

# Seed motifs

Melina Klostermann

02 June, 2023

## Contents

1	Libraries and settings	1
2	What was done?	2
3	Files	2
4	Seed position and distribution	3
5	Save table	15

## 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 <- "#008EAOFF"
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 adjacent nt, and a 8mer with two adjacent nts.

## 3 Files

```

# -----
# MREs
# -----

mir181_bs <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/1")
mir_crosslinks <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/xx_down_stream_R/02_BS_definition_WT_r")
load("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/xx_down_stream_R/02_BS_definition_WT_r")

nrow(mir181_bs)

## [1] 10473

#####
# 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, seed_7mer_m8_wobble, seed_7mer_a1_wobble)

seed_names_list <- list("seed_8mer", "seed_7mer_m8", "seed_7mer_a1", "seed_6mer", "seed_6mer_wobble", "seed_8mer_wobble", "seed_7mer_m8_wobble", "seed_7mer_a1_wobble")

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

```

## 4 Seed position and distribution

### 4.1 200nt after the binding site

```
#####  
# get seed in 200er window  
#####  
  
# get sequence 200nt downstream of binding site  
mir181_bs_200down <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T) %>% resize(., width = 200)  
mir181_bs_200down_seq <- getSeq(mir181_bs_200down, x = BSgenome.Mmusculus.UCSC.mm10) %>%  
  RNAStringSet()  
  
# count occurrences 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"))

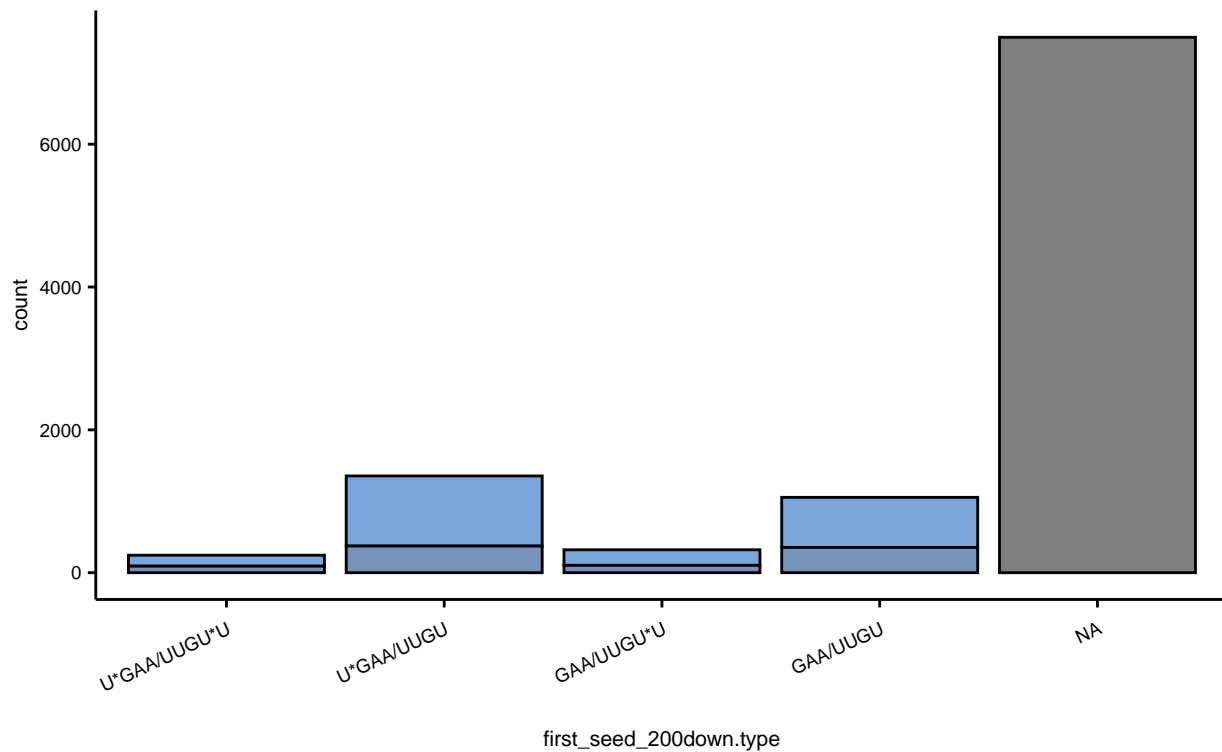
#####
# plots
#####

# plot seed variations
p <- ggplot(mir181_bs, aes(x = first_seed_200down.type, fill = first_seed_200down.wobble))+
  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))+
  ggtitle("mir181 seed variations in 200nt after the binding site",
    subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")

