Figure 7 A and S10 Inr variants and paused Pol II half-lives

Wanqing Shao(was@stowers.org)

Contents

Description

To test if initiator (Inr) sequence variants can contribute to paused Pol II stability, we separated Inr G and non-G variants in our core promoter combination analysis.

Enviroment setup

```
library(GenomicRanges, warn.conflicts=F)
library(magrittr)
library(Rmisc)
library(magrittr)
library(Biostrings)
library(BSgenome.Dmelanogaster.UCSC.dm3)

setwd("/data/analysis_code")
options(knitr.figure_dir =
        "Figure7A_S10_inr_variants_and_paused_polii_halflives"
        )

source("shared_code/knitr_common.r")
source("shared_code/ggplot_common.r")
source("shared_code/granges_common.r")
source("shared_code/metapeak_common.r")
source("shared_code/sample_common.r")
```

Analysis

Checking core promoter elements

```
half_life_df <- get(load("rdata/half_life_df.RData"))
tss <- get(load("rdata/dme_mrna_unique_tss.RData"))

find_motif <- function(motif_name, fb_t_id,  mismatch=0) {

    motif_info <- subset(promoter_table, name == motif_name)
    motif <- DNAString(motif_info$motif)
    up_dis <- motif_info$window_start
    down_dis <- motif_info$window_end

    gene_tss <- tss[tss$fb_t_id %in% fb_t_id]

    if(up_dis >= 0 & down_dis >=0){
      tss_r <- resize(gene_tss, down_dis, "start") %>%
               resize(., down_dis - up_dis, "end")
    }
    if(up_dis < 0 & down_dis >=0){
      tss_r <- resize(gene_tss, down_dis, "start") %>%
               resize(., abs(up_dis)+down_dis, "end")
    }
    if(up_dis < 0 & down_dis <0){
      tss_r <- resize(gene_tss, abs(up_dis), "end") %>%
               resize(., abs(up_dis)-abs(down_dis), "start")
    }

    promoter_seq <- getSeq(Dmelanogaster, tss_r)
    names(promoter_seq) <- tss_r$fb_t_id

    count_df <- vcountPattern(motif, promoter_seq, fixed = FALSE,
                              min.mismatch = 0, max.mismatch = mismatch) %>%
                data.frame(fb_t_id = fb_t_id, count =.)

    count_df$count <- ifelse(count_df$count >0, T, F)
    colnames(count_df)[2] <- motif_name
    count_df
}

promoter_table <- read.table("promoter_elements.txt", header=T)
motifs <- c("TATA", "DPE", "MTE", "PB", "Inr")
half_life_tss <- tss[tss$fb_t_id %in% half_life_df$fb_t_id]

motif_list <- lapply(as.character(motifs), function(x){
  motif <- find_motif(motif_name=x, half_life_tss$fb_t_id, mismatch = 1)
  motif
  })

motif_df <- reshape::merge_recurse(motif_list)
all_info_df <- merge(half_life_df, motif_df)
all_info_df$half_life <- ifelse(all_info_df$half_life >=0 & all_info_df$half_life<= 60,
```

```
                                all_info_df$half_life, 60)

new_info_df <-
  with(all_info_df, data.frame(fb_t_id = fb_t_id, gene = gene,
                               half_life = half_life, TATA = TATA,
                               Inr = Inr, pausing_elements = DPE | MTE | PB))
```

