An exploration of the Gene Curation Coalition Database of Gene-Disease Mappings

Introduction

The Gene Curation Coalition is a unified database of the results and confidence classifications of many different groups and resources. There was a lot of data produced experimentally but the publication mediums and organisations were not standardised which impeded efforts to amalgamate different organisations' research. They took 241 survey responses to help standardise how they would grade the validity of pieces of data and organise it within the database [4]. As of the 17th of November 2023, the GenCC database has 18,504 submitted classifications with almost 4,888 unique genes from 12 different submitters. The company with the most submissions was Ambry Genetics who mostly submitted supportive/limited. They were followed by ClinGen who submitted a lot less, but what they did submit was overwhelmingly definitive. This report investigates some of the disease relationships present in the GenCC Database as of November 2023 and explores some of the analytics that can be performed with it.

Part 1

Methods

For the first task, I wrote a Python script to read the GenCC file as a tsv to a Pandas data frame. From there I used the inbuilt panda methods to gain a name-value pair of unique “disease_title” names and their frequency in the column to ascertain the top 10 most frequent diseases without the “No known disease relationship” classification title. I then visualised the results with a bar chart produced in Microsoft Excel.

For the next tasks, I used the same method but accessed the classification confidence category and the provenance submission category to count occurrences and visualised the results with another Excel bar chart.

In the extension, I looked at the paper submission dates versus the dates they were run. I used the panda's data analysis library to get statistics of the differences and visualised the data with Matplotlib.

Results

1.

Disease Name	Frequency
Complex Neurodevelopmental Disorder	156
Leigh Syndrome	127
Retinis Pigmentosa	104
Nonsyndromatic Genetic Hearing Loss	93
Hearing Loss, Autosomal Recessive	87
Mitochondrial Disease	76
Syndromic Intellectual Disability	73
Primary Ciliary Dyskinesia	58
Autosomal Dominant Nonsyndromic Hearing Loss	55
Male Infertility with Azoospermia or Oligozoospermia	54

Table 1 Top ten diseases with the most genes affiliated with them in the GenCC flat file

2.

