Ionomics analysis for human data set using Ionflow

Wanchang Lin

11-12-2020, Fri

Contents

Data preparation

The human ionomics data set has been pre-processed. We need to get the symbolic data:

```
dat <- read.table("./test-data/human.csv", header = T, sep = ",")
dat <- dat[!duplicated(dat[, 1]), ]
colnames(dat)[1] <- "Line"
dat_symb <- symbol_data(x = dat, thres_symb = 3)
```

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"))
```

Table 1: Ionomics data

Line	As	B	Ca	Cd	Co	Cu	Fe	K	Li	Mg	Mn	Mo	Na	Ni	P	S	Se	Zn
EIF2B1	-3.86	0.65	-0.34	2.61	0.05	0.75	0.18	1.42	2.57	1.47	-3.91	0.04	0.05	-0.23	3.29	0.06	1.78	4.51
CALML3	-0.11	-0.58	-0.73	-0.10	-0.30	-2.76	-1.98	-1.45	1.06	-2.09	-0.11	-1.19	-1.12	0.31	-1.78	-0.10	1.35	-1.16
DNAJB11	-1.30	-0.53	-2.05	0.09	-2.83	-0.38	-0.19	0.47	-1.56	0.48	-2.17	3.60	-1.32	0.96	1.29	-1.97	0.24	1.59
UBE2I	-0.54	0.80	2.60	3.49	3.29	-0.15	0.39	2.03	-0.76	0.85	1.49	1.10	1.78	-0.13	1.05	2.74	-1.66	-0.20
ZDHHC14	1.49	-0.14	-1.68	-0.69	0.62	-2.09	-2.51	-2.57	0.43	-2.25	0.12	0.59	-1.57	-0.59	-2.10	0.84	1.69	-0.96
DUSP14	-0.52	-0.05	-0.43	0.16	-0.96	-0.80	-0.42	0.20	-0.99	-1.18	0.95	-0.48	-0.76	-0.15	-0.66	-1.05	0.17	-1.53
ABCA2	2.46	2.42	2.08	-0.67	2.12	-1.92	-1.50	-1.79	1.89	-2.60	1.26	2.28	1.74	0.96	-1.45	0.05	0.46	-1.48
MRPL52	0.43	1.05	0.27	2.15	0.55	0.86	0.76	-1.07	1.20	-0.63	-4.20	0.12	0.40	-0.49	0.64	0.84	2.81	1.51
RHOA	-6.32	1.95	-0.94	0.16	0.49	-1.86	2.34	-7.72	4.25	3.08	0.56	2.09	-0.19	3.09	0.22	3.70	-0.12	-1.01
SBDS	-2.05	-2.90	-2.33	-3.72	-2.05	-1.26	-1.65	1.04	-1.99	1.54	0.80	-2.21	-2.83	-1.21	0.57	-2.47	-0.80	-0.01

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

Table 2: Symbolic data

Line	As	B	Ca	Cd	Co	Cu	Fe	K	Li	Mg	Mn	Mo	Na	Ni	P	S	Se	Zn
PKLR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARV1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RPS5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TARS	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
GARS	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
OATL1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NMT1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OPA1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CALM1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SOAT1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

These data are filtered, i.e. remove all zero genes in symbolic data set:

```
idx <- rowSums(abs(dat_symb[, -1])) > 0
dat <- dat[idx, ]
dat_symb <- dat_symb[idx, ]
dim(dat)
#> [1] 195  19
```

Data clustering

The hierarchical cluster analysis is the key part of gene network and gene enrichment analysis. The methodology is as follow:

- Compute the distance of symbolic data
- Hierarchical cluster analysis on the distance
- Identify clusters/groups with a threshold of minimal number of cluster size

One example is:

```
min <- 8
clust <- gene_clus(dat_symb[, -1], min_clust_size = min)
names(clust)
#> [1] "clus"    "idx"     "tab"     "tab_sub"
clust$tab_sub
#>   cluster nGenes
#> 1      17     12
#> 2       4     11
```

Gene network

The gene network uses both the ionomics and symbolic data. The similarity measures on ionomics data are used to construct the network. Before creating a network, these analyses are further filtered by:

- clustering of symbolic data;
- and the similarity threshold located between 0 and 1;

The methods implemented are: *pearson*, *spearman*, *kendall*, *cosine*, *mahal_cosine* or *hybrid_mahal_cosine*.

We use the Pearson correlation as similarity measure for network analysis:

```
net <- GeneNetwork(data = dat,
                   data_symb = dat_symb,
                   min_clust_size = min,
                   thres_corr = 0.6,
                   method_corr = "pearson")
```

The network with nodes coloured by the symbolic data clustering is:

```
net$plot.pnet1
```

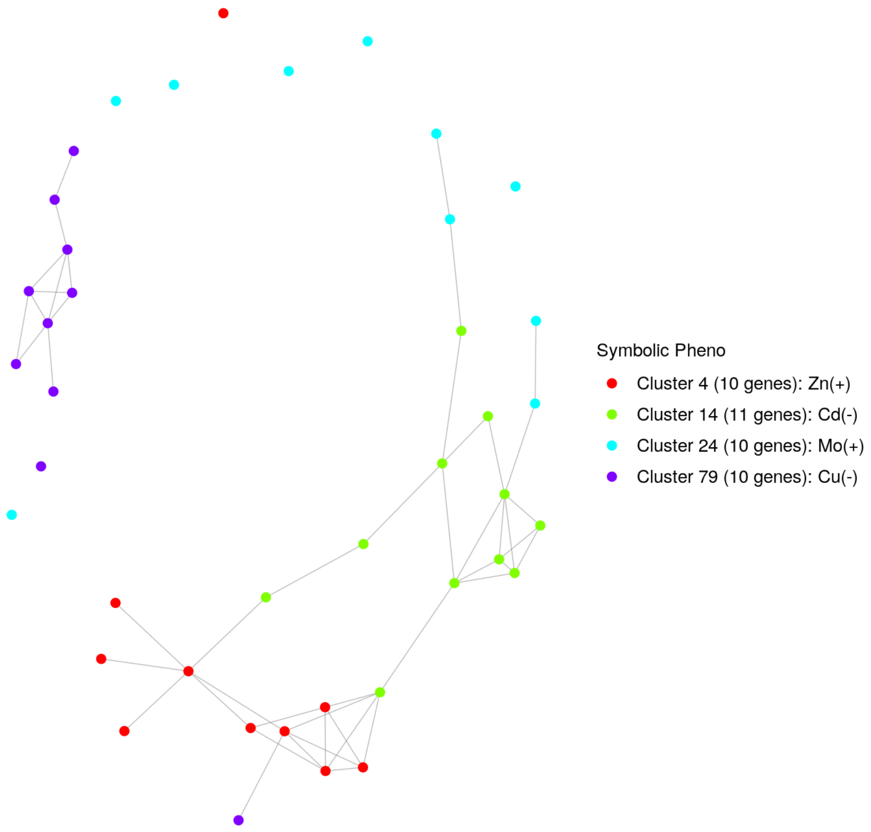

Figure 1: **Network with Pearson correlation: symbolic clustering**

The same network, but nodes are coloured by the network community detection:

```
net$plot.pnet2
```

The network analysis also returns a network impact and betweenness plot:

```
net$plot.impact_betweenness
```

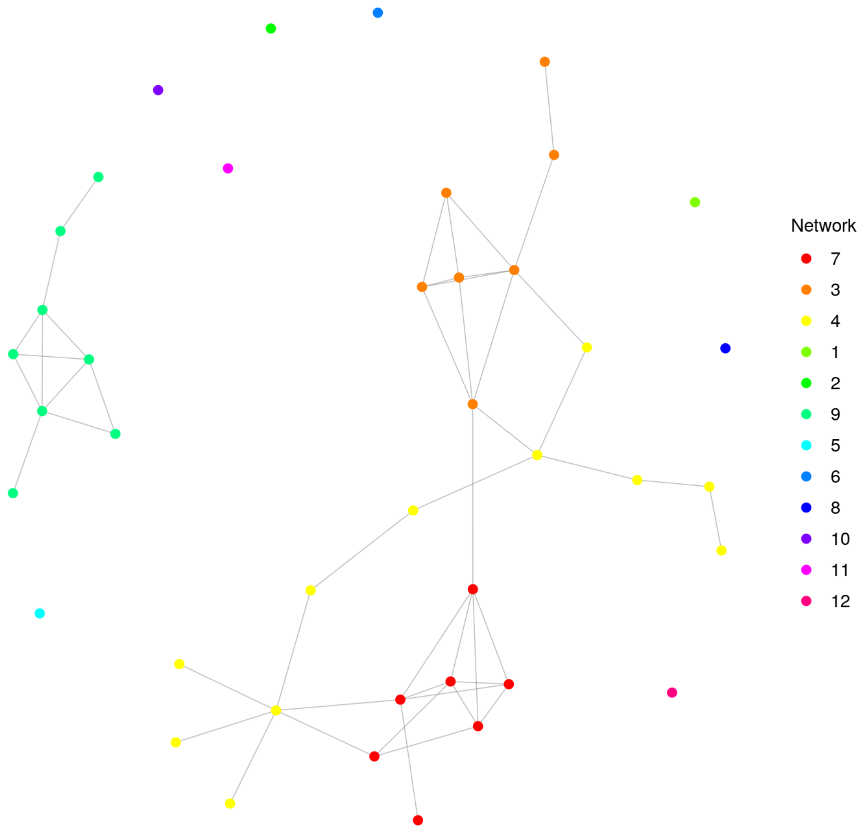

Figure 2: Network with Pearson correlation: community detction

Ionomics analysis for human data set using Ionflow

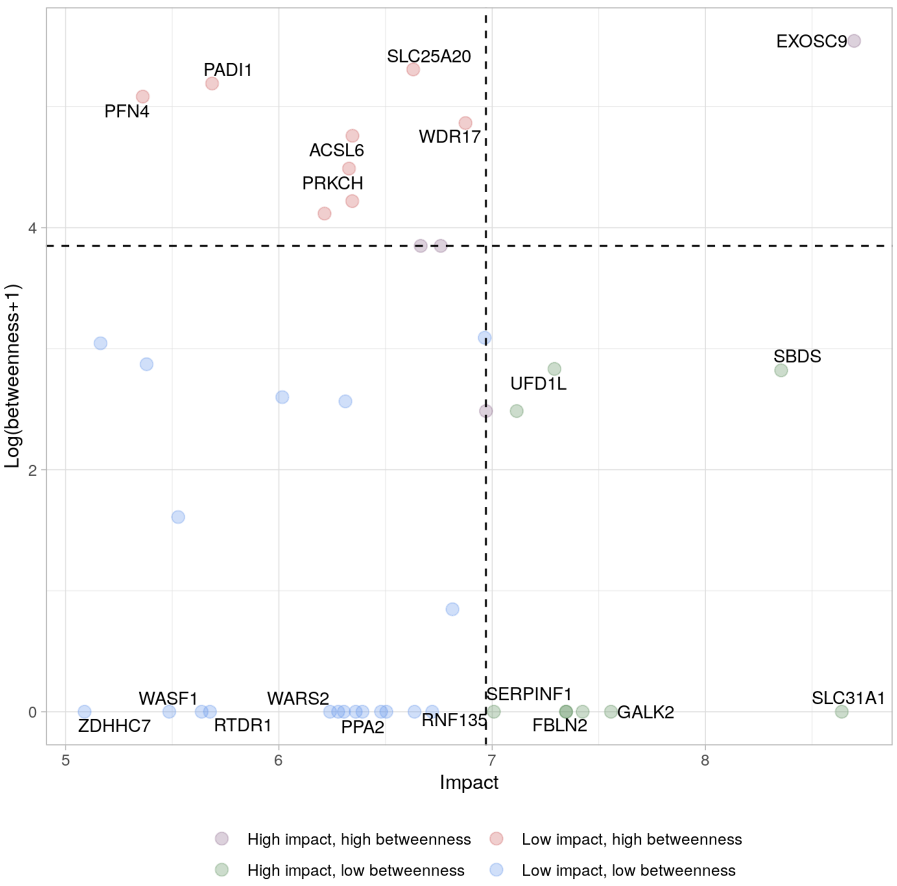

Figure 3: Network with Pearson correlation: impact and betweenness

For comparison purposes, we use *Mahalanobis Cosine*:

```
net_2 <- GeneNetwork(data = dat,
                     data_symb = dat_symb,
                     min_clust_size = min,
                     thres_corr = 0.6,
                     method_corr = "mahal_cosine")
net_2$plot.pnet1
```

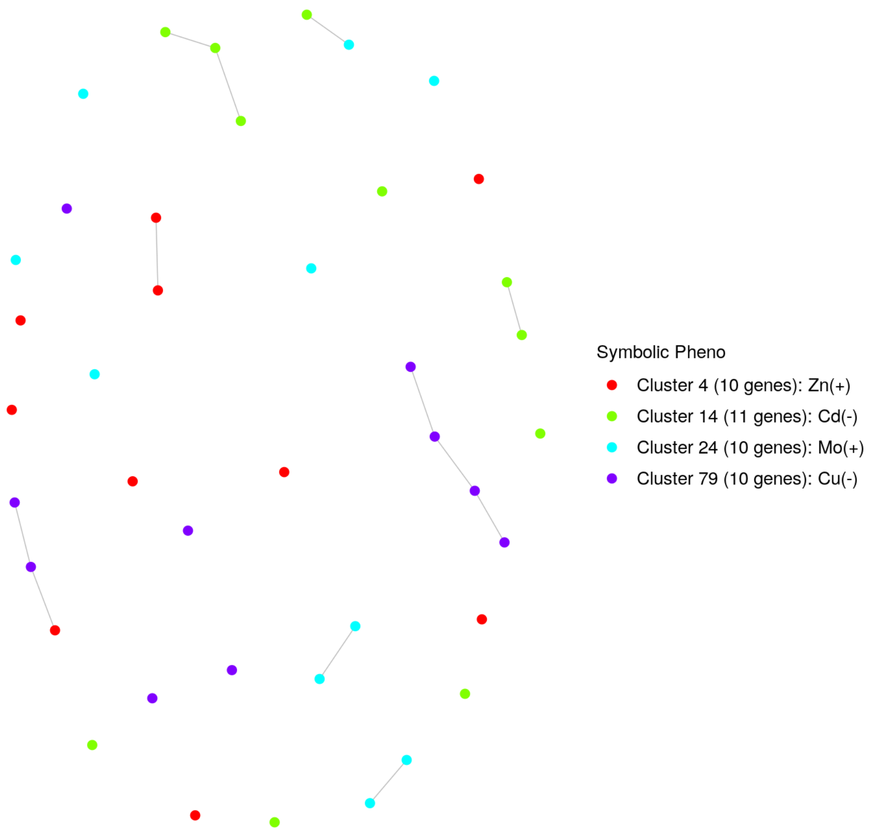

Figure 4: Network with Mahalanobis Cosine