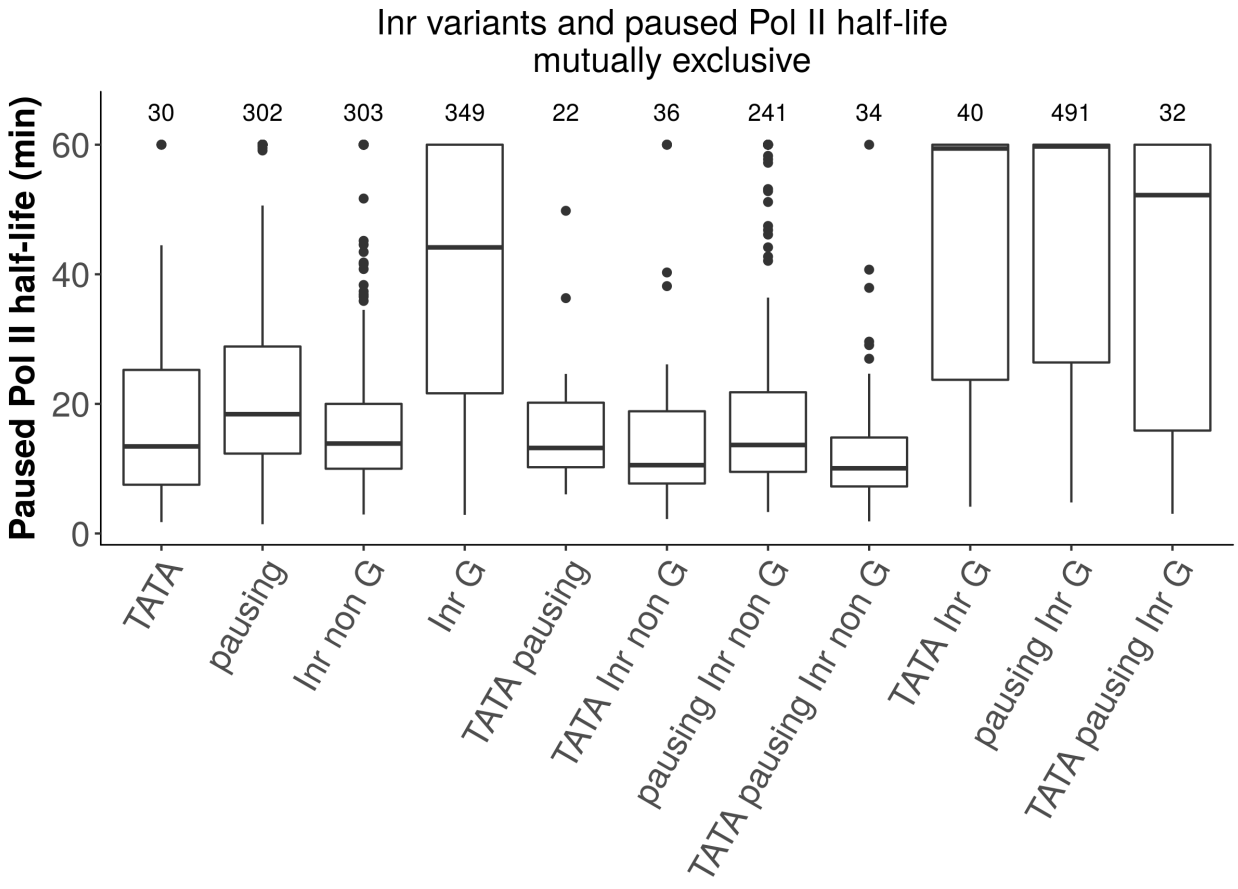
Mutually exclusive model

We put genes into the following group:

-TATA only -TATA + Inr_non_G -TATA + Inr_G -Pausing only -Pausing + Inr_non_G -Pausing + Inr_G -TATA + pasuing -TATA + pausing + Inr_non_G -TATA + pausing + Inr_G -Inr_non_G only -Inr_G only

```
inr_G_region_seq <- resize(half_life_tss, 2, "start") %>%
  resize(., 1, "end") %>%
  getSeq(Dmelanogaster, .) %>%
  as.character()

inr_info_df <- data.frame(fb_t_id = half_life_tss$fb_t_id,
                          G_at_2 = inr_G_region_seq == "G")

new_info_df <- merge(new_info_df, inr_info_df)
new_info_df$Inr_G <- new_info_df$Inr &  new_info_df$G_at_2
new_info_df$Inr_non_G <- new_info_df$Inr &  !new_info_df$G_at_2

tata <- subset(new_info_df, TATA & !(Inr | pausing_elements)) %>%
  data.frame(type = "TATA")
tata_inr_non_g <- subset(new_info_df, TATA & Inr_non_G  & ! pausing_elements)%>%
  data.frame(type = "TATA Inr non G")
tata_inr_g <- subset(new_info_df, TATA & Inr_G  & ! pausing_elements)%>%
  data.frame(type = "TATA Inr G")

tata_pausing <- subset(new_info_df, TATA & pausing_elements & !Inr)%>%
  data.frame(type = "TATA pausing")
tata_pausing_inr_non_g <- subset(new_info_df, TATA & pausing_elements & Inr_non_G) %>%
  data.frame(type = "TATA pausing Inr non G")
tata_pausing_inr_g <- subset(new_info_df, TATA & pausing_elements & Inr_G)%>%
  data.frame(type = "TATA pausing Inr G")

pausing <- subset(new_info_df,  pausing_elements & !(Inr|TATA))%>%
  data.frame(type = "pausing")
pausing_inr_non_g <- subset(new_info_df,  pausing_elements & Inr_non_G & !TATA)%>%
  data.frame(type = "pausing Inr non G")
pausing_inr_g <- subset(new_info_df,  pausing_elements & Inr_G & !TATA)%>%
  data.frame(type = "pausing Inr G")

inr_non_g <- subset(new_info_df,  Inr_non_G & !(pausing_elements | TATA)) %>%
  data.frame(type = "Inr non G")
inr_g <- subset(new_info_df,  Inr_G & !(pausing_elements | TATA)) %>%
  data.frame(type = "Inr G")
```

```
all_combined <- rbind(tata,pausing, inr_non_g, inr_g, tata_pausing, tata_inr_non_g,
                      pausing_inr_non_g, tata_pausing_inr_non_g,
                      tata_inr_g, pausing_inr_g, tata_pausing_inr_g)

half_life_boxplot <- function(combined_df, title){
  count_info <- table(combined_df$type) %>% as.data.frame()
  ggplot(combined_df, aes(x = type, y = half_life)) +
  geom_boxplot() +
  geom_text(data = count_info, aes(x = count_info$Var1, label = Freq, y = 65),
              position = position_dodge(width = .75),
              show.legend = FALSE )+
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  ylab("Paused Pol II half-life (min)") +
  xlab("") +
  ggtitle(title)
}


half_life_boxplot(all_combined,
        "Inr variants and paused Pol II half-life \n mutually exclusive")
```




```
pairwise_type <- unique(all_combined$type) %>%
  as.character(.) %>%
  combn(., 2, simplify = F)
```

```
calc_pval <- function(type1, type2, df){
  type1_df <- subset(df, type == type1)
  type2_df <- subset(df, type == type2)
  pval = wilcox.test(type1_df$half_life, type2_df$half_life)$ p.value
  data.frame(compare = paste(type1, "vs.", type2),
             pval = pval)
}
pval_df <- lapply(pairwise_type, function(x)calc_pval(x[1], x[2], all_combined)) %>%
  do.call(rbind, .)
pander(pval_df)
```

