batlas_formatting_averaged

Sarah Hp

2022-09-02

```
here::i_am("R/02_batlas_formatting_averaged.Rmd")
```

```
## here() starts at /projects/imb-pkbphil/sp/rnaseq/six_donor_trans/splicing_paper
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(ggplot2)
library(here)
figs = here("R/plots")
```

Now adding a BATLAS heatmap

```
fpkm = read.delim(here("03limma/beige_day15_rpkm_tmm_means.tab"))
head(fpkm)
```

```
##            Geneid Length gene_name
## 1 ENSG00000000003   4536    TSPAN6
## 2 ENSG00000000005   1476      TNMD
## 3 ENSG00000000419   1207      DPM1
## 4 ENSG00000000457   6883     SCYL3
## 5 ENSG00000000460   5970  C1orf112
## 6 ENSG00000000938   3382       FGR
##                                                    description   gene_biotype
## 1                                              tetraspanin 6  protein_coding
## 2                                                tenomodulin  protein_coding
## 3 dolichyl-phosphate mannosyltransferase subunit 1, catalytic  protein_coding
## 4                                    SCY1 like pseudokinase 3  protein_coding
## 5                          chromosome 1 open reading frame 112  protein_coding
```

```
## 6              FGR proto-oncogene, Src family tyrosine kinase  protein_coding
##   ensembl_gene_id_version subject1.beige subject1.white subject2.beige
## 1      ENSG00000000003.15       7.015370      5.0752183     12.5300941
## 2       ENSG00000000005.6       1.627583      0.3200113     30.6612441
## 3      ENSG00000000419.12      30.635853     33.4089794     32.3055705
## 4      ENSG00000000457.14       1.862880      1.4176901      2.3720554
## 5      ENSG00000000460.17       0.393713      0.4077405      0.4788014
## 6      ENSG00000000938.13       2.977751      0.7902128      4.6470874
##   subject2.white subject3.beige subject3.white subject4.beige subject4.white
## 1      6.4851444      9.7247952      6.2324136      9.3056202      9.0232992
## 2     30.2448972      2.7906896      0.4912616      2.3021246      0.6401144
## 3     29.1252400     34.1100563     34.7548040     32.7208903     34.6618668
## 4      1.5836657      2.1136964      1.5022350      2.1423301      1.7105704
## 5      0.3762281      0.4816479      0.4403470      0.4698593      0.5608791
## 6      0.8865716      7.1807572      0.9509210      5.8448505      0.6362375
##   subject5.beige subject5.white subject6.beige subject6.white
## 1     10.0681020      6.3019890     11.7900221      8.3815053
## 2      6.2119587      1.6856768      8.8600561      2.1400068
## 3     31.6549387     30.9930433     38.2724535     35.8339890
## 4      1.9104027      1.5130770      2.0849212      1.6548458
## 5      0.3557019      0.4699746      0.3751493      0.4361585
## 6      5.6321927      0.7380337     12.7832922      1.9632742
```

```
markrs = read.delim(here("annotations/batlas_MarkersList_V1.txt"))
head(markrs)
```

```
##   geneid              mouse           human marker.type
## 1  ACSL5 ENSMUSG00000024981 ENSG00000197142       brown
## 2 MT-ND3 ENSMUSG00000064360 ENSG00000198840       brown
## 3  PANK1 ENSMUSG00000033610 ENSG00000152782       brown
## 4   UCP1 ENSMUSG00000031710 ENSG00000109424       brown
## 5 MT-CO3 ENSMUSG00000064358 ENSG00000198938       brown
## 6 LETMD1 ENSMUSG00000037353 ENSG00000050426       brown
```

Remove duplicate IDs + remove additional columns

```
fpkm = fpkm[!duplicated(fpkm$Geneid),c(1,6:ncol(fpkm))]
dim(fpkm)
```

```
## [1] 18058    14
```

check marker genes are in list

```
summary(markrs$human %in% fpkm$Geneid)
```

```
##    Mode    TRUE
## logical     119
```

```
length(markrs$human)
```

```
## [1] 119
```

