EDA of mock TCR:pMHC complexes

```
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(RColorBrewer)
```

```
df = fread("../preprocessing/output/structure.txt", header=T, sep="\t")[tcr_region %in% c("CDR3")]

df$tcr_gene = as.factor(substr(as.character(df$tcr_v_allele), 1, 3))
df$tcr_v_allele = NULL
df$mhc_a_allele = NULL
df$mhc_b_allele = NULL
df$energy = NULL
```

```
## Warning in `[<-.data.table`(x, j = name, value = value): Adding new column
## 'energy' then assigning NULL (deleting it).
```

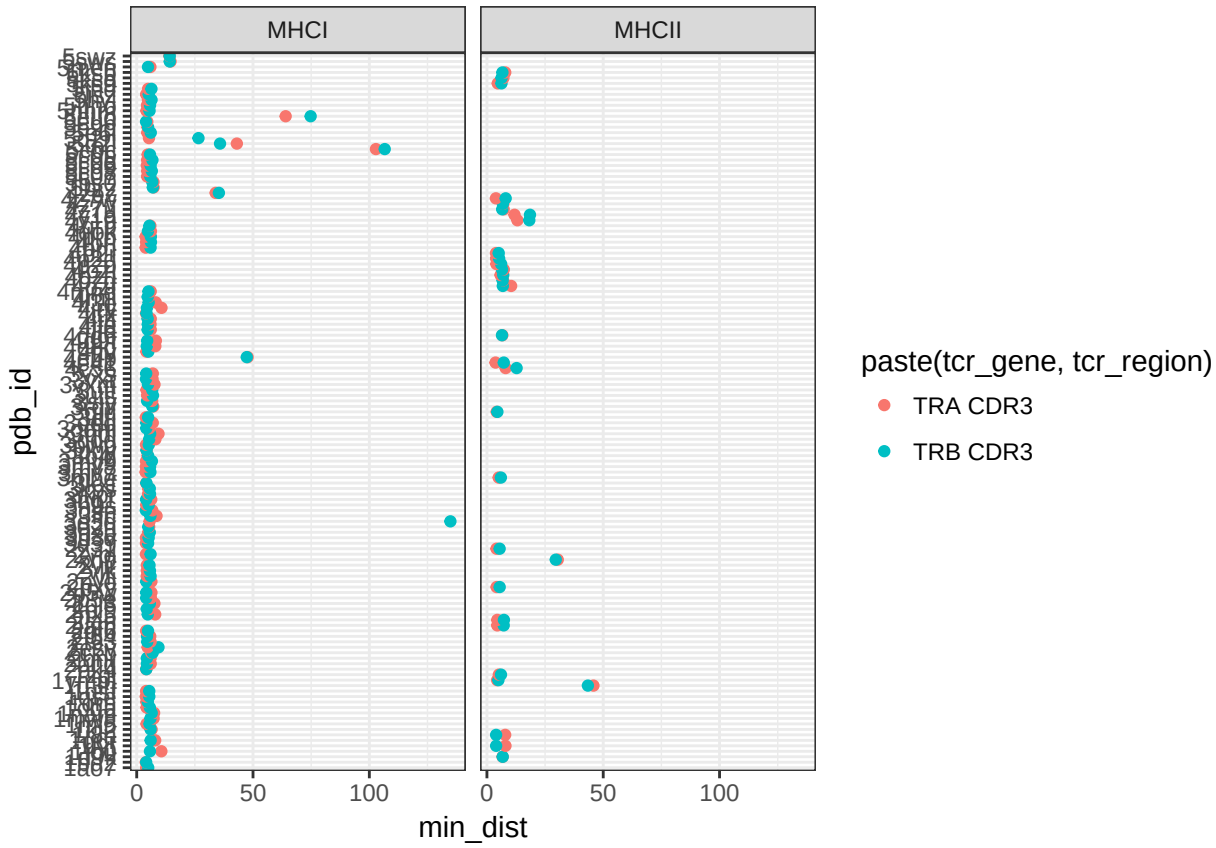
```
df$mhc_type = as.factor(df$mhc_type)

df$tcr_region = as.factor(df$tcr_region)
df$aa_tcr = as.factor(df$aa_tcr)
df$aa_antigen = as.factor(df$aa_antigen)
df$species = as.factor(df$species)
df$pdb_id = as.factor(df$pdb_id)

df$contact = df$distance <= 4.5
#df = df[pdb_id != "4p46"] # this one has covlinked peptide

summary(df)
```

```
##      pdb_id              species       mhc_type     antigen_seq
##  4p2o   :  580   Homo_sapiens:30645   MHCI :25557   Length:37245
##  3mbe   :  476   Mus_musculus: 6600   MHCII:11688   Class :character
```

```
##  5ks9    :  464                                      Mode  :character
##  1ymm    :  420
##  4y1a    :  420
##  4z7u    :  420
##  (Other):34465
##  tcr_gene    tcr_region   tcr_region_seq         aa_tcr
##  TRA:18287   CDR3:37245   Length:37245       S      : 4371
##  TRB:18958                Class :character   A      : 4156
##                           Mode  :character   G      : 4151
##                                              F      : 3563
##                                              C      : 2748
##                                              L      : 2438
##                                              (Other):15818
##    aa_antigen       len_tcr       len_antigen       pos_tcr
##  L      : 4769   Min.   : 5.00   Min.   : 8.00   Min.   : 0.000
##  G      : 3869   1st Qu.:13.00   1st Qu.: 9.00   1st Qu.: 3.000
##  P      : 3390   Median :14.00   Median :10.00   Median : 6.000
##  A      : 2753   Mean   :13.79   Mean   :10.97   Mean   : 6.394
##  F      : 2560   3rd Qu.:15.00   3rd Qu.:13.00   3rd Qu.:10.000
##  V      : 2461   Max.   :18.00   Max.   :20.00   Max.   :17.000
##  (Other):17443
##   pos_antigen        distance        distance_CA       contact
##  Min.   : 0.000   Min.   :  2.355   Min.   :  3.696   Mode :logical
##  1st Qu.: 2.000   1st Qu.: 10.136   1st Qu.: 13.417   FALSE:35647
##  Median : 5.000   Median : 15.026   Median : 18.184   TRUE :1598
##  Mean   : 4.985   Mean   : 17.552   Mean   : 20.678   NA's :0
##  3rd Qu.: 7.000   3rd Qu.: 20.096   3rd Qu.: 23.181
##  Max.   :19.000   Max.   :146.838   Max.   :150.204
##
```

