MS Biostatistics Capstone Project: Differential Gene Expression on GSE98692 dataset

Sumia Tahir

4/24/2020

BACKGROUND

Glioma is the general term for tumors originating in the glial cells of the brain. Glial cells, including astrocytes, oligodendrocytes and ependymal cells, surround and support neurons. About a third of all brain tumors (primary and secondary) are gliomas. Glioblastoma multiforme (GBM, or glioblastoma) is a type of malignant, grade IV glioma, consisting of tumors of astrocytes. GBM is the most aggressive and most common malignant tumor of the central nervous system, accounting for 52% of all primary brain tumors. The incidence of GBM is 2-3 per 100,000 adults per year, with a very poor prognosis of only 15 months life-expectancy after diagnosis. Currently, there is no standard therapy for glioblastoma and treatments center around supportive care.

Glioblastoma is characterized by high heterogeneity, meaning that cells may exhibit differences in cellular morphology, gene expression, metabolism, motility, proliferation, and metastatic potential. This is one of the reasons that GBMs present treatment challenges. Other complicating factors include the brain's limited capacity for self-repair, migration of malignant cells into adjacent brain tissue, neurotoxicity of treatments, and tumor capillary leakage resulting in intracranial hypertension.

HSR-GBM1 refers to cancer cells from the human glioblastoma multiforme cell line (Sun et al, 2018). These stem-like cancer cells fail to differentiate completely and are stuck replicating in a pseudo-differentiated state. One proposed factor is global methylation of the cell's DNA. DNA methylation is a form of epigenetic modification in which a methly group is added to the cytosine base within a cytosine-guanine nucleotide base-pair to form 5-methylcytosine. This can alter the transcriptional activity of certain DNA segments, without changing their nucleotide sequence.

HSR-GBM1 cells are characterized by a hypermethylated genome. However, global DNA methylation may be reversed by the addition of demethylation agents such as 5-Azacitadine and Vitamin C, allowing cells to undergo differentiation into mature cells and thereby, blocking tumorigenesis. 5-Azacitadine is a chemical analogue of cytosine that has been used in cancer therapies to block DNA methylation. It acts by inhibiting the enzyme, DNA methlytransferase, that causes global DNA methylation. Vitamin C can also be used as a demethylating agent. It increases the conversion of 5-methylcytosine to the DNA demethylated state of 5-hydroxy-methylcytosine through enhancement of the activity of the TET1 enzyme.

Introduction to Dataset

This dataset was taken from the Gene Expression Omnibus (GEO) repository as series GSE98692.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98692

The data includes RNA-sequence counts from samples of Glioblastoma HSR-GBM1 cells, under three conditions: untreated control group, treated with 0.5 M of 5-Azacitadine for 48 hours, and treated with 100 g/ml of Vitamin C (l-ascorbic acid) for 72 hours.

For each condition, there are 3 biological replicates. The total number of samples is 9, and the total number of genes is 28395.

The following datafile was downloaded and unzipped. https://ftp.ncbi.nlm.nih.gov/geo/series/GSE98nnn/GSE98692/suppl/GSE98692_Results.csv.gz

In the proceeding analysis, our goal is to identify genes that are differentially expressed between the untreated (control) glioblastoma cells, and glioblastoma cells treated with either Vitamin C or 5-Azacitadine.

Differential Gene Expression Analyses with *edgeR*

The *edgeR* package is available through the Bioconductor website for differential analysis of read counts for pre-defined genomic features including genes, exons, transcripts or tags. The read counts follow quadratic mean-variance relationships that are captured by negative binomial based models in *edgeR*. The package features two complementary sets of methods: *classic edgeR* uses exact statistical methods for multigroup experiments, while *glm edgeR* employs statistical methods based on generalized linear models (glms) with the capability to handle complex experimental designs with multiple treatment factors. Both approaches provide estimation of gene-specific biological variation through empirical Bayes methods, even in cases of minimal biological replicates. Under the glm framework, likelihood ratio tests or quasi-likelihood F-tests can be used to test for differential expression.

DATA ANALYSIS & RESULTS

Installing R packages

```
#if (!requireNamespace("BiocManager", quietly = TRUE))
#  install.packages("BiocManager")
#BiocManager::install("edgeR")
#BiocManager::install("org.Hs.eg.db")
#BiocManager::install("GO.db")
#BiocManager::install("KEGGREST")
#install.packages("ggplot2")
#install.packages("RColorBrewer")
#install.packages("gplots")

library("edgeR")
```

```
## Loading required package: limma
```

```
library("org.Hs.eg.db")
```

```
## Loading required package: AnnotationDbi

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
```

```
##     parLapplyLB, parRapply, parSapply, parSapplyLB
## The following object is masked from 'package:limma':
##
##     plotMA
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
## Loading required package: IRanges
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:base':
##
##     expand.grid
##
## Attaching package: 'IRanges'
## The following object is masked from 'package:grDevices':
##
##     windows
##
library("GO.db")
##
library("KEGGREST")
library("ggplot2")
library("RColorBrewer")
library("gplots")
##
## Attaching package: 'gplots'
## The following object is masked from 'package:IRanges':
##
##     space
```

```
## The following object is masked from 'package:S4Vectors':
##
##     space
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

Loading the datafile

```
csvfile <- read.csv("GSE98692_results.csv")
csvfile[1:2,]
```

```
##   GeneID logFC.GBM.VitC logFC.GBM.5Aza   logCPM       LR PValue FDR GBM1 GBM2
## 1  10397      -3.677827      -4.091513 6.984862 1721.743      0   0 6981 5867
## 2    768      -5.699938      -6.126855 3.339448 1535.649      0   0  591  534
##   GBM3 GBM.VitC1 GBM.VitC2 GBM.VitC3 GBM.5Aza1 GBM.5Aza2 GBM.5Aza3
## 1 7275       539       547       594       439       369       375
## 2  639         8        11        17         7         8        10
```

The datafile already included some columns of analyzed data (columns 2-7 above) so those were removed after importing into R. The updated dataset only includes the GeneID and gene counts for the 9 samples.

```
data0 <- csvfile[ , c(1, 8:16)]
data0[1:2,]
```

```
##   GeneID GBM1 GBM2 GBM3 GBM.VitC1 GBM.VitC2 GBM.VitC3 GBM.5Aza1 GBM.5Aza2
## 1  10397 6981 5867 7275       539       547       594       439       369
## 2    768  591  534  639         8        11        17         7         8
##   GBM.5Aza3
## 1       375
## 2        10
```

```
dim(data0)
```

```
## [1] 28395    10
```

Gene Annotation

We will add some information about the genes by using the GeneID column which are the Entrez Gene Identifiers. These are mapped to Gene Symbols and Gene Names in the NCBI database by the "org.Hs.eg.db" package.

```
gene_tab <- toTable(org.Hs.egSYMBOL)
#head(gene_tab)
match <- match(data0$GeneID, gene_tab$gene_id)
data0$GeneSymbol <- gene_tab$symbol[match]

name_tab <- toTable(org.Hs.egGENENAME)
#head(name_tab)
match <- match(data0$GeneID, name_tab$gene_id)
data0$GeneName <- name_tab$gene_name[match]
data0[1:2, c(1,11:12)]
```

```
##   GeneID GeneSymbol                     GeneName
## 1  10397      NDRG1 N-myc downstream regulated 1
```

```
## 2    768        CA9         carbonic anhydrase 9
```

Creating DGEList object

A DGEList object is created from the data for use with the "edgeR" package. The library size (lib.size) is the sum of all the genewise counts for one sample.

```
group_labs <- c(rep("Control",3), rep("VitaminC", 3), rep("Azacitadine", 3))
dlist <- DGEList(counts=data0[,2:10], genes=data0[ ,c(1,11,12)], group=group_labs)
dlist
```

```
## An object of class "DGEList"
## $counts
##    GBM1  GBM2  GBM3 GBM.VitC1 GBM.VitC2 GBM.VitC3 GBM.5Aza1 GBM.5Aza2 GBM.5Aza3
## 1  6981  5867  7275       539       547       594       439       369       375
## 2   591   534   639         8        11        17         7         8        10
## 3 51108 44585 51087      4779      4675      5358      5293      4377      4246
## 4   160   147   189         3         4         3         2         1         3
## 5   496   431   548        34        20        17        48        21        26
## 28390 more rows ...
##
## $samples
##                 group lib.size norm.factors
## GBM1          Control 20354614            1
## GBM2          Control 18655424            1
## GBM3          Control 21514343            1
## GBM.VitC1    VitaminC 20749727            1
## GBM.VitC2    VitaminC 19979574            1
## GBM.VitC3    VitaminC 23372270            1
## GBM.5Aza1 Azacitadine 20434068            1
## GBM.5Aza2 Azacitadine 20421040            1
## GBM.5Aza3 Azacitadine 19492425            1
##
## $genes
##   GeneID GeneSymbol                                       GeneName
## 1  10397      NDRG1                  N-myc downstream regulated 1
## 2    768        CA9                          carbonic anhydrase 9
## 3   1116     CHI3L1                            chitinase 3 like 1
## 4   7052       TGM2                            transglutaminase 2
## 5  56901   NDUFA4L2 NDUFA4 mitochondrial complex associated like 2
## 28390 more rows ...
```

Some Entrez Gene Identifiers map to the same Gene Symbol. Therefore, for multiple entries of Gene symbol, the RNA-sequence with the highest counts are kept. Also, if an Entrez Gene ID does not map to an official gene symbol, that is also dropped from the analysis.

```
sum(duplicated(dlist$genes$GeneSymbol))
```

```
## [1] 509
```

```
s <- order(rowSums(dlist$counts), decreasing=TRUE)
dlist <- dlist[s,]
dupl <- duplicated(dlist$genes$GeneSymbol)
dlist <- dlist[!dupl,]
dim(dlist)
```

```
## [1] 27886     9
```

```
dlist <- dlist[!is.na(dlist$genes$GeneSymbol), ]
dim(dlist)
```

```
## [1] 27885     9
```

A total of 509 RNA-sequences(rows) were removed due to multiple entries of Gene symbol. One Entrez Gene ID did not map to a gene symbol so was removed.

Filtering of lowly-expressed genes

The counts data is further simplified by filtering out genes with relatively low counts across all samples. Biologically, if the number of gene transcripts is low, then that gene will not likely be translated into a protein. Statistically, for differential expression analysis, genes with low counts are unlikely to be signficant because there is not enough statistical evidence for making reliable judgments.

Only those genes are kept that have adequate counts in at least three samples (because the smallest group size is 3 for this dataset). A rough estimate for the cut-off count is 10/L where L is the smallest library size in counts per million. The filterByExpr() function in edgeR automatically determines the lowly expressed genes.

```
keep <- filterByExpr(dlist, group=dlist$samples$group)
summary(keep)
```

```
##    Mode   FALSE    TRUE
## logical   13873   14012
```

```
dlist2 <- dlist[keep, , keep.lib.sizes=FALSE]
dim(dlist2)
```

```
## [1] 14012     9
```

About half of the genes have been removed due to low counts, and the library sizes have been recalculated to account for this. 14074 genes are remaining.

We can compare density plots with the original gene pool (raw data) and the filtered gene pool.

```
#Before applying low count filter
L <- mean(dlist$samples$lib.size) * 1e-6
M <- median(dlist$samples$lib.size) * 1e-6
lcpm.cutoff <- log2(10/M + 2/L)

x <- dlist
lcpm <- cpm(x, log=TRUE)
nsamples <- ncol(x)
col <- brewer.pal(nsamples, "Paired")
par(mfrow=c(1,2))

plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="", xlab="")
title(main="Full gene set", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)

for (i in 2:nsamples){
  den <- density(lcpm[,i])
  lines(den$x, den$y, col=col[i], lwd=2)
}
```

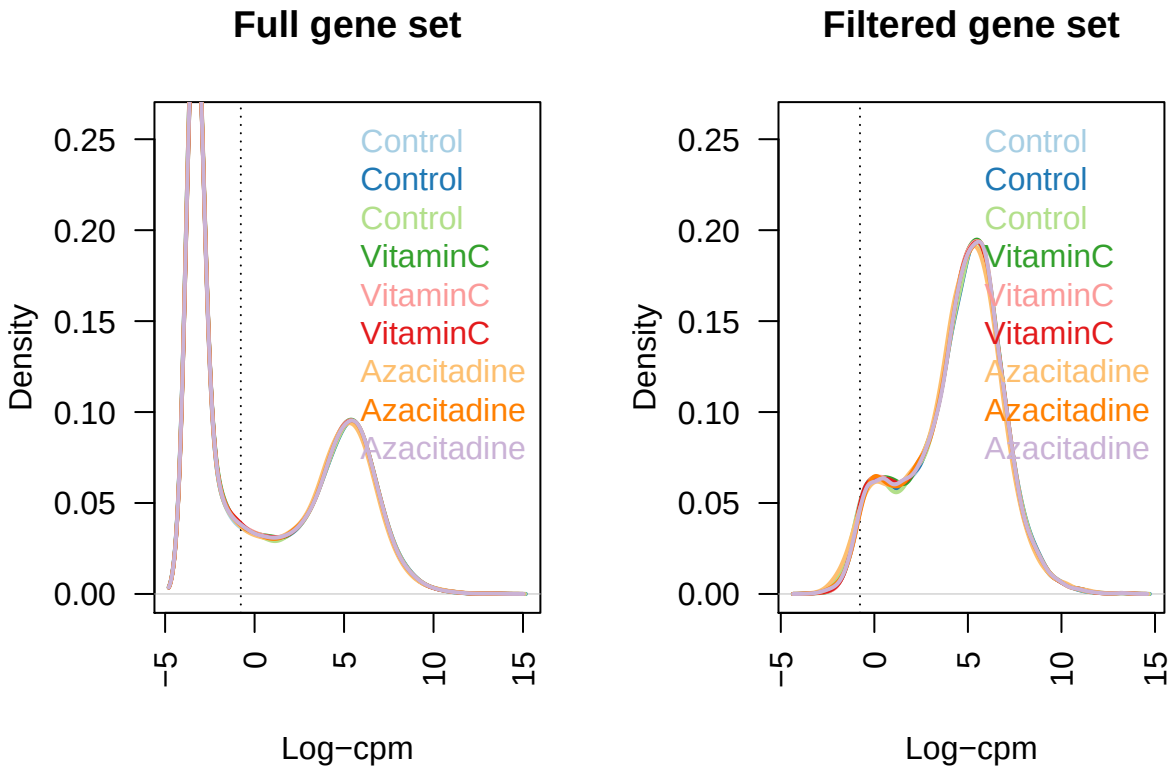
```
snames <- dlist$samples$group
legend("topright", text.col=col, bty="n", legend=snames)