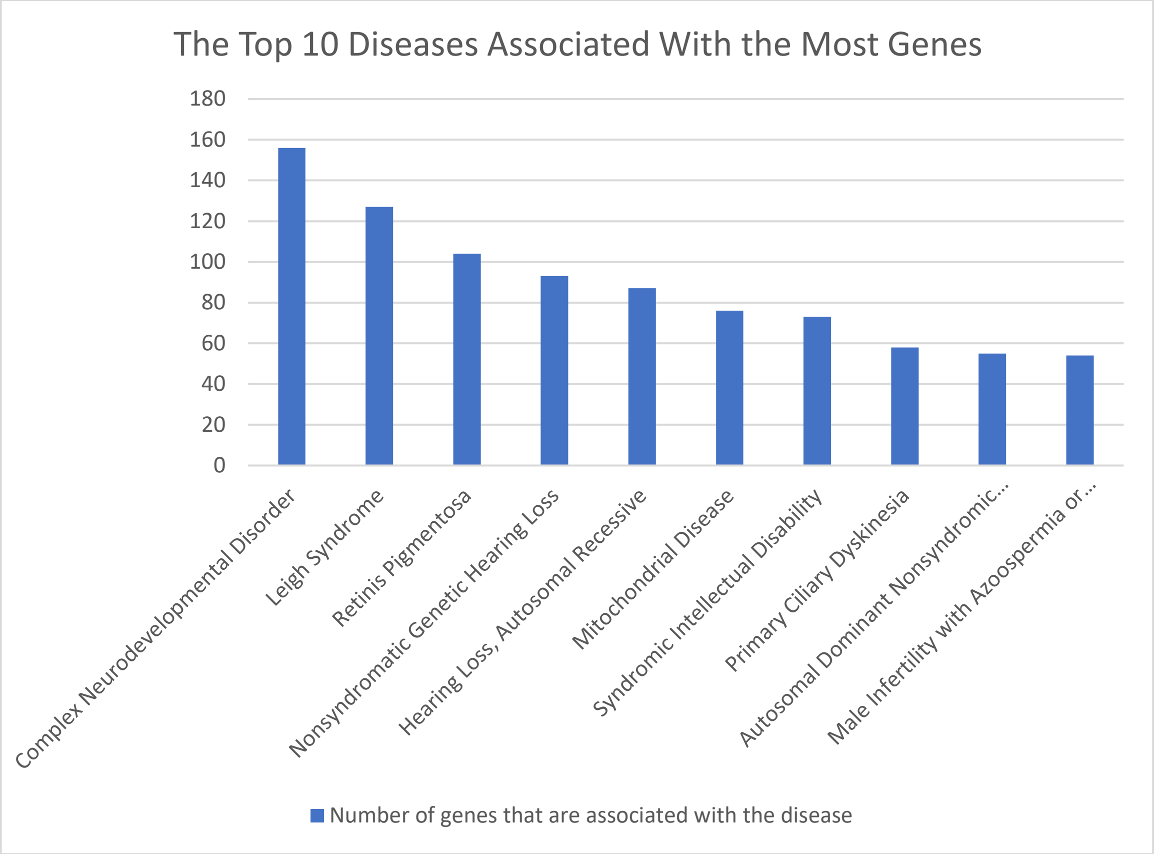

Figure 1 A graphic representation of Table 1

3.

Classifications	**Count**
Definitive	4178
Strong	4720
Supportive	5330
Moderate	1791
Limited	2030
Disputed	182
Refuted	27
No Known Disease Relationship	246

Table 2 Frequency of different evidence strength classifications for each gene-disease entry

4.

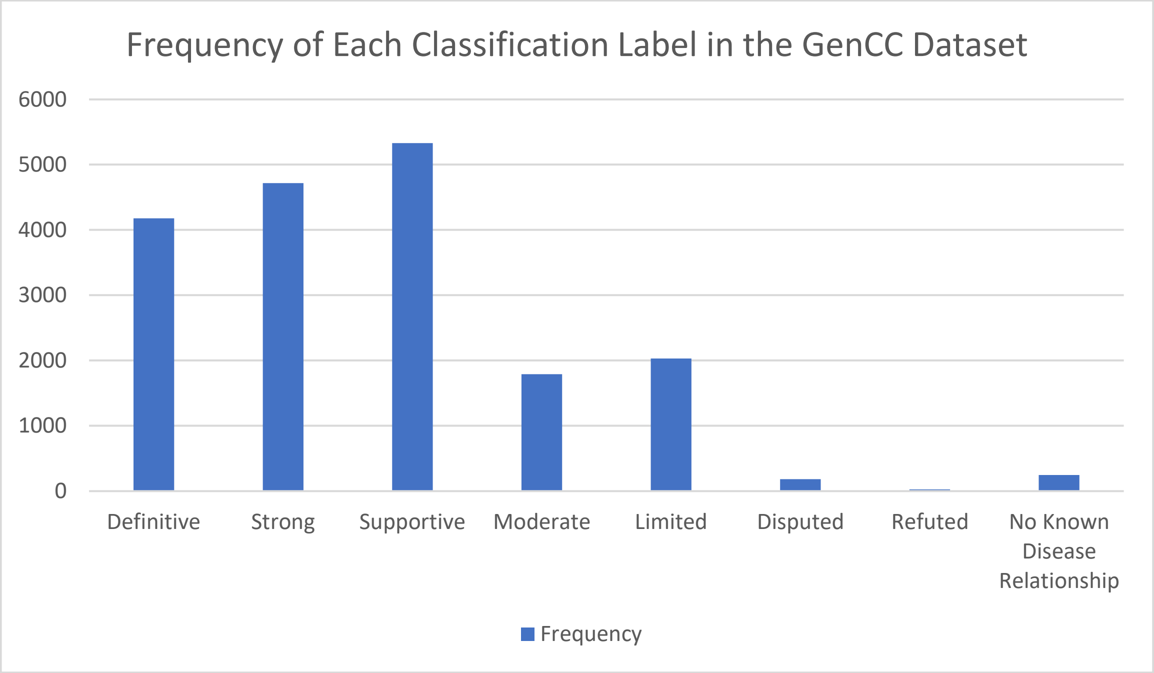

Figure 2 Graphic representation of Table 2

5.

