Phylogenetic analysis

D.Smutin

2024-10-31

Setup

Just tree

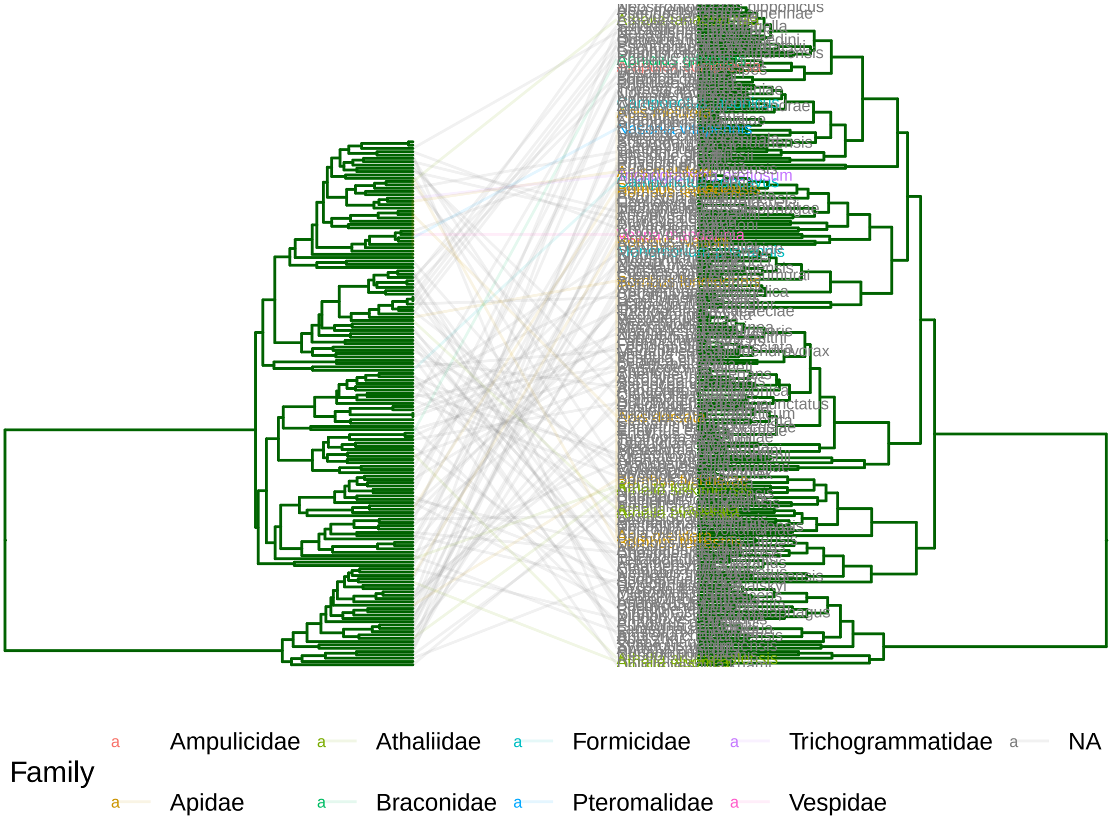
Read tree

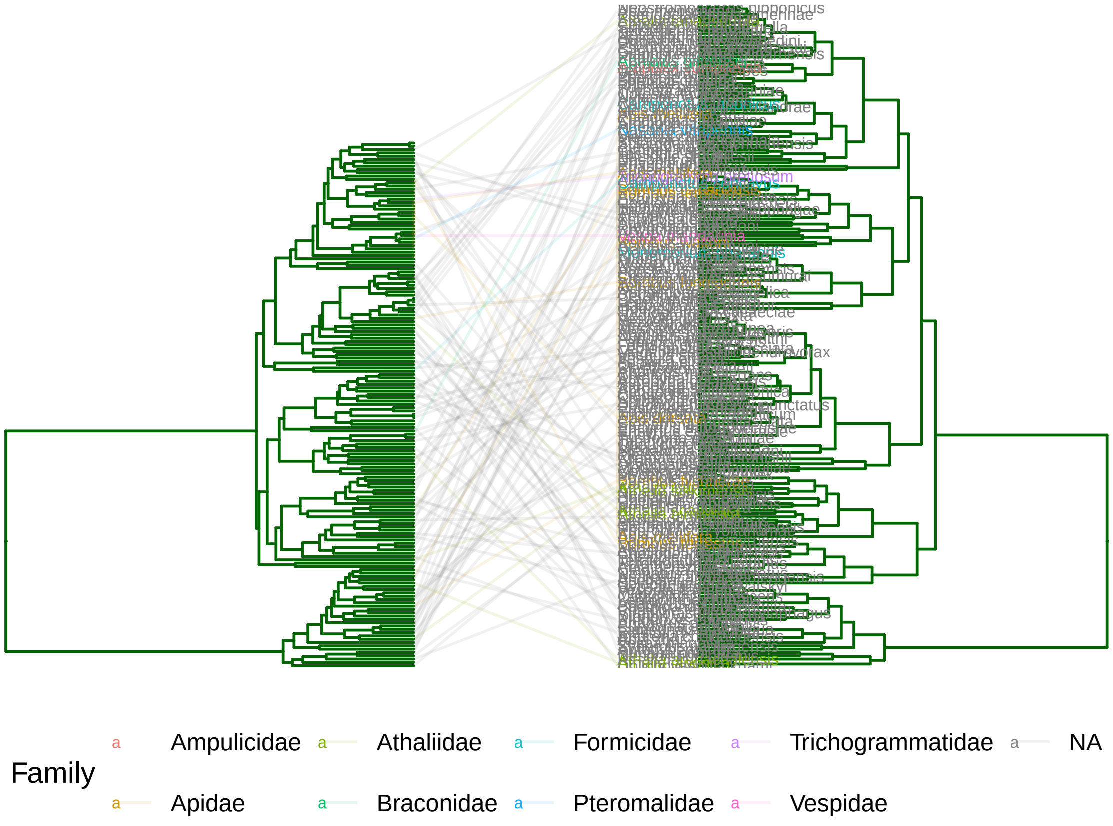
```r
leg <- readxl::read_xlsx("../data/legends/TE_101_annot.xlsx") %>%
  rename("label"="SP") %>%
  mutate(repeats = apply(.[,-c(1:4)], 1, sum))
colnames(leg) <- colnames(leg) %>% str_remove_all("[?:]")

# Read files, combine with legend and drop tips which are not represented
rb <- function(path) {
  tmp <- (read.nexus(path) %>% ggtree)$data %>%
    mutate(label = label %>%
             str_replace("_", " ") %>%
             str_remove("_.*")) %>%
    left_join(leg, by = "label")
}

rn <- function(path) {
  tmp <- (read.tree(path) %>% ggtree)$data %>%
    mutate(label = label %>%
             str_replace("_", " ") %>%
             str_remove("_.*")) %>%
    left_join(leg, by = "label")
}

rt <- function(path) {
  tmp <- (read.nexus(path) %>% ggtree(
    #branch.length = 'none'
    ))$data %>%
    mutate(label = label %>%
             str_replace("_", " ") %>%
             str_remove("_.*")) %>%
    left_join(leg, by = "label")
}

d1 <- rb("../data/tree/beast/hymen16s.out.nex")
```

```
d2 <- rb("../data/tree/beast/hymenCOX.out.nex")
d3 <- rn("../data/tree/nj/hymen16s.nwk")
d4 <- rn("../data/tree/nj/hymenCOXI.nwk")

d5 <- rt("../data/tree/new/mrbayes.tre")
d6 <- rn("../data/tree/new/tree.nwk")
```

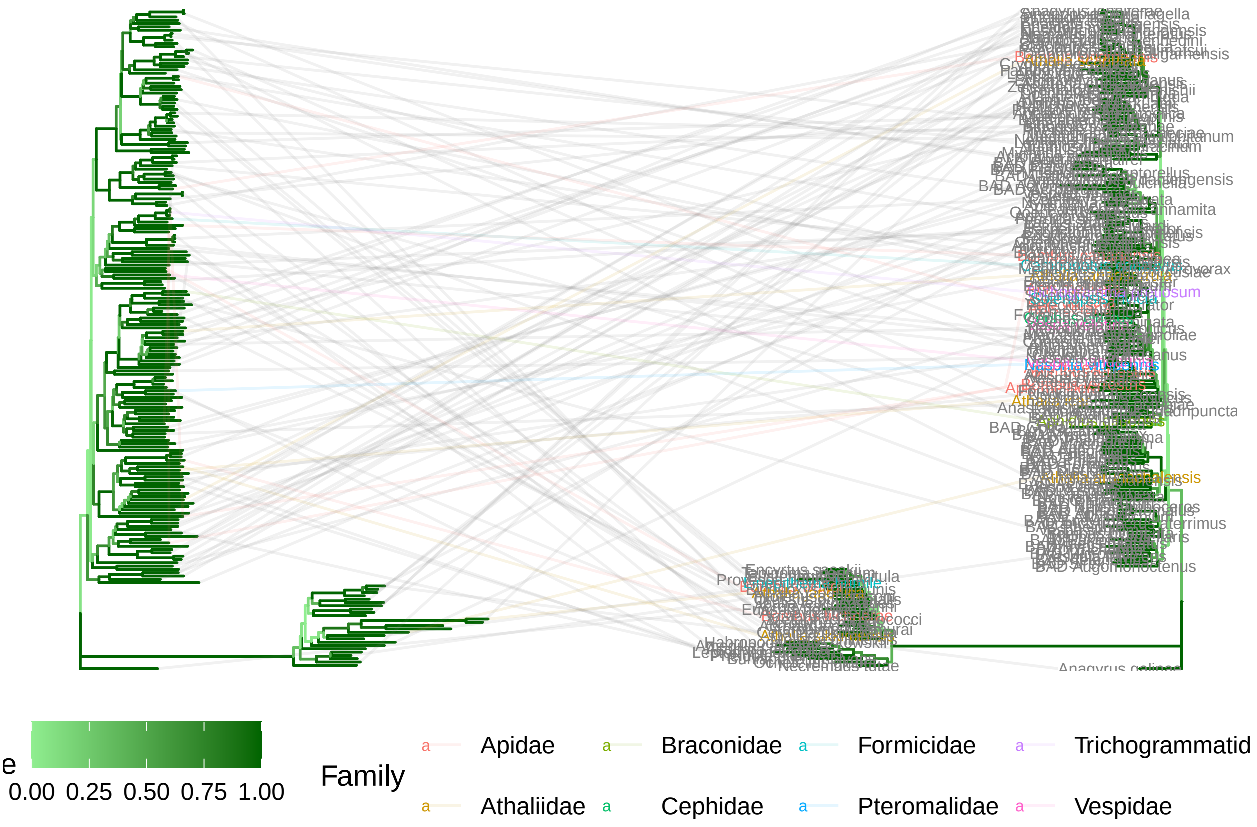
Compare phylo

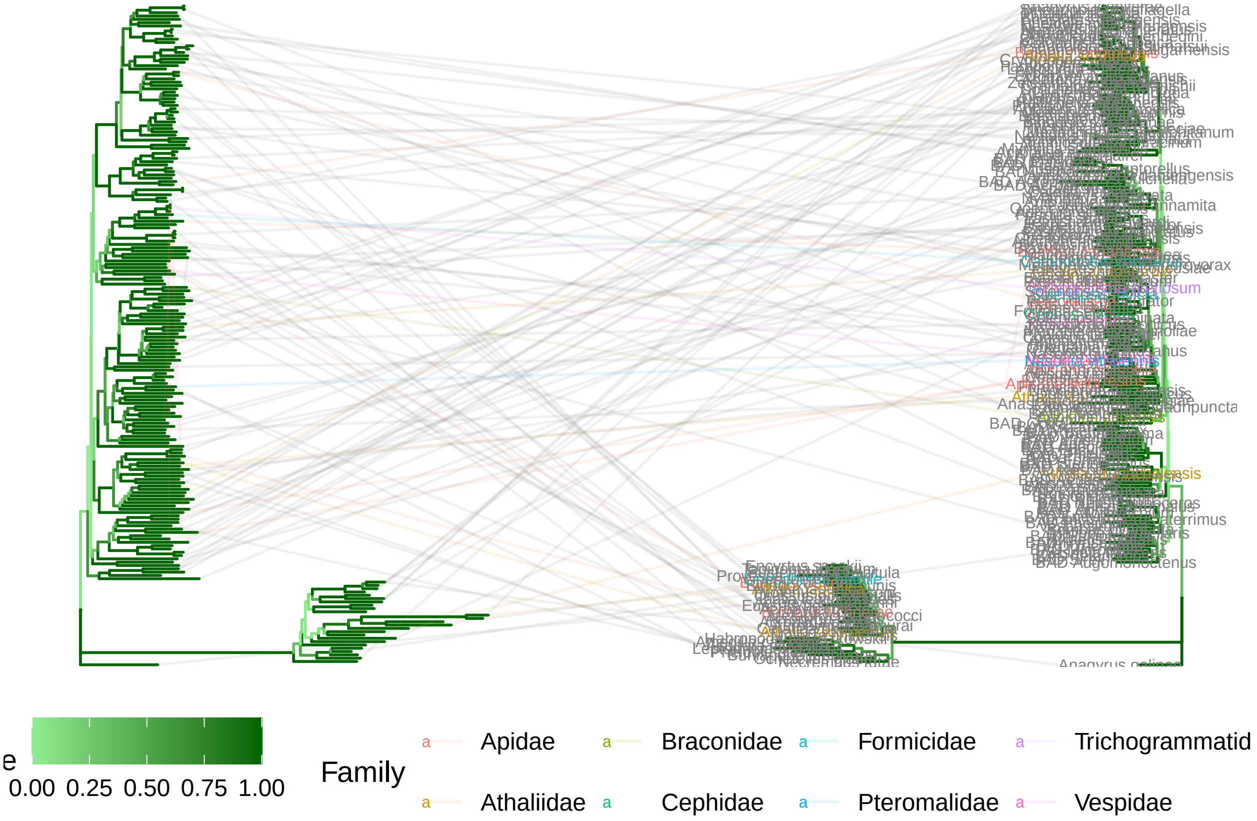
```
ggcompare <- function(d1, d2) {
  # main code from https://yulab-smu.top/treedata-book/chapter2.html
  d1$x <- mms(d1$x)
  d2$x <- mms(d2$x)

  ## reverse x-axis and
  d2$x <- max(d2$x) - d2$x + max(d1$x) + .5

  pp <- ggtree(d1, aes(color = as.numeric(label))) +
    geom_tree(data = d2, aes(color = as.numeric(label), x = x + .2)) +
    theme(legend.position = "bottom") +
    scale_color_gradient(name = "confidence", low = "lightgreen", high = "darkgreen", na.value = "darkg

    dd <- bind_rows(
      d1[d1$label %in% d2$label, ],
      d2[d2$label %in% d1$label, ]
    ) %>%
      filter(!is.na(label)) %>%
      subset(isTip)

    gg <- pp +
      ggnewscale::new_scale_color() +
      geom_line(
        aes(x, y,
          group = label,
          color = Family
        ),
        data = dd, alpha = .1
      ) +
      #scale_color_brewer("", palette = "Set2", na.value = "gray") +
      geom_tiplab(data = d2, hjust = 0, aes(x = x, color = Family), size = 2)
    print(gg)
    gg
}

ggcompare(
  d2,d1
)
```


Family
Ampulicidae
Apidae
Athaliidae
Braconidae
Formicidae
Pteromalidae
Trichogrammatidae
Vespidae
NA