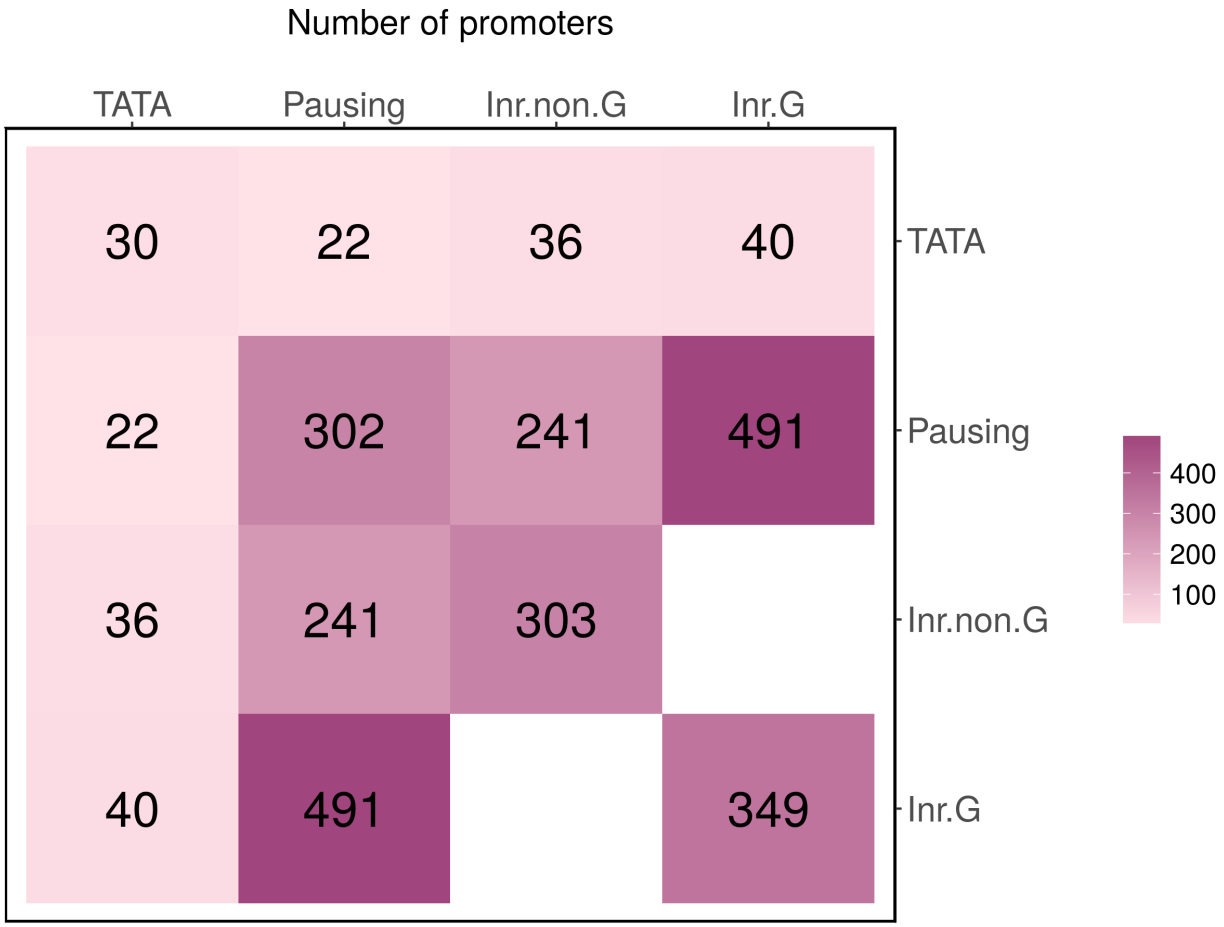
compare	pval
TATA vs. pausing	0.0167
TATA vs. Inr non G	0.6568
TATA vs. Inr G	1.581e-08
TATA vs. TATA pausing	0.9926
TATA vs. TATA Inr non G	0.4866
TATA vs. pausing Inr non G	0.735
TATA vs. TATA pausing Inr non G	0.2259
TATA vs. TATA Inr G	1.941e-05
TATA vs. pausing Inr G	3.207e-11
TATA vs. TATA pausing Inr G	9.914e-05
pausing vs. Inr non G	1.267e-08
pausing vs. Inr G	1.213e-25
pausing vs. TATA pausing	0.01107
pausing vs. TATA Inr non G	0.0001798
pausing vs. pausing Inr non G	1.092e-07
pausing vs. TATA pausing Inr non G	3.377e-06
pausing vs. TATA Inr G	2.146e-06
pausing vs. pausing Inr G	1.823e-43
pausing vs. TATA pausing Inr G	5.3e-05
Inr non G vs. Inr G	8.048e-49
Inr non G vs. TATA pausing	0.6852
Inr non G vs. TATA Inr non G	0.09065
Inr non G vs. pausing Inr non G	0.8086
Inr non G vs. TATA pausing Inr non G	0.003634
Inr non G vs. TATA Inr G	1.56e-10
Inr non G vs. pausing Inr G	5.416e-72
Inr non G vs. TATA pausing Inr G	8.844e-09
Inr G vs. TATA pausing	5.612e-09
Inr G vs. TATA Inr non G	6.905e-12
Inr G vs. pausing Inr non G	2.344e-43
Inr G vs. TATA pausing Inr non G	7e-14
Inr G vs. TATA Inr G	0.4826
Inr G vs. pausing Inr G	0.001602
Inr G vs. TATA pausing Inr G	0.9219
TATA pausing vs. TATA Inr non G	0.3568
TATA pausing vs. pausing Inr non G	0.8046
TATA pausing vs. TATA pausing Inr non G	0.1134
TATA pausing vs. TATA Inr G	6.893e-06
TATA pausing vs. pausing Inr G	1.445e-11
TATA pausing vs. TATA pausing Inr G	1.658e-05
TATA Inr non G vs. pausing Inr non G	0.1553

