Figure 5

Sergio Garcia-Moreno Alcantara (sga@stowers.org)

August 24, 2021

Aim

The aim of this analysis is to investigate whether there are differences in the occupancy of the SAGA core module at different promoter types in Drosophila kc167 cells

Enviroment setup

Set working directory and load required libraries and lab functions

```
setwd("/n/projects/sga/analysis/SAGA/saga_publication/")
options(knitr.figure_dir = "plots/figure_5/")

# Standard packages
library(GenomicAlignments)
library(GenomicRanges)
library(Biostrings)
library(BSgenome.Dmelanogaster.UCSC.dm6)
library(TxDb.Dmelanogaster.UCSC.dm6.ensGene)
library(dplyr)
library(reshape2)
library(plyranges)
library(CAGEr)
library(magrittr)
library(ggplot2)
library(cowplot)
library(ggseqlogo)
library(gridExtra)
library(ggpubr)

# Lab sources
source("./shared_code/granges_common.r")
source("./shared_code/metapeak_common.r")
source("./shared_code/knitr_common.r")
```

Analysis

1. Loading samples and necessary data sets

```
tbp_bw <- list(pos = "./bw/kc167_nt_tbp_nexus_1_normalized_positive.bw", neg = "./bw/kc167_nt_tbp_nexus
wda_bw <- list(pos = "./bw/kc167_wda_nexus_2_normalized_positive.bw", neg = "./bw/kc167_wda_nexus_2_nor
adab2_bw <- list(pos = "./bw/kc167_adab2_nexus_1_normalized_positive.bw", neg = "./bw/kc167_adab2_nexus
saf11_bw <- list(pos = "./bw/kc167_saf11_nexus_1_normalized_positive.bw", neg = "./bw/kc167_saf11_nexus
saf6_bw <- list(pos = "./bw/kc167_saf6_nexus_1_normalized_positive.bw", neg = "./bw/kc167_saf6_nexus_1_
spt3_bw <- list(pos = "./bw/kc167_spt3_nexus_1_normalized_positive.bw", neg = "./bw/kc167_spt3_nexus_1_
trf2_bw <- list(pos = "./bw/kc167_dmso_trf2_nexus_merged_normalized_positive.bw",
    neg = "./bw//kc167_dmso_trf2_nexus_merged_normalized_negative.bw")

patchcap_bw <- list(pos = "./bw/kc_patchcap_nexus_positive.bw", neg = "./bw/kc_patchcap_nexus_negative.

motif_list_kc <- get(load("./rdata/motif_list_kc167.RData"))
```

