bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

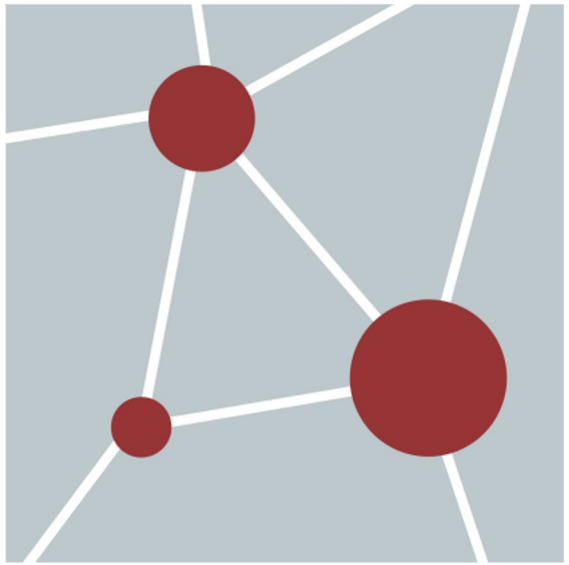
bioinformaticsdotca.github.io

Module 2: Finding over-represented pathways in gene lists lab

Ruth Isserlin

Pathway and Network Analysis

June 26-28, 2024



BADER LAB
UNIVERSITY OF TORONTO
Donnelly Centre
Cellular & Biomolecular Research
UNIVERSITY OF TORONTO

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.

Part 1: g:Profiler

Part 2: GSEA Gene Set Enrichment Analysis

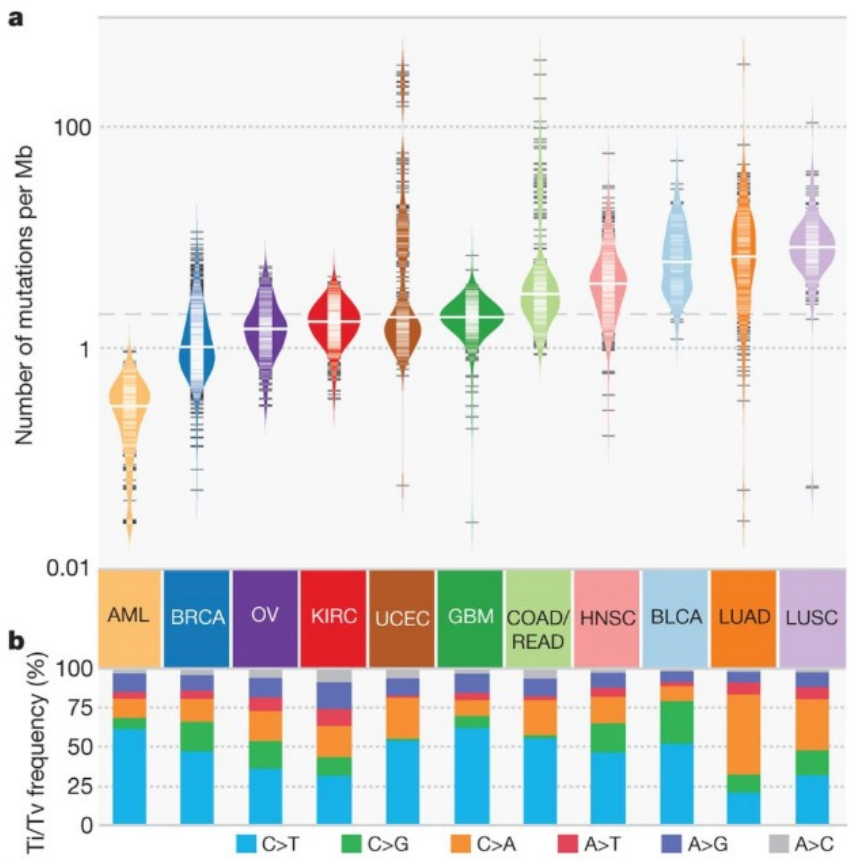
Characteristics:	g:Profiler	GSEA
Input	gene list (thresholded)	ranked gene list (non thresholded)
Statistics	**Fisher's exact test** (can upload specific background), **minimum hypergeometric test**	modified **Kolmogorov-Smirnov** test
Multiple hypothesis testing correction	yes (FDR, Bonferroni, custom)	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. link to Baderlab gene-sets 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:

g:Profiler

Data used for practical lab:

Dataset: Mutational landscape and significance across 12 major cancer types



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

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

g:GOSt
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

Query | Upload query | Upload bed file

Input is whitespace-separated list of genes

EGR3
ACVR2A
MECOM
LIFR
SMC3
NCOR1
RPL5
SMAD2
SPOP
AXIN2
MIR142
RAD21
ERCC2
CDKN2C
EZH2
PCBP1

gene list

Run query | random | example

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

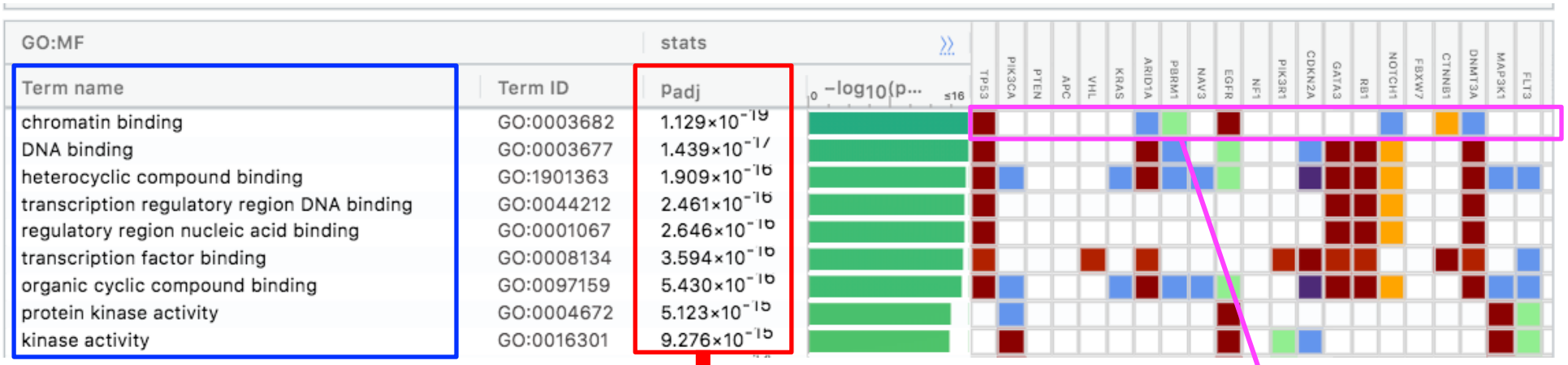
gene sets

Data sources

Custom GMT

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 ParaSite. 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



GO:MF		stats
Term name	Term ID	Padj
chromatin binding	GO:0003682	1.129×10^{-19}
DNA binding	GO:0003677	1.439×10^{-17}
heterocyclic compound binding	GO:1901363	1.909×10^{-16}
transcription regulatory region DNA binding	GO:0044212	2.461×10^{-16}
regulatory region nucleic acid binding	GO:0001067	2.646×10^{-16}
transcription factor binding	GO:0008134	3.594×10^{-16}
organic cyclic compound binding	GO:0097159	5.430×10^{-16}
protein kinase activity	GO:0004672	5.123×10^{-15}
kinase activity	GO:0016301	9.276×10^{-15}

each row is a gene-set (pathway)

Result of Fisher's exact test + multiple hypothesis correction:
gene-sets (pathways) are significantly enriched at FDR <0.05 (scientific notation: 5×10^{-2})

colored boxes: genes in our gene list that overlap with the tested gene-set

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

Time to start practical part:

g:Profiler

- Go the 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.