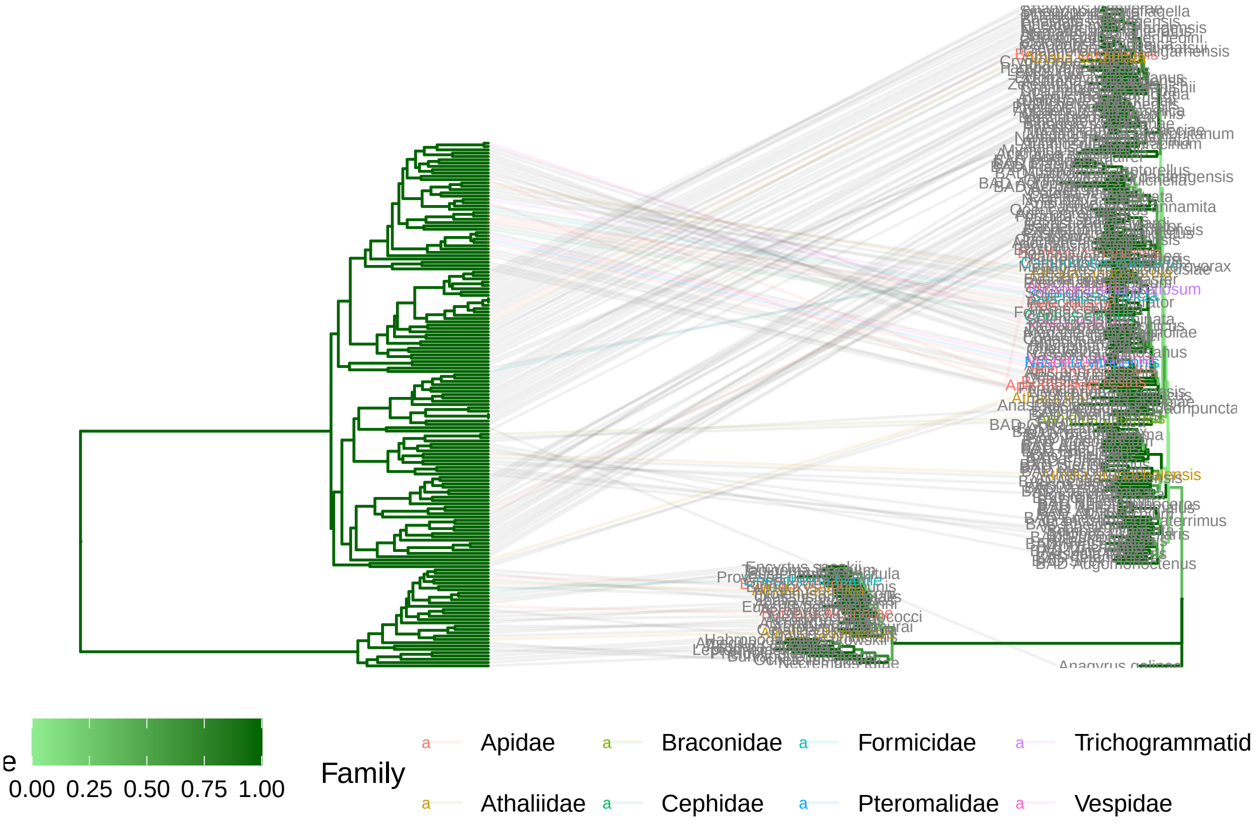


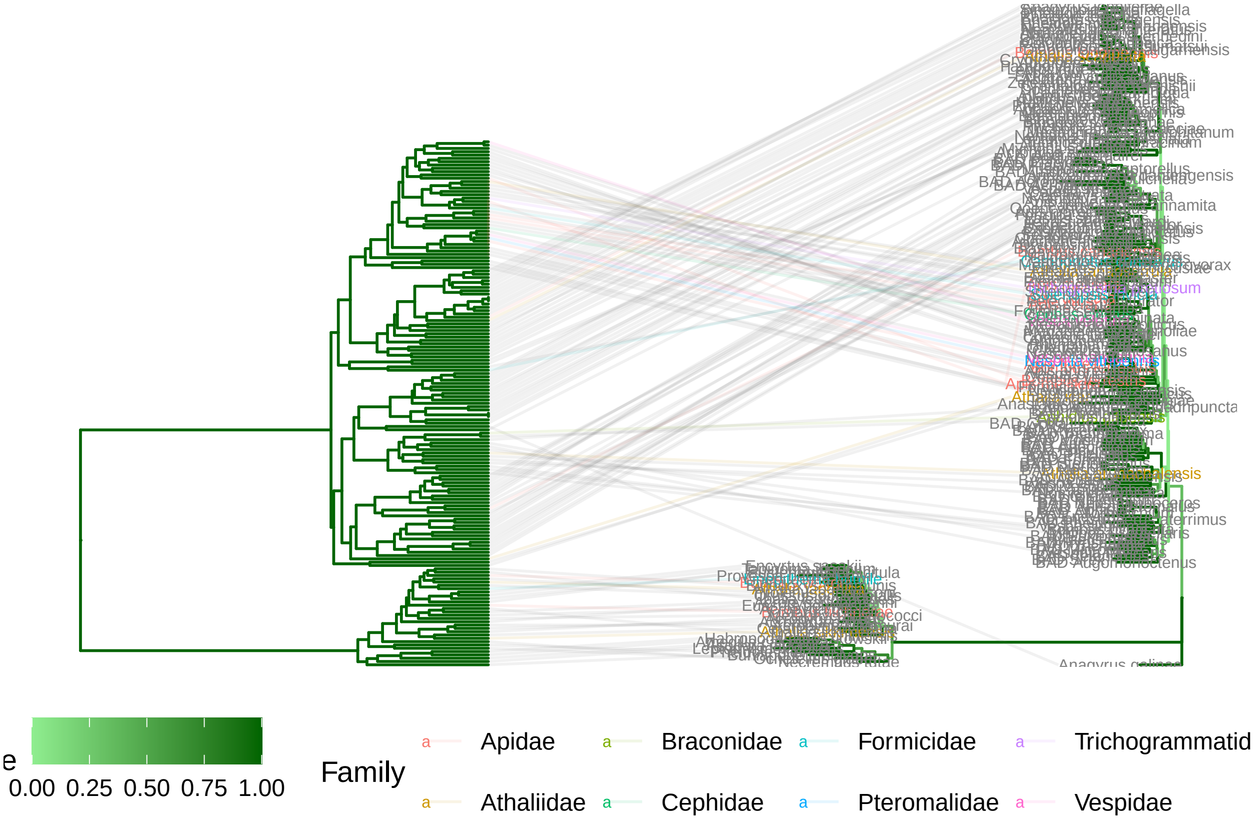
```
ggcompare(
  d3,d4
)
```


BAD Augomonoctenus
e
0.00
0.25
0.50
0.75
1.00
Family
Apidae
Braconidae
Formicidae
Trichogrammatid
Athaliidae
Cephidae
Pteromalidae
Vespidae

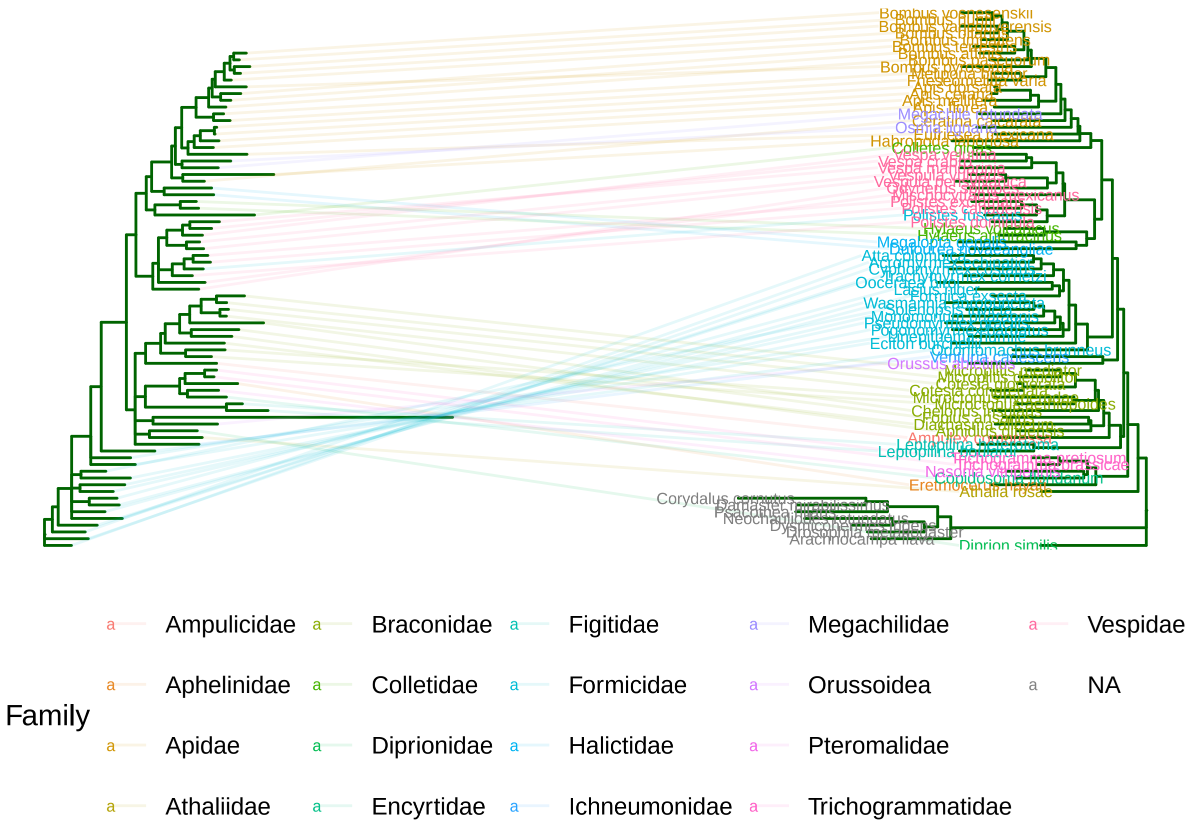


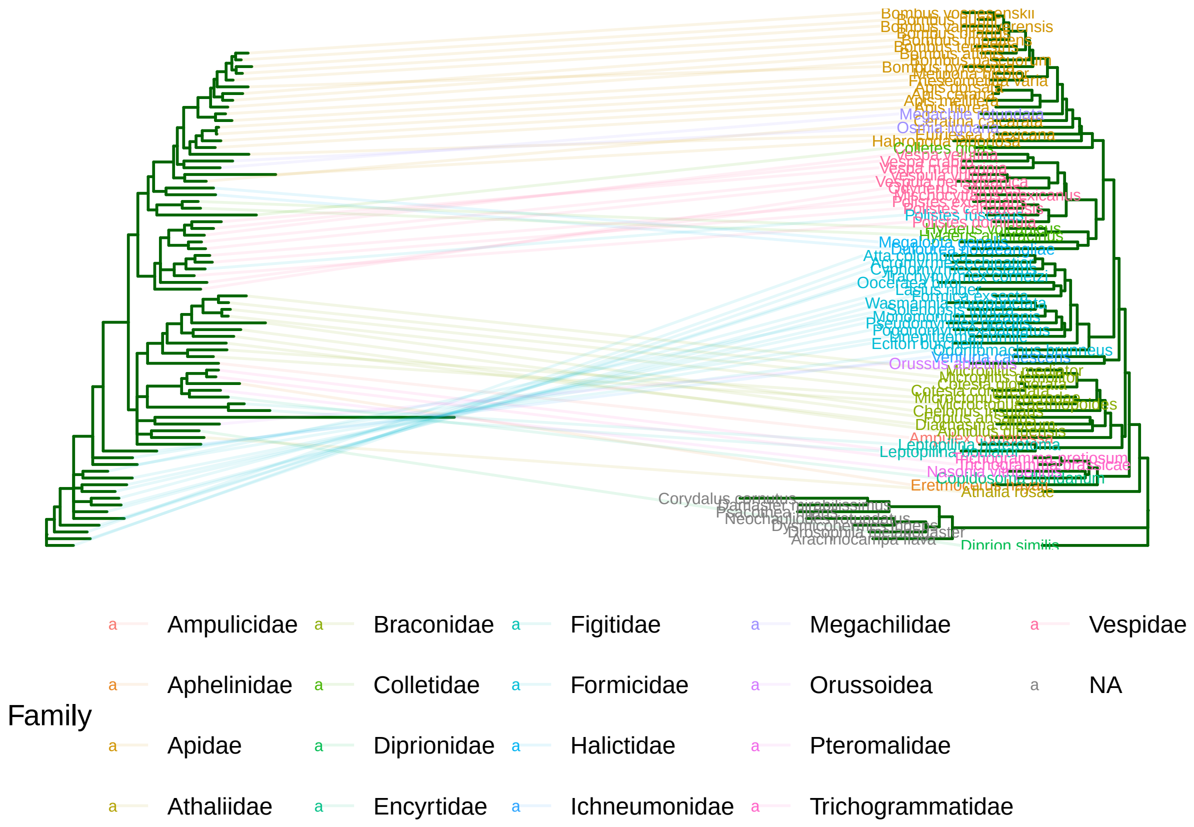
```
ggcompare(
  d2,d4
)
```


BAD Augomonoctenus
e
0.00
0.25
0.50
0.75
1.00
Family
a Apidae
a Athaliidae
a Braconidae
a Cephidae
a Formicidae
a Pteromalidae
a Trichogrammatid
a Vespidae



```
ggcompare(
  d5,d6
)
```

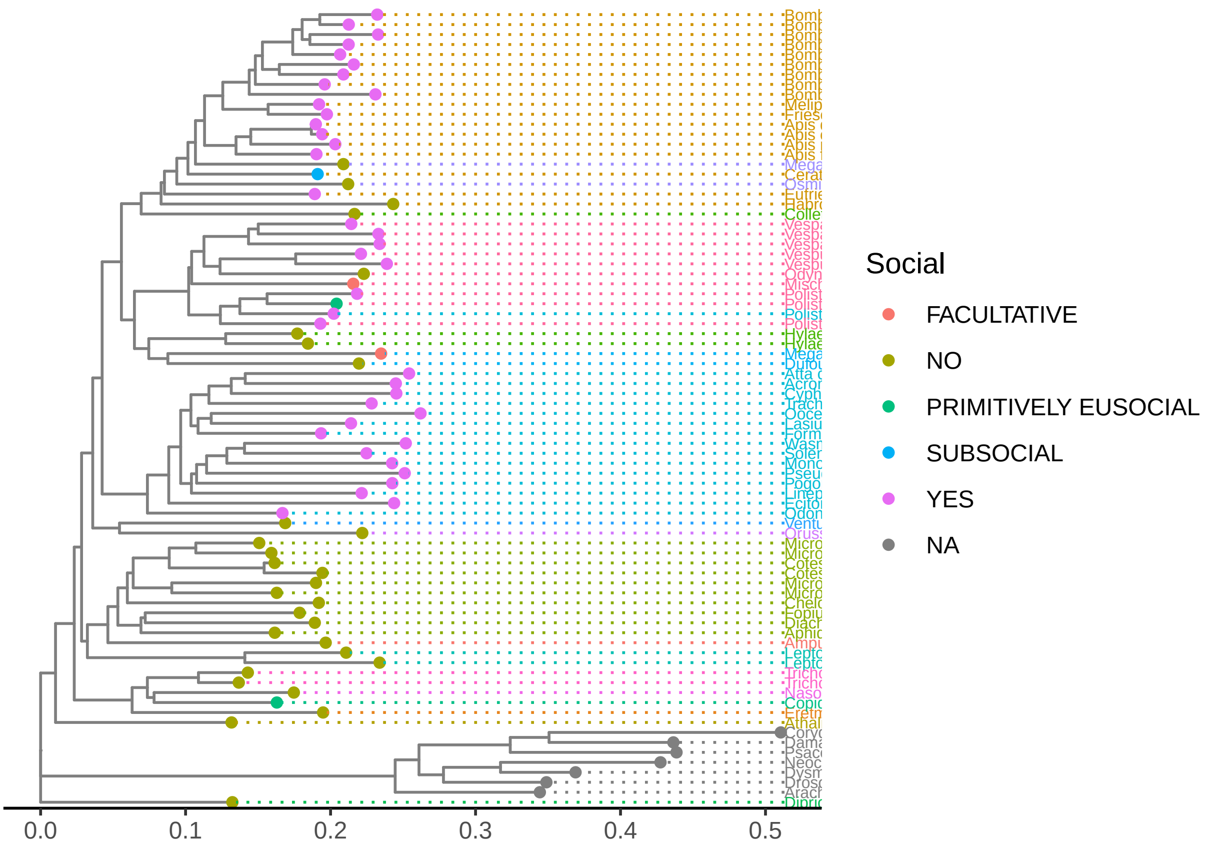

Bombus vosnesenskii
Bombus huntii
Bombus vancouverensis
Bombus bifarius
Bombus impatiens
Bombus terrestris
Bombus affinis
Bombus pascuorum
Bombus pyrosoma
Melipona bicolor
Frieseomelitta varia
Apis dorsata
Apis cerana
Apis mellifera
Apis florea
Megachile rotundata
Ceratina calcarata
Osmia lignaria
Eufriesea mexicana
Habropoda laboriosa
Colletes gigas
Vespa velutina
Vespa crabro
Vespa mandarinia
Vespula vulgaris
Vespula pensylvanica
Odynerus spinipes
Mischocyttarus mexicanus
Polistes exclamans
Polistes canadensis
Polistes fuscatus
Polistes dominula
Hylaeus volcanicus
Hylaeus anthracinus
Megalopta genalis
Dufourea novaeangliae
Atta colombica
Acromyrmex echinatior
Cyphomyrmex costatus
Trachymyrmex cornetzi
Ooceraea biroi
Lasius niger
Formica exsecta
Wasmannia auropunctata
Solenopsis invicta
Monomorium pharaonis
Pseudomyrmex gracilis
Pogonomyrmex barbatus
Linepithema humile
Eciton burchellii
Odontomachus brunneus
Venturia canescens
Orussus abietinus
Microplitis mediator
Microplitis demolitor
Cotesia glomerata
Cotesia congregata
Microctonus hyperodae
Microctonus aethiopoides
Chelonus insularis
Fopius arisanus
Diachasma alloeum
Aphidius gifuensis
Ampulex compressa
Leptopilina heterotoma
Leptopilina boulardi
Trichogramma pretiosum
Trichogramma brassicae
Nasonia vitripennis
Copidosoma floridanum
Eretmocerus hayati
Athalia rosae
Corydalus cornutus
Damaster mirabilissimus
Psacothea hilaris
Neochauliodes rotundatus
Dysmicohermes ingens
Drosophila melanogaster
Arachnocampa flava
Diprion similis
Family
a Ampulicidae
a Aphelinidae
a Apidae
a Athaliidae
a Braconidae
a Colletidae
a Diprionidae
a Encyrtidae
a Figitidae
a Formicidae
a Halictidae
a Ichneumonidae
a Megachilidae
a Orussoidea
a Pteromalidae
a Trichogrammatidae
a Vespidae
a NA



