Step 4: Noncoding RNA analysis

Carlos Gallardo & Christian Oertlin

17 April, 2023

```
# Import libraries and helper functions
source("code/helper_functions.R")
library(tidyverse)
library(magrittr)
library(ggrepel)
library(RColorBrewer)
library(pheatmap)
library(viridis)
library(patchwork)
library(ComplexHeatmap)

# Colors
colPals <- vector(mode = "list")
colPals$time <- setNames(c("#FBAA3E", "#2C83BE", "#3EB6BD", "#A3D5B3", "#CD71A8"),
                         nm = c("day0", "day7", "day14", "day21", "day28"))
colPals$time_light <- setNames(c("#FDD6A1", "#A2CDE9", "#AFE2E5", "#DDF0E3", "#E8BDD6"),
                               nm = c("day0", "day7", "day14", "day21", "day28"))
colPals$time_dark <- setNames(c("#D87E04", "#174564", "#1F5C60", "#49A065", "#AA3C7E"),
                              nm = c("day0", "day7", "day14", "day21", "day28"))
colPals$inferno <- c("#000004", "#420A68", "#932667", "#DD513A", "#FCA50A", "#FCFFA4")
colPals$blood_cells <- setNames(c("#E54D34","#77A2D5","#B58B80"),
                                nm = c("granulocytes", "lymphocytes", "monocytes"))
colPals$cell_types <- setNames(c("#83D1F6","#FBAA3E","#FCCA7C","#B58B80","#E54D34",
                                 "#B3177E","#9A509F","#77A2D5","#CAC1DD","#36B449","#C1C1C1"),
                               nm = c("B cell", "Macrophage M1", "Macrophage M2",
                                      "Monocyte", "Neutrophil", "NK cell",
                                      "T cell CD4+ (non-regulatory)", "T cell CD8+",
                                      "T cell regulatory (Tregs)", "Myeloid dendritic cell",
                                      "uncharacterized cell"))
colPals$RdBu <- brewer.pal(11, name = "RdBu")
colPals$biotype <- setNames(c("#395982","#49BED9","#18A38A","#36B449","#826F99",
                              "#9852A5","#FBAA3E","#FCCA7C","#FCFFA4","#C1C1C1"),
                            nm = c("protein_coding", "lncRNA", "miRNA", "snoRNA",
                                   "IG_C_gene", "IG_V_gene", "TR_C_gene",
                                   "TR_J_gene", "TR_V_gene", "other"))
colPals$rtqpcr_rnaseq <- setNames(c("#B80D48","#2B6A6C"),
                                  nm = c("rtqpcr", "rnaseq"))
```

Load data

```
# Gene expression and DEGs
RNAseq <- readRDS(file='data/rnaseq/rnaseq_volunteers_9&10_excl.rds')
DESeq2_DEGs <- readRDS(file='data/rnaseq/DESeq2_DEGs_unfilt_volunteers_9&10_excl.rds')
DESeq2_DEGs_filt <- readRDS(file='data/rnaseq/DESeq2_DEGs_filt_volunteers_9&10_excl.rds')

# Normalized k-mer counts
kmer_counts <- vector("list")

kmer_counts[["canonical_background_lncRNA_3mers"]] <- read.table(
  file = 'data/ncrna_analysis/kmer_counts/canonical_background_lncRNA_3mers.txt',
  sep = "\t",
  header = TRUE)

kmer_counts[["canonical_background_lncRNA_4mers"]] <- read.table(
  file = 'data/ncrna_analysis/kmer_counts/canonical_background_lncRNA_4mers.txt',
  sep = "\t",
  header = TRUE)

kmer_counts[["canonical_background_lncRNA_5mers"]] <- read.table(
  file = 'data/ncrna_analysis/kmer_counts/canonical_background_lncRNA_5mers.txt',
  sep = "\t",
```

```
  header = TRUE)

kmer_counts[["canonical_background_lncRNA_6mers"]] <- read.table(
  file = 'data/ncrna_analysis/kmer_counts/canonical_background_lncRNA_6mers.txt',
  sep = "\t",
  header = TRUE)
```

