main

August 4, 2021

1 Plot boxplots

[1]:
```
#library(repr)
library(ggpubr)
library(ggsignif)
library(tidyverse)
```

```
Loading required package: ggplot2

  Attaching packages                                    tidyverse
1.3.1

  tibble  3.1.2      dplyr   1.0.7
  tidyr   1.1.3      stringr 1.4.0
  readr   1.4.0      forcats 0.5.1
  purrr   0.3.4

  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
```

1.1 Functions

[2]:
```
add_symnum <- function(res){
    symnum.args <- list(cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05,  1),
                        symbols = c("****", "***", "**", "*",  "ns"))
    symnum.args$x <- res$adj.P.Val
    pvalue.signif <- do.call(stats::symnum, symnum.args) %>%
    as.character()
    pvalue.format <- format.pval(res$adj.P.Val, digits = 2)
    res <- res %>%
    dplyr::ungroup() %>%
    mutate(FDR = pvalue.format, p.signif = pvalue.signif)
    return(res)
}
```

```
save_ggplots <- function(fn, p, w, h){
    for(ext in c('.pdf', '.png', '.svg')){
        ggsave(paste0(fn, ext), plot=p, width=w, height=h)
    }
}
```

1.2 Gene annotation

1.2.1 Caudate vs DLPFC

[3]:
```
deg_file = '../../_m/genes/diffExpr_CvD_sex_full.txt'
deg_cd = read.delim(deg_file, row.names=1) %>% filter(Symbol %in% c("ANK3",␣
↪"EML1", "PKIG")) %>%
    select(-c(Length, Class, meanExprs, NumTx, gene_type, gencodeTx)) %>%␣
↪mutate(Comparison="C_vs_D")
deg_cd = add_symnum(deg_cd)
deg_cd
```

A data.frame: 3 × 13

	gencodeID <chr>	ensemblID <chr>	Symbol <chr>	EntrezID <int>
ENSG00000066629.16	ENSG00000066629.16	ENSG00000066629	EML1	2009
ENSG00000151150.21	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000168734.13	ENSG00000168734.13	ENSG00000168734	PKIG	11142

1.2.2 Caudate vs Hippocampus

[4]:
```
deg_file = '../../_m/genes/diffExpr_CvH_sex_full.txt'
deg_ch = read.delim(deg_file, row.names=1) %>% filter(Symbol %in% c("ANK3",␣
↪"EML1", "PKIG")) %>%
    select(-c(Length, Class, meanExprs, NumTx, gene_type, gencodeTx)) %>%␣
↪mutate(Comparison="C_vs_H")
deg_ch = add_symnum(deg_ch)
deg_ch
```

A data.frame: 3 × 13

	gencodeID <chr>	ensemblID <chr>	Symbol <chr>	EntrezID <int>
ENSG00000066629.16	ENSG00000066629.16	ENSG00000066629	EML1	2009
ENSG00000151150.21	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000168734.13	ENSG00000168734.13	ENSG00000168734	PKIG	11142

1.2.3 DLPFC vs Hippocampus

[5]:
```
deg_file = '../../_m/genes/diffExpr_DvH_sex_full.txt'
deg_dh = read.delim(deg_file, row.names=1) %>% filter(Symbol %in% c("ANK3",␣
↪"EML1", "PKIG")) %>%
    select(-c(Length, Class, meanExprs, NumTx, gene_type, gencodeTx)) %>%␣
↪mutate(Comparison="D_vs_H")
```

```
deg_dh = add_symnum(deg_dh)
deg_dh
```

A data.frame: 3 × 13

	gencodeID <chr>	ensemblID <chr>	Symbol <chr>	EntrezID <int>
ENSG00000168734.13	ENSG00000168734.13	ENSG00000168734	PKIG	11142
ENSG00000151150.21	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000066629.16	ENSG00000066629.16	ENSG00000066629	EML1	2009

