Biological Theme Comparison

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

March 30, 2012

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(Yu et al., 2010).
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*, for comparing and visulizing 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), in press.

3 Functional Profiles

In *clusterProfiler*, we implemented three functions to explore the functional profiles of a collection of genes.

- groupGO for gene classification based on GO distribution at a specific level

```
> data(gcSample)
> x <- groupGO(gene=gcSample[[1]],
+              organism="human",
+              ont="CC",
+              level=2,
+              readable=TRUE)
> head(summary(x))
```

```
                   ID             Description Count
GO:0016020 GO:0016020                membrane     6
GO:0005576 GO:0005576    extracellular region     1
GO:0005623 GO:0005623                    cell    13
GO:0019012 GO:0019012                  virion     0
GO:0030054 GO:0030054           cell junction     1
GO:0031974 GO:0031974 membrane-enclosed lumen     7
                                                                                ge
GO:0016020 SDF2L1/ERGIC1/PA2G4/PEBP1/RAD50/RUVBL2/RPS23/LONP1/CYC1/IARS/RPL4/MCM6/
GO:0005576                                       SDF2L1/ERGIC1/PEBP1/RUVBL2/CYC1/
GO:0005623                                                                       P
GO:0019012
GO:0030054
GO:0031974                                  SDF2L1/PA2G4/RAD50/RUVBL2/LONP1/RPL4/
```

- `enrichGO` for GO enrichment analysis

```
> y <- enrichGO(gene=gcSample[[2]],
+               organism="human",
+               ont="MF",
+               pvalueCutoff=0.01,
+               qvalueCutoff=0.05,
+               readable=TRUE)
> head(summary(y))
```

```
                   ID
GO:0003924 GO:0003924
GO:0008135 GO:0008135
GO:0003746 GO:0003746
GO:0000166 GO:0000166
GO:0036094 GO:0036094
GO:0005525 GO:0005525
                                                 Description
GO:0003924                                   GTPase activity
GO:0008135 translation factor activity, nucleic acid binding
GO:0003746            translation elongation factor activity
GO:0000166                                nucleotide binding
GO:0036094                            small molecule binding
GO:0005525                                       GTP binding
           GeneRatio    BgRatio       pvalue      qvalue
GO:0003924      4/18  231/17980 7.050766e-05 0.002675400
GO:0008135      3/18   87/17980 8.472100e-05 0.002675400
GO:0003746      2/18   20/17980 1.779433e-04 0.003040629
GO:0000166      9/18 2379/17980 1.925732e-04 0.003040629
GO:0036094      9/18 2547/17980 3.271587e-04 0.004097529
GO:0005525      4/18  377/17980 4.611144e-04 0.004097529
                                    geneID Count
GO:0003924 EEF1A2/EEF2/EIF4A1/RAB5A/EFTUD2     4
GO:0008135                                     3
GO:0003746        EEF1A2/EEF2/RAB5A/EFTUD2     2
GO:0000166        EEF1A2/EEF2/RAB5A/EFTUD2     9
```

```
GO:0036094                                       9
GO:0005525                                       4
```

- `enrichKEGG` for KEGG pathway enrichment analysis.

```
> z <- enrichKEGG(gene=gcSample[[3]],
+                 organism="human",
+                 pvalueCutoff=0.05,
+                 qvalueCutoff=0.05,
+                 readable=TRUE)
> head(summary(z))
```

```
               ID                                 Description
hsa05130 hsa05130    Pathogenic Escherichia coli infection
hsa04145 hsa04145                                   Phagosome
hsa04540 hsa04540                                Gap junction
hsa04962 hsa04962 Vasopressin-regulated water reabsorption
         GeneRatio  BgRatio       pvalue       qvalue
hsa05130      4/17  58/5894 1.826892e-05 0.0002115348
hsa04145      5/17 156/5894 5.827611e-05 0.0003373880
hsa04540      4/17  90/5894 1.039489e-04 0.0004012064
hsa04962      2/17  44/5894 6.898981e-03 0.0199707355
         geneID Count
hsa05130           4
hsa04145           5
hsa04540           4
hsa04962           2
```

With the demise of KEGG (at least without subscription), the pathway data used in *clusterProfiler* will not update, and we encourage user to use `enrichPathway` in Bioconductor package *ReactomePA*, which use Reactome as a source of pathway data.

