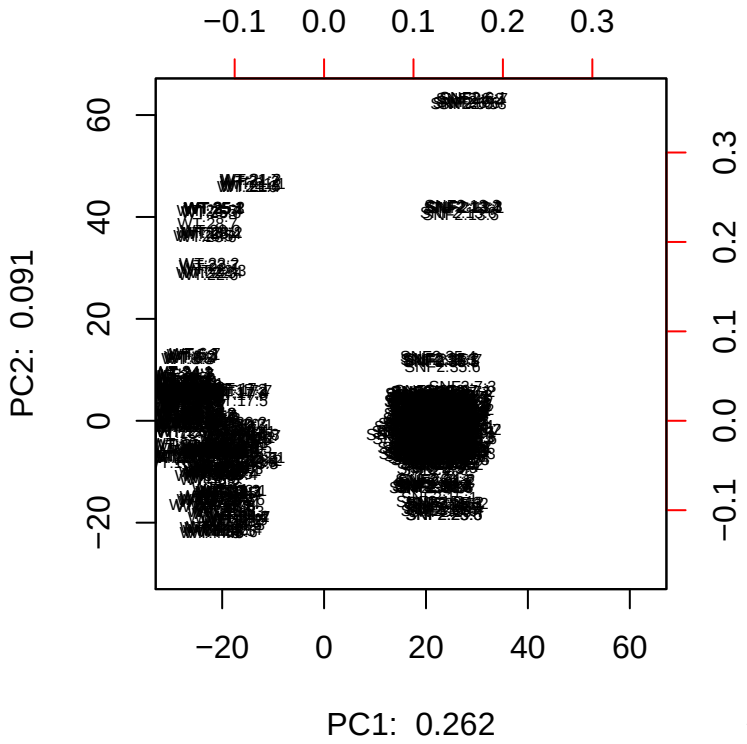
  this is true in linear space, but not in clr space: will need Andrew, David or Vera/Juanjo's help on this
- that the presence of 'bad' replicates breaks the negative binomial assumption

  so test how the bad replicates affect the ALDEx and edgeR approaches

Technical replication

The basic message here is that the technical replication is tight and Multivariate Poisson distributed. We have seen that, they see it, check it, reference it.