Bonus – Run g:Profiler programmatically from R

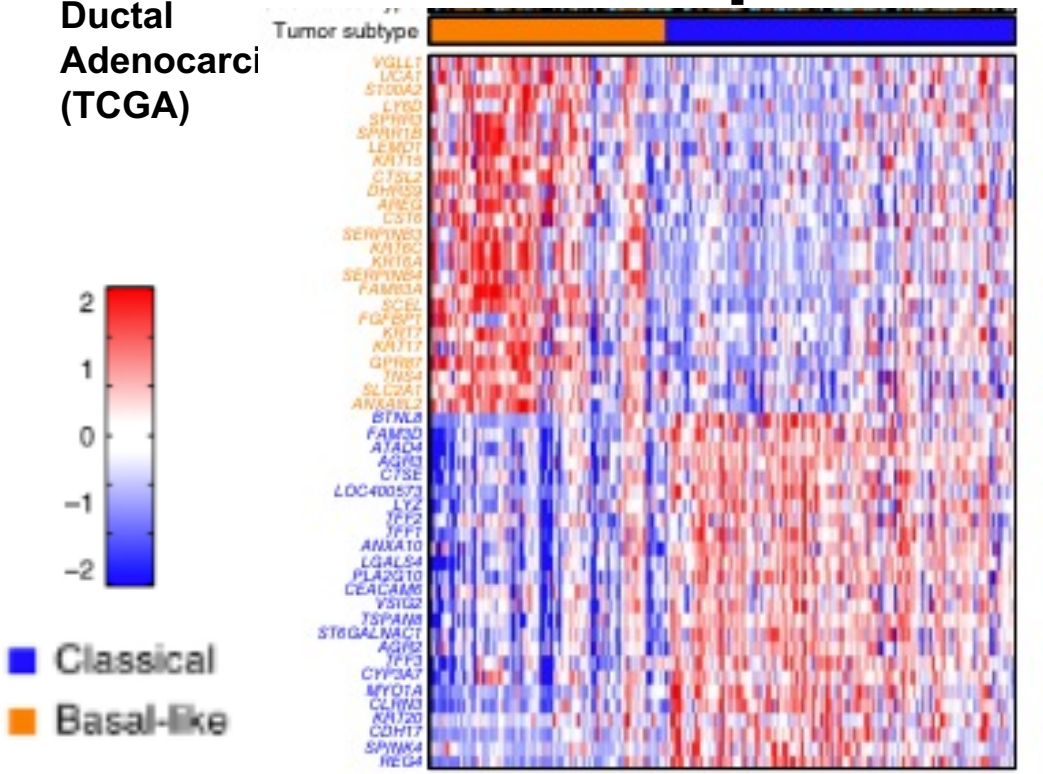
- See example code - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/run-gprofiler-from-r.html

- For instructions on how to set up R so you can run the above notebooks - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/setup.html

Part 2:

Data used for practical lab: RNAseq workflow

Dataset Pancreatic Ductal Adenocarci (TCGA)



Basal vs **Classical**

↓

Differential expression (edgeR)

↓

Rank file

↓

Gene-Set Enrichment Analysis (GSEA)

Moffitt, R., Marayati, R., Flate, E. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168–1178 (2015)

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	LR	PValue	FDR
S100A2	3.64836808	6.92509974	109.253171	1.43E-25	2.23E-21
TP63	1.94455007	2.70442944	50.9005836	9.72E-13	2.03E-10
HEATR1	8.99E-05	5.23106622	1.03E-06	0.9991911	0.99964026
LCMT2	4.96E-05	3.34285388	4.64E-07	0.99945676	0.9996645
CPB1	-1.2706506	12.4548512	-1.2095769	1	1
DKC1	-1.11E-06	5.8750225	2.63E-10	0.99998705	1
NUS1	-3.21E-05	4.79657456	2.75E-07	0.99958166	0.99971
CYP2S1	-1.5319202	6.71168783	46.4828517	9.24E-12	1.48E-09
HNF4A	-0.9156346	6.02217321	47.3402881	5.97E-12	9.79E-10
FMO5	-1.7405263	4.75672333	101.514393	7.09E-24	5.53E-20

