RNA-seq differential analysis with CummeRbund

Lavinia Carabet

Install the CummeRbund package

Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data

```
#if (!require("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")

#BiocManager::install("cummeRbund")
```

Load the CummeRbund package

```
library(cummeRbund)
```

Create a CummeRbund database from the Cuffdiff output

```
print(readCufflinks)
```

```
## function (dir = getwd(), dbFile = "cuffData.db", gtfFile = NULL,
##     runInfoFile = "run.info", repTableFile = "read_groups.info",
##     geneFPKM = "genes.fpkm_tracking", geneDiff = "gene_exp.diff",
##     geneCount = "genes.count_tracking", geneRep = "genes.read_group_tracking",
##     isoformFPKM = "isoforms.fpkm_tracking", isoformDiff = "isoform_exp.diff",
##     isoformCount = "isoforms.count_tracking", isoformRep = "isoforms.read_group_tracking",
##     TSSFPKM = "tss_groups.fpkm_tracking", TSSDiff = "tss_group_exp.diff",
##     TSSCount = "tss_groups.count_tracking", TSSRep = "tss_groups.read_group_tracking",
##     CDSFPKM = "cds.fpkm_tracking", CDSExpDiff = "cds_exp.diff",
##     CDSCount = "cds.count_tracking", CDSRep = "cds.read_group_tracking",
##     CDSDiff = "cds.diff", promoterFile = "promoters.diff", splicingFile = "splicing.diff",
##     varModelFile = "var_model.info", driver = "SQLite", genome = NULL,
##     rebuild = FALSE, verbose = FALSE, ...)
## {
##     dbFile = file.path(dir, dbFile)
##     runInfoFile = file.path(dir, runInfoFile)
##     repTableFile = file.path(dir, repTableFile)
##     geneFPKM = file.path(dir, geneFPKM)
##     geneDiff = file.path(dir, geneDiff)
##     geneCount = file.path(dir, geneCount)
##     geneRep = file.path(dir, geneRep)
##     isoformFPKM = file.path(dir, isoformFPKM)
```

```
##     isoformDiff = file.path(dir, isoformDiff)
##     isoformCount = file.path(dir, isoformCount)
##     isoformRep = file.path(dir, isoformRep)
##     TSSFPKM = file.path(dir, TSSFPKM)
##     TSSDiff = file.path(dir, TSSDiff)
##     TSSCount = file.path(dir, TSSCount)
##     TSSRep = file.path(dir, TSSRep)
##     CDSFPKM = file.path(dir, CDSFPKM)
##     CDSExpDiff = file.path(dir, CDSExpDiff)
##     CDSCount = file.path(dir, CDSCount)
##     CDSRep = file.path(dir, CDSRep)
##     CDSDiff = file.path(dir, CDSDiff)
##     promoterFile = file.path(dir, promoterFile)
##     splicingFile = file.path(dir, splicingFile)
##     varModelFile = file.path(dir, varModelFile)
##     if (!file.exists(dbFile) || rebuild == TRUE) {
##         write(paste("Creating database ", dbFile, sep = ""),
##             stderr())
##         dbConn <- createDB_noIndex(dbFile)
##         if (file.exists(runInfoFile)) {
##             loadRunInfo(runInfoFile, dbConn)
##         }
##         if (file.exists(repTableFile)) {
##             loadRepTable(repTableFile, dbConn)
##         }
##         if (file.exists(varModelFile)) {
##             loadVarModelTable(varModelFile, dbConn)
##         }
##         if (!is.null(gtfFile)) {
##             if (!is.null(genome)) {
##                 .loadGTF(gtfFile, genome, dbConn)
##             }
##             else {
##                 stop("'genome' cannot be NULL if you are supplying a .gtf file!")
##             }
##         }
##         loadGenes(geneFPKM, geneDiff, promoterFile, countFile = geneCount,
##             replicateFile = geneRep, dbConn)
##         loadIsoforms(isoformFPKM, isoformDiff, isoformCount,
##             isoformRep, dbConn)
##         loadTSS(TSSFPKM, TSSDiff, splicingFile, TSSCount, TSSRep,
##             dbConn)
##         loadCDS(CDSFPKM, CDSExpDiff, CDSDiff, CDSCount, CDSRep,
##             dbConn)
##         write("Indexing Tables...", stderr())
##         createIndices(dbFile, verbose = verbose)
##     }
##     dbConn <- dbConnect(dbDriver(driver), dbFile)
##     return(new("CuffSet", DB = dbConn, genes = new("CuffData",
##         DB = dbConn, tables = list(mainTable = "genes", dataTable = "geneData",
##             expDiffTable = "geneExpDiffData", featureTable = "geneFeatures",
##             countTable = "geneCount", replicateTable = "geneReplicateData"),
##         filters = list(), type = "genes", idField = "gene_id"),
##         isoforms = new("CuffData", DB = dbConn, tables = list(mainTable = "isoforms",
```

```
##             dataTable = "isoformData", expDiffTable = "isoformExpDiffData",
##             featureTable = "isoformFeatures", countTable = "isoformCount",
##             replicateTable = "isoformReplicateData"), filters = list(),
##             type = "isoforms", idField = "isoform_id"), TSS = new("CuffData",
##             DB = dbConn, tables = list(mainTable = "TSS", dataTable = "TSSData",
##                 expDiffTable = "TSSExpDiffData", featureTable = "TSSFeatures",
##                 countTable = "TSSCount", replicateTable = "TSSReplicateData"),
##             filters = list(), type = "TSS", idField = "TSS_group_id"),
##         CDS = new("CuffData", DB = dbConn, tables = list(mainTable = "CDS",
##             dataTable = "CDSData", expDiffTable = "CDSExpDiffData",
##             featureTable = "CDSFeatures", countTable = "CDSCount",
##             replicateTable = "CDSReplicateData"), filters = list(),
##             type = "CDS", idField = "CDS_id"), promoters = new("CuffDist",
##             DB = dbConn, table = "promoterDiffData", type = "promoter",
##             idField = "gene_id"), splicing = new("CuffDist",
##             DB = dbConn, table = "splicingDiffData", type = "splicing",
##             idField = "TSS_group_id"), relCDS = new("CuffDist",
##             DB = dbConn, table = "CDSDiffData", type = "relCDS",
##             idField = "gene_id")))
## }
## <bytecode: 0x000000004ee8a9e0>
## <environment: namespace:cummeRbund>
```

```
cuff_data <- readCufflinks(dir = '../cuffdiff/diff_out/')
cuff_data@genes # or genes(cuff_data)
```

```
## CuffData instance with:
##   37178 features and 2 samples
```

```
getLevels(cuff_data)
```

```
## [1] "T" "U"
```

```
replicates(cuff_data)
```

```
##                                          file sample_name replicate rep_name
## 1 ../tophat/T_SRR5272677_thout/accepted_hits.bam           T         0      T_0
## 2 ../tophat/T_SRR5272678_thout/accepted_hits.bam           T         1      T_1
## 3 ../tophat/T_SRR5272679_thout/accepted_hits.bam           T         2      T_2
## 4 ../tophat/U_SRR5272674_thout/accepted_hits.bam           U         0      U_0
## 5 ../tophat/U_SRR5272675_thout/accepted_hits.bam           U         1      U_1
## 6 ../tophat/U_SRR5272676_thout/accepted_hits.bam           U         2      U_2
##   total_mass norm_mass internal_scale external_scale
## 1   51717000  52101400       1.054330              1
## 2   58200700  52101400       1.146160              1
## 3   41277800  52101400       0.789822              1
## 4   44710200  52101400       0.824518              1
## 5   54362200  52101400       1.009880              1
## 6   67322000  52101400       1.282890              1
```

```
cuff_data
```