```
net_2$plot.pnet2
```

Ionomics analysis for human data set using Ionflow

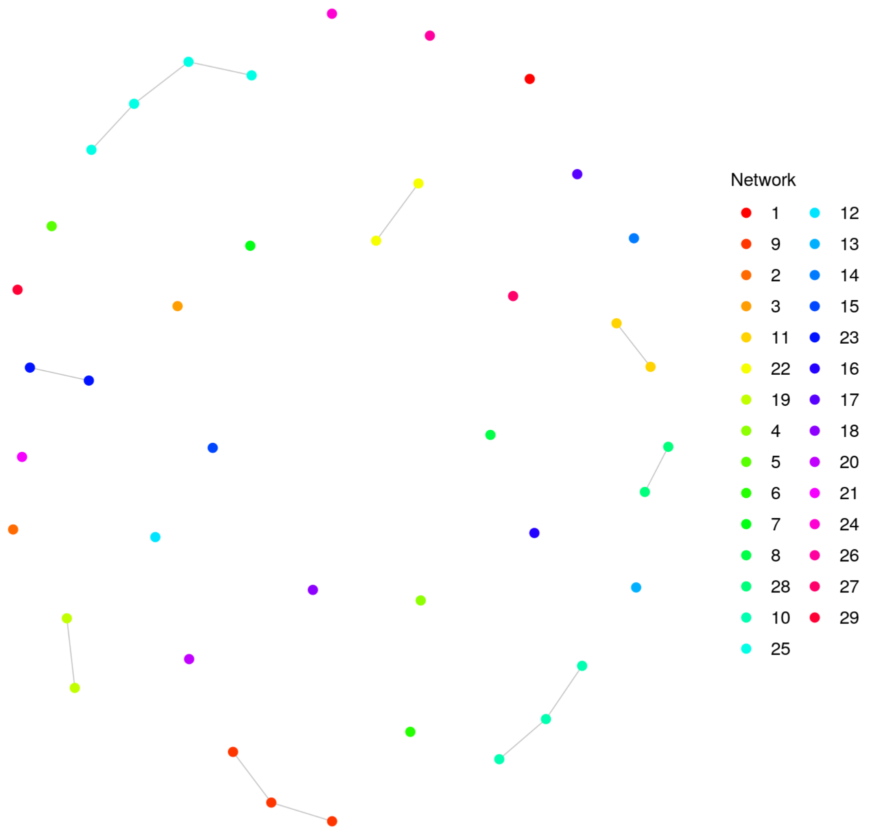

Figure 5: Network with Mahalanobis Cosine

Again, we use *Hybrid Mahalanobis Cosine*:

```
net_3 <- GeneNetwork(data = dat,
                     data_symb = dat_symb,
                     min_clust_size = min,
                     thres_corr = 0.6,
                     method_corr = "hybrid_mahal_cosine")
net_3$plot.pnet1
```

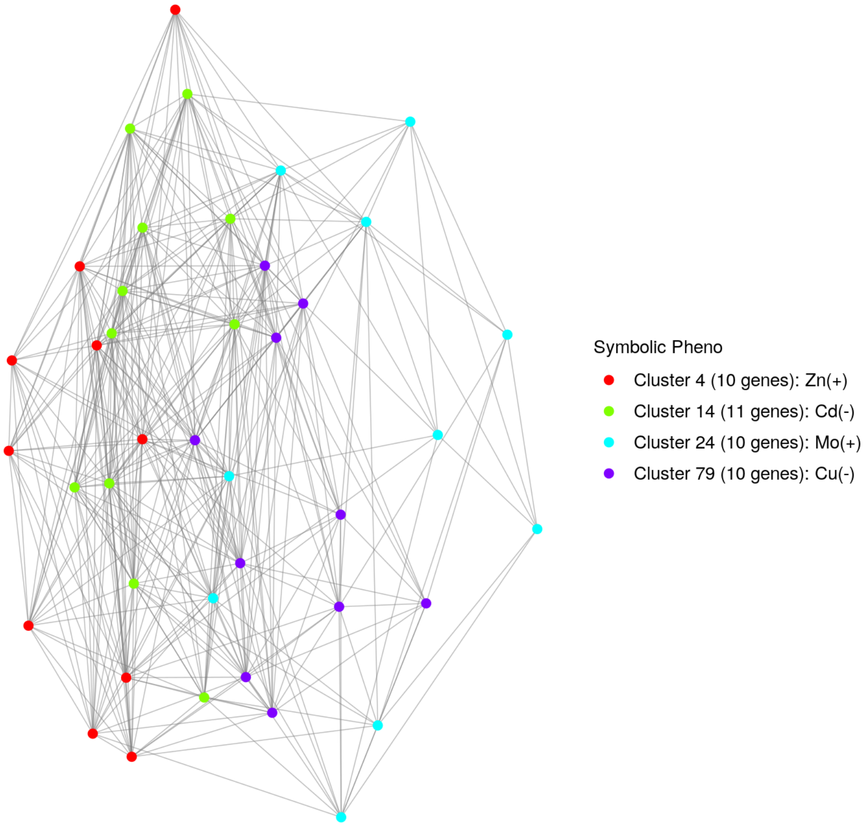
Symbolic Pheno
● Cluster 4 (10 genes): Zn(+)
● Cluster 14 (11 genes): Cd(-)
● Cluster 24 (10 genes): Mo(+)
● Cluster 79 (10 genes): Cu(-)

Figure 6: Network with Hybrid Mahalanobis Cosine