```

p

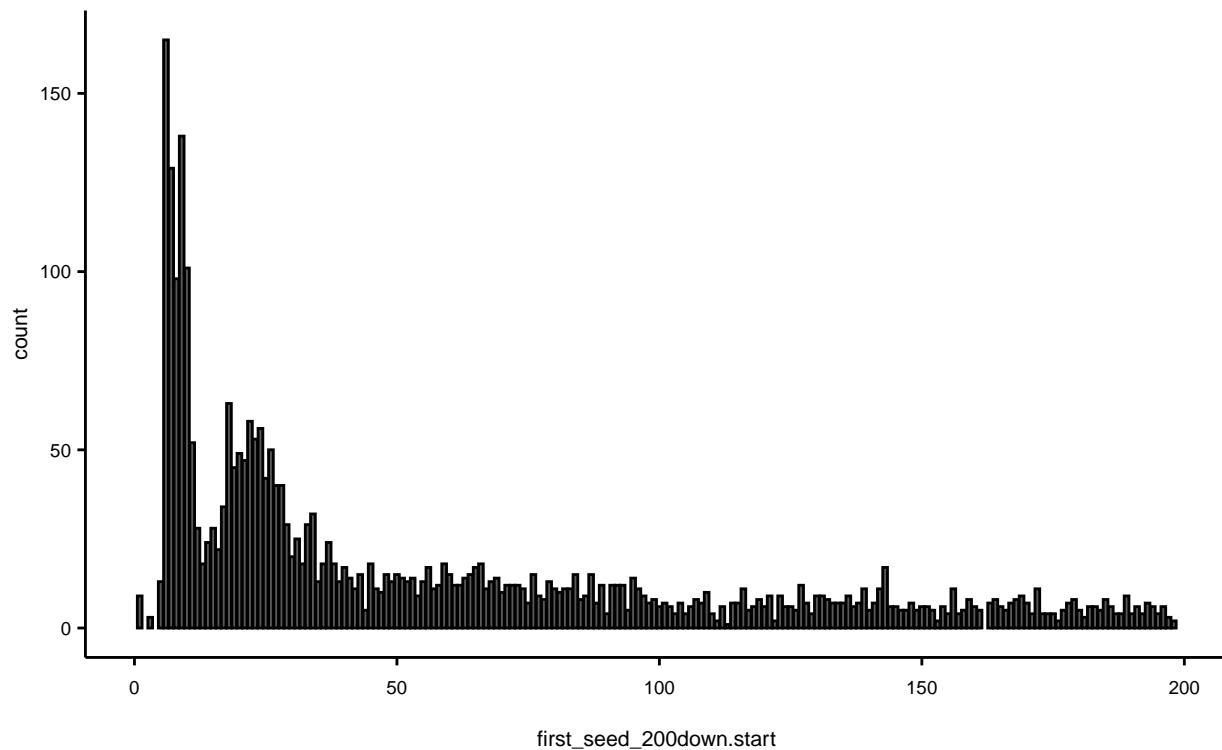
mir181 seed variations in 200nt after the binding site  
in case of mutiple seeds the seed nearest to the bindingsite in used



```
# plot seed distributions
p2 <- ggplot(mir181_bs, aes(x = first_seed_200down.start))+
  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
```