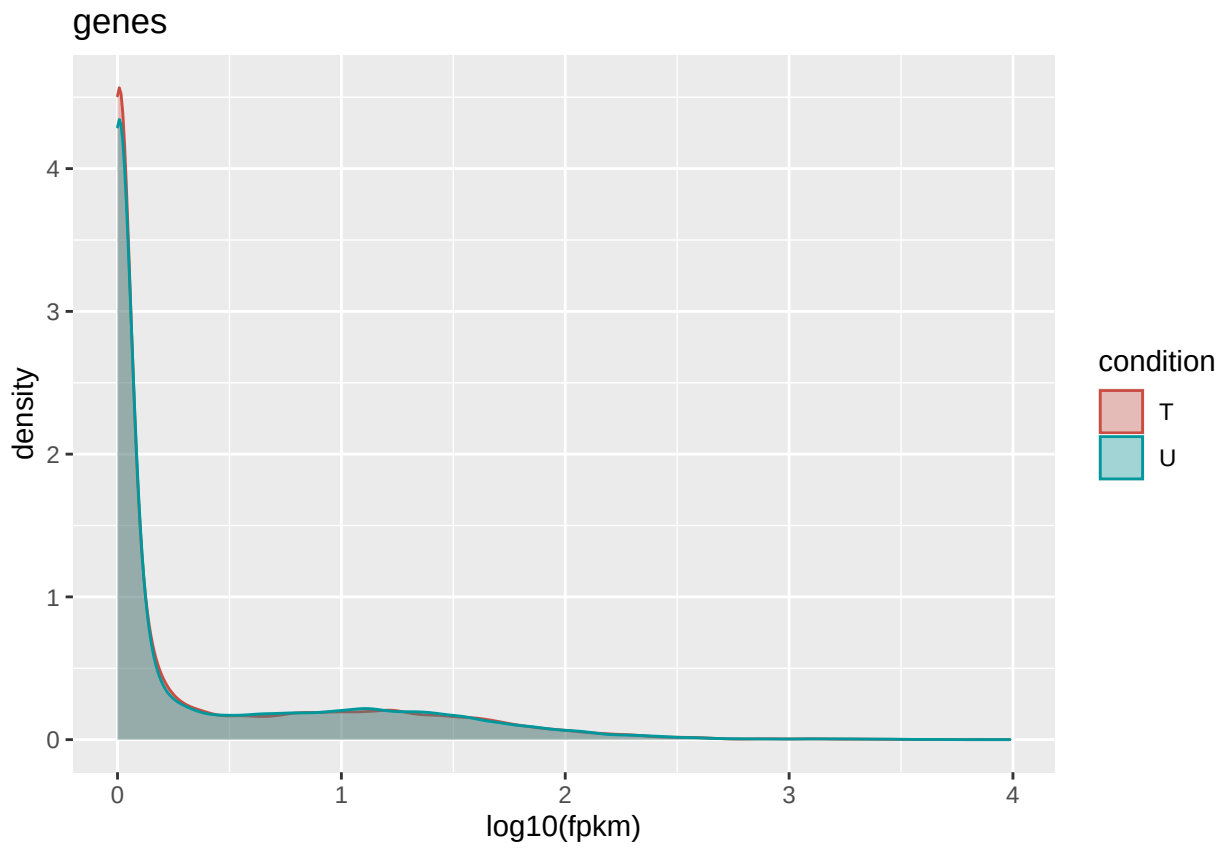
```
## CuffSet instance with:
##   2 samples
##   37178 genes
##   0 isoforms
##   0 TSS
##   0 CDS
##   0 promoters
##   0 splicing
##   0 relCDS
```

