Epitope length

```
library(data.table)
library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## ------------------------------------------------------------------------------
```

```
##
## 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(stringr)
```

```
df.raw = fread("rearr_model/VDJDB_fullP_rob_ageing.txt") %>%
  filter(mhc.class == "MHCI", species == "HomoSapiens", gene == "TRB") %>%
  mutate(epi.len = nchar(as.character(antigen.epitope)),
         cdr3.len = nchar(as.character(cdr3)),
         hla = str_split_fixed(mhc.a, fixed("*"), 10)[,1]) %>%
  select(species, gene, cdr3, v.segm, j.segm, hla, cdr3.len, epi.len, cdr3.len, antigen.epitope, genP_1
  unique

df.summary = df.raw %>%
  group_by(species, antigen.epitope, hla, epi.len, cdr3.len) %>%
  dplyr::summarise(count = n()) %>%
  arrange(-count)

head(df.summary)
```

```
## Source: local data frame [6 x 6]
## Groups: species, antigen.epitope, hla, epi.len [3]
##
```

```
##       species antigen.epitope   hla epi.len cdr3.len count
##         <chr>           <chr> <chr>   <int>    <int> <int>
## 1 HomoSapiens      ELAGIGILTV HLA-A      10       14   251
## 2 HomoSapiens       GILGFVFTL HLA-A       9       13   227
## 3 HomoSapiens      ELAGIGILTV HLA-A      10       15   197
## 4 HomoSapiens       GLCTLVAML HLA-A       9       13   176
## 5 HomoSapiens      ELAGIGILTV HLA-A      10       13   163
## 6 HomoSapiens       GLCTLVAML HLA-A       9       15   159
```

```
mean(df.summary$count)
```

```
## [1] 11.77612
```

```
print(df.summary %>%
  group_by(species, epi.len, hla) %>%
  dplyr::summarise(count = sum(count), n.epi = length(unique(antigen.epitope)), cdr3.len.mean=mean(cdr3
```

```
## Source: local data frame [12 x 6]
## Groups: species, epi.len [7]
##
##        species epi.len   hla count n.epi cdr3.len.mean
##          <chr>   <int> <chr> <int> <int>         <dbl>
## 1  HomoSapiens       9 HLA-A  3168    41      14.91787
## 2  HomoSapiens      10 HLA-A  1065    11      15.20000
## 3  HomoSapiens      10 HLA-B   751     7      14.69231
## 4  HomoSapiens       9 HLA-B   713    23      14.47101
## 5  HomoSapiens       8 HLA-B   331     8      15.34884
## 6  HomoSapiens      11 HLA-B   252     6      14.42857
## 7  HomoSapiens      13 HLA-B    13     9      13.81818
## 8  HomoSapiens      11 HLA-A     9     1      14.25000
## 9  HomoSapiens      15 HLA-A     6     3      13.40000
## 10 HomoSapiens      12 HLA-B     2     1      14.00000
## 11 HomoSapiens       8 HLA-A     1     1      16.00000
## 12 HomoSapiens       9 HLA-E     1     1      12.00000
```

```
df = df.raw %>% filter(epi.len %in% 8:11)
```

Comparing length distributions

```
df.s = df %>%
  group_by(epi.len) %>%
  summarise(cdr3.len.m = mean(cdr3.len))

p8=ggplot(df,
       aes(x=cdr3.len, group=epi.len, fill = as.factor(epi.len))) +
  #geom_histogram(binwidth = 1, aes(y=..density..), color = "black") +
  geom_area(binwidth = 1, aes(y = ..density..), stat = "bin", position = "stack", color = "black") +
  geom_vline(data=df.s, aes(xintercept = cdr3.len.m), linetype = "dashed") +
  #geom_freqpoly(binwidth = 1, aes(y=..density..), color = "black", position = "stack") +
  scale_x_continuous("CDR3 length", limits = c(7.5,21.5), breaks = seq(8,22,by=3)) + ylab("Fraction of
  scale_fill_brewer("Epitope length", palette = "RdBu") +
  facet_grid(epi.len~.) +
  theme_bw() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        strip.background = element_blank(),
        strip.text.y = element_blank())
```

```
kruskal.test(cdr3.len ~ epi.len, df)
```

