Race

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

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

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

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

Load metadata

```r
dt.hip.stats = fread("annotations/hip_stats.txt") %>%
  mutate(race = ifelse(is.na(race), "Unknown,Unknown", race)) %>%
  mutate(count_total = count, occurrences_total = diversity) %>%
  select(sample_id, race)

# split race and origin

tmp = str_split_fixed(dt.hip.stats$race, ",", n = 2)
tmp[,2] = ifelse(tmp[,2] == "", tmp[,1], tmp[,2])
tmp[,1] = ifelse(tmp[,2] == tmp[,1], "Unknown", tmp[,1])
dt.hip.stats$race = tmp[,1]

dt.hip.stats = dt.hip.stats %>% filter(race != "Unknown")

summary(as.factor(dt.hip.stats$race))
```

```
## asian or pacific islander black or african american
##                        47                        12
##                 caucasian
##                       465
```

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.vdjdb)
```

Summarise by race

```
dt.hip.s = as.data.table(dt.hip.m) %>%
  group_by(sample_id, race, 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_c = 1, freq.rati
  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$antigen.epitope[i])
  tmp$race = as.factor(tmp$race)
  dt.p$freq.ratio_c[i] = with(tmp, mean(occurrences_share[which(race=="caucasian")]) / mean(occurrences_
  dt.p$freq.ratio_b[i] = with(tmp, mean(occurrences_share[which(race=="black or african american")]) / m
  dt.p$freq.ratio_a[i] = with(tmp, mean(occurrences_share[which(race=="asian or pacific islander")]) / m
  a = aov(occurrences_share ~ race, tmp)
  dt.p$p[i] = summary(a)[[1]][["Pr(>F)"]][1]
  #dt.p$p[i] = kruskal.test(occurrences_share ~ race, 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_c	freq.ratio_b	freq.ratio_a	antigen.species	len
GPGHKARVL	0.0031257	0.9928759	0.8924625	1.0979395	HIV-1	9
KLVALGINAV	0.0031257	0.9941108	1.0160172	1.0541764	HCV	10
KRWIILGLNK	0.0031257	0.9960243	1.0080219	1.0372860	HIV-1	10

antigen.epitope	p	freq.ratio_c	freq.ratio_b	freq.ratio_a	antigen.species	len
TPRVTGGGAM	0.0031257	1.0086070	0.9828783	0.9192171	CMV	10
KAFSPEVIPMF	0.0062588	1.0065380	0.9696819	0.9430560	HIV-1	11
EIYKRWII	0.0091940	0.9898537	1.1061235	1.0732886	HIV-1	8
HSKKKCDEL	0.0091940	1.0181298	0.9496281	0.8342629	HCV	9
VTEHDTLLY	0.0115844	0.9977734	0.9667804	1.0305104	CMV	9
LPPIVAKEI	0.0157459	0.9934615	1.1069874	1.0373736	HIV-1	9
FLYNLLTRV	0.0737289	1.0047088	0.9742584	0.9599855	HomoSapiens	9
LPRRSGAAGA	0.1541243	0.9983015	0.9824625	1.0212824	InfluenzaA	10
CINGVCWTV	0.1554756	1.0062263	0.9397990	0.9537703	HCV	9
ISPRTLNAW	0.1554756	0.9941388	1.0751480	1.0388022	HIV-1	9
NLVPMVATV	0.1727873	1.0020015	0.9835863	0.9843888	CMV	9
FPRPWLHGL	0.2273192	1.0035115	1.0445965	0.9539470	HIV-1	9
KRWIIMGLNK	0.3251211	0.9951243	1.0289588	1.0408442	HIV-1	10
GILGFVFTL	0.3965058	1.0016895	1.0024012	0.9826721	InfluenzaA	9
LLLGIGILV	0.4854929	0.9985625	1.0167356	1.0099488	HomoSapiens	9
AMFWSVPTV	0.7671592	0.9976709	1.0112606	1.0201682	HomoSapiens	9
ATDALMTGY	0.9484594	1.0022560	0.9810788	0.9825108	HCV	9
EPLPQGQLTAY	0.9557428	1.0033693	1.0138007	0.9632132	EBV	11
FLKEKGGL	0.9557428	0.9994419	1.0208198	1.0002058	HIV-1	8
GLCTLVAML	0.9557428	0.9993633	1.0113999	1.0033884	EBV	9
HPKVSSEVHI	0.9557428	1.0015663	1.0152859	0.9806006	HIV-1	10
IIKDYGKQM	0.9557428	1.0003727	1.0325668	0.9879974	HIV-1	9
IPSINVHHY	0.9557428	0.9991849	0.9819760	1.0126657	CMV	9
LLWNGPMAV	0.9557428	0.9992480	1.0128237	1.0041655	YellowFeverVirus	9
RAKFKQLL	0.9557428	1.0009982	0.9847979	0.9940055	EBV	8
RPRGEVRFL	0.9557428	1.0008575	1.0490814	0.9790218	HSV-2	9
TPGPGVRYPL	0.9557428	1.0016117	0.9961808	0.9850295	HIV-1	10
YVLDHLIVV	0.9557428	0.9987287	1.0175173	1.0081052	EBV	9
ELAGIGILTV	0.9870871	0.9999765	1.0013629	0.9998846	HomoSapiens	10
SLYNTVATL	0.9870871	1.0000705	1.0100585	0.9967363	HIV-1	9
TPQDLNTML	0.9870871	1.0002085	1.0046113	0.9967597	HIV-1	9

