Using *clusterProfiler* to identify and compare functional profiles of gene lists

Guangchuang Yu
School of Biological Sciences
The University of Hong Kong, Hong Kong SAR, China
email: gcyu@connect.hku.hk

November 12, 2013

Contents

1 Introduction

In recently years, high-throughput experimental techniques such as microarray, RNA-Seq and mass spectrometry can detect cellular moleculars at systems-level. These kinds of analyses generate huge quantitaties of data, which need to be

given a biological interpretation. A commonly used approach is via clustering in the gene dimension for grouping different genes based on their similarities [1].

To search for shared functions among genes, a common way is to incorporate the biological knowledge, such as Gene Ontology (GO) and Kyoto Encyclopedia of genes and Genomes (KEGG), for identifying predominant biological themes of a collection of genes.

After clustering analysis, researchers not only want to determine whether there is a common theme of a particular gene cluster, but also to compare the biological themes among gene clusters. The manual step to choose interesting clusters followed by enrichment analysis on each selected cluster is slow and tedious. To bridge this gap, we designed *clusterProfiler* [2], for comparing and visualizing functional profiles among gene clusters.

2 Citation

Please cite the following articles when using *clusterProfiler*.

G Yu, LG Wang, Y Han, QY He. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*. 2012, 16(5), 284-287.

3 Gene Ontology Classification

In *clusterProfiler*, `groupGO` is designed for gene classification based on GO distribution at a specific level.

```
require(DOSE)
data(geneList)
gene <- names(geneList)[abs(geneList) > 2]
head(gene)
```

```
## [1] "4312"  "8318"  "10874" "55143" "55388" "991"
```

```
ggo <- groupGO(gene = gene, organism = "human", ont = "BP",
    level = 3, readable = TRUE)
head(summary(ggo))
```

```
##                    ID                          Description Count
## GO:0019953 GO:0019953               sexual reproduction     9
## GO:0019954 GO:0019954             asexual reproduction      0
## GO:0022414 GO:0022414             reproductive process     19
```

```
## GO:0032504 GO:0032504       multicellular organism reproduction    10
## GO:0032505 GO:0032505  reproduction of a single-celled organism     0
## GO:0048610 GO:0048610 cellular process involved in reproduction     8
##
## GO:0019953                                                          ASPM/CDK1
## GO:0019954
## GO:0022414 ASPM/CDK1/TRIP13/IDO1/CCNB1/CSN3/PTTG1/COL16A1/DACH1/CORIN/GAMT/BMP4/
## GO:0032504                                                        ASPM/TRIP13/C
## GO:0032505
## GO:0048610                                                                CDC20
```

4 Enrichment Analysis

4.1 Hypergeometric model

Enrichment analysis [3] is a widely used approach to identify biological themes. Here we implement hypergeometric model to assess whether the number of selected genes associated with disease is larger than expected.

To determine whether any terms annotate a specified list of genes at frequency greater than that would be expected by chance, *clusterProfiler* calculates a p-value using the hypergeometric distribution:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

In this equation, *N* is the total number of genes in the background distribution, *M* is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, *n* is the size of the list of genes of interest and *k* is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes that have annotation.

P-values were adjusted for multiple comparison, and q-values were also calculated for FDR control.