compare	pval
TATA Inr non G vs. TATA pausing Inr non G	0.3975
TATA Inr non G vs. TATA Inr G	7.297e-07
TATA Inr non G vs. pausing Inr G	2.765e-15
TATA Inr non G vs. TATA pausing Inr G	3.125e-06
pausing Inr non G vs. TATA pausing Inr non G	0.01548
pausing Inr non G vs. TATA Inr G	1.179e-10
pausing Inr non G vs. pausing Inr G	2.629e-63
pausing Inr non G vs. TATA pausing Inr G	1.176e-08
TATA pausing Inr non G vs. TATA Inr G	5.768e-08
TATA pausing Inr non G vs. pausing Inr G	2.904e-17
TATA pausing Inr non G vs. TATA pausing Inr G	1.552e-07
TATA Inr G vs. pausing Inr G	0.5998
TATA Inr G vs. TATA pausing Inr G	0.697
pausing Inr G vs. TATA pausing Inr G	0.2886

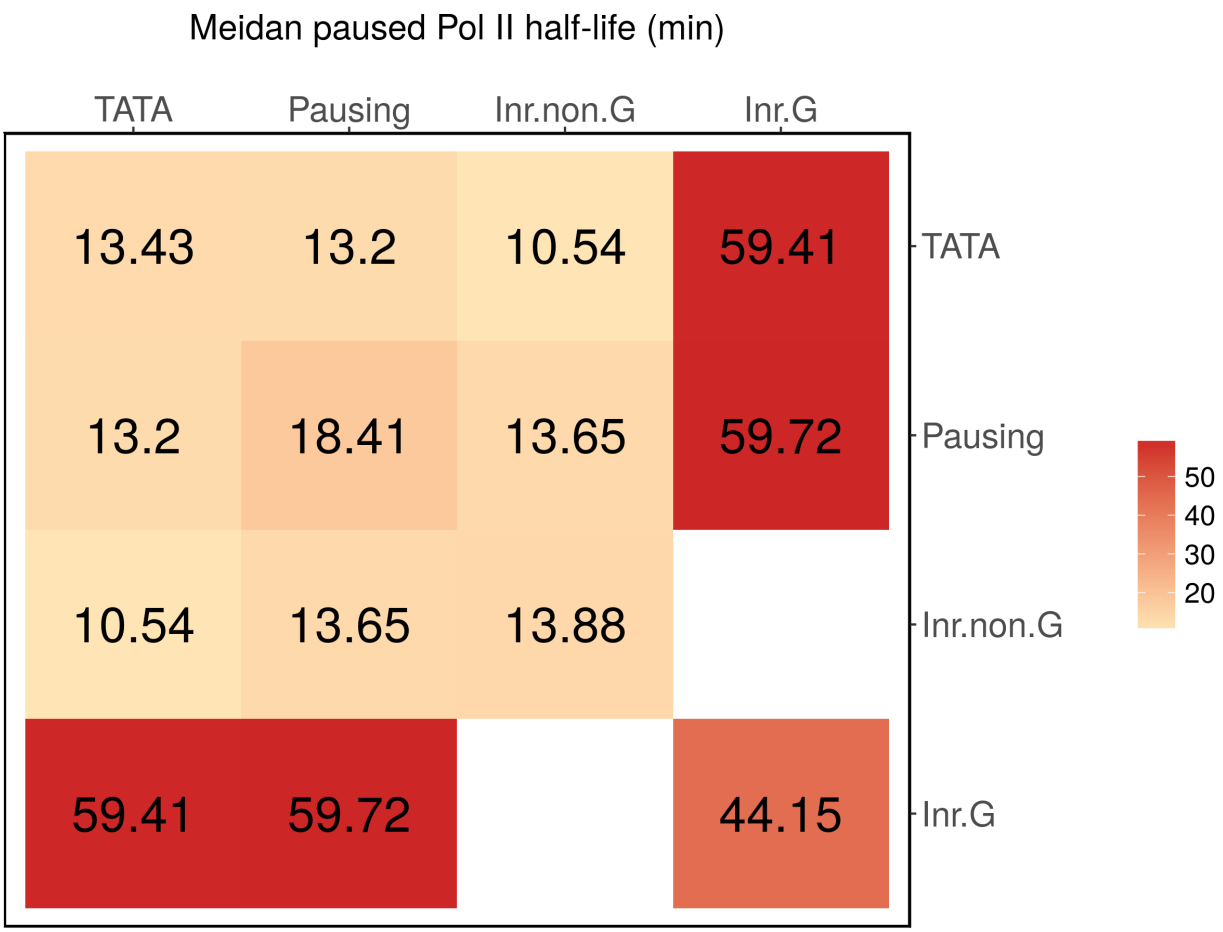
```
write.csv(pval_df, file = "Inr_variant_halflife_wilcox_test_mutually_exclusive.csv")
```

```
info_list <- list(TATA_TATA = tata,
                  Inr.non.G_Inr.non.G = inr_non_g,
                  Inr.G_Inr.G = inr_g,
                  Pausing_Pausing = pausing,
                  TATA_Inr.non.G = tata_inr_non_g,
                  TATA_Inr.G = tata_inr_g,
                  Pausing_Inr.non.G = pausing_inr_non_g,
                  Pausing_Inr.G = pausing_inr_g,
                  TATA_Pausing = tata_pausing)

generate_pairwise_df <- function(info_list){
  pairwise_df <- lapply(names(info_list), function(x){
  info_df <- data.frame(motif1 = gsub("_.*", "", x),
                        motif2 = gsub(".*_", "", x),
                        half_life = median(info_list[[x]]$half_life),
                        count = nrow(info_list[[x]]))
  info_df
  }) %>% do.call(rbind, .)
  pairwise_df_flip <- pairwise_df
  pairwise_df_flip$motif1 <- pairwise_df$motif2
  pairwise_df_flip$motif2 <- pairwise_df$motif1

  pairwise_df <- rbind(pairwise_df,pairwise_df_flip ) %>% unique()
  pairwise_df
}


promoter_number_heatmap <- function(pairwise_df){
  ggplot(pairwise_df, aes(x = motif1, y = motif2, fill = count)) +
  scale_fill_gradient(low="#ffe1e8", high= "#a0457e") +
  geom_tile() + xlab("") + ylab("") +
  geom_text(aes(label=count), color="black", size=8) +
  scale_x_discrete(position = "top") +
```

```
  scale_y_discrete(position = "right") +
  ggtitle("Number of promoters") +
  theme(panel.border = element_rect( colour = "black", fill = NA, size=1))
}

half_life_heatmap <- function(pairwise_df){
  ggplot(pairwise_df, aes(x = motif1, y = motif2, fill = half_life)) +
  scale_fill_gradient(low="moccasin", high= "firebrick3") +
  geom_tile() + xlab("") + ylab("") +
  geom_text(aes(label=round(half_life, digits = 2)), color="black", size=8) +
  scale_x_discrete(position = "top") +
  scale_y_discrete(position = "right") +
  ggtitle("Meidan paused Pol II half-life (min)") +
  theme(panel.border = element_rect( colour = "black", fill = NA, size=1))
}

pairwise_df <- generate_pairwise_df(info_list)
pairwise_df$motif1 <-
  factor(pairwise_df$motif1, levels = c("TATA", "Pausing", "Inr.non.G", "Inr.G"))
pairwise_df$motif2 <-
  factor(pairwise_df$motif2, levels = rev(c("TATA", "Pausing", "Inr.non.G", "Inr.G")))

promoter_number_heatmap(pairwise_df)
```



