Fig4 ECDF plots

Nikita Verheyden

2023-05-16

Setup

dir

```
setwd("D:/Krueger_Lab/Publications/miR181_paper/Figure4")
```

packages

```
library(ggplot2)
library(rtracklayer)
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
## Loading required package: GenomeInfoDb
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
```

```
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
```

```
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union
```

```
## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union
```

```
## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

data

```
#Ribo profiling
RNA <- read.csv("D:/Krueger_Lab/Publications/miR181_paper/Figure3/RNA_masterframe.csv")
RPF <- read.csv("D:/Krueger_Lab/Publications/miR181_paper/Figure3/RPF_masterframe.csv")

#load the gtf file to compare genes
gff23 <- import.gff3("D:/Krueger_Lab/Ribo_Profiling/run15112022M23/ref_genome/gencode.vM23.annotation.g

#targets
larget <- readRDS("D:/Krueger_Lab/Publications/miR181_paper/Figure3/mir181_bs_with_seeds.rds")
largetframe <- as.data.frame(larget)
#MMsat4
repeat_masker <- readRDS("D:/Krueger_Lab/Publications/miR181_paper/Figure2/MMsat4/repeat_masker.rds")
MMSAT4 <- repeat_masker[repeat_masker$repName == "MMSAT4"]
```

colours

```
#colours
farbeneg <- "#b4b4b4"
farbe1 <- "#0073C2FF"
farbe2 <- "#EFC000FF"
farbe3 <- "#CD534CFF"
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"

RNApcol <- "#b56504"
RNAncol <- "#027d73"
RPFpcol <- "#c4c404"
RPFncol <- "#8d0391"
```

inspect targetdata

We're keeping all of those targets for now but will analyze the in ecdf plots

```
table(largetframe$set)
```

```
##
##                 ago_bs_mir181_chi ago_bs_mir181_chi&mir181_enriched
##                              5815                              1082
##                   mir181_enriched
##                              3576
```

```
colnames(largetframe)
```

```
##  [1] "seqnames"                      "start"
##  [3] "end"                           "width"
##  [5] "strand"                        "scoreSum"
##  [7] "scoreMean"                     "scoreMax"
##  [9] "geneType"                      "geneName"
## [11] "geneID"                        "region"
## [13] "mir_IP"                        "n_mir181"
## [15] "n_mir181a"                     "n_mir181b"
## [17] "n_mir181c"                     "n_mir181d"
## [19] "set"                           "mir181BS_ID"
## [21] "WT"                            "KO"
## [23] "geneID.2"                      "geneName.1"
## [25] "region.1"                      "counts.bs.1_KO"
## [27] "counts.bs.2_KO"                "counts.bs.3_KO"
## [29] "counts.bs.4_WT"                "counts.bs.5_WT"
```

```
## [31] "counts.bs.6_WT"                  "geneID.1"
## [33] "counts.bg.1_KO"                  "counts.bg.2_KO"
## [35] "counts.bg.3_KO"                  "counts.bg.4_WT"
## [37] "counts.bg.5_WT"                  "counts.bg.6_WT"
## [39] "resBs.baseMean"                  "resBs.log2FoldChange"
## [41] "resBs.lfcSE"                     "resBs.stat"
## [43] "resBs.pvalue"                    "resBs.padj"
## [45] "resBg.baseMean"                  "resBg.log2FoldChange"
## [47] "resBg.lfcSE"                     "resBg.stat"
## [49] "resBg.pvalue"                    "resBg.padj"
## [51] "tpm.counts.bg.1_KO"              "tpm.counts.bg.2_KO"
## [53] "tpm.counts.bg.3_KO"              "tpm.counts.bg.4_WT"
## [55] "tpm.counts.bg.5_WT"              "tpm.counts.bg.6_WT"
## [57] "BS_ID"                           "tpm_support_KO"
## [59] "tpm_support_WT"                  "tpm_supported"
## [61] "down"                            "all_seeds_200down"
## [63] "first_seed_200down.start"        "first_seed_200down.end"
## [65] "first_seed_200down.width"        "first_seed_200down.type"
## [67] "first_seed_200down.wobble"       "seed_repetitions.200down"
## [69] "seed_repetitions.200down.wobble" "all_seeds_200up"
## [71] "first_seed_200up.start"          "first_seed_200up.end"
## [73] "first_seed_200up.width"          "first_seed_200up.type"
## [75] "first_seed_200up.wobble"         "seed_repetitions.200up"
## [77] "seed_repetitions.200up.wobble"
```

ECDF plots

each code chunk is a split of the main target table that is then used for a specific ecdf plot

datasets

