# Disease Ontology Semantic and Enrichment analysis

Guangchuang Yu, Li-Gen Wang

May 29, 2013

## Contents

# 1 Introduction

Disease Ontology (DO) provides an open source ontology for the integration of biomedical data that is associated with human disease. DO analysis can lead to interesting discoveries that deserve further clinical investigation.

*DOSE* was designed for semantic similarity measure and enrichment analysis.

Four information content (IC)-based methods, proposed by Resnik [**?**], Jiang [**?**], Lin [**?**] and Schlicker [**?**], and one graph structure-based method, proposed by Wang [**?**], were implemented. These methods were also implemented in our *GOSemSim* [**?**] package for measuring GO-term semantic similarities. Hypergeometric test [**?**] was implemented for enrichment analysis.
To start with *DOSE* package, type following code below:

```
library(DOSE)
help(DOSE)
```

# 2 Enrichment Analysis

Enrichment analysis 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. We also implement a bar plot and gene-category-network for visualization.

To determine whether any DO terms annotate a specified list of genes at frequency greater than that would be expected by chance, 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 DO annotation.

```
data(geneList)
gene <- names(geneList)[geneList > 1 | geneList < 1]
x <- enrichDO(gene, pvalueCutoff=0.05)
head(summary(x))
```

```
                    ID         Description GeneRatio  BgRatio   pvalue p.adjust
DOLite:100 DOLite:100              Cancer  665/3396 736/4051 1.65e-08 6.67e-06
DOLite:44   DOLite:44 Alzheimer's disease  184/3396 193/4051 4.26e-07 8.63e-05
DOLite:173 DOLite:173       Endometriosis  138/3396 145/4051 1.85e-05 2.50e-03
DOLite:156 DOLite:156   Diabetes mellitus  329/3396 362/4051 3.53e-05 3.57e-03
DOLite:376 DOLite:376 Neoplasm metastasis  143/3396 152/4051 9.05e-05 7.33e-03
```

```
DOLite:320 DOLite:320         Lung cancer  189/3396 205/4051 2.39e-04 1.38e-02
             qvalue
DOLite:100 5.97e-06
DOLite:44  7.72e-05
DOLite:173 2.23e-03
DOLite:156 3.19e-03
DOLite:376 6.55e-03
DOLite:320 1.24e-02

DOLite:100 9052/9/6286/4582/4583/10549/2099/6241/1894/2261/11065/1509/4072/9232/9833/
DOLite:44
DOLite:173
DOLite:156
DOLite:376
DOLite:320
           Count
DOLite:100   665
DOLite:44    184
DOLite:173   138
DOLite:156   329
DOLite:376   143
DOLite:320   189
```

```
barplot(x)
```

