TEdensity

sharonX

2024-10-11

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

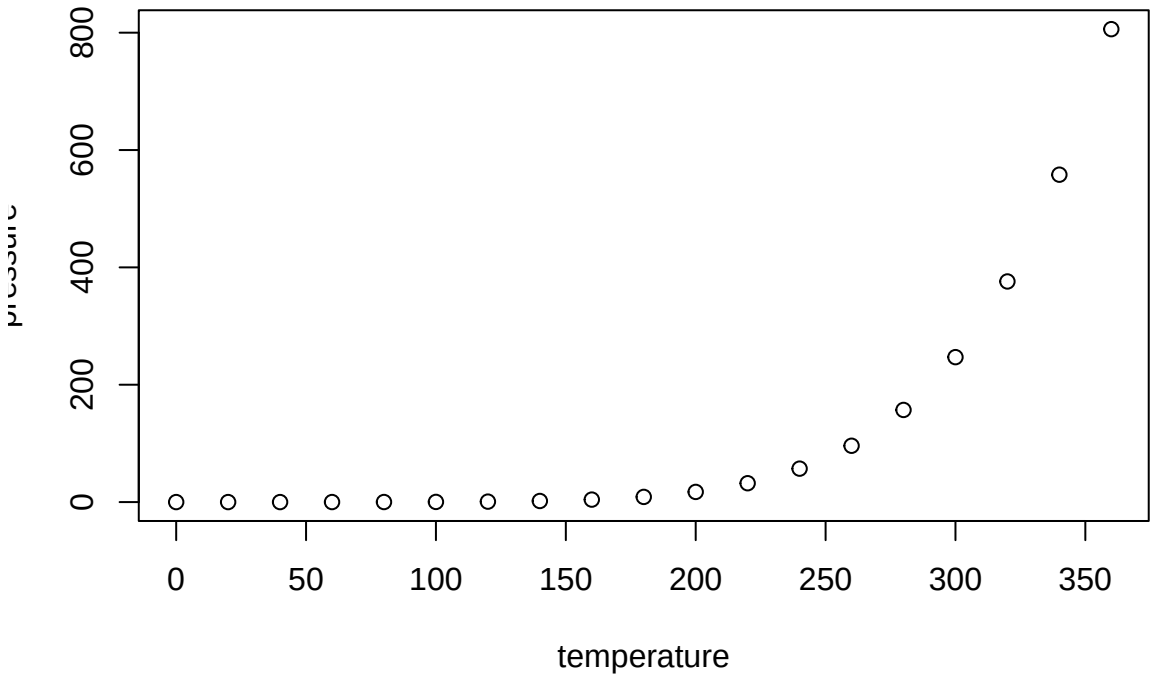
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
R0repeats<-read_tsv("~/bigdata/Genomecontent/GCcontent/Emus/R0repeats.gff3",col_names = F)
```

```
## Rows: 50747 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (7): X1, X2, X3, X6, X7, X8, X9
## dbl (5): X4, X5, X10, X11, X12
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
R0repeats<-R0repeats %>% mutate(length = X5-X4+1)
R0<-R0repeats %>% group_by(X8,length) %>% summarise(overlap=sum(X12))
```

```
## `summarise()` has grouped output by 'X8'. You can override using the `.groups`
## argument.
```

```
R0 <- R0 %>% mutate(perc = overlap/length)
R0$region<-"R0"
```

```
R1repeats<-read_tsv("~/bigdata/Genomecontent/GCcontent/Emus/R1repeats.gff3",col_names = F)
```

```
## Rows: 340118 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (7): X1, X2, X3, X6, X7, X8, X9
## dbl (5): X4, X5, X10, X11, X12
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
R1repeats<-R1repeats %>% mutate(length = X5-X4+1)
R1<-R1repeats %>% group_by(X8,length) %>% summarise(overlap=sum(X12))
```