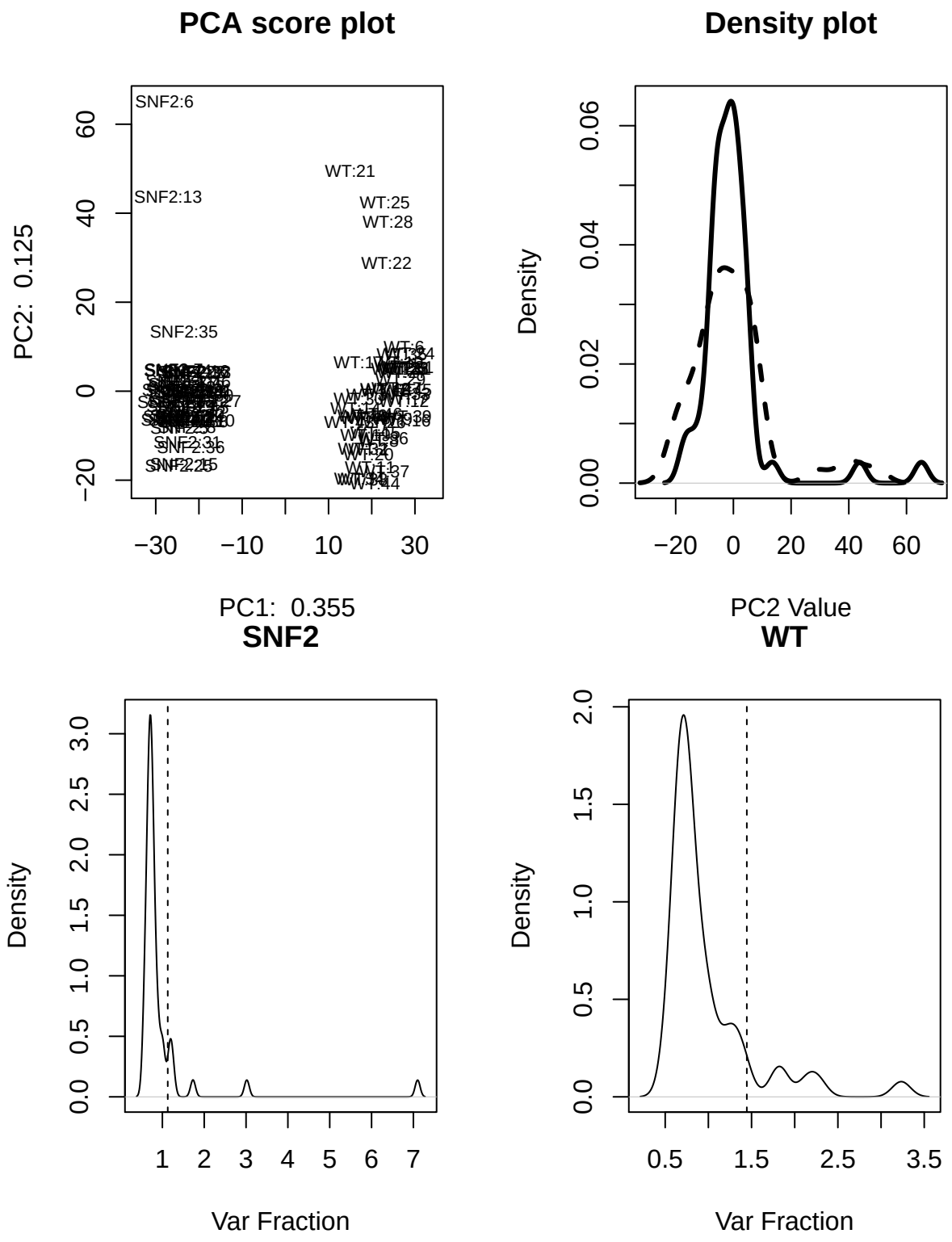
```
## No. corrected values:  2167
```



We can see that the samples cluster in sets of 7. Color each sample replicate similarly to observe the overlap. The alternative would be to determine the distance across replicates within samples or to determine the metric variance across replicates within samples.

Identifying outliers

- First aggregate all samples by summing the values for all replicates to give 96 samples
- Then remove genes with 0 reads across all samples.
- Then plot the PCA of log-ratio transformed data. They are discrete for the two groups — as expected



