Tissue specificity index analysis - tau

Deisiany

2024-01-13

# Load libraries

# read in data

For Tissue specificity index analysis - tau - we will use the 133 genes which had average expression across all tissues equal or greater than 2, and was used in our PCA analysis.

# Load full genes  
data <- read\_csv("Tau\_133\_genes\_with\_expression\_across\_tissues\_are\_greater\_than\_2\_SoyARCs.csv")

## Rows: 133 Columns: 20  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Transcript ID, heatmap\_label, Family, Class, Clade  
## dbl (15): AM, OF, IAM, IBM, RootTip, Cotyledon, Hypocotyl, SAM6D, SAM17D, SA...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data <- data %>% select(-c(Class, Clade, expr\_median))

# Calculate the component value

X(^) = Xi/ max(Xi) for 1<=i <=n

# lets start by assigning our data to df  
max\_component\_df <- data  
  
for (Xi in 1:nrow(max\_component\_df)) {  
 max\_component\_df[Xi, 4:17] <- max\_component\_df[Xi, 4:17] / max(max\_component\_df[Xi, 4:17], na.rm = TRUE)  
}  
  
max\_component\_df

## # A tibble: 133 × 17  
## `Transcript ID` heatmap\_label Family AM OF IAM IBM RootTip  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Glyma.01G002100.1 Glyma.01G002100|G… ARF 0.816 0.890 0.797 0.945 0.486   
## 2 Glyma.01G019400.1 Glyma.01G019400|G… IAA 0.382 0.718 0 0.140 0.349   
## 3 Glyma.01G019400.2 Glyma.01G019400|G… IAA 0.175 0 0.266 0.339 0.0983  
## 4 Glyma.01G019400.3 Glyma.01G019400|G… IAA 0.254 0.404 0.234 0.301 0.182   
## 5 Glyma.01G098000.3 Glyma.01G098000|G… IAA 0.396 0.298 0.311 0.559 0.190   
## 6 Glyma.01G103500.1 Glyma.01G103500|G… ARF 0.694 0.214 0.420 0.839 0.146   
## 7 Glyma.02G000500.1 Glyma.02G000500|G… IAA 0.252 0.315 0.137 0.114 0.271   
## 8 Glyma.02G065300.1 Glyma.02G065300|G… AFB/T… 0.522 0.424 0.506 0.772 0.251   
## 9 Glyma.02G142500.3 Glyma.02G142500|G… IAA 0.0277 1 0.138 0.0538 0.0816  
## 10 Glyma.02G142600.1 Glyma.02G142600|G… IAA 0.0541 1 0.243 0.266 0.238   
## # ℹ 123 more rows  
## # ℹ 9 more variables: Cotyledon <dbl>, Hypocotyl <dbl>, SAM6D <dbl>,  
## # SAM17D <dbl>, SAM38D <dbl>, Callus <dbl>, Leaf <dbl>, Root <dbl>,  
## # Nodule <dbl>

# Calculate Tau (Tissue Specificity Index) for each gene  
tau\_df <- max\_component\_df  
  
tau\_df$tau <- numeric(nrow(tau\_df)) # Create an empty column  
  
for (i in 1:nrow(tau\_df)) {  
 tau\_df$tau[i] <- sum(1 - tau\_df[i, 4:17]) / 13  
}  
  
# write\_csv(tau\_df, "tau\_df\_contains\_maximal\_component\_values.csv")

(tau\_df\_values <- tau\_df %>% select(`Transcript ID`, heatmap\_label, tau))

## # A tibble: 133 × 3  
## `Transcript ID` heatmap\_label tau  
## <chr> <chr> <dbl>  
## 1 Glyma.01G002100.1 Glyma.01G002100|GmARF7/19\_F.1 0.376  
## 2 Glyma.01G019400.1 Glyma.01G019400|GmIAA8-9-B.1 0.615  
## 3 Glyma.01G019400.2 Glyma.01G019400|GmIAA8-9-B.2 0.796  
## 4 Glyma.01G019400.3 Glyma.01G019400|GmIAA8-9-B.3 0.800  
## 5 Glyma.01G098000.3 Glyma.01G098000|GmIAA8-9-D.3 0.643  
## 6 Glyma.01G103500.1 Glyma.01G103500|GmARF9\_A.1 0.711  
## 7 Glyma.02G000500.1 Glyma.02G000500|GmIAA1-4-D.1 0.831  
## 8 Glyma.02G065300.1 Glyma.02G065300|GmAFB2/3\_C.1 0.619  
## 9 Glyma.02G142500.3 Glyma.02G142500|GmIAA16-F.3 0.944  
## 10 Glyma.02G142600.1 Glyma.02G142600|GmIAA1-4-H.1 0.895  
## # ℹ 123 more rows

#write\_csv(tau\_df\_values, "tau\_df\_values.csv")

