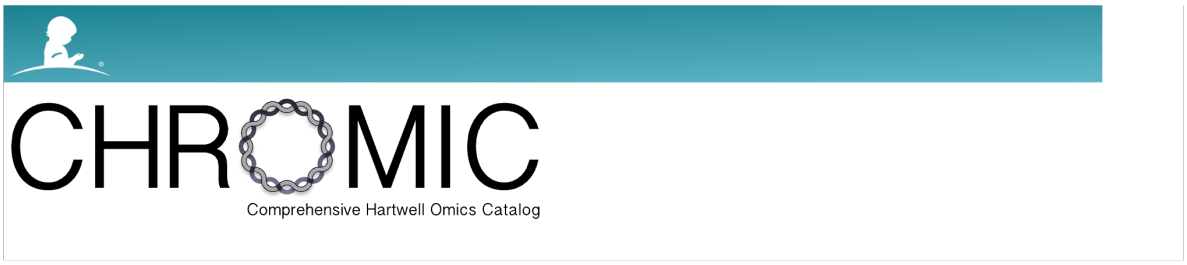


Data exploratory analysis

Antonia Chroni achroni@stjude.org for St. Jude Children's Research Hospital BioHackathon Team 1

Contents


```
## The following object is masked _by_ .GlobalEnv:
##
##     root_dir
```

Project: Comprehensive Omics Catalogue for Hartwell

St. Jude Children's Research Hospital BioHackathon Team 1

Date started: 09/04/2024 Date completed: 9/06/2024 Report generated: 13:57:51 CDT 09/11/2024

1 Information about this notebook

This is an exploratory analysis of the data availability in terms of assays in the Comprehensive Omics Catalogue for Hartwell. This is critical for mitigating duplicate sequencing requests and efforts through Hartwell. This notebook aims to showcase: (1) which samples have already been sequenced by Hartwell, and (2) what omics data are available per sample.

For demo purposes, we use dummy data cohort and subset by human brain tumor samples. Finally, we investigate the number of samples per `cancer_type_brain` and `Assay`.

2 Set up

```
suppressPackageStartupMessages({
  library(tidyverse)
})
```

3 Directories and paths to file Inputs/Outputs

```
attach(params)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     root_dir
```

```
analysis_dir <- file.path(root_dir, "analyses", "data-exploratory-analysis")
input_dir <- file.path(analysis_dir, "input")

# We will first read in metadata file as we need to define sample_name
metadata_file <- file.path(input_dir, input_file) # metadata input file
palette_file <- file.path(root_dir, "figures", "palettes", "assay_color_palette.tsv")
tumor_palette_file <- file.path(root_dir, "figures", "palettes", "tumor_color_palette.tsv")

# File path to `plots` directory
plots_dir <- file.path(analysis_dir, "plots")
if (!dir.exists(plots_dir)) {
  dir.create(plots_dir)}

figures_plots_dir <- file.path(plots_dir, "figures")
if (!dir.exists(figures_plots_dir)) {
  dir.create(figures_plots_dir)}

source(paste0(analysis_dir, "/util/function-create-barplot.R"))
source(paste0(root_dir, "/figures/scripts/theme_plot.R"))
```

4 Read metadata file

We will subset by human brain tumor samples.

```
# Read metadata
df <- read.csv(metadata_file, stringsAsFactors=FALSE)

# Number of samples per cancer_type_brain
```

```r
assays_number <- length(df$SJUID)
samples_number <- length(unique(df$SJUID))
```

```r
cancer_type_brain_order <- c("Ependymoma", "HGG", "LGG", "Medulloblastoma")

# Re-order df
f <- c("WES", "WGBS", "WGS", "RNAseq", "ATACseq", "ChIPseq", "Methylation")# Level df by assay
df <- df %>%
  dplyr::mutate(Assay = factor(Assay),
                Assay = fct_relevel(Assay, f)) %>%
  arrange(cancer_type_brain, Assay)
```

