enrichmotifpairR: An R package/shiny app for finding enriched transcription factor motifs and their binding partners

Naresh Doni Jayavelu, Nicholas Moss, and R. David Hawkins

04/13/2022

Contents

Quick start

```
# load the package and other useful packages
library(enrichmotifpairR)
# load the example data provided with package
data("example_peaks_data")
```

Finding enriched TF motifs and their binding partners in H1-ESC at DHS peaks

```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPair(
  target_data = example_peaks_data$`H1-ESC_DHS_peaks`,
  background_data = example_peaks_data$`H1-ESC_DHS_peaks_matched_background`,
  genome_ver = "hg38",
  scramble_data = F,
  motif_database = "ENCODE",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
  Pvalue_adjust_method = "BH"
)
```

The enriched TF motifs are stored in the `results$motif_enrich` and their binding partners in the `results$motif_pair_enrich`.

```
# assign the results data to individual objects
enrich_motifs <- results$motif_enrich
enrich_motif_pairs <- results$motif_pair_enrich
# The output of the enriched motifs top five
enrich_motifs %>% head(5) %>% kable(., caption="Top 5 enriched motifs")
```

Table 1: Top 5 enriched motifs

motif_name	TF_name	tg_motif_count	bg_motif_count	fold_enrich	pval	pval_adj
NFY_disc1	NFY	1821	179	10.1731844	0	0
CTCF_disc1	CTCF	2837	326	8.7024540	0	0

motif_name	TF_name	tg_motif_count	bg_motif_count	fold_enrich	pval	pval_adj
RAD21_disc1	RAD21	2792	398	7.0150754	0	0
IRF_disc1	IRF	1664	166	10.0240964	0	0
SP1_disc1	SP1	1716	186	9.2258065	0	0

```
# The output of the enriched motif pairs, top five binding partners for **NFY**
enrich_motif_pairs %>% head(5) %>%
    kable(., caption="Top 5 binding partners for NFY")
```

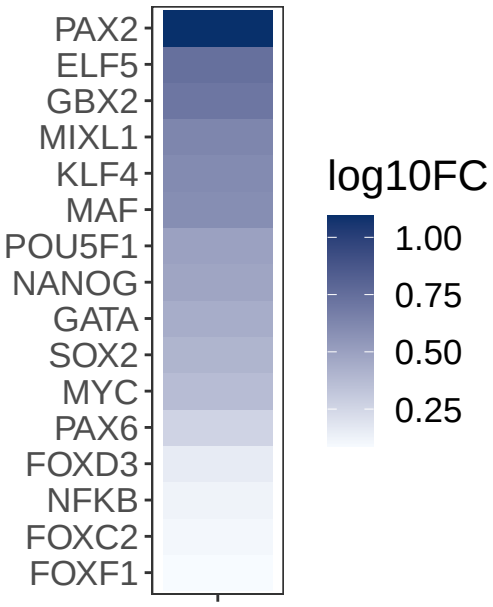
Table 2: Top 5 binding partners for NFY

motif_name_1	TF_name_1	motif_name_2	TF_name_2	fold_enrich	pval	pval_adj
NFY_disc1	NFY	TATA_disc6	TATA	1.5677851	0.0e+00	0.00000003
NFY_disc1	NFY	EN1_4	EN1	5.8978583	0.0e+00	0.00000017
NFY_disc1	NFY	SP1_disc1	SP1	1.4253158	0.0e+00	0.00000034
NFY_disc1	NFY	TATA_disc4	TATA	2.4936559	3.9e-07	0.00013382
NFY_disc1	NFY	TCF12_disc3	TCF12	7.3067911	4.8e-07	0.00014137

Before interpreting these results, either enriched TF motifs or their binding partners we strongly recommend to filter them based on their expression level, for instance at RPKM (FPKM) or TPM > 1. Next, we can choose selected TFs that are known to be involved based on prior knowledge or highly enriched ones. If you are interested to filter for only TF genes as defined by Lambert et al., 2018, you can do so as well.

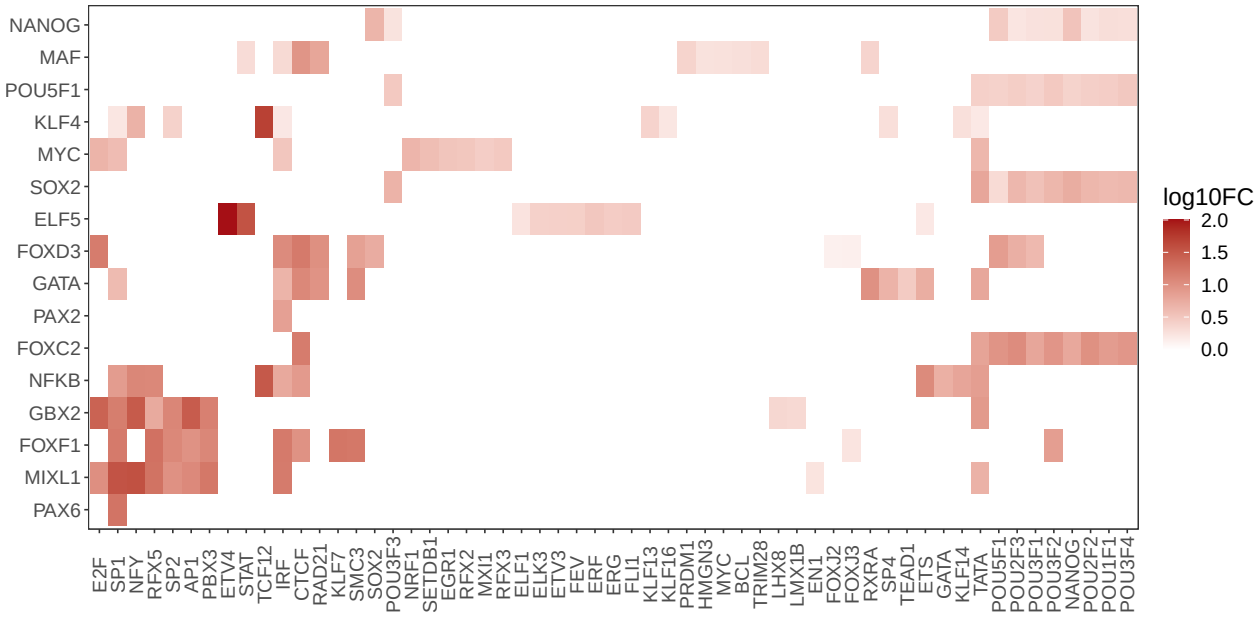
Now we can visualize the enriched motifs and enriched motif pairs using heatmaps.

```
# Select TFs for visualization. These are some well known TFs in
# H1-ESCs
TF_list <- c("SOX2", "NANOG", "MYC", "POU5F1", "STAT3", "TCF3", "KLF4", "SMAD1",
             "SMAD2", "SMAD3", "SMAD4", "SMAD5", "SMAD8", "TFAP2C", "KLF2",
             "KLF4", "KLF5", "ELF5", "ESRRB", "PRDM14", "DPPA4", "FOXC2",
             "FOXD3", "FOXF1", "GATA", "GBX2", "MAF", "MEF2C", "MAX", "JUN",
             "MIXL1", "NFKB", "PAX2", "PAX6", "SNAI1", "TBX6", "SUZ12")

# Plot motif enrichment heatmap
plotEnrichment(enrich_motifs = enrich_motifs,
               tfs = TF_list)
```



