BUSTED_MH

sadie

1/10/2020

```
library(readr)
library(dplyr)
library(ggplot2)
library(stringr)
library(plotly)
library(tidyr)
```

load data

```
BUSTED_MH <- read_csv("~/Downloads/selectome_subset_THvDH_1_10_2020")

BUSTED.SRV.dat.selectome <- read_csv("~/Downloads/BUSTED-SRV_selectome_5_20_19")

BUSTED.dat.selectome <- read_csv("~/Downloads/BUSTED_selectome_5_20_19")
```

get basenames for files since full paths aren't going to match

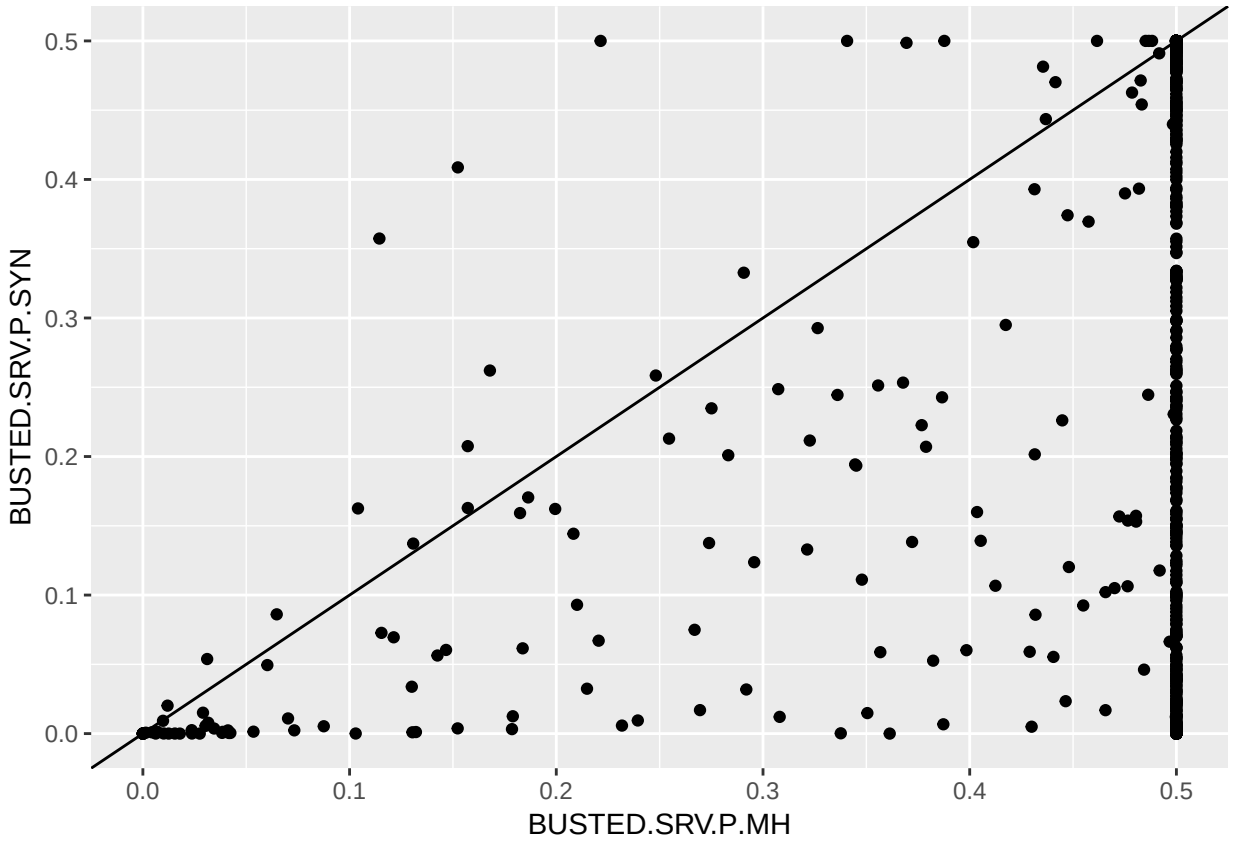
```
BUSTED_MH$basename <- BUSTED_MH$FILE %>% basename()

BUSTED.SRV.dat.selectome$basename <-BUSTED.SRV.dat.selectome$FILE %>% basename()

BUSTED.dat.selectome$basename <-BUSTED.dat.selectome$FILE %>% basename()
```

create one df with both BUSTED[SMH] and BUSTED[S] results matching the results by basename, number of sites and number of sequences.

```
dat <- inner_join(BUSTED_MH,BUSTED.SRV.dat.selectome,  by = c("basename", "Sites", "Sequences"), suffix

dat.sel <- inner_join(BUSTED.dat.selectome,BUSTED.SRV.dat.selectome, by = c("basename", "Sites", "Seque

dat_3 <- inner_join(dat, BUSTED.dat.selectome, by = c("basename", "Sites", "Sequences"))

attach(dat)
```

