Whitney Domico

The dopamine hypothesis is outdated: A meta-analysis of the complexities in the biochemistry of schizophrenia

R Script Instructions

Go to <https://github.com/domicowe/BIO247> -> BIO247Project -> BIO247ProjectSubmissions and download zip file “DomicoProjectZip.zip”

*Note: You can also look at the README, which is also in the file.*

Open “DomicoProjectScript.R” in R software environment and set working directory to DomicoProjectZip.zip

Run code through line 664. Here is what you should expect it to do:

* Import data from PMC articles (which are provided in zip file)
  + The data was pulled from the following articles
    - [PMC3077530](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3077530)
    - [PMC2775422](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775422)
    - [PMC3912837](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912837)
    - [PMC2890845](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2890845)
    - [PMC4724864](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4724864)
    - [PMC6927206](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6927206)
    - [PMC3872086](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872086)
    - [PMC3827979](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3827979)
    - [PMC4059435](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4059435)
    - [PMC3905728](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3905728)
* Clean and transform data into a singular data frame (“MainData” df in environment”)
* Create scatterplot and bar graph of relative frequency of SNPs per chromosome
  + *Note:* *Chrs 6, 11, and 22 are >2 SDs above mean.*
* Import and clean NCBI data for Chr 6 (“Chr6DataNCBI” df in environment) which includes genes with their positions along the chromosome.
* Do statistical analysis on these to compare against NCBI data and find genes where these SNPs overlap.
* Repeat last two steps with Chrs 11 and 22.
* Combine data from Chrs 6 and 11 (the data found no significant genes for Chr 22 so its data was dropped).
* Clean data (“genedata” df in environment) and import and clean NDEx pathway data.
* Create excel file (“diseasesdf.xlsx”) of common disorders associated with these pathways.

You should now have a file “diseasesdf.xlsx” saved to the DomicoProjectZip folder. Open this and define new column “Immune.” If the disease in the previous column is autoimmune, put a “1” in the “Immune” column. If not, put a “0.” Save the file (same name 🡪 diseasesdf.xlsx) into DomicoProjectZip.zip. The “answers” are provided below if you don’t want to look it up.

|  |  |  |
| --- | --- | --- |
| **Var1** | **Freq** | **Immune** |
| addison disease | 1 | 1 |
| adrenal gland pheochromocytoma | 1 | 0 |
| alopecia | 1 | 1 |
| arbd | 1 | 0 |
| bladder papillary urothelial neoplasm | 1 | 0 |
| blepharocheilodontic syndrome 2 | 1 | 0 |
| breast cancer | 1 | 0 |
| cardiomyopathy | 1 | 1 |
| crohn disease | 1 | 1 |
| deafness | 1 | 0 |
| entropion | 1 | 0 |
| fanconi renotubular syndrome 2 | 1 | 0 |
| gout | 2 | 0 |
| graves disease | 1 | 1 |
| hashimoto thyroiditis | 1 | 1 |
| hydrolethalus syndrome | 1 | 0 |
| hyperuricemia | 1 | 0 |
| idiopathic pulmonary fibrosis | 1 | 1 |
| joubert syndrome 39 | 1 | 0 |
| juvenile idiopathic arthritis | 1 | 0 |
| leprosy 4 | 1 | 0 |
| lyme disease | 1 | 1 |
| lysosomal storage disease | 1 | 0 |
| mechel syndrome | 1 | 0 |
| multiple sclerosis | 2 | 1 |
| myocardial infarction | 1 | 1 |
| narcolepsy | 1 | 0 |
| nephrolithiasis | 1 | 1 |
| neuromyelitis optica spectrum disorder | 1 | 1 |
| olecranon bursitis | 1 | 0 |
| psoriatic arthritis | 2 | 1 |
| rheumatoid arthritis | 2 | 1 |
| rosacea | 1 | 0 |
| tonsilitis | 1 | 0 |
| toxic shock syndrome | 1 | 0 |
| type 1 diabetes | 2 | 1 |

Run code lines 670-935. Here’s what you should expect it to do:

* Calculate percentage of disorders that are autoimmune.
* Import gene data about Chr 6 and compare known SNPs against genes for the histocompatibility complex.
* Calculate percentage of SNPs that overlap with histocompatibility genes.
* Calculate percentage of each pathway in implicated genes.
* Find common words in pathway titles (“finaldata” df in environment) and cut to only include statistically significant ones (>2 SDs).
* Save excel file “finaldata.xlsx.”

Open file “finaldata.xlsx.” Define new column “kept.” If you decide that the word given is relevant, put a “1” in kept column. If not, put a “0.” The words that I decided were important for my purposes are provided below, but it could be argued that these are not important or that others are important. Save the file (same name 🡪 finaldata.xlsx) in DomicoProjectZip folder.

|  |  |  |  |
| --- | --- | --- | --- |
| **Var1** | **Freq** | **relfreq** | **kept** |
| rna | 164 | 1 | 0 |
| human | 159 | 0.969512 | 0 |
| protein | 117 | 0.713415 | 0 |
| mirna | 107 | 0.652439 | 0 |
| interactions | 104 | 0.634146 | 0 |
| network | 101 | 0.615854 | 0 |
| atlas | 99 | 0.603659 | 0 |
| blood | 96 | 0.585366 | 0 |
| cell | 92 | 0.560976 | 0 |
| expression | 78 | 0.47561 | 0 |
| brain | 58 | 0.353659 | 1 |
| hpa | 58 | 0.353659 | 1 |
| v21.0 | 57 | 0.347561 | 0 |
| regulatory | 56 | 0.341463 | 0 |
| types | 54 | 0.329268 | 0 |
| gene | 53 | 0.323171 | 0 |
| immune | 51 | 0.310976 | 1 |
| response | 51 | 0.310976 | 0 |
| tissue | 50 | 0.304878 | 0 |
| sapiens | 48 | 0.292683 | 0 |
| full | 46 | 0.280488 | 0 |
| string | 46 | 0.280488 | 0 |
| associations | 41 | 0.25 | 0 |
| only | 40 | 0.243902 | 0 |
| consensome | 36 | 0.219512 | 1 |
| transcriptomic | 36 | 0.219512 | 1 |
| bioplex | 32 | 0.195122 | 0 |
| humannet | 32 | 0.195122 | 0 |
| proteinprotein | 30 | 0.182927 | 0 |
| from | 28 | 0.170732 | 0 |
| biogrid: | 27 | 0.164634 | 0 |
| composite | 26 | 0.158537 | 0 |
| normal | 26 | 0.158537 | 0 |
| high | 25 | 0.152439 | 0 |
| scored | 25 | 0.152439 | 0 |
| disgenet | 24 | 0.146341 | 0 |
| links | 24 | 0.146341 | 0 |
| lines | 22 | 0.134146 | 0 |
| pathway | 21 | 0.128049 | 0 |

Run code lines 936-end. Here’s what you should expect it to do:

* Cut user-cited unimportant words.
* Create second bar plot of only user-cited important words.