Gene expression analysis using A-SBF

Amal Thomas

Contents

1 Introduction

Consider a set of k real matrices $D_i \in \mathbb{R}^{m_i \times n}$, each with full column rank. Each D_i represents the mean expression of n tissues (columns) in species i with m_i genes (rows) annotated for that species. The p-th column of each D_i matrix corresponds to the average gene-expression profile of a functionally equivalent tissue representing a similar phenotype across species. The number of rows is different for each D_i matrix, but they all have the same number of columns.

In A-SBF, we estimate an orthonormal basis V for the common expression space based on the inter-tissue correlations within each species. Let $X_i \in \mathbb{R}^{m_i \times n}$ be the standardized gene expression matrix for species i. We can write

$$X_i = C_i D_i {S_i}^{-1},$$

where

$$C_i = I_{m_i} - {m_i}^{-1} 1_{m_i} {1_{m_i}}^T$$

is a centering matrix and $S_i = \mathrm{diag}(s_1, \ldots, s_n)$ is a diagonal scaling matrix, where s_j is the standard deviation of j-th column of D_i. Within species i, the correlation between expression profiles for the n tissues is

$$R_i = X_i^T X_i / m_i.$$

Each R_i matrix has a dimension of $n \times n$ independent of the number of genes annotated in the species. Since the corresponding column of each D_i matrix represents a similar phenotype, we can define an expected correlation matrix across the species:

$$E = \mathrm{E}(R) = \tfrac{1}{k} \textstyle\sum_{i=1}^{k} R_i.$$

The shared right basis matrix V represents the common expression space and is determined from the eigenvalue decomposition of E, where $E = V\Lambda V^T$. The right basis matrix V is identical in all the k matrix factorizations, and the different dimensions of V represent the correlation relationship between tissues independent of the fact to which species tissues belong. Once the V is estimated, the species-specific left basis matrix with orthonormal columns U_i (eigengenes) and the Δ_i that minimize the factorization error are computed

```
# load SBF package
library(SBF)
```

Additional package required for the vignette

```
# install packages
pkgs <- c("data.table", "dplyr", "matrixStats")
require_install <- pkgs[!(pkgs %in% row.names(installed.packages()))]
if (length(require_install))
  install.packages(require_install)
suppressPackageStartupMessages({
  library(data.table)
  library(dplyr)
  library(matrixStats)
})
```

2 Species and organs

In this workflow, we will work with gene expression profile of six tissues from eight species.

```
species <- c("Homo_sapiens", "Pan_troglodytes", "Macaca_mulatta",
             "Mus_musculus", "Rattus_norvegicus", "Bos_taurus",
             "Sus_scrofa", "Gallus_gallus")
# function to create short name for species
getSpeciesShortName <- function(species_list) {
    require(data.table)
    shortname <- c()
    for (sp in species_list) {
       if (!grepl("_", sp)) {
          cat("\nspecies:", sp, "\n")
          stop("Invalid species full name")
       }
       shortname <- c(shortname,
                      paste0(tolower(substr(data.table::tstrsplit(sp, "_")[[1]],
                                            1, 1)),
                      tstrsplit(sp, "_")[[2]]))
    }
    return(shortname)
}
species_short <- sapply(species, getSpeciesShortName)
species_short
#>      Homo_sapiens   Pan_troglodytes    Macaca_mulatta      Mus_musculus
#>         "hsapiens"    "ptroglodytes"        "mmulatta"       "mmusculus"
#> Rattus_norvegicus        Bos_taurus        Sus_scrofa     Gallus_gallus
#>      "rnorvegicus"         "btaurus"         "sscrofa"         "ggallus"
# common tissues present in all 8 species
tissues <-  c("brain", "heart", "kidney", "liver", "lung", "testis")
```

3 Load gene expression profiles

Download the processed RNA-Seq counts file from https://figshare.com/s/4d8d5cf9c1362bcbe90d

```
# set the path to the counts. Change it accordingly
counts_path <- "~/Dropbox/0.Analysis/0.paper/counts/"
counts_list <- metadata_list <- list()
require(data.table)
for (sp in species) {
  counts <- data.table::fread(paste0(counts_path, sp, "_logTPM.tsv"),
                              sep = "\t", header = T, data.table = FALSE,
                              nThread = 4)
  row.names(counts) <- counts$V1
  counts$V1 <- NULL
  col_fields <- data.table::tstrsplit(colnames(counts), "_")
  metadata <- data.frame(
      project = col_fields[[1]],
      species = col_fields[[2]],
      tissue = col_fields[[3]],
      gsm = col_fields[[4]],
      name = colnames(counts),
      stringsAsFactors = F)
  metadata_sel <- metadata[metadata$tissue %in% tissues, , drop = F]
  counts_sel <- counts[, colnames(counts) %in% metadata_sel$name, drop = F]
  metadata_sel$ref <- seq_len(nrow(metadata_sel))
  metadata_sel$key <- paste0(metadata_sel$species, "_", metadata_sel$tissue)

  counts_list[[sp]] <- counts_sel
  metadata_list[[sp]] <- metadata_sel
}
```

The dimensions and the number of RNA-Seq profiles for different species.

```
sapply(counts_list, dim)
#>      Homo_sapiens Pan_troglodytes Macaca_mulatta Mus_musculus Rattus_norvegicus
#> [1,]        58676           31373          30807        54446              32623
#> [2,]          111              61            124          107                 64
#>      Bos_taurus Sus_scrofa Gallus_gallus
#> [1,]      24596      25880         19152
#> [2,]         53        165           117
```

Now, for each species, let us compute the average expression profile for each tissue.

```
#function to calculate mean expression values for each tissue/group
calcAvgCounts <- function(counts, metadata, ndecimal = 4) {
   require(dplyr)
   counts.avg <- list()
   if ("key" %in% colnames(metadata)) {
     for (species.organ in sort(unique(metadata$key))) {
        no.of.libs <- length(colnames(counts)[grepl(species.organ,
                                                    colnames(counts))])
        if (no.of.libs > 1)
           counts.avg[[species.organ]] <- round(rowMeans(counts[, colnames(counts)[
                                  grepl(species.organ, colnames(counts))]],
                                  dim = 1), ndecimal)
        else if (no.of.libs == 1)
```

```
          counts.avg[[species.organ]] <- round(counts[, colnames(counts)[grepl(
                              species.organ, colnames(counts))]], drop = T], ndecimal)
    }
    avg.counts <- as.data.frame(counts.avg %>% dplyr::bind_cols())
    row.names(avg.counts) <- row.names(counts)
    return(avg.counts)
  } else {
    stop("'key' column not found in the metadata!Exiting!")
  }
}

avg_counts <- list()
for (sp in species) {
    avg_counts[[sp]] <- calcAvgCounts(counts_list[[sp]],
                                      metadata_list[[sp]])
}
# check tissue columns are matching in each species
c_tissues <- as.data.frame(sapply(avg_counts, function(x) {
            data.table::tstrsplit(colnames(x), "_")[[2]]
  }))
if (!all(apply(c_tissues, 1, function(x) all(x == x[1])))) {
        stop("Error! tissues not matching")
}
```

The dimension of mean expression profiles

```
sapply(avg_counts, dim)
#>      Homo_sapiens Pan_troglodytes Macaca_mulatta Mus_musculus Rattus_norvegicus
#> [1,]        58676           31373          30807        54446              32623
#> [2,]            6               6              6            6                  6
#>      Bos_taurus Sus_scrofa Gallus_gallus
#> [1,]      24596      25880         19152
#> [2,]          6          6             6
```

Remove genes not expressed.

```
removeZeros <- function(df) {
    return(df[rowSums(df) > 0, ])
}

avg_counts <- lapply(avg_counts, removeZeros)
sapply(avg_counts, dim)
#>      Homo_sapiens Pan_troglodytes Macaca_mulatta Mus_musculus Rattus_norvegicus
#> [1,]        47301           26173          27762        39097              23004
#> [2,]            6               6              6            6                  6
#>      Bos_taurus Sus_scrofa Gallus_gallus
#> [1,]      20028      21448         17810
#> [2,]          6          6             6
# update counts_list
counts_list_sub <- list()
for (sp in names(avg_counts)) {
  counts_list_sub[[sp]] <- counts_list[[sp]][row.names(avg_counts[[sp]]), ,
                                             drop = F]
}
```

4 A-SBF

We will perform A-SBF in two ways.