1.2.4 Merge gene annotation

[6]:
```
gene_annot = rbind(deg_cd, deg_ch, deg_dh)
gene_annot
```

A data.frame: 9 × 13

	gencodeID <chr>	ensemblID <chr>	Symbol <chr>	EntrezID <int>
ENSG00000066629.16	ENSG00000066629.16	ENSG00000066629	EML1	2009
ENSG00000151150.21	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000168734.13	ENSG00000168734.13	ENSG00000168734	PKIG	11142
ENSG00000066629.161	ENSG00000066629.16	ENSG00000066629	EML1	2009
ENSG00000151150.211	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000168734.131	ENSG00000168734.13	ENSG00000168734	PKIG	11142
ENSG00000168734.132	ENSG00000168734.13	ENSG00000168734	PKIG	11142
ENSG00000151150.212	ENSG00000151150.21	ENSG00000151150	ANK3	288
ENSG00000066629.162	ENSG00000066629.16	ENSG00000066629	EML1	2009

1.3 Load residualized data

[7]:
```
res_file = '../../_m/genes/residualized_expression.tsv'
resdf0 = data.table::fread(res_file) %>%
    filter(Geneid %in% gene_annot$gencodeID) %>%
    column_to_rownames("Geneid") %>% t %>% data.frame
resdf0 %>% head(2)
```

A data.frame: 2 × 3

	ENSG00000151150.21 <dbl>	ENSG00000066629.16 <dbl>	ENSG00000168734.13 <dbl>
R10424	0.1373256	-0.5566796	0.52761320
R12195	0.0542371	-0.3045881	0.07979308

1.4 Load phenotype data

[8]:
```
pheno_file = '/ceph/projects/v4_phase3_paper/inputs/phenotypes/_m/
 ↪merged_phenotypes.csv'
pheno = read.csv(pheno_file, row.names=1) %>%
    mutate_if(is.character, as.factor)
head(pheno, 2)
```

A data.frame: 2 × 20

	Sex <fct>	Race <fct>	Dx <fct>	Age <dbl>	mitoRate <dbl>	rRNA_rate <dbl>	totalAssignedGen <dbl>
R11135	Male	EA	CTL	18.77	0.2572796	0.0001690954	0.5231321
R11137	Male	EA	CTL	41.44	0.3840272	0.0000884558	0.5933431

1.5 Merge dataframe

[9]:
```
resdf <- inner_join(rownames_to_column(pheno),
                    rownames_to_column(resdf0),
                    by="rowname")
dim(resdf)
resdf[1:2, 1:11]
```

1. 1127 2. 24

A data.frame: 2 × 11

	rowname <chr>	Sex <fct>	Race <fct>	Dx <fct>	Age <dbl>	mitoRate <dbl>	rRNA_rate <dbl>	totalAssigne <dbl>
1	R11135	Male	EA	CTL	18.77	0.2572796	0.0001690954	0.5231321
2	R11137	Male	EA	CTL	41.44	0.3840272	0.0000884558	0.5933431

1.6 Melt data frame

[10]:
```
df = resdf %>% select(c('rowname', 'Sex', "Region", starts_with('ENSG'))) %>%
    pivot_longer(-c(rowname, Sex, Region), names_to = "gencodeID", values_to =␣
↪"Res") %>%
    inner_join(gene_annot, by='gencodeID') %>%
    select(rowname, Sex, Region, gencodeID, Res, Symbol) %>% distinct %>%
    mutate_at(vars("Symbol", "Region", "gencodeID", "Sex"), as.factor)
levels(df$Sex) <- c("Female", "Male")
levels(df$Region) <- c("Caudate", "DLPFC", "Hippocampus")
head(df, 3)
```

