Annex 2: Duplicates elimination

Juanma Medina

UNIX

```
# Keep only chromosome, start, end, window and score
awk '{print $1,$2,$3,$4,$5}' windows_200.bed > windows_200_bedtools.bed

# Sort by chromosome and x1
sort -k1,1 -k2,2n windows_200_bedtools.bed > sorted_windows_200_bedtools.bed

# Tab-separate the file
sed 's/ /\t/g' sorted_windows_200_bedtools.bed > tab_sorted_windows_200_bedtools.bed

# Detection of overlapping with BEDtools (11907 original, 11045 unique, 862 overlapping)
bedtools merge -nms -i tab_sorted_windows_200_bedtools.bed > duplicates.txt

# Save duplicates
cut -f4 duplicates.txt | grep -o ";.*" | awk '{print substr($1,2); }' > my_duplicates_1.txt

cut -f4 duplicates.txt | grep -o ";.*" | awk '{print substr($1,2); }' | grep -o ";.*" |
awk '{print substr($1,2);}' > my_duplicates_2.txt
```

R

```
setwd("~/ProjectX")

# Initial 200-sized DHSs
original <- read.table("windows_200.bed", h = T)

# Duplicated windows detected with Bedtools
dups_1 <- read.table("my_duplicates_1.txt")
dups_2 <- read.table("my_duplicates_2.txt")

# Regex to parse out triplicates
myregex <- "(;.*)"

# Keep only what is before ; once (for duplicates)...
dups_1_fixed <- data.frame(gsub(pattern = myregex, "\\2", x = dups_1$V1))

# ... and again (for triplicates)
dups_2_fixed <- data.frame(gsub(pattern = myregex, "\\2", x = dups_2$V1))

# 1 quadriplicate left
dups_3_fixed <- data.frame("4:981220-981370")

colnames(dups_1_fixed) <- "duplicated_windows"
colnames(dups_2_fixed) <- "duplicated_windows"
colnames(dups_3_fixed) <- "duplicated_windows"
```

```r
# Merge all the duplicates
all_dups <- rbind(dups_1_fixed, dups_2_fixed, dups_3_fixed)

# Label duplicates accordingly
original$rep <- ifelse(original$chr_x1_x2 %in% all_dups$duplicated_windows,
                       "REP", "UNI")

# # Sanity check
# sum(original$rep == "REP")

# Final .bed file with no overlapping windows
windows_200_unique <- original[original$rep != "REP", c(1:5)]

# Write output
write.table(windows_200_unique, file = "windows_200_unique.bed", row.names = F, quote = F)
```