

# istotype\_ext\_analysis.Rmd

```
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:data.table':
##
##   between, first, last
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(stringr)
library(ggplot2)
library(NMF)
```

```
## Loading required package: pkgmaker
## Loading required package: registry
##
## Attaching package: 'pkgmaker'
## The following object is masked from 'package:base':
##
##   isNamespaceLoaded
## Loading required package: rngtools
## Loading required package: cluster
## NMF - BioConductor layer [NO: missing Biobase] | Shared memory capabilities [NO: bigmemory] | Cores !
## To enable the Bioconductor layer, try: install.extras('
## NMF
## ') [with Bioconductor repository enabled]
## To enable shared memory capabilities, try: install.extras('
## NMF
## ')
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
## The following objects are masked from 'package:data.table':
##
##   dcast, melt
```

```
library(parallel)
library(RColorBrewer)
library(scales)
```

```

summarise = dplyr::summarise

load("shm_rep12.rda")

df = shm %>%
  mutate(replacement = ifelse(as.character(from.aa) != as.character(to.aa), "replacement", "silent"),
         i = isotype) %>%
  group_by(clone, sample, proj, i, replacement) %>%
  summarise(count = n())

df.s = df %>%
  group_by(proj, i) %>%
  summarise(count = n())

print(df.s)

## # A tibble: 17 x 3
## # Groups:   proj [?]
##   proj i      count
##   <chr> <chr> <int>
## 1 old  ""         2
## 2 old  IGHA1    2145
## 3 old  IGHD      8
## 4 old  IGHE     125
## 5 old  IGHG1   2456
## 6 old  IGHG2      2
## 7 old  IGHG3    118
## 8 old  IGHGP     10
## 9 old  IGHM    3000
## 10 young IGHA1   3717
## 11 young IGHD      69
## 12 young IGHE     108
## 13 young IGHG1   2545
## 14 young IGHG2      2
## 15 young IGHG3    186
## 16 young IGHGP     19
## 17 young IGHM   3174

df = df %>% filter(i != "IGHD", !is.na(i), i != "") %>%
  mutate(isotype.full = i, isotype = str_sub(i, 4, 4))

df$proj = factor(df$proj, levels = c('young', 'old'))

dt.p = data.table()

for (iso in unique(df$isotype)) {
  tmp = df %>% filter(isotype == iso)
  x = (tmp %>% filter(proj == "old"))$count
  y = (tmp %>% filter(proj != "old"))$count
  kk = ks.test(x, y)
  p = kk$p.value
  dt.p = rbind(dt.p,
               data.table(isotype = iso, p=p))
}

```

```
## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties
```

```
## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties
```

```
## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties
```

```
## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties
```

```
dt.p$p.adj = p.adjust(dt.p$p, method = "BH")
print(dt.p %>% arrange(p.adj))
```

```
##      isotype      p      p.adj
## 1      M 0.000000e+00 0.000000e+00
## 2      G 1.520080e-09 3.040159e-09
## 3      A 3.476411e-03 4.635215e-03
## 4      E 1.138583e-01 1.138583e-01
```