Provenance Category	**Count (Total: 18504)**
PubMed	1706
Digital Object Identifiers	15550
None	1248
Other	0

Table 3 Frequency of types of provenance entry

Discussion and Extension – Rate of Table Additions over the Years

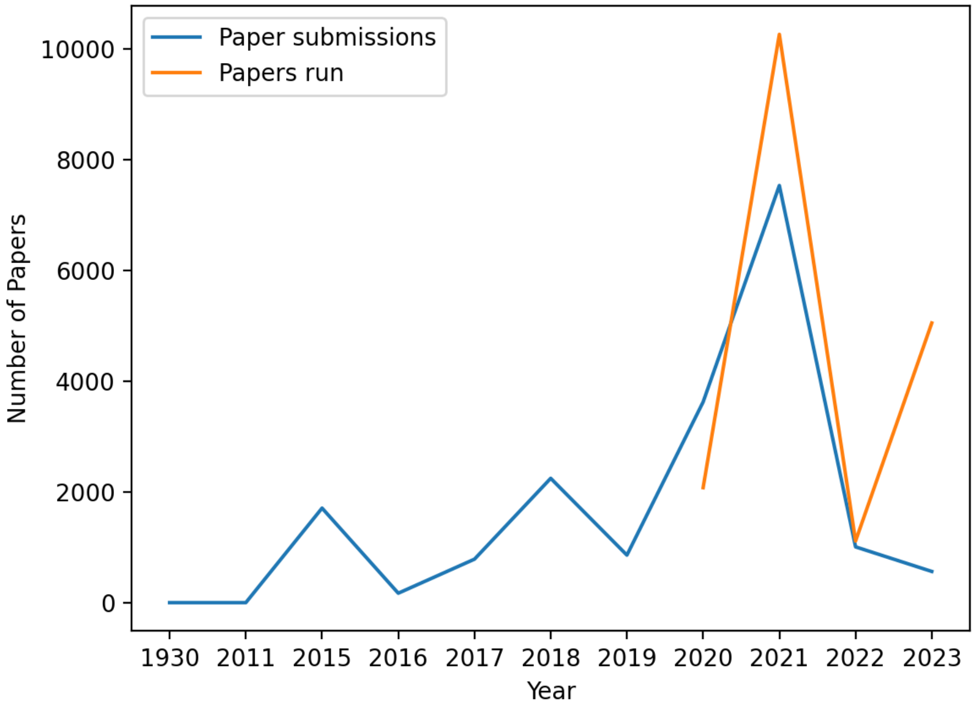

Figure 3 How many papers were submitted to the data set and run for each year?

It should be noted there is a non-linear x-axis as the one paper included from 1930 would have severely compressed the graph to the right. Upon analysing the delays between paper submission and running (excluding the 1930 entry), I found the average latency to be 686 days with a standard deviation and median of 888 days and 333 days respectively.

Part 2

Methods

For the first part, I used the same methods for extracting the flat file data as in Part 1. I then counted all of the rows and the number of them with an MOI of the form x:y where x,y could be anything from which I formed a proportion.

For the second and third tasks, I used the Pronto library to read the data from the Obo file. I expanded the code in the first question to also generate a list of unique MOIs and their corresponding frequency counts which was then used as a reference point to find the corresponding names from the obo file. The difference between different Modes of Inheritance was massive so to demonstrate the data better I elected to use a logarithmic scale for the y-axis.

Results

1. There are 18504 entries and every row has an MOI curie.
2.

MOI Curie	Mode of Inheritance Name	Number of genes
HP:0000005	Mode of Inheritance	329
HP:0000006	Autosomal Dominant Inheritance	7677
HP:0000007	Autosomal Recessive Inheritance	9122
HP:0001417	X-Linked Inheritance	952
HP:0001419	X-Linked Recessive Inheritance	144
HP:0001423	X-Linked Dominant Inheritance	26
HP:0001427	Mitochondrial Inheritance	100
HP:0001442	Typified by Somatic Mosaicism	9
HP:0001450	Y-Linked Inheritance	2
HP:0010984	Digenic Inheritance	1
HP:0012274	Autosomal Dominant Inheritance with Paternal Imprinting	5
HP:0012275	Autosomal Dominant Inheritance with Maternal Imprinting	4
HP:0032113	Semidominant Inheritance	133