BATLAS Heatmap

```
library(ComplexHeatmap)
```

```
## Loading required package: grid
```

```
## ========================================
## ComplexHeatmap version 2.14.0
## Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
## Github page: https://github.com/jokergoo/ComplexHeatmap
## Documentation: http://jokergoo.github.io/ComplexHeatmap-reference
##
## If you use it in published research, please cite either one:
## - Gu, Z. Complex Heatmap Visualization. iMeta 2022.
## - Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
##     genomic data. Bioinformatics 2016.
##
##
## The new InteractiveComplexHeatmap package can directly export static
## complex heatmaps into an interactive Shiny app with zero effort. Have a try!
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(ComplexHeatmap))
## ========================================
```

```
ma = merge(fpkm, markrs, by.x="Geneid", by.y="human")
nrow(ma)
```

```
## [1] 119
```

```
head(ma)
```

```
##            Geneid ensembl_gene_id_version subject1.beige subject1.white
## 1 ENSG00000004142      ENSG00000004142.12      27.670863      21.183021
## 2 ENSG00000004779      ENSG00000004779.10      12.401646       9.727229
## 3 ENSG00000004961      ENSG00000004961.15       9.806021       8.064562
## 4 ENSG00000005194      ENSG00000005194.15       8.061360       6.933250
## 5 ENSG00000006695      ENSG00000006695.12       2.216775       1.562307
## 6 ENSG00000007923      ENSG00000007923.16      11.248144       9.341558
##   subject2.beige subject2.white subject3.beige subject3.white subject4.beige
## 1      48.327490      27.693305       43.60495      30.705074      41.344838
## 2      25.746909      13.966258       20.46305      11.224930      17.454139
## 3      13.790138      10.586403       11.34719       8.747039      11.869558
## 4      11.958195       7.909172       10.60540       9.457793      10.271900
## 5       3.727428       1.997453        3.14132       1.863222       3.421081
## 6      13.886272      10.356216       13.92478      12.478498      13.941201
##   subject4.white subject5.beige subject5.white subject6.beige subject6.white
## 1      35.028689      43.075900      23.949367      56.423883      43.164307
## 2      11.579449      23.681495       9.149905      21.573128      16.098597
## 3       8.173349      11.263196       8.889168      13.158955       9.509081
## 4      11.768426      10.678632       7.339716      10.631814      12.653139
```

```
## 5       2.321938       3.374354       2.098090       3.837377       2.570708
## 6      10.661073      13.984772      11.072825      13.969384      11.386226
##    geneid              mouse marker.type
## 1 POLDIP2 ENSMUSG00000001100       brown
## 2 NDUFAB1 ENSMUSG00000030869       brown
## 3    HCCS ENSMUSG00000031352       brown
## 4 CIAPIN1 ENSMUSG00000031781       brown
## 5   COX10 ENSMUSG00000042148       brown
## 6 DNAJC11 ENSMUSG00000039768       brown
```

```
rownames(ma) = ma$geneid
ma = as.matrix(ma[3:(ncol(ma)-3)])
#head(ma)
```

```
sm = scale(t(ma))
#head(sm)
```

```
markrs = markrs[order(markrs$geneid),]
head(markrs)
```

```
##    geneid              mouse           human marker.type
## 88  ACADS ENSMUSG00000029545 ENSG00000122971       brown
## 29 ACADVL ENSMUSG00000018574 ENSG00000072778       brown
## 98  ACAT1 ENSMUSG00000032047 ENSG00000075239       brown
## 51   ACO2 ENSMUSG00000022477 ENSG00000100412       brown
## 93  ACSF2 ENSMUSG00000076435 ENSG00000167107       brown
## 1   ACSL5 ENSMUSG00000024981 ENSG00000197142       brown
```

```
bvw = HeatmapAnnotation(marker_for=markrs$marker.type[order(markrs$geneid)],
                        col = list(marker_for=c("brown"="sienna4", "white"="grey75")))
plot(bvw)
```


```
head(colnames(sm))
```

```
## [1] "POLDIP2" "NDUFAB1" "HCCS"    "CIAPIN1" "COX10"   "DNAJC11"
```

```
# sort for annotation
sm = sm[,order(colnames(sm))]
```