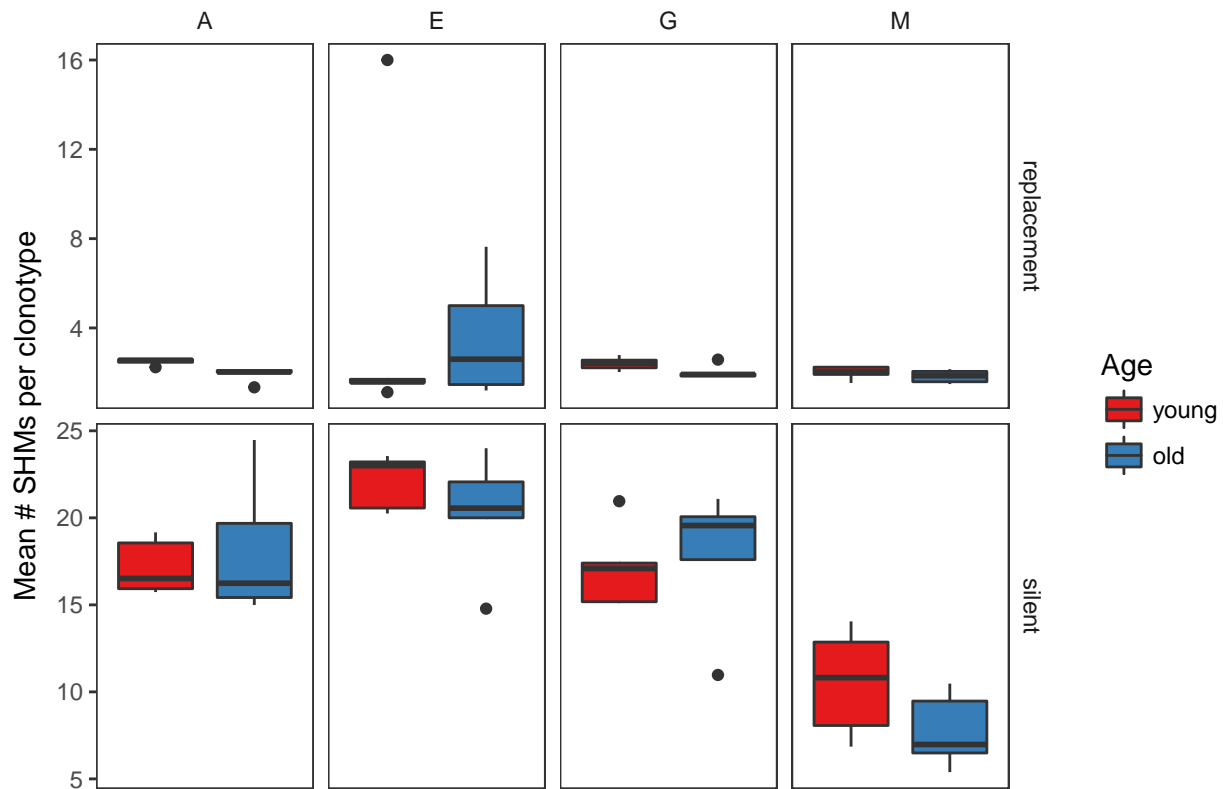
```
p13=ggplot(df, aes(x = count, fill = proj)) +
  geom_density(alpha = 0.9, color = NA) +
  facet_wrap(~isotype) +
  scale_fill_brewer("Age", palette = "Set1") +
  xlab("SHMs per clonotype") + ylab("") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        strip.background = element_blank())
ggsave("figures/p13.pdf", p13)
```

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

```
df.1 = df %>%
  group_by(sample, proj, replacement, isotype) %>%
  summarise(shms = mean(count))
```

```
p10=ggplot(df.1, aes(x=proj, fill = proj, y = shms)) +
  geom_boxplot() +
  facet_grid(replacement~isotype, scales = "free") +
  scale_fill_brewer("Age", palette = "Set1") +
  xlab("") + ylab("Mean # SHMs per clonotype") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        strip.background = element_blank())
```

```
p10
```



```
ggsave("figures/p10.pdf", p10)
```

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

```
a = aov(shms ~ replacement + isotype + proj, df.1)
summary(a)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## replacement  1   3808    3808  352.02 < 2e-16 ***
## isotype      3    513     171   15.80 4.88e-08 ***
## proj         1      7        7    0.68  0.412
## Residuals   74    800      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(a, "proj")
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = shms ~ replacement + isotype + proj, data = df.1)
##
## $proj
##           diff       lwr       upr     p adj
## old-young -0.6065988 -2.071921 0.8587229 0.4121063
```

