rTRM: an R package for the identification of transcription regulatory modules (TRMs)

Diego Diez

February 14, 2014

1 Introduction

Transcription factors (TFs) bind to short motifs in the DNA and regulate the expression of target genes in a cell type and time dependent fashion. TFs do so by cooperating with other TFs in what it is called Transcriptional Regulatory Modules (TRMs). These TRMs contain different TFs and form a combinatorial code that explains TF specificity. We have implemented a method for the identification of TRMs that integrates information about binding locations from a single ChIP-seq experiment, computational estimation of TF binding, gene expression and protein-protein interaction (PPI) data (Diez et al. submitted; see workflow figure). rTRM implements the methods required for the integration of PPI information (step 4 in workflow). To do so, rTRM tries to identify TFs that are bound to a target TF (the one with experimental evidence- i.e. ChIP-seq data) either directly or through a bridge protein. This package has been used to identify cell-type independent and dependent TRMs associated with Stat3 functions [1]. Also, it has been used to identify TRMs in embryonic and hematopoietic stem cells as part of the publication presenting the methodology (Diez et al. submitted). Here we present the basic capabilities of rTRM with a naive example, a case study showing the identification of Sox2 related TRM in ESCs as performed in the paper describing rTRM (Diez et al. submitted), and a complete workflow in R using the PWMEnrich package for the motif enrichment step.

2 Minimal example

In this minimal example a fake network is search to identify TRMs focused around a target node, *N6*, with query nodes being *N7*, *N12* and *N28*. By default *findTRM* find nodes separated a max distance of 0 (i.e. nodes directly connected). We change this with parameter *max.bridge = 1*. Because node *N28* is separated by two other nodes from the target node *N6*, it is not included in the predicted TRM. By default findTRM() returns an object of class igraph, which can be used with plotTRM(), plotTRMlegend() and other rTRM functions. However it is possible to directly obtain a graphNEL object (from the Bioconductor package graph), setting the type argument to "graphNEL". Of course it is possible to also use the igraph.to.graphNEL() function in the igraph package to transform an igraph object into a graphNEL object.

```
> # load the rTRM package
> library(rTRM)
>
> # load network example.
> load(system.file(package = "rTRM", "extra/example.rda"))
>
> # plot network
```

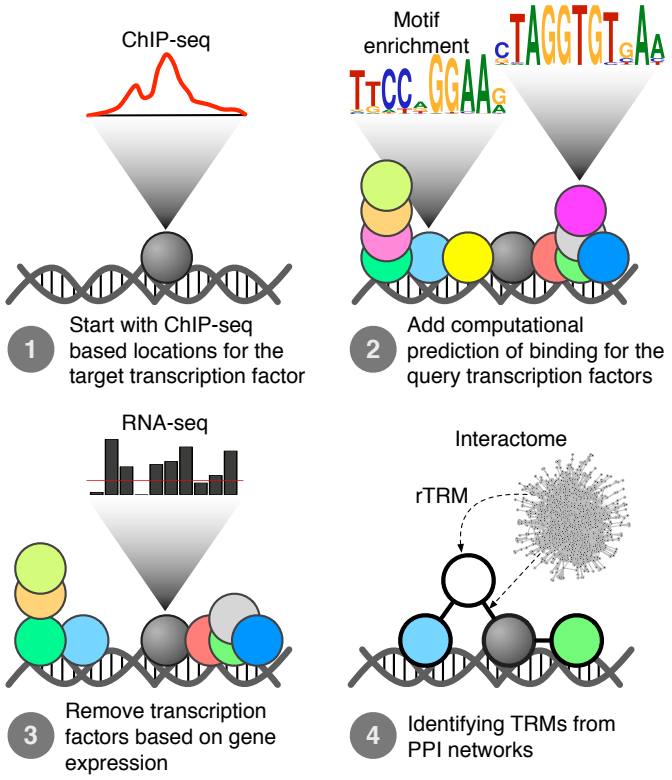

Figure 1: Workflow for the identification of TRMs. Steps 1-3 can be performed with standard Bioconductor approaches. rTRM implements a method to perform step 4.

```
> plotGraph(g, vertex.cex = 5)
>
> # define target and query nodes:
> target <- "N6"
> query <- c("N7", "N12", "N28")
>
> # find TRM:
> s <- findTRM(g, target = target, query = query, method = "nsa", max.bridge = 1)

removing:  4 nodes out of 8 [keeping 4 nodes]

>
> # annotate nodes:
> V(s)$color <- "white"
> V(s)[name %in% query]$color <- "steelblue2"
> V(s)[name %in% target]$color <- "steelblue4"
>
> # plot:
> plotGraph(s, mar = 5, vertex.cex = 5)
```