Add a rowannotation

```
treat = gsub(".*\\.","",row.names(sm))
treat[treat == "b"] = "beige"
treat[treat == "w"] = "white"
treat
```

```
##  [1] "beige" "white" "beige" "white" "beige" "white" "beige" "white" "beige"
## [10] "white" "beige" "white"
```

```
treatann = rowAnnotation(treat=treat, col=list(treat =c(white="#74B2CD",beige="#C1A486")),
                         annotation_name_side = "top")
draw(treatann)
```

Filter on significance

```
sig = read.delim(here( "03limma/any_and_all_donor_DGE.tsv"))
head(sig)
```

```
##            Geneid gene_name
## 1 ENSG00000000003    TSPAN6
## 2 ENSG00000000005      TNMD
## 3 ENSG00000000419      DPM1
## 4 ENSG00000000457     SCYL3
## 5 ENSG00000000460  C1orf112
## 6 ENSG00000000938       FGR
##                                                  description Length
## 1                                              tetraspanin 6   4536
## 2                                                tenomodulin   1476
## 3 dolichyl-phosphate mannosyltransferase subunit 1, catalytic   1207
## 4                                SCY1 like pseudokinase 3   6883
## 5                          chromosome 1 open reading frame 112   5970
## 6              FGR proto-oncogene, Src family tyrosine kinase   3382
##     gene_biotype    logFC.s1   logFC.s2    logFC.s3    logFC.s4    logFC.s5
## 1 protein_coding  0.49422154  0.8603088  0.57278654  0.37343465  0.56429720
## 2 protein_coding  2.38787369 -0.2720670  2.58660597  1.81772689  1.94504039
## 3 protein_coding -0.11790050  0.1194288 -0.06185359  0.06689829 -0.01815705
## 4 protein_coding  0.39439050  0.5484816  0.49896973  0.32546768  0.35682721
```

```
## 5 protein_coding -0.05290602  0.3071896  0.13017667 -0.17866146 -0.36021700
## 6 protein_coding  1.91496687  2.3497110  2.88724528  3.24138536  2.88754674
##     logFC.s6  AveExpr          F all.donors.P.Value all.donors.adj.P.Val
## 1  0.6439646 5.212308 31.7140824       2.969288e-11         7.806158e-10
## 2  1.9785258 1.961126 20.9163666       3.809066e-09         5.434087e-08
## 3  0.1710766 5.314282  0.4712428       8.237264e-01         8.380174e-01
## 4  0.3469994 3.626998 10.0207970       6.300732e-06         3.633461e-05
## 5 -0.2082301 1.367959  1.3456734       2.705375e-01         3.177382e-01
## 6  2.7442577 2.951839 95.9233119       2.476214e-17         5.025045e-15
##   all.donors.AvelogFC   P.Value.s1 adj.P.Val.s1 AveExpr.s1   P.Value.s2
## 1           0.5848355 7.423497e-05 2.280200e-03   5.212308 1.031683e-08
## 2           1.7406176 4.606466e-06 2.307037e-04   1.961126 5.374467e-01
## 3           0.0265821 4.216597e-01 7.813318e-01   5.314282 4.196868e-01
## 4           0.4118560 5.556707e-03 6.606388e-02   3.626998 3.317889e-04
## 5          -0.0604414 7.868527e-01 9.473852e-01   1.367959 1.314679e-01
## 6           2.6708521 8.493511e-08 8.063249e-06   2.951839 3.346207e-09
##   adj.P.Val.s2 AveExpr.s2   P.Value.s3 adj.P.Val.s3 AveExpr.s3   P.Value.s4
## 1 4.254162e-07   5.212308 1.180377e-05 2.253571e-04   5.212308 1.164889e-02
## 2 6.575098e-01   1.961126 1.956120e-06 5.218110e-05   1.961126 