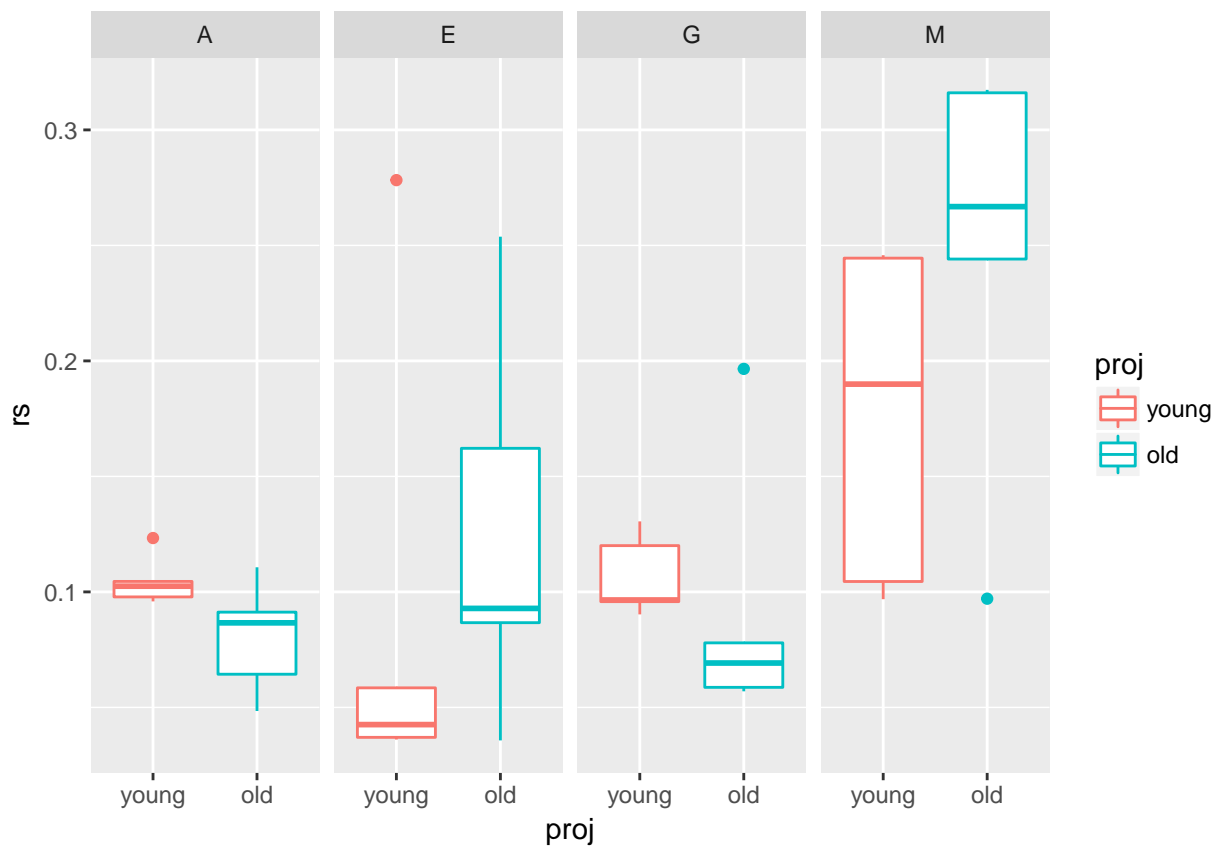
```
TukeyHSD(a, "isotype")
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
```

```
## Fit: aov(formula = shms ~ replacement + isotype + proj, data = df.1)
##
## $isotype
##          diff          lwr          upr      p adj
## E-A  2.65255106 -0.08101161  5.386113719 0.0604525
## G-A -0.08212375 -2.81568641  2.651438914 0.9998218
## M-A -4.40984410 -7.14340677 -1.676281440 0.0003651
## G-E -2.73467481 -5.46823747 -0.001112141 0.0498681
## M-E -7.06239516 -9.79595782 -4.328832495 0.0000000
## M-G -4.32772035 -7.06128302 -1.594157690 0.0004819
```

```
df.2 = df %>%
  group_by(sample, proj, isotype) %>%
  summarise(rs = sum(count[which(replacement == "replacement")]) / sum(count[which(replacement != "replacement")]))

ggplot(df.2, aes(x=proj, color = proj, y = rs)) +
  geom_boxplot() +
  facet_grid(.~isotype, scales = "free")
```



