Seed motifs

Melina Klostermann

24 June, 2023

Contents

1 Libraries and settings

```
# ----------------------------------------
# libraries
# ----------------------------------------
library(tidyverse)
library(GenomicRanges)
library(colorspace)
library(gghalves)
library(BSgenome.Mmusculus.UCSC.mm10)
library(Biostrings)

# ----------------------------------------
# settings
# ----------------------------------------
out <- "/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/Figure2/Seed_m
source("/Users/melinaklostermann/Documents/projects/R_general_functions/theme_paper.R")
source("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/mirko_files/mirECLIP/DifferentialBin


# farben
farbeneg <- "#B4B4B4"
farbe1 <- "#0073C2FF" # WT farbe
farbe2 <- "#EFC000FF" # mir181 enriched
farbe3 <- "#CD534CFF" #miR181KO farbe
farbe4 <- "#7AA6DCFF"
farbe5 <- "#868686FF"
farbe6 <- "#003C67FF"
```

```
farbe7 <- "#8F7700FF"
farbe8 <- "#3B3B3BFF"
farbe9 <- "#A73030FF"
farbe10 <- "#4A6990FF"
farbe11 <- "#FF6F00FF"
farbe12 <- "#C71000FF"
farbe13 <- "#008EA0FF"
farbe14 <- "#8A4198FF"
farbe15 <- "#5A9599FF"
farbe16 <- "#FF6348FF"
```

2 What was done?

- I count different versions of the miR181 seed in the 200nt before and after mir181 binding sites.
- I use the seed 6mer, 7mers with one adjecent nt, and a 8mer with two adjecent nts.

3 Files

```
# ----------------------------------------
# MREs
# ----------------------------------------

mir181_bs <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/I

mir_crosslinks <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/xx_down_stream_R/0

load("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/xx_down_stream_R/02_BS_definition_WT_re

mir181_enriched_set <- mir181_bs %>%
  as.data.frame(.) %>%
  subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched"))
```

4 XSTREME de novo motif discovery

```
# get seqeunce 200nt around binding sites
mir181_bs_200_both_sides <- makeGRangesFromDataFrame(mir181_enriched_set, keep.extra.columns = T)  + 200
names(mir181_bs_200_both_sides) <- 1:NROW(mir181_bs_200_both_sides)
mir181_bs_200_both_sides_seq<- getSeq(mir181_bs_200_both_sides, x = BSgenome.Mmusculus.UCSC.mm10) %>%
  RNAStringSet()

# write fasta file for XSTREME
writeXStringSet(mir181_bs_200_both_sides_seq, filepath = paste0(out, "mirBS_200_both_sides.fasta" ))
```

XSTREME is executed on the fasta file from above via the MEME SUITE webpage (https://meme-suite.org/meme/tools/xstreme) with the following parameters:

- E-value <= 0.05
- Width 5-10
- background: model control sequences
- STREME limit: Number of motifes = 20
- MEME options: Default E-value, Zero or one occurence per sequence

- SEA: Output the matching sequences in a TSV file

```
# plot logo
motif <- read.table("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/F

colnames(motif) <- c("A", "C", "G", "U")
motif <- t(motif)

logo <- ggseqlogo::ggseqlogo(motif)

logo
```


```
ggsave(logo, filename = paste0(out, "xstreme_logo.pdf"), height = 6, width = 8, units = "cm")
```

```
#################
# the mir181 seed and interesting seed variations
#################

seed_8mer <- "UGAAUGUU"
seed_7mer_m8 <- "UGAAUGU"
seed_7mer_a1 <- "GAAUGUU"
seed_6mer <- "GAAUGU"
seed_6mer_wobble <- "GAUUGU"
seed_8mer_wobble <- "UGAUUGUU"
seed_7mer_m8_wobble <- "UGAUUGU"
seed_7mer_a1_wobble <- "GAUUGUU"

# make a list of all seeds
seed_list <- list(seed_8mer, seed_7mer_m8, seed_7mer_a1,seed_6mer, seed_6mer_wobble, seed_8mer_wobble, s
```

```
seed_names_list <- list( "seed_8mer",  "seed_7mer_m8", "seed_7mer_a1",  "seed_6mer", "seed_6mer_wobble"

# hierarchy order, to decide which seed to use if several ar present
seed_importance_order <- c("seed_8mer",  "seed_7mer_m8", "seed_7mer_a1",  "seed_6mer")
```

