Biological Theme Comparison

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

March 1, 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=T
> summary(x)
```

```
                 GOID               Description Count
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
GO:0032991 GO:0032991    macromolecular complex     6
GO:0043226 GO:0043226                 organelle    13
GO:0044421 GO:0044421 extracellular region part     1
GO:0044422 GO:0044422            organelle part    12
GO:0044423 GO:0044423               virion part     0
GO:0044456 GO:0044456              synapse part     1
GO:0044464 GO:0044464                 cell part    13
GO:0045202 GO:0045202                   synapse     1
GO:0055044 GO:0055044                  symplast     0
                                                                                Ge
GO:0005576                                                                       P
GO:0005623 SDF2L1/ERGIC1/PA2G4/PEBP1/RAD50/RUVBL2/RPS23/LONP1/CYC1/IARS/RPL4/MCM6/
GO:0019012
GO:0030054                                                                       
GO:0031974                                      SDF2L1/PA2G4/RAD50/RUVBL2/LONP1/RPL4/
GO:0032991                                             PA2G4/RAD50/RUVBL2/RPS23/RPL4/
GO:0043226 SDF2L1/ERGIC1/PA2G4/PEBP1/RAD50/RUVBL2/RPS23/LONP1/CYC1/IARS/RPL4/MCM6/
GO:0044421                                                                       P
GO:0044422      SDF2L1/ERGIC1/PA2G4/PEBP1/RAD50/RUVBL2/RPS23/LONP1/CYC1/RPL4/MCM6/
GO:0044423
GO:0044456                                                                       P
GO:0044464 SDF2L1/ERGIC1/PA2G4/PEBP1/RAD50/RUVBL2/RPS23/LONP1/CYC1/IARS/RPL4/MCM6/
GO:0045202                                                                       P
GO:0055044
```

- `enrichGO` for GO enrichment analysis

```
> y <- enrichGO(gene=gcSample[[2]], organism="human", ont="MF", pvalueCutoff=0.01,
> summary(y)
```

```
                GOID
GO:0003924 GO:0003924
GO:0008135 GO:0008135
GO:0003746 GO:0003746
GO:0000166 GO:0000166
GO:0005525 GO:0005525
GO:0019001 GO:0019001
GO:0032561 GO:0032561
GO:0004757 GO:0004757
GO:0016768 GO:0016768
GO:0017111 GO:0017111
GO:0016462 GO:0016462
GO:0016818 GO:0016818
GO:0016817 GO:0016817
GO:0003723 GO:0003723
GO:0004766 GO:0004766
GO:0035639 GO:0035639
GO:0032555 GO:0032555
GO:0032553 GO:0032553
GO:0017076 GO:0017076
```

GO:0030292 GO:0030292
GO:0017016 GO:0017016
GO:0030274 GO:0030274
GO:0031267 GO:0031267
GO:0000339 GO:0000339

GO:0003924 GTPa
GO:0008135 translation factor activity, nucleic a
GO:0003746 translation elongation fact
GO:0000166 nucleot
GO:0005525
GO:0019001 guanyl nucleot
GO:0032561 guanyl ribonucleot
GO:0004757 sepiapterin reducta
GO:0016768 spermine syntha
GO:0017111 nucleoside-triphosphata
GO:0016462 pyrophosphata
GO:0016818 hydrolase activity, acting on acid anhydrides, in phosphorus-containing
GO:0016817 hydrolase activity, acting on acid
GO:0003723
GO:0004766 spermidine syntha
GO:0035639 purine ribonucleoside triphospha
GO:0032555 purine ribonucleot
GO:0032553 ribonucleot
GO:0017076 purine nucleot
GO:0030292 protein tyrosine kinase inhibit
GO:0017016 Ras GTP
GO:0030274 LIM dom
GO:0031267 small GTP
GO:0000339 RNA

	GeneRatio	BgRatio	pvalue	qvalue
GO:0003924	4/18	230/15190	0.0001326109	0.004907085
GO:0008135	3/18	87/15190	0.0001391562	0.004907085
GO:0003746	2/18	20/15190	0.0002488272	0.005849621
GO:0000166	9/18	2270/15190	0.0004925813	0.008684986
GO:0005525	4/18	375/15190	0.0008500242	0.009285888
GO:0019001	4/18	388/15190	0.0009653090	0.009285888
GO:0032561	4/18	388/15190	0.0009653090	0.009285888
GO:0004757	1/18	1/15190	0.0011849901	0.009285888
GO:0016768	1/18	1/15190	0.0011849901	0.009285888
GO:0017111	5/18	757/15190	0.0015098609	0.009958066
GO:0016462	5/18	784/15190	0.0017645930	0.009958066
GO:0016818	5/18	787/15190	0.0017947502	0.009958066
GO:0016817	5/18	791/15190	0.0018355540	0.009958066
GO:0003723	5/18	831/15190	0.0022824963	0.011136829
GO:0004766	1/18	2/15190	0.0023686540	0.011136829
GO:0035639	7/18	1833/15190	0.0034980320	0.014782893
GO:0032555	7/18	1862/15190	0.0038273074	0.014782893
GO:0032553	7/18	1863/15190	0.0038390788	0.014782893
GO:0017076	7/18	1875/15190	0.0039825554	0.014782893
GO:0030292	1/18	4/15190	0.0047320084	0.016686556