Remove bad regions/complexes

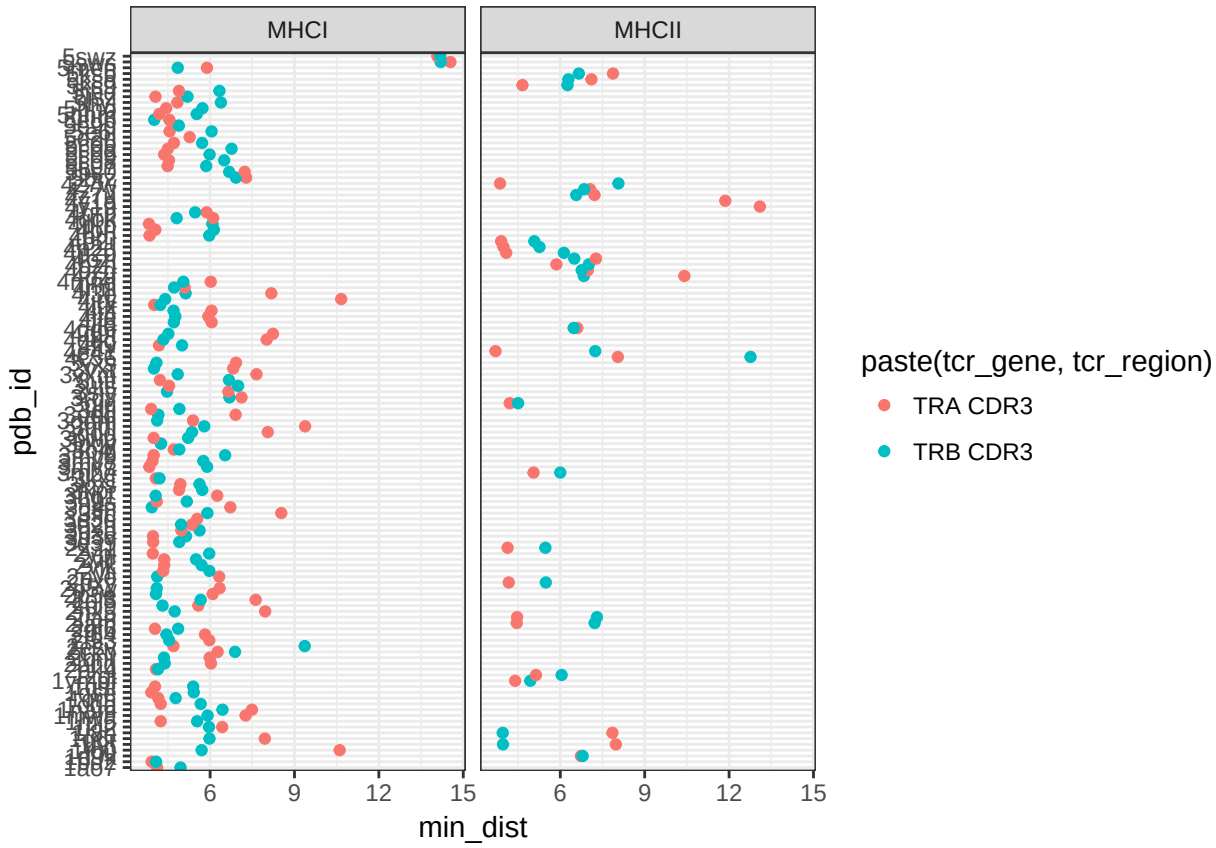
```
df.dist.min = df %>%
  group_by(pdb_id, tcr_gene, tcr_region, mhc_type) %>%
  summarize(min_dist = min(distance_CA), mean_dist = mean(distance_CA))

ggplot(df.dist.min, aes(x=pdb_id, y = min_dist, color = paste(tcr_gene, tcr_region))) +
  geom_point() +
  coord_flip() +
  facet_wrap(~mhc_type) +
  theme_bw()
```




```
good_regions = df.dist.min %>% filter(min_dist <= 15) %>%
  select(pdb_id, tcr_gene, tcr_region)

good_regions$good_region = T

good_pdb = unique(good_regions$pdb_id)

print(good_pdb)
```

```
##   [1] 1ao7 1bd2 1d9k 1fo0 1fyt 1g6r 1j8h 1kj2 1mi5 1mwa 1nam 1oga 1qrn 1qse
##  [15] 1qsf 1ymm 1zgl 2ak4 2bnq 2bnr 2ckb 2esv 2f53 2f54 2gj6 2iam 2ian 2nx5
##  [29] 2oi9 2ol3 2p5e 2p5w 2pxy 2pye 2vlj 2vlk 2vlr 2ypl 2z31 3d39 3d3v 3dxa
##  [43] 3e2h 3e3q 3ffc 3gsn 3h9s 3hg1 3kpr 3kps 3kxf 3mbe 3mv7 3mv8 3mv9 3o4l
##  [57] 3pqy 3pwp 3qdg 3qdj 3qdm 3qeq 3qfj 3qiu 3rgv 3sjv 3uts 3utt 3vxm 3vxr
##  [71] 3vxs 4c56 4e41 4ftv 4g8g 4g9f 4gg6 4jfd 4jfe 4jff 4jrx 4jry 4l3e 4mji
##  [85] 4mnq 4ozf 4ozg 4ozh 4ozi 4p2o 4p2q 4p2r 4prh 4pri 4prp 4qok 4qrp 4y19
##  [99] 4y1a 4z7u 4z7v 4z7w 5brz 5bs0 5c07 5c08 5c09 5c0a 5c0b 5d2n 5e6i 5e9d
## [113] 5eu6 5hhm 5hho 5hyj 5isz 5jhd 5ks9 5ksa 5ksb 5men 5sws 5swz
## 131 Levels: 1ao7 1bd2 1d9k 1fo0 1fyt 1g6r 1j8h 1kj2 1mi5 1mwa 1nam ... 5swz
```

```
ggplot(subset(df.dist.min, min_dist<=15), aes(x=pdb_id, y = min_dist, color = paste(tcr_gene, tcr_regio
  geom_point() +
  coord_flip() +
  facet_wrap(~mhc_type) +
  theme_bw()
```




```
df = merge(df, good_regions, all.x = T, by = c("pdb_id", "tcr_gene", "tcr_region"))
df = subset(df, good_region)
```

Load mock data

```
df.mock = fread("../preprocessing/output/mock_structure.txt", header=T, sep="\t")[tcr_region %in% c("CD
```