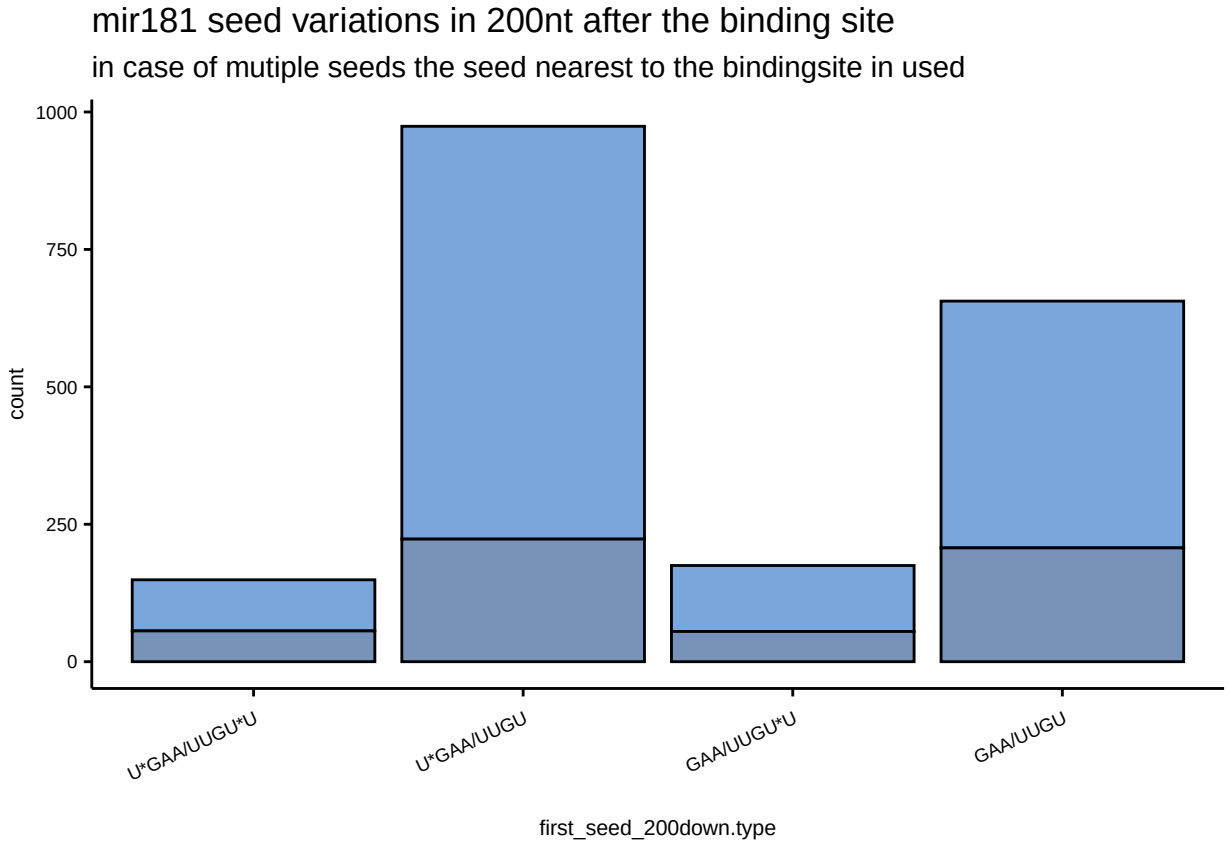
5 Seed position and distribution

5.1 200nt after the binding site

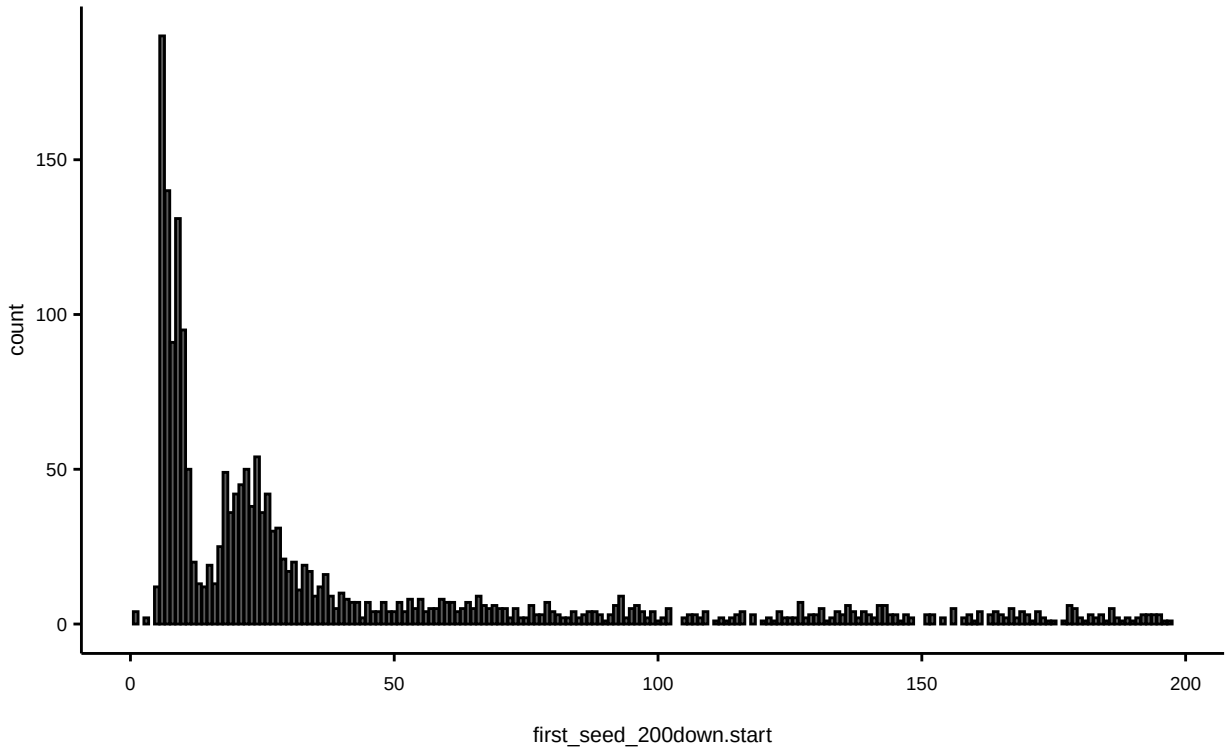
```
#######################
# get seed in 200er window
#######################

mir181_bs_200down <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T) %>% resize(., width = 1
mir181_bs_200down_seq <- getSeq(mir181_bs_200down, x = BSgenome.Mmusculus.UCSC.mm10) %>%
  RNAStringSet()

# count occurences of all seed variations
Seeds_200down <- lapply(seed_list, function(x) {
  vmatchPattern(pattern = x, mir181_bs_200down_seq) %>%
  lapply(., function(x) as.data.frame(x))})

# add the binding site id to the seeds and make a df per seed type
BS_ID_list <- as.list(mir181_bs$mir181BS_ID)

Seeds_200down <- map(Seeds_200down,
                     ~map2(.x, BS_ID_list, ~mutate(.x, mir181BS_ID = .y) ) %>%
               map_dfr(~.x))


# add the seed type names and make one df of all
Seeds_200down <- map2(Seeds_200down, seed_names_list, ~mutate(.x, seed = .y) ) %>% map_dfr(~.x)


# extract wobble positions
Seeds_1_per_BS <- Seeds_200down %>%
  mutate(wobble = grepl("wobble", seed),
         seed = case_when(wobble  ~ substr(seed, 1, nchar(seed)-7), T ~ seed))


# order seeds by importance
Seeds_1_per_BS$seed <- factor(Seeds_1_per_BS$seed, levels = seed_importance_order )

# select 1 seed per BS --> closest seed with highest importance
Seeds_1_per_BS <- Seeds_1_per_BS %>%
   group_by(mir181BS_ID) %>%
  arrange(start, seed ) %>%
  dplyr::slice(1) %>%
  ungroup(.)

########################
# combine the closest seed, and all found seeds to the Binding site data.frame
```

```
#######################

# add all as list column

colnames(Seeds_200down) <- c("Seeds_200down.start",
                             "Seeds_200down.end",
                             "Seeds_200down.width",
                             "mir181BS_ID",
                             "Seeds_200down.type")


mir181_bs <- left_join(mir181_bs, Seeds_200down, by = "mir181BS_ID") %>%
  tidyr::nest(all_seeds_200down = c("Seeds_200down.start",
                             "Seeds_200down.end",
                             "Seeds_200down.width",
                             "Seeds_200down.type"))

# add closest mir
colnames(Seeds_1_per_BS) <- c("first_seed_200down.start",
                             "first_seed_200down.end",
                             "first_seed_200down.width",
                             "mir181BS_ID",
                             "first_seed_200down.type",
                             "first_seed_200down.wobble")

mir181_bs <- left_join(mir181_bs, Seeds_1_per_BS, by = "mir181BS_ID")

mir181_bs <- mir181_bs %>%
  rowwise() %>%
  mutate(seed_repetitions.200down = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer"),
         seed_repetitions.200down.wobble = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer_wobble
```

