



# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](http://bioinformaticsdotca.github.io)

Supported by



Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto  
Castellano Castellano (AR) Español (CL) Español (Ecusador) Castellano (MX) Castellano (PE)  
Euskara Suomi مکالمه français français (CA) Galego മലബാറി മരാറ്റി മന്ത്യിൽ Italiano 日本語 한국어 Macedonian Melayu  
Nederlands Norsk Sesotho sa Leboa polski Português română slovenščiški jezik српски (latinica) Sotho svenska  
中文 草語 (臺灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



**Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:

English French

Learn how to distribute your work using this licence

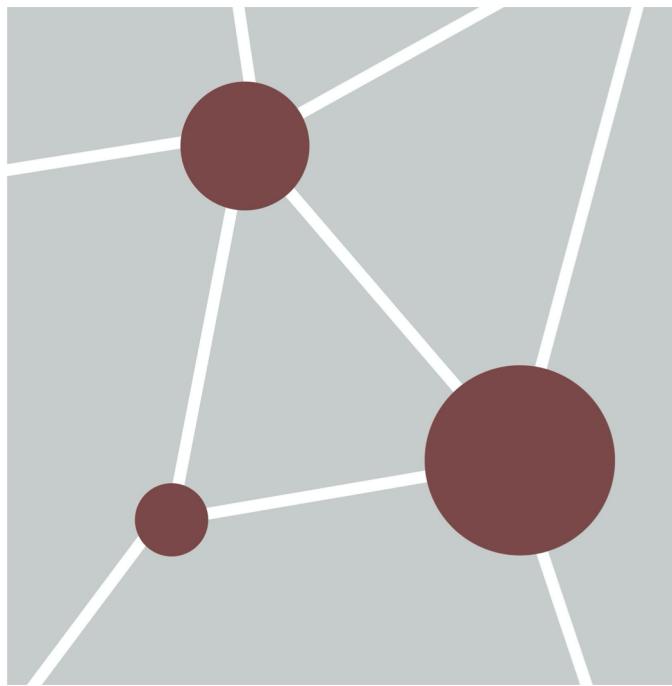
# Finding over-represented pathways in gene lists: practical lab



Ruth Isserlin

Pathway and Network Analysis of -omics Data

July 27-29, 2020



# Learning Objectives of Module

- By the end of this lab, you will:
  - Be able to run a simple enrichment tool like **g:Profiler** using a **gene list** and understand the main parameters and output results.
  - Be able to run **GSEA** (Gene Set Enrichment Tool) on a **ranked gene list** and understand the main parameters and output results.

# Learning Objectives of Module

- Be able to run a simple enrichment tool like **g:Profiler** using a **gene list** and understand the main parameters and output results.
- Be able to run **GSEA** (Gene Set Enrichment Tool) on a **ranked gene list** and understand the main parameters and output results.

# Part 1:



# Part 2:



Characteristics:	g:Profiler	GSEA
Input	gene list (thresholded)	ranked gene list (non thresholded)
Statistics	Fisher's exact test (can upload specific background), minimum hypergeometric test	modified Kolmogorov-Smirnov test
Multiple hypothesis testing correction	yes (FDR, Bonferroni, custom)	yes (FDR)
Pathway databases (gene-sets) (choice/ up to date?)	several databases, can check the ones we are interested in, frequently updated	Several choices from MSigDB from GSEA or upload custom ones. <a href="#">link to Baderlab gene-sets</a> both frequently updated
Model organisms	multiple, directly from Ensembl	mostly human through MSigDB but compatible with any model organisms using the custom upload function.
Output	Graphic image or table and compatible with Cytoscape/EnrichmentMap	Table and Compatible with Cytoscape/EnrichmentMap
Software type	Website and R package	Standalone (java) / or can be called and run from command line

