



Journal Article:
Dissertation

MSc. Health DataScience 2024-2025

Table of Contents

Abstract.....	3
Introduction	3
Methods	6
Identify carriers of Pathogenic Variants in ACMG Genes	6
Study Design	6
Descriptive Statistics and Data Pre-Processing	7
Comparing burden of disease in Cases vs controls	9
Disease Transitions	10
Disease Trajectories	11
Software	12
Results	13
Disease Transition Pairs	17
Disease Trajectories	21
Discussion	23
Strengths, Limitations, and Challenges	25
Strengths	25
Limitations	25
Challenges	25
Future work	25
Conclusion	26
Ethical Issues	26
References	27
Appendix	33

Exploring disease trajectories in participants with pathogenic variants in medically actionable genes: A Genomics England Study

Abstract

Background: Disease trajectories are increasingly being used to map disease progression and disease transitions across various fields like mental health and chronic illnesses. This is enabled by the widespread availability of linked patient records using Electronic Health Records (EHRs). Finding hidden patterns in disease trajectories can empower healthcare systems by promoting research and allowing early disease intervention. The aim of this study is to explore disease trajectories in carriers of pathogenic variants in American College of Medical Genetics and Genomics (ACMG) genes and compare it with controls. The analysis for this study was done entirely using the Genomics England (GE) dataset and research environment (RE).

Methods: Case and control cohorts were built using several inclusion and exclusion criteria. For the cases, burden of disease, and simple and complex disease trajectories were mapped and compared to controls. Complex trajectories were built on network graphs architecture.

Results: For all carriers of pathogenic variants in ACMG genes (cases), burden of disease for cancer, ear, and perinatal diseases was found to be higher than controls. Contrastingly, benign neoplasm and endocrine diseases showed lower odds in cases than controls. Additionally, simple disease transitions and complex disease trajectories were more common in cases and less common in controls. For example, in simple disease transitions, out of top 10 disease transition pairs, on average 6 pairs were more common in gene carriers than controls. The only exceptions to this were the carriers of gene *SCN5A* which had 6 out of 10 transitions more common in controls. For more complex disease trajectories across all the studied genes, the trajectories were more frequent in gene carriers than controls.

Discussion: This study illustrates a marked difference in disease burden, simpler trajectories, and more complex disease trajectories between cases and controls. A study of this kind has not been done before and hence it can provide unique insights into progression of disease trajectories in carriers of pathogenic mutations.

Keywords: acmg genes, cancer, rare disease, genomics england, disease trajectories, disease transition

Introduction

Study of disease trajectories is an exciting and evolving field of research enabled by the widespread use of EHRs. EHRs longitudinally record complete patient information and have been increasingly adopted for clinical and healthcare research over the last two decades^{1,2}. Longitudinal patient data enables tracking a patient's disease history and comorbidities over time. Comorbidities are simply one or more diseases or health conditions that exist simultaneously in an individual³. However, much is still not understood about how the order of occurrence of diseases impacts general health outcomes. A valid

method to map comorbidities is by using disease trajectories which establish a temporal order for the occurrence of the diseases one after the other. The temporality is crucial to ascertain the order of occurrence in the disease pairs which helps establish the burden of comorbidity⁴. The order of the diseases matters, and not just the presence of the two diseases existing simultaneously. This difference is exemplified in a study which found that patients diagnosed with sleep apnea before Type 2 Diabetes (T2D) had a lesser burden of disease than patients diagnosed with T2D before sleep apnea. The burden of disease is measured by the number of comorbidities associated with the patient. Patients who were diagnosed with T2D first, had 3 more comorbidities recorded compared to those patients who were diagnosed with sleep apnea first⁵.

Most diseases, and by extension, comorbidities, have some degree of genetic factors involved in how the disease develops, and may possess one or more genetic factor associations. These genes are called disease-associated genes. Disease-associated genes can help understand comorbidities and their patterns better⁶. Furthermore, some of these disease-associated genes are “Actionable” in nature. Actionable genes, as defined by the American College of Medical Genetics and Genomics (ACMG), are pathogenic mutations associated with a disease for which complete, or partial, medical intervention is possible. These genes will be referred as ACMG genes hereafter. To ensure, that these genes are strongly associated with a disease, ACMG has put forth two criteria to assess the association. First, the genetic variant has shown prior evidence of causing a disease. Second, if the genetic variant has not shown prior evidence of being causative of a disease, the gene is still likely expected to cause a disease^{7,8}. These findings are applicable only to the 81 unique ACMG genes, as per the 2023, ACMG Secondary Findings Version 3.2. Additionally, these ACMG genes show high penetrance (people carrying the mutation also having the disease)^{7,9}. When genetic testing is performed by clinicians, any pathogenic variants found in the tests are recommended by the ACMG to be reported back to the patients even though the genetic testing might not have been primarily done to detect these mutations. These are termed as incidental findings⁷. However, it is important to note that since these incidental findings are recommendations from an organisation, these will not always be fully followed. Genomics England (GE), the dataset and environment used for this study, for instance, only reports on 11 ACMG genes out of the 81 ACMG genes. Additionally, GE terms these findings, if found, as “additional findings” for which participants have agreed to be informed about¹⁰. A list of these 11 genes from GE additional findings and the complete list of 81 ACMG genes is attached in Appendix 2.

Despite the considerable work done to understand the impact of comorbidities, there is significantly less research done to understand disease trajectories in comorbidities, especially in carriers of these ACMG genes. When searching on PubMed for relevant studies on disease trajectories in pathogenic variants, only 132 studies were found, many of which were related to COVID-19. The MeSH terms used as search strategy can be checked in Appendix 1. Out of these 132 studies, only 15 studies explored actual disease trajectories without focusing on pathogenic variants or ACMG genes. Similar search was done using the search term “disease progressions disease trajectory pathogenic variants”. It yielded only 11 papers which were not more relevant than the previous search. Doing a

manual search on Google Scholar and other journal aggregators, yielded some more papers. However, most papers that talked about disease progressions, focused only on one gene or disease, or explored pathogenic mutations but not specifically in ACMG genes¹¹. Furthermore, some studies employed survival analysis as a major component of exploring disease progressions¹². It is imperative to note that these studies had vastly different objectives than this research project, and they utilised different ways of employing disease trajectories. Most studies used disease trajectories to map disease progressions which can include natural history of a patient or disease or a particular disease domain specific trajectory. Disease transitions are used for mapping discrete diseases over the lifespan of a participant, as done in this project. Therefore, disease transitions have a much broader range of exploration compared to disease progressions. The focus of this study is to explore disease transitions in carriers of pathogenic variants and compare it with controls.

From the above literature search, it can be inferred that there is a considerable lack of published studies exploring disease trajectories in pathogenic ACMG genes.

The aim of this research project is to explore the disease trajectories in rare disease and cancer patients with pathogenic variants in ACMG genes (cases or gene carriers) and compare it with a control group. To achieve this, a study has been conducted within the GE RE with the following objectives:

- Identify carriers with mutations in the ACMG genes to form the case population and apply various exclusion criteria to establish a control population that does not carry the mutations.
- Compare the burden of phenotypic diseases based on patient's assigned disease categories by assessing these disease frequencies within case and control populations.
- Map and compare simple disease trajectories between gene carriers and controls and identify which disease transitions, from one disease to the next, are more common in either case or control group.
- Explore and compare more complex and longer disease trajectories and identify which transitions are more common in gene carriers than controls.

Rationale:

As evidenced above, there is a lack of published studies exploring disease trajectories in individuals carrying pathogenic variants and ACMG genes.

Disease-associated genes and pathogenic variants are strongly associated with comorbidities. And, a study of this kind can help discover novel trajectories, assess the burden of disease based on pathogenic mutations, and ultimately broaden the knowledgebase of disease patterns affecting patient health outcomes. By establishing novel trajectories, this study can promote proactive care in terms of early screening and intervention. The study has the potential to positively impact the lives of millions of patients suffering from cancers and rare diseases. Hence, the need for a comprehensive analysis of such nature is pertinent.

Hypothesis:

Patients with pathogenic variants in ACMG genes will share specific health outcomes which will be reflected in a difference in burden of specific disease categories, and some shorter disease transition and longer disease trajectories being more frequent in cases than controls.

Methods

Genomics England (GE) 100,000 Genome Project has participant level genomic and EHR data of over 90,000 individuals. This report uses the Version 18 of the dataset, released in 2023, with the data cutoff year being 2023. GE combines clinical and secondary data from different sources like mortality data from Office of National Statistics (ONS) and medical history using Hospital Episode Statistics (HES) episodes from National Health Services (NHS) England. GE has participants primarily with diagnoses for rare diseases and cancer recruited by the NHS through its Genomic Medicine Service (GMS) program¹³. NHS GMS was set up to provide Whole Genome Sequencing (WGS) as a routine part of England's healthcare system by promoting research and enabling early disease detection¹⁴. GE is based in the UK, and all participant data is recorded or collated in the UK as well.

Identify carriers of Pathogenic Variants in ACMG Genes

Missense variants are single genetic alterations in the DNA sequence of a gene coding for a protein. The resulting protein has a different amino-acid structure than usual. Hence, protein functioning is hampered¹⁵. Pathogenicity of missense variants in the ACMG genes was assessed using AlphaMissense scores. AlphaMissense (AM) is the best-in-class model which uses the AlphaFold Machine Learning (ML) model to predict the pathogenicity of missense variants. Pathogenicity scores were calculated using an Ensemble ML model with a base estimator of Support Vector Machine ML model. The AlphaMissense scores were just probabilities for each missense variant out of 1¹⁶. ClinVar was used for the training of missense mutations and their validation as pathogenic. ClinVar is a public dataset which aggregates genetic variants and associated diseases submitted by researchers from all over the world¹⁷. Pathogenicity scores range on a scale of 0 to 1, where likely benign (<0.34), ambiguous ($0.34 < x < 0.56$), and likely pathogenic (>0.56) are the standard benchmarks scores used¹⁵. However, for this report, only pathogenicity scores over 0.8 were used to eliminate any ambiguous variants and adhere to the list of ACMG genes. The Alpha Missense dataset used for filtering of the gene carriers was generated in the High-Performance Cluster (HPC) of the GE RE. Based on the mutations in all the genes, a list of participants with AlphaMissense pathogenicity scores greater than 0.80 were selected along with their mutated genes from the public dataset 'AlphaMissense_hg38.tsv.gz'. Amongst these, only the 81 ACMG genes were kept, rest were filtered out. This forms the basis of case or carrier dataset.

Study Design

The cohort formation was done using strict inclusion and exclusion criteria. Firstly, only those participants were kept who had both WGS and demographic information available.

For the cases, only the participants who had a pathogenic mutation in one of the ACMG genes were included. For controls, unaffected relatives of participants with a rare disease were included.