Eusociality

```
gg <- ggtree(d6, aes(color = as.numeric(label)),
            # layout = "circular"
             ) +
  ggnewscale::new_scale_color() +
  geom_tippoint(aes(color = Social)) +
  ggnewscale::new_scale_color() +
  geom_tiplab(size = 2,align=TRUE, aes(color = Family), show.legend = F) +
  ggnewscale::new_scale_color() +
  theme_tree2()
```

```
gg
```



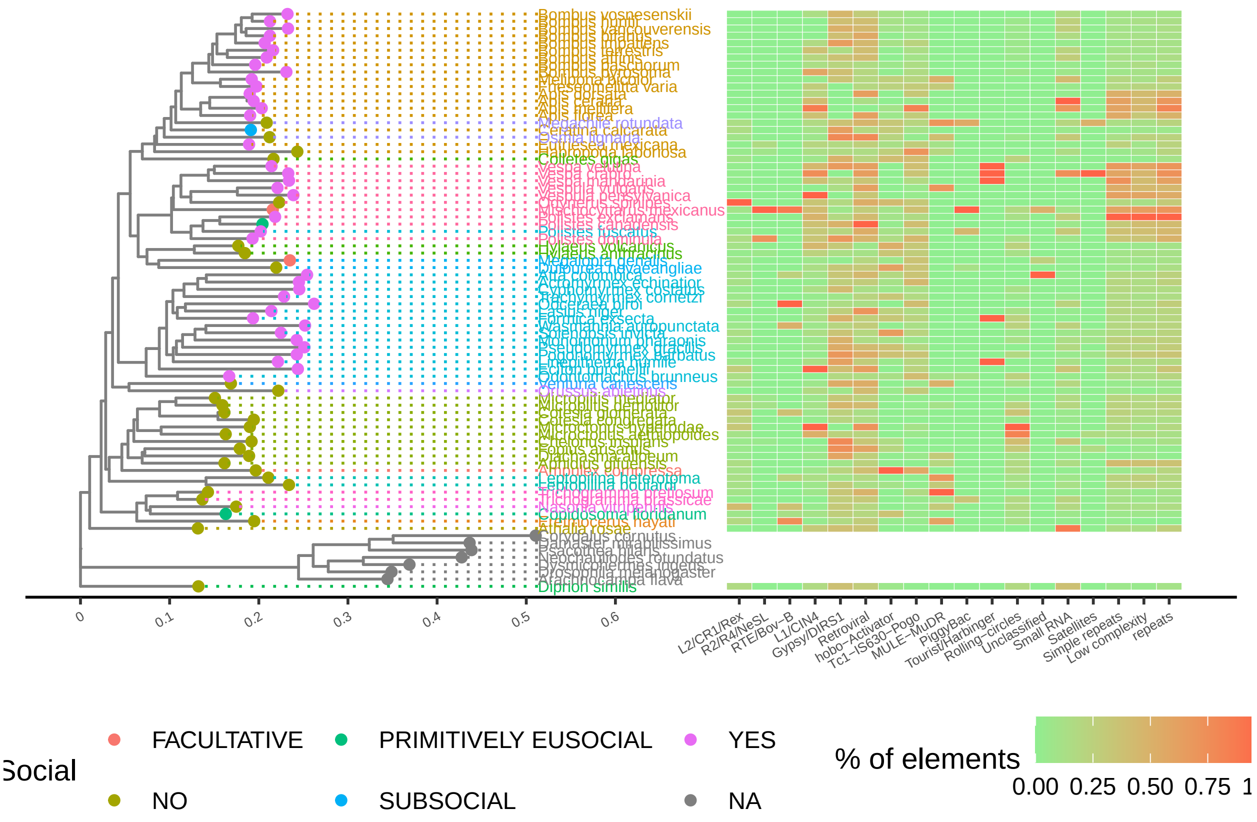
Add data

Heatmap

```
hmp <- (leg %>% column_to_rownames("label"))[,-c(1:4, 23)] %>%
  na.omit %>%
  apply(2, mms)

gg2 <- gheatmap(gg, hmp, offset=.2, width=1,
        colnames=F, legend_title="amount") +
  scale_fill_gradient("% of elements",
                      low = "lightgreen", high="tomato",
                      na.value = "transparent") +
    scale_x_ggtree() +
    scale_y_continuous(expand=c(0, 1)) +
  theme(axis.text.x = element_text(angle = 30, hjust= 1, size = 5),
        legend.position = "bottom")
```

```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```

```
gg2
```

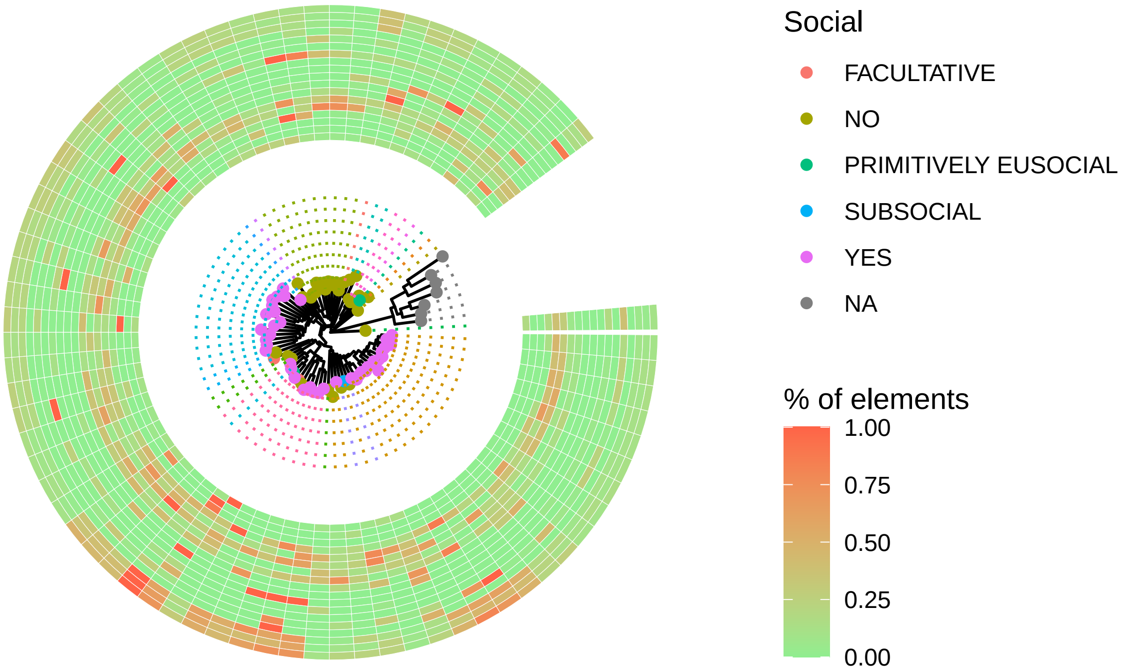


Hmp circ

```
gg <- ggtree(d6, layout = "circular",
            branch.length = "none"
             ) +
  ggnewscale::new_scale_color() +
  geom_tippoint(aes(color = Social)) +
  ggnewscale::new_scale_color() +
  geom_tiplab(data = d6, size = 2,align=TRUE, aes(color = Family, label = ""), show.legend = F) +
  ggnewscale::new_scale_color()


gg2 <- gheatmap(gg, hmp, offset=.2, width=1,
        colnames=F, legend_title="amount") +
  scale_fill_gradient("% of elements",
                      low = "lightgreen", high="tomato",
                      na.value = "transparent") +
    scale_y_continuous(expand=c(0, .1)) +
  theme_tree() +
  theme(axis.text = element_text())
```

```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```

```
gg2
```



Permanova

Cophenetic

```
hmp <- (leg %>% column_to_rownames("label"))[,-c(1:4, 23)] %>%
  na.omit
colnames(hmp) <- colnames(hmp) %>%
  str_replace_all("[/ -]", "_")

coph <- cophenetic.phylo(d6 %>% as.phylo) %>%
```

```r
  as.data.frame() %>%
  subset(rownames(.) %in% rownames(hmp)) %>% t %>%
  subset(rownames(.) %in% rownames(hmp)) %>%
  as.data.frame()

hmp <- hmp %>%
  subset(rownames(.) %in% rownames(coph)) %>%
  as.data.frame()

#colnames(hmp) %>% paste0(collapse = " + ") %>%  cat

hmp3 <- hmp %>% apply(2, mms) %>% as.data.frame()
for ( i in colnames(hmp) ) {
  adform <- as.formula(paste0("coph ~ ", i))
  adres <- adonis2(adform,data = hmp3)
  pv1 = adres$`Pr(>F)`[1]
  pv2 = rep("*", -log10(adres$`Pr(>F)`[1]) %>% round)[-1] %>%
    paste0(collapse = "")

  cat(i,
      "\nR2:", adres$R2[1] %>% round(2),
      "\np:",adres$`Pr(>F)`[1] %>% round(3), pv2)
  cat("\n####\n\n")
}
```

```
## L2_CR1_Rex
## R2: 0.02
## p: 0.04
## ####
##
## R2_R4_NeSL
## R2: 0.02
## p: 0.268
## ####
##
## RTE_Bov_B
## R2: 0.02
## p: 0.176
## ####
##
## L1_CIN4
## R2: 0.01
## p: 0.699
## ####
##
## Gypsy_DIRS1
## R2: 0.01
## p: 0.381
## ####
##
## Retroviral
## R2: 0.02
## p: 0.063
## ####
##
## hobo_Activator
## R2: 0.01
## p: 0.594
## ####
##
## Tc1_IS630_Pogo
## R2: 0.01
## p: 0.334
## ####
##
## MULE_MuDR
## R2: 0.02
## p: 0.05
## ####
##
## PiggyBac
## R2: 0.02
## p: 0.05
## ####
##
## Tourist_Harbinger
## R2: 0.02
## p: 0.06
## ####
##
## Rolling_circles
## R2: 0.03
## p: 0.007 *
## ####
##
## Unclassified
## R2: 0.01
## p: 0.615
## ####
##
## Small_RNA
## R2: 0.02
## p: 0.259
## ####
##
## Satellites
## R2: 0.01
## p: 0.384
## ####
##
## Simple_repeats
## R2: 0.03
## p: 0.02 *
## ####
##
## Low_complexity
## R2: 0.03
## p: 0.009 *
## ####
##
## repeats
## R2: 0.03
## p: 0.02 *
## ####
```

