

main

July 10, 2021

1 Boxplot of top five up- and downregulated genes

```
[1]: 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 Genes

1.2.1 Differential expressed genes

```

[3]: drd2 = read.delim('../_m/genes/diffExpr_maleVfemale_full.txt') %>%
      filter(Symbol == 'DRD2') %>%
      select(gencodeID, ensemblID, Symbol, logFC, t, adj.P.Val) %>%
      add_symnum() %>% mutate_if(is.character, as.factor)
drd2 %>% head(2)

```

A data.frame: 1 × 8	gencodeID <fct>	ensemblID <fct>	Symbol <fct>	logFC <dbl>	t <dbl>	adj.P.Val <dbl>
	1	ENSG00000149295.13	ENSG00000149295	DRD2	0.06437243	1.806703 0.429904

1.2.2 Load residualized expression

```

[4]: res_file = '../_m/genes/residualized_expression.tsv'
resdf0 = data.table::fread(res_file) %>%
      filter(V1 %in% drd2$gencodeID) %>% column_to_rownames("V1") %>%
      t %>% data.frame
resdf0 %>% head(2)

```

Warning message in data.table::fread(res_file):
 "Detected 393 column names but the data has 394 columns (i.e. invalid file).
 Added 1 extra default column name for the first column which is guessed to be
 row names or an index. Use setnames() afterwards if this guess is not correct,
 or fix the file write command that created the file to create a valid file."

A data.frame: 2 × 1	ENSG00000149295.13 <dbl>
	R12864 -1.903295
	R12865 1.520050

1.2.3 Load pheno data

```

[5]: pheno_file = '/ceph/users/jbenja13/projects/sex_sz_ria/input/phenotypes/_m/
      ↪phenotypes.csv'
pheno = read.csv(pheno_file) %>%
      mutate_if(is.character, as.factor) %>%
      column_to_rownames("RNum") %>%
      filter(Region=="Caudate") %>%
      select(Sex, Dx, Age, Region, BrNum)
levels(pheno$Sex) <- c("Female", "Male")
head(pheno, 2)

```

		Sex <fct>	Dx <fct>	Age <dbl>	Region <fct>	BrNum <fct>
A data.frame: 2 × 5	R12864	Female	Schizo	42.98	Caudate	Br1303
	R12865	Male	Schizo	53.12	Caudate	Br1320

1.2.4 Merge dataframe

```
[6]: resdf <- inner_join(rownames_to_column(pheno),
                        rownames_to_column(resdf0),
                        by="rowname")
dim(resdf)
resdf %>% head(2)
```

1. 393 2. 7

		rowname <chr>	Sex <fct>	Dx <fct>	Age <dbl>	Region <fct>	BrNum <fct>	ENSG00000149295.13 <dbl>
A data.frame: 2 × 7	1	R12864	Female	Schizo	42.98	Caudate	Br1303	-1.903295
	2	R12865	Male	Schizo	53.12	Caudate	Br1320	1.520050

1.2.5 Melt data frame

```
[7]: df = resdf %>% select(c('rowname', 'Sex', starts_with('ENSG')) %>%
                        pivot_longer(-c(rowname, Sex), names_to = "gencodeID", values_to = "Res"))
      ↪ %>%
inner_join(drd2, by='gencodeID') %>%
mutate_at(vars("Symbol", "gencodeID"), as.factor)

df$Symbol = with(df, reorder(Symbol, adj.P.Val, median))
head(df, 2)
```

	rowname <chr>	Sex <fct>	gencodeID <fct>	Res <dbl>	ensemblID <fct>	Symbol <fct>	logFC <dbl>
A tibble: 2 × 11	R12864	Female	ENSG00000149295.13	-1.903295	ENSG00000149295	DRD2	0.06437
	R12865	Male	ENSG00000149295.13	1.520050	ENSG00000149295	DRD2	0.06437

1.2.6 Initial ggplot with ggpubr