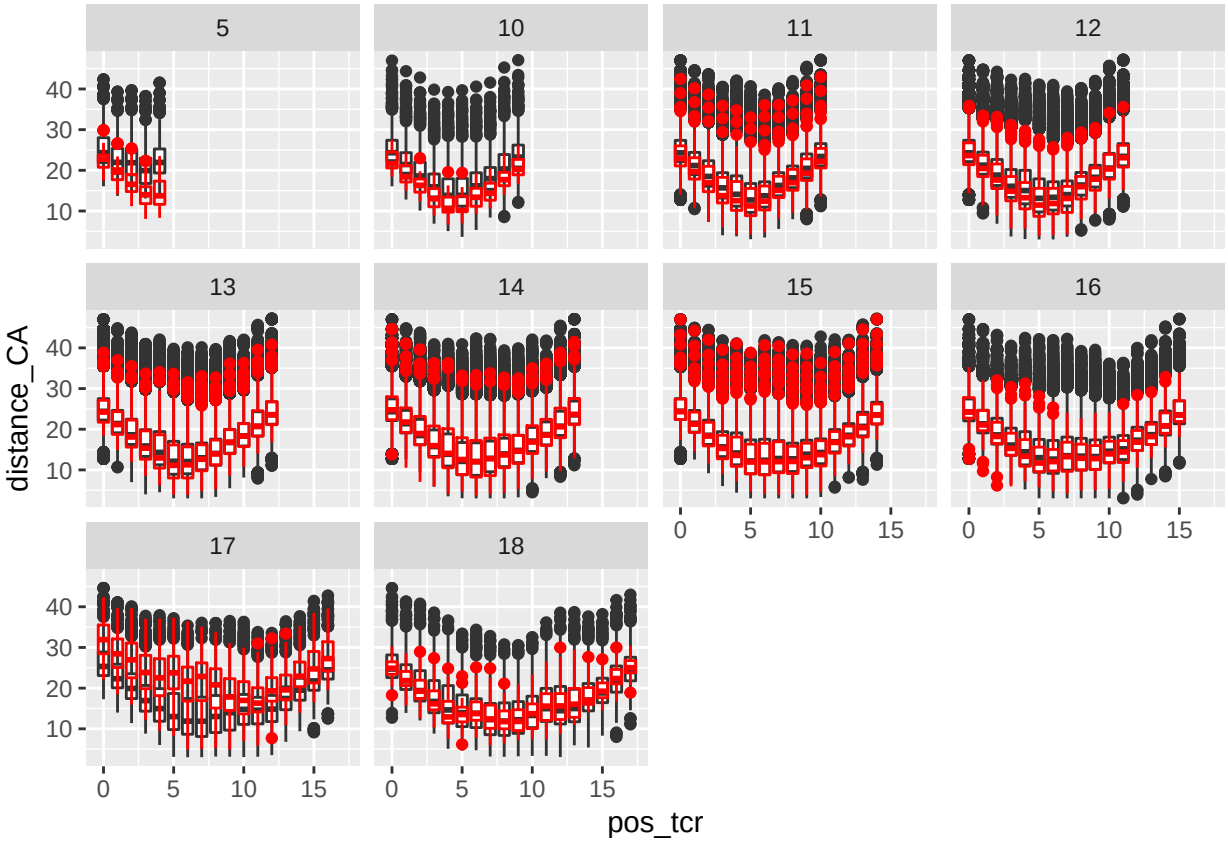
```
##
Read 4.9% of 4710696 rows
Read 25.5% of 4710696 rows
Read 46.1% of 4710696 rows
Read 66.7% of 4710696 rows
Read 87.2% of 4710696 rows
Read 4710696 rows and 17 (of 17) columns from 0.562 GB file in 00:00:07
```

```
df.dist.min.m = df.mock %>%
  group_by(pdb_id_a, pdb_id_t, tcr_gene, tcr_region, mhc_type) %>%
  summarize(min_dist = min(distance_CA), mean_dist = mean(distance_CA))

good_regions = df.dist.min.m %>% filter(min_dist <= 15 & min_dist > 3) %>%
  select(pdb_id_a, pdb_id_t, tcr_gene, tcr_region, mhc_type)

good_regions$good_region = T

df.mock = merge(df.mock, good_regions, all.x = T,
                by = c("pdb_id_a", "pdb_id_t", "tcr_gene", "tcr_region", "mhc_type"))
df.mock = subset(df.mock, good_region)
```

1D parameters

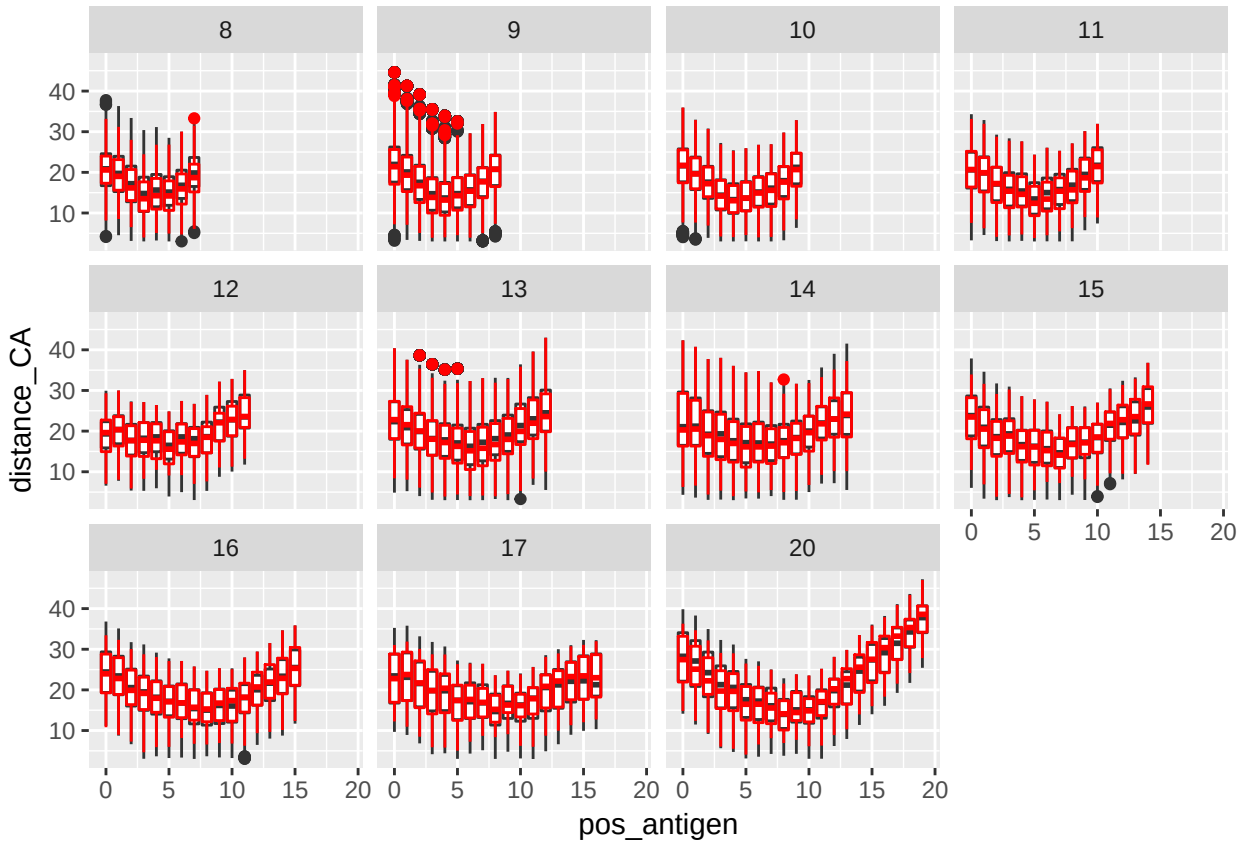
Distribution of CA distances from TCR point of view

```
ggplot() +
  geom_boxplot(data=df.mock,aes(x=pos_tcr, group=pos_tcr, y=distance_CA)) +
  geom_boxplot(data=df,aes(x=pos_tcr, group=pos_tcr, y=distance_CA), color="red", fill=NA) +
  facet_wrap(~len_tcr)
```