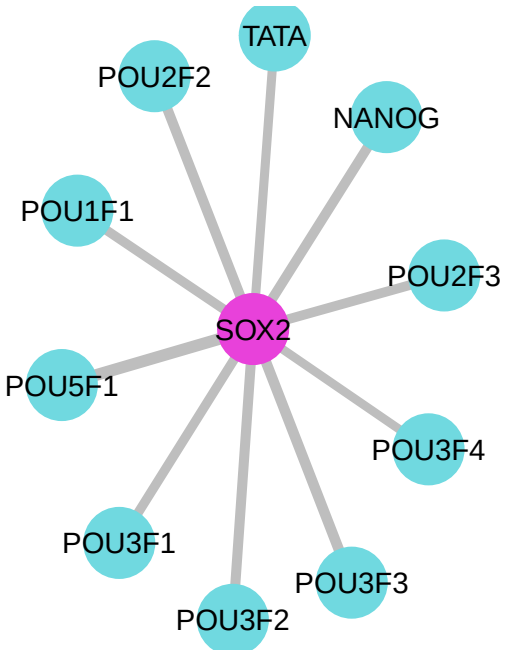
Now plot the top 10 binding partners for the above TF motifs.

```
plotEnrichPair(enrich_pairs = enrich_motif_pairs,
               tfs = TF_list)
```

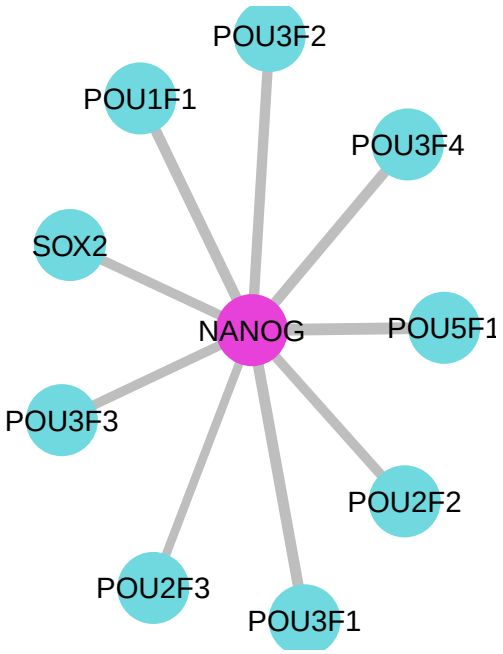


Now we can make colorful networks for select TFs of interest

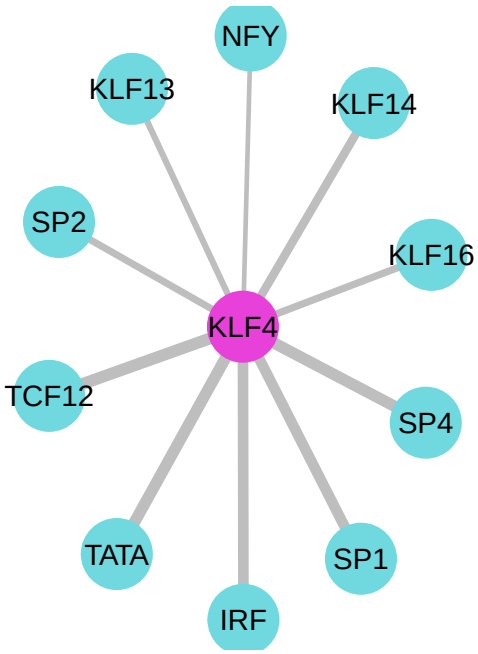
```
# select TF "SOX2" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "SOX2"
            )
```




```
# select TF "NANOG" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "NANOG"
            )
```




```
# select TF "KLF4" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "KLF4"
            )
```



Additional example use cases:

Example use case 1: Genomic regions from two conditions

Here, we have genomic regions from two conditions, for instance, cells treated vs untreated. To demonstrate this example, we obtained the ATAC-seq data in Th17 cells treated with a stimulus and untreated Th17 cells. We can find TFs and their binding partner TFs specifically enriched in stimulated cells relative to untreated cells. To do so, we need to provide ATAC-seq peaks from stimulated cells as the input set and ATAC-seq peaks from untreated cells as the control set in the `enrichmotifpairR` package.