Plot the distribution of expression levels for all genes for each sample/condition

T - treated U - untreated

`csDensity` method creates a smoothed density plot, by sample, for log10 FPKM values from a cuffdiff run FPKM - fragments per kilobase of transcript per million fragments mapped

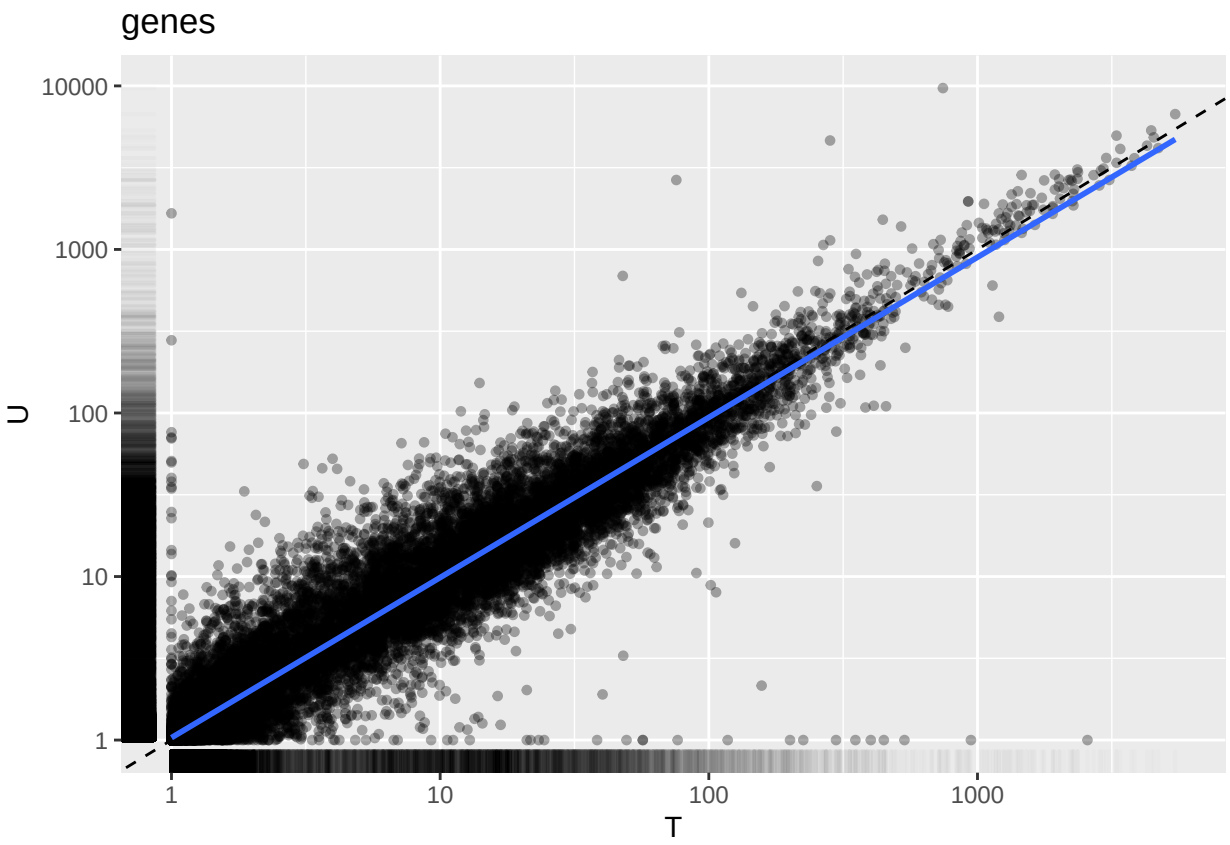
```
csDensity(cuff_data@genes, pseudocount=1.0)
```



Compare the expression of each gene in the two conditions

`csScatter` method creates a scatter plot comparing the FPKM values from the two samples in a cuffdiff run

```
csScatter(cuff_data@genes, 'T', 'U', pseudocount=1.0, colorByStatus=TRUE, smooth=TRUE)
```



Inspect differentially expressed genes