from antigen point of view

```
ggplot() +
  geom_boxplot(data=df.mock,aes(x=pos_antigen, group=pos_antigen, y=distance_CA)) +
  geom_boxplot(data=df,aes(x=pos_antigen, group=pos_antigen, y=distance_CA), color="red", fill=NA) +
  facet_wrap(~len_antigen)
```

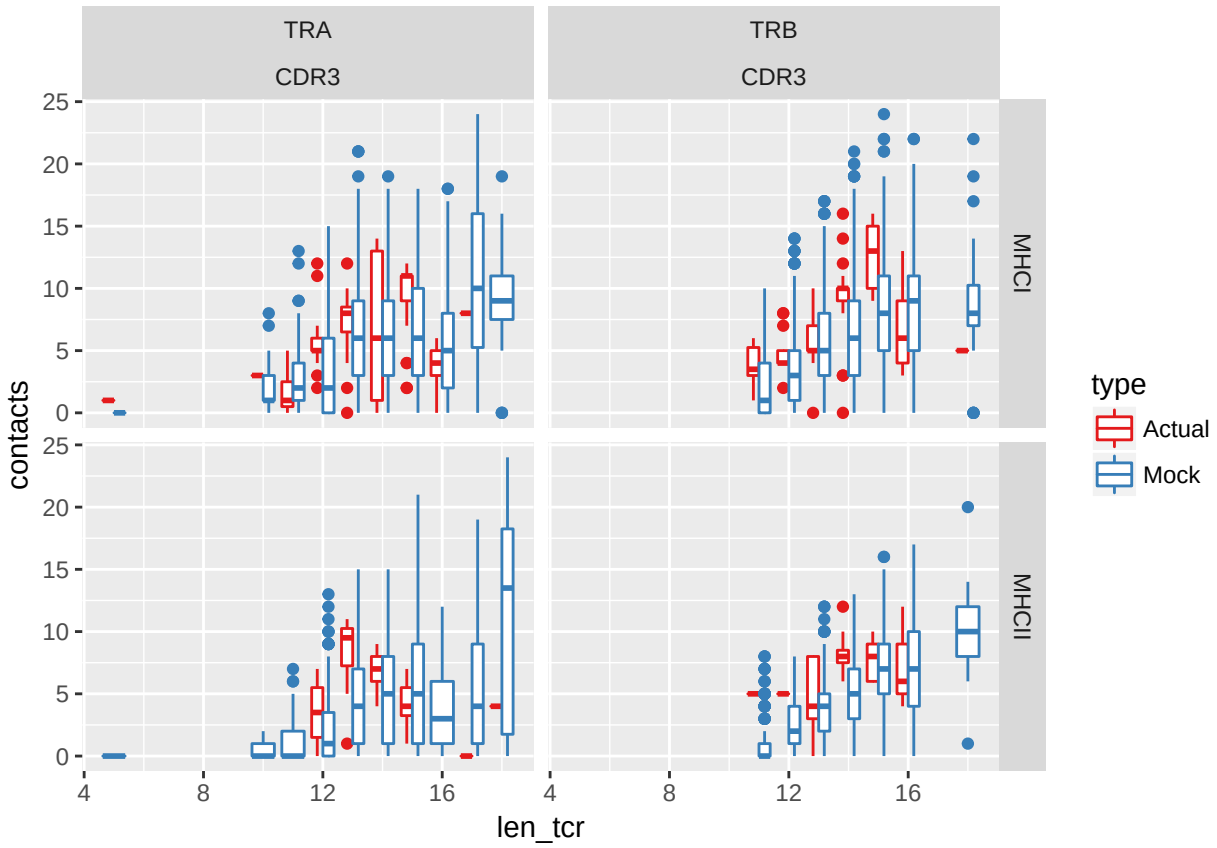


Number of contacts

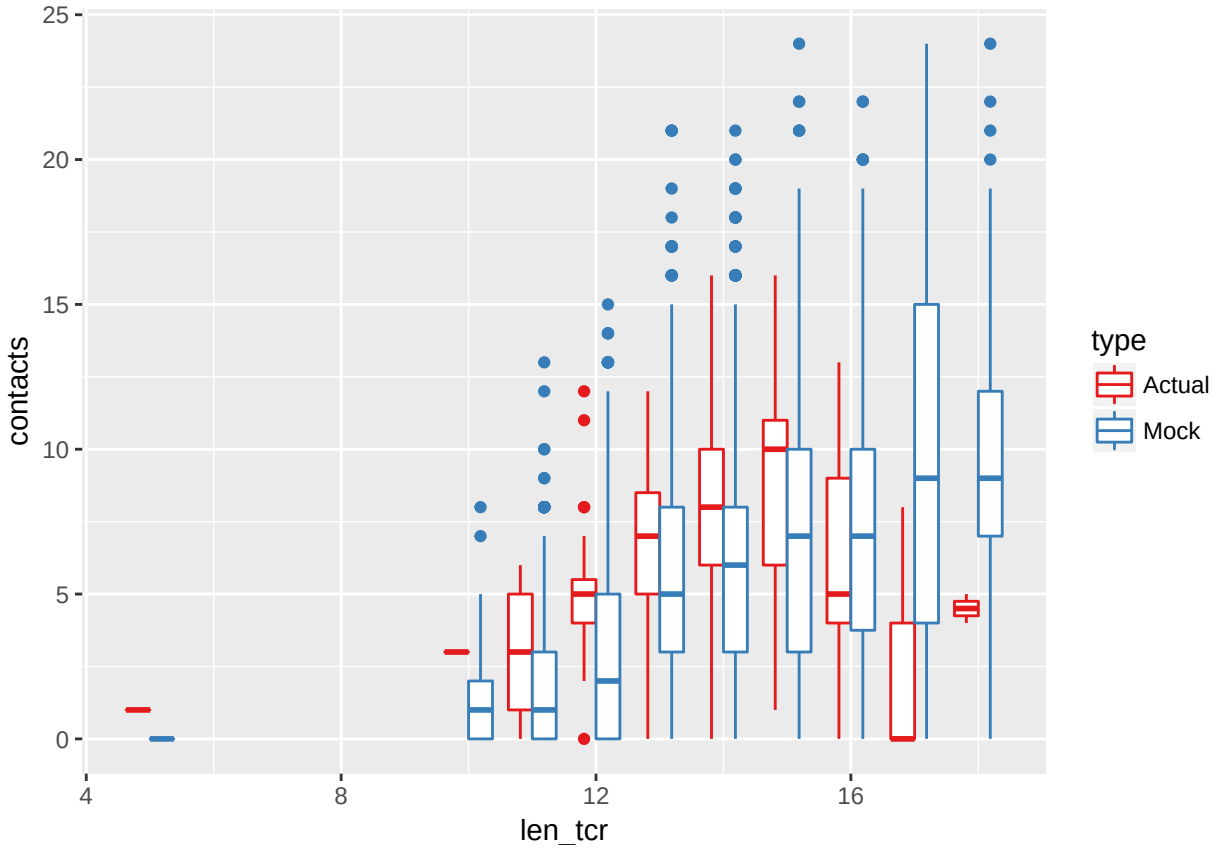
```
df.contsum = df %>%
  group_by(pdb_id, tcr_gene, tcr_region, mhc_type, len_tcr) %>%
  summarise(contacts = sum(distance <= 4.5))
df.contsum$type = "Actual"
df.contsum$pdb_id = NULL
df.contsum = as.data.frame(df.contsum)

df.contsum.m = df.mock %>%
  group_by(pdb_id_a, pdb_id_t, tcr_gene, tcr_region, mhc_type, len_tcr) %>%
  summarise(contacts = sum(distance <= 4.5)) %>%
  dplyr::select(tcr_gene, tcr_region, mhc_type, len_tcr, contacts)
```

```
## Adding missing grouping variables: `pdb_id_a`, `pdb_id_t`
```

```
df.contsum.m$type = "Mock"
df.contsum.m$pdb_id_a = NULL
df.contsum.m$pdb_id_t = NULL
df.contsum.m = as.data.frame(df.contsum.m)

ggplot(rbind(df.contsum, df.contsum.m), aes(x=len_tcr, group = interaction(type,len_tcr), y=contacts, c
  geom_boxplot() +
  facet_grid(mhc_type~tcr_gene+tcr_region) +
  scale_color_brewer(palette = "Set1")
```




```
ggplot(rbind(df.contsum, df.contsum.m), aes(x=len_tcr, group = interaction(type,len_tcr), y=contacts, c
  geom_boxplot() +
  scale_color_brewer(palette = "Set1")
```