## mir181 seed positions

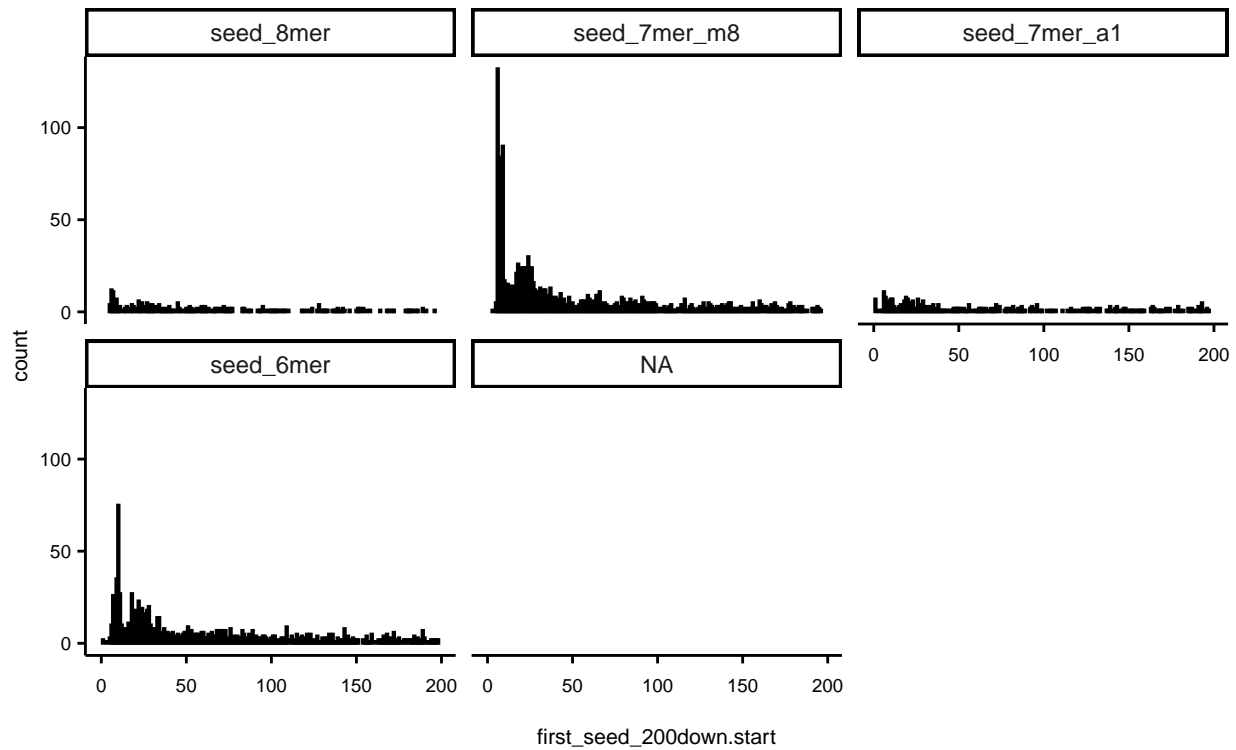
in case of mutiple seeds the seed nearest to the bindingsite in used



```
p3 <- ggplot(mir181_bs, aes(x = first_seed_200down.start))+  
  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
```

## mir181 seed positions

in case of mutiple seeds the seed nearest to the bindingsite in used

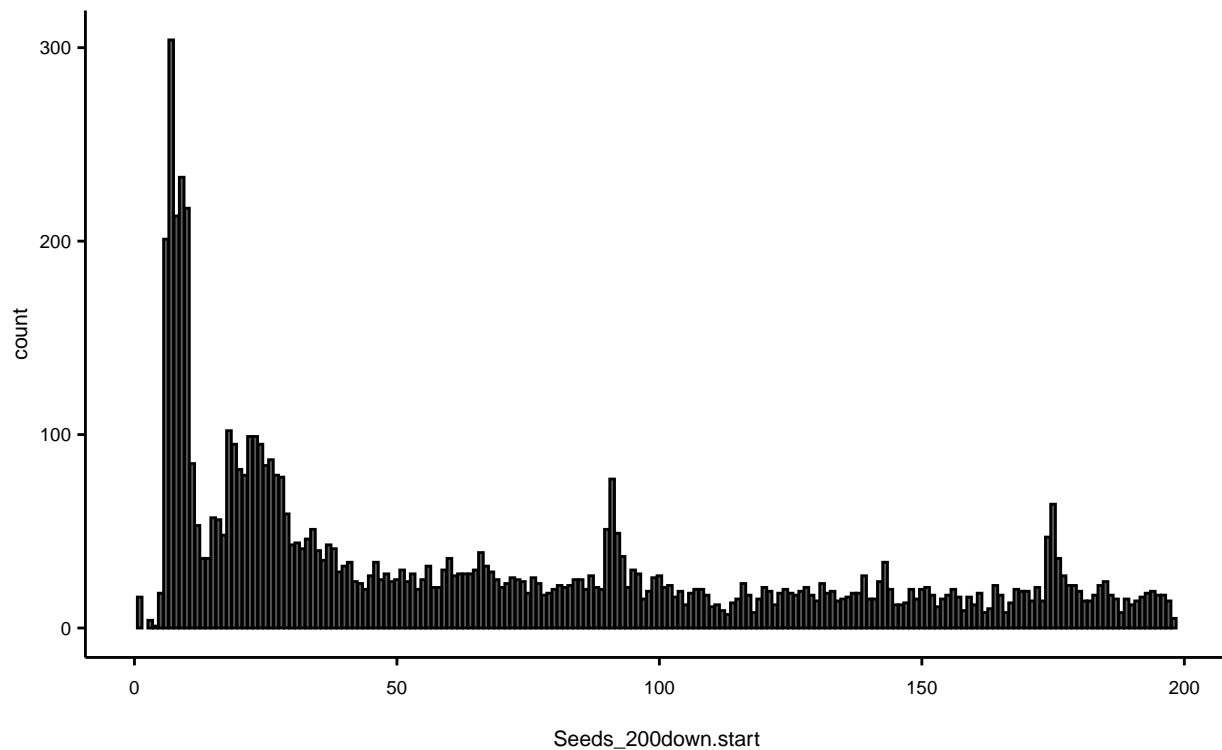


```
# all seeds per BS
```

```
p4 <- ggplot(unnest(mir181_bs), aes(x = Seeds_200down.start))+  
  geom_bar(color = "black")+  
  theme_paper()+  
  ggtitle("mir181 seed positions",  
    subtitle = "in case of mutiple seeds all are used")  
p4
```