#After filtering
x <- dlist2
L <- mean(dlist2$samples$lib.size) * 1e-6
M <- median(dlist2$samples$lib.size) * 1e-6
lcpm.cutoff <- log2(10/M + 2/L)

lcpm <- cpm(x, log=TRUE)
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="", xlab="")
title(main="Filtered gene set", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)

for (i in 2:nsamples){
  den <- density(lcpm[,i])
  lines(den$x, den$y, col=col[i], lwd=2)
}

legend("topright", legend=snames, text.col=col, bty="n")
```



The gene counts are transformed into log-2 counts per million (Log-cpm). The plot of raw data shows that there are many genes that have very low counts (below CPM = 0.2, the vertical line). After these genes are removed, the distribution appears unimodal and a larger density of genes have higher expressions.

Normalization

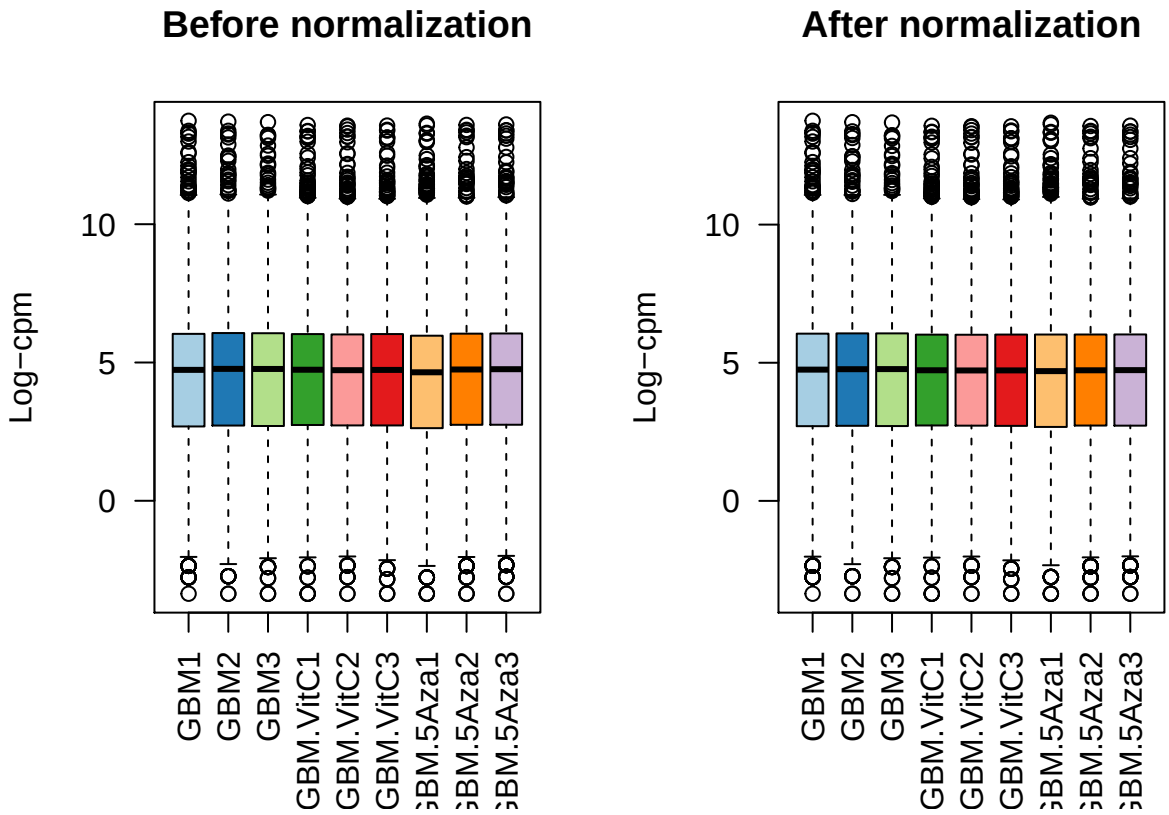
We proceed to normalize the gene counts to eliminate composition biases between the samples (libraries). This is done by the trimmed mean of M values (TMM) method. The norm.factor column is updated with the normalization factors for the samples. Normalization reduces the effect of genes that seem to be artificially overexpressed due to a larger library size, or underexpressed due to a smaller library size. It ensures that all samples have a similar range and distribution of expression counts.

```
par(mfrow=c(1,2))
x <- dlist2
lcpm <- cpm(x, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
title(main="Before normalization", ylab="Log-cpm")

dlist3 <- calcNormFactors(dlist2)
dlist3$samples
```

```
##                 group lib.size norm.factors
## GBM1          Control 20338100    0.9877836
## GBM2          Control 18639808    1.0022595
## GBM3          Control 21496819    0.9974865
## GBM.VitC1    VitaminC 20730283    1.0078265
## GBM.VitC2    VitaminC 19960731    1.0014803
## GBM.VitC3    VitaminC 23350628    1.0062080
## GBM.5Aza1 Azacitadine 20416353    0.9649161
## GBM.5Aza2 Azacitadine 20401954    1.0147984
## GBM.5Aza3 Azacitadine 19474119    1.0182763
```

```
x <- dlist3
lcpm <- cpm(x, log=TRUE)
boxplot(lcpm, las=2, col=col, main="")
title(main="After normalization", ylab="Log-cpm")
```

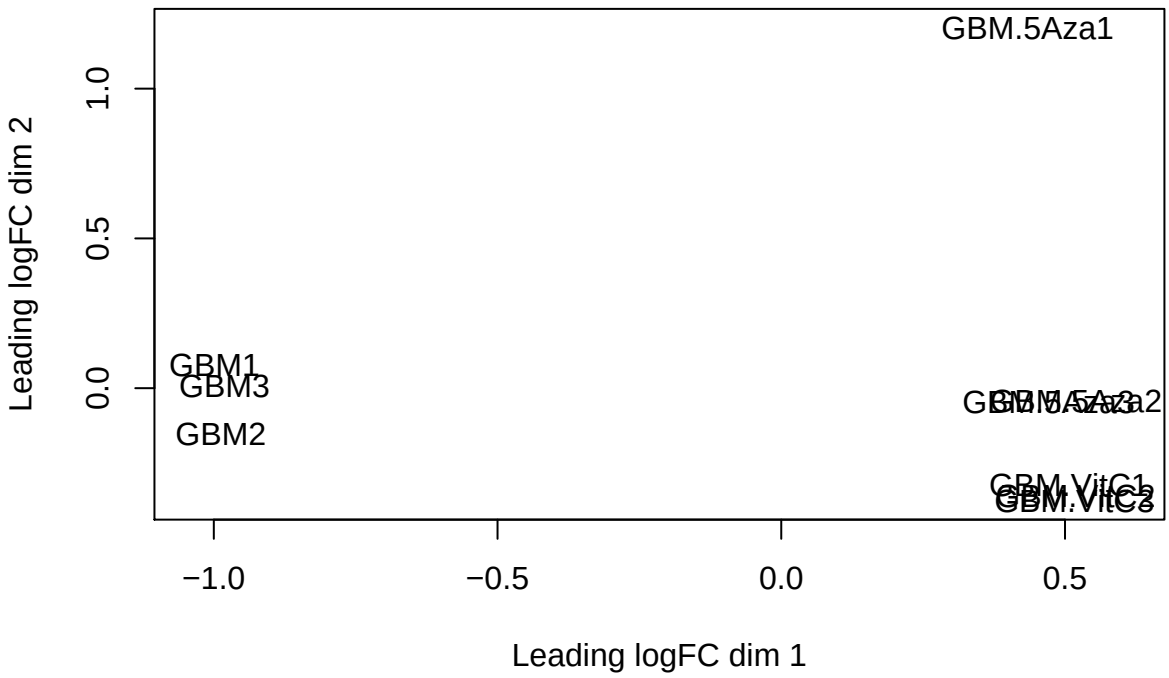


All of the normalization factors are close to 1, and the box plots show no difference pre- and post-normalization.

Multidimensional Scaling plots

Multidimensional scaling (MDS) plots visualize the principal components that give rise to the variation in expression profiles between the samples in two dimensions. The distance between each pair of samples is calculated as the root-mean-square of the largest 500 log-fold changes between genes from the two samples. The log-fold change is in base 2.

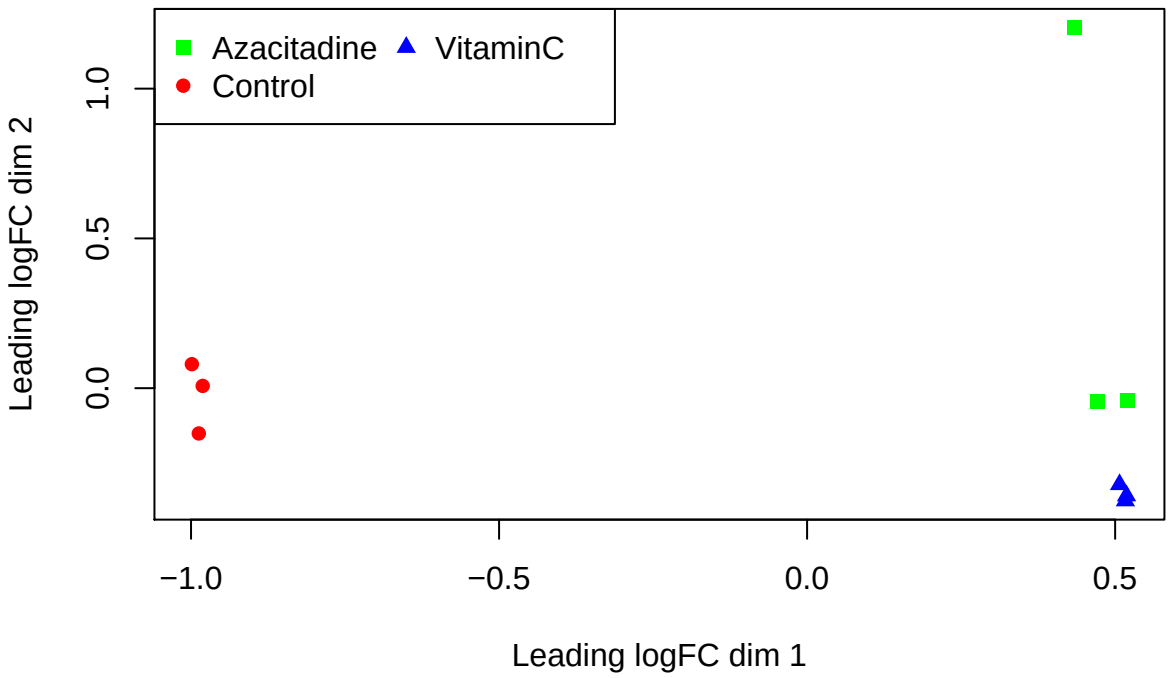
```
plotMDS(dlist3)
```



The plot above shows that the differences between the groups are larger than the differences within the samples in each group. Most noticeable, the control group is 1.5 units away from both of the treated groups along the first dimension, which corresponds to a 3-fold difference. In dimension 2, the samples within each group are clustered very close together except for sample 1 from the 5-Azacitadine group. That appears to be an outlier.

We can see the clustering of the groups more clearly in the plot below. It is worth noting that the two treated groups are fairly similar (except for the outlier).

```
pch <- c(15,16,17)
colors <- c("green", "red", "blue")
group <- dlist3$samples$group
plotMDS(dlist3, col=colors[group], pch=pch[group])
legend("topleft", legend=levels(group), pch=pch, col=colors, ncol=2)
```



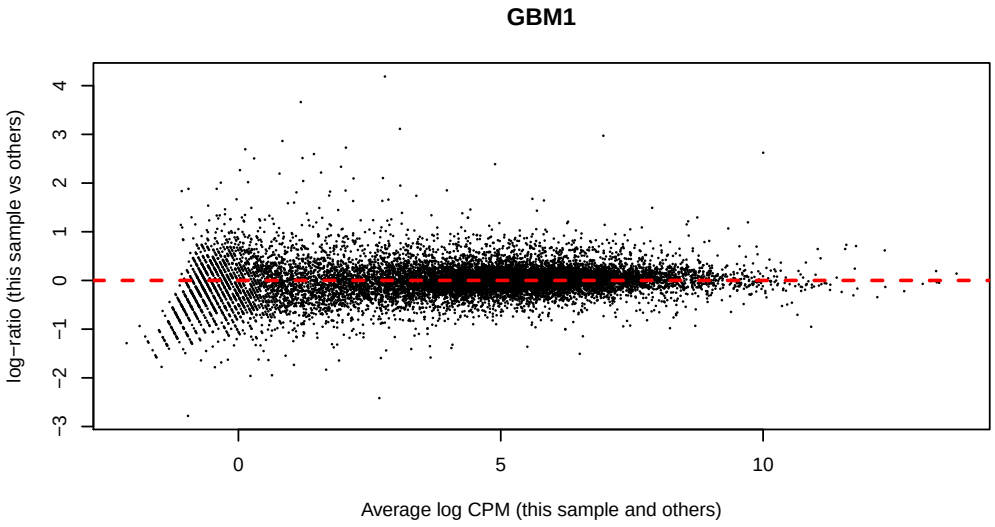
Mean Difference plots

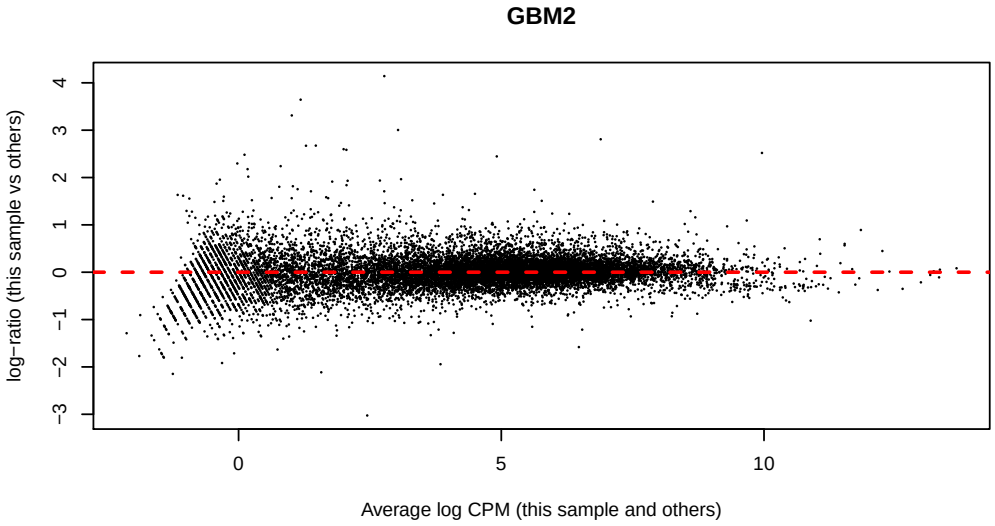
We can also look at mean-difference (MD) plots for all the samples. This plot shows the difference between two libraries (the library size-adjusted log-fold change) against the mean of the libraries (the average log-expression across those libraries). Each sample is compared to a reference library which is the average of all the other samples. The red line (at 0) indicates zero log-fold change.

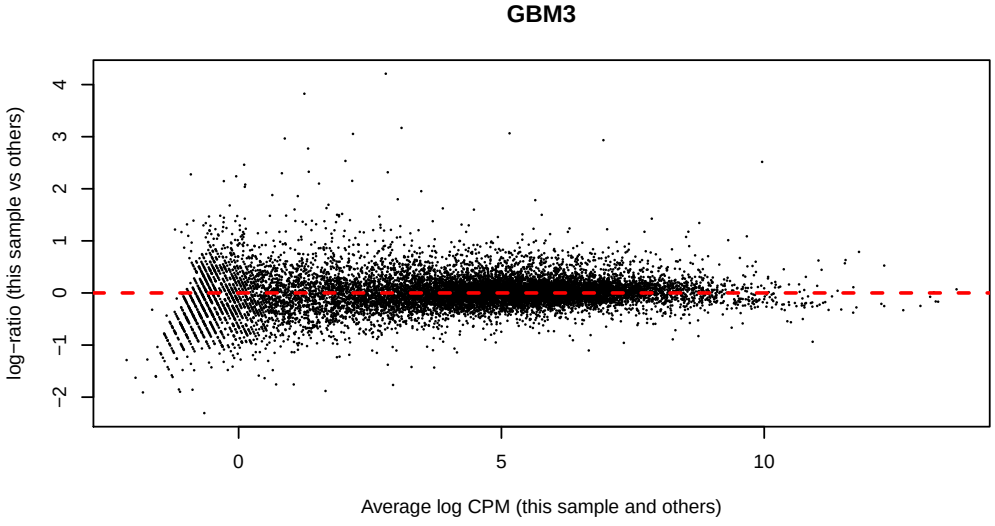
HSR-GBM1 Control group