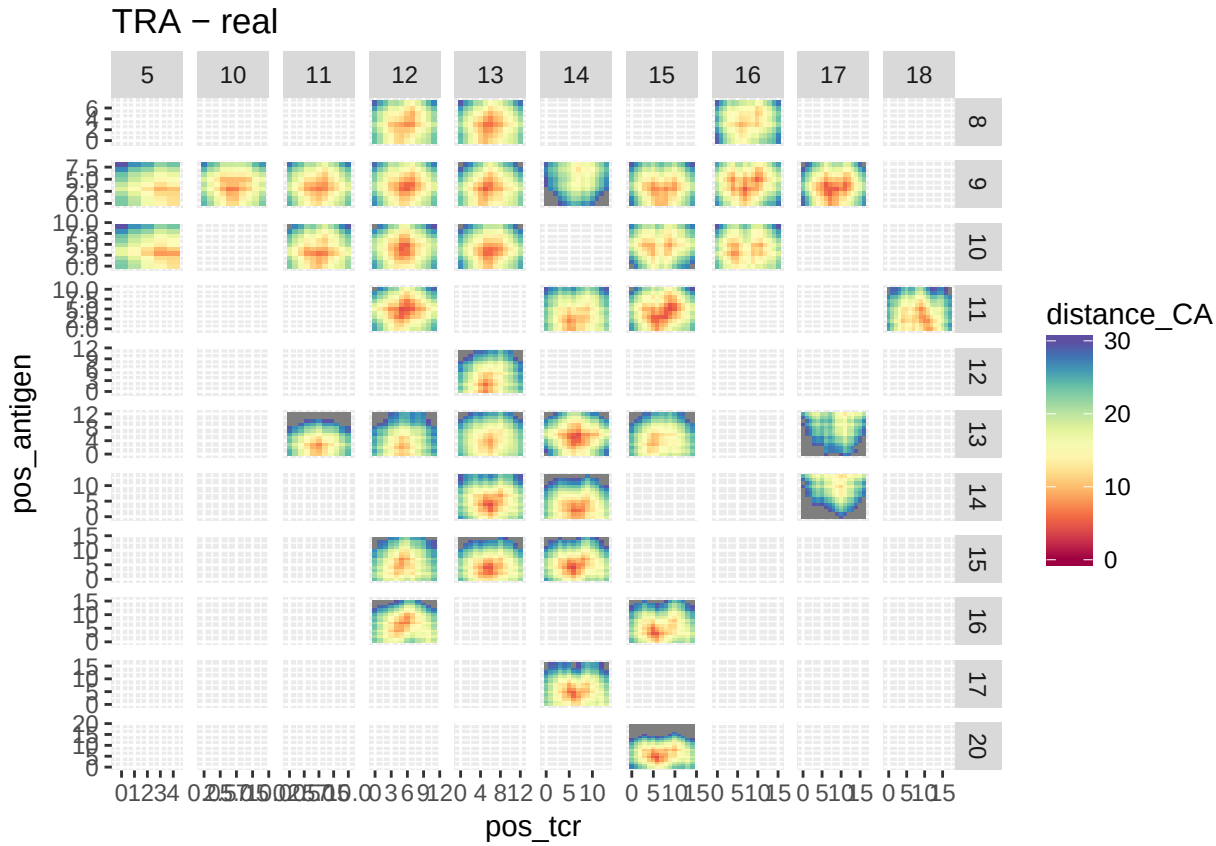


2D distance and contact plots

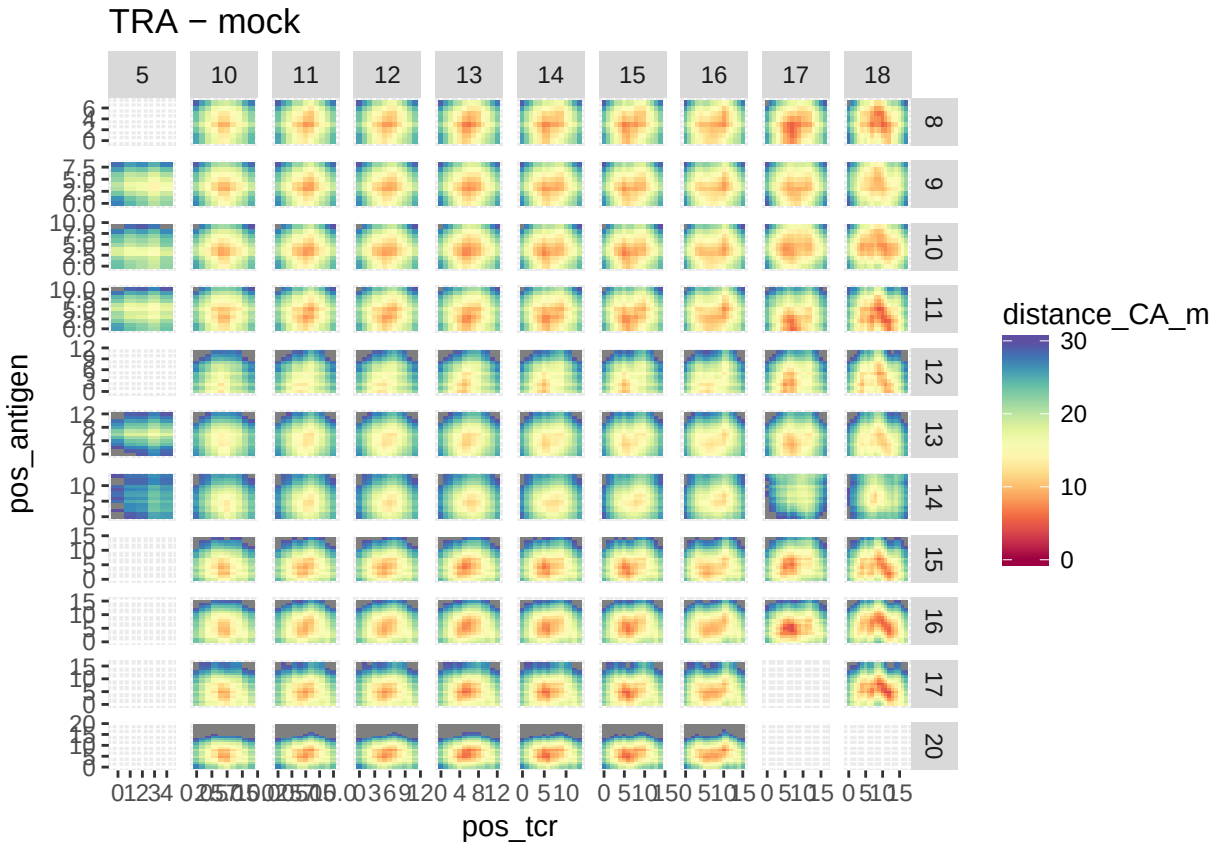
```
df.dist = df %>%
  group_by(pos_tcr, pos_antigen, len_antigen, len_tcr, tcr_gene, tcr_region) %>%
  summarize(distance_CA = mean(distance_CA))

df.dist.m = df.mock %>%
  group_by(pos_tcr, pos_antigen, len_antigen, len_tcr, tcr_gene, tcr_region) %>%
  summarize(distance_CA_m = mean(distance_CA))