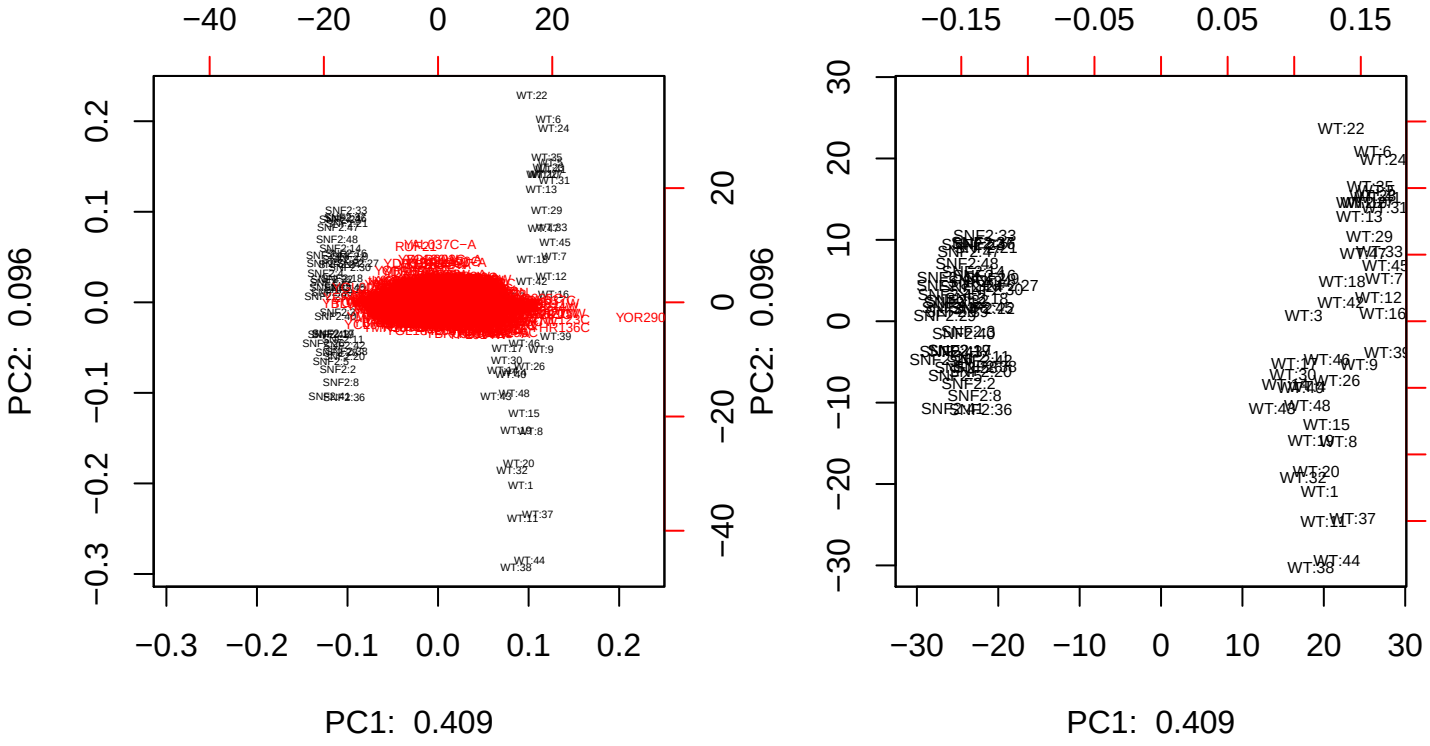
The underlying observation here is that there may be two viable strategies to detect outliers. The first would be to make a cutpoint at PC2=20 and remove samples that have a PC2 score greater than 20. This would be specific to each experiment and would not work if the data were rotated. Thus, I favour the second which would be to determine which samples are contributing more than expected to the total variance of the group. In either case, this would have to be determined empirically. I suggest that an appropriate cutpoint would be to remove samples that are contributing at least the median plus twice the IQR of variance to the group:

this would remove samples SNF2:10, SNF2:13, SNF2:15, SNF2:25, SNF2:31, SNF2:35, SNF2:6 from the SNF2 group and samples WT:21, WT:25, WT:28, WT:34, WT:36 from the wildtype group. This is not in perfect concordance with Barton, but is likely defensible on the grounds that we are looking at excluding those samples that contribute more than expected to the total variance of the group. We could go to a more stringent cutoff for sure by reducing the difference from the median, and it might make sense to use 1.5 as a cutoff. This depends on the number of samples that an investigator has to burn.