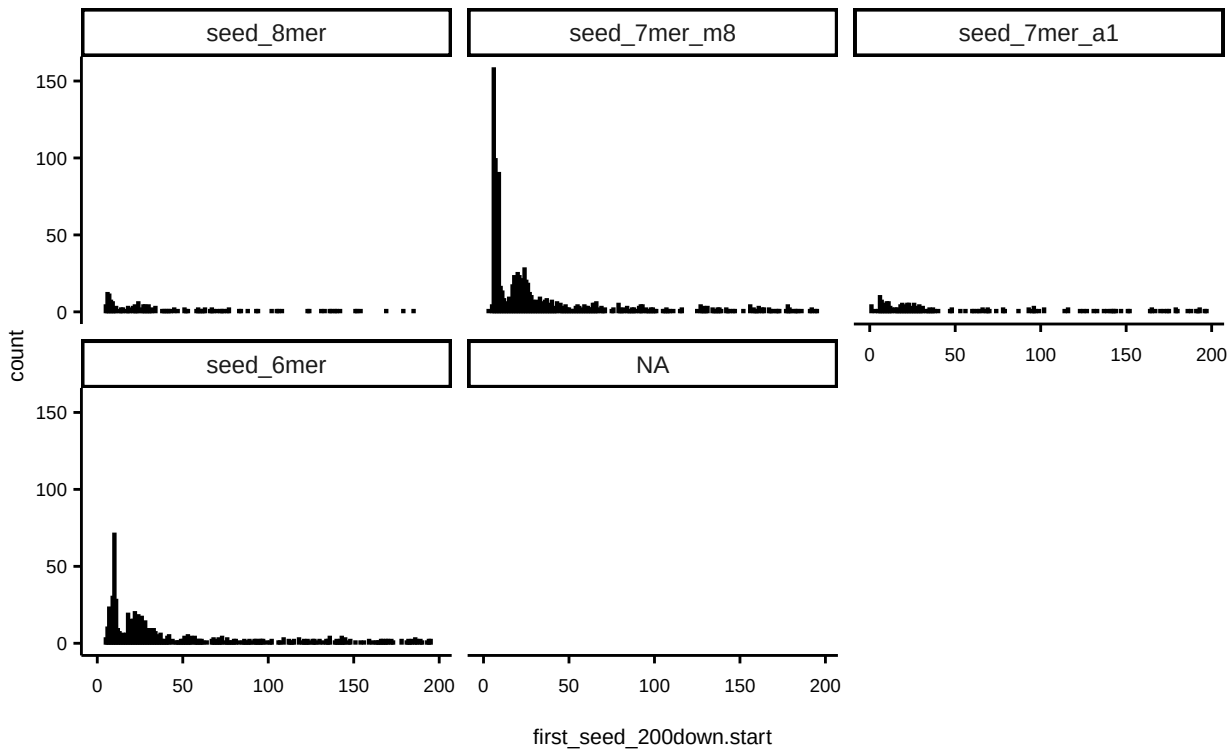
```
####################
# plots
####################

# plot seed variations
p <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")) %>%
  geom_bar(color = "black")+
  theme_paper()+
  scale_fill_manual(values = c(farbe4, darken(farbe4)))+
  theme(legend.position = "None") +
  scale_x_discrete(labels=c(seed_8mer = "U*GAA/UUGU*U",
                            seed_7mer_m8 = "U*GAA/UUGU",
                            seed_7mer_a1 = "GAA/UUGU*U",
                            seed_6mer = "GAA/UUGU"),
                   guide = guide_axis(angle = 25))

p+
  ggtitle("mir181 seed variations in 200nt after the binding site",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
```



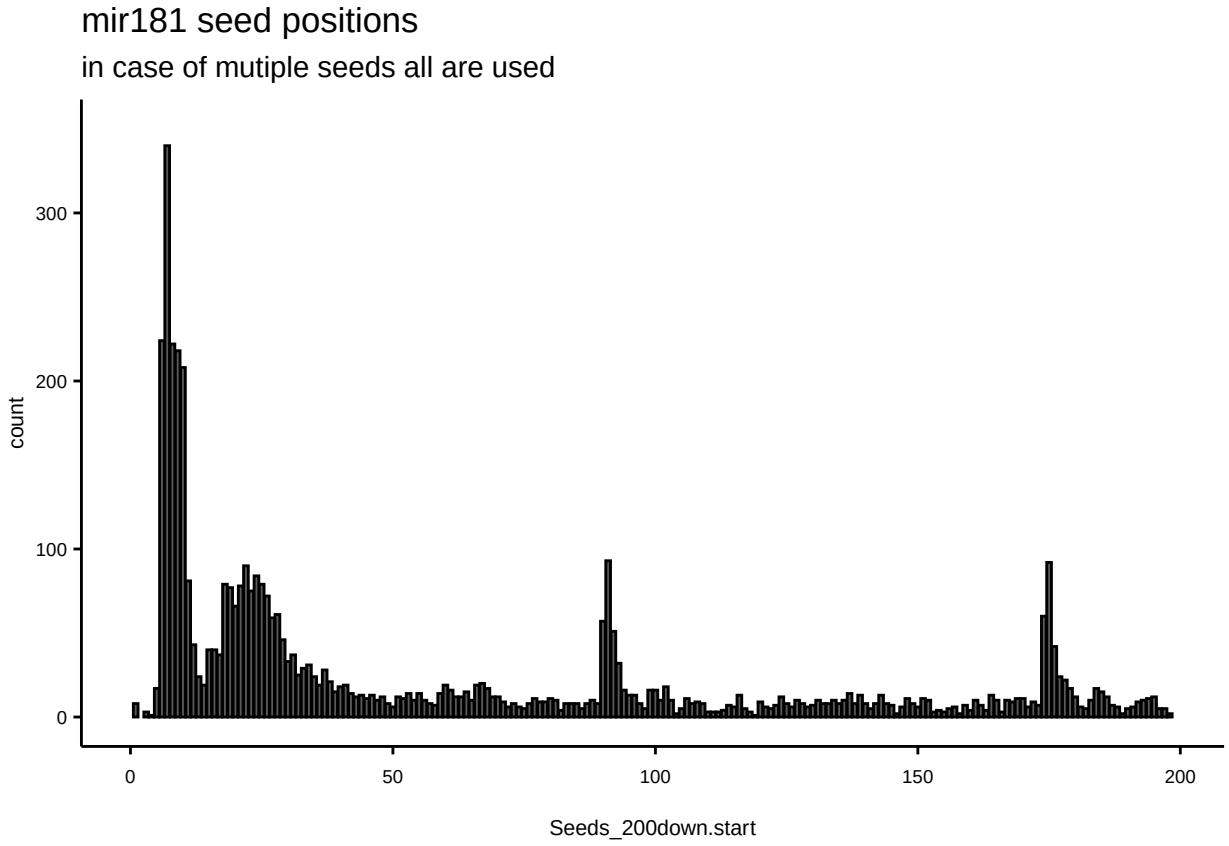

```
# plot seed distributions
p2 <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")), a
  geom_bar(color = "black")+
  theme_paper()+
  ggtitle("mir181 seed positions",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
p2
```