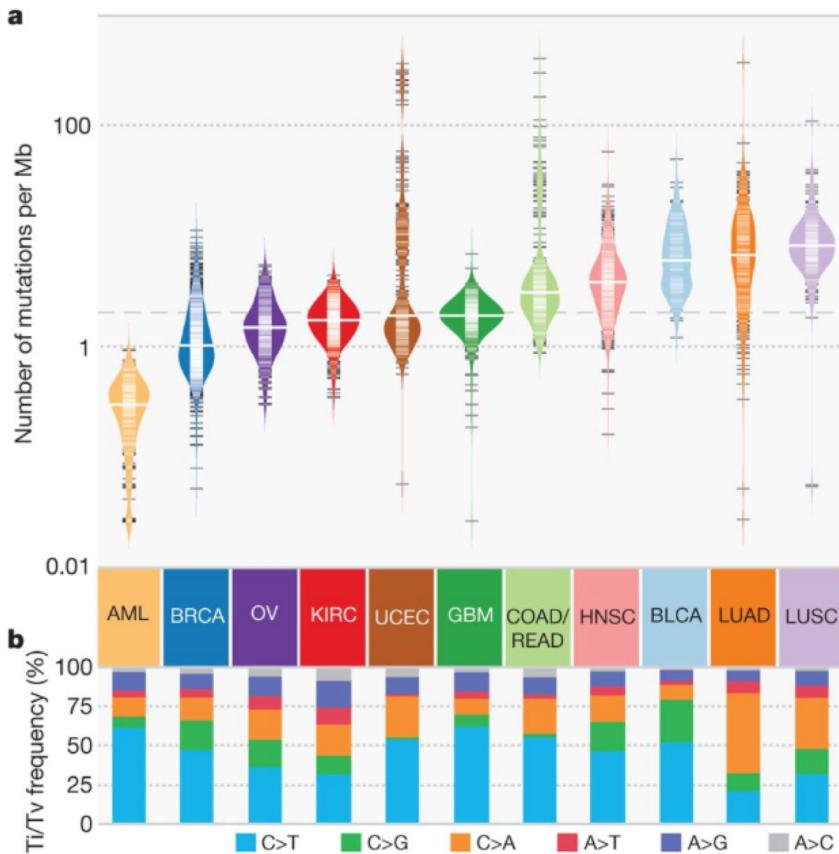
# Part 1:



The logo for g:Profiler features the word "g:Profiler" in a bold, sans-serif font. The "g:" is orange, while "Profiler" is blue. The letter "g" has a small orange square at its top left corner. Below the main text are two horizontal bars: a shorter one above and a longer one below, both composed of several colored segments (orange, yellow, light blue, white) of varying widths.

# Data used for practical lab:

## Dataset: Mutational landscape and significance across 12 major cancer types



Exome sequencing  
Tumor samples and  
matched control tissues

Detection of points mutations and small  
insertions/deletions: somatic variant calls in each  
cancer type and in each tumor

Calculation of mutation frequency: genes  
mutated in at least 5% of tumors were  
selected

Genes positively correlated with number  
of mutation per sample

127 'significantly mutated genes'

gene list

<https://www.nature.com/articles/nature12634> (2013)

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes 

EEF1A  
ACVR2A  
MECOM  
LIFR  
SMC3  
NCOR1  
RPL5  
SMAD2  
SPOP  
AXIN2  
MIR142  
RAD21  
ERCC2  
CDKN2C  
EZH2  
PCBP1

**Run query**

random

example

## gene list

## gene sets

**g:GOST** performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. We regularly retrieve data from Ensembl database and fungi, plants or metazoa specific versions of Ensembl Genomes, and parasite specific data from WormBase Par-

## Options

Organism: 

Homo sapiens (Human)

 Ordered query  Run as multiquery 

ranked gene list:  
minimum hypergeometric test

Advanced options  All results  Measure underrepresentation Statistical domain scope Only annotated genes 

