bioinformatics.ca

Canadian Bioinformatics Workshops

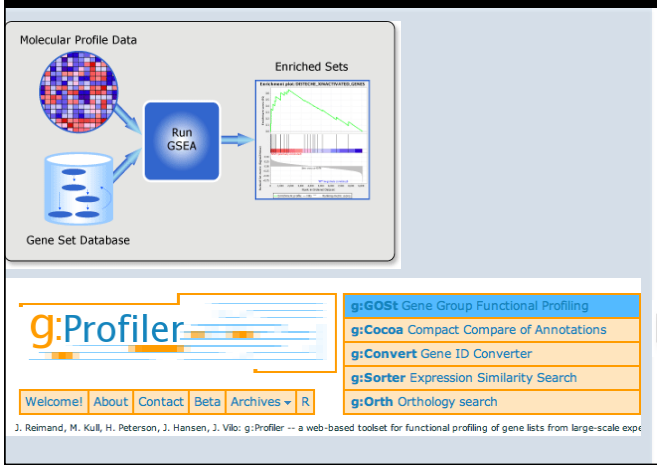
www.bioinformatics.ca

Module 2
Over Representation Analysis
Lab Practical

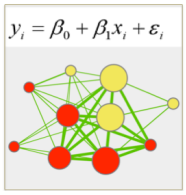
Veronique Voisin

Pathway and Network Analysis of –omics Data
June 1-3, 2015

bioinformatics.ca



http://baderlab.org



Donnelly Centre
for Cellular + Biomolecular Research

UNIVERSITY OF TORONTO

Learning Objectives of Module

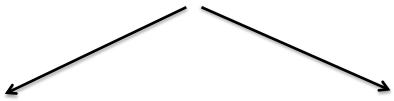
- Be able to run GSEA (Gene Set Enrichment Tool) and understand the main parameters
- Be able to run a simple enrichment tool like g:Profiler and understand the main parameters

DATA-SET

- **Model**
  MCF7 cells, a human breast cancer line, treated or non treated with estradiol.
- **Time points**
  The cells were treated with estradiol for 12, 24 or 48 hours.
- **Protocol**
  Total RNA extracted from the cells was amplified, labeled and hybridized to microarrays.
- **Chip model**
  Affymetrix GeneChip U133 Plus 2.0 microarrays
- **GEO accession number**
  GSE11352

ORA lab

Gene-set enrichment analysis

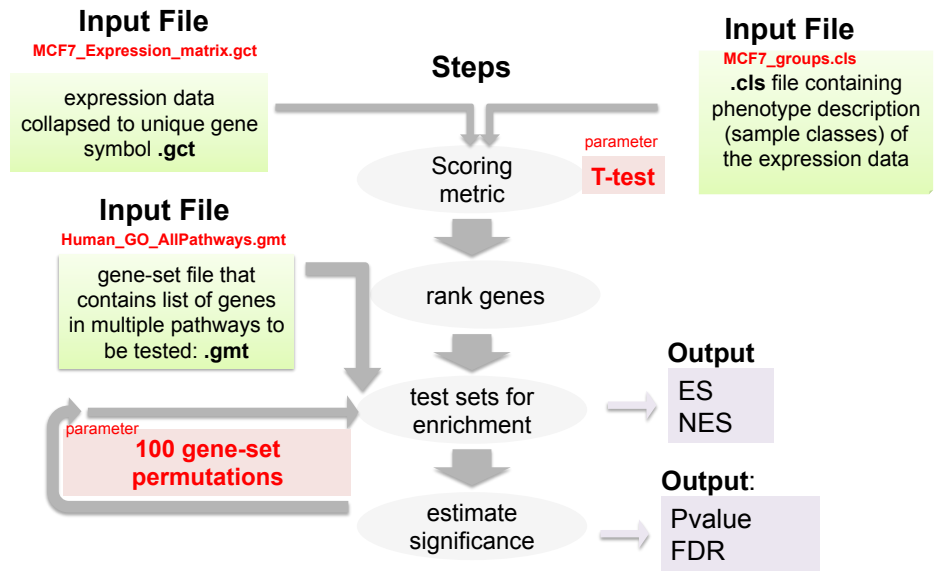


Exercise 1: GSEA

Exercise 2: g:Profiler

Note: GSEA and g:Profiler gene-set enrichment results will also be visualized as a network using the EnrichmentMap application of the Cytoscape software in the next module of the workshop.