4.1 Color palette for plotting

```r
# Read color palette
palette_df <- readr::read_tsv(palette_file, guess_max = 100000, show_col_types = FALSE)

# Define and order palette
palette <- palette_df$hex_codes
names(palette) <- palette_df$color_names

# Read color palette for tumor type
tumor_palette_df <- readr::read_tsv(tumor_palette_file, guess_max = 100000, show_col_types = FALSE)

# Define and order palette
tumor_palette <- tumor_palette_df$hex_codes
names(tumor_palette) <- tumor_palette_df$color_names
```

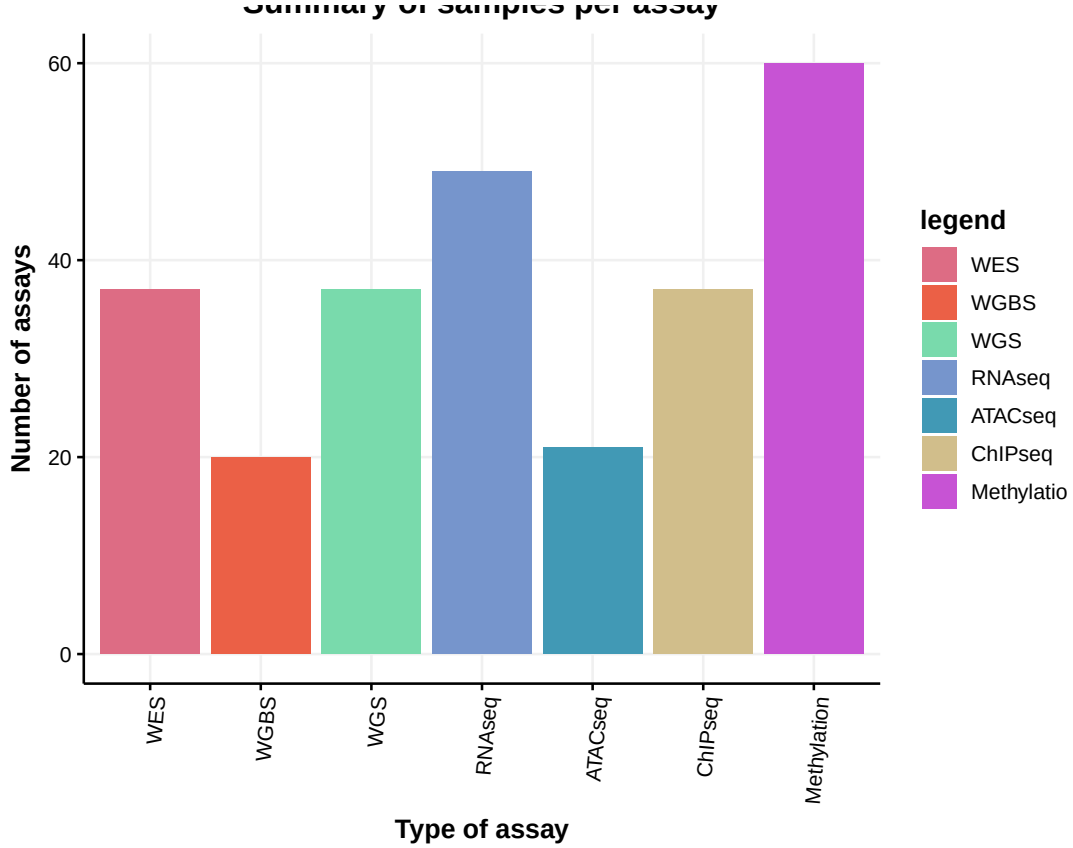
5 Number of samples with assay information

Table 1: Summary of samples per assay

Assay	n
WES	37
WGBS	20
WGS	37
RNAseq	49
ATACseq	21
ChIPseq	37
Methylation	60

```r
# Define parameters for function
ylim <- max(tables1$n)

# Run function
fname <- paste0(figures_plots_dir, "/", "samples-per-assay.pdf")
p <- create_barplot(plot_df = tables1,
                    ylim = ylim,
                    x_value = tables1$Assay,
                    use_palette = palette,
                    xtitle ="Type of assay",
                    title_value = caption_value)
```




```r
pdf(file = fname, width = 6, height = 5)
print(p)
dev.off()
```

```
## pdf
##   2
```