## mir181 seed positions

in case of mutiple seeds all are used



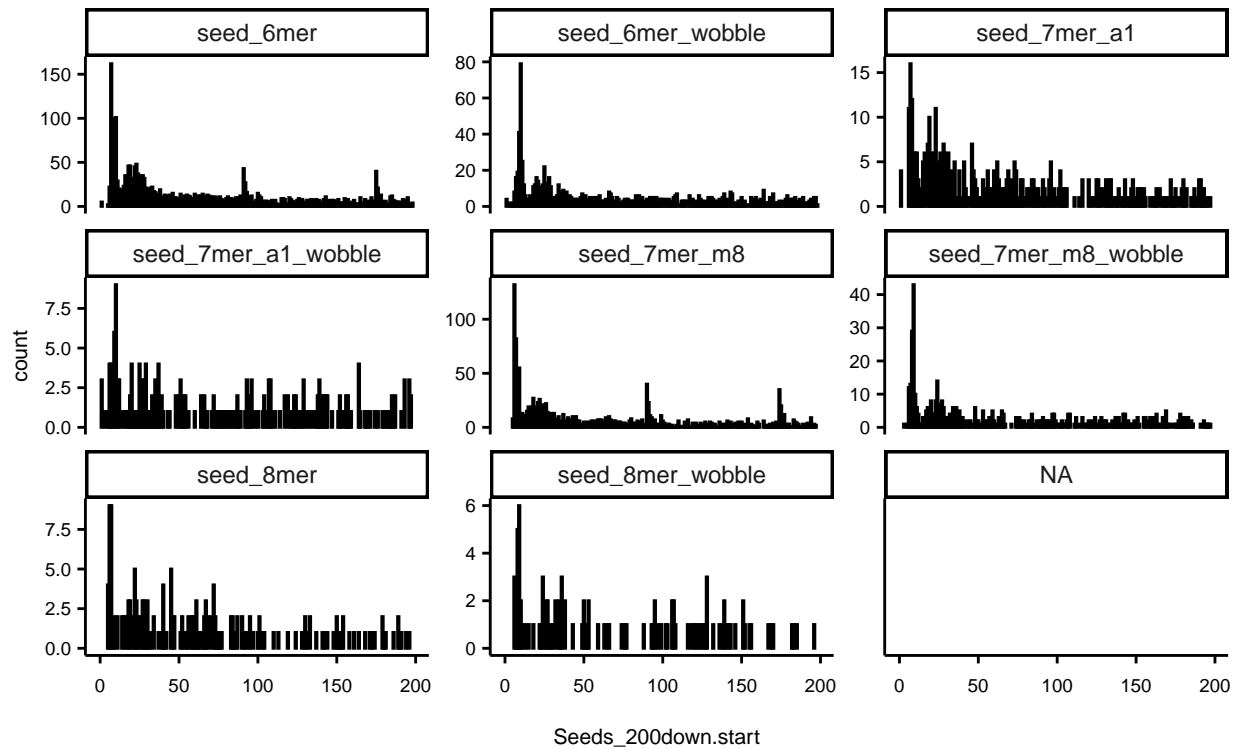
```
p5 <- ggplot(unnest(mir181_bs), aes(x = Seeds_200down.start))+  
  geom_bar(color = "black")+  
  theme_paper()+  
  facet_wrap(~Seeds_200down.type, scales = "free_y")+  
  ggtitle("mir181 seed positions",  
          subtitle = "in case of mutiple seeds all are used")
```

p5



## mir181 seed positions

in case of mutiple seeds all are used



```
nrow(mir181_bs)
```

```
## [1] 10473
```