```
## `summarise()` has grouped output by 'X8'. You can override using the `.groups`
## argument.
```

```
R1 <- R1 %>% mutate(perc = overlap/length)
R1$region<-"R1"
```

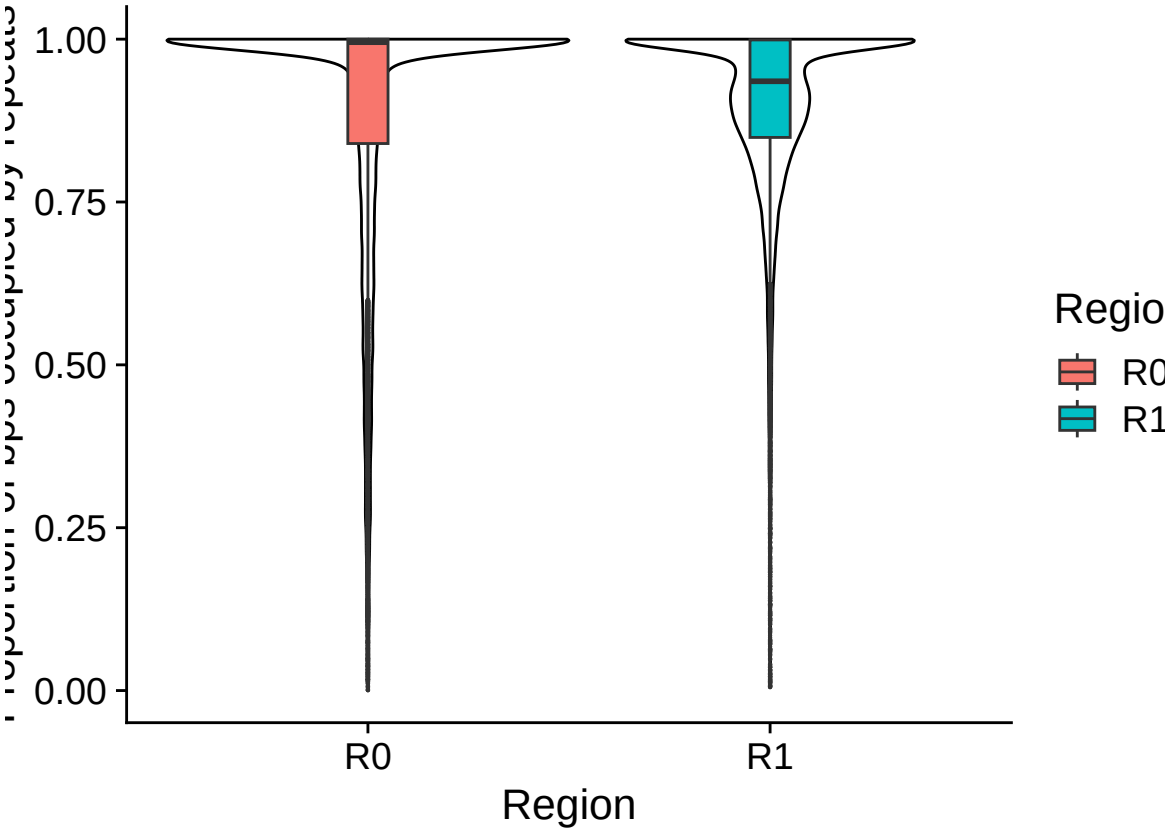
```
library(cowplot)
```

```
##
```

```
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
total<-rbind(R0,R1)
figure3b<-total %>% ggplot(aes(x=region, y=perc, fill=region)) +
  geom_violin(width =1,color="black",fill="white") +
  geom_boxplot(width = 0.1,outlier.size = 0.1)+
  theme_cowplot(16)+
  ylab("Proportion of bps occupied by repeats")+
  xlab("Region")+
  labs(fill = "Region")
figure3b
```




```
t.test(R0$perc,R1$perc,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  R0$perc and R1$perc
## t = -9.4317, df = 38165, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.01810881         Inf
## sample estimates:
## mean of x mean of y
```

