H2afy_paper_deg

Sridhar

7/7/2021

Gene Expression analysis

DEG code used in paper Mutant U2AF1-Induced Alternative Splicing of H2afy (macroH2A1) Regulates B-Lymphopoiesis in Mice

Gene expression

```
suppressPackageStartupMessages({
library(DESeq2)
library(ggplot2)
library(pheatmap)
})

blp <- read.table('mergedCounts.tsv', header = T, sep='\t', stringsAsFactors = F)
head(blp )
```

```
##                    KO_4 KO_2 WT_3 KO_1 WT_4 WT_2 KO_3 WT_1
## ENSMUSG00000000001 2076 2198 2178 3246 2230 2322 1407 2955
## ENSMUSG00000000003    0    0    0    0    0    0    0    0
## ENSMUSG00000000028 2067 1749 1932 2916 2850 2513 1111 3337
## ENSMUSG00000000031    0    0    0    0    0    1    0    0
## ENSMUSG00000000037   86   88   64  101   97   87   59   82
## ENSMUSG00000000049    1    0    3    0    0    0    4    0
```

```
get_signi_deseq2 <- function(df,mt,wt){
  metak.s <- read.table('mm9_gene_biotype.tsv',header = T, stringsAsFactors = F, sep = '\t')
  print("## print before ordering")
  print(head(df))
  df <- df[,order(names(df))]
  df2 <- df
  df2$ID <- rownames(df2)
  print("## print after ordering")
  print(head(df))
  condition <- factor(c(rep("MT", mt), rep("WT", wt)))
  coldata.s <- data.frame(row.names =colnames(df), condition)
  coldata.s
  dds.s <- DESeqDataSetFromMatrix(countData = df, colData = coldata.s, design = ~condition)
  dds.s <- estimateSizeFactors(dds.s)
  dds.s <- DESeq(dds.s)
  dim(dds.s)
  dds.s$condition <- factor(dds.s$condition, levels = c("MT","WT"))
  png(paste0("qc-dispersions.png"), 1000, 1000, pointsize=20)
  plotDispEsts(dds.s, main="Dispersion plot")
```

```
  dev.off()
  res.s <- results(dds.s,independentFiltering = T, contrast = c("condition", "MT", "WT"))
  print("###total number of genes significant at <0.1")
  print(table(res.s$padj<0.1))
  res.s <- res.s[order(res.s$padj), ]
  resdata.s <- merge(as.data.frame(res.s), as.data.frame(counts(dds.s,normalize=TRUE)), by="row.names",
  print("#normalize resdata.s")
  print(head(resdata.s))
  resdata.s.r <- merge(as.data.frame(res.s), as.data.frame(counts(dds.s)), by="row.names", sort=FALSE)
  print("#normalize resdata.s.r")
  print(head(resdata.s.r))
  finalk.s <- merge(resdata.s, metak.s, by.x="Row.names", by.y="ensembl_gene_id")
  finalk.s.r <- merge(resdata.s.r, metak.s, by.x="Row.names", by.y="ensembl_gene_id")
  final.all <- merge(finalk.s, df2, by.x="Row.names",by.y="ID")
  print("### writing to file")
  write.csv(final.all, file=paste0("diffexpr-results_combined.csv"), row.names = F, quote = F)
  library(pheatmap)
  df.f.final.f <- dplyr::filter(resdata.s,  padj < 0.1)
  dim(df.f.final.f)
  print(head(df.f.final.f))
  df.f.final.f <- df.f.final.f[,-c(2:7)]
  names(df.f.final.f)
  print(head(df.f.final.f))
  d1 <- dim(df.f.final.f)
  head(df.f.final.f)
  heat <- pheatmap(df.f.final.f[-1], clustering_method = "average",
                   clustering_distance_rows = "euclidean",
                   clustering_distance_cols = "euclidean",
                   show_rownames = F,fontsize_row=10,
                   show_colname = T,
                   drop_levels = TRUE, scale = "row",
                   # color = colorRampPalette(c("green","red"))(70),
                   color = colorRampPalette(c("midnightblue","white","goldenrod1"))(70),
                   main=paste0("UPGMA clustering method "," n=", d1))
  pdf(paste0("heatmap_fdr_less_0.01.pdf"))
  print(heat)
  dev.off()

}
# prefilter low expressed genes
blp.pre <- rowSums(blp > 5) >= 4
data.cnt <- blp[blp.pre,]
dim(data.cnt)
```

