Final Project

PBX

2025-08-19

Step0:Environment

Step1:Read and Adjust data

```
meta_raw  <- readr::read_csv("QBS103_GSE157103_series_matrix-1.csv", show_col_types = FALSE)
genes_raw <- readr::read_csv("QBS103_GSE157103_genes (1).csv", show_col_types = FALSE)

#cleaning names in dataset
meta <- meta_raw %>% clean_names()
glimpse(meta, width = 80)
```

```
## Rows: 126
## Columns: 25
## $ participant_id                          <chr> "COVID_01_39y_male_NonICU", "C~
## $ geo_accession                           <chr> "GSM4753021", "GSM4753022", "G~
## $ status                                  <chr> "Public on Aug 29 2020", "Publ~
## $ sample_submission_date                  <chr> "Aug 28 2020", "Aug 28 2020", ~
## $ last_update_date                        <chr> "Aug 29 2020", "Aug 29 2020", ~
## $ type                                    <chr> "SRA", "SRA", "SRA", "SRA", "S~
## $ channel_count                           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ source_name_ch1                         <chr> "Leukocytes from whole blood",~
## $ organism_ch1                            <chr> "Homo sapiens", "Homo sapiens"~
## $ disease_status                          <chr> "disease state: COVID-19", "di~
## $ age                                     <chr> "39", "63", "33", "49", "49", ~
## $ sex                                     <chr> "male", "male", "male", "male"~
## $ icu_status                              <chr> "no", "no", "no", "no", "no", ~
## $ apacheii                                <chr> "15", "unknown", "unknown", "u~
## $ charlson_score                          <dbl> 0, 2, 2, 1, 1, 1, 7, 7, 2, 1, ~
## $ mechanical_ventilation                  <chr> "yes", "no", "no", "no", "yes"~
## $ ventilator_free_days                    <dbl> 0, 28, 28, 28, 23, 28, 28, 0, ~
## $ hospital_free_days_post_45_day_followup <dbl> 0, 39, 18, 39, 27, 36, 42, 0, ~
## $ ferritin_ng_ml                          <chr> "946", "1060", "1335", "583", ~
## $ crp_mg_l                                <chr> "73.1", "unknown", "53.2", "25~
## $ ddimer_mg_l_feu                         <chr> "1.3", "1.03", "1.48", "1.32",~
## $ procalcitonin_ng_ml                     <chr> "36", "0.37", "0.07", "0.98", ~
## $ lactate_mmol_l                          <chr> "0.9", "unknown", "unknown", "~
## $ fibrinogen                              <chr> "513", "unknown", "513", "949"~
## $ sofa                                    <chr> "8", "unknown", "unknown", "un~
```

```
meta <- meta %>%
  mutate(across(where(is.character), ~str_trim(.x))) %>%
```

```r
  mutate(across(where(is.character),
                ~ifelse(str_to_lower(.x) %in% c("unknown","na",""),
                        NA_character_, .x)))

#creating 3 continuous variables + categorical variables
clean_num <- function(x) {
  x %>%
    str_trim() %>%
    na_if("") %>% na_if("NA") %>% na_if("na") %>% na_if("unknown") %>% na_if(":") %>%
    readr::parse_number(locale = readr::locale(decimal_mark = ".", grouping_mark = ","))
}

meta <- meta %>%
  mutate(
    age      = clean_num(age),
    ferritin = clean_num(ferritin_ng_ml),
    crp      = clean_num(crp_mg_l)
  )

meta <- meta %>%
  mutate(
    sex = case_when(
      str_to_lower(sex) == "female" ~ "Female",
      str_to_lower(sex) == "male"   ~ "Male",
      TRUE                          ~ "Unknown"
    ),
    icu_status = case_when(
      str_to_lower(icu_status) %in% c("yes","icu")                    ~ "ICU",
      str_to_lower(icu_status) %in% c("no","nonicu","non-icu")        ~ "Non-ICU",
      TRUE                                                            ~ "Unknown"
    ),
    disease_status = case_when(
      !is.na(disease_status) & str_detect(str_to_lower(disease_status), "covid") ~ "COVID-19",
      TRUE                                                                       ~ "Non-COVID"
    )
  ) %>%
  mutate(
    sex            = factor(sex,            levels = c("Female","Male","Unknown")),
    icu_status     = factor(icu_status,     levels = c("ICU","Non-ICU","Unknown")),
    disease_status = factor(disease_status, levels = c("COVID-19","Non-COVID"))
  )

#transpose matrix and combining dataset
genes_t <- genes_raw %>%
  rename(gene = 1) %>%
  column_to_rownames("gene") %>%
  t() %>% as.data.frame() %>%
  rownames_to_column("participant_id") %>%
  as_tibble()

full_data <- meta %>% left_join(genes_t, by = "participant_id")

#checking
```

```r
dim(full_data)
```

```
## [1] 126 127
```

