main

September 22, 2021

1 Plot and comparisons

[1]:
```
library(tidyverse)
library(ggpubr)
```

```
  Attaching packages                                   tidyverse
1.3.1

 ggplot2 3.3.5      purrr   0.3.4
 tibble  3.1.4      dplyr   1.0.7
 tidyr   1.1.3      stringr 1.4.0
 readr   2.0.1      forcats 0.5.1

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

1.1 Functions

[2]:
```
save_plot <- function(p, fn, w=7, h=6){
    for(ext in c(".pdf", ".png", ".svg")){
        ggsave(filename=paste0(fn,ext), plot=p, width=w, height=h)
    }
}


get_ml_summary <- function(fn){
    ml_df = data.table::fread(fn) %>% mutate_at("fold", as.character) %>%
        select(fold, n_features, n_redundant, starts_with("test_score_r2")) %>%
        pivot_longer(-fold) %>% group_by(name) %>%
        summarise(Mean=mean(value), Median=median(value), Std=sd(value), .
 ↪groups = "keep")
    return(ml_df)
}
```

```
get_metrics <- function(filename, tissue){
    datalist = list()
    for(fn in Sys.glob(filename)){
        gene_id = str_extract(fn, "ENSG\\d+_\\d+")
        dat <- get_ml_summary(fn)
        dat["Geneid"] = gene_id
        datalist[[gene_id]] <- dat
    }
    ml_df <- bind_rows(datalist)
    ml_df["Tissue"] = tissue
    return(ml_df)
}
```

1.2 Load metrics

1.2.1 Random forest

[3]:
```
rf = data.table::fread("../../rf/summary_10Folds_allTissues.tsv") %>%
    as.data.frame %>% mutate_if(is.character, as.factor) %>%
    mutate_at("fold", as.character) %>%
    select(tissue, feature, fold, n_features, starts_with("test_score_r2")) %>%
    pivot_longer(-c(tissue, feature, fold), names_to="metric",␣
 ↪values_to="score") %>%
    group_by(tissue, feature, metric) %>%
    summarise(Mean=mean(score), Median=median(score), Std=sd(score), .groups =␣
 ↪"keep") %>%
    filter(metric == "test_score_r2") %>% mutate("model"="Random Forest")
dim(rf)
rf %>% head(2)
```

1. 9304 2. 7

A grouped_df: 2 × 7

tissue <fct>	feature <fct>	metric <chr>	Mean <dbl>	Median <dbl>	Std <dbl>
Caudate	ENSG00000003249_13	test_score_r2	-0.04040379	-0.01089123	0.1993740
Caudate	ENSG00000003509_15	test_score_r2	-0.09541787	-0.03918813	0.1591028

1.2.2 Elastic net

[4]:
```
enet = data.table::fread("../../enet/summary_10Folds_allTissues.tsv") %>%
    as.data.frame %>% mutate_if(is.character, as.factor) %>%
    mutate_at("fold", as.character) %>%
    select(tissue, feature, fold, n_features, starts_with("test_score_r2")) %>%
    pivot_longer(-c(tissue, feature, fold), names_to="metric",␣
 ↪values_to="score") %>%
    group_by(tissue, feature, metric) %>%
    summarise(Mean=mean(score), Median=median(score), Std=sd(score), .groups =␣
 ↪"keep") %>%
    filter(metric == "test_score_r2") %>% mutate("model"="Elastic Net")
```

```
dim(enet)
enet %>% head(2)
```

1. 9324 2. 7

A grouped_df: 2 × 7

tissue <fct>	feature <fct>	metric <chr>	Mean <dbl>	Median <dbl>	Std <dbl>
Caudate	ENSG00000003249_13	test_score_r2	-0.09865048	-0.07496671	0.1878822
Caudate	ENSG00000003509_15	test_score_r2	-0.03012670	-0.01017748	0.0723698

1.3 Annotate

[5]:
```
dtu = data.table::fread(paste0("../../../../../differential_analysis/
 ↪tissue_comparison/",
                               "ds_summary/_m/
 ↪diffSplicing_ancestry_FDR05_4regions.tsv")) %>%
    select(gene, Tissue) %>% distinct %>% rename("gene_name"="gene")