Table 4 What does each MOI Curie mean and how many genes do they apply to in the GenCC dataset

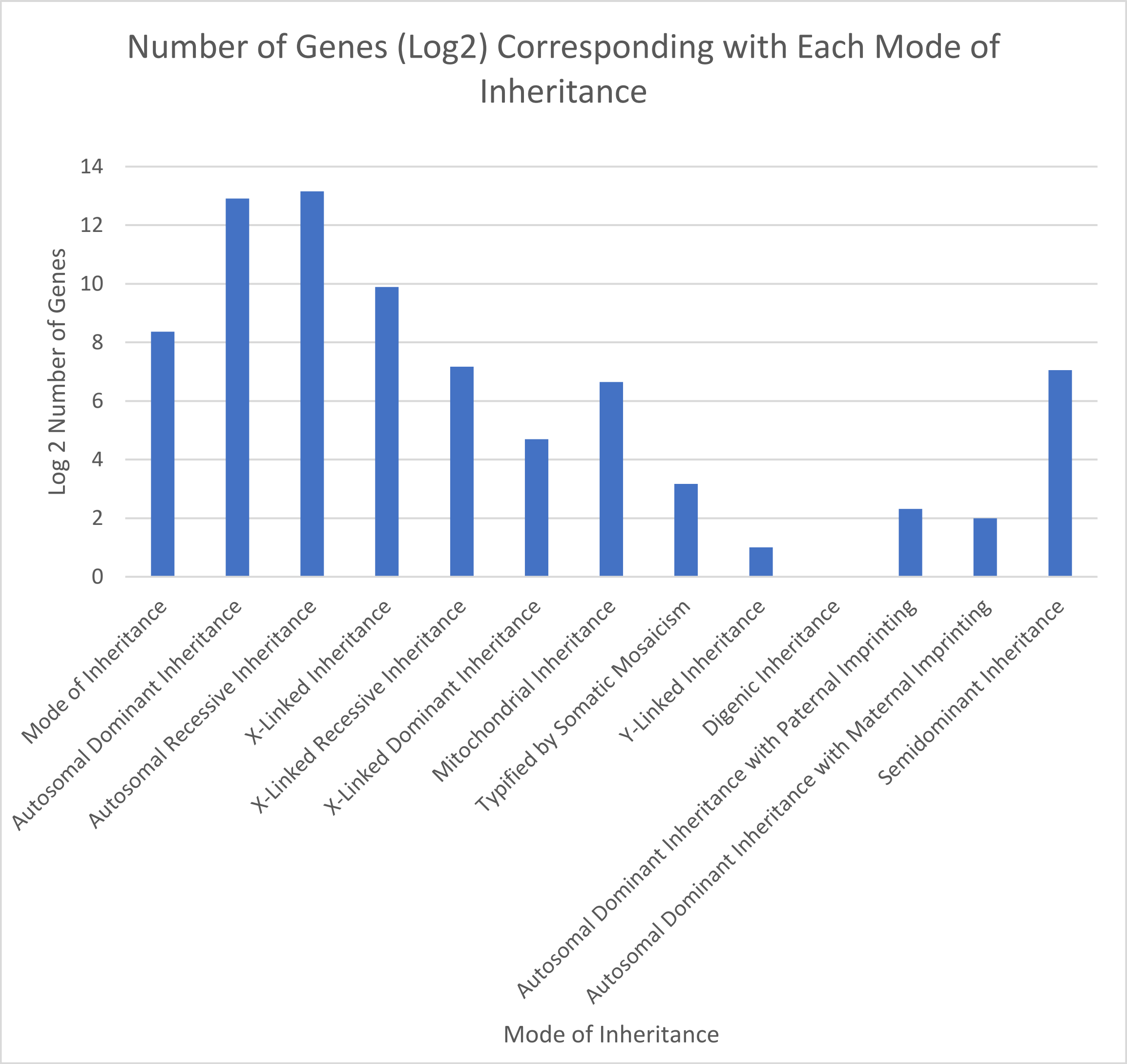

Table 5 Bar chart to show the number of genes for each mode of inheritance. I have taken the log of each number to ease visualisation given the large range of values.