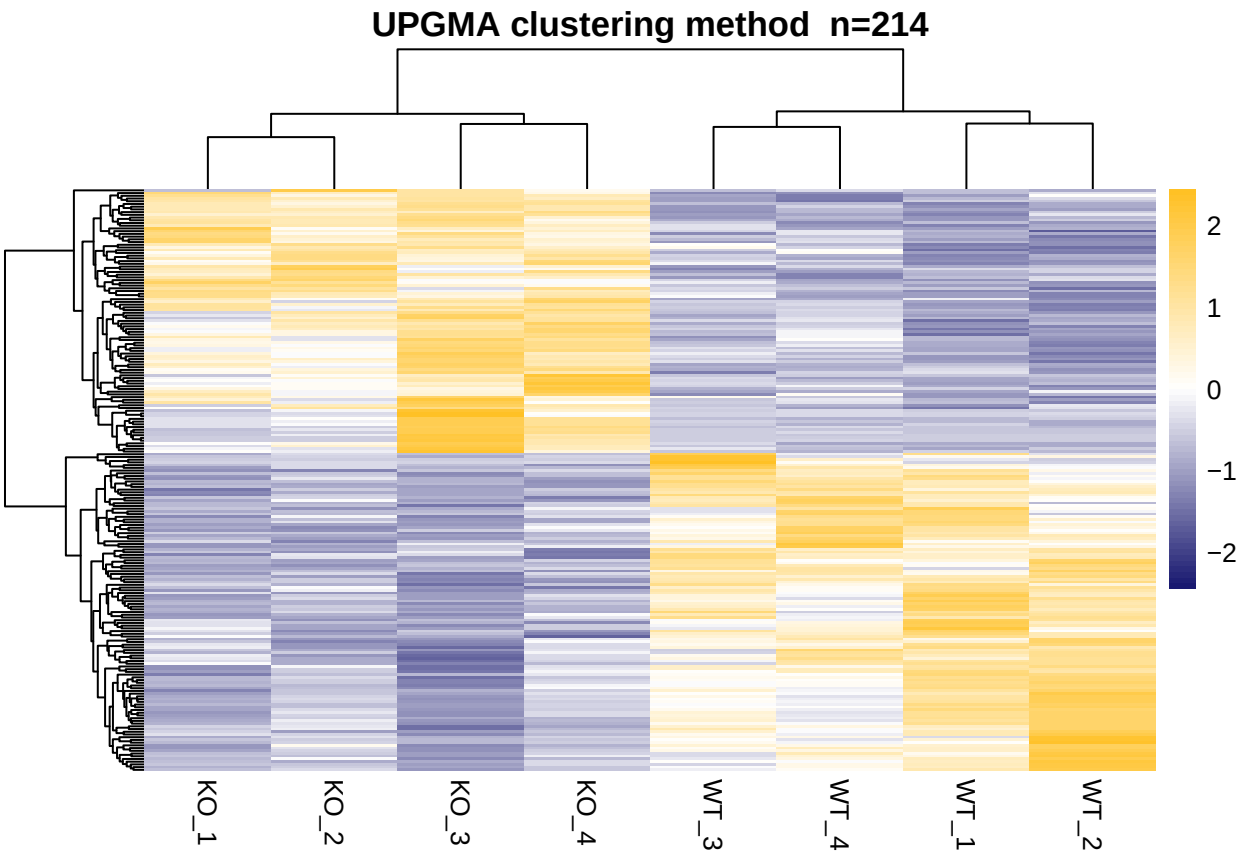
```
## [1] 14905     8
```

```
blp.whole.pre <-  get_signi_deseq2(data.cnt,4,4)
```

```
## [1] "## print before ordering"
##                     KO_4 KO_2 WT_3 KO_1 WT_4 WT_2 KO_3 WT_1
## ENSMUSG00000000001 2076 2198 2178 3246 2230 2322 1407 2955
## ENSMUSG00000000028 2067 1749 1932 2916 2850 2513 1111 3337
## ENSMUSG00000000037   86   88   64  101   97   87   59   82
## ENSMUSG00000000056 1640 1251 1269 1455 1678 1230  925 1657
```

```
## ENSMUSG00000000058  130  102   92   60  126  190   23  168
## ENSMUSG00000000078 3063 1036 1100 1224 2386 1489 1071 1457
## [1] "## print after ordering"
##                    KO_1 KO_2 KO_3 KO_4 WT_1 WT_2 WT_3 WT_4
## ENSMUSG00000000001 3246 2198 1407 2076 2955 2322 2178 2230
## ENSMUSG00000000028 2916 1749 1111 2067 3337 2513 1932 2850
## ENSMUSG00000000037  101   88   59   86   82   87   64   97
## ENSMUSG00000000056 1455 1251  925 1640 1657 1230 1269 1678
## ENSMUSG00000000058   60  102   23  130  168  190   92  126
## ENSMUSG00000000078 1224 1036 1071 3063 1457 1489 1100 2386

## using pre-existing size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## [1] "###total number of genes significant at <0.1"
##
## FALSE  TRUE
## 12344   214
## [1] "#normalize resdata.s"
##            Row.names   baseMean log2FoldChange     lfcSE      stat       pvalue
## 1 ENSMUSG00000032125   79.16034     -3.7145793 0.4616222 -8.046796 8.498989e-16
## 2 ENSMUSG00000036330  784.51657     -1.1688855 0.1598838 -7.310846 2.654659e-13
## 3 ENSMUSG00000026009  487.77743     -1.2720307 0.1822670 -6.978943 2.974093e-12
## 4 ENSMUSG00000018774 1323.72712      1.4535211 0.2211005  6.574029 4.897167e-11
## 5 ENSMUSG00000032503 6107.20294     -0.9132838 0.1426515 -6.402201 1.531528e-10
## 6 ENSMUSG00000022240  220.65069      1.3720072 0.2188991  6.267761 3.662766e-10
##           padj        KO_1      KO_2        KO_3       KO_4       WT_1
## 1 1.067303e-11    5.477708   11.7714    4.469999   22.77816  155.12632
## 2 1.666860e-09  505.514236  462.2951  