background

Significance threshold Benjamini-Hochberg FDR 

multi hypothesis testing

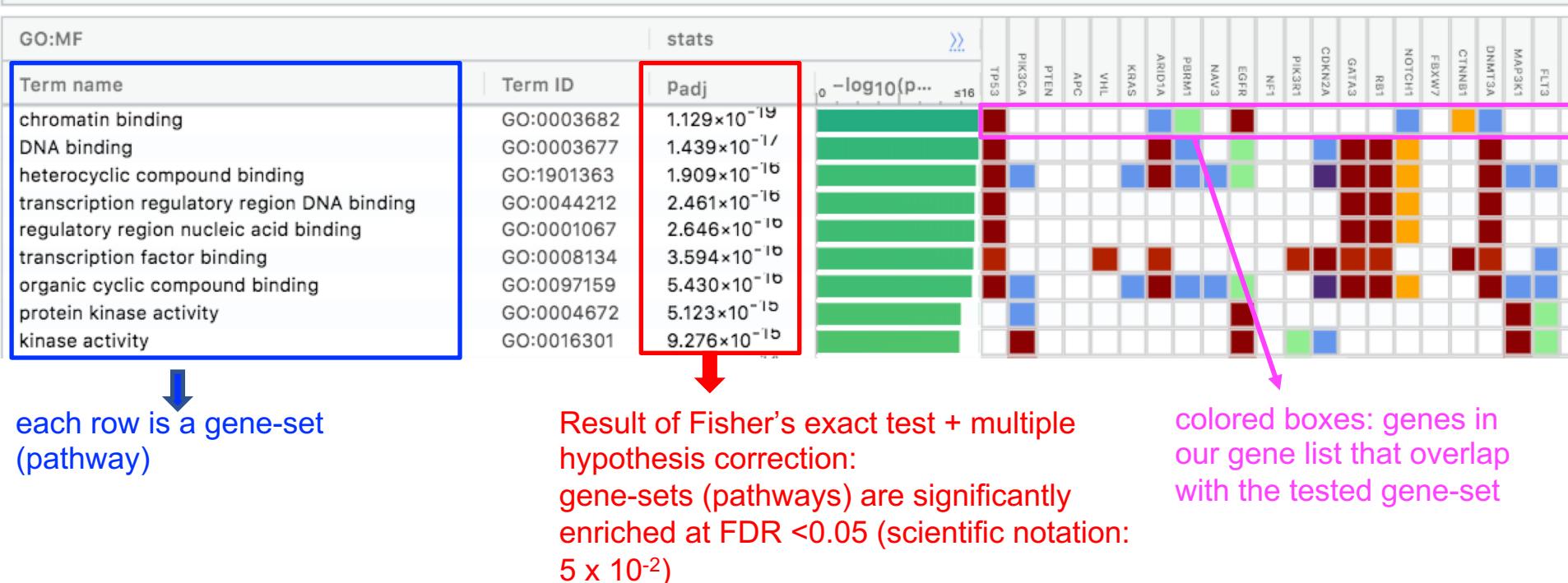
User threshold 

0.05

Numeric IDs treated as ENTREZGENE\_ACC Data sources Custom GMT 

aSite. In addition to Gene Ontology, we include pathways from KEGG Reactome and WikiPathways; miRNA targets from miRTarBase and regulatory motif matches from TRANSFAC; tissue specificity from Human Protein Atlas; protein complexes from CORUM and human disease phenotypes from Human Phenotype Ontology. g:GOST supports close to 500 organisms and accepts hundreds of identifier types.

# Explore results



Note: observe that same genes are included in several enriched gene-sets (pathways).



Time to start practical part:



- Go to the CBW course page and go to module 2.
- Open the 'Lab practical part 1 (g:Profiler)' document.
- Download required files on your computer.
- Do the exercise at your own pace and ask teaching assistants for help or questions.

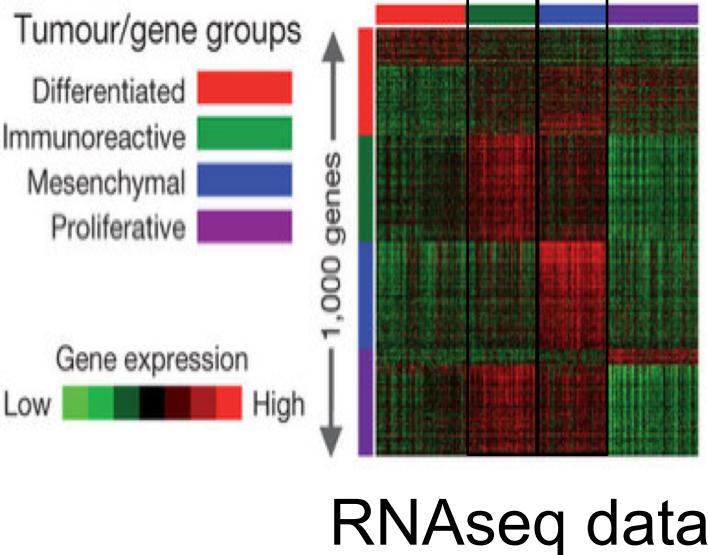
# Part 2:



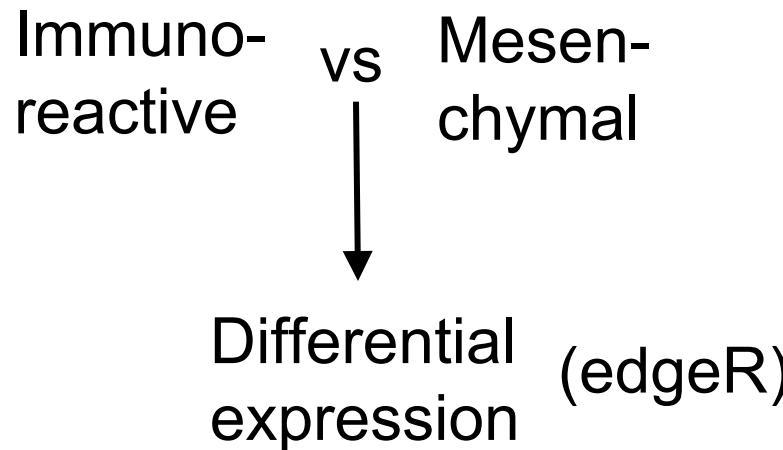
# Data used for practical lab: RNAseq workflow

