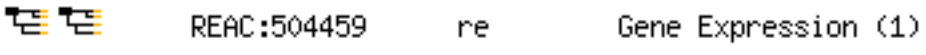
g:Profiler — functional enrichment analysis of gene lists

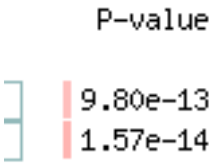
Jüri Reimand
Juri.Reimand@ut.ee

g:Profiler -- functional enrichment analysis of gene lists

- Open a browser window at the URL http://biit.cs.ut.ee/gprofiler/ .
- Open another browser window and point it to the online course materials at http://emu.at.mt.ut.ee/u/reimand/roadshow2009/.
- Copy the contents of `ISW1_up_tgts.txt` to g:Profiler query field. It contains 188 yeast genes that are upregulated in ISW1 knockout mutant. ISW1 is a component of several chromatin remodelling complexes with roles in regulating transcriptional initiation and elongation.
- Change g:Profiler organism dropdown value to yeast (`Saccharomyces cerevisiae`) and click on `g:Profile` start the analysis.
- Investigate the output of genes and corresponding functional terms. Navigate to the right to see functional terms and corresponding statistics. Functional terms from the Gene Ontology (GO) are shown first: biological processes (bp), cell components (cc) and molecular functions (mf). Matches for pathways (KEGG, Reactome) are shown below. At the bottom of the page, you can see your input gene list with brief descriptions.
- What is the meaning of coloured boxes? Why do you see mostly blue and black boxes?
- Visualise the hierarchy of functional terms below `REAC:504522 Gene Expression.`
  hint: click the icon just next to term ID.

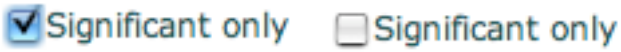


- Enrichment p-values are shown right of the last gene in the query. Shades of pink and red highlight strength of p-values.

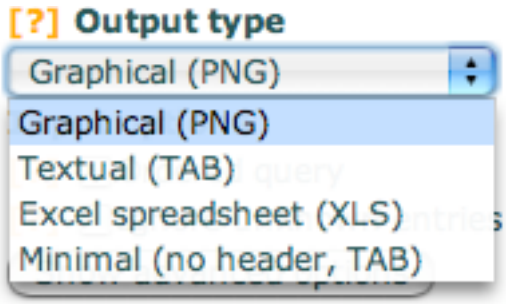


- Terms are grouped according to logical and hierarchical rules. You can also order the terms according to statistical significance. Remove the tick from `Hierarchical sorting` and restart the analysis.

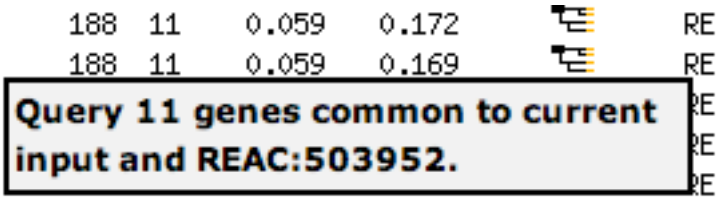
  What is the term with the most significant enrichment? What is the corresponding p-value?
- Only statistically significant functional enrichments are shown in standard output. You can also browse around the list of 1534 functions and pathways that have any overlap with the input gene list. To do so, remove the tick from `Significant only` and restart the analysis. (This may take a while)



- The output probably changed from a picture (PNG) to text, because the list of 1534 functional terms is too large to fit into a figure. Textual output is also useful for other purposes, e.g. search, copy-paste and further automated analysis. Try out the other options under `Output type`[1].



- Tick `Significant only` and `Hierarchical sorting` again to continue browsing the shorter list of functional terms.
- Three columns on the right of the p-value reveal important numbers about the enrichment:
  - Column T : number of all known genes with a specific function T
  - Column Q : number of genes in your list Q (currently 188)
  - Column Q&T : number of genes in your list Q that have function T
- You can click on any[2] number to create a new g:Profiler analysis with the list of genes behind this number. For instance, click on `11` to see which genes have the Reactome function `REAC:503952 ribosomal scanning`.



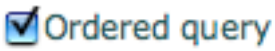
- The two following columns of numbers quantify the sensitivity and specificity of the enrichments:
  - Column Q&T/Q : proportion of genes in your list Q with function T
  - Column Q&T/T : proportion of genes with function T that are covered in your list Q.

  For example, note that more than one third (0.378) of the list's genes are related to the admittedly rather general GO term `GO:0009058 biosynthetic process`, while the list includes about 4.4% of all known yeast genes with such function.
- Clicking on term IDs opens a new browser window and directs you to websites of Gene Ontology, Reactome or KEGG. For instance, click on `GO:0006412` to visit the GO website and read a brief discussion on translation.