Exercise 1: GSEA



(for your information)

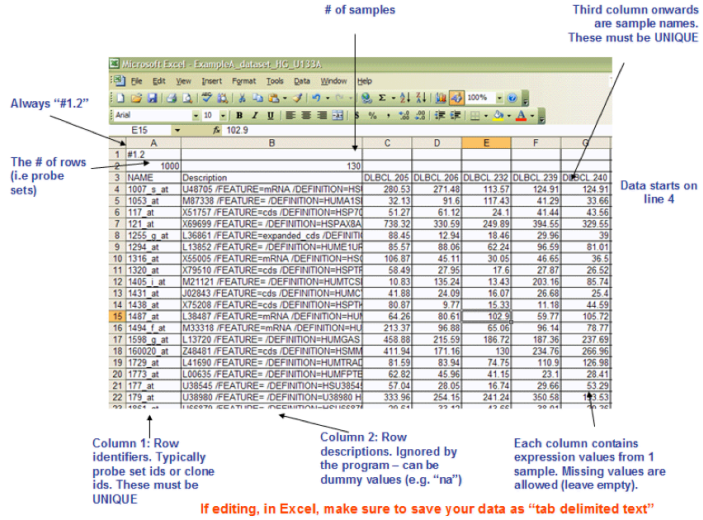
Help on GSEA format

http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

GCT: Gene Cluster Text file format (*.gct)

(for your information)

The GCT format is a tab delimited file format that describes an expression dataset. It is organized as follows:



The **first line** contains the version string and is always the same for this file format. Therefore, the first line must be as follows:

```
#1.2
```

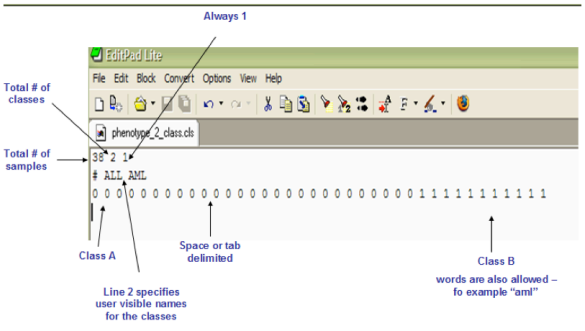
The **second line** contains numbers indicating the size of the data table that is contained in the remainder of the file. Note that the name and desc

CLS: Categorical (e.g tumor vs normal) class file format (*.cls)

(for your information)

The CLS file format defines phenotype (class or template) labels and associates each sample in the expression data with a label. The CLS file format uses spaces or tabs to separate the fields.

The CLS file format differs somewhat depending on whether you are defining categorical or continuous phenotypes. Categorical labels define discrete phenotypes; for example, normal vs tumor. For categorical labels CLS file format is organized as follows:



Example of a 3 class cls file

```
phenotype_2_class.cls   P53.cls
# 3 1
#KRAS_MUT WT MYC_MUT
KMUT KMUT KMUT WT WT WT myc myc myc
```

The **first line** of a CLS file contains numbers indicating the number of samples and number of classes. The number of samples should correspond to the number of samples in the associated RES or GCT data file.