4. In discussion section

Discussion

The sex chromosomes consist of an X and a Y. The X chromosome is the bigger of the two and contains over 1400 genes [1] whereas the Y is much smaller and contains only around 200 genes. An autosome on the other hand is one of the other 22 pairs of chromosomes and can contain 750 – 3000+ genes on each [2]. The disparity between the number of autosomal MOIs and X/Y MOIs is reflected here as Sex-related genes represent about 10% of the genes in the file and about 5% in the human genome [3].

Part 3

Methods

For the exercises in this part, I used the Pronto library to read the Mondo Obo file and the panda library again to read through the GenCC tsv file. For each count, I also excluded any rows that stated their classification as “No known disease relationship”.

Results

1. The Mondo ID for nervous system disorder is MONDO:0005071.
2. There are 5588 subclasses of MONDO:0005071.
3.

MONDO ID	Name
MONDO:0002602	Congenital Nervous System Disorder
MONDO:0002320	Central Nervous System Disorder
MONDO:0002977	Autoimmune Disorder of the Nervous System
MONDO:0003569	Cranial Nerve Neuropathy
MONDO:0003620	Peripheral Nervous System Disorder
MONDO:0004466	Neuronitis
MONDO:0004618	Diplegia of Upper Limb
MONDO:0005283	Retinal Disorder
MONDO:0005287	Developmental Disability
MONDO:0005391	Restless Legs Syndrome

Table 6 First 10 MONDO terms retrieved under nervous system disorder.

4.

MONDO ID	Disease Name	Gene Count
MONDO:0100038	Complex Neurodevelopmental Disorder	156
MONDO:0009723	Leigh Syndrome	127
MONDO:0019200	Retinitis Pigmentosa	104
MONDO:0019497	Non-syndromic Genetic Hearing Loss	93
MONDO:0019588	Autosomal Recessive Hearing Loss	87
MONDO:0000508	Syndromic Intellectual Disability	73
MONDO:0019587	Autosomal Dominant Non-syndromic Hearing Loss	55
MONDO:0019502	Autosomal Recessive Non-syndromic Intellectual Disability	53
MONDO:0100062	Developmental and Epileptic Encephalopathy	52
MONDO:0001071	Intellectual Disability	50

Table 7 The MONDO terms under nervous system disorder with the most associated genes in the GenCC dataset

5.

Gene	Number of NSD_GenCC Entries
SCN4A	17
MECP2	16
POMGNT1	15
ARX	15
SCN1A	15
TTN	15
COL6A3	15
MYO7A	14
PLP1	14
ATP1A3	14

Table 8 The top 10 genes by number of entries relating to one of the terms under nervous system disorder.

6. There are 2085 genes in the NSD_GenCC Dataset that are not labelled with no known disease relationship.

Extension – Analysis of Disorder of the Visual System

The same method was carried out for this extension as in the rest of the part.

1. MONDO:0024458 is the accession ID for the disorder of the visual system
2. There are 1932 subclasses of MONDO:0024458.
3.

MONDO ID	Name
MONDO:0002135	Optic Nerve Disorder
MONDO:0004746	Myopathy of Extraocular Muscle
MONDO:0005328	Eye Disorder
MONDO:0021084	Vision Disorder
MONDO:0001746	Optic Disk Drusen
MONDO:0002003	Papilledema
MONDO:0002640	Optic Nerve Neoplasm
MONDO:0003608	Optic Atrophy
MONDO:0005885	Optic Neuritis
MONDO:0006649	Anterior Ischemic Optic Neuropathy

Table 9 First 10 MONDO terms retrieved under the disorder of visual system.

