An Exploration of the Gene Curation Consortium (GenCC) Database of Gene-Disease Mappings

Bioinformatics 1 (INFR11160) – Coursework 2

Exam No. B182394

Introduction

The Gene Curation Coalition (GenCC) [1], is a collection of gene-disease relationships and importantly includes information on the validity of these relationships. This is useful as it not only collates the data from many groups, giving us more data to make inferences from, but also helps evaluate the trustworthiness of entries so that we can define levels of evidence for different genes' role in disease. The dataset is publicly available and thus its entries can be scrutinised more thoroughly but due to the large number of entries and the quantity of new data being added, in reality it is probably much harder to validate and verify every single entry, as we will find later on. In this report I aim to explore the features of the GenCC database, and how it can be utilised by a bioinformatician to assist in their research efforts, as well as try to uncover some broader trends within the dataset.

Data & Methods

Datasets:

Firstly, we use this GenCC database, as described in the intro this is a public gene-disease relation database, with a focus on evaluating gene-disease validity, and developing consistent assessment terminology through the centralisation of information from many different groups. We use a flat-file obtained from GenCC listed in the coursework specification.

We also use a human phenotype ontology file obtained from the NCBO BioPortal. The BioPortal is the world's most comprehensive repository of biomedical ontologies [6]. Ontologies provide essential domain knowledge, and we will use it particularly here for finding official names for modes of inheritance. [7]

Finally, we use a MONDO disease ontology file. This is a disease definitions database which aims to "harmonize disease definitions across the world" [3]. This database also contains the existing data models for diseases, making it useful cross referencing a disease across multiple databases.

Methods:

All tasks conducted during this project were completed in Jupyter Notebooks using a Python 3.9 kernel. Several packages were employed in these notebooks: Pandas for data handling, NumPy for large-scale array operations, Matplotlib Pyplot and Seaborn for plotting graphs and creating figures, and Pronto for reading ontology files.

Part 1

In this part, Pandas was used to read and manipulate our GenCC file. The only major design decision made in this part was how to handle duplicate elements found in the GenCC file. I've chosen to only count unique disease-gene entries. For the tasks I utilised the Pandas groupby function (on different columns) with the count function, then filtered out all counts apart from uuid. This sequence of functions groups a column of a pandas dataframe, producing counts for each column i.e. how many elements were in that collapsed column. I filter by uuid to only view the counts from the uuid column, this is because every element in the DB has a uuid value so will ensure that all rows in a group contribute to the count. In Task 5 I made the assumption that all 'http…' links had DOI provenances, and my totals of PMID, DOI and No Provenance summed to the total size of the table, so this seemed accurate. While I could have only counted an entry as a DOI provenance if it included an explicit DOI in its URL, this is not technically true as there exist entries with non-PMID, non-DOI sources but these are still from more trustworthy sources (for example PanelApp and the NHS Genomic Medicine Service Panels Resource [11]) I've chosen to categorise them as DOI as I think they are in a similar level of trustworthiness to something with a DOI. To categorise each provenance, I first checked if the value was a float, this is because no provenance is of NaN type which is technically a float, then I checked the first 4 characters of the provenance string and if it was 'PMID' or 'http'.

Part 2
When completing this part, the GenCC database was a bit deceptive. The database seems to use default values for moi_curie, so performing an intuitive operation like dropping all rows with no moi_curie would end up dropping no rows. After plotting my modes of inheritance, I noticed one of the modes of inheritance wasn't a mode of inheritance at all but instead an entry from the Human Phenotype Ontology file that was simply "Mode of Inheritance", after discovering this I found it is equivalent to "Unknown Inheritance" [2], so I cleaned all entries with where any moi related column (moi_curie, submitted_as_moi_id, submitted_as_moi_name, moi_title) in the dataset was set to 'Unknown', 'Unknown inheritance' or had a NaN value. This removed the unknown inheritance values, and from here I find cross-references and combine datasets, recording and plotting results.

Part 3
First, I found the nervous system disorder ID by iterating through all terms in the Mondo ontology file. After this, I used Pronto to get a list of the subclasses of nervous system disorder, this included the nervous system disorder element which was removed. In task 4 to generate a Nervous System Disorder GenCC (NSD_GCC) dataframe I used the NumPy function 'isin' to create a query for the GenCC dataframe, where I filtered only rows that included a disease_curie that was in an array of all the obtained NSD subclass disease_curies. I used a similar method to count genes as I used in previous parts of grouping counting and this time sorting values so I could see the top 10. Finally, to find the number of unique genes in this new NSD_GCC dataframe, I used the Pandas 'unique' function which returns an array of unique elements in a given column, then finding the length of this was trivial. Note that the MONDO ontology file includes obsolete data, where an entry has been replaced with another, I chose to keep these obsolete ontologies just in case the GenCC file included obsolete MONDO IDs.

Part 4
For this part, I retrieved the GenCC dataset and Mondo ontology file as before, using pronto to retrieve the subclasses of Nervous System Disorder (NSD). Following this, I again used the 'isin' operator to filter only entries of these diseases in the GenCC dataframe, here I also used the 'unique()' method provided by pandas to filter out duplicate entries. I then use the itertools function, 'combinations()' to get all possible disease pairs from the remaining GenCC entries. I then create a python dictionary, where a disease_curie is the key and the value is a set of its genes (in gene curie format), this is more efficient than repeatedly calculating the sets for each disease when within the main loop while now we can just look them up in the dictionary. I then loop through all previously generated disease sequences, I then calculate the intersection of the pair and add the length of this to another dictionary where the key is (disease1_curie, disease2_curie). Finally, I sort these values so I can see the largest disease pair intersections.

From this data I then plotted networks, only including edges, where the disease pair intersection is greater than or equal to 3, this graph was not very clear or useful due to the large number of nodes, so I reduced the graph to only nodes with degree greater than 1 (Figure 9), to see the full graph see Figure 10 in the Appendix. After this I grouped the graph nodes into communities, recording the top ten largest communities. Finally, I plot just the communities that are greater than 1 aka the 'not_lonely_communities' as to avoid cluttering the new clustered graph with communities that only contain a single element. I tweaked the spring layout parameters to get a sparser graph, so communities were easier to identify, and ran multiple times till it produced a clear result.

Results

(Note: all non-table, non-plot results are included in the Discussion section.)

Part One – Summary Analysis of GenCC Data:

Disease Name	Number of Associated Genes
male infertility with azoospermia or oligozoospermia due to single gene mutation	53
Syndromic intellectual disability	66
mitochondrial disease	76
Tourette syndrome	76
hearing loss, autosomal recessive	72
schizophrenia	88
nonsyndromic genetic hearing loss	85
retinitis pigmentosa	90
Leigh syndrome	113
complex neurodevelopmental disorder	149

Table 1 - (Top 10) Number of Genes associated with each Disease.