1. Keeping the initial estimate of V the same while updating U_i and Δ_i to minimize the factorization error. By keeping the V same, the initial V estimated based inter-sample correlation is maintained.
2. Update V, U_i and Δ_i to minimize the factorization error.

```
cat("\n=========================================================")
#>
#> =========================================================
cat("\nASBF with optimizing V = FALSE started\n")
#>
#> ASBF with optimizing V = FALSE started
cat(format(Sys.time(), "%a %b %d %X %Y"), "\n")
#> Fri Apr 22 12:45:13 PM 2022
t1 <- proc.time()
sbf_noVupdate <- SBF(avg_counts, transform_matrix = TRUE, approximate = TRUE,
                     optimizeV = FALSE, tol = 1e-3)#, tol = 1e-10)
#>
#> A-SBF optimizing factorization error
t2 <- proc.time()
cat(format(Sys.time(), "%a %b %d %X %Y"), "\n")
#> Fri Apr 22 12:45:14 PM 2022
cat("ASBF with optimizing V = FALSE finished\n")
#> ASBF with optimizing V = FALSE finished
cat("Time taken:\n")
#> Time taken:
t2 - t1
#>    user  system elapsed
#>   0.657   0.084   0.742
```

```
cat("\n=========================================================")
#>
#> =========================================================
cat("\nASBF with optimizing V=TRUE started\n")
#>
#> ASBF with optimizing V=TRUE started
cat(format(Sys.time(), "%a %b %d %X %Y"), "\n")
#> Fri Apr 22 12:45:14 PM 2022
t1 <- proc.time()
sbf <- SBF(avg_counts, transform_matrix = TRUE, approximate = TRUE,
                     optimizeV = TRUE, tol = 1e-3)#, tol = 1e-10)
#>
#> A-SBF optimizing factorization error
t2 <- proc.time()
cat(format(Sys.time(), "%a %b %d %X %Y"), "\n")
#> Fri Apr 22 12:45:28 PM 2022
cat("ASBF with optimizing V=TRUE finished\n")
#> ASBF with optimizing V=TRUE finished
cat("Time taken:\n")
#> Time taken:
t2 - t1
#>    user  system elapsed
#>  14.156   0.084  14.243
```

Note: We use tolerance threshold = 1e-3 for fast computation. Reduce the tolerance threshold (default value 1e-10) to minimize the factorization error futher. Lowering the threshold value increases the computing time.

The final factorization error and number of updates taken:

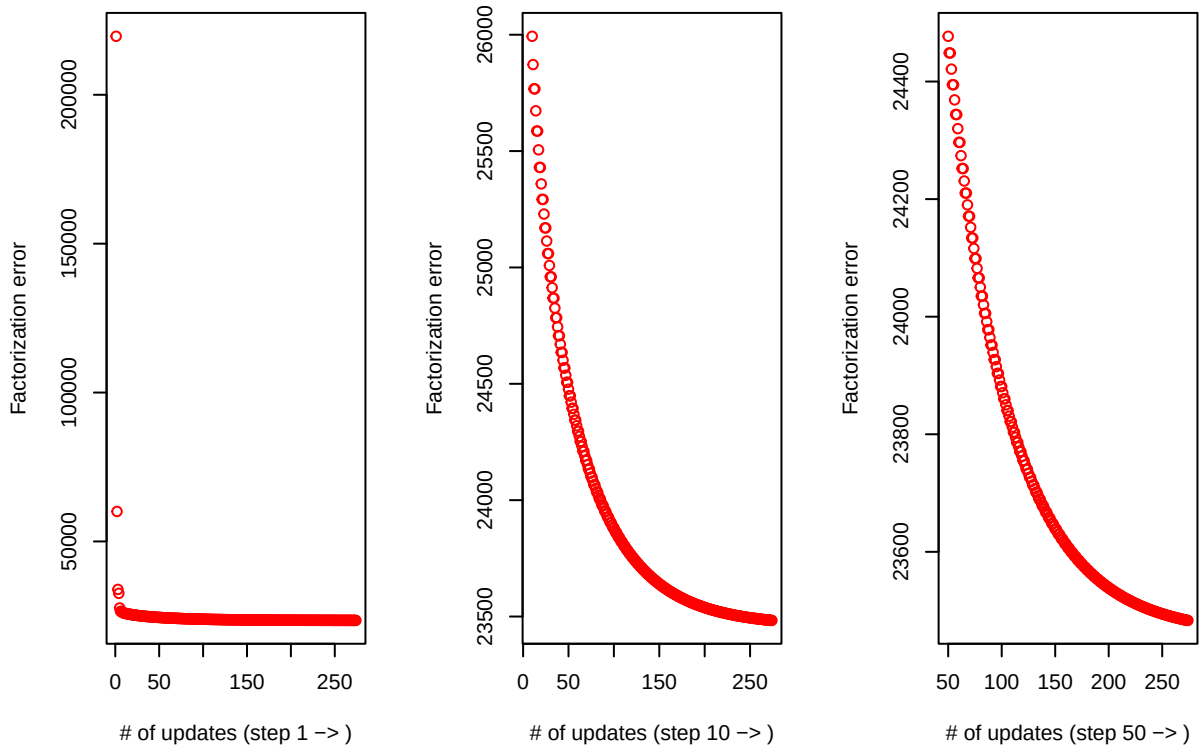
```
cat("\n", sprintf("%-27s:", "Final error [No V update]"), sprintf("%16.2f",
                                                        sbf_noVupdate$error))
#>
#>  Final error [No V update]  :        219507.66
cat("\n", sprintf("%-27s:", "Final error [With V update]"), sprintf("%16.2f",
                                                                sbf$error))
#>
#>  Final error [With V update]:         23483.37
cat("\n", sprintf("%-27s:", "# of update [No V update]"), sprintf("%16d",
                                                      sbf_noVupdate$error_pos))
#>
#>  # of update [No V update]  :                5
cat("\n", sprintf("%-27s:", "# of update [With V update]"), sprintf("%16d",
                                                            sbf$error_pos))
#>
#>  # of update [With V update]:              274
```

Optimization with updating V achieves a lower decomposition error.

```
sbf_noVupdate$error / sbf$error
#> [1] 9.347368
```

Let us plot the decomposition error vs. updates.

```
par(mfrow = c(1, 3))
plot(x = seq_len(length(sbf$error_vec)), y = sbf$error_vec,
     xlab = "# of updates (step 1 -> )",
     ylab = "Factorization error", col = "red")
plot(x = 10:length(sbf$error_vec),
     y = sbf$error_vec[10:length(sbf$error_vec)],
     xlab = "# of updates (step 10 -> )",
     ylab = "Factorization error", col = "red")
plot(x = 50:length(sbf$error_vec),
     y = sbf$error_vec[50:length(sbf$error_vec)],
     xlab = "# of updates (step 50 -> )",
     ylab = "Factorization error", col = "red")
```



Percentage of information (p_{ij}) represented by a common space dimension is defined as $p_{ij} = \delta_{ij}^2 / \sum_{j=1}^{6} \delta_{ij}^2 \times 100$, where $\Delta_i = \text{diag}(\delta_{i1}, \ldots, \delta_{i6})$

```
percentInfo_noVupdate <- lapply(sbf_noVupdate$d, function(x) {
  round(x^2 / sum(x^2) * 100, 2)})
cat("\nPercentage for each delta [No V update]:")
#>
#> Percentage for each delta [No V update]:
for (i in names(sbf_noVupdate$d)) {
  cat("\n", sprintf("%-25s:", i), sprintf("%8.2f", percentInfo_noVupdate[[i]]))
}
#>
#>  Homo_sapiens             :    86.56     5.52     3.01     2.03     1.62     1.26
#>  Pan_troglodytes          :    88.57     4.87     2.61     1.71     1.30     0.95
#>  Macaca_mulatta           :    87.44     5.09     2.99     1.75     1.47     1.26
#>  Mus_musculus             :    81.30     8.11     4.56     2.64     1.73     1.66
#>  Rattus_norvegicus        :    83.84     6.53     3.96     2.40     1.77     1.49
#>  Bos_taurus               :    86.99     5.19     2.94     1.85     1.77     1.26
#>  Sus_scrofa               :    85.28     5.38     3.52     2.31     2.00     1.51
#>  Gallus_gallus            :    87.70     4.49     2.93     1.90     1.65     1.33


percentInfo <- lapply(sbf$d, function(x) {
  round(x^2 / sum(x^2) * 100, 2) })
cat("\nPercentage for each delta:")
#>
#> Percentage for each delta:
for (i in names(sbf$d)) {
  cat("\n", sprintf("%-25s:", i), sprintf("%8.2f", percentInfo[[i]]))
}
```

```
#>
#>  Homo_sapiens              :    87.23     5.05     2.99     2.13     1.41     1.19
#>  Pan_troglodytes           :    88.86     4.64     2.70     1.60     1.24     0.96
#>  Macaca_mulatta            :    87.77     4.87     2.98     1.79     1.36     1.22
#>  Mus_musculus              :    81.64     7.99     4.43     2.67     1.67     1.59
#>  Rattus_norvegicus         :    84.10     6.48     3.86     2.51     1.62     1.44
#>  Bos_taurus                :    87.37     4.94     2.96     1.78     1.71     1.24
#>  Sus_scrofa                :    85.53     5.39     3.38     2.41     1.82     1.47
#>  Gallus_gallus             :    88.21     4.15     2.98     1.86     1.56     1.24
```

The percentage of information represented by different dimensions of the two approaches looks very similar.