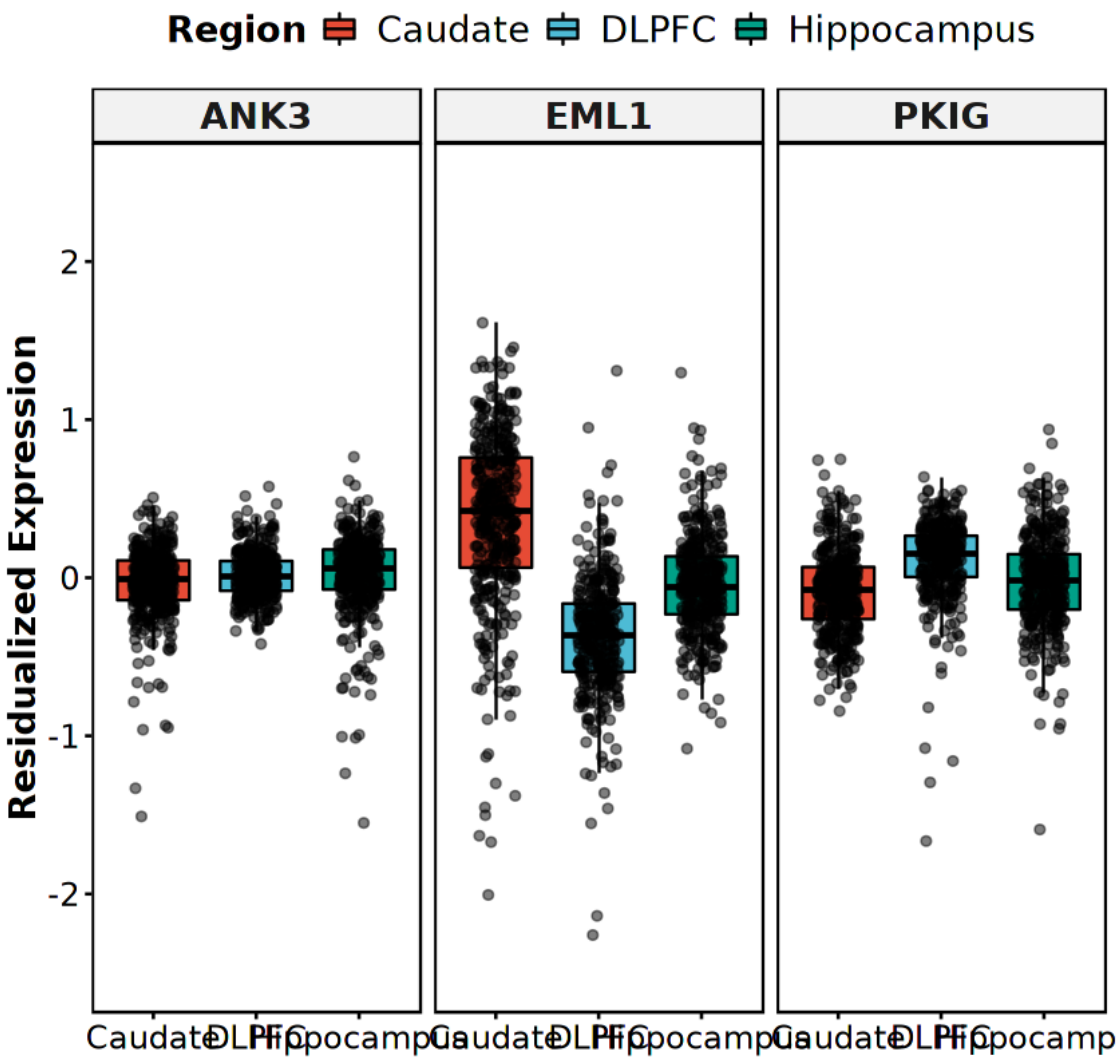
A tibble: 3 × 6

rowname <chr>	Sex <fct>	Region <fct>	gencodeID <fct>	Res <dbl>	Symbol <fct>
R11135	Male	Hippocampus	ENSG00000151150.21	-0.1901582	ANK3
R11135	Male	Hippocampus	ENSG00000066629.16	-0.2084132	EML1
R11135	Male	Hippocampus	ENSG00000168734.13	-0.5161214	PKIG

1.6.1 Initial ggplot with ggpubr

[11]:
```
tmp = gene_annot %>% mutate(group1='Female', group2='Male', y_pos=3)

bxp_r <- ggboxplot(df, x="Region", y="Res", facet.by=c("Symbol"), ncol=4,
                   fill="Region", xlab='', palette="npg", outlier.shape=NA,
                   panel.labs.font=list(face='bold', size = 16),
                   ylab='Residualized Expression', add='jitter', ylim=c(-2.5, 2.
↪5),
                  add.params=list(alpha=0.5)) +
        font("xy.title", size=16, face="bold") + font("xy.text", size=14) +
```