Step3: Generating Latex summary table (stratified by icu_status)

```r
# Step3: Generate LaTeX summary table (stratified by icu_status)

#selecting variables
table_dat <- full_data %>%
  dplyr::select(icu_status, sex, disease_status, age, crp, ferritin)

tab1 <- gtsummary::tbl_summary(
  data = table_dat,
  by = icu_status,
  statistic = list(
    gtsummary::all_continuous()  ~ "{mean} ({sd})",
    gtsummary::all_categorical() ~ "{n} ({p}%)"
  ),
  digits  = gtsummary::all_continuous() ~ 1,
  missing = "ifany",
  label = list(
    age            ~ "Age (years)",
    crp            ~ "CRP (mg/L)",
    ferritin       ~ "Ferritin (ng/mL)",
    sex            ~ "Sex",
    disease_status ~ "Disease status"
  )
) %>%
  gtsummary::add_overall(last = TRUE) %>%
  gtsummary::bold_labels() %>%
  gtsummary::modify_caption("**Summary statistics stratified by ICU status**")

#export Latex
library(kableExtra)
dir.create("tables", showWarnings = FALSE)

invisible(tab1)

latex_tab1 <- gtsummary::as_kable_extra(
  x = tab1, format = "latex", booktabs = TRUE, escape = FALSE
) %>%
  kable_styling(latex_options = "HOLD_position")

invisible(save_kable(latex_tab1, "tables/summary_table.tex"))
```

Step4: Making previous graph

```r
#select main gene
gene_main <- "AAK1"

#histogram
```

```
p_hist <- ggplot(full_data, aes(x = .data[[gene_main]])) +
  geom_histogram(binwidth = 0.5, color = "white") +
  labs(
    title = glue("Histogram of {gene_main} Expression"),
    x = glue("{gene_main} Expression (a.u.)"),
    y = "Count"
  ) +
  theme(plot.title = element_text(hjust = 0.04))

ggsave("figs/fig1_histogram_AAK1.png", p_hist, width = 6, height = 4, dpi = 300)

#scatter plot
p_scatter <- ggplot(full_data, aes(x = ferritin, y = .data[[gene_main]], color = icu_status)) +
  geom_point() +
  labs(
    title = glue("{gene_main} vs Ferritin"),
    x = "Ferritin (ng/mL)",
    y = glue("{gene_main} Expression"),
    color = "ICU"
  ) +
  scale_x_continuous(labels = scales::label_comma()) +
  scale_color_brewer(palette = "Set1") +
  theme(plot.title = element_text(hjust = 0.04))

ggsave("figs/fig2_scatter_AAK1_ferritin.png", p_scatter, width = 6, height = 4, dpi = 300)

#boxplot
p_box <- ggplot(full_data, aes(x = sex, y = .data[[gene_main]], fill = icu_status)) +
  geom_boxplot() +
  labs(
    title = glue("{gene_main} Expression by Sex and ICU Status"),
    x = "Sex",
    y = glue("{gene_main} Expression"),
    fill = "ICU"
  ) +
  scale_fill_brewer(palette = "Set2") +
  theme(plot.title = element_text(hjust = 0.04))

ggsave("figs/fig3_box_AAK1_sex_icu.png", p_box, width = 6, height = 4, dpi = 300)
```

Step5:Heatmap

Step6: hexbin plot

Introduction

This report analyzes gene expression profiles from the dataset with clinical covariates. I focus on AAK1 for the main figures and summarize key continuous (age, CRP, ferritin) and categorical variables (sex, ICU status, disease status). Data were cleaned (string trimming, standardization of categories), merged with the gene matrix, and visualized.

Methods

I analyzed a patient dataset containing both clinical variables and gene expression data. The clinical variables included three continuous measures (age, CRP, and ferritin) and three categorical variables (sex, ICU status, and disease status). For descriptive analysis, I created a LaTeX-formatted summary table stratified by ICU status.