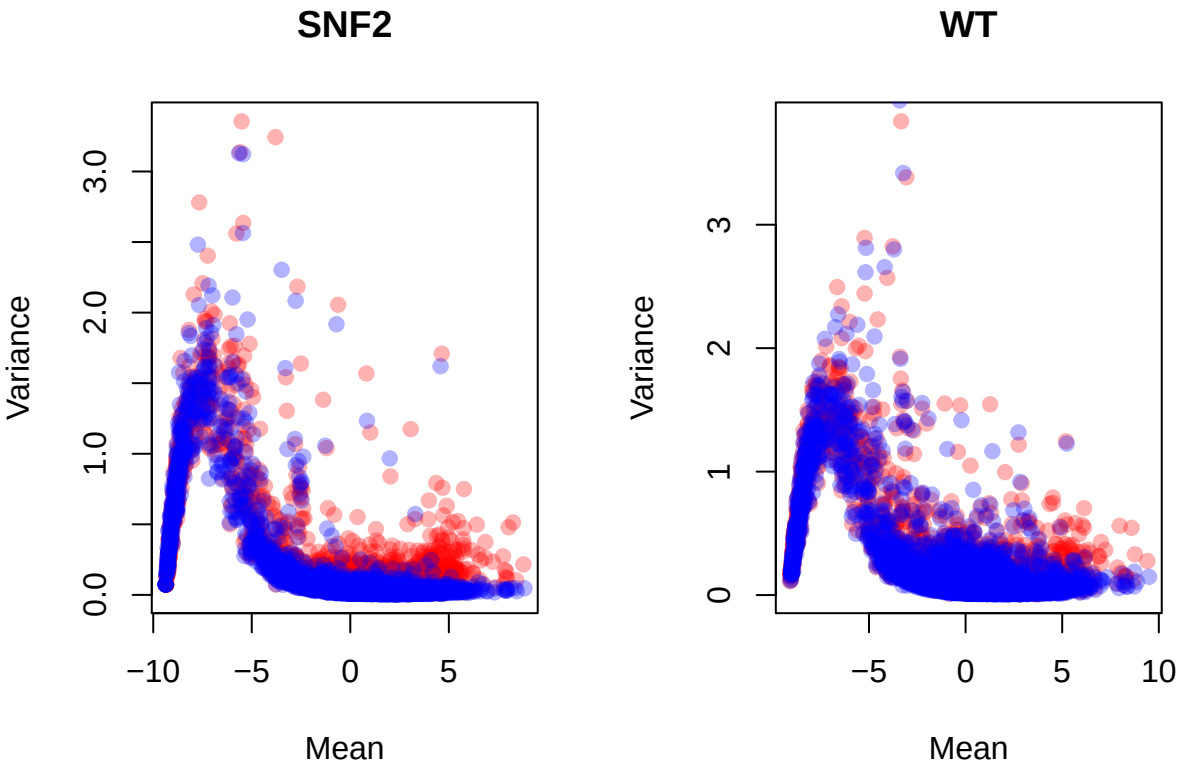
Mean Variance relationship

Finally, the Barton group demonstrated that the mean-variance relationship across samples could be modelled very well using the negative binomial distribution. This relationship is true in linear space — when there are no outlier samples, and in this idealized dataset. However, they go on to make the strong point that therefore, we should use the negative binomial in linear space as the only method of performing differential abundance tests: even though they demonstrate that the data are not negatively binomial when there are outliers, and even though they did not test any alternatives.

Thus, this is a fallacy because log-transformed data do not follow the negative binomial. The next figure shows the mean-variance relationship in the entire dataset as log-ratios, and in the dataset with the outliers removed.



We see that removing the outliers has increased the separation on component 1, and reduced it on component 2. The SNF2 dataset is obviously much less variable than is the wt dataset - there is no information in the paper as to why this would be.



These plots show the mean variance relationship as for the entire dataset (red) and for the good dataset (blue). In general, there is less variance in the good than in the entire dataset, and there is a set of genes with high expression that appear to have much lower variance in both datsets. So what distribution does the log-ratio data fit?

We can test to determine if the data fits a log-normal distribution using the Shapiro-Wilks test. Since the data are log-ratio transformed then they should fit a normal distribution if the underlying data are log-normal. One caveat is that genes that are largely composed of 0 counts cannot fit a log-normal distribution because they are heavily left-censored. Such genes are excluded from the analysis with filter that excludes genes with a mean value of less than -7. These genes are observed to have extremely high within-condition variation when Bayesian estimation is used to determine the distribution of their underlying log-ratio values (Fernandes PLoS ONE 2013). This high variation excludes them from being observed as 'significant' regardless of the statistical test.