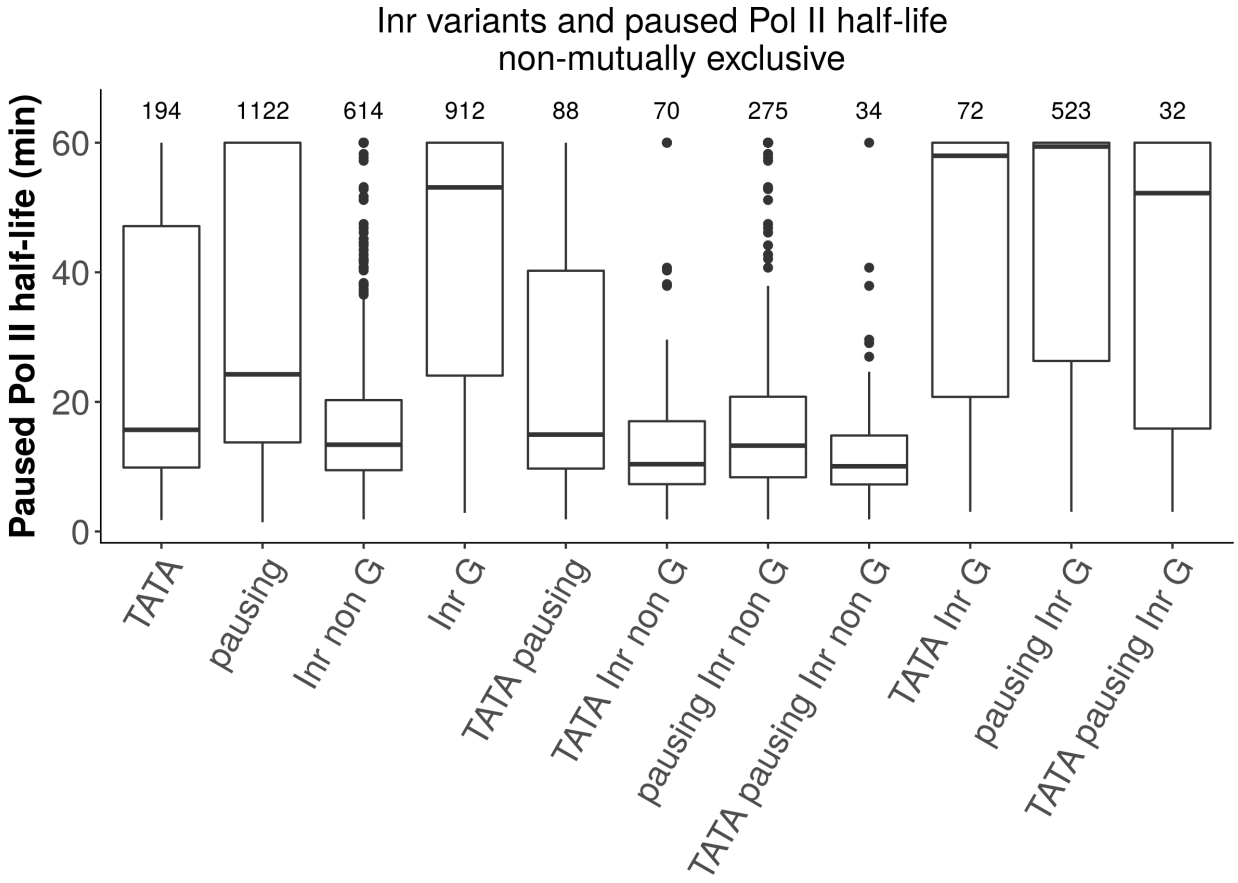
```
half_life_heatmap(pairwise_df)
```



Non-Mutually exclusive model

```
tata <- subset(new_info_df, TATA) %>%
  data.frame(type = "TATA")
tata_inr_non_g <- subset(new_info_df, TATA & Inr_non_G  )%>%
  data.frame(type = "TATA Inr non G")
tata_inr_g <- subset(new_info_df, TATA & Inr_G)%>%
  data.frame(type = "TATA Inr G")

tata_pausing <- subset(new_info_df, TATA & pausing_elements )%>%
  data.frame(type = "TATA pausing")
tata_pausing_inr_non_g <- subset(new_info_df, TATA & pausing_elements & Inr_non_G) %>%
  data.frame(type = "TATA pausing Inr non G")
tata_pausing_inr_g <- subset(new_info_df, TATA & pausing_elements & Inr_G)%>%
  data.frame(type = "TATA pausing Inr G")

pausing <- subset(new_info_df,  pausing_elements)%>%
  data.frame(type = "pausing")
pausing_inr_non_g <- subset(new_info_df,  pausing_elements & Inr_non_G )%>%
