Network and enrichment analysis for the ionomics data in a publication

Wanchang Lin

2021-12-01

Contents

Data preparation

This document explains how to perform network and enrichment analysis on the processed data set in publication:

Yu, D., Danku, J.M.C., Baxter, I. et al. **High-resolution genome-wide scan of genes, gene-networks and cellular systems impacting the yeast ionome**. BMC Genomics 13, 623 (2012). (https://doi.org/10.1186/1471-2164-13-623)

The authors identified 1065 strains with an altered ionome, including 584 haploid and 35 diploid deletion strains, and 446 over expression strains.

To explore the ionomics pipeline, we'll use the ionomics data set from the paper:

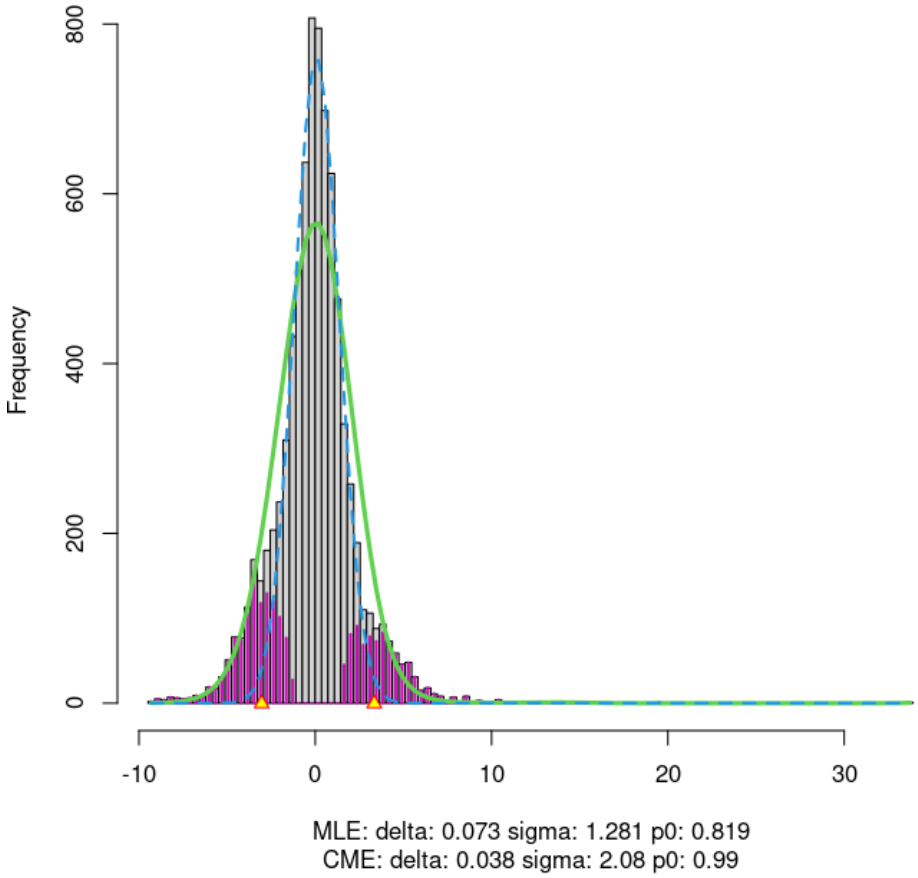
```
## identified 584 haploid
dat <- read_csv("paper_ko.csv")
#> Rows: 584 Columns: 15
#> -- Column specification --------------------------------------------------------
#> Delimiter: ","
#> chr  (1): gene
#> dbl (14): Ca44, Cd111, Co59, Cu65, Fe57, K39, Mg25, Mn55, Mo95, Na23, Ni60, ...
#>
#> i Use `spec()` to retrieve the full column specification for this data.
#> i Specify the column types or set `show_col_types = FALSE` to quiet this message.
dim(dat)
#> [1] 584  15
```

