Coursework 2

Important Note

This coursework **must be completed individually**. We will check for plagiarism in your coursework. Plagiarism will lead to mark deductions for poor academic practice or more formal investigations, which can lead to major mark penalties.

If you have questions about requirements of the coursework, please ask these on the discussion board on the Bioinformatics 1 Learn site. If there is any ambiguity in the hand-out, course staff will be happy to help. All students on the course can see the discussion board, so this helps keep things fair.

Introduction

Exploring Autism Genes

In this coursework you are going to undertake a mini-research project looking at some of what is known about Autism Spectrum Disorder. We are going to use a gene driven strategy which will begin by downloading a list of genes that have been linked to Autism that has been curated by the Simons Foundation Autism Research Initative (SFARI) - https://www.sfari.org/.

In **Part One** the project we want to see how the literature is evolving for these genes and where the weight of evidence is from published articles. In a bigger project we would want to perform some in depth analysis of the content of these papers, but for this project we will be using the admittedly crude method of looking at prevalence & citation data. In **Part Two** of the project we will perform some analysis on the genes themselves to try to learn a little more about them. In **Part Three** of the project we will perform some basic network analysis to identify clusters of genes that might be of interest and use a web tool to gain some insight into the functions of those groups of genes.

All that you need to successfully complete this project has been covered in the course using web-tools and/or programming approaches. It is fine for you to use either approach or a hybrid of both, but however you aim to undertake the work everyone needs to write up a report of the research. Throughout the instructions, below, we guide you through the project and provide tips, especially where there is an opportunity to do something a little more complex if you choose. We hope you enjoy the project and exploring in a real-world example how you can bring together different sources of information to anayse and gain insight into a research problem.

Research Instructions

Part One - *Autism Literature*

First, download the SFARI gene list from https://gene.sfari.org/database/human-gene/

- **Task 1** - Plot a bar chart of the number of genes in each SFARI gene-score category (2 marks).
- **Task 2** - Rank the genes by 'number-of-reports' and find the top 5 SFARI genes that are in gene-score category 1 (2 marks).
- **Task 3** - For each of these genes find the number of papers in PubMed that include the gene AND are related to Autism (5 marks).
- **Task 4** - From these data fill a table with genes as rows and paper count by year as column (3 marks).
- **Task 5** - Plot a single stacked histogram displaying the data from the table (3 marks).

Comment on your results. You might want to think about the timescale over which papers have appeared, how different genes have fared over the years, how representative these 5 genes might be. Other factors to consider might be; how confident are you that the papers you have identified are related to Autism and why?

Possible extensions here might be to look at citation data, extend to more of the genes or even all of them.

Part Two - *Autism Genes*

Next we want to look at some of the functional terms that have been annotated to SFARI genes separating them out by their gene-score.

- **Task 1** – Map the gene-symbol for every gene in the SFARI gene list to an NCBI UID (unique Entrez Gene identifier) (2 marks).
- **Task 2** – Using the gene2go file from NCBI that we worked with in week 8 find the Gene Ontology terms that have been annotated to all of the SFARI genes (4marks).
- **Task 3** – Now split the genes up into three lists by their SFARI gene-score (1 mark).
- **Task 4** – Create tables of the 10 most commonly annotated terms for each gene list. The tables should have the following columns: GO term ID, GO term Description, GO term count (3 marks).
- **Task 5** – Take the three lists of UIDs created above and use the PantherDB tool - http://pantherdb.org/ to perform a "Functional classification viewed in graphic charts" analysis reporting and select the "Bar chart" display option, using default settings. Once on the results page displaying the bar chart use the pull-down menu to change the ontology to "Biological Process". Click the small (!) "export" button to download the results as a text file and us this to make your own bar chart (5 marks).

Compare and contrast your results from Tasks 4 and 5 and any differences you find between the gene lists from the different gene-score categories in the SFARI gene lists. What can you say about Autism genes based on the results from these analyses?

Possible extensions here might be to explore other pathway analysis tools and websites such as KEGG and Reactome. You could also chose to perform a statistical analysis directly on the results from Task 4.

Part Three - *Autism Networks*

We now want to see if there's any evidence that these proteins work together. One way to look at this is to look at whether any of their proteins physically interact in a co-ordinated way. To assess this, you will now do some basic network analysis on protein-protein interaction data.

- **Task 1** – Create a plain text file of the NCBI UIDs for all of the gene-score 1 SFARI genes. Then, using the STRING website (https://string-db.org/) upload the gene-score 1 gene list, select "homo sapiens" as the species and hit search then click continue to visualise the protein-protein interaction network. In this network the genes are the nodes and the interactions between them the edges (connections). Click the "analysis" option near the bottom and report the following statistics: "number of nodes", "number of edges" and "average node degree" (3 marks).
- **Task 2** – Click the "cluster" option and select "MCL clustering" with the default option. Download the "MCL clusters in TSV format" file. From this, find the two biggest clusters produced by the MCL clustering and use the PantherDB tool as in Part Two to analyse the function of genes in these clusters. This time instead of selecting "Biological Process" ontology use the "Pathway" ontology (6 marks).
- **Task 3** – Click the "exports" option and download the network as a "bitmap image" file (1 mark).

Compare and contrast your results from the two clusters above, commenting on the functional differences highlighted by the Panther tool. Also, compare back to your findings from Part Two.

Possible extensions would be to repeat these analyses with other gene-score restricted lists and with them all combined. Also, using other ontologies with the Panther tool to explore cluster functions.

Structure of the Report

Write a report of your project including figures, tables and results from your research in Parts One-Three above in the relevant sections. The report should follow the structure described below.

- Completed Cover Sheet - download here - (pdf , word)
- Introduction
- Data & Methods
- Results
  - Part One – Autism Literature
  - Part Two – Autism Genes
  - Part Three – Autism Gene Networks
- Discussion
- References
- Appendices (optional)

There are no hard specific upper or lower length limits. We expect your report to use an 11-point font for the main text, and to be 5-8 pages in total (including everything except the cover page, references and optional appendices). You need to upload your report as a single **PDF** file **including the single page cover page** using the link at the top of this section. To help us mark anonymously, please **do not** include your name anywhere in your report.

Learn Help

Marking Scheme

Your mark for this coursework assignment will **contribute 35% to your mark for the Bioinformatics 1 course**.

The work will be **marked out of 100** with the following scheme:

Overall Report [**30 marks**]

- Data, Methods, Presentation of Results and their interpretation clearly described paying close attention to keeping track of versions and dates. Think of how able another researcher would be to look at your report and repeat what you have done - reproducibility [10 marks]
- Correct overall structure of the report [5 marks]
- References and any Appendices in good order [5 marks]
- Clear, well-labelled graphs throughout [10 marks]

Part One [**25 marks**]

- 15 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.

Part Two [**25 marks**]

- 15 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.

Part Three [**20 marks**]

- 10 marks for successfully completing the tasks.
- up to 6 marks for including one additional piece of analysis.
- up to 4 marks for exceptionally well organised and executed approach.