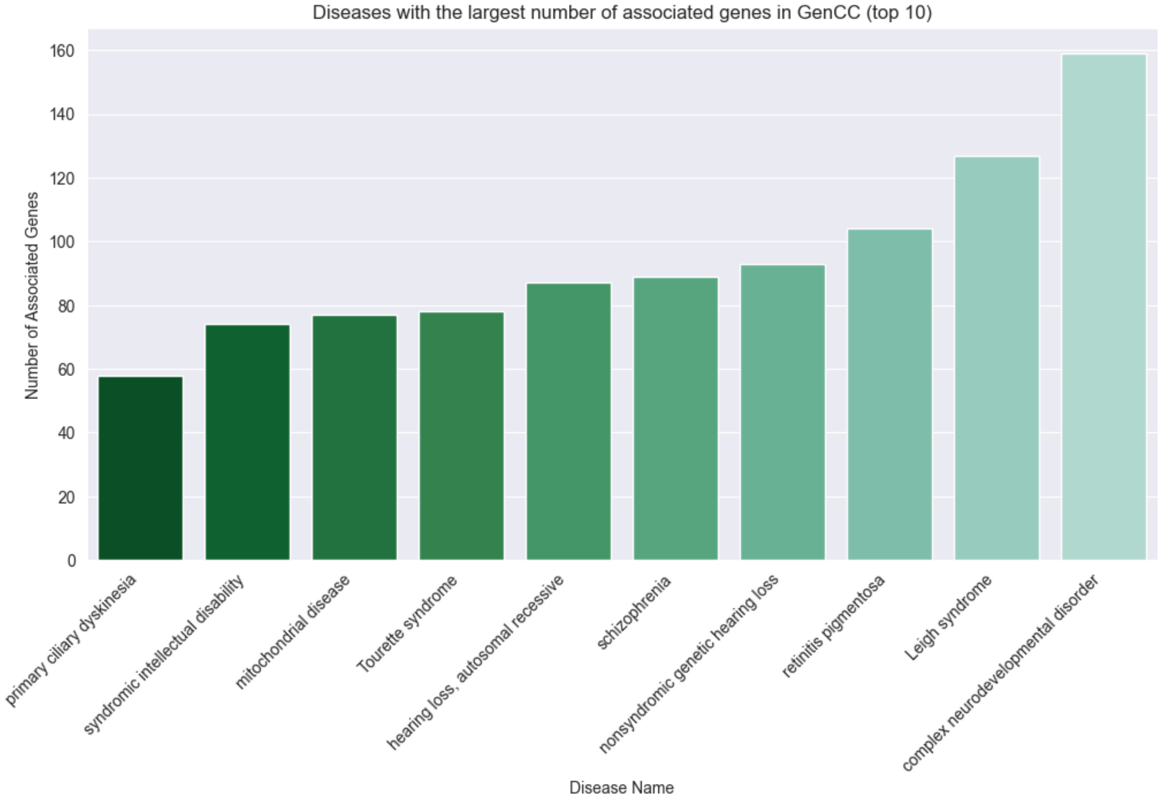

Figure 1 - Plot of Table 1

Classification Title	Number of Associated Genes
Definitive	4178
Strong	4720
Supportive	5330
Moderate	1791
Limited	2030
Disputed Evidence	182
Refuted Evidence	27
No Known Disease Relationship	246