```
## 0.8779021 0.8933217
```

```
# They do not follow normal distribution do the normality test first
# Perform a one-sided Wilcoxon rank-sum test (non-parametric)
wilcox.test(R0$perc, R1$perc, alternative = "greater")
```

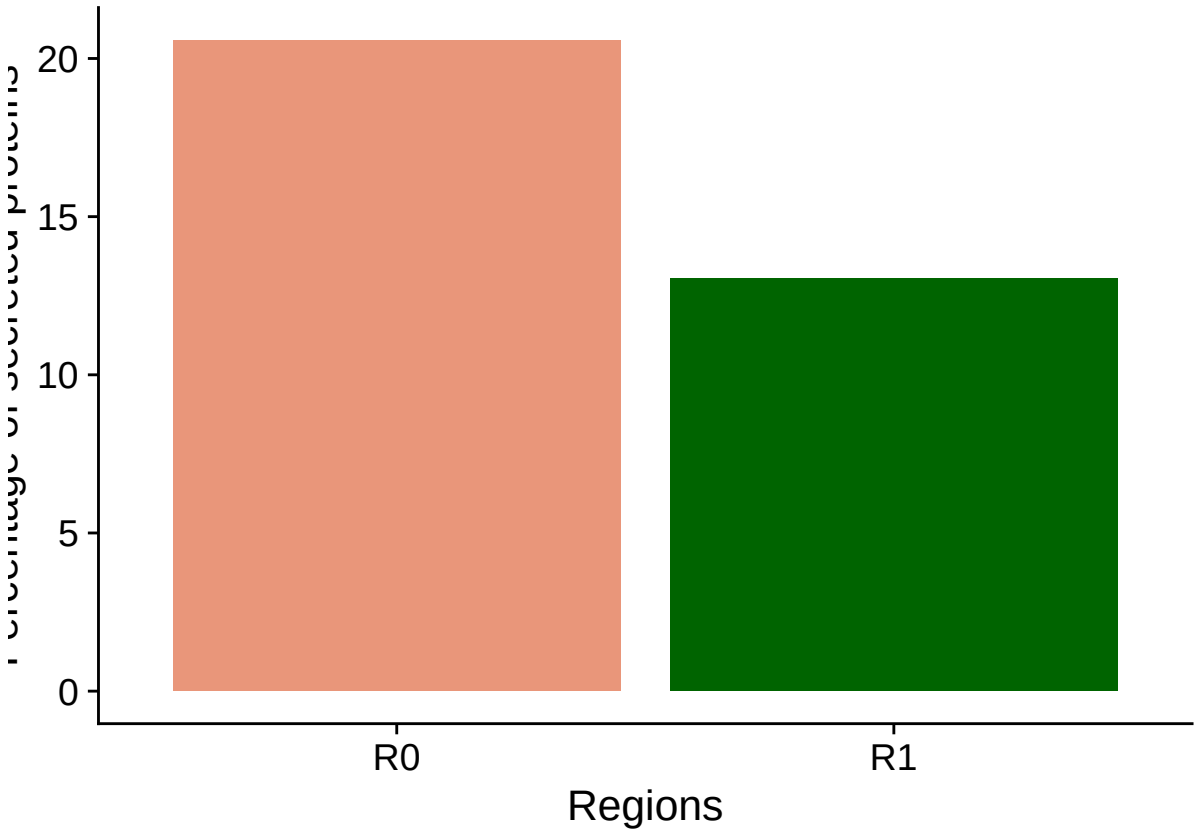
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  R0$perc and R1$perc
## W = 380508332, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

1. Test Statistic (W = 380508332): The test statistic W is the sum of the ranks for the group with the smaller total number of ranks. While you don't interpret this value directly in practice (it is not as intuitive as the t-value in a t-test), it is used to compute the p-value. The exact value of W is less important than the p-value and interpretation of the result.
2. p-value (p-value $<$ 2.2e-16): The p-value indicates the probability of observing a difference between R0$percandR1$perc as extreme as the one observed, assuming the null hypothesis is true (i.e., that there is no difference between the distributions of R0$percandR1$perc). A p-value smaller than 0.05 (which is extremely small here, i.e., $<$ 2.2e-16) suggests that the difference between the groups is statistically significant. Interpretation: Since your p-value is much smaller than any conventional significance level (e.g., 0.05), you can reject the null hypothesis. This means there is very strong evidence that the distribution of R0$percisshiftedhigherthanR1$perc.