`csVolcano` method creates a volcano plot of log fold change in expression vs significance (-log(pval)) for the two samples in a cuffdiff run

```
gene_diff_data <- diffData(cuff_data@genes, 'T', 'U')
gene_diff_data
```

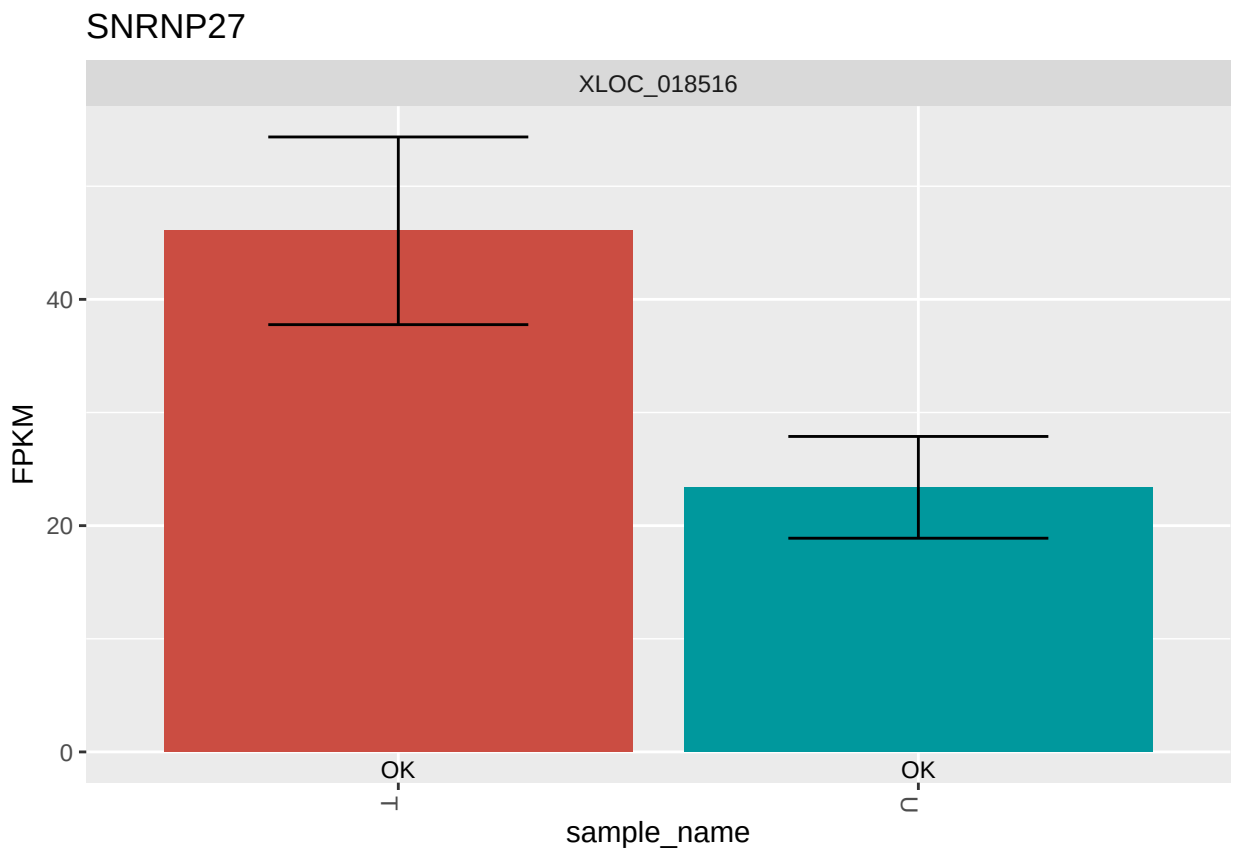
```
##  [1] gene_id          gene_id          sample_1         sample_2
##  [5] status           value_1          value_2          log2_fold_change
##  [9] test_stat        p_value          q_value          significant
## <0 rows> (or 0-length row.names)
```

```
#csVolcano(cuff_data@genes, 'T', 'U', alpha=0.05, showSignificant=T)
# The call to csVolcano fails with
#Error in `$<-.data.frame`(`*tmp*`, "significant", value = "no") :
#replacement has 1 row, data has 0
```