Line format: `(number of samples) (space) (number of classes) (space) 1`

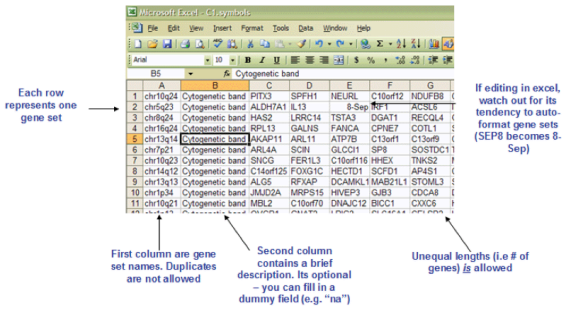
Example: `58 2 1`

The **second line** in a CLS file contains a user-visible name for each class. These are the class names that appear in analysis reports. The line should begin with a pound sign (#) followed by a space.

(for your information)

GMT: Gene Matrix Transposed file format (*.gmt)

The GMT file format is a tab delimited file format that describes gene sets. In the GMT format, each row represents a gene set; in the GMX format, each column represents a gene set. The GMT file format is organized as follows:



GMT format is convenient to store large databases of gene sets. For a handful of sets (<256) the gmx format offers greater excel-editability

Each gene set is described by a name, a description, and the genes in the gene set. GSEA uses the description field to determine what hyperlink to provide in the report for the gene set description: if the description is "na", GSEA provides a link to the named gene set in MSigDB; if the description is a URL, GSEA provides a link to that URL.

Important GSEA Parameters

- permutation type:
  - "phenotype" only if ≥ 7 samples per class are available
  - "gene_set" works also with fewer samples
- collapse only if chip-annotation file is used (probe id and no gene names
- scoring scheme: weighted, change to p2 if noisy data
- metric
  - Ratio_of_Classes ← use with log2 expression data
  - log2_Ratio_of_Classes ← use with linear expression data
  - t-Test
  - Signal2Noise
- Min/Max size of Gene Sets

GSEA : chosen parameters

Gsea: Set parameters and run enrichment tests

Required fields

Expression dataset	MCF7_Expression_matrix [20326x18 (ann: 20326,18,chip na)]
Gene sets database	)RAlab/GSEA_tutorial/GSEATutorial/Human_GO_AllPathways.gmt
Number of permutations	100
Phenotype labels	:SEA_tutorial/GSEATutorial/MCF7_groups.cls#ES12_versus_NT12
Collapse dataset to gene symbols	false
Permutation type	gene_set
Chip platform(s)	

GSEA : chosen parameters

Basic fields Hide

Analysis name	ES12_vs_NT12_CBW
Enrichment statistic	weighted
Metric for ranking genes	tTest
Gene list sorting mode	real
Gene list ordering mode	descending
Max size: exclude larger sets	500
Min size: exclude smaller sets	15
Save results in this folder	/Users/veroniquevoisin/Downloads

GSEA : chosen parameters

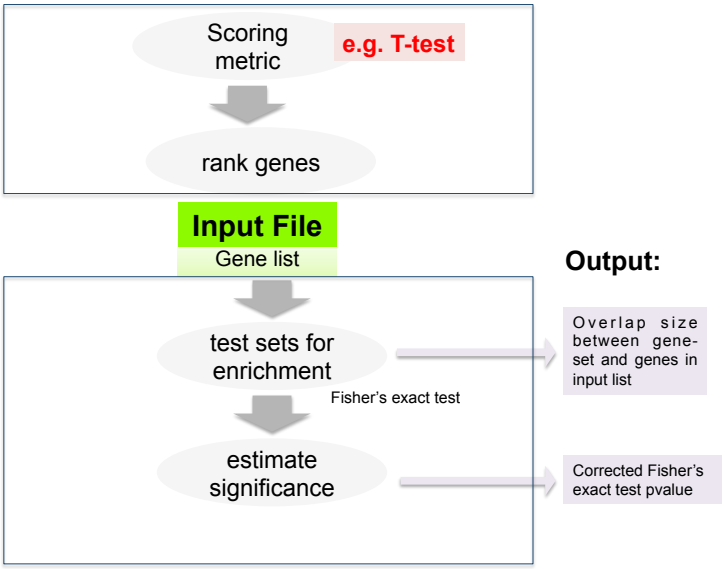
Advanced fields — Hide

Collapsing mode for probe sets => 1 gene	Max_probe
Normalization mode	meandiv
Randomization mode	no_balance
Omit features with no symbol match	true
Make detailed gene set report	true
Median for class metrics	false
Number of markers	100
Plot graphs for the top sets of each phenotype	20
Seed for permutation	timestamp
Save random ranked lists	false
Make a zipped file with all reports	false