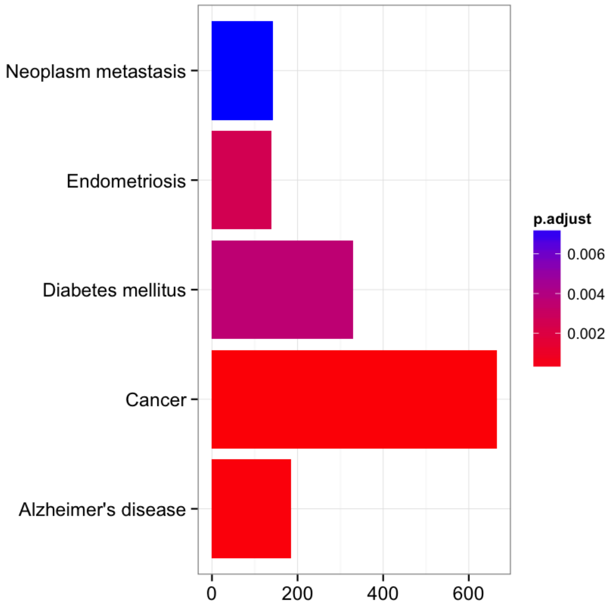

Figure 1: barplot of DO enrichment result

# 3 Enrichment Analysis

```
require(DOSE)
data(geneList)
y <- gseaAnalyzer(geneList, setType="DO", nPerm=100, minGSSize=120, pvalueCutoff=0.05,
res <- summary(y)
head(res)
```

```
                       ID                       Description enrichmentScore
DOID:0050687 DOID:0050687                  cell type cancer          -0.335
DOID:1240       DOID:1240                          leukemia          -0.347
DOID:14566     DOID:14566 disease of cellular proliferation          -0.318
DOID:150         DOID:150          disease of mental health          -0.389
DOID:16           DOID:16      integumentary system disease          -0.432
DOID:162         DOID:162                            cancer          -0.309
             pvalues p.adjust qvalues
DOID:0050687       0        0       0
DOID:1240          0        0       0
DOID:14566         0        0       0
DOID:150           0        0       0
DOID:16            0        0       0
DOID:162           0        0       0
```

```
topID <- res[1,1]
```

```
gseaplot(y, geneSetID = topID)
```

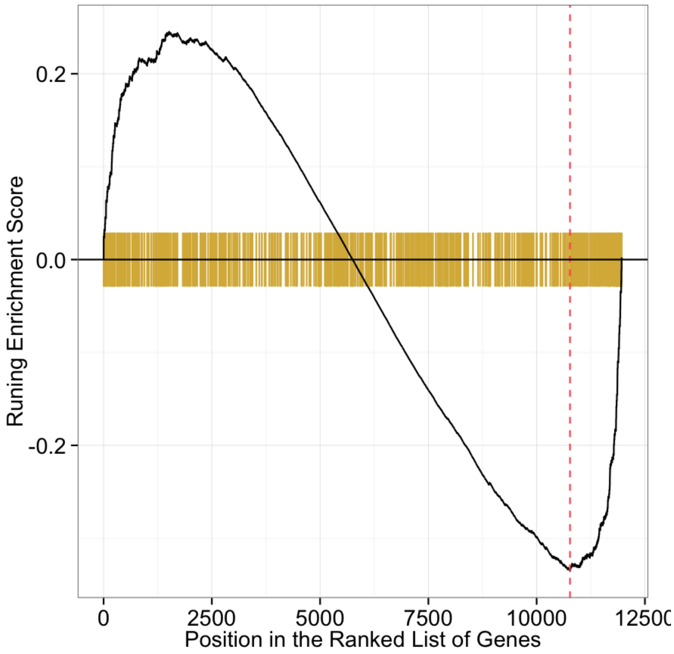

Figure 2: gseaplot example

## 4 Session Information

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

- R version 3.0.0 (2013-04-03), `x86_64-apple-darwin10.8.0`
- Locale: `C/UTF-8/C/C/C/C`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationDbi 1.22.5, Biobase 2.20.0, BiocGenerics 0.6.0, DBI 0.2-7, DO.db 2.6.0, DOSE 1.99.0, RSQLite 0.11.3, cacheSweave 0.6-1, filehash 2.2-1, stashR 0.3-5
- Loaded via a namespace (and not attached): GO.db 2.9.0, GOSemSim 1.18.0, IRanges 1.18.1, MASS 7.3-26, RColorBrewer 1.0-5, colorspace 1.2-2, dichromat 2.0-0, digest 0.6.3, ggplot2 0.9.3.1, grid 3.0.0, gtable 0.1.2, igraph 0.6.5-2, labeling 0.1, munsell 0.4, plyr 1.8, proto 0.3-10, qvalue 1.34.0, reshape2 1.2.2, scales 0.2.3, stats4 3.0.0, stringr 0.6.2, tcltk 3.0.0, tools 3.0.0

## References

[1] Resnik Philip. Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[2] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of 10th International Conference on Research In Computational Linguistics*, 1997.

[3] Dekang Lin. An Information-Theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296—304, 1998.

[4] Andreas Schlicker, Francisco S Domingues, JÃűrg RahnenfÃijhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302, 2006. PMID: 16776819.

[5] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics (Oxford, England)*, 23:1274–81, May 2007. PMID: 17344234.

[6] 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. PMID: 20179076.

[7] 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.