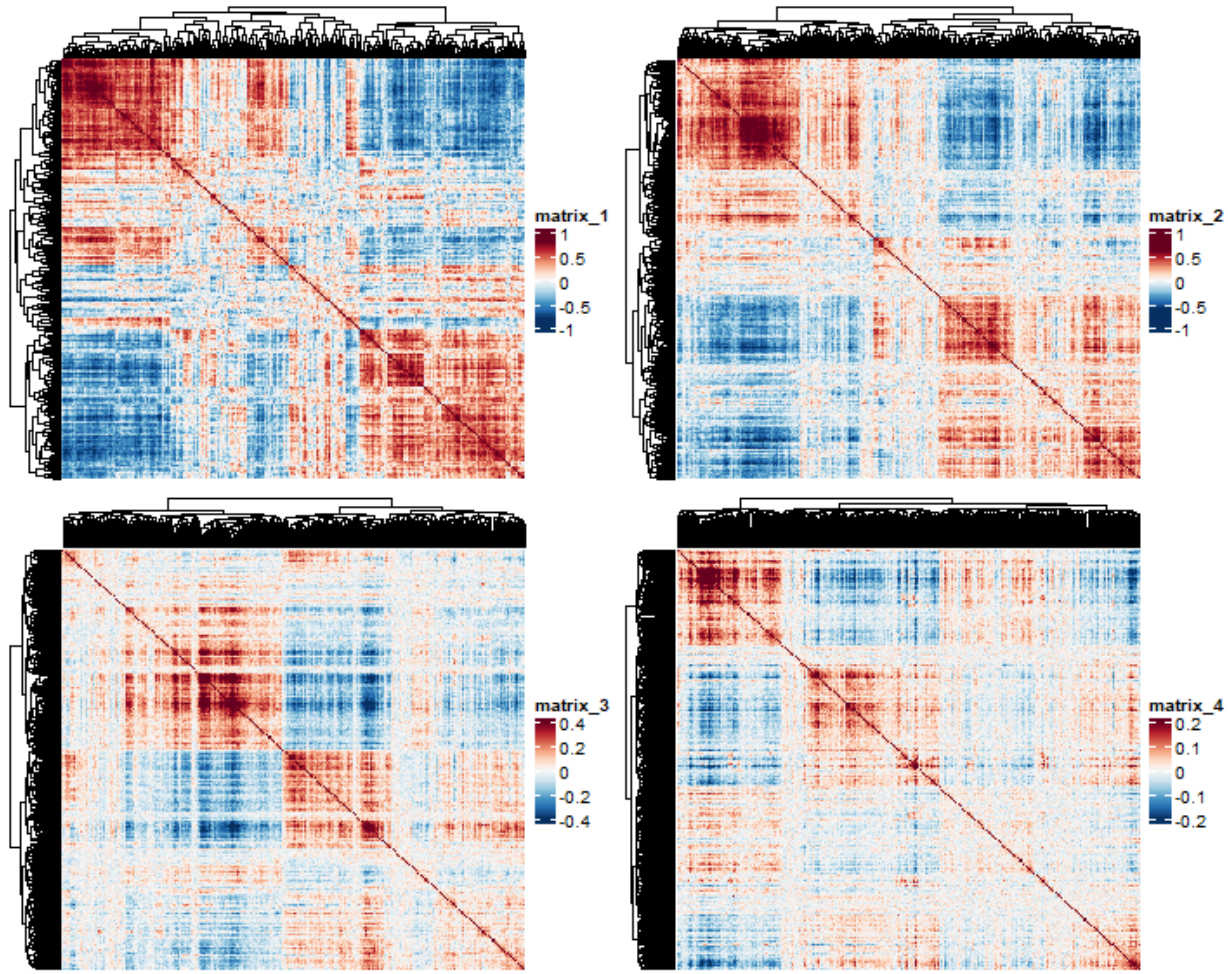
Filter DE lncRNA from pairwise wald tests

```
# Pairwise comparisons
lncRNA <- RNAseq$unfilt$annotation %>%
  filter(gene_biotype == "lncRNA") %>%
  pull(ensembl_gene_id_version) %>%
  intersect(., lapply(DESeq2_DEGs_filt[2:5], function(x) x$ensembl_gene_id_version) %>% unlist() %>% unique())

kmer_counts_pairwise_DE <- lapply(kmer_counts, function(x) filter(x, Geneid %in% lncRNA))
kmer_counts_pairwise_DE <- lapply(kmer_counts_pairwise_DE, function(x) column_to_rownames(x, var = 'Geneid'))
```

```
# Pairwise comparisons
lncRNA <- RNAseq$unfilt$annotation %>%
  filter(gene_biotype == "lncRNA") %>%
  pull(ensembl_gene_id_version) %>%
  intersect(., lapply(DESeq2_DEGs_filt[2:5], function(x) x$ensembl_gene_id_version) %>% unlist() %>% unique())

kmer_counts_pairwise_DE <- lapply(kmer_counts, function(x) filter(x, Geneid %in% lncRNA))
kmer_counts_pairwise_DE <- lapply(kmer_counts_pairwise_DE, function(x) column_to_rownames(x, var = 'Geneid'))

kmer_counts_pairwise_DE_cor <- lapply(kmer_counts_pairwise_DE, function(x) cor(t(x), method = 'pearson'))
```