Table 2 - Classification categories of GenCC entries

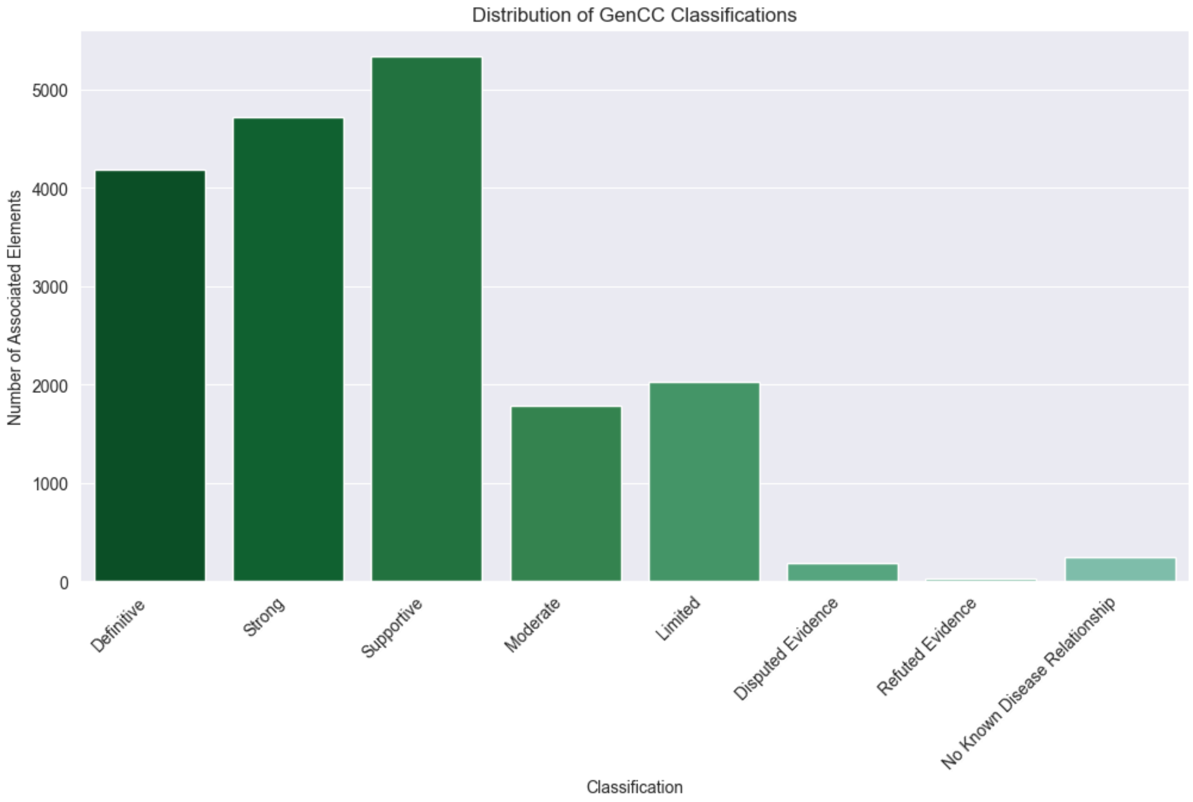

Figure 2 - Plot of Table 2

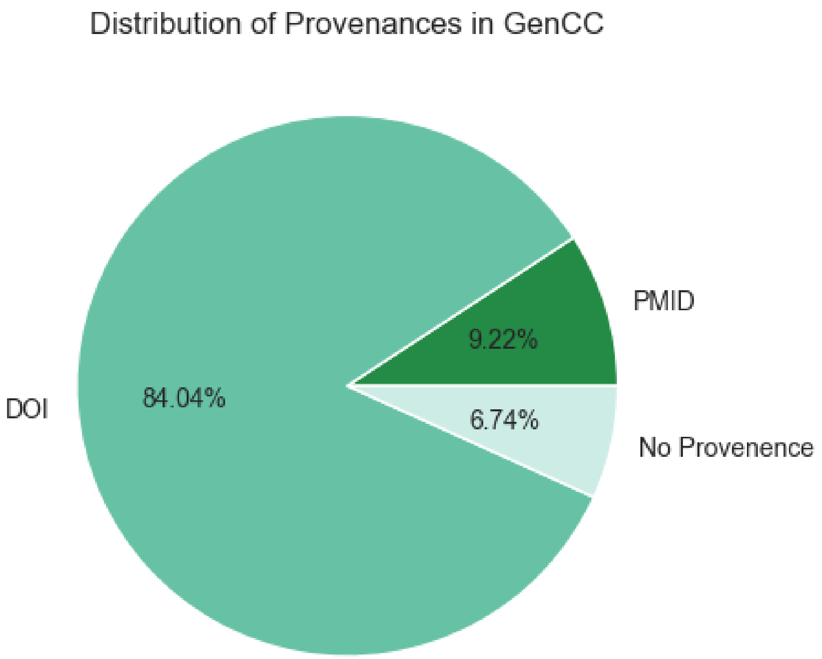

Figure 3

Part Two – Modes of inheritance:

MOI Curie	Mode of Inheritance Name
HP:0000006	Autosomal dominant inheritance
HP:0000007	Autosomal recessive inheritance
HP:0001417	X-linked inheritance
HP:0001419	X-linked recessive inheritance
HP:0001423	X-linked dominant inheritance
HP:0012275	Autosomal dominant inheritance with maternal imprinting
HP:0001442	Typified by somatic mosaicism
HP:0012274	Autosomal dominant inheritance with paternal imprinting
HP:0010984	Digenic inheritance
HP:0001427	Mitochondrial inheritance
HP:0032113	Semidominant inheritance
HP:0001450	Y-linked inheritance

Table 3 - unique moi curies in GenCC data and their names

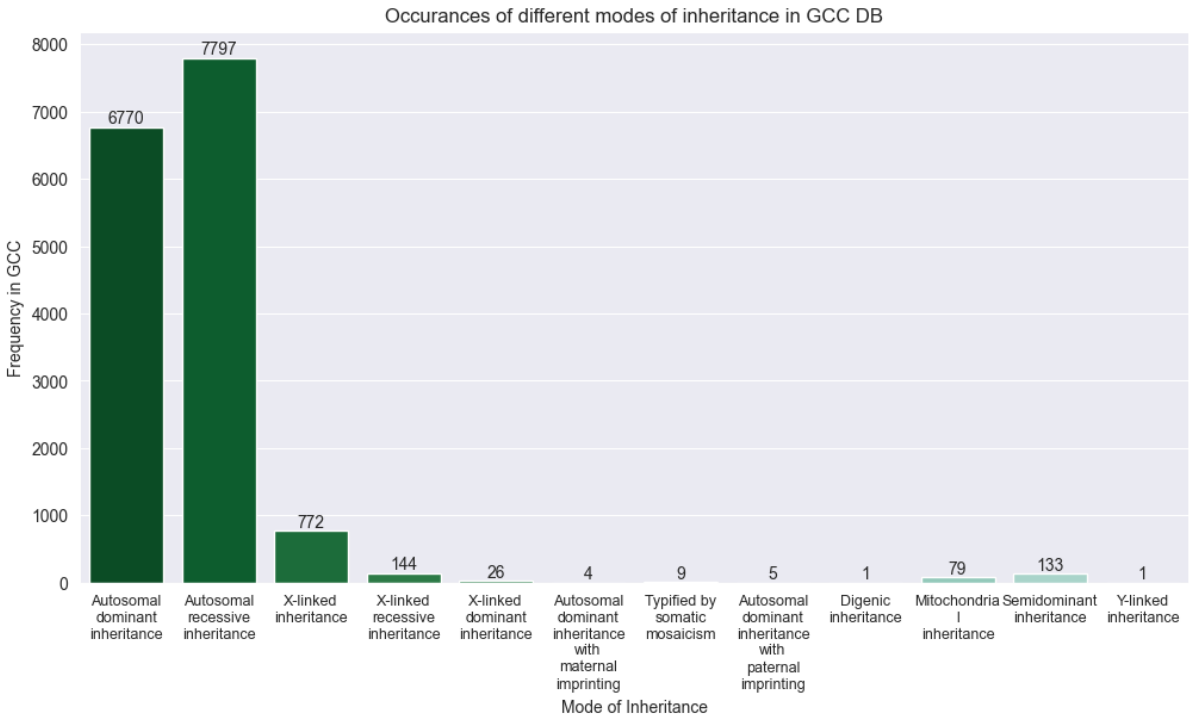

Figure 4

Part Three – Disease Groupings & their Associated Genes:

MONDO ID	MONDO Name
MONDO:0002320	Congenital nervous system disorder
MONDO:0002602	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 4 - first 10 MONDO terms below NSD node in MONDO ontology

MONDO ID	Disease Name	Gene Count
MONDO:0019502	Schizophrenia	35
MONDO:0019587	Syndromic intellectual disability	55
MONDO:0000508	Tourette syndrome	74
MONDO:0007661	Leigh syndrome	78
MONDO:0019588	Nonsyndromic genetic hearing loss	87
MONDO:0005090	Autosomal recessive non-syndromic intellectual disability	89

MONDO:0019497	Autosomal dominant nonsyndromic hearing loss	93
MONDO:0019200	Hearing loss, autosomal recessive	104
MONDO:0009723	Complex neurodevelopmental disorder	127
MONDO:0100038	Retinitis pigmentosa	159

Table 5 - top 10 NSD diseases by gene count

Gene ID	Number of NSD GenCC Entries
HGNC:9086	14
HGNC:801	14
HGNC:7606	14
HGNC:12403	14
HGNC:18060	15
HGNC:2213	15
HGNC:19139	15
HGNC:10585	15
HGNC:6990	16
HGNC:10591	17

Table 6 - top 10 genes by number of entries, from NSD related disease data

Part Four – Building a Simple GenCC Disease Network:

Disease Pair	Common Gene Count
MONDO:0019588, MONDO:0019497	54
MONDO:0019587, MONDO:0019497	29
MONDO:0009723, MONDO:0016815	28
MONDO:0019609, MONDO:0019234	13
MONDO:0100038, MONDO:0015802	11
MONDO:0019587, MONDO:0019588	11
MONDO:0000508, MONDO:0014699	10
MONDO:0000508, MONDO:0100038	10
MONDO:0019200, MONDO:0018998	9
MONDO:0019181, MONDO:0100148	8

Table 7 - top 10 disease pairs by common gene count

Community Number	Number of Diseases
1	12
2	7
3	6
4	6
5	5
6	5
7	5
8	4
9	4

10	3
11	3
12	3
13	3
14	2
15	2
16	2
17	2
18	2

Table 8 - top 10 communities by size

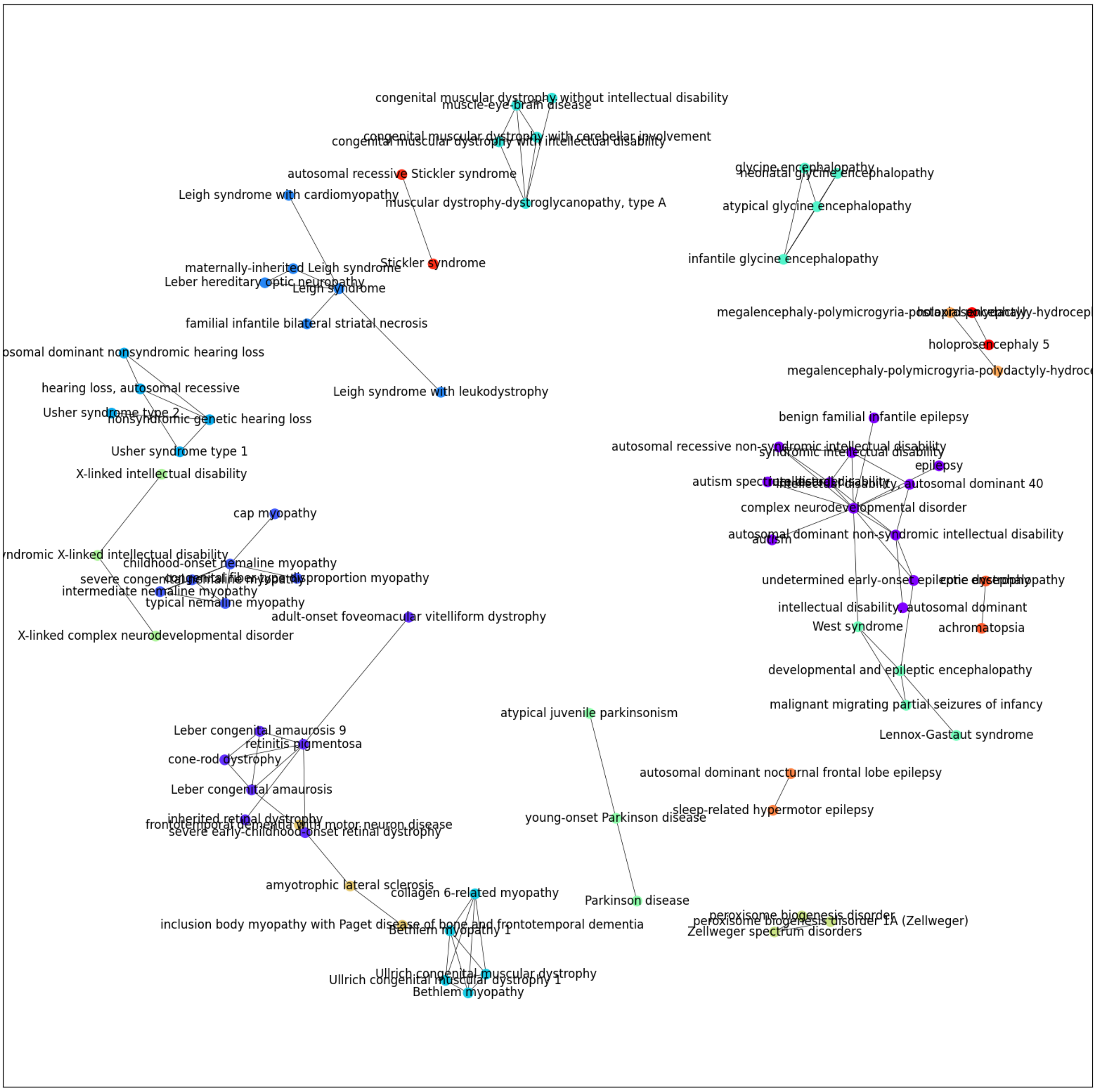
congenital muscular dystrophy without intellectual disability
muscle-eye-brain disease
congenital muscular dystrophy with cerebellar involvement
congenital muscular dystrophy with intellectual disability
autosomal recessive Stickler syndrome
Leigh syndrome with cardiomyopathy
muscular dystrophy-dystroglycanopathy,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

Figure 5 - a clustered network of nodes that belong to communities of size > 1

Discussion

Part One – Summary and Analysis of GenCC Data:

Table 1 and Figure 1 show the counts and distribution of the number of genes associated with each disease (note only the top ten are shown), from this, we can see that the disease with the largest number of genes is Complex Neurodevelopmental Disorder. Figure 1 shows that all these top diseases have associated gene numbers in the same orders of magnitude, but as we approach the diseases with larger numbers of associated genes they become more spread out and the difference between the next largest disease becomes more significant. Despite this, the numbers of related genes are relatively low, so it would not be impossible for a person to sift through them individually.

In Table 2 and Figure 2, we can see that most entries in the dataset are listed in higher confidence categories: Definitive, Strong, or Supportive, there is a good amount but significantly fewer entries in middle confidence categories: Moderate and Limited, and low numbers of entries with low confidence: Refuted evidence, Disputed evidence, and No evidence. This makes sense as the goal of GenCC is to harmonise gene-level resources, so the different groups committing to this dataset will probably tend to submit trustworthy entries. Also, because it is an open collection, this allows for other groups to view and validate entries, so lower confidence entries will probably be critically reviewed, or further researched and thus have their disease/gene association validated or invalidated and provide evidence to support changing its confidence level.

In Figure 3, we can see the distribution of sources for the GenCC file. We can see that only 9% of entries come from PubMed sources while the majority of entries are from non-PubMed articles and papers. Finally, there is a significant amount of data with no provenance. PubMed entries are probably the most reliable, as they have high visibility and PubMed articles are generally peer reviewed, while entries with a DOI source while still coming from a paper the process of review is less known as its source is less specific, so could include very low visibility papers that might not have been reviewed. No provenance data is the most unreliable of the 3 classifications, as it means that the entry has no known associated paper, this makes it harder to critique and validate these entries, so a very sceptical user of the GenCC file might choose to avoid using these. Over 93% of entries have traceable sources, meaning that it is possible for these to have been externally validated, making them more trustworthy.