The function calls of `groupGO`, `enrichGO` and `enrichKEGG` are similar. The input parameters of *gene* is a vector of entrezgene (for human and mouse) or ORF (for yeast) IDs, and *organism* must be one of "human", "mouse", and "yeast", according to the gene IDs.

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 *qvalueCutoff* is to control false discovery rate (FDR) to prevent high FDR in multiple testing. The *readable* is a logical parameter to indicate the input gene IDs will map to gene symbols or not.

4 Biological theme comparison

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

As we demonstrated in Yu et al. (2012), we analyzed the publicly available expression dataset of breast tumour tissues from 200 patients (GSE11121, Gene Expression Omnibus) (Schmidt et al., 2008). We identified 8 gene clusters from differentially expressed genes, and using `compareCluster` to compare these gene clusters by their enriched biological process, with the strict cutoff of p-values < 0.01 and q-values < 0.05. The analysis result was illustrated in Figure 1. More details of this analysis are described in Yu et al. (2012).

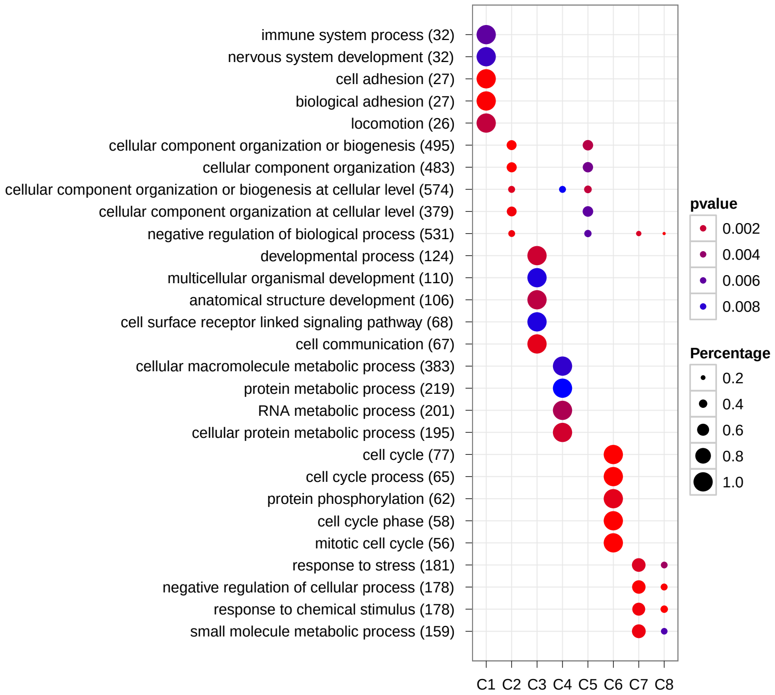

Figure 1: Comparison of GO enrichment of gene clusters

Another example was shown in Yu and He (2011), we calculated functional similarities among viral miRNAs using method described in Yu et al. (2011), and compared significant KEGG pathways regulated by different viruses using *clusterProfiler*.

The comparison function was designed as a general-package for comparing gene clusters of any kind of gene-ontology associations, not only GO and KEGG this package provided, but also other biological and biomedical ontologies.

For example, `compareCluster` can cooperate seamless with *DOSE* and *ReactomePA* and compare gene cluster in the context of disease and reactome pathway as demonstrated in the online vignette of *DOSE* and *ReactomePA* respectively.

5 Visualization

clusterProfiler implemented serveral methods for visualizing analyzed result.

- Bar Plot

  Bar plot was used to visualized functional profile of the given collection of genes.

  The plot function call was consistent for analysis results generated by `groupGO`, `enrichGO` and `enrichKEGG`.