This data has missing values. We use a simple missing value filling function to impute the missing value:

```
## missing valuee filling with mean
dat <- dat %>%
  mutate(across(where(is.numeric), function(x) {
    m <- mean(x, na.rm = TRUE)
    x[is.na(x)] <- m
    x
  }))
dat
#> # A tibble: 584 x 15
#>    gene        Ca44  Cd111    Co59   Cu65     Fe57     K39   Mg25   Mn55   Mo95
#>    <chr>      <dbl>  <dbl>   <dbl>  <dbl>    <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
#>  1 YAL002W    3.83   5.07  -0.367   0.944 -0.0847  -1.32   -1.36   2.65  -1.08
#>  2 YAL016W    2.82   4.77  -8.35   -1.04  -1.16    -1.54   -6.34   7.70  -4.13
#>  3 YAL048C    0.163 -0.786 -0.0411  0.165 -0.0847  -4.83   -0.648 -5.01  -1.99
#>  4 YAL067W-A -4.91  -0.213  0.622   0.417 -0.00964  2.99    0.961  0.893  0.832
#>  5 YBL007C   -1.19  -1.94  -3.04   -1.29  -0.241   -2.71   -3.42   1.46   1.37
#>  6 YBL017C    3.17   4.29   0.826   0.424  0.607   -0.915  -0.518  4.71   2.55
#>  7 YBL024W    1.55   8.61   1.06    0.190 -2.59     0.680   3.16  -0.478 -1.94
#>  8 YBL027W   -0.365 -5.04  -4.48   -2.14   0.137   -1.19   -1.94  -0.786 -1.00
#>  9 YBL047C    4.29   1.75  -0.0760 -0.305  1.41    -0.850  -0.442  1.18   2.28
#> 10 YBL068W    1.93   4.09   0.275  -0.802  1.00    -0.0608 -0.262  3.12   4.08
#> # ... with 574 more rows, and 5 more variables: Na23 <dbl>, Ni60 <dbl>,
#> #   P31 <dbl>, S34 <dbl>, Zn66 <dbl>
```

To symbolise the data set, we use local FDR to estimate the symbolisation threshold:

```
res <- locfdr_filter(t(dat[, -1]), plot = 1)
#> Warning in locfdr::locfdr(vec, nulltype = 1, plot = plot, ...): f(z) misfit =
#> 7.2. Rerun with increased df
```




```
res$thres
#> [1] -3.039761  3.345607
## dat <- dat[res$idx, , drop = FALSE]
## symbolise data
dat_sym <- dat %>%
  mutate(across(where(is.numeric), ~ dat_symb(., thres = res$thres)))
```

Some of ionomics data and symbolic data are like:

```
dat %>% sample_n(10) %>%
  kable(caption = 'Ionomics data', digits = 2, booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

```
dat_sym %>% sample_n(10) %>%
  kable(caption = 'Symbolic data', booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
```

Table 1: Ionomics data

gene	Ca44	Cd111	Co59	Cu65	Fe57	K39	Mg25	Mn55	Mo95	Na23	Ni60	P31	S34	Zn66
YLR146C	-2.83	3.50	-0.16	0.93	-0.34	-0.41	1.45	1.00	0.39	-2.02	-0.05	0.69	1.64	0.42
YGR104C	2.01	-6.90	-5.65	1.21	0.48	-1.40	-3.18	0.00	3.08	-3.61	0.29	-2.91	-1.79	0.70
YJL208C	1.84	4.80	0.99	-1.08	0.80	0.19	0.62	1.90	0.59	-1.04	1.71	0.59	0.74	0.84
YDR176W	-0.11	4.89	0.81	-0.41	0.25	-0.42	-1.69	-1.97	1.50	-2.95	-0.94	-0.72	-2.52	-0.65
YGL212W	0.64	2.70	0.96	-0.08	-1.91	0.24	-5.03	2.90	-1.76	6.64	-2.41	-2.88	-0.68	-1.38
YJL204C	3.30	-0.63	-3.93	-1.56	1.71	-0.97	-1.03	2.29	1.94	0.57	-1.59	-2.23	-0.46	-0.10
YLR034C	1.63	4.49	1.12	0.48	0.70	-0.95	-0.06	8.49	2.87	-1.50	1.07	0.75	0.28	0.33
YPL104W	-2.61	-0.11	2.10	-0.56	0.88	-3.33	1.95	-2.16	-0.65	-0.05	0.54	1.24	0.14	-0.27
YMR084W	0.60	-0.51	1.09	-1.35	0.32	-3.60	-0.05	-3.68	-0.88	-0.19	0.70	-1.68	-0.46	0.11
YLR025W	-3.07	3.07	-1.32	0.21	-1.69	-4.06	-0.33	1.82	-1.45	2.44	-1.25	-2.23	0.97	-2.66

