Using clusterProfiler to identify and compare functional profiles of gene lists

Guangchuang Yu
College of Life Science and Technology
Jinan University, Guangzhou, China
email: guangchuangyu@gmail.com

June 14, 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 analysis 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    23
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
GO:0019954
GO:0022414 MMP1/CDC20/TOP2A/ASPM/CDK1/TRIP13/IDO1/CCNB1/CSN3/PTTG1/COL16A1/DACH1/CORI
```

```
GO:0032504
GO:0032505
GO:0048610
```

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    23/195
GO:0015630 GO:0015630                 microtubule cytoskeleton    37/195
GO:0000793 GO:0000793                     condensed chromosome    16/195
GO:0000779 GO:0000779 condensed chromosome, centromeric region    12/195
GO:0044430 GO:0044430                        cytoskeletal part    41/195
GO:0005876 GO:0005876                      spindle microtubule     9/195
             BgRatio   pvalue p.adjust   qvalue
GO:0005819 198/11807 1.56e-13 1.26e-11 7.37e-12
GO:0015630 666/11807 5.13e-11 2.08e-09 1.22e-09
```

```
GO:0000793 136/11807 7.63e-10 2.06e-08 1.20e-08
GO:0000779  69/11807 1.14e-09 2.30e-08 1.35e-08
GO:0044430 931/11807 4.67e-09 7.56e-08 4.42e-08
GO:0005876  37/11807 6.34e-09 8.56e-08 5.01e-08

GO:0005819
GO:0015630                         KIF20A/TACC3/CENPE/CHEK1/KIF18B/SKA1/TPX2/NCAPH/KI
GO:0000793
GO:0000779
GO:0044430 KIF20A/TACC3/CENPE/CHEK1/KIF18B/SKA1/TPX2/PSD3/KIF4A/ASPM/AK5/BIRC5/KIF11/
GO:0005876
           Count
GO:0005819    23
GO:0015630    37
GO:0000793    16
GO:0000779    12
GO:0044430    41
GO:0005876     9
```

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 similar. 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)
```

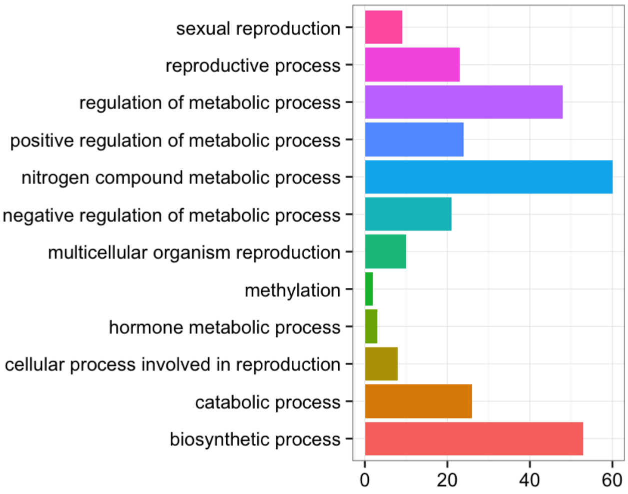

Figure 1: barplot of GO classification Result

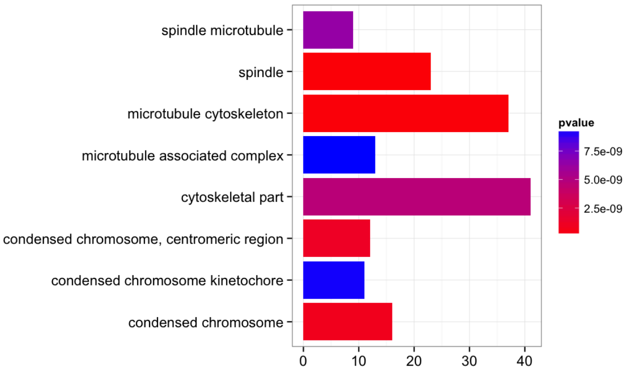

Figure 2: barplot of GO enrichment Result

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 viewKEGG

We developed `viewKEGG`, which extend functions of *Pathview* to support output of `enrichKEGG`, to automate visualize KEGG pathway.

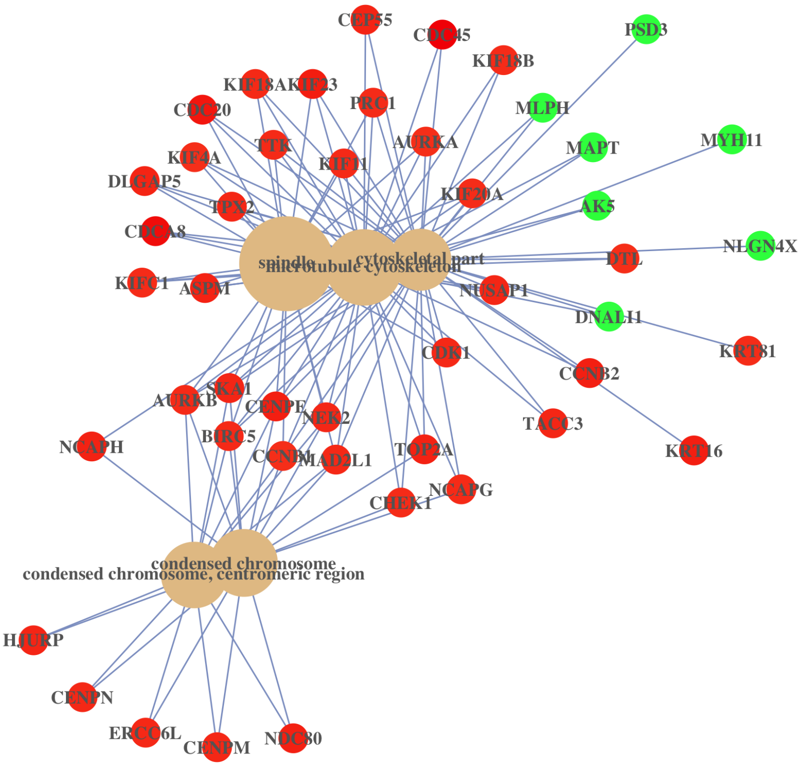

Figure 3: cnetplot of GO enrichment result

