A complete analysis of peptide microarray binding data using the pepStat framework

Greg Imholte[*], Renan Sauteraud[†], Mike Jiang[‡] and Raphael Gottardo[§]

October 24, 2023

This document present a full analysis, from reading the data to displaying the results that makes use of all the packages we developped for peptide microarray.

Contents

[*]gimholte@uw.edu
[†]rsautera@fhcrc.org
[‡]wjiang2@fhcrc.org
[§]rgottard@fhcrc.org

1 Introduction

The pepStat package offers a complete analytical framework for the analysis of peptide microarray data. It includes a novel normalization method to remove non-specific peptide binding activity of antibodies, a data smoothing reducing step to reduce background noise, and subject-specific positivity calls.

1.1 Requirements

The pepStat package requires GSL, an open source scientific computing library. This library is freely available at http://www.gnu.org/software/gsl/.

In this vignette, we make use of the samples and examples available in the data package pepDat.

2 Generating a peptideSet

```
library(pepDat)
library(pepStat)
```

2.1 Reading in .gpr files

The reading function, makePeptideSet, takes a path as its argument and parses all the *.gpr* files in the given directory. Alternatively, one may specify a character vector of paths to individual *.gpr* files.

By default channels F635 Median and B635 Median are collected, and the 'normexp' method of the backgroundCorrect function in the limma package corrects probe intensities for background fluorescence. Other methods may be selected, see documentation.

```
mapFile <- system.file("extdata/mapping.csv", package = "pepDat")
dirToParse <- system.file("extdata/gpr_samples", package = "pepDat")
pSet <- makePeptideSet(files = NULL, path = dirToParse,
                       mapping.file = mapFile, log=TRUE)
```

While optional, it is strongly recommended to provide a mapping.file giving annotations data for each slide, such as treatment status or patient information. If provided, the mapping.file should be a .csv file. It must include columns labeled filename, ptid, and visit. Elements in column filename must correspond to the filenames of slides to be read in, without the .gpr extension. Column ptid is a subject or slide identifier. Column visit indicates a case or control condition, such as pre/post vaccination, pre/post infection, or healthy/infected status. Control conditions must be labelled *pre*, while case conditions must be labelled *post*. Alternatively, one may input a data.frame satisfying the same requirements.

This minimal information is required by pepStat's functions further in the analysis. Any additional information (column) will be retained and can be used as a grouping variable.

If no mapping file is included, the information will have to be added later on to the peptideSet object.

For our example, we use a toy dataset of 8 samples from 4 patients and we are interested in comparing the antibody binding in placebo versus vaccinated subjects.

```
read.csv(mapFile)

##   filename ptid visit treatment
## 1     f1_1    1   Pre   PLACEBO
## 2     f1_2    1  Post   PLACEBO
## 3     f2_1    2   Pre   PLACEBO
## 4     f2_2    2  Post   PLACEBO
## 5     f3_1    3   Pre   VACCINE
## 6     f3_2    3  Post   VACCINE
## 7     f4_1    4   Pre   VACCINE
## 8     f4_2    4  Post   VACCINE
```

2.2 Additional arguments

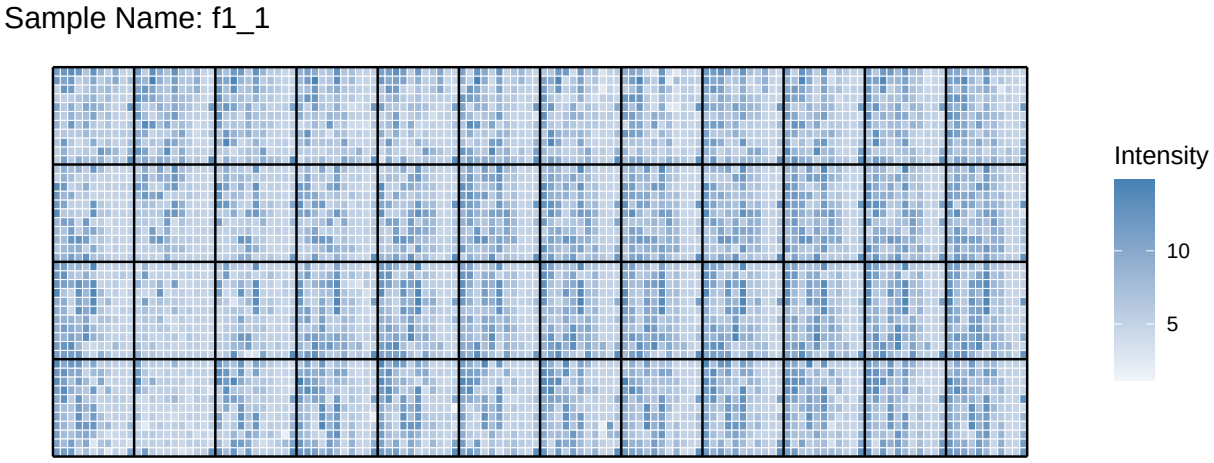
The empty spots should be listed in order to background correct the intensities. It is also useful to remove the controls when reading the data. Here we have the JPT controls, human Ig (A, E and M) and dye controls.

```
pSetNoCtrl <- makePeptideSet(files = NULL, path = dirToParse,
                        mapping.file = mapFile, log = TRUE,
                        rm.control.list = c("JPT-control", "Ig", "Cy3"),
                        empty.control.list= c("empty", "blank control"))
```

