A BIOINFORMATICS TUTORIAL RESOURCE

on June 20, 2011 by Simon Cockell in [Transcriptomics](http://bioinformatics.knowledgeblog.org/category/tutorials/transcriptomics/%22%20%5Co%20%22View%20all%20posts%20in%20Transcriptomics), [Comments (4)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#comments)

Downstream functional analysis of an ‘omics experiment using freely available tools

1. Introduction

The direct analysis of data from ‘omics experiments (transcriptomics and proteomics especially), focusses on producing lists of genes or proteins of interest (i.e. those that vary in expression, or some other character, between experimental conditions). However, merely identifying those genes that are altered in the experiment does not generally advance the understanding of the processes being studied. Downstream functional analysis of ‘omics data focusses on connecting the genes in a gene list by related function, with the aim of shedding more light on the processes being affected by coordinated changes in gene expression.

In this tutorial, a number of tools for functional analysis will be introduced. We will focus on the downstream analysis of the gene list produced by the [Analysing microarray data in R and BioConductor](http://bioinformatics.knowledgeblog.org/2011/06/20/analysing-microarray-data-in-bioconductor/) tutorial.

1. Hypergeometric tests for over-representation of functional categories

The Hypergeometric distribution is a statistical distribution that describes the number of success in a series of *n*draws without replacement, from a finite population, *N*.
A test against this distribution determines the probability that the number of successes observed could be obtained by chance. It is an appropriate test for the kind of functional analysis described here because a gene list can be seen as a series of draws from the finite population of Gene Ontology terms that annotate the whole of a genome, therefore a Hypergeometric test determines the possibility that a certain term is over-represented in the gene list with respect to the parent population.

***Tools***

There are many tools that implement the hypergeometric test for over-representation of terms, often from the Gene Ontology, but also other sets of terms such as KEGG or Reactome pathways. Below is a brief tutorial for the use of a few of these tools.

*DAVID (Database for Annotation, Visualization and Integrated Discovery)*

[DAVID](http://david.abcc.ncifcrf.gov/), the Database for Annotation, Visualization and Integrated Discovery, is a suite of online tools which provide a number of different analysis methods [(Huang, Sherman, and Lempicki 2008)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-0)[\*](http://dx.doi.org/10.1038/nprot.2008.211). For the purposes of this tutorial, we are going to focus on the ‘functional annotation clustering’ tool.

Go to the [DAVID web site](http://david.abcc.ncifcrf.gov/) and click on ‘Start Analysis’, the link can be found in the menu bar beneath the web page header. The first step in any analysis with DAVID is to upload your gene list. Download the list of probeset identifiers from [this link](http://bioinformatics.knowledgeblog.org/wp-content/uploads/bioinf/gsea/gene_names.txt), and either upload it into DAVID using the ‘Upload a File’ option, or paste the contents into the text box in the left-hand side of the page. Select ‘OFFICIAL\_GENE\_SYMBOL’ as the Identifier Type, select the ‘Gene List’ radio button, and press ‘Submit List’. DAVID warns us that it can map the identifiers to multiple species, dismiss the warning, but make sure you select the *Homo sapiens* from the list of species in the left hand side panel. DAVID presents an ‘Analysis Wizard’ which lets us quickly submit our list of genes to one of DAVID’s analysis tools. Analysing functional over-representation is as simple as clicking on the ‘Functional Annotation Tool’ link.

The 194 differentially expressed probesets from the [microarray analysis tutorial](http://bioinformatics.knowledgeblog.org/2011/06/20/analysing-microarray-data-in-bioconductor/) map to 133 human genes in the DAVID database. The Annotation summary gives lists of the functional terms that DAVID has analysed for over-representation. As an example, click on the ‘+’ icon next to ‘Gene Ontology’ to expand that list, and then click the ‘Chart’ button next to ‘GOTERM\_BP\_FAT’ to look at Biological Process enrichment in more detail. There are 126 biological process terms in this list, all with a P-value < 0.05. We are, however, performing multiple tests here, so we must correct these P-values to allow for that. Fortunately DAVID has already done this, and we can see these corrected P-values in the final column of the chart (headed ‘Benjamini’). 10 of the 126 terms are still statistically over-represented if we use this corrected P-value