4. There are 773 diseases associated with disorder of visual system

MONDO ID	Disease Name	Gene Count
MONDO:0019200	Retinitis Pigmentosa	104
MONDO:0018997	Noonan syndrome	34
MONDO:0015993	Cone-rod Dystrophy	29
MONDO:0018998	Leber Congenital Amaurosis	24
MONDO:0020376	Early-onset Nuclear Cataract	18
MONDO:0020344	Postsynaptic Congenital Myasthenic Syndrome	16
MONDO:0021548	Total Early-onset Cataract	15
MONDO:0010788	Leber Hereditary Optic Neuropathy	15
MONDO:0010168	Usher Syndrome Type 1	14
MONDO:00	Aniridia-cerebellar Ataxia-intellectual Disability Syndrome	14

Table 10 The MONDO terms under the disorder of the visual system with the most associated genes in the GenCC dataset

5.

Gene	DVS_GenCC Entries
PAX6	17
GBA1	13
BEST1	13
ITPR1	12
CRYBB2	11
PRPH2	11
TGFBI	11
LZTR1	10
GLB1	10
IDUA	10

Table 11 The top 10 genes by number of entries relating to one of the terms under disorder of the visual system.

6. There are 650 genes in the DVS_GenCC dataset that are not labelled with no known disease relationship

Part 4

Methods

For the first part, I gathered corresponding arrays, one of which held the gene name, and the other was a list of all diseases associated with it. I then through all of the different lists of gene-affected diseases and constructed a pair name based on the sorted concatenation of the two names. Sorted was important so that the order they were registered in did not separate identical pairs. I again excluded any rows that were classified with “No Disease Relationship”.

Results

1.

Disease Pair	Gene Count
MONDO:0019497,MONDO:0019588	54
MONDO:0019497,MONDO:0019587	29
MONDO:0009723,MONDO:0016815	28
MONDO:0019234,MONDO:0019609	13
MONDO:0019587,MONDO:0019588	11
MONDO:0015802,MONDO:0100038	11
MONDO:0000508,MONDO:0014699	10
MONDO:0000508,MONDO:0100038	10
MONDO:0018998,MONDO:0019200	9
MONDO:0014699,MONDO:0015802	8

Table 12 The top 10 MONDO term pairs with the most mutually associated genes

2. There are 92 edges in the graph as stated by number_of_edges()

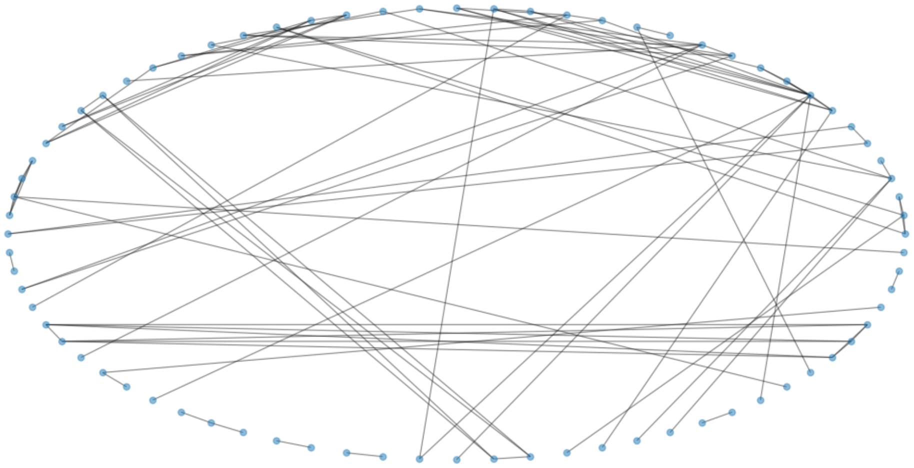

Figure 4 Unlabelled and non-clustered diagram of the nodes in the network

3.

Community Number	Number of diseases
1	12
2	7
3,4	6
5,6,7	5
8,9	4
10,11,12,13	3
14,15,16,17,18	2

Table 13 An ordered list of all of the identified communities in the network with the number of diseases in each

4.