```
net_3$plot.pnet2
```

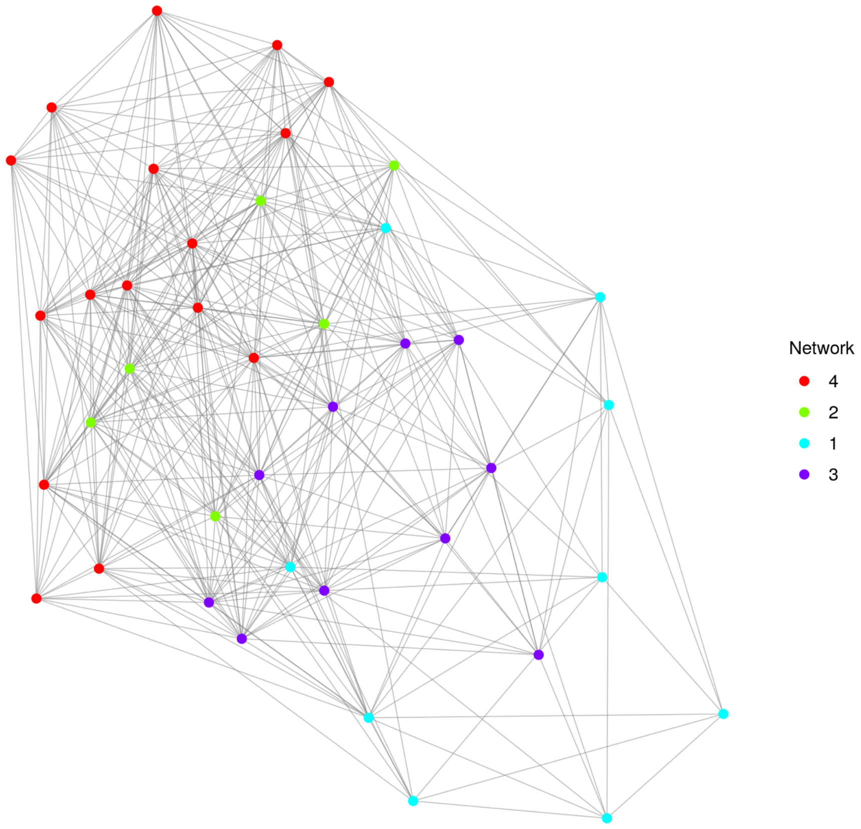

Figure 7: Network with Hybrid Mahalanobis Cosine

Enrichment analysis

The enrichment analysis is based on symbolic data clustering. The genes in clusters are considered target gene sets while genes in the whole data set is the universal gene set.

The KEGG enrichment analysis with a p-values of 0.05:

```
kegg <- kegg_enrich(data = dat_symb, min_clust_size = min, pval = 0.05,
                    annot_pkg =  "org.Hs.eg.db")

#' kegg
kegg %>%
  kable(caption = 'KEGG enrichment analysis',
        digits = 3, booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

Table 3: **KEGG enrichment analysis**

Cluster	KEGGID	Pvalue	Count	Size	Term
Cluster 17 (12 genes)	03018	0.003	2	2	RNA degradation
Cluster 4 (11 genes)	03010	0.007	2	5	Ribosome
Cluster 9 (6 genes)	03040	0.026	2	4	Spliceosome
Cluster 35 (6 genes)	00520	0.001	2	2	Amino sugar and nucleotide sugar metabolism

Note that there could be no results returned for KEGG enrichment analysis. Arguments such as `min_clust_size` can be changed as appropriate.