# tau values greater than .80 are a cut off utilized in the literature to denote tissues specificity.

tau\_greater\_80 <- tau\_df\_values[tau\_df\_values$tau >= 0.80, ]  
  
#write\_csv(tau\_greater\_80, "tau\_greater\_80.csv")

Here <https://rdrr.io/github/roonysgalbi/tispec/f/vignettes/UserGuide.Rmd> we can access some useful information about the tau measurement. And these papers also mention these same analysis: <https://academic.oup.com/bioinformatics/article/21/5/650/220059#394030419> <https://academic.oup.com/bib/article/18/2/205/2562739?login=true> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7791837/#CR75>

# tau distribution for all 14 tissues

tau\_df <- read\_csv( "tau\_df\_contains\_maximal\_component\_values.csv")

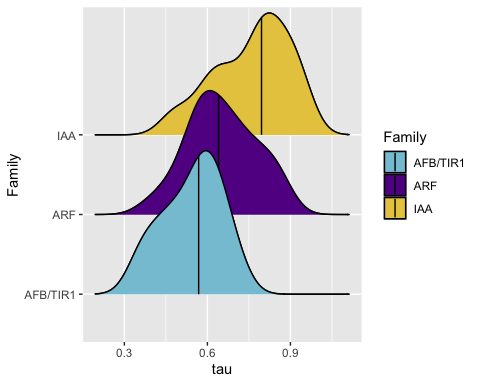
## Rows: 133 Columns: 18  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): Transcript ID, heatmap\_label, Family  
## dbl (15): AM, OF, IAM, IBM, RootTip, Cotyledon, Hypocotyl, SAM6D, SAM17D, SA...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

library(ggplot2)  
library(ggridges)

## Warning: package 'ggridges' was built under R version 4.3.2

library(dplyr)  
  
# Calculate medians  
medians <- tau\_df %>%   
 group\_by(Family) %>%   
 summarise(median\_tau = median(tau))  
  
ggplot(tau\_df, aes(x = tau, y = Family, fill = Family)) +  
 scale\_fill\_manual(values = c("#86C5D8", "#620093", "#E7C94C")) +  
 ggridges::geom\_density\_ridges() +  
 ggridges::stat\_density\_ridges(  
 quantile\_lines = TRUE,   
 quantiles = 2) +  
 coord\_cartesian(ylim = c(1, length(unique(tau\_df$Family)) + 1))

## Picking joint bandwidth of 0.0518  
## Picking joint bandwidth of 0.0518



# statistical testing

stats\_tau <- tau\_df %>% select(Family, tau)  
  
summary(stats\_tau)

## Family tau   
## Length:133 Min. :0.3506   
## Class :character 1st Qu.:0.5807   
## Mode :character Median :0.6743   
## Mean :0.6828   
## 3rd Qu.:0.7964   
## Max. :0.9559

library(doBy)

## Warning: package 'doBy' was built under R version 4.3.2

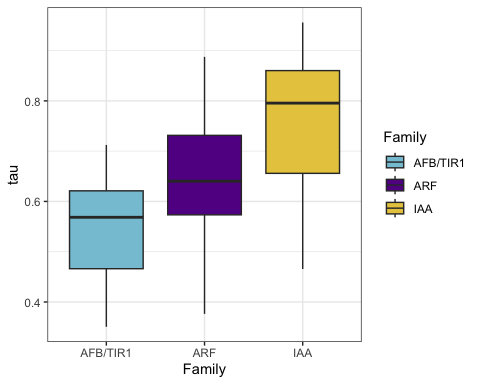
##   
## Attaching package: 'doBy'

## The following object is masked from 'package:dplyr':  
##   
## order\_by

summaryBy(tau ~ Family,  
 data = stats\_tau,  
 FUN = median)

## # A tibble: 3 × 2  
## Family tau.median  
## <chr> <dbl>  
## 1 AFB/TIR1 0.568  
## 2 ARF 0.640  
## 3 IAA 0.796

ggplot(stats\_tau) +  
 aes(x = Family, y = tau, fill = Family) +  
 geom\_boxplot() +  
 theme(legend.position = "none") +  
 scale\_fill\_manual(values = c("#86C5D8", "#620093", "#E7C94C")) +  
 theme\_bw()

 The null and alternative hypotheses of the Kruskal-Wallis test are:

𝐻0: The 3 families are equal in terms of tau 𝐻1: At least one Family is different from the other 2 Families in terms of tau

# kruskal test

kruskal.test(tau ~ Family, data = stats\_tau)

##   
## Kruskal-Wallis rank sum test  
##   
## data: tau by Family  
## Kruskal-Wallis chi-squared = 32.403, df = 2, p-value = 9.201e-08

based on the kruskal test we reject the null hypothesis and conclude that at least one family is different in terms of tau values.