```
#RNA
RNA$targetset <- "Non-target"
RNA$targetset[RNA$gene_symbol %in% largetframe[largetframe$set == "ago_bs_mir181_chi", "geneName"]] <-
RNA$targetset[RNA$gene_symbol %in% largetframe[largetframe$set == "mir181_enriched", "geneName"]] <- "mi
RNA$targetset[RNA$gene_symbol %in% largetframe[largetframe$set == "ago_bs_mir181_chi&mir181_enriched",

table(RNA$targetset)
```

```
##
## ago_bs_mir181_chi              both   mir181_enriched        Non-target
##               783               667              1521             10330
```

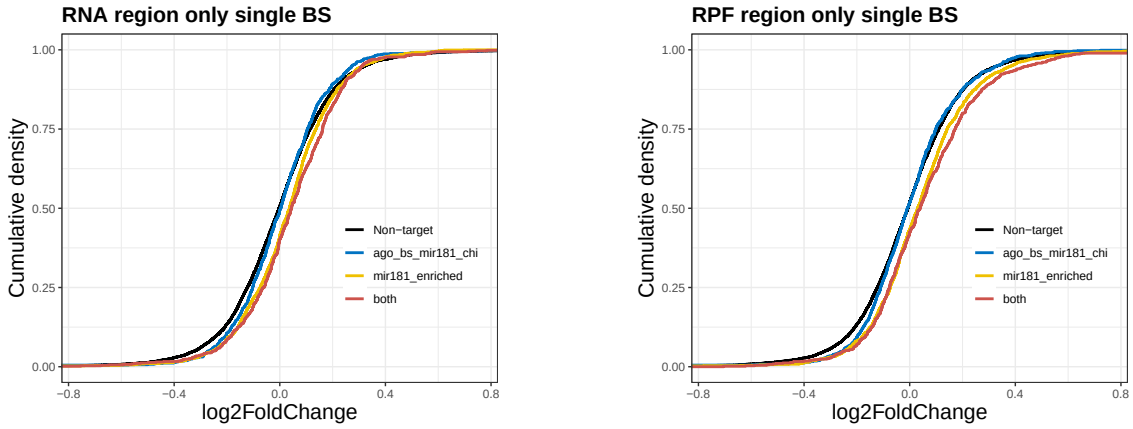
```
#RPF
RPF$targetset <- "Non-target"
RPF$targetset[RPF$gene_symbol %in% largetframe[largetframe$set == "ago_bs_mir181_chi", "geneName"]] <-
RPF$targetset[RPF$gene_symbol %in% largetframe[largetframe$set == "mir181_enriched", "geneName"]] <- "mi
RPF$targetset[RPF$gene_symbol %in% largetframe[largetframe$set == "ago_bs_mir181_chi&mir181_enriched",

table(RPF$targetset)
```

```
##
## ago_bs_mir181_chi              both   mir181_enriched        Non-target
##               782               667              1508              8412
```

```
# ecdf plots
#RNA
setECDFRNA <- ggplot(RNA, aes(as.numeric(log2FoldChange), colour=factor(targetset, levels = c("Non-targ
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RNA region only single BS")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
setECDFRNA

#RPF
setECDFRPF <- ggplot(RPF, aes(as.numeric(log2FoldChange), colour=factor(targetset, levels = c("Non-targ
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RPF region only single BS")

setECDFRPF
```



region (single targets)

```
#get number of binding sites per gene to be able to sort for singles
bsnum <- as.data.frame(table(largetframe$geneName))
colnames(bsnum) <- c("geneName", "BS_number")