```
GO:0017016      2/18  106/15190 0.0068617051 0.023044323
GO:0030274      1/18    7/15190 0.0082671279 0.025881102
GO:0031267      2/18  118/15190 0.0084403296 0.025881102
GO:0000339      1/18    8/15190 0.0094428634 0.027748765
                                                      geneID
GO:0003924                                 RAB5A/EEF2/EFTUD2/EEF1A2
GO:0008135                                       EIF4A1/EEF2/EEF1A2
GO:0003746                                             EEF2/EEF1A2
GO:0000166 SNRPB2/CCT2/NDUFA10/SPR/RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0005525                                 RAB5A/EEF2/EFTUD2/EEF1A2
GO:0019001                                 RAB5A/EEF2/EFTUD2/EEF1A2
GO:0032561                                 RAB5A/EEF2/EFTUD2/EEF1A2
GO:0004757                                                      SPR
GO:0016768                                                      SMS
GO:0017111                          RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0016462                          RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0016818                          RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0016817                          RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0003723                          SNRPB2/SF3A1/EIF4A1/EEF2/EEF1A2
GO:0004766                                                      SMS
GO:0035639            CCT2/NDUFA10/RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0032555            CCT2/NDUFA10/RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0032553            CCT2/NDUFA10/RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0017076            CCT2/NDUFA10/RAB5A/EIF4A1/EEF2/EFTUD2/EEF1A2
GO:0030292                                                   GNB2L1
GO:0017016                                                IPO5/PFN1
GO:0030274                                                     TLN1
GO:0031267                                                IPO5/PFN1
GO:0000339                                                   EIF4A1
           Count
GO:0003924     4
GO:0008135     3
GO:0003746     2
GO:0000166     9
GO:0005525     4
GO:0019001     4
GO:0032561     4
GO:0004757     1
GO:0016768     1
GO:0017111     5
GO:0016462     5
GO:0016818     5
GO:0016817     5
GO:0003723     5
GO:0004766     1
GO:0035639     7
GO:0032555     7
GO:0032553     7
GO:0017076     7
GO:0030292     1
GO:0017016     2
```

```
GO:0030274     1
GO:0031267     2
GO:0000339     1
```

- `enrichKEGG` for KEGG pathway enrichment analysis.

```
> z <- enrichKEGG(gene=gcSample[[3]], organism="human", pvalueCutoff=0.05, qvalueC
> summary(z)
```

```
      pathwayID                                Description
05130  hsa05130    Pathogenic Escherichia coli infection
04145  hsa04145                               Phagosome
04540  hsa04540                             Gap junction
04962  hsa04962 Vasopressin-regulated water reabsorption
      GeneRatio  BgRatio       pvalue       qvalue
05130      4/17  58/5894 1.826892e-05 0.0002115348
04145      5/17 156/5894 5.827611e-05 0.0003373880
04540      4/17  90/5894 1.039489e-04 0.0004012064
04962      2/17  44/5894 6.898981e-03 0.0199707355
                              geneID Count
05130       TUBB2C/TUBB2A/TUBB3/TUBB6     4
04145 TUBB2C/TUBB2A/TUBB3/RAB5B/TUBB6     5
04540       TUBB2C/TUBB2A/TUBB3/TUBB6     4
04962                      NSF/RAB5B     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 *rPA* `https://github.com/GuangchuangYu/rPA`, 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).