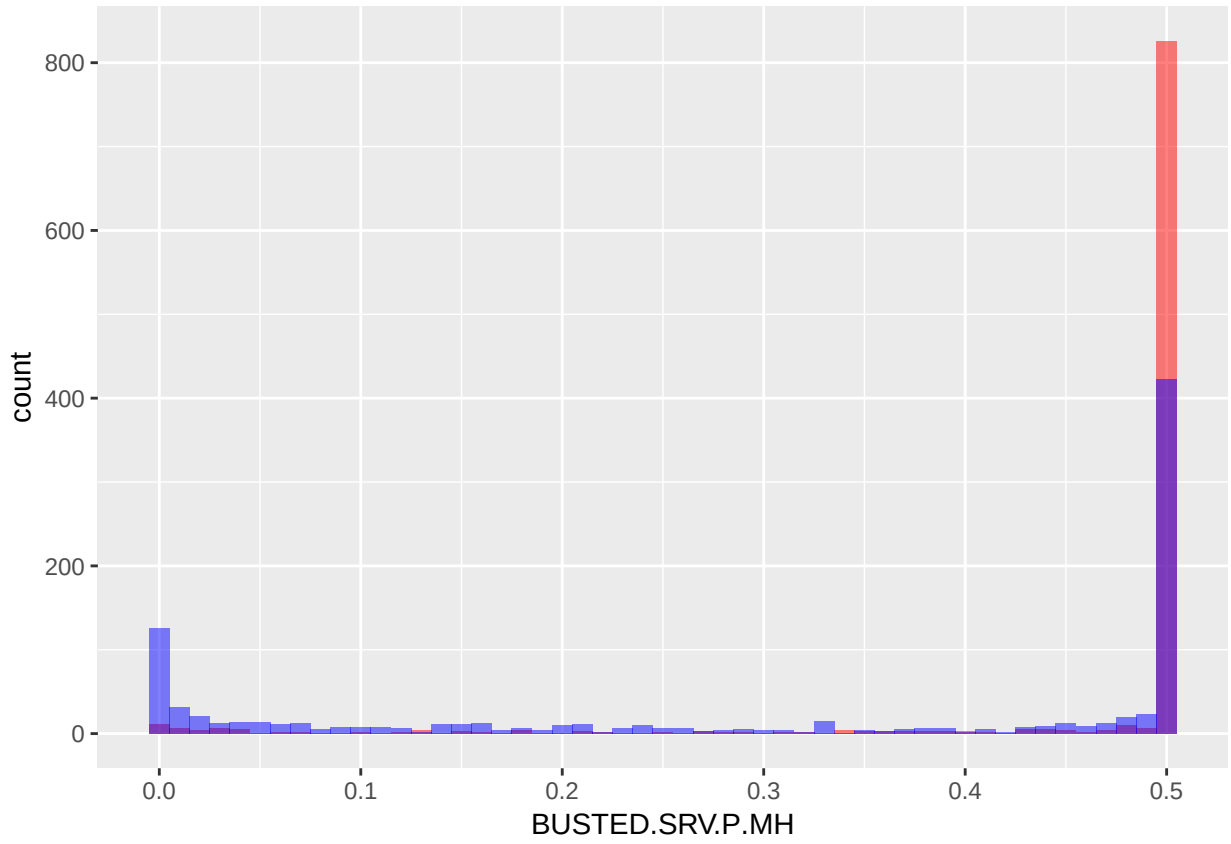
line is x = y

```
dat %>% ggplot() + geom_point(aes(x=BUSTED.SRV.P.MH, y = BUSTED.SRV.P.SYN)) + geom_abline(slope = 1, in
```

