

# Microarray1, 2, RNA seq datasets combined

Seoyeon Oh, Tobias Heyman

12/20/2021

```
library("mogene10sttranscriptcluster.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 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.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 objects are masked from 'package:base':
##       expand.grid, I, uname

## Loading required package: org.Mm.eg.db

## 

## 

library("ArrayExpress")
library("arrayQualityMetrics")
library("ggplot2")
library("huex10sttranscriptcluster.db")

## Loading required package: org.Hs.eg.db

## 

## 

library("limma")

## 
## Attaching package: 'limma'

## The following object is masked from 'package:BiocGenerics':
##       plotMA

library("oligo")

## Loading required package: oligoClasses

## Welcome to oligoClasses version 1.54.0

## Loading required package: Biostrings

## Loading required package: XVector

## Loading required package: GenomeInfoDb

```

```

## 
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##       strsplit

## =====

## Welcome to oligo version 1.56.0

## =====

## 
## Attaching package: 'oligo'

## The following object is masked from 'package:limma':
##       backgroundCorrect

library("siggenes")

## Loading required package: multtest

## Loading required package: splines

library("affy")

## 
## Attaching package: 'affy'

## The following objects are masked from 'package:oligo':
##       intensity, MApplot, mm, mm<-, mmindex, pm, pm<-, pmindex,
##       probeNames, rma

## The following object is masked from 'package:oligoClasses':
##       list.celfiles

#library("pd.huex.1.0.st.v2")
library("wateRmelon")

## Loading required package: matrixStats

## 
## Attaching package: 'matrixStats'

```

```

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## Loading required package: methylumi

## Loading required package: scales

## Loading required package: reshape2

## Loading required package: FDb.InfiniumMethylation.hg19

## Loading required package: GenomicFeatures

## Loading required package: GenomicRanges

## Loading required package: TxDb.Hsapiens.UCSC.hg19.knownGene

## Loading required package: minfi

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnyNs, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnyNs, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## The following object is masked from 'package:Biobase':
##
##     rowMedians

## Loading required package: bumphunter

```

```

## Loading required package: foreach

## Parallel computing support for 'oligo/crlmm': Disabled
##   - Load 'ff'
##   - Load and register a 'foreach' adaptor
##     Example - Using 'multicore' for 2 cores:
##       library(doMC)
##       registerDoMC(2)
## =====

## Loading required package: iterators
## Loading required package: locfit
## locfit 1.5-9.4    2020-03-24
## Setting options('download.file.method.GEOquery'='auto')
## Setting options('GEOquery.inmemory.gpl'=FALSE)
##
## Attaching package: 'minfi'
##
## The following object is masked from 'package:oligo':
##
##      getProbeInfo
##
## The following object is masked from 'package:oligoClasses':
##
##      getM
##
## Loading required package: lumi
## No methods found in package 'RSQLite' for request: 'dbListFields' when loading 'lumi'
##
## Attaching package: 'lumi'
##
## The following objects are masked from 'package:methylumi':
##
##      estimateM, getHistory
##
## The following objects are masked from 'package:affy':
##
##      MAplot, plotDensity
##
## The following object is masked from 'package:oligo':
##
##      MAplot
##
## Loading required package: ROC
## Loading required package: IlluminaHumanMethylation450kanno.ilmn12.hg19
## Loading required package: illuminaio

library("affy")
library("arrayQualityMetrics")
library("ArrayExpress")
library("RSQLite")
library("DBI")
library("htmltools")
library("biomaRt")
library("tximport")

```

```

library("edgeR")
library("rhd5")

setwd("/Users/seoyeon/Desktop/MSc Bioinformatics Year 1/Applied High-throughput Analysis/Project/Datasets")

#E-GEOD-57452

```

## General info

The array used for this dataset is A-AFFY-130 - Affymetrix GeneChip Mouse Gene 1.0 ST Array [MoGene-1\_0-st-v1]. Mice were infected with influenza and RNA was extracted from the lungs after 10 days. We used samples involving susceptible mice after 10 days of infection with influenza from this dataset.

## Intensity values

Read in the microarray data and display the head and dimensions of the intensity value matrix.

```

id_1 <- "E-GEOD-57452"
exonCEls <- list.celfiles("../Datasets/Microarray1/")
data.raw_1 <- read.celfiles(paste(rep("../Datasets/Microarray1/", length(exonCEls)), exonCEls, sep=""))

## Reading in : ../Datasets/Microarray1/GSM1382971_lung_T-9_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382972_lung_T-9_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382973_lung_T-9_rep3.CEL
## Reading in : ../Datasets/Microarray1/GSM1382974_lung_T0_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382975_lung_T1_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382976_lung_T2_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382977_lung_T3_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382978_lung_T3_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382979_lung_T4_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382980_lung_T4_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382981_lung_T5_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382982_lung_T5_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382983_lung_T5_rep3.CEL
## Reading in : ../Datasets/Microarray1/GSM1382984_lung_T6_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382985_lung_T7_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382986_lung_T7_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382987_lung_T8_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382988_lung_T8_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382989_lung_T9_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382990_lung_T9_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382991_lung_T10_rep1.CEL
## Reading in : ../Datasets/Microarray1/GSM1382992_lung_T10_rep2.CEL
## Reading in : ../Datasets/Microarray1/GSM1382993_lung_T10_rep3.CEL

# make vector containing the sample class
samples <- c(replicate(3, "control"), "day0", "day1", "day2", replicate(2, "day3"), replicate(2, "day4"))

# add the sample classes to the pData object
pData(data.raw_1)[,2] <- samples

```

```

colnames(pData(data.raw_1)) <- c("index", "treatment")

# filter control samples and samples taken after 3 days of infection
filter <- colnames(data.raw_1)[data.raw_1@phenoData@data$treatment=="control" | data.raw_1@phenoData@da]

# apply filter
filtered <- data.raw_1[,filter]

# check dimentions of filtered object
dim(exprs(filtered))

## [1] 1102500      6

## arrayQualityMetrics
#arrayQualityMetrics(filtered,outdir="..../Datasets/microarray1/raw1",force=T)
#arrayQualityMetrics(filtered,outdir="..../Datasets/microarray1/rawlog1",force=T,do.logtransform=T)

# Preprocessing (using the oligo function because affy didnt work)
MouseRMA<- oligo::rma(filtered,background=T)

## Background correcting
## Normalizing
## Calculating Expression

## QC post preprocessing
#arrayQualityMetrics(MouseRMA,outdir="..../Datasets/microarray1/rma1",force=T)          #RMA produces l

##Data exploration

# transpose the data before Pca as this function requires the variables to b columns
data <- t(as.data.frame(MouseRMA@assayData$exprs))
pca <- prcomp(data, center = T, scale. = T)

summary(pca)

## Importance of components:
##           PC1       PC2       PC3       PC4       PC5       PC6
## Standard deviation 117.2632 86.7380 83.3066 63.5689 57.45316 6.844e-13
## Proportion of Variance 0.3867 0.2116 0.1952 0.1137 0.09284 0.000e+00
## Cumulative Proportion 0.3867 0.5983 0.7935 0.9072 1.00000 1.000e+00

# save as dataframe and add treatment variable
pca_out <- as.data.frame(pca$x)
pca_out$treatment <- as.character(MouseRMA@phenoData@data$treatment)

# get labels
percentage <- round(pca$sdev / sum(pca$sdev) * 100, 2)
percentage <- paste( colnames(pca_out), "(", paste( as.character(percentage), "%", ")"), sep="") )

ggplot(data = pca_out)+
```

```

geom_point(aes(x = PC1, y = PC2, colour = treatment, label=''), size=3)+  

  geom_text(aes(x = PC1, y = PC2, colour = treatment, label=''), hjust=0.5, vjust=1.15)+  

  theme_bw() +  

  xlab(percentage[1]) +  

  ylab(percentage[2]) +  

  labs(colour = "treatment") +  

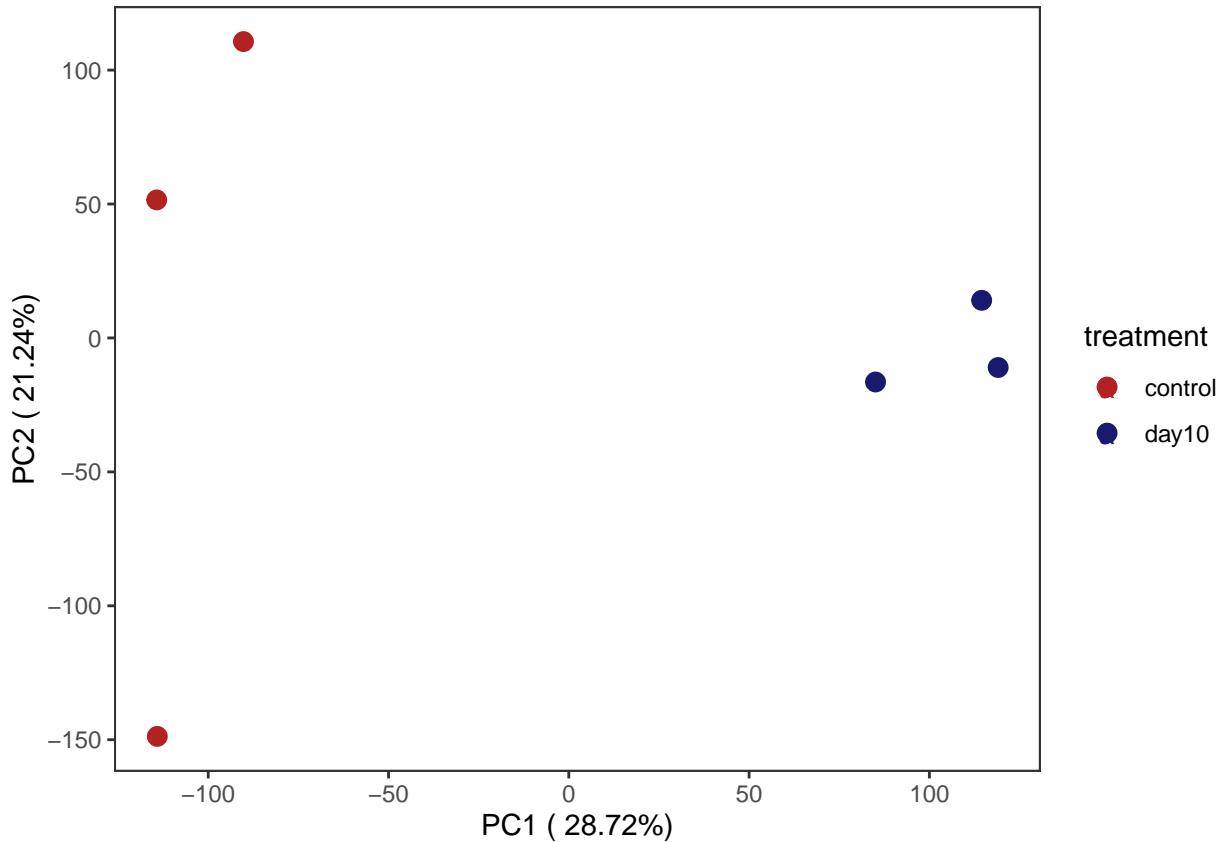
  theme(plot.title = element_text(hjust = 0.5)) +  

  scale_colour_manual(values = c("firebrick", "midnightblue")) +  

  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

```

## Warning: Ignoring unknown aesthetics: label



```
ggsave("PCA_array1.png", dpi=750, width=8, height = 5)
```

determine differential expression

```

annot <- factor(pData(MouseRMA)[,2])

## Differential expression by LIMMA
# Method as stated in limma package (no intercept, easy for simple model designs)
design <- model.matrix(~0+annot)
colnames(design)<-c("control","infected")

# make linear model

```

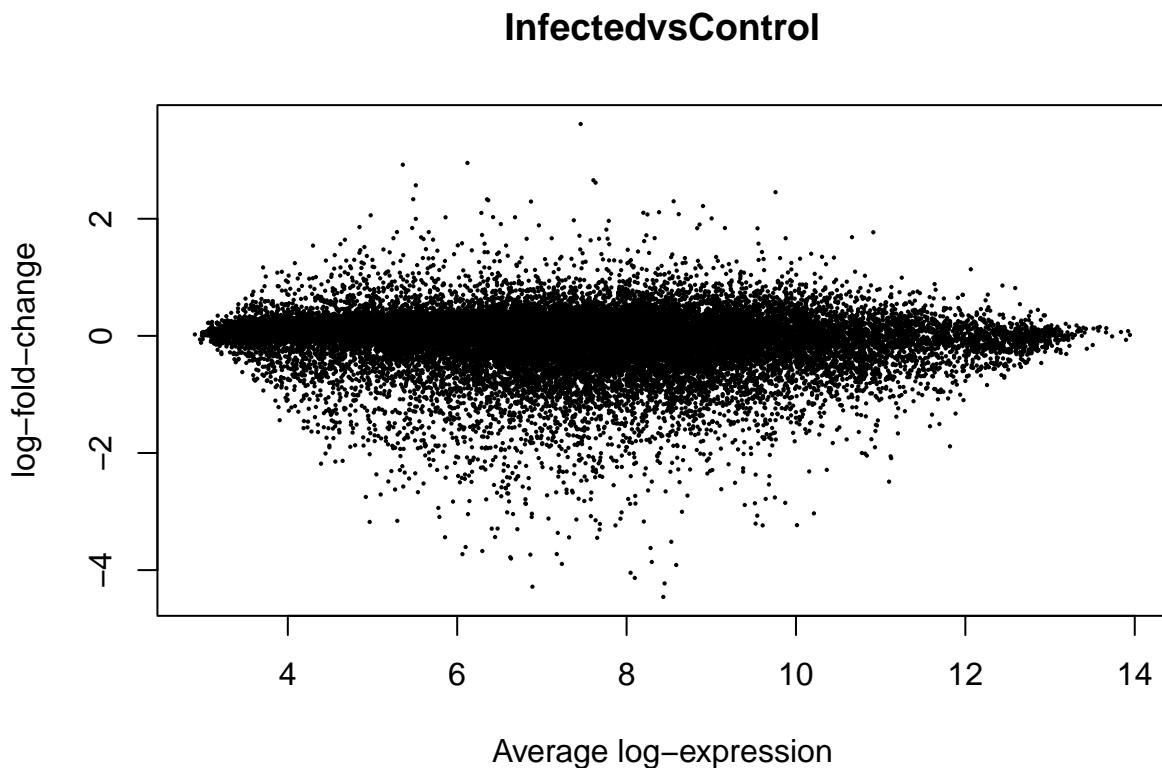
```

fit <- lmFit(MouseRMA,design)

# create contrast matrix to get the differential expression between samples from infected mice and uninfect
cont.matrix <- makeContrasts(InfectedvsControl=control-infected,levels=design)
fit2 <- contrasts.fit(fit,cont.matrix)
fit2 <- eBayes(fit2)

# make MA plot for model with applied contrast matrix
limma:::plotMA(fit2)

```



```

library(ggplot2)
# DE results with multiple testing correction (Benjamini-Hochberg = BH)
LIMMAout <- topTable(fit2,adjust="BH",number=nrow(exprs(MouseRMA)))

# add column indicating for all differentially expressed genes (adjusted p-value < 0.05) whether they're up or down
LIMMAout$diffexpressed <- "NO"
LIMMAout$diffexpressed[LIMMAout$logFC > 0 & LIMMAout$adj.P.Val < 0.05] <- "UP"
LIMMAout$diffexpressed[LIMMAout$logFC < 0 & LIMMAout$adj.P.Val < 0.05] <- "DOWN"

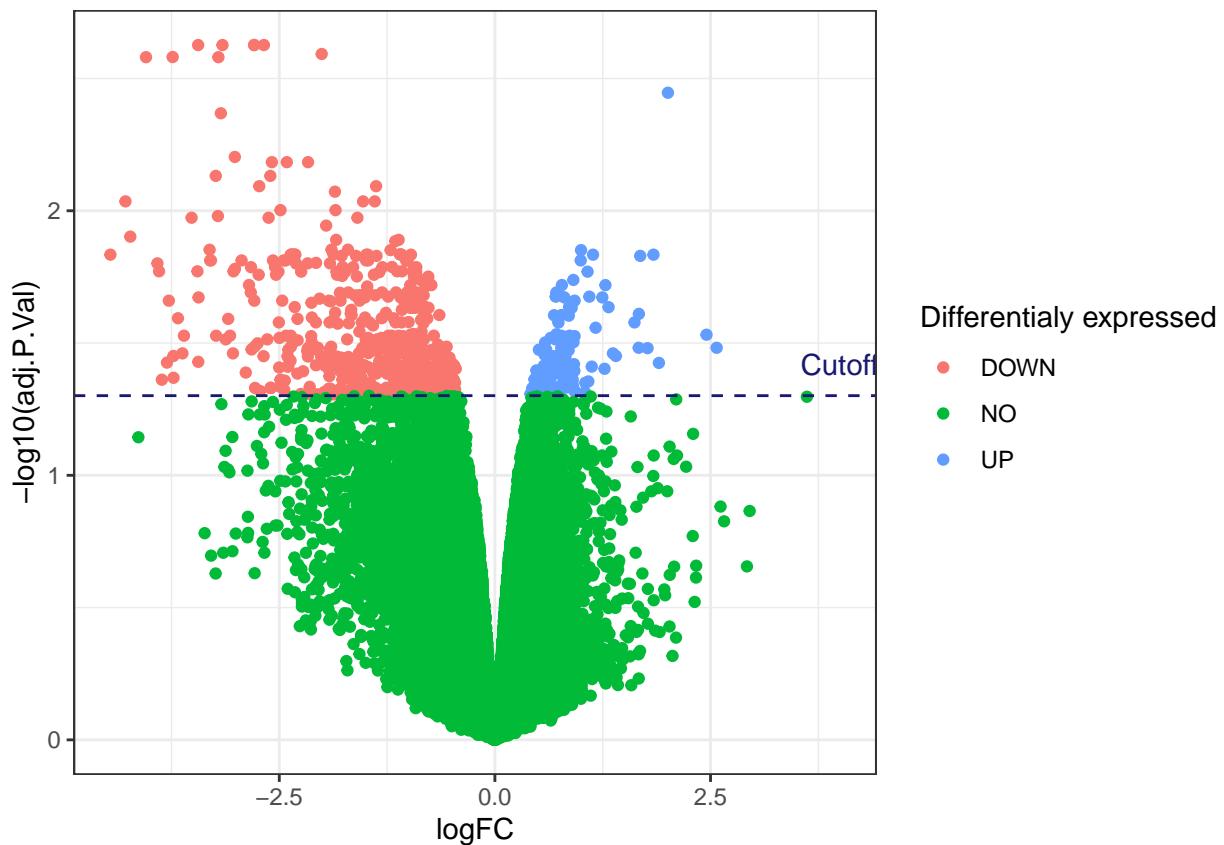
# do the same but if we would not correct for multiple testing.
LIMMAout$diffexpressed_no_BH <- "NO"
LIMMAout$diffexpressed_no_BH[LIMMAout$logFC > 0 & LIMMAout$P.Value < 0.05] <- "UP"