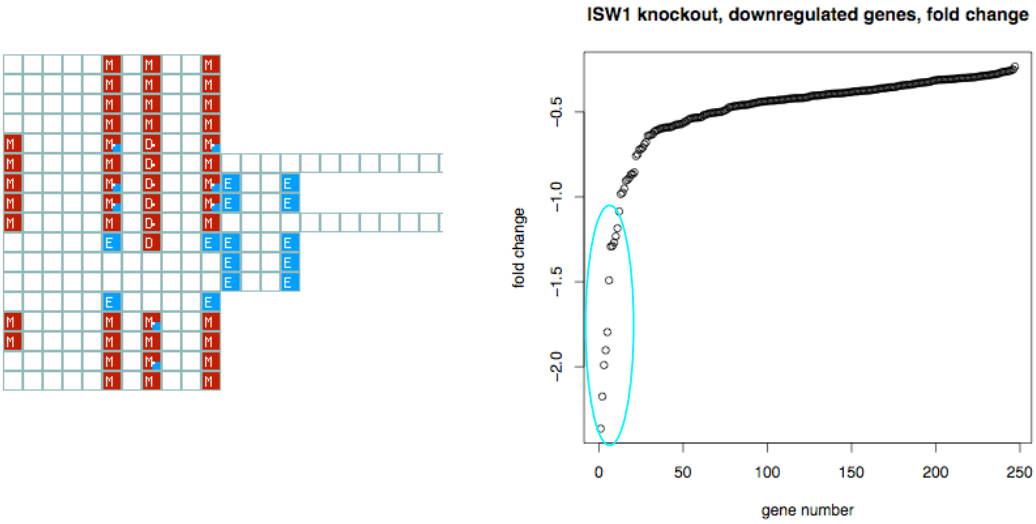
[1] In particular, minimal output can be combined with automatic download tools (wget, curl) to construct a pipeline that analyses any number of gene lists.
[2] Except if the number is greater than 1000 – browser standards restrict lengthy URLs.

g:Profiler – functional analysis of sorted gene lists

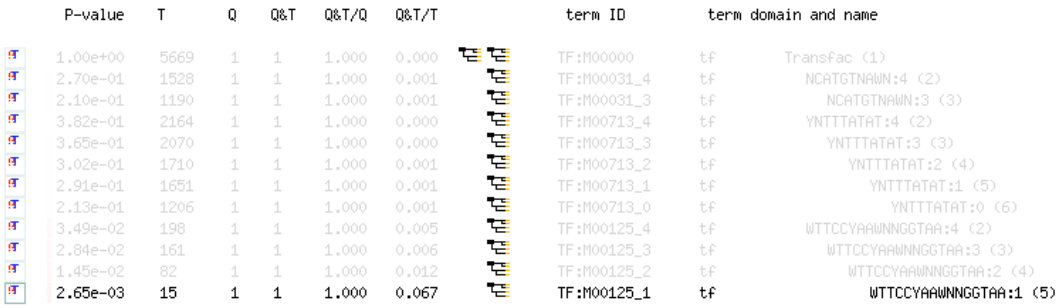
- Open the browser window with online course materials at http://emu.at.mt.ut.ee/u/reimand/roadshow2009/ and copy the contents of `ISW1_down_tgts_ranked_by_fold_change.txt` to g:Profiler query field. It contains 247 yeast genes that are downregulated in ISW1 knockout mutant. Here the genes are ordered according to fold change – the genes with the most dramatic change in expression level in knockout mutant are ranked first.
- Tick the box `Ordered query` and click `g:Profile` to start the analysis.



- Scroll around in the resulting page and observe how functional terms produce rows of varying length. This type of analysis detects the portion of the ordered list that has the best enrichment p-value. For instance, first 11 genes from the top have strong enrichments in terms like `mating` and `response to pheromone`. Numbers of top genes are shown in column Q.



- To see all annotations of a single gene, click the gene name on the top of the image. For example, click on the first gene `YJR004C`.
- Scroll to the bottom of the page to observe transcription factor binding sites from the TRANSFAC database.