degs = data.table::fread("../../../../_m/degs_annotation.txt") %>%
    select(V1, ensemblID, gene_name, Tissue) %>% distinct %>%
    rename("Feature"="V1") %>% inner_join(dtu, by=c("Tissue", "gene_name")) %>%
    rename("tissue"="Tissue") %>% mutate("DTU"="DTU")
```

[6]:
```
df = bind_rows(rf, enet) %>% mutate(Feature=gsub("_", ".", feature)) %>%
    left_join(degs, by=c("tissue", "Feature")) %>% as.data.frame %>%
    mutate(DTU = replace_na(DTU, "DEG")) %>%
    mutate_if(is.character, as.factor)
dim(df)
df %>% head(2)
```

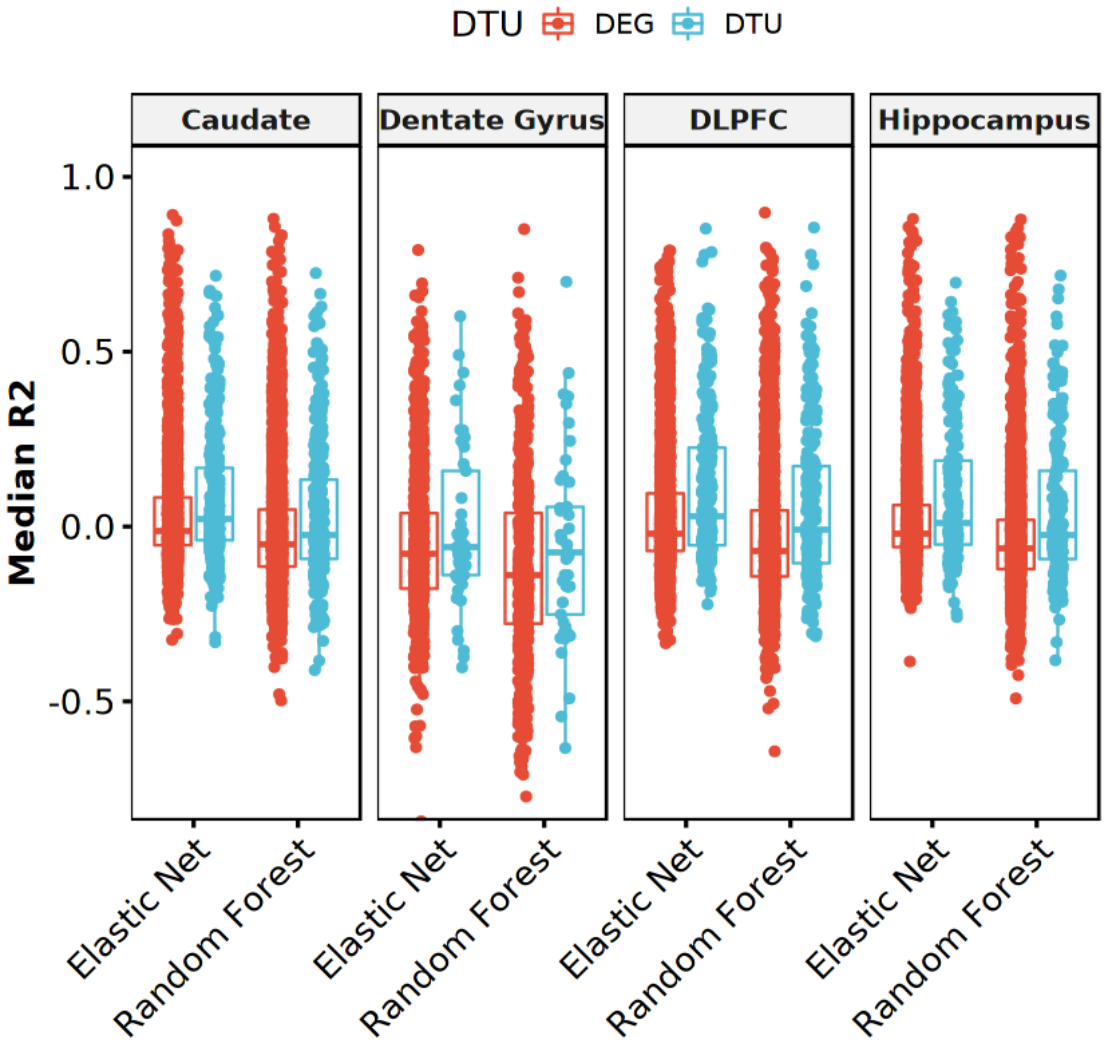
1. 18628 2. 11

A data.frame: 2 × 11

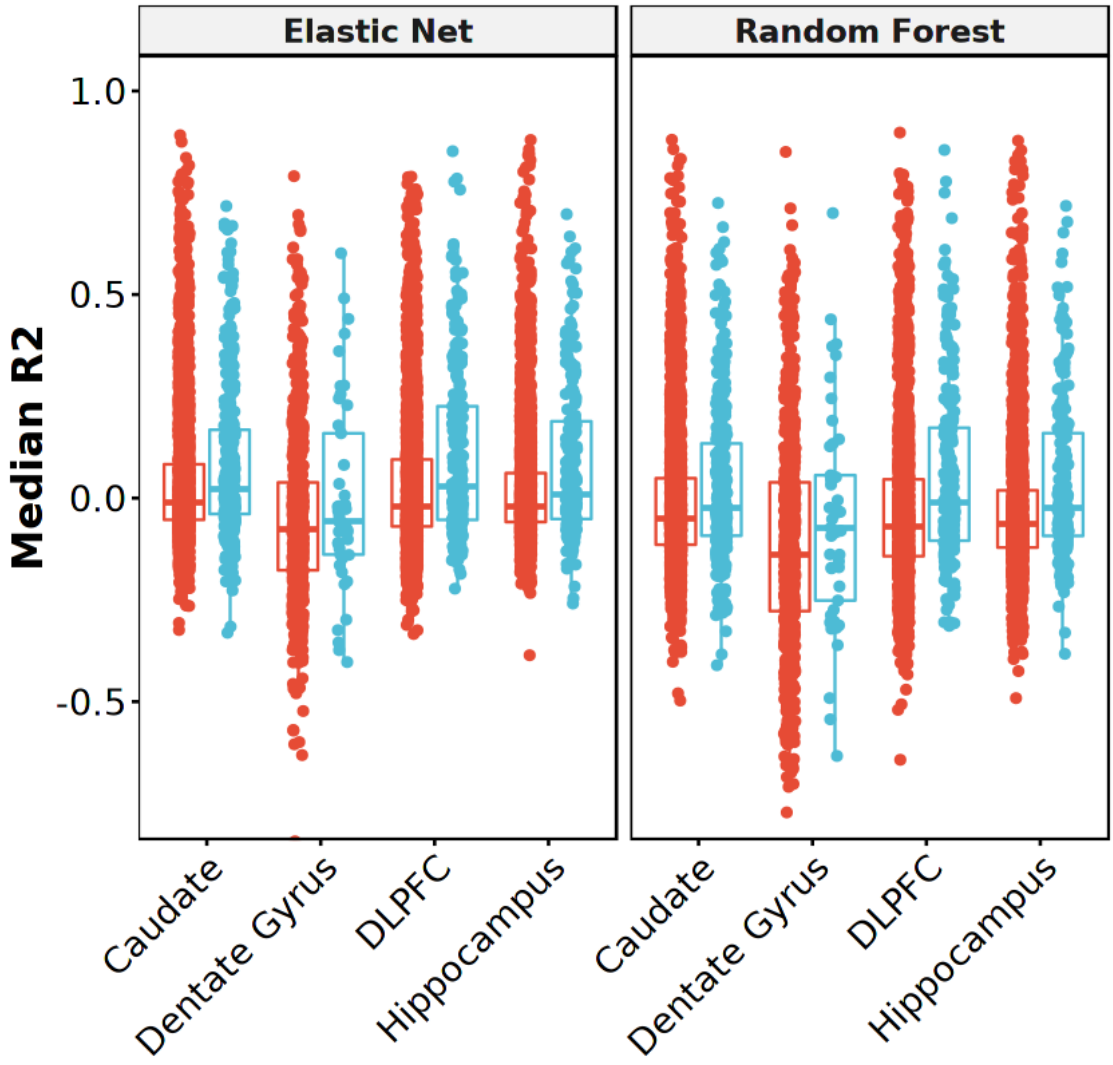
	tissue <fct>	feature <fct>	metric <fct>	Mean <dbl>	Median <dbl>	Std <dbl
1	Caudate	ENSG00000003249_13	test_score_r2	-0.04040379	-0.01089123	0.1993