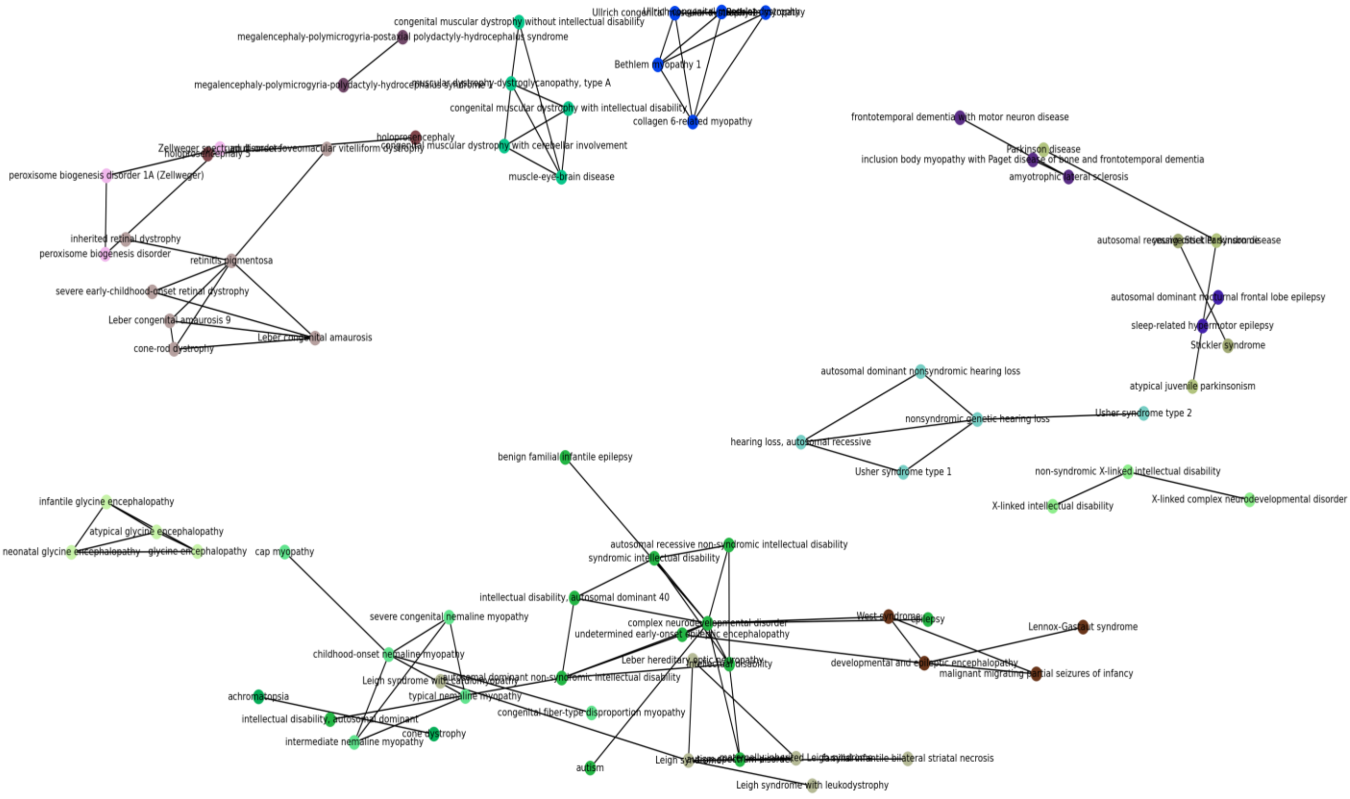
congenital muscular dystrophy without intellectual disability
megalencephaly-polymicrogyria-postaxial polydactyly-hydrocephalus syndrome
megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 1
muscular dystrophy-dystroglycanopathy, type A
congenital muscular dystrophy with intellectual disability
holoprosencephaly
congenital muscular dystrophy with cerebellar involvement
muscle-eye-brain disease
Zellweger spectrum disorder
holoprosencephaly 5
adult-onset foveomacular vitelliform dystrophy
peroxisome biogenesis disorder 1A (Zellweger)
inherited retinal dystrophy
peroxisome biogenesis disorder
retinitis pigmentosa
severe early-childhood-onset retinal dystrophy
Leber congenital amaurosis 9
Leber congenital amaurosis
cone-rod dystrophy
Ullrich congenital muscular dystrophy
Bethlem myopathy 1
collagen 6-related myopathy
frontotemporal dementia with motor neuron disease
Parkinson disease
inclusion body myopathy with Paget disease of bone and frontotemporal dementia
amyotrophic lateral sclerosis
Stickler syndrome
autosomal dominant nocturnal frontal lobe epilepsy
sleep-related hypermotor epilepsy
atypical juvenile parkinsonism
autosomal dominant nonsyndromic hearing loss
nonsyndromic genetic hearing loss
Usher syndrome type 2
hearing loss, autosomal recessive
Usher syndrome type 1
non-syndromic X-linked intellectual disability
X-linked intellectual disability
X-linked complex neurodevelopmental disorder
benign familial infantile epilepsy
infantile glycine encephalopathy
atypical glycine encephalopathy
neonatal glycine encephalopathy
glycine encephalopathy
cap myopathy
autosomal recessive non-syndromic intellectual disability
syndromic intellectual disability
intellectual disability, autosomal dominant 40
severe congenital nemaline myopathy
complex neurodevelopmental disorder
undetermined early-onset epileptic encephalopathy
West syndrome
epilepsy
Lennox-Gastaut syndrome
childhood-onset nemaline myopathy
Leber hereditary optic neuropathy
intellectual disability
developmental and epileptic encephalopathy
malignant migrating partial seizures of infancy
Leigh syndrome with cardiomyopathy
autosomal dominant non-syndromic intellectual disability
typical nemaline myopathy
achromatopsia
congenital fiber-type disproportion myopathy
intellectual disability, autosomal dominant
cone dystrophy
intermediate nemaline myopathy
autism
Leigh syndrome
autism spectrum disorder
infantile bilateral striatal necrosis
Leigh syndrome with leukodystrophy