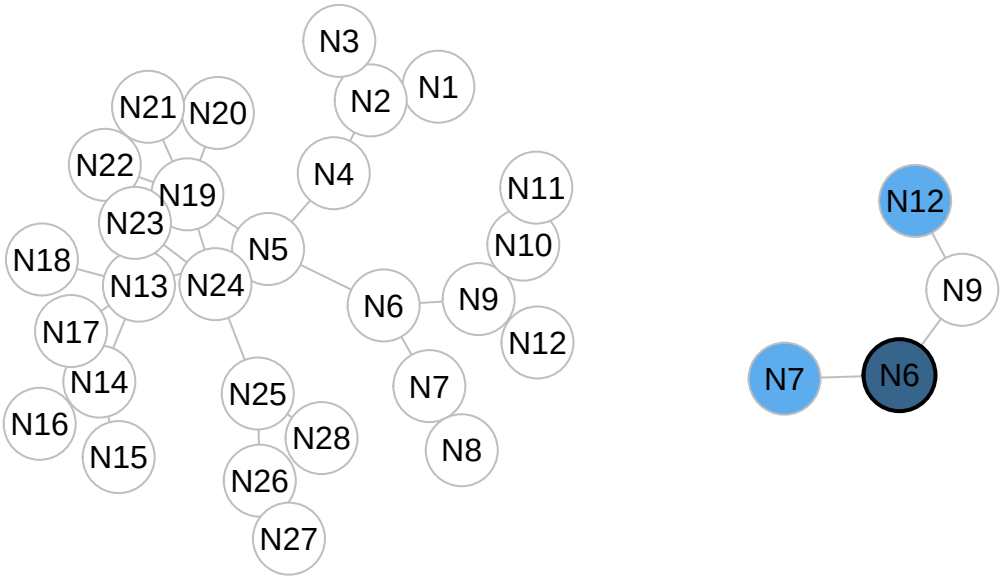

Figure 2: Identification of a TRM from a test network (left). In the resulting TRM (right) dark blue indicates the target node light blue are query nodes and white nodes are bridge nodes

3 Introduction to the rTRM package

rTRM relies on a series of optimizations. For example in the publication we used PWMs for vertebrate species compiled from different sources. This assumes the binding specificities of TFs will be conserved on all these species. Recent comparison between mouse and human PWMs suggests that this is the case for most TFs [9]. rTRM also relies on protein-protein interaction data, and so provides utilities to download data from the BioGRID database (see below). As some of these functionalities are further integrated with existing Bioconductor functionality they may be defunct in the future.

3.1 Database

Information about TFs, including Position Specific Weight (PWMs) matrices, mapping to Entrez Gene identifiers, orthologs in mouse and human and other annotations are stored as a SQLite database. rTRM provides a basic API for accessing the data. Below there are some examples.

To obtain PWMs:

```
> pwm <- getMatrices()
> head(pwm, 1)
$MA0009.1
     1     2 3 4 5 6 7    8     9 10    11
A 0.05 0.025 1 0 0 0 0 0.00 0.025  1 0.775
C 0.70 0.025 0 0 0 0 0 0.05 0.175  0 0.125
G 0.20 0.000 0 1 1 0 1 0.00 0.700  0 0.000
T 0.05 0.950 0 0 0 1 0 0.95 0.100  0 0.100
```

To get annotations:

```
> ann <- getAnnotations()
> head(ann)
```

```
  row_names   pwm_id   symbol                    family domain binding source
1         1 MA0009.1        T                         T        monomer jaspar
2         2 MA0059.1 MYC::MAX          Helix-Loop-Helix          dimer jaspar
3         3 MA0146.1      Zfx BetaBetaAlpha-zinc finger        monomer jaspar
4         4 MA0132.1     Pdx1                      Homeo        monomer jaspar
5         5 MA0162.1     Egr1 BetaBetaAlpha-zinc finger        monomer jaspar
6         6 MA0093.1     USF1          Helix-Loop-Helix        monomer jaspar
  note
1 2010
2 2010
3 2010
4 2010
5 2010
6 2010
```

