**RNA SEQUENCING REPORT**

*PART A- FROM RAW DATA TO GENE COUNT TABLE:*

*Aim:*

This report aims to present the sample specific statistics of the reads for raw, preprocessed and aligned data. The report also comments on the choice of flags for each software used in the pipeline and how the flags would have to be changed if the reads were paired end instead of single end.

*Sample specific read statistics:*

| Sample | Contaminated reads | Total reads | Contamination rate | Reads Kept | Reads Discarded | Aligned once(%) | Overall alignment(%) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| s1.c2 | 1549 | 5005476 | 0.000309 | 5004437 | 2588 | 89.32 | 100.00 |
| s10.c2 | 1612 | 4787005 | 0.000337 | 4784964 | 3653 | 88.03 | 100.00 |
| s11.c2 | 1448 | 4270489 | 0.000339 | 4270088 | 1849 | 89.38 | 100.00 |
| s12.c2 | 1327 | 4825269 | 0.000275 | 4824954 | 1642 | 88.32 | 100.00 |
| s2.c2 | 1353 | 4508122 | 0.000300 | 4506808 | 2667 | 87.43 | 100.00 |
| s3.c2 | 1040 | 4778803 | 0.000218 | 4777339 | 2504 | 87.32 | 100.00 |
| s4.c2 | 1407 | 5626462 | 0.000250 | 5626056 | 1813 | 88.96 | 100.00 |
| s5.c2 | 1632 | 5953416 | 0.000274 | 5952253 | 2795 | 87.35 | 100.00 |
| s6.c2 | 1164 | 5019472 | 0.000232 | 5019101 | 1535 | 89.22 | 100.00 |
| s7.c2 | 1179 | 4235397 | 0.000278 | 4235585 | 991 | 87.25 | 100.00 |
| s8.c2 | 2165 | 7707503 | 0.000281 | 7707368 | 2300 | 84.96 | 100.00 |
| s9.c2 | 1407 | 5626462 | 0.000250 | 5626056 | 1813 | 88.96 | 100.00 |

The above table depicts the sample specific read statistics of each sample. All samples achieved an overall alignment score of 100%, Most reads survived the quality control process conducted using scythe for adapter trimming and sickle for base quality trimming, highlighting that the contamination rates of the samples were low and the above table shows that the range of contamination was close to 10-6. Over 85% of all reads per sample aligned uniquely. The Flags used are discussed below.

*Flag selection in the various software used in the pipeline:*

Adapter and Quality trimming was done using:

1. *Scythe:*

The flags used in scythe were (***-a***) for the adapter file and (***-q***) with sanger

since the data was sanger encoded.

1. *Sickle:*

The flags used in sickle were (***-t***) to declare the encoding as sanger and a

quality and length threshold of 10 and 50 were declared using the (***-q***) and

(***-l***) Flags respectively.

The quality threshold of 10 was used to ensure bases of low confidence were

removed while keeping reads of a minimum length of 50, thus preventing

short sequences from hindering alignment.

Read alignment Using HISAT2:

The flags used in HISAT2 were ***--rna-strandness RF***, ***--phred33*** for quality encoding and

***-p 4*** which tells HISAT2 to use four threads. The reads were prepared using the stranded protocol in Illumina, this requires the use of the flag ***--rna-strandness RF*** to guarantee proper strand specific alignment. ***-U*** was used to inform the HISAT2 software that the reads were unpaired.

Conversion of SAM filetype to BAM and sorting using SAMtools:

The flags used in SAMtools were ***view -b -o*** to convert the .sam files to .bam and the flag ***sort -o*** to sort the .bam files. This step was performed to improve both storage and the speed at which the analysis was performed, this step is also necessary for stringtie to properly assemble the transcriptome.

Transcript Assembly using Stringtie:

We used the flags ***-p 4 -e -B -t --rf -G*** these flags inform stringtie to perform a guided assembly using 4 threads with the proper strand orientation, based on the provided GTF file.

The ***-e*** flag makes sure that only known transcripts are estimated and the ***-B*** flag ensures the output is in ballgown format, which is key to further downstream analysis using Ballgown.

*How would the flags change for paired end reads?*

Suppose the dataset was composed of paired end reads, then we would chane the following flags in our RNA-seq pipeline:

1. HISAT2:

In our script HISAT2 uses the flag ***-U*** to establish that the reads were unpaired, if this was not the case then we would use ***-1*** and ***-2*** to establish the forward paired read and reverse paired read respectively.

Ex: hisat2 -p 4 --rna-strandness RF --phred33 -x ${Genome\_dir} -1 read\_1.fq -2

read\_2.fq -S output.sam

1. SAMtools:

SAMtools will need an additional flag to retain the exact paired reads while converting the .sam file to.bam and this flag is ***-f 2***.

1. StringTie:

Does not need any change or additional flags as it cross references the provided reference annotation.

*Conclusion:*

The quality control which included the adapter trimming and quality base trimming using scythe and sickle respectively, alignment using HISAT2, transcript assembly using StringTie resulted in the procurement of the gene and transcript count matrices of all samples. The statistics showed that the adapter contamination was minimal for all samples and approx >85% of all reads mapped uniquely for all samples. Using the appropriate flags ensured the data quality and accuracy was optimal. The report also discusses how the flags would change if the reads were paired end.

*PART B- DIFFERENTAL EXPRESSION ANALYSIS USING DESeq2:*

*Introduction:*

Identifying genes that show significant changes in expression across different conditions is a crucial part of transcriptomics and Differential Expression analysis allows us to identify these genes. In this analysis we use the DESeq2 package in R to perform the DE analysis, to assess the difference in expression between the groups *B vs A* and *C vs A*. The analysis focused on a variety of statistical tests such as dispersion estimates, Principal Component Analysis (PCA), variance stabilization and log-fold change (LFC) shrinkage.

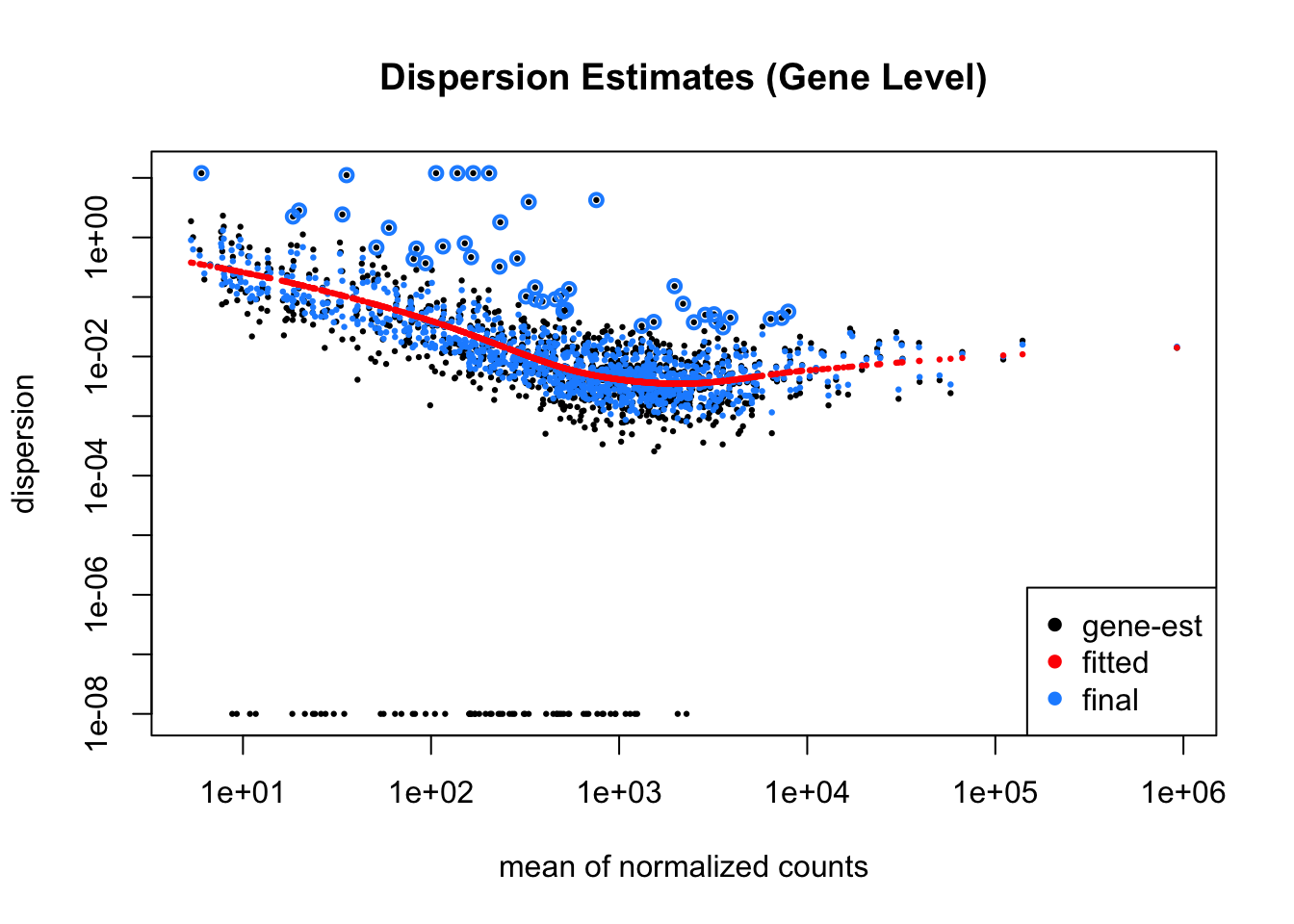
Through the use of MA plots, dispersion plots and PCA plots, we examined the variation in gene expression and how batch correction affects the results. Variance stabilization through the use of rlog transform was also performed to understand the best method for analysis.

