Task8

2025-09-30

```
#install.packages("devtools")
#devtools::install_github("hirscheylab/tidybiology")
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.1     ✔ stringr   1.5.2
## ✔ ggplot2   4.0.0     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(tidybiology)

data("chromosome")
#view(chromosome)

# Summarise selected variables
chromosome_summary <- chromosome %>%
  summarise(
    across(c(variations, protein_codinggenes, mi_rna),
      list(
        mean = ~mean(.x, na.rm = TRUE),
        median = ~median(.x, na.rm = TRUE),
        max = ~max(.x, na.rm = TRUE)
      ),
      .names = "{.col}_{.fn}"
    )
  )

# Reshape to long format
summary_long <- chromosome_summary %>%
  pivot_longer(
    everything(),
    names_to = c("Variable", "Statistic"),
    names_sep = "_",
    values_to = "Value")
```

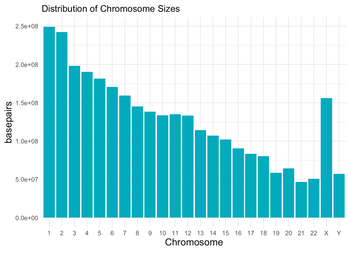
```
## Warning: Expected 2 pieces. Additional pieces discarded in 6 rows [4, 5, 6, 7,
## 8, 9].
```

```
# View result
summary_long
```

```
## # A tibble: 9 × 3
##   Variable   Statistic         Value
##   <chr>      <chr>             <dbl>
## 1 variations mean           6484572.
## 2 variations median         6172346
## 3 variations max           12945965
## 4 protein    codinggenes       850.
## 5 protein    codinggenes       836
## 6 protein    codinggenes      2058
## 7 mi         rna               73.2
## 8 mi         rna               75
## 9 mi         rna              134
```

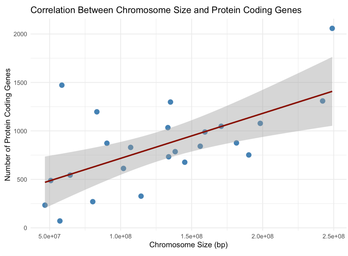
```
chromosome_size <- ggplot(chromosome, aes(x = id, y = basepairs)) +
  geom_bar(stat = "identity", fill = "#00AEBE") +
  labs(title = "Distribution of Chromosome Sizes", x = "Chromosome", y = "basepairs") +
  theme_minimal() +
  theme(axis.title.x = element_text(size = 14, family = "Arial"),
        axis.title.y = element_text(size = 14, family = "Arial"),
        axis.text.x = element_text(size = 9, family = "Arial"),
        axis.text.y = element_text(size = 9, family = "Arial"))

chromosome_size
```



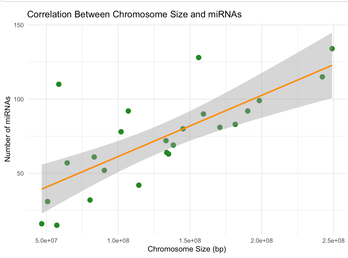

```
# Plot 1: Chromosome size vs. number of protein coding genes
ggplot(chromosome, aes(x = basepairs, y = protein_codinggenes)) +
  geom_point(color = "steelblue", size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  labs(
    title = "Correlation Between Chromosome Size and Protein Coding Genes",
    x = "Chromosome Size (bp)",
    y = "Number of Protein Coding Genes"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Plot 2: Chromosome size vs. number of miRNAs
ggplot(chromosome, aes(x = basepairs, y = mi_rna)) +
  geom_point(color = "forestgreen", size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "darkorange") +
  labs(
    title = "Correlation Between Chromosome Size and miRNAs",
    x = "Chromosome Size (bp)",
    y = "Number of miRNAs"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
data("proteins")
#view(proteins)

# Summarise selected variables
proteins_summary <- proteins %>%
  summarise(
    across(
      c(length, mass),
      list(
        mean = ~mean(.x, na.rm = TRUE),
        median = ~median(.x, na.rm = TRUE),
        max = ~max(.x, na.rm = TRUE)
      ),
      .names = "{.col}_{.fn}"
    )
  )

# Reshape to long format
proteins_long <- proteins_summary %>%
  pivot_longer(
    everything(),
    names_to = c("Variable", "Statistic"),
    names_sep = "_",
    values_to = "Value")

proteins_long
```

```
## # A tibble: 6 × 3
##   Variable Statistic    Value
##   <chr>    <chr>        <dbl>
## 1 length   mean          557.
## 2 length   median        414
## 3 length   max         34350
## 4 mass     mean        62061.
## 5 mass     median      46140.
## 6 mass     max       3816030
```

```
ggplot(proteins, aes(x = length, y = mass)) +
  geom_point(color = "#F7B1AB", size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "#807182") +
  labs(
    title = "Correlation Between length and mass",
    x = "length",
    y = "mass"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```