Babraham Bioinformatics

Exercises:
Running Gene Lists

Version 2019-02

Licence

.

Introduction

In this practical you are going to take a list of experimentally determined differentially expressed genes and will run them through a few different Gene Ontology search tools. In each case you can see the specifics of how the different tools work, and can see what results are produced. In each case try to identify the common major measures which should be produced by this type of analysis:

1. The name of the enriched gene set
2. The source of the originally curated list of genes
3. The significance of the statistical test used. Be sure you know whether the p-values you see are raw or are corrected for multiple testing.
4. The degree of enrichment (absolute change, odds ratio etc)

Ideally we'd also like to know:

1. Which genes from your query list were found in the gene set.
2. Which genes in the gene set were not found in your query

Sometimes these bits of information will be available, but not in all cases.

The rest of the sections below are split based on the tool we're going to use. Because we don't want everyone hitting the same tool at the same time pick a random place to start rather than necessarily starting with the first tool, and then work your way through them.

Your starting list of genes is in "human_gene_list.txt". It is a qualitative gene list, but it is ordered by significance, so the most significant genes are at the top of the list.

Panther: http://www.pantherdb.org

Panther allows you to do two types of analysis:

1. An assessment of the division of your gene list into different gene ontology categories
2. As assessment of enrichment in your gene list compared to a background

We want to do the second (but we'll look at the first a bit too).

- Paste your list of genes into the search box on the Panther front page
- Select that these are Human genes from the organism options
- Select "Statistical overrepresentation test" in the analysis section
- Press submit to run the analysis

By default Panther will only analyse the Biological Process subset of the cut down "GO-Slims" Gene Ontology data. If you want to analyse the other GO sections you need to change the options in the "Annotation Data Set" at the top of the page.

When looking at the results try to answer the following questions:

- How many of the original genes were recognised?
- What terms were most changing? Were they enriched or depleted (Panther identifies both)
- What effect does switching to the full GO, rather than GO-slims have on the results? Change this in the "Annotation Data Sets" at the top.
- For any category which was identified as interesting, can you find out what the description for that category is?

As well as running the analysis you can also draw some figures which summarise what you saw. These only work for the GO-Slims results. They can be found just above the top of the results table. Try drawing the multiple pie charts, and the bar charts of gene counts and see if you understand what they say about your list.

You can export the results by pressing the "Export Results" button. Look at the text file you get back and see if the information matches up with what you see in the web based report.

DAVID: https://david.ncifcrf.gov/

DAVID is the most popular gene set enrichment tool based on references and publications. Not necessarily because it does anything better than any other tool, but it's been around a long time and is generally quite easy to use.

- From the home page of DAVID select "Functional Annotation" to go to the enrichment tool.
- Paste your list of genes into the search box on the left
    - Change the Identifier to OFFICIAL_GENE_SYMBOL
    - Select "GeneList" as the type for the list
- Press "Submit list"

Since you are supplying a list of gene names there is potentially some ambiguity in the species from which they came. DAVID will therefore give you a warning to say that your list could have come from a number of different species, and give you a list to select from.

- Select "Homo sapiens" from the list you are given and press "Select species" to activate it
- Look in the Annotation Summary Results and check that "Homo sapiens" is the background list it is using.
- Look in the Annotation Summary Results to see how many IDs it is working with. You started with a list of 1432 genes. How many are left?
    - You can see which genes weren't found by clicking on the "View Unmapped IDs" link on the left hand toolbar.
- Under the Annotation Summary you can see the sets of Gene Lists which DAVID can analyse. You can expand these by clicking on the small + signs. Have a look which sets are analysed by default, and look at which others you could add. Can you tell what all of the lists are?

DAVID has a couple of different types of analysis it can run. The simplest is the Functional Annotation Chart where each gene list is analysed independently. Run this and look at the results.

- Do you see the list source, p-value, enrichment and gene name lists in the chart?
    - If not then can you get more of these by changing the options at the top of the page?
- Can you find the details of what each gene set means?

You can also run a version of the analysis where DAVID tries to group together related gene sets to give you a view which makes it easier to get at the overall biological themes. Run the "Functional Annotation Clustering" tool to get this view.