To get map of TFs to genes:

```
> map <- getMaps()
> head(map)
  row_names   pwm_id entrezgene organism
1         1 MA0009.1      20997    mouse
2         2 MA0059.1       4609    human
3         3 MA0059.1       4149    human
4         4 MA0146.1      22764    mouse
5         5 MA0132.1      18609    mouse
6         6 MA0162.1      13653    mouse
```

To get map of TFs to ortholog genes:

```
> o <- getOrthologs(organism = "mouse")
> head(o)
  row_names entrezgene organism map_entrezgene map_organism
1       775      20997    mouse          20997        mouse
2       776       4609    human         107771        mouse
3       777       4149    human          17187        mouse
4       778      22764    mouse          22764        mouse
5       779      18609    mouse          18609        mouse
6       780      13653    mouse          13653        mouse
```

It is possible to map motif ids to entrezgene ids in the target organism (only between human and mouse). This is useful when all the information about existing PWMs is desired, as some TF binding affinities have only been studied in one organism.

```
> getOrthologFromMatrix("MA0009.1", organism = "human")
[1] "6862"
> getOrthologFromMatrix("MA0009.1", organism = "mouse")
[1] "20997"
```

3.2 Interactome data

rTRM requires information about protein-protein interactions (PPIs) for its predictions and includes interactome (PPI network) data from the BioGRID database [2]. Currently mouse and human interactomes are supported. The networks are provided as *igraph* format. To access the data use:

```
> # check statistics about the network.
> biogrid_mm()

Mouse PPI network data of class igraph
Number of nodes:  4536
Number of edges:  9590
Source:  The BioGRID (http://www.thebiogrid.org)
Release:  3.2.105
Downloaded:  2013-10-01
Use data(biogrid_mm) to load it

> # load mouse PPI network:
> data(biogrid_mm)
```

The amount of available PPI data increases rapidly so it is desirable to have a way to access the newest data conveniently. rTRM includes support for direct download and processing of PPI data from the BioGRID database. The PPI network is stored as an *igraph* object that can be readily used with rTRM or stored for later use. Below there is an example of the BioGRID database update procedure.

```
> # obtain dataset (currently need to state the version number explicitly, see
> # http://www.thebiogrid.org to identify the latest version:
> db <- getBiogridData("3.2.96")
>
> # process PPI data for different organisms (currently supported human and mouse):
> biogrid_hs <- processBiogrid(db, org = "human")
> biogrid_mm <- processBiogrid(db, org = "mouse")
```

PPI data from other databases could be used as long as it is formatted as an *igraph* object with the 'name' attribute containing entrezgene identifiers and the 'label' attribute containing the symbol.

3.2.1 Using PSICQUIC package to obtain protein-protein interactions

One possibility available from Bioconductor 2.13 is to use the package PSIQUIC to obtain PPI data. PSICQUIC provides access to different databases of PPIs, including BioGRID and STRINGS, and databases of cellular networks like KEGG or Reactome. For exaple, to obtain the human BioGRID data (note is named BioGrid at PSICQUIC):

```
> library(PSICQUIC)
> psicquic <- PSICQUIC()
> providers(psicquic)
>
> # obtain BioGrid human PPIs (as data.frame):
> tbl <- interactions(psicquic, species = "9606", provider = "BioGrid")
>
> # the target and source node information needs to be polished (i.e. must be
> # Entrez gene id only)
> biogrid_hs <- data.frame(source = tbl$A, target = tbl$B)
> biogrid_hs$source <- sub(".*locuslink:(.*)\\|BIOGRID:.*", "\\1", biogrid_hs$source)
> biogrid_hs$target <- sub(".*locuslink:(.*)\\|BIOGRID:.*", "\\1", biogrid_hs$target)
>
> # create graph.
> library(igraph)
```

```
> biogrid_hs <- graph.data.frame(biogrid_hs, directed = FALSE)
> biogrid_hs <- simplify(biogrid_hs)
>
> # annotate with symbols.
> library(org.Hs.eg.db)
> V(biogrid_hs)$label <- select(org.Hs.eg.db, keys = V(biogrid_hs)$name, columns = c("SYMBOL"))$SYMBO
```

Facilities will be added to automatically process the data from PSICQUIC to produce igraph objects suitable for rTRM.