*Dispersion Estimation and Shrinkage:*

*What is dispersion in RNA-seq?*

Dispersion estimates the amount of variation in gene expression across biological replicates. Dispersion and variance are directly proportionate i.e. high dispersion implies high variability and low dispersion implies that the replicates are consistent. RNA-seq count data follows a negative binomial distribution and DESeq 2 models them accordingly.

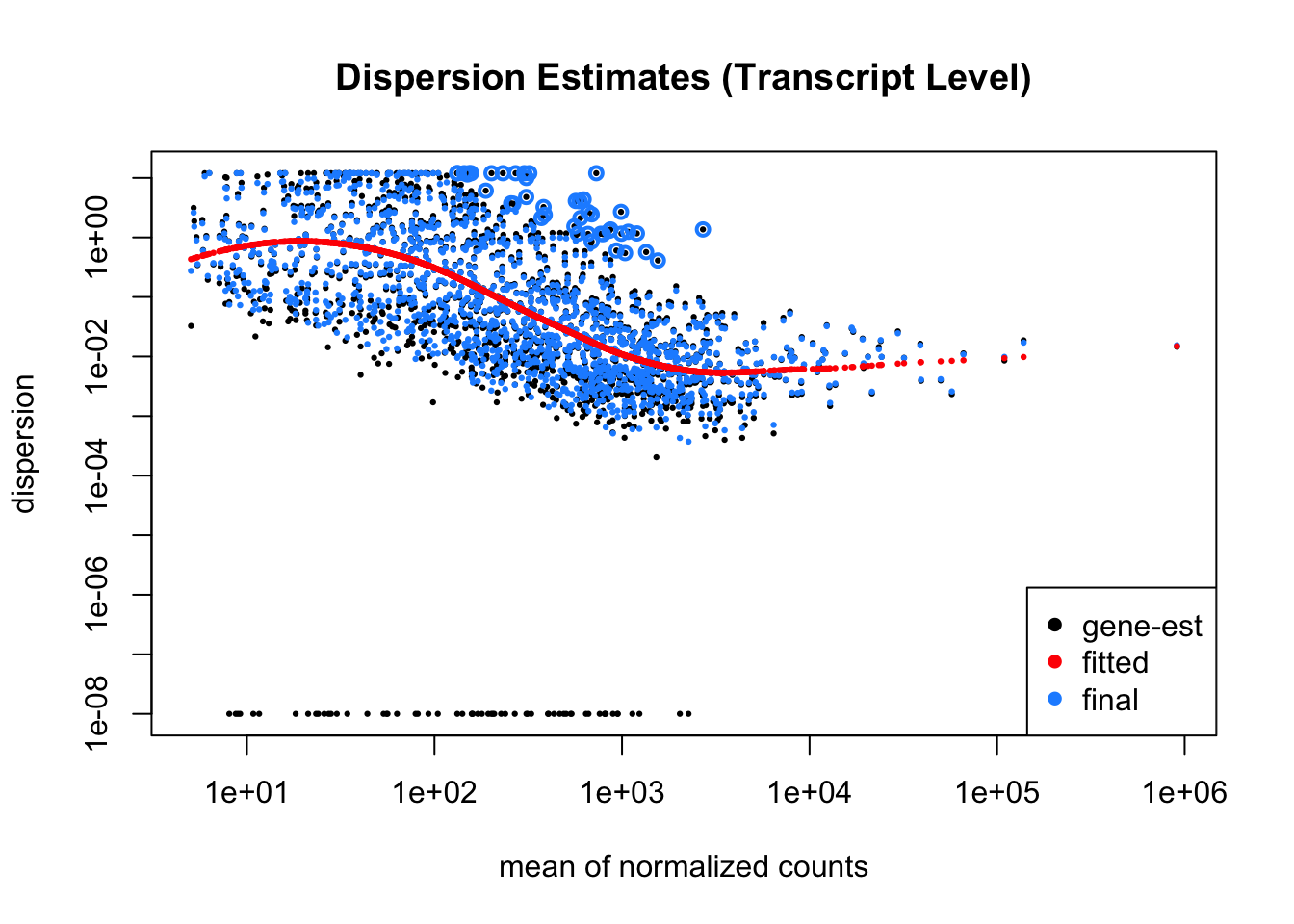
*Gene level Dispersion Plot:*



The above plot depicts the dispersion estimates at the gene level. The black dots represent the raw dispersion estimate per gene and the blue circles indicate the final dispersion values after shrinkage. The red line indicates the trend of dispersion.

Here the trend indicates that the genes with low mean counts have higher dispersion, by applying shrinkage the values are stabilized, which prevents variance inflation in lowly expressed genes.