5 Project libraries into common space

Let us first compute Δ^{-1}

```
d_inv_NoVupdate <- list()
for (sp in names(avg_counts)) {
  if (length(sbf_noVupdate$d[[sp]]) == 1) {
    d_inv_NoVupdate[[sp]] <- as.matrix(diag(as.matrix(1 / sbf_noVupdate$d[[sp]])))
    } else {
      d_inv_NoVupdate[[sp]] <- as.matrix(diag(1 / sbf_noVupdate$d[[sp]]))
    }
}
d_inv <- list()
for (sp in names(avg_counts)) {
  if (length(sbf$d[[sp]]) == 1) {
    d_inv[[sp]] <- as.matrix(diag(as.matrix(1 / sbf$d[[sp]])))
    } else {
      d_inv[[sp]] <- as.matrix(diag(1 / sbf$d[[sp]]))
    }
}
```

Project individual profiles and average counts to common space by computing $D_i^T U_i \Delta^{-1}$

```
# project using no V update estimates
avgcounts_projected_noVupdate  <- counts_projected_noVupdate <- list()
n_noVupdate <- n1_noVupdate <- c()
for (sp in names(avg_counts)) {
  avgcounts_projected_noVupdate[[sp]] <- as.matrix(t(avg_counts[[sp]])) %*%
    as.matrix(sbf_noVupdate$u[[sp]]) %*% d_inv_NoVupdate[[sp]]
  n_noVupdate <- c(n_noVupdate, row.names(avgcounts_projected_noVupdate[[sp]]))
  counts_projected_noVupdate[[sp]] <- as.matrix(t(counts_list_sub[[sp]])) %*%
    as.matrix(sbf_noVupdate$u[[sp]]) %*% d_inv_NoVupdate[[sp]]
  n1_noVupdate <- c(n1_noVupdate, row.names(counts_projected_noVupdate[[sp]]))
}

# combine projected profiles
df_proj_avg_noVupdate <- as.data.frame(do.call(rbind, avgcounts_projected_noVupdate))
rownames(df_proj_avg_noVupdate) <- n_noVupdate
colnames(df_proj_avg_noVupdate) <- paste0("Dim", 1:ncol(df_proj_avg_noVupdate))
meta <- tstrsplit(row.names(df_proj_avg_noVupdate), "_")
df_proj_avg_noVupdate$tissue <- factor(meta[[2]])
df_proj_avg_noVupdate$species <- factor(meta[[1]])
df_proj_avg_noVupdate <- df_proj_avg_noVupdate %>% mutate(species = factor(species,
```

```
                            levels = species_short))


df_proj_noVupdate <- as.data.frame(do.call(rbind, counts_projected_noVupdate))
rownames(df_proj_noVupdate) <- n1_noVupdate
colnames(df_proj_noVupdate) <- paste0("Dim", 1:ncol(df_proj_noVupdate))
meta1 <- data.table::tstrsplit(row.names(df_proj_noVupdate), "_")
df_proj_noVupdate$tissue <- factor(meta1[[3]])
df_proj_noVupdate$species <- factor(meta1[[2]])
df_proj_noVupdate <- df_proj_noVupdate %>% mutate(species = factor(species,
                            levels = species_short))

# project using V update estimates
avgcounts_projected  <- counts_projected <- list()
n <- n1 <- c()
for (sp in names(avg_counts)) {
  avgcounts_projected[[sp]] <- as.matrix(t(avg_counts[[sp]])) %*%
    as.matrix(sbf$u[[sp]]) %*% d_inv[[sp]]
  n <- c(n, row.names(avgcounts_projected[[sp]]))
  counts_projected[[sp]] <- as.matrix(t(counts_list_sub[[sp]])) %*%
    as.matrix(sbf$u[[sp]]) %*% d_inv[[sp]]
  n1 <- c(n1, row.names(counts_projected[[sp]]))
}

# combine projected profiles
df_proj_avg <- as.data.frame(do.call(rbind, avgcounts_projected))
rownames(df_proj_avg) <- n
colnames(df_proj_avg) <- paste0("Dim", 1:ncol(df_proj_avg))
meta <- data.table::tstrsplit(row.names(df_proj_avg), "_")
df_proj_avg$tissue <- factor(meta[[2]])
df_proj_avg$species <- factor(meta[[1]])
df_proj_avg <- df_proj_avg %>% mutate(species = factor(species,
                            levels = species_short))


df_proj <- as.data.frame(do.call(rbind, counts_projected))
rownames(df_proj) <- n1
colnames(df_proj) <- paste0("Dim", 1:ncol(df_proj))
meta1 <- tstrsplit(row.names(df_proj), "_")
df_proj$tissue <- factor(meta1[[3]])
df_proj$species <- factor(meta1[[2]])
df_proj <- df_proj %>% mutate(species = factor(species,
                            levels = species_short))
```

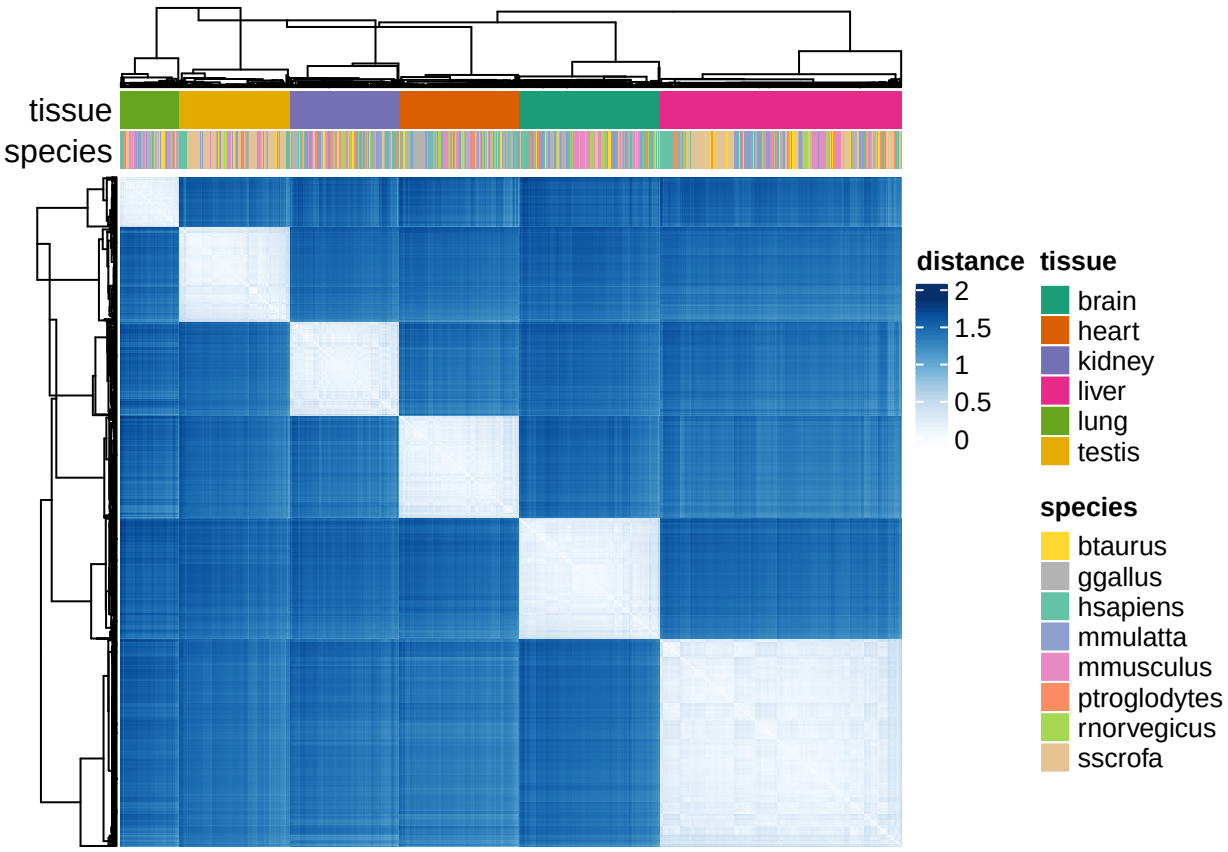
6 Clustering in the common space

```
# install packages
pkgs <- c("ComplexHeatmap", "RColorBrewer")
require_install <- pkgs[!(pkgs %in% row.names(installed.packages()))]
if (length(require_install))
  install.packages(require_install)
suppressPackageStartupMessages({
```

```
  library(ComplexHeatmap)
  library(RColorBrewer)
})
```

Compute distances between projected profiles and perform clustering.