Exercise 2: g:Profiler

Steps:

outside g:Profiler



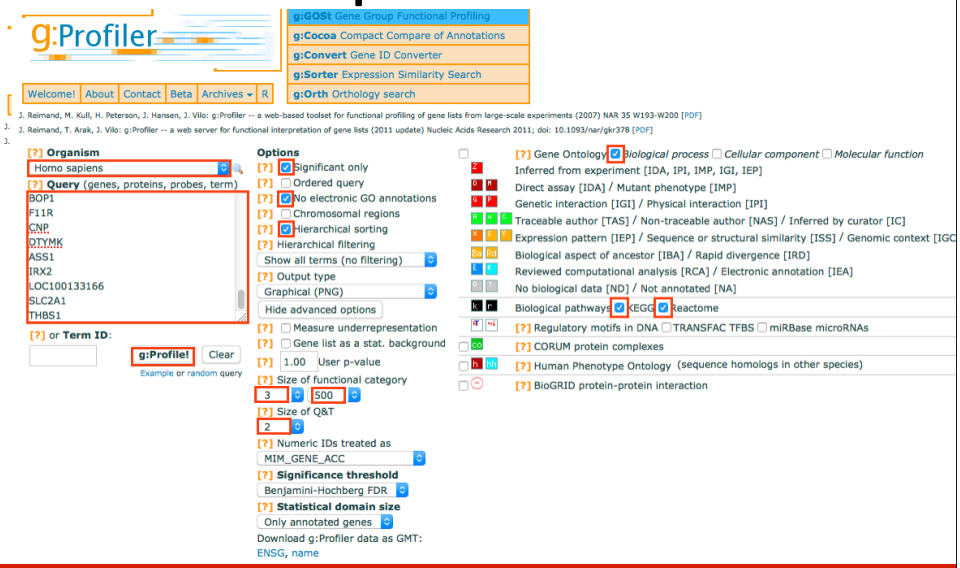
inside g:Profiler

Output:

Overlap size between gene-set and genes in input list

Corrected Fisher's exact test pvalue

g:Profiler chosen parameters



Module 2 bioinformatics.ca

Time to start practical part:

1 hour suggested

- Go the the CBW wiki and download the ORA lab document.
- Download required files on your computer.
- Do the 2 exercises at your own pace and ask teaching assistant for help if required.

Module 2 bioinformatics.ca

Exploring GSEA results

GSEA Report for Dataset MCF7_Expression_matrix

Enrichment in phenotype: ES12 (3 samples)

gene-sets enriched in genes up-regulated in treated cells compared to non-treated samples

- 2120 / 4756 gene sets are upregulated in phenotype **ES12**
- 665 gene sets are significant at FDR < 25%
- 422 gene sets are significantly enriched at nominal pvalue < 1%
- 612 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: NT12 (3 samples)

gene-sets enriched in genes down-regulated in treated cells compared to non-treated samples

- 2636 / 4756 gene sets are upregulated in phenotype **NT12**
- 445 gene sets are significantly enriched at FDR < 25%
- 337 gene sets are significantly enriched at nominal pvalue < 1%
- 601 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 20323 features (genes)
- No probe set => gene symbol collapsing was requested, so all 20323 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 12503 / 17259 gene sets
- The remaining 4756 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the ES12 *versus* NT12 **comparison**

- The dataset has 20323 features (genes)
- # of markers for phenotype **ES12**: 9758 (48.0%) with correlation area 49.7%
- # of markers for phenotype **NT12**: 10565 (52.0%) with correlation area 50.3%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset

Index.html summary of results

- Give the number or significant gene-sets (pathways) for the E12 or the NT12 phenotype.
- Link to the GSEA plots (snapshots)
- Link to the GSEA results as tabular format (html or excel format)

Note: you can access the index.html file using the **'Success 5'** link or locate it in the GSEA folder result.