```

```
LIMMAout$diffexpressed_no_BH[LIMMAout$logFC < 0 & LIMMAout$P.Value < 0.05] <- "DOWN"

# code to make volcano plots

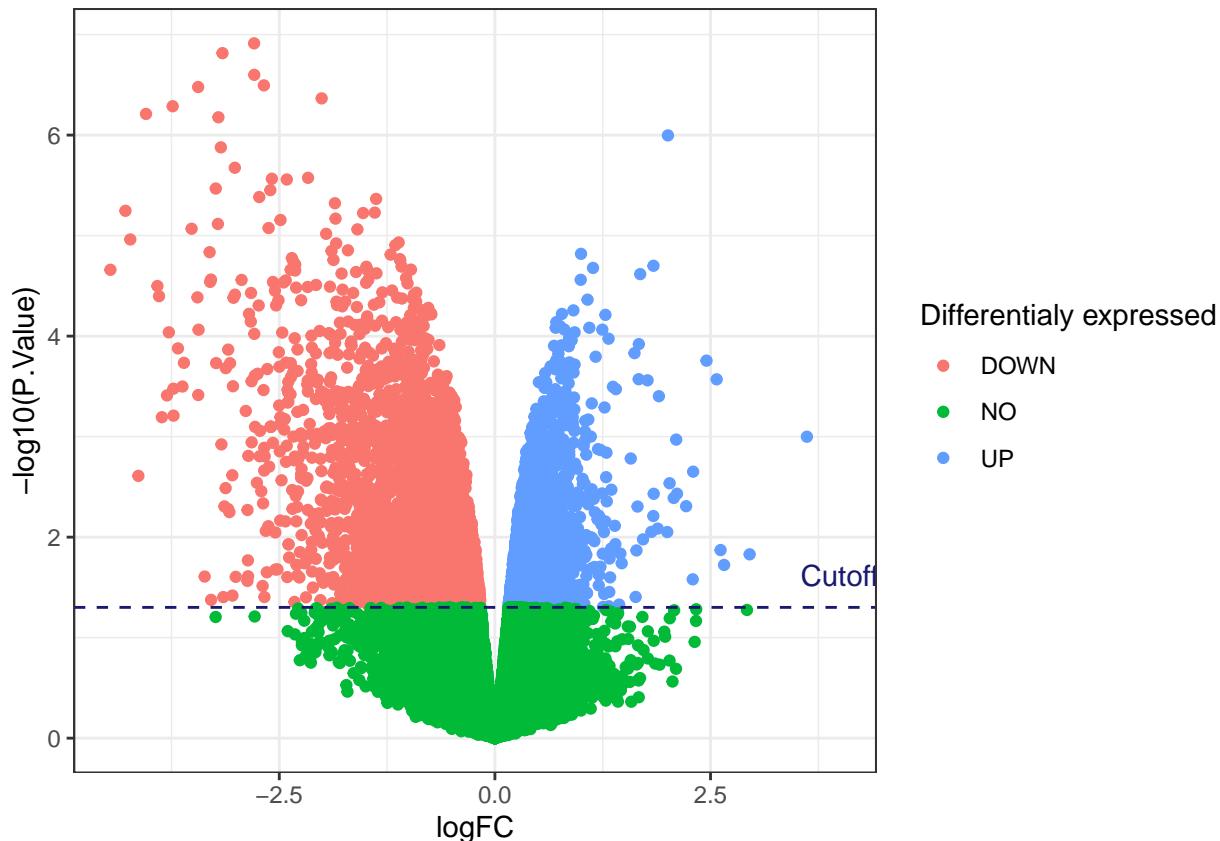
ggplot(data = LIMMAout, aes(x= logFC, y = -log10(adj.P.Val), colour = diffexpressed)) +
  geom_point()+
  theme_bw()+
  geom_hline(yintercept = -log10(0.05), linetype="dashed", color="midnightblue")+
  annotate("text", min(4), 1.3, vjust = -1, label = "Cutoff", color="midnightblue")+
  labs(colour = "Differentially expressed")
```



```
ggsave("volcanoplot.png", dpi=750)
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(data = LIMMAout, aes(x= logFC, y = -log10(P.Value), colour = diffexpressed_no_BH)) +
  geom_point()+
  theme_bw()+
  geom_hline(yintercept = -log10(0.05), linetype="dashed", color="midnightblue")+
  annotate("text", min(4), 1.3, vjust = -1, label = "Cutoff", color="midnightblue")+
  labs(colour = "Differentially expressed")
```



```
table(LIMMAout$diffexpressed)
```

```
##
##    DOWN      NO      UP
##    570  34863   123
```

### Annotation

```
# get the annotation through package (mogene10sttranscriptcluster.db) found at https://www.biostars.org/
columns(mogene10sttranscriptcluster.db)
```

```
## [1] "ACNUM"          "ALIAS"           "ENSEMBL"          "ENSEMLPROT"       "ENSEMLTRANS"
## [6] "ENTREZID"        "ENZYME"          "EVIDENCE"         "EVIDENCEALL"     "GENENAME"
## [11] "GENETYPE"        "GO"               "GOALL"            "IPI"              "MGI"
## [16] "ONTOLOGY"        "ONTOLOGYALL"     "PATH"             "PFAM"             "PMID"
## [21] "PROBEID"         "PROSITE"          "REFSEQ"           "SYMBOL"           "UNIPROT"
```

```
annotTable <- select(
  mogene10sttranscriptcluster.db,
  keys = keys(mogene10sttranscriptcluster.db),
  column = c('PROBEID', 'SYMBOL', 'ENTREZID', 'ENSEMBL', 'GENENAME', 'PROSITE'),
  keytype = 'PROBEID')
```

```

## 'select()' returned 1:many mapping between keys and columns

## sort annotation data alphabetically on probe name

annotTable.filt <- annotTable[sort(annotTable$PROBEID, index.return=T)$ix,]

# merge information from multiple lines describing the same probe
probe <- "start"
position <- 0
for (i in 1:dim(annotTable.filt)[1]){
  if (annotTable.filt[i, 1] != probe){
    probe <- annotTable.filt[i, 1]
    position <- i
  }
  else{
    # concatenate the information of the 2 lines with a ; as separator
    annotTable.filt[position,2:5] <- paste(annotTable.filt[position,2:5], annotTable.filt[i, 2:5], sep="")
    # mark the line
    annotTable.filt[i,1] <- NA
  }
}

annotTable.filt <- annotTable.filt[!is.na(annotTable.filt$PROBEID),]

## Check if all probes are present in both sets
dim(annotTable.filt)

## [1] 35556      6

dim(LIMMAout)

## [1] 35556      8

## Double check => "Assumption is the mother of all fuck up's ;)"
sum(annotTable.filt$PROBEID!=sort(rownames(LIMMAout)))

## [1] 0

## Sort LIMMA output alphabetically on probe name
LIMMAout_sorted <- LIMMAout[sort(rownames(LIMMAout), index.return=T)$ix,]

## Add gene names to LIMMA output
LIMMAout_sorted$gene <- annotTable.filt$SYMBOL
LIMMAout_annot <- LIMMAout_sorted[sort(LIMMAout_sorted$adj.P.Val, index.return=T)$ix,]

# determine how many differentially expressed probes have an annotated gene
table(is.na(LIMMAout_annot[LIMMAout_annot$diffexpressed != "NO",9]))


## 
## FALSE  TRUE
##   510   183

```

```

## alternative annotation method:

# annotation file from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6246
annotation_MA <- read.delim("./Microarray1/GPL6246.annot", header=T, sep="\t", skip=27, fill=T)
print(head(annotation_MA))

##           ID      Gene.title Gene.symbol   Gene.ID UniGene.title
## 1 10344614 predicted gene    2889       Gm2889 100040658
## 2 10344616
## 3 10344618
## 4 10344620 predicted gene    10568      Gm10568 100038431
## 5 10344622
## 6 10344624 lysophospholipase 1      Lypla1     18777
##   UniGene.symbol UniGene.ID
## 1
## 2
## 3
## 4
## 5
## 6
## 
## 1
## 2
## 3
## 4
## 5
## 6 Mus musculus lysophospholipase 1 (Lypla1), mRNA///PREDICTED: Mus musculus lysophospholipase 1 (Lyp
## 
## 1
## 2
## 3
## 4
## 5
## 6 31543132///1039727931///26341311///74191027///15488807///31127306///71059730///1864158///12854598///
## 
## 1
## 2
## 3
## 4
## 5
## 6 NM_008866///XM_006495471///AK050549///AK167231///BC013536///BC052848///CT010201///U89352///AK01602
## Platform_CLONEID Platform_ORF      Platform_SPOTID Chromosome.location
## 1          NA          NA chr1:3054233-3054733          18
## 2          NA          NA chr1:3102016-3102125
## 3          NA          NA chr1:3276323-3277348
## 4          NA          NA chr1:3680571-3680912
## 5          NA          NA chr1:4771131-4772199
## 6          NA          NA chr1:4807862-4846736          1 A1
## 
## Chromosome.annotation
## 1
## 2
## 3
## 4          Chromosome 1

```

```

## 5
## 6 Chromosome 1, NC_000067.6 (4807560..4846739)
##
## 1
## 2
## 3
## 4
## 5
## 6 2,4,4-trimethyl-3-oxopentanoyl-CoA thioesterase activity///2-oxoglutaryl-CoA thioesterase activity
##
## 1
## 2
## 3
## 4
## 5
## 6 fatty acid metabolic process///lipid metabolic process///negative regulation of Golgi to plasma mem
## GO.Component
##
## 1
## 2
## 3
## 4
## 5
## 6 cytoplasm///cytosol///extracellular exosome///mitochondrion
##
## 1
## 2
## 3
## 4
## 5
## 6 GO:0034869///GO:0034843///GO:0034946///GO:0016289///GO:0047617///GO:0052689///GO:0044466///GO:0016
## GO.Process.ID
##
## 1
## 2
## 3
## 4
## 5
## 6 GO:0006631///GO:0006629///GO:0042997///GO:0002084
## GO.Component.ID
##
## 1
## 2
## 3
## 4
## 5
## 6 GO:0005737///GO:0005829///GO:0070062///GO:0005739

annotation_MA <- annotation_MA[sort(annotation_MA$ID, index.return=T)$ix,]

dim(annotation_MA)

## [1] 35558      21

dim(LIMMAout)

## [1] 35556      8

```

```

## the dimensions don't match but we can still check whether the information for the differentially expressed genes is correct
#annotation_MA[annotation_MA$ID %in% rownames(LIMMAout_annot[LIMMAout_annot$diffexpressed != "NO",]),]

#get DE genes symbols
DEgenes_symbols1 <- unique(annotation_MA[annotation_MA$ID %in% rownames(LIMMAout_annot[LIMMAout_annot$diffexpressed != "NO",]) & annotation_MA$Gene.symbol != "",]$Gene.symbol)
head(DEgenes_symbols1)

## [1] "Cspp1"   "Cetn4"    "Il1r2"    "Il18r1"   "Ormdl1"   "Nif3l1"

```

## Microarray2: E-GEO-D-64750

### General info

The array used for this dataset is A-AFFY-45 - Affymetrix GeneChip Mouse Genome 430 2.0 [Mouse430\_2]. In this experiment, susceptible mice were infected with H5N1 influenza. After 72h RNA was extracted from the lungs of the mice. We used 9 samples of this experiment (susceptible mice).

### Intensity values

Read in the microarray data and examine dimensionality of the intensity value matrix.

```

id <- "E-GEO-D-64750"
exonCEls <- list.celfiles("../Datasets/Microarray2/")
data.raw_2 <- read.celfiles(paste(rep("../Datasets/Microarray2/", length(exonCEls)), exonCEls, sep=""))

## Reading in : ../Datasets/Microarray2/GSM1579245_jac013-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579246_jac021-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579247_jac044-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579248_jac045-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579249_jac046-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579250_jac007-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579251_jac008-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579252_jac036-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579253_jac047-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579254_jac014-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579255_jac024-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579256_jac025-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579257_jac011-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579258_jac012-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579259_jac037-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579260_jac048-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579261_jac015-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579262_jac016-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579263_jac017-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579264_jac018-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579265_jac019-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579266_jac020-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579267_jac027-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579268_jac028-430v2.CEL

```

```

## Reading in : ../Datasets/Microarray2/GSM1579269_jac034-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579270_jac033-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579271_jac029-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579272_jac035-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579273_jac041-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579274_jac042-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579275_jac043-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579276_jac030-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579277_jac031-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579278_jac032-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579279_jac038-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579280_jac039-430v2.CEL
## Reading in : ../Datasets/Microarray2/GSM1579281_jac040-430v2.CEL

```

```
dim(exprs(data.raw_2))
```

```
## [1] 1004004      37
```

```
head(exprs(data.raw_2))
```

	GSM1579245_jac013-430v2.CEL	GSM1579246_jac021-430v2.CEL
## 1	62	191
## 2	5707	16272
## 3	69	265
## 4	5913	16107
## 5	57	143
## 6	54	143
## GSM1579247_jac044-430v2.CEL	GSM1579248_jac045-430v2.CEL	
## 1	198	134
## 2	7485	5697
## 3	200	144
## 4	7664	6184
## 5	154	121
## 6	162	128
## GSM1579249_jac046-430v2.CEL	GSM1579250_jac007-430v2.CEL	
## 1	136	123
## 2	6036	7786
## 3	134	162
## 4	6147	7832
## 5	126	111
## 6	113	105
## GSM1579251_jac008-430v2.CEL	GSM1579252_jac036-430v2.CEL	
## 1	139	80
## 2	8743	6310
## 3	154	102
## 4	8940	6335
## 5	102	70
## 6	96	88
## GSM1579253_jac047-430v2.CEL	GSM1579254_jac014-430v2.CEL	
## 1	92	105
## 2	4988	8341
## 3	87	159
## 4	5275	8721

## 5	87	109
## 6	79	96
## GSM1579255_jac024-430v2.CEL	GSM1579256_jac025-430v2.CEL	
## 1	125	153
## 2	9656	8775
## 3	200	180
## 4	10228	8859
## 5	94	101
## 6	118	121
## GSM1579257_jac011-430v2.CEL	GSM1579258_jac012-430v2.CEL	
## 1	182	130
## 2	13154	9498
## 3	218	195
## 4	13689	9705
## 5	104	108
## 6	132	128
## GSM1579259_jac037-430v2.CEL	GSM1579260_jac048-430v2.CEL	
## 1	176	92
## 2	15363	4973
## 3	251	72
## 4	15737	5058
## 5	142	72
## 6	173	63
## GSM1579261_jac015-430v2.CEL	GSM1579262_jac016-430v2.CEL	
## 1	59	69
## 2	5959	6331
## 3	73	72
## 4	6140	6388
## 5	66	65
## 6	51	71
## GSM1579263_jac017-430v2.CEL	GSM1579264_jac018-430v2.CEL	
## 1	58	69
## 2	5856	6670
## 3	78	82
## 4	5730	6455
## 5	61	92
## 6	59	47
## GSM1579265_jac019-430v2.CEL	GSM1579266_jac020-430v2.CEL	
## 1	87	202
## 2	6029	6455
## 3	78	74
## 4	6409	6648
## 5	65	102
## 6	67	58
## GSM1579267_jac027-430v2.CEL	GSM1579268_jac028-430v2.CEL	
## 1	161	123
## 2	7684	8069
## 3	230	156
## 4	7162	8064
## 5	123	97
## 6	139	146
## GSM1579269_jac034-430v2.CEL	GSM1579270_jac033-430v2.CEL	
## 1	193	150
## 2	14103	7298

```

## 3 279 231
## 4 14199 7447
## 5 156 145
## 6 197 139
## GSM1579271_jac029-430v2.CEL GSM1579272_jac035-430v2.CEL
## 1 242 149
## 2 18215 7908
## 3 283 150
## 4 17438 8798
## 5 160 137
## 6 151 141
## GSM1579273_jac041-430v2.CEL GSM1579274_jac042-430v2.CEL
## 1 132 184
## 2 8993 8146
## 3 185 169
## 4 9202 8344
## 5 153 180
## 6 117 152
## GSM1579275_jac043-430v2.CEL GSM1579276_jac030-430v2.CEL
## 1 187 101
## 2 6867 6737
## 3 223 92
## 4 7165 6788
## 5 101 61
## 6 138 83
## GSM1579277_jac031-430v2.CEL GSM1579278_jac032-430v2.CEL
## 1 74 165
## 2 7379 7416
## 3 67 201
## 4 7312 7929
## 5 65 157
## 6 77 141
## GSM1579279_jac038-430v2.CEL GSM1579280_jac039-430v2.CEL
## 1 72 72
## 2 4883 4924
## 3 68 74
## 4 4814 4988
## 5 68 65
## 6 71 57
## GSM1579281_jac040-430v2.CEL
## 1 81
## 2 4475
## 3 78
## 4 4646
## 5 59
## 6 67

```

## Annotation

Here we provide basic sample annotation, including the phenotype of interest and relevant other features (e.g. confounders). This dataset contains array data (A-AFFY-45) of different mice strains (BXD98, BXD97, BXD83, BXD73, BXD68, BXD67 ,BXD43, C57BL/6J, DBA/2J) infected with influenza virus H5N1.

```

sdrf <- read.delim("./Microarray2/E-GEO-64750.sdrf.txt")
print(sdrf[,c("Source.Name", "Comment..Sample_source_name.", "Array.Design.REF", "Characteristics..strain")]

```

	Source.Name	Comment..Sample_source_name.
## 1	GSM1579281	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 2	GSM1579280	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 3	GSM1579279	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 4	GSM1579278	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 5	GSM1579277	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 6	GSM1579276	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 7	GSM1579275	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 8	GSM1579274	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 9	GSM1579273	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 10	GSM1579272	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 11	GSM1579271	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 12	GSM1579270	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 13	GSM1579269	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 14	GSM1579268	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 15	GSM1579267	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 16	GSM1579266	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 17	GSM1579265	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 18	GSM1579264	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 19	GSM1579263	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 20	GSM1579262	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 21	GSM1579261	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 22	GSM1579260	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 23	GSM1579259	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 24	GSM1579258	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 25	GSM1579257	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 26	GSM1579256	Uninfected control
## 27	GSM1579255	Uninfected control
## 28	GSM1579254	Uninfected control
## 29	GSM1579253	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 30	GSM1579252	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 31	GSM1579251	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 32	GSM1579250	Highly Pathogenic H5N1 Influenza A virus infected; 72 hours
## 33	GSM1579249	Uninfected control
## 34	GSM1579248	Uninfected control
## 35	GSM1579247	Uninfected control
## 36	GSM1579246	Uninfected control
## 37	GSM1579245	Uninfected control
##	Array.Design.REF	Characteristics..strain.
## 1	A-AFFY-45	BXD98
## 2	A-AFFY-45	BXD98
## 3	A-AFFY-45	BXD98
## 4	A-AFFY-45	BXD97
## 5	A-AFFY-45	BXD97
## 6	A-AFFY-45	BXD97
## 7	A-AFFY-45	BXD83
## 8	A-AFFY-45	BXD83
## 9	A-AFFY-45	BXD83
## 10	A-AFFY-45	BXD73
## 11	A-AFFY-45	BXD73

```

## 12      A-AFFY-45          BXD73
## 13      A-AFFY-45          BXD68
## 14      A-AFFY-45          BXD68
## 15      A-AFFY-45          BXD68
## 16      A-AFFY-45          BXD67
## 17      A-AFFY-45          BXD67
## 18      A-AFFY-45          BXD67
## 19      A-AFFY-45          BXD43
## 20      A-AFFY-45          BXD43
## 21      A-AFFY-45          BXD43
## 22      A-AFFY-45          C57BL/6J
## 23      A-AFFY-45          C57BL/6J
## 24      A-AFFY-45          C57BL/6J
## 25      A-AFFY-45          C57BL/6J
## 26      A-AFFY-45          C57BL/6J
## 27      A-AFFY-45          C57BL/6J
## 28      A-AFFY-45          C57BL/6J
## 29      A-AFFY-45          DBA/2J
## 30      A-AFFY-45          DBA/2J
## 31      A-AFFY-45          DBA/2J
## 32      A-AFFY-45          DBA/2J
## 33      A-AFFY-45          DBA/2J
## 34      A-AFFY-45          DBA/2J
## 35      A-AFFY-45          DBA/2J
## 36      A-AFFY-45          DBA/2J
## 37      A-AFFY-45          DBA/2J
##                                     Comment..Sample_description.
## 1  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 2  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 3  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 4  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 5  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 6  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 7  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 8  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 9  Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 10 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 11 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 12 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 13 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 14 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 15 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 16 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 17 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 18 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 19 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 20 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 21 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 22 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 23 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 24 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 25 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 26                                     Gene expression data from lungs of uninfected mice
## 27                                     Gene expression data from lungs of uninfected mice

```

```

## 28 Gene expression data from lungs of uninfected mice
## 29 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 30 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 31 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 32 Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected mice
## 33 Gene expression data from lungs of uninfected mice
## 34 Gene expression data from lungs of uninfected mice
## 35 Gene expression data from lungs of uninfected mice
## 36 Gene expression data from lungs of uninfected mice
## 37 Gene expression data from lungs of uninfected mice

```

### Which samples we are using, and not using: We will be using samples involving susceptible and resistant mouse strain, DBA/2J (GSM1579245 - GSM1579253) and C57BL/6J (GSM1579254 - GSM1579260) respectively. Each strain was inoculated with H5N1 influenza A virus. We are not using the data from other strains (BXD98, BXD97, BXD83, BXD73, BXD68, BXD67, BXD43) which do not contain non-infected control samples.

```

#Load in the ExpressionFeatureSet object
setwd("/Users/seoyeon/Desktop/MSc Bioinformatics Year 1/Applied High-throughput Analysis/Project/Datasets")
MouseExp_AE2 <- ArrayExpress("E-GEO-D-64750")