When using the GenCC database, depending on the choices of the user, they could choose to use this data to filter entries based on trustworthiness, so that they do not make assumptions based on possibly incorrect relations. Although we can see that any of these entries that could be deemed less trustworthy also make up a minority of the dataset, their effect will most likely be insignificant. It is also easy for a user to focus on certain diseases as the number of related genes is relatively small, meaning it will be easier to review just the relevant genes.

Part Two – Modes of Inheritance:

There are 18504 entries in the GenCC dataframe, and 15741 (85.1%) entries have valid MOIs. This means just under 15% of entries in the dataset have unknown inheritances. From Table 3 and Figure 4, we can see that by far the greatest modes of inheritance for the GenCC dataset are autosomal dominant and recessive inheritance. There is also a much smaller but still significant amount of X-linked inheritance. Also, for X-linked inheritance, where we can see 3 different types of x-linked inheritance, it is important to note that the use of dominant and recessive has been discouraged [5], and that these columns should perhaps be merged into one value. I haven't done this, as our study is more about exploring traits of the dataset rather than conducting in-depth research, so this is useful as it shows an important trait about the public nature of the dataset and that perhaps even though they shouldn't, that people can still input X-linked inheritance data as dominant or recessive.

The reason that autosomal inheritance (both dominant and recessive) is the most common mode of inheritance in GenCC, could be because of the number of autosomes. Autosomal inheritance is where a disease is inherited through the passed-on autosomes of a parent [8], where the autosomes are one of the numbers chromosomes, making up the majority of genetic material [9]. Therefore, this means there are many more entries as this refers to most genes for a person meaning more entries in our disease-gene relation database.

Part Three – Disease Groupings and their Associated Genes:

Nervous System Disorder Mondo ID: MONDO:0005071. There are 5587 Mondo terms below this node in the Mondo ontology file and there are 2208 unique gene curies in the NSD GenCC Dataframe.

From Table 5 we can see that there are many shared entries to Table 1, these are schizophrenia, syndromic intellectual disability, Tourette Syndrome, Leigh Syndrome, Nonsyndromic genetic hearing loss, complex neurodevelopmental disorder, retinitis pigmentosa, also called by different names but still included in both tables is non syndromic genetic hearing loss & autosomal recessive hearing loss. So clearly in GenCC Nervous System disorders tend to have the most associated genes out of all recorded diseases, as in Table 1 we don't limit to just NSDs but compare to all diseases.

We can also see comparing Table 1 and 5 that some diseases have larger number of associated genes in this part. This shows us the usefulness of using an ontology database, as in part 1 we group by the disease title, but diseases have synonyms so this means that we may not accurately group all diseases and we have no check for equivalence. After using the MONDO dataset though we can work around the use of synonyms, thus resulting in an increase in gene count for some diseases, such as Retinitis Pigmentosa.

In Table 6, we can see that there is not much difference between the number of entries associated with each gene, and considering that there are over 2000 different genes, the counts are relatively low and similar. This table tells us the genes which could play a role in many different diseases, this could be useful for focusing research on critical genes and what their specific role is, but as there are no particularly prominent genes with high counts this unfortunately means that this is probably not possible for now.

Part Four – Building a Simple GenCC Disease Network:

There are 93 edges in the graph shown in Figure 5. This shows us that there are not many diseases with large numbers of common genes, particularly as we are considering over 2000 different diseases. As most diseases do not have many genes in common with any other diseases, this helps us visualise the wide range of diseases in GenCC, and how they generally may not often share much in common with other diseases. Also, looking at Figure 6 we can see that clusters rarely connect to other clusters, except in one instance where we have the largest community in our network (community 1), which does have some connections to a smaller cluster (see West Syndrome and Developmental and Epileptic Encephalopathy). This makes sense as if we know that diseases usually don't share many genes it would be suspicious if these clustered diseases also connect to other clusters. We can see just from the names of the diseases that the clustered diseases are usually closely related, such the Parkinson's cluster of 3 nodes.

Extensions

Part 1 – Pattern Analysis of GenCC Entry Submission Dates:

In this extension task I chose to try to view patterns relating to submission dates in the GenCC dataset, and trying to give some possible explanations for why these trends might exist. To do this I took all entries in the dataset and converted their submission date into a workable format, I sorted them by date, removing outliers from the start and end of the dataset – for example, one entry was submitted in the year 3016. This is slightly concerning as it is the goal of GenCC if the goal of GenCC is to be trusted, you would expect more input validation on obviously erroneous data. I then plotted the submission years in the line plots shown below. The GenCC was officially formed in February 2018 [4], so data inputted before this time probably did not have the same verification that later entries might have.