2. Plot a metapeak for each SAGA core subunit at the different promoter types

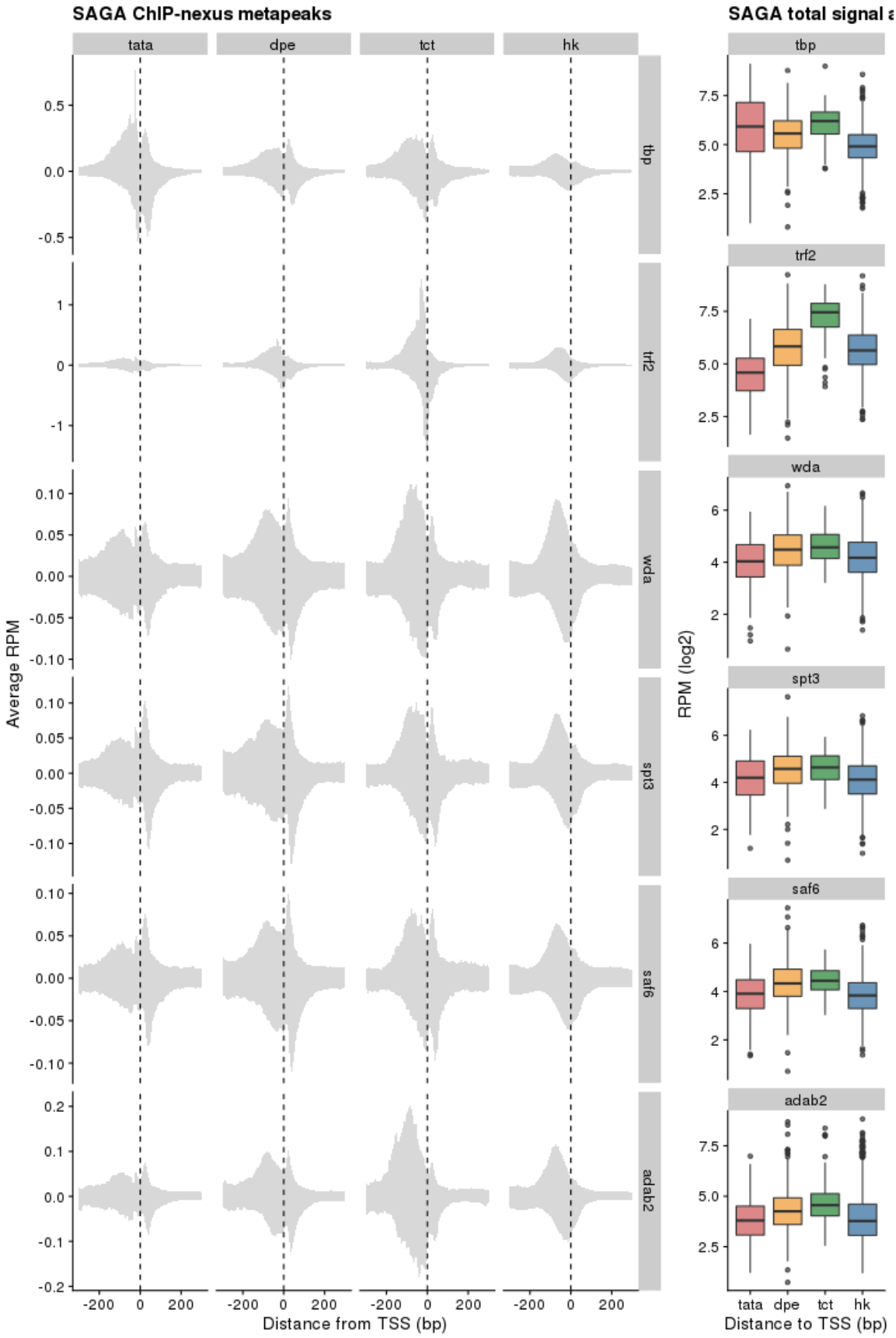
```
bw_list <- list(wda = wda_bw, adab2 = adab2_bw, saf6 = saf6_bw, spt3 = spt3_bw, tbp = tbp_bw,
    trf2 = trf2_bw)

# bw_list <- list(tbp=tbp_bw, patchcap=patchcap_bw)

## Calculate the average signal per factor per base pair at different promoter
## types

promoter_type_metapeak_df <- mclapply(names(motif_list_kc), function(x) {
    motif <- motif_list_kc[[x]]
    mclapply(names(bw_list), function(y) {
        bw <- bw_list[[y]]
        exo_metapeak(motif, bw, 300, 301, paste(x, "at", y), 5)
    }, mc.cores = 5)
}, mc.cores = 5) %>%
    do.call(c, .) %>%
    bind_rows()

promoter_type_metapeak_df %<>%
    mutate(., factor = gsub(".* ", "", promoter_type_metapeak_df$sample_name)) %<>%
    mutate(., motif = gsub(" .*", "", promoter_type_metapeak_df$sample_name))

## Setting the plotting order

sample_levels <- c("tbp", "trf2", "wda", "spt3", "saf6", "adab2")
motif_levels <- c("tata", "dpe", "tct", "hk")

## Create a plotting function

plot_func <- function(df, name, color) {

    df$motif <- factor(df$motif, levels = c(motif_levels))
```

```
    df$factor <- factor(df$factor, levels = c(sample_levels))

    ggplot(df, aes(x = tss_distance, y = reads)) + geom_area(aes(fill = strand),
        alpha = 0.6, show.legend = F) + scale_fill_manual(values = color) + geom_vline(xintercept = 0,
        linetype = 2) + facet_grid(factor ~ motif, scales = "free") + ggtitle(name) +
        theme_cowplot() + theme(plot.title = element_text(size = 15, face = "bold")) +
        xlab("Distance from TSS (bp)") + ylab("Average RPM")
}

metapeak <- plot_func(promoter_type_metapeak_df, "SAGA ChIP-nexus metapeaks", c("grey",
    "grey"))
```