2	Caudate	ENSG00000003509_15	test_score_r2	-0.09541787	-0.03918813	0.1591

1.4 Merge and plot

[7]:
```
df %>% #filter(DTU == "DTU") %>%
    ggboxplot(x="model", y="Median", color="DTU", add="jitter",
              facet.by="tissue", palette="npg", ylim=c(-0.75, 1),
              ylab="Median R2", xlab="",
              panel.labs.font=list(face='bold'), ncol=4,
              ggtheme=theme_pubr(base_size=15, border=TRUE)) +
    rotate_x_text(45) + font("xy.title", face="bold")
```



[8]:
```
bxp = df %>% ggboxplot(x="tissue", y="Median", color="DTU", add="jitter",
                       facet.by="model", palette="npg", ylim=c(-0.75, 1),
                       ylab="Median R2", xlab="", legend="None",
                       panel.labs.font=list(face='bold', size = 14)) +
    rotate_x_text(45) + font("xy.title", size=18, face="bold") +
    font("xy.text", size=16) + font("legend.text", size=16)
save_plot(bxp, "summary_boxplots_r2_2methods", 9, 6)
bxp
```



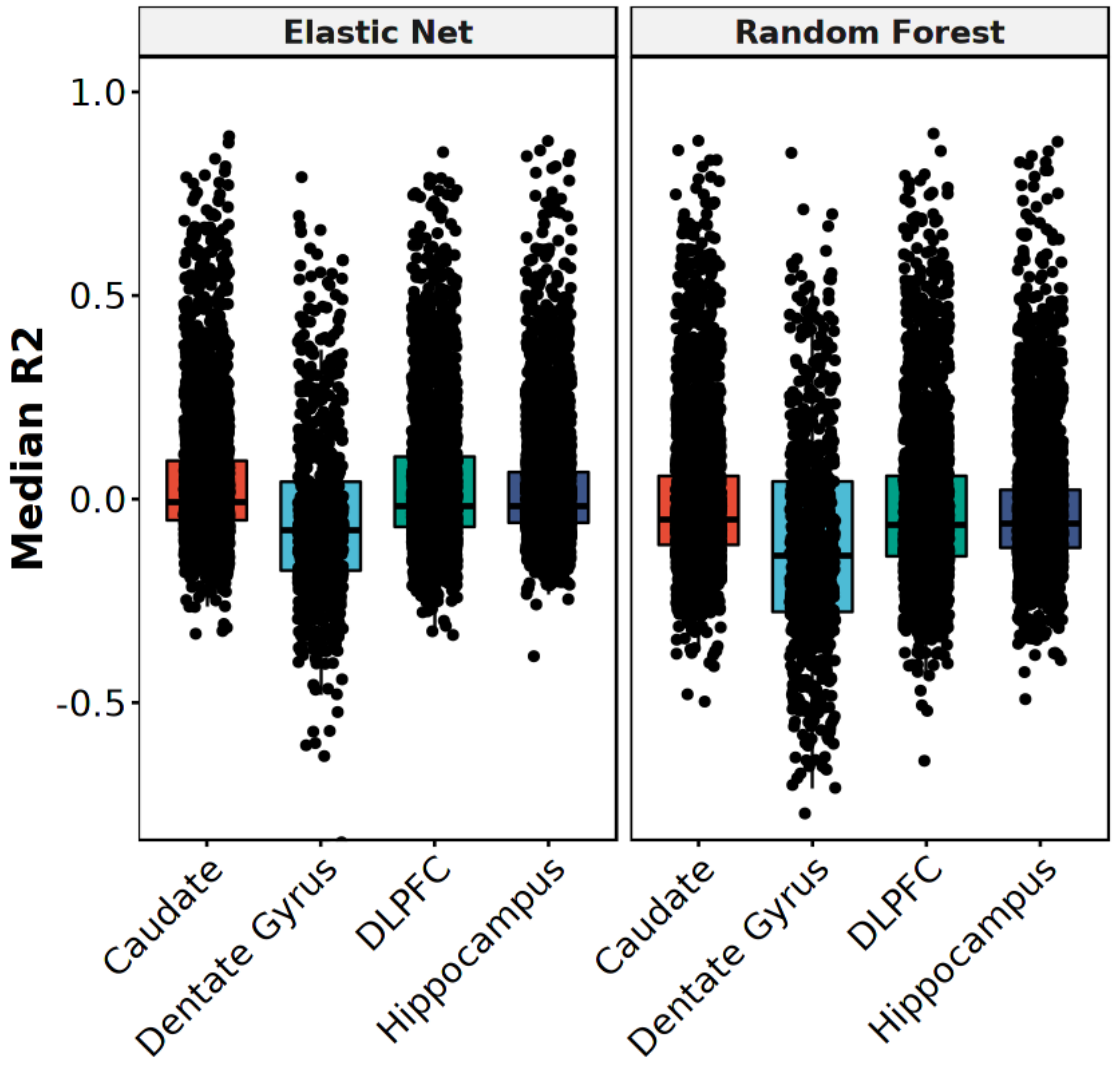
[9]:
```
df2 = bind_rows(rf, enet)
df2 %>% head(2)
```

A grouped_df: 2 × 7

tissue <fct>	feature <fct>	metric <chr>	Mean <dbl>	Median <dbl>	Std <dbl>
Caudate	ENSG00000003249_13	test_score_r2	-0.04040379	-0.01089123	0.1993740
Caudate	ENSG00000003509_15	test_score_r2	-0.09541787	-0.03918813	0.1591028

[10]:
```
bxp = df2 %>% ggboxplot(x="tissue", y="Median", fill="tissue", add="jitter",
                        facet.by="model", palette="npg", ylim=c(-0.75, 1),
                        ylab="Median R2", xlab="", legend="None",
                        panel.labs.font=list(face='bold', size = 14)) +
    rotate_x_text(45) + font("xy.title", size=18, face="bold") +
```