Group lncRNA based on k-mer enrichment

```
p1 <- Heatmap(kmer_counts_pairwise_DE_cor$canonical_background_lncRNA_3mers,
        show_row_names = F,
        show_column_names = F,
        clustering_distance_rows = function(x) as.dist(1-x),
        clustering_distance_columns = function(x) as.dist(1-x),
        col = circlize::colorRamp2(breaks=seq(-0.8, 0.8, length.out=21),
                                   colors=colorRampPalette(rev(RColorBrewer::brewer.pal(n=11, name = "RdBu")))(21)))

p2 <- Heatmap(kmer_counts_pairwise_DE_cor$canonical_background_lncRNA_4mers,
        show_row_names = F,
        show_column_names = F,
        clustering_distance_rows = function(x) as.dist(1-x),
        clustering_distance_columns = function(x) as.dist(1-x),
        col = circlize::colorRamp2(breaks=seq(-0.6, 0.6, length.out=21),
                                   colors=colorRampPalette(rev(RColorBrewer::brewer.pal(n=11, name = "RdBu")))(21)))

p3 <- Heatmap(kmer_counts_pairwise_DE_cor$canonical_background_lncRNA_5mers,
        show_row_names = F,
        show_column_names = F,
        clustering_distance_rows = function(x) as.dist(1-x),
        clustering_distance_columns = function(x) as.dist(1-x),
        col = circlize::colorRamp2(breaks=seq(-0.4, 0.4, length.out=21),
                                   colors=colorRampPalette(rev(RColorBrewer::brewer.pal(n=11, name = "RdBu")))(21)))

p4 <- Heatmap(kmer_counts_pairwise_DE_cor$canonical_background_lncRNA_6mers,
        show_row_names = F,
        show_column_names = F,
        clustering_distance_rows = function(x) as.dist(1-x),
        clustering_distance_columns = function(x) as.dist(1-x),
        col = circlize::colorRamp2(breaks=seq(-0.2, 0.2, length.out=21),
                                   colors=colorRampPalette(rev(RColorBrewer::brewer.pal(n=11, name = "RdBu")))(21)))

(ggplotify::as.ggplot(p1) | ggplotify::as.ggplot(p2)) / (ggplotify::as.ggplot(p3) | ggplotify::as.ggplot(p4))
```




```
pdf("plots/figS7_lncRNA_cor_mat_3mers.pdf", width = 6, height = 5)
draw(p1)
dev.off()
```

```
## png
##   2
```

```
pdf("plots/figS7_lncRNA_cor_mat_4mers.pdf", width = 6, height = 5)
draw(p2)
dev.off()
```

```
## png
##   2
```

```
pdf("plots/figS7_lncRNA_cor_mat_5mers.pdf", width = 6, height = 5)
draw(p3)
dev.off()
```

```
## png
##   2
```