```
                latex_options = c("striped", "scale_down"))
```

Table 2: Symbolic data

gene	Ca44	Cd111	Co59	Cu65	Fe57	K39	Mg25	Mn55	Mo95	Na23	Ni60	P31	S34	Zn66
YLR089C	0	-1	0	0	0	0	0	0	0	0	0	0	0	0
YLR429W	0	1	0	0	0	0	0	0	0	-1	0	0	0	0
YPL179W	0	0	0	0	0	0	1	0	0	0	1	1	0	0
YIL097W	0	0	0	0	0	0	1	0	0	1	0	0	1	0
YGR045C	0	1	0	0	0	0	0	0	0	0	0	0	0	0
YDR455C	1	0	0	-1	0	0	0	-1	0	0	0	0	0	0
YMR157C	-1	0	0	0	0	0	0	0	0	0	0	0	0	0
YDR436W	-1	0	0	-1	0	0	0	0	0	0	0	0	-1	0
YLR403W	0	0	0	0	0	0	1	0	0	0	0	0	0	0
YGR159C	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1

The data has been processed and the gene are significant. They are ready for network and enrichment analysis.

Clustering

The gene network and enrichment analysis are based on gene data clustering. The available similarity measures are *correlation*, *correlation*, *cosine*,*eJaccard* and *Mahalanobis*. The last one is converted from distance and will take a considerable time to run. The method for clustering are from R base's `hclust` such as *complete* and *average*.

One of example:

```
clus <- simil_clus(x = dat[, -1], method_simil = "correlation",
                   method_hclus = "ave", thres_simil = 0.6, thres_clus = 5)
names(clus)
#> [1] "sim"     "clus"    "idx"     "tab"     "tab_sub"
```

The cluster and its number of elements(larger than `thres_clus`) are:

```
clus$tab_sub
#>    cluster nGenes
#> 1        1    154
#> 2        3     34
#> 3        6     30
#> 4       22     28
#> 5       29     27
```

```
#> 6       17     23
#> 7        4     20
#> 8       25     18
#> 9       16     16
#> 10      21     15
#> 11      20     14
#> 12       7     12
#> 13      35     12
#> 14      11     10
#> 15      14     10
#> 16      12      9
#> 17      53      9
#> 18      66      8
#> 19      39      7
#> 20      45      7
#> 21      52      7
#> 22       2      6
#> 23      13      6
#> 24      27      6
#> 25      42      6
```

The function `gene_clus` is also used for clustering but only for symbolic data:

```
clus_1 <- gene_clus(x = dat_sym[, -1], thres_clus = 5)
names(clus_1)
#> [1] "clus"    "idx"     "tab"     "tab_sub"
clus_1$tab_sub
#>    cluster nGenes
#> 1        7     97
#> 2        3     29
#> 3       16     23
#> 4        4     22
#> 5       18     16
#> 6       27     16
#> 7       46     16
#> 8        6     14
#> 9       14     14
#> 10      40     14
#> 11      32     12
#> 12      41     10
#> 13      54      7
#> 14     102      7
#> 15      34      6
#> 16      92      6
```

Gene network

Here we set up the parameters for network and enrichment analysis. The similarity measure method is one of *correlation*, *cosine*, *eJaccard* and *Mahalanobis*. The hierarchical clustering method is from `hclust`. This should be (an unambiguous abbreviation of) one of *ward.D*, *ward.D2*, *single*, *complete*, *average*, *mcquitty*, *median* or *centroid*.

```
method_simil <- "correlation"
method_hclus <- "ave"
thres_simil  <- 0.70
thres_clus   <- 10
```

Network analysis with ionomics data only

The gene network is based on the pre-processed data and use the similarity measure results to build up the network.

```
net <- gene_net(data = dat,
                method_simil = method_simil,
                method_hclus = method_hclus,
                thres_simil = thres_simil,
                thres_clus = thres_clus)
```

The node colours are indicated by the similarity measures and the network community detection, i.e. clustering.