OOPS! No significantly differential expressed genes found?! Figuring out the issue ...

Explore expression levels for some gene

```
mygene <- getGene(cuff_data, 'SNRNP27')
expressionBarplot(mygene)
```



Look at the `gene_exp.diff` and `iso_exp.diff` files generated by cuffdiff

Issue found

```
gene_expr_diff <- read.delim2('../cuffdiff/diff_out/gene_exp.diff')
head(gene_expr_diff)
```

```
##       test_id     gene_id         gene              locus sample_1 sample_2
## 1 XLOC_000001 XLOC_000001      DDX11L1   chr1:11873-29370     TRUE        U
## 2 XLOC_000002 XLOC_000002    MIR1302-2   chr1:30365-30503     TRUE        U
## 3 XLOC_000003 XLOC_000003        OR4F5   chr1:69090-70008     TRUE        U
## 4 XLOC_000004 XLOC_000004 LOC100287934 chr1:764864-810022     TRUE        U
## 5 XLOC_000005 XLOC_000005 LOC100287934 chr1:764864-810022     TRUE        U
## 6 XLOC_000006 XLOC_000006       FAM87B chr1:817370-819834     TRUE        U
##   status   value_1   value_2 log2.fold_change. test_stat p_value    q_value
```

```
## 1 NOTEST  0.186969  0.119911         -0.640841         0       1           1
## 2 NOTEST         0         0                 0         0       1           1
## 3 NOTEST         0         0                 0         0       1           1
## 4     OK   0.37868  0.059133          -2.67894  -1.96996  0.0027 0.00584503
## 5     OK  0.452312  0.572217          0.339244  0.414573  0.4708   0.545048
## 6 NOTEST 0.0173135 0.0384379           1.15063         0       1           1
##   significant
## 1          no
## 2          no
## 3          no
## 4         yes
## 5          no
## 6          no
```

```
# Issue: the 'T' character label is the files is converted behind the scenes to logical TRUE
# in data.frame(s) OOPS!
# Lesson learned: Never use 'T' or 'F' to label conditions/samples
gene_expr_diff$sample_1 <- 'T' # replace logical TRUE interpretation for the original character label T
iso_expr_diff <- read.delim2('../cuffdiff/diff_out/isoform_exp.diff')
iso_expr_diff$sample_1 <- 'T'

dim(gene_expr_diff[gene_expr_diff$status == 'OK' & gene_expr_diff$significant == 'yes', ])
```

```
## [1] 10125    14
```

```
head(gene_expr_diff[gene_expr_diff$status == 'OK' & gene_expr_diff$significant == 'yes', ])
```

```
##        test_id     gene_id         gene                locus sample_1 sample_2
## 4  XLOC_000004 XLOC_000004 LOC100287934   chr1:764864-810022        T        U
## 11 XLOC_000011 XLOC_000011      PLEKHN1   chr1:966491-982093        T        U
## 12 XLOC_000012 XLOC_000012        ISG15 chr1:1013466-1014540        T        U
## 13 XLOC_000013 XLOC_000013         AGRN chr1:1020122-1056119        T        U
## 14 XLOC_000014 XLOC_000014 LOC100288175 chr1:1059714-1067264        T        U
## 16 XLOC_000016 XLOC_000016            - chr1:1151935-1157301        T        U
##    status  value_1  value_2 log2.fold_change. test_stat p_value     q_value
## 4      OK  0.37868 0.059133          -2.67894  -1.96996  0.0027  0.00584503
## 11     OK 0.778295  1.58586           1.02688   2.82337   5e-05 0.000141598
## 12     OK  14.0938  24.7979          0.815152   3.32746   5e-05 0.000141598
## 13     OK  1.49063  5.56731           1.90106   7.75049   5e-05 0.000141598
## 14     OK  1.97514  3.18071          0.687392   2.09137 0.00025 0.000645331
## 16     OK 0.298565    1.018           1.76962   5.66169   5e-05 0.000141598
##    significant
## 4          yes
## 11         yes
## 12         yes
## 13         yes
## 14         yes
## 16         yes
```

```
dim(iso_expr_diff[iso_expr_diff$status == 'OK' & iso_expr_diff$significant == 'yes', ])
```

```
## [1] 11111    14
```