3. Alternative Hypothesis (true location shift is greater than 0): The alternative hypothesis is that the location (median or central tendency) of R0$percisgreaterthanR1$perc. Since your p-value is extremely small, the data strongly supports this alternative hypothesis. Interpretation: The test result suggests that the central tendency (often interpreted as the median) of R0$percissignificantlygreaterthanthatofR1$perc.
4. Continuity Correction: In large sample sizes, a continuity correction is applied when approximating the distribution of the Wilcoxon test statistic by a normal distribution. This correction adjusts for the fact that ranks are discrete, while the normal distribution is continuous. The presence of the continuity correction doesn't change the interpretation, but it ensures that the p-value is more accurate, especially for large datasets like yours. Conclusion: Statistical significance: There is strong evidence (based on the p-value) to suggest that the distribution of R0$percissignificantlygreaterthanthatofR1$perc. Effect size: Since this is a non-parametric test, it compares ranks rather than means. This means the result is robust to outliers and non-normal distributions. Would you like to further explor

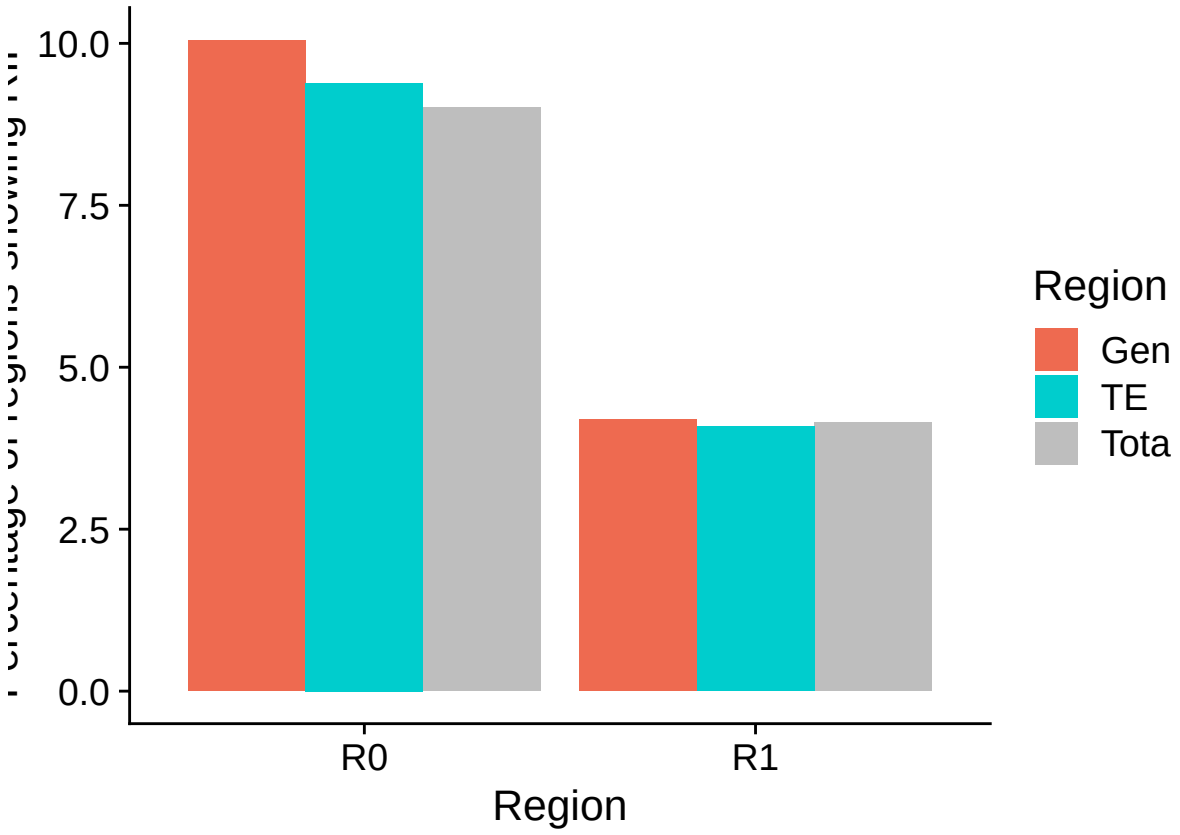
```
library(tidyverse)
library(cowplot)
library(viridis)
```