## Dataset

### Ovarian cancer (TCGA)



Integrated genomic analyses of ovarian carcinoma, PMID:21720365



Rank file

Gene-Set  
Enrichment  
Analysis  
(GSEA)

# Which files do we need to run GSEA?

- A **ranked list of genes** called the rank file
  - this is a text file (tab separated) that should be renamed to end with the extension .rnk
  - This file has 2 columns :
    - gene identifier
    - ranking values
- A file called a **.gmt** file that contains **the pathway data base (the gene-sets)**
  - this is a text file (tab separated) that should end with the extension .gmt
  - the first column contains gene-set names and the additional columns contains the gene names included in each gene-set

# How to generate the rank file

genenames	logFC	logCPM	PValue	FDR
BGN	1.75	9.05	1.73E-33	2.50E-29
ANTXR1	1.55	7.50	4.39E-31	3.18E-27
FZD1	1.28	5.52	4.41E-30	2.13E-26
COL16A1	1.62	5.09	1.33E-29	4.81E-26
KLF3	0.13	6.37	8.32E-02	2.04E-01
RASEF	0.02	2.38	9.01E-01	9.49E-01
ISOC1	0.01	5.24	9.01E-01	9.50E-01
ANO1	0.03	4.93	9.02E-01	9.50E-01
CBWD3	-0.27	3.74	8.18E-02	2.02E-01
GBP4	-1.67	6.63	2.45E-16	2.57E-14
TAP1	-1.40	7.80	1.04E-19	2.38E-17
PSMB9	-1.55	6.52	1.84E-20	5.12E-18

edgeR output

gene name	score
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.61
TAP1	-18.98
PSMB9	-19.73

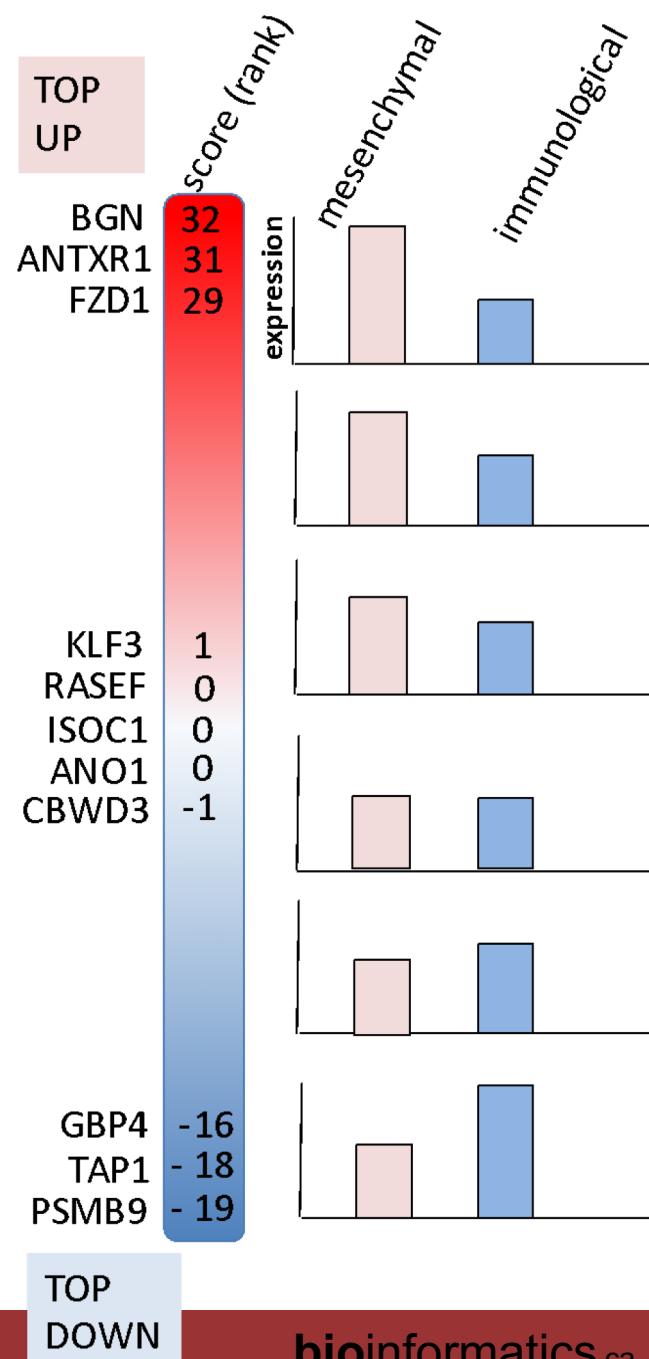
1. Calculate the ranking score:

Using Excel:  
`=SIGN(logFC)*-LOG10(pvalue)`

Using R:  
`sign(logFC)*-log10(pvalue)`

2. Save the file as a **tab** delimited text and with the extension **.rnk**

3. Do keep all genes in the rank files (e.g. 15,000 genes) !  
Do not remove non significant ones.



# ranked list (.rnk)

gene name	score
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.61
TAP1	-18.98
PSMB9	-19.73

Save the file as a **tab** delimited text and with the extension **.rnk**

Do keep all genes in the rank files (e.g. 15,000 genes) !  
Do not remove non significant ones.

# What does a .gmt file look like?

Gene-set name

MOLYBDENUM COFACTOR BIOSYNTHESIS%HUMANCYC%PWY-6823

GLYCEROL DEGRADATION I%HUMANCYC%PWY-4261

OXIDATIVE ETHANOL DEGRADATION III%HUMANCYC%PWY66-161

TETRAPYRROLE BIOSYNTHESIS II%HUMANCYC%PWY-5189

Gene-set name

molybdenum cofactor biosynthesis

glycerol degradation I

oxidative ethanol degradation III

tetrapyrrole biosynthesis I

gene

gene

gene

gene

gene

gene

gene

gene

NFS1

MOCS2

GPHN

MOCS3

GK5

GK

GK2

CYP2E1

ACSS2

ACSS3

ALDH3A2

ACSS1

ALDH2

ALAS2

ALAD

UROS

HMBS

ALAS1

\* Save as tab delimited text with extension .gmt

# Where to find a .gmt file?

If your model organism is *Homo sapiens*, you don't need to create your own:

- you can use directly the MSigDB within GSEA
- you can use the Baderlab gene-set file which is a frequently updated .gmt file which gathers public Gene Ontology and pathways from different sources.

If your model organism is *Mus musculus*:

- you can use the Baderlab gene-set file

If your model organism is different and you need to run GSEA:

- get (access or download) the Gene ontology database directly from biomart / Ensembl and parse it as a .gmt file (see last slide for example code).