```
pv.res <- viewKEGG(kk, pathwayID="hsa04110", foldChange=geneList, kegg.native=TRUE)
pv.res <- viewKEGG(kk, pathwayID=1, foldChange=geneList, kegg.native=FALSE)
```

The parameter *pathwayID* can be numeric vector or character vector. If *pathwayID* is numeric value, `viewKEGG` will use it as index of *KEGG enrichment result* and convert it to corresponding pathway ID. The *pathwayID* can also set to "all", and all the maps of significant pathways will be generated.

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)
```

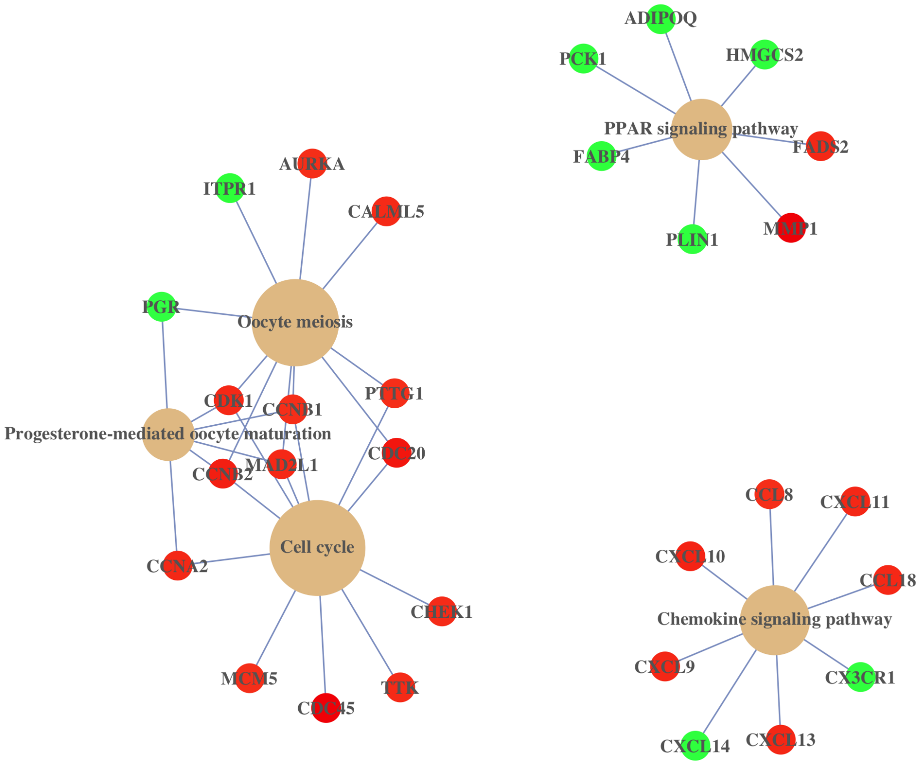

Figure 4: cnetplot of KEGG enrichment result

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*.