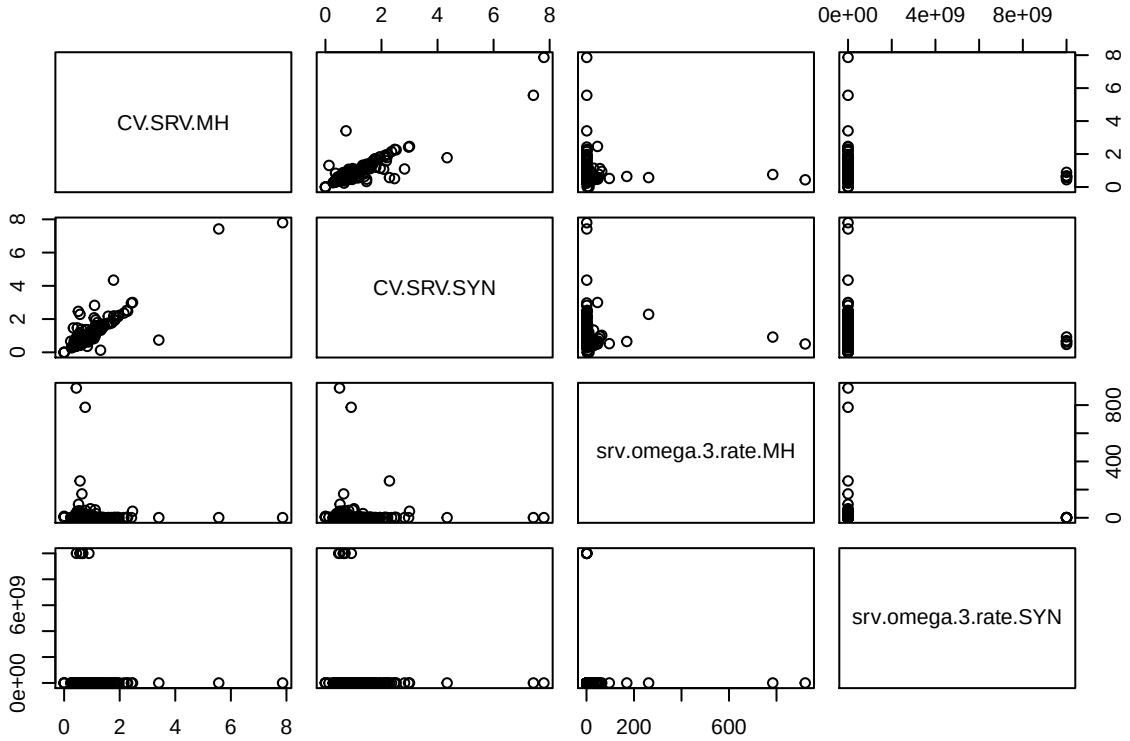


BUSTED[S] p values are typically lower than BUSTED[SMH] p values

```
dat %>% ggplot() + geom_histogram(aes(x = BUSTED.SRV.P.MH), fill = "red", alpha = 0.5, binwidth = 0.01)
  geom_histogram(aes(x = BUSTED.SRV.P.SYN), fill = 'blue' , alpha = 0.5, binwidth = 0.01)
```




```
pairs(data = dat, ~ CV.SRV.MH + CV.SRV.SYN + srv.omega.3.rate.MH + srv.omega.3.rate.SYN)
```



oh how do the two way tables compare???

```
source("/Volumes/GoogleDrive/My Drive/BUSTED-SRV/R/useful_functions.R")

gen.sig.table.mh <- function(dat = x, p = 0.05, Bon = FALSE){
  #prints out a table that tells what proportion of data files
  #are under selection/not under selection according to BUSTED
  #and BUSTED+SRV
  require("xtable")
  library("xtable")

  #bonferroni correction
  if(Bon ==TRUE){
  alpha <- p/nrow(dat)
  }
  if(Bon == FALSE){
    alpha <- p
  }

  under.sel.busted = which(dat$BUSTED.SRV.P.MH<=alpha)
  under.sel.srv = which(dat$BUSTED.SRV.P.SYN<=alpha)


  mat = matrix( rep(0),nrow=length(dat$BUSTED.SRV.P.MH),ncol = 2,
                dimnames= list(1:length(dat$BUSTED.SRV.P.MH),
                               c("BUSTED[MH]", "BUSTED[S]")))
  mat[under.sel.busted,1] = 1
  mat[under.sel.srv,2] = 1
  sel.tab = table(mat[,1],mat[,2], dnn = colnames(mat))
  t.names <- c("No Selection", "Selection")
  row.names(sel.tab) <- t.names[rownames(sel.tab) %>% as.numeric()+1]
  colnames(sel.tab) <- t.names[colnames(sel.tab) %>% as.numeric()+1]
  Sel.Prop =prop.table(sel.tab)
  test.table = xtable(Sel.Prop)


  print.xtable(test.table, type = "latex", file = "sel_table.txt")
  return(Sel.Prop)
}

dat %>% gen.sig.table.mh()
```

```
## Loading required package: xtable
```

```
##                BUSTED[S]
## BUSTED[MH]      No Selection   Selection
##   No Selection  0.780185759 0.186790506
##   Selection     0.001031992 0.031991744
```