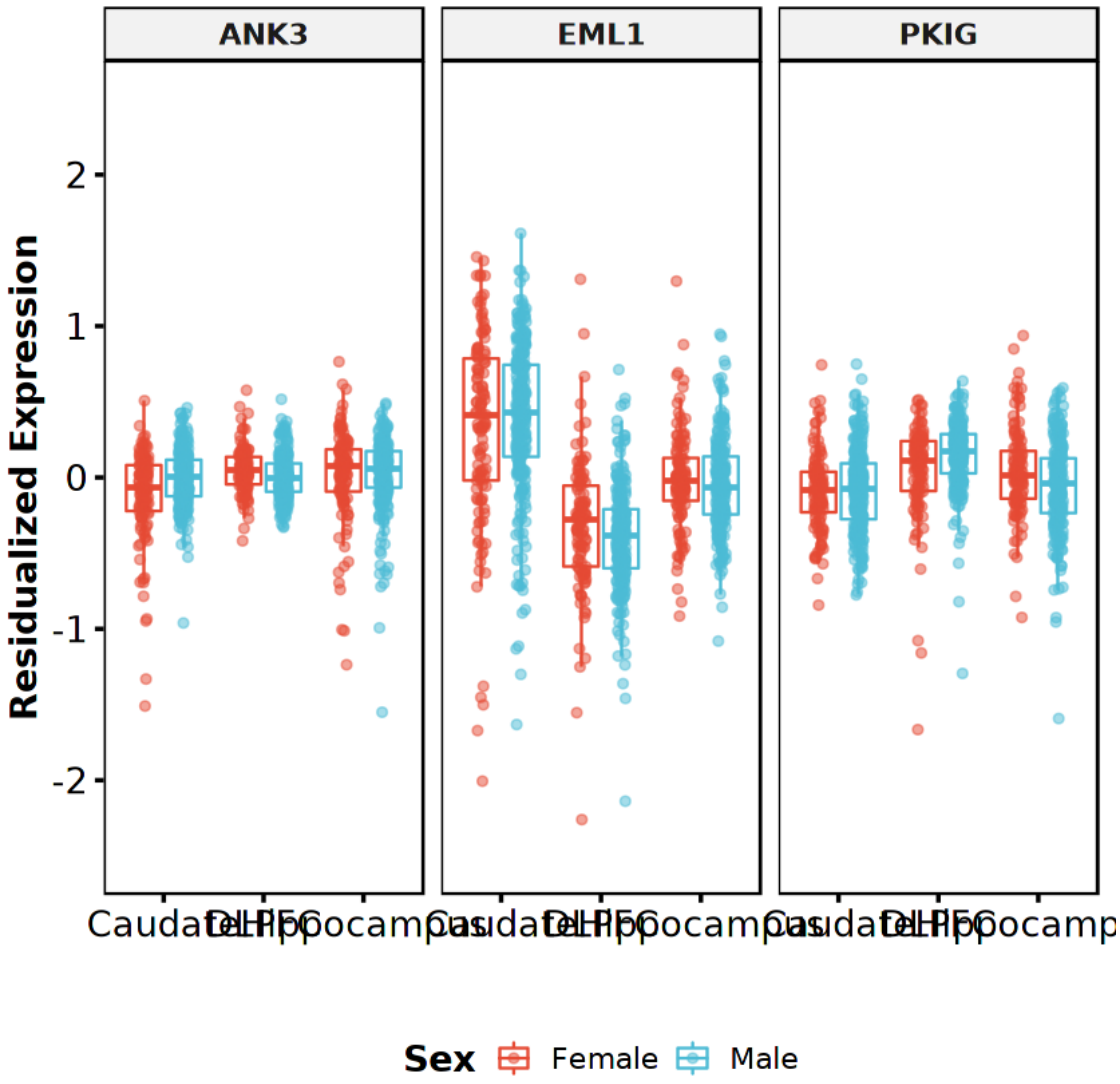
```
        font("legend.title", size=16, face="bold") +
        font("legend.text", size=16)

bxp_r
```