1.864017e-03
## 3 5.530398e-01   5.314282 6.755915e-01 8.196868e-01   5.314282 7.245401e-01
## 4 2.112990e-03   3.626998 8.348144e-04 6.877477e-03   3.626998 6.456487e-02
## 5 2.355830e-01   1.367959 5.128644e-01 7.062247e-01   1.367959 4.826271e-01
## 6 1.798686e-07   2.951839 3.418267e-11 6.015586e-09   2.951839 1.506080e-10
##   adj.P.Val.s4 AveExpr.s4   P.Value.s5 adj.P.Val.s5 AveExpr.s5   P.Value.s6
## 1 4.171991e-02   5.212308 1.989400e-05 3.539955e-04   5.212308 5.072093e-06
## 2 1.155716e-02   1.961126 1.460798e-04 1.739187e-03   1.961126 1.584704e-04
## 3 8.149446e-01   5.314282 9.043584e-01 9.430495e-01   5.314282 2.804584e-01
## 4 1.441980e-01   3.626998 1.359659e-02 5.250544e-02   3.626998 2.076997e-02
## 5 6.174632e-01   1.367959 8.133352e-02 1.835518e-01   1.367959 3.288567e-01
## 6 5.681113e-08   2.951839 6.097364e-11 1.596007e-08   2.951839 2.754188e-10
##   adj.P.Val.s6 AveExpr.s6
## 1 1.086679e-04   5.212308
## 2 1.496932e-03   1.961126
## 3 4.204665e-01   5.314282
## 4 6.094661e-02   3.626998
## 5 4.715745e-01   1.367959
## 6 5.292092e-08   2.951839
```

```
any_sig = sig[sig$all.donors.adj.P.Val < 0.01,]
nrow(any_sig)
```

```
## [1] 7554
```

```
summary(markrs$human %in% any_sig$Geneid)
```

```
##    Mode   FALSE    TRUE
## logical      21      98
```

```
summary(sig$all.donors.adj.P.Val[sig$Geneid[sig$all.donors.adj.P.Val > 0.01] %in%  markrs$human])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0002594 0.0165473 0.1520341 0.1547173 0.8526574
```

```
markrs[markrs$human %in% sig$Geneid[sig$all.donors.adj.P.Val > 0.01] ,] #21 BATLAS gene are not DE, they
```

```
##       geneid              mouse           human marker.type
## 119   ACVR1C ENSMUSG00000026834 ENSG00000123612       white
## 22     AMACR ENSMUSG00000022244 ENSG00000242110       brown
## 97  AURKAIP1 ENSMUSG00000065990 ENSG00000175756       brown
## 107    CCND2 ENSMUSG00000000184 ENSG00000118971       white
## 112    DMRT2 ENSMUSG00000048138 ENSG00000173253       white
## 109  GADD45A ENSMUSG00000036390 ENSG00000116717       white
## 101     IGF1 ENSMUSG00000020053 ENSG00000017427       white
## 104      LEP ENSMUSG00000059201 ENSG00000174697       white
## 113   LPGAT1 ENSMUSG00000026623 ENSG00000123684       white
## 116     LRP1 ENSMUSG00000040249 ENSG00000123384       white
## 87    MARCH5 ENSMUSG00000023307 ENSG00000198060       brown
## 5     MT-CO3 ENSMUSG00000064358 ENSG00000198938       brown
## 11    MT-ND2 ENSMUSG00000064345 ENSG00000198763       brown
## 64   NDUFA13 ENSMUSG00000036199 ENSG00000186010       brown
## 37    NDUFB7 ENSMUSG00000033938 ENSG00000099795       brown
## 115    NUPR1 ENSMUSG00000030717 ENSG00000176046       white
## 66      POLN ENSMUSG00000045102 ENSG00000130997       brown
## 100  PRKCDBP ENSMUSG00000037060 ENSG00000170955       white
## 110     PYGB ENSMUSG00000033059 ENSG00000100994       white
## 73     THEM4 ENSMUSG00000028145 ENSG00000159445       brown
## 94    TIMM50 ENSMUSG00000003438 ENSG00000105197       brown
```

```
table(markrs$marker.type)
```

```
##
## brown white
##    98    21
```