  Users can try the following command:

```
> plot(x, type="bar", order=FALSE, drop=TRUE)
> plot(z, type="bar", font.size=12)
> plot(y, type="bar", title="MF Enrichment analysis", showCategory=10)
```

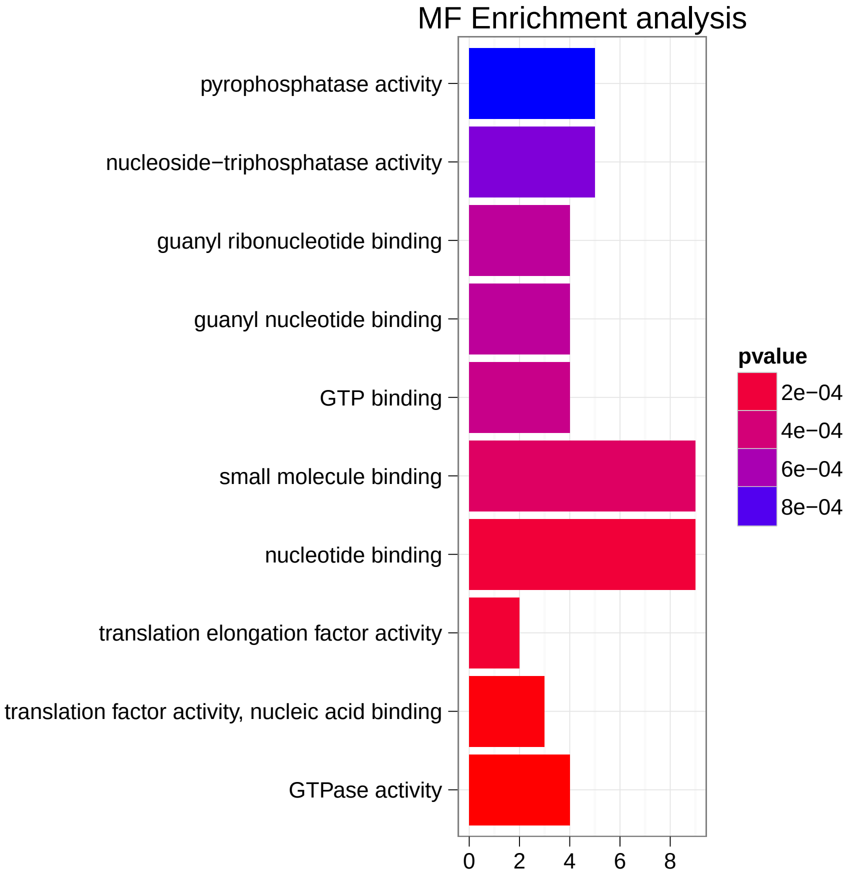

Figure 2: Example of plotting functional profiles

- Category Net Plot

  Category-gene network model was also implmented to extract the complex relationships between genes and associated categories. It provides a high-level model to understand the functionalities of genes.

  The plot function call was consistent for analysis results generated by `groupGO`, `enrichGO` and `enrichKEGG`. Users can try the following command:

```
> plot(y, type="cnet", categorySize="geneNum")
> plot(z, type="cnet", categorySize="pvalue", output="interactive")
```

- Dot Plot

  Dot plot was implemented for cluster comparison as shown in Figure 1. Here, we demonstrated the functional call of `compareCluster`.

```
> xx <- compareCluster(gcSample,
+                      fun="enrichGO",
> plot(x, type="cnet", showCategory=5, output="fixed")
```

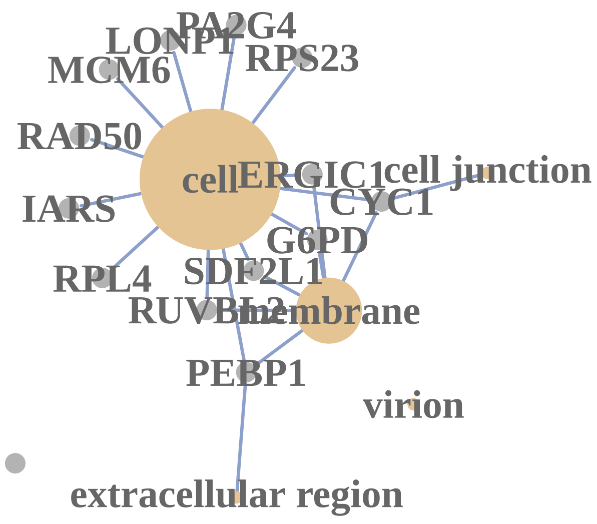

Figure 3: Example of plotting GO profiles using cnetplot

```
+                            ont="CC",
+                            organism="human",
+                            pvalueCutoff=0.05,
+                            qvalueCutoff=0.05)
> plot(xx)
```

Bar plot was also supported to visualize cluster comparision. User can try the following command to explore the usage:

```
> plot(xx, type="bar", by="percentage")
> plot(xx, type="bar", by="count")
```