4.2 GO enrichment analysis

```
ego <- enrichGO(gene = gene, universe = names(geneList),
    organism = "human", ont = "CC", pvalueCutoff = 0.01,
    readable = TRUE)
head(summary(ego))
```

```
##                    ID                          Description GeneRatio
## GO:0005819 GO:0005819                              spindle    23/194
```

```
## GO:0015630 GO:0015630                 microtubule cytoskeleton    37/194
## GO:0000793 GO:0000793                     condensed chromosome    16/194
## GO:0000779 GO:0000779 condensed chromosome, centromeric region    12/194
## GO:0000777 GO:0000777         condensed chromosome kinetochore    11/194
## GO:0044430 GO:0044430                        cytoskeletal part    41/194
##              BgRatio   pvalue p.adjust   qvalue
## GO:0005819 202/11815 2.11e-13 1.71e-11 1.09e-11
## GO:0015630 692/11815 1.29e-10 5.24e-09 3.34e-09
## GO:0000793 136/11815 7.00e-10 1.89e-08 1.20e-08
## GO:0000779  69/11815 1.06e-09 2.15e-08 1.37e-08
## GO:0000777  66/11815 8.61e-09 1.15e-07 7.35e-08
## GO:0044430 960/11815 9.50e-09 1.15e-07 7.35e-08
##
## GO:0005819
## GO:0015630                        KIF20A/TACC3/CENPE/CHEK1/KIF18B/SKA1/TPX2/NCA
## GO:0000793
## GO:0000779
## GO:0000777
## GO:0044430 KIF20A/TACC3/CENPE/CHEK1/KIF18B/SKA1/TPX2/PSD3/KIF4A/ASPM/AK5/BIRC5/K
##            Count
## GO:0005819    23
## GO:0015630    37
## GO:0000793    16
## GO:0000779    12
## GO:0000777    11
## GO:0044430    41
```

4.3 KEGG pathway enrichment analysis

```
kk <- enrichKEGG(gene = gene, organism = "human", pvalueCutoff = 0.01,
    readable = TRUE)
head(summary(kk))
```

```
##                ID                            Description GeneRatio  BgRatio
## hsa04110 hsa04110                              Cell cycle     11/74 128/5894
## hsa04114 hsa04114                          Oocyte meiosis     10/74 114/5894
## hsa03320 hsa03320                 PPAR signaling pathway      7/74  70/5894
## hsa04914 hsa04914 Progesterone-mediated oocyte maturation      6/74  87/5894
## hsa04062 hsa04062             Chemokine signaling pathway      8/74 189/5894
## hsa04060 hsa04060  Cytokine-cytokine receptor interaction      9/74 265/5894
##            pvalue p.adjust   qvalue
## hsa04110 4.31e-07 3.02e-06 4.54e-07
## hsa04114 1.25e-06 4.38e-06 6.59e-07
## hsa03320 2.35e-05 5.49e-05 8.25e-06
## hsa04914 7.21e-04 1.26e-03 1.90e-04
```

```
## hsa04062 2.37e-03 3.32e-03 5.00e-04
## hsa04060 5.58e-03 6.51e-03 9.79e-04
##                                                                 geneID Count
## hsa04110 CDC45/CDC20/CCNB2/CCNA2/CDK1/MAD2L1/TTK/CHEK1/CCNB1/MCM5/PTTG1    11
## hsa04114     CDC20/CCNB2/CDK1/MAD2L1/CALML5/AURKA/CCNB1/PTTG1/ITPR1/PGR    10
## hsa03320                      MMP1/FADS2/ADIPOQ/PCK1/FABP4/HMGCS2/PLIN1     7
## hsa04914                           CCNB2/CCNA2/CDK1/MAD2L1/CCNB1/PGR     6
## hsa04062            CXCL10/CXCL13/CXCL11/CXCL9/CCL18/CCL8/CXCL14/CX3CR1     8
## hsa04060      CXCL10/CXCL13/CXCL11/CXCL9/CCL18/IL1R2/CCL8/CXCL14/CX3CR1     9
```

4.4 DO enrichment analysis

Disease Ontology (DO) enrichment analysis is implemented in *DOSE*, please refer to the package vignettes. The `enrichDO` function is very useful for identifying disease association of interesting genes.

4.5 Reactome pathway enrichment analysis

With the demise of KEGG (at least without subscription), the KEGG pathway data in Bioconductor will not update and we encourage user to analyze pathway using *ReactomePA* which use Reactome as a source of pathway data. The function call of `enrichPathway` in *ReactomePA* is consistent with `enrichKEGG`.