Eusoc

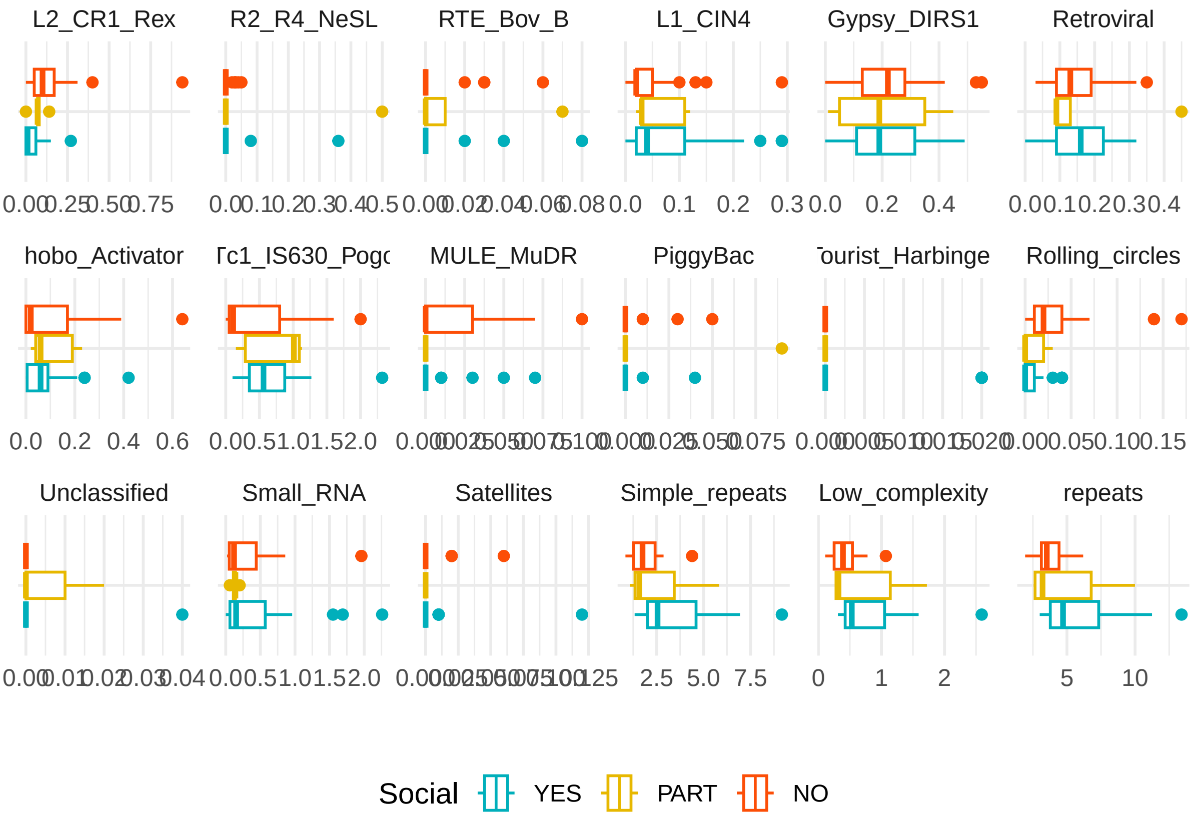
```
socvec <- leg[match(rownames(hmp3),leg$label),"Social"] %>%
  mutate(Social = ifelse(
    Social %in% c("YES", "NO"),
    Social,
    "PART"
  ))

res <- adonis2(hmp3~Social,
               data = socvec,
               #method = "euclidian",
               by = "onedf")

res
```

```
## Permutation test for adonis under reduced model
## Sequential test for contrasts
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = hmp3 ~ Social, data = socvec, by = "onedf")
##            Df SumOfSqs      R2      F Pr(>F)
## SocialPART  1   0.1535 0.01536 1.1811  0.275
## SocialYES   1   0.7404 0.07409 5.6954  0.001 ***
## Residual   70   9.1000 0.91055
## Total      72   9.9939 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

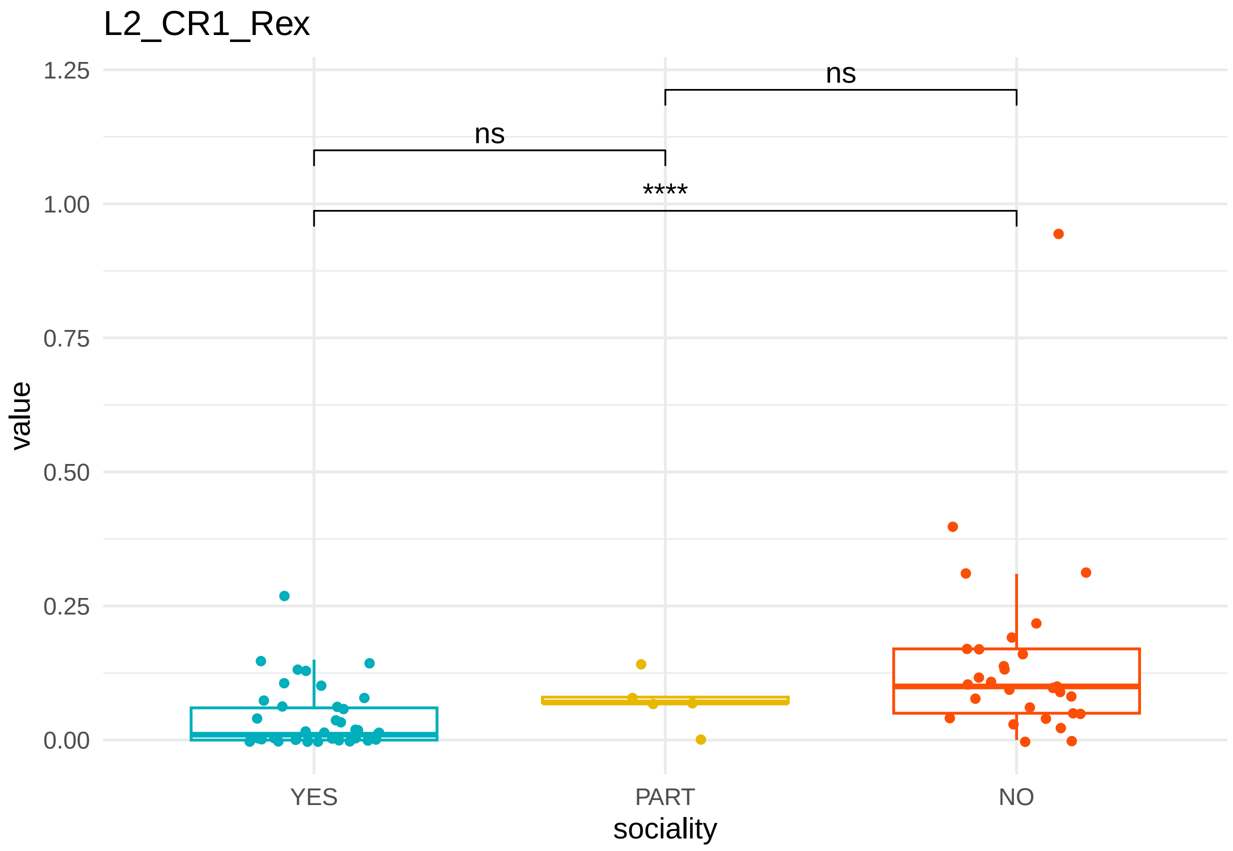
Barplots

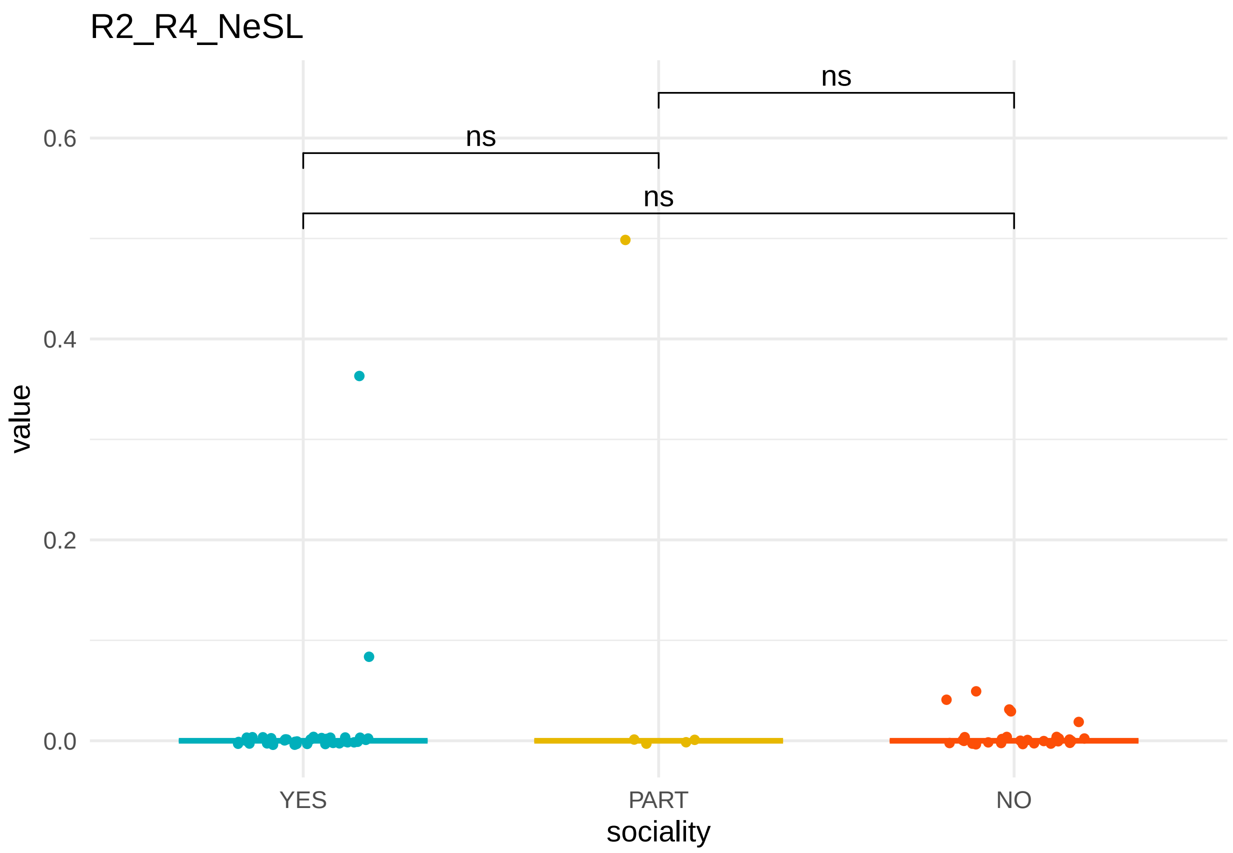
```
hmp2 <- hmp %>%
  rownames_to_column("sp") %>%
  mutate(Social = socvec$Social) %>%
  pivot_longer(cols = -c(1, 20))

gg <- hmp2 %>%
  mutate(name = fct_inorder(name)) %>%
  mutate(Social = fct_rev(Social)) %>%
  ggplot(aes(color = Social, y = "", x = value)) +
  geom_boxplot() +
```

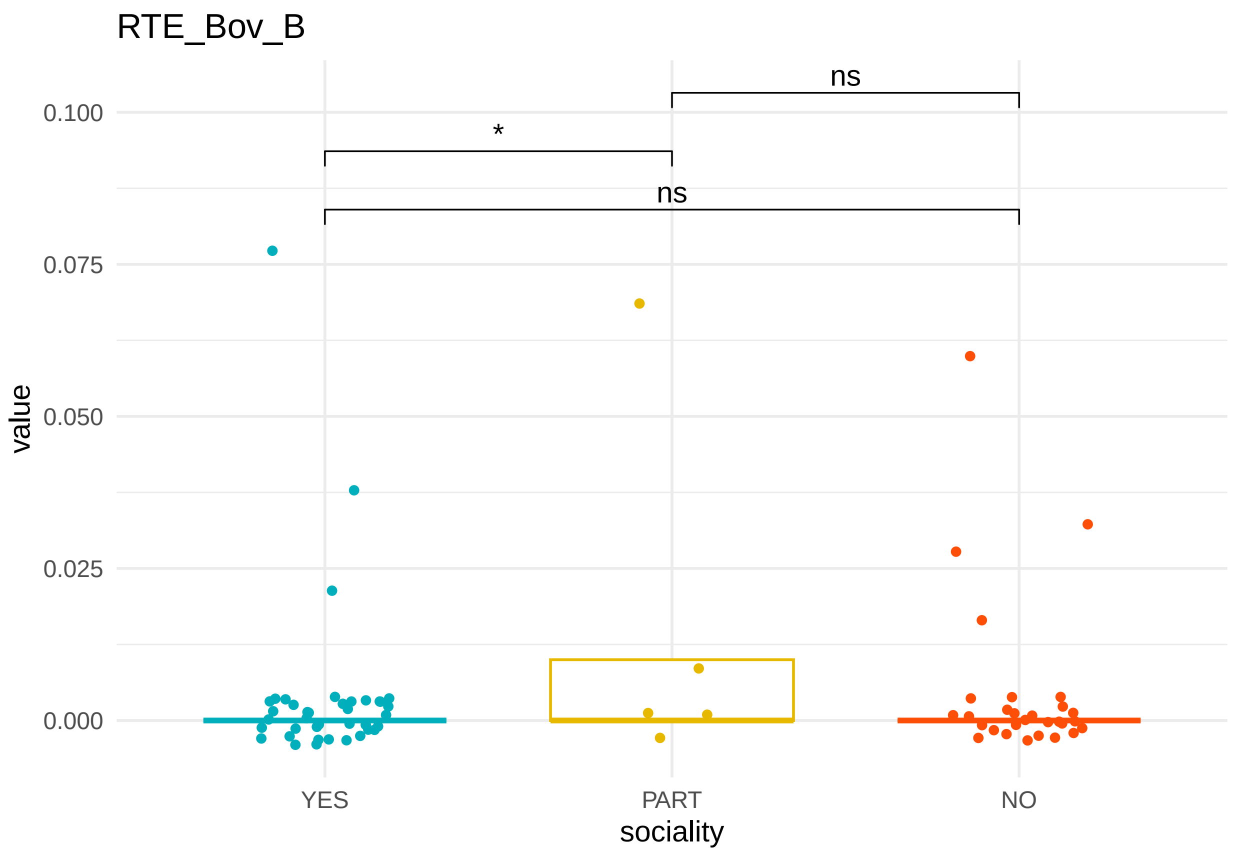
```r
  facet_wrap(~name, scales = "free", nrow = 3) +
  theme_minimal() +
  scale_color_manual(values = c("#00AFBB", "#E7B800", "#FC4E07")) +
  xlab("") + ylab("") +
  theme(legend.position = "bottom",
        strip.clip = "none")

gg
```

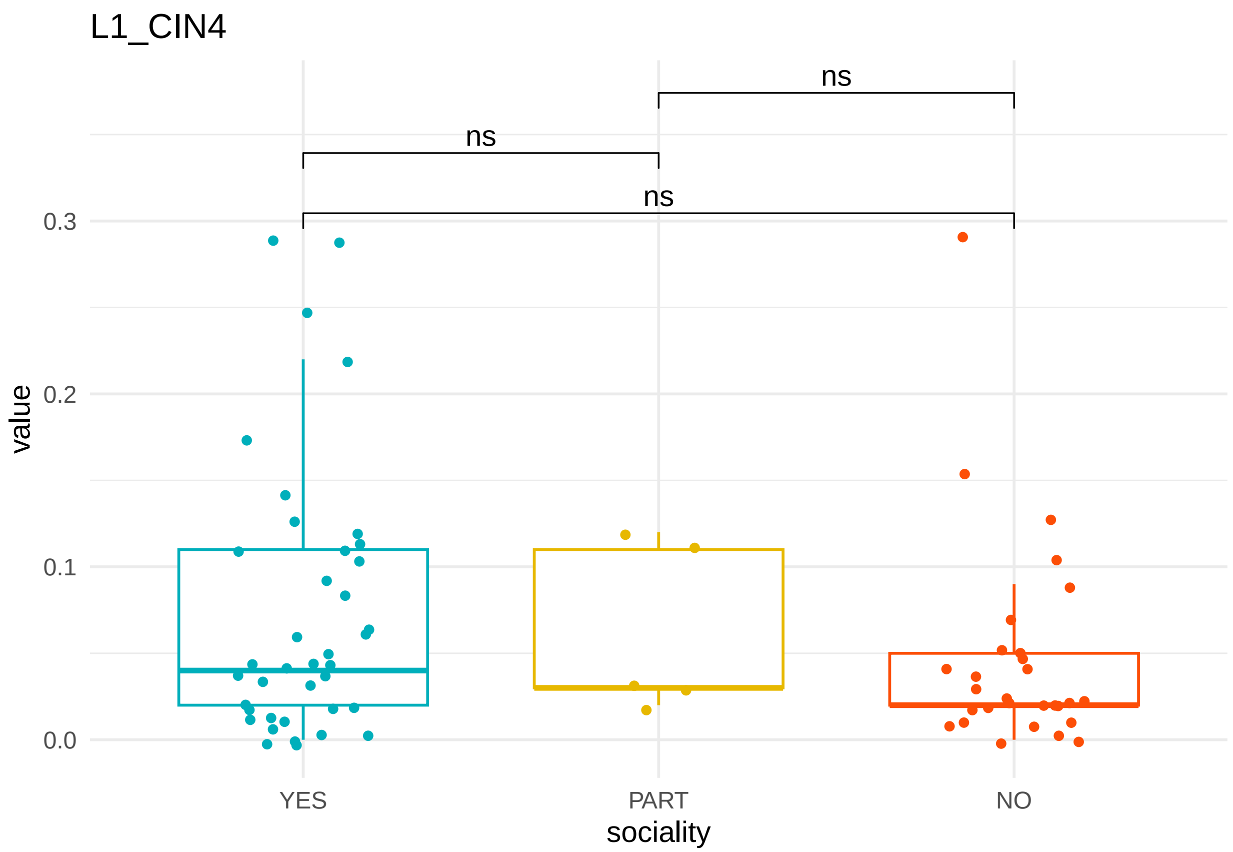


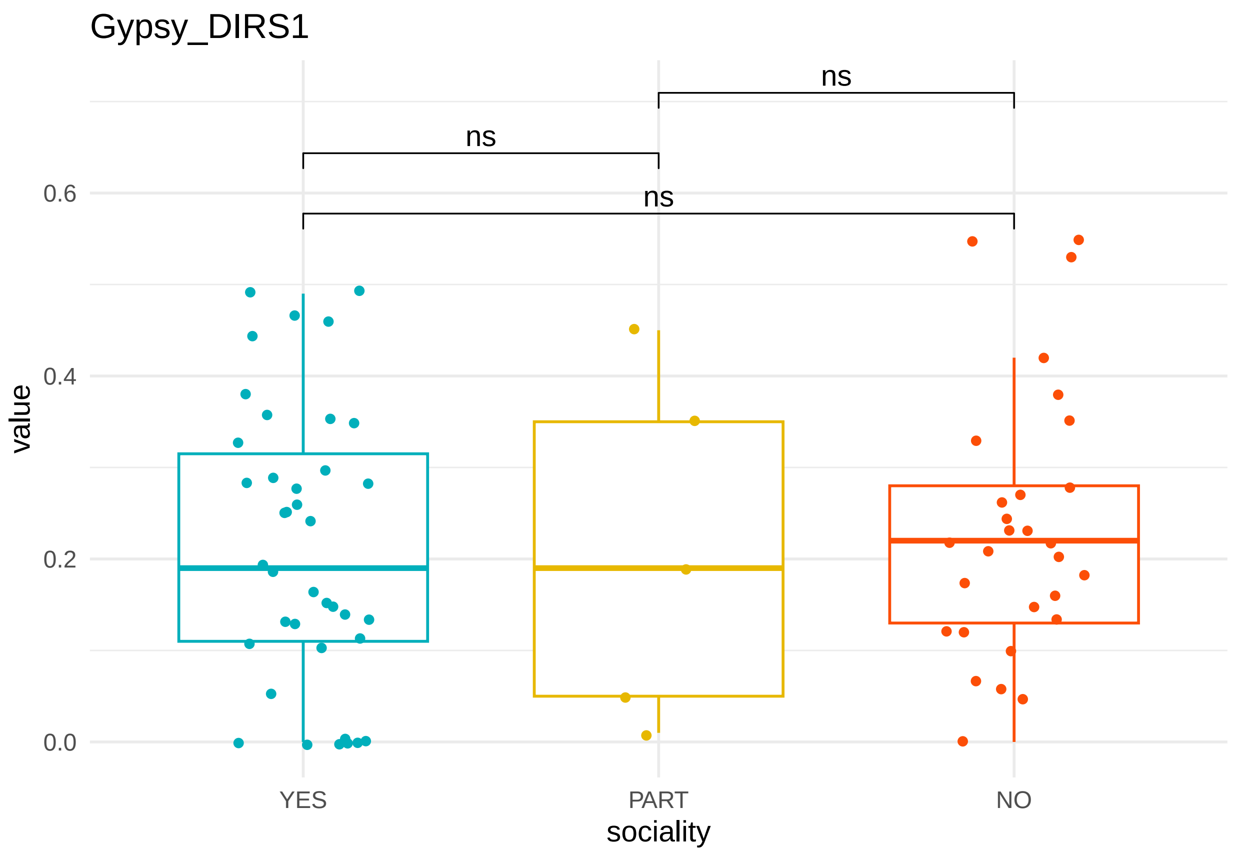
```r
cmp <- list(c("YES", "NO"), c("YES", "PART"), c("NO", "PART"))