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

## 4.2 200nt before the binding site

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

# get sequeunce 200nt upstream of binding site
mir181_bs_200up <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T) %>% resize(., width = 200)
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(., BS_ID_list, ~mutate(., 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")

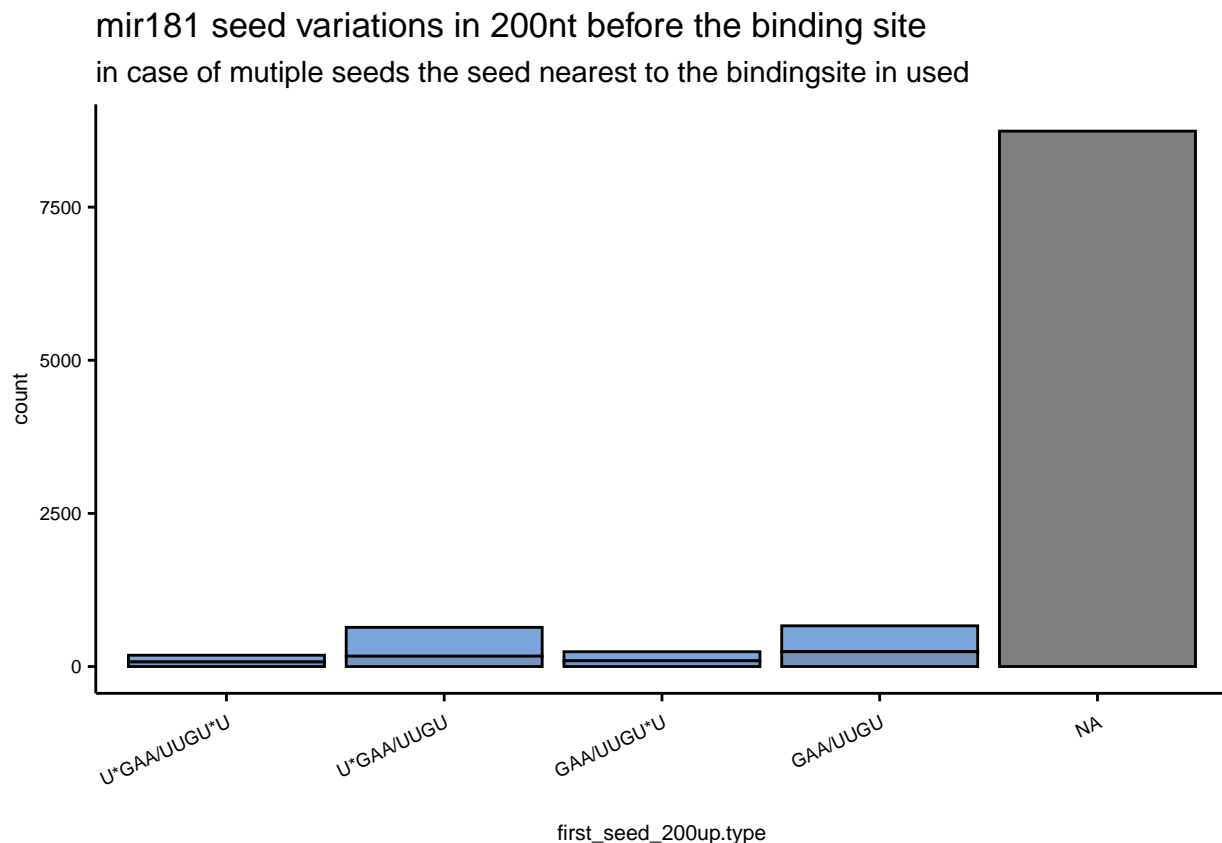
mir181_bs <- mir181_bs %>%
  rowwise() %>%
  mutate(seed_repetitions.200down = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer"),
         seed_repetitions.200down.wobble = sum(all_seeds_200down$Seeds_200down.type == "seed_6mer_wobble"))

```

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

# plot seed variations
p <- ggplot(mir181_bs, aes(x = first_seed_200up.type, fill = first_seed_200up.wobble))+
  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))+
  ggtitle("mir181 seed variations in 200nt before the binding site",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")

p
```

