UCB

```
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)
```

```
dt.aging.stats = fread("annotations/aging_stats.txt") %>%
  mutate(count_total = count, occurrences_total = diversity, ucb = age == 0) %>%
  select(sample_id, ucb, count_total, occurrences_total)
```

Load VDJdb annotations with 1 mismatch for aging data

```
dt.aging = rbindlist(mclapply(as.list(dt.aging.stats$sample_id),
                   function(x) fread(paste0("annotations/aging_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, hla_spec, antigen.epitope, antigen.species, unique_cdrs)
```

Merge datasets

```
dt.aging.m = dt.aging %>%
  merge(dt.vdjdb) %>%
  merge(dt.aging.stats)
```

Summarise by epitope

```
dt.aging.s = dt.aging.m %>%
  group_by(ucb, antigen.epitope, antigen.species, unique_cdrs) %>%
  summarise(occurrences = sum(occurrences) / unique_cdrs[1]) %>%
  merge(dt.aging.stats %>%
          group_by(ucb) %>%
          summarise(occurrences_total = sum(as.numeric(occurrences_total))))
```

Distribution of epitopes in UCB and PBMC samples

```
fancy_scientific = function(l) {
     # turn in to character string in scientific notation
     l = format(l, scientific = TRUE)
     # quote the part before the exponent to keep all the digits
     l = gsub("^(.*)e", "'\\1'e", l)
     # turn the 'e+' into plotmath format
     l = gsub("e", "%*%10^", l)
     # return this as an expression
     parse(text=l)
}

dt.aging.s.s = dt.aging.s %>%
  filter(ucb == T) %>%
  group_by(antigen.epitope) %>%
  summarise(freq = sum(occurrences / unique_cdrs) / sum(occurrences_total))

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

tmp = dt.aging.s %>%
  group_by(ucb, antigen.epitope) %>%
  summarise(freq = sum(occurrences / unique_cdrs) / sum(occurrences_total)) %>%
  dcast(antigen.epitope~ucb, value.var= "freq")
freq.ratios = tmp[,3] / tmp[,2]
m=mean(freq.ratios)
ci = qnorm(0.975)*sd(freq.ratios)/sqrt(length(freq.ratios))
paste(round(m,2), round(m-ci,2), round(m+ci,2))
```

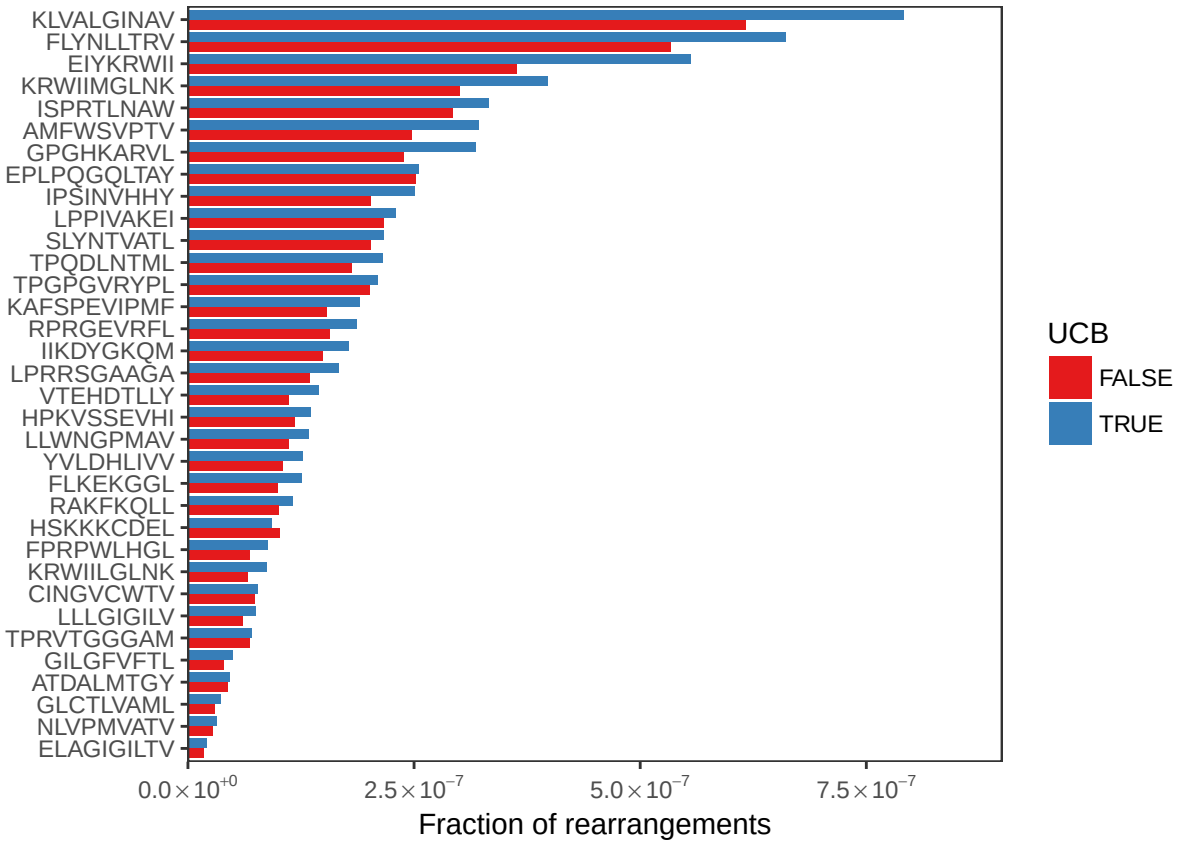
```
## [1] "1.2 1.16 1.24"
```

```
t.test(occurrences / unique_cdrs / occurrences_total ~ ucb, dt.aging.s, paired=T)
```

```
##
##  Paired t-test
##
## data:  occurrences/unique_cdrs/occurrences_total by ucb
## t = -4.5646, df = 33, p-value = 6.613e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.323222e-08 -2.040925e-08
```

```
## sample estimates:
## mean of the differences
##           -3.682073e-08
```

```
p17=ggplot(dt.aging.s, aes(x = antigen.epitope, fill = ucb,
                           y = occurrences / unique_cdrs / occurrences_total)) +
  geom_bar(stat="identity", position = "dodge") +
  coord_flip() +
  scale_fill_brewer("UCB", palette = "Set1") +
  xlab("") + scale_y_continuous("Fraction of rearrangements", labels = fancy_scientific,
                                expand = c(0,0), limits = c(0, 9e-7)) +
  theme_bw()  +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
p17
```