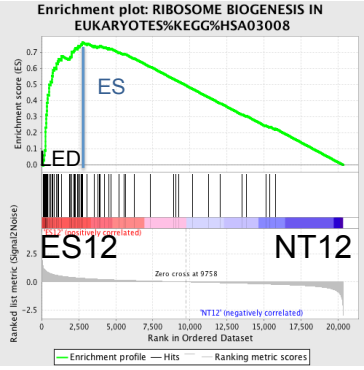
Exploring GSEA Results

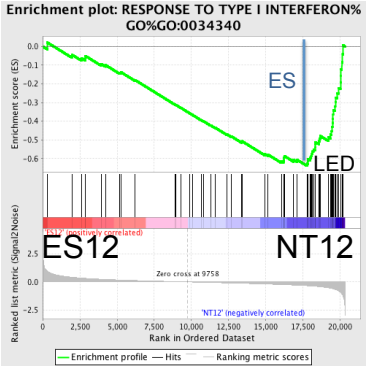
NES FDR

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	RIBOSOME BIOGENESIS IN EUKARYOTES%KEGG%HSA03008	Details ...	69	0.76	2.71	0.000	0.000	0.000	2778	tags=65%, list=14%, signal=75%
2	RIBOSOME BIOGENESIS%GO%GO:0042254	Details ...	61	0.77	2.68	0.000	0.000	0.000	2454	tags=48%, list=12%, signal=54%
3	RRNA PROCESSING%GO%GO:0006364	Details ...	42	0.80	2.64	0.000	0.000	0.000	2438	tags=45%, list=12%, signal=51%
4	NCRNA PROCESSING%GO%GO:0034470	Details ...	86	0.69	2.59	0.000	0.000	0.000	3038	tags=43%, list=15%, signal=50%
5	NCRNA METABOLIC PROCESS%GO%GO:0034660	Details ...	158	0.62	2.53	0.000	0.000	0.000	3311	tags=42%, list=16%, signal=50%
6	RRNA METABOLIC PROCESS%GO%GO:0016072	Details ...	47	0.76	2.52	0.000	0.000	0.000	2438	tags=43%, list=12%, signal=48%
7	RIBONUCLEOPROTEIN COMPLEX BIOGENESIS%GO%GO:0022613	Details ...	123	0.64	2.52	0.000	0.000	0.000	3476	tags=46%, list=17%, signal=55%
8	DNA STRAND ELONGATION%GO%GO:0022616	Details ...	34	0.80	2.50	0.000	0.000	0.000	3149	tags=82%, list=15%, signal=97%

NES: normalized enrichment score
FDR: false discovery rate

Exploring GSEA Results





NES:2.71
FDR:0.0005

NES:-2.34
FDR: 0.0005

ES: enrichment score; NES: normalized enrichment score;
LED: leading edge genes; FDR false discovery rate

A GSEA result folder contains multiple files:

- **Index.html** will guide you to main result file
- The **edb folder** contains the input files filtered by GSEA
- **.rpt file** can be used in EnrichmentMap to built a network
- The main GSEA results are in 2 excel files :
  - **gsea_report**_for_ES12_1401563306908.xls
  - **gsea_report**_for_NT12_1401563306908.xls

quick guide: GSEA scores and results

- **ES (enrichment score):** reflects the degree to which a gene-set is overrepresented at the top or bottom of a ranked list of genes.
- **NES (normalized enrichment score):** NES corrects for differences in ES between gene-sets due to differences in gene-set sizes. It enables to compare the scores of the different tested gene-sets with each other.