4.6 Function call

The function calls of `groupGO`, `enrichGO`, `enrichKEGG`, `enrichDO` and `enrichPathway` are consistent. The input parameters of *gene* is a vector of entrezgene (for human and mouse) or ORF (for yeast) IDs, and *organism* should be supported species (please refer to the manual of the specific function).

For GO analysis, *ont* must be assigned to one of "BP", "MF", and "CC" for biological process, molecular function and cellular component, respectively. In `groupGO`, the *level* specify the GO level for gene projection.

In enrichment analysis, the *pvalueCutoff* is to restrict the result based on their pvalues and the adjusted p values. *Q-values* were also calculated for controlling false discovery rate (FDR).

The *readable* is a logical parameter to indicate the input gene IDs will map to gene symbols or not.

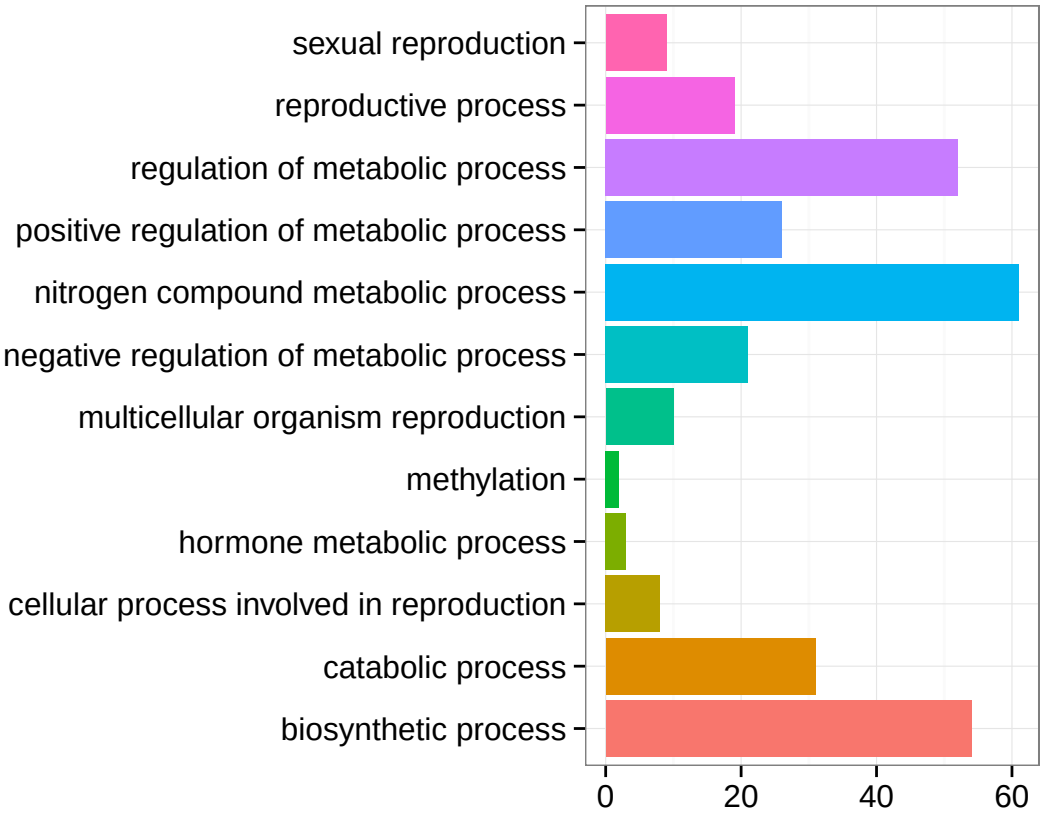
4.7 Visualization

The output of `groupGO`, `enrichGO` and `enrichKEGG` can be visualized by bar plot and category-gene-network plot. It is very common to visualize the enrichment result

