Isotype usage and hypermutation burden analysis

Mikhail Shugay, Anna Obraztsova

11/17/2016

Load raw MIXCR data

```
library(plyr)
library(dplyr)
library(ggplot2)
library(stringr)
library(reshape2)

rna <- data.frame()

old_rna = c("Abdulain", "Ilgen", "Mamaev", "Smirnov", "Vlasov")
young_rna = c("Antipyat", "Epifancev", "Hadjibekov", "Koshkin", "Kovalchuk")

for (sample in old_rna){
  .df <- read.table(paste('data/mixcr_yf_old_RNA/', sample, ".txt.gz", sep = ""), header=T, sep="\t", s
  .df$proj <- "old"
    .df$sample <- sample
    rna <- rbind(rna, .df)
}

for (sample in young_rna){
    .df <- read.table(paste('data/mixcr_yf_young_RNA/', sample, ".txt.gz", sep = ""), header=T, sep="\t
    .df$proj <- "young"
    .df$sample <- sample
    rna <- rbind(rna, .df)
}

new_colnames = c('clone.id','clone.count','clone.fraction','clonal.seq','clonal.seq.qual','all.v.hits',
    'all.j.hits','all.c.hits','all.v.alignments','all.d.alignments','all.j.alignments','all.c.alignment
    'nt.seq.FR1','min.qual.FR1','nt.seq.CDR1','min.qual.CDR1','nt.seq.FR2','min.qual.FR2','nt.seq.CDR2'
    'nt.seq.FR3','min.qual.FR3','nt.seq.CDR3','min.qual.CDR3','nt.seq.FR4','min.qual.FR4','aa.seq.FR1',
    'aa.seq.FR2','aa.seq.CDR2','aa.seq.FR3','aa.seq.CDR3','aa.seq.FR4','ref.points', 'proj', 'sample')

colnames(rna) <- new_colnames
rna <- mutate(rna, isotype = str_sub(all.c.hits, 1, 4))
rna$shm.count <- unlist(lapply(str_split(rna$all.v.alignments, ";"), function(x) str_count(x[1], "S")))
rna <- subset(rna, nchar(isotype) > 0)
rna$isotype <- factor(rna$isotype, levels = c("IGHM", "IGHG", "IGHA", "IGHE", "IGHD"))
```

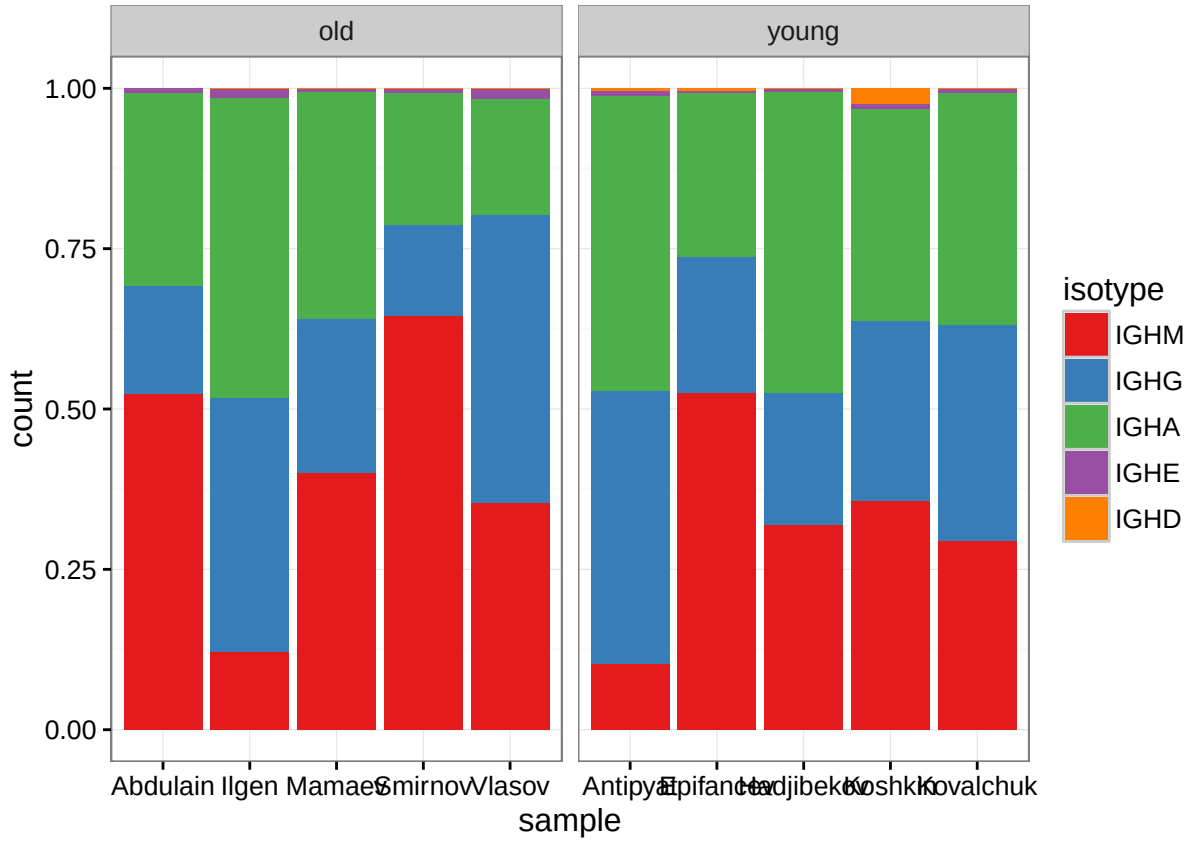
Summarize frequency, diversity and hypermutation type for each isotype

```
rna.2 <- ddply(rna, .(proj, sample, isotype), summarize,
               share = sum(clone.fraction), clonotypes = length(clone.fraction), shm.count = sum(shm.cou
```