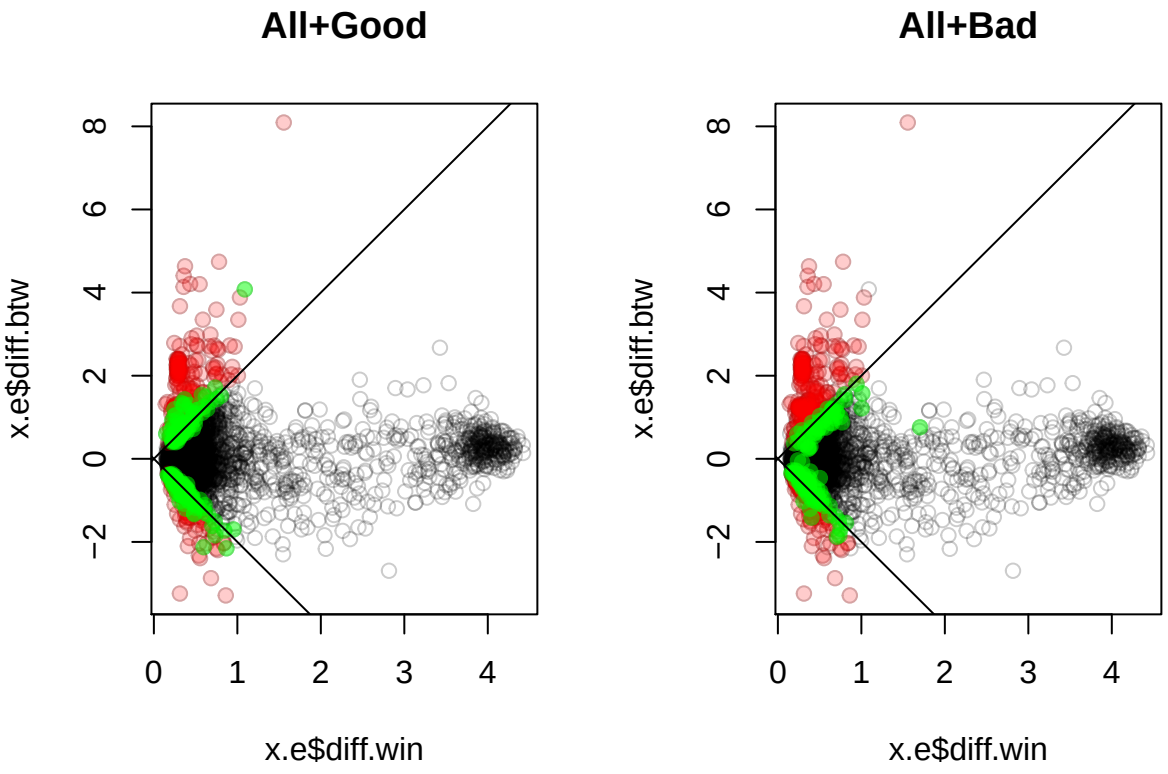
So in the entire dataset of 5904 and 5905 abundant genes in the entire and good dataset, there are 1327 in the SNF2 deletion and 422 in the WT set. When the dataset is reduced by removing the outlier samples, then there are 9 and 52. Thus, while the majority of the genes have a normal distribution following center log-ratio transformation, some do not. With the the number of non-normally-distributed genes being greatly reduced when the data are filtered to remove outlying samples. Inspection of the non-normally distributed genes reveals that they are either left-skewed (as would happen if a number of the values were 0 values) or they are bimodal (as would happen if there was a sub-structure to the gene expression of some of these genes in one condition or the other).

