BCB420 - Computational Systems Biology

Lecture 7 - Recap and GSEA

Ruth Isserlin

2020-03-22

Before we start

It is not too late to fill this out (If you haven't filled this out yet please do):

Mid course Feedback : https://forms.gle/maGA529V7pxvgBXC6

Journal (Clarifications)

- **Main purpose** : to develop good habits.
- Things to remember:
  - Often the person we are writing notes for is our future selves when we revisit a project, need to write up all the details of a given project for publication.
  - data transformations, parameters, code version are good details to include.
  - Errors that you encountered and how you fixed them! **They will come up again. I guarentee it!**
  - See journal course prepratory material for more details and template of a journal entry.
- What should be in your journal? (minimally:)
  - Plaigarism unit
  - work associated with Assignment
  - attempts at using docker
  - annotation source homework
  - gprofiler homework
  - any future homework
  - If there are assigned readings - enter your notes on the article as a journal entry.

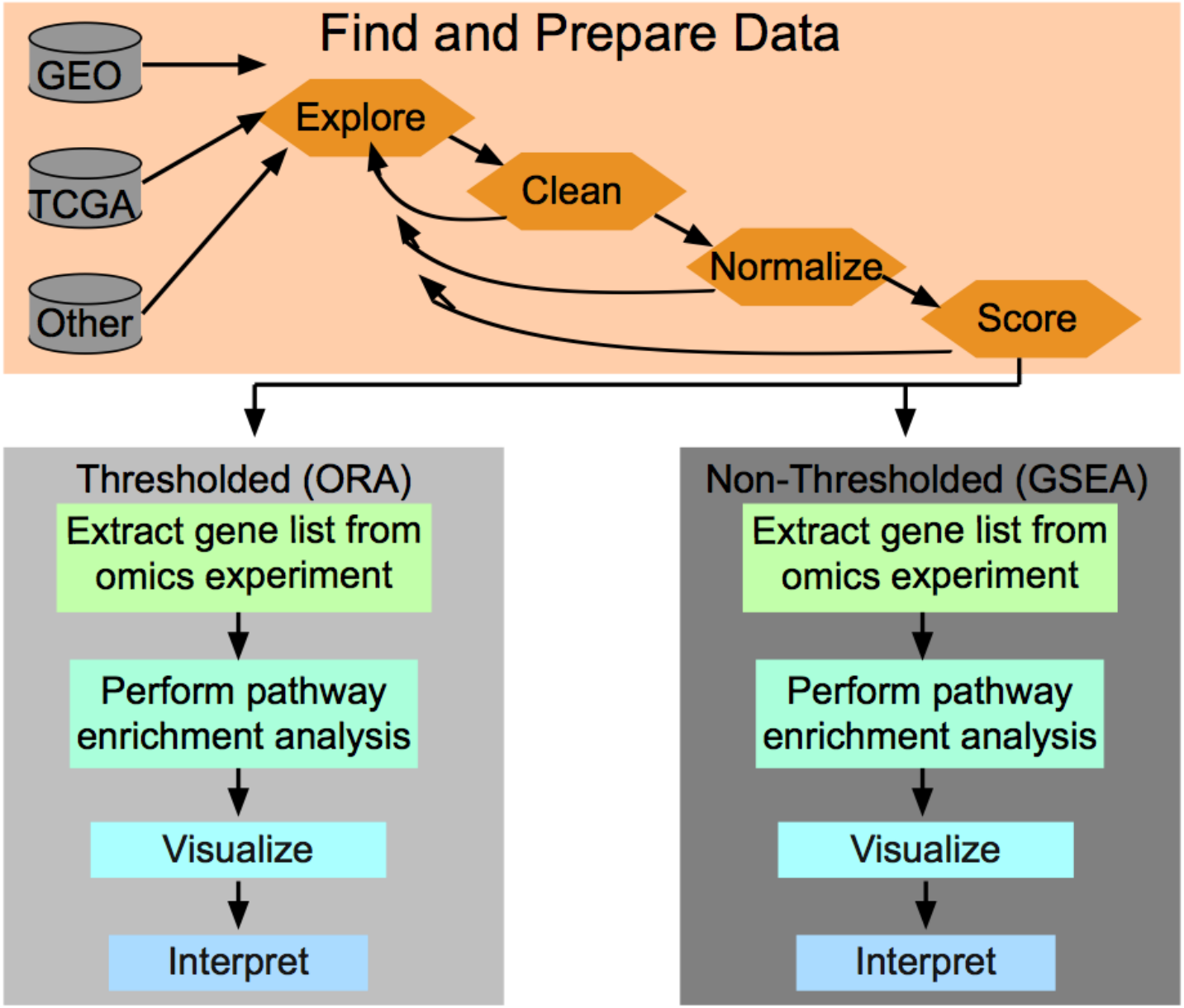
Find and Prepare Data
GEO
TCGA
Other
Explore
Clean
Normalize
Score
Thresholded (ORA)
Extract gene list from omics experiment
Perform pathway enrichment analysis
Visualize
Interpret
Non-Thresholded (GSEA)
Extract gene list from omics experiment
Perform pathway enrichment analysis
Visualize
Interpret