The GO Terms enrichment analysis with ontology of *BP* (other two are *MF* and *CC*):

```
go <- go_enrich(data = dat_symb, min_clust_size = min, pval = 0.05,
                ont = "BP", annot_pkg =  "org.Hs.eg.db")
#' go
go %>% head() %>%
  kable(caption = 'GO Terms enrichment analysis',
        digits = 3, booktabs = T) %>%
  kable_styling(full_width = F, font_size = 10,
                latex_options = c("striped", "scale_down"))
```

Table 4: **GO Terms enrichment analysis**

Cluster	ID	Description	Pvalue	Count	CountUniverse	Ontology
Cluster 17 (12 genes)	GO:0016180	snRNA processing	0.0036	2	2	BP
Cluster 17 (12 genes)	GO:0034427	nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'	0.0036	2	2	BP
Cluster 17 (12 genes)	GO:0034475	U4 snRNA 3'-end processing	0.0036	2	2	BP
Cluster 17 (12 genes)	GO:0043928	exonucleolytic catabolism of deadenylated mRNA	0.0036	2	2	BP
Cluster 17 (12 genes)	GO:0090503	RNA phosphodiester bond hydrolysis, exonucleolytic	0.0036	2	2	BP
Cluster 17 (12 genes)	GO:0000460	maturation of 5.8S rRNA	0.0105	2	3	BP

Exploratory analysis

The explanatory analysis performs PCA and correlation analysis for ions in terms of genes. Note that this analysis treats ions as samples/replicates while genes are treated as variables/features. The explanatory analysis is initially employed at an early stage of the analysis.

We apply it to the pre-processed data `dat` before any other analysis:

```
expl <- ExploratoryAnalysis(data = dat)
names(expl)
#> [1] "plot.pca"       "data.pca.load"  "plot.corr"      "plot.corr.heat"
#> [5] "plot.heat"      "plot.net"
```

The PCA plot is:

```
expl$plot.pca
```

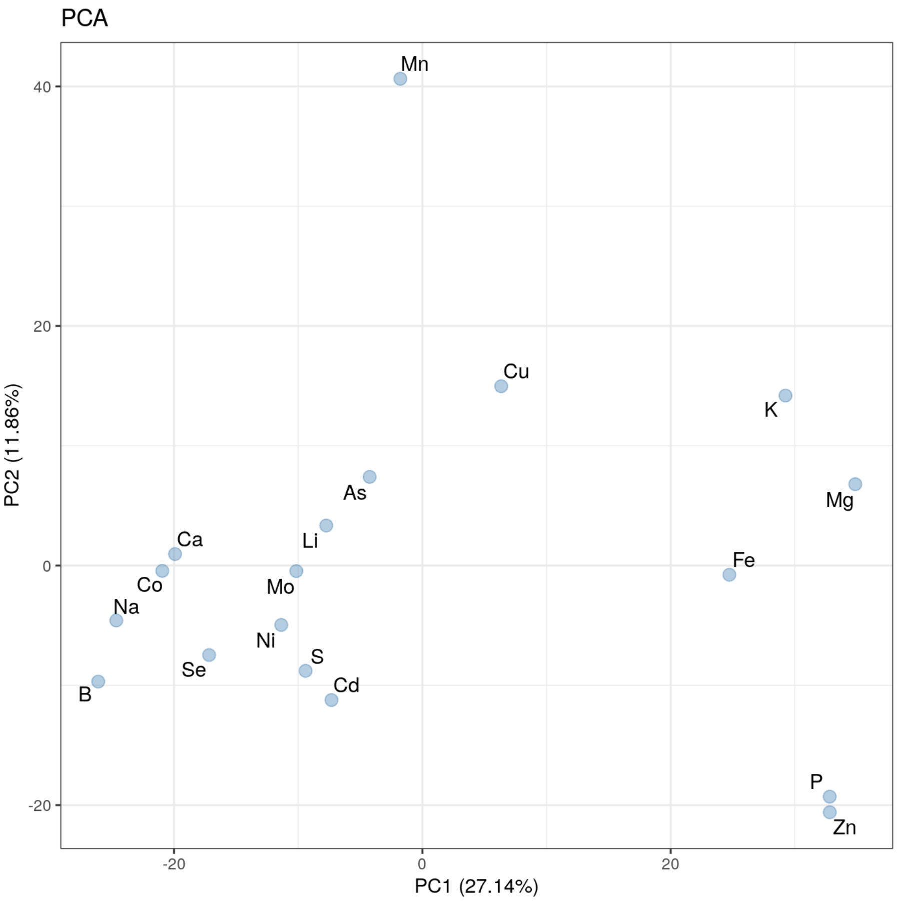

Figure 8: **Ion PCA plot on pre-processed data**

The Person correlation of ions are shown in correlation plot, heatmap and network plot:

```
expl$plot.corr
```

