Blind hackathon to assess tools for biological interpretation

List of tools

- DAVID https://david.ncifcrf.gov/ -
- g:Profiler https://biit.cs.ut.ee/gprofiler
- StringDB http://string-db.org/ -
- EnrichNet www.enrichnet.org
- GeneTrail https://genetrail2.bioinf.uni-sb.de/
- clusterProfiler
- Cytoscape app ClueGO?
- MetaScape?
- GOrilla -
- Reactome https://reactome.org
- Panther www.pantherdb.org/
- IMPaLA http://impala.molgen.mpg.de/
- Use your own way

Backup:
WebGestalt
ClueGO
L2L
Gage
GOseq
SeqGSA
IMPaLA http://impala.molgen.mpg.de/

Tasks

Selection and testing

- Select 2-3 tools from the list and use their included examples to understand what they provide
- Which species are you dealing with
- Before you start, write down a sentence for each tool: How do you expect the tool to perform?
- What does functional enrichment mean?
- How are the functions given?
- What is enriched?

Q8N112

Data preparation

- Download files from

Proteomics data set I: Vasilis

1) http://computproteomics.bmb.sdu.dk/tmp/UniprotAccsComma.txt
platelet granule proteome taking the study https://www.ncbi.nlm.nih.gov/pubmed/24549006, most likely outcome (KEGG pathway): platelet activation

Proteomics data set II:

2) http://computproteomics.bmb.sdu.dk/tmp/UniprotAccs.csv - Patrick & Dainel
Malaria study of blood) from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5137300/

RNA-seq data set:

3) http://computproteomics.bmb.sdu.dk/tmp/RNAseq_dataset.csv or http://computproteomics.bmb.sdu.dk/tmp/RNAseq_dataset.xlsx

Transcriptomics data set(HOXA1 knock-down in lung fibroblasts) from:
https://www.ncbi.nlm.nih.gov/pubmed/23222703
Data: *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37704*

- Think about the available data format. What does it contain?
- Would you use a similar table for subsequent data interpretation?
- What are the criteria to select genes/proteins for further processing?
- Give 3 statements on the criteria

Run tools, compare and benchmark

- Apply the tools and answer the following questions:
    - How easy was it?
    - Do the results compare?
    - Which tools would you prefer and why?
    - Which annotations do you mainly use for interpretation?
- Give your top three annotations for each data set.
- What are your guesses on the data?
    - Which cell line / tissue?
    - Any idea on disease or treatment?

Disclosure and discussion

Patrick & Daniel

- Select 2-3 tools from the list and use their included examples to understand what they provide
    - GOrilla
    - String
    -
- Which species are you dealing with

Human
Mouse
Human

- Before you start, write down a sentence for each tool: How do you expect the tool to perform?
    - **String**: A protein interaction prediction tool.
    - **GOrilla**: Go term enrichment, identifies which go terms are enriched in a list of genes
- What does functional enrichment mean?
    - Genes or proteins that are overrepresented in a subset of data with similar or connected function
- How are the functions given?
    - The functions are given in the form of gene ontology terms (which are collected in databases)
- What is enriched?
    -

- Think about the available data format. What does it contain?
    - Data set II: csv file with iTRAQ data. 3 data sets (mice subjected to different treatments?). Accession numbers of proteins with increased expression levels from each treatment with description of the protein, fold-change, p-values.
- Would you use a similar table for subsequent data interpretation?
    - No, we have to extract only the accession numbers for each condition individually. Possibly make it long-format
- What are the criteria to select genes/proteins for further processing?
    - Fold-change, p-value, confidence of protein identification
- Give 3 statements on the criteria
    -