edgeR output

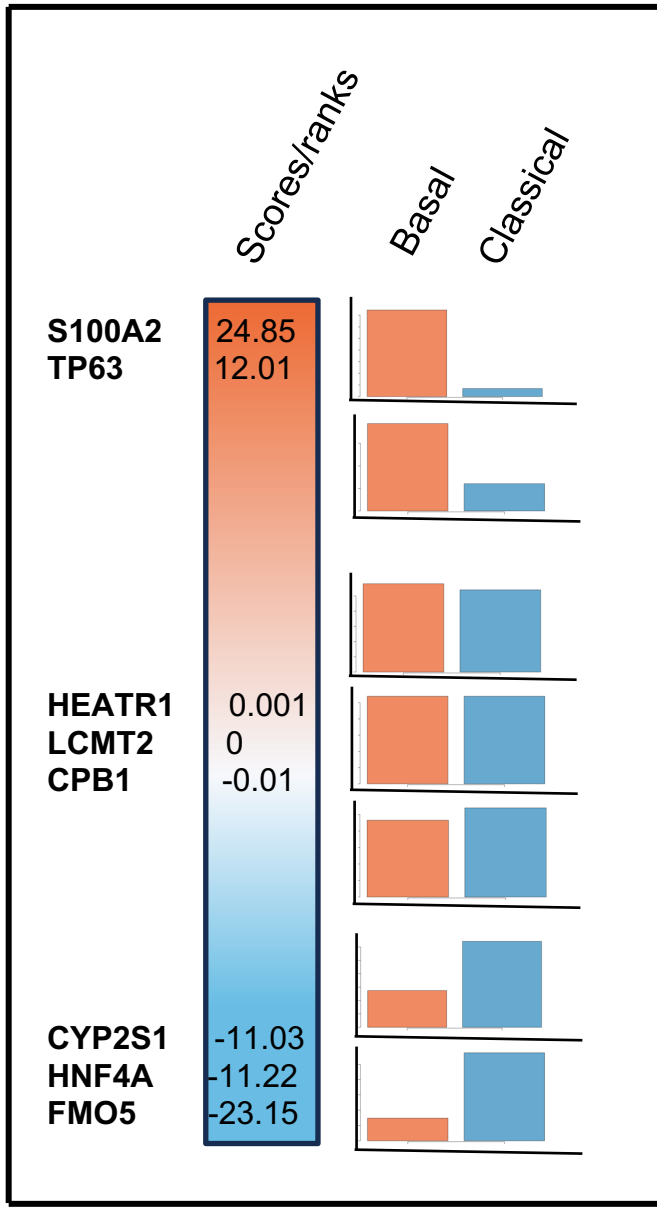
Calculate the ranking score:

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.



What does a .gmt file look like?

Gene-set name	Gene-set name	gene	gene	gene	gene	gene	gene
MOLYBDENUM COFACTOR BIOSYNTHESIS%HUMANCYC%PWY-6823	molybdenum cofactor biosynthesis	NFS1	MOCS2	GPHN	MOCS3		
GLYCEROL DEGRADATION I%HUMANCYC%PWY-4261	glycerol degradation I	GK5	GK	GK2			
OXIDATIVE ETHANOL DEGRADATION III%HUMANCYC%PWY66-161	oxidative ethanol degradation III	CYP2E1	ACSS2	ACSS3	ALDH3A2	ACSS1	ALDH2
TETRAPYRROLE BIOSYNTHESIS II%HUMANCYC%PWY-5189	tetrapyrrole biosynthesis I	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 coding example - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/create-gmt-file-from-ensembl.html).