[12]:
```
save_ggplots('region_interaction_sex_byRegion', bxp_r, 12, 5)
```

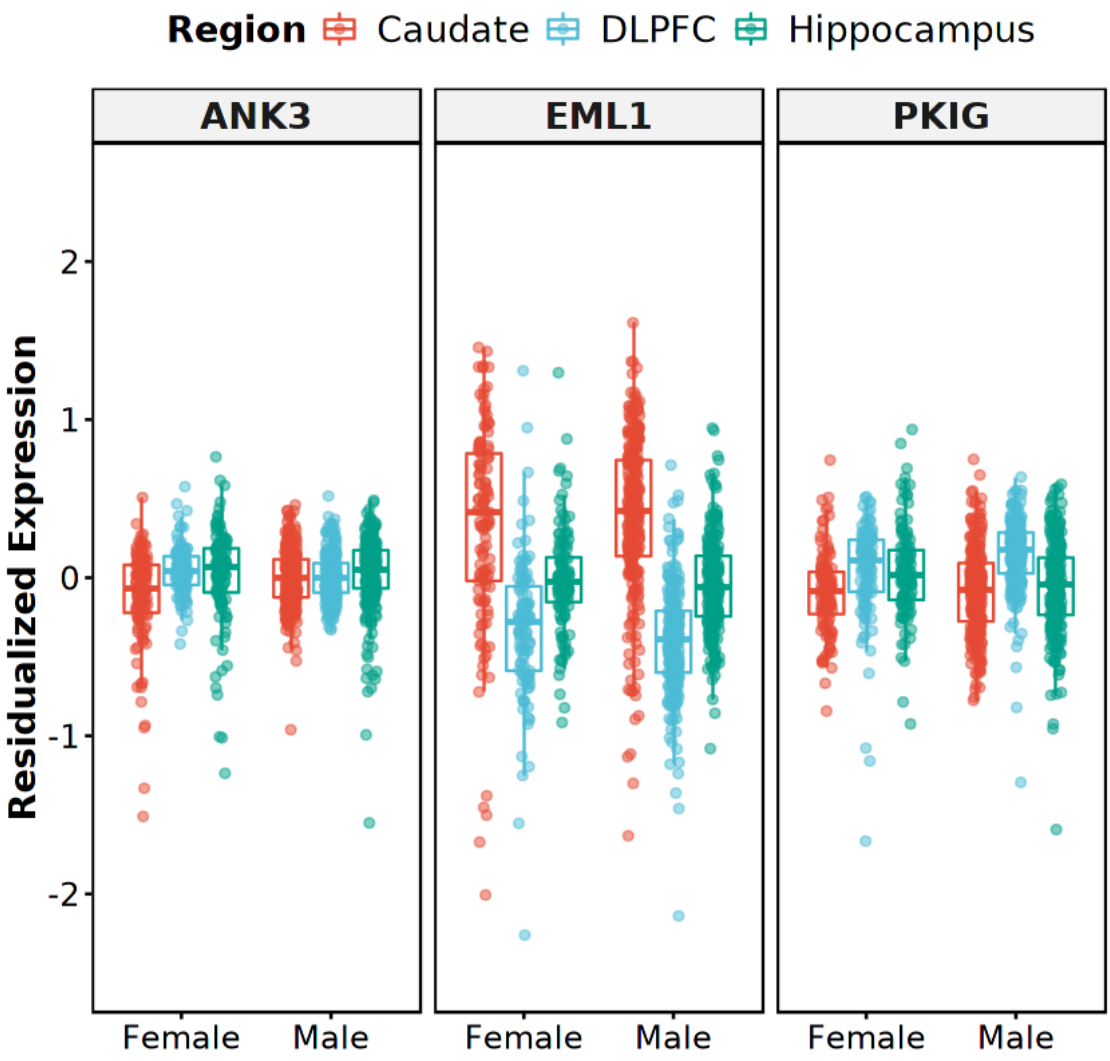
[13]:
```
tmp = gene_annot %>% mutate(group1='Female', group2='Male', y_pos=3)
bxp_x <- ggboxplot(df, x="Region", y="Res", facet.by=c("Symbol"), ncol=3,
                   color="Sex", xlab='', palette="npg", outlier.shape=NA,
                   panel.labs.font=list(face='bold'), add='jitter',
                   ylab='Residualized Expression', ylim=c(-2.5, 2.5),
                   add.params=list(alpha=0.5), legend="bottom",
```

```
                    ggtheme=theme_pubr(base_size=16, border=TRUE)) +
    font("xy.title", face="bold") + font("legend.title", face="bold")
save_ggplots('region_interaction_sex_bySex', bxp_x, 12, 5)
bxp_x
```