```
layout.matrix <- matrix(1:3, nrow = 3, ncol = 2)
layout(mat=layout.matrix, heights = c(4,4,4), widths = 2)

for (i in 1:3) {
plotMD(dlist3, column=i)
abline(h=0, col="red", lty=2, lwd=2)
}
```


GBM1
log−ratio (this sample vs others)
4
3
2
1
0
−1
−2
−3
0
5
10
Average log CPM (this sample and others)

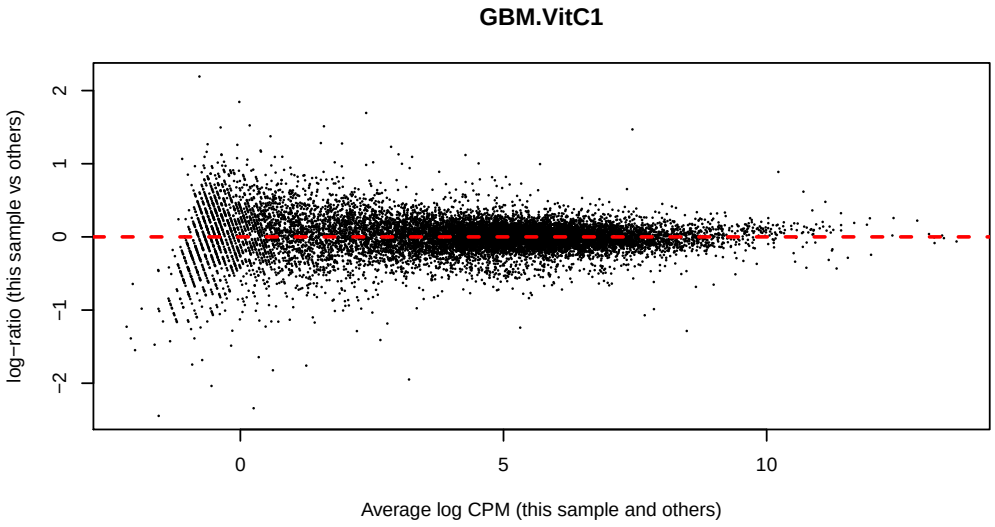

GBM2
log−ratio (this sample vs others)
4
3
2
1
0
−1
−2
−3
0
5
10
Average log CPM (this sample and others)

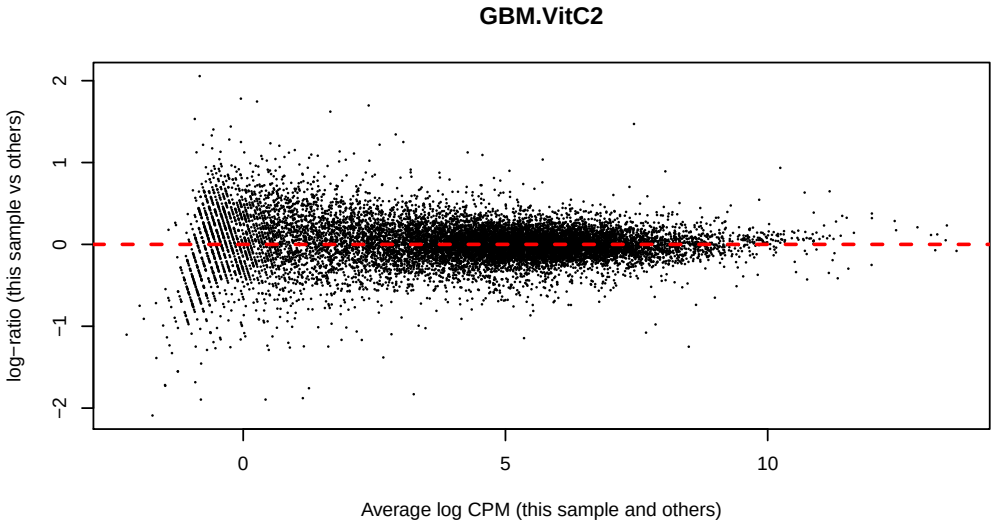

GBM3
log−ratio (this sample vs others)
4
3
2
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)

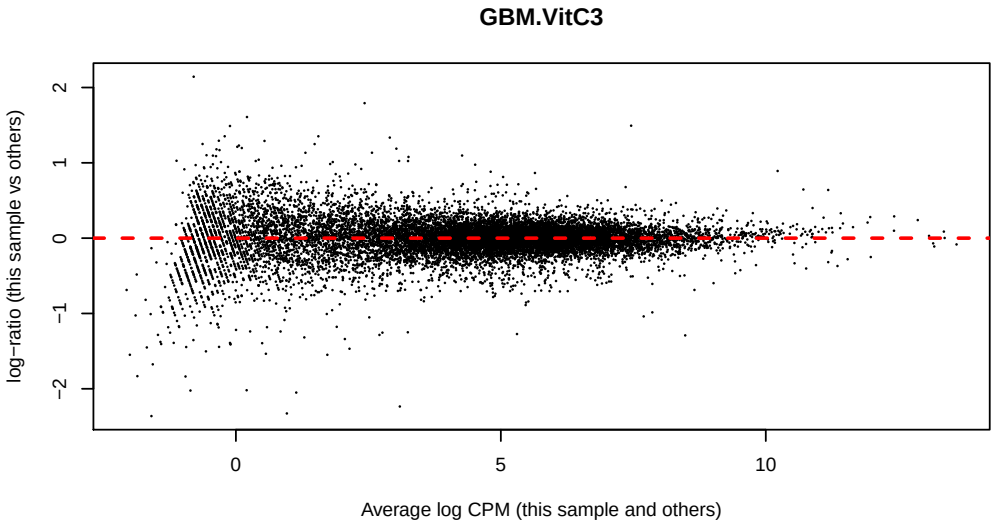
In all the plots, most of the genes are close to the zero log-fold change. The distribution appears similar in all the control replicates.

HSR-GBM1 Vitamin C-treated group

```
layout.matrix <- matrix(1:3, nrow = 3, ncol = 2)
layout(mat=layout.matrix, heights = c(4,4,4), widths = 2)
for (i in 4:6) {
plotMD(dlist3, column=i)
abline(h=0, col="red", lty=2, lwd=2)
}
```


GBM.VitC1
log−ratio (this sample vs others)
2
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)

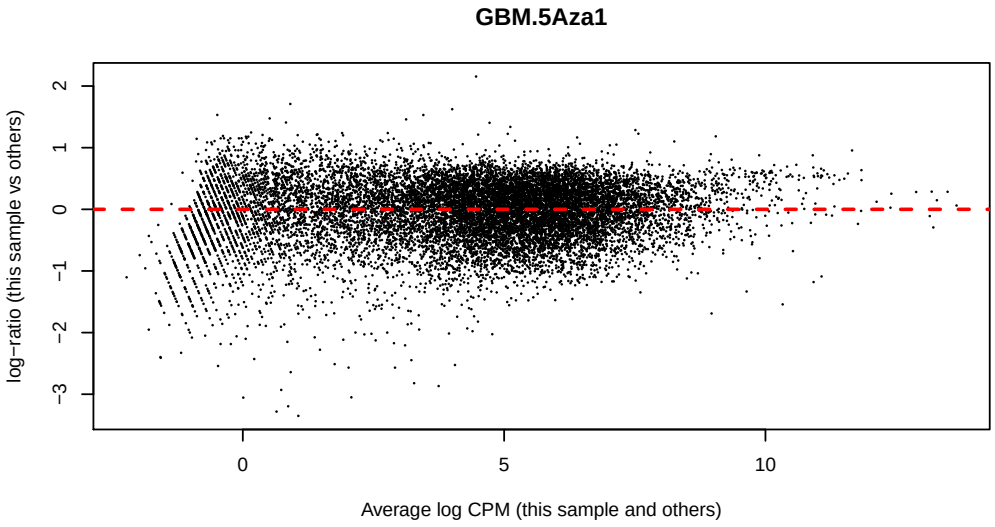

GBM.VitC2
log−ratio (this sample vs others)
2
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)

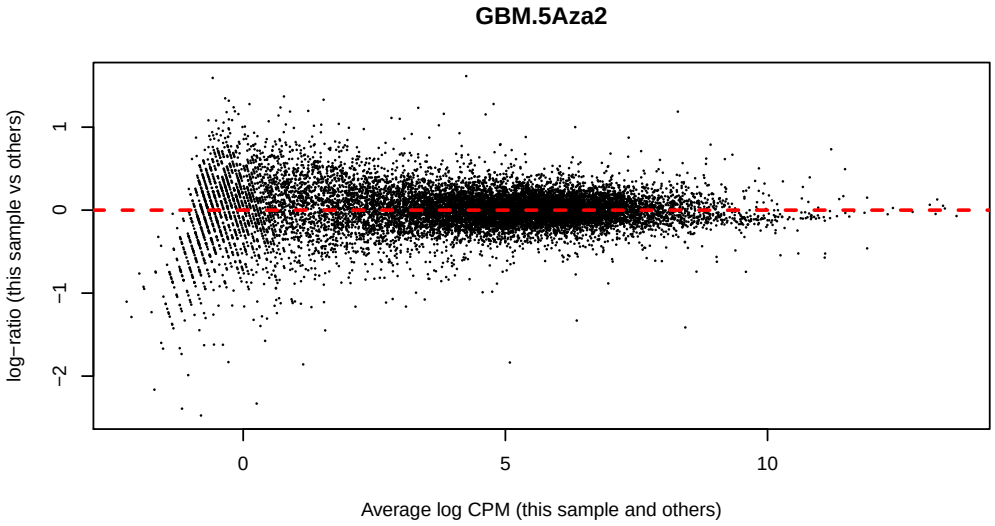

GBM.VitC3
log−ratio (this sample vs others)
2
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)

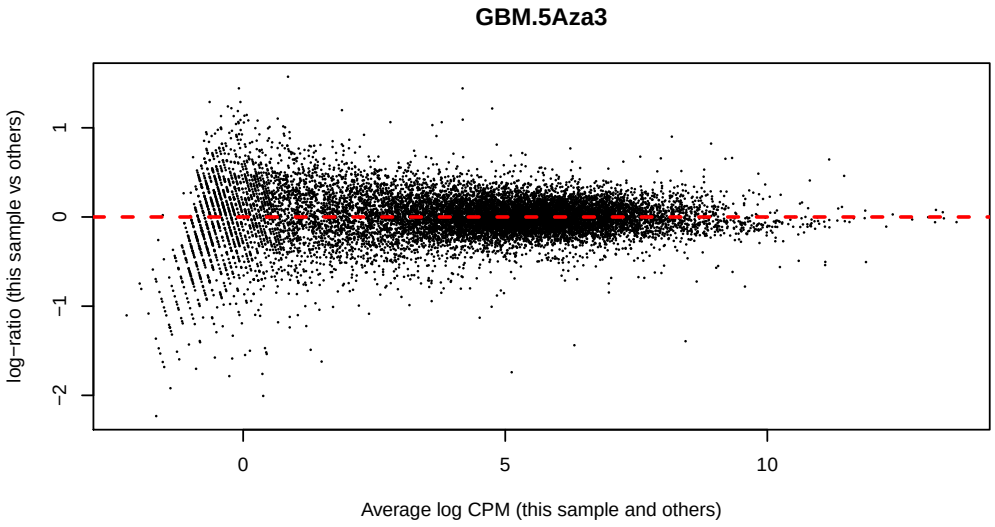
All three samples from the Vitamin C treated group appear similar. The log-fold changes of most genes are close to zero.

HSR-GBM1 5-Azacitadine-treated group

```
layout.matrix <- matrix(1:3, nrow = 3, ncol = 2)
layout(mat=layout.matrix, heights = c(4,4,4), widths = 2)
for (i in 7:9) {
plotMD(dlist3, column=i)
abline(h=0, col="red", lty=2, lwd=2)
}
```


GBM.5Aza1
log−ratio (this sample vs others)
2
1
0
−1
−2
−3
0
5
10
Average log CPM (this sample and others)


GBM.5Aza2
log−ratio (this sample vs others)
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)


GBM.5Aza3
log−ratio (this sample vs others)
1
0
−1
−2
0
5
10
Average log CPM (this sample and others)

The plot of the sample "GBM.5Aza1" has a greater number of genes that are overexpressed (larger log-fold changes) than other samples. This sample was the outlier in the MDS plot, so we will remove it from the dataset before further analysis.

Removal of outlier - GBM.5Aza1

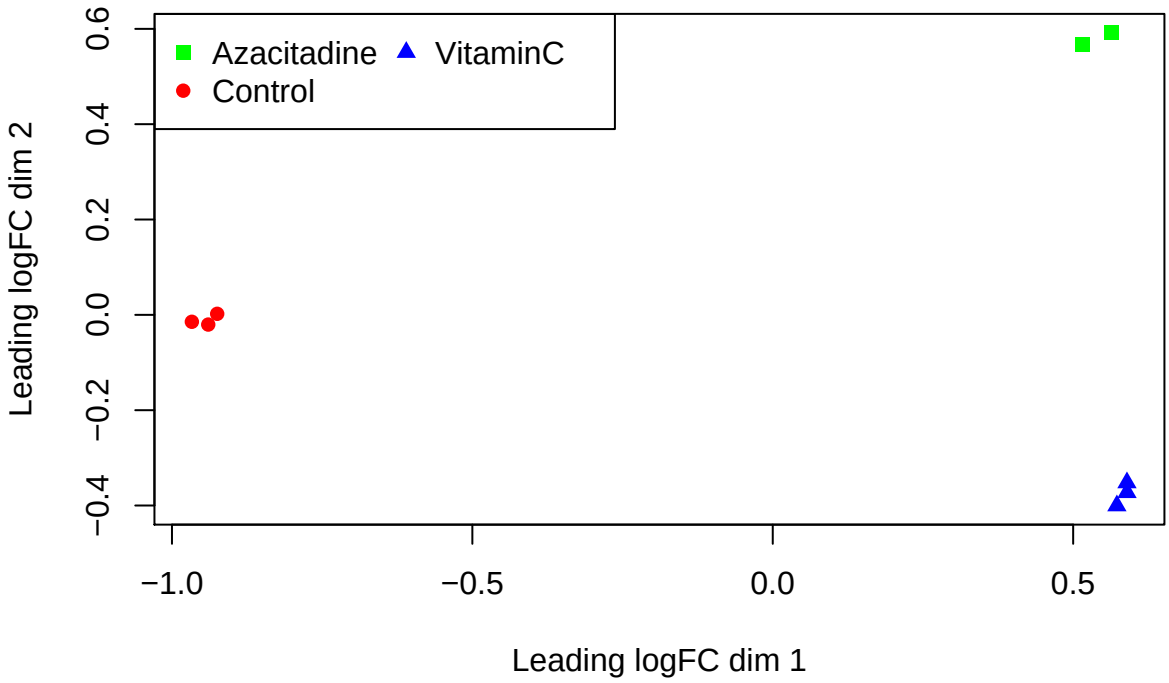
```
dlist3$counts <- dlist3$counts[,-7]
dlist3$samples <- dlist3$samples[-7,]
dlist3$samples
```