```

```

## Copying raw data files

## Unpacking data files

## ArrayExpress: Reading pheno data from SDRF

## ArrayExpress: Reading data files

## Platform design info loaded.

## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579281_jac040-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579280_jac039-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579279_jac038-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579278_jac032-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579277_jac031-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579276_jac030-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579275_jac043-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579274_jac042-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579273_jac041-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579272_jac035-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579271_jac029-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579270_jac033-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579269_jac034-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579268_jac028-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579267_jac027-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579266_jac020-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579265_jac019-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579264_jac018-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579263_jac017-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579262_jac016-430v2.CER
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579261_jac015-430v2.CER

```

```

## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579260_jac048-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579259_jac037-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579258_jac012-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579257_jac011-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579256_jac025-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579255_jac024-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579254_jac014-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579253_jac047-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579252_jac036-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579251_jac008-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579250_jac007-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579249_jac046-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579248_jac045-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579247_jac044-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579246_jac021-430v2.CEL
## Reading in : /var/folders/my/9t5kkz3n4x15jqxc9cf6hy740000gn/T//RtmpbPVBkJ/GSM1579245_jac013-430v2.CEL

##
## E-GEOD-64750 was successfully loaded into ExpressionFeatureSet

#Load in the Affybatch object
setwd("/Users/seoyeon/Desktop/MSc Bioinformatics Year 1/Applied High-throughput Analysis/Project/Datasets")
MouseExp_pheno2 <- ReadAffy(phenoData=pData(MouseExp_AE2))

#load first lines of output from the object
head(exprs(MouseExp_pheno2))

##      GSM1579281_jac040-430v2.CEL GSM1579280_jac039-430v2.CEL
## 1          81                  72
## 2         4475                 4924
## 3          78                  74
## 4         4646                 4988
## 5          59                  65
## 6          67                  57
##      GSM1579279_jac038-430v2.CEL GSM1579278_jac032-430v2.CEL
## 1          72                  165
## 2         4883                 7416
## 3          68                  201
## 4         4814                 7929
## 5          68                  157
## 6          71                  141
##      GSM1579277_jac031-430v2.CEL GSM1579276_jac030-430v2.CEL
## 1          74                  101
## 2         7379                 6737
## 3          67                  92
## 4         7312                 6788
## 5          65                  61
## 6          77                  83
##      GSM1579275_jac043-430v2.CEL GSM1579274_jac042-430v2.CEL
## 1          187                 184
## 2         6867                 8146
## 3          223                 169
## 4         7165                 8344

```

## 5	101	180
## 6	138	152
## GSM1579273_jac041-430v2.CEL	GSM1579272_jac035-430v2.CEL	
## 1	132	149
## 2	8993	7908
## 3	185	150
## 4	9202	8798
## 5	153	137
## 6	117	141
## GSM1579271_jac029-430v2.CEL	GSM1579270_jac033-430v2.CEL	
## 1	242	150
## 2	18215	7298
## 3	283	231
## 4	17438	7447
## 5	160	145
## 6	151	139
## GSM1579269_jac034-430v2.CEL	GSM1579268_jac028-430v2.CEL	
## 1	193	123
## 2	14103	8069
## 3	279	156
## 4	14199	8064
## 5	156	97
## 6	197	146
## GSM1579267_jac027-430v2.CEL	GSM1579266_jac020-430v2.CEL	
## 1	161	202
## 2	7684	6455
## 3	230	74
## 4	7162	6648
## 5	123	102
## 6	139	58
## GSM1579265_jac019-430v2.CEL	GSM1579264_jac018-430v2.CEL	
## 1	87	69
## 2	6029	6670
## 3	78	82
## 4	6409	6455
## 5	65	92
## 6	67	47
## GSM1579263_jac017-430v2.CEL	GSM1579262_jac016-430v2.CEL	
## 1	58	69
## 2	5856	6331
## 3	78	72
## 4	5730	6388
## 5	61	65
## 6	59	71
## GSM1579261_jac015-430v2.CEL	GSM1579260_jac048-430v2.CEL	
## 1	59	92
## 2	5959	4973
## 3	73	72
## 4	6140	5058
## 5	66	72
## 6	51	63
## GSM1579259_jac037-430v2.CEL	GSM1579258_jac012-430v2.CEL	
## 1	176	130
## 2	15363	9498

```

## 3          251          195
## 4          15737         9705
## 5          142          108
## 6          173          128
##   GSM1579257_jac011-430v2.CEL GSM1579256_jac025-430v2.CEL
## 1          182          153
## 2          13154         8775
## 3          218          180
## 4          13689         8859
## 5          104          101
## 6          132          121
##   GSM1579255_jac024-430v2.CEL GSM1579254_jac014-430v2.CEL
## 1          125          105
## 2          9656          8341
## 3          200          159
## 4          10228         8721
## 5          94           109
## 6          118           96
##   GSM1579253_jac047-430v2.CEL GSM1579252_jac036-430v2.CEL
## 1          92            80
## 2          4988          6310
## 3          87            102
## 4          5275          6335
## 5          87            70
## 6          79            88
##   GSM1579251_jac008-430v2.CEL GSM1579250_jac007-430v2.CEL
## 1          139          123
## 2          8743          7786
## 3          154          162
## 4          8940          7832
## 5          102          111
## 6          96           105
##   GSM1579249_jac046-430v2.CEL GSM1579248_jac045-430v2.CEL
## 1          136          134
## 2          6036          5697
## 3          134          144
## 4          6147          6184
## 5          126          121
## 6          113          128
##   GSM1579247_jac044-430v2.CEL GSM1579246_jac021-430v2.CEL
## 1          198          191
## 2          7485          16272
## 3          200          265
## 4          7664          16107
## 5          154          143
## 6          162          143
##   GSM1579245_jac013-430v2.CEL
## 1          62
## 2          5707
## 3          69
## 4          5913
## 5          57
## 6          54

```

```

head(pData(MouseExp_pheno2)) #source name, comment sample description, sample source name, sample titl

##                               Source.Name
## GSM1579281_jac040-430v2.CEL GSM1579281 1
## GSM1579280_jac039-430v2.CEL GSM1579280 1
## GSM1579279_jac038-430v2.CEL GSM1579279 1
## GSM1579278_jac032-430v2.CEL GSM1579278 1
## GSM1579277_jac031-430v2.CEL GSM1579277 1
## GSM1579276_jac030-430v2.CEL GSM1579276 1
##                                         Comment..Sample_source_name.
## GSM1579281_jac040-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
## GSM1579280_jac039-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
## GSM1579279_jac038-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
## GSM1579278_jac032-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
## GSM1579277_jac031-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
## GSM1579276_jac030-430v2.CEL Gene expression data from lungs of highly pathogenic H5N1 influenza A virus infected; 72 hours post H5N1 infection
##                                         Comment..Sample_source_name.
## GSM1579281_jac040-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
## GSM1579280_jac039-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
## GSM1579279_jac038-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
## GSM1579278_jac032-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
## GSM1579277_jac031-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
## GSM1579276_jac030-430v2.CEL Highly Pathogenic H5N1 Influenza A virus infected; 72 hours post H5N1 infection
##                                         Characteristics..age. Characteristics..organism.
## GSM1579281_jac040-430v2.CEL          6-8wk old           Mus musculus
## GSM1579280_jac039-430v2.CEL          6-8wk old           Mus musculus
## GSM1579279_jac038-430v2.CEL          6-8wk old           Mus musculus
## GSM1579278_jac032-430v2.CEL          6-8wk old           Mus musculus
## GSM1579277_jac031-430v2.CEL          6-8wk old           Mus musculus
## GSM1579276_jac030-430v2.CEL          6-8wk old           Mus musculus
##                                         Term.Source.REF
## GSM1579281_jac040-430v2.CEL          EFO
## GSM1579280_jac039-430v2.CEL          EFO
## GSM1579279_jac038-430v2.CEL          EFO
## GSM1579278_jac032-430v2.CEL          EFO
## GSM1579277_jac031-430v2.CEL          EFO
## GSM1579276_jac030-430v2.CEL          EFO
##                                         Term.Accession.Number
## GSM1579281_jac040-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
## GSM1579280_jac039-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
## GSM1579279_jac038-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
## GSM1579278_jac032-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
## GSM1579277_jac031-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
## GSM1579276_jac030-430v2.CEL http://purl.obolibrary.org/obo/NCBITaxon_10090
##                                         Characteristics..sex. Term.Source.REF.1
## GSM1579281_jac040-430v2.CEL          female            EFO

```

```

## GSM1579280_jac039-430v2.CEL female EFO
## GSM1579279_jac038-430v2.CEL female EFO
## GSM1579278_jac032-430v2.CEL female EFO
## GSM1579277_jac031-430v2.CEL female EFO
## GSM1579276_jac030-430v2.CEL female EFO
## Term.Accession.Number.1 Characteristics..strain.
## GSM1579281_jac040-430v2.CEL EFO_0001265 BXD98
## GSM1579280_jac039-430v2.CEL EFO_0001265 BXD98
## GSM1579279_jac038-430v2.CEL EFO_0001265 BXD98
## GSM1579278_jac032-430v2.CEL EFO_0001265 BXD97
## GSM1579277_jac031-430v2.CEL EFO_0001265 BXD97
## GSM1579276_jac030-430v2.CEL EFO_0001265 BXD97
## Term.Source.REF.2 Term.Accession.Number.2
## GSM1579281_jac040-430v2.CEL
## GSM1579280_jac039-430v2.CEL
## GSM1579279_jac038-430v2.CEL
## GSM1579278_jac032-430v2.CEL
## GSM1579277_jac031-430v2.CEL
## GSM1579276_jac030-430v2.CEL
## Protocol.REF Term.Source.REF.3 Protocol.REF.1
## GSM1579281_jac040-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## GSM1579280_jac039-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## GSM1579279_jac038-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## GSM1579278_jac032-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## GSM1579277_jac031-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## GSM1579276_jac030-430v2.CEL P-GSE64750-2 ArrayExpress P-GSE64750-3
## Term.Source.REF.4 Extract.Name
## GSM1579281_jac040-430v2.CEL ArrayExpress GSM1579281 extract 1
## GSM1579280_jac039-430v2.CEL ArrayExpress GSM1579280 extract 1
## GSM1579279_jac038-430v2.CEL ArrayExpress GSM1579279 extract 1
## GSM1579278_jac032-430v2.CEL ArrayExpress GSM1579278 extract 1
## GSM1579277_jac031-430v2.CEL ArrayExpress GSM1579277 extract 1
## GSM1579276_jac030-430v2.CEL ArrayExpress GSM1579276 extract 1
## Material.Type Protocol.REF.2 Term.Source.REF.5
## GSM1579281_jac040-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## GSM1579280_jac039-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## GSM1579279_jac038-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## GSM1579278_jac032-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## GSM1579277_jac031-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## GSM1579276_jac030-430v2.CEL total RNA P-GSE64750-4 ArrayExpress
## Labeled.Extract.Name Label Protocol.REF.3
## GSM1579281_jac040-430v2.CEL GSM1579281 LE 1 biotin P-GSE64750-5
## GSM1579280_jac039-430v2.CEL GSM1579280 LE 1 biotin P-GSE64750-5
## GSM1579279_jac038-430v2.CEL GSM1579279 LE 1 biotin P-GSE64750-5
## GSM1579278_jac032-430v2.CEL GSM1579278 LE 1 biotin P-GSE64750-5
## GSM1579277_jac031-430v2.CEL GSM1579277 LE 1 biotin P-GSE64750-5
## GSM1579276_jac030-430v2.CEL GSM1579276 LE 1 biotin P-GSE64750-5
## Term.Source.REF.6 Assay.Name Array.Design.REF
## GSM1579281_jac040-430v2.CEL ArrayExpress GSM1579281 A-AFFY-45
## GSM1579280_jac039-430v2.CEL ArrayExpress GSM1579280 A-AFFY-45
## GSM1579279_jac038-430v2.CEL ArrayExpress GSM1579279 A-AFFY-45
## GSM1579278_jac032-430v2.CEL ArrayExpress GSM1579278 A-AFFY-45
## GSM1579277_jac031-430v2.CEL ArrayExpress GSM1579277 A-AFFY-45
## GSM1579276_jac030-430v2.CEL ArrayExpress GSM1579276 A-AFFY-45

```

```

##                               Term.Source.REF.7 Technology.Type Protocol.REF.4
## GSM1579281_jac040-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
## GSM1579280_jac039-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
## GSM1579279_jac038-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
## GSM1579278_jac032-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
## GSM1579277_jac031-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
## GSM1579276_jac030-430v2.CEL      ArrayExpress    array assay   P-GSE64750-6
##                               Term.Source.REF.8          Array.Data.File
## GSM1579281_jac040-430v2.CEL      ArrayExpress  GSM1579281_jac040-430v2.CEL
## GSM1579280_jac039-430v2.CEL      ArrayExpress  GSM1579280_jac039-430v2.CEL
## GSM1579279_jac038-430v2.CEL      ArrayExpress  GSM1579279_jac038-430v2.CEL
## GSM1579278_jac032-430v2.CEL      ArrayExpress  GSM1579278_jac032-430v2.CEL
## GSM1579277_jac031-430v2.CEL      ArrayExpress  GSM1579277_jac031-430v2.CEL
## GSM1579276_jac030-430v2.CEL      ArrayExpress  GSM1579276_jac030-430v2.CEL
##                               Protocol.REF.5 Term.Source.REF.9
## GSM1579281_jac040-430v2.CEL      P-GSE64750-1    ArrayExpress
## GSM1579280_jac039-430v2.CEL      P-GSE64750-1    ArrayExpress
## GSM1579279_jac038-430v2.CEL      P-GSE64750-1    ArrayExpress
## GSM1579278_jac032-430v2.CEL      P-GSE64750-1    ArrayExpress
## GSM1579277_jac031-430v2.CEL      P-GSE64750-1    ArrayExpress
## GSM1579276_jac030-430v2.CEL      P-GSE64750-1    ArrayExpress
##                               Normalization.Name
## GSM1579281_jac040-430v2.CEL  GSM1579281_sample_table.txt norm
## GSM1579280_jac039-430v2.CEL  GSM1579280_sample_table.txt norm
## GSM1579279_jac038-430v2.CEL  GSM1579279_sample_table.txt norm
## GSM1579278_jac032-430v2.CEL  GSM1579278_sample_table.txt norm
## GSM1579277_jac031-430v2.CEL  GSM1579277_sample_table.txt norm
## GSM1579276_jac030-430v2.CEL  GSM1579276_sample_table.txt norm
##                               Derived.Array.Data.File
## GSM1579281_jac040-430v2.CEL  GSM1579281_sample_table.txt
## GSM1579280_jac039-430v2.CEL  GSM1579280_sample_table.txt
## GSM1579279_jac038-430v2.CEL  GSM1579279_sample_table.txt
## GSM1579278_jac032-430v2.CEL  GSM1579278_sample_table.txt
## GSM1579277_jac031-430v2.CEL  GSM1579277_sample_table.txt
## GSM1579276_jac030-430v2.CEL  GSM1579276_sample_table.txt
##                               Protocol.REF.10
## GSM1579281_jac040-430v2.CEL      BXD98
## GSM1579280_jac039-430v2.CEL      BXD98
## GSM1579279_jac038-430v2.CEL      BXD98
## GSM1579278_jac032-430v2.CEL      BXD97

```

```

## GSM1579277_jac031-430v2.CEL           BXD97
## GSM1579276_jac030-430v2.CEL           BXD97
##                                         Term.Accession.Number.3
## GSM1579281_jac040-430v2.CEL
## GSM1579280_jac039-430v2.CEL
## GSM1579279_jac038-430v2.CEL
## GSM1579278_jac032-430v2.CEL
## GSM1579277_jac031-430v2.CEL
## GSM1579276_jac030-430v2.CEL

filter2 <- colnames(data.raw_2)[data.raw_2@phenoData@data$index <= 16]
filter2

## [1] "GSM1579245_jac013-430v2.CEL" "GSM1579246_jac021-430v2.CEL"
## [3] "GSM1579247_jac044-430v2.CEL" "GSM1579248_jac045-430v2.CEL"
## [5] "GSM1579249_jac046-430v2.CEL" "GSM1579250_jac007-430v2.CEL"
## [7] "GSM1579251_jac008-430v2.CEL" "GSM1579252_jac036-430v2.CEL"
## [9] "GSM1579253_jac047-430v2.CEL" "GSM1579254_jac014-430v2.CEL"
## [11] "GSM1579255_jac024-430v2.CEL" "GSM1579256_jac025-430v2.CEL"
## [13] "GSM1579257_jac011-430v2.CEL" "GSM1579258_jac012-430v2.CEL"
## [15] "GSM1579259_jac037-430v2.CEL" "GSM1579260_jac048-430v2.CEL"

filtered2 <- data.raw_2[,filter2]
filtered2

## ExpressionFeatureSet (storageMode: lockedEnvironment)
## assayData: 1004004 features, 16 samples
##   element names: exprs
## protocolData
##   rowNames: GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL ...
##             GSM1579260_jac048-430v2.CEL (16 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL ...
##             GSM1579260_jac048-430v2.CEL (16 total)
##   varLabels: index
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: pd.mouse430.2

dim(exprs(filtered)) #1004004 features          9 samples

## [1] 1102500      6

head(exprs(filtered2))

##   GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL
## 1                               62                           191
## 2                             5707                         16272

```

## 3	69	265
## 4	5913	16107
## 5	57	143
## 6	54	143
## GSM1579247_jac044-430v2.CEL	GSM1579248_jac045-430v2.CEL	
## 1	198	134
## 2	7485	5697
## 3	200	144
## 4	7664	6184
## 5	154	121
## 6	162	128
## GSM1579249_jac046-430v2.CEL	GSM1579250_jac007-430v2.CEL	
## 1	136	123
## 2	6036	7786
## 3	134	162
## 4	6147	7832
## 5	126	111
## 6	113	105
## GSM1579251_jac008-430v2.CEL	GSM1579252_jac036-430v2.CEL	
## 1	139	80
## 2	8743	6310
## 3	154	102
## 4	8940	6335
## 5	102	70
## 6	96	88
## GSM1579253_jac047-430v2.CEL	GSM1579254_jac014-430v2.CEL	
## 1	92	105
## 2	4988	8341
## 3	87	159
## 4	5275	8721
## 5	87	109
## 6	79	96
## GSM1579255_jac024-430v2.CEL	GSM1579256_jac025-430v2.CEL	
## 1	125	153
## 2	9656	8775
## 3	200	180
## 4	10228	8859
## 5	94	101
## 6	118	121
## GSM1579257_jac011-430v2.CEL	GSM1579258_jac012-430v2.CEL	
## 1	182	130
## 2	13154	9498
## 3	218	195
## 4	13689	9705
## 5	104	108
## 6	132	128
## GSM1579259_jac037-430v2.CEL	GSM1579260_jac048-430v2.CEL	
## 1	176	92
## 2	15363	4973
## 3	251	72
## 4	15737	5058
## 5	142	72
## 6	173	63

```

#arrayQualityMetrics(filtered2, outdir=".~/raw2", force=T)
#arrayQualityMetrics(filtered2, outdir=".~/rawlog2", force=T, do.logtransform=T)

miceRMA <- oligo:::rma(filtered2, background=T)

## Background correcting
## Normalizing
## Calculating Expression

head(miceRMA)

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 6 features, 16 samples
##   element names: exprs
## protocolData
##   rowNames: GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL ...
##   GSM1579260_jac048-430v2.CEL (16 total)
##   varLabels: exprs dates
##   varMetadata: labelDescription channel
## phenoData
##   rowNames: GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL ...
##   GSM1579260_jac048-430v2.CEL (16 total)
##   varLabels: index
##   varMetadata: labelDescription channel
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation: pd.mouse430.2

#arrayQualityMetrics(miceRMA, outdir=".~/rma2", force=TRUE)

## Differential expression analysis with RMA preprocessed data
#####
## Additional preprocessing
samples <- c(replicate(5, "DBA/2J control"), replicate(4, "DBA/2J infected"), replicate(3, "C57/BL6J control"))
samples

## [1] "DBA/2J control"      "DBA/2J control"      "DBA/2J control"
## [4] "DBA/2J control"      "DBA/2J control"      "DBA/2J infected"
## [7] "DBA/2J infected"     "DBA/2J infected"     "DBA/2J infected"
## [10] "C57/BL6J control"    "C57/BL6J control"    "C57/BL6J control"
## [13] "C57/BL6J infected"   "C57/BL6J infected"   "C57/BL6J infected"
## [16] "C57/BL6J infected"

condition <- c(replicate(5, "control"), replicate(4, "infected"), replicate(3, "control"), replicate(4, "infected"))

pData(miceRMA)[,2] <- condition
pData(miceRMA)[,3] <- c(replicate(9, "DBA/2J"), replicate(7, "C57/BL6J"))
pData(miceRMA)[,4] <- samples

colnames(pData(miceRMA)) <- c("index", "condition", "strain", "samples")
pData(miceRMA)

