Exercise 2 – Explore the new enrichment detection in STRING

Objectives:

- to learn about the two different modes of "functional enrichment detection" in the STRING database
- to search for published scientific papers that show an enrichment signal, relative to a user-provided, ranked gene list
- to search for enrichments with a gene-set of your own.

A part of the exercise will be working with a new feature in STRING that has only recently been released (for version 11.5); some things may still be a bit rough around the edges.

1). standard gene-set enrichment in STRING

- we will be working with two published gene-lists, from two different studies:
  a) https://string-db.org/string_course/melanoma.log_fold_changes.txt
  b) https://string-db.org/string_course/liver.gender.log_fold_changes.txt

  both are gene-expression studies; the first one describes bulk expression differences between primary melanoma tumors and melanoma metastases (Riker et al., BMC Med Genomics; 1:13. doi: 10.1186/1755-8794-1-13) and the second one aims to identify gender-biased gene expression in human liver, averaging over a large number of liver samples from both sexes (Zhang et al., PLoS One. 6(8): e23506. doi: 10.1371/journal.pone.0023506).

- download the first file (melanoma expression), and save it on your Desktop.
- open it with Excel, and sort it by the log-fold-expression column, ascending.
- now, let's submit the first 100 proteins to STRING: point your browser to the STRING site (https://string-db.org/), and choose the input option "Multiple proteins". Then, use copy-and-paste to copy the first 100 names of the sorted list from Excel to the browser. In this case, make sure to copy only the protein names at this point, not the fold-change values. And, specify the organism "Homo sapiens" before submitting.
- submit the list, and proceed until you arrive at the network view in STRING.
- then, below the network, select the tab labeled "Analysis".
- first, inspect the "Network Stats". Does your network have more edges than expected? If so, what would be the biological explanation?
- Below the network stats are the "functional enrichments". Inspect those now … do they make sense for a melanoma-related dataset?
- can you determine what the directionality of the fold-change values in the Excel file means? We took the top of the list, sorted by fold-change in ascending order … looking at the functional enrichments, are those the genes that are biased towards metastases, or towards the primary tumor?
- now, proceed in the same way as before, but choose the bottom 100 proteins in the list. Are you getting different enrichment results for these?
- lastly, repeat the procedure for the other dataset (human liver, male vs. female). Try separately with the 100 top-ranked proteins, and with the 100 bottom-ranked proteins.

2). rank/value-based gene-set enrichment in STRING

- now, we will try the new “rank/value-based” enrichment feature in STRING. You may want to keep the previous results in the browser for comparison, so open a new browser tab now.
- In the new window, open STRING again: https://string-db.org/
- select the new input option “Proteins with Values/Ranks”
- then, submit the entire “melanoma” dataset as downloaded in the previous exercise. You do not need to take it via Excel, instead you can just submit the entire file as it is. Do not forget to specify the organisms “human”, then click Search.
- notice how this now takes a bit longer – STRING needs to map all genes in the input set, and then run rank-based statistical tests including permutations.
- when the results come back, compare them to the results from the previous exercise. Did the same terms come up?
- one of the terms that did not come up before is “multicellular organismal water homeostasis” (under the section “Biological Process”). Can you guess why that term did not show up when looking at the 100 top proteins only?
- notice how on the bottom left of the page there is a zoomable STRING network, showing all human proteins. Zooming all the way in, while repeatedly hovering over the enrichment term “keratinization”, can you identify and show the protein domain structure of one of the constituent proteins of that term?
- next, repeat the entire procedure with the liver-gender dataset. Recall that it revealed little functional enrichments previously. Do you observe anything now? Does it make sense?

3). try your own ranked list of genes.

- now, since this is a new feature, it would be nice to see how well it performs on other, real-world datasets.
- identify a dataset or study that is of interest to you. It could be either published or private, but it should consist of human genes and some measured property of those. The measured property does not have to be a (log) fold-change … it could be anything from mutation counts to phenotypic screening results to protein abundances.
- use Excel to format the list in analogy to the input examples studied above: one gene name per line, followed by a numerical value. Note that the list does not have to be sorted … note also that STRING understands a number of different gene-names/identifiers (if in doubt, choose Uniprot accessions or HUGO gene names).
- given what you know about the process / the study, do the results make any sense?
- can you identify ways to improve this new STRING functionality further? If so, please let Dr. von Mering know (or send an email to STRING’s helpdesk).