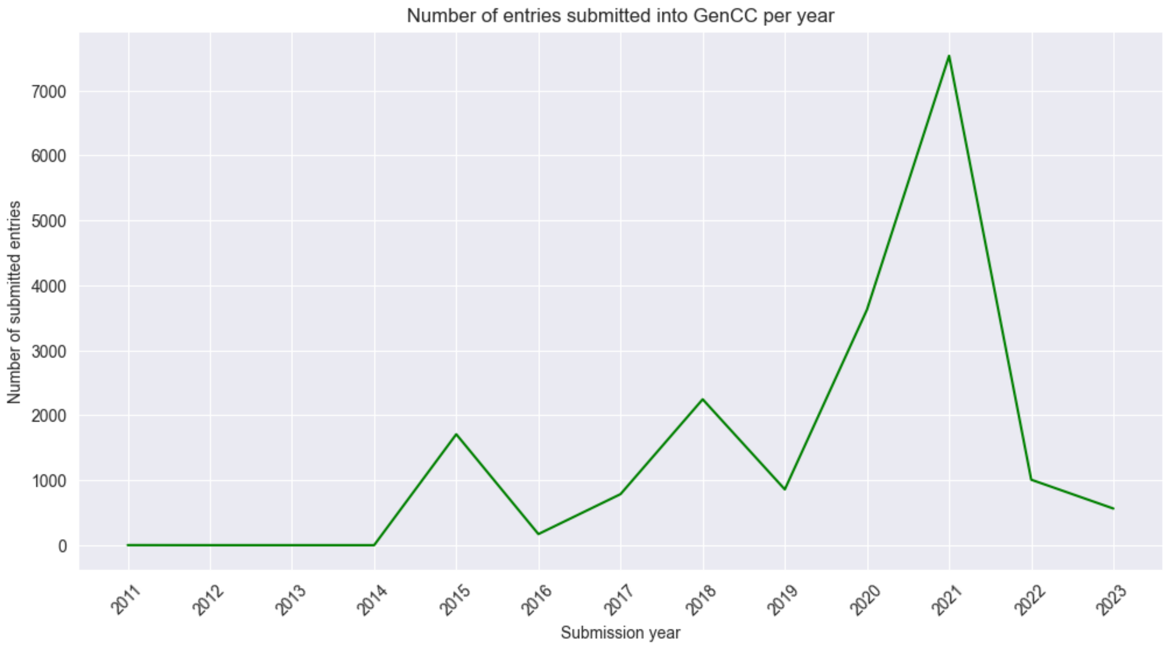

Figure 6

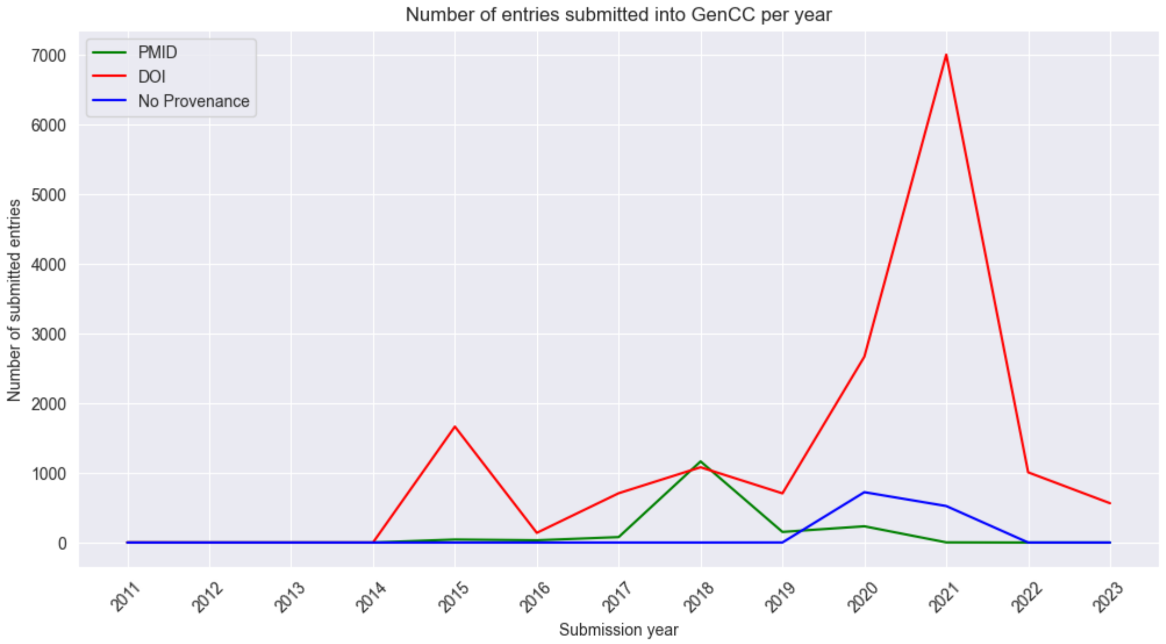

Figure 7

Firstly, as for patterns shown in Figure 6, the most obvious trend is that we can see a big spike in submissions in the years 2020 and 2021. We can also see that after this spike we appear to return to a similar level of submission as pre-spike. Before plotting this data, I assumed that we would see a trend of the number of submissions steadily increasing each year, which could be true but is hard to verify without more years of data. We can also so smaller peaks in submissions in the years 2015 and 2018.

As for patterns in Figure 7, we can see that the DOI source entries have contributed the most except for in the year 2018 when there was a spike in PMID entries. We can also see that the large spike in 2020-21 is mostly made up of DOI entries, but we can also see a significant spike in entries with no listed provenance during this time. Another interesting trend is that PMID entries seem to be decreasing since 2018 (the year GenCC went official).

As for reasons for the large spike in 2020-21 this most obvious reason for this could be the Covid-19 pandemic, for most fields you would expect to see slumps around these years caused by major disruptions from counterefforts such as lockdowns, but in GenCC we can see an increase during these years. But most obviously, diseases and their related genes obviously became a much higher global priority due to the prominence of a global pandemic, so this may have fuelled research efforts and thus usage and submissions to the GenCC database. One source I found supporting this was the J. Park et al series [10] that describes the new norm of medical research publishing as “quantity over quality”. This could also be why there is a large increase in entries without a provenance, as the emphasise on sharing work quickly, surpasses sharing it through published articles and papers, that would take more time.

As for the trend of PMID sourced articles decreasing since the year 2018, we first know that GenCC became official in early 2018, so this might have spurred on efforts to submit existing data from PMID into GenCC, but as the years progressed the focus became more on submitting current data as it is found, so the numbers of PMID sourced entries decreases.

The significant drop in submissions in 2022 was expected as the pandemic was overcome, but this it is still unexpected as it returns to levels similar to that of pre-pandemic years. This shows that the uptick in submissions during the years 2020-21 was temporary and there was no large lasting effect of this spike in terms of number of submissions. This is unexpected as you would expect GenCC to become more well known and well used after this big spike, but without more data we do not really see this. This could be just because there was so much pressure in the years 2020-21 that researchers have moved on to other areas for now and we might see a decent increase in the future. Or perhaps it shows some discontent with the GenCC database, although this would require far more research and evidence to confirm.

Part 3 – Distribution of Groups of Diseases in MONDO ontology:

In the MONDO ontology file each disease name should be unique but the entries do contain references what type of disorder it is, that we can use to visualise the distribution of groups of disease within MONDO. For example, Infantile Liver Failure is a hereditary parenchymatous liver disease, which is a liver disorder (note this is not a great example as the MONDO entry of hereditary parenchymatous liver disease is being made obsolete). MONDO is also hierarchical so includes many different levels of disease, since I want groupings, I need to choose a suitable level in the database of groups of disease, too high a level and the information is too broad, too low and information is too specific. If we organise the MONDO database into a tree, with the root node as the 0th level, I chose the 2nd level of nodes. I then counted the sizes of their subclass sets and sorted them using this information. Below you can see the top 10 and a corresponding bar chart in Table 9 and Figure 8, and you can find code used for this task in Figure 11.