All analyses were performed in **R version 4.3.1**. I used the following R packages: `tidyverse` for data manipulation and visualization, `janitor` for data cleaning, `gtsummary` and `kableExtra` for generating formatted summary tables, `scales` for axis scaling, `ComplexHeatmap` and `circlize` for heatmap generation, `glue` and `patchwork` for combining plots, and `hexbin` for hexagonal binning of dense scatterplots.

For the main plots, I produced:
(i) a histogram of AAK1 expression to examine its distribution;
(ii) a scatterplot of AAK1 expression versus ferritin, colored by ICU status;
(iii) a boxplot of AAK1 expression by sex with ICU status as the grouping variable.
For the heatmap, I first selected the top 10 most variable genes (by variance), while ensuring that `AAK1` was included. Then I constructed a gene-by-sample matrix and performed row-wise z-score normalization of gene expression. I also applied hierarchical clustering to both rows (genes) and columns (samples), using Euclidean distance as the distance metric and complete linkage as the clustering method and added two annotation bars to the heatmap to indicate patient sex and ICU status. To reduce visual complexity, I restricted the heatmap to the first 20 patients as an illustrative example.

Finally, I introduced an additional plot type, the `hexbin` plot, to explore the joint distribution of CRP and ferritin values. This method bins data into hexagonal cells and encodes cell counts with color intensity, allowing dense two-dimensional relationships to be visualized while minimizing overplotting.

Results

Summary Table

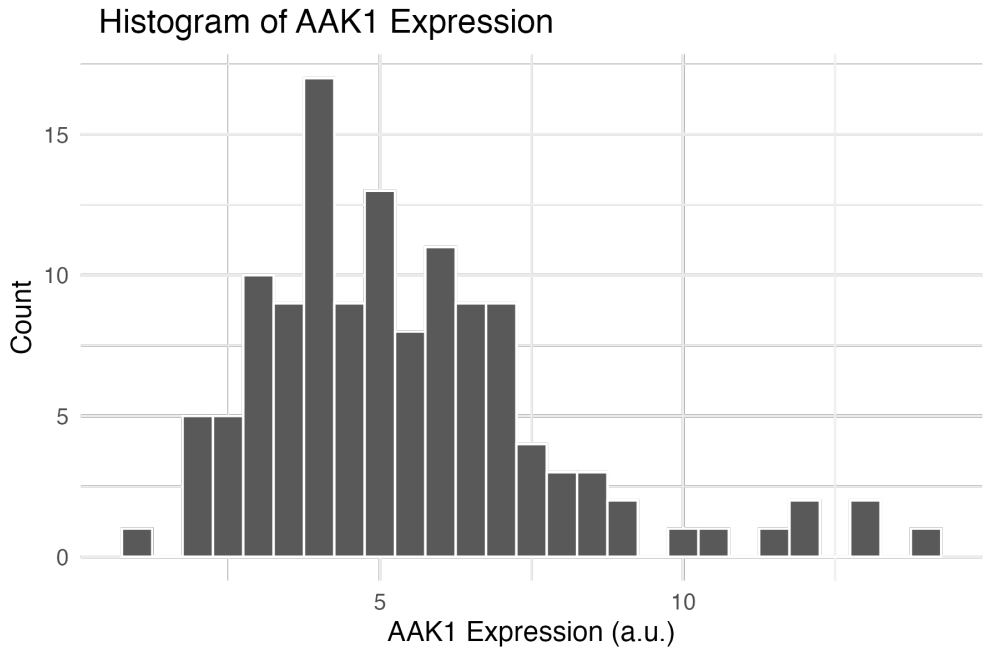
Table 1: **Summary statistics stratified by ICU status**

Characteristic	**ICU** N = 66	**Non-ICU** N = 60	**Unknown** N = 0	**Overall** N = 126
Sex				
Female	24 (36%)	27 (45%)	0 (NA%)	51 (40%)
Male	41 (62%)	33 (55%)	0 (NA%)	74 (59%)
Unknown	1 (1.5%)	0 (0%)	0 (NA%)	1 (0.8%)
Disease status				
COVID-19	66 (100%)	60 (100%)	0 (NA%)	126 (100%)
Non-COVID	0 (0%)	0 (0%)	0 (NA%)	0 (0%)
Age (years)	63.5 (14.0)	59.7 (18.4)	NA (NA)	61.7 (16.2)
Unknown	0	1	0	1
CRP (mg/L)	149.6 (105.5)	109.4 (94.4)	NA (NA)	131.2 (102.1)
Unknown	8	11	0	19
Ferritin (ng/mL)	935.3 (1,019.0)	715.7 (1,067.6)	NA (NA)	833.5 (1,042.8)
Unknown	7	9	0	16