so about 19% of the data sets are under selection according to BUSTED[S] but not BUSTED[SMH].

```
dat.sel %>% gen.sig.table()
```

```
##                BUSTED-SRV
## BUSTED          No Selection Selection
##   No Selection    0.6341969 0.0880829
##   Selection       0.1471503 0.1305699
```

about 15% of the data sets are under selection according to BUSTED but not BUSTED[S]. Overall, more data sets have no evidence of positive selection according to both tests for BUSTED[S] and BUSTED[SMH] verse BUSTED and BUSTED[S]

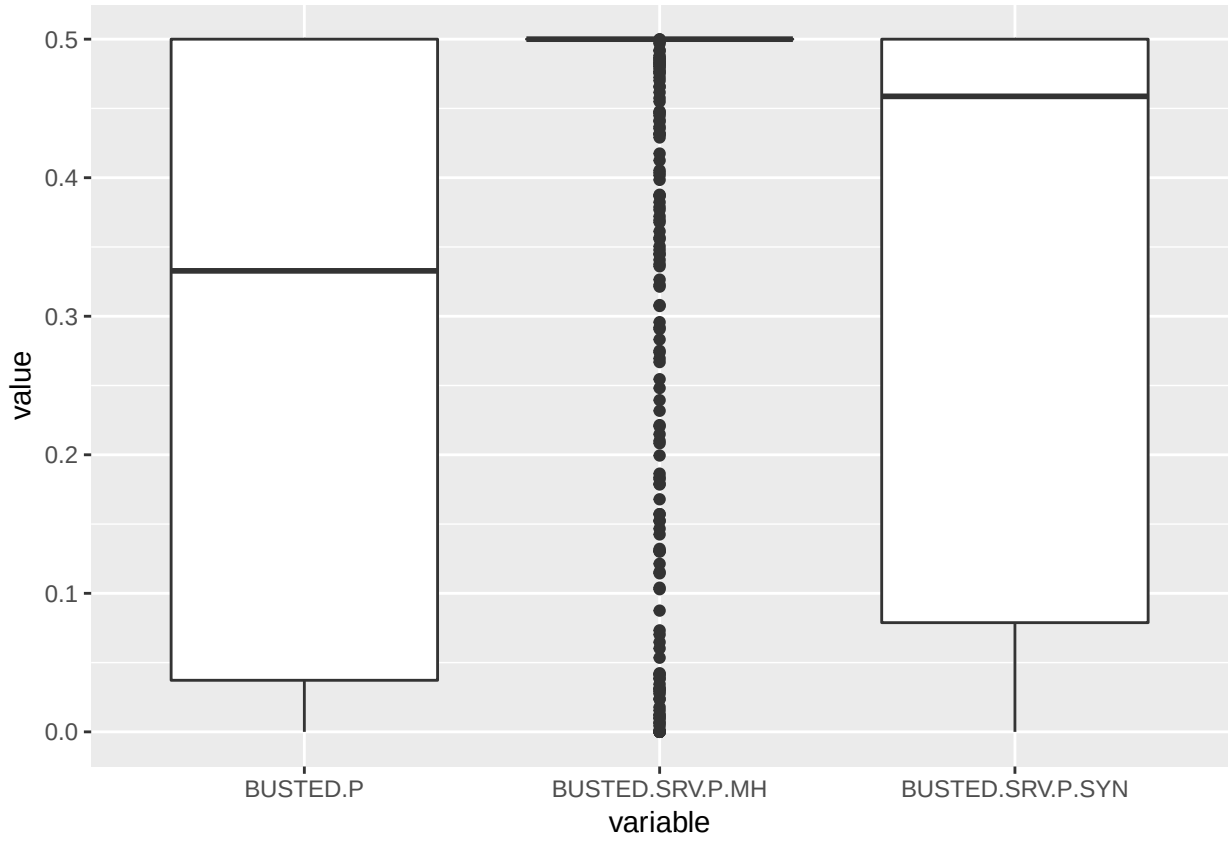