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*.

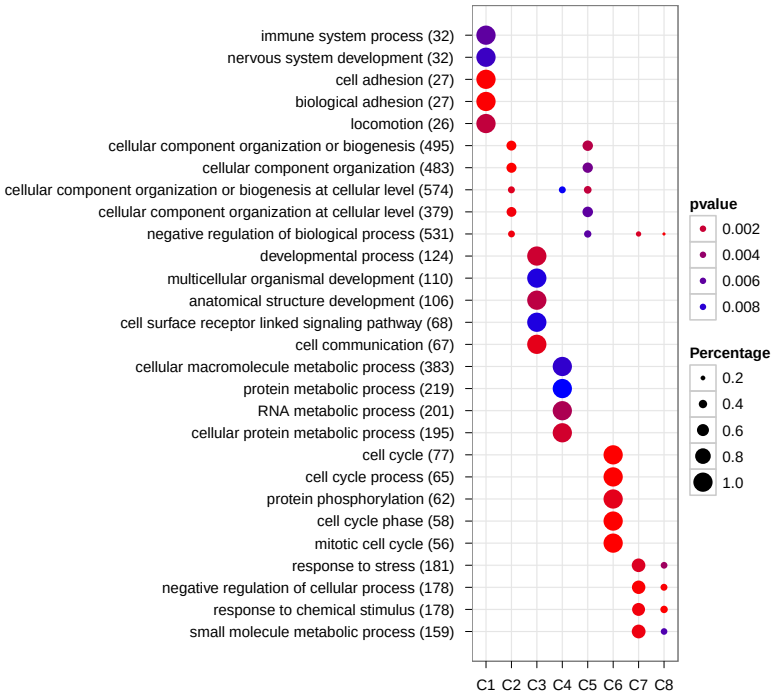

Figure 1: Comparison of GO enrichment of gene clusters

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 *rPA* and compare gene cluster in the context of disease and reactome pathway as demonstrated in the online vignette of *DOSE* and *rPA* respectively.

5 Visualization

clusterProfiler implemented serveral methods for visualizing analyzed result.

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)
> plot(z)
```

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, ont="CC", organism="human", pvalueCutof
> plot(xx)
```

```
> plot(y, titile="MF Enrichment analysis", showCategory=10)
```