6 Number of samples per brain cancer type and assay

6.1 Overall assays

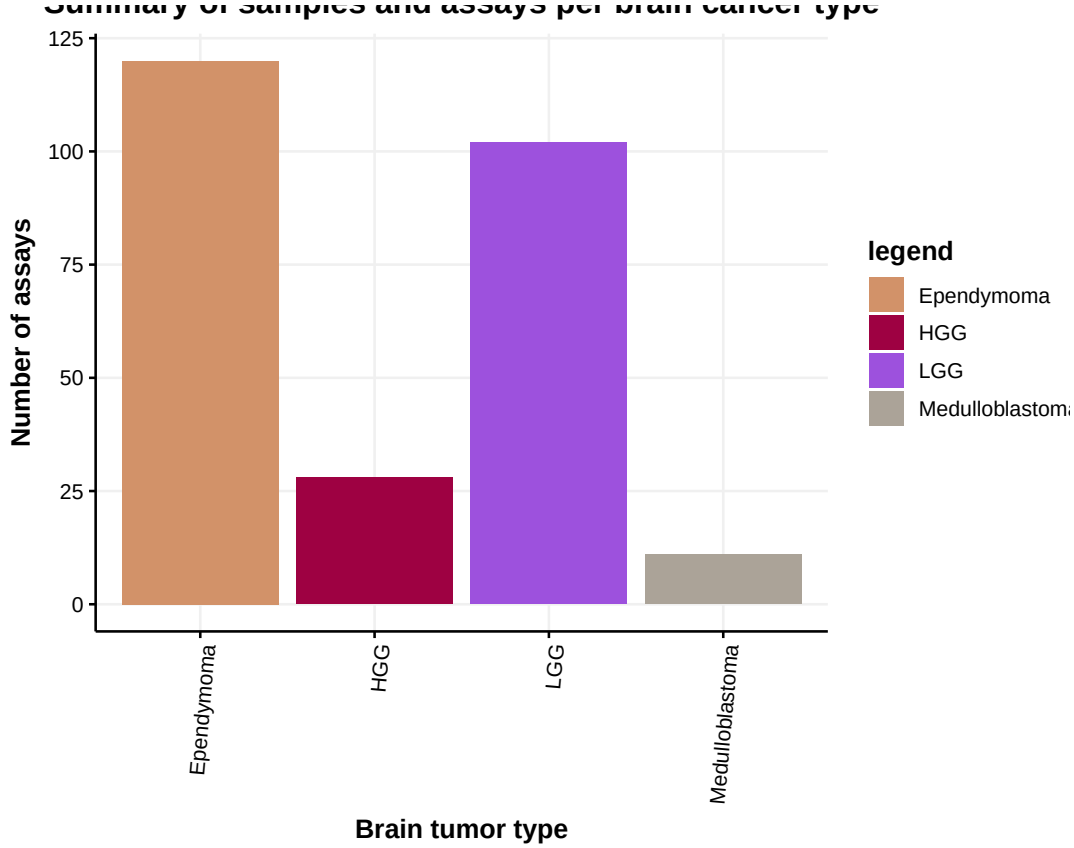
There are 60 brain tumor samples with 261 assays in total.

Table 2: Summary of samples and assays per brain cancer type

cancer_type_brain	n
Ependymoma	120
HGG	28
LGG	102
Medulloblastoma	11

```
# Define parameters for function
ylim <- max(tables1$n)

# Run function
fname <- paste0(figures_plots_dir, "/", "cancer-type-brain-assay-overall.pdf")
p <- create_barplot(plot_df = tables1,
                    ylim = ylim,
                    x_value = tables1$cancer_type_brain,
                    use_palette = tumor_palette,
                    xtitle ="Brain tumor type",
                    title_value = caption_value)
```




```
pdf(file = fname, width = 6, height = 5)
print(p)
dev.off()
```

```
## pdf
##   2
```