4 Case study: TRM associated with Sox2 in embryonic stem cells (ESCs)

Sox2 is a TF involved in the determination and maintainance of pluripotency in embryonic stem cells (ESCs). It is known that Sox2 works together with Nanog and Pou5f1 (Oct4). Other TFs important are Erssb, and Klf4. In this case study we want to identify TRMs associated with Sox2. ChIP-seq data for Sox2 was obtained from Chen et al. (Cell 2008) [3] and motif enrichment analysis performed with HOMER [4], followed by matching against our library of PWMs using TOMTOM [5]. The starting dataset is the TOMTOM output file with the motifs enriched in the Sox2 binding regions.

```
> # library(org.Mm.eg.db)
>
> # read motif enrichment results.
> motif_file <- system.file("extra/sox2_motif_list.rda", package = "rTRM")
> load(motif_file)
> length(motif_list)
[1] 177
> head(motif_list)
[1] "MA0039.2" "MA0071.1" "MA0075.1" "MA0077.1" "MA0078.1" "MA0112.2"
```

First, we read the motifs and convert them into gene identifiers (i.e. Entrez Gene identifier). To do this we use the function **getOrthologFromMatrix**, which takes a list of motif identifiers and the target organism as parameters. The function returns a list with the Entrez Gene ids.

```
> # get the corresponding gene.
> tfs_list <- getOrthologFromMatrix(motif_list, organism = "mouse")
> tfs_list <- unique(unlist(tfs_list, use.names = FALSE))
> length(tfs_list)
[1] 98
> head(tfs_list)
[1] "18609" "20682" "13983" "20665" "18291" "18227"
```

Next, we need a list of genes expressed in ESC. For this, the dataset was obtained from GEO (GSE27708; [6]) and processed using the custom CDFs from the BrainArray project [7] and the **rma** function from the package *affy* [8]. Genes expressed were determined by removing all genes with log2 expression ¡ 5 in all samples.

```
> # load expression data.
> eg_esc_file <- system.file("extra/ESC-expressed.txt", package = "rTRM")
> eg_esc <- scan(eg_esc_file, what = "")
> length(eg_esc)
```

```
[1] 8734
> head(eg_esc)
[1] "100008567" "100017"    "100019"    "100037258" "100038489" "100039781"
>
> tfs_list_esc <- tfs_list[tfs_list %in% eg_esc]
> length(tfs_list_esc)
[1] 22
> head(tfs_list_esc)
[1] "26380" "18999" "20674" "16600" "26379" "13984"
```

Next, we load the PPI network and filter out potential degree outliers and proteins not expressed in the paired expression data.

```
> # load and process PPI data.
> biogrid_mm()

Mouse PPI network data of class igraph
Number of nodes:  4536
Number of edges:  9590
Source:  The BioGRID (http://www.thebiogrid.org)
Release:  3.2.105
Downloaded:  2013-10-01
Use data(biogrid_mm) to load it

> data(biogrid_mm)
> ppi <- biogrid_mm
> vcount(ppi)
[1] 4536
> ecount(ppi)
[1] 9590
>
> # remove outliers.
> f <- c("Ubc", "Sumo1", "Sumo2", "Sumo3")
> f <- unlist(mget(f, revmap(org.Mm.egSYMBOL)))
> ppi <- removeVertices(ppi, f)
> vcount(ppi)
[1] 4205
> ecount(ppi)
[1] 9065
>
> # filter by expression.
> ppi_esc <- induced.subgraph(ppi, V(ppi)[name %in% eg_esc])
> vcount(ppi_esc)
[1] 2541
> ecount(ppi_esc)
[1] 3668
>
> # ensure a single component.
> ppi_esc <- getLargestComp(ppi_esc)
> vcount(ppi_esc)
[1] 1904
> ecount(ppi_esc)
[1] 3610
```

To identify TRMs we define a target TF (the one the ChIP-seq data comes from) and some query TFs (the ones with enriched binding sites in the neighborhood of the target TF).

```
> # define target.
> target <- unlist(mget("Sox2", org.Mm.egSYMBOL2EG))
> target
   Sox2
"20674"
>
> # find TRM.
> s <- findTRM(ppi_esc, target, tfs_list_esc, method = "nsa", max.bridge = 1)

11 genes NOT FOUND in network-- removed from query
removing:  319 nodes out of 327 [keeping 8 nodes]

> vcount(s)
[1] 8
> ecount(s)
[1] 15
```