536.399841  428.64355 1034.72554
## 3 1.244955e-08  298.143845  329.5993  248.829926  265.05495  806.98690
## 4 1.537466e-07 1702.002266 1727.1858 2430.189278 1897.83485  547.89295
## 5 3.846584e-07 4439.291427 5058.4928 3498.518960 3945.79848 8007.15838
## 6 7.666169e-07  314.576971  314.6175  368.029891  276.44403   99.84194
##        WT_2      WT_3      WT_4
## 1  142.0047  109.8661  181.7882
## 2  916.0697 1191.9440 1200.5406
## 3  601.4318  608.4097  743.7631
## 4  552.2406  971.1752  761.2959
## 5 8064.5688 7887.5599 7956.2347
## 6  144.7891  105.7203  141.1858
## [1] "#normalize resdata.s.r"
##            Row.names   baseMean log2FoldChange     lfcSE      stat       pvalue
## 1 ENSMUSG00000032125   79.16034     -3.7145793 0.4616222 -8.046796 8.498989e-16
## 2 ENSMUSG00000036330  784.51657     -1.1688855 0.1598838 -7.310846 2.654659e-13
## 3 ENSMUSG00000026009  487.77743     -1.2720307 0.1822670 -6.978943 2.974093e-12
## 4 ENSMUSG00000018774 1323.72712      1.4535211 0.2211005  6.574029 4.897167e-11
## 5 ENSMUSG00000032503 6107.20294     -0.9132838 0.1426515 -6.402201 1.531528e-10
## 6 ENSMUSG00000022240  220.65069      1.3720072 0.2188991  6.267761 3.662766e-10
```

```
##           padj KO_1 KO_2 KO_3 KO_4 WT_1 WT_2 WT_3 WT_4
## 1 1.067303e-11    7   11    3   22  188  153  106  197
## 2 1.666860e-09  646  432  360  414 1254  987 1150 1301
## 3 1.244955e-08  381  308  167  256  978  648  587  806
## 4 1.537466e-07 2175 1614 1631 1833  664  595  937  825
## 5 3.846584e-07 5673 4727 2348 3811 9704 8689 7610 8622
## 6 7.666169e-07  402  294  247  267  121  156  102  153
## [1] "### writing to file"
##            Row.names   baseMean log2FoldChange     lfcSE      stat       pvalue
## 1 ENSMUSG00000032125   79.16034     -3.7145793 0.4616222 -8.046796 8.498989e-16
## 2 ENSMUSG00000036330  784.51657     -1.1688855 0.1598838 -7.310846 2.654659e-13
## 3 ENSMUSG00000026009  487.77743     -1.2720307 0.1822670 -6.978943 2.974093e-12
## 4 ENSMUSG00000018774 1323.72712      1.4535211 0.2211005  6.574029 