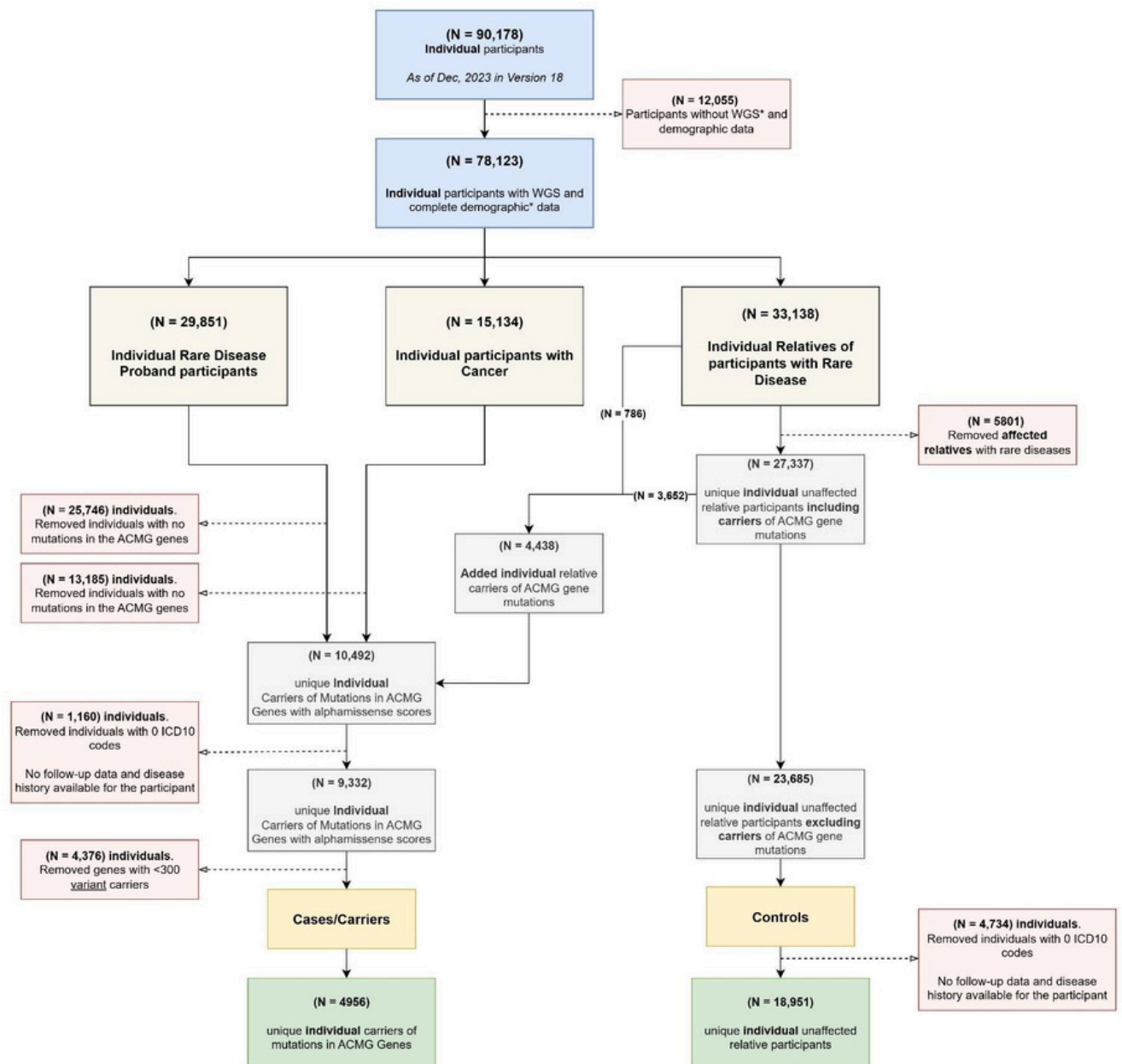


Figure 1: Flowchart explaining the Study design for the cohort formation. WGS means Whole Genome Sequence. Complete demographic information includes year of birth, phenotypic sex, affection status, ethnicity, and ancestry.

Furthermore, amongst the carriers, only genes that had more than 300 carriers of pathogenic variants were included in the study.

Descriptive Statistics and Data Pre-Processing

A table to show the demographic and other important variables associated with each participant was made using the tableone python library¹⁸.

The age for each participant was calculated based on the year of birth and the cutoff year, 2023. If a person was dead, their age was calculated based on the year of their death. Furthermore, age brackets of 10 years were used to categorise age groups.

Index of Multiple Deprivation quantiles with their IMD score range for the brackets. IMD score categorises participants from most deprived to least deprived in the UK based on their socio-economic status according to the pin code of the area they live in¹⁹. One participant has three IMD scores: latest, lowest, and highest. The IMD quantiles of the participants were calculated based on their latest IMD score. The participants were equally divided, based on their frequency distributions, into 5 quantiles for the entire cohort. The range of scores that formed the brackets are mentioned in the participant demographics table (Table 4).

ICD-10 (International Classification of Disease Version 10) codes were derived for each participant along with the dates of diagnosis. This formed the ICD-10 table. A disease phenotype table was generated using a list of 308 disease phenotypes across 16 disease categories identified by Kuan et al, 2019. In their paper, using phenotyping algorithms from CALIBER, the researchers used primary and secondary care data of 4 million NHS patients to identify only those diseases that had a prevalence of at least 0.01% and logged at least 10,000 HES consulting episodes (monitoring hospital activity like admissions, diagnoses, and more)²⁰. The resulting 308 diseases have been mapped to each participant and utilised in this report forming the Pheno308 table.

In the demographics table, the median number of ICD-10 codes assigned variable has two components, one calculated using the ICD-10 table, and the other calculated using the Pheno308 table.

Due to skewed distribution of the number of ICD-10 codes assigned to each participant, for the cohort, the median was used instead of mean.

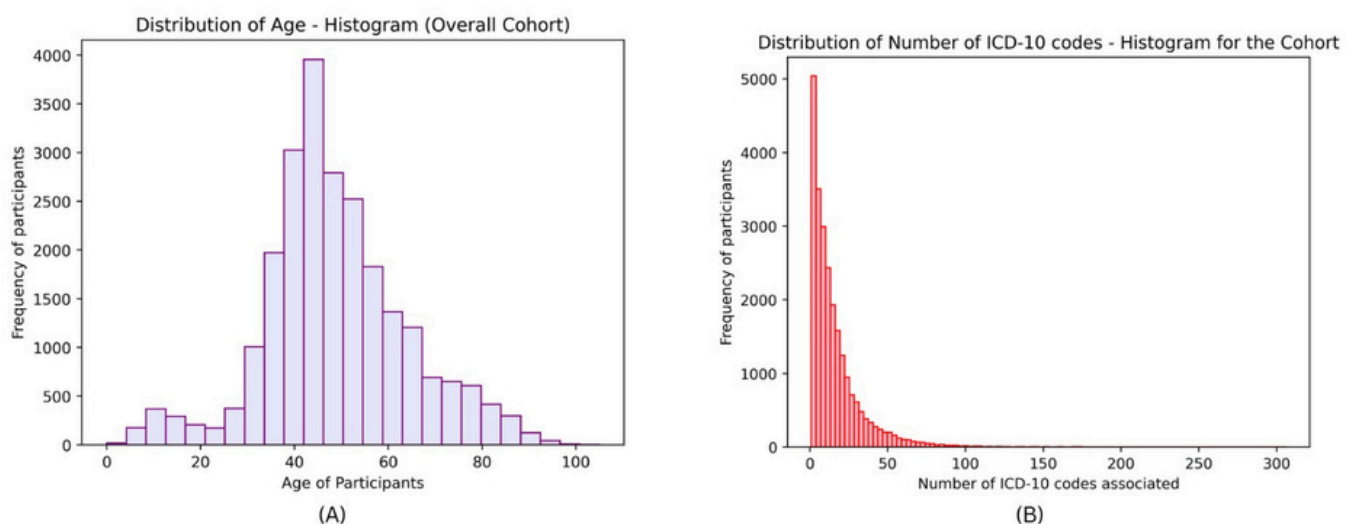


Figure 2: Histograms to show the distribution of Age and Number of ICD-10 codes per participant for overall Cohort. (A) The Age is normally distributed (left plot). (B) The Number of ICD-10 codes follows a left skewed distribution (right plot).

Comparing burden of disease in Cases vs controls

The participants across the cohort, had their associated disease and their corresponding categories mapped using the Pheno308 table. Complete list of these 308 diseases and disease categories can be found in the supplementary materials with the title, 'phenotype_categories_mapping.csv'. The broader categories for these diseases were also listed for each participant. There was a total of 16 disease categories mentioned in the below plots and supplementary phenotype table.

The table contains the following columns.

Index	participant_id	gene	disease	category	date of CD-10 code assignment
0	Participant1	APC	Asthma	Respiratory	2012-10-19
1	Participant1	APC	Sickle-cell trait	Haematological or immunological	2013-01-10
2	Participant1	APC	Bacterial Diseases (excluding TB)	Infections	2013-05-09
3	Participant1	APC	Infections of Other or unspecified organs	Infections	2014-09-21
4	Participant2	Control	Ear and Upper Respiratory Tract Infection	Neurological	2009-08-15
5	Participant2	Control	Epilepsy	Digestive	2009-12-16
6	Participant2	Control	Irritable bowel syndrome	Psychiatric	2011-04-05
7	Participant2	Control	Anxiety disorders		2012-01-06

Table 1: Artificial dataset with same columns as original, used to create the OR plots and disease trajectories. The *date* and *participant id* columns are changed to conform with the patient Personal Identifiable Information (PII) guidelines of GE.

The dataset had a sample size of 114,477 rows spread across 5 variables mentioned in Table 1. The dataset was validated for correct data types and cleaned by removing missing values like dates and diseases for any participant.

Burden of disease is described here as the percentage of a specific disease being present in each gene group or control out of total diseases present.

To assess the burden of broader disease categories for carriers and controls, a side-by-side bar plot was generated. The plot shows the frequency distributions of diseases by categories. The stratification was done by genes hence there are 10 subplots, one for each gene. To statistically explore these findings, Fisher's Exact test (FET) was used. A contingency table was constructed for all valid combinations of gene carriers, controls, disease categories, and no disease categories.

	Has the disease category	Does not have the disease category
carriers(per gene)	W	X
controls	Y	Z

Table 2: Odds Ratio Contingency Table

Hence, OR is calculated as follows,

$$OR=(W/X)/(Y/Z)$$

$$OR=(W*Z)/(Y*X)$$

Using the Odds Ratios (OR) values obtained using FET for each gene and controls, a line plot was generated. This line plot mapped the 16 disease categories based on their log OR values. Even though, the disease categories are unrelated, a line plot can help better visualise the overall trends. Log OR = 0 formed the baseline indicating no difference between gene carriers and controls. Log OR values closer to 0.5 or greater than 0.5 indicate carriers are at a higher risk for that disease category than controls and log OR values ≤ -0.5 indicate that the gene carriers are protected or have lower odds for that disease category.

Another similar method of calculating OR is using Chi-Square test. However, due to smaller sample size of some of the gene carriers, and Chi-Square not being able to handle counts less than 5, it is apt to use FET.

Disease Transitions

Disease transitions are described here as simple or short disease trajectories of each participant from one disease to the next based on the ascending order of dates. Only two disease transition pairs are utilised for the simpler trajectory plots. Disease transitions were plotted using the Pheno308 table.

The concept of next disease is introduced here for mapping shorter trajectories. For each participant, their next immediate disease diagnosis is simply mapped next to their current disease. This can be better understood by a sample of the dataset displayed in Table 3 below. Hence, the disease transition pairs were simply a combination of current disease and next disease. The disease pairs for each participant were non-unique, meaning, if a participant had multiple instances of the same disease at different dates, it was counted multiple times. This approach leads to duplicate disease entries, but it also allows for a more realistic modelling of the data as it is presented from the EHR.

Index	participant_id	gene	current disease	next disease	date of ICD-10 code assignment
0	Participant1	APC	Asthma	Sickle-cell trait	2012-10-19
1	Participant1	APC	Sickle-cell trait	Bacterial Diseases (excluding TB)	2013-01-10
2	Participant1	APC	Bacterial Diseases (excluding TB)	Infections of Other or unspecified organs	2013-05-09
3	Participant1	APC	Infections of Other or unspecified organs	none	2014-09-21
4	Participant2	Control	Ear and Upper Respiratory Tract Infection	Epilepsy	2009-08-15
5	Participant2	Control	Epilepsy	Irritable bowel syndrome	2009-12-16
6	Participant2	Control	Irritable bowel syndrome	Anxiety disorders	2011-04-05
7	Participant2	Control	Anxiety disorders	none	2012-01-06

Table 3: Artificially generated dataset explaining the concept of next disease.