```
net$plot_net_1
```

```
net$plot_net_2
```

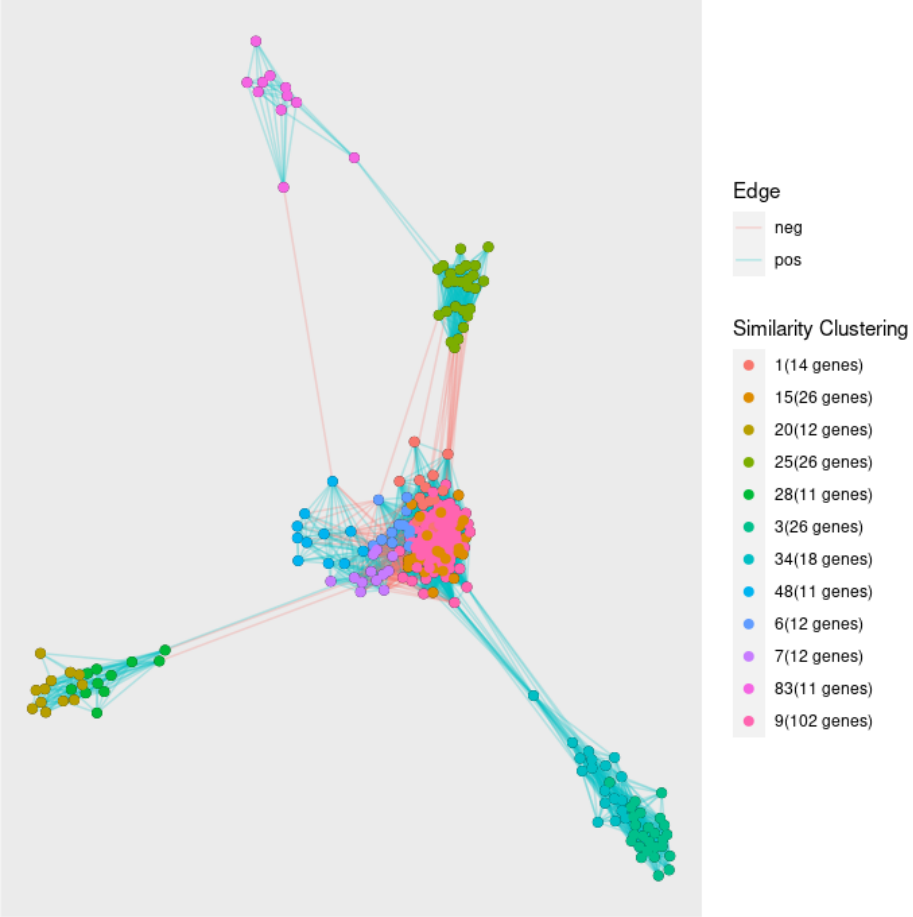


Figure 1: Netwok analysis based on ionomics data

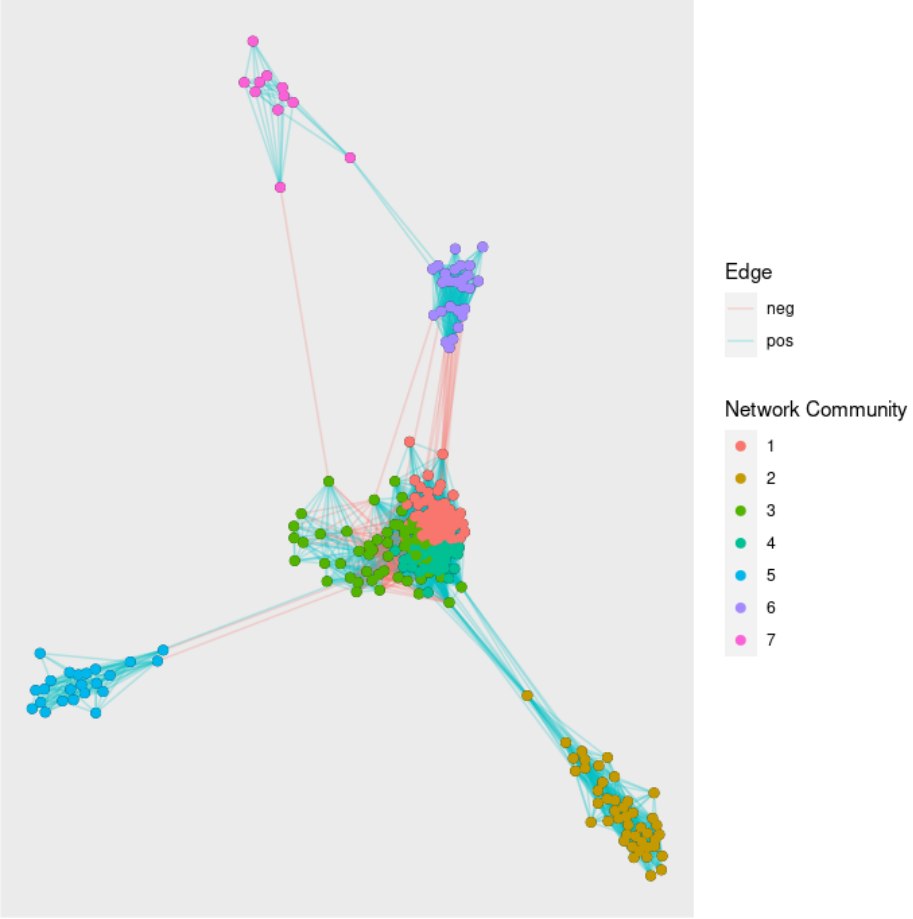


Figure 2: Netwok analysis based on ionomics data

Network analysis with ionomics data and symbolic data

`gene_network` need both ionomics data and its symbolic data. The later is used for hierarchical clustering while the former is used for correlation analysis. The results of two analysis are used for network build up.