Finally, we layout the network using a customized concentric layout and plot the network and the legend.

```
> # generate layout (order by cluster, then label)
> cl <- getConcentricList(s, target, tfs_list_esc)
> l <- layout.concentric(s, cl, order = "label")
>
> # plot TRM.
> plotTRM(s, layout = l, vertex.cex = 15, label.cex = 0.8)
> plotTRMlegend(s, title = "ESC Sox2 TRM", cex = 0.8)
```

5 A complete workflow in R

In this section we will identify Sox2 TRMs using a workflow performed completely in R. For this the MotifDb package will be used to obtain the information about PWMs, and PWMEnrich package for identifying enriched motifs. PWMEnrich requires the computation of background models and the enrichment analysis *per se*, which are computational intensive. Therefore these steps were not run during the compilation of this vignette.

The first step is to retrieve a set of PWMs. Here we will use the MotifDb package available in Bioconductor. We will use only mouse PWMs (i.e. PWMs for the target organism). It could be possible to use matrices from other species but then the user has to obtain the orthologs in the target organism (e.g. using getOrthologsFromBiomart() in rTRM or using the Biomart package directly).

```
> library(rTRM)
> library(BSgenome.Mmusculus.UCSC.mm8)  # Sox2 peaks found against mm8

Loading required package:  BSgenome
Loading required package:  IRanges
   Attaching package:  'IRanges'
   The following objects are masked from 'package:igraph':
       compare, simplify
```

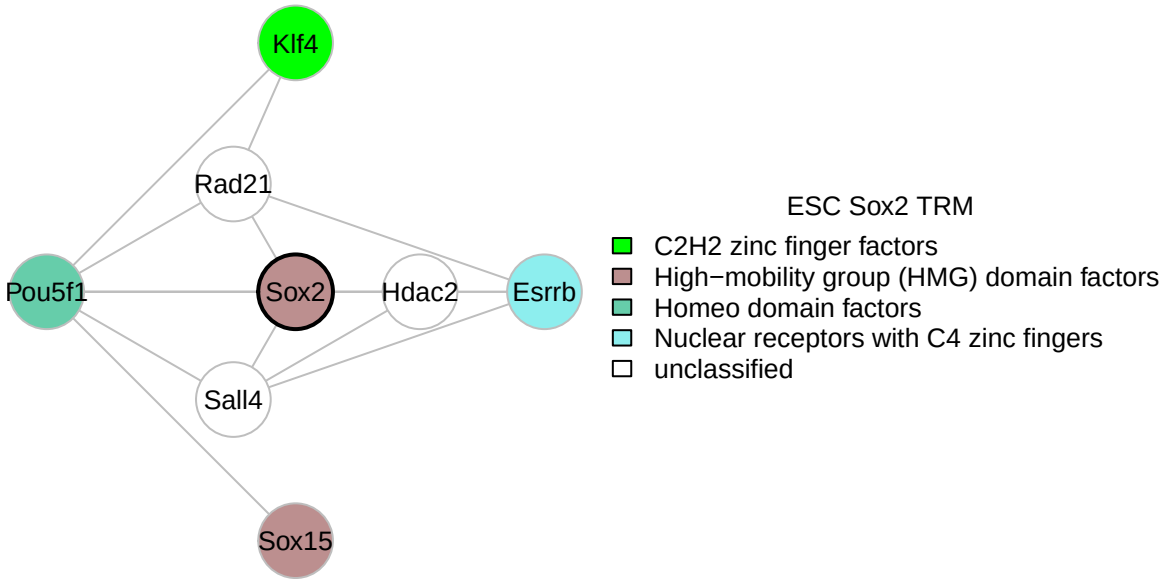

Figure 3: Sox2 specific TRM in ESCs.

```
   Loading required package:  GenomicRanges
Loading required package:  XVector
Loading required package:  Biostrings
   Attaching package:  'BSgenome'
   The following object is masked from 'package:annotate':
       organism
   The following object is masked from 'package:AnnotationDbi':
       species


Loading required package:  grid



See system.file("LICENSE", package="MotifDb") for use restrictions.
```