Volcano Plot

```
ged <- gene_expr_diff[gene_expr_diff$status == 'OK' &
                        !is.infinite(as.numeric(gene_expr_diff$log2.fold_change.)), ]
ged$log2.fold.change <- as.numeric(ged$log2.fold_change.)
ged$adj.p.value <- as.numeric(ged$q_value)
head(ged)
```

```
##        test_id     gene_id         gene                locus sample_1 sample_2
## 4  XLOC_000004 XLOC_000004 LOC100287934   chr1:764864-810022        T        U
## 5  XLOC_000005 XLOC_000005 LOC100287934   chr1:764864-810022        T        U
## 7  XLOC_000007 XLOC_000007    LINC01128   chr1:827590-859446        T        U
## 10 XLOC_000010 XLOC_000010       KLHL17   chr1:960586-965897        T        U
## 11 XLOC_000011 XLOC_000011      PLEKHN1   chr1:966491-982093        T        U
## 12 XLOC_000012 XLOC_000012        ISG15 chr1:1013466-1014540        T        U
##    status  value_1  value_2 log2.fold_change. test_stat p_value     q_value
## 4      OK  0.37868 0.059133          -2.67894  -1.96996  0.0027  0.00584503
## 5      OK 0.452312 0.572217           0.339244  0.414573  0.4708    0.545048
## 7      OK  3.91537  3.83394        -0.0303199 -0.130856  0.8162    0.855538
## 10     OK  3.13695  3.44618          0.135638  0.534257 0.34895    0.424701
## 11     OK 0.778295  1.58586           1.02688   2.82337   5e-05 0.000141598
## 12     OK  14.0938  24.7979          0.815152   3.32746   5e-05 0.000141598
##    significant log2.fold.change adj.p.value
## 4          yes       -2.6789400 0.005845030
## 5           no        0.3392440 0.545048000
## 7           no       -0.0303199 0.855538000
## 10          no        0.1356380 0.424701000
## 11         yes        1.0268800 0.000141598
## 12         yes        0.8151520 0.000141598
```

```
h.line <- -log10(0.05)
v.line <- log2(2)

ged$diff.expr <- "No"
ged$diff.expr[ged$log2.fold.change > v.line & -log10(ged$adj.p.value) > h.line ] <- "Up"
ged$diff.expr[ged$log2.fold.change < -v.line & -log10(ged$adj.p.value) > h.line ] <- "Down"

head(ged)
```