for (i in unique(hmp2$name)) {
  ggp <- hmp2 %>%
    mutate(Social = fct_rev(Social)) %>%
    subset(name == i) %>%
    ggboxplot(
      x = "Social", y = "value",
      color = "Social",
      palette = c("#00AFBB", "#E7B800", "#FC4E07"),
      add = "jitter", shape = "name"
    ) +
    # facet_wrap(~name, scales = "free") +
    theme_minimal() +
    theme(
      legend.position = "none",
      # axis.text.x = element_blank()
    ) +
    stat_compare_means(comparisons = cmp, label = "p.signif") +
    ggtitle(i) +
    xlab("sociality")

  print(ggp)

  # ggsave(paste0("pubr/",i,".png"), ggp, width = 1000, height = 1000, units = "px")
}
```

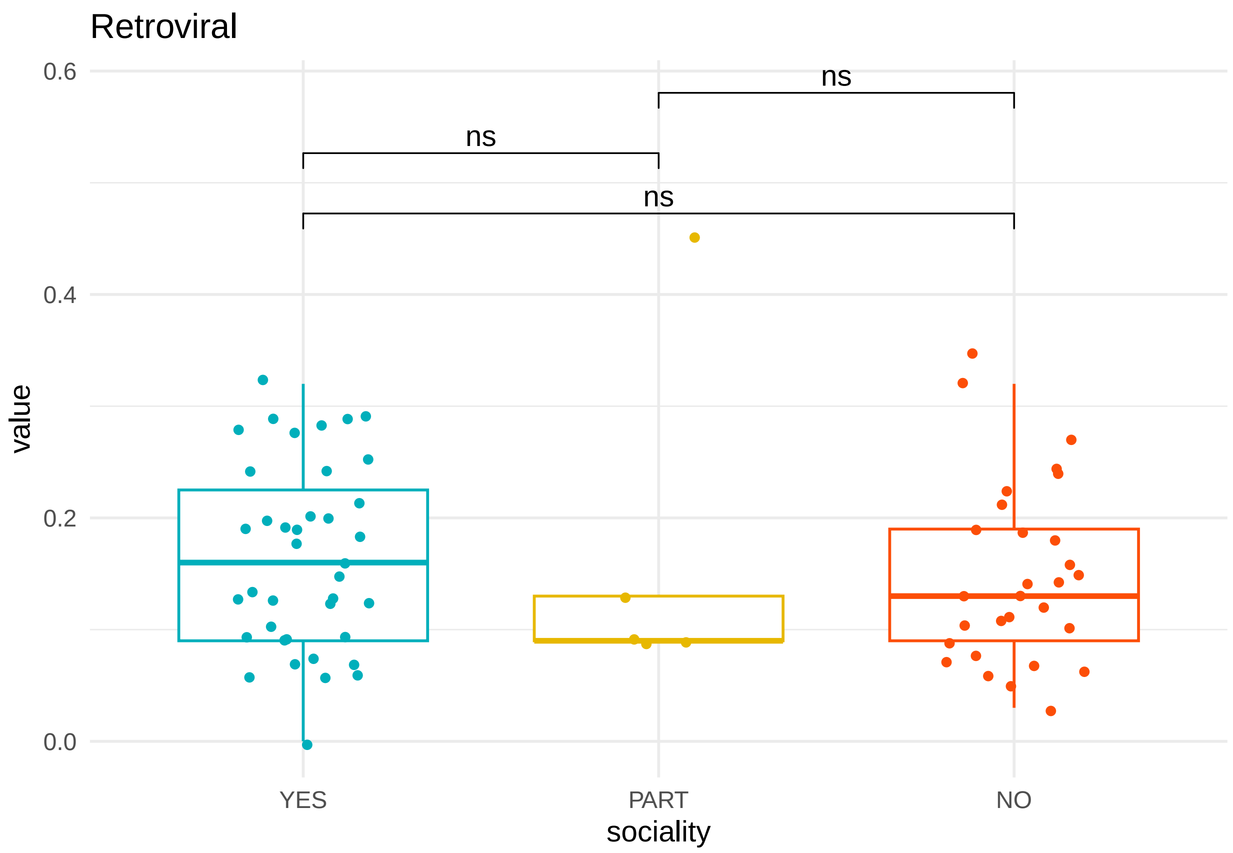


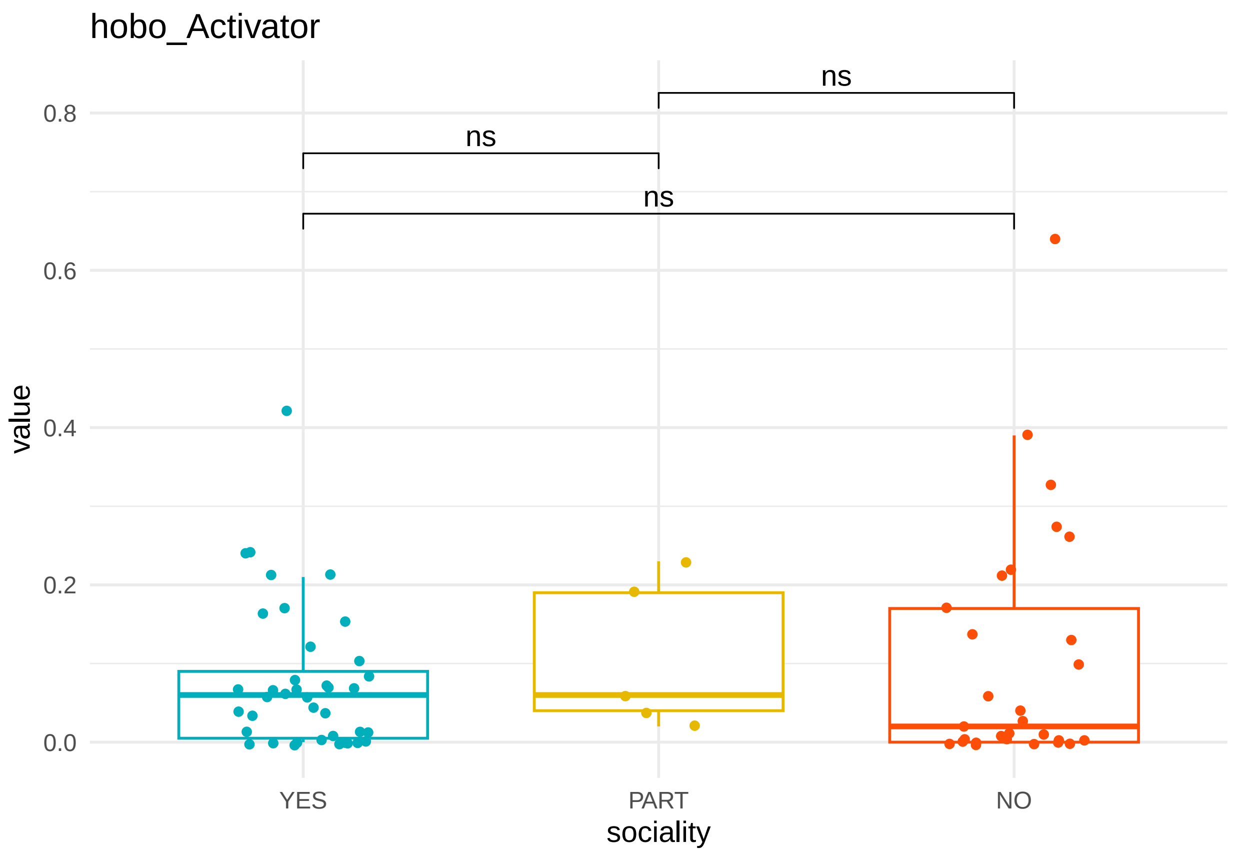

R2_R4_NeSL
ns
ns
ns
value
0.6
0.4
0.2
0.0
YES
PART
NO
sociality

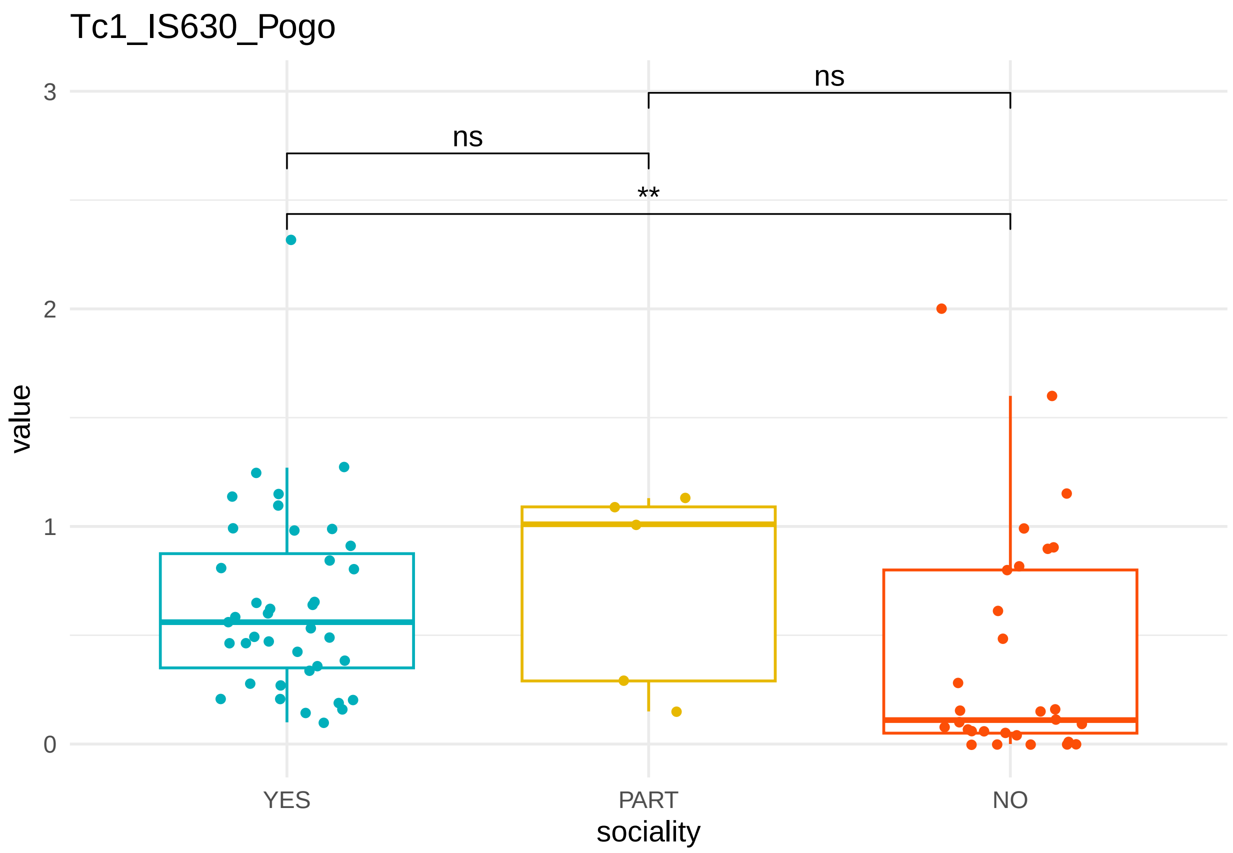

RTE_Bov_B
ns
*
ns
0.100
0.075
0.050
0.025
0.000
value
YES
PART
NO
sociality

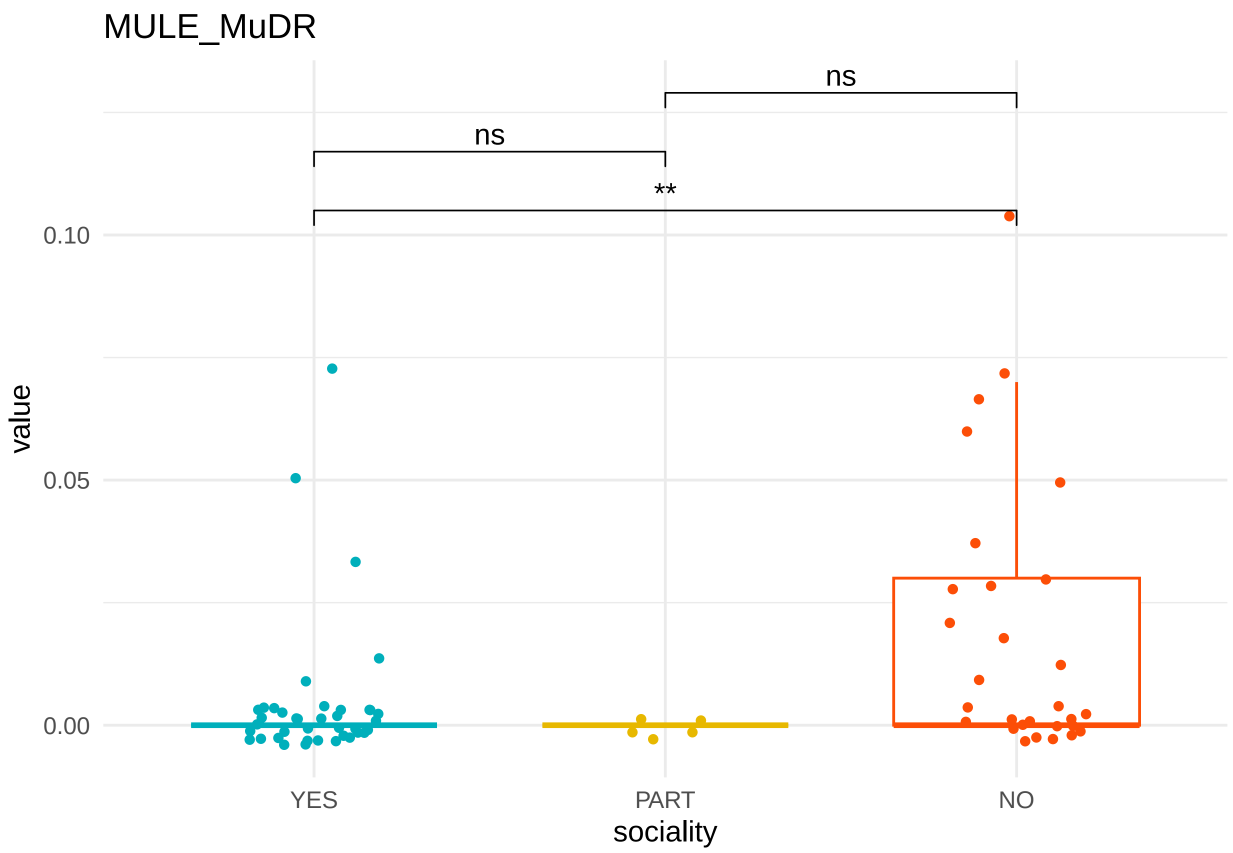

L1_CIN4
ns
ns
ns
0.3
0.2
0.1
0.0
value
YES
PART
NO
sociality

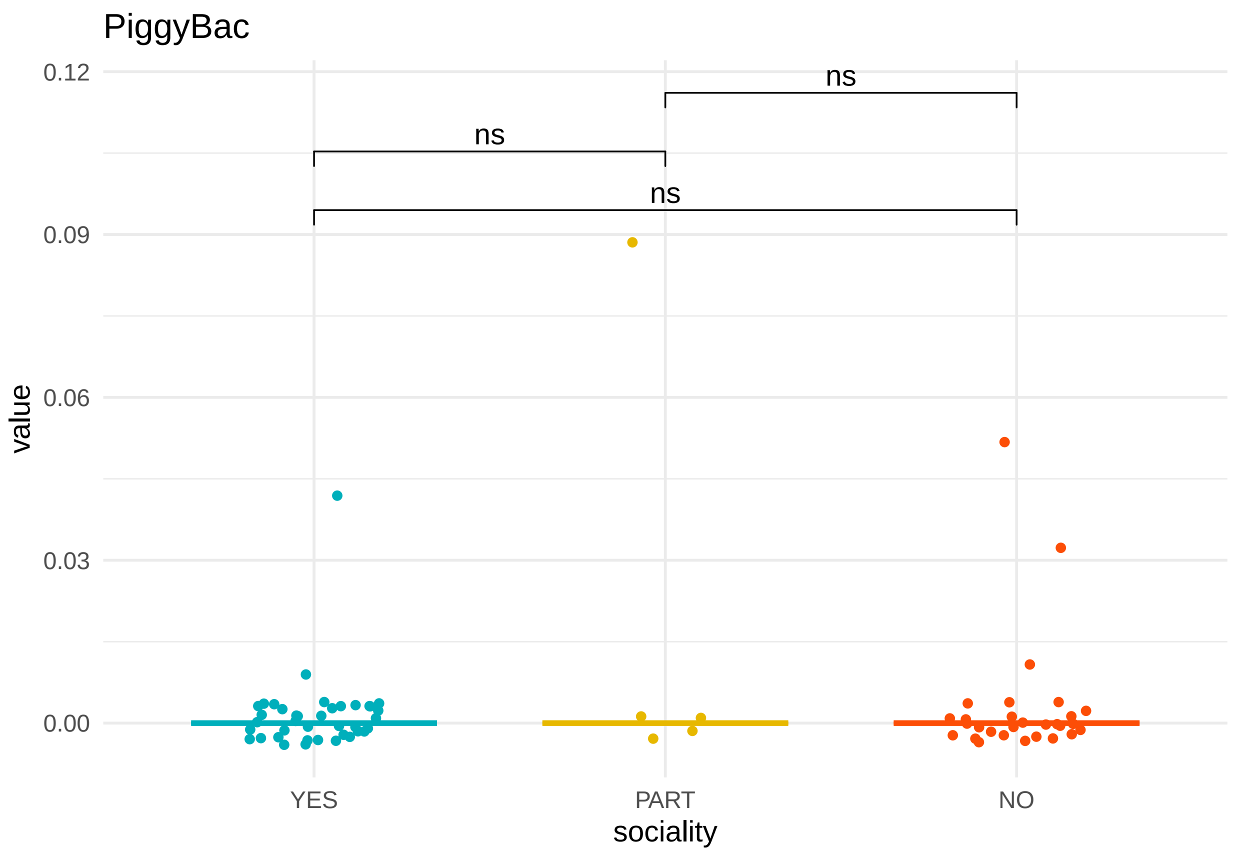

Gypsy_DIRS1
ns
ns
ns
0.6
0.4
0.2
0.0
value
YES
PART
NO
sociality

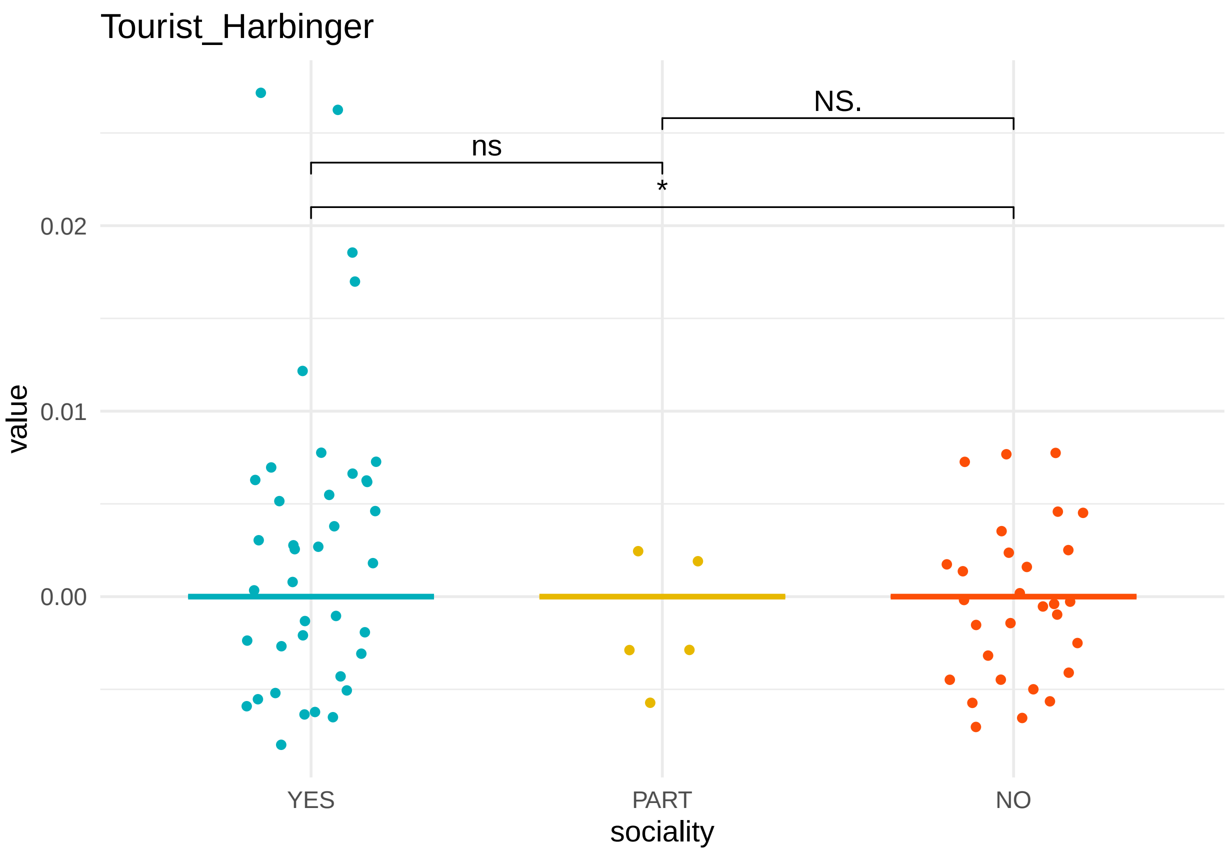