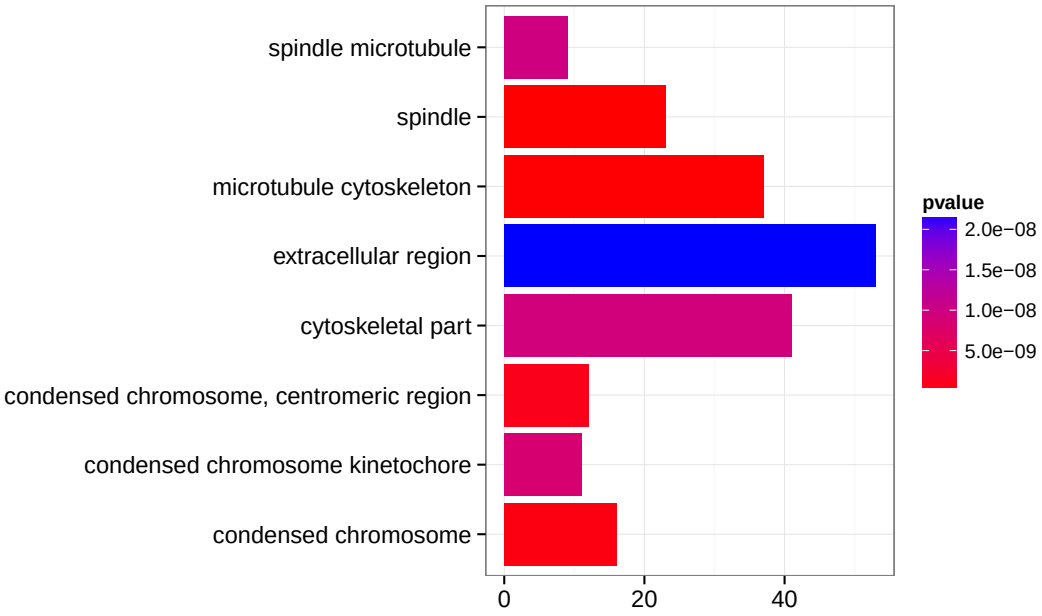
in bar or pie chart. We believe the pie chart is misleading and only provide bar chart.