NES = actual ES / mean of all ESs obtained from all random permutations for the single gene-set that is being tested

- **nom p-value**: The nominal p value estimates the statistical significance of the enrichment score for a single gene set. The p-value is calculated from the null distribution (all ES obtained from the permutation).

Using gene-set permutation, the null distribution is created by generating, for each permutation, a random gene set the same size as your specified gene set by selecting that number of genes from all of the genes in your expression data set (or pre-ranked list), and then calculating the enrichment score for that randomly selected gene set. The distribution of those enrichment scores across all of the permutations constitutes the null distribution.

- **FDR (false discovery rate)**: corrects for multiple hypothesis testing and enable a more correct comparison of the different tested gene-sets with each other.

note: for a given gene-set S and observed NES, called NES*, FDR is [% of all NES (including permutations) >= NES*] / [% of all observed NES (=NES for all tested gene-sets) >= NES*]

Exploring g:Profiler output

Fisher's exact test

Sizes of gene sets and lists (click starts new query)

Info about processes, pathways (click opens external DB site)

FDR-corrected p-value (q-value)

How confident I can be that 9 genes are included in a specific geneset/ pathway not by random chance only

=

How confident I can be that this gene-set / pathway is significantly enriched in my gene list

Gene Ontology

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Gene Ontology (Biological process)					
BP	regulation of protein kinase B signaling	GO:0051896	82	400	9	3.31e-02
BP	positive regulation of protein kinase B signaling	GO:0051897	56	400	9	1.45e-03
BP	negative regulation of cellular component organization	GO:0051129	385	400	22	2.79e-02
BP	mitochondrial translation	GO:0032543	111	400	11	1.52e-02
BP	mitochondrial translational initiation	GO:0070124	85	400	9	4.38e-02
BP	mitochondrial translational elongation	GO:0070125	85	400	9	4.38e-02
BP	mitochondrial translational termination	GO:0070126	85	400	9	4.38e-02
BP	regulation of microvillus organization	GO:0032530	5	400	3	3.45e-02
BP	regulation of microvillus assembly	GO:0032534	4	400	3	1.40e-02
BP	positive regulation of apoptotic process	GO:0043065	401	400	22	5.00e-02
BP	positive regulation of cell death	GO:0010942	423	400	23	4.07e-02
BP	cellular protein complex disassembly	GO:0043624	236	400	16	3.68e-02
BP	reactive oxygen species metabolic process	GO:0072593	167	400	13	4.04e-02

KEGG

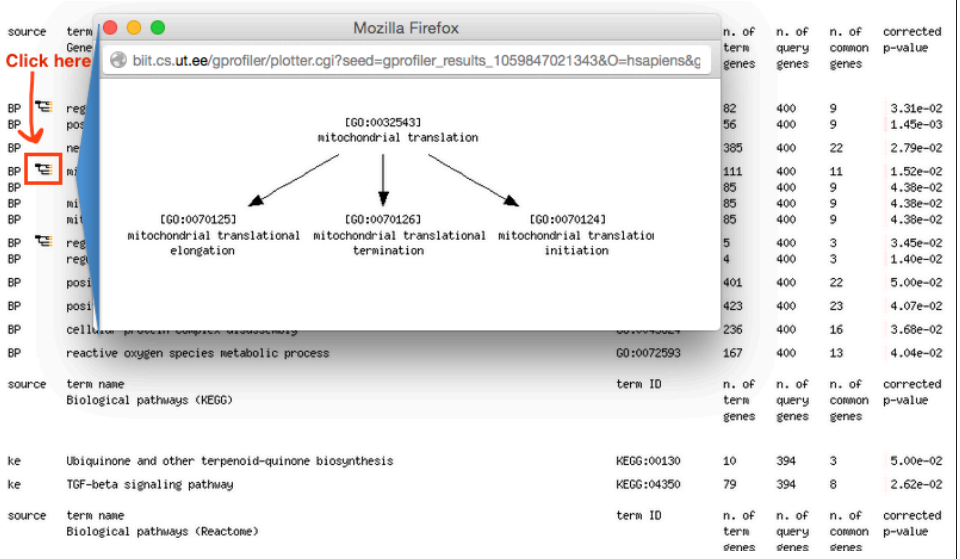
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
ke	Ubiquinone and other terpenoid-quinone biosynthesis	KEGG:00130	10	394	3	5.00e-02
ke	TGF-beta signaling pathway	KEGG:04350	79	394	8	2.62e-02