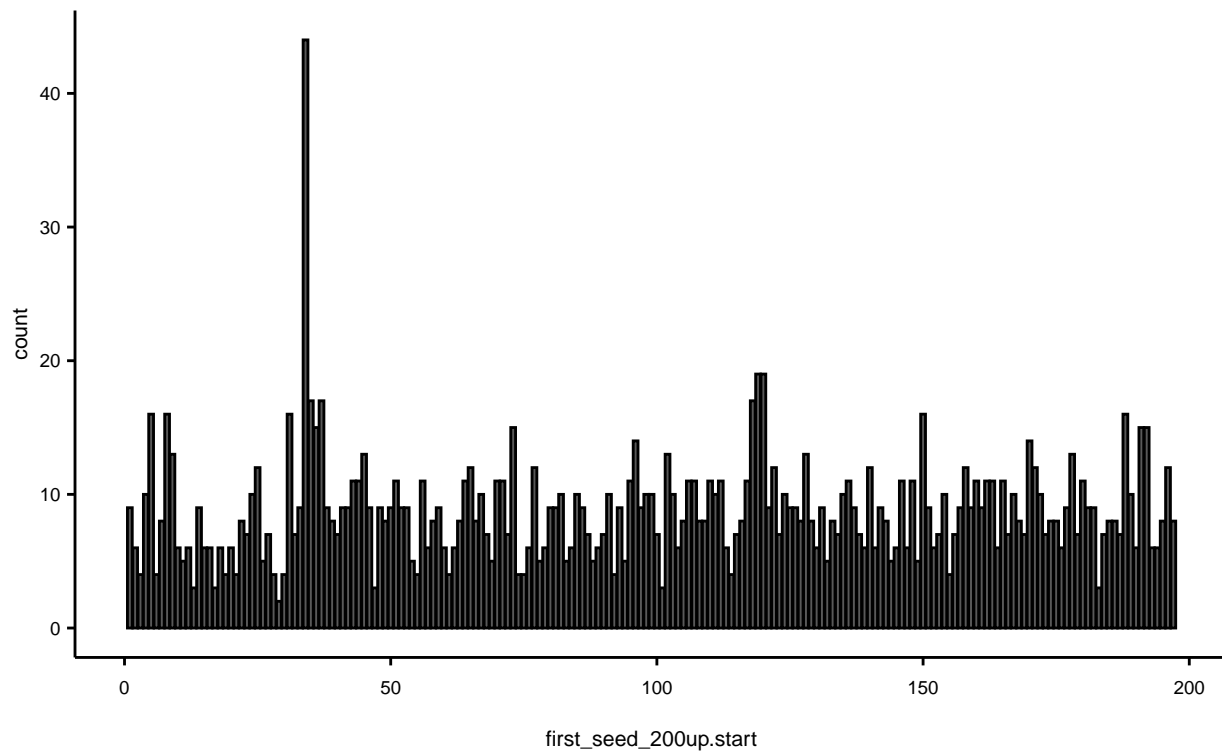


```
# plot seed distributions
p2 <- ggplot(mir181_bs, aes(x = first_seed_200up.start))+
  geom_bar(color = "black")+
  theme_paper()+
  ggtitle("mir181 seed positions before binding sites",
          subtitle = "in case of mutiple seeds the seed nearest to the bindingsite in used")

p2
```