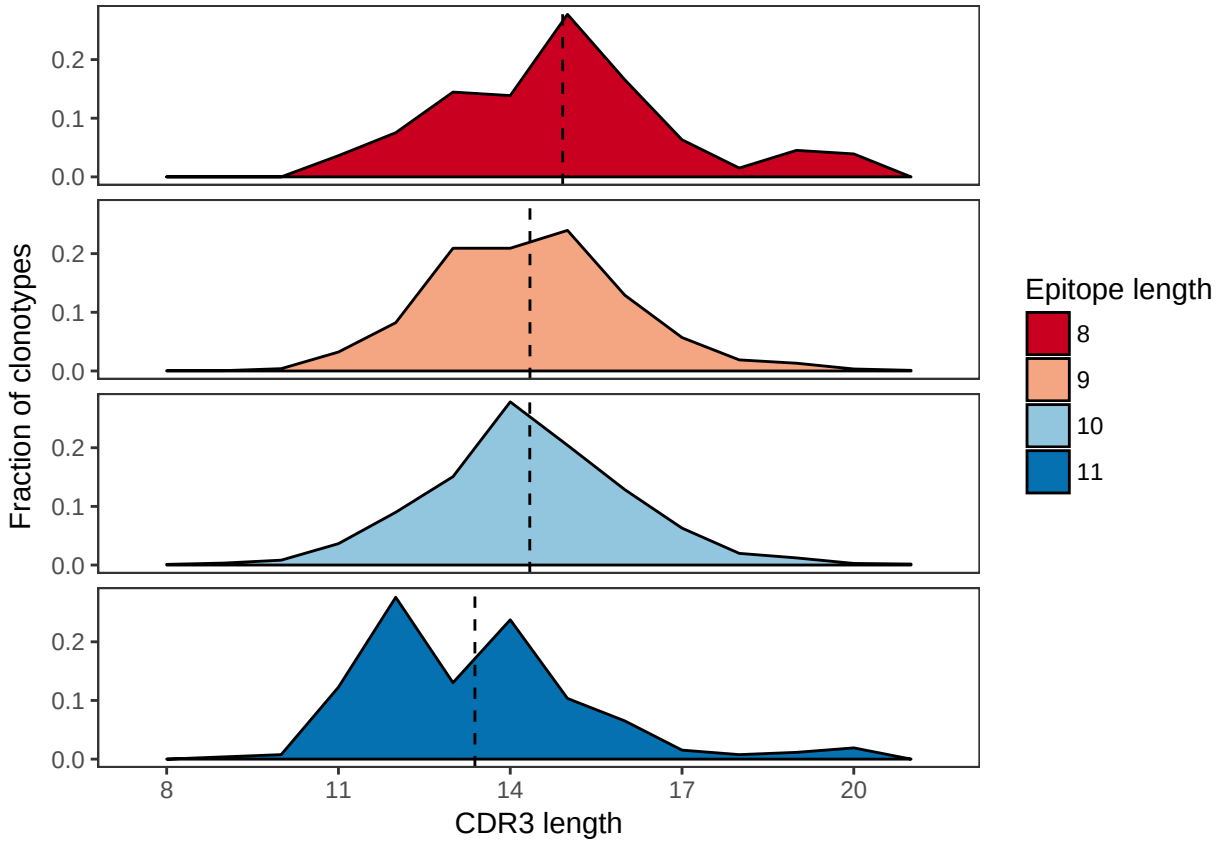
```
##
##  Kruskal-Wallis rank sum test
##
## data:  cdr3.len by epi.len
## Kruskal-Wallis chi-squared = 113.34, df = 3, p-value < 2.2e-16
```

```
summary(aov(cdr3.len ~ epi.len, df))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## epi.len        1    179  179.20   56.57 6.19e-14 ***
## Residuals   6289  19923    3.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p8
```

```
## Warning: Removed 9 rows containing non-finite values (stat_bin).
```




```
p9=ggplot(df.raw %>% filter(nchar(cdr3) %in% 7:22), aes(x = nchar(cdr3),  y = genP_1mism_rob)) +
  stat_density_2d(geom = "tile", aes(fill=..density..), contour = F) +
  geom_smooth(color = "red") +
  scale_fill_gradient("Density", low = "white", high = "black") +
  scale_x_continuous("CDR3 length", limits = c(7.5,21.5), breaks = seq(8,22,by=3)) +
  scale_y_log10("Theoretical rearrangement probability",
                breaks = 10^(-15:-1),
                label= function(x) {ifelse(x==0, "0", parse(text=gsub("[+]", "", gsub(".+e", "10^", sci
```

```
  theme_bw()  +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

kruskal.test(genP_1mism_rob ~ I(nchar(cdr3)), df.raw %>% filter(nchar(cdr3) %in% 7:22))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  genP_1mism_rob by I(nchar(cdr3))
## Kruskal-Wallis chi-squared = 2167.1, df = 15, p-value < 2.2e-16
```

```
summary(aov(log(genP_1mism_rob) ~ I(nchar(cdr3)), df.raw %>% filter(nchar(cdr3) %in% 7:22, genP_1mism_r
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## I(nchar(cdr3))     1  21291   21291    2855 <2e-16 ***
## Residuals       5973  44545       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p9
```