```
## Loading required package: viridisLite
```

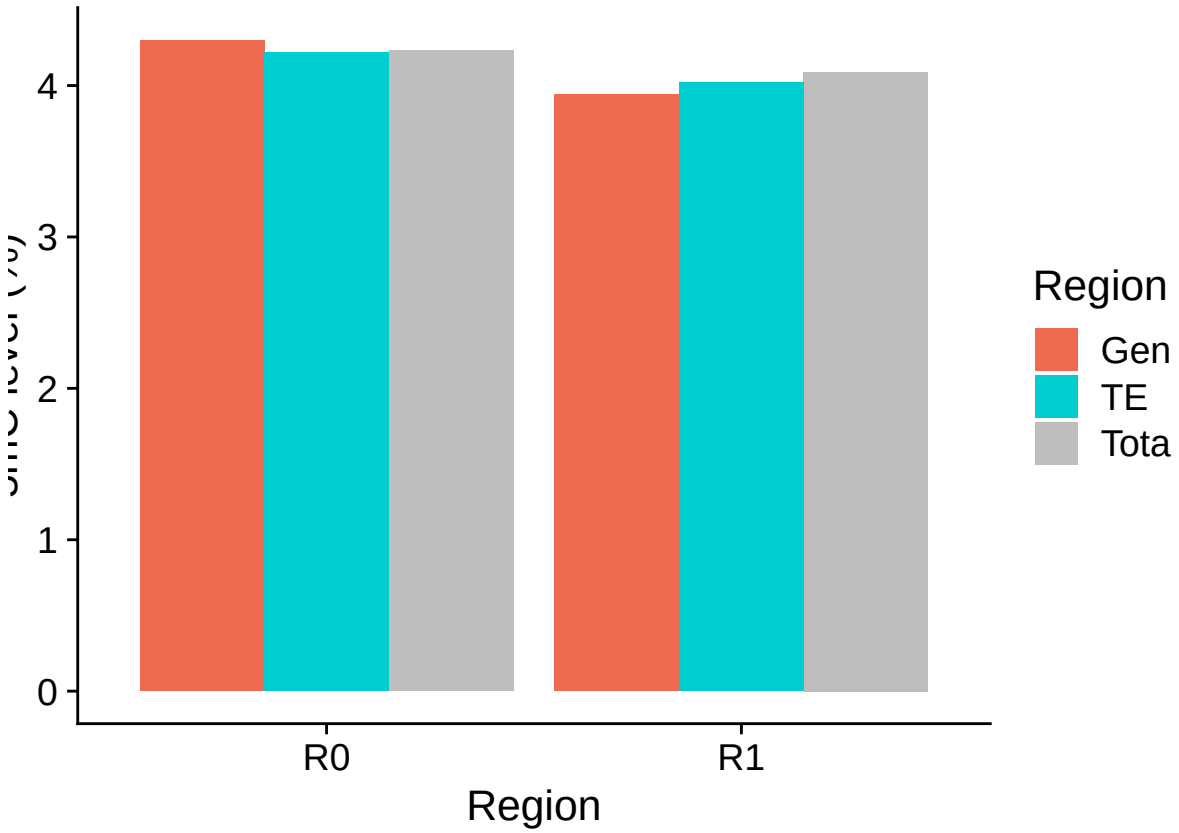
```
region<-c("R0","R1")
percentage<-c(20.58, 13.04)
secreted<-data.frame(region = unlist(region),percentage = unlist(percentage))
figure3c<-secreted %>% ggplot(aes(x=region,y=percentage,fill=region)) + geom_bar(stat = "identity",fill
  xlab("Regions") +
  ylab("Percentage of secreted proteins")+
  labs(fill = "Region")