```

For TRA

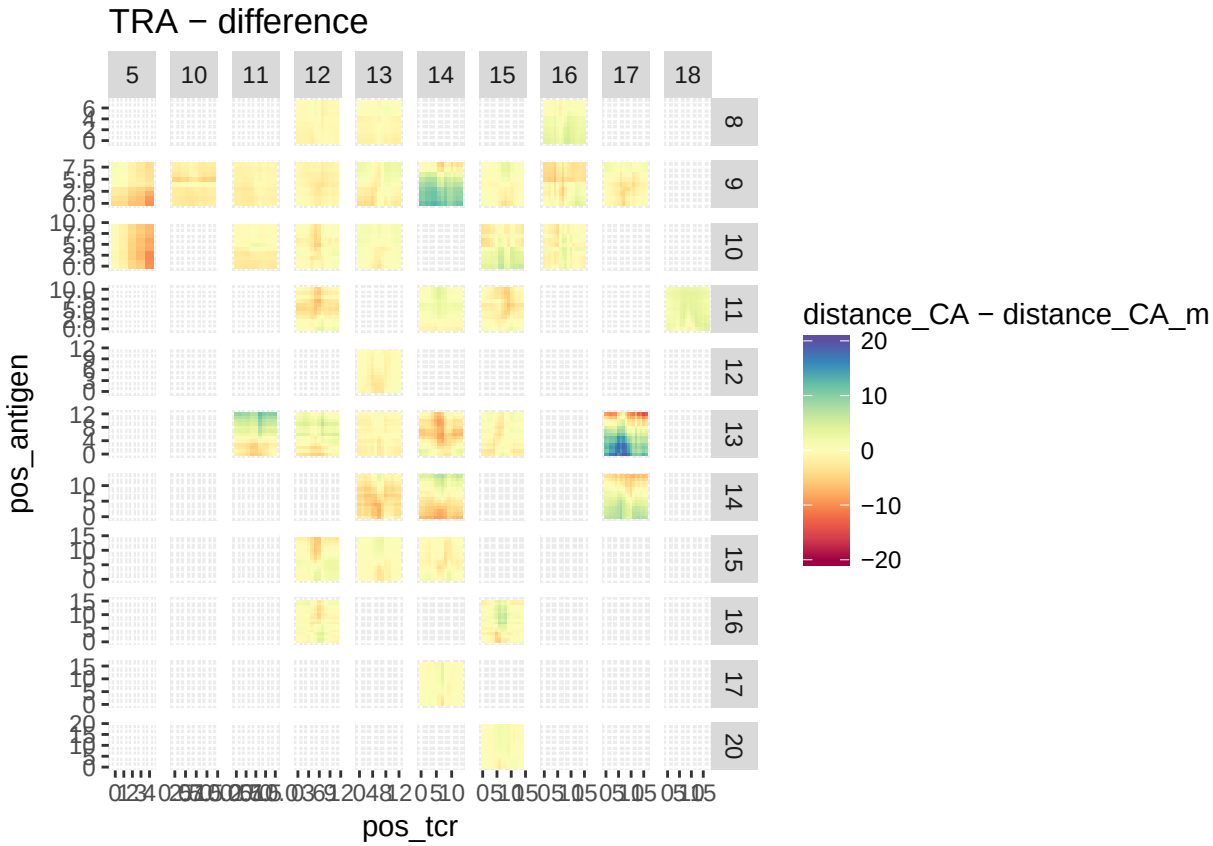
```
ggplot(df.dist %>% filter(tcr_gene == "TRA" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(0, 30)) +
  ggtitle("TRA - real")
```




```
ggplot(df.dist.m %>% filter(tcr_gene == "TRA" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA_m)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(0, 30)) +
  ggtitle("TRA - mock")
```