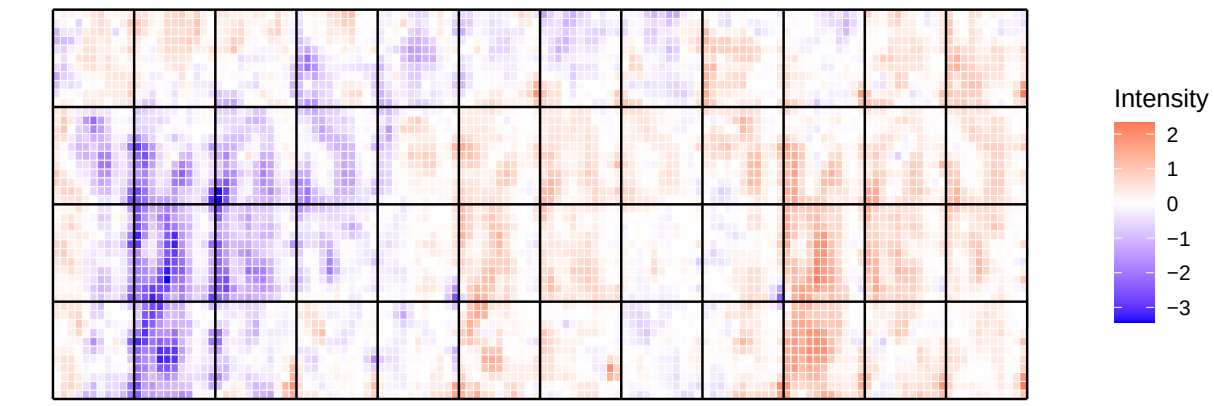
2.3 Visualize slides

We include two plotting functions to detect possible spatial slide artifacts. Since the full plate is needed for this visualization, the functions will work best with rm.contol.list and empty.control.list set to NULL in makePeptideSet.

```
plotArrayImage(pSet, array.index = 1)
```




```
plotArrayResiduals(pSet, array.index = 1, smooth = TRUE)
```



3 Adding peptide informations

At this point, the peptideSet contain only the peptide sequences and the associated background corrected intensities. To continue with the analysis, we need to add the position information, as well as physicochemical properties of the peptides summarized by their z-scales.

The slides used in this example are the enveloppe of HIV-1 and peptide collections are available for this in our pepDat package (please refere to the vignette and `?pep_hxb2` for more information). However, we will pretend that this is not the case to show an example of how to build a custom peptide collection.

3.1 Creating a custom peptide collection

Here, we load a data.frame that contains the peptides used on the array as well as their start and end coordinates, and clade information.

```
peps <- read.csv(system.file("extdata/pep_info.csv", package = "pepDat"))
head(peps)

##   start end         peptide clade
## 1     1  16 MRVKETQMNWPNLWK CRF01
## 2     1  16 MRVMGIQKNYPLLWR CRF02
## 3     1  16 MRVMGIQRNCQHLWR     A
## 4     1  16 MRVKGIRKNYQHLWR     B
## 5     1  16 MRVRGILRNWQQWWI     C
## 6     1  16 MRVRGIERNYQHLWR     D
```

Then we call the constructor that will create the appropriate collection.

```
pep_custom <- create_db(peps)
```

pep_custom is a `GRanges` object with the required "peptide" metadata column and the physiochemical properties of each peptide sequence summarized by z-scores.