**Table 1** – Statistically over-represented GO Biological Process terms, according to DAVID.

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **Number of genes with annotation** | **P-value** | **Benjamini-Hochberg corrected P-Value** |
| collagen fibril organization | 6 | 2.9E-6 | 3.4E-3 |
| extracellular structure organization | 10 | 5.6E-6 | 3.3E-3 |
| extracellular matrix organization | 8 | 1.8E-5 | 7.1E-3 |
| skeletal system development | 12 | 4.4E-5 | 1.3E-2 |
| negative regulation of signal transduction | 10 | 6.3E-5 | 1.5E-2 |
| response to nutrient | 8 | 1.2E-4 | 2.4E-2 |
| response to steroid hormone stimulus | 9 | 1.4E-4 | 2.4E-2 |
| negative regulation of cell communication | 10 | 1.5E-4 | 2.2E-2 |
| response to estrogen stimulus | 7 | 1.8E-4 | 2.4E-2 |
| tube development | 9 | 3.5E-4 | 4.1E-2 |

This list of over-represented terms gives us a lot of information about the genes that have changed expression in our experiment. But we have no information about how the individual terms are related. DAVID provides a functional annotation clustering tool that can help with this, but the results are still represented as HTML tables. A more visual representation of the relationships between terms may help us with our interpretation.

*BiNGO*

BiNGO [(Maere, Heymans, and Kuiper 2005)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-1)[\*](http://dx.doi.org/10.1093/bioinformatics/bti551) is a plugin for the popular graph visualisation tool Cytoscape[(Smoot et al. 2011)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/%22%20%5Cl%20%22ITEM-45-2)[\*](http://dx.doi.org/10.1093/bioinformatics/btq675). BiNGO performs essentially the same calculations as DAVID, but displays the results in a much more visual format, a network representation of the over-represented GO terms and their connections within the Directed Acyclic Graph (DAG) of the Gene Ontology.

To run BiNGO, we first need to download and install [Cytoscape](http://cytoscape.org/) (version at time of writing: 2.8.1), then run it. Once Cytoscape is running, click on the ‘Plugins’ menu, and select ‘Manage Plugins’. This will bring up the plugin manager. Search for BiNGO, and install the latest version (currently 2.44). Now from the ‘Plugins’ menu, select ‘Start BiNGO’ to bring up the options box for the plugin.

Fill in the box as illustrated (Figure 1). Paste the list of gene names we used for DAVID into the text box (after selecting ‘Paste Genes from Text’). Select ‘GO\_Biological\_Process’ from the ‘Select Ontology File’ dropdown, and select ‘Homo sapiens’ from the list of organisms. Now click ‘Start BiNGO’ and wait for the analysis to complete.



*Figure 1 - The settings for BiNGO used here*

BiNGO produces a network (Figure 2), and a table of results (excerpt – Table 2). The list of terms produced by BiNGO looks rather different from that produced by DAVID, but this can be largely explained by the fact that BiNGO analyses every term in the Gene Ontology section selected, whereas DAVID cuts the DAG off at a certain level, and only analyses terms with a particular specificity. Terms, such as ‘collagen fibril organzation’ which score highly with DAVID still score highly with BiNGO. Further differences are accounted for by BiNGO not being as successful as DAVID at mapping Gene Names to entities in the Gene Ontology (110 vs 133 genes mapped), and BiNGO shipping with an outdated version of the Gene Ontology, which is not as complete as the more up-to-date version used by DAVID (it is perfectly possible to update the annotation used by BiNGO).



*Figure 2 - Network output from BiNGO when run using the settings shown in figure 1.*

The network is a representation of the segment of the GO DAG which contains the over-represented terms, so not only do we see the terms and the magnitude of their statistical over-representation (signified by the ‘redness’ of a given node in the network), we can also see how those over-represented terms are related within the Gene Ontology.

**Table 2** – Top 10 statistically over-represented GO Biological Process terms, according to BiNGO.

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **Number of genes with annotation** | **P-value** | **Benjamini-Hochberg corrected P-Value** |
| Anatomical Structure Development | 47 | 3.7633E-9 | 5.7353E-6 |
| System Development | 43 | 2.4002E-8 | 1.8290E-5 |
| Anatomical Structure Morphogenesis | 28 | 9.4070E-8 | 4.7788E-5 |
| Developmental Process | 49 | 2.6029E-7 | 9.9169E-5 |
| Multicellular Organismal Development | 46 | 4.2036E-7 | 1.2813E-4 |
| Organ Development | 33 | 9.1325E-7 | 2.3196E-4 |
| Response to Stimulus | 51 | 1.4705E-6 | 3.2015E-4 |
| Collagen Fibril Organization | 5 | 1.7326E-6 | 3.3007E-4 |
| Response to Estrogen Stimulus | 8 | 2.6527E-6 | 4.4919E-4 |
| Tissue Development | 19 | 3.8427E-6 | 5.3740E-4 |

*Other Tools*

There are plenty of other tools for looking at term enrichment within a gene list. Many of which are freely available.