6.2 Per assay

Table 3: Summary of samples and assays per brain cancer type and per assay

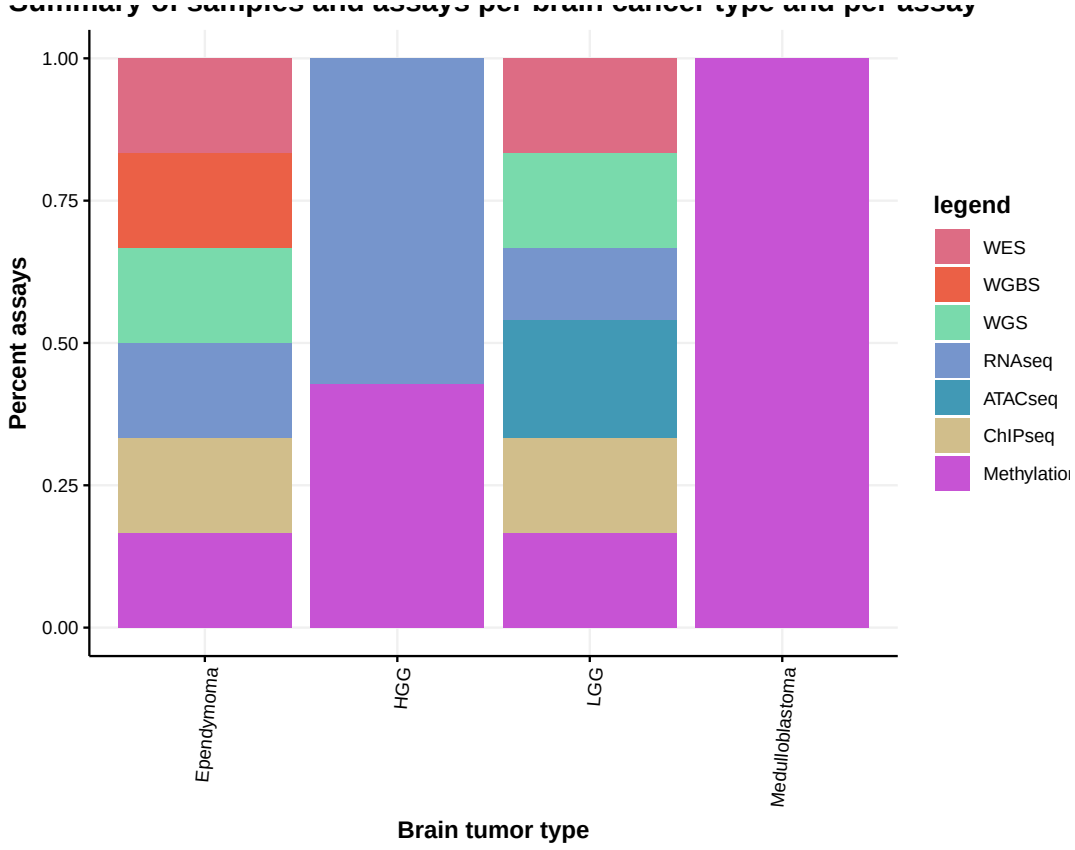
cancer_type_brain	WES	WGBS	WGS	RNAseq	ChIPseq	Methylation	ATACseq
Ependymoma	20	20	20	20	20	20	0

cancer_type_brain	WES	WGBS	WGS	RNAseq	ChIPseq	Methylation	ATACseq
HGG	0	0	0	16	0	12	0
LGG	17	0	17	13	17	17	21
Medulloblastoma	0	0	0	0	0	11	0

```
tables1 <- df %>% count(cancer_type_brain, Assay) %>%
  as.data.frame() %>%
  mutate_all(funs(replace_na(.,0)))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
# Plot stacked barplot
fname <- paste0(figures_plots_dir, "/", "cancer-type-brain-per-assay.pdf")
p <- create_stacked_barplot(plot_df = tables1,
                            x_value = tables1$cancer_type_brain,
                            use_palette = palette,
                            xtitle ="Brain tumor type",
                            legend = tables1$Assay,
                            title_value = caption_value)
```




```
pdf(file = fname, width = 6, height = 5)
print(p)
dev.off()
```

```
## pdf
```