Evenness of epitope-specific TCR occurrences

```
dt.aging.s2 = as.data.table(dt.aging.m) %>%
  group_by(sample_id, ucb, 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.aging.s2$antigen.epitope), p = 1, freq.ratio = 1) %>%
  merge(dt.aging.s2 %>% ungroup %>% select(antigen.species, antigen.epitope) %>% unique)
```

```
for (i in 1:nrow(dt.p)) {
  tmp = dt.aging.s2 %>% filter(antigen.epitope == dt.p$antigen.epitope[i])
  dt.p$freq.ratio[i] = with(tmp, mean(occurrences_share[which(ucb)]) / mean(occurrences_share[which(!ucb
  dt.p$p[i] = t.test(occurrences_share ~ ucb, tmp)$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
ATDALMTGY	0.0034813	0.8779157	HCV	9
TPGPGVRYPL	0.0034813	0.8452927	HIV-1	10
TPRVTGGGAM	0.0049082	0.8316385	CMV	10
GLCTLVAML	0.0064975	0.9659187	EBV	9
SLYNTVATL	0.0122898	0.8943390	HIV-1	9
ISPRTLNAW	0.0123381	0.9214290	HIV-1	9
AMFWSVPTV	0.0472048	1.0644062	HomoSapiens	9
EPLPQGQLTAY	0.0657514	0.8396878	EBV	11
EIYKRWII	0.0677543	1.1979569	HIV-1	8
RAKFKQLL	0.0841447	0.9341371	EBV	8
CINGVCWTV	0.1101124	0.8749749	HCV	9
KRWIILGLNK	0.1370102	1.1179089	HIV-1	10
NLVPMVATV	0.1370102	0.9584504	CMV	9
LPPIVAKEI	0.1419204	0.9029855	HIV-1	9
HSKKKCDEL	0.1768884	0.7849616	HCV	9
VTEHDTLLY	0.1797523	1.0544656	CMV	9
GILGFVFTL	0.1993188	1.0350935	InfluenzaA	9
FPRPWLHGL	0.3332847	1.0596474	HIV-1	9
FLKEKGGL	0.3585158	1.0773967	HIV-1	8
IIKDYGKQM	0.3987301	0.9466931	HIV-1	9
LLWNGPMAV	0.3987301	0.9718025	YellowFeverVirus	9
IPSINVHHY	0.4121486	1.0608830	CMV	9
KLVALGINAV	0.4121486	1.0260608	HCV	10
KRWIIMGLNK	0.4121486	1.0562421	HIV-1	10
LLLGIGILV	0.4219274	1.0190819	HomoSapiens	9
FLYNLLTRV	0.6465261	1.0311266	HomoSapiens	9