*Transcript level Dispersion Plot:*

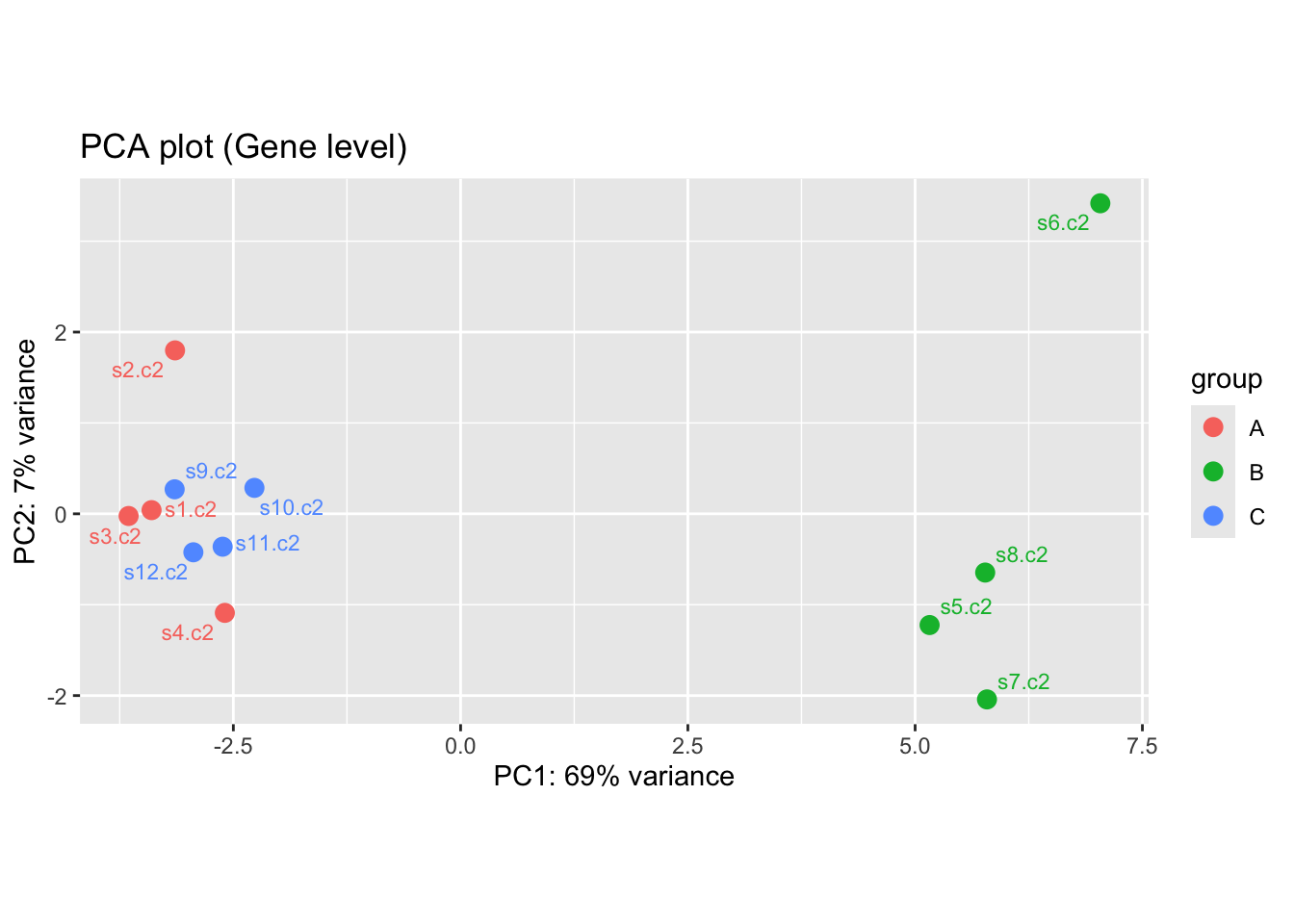


In the above plot it is clear that the trend is higher at low mean counts, showing that there is greater variability in the expression of transcripts. The variation is higher than at the gene level, which is obvious because of the complexity that comes from alternative splicing.

From the above discussion it is clear that shrinkage is necessary to control dispersion especially for low expression transcripts. Gene level trends are more uniform than at the transcript level.

*Principal Component Analysis:*

PCA helps us to understand and comprehend the main sources of variance. If the groups separate well then we can say that there could be strong differential expression patterns.

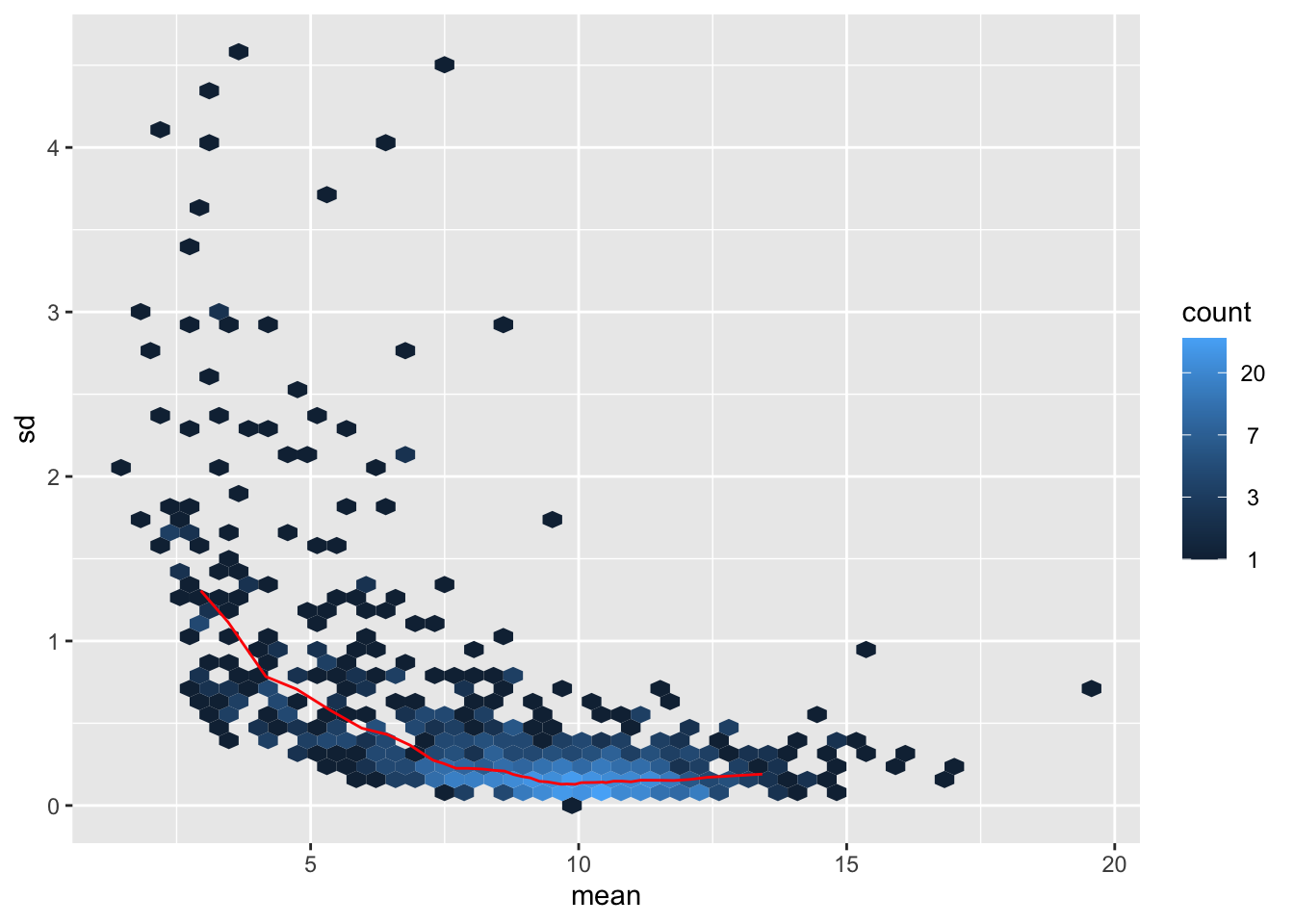


The above plot depicts the PCA analysis at the gene level, PC1 explains 69% of the variance and shows the differences in expression between groups. The samples cluster by group, which implies biological separation. Group B seems to be the main driving force behind the variation along PC1 since it clustered separately and group A and group C may be sharing a more similar biological relationship.

*Effect of rlog Transformation on Variance:*

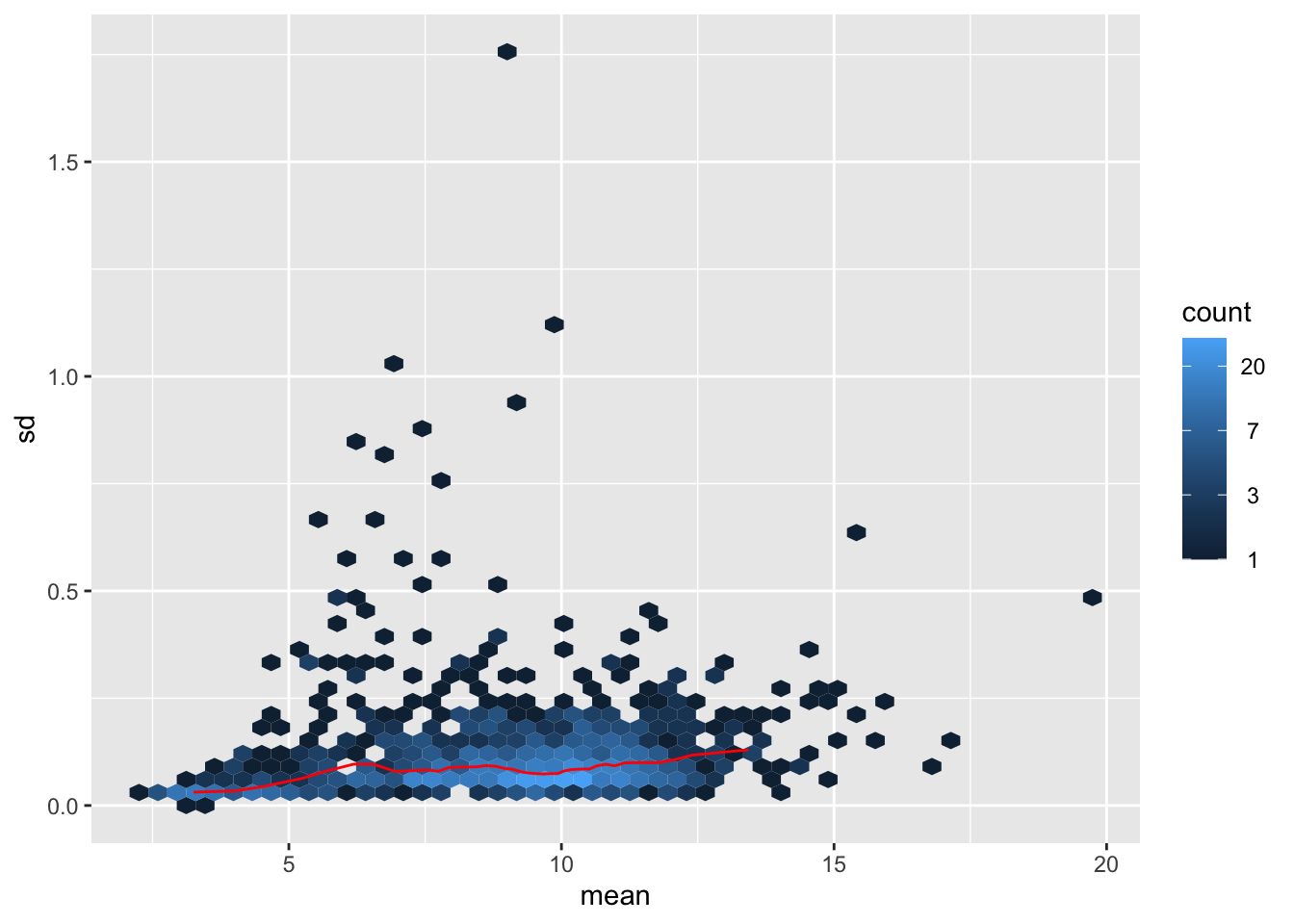
This effect was explored by comparing meanSdPlots for log2 normalized counts and rlog transformations.