```
[8]: tmp = drd2 %>% mutate(group1='Female', group2='Male', y_pos=4) %>%
      mutate_if(is.character, as.factor)

bxp_g <- ggboxplot(df, x="Sex", y="Res", facet.by=c("Symbol"), fill="Sex",
      ↪ xlab='',
      panel.labs.font=list(face='bold', size = 16), palette="npg",
      outlier.shape=NA, ylab='Residualized Expression', add='jitter',
      ylim=c(-10, 5), add.params=list(alpha=0.5), legend="bottom",
      ggtheme=theme_pubr()) +
geom_signif(data=tmp,
```

```

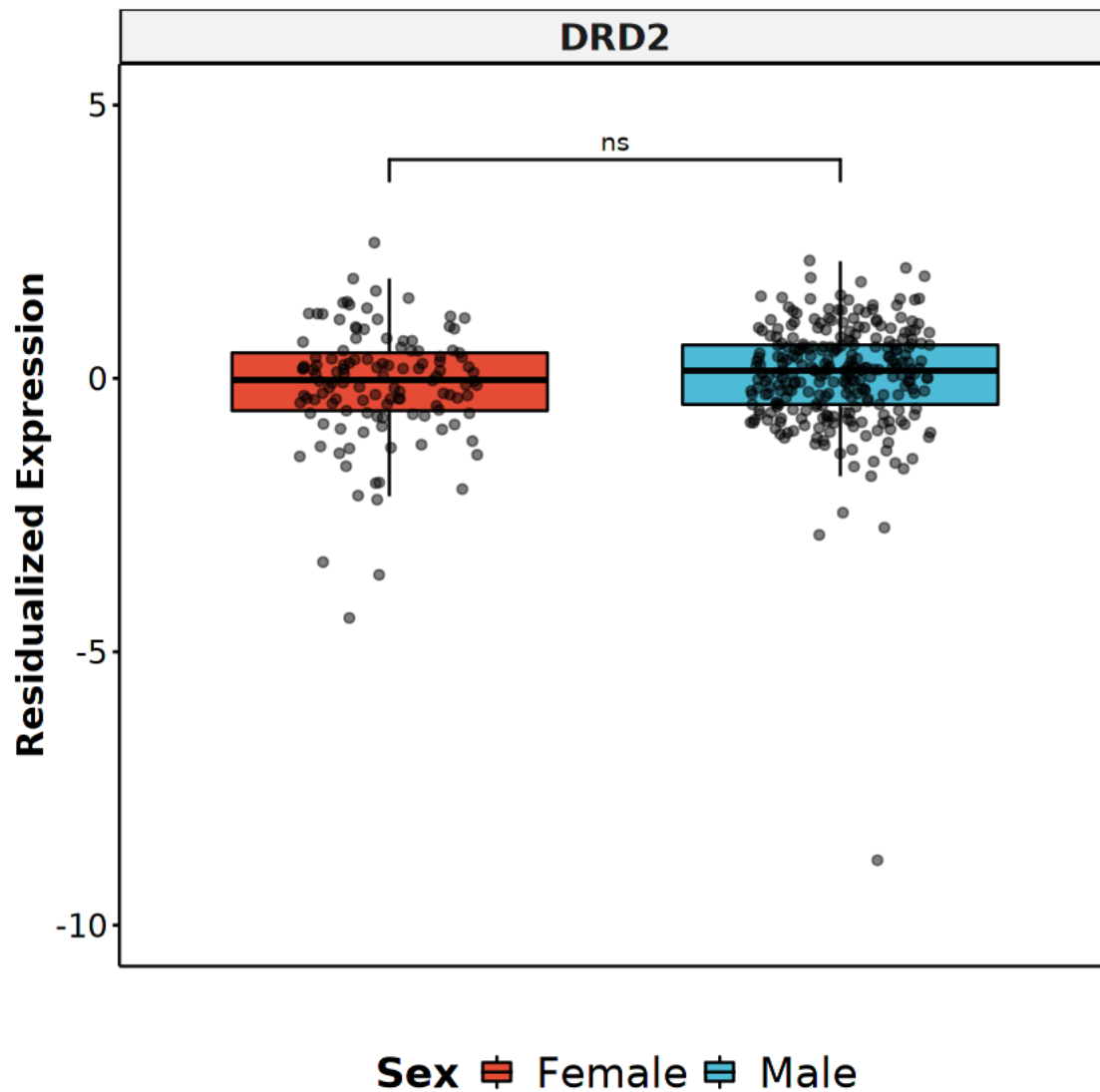
aes(xmin=group1, xmax=group2, annotations=p.signif, y_position_
↪= y_pos),
  manual=TRUE) +
font("xy.title", size=16, face="bold") + font("xy.text", size=14) +
font("legend.title", size=18, face="bold") +
font("legend.text", size=18) + rremove("x.text") + rremove("x.ticks")