Retroviral
0.6
0.4
0.2
0.0
value
ns
ns
ns
YES
PART
NO
sociality

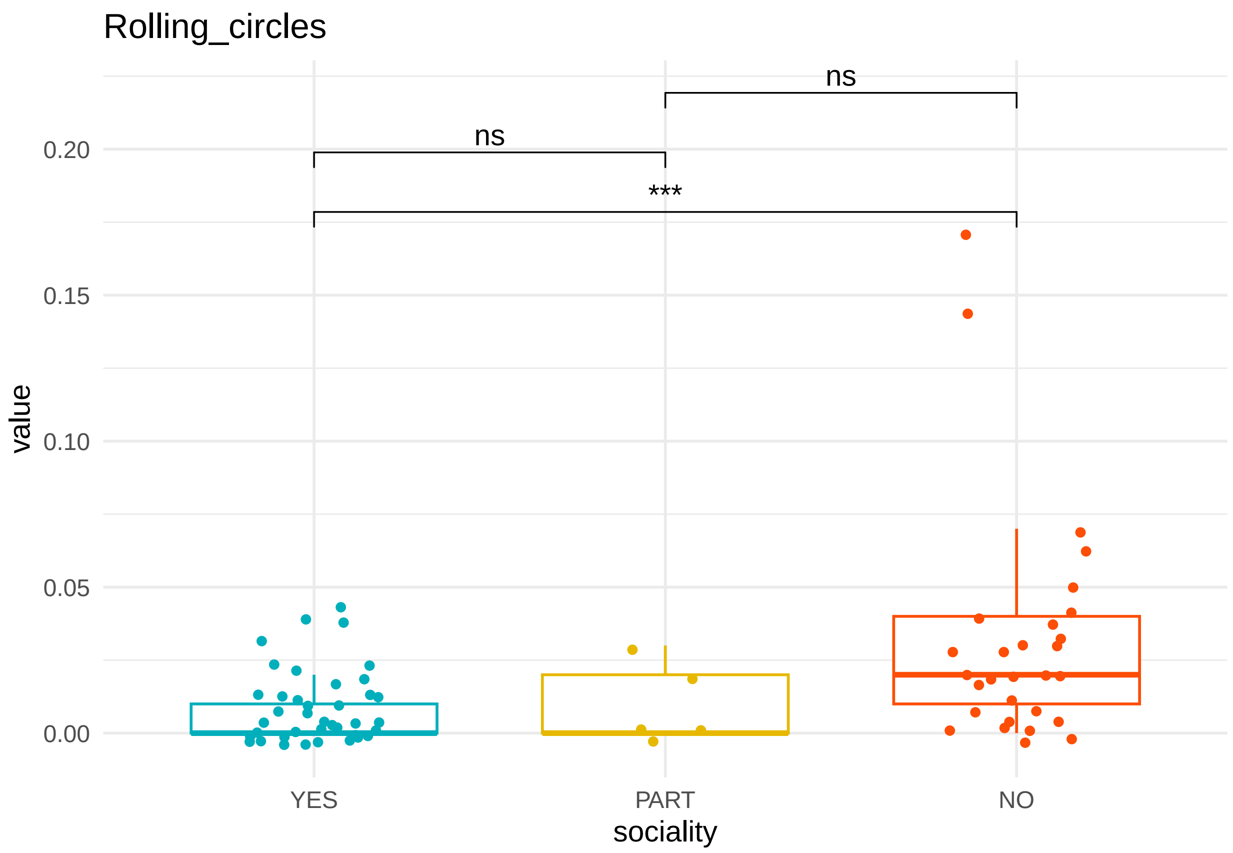

hobo_Activator
value
0.8
0.6
0.4
0.2
0.0
ns
ns
ns
YES
PART
NO
sociality

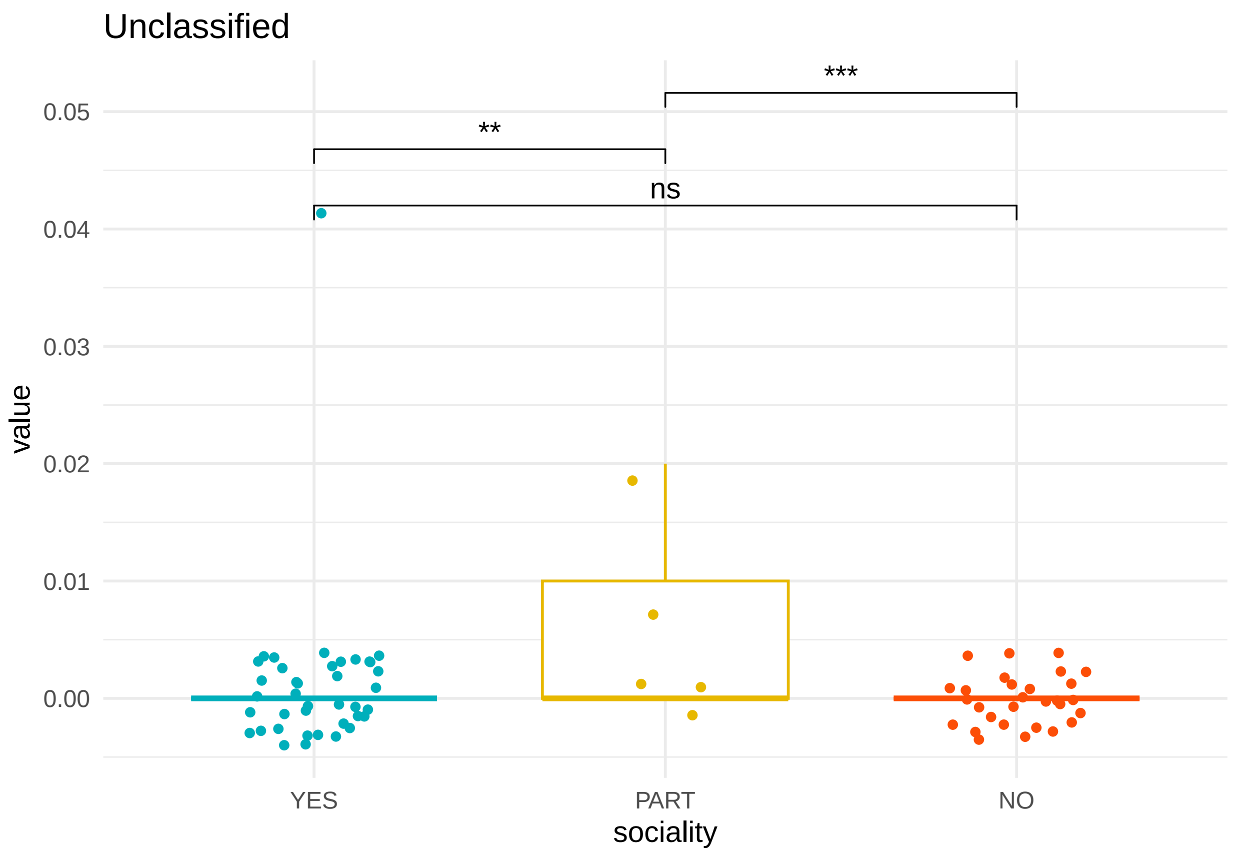

Tc1_IS630_Pogo
value
3
2
1
0
ns
ns
**
YES
PART
NO
sociality

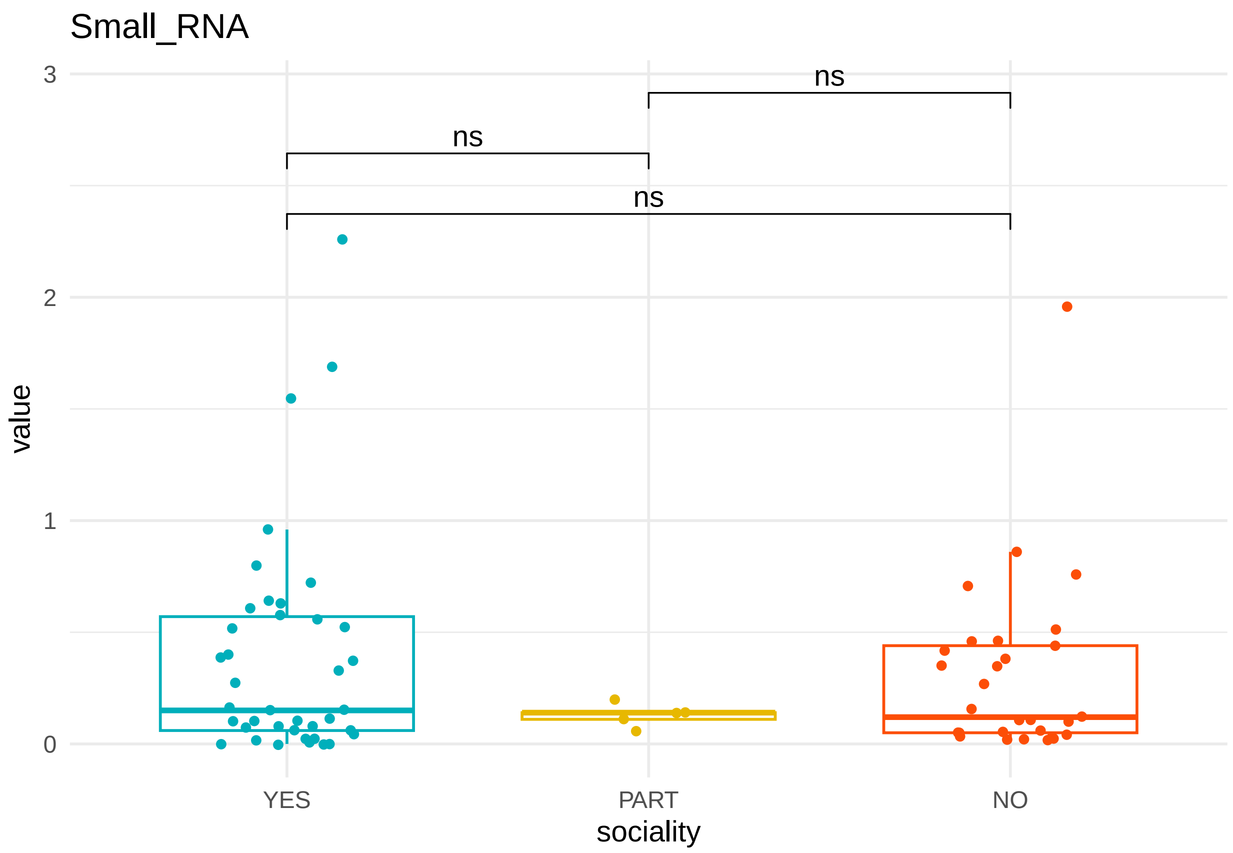

MULE_MuDR
ns
ns
**
value
0.10
0.05
0.00
YES
PART
NO
sociality

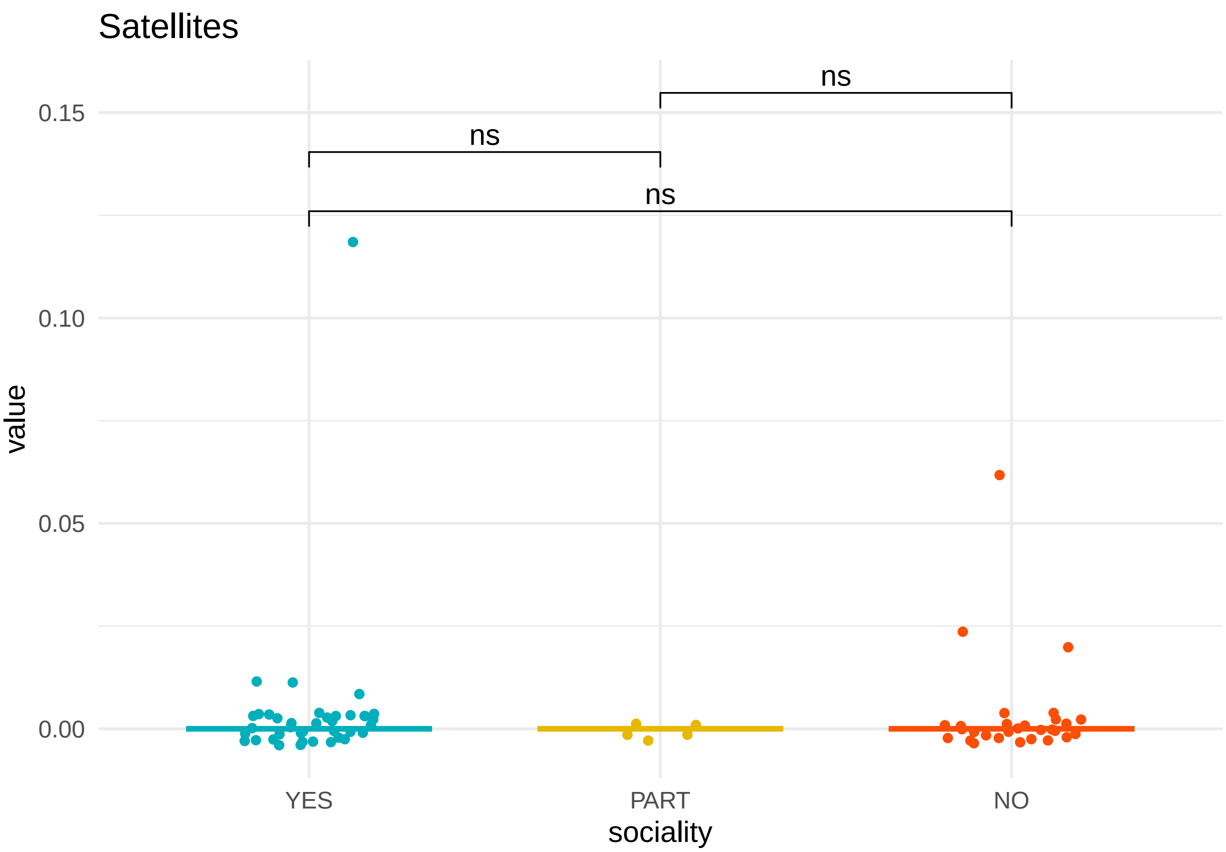

PiggyBac
0.12
0.09
0.06
0.03
0.00
value
ns
ns
ns
YES
PART
NO
sociality

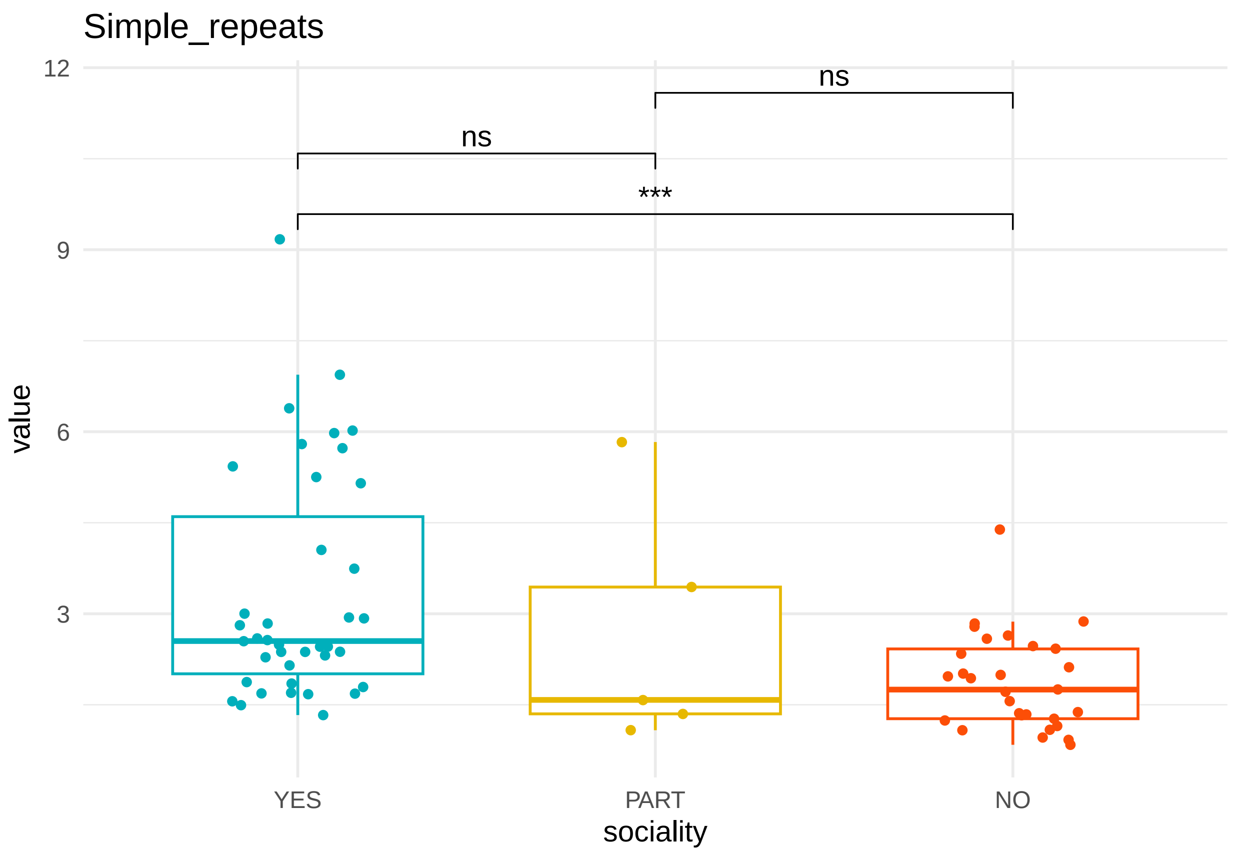

Tourist_Harbinger
NS.
ns
*
0.02
0.01
0.00
value
YES
PART
NO
sociality

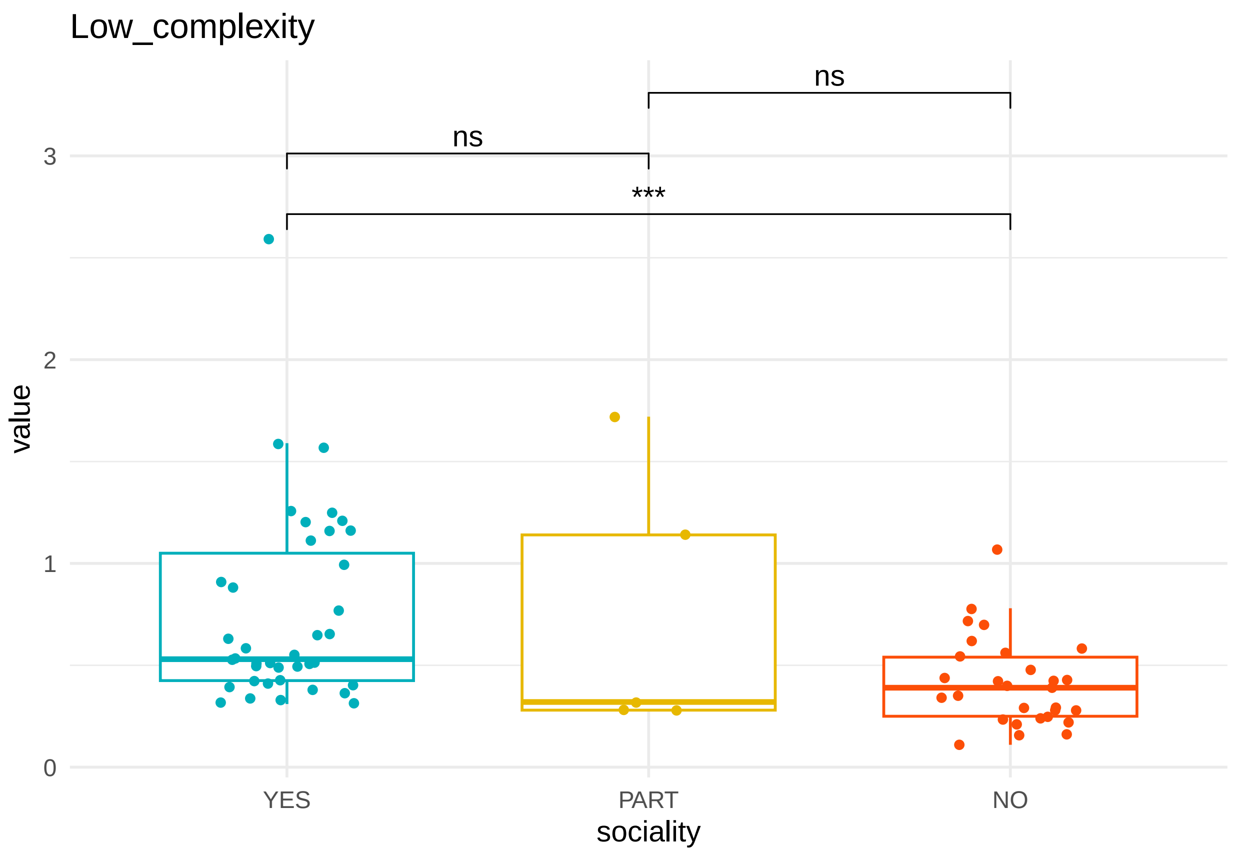

Rolling_circles
value
0.20
0.15
0.10
0.05
0.00
ns
ns
***
YES
PART
NO
sociality

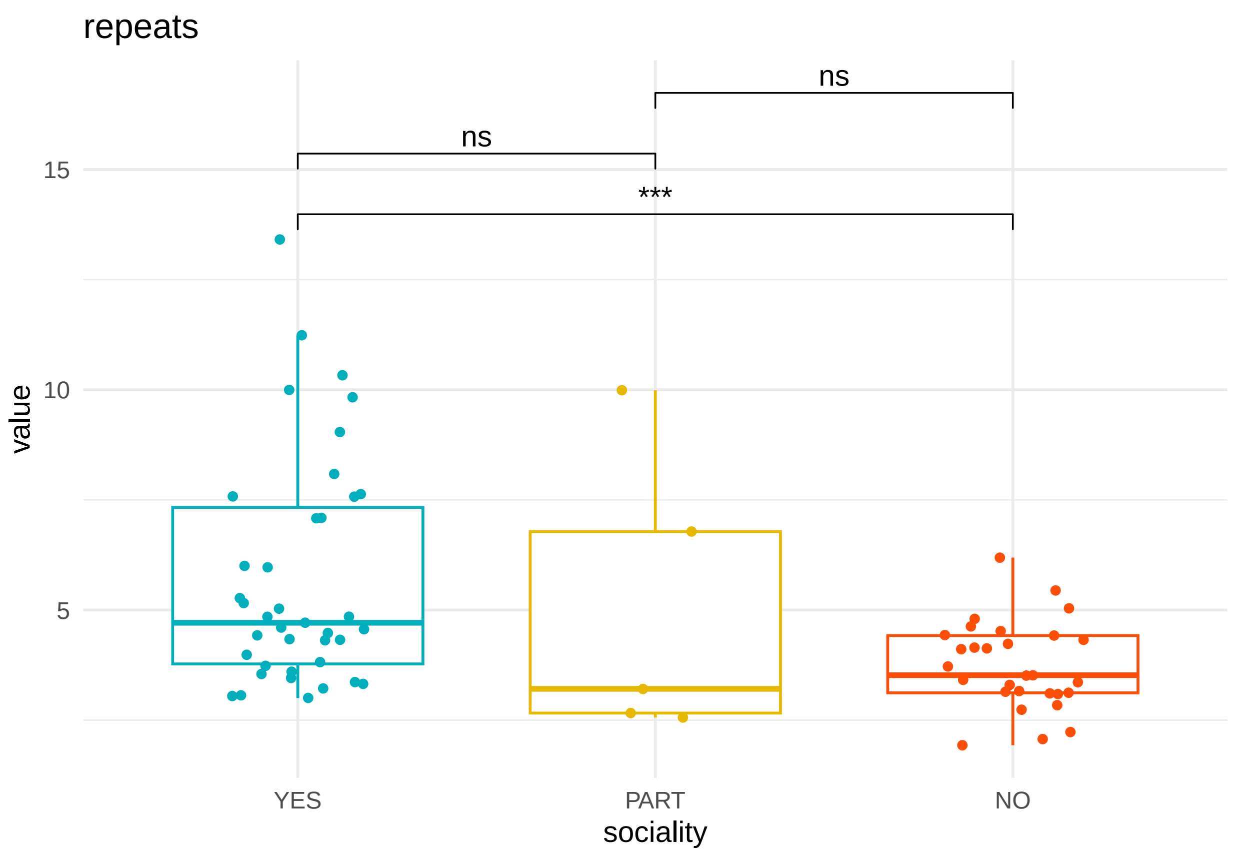

Unclassified
***
**
ns
0.05
0.04
0.03
0.02
0.01
0.00
value
YES
PART
NO
sociality


Small_RNA
value
3
2
1
0
ns
ns
ns
YES
PART
NO
sociality


Satellites
0.15
0.10
0.05
0.00
value
ns
ns
ns
YES
PART
NO
sociality

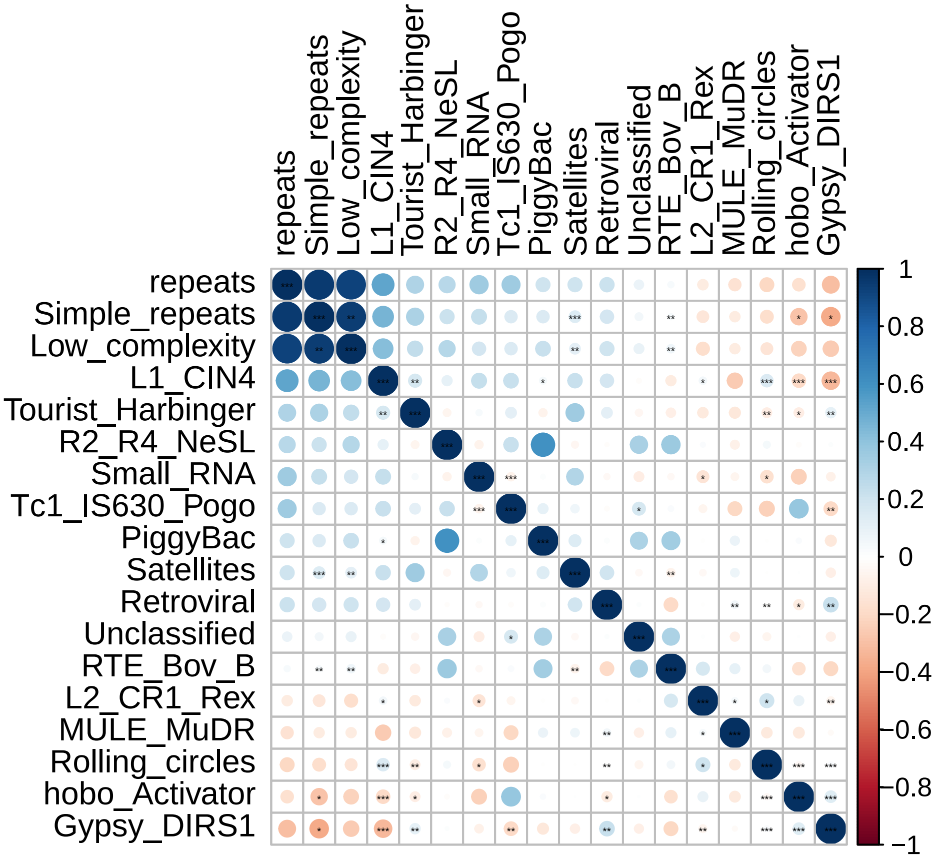

Simple_repeats
12
9
6
3
value
ns
ns
***
YES
PART
NO
sociality


Low_complexity
ns
ns
***
value
3
2
1
0
YES
PART
NO
sociality



Correllations