  data.frame(type = "pausing Inr non G")
```

```
pausing_inr_g <- subset(new_info_df,  pausing_elements & Inr_G)%>%
  data.frame(type = "pausing Inr G")

inr_non_g <- subset(new_info_df,  Inr_non_G) %>%
  data.frame(type = "Inr non G")
inr_g <- subset(new_info_df,  Inr_G ) %>%
  data.frame(type = "Inr G")

all_combined <- rbind(tata,pausing, inr_non_g, inr_g, tata_pausing, tata_inr_non_g,
                      pausing_inr_non_g, tata_pausing_inr_non_g,
                      tata_inr_g, pausing_inr_g, tata_pausing_inr_g)

half_life_boxplot(all_combined,
        "Inr variants and paused Pol II half-life \n non-mutually exclusive")
```



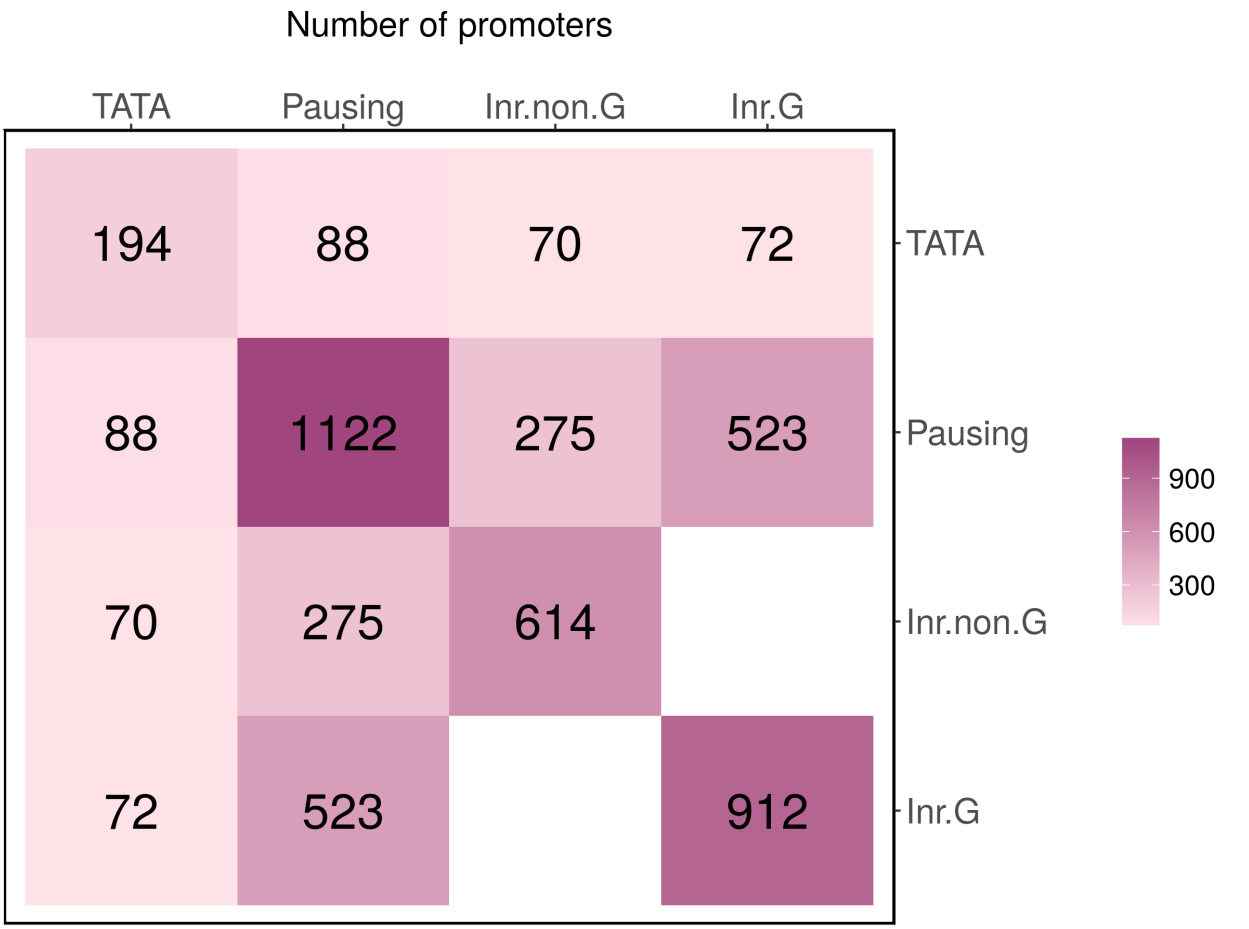

```
pairwise_type <- unique(all_combined$type) %>%
  as.character(.) %>%
  combn(., 2, simplify = F)

pval_df <- lapply(pairwise_type, function(x)calc_pval(x[1], x[2], all_combined)) %>%
  do.call(rbind, .)
pander(pval_df)
```