```
##                 group lib.size norm.factors
## GBM1          Control 20338100    0.9877836
## GBM2          Control 18639808    1.0022595
## GBM3          Control 21496819    0.9974865
## GBM.VitC1    VitaminC 20730283    1.0078265
## GBM.VitC2    VitaminC 19960731    1.0014803
## GBM.VitC3    VitaminC 23350628    1.0062080
## GBM.5Aza2 Azacitadine 20401954    1.0147984
## GBM.5Aza3 Azacitadine 19474119    1.0182763
```

The samples can be re-normalized and we can make the MDS plot without the outlier sample.

MDS plot without outlier

```
dlist3 <- calcNormFactors(dlist3)
pch <- c(15,16,17)
colors <- c("green", "red", "blue")
group <- dlist3$samples$group
plotMDS(dlist3, col=colors[group], pch=pch[group])
legend("topleft", legend=levels(group), pch=pch, col=colors, ncol=2)
```



The MDS plot shows three tight clusters of groups, indicating that the within-group variation is very small. This is what we would expect of replicates. The first dimension separates the untreated group from the treated groups. Therefore, we can infer that the primary factor that causes variation between the groups is their DNA methylation status. The second dimension distinguishes the Vitamin C-treated group from the 5-Azacitadine group, indicating that the two treatments are not identical in their effects on DNA demethylation.

Estimating dispersion of gene counts between replicates

The read counts for each gene in each sample are modelled by a negative binomial distribution. The dispersion of the negative binomial distribution is also known the biological coefficient of variation (BCV), the variability between biological replicates. The BCV (square-root of dispersion parameter) needs to be estimated before a negative binomial model can be fitted.

We implement the estimateDisp() function in edgeR to estimate three types of dispersions: 1) the tagwise dispersion is the gene-specific dispersion after an empirical Bayes strategy is applied to squeeze the tagwise dispersions towards a global trend or common value, 2) the common dispersion assumes all genes have the same dispersion (average over all genes), and 3) the trended dispersion assumes that all gene's with the same abundance have the same variance

A design matrix for the samples is first generated.

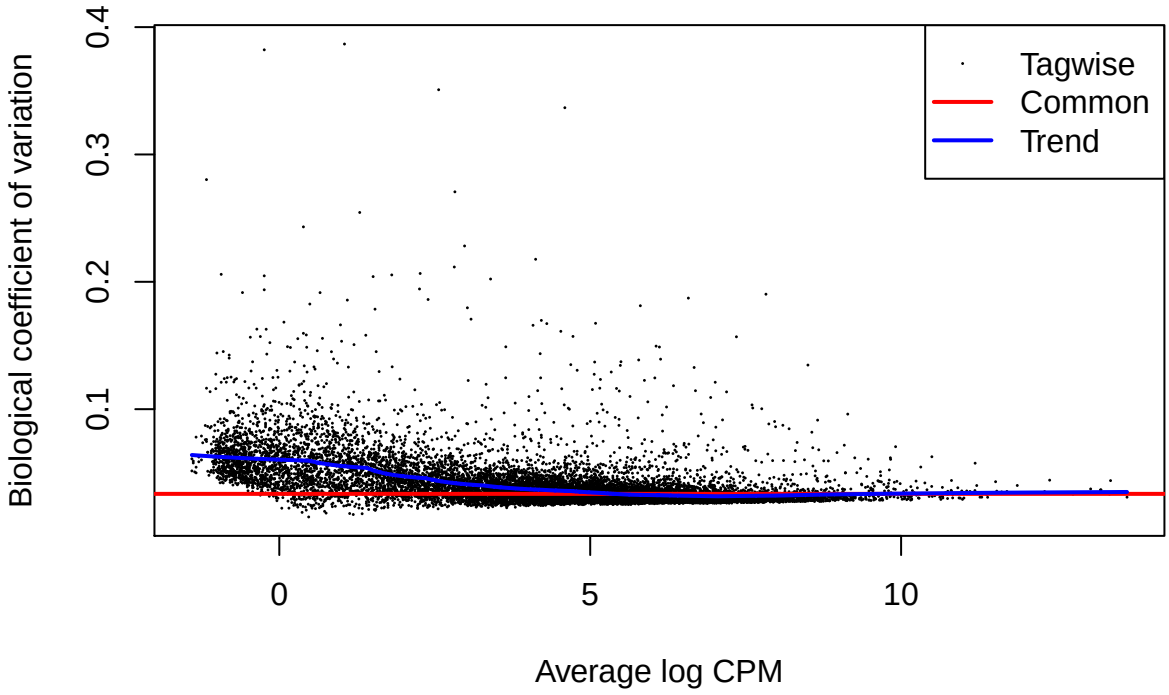
```
group <- dlist3$samples$group
design <- model.matrix(~0+group)
colnames(design) <- levels(group)
design
```

```
##   Azacitadine Control VitaminC
```

```
## 1           0       1       0
## 2           0       1       0
## 3           0       1       0
## 4           0       0       1
## 5           0       0       1
## 6           0       0       1
## 7           1       0       0
## 8           1       0       0
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Dispersion Plot

```
dlist3 <- estimateDisp(dlist3, design, robust=TRUE)
plotBCV(dlist3)
```



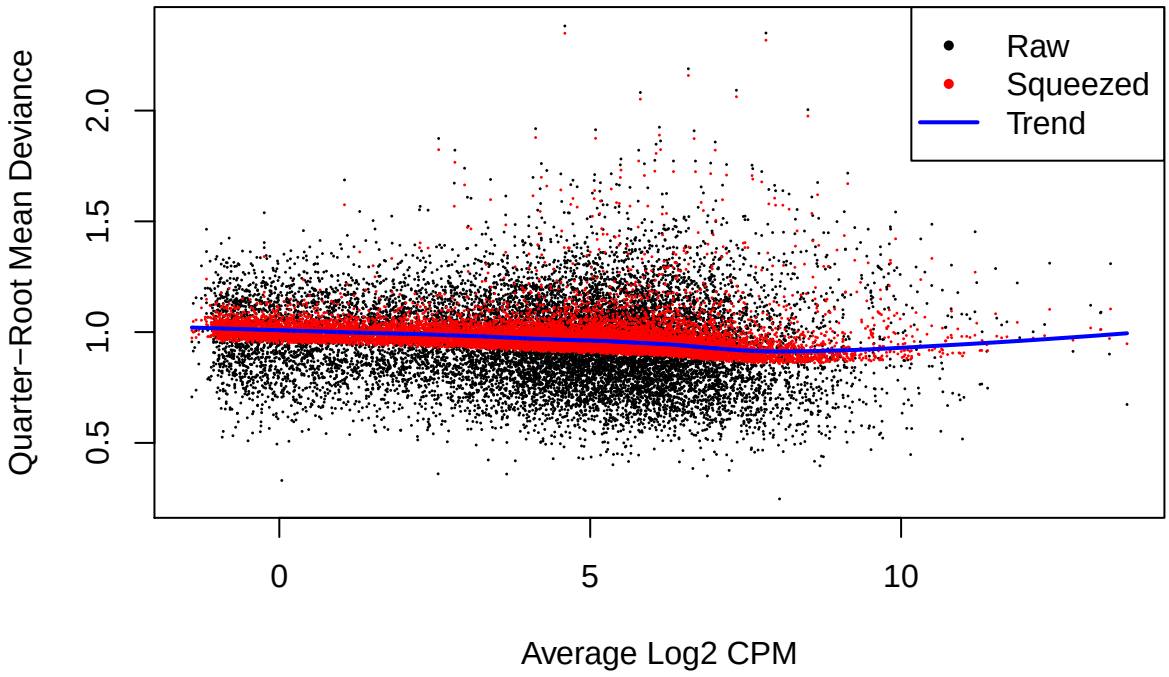
The BCV plot visualizes the dispersion estimates. The individual dispersions tend to be higher for the low gene counts. The trended dispersion asymptotically approaches the common dispersion as the gene abundance increases. The common BCV estimate between replicates is near zero, indicating that there is no biological variation in the true abundance between genes in replicate samples.

Along with variation due to biological sources, variation in gene abundance may also arise from technical sources. To account for the effects of both sources on gene-specific variability, quasi-Likelihood (QL) methods can be applied to the negative binomial (NB) model. The quasi-likelihood NB model includes a global

dispersion parameter as well as a gene-specific dispersion parameter. The global parameter is estimated from the NB trended dispersion to describe overall biological variability across all genes, and the QL dispersion parameter adds gene-specific variability above and below the overall level. The gene-specific dispersion estimates are stabilized by an Empirical Bayes approach by squeezing the raw estimates towards a global trend.

Quasi-Likelihood dispersion estimates

```
fit <- glmQLFit(dlist3, design, robust=TRUE)
plotQLDisp(fit)
```



The QL gene-specific dispersion estimates are fairly uniformly distributed across all gene counts, as seen by the red points clustered around the trend line.

Differential Expression Analysis

The differential expression analysis is performed after obtaining dispersion estimates and fitting NB generalized linear models. The QL F-test is preferred over the LRT in the case of a small number of replicates because it provides more robust and reliable error rate control by accounting for the uncertainty in gene-specific dispersion estimation.

The differential expression is performed between the different treatment groups. Differentially-expressed genes are ranked according to the p-value for that gene.

Control group vs Vitamin C-treated group

```
cc <- makeContrasts(Control - VitaminC, levels=design)
de_cc <- glmQLFTest(fit, contrast=cc)
is.de <- decideTestsDGE(de_cc)
summary(is.de)
```

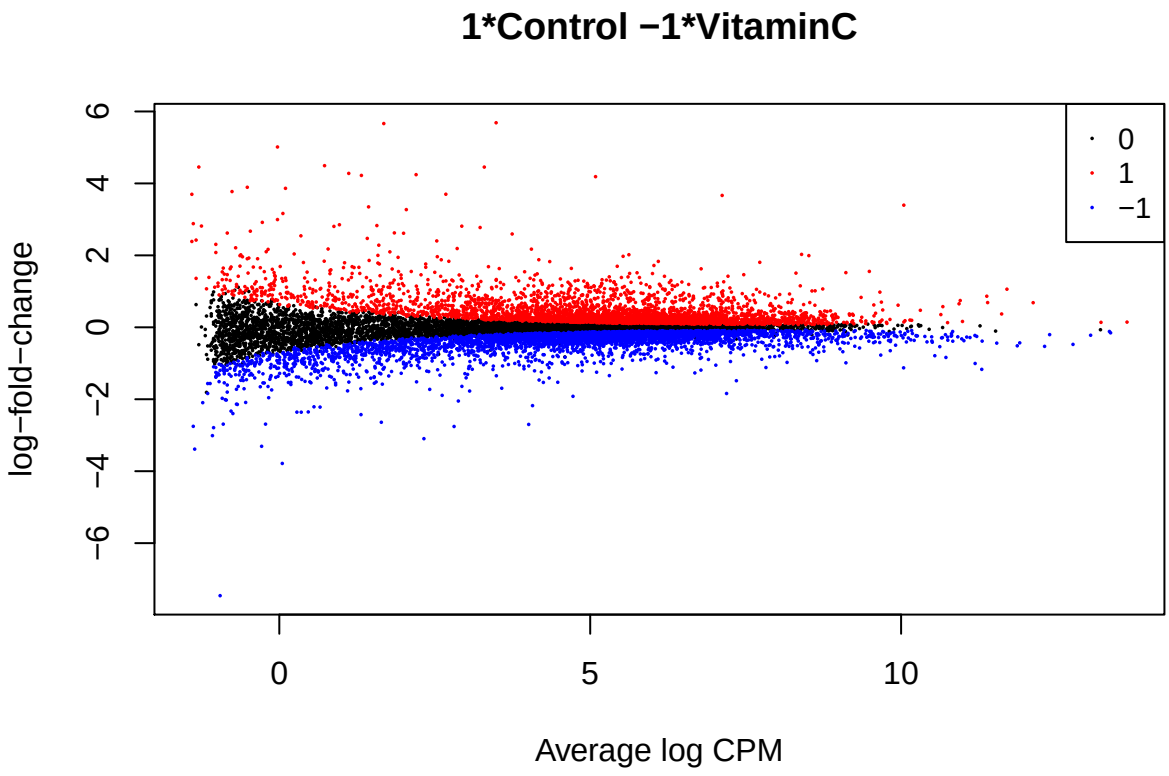
```
##        1*Control -1*VitaminC
## Down                    3669
## NotSig                  6630
## Up                      3713
```

The function decideTestsDGE() shows the total number of DE genes identified at a false decovery rate (FDR) of 5%, corrected for multiple testing by the Benjamini-Hochberg method. There are a total of 7382 genes that are differentially-expressed between the control group and the Vitamin C-treated group.

MD plot - Control group vs Vitamin C-treated group

(FDR <= 0.05, absolute fold change >= 1)

```
plotMD(de_cc, status=is.de, values=c(0, 1,-1), col=c("black","red","blue"),
       cex = 0.25, legend="topright")
```




```
plotMD
```

```
## function (object, ...)
## UseMethod("plotMD")
```

```
## <bytecode: 0x00000000093ad4a8>
## <environment: namespace:limma>
```

The MD plots show the log-fold change in the gene expressions between two treatment groups. The genes in red are upregulated, genes in blue are downregulated and the genes in black are not statistically significant. A higher proportion of genes that have low abundances are non-DE genes. The number of DE genes seems to increase with gene abundance.

Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 1)

```
topTags(de_cc)
```

```
## Coefficient:  1*Control -1*VitaminC
##    GeneID GeneSymbol                                      GeneName     logFC
## 3    1116     CHI3L1                             chitinase 3 like 1  3.395006
## 1   10397      NDRG1                 N-myc downstream regulated 1  3.665712
## 14   7422      VEGFA         vascular endothelial growth factor A  2.026528
## 6       2        A2M                           alpha-2-macroglobulin  1.806370
## 44   3485     IGFBP2 insulin like growth factor binding protein 2  1.992508
## 35   3939       LDHA                       lactate dehydrogenase A  1.519519
## 2     768        CA9                          carbonic anhydrase 9  5.686109
## 23   5230       PGK1                    phosphoglycerate kinase 1  1.553837
## 7    1728       NQO1              NAD(P)H quinone dehydrogenase 1 -1.484934
## 51    183        AGT                                angiotensinogen  1.506213
##       logCPM        F       PValue          FDR
## 3  10.043626 8064.332 4.965166e-30 6.957191e-26
## 1   7.121955 6082.205 1.140579e-28 7.990895e-25
## 14  8.399366 3398.285 7.265424e-26 3.393437e-22
## 6   7.727273 2522.563 1.964038e-24 6.880025e-21
## 44  8.516434 2472.373 6.565465e-24 1.839906e-20
## 35  9.111681 2013.193 2.369263e-23 4.573918e-20
## 2   3.486755 2005.625 2.469744e-23 4.573918e-20
## 23  9.489094 1995.505 2.611429e-23 4.573918e-20
## 7   7.350772 1918.566 4.028444e-23 6.271840e-20
## 51  8.309069 1890.433 4.740724e-23 6.642702e-20
```

Above are the top 10 DE genes ranked by p-value. The genes CHI3L1, NDRG1 and CA9 show the highest upregulation (logFC) in the control group as compared to the Vitamin C group.

Control group vs 5-Azacitadine-treated group

```
caz <- makeContrasts(Control - Azacitadine, levels=design)
de_caz <- glmQLFTest(fit, contrast=caz)
is.de2 <- decideTestsDGE(de_caz)
summary(is.de2)
```

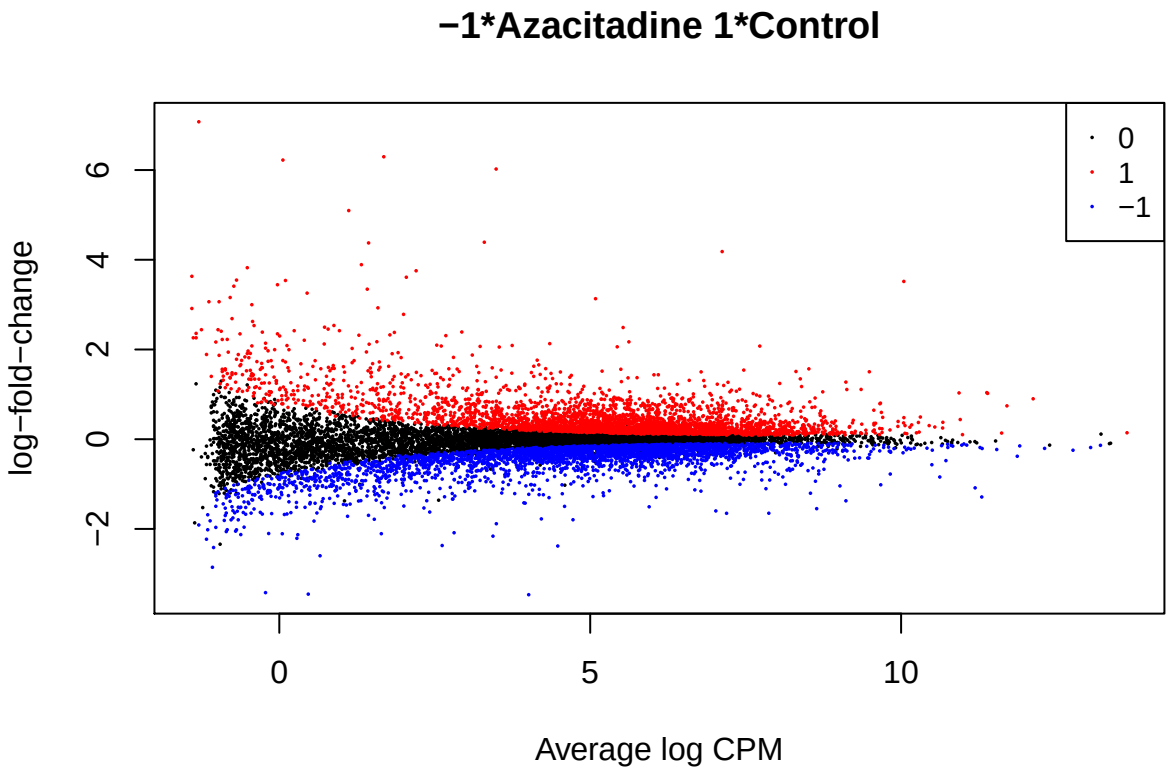
```
##        -1*Azacitadine 1*Control
## Down                     3186
## NotSig                   7367
## Up                       3459
```

There are a total of 6645 DE genes between the control group and the 5-Azacitadine group.

MD plot - Control group vs 5-Azacitadine-treated group

(FDR <= 0.05, absolute fold change >= 1)

```
plotMD(de_caz, status=is.de2, values=c(0, 1,-1), col=c("black","red","blue"),
       cex = 0.25, legend="topright")
```




```
plotMD
```

```
## function (object, ...)
## UseMethod("plotMD")
## <bytecode: 0x00000000093ad4a8>
## <environment: namespace:limma>
```

Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 1)

```
topTags(de_caz)
```

```
## Coefficient:  -1*Azacitadine 1*Control
##    GeneID GeneSymbol                                                GeneName
## 3    1116     CHI3L1                                       chitinase 3 like 1
## 1   10397      NDRG1                             N-myc downstream regulated 1
## 6       2        A2M                                    alpha-2-macroglobulin
## 33   2353        FOS Fos proto-oncogene, AP-1 transcription factor subunit
## 10   6533     SLC6A6                      solute carrier family 6 member 6
## 2     768        CA9                                    carbonic anhydrase 9
## 51    183        AGT                                         angiotensinogen
```

```
## 23   5230       PGK1                              phosphoglycerate kinase 1
## 8    5210     PFKFB4 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4
## 98   5351      PLOD1     procollagen-lysine,2-oxoglutarate 5-dioxygenase 1
##        logFC    logCPM        F       PValue          FDR
## 3   3.516983 10.043626 6079.021 1.147233e-28 1.607502e-24
## 1   4.182366  7.121955 5037.498 9.242676e-28 6.475419e-24
## 6   2.075743  7.727273 2371.408 3.887281e-24 1.815620e-20
## 33 -1.548018  8.642403 1898.565 4.521670e-23 1.583941e-19
## 10  2.489768  5.529005 1634.420 2.353559e-22 6.595614e-19
## 2   6.022555  3.486755 1454.385 8.489556e-22 1.982594e-18
## 51  1.509630  8.309069 1424.920 1.062928e-21 2.127678e-18
## 23  1.502875  9.489094 1407.244 1.219056e-21 2.135176e-18
## 8   2.170414  5.622134 1341.873 2.054633e-21 2.920088e-18
## 98  1.344907  8.378773 1340.139 2.083991e-21 2.920088e-18
```

The genes with the highest upregulation in the control group as compared to the 5-Azacitadine group are CHI3L1, NDRG1 and CA9.
(These are the same as in the comparison of the control group to the Vitamin C group.)

Vitamin C vs 5-Azacitadine treated group

```
ca <- makeContrasts(VitaminC - Azacitadine, levels=design)
de_ca <- glmQLFTest(fit, contrast=ca)
is.de3 <- decideTestsDGE(de_ca)
summary(is.de3)
```

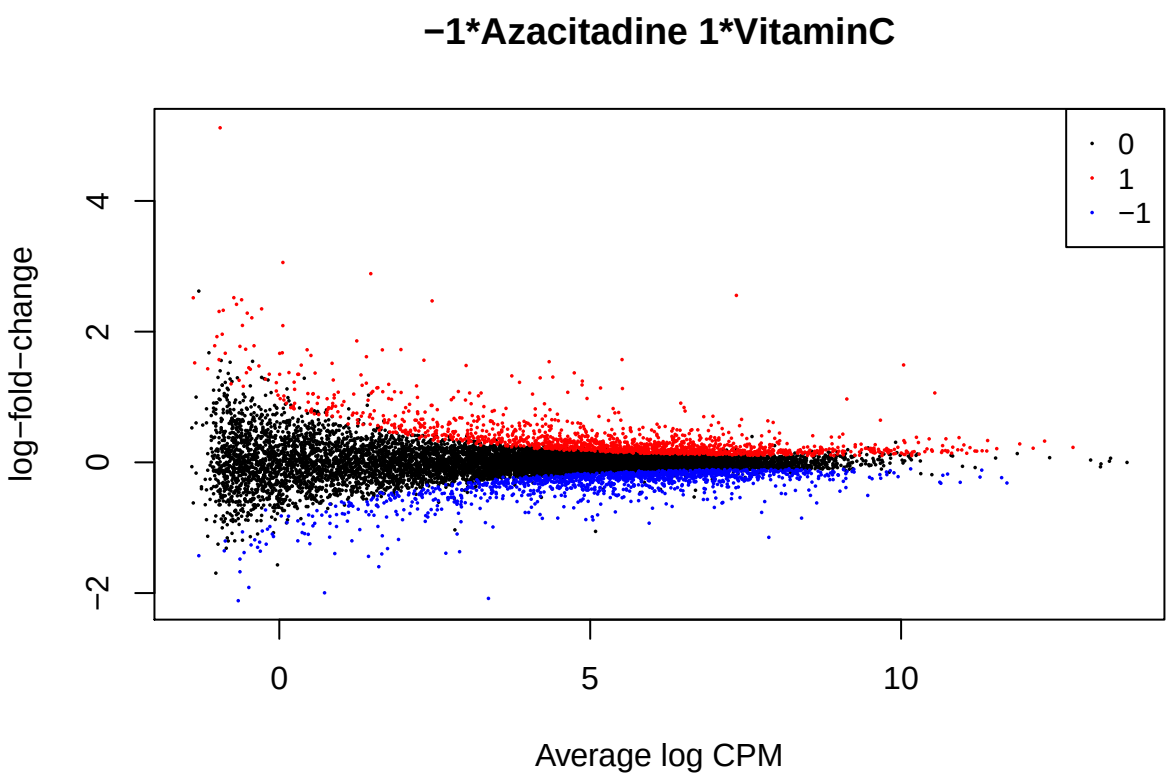
```
##        -1*Azacitadine 1*VitaminC
## Down                         1311
## NotSig                      11200
## Up                           1501
```

There are a total of 2812 DE genes between the Vitamin C-treated group and the 5-Azacitadine-treated group.

MD plot - Vitamin C-treated group vs 5-Azacitadine-treated group

(FDR <= 0.05, absolute fold change >= 1)

```
plotMD(de_ca, status=is.de3, values=c(0, 1,-1), col=c("black","red","blue"),
       cex=0.25, legend="topright")
```




```
plotMD
```

```
## function (object, ...)
## UseMethod("plotMD")
## <bytecode: 0x00000000093ad4a8>
## <environment: namespace:limma>
```

Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 1)

```
topTags(de_ca)
```

```
## Coefficient:  -1*Azacitadine 1*VitaminC
##     GeneID GeneSymbol                                                GeneName
## 7     1728       NQO1                     NAD(P)H quinone dehydrogenase 1
## 193   2512        FTL                                ferritin light chain
## 429   7086        TKT                                        transketolase
## 92  146802    SLC47A2                  solute carrier family 47 member 2
## 126   3488     IGFBP5         insulin like growth factor binding protein 5
## 14    7422      VEGFA                vascular endothelial growth factor A
## 272   2897      GRIK1 glutamate ionotropic receptor kainate type subunit 1
## 142  51655      RASD1                ras related dexamethasone induced 1
## 313    721        C4B                 complement C4B (Chido blood group)
## 60   23544      SEZ6L                     seizure related 6 homolog like
##          logFC    logCPM         F       PValue          FDR
## 7    2.5546851  7.350772 3556.7913 4.384812e-26 6.143999e-22
## 193  1.4902013 10.040229 1096.8100 1.870301e-20 1.310333e-16
```

```
## 429  1.0620019 10.542053  694.4139 2.698350e-18 1.191071e-14
## 92   1.5723634  5.511778  672.4530 3.818856e-18 1.191071e-14
## 126  0.9690846  9.127131  665.8233 4.250180e-18 1.191071e-14
## 14  -0.8529869  8.399366  496.3092 9.989106e-17 2.332789e-13
## 272  0.9059765  6.456186  427.2236 4.930508e-16 9.869468e-13
## 142  1.5399799  4.339798  408.4173 7.947009e-16 1.370092e-12
## 313  1.1296870  5.518375  403.4907 9.036775e-16 1.370092e-12
## 60  -2.0819292  3.363546  400.4965 9.777993e-16 1.370092e-12
```

The genes NQO1 and SEZ6L show the greatest difference in absolute log-fold change. However, these differences are not as large as between DE genes in the control vs treated groups.

Differential Expression analysis *(with logFC cutoff)*

So far, we've included all of the genes that were statistically significant, regardless of the magnitude of the difference in gene expressions.
We can add a parameter (lfc = log-fold change) which would only look at genes that have significantly larger differences than the specified lfc-cutoff.
Control group vs Vitamin C-treated group* # *(FDR <= 0.05, absolute fold change >= 2)*

```
tr1 <- glmTreat(fit, contrast=cc, lfc=1)
is.de4 <- decideTestsDGE(tr1)
summary(is.de4)
```

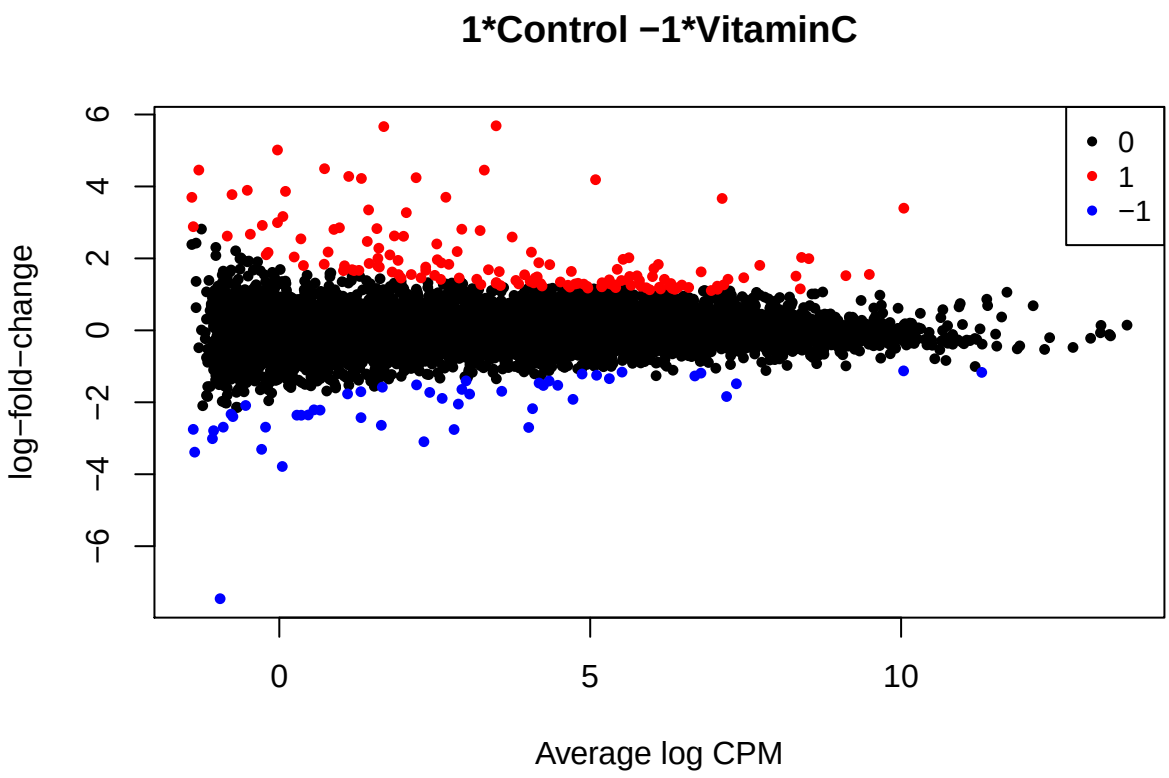
```
##        1*Control -1*VitaminC
## Down                      49
## NotSig                 13819
## Up                       144
```

The lfc is log-fold change in base 2, so an lfc = 1 means a two-fold difference in gene expression. After applying the cutoff for absolute log-fold change, 193 genes are DE between the control group and the Vitamin C-treated group.