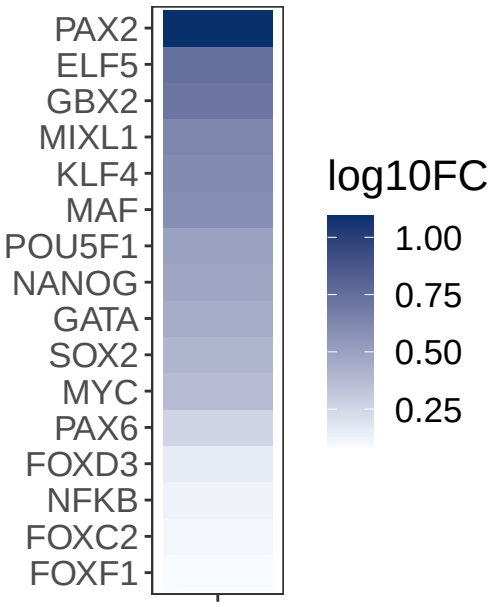
```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPair(
  target_data = example_peaks_data$Th17_stimulated_ATAC_seq_peaks,
  background_data = example_peaks_data$Th17_ATAC_seq_peaks,
  genome_ver = "hg19",
  scramble_data = FALSE,
  motif_database = "ENCODE",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.05,
  Pvalue_adjust_method = "BH"
)
```

Assign the resulting data to individual objects and filter motifs for only TFs genes. TF genes are defined by Lambert et al., 2018.

```
# assign the results data to individual objects
enrich_motifs <- results$motif_enrich
enrich_motif_pairs <- results$motif_pair_enrich
```

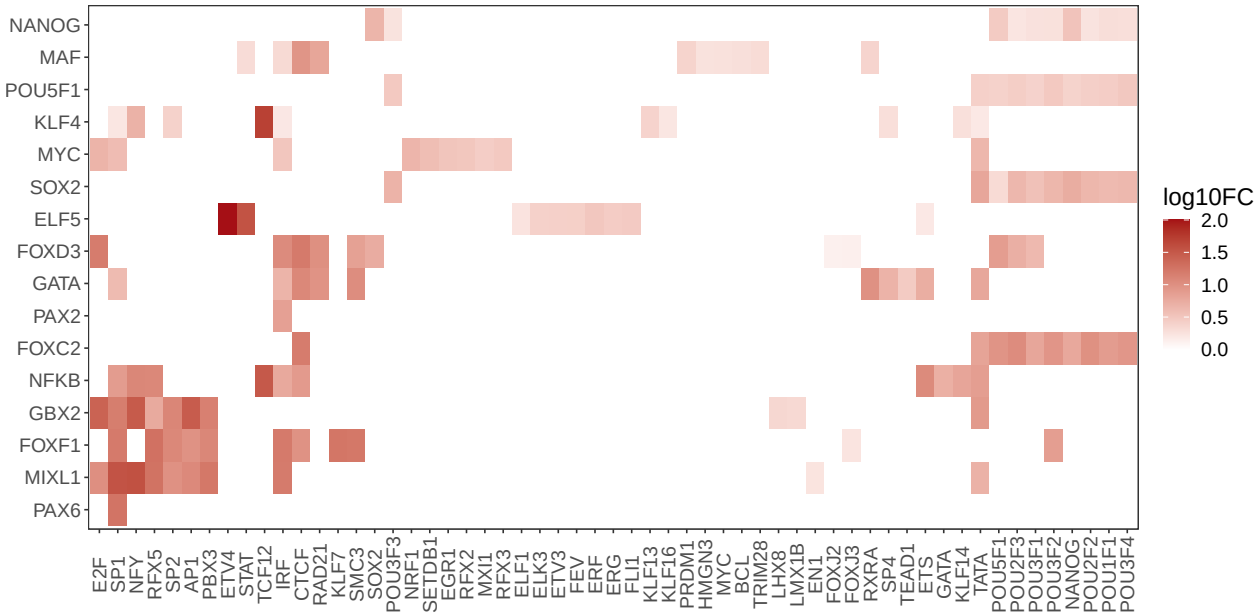
Plot the heatmap of enriched TF motifs

```
# plot enrichment for list of TF genes

plotEnrichment(enrich_motifs = enrich_motifs,
               tfs = TF_list)
```



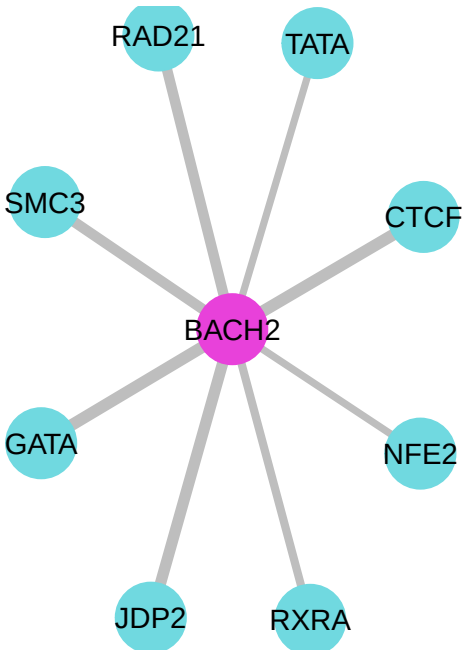
Now plot the binding partners for the above TF motifs

```
plotEnrichPair(enrich_pairs = enrich_motif_pairs,
               tfs = TF_list)
```

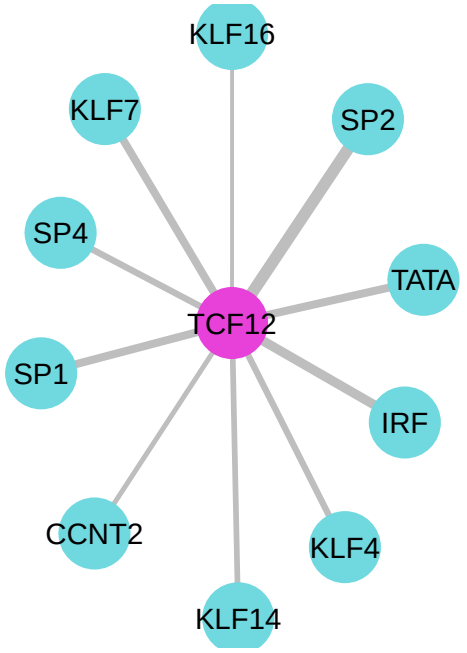


Now we can make colorful networks for select TFs of interest.