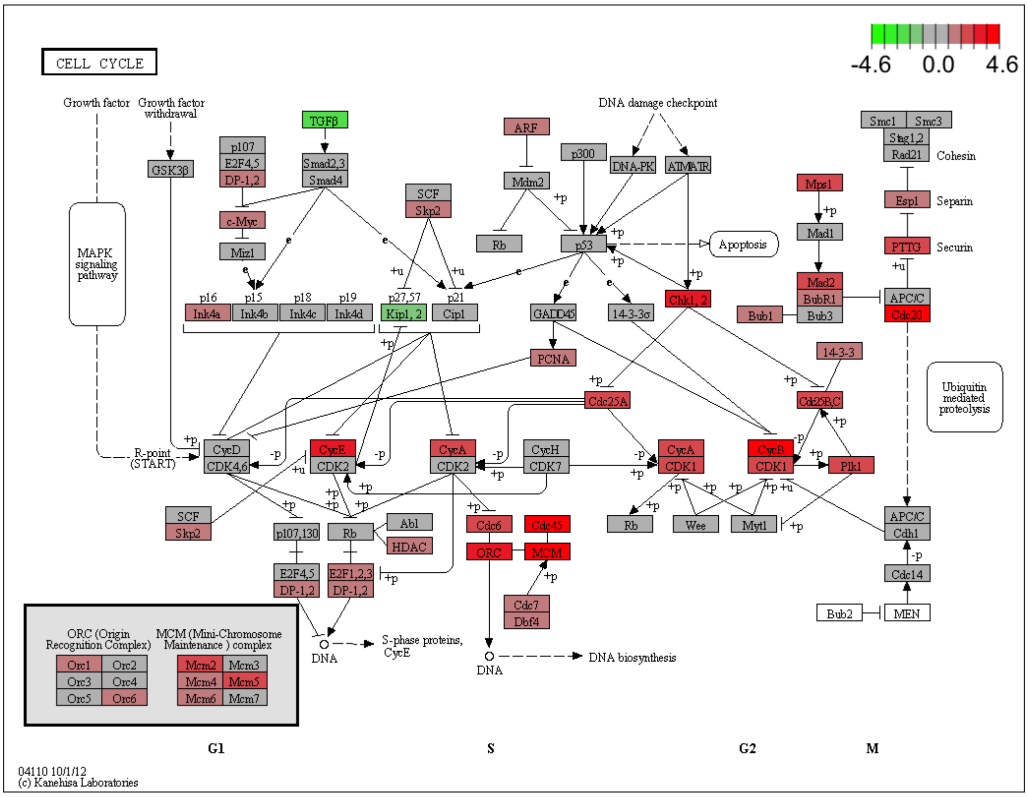

Figure 5: viewKEGG in KEGG view

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) [?]. 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 [?], we calculated functional similarities among viral miRNAs using method described in [?], 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 3.0.1 (2013-05-16), x86_64-apple-darwin10.8.0
- Locale: C/UTF-8/C/C/C/C