[14]:
```
tmp = gene_annot %>% mutate(group1='Female', group2='Male', y_pos=3)

bxp <- ggboxplot(df, x="Sex", y="Res", facet.by=c("Symbol"), ncol=3,
                 color="Region", xlab='', palette="npg", outlier.shape=NA,
                 panel.labs.font=list(face='bold', size = 16),
                 ylab='Residualized Expression', add='jitter', ylim=c(-2.5, 2.
 ↪5),
```

```
                  add.params=list(alpha=0.5)) +
          font("xy.title", size=16, face="bold") + font("xy.text", size=14) +
          font("legend.title", size=16, face="bold") +
          font("legend.text", size=16)

bxp
```



[15]:
```
save_ggplots('region_interaction_sex', bxp, 12, 5)
```

1.7 Reproducibility information

[16]:
```
print("Reproducibility Information:")
Sys.time()
proc.time()
options(width=120)
sessioninfo::session_info()
```

```
[1] "Reproducibility Information:"
```

```
[1] "2021-08-04 14:58:15 EDT"
```

```
   user  system elapsed
 17.001   3.880  16.153
```

```
 Session info
 setting  value
 version  R version 4.0.3 (2020-10-10)
 os       Arch Linux
 system   x86_64, linux-gnu
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/New_York
 date     2021-08-04

 Packages
 package     * version  date       lib source
 abind         1.4-5    2016-07-21 [1] CRAN (R 4.0.2)
 assertthat    0.2.1    2019-03-21 [1] CRAN (R 4.0.2)
 backports     1.2.1    2020-12-09 [1] CRAN (R 4.0.2)
 base64enc     0.1-3    2015-07-28 [1] CRAN (R 4.0.2)
 broom         0.7.8    2021-06-24 [1] CRAN (R 4.0.3)
 Cairo         1.5-12.2 2020-07-07 [1] CRAN (R 4.0.2)
 car           3.0-11   2021-06-27 [1] CRAN (R 4.0.3)
 carData       3.0-4    2020-05-22 [1] CRAN (R 4.0.2)
 cellranger    1.1.0    2016-07-27 [1] CRAN (R 4.0.2)
 cli           3.0.0    2021-06-30 [1] CRAN (R 4.0.3)
 colorspace    2.0-2    2021-06-24 [1] CRAN (R 4.0.3)
 crayon        1.4.1    2021-02-08 [1] CRAN (R 4.0.3)
 curl          4.3.2    2021-06-23 [1] CRAN (R 4.0.3)
 data.table    1.14.0   2021-02-21 [1] CRAN (R 4.0.3)
 DBI           1.1.1    2021-01-15 [1] CRAN (R 4.0.2)
 dbplyr        2.1.1    2021-04-06 [1] CRAN (R 4.0.3)
 digest        0.6.27   2020-10-24 [1] CRAN (R 4.0.2)
 dplyr       * 1.0.7    2021-06-18 [1] CRAN (R 4.0.3)
 ellipsis      0.3.2    2021-04-29 [1] CRAN (R 4.0.3)
 evaluate      0.14     2019-05-28 [1] CRAN (R 4.0.2)
 fansi         0.5.0    2021-05-25 [1] CRAN (R 4.0.3)
```

```
farver        2.1.0    2021-02-28 [1] CRAN (R 4.0.3)
forcats     * 0.5.1    2021-01-27 [1] CRAN (R 4.0.2)
foreign       0.8-80   2020-05-24 [2] CRAN (R 4.0.3)
fs            1.5.0    2020-07-31 [1] CRAN (R 4.0.2)
generics      0.1.0    2020-10-31 [1] CRAN (R 4.0.2)
ggplot2     * 3.3.5    2021-06-25 [1] CRAN (R 4.0.3)