3. Plot the total signal distribution of each SAGA core subunit at the different promoter types

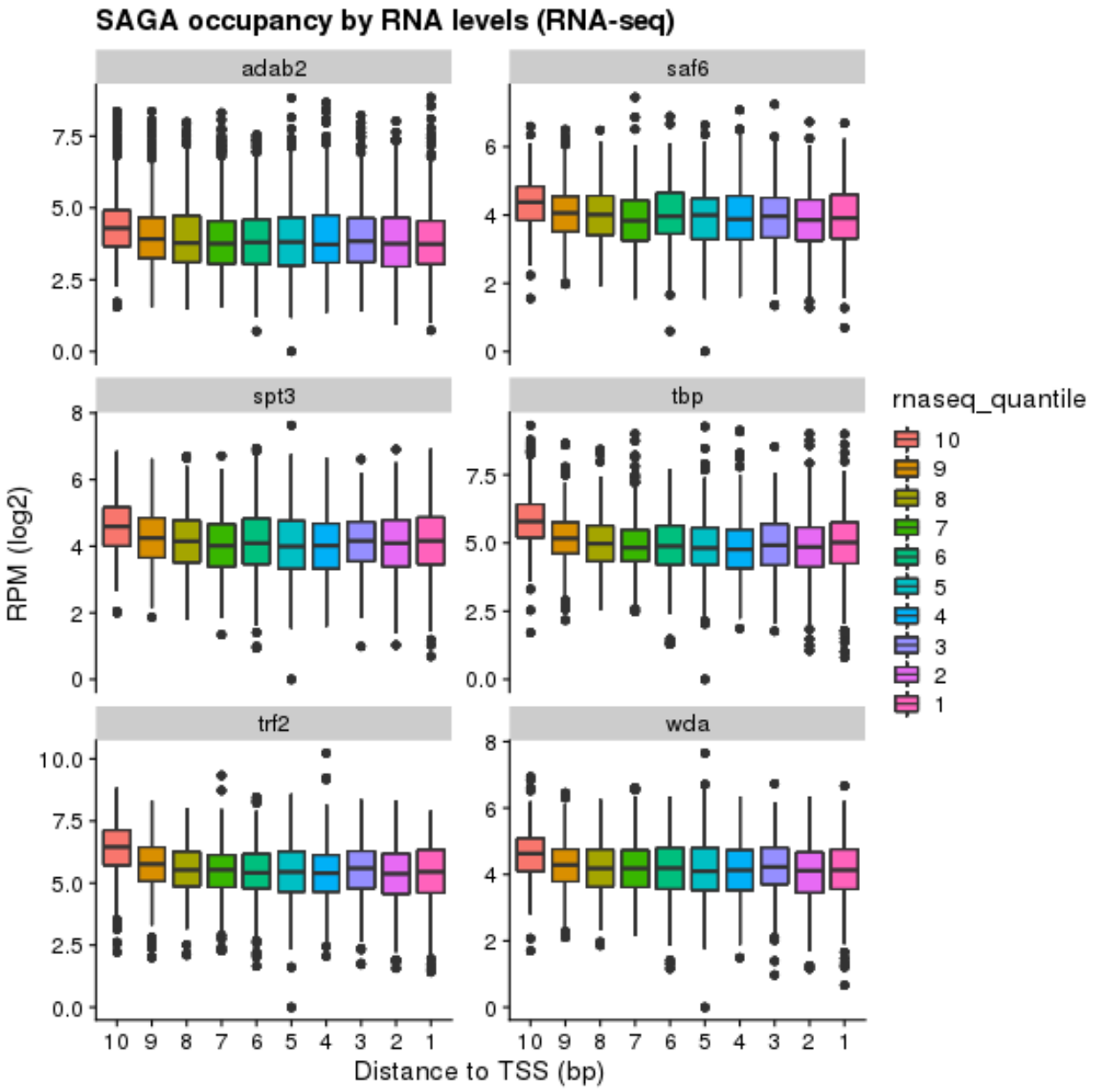
Total signal was calculated as sum of the signal for each promoter and factor in 200 bp window centered at the TSS

```
# Make a data frame containing transcript ID and total signal per gene and
# promoter type

sig_df <- mclapply(names(motif_list_kc), function(x) {

    motif <- motif_list_kc[[x]]
    mclapply(names(bw_list), function(y) {
        bw <- bw_list[[y]]

        df <- data.frame(fb_t_id = motif$fb_t_id, signal = nexus_regionSums(resize(motif,
            201, "center"), bw), sample = y, motif = x)
        df
    }, mc.cores = 4)
}, mc.cores = 4) %>%
    do.call(c, .) %>%
    bind_rows()

sig_df$sample <- factor(sig_df$sample, levels = sample_levels)

sig_df$motif <- factor(sig_df$motif, levels = motif_levels)
sig_df$sample <- factor(sig_df$sample, levels = sample_levels)

boxplot <- ggplot(sig_df, aes(motif, log2(signal + 1), fill = motif)) + geom_boxplot(alpha = 0.7,
    show.legend = F) + theme_cowplot() + scale_fill_manual(values = c("indianred3",
    "#EE962B", "#228232", "#2C699B")) + ggtitle("SAGA total signal at promoter types") +
    facet_wrap(~sample, scales = "free_y", ncol = 1) + theme(plot.title = element_text(size = 15,
    face = "bold")) + xlab("Distance to TSS (bp)") + ylab("RPM (log2)")

# boxplot

grid.arrange(metapeak, boxplot, widths = c(3, 1))
```