compare	pval
TATA vs. pausing	4.347e-06
TATA vs. Inr non G	0.0008458
TATA vs. Inr G	3.747e-26
TATA vs. TATA pausing	0.656
TATA vs. TATA Inr non G	0.0001794
TATA vs. pausing Inr non G	0.001123
TATA vs. TATA pausing Inr non G	0.001044
TATA vs. TATA Inr G	1.168e-07
TATA vs. pausing Inr G	1.137e-26
TATA vs. TATA pausing Inr G	0.0001948
pausing vs. Inr non G	1.644e-47
pausing vs. Inr G	7.419e-35
pausing vs. TATA pausing	9.074e-05
pausing vs. TATA Inr non G	1.083e-13
pausing vs. pausing Inr non G	1.625e-28
pausing vs. TATA pausing Inr non G	9.547e-09
pausing vs. TATA Inr G	0.0001728
pausing vs. pausing Inr G	1.221e-31
pausing vs. TATA pausing Inr G	0.01898
Inr non G vs. Inr G	1.818e-125
Inr non G vs. TATA pausing	0.05769
Inr non G vs. TATA Inr non G	0.007377
Inr non G vs. pausing Inr non G	0.6834
Inr non G vs. TATA pausing Inr non G	0.01133
Inr non G vs. TATA Inr G	1.542e-18
Inr non G vs. pausing Inr G	4.4e-105
Inr non G vs. TATA pausing Inr G	1.996e-09
Inr G vs. TATA pausing	7.031e-17
Inr G vs. TATA Inr non G	1.005e-27
Inr G vs. pausing Inr non G	4.591e-76
Inr G vs. TATA pausing Inr non G	2.507e-16
Inr G vs. TATA Inr G	0.7273
Inr G vs. pausing Inr G	0.1279
Inr G vs. TATA pausing Inr G	0.5788
TATA pausing vs. TATA Inr non G	0.00247
TATA pausing vs. pausing Inr non G	0.04015
TATA pausing vs. TATA pausing Inr non G	0.004143
TATA pausing vs. TATA Inr G	2.669e-07
TATA pausing vs. pausing Inr G	4.596e-18
TATA pausing vs. TATA pausing Inr G	0.0001036
TATA Inr non G vs. pausing Inr non G	0.03007
TATA Inr non G vs. TATA pausing Inr non G	0.6178
TATA Inr non G vs. TATA Inr G	5.848e-13
TATA Inr non G vs. pausing Inr G	1.906e-28
TATA Inr non G vs. TATA pausing Inr G	1.461e-08
pausing Inr non G vs. TATA pausing Inr non G	0.03247
pausing Inr non G vs. TATA Inr G	4.424e-17
pausing Inr non G vs. pausing Inr G	1.863e-70
pausing Inr non G vs. TATA pausing Inr G	3.817e-09
TATA pausing Inr non G vs. TATA Inr G	5.145e-10
TATA pausing Inr non G vs. pausing Inr G	3.874e-17