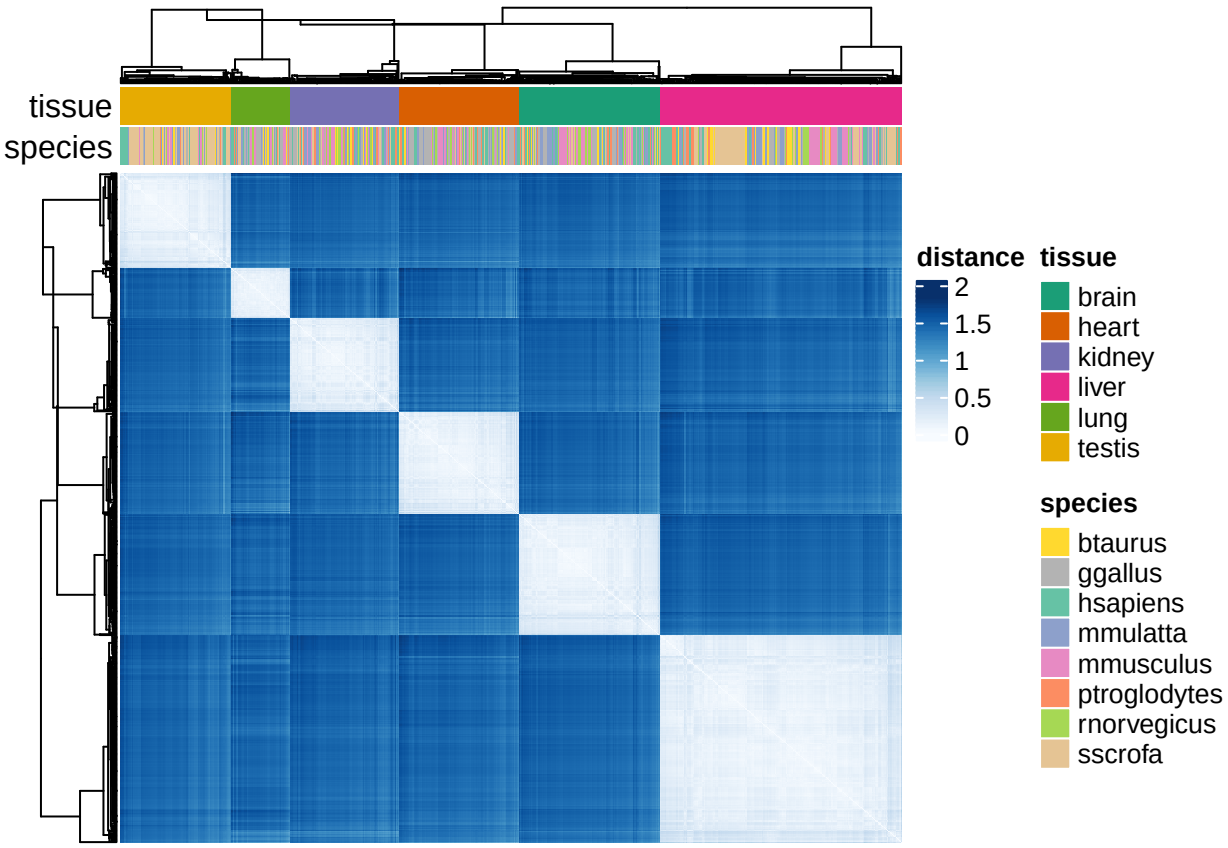
```
data_noVupdate <- df_proj_noVupdate
data_noVupdate$tissue <- NULL
data_noVupdate$species <- NULL
data_noVupdate <- as.matrix(data_noVupdate)
data_noVupdate_dist <- as.matrix(dist(data_noVupdate, method = "euclidean"))
meta <- data.table::tstrsplit(colnames(data_noVupdate_dist), "_")
ht <- ComplexHeatmap::HeatmapAnnotation(tissue = meta[[3]], species = meta[[2]],
                        col = list(tissue = c("brain" = "#1B9E77",
                                              "heart" = "#D95F02",
                                              "kidney" = "#7570B3",
                                              "liver" = "#E7298A",
                                              "lung" = "#66A61E",
                                              "testis" = "#E6AB02"),
                                   species = c("hsapiens" = "#66C2A5",
                                               "ptroglodytes" = "#FC8D62",
                                               "mmulatta" = "#8DA0CB",
                                               "mmusculus" = "#E78AC3",
                                               "rnorvegicus" = "#A6D854",
                                               "btaurus" = "#FFD92F",
                                               "sscrofa" = "#E5C494",
                                               "ggallus" = "#B3B3B3")),
                        #show_annotation_name = F,
                        annotation_name_side = "left")
mypalette <- RColorBrewer::brewer.pal(9, "Blues")
morecolors <- colorRampPalette(mypalette)

myheatmap1 <- ComplexHeatmap::Heatmap(as.matrix(data_noVupdate_dist), cluster_rows = T,
                      clustering_method_rows = "centroid",
                      cluster_columns = T,
                      clustering_method_columns = "centroid",
                      top_annotation = ht, col = morecolors(50),
                      show_row_names = F, show_column_names = F,
                      name = "distance")
myheatmap1
```



Compute distances between projected profiles in the common space estimated using optimized V and perform clustering.

```
data <- df_proj
data$tissue <- NULL
data$species <- NULL
data <- as.matrix(data)
data_dist <- as.matrix(dist(data, method = "euclidean"))
meta <- data.table::tstrsplit(colnames(data_dist), "_")
ht <- ComplexHeatmap::HeatmapAnnotation(tissue = meta[[3]], species = meta[[2]],
                        col = list(tissue = c("brain" = "#1B9E77",
                                              "heart" = "#D95F02",
                                              "kidney" = "#7570B3",
                                              "liver" = "#E7298A",
                                              "lung" = "#66A61E",
                                              "testis" = "#E6AB02"),
                                   species = c("hsapiens" = "#66C2A5",
                                               "ptroglodytes" = "#FC8D62",
                                               "mmulatta" = "#8DA0CB",
                                               "mmusculus" = "#E78AC3",
                                               "rnorvegicus" = "#A6D854",
                                               "btaurus" = "#FFD92F",
                                               "sscrofa" = "#E5C494",
                                               "ggallus" = "#B3B3B3")),
                        #show_annotation_name = F,
                        annotation_name_side = "left")
mypalette <- RColorBrewer::brewer.pal(9, "Blues")
```

```
morecolors <- colorRampPalette(mypalette)

myheatmap <- ComplexHeatmap::Heatmap(as.matrix(data_dist), cluster_rows = T,
                      clustering_method_rows = "centroid",
                      cluster_columns = T,
                      clustering_method_columns = "centroid",
                      top_annotation = ht, col = morecolors(50),
                      show_row_names = F, show_column_names = F,
                      name = "distance")
myheatmap
```



Gene expression profiles cluster by tissue type independent of the species of origin.

7 Two-dimensional projection plots in the common space

Next, we will explore the 2D projection plots in the common space. We will first define a custom theme that we will use for the plots.

```
# install packages
pkgs <- c("grid", "ggthemes", "ggplot2")
require_install <- pkgs[!(pkgs %in% row.names(installed.packages()))]
if (length(require_install))
  install.packages(require_install)
suppressPackageStartupMessages({
  library(grid)
  library(ggthemes)
  library(ggplot2)
```

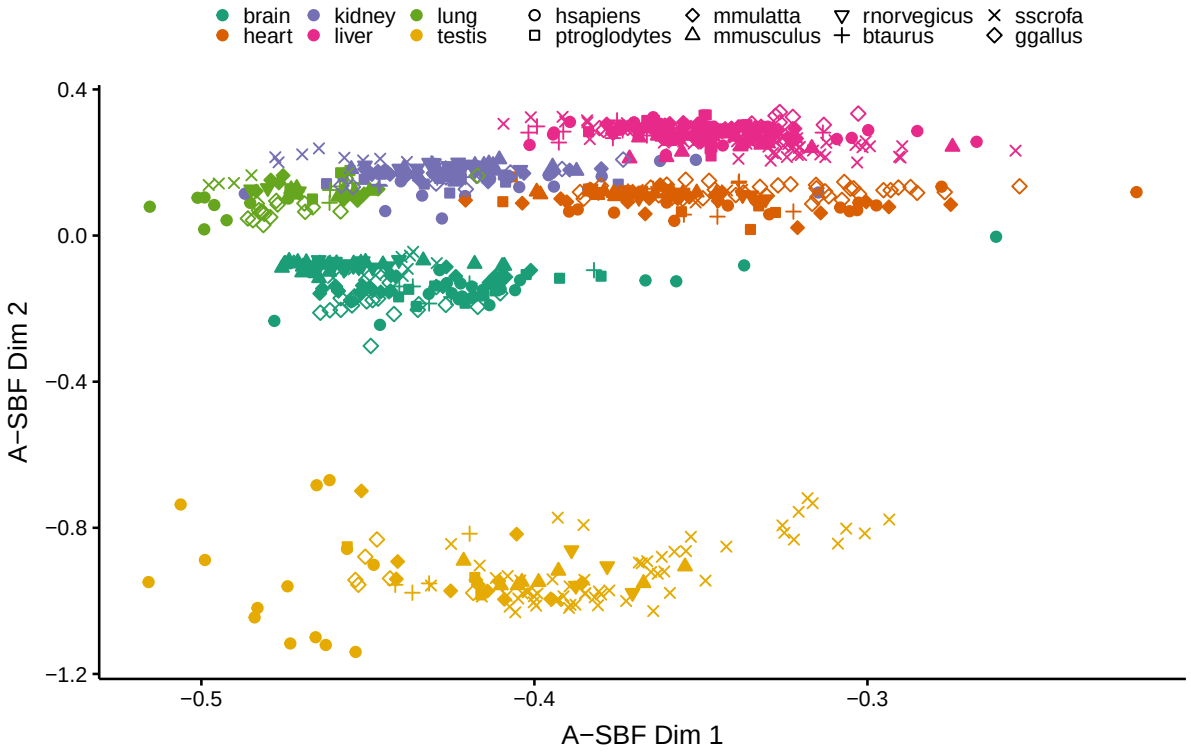
```
})
```

We will use the following custom theme for ggplots.

```
# custom theme function for ggplot2
customTheme <- function(base_size = 10, base_family="helvetica") {
   require(grid)
   require(ggthemes)
   (ggthemes::theme_foundation(base_size = base_size)
   + ggplot2::theme(plot.title = element_text(face = "bold",
                                    size = rel(1.2), hjust = 0.5),
         text = element_text(),
         panel.background = element_rect(colour = NA),
         plot.background = element_rect(colour = NA),
         panel.border = element_rect(colour = NA),
         axis.title = element_text(size = rel(1)),
         axis.title.y = element_text(angle = 90, vjust = 2),
         axis.title.x = element_text(vjust = -0.2),
         axis.text = element_text(),
         axis.line = element_line(colour = "black"),
         axis.ticks = element_line(),
         panel.grid.major = element_blank(),
         panel.grid.minor = element_blank(),
         legend.key = element_rect(colour = NA),
         legend.position = "top",
         legend.direction = "horizontal",
         legend.key.size = unit(0.2, "cm"),
         legend.spacing = unit(0, "cm"),
         legend.title = element_text(face = "italic"),
         plot.margin = unit(c(10, 5, 5, 5), "mm"),
         strip.background = element_rect(colour = "#f0f0f0", fill = "#f0f0f0"),
         strip.text = element_text(face = "bold")))
}
```