Isotype usage, displays high variance across donors, and, hopefully, in line with the usage observed in plasma B cell subset

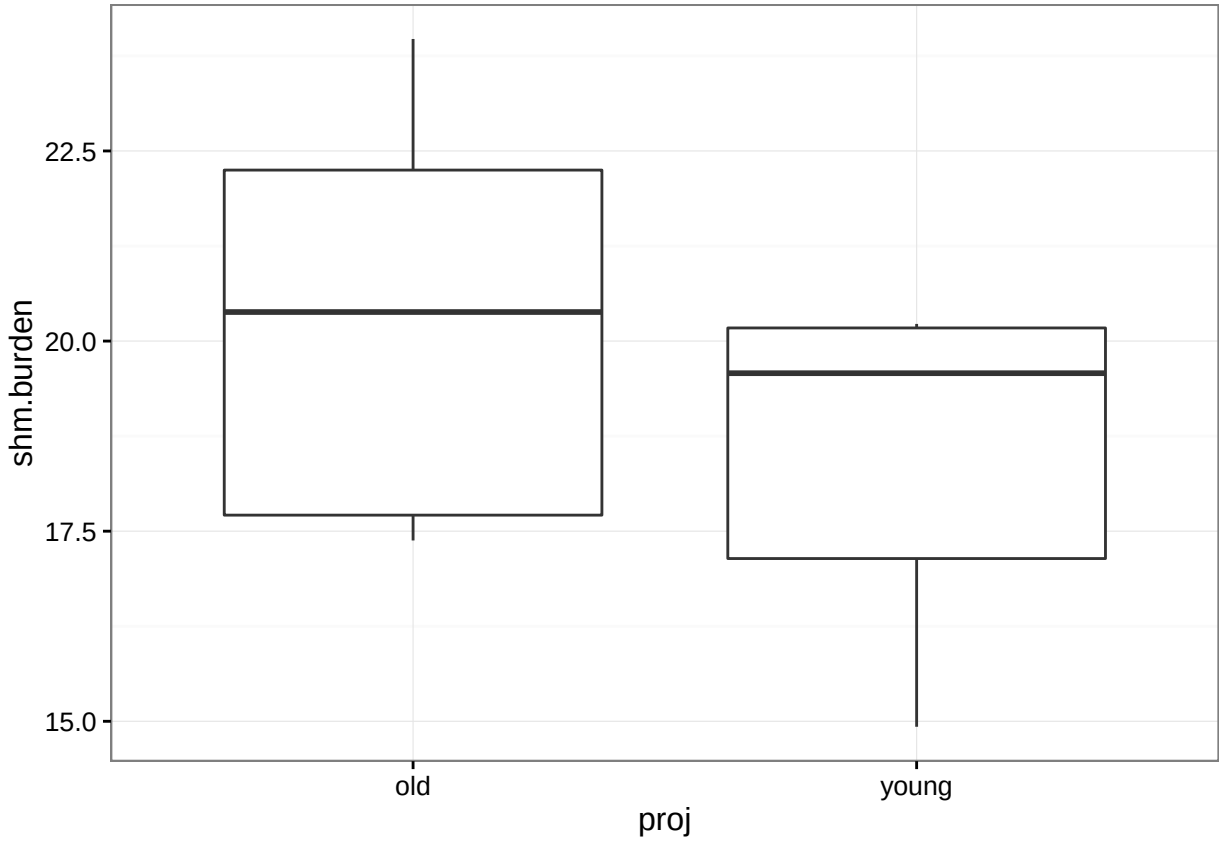
```
ggplot(rna.2, aes(x=sample, weight=share, fill = isotype)) +
  geom_bar() +
```

```
  facet_wrap(~proj, scales = "free_x") +
  scale_fill_brewer(palette = "Set1") +
  theme_bw()
```