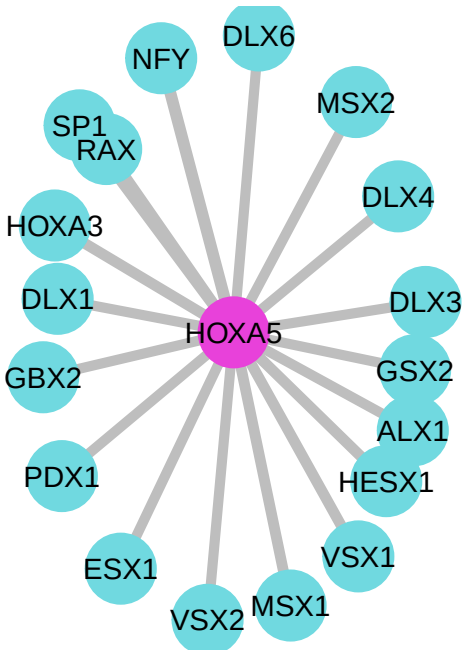
```
# select TF "BACH2" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "BACH2"
            )
```




```
# select TF "TCF12" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "TCF12"
            )
```




```
# select TF "HOXA5" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "HOXA5"
            )
```



Example use case 2: Genomic regions from two conditions

Here, we have genomic regions from two conditions, for instance, cells differentiating from one cell fate to another. To demonstrate this example, we obtained ATAC-seq data in Th0 cells (activated T cells) moving to Th1. We can find TFs and their binding partner TFs specifically enriched in Th1 cells relative to Th0 cells. To do so, we need to provide ATAC-seq peaks from Th1 cells as the input set and ATAC-seq peaks from Th0 cells as the control set in the `enrichmotifpairR` package. In this case, we are selecting motifs from the `CISBP` database.

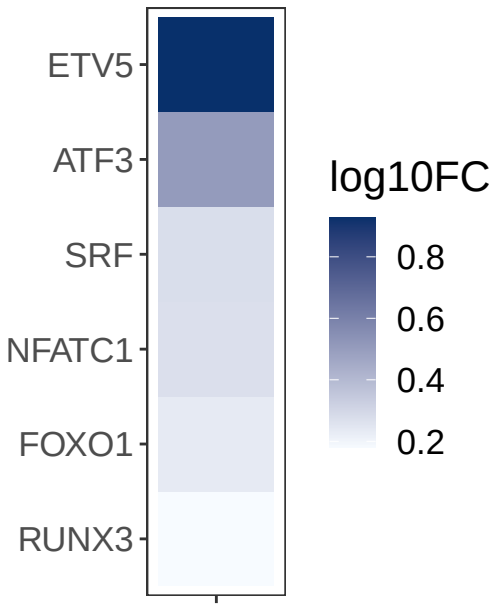
```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPair(
  target_data = example_peaks_data$Th1_ATAC_seq_peaks,
  background_data = example_peaks_data$Th0_ATAC_seq_peaks,
  genome_ver = "hg19",
  scramble_data = FALSE,
  motif_database = "CISBP",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
  Pvalue_adjust_method = "BH"
)
```

Assign the resulting data to individual objects and filter motifs for only TFs genes. TF genes are defined by Lambert et al., 2018.

```
# assign the results data to individual objects
enrich_motifs <- results$motif_enrich
enrich_motif_pairs <- results$motif_pair_enrich

TF_list <- unique(sort(c("STAT1", "STAT3", "STAT4", "STAT5A", "STAT5B", "ATF3",
                         "JUN", "JUNB", "FOXO1", "HAND1", "NFATC1", "NFATC2",
                         "NFATC3", "NFATC4", "SRF", "RUNX3", "IRF1", "IRF6",
                         "ETV5")))
```

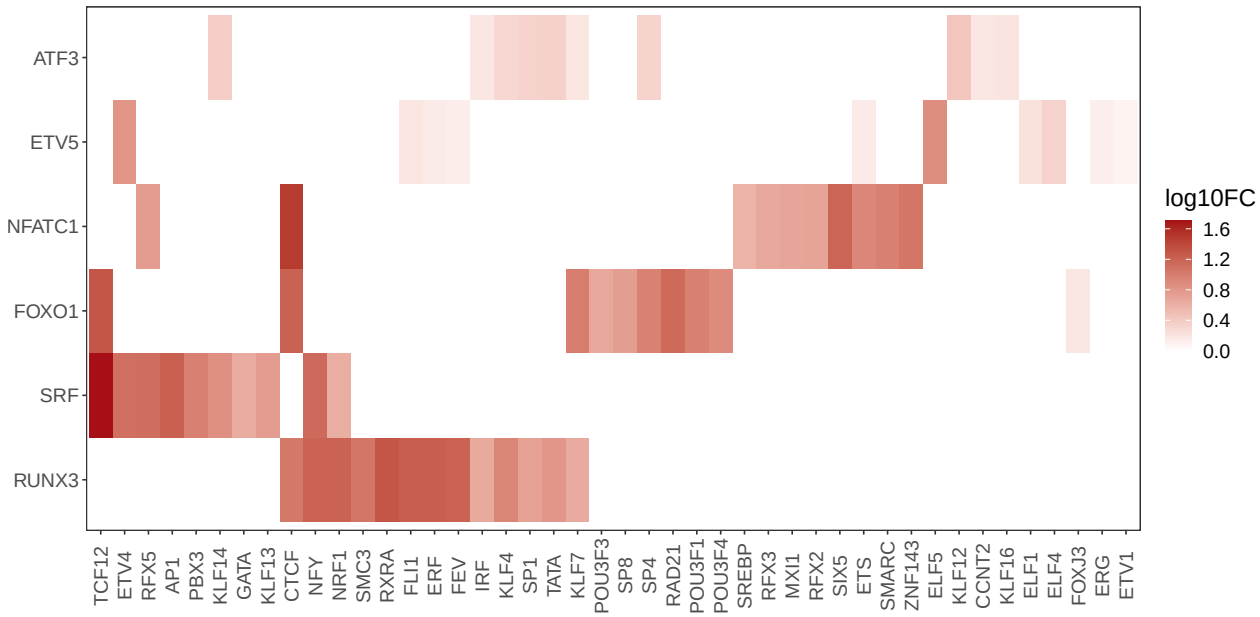
Plot the heatmap of enriched TF motifs

```
plotEnrichment(enrich_motifs = enrich_motifs,
               tfs = TF_list)
```