Note that the function will also accept `GRanges` input.

```
pep_custom <- create_db(pep_custom)
```

3.2 Summarize the information

The function `summarizePeptides` summarizes within-slide replicates by either their mean or median. Additionaly, with the newly constructed peptide collection, peptides positions and annotations can be passed on to the existing peptideSet. Alternately, the function could be callled directly on the `data.frame` object. Internally, `summarizePeptides` will call `create_db` to make sure the input is formatte appropriately.

```
psSet <- summarizePeptides(pSet, summary = "mean", position = pep_custom)

## Some peptides have no match in the GRanges object rownames and are removed from
the peptideSet!
```

Now that all the required information is available, we can proceed with the analysis.

4 Normalization

The primary goal of the data normalization step is to remove non-biological source of bias and increase the comparability of true positive signal intensities across slides. The method developped for this package uses physiochemical properties of individual peptides to model non-specific antibody binding to arrays.

```
pnSet <- normalizeArray(psSet)
```

An object of class `peptideSet` containing the corrected peptides intensities is returned.

5 Data smoothing

The optional data smoothing step takes advantage of the overlapping nature of the peptides on the array to remove background noise caused by experimental variation. It is likely that two overlapping peptides will share common binding signal, when present. `pepStat` use a sliding mean technique technique to borrow strength across neighboring peptides and to reduce signal variability. This statistic increases detection of binding *hotspots* that noisy signals might otherwise obscure. Peptides are smoothed according to their sequence alignment position, taken from `position(psSet)`.

From here on, two types of analyses are possible. The peptides can be aggregated by position or split by clade. When aggregating by position, the sliding mean will get information from surrounding peptides as well as peptides located around their coordinates in other clades. This increase the strength of calls but the clade specificity is lost.

It is common to do a first run with aggregated clades to detect binding hotspots and then do a second run to look for clade specificity in the peaks found during the first run.

This is decided by the `split.by.clade` argument. By default it is set to TRUE for a clade specific analysis.

```
psmSet <- slidingMean(pnSet, width = 9)
```

For the aggregated `peptideSet` we set it to FALSE.

```
psmSetAg <- slidingMean(pnSet, width = 9, split.by.clade = FALSE)
```

6 Making calls

The final step is to make the positivity calls. The function `makeCalls` automatically uses information provided in the mapping file, accessed via `pData(pSet)`. It detects whether samples are paired or not. If samples are paired, POST intensities are subtracted from PRE intensities, then thresholded. Otherwise, PRE samples are averaged, and then subtracted from POST intensities. These corrected POST intensities are thresholded.

The `freq` argument controls whether we return the percentage of responders against each peptide, or a matrix of subject specific call. When `freq` is `TRUE`, we may supply a `group` variable from `pData(psmSet)` on which we split the frequency calculation.

```
calls <- makeCalls(psmSet, freq = TRUE, group = "treatment",
                   cutoff = .1, method = "FDR", verbose = TRUE)

## You have paired PRE/POST samples

## The selected threshold T is  1.100119
```

The function automatically selected an appropriate FDR threshold.

```
callsAg <- makeCalls(psmSetAg, freq = TRUE, group = "treatment",
                     cutoff = .1, method = "FDR")
```

7 Results

7.1 summary

To get a summary of the analysis, for each peptide, the package provides the function `restab` that combines a `peptideSet` and the result of `makeCalls` into a single `data.frame` with one row per peptide and per clade.

```
summary <- restab(psmSet, calls)
head(summary)
```

```
##                              peptide position start end width clade PLACEBO
## MRVKETQMNWPNLWK_CRF01 MRVKETQMNWPNLWK        8     1  16    16 CRF01       0
## MRVKGIRKNYQHLWR_B     MRVKGIRKNYQHLWR        8     1  16    16     B       0
## MRVMGIQKNYPLLWR_CRF02 MRVMGIQKNYPLLWR        8     1  16    16 CRF02       0
## MRVMGIQRNCQHLWR_A     MRVMGIQRNCQHLWR        8     1  16    16     A       0
## MRVMGIQRNWQHLWR_M     MRVMGIQRNWQHLWR        8     1  16    16     M       0
## MRVRGIERNYQHLWR_D     MRVRGIERNYQHLWR        8     1  16    16     D     100
##                       VACCINE
## MRVKETQMNWPNLWK_CRF01       0
## MRVKGIRKNYQHLWR_B           0
## MRVMGIQKNYPLLWR_CRF02       0
## MRVMGIQRNCQHLWR_A           0
## MRVMGIQRNWQHLWR_M           0
## MRVRGIERNYQHLWR_D           0
```