ggpubr      * 0.4.0    2020-06-27 [1] CRAN (R 4.0.2)
ggsci         2.9      2018-05-14 [1] CRAN (R 4.0.2)
ggsignif    * 0.6.2    2021-06-14 [1] CRAN (R 4.0.3)
glue          1.4.2    2020-08-27 [1] CRAN (R 4.0.2)
gtable        0.3.0    2019-03-25 [1] CRAN (R 4.0.2)
haven         2.4.1    2021-04-23 [1] CRAN (R 4.0.3)
hms           1.1.0    2021-05-17 [1] CRAN (R 4.0.3)
htmltools     0.5.1.1  2021-01-22 [1] CRAN (R 4.0.2)
httr          1.4.2    2020-07-20 [1] CRAN (R 4.0.2)
IRdisplay     1.0      2021-01-20 [1] CRAN (R 4.0.2)
IRkernel      1.2      2021-05-11 [1] CRAN (R 4.0.3)
jsonlite      1.7.2    2020-12-09 [1] CRAN (R 4.0.2)
labeling      0.4.2    2020-10-20 [1] CRAN (R 4.0.2)
lifecycle     1.0.0    2021-02-15 [1] CRAN (R 4.0.3)
lubridate     1.7.10   2021-02-26 [1] CRAN (R 4.0.3)
magrittr      2.0.1    2020-11-17 [1] CRAN (R 4.0.2)
modelr        0.1.8    2020-05-19 [1] CRAN (R 4.0.2)
munsell       0.5.0    2018-06-12 [1] CRAN (R 4.0.2)
openxlsx      4.2.4    2021-06-16 [1] CRAN (R 4.0.3)
pbdZMQ        0.3-5    2021-02-10 [1] CRAN (R 4.0.3)
pillar        1.6.1    2021-05-16 [1] CRAN (R 4.0.3)
pkgconfig     2.0.3    2019-09-22 [1] CRAN (R 4.0.2)
purrr       * 0.3.4    2020-04-17 [1] CRAN (R 4.0.2)
R6            2.5.0    2020-10-28 [1] CRAN (R 4.0.2)
Rcpp          1.0.7    2021-07-07 [1] CRAN (R 4.0.3)
readr       * 1.4.0    2020-10-05 [1] CRAN (R 4.0.2)
readxl        1.3.1    2019-03-13 [1] CRAN (R 4.0.2)
repr          1.1.3    2021-01-21 [1] CRAN (R 4.0.2)
reprex        2.0.0    2021-04-02 [1] CRAN (R 4.0.3)
rio           0.5.27   2021-06-21 [1] CRAN (R 4.0.3)
rlang         0.4.11   2021-04-30 [1] CRAN (R 4.0.3)
rstatix       0.7.0    2021-02-13 [1] CRAN (R 4.0.3)
rstudioapi    0.13     2020-11-12 [1] CRAN (R 4.0.2)
rvest         1.0.0    2021-03-09 [1] CRAN (R 4.0.3)
scales        1.1.1    2020-05-11 [1] CRAN (R 4.0.2)
sessioninfo   1.1.1    2018-11-05 [1] CRAN (R 4.0.2)
stringi       1.7.3    2021-07-16 [1] CRAN (R 4.0.3)
stringr     * 1.4.0    2019-02-10 [1] CRAN (R 4.0.2)
svglite       2.0.0    2021-02-20 [1] CRAN (R 4.0.3)
systemfonts   1.0.2    2021-05-11 [1] CRAN (R 4.0.3)
tibble      * 3.1.2    2021-05-16 [1] CRAN (R 4.0.3)
tidyr       * 1.1.3    2021-03-03 [1] CRAN (R 4.0.3)
```

```
 tidyselect    1.1.1    2021-04-30 [1] CRAN (R 4.0.3)
 tidyverse   * 1.3.1    2021-04-15 [1] CRAN (R 4.0.3)
 utf8          1.2.1    2021-03-12 [1] CRAN (R 4.0.3)
 uuid          0.1-4    2020-02-26 [1] CRAN (R 4.0.2)
 vctrs         0.3.8    2021-04-29 [1] CRAN (R 4.0.3)
 withr         2.4.2    2021-04-18 [1] CRAN (R 4.0.3)
 xml2          1.3.2    2020-04-23 [1] CRAN (R 4.0.2)
 zip           2.2.0    2021-05-31 [1] CRAN (R 4.0.3)

[1] /home/jbenja13/R/x86_64-pc-linux-gnu-library/4.0
[2] /usr/lib/R/library
```