#RNA
```

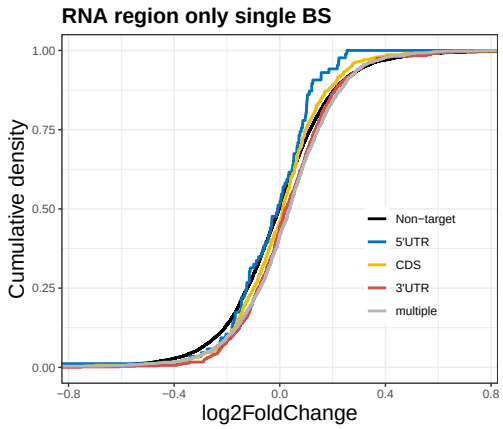
```
RNA$region_single <- "Non-target"
RNA$region_single[RNA$gene_symbol %in% largetframe[largetframe$region == "utr5", "geneName"]] <- "5'UTR
RNA$region_single[RNA$gene_symbol %in% largetframe[largetframe$region == "cds", "geneName"]] <- "CDS"
RNA$region_single[RNA$gene_symbol %in% largetframe[largetframe$region == "utr3", "geneName"]] <- "3'UTR
RNA$region_single[RNA$gene_symbol %in% bsnum[bsnum$BS_number > 1, "geneName"]] <- "multiple"

table(RNA$region_single)
```

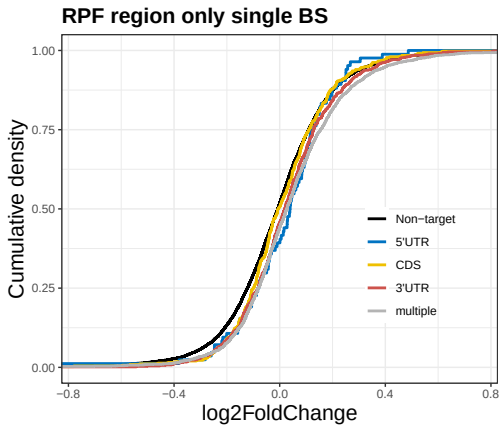
```
##
##      3'UTR      5'UTR        CDS   multiple Non-target
##        659         86        451       1775      10330
```

```
#RPF
RPF$region_single <- "Non-target"
RPF$region_single[RPF$gene_symbol %in% largetframe[largetframe$region == "utr5", "geneName"]] <- "5'UTR
RPF$region_single[RPF$gene_symbol %in% largetframe[largetframe$region == "cds", "geneName"]] <- "CDS"
RPF$region_single[RPF$gene_symbol %in% largetframe[largetframe$region == "utr3", "geneName"]] <- "3'UTR
RPF$region_single[RPF$gene_symbol %in% bsnum[bsnum$BS_number > 1, "geneName"]] <- "multiple"

table(RPF$region_single)
```

```
##
##      3'UTR      5'UTR        CDS   multiple Non-target
##        656         84        450       1767       8412
```

```
# ECDF plots
#RNA
regsingECDFRNA <- ggplot(RNA, aes(as.numeric(log2FoldChange), colour=factor(region_single, levels = c("
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3, farbeneg)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RNA region only single BS")

regsingECDFRNA

#RPF
regsingECDFRPF <- ggplot(RPF, aes(as.numeric(log2FoldChange), colour=factor(region_single, levels = c("
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3, farbeneg)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RPF region only single BS")

regsingECDFRPF
```

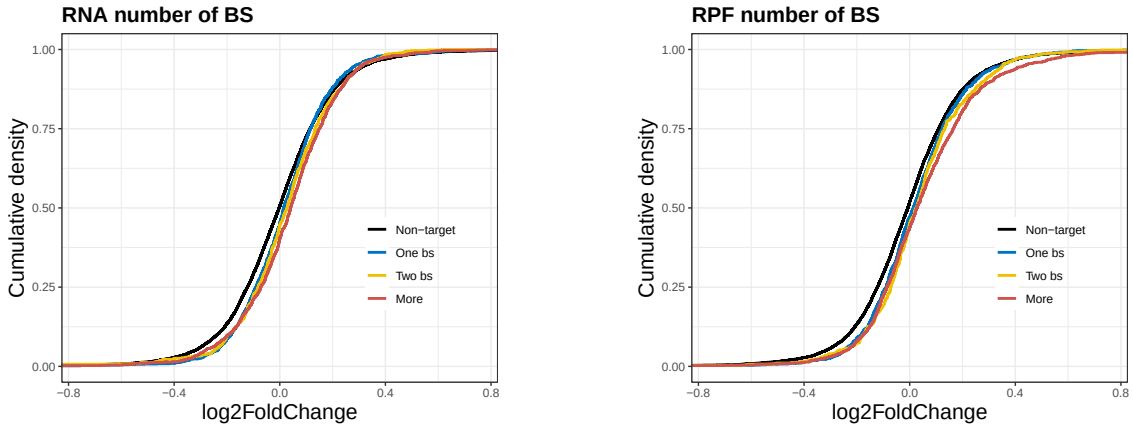




number of target sites