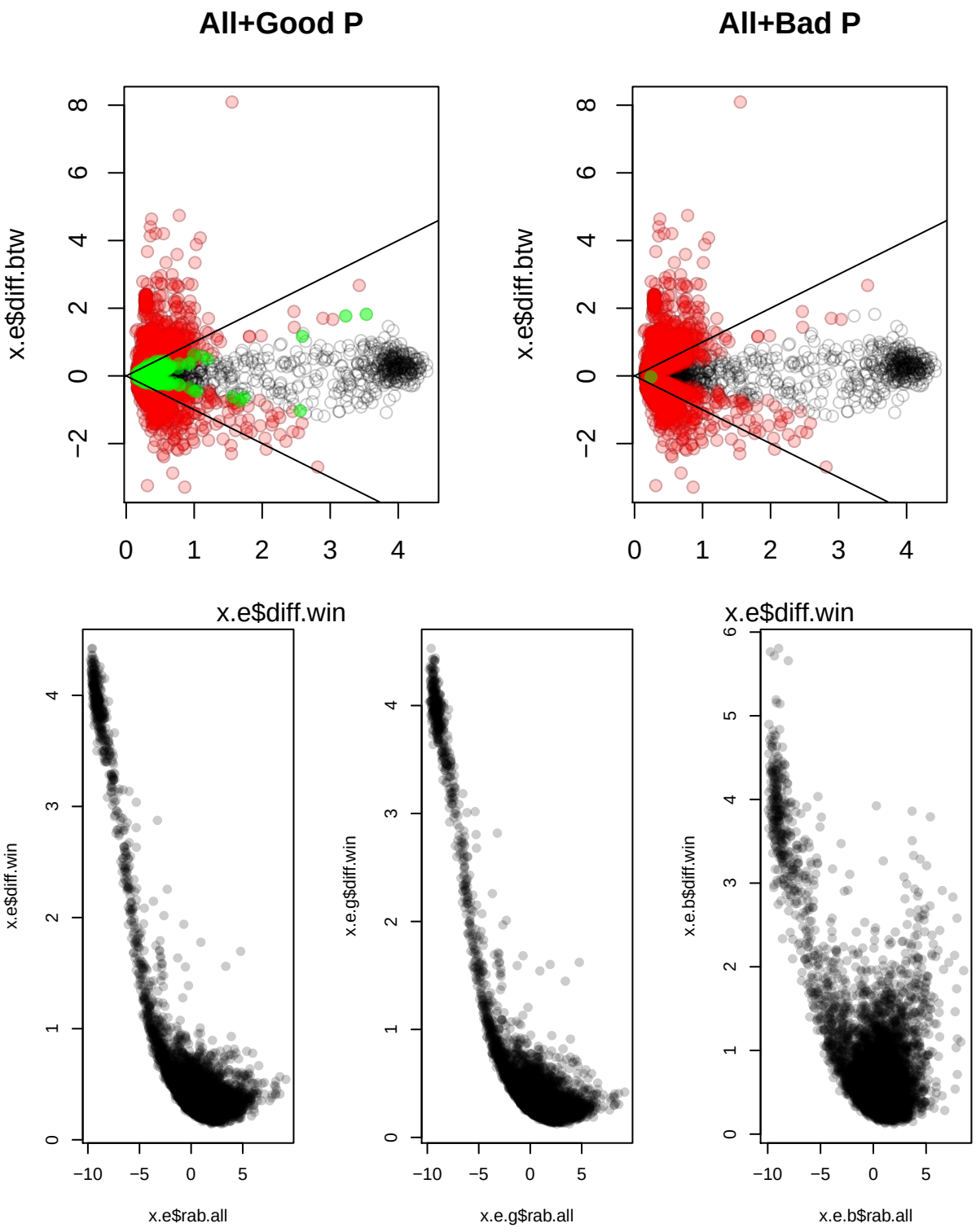
This reinforces the notion that whenever possible it is prudent to conduct statistical tests between conditions using non-parametric tests, but that in the worst case scenario a parametric test is probably OK, as long as outliers in the data are removed. This constrains the sample size to about 7-10 samples per group minimum for sufficient statistical power.

So what effect if any, do bad replicates have on the log-ration abundance test?

```
FALSE [1] "operating in serial mode"
```

```
FALSE [1] "operating in serial mode"

FALSE [1] "operating in serial mode"

FALSE [1] "operating in serial mode"

FALSE [1] "operating in serial mode"

FALSE [1] "operating in serial mode"
```