[1] n (%); Mean (SD)

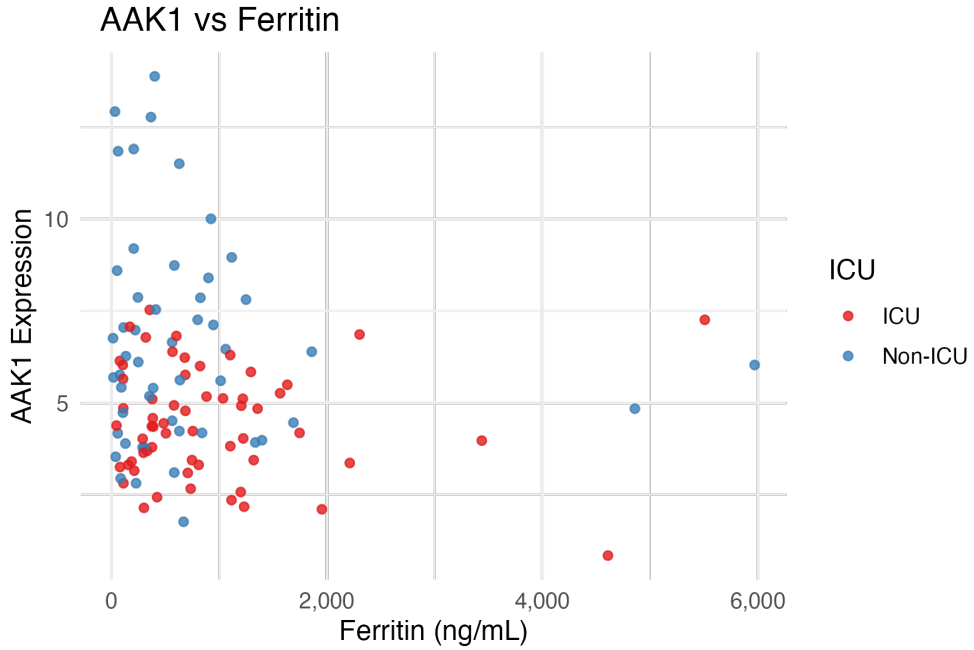
Figures

Figure 1. Histogram of AAK1 expression.



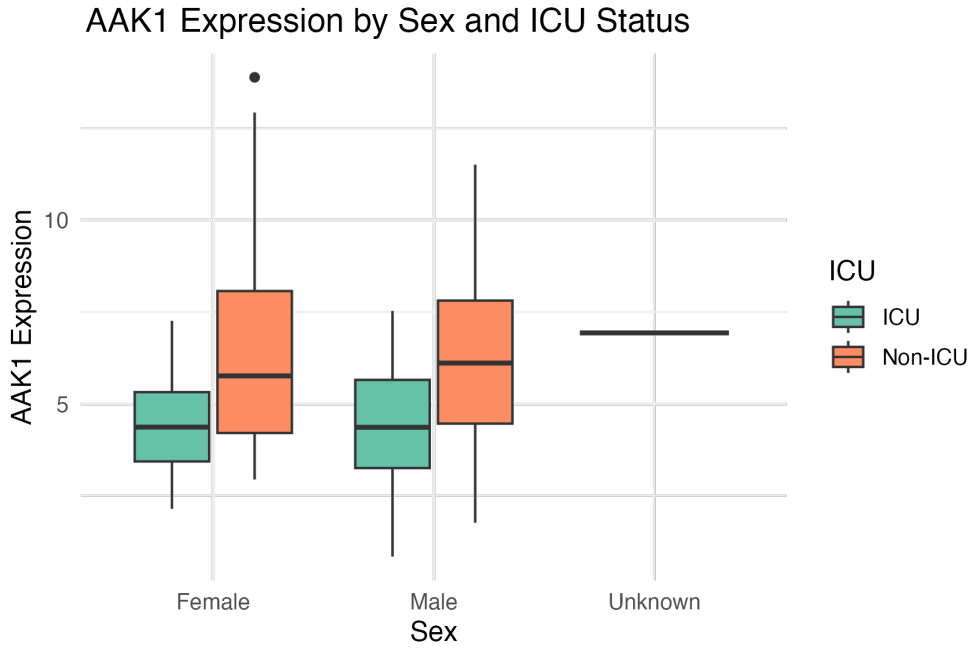
AAK1 expression is right-skewed, with most samples concentrated between 3–7 units, suggesting variability across patients.