REACTOME

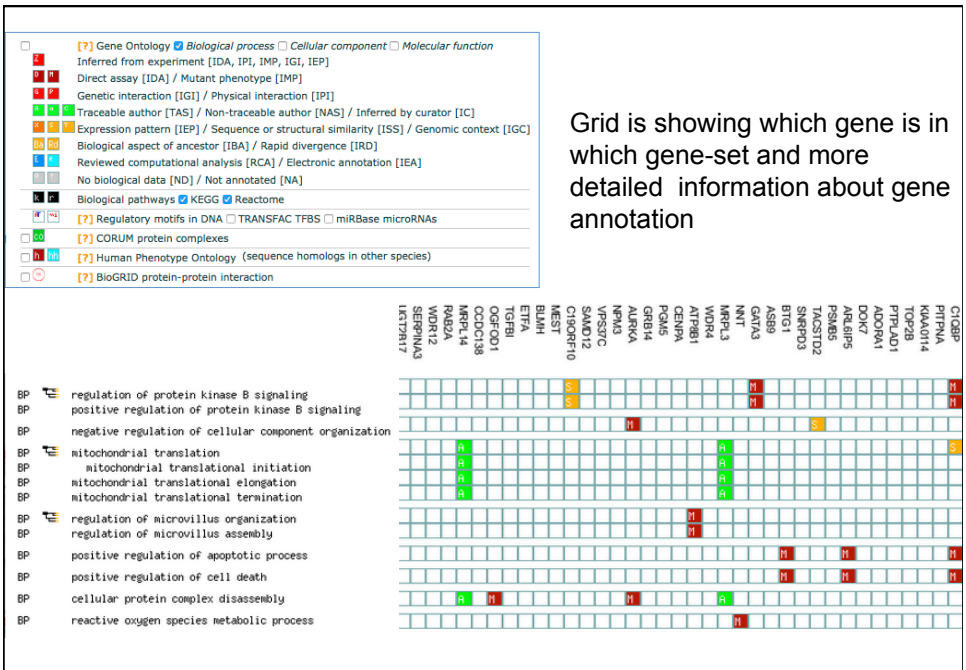
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
re	Mitochondrial translation	REAC:5368287	93	395	9	3.15e-02
re	Mitochondrial translation initiation	REAC:5368286	87	395	9	1.89e-02
re	Mitochondrial translation elongation	REAC:5389840	87	395	9	1.89e-02
re	Mitochondrial translation termination	REAC:5419276	87	395	9	1.89e-02
re	Nuclear signaling by ERBB4	REAC:1251985	39	395	6	2.65e-02
re	binding of TCF/LEF:CTNNB1 to target gene promoters	REAC:4411364	7	395	3	4.18e-02
re	Regulation of mitotic cell cycle	REAC:453276	85	395	9	1.58e-02

NB! g:Profiler switches to text automatically if image too large

Graph shows how processes are related (Gene Ontology and Reactome)