```
    font("xy.text", size=16) + font("legend.text", size=16)
save_plot(bxp, "summary_boxplots_r2_2methods", 6, 5)
bxp
```



1.5 Reproducibility Information

[11]:
```
Sys.time()
proc.time()
options(width = 120)
sessioninfo::session_info()
```

```
[1] "2021-09-22 18:07:45 EDT"
```

```
  user  system elapsed
21.363   0.982  19.906
```

```
 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     2021-09-22
```

```
 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         0.7.9    2021-07-27 [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           3.0.1    2021-07-17 [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)
fastmap       1.1.0    2021-01-25 [1] CRAN (R 4.0.2)
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         2.4.3    2021-08-04 [1] CRAN (R 4.0.3)
hms           1.1.0    2021-05-17 [1] CRAN (R 4.0.3)
htmltools     0.5.2    2021-08-25 [1] CRAN (R 4.0.3)
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        1.6.2    2021-07-29 [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            2.5.1    2021-08-19 [1] CRAN (R 4.0.3)
Rcpp          1.0.7    2021-07-07 [1] CRAN (R 4.0.3)
readr       * 2.0.1    2021-08-10 [1] CRAN (R 4.0.3)
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        2.0.1    2021-08-05 [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         1.0.1    2021-07-26 [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       1.7.4    2021-08-25 [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      * 3.1.4    2021-08-25 [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)
tzdb          0.1.2    2021-07-20 [1] CRAN (R 4.0.3)
utf8          1.2.2    2021-07-24 [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
```