*Log2 transformation:*

**

The above plot depicts the meanSD plot obtained after applying the log2 transform, it shows a strong correlation between mean expression and standard deviation(SD). Higher mean values imply greater variance. This impacts further analysis.

*Rlog transformation:*

**

In the meanSD plot obtained from rlog-transformed data, there is a more uniform distribution across expression levels, thus implying that the rlog transformation resulted in a more homoskedastic distribution as compared to the log2 transformed data.

*What are NULL Hypotheses of LFC=0 and LFC<1?*

In DE analysis the null hypothesis decides whether a gene is significantly differentially expressed or not based on its LFC and Adjusted p-value.

We explored differential expression under two different hypotheses:

1. LFC = 0:

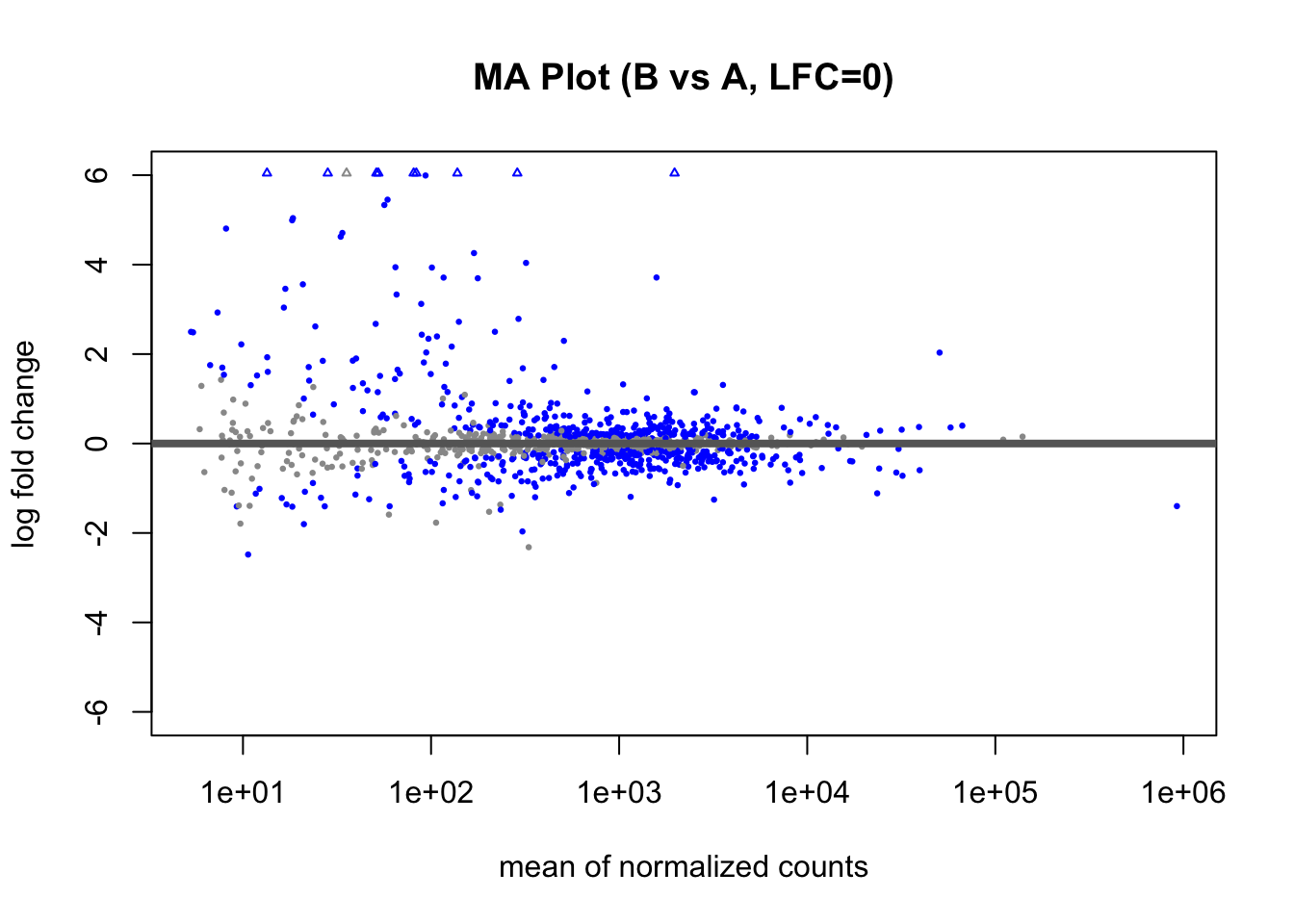
In this case we tested whether the log2fold change is significantly different from zero. It assumes that there is no difference in expression between the groups. Genes with an adjusted p value below 0.05 are considered differentially expressed.

1. LFC < 1:

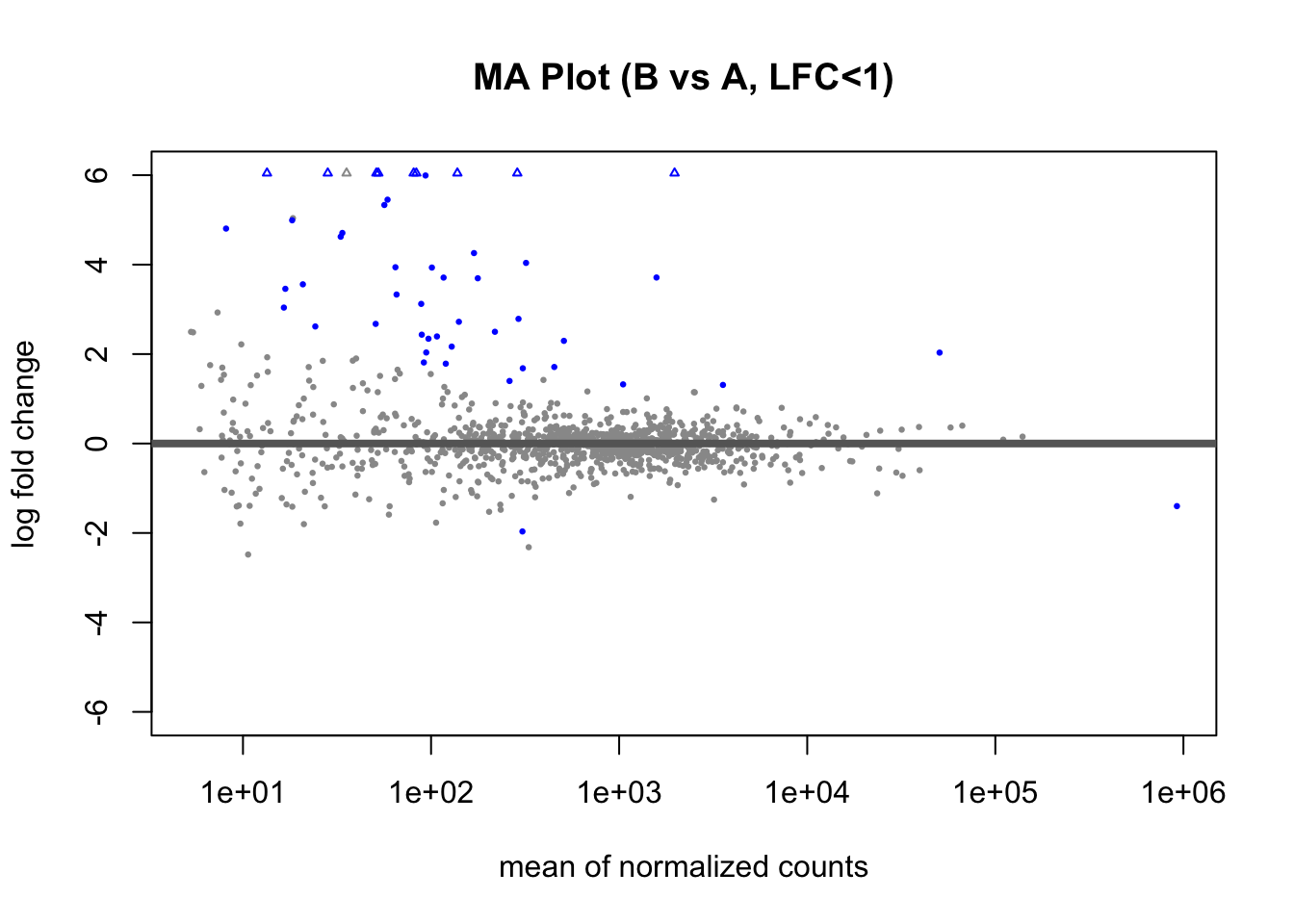
In this case we do not assume any non zero LFC as significant, the hypothesis now states that genes with an absolute LFC below 1 are not relevant. Genes with an LFC > 1 are considered significantly expressed. This prevents small fold changes from being considered as significant, reducing noise.