- Can you see the groups of hits, rather than individual lists.
- Can you see all of the metrics you would like?
- Save the results using the "Download File" link
    - What level of detail do you get in the downloaded data?

GOrilla: http://cbl-gorilla.cs.technion.ac.il

GOrilla is a fairly straight forward Gene Ontology search tool, with some interesting visualisation options. Unlike DAVID and Panther it requires an **ordered** gene list as input (which the one you have is).

You should be able to specifiy the Species and Genes for the search. As with Panther you have to search each Gene Ontology section separately so start with Biological Process since that's often the one which gives the most relevant information (you can try others later if you want).

The output from GOrilla comes in two parts, a graphical view of the Gene Ontology structure and a table of hits.

- Make sure you understand what the graphical view of the ontology means, and can identify groups of hits from your data.
- How well do you think this presentation of data works for different numbers of hits?
- In the table can you see all of the metrics you would want (p-value, enrichment and gene lists)?
- Can you easily get to the information about what the gene ontology section means?
- Can you see how many of your original genes were recognised by GOrilla?
- Can you save a table of results?

GOrilla also offers the ability to link out to Revigo, which is a tool to try to collate large numbers of GO hits. This option isn't selected by default.

Go back to the main GOrilla page and re-run the analysis, but this time scroll down to the Advanced Parameters and turn on "Show output also in REViGO". After re-running GOrilla (which will look the same), select "Visualise output in REViGO" (underneath the hit table).

*NB Revigo is a Flash web application. Some computers will require that you specifically allow it to run. Depending on your settings it's possible that you won't be allowed to run Flash at all, in which case you'll have to skip this bit and look at the summarised results later.

- Take the default REViGO options and start the tool.
- The useful plots are the "Scatterplot and Table" and the "Tree Map"
    - Look at both of these and check you understand what they are showing
    - Try altering some of the options in the scatterplot to see if you can make it clearer

If you cannot run Flash on your computer, or you would like to export a version of the plot, and you are comfortable using R, Revigo does provide the option of creating an R script that can be downloaded and run in an R session to produce the plot. You should see the option "Make R script for plotting" at the top right of the table of results.

GProfiler (gGOST): https://biit.cs.ut.ee/gprofiler/

GProfiler has quite a wide variety of gene sets which it can analyse and an interesting presentation of results which works well if you have small gene lists.

Start by running the tool using the full gene set and the default options. Note that the default output is "Graphical (PNG)" so you should get an image result.

- Can you see at the top the full list of sources for gene sets which the tool is using?
- Do you actually get a graphical result from your search?
    - If not, can you see why not?
- How easy is it to see the metrics you would like (enrichment, p-value, gene lists)?
- Can you understand the way the results are structured?

Instead of using the entire gene list, try rerunning the tool using the top 50 or so hits (you don't need to be exact).

- What difference does this make to the way the results are presented?
- Do you think this presentation of results is useful?
- Can you see how to tell the level of annotation for each of the GO category assignments (look at the colour key at the top of the page)

Finally, run the tool opting to generate an Excel file as output. Have a look at the information which is in there. Do you have everything you'd want?

EnrichR: http://amp.pharm.mssm.edu/Enrichr/

EnrichR is about the simplest gene set enrichment tool to use, and has an impressively large list of gene sets to work with. It also has very few options to set, which could be considered to be a good and bad thing.

To run the tool just paste in your gene list and press Submit. There are no additional options to set.

The results come in groups of lists. These are categorised into major groups at the top of the page, and then split into specific studies or groups in the main view. Have a look through some of the categories to get a general impression of what the tool is showing you.

- How easy is it to get an impression of the relative level of importance of the hits in the different categories?
- Do the top hits in the Gene Ontology categories match with what you have seen before (if this isn't the first tool you've run)
- These hits came from a comparison of two cell lines – T47D and Ish. Under the Cell Types group you can see hits mapped to marker genes. Do the right cell lines come up?
- Click on the hits for GO Biological Process to see the more detailed view
    - Do you find all of the views useful?
    - Can you export the table of hits?
    - Does the exported table have all of the information you'd want?