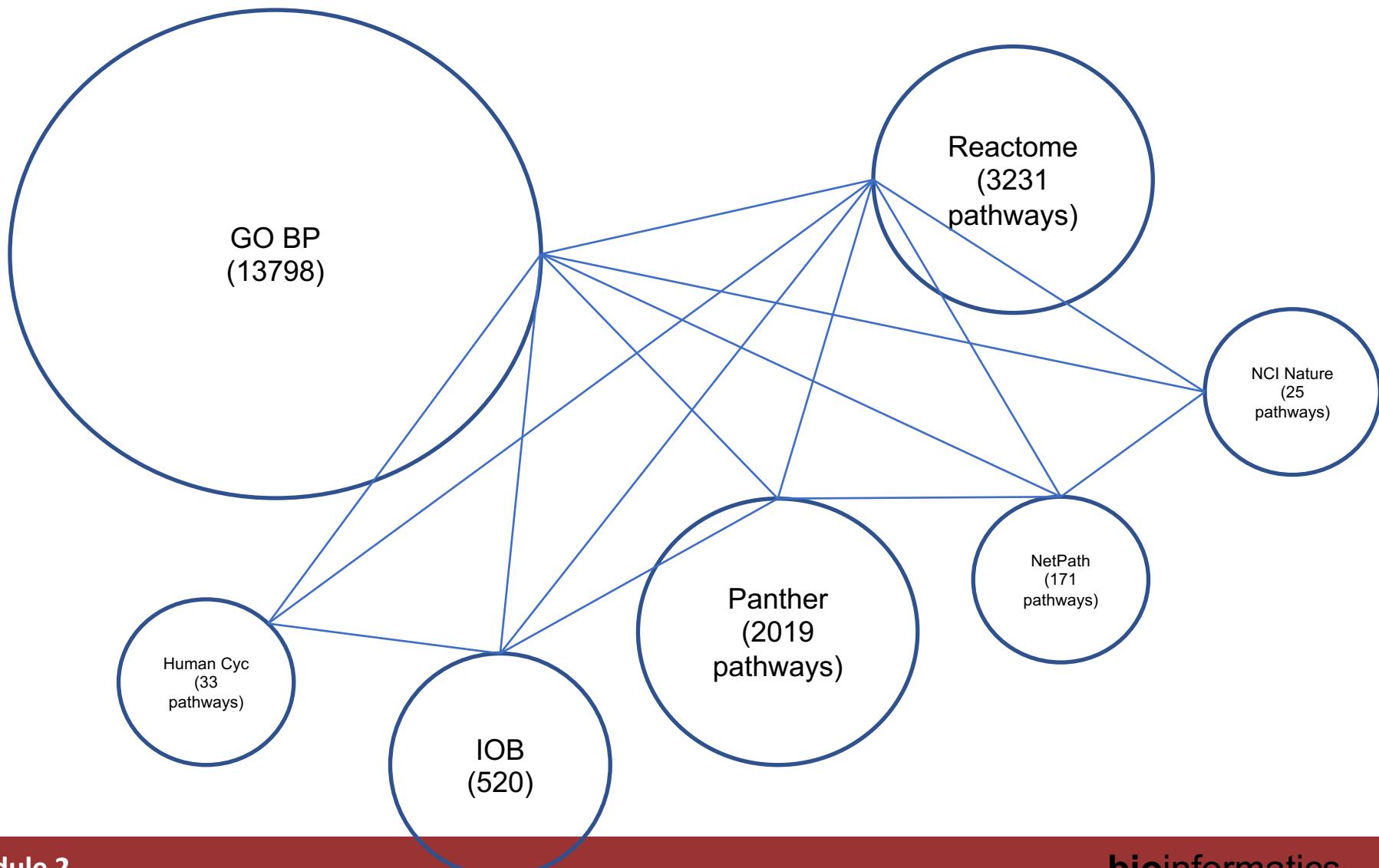
# MSigDB database

<https://software.broadinstitute.org/gsea/msigdb/>

<b>C2: curated gene sets</b> <a href="#">(browse 4738 gene sets)</a>	Gene sets curated from various sources such as online pathway databases, the biomedical literature, and knowledge of domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into two sub-collections: CGP and CP. <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">gene symbols</a> <a href="#">entrez genes ids</a>
CP:REACTOME: Reactome gene sets <a href="#">(browse 674 gene sets)</a>	Gene sets derived from the Reactome pathway database.	<a href="#">Download GMT Files</a> <a href="#">gene symbols</a> <a href="#">entrez genes ids</a>
<b>C5: GO gene sets</b> <a href="#">(browse 5917 gene sets)</a>	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: BP, CC, and MF. <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">gene symbols</a> <a href="#">entrez genes ids</a>
BP: GO biological process <a href="#">(browse 4436 gene sets)</a>	Gene sets derived from the GO Biological Process Ontology.	<a href="#">Download GMT Files</a> <a href="#">gene symbols</a> <a href="#">entrez genes ids</a>
<b>H: hallmark gene sets</b> <a href="#">(browse 50 gene sets)</a>	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. <a href="#">details</a>	<a href="#">Download GMT Files</a> <a href="#">gene symbols</a> <a href="#">entrez genes ids</a>

# BaderLab EM\_Genesets

[http://download.baderlab.org/EM\\_Genesets/](http://download.baderlab.org/EM_Genesets/)



# BaderLab EM\_Genesets

- go to [http://download.baderlab.org/EM\\_Genesets/](http://download.baderlab.org/EM_Genesets/)
  - select current release/
    - Human/
      - symbol/
        - save the Human\_GOPP\_AllPathways\_no\_GO\_iea....gmt file on your computer (right click on the link to save it)

## Index of /EM\_Genesets/current\_release/Human/symbol

Name	Last modified	Size	Description
<a href="#">Parent Directory</a>			-
<a href="#">symbol translation summary.log</a>	2017-03-02 06:37	399	
<a href="#">Human GOBP AllPathways no GO iea March 01 2017 symbol.gmt</a>	2017-03-02 06:37	7.5M	
<a href="#">Human GOBP AllPathways with GO iea March 01 2017 symbol.gmt</a>	2017-03-02 06:37	10M	
<a href="#">Human GO AllPathways no GO iea March 01 2017 symbol.gmt</a>	2017-03-02 06:37	11M	
<a href="#">Human GO AllPathways with GO iea March 01 2017 symbol.gmt</a>	2017-03-02 06:36	14M	
<a href="#">Human AllPathways March 01 2017 symbol.gmt</a>	2017-03-02 06:36	1.2M	
<a href="#">Misc/</a>	2017-03-02 06:36	-	
<a href="#">DrugTargets/</a>	2017-03-02 06:36	-	
<a href="#">DiseasePhenotypes/</a>	2017-03-02 06:36	-	
<a href="#">TranscriptionFactors/</a>	2017-03-02 06:36	-	
<a href="#">miRs/</a>	2017-03-02 06:36	-	
<a href="#">Pathways/</a>	2017-03-02 06:36	-	
<a href="#">GO/</a>	2017-03-02 06:36	-	