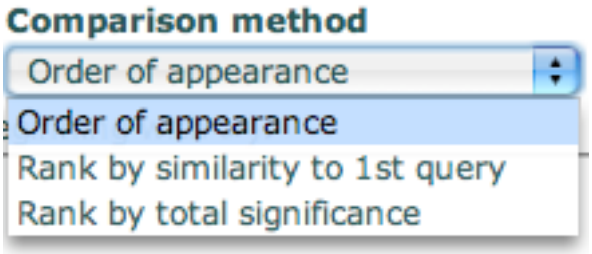
g:Cocoa – functional enrichment analysis of multiple gene lists

- Swich to `g:Cocoa` from the top menu and select S.cerevisiae from the `organism` menu.

- Open the browser window with online course materials at http://emu.at.mt.ut.ee/u/reimand/roadshow2009/ and copy the contents of `yeast_cc_tf_knockout.txt` to the `Gene queries` box. The file contains 11 lists of genes that are differentially expressed in yeast knockout mutants of cell cycle transcription factors SWI4, SWI5, SWI6, MBP1, ACE2, YOX1,STB1, STP1, SOK2, SWI5, FKH2. Note how different lists are separated from one another with the symbol >. List names can be provided for reference.

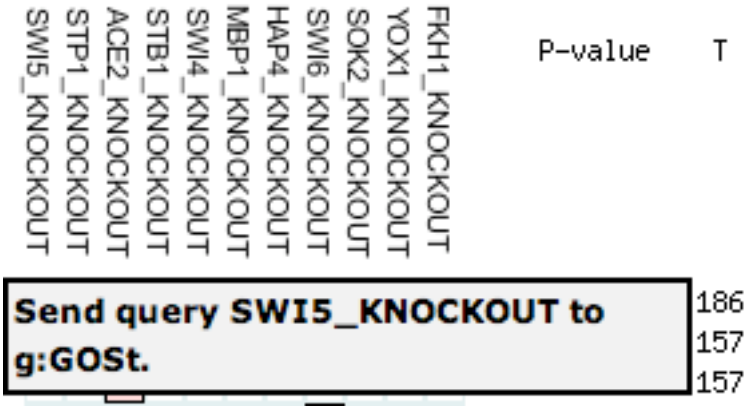
```
> swi4 knockout
YBL113C YBR004C ...
> mbp1 knockout
YLR216C YOL017W ...
> swi6 knockout
YCR038C YDR368W ...
...
```

- Get the list into g:Cocoa and click submit.
- Look around in the following page. Displaying results for multiple lists is similar to single lists. Every coloured box now corresponds to a single list. Shades of pink and red denote statistically significant enrichments between lists and functional terms.
- Find the gene list that has most related statistical significance from enrichments. To do so, select `Order by total significance` from the `Comparison method` menu.



- Which transcription factor knockout affects translation most? Scroll downwards and find the pink block of boxes at the Reactome pathways.

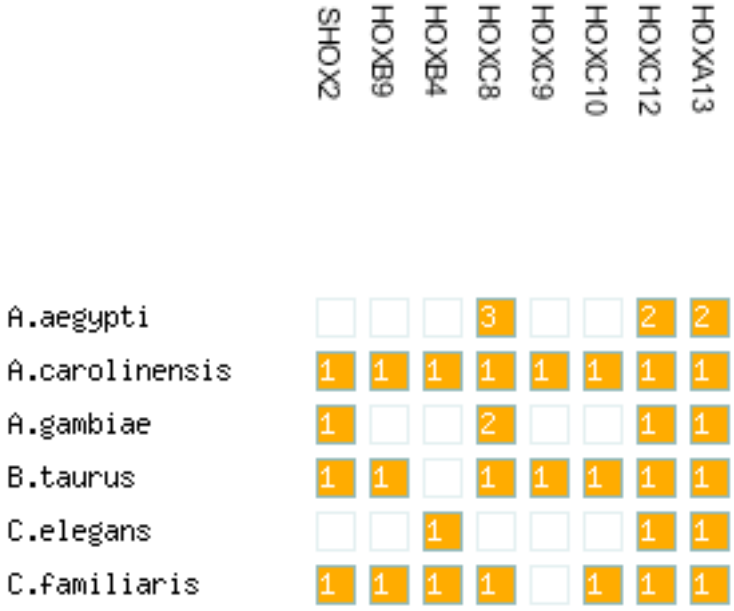
- Click on the TF on the top of the page to send the list to g:Profiler.



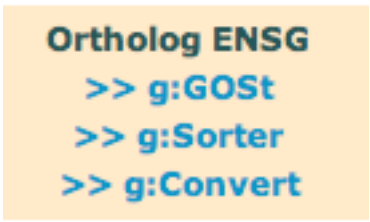
g:Orth – find orthologs and study the conservation of gene lists

- Swich to `g:Orth` from the top menu and select mouse `M.musculus` from the `organism` menu.