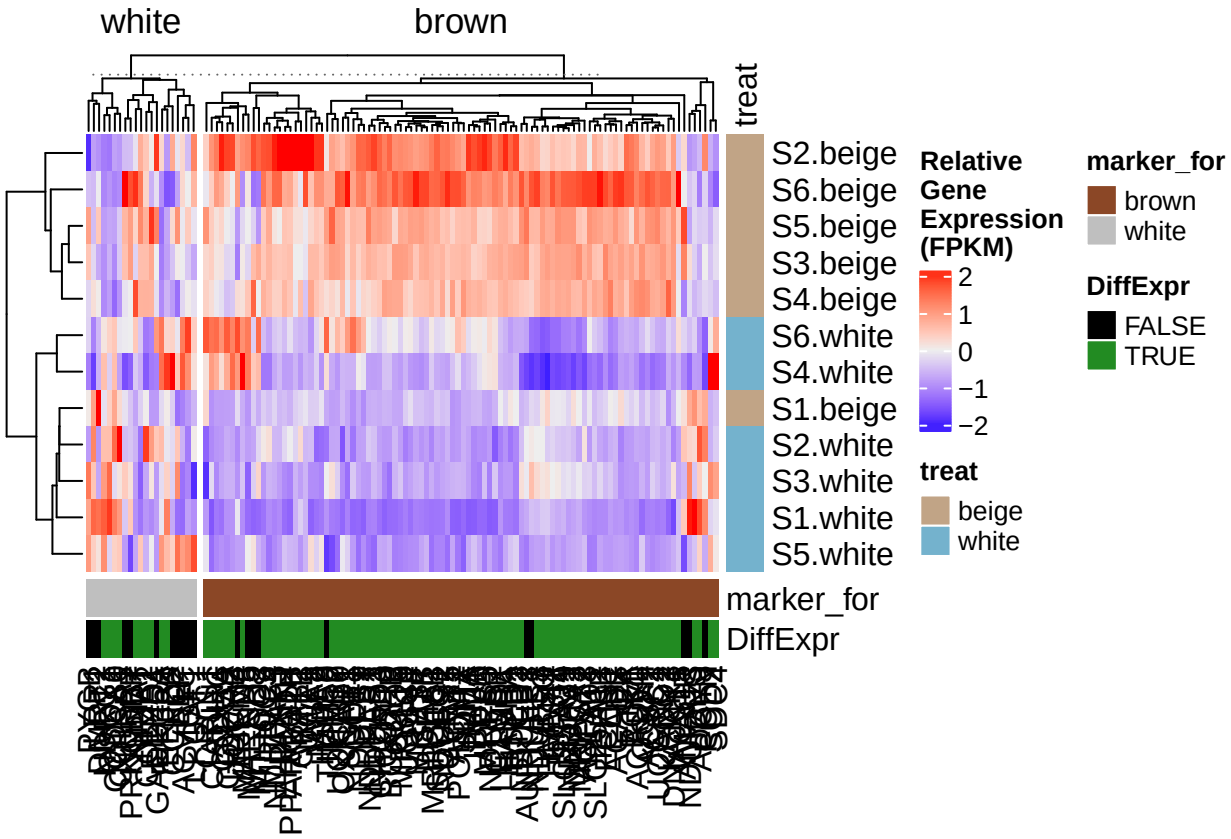
```
table(markrs$marker.type[!markrs$human %in% any_sig$Geneid])
```

```
##
## brown white
##    10    11
```

Hmm since half of the non-differential genes are white, I'd like to show that...

```
rownames(sm) = gsub("subject","S", rownames(sm))
is_sig = HeatmapAnnotation(marker_for=markrs$marker.type[order(markrs$geneid)],
                           DiffExpr=markrs$human %in% any_sig$Geneid,
                         col = list(DiffExpr=c("TRUE"="forestgreen", "FALSE"="black"),
                                    marker_for=c("brown"="sienna4", "white"="grey75")))
```

```
Heatmap(sm, bottom_annotation = is_sig, name="Relative \nGene \nExpression \n(FPKM)",
             right_annotation = treatann, column_title_side ="top",
           column_split = markrs$marker.type)
```