The matrices need to be passed as counts, that is the PFM need to be converted to counts. The easiest way is to multiply by 100 and round the results. We also need to convert it to integer.

```
> # generate logn background model of PWMs:
> p <- as.list(pwm.mm)
> p <- lapply(p, function(x) round(x * 100))
> p <- lapply(p, function(x) t(apply(x, 1, as.integer)))
```

With the PFMs we compute the background model using the makeBackground() function from the

PWMEnrich package, which returns the corresponding PWMs. This requires a list with the PFMs as counts, the organisms to obtain the sequences to compute the background and the type of background model (here "logn" model is used).

```
> pwm_logn <- makeBackground(p, Mmusculus, type = "logn")
```

Next we read the peak information from the Sox2 Chip-seq data. This is the original coordinates obtained from Chen et al. (Cell 2008), which were obtained for Mmusculus mm8 genome. The function getSequencesFromGenome() is an utility wrapper to getSeq() that facilitates appending a label to the sequences' ids. PWMEnrich requires sequences the same size or longer to the motifs so we check what is the largest motif and filter the sequences accordingly.

```
> sox2_bed <- read.table(system.file("extra/ESC_Sox2_peaks.txt", package = "rTRM"))
>
> colnames(sox2_bed) <- c("chr", "start", "end")
>
> sox2_seq <- getSequencesFromGenome(sox2_bed, Mmusculus, append.id = "Sox2")
> # PWMEnrich throws an error if the sequences are shorter than the motifs so we
> # filter those sequences.
> min.width <- max(sapply(p, ncol))
> sox2_seq_filter <- sox2_seq[width(sox2_seq) >= min.width]
```

Next, enrichment is computed with the sequences and the PWMs with the background model as parameters.

```
> # find enrichment:
> sox2_enr <- motifEnrichment(sox2_seq_filter, pwms = pwm_logn, group.only = TRUE)
```

Next, retrieve the enriched motifs and obtain the Entrezgene ids. Then proceed with the same steps as in the Use Case example shown in the previous section. The resulting TRM looks very similar to the one in the Use Case. The main difference is the absence of Klf4 and the presence of Tcf7l1. A possibility could be that different matrices were used during the enrichment analysis since our approach uses compiled vertebrate PWMs that are mapped to the target species (Diez et al. NAR 2013). However in this case there is a mouse PWM for Klf4 in MotifDb. The reason for this discrepacy is in the different approaches used for motif enrichment. Whereas HOMER, which was used to obtain the list of enriched motifs in the Use Case, uses random sets of sequences with similar composition to the ChIP-seq peaks provided to generate the background, PWMEnrich uses 2000 bp upstream of the promoter of all genes in the genome. Generally, using different strategies for enrichment will tend to produce slightly different TRMs.

```
> res <- motifRankingForGroup(sox2_enr)
> res.gene <- unique(values(MotifDb[names(res[res < 0.05])])$geneId)
> res.gene <- res.gene[!is.na(res.gene)]
>
> data(biogrid_mm)
> ppi <- biogrid_mm
> vcount(ppi)
[1] 4536
> ecount(ppi)
[1] 9590
>
> f <- c("Ubc", "Sumo1", "Sumo2", "Sumo3")
```

```
> f <- unlist(mget(f, revmap(org.Mm.egSYMBOL)))
> ppi <- removeVertices(ppi, f)
> vcount(ppi)
[1] 4205
> ecount(ppi)
[1] 9065
>
> # filter by expression.
> eg_esc <- scan(system.file("extra/ESC-expressed.txt", package = "rTRM"), what = "")
> ppi_esc <- induced.subgraph(ppi, V(ppi)[name %in% eg_esc])
> vcount(ppi_esc)
[1] 2541
> ecount(ppi_esc)
[1] 3668
>
> # ensure a single component.
> ppi_esc <- getLargestComp(ppi_esc)
> vcount(ppi_esc)
[1] 1904
> ecount(ppi_esc)
[1] 3610
>
> sox2.gene <- unlist(mget("Sox2", org.Mm.egSYMBOL2EG))
> sox2_trm <- findTRM(ppi_esc, target = sox2.gene, query = res.gene)

40 genes NOT FOUND in network-- removed from query
removing:  347 nodes out of 358 [keeping 11 nodes]

>
> cl <- getConcentricList(sox2_trm, t = sox2.gene, e = res.gene)
> l <- layout.concentric(sox2_trm, concentric = cl, order = "label")
> plotTRM(sox2_trm, layout = l, vertex.cex = 15, label.cex = 0.8)
> plotTRMlegend(sox2_trm, title = "ESC Sox2 TRM", cex = 0.8)
```