The LFC = 0 condition is less stringent and identifies a larger cohort of genes as significant due to the addition of genes with small effect sizes. The LFC < 1 condition reduces the number of significant genes and focuses only on the ones with a higher impact.

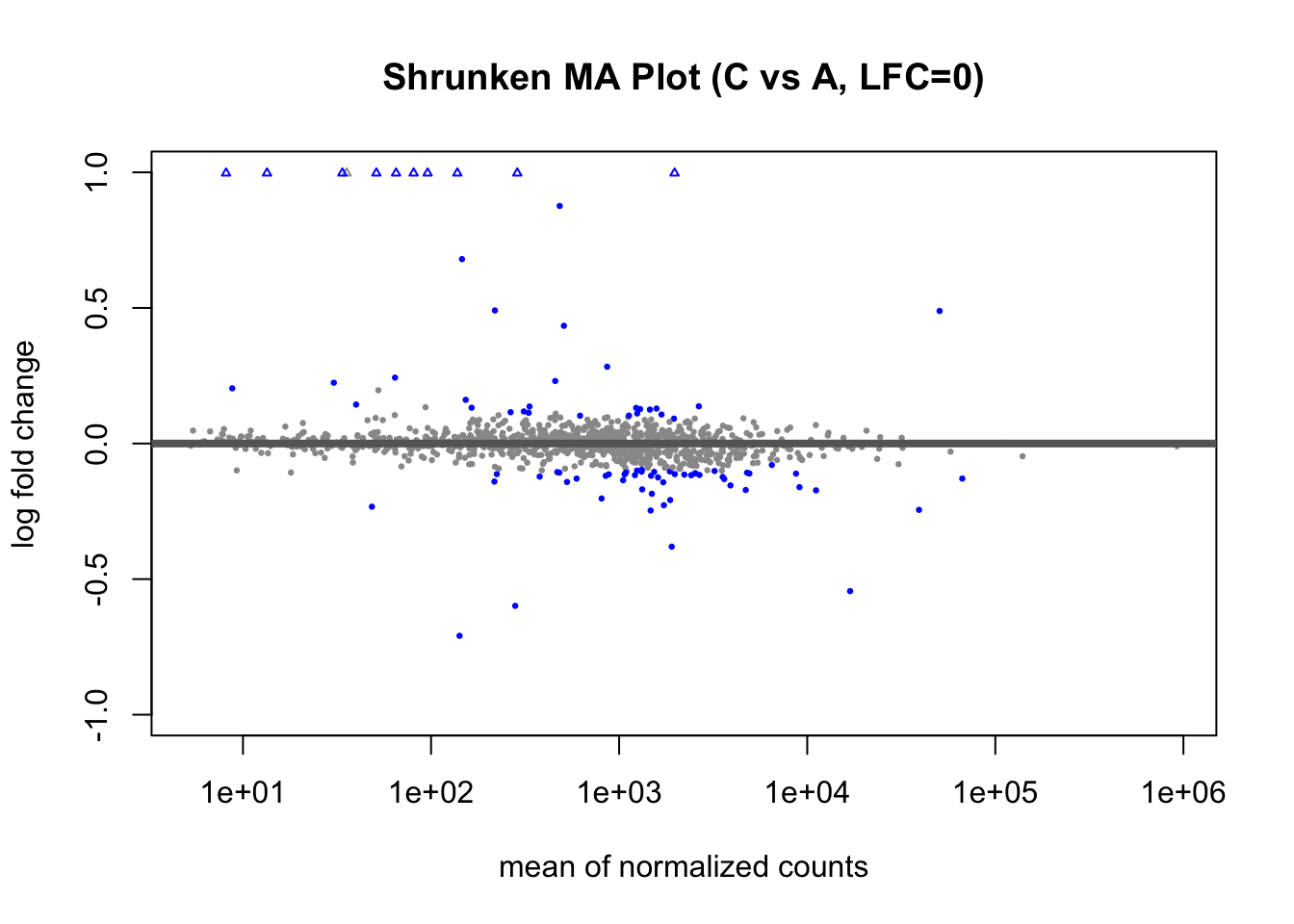
*MA plots for BvsA(LFC =0 and LFC < 1):*

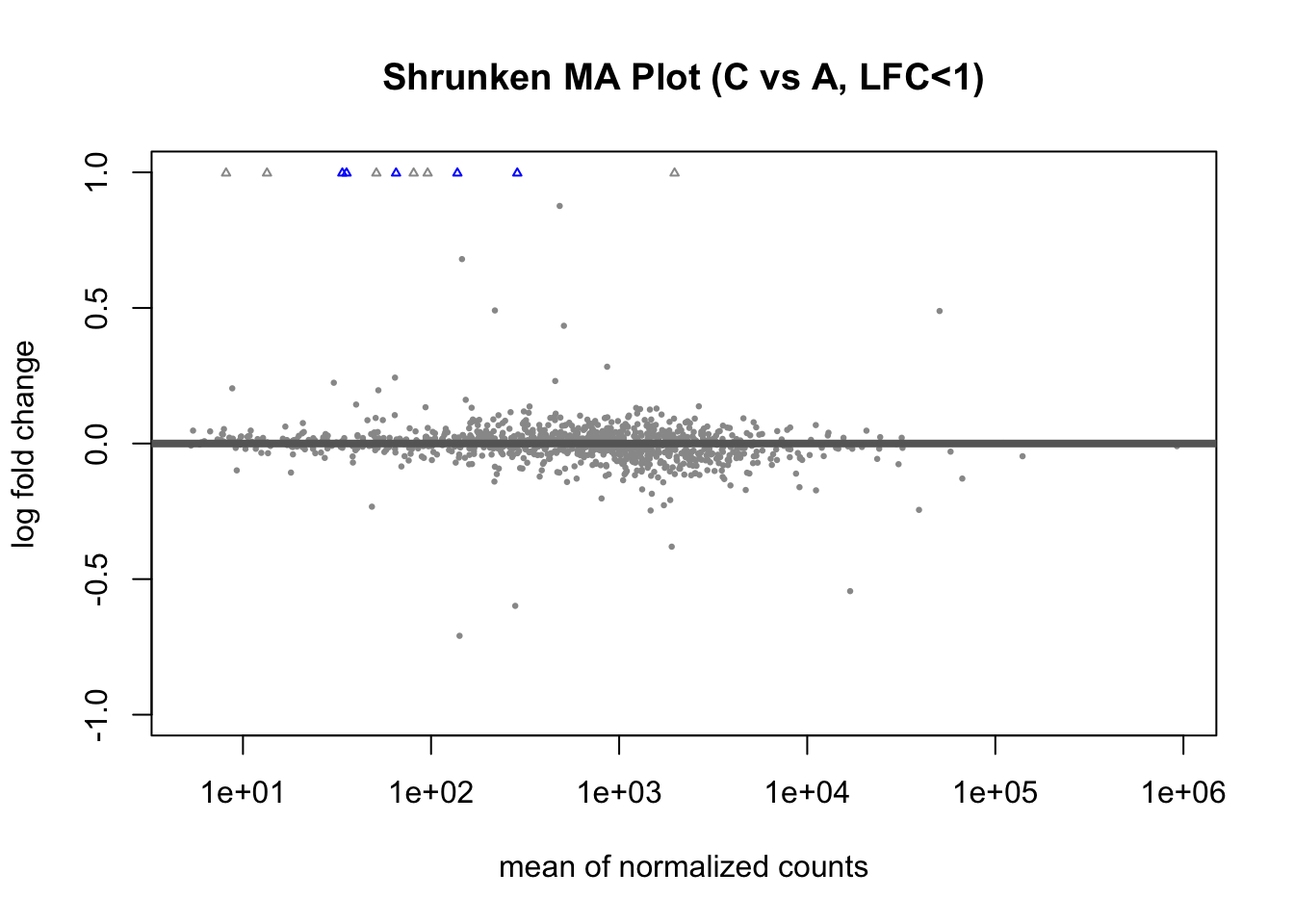


In the above MA plot where LFC = 0, there are a high number of differentially expressed genes (blue dots). Genes with low expression also show high LFC values.