```

```

##          index condition strain      samples
## GSM1579245_jac013-430v2.CEL    1   control DBA/2J DBA/2J control
## GSM1579246_jac021-430v2.CEL    2   control DBA/2J DBA/2J control
## GSM1579247_jac044-430v2.CEL    3   control DBA/2J DBA/2J control
## GSM1579248_jac045-430v2.CEL    4   control DBA/2J DBA/2J control
## GSM1579249_jac046-430v2.CEL    5   control DBA/2J DBA/2J control
## GSM1579250_jac007-430v2.CEL    6 infected DBA/2J DBA/2J infected
## GSM1579251_jac008-430v2.CEL    7 infected DBA/2J DBA/2J infected
## GSM1579252_jac036-430v2.CEL    8 infected DBA/2J DBA/2J infected
## GSM1579253_jac047-430v2.CEL    9 infected DBA/2J DBA/2J infected
## GSM1579254_jac014-430v2.CEL   10 control C57/BL6J C57/BL6J control
## GSM1579255_jac024-430v2.CEL   11 control C57/BL6J C57/BL6J control
## GSM1579256_jac025-430v2.CEL   12 control C57/BL6J C57/BL6J control
## GSM1579257_jac011-430v2.CEL   13 infected C57/BL6J C57/BL6J infected
## GSM1579258_jac012-430v2.CEL   14 infected C57/BL6J C57/BL6J infected
## GSM1579259_jac037-430v2.CEL   15 infected C57/BL6J C57/BL6J infected
## GSM1579260_jac048-430v2.CEL   16 infected C57/BL6J C57/BL6J infected

```

The variability of the strain is encompassed in the model but you do not test for it

```

condition <- factor(pData(miceRMA)[,2])
strain <- factor(pData(miceRMA)[,3])
condition

## [1] control control control control control infected infected infected
## [9] infected control control control infected infected infected infected
## Levels: control infected

strain

## [1] DBA/2J DBA/2J DBA/2J DBA/2J DBA/2J DBA/2J DBA/2J DBA/2J
## [9] DBA/2J C57/BL6J C57/BL6J C57/BL6J C57/BL6J C57/BL6J C57/BL6J C57/BL6J
## Levels: C57/BL6J DBA/2J

design <- model.matrix(~0+condition*strain)
colnames(design)<-c("Control", "Infected", "strain", "interaction")
# for strain, 1 represents DBA/2J, 0 represents c57/BL6
design

##   Control Infected strain interaction
## 1       1       0     1         0
## 2       1       0     1         0
## 3       1       0     1         0
## 4       1       0     1         0
## 5       1       0     1         0
## 6       0       1     1         1
## 7       0       1     1         1
## 8       0       1     1         1
## 9       0       1     1         1
## 10      1       0     0         0
## 11      1       0     0         0
## 12      1       0     0         0

```

```

## 13      0      1      0      0
## 14      0      1      0      0
## 15      0      1      0      0
## 16      0      1      0      0
## attr(),"assign")
## [1] 1 1 2 3
## attr(),"contrasts")
## attr(),"contrasts")$condition
## [1] "contr.treatment"
##
## attr(),"contrasts")$strain
## [1] "contr.treatment"

fit_m2 <- lmFit(miceRMA, design)
cont.matrix <- makeContrasts(InfectedvsControl="Infected-Control", levels=design)
cont.matrix

##           Contrasts
## Levels      InfectedvsControl
## Control          -1
## Infected          1
## strain            0
## interaction       0

fit2_m2 <- contrasts.fit(fit_m2,cont.matrix)
fit2_m2 <- eBayes(fit2_m2)
fit2_m2

## An object of class "MArrayLM"
## $coefficients
##           Contrasts
##           InfectedvsControl
## 1415670_at      0.295615220
## 1415671_at      0.108222965
## 1415672_at     -0.001215109
## 1415673_at     -0.023467868
## 1415674_a_at    0.173622489
## 45096 more rows ...
##
## $rank
## [1] 4
##
## $assign
## [1] 1 1 2 3
##
## $qr
## $qr
##           Control  Infected   strain interaction
## 1 -2.8284271  0.000000 -1.7677670  0.0000000
## 2  0.3535534 -2.828427 -1.4142136 -1.4142136
## 3  0.3535534  0.000000 -1.9685020 -1.0160010
## 4  0.3535534  0.000000  0.1407408  0.9837388
## 5  0.3535534  0.000000  0.1407408  0.1274229

```

```

## 11 more rows ...
##
## $qraux
## [1] 1.353553 1.000000 1.140741 1.127423
##
## $pivot
## [1] 1 2 3 4
##
## $tol
## [1] 1e-07
##
## $rank
## [1] 4
##
## $df.residual
## [1] 12 12 12 12 12
## 45096 more elements ...
##
## $sigma
##    1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
##    0.3165328    0.2033128    0.1071823    0.3166509    0.1363164
## 45096 more elements ...
##
## $cov.coefficients
##          Contrasts
## Contrasts           InfectedvsControl
##   InfectedvsControl      0.5833333
##
## $stdev.unscaled
##          Contrasts
##          InfectedvsControl
##    1415670_at      0.7637626
##    1415671_at      0.7637626
##    1415672_at      0.7637626
##    1415673_at      0.7637626
##    1415674_a_at    0.7637626
## 45096 more rows ...
##
## $pivot
## [1] 1 2 3 4
##
## $Amean
##    1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
##    8.995360     10.365534    10.828212    7.422040    9.235163
## 45096 more elements ...
##
## $method
## [1] "ls"
##
## $design
##   Control Infected strain interaction
## 1       1       0       1       0
## 2       1       0       1       0

```

```

## 3      1      0      1      0
## 4      1      0      1      0
## 5      1      0      1      0
## 11 more rows ...
##
## $contrasts
##           Contrasts
## Levels      InfectedvsControl
##   Control          -1
##   Infected          1
##   strain            0
##   interaction       0
##
## $df.prior
## [1] 3.078919
##
## $s2.prior
## [1] 0.03390424
##
## $var.prior
## [1] 61.79253
##
## $proportion
## [1] 0.01
##
## $s2.post
##    1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
## 0.08665770  0.03981860  0.01606514  0.08671723  0.02171072
## 45096 more elements ...
##
## $t
##           Contrasts
##           InfectedvsControl
##   1415670_at      1.31481549
##   1415671_at      0.71009763
##   1415672_at     -0.01255205
##   1415673_at     -0.10434281
##   1415674_a_at     1.54280288
## 45096 more rows ...
##
## $df.total
## [1] 15.07892 15.07892 15.07892 15.07892 15.07892
## 45096 more elements ...
##
## $p.value
##           Contrasts
##           InfectedvsControl
##   1415670_at      0.2082185
##   1415671_at      0.4884808
##   1415672_at      0.9901497
##   1415673_at      0.9182725
##   1415674_a_at     0.1436020
## 45096 more rows ...
##

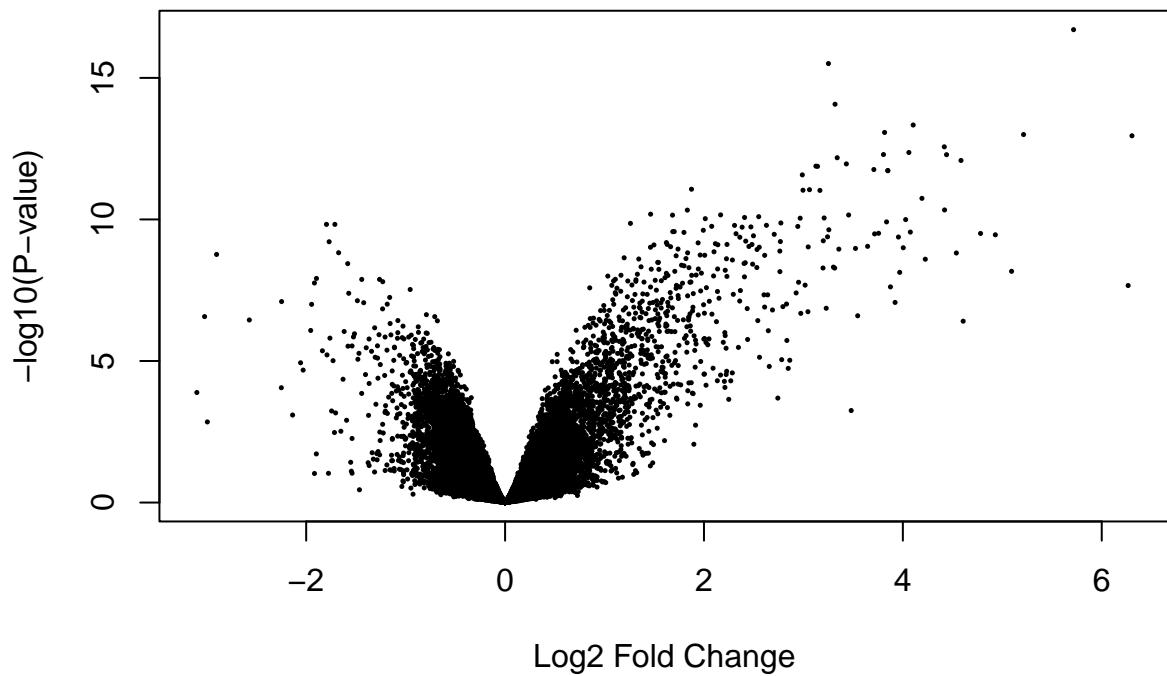
```

```

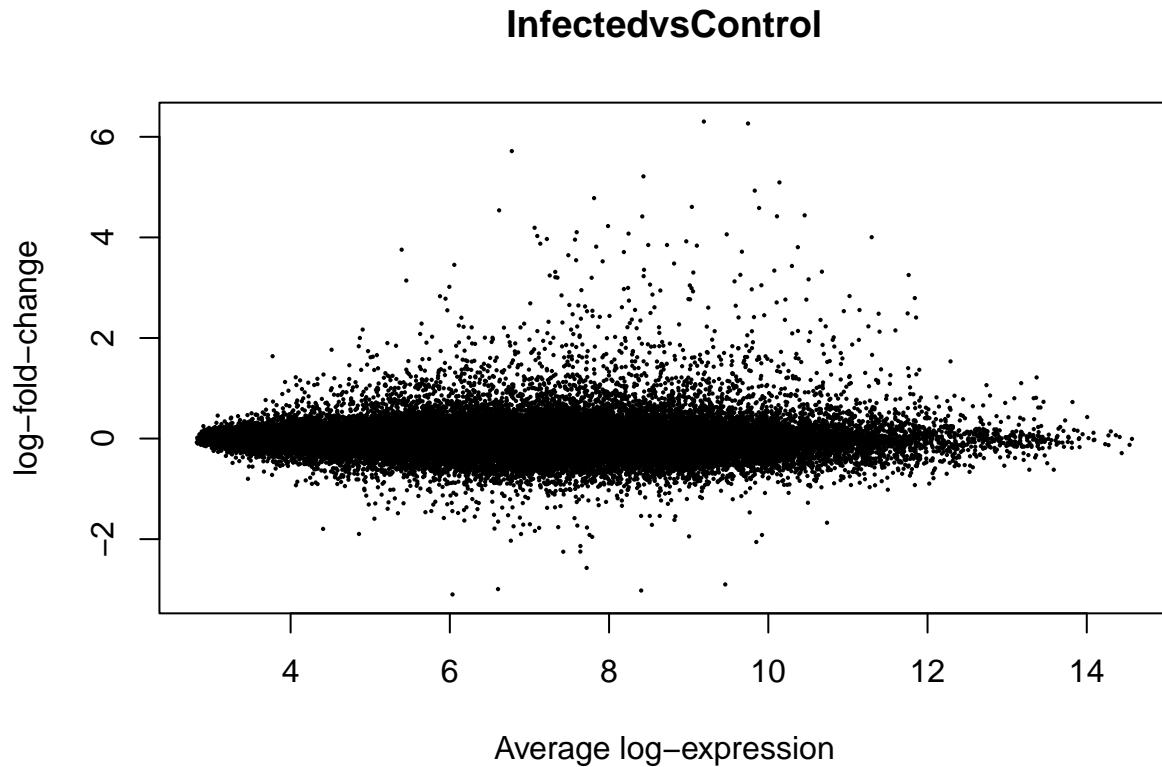
## $lods
##           Contrasts
##           InfectedvsControl
##   1415670_at      -6.067244
##   1415671_at      -6.669279
##   1415672_at      -6.931124
##   1415673_at      -6.925459
##   1415674_a_at    -5.764749
## 45096 more rows ...
##
## $F
## [1] 1.7287397757 0.5042386432 0.0001575539 0.0108874228 2.3802407312
## 45096 more elements ...
##
## $F.p.value
## [1] 0.2082185 0.4884808 0.9901497 0.9182725 0.1436020
## 45096 more elements ...

```

```
volcanoplot(fit2_m2)
```



```
limma::plotMA(fit2_m2)
```



```
# DE results
LIMMAout2 <- topTable(fit2_m2, adjust="BH", number=nrow(exprs(miceRMA)))
#head(LIMMAout)

## Check intensity values for top results
head(exprs(miceRMA)[rownames(exprs(miceRMA)) %in% rownames(head(LIMMAout2)),])
```

```
##          GSM1579245_jac013-430v2.CEL GSM1579246_jac021-430v2.CEL
## 1418293_at           8.850220          8.697405
## 1419697_at           4.873087          5.196294
## 1424518_at           5.610642          5.255470
## 1449025_at           9.824438          9.841855
## 1450297_at           3.450449          3.409701
## 1451905_a_at         5.271445          5.483409
##          GSM1579247_jac044-430v2.CEL GSM1579248_jac045-430v2.CEL
## 1418293_at           8.628802          8.583339
## 1419697_at           4.944273          4.957930
## 1424518_at           5.033368          5.491428
## 1449025_at           9.930787          9.733168
## 1450297_at           3.628502          3.614535
## 1451905_a_at         5.814011          5.916976
##          GSM1579249_jac046-430v2.CEL GSM1579250_jac007-430v2.CEL
## 1418293_at           8.979707          13.280513
```

```

## 1419697_at      5.156156      12.484048
## 1424518_at      5.391127      9.926284
## 1449025_at      10.059654     13.751370
## 1450297_at      3.661825      10.544846
## 1451905_a_at    6.290288     12.039855
##          GSM1579251_jac008-430v2.CEL GSM1579252_jac036-430v2.CEL
## 1418293_at      13.17569      13.17675
## 1419697_at      12.58070      12.54747
## 1424518_at      10.07358      10.38063
## 1449025_at      13.69643      13.85968
## 1450297_at      10.77543      10.84424
## 1451905_a_at    11.65031      12.11285
##          GSM1579253_jac047-430v2.CEL GSM1579254_jac014-430v2.CEL
## 1418293_at      13.07904      8.738460
## 1419697_at      12.59532      4.815025
## 1424518_at      10.27056      5.208682
## 1449025_at      13.77391      10.119994
## 1450297_at      10.40172      3.770888
## 1451905_a_at    11.95163      5.430919
##          GSM1579255_jac024-430v2.CEL GSM1579256_jac025-430v2.CEL
## 1418293_at      8.723118      8.699255
## 1419697_at      5.156243      4.948877
## 1424518_at      5.235854      5.688553
## 1449025_at      10.063244     10.129423
## 1450297_at      3.496160      3.556766
## 1451905_a_at    5.635768      5.014256
##          GSM1579257_jac011-430v2.CEL GSM1579258_jac012-430v2.CEL
## 1418293_at      12.021105     12.276407
## 1419697_at      8.623987      9.287277
## 1424518_at      9.305259      9.485391
## 1449025_at      13.289408     13.458878
## 1450297_at      9.421028      9.491732
## 1451905_a_at    10.501785     10.709535
##          GSM1579259_jac037-430v2.CEL GSM1579260_jac048-430v2.CEL
## 1418293_at      11.864963     11.991672
## 1419697_at      8.614776      8.632264
## 1424518_at      9.530186      9.606000
## 1449025_at      13.312026     13.363678
## 1450297_at      9.073683      9.311915
## 1451905_a_at    10.413056     10.672276

```

```

#mean expression of control/DBA2J
rowMeans(exprs(miceRMA)[rownames(exprs(miceRMA))%in%rownames(head(LIMMAout2)), 1:5])

```

```

## 1418293_at    1419697_at    1424518_at    1449025_at    1450297_at    1451905_a_at
##     8.747894    5.025548    5.356407    9.877981    3.553002    5.755226

```

```

#mean expression of infected/DBA2J
rowMeans(exprs(miceRMA)[rownames(exprs(miceRMA))%in%rownames(head(LIMMAout2)), 6:9])

```

```

## 1418293_at    1419697_at    1424518_at    1449025_at    1450297_at    1451905_a_at
##    13.17800    12.55189    10.16277    13.77035    10.64156    11.93866