GPGHKARVL	0.6465261	1.0431526	HIV-1	9
KAFSPEVIPMF	0.6465261	0.9601713	HIV-1	11
RPRGEVRFL	0.6465261	0.9606192	HSV-2	9
TPQDLNTML	0.6747572	1.0231133	HIV-1	9
YVLDHLIVV	0.6747572	0.9804393	EBV	9
HPKVSSEVHI	0.9312795	1.0086373	HIV-1	10
ELAGIGILTV	0.9868604	0.9993951	HomoSapiens	10
LPRRSGAAGA	0.9880312	0.9996816	InfluenzaA	10

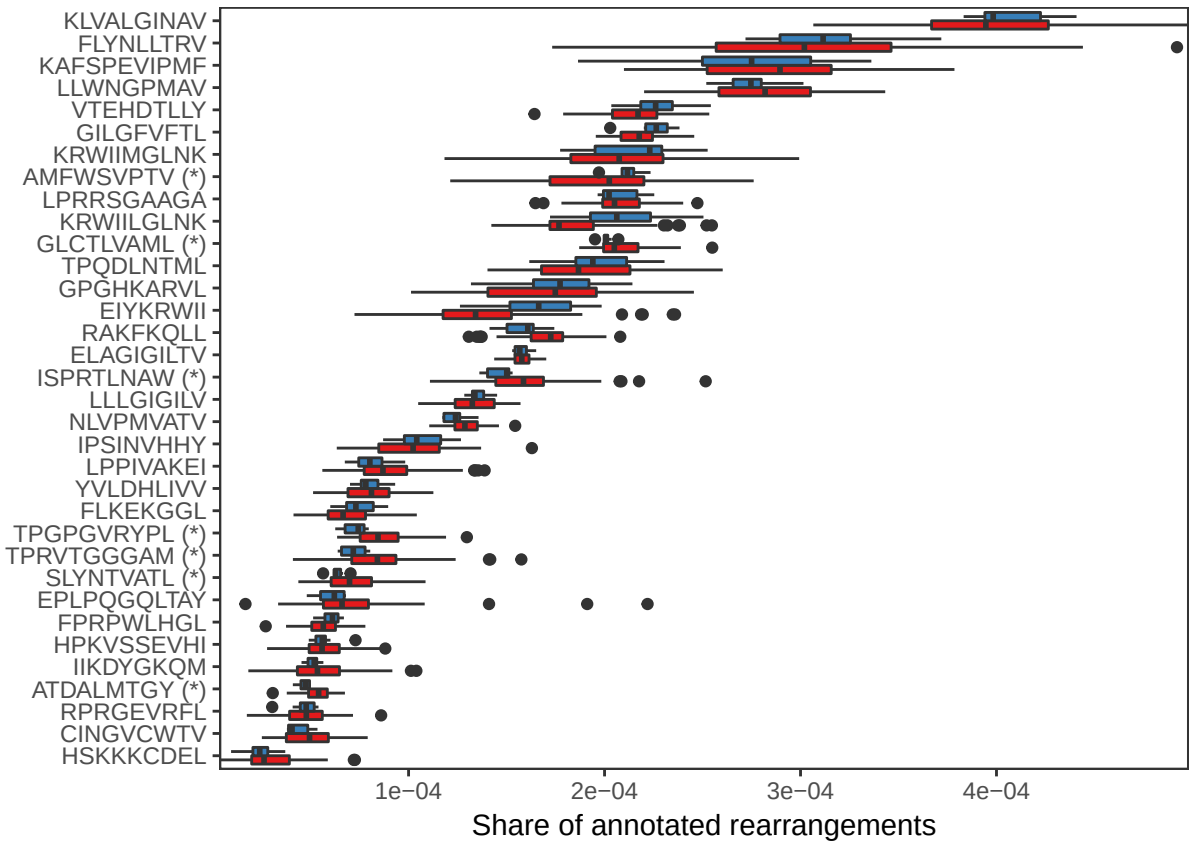
```
good_epi = (dt.p %>% filter(p < 0.05))$antigen.epitope
dt.aging.s2 = dt.aging.s2 %>%
  mutate(antigen.epitope = ifelse(antigen.epitope %in% good_epi, paste(antigen.epitope, "(*)"),antigen.
```

```
dt.aging.s2.s = dt.aging.s2 %>%
  #filter(ucb == T) %>%
  group_by(antigen.epitope) %>%
  summarise(freq = mean(occurrences_share[which(ucb)]))

dt.aging.s2$antigen.epitope = factor(dt.aging.s2$antigen.epitope,
                                     levels = dt.aging.s2.s$antigen.epitope[order(dt.aging.s2.s$freq)])

dt.aging.s2$epi.len = nchar(as.character(dt.aging.s2$antigen.epitope))

p18=ggplot(dt.aging.s2, aes(x = antigen.epitope, group = paste(antigen.epitope,ucb),
                            fill = ucb,
                           y = occurrences_share)) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_brewer(guide = F, palette = "Set1") +
  xlab("") + scale_y_continuous("Share of annotated rearrangements",
                                expand = c(0,0)) +
  theme_bw()  +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
p18
```




```
ggsave("figures/p17.pdf", p17)
```

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

```r
ggsave("figures/p18.pdf", p18)
```

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