The disease transition pairs were assessed for their frequency out of total transition pairs within the gene or control group. This provided the percentage value depicting how frequent or common that disease transition pair was across that gene or control.

A horizontal side-by-side bar chart was generated for the top 10 most common disease transition pairs across genes carriers and controls. The x-axis showed the percentage values, and the y-axis showed the transition pairs.

Disease Trajectories

More complex or longer disease trajectories were constructed, now working with up to 4 diseases, for each gene and control group. There are multiple ways to plot disease transitions and visualise disease trajectories like network graphs, Sankey Diagrams, chord diagrams, linear trajectories, and more⁴. For this report, network graph has been chosen to plot the trajectories. Network graphs can map complex trajectories and quantify participant flow unlike Sankey diagrams, chord diagrams or linear trajectories. Additionally, for this report, other trajectory visualisation methods would not have been able to handle the long disease names without making the plot cluttered. Network graph is a simple yet effective tool that plots disease names in a circular node and shows the disease transition from *disease 1* to *disease 2* using directional (left to right) arrows. Two papers have successfully used network graphs for generating disease trajectories, one for mapping trajectories of 103 codes associated with sleep apnea disease⁵, another one mapping chronic conditions like heart disease, lung diseases using a network graph²¹. Another paper plotted cluster network analysis plots for Alzheimer's disease²².

For a participant who has four disease diagnoses A, B, C, and D respectively, with dates sorted in ascending order, the disease transition pairs, and overall disease trajectory can be represented like this:

$$A \rightarrow B, \quad B \rightarrow C, \quad C \rightarrow D$$

$$A \rightarrow B \rightarrow C \rightarrow D$$

Using **Table 1**, for *Participant1* and *Participant2*, their respective disease trajectories would look like this:

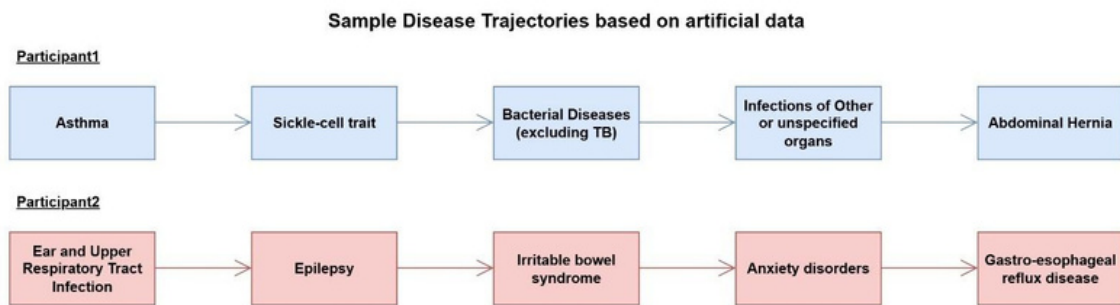


Figure 3: Sample disease trajectories generated using **Table 1** for two participants, one each from gene carriers and controls.

The width of the arrow indicates the number or count of participants that have the disease. The size of the node counts both incoming and outgoing flow of participants for that disease. Since different transition sequences can have common diseases, there is some overlap between disease transition sequences that are otherwise discreet. This results in overlapping arrows across the disease transition sequences.

Using the Networkx library²³, a network diagram was generated comparing the top 10 disease trajectories for each gene carrier and controls. Top 10 disease trajectories were created using frequency of participants for a disease sequence. Only those participants were plotted who had more than 3 disease diagnoses and could result in disease trajectory sequences of up to 4 diseases. A sliding window mechanism was employed which ensures all possible disease transitions are analysed as discreet transitions. Sample disease transition pairs can be better understood by Figure 4 below. There can be 3 distinct transitions for a participant who has 6 associated diseases.

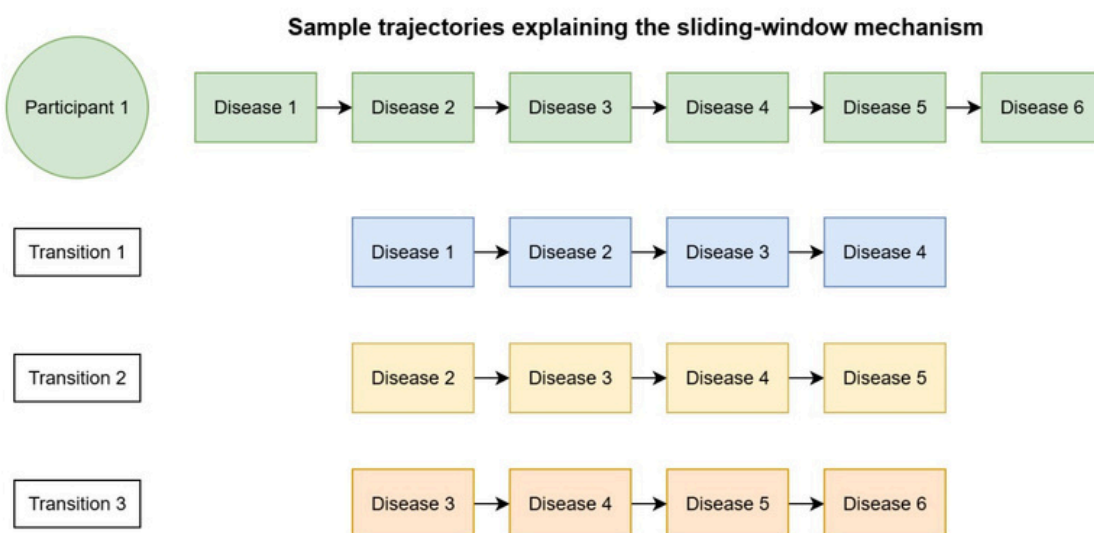


Figure 4: Sample disease transition pairs explained using an example of a participant who has 6 diseases.

Software

All the analysis was performed on Python using Jupyter Notebook inside the GE Research Environment. For data preparation, R Studio, Bash and Linux were used.

A detailed list of the Python libraries used for the analysis can be found in the Appendix 3.

Results

The threshold (>300 carriers) for determining which genes were carried by the case population yielded 10 genes with substantial number of carriers. The 10 genes are: *APC*, *ATP7B*, *CACNA1S*, *DSP*, *FLNC*, *MSH2*, *MYH7*, *RYR1*, *SCN5A*, *TTN* (Figure 5).

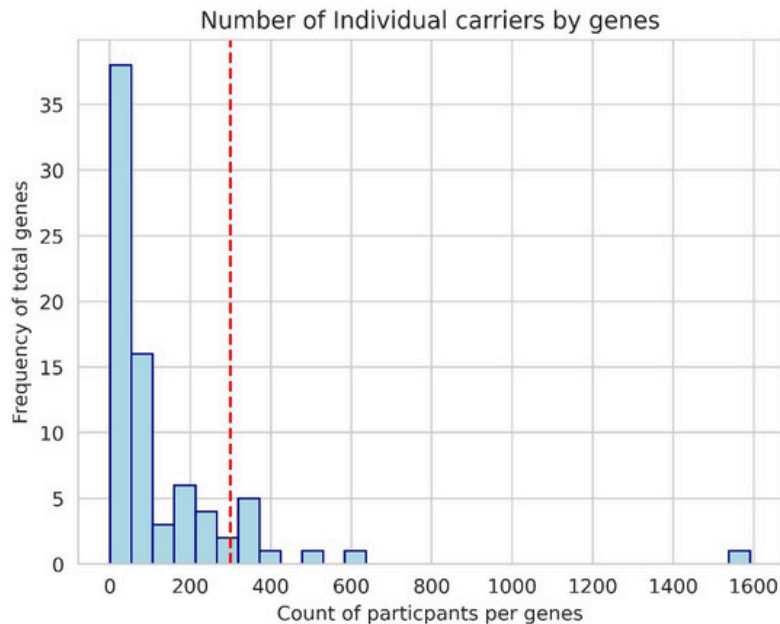


Figure 5: Histogram of counts of carrier participants by their associated ACMG genes. The red dotted line represents the threshold of inclusion, i.e. more than 300 carriers.

The total cohort sample size was 24,100 participants with 18,951 controls and 4,956 unique gene carriers. Since one participant can be associated with more than one gene mutation, the number of carriers for each gene were 5,149. *TTN* gene had the most carriers (N = 1,591) followed by *SCN5A* (N = 598); the rest of the genes had between 300 and 400 carriers.

Full demographic breakdown for the overall samples stratified by gene, is given in Table 4 below. There is no missing data except for the participant 'length of stay' variable (N = 30). The sex of participants consistently follows an approximate Female to Male split of 55 to 45% $\pm$ 5% across all carriers and 63.6% to 36.4% for controls. The average age of the carriers is consistently around 47 $\pm$ 2 years and that of controls is 49.3 years. controls have a much higher percentage of participants falling in the 40-50 (34.9%) years bracket than all the carriers (18.5% $\pm$ 6). The participants have a majority self-reported White ethnicity (68.8%) which is inconsistent across most carriers and controls. However, carriers of *CACNA1S* have a much higher percentage of Asian ethnicity (17%) and a much lower White population (56.3%), lowest across the cohort. Majority of participants across the cohort had a European ancestry followed by South Asians. Although, carriers of gene *MYH7* had 82% European population, highest among genes and controls.