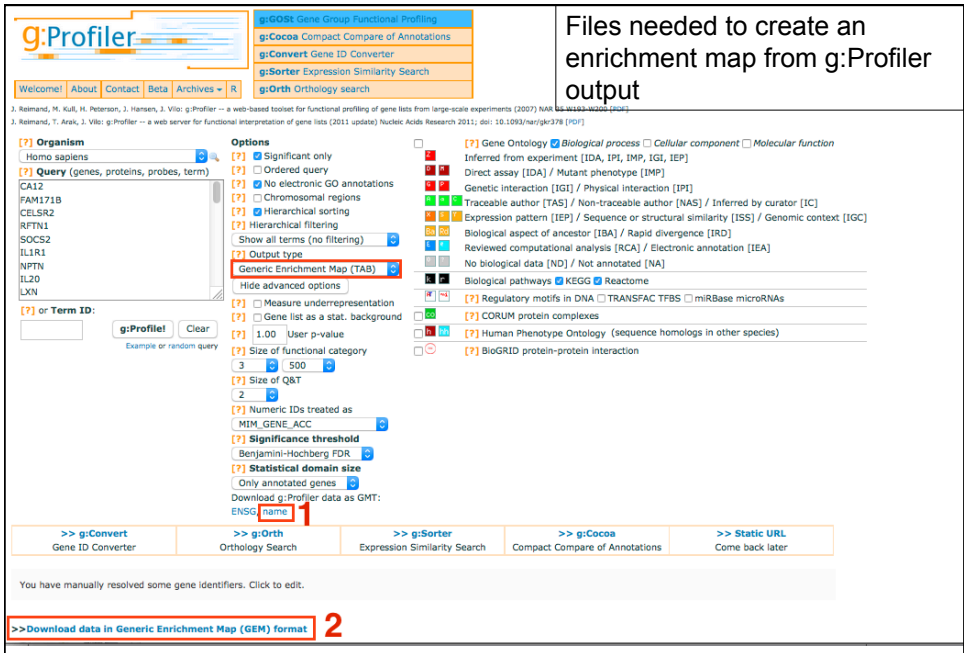


NB! g:Profiler switches to text automatically if image too large

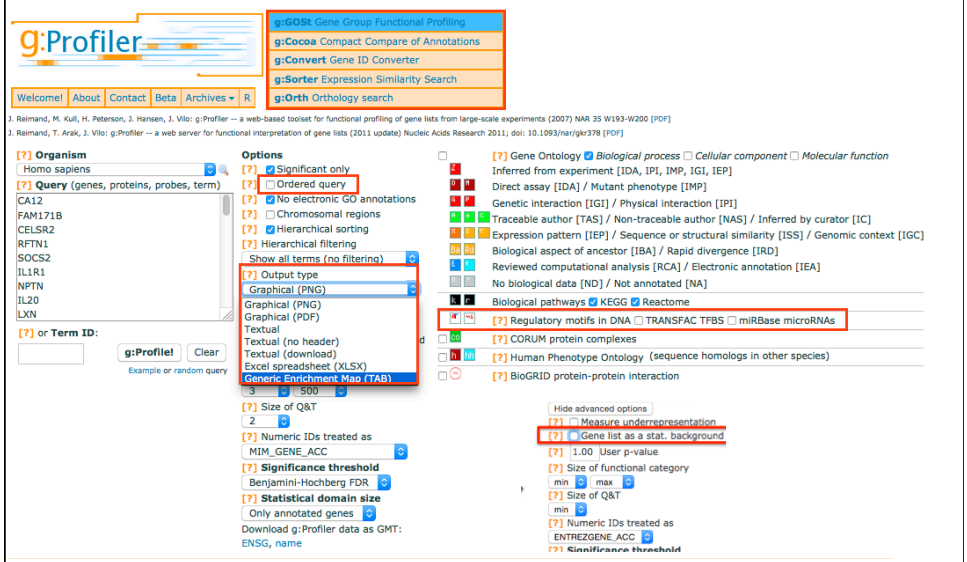


Grid is showing which gene is in which gene-set and more detailed information about gene annotation

Files needed to create an enrichment map from g:Profiler output



Selected g:Profiler additional features



We are on a Coffee Break & Networking Session