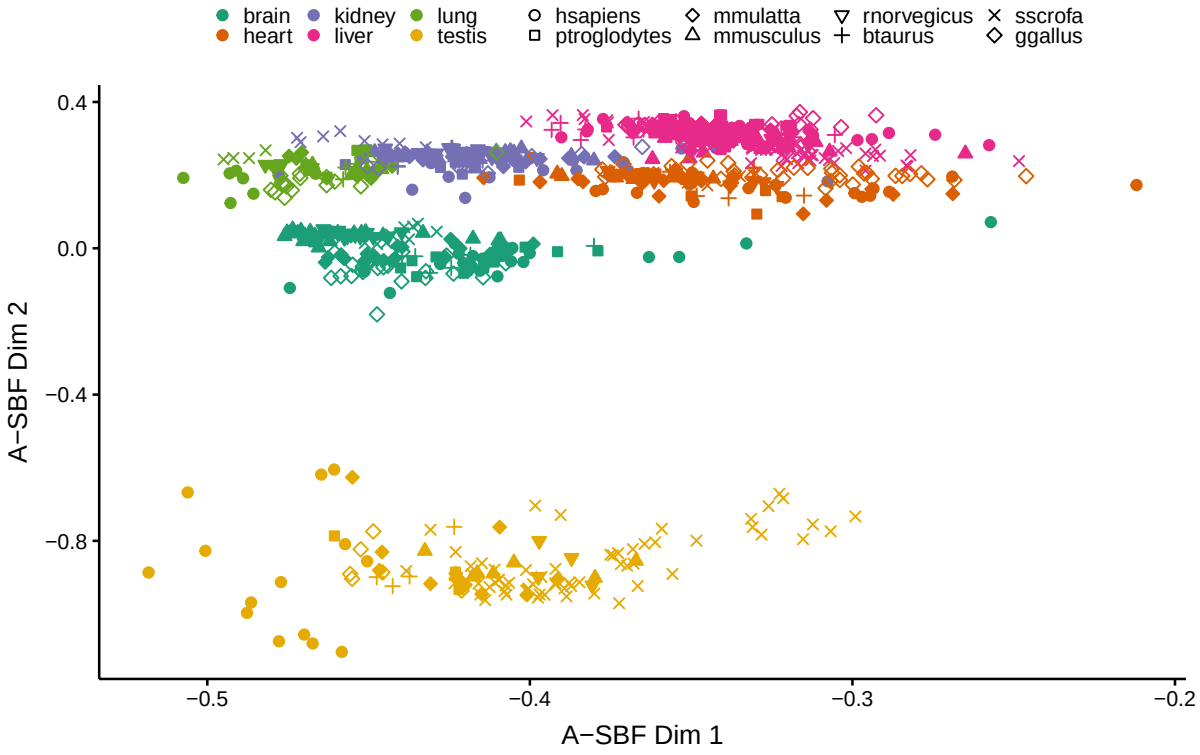
Let us first check the projected libraries in dimension 1 and 2.

```
# 2D plot for Dim1 and Dim2 [No V update]
sel_colors <- RColorBrewer::brewer.pal(8, "Dark2")[seq_len(length(unique(df_proj_noVupdate$tissue)))]
i <- 1
j <- 2
ggplot2::ggplot(df_proj_noVupdate, aes(x = df_proj_noVupdate[, i],
                              y = df_proj_noVupdate[, j], col = tissue,
                              shape = species, fill = tissue)) +
  xlab(paste("A-SBF Dim", i)) + ylab(paste("A-SBF Dim", j)) +
  geom_point(size = 1.5) + scale_color_manual(values = sel_colors) +
  scale_shape_manual(values = c(21:25, 3:7)) +
  scale_fill_manual(values = sel_colors) +
  customTheme() +
  theme(legend.title = element_blank())
```



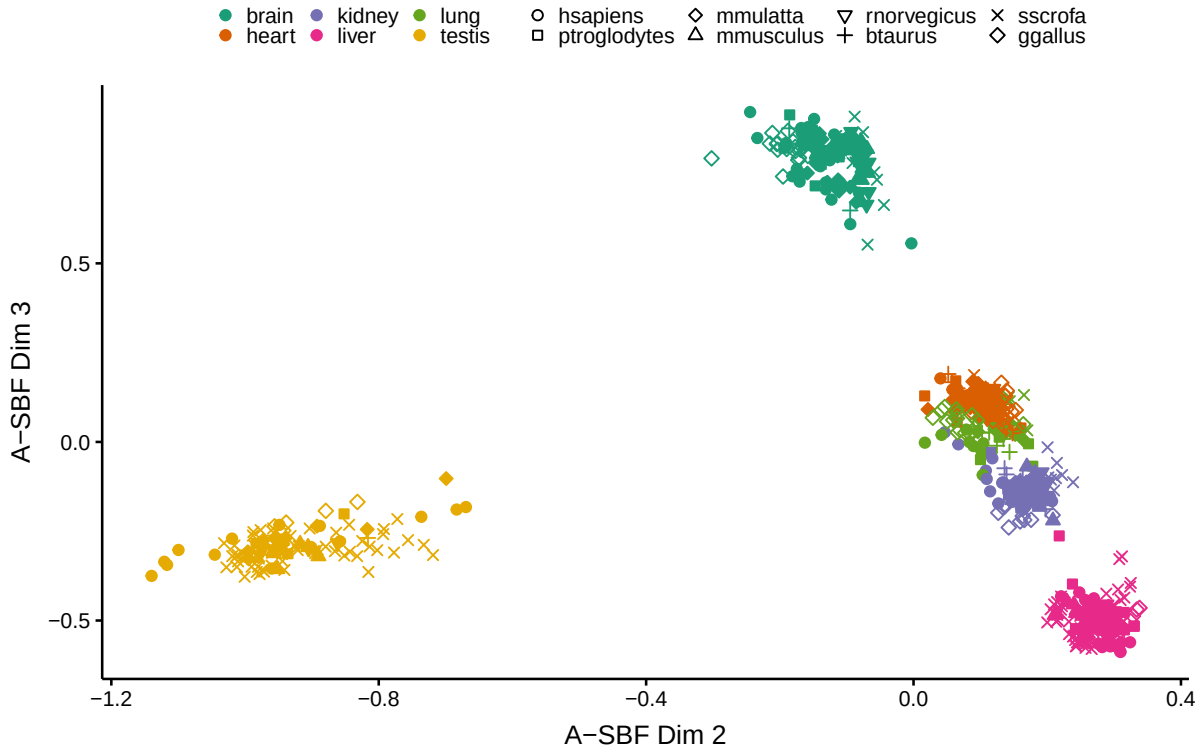
The same with optimized V

```
# 2D plot for Dim1 and Dim2 [With V update]
sel_colors <- RColorBrewer::brewer.pal(8, "Dark2")[seq_len(length(unique(df_proj$tissue)))]
i <- 1
j <- 2
ggplot2::ggplot(df_proj, aes(x = df_proj[, i],
                             y = df_proj[, j], col = tissue,
                             shape = species, fill = tissue)) +
  xlab(paste("A-SBF Dim", i)) + ylab(paste("A-SBF Dim", j)) +
  geom_point(size = 1.5) + scale_color_manual(values = sel_colors) +
  scale_shape_manual(values = c(21:25, 3:7)) +
  scale_fill_manual(values = sel_colors) +
  customTheme() +
  theme(legend.title = element_blank())
```