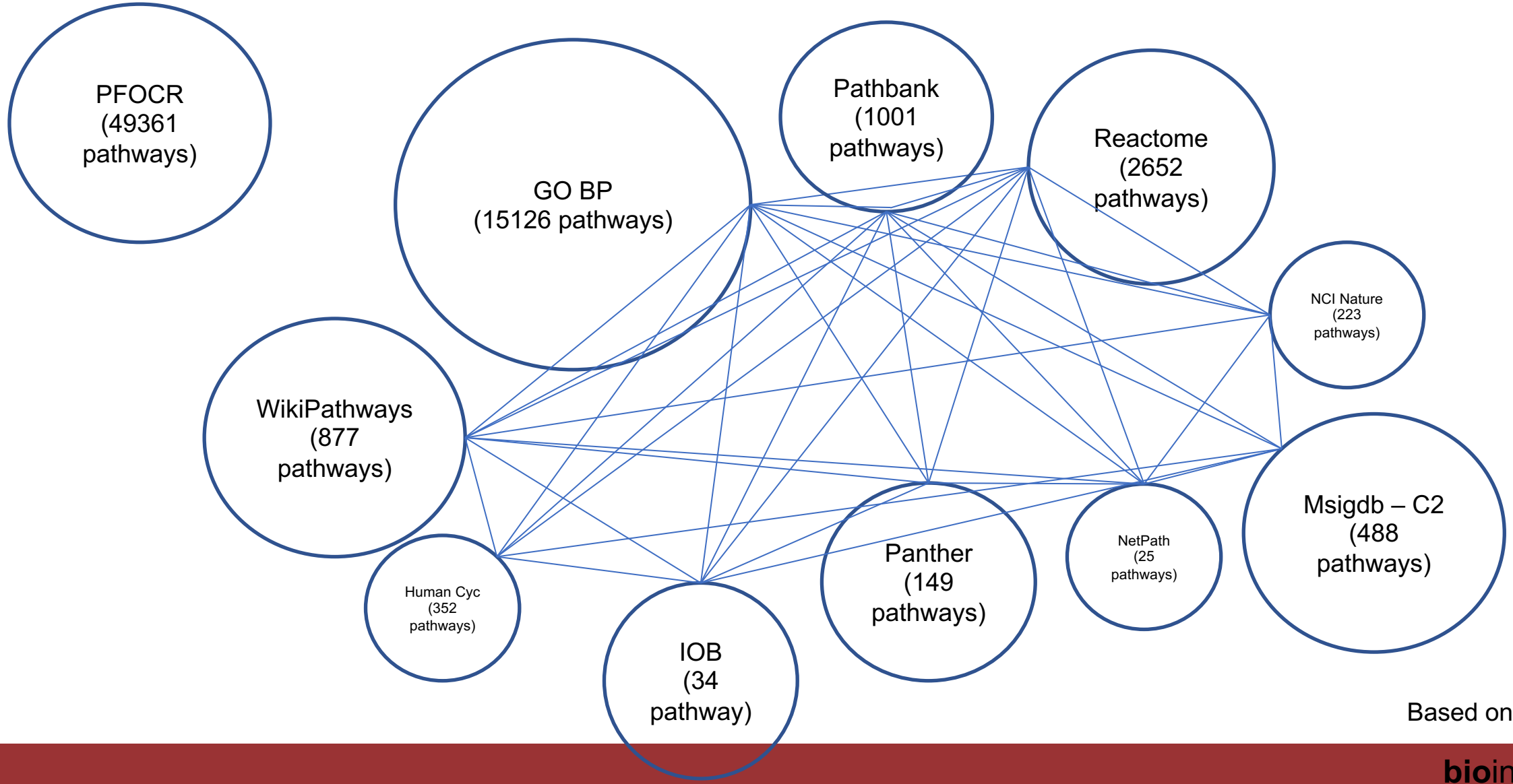
MSigDB database

https://www.gsea-msigdb.org/gsea/msigdb/

C2: curated gene sets (browse 7233 gene sets)	Gene sets in this collection are curated from various sources, including online pathway databases and the biomedical literature. Many sets are also contributed by individual domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into the following two subcollections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP). details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
Reactome subset of CP (browse 1692 gene sets)	Canonical Pathways gene sets derived from the Reactome pathway database.	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
C5: ontology gene sets (browse 16008 gene sets)	Gene sets that contain genes annotated by the same ontology term. The C5 collection is divided into two subcollections, the first derived from the Gene Ontology resource (GO) which contains BP, CC, and MF components and a second derived from the Human Phenotype Ontology (HPO). details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
BP: subset of GO (browse 7647 gene sets)	Gene sets derived from the GO Biological Process ontology.	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
H: hallmark gene sets (browse 50 gene sets)	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. details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle

BaderLab EM_Genesets

http://download.baderlab.org/EM_Genesets/



Based on June 2024 build

BaderLab EM_Genesets

- go to http://download.baderlab.org/EM_Genesets/
  - select current release/
    - Human/
      - symbol/
        - save the Human_GOBP_AllPathways_noPFOCR_no_GO_iea….gmt file on your computer (right click on the link to save it)

Index of /EM_Genesets/June_01_2024/Human/symbol

Name	Last modified	Size	Description
Parent Directory		-	
symbol_translation_summary.log	2024-06-01 13:45	466	
Human_GOBP_AllPathways_noPFOCR_no_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	7.9M	
Human_GOBP_AllPathways_noPFOCR_with_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	9.6M	
Human_GOBP_AllPathways_withPFOCR_no_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	15M	
Human_GOBP_AllPathways_withPFOCR_with_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	16M	
Human_GO_AllPathways_noPFOCR_no_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	12M	
Human_GO_AllPathways_noPFOCR_with_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	15M	
Human_GO_AllPathways_withPFOCR_no_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	19M	
Human_GO_AllPathways_withPFOCR_with_GO_iea_June_01_2024_symbol.gmt	2024-06-01 13:45	22M	
Human_AllPathways_withPFOCR_June_01_2024_symbol.gmt	2024-06-01 13:45	8.5M	
Human_AllPathways_noPFOCR_June_01_2024_symbol.gmt	2024-06-01 13:45	1.8M	
Misc/	2024-06-01 13:45	-	
DrugTargets/	2024-06-01 13:45	-	
DiseasePhenotypes/	2024-06-01 13:45	-	
TranscriptionFactors/	2024-06-01 13:45	-	
miRs/	2024-06-01 13:45	-	
Pathways/	2024-06-01 13:45	-	
GO/	2024-06-01 13:45	-	

GSEA preranked

GSEA 4.3.2 (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

Analysis history

GSEA reports
Processes: click 'status' field for results

Name	Status

Show results folder

Home | 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
Collapse/Remap to gene symbols	Remap_Only
Chip platform	

Basic fields (Hide)

Analysis name	my_analysis
Enrichment statistic	weighted
Max size: exclude larger sets	200
Min size: exclude smaller sets	15
Save results in this folder	/Users/ruthisserlin/gsea_home/output/mar29

Advanced fields (Show)

Reset | Last | Command | Run

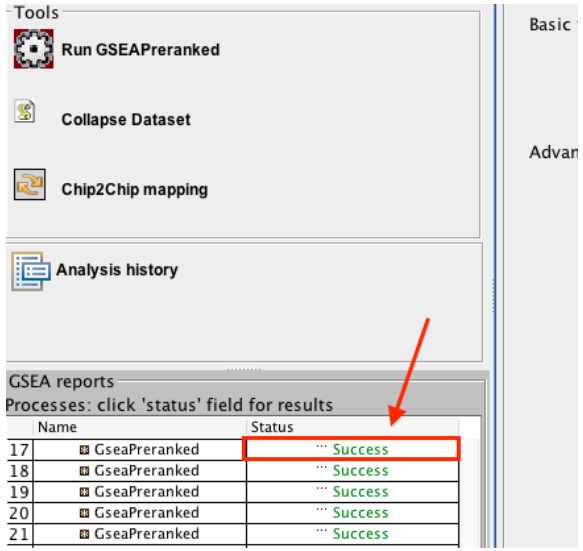
11:12:07 a.m. | 83M of 256M

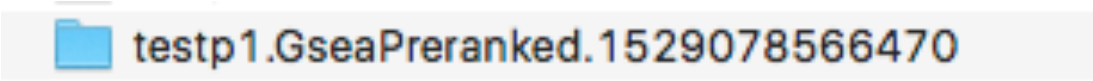
Each gene-set will be permutated 1000 with random genes to build the null distribution

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

Exploring GSEA results

How to access GSEA results?