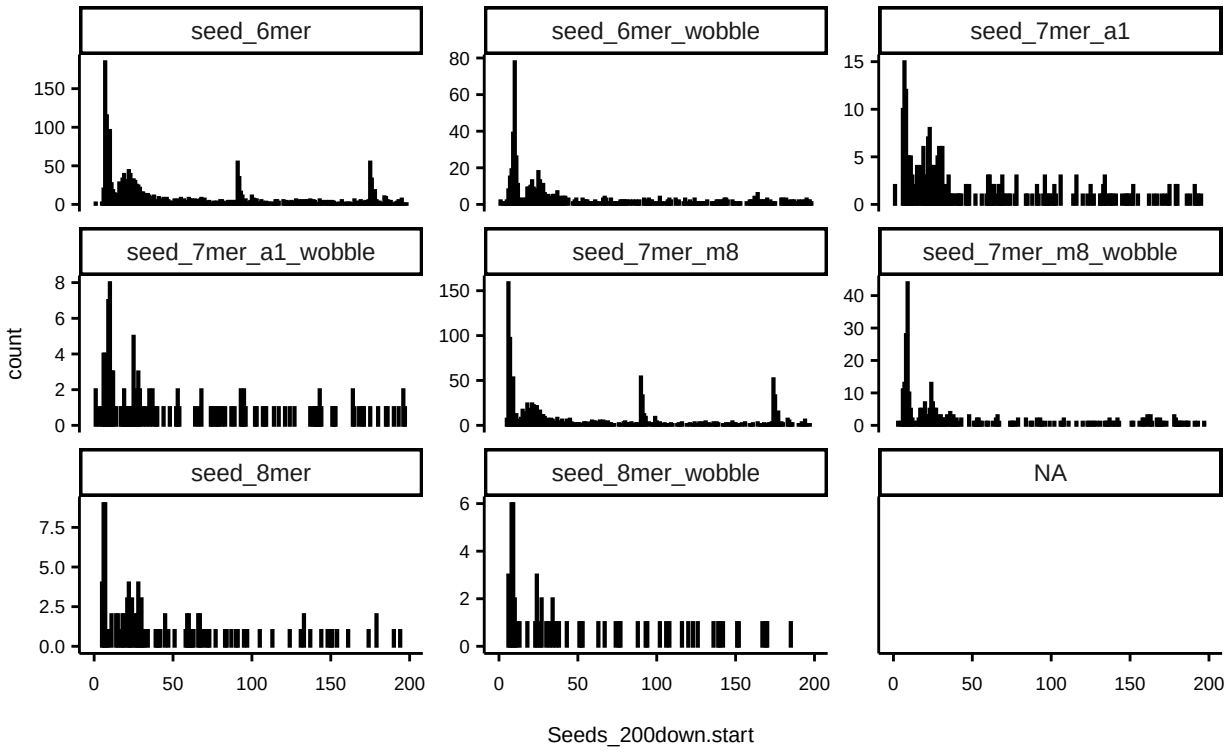

```
p3 <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")), a
  geom_bar(color = "black")+
  theme_paper()+
  facet_wrap(~first_seed_200down.type)+
  ggtitle("mir181 seed positions",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
p3
```



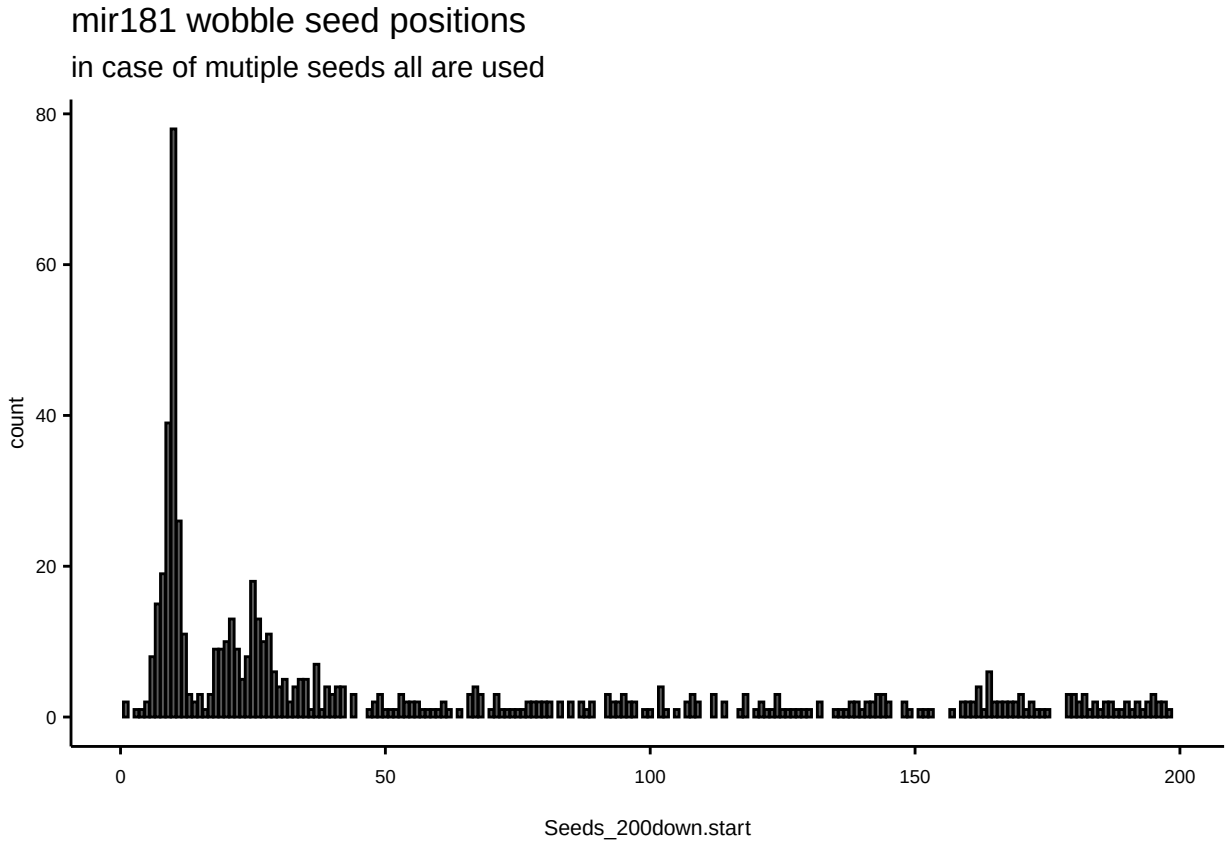

```
# all seeds per BS

p4 <- ggplot(unnest(mir181_bs) %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enrich
  geom_bar(color = "black")+
  theme_paper()+
  ylim(c(0,350))
p4+
  ggtitle("mir181 seed positions",
          subtitle = "in case of mutiple seeds all are used")
```