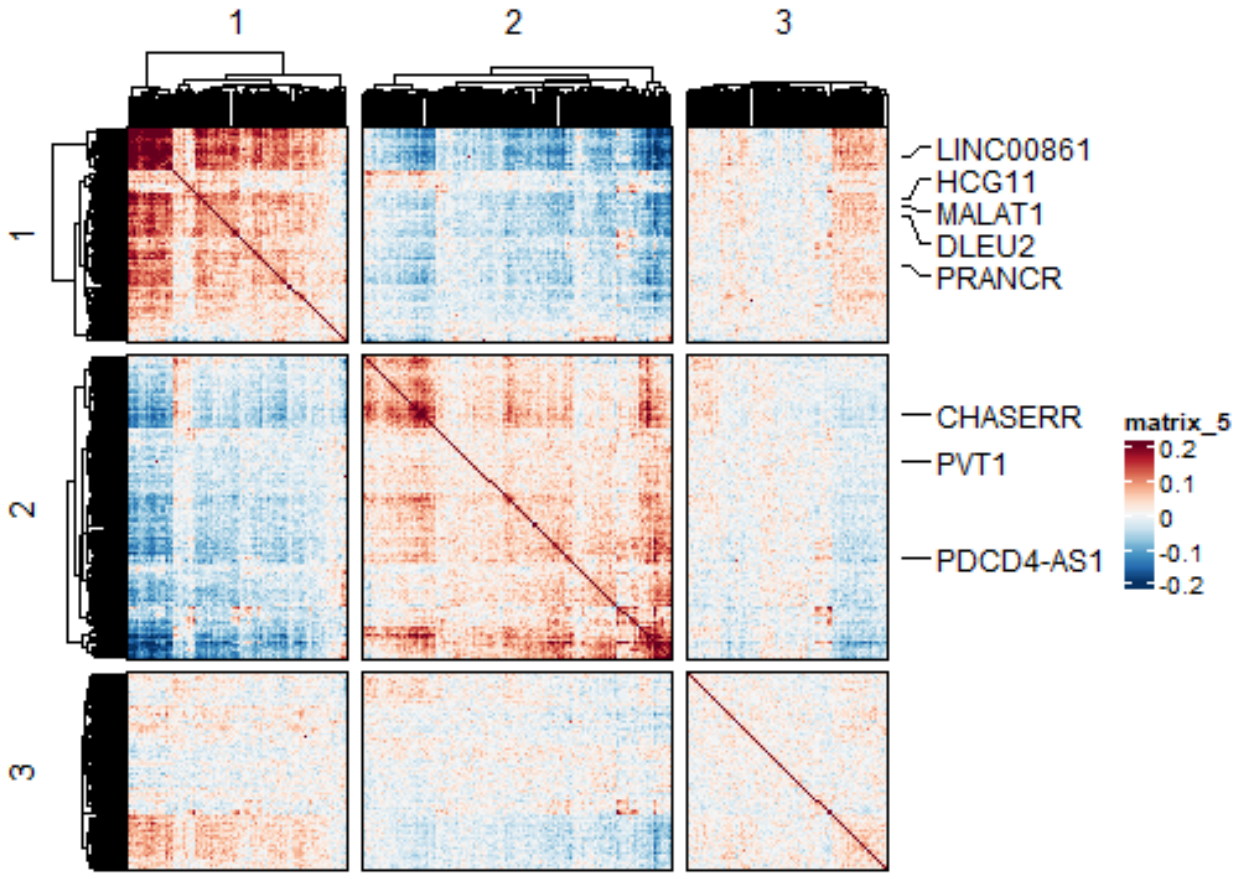
```
pdf("plots/figS7_lncRNA_cor_mat_6mers.pdf", width = 6, height = 5)
draw(p4)
dev.off()
```

```
## png
##   2
```

Clustering on 6-mer enrichment profiles

```
mark_genes <- c('PVT1', 'MALAT1', 'PRANCR', 'DLEU2', 'HCG11', 'CHASERR', 'PDCD4-AS1', 'LINC00861')

set.seed(2)
```

```
lncRNA_cor_mtx <- cor(t(kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers), method = 'pearson')
colnames(lncRNA_cor_mtx) <- plyr::mapvalues(x = colnames(lncRNA_cor_mtx),
                                            from = DESeq2_DEGs$day7Vsday0$ensembl_gene_id_version,
                                            to = DESeq2_DEGs$day7Vsday0$GeneSymbol,
                                            warn_missing = F)
rownames(lncRNA_cor_mtx) <- colnames(lncRNA_cor_mtx)
lncRNA_cl <-kmeans(kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers, centers=3)
p1 <- Heatmap(lncRNA_cor_mtx,
              cluster_rows = T,
              cluster_row_slices = F,
              cluster_columns = T,
              cluster_column_slices = F,
              split = lncRNA_cl$cluster,
              column_split = lncRNA_cl$cluster,
              show_row_names = F,
              show_column_names = F,
              show_row_dend = T,
              show_column_dend = T,
              border = T,
              gap = unit(2, "mm"),
              column_gap = unit(2, "mm"),
              row_dend_gp = gpar(lwd=unit(0.1, "mm")),
              column_dend_gp = gpar(lwd=unit(0.1, "mm")),
              col = circlize::colorRamp2(breaks=seq(-0.2, 0.2, length.out=21),
                                         colors=colorRampPalette(rev(RColorBrewer::brewer.pal(n=11, name = "RdBu")))(21))
) +
  rowAnnotation(mark = anno_mark(at=which(rownames(lncRNA_cor_mtx) %in% mark_genes),
                                 labels = rownames(lncRNA_cor_mtx)[which(rownames(lncRNA_cor_mtx) %in% mark_genes)]))

p1
```