Table 9

Disease Group	Number of MONDO sub diseases
Hereditary disease	10622
Nervous system disorder	5587
Cancer or benign tumor	4573
Disorder of development or morphogenesis	4209
Musculoskeletal system disorder	3220
Syndromic disease	2987
Metabolic disease	2313
Disorder of the visual system	1931
Disorder of the orbital region	1902
Integumentary system disorder	1731

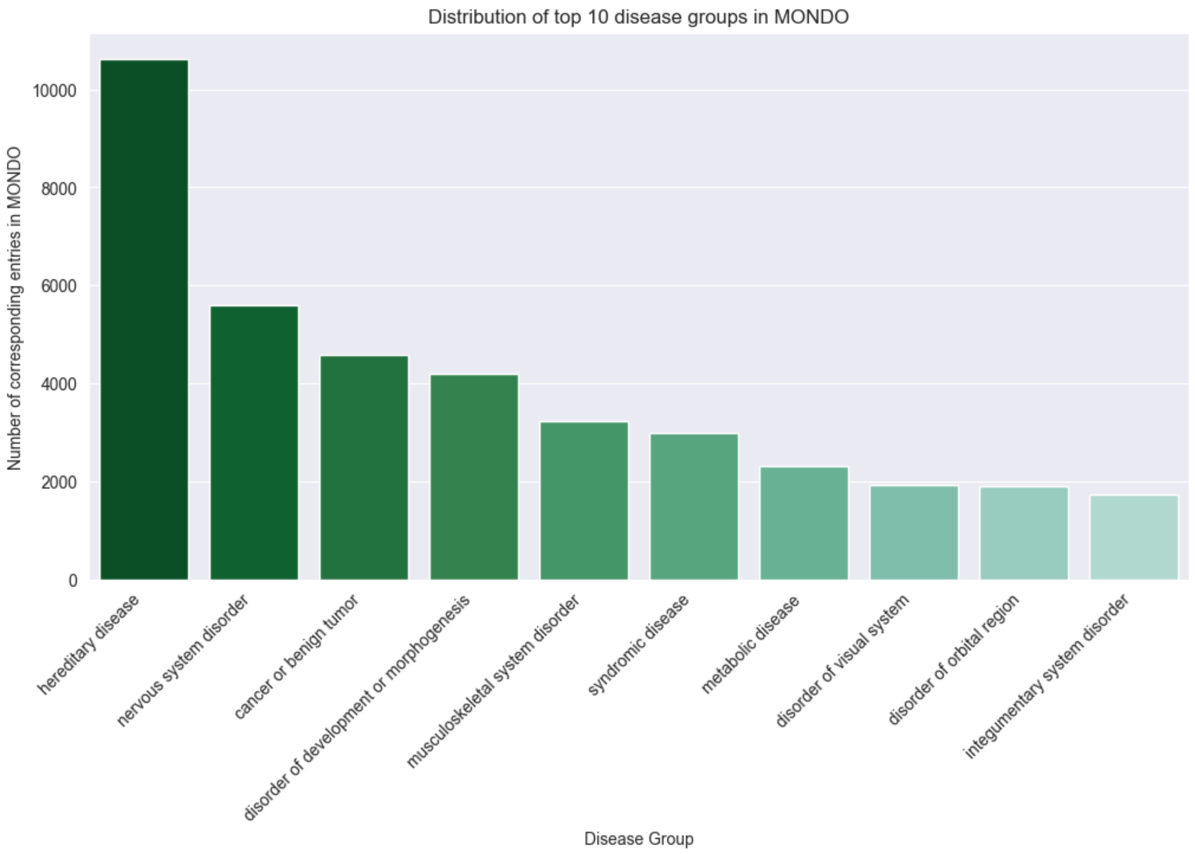

Figure 8

From this we can see that the most represented disease group is by far hereditary disease, and we can see a general trend in the number of correlated diseases for each disease group, that the difference between the groups is shortening as we continue down the top 10.

References

1. GenCC FAQs: https://thegencc.org/faq.html, last accessed 17/11/23
2. GenCC Submission Directions: https://thegencc.org/submission-directions, last accessed 17/11/23
3. Mondo Disease Ontology homepage: https://mondo.monarchinitiative.org/, last accessed 17/11/23
4. DiStefano M.T. et al. (2022) The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. Genetics in Medicine Aug 2004, Volume 24, Issue 8, p1732-1742. doi: https://doi.org/10.1016/j.gim.2022.04.017
5. Dobyns W.B., Filauro A., Tomson B.N., Chan A.S., Ho A.W., Ting N.T., Oosterwijk J.C., Ober C. (2004) Inheritance of most X-linked traits is not dominant or recessive, just X-linked. Am J Med Genet A. 2004 Aug 30;129A(2):136-43. doi: https://doi.org/10.1002/ajmg.a.30123. PMID: 15316978.
6. BioPortal homepage: https://bioportal.bioontology.org/, last accessed 18/11/23
7. Noy N.F. et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37(Web Server issue), W170-173. doi: https://doi.org/10.1093/nar/gkp440. PMID: 19483092
8. Dorschner M.O. (2013) Autosomal Inheritance. Brenner's Encyclopedia of Genetics (second edition). doi: https://doi.org/10.1016/B978-0-12-374984-0.00121-2.
9. Autosome entry in the NIH Glossary of genomic and genetic terms: https://www.genome.gov/genetics-glossary/Autosome, last accessed 20/11/23
10. Park J.J.H. et al. (2021) How COVID-19 has fundamentally changed clinical research in global health. The Lancet Global Health, Volume 9, Issue 5, e711-e720. doi: https://doi.org/10.1016/S2214-109X(20)30542-8
11. Genomics England PanelApp homepage: https://panelapp.genomicsengland.co.uk/, last accessed 18/11/23