* [GOStats](http://www.bioconductor.org/packages/2.3/bioc/html/GOstats.html) is a Bioconductor package that will allow for enrichment analysis to be added simply to an R workflow for microarray analysis [(Falcon and Gentleman 2007)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-3)[\*](http://dx.doi.org/10.1093/bioinformatics/btl567).
* [g:Profiler](http://biit.cs.ut.ee/gprofiler) is a web tool for enrichment analysis, results are coded by evidence type from GO, for visually classifying by reliability [(Reimand et al. 2007)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-4)[\*](http://dx.doi.org/10.1093/nar/gkm226).
* [GeneMANIA](http://www.genemania.org/) is another Cytoscape plugin that finds many types of relationship between the genes of a gene list [(Warde-Farley et al. 2010)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-5)[\*](http://dx.doi.org/10.1093/nar/gkq537). Among these relationships are physical interactions (i.e protein-protein interactions), pathway membership, and shared GO terms (including a calculation of over-representation).
* [GOrilla](http://cbl-gorilla.cs.technion.ac.il/) is another web tool which allows terms over-represented within a gene list to be calculated [(Eden et al. 2009)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-6)[\*](http://dx.doi.org/10.1186/1471-2105-10-48). Results can be sent to [REViGO](http://revigo.irb.hr/) for subsequent visualisation [(Supek et al. 2011)](http://bioinformatics.knowledgeblog.org/2011/06/20/downstream-functional-analysis-of-an-omics-experiment-using-freely-available-tools/#ITEM-45-7)[\*](http://dx.doi.org/10.1371/journal.pone.0021800).

Analysis for enrichment of functional terms in a list of genes is a relatively simple operation, with a number of tools available for achieving the same results, each with different strengths and weaknesses. It can be argued that by only focussing on those genes which have significantly altered expression between conditions a large amount of information is being lost, and by not relating terms to one another meaningfully, the underlying biology is ignored. Another class of tools has been developed that aim to address these issues, and are the subject of another article: [Gene Set Enrichment Analysis](http://bioinformatics.knowledgeblog.org/2011/06/20/gene-set-enrichment-analysis/).

1. Bibliography

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*. Springer Science + Business Media. <http://dx.doi.org/10.1186/1471-2105-10-48>.

Falcon, S., and R. Gentleman. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics*. Oxford University Press (OUP), January 15. <http://dx.doi.org/10.1093/bioinformatics/btl567>.

Huang, Da Wei, Brad T Sherman, and Richard A Lempicki. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. Nature Publishing Group, December.<http://dx.doi.org/10.1038/nprot.2008.211>.

Maere, S., K. Heymans, and M. Kuiper. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*. Oxford University Press (OUP), August 15.<http://dx.doi.org/10.1093/bioinformatics/bti551>.

Reimand, J., M. Kull, H. Peterson, J. Hansen, and J. Vilo. 2007. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*. Oxford University Press (OUP), May 8. <http://dx.doi.org/10.1093/nar/gkm226>.

Smoot, M. E., K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. Oxford University Press (OUP), February 1.<http://dx.doi.org/10.1093/bioinformatics/btq675>.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Ed. Cynthia Gibas. *PLoS ONE*. Public Library of Science (PLoS), July 18.<http://dx.doi.org/10.1371/journal.pone.0021800>.

Warde-Farley, D., S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, et al. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*. Oxford University Press (OUP), July 1. <http://dx.doi.org/10.1093/nar/gkq537>.

Tags: [analysis](http://bioinformatics.knowledgeblog.org/tag/analysis/), [gene-ontology](http://bioinformatics.knowledgeblog.org/tag/gene-ontology/), [hypergeometic-distribution](http://bioinformatics.knowledgeblog.org/tag/hypergeometic-distribution/), [microarray](http://bioinformatics.knowledgeblog.org/tag/microarray/)