three-way comparison of p values

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

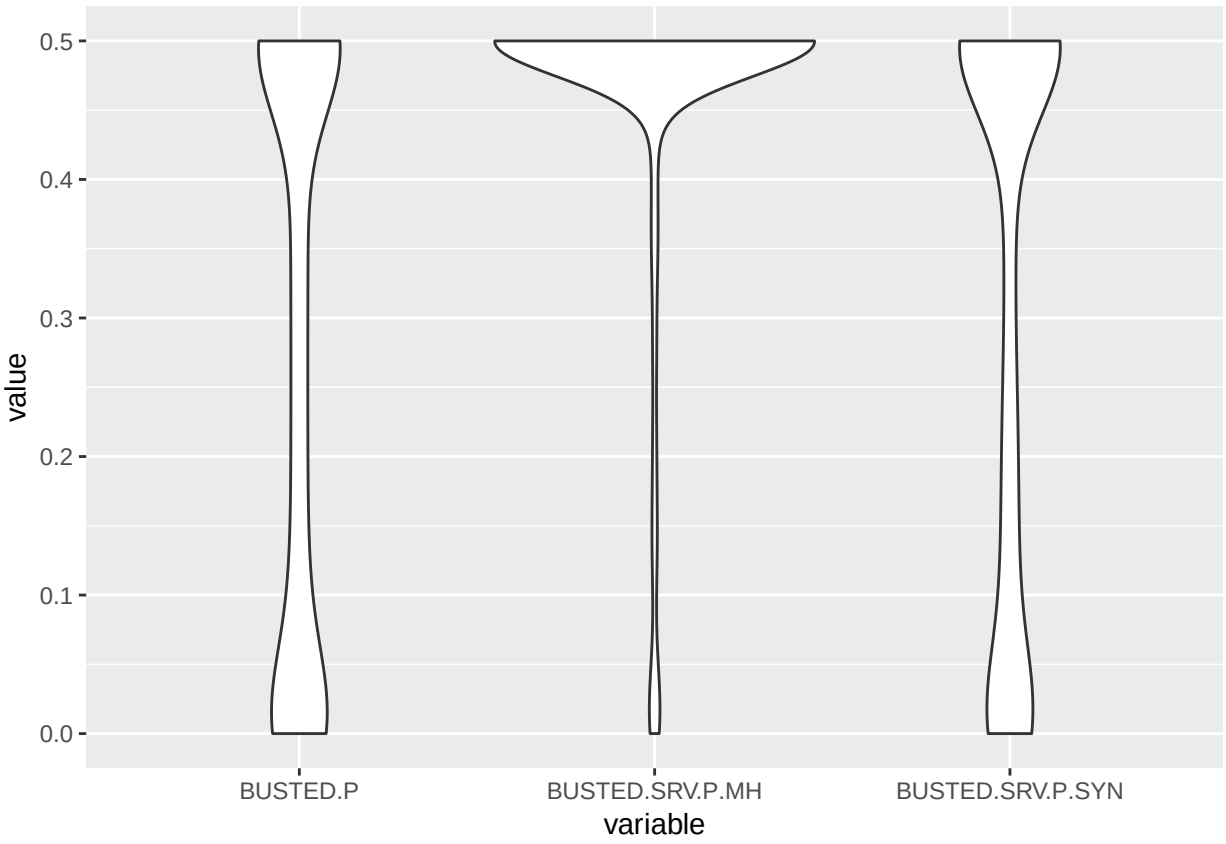
```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
dat_3 %>% melt(measure.vars = c("BUSTED.P", "BUSTED.SRV.P.MH", "BUSTED.SRV.P.SYN"), id.vars = c("basenar

temp %>% ggplot()+ geom_boxplot( aes(y = value, x = variable))
```



or violin plot

```
temp %>% ggplot()+ geom_violin( aes(y = value, x = variable))
```



basically shows that for BUSTED[SMH] p values are skewed much higher in comparison to both BUSTED and BUSTED[S]