```
colnames(bsnum) <- c("gene_symbol", "BS_number")

#RNA
RNAnum <- left_join(RNA, bsnum, by="gene_symbol")
RNAnum$BS_number[is.na(RNAnum$BS_number)] <- "Non-target"
RNAnum$BS_num_plot <- ifelse(RNAnum$BS_number == "Non-target", "Non-target",
                             ifelse(RNAnum$BS_number == 1, "One bs",
                                    ifelse(RNAnum$BS_number == 2, "Two bs", "More")))


#RPF
RPFnum <- left_join(RPF, bsnum, by="gene_symbol")
RPFnum$BS_number[is.na(RPFnum$BS_number)] <- "Non-target"
RPFnum$BS_num_plot <- ifelse(RPFnum$BS_number == "Non-target", "Non-target",
                             ifelse(RPFnum$BS_number == 1, "One bs",
                                    ifelse(RPFnum$BS_number == 2, "Two bs", "More")))

#ecdf plots
#RNA
numECDFRNA <- ggplot(RNAnum, aes(as.numeric(log2FoldChange), colour=factor(BS_num_plot, levels = c("Non
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RNA number of BS")

numECDFRNA

#RPF
numECDFRPF <- ggplot(RPFnum, aes(as.numeric(log2FoldChange), colour=factor(BS_num_plot, levels = c("Non
  stat_ecdf(geom="step", size=1) +
  scale_colour_manual(values = c("black", farbe1, farbe2, farbe3)) +
  coord_cartesian(xlim = c(-0.75, 0.75)) + theme_bw() +
  theme(legend.position = c(0.8, 0.35), legend.title = element_blank(),
```

```
        legend.background = element_rect(colour = "transparent", fill="transparent"),
        axis.title=element_text(size=16),plot.title = element_text(size=16, face = "bold"), aspect.ratio
  scale_y_continuous("Cumulative density") + scale_x_continuous("log2FoldChange") +
  ggtitle("RPF number of BS")

numECDFRPF
```



session info

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
## [3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.utf8
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] dplyr_1.1.2          rtracklayer_1.58.0   GenomicRanges_1.50.2
## [4] GenomeInfoDb_1.34.9  IRanges_2.32.0       S4Vectors_0.36.2
## [7] BiocGenerics_0.44.0  ggplot2_3.4.2
##
## loaded via a namespace (and not attached):
##  [1] SummarizedExperiment_1.28.0 tidyselect_1.2.0
##  [3] xfun_0.39                   lattice_0.20-45
##  [5] colorspace_2.1-0            vctrs_0.6.2
##  [7] generics_0.1.3              htmltools_0.5.4
##  [9] yaml_2.3.7                  utf8_1.2.3
## [11] XML_3.99-0.14               rlang_1.1.0
```

```
## [13] pillar_1.9.0               glue_1.6.2
## [15] withr_2.5.0                BiocParallel_1.32.6
## [17] matrixStats_0.63.0         GenomeInfoDbData_1.2.9
## [19] lifecycle_1.0.3            zlibbioc_1.44.0
## [21] MatrixGenerics_1.10.0      Biostrings_2.66.0
## [23] munsell_0.5.0              gtable_0.3.3
## [25] codetools_0.2-19           evaluate_0.21
## [27] restfulr_0.0.15            labeling_0.4.2
## [29] Biobase_2.58.0             knitr_1.42
## [31] fastmap_1.1.1              parallel_4.2.3
## [33] fansi_1.0.4                scales_1.2.1
## [35] DelayedArray_0.23.2        XVector_0.38.0
## [37] farver_2.1.1               Rsamtools_2.14.0
## [39] rjson_0.2.21               digest_0.6.31
## [41] BiocIO_1.8.0               grid_4.2.3
## [43] cli_3.6.0                  tools_4.2.3
## [45] bitops_1.0-7               magrittr_2.0.3
## [47] RCurl_1.98-1.12            tibble_3.2.1
## [49] crayon_1.5.2               pkgconfig_2.0.3
## [51] Matrix_1.5-3               rmarkdown_2.21
## [53] rstudioapi_0.14            R6_2.5.1
## [55] GenomicAlignments_1.34.1   compiler_4.2.3
```