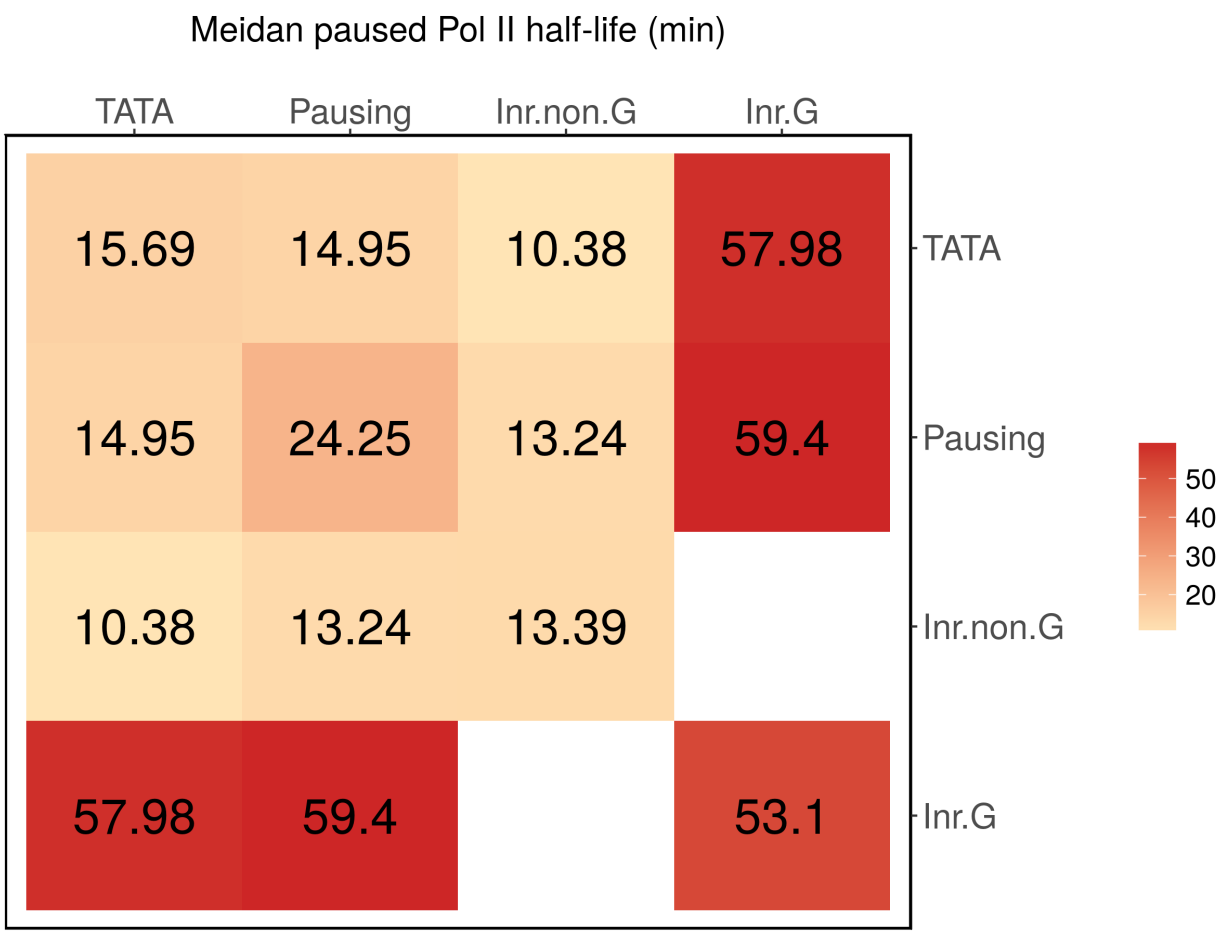
compare	pval
TATA pausing Inr non G vs. TATA pausing Inr G	1.552e-07
TATA Inr G vs. pausing Inr G	0.3302
TATA Inr G vs. TATA pausing Inr G	0.8095
pausing Inr G vs. TATA pausing Inr G	0.3189

```
write.csv(pval_df, file = "Inr_variant_halflife_wilcox_test_non_mutually_exclusive.csv")
```

```
info_list <- list(TATA_TATA = tata,
                  Inr.non.G_Inr.non.G = inr_non_g,
                  Inr.G_Inr.G = inr_g,
                  Pausing_Pausing = pausing,
                  TATA_Inr.non.G = tata_inr_non_g,
                  TATA_Inr.G = tata_inr_g,
                  Pausing_Inr.non.G = pausing_inr_non_g,
                  Pausing_Inr.G = pausing_inr_g,
                  TATA_Pausing = tata_pausing)

pairwise_df <- generate_pairwise_df(info_list)
pairwise_df$motif1 <-
  factor(pairwise_df$motif1, levels = c("TATA", "Pausing", "Inr.non.G", "Inr.G"))
pairwise_df$motif2 <-
  factor(pairwise_df$motif2, levels = rev(c("TATA", "Pausing", "Inr.non.G", "Inr.G")))

promoter_number_heatmap(pairwise_df)
```




```
half_life_heatmap(pairwise_df)
```



Session Info

This analysis was performed with the following R/Bioconductor session:

```
sessionInfo()
```

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.4 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## 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] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] reshape2_1.4.3                      ggplot2_2.2.1
##  [3] pander_0.6.1                        BSgenome.Dmelanogaster.UCSC.dm3_1.4.0
##  [5] BSgenome_1.46.0                     rtracklayer_1.38.3
##  [7] Biostrings_2.46.0                   XVector_0.18.0
##  [9] Rmisc_1.5                           plyr_1.8.4
## [11] lattice_0.20-35                     magrittr_1.5
## [13] GenomicRanges_1.30.3                GenomeInfoDb_1.14.0
## [15] IRanges_2.12.0                      S4Vectors_0.16.0
## [17] BiocGenerics_0.24.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.17               pillar_1.2.3
##  [3] compiler_3.4.4             bitops_1.0-6
##  [5] tools_3.4.4                zlibbioc_1.24.0
##  [7] digest_0.6.15              tibble_1.4.2
##  [9] gtable_0.2.0               evaluate_0.10.1
## [11] rlang_0.2.1                Matrix_1.2-14
## [13] DelayedArray_0.4.1         yaml_2.1.19
## [15] GenomeInfoDbData_1.0.0     stringr_1.3.1
## [17] knitr_1.20                 rprojroot_1.3-2
## [19] grid_3.4.4                 reshape_0.8.7
## [21] Biobase_2.38.0             XML_3.98-1.11
## [23] BiocParallel_1.12.0        rmarkdown_1.10
## [25] scales_0.5.0               backports_1.1.2
## [27] Rsamtools_1.30.0           htmltools_0.3.6
## [29] matrixStats_0.53.1         GenomicAlignments_1.14.2
## [31] SummarizedExperiment_1.8.1 colorspace_1.3-2
## [33] labeling_0.3               stringi_1.2.3
## [35] lazyeval_0.2.1             munsell_0.5.0
## [37] RCurl_1.95-4.10
```