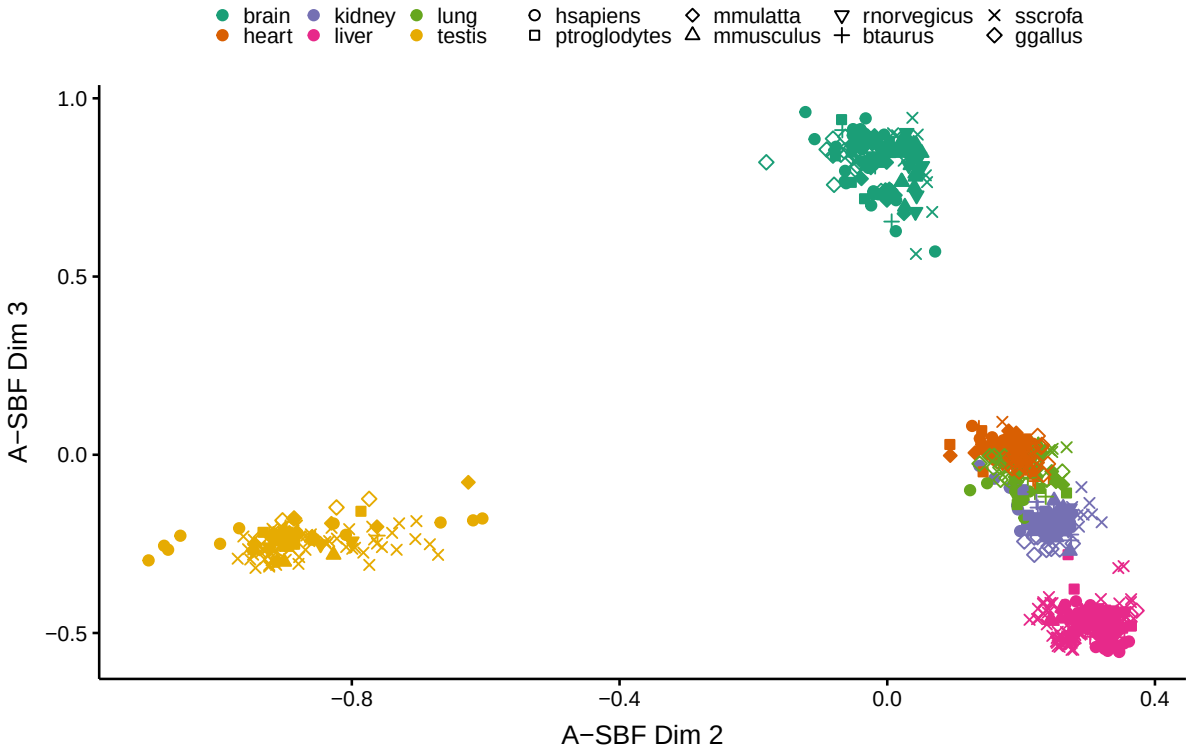


Let us plot the projected libraries in dimensions 2 and 3.

```
# 2D plot for Dim1 and Dim2 [No V update]
sel_colors <- RColorBrewer::brewer.pal(8, "Dark2")[seq_len(length(unique(df_proj_noVupdate$tissue)))]
i <- 2
j <- 3
ggplot2::ggplot(df_proj_noVupdate, aes(x = df_proj_noVupdate[, i],
                                       y = df_proj_noVupdate[, j], col = tissue,
                                       shape = species, fill = tissue)) +
  xlab(paste("A-SBF Dim", i)) + ylab(paste("A-SBF Dim", j)) +
  geom_point(size = 1.5) + scale_color_manual(values = sel_colors) +
  scale_shape_manual(values = c(21:25, 3:7)) +
  scale_fill_manual(values = sel_colors) +
  customTheme() +
  theme(legend.title = element_blank())
```




```
# 2D plot for Dim1 and Dim2 [With V update]
sel_colors <- RColorBrewer::brewer.pal(8, "Dark2")[seq_len(length(unique(df_proj$tissue)))]
i <- 2
j <- 3
ggplot2::ggplot(df_proj, aes(x = df_proj[, i], y = df_proj[, j], col = tissue,
                             shape = species, fill = tissue)) +
  xlab(paste("A-SBF Dim", i)) + ylab(paste("A-SBF Dim", j)) +
  geom_point(size = 1.5) + scale_color_manual(values = sel_colors) +
  scale_shape_manual(values = c(21:25, 3:7)) +
  scale_fill_manual(values = sel_colors) +
  customTheme() +
  theme(legend.title = element_blank())
```



Similarly, we can check for other dimensions of the common space. We observe that both clustering and 2D projections plots in the common space with optimized V are similar to those without V update. So we will use the common space with optimized V for future analysis.

8 Expression specificity and eigengene loadings

Functions to compute tissue specificity (τ) and scaled average expression profile.

```
# function to compute Tau
calc_tissue_specificity <- function(a) {
    a <- as.matrix(a)
    b <- a / matrixStats::rowMaxs(a)
    return(rowSums(1 - b) / (ncol(b) - 1))
}

Tau <- lapply(avg_counts, function(x) { calc_tissue_specificity(x)})
avg_counts_scaled <- lapply(avg_counts, function(x) { t(scale(t(x)))})

combine_expr <- list()
for (sp in names(avg_counts_scaled)) {
  x <- as.data.frame(avg_counts_scaled[[sp]])
  x[["Tau"]] <- Tau[[sp]]
  combine_expr[[sp]] <- x
}
```

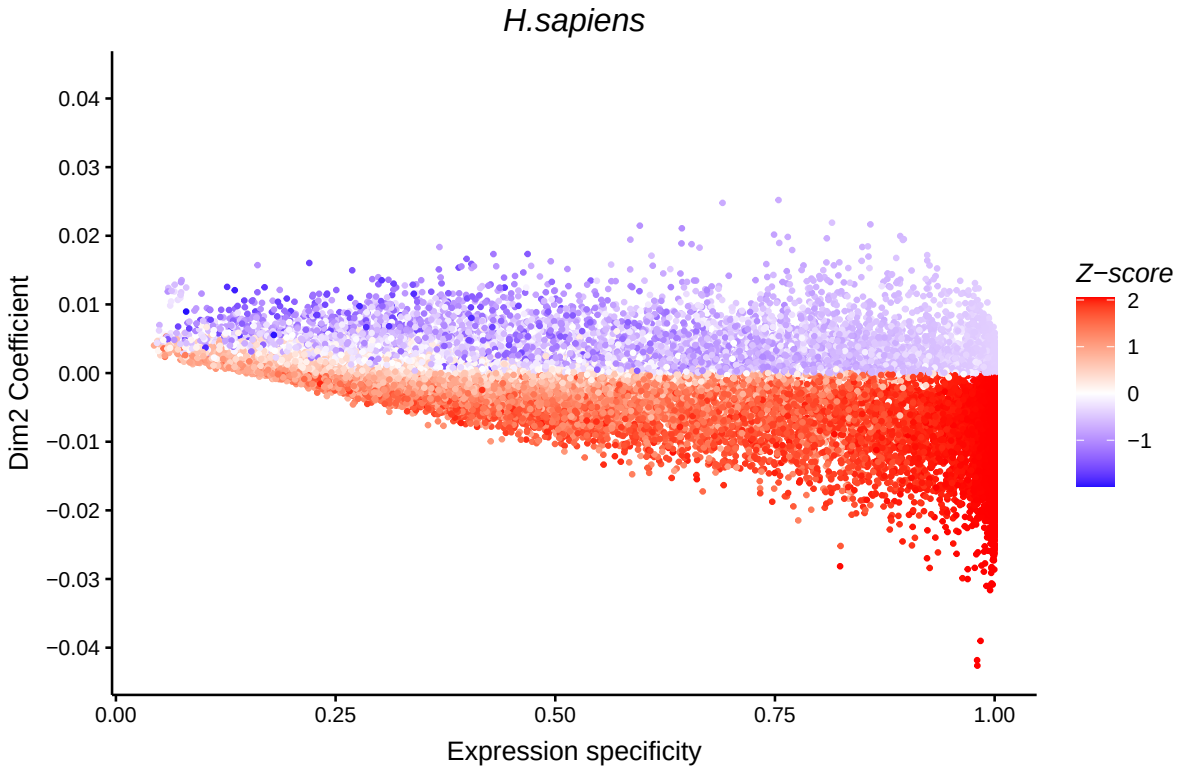
From the 2D projection plot, we find that along the dimension 2, testis profiles lie on the negative axis and separate from the rest of the tissues. We will plot the relationship between U_i loadings in dimension 2 and the testis expression.

```
sel_dim <- 2
sel_tissue <- "testis"
```

```
species <- "Homo_sapiens"
expr <- combine_expr[[species]]
asbf_coef <- sbf$u[[species]]
expr[["coef"]] <- asbf_coef[, sel_dim, drop = T]
expr1 <- expr[, c(paste0(getSpeciesShortName(species), "_", sel_tissue),
                  "Tau", "coef")]
colnames(expr1) <- c("tissue_zscore", "Tau", "coef")
head(expr1)
#>                 tissue_zscore       Tau          coef
#> ENSG00000000003    1.15812809 0.3824796 -0.0027709444
#> ENSG00000000005   -0.05534986 0.8531503  0.0002702072
#> ENSG00000000419    1.43944024 0.1848732  0.0002641079
#> ENSG00000000457    1.15751198 0.3444267 -0.0017597439
#> ENSG00000000460    1.95491402 0.8630708 -0.0106484590
#> ENSG00000000938   -0.54023331 0.6759041  0.0076031503
```