bxp_g

```

Warning message:

"Ignoring unknown aesthetics: xmin, xmax, annotations, y_position"



1.3 Junction

1.3.1 Differential expressed genes

```
[9]: drd2 = data.table::fread('../_m/junctions/diffExpr_maleVfemale_full.txt') %>%
      filter(Symbol == 'DRD2') %>%
      select(V1, gencodeGeneID, ensemblID, Symbol, logFC, t, adj.P.Val) %>%
      rename("gencodeID"="gencodeGeneID", "Feature"="V1") %>% add_symnum()
      drd2 %>% head(2)
```

	Feature	gencodeID	ensemblID	Symbol	logF
	<chr>	<chr>	<chr>	<chr>	<dbl>
A data.table: 2 × 9	chr11:113418137-113424366(-)	ENSG00000149295.13	ENSG00000149295	DRD2	0.08
	chr11:113414462-113415420(-)	ENSG00000149295.13	ENSG00000149295	DRD2	0.08

1.3.2 Load residualized expression

```
[10]: res_file = '../_m/junctions/residualized_expression.tsv'
      resdf0 = data.table::fread(res_file) %>%
      inner_join(drd2, by=c("V1"="Feature")) %>%
      mutate(ID=paste0("j",1:dim(drd2)[1])) %>%
      column_to_rownames("ID") %>% select(starts_with("R")) %>%
      t %>% data.frame
      resdf0 %>% head(2)
      jxn_ann = data.table::fread(res_file) %>%
      inner_join(drd2, by=c("V1"="Feature")) %>%
      mutate(ID=paste0("j",1:dim(drd2)[1])) %>%
      select(!starts_with("R"))
      jxn_ann %>% head(2)
```

Warning message in data.table::fread(res_file):
 "Detected 393 column names but the data has 394 columns (i.e. invalid file).
 Added 1 extra default column name for the first column which is guessed to be
 row names or an index. Use setnames() afterwards if this guess is not correct,
 or fix the file write command that created the file to create a valid file."

	j1	j2	j3	j4	j5	j6	j7		
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>		
A data.frame: 2 × 10	R12864	-0.8127860	-0.6714232	-0.1656533	-0.8511028	-0.7196601	-0.6241724	-0.0	-0.0
	R12865	0.3310892	-0.6067889	-1.1151507	0.2102291	0.3115348	0.3511729	-0.0	-0.0

Warning message in data.table::fread(res_file):
 "Detected 393 column names but the data has 394 columns (i.e. invalid file).
 Added 1 extra default column name for the first column which is guessed to be
 row names or an index. Use setnames() afterwards if this guess is not correct,
 or fix the file write command that created the file to create a valid file."

	V1 <chr>	gencodeID <chr>	ensemblID <chr>	Symbol <chr>	log <dbl>
A data.table: 2 × 10	chr11:113410921-113412555(-)	ENSG00000149295.13	ENSG00000149295	DRD2	0.07
	chr11:113412884-113414374(-)	ENSG00000149295.13	ENSG00000149295	DRD2	0.08

1.3.3 Load pheno data