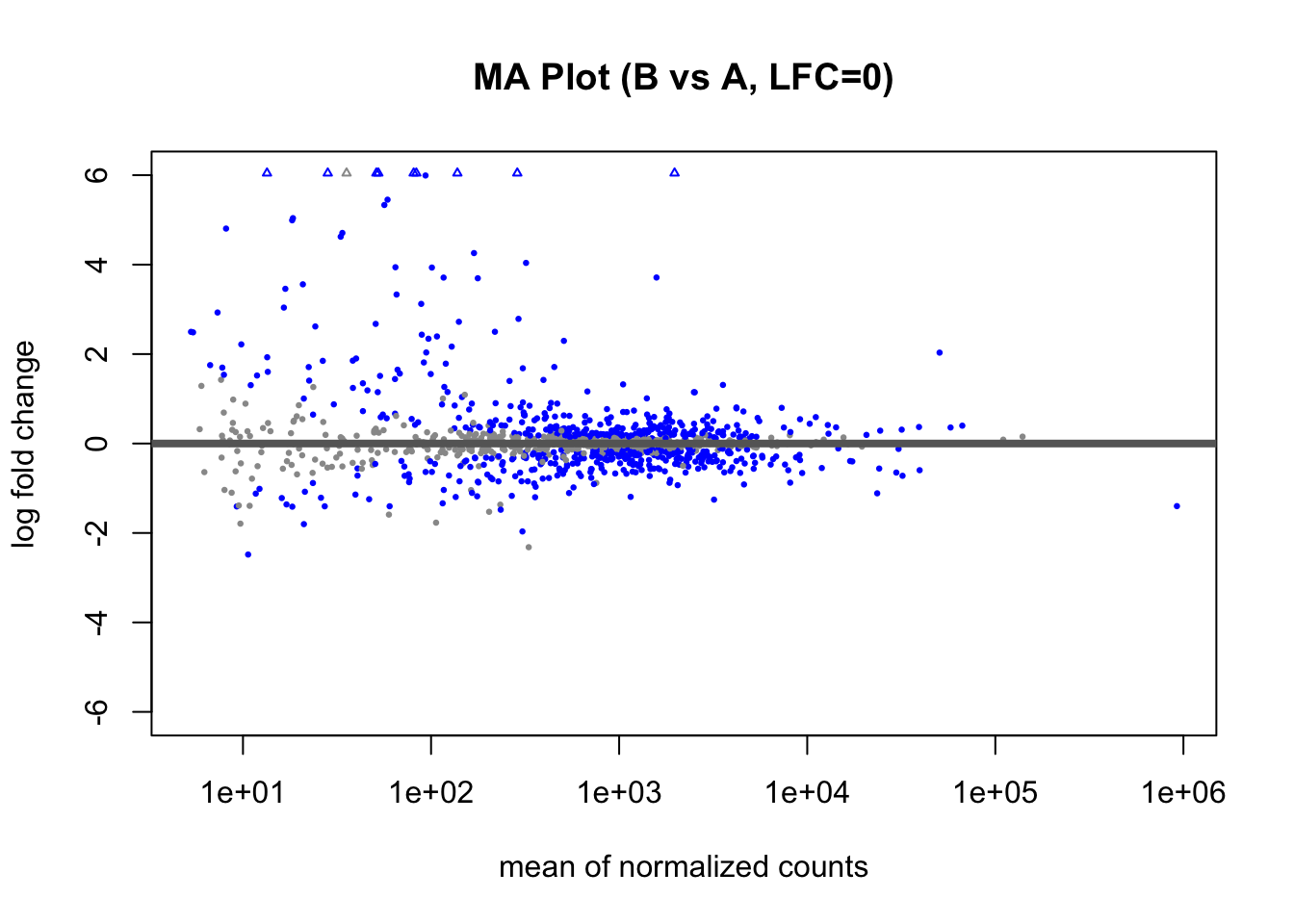
In the above MA plot where LFC < 1, many genes that were blue in the LFC = 0 condition have now turned black, but the ones with high log fold changes have stayed significant. Clearly, the lowly expressed genes with small fold changes have been disregarded as insignificant. Below are the MA plots of CvsA for both LFC = 0 and LFC < 1, similar results are depicted in them. As seen from the PCA the number of significant genes in the case of C vs A are far fewer than B vs A.



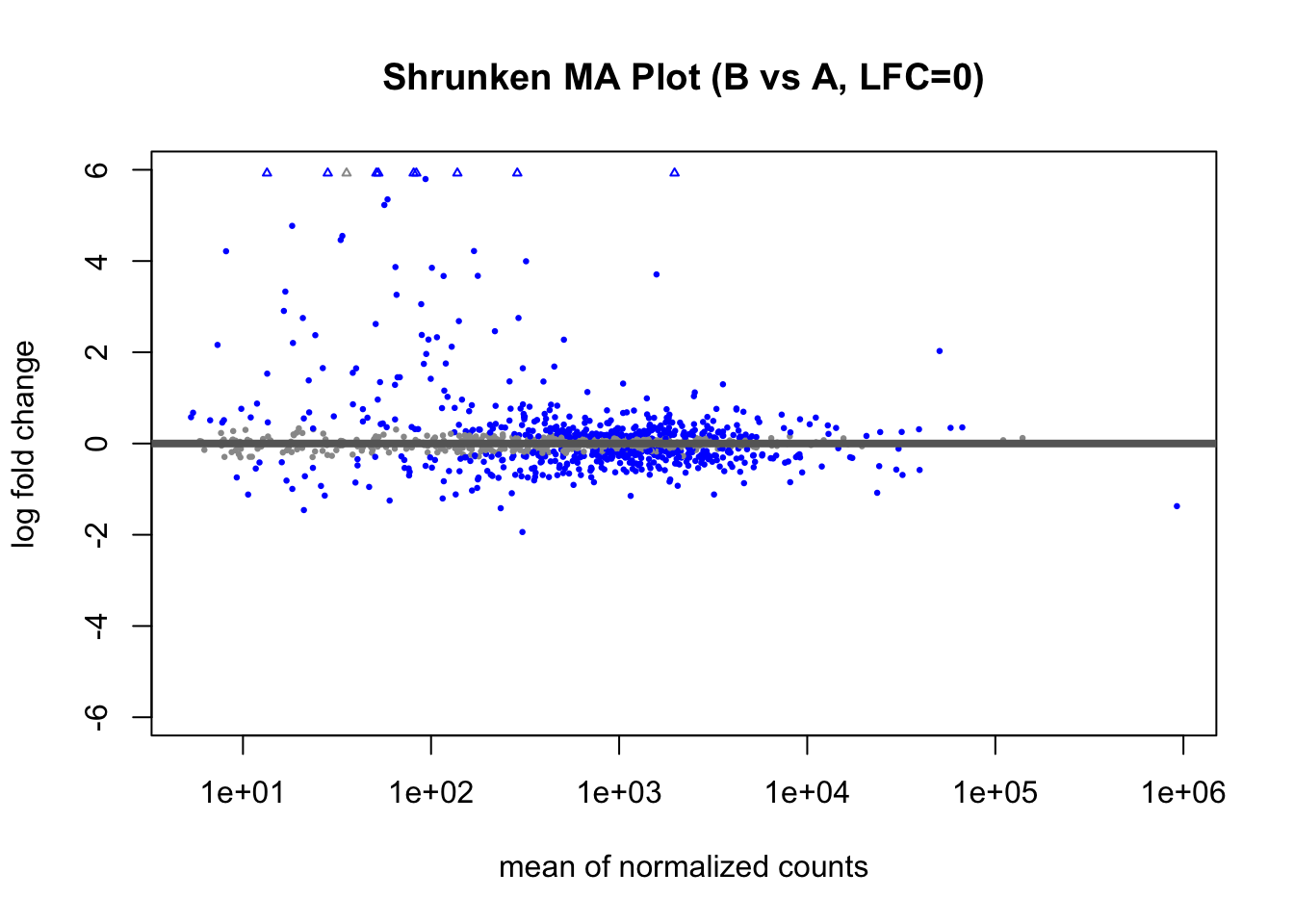


*Shrunken vs Unshrunken MA plots in standard Null Hypothesis or LFC = 0:*

LFC shrinkage is applied to increase the capability of handling genes with low counts, this reduces noise.