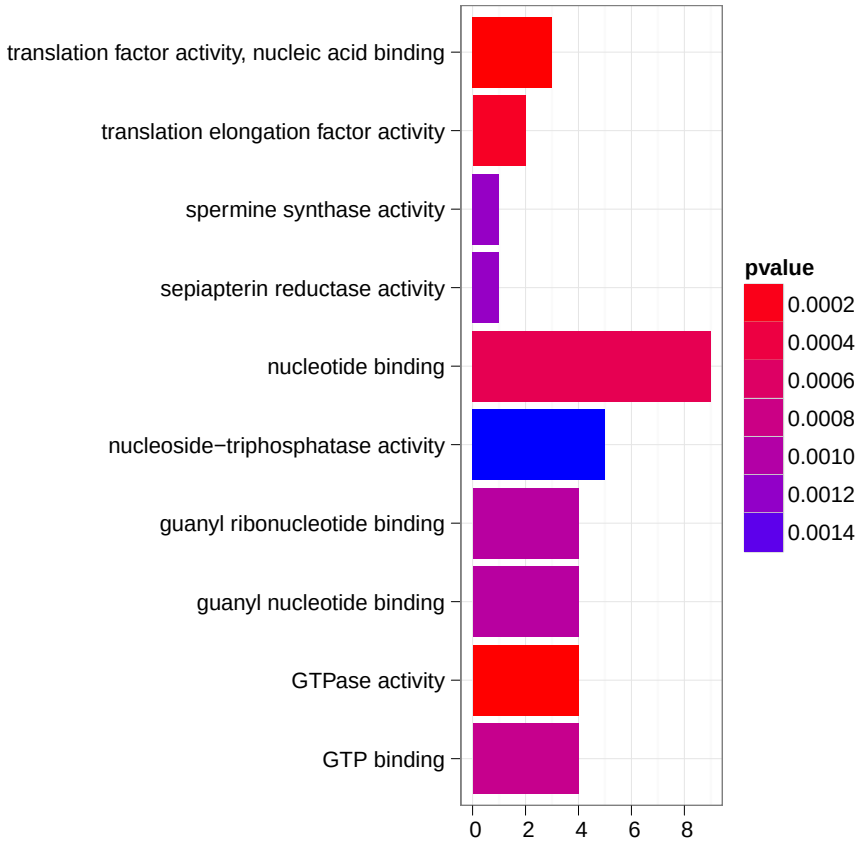

Figure 2: Example of plotting functional profiles

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 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-

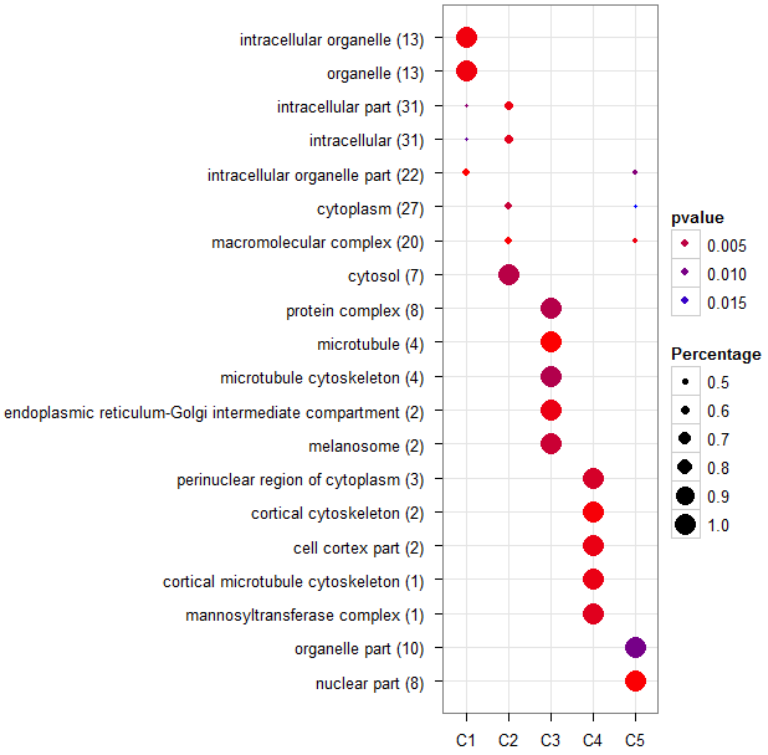

Figure 3: GO Enrichment Comparison

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 *rPA*). More details can be found in the vignettes of *DOSE* and *rPA*.

6 Session Information

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

```
R version 2.14.1 (2011-12-22)
Platform: i386-redhat-linux-gnu (32-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_2.6.1            clusterProfiler_1.3.11
[3] AnnotationDbi_1.17.22  Biobase_2.15.3
[5] BiocGenerics_0.1.6     RSQLite_0.11.1
[7] DBI_0.2-5              ggplot2_0.9.0

loaded via a namespace (and not attached):
 [1] IRanges_1.13.25     KEGG.db_2.6.1
 [3] MASS_7.3-16         RColorBrewer_1.0-5
 [5] colorspace_1.1-1    dichromat_1.2-4
 [7] digest_0.5.1        grid_2.14.1
 [9] igraph_0.5.5-4      memoise_0.1
[11] munsell_0.3         org.Hs.eg.db_2.6.4
[13] org.Mm.eg.db_2.6.4  org.Sc.sgd.db_2.6.4
[15] plyr_1.7.1          proto_0.3-9.2
[17] qvalue_1.29.0       reshape2_1.2.1
[19] scales_0.1.0        stringr_0.6
[21] tcltk_2.14.1        tools_2.14.1
```

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.