6 Ploting parameters

The most important parameter determining the appearance of your network will be the layout. When networks contain many nodes and edges are difficult to interpret. rTRM implements two igraph layouts that try improve the visualization and interpretation of the identified TRMs. The layout *layout.concentric* is a circular layout with multiple concentric layers that places the target TFs in the center, the enriched (or query) TFs in the outer circle and the bridge TFs in the middle circle. Another layut is *layout.arc* that tries to mimic the layout presented in the rTRM description (Fig. 1). In this case all nodes are plotted in a liner layout, with the targets in the center, and the enriched (query) nodes at each side. Those enriched nodes connected directly to any of the target nodes are placed in the left side. Those connected through a bridge node are placed in the right side, with the bridge node placed in between. The following figure compares the concentric layout obtained in the previous section with a layout using the *layout.arc* function.

```
> plotTRM(sox2_trm, layout = l, vertex.cex = 15, label.cex = 0.7)
> l <- layout.arc(sox2_trm, target = sox2.gene, query = res.gene)
```

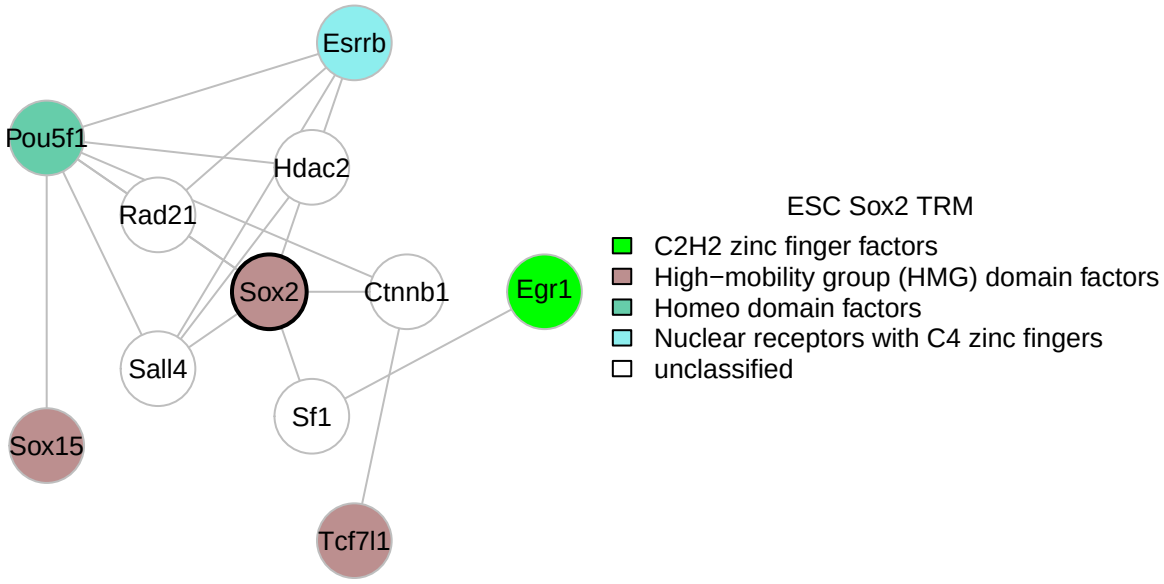

Figure 4: Sox2 TRM identified using PWMEnrich for the motif enrichment. In this TRM Klf4 is missing but Tcf7l1 is added to the list of enriched motifs

```
> plotTRM(sox2_trm, layout = l, vertex.cex = 15, label.cex = 0.7)
```

7 Citation

If you use *rTRM* in your research please include the following reference:

```
> citation(package = "rTRM")

To cite rTRM in publications use:

  Diego Diez, Andrew P. Hutchins, Diego Miranda-Saavedra. Systematic
  identification of transcriptional regulatory modules from
  protein-protein interaction networks. 2013, Nucleic Acids Research
  (in press)

A BibTeX entry for LaTeX users is

  @Article{,
    title = {Systematic identification of transcriptional regulatory modules from protein-protein int
    author = {Diego Diez and Andrew P. Hutchins and Diego Miranda-Saavedra},
    year = {2013 (in press)},
    journal = {Nucleic Acids Research},
  }
```

