istoype_ext_analysis.Rmd

```r
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
```

```r
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
## ')
```

```r
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

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

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

```
summarise = dplyr::summarise
```

```
load("shm_rep2.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: 15 x 3
## # Groups:   proj [?]
##    proj  i     count
##    <chr> <chr> <int>
##  1 old   ""        2
##  2 old   IGHA1  1592
##  3 old   IGHD      7
##  4 old   IGHE     82
##  5 old   IGHG1  1799
##  6 old   IGHG3   108
##  7 old   IGHGP     4
##  8 old   IGHM   2497
##  9 young IGHA1  2055
## 10 young IGHD     38
## 11 young IGHE     70
## 12 young IGHG1  1424
## 13 young IGHG3   112
## 14 young IGHGP     8
## 15 young IGHM   1956
```

```
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): cannot compute exact p-value with ties
```

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

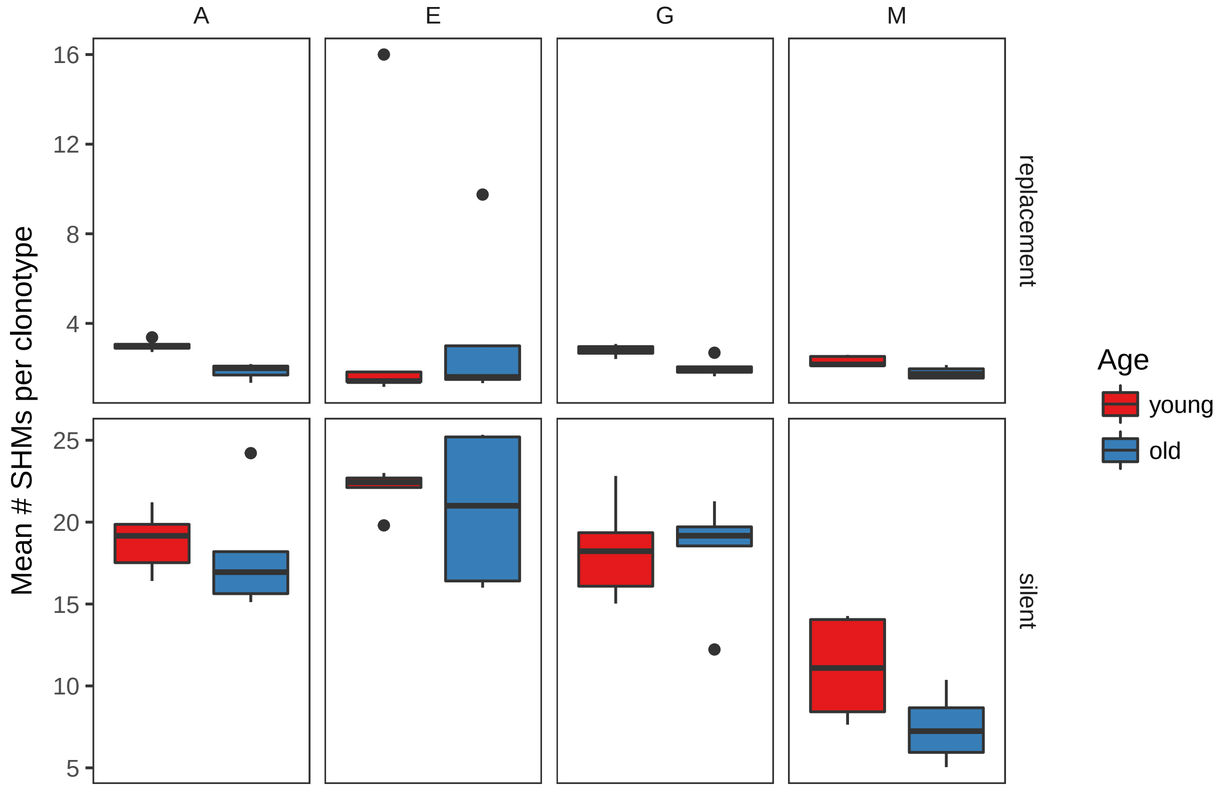
```
##   isotype            p        p.adj
## 1       M 0.000000e+00 0.000000e+00
## 2       G 9.469078e-07 1.893816e-06
## 3       A 8.460576e-06 1.128077e-05
## 4       E 3.761918e-01 3.761918e-01
```