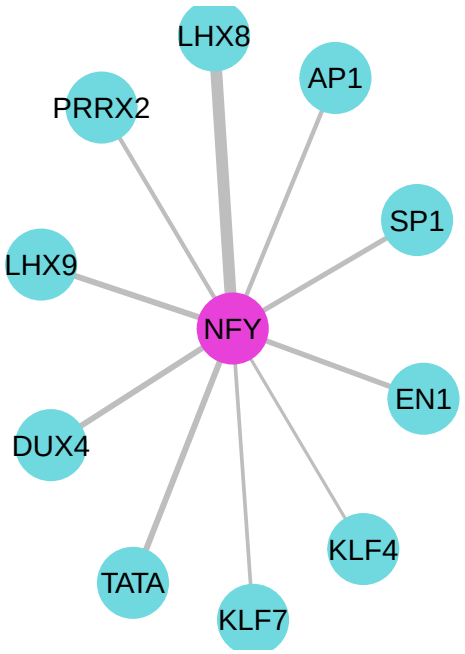
Now plot the binding partners for the above TF motifs.

```
plotEnrichPair(enrich_pairs = enrich_motif_pairs,
               tfs = TF_list)
```

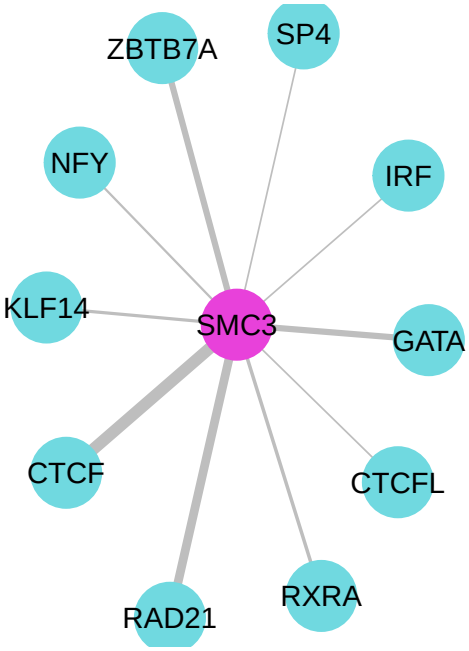


Now we can make colorful networks for select TFs of interest using the `igraph` package.

```
# select TF "NFY" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "NFY"
            )
```




```
# select TF "HAND1" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "SMC3"
            )
#> Warning in GGally::ggnet2(network, node.size = 12, node.color = my_color, :
#> ggnet2 does not know how to handle self-loops
```



Example use case 3: Genomic regions from two chromatin states

Here, we have genomic regions from two chromatin states, for instance, active genomic regions vs repressive genomic regions. To demonstrate this example, we obtained H3K27ac peaks representing active regions and

H3K27me3 peaks representing repressive genomic regions in CD8+ cells. We can find TFs and their binding partner TFs specifically enriched at active genomic regions relative to repressive regions. To do so, we need to provide H3K27ac peaks as the input set and H3K27me3 peaks as the control set in the `enrichmotifpairR` package. In this case, we are selecting motifs from the `JASPAR-CORE` database.

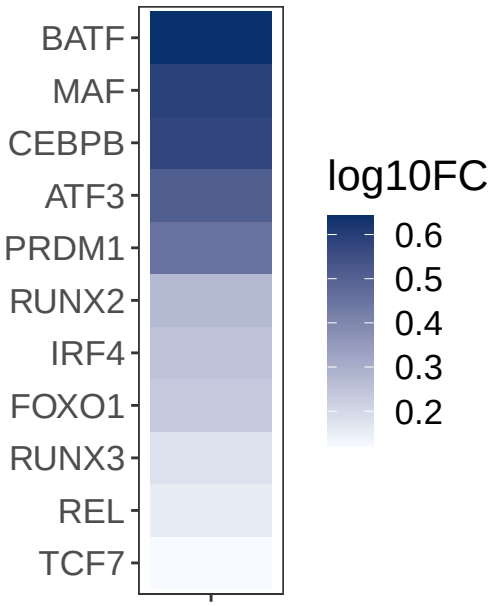
```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPair(
  target_data = example_peaks_data$`CD8+_H3K27ac_peaks`,
  background_data = example_peaks_data$`CD8+_H3K27me3_peaks`,
  genome_ver = "hg38",
  scramble_data = FALSE,
  motif_database = "JASPAR_CORE",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
  Pvalue_adjust_method = "BH"
)
```

Assign the resulting data to individual objects and filter motifs for only TFs genes. TF genes are defined by Lambert et al., 2018.

```
# assign the results data to individual objects
enrich_motifs <- results$motif_enrich
enrich_motif_pairs <- results$motif_pair_enrich

TF_list <- unique(sort(c("PRDM1", "TBX21", "TBX20", "EOMES", "BLIMP1", "ESRRB",
                         "ID2", "ID3", "CEBPB", "STAT3", "STAT5", "STAT6",
                         "ATF3", "FOXO1", "RUNX1", "RUNX2", "GATA1", "GATA2",
                         "GATA3", "GATA4", "BCL6", "BATF", "RORA", "RORC",
                         "IRF4", "REL", "TCF1", "TCF7", "IRF4", "TCF4", "EBF1",
                         "MAF", "ETS2", "RUNX3")))
```