```

```

#mean expression of control/C57BL6
rowMeans(exprs(miceRMA)[rownames(exprs(miceRMA))%in%rownames(head(LIMMAout2)), 10:12])

##    1418293_at    1419697_at    1424518_at    1449025_at    1450297_at 1451905_a_at
##    8.720277     4.973382     5.377696    10.104220     3.607938     5.360314

#mean expression of infected/C57BL6
rowMeans(exprs(miceRMA)[rownames(exprs(miceRMA))%in%rownames(head(LIMMAout2)), 13:16])

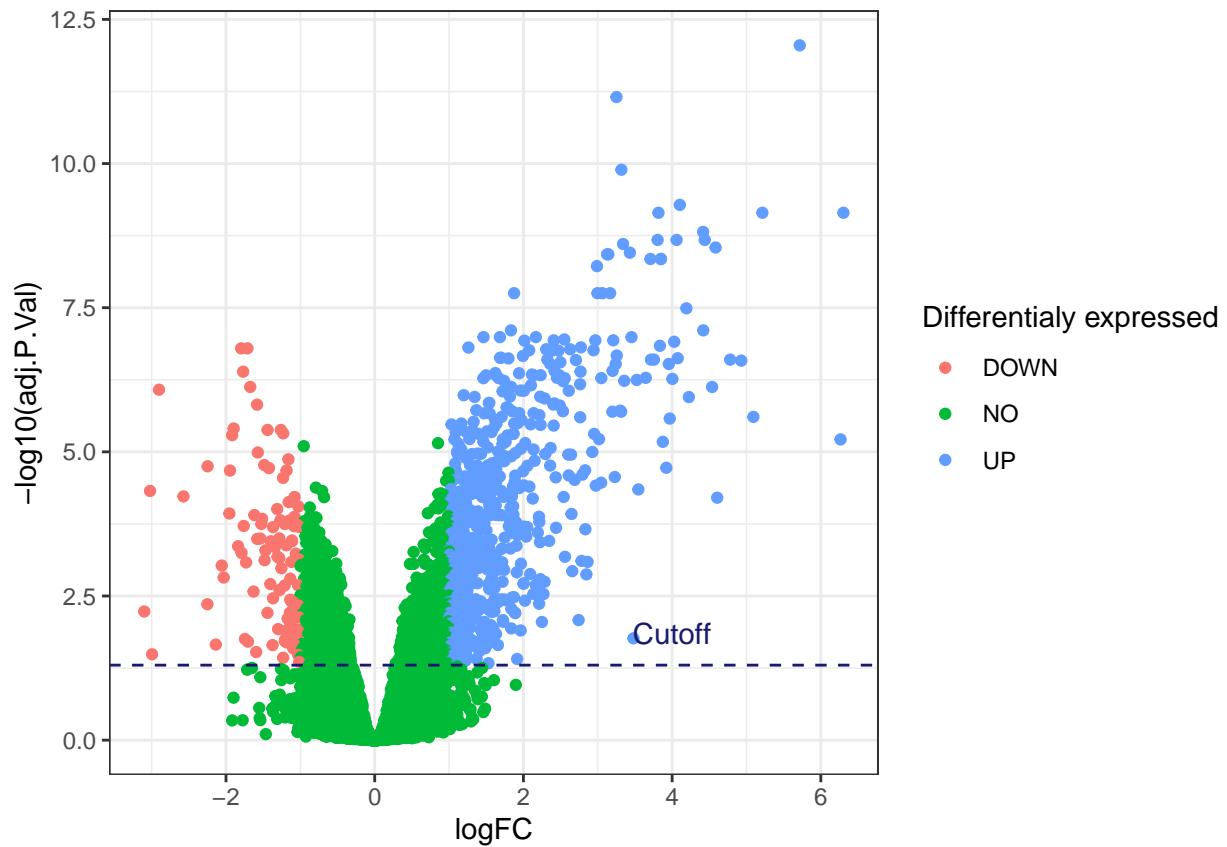
##    1418293_at    1419697_at    1424518_at    1449025_at    1450297_at 1451905_a_at
##    12.038537     8.789576     9.481709    13.355997     9.324589     10.574163

#Adjustments on p values using Benjamini-Hochberg
LIMMAout2$diffexpressed <- "NO"
LIMMAout2$diffexpressed[LIMMAout2$logFC > 1 & LIMMAout2$adj.P.Val < 0.05] <- "UP"
LIMMAout2$diffexpressed[LIMMAout2$logFC < -1 & LIMMAout2$adj.P.Val < 0.05] <- "DOWN"

#No adjustments on pvalues
LIMMAout2$diffexpressed_no_BH <- "NO"
LIMMAout2$diffexpressed_no_BH[LIMMAout2$logFC > 1 & LIMMAout2$P.Value < 0.05] <- "UP"
LIMMAout2$diffexpressed_no_BH[LIMMAout2$logFC < -1 & LIMMAout2$P.Value < 0.05] <- "DOWN"

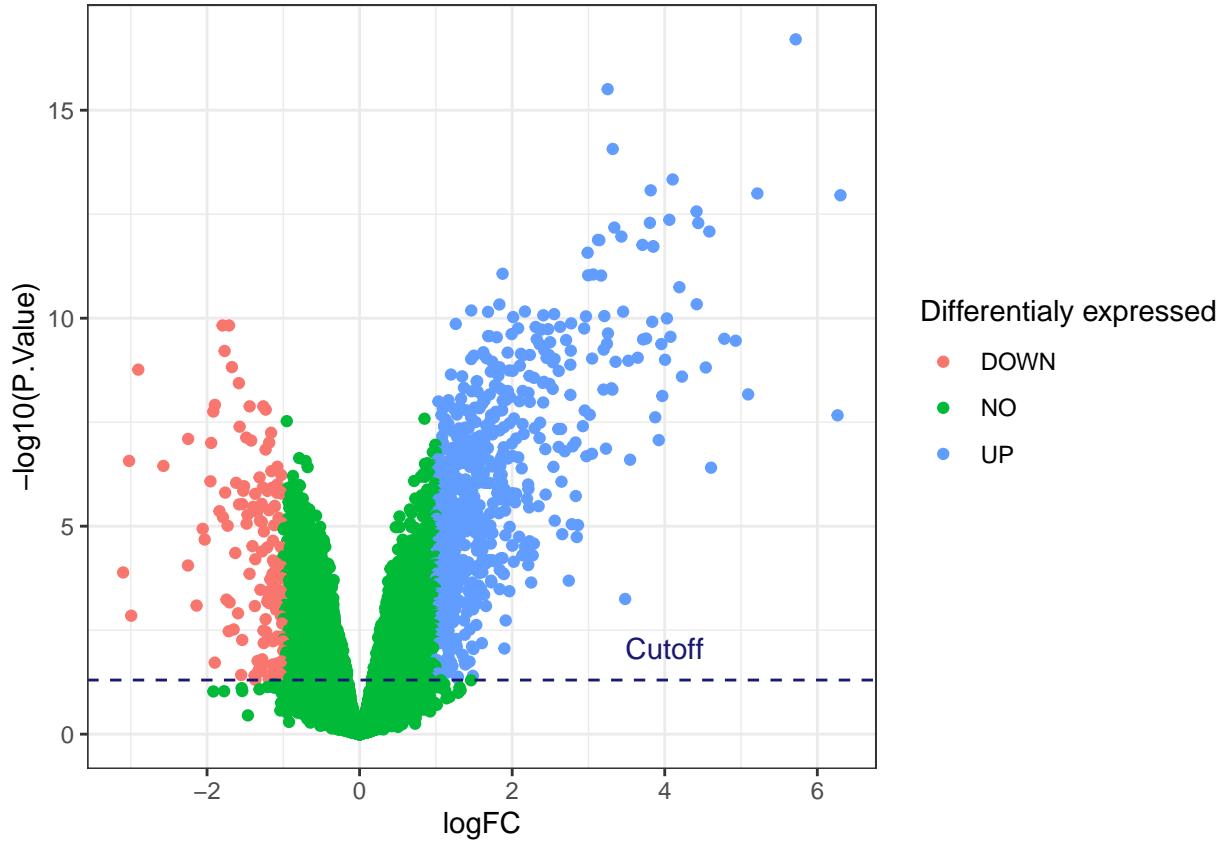
#jpeg("microarray2_volcanoplot.jpg")
ggplot(data = LIMMAout2, aes(x= logFC, y = -log10(adj.P.Val), colour = diffexpressed)) +
  geom_point()+
  theme_bw()+
  geom_hline(yintercept = -log10(0.05), linetype="dashed", color="midnightblue")+
  annotate("text", min(4), 1.3, vjust = -1, label = "Cutoff", color="midnightblue")+
  #ggtitle("Differentiall (unadjusted P-value)") +
  labs(colour = "Differentially expressed")

```



```
#theme(plot.title = element_text(hjust = 0.5, face = "bold.italic"))dev.off()

ggplot(data = LIMMAout2, aes(x= logFC, y = -log10(P.Value), colour = diffexpressed_no_BH)) +
  geom_point()+
  theme_bw()+
  geom_hline(yintercept = -log10(0.05), linetype="dashed", color="midnightblue")+
  annotate("text", min(4), 1.3, vjust = -1, label = "Cutoff", color="midnightblue")+
  #ggtitle("Differentiall (unadjusted P-value)") +
  labs(colour = "Differentially expressed")
```



```

length(which(LIMMAout2$diffexpressed=="UP")) #641 upregulated

## [1] 641

length(which(LIMMAout2$diffexpressed=="DOWN")) #100 downregulated

## [1] 100

length(which(LIMMAout2$diffexpressed_no_BH=="UP")) #701 upregulated

## [1] 701

length(which(LIMMAout2$diffexpressed_no_BH=="DOWN")) #136 downregulated

## [1] 136

## Load annotation and sort alphabetically on probe name
setwd("../Datasets/Microarray2/")
annotation_MA2 <- read.table("A-AFFY-45.adf.txt", header=T, sep="\t", skip=17, fill=T)
print(head(annotation_MA2))

```

```

## Composite.Element.Name Composite.Element.Database.Entry.interpro.
## 1          AFFX-BioB-5_at                               IPR007197
## 2          AFFX-BioB-M_at                             IPR007197
## 3          AFFX-BioB-3_at                             IPR007197
## 4          AFFX-BioC-5_at
## 5          AFFX-BioC-3_at
## 6          AFFX-BioDn-5_at                            IPR002586

## Composite.Element.Database.Entry.embl.
## 1          AFFX-BioB-5
## 2          AFFX-BioB-M
## 3          AFFX-BioB-3
## 4          AFFX-BioC-5
## 5          AFFX-BioC-3
## 6          AFFX-BioDn-5

## Composite.Element.Database.Entry.affymetrix_netaffx.
## 1          AFFX-BioB-5_at
## 2          AFFX-BioB-M_at
## 3          AFFX-BioB-3_at
## 4          AFFX-BioC-5_at
## 5          AFFX-BioC-3_at
## 6          AFFX-BioDn-5_at

## Composite.Element.Database.Entry.genbank.
## 1
## 2
## 3
## 4
## 5
## 6                               NP_752788

## Composite.Element.Database.Entry.ec. Composite.Element.Database.Entry.refseq.
## 1
## 2
## 3
## 4
## 5
## 6

## Composite.Element.Database.Entry.swall.
## 1
## 2
## 3
## 4
## 5
## 6

## Composite.Element.Database.Entry.ensembl.
## 1
## 2
## 3
## 4
## 5
## 6

## Composite.Element.Database.Entry.go.
## 1
## 2
## 3
## 4

```

```

## 5
## 6
##   Composite.Element.Database.Entry.unigene.
## 1
## 2
## 3
## 4
## 5
## 6
##   Composite.Element.Database.Entry.mgd. Composite.Element.Database.Entry.locus.
## 1          NA          NA
## 2          NA          NA
## 3          NA          NA
## 4          NA          NA
## 5          NA          NA
## 6          NA          NA
##   Composite.Element.Database.Entry.pkr_hanks.
## 1
## 2
## 3
## 4
## 5
## 6
##   Composite.Element.Database.Entry.scop.
## 1
## 2
## 3
## 4
## 5
## 6
##   Composite.Element.Database.Entry.cp450.
## 1
## 2
## 3
## 4
## 5
## 6

annotation_MA2 <- annotation_MA2[sort(annotation_MA2$Composite.Element.Name, index.return=T)$ix,]

## Check if all probes are present in both sets
dim(annotation_MA2)

## [1] 45101      16

dim(LIMMAout2)

## [1] 45101      8

## Double check => "Assumption is the mother of all fuck up's ;)"
sum(annotation_MA2$Composite.Element.Name==sort(rownames(LIMMAout2)))

## [1] 45101

```

```

## Sort LIMMA output alphabetically on probe name
LIMMAout_sorted2 <- LIMMAout2[sort(rownames(LIMMAout2), index.return=T)$ix,]

## Add gene names to LIMMA output
LIMMAout_sorted2$gene <- annotation_MA2$Composite.Element.Database.Entry.ensembl.

LIMMAout_annot2 <- LIMMAout_sorted2[sort(LIMMAout_sorted2$adj.P.Val, index.return=T)$ix,]

#sort by adjusted p value from most significant to least
LIMMAout_sorted2 <- LIMMAout_sorted2[order(LIMMAout_sorted2$adj.P.Val, decreasing= F),]

#extract top 50 significant DE genes
LIMMAout_sorted2[1:50,]$gene

## [1] "ENSMUSG00000025746" "ENSMUSG00000045303" "ENSMUSG00000045932"
## [4] "ENSMUSG00000051925" "ENSMUSG00000034855" "ENSMUSG00000029419"
## [7] "ENSMUSG00000023341" "ENSMUSG00000041827" "ENSMUSG00000022548"
## [10] "" "ENSMUSG00000047610" "ENSMUSG00000054261"
## [13] "" "ENSMUSG00000030107" "ENSMUSG00000025165"
## [16] "ENSMUSG00000020638" "ENSMUSG00000015947" "ENSMUSG00000022586"
## [19] "ENSMUSG00000027514" "" "ENSMUSG00000035152"
## [22] "ENSMUSG00000017830" "ENSMUSG00000029561" ""
## [25] "ENSMUSG00000048806" "ENSMUSG00000025498" "ENSMUSG00000035208"
## [28] "ENSMUSG00000022906" "ENSMUSG00000029379" ""
## [31] "ENSMUSG00000010358" "ENSMUSG00000039364" "ENSMUSG00000046031"
## [34] "" "ENSMUSG00000030921" "ENSMUSG00000019910"
## [37] "" "ENSMUSG00000001131" ""
## [40] "ENSMUSG00000003617" "ENSMUSG00000009670" "ENSMUSG00000028957"
## [43] "ENSMUSG00000023341" "ENSMUSG00000055116" "ENSMUSG00000015947"
## [46] "" ""
## [49] "ENSMUSG00000024371" ""

# Have a look at the results and search for other probesets for your DE genes
head(LIMMAout_annot2)

##          logFC    AveExpr      t     P.Value   adj.P.Val      B
## 1450297_at 5.716651  6.778338 44.60100 1.977975e-17 8.920867e-13 25.91879
## 1449025_at 3.251777 11.762997 37.08774 3.118352e-16 7.032040e-12 24.48160
## 1418293_at 3.318259 10.672902 29.69841 8.518266e-15 1.280608e-10 22.40893
## 1424518_at 4.104013  7.593314 26.49470 4.619705e-14 5.208833e-10 21.21036
## 1418930_at 6.303782  9.190682 24.97055 1.107804e-13 7.137579e-10 20.55617
## 1419697_at 3.816194  7.838359 25.43495 8.441607e-14 7.137579e-10 20.76185
##          diffexpressed diffexpressed_no_BH      gene
## 1450297_at        UP          UP ENSMUSG00000025746
## 1449025_at        UP          UP ENSMUSG00000045303
## 1418293_at        UP          UP ENSMUSG00000045932
## 1424518_at        UP          UP ENSMUSG00000051925
## 1418930_at        UP          UP ENSMUSG00000034855
## 1419697_at        UP          UP ENSMUSG00000029419

LIMMAout_annot2[LIMMAout_annot2$gene==" ENSMUSG00000025746",]

```

```

## [1] logFC          AveExpr          t
## [4] P.Value         adj.P.Val        B
## [7] diffexpressed   diffexpressed_no_BH gene
## <0 rows> (or 0-length row.names)

ensembl <- useEnsembl(biomart = "genes")
#listDatasets(ensembl)
searchDatasets(mart = ensembl, pattern = "musculus")

##                                     dataset      description      version
## 18  bmusculus_gene_ensembl Blue whale genes (mBalMus1.v2) mBalMus1.v2
## 107 mmusculus_gene_ensembl           Mouse genes (GRCm39)    GRCm39

#mmusculus_gene_ensembl

library('biomaRt')
mart <- useMart("ENSEMBL_MART_ENSEMBL")
mart <- useDataset("mmusculus_gene_ensembl", mart)

## Ensembl site unresponsive, trying asia mirror

ensLookup <- gsub("\\.\\[0-9]*$", "", c(LIMMAout_sorted2$gene))

annotLookup <- getBM(
  mart=mart,
  attributes=c("ensembl_transcript_id", "ensembl_gene_id",
  "gene_biotype", "external_gene_name"),
  filter="ensembl_gene_id",
  values=ensLookup,
  uniqueRows=TRUE)

#retrieved external gene names
#unique(annotLookup$external_gene_name) #11449 genes

#retrieve the list of gene names from the limma output genes
gene_list <- unique(annotLookup$ensembl_gene_id[annotLookup$ensembl_gene_id %in% LIMMAout_sorted2$gene])
length(gene_list)

## [1] 11449

#extract DE genes
DEgeneIDs <- LIMMAout_sorted2$gene[LIMMAout_sorted2$adj.P.Val <= 0.05]
DEgeneIDs <- DEgeneIDs[DEgeneIDs != ""]
DEgeneIDs <- DEgeneIDs[DEgeneIDs %in% unique(annotLookup$ensembl_gene_id)]
length(DEgeneIDs)

## [1] 1382

```

```

#get gene symbols from ENSEMBL gene ids
DEgene_symbols2 <- NULL
for (gene in DEgeneIDs){
  n <- which(gene == gene_list)
  DEgene_symbols2 <- c(DEgene_symbols2, unique(annotLookup$external_gene_name)[n])
}

#top 10 DE genes
DEgene_symbol_and_ID <- cbind(DEgene_symbols2, DEgeneIDs)
head(DEgene_symbol_and_ID , 10)

##          DEgene_symbols2 DEgeneIDs
## [1,] "Il6"           "ENSMUSG00000025746"
## [2,] "Ifit2"         "ENSMUSG00000045932"
## [3,] "Cxcl10"        "ENSMUSG00000034855"
## [4,] "Ajm1"          "ENSMUSG00000029419"
## [5,] "Mx2"           "ENSMUSG00000023341"
## [6,] "Oasl1"          "ENSMUSG00000041827"
## [7,] "Apod"          "ENSMUSG00000022548"
## [8,] "Usp18"          "ENSMUSG00000030107"
## [9,] "Sectm1a"        "ENSMUSG00000025165"
## [10,] "Cmpk2"         "ENSMUSG00000020638"

# transpose the data before PCA as this function requires the variables to be columns
data <- t(as.data.frame(miceRMA@assayData$exprs))
pca <- prcomp(data, center = T, scale. = T)

summary(pca)

## Importance of components:
##                PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation 109.5792 94.5875 70.5787 58.29125 49.1200 45.97059
## Proportion of Variance 0.2662 0.1984 0.1105 0.07534 0.0535 0.04686
## Cumulative Proportion 0.2662 0.4646 0.5751 0.65040 0.7039 0.75075
##                  PC7      PC8      PC9      PC10     PC11     PC12
## Standard deviation 42.54130 41.32187 37.95470 35.78304 34.98692 32.96254
## Proportion of Variance 0.04013 0.03786 0.03194 0.02839 0.02714 0.02409
## Cumulative Proportion 0.79088 0.82874 0.86068 0.88907 0.91621 0.94030
##                  PC13     PC14     PC15     PC16
## Standard deviation 31.05439 30.60996 28.12686 4.935e-13
## Proportion of Variance 0.02138 0.02077 0.01754 0.000e+00
## Cumulative Proportion 0.96168 0.98246 1.00000 1.000e+00

# save as dataframe and add treatment variable
pca_out <- as.data.frame(pca$x)

pca_out$condition <- as.character(miceRMA@phenoData$data$condition)

# get labels
percentage <- round(pca$sdev / sum(pca$sdev) * 100, 2)
percentage <- paste( colnames(pca_out), "(", paste( as.character(percentage), "%", ")\"", sep="" ) )

```

```

ggplot(data = pca_out)+  

  ggtitle("DBA/2J vs C57/BL6") +  

  geom_point(aes(x = PC1, y = PC2, colour = condition, label='', size=strain)) +  

  geom_text(aes(x = PC1, y = PC2, colour = condition, label=''), hjust=0.5, vjust=1.15) +  

  theme_bw() +  

  xlab(percentage[1]) +  

  ylab(percentage[2]) +  

  labs(colour = "condition") +  

  theme(plot.title = element_text(hjust = 0.5)) +  

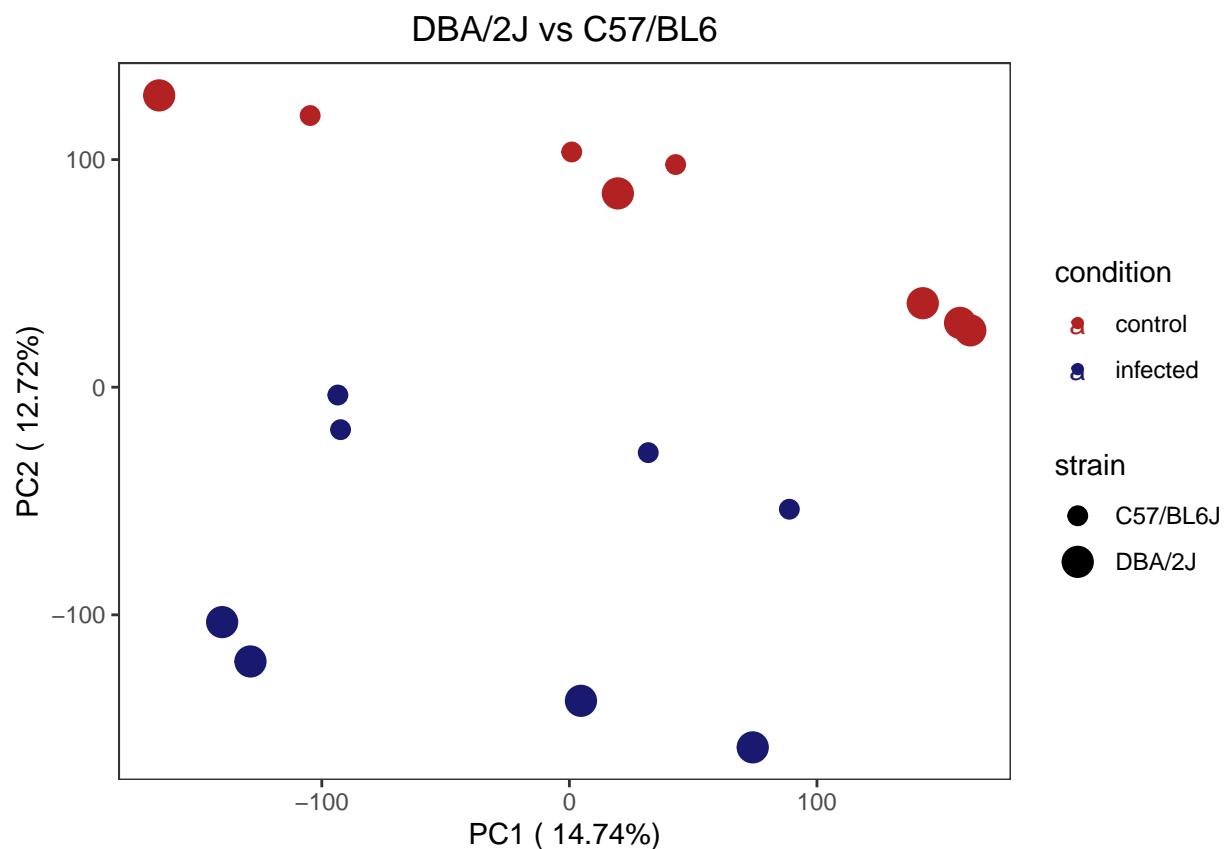
  scale_size_manual(values = c(3, 5)) +  

  scale_colour_manual(values = c("firebrick", "midnightblue")) +  

  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