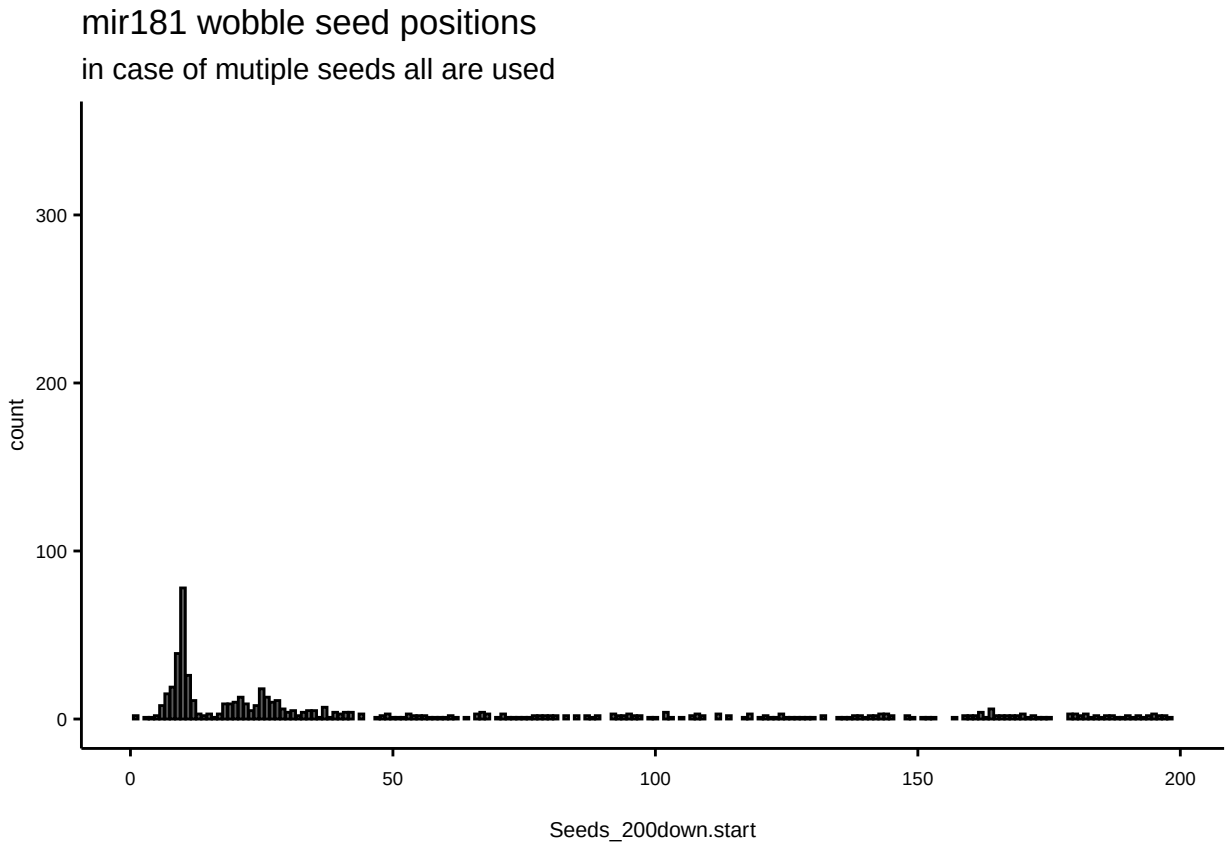

```
p5 <- ggplot(unnest(mir181_bs) %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enrich
  geom_bar(color = "black")+
  theme_paper()+
  facet_wrap(~Seeds_200down.type, scales = "free_y")+
  ggtitle("mir181 seed positions",
          subtitle = "in case of mutiple seeds all are used")
p5
```




```
p6 <- ggplot(unnest(mir181_bs) %>% subset(Seeds_200down.type == "seed_6mer_wobble") %>% subset(set %in%
  geom_bar(color = "black")+
  theme_paper()
p6+
  ggtitle("mir181 wobble seed positions",
          subtitle = "in case of mutiple seeds all are used")
```




```
p7 <- ggplot(unnest(mir181_bs) %>% subset(Seeds_200down.type == "seed_6mer_wobble") %>% subset(set %in%
  geom_bar(color = "black")+
  theme_paper()+
  ylim(c(0,350))

p7+
  ggtitle("mir181 wobble seed positions",
          subtitle = "in case of mutiple seeds all are used")
```




```
nrow(mir181_bs)
```

```
## [1] 10989
```

```
ggsave(p, filename = paste0(out, "seed_versions_Figure2E.pdf"), width = 6, height = 6, units = "cm"  )
ggsave(p4, filename = paste0(out, "seed_position_after_BS.pdf"), width = 6, height = 6, units = "cm"  )
ggsave(p6, filename = paste0(out, "wobbleseed_position_after_BS.pdf"), width = 6, height = 4, units = "
ggsave(p7, filename = paste0(out, "wobbleseed_position_after_BS_achse_angepasst.pdf"), width = 6, heigh
```