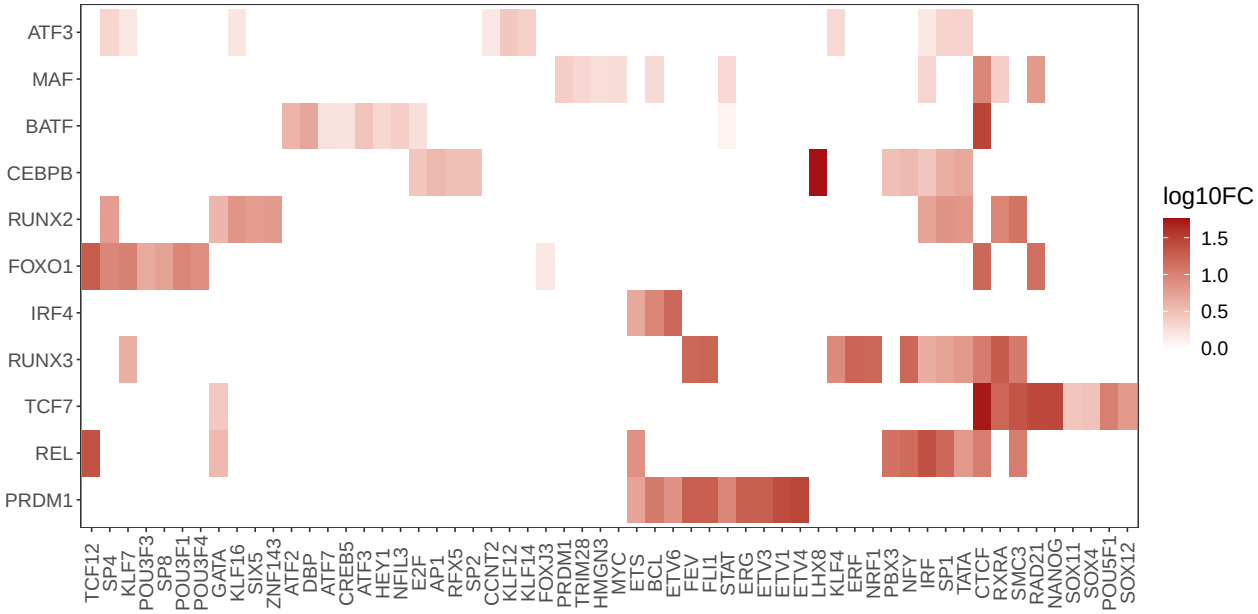
Plot the heatmap of enriched TF motifs

```
# replace very low values so that it is easy to visualize
enrich_motifs[enrich_motifs < 1e-300] <- 1e-300

plotEnrichment(enrich_motifs = enrich_motifs,
               tfs = TF_list)
```



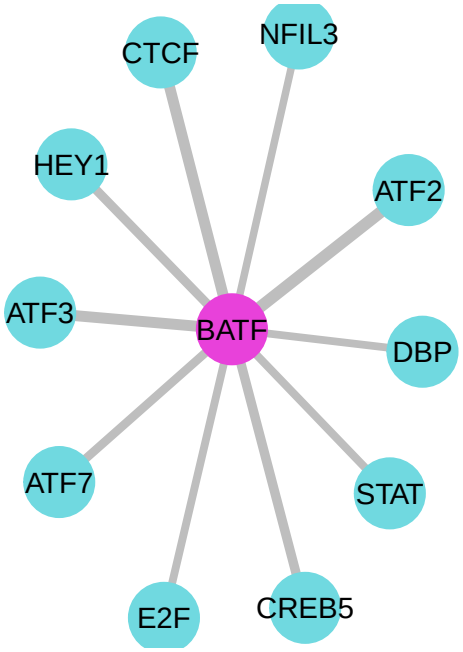
Now plot the binding partners for the above TF motifs

```
# replace very low values so that it is easy to visualize
enrich_motif_pairs[enrich_motif_pairs < 1e-100] <- 1e-100

plotEnrichPair(enrich_pairs = enrich_motif_pairs,
               tfs = TF_list)
```

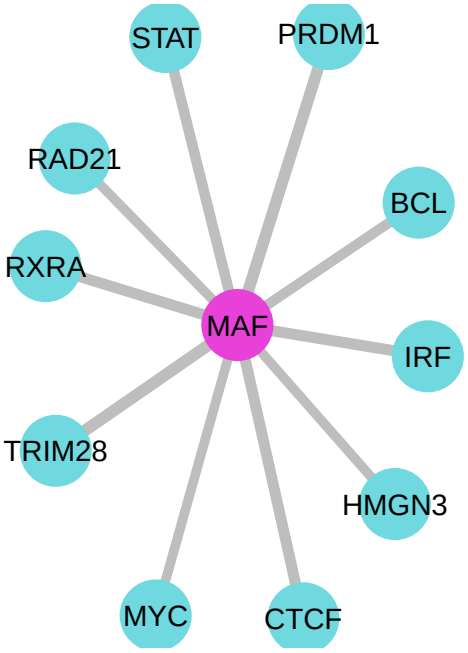


Now we can make colorful networks for select TFs of interest using the `igraph` package.

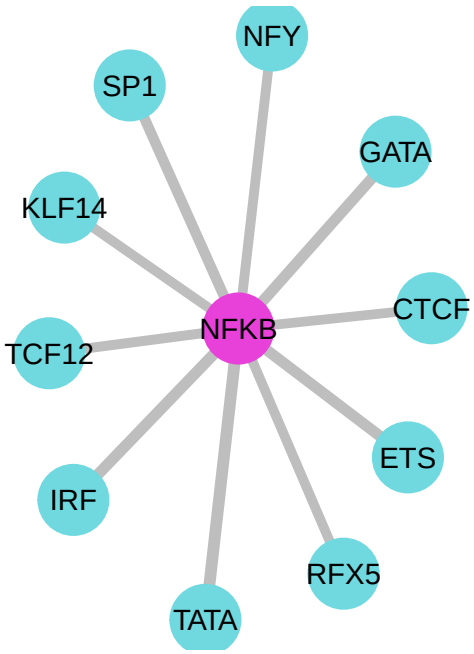
```
# select TF "BATF" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "BATF"
            )
```




```
# select TF "MAF" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "MAF"
            )
```




```
# select TF "NFKB" and extract its connections
plotNetwork(enrich_pairs = enrich_motif_pairs,
            TF_name = "NFKB"
            )
```



Special use cases:

Finding enriched TF motif pairs from Jolma et al.,2015.