7 Number of samples per brain cancer type, assay, and SJUID

Table 4: Summary of samples and assays per brain cancer type, per assay and per SJUID

cancer_type_brain	SJUID	WES	WGBS	WGS	RNAseq	ChIPseq	Methylation	ATAC-seq
Ependymoma	SJH0H5WYREP	1	1	1	1	1	1	0
Ependymoma	SJH2HPPEEKM	1	1	1	1	1	1	0
Ependymoma	SJH51B396IW	1	1	1	1	1	1	0
Ependymoma	SJH5HCKPC97	1	1	1	1	1	1	0
Ependymoma	SJH98QNKIUU	1	1	1	1	1	1	0
Ependymoma	SJH9YML-FOTI	1	1	1	1	1	1	0
Ependymoma	SJHA6KC56J6	1	1	1	1	1	1	0
Ependymoma	SJH-BKS7QSFO	1	1	1	1	1	1	0
Ependymoma	SJHC70DZRJS	1	1	1	1	1	1	0
Ependymoma	SJHCN03RCVD	1	1	1	1	1	1	0
Ependymoma	SJHD1KI-UMM6	1	1	1	1	1	1	0
Ependymoma	SJHFG-GJRHYO	1	1	1	1	1	1	0
Ependymoma	SJH-HZH67WMF	1	1	1	1	1	1	0
Ependymoma	SJHI8CC7QT8	1	1	1	1	1	1	0
Ependymoma	SJHIQT-NAYIF	1	1	1	1	1	1	0
Ependymoma	SJHKVQE-BOCP	1	1	1	1	1	1	0
Ependymoma	SJHKYUIYAME	1	1	1	1	1	1	0
Ependymoma	SJHPVXLA-CLM	1	1	1	1	1	1	0
Ependymoma	SJHUMKP2L6V	1	1	1	1	1	1	0
Ependymoma	SJHXTBTQ5ZT	1	1	1	1	1	1	0
HGG	SJH5QHZM4US	0	0	0	1	0	1	0
HGG	SJHADL5OJME	0	0	0	1	0	1	0
HGG	SJH-BISK5KBU	0	0	0	1	0	1	0
HGG	SJHBL7CDYRN	0	0	0	1	0	1	0
HGG	SJHD-DTE0SYL	0	0	0	1	0	0	0
HGG	SJHEOVURBMJ	0	0	0	1	0	1	0
HGG	SJHEQG3P4FK	0	0	0	1	0	1	0
HGG	SJHHPN-QERSQ	0	0	0	1	0	1	0
HGG	SJHHW23UJ9P	0	0	0	1	0	1	0
HGG	SJHJ59RLHSU	0	0	0	1	0	1	0
HGG	SJHKKDDX-OYH	0	0	0	1	0	1	0