# Create a .gmt file for your model organism

Example in R (needs to be adapted to each case and needs):

You need to install and load the biomaRt library

Here is an example using horse.

```
library(biomaRt)

### get annotations
mart=useMart(biomart="ensembl",dataset="ecaballus_gene_ensembl")

go_annotation <-
getBM(attributes=c("uniprot_gn","ensembl_gene_id","ensembl_transcript_id","go_id","name_1006","namespace_1003","go_linkage_type"),filters=list(biotype
='protein_coding'),mart=mart);

go_annotation_bp <- go_annotation[which(go_annotation$namespace_1003=="biological_process"),]
head(go_annotation_bp)

##create gmt
go_pathway_sets <- aggregate(go_annotation_bp[1],by=list(go_annotation_bp$go_id),FUN=function(x){list(unique(x))})

m = match( go_pathway_sets[,1], go_annotation_bp$go_id)
go_pathway_names <- go_annotation_bp$name_1006[m]

### write the gmt

fname = "gobp_horse.gmt"
object = go_pathway_sets[,2]
for ( e in 1: length(object) ){

    write.table( t(c(go_pathway_sets[e,1],
go_pathway_names[e],object[[e]])),sep="\t",quote=FALSE,file=fname,append=TRUE,col.names=FALSE,row.names=FALSE)

}
```

# GSEA preranked

GSEA 3.0 (Gene set enrichment analysis)

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis
- Enrichment Map Visualization

Tools

- Run GSEAPreranked
- Collapse Dataset
- Chip2Chip mapping

GSEA reports

Processes: click 'status' field for results

Name	Status

Show results folder

Home | Load data | Run Gsea on a Pre-Ranked gene list

GseaPreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database .gmt  
Number of permutations 1000

Ranked List .rnk

Basic fields

Analysis name my\_analysis  
Enrichment statistic weighted

Max size: exclude larger sets 500  
Min size: exclude smaller sets 15  
Save results in this folder /Users/veroniquevoisin/gsea\_home/output/jun14

Advanced fields

Each gene-set will be permuted 1000 with random genes to build the null distribution

weighted = p1  
weighted = p1.5  
weighted = p2  
classic weight = 0

Reset Last Command Run

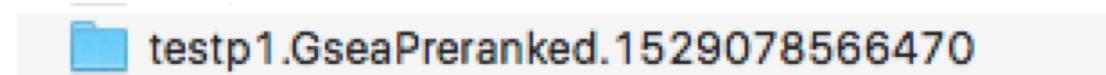
7:29:13 PM | 9193 [INFO] Made Vdb dir: /Users/veroniquevoisin/gsea\_home/output/jun14 | 46M of 619M

# Exploring GSEA results

# How to access GSEA results?

The screenshot shows the GSEA software interface. On the left, there's a sidebar with 'Tools' containing 'Run GSEAPreranked', 'Collapse Dataset', 'Chip2Chip mapping', and 'Analysis history'. The 'Analysis history' section is expanded, showing a table of processes. The table has columns 'Name' and 'Status'. A red arrow points from the text 'Processes: click 'status' field for results' to the 'Status' column. The table data is as follows:

	Name	Status
17	GseaPreranked	Success
18	GseaPreranked	Success
19	GseaPreranked	Success
20	GseaPreranked	Success
21	GseaPreranked	Success



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\_pos\_1401563306908.xls**
  - **gsea\_report\_for\_neg\_1401563306908.xls**

## 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 of significant gene-sets (pathwaysLink 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