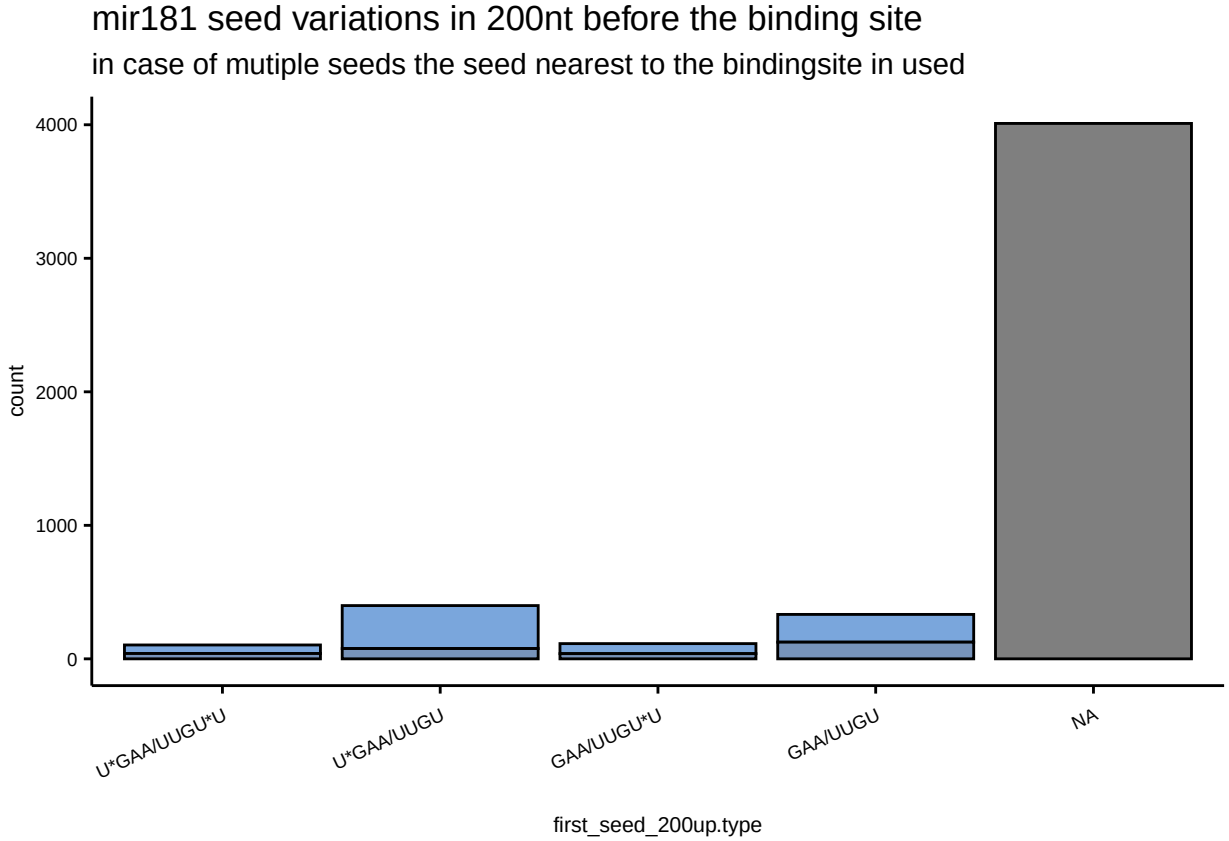
5.2 200nt before the binding site

```
########################
# get seed in 200er window upstream
########################

# get seqeunce 200nt upstream of binding site
mir181_bs_200up <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T) %>% resize(., width = 20
mir181_bs_200up_seq <- getSeq(mir181_bs_200up, x = BSgenome.Mmusculus.UCSC.mm10) %>%
  RNAStringSet()

# count occurences of all seed variations
Seeds_200up <- lapply(seed_list, function(x) {
  vmatchPattern(pattern = x, mir181_bs_200up_seq) %>%
  lapply(., function(x) as.data.frame(x))})
```

```
Seeds_200up <- map(Seeds_200up,
                   ~map2(.x, BS_ID_list, ~mutate(.x, mir181BS_ID = .y) ) %>%
               map_dfr(~.x))


# add the seed type names and make one df of all
Seeds_200up <- map2(Seeds_200up, seed_names_list, ~mutate(.x, seed = .y) ) %>% map_dfr(~.x)


# extract wobble positions
Seeds_1_per_BS <- Seeds_200up %>%
  mutate(wobble = grepl("wobble", seed),
         seed = case_when(wobble  ~ substr(seed, 1, nchar(seed)-7), T ~ seed))


# order seeds by importance
Seeds_1_per_BS$seed <- factor(Seeds_1_per_BS$seed, levels = seed_importance_order )

# select 1 seed per BS --> closest seed with highest importance
Seeds_1_per_BS <- Seeds_1_per_BS %>%
   group_by(mir181BS_ID) %>%
  arrange(start, seed ) %>%
  dplyr::slice(1) %>%
  ungroup(.)

#########################
# combine the closest seed, and all found seeds to the Binding site data.frame
#########################

# add all as list column

colnames(Seeds_200up) <- c("Seeds_200up.start",
                           "Seeds_200up.end",
                           "Seeds_200up.width",
                           "mir181BS_ID",
                           "Seeds_200up.type")


mir181_bs <- left_join(mir181_bs, Seeds_200up, by = "mir181BS_ID") %>%
  tidyr::nest(all_seeds_200up = c("Seeds_200up.start",
                           "Seeds_200up.end",
                           "Seeds_200up.width",
                           "Seeds_200up.type"))

# add closest mir
colnames(Seeds_1_per_BS) <- c("first_seed_200up.start",
                           "first_seed_200up.end",
                           "first_seed_200up.width",
                           "mir181BS_ID",
                           "first_seed_200up.type",
                           "first_seed_200up.wobble")

mir181_bs <- left_join(mir181_bs, Seeds_1_per_BS, by = "mir181BS_ID")
```