```
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   3998    3998  340.77  < 2e-16 ***
## isotype      3    519     173   14.74 1.28e-07 ***
## proj         1     26      26    2.22    0.141    
## Residuals   74    868      12                     
## ---
## 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 -1.141036 -2.667091 0.3850185 0.1405186
```

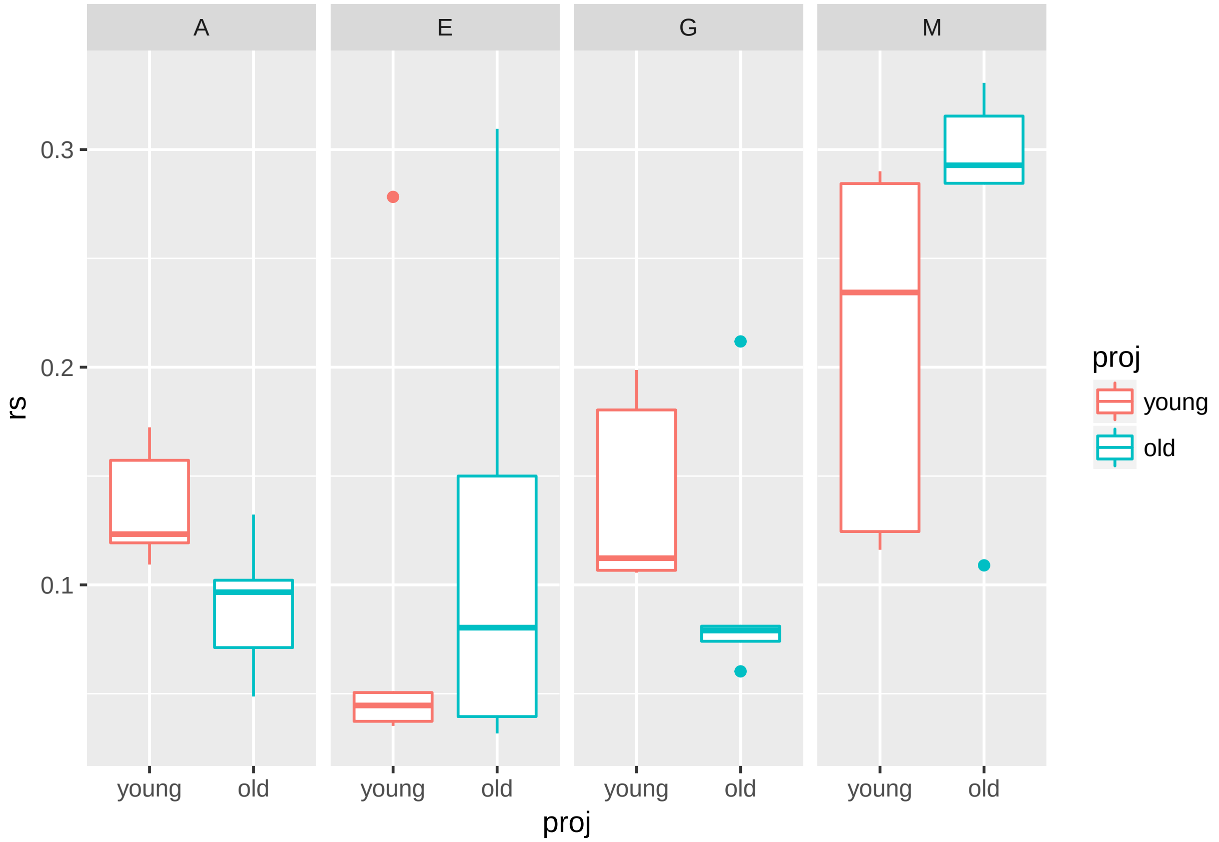
```
TukeyHSD(a, "isotype")
```