```
net_1 <- gene_network(data = dat, data_symb = dat_sym,
                      method_simil = method_simil,
                      thres_simil = thres_simil,
                      thres_clus = thres_clus)

net_1$plot_net_1
```

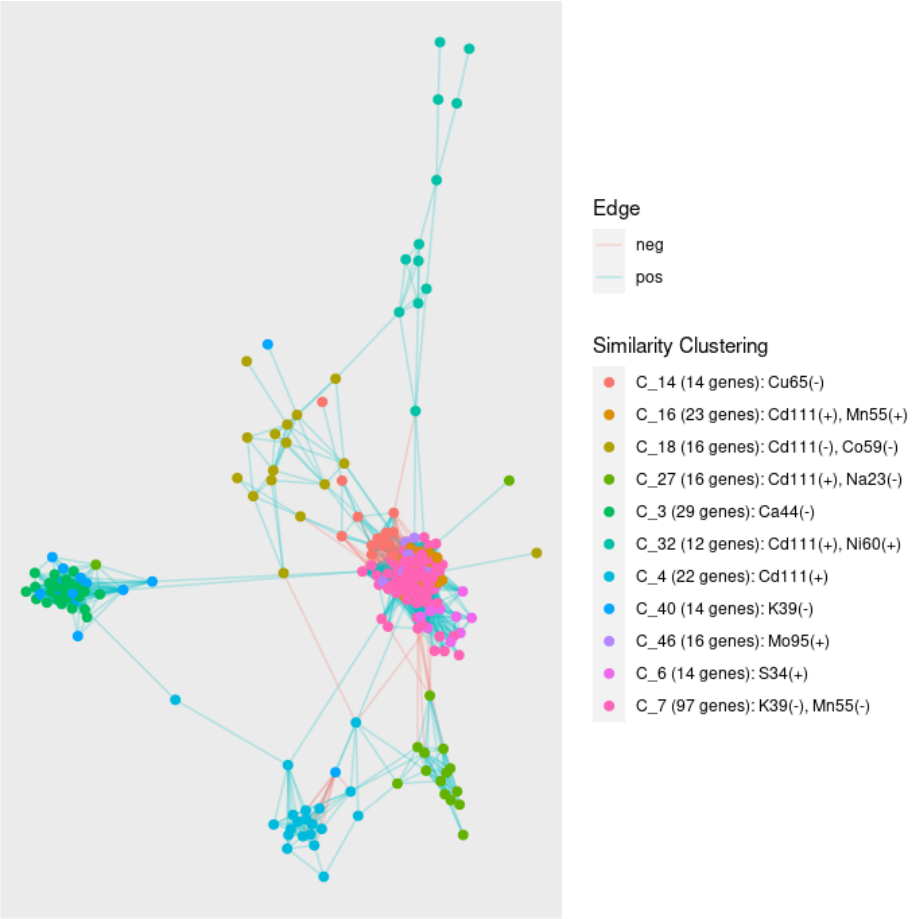


Figure 3: Netwok analysis based on symbolic data

```
net_1$plot_net_2
```

```
net_1$plot_impact_betweenness_1
```

```
net_1$plot_impact_betweenness_2
#> Warning: ggrepel: 184 unlabeled data points (too many overlaps). Consider
#> increasing max.overlaps
```

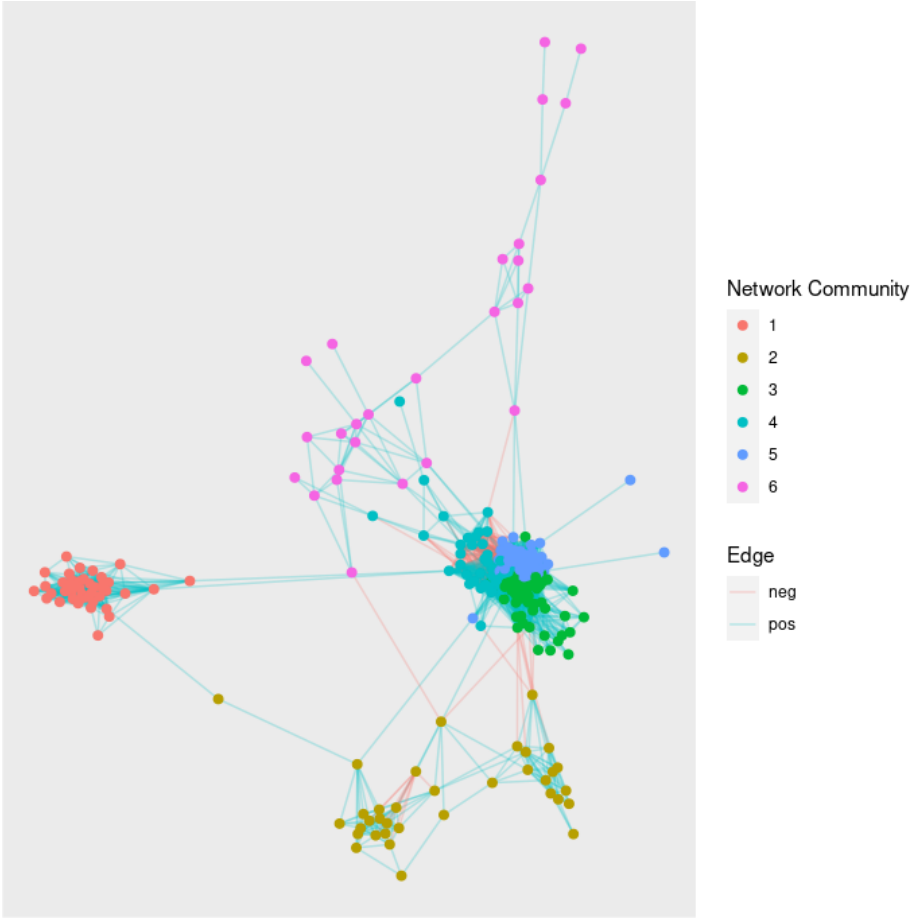


Figure 4: Netwok analysis based on symbolic data

