Babraham Bioinformatics

Exercises:
Running Gene Lists

Version 2021-11

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. For the tools which allow you to specify a background gene list this is contained in "**`Human_background_list.txt`**" and you can use that too if you wish. We'll talk about the relevance of a background list in the next section.

1. 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. An 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
- You will be offered the chance to provide a custom background list. For this initial analysis you should use the default human gene list from the drop down on the right.

Panther will only analyse one section of the Gene Ontology at a time and also offers 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 Set" option. You can also select pathway databases to search against.

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.

2. 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 "Start Analysis" 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
    - Type homo sapiens into the species selection
    - Select "GeneList" as the type for the list
- Press "Submit list"

- Look in the gene list manager to see how many genes were imported, and whether any were not recognised as human genes (will be reported as Unknown. You started with a list of 1432 genes.
- You can see which genes weren't found by clicking on the "View Unmapped IDs" link at the bottom of the left hand toolbar. It's a blue link on a blue background so it's not easy to see!
- You can start the analysis by clicking on the "Functional Annotation Tool" link. This will be in the main window when you first upload your gene list, but if you've viewed an uploaded list you'll need to select it from the top menu under "Shortcut to DAVID tools"

- At the top of the Annotation Summary Page you will see all of the gene sets DAVID can use for its analysis. You can expand these to see what's available and what is selected by default. You can click on the "Chart" button to see the lists in each group.

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. Go back to the main DAVID window and 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?

3. 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), or you can supply an unordered list along with a custom background list. We also supplied a background list so you could try that too.

You should be able to specify 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).

The Revigo results will show you a few different visualisations of the hit categories from your Gorilla search. In all cases they are trying to indicate which of the hit categories are semantically similar to each other to reduce the complexity of the results you need to interpret.

- All three of the main plots are interesting ways to review this information.
    - Look at all of the plots and check you understand what they are showing
    - Look at some of the closely linked groups and see if you can see how their functionality is related

If 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. If you download this you can either open in in a program such as RStudio and run the code to see the plot, or you can generate the plot in a terminal by running:

```
Rscript REVIGO.r
```

Which will generate a PDF you can then view.

4. 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 really well if you have small gene lists.

Start by running the tool using the full gene set and the default options. You will be asked about identifiers with ambiguous names, you can choose the option to “Select the Ensembl ID with the most GO annotations”, and then press the “Rerun” button.

The “Results” tab should show you a dynamic summary graph of your hits. You can mouse over hits to see what they are, and click on them to build up a table of interesting categories below the main plot. See how the hits are grouped by data source, and check you know what each source is.

The “Detailed Results” tab gives more information about the hits and allows for additional filtering of the hits and the presentation of additional hit metrics.

- Can you see at the top the full list of data sources for gene sets which the tool is using?
- Do you actually get a graphical result from your search?
- How easy is it to see the metrics you would like (enrichment, p-value, gene lists)? You may need to expand the “stats” section by clicking on the >> symbol.

- Try using the filter options in the detailed results to show only categories with between 3 and 100 genes and see what effect this has on the specificity of the categories shown.
- Press the “Show evidence codes” link to load in the details of which genes are in which categories and the evidence on which the assignment was made. You’ll need to scroll to the right of the results tables to see this option.

Finally, go to the “Detailed Results” tab and export as a CSV file to generate an Excel file as output. Have a look at the information which is in there. Do you have everything you’d want?

5. EnrichR: https://maayanlab.cloud/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?

6. GOliath: https://www.bioinformatics.babraham.ac.uk/goliath/

GOliath is a simple search tool, but which can also identify artefacts and biases within your results. We will discuss these in the next talk, but you can run the tool now and see what you find.

To run the tool just paste in your gene list, select Human as the species and press “Analyse my list”.

Have a look at the list of hits and see what was found.

Do any of the hits have potential biases associated with them? Click on the bias to see what it might mean.

Select the “Properties” tab to look at some overall properties of your gene list. How do the properties of the hits compare to those of all of the genes in the genome. Can you think of any reasons you might see a difference?

Under the biases tab see what biases are being tested for in this species, and which were found in your hits. You can look at the details of how the biased sets were collected.

Finally, export your list of hits as an excel (or csv if you prefer) file and check what information is given in there.

7. Reactome: https://reactome.org/

Reactome is a pathway based gene set analysis whichcan analyse either categorical or quantitative data. It doesn't have much customisation and won't allow you to customise a background list. Some of the other tools have pathways as one of their data sources but this is the main focus of reactome.

To get to the gene set search click on the "Analyze Data" button on the home page. You can then paste your gene list into the search box and press continue.

You can leave the "Project to human" option ticked (these should be human genes anyway) and don't use the option to extend your gene set using IntAct interactions.

In the main display you should see a list of different pathway categories in a window on the left. A graphical overview of the pathways at the top and a list of specific pathways and their statistical results at the bottom.

On the top right of the statistical results table is a round button which if you hover your mouse over it says "Filter your results". Use this to show only results which are significant (p<0.05).

You can click on an individual hit to see the area of the pathway diagram from which it comes, and see if other similar pathways have been detected too. You can click twice on a node in the graph to open up a more detailed view of the pathway. You can go back by clicking on the round "Pathway overview" button at the top of the graphical window.

Probably the nicest view of the data is the voronoi diagram generated by FoamTree. From the graphical pathway overview you can click the round button at the top to generate this view in a new browser tab. Pathways with significant enrichment will be coloured yellow. Have a look at this view and see which areas of the pathway landscape are particularly activated in this dataset.