N	Variables (categorical)	Overall 24100 14912 (61.9)	APC 356 203 (57.0)	ATP7B 314 169 (53.8)	CACNA1S 366 197 (53.8)	DSP 337 189 (56.1)	FLNC 523 280 (53.5)	MSH2 333 195 (58.6)	MYH7 392 202 (51.5)	RYR1 339 193 (56.9)	SCN5A 598 333 (55.7)	TTN 1591 896 (56.3)	controls 18951 12055 (63.6)
Sex, n (%)	Female	9188 (38.1)											6896
	Male		153 (43.0)	145 (46.2)	169 (46.2)	148 (43.9)	243 (46.5)	138 (41.4)	190 (48.5)	146 (43.1)	265 (44.3)	695 (43.7)	(36.4) 49.3
Age, mean (SD)		48.6 (15.4)	45.2 (22.0)	44.3 (22.3)	45.0 (21.4)	49.5 (21.7)	45.7 (22.5)	46.4 (20.5)	47.9 (22.3)	46.0 (23.7)	47.2 (22.1)	46.1 (22.0)	(12.9) 28 (0.1)
Age group, n (%)	0-10	371 (1.5)	22 (6.2)	22 (7.0)	27 (7.4)	16 (4.7)	45 (8.6)	15 (4.5)	24 (6.1)	30 (8.8)	36 (6.0)	106 (6.7)	72 (0.4)
	10-20	685 (2.8)	47 (13.2)	43 (13.7)	43 (11.7)	30 (8.9)	59 (11.3)	35 (10.5)	41 (10.5)	47 (13.9)	62 (10.4)	206 (12.9)	430 (2.3)
	20-30	719 (3.0)	20 (5.6)	20 (6.4)	20 (5.5)	23 (6.8)	34 (6.5)	25 (7.5)	21 (5.4)	12 (3.5)	37 (6.2)	77 (4.8)	4350
	30-40	5000 (20.7)	43 (12.1)	41 (13.1)	53 (14.5)	36 (10.7)	62 (11.9)	39 (11.7)	57 (14.5)	43 (12.7)	83 (13.9)	193 (12.1)	(23.0) 6614
	40-50	7572 (31.4)	85 (23.9)	63 (20.1)	75 (20.5)	56 (16.6)	84 (16.1)	71 (21.3)	57 (14.5)	59 (17.4)	113 (18.9)	295 (18.5)	(34.9) 4249
	50-60	5070 (21.0)	43 (12.1)	51 (16.2)	61 (16.7)	58 (17.2)	84 (16.1)	64 (19.2)	62 (15.8)	45 (13.3)	93 (15.6)	260 (16.3)	(22.4)
	60-70	2382 (9.9)	40 (11.2)	24 (7.6)	37 (10.1)	57 (16.9)	86 (16.4)	44 (13.2)	64 (16.3)	41 (12.1)	74 (12.4)		
	70-80	1527 (6.3)	39 (11.0)	29 (9.2)	36 (9.8)	35 (10.4)	41 (7.8)	30 (9.0)	42 (10.7)	42 (12.4)	61 (10.2)	207 (13.0)	1708 (9.0)
	>80	774 (3.2)	17 (4.8)	21 (6.7)	14 (3.8)	26 (7.7)	28 (5.4)	10 (3.0)	24 (6.1)	20 (5.9)	39 (6.5)	180 (11.3)	992 (5.2)
Participant type, n (%)	Cancer	1089 (4.5)	64 (18.0)	59 (18.8)	135 (36.9)	120 (35.6)	225 (43.0)	123 (36.9)	158 (40.3)	136 (40.1)	242 (40.5)	331 (20.8)	0
	Proband	2041 (8.5)	145 (40.7)	126 (40.1)	162 (44.3)	134 (39.8)	194 (37.1)	143 (42.9)	142 (36.2)	120 (35.4)	219 (36.6)	631 (39.7)	0
	Relative	20970 (87.0)	147 (41.3) 13 (3.7)	129 (41.1)	9 (2.5)	31 (9.2)	60 (11.5)	8 (2.4)	8 (2.0)	24 (7.1)	36 (6.0)	629 (39.5) 40 (2.5)	18951 (100.0)
			<5		<5	6 (1.8)	<5	<5	0	<5	<5		391 (2.1)
Ancestry, n (%)	African	625 (2.6)	267 (75.0)	5 (1.6)	252 (88.9)	246 (73.0)	353 (87.5)	236 (70.9)	324 (82.7)	257 (75.8)		19 (1.2)	92 (0.5)
	East Asian	137 (0.6)	20 (5.6)	<5	35 (9.6)	23 (6.8)	67 (12.8)	23 (6.9)	33 (8.4)	31 (9.1)			14974
	European	18647 (77.4)	54 (15.2)	218 (69.4)	<5	0	<5	0	0	<5	466 (77.9)	1054 (66.2)	(79.0)
	Other	1841 (7.6)	0	42 (13.4)	62 (16.9)	24 (7.1)	27 (5.2)	55 (16.5)	25 (6.4)	19 (5.6)			
					13 (3.6)	25 (7.4)	60 (11.5)	7 (2.1)	7 (1.8)	15 (4.4)	56 (9.4)	220 (13.8)	1291 (6.8)
	South Asian	2780 (11.5)	46 (12.9)	47 (15.0)	70 (19.1)	5 (1.5)	23 (4.4)	6 (1.8)	6 (1.5)	16 (4.7)	37 (6.2)		
	American	70 (0.3)	11 (3.1)	0	9 (2.5)	11 (3.3)	19 (3.6)	<5	10 (2.6)	11 (3.2)			
Ethnic category, n (%)	Asian or Asian British	2524 (10.5)	9 (2.5)	47 (15.0)									
	Black or Black British	627 (2.6)	48 (13.5)	6 (1.9)									
	Mixed	411 (1.7)	<5	5 (1.6)									
	Unknown / Missing	3541 (14.7)	238 (66.9)	61 (19.4)									
	Other	412 (1.7)		<5									
	White	16585 (68.8)		191 (60.8)									
		4518 (18.7)											
Related, n (%)	Unknown	19582 (81.3)	116 (32.6)	96 (30.6)	115 (31.4)	84 (24.9)	163 (31.2)	99 (29.7)	105 (26.8)	99 (29.2)	163 (27.3)	485 (30.5)	(15.8) 15958
	Unrelated		240 (67.4)	218 (69.4)	251 (68.6)	253 (75.1)	360 (68.8)	234 (70.3)	287 (73.2)	240 (70.8)	435 (72.7)	1106 (69.5)	(84.2) 10 (1 – 173)
N_ICD-10, median (range)	ICD10 Table	11 (1 – 306)	16 (1 – 106)	14.5 (1 – 153)	15 (1 – 178)	17 (1 – 104)	15 (1 – 173)	13 (1 – 162)	17 (1 – 118)	15 (1 – 149)	15 (1 – 306)	15 (1 – 171)	8 (1 - 44)
	Pheno308 Table	9 (1 - 49)	12 (1 - 30)	12 (1 - 48)	13 (1 - 41)	13 (1 - 31)	13 (1 - 38)	11 (1 - 37)	14 (1 - 33)	13 (1 - 42)	13 (1 - 49)	12 (1 - 49)	11.4
Length of stay, mean (SD)	Missing (n = 30)	14.8 (48.2)	22.7 (48.5)	34.5 (90.2)	26.7 (54.5)	28.1 (110.1)	31.3 (110.9)	21.4 (49.7)	26.4 (65.7)	41.1 (216.8)	30.0 (81.0)	23.3 (52.6)	(26.4) 18619
Status, n (%)	Alive	23307 (96.7)	333 (93.5) 23 (6.5)	293 (93.3)	329 (89.9) 37 (10.1)	296 (87.8) 41 (12.2)	482 (92.2) 41 (7.8)	302 (90.7) 31 (9.3)	347 (88.5) 45 (11.5)	303 (89.4) 36 (10.6)	539 (90.1)	1464 (92.0)	332 (1.8)
	Died	793 (3.3)	240 (67.4) 116 (32.6)	21 (6.7)	237 (64.8) 129 (35.2)	227 (67.4) 110 (32.6)	368 (70.4) 155 (29.6)	209 (62.8) 124 (37.2)	275 (70.2) 117 (29.8)	236 (69.6) 103 (30.4)	59 (9.9)	127 (8.0)	0
Affection status, n (%)	Unknown	3485 (14.5)		203 (64.6)									18951
	Unaffected	20615 (85.5)		111 (35.4)							186 (31.1)	513 (32.2)	(100.0) 3173
IMD Quantiles, n (%)	Q1 (0.53 – 9.03)	4046 (16.8)	53 (14.9)	62 (19.7)	53 (14.5)	63 (18.7)	76 (14.5)	59 (17.7)	70 (17.9)	49 (14.5)	111 (18.6)	277 (17.4)	(16.7) 3148
	Q2 (9.03 – 15.37)	4040 (16.8)	76 (21.3)	56 (17.8)	66 (18.0)	58 (17.2)	93 (17.8)	41 (12.3)	77 (19.6)	50 (14.7)	104 (17.4)	271 (17.0)	(16.6) 3136
	Q3 (15.37 – 23.93)	4037 (16.8)	59 (16.6)	47 (15.0)	67 (18.3)	56 (16.6)	91 (17.4)	63 (18.9)	61 (15.6)	67 (19.8)	109 (18.2)	281 (17.7)	(16.5) 3104
	Q4 (23.93 – 37.76)	4039 (16.8)	66 (18.5)	51 (16.2)	67 (18.3)	57 (16.9)	103 (19.7)	60 (18.0)	72 (18.4)	64 (18.9)	99 (16.6)	296 (18.6)	(16.4) 3187
	Q5 (37.76 – 84.22)	4041 (16.8)	54 (15.2)	57 (18.2)	60 (16.4)	61 (18.1)	93 (17.8)	61 (18.3)	56 (14.3)	57 (16.8)	94 (15.7)	261 (16.4)	(16.8) 3203
	Unknown	3897 (16.2)	48 (13.5)	41 (13.1)	53 (14.5)	42 (12.5)	67 (12.8)	49 (14.7)	56 (14.3)	52 (15.3)	81 (13.5)	205 (12.9)	(16.9)

Table 4: Demographics of the Cohort. The stratification is by ACMG genes with more than 300 carriers. Description of the variables a) Participant Type – Three types of participants based on

their disease being cancer, rare disease or relative (not cancer or rare disease), b) Ancestry – Ancestry of Participants, c) Ethnic Category – Simplified self-reported ethnic category of participants, d) Related – Related to a patient with rare disease, or Unknown relation d) Number of ICD10 codes – Total number of unique ICD-10 Codes associated with a participant, e) Length of Stay – how long the participant was admitted to a hospital for, f) Status – Patient Status Alive or Dead, g) Affection Status – Affected participants having a rare disease or cancer diagnosis or Unaffected participants, h) IMD Quantiles – IMD scores of participants put into 5 quantiles. Counts equal to or below 5 are masked as <5 as per GE RE policy. Higher IMD score means higher deprivation.

Disease category distributions help analyse the burden of disease. In Figure 6, the side-by-side bar chart compares the disease category distribution for each gene as compared to controls. Digestive and Infectious diseases constitute either higher or equal percentage of total diseases for controls than most gene carriers, excluding *CACNA1S* and *ATP7B* for Infections. Across all gene carriers, percentage of participants with cancers and perinatal diseases is at least twice as high, or more, compared to controls. On the contrary, benign neoplasms or carcinoma in situ diseases have a slightly higher presence in controls than all gene carriers.

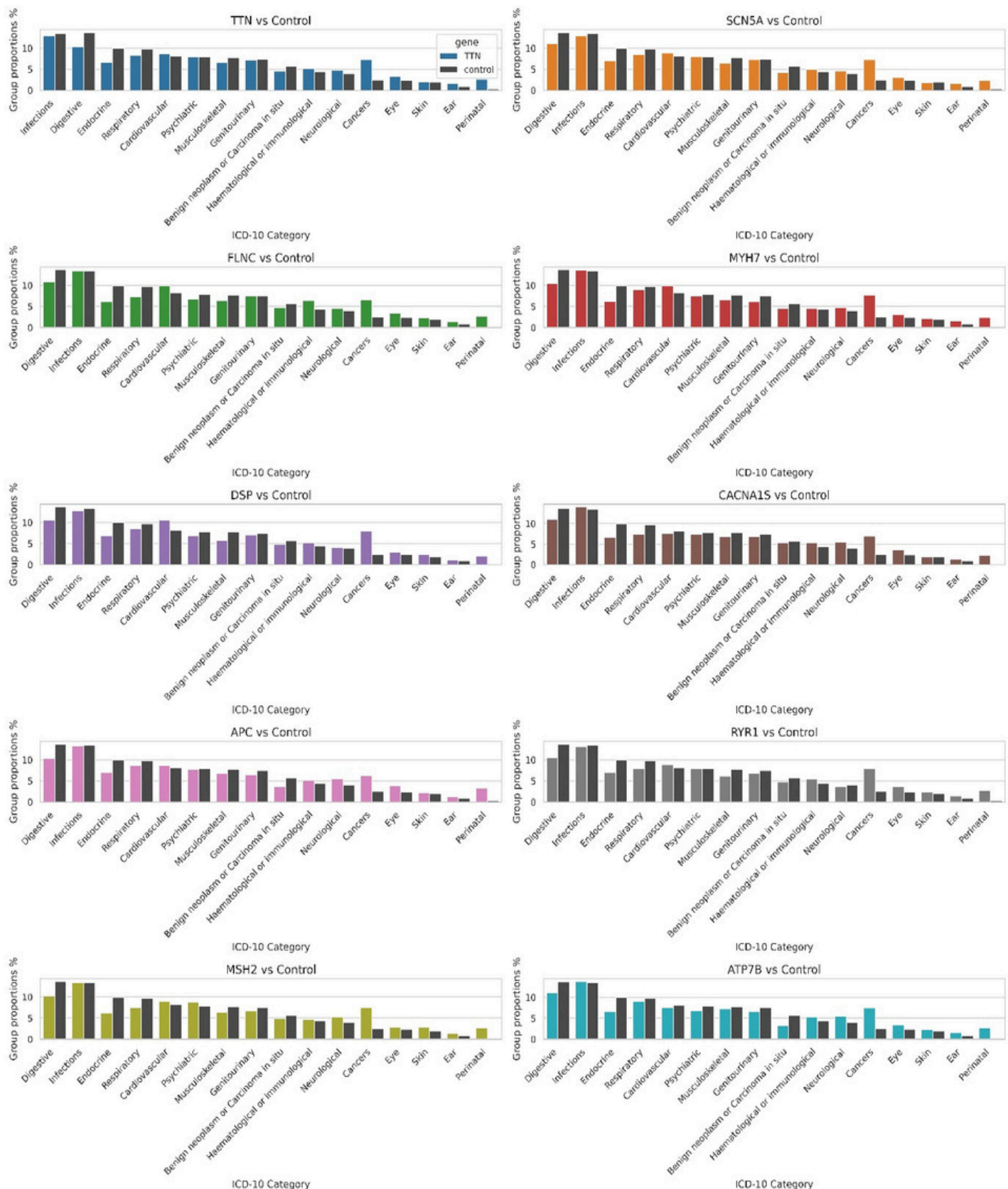


Figure 6: Bar plot for disease categories comparison between gene carriers and controls. Each gene is represented by a different colour, and controls are shown in dark grey. X-axis has the ICD-10 disease categories, y-axis shows the within group proportions or percentage of participants with that disease category.