MD plot - Control group vs Vitamin C-treated group

(FDR <= 0.05, absolute fold change >= 2)

```
plotMD(tr1, status=is.de4, values=c(0, 1,-1), col=c("black", "red","blue"),
       cex=0.75,legend="topright")
```



The vast majority of genes are now non-DE.

Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 2)

```
topTags(tr1)
```

```
## Coefficient:  1*Control -1*VitaminC
##    GeneID GeneSymbol                                                GeneName
## 3    1116     CHI3L1                                      chitinase 3 like 1
## 1   10397      NDRG1                          N-myc downstream regulated 1
## 2     768        CA9                                   carbonic anhydrase 9
## 14   7422      VEGFA                 vascular endothelial growth factor A
## 44   3485     IGFBP2          insulin like growth factor binding protein 2
## 6       2        A2M                                  alpha-2-macroglobulin
## 5   56901   NDUFA4L2        NDUFA4 mitochondrial complex associated like 2
## 8    5210     PFKFB4 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4
## 10   6533     SLC6A6                      solute carrier family 6 member 6
## 16   5033      P4HA1                  prolyl 4-hydroxylase subunit alpha 1
##       logFC unshrunk.logFC    logCPM       PValue          FDR
## 3  3.395006       3.395040 10.043626 4.627495e-28 6.484045e-24
## 1  3.665712       3.666019  7.121955 1.096182e-26 7.679854e-23
## 2  5.686109       5.701290  3.486755 9.864529e-22 4.607393e-18
## 14 2.026528       2.026574  8.399366 5.898473e-21 2.066235e-17
## 44 1.992508       1.992547  8.516434 5.243519e-19 1.469444e-15
## 6  1.806370       1.806426  7.727273 2.739969e-18 6.398740e-15
## 5  4.453948       4.461461  3.296732 5.127177e-18 1.026314e-14
## 8  2.018267       2.018550  5.622134 3.476388e-17 6.088894e-14
```

```
## 10 1.973336       1.973622  5.529005 7.985483e-17 1.230381e-13
## 16 1.831870       1.832058  6.095390 8.780908e-17 1.230381e-13
```

Some notable genes that are significantly upregulated in the Control group as compared to the Vitamin C-treated group are *CHI3L1, NDRG1, CA9, VEGFA, IGFBP2 and NDUFA4L2.*

CHI3L1: The chitinase 3-like-1 gene (CHi3L1) is involved in tissue remodelling and inflammatory response. It showed a 10.5-fold greater expression in the untreated control group as compared to the Vitamin C treated group. Higher expressions of this gene have been linked to shorter overall survival in glioblastoma patients (Steponaitis et al, 2016).

NDRG1: The N-myc downstream regulated 1 gene (NDRG1) showed almost a 13-fold increased expression in the control group. This gene plays important roles in specific stress responses, hormone responses, cell growth and differentiation. However, its role in cancer development is ambiguous. In some cancers, such as prostate and breast cancer, its overexpression has been shown to suppress tumor metastasis but in other cancers such as hepatocellular carcinoma, its upregulation is correlated with poor prognosis (Chua et al, 2007). Therefore, further studies can be conducted to investigate the function of this gene and its expression profile in different environments and tissues.

CA9: The carbonic anhydrase isoform 9 gene (CA9) showed a 50-fold increase in the control group. Overexpression of this gene has been linked to many cancers expecially glioblastoma (Said et al, 2010). CA9 increases the acidity of the tumour environment, resulting in resistance to weaky basic anticancer drugs. Agents that inhibit CA9 such as sulfonamide derivative compounds are being studied for glioblastoma therapy.

VEGFA: The vascular endothelial growth factor A (VEGFA) gene is involved in angiogenesis, vasculogenesis and endothelial cell growth. It showed a 4.5-fold increase in expression in the control group compared to the Vitamin C treated group. Its expression is positively correlated with tumor stage and progression (NCB1, 2015) (This gene is upregulated in many known tumors and its expression is correlated with tumor stage and progression. Induces endothelial cell proliferation, promotes cell migration, inhibits apoptosis and induces permeabilization of blood vessels.

IGFBP2: The insulin-like growth factor binding protein 2 (IGFBP2) gene showed a 4-fold increased expression in the control group. Overexpression of this gene has been shown to promote glioma progression and is linked to poor survival in gioblastoma (Liu et al, 2019). In mouse GBM cells, blocking IGFBP2 suppressed tumor growth and improved survival. Therefore, an IGFBP2-blocking agent may be an effective therapy for GBMs.

NDUFA4L2: The NDUFA4 mitochondrial complex associated like 2 (NDUFA4L2) gene showed a 22-fold increase in expression in the control group. In studies of human hepatocellular carcinoma, the overexpression of NDUF4AL2 was found to be linked to tumor microsatellite formation, absence of tumor encapsulation, and poor overall survival (Lai et al, 2016). The gene HIF is a regulator of NDUF4AL2. Inhibition of HIF by digoxin lead to inhibition of the NDUF4AL2 pathway, resulting in suppression of tumor growth. Therefore, HIF inhibitors (that lead to NDUF4AL2 inactivation) may be potential candidates for cancer therapy.

Control group vs 5-Azacitadine-treated group

(FDR <= 0.05, absolute fold change >= 2)

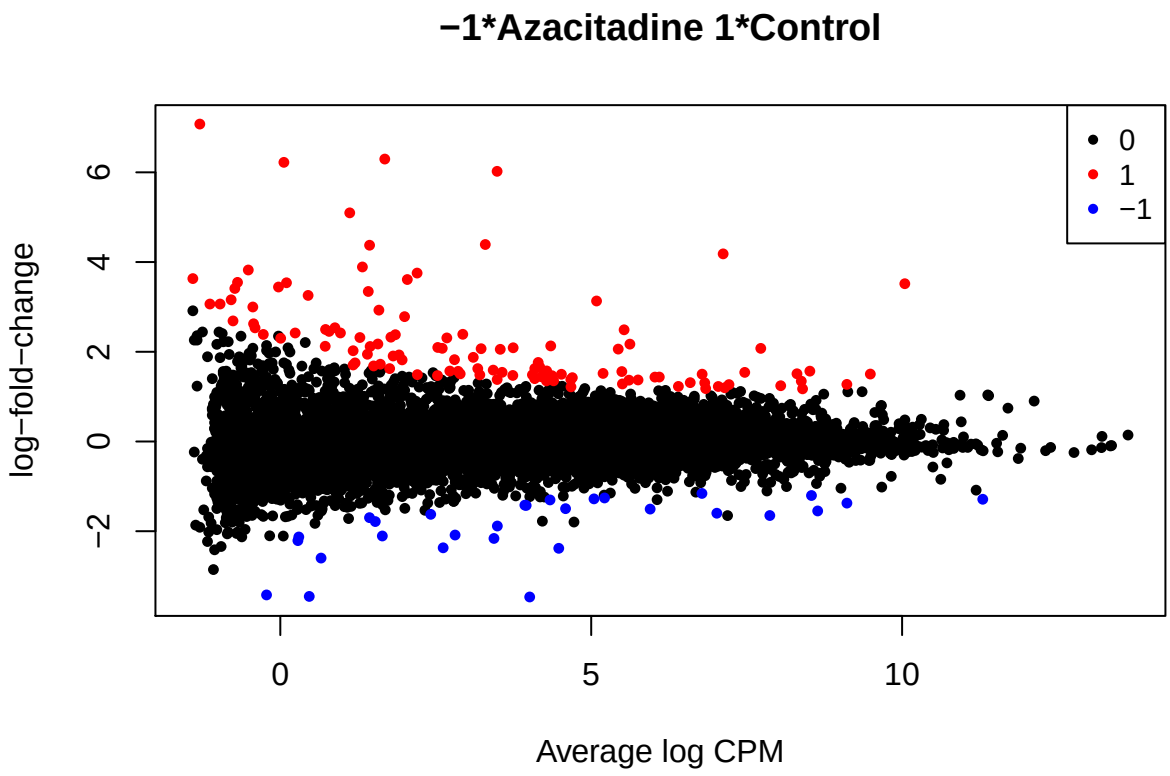
```
tr2 <- glmTreat(fit, contrast=caz, lfc=1)
is.de5 <- decideTestsDGE(tr2)
summary(is.de5)
```

```
##        -1*Azacitadine 1*Control
## Down                       29
## NotSig                  13868
## Up                        115
```

There are a total of 144 DE genes between the control group and the 5-Azacitadine-treated group with an absolute fold change of 2 or more.

MD plot - Control group vs 5-Azacitadine-treated group (FDR <= 0.05, absolute fold change >= 2)

```
plotMD(tr2, status=is.de5, values=c(0, 1,-1), col=c("black", "red","blue"),
       cex=0.75, legend="topright")
```



Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 2)

```
topTags(tr2)
```

```
## Coefficient:  -1*Azacitadine 1*Control
##    GeneID GeneSymbol                                                GeneName
## 3    1116     CHI3L1                                      chitinase 3 like 1
## 1   10397      NDRG1                           N-myc downstream regulated 1
## 2     768        CA9                                   carbonic anhydrase 9
## 6       2        A2M                                  alpha-2-macroglobulin
## 10   6533     SLC6A6                        solute carrier family 6 member 6
## 8    5210     PFKFB4  6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4
## 5   56901   NDUFA4L2         NDUFA4 mitochondrial complex associated like 2
## 31   2354       FOSB FosB proto-oncogene, AP-1 transcription factor subunit
## 4    7052       TGM2                                      transglutaminase 2
## 33   2353        FOS  Fos proto-oncogene, AP-1 transcription factor subunit
```

```
##        logFC unshrunk.logFC    logCPM       PValue          FDR
## 3   3.516983       3.517021 10.043626 6.617530e-27 9.272483e-23
## 1   4.182366       4.182817  7.121955 3.844548e-26 2.693490e-22
## 2   6.022555       6.041836  3.486755 3.573951e-20 1.669273e-16
## 6   2.075743       2.075815  7.727273 1.637744e-19 5.737017e-16
## 10  2.489768       2.490218  5.529005 1.000708e-18 2.804385e-15
## 8   2.170414       2.170739  5.622134 7.012490e-17 1.637650e-13
## 5   4.390959       4.398131  3.296732 2.465075e-16 4.934377e-13
## 31 -2.380959      -2.381843  4.478367 1.151727e-15 2.017251e-12
## 4   6.295814       6.380839  1.680235 8.513618e-15 1.325476e-11
## 33 -1.548018      -1.548044  8.642403 1.085103e-14 1.520447e-11
```

The genes that show the highest upregulation in the control group as compared to the 5-Azacitadine-treated group are CHI3L1, NDRG1, CA9 and SLC6A6, NDUFA4L2 and TGM2. Several of these genes were also overexpressed in the control group as compared to the Vitamin C group (section above), and their magnitudes of fold change are similar.

SLC6A6: The solute carrier family 6 member 6 (SLC6A6) gene encodes for a taurine transporter. Taurine is a chemical involved in osmoregulation, membrane stabilization, antioxidation and neurotransmission. Its expression was increased by 5.5-fold in the control group compared to the 5-Azacitadine-treated group. A study on colorectal cancer(CRC) cells found that SLC6A6 was highly expressed in CRC cells compared to normal colonocytes and increased the cell survival while enhancing multidrug resistance (Yasunaga et al, 2016). However, after knockdown of SLC6A6, cell survival was shortened and resistance to 5-fluorouracil (5-FU), doxycycline (DOX) and SN-38 was decreased.

TGM2: The transglutaminase 2 (TGM2) gene plays an important role in inflammation through its effects on the extracellular matrix. In the control group, expression of TGM2 was increased by almost 80-fold. In tumor progression, TGM2 promotes malignant cell mobility, invasion, and metastasis, and induces chemo-resistance of cancer cells (Huang et al, 2015).