Take out non-DE genes and white markers

```
sm = t(sm)
sm = sm[markrs$geneid[markrs$marker.type == "brown"],]
nrow(sm)
```

```
## [1] 98
```

```
nrow(any_sig)
```

```
## [1] 7554
```

```
sm = sm[rownames(sm) %in% any_sig$gene_name,]
nrow(sm)
```

```
## [1] 88
```

```
treatann = HeatmapAnnotation(treat=treat,
                                #  subject = gsub("\\..*","",colnames(sm)),
                               col=list(treat =c(white="#74B2CD",beige="#C1A486")))
```

```
donors = c("S1","S3","S4","S5","S6","S2")
pdf(here("R/plots", "BATLAS_heatmap.pdf"), width=7, height=7)
Heatmap(sm,name="Relative \nGene \nExpression \n(FPKM)",
            bottom_annotation = treatann, column_title_side ="top",
          column_order =  paste( c(rev(donors), donors),  sep=".", rep(c("white","beige"),each=6)),
        column_labels = gsub("\\..*","",colnames(sm)), column_names_side = "top",
        column_names_rot=0, column_names_centered = T,
)
dev.off
```

```
## function (which = dev.cur())
## {
##     if (which == 1)
##         stop("cannot shut down device 1 (the null device)")
##     .External(C_devoff, as.integer(which))
##     dev.cur()
## }
## <bytecode: 0x4173880>
## <environment: namespace:grDevices>
```

BATLAS median expression

```
load(here("03limma/rpkm_rep_for_plotting.RData"))
```

```
batlas_genes  = filter(long, Geneid %in% markrs$human[markrs$marker.type == "brown"])
head(long)
```

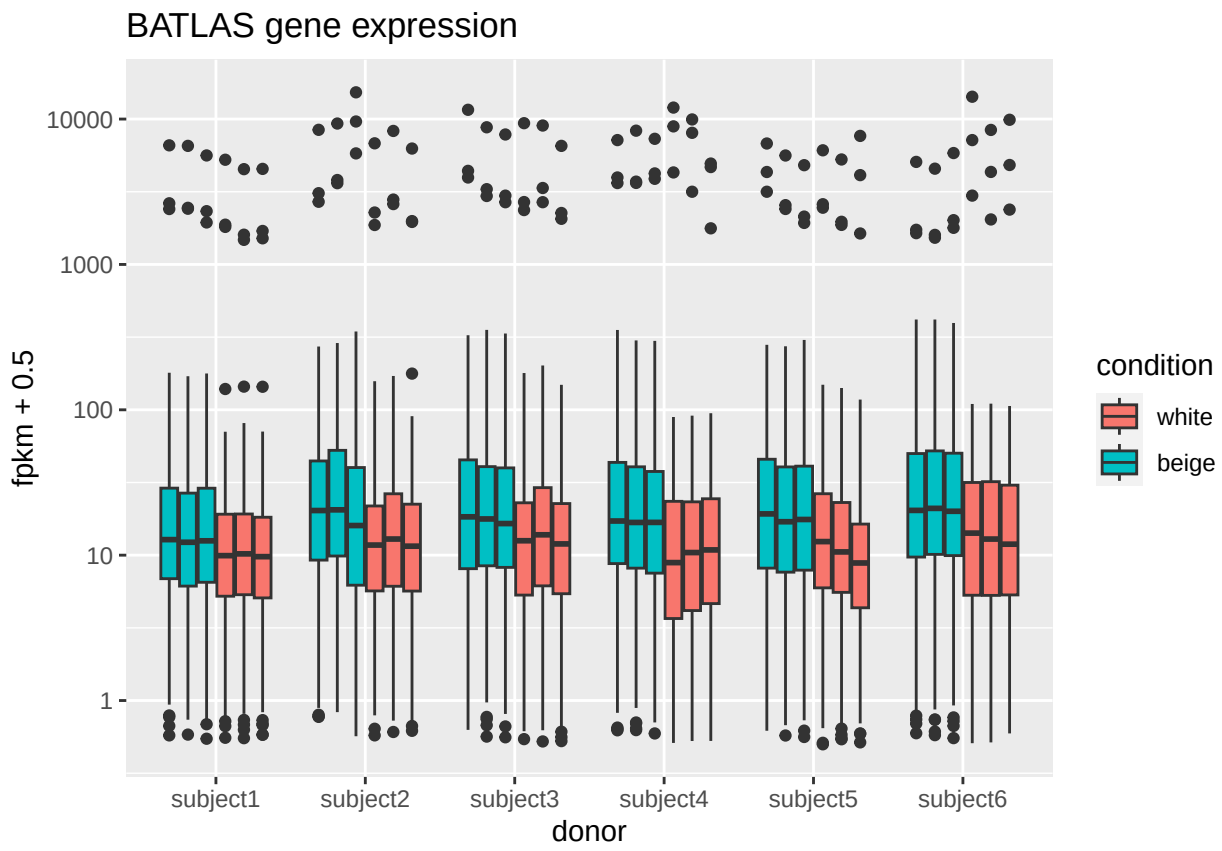
```
## # A tibble: 6 x 13
## # Groups:   Geneid [1]
##   Geneid Length gene_name description gene_biotype ensembl_gene_id_vers~1 biorep
##   <chr>   <dbl> <chr>     <chr>       <chr>        <chr>                   <chr>
## 1 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## 2 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## 3 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## 4 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## 5 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## 6 ENSG0~   4536 TSPAN6    "tetraspan~ protein_cod~ ENSG00000000003.15     subje~
## # i abbreviated name: 1: ensembl_gene_id_version
## # i 6 more variables: fpkm <dbl>, donor <fct>, condition <fct>, rep <chr>,
## #   donor.condition <fct>, zscore <dbl[,1]>
```

```
length(unique(batlas_genes$gene_name))
```

```
## [1] 98
```

```
batlas_genes$donor = factor(batlas_genes$donor, levels=levels(batlas_genes$donor)[order(levels(batlas_g
```