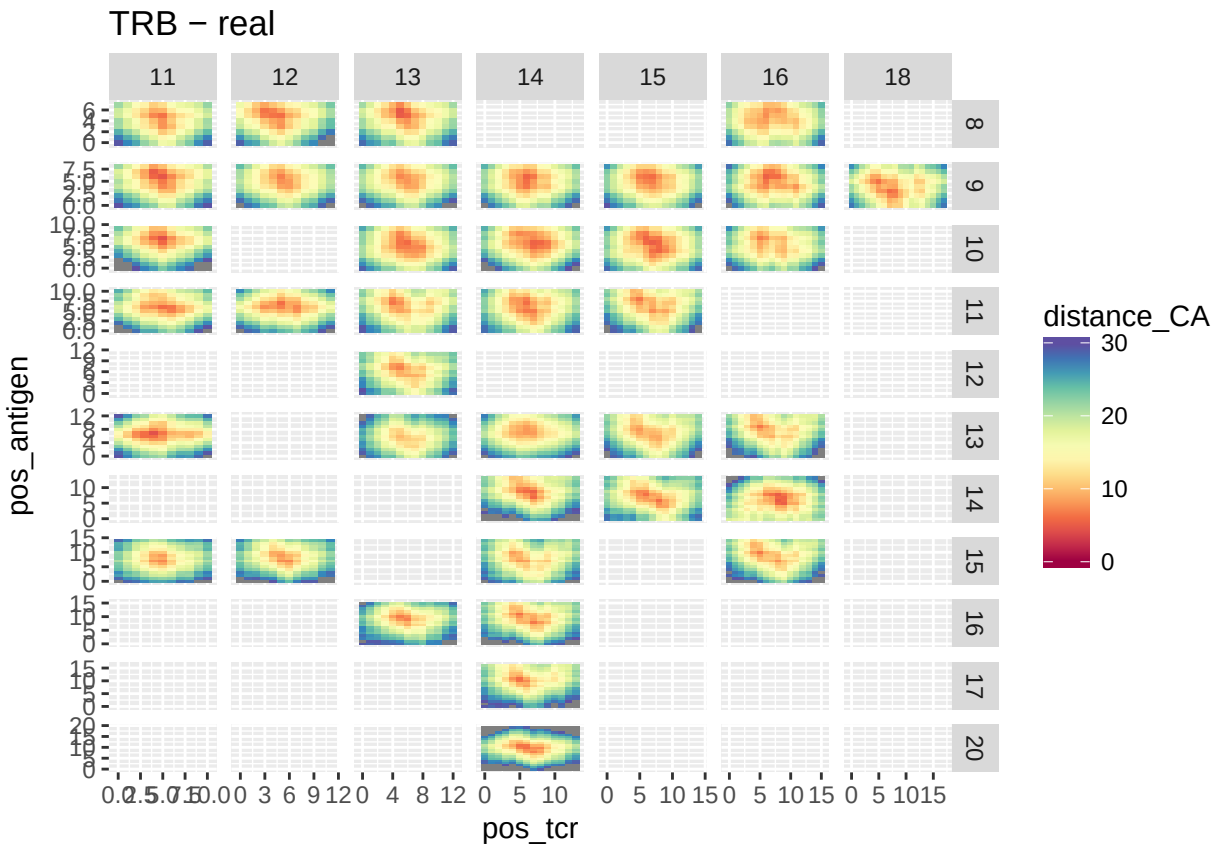



```
ggplot(merge(df.dist, df.dist.m) %>% filter(tcr_gene == "TRA" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA - distance_CA_m)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(-20, 20)) +
  ggtitle("TRA - difference")
```