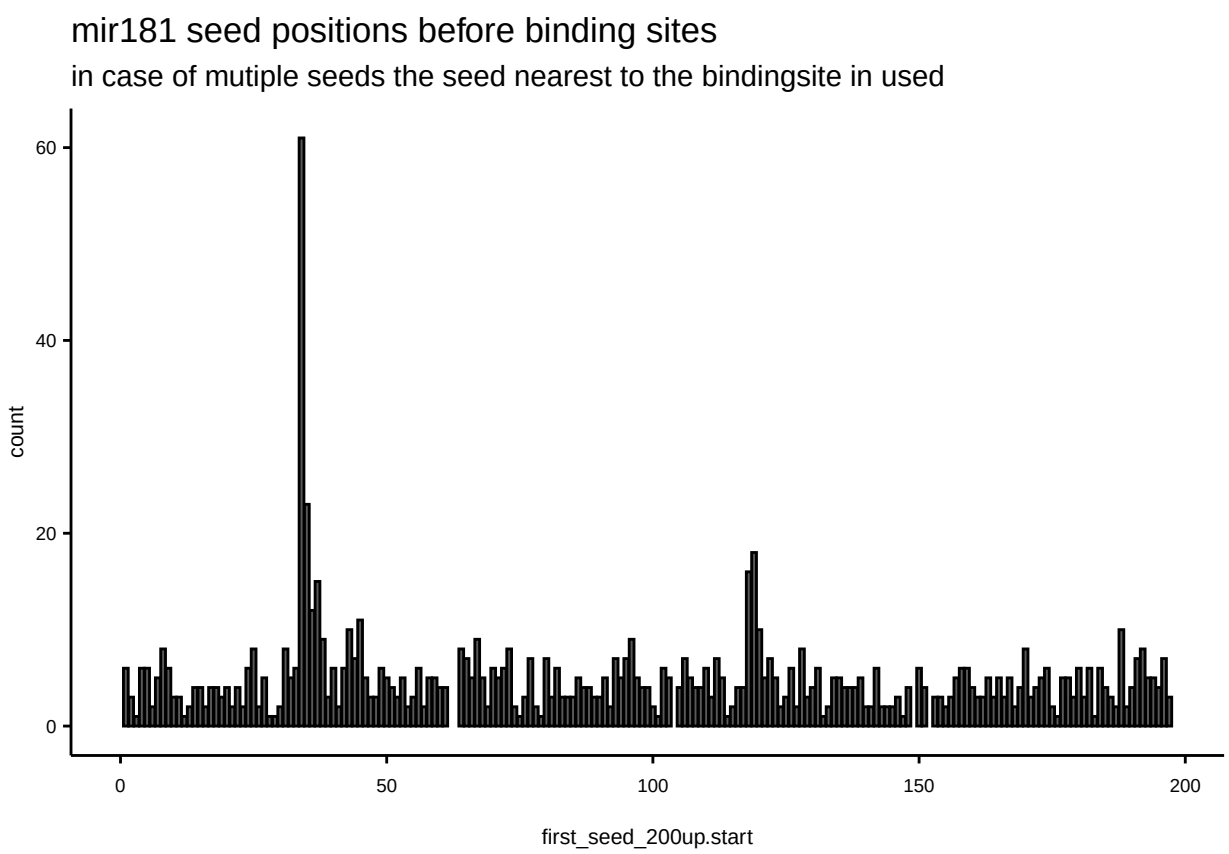
```
mir181_bs <- mir181_bs %>%
  rowwise() %>%
  mutate(seed_repetitions.200down = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer"),
         seed_repetitions.200down.wobble = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer_wobble
```

```
####################
# plots
####################

# plot seed variations
p <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")), a
  geom_bar(color = "black")+
  theme_paper()+
  scale_fill_manual(values = c(farbe4, darken(farbe4)))+
  theme(legend.position = "None") +
  scale_x_discrete(labels=c(seed_8mer = "U*GAA/UUGU*U",
                            seed_7mer_m8 = "U*GAA/UUGU",
                            seed_7mer_a1 = "GAA/UUGU*U",
                            seed_6mer = "GAA/UUGU"),
                   guide = guide_axis(angle = 25))

p +
  ggtitle("mir181 seed variations in 200nt before the binding site",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
```

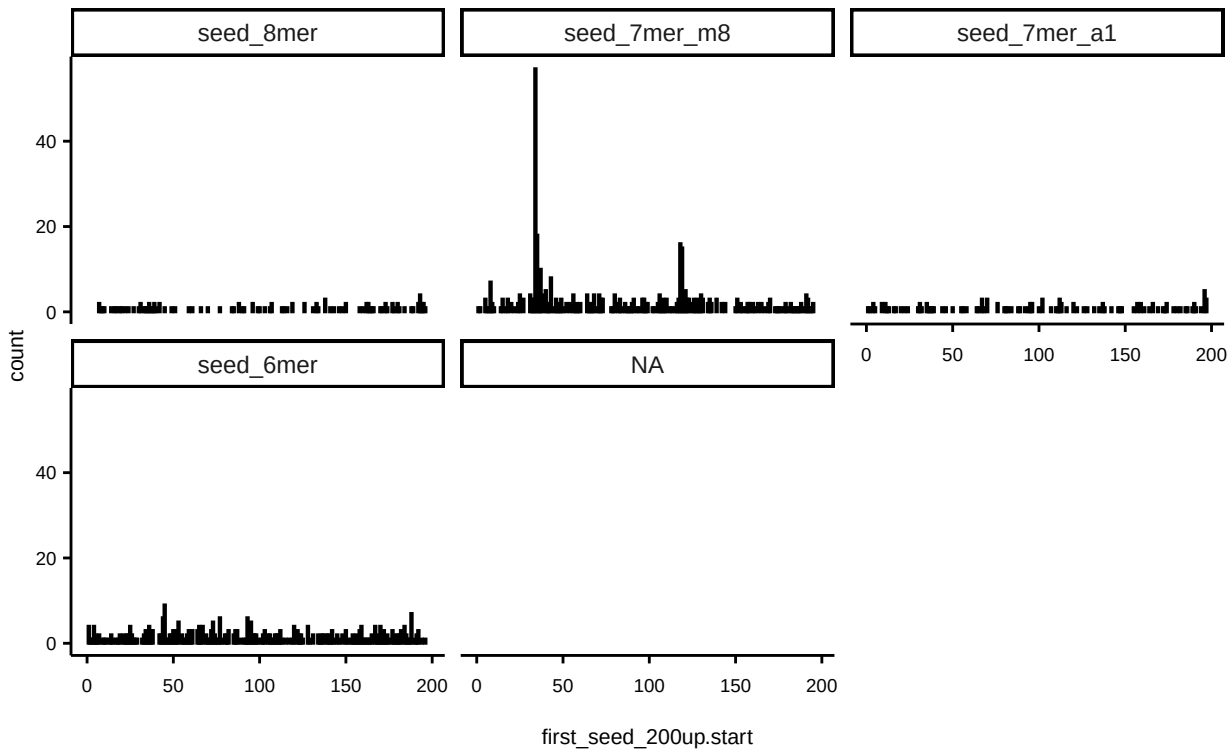



```
# plot seed distributions
p2 <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")), a
```

