Practical 8

**3. Using the provided file experiments.txt containing gene sets (one set per row) of DEGs from the experiments in S.cerevisiae S228C, find two experimental gene sets that overlap the most with the genes in your eukaryotic chromosome. Save the two gene sets(from experiments.txt) with most overlap for further analysis.**

**Tip 4: You will need the gene symbol, i.e. the first part of the third string (example: YJU6 in sp|P39529|YJU6\_YEAST; sp | uniprot accession number| uniprot entry name) for yoursearch of overlapping genes.**

Wrote the script <exp\_parse.py>, that produced output file <overlap\_count>, which contains a count of the number of overlapping genes displayed with the gene set. There were a total of 3 sets which presented a total of 13 overlapping genes:

ID:26 count:13: ['ATP1', 'FAS2', 'FAS1', 'CEM1', 'BNA2', 'AMD2', 'LIP2', 'SEC59', 'ILV5', 'MTR3', 'FAS1', 'PRP28', 'RPB3', 'SNU13', 'COX13', 'HAM1', 'FAS2', 'PMS1', 'FAS2', 'ARO8', 'FAD1', 'OAR1', 'MRL1', 'PRP18', 'RRP4', 'FAS2', 'NTE1', 'URA1', 'DPB2', 'ALG2', 'DUT1', 'ADE1', 'HIS4', 'LIP2', 'LSM5', 'YMR1', 'ATP14', 'SUB2', 'PRI1', 'PMI40']

ID:38 count:13: ['DCD1', 'DUT1', 'YPC1', 'RRP4', 'OAR1', 'IFA38', 'DPB4', 'ARO8', 'HIS4', 'SGS1', 'PAN6', 'GAD1', 'ACO2', 'SPE2', 'BNA2', 'PGK1', 'SPC3', 'ARO8', 'SSS1', 'FCY1', 'VMA8', 'SUB2', 'HAM1', 'FAS2', 'PRO2', 'GDH2', 'GLN1', 'HIS5', 'LIP5', 'FAS1', 'LYS2', 'PAB1', 'FAS1', 'GAL10', 'NPY1', 'PFS2', 'PPX1', 'PRI1', 'ARO8', 'HAM1']

ID:52 count:13:['RAD4', 'PHA2', 'PRE9', 'DFR1', 'GDH2', 'ERG27', 'MNN9', 'RPL29', 'AGX1', 'PHS1', 'RAD28', 'GAL10', 'YET3', 'RAD59', 'SKI2', 'ALG1', 'RIO2', 'RPB8', 'GLN4', 'ARO8', 'PUT2', 'KIN28', 'TRP5', 'TAZ1', 'GUT1', 'OXA1', 'APN1', 'TSC13', 'BNA2', 'PGC1', 'THR4', 'SSL1', 'MRPS28', 'RPA14', 'SPC3', 'RNH1', 'UTP18', 'TFB5', 'MNN11', 'PRP28']

3 output files were also produced, one for each individual gene set (outID26, outID38, and outID52).

