

# Sex

```
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(ggplot2)
library(reshape2)
```

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

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

```
library(scales)
library(parallel)
library(stringr)
library(knitr)
```

Load metadata

```
dt.hip.stats = fread("annotations/hip_stats.txt") %>%
  filter(!is.na(sex)) %>%
  mutate(count_total = count, occurrences_total = diversity) %>%
  select(sample_id, sex, count_total, occurrences_total)
```

Load VDJdb annotations with 1 mismatch for HIP data (time consuming, ~ 2mln clonotypes)

```
dt.hip = rbindlist(mclapply(as.list(dt.hip.stats$sample_id),
  function(x) fread(paste0("annotations/split_1mm/", x, ".annot.txt")) %>%
    mutate(sample_id = x), mc.cores = 40)) %>%
  group_by(sample_id, cdr3) %>%
  summarise(count = sum(count), occurrences = n())
```

VDJdb data

```
dt.vdjdb = fread("rearr_model/VDJDB_fullP_rob_ageing.txt") %>%
  filter(gene == "TRB", mhc.class == "MHCI") %>%
  mutate(hla_spec = str_split_fixed(mhc.a, pattern = "[:,]", 2)[,1]) %>%
  select(cdr3, hla_spec, antigen.epitope, antigen.species) %>%
  group_by(antigen.epitope) %>%
  mutate(unique_cdrs = n()) %>%
```

```
filter(unique_cdrs > 30) %>%
select(cdr3, antigen.epitope, antigen.species, unique_cdrs)
```

Merge

```
dt.hip.m = dt.hip %>%
  merge(dt.hip.stats) %>%
  merge(dt.vdjdbc)
```

Summarise by sex

```
dt.hip.s = as.data.table(dt.hip.m) %>%
  group_by(sample_id, sex, antigen.epitope, antigen.species, unique_cdrs) %>%
  summarise(occurrences = sum(occurrences)) %>%
  group_by(sample_id) %>%
  mutate(occurrences_share = occurrences / sum(occurrences) / unique_cdrs)
```

```
dt.p = data.table(antigen.epitope = unique(dt.hip.s$antigen.epitope), p = 1, freq.ratio = 1) %>%
  merge(dt.hip.s %>% ungroup %>% select(antigen.species, antigen.epitope) %>% unique)
```

```
for (i in 1:nrow(dt.p)) {
  tmp = dt.hip.s %>% filter(antigen.epitope == dt.p$antigen.epitope[i])
  dt.p$freq.ratio[i] = with(tmp, mean(occurrences_share[which(sex=="male")]) / mean(occurrences_share[w
  dt.p$p[i] = t.test(occurrences_share ~ sex, tmp)$p.value
}
```

```
dt.p$p = p.adjust(dt.p$p, method = "BH")
dt.p$len = nchar(as.character(dt.p$antigen.epitope))
```

```
kable(dt.p %>% arrange(p))
```

| antigen.epitope | p         | freq.ratio | antigen.species  | len |
|-----------------|-----------|------------|------------------|-----|
| KAFSPEVIPMF     | 0.0003249 | 0.9633999  | HIV-1            | 11  |
| NLVPMTATV       | 0.0217787 | 0.9872201  | CMV              | 9   |
| RAKFKQLL        | 0.0614911 | 1.0185100  | EBV              | 8   |
| ATDALMTGY       | 0.0691821 | 1.0262347  | HCV              | 9   |
| KRWIHLGLNK      | 0.1478753 | 1.0110325  | HIV-1            | 10  |
| TPQDLNMTL       | 0.1478753 | 0.9827688  | HIV-1            | 9   |
| HSKKKCDEL       | 0.2400929 | 1.0502778  | HCV              | 9   |
| EIYKRWH         | 0.3729144 | 0.9770035  | HIV-1            | 8   |
| FPRPWLHGL       | 0.4941968 | 1.0188903  | HIV-1            | 9   |
| GILGFVFTL       | 0.4969108 | 1.0073121  | InfluenzaA       | 9   |
| HPKVSSEVHI      | 0.4969108 | 0.9822849  | HIV-1            | 10  |
| IPSINVHHY       | 0.4969108 | 0.9845237  | CMV              | 9   |
| YVLDHLIVV       | 0.4969108 | 0.9883074  | EBV              | 9   |
| CINGVCWTV       | 0.6009167 | 0.9858520  | HCV              | 9   |
| KRWIIMGLNK      | 0.6009167 | 1.0125451  | HIV-1            | 10  |
| LLWNGPMAV       | 0.6009167 | 1.0056339  | YellowFeverVirus | 9   |
| LPRRSGAAGA      | 0.6220877 | 1.0048834  | InfluenzaA       | 10  |
| RPRGEVRFL       | 0.6220877 | 0.9822155  | HSV-2            | 9   |
| AMFWSVPTV       | 0.6506679 | 1.0081599  | HomoSapiens      | 9   |
| GLCTLVAML       | 0.6506679 | 1.0036908  | EBV              | 9   |
| SLYNTVATL       | 0.6648368 | 0.9871165  | HIV-1            | 9   |
| EPLPQGQLTAY     | 0.6868217 | 1.0329359  | EBV              | 11  |
| LPPIVAKEI       | 0.7270800 | 1.0076104  | HIV-1            | 9   |

| antigen.epitope | p         | freq.ratio | antigen.species | len |
|-----------------|-----------|------------|-----------------|-----|
| KLVALGINAV      | 0.7700850 | 0.9959509  | HCV             | 10  |
| ISPRTLNAW       | 0.8539696 | 0.9942793  | HIV-1           | 9   |
| GPGHKARVL       | 0.8609838 | 0.9946194  | HIV-1           | 9   |
| IIKDYGKQM       | 0.8609838 | 0.9921577  | HIV-1           | 9   |
| TPRVTGGGAM      | 0.8609838 | 1.0050668  | CMV             | 10  |
| ELAGIGILTV      | 0.9125300 | 1.0004072  | HomoSapiens     | 10  |
| FLKEKGGL        | 0.9125300 | 0.9977665  | HIV-1           | 8   |
| TPGPGVRYPL      | 0.9125300 | 1.0021031  | HIV-1           | 10  |
| VTEHDTLLY       | 0.9125300 | 1.0009461  | CMV             | 9   |
| FLYNLLTRV       | 0.9877709 | 1.0001565  | HomoSapiens     | 9   |
| LLLIGILV        | 0.9877709 | 0.9999372  | HomoSapiens     | 9   |

```

good_epi = (dt.p %>% filter(p < 0.05))$antigen.epitope
dt.hip.s = dt.hip.s %>%
  mutate(antigen.epitope = ifelse(antigen.epitope %in% good_epi, paste(antigen.epitope, "(*)"), antigen.epitope))

dt.hip.s.s = dt.hip.s %>%
  #filter(ucb == T) %>%
  group_by(antigen.epitope) %>%
  summarise(freq = mean(occurrences_share[which(sex == "male")]))

dt.hip.s$antigen.epitope = factor(dt.hip.s$antigen.epitope,
  levels = dt.hip.s.s$antigen.epitope[order(dt.hip.s.s$freq)])

ggplot(dt.hip.s, aes(x = antigen.epitope, group = paste(antigen.epitope, sex),
  fill = sex, color = sex,
  y = occurrences_share)) +
  geom_boxplot(color = "black") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1") +
  scale_color_brewer(palette = "Set1") +
  xlab("") + scale_y_continuous("Share of annotated rearrangements",
    expand = c(0,0)) +
  theme_bw() +
  theme(aspect = 1.1,
    panel.grid.major = element_blank(), panel.grid.minor = element_blank())

```