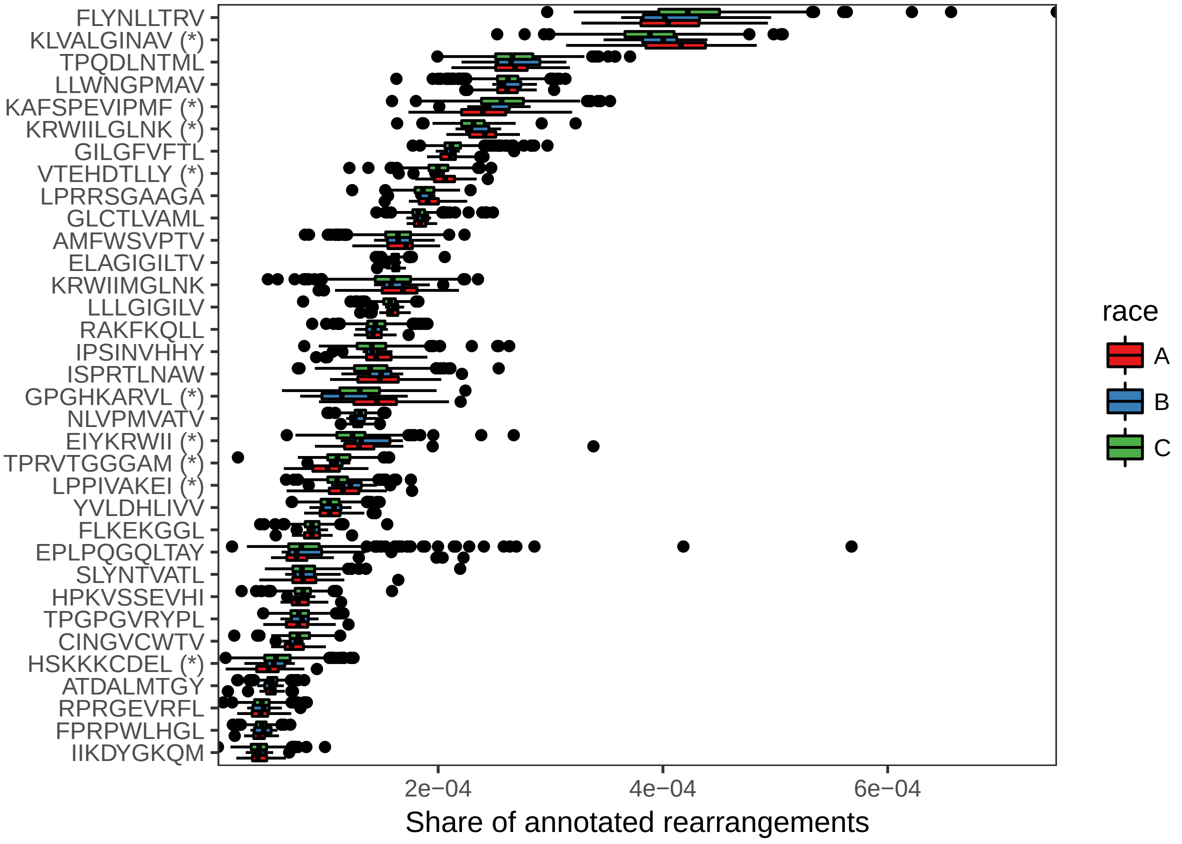
```
good_epi = (dt.p %>% filter(p < 0.05))$antigen.epitope
dt.hip.s = dt.hip.s %>%
  mutate(antigen.epitope = ifelse(antigen.epitope %in% good_epi, paste(antigen.epitope, "(*)"),antigen.
```

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

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

dt.hip.s$race = toupper(substr(as.character(dt.hip.s$race), 1, 1))

p19=ggplot(dt.hip.s, aes(x = antigen.epitope, group = paste(antigen.epitope,race),
                        fill = race,
                       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(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
p19
```



```
ggsave("figures/p19.pdf", p19)
```

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