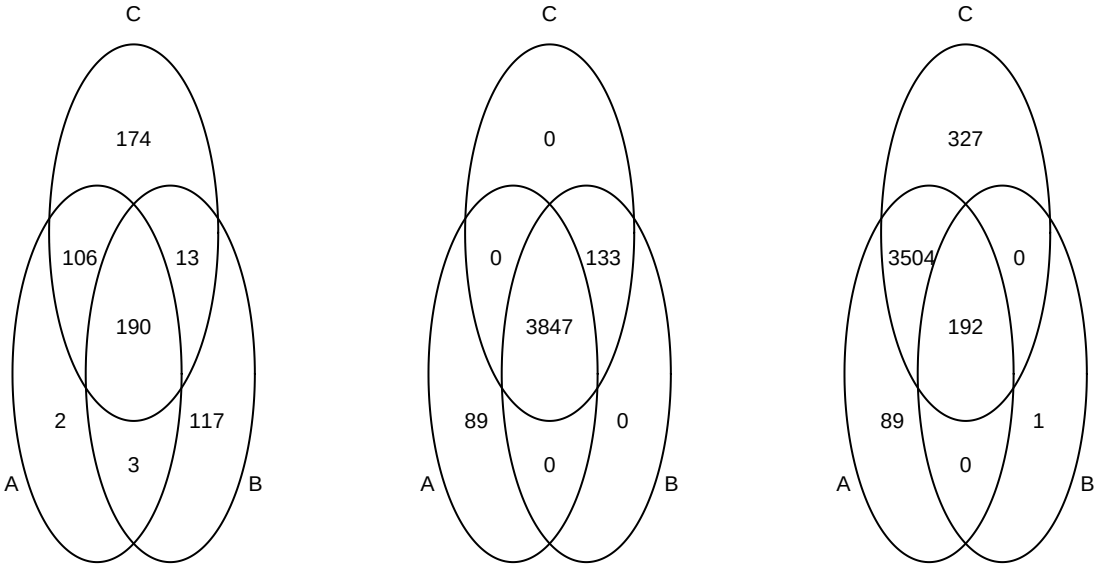


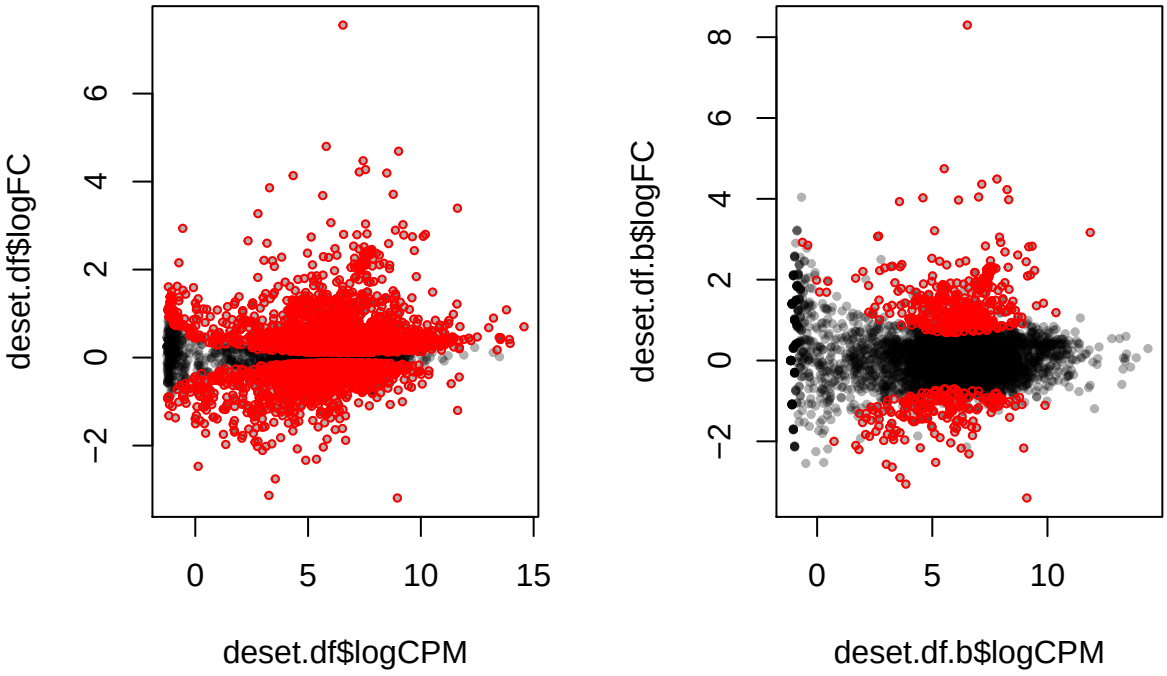
As we can see here the primary effect of including samples that contribute a lot of variance is to reduce power. This is shown in two ways. The top plot uses effect sizes, which is a more stable estimate of difference between groups than P values. The green circles are what is found in the good (or bad) dataset, that was not found in the entire dataset. You can see that the good dataset gets more genes with effects greater than 2, and the bad dataset includes more genes with effects less than 2. The entire dataset contains 301 genes with an effect >=2, the clean dataset contains 483, and the removed set of samples contains 323.