Figure 5 Network of the diseases that share the most common genes relating to nervous system disorder. They are colour-coded by the community

References

[1] National Human Genome Research Institute. (2021, July 22). X Chromosome. National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/X-Chromosome-facts

[2] National Human Genome Research Institute. (2023, November 20). Autosome. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Autosome

[3] National Center for Biotechnology Information (US). Genes and Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1998-. Chromosome Map. Available from:https://www.ncbi . nlm.nih.gov/books/NBK22266 /

[4] Marina T. DiStefano, Scott Goehringer, Lawrence Babb, Fowzan S. Alkuraya, Joanna Amberger, Mutaz Amin, Christina Austin-Tse, Marie Balzotti, Jonathan S. Berg, Ewan Birney, Carol Bocchini, Elspeth A. Bruford, Alison J. Coffey, Heather Collins, Fiona Cunningham, Louise C. Daugherty, Yaron Einhorn, Helen V. Firth, David R. Fitzpatrick, Rebecca E. Foulger, Jennifer Goldstein, Ada Hamosh, Matthew R. Hurles, Sarah E. Leigh, Ivone U.S. Leong, Sateesh Maddirevula, Christa L. Martin, Ellen M. McDonagh, Annie Olry, Arina Puzriakova, Kelly Radtke, Erin M. Ramos, Ana Rath, Erin Rooney Riggs, Angharad M. Roberts, Charlotte Rodwell, Catherine Snow, Zornitza Stark, Jackie Tahiliani, Susan Tweedie, James S. Ware, Phillip Weller, Eleanor Williams, Caroline F. Wright, Thabo Michael Yates, Heidi L. Rehm,The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources,Genetics in Medicine,Volume 24, Issue 8,2022,Pages 1732-1742,ISSN 1098-3600, https://doi.org/10.1016/j.gim.2022.04.017.(https://www.sciencedirect.com/science/article/pii/S109836002200746)

[5] Genetic Counseling Collective. (2023, November 17). Statistics. The Genetic Counseling Collective. https://search.thegencc.org/statistics

[6] (Introduction Image) Genetic Counseling Collective. (2023, November 22). Genetic Counseling Collective. https://search.thegencc.org/

Library Versions

Matplotlib – v3.7.2, NetworkX – v3.2.1, Numpy – v1.23.5, Pandas – v2.0.3, Pronto – v2.5.5

Data Version

Mondo Data – 2023-09-12, HP Data – 2023-10-09

GenCC Data – 2023-09-11 https://search.thegencc.org/download/action/submissions-export-tsv