- Look at the online course materials at http://emu.at.mt.ut.ee/u/reimand/roadshow2009/ and copy the contents of `hox_clusters_mouse.txt` to the `Query` field. The file includes a list of mouse transcription factors of the HOX family that define basic body layout during early embryonic development.
- Set `target organism` to chicken `Gallus gallus`. Click on `Find orthologs` to find conserved HOX genes in chicken.
- Browse around in the results page. The table of orange boxes shows the conservation of given genes in a variety of species. Numbers in boxes indicate presence of single or multiple orthologous genes in target organism.



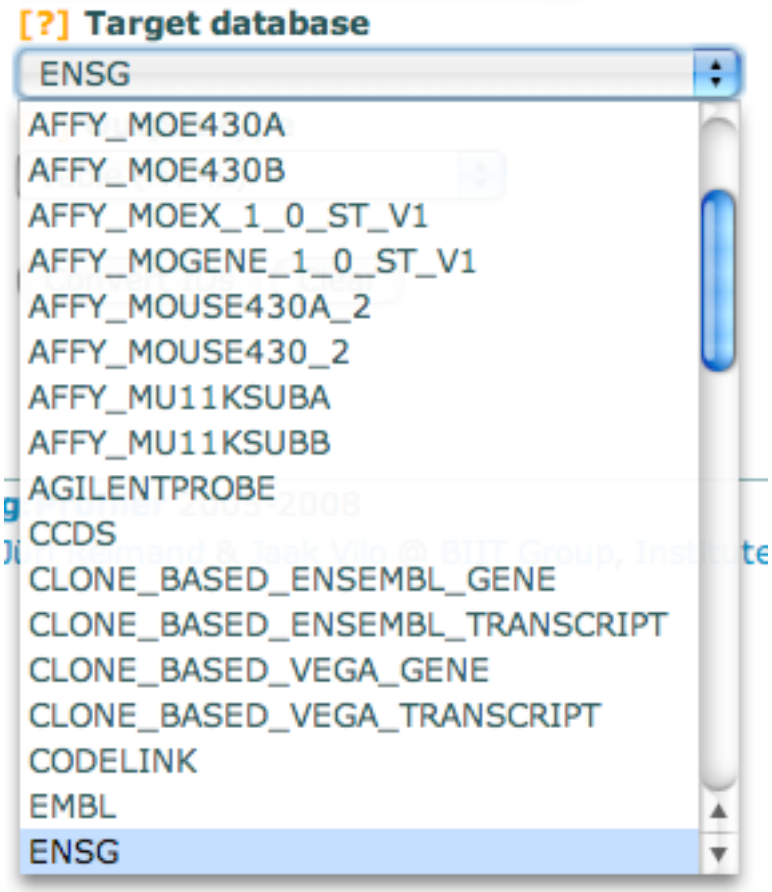
- Below the coloured table, the list of mouse genes and corresponding chicken genes is shown. In order to perform the functional analysis of chicken genes, click on `g:Gost` in the header of `Ortholog ENSG`.



g:Convert – convert names and database IDs of genes, proteins, microarray probesets

- Swich to `g:Convert` from the top menu and select mouse `M.musculus` from the `organism` menu.

- Look at the online course materials at http://emu.at.mt.ut.ee/u/reimand/roadshow2009/ and copy the contents of `hox_clusters_mouse.txt` to the `Query` field.
- Convert the list of HOX genes to Refseq IDs. Set `Target database` to `REFSEQ_DNA` and click on `Convert IDs` to start the process.

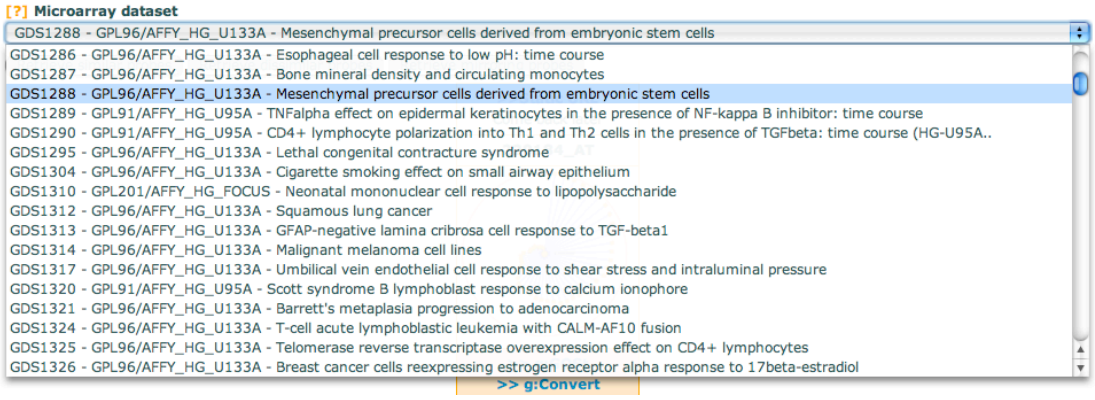


- Convert the HOX genes to Affymetrix probeset IDs `AFFY_MG_U47A`. Copy and paste some of the probeset IDs into your initial list. Observe the fact that g:Convert (as all other g:Profiler tools ☺) accepts a mixture of different IDs and you don't need to define the type of names or IDs you handle.