4.7.1 barplot

```
barplot(ggo, drop = TRUE, showCategory = 12)
```



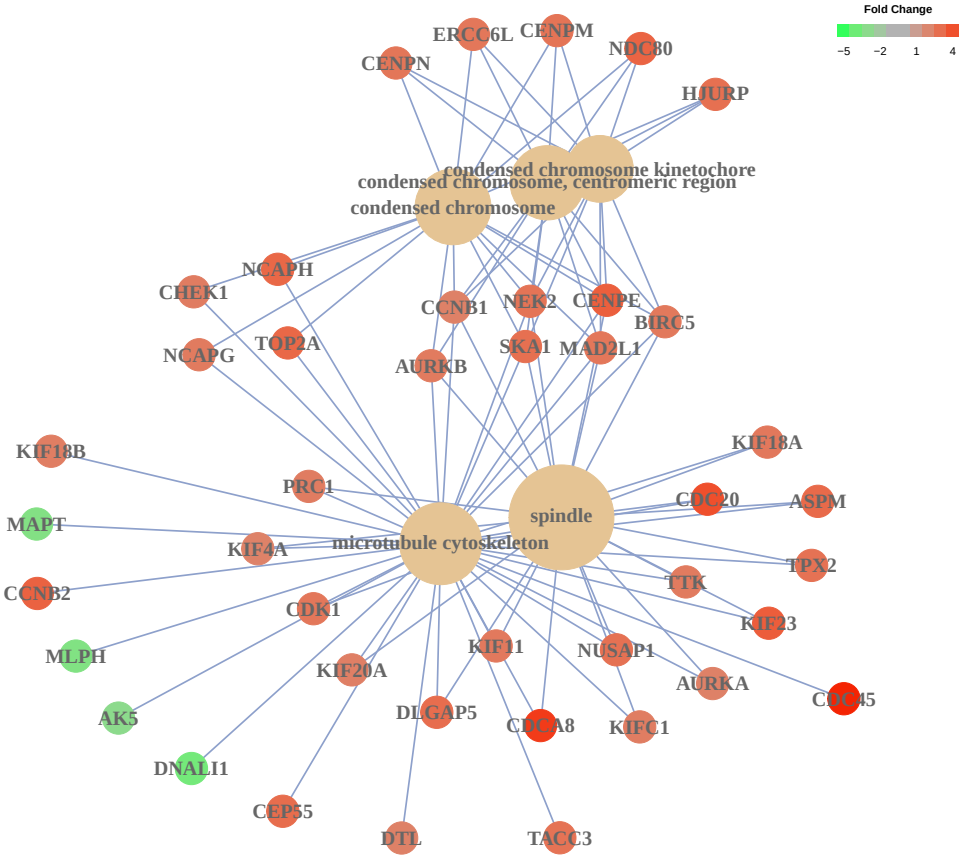

```
barplot(ego, showCategory = 8)
```



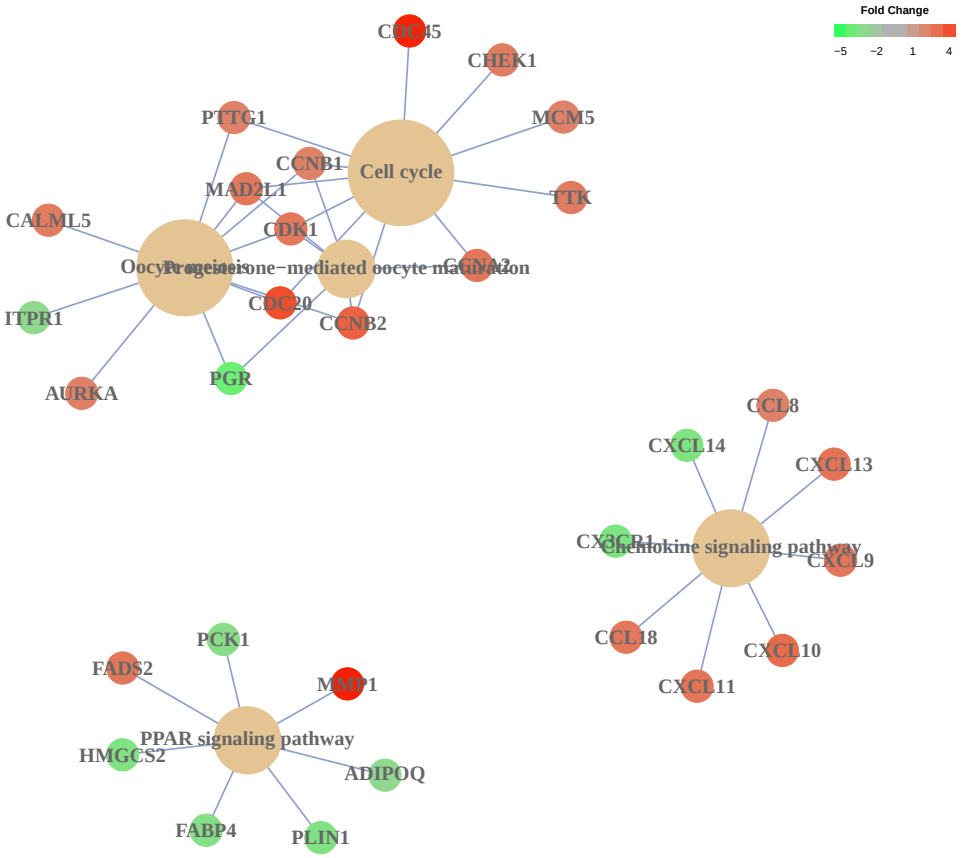
4.7.2 cnetplot

In order to consider the potentially biological complexities in which a gene may belong to multiple annotation categories and provide information of numeric changes if available, we developed `cnetplot` function to extract the complex association.

```
cnetplot(ego, categorySize = "pvalue", foldChange = geneList)
```




```
cnetplot(kk, categorySize = "geneNum", foldChange = geneList)
```



4.7.3 pathview from pathview package

clusterProfiler users can also use `pathview` from the *pathview* [4] to visualize KEGG pathway.