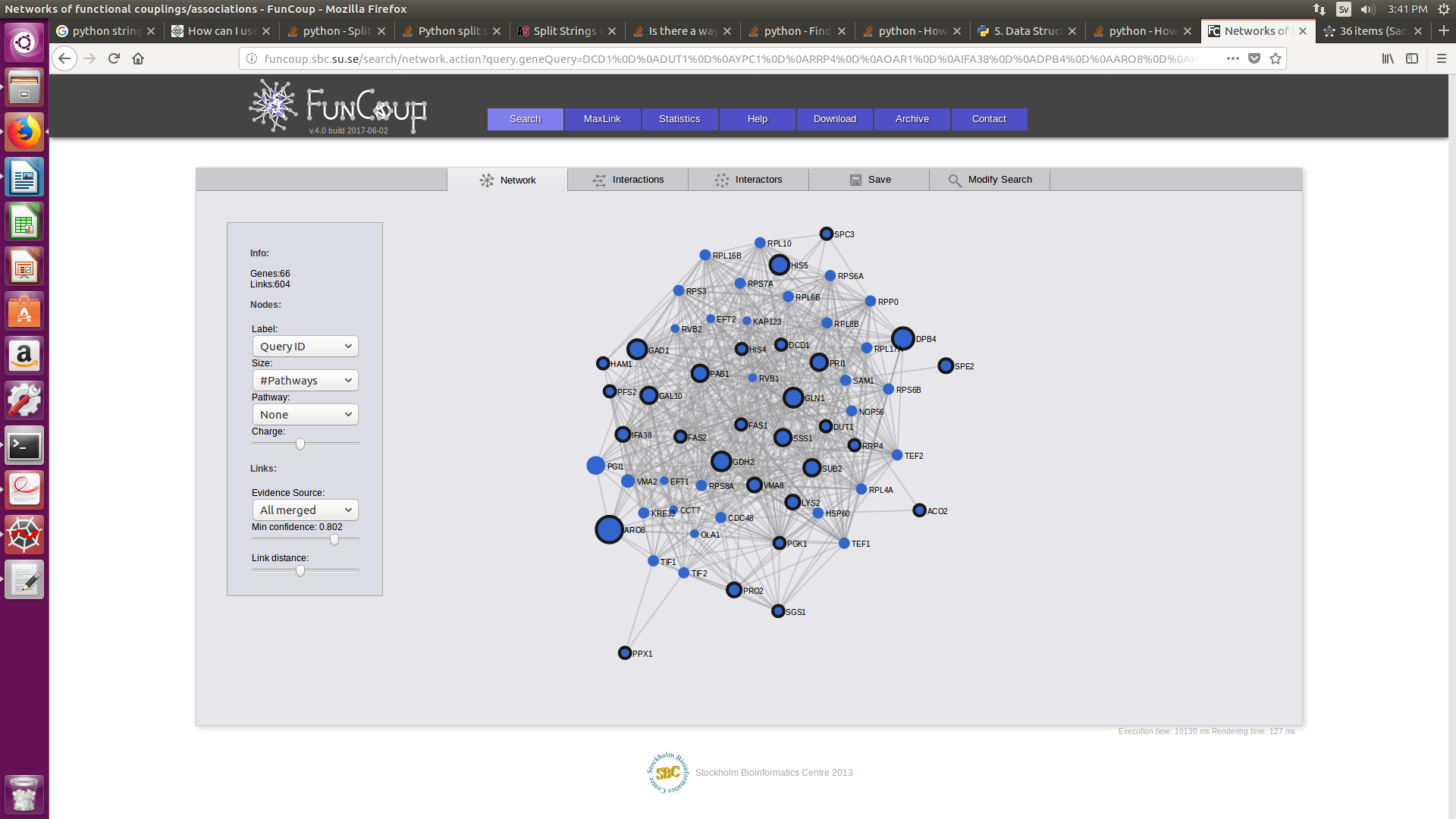
**Comparative network analysis using FunCoup and STRING**

**4. Compare your results gained with both tools based on the same input data (your two genes sets).**

Decided to work on outID38 and outID52.

a. **How do these networks differ in terms of nodes, links, and hubs (the three nodes with the highest degree) for both of your gene sets?**

OutID38 FunCoup:

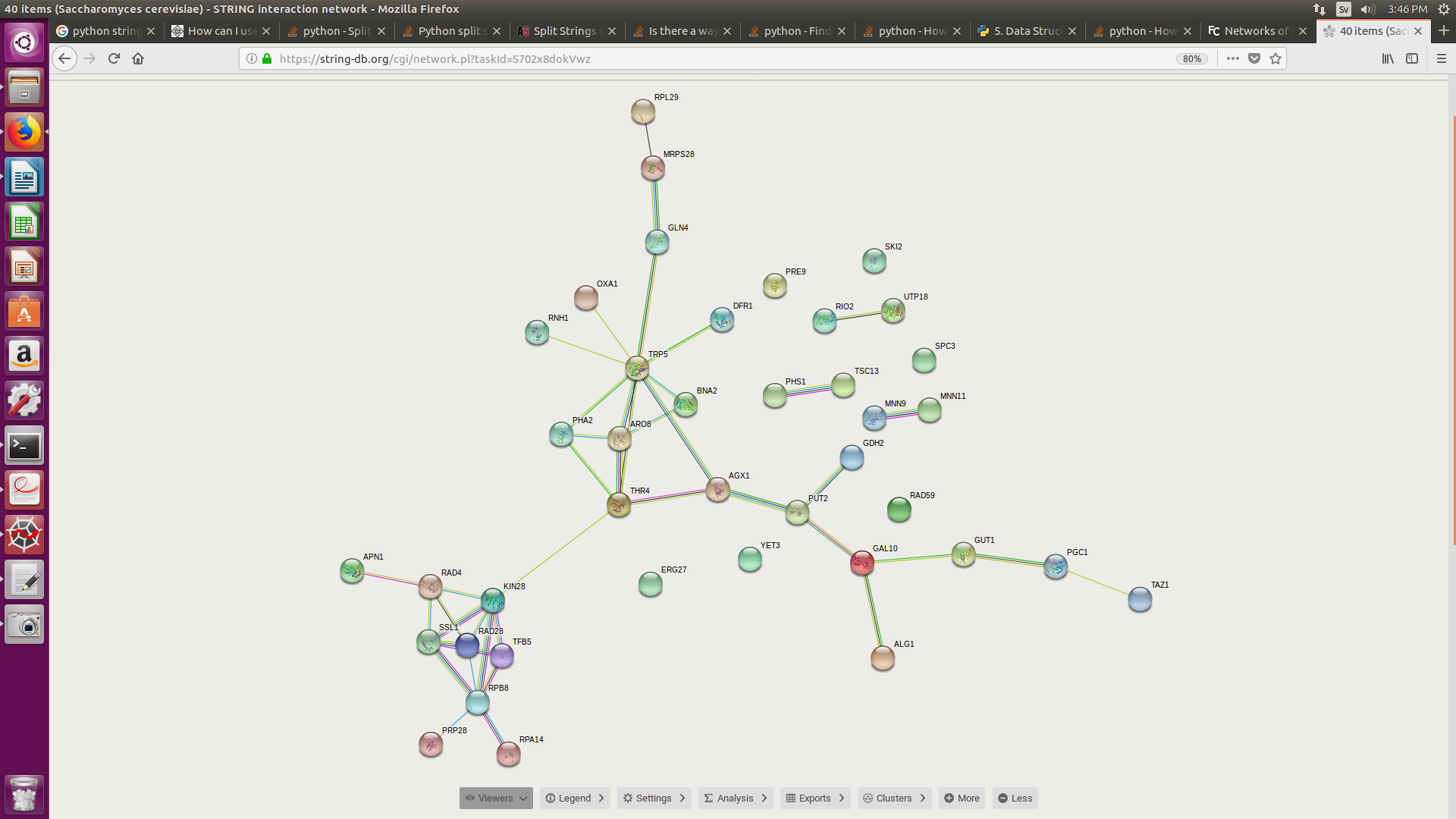


OutID38 STRING:



OutID52 Funcoup:

OutID52 STRING:



For OutID38, the 3 nodes of highest degrees were PAB1 (46), FAS2 (40), and VMA8 (40) in Funcoup, and FAS1, FAS2, HIS4 in STRING. For OutIDt52, the 3 nodes of highest degrees were TRP5 (32), PRE9 (32), RPB8 (28) in Funcoup, and TRP5, RPB8, and RAD28 in STRING. In general, all 4 networks produced vary quite significantly in terms of nodes, links, and hubs, with OutID38 producing a network with a denser network.

b. Which is the most common evidence type with a high confidence (>0.9)?

The protein interaction network (PIN) is the most common evidence type with high confidence.

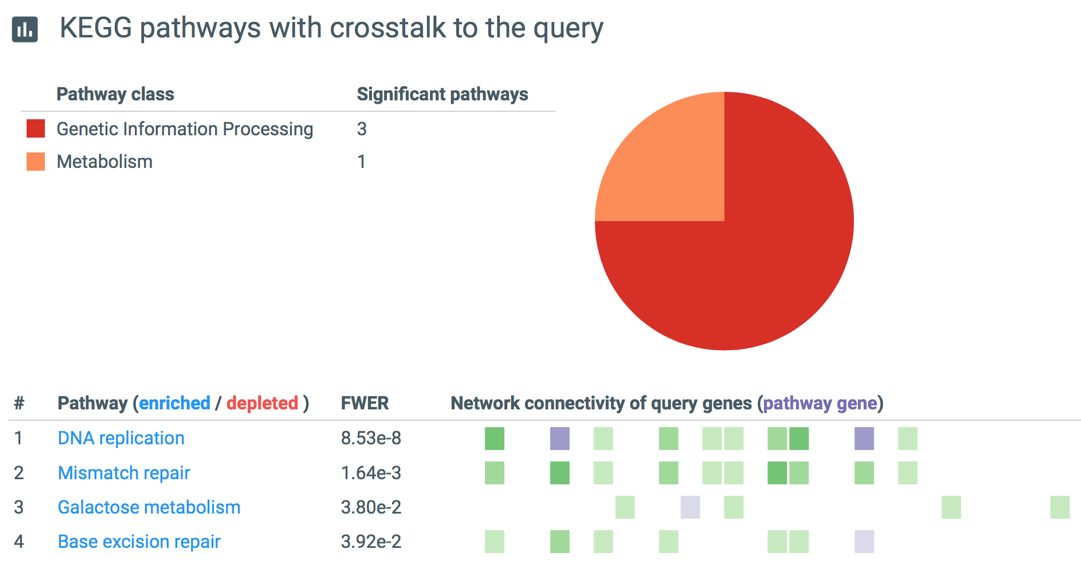
c. What are the differences in terms of underlying data sources in the two databases? Explain them!