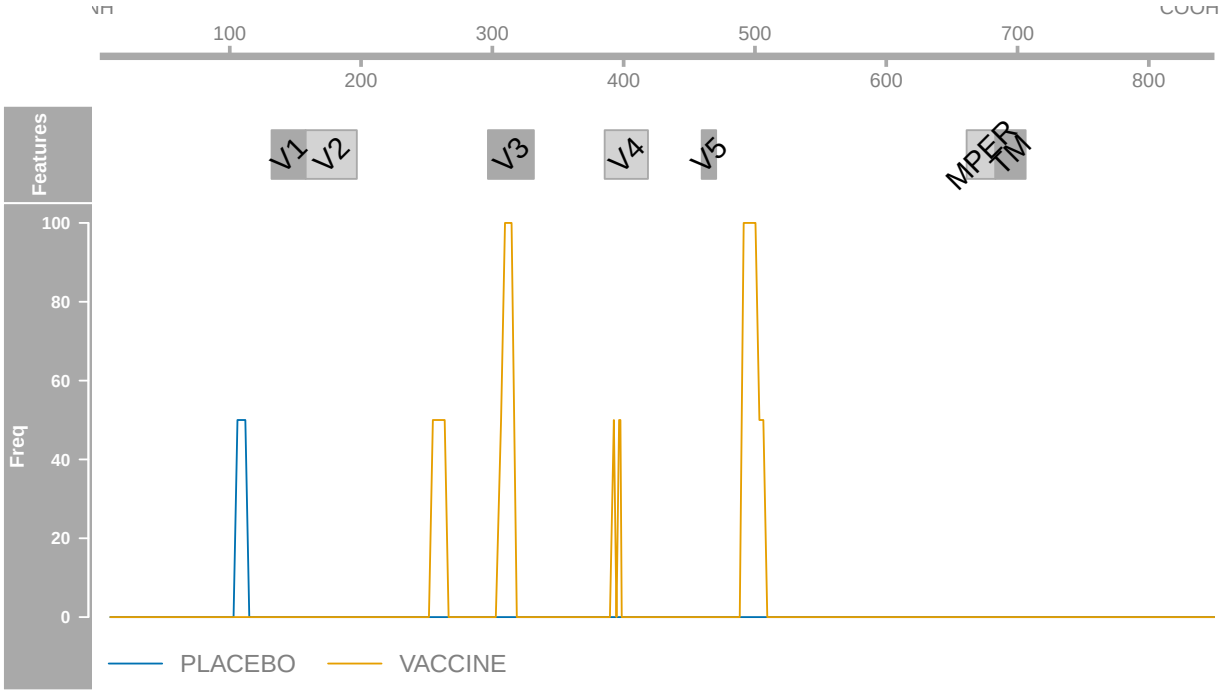
Note that if calls are made with a `peptideSet` that has been normalized with `split.by.clade` set to FALSE, the table will have one row per peptide. Peptides that are identical accross clades will only have one entry.

7.2 Plots

As part of the pipeline for the analysis of peptide microarray data, the `Pviz` package includes a track that can use the result of an experiment to generate plots.