The findings from the above plot can be statistically explored using a line plot with Odds Ratios plotted for each disease category.

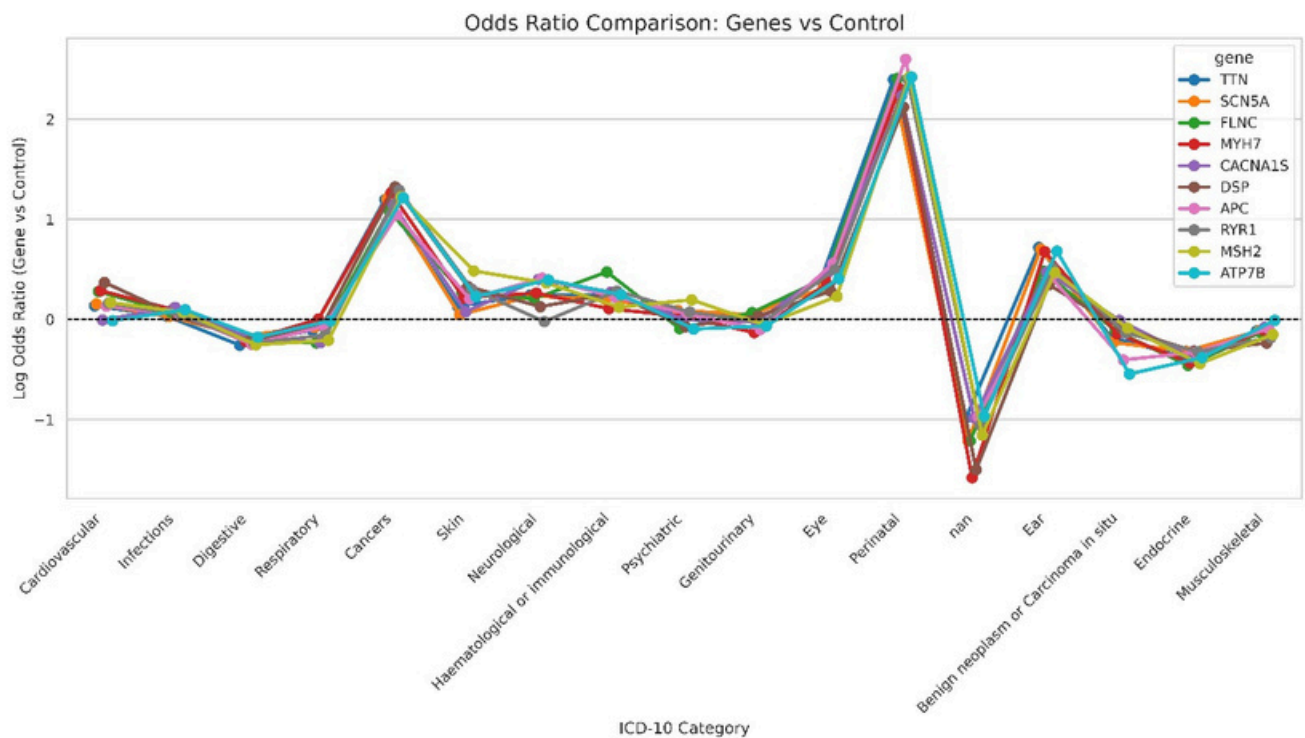


Figure 7: Line plot for disease categories showing comparison of OR between gene carriers and controls. X-axis has the ICD-10 disease categories; Y-axis has the Log OR values. Dotted line represents log OR = 0 baseline.

In Figure 7, values falling above the baseline at log OR = 0 indicate gene carriers have higher odds associated with that disease than controls. Log OR values below 0 indicate a protective effect, or the disease having lower odds of presenting itself in carriers than controls. Using a natural log conversion table, log OR values of 1 and 2 convert to OR values of 2.7 and 7.4 respectively between the groups being compared. Hence, cancer disease category indicates 2.7 times higher odds and ear diseases have 1.67 times higher odds of having the disease in all gene carriers than controls. Similarly, for perinatal diseases, gene carriers have 7.4 times higher odds of having the disease compared to controls. Disease categories, like cardiovascular, neurological, and hematological and immunological, have log OR values slightly higher than 1 which corresponds to slightly higher odds for the disease in gene carriers than controls. On the other hand, benign neoplasms or carcinoma in situ has OR value of -0.6 for genes *ATP7B* and *APC* which corresponds to 62.5% reduced odds of having the disease compared to controls. Similarly, for all gene carriers, the odds of having endocrine diseases are reduced by 62.5% (OR = -0.6) compared to controls. Lastly, nan values were assigned to participants without disease categories in either group. Here, these disease categories are not present in gene carriers while being present in the controls.

Disease Transition Pairs

Simpler disease trajectories can be analysed using disease transitions. Disease transition pairs are the most frequent successive disease pairs observed within each gene and

control group. Figures 8 and 9 show the percentages of top 10 disease transition pairs per gene compared with controls. The top disease transition pair across all gene carriers was “*Lower Respiratory Tract Infections → Other or unspecified infectious organisms*” (1.5% ± 0.5%) except for the *FLNC* gene carriers. The same transition pair was more frequently present in all gene carriers (except *FLNC*) than controls (1.35%).

Either first or second most common transition pairs observed for carriers of genes *FLNC*, *CACNA1S*, *ATP7B*, *MYH7*, *MSH2*, *RYR1*, *SCN5A*, and *TTN* was “*Ear and Upper Respiratory Tract Infections → Other or unspecified infectious organisms*”. *RYR1* gene shows 8 out of 10 trajectory pairs that are more common in carriers than controls, most for any gene carrier. In contrast, for *SCN5A* and *DSP* genes, 6 out of 10, and 7 out of 10, disease transition pairs were more common in controls than the gene itself respectively. Additionally, genes *FLNC* and *MYH7* had equal number of transition pairs, 5 each, common between gene carriers and controls. For all genes, the transition pairs were dominated by Infectious diseases with at least 1 disease being an infectious disease out of the 10 disease transition pairs.

Top 10 disease-pair trajectories per gene vs Controls (%) - First 5 genes

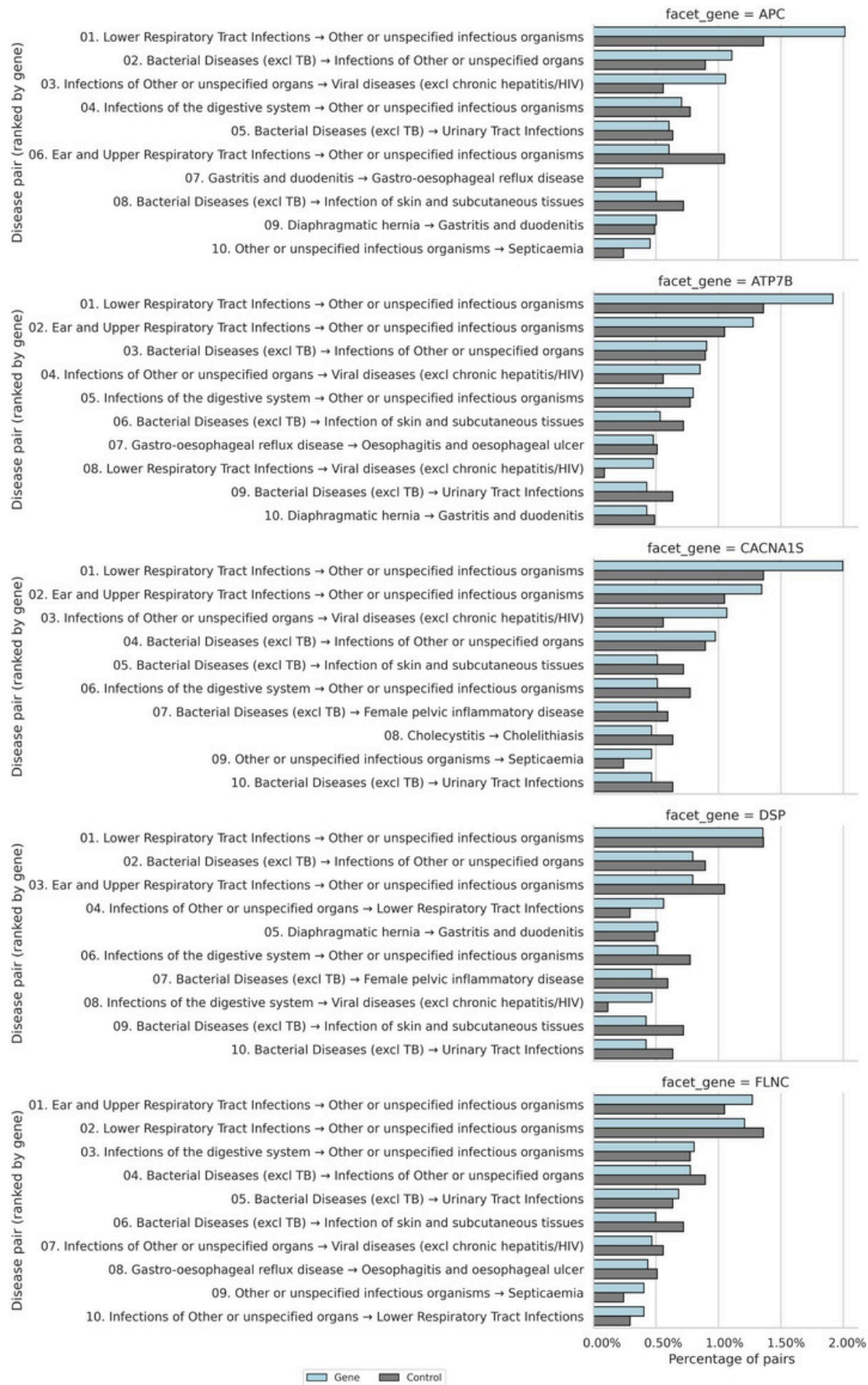


Figure 8: Side-by-side bar plot showing the disease transition pairs percentage of occurrence of first 5 gene carriers compared to controls. Genes are light blue colour, controls are grey.