```

## Warning: Ignoring unknown aesthetics: label



```
#ggsave("PCA_array2_bothstrains.png", dpi=750, width=8, height = 5)
```

## RNAseq dataset E-MTAB-5337

RNA-seq of whole lungs from Irgm1-/- and wildtype littermates mice that were either uninfected or infected with influenza for 10 days. Here, we only used wild type 3 infected vs 3 non-infected samples. For each sample, two technical replicates were present (total 12)



```

## ENSMUSG00000000037.16 2808.0003 2807.9994 3651.758 2807.9998 3442.499 3510.001
## ENSMUSG00000000049.11 443.4987 497.3788 318.000 635.9999 682.000 318.000
## ENSMUSG00000000056.7 2890.5020 4067.4717 3477.942 3140.7988 3472.230 2859.385

```

118489 features 12 samples in transcript level.36047 features in gene level.

```

sdrg <- read.delim("./RNAseq/E-MTAB-5337.sdrf.txt")
head(sdrg)

```

```

##           Source.Name Comment.ENA_SAMPLE. Comment.BioSD_SAMPLE.
## 1 Infected_Irgm1/-_Day10_1      ERS1471225      SAMEA24832918
## 2 Infected_Irgm1/-_Day10_1      ERS1471225      SAMEA24832918
## 3 Infected_Irgm1/-_Day10_2      ERS1471226      SAMEA24833668
## 4 Infected_Irgm1/-_Day10_2      ERS1471226      SAMEA24833668
## 5 Infected_Irgm1/-_Day10_3      ERS1471227      SAMEA24834418
## 6 Infected_Irgm1/-_Day10_3      ERS1471227      SAMEA24834418
##   Characteristics.organism. Characteristics.strain. Characteristics.age.
## 1             Mus musculus          C57BL/6       8 to 12
## 2             Mus musculus          C57BL/6       8 to 12
## 3             Mus musculus          C57BL/6       8 to 12
## 4             Mus musculus          C57BL/6       8 to 12
## 5             Mus musculus          C57BL/6       8 to 12
## 6             Mus musculus          C57BL/6       8 to 12
##   Unit.time.unit. Characteristics.genotype.
## 1        week     Irgm1/- knockout
## 2        week     Irgm1/- knockout
## 3        week     Irgm1/- knockout
## 4        week     Irgm1/- knockout
## 5        week     Irgm1/- knockout
## 6        week     Irgm1/- knockout
##   Characteristics.phenotype. Characteristics.organism.part.
## 1 protected from influenza-induced mortality                  lung
## 2 protected from influenza-induced mortality                  lung
## 3 protected from influenza-induced mortality                  lung
## 4 protected from influenza-induced mortality                  lung
## 5 protected from influenza-induced mortality                  lung
## 6 protected from influenza-induced mortality                  lung
##   Material.Type Protocol.REF    Performer Protocol.REF.1 Performer.1
## 1 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
## 2 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
## 3 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
## 4 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
## 5 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
## 6 organism part P-MTAB-53254 Ashley Steed    P-MTAB-53248 Ashley Steed
##   Protocol.REF.2 Performer.2 Protocol.REF.3
## 1    P-MTAB-53249 Ashley Steed    P-MTAB-53250
## 2    P-MTAB-53249 Ashley Steed    P-MTAB-53250
## 3    P-MTAB-53249 Ashley Steed    P-MTAB-53250
## 4    P-MTAB-53249 Ashley Steed    P-MTAB-53250
## 5    P-MTAB-53249 Ashley Steed    P-MTAB-53250
## 6    P-MTAB-53249 Ashley Steed    P-MTAB-53250
##           Performer.3
## 1 Genome Technology Access Center at Washington University

```

```

## 2 Genome Technology Access Center at Washington University
## 3 Genome Technology Access Center at Washington University
## 4 Genome Technology Access Center at Washington University
## 5 Genome Technology Access Center at Washington University
## 6 Genome Technology Access Center at Washington University
##           Extract.Name Comment.LIBRARY_LAYOUT. Comment.LIBRARY_SELECTION.
## 1 Infected_Irgm1/-_Day10_1             SINGLE          Inverse rRNA
## 2 Infected_Irgm1/-_Day10_1             SINGLE          Inverse rRNA
## 3 Infected_Irgm1/-_Day10_2             SINGLE          Inverse rRNA
## 4 Infected_Irgm1/-_Day10_2             SINGLE          Inverse rRNA
## 5 Infected_Irgm1/-_Day10_3             SINGLE          Inverse rRNA
## 6 Infected_Irgm1/-_Day10_3             SINGLE          Inverse rRNA
##           Comment.LIBRARY_SOURCE. Comment.LIBRARY_STRAND. Comment.LIBRARY_STRATEGY.
## 1      TRANSCRIPTOMIC    not applicable        RNA-Seq
## 2      TRANSCRIPTOMIC    not applicable        RNA-Seq
## 3      TRANSCRIPTOMIC    not applicable        RNA-Seq
## 4      TRANSCRIPTOMIC    not applicable        RNA-Seq
## 5      TRANSCRIPTOMIC    not applicable        RNA-Seq
## 6      TRANSCRIPTOMIC    not applicable        RNA-Seq
##           Protocol.REF.4                    Performer.4
## 1 P-MTAB-53251 Genome Technology Access Center at Washington University
## 2 P-MTAB-53251 Genome Technology Access Center at Washington University
## 3 P-MTAB-53251 Genome Technology Access Center at Washington University
## 4 P-MTAB-53251 Genome Technology Access Center at Washington University
## 5 P-MTAB-53251 Genome Technology Access Center at Washington University
## 6 P-MTAB-53251 Genome Technology Access Center at Washington University
##           Assay.Name Comment.technical.replicate.group.
## 1 Infected_Irgm1/-_Day10_1             group 24
## 2 Infected_Irgm1/-_Day10_1_1          group 24
## 3 Infected_Irgm1/-_Day10_2             group 23
## 4 Infected_Irgm1/-_Day10_2_1          group 23
## 5 Infected_Irgm1/-_Day10_3             group 22
## 6 Infected_Irgm1/-_Day10_3_1          group 22
##           Technology.Type Comment.ENA_EXPERIMENT.
## 1 sequencing assay                  ERX1820995
## 2 sequencing assay                  ERX1820995
## 3 sequencing assay                  ERX1820996
## 4 sequencing assay                  ERX1820996
## 5 sequencing assay                  ERX1820997
## 6 sequencing assay                  ERX1820997
##           Scan.Name
## 1 run_1653_s_3_withindex_sequence.txt_CGAAAGT.fq.gz
## 2 run_1653_s_4_withindex_sequence.txt_CGAAAGT.fq.gz
## 3 run_1653_s_3_withindex_sequence.txt_CAGAGTC.fq.gz
## 4 run_1653_s_4_withindex_sequence.txt_CAGAGTC.fq.gz
## 5 run_1653_s_3_withindex_sequence.txt_ACTGGAT.fq.gz
## 6 run_1653_s_4_withindex_sequence.txt_ACTGGAT.fq.gz
##           Comment.SUBMITTED_FILE_NAME. Comment.ENA_RUN.
## 1 run_1653_s_3_withindex_sequence.txt_CGAAAGT.fq.gz      ERR1753604
## 2 run_1653_s_4_withindex_sequence.txt_CGAAAGT.fq.gz      ERR1753605
## 3 run_1653_s_3_withindex_sequence.txt_CAGAGTC.fq.gz      ERR1753606
## 4 run_1653_s_4_withindex_sequence.txt_CAGAGTC.fq.gz      ERR1753607
## 5 run_1653_s_3_withindex_sequence.txt_ACTGGAT.fq.gz      ERR1753608
## 6 run_1653_s_4_withindex_sequence.txt_ACTGGAT.fq.gz      ERR1753609

```

```

## Comment.FASTQ.URI.
## 1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/004/ERR1753604/ERR1753604.fastq.gz
## 2 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/005/ERR1753605/ERR1753605.fastq.gz
## 3 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/006/ERR1753606/ERR1753606.fastq.gz
## 4 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/007/ERR1753607/ERR1753607.fastq.gz
## 5 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/008/ERR1753608/ERR1753608.fastq.gz
## 6 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR175/009/ERR1753609/ERR1753609.fastq.gz
## Protocol.REF.5 Performer.5
## 1 P-MTAB-53252 Genome Technology Access Center at Washington University
## 2 P-MTAB-53252 Genome Technology Access Center at Washington University
## 3 P-MTAB-53252 Genome Technology Access Center at Washington University
## 4 P-MTAB-53252 Genome Technology Access Center at Washington University
## 5 P-MTAB-53252 Genome Technology Access Center at Washington University
## 6 P-MTAB-53252 Genome Technology Access Center at Washington University
## Protocol.REF.6 Performer.6
## 1 P-MTAB-53253 Genome Technology Access Center at Washington University
## 2 P-MTAB-53253 Genome Technology Access Center at Washington University
## 3 P-MTAB-53253 Genome Technology Access Center at Washington University
## 4 P-MTAB-53253 Genome Technology Access Center at Washington University
## 5 P-MTAB-53253 Genome Technology Access Center at Washington University
## 6 P-MTAB-53253 Genome Technology Access Center at Washington University
## Derived.Array.Data.File
## 1 all.gene_counts_D0-and-D10-samples.txt
## 2 all.gene_counts_D0-and-D10-samples.txt
## 3 all.gene_counts_D0-and-D10-samples.txt
## 4 all.gene_counts_D0-and-D10-samples.txt
## 5 all.gene_counts_D0-and-D10-samples.txt
## 6 all.gene_counts_D0-and-D10-samples.txt
## Comment..Derived.ArrayExpress.FTI
## 1 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 2 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 3 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 4 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 5 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 6 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## Protocol.REF.7 Performer.7
## 1 P-MTAB-53252 Genome Technology Access Center at Washington University
## 2 P-MTAB-53252 Genome Technology Access Center at Washington University
## 3 P-MTAB-53252 Genome Technology Access Center at Washington University
## 4 P-MTAB-53252 Genome Technology Access Center at Washington University
## 5 P-MTAB-53252 Genome Technology Access Center at Washington University
## 6 P-MTAB-53252 Genome Technology Access Center at Washington University
## Protocol.REF.8 Performer.8
## 1 P-MTAB-53253 Genome Technology Access Center at Washington University
## 2 P-MTAB-53253 Genome Technology Access Center at Washington University
## 3 P-MTAB-53253 Genome Technology Access Center at Washington University
## 4 P-MTAB-53253 Genome Technology Access Center at Washington University
## 5 P-MTAB-53253 Genome Technology Access Center at Washington University
## 6 P-MTAB-53253 Genome Technology Access Center at Washington University
## Derived.Array.Data.File.1
## 1 all.transcript_counts_D0-and-D10-samples.txt
## 2 all.transcript_counts_D0-and-D10-samples.txt
## 3 all.transcript_counts_D0-and-D10-samples.txt
## 4 all.transcript_counts_D0-and-D10-samples.txt

```

```

## 5 all.transcript_counts_D0-and-D10-samples.txt
## 6 all.transcript_counts_D0-and-D10-samples.txt
##
## 1 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 2 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 3 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 4 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 5 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 6 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
##   Protocol.REF.9                               Comment..Derived.ArrayExpress.FTP...
## 1 P-MTAB-53252 Genome Technology Access Center at Washington University
## 2 P-MTAB-53252 Genome Technology Access Center at Washington University
## 3 P-MTAB-53252 Genome Technology Access Center at Washington University
## 4 P-MTAB-53252 Genome Technology Access Center at Washington University
## 5 P-MTAB-53252 Genome Technology Access Center at Washington University
## 6 P-MTAB-53252 Genome Technology Access Center at Washington University
##   Protocol.REF.10                             Performer.9
## 1 P-MTAB-53253 Genome Technology Access Center at Washington University
## 2 P-MTAB-53253 Genome Technology Access Center at Washington University
## 3 P-MTAB-53253 Genome Technology Access Center at Washington University
## 4 P-MTAB-53253 Genome Technology Access Center at Washington University
## 5 P-MTAB-53253 Genome Technology Access Center at Washington University
## 6 P-MTAB-53253 Genome Technology Access Center at Washington University
##   Derived.Array.Data.File.2
## 1 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
## 2 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
## 3 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
## 4 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
## 5 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
## 6 all.gene_moderated_log2cpm_D0-and-D10-samples.txt
##   Comment..Derived.ArrayExpress.FTP...
## 1 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 2 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 3 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 4 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 5 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
## 6 ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/MTAB/E-MTAB-5337/E-MTAB-5337.processed
##   Factor.Value.genotype. Factor.Value.infect. Factor.Value.sampling.time.point.
## 1 Irgm1/- knockout influenza A (H1N1)          10
## 2 Irgm1/- knockout influenza A (H1N1)          10
## 3 Irgm1/- knockout influenza A (H1N1)          10
## 4 Irgm1/- knockout influenza A (H1N1)          10
## 5 Irgm1/- knockout influenza A (H1N1)          10
## 6 Irgm1/- knockout influenza A (H1N1)          10
##   Unit.time.unit..1
## 1 day
## 2 day
## 3 day
## 4 day
## 5 day
## 6 day

```

```
print(sdrf[,c("Source.Name", "Technology.Type", "Characteristics.phenotype.", "Characteristics.organism.")]
```

```

##           Source.Name Technology.Type
## 1 Infected_Irgm1-/-'_Day10_1 sequencing assay
## 2 Infected_Irgm1-/-'_Day10_1 sequencing assay
## 3 Infected_Irgm1-/-'_Day10_2 sequencing assay
## 4 Infected_Irgm1-/-'_Day10_2 sequencing assay
## 5 Infected_Irgm1-/-'_Day10_3 sequencing assay
## 6 Infected_Irgm1-/-'_Day10_3 sequencing assay
## 7 Uninfected_Irgm1-/-'_Day0_1 sequencing assay
## 8 Uninfected_Irgm1-/-'_Day0_1 sequencing assay
## 9 Uninfected_Irgm1-/-'_Day0_2 sequencing assay
## 10 Uninfected_Irgm1-/-'_Day0_2 sequencing assay
## 11 Uninfected_Irgm1-/-'_Day0_3 sequencing assay
## 12 Uninfected_Irgm1-/-'_Day0_3 sequencing assay
## 13 Uninfected_WT_Day0_1 sequencing assay
## 14 Uninfected_WT_Day0_1 sequencing assay
## 15 Uninfected_WT_Day0_2 sequencing assay
## 16 Uninfected_WT_Day0_2 sequencing assay
## 17 Uninfected_WT_Day0_3 sequencing assay
## 18 Uninfected_WT_Day0_3 sequencing assay
## 19 Infected_WT_Day10_1 sequencing assay
## 20 Infected_WT_Day10_1 sequencing assay
## 21 Infected_WT_Day10_2 sequencing assay
## 22 Infected_WT_Day10_2 sequencing assay
## 23 Infected_WT_Day10_3 sequencing assay
## 24 Infected_WT_Day10_3 sequencing assay
## 25 Infected_Irgm1-/-'_Day3_1 sequencing assay
## 26 Infected_Irgm1-/-'_Day3_1 sequencing assay
## 27 Infected_Irgm1-/-'_Day3_2 sequencing assay
## 28 Infected_Irgm1-/-'_Day3_2 sequencing assay
## 29 Infected_Irgm1-/-'_Day3_3 sequencing assay
## 30 Infected_Irgm1-/-'_Day3_3 sequencing assay
## 31 Infected_Irgm1-/-'_Day6_1 sequencing assay
## 32 Infected_Irgm1-/-'_Day6_1 sequencing assay
## 33 Infected_Irgm1-/-'_Day6_2 sequencing assay
## 34 Infected_Irgm1-/-'_Day6_2 sequencing assay
## 35 Infected_Irgm1-/-'_Day6_3 sequencing assay
## 36 Infected_Irgm1-/-'_Day6_3 sequencing assay
## 37 Infected_WT_Day3_1 sequencing assay
## 38 Infected_WT_Day3_1 sequencing assay
## 39 Infected_WT_Day3_2 sequencing assay
## 40 Infected_WT_Day3_2 sequencing assay
## 41 Infected_WT_Day3_3 sequencing assay
## 42 Infected_WT_Day3_3 sequencing assay
## 43 Infected_WT_Day6_1 sequencing assay
## 44 Infected_WT_Day6_1 sequencing assay
## 45 Infected_WT_Day6_2 sequencing assay
## 46 Infected_WT_Day6_2 sequencing assay
## 47 Infected_WT_Day6_3 sequencing assay
## 48 Infected_WT_Day6_3 sequencing assay
##
##           Characteristics.phenotype Characteristics.organism.part.
## 1 protected from influenza-induced mortality lung
## 2 protected from influenza-induced mortality lung
## 3 protected from influenza-induced mortality lung
## 4 protected from influenza-induced mortality lung

```

```

## 5 protected from influenza-induced mortality          lung
## 6 protected from influenza-induced mortality          lung
## 7 protected from influenza-induced mortality          lung
## 8 protected from influenza-induced mortality          lung
## 9 protected from influenza-induced mortality          lung
## 10 protected from influenza-induced mortality         lung
## 11 protected from influenza-induced mortality         lung
## 12 protected from influenza-induced mortality         lung
## 13 susceptible to influenza-induced mortality        lung
## 14 susceptible to influenza-induced mortality        lung
## 15 susceptible to influenza-induced mortality        lung
## 16 susceptible to influenza-induced mortality        lung
## 17 susceptible to influenza-induced mortality        lung
## 18 susceptible to influenza-induced mortality        lung
## 19 susceptible to influenza-induced mortality        lung
## 20 susceptible to influenza-induced mortality        lung
## 21 susceptible to influenza-induced mortality        lung
## 22 susceptible to influenza-induced mortality        lung
## 23 susceptible to influenza-induced mortality        lung
## 24 susceptible to influenza-induced mortality        lung
## 25 protected from influenza-induced mortality        lung
## 26 protected from influenza-induced mortality        lung
## 27 protected from influenza-induced mortality        lung
## 28 protected from influenza-induced mortality        lung
## 29 protected from influenza-induced mortality        lung
## 30 protected from influenza-induced mortality        lung
## 31 protected from influenza-induced mortality        lung
## 32 protected from influenza-induced mortality        lung
## 33 protected from influenza-induced mortality        lung
## 34 protected from influenza-induced mortality        lung
## 35 protected from influenza-induced mortality        lung
## 36 protected from influenza-induced mortality        lung
## 37 susceptible to influenza-induced mortality       lung
## 38 susceptible to influenza-induced mortality       lung
## 39 susceptible to influenza-induced mortality       lung
## 40 susceptible to influenza-induced mortality       lung
## 41 susceptible to influenza-induced mortality       lung
## 42 susceptible to influenza-induced mortality       lung
## 43 susceptible to influenza-induced mortality       lung
## 44 susceptible to influenza-induced mortality       lung
## 45 susceptible to influenza-induced mortality       lung
## 46 susceptible to influenza-induced mortality       lung
## 47 susceptible to influenza-induced mortality       lung
## 48 susceptible to influenza-induced mortality       lung

```

```

#get annotation data
mart = useMart(host="useast.ensembl.org",
               biomart="ENSEMBL_MART_ENSEMBL",
               dataset="mmusculus_gene_ensembl")

mmusculus <- getBM(attributes=c('ensembl_transcript_id',
                                 'ensembl_gene_id',
                                 'external_gene_name'),
                     mart = mart)

```

```

head(mmusculus)

##   ensembl_transcript_id   ensembl_gene_id external_gene_name
## 1   ENSMUST00000082387 ENSMUSG00000064336          mt-Tf
## 2   ENSMUST00000082388 ENSMUSG00000064337          mt-Rnr1
## 3   ENSMUST00000082389 ENSMUSG00000064338          mt-Tv
## 4   ENSMUST00000082390 ENSMUSG00000064339          mt-Rnr2
## 5   ENSMUST00000082391 ENSMUSG00000064340          mt-Tl1
## 6   ENSMUST00000082392 ENSMUSG00000064341          mt-Nd1

# What are the available attributes
atr <- listAttributes(mart)

data <- getBM(attributes = c('ensembl_gene_id', 'ensembl_transcript_id',
                           'external_gene_name'),
              mart = mart)

```

### Which samples we are using, and not using: We will only be using samples involving susceptible (Wild Type) after 10 days of infection with influenza.