```
head(net_1$impact_betweenness)
#>              var_id impact log.betweenness                      position
#> YAL048C     YAL048C  7.653          5.159   Low impact, low betweenness
#> YAL067W-A YAL067W-A  6.265          4.670   Low impact, low betweenness
#> YBL017C     YBL017C  8.321          6.473 High impact, high betweenness
#> YBL024W     YBL024W 10.870          3.562  High impact, low betweenness
#> YBL093C     YBL093C  7.861          5.598 High impact, high betweenness
#> YBL102W     YBL102W  7.135          4.576   Low impact, low betweenness
#>                                  cluster
#> YAL048C           C_3 (29 genes): Ca44(-)
#> YAL067W-A        C_4 (22 genes): Cd111(+)
#> YBL017C            C_6 (14 genes): S34(+)
#> YBL024W   C_7 (97 genes): K39(-), Mn55(-)
#> YBL093C          C_14 (14 genes): Cu65(-)
#> YBL102W   C_7 (97 genes): K39(-), Mn55(-)
```

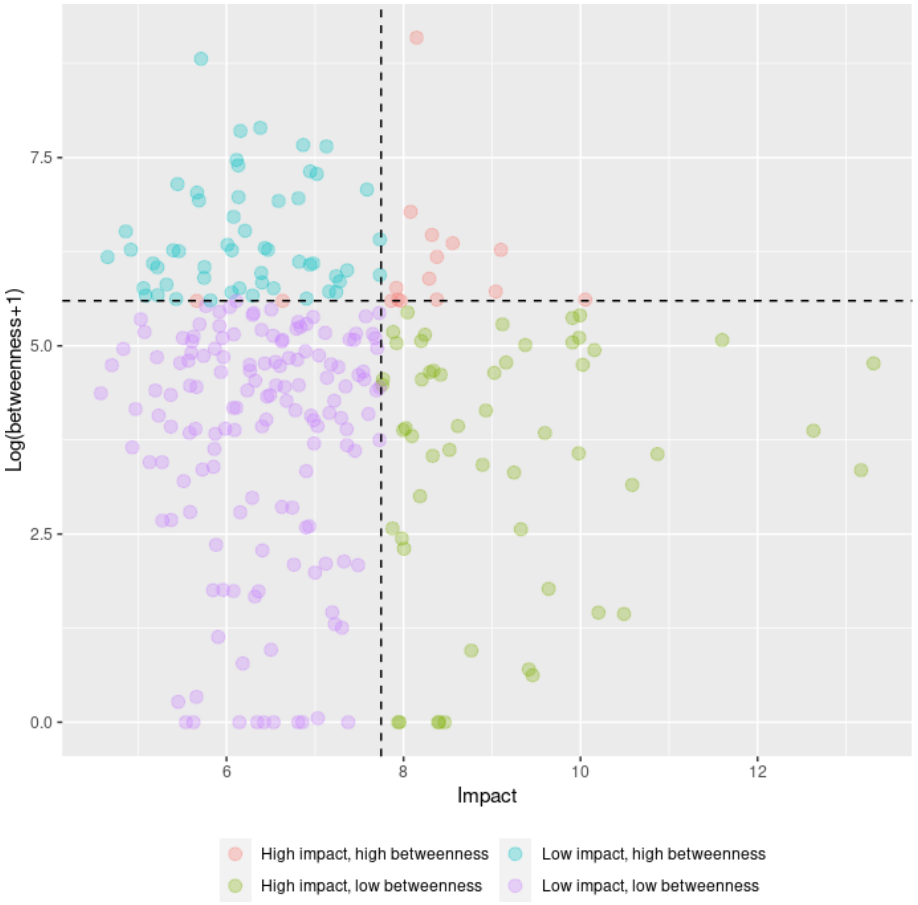


Figure 5: Netwok analysis based on symbolic data

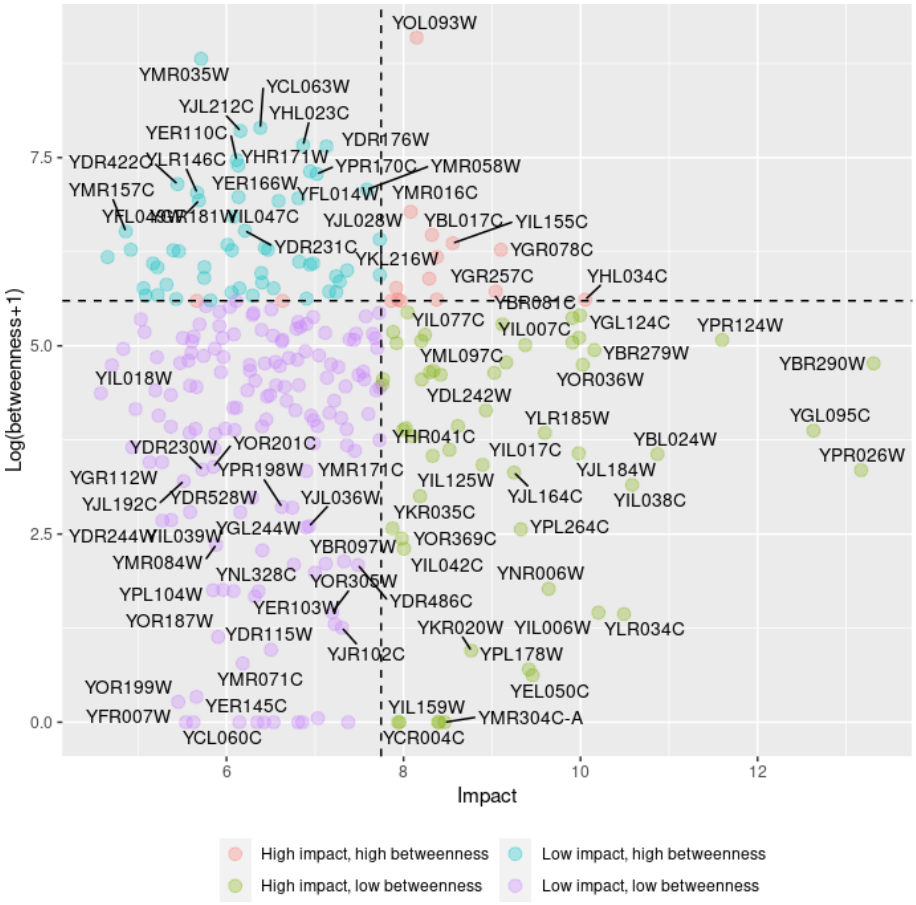


Figure 6: Netwok analysis based on symbolic data

Enrichment analysis

The results of hierarchical clustering either on ionomics data or on symbolic data are for enrichment analysis.

Enrichment analysis on ionomics data

The KEGG enrichment analysis:

```
kegg <- kegg_enrich(data = dat,
                    method_simil = method_simil,
                    method_hclus = method_hclus,
                    thres_simil = thres_simil,
                    thres_clus = thres_clus,
                    pval = 0.05,
                    annot_pkg = "org.Sc.sgd.db",
                    is_symb = F)
## kegg
kegg %>%
  kable(caption = 'KEGG enrichment analysis: ionomics data',
        digits = 3, booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

Table 3: KEGG enrichment analysis: ionomics data

Cluster	KEGGID	Pvalue	Count	Size	Term
Cluster 25 (26 genes)	03010	0.005	4	20	Ribosome
Cluster 25 (26 genes)	04141	0.023	2	6	Protein processing in endoplasmic reticulum
Cluster 34 (18 genes)	01110	0.019	7	19	NA
Cluster 34 (18 genes)	00030	0.026	2	2	Pentose phosphate pathway
Cluster 34 (18 genes)	00520	0.026	2	2	Amino sugar and nucleotide sugar metabolism
Cluster 34 (18 genes)	00052	0.032	3	5	Galactose metabolism
Cluster 6 (12 genes)	00190	0.000	5	12	Oxidative phosphorylation
Cluster 6 (12 genes)	04145	0.000	5	14	Phagosome
Cluster 6 (12 genes)	01100	0.006	5	58	Metabolic pathways
Cluster 7 (12 genes)	03040	0.039	2	6	Spliceosome
Cluster 20 (12 genes)	00190	0.000	5	12	Oxidative phosphorylation
Cluster 20 (12 genes)	04145	0.000	5	14	Phagosome
Cluster 20 (12 genes)	01100	0.006	5	58	Metabolic pathways
Cluster 28 (11 genes)	00970	0.017	2	6	Aminoacyl-tRNA biosynthesis

Note that there can be none results for KRGG enrichment analysis. Change arguments such as `thres_clus` as appropriate.

The GO Terms enrichment analysis:

```
go <- go_enrich(data = dat,
                method_simil, method_hclus, thres_simil, thres_clus,
                pval = 0.05, ont = "BP", annot_pkg =  "org.Sc.sgd.db",
                is_symb = F)
## go
go %>% head() %>%
  kable(caption = 'GO Terms enrichment analysis: ionomics data',
        digits = 3, booktabs = T) %>%
```

```
        kable_styling(full_width = F, font_size = 10,
                      latex_options = c("striped", "scale_down"))
```

Table 4: GO Terms enrichment analysis: ionomics data

Cluster	ID	Description	Pvalue	Count	CountUniverse	Ontology
Cluster 9 (102 genes)	GO:0072594	establishment of protein localization to organelle	0.0012	5	42	BP
Cluster 9 (102 genes)	GO:0042886	amide transport	0.0027	7	94	BP
Cluster 9 (102 genes)	GO:0015031	protein transport	0.0052	6	83	BP
Cluster 9 (102 genes)	GO:0033036	macromolecule localization	0.0071	7	110	BP
Cluster 9 (102 genes)	GO:0071702	organic substance transport	0.0071	7	110	BP
Cluster 9 (102 genes)	GO:0000011	vacuole inheritance	0.0105	2	7	BP

Enrichment analysis on symbolic data

The KEGG enrichment analysis:

```
kegg_1 <- kegg_enrich(data = dat_sym,
                      method_simil = method_simil,
                      method_hclus = method_hclus,
                      thres_simil = thres_simil,
                      thres_clus = thres_clus,
                      pval = 0.05,
                      annot_pkg = "org.Sc.sgd.db",
                      is_symb = T)
## kegg_1
kegg_1 %>%
  kable(caption = 'KEGG enrichment analysis: Symbolic data', digits = 3,
        booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

Table 5: KEGG enrichment analysis: Symbolic data

Cluster	KEGGID	Pvalue	Count	Size	Term
Cluster 16 (23 genes)	04144	0.005	3	18	Endocytosis
Cluster 16 (23 genes)	04145	0.040	2	14	Phagosome
Cluster 4 (22 genes)	01100	0.039	11	58	Metabolic pathways
Cluster 18 (16 genes)	04141	0.023	2	6	Protein processing in endoplasmic reticulum
Cluster 18 (16 genes)	03010	0.044	3	20	Ribosome
Cluster 46 (16 genes)	03010	0.042	2	20	Ribosome
Cluster 32 (12 genes)	04113	0.001	3	6	NA

The GO Terms enrichment analysis:

```
go_1 <- go_enrich(data = dat_sym,
                  method_simil, method_hclus, thres_simil, thres_clus,
                  pval = 0.05, ont = "BP", annot_pkg =  "org.Sc.sgd.db",
                  is_symb = T)
## go_1
go_1 %>% head() %>%
  kable(caption = 'GO Terms enrichment analysis: Symbolic data',
        digits = 3, booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

Table 6: GO Terms enrichment analysis: Symbolic data

Cluster	ID	Description	Pvalue	Count	CountUniverse	Ontology
Cluster 7 (97 genes)	GO:0032543	mitochondrial translation	0	9	25	BP
Cluster 7 (97 genes)	GO:0043043	peptide biosynthetic process	0	12	71	BP
Cluster 7 (97 genes)	GO:0010821	regulation of mitochondrion organization	0.0024	2	2	BP
Cluster 7 (97 genes)	GO:0044267	cellular protein metabolic process	0.0112	13	152	BP
Cluster 7 (97 genes)	GO:0009060	aerobic respiration	0.0136	2	4	BP
Cluster 7 (97 genes)	GO:0034645	cellular macromolecule biosynthetic process	0.0195	12	142	BP