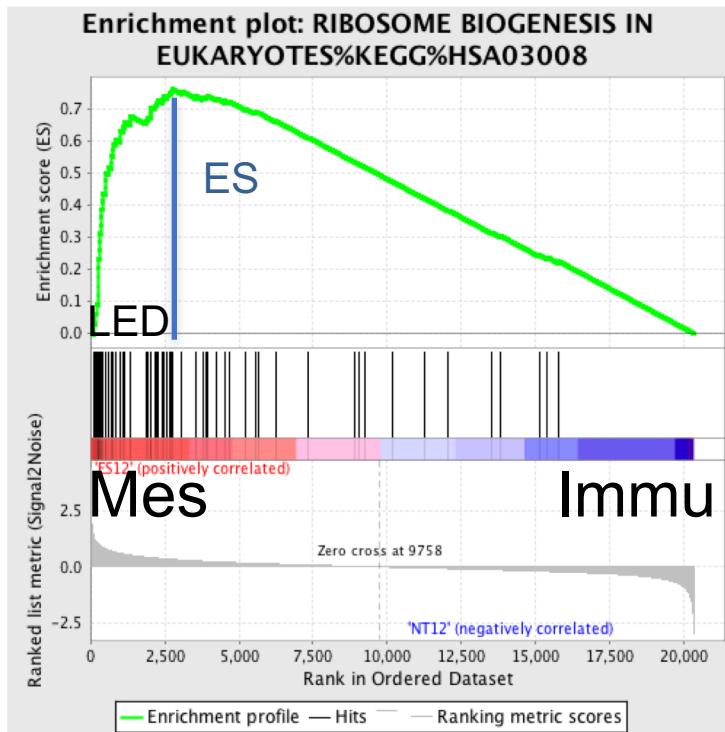
NES FDR

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	POWER p-val	RANK AT MAX	LEADING EDGE
1	RIBOSOME BIOGENESIS IN EUKARYOTES%KEGG%HSA03008	<a href="#">Details ...</a>	69	0.76	2.71	0.000	0.000	0.000	2778	tags=65%, list=14%, signal=75%
2	RIBOSOME BIOGENESIS%GO%GO:0042254	<a href="#">Details ...</a>	61	0.77	2.68	0.000	0.000	0.000	2454	tags=48%, list=12%, signal=54%
3	RRNA PROCESSING%GO%GO:0006364	<a href="#">Details ...</a>	42	0.80	2.64	0.000	0.000	0.000	2438	tags=45%, list=12%, signal=51%
4	NCRNA PROCESSING%GO%GO:0034470	<a href="#">Details ...</a>	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	<a href="#">Details ...</a>	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	<a href="#">Details ...</a>	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	<a href="#">Details ...</a>	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	<a href="#">Details ...</a>	34	0.80	2.50	0.000	0.000	0.000	3149	tags=82%, list=15%, signal=97%

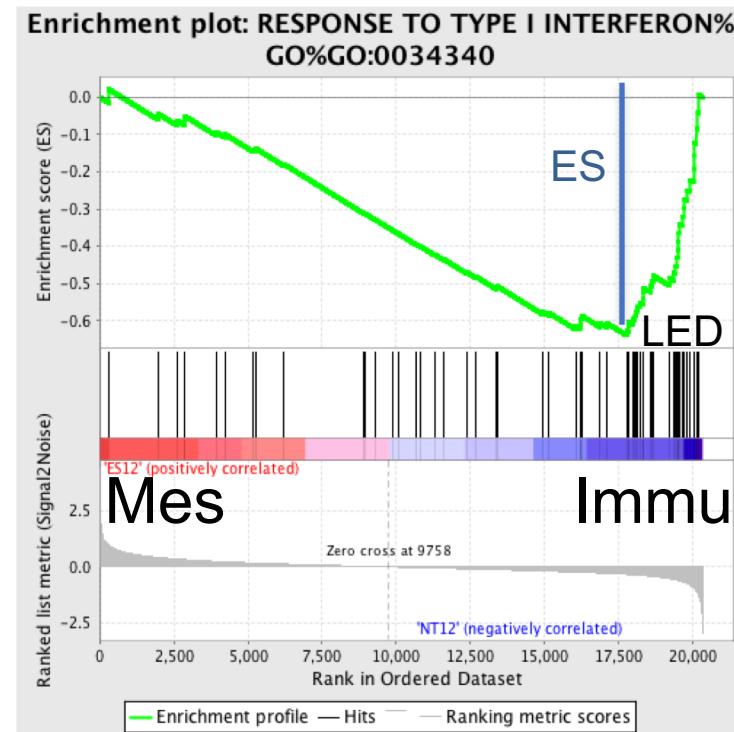
NES: normalized enrichment score  
FDR: false discovery rate

Excel tables are going to be  
exported and uploaded in  
Cytoscape/EM (module 3)

# 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



Time to start practical part:



- Go to the CBW course page.
- Download or open the Module 2 Lab practical documents.
- Download required files on your computer.
- Do the exercise at your own pace and ask teaching assistant for help or questions.

# Links to more tutorials

Step by Step Protocol: Pathway enrichment analysis of -omics data:

<https://www.nature.com/articles/s41596-018-0103-9>

Notebooks of the protocol:

[https://github.com/BaderLab/Cytoscape\\_workflows/tree/master/EnrichmentMapPipeline](https://github.com/BaderLab/Cytoscape_workflows/tree/master/EnrichmentMapPipeline)

# We are on a Coffee Break & Networking Session

**compute | calcul**  
canada | canada



Workshop Sponsors:

Canadian Centre for  
Computational  
Genomics

**MicM** McGill initiative in  
Computational Medicine