Figure 2. Scatterplot of AAK1 expression vs ferritin, colored by ICU status.



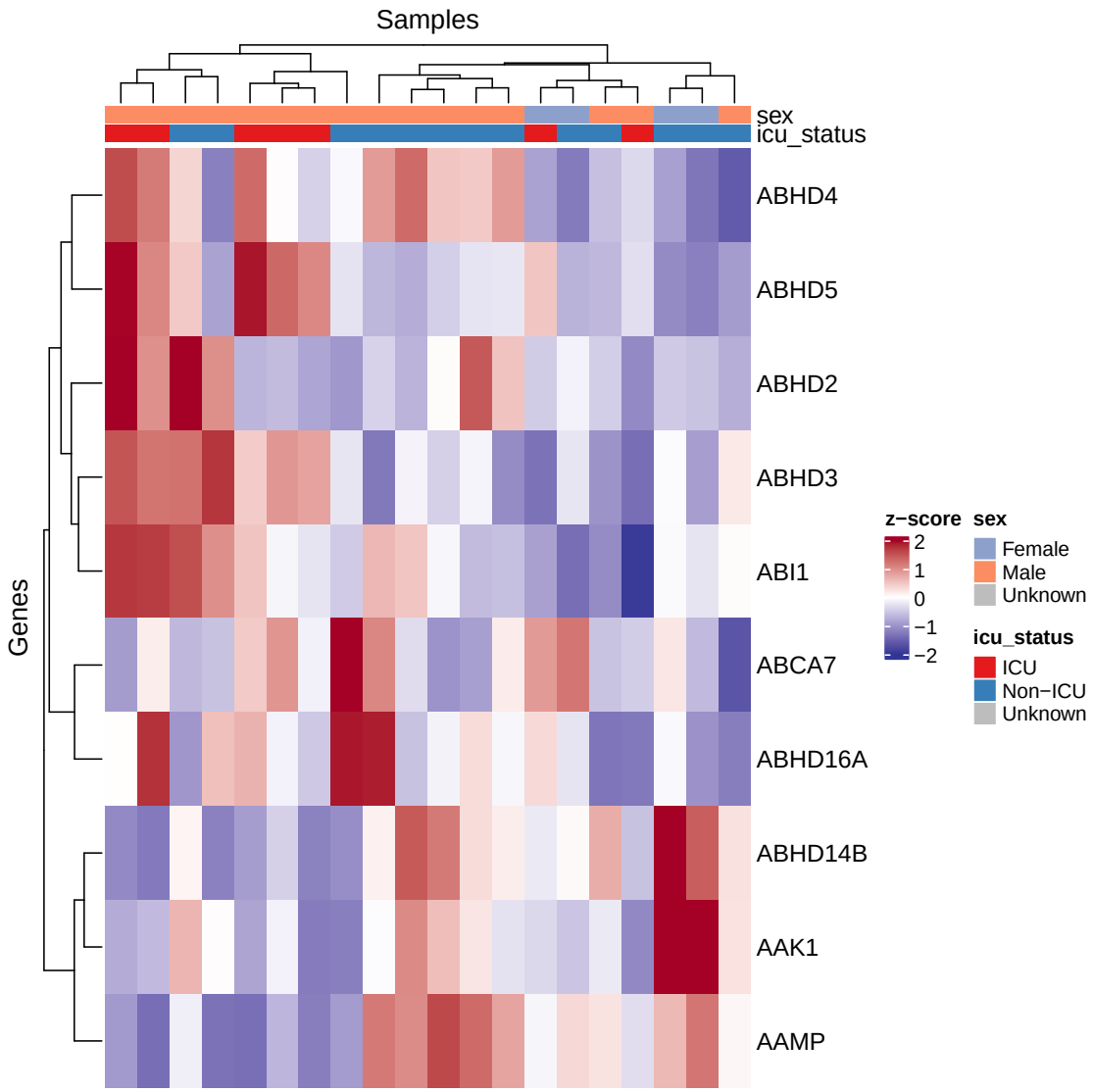
The scatterplot shows no clear linear relationship between ferritin and AAK1 expression, although most patients cluster at low ferritin values.

Figure 3. Boxplot of AAK1 expression by sex and ICU status.