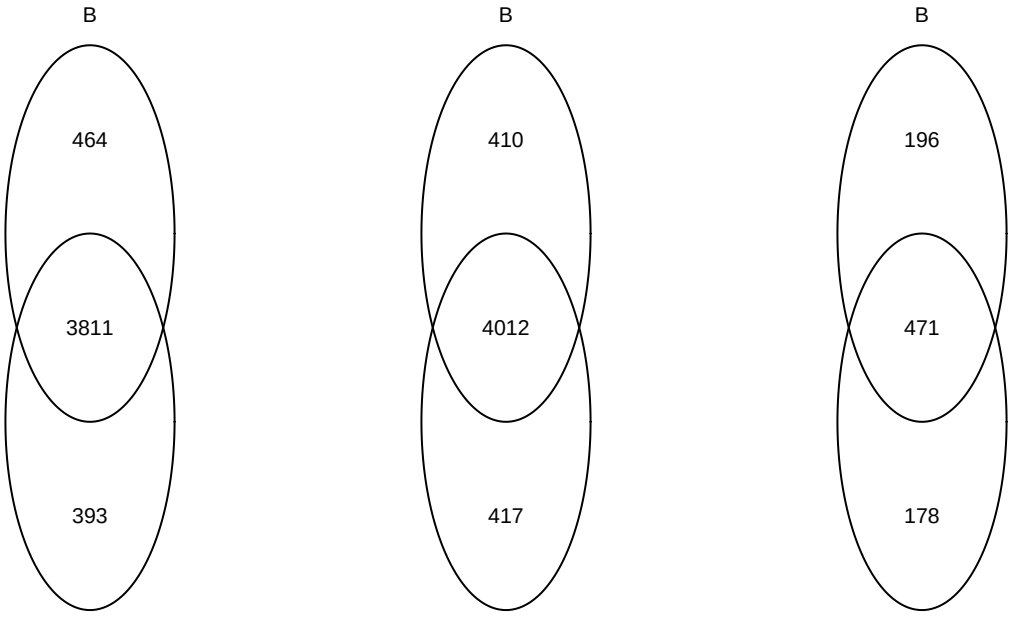
The next plots show the same thing for hypothesis tests, and we see here that the good (clean) dataset is providing more power, and the bad dataset has almost no power. See the Venn diagrams below.

The next set of plots shows why. The variance (diff.win) is plotted vs abundance. We can see that the all and good datasets have a nice relationship between variance, and log-ratio abundance, but that the bad dataset is quite ugly. Thus, the loss of power.

The first Venn diagram shows that the significant gene calls (effect > 2) from the set of all samples (A) is almost entirely contained within the clean dataset (C), and the bad dataset identifies many that are not in the all or clean dataset. The second venn shows the same picture when we use a Wilcox rank test, and the third shows what happens with a Welch's t-test. In both these cases the clean and all sets find almost the same thing, and the bad dataset has almost no power to find anything but the most obvious genes.







Finally, lets look at edgeR. The top plot shows that there are lots of genes with adjusted P less than 0.05 in the entire dataset, and a lot fewer in the bad dataset. We can get a better idea using Venn diagrams, and the three Venn diagrams show what we see for the all, good and bad datasets using edgeR (A) and ALDEx2 (B). Thus, in this idealized dataset, we are finding almost the same stuff with two tools with completely different assumptions. Hurray! If they did not agree, then you need to decide. However, it is important to point out that ALDEx2 has very few assumptions, and they are weak, whereas edgeR (DESeq) contain many strong assumptions. In stats, the fewer and weaker the assumptions you have to make, the generally more widely useful a tool is.