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_na_neg_1717773429384**.tsv
  - **gsea_report_for_na_neg_1717773429384**.tsv

GSEA Report for Dataset TCGA-PAAD_GDC_Subtype...Moffitt_BasalvsClassical_

Enrichment in phenotype: na

Basal

- 3591 / 5825 gene sets are upregulated in phenotype **na_pos**
- 970 gene sets are significant at FDR < 25%
- 441 gene sets are significantly enriched at nominal pvalue < 1%
- 910 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

gene-sets enriched in genes up-regulated in treated cells compared to non-treated samples or condition A vs condition B

Enrichment in phenotype: na

Classical

- 2234 / 5825 gene sets are upregulated in phenotype **na_neg**
- 285 gene sets are significantly enriched at FDR < 25%
- 141 gene sets are significantly enriched at nominal pvalue < 1%
- 387 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

gene-sets enriched in genes down-regulated in treated cells compared to non-treated samples or condition B vs condition A

Dataset details

- The dataset has 15579 features (genes)
- No probe set => gene symbol collapsing was requested, so all 15579 features were used

Gene set details

- Gene set size filters (min=15, max=200) resulted in filtering out 13695 / 19520 gene sets
- The remaining 5825 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Index.html
summary of results

- Give the number or significant gene-sets (pathways)
- 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'** link or locate it in the GSEA folder result.