4.897167e-11
## 5 ENSMUSG00000032503 6107.20294     -0.9132838 0.1426515 -6.402201 1.531528e-10
## 6 ENSMUSG00000022240  220.65069      1.3720072 0.2188991  6.267761 3.662766e-10
##           padj        KO_1      KO_2        KO_3       KO_4       WT_1
## 1 1.067303e-11    5.477708   11.7714    4.469999   22.77816  155.12632
## 2 1.666860e-09  505.514236  462.2951  536.399841  428.64355 1034.72554
## 3 1.244955e-08  298.143845  329.5993  248.829926  265.05495  806.98690
## 4 1.537466e-07 1702.002266 1727.1858 2430.189278 1897.83485  547.89295
## 5 3.846584e-07 4439.291427 5058.4928 3498.518960 3945.79848 8007.15838
## 6 7.666169e-07  314.576971  314.6175  368.029891  276.44403   99.84194
##        WT_2      WT_3      WT_4
## 1  142.0047  109.8661  181.7882
## 2  916.0697 1191.9440 1200.5406
## 3  601.4318  608.4097  743.7631
## 4  552.2406  971.1752  761.2959
## 5 8064.5688 7887.5599 7956.2347
## 6  144.7891  105.7203  141.1858
##            Row.names        KO_1      KO_2        KO_3       KO_4       WT_1
## 1 ENSMUSG00000032125    5.477708   11.7714    4.469999   22.77816  155.12632
## 2 ENSMUSG00000036330  505.514236  462.2951  536.399841  428.64355 1034.72554
## 3 ENSMUSG00000026009  298.143845  329.5993  248.829926  265.05495  806.98690
## 4 ENSMUSG00000018774 1702.002266 1727.1858 2430.189278 1897.83485  547.89295
## 5 ENSMUSG00000032503 4439.291427 5058.4928 3498.518960 3945.79848 8007.15838
## 6 ENSMUSG00000022240  314.576971  314.6175  368.029891  276.44403   99.84194
##        WT_2      WT_3      WT_4
## 1  142.0047  109.8661  181.7882
## 2  916.0697 1191.9440 1200.5406
## 3  601.4318  608.4097  743.7631
## 4  552.2406  971.1752  761.2959
## 5 8064.5688 7887.5599 7956.2347
## 6  144.7891  105.7203  141.1858
## Warning in if (is.na(main)) {: the condition has length > 1 and only the first
## element will be used
## Warning in if (!is.na(main)) {: the condition has length > 1 and only the first
## element will be used
```


UPGMA clustering method n=214
2
1
0
−1
−2
KO_1
KO_2
KO_3
KO_4
WT_3
WT_4
WT_1
WT_2