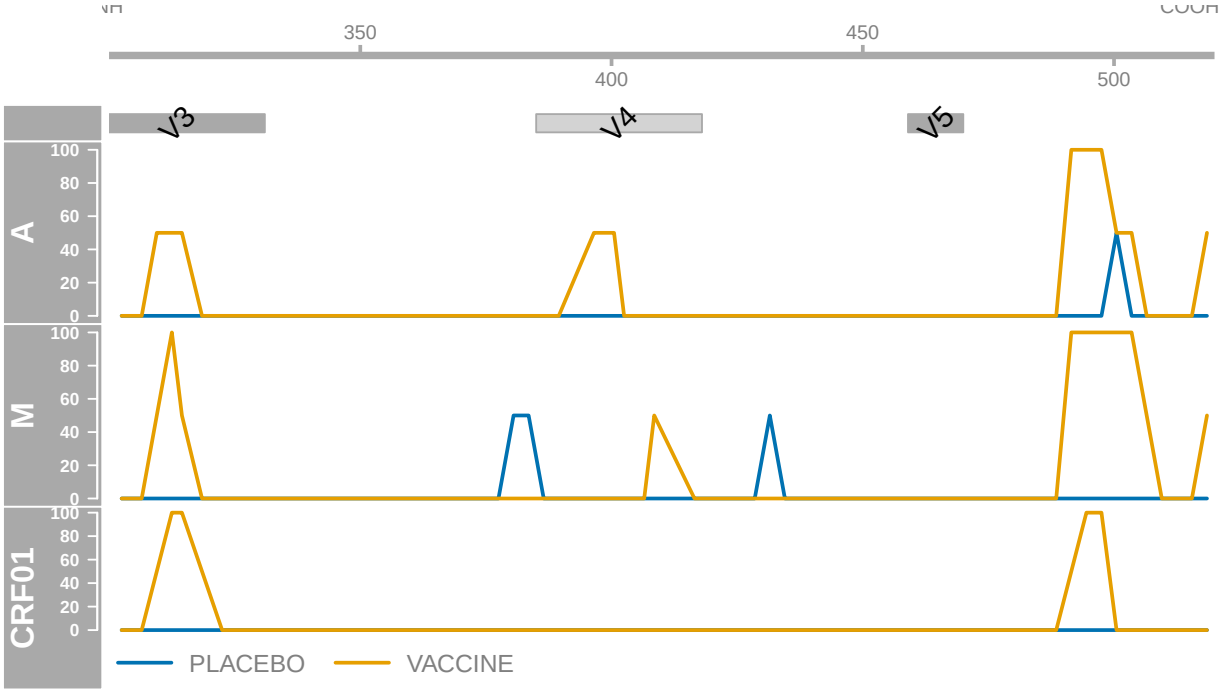
When analysing all clades at once, the `plot_inter` function can be used to easily identify binding peaks. It gives an overview of the differences between the selected groups. In this case, comparing placebo and vaccine.

```
library(Pviz)
summaryAg <- restab(psmSetAg, callsAg)
plot_inter(summaryAg)
```



When clade specific calls have been made, it is more interesting to plot each clade on a separate track.

```
plot_clade(summary, clade=c("A", "M", "CRF01"), from = 300, to = 520)
```



Much more complex plots can be made, custom tracks can be added and every graphical parameter can be tweaked. Refer to the `Pviz` documentation as well as the `Gviz` package for detailed information on all tracks and display paramters.

8 shinyApp

As part of the package, a shinyApp provides a user interface for peptide microarray analysis. After making the calls, the results can be downloaded and the app displays plots as shown in the previous sections.

The app can be started from the command line using the `shinyPepStat` function.

```
shinyPepStat()
```

9 Quick analysis

Here we showcase a quick analysis of peptide microarray data for HIV-1 gp160. This displays the minimal amount of code required to go from raw data file to antibody binding positivity call.

```
library(pepStat)
library(pepDat)
mapFile <- system.file("extdata/mapping.csv", package = "pepDat")
dirToParse <- system.file("extdata/gpr_samples", package = "pepDat")
ps <- makePeptideSet(files = NULL, path = dirToParse, mapping.file = mapFile)
data(pep_hxb2)
ps <- summarizePeptides(ps, summary = "mean", position = pep_hxb2)
ps <- normalizeArray(ps)
ps <- slidingMean(ps)
calls <- makeCalls(ps, group = "treatment")
summary <- restab(ps, calls)
```