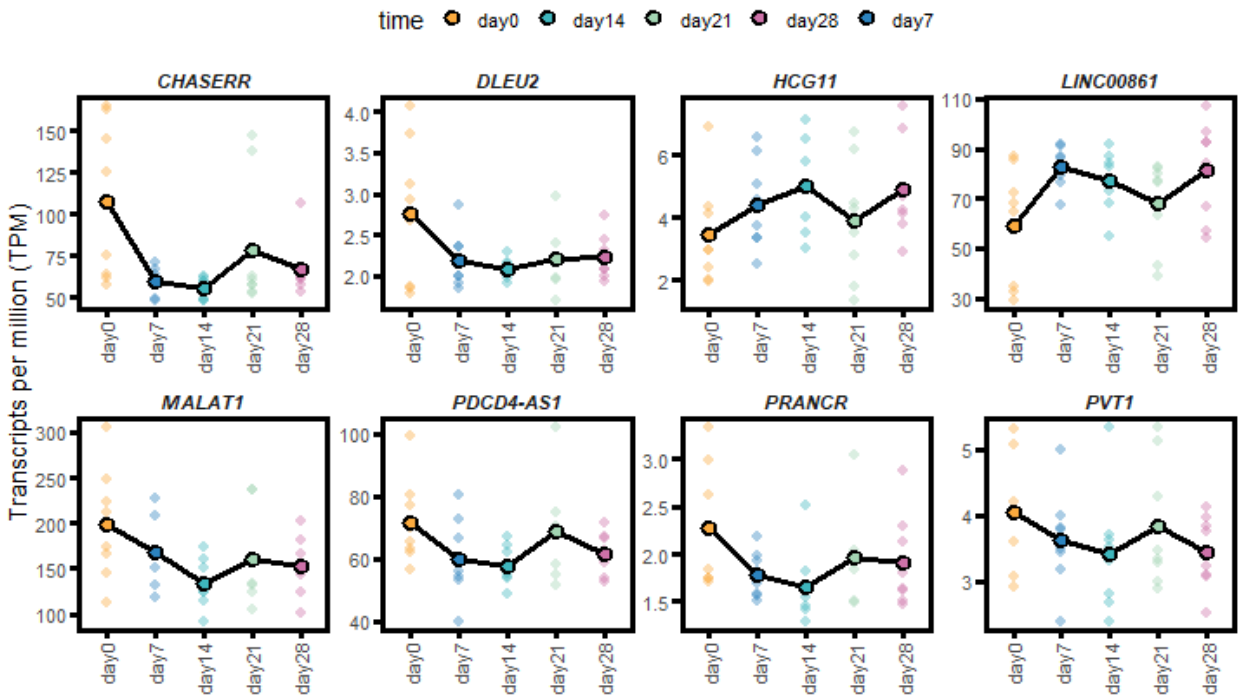

```
pdf("plots/fig4A_lncRNA_cor_mat_6mers_clusters.pdf", width = 7, height = 5)
draw(p1)
dev.off()
```

```
## png
##   2
```