The following example illustrate how to visualize "hsa04110" pathway, which was enriched in our previous analysis.

```
require(pathview)
hsa04110 <- pathview(gene.data = geneList, pathway.id = "hsa04110",
    species = "hsa", limit = list(gene = max(abs(geneList)),
        cpd = 1))

## [1] "Downloading xml files for hsa04110, 1/1 pathways.."
## [1] "Downloading png files for hsa04110, 1/1 pathways.."

## Working in directory /tmp/RtmpIcPvXQ/Rbuild266d61d64c17/clusterProfiler/vignet
## Writing image file hsa04110.pathview.png
```

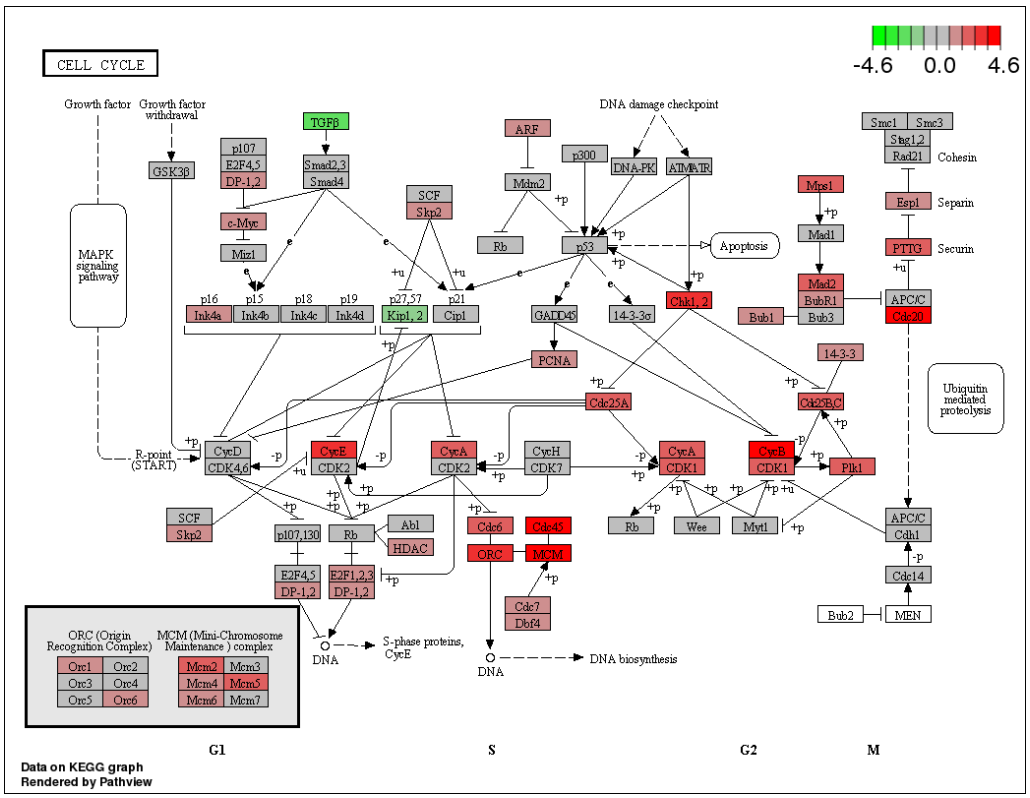


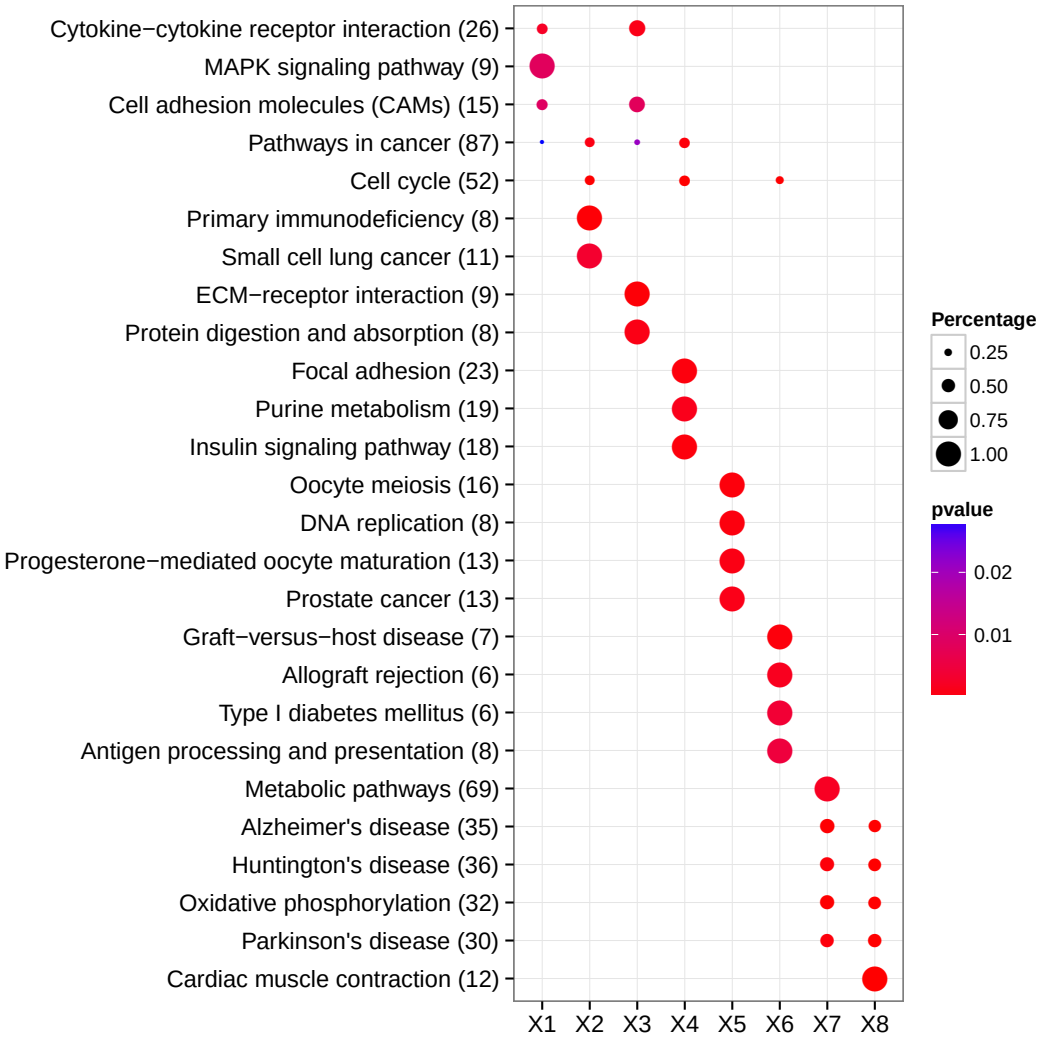
Figure 1: visualize KEGG pathway using pathview

For further information, please refer to the vignette of *pathview* [4].

5 Biological theme comparison

clusterProfiler was developed for biological theme comparison, and it provides a function, `compareCluster`, to automatically calculate enriched functional categories of each gene clusters.

```
data(gcSample)
ck <- compareCluster(geneCluster = gcSample, fun = "enrichKEGG")
plot(ck)
```



By default, only top 5 (most significant) categories of each cluster was plotted. User can changes the parameter *showCategory* to specify how many categories of each cluster to be plotted, and if *showCategory* was set to *NULL*, the whole result will be plotted.

The dot sizes were based on their corresponding row percentage by default, and user can set the parameter *by* to ”count” to make the comparison based on gene counts. We choose ”percentage” as default parameter to represent the size of dots, since some categories may contain a large number of genes, and make the dot sizes of those small categories too small to compare. To provide the full information, we also provide number of identified genes in each category (numbers in parentheses), as shown in Figure 3. If the dot sizes were based on ”count”, the row numbers will not shown.