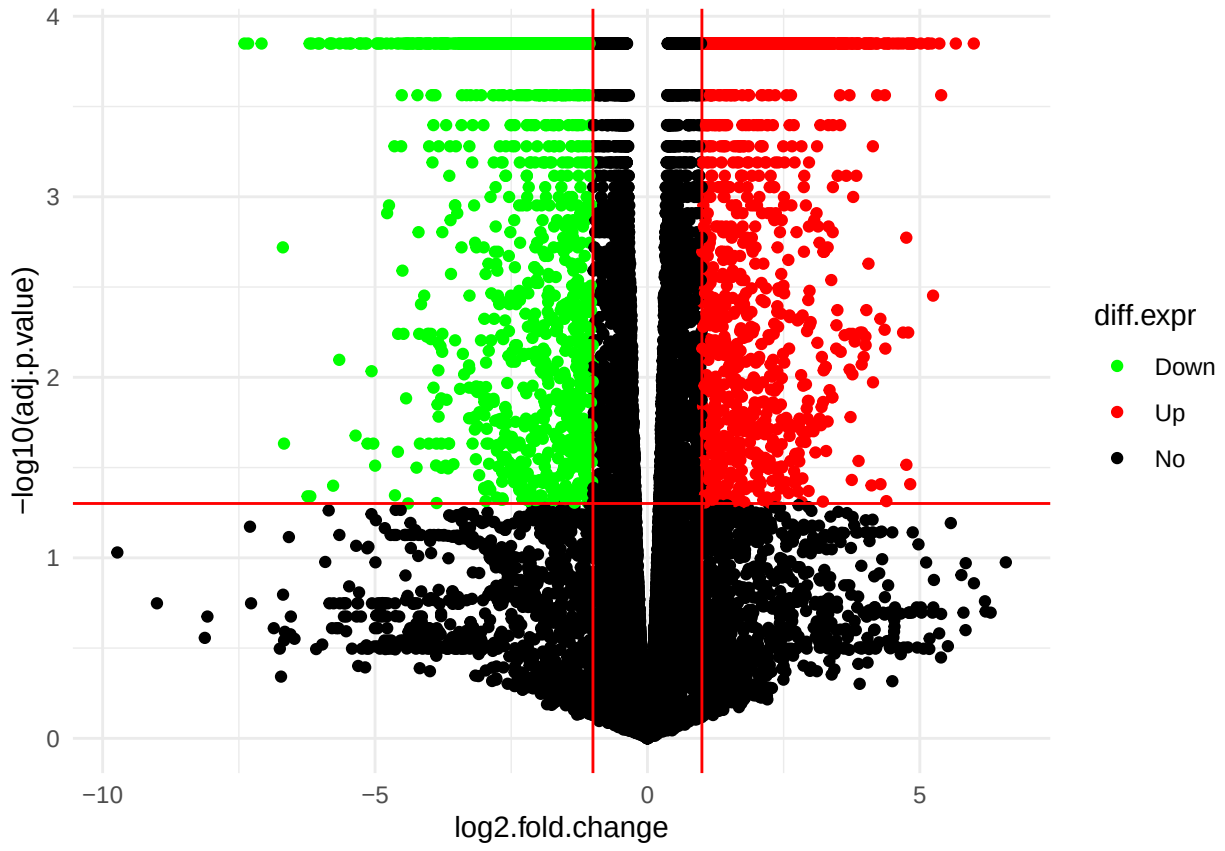
```
##        test_id     gene_id         gene                locus sample_1 sample_2
## 4  XLOC_000004 XLOC_000004 LOC100287934   chr1:764864-810022        T        U
## 5  XLOC_000005 XLOC_000005 LOC100287934   chr1:764864-810022        T        U
## 7  XLOC_000007 XLOC_000007    LINC01128   chr1:827590-859446        T        U
## 10 XLOC_000010 XLOC_000010       KLHL17   chr1:960586-965897        T        U
## 11 XLOC_000011 XLOC_000011      PLEKHN1   chr1:966491-982093        T        U
## 12 XLOC_000012 XLOC_000012        ISG15 chr1:1013466-1014540        T        U
##    status  value_1  value_2 log2.fold_change. test_stat p_value     q_value
## 4      OK  0.37868 0.059133          -2.67894  -1.96996  0.0027  0.00584503
## 5      OK 0.452312 0.572217           0.339244  0.414573  0.4708    0.545048
## 7      OK  3.91537  3.83394        -0.0303199 -0.130856  0.8162    0.855538
## 10     OK  3.13695  3.44618          0.135638  0.534257 0.34895    0.424701
## 11     OK 0.778295  1.58586           1.02688   2.82337   5e-05 0.000141598
```

```
## 12     OK  14.0938  24.7979          0.815152   3.32746   5e-05 0.000141598
##    significant log2.fold.change adj.p.value diff.expr
## 4          yes       -2.6789400 0.005845030      Down
## 5           no        0.3392440 0.545048000        No
## 7           no       -0.0303199 0.855538000        No
## 10          no        0.1356380 0.424701000        No
## 11         yes        1.0268800 0.000141598        Up
## 12         yes        0.8151520 0.000141598        No
```

```
cols <- c("green", "red", "black")
names(cols) <- c("Down", "Up", "No")

vp <- ggplot(data=ged, aes(x=log2.fold.change, y=-log10(adj.p.value ), col=diff.expr)) +
  geom_point() + theme_minimal() +
  scale_color_manual(values=cols) +
  geom_vline(xintercept=c(-v.line, v.line), col="red") +
  geom_hline(yintercept=h.line, col="red")


vp
```



