Bioinformatics @ Data Science A.Y. 2018-2019

Network Biology

GROUP 3:

GUNTURI VAMSI KRISHNA VARMA 1794653
DUDEKULA DASTAGIRI 1826239
ALIKANA PAVAN KUMAR 1826777

Abstract

Crohn’s Disease is one of the inherited disorders of the peripheral nervous system. We try to study the possible neuronal genes involved in the mutation for this disease by documenting their possible interactions and relationships. We have three interactions sources namely IID data, Bio grid, and Innate DB. Later, we perform a detailed analysis which involves the enrichment analysis associated with these genes.

1 Basic introduction to the disease/process

Crohn's disease is a type of inflammatory bowel disease (IBD) that may affect any part of the gastrointestinal tract from mouth to anus. Signs and symptoms often include abdominal pain, diarrhea (which may be bloody if inflammation is severe), fever, and weight loss. Other complications may occur outside the gastrointestinal tract and include anemia, skin rashes, arthritis, inflammation of the eye, and tiredness. The skin rashes may be due to infections as well as pyoderma gangrenosum or erythema nodosum. Bowel obstruction may occur as a complication of chronic inflammation, and those with the disease are at greater risk of bowel cancer.
Crohn's disease affects about 3.2 per 1,000 people in Europe and North America. It is less common in Asia and Africa. It has historically been more common in the developed world.

2 Seed genes

Possible genes that could be involved in this disease are represented below in the table. For each of these genes, we have the information about the Gene Symbol, UniProt AC, GeneID, main protein, Functionality.

Gene Symbol	Uniprot AC	Protein Name	Gene ID	Description
ATG16L1	Q676U5	autophagy related 16 like 1	55054	necessary for autophagy, the major process by which intracellular components are targeted to lysosomes for degradation. Defects in this gene are a cause of susceptibility to inflammatory bowel disease type 10 (IBD10)
UCN	P55089	Urocortin	7349	This gene encodes a member of the sauvagine/corticotropin-releasing factor/urotensin I family
TNFSF18	Q9UNG2	TNF superfamily member 18	8995	The protein encoded by this gene is a cytokine that belongs to the tumor necrosis factor (TNF) ligand family.
TAGAP	Q8N103	T cell activation RhoGTPase activating protein	117289	This gene encodes a member of the Rho GTPase-activator protein superfamily
SPRED1	7Z699	sprouty related EVH1 domain containing 1	161742	The protein encoded by this gene is a member of the Sprouty family of proteins and is phosphorylated by tyrosine kinase in response to several growth factors
SP140	Q13342	SP140 nuclear body protein	11262	ariants of this gene have been associated with multiple sclerosis, Crohn's disease, and chronic lymphocytic leukemia. Alternative splicing results in multiple variants
IL23R	Q5VWK5	interleukin 23 receptor	149233	This protein pairs with the receptor molecule IL12RB1/IL12Rbeta1, and both are required for IL23A signaling
IRGM	A1A4Y4	immunity related GTPase M	345611	The encoded protein may play a role in the innate immune response by regulating autophagy formation in response to intracellular pathogens
ADAM30	Q9UKF2	ADAM metallopeptidase domain 30	11085	This is Adam family - are membrane-anchored proteins structurally related to snake venom disintegrins, and have been implicated in a variety of biological processes involving cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesi
CPEB4	Q17RY0	cytoplasmic polyadenylation element binding protein 4	80315	Sequence-specific RNA-binding protein that binds to the cytoplasmic polyadenylation element (CPE), an uridine-rich sequence element (consensus sequence 5'-UUUUUAU-3') within the mRNA 3'-UTR (PubMed:24990967)
CREB5	Q02930	cAMP responsive element binding protein 5	9586	Binds to the cAMP response element and activates transcription