SAGA ChIP-nexus metapeaks
tata
dpe
tct
hk
tbp
trf2
wda
spt3
saf6
adab2
Average RPM
Distance from TSS (bp)
SAGA total signal a
RPM (log2)
tata dpe tct hk
Distance to TSS (bp)

4. Plot distribution of SAGA occupancy levels across quantiles of RNA-seq expression data

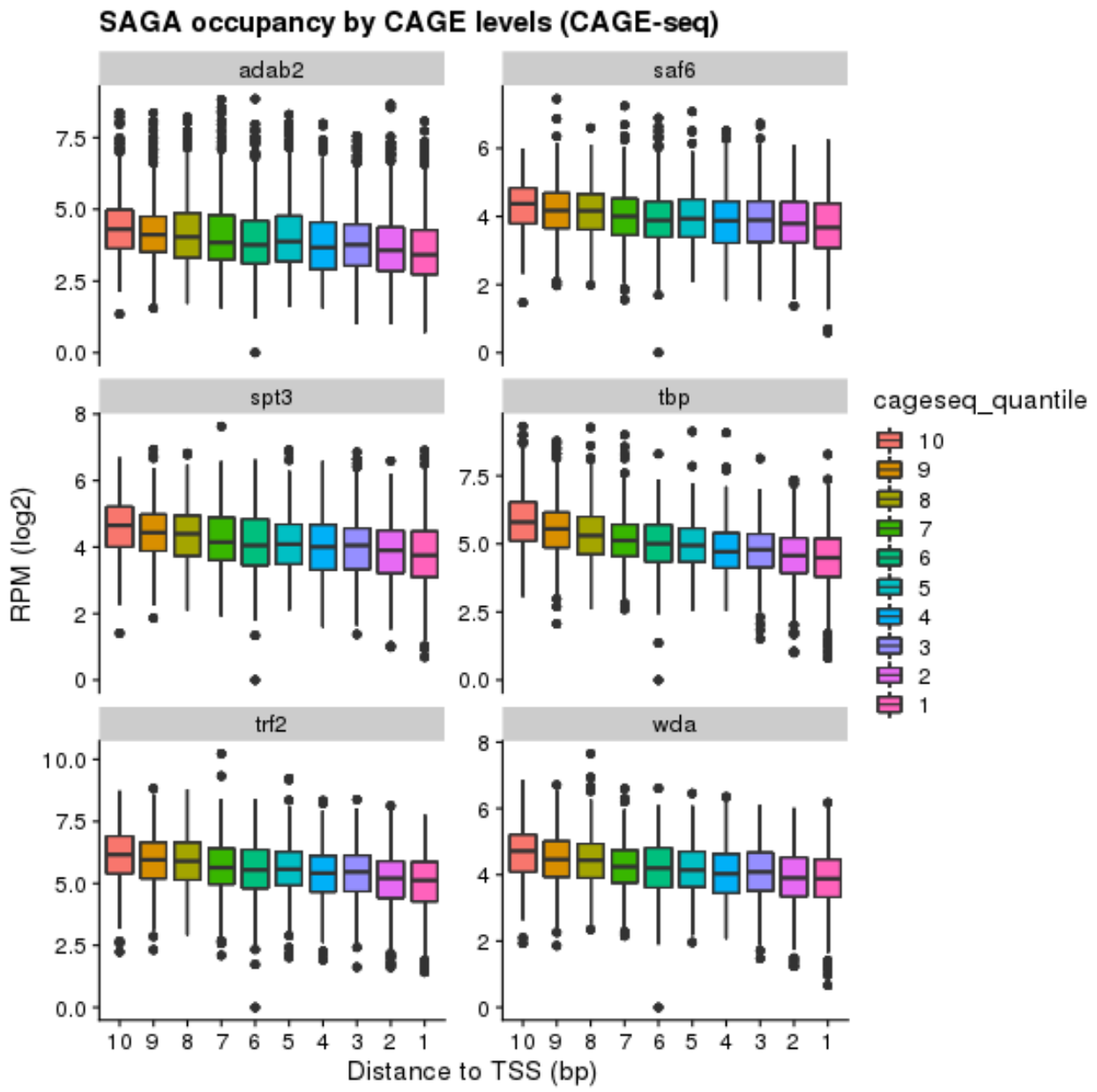
```
tss <- get(load("./rdata/cage_kc167_tss_pLaw_3tpm.RData"))
rna_tss <- tss[order(tss$RNAseq_tpm, decreasing = T)]
rna_tss$rnaseq_quantile <- ntile(rna_tss$RNAseq_tpm, 10)

# Make a data frame containing transcript ID and total signal per gene and
# promoter type

sig_df <- mclapply(levels(as.factor(rna_tss$rnaseq_quantile)), function(x) {

    quantile_gr <- subset(rna_tss, rnaseq_quantile == x)
    mclapply(names(bw_list), function(y) {
        bw <- bw_list[[y]]

        df <- data.frame(fb_t_id = quantile_gr$fb_t_id, signal = nexus_regionSums(resize(quantile_gr,
            201, "center"), bw), sample = y, rnaseq_quantile = x)
        df
    }, mc.cores = 4)
}, mc.cores = 4) %>%
    do.call(c, .) %>%
    bind_rows()

sig_df$rnaseq_quantile <- factor(sig_df$rnaseq_quantile, levels = c("10", "9", "8",
    "7", "6", "5", "4", "3", "2", "1"))
# sig_df$sample <- factor(sig_df$sample, levels = c('wda', 'saf6', 'spt3',
# 'ada2b')) sig_df<-filter(sig_df,signal>=0)

boxplot <- ggplot(sig_df, aes(rnaseq_quantile, log2(signal + 1), fill = rnaseq_quantile)) +
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
ggtitle("SAGA occupancy by RNA levels (RNA-seq)") + facet_wrap(~sample, scales = "free_y",
    ncol = 2) + theme(plot.title = element_text(size = 15, face = "bold")) + xlab("Distance to TSS (bp)
    ylab("RPM (log2)")

boxplot
```