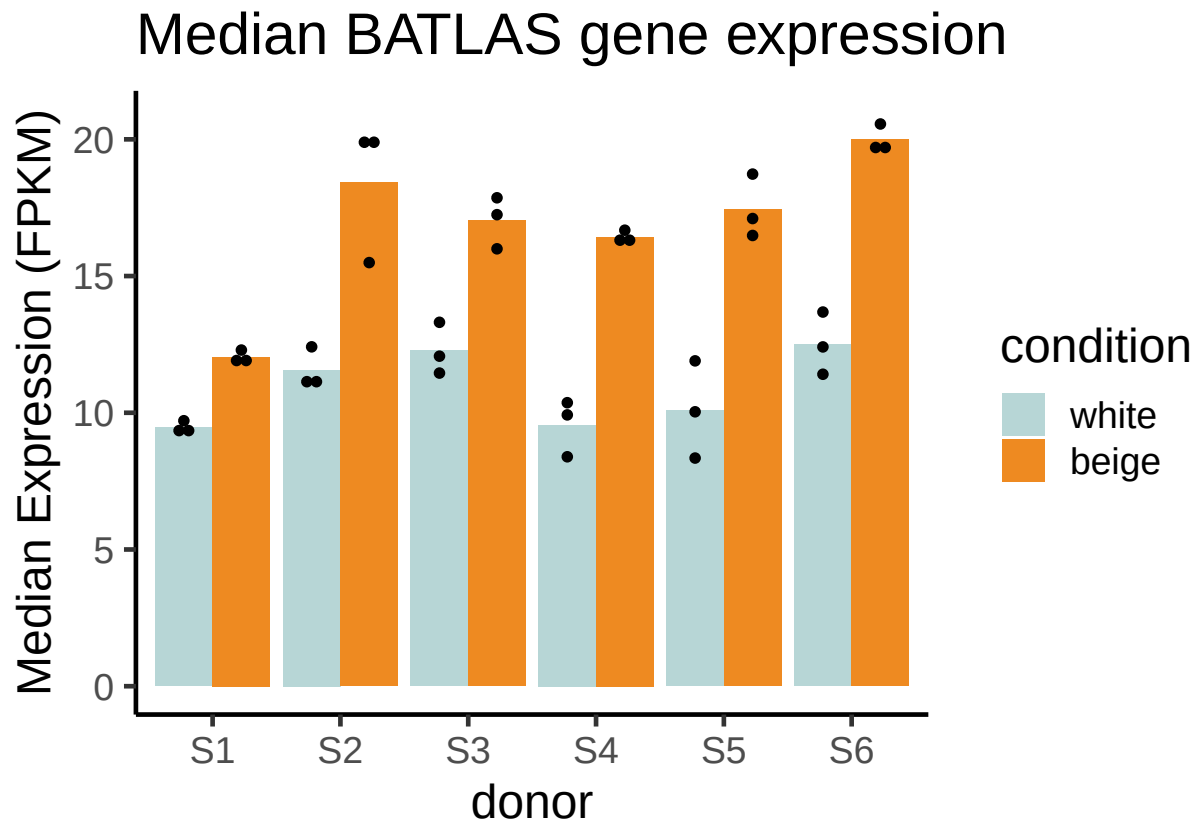
```
ggplot(batlas_genes,
         aes(x=donor, y=fpkm +0.5, fill=condition, group=biorep)) +
    geom_boxplot()  +
    ggtitle("BATLAS gene expression") + scale_y_log10()
```




```
to_plot =  group_by(batlas_genes, biorep, donor, condition, donor.condition,rep) %>% summarise(average =
```

```
## `summarise()` has grouped output by 'biorep', 'donor', 'condition',
## 'donor.condition'. You can override using the `.groups` argument.
```

```
ggplot(to_plot,
         aes(x=donor, y=median, fill=condition, group=donor.condition)) +
  geom_bar(stat = "summary", fun="mean", position="dodge")  +
    geom_dotplot(binaxis="y", stackdir = "center", position=position_dodge(0.9), fill="black", binwidth=
    labs(title="Median BATLAS gene expression", y="Median Expression (FPKM)") + theme_classic(base_size=
  scale_fill_manual(values =c("#B7D6D6", "#EE8A21")) +
                      scale_x_discrete(labels = gsub("subject","S",levels(to_plot$donor)))
```




```
ggsave(file= here(figs, "median_batlas_gene_expression.pdf"))
```