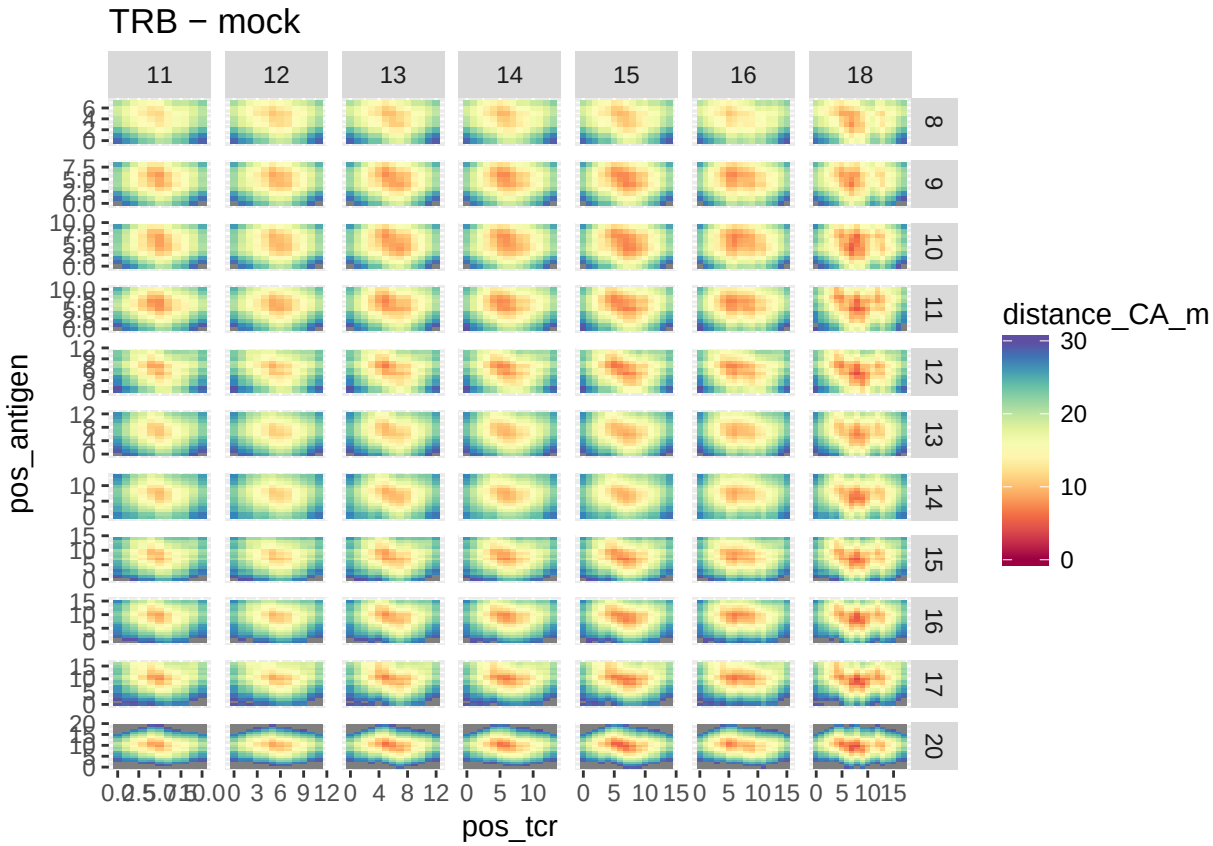


For TRB

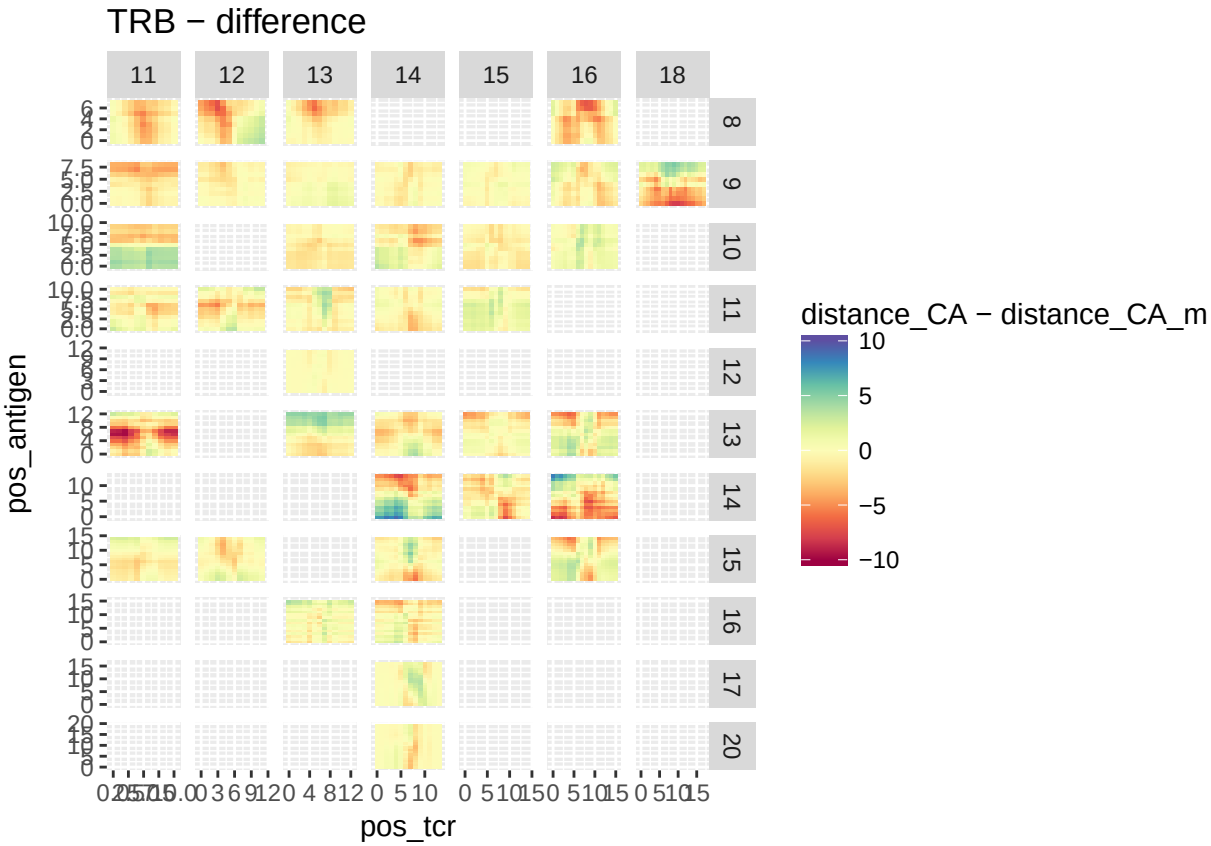
```
ggplot(df.dist %>% filter(tcr_gene == "TRB" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(0, 30)) +
  ggtitle("TRB - real")
```




```
ggplot(df.dist.m %>% filter(tcr_gene == "TRB" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA_m)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(0, 30)) +
  ggtitle("TRB - mock")
```




```
ggplot(merge(df.dist, df.dist.m) %>% filter(tcr_gene == "TRB" & tcr_region == "CDR3"),
       aes(x=pos_tcr, y=pos_antigen, fill = distance_CA - distance_CA_m)) +
  geom_tile() +
  facet_grid(len_antigen~len_tcr, scales="free") +
  scale_fill_gradientn(colors=colorRampPalette(brewer.pal(11, 'Spectral'))(32), limits=c(-10, 10)) +
  ggtitle("TRB - difference")
```


TRB − difference
11
12
13
14
15
16
18
8
9
10
11
12
13
14
15
16
17
20
pos_antigen
pos_tcr
distance_CA − distance_CA_m
10
5
0
−5
−10