10 sessionInfo

```
sessionInfo()

## R version 4.3.1 (2023-06-16)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.3 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.18-bioc/R/lib/libRblas.so
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB              LC_COLLATE=C
##  [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/New_York
## tzcode source: system (glibc)
##
## attached base packages:
## [1] grid      stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] Pviz_1.36.0          Gviz_1.46.0          GenomicRanges_1.54.0
##  [4] pepStat_1.36.0       IRanges_2.36.0       S4Vectors_0.40.0
##  [7] Biobase_2.62.0       BiocGenerics_0.48.0  pepDat_1.21.0
## [10] knitr_1.44
##
## loaded via a namespace (and not attached):
##   [1] RColorBrewer_1.1-3          rstudioapi_0.15.0
##   [3] magrittr_2.0.3              GenomicFeatures_1.54.0
##   [5] farver_2.1.1                rmarkdown_2.25
##   [7] BiocIO_1.12.0               fields_15.2
##   [9] zlibbioc_1.48.0             vctrs_0.6.4
##  [11] memoise_2.0.1               Rsamtools_2.18.0
##  [13] RCurl_1.98-1.12             base64enc_0.1-3
##  [15] htmltools_0.5.6.1           S4Arrays_1.2.0
##  [17] progress_1.2.2              curl_5.1.0
##  [19] SparseArray_1.2.0           Formula_1.2-5
##  [21] htmlwidgets_1.6.2           plyr_1.8.9
```

```
##  [23] cachem_1.0.8               GenomicAlignments_1.38.0
##  [25] lifecycle_1.0.3            pkgconfig_2.0.3
##  [27] Matrix_1.6-1.1             R6_2.5.1
##  [29] fastmap_1.1.1              GenomeInfoDbData_1.2.11
##  [31] MatrixGenerics_1.14.0      digest_0.6.33
##  [33] colorspace_2.1-0           AnnotationDbi_1.64.0
##  [35] Hmisc_5.1-1                RSQLite_2.3.1
##  [37] filelock_1.0.2             labeling_0.4.3
##  [39] fansi_1.0.5                httr_1.4.7
##  [41] abind_1.4-5                compiler_4.3.1
##  [43] bit64_4.0.5                withr_2.5.1
##  [45] htmlTable_2.4.1            backports_1.4.1
##  [47] BiocParallel_1.36.0        DBI_1.1.3
##  [49] highr_0.10                 maps_3.4.1
##  [51] biomaRt_2.58.0             rappdirs_0.3.3
##  [53] DelayedArray_0.28.0        rjson_0.2.21
##  [55] tools_4.3.1                foreign_0.8-85
##  [57] nnet_7.3-19                glue_1.6.2
##  [59] restfulr_0.0.15            checkmate_2.2.0
##  [61] cluster_2.1.4              generics_0.1.3
##  [63] gtable_0.3.4               BSgenome_1.70.0
##  [65] ensembldb_2.26.0           data.table_1.14.8
##  [67] hms_1.1.3                  xml2_1.3.5
##  [69] utf8_1.2.4                 XVector_0.42.0
##  [71] pillar_1.9.0               stringr_1.5.0
##  [73] spam_2.10-0                limma_3.58.0
##  [75] dplyr_1.1.3                BiocFileCache_2.10.0
##  [77] lattice_0.22-5             deldir_1.0-9
##  [79] rtracklayer_1.62.0         bit_4.0.5
##  [81] biovizBase_1.50.0          tidyselect_1.2.0
##  [83] Biostrings_2.70.0          gridExtra_2.3
##  [85] ProtGenerics_1.34.0        SummarizedExperiment_1.32.0
##  [87] xfun_0.40                  statmod_1.5.0
##  [89] matrixStats_1.0.0          stringi_1.7.12
##  [91] lazyeval_0.2.2             yaml_2.3.7
##  [93] evaluate_0.22              codetools_0.2-19
##  [95] interp_1.1-4               tibble_3.2.1
##  [97] cli_3.6.1                  rpart_4.1.21
##  [99] munsell_0.5.0              dichromat_2.0-0.1
## [101] Rcpp_1.0.11                GenomeInfoDb_1.38.0
## [103] dbplyr_2.3.4               png_0.1-8
## [105] XML_3.99-0.14              parallel_4.3.1
## [107] ggplot2_3.4.4              blob_1.2.4
## [109] prettyunits_1.2.0          dotCall64_1.1-0
## [111] jpeg_0.1-10                latticeExtra_0.6-30
```

```
## [113] AnnotationFilter_1.26.0     bitops_1.0-7
## [115] viridisLite_0.4.2           VariantAnnotation_1.48.0
## [117] scales_1.2.1                crayon_1.5.2
## [119] rlang_1.1.1                 KEGGREST_1.42.0
```