The above plot is the unshrunken MA plot; it shows higher variability in the LFC estimates, especially for lowly expressed genes.



The above MA plot is after the shrink is applied to the LFC values, it narrows the range of

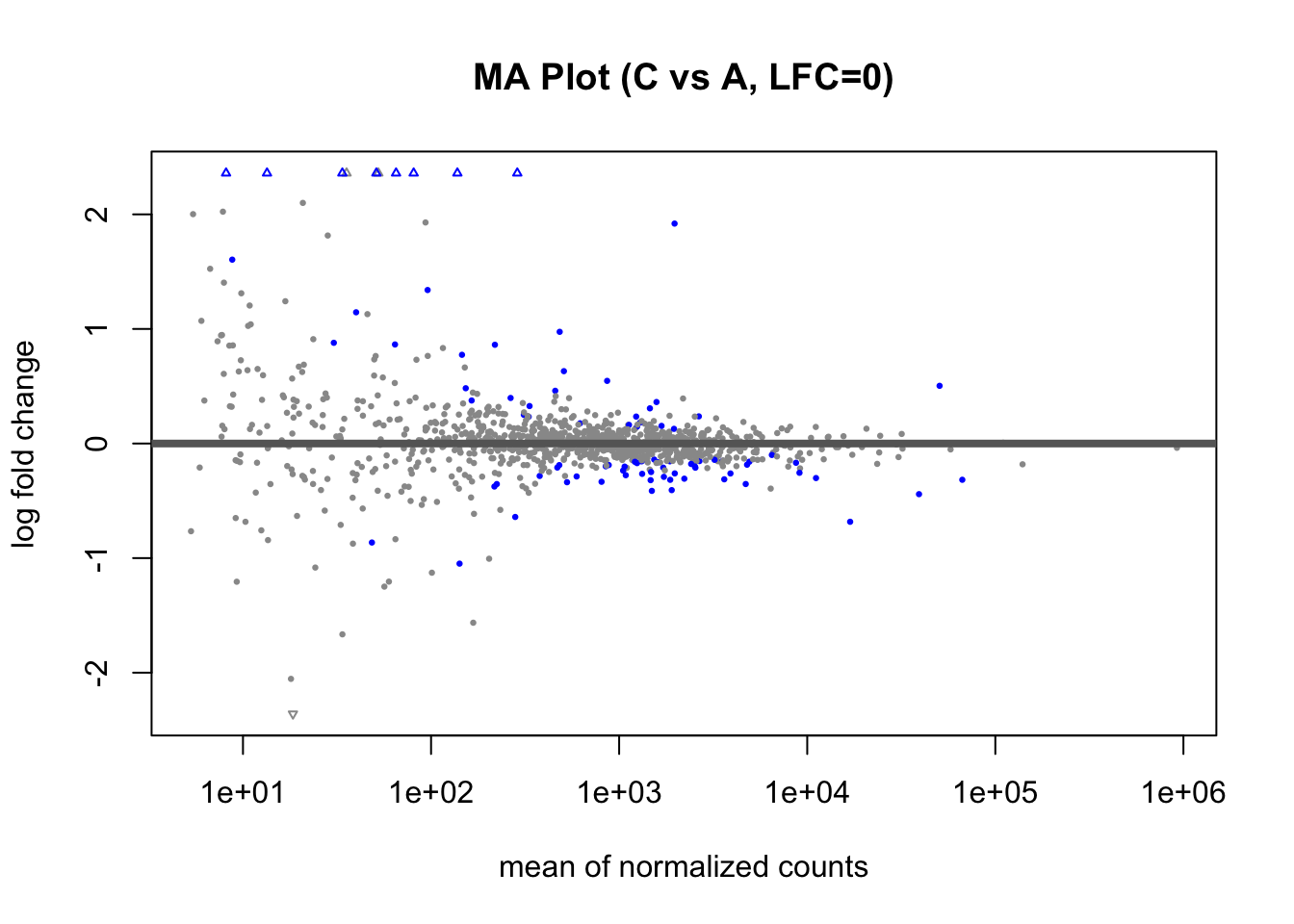
the detected changes and reduces the number of false positives while keeping the truly

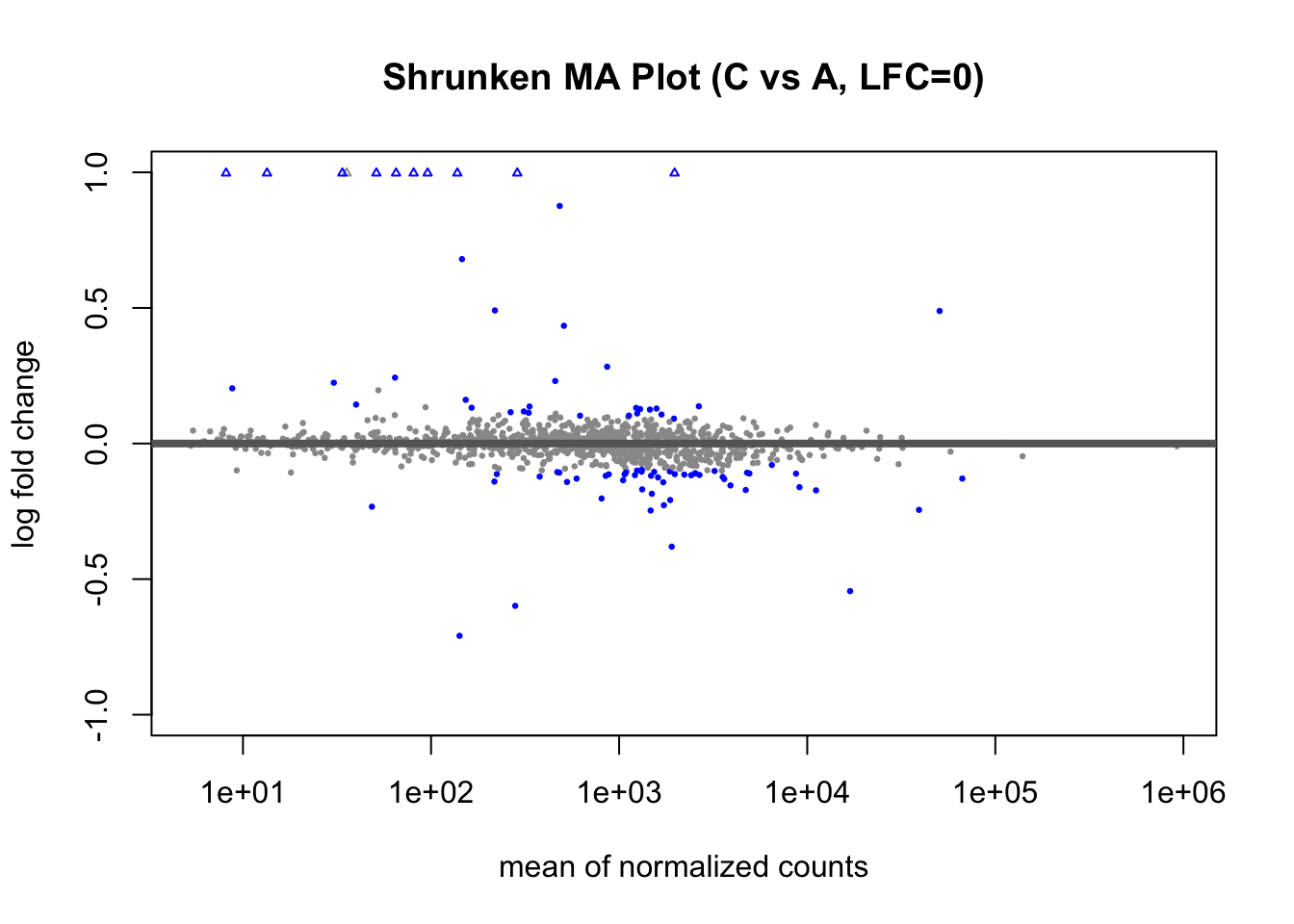
significant results. This is more useful to rank genes based on their effect sizes when the

Sample size is small.