8 Session Information

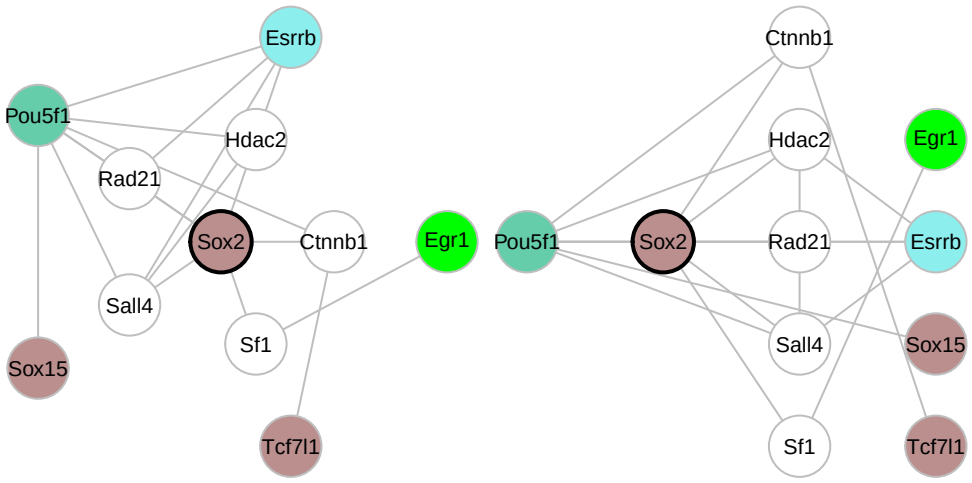

Figure 5: Sox2 TRM obtained with PWMEnrich workflow and layout.concentric is shown in the left. Same TRM with layout.arc is shown in the right.

```
> sessionInfo()
R version 3.0.2 Patched (2013-12-18 r64488)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] MotifDb_1.4.1                      PWMEnrich_2.6.2
 [3] BSgenome.Mmusculus.UCSC.mm8_1.3.17 BSgenome_1.30.0
 [5] Biostrings_2.30.1                  GenomicRanges_1.14.4
 [7] XVector_0.2.0                      IRanges_1.20.6
 [9] codetools_0.2-8                    org.Mm.eg.db_2.10.1
[11] rTRM_1.0.5                         annotate_1.40.0
[13] AnnotationDbi_1.24.0               Biobase_2.22.0
[15] BiocGenerics_0.8.0                 RSQLite_0.11.4
[17] DBI_0.2-7                          igraph_0.7.0
[19] knitr_1.5

loaded via a namespace (and not attached):
 [1] RCurl_1.95-4.1     Rsamtools_1.14.3   XML_3.98-1.1       bitops_1.0-6
 [5] evaluate_0.5.1     evd_2.3-0          formatR_0.10       gdata_2.13.2
```

```
 [9] gtools_3.3.0       highr_0.3          rtracklayer_1.22.3 seqLogo_1.28.0
[13] stats4_3.0.2       stringr_0.6.2      tools_3.0.2        xtable_1.7-1
[17] zlibbioc_1.8.0
```

References

[1] Hutchins, A. P., D. Diez, et al. (2013). "Distinct transcriptional regulatory modules underlie STAT3's cell type-independent and cell type-specific functions." Nucleic Acids Res.

[2] Stark, C., B. J. Breitkreutz, et al. (2011). "The BioGRID Interaction Database: 2011 update." Nucleic Acids Res 39(Database issue): D698-704.

[3] Chen, X., H. Xu, et al. (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell 133(6): 1106-1117.

[4] Heinz, S., C. Benner, et al. (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Mol Cell 38(4): 576-589.

[5] Gupta, S., J. A. Stamatoyannopoulos, et al. (2007). "Quantifying similarity between motifs." Genome Biol 8(2): R24.

[6] Ho, L., E. L. Miller, et al. (2011). "esBAF facilitates pluripotency by conditioning the genome for LIF/STAT3 signalling and by regulating polycomb function." Nat Cell Biol 13(8): 903-913.

[7] Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Res 33(20): e175.

[8] Gautier, L., L. Cope, et al. (2004). "affy–analysis of Affymetrix GeneChip data at the probe level." Bioinformatics 20(3): 307-315.

[9] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell, 152(1-2), 327–339.