Table 2

Gene set analysis tools

Tool	Author	Year	Citations[1]	Availability	Gene sets	Methods[2]
WEBGESTALT	Zhang *et al.* [73]	2005	1423	Web server	GO, KEGG, +20 more	ORA, GSEA
GOSTATS	Falcon and Gentleman [74]	2007	1437	R package	GO	ORA
G:PROFILER	Reimand *et al.* [75]	2007	534	Web server	GO, KEGG, +7 more	ORA
GENETRAIL	Backes *et al.* [76]	2007	360	Web server	GO, KEGG, +28 more	ORA, GSEA
DAVID	Huang *et al.* [8]	2009	19 569	Web server	GO, KEGG, +38 more	ORA
GORILLA	Eden *et al.* [77]	2009	1881	Web server	GO	ORA
TOPPGENE	Chen *et al.* [78]	2009	1200	Web server	GO, KEGG, +45 more	ORA
CLUSTER-PROFILER	Yu *et al.* [10]	2012	1305	R package	GO, KEGG, +8 more	ORA, GSEA
PANTHER	Mi *et al.* [79]	2013	1405	Web server	GO, +2 more	ORA, GSEA
ENRICHR	Chen *et al.* [9]	2013	1246	Web server	GO, KEGG, +33 more	ORA

Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, Law C,Davis S, Carey V, Morgan M, Zimmer R, Waldron L. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform. 2020 Feb 6 [PMID] (https://www.ncbi.nlm.nih.gov/pubmed/32026945)

Homework from last week

Use this list of genes:genelist.txt as your query set and run a g:profiler enrichment analysis with the following parameters:

1.Data sources : Reactome, Go biologoical process, and Wiki pathways 1.Multiple hypothesis testing - Benjamini hochberg

Answer the questions below:

1. What is the top term returned in each data source?
2. How many genes are in each of the above genesets returned?
3. How many genes from our query are found in the above genesets?
4. Change g:profiler settings so that you limit the size of the returned genesets. Make sure the returned genesets are between 5 and 200 genes in size. Did that change the results?
5. Which of the 4 ovarian cancer expression subtypes do you think this list represents?
6. **Bonus**: The top gene returned for this comparison is TFEC (ensembl gene id:ENSG00000105967). Is it found annotated in any of the pathways returned by g:profiler for our query? What terms is it associated with in g:profiler?

Let's go through the answers

www.kahoot.it

Bonus: The top gene returned for this comparison is TFEC (ensembl gene id:ENSG00000105967). Is it found annotated in any of the pathways returned by g:profiler for our query? What terms is it associated with in g:profiler?

Bader lab genesets

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

- Automatically download the latest geneset file for your analysis

```
gmt_url =
"http://download.baderlab.org/EM_Genesets/current_release/Human/symbo

# list all the files on the server
filenames = getURL(gmt_url)
tc = textConnection(filenames)
contents = readLines(tc)
close(tc)

# get the gmt that has all the pathways and does not include terms
inferred from
# electronic annotations(IEA) start with gmt file that has pathways
only
rx = gregexpr("(?<=<a href=\")(.*.GOBP_AllPathways_no_GO_iea.*.)
(.gmt)(?=\">)", contents,
    perl = TRUE)
gmt_file = unlist(regmatches(contents, rx))

dest_gmt_file <- file.path(data_dir, gmt_file)

download.file(paste(gmt_url, gmt_file, sep = ""), destfile =
dest_gmt_file)
```

Table 1
Gene set analysis methods under benchmark

Method	Author	Year	Citations[1]	RNA-seq	Gene statistic[2]	Set statistic	Significance estimation
ORA	–[3]	–[3]	–[3]	✓	user-defined	DE / GS overlap	Fisher's exact test
GLOBALTEST	Goeman *et al.* [68]	2004	983	–	–	Q statistic	Empirical Bayes GLM
GSEA	Subramanian *et al.* [7]	2005	16 730	–	t_{S2N}	KS statistic	Sample permutation
SAFE	Barry *et al.* [54]	2005	350	–	Student's t	Wilcoxon rank sum	Sample permutation
GSA	Efron and Tibshirani [62]	2007	798	–	t_{SAM}	Maxmean	Sample permutation
SAMGS	Dinu *et al.* [69]	2007	270	–	t_{SAM}	Hotelling's T^2	Sample permutation
ROAST	Wu *et al.* [70]	2010	253	✓	t_{MOD}	Weighted mean	Rotation
CAMERA	Wu and Smyth [66]	2012	246	✓	t_{MOD}	t_{IGC}	Two-sample t-test
PADOG	Tarca *et al.* [25]	2012	71	–	$\lvert t_{MOD} \rvert$	Weighted mean	Sample permutation
GSVA	Hänzelmann *et al.* [71]	2013	471	✓	–	KS statistic	Empirical Bayes GLM

Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, Law C,Davis S, Carey V, Morgan M, Zimmer R, Waldron L. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform. 2020 Feb 6 [PMID] (https://www.ncbi.nlm.nih.gov/pubmed/32026945)

Running and Exploring GSEA

Assignment #2

- differentail gene expression and preliminary ORA
- Due March 3, 2020! @ 20:00

What to hand in?

- **html rendered RNotebook** - you should submit this through quercus
- Make sure the notebook and all associated code is checked into your github repo as I will be pulling all the repos at the deadline and using them to compile your code. - Your checked in code must replicate the handed in notebook.
- Document your work and your code directly in the notebook.
- **Reference the paper associated with your data!**
- **Introduce your paper and your data again**
- You are allowed to use helper functions or methods but make sure when you source those files the paths to them are relative and that they are checked into your repo as well.

Homework for next week

Practise using GSEA. Given the ranked list comparing mesenchymal and immunoreactive ovarian cancer (mesenchymal genes have positive scores, immunoreactive have negative scores). perform a GSEA preranked analysis using the following parameters:

- genesets from the baderlab geneset collection from February 1, 2020 containing GO biological process, no IEA and pathways.
- maximum geneset size of 200
- minimum geneset size of 15
- gene set permutation

and answer the following questions in your journal:

1. What is the top gene set returned for the Mesenchymal sub type? What is the top gene set returned for the Immunoreactive subtype?
2. What is its pvalue, ES, NES and FDR associated with it.
3. How many genes in its leading edge?
4. What is the top gene associated with this geneset