figure3c
```



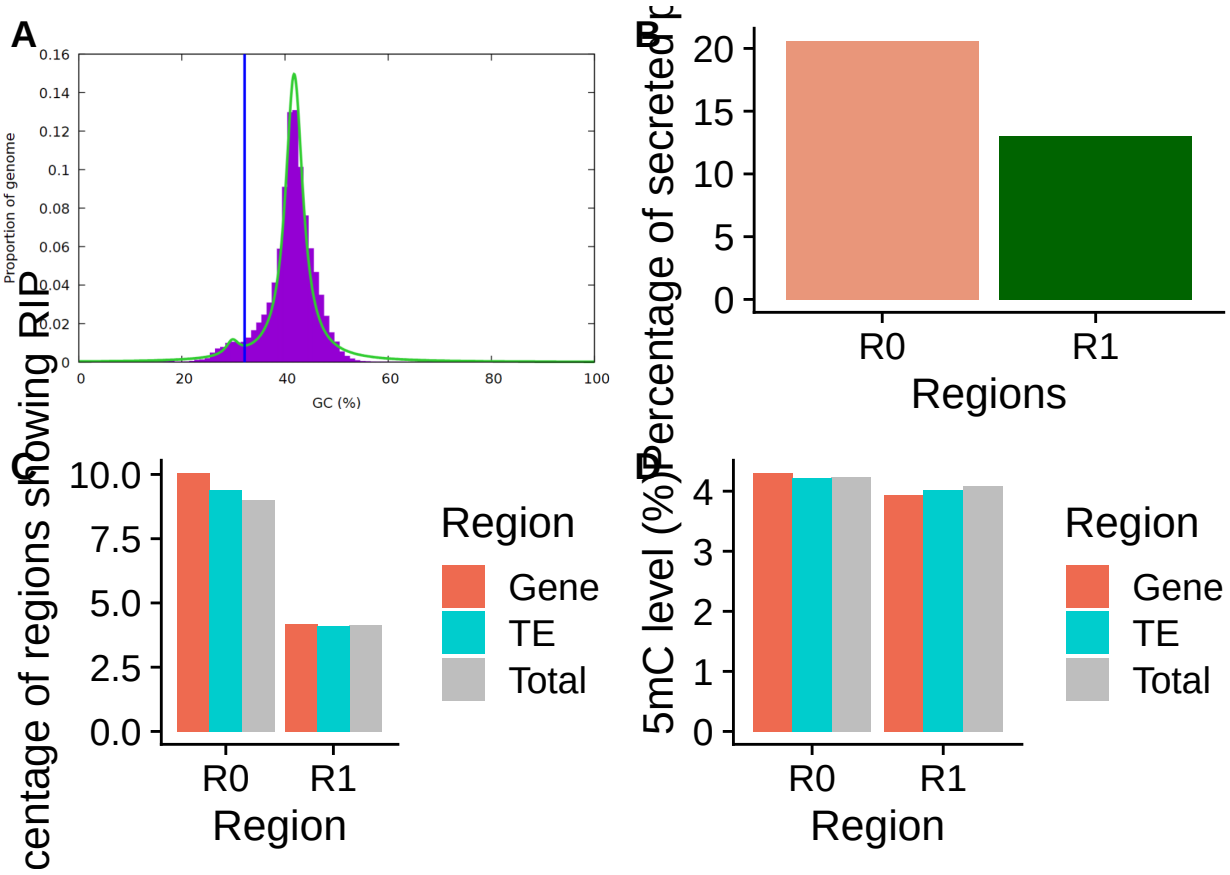

```
library(RColorBrewer)
rowname<-c("Total","Gene","TE","Total","Gene","TE")
percentage<-c(9.01,10.05,9.39,4.15,4.19,4.09)
regions<-c("R0","R0","R0","R1","R1","R1")
RIP<-data.frame(region = unlist(rowname),percentage = unlist(percentage), regions = unlist(regions))
my_pal<-c("coral2","cyan3","grey")
figure3d<-RIP %>% ggplot(aes(x=regions, y= percentage,fill = region)) +
  geom_bar(stat = "identity",position = "dodge")+
  theme_cowplot(16)+
  xlab("Region") +
  ylab("Percentage of regions showing RIP") +
  labs(fill = "Region")+
  scale_fill_manual(values=my_pal)
figure3d
```




```
library(RColorBrewer)
rowname<-c("Total","Gene","TE","Total","Gene","TE")
mc5<-c(4.23,4.30,4.22,4.09,3.94,4.02)
regions<-c("R0","R0","R0","R1","R1","R1")
methyl<-data.frame(region = unlist(rowname),mc5 = unlist(mc5), regions = unlist(regions))
my_pal<-c("coral2","cyan3","grey")
figure3e<-methyl %>% ggplot(aes(x=regions, y= mc5,fill = region)) +
  geom_bar(stat = "identity",position = "dodge")+
  theme_cowplot(16)+
  xlab("Region") +
  ylab("5mC level (%)") +
  labs(fill = "Region")+
  scale_fill_manual(values=my_pal)
figure3e
```




```
figure3a<- ggdraw() + draw_image("~/bigdata/Genomecontent/GCcontent/Emus/Emusgenome.png")

figure3<-plot_grid(figure3a,figure3c,figure3d,figure3e,