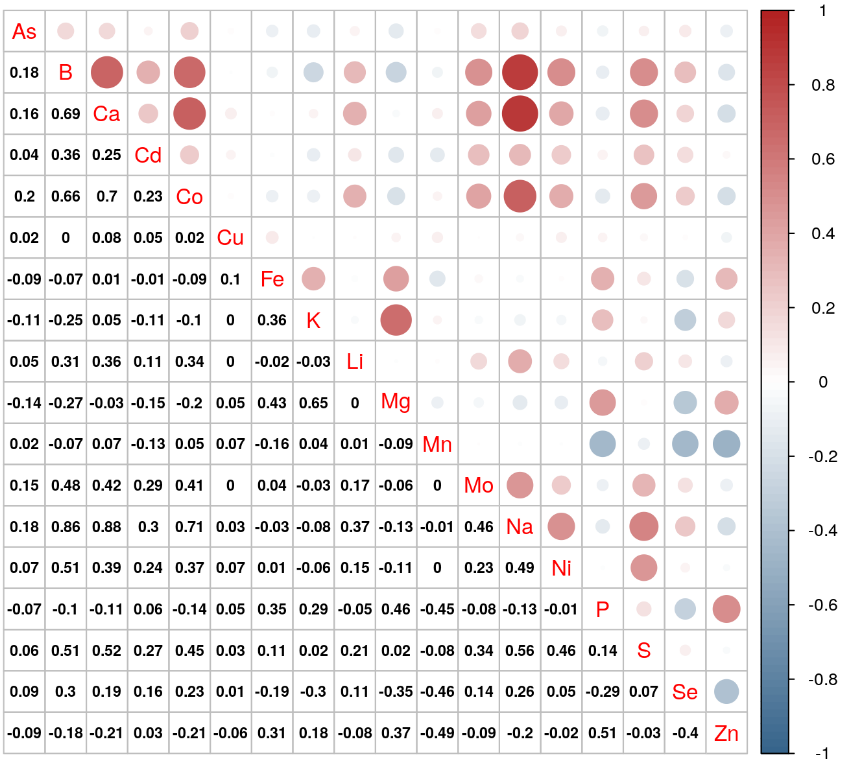

Figure 9: Ion correlation plots on pre-processed data

```
expl$plot.corr.heat
```

```
expl$plot.net
```

The correlation between ions and genes are shown in heatmap with dendrogram:

```
expl$plot.heat
```