```
##   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  2.218124 -0.6287357  5.0649844 0.1801093
## G-A -0.105469 -2.9523290  2.7413911 0.9996663
## M-A -4.773404 -7.6202637 -1.9265436 0.0002010
## G-E -2.323593 -5.1704534  0.5232667 0.1485469
## M-E -6.991528 -9.8383880 -4.1446680 0.0000001
## M-G -4.667935 -7.5147947 -1.8210746 0.0002852
```

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

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



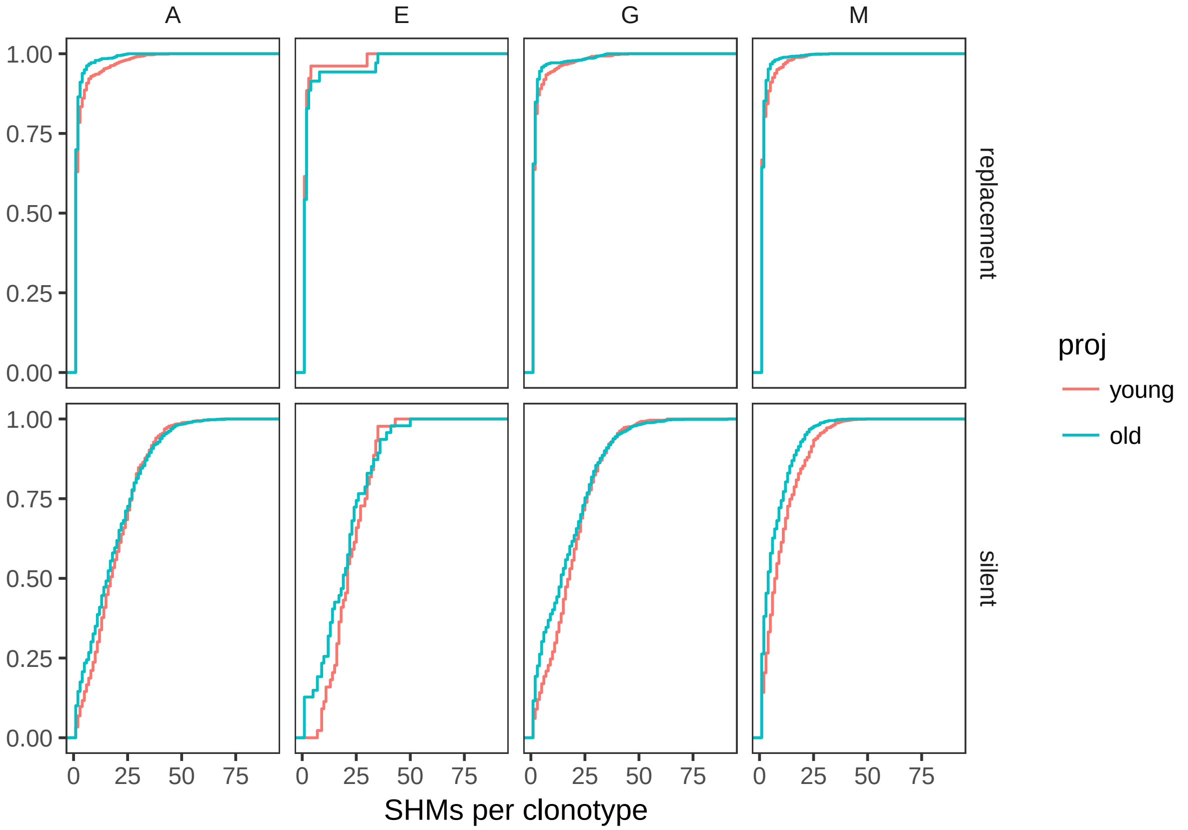
```
a = aov(rs ~ isotype + proj, df.2)
summary(a)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## isotype      3 0.11810 0.03937   6.590 0.0012 **
## proj         1 0.00001 0.00001   0.002 0.9673
## Residuals   35 0.20909 0.00597
## ---
## 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 0.001008596 -0.04861054 0.05062773 0.9673188
```

```
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): 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

## 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
```

```
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 2.236866e-08 8.947463e-08
## 3       A      silent 6.442916e-04 1.718111e-03
## 4       M replacement 5.332661e-03 1.066532e-02
## 5       A replacement 1.033748e-02 1.653996e-02
## 6       G replacement 1.775563e-01 2.367417e-01
## 7       E      silent 3.250029e-01 3.714319e-01
## 8       E replacement 9.999987e-01 9.999987e-01
```