```
[11]: pheno_file = '/ceph/users/jbenja13/projects/sex_sz_ria/input/phenotypes/_m/
       ↪phenotypes.csv'
pheno = read.csv(pheno_file) %>%
  mutate_if(is.character, as.factor) %>%
  column_to_rownames("RNum") %>%
  filter(Region=="Caudate") %>%
  select(Sex, Dx, Age, Region, BrNum)
levels(pheno$Sex) <- c("Female", "Male")
head(pheno, 2)
```

	Sex <fct>	Dx <fct>	Age <dbl>	Region <fct>	BrNum <fct>
A data.frame: 2 × 5	R12864	Female	Schizo	42.98	Caudate
	R12865	Male	Schizo	53.12	Caudate

1.3.4 Merge dataframe

```
[12]: resdf <- inner_join(rownames_to_column(pheno),
                        rownames_to_column(resdf0),
                        by="rowname")
dim(resdf)
resdf %>% head(2)
```

1. 393 2. 16

	rowname <chr>	Sex <fct>	Dx <fct>	Age <dbl>	Region <fct>	BrNum <fct>	j1 <dbl>	j2 <dbl>	j3 <dbl>	
A data.frame: 2 × 16	1	R12864	Female	Schizo	42.98	Caudate	Br1303	-0.8127860	-0.6714232	-0.6714232
	2	R12865	Male	Schizo	53.12	Caudate	Br1320	0.3310892	-0.6067889	-0.6067889

1.3.5 Melt data frame