## Check for duplicate rows

```

# no duplicate rows
sum(duplicated(rownames(txi.kallisto$counts)))

## [1] 0

setwd("./RNAseq_output")

## Make tpm values compatible with edgeR
cts <- txi.kallisto$counts
normMat <- txi.kallisto$length
head(normMat)

##           sample1 sample2 sample3 sample4 sample5 sample6
## ENSMUSG00000000001.4 3213.0000 3213.000 3213.0000 3213.000 3213.0000
## ENSMUSG00000000003.15 750.5000 750.500 750.5000 750.5000 750.5000 750.5000
## ENSMUSG00000000028.15 1495.8714 1480.561 1462.6724 1460.3967 1465.040 1479.9446
## ENSMUSG00000000037.16 1638.0000 2340.002 2573.9997 3276.0010 3602.508 3391.0000
## ENSMUSG00000000049.11 443.4987 433.752 443.4987 443.4987 318.000 497.3792
## ENSMUSG00000000056.7 3443.1809 3146.751 2284.8596 2217.8745 2907.992 2752.3972
##           sample7 sample8 sample9 sample10 sample11 sample12
## ENSMUSG00000000001.4 3213.0000 3213.0000 3213.0000 3213.000 3213.0000
## ENSMUSG00000000003.15 750.5000 750.5000 750.5000 750.5000 750.5000 750.5000
## ENSMUSG00000000028.15 1540.7317 1677.0540 1743.725 1699.3468 1683.682 1591.334
## ENSMUSG00000000037.16 2808.0003 2807.9994 3651.758 2807.9998 3442.499 3510.001
## ENSMUSG00000000049.11 443.4987 497.3788 318.000 635.9999 682.000 318.000
## ENSMUSG00000000056.7 2890.5020 4067.4717 3477.942 3140.7988 3472.230 2859.385

```

```

# Obtaining per-observation scaling factors for length, adjusted to avoid
# changing the magnitude of the counts.
normMat <- normMat/exp(rowMeans(log(normMat)))
normCts <- cts/normMat
head(normCts)

##          sample1    sample2    sample3    sample4    sample5
## ENSMUSG000000000001.4 686.0000 618.000000 730.0000 679.000000 618.000000
## ENSMUSG000000000003.15 0.0000 0.000000 0.0000 0.000000 0.000000
## ENSMUSG000000000028.15 153.4060 121.642167 231.9512 204.354916 178.065788
## ENSMUSG000000000037.16 12.4827 6.241345 12.4827 6.241352 4.054054
## ENSMUSG000000000049.11 0.0000 2.044942 0.0000 0.000000 2.789300
## ENSMUSG000000000056.7 158.4009 176.174112 222.8767 225.495887 150.201141
##          sample6    sample7    sample8    sample9    sample10
## ENSMUSG000000000001.4 586.000000 603.000000 610.000000 1038.000000 991.000000
## ENSMUSG000000000003.15 0.000000 0.000000 0.000000 0.000000 0.000000
## ENSMUSG000000000028.15 223.130622 106.616330 55.786146 170.16046 181.10422
## ENSMUSG000000000037.16 4.306915 6.241344 12.482701 11.19825 12.48269
## ENSMUSG000000000049.11 1.783343 0.000000 1.783344 1.39465 1.39465
## ENSMUSG000000000056.7 133.225844 202.129337 125.939701 188.81850 203.66169
##          sample11   sample12
## ENSMUSG000000000001.4 584.000000 621.000000
## ENSMUSG000000000003.15 0.000000 0.000000
## ENSMUSG000000000028.15 147.9495762 125.32669
## ENSMUSG000000000037.16 6.7879752 12.48269
## ENSMUSG000000000049.11 0.6502914 1.39465
## ENSMUSG000000000056.7 144.0585049 162.63249

# Computing effective library sizes from scaled counts, to account for
# composition biases between samples.
eff.lib <- calcNormFactors(normCts) * colSums(normCts)
head(eff.lib)

## sample1 sample2 sample3 sample4 sample5 sample6
## 6809954 6625648 7294221 7122479 6301408 6145094

# Combining effective library sizes with the length factors, and calculating
# offsets for a log-link GLM.

#merge every two columns which correspond to each 2 technical replicates
merged_normMat <- NULL
for (x in seq(1,12,2)){
  merged_normMat <- cbind(merged_normMat,normMat[,x]+normMat[,x+1])
}

normMat <- merged_normMat
normMat <- sweep(normMat, 2, eff.lib, "*")

## Warning in sweep(normMat, 2, eff.lib, "*"): STATS is longer than the extent of
## 'dim(x)[MARGIN]'
```

```

normMat <- log(normMat)

new_cts <- NULL
for (c in seq(1,12,2)){
  new_cts <- cbind(new_cts, cts[,c]+cts[,c+1])
}
cts <- new_cts

y <- DGEList(cts)
y <- scaleOffset(y, normMat)
head(y)

## An object of class "DGEList"
## $counts
##           Sample1   Sample2   Sample3   Sample4   Sample5
## ENSMUSG000000000001.4 1304.00000 1409.00000 1204.00000 1213.00000 2029.000000
## ENSMUSG000000000003.15 0.00000 0.00000 0.00000 0.00000 0.000000
## ENSMUSG000000000028.15 262.26604 408.34856 378.5002 165.0942 387.066300
## ENSMUSG000000000037.16 12.00001 18.00001 10.0000 18.0000 26.000006
## ENSMUSG000000000049.11 2.00000 0.00000 4.0000 2.0000 2.999999
## ENSMUSG000000000056.7 364.97800 334.97230 266.6449 363.8936 430.216440
##           Sample6
## ENSMUSG000000000001.4 1205.0000
## ENSMUSG000000000003.15 0.0000
## ENSMUSG000000000028.15 287.2151
## ENSMUSG000000000037.16 23.0000
## ENSMUSG000000000049.11 2.0000
## ENSMUSG000000000056.7 320.3270
##
## $samples
##      group lib.size norm.factors
## Sample1 1 13413997 1
## Sample2 1 14385696 1
## Sample3 1 12362925 1
## Sample4 1 13767983 1
## Sample5 1 18090430 1
## Sample6 1 11097351 1
##
## $offset
##           [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## ENSMUSG000000000001.4 16.44816 16.42072 16.51685 16.49303 16.37055 16.34543
## ENSMUSG000000000003.15 16.42970 16.40437 16.66884 16.65421 16.22957 16.20804
## ENSMUSG000000000028.15 16.39974 16.35421 16.45782 16.52258 16.46777 16.39261
## ENSMUSG000000000037.16 16.04006 16.40039 16.84341 16.60941 16.32474 16.37672
## ENSMUSG000000000049.11 16.41103 16.39464 16.40659 16.52592 16.41729 16.43926
## ENSMUSG000000000056.7 16.51539 16.10921 16.60249 16.79426 16.31963 16.25375

## Library sizes
infection <- factor(c("10dpi","10dpi","10dpi","control", "control", "control"))
lib <- NULL
n <- 1
for (x in seq(1,12,2)){
  lib[n] <- eff.lib[x]+ eff.lib[x+1]
}

```

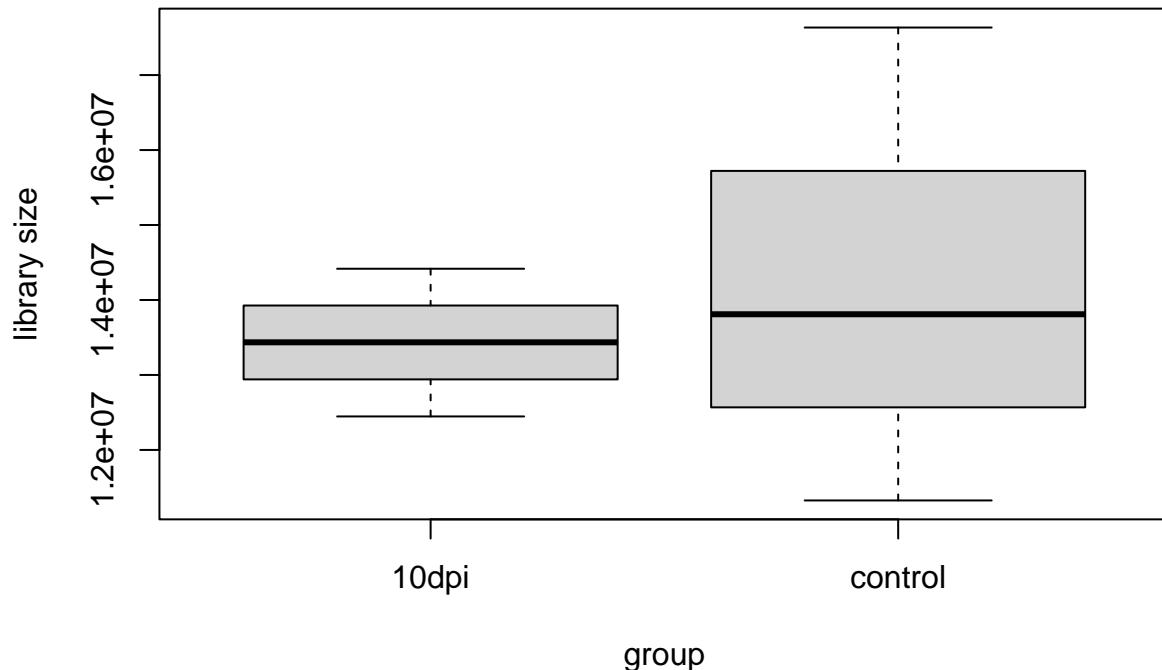
```

n <- n + 1
}
lib

## [1] 13435602 14416700 12446502 13809810 17633891 11326350

#jpeg("library_sizes_musmusculus.jpg")
boxplot(lib~as.factor(infection),xlab="group",ylab="library size")

```



```

#dev.off()

wilcox.test(lib~as.factor(infection))

##
## Wilcoxon rank sum exact test
##
## data: lib by as.factor(infection)
## W = 4, p-value = 1
## alternative hypothesis: true location shift is not equal to 0

## Optionally filter on counts mean
cutoff <- 3/(mean(y$samples$lib.size)/1000000)
keep <- rowSums(cpm(y)>cutoff) >= 3
y <- y[keep, ,keep.lib.sizes=FALSE]
summary(keep) #FALSE 15672 TRUE 20375

```

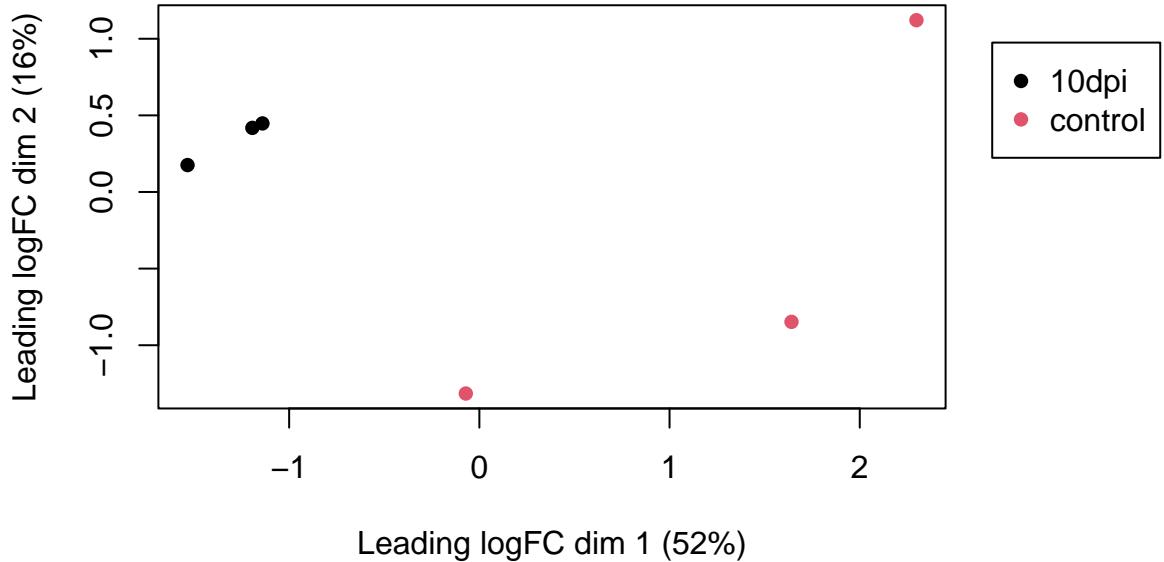
```

##      Mode   FALSE    TRUE
## logical 15672 20375

## MDS plot
#jpeg("plotMDS_mmusculus.jpg")
par(mar=c(6,6,6,6))
plotMDS(y,col=as.numeric(as.factor(infection)), pch=16, main="3 infected vs 3 controls")
par(xpd=T)
legend(par("usr") [2]*1.1,par("usr") [4]*0.8,sort(unique(infection)),
      pch=c(16),col=as.double(as.factor(sort(unique(infection)))))


```

### 3 infected vs 3 controls



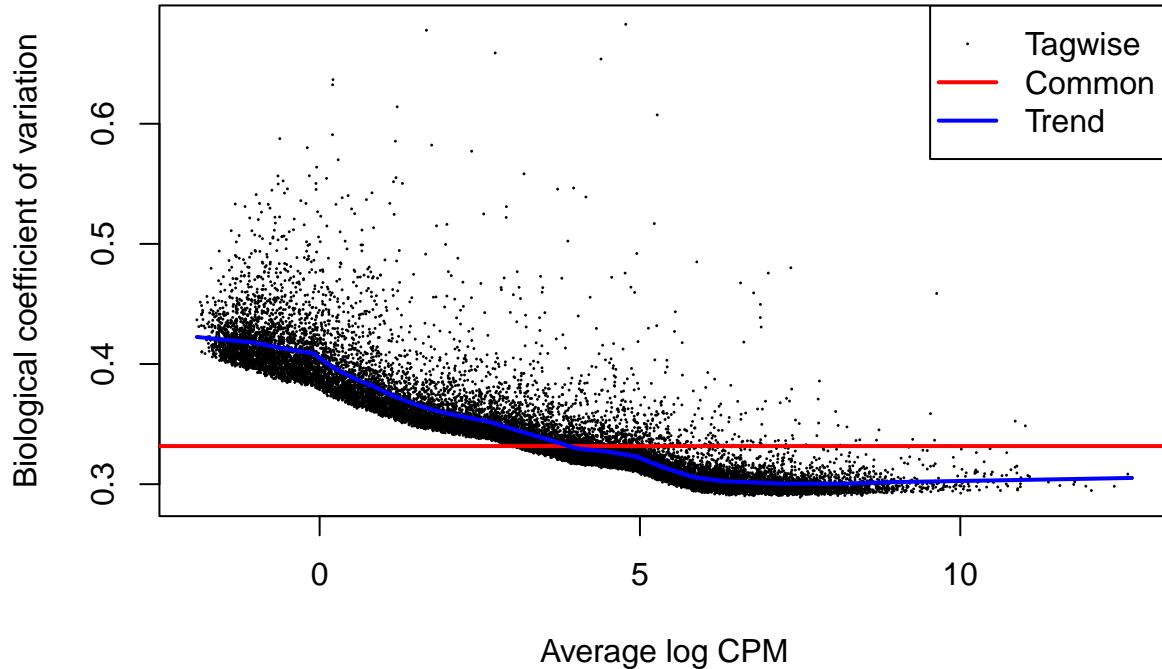
```
#dev.off()
```

```

## Differential expression analysis
design <- model.matrix(~infection)
rownames(design) <- colnames(y)

y <- estimateDisp(y,design)
#jpeg("BCVplot_mmusculus.jpg")
plotBCV(y)

```



```

#dev.off()
#there is a trend!

fit_edgeR <- glmQLFit(y,design)
qlf_edgeR <- glmQLFTest(fit_edgeR,coef=2)
# coef=2 => infected vs control, see "design" object
# note: standard, the last column is considered as the contrast of interest
res_edgeR <- topTags(qlf_edgeR,n=nrow(qlf_edgeR$table))$table

## Add gene symbols
data_sorted_edgeR <- data[sort(data$ensembl_transcript_id,index.return=T)$ix,]
data_sorted_edgeR <- data_sorted_edgeR[duplicated(data_sorted_edgeR$ensembl_gene_id)==F,]

res_edgeR <- cbind(rownames(res_edgeR),res_edgeR)
colnames(res_edgeR)[1] <- "Ensembl_gene_id"

res_edgeR$diffexpressed <- "NO"
res_edgeR$diffexpressed[res_edgeR$logFC > 0 & res_edgeR$FDR < 0.05] <- "UP"
res_edgeR$diffexpressed[res_edgeR$logFC < 0 & res_edgeR$FDR < 0.05] <- "DOWN"

res_edgeR_sorted <- res_edgeR[sort(res_edgeR$Ensembl_gene_id,index.return=T)$ix,]
head(res_edgeR_sorted)

##                                     Ensembl_gene_id      logFC      logCPM          F
## ENSMUSG000000000001.4    ENSMUSG000000000001.4  0.2742049 6.650000 0.5884430

```

```

## ENSMUSG00000000028.15 ENSMUSG00000000028.15 -0.3937051 4.530961 1.0743186
## ENSMUSG00000000037.16 ENSMUSG00000000037.16 0.7040886 0.522503 1.8979925
## ENSMUSG00000000056.7 ENSMUSG00000000056.7 0.1279611 4.663505 0.1152170
## ENSMUSG00000000058.6 ENSMUSG00000000058.6 -0.5199073 7.047578 2.0858873
## ENSMUSG00000000078.7 ENSMUSG00000000078.7 0.1639739 7.977682 0.2053453
## PValue FDR diffexpressed
## ENSMUSG00000000001.4 0.4467346 0.8271373 NO
## ENSMUSG00000000028.15 0.3051047 0.7431571 NO
## ENSMUSG00000000037.16 0.1746180 0.6024115 NO
## ENSMUSG00000000056.7 0.7357444 0.9338949 NO
## ENSMUSG00000000058.6 0.1550858 0.5732253 NO
## ENSMUSG00000000078.7 0.6524583 0.9108295 NO

original <- res_edgeR$Ensembl_gene_id
#substr(original[1], 1, 18)
n <- 1
for (id in original){
  res_edgeR$Ensembl_gene_id[n] <- substr(id, 1, 18)
  n <- n + 1
}
head(res_edgeR$Ensembl_gene_id)

## [1] "ENSMUSG00000076612" "ENSMUSG00000100131" "ENSMUSG00000029417"
## [4] "ENSMUSG00000076613" "ENSMUSG00000094708" "ENSMUSG00000095937"

ids <- res_edgeR_sorted$Ensembl_gene_id
#substr(original[1], 1, 18)
n <- 1
for (id in ids){
  res_edgeR_sorted$Ensembl_gene_id[n] <- substr(id, 1, 18)
  n <- n + 1
}
#gsub
data_sorted_edgeR <- data_sorted_edgeR[data_sorted_edgeR$ensembl_gene_id %in% res_edgeR_sorted$Ensembl_gene_id]
res_edgeR_sorted <- res_edgeR_sorted[res_edgeR_sorted$Ensembl_gene_id %in% data_sorted_edgeR$ensembl_gene_id]

dim(res_edgeR_sorted) #19174

## [1] 20321      7

dim(data_sorted_edgeR) #19174

## [1] 20321      3

#which(res_edgeR_sorted$Ensembl_gene_id == data_sorted_edgeR$ensembl_transcript_id)

res_edgeR_sorted$Gene_symbol <- data_sorted_edgeR$external_gene_name

## Resort and save results
res_edgeR <- res_edgeR_sorted[sort(res_edgeR_sorted$PValue, index.return=T)$ix,]
head(res_edgeR[,c(1,7,2,5,6)],10) #geneID, symbol, logFC, pvalue, FDR