FASLG	P48023	Fas ligand	56	member of the tumor necrosis factor superfamily. The primary function of the encoded transmembrane protein is the induction of apoptosis triggered by binding to FA
FUT2	Q10981	fucosyltransferase 2	2524	The protein encoded by this gene is a Golgi stack membrane protein that is involved in the creation of a precursor of the H antigen, which is required for the final step in the soluble A and B antigen synthesis pathway
GPX4	P36969	glutathione peroxidase 4	2879	The protein encoded by this gene belongs to the glutathione peroxidase family, members of which catalyze the reduction of hydrogen peroxide, organic hydroperoxides and lipid hydroperoxides, and thereby protect cells against oxidative damage
HMHA1	Q8N103	T cell activation RhoGTPase activating protein	117289	This gene encodes a member of the Rho GTPase-activator protein superfamily. The encoded protein may function as a Rho GTPase-activating protein
IFNAR1	P17181	interferon alpha and beta receptor subunit 1	3454	The protein encoded by this gene is a type I membrane protein that forms one of the two chains of a receptor for interferons alpha and beta. Binding and activation of the receptor stimulates Janus protein kinases, which in turn phosphorylate several proteins, including STAT1 and STAT2
IFNGR2	P38484	interferon gamma receptor 2	3460	This gene (IFNGR2) encodes the non-ligand-binding beta chain of the gamma interferon receptor. Human interferon-gamma receptor is a heterodimer of IFNGR1 and IFNGR2
IL31RA	Q8NI17	interleukin 31 receptor A	133396	The protein encoded by this gene belongs to the type I cytokine receptor family. This receptor, with homology to gp130, is expressed on monocytes, and is involved in IL-31 signaling via activation of STAT-3 and STAT-5
IL6ST	P40189	interleukin 6 signal transducer	3572	he protein encoded by this gene is a signal transducer shared by many cytokines, including interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF), and oncostatin M (OSM)
JAZF1	86VZ6	JAZF zinc finger 1	221895	This gene encodes a nuclear protein with three C2H2-type zinc fingers, and functions as a transcriptional repressor. Chromosomal aberrations involving this gene are associated with endometrial stromal tumors
LACC1	Q8IV20	laccase domain containing 1	144811	This gene encodes an oxidoreductase that promotes fatty-acid oxidation, with concomitant inflammasome activation, mitochondrial and NADPH-oxidase-dependent reactive oxygen species production, and bactericidal activity of macrophage
LGALS9	O00182	galectin 9	3965	The galectins are a family of beta-galactoside-binding proteins implicated in modulating cell-cell and cell-matrix interactions. The protein encoded by this gene is an S-type lectin

NOD2	Q9HC29	nucleotide binding oligomerization domain containing 2	64127	This gene is a member of the Nod1/Apaf-1 family and encodes a protein with two caspase recruitment (CARD) domains and six leucine-rich repeats (LRRs).
NOS2	P35228	nitric oxide synthase 2	4843	Nitric oxide is a reactive free radical which acts as a biologic mediator in several processes, including neurotransmission and antimicrobial and antitumoral activities
PTPN22	Q9Y2R2	protein tyrosine phosphatase, non-receptor type 22	26191	The encoded protein is a lymphoid-specific intracellular phosphatase that associates with the molecular adapter protein CBL and may be involved in regulating CBL function in the T-cell receptor signaling pathway.
RASGRP1	O95267	RAS guanyl releasing protein 1	10125	This gene is a member of a family of genes characterized by the presence of a Ras superfamily guanine nucleotide exchange factor (GEF) domain
RIPK2	O43353	receptor interacting serine/threonine kinase 2	8767	This gene encodes a member of the receptor-interacting protein (RIP) family of serine/threonine protein kinases.

3 Summary of interaction data

We got the Interaction data for all our seed Genes from two sources. We accessed the first source programmatically (from the bio grid database) while the second one was accessed manually through a CSV file downloaded from the IID website. We saved data sets in two different files BioGrid.csv and IID.csv respectively.

3.1 & 3.2 Data Sources and Data collection:

As mentioned above, the interaction data was saved in two different files corresponding to two sources. The names are “Bio_Grid.csv” and “IID.csv” respectively. We attach a few samples from each of the tables respectively.

Bio_Grid.csv

Seed_Gene	**GeneA**	**GeneB**
ATG16L1	55054	55054
ATG16L1	9140	55054
ATG16L1	8517	55054
ATG16L1	55054	1213

IID.csv

UniProt1	**UniProt2**	**symbol1**	**symbol2**
A1A4Y4	O95786	IRGM	DDX58
A1A4Y4	O75385	IRGM	ULK1
A1A4Y4	Q9Y239	IRGM	NOD1
A1A4Y4	Q9NP85	IRGM	NPHS2
A1A4Y4	A1A4Y4	IRGM	IRGM
A1A4Y4	Q9C0C7	IRGM	AMBRA1
A1A4Y4	O15455	IRGM	TLR3

The first table corresponds to data from BioGrid and saves the Gene IDs for each of the interacting proteins. The second table corresponds to data from IID and contains UniProt AC and Gene symbol for each of the interacting proteins.