Vitamin C treated group vs 5-Azacitadine-treated group (FDR <= 0.05, absolute fold change >= 1.5)

The difference between the Vitamin C treated group and 5-Azacitadine treated group was not as great as the differences between the Vitamin C and control, or 5-Azacitadine and Control groups. This was apparent from the multidimensional plot. Therefore, we can select a lower threshold for lfc (lfc=log2(1.5)) so that more DE genes are shown.

```
tr3 <- glmTreat(fit, contrast=ca, lfc=log2(1.5))
is.de6 <- decideTestsDGE(tr3)
summary(is.de6)
```

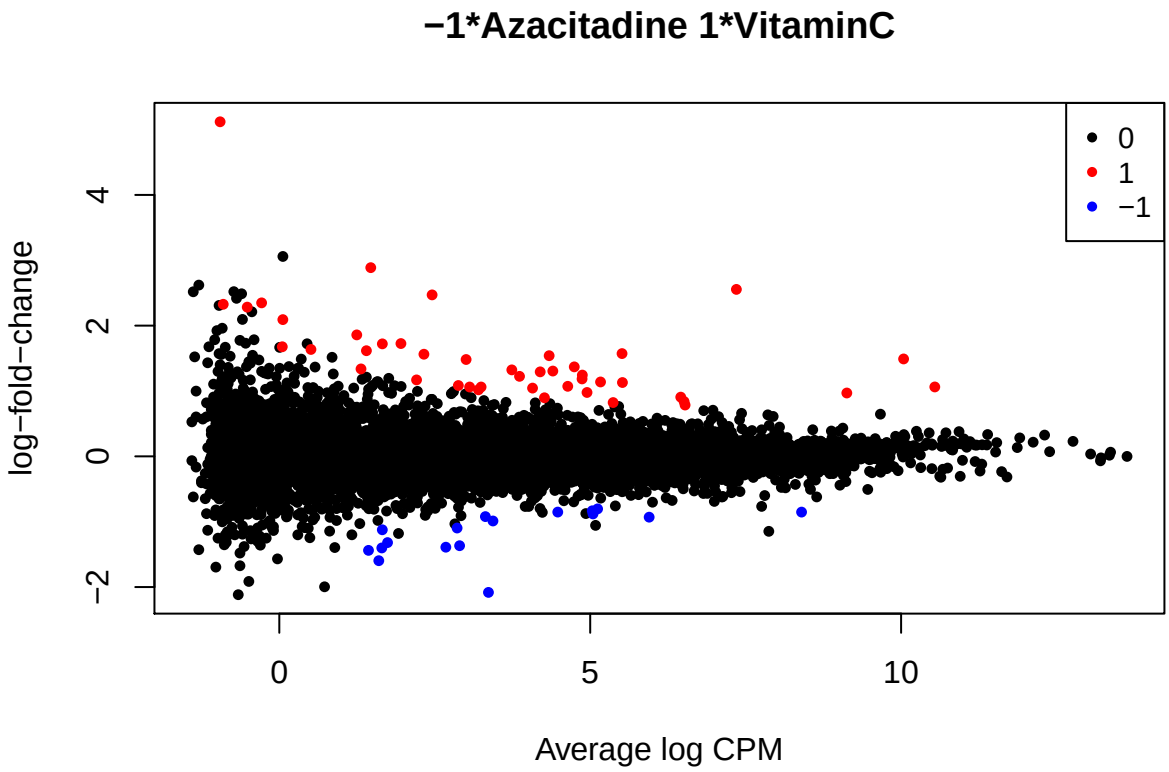
```
##        -1*Azacitadine 1*VitaminC
## Down                           17
## NotSig                      13951
## Up                             44
```

There are a total of 63 DE genes between the Vitamin C-treated and 5-Azacitadine-treated group with an absolute fold change >= 1.5

MD plot - Vitamin-C vs 5-Azacitadine-treated group

(FDR <= 0.05, absolute fold change >= 1.5)

```
plotMD(tr3, status=is.de6, values=c(0, 1,-1), col=c("black", "red","blue"),
       cex=0.75, legend="topright")
```



Top 10 Differentially-Expressed genes (FDR <= 0.05, absolute FC >= 1.5)

```
topTags(tr3)
```

```
## Coefficient:  -1*Azacitadine 1*VitaminC
##     GeneID GeneSymbol                                      GeneName      logFC
## 7     1728       NQO1             NAD(P)H quinone dehydrogenase 1  2.5546851
## 193   2512        FTL                         ferritin light chain  1.4902013
## 92  146802    SLC47A2           solute carrier family 47 member 2  1.5723634
## 60   23544      SEZ6L              seizure related 6 homolog like -2.0819292
## 142  51655      RASD1         ras related dexamethasone induced 1  1.5399799
## 429   7086        TKT                                 transketolase  1.0620019
## 138    218    ALDH3A1   aldehyde dehydrogenase 3 family member A1  2.4704973
## 126   3488     IGFBP5 insulin like growth factor binding protein 5  0.9690846
## 160   2052      EPHX1                           epoxide hydrolase 1  1.3050920
## 279   9890     PLPPR4          phospholipid phosphatase related 4  1.3689762
##     unshrunk.logFC    logCPM       PValue          FDR
## 7        2.5548302  7.350772 1.192619e-24 1.671098e-20
## 193      1.4902109 10.040229 2.323367e-17 1.627751e-13
```

```
## 92       1.5726032  5.511778 3.705290e-15 1.730617e-11
## 60      -2.0836467  3.363546 7.869294e-14 2.756614e-10
## 142      1.5404847  4.339798 9.261380e-13 2.595409e-09
## 429      1.0620065 10.542053 1.462916e-12 3.416395e-09
## 138      2.4751058  2.457228 2.859073e-12 5.723046e-09
## 126      0.9690984  9.127131 2.709493e-11 4.745677e-08
## 160      1.3055187  4.398832 5.547541e-11 8.588417e-08
## 279      1.3693499  4.743976 6.129330e-11 8.588417e-08
```

The differences in absolute fold change of gene expressions between the Vitamin C treated group and 5-Azacitadine treated group were not as large as with the control vs treated groups. Two genes showed a 5.5-fold increase in expression in Vitamin C compared to 5-Azacitidine group.

NQ01: The NAD(P)H quinone dehydrogenase 1 (NQO1) gene encodes for a plasma membrane redox enzyme. Overexpression of NQO1 results in cells with higher levels of oxygen consumption and ATP production, and less oxidative/nitrative damage and less apoptotic cell death (Kim et al, 2013). In neuroblastoma cells, elevated levels of NQO1 prevented apoptosis, suggesting that cell death may be enhanced by therapeutic agents that inhibit NQO1. Glioblastoma cells also show higher expression of NQO1.

ALDH3A1: The aldehyde dehydrogenase 3 family member A1 (ALDH3A1) enzyme converts 4-hydroxyonenal to fatty acids with NADH production. Gastric cancer cells showed increased fatty acid oxidation and increased ALDH3A1 expression. Inhibition of this enzyme reduced ATP production and lead to apoptosis of cancer cells (Lee et al, 2019).

Heat Maps

The heatmap.2() function clusters genes and samples based on Euclidean distance between the expression values. Since the rows of the logCPM matrix are pre-standardized, the Euclidean distance between each pair of genes is proportional to (1-r)^2, where r is the Pearson correlation coefficient between the genes. This means that genes that have positively correlated logCPM values (large r), will have smaller Euclidean distances and will cluster together on a heatmap. The selection of the genes affects the positioning of the samples. If we are displaying genes that are most DE between two groups, then those groups will be well separated on the plot. Therefore, samples are clustered together based on the similarity of their gene expression profiles. The replicate samples are clustered together, as expected.

Heatmap of top 30 differentially-expressed genes between Control group and Vitamin C treated group

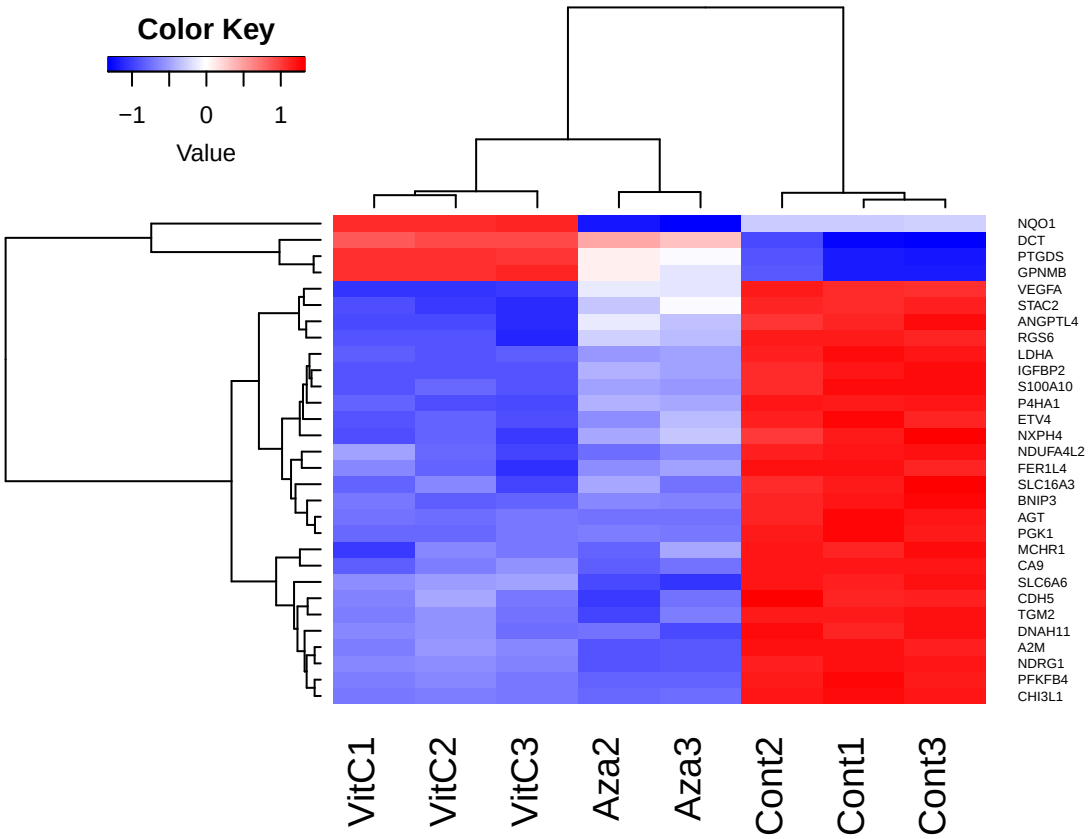
```
logCPM <- cpm(dlist3, prior.count=2, log=TRUE)
rownames(logCPM) <- dlist3$genes$GeneSymbol
colnames(logCPM) <- c("Cont1", "Cont2", "Cont3","VitC1","VitC2","VitC3","Aza2","Aza3")

ord1 <- order(tr1$table$PValue)
logCPM1 <- logCPM[ord1[1:30],]

logCPM1 <- t(scale(t(logCPM1)))

col.pan <- colorpanel(100, "blue", "white", "red")

heatmap.2(logCPM1, col=col.pan, scale="none",
          trace="none", dendrogram="both", Rowv=TRUE,
          cexRow=0.5, cexCol=1.4, density.info="none",
          margin=c(5,8), lhei = c(1,3))
```



The colour and intensity of the tiles represents changes (not absolute values) of gene expression. In this case, red represents up-regulated genes, white represents unchanged expression and blue indicates down-regulated genes. Almost all of the genes are up-regulated in the control group compared to the treated groups. This may be a result of DNA methylation, where methylation increases the transcription of certain genes. However, a few genes near the top of the heatmap (DCT, PTGDS, GPNMB) are down-regulated in the control group compared to the Vitamin C group.

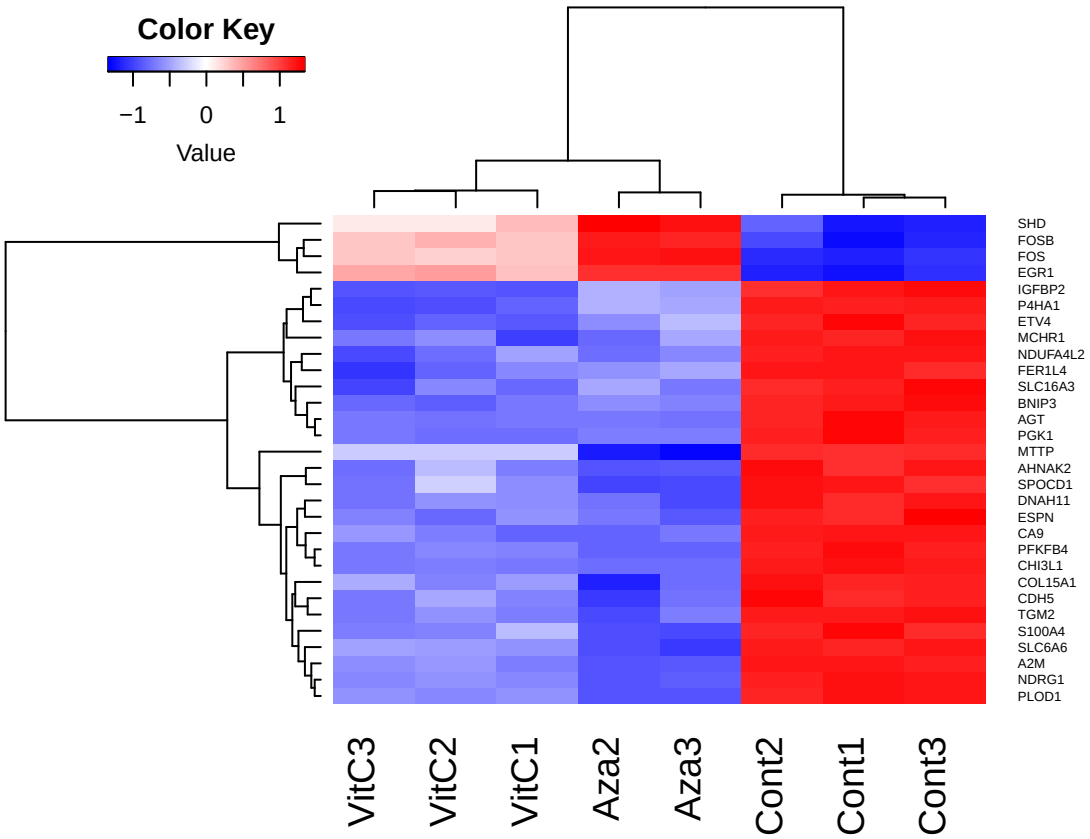
Heatmap of top 30 differentially-expressed genes between Control group and 5-Azacitadine treated group

```
ord2 <- order(tr2$table$PValue)
logCPM2 <- logCPM[ord2[1:30],]

logCPM2 <- t(scale(t(logCPM2)))

col.pan <- colorpanel(100, "blue", "white", "red")

heatmap.2(logCPM2, col=col.pan, scale="none",
          trace="none", dendrogram="both", Rowv=TRUE,
          cexRow=0.5, cexCol=1.4, density.info="none",
          margin=c(5,8), lhei = c(1,3))
```



This heatmap shows a similar pattern to the one above but there are 4 genes (at the top) that are down-regulated in the control group compared to the 5-Azacitadine group (SHD, FOSB, FOS, EGR1).