                   labels = c("A","B","C","D"),
                   nrow =2)
figure3
```




```
ggsave(filename = "~/bigdata/Genomecontent/GCcontent/Emus/R0R1regions.pdf",
       plot = figure3,
       width = 12,
       height = 10,
       dpi=600)
```

```
library(tidyverse)
R0repeats<-read_tsv("~/bigdata/Genomecontent/GCcontent/Emus/R0repeats1.gff3",col_names = F)
```

```
## Rows: 90403 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (12): X1, X2, X3, X6, X7, X8, X9, X10, X11, X15, X16, X17
## dbl  (6): X4, X5, X12, X13, X14, X18
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
R0repeats<-R0repeats %>% mutate(length = abs(X13-X12)+1)
R0repeats<-R0repeats %>% mutate(overlap=X18/length)
R0sum<-sum(R0repeats$overlap)
R0density<-R0sum/(22589*2.82)
#1.07 TE /kbp
```

```
R1repeats<-read_tsv("~/bigdata/Genomecontent/GCcontent/Emus/R1repeats1.gff3",col_names = F)
```

```
## Rows: 760947 Columns: 18
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: "\t"
## chr (12): X1, X2, X3, X6, X7, X8, X9, X10, X11, X15, X16, X17
## dbl  (6): X4, X5, X12, X13, X14, X18
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
R1repeats<-R1repeats %>% mutate(length = abs(X13-X12)+1)
R1repeats<-R1repeats %>% mutate(overlap=X18/length)
R1sum<-sum(R1repeats$overlap)
R1density<-R1sum/(28910*33.5)
# 0.76811 TE/kbp
```