In the paper Jolma et al.,2015, the authors report PWM for pairs of TFs and if you want to find enrichment of these TF pairs in the input peaks one can use the function `findEnrichMotifPair_Jolma2015`.

Here, we demonstrate this functionality by using genomic regions from two conditions, for instance, cells differentiating from one cell fate to another. We obtained ATAC-seq data in Th0 cells (activated T cells) moving to Th2. We can find enriched TF pairs in Th2 cells relative to Th0 cells. To do so, we need to provide ATAC-seq peaks from Th2 cells as the input set and ATAC-seq peaks from Th0 cells as the control set in the `enrichmotifpairR` package.

```
# Finding the enriched motif pairs
results <- findEnrichMotifPair_Jolma2015(
  target_data = example_peaks_data$Th2_ATAC_seq_peaks,
  background_data = example_peaks_data$Th0_ATAC_seq_peaks,
  genome_ver = "hg19",
  scramble_data = FALSE,
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
  Pvalue_adjust_method = "BH"
)
```

The enriched TF motif pairs are stored in the `results`.

```
enrich_motif_pairs <- results$motif_pair_enrich
# The output of the enriched motif pairs
enrich_motif_pairs[, 2:7] %>% head(10) %>%
    knitr::kable(., caption="Top 10 motif pairs")
```

Table 3: Top 10 motif pairs

TF_name_1	motif_name_2	TF_name_2	fold_enrich	pval	pval_adj
NFY	TATA_disc6	TATA	1.5677851	0.00e+00	0.00000003

TF_name_1	motif_name_2	TF_name_2	fold_enrich	pval	pval_adj
NFY	EN1_4	EN1	5.8978583	0.00e+00	0.00000017
NFY	SP1_disc1	SP1	1.4253158	0.00e+00	0.00000034
NFY	TATA_disc4	TATA	2.4936559	3.90e-07	0.00013382
NFY	TCF12_disc3	TCF12	7.3067911	4.80e-07	0.00014137
NFY	RFX5_disc2	RFX5	1.2707848	9.10e-07	0.00022560
NFY	ESX1_3	ESX1	4.9542010	1.15e-06	0.00023781
NFY	MIXL1_1	MIXL1	15.6293245	2.87e-06	0.00053861
NFY	LBX2_3	LBX2	4.0793520	6.60e-06	0.00113538
NFY	E2F_disc4	E2F	1.3377518	1.33e-05	0.00211175

```
# remove the duplicate motif pairs
enrich_motif_pairs_filtered <- enrich_motif_pairs %>%
    dplyr::group_by(TF_name_1, TF_name_2) %>%
    dplyr::distinct(TF_name_1, TF_name_2, .keep_all = TRUE)
enrich_motif_pairs_filtered[, 2:7] %>% head(10) %>%
    knitr::kable(., caption="Top 10 motif pairs after removing duplicates")
```

Table 4: Top 10 motif pairs after removing duplicates

TF_name_1	motif_name_2	TF_name_2	fold_enrich	pval	pval_adj
NFY	TATA_disc6	TATA	1.5677851	0.000e+00	0.00000003
NFY	EN1_4	EN1	5.8978583	0.000e+00	0.00000017
NFY	SP1_disc1	SP1	1.4253158	0.000e+00	0.00000034
NFY	TCF12_disc3	TCF12	7.3067911	4.800e-07	0.00014137
NFY	RFX5_disc2	RFX5	1.2707848	9.100e-07	0.00022560
NFY	ESX1_3	ESX1	4.9542010	1.150e-06	0.00023781
NFY	MIXL1_1	MIXL1	15.6293245	2.870e-06	0.00053861
NFY	LBX2_3	LBX2	4.0793520	6.600e-06	0.00113538
NFY	E2F_disc4	E2F	1.3377518	1.330e-05	0.00211175
NFY	SP4_2	SP4	2.2194572	1.709e-05	0.00251980

Directly finding all possible enriched TF motif pairs without first looking for individual enriched motifs.

If users are interested in finding all possible enriched TF motif pairs without first looking for individual enriched motifs, one can use the function `findEnrichMotifPairAll`.

Here, we demonstrate this functionality by using genomic regions from two conditions, for instance, cells differentiating from one cell fate to another. We obtained ATAC-seq data in Th0 cells (activated T cells) moving to Th1. We can find all enriched TFs pairs in Th1 cells relative to Th0 cells. To do so, we need to provide ATAC-seq peaks from Th1 cells as the input set and ATAC-seq peaks from Th0 cells as the control set in the `enrichmotifpairR` package. In this case, we are selecting motifs from `JASPAR-UNVALIDATED` database.

```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPairAll(
  target_data = example_peaks_data$Th1_ATAC_seq_peaks,
  background_data = example_peaks_data$Th0_ATAC_seq_peaks,
  genome_ver = "hg19",
  scramble_data = FALSE,
  motif_database = "JASPAR_UNVALIDATED",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
```

```
  Pvalue_adjust_method = "BH"
)
```