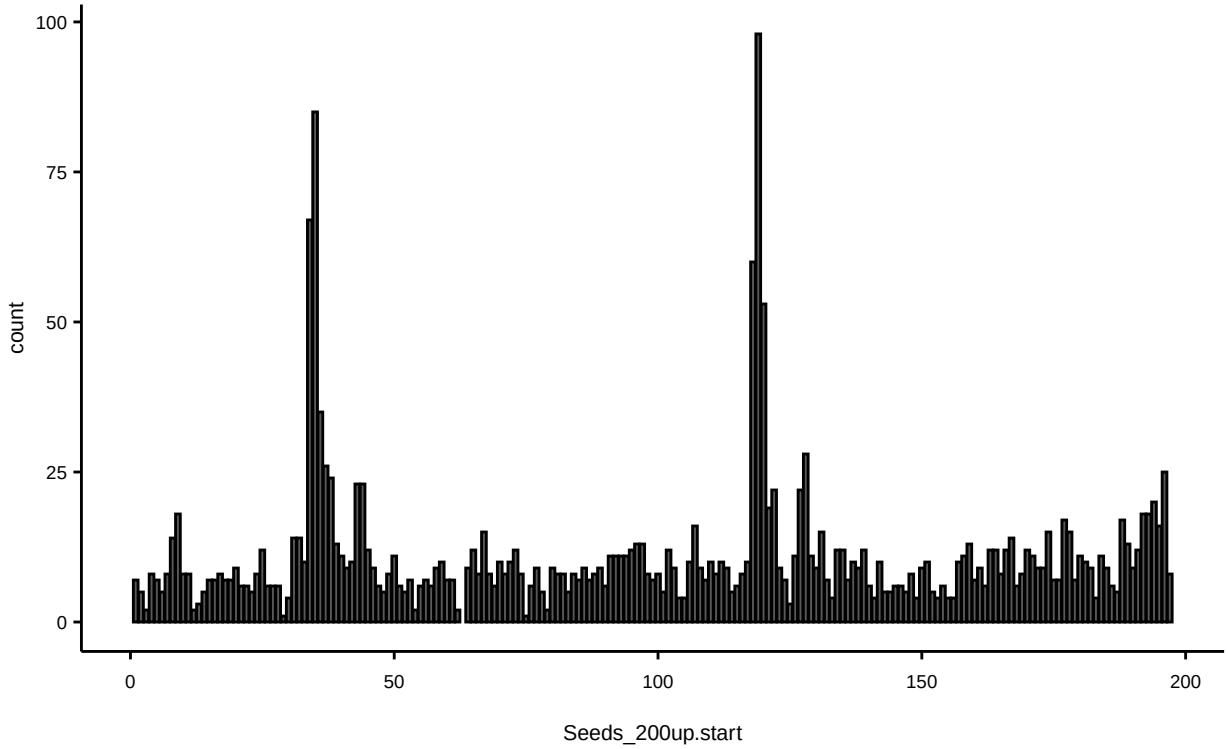
```
  geom_bar(color = "black")+
  theme_paper()+
  ggtitle("mir181 seed positions before binding sites",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
p2
```



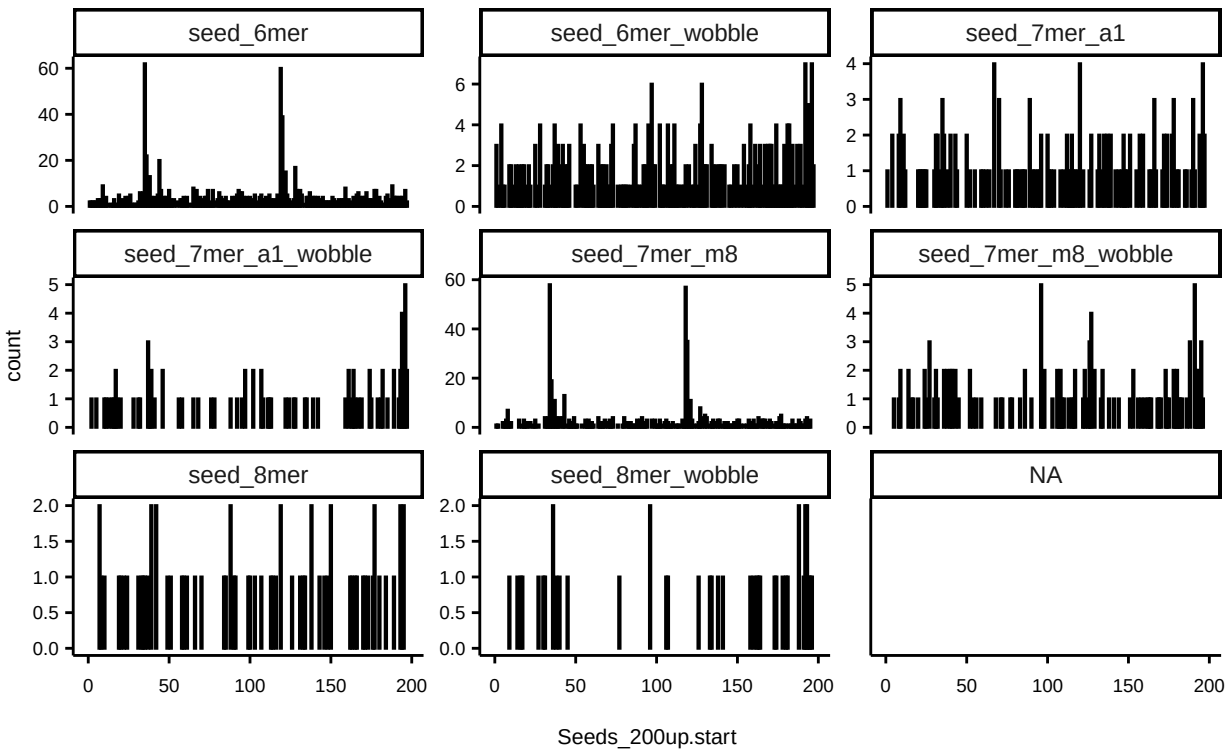

```
p3 <- ggplot(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")), a
  geom_bar(color = "black")+
  theme_paper()+
  facet_wrap(~first_seed_200up.type)+
  ggtitle("mir181 seed positions before binding sites",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")
p3
```




```
# all seeds per BS

p4 <- ggplot(unnest(mir181_bs, cols = all_seeds_200up) %>% subset(set %in% c("ago_bs_mir181_chi&mir181_
  geom_bar(color = "black")+
  theme_paper()
p4+
  ggtitle("mir181 seed positions before binding sites",
          subtitle = "in case of mutiple seeds all are used")
```




```
p5 <- ggplot(unnest(mir181_bs, cols = all_seeds_200up) %>% subset(set %in% c("ago_bs_mir181_chi&mir181_
  geom_bar(color = "black")+
  theme_paper()+
  facet_wrap(~Seeds_200up.type, scales = "free_y")+
  ggtitle("mir181 seed positions",
          subtitle = "in case of mutiple seeds all are used")
p5
```




```
#ggsave(p, filename = paste0(out, "seed_versions_Figure2E.pdf"), width = 6, height = 6, units = "cm"  )
```

5.2.1 percent binding sites with a seed downstream

```
nrow(mir181_bs %>% subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")) %>% subse
```

```
## [1] 0.1778142
```

6 Save table

```
saveRDS(mir181_bs, file = paste0(out, "mir181_bs_with_seeds.rds"))
```