Both STRING and FunCoup are integrated databases that serve to merge data from various sources, but differ in terms of of how the information is combined. FunCoup primarily uses the InParanoid database as a reference for orthologous information transfer and a large variety of datasets from various sources such as Gene Ontology (Ogris., et al. 2018), while STRING uses the KEGG database, and PubMed literature for protein association data (Mering., et al. 2005).

***Enrichment analysis using PathwAX and DAVID***

5. Which pathways are enriched ?

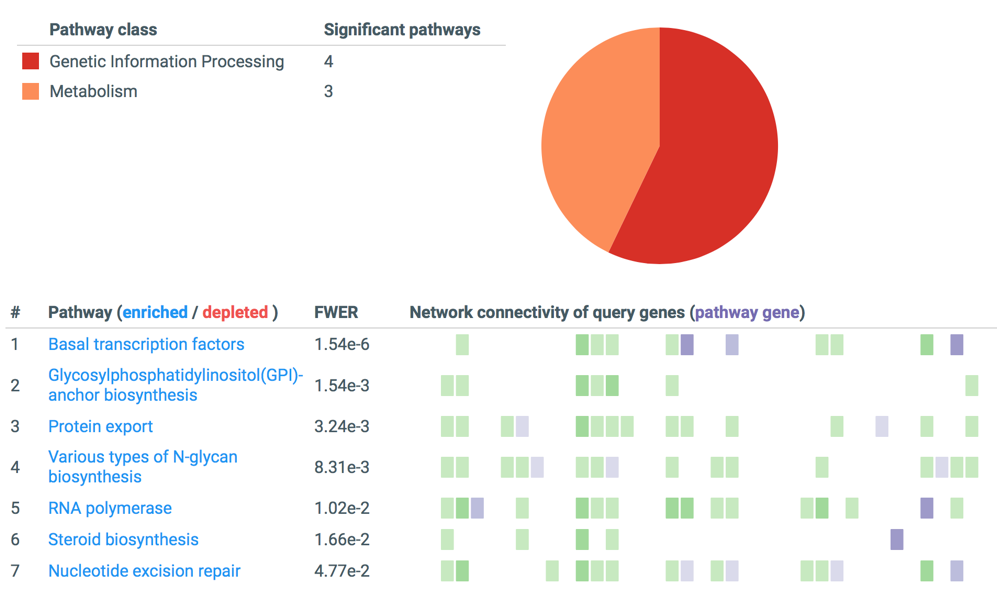
OutID38, PathwAX:



OutID38, DAVID:

Included DAVID output cluster as file outID38DAVID\_Cluster. The most enriched pathways were metabolic pathways and catalytic activity.

OutID52, PathwAX:



OutID52, DAVID:

Included DAVID output cluster as file outID52DAVID\_Cluster. The most enriched pathways were metabolic pathways and biosynthesis.

6. What differences do you observe in the results from PathwAX and DAVID?

The enriched pathways given by the 2 programs differ significantly, for both gene sets. However both programs did state Metabolic Pathways as enriched.

7. Does the number of (input) genes matter for the results? If so, explain why!

Yes, depending on the number of links between the gene set and pathways (crosstalk) the pathway is determined to be enriched or depleted (Ogris., et al. 2016). Therefore a larger number of input genes improves the possibility of crosstalk, and such will affect the results.

**References:**

Christoph Ogris, Dimitri Guala, Mateusz Kaduk, Erik L L Sonnhammer; FunCoup 4: new species, data, and visualization, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D601–D607, <https://doi.org/10.1093/nar/gkx1138>

Von Mering C, Jensen LJ, Snel B, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*. 2005;33(Database Issue):D433-D437. doi:10.1093/nar/gki005.

Ogris C., Helleday T., Sonnhammer E.L.L.. PathwAX: a web server for network crosstalk based pathway annotation. Nucleic Acids Res. 2016;44:W105−W109.