Appendix

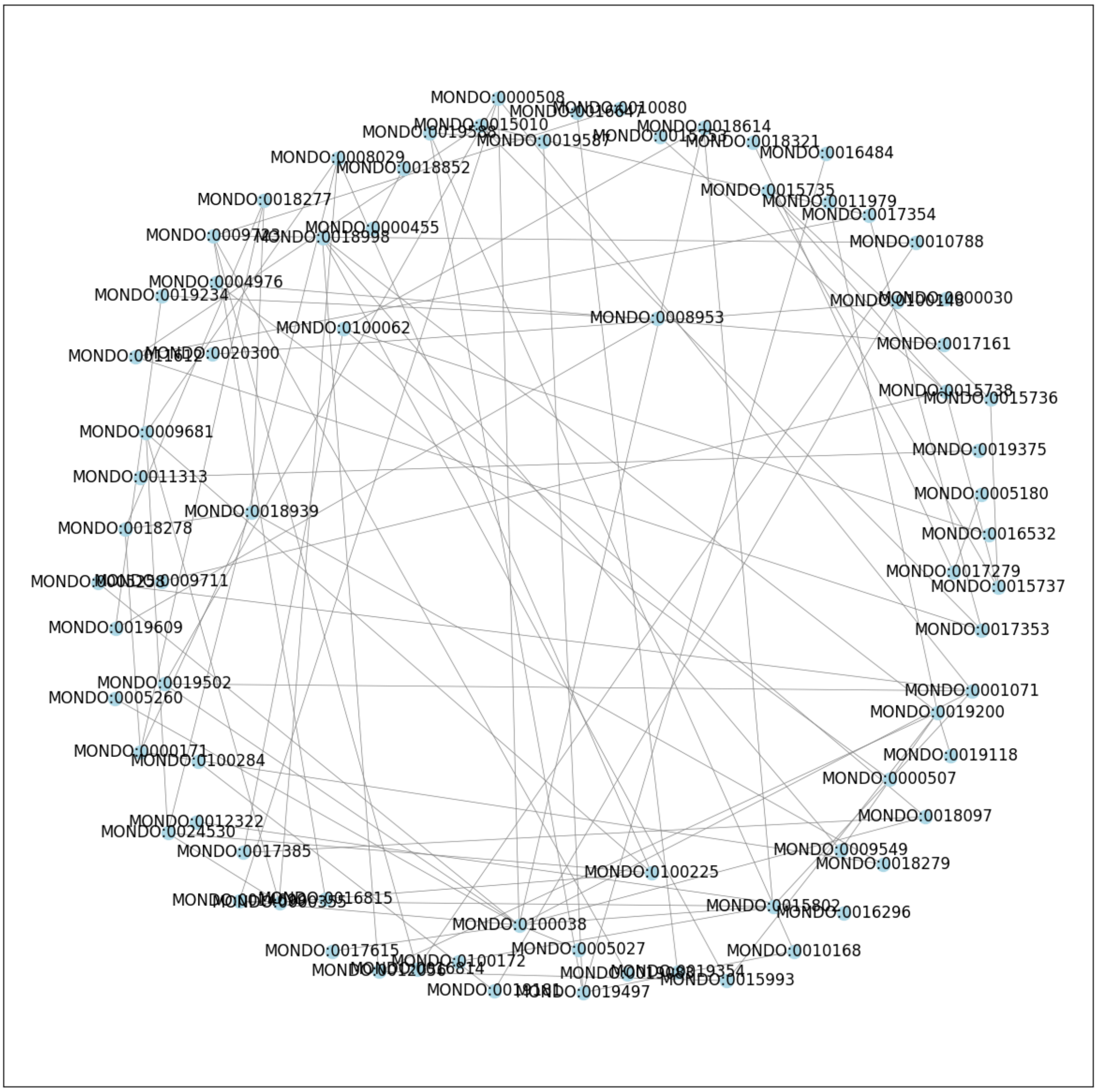

Figure 9

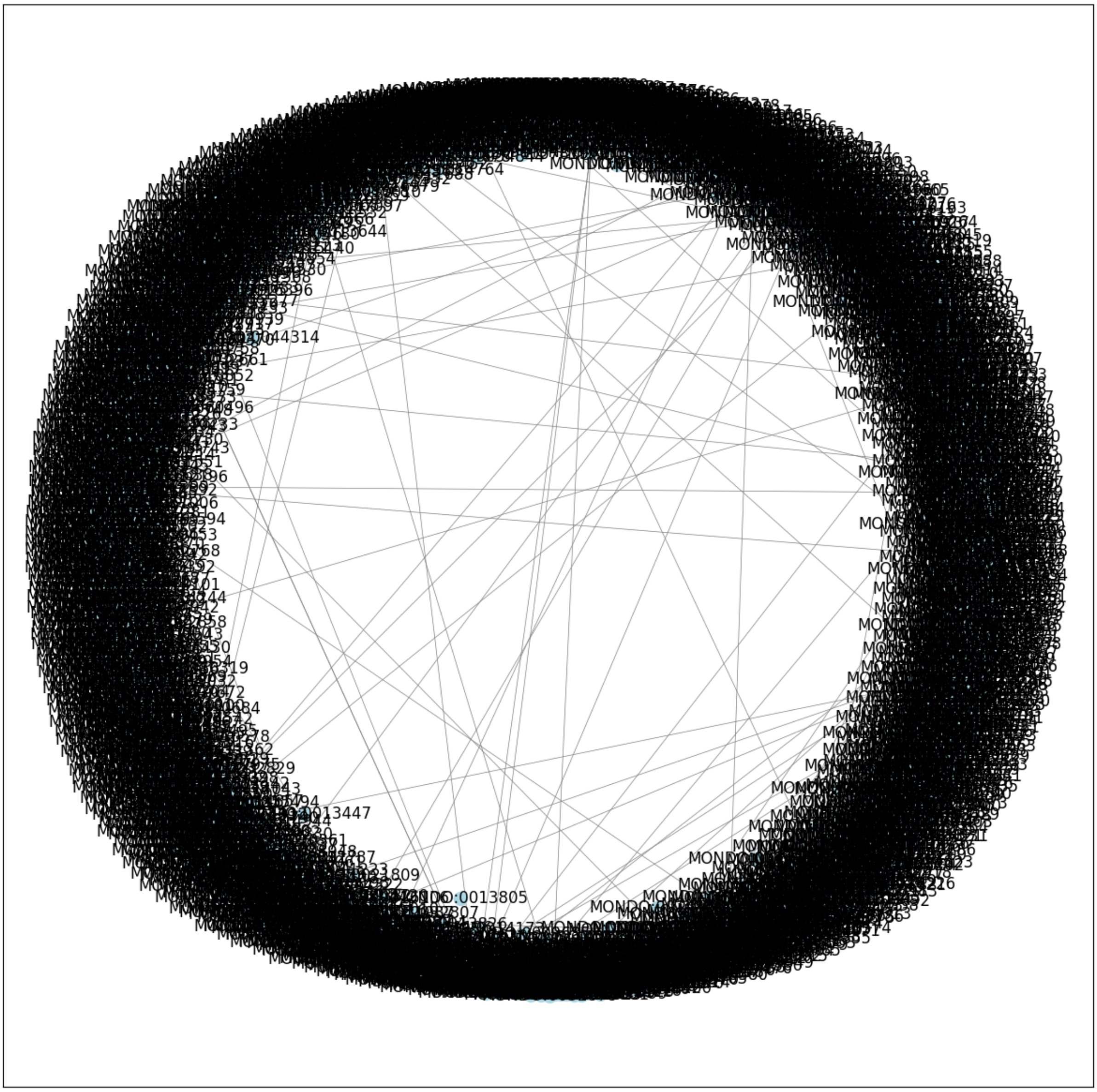

Figure 10

```python
from itertools import chain

mondo=pronto.Ontology("mondo.obo")

# working from root down...
root = mondo.get_term("MONDO:0000001") # disease
subclasses_l1 = list(root.subclasses(distance=1, with_self=False))
subclasses_l2 = list(set(chain(*[list(subclasses_l1[i].subclasses(distance=1, with_self=False)) for i in range(len(subclasses_l1))])))

# now getting the number of subclasses for each of these terms so we can see the most popular
subclass_size = []
for sc in subclasses_l2:
```

```
8.     subclass_size.append((sc, len(list(sc.subclasses())) - 1))

9. sorted_disease_groups = list(sorted(subclass_size, key=lambda x: x[1], reverse=True))
```

Figure 11 – line numbers have been added in post so that code that spans more than one line in the document can be identified more easily