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

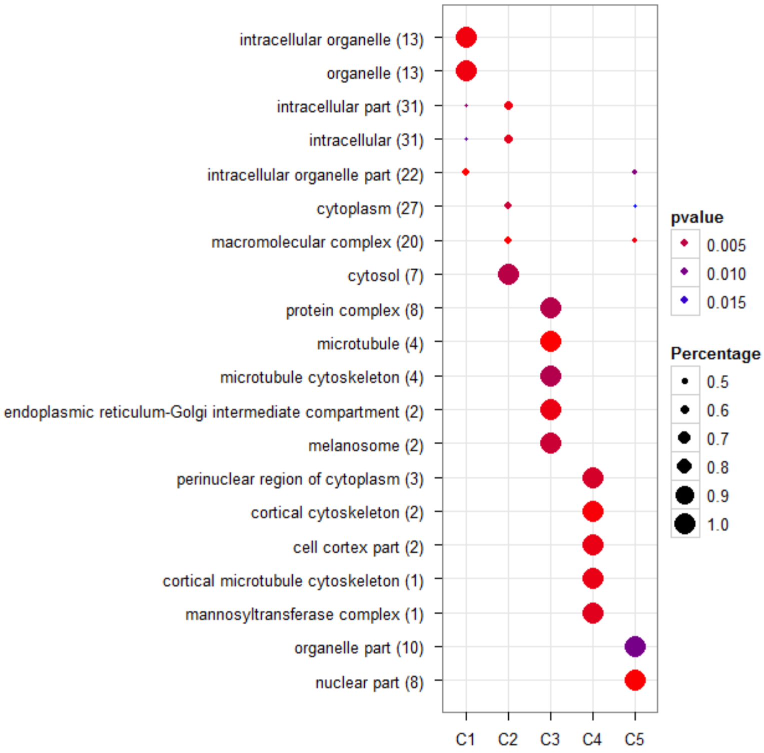

Figure 4: GO Enrichment Comparison

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 were filtered out by the threshold giving by parameter *pvalueCutoff*, and FDR was control by parameter *qvalueCutoff*.

`compareCluster` was designed as a general function for comparing gene clusters of any kind of gene-ontology associations, not only GO (`groupGO` and `enrichGO`) and KEGG (`enrichKEGG`) provided in this package, but also other biological or biomedical ontologies, including Disease Ontology (via `enrichDO` in *DOSE*) and Reactome Pathway (via `enrichPathway` in *ReactomePA*). 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 2.15.0 (2012-03-30)
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=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
[1] GO.db_2.7.1           org.Hs.eg.db_2.7.1
[3] clusterProfiler_1.4.0 AnnotationDbi_1.18.0
[5] Biobase_2.16.0        BiocGenerics_0.2.0
[7] RSQLite_0.11.1        DBI_0.2-5

loaded via a namespace (and not attached):
 [1] DO.db_2.4          DOSE_1.2.0
 [3] IRanges_1.14.0     KEGG.db_2.7.1
 [5] MASS_7.3-17        RColorBrewer_1.0-5
 [7] colorspace_1.1-1   dichromat_1.2-4
 [9] digest_0.5.2       ggplot2_0.9.0
[11] grid_2.15.0        igraph_0.5.5-4
[13] memoise_0.1        munsell_0.3
[15] plyr_1.7.1         proto_0.3-9.2
[17] qvalue_1.30.0      reshape2_1.2.1
[19] scales_0.2.0       stats4_2.15.0
[21] stringr_0.6        tcltk_2.15.0
[23] tools_2.15.0
```

References

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. doi: 10.1158/0008-5472.CAN-07-5206. URL http://cancerres.aacrjournals.org/content/68/13/5405.abstract.

Guangchuang Yu and Qing-Yu He. Functional similarity analysis of human virus-encoded miRNAs. *Journal of Clinical Bioinformatics*, 1(1):15, May 2011. ISSN 2043-9113. doi: 10.1186/2043-9113-1-15. URL http://www.jclinbioinformatics.com/content/1/1/15.

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*, 26:976–978, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq064. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/7/976. PMID: 20179076.

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. ISSN 2182-0287. doi: 10.5584/jiomics.v1i1.21. URL `http://www.jiomics.com/index.php/jio/article/view/21`.

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:in press, 2012. ISSN 1536-2310.