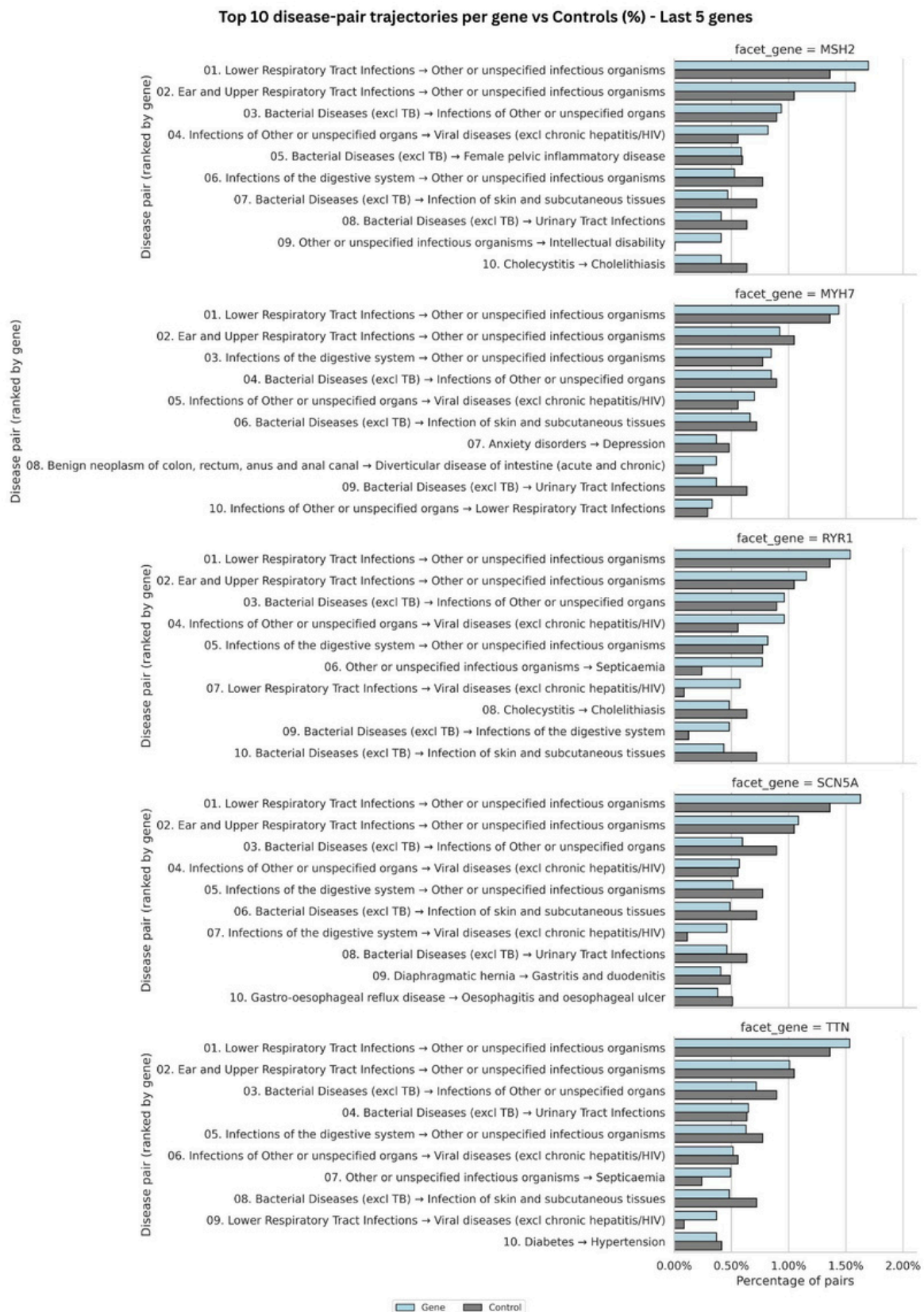


Figure 9: Side-by-side bar plot showing the disease transition pairs percentage of occurrence of remaining 5 gene carriers compared to controls. Genes are light blue colour, controls are grey.

Disease Trajectories

More complex disease trajectories are used to analyse longer sequences of disease transitions. Each node represents a disease, and the node size represents the frequency of participants for that disease. Each arrow connecting the nodes is thicker for higher frequency of participants having the disease and thinner vice-a-versa.

Disease trajectories for *RYR1* gene demonstrate different patterns in Top 10 disease transitions than controls (Figure 10). For the same disease transition sequences in *RYR1*, there are more prominent transitions flowing through the disease nodes than controls.

The node size is also much bigger in *RYR1* than controls, for the same nodes along with thicker arrows connecting the nodes.

For example, the below sequence in *RYR1* has thicker arrows and bigger node sizes than their control counterparts.

*Infections of Other of unspecified organs -->
Viral diseases (excl. hepatitis/HIV) -->
Lower respiratory tract infections*

Overall, the trajectory analysis for *RYR1* shows that for most diseases in the Top 10 disease sequences, the transition pairs are much more prominent in *RYR1* than controls. Some transition pairs were more common in the controls than *RYR1* carriers. For example,

Disease trajectory network graphs for other genes are attached in the Appendix 4.1 – 4.9.

Overall, for rest of the genes, their trajectory plots show more prominent transition pairs for respective diseases sequences, in terms of participant frequencies and node sizes, in gene carriers than controls. However, in some instances, the controls do show more frequent transitions compared to gene carriers. E.g., the below transition sequence is more prominent in the controls than the carriers of *FLNC* gene.

*Diaphragmatic Hernia -->
Gastritis and Duodenitis-->
Gastro-oesophageal reflux disease -->
Oesophagitis and oesophageal ulcer*

Disease Network Graph: RYR1 vs Controls — top 10 sequences

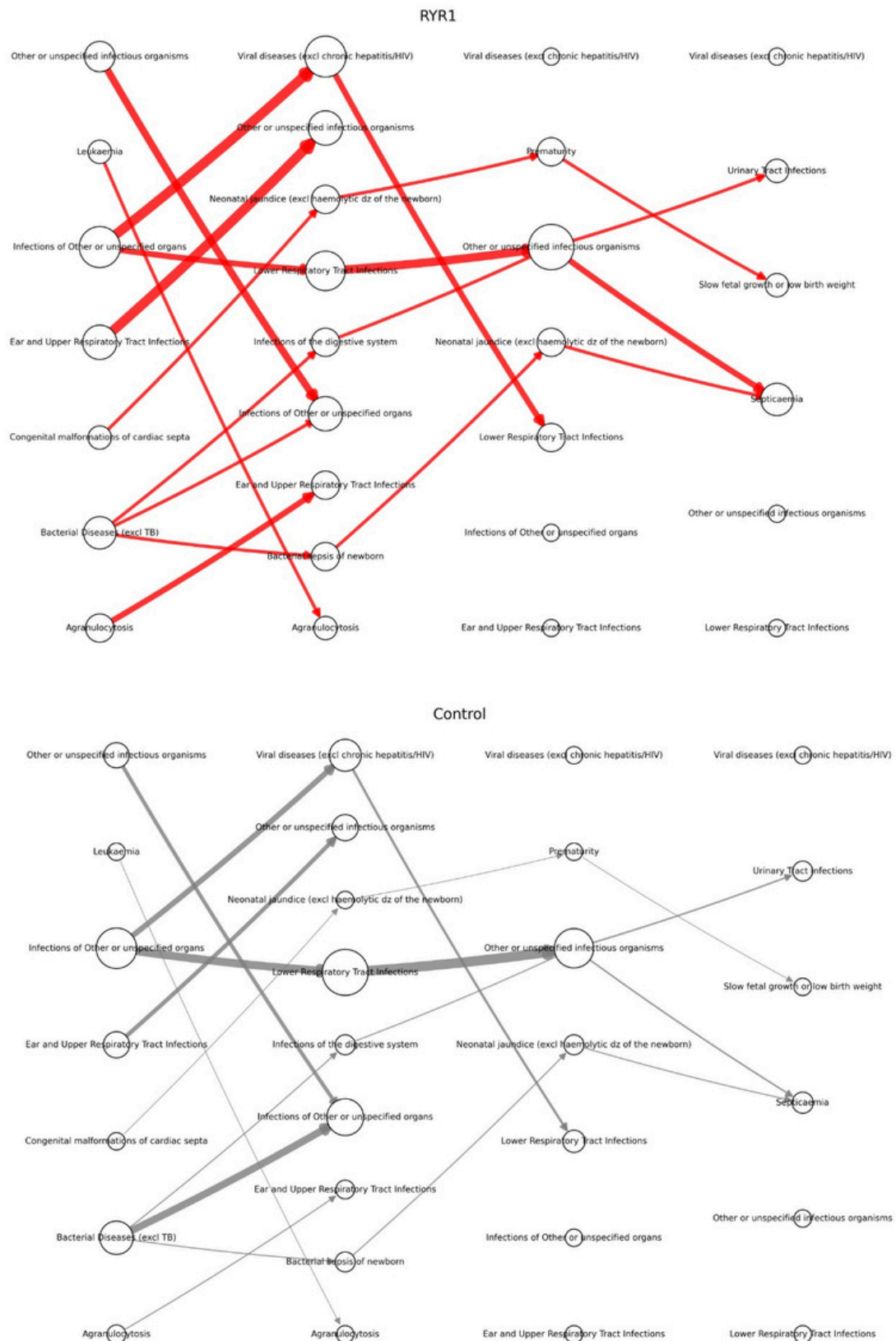


Figure 10: Disease trajectories for Top 10 disease transitions for the carriers of gene *RYR1* compared with controls. Red colour is for *RYR1* progressions, grey is for controls.

Discussion

This report yielded interesting results comparing each gene carrier to controls assessing the burden of disease categories, disease transitions, and complex disease trajectories.

The demographic table (Table 4) showed that most variables displayed similar trends across the gene carriers and the controls except age. Controls (mean age 49.3 years), on average, were older than the gene carriers (mean age 47 ± 2 years). Although it is hard to ascertain the reason behind this discovery as the study design was balanced in terms of participant origins (same dataset) and segregation into carriers and controls. Age is a strong risk factor for several chronic diseases, and it can allow for a different manifestation of diseases and patterns at different age groups. Hence, controls might share different disease patterns compared to gene carriers due to difference in age. Future studies should consider adjusting for age, sex, and other confounding.

Disease burden for all disease categories was visualised and statistically explored in a side-by-side bar chart comparing gene carriers with controls. The difference in higher or lower odds for specific disease categories in cases and controls was observed. Gene carriers had higher odds of cancer (log OR = 2.7), ear (log OR = 0.5) and perinatal diseases (log OR = 7.4). On the other hand, for digestive and infectious diseases, gene carriers were protected or had lower odds. The higher cancer odds in gene carriers can be contributed to the pathogenic nature of mutations in the ACMG genes as discussed below.

A disease is considered rare if it has a low prevalence rate. However, this threshold for recognising rare diseases is highly debated among different geographies and health organisation, hence, there is no single consensus. Therefore, a prevalence threshold of less than 40-50 per 100,000 people is being used in this report to accommodate both more common rare diseases and ultra rare disease. Through Table 5, a pattern of disease-gene associations can be observed using other independent research studies. Pathogenic variants in all the examined ACMG genes are associated with an increased risk of developing a cancer or a rare disease. Carriers of ACMG genes, *APC* and *MSH2* have shown increased risk of developing cancers. While carriers of the rest of the 8 ACMG genes have shown a higher risk of developing various rare diseases.

Cardiac diseases are the most frequently observed diseases with genes *TTN*, *MYH7*, *RYR2*, *Brugada*, *SCN5A* and *ILNC* being associated with cardiomyopathies and *SCN5A* Syndrome, a rare cardiac disorder. Other genes are associated with musculoskeletal diseases and cancers. Hence, this strong association of the shortlisted ACMG genes with cancer and cardiac diseases can explain the higher odds associated with those disease categories in cases than controls. Additionally, the gene carriers showed lower odds of having 'benign neoplasms or carcinoma in situ' (benign tumours or pre-malignant tumours) than controls. This finding is inconsistent with the previous discussion point because cancer and pre-malignant neoplasms should ideally follow similar patterns of disease distribution. This finding needs to be explored further by dissecting all the disease transition sequences for each gene carrier. Furthermore, to add more robustness, the analysis can be conducted using other independent datasets.