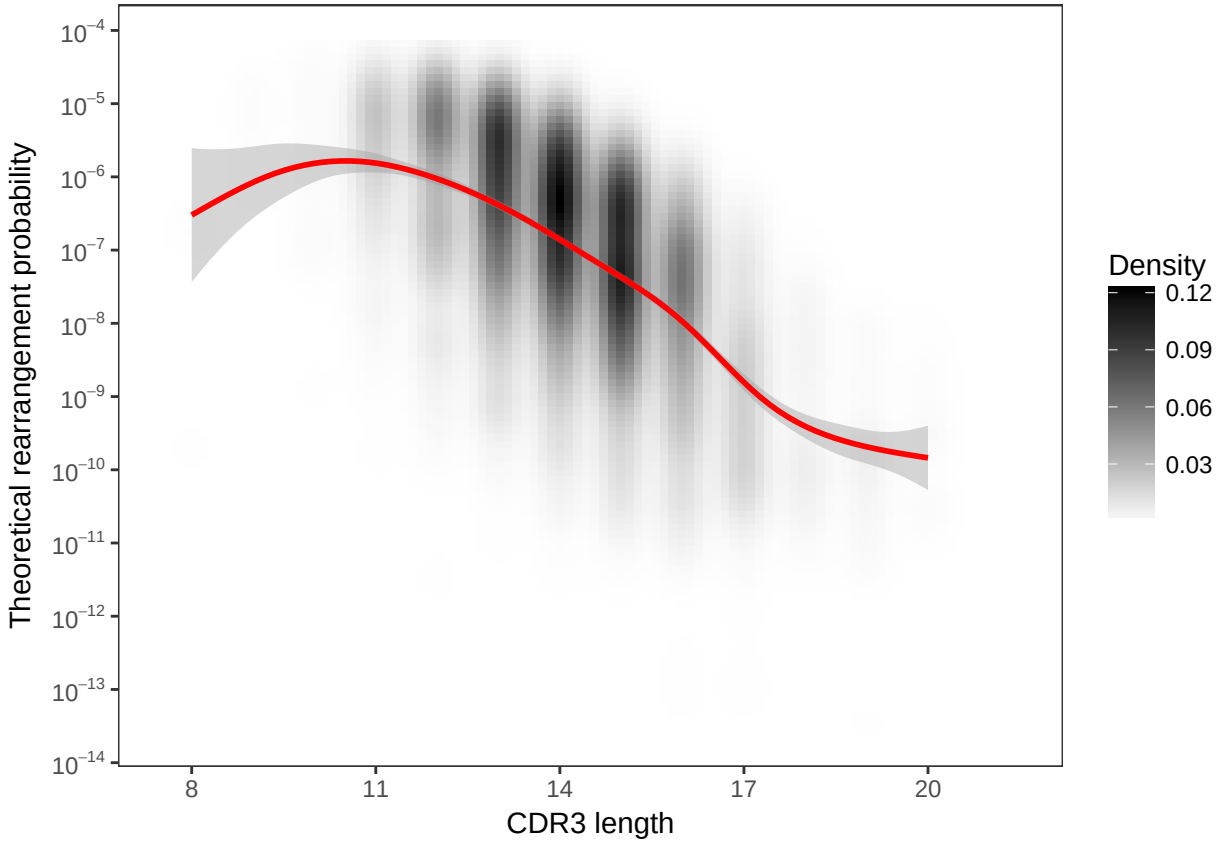
```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 335 rows containing non-finite values (stat_density2d).
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 335 rows containing non-finite values (stat_smooth).
```