## mir181 seed positions before binding sites

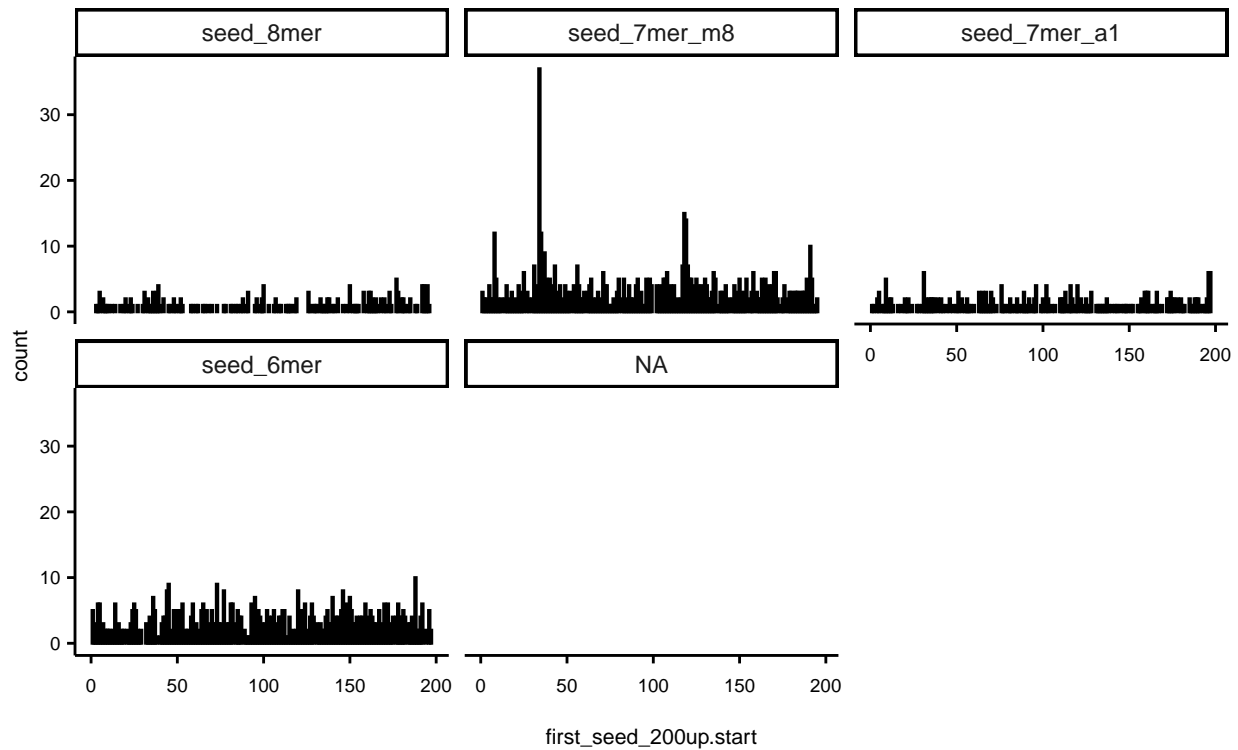
in case of mutiple seeds the seed nearest to the bindingsite in used



```
p3 <- ggplot(mir181_bs, aes(x = first_seed_200up.start))+  
  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
```