cancer_type_brain	SJUID	WES	WGBS	WGS	RNAseq	ChIPseq	Methylation	ATAC-seq
HGG	SJHMI643DMD	0	0	0	1	0	1	0
HGG	SJHQBOPS9FD	0	0	0	1	0	0	0
HGG	SJHRO-JUQZAP	0	0	0	1	0	1	0
HGG	SJHUT-PISXFQ	0	0	0	1	0	0	0
HGG	SJHVIS5Q8HX	0	0	0	1	0	0	0
LGG	SJH2W47P7DG	1	0	1	1	1	1	1
LGG	SJHAY4GX4AN	1	0	1	1	1	1	1
LGG	SJHBN-JSZHW6	1	0	1	1	1	1	1
LGG	SJHBV3Q6UVR	1	0	1	1	1	1	1
LGG	SJH-CLGFJTIG	1	0	1	1	1	1	1
LGG	SJHD-DTE0SYL	1	0	1	0	1	1	1
LGG	SJHI52BLNWK	1	0	1	1	1	1	1
LGG	SJHN-PJTQHIT	1	0	1	1	1	1	1
LGG	SJHO-FORRR7C	1	0	1	1	1	1	1
LGG	SJHOY05OJJN	1	0	1	1	1	1	1
LGG	SJHQBOPS9FD	1	0	1	0	1	1	1
LGG	SJHUT-PISXFQ	1	0	1	0	1	1	1
LGG	SJHVIS5Q8HX	1	0	1	0	1	1	1
LGG	SJHWS0NRZVA	1	0	1	1	1	1	1
LGG	SJHXIKCWNKY	1	0	1	1	1	1	1
LGG	SJHYP-KTG3P5	1	0	1	1	1	1	1
LGG	SJHZW7GYEF9	1	0	1	1	1	1	1
LGG	SJH5HCKPC97	0	0	0	0	0	0	1
LGG	SJHC70DZRJS	0	0	0	0	0	0	1
LGG	SJHCN03RCVD	0	0	0	0	0	0	1
LGG	SJHIQT-NAYIF	0	0	0	0	0	0	1
Medulloblastoma	SJH77NRD-WUX	0	0	0	0	0	1	0
Medulloblastoma	SJH8HIE3P0P	0	0	0	0	0	1	0
Medulloblastoma	SJHAF-TIOMPQ	0	0	0	0	0	1	0
Medulloblastoma	SJHC1QST5GR	0	0	0	0	0	1	0
Medulloblastoma	SJHF-CYGKSDY	0	0	0	0	0	1	0
Medulloblastoma	SJHJAC-CGA3S	0	0	0	0	0	1	0
Medulloblastoma	SJHQS0D51KH	0	0	0	0	0	1	0
Medulloblastoma	SJHXVMEU21L	0	0	0	0	0	1	0
Medulloblastoma	SJHY4W1ZWCY	0	0	0	0	0	1	0
Medulloblastoma	SJHZ40PT-CYH	0	0	0	0	0	1	0

cancer_type_brain	SJUID	WES	WGBS	WGS	RNAseq	ChIPseq	Methylation	ATAC-seq
Medulloblastoma	SJHZZLR-CGJ6	0	0	0	0	0	1	0

8 Future directions

The current exploratory data analysis module can be expanded by investigating samples with paired assays. Moreover, if other metadata are available, e.g., disease_stage, treatment, this will build large, longitudinal cohorts with multi-omic sequencing data. Such an analysis permits consideration of samples according to the condition(s) of the experiment and research aims. In addition, it can be used to refine research questions and/or generate new ones.

This will facilitate collaboration across departments at St. Jude, expedite discoveries, and find cures for children with cancer and other catastrophic diseases.

9 Session Info

```
## R version 4.4.0 (2024-04-24)
## Platform: x86_64-pc-linux-gnu
## Running under: Red Hat Enterprise Linux 8.8 (Ootpa)
##
## Matrix products: default
## BLAS:   /research/rgs01/applications/hpcf/authorized_apps/rhel8_apps/lapack/3.10.1/install/lib64/libl
## LAPACK: /research/rgs01/applications/hpcf/authorized_apps/rhel8_apps/lapack/3.10.1/install/lib64/libl
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Chicago
## tzcode source: system (glibc)
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] ggthemes_5.1.0  lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] sass_0.4.9        utf8_1.2.4        generics_0.1.3    stringi_1.8.4
##  [5] hms_1.1.3         digest_0.6.37     magrittr_2.0.3    evaluate_0.24.0
##  [9] timechange_0.3.0  fastmap_1.2.0     jsonlite_1.8.8    tinytex_0.52
## [13] fansi_1.0.6       scales_1.3.0      jquerylib_0.1.4   cli_3.6.3
## [17] rlang_1.1.4       crayon_1.5.3      bit64_4.0.5       munsell_0.5.1
## [21] withr_3.0.1       cachem_1.1.0      yaml_2.3.10       tools_4.4.0
## [25] parallel_4.4.0    tzdb_0.4.0        colorspace_2.1-1  vctrs_0.6.5
## [29] R6_2.5.1          mime_0.12         lifecycle_1.0.4   bit_4.0.5
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      bslib_0.8.0
## [37] gtable_0.3.5      glue_1.7.0        highr_0.11        xfun_0.47
## [41] tidyselect_1.2.1  knitr_1.48        farver_2.1.2      htmltools_0.5.8.1
## [45] rmarkdown_2.28    labeling_0.4.3    compiler_4.4.0
```