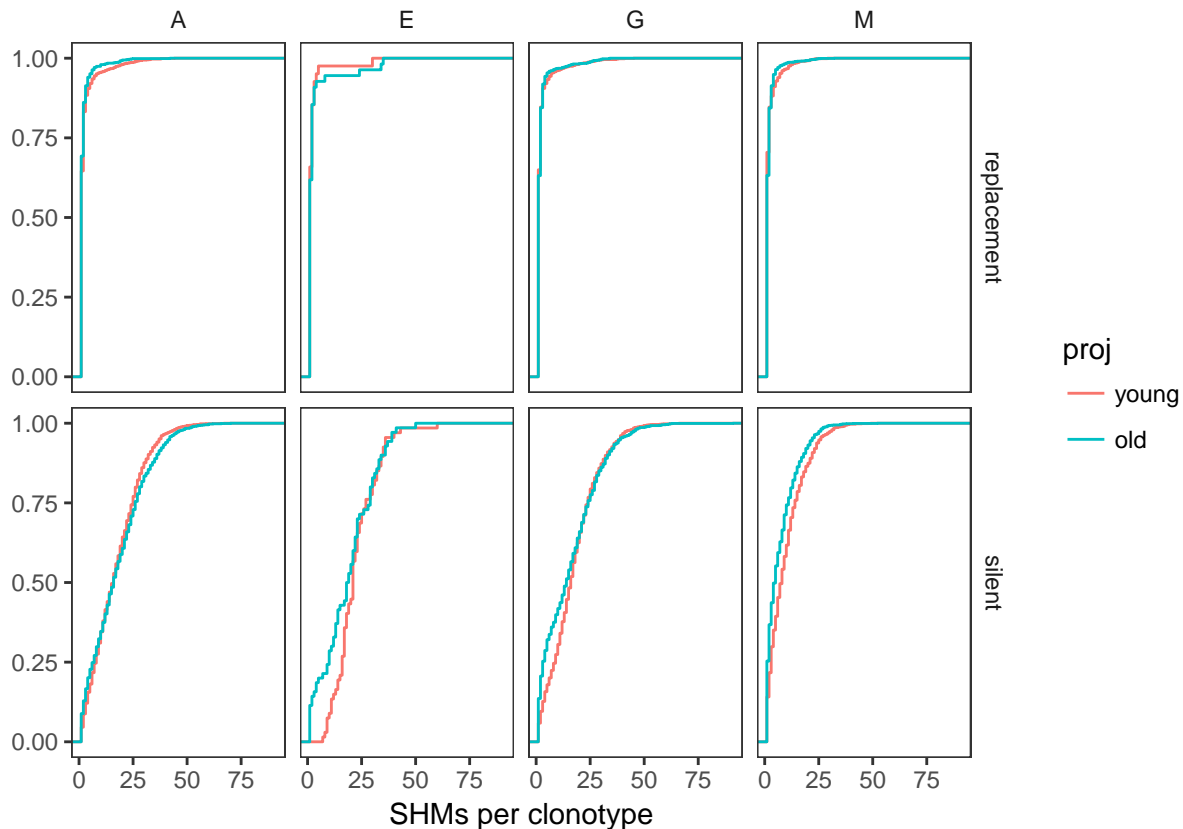
```
a = aov(rs ~ isotype + proj, df.2)
summary(a)
```

```
##          Df  Sum Sq Mean Sq F value    Pr(>F)
## isotype    3  0.09575  0.03192    6.987 0.000834 ***
## proj        1  0.00293  0.00293    0.640 0.428943
## Residuals   35  0.15987  0.00457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(a, "proj")
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = rs ~ isotype + proj, data = df.2)
##
## $proj
##              diff              lwr              upr              p adj
## old-young 0.01710374 -0.02628374 0.06049123 0.4289433
```

```
ggplot(df, aes(x=count, color = proj)) +
  stat_ecdf() +
  facet_grid(replacement~isotype) + #, scales = "free") +
  #scale_fill_brewer("Age", palette = "Set1") +
  xlab("SHMs per clonotype") + ylab("") +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        strip.background = element_blank())
```



```
dt.p = data.table()

for (iso in unique(df$isotype)) {
  for (rr in unique(df$replacement)) {
    tmp = df %>% filter(isotype == iso, replacement == rr)
    x = (tmp %>% filter(proj == "old"))$count
    y = (tmp %>% filter(proj != "old"))$count
    kk = ks.test(x, y)
```

```

p = kk$p.value
dt.p = rbind(dt.p,
             data.table(isotype = iso, replacement = rr, p=p))
}
}

## Warning in ks.test(x, y): cannot compute exact p-value with ties

## Warning in ks.test(x, y): cannot compute exact p-value with ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

## Warning in ks.test(x, y): p-value will be approximate in the presence of
## ties

dt.p$p.adj = p.adjust(dt.p$p, method = "BH")
print(dt.p %>% arrange(p.adj))

##   isotype replacement          p          p.adj
## 1      M      silent 0.000000e+00 0.000000e+00
## 2      G      silent 1.135758e-13 4.543033e-13
## 3      M replacement 9.633436e-04 2.568916e-03
## 4      E      silent 7.219225e-02 1.155076e-01
## 5      A      silent 6.004913e-02 1.155076e-01
## 6      A replacement 1.689919e-01 2.253225e-01
## 7      E replacement 1.000000e+00 1.000000e+00
## 8      G replacement 9.318338e-01 1.000000e+00

```