The p-values indicate that which categories are more likely to have biological meanings. The dots in the plot are color-coded based on their corresponding p-values. Color gradient ranging from red to blue correspond to in order of increasing p-values. That is, red indicate low p-values (high enrichment), and blue indicate high p-values (low enrichment). P-values and adjusted p-values were

filtered out by the threshold giving by parameter *pvalueCutoff*, and FDR can be estimated by *qvalue*.

User can refer to the example in [2]; we analyzed the publicly available expression dataset of breast tumour tissues from 200 patients (GSE11121, Gene Expression Omnibus) [5]. We identified 8 gene clusters from differentially expressed genes, and using `compareCluster` to compare these gene clusters by their enriched biological process.

Another example was shown in [6], we calculated functional similarities among viral miRNAs using method described in [7], and compared significant KEGG pathways regulated by different viruses using `compareCluster`.

The comparison function was designed as a general-package for comparing gene clusters of any kind of ontology associations, not only `groupGO`, `enrichGO`, and `enrichKEGG` this package provided, but also other biological and biomedical ontologies, for instance, `enrichDO` from *DOSE* and `enrichPathway` from *ReactomePA* work fine with `compareCluster` for comparing biological themes in disease and reactome pathway perspective. More details can be found in the vignettes of *DOSE* and *ReactomePA*.

6 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 3.0.2 (2013-09-25), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationDbi˜1.24.0, Biobase˜2.22.0, BiocGenerics˜0.8.0, DBI˜0.2-7, DOSE˜2.0.0, GO.db˜2.10.1, KEGGgraph˜1.20.0, RSQLite˜0.11.4, XML˜3.98-1.1, clusterProfiler˜1.11.1, ggplot2˜0.9.3.1, graph˜1.40.0, knitr˜1.5, org.Hs.eg.db˜2.10.1, pathview˜1.2.0
- Loaded via a namespace (and not attached): Biostrings˜2.30.0, DO.db˜2.7, GOSemSim˜1.20.0, IRanges˜1.20.5, KEGG.db˜2.10.1, KEGGREST˜1.2.0, MASS˜7.3-29, RColorBrewer˜1.0-5, RCurl˜1.95-4.1, Rgraphviz˜2.6.0, XVector˜0.2.0, codetools˜0.2-8, colorspace˜1.2-4, dichromat˜2.0-0, digest˜0.6.3, evaluate˜0.5.1, formatR˜0.10, grid˜3.0.2, gtable˜0.1.2, highr˜0.3, httr˜0.2, igraph˜0.6.6, labeling˜0.2, munsell˜0.4.2, plyr˜1.8, png˜0.1-6, proto˜0.3-10, qvalue˜1.36.0, reshape2˜1.2.2, scales˜0.2.3, stats4˜3.0.2, stringr˜0.6.2, tcltk˜3.0.2, tools˜3.0.2

References

[1] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010. PMID: 20179076.

[2] Guangchuang Yu, Le-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, May 2012.

[3] Elizabeth˜I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J˜Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18):3710–3715, December 2004. PMID: 15297299.

[4] Weijun Luo and Cory Brouwer. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. 29:1830–1831. PMID: 23740750.

[5] Marcus Schmidt, Daniel B?hm, Christian von T?rne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan˜G. Hengstler, Heinz K?lbl, and Mathias Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13):5405 –5413, July 2008.

[6] Guangchuang Yu and Qing-Yu He. Functional similarity analysis of human virus-encoded miRNAs. *Journal of Clinical Bioinformatics*, 1(1):15, May 2011.

[7] Guangchuang Yu, Chuan-Le Xiao, Xiaochen Bo, Chun-Hua Lu, Yide Qin, Sheng Zhan, and Qing-Yu He. A new method for measuring functional similarity of microRNAs. *Journal of Integrated OMICS*, 1(1):49–54, February 2011.