5. Plot distribution of SAGA occupancy levels across quantiles of CAGE-seq expression data

```
cage_tss <- tss[order(tss$score, decreasing = T)]
cage_tss$cageseq_quantile <- ntile(cage_tss$score, 10)

# Make a data frame containing transcript ID and total signal per gene and
# promoter type

sig_df <- mclapply(levels(as.factor(cage_tss$cageseq_quantile)), function(x) {

    quantile_gr <- subset(cage_tss, cageseq_quantile == x)
```

```
    mclapply(names(bw_list), function(y) {
        bw <- bw_list[[y]]

        df <- data.frame(fb_t_id = quantile_gr$fb_t_id, signal = nexus_regionSums(resize(quantile_gr,
            201, "center"), bw), sample = y, cageseq_quantile = x)
        df
    }, mc.cores = 4)
}, mc.cores = 4) %>%
    do.call(c, .) %>%
    bind_rows()

sig_df$cageseq_quantile <- factor(sig_df$cageseq_quantile, levels = c("10", "9",
    "8", "7", "6", "5", "4", "3", "2", "1"))
# sig_df$sample <- factor(sig_df$sample, levels = c('wda', 'saf6', 'spt3',
# 'ada2b')) sig_df<-filter(sig_df,signal>=0)

boxplot <- ggplot(sig_df, aes(cageseq_quantile, log2(signal + 1), fill = cageseq_quantile)) +
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
    geom_boxplot(alpha = 0.7) + theme_cowplot() + #scale_fill_manual(values = c('indianred3', '#EE962B'
ggtitle("SAGA occupancy by CAGE levels (CAGE-seq)") + facet_wrap(~sample, scales = "free_y",
    ncol = 2) + theme(plot.title = element_text(size = 15, face = "bold")) + xlab("Distance to TSS (bp)
    ylab("RPM (log2)")

boxplot
```



Session Info

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS Linux 7 (Core)
##
## Matrix products: default
## BLAS:   /n/apps/CentOS7/install/r-4.1.0/lib64/R/lib/libRblas.so
## LAPACK: /n/apps/CentOS7/install/r-4.1.0/lib64/R/lib/libRlapack.so
##
```

