# This markdown file contains all script and annotations used for this project

## Load required packages:

```{r} ### Import Excel data install.packages(“readxl”)

### For creating heat map

install.packages(“ggplot2”)

install.packages(“dplyr”)

### Ensure tidyverse is installed

install.packages(“tidyr”)

### Load packages

library(“readxl”)

library(“ggplot2”)

library(“dplyr”)

## Set working directory to the local repository and import Excel data:  
```{r}  
### Set the working directory to the local repository   
setwd("~/Documents/ISC/Final\_project/biol-4386-course-project-tarahicks/Data/")  
  
### Get the current working directory  
current\_dir <- getwd()  
  
### Print the current working directory to the console  
print(current\_dir)  
  
### Import Excel data into RStudio:  
data <- read\_excel("DRIP\_seq\_data\_IIHG-tidied-2.28.2023-Sherrin.xlsx")  
  
### Check imported data:  
head(data)

## Clean up “Chromosome” column so R can recognize mitochondrial DNA and X chromosome:

{r} data\_clean <- data %>% mutate(., Chromosome = case\_when( as.character(Chromosome) == "Mt" ~ "Mt", as.character(Chromosome) == "X" ~ "X", TRUE ~ as.character(Chromosome) ))

## Convert “Range\_of\_position” column into two separate columns for the start and end position:

{r} data\_clean <- data\_clean %>% separate(Range\_of\_position, into = c("start\_pos", "end\_pos"), sep = "\\.\\.")

## Convert the “read\_count” column into a numeric value:

```{r} data\_cleanRead\_count)

## Calculate the percentage of reads for each locus using "Total\_read\_number" column:  
```{r}  
data\_clean <- data\_clean %>%  
 mutate(Percent\_reads = Read\_count / Total\_read\_number \* 100)

## Create heatmap using “Chromosome” column as the vertical axis, the “start\_pos” and “end\_pos” columns for horizontal axis, and the “Percent\_reads” column for the fill color.

```{r} ggplot(data\_clean, aes(x = start\_pos, y = Chromosome, fill = Percent\_reads)) + geom\_tile() + scale\_fill\_gradientn(colors = c(“blue”, “yellow”, “red”), values = c(0, 0.001, 0.01), name = “% Reads”) + labs(title = “Chromosome Heatmap”, x = “Position Range”, y = “Chromosome”)

## Try again :(  
```{r}  
ggplot(data = data\_clean, aes(x = Chromosome, y = start\_pos, fill = Percent\_reads)) +  
 geom\_tile() +  
 scale\_fill\_gradientn(colors = c("blue", "yellow", "red"), values = c(0, 0.001, 0.01), na.value = "gray") +  
 labs(title = "Distribution of DRIP-seq reads across genomic loci", x = "Chromosome", y = "", fill = "% Reads") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1)) +  
 coord\_fixed(ratio = 1/15)

## And again!

{r} ggplot(data\_clean, aes(x = Chromosome, y = start\_pos, fill = Percent\_reads)) + geom\_tile() + scale\_fill\_gradientn(colors = c("blue", "yellow", "red"), values = c(0, 0.001, 0.01), na.value = "white") + labs(x = "Chromosome", y = "Position", fill = "% Reads") + theme(legend.position = "right", panel.grid = element\_blank()) ## “one more time” (data\_clean only calculated % reads for the first row because the array wasn’t the same scale in the total reads column) {r} data\_clean <- data\_clean %>% mutate(Total\_read\_number = 1456985) data\_clean <- data\_clean %>% mutate(Percent\_reads = Read\_count / Total\_read\_number \* 100)

## Trying 1st code again after realizing issue with df

{r} ggplot(data\_clean, aes(x = start\_pos, y = Chromosome, fill = Percent\_reads)) + geom\_tile() + scale\_fill\_gradientn(colors = c("blue", "yellow", "red"), values = c(0, 0.001, 0.01), name = "% Reads") + labs(title = "Chromosome Heatmap", x = "Position Range", y = "Chromosome")

## Better! Now the key.

{r} ggplot(data\_clean, aes(x = start\_pos, y = Chromosome, fill = Percent\_reads)) + geom\_tile() + scale\_fill\_gradientn(colors = c("blue", "yellow", "red"), values = c(0, 0.001, 0.01), breaks = c(0, 0.001, 0.02, 0.1), name = "% Reads") + labs(title = "Chromosome Heatmap", x = "Position Range", y = "Chromosome")

## Oh boy…

```{r} library(scales)

# Rescale the data

dfPercent\_reads, to = c(0, 1))

# Define the breaks

my\_breaks <- seq(0, 1, length.out = 6)

# Define the colors

my\_colors <- c(“blue”, “green”, “yellow”, “orange”, “red”)

# Create the plot

ggplot(df, aes(x = Chromosome, y = Range\_of\_position, fill = Percent\_reads)) + geom\_tile() + scale\_fill\_gradientn(colors = my\_colors, values = my\_breaks, guide = guide\_colorbar(title = “% Reads”, barwidth = 10, barheight = 1)) + labs(title = “Heatmap of Percent Reads by Chromosome and Position”, x = “Chromosome”, y = “Position”) + theme\_minimal()

```

At this point, I think I’m making it worse. I have to admit defeat here, but truthfully I’m honestly just excited I got something that resembles what I was going for to some degree…