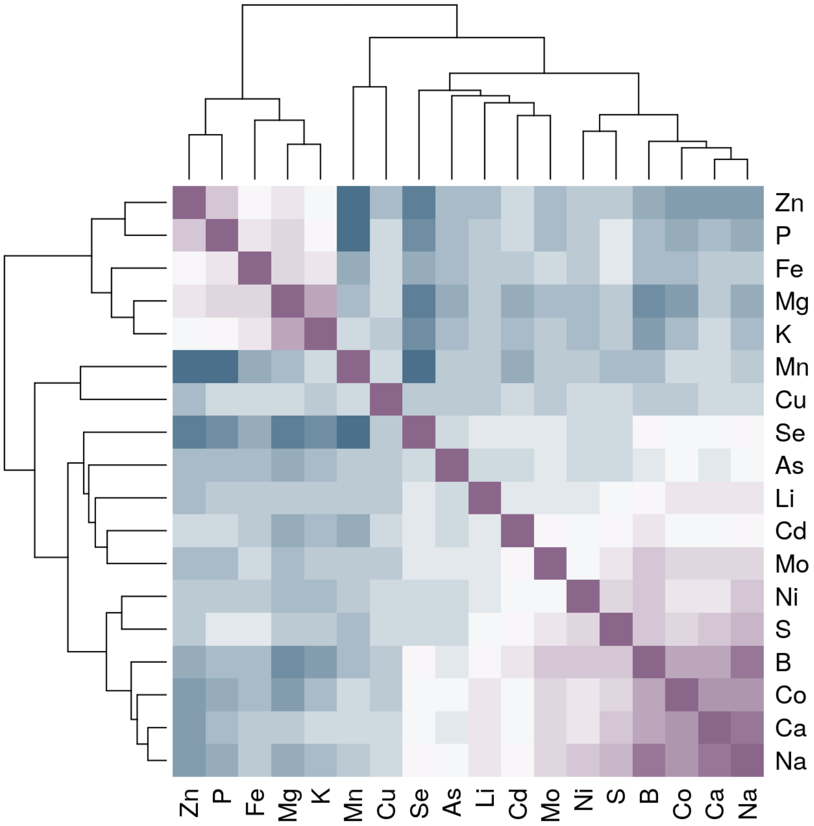

Figure 10: Ion correlation plots on pre-processed data

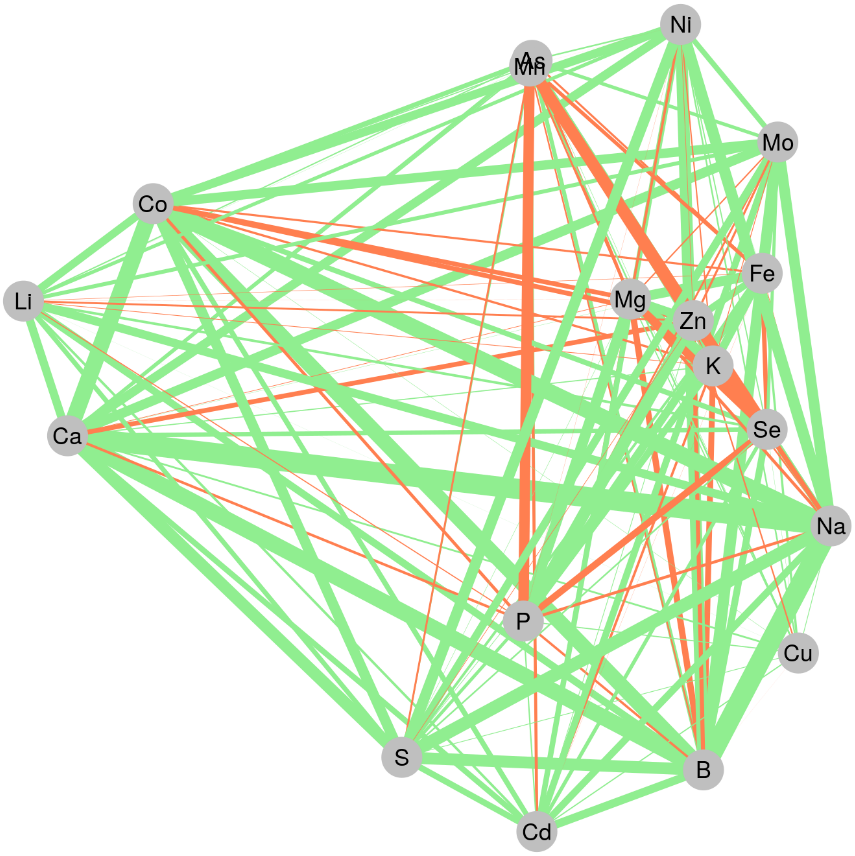

Figure 11: Ion correlation plots on pre-processed data

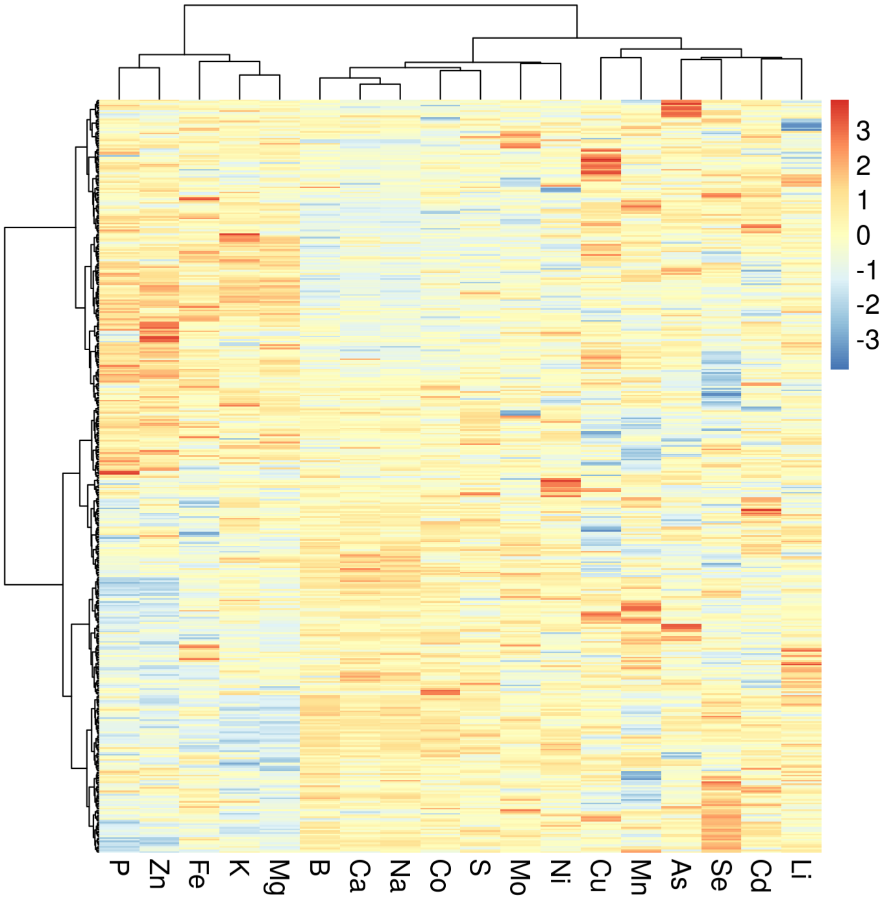

Figure 12: Correlation between ions and genes on pre-processed data