```
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] digest_0.6.27
##  [2] pander_0.6.3
##  [3] data.table_1.14.0
##  [4] lattice_0.20-44
##  [5] ggpubr_0.4.0
##  [6] gridExtra_2.3
##  [7] ggseqlogo_0.1
##  [8] cowplot_1.1.1
##  [9] ggplot2_3.3.3
## [10] magrittr_2.0.1
## [11] CAGEr_1.34.0
## [12] MultiAssayExperiment_1.18.0
## [13] plyranges_1.12.0
## [14] reshape2_1.4.4
## [15] dplyr_1.0.6
## [16] TxDb.Dmelanogaster.UCSC.dm6.ensGene_3.12.0
## [17] GenomicFeatures_1.44.0
## [18] AnnotationDbi_1.54.0
## [19] BSgenome.Dmelanogaster.UCSC.dm6_1.4.1
## [20] BSgenome_1.60.0
## [21] rtracklayer_1.52.0
## [22] GenomicAlignments_1.28.0
## [23] Rsamtools_2.8.0
## [24] Biostrings_2.60.1
## [25] XVector_0.32.0
## [26] SummarizedExperiment_1.22.0
## [27] Biobase_2.52.0
## [28] MatrixGenerics_1.4.0
## [29] matrixStats_0.59.0
## [30] GenomicRanges_1.44.0
## [31] GenomeInfoDb_1.28.0
## [32] IRanges_2.26.0
## [33] S4Vectors_0.30.0
## [34] BiocGenerics_0.38.0
##
## loaded via a namespace (and not attached):
##   [1] VGAM_1.1-5             colorspace_2.0-1       ggsignif_0.6.1
##   [4] rjson_0.2.20           rio_0.5.26             ellipsis_0.3.2
##   [7] som_0.3-5.1            farver_2.1.0           bit64_4.0.5
##  [10] fansi_0.5.0            splines_4.1.0          cachem_1.0.5
##  [13] knitr_1.33             broom_0.7.6            cluster_2.1.2
```

```
##  [16] dbplyr_2.1.1           png_0.1-7              compiler_4.1.0
##  [19] httr_1.4.2             backports_1.2.1        assertthat_0.2.1
##  [22] Matrix_1.3-4           fastmap_1.1.0          formatR_1.11
##  [25] htmltools_0.5.1.1      prettyunits_1.1.1      tools_4.1.0
##  [28] gtable_0.3.0           glue_1.4.2             GenomeInfoDbData_1.2.6
##  [31] rappdirs_0.3.3         Rcpp_1.0.6             carData_3.0-4
##  [34] cellranger_1.1.0       vctrs_0.3.8            nlme_3.1-152
##  [37] xfun_0.23              stringr_1.4.0          openxlsx_4.2.3
##  [40] lifecycle_1.0.0        restfulr_0.0.13        formula.tools_1.7.1
##  [43] gtools_3.9.2           rstatix_0.7.0          XML_3.99-0.6
##  [46] beanplot_1.2           stringdist_0.9.6.3     zlibbioc_1.38.0
##  [49] MASS_7.3-54            scales_1.1.1           hms_1.1.0
##  [52] yaml_2.2.1             curl_4.3.1             memoise_2.0.0
##  [55] biomaRt_2.48.0         reshape_0.8.8          stringi_1.6.2
##  [58] RSQLite_2.2.7          highr_0.9              BiocIO_1.2.0
##  [61] permute_0.9-5          filelock_1.0.2         zip_2.2.0
##  [64] BiocParallel_1.26.0    operator.tools_1.6.3   rlang_0.4.11
##  [67] pkgconfig_2.0.3        bitops_1.0-7           evaluate_0.14
##  [70] purrr_0.3.4            labeling_0.4.2         bit_4.0.4
##  [73] tidyselect_1.1.1       plyr_1.8.6             R6_2.5.0
##  [76] generics_0.1.0         DelayedArray_0.18.0    DBI_1.1.1
##  [79] haven_2.4.1            foreign_0.8-81         pillar_1.6.1
##  [82] withr_2.4.2            mgcv_1.8-36            abind_1.4-5
##  [85] KEGGREST_1.32.0        RCurl_1.98-1.3         tibble_3.1.2
##  [88] crayon_1.4.1           car_3.0-10             KernSmooth_2.23-20
##  [91] utf8_1.2.1             BiocFileCache_2.0.0    rmarkdown_2.8
##  [94] progress_1.2.2         readxl_1.3.1           grid_4.1.0
##  [97] blob_1.2.1             vegan_2.5-7            forcats_0.5.1
## [100] tidyr_1.1.3            munsell_0.5.0
```