The below plots depict the same findings from BvsA but in the other comparison group

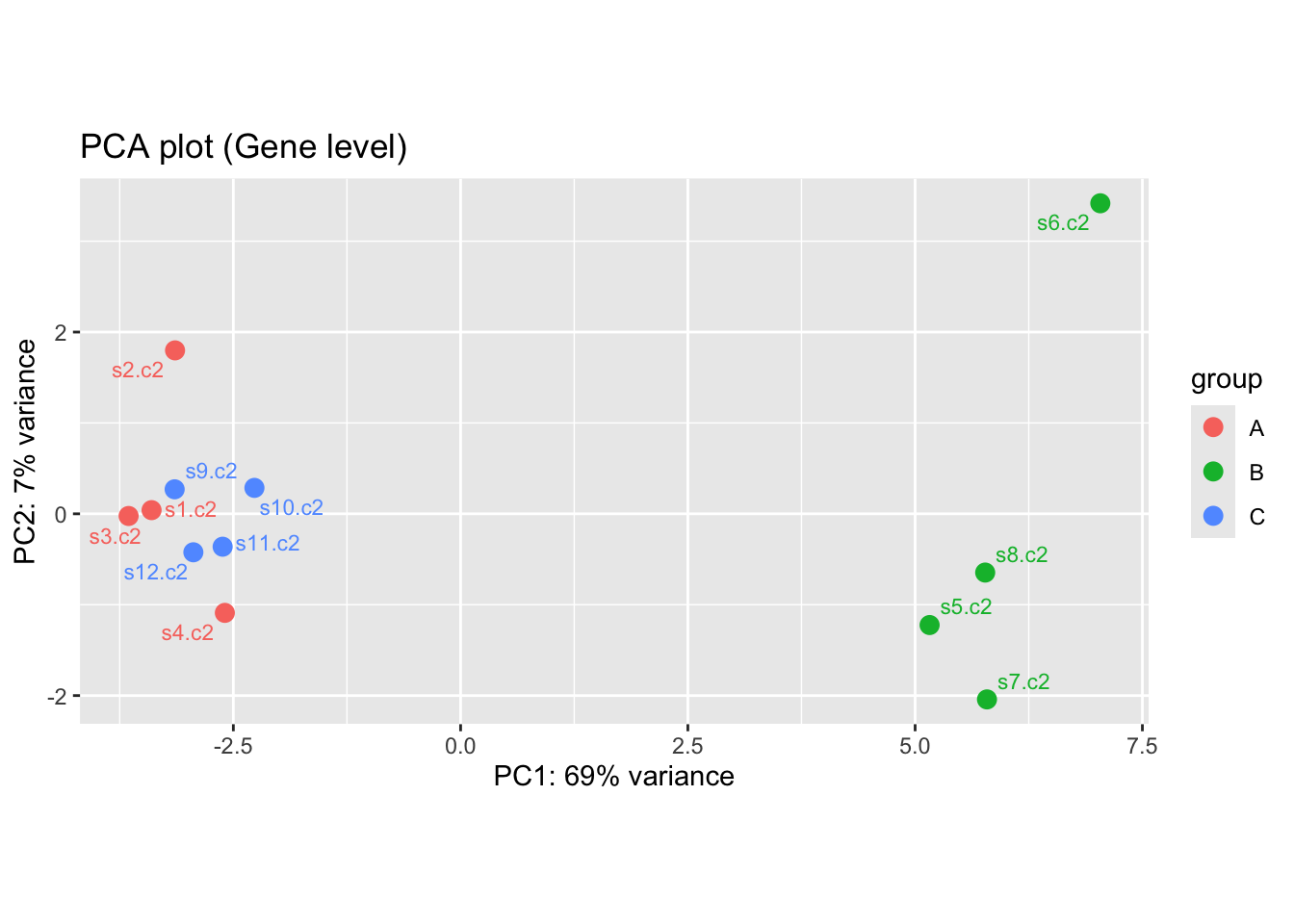
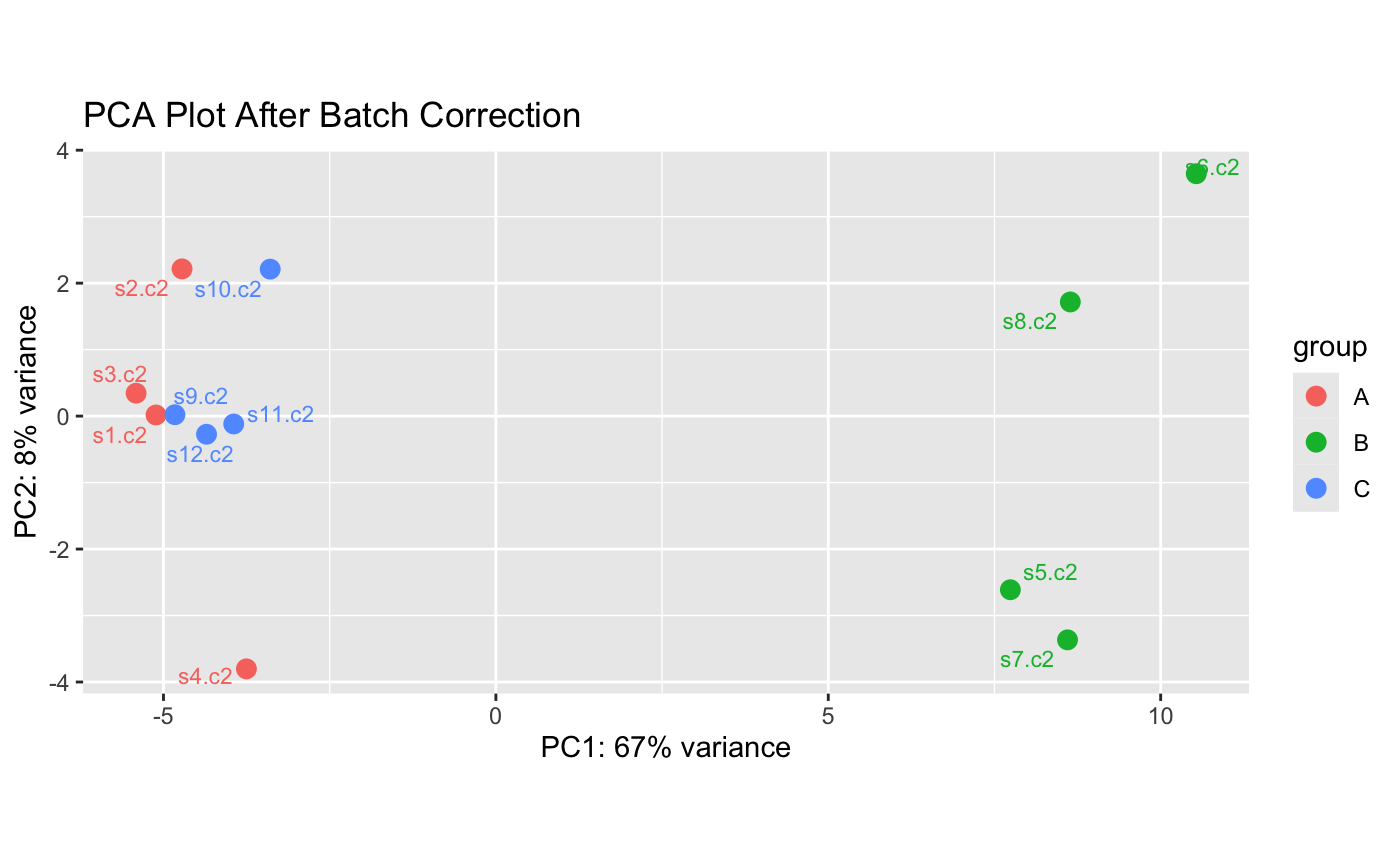
CvsA.





*PCA before and after applying batch correction:*

Batch effects can cause surprising effects during differential expression analysis, through batch correction such technical variability is reduced. By comparing the PCA plots from before and after applying batch correction we can explore the batch effect.



The above plots depict the PCA plots after and before batch correction is applied, before batch correction there was a clear overlap between the groups A and C, but after the correction was applied variance was better spread out, PC1 went from 69% before correction to 67% after correction and PC2 went from 7% to 8%. The difference between group A and C, though slight, became clearer after the correction was applied.

*Conclusion:*

The analysis shows that variance stabilisation, dispersion shrinkage and LFC shrinkage play key roles in differential expression analysis of genes and transcripts. The PCA and MA plots provide real evidence as to how different processes on the data can affect its interpretation.

Using rlog transformations, batch correction and the right null hypotheses plays an important role in accurately spotting and understanding relevant changes in expression