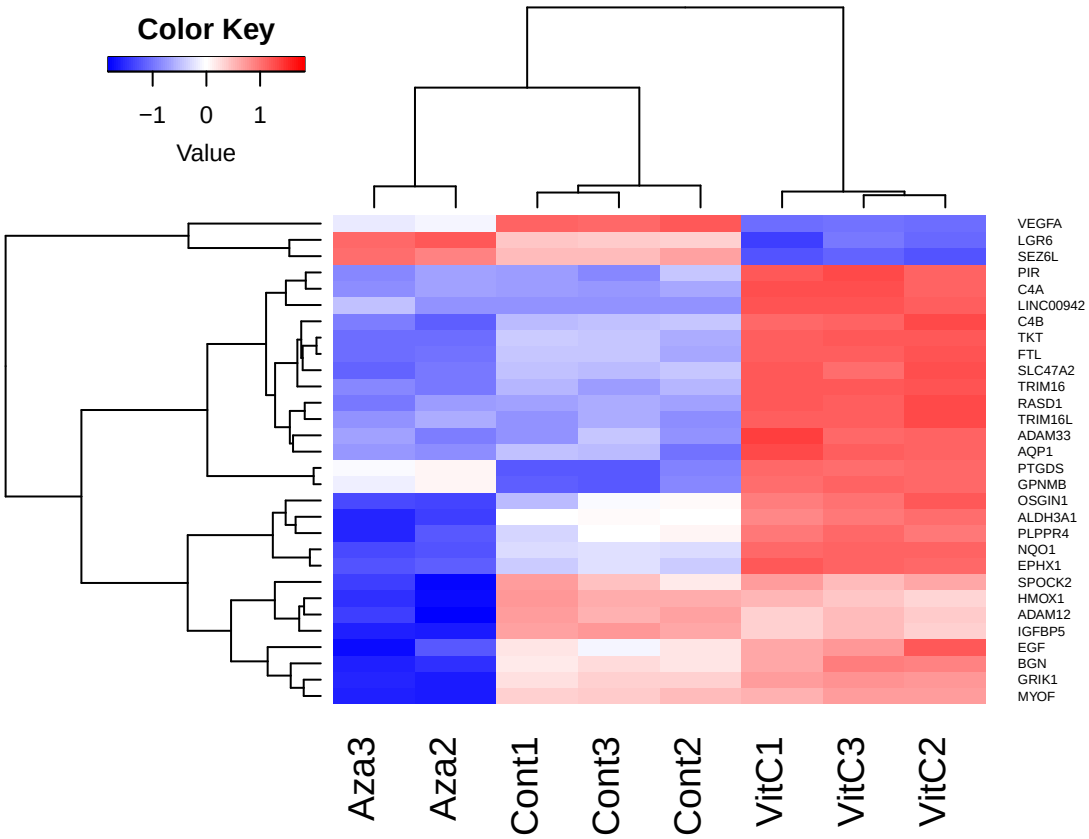
Heatmap of top 30 differentially-expressed genes between Vitamin C treated group and 5-Azacitadine treated group

```
ord3 <- order(tr3$table$PValue)
logCPM3 <- logCPM[ord3[1:30],]

logCPM3 <- t(scale(t(logCPM3)))

col.pan <- colorpanel(100, "blue", "white", "red")

heatmap.2(logCPM3, col=col.pan, scale="none",
          trace="none", dendrogram="both", Rowv=TRUE,
          cexRow=0.5, cexCol=1.4, density.info="none",
          margin=c(5,8), lhei = c(1,3))
```



The shading of the tiles indicate that the differences in the gene expressions are not that large between Vitamin C group and 5-Azacitadine group. There are very few genes that are up-regulated (red) in one treated group and down-regulated (blue) in the other treated group.

Gene Ontology and Pathway Analysis

We can look at the results of the differential expression analysis in terms of higher-order biological processes and molecular pathways. Using gene ontology (GO) databases and the Kyoto Encyclopedia of Genes and Genomes (KEGG), each DE gene is annotated with a biological process and pathway. The number of DE genes associated with each GO term and KEGG term is counted. Terms occurring more frequently are said to be over-represented or enriched.

We will only look at the KEGG molecular pathways.

Top 10 molecular pathways that are UP-REGULATED in Control group compared to Vitamin C treated group

```
keg <- kegga(tr1, species="Hs")
topKEGG(keg, n=10, sort="up")
```

```
##                                                  Pathway   N Up Down
## path:hsa01100                          Metabolic pathways 797 38    9
## path:hsa00071                     Fatty acid degradation  37 11    1
## path:hsa04022                cGMP-PKG signaling pathway 134 15    2
## path:hsa00280 Valine, leucine and isoleucine degradation  33  9    0
## path:hsa04261     Adrenergic signaling in cardiomyocytes 114 13    2
```

```
## path:hsa01212                      Fatty acid metabolism  32  8    1
## path:hsa04970                         Salivary secretion  77 10    2
## path:hsa04971                       Gastric acid secretion  59  9    1
## path:hsa04911                            Insulin secretion  65  9    1
## path:hsa05414                 Dilated cardiomyopathy (DCM)  70  9    2
##                         P.Up      P.Down
## path:hsa01100 6.032043e-16 0.001568713
## path:hsa00071 6.242119e-14 0.121713958
## path:hsa04022 6.930836e-12 0.079802529
## path:hsa00280 3.108225e-11 1.000000000
## path:hsa04261 1.472002e-10 0.060245969
## path:hsa01212 8.738597e-10 0.106156534
## path:hsa04970 5.851195e-09 0.029678253
## path:hsa04971 8.075576e-09 0.187073408
## path:hsa04911 1.949893e-08 0.204055706
## path:hsa05414 3.798014e-08 0.024875679
```

The category of "metabolic pathways" includes 38 genes that are upregulated in the control group compared to the Vitamin C group. These pathways include glycolysis, which is increased in cancer cells (Liberti et al, 2016). The pathway of fatty acid degradation includes fatty acid beta-oxidation, which is heavily used by cancer cells for proferation, survival, drug resistance and metastatic progression (Ma et al, 2018). In the control group, 11 genes from this pathway were seen to be upregulated. Branched-chain amino acids are preferentially uptaken by cancer cells. Therefore, "valine, leucine and isoleucine degradation" pathways are upregulated in cancer cells (Ananieva et al, 2017). In the control group, 9 genes are overexpressed from this pathway. The pathways of fatty acid metabolism include fatty acid synthesis and beta-oxidation. Fatty acid synthesis is greatly increased in cancer tissue (Chen et al, 2019). 8 genes from this pathway were upregulated in the control group compared to the Vitamin C treated group.

Top 10 molecular pathways that are UP-REGULATED in Control group compared to 5-Azacitadine treated group

```
keg <- kegga(tr2, species="Hs")
topKEGG(keg, n=10, sort="up")
```

```
##                                                 Pathway   N Up Down
## path:hsa00280 Valine, leucine and isoleucine degradation  33 10    1
## path:hsa01100                          Metabolic pathways 797 29   10
## path:hsa00071                       Fatty acid degradation  37  9    2
## path:hsa00640                        Propanoate metabolism  22  7    1
## path:hsa01212                        Fatty acid metabolism  32  7    2
## path:hsa00620                          Pyruvate metabolism  29  6    2
## path:hsa00380                        Tryptophan metabolism  30  6    1
## path:hsa02010                             ABC transporters  28  5    0
## path:hsa00650                       Butanoate metabolism  15  4    0
## path:hsa00630    Glyoxylate and dicarboxylate metabolism  18  4    0
##                         P.Up       P.Down
## path:hsa00280 7.356007e-14 6.615765e-02
## path:hsa01100 6.028713e-12 2.493865e-06
## path:hsa00071 1.262332e-11 2.633632e-03
## path:hsa00640 3.213189e-10 4.458914e-02
## path:hsa01212 5.926769e-09 1.973988e-03
## path:hsa00620 1.091427e-07 1.622030e-03
## path:hsa00380 1.355200e-07 6.032241e-02
## path:hsa02010 2.884067e-06 1.000000e+00
```

```
## path:hsa00650 5.481110e-06 1.000000e+00
## path:hsa00630 1.205570e-05 1.000000e+00
```

The main pathways that were up-regulated in the control group vs Vitamin C group, are also up-regulated in the control group vs 5-Azacitadine group. 10 genes are up-regulated in the control group from the "valine, leucine and isoleucine degradation" pathway. 29 genes are up-regulated from the "metabolic pathways" 9 genes are up-regulated from the "fatty acid degradation" pathway. 7 genes are up-regulated from the "fatty acid metabolism" pathway.

Top 10 molecular pathways that are UP-REGULATED in Vitamin C treated group compared to 5-Azacitadine treated group

```
keg <- kegga(tr3, species="Hs")
topKEGG(keg, n=10, sort="up")
```

```
##                                                                  Pathway   N Up
## path:hsa00350                                        Tyrosine metabolism  32  3
## path:hsa01100                                         Metabolic pathways 797 10
## path:hsa00360                                    Phenylalanine metabolism  16  2
## path:hsa04020                                   Calcium signaling pathway 147  4
## path:hsa00230                                           Purine metabolism  78  3
## path:hsa00410                                     beta-Alanine metabolism  22  2
## path:hsa04024                                      cAMP signaling pathway 175  4
## path:hsa00260                  Glycine, serine and threonine metabolism  24  2
## path:hsa04550 Signaling pathways regulating pluripotency of stem cells  96  3
## path:hsa05418                    Fluid shear stress and atherosclerosis 104  3
##               Down         P.Up      P.Down
## path:hsa00350    0 0.0001344634 1.000000000
## path:hsa01100    1 0.0001419492 0.630694963
## path:hsa00360    0 0.0011245662 1.000000000
## path:hsa04020    0 0.0011394413 1.000000000
## path:hsa00230    0 0.0018650061 1.000000000
## path:hsa00410    0 0.0021390443 1.000000000
## path:hsa04024    1 0.0021618379 0.192472250
## path:hsa00260    0 0.0025455773 1.000000000
## path:hsa04550    1 0.0033679151 0.110360920
## path:hsa05418    2 0.0042183675 0.006900313
```

Except for genes in "metabolic pathways", none of the other pathways exhibit much difference between the Vitamin C group and 5-Azacitadine group.

CONCLUSION

We investigated the genes that were differentially-expressed in three treatment groups of the human glioblastoma multiforme cancer cell-line (HSR-GMB1). The first group was control (untreated HSR-GBM), the second group was HSR-GBM cells treated with Vitamin C, and the third group was HSR-GBM cells treated with 5-Azacitadine. It has been speculated that global methylation of DNA is a factor in tumorigenesis. Therefore, the effects of demethylating agents (Vitamin C and 5-Azacitadine) on gene expression were studied.

The differential expression analysis showed that many genes that are involved in tumor progression and metastasis (eg. CHI3L1, CA9, VEGFA, NDUFA4L2, SLC6A6, and TGM2) were upregulated in the control (untreated) group as compared to either of the treatment groups. Therefore, the DNA demethylating agents, 5-Azacitadine and Vitamin C, may inhibit tumorigenesis by promoting differential of cancerous cells into

mature cells. Also, agents that are inhibitors of these upregulated genes may be potential candidates for cancer therapy.

The biochemical pathways (KEGG) that were up-regulated in the control group compared to both treatment groups included metabolic pathways (including glycolysis), fatty acid degradation pathway, fatty acid metabolism pathway (including beta-oxidation) and branched-chain amino acid (val, leu, iso-leu) degradation pathways. Activity of these pathways is increased in cancer cells. Therefore, the addition of Vitamin C or 5-Azacitadine resulted in a gene expression profile that was associated with decreased tumor growth.

Future investigations could include analysis of more samples from these treatment groups to corroborate these findings. Since glioblastoma cells show high heterogeneity, a larger number of samples would provide more reliable results. In fact, there is a dataset from the GEO omnibus reposity under series GSE98693 which includes raw RNA-sequence data from HSR-GBM cells under these same conditions. This dataset needs to be first converted into counts data. Combining the samples from both datasets would increase the number of replicates from 3 to 6 for each treatment group. This would also give a better estimate of dispersion of gene abundance.

Further studies can also elucidate the roles of genes that were down-regulated in the untreated control cells compared to either of the treatment groups. From looking at the heatmaps, these genes included DCT, PTGDS, GPNMB, SHD, FOSB, FOS, EGR1. Perhaps agents that enhance these genes may have therapeutic benefit in cancer treatment.

In an effort to discover novel agents for cancer therapy, other DNA demethylating agents can be studied to see their effects on overall gene expression profiles. From this analysis, we found that the effects of Vitamin C or 5-Azacitadine were similar but there were some differences. It is not clear which of the two was better in terms of tumor growth.

To increase knowledge of cancer prevention, gene expression profiles from other health conditions (eg. diabetes, autoimmune disorder, Crohn's disease, etc) can be examined to see how they compare to gene profiles from different cancer cells.

REFERENCES

American Association of Neurological Surgeons. Glioblastoma Multiforme. https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme. 2020.

Ananieva E and Wilkinson A. Branched-chain amino acid metabolism in cancer. *Curr Opin Clin Nutr Metab Care.* 2018 Jan; 21(1): 64–70.

Chua MS, Sun H, Cheung ST, Mason V, Higgins J, Ross DT, Fan ST and So S: Overexpression of NDRG1 is an indicator of poor prognosis in hepatocellular carcinoma. *Mod Pathol.* 20:76–83. 2007.

Ellen TP, Ke Q, Zhang P, Costa M. NDRG1, a growth and cancer related gene: regulation of gene expression and function in normal and disease states.*Carcinogenesis.* 2008 Jan;29(1):2-8.

Huang L, Xu A and Liu W. Transglutaminase 2 in cancer. *Am J Cancer Res.* 2015; 5(9): 2756–2776.

KEGG database. https://www.genome.jp/kegg/kegg2.html.

Kim J, Kim SK, Kim HK, et al. Mitochondrial Function in Human Neuroblastoma Cells Is Up-Regulated and Protected by NQO1, a Plasma Membrane Redox Enzyme. *PLoS One.* 2013; 8(7).

Lai RK, Xu IM, Chiu DK, et al.NDUFA4L2 Fine-tunes Oxidative Stress in Hepatocellular Carcinoma. *Clin Cancer Res.* 2016 Jun 15;22(12):3105-17.

Lee J, Kim S, Lee S, et al. Gastric cancer depends on aldehyde dehydrogenase 3A1 for fatty acid oxidation. November 2019. *Scientific Reports* 9(1):16313.

Liberti M, Locasale J. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci.* 2016 Mar; 41(3): 211–218.

Liu Y, Song C, Shen F, Zhang J, Song SW. IGFBP2 promotes immunosuppression associated with its mesenchymal induction and Fc RIIB phosphorylation in glioblastoma.*PLoS One*. 2019 Sep 27;14(9).

NCBI. VEGFA Gene. https://ghr.nlm.nih.gov/gene/VEGFA. Nov 2015.

Said HM1, Supuran CT, Hageman C, Staab A, Polat B, Katzer A, Scozzafava A, Anacker J, Flentje M, Vordermark D. Modulation of carbonic anhydrase 9 (CA9) in human brain cancer. *Curr Pharm Des.* 2010;16(29):3288-99.

Steponaitis G, Skiriutė D, Kazlauskas A, Golubickaitė I, Stakaitis R, Tamašauskas T, and Vaitkienė P. High CHI3L1 expression is associated with glioma patient survival. *Diagn Pathol.* 2016; 11: 42.

Sun X, Johnson J, and St.John J. Global DNA methylation synergistically regulates the nuclear and mitochondrial genomes in glioblastoma cells.
Nucleic Acids Res. 2018 Jul 6; 46(12): 5977–5995.

Yasunaga M and Matsumura Y. Role of SLC6A6 in promoting the survival and multidrug resistance of colorectal cancer.*Sci Rep.* 2014; 4: 4852.

edgeR:

Chen Y, Lun A and Smyth G. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR (Ch3).*Statistical Analysis of Next Generation Sequencing Data.* June 2014.

Chen Y, Lun A and Smyth G. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. https://f1000research.com/articles/5-1438/v2. 2016.

CHen Y, McCarthy D, et al. edgeR: differential expression analysis of digital gene expression data. User's Guide. https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf. 21 October 2019.