Top 10 down- and up-regulated genes

```
ged.diff.expr <- ged[ged$diff.expr %in% c("Up", "Down"),]
dim(ged.diff.expr)
```

```
## [1] 4646   17
```

```
dim(ged.diff.expr[ged.diff.expr$significant == "no",])
```

```
## [1]  0 17
```

```
ged.diff.expr.order <- ged.diff.expr[order(ged.diff.expr[, 15]),]
head(ged.diff.expr.order, n=5)
```

```
##           test_id     gene_id         gene                      locus sample_1
## 31661 XLOC_031661 XLOC_031661         MGAM  chr7:141911493-142106747        T
## 31969 XLOC_031969 XLOC_031969       NPC1L1    chr7:44512432-44543696        T
## 9661  XLOC_009661 XLOC_009661 LOC102723479 chr14:106399794-106496343        T
## 19407 XLOC_019407 XLOC_019407         GALM    chr2:38665909-38734767        T
## 9955  XLOC_009955 XLOC_009955     SYNDIG1L   chr14:74405892-74426102        T
##       sample_2 status  value_1    value_2 log2.fold_change. test_stat p_value
## 31661        U     OK  2.03008   0.012047          -7.39671  -6.57524   5e-05
## 31969        U     OK  4.71272  0.0292933          -7.32985  -7.51378   5e-05
## 9661         U     OK   156.26    1.15253            -7.083  -8.67029   5e-05
## 19407        U     OK 0.908096 0.00878361          -6.69189 -0.786757  0.0008
## 9955         U     OK 0.722329 0.00708826          -6.67108    -4.504 0.01235
##           q_value significant log2.fold.change adj.p.value diff.expr
## 31661 0.000141598         yes         -7.39671 0.000141598      Down
## 31969 0.000141598         yes         -7.32985 0.000141598      Down
## 9661  0.000141598         yes         -7.08300 0.000141598      Down
## 19407  0.00190649         yes         -6.69189 0.001906490      Down
## 9955     0.023295         yes         -6.67108 0.023295000      Down
```

```
tail(ged.diff.expr.order, n=5)
```

```
##           test_id     gene_id   gene                     locus sample_1 sample_2
## 14418 XLOC_014418 XLOC_014418  GNGT2  chr17:49204331-49223325        T        U
## 27334 XLOC_027334 XLOC_027334 GPRIN1 chr5:176595801-176617357        T        U
## 8089  XLOC_008089 XLOC_008089      -    chr12:1527089-1529969        T        U
## 33426 XLOC_033426 XLOC_033426 NUDT18   chr8:22104944-22109419        T        U
## 23788 XLOC_023788 XLOC_023788 PFKFB4   chr3:48517659-48556835        T        U
##       status    value_1  value_2 log2.fold_change. test_stat p_value
## 14418     OK 0.00910793 0.344963           5.24318  0.715596 0.00155
## 27334     OK  0.0646707  2.65464           5.35926   11.1728   5e-05
## 8089      OK  0.0141094 0.592514           5.39212   4.31488  0.0001
## 33426     OK  0.0942486   4.7686           5.66095   6.04928   5e-05
## 23788     OK   0.106216  6.75489           5.99085   13.8721   5e-05
##           q_value significant log2.fold.change adj.p.value diff.expr
## 14418    0.003528         yes          5.24318 0.003528000        Up
## 27334 0.000141598         yes          5.35926 0.000141598        Up
## 8089  0.000273729         yes          5.39212 0.000273729        Up
## 33426 0.000141598         yes          5.66095 0.000141598        Up
## 23788 0.000141598         yes          5.99085 0.000141598        Up
```