Overall mutation burden by donor

```
rna.1 <- ddply(rna.2, .(proj, sample), summarize, shm.burden = sum(shm.count)/sum(clonotypes))

ggplot(rna.1, aes(x=proj,group=proj,y=shm.burden)) +
  geom_boxplot() +
  theme_bw()
```

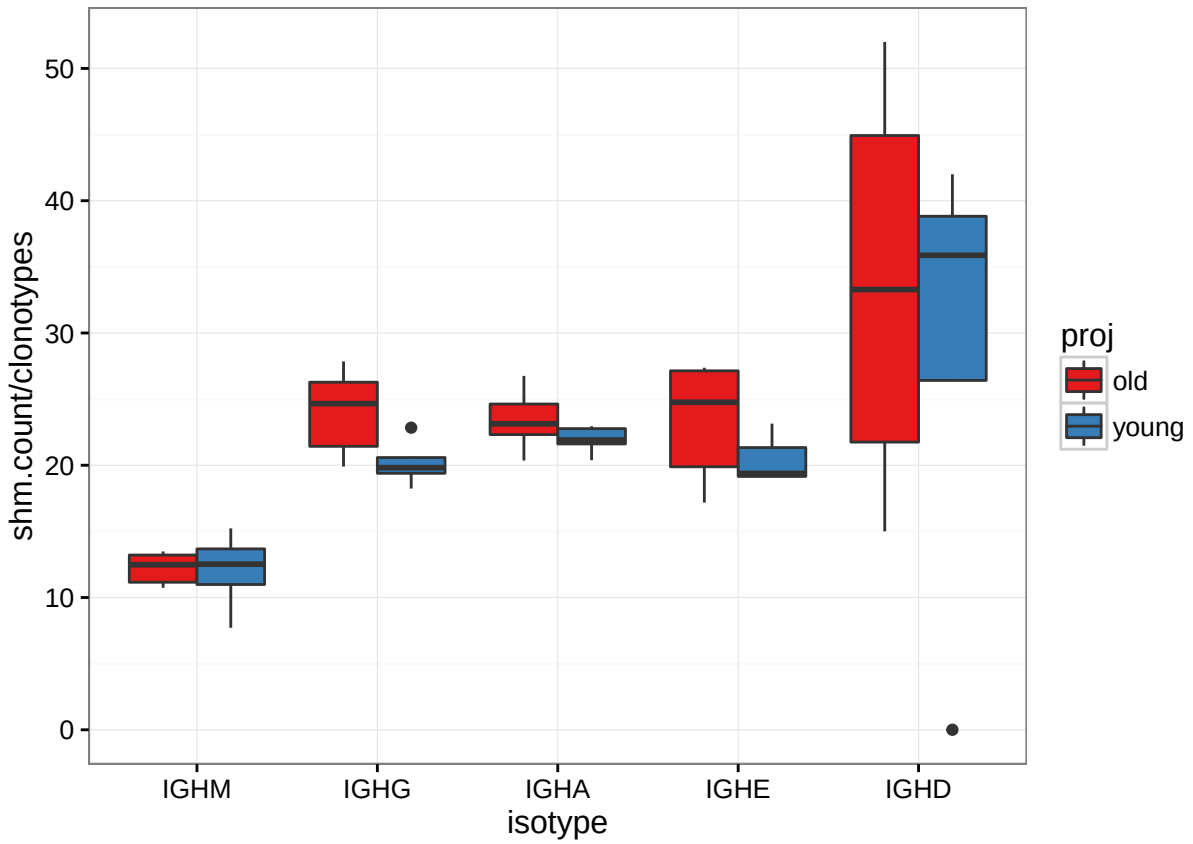



```
t.test(shm.burden ~ proj, rna.1)
```

```
##
##  Welch Two Sample t-test
##
## data:  shm.burden by proj
## t = 1.1741, df = 7.6803, p-value = 0.2755
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.887988  5.747440
## sample estimates:
##   mean in group old mean in group young
##            20.33853            18.40881
```

Mutation burden by isotype, old donors have higher number of hypermutations per clonotypes as demonstrated in next section

```
ggplot(rna.2, aes(x=isotype, group = interaction(isotype, proj), y=shm.count / clonotypes, fill=proj)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Set1") +
  theme_bw()
```




```
a <- aov(shm.count / clonotypes ~ isotype + proj, rna.2)
summary(a)
```