```
[13]: df = resdf %>% select(c('rowname', 'Sex', starts_with('j'))) %>%
  pivot_longer(-c(rowname, Sex), names_to = "ID", values_to = "Res") %>%
  inner_join(jxn_ann, by="ID") %>% mutate(New_ID=paste(Symbol, V1, sep='\n'))
  ↪%>%
  mutate_at(vars("Symbol", "New_ID"), as.factor)
df$New_ID = with(df, reorder(New_ID, adj.P.Val, median))
df %>% head(2)
```

	rowname	Sex	ID	Res	V1	gencodeID
	<chr>	<fct>	<chr>	<dbl>	<chr>	<chr>
A tibble: 2 × 14	R12864	Female	j1	-0.8127860	chr11:113410921-113412555(-)	ENSG00000149295.1
	R12864	Female	j2	-0.6714232	chr11:113412884-113414374(-)	ENSG00000149295.1

1.3.6 Initial ggplot with ggpubr

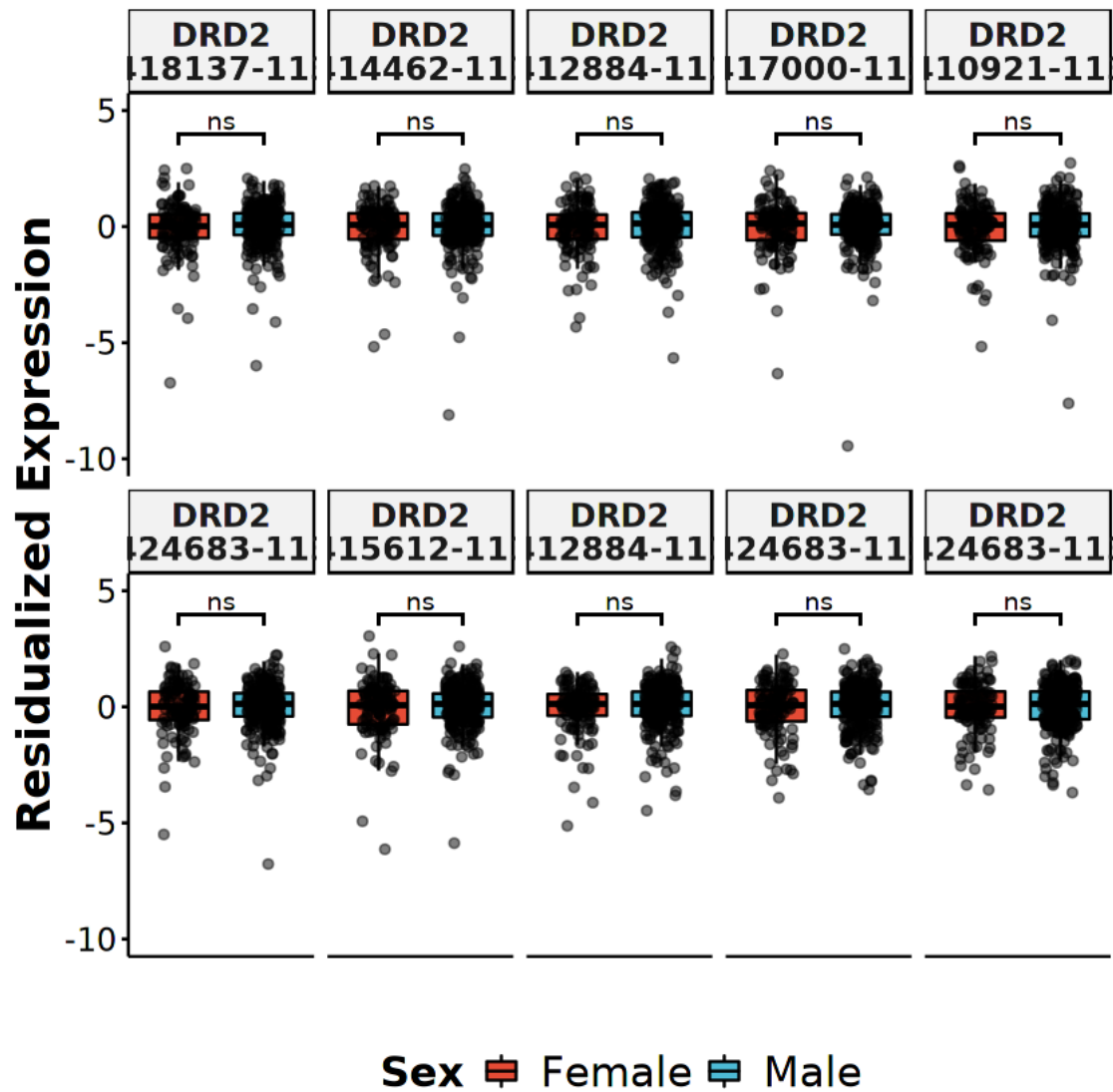
```
[14]: tmp = drd2 %>% mutate(group1='Female', group2='Male', y_pos=4) %>%
      mutate_if(is.character, as.factor)

bxp_j <- ggboxplot(df, x="Sex", y="Res", facet.by=c("New_ID"), fill="Sex",
  xlab='',
  panel.labs.font=list(face='bold', size = 14), palette="npg",
  outlier.shape=NA, ylab='Residualized Expression', add='jitter',
  ylim=c(-10, 5), add.params=list(alpha=0.5), legend="bottom",
  ggtheme=theme_pubr(), ncol=5) +
  geom_signif(data=tmp,
    aes(xmin=group1, xmax=group2, annotations=p.signif, y_position=
  y_pos),
    manual=TRUE) +
  font("xy.title", size=20, face="bold") + font("xy.text", size=14) +
  font("legend.title", size=18, face="bold") +
  font("legend.text", size=18) + rremove("x.text") + rremove("x.ticks")

bxp_j
```

Warning message:

"Ignoring unknown aesthetics: xmin, xmax, annotations, y_position"



1.3.7 Annotate figure

```
[15]: save_ggplots('drd2_gene', bxp_g, 6, 6)
      save_ggplots('drd2_junctions', bxp_j, 16, 10)
```

1.4 Session Info

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

```
[1] "2021-07-10 10:20:31 EDT"
```



```

    user  system elapsed
25.141 25.926 27.737

```

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     2021-07-10

```

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	1.6.2	2021-05-17	[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