```
hcor <- hmp3 %>% cor
hp <- corrplot::cor.mtest(hmp3)
corrplot::corrplot(hmp3 %>% cor,
                   p.mat = hp$p,
                   insig = 'label_sig',
                   sig.level = c(0.001, 0.01, 0.05),
                   pch.cex = 0.3,
                   is.corr = T,
                   tl.col = "black",
                   method = 'circle',
                   order = "FPC",
                   #col = c('white', 'black'), bg = 'gold2'
                   )
```



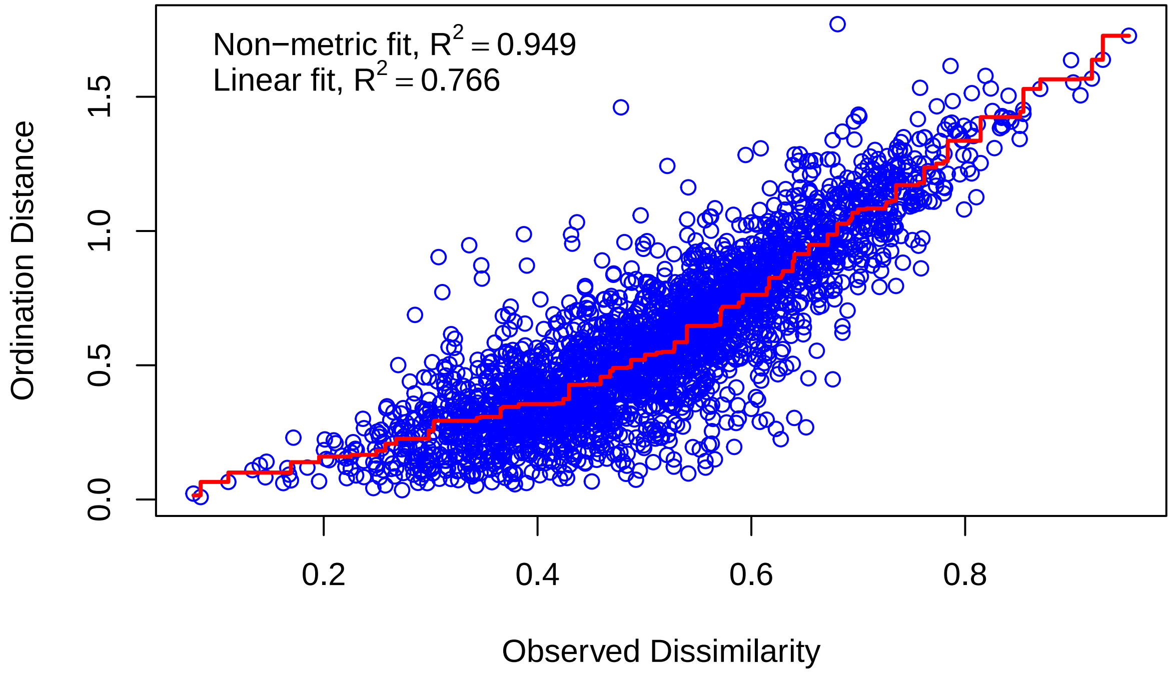
NMDS

Coph as distance