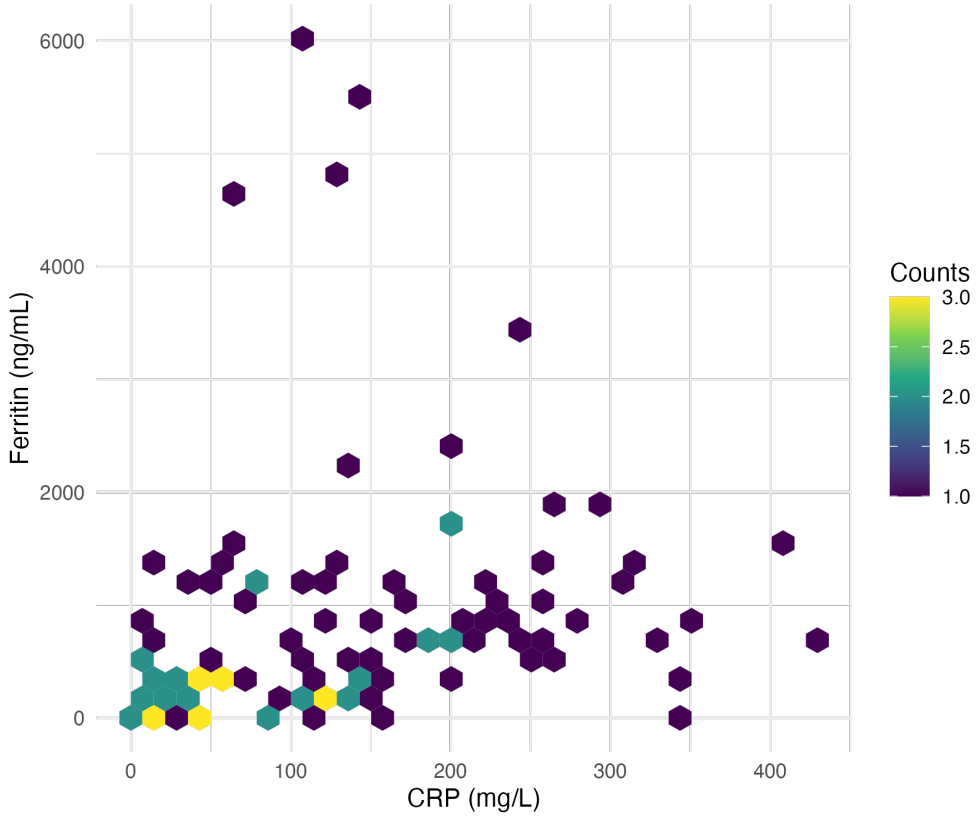
The boxplot suggests that ICU patients may have slightly higher AAK1 expression than non-ICU patients, with no major differences across sex groups.

Figure 4. Heatmap of the top 10 most variable genes.



The heatmap of the first 20 patients shows distinct clustering patterns, indicating that both ICU status and sex are associated with gene expression variation.

Figure 5. Hexbin plot of CRP vs Ferritin.



The hexbin plot shows most patients clustered in the low CRP and low ferritin range, while a few outliers display very high ferritin levels.

References

- Wickham H et al. (2019). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.3.1. Available at: https://CRAN.R-project.org/package=tidyverse
- Firke S. (2023). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.0. Available at: https://CRAN.R-project.org/package=janitor
- Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. (2021). gtsummary: Presentation-Ready Data Summaries and Analytic Result Tables. Available at: https://CRAN.R-project.org/package=gtsummary
- Zhu H. (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. Available at: https://CRAN.R-project.org/package=kableExtra
- Wickham H. (2022). scales: Scale Functions for Visualization. Available at: https://CRAN.R-project.org/package=scales
- Gu Z, Eils R, Schlesner M. (2016). ComplexHeatmap: Making Complex Heatmaps Simple. *Bioinformatics* 32(18):2847-2849. Available at: https://bioconductor.org/packages/ComplexHeatmap
- Gu Z, Gu L. (2014). circlize: Circular Visualization in R. *Bioinformatics* 30(19):2811-2812. Available at: https://CRAN.R-project.org/package=circlize

- Hester J, Wickham H, Chang W, Bryan J. (2023). glue: Interpreted String Literals. Available at: https://CRAN.R-project.org/package=glue
- Pedersen TL. (2020). patchwork: The Composer of Plots. Available at: https://CRAN.R-project.org/package=patchwork
- Carr DB, Lewin-Koh NJ, Maechler M. (2022). hexbin: Hexagonal Binning Routines. Available at: https://CRAN.R-project.org/package=hexbin