Expression plots for selected lncRNAs

```
RNAseq_expr_tpm <- RNAseq$filt$rawdata
rownames(RNAseq_expr_tpm) <- NULL
RNAseq_expr_tpm <- RNAseq_expr_tpm %>%
  column_to_rownames(var = 'Geneid_version') %>%
  select(-c('Geneid', 'Length', 'GeneSymbol')) %>%
  tpm.normalize(., RNAseq$filt$rawdata$Length) %>%
  rownames_to_column(var = 'Geneid_version')
colnames(RNAseq_expr_tpm) <- c('Geneid_version', paste(RNAseq$filt$design$volunteer,
                                                       RNAseq$filt$design$time,
                                                       sep = '_'))

lncRNA_expr <- as.data.frame(lncRNA_cl$cluster) %>%
  dplyr::rename(Cluster = 'lncRNA_cl$cluster') %>%
  rownames_to_column(var = 'Geneid_version') %>%
  left_join(RNAseq_expr_tpm, by = 'Geneid_version') %>%
  mutate(Geneid_version = plyr::mapvalues(x = Geneid_version,
                                          from = DESeq2_DEGs$day7Vsday0$ensembl_gene_id_version,
                                          to = DESeq2_DEGs$day7Vsday0$GeneSymbol,
                                          warn_missing = F)) %>%
  dplyr::rename(Gene = Geneid_version)

lncRNA_expr_mean <- lncRNA_expr %>%
  column_to_rownames(var = 'Gene') %>%
  select(-Cluster) %>%
  group.transform(group = RNAseq$filt$design$time,
                  FUN = function(x) apply(x, 1, mean)) %>%
  rownames_to_column(var = 'Gene') %>%
  add_column(Cluster = lncRNA_expr$Cluster, .after = 'Gene')

df <- lapply(setNames(mark_genes, mark_genes), function(x) {

  lncRNA_expr %>%
    filter(Gene == x) %>%
    column_to_rownames(var = 'Gene') %>%
    select(-Cluster) %>%
    t() %>%
    as.data.frame() %>%
    dplyr::rename(TPM = x) %>%
    mutate(time = RNAseq$filt$design$time,
           gene = x)

}) %>% bind_rows()

df2 <- lapply(setNames(mark_genes, mark_genes), function(x) {

  lncRNA_expr_mean %>%
    filter(Gene == x) %>%
    column_to_rownames(var = 'Gene') %>%
    select(-Cluster) %>%
    t() %>%
    as.data.frame() %>%
    dplyr::rename(TPM = x) %>%
    rownames_to_column(var = 'time') %>%
    mutate(gene = x)

}) %>% bind_rows()

ggplot(df, aes(x=time, y=TPM, color=time)) +
  geom_point(shape=16, size=3, stroke=0, alpha=0.4) +
  facet_wrap(~gene, ncol=4, scales = 'free') +
  geom_line(data = df2, aes(x=time, y=TPM, group=1), color='black', size = 1.25) +
  geom_point(data = df2, aes(x=time, y=TPM, group=1, fill=time),
             shape=21, size=3.5, stroke=1.25, color='black') +
  xlab('') +
  ylab('Transcripts per million (TPM)') +
  scale_color_manual(values = colPals$time) +
  scale_fill_manual(values = colPals$time) +
  guides(color = F) +
  theme_bw() +
  theme(
    text = element_text(family = 'Arial', size = 14),
    axis.text.x.bottom = element_text(angle = 90, hjust = 1, vjust = 0.5),
    panel.border = element_rect(color = "black", fill = NA, size = 2),
    axis.ticks = element_line(color = "black", size = 1.25),
    axis.ticks.length = unit(1.5, 'mm'),
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank(),
```

```
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank(),
    legend.position = 'top',
    strip.background = element_blank(),
    strip.text = element_text(face = "bold.italic"),
    strip.text.y = element_text(angle = 90)
  )
```




```
ggsave("plots/fig4B_lncRNA_selected_expression_tpm.pdf", width = 10, height = 6, units = "in", dpi = 300, device = cairo_pdf)
```

Motif enrichemnt in lncRNA clusters

```
lncRNA <- rownames(kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers)
lncRNA_cl1 <- names(lncRNA_cl$cluster[lncRNA_cl$cluster == 1])
lncRNA_cl2 <- names(lncRNA_cl$cluster[lncRNA_cl$cluster == 2])
lncRNA_cl3 <- names(lncRNA_cl$cluster[lncRNA_cl$cluster == 3])

lncRNA_cl1_pwm <- kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers[lncRNA %in% lncRNA_cl1, ] %>%
  colSums(.) %>%
  sort(., decreasing = T) %>%
  head(100) %>%
  names(.) %>%
  paste0(., collapse = "") %>%
  str_split(., pattern = '') %>%
  unlist() %>%
  matrix(., ncol = 6, byrow = T) %>%
  apply(., MARGIN = 2, FUN = function(x) table(x)) %>%
  apply(., MARGIN = 2, FUN = function(x) x/100)

lncRNA_cl2_pwm <- kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers[lncRNA %in% lncRNA_cl2, ] %>%
  colSums(.) %>%
  sort(., decreasing = T) %>%
  head(100) %>%
  names(.) %>%
  paste0(., collapse = "") %>%
  str_split(., pattern = '') %>%
  unlist() %>%
  matrix(., ncol = 6, byrow = T) %>%
  apply(., MARGIN = 2, FUN = function(x) table(x)) %>%
  apply(., MARGIN = 2, FUN = function(x) x/100)
```

```
lncRNA_cl3_pwm <- kmer_counts_pairwise_DE$canonical_background_lncRNA_6mers[lncRNA %in% lncRNA_cl3, ] %>%
  colSums(.) %>%
  sort(., decreasing = T) %>%
  head(100) %>%
  names(.) %>%
  paste0(., collapse = "") %>%
  str_split(., pattern = '') %>%
  unlist() %>%
  matrix(., ncol = 6, byrow = T) %>%
  apply(., MARGIN = 2, FUN = function(x) table(x)) %>%
  apply(., MARGIN = 2, FUN = function(x) x/100)

lncRNA_cl1_pwm
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6]
## A 0.30 0.35 0.40 0.35 0.32 0.36
## C 0.08 0.03 0.04 0.03 0.04 0.07
## G 0.09 0.09 0.09 0.09 0.10 0.11
## T 0.53 0.53 0.47 0.53 0.54 0.46
```

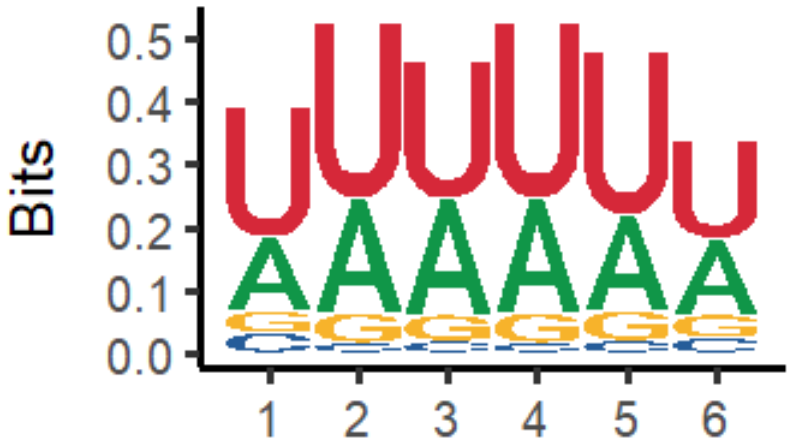
```
lncRNA_cl2_pwm
```

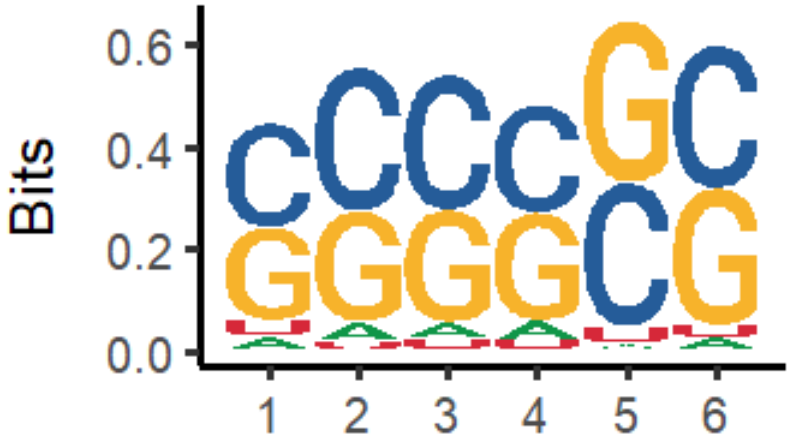
```
##   [,1] [,2] [,3] [,4] [,5] [,6]
## A 0.06 0.07 0.06 0.08 0.02 0.04
## C 0.46 0.51 0.49 0.44 0.44 0.47
## G 0.41 0.39 0.41 0.44 0.49 0.45
## T 0.07 0.03 0.04 0.04 0.05 0.04
```

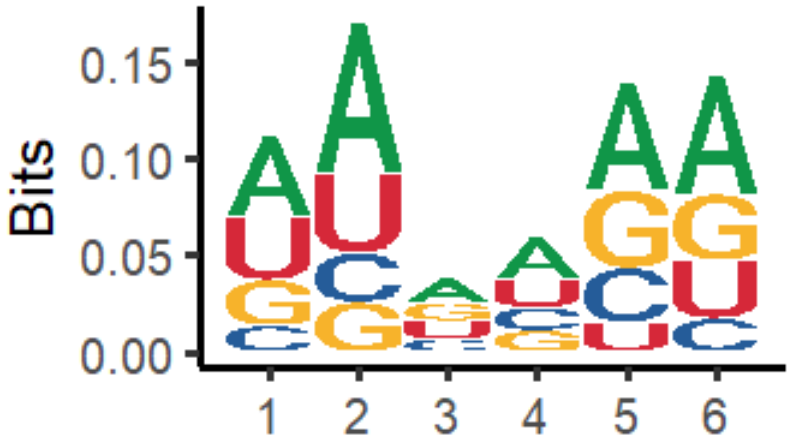
```
lncRNA_cl3_pwm
```

```
##   [,1] [,2] [,3] [,4] [,5] [,6]
## A 0.38 0.46 0.33 0.37 0.40 0.43
## C 0.12 0.15 0.17 0.20 0.20 0.12
## G 0.21 0.15 0.25 0.18 0.29 0.24
## T 0.29 0.24 0.25 0.25 0.11 0.21
```

```
rownames(lncRNA_cl1_pwm) <- gsub('T','U', rownames(lncRNA_cl1_pwm))
rownames(lncRNA_cl2_pwm) <- gsub('T','U', rownames(lncRNA_cl2_pwm))
rownames(lncRNA_cl3_pwm) <- gsub('T','U', rownames(lncRNA_cl3_pwm))

p1 <- ggseqlogo::ggseqlogo(lncRNA_cl1_pwm, method = 'bits' ) +
  theme_classic(base_size = 25, base_family = 'Arial')

p2 <- ggseqlogo::ggseqlogo(lncRNA_cl2_pwm, method = 'bits' ) +
  theme_classic(base_size = 25, base_family = 'Arial')

p3 <- ggseqlogo::ggseqlogo(lncRNA_cl3_pwm, method = 'bits' ) +
  theme_classic(base_size = 25, base_family = 'Arial')

p1/p2/p3
```


Bits
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6


Bits
0.6
0.4
0.2
0.0
1 2 3 4 5 6


Bits
0.15
0.10
0.05
0.00
1 2 3 4 5 6

```r
ggsave("plots/fig4C_lncRNA_6mers_logo_cl1_withUs.pdf", plot = p1, width = 6, height = 4, units = "in", dpi = 300, device = cairo_pdf)
ggsave("plots/fig4D_lncRNA_6mers_logo_cl2_withUs.pdf", plot = p2, width = 6, height = 4, units = "in", dpi = 300, device = cairo_pdf)
ggsave("plots/fig4E_lncRNA_6mers_logo_cl3_withUs.pdf", plot = p3, width = 6, height = 4, units = "in", dpi = 300, device = cairo_pdf)
```

SessionInfo

```r
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] ComplexHeatmap_2.14.0 patchwork_1.1.2       viridis_0.6.2
##  [4] viridisLite_0.4.1     pheatmap_1.0.12       RColorBrewer_1.1-3
##  [7] ggrepel_0.9.3         magrittr_2.0.3        forcats_1.0.0
## [10] stringr_1.5.0         dplyr_1.1.1           purrr_1.0.1
## [13] readr_2.1.4           tidyr_1.3.0           tibble_3.2.1
## [16] ggplot2_3.4.2         tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] matrixStats_0.63.0  fs_1.6.1            lubridate_1.9.2
##  [4] doParallel_1.0.17   httr_1.4.5          tools_4.2.1
##  [7] backports_1.4.1     utf8_1.2.3          R6_2.5.1
## [10] DBI_1.1.3           BiocGenerics_0.44.0 colorspace_2.1-0
## [13] GetoptLong_1.0.5    withr_2.5.0         tidyselect_1.2.0
## [16] gridExtra_2.3       compiler_4.2.1      cli_3.6.1
## [19] rvest_1.0.3         xml2_1.3.3          labeling_0.4.2
## [22] scales_1.2.1        digest_0.6.31       yulab.utils_0.0.6
## [25] rmarkdown_2.21      pkgconfig_2.0.3     htmltools_0.5.5
## [28] dbplyr_2.3.2        fastmap_1.1.1       highr_0.10
## [31] rlang_1.1.0         GlobalOptions_0.1.2 readxl_1.4.2
## [34] rstudioapi_0.14     shape_1.4.6         gridGraphics_0.5-1
## [37] generics_0.1.3      farver_2.1.1        jsonlite_1.8.4
## [40] googlesheets4_1.1.0 ggplotify_0.1.0     Rcpp_1.0.10
## [43] munsell_0.5.0       S4Vectors_0.36.2    fansi_1.0.4
## [46] lifecycle_1.0.3     stringi_1.7.12      yaml_2.3.7
## [49] plyr_1.8.8          ggseqlogo_0.1       parallel_4.2.1
## [52] crayon_1.5.2        haven_2.5.2         circlize_0.4.15
## [55] hms_1.1.3           knitr_1.42          pillar_1.9.0
## [58] rjson_0.2.21        codetools_0.2-18    stats4_4.2.1
## [61] reprex_2.0.2        glue_1.6.2          evaluate_0.20
## [64] modelr_0.1.11       png_0.1-8           vctrs_0.6.1
## [67] tzdb_0.3.0          foreach_1.5.2       cellranger_1.1.0
## [70] gtable_0.3.3        clue_0.3-64         xfun_0.38
## [73] broom_1.0.4         googledrive_2.1.0   gargle_1.3.0
## [76] iterators_1.0.14    IRanges_2.32.0      cluster_2.1.3
## [79] timechange_0.2.0
```