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

```
library(ggpubr)
head(to_plot)
```

```
## # A tibble: 6 x 8
## # Groups:   biorep, donor, condition, donor.condition [6]
##   biorep            donor condition donor.condition rep   average median ave_log
##   <chr>             <fct> <fct>     <fct>           <chr>   <dbl>  <dbl>   <dbl>
## 1 subject1.beige_r~ subj~ beige     subject1.beige  rep1    140.   12.3     3.63
## 2 subject1.beige_r~ subj~ beige     subject1.beige  rep2    137.   11.8     3.58
## 3 subject1.beige_r~ subj~ beige     subject1.beige  rep3    122.   12.1     3.59
## 4 subject1.white_r~ subj~ white     subject1.white  rep1    105.    9.43    3.11
## 5 subject1.white_r~ subj~ white     subject1.white  rep2     92.0   9.71    3.13
## 6 subject1.white_r~ subj~ white     subject1.white  rep3     92.9   9.27    3.05
```

```
table(to_plot$group)
```

```
## Warning: Unknown or uninitialised column: `group`.
```

```
## < table of extent 0 >
```

```r
summary(aov(median~condition*donor, data=to_plot))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## condition        1  322.2   322.2 239.660 5.45e-14 ***
## donor            5  108.0    21.6  16.063 5.58e-07 ***
## condition:donor  5   28.3     5.7   4.205  0.00693 **
## Residuals       24   32.3     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
compare_means(median~condition, group.by="donor", data=to_plot, method="t.test", p.adjust.method = "fdr
```

```
## Adding missing grouping variables: `biorep`, `donor.condition`
```

```
## # A tibble: 6 x 9
##   donor    .y.    group1 group2        p  p.adj p.format p.signif method
##   <fct>    <chr>  <chr>  <chr>     <dbl>  <dbl> <chr>    <chr>    <chr>
## 1 subject1 median white  beige  0.000240 0.0014 0.00024  ***      T-test
## 2 subject2 median white  beige  0.0343   0.034  0.03434  *        T-test
## 3 subject3 median white  beige  0.00356  0.0071 0.00356  **       T-test
## 4 subject4 median white  beige  0.00584  0.0075 0.00584  **       T-test
## 5 subject5 median white  beige  0.00622  0.0075 0.00622  **       T-test
## 6 subject6 median white  beige  0.00264  0.0071 0.00264  **       T-test
```