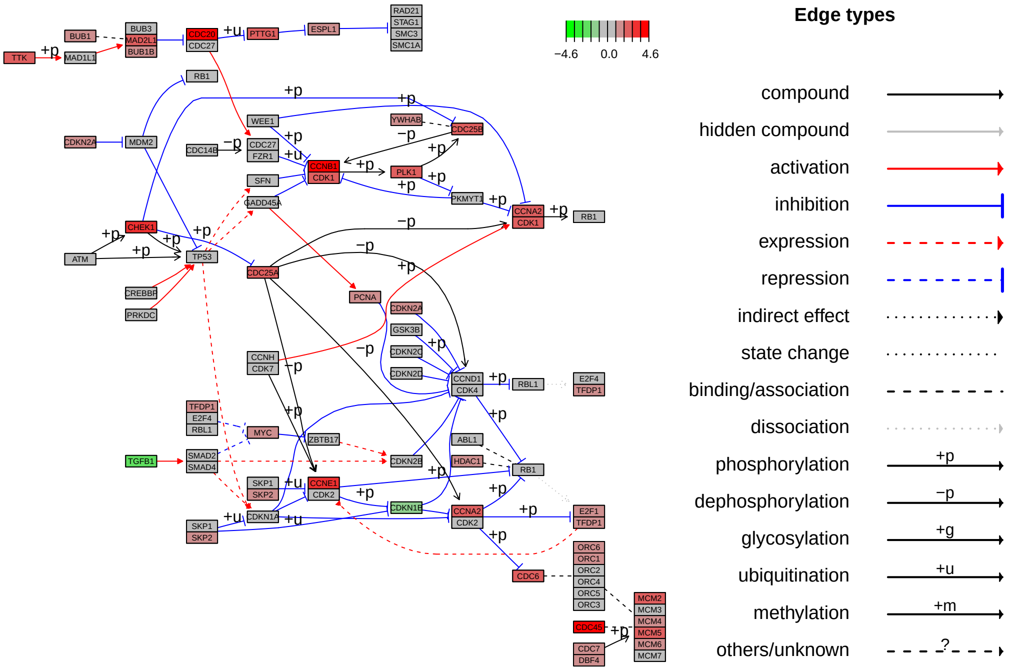

Figure 6: viewKEGG in Graphviz view

- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationDbi 1.22.6, Biobase 2.20.0, BiocGenerics 0.6.0, DBI 0.2-7, DOSE 1.99.0, GO.db 2.9.0, KEGGgraph 1.16.0, RSQLite 0.11.4, XML 3.95-0.2, cacheSweave 0.6-1, clusterProfiler 1.9.1, filehash 2.2-1, ggplot2 0.9.3.1, graph 1.38.2, org.Hs.eg.db 2.9.0, pathview 1.1.2, stashR 0.3-5
- Loaded via a namespace (and not attached): DO.db 2.6.0, GOSemSim 1.19.0, IRanges 1.18.1, KEGG.db 2.9.1, MASS 7.3-26, RColorBrewer 1.0-5, Rgraphviz 2.4.0, colorspace 1.2-2, dichromat 2.0-0, digest 0.6.3, grid 3.0.1, gtable 0.1.2, igraph 0.6.5-2, labeling 0.1, munsell 0.4, plyr 1.8, png 0.1-5, proto 0.3-10, qvalue 1.34.0, reshape2 1.2.2, scales 0.2.3, stats4 3.0.1, stringr 0.6.2, tcltk 3.0.1, tools 3.0.1

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

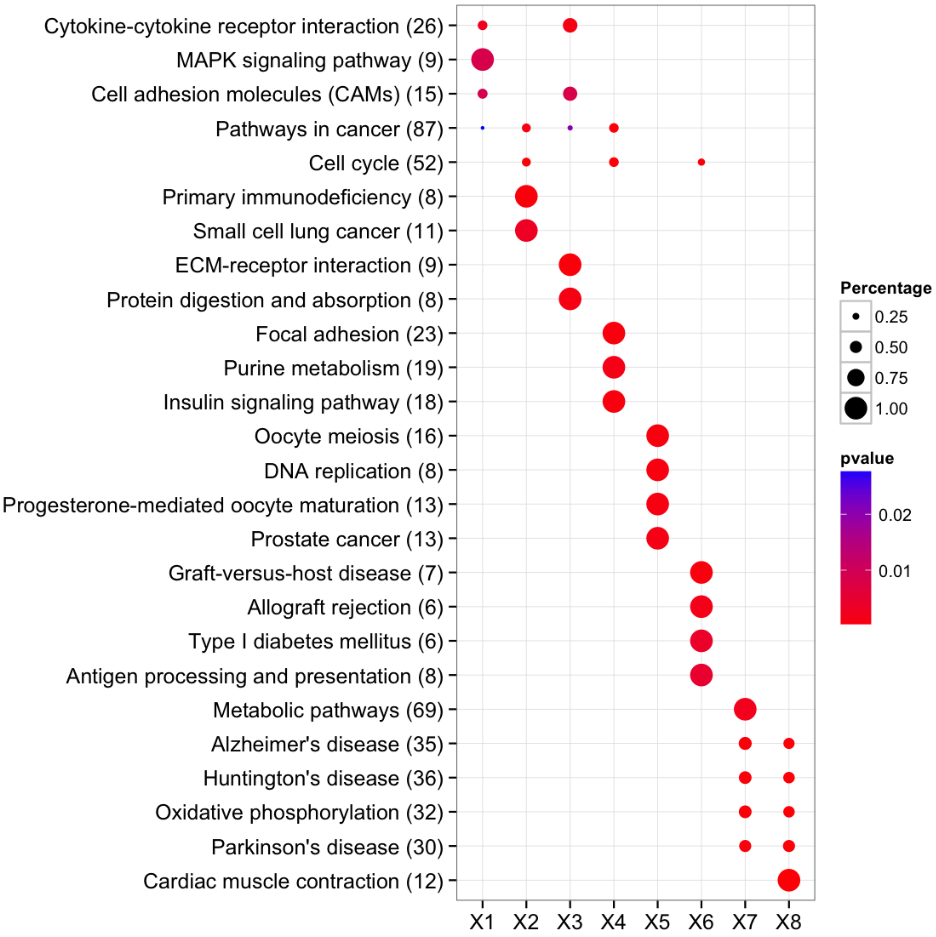

Figure 7: Comparison of KEGG enrichment of gene clusters

associated with a list of genes. *Bioinformatics (Oxford, England)*, 20(18):3710–3715, December 2004. PMID: 15297299.