CDR3 length and rearrangement probability

Structural basis

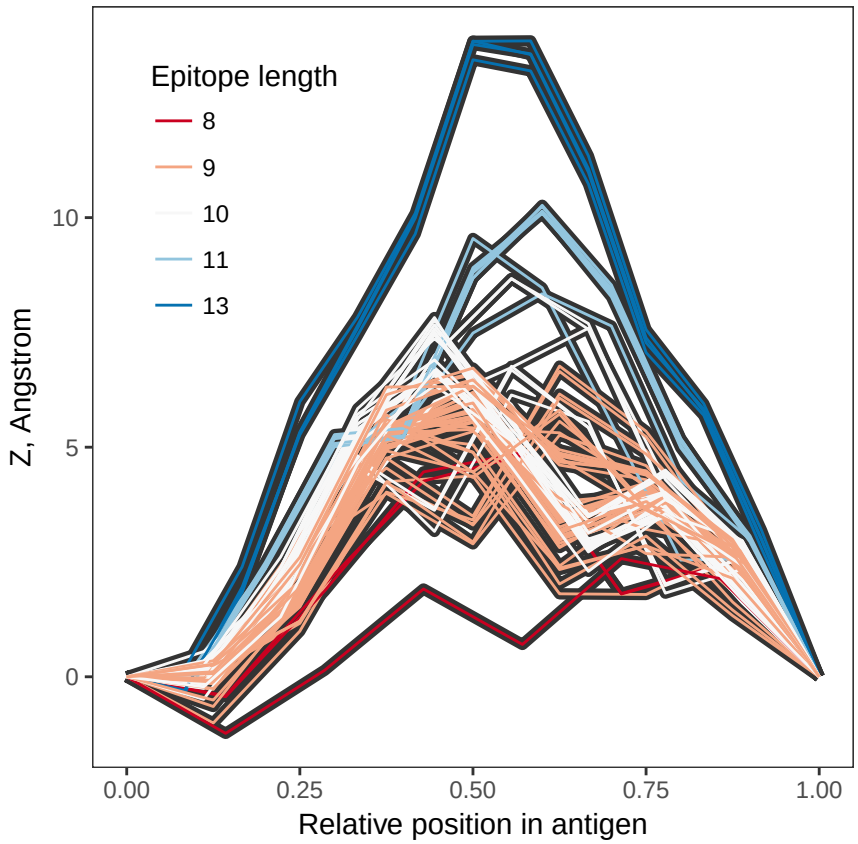
Epitope "bulging", note dist between C and N (X axis) is conserved for MHCI but not MHCII.

```
df.struct.annot = fread("zcat struct/final.annotations.txt.gz") %>%
  select(pdb_id, species, mhc_type, mhc_a_allele, mhc_b_allele) %>%
  mutate(hla = ifelse(mhc_type == "MHCI",
                      str_split_fixed(mhc_a_allele, fixed("*"), 10)[,1],
                      paste(str_split_fixed(mhc_a_allele, fixed("*"), 10)[,1], str_split_fixed(mhc_b_al

df.ag.coords = fread("zcat struct/backbone_ag.txt.gz") %>%
  merge(df.struct.annot, allow.cartesian=T) %>%
  select(pdb_id, species, mhc_type, hla, len_ag, pos_ag, x, y, z)

colnames(df.ag.coords) = c("pdb_id", "species", "mhc_type", "hla", "len_ag", "pos_ag", "x_ag", "y_ag",

p10=ggplot(df.ag.coords %>% filter(mhc_type == "MHCI" & species == "Homo_sapiens"), aes(x=pos_ag/(len_a
  geom_line(aes(group = pdb_id), size = 2, color="grey20") +
  geom_line(aes(group = pdb_id, color = factor(len_ag))) +
  scale_x_continuous("Relative position in antigen") +
  scale_y_continuous("Z, Angstrom") +
  scale_color_brewer("Epitope length", palette = "RdBu") +
  theme_bw() +
  theme(aspect = 1,
        panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        legend.position = c(0.2, 0.75),
        legend.background = element_blank())
p10
```