Exploring GSEA Results

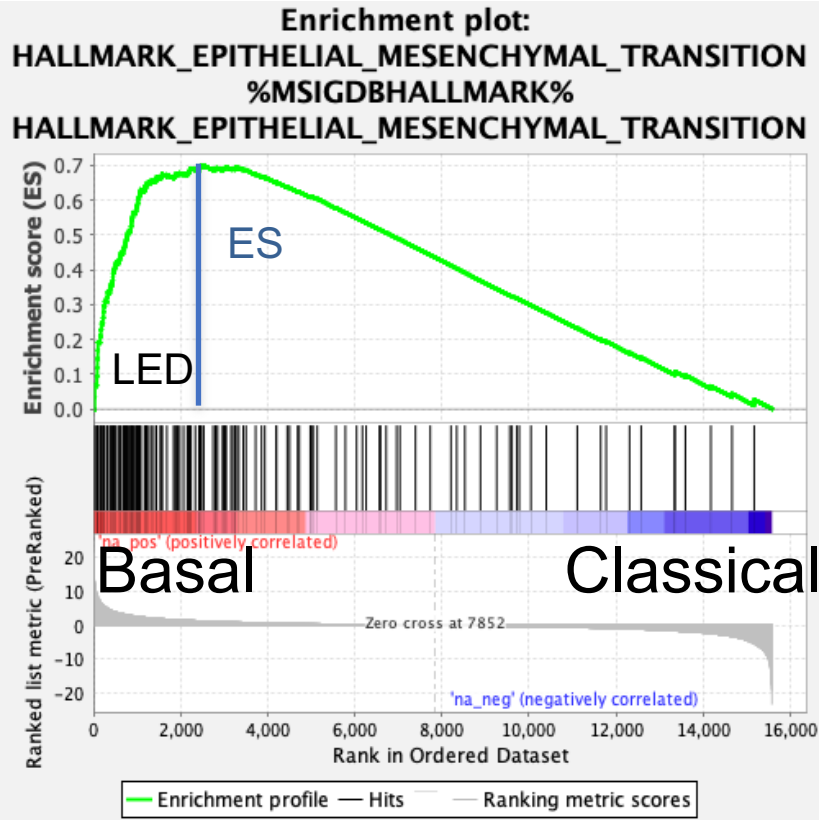
NES FDR

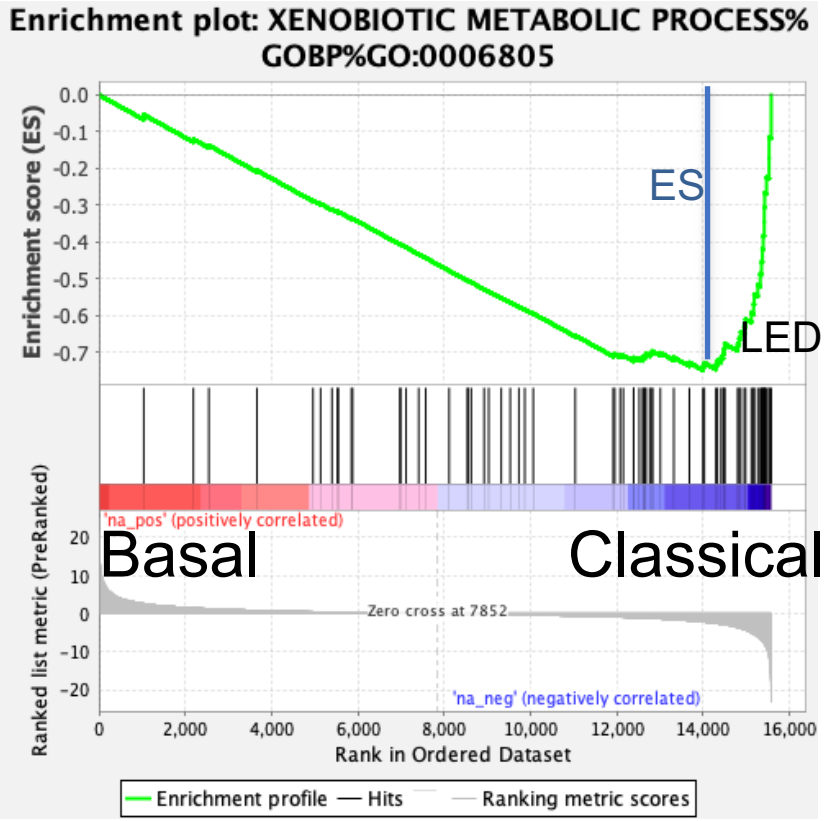
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION%MSIGDBHALLMARK%HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	Details ...	149	0.70	2.48	0.000	0.000	0.000	2536	tags=56%, list=16%, signal=67%
2	KERATINIZATION%REACTOME%R-HSA-6805567.5	Details ...	47	0.81	2.39	0.000	0.000	0.000	178	tags=28%, list=1%, signal=28%
3	SKIN DEVELOPMENT%GOBP%GO:0043588	Details ...	87	0.71	2.26	0.000	0.000	0.000	1285	tags=39%, list=8%, signal=42%
4	KERATINOCYTE DIFFERENTIATION%GOBP%GO:0030216	Details ...	41	0.79	2.21	0.000	0.000	0.000	1213	tags=37%, list=8%, signal=40%
5	TGF-BETA RECEPTOR SIGNALING ACTIVATES SMADS%REACTOME%R-HSA-2173789.6	Details ...	45	0.73	2.20	0.000	0.000	0.000	1133	tags=33%, list=7%, signal=36%
6	EPIDERMIS DEVELOPMENT%GOBP%GO:0008544	Details ...	117	0.64	2.15	0.000	0.000	0.000	1352	tags=34%, list=9%, signal=37%
7	INTERMEDIATE FILAMENT ORGANIZATION%GOBP%GO:0045109	Details ...	30	0.82	2.15	0.000	0.000	0.000	151	tags=33%, list=1%, signal=34%

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:2.48
FDR:0.000

NES:-2.24
FDR: 0.000

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

Time to start practical part:

- Go the 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://baderlab.github.io/Cytoscape_workflows/EnrichmentMapPipeline/index.html

Bonus – Run GSEA programmatically from R

- See example code - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/run-gsea-from-within-r.ht

- For instructions on how to set up R so you can run the above notebooks - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/setup.html

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR Ontario Institute for Cancer Research

Ontario Genomics

GenomeCanada