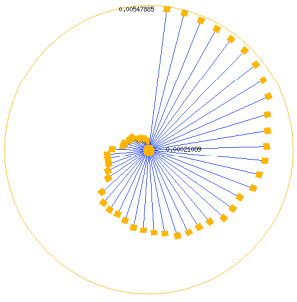
g:Sorter– find similarly expressed genes in microarray datasets

- Swich to `g:Sorter` from the top menu.

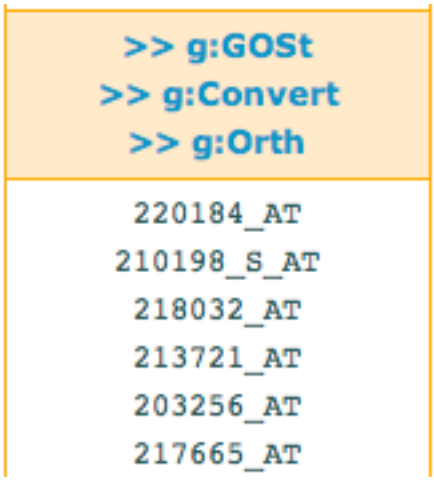
- g:Sorter allows you to search for genes with similar expression patterns in public microarray datasets from Gene Expression Omnibus (GEO).
- The default similarity measure is the commonly used Pearson correlation. By default, 50 most similar microarray probesets are retrieved. There is no need to change these settings.
- Set the `Query` field to `NANOG`. NANOG is one of the three core transcription factors that guide the maintenance of pluripotency and differentiation in embryonic stem cells (ESC). The other two factors are SOX2 and POU5F1 (OCT4).
- Set the following dataset from the `Microarray dataset` dropdown menu. In order to do that efficiently, type the first letters G-D-S-1-2-8-8 rapidly while the menu is open.
    - GDS1288 - GPL96/AFFY_HG_U133A - Mesenchymal precursor cells derived from embryonic stem cells



- Click on `Retrieve similar probes` to start the analysis.
- A list of most similar probesets is soon delivered. Look at the distance starplot to study the relative similarity of probesets. Distance 0 denotes perfect similarity (correlation), while distance 1 denotes dissimilarity (anticorrelation).



- Click on `g:Gost` to send the list of microarray probesets to functional analysis.



- g:Profiler will disregard ambiguous probesets (cases where the same probeset is known to bind several different genes). It also counts duplicate probesets (that represent the same gene) as one, so that the statistics is kept intact.

```
WARNING Gene 37424_AT in query [1/9] is ambiguous (3 gene matches in Ensembl), skipping!
WARNING Gene 208295_X_AT in query [1/31] is ambiguous (3 gene matches in Ensembl), skipping!
```

- In g:Profiler, notice how the list of NANOG neighbours is enriched in functions relevant to embryonic development, e.g. `anatomical structure development` and `nervous system development`.
- Browse towards the bottom of the page to see the list of retrieved genes. Note that the other ES factor SOX2 is present in the sorted list on position 4.

220184_AT	NANOGP8	Homeobox protein NANOG (Homeobox transcription factor Nanog)(hNanog) [Source:UniProtKB/Swiss-Prot;Acc:Q9H9S0]
210198_S_AT	PLP1	Myelin proteolipid protein (PLP)(Lipophilin) [Source:UniProtKB/Swiss-Prot;Acc:P60201]
218032_AT	SNN	Stannin (AG8_1) [Source:UniProtKB/Swiss-Prot;Acc:075324]
213721_AT	SOX2	Transcription factor SOX-2 [Source:UniProtKB/Swiss-Prot;Acc:P48431]
203256_AT	CDH3	Cadherin-3 Precursor (Placental cadherin)(P-cadherin) [Source:UniProtKB/Swiss-Prot;Acc:P22223]

Questions or comments?
Juri.Reimand@ut.ee
http://biit.cs.ut.ee

Publication:
J. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) NAR 35 W193-W200.