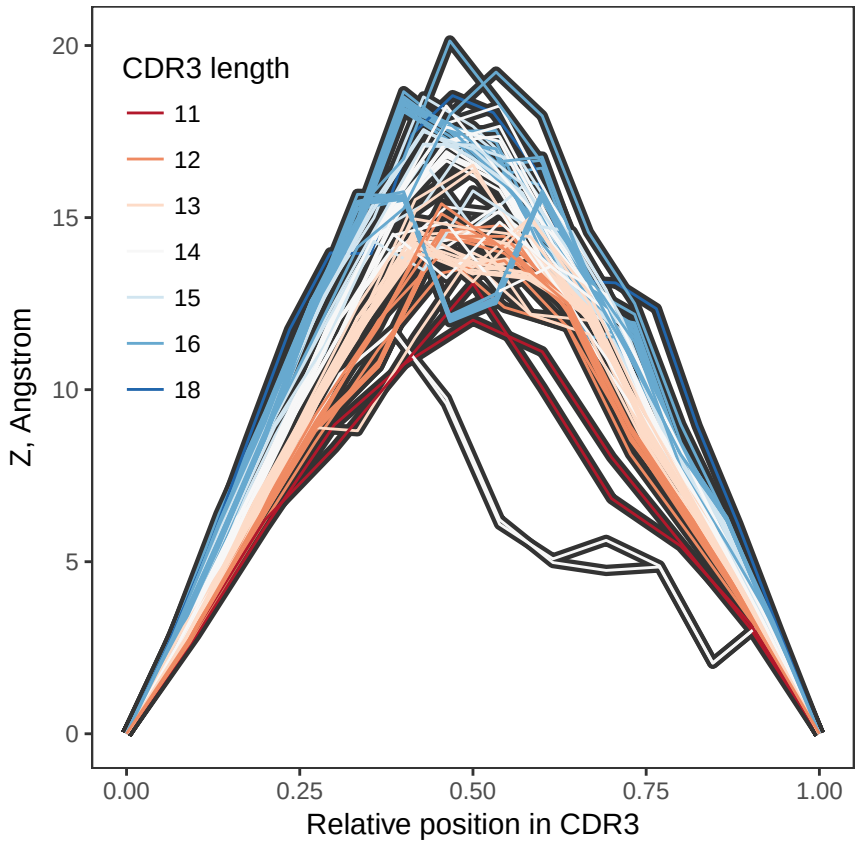


CDR3 bulging

```
df.tcr.coords = fread("zcat struct/backbone.txt.gz") %>%
  merge(df.ag.coords %>% select(pdb_id, len_ag) %>% unique) %>% filter(tcr_region == "CDR3" & mhc_type =
  mutate(tcr_chain = substr(tcr_v_allele,1,3)) %>%
  select(pdb_id, species, mhc_type, len_ag, tcr_chain, len_tcr, pos_tcr, x, y, z)

colnames(df.tcr.coords) = c("pdb_id", "species", "mhc_type", "len_ag", "tcr_chain", "len_tcr", "pos_tcr

p11=ggplot(df.tcr.coords %>% filter(species == "Homo_sapiens", tcr_chain == "TRB"), aes(x=pos_tcr/(len_t
  geom_line(aes(group = pdb_id), size = 2, color = "grey20") +
  geom_line(aes(group = pdb_id, color = factor(len_tcr))) +
  scale_x_continuous("Relative position in CDR3") +
  scale_y_continuous("Z, Angstrom") +
  scale_color_brewer("CDR3 length", palette = "RdBu") +
  theme_bw() +
  theme(aspect = 1,
        panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        legend.position = c(0.15, 0.7),
        legend.background = element_blank())
p11
```



Figures

```
ggsave("figures/p8.pdf", p8, width = 4*2, height = 4)
```

```
## Warning: Removed 9 rows containing non-finite values (stat_bin).
```

```
ggsave("figures/p9.pdf", p9, width = 4*2, height = 4)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 335 rows containing non-finite values (stat_density2d).
## `geom_smooth()` using method = 'gam'
## Warning: Removed 335 rows containing non-finite values (stat_smooth).
```

```
ggsave("figures/p10.pdf", p10, width = 4, height = 4)
ggsave("figures/p11.pdf", p11, width = 4, height = 4)
```