3.3 Summarize the Main result

No. of Seed Genes in Each DB Respectively	No. of Interacting Proteins in each DB Respectively	Number of Interactions found in each DB Respectively
23	713	1132
26	2379	3844

4 Intersection interactome

We wrote a python script to analyze interaction in the above 2 data sets. As mentioned in the homework PDF file we divided our code for calculating seed genes interactome, union interactome, and intersection interactome

Seed genes interactome:
We built a table which involves interactions from all the Database sources but both proteins involved in the interactions are in the list of Genes provided. We saved the result in the 4.1.csv file

Union interactome:
We built a table which involves interactions from all the Database sources with at least one protein involved in the interactions to be in the list of Genes provided. We saved the result in the 4.2.csv file

Intersection interactome:
We built a table to report the interactions which are confirmed by all three Database sources. We saved the result in the 4.3.csv file

Note: The format of all the tables is same.

Here we attach a few samples from the 4.1.csv. There are some instances in the dataset where the same interaction is present multiple times. This is because two genes can interact with each other in a lot of possible ways, under different conditions, atmosphere and catalyzers. We have a total of 4204 (Not Necessarily Unique) interactions in 4.3.csv. Below is a sample from table 4.3.csv.

GeneA	GeneB	UniprotA	UniprotB
ATG16L1	ATG16L1	Q676U5	Q676U5
ATG16L1	ATG16L1	Q676U5	Q676U5
ATG16L1	NOD2	Q676U5	Q9HC29
ATG16L1	NOD2	Q676U5	Q9HC29
NOD2	ATG16L1	Q9HC29	Q676U5
IRGM	ATG16L1	A1A4Y4	Q676U5
ATG16L1	IRGM	Q676U5	A1A4Y4

5 Enrichment analysis

In Question 4, we built three tables namely 4.1.csv, 4.2.csv , and 4.3.csv. In Question 5, we were supposed to perform Gene Ontology and Pathway Analysis to each of these tables. More specifically, we were supposed to get the unique set of all the Genes that are involved in the interactions for each of these tables, and then use the online portal to perform these tasks. To get the unique set of all the Genes for each of the tables, we wrote a python script using which we are able to get a set(unique Gene ids) for each of the above tables, and then save it into a new file named according to the sources. Thus, we eventually had three new tables named 4.1GO.csv, 4.2GO.csv , and 4.3GO.csv.

To perform the enrichment analysis, we chose Innate DB as our choice. We presented each of three tables (returned by our Software) as input to Innate DB and we received two outputs corresponding to Gene Ontology Analysis and Pathway Analysis. Thus, after we had performed Enrichment analysis on all the tables, we had, as a result, six new tables namely 4.1 GO Analysis.txt, 4.1 Pathway Analysis.xls, 4.2 GO Analysis.xls, 4.2 Pathway Analysis.xls, 4.3 GO Analysis.xls and 4.3 Pathway Analysis.xls. The format for each pair of the tables is same, meaning for example 4.2 Pathway Analysis.xls and 4.3 Pathway Analysis.xls would have the same format as it is the same format returned by Innate DB, however, the entries in each of the tables could be different. The above-mentioned files are attached to our submission (In folder 5).

6 Notes and comments:

Important notes regarding the entire project are presented below:

- There's is considerable difference between the number of interactions returned by Bio Grid (1132) and IID (3844) datasets.
- We observed some inconsistencies between the results that we obtained in steps 3 and 4 using Bio Grid and IID data sets and the results that we saw in the Innate DB site because not all genes are present in the Ontology and Pathway analysis as we observed a different source data source (other than IID and BioGrid) which might lead to some inconsistencies in the analysis and results may be misleading as some genes which might be important for analyzing the pathological condition might be missing in the Innate DB resource
- Some of the interactions are repeated more than once, as it is possible under the presence of a different environment.
- Using different Gene Symbol may return different interactions.
- The code and the related data are present in the corresponding folders (3,4 and 5).
- For retrieving Gene symbols for non seed genes we used a python package mygene and for retriving the Uniprot IDs for non-seed genes to be updated in 4.2.csv and 4.3.csv we used data available in this link. We downloaded the Gene symbol to uniport AC map available in JSON format and used it to update the uniport ids in 4.2.csv and 4.3.csv
- Additional metadata that we used for our analysis is included in the metadata folder, which includes Gene Map.xls for mapping between Gene symbol, Uniprot ID and Gene ID
- As per our analysis, there is similar gene set for both union interactome and intersection interactome so the results of enrichment (both Gene Ontology and Pathway) analysis will be more or less similar.