# Post-hoc test

pairwise.wilcox.test(stats\_tau$tau, stats\_tau$Family,  
 p.adjust.method = "holm"  
)

##   
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction   
##   
## data: stats\_tau$tau and stats\_tau$Family   
##   
## AFB/TIR1 ARF   
## ARF 0.0042 -   
## IAA 1.1e-05 3.2e-05  
##   
## P value adjustment method: holm

all 3 families are differ significantly in terms of tau, p<0.05

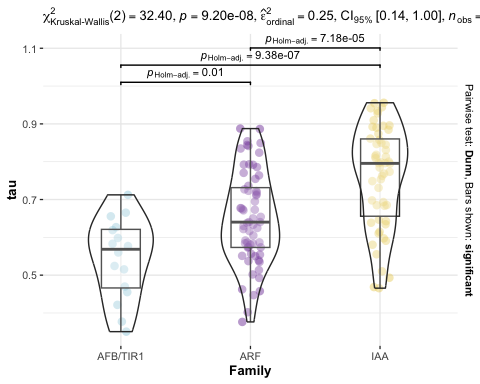
# install newest version of ggplot2 and load all packages  
#library(remotes)  
# remotes::install\_version("ggplot2", version = "3.5.1")  
#install.packages("ggstatsplot")  
library(ggplot2)  
library(ggstatsplot)

## Warning: package 'ggstatsplot' was built under R version 4.3.2

## You can cite this package as:  
## Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.  
## Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

ggstatsplot::ggbetweenstats(  
 data = stats\_tau,  
 x = Family,  
 y = tau,  
 type = "nonparametric", # ANOVA or Kruskal-Wallis  
 plot.type = "box",  
 pairwise.comparisons = TRUE,  
 pairwise.display = "significant",  
 centrality.plotting = FALSE,  
 bf.message = FALSE) +  
 scale\_color\_manual(values =   
 c("#86C5D8", "#620093", "#E7C94C"))

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



# repeat analysis removing tissues that were not used in our final analysis

Roots, callus, Inflorescence and so on, were not taken into account in our final analysis. Thus, lets repeat this without this tissues.

# lets start by assigning our data to df  
max\_component\_df\_aerial <- data %>%   
 select(`Transcript ID`, heatmap\_label, SAM6D, SAM17D, SAM38D, AM, Cotyledon, Leaf, Hypocotyl)  
  
for (Xi in 1:nrow(max\_component\_df\_aerial)) {  
 max\_component\_df\_aerial[Xi, 3:9] <- max\_component\_df\_aerial[Xi, 3:9] / max(max\_component\_df\_aerial[Xi, 3:9], na.rm = TRUE)  
}  
  
max\_component\_df\_aerial

## # A tibble: 133 × 9  
## `Transcript ID` heatmap\_label SAM6D SAM17D SAM38D AM Cotyledon Leaf  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Glyma.01G002100.1 Glyma.01G00210… 0.718 0.650 0.461 0.816 0.505 0.392   
## 2 Glyma.01G019400.1 Glyma.01G01940… 0.283 0.747 0.327 0.382 0.0952 1   
## 3 Glyma.01G019400.2 Glyma.01G01940… 0.795 0 0.211 0.334 0.177 1   
## 4 Glyma.01G019400.3 Glyma.01G01940… 0.283 0.299 0.328 0.254 0.0454 0.0695  
## 5 Glyma.01G098000.3 Glyma.01G09800… 0.590 0.628 0.334 0.396 0.455 0.419   
## 6 Glyma.01G103500.1 Glyma.01G10350… 0.790 1 0.488 0.694 0.0621 0   
## 7 Glyma.02G000500.1 Glyma.02G00050… 0.220 0.215 0.415 0.252 0.0604 0.0656  
## 8 Glyma.02G065300.1 Glyma.02G06530… 0.859 1 0.430 0.522 0.0816 0.148   
## 9 Glyma.02G142500.3 Glyma.02G14250… 0.0641 0.0579 0.0650 0.128 0.00176 0.179   
## 10 Glyma.02G142600.1 Glyma.02G14260… 0.220 0.352 0.130 0.321 0.436 0.541   
## # ℹ 123 more rows  
## # ℹ 1 more variable: Hypocotyl <dbl>

# Calculate Tau (Tissue Specificity Index) for each gene  
tau\_df\_aerial <- max\_component\_df\_aerial  
  
tau\_df\_aerial$tau <- numeric(nrow(tau\_df\_aerial)) # Create an empty column  
  
for (i in 1:nrow(tau\_df\_aerial)) {  
 tau\_df\_aerial$tau[i] <- sum(1 - tau\_df\_aerial[i, 3:9]) / 6  
}  
  
#write\_csv(tau\_df\_aerial, "tau\_df\_values\_aerial\_tissues.csv")  
  
tau\_df\_aerial\_values <- tau\_df\_aerial %>% select(`Transcript ID`, heatmap\_label, tau)  
  
# Now lets select values greater than 80 again  
tau\_aerial\_greater\_80 <-   
 tau\_df\_aerial\_values[tau\_df\_aerial\_values$tau >= 0.80, ]