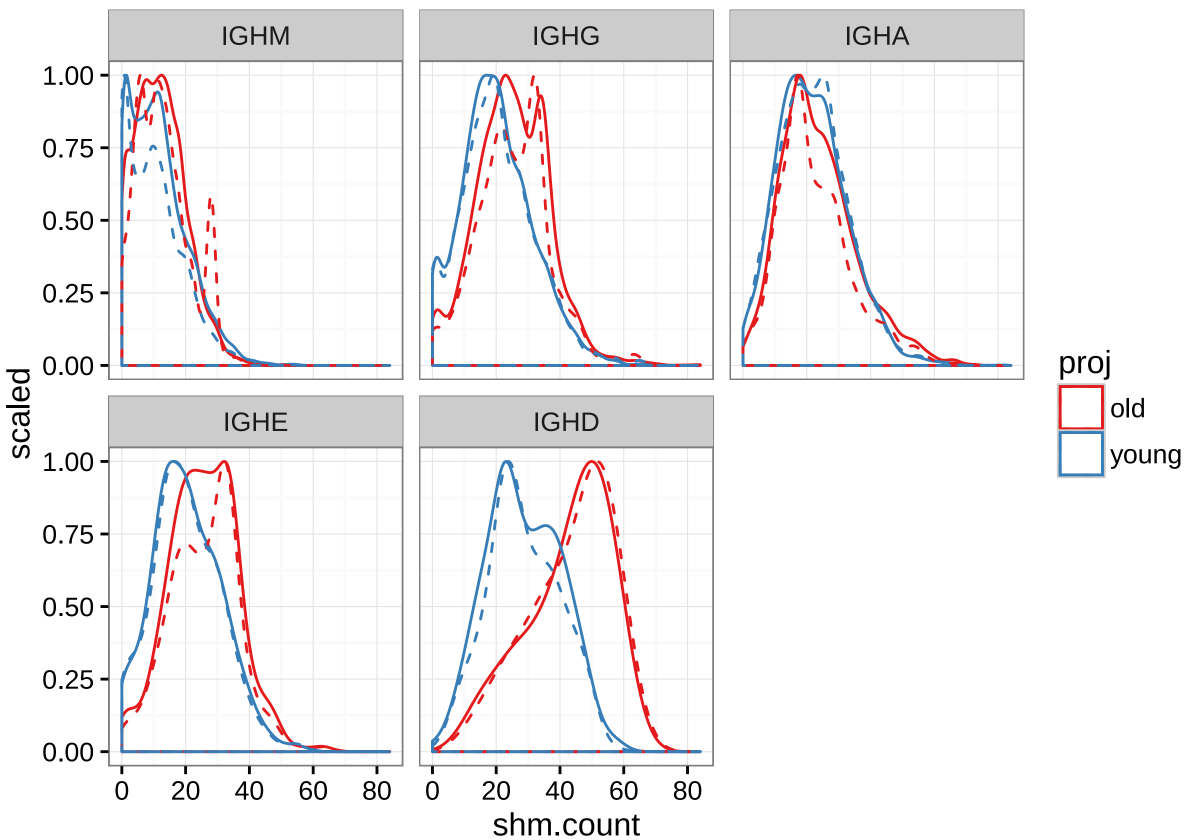
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## isotype      4 1665.4   416.4   7.901 7.24e-05 ***
## proj         1   81.6    81.6   1.549     0.22
## Residuals   43 2266.0    52.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

More detalization, scaled distributions of hypermutation count per clonotype for old and young. Solid lines - unweighted, dashed lines - weighted by clonotype frequency. Old donors have more hypermutation burden, especially for IGHD.

```
ggplot(rna, aes(x=shm.count, color = proj)) +
  geom_density(aes(y=..scaled..), linetype = "solid") +
  geom_density(aes(weight = clone.fraction, y=..scaled..), linetype = "dashed") +
  facet_wrap(~isotype) +
  scale_color_brewer(palette = "Set1") +
  theme_bw()
```

```
## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
```

```
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density

## Warning in density.default(x, weights = w, bw = bw, adjust = adjust, kernel
## = kernel, : sum(weights) != 1 -- will not get true density
```



Just a nice pic showing that we are doing everything right - SHM distribution by isotype, IgM having lowest number of SHMs

```r
rna.3 <- rna
rna.3$isotype <- NULL
```

```
ggplot(rna) +
  geom_density(data=rna.3, aes(x=shm.count), fill="grey", linetype="dashed") +
  geom_density(aes(x=shm.count, color = isotype)) +
  facet_wrap(~isotype) +
  theme_bw()
```