```

```

##                                     Ensembl_gene_id diffexpressed      logFC      PValue
## ENSMUSG00000076612.8   ENSMUSG00000076612          UP  6.251235 3.580705e-17
## ENSMUSG00000100131.1   ENSMUSG00000100131         DOWN -12.409971 6.470052e-17
## ENSMUSG0000029417.9    ENSMUSG0000029417          UP  7.421845 6.322303e-16
## ENSMUSG0000076613.4    ENSMUSG0000076613          UP  5.750282 1.958567e-15
## ENSMUSG0000094708.1    ENSMUSG0000094708         DOWN -9.590055 4.942463e-15
## ENSMUSG0000095937.1    ENSMUSG0000095937         DOWN -9.514044 6.419399e-15
## ENSMUSG0000099875.1    ENSMUSG0000099875          UP  9.619127 1.212535e-14
## ENSMUSG0000031972.5    ENSMUSG0000031972          UP  8.506408 8.774228e-14
## ENSMUSG0000034855.13   ENSMUSG0000034855          UP  6.094653 2.401049e-13
## ENSMUSG0000042385.14   ENSMUSG0000042385          UP  6.964233 2.638372e-13
##                                     FDR
## ENSMUSG0000076612.8    6.591366e-13
## ENSMUSG00000100131.1   6.591366e-13
## ENSMUSG0000029417.9    4.293898e-12
## ENSMUSG0000076613.4    9.976453e-12
## ENSMUSG0000094708.1    2.014054e-11
## ENSMUSG0000095937.1    2.179921e-11
## ENSMUSG0000099875.1    3.529343e-11
## ENSMUSG0000031972.5    2.234686e-10
## ENSMUSG0000034855.13   5.375683e-10
## ENSMUSG0000042385.14   5.375683e-10

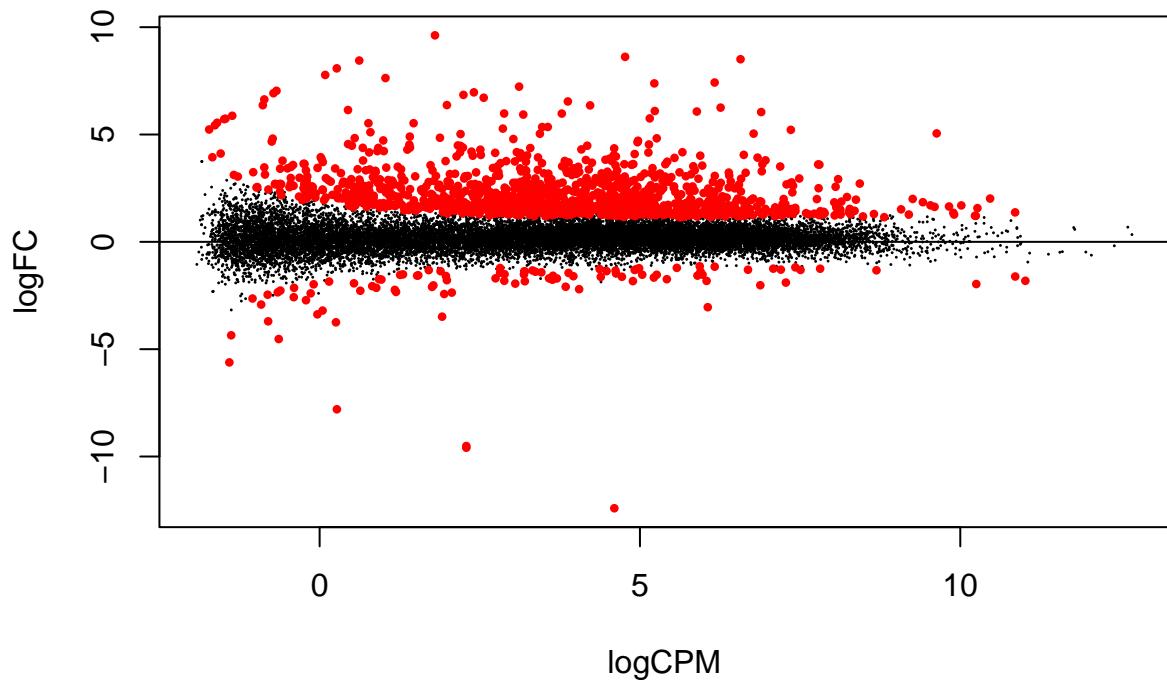
write.table(res_edgeR,file="res_edgeR_mmusculus.txt",col.names=T,row.names=T,sep="\t",quote=F)
# top 1000 loci, also contain non-significant loci (filtered out in next step)
res_edgeR_sign <- res_edgeR[res_edgeR$FDR<0.05,]
dim(res_edgeR_sign)

## [1] 1329     8

# last column of res (ncol(res)) contains FDRs => filtering at 5% level

## MA plot
#jpeg("res_edgeR_mmusculus_MA.png")
## MA-plot
with(res_edgeR,plot(logCPM,logFC,pch=16,cex=0.2), title="")
# MApplot: all data points
with(res_edgeR,points(logCPM[FDR<0.05],logFC[FDR<0.05],pch=16,col="red",cex=0.6))
# MA-plot: significant loci
abline(0,0)

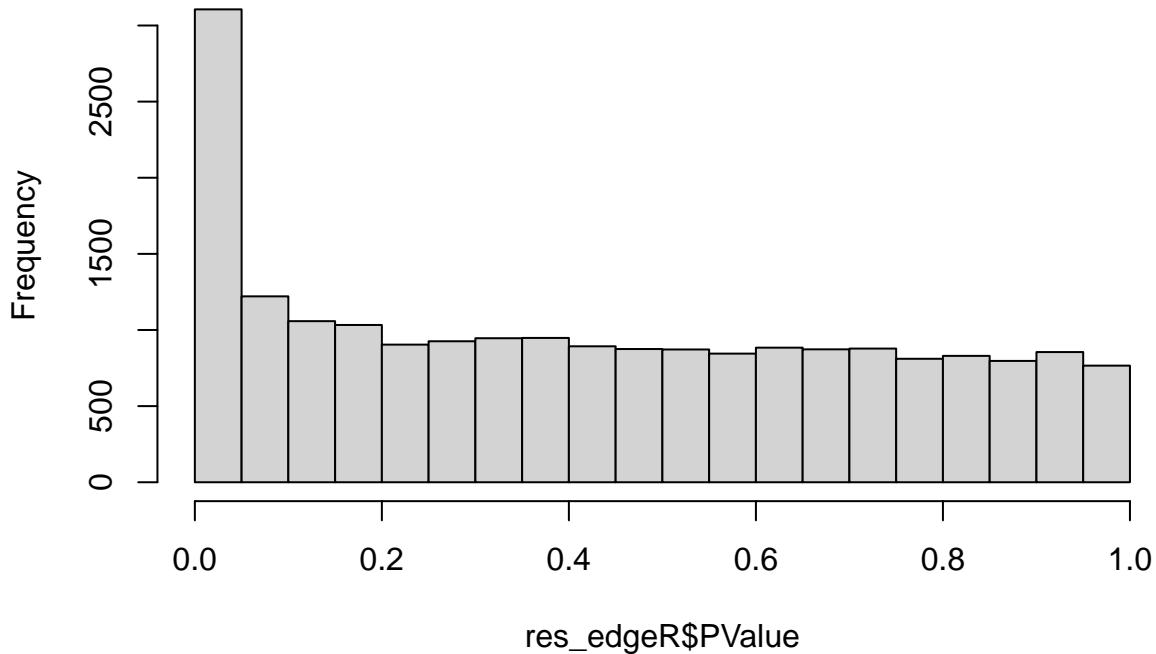
```



```
# X-axis
#dev.off()

## Pvalue distribution
#jpeg("res_edgeR_mmusculus_histogram.jpg")
hist(res_edgeR$PValue)
```

## Histogram of res\_edgeR\$PValue



```
#dev.off()

## Plot boxplots for top 20 loci (first make the folder where you want to put them!)
setwd("./RNAseq_output/Boxplots/")
counts_k <- txi.kallisto$counts[keep,]
for (i in 1:20){
  counts_part <- as.numeric(cpm(y)[rownames(counts_k)==rownames(res_edgeR)[i],])
  dat_boxplot <- data.frame(counts=counts_part,group=infection)
  jpeg(paste(i,"_",rownames(res_edgeR)[i],".jpg",sep=""))
  if (res_edgeR$Gene_symbol[i]!=""){
    boxplot(counts~group,dat_boxplot,main=paste(rownames(res_edgeR)[i], " (",res_edgeR$Gene_symbol[i],")"))
  } else {
    boxplot(counts~group,dat_boxplot,main=paste(rownames(res_edgeR)[i], " (NA)",sep=""))
  }
  dev.off()
}

RNAseq_DEgenes <- res_edgeR_sign$Gene_symbol
head(RNAseq_DEgenes)

## [1] "Thy1"      "Hmgb1-ps8"  "Rhbdd3"     "Uhrf1bp1"   "Gm17235"   "Hspe1-ps3"
```

```

length(RNAseq_DEgenes)

## [1] 1329

length(which(res_edgeR$diffexpressed=="UP")) #99

## [1] 1230

length(which(res_edgeR$diffexpressed=="DOWN")) #1230

## [1] 99

#List all down and upregulated genes for each dataset

#Microarray1
Array1_upregulated <- unique(annotation_MA[annotation_MA$ID %in% rownames(LIMMAout_annot)[LIMMAout_annot$diffexpressed=="UP"]])

Array1_downregulated <- unique(annotation_MA[annotation_MA$ID %in% rownames(LIMMAout_annot)[LIMMAout_annot$diffexpressed=="DOWN"]])

#Microarray2
Array2_up <- unique(LIMMAout_sorted2$gene[LIMMAout_sorted2$logFC > 1 & LIMMAout_sorted2$adj.P.Val <= 0.05])
Array2_upregulated <- NULL
for (gene in Array2_up){
  n <- which(gene == gene_list)
  if (gene != ""){
    Array2_upregulated <- c(Array2_upregulated, unique(annotLookup$external_gene_name)[n])
  }
}

Array2_down <- unique(LIMMAout_sorted2$gene[LIMMAout_sorted2$logFC < -1 & LIMMAout_sorted2$adj.P.Val <= 0.05])
Array2_downregulated <- NULL
for (gene in Array2_down){
  n <- which(gene == gene_list)
  if (gene != ""){
    Array2_downregulated <- c(Array2_downregulated, unique(annotLookup$external_gene_name)[n])
  }
}

#RNAseq
RNA_upregulated <- unique(res_edgeR$Gene_symbol[res_edgeR$diffexpressed=="UP"])
RNA_downregulated <- unique(res_edgeR$Gene_symbol[res_edgeR$diffexpressed=="DOWN"])

head(Array1_upregulated)

## [1] "Dgkd"    "Fcmr"    "Kif21b"   "Parp1"    "Acbd3"    "Bcl2"

length(Array1_upregulated) #102

## [1] 102

```

```

head(Array1_downregulated)

## [1] "Cspp1"   "Cetn4"    "Il1r2"    "Il18r1"  "Ormdl1"  "Nif3l1"

length(Array1_downregulated)      #357

## [1] 357

head(Array2_upregulated)

## [1] "Il6"      "Ifit2"    "Cxcl10"   "Ajm1"    "Mx2"     "Oasl1"

length(Array2_upregulated)

## [1] 334

head(Array2_downregulated)

## [1] "Tex11"   "Per3"    "Nr1d2"   "Sspn"    "Ces1f"   "Tef"

length(unique(Array2_downregulated))

## [1] 40

head(RNA_upregulated)

## [1] "Thy1"     "Rhbdd3"   "Uhrf1bp1" "Gm7389"  "Nbas"    "Mreg"

length(RNA_upregulated)

## [1] 1230

head(RNA_downregulated)

## [1] "Hmgb1-ps8" "Gm17235"  "Hspe1-ps3" "Gm12208"  "Gm18094"  "Fyn"

length(RNA_downregulated)

## [1] 99

library(ggvenn)

## Loading required package: dplyr

```

```

## 
## Attaching package: 'dplyr'

## The following object is masked from 'package:biomaRt':
## 
##     select

## The following object is masked from 'package:lumi':
## 
##     combine

## The following object is masked from 'package:methylumi':
## 
##     combine

## The following object is masked from 'package:minfi':
## 
##     combine

## The following objects are masked from 'package:GenomicRanges':
## 
##     intersect, setdiff, union

## The following object is masked from 'package:matrixStats':
## 
##     count

## The following object is masked from 'package:oligo':
## 
##     summarize

## The following objects are masked from 'package:Biostrings':
## 
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
## 
##     intersect

## The following object is masked from 'package:XVector':
## 
##     slice

## The following object is masked from 'package:AnnotationDbi':
## 
##     select

## The following objects are masked from 'package:IRanges':
## 
##     collapse, desc, intersect, setdiff, slice, union

```

```

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following object is masked from 'package:Biobase':
##
##     combine

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: grid

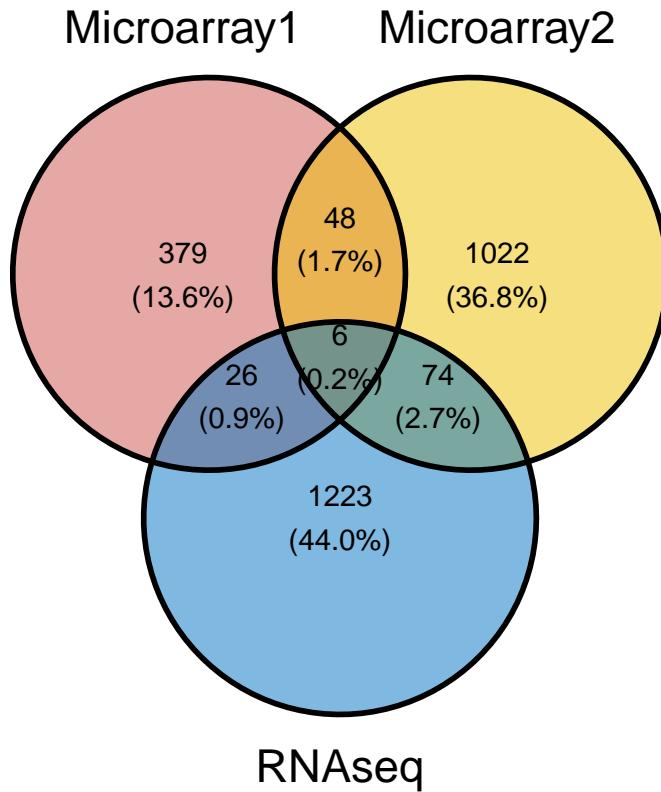
##
## Attaching package: 'grid'

## The following object is masked from 'package:Biostrings':
##
##     pattern

overlap <-list('Microarray1'= DEgenes_symbols1, 'Microarray2'=DEgene_symbols2, 'RNaseq'=RNaseq_DEgenes)

#Create venn diagram and display all the sets
ggvenn(overlap, fill_color = c("#CD534CFF", "#EFC000FF", "#0073C2FF"))
)

```



```
common_DEgenes <- Reduce(intersect, list(DEgenes_symbols1, DEgene_symbols2, RNaseq_DEgenes))
array1_RNA <- Reduce(intersect, list(DEgenes_symbols1, RNaseq_DEgenes))
array1_RNA
```

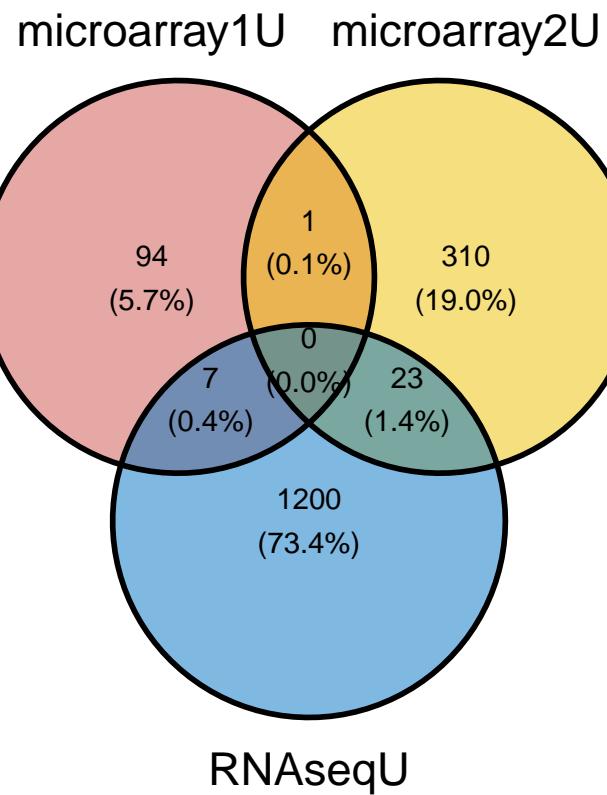
```
## [1] "Il1r2"          "Rpa3"           "Kansl1l"         "Slc25a5"
## [5] "Fuca2"          "Slfn4"           "Zbtb1"           "Commd6"
## [9] "C730034F03Rik" "Prdx1"          "Samsn1"          "Tnfsf9"
## [13] "Gca"             "Lpcat4"          "Il1b"            "Slc16a1"
## [17] "Pitx2"          "Lef1"            "Ssu72"           "Ndufb6"
## [21] "Cux1"            "Gpnmb"          "Tnip3"           "Ccnd2"
## [25] "Cd79a"          "Pde8a"           "Swap70"          "Zpr1"
## [29] "Nt5e"            "Bbs4"            "Dynlt3"          "Gla"
```

```
head(common_DEgenes)
```

```
## [1] "Il1r2"   "Samsn1"  "Il1b"    "Ccnd2"   "Swap70"  "Gla"
```

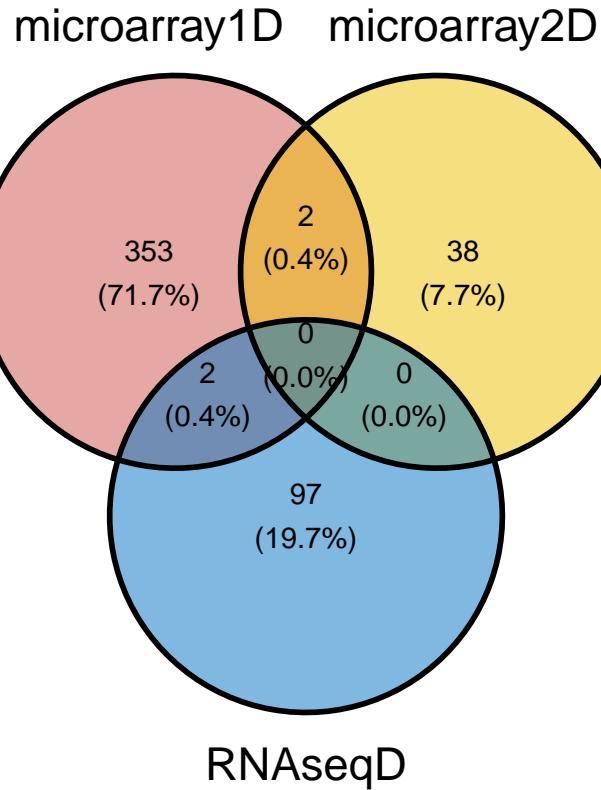
```
UP <- list("microarray1U"=Array1_upregulated, "microarray2U"=Array2_upregulated, "RNaseqU"=RNA_upregulated)
DOWN <- list("microarray1D"= Array1_downregulated, "microarray2D"=Array2_downregulated, "RNaseqD"=RNA_downregulated)
```

```
# upregulated genes for all 3 datasets
ggvenn(UP, fill_color = c("#CD534CFF", "#EFC000FF", "#0073C2FF"))
```



```
common_upreg <- Reduce(intersect, list(Array1_upregulated, Array2_upregulated, RNA_upregulated))

# downregulated genes for all 3 datasets
ggvenn(DOWN, fill_color = c("#CD534CFF", "#EFC000FF", "#0073C2FF"))
```



```

common_downreg <- Reduce(intersect, list(Array1_downregulated, Array2_downregulated, RNA_downregulated))

Array1_RNA_up <- Reduce(intersect, list(Array1_upregulated, RNA_upregulated))
Array1_RNA_up

## [1] "Lpcat4" "Pitx2"   "Lef1"     "Ccnd2"    "Cd79a"   "Swap70"   "Zpr1"

Array1_RNA_down <- Reduce(intersect, list(Array1_downregulated, RNA_downregulated))
Array1_RNA_down

## [1] "Slc25a5" "Prdx1"

# 6 significant genes for all datasets
common_DEgenes

## [1] "Il1r2"   "Samsn1"  "Il1b"     "Ccnd2"    "Swap70"   "Gla"

for (x in common_DEgenes){
  if (x %in% RNAseq_DEgenes){
    print("UP")
  }
  else{
    print("DOWN")
  }
}

```

```
## [1] "UP"

#array1 d,d,d,u,u,d
#array2 u u u d d u
#rna u u u u u u u
```