```
MMDS <- metaMDS(comm = hmp3,
        dist = coph,
        distfun = function(x) vegdist(x,method = "bray"))
```

```
## Run 0 stress 0.2277812
## Run 1 stress 0.226357
## ... New best solution
## ... Procrustes: rmse 0.02646725  max resid 0.1866783
## Run 2 stress 0.2362722
## Run 3 stress 0.2327406
## Run 4 stress 0.2503912
## Run 5 stress 0.2491381
## Run 6 stress 0.2297638
## Run 7 stress 0.2299896
## Run 8 stress 0.245215
## Run 9 stress 0.226357
## ... Procrustes: rmse 4.017085e-05  max resid 0.0002076848
## ... Similar to previous best
## Run 10 stress 0.226357
## ... Procrustes: rmse 6.059361e-05  max resid 0.0003327073
## ... Similar to previous best
## Run 11 stress 0.2601206
## Run 12 stress 0.2540919
## Run 13 stress 0.2440955
## Run 14 stress 0.2456981
## Run 15 stress 0.2310447
## Run 16 stress 0.2321069
## Run 17 stress 0.2325113
## Run 18 stress 0.2325226
## Run 19 stress 0.2544167
## Run 20 stress 0.2312615
## *** Best solution repeated 2 times
```

```
#goodness(MMDS)
stressplot(MMDS)
```



```
envfit(MMDS, socvec)
```

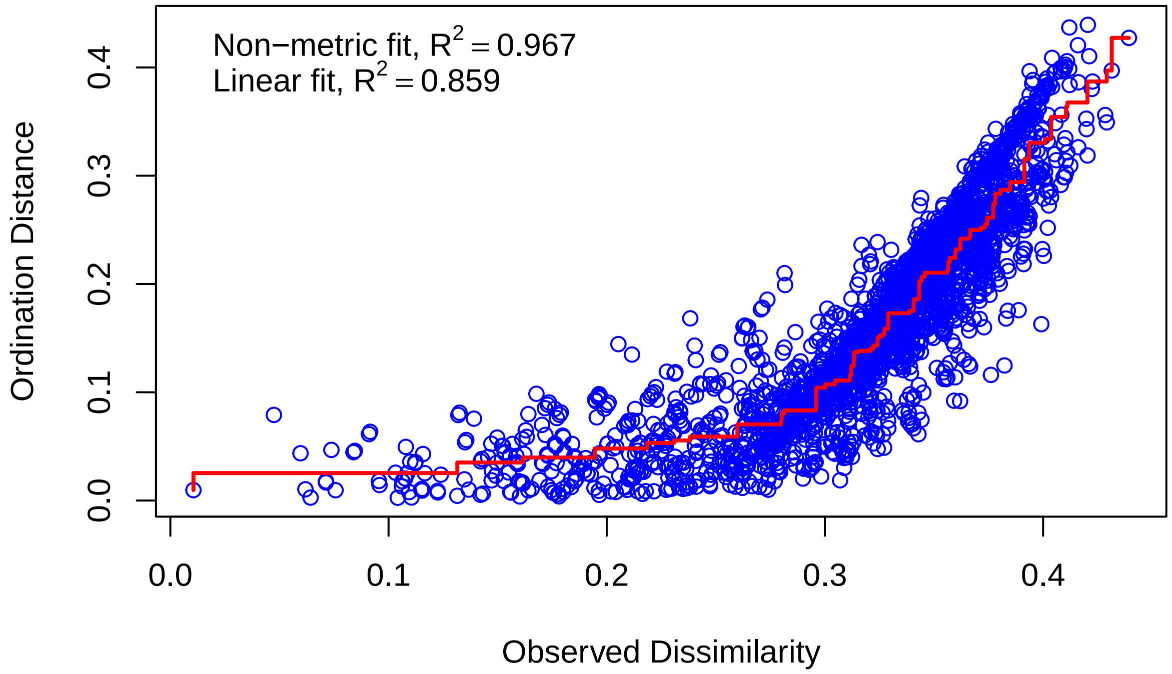
```
##
## ***FACTORS:
##
## Centroids:
##               NMDS1   NMDS2
## SocialNO     0.1915  0.1066
## SocialPART -0.0392 -0.2678
## SocialYES  -0.1373 -0.0449
##
## Goodness of fit:
##            r2 Pr(>r)
## Social 0.1562  0.001 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## Permutation: free
## Number of permutations: 999
```

Coph as matrix

```
MMDS <- metaMDS(comm = coph,
        distfun = function(x) vegdist(x,method = "bray"))
```

```
## Run 0 stress 0.18142
## Run 1 stress 0.1822556
## Run 2 stress 0.1822798
## Run 3 stress 0.1993977
## Run 4 stress 0.2409072
## Run 5 stress 0.2110128
## Run 6 stress 0.1996611
## Run 7 stress 0.1867918
## Run 8 stress 0.1923358
## Run 9 stress 0.2405974
## Run 10 stress 0.4088288
## Run 11 stress 0.1892917
## Run 12 stress 0.1822003
## Run 13 stress 0.2074174
## Run 14 stress 0.1817423
## ... Procrustes: rmse 0.07105617  max resid 0.1846668
## Run 15 stress 0.1872247
## Run 16 stress 0.2026956
## Run 17 stress 0.1883673
## Run 18 stress 0.2349414
## Run 19 stress 0.1819521
## Run 20 stress 0.2263268
## *** Best solution was not repeated -- monoMDS stopping criteria:
##     20: stress ratio > sratmax
```

```
#goodness(MMDS)
stressplot(MMDS)
```



```
envfit(MMDS, hmp %>% cbind(socvec))
```

```
##
## ***VECTORS
##
##                      NMDS1    NMDS2     r2 Pr(>r)
## L2_CR1_Rex         0.99943 -0.03385 0.0356  0.267
## R2_R4_NeSL         0.98484 -0.17348 0.0219  0.469
## RTE_Bov_B          0.20436  0.97890 0.0218  0.458
## L1_CIN4            0.93157 -0.36356 0.0075  0.773
## Gypsy_DIRS1       -0.79641  0.60476 0.0054  0.843
## Retroviral         0.14852  0.98891 0.0563  0.147
## hobo_Activator    -0.94599 -0.32421 0.0263  0.391
## Tc1_IS630_Pogo    -0.91471  0.40411 0.0295  0.349
## MULE_MuDR          0.73736  0.67550 0.0604  0.126
## PiggyBac           0.99095  0.13426 0.0807  0.048 *
## Tourist_Harbinger  0.29972  0.95403 0.0174  0.526
## Rolling_circles    0.91610 -0.40095 0.0885  0.044 *
## Unclassified       0.42122 -0.90696 0.0467  0.199
## Small_RNA         -0.15884 -0.98730 0.0082  0.741
## Satellites         0.64629  0.76309 0.0326  0.331
## Simple_repeats     0.42534  0.90503 0.0284  0.368
## Low_complexity     0.46214  0.88681 0.0495  0.160
## repeats            0.32442  0.94591 0.0247  0.420
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## Permutation: free
## Number of permutations: 999
##
## ***FACTORS:
##
## Centroids:
##               NMDS1   NMDS2
## SocialNO    0.0069 -0.0101
## SocialPART  0.0415 -0.0078
## SocialYES  -0.0105  0.0085
##
## Goodness of fit:
##             r2 Pr(>r)
## Social 0.0194  0.615
## Permutation: free
## Number of permutations: 999
```

Phylogenetic GLM

phylolm test

```
hmp4 <- hmp3 %>%
  mutate(social = socvec %>%
           unlist) %>%
  mutate(socbin = !(social == "NO"))

phy <- d6 %>%
  as.phylo() %>%
  drop.tip(tip = .$tip.label[!(.$tip.label %in% rownames(hmp4))])

signinf <- c()

for ( i in colnames(hmp3) ) {
  lmform <- as.formula(paste0("socbin ~ ", i))
  fit <- try(suppressWarnings(phylolm::phyloglm(lmform,
                    phy = phy,
                    method = "logistic_IG10",
                    data = hmp4,
                    boot = 100) ),
    silent = T)

  sfit <- summary(fit)

  rv = fit$coefficients[2]
  pv1 = sfit$coefficients[,"p.value"][2]

  if(pv1 < 0.1) signif <- c( signif, as.character(i) )

  pv2 = rep("*", -log10(pv1) %>% round) %>%
    paste0(collapse = "")
  cat("\n",i,
      "\ncoef: ", rv %>% round(2),
      "\np: ", pv1 %>% round(3), pv2,
      "\n##########\n", sep = "")
}
```

```
##
## L2_CR1_Rex
## coef: -7.69
## p: 0.022**
## ##########
##
## R2_R4_NeSL
## coef: 1.14
## p: 0.475
## ##########
##
## RTE_Bov_B
## coef: 0.08
## p: 0.917
## ##########
##
## L1_CIN4
## coef: 0.88
## p: 0.299*
## ##########
##
## Gypsy_DIRS1
## coef: -0.64
## p: 0.327
## ##########
##
## Retroviral
## coef: -0.27
## p: 0.707
## ##########
##
## hobo_Activator
## coef: -1.37
## p: 0.227*
## ##########
##
## Tc1_IS630_Pogo
## coef: 0.27
## p: 0.737
## ##########
##
## MULE_MuDR
## coef: -2.47
## p: 0.065*
## ##########
##
## PiggyBac
## coef: -0.08
## p: 0.939
## ##########
##
## Tourist_Harbinger
## coef: 1.89
## p: 0.138*
## ##########
##
## Rolling_circles
## coef: -3.48
## p: 0.068*
## ##########
##
## Unclassified
## coef: 2.22
## p: 0.346
## ##########
##
## Small_RNA
## coef: -0.01
## p: 0.974
## ##########
##
## Satellites
## coef: -0.08
## p: 0.942
## ##########
##
## Simple_repeats
## coef: 6.81
## p: 0.002***
## ##########
##
## Low_complexity
## coef: 6.72
## p: 0.002***
## ##########
##
## repeats
## coef: 5.63
## p: 0.01**
## ##########
```