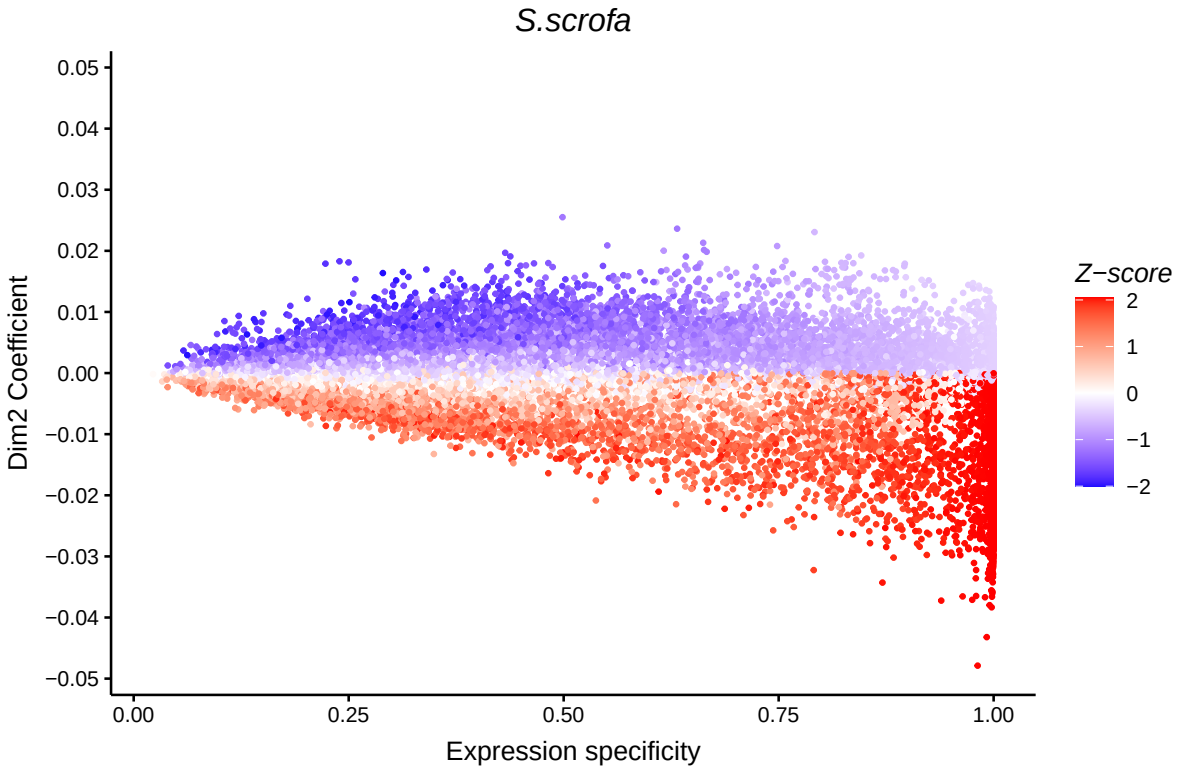
Dimension 2 U_i loadings vs expression specificity (τ) for humans

```
# function to return scientific name
species_scientific <- function(x) {
  parts <- data.table::tstrsplit(x, "_")
  return(paste0(substr(x, 1, 1), ".", parts[[2]]))
}

# plot scatter
mid <- 0
p1 <- ggplot2::ggplot(expr1, aes(x = Tau, y = coef, col = tissue_zscore)) + theme_bw() +
  geom_point(size = 0.5) + xlab("Expression specificity") +
  ylab(paste0("Dim", sel_dim, " Coefficient")) +
  scale_color_gradient2(midpoint = mid, low = "blue", mid = "white",
                        high = "red", space = "Lab") +
  scale_y_continuous(limits = c(-1 * max(abs(expr1$coef)), max(abs(expr1$coef))),
                     breaks = seq(-1 * round(max(abs(expr$coef)), 2),
                                  round(max(abs(expr$coef)), 2), by = 0.01)) +
  customTheme() +  theme(legend.position = "right",
                         legend.direction = "vertical") +
  labs(title = paste0(species_scientific(species)), color = "Z-score") +
  theme(legend.key.size = unit(0.5, "cm"),
        plot.title = element_text(face = "italic"))
p1
```



Dimension 2 U_i loadings vs expression specificity (τ) for pig.

```
sel_dim <- 2
sel_tissue <- "testis"
species <- "Sus_scrofa"
expr <- combine_expr[[species]]
asbf_coef <- sbf$u[[species]]
expr[["coef"]] <- asbf_coef[, sel_dim, drop = T]
expr1 <- expr[, c(paste0(getSpeciesShortName(species), "_", sel_tissue),
                  "Tau", "coef")]
colnames(expr1) <- c("tissue_zscore", "Tau", "coef")

# plot scatter
mid <- 0
p2 <- ggplot2::ggplot(expr1, aes(x = Tau, y = coef, col = tissue_zscore)) + theme_bw() +
  geom_point(size = 0.5) + xlab("Expression specificity") +
  ylab(paste0("Dim", sel_dim, " Coefficient")) +
  scale_color_gradient2(midpoint = mid, low = "blue", mid = "white",
                        high = "red", space = "Lab") +
  scale_y_continuous(limits = c(-1 * max(abs(expr1$coef)), max(abs(expr1$coef))),
                     breaks = seq(-1 * round(max(abs(expr$coef)), 2),
                                  round(max(abs(expr$coef)), 2), by = 0.01)) +
  customTheme() +  theme(legend.position = "right",
                         legend.direction = "vertical") +
  labs(title = paste0(species_scientific(species)), color = "Z-score") +
  theme(legend.key.size = unit(0.5, "cm"),
        plot.title = element_text(face = "italic"))
p2
```



9 GO analysis

Download the GO files from: https://figshare.com/s/d96c586d5e53199d5370 We will perform the gene ontology analysis for genes with high coefficients. For GO analysis, we will use shared 1-to-1 orthologs in eight species as the background.

```
# load GO analysis files
# go_path <- "path to shared 1-to-1 orthologs"
go_path <- "~/Dropbox/0.Analysis/0.paper/GOKeggFiles/"
file <- "allwayOrthologs_hsapiens-ptroglodytes-mmulatta-mmusculus-rnorvegicus-btaurus-sscrofa-ggallus_e
orthologs <- read.table(file.path(go_path, file), header = T, sep = "\t")
head(orthologs)
#>          hsapiens external_gene_name   gene_biotype       ptroglodytes
#> 1 ENSG00000223224            SNORD71         snoRNA ENSPTRG00000036308
#> 2 ENSG00000199574           SNORD18C         snoRNA ENSPTRG00000026302
#> 3 ENSG00000284202             MIR137          miRNA ENSPTRG00000027612
#> 4 ENSG00000207797             MIR187          miRNA ENSPTRG00000050189
#> 5 ENSG00000198888             MT-ND1 protein_coding ENSPTRG00000042641
#> 6 ENSG00000198763             MT-ND2 protein_coding ENSPTRG00000042626
#>             mmulatta          mmusculus        rnorvegicus             btaurus
#> 1 ENSMMUG00000034773 ENSMUSG00000077549 ENSRNOG00000059926 ENSBTAG00000043472
#> 2 ENSMMUG00000024353 ENSMUSG00000064514 ENSRNOG00000059435 ENSBTAG00000043531
#> 3 ENSMMUG00000026871 ENSMUSG00000065569 ENSRNOG00000035491 ENSBTAG00000029809
#> 4 ENSMMUG00000027111 ENSMUSG00000065532 ENSRNOG00000035521 ENSBTAG00000029796
#> 5 ENSMMUG00000028699 ENSMUSG00000064341 ENSRNOG00000030644 ENSBTAG00000043558
#> 6 ENSMMUG00000028695 ENSMUSG00000064345 ENSRNOG00000031033 ENSBTAG00000043571
#>              sscrofa            ggallus hsapiens_genelength
#> 1 ENSSSCG00000019395 ENSGALG00000025302                  86
#> 2 ENSSSCG00000018926 ENSGALG00000043763                  69
```

```
#> 3 ENSSSCG00000019038 ENSGALG00000018336                 102
#> 4 ENSSSCG00000019111 ENSGALG00000018285                 109
#> 5 ENSSSCG00000018065 ENSGALG00000042750                 956
#> 6 ENSSSCG00000018069 ENSGALG00000043768                1042
#>   ptroglodytes_genelength mmulatta_genelength mmusculus_genelength
#> 1                      86                  86                   85
#> 2                      69                  70                   71
#> 3                     102                 102                   73
#> 4                     109                 109                   61
#> 5                     957                 955                  957
#> 6                    1044                1042                 1038
#>   rnorvegicus_genelength btaurus_genelength sscrofa_genelength
#> 1                     84                 86                 86
#> 2                     71                 70                 70
#> 3                    102                102                102
#> 4                    104                109                104
#> 5                    955                956                955
#> 6                   1039               1042               1042
#>   ggallus_genelength
#> 1                 87
#> 2                 71
#> 3                 96
#> 4                 86
#> 5                975
#> 6               1041
```

We will use the goseq Bioconductor package to perform GO enrichment analysis.

```
# install packages
pkgs <- c("goseq")
require_install <- pkgs[!(pkgs %in% row.names(installed.packages()))]
if (length(require_install)) {
  if (!require("BiocManager", quietly = TRUE))
      install.packages("BiocManager")
  BiocManager::install("goseq")
}
suppressPackageStartupMessages({
  library(goseq)
})
```

Let us perform GO analysis for top testis-specific genes in Dimension 2. The testis-specific genes have negative loadings.

```
sel_dim <- 2
sel_tissue <- "testis"
top_genes <- 100
sel_sign <- "neg"
```

We will perform GO analysis for human first.

```
species <- "Homo_sapiens"
expr <- combine_expr[[species]]
asbf_coef <- sbf$u[[species]]
expr[["coef"]] <- asbf_coef[, sel_dim, drop = T]
expr1 <- expr[, c(paste0(getSpeciesShortName(species), "_", sel_tissue), "Tau",
                  "coef")]
```