The enriched TF motif pairs are stored in the `results`. These pairs contain duplicates (redundant entries) and should be removed. To do so one can make use of the function `removeDuplicateTFPairs`. Also filter for only TF genes. TF genes are defined by Lambert et al., 2018.

```
# assign the results data to individual objects
enrich_motif_pairs <- results$motif_pair_enrich
# remove duplicate pairs
removeDuplicateTFPairs <- function(data = data){
    data <- data %>% dplyr::arrange(pval_adj) %>%
        dplyr::mutate(key = paste0(pmin(TF_name_1, TF_name_2),
                                   pmax(TF_name_1, TF_name_2), sep = "")) %>%
        dplyr::distinct(key, .keep_all = TRUE) %>%
        dplyr::filter(TF_name_1 != TF_name_2) %>%
        dplyr::select(-key)
    return(data)
}
enrich_motif_pairs <- removeDuplicateTFPairs(data = enrich_motif_pairs)
# get the list of TFs genes
TF_df <- TF_df %>% dplyr::filter(TF_Yes_No == "Yes")
# filter TFs motifs for only TFs genes based on Lambert et al., 2018
enrich_motif_pairs_filtered <- enrich_motif_pairs %>%
    dplyr::distinct(TF_name_1, TF_name_2, .keep_all = TRUE) %>%
    dplyr::filter(TF_name_1 %in% TF_df$Name) %>%
    dplyr::filter(TF_name_2 %in% TF_df$Name) %>%
    dplyr::filter(pval_adj < 1e-05)
```

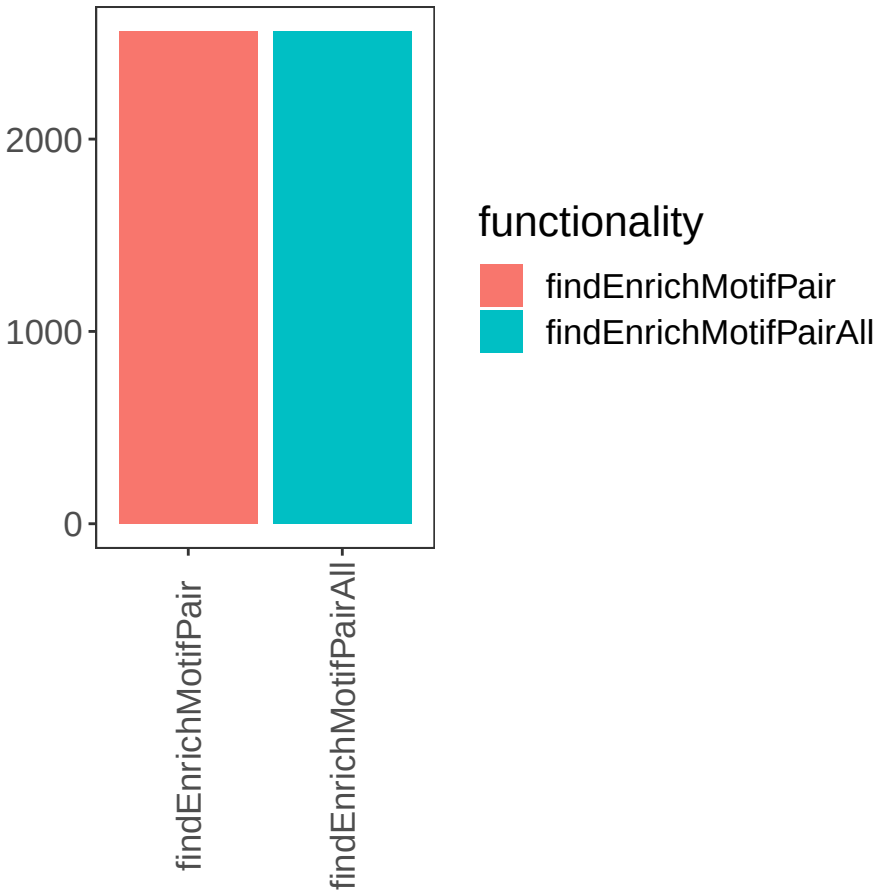
Compare these TF motif pairs with the motif pairs from the functionality of `findEnrichMotifPair`.

```
# Finding the enriched motifs and their partners
results <- findEnrichMotifPair(
  target_data = example_peaks_data$Th1_ATAC_seq_peaks,
  background_data = example_peaks_data$Th0_ATAC_seq_peaks,
  genome_ver = "hg19",
  scramble_data = FALSE,
  motif_database = "JASPAR_UNVALIDATED",
  Pvalue_computation = "hyper",
  Pvalue_threshold = 0.01,
  Pvalue_adjust_method = "BH"
)

# assign the results data to individual objects
enrich_motif_pairs_2 <- results$motif_pair_enrich
# remove duplicate pairs
removeDuplicateTFPairs <- function(data = data){
    data <- data %>% dplyr::arrange(pval_adj) %>%
        dplyr::mutate(key = paste0(pmin(TF_name_1, TF_name_2),
                                   pmax(TF_name_1, TF_name_2), sep = "")) %>%
        dplyr::distinct(key, .keep_all = TRUE) %>%
        dplyr::filter(TF_name_1 != TF_name_2) %>%
        dplyr::select(-key)
    return(data)
}
enrich_motif_pairs_2 <- removeDuplicateTFPairs(data = enrich_motif_pairs_2)
```

```
# get the list of TFs genes
TF_df <- TF_df %>% dplyr::filter(TF_Yes_No == "Yes")
# filter TFs motifs for only TFs genes based on Lambert et al., 2018
enrich_motif_pairs_filtered_2 <- enrich_motif_pairs_2 %>%
    dplyr::distinct(TF_name_1, TF_name_2, .keep_all = TRUE) %>%
    dplyr::filter(TF_name_1 %in% TF_df$Name) %>%
    dplyr::filter(TF_name_2 %in% TF_df$Name) %>%
    dplyr::filter(pval_adj < 1e-05)
```

Comparison plot

```
# create data frame to plot the number of TF pairs comparison plot
data <- data.frame(functionality = c("findEnrichMotifPair", "findEnrichMotifPairAll"), value = c(nrow(en

# bar plot
pp <- ggplot(data = data,
             aes(x = functionality, y = value, fill = functionality)) +
    geom_bar(stat = "identity") +
    ylab("") + xlab("") + theme_bw() +
    theme(plot.background = element_blank()
    ,panel.grid.major = element_blank()
    ,panel.grid.minor = element_blank()
    ,text = element_text(size=16), axis.text.x = element_text(angle=90, vjust=0.5)
)
pp
```



As you can see, `findEnrichMotifPair` determines relatively fewer enriched motif pairs than

`findEnrichMotifPairAll`, as the former finds only the pairs associated with individually enriched motifs.