## mir181 seed positions before binding sites

in case of mutiple seeds the seed nearest to the bindingsite in used

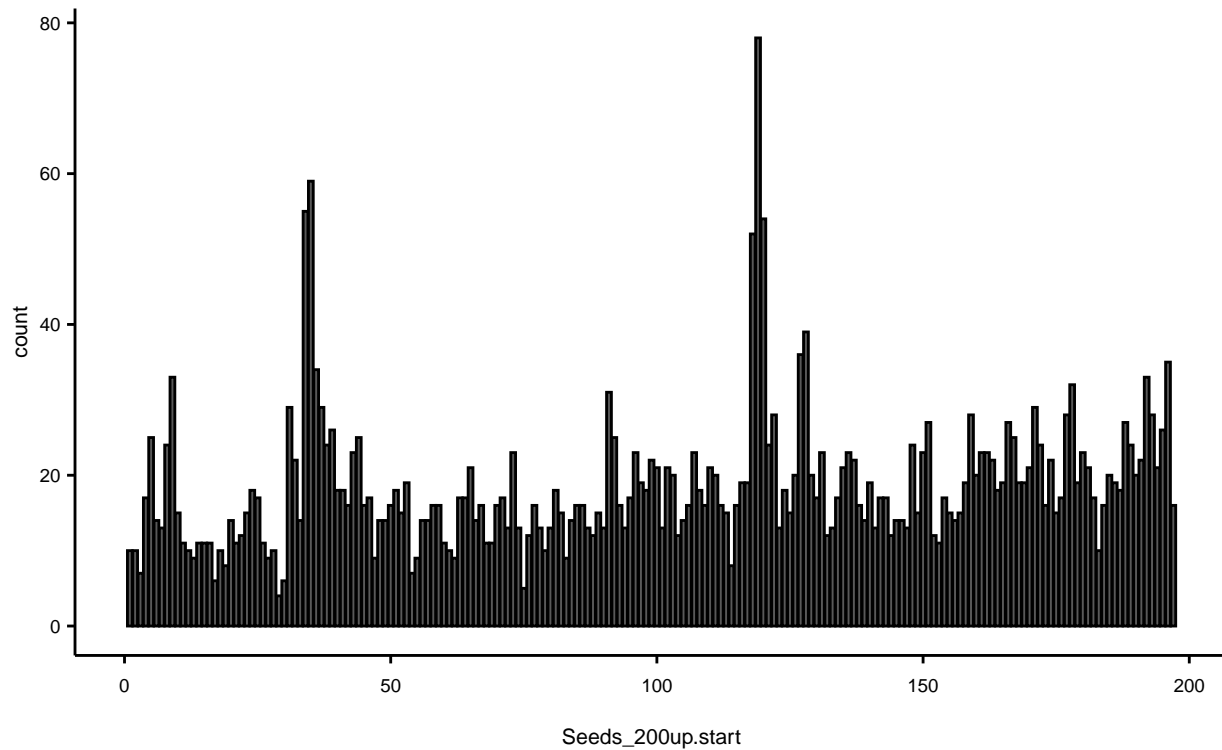


*# all seeds per BS*

```
p4 <- ggplot(unnest(mir181_bs, cols = all_seeds_200up), aes(x = Seeds_200up.start))+
  geom_bar(color = "black")+
  theme_paper()+
  ggtitle("mir181 seed positions before binding sites",
    subtitle = "in case of mutiple seeds all are used")
p4
```

## mir181 seed positions before binding sites

in case of mutiple seeds all are used

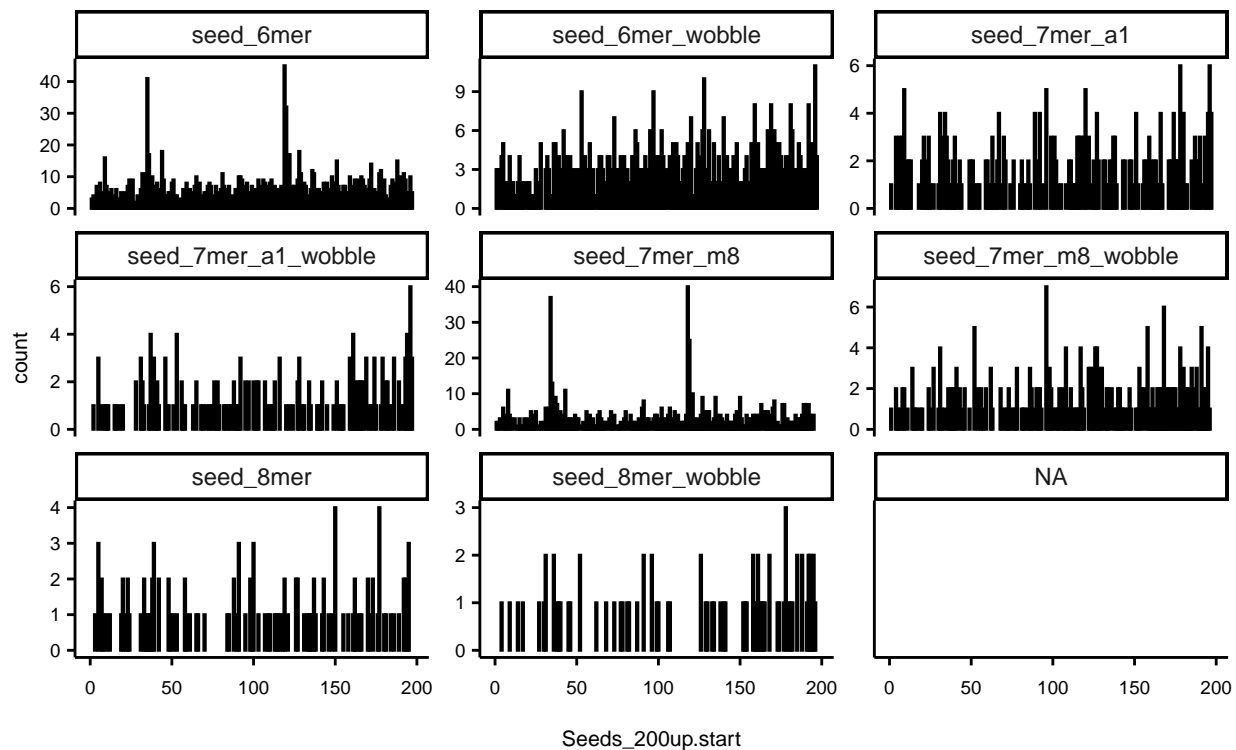


```
p5 <- ggplot(unnest(mir181_bs, cols = all_seeds_200up), aes(x = Seeds_200up.start))+  
  geom_bar(color = "black")+  
  theme_paper()+  
  facet_wrap(~Seeds_200up.type, scales = "free_y")+  
  ggtitle("mir181 seed positions",  
          subtitle = "in case of mutiple seeds all are used")
```

p5

## mir181 seed positions

in case of mutiple seeds all are used



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

### 4.2.1 percent binding sites with a seed downstream

```
nrow(mir181_bs %>% subset(!is.na(first_seed_200down.start)))/ nrow(mir181_bs)
```

```
## [1] 0.2841593
```

## 5 Save table

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