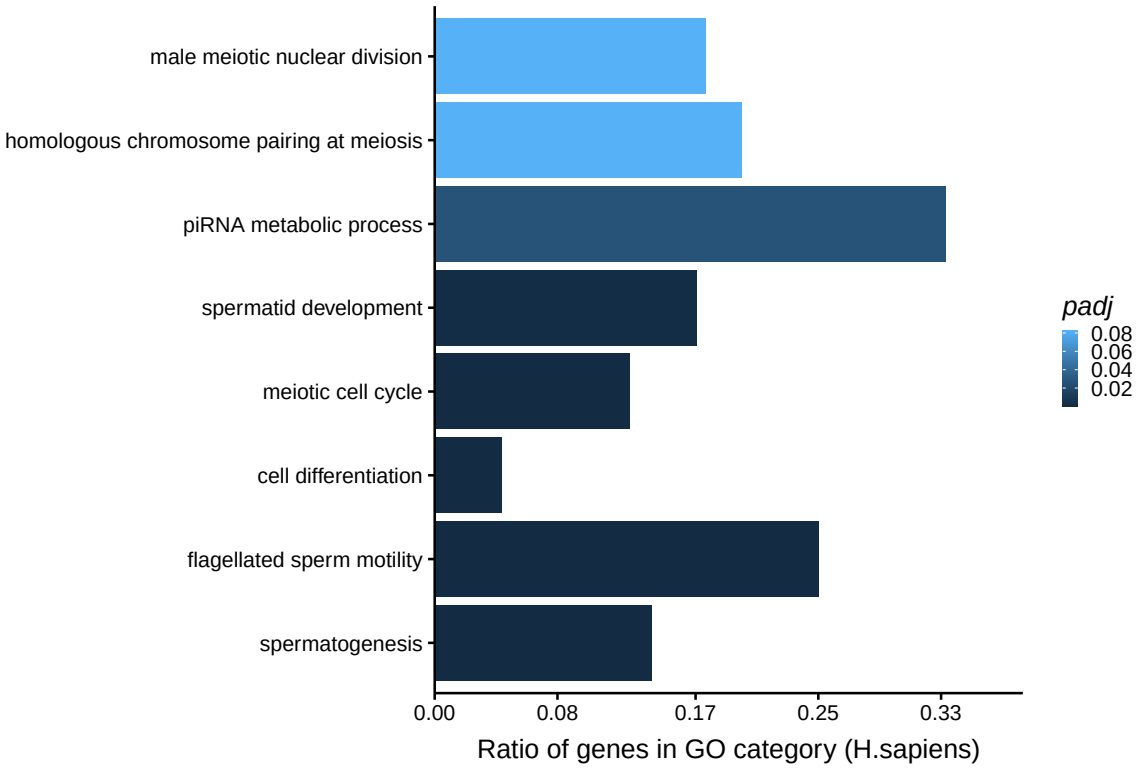
```
colnames(expr1) <- c("tissue_zscore", "Tau", "coef")
if (sel_sign == "neg") {
  cat("\n selecting negative loadings")
  expr1_selsign <- expr1[expr1$coef < 0, ]
} else {
  cat("\n selecting positive loadings")
  expr1_selsign <- expr1[expr1$coef >= 0, ]
}
#>
#>  selecting negative loadings
expr1_selsign$score <- expr1_selsign$Tau * abs(expr1_selsign$coef)
expr1_selsign$rank <- rank(-1 * expr1_selsign$score)
expr1_selsign <- expr1_selsign[order(expr1_selsign$rank), ]
genes_fg <- row.names(expr1_selsign[expr1_selsign$rank <= top_genes, ])
if (species == "Homo_sapiens") {
  cat("\nUsing orthologs (human IDs) as background")
  genes_bg <- orthologs$hsapiens
}
#>
#> Using orthologs (human IDs) as background
if (species == "Mus_musculus") {
  cat("\nUsing orthologs (mouse IDs) as background")
  genes_bg <- orthologs$mmusculus
}
genes_bg <- genes_bg[!genes_bg %in% genes_fg]
genome <- "hg38"
total_genes <- unique(c(genes_fg, genes_bg))
up_genes <- as.integer(total_genes %in%  genes_fg)
names(up_genes) <- total_genes
```

```
#> Using manually entered categories.
#> For 183 genes, we could not find any categories. These genes will be excluded.
#> To force their use, please run with use_genes_without_cat=TRUE (see documentation).
#> This was the default behavior for version 1.15.1 and earlier.
#> Calculating the p-values...
#> 'select()' returned 1:1 mapping between keys and columns
```

```
head(go.sub)
#>        category numDEInCat numInCat                        term ontology
#> 2863 GO:0007283         28      198            spermatogenesis       BP
#> 5356 GO:0030317          8       32 flagellated sperm motility       BP
#> 5253 GO:0030154         22      500       cell differentiation       BP
#> 9828 GO:0051321          7       55         meiotic cell cycle       BP
#> 2865 GO:0007286          7       41      spermatid development       BP
#> 6760 GO:0034587          3        9    piRNA metabolic process       BP
#>              padj  ratio
#> 2863 2.153710e-19 0.1414
#> 5356 2.322728e-05 0.2500
#> 5253 1.169276e-04 0.0440
#> 9828 3.729645e-04 0.1273
#> 2865 1.555172e-03 0.1707
#> 6760 2.642197e-02 0.3333
```

Barplot with top human GO terms and their p-value.

```
# GO enrichment plot for human
go_out <- head(go.sub, n = 8)
go_out$padj <- as.numeric(go_out$padj)
go_out$term <- factor(go_out$term, levels = go_out$term)
breaks <- round(c(0, 1 / 4, 2 / 4, 3 / 4, 1) * max(go_out[["ratio"]]), 2)
go_plot <- ggplot2::ggplot(go_out, aes(x = term, y = ratio, fill = padj)) + geom_col() +
  scale_y_continuous(expand = c(0, 0), breaks = breaks,
                     limits = c(0, max(go_out[["ratio"]] + 0.05))) +
  scale_x_discrete() + coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  ylab(paste0("Ratio of genes in GO category (",
              species_scientific(species), ")")) +
  xlab("") + customTheme() + theme(legend.position = "right",
                                   legend.direction = "vertical",
                                   plot.margin = unit(c(10, 5, 5, 5), "mm"))
go_plot
```



GO analysis for mouse.

```
species <- "Mus_musculus"
expr <- combine_expr[[species]]
asbf_coef <- sbf$u[[species]]
expr[["coef"]] <- asbf_coef[, sel_dim, drop = T]
expr1 <- expr[, c(paste0(getSpeciesShortName(species), "_", sel_tissue), "Tau",
                  "coef")]
colnames(expr1) <- c("tissue_zscore", "Tau", "coef")
if (sel_sign == "neg") {
  cat("\n selecting negative loadings")
  expr1_selsign <- expr1[expr1$coef < 0, ]
} else {
```

```
  cat("\n selecting positive loadings")
  expr1_selsign <- expr1[expr1$coef >= 0, ]
}
#>
#>  selecting negative loadings
expr1_selsign$score <- expr1_selsign$Tau * abs(expr1_selsign$coef)
expr1_selsign$rank <- rank(-1 * expr1_selsign$score)
expr1_selsign <- expr1_selsign[order(expr1_selsign$rank), ]
genes_fg <- row.names(expr1_selsign[expr1_selsign$rank <= top_genes, ])
if (species == "Homo_sapiens") {
  cat("\nUsing orthologs (human IDs) as background")
  genes_bg <- orthologs$hsapiens
}
if (species == "Mus_musculus") {
  cat("\nUsing orthologs (mouse IDs) as background")
  genes_bg <- orthologs$mmusculus
}
#>
#> Using orthologs (mouse IDs) as background
genes_bg <- genes_bg[!genes_bg %in% genes_fg]
genome <- "mm10"
total_genes <- unique(c(genes_fg, genes_bg))
up_genes <- as.integer(total_genes %in%  genes_fg)
names(up_genes) <- total_genes
```

```
#> Using manually entered categories.
#> For 133 genes, we could not find any categories. These genes will be excluded.
#> To force their use, please run with use_genes_without_cat=TRUE (see documentation).
#> This was the default behavior for version 1.15.1 and earlier.
#> Calculating the p-values...
#> 'select()' returned 1:1 mapping between keys and columns
```

```
head(go.sub)
#>        category numDEInCat numInCat                       term ontology
#> 2794 GO:0007283         17      185            spermatogenesis       BP
#> 5203 GO:0030317          9       35 flagellated sperm motility       BP
#> 2965 GO:0008150         29      919         biological_process       BP
#> 2796 GO:0007286          7       48      spermatid development       BP
#> 3346 GO:0009566          4       22              fertilization       BP
#> 5171 GO:0030261          5       13    chromosome condensation       BP
#>              padj  ratio
#> 2794 1.404847e-06 0.0919
#> 5203 7.702713e-06 0.2571
#> 2965 2.268008e-04 0.0316
#> 2796 1.056220e-02 0.1458
#> 3346 7.827071e-02 0.1818
#> 5171 7.827071e-02 0.3846
```

Barplot with top mouse GO terms and their p-value.

```
# GO enrichment plot for mouse
go_out <- head(go.sub, n = 8)
go_out$padj <- as.numeric(go_out$padj)
go_out$term <- factor(go_out$term, levels = go_out$term)
breaks <- round(c(0, 1 / 4, 2 / 4, 3 / 4, 1) * max(go_out[["ratio"]]), 2)
```

```
go_plot <- ggplot2::ggplot(go_out, aes(x = term, y = ratio, fill = padj)) + geom_col() +
  scale_y_continuous(expand = c(0, 0), breaks = breaks,
                     limits = c(0, max(go_out[["ratio"]] + 0.05))) +
  scale_x_discrete() + coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  ylab(paste0("Ratio of genes in GO category (",
              species_scientific(species), ")")) +
  xlab("") + customTheme() + theme(legend.position = "right",
                                   legend.direction = "vertical",
                                   plot.margin = unit(c(10, 5, 5, 5), "mm"))
go_plot
```