Gene	Associated Disease	Cancer/ Rare Disease	Gene-Disease Description	Source
APC	Colorectal Cancer	Cancer	APC has been associated with colorectal cancer.	26–28
ATP7B	Wilson Disease	Rare Disease	Wilson's disease is rare disorder which accumulates copper in the body affecting mainly brain and liver. It has a prevalence of 1 in 30,000 globally.	29
CACNA1S	Hypokalemic periodic paralysis (HypoPP)	Rare Disease	HypoPP is rare skeletal muscle disorder that has a prevalence of 1 in 100,000 people.	30
	Congenital Myopathy		Congenital Myopathy group of disorders is a congenital muscle disorder and affects motor functions. It has a prevalence of 1.62 per 100,000 people. It is a disease which scars or affects the lung tissue. It has a prevalence of 16.5 per 100,000 people.	31
DSP	Idiopathic pulmonary fibrosis	Rare Disease	FLNC gene is associated with all types of cardiomyopathies which are general inherited cardiac disorders. These weaken the cardiac wall muscles and hamper the functioning of heart. These diseases have a collective prevalence of 40 in 100,000 people.	32,33
FLNC	Cardiomyopathies	Rare Disease	Lynch Syndrome is a genetic phenomenon which is caused by a mutation in certain genes. It highly increases the risk of developing colorectal and other cancers.	34,35
MSH2	Lynch Syndrome	Cancer		36,37
MYH7	Hypertrophic cardiomyopathy	Rare Disease	Hypertrophic cardiomyopathy has a prevalence of 33 per 100,000 people.	38
RYR1	Central Core Disease (CCD)	Rare Disease	CCD is a type of cardiomyopathy that has an unknown prevalence. However, overall prevalence of cardiomyopathies is 40 per 100,000 people.	35,39,40
SCN5A	Brugada Syndrome	Rare Disease	Brugada Syndrome is a cardiac disorder which causes tachycardia and increases the chances of heart attack. It has a prevalence of 5 per 10,000 people.	41,42
TTN	Cardiomyopathy	Rare Disease	TTN is associated with several types of cardiomyopathies, and they have a collective prevalence of 40 per 100,000 people.	35,43
	Limb-girdle muscular dystrophy (LGMD)		LGMD is a rare type of muscular dystrophy with a prevalence of 5.5 per 100,000 people.	44,45

Table 5: Gene-disease associations for each gene carrier; focusing on rare disease or cancers.

In simple trajectories exploring shorter disease transition sequences, it was observed that consistently, the disease transition pairs that were common in gene carriers were not as common in controls. Only exception was the *DSP* gene where 7 out of 10 transition pairs were more common in controls than the gene carriers of *DSP*.

For longer and more complex disease trajectories, the transitions in gene carriers were more prominent than controls. This was true for all the gene carriers. Some shorter transition sequences of 1 disease pair or the corresponding two diseases were found to be more common amongst controls than gene carriers. However, these were exceptions and not the norms.

These trajectories have not been statistically verified; hence, the only inference that can be drawn is how common the disease trajectories were amongst gene carriers than controls. To statistically verify these claims, further tests like Markov analysis can be used. Markov test is further elaborated in the future works section of this report.

Strengths, Limitations, and Challenges

Strengths

A study of this kind has not been done on pathogenic variants, especially in ACMG genes. Hence, the insights gathered here can help understand disease transitions, and longer patient disease transition sequences better. Another strength of this report lies in its direct visual comparison of gene carriers and controls. The report presents a unique way to make strong inferences about disease burden, simpler or shorter disease transitions and longer or more complex disease trajectories.

Limitations

First limitation of the study was focusing only on the top 10 disease trajectory pairs which does not allow the analysis of rarer disease transition pairs that might still hold clinical relevance. Secondly, comparison of difference in disease trajectories between gene carriers and controls was not statistically validated which prevents making substantial claims about the results.

Lastly, for the OR calculations, adjustment for sex, age, and other variables was not performed, hence confounding of variables cannot be eliminated and the OR values are unadjusted. Further replication of this study should involve using a more robust statistical approach.

Challenges

Most papers implementing disease trajectories focus on one disease or a certain medical niche. Many research papers mapping disease trajectories explore natural histories of the patients, which is different from disease trajectories explored in this study. 'Disease progressions' usually analyse progression through various stages of the same disease. Disease trajectories analyse connections between different diseases.

The task of creating a visually compelling disease trajectory is hence much simpler for a restricted use case. In this report, since the focus is on 300 diseases for around 24,000 participants, it is difficult to produce longer trajectories or a network diagram with all the possible diseases transitions in one. Lastly, the results of this study are hard to compare with existing literature as no such study has been done earlier for disease trajectories in pathogenic genes.

Future work

Although different approaches have been utilised to map disease trajectories in this report, many other interesting analyses can be implemented which are beyond the scope of this current study.

Statistical quantification of the disease burden analysis and mapped disease trajectories is of utmost significance to establish a difference in the explored variables for cases and controls. For disease trajectories, Markov Analysis can be utilised to statistically quantify the findings. Markov analysis is widely used in finance and predicting stock market prices,

but it has since also been adapted for analysing disease progressions^{46,47}. Markov analysis could interpret whether the disease transitions are independent or dependent on previously identified disease transitions. In other words, whether non-immediate disease history of the patient affects which disease that person is diagnosed with later in life. Furthermore, based on a non-exhaustive but in-depth literature search for this report, a Markov model has not yet been applied to assess longer disease transitions or trajectories for pathogenic genes.

More exploratory and statistical analysis of the disease transition pairs can be done using a cluster analysis like the one implemented by Fu and Chang, 2024²². The 'greedy method' used to combine shorter transitions that share a path with longer trajectories into one longer trajectory would be especially useful for this report. Clustering can provide unique insights into hidden disease patterns and help analyse clusters with high similarity. Both tests will help validate the findings of this report.

Conclusion

This study has successfully explored burden of disease and mapped shorter and longer disease transitions between carriers of pathogenic variants in ACMG genes and controls to detect clear deviation in disease transition patterns. Burden of disease was found to be higher in gene carriers for cancer, ear, and perinatal disease categories. Higher burden was also observed for digestive, endocrine, and pre-malignant cancers amongst controls than gene carriers. Additionally, the shorter disease transitions showed prominent disease trajectory patterns amongst gene carriers than controls. Similar findings were observed for longer disease trajectories for the gene carriers than controls. Overall, this study was able to establish clear differences in disease progression patterns amongst carriers of ACMG genes compared to controls. However, the study can be made more robust with conducting relevant statistical analysis in the future.

Ethical Issues

This project was strictly done entirely in the GE secure RE, this includes access to all the datasets, writing of scripts, and generation of plots.

Training was performed to access the GE secure RE. I undertook training to deal with cyber-safety issues like phishing, data theft, etc. and how to prevent and report them. I have also been approved as a GE researcher working on a pre-approved project of my supervisor.

The ethical approval for patient data to be used in research was acquired from the participants in 2015 by the Health Research Authority Research Ethics Committee⁴⁸. The plots and tables exported from GE RE must go through an airlock system. It ensures participant safety by masking identifiable data and hiding counts smaller than five. The export requests are reviewed and approved by a member of the GE staff.

It is difficult to take out code and scripts from the environment due to Genomics England Limited's strict privacy policies. However, my supervisor has full editorial access to my code in the RE.

Lastly, no generative AI was used in this report in any capacity.

References

1. Shen, Y., Yu, J., Zhou, J. & Hu, G. Twenty-Five Years of Evolution and Hurdles in Electronic Health Records and Interoperability in Medical Research: Comprehensive Review. *J. Med. Internet Res.* **27**, e59024 (2025) Available at <https://doi.org/10.2196/59024>.
2. Pan, M. *et al.* Genetic variants for Alzheimer's disease and comorbid conditions. *J. Alzheimers Dis. JAD* **102**, 470–479 (2024) Available at <https://doi.org/10.1177/13872877241289054>.
3. Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C. & Roland, M. Defining Comorbidity: Implications for Understanding Health and Health Services. *Ann. Fam. Med.* **7**, 357–363 (2009) Available at <https://doi.org/10.1370/afm.983>.
4. Jørgensen, I. F., Haue, A. D., Placido, D., Hjaltelin, J. X. & Brunak, S. Disease Trajectories from Healthcare Data: Methodologies, Key Results, and Future Perspectives. *Annu. Rev. Biomed. Data Sci.* **7**, 251–276 (2024) Available at <https://doi.org/10.1146/annurev-biodatasci-110123-041001>.
5. Beck, M. K., Westergaard, D., Jensen, A. B., Groop, L. & Brunak, S. TEMPORAL ORDER OF DISEASE PAIRS AFFECTS SUBSEQUENT DISEASE TRAJECTORIES: THE CASE OF DIABETES AND SLEEP APNEA. in *Biocomputing 2017* 380–389 (WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA, 2017).
doi:10.1142/9789813207813_0036 Available at https://doi.org/10.1142/9789813207813_0036.
6. Rubio-Perez, C. *et al.* Genetic and functional characterization of disease associations explains comorbidity. *Sci. Rep.* **7**, 6207 (2017) Available at <https://doi.org/10.1038/s41598-017-04939-4>.

7. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013) Available at <https://doi.org/10.1038/gim.2013.73>.
8. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008) Available at <https://doi.org/10.1097/GIM.0b013e31816b5cae>.
9. ACMG SF v3.2. ACMG Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing. (2023).
10. Additional findings. *Genomics England*
<https://www.genomicsengland.co.uk/initiatives/100000-genomes-project/additional-findings>.
11. Giannoula, A., Centeno, E., Mayer, M.-A., Sanz, F. & Furlong, L. I. A system-level analysis of patient disease trajectories based on clinical, phenotypic and molecular similarities. *Bioinformatics* **37**, 1435–1443 (2021) Available at <https://doi.org/10.1093/bioinformatics/btaa964>.
12. Liu, J. *et al.* Clinical outcomes associated with pathogenic genomic instability mutations in prostate cancer: a retrospective analysis of US pharmacy and medical claims data. *J. Med. Econ.* **22**, 1080–1087 (2019) Available at <https://doi.org/10.1080/13696998.2019.1649267>.
13. NHS Genomic Medicine Service. *Genomics England*
<https://www.genomicsengland.co.uk/genomic-medicine/nhs-gms>.
14. NHS England. NHS England » NHS Genomic Medicine Service.
<https://www.england.nhs.uk/genomics/nhs-genomic-med-service/>.
15. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023) Available at <https://doi.org/10.1126/science.adg7492>.

16. Tordai, H. *et al.* Analysis of AlphaMissense data in different protein groups and structural context. *Sci. Data* **11**, 495 (2024) Available at <https://doi.org/10.1038/s41597-024-03327-8>.
17. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res.* **53**, D1313 (2024) Available at <https://doi.org/10.1093/nar/gkae1090>.
18. Pollard, T. J., Johnson, A. E. W., Raffa, J. D. & Mark, R. G. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open* **1**, 26–31 (2018) Available at <https://doi.org/10.1093/jamiaopen/ooy012>.
19. UK Ministry of Housing, Communities, and Local Government. The English Indices of Deprivation 2019 - FAQs. (2019).
20. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit. Health* **1**, e63–e77 (2019) Available at [https://doi.org/10.1016/S2589-7500\(19\)30012-3](https://doi.org/10.1016/S2589-7500(19)30012-3).
21. Chen, J. *et al.* Trajectories network analysis of chronic diseases among middle-aged and older adults: evidence from the China Health and Retirement Longitudinal Study (CHARLS). *BMC Public Health* **24**, 559 (2024) Available at <https://doi.org/10.1186/s12889-024-17890-7>.
22. Fu, M. & Chang, T. S. Identifying common disease trajectories of Alzheimer’s disease with electronic health records. 2024.07.26.24311084 Preprint at <https://doi.org/10.1101/2024.07.26.24311084> (2024) Available at <https://doi.org/10.1101/2024.07.26.24311084>.
23. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *scipy* (2008) doi:10.25080/TCWV9851 Available at <https://doi.org/10.25080/TCWV9851>.

24. Guo, J. *et al.* Aging and aging-related diseases: from molecular mechanisms to interventions and treatments. *Signal Transduct. Target. Ther.* **7**, 391 (2022) Available at <https://doi.org/10.1038/s41392-022-01251-0>.
25. Abozaid, G. M., Kerr, K., McKnight, A. & Al-Omar, H. A. Criteria to define rare diseases and orphan drugs: a systematic review protocol. *BMJ Open* **12**, e062126 (2022) Available at <https://doi.org/10.1136/bmjopen-2022-062126>.
26. Siraj, A. K. *et al.* Whole Exome-Wide Association Identifies Rare Variants in APC Associated with High-Risk Colorectal Cancer in the Middle East. *Cancers* **16**, 3720 (2024) Available at <https://doi.org/10.3390/cancers16213720>.
27. Zhan, Q. *et al.* An APC Mutation in a Large Chinese Kindred With Familial Adenomatous Polyposis Was Identified Using Both Next Generation Sequencing and Simple STR Marker Haplotypes. *Front. Genet.* **11**, (2020) Available at <https://doi.org/10.3389/fgene.2020.00191>.
28. APC gene: MedlinePlus Genetics. <https://medlineplus.gov/genetics/gene/apc/>.
29. Olsson, C., Waldenström, E., Westermarck, K., Landegren, U. & Syvänen, A.-C. Determination of the frequencies of ten allelic variants of the Wilson disease gene (ATP7B), in pooled DNA samples. *Eur. J. Hum. Genet.* **8**, 933–938 (2000) Available at <https://doi.org/10.1038/sj.ejhg.5200566>.
30. Hypokalemic periodic paralysis: MedlinePlus Genetics. <https://medlineplus.gov/genetics/condition/hypokalemic-periodic-paralysis/>.
31. Huang, K., Bi, F.-F. & Yang, H. A Systematic Review and Meta-Analysis of the Prevalence of Congenital Myopathy. *Front. Neurol.* **12**, 761636 (2021) Available at <https://doi.org/10.3389/fneur.2021.761636>.
32. Mathai, S. K. *et al.* Desmoplakin Variants Are Associated with Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **193**, 1151–1160 (2016) Available at <https://doi.org/10.1164/rccm.201509-1863OC>.

33. Idiopathic pulmonary fibrosis: MedlinePlus Genetics.
<https://medlineplus.gov/genetics/condition/idiopathic-pulmonary-fibrosis/>.
34. Muravyev, A. *et al.* Rare clinical phenotype of filaminopathy presenting as restrictive cardiomyopathy and myopathy in childhood. *Orphanet J. Rare Dis.* **17**, 358 (2022)
Available at <https://doi.org/10.1186/s13023-022-02477-5>.
35. Brownrigg, J. R. *et al.* Epidemiology of cardiomyopathies and incident heart failure in a population-based cohort study. *Heart* **108**, 1383–1391 (2022) Available at
<https://doi.org/10.1136/heartjnl-2021-320181>.
36. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam. Cancer* **10**, 297–301 (2011) Available at
<https://doi.org/10.1007/s10689-011-9427-0>.
37. Dominguez-Valentin, M. *et al.* Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. *Genet. Med.* **22**, 15–25 (2020) Available at <https://doi.org/10.1038/s41436-019-0596-9>.
38. Franke, M. *et al.* A MYH7 variant in a five-generation-family with hypertrophic cardiomyopathy. *Front. Genet.* **15**, (2024) Available at
<https://doi.org/10.3389/fgene.2024.1306333>.
39. O'Connor, T. N. *et al.* RYR-1-Related Diseases International Research Workshop: From Mechanisms to Treatments Pittsburgh, PA, U.S.A., 21–22 July 2022. *J. Neuromuscul. Dis.* **10**, 135–154 (2023) Available at <https://doi.org/10.3233/JND-221609>.
40. RYR1-Related Diseases - Symptoms, Causes, Treatment | NORD.
<https://rarediseases.org/rare-diseases/ryr-1-related-diseases/>.

41. Vutthikraivit, W. *et al.* Worldwide Prevalence of Brugada Syndrome: A Systematic Review and Meta-Analysis. *Acta Cardiol. Sin.* **34**, 267–277 (2018) Available at [https://doi.org/10.6515/ACS.201805_34\(3\).20180302B](https://doi.org/10.6515/ACS.201805_34(3).20180302B).
42. Campuzano, O. *et al.* Update on Genetic Basis of Brugada Syndrome: Monogenic, Polygenic or Oligogenic? *Int. J. Mol. Sci.* **21**, 7155 (2020) Available at <https://doi.org/10.3390/ijms21197155>.
43. Barrett, K. M. S. *et al.* Cardiomyopathy prevalence exceeds 30% in individuals with TTN variants and early atrial fibrillation. *Genet. Med.* **25**, (2023) Available at <https://doi.org/10.1016/j.gim.2023.100012>.
44. Khan, A. *et al.* Homozygous missense variant in the TTN gene causing autosomal recessive limb-girdle muscular dystrophy type 10. *BMC Med. Genet.* **20**, 166 (2019) Available at <https://doi.org/10.1186/s12881-019-0895-7>.
45. Alghamdi, O. A., Obaid, O., Sayed, A. G., Farhan, H. & Sayed, J. Autosomal Recessive Limb-Girdle Muscular Dystrophy Type 10 (LGMD,10), Caused by a Novel Homozygous Variant in the TTN Gene. *Int. Med. Case Rep. J.* **18**, 683–689 (2025) Available at <https://doi.org/10.2147/IMCRJ.S483508>.
46. Pasman, H. J. & Rogers, W. J. Chapter Three - Logic based methods for dynamic risk assessment. in *Methods in Chemical Process Safety* (eds. Khan, F. I. & Amyotte, P. R.) vol. 4 61–122 (Elsevier, 2020). Available at <https://doi.org/10.1016/bs.mcps.2020.02.001>.
47. Martino, A., Guatterri, G. & Paganoni, A. M. Multivariate Hidden Markov Models for disease progression. *Stat. Anal. Data Min. ASA Data Sci. J.* **13**, 499–507 (2020) Available at <https://doi.org/10.1002/sam.11479>.
48. Genomics England Ethical Approval. 100,000 Genomes Project gains ethical approval to offer NHS patients further information about their genomic results.

<https://www.genomicsengland.co.uk/news/100000-genomes-project-gains-ethical-approval> (2015).

Appendix

1. MeSH terms for literature search of disease trajectories in pathogenic variants:

Search: disease trajectory in pathogenic variants Filters: English, Humans, Exclude preprints

((("disease"[MeSH Terms] OR "disease"[All Fields] OR "diseases"[All Fields] OR "disease s"[All Fields] OR "diseased"[All Fields]) AND ("trajectories"[All Fields] OR "trajectory"[All Fields] OR "trajectory s"[All Fields]) AND ("pathogen"[All Fields] OR "pathogen s"[All Fields] OR "pathogene"[All Fields] OR "pathogenes"[All Fields] OR "pathogenic"[All Fields] OR "pathogenically"[All Fields] OR "pathogenicities"[All Fields] OR "pathogenicity"[MeSH Subheading] OR "pathogenicity"[All Fields] OR "virulence"[MeSH Terms] OR "virulence"[All Fields] OR "pathogenity"[All Fields] OR "pathogenous"[All Fields] OR "pathogens"[All Fields]) AND ("variant"[All Fields] OR "variant s"[All Fields] OR "variants"[All Fields])) AND ((excludepreprints[Filter]) AND (humans[Filter]) AND (english[Filter]))

2. GE ACMG list of genes for “additional findings”:

APC, APOB, BRCA1, LDLR, MEN1, MLH1, MSH2, MUTYH, PCSK9, RET, VHL

Complete ACMG list of 81 genes according to ACMG Secondary Findings Version 3.2

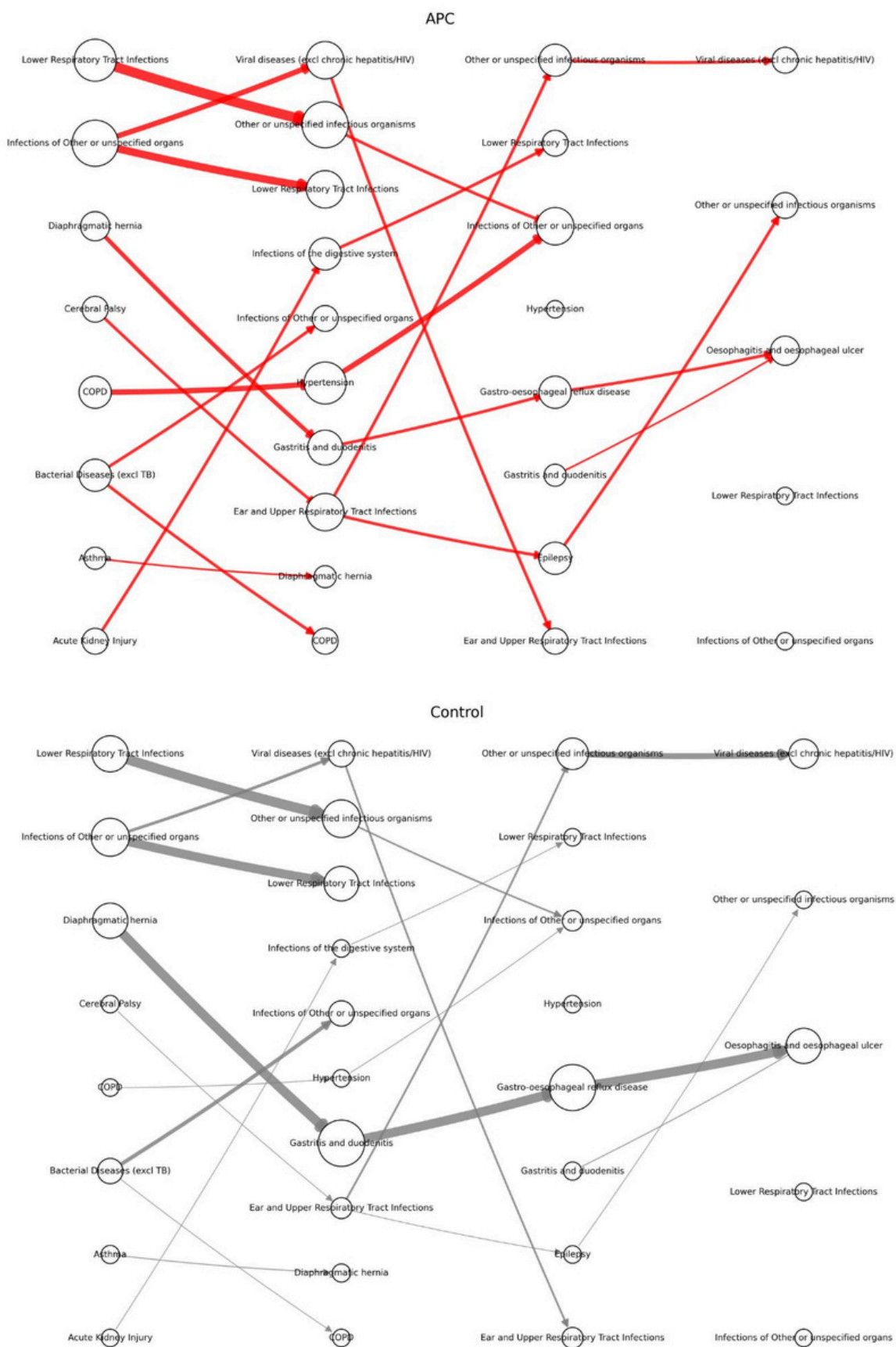
ACTA2, ACTC1, ACVRL1, APC, APOