Annex 8: Overlapping with STARR peaks and TSSs

Juanma Medina

```r
# Run pca.R until obtaining the FULL dataframe
DHS_overlap <- full

# Rownames as one column
DHS_overlap$names <- rownames(DHS_overlap)

# Take only names and characterization
DHS_overlap <- DHS_overlap[, c("names", "final")]

# Separate location into chromosome, start and end by pattern finding
myregex <- "([0-9RHetLUXUextra]*):([0-9]*)-([0-9]*)"

DHS_overlap$chromosome <- gsub(pattern = myregex, "\\1", x = DHS_overlap$names)
DHS_overlap$start <- as.numeric(gsub(pattern = myregex, "\\2", x = DHS_overlap$names))
DHS_overlap$end <- as.numeric(gsub(pattern = myregex, "\\3", x = DHS_overlap$names))

# Take only chr, start, end and classification of each DHS
DHS_overlap <- DHS_overlap[, c("chromosome", "start", "end", "final")]

# Add "chr" to chromosome names
DHS_overlap$chromosome <- paste("chr", as.character(DHS_overlap$chromosome), sep = "")

# Write as .bed
write.table(DHS_overlap, file = "/home/juanma/ProjectX/DHSs.bed",
            row.names = F, quote = F, sep = "\t", col.names = F)

# BEDTOOLS overlapping with STARR-peaks with a 90% fraction (2300 out of 9538)

# Read overlapping results
starr <- read.table(file = "/home/juanma/ProjectX/DHS_STARR_intersect.bed")

# Reassemble chromosomic positions
starr$position <- apply(starr[ ,c("V2", "V3")], 1, paste, collapse = "-")
starr$position <- apply(starr[ ,c("V1", "position")], 1, paste, collapse = ":")

# Take specified columns
starr <- starr[, c("position", "V15")]

# Remove "chr"
starr$position <- gsub("chr", "", starr$position)

# First column as rownames
starr <- data.frame(starr[, -1], row.names = starr[, 1])

# Merge with characterized DHSs by rownames
starr_peaks <- merge(full, starr, by = 0, all = TRUE)

# Change last colname: indicator of STARR peak (enhancer)
colnames(starr_peaks)[21] <- "starr"
```

```r
# Remove annoying NAs
starr_peaks$starr <- factor(starr_peaks$starr, levels = c(levels(starr_peaks$starr), 0))
starr_peaks$starr[is.na(starr_peaks$starr)] <- 0
starr_peaks[is.na(starr_peaks)] <- 0

# Remove 56 DHSs not present in original dataset
starr_peaks <- starr_peaks[rowSums(starr_peaks == 0) <= 4, ]

# Enhancers
qplot(data = data.frame(pca_res$x), x = PC1, y = PC2, color = starr_peaks$starr,
      size=I(0.4)) +
  labs(x = paste("PC1", round(prop_var[1] * 100, digits = 2), "%", sep = " "),
       y = paste("PC2", round(prop_var[2] * 100, digits = 2), "%", sep = " ")) +
  xlim(-3, 3) + ylim(-2.5, 2.5) + labs(colour = 'STARR overlap') +
  scale_color_manual(labels = c("Enhancer", "No overlap"), values = c("red", "cyan")) +
  ggtitle("PCA of STARR peaks overlapping DHSs") +
  theme_bw(base_size = 8) +
  theme(legend.justification = "right",
        legend.margin = margin(-6, -6, -6, -6),
        legend.box.margin=margin(3, 3, 3, 3))

# Overlapping with promoters
# Parsing of the .vcf file in search of promoters
TSS <- read.table(file = "/home/juanma/ProjectX/fly_sorted.gtf", header = F)

# Keep only columns 1, 3, 4, and 5 (CHR, FEATURE, START, END)
TSS <- TSS[, c(1, 3, 4, 5)]

# Re-name the columns
names(TSS) <- c("CHR", "FEATURE", "START", "END")

# Keep only chr2L, chr2R, chr3L, chr3R and chrX
TSS <- TSS[TSS$CHR == "chr2L" | TSS$CHR == "chr2R" | TSS$CHR == "chr3L" |
             TSS$CHR == "chr3R" | TSS$CHR == "chrX", ]

# Keep unique regions
TSS <- TSS[!duplicated(TSS), ]

# As almost every CDS is overlapping 100% with the exon, remove them
TSS <- TSS[TSS$FEATURE != "CDS", ]

# Select the exonic regions previous to start codons (5 UTR)
UTR <- TSS[which(TSS$FEATURE == "start_codon") - 1, ]

# Filter out alternative start codons
UTR <- UTR[UTR$FEATURE == "exon", ]

# Add the midpoints of the UTRs as another column
UTR$MIDS <- round(UTR$END - ((UTR$END - UTR$START) / 2))

# Re-calculate the windows (+- 200 from the midpoints) to extract potential TSS
UTR$START <- UTR$MIDS - 200
UTR$END <- UTR$MIDS + 200
```

```r
# Remove exon and midpoints columns
UTR <- UTR[, c("CHR", "START", "END")]

# Write as .bed file
write.table(UTR, file = "/home/juanma/ProjectX/TSS.bed", row.names = F, quote = F, sep = "\t")

# BEDTOOLS overlapping with STARR_DHSs regions with a 30% fraction (148 out of 2300)

# Read overlapping results
tss_starr <- read.table(file = "/home/juanma/ProjectX/TSS_STARR_DHS.bed")

# Reassemble chromosomic positions
tss_starr$position <- apply(tss_starr[ ,c("V2", "V3")], 1, paste, collapse = "-")
tss_starr$position <- apply(tss_starr[ ,c("V1", "position")], 1, paste, collapse = ":")

# Take specified columns
tss_starr <- tss_starr[, c("position", "V18")]

# Remove "chr"
tss_starr$position <- gsub("chr", "", tss_starr$position)

# Remove a couple of duplicates
tss_starr <- tss_starr[!duplicated(tss_starr$position), ]

# First column as rownames
tss_starr <- data.frame(tss_starr[, -1], row.names = tss_starr[, 1])

# Merge with characterized DHSs by rownames
starr_peaks_tss <- merge(full, tss_starr, by = 0, all = TRUE)

# Change last colname: indicator of promoter
colnames(starr_peaks_tss)[21] <- "TSS"

# Remove annoying NAs
starr_peaks_tss$TSS <- factor(starr_peaks_tss$TSS, levels = c(levels(starr_peaks_tss$TSS), 0))
starr_peaks_tss$TSS[is.na(starr_peaks_tss$TSS)] <- 0
starr_peaks_tss[is.na(starr_peaks_tss)] <- 0

# Remove 44 DHSs not present in original dataset
starr_peaks_tss <- starr_peaks_tss[rowSums(starr_peaks_tss == 0) <= 4, ]

# Promoters
qplot(data = data.frame(pca_res$x), x = PC1, y = PC2, color = starr_peaks_tss$TSS,
      size=I(0.4)) +
  labs(x = paste("PC1", round(prop_var[1] * 100, digits = 2), "%", sep = " "),
       y = paste("PC2", round(prop_var[2] * 100, digits = 2), "%", sep = " ")) +
  xlim(-3, 3) + ylim(-2.5, 2.5) + labs(colour = 'Promoter overlap') +
  scale_color_manual(labels = c("Promoter", "No overlap"), values = c("red", "cyan")) +
  ggtitle("PCA of STARR peaks overlapping DHSs and promoters") +
  theme_bw(base_size = 8) +
  theme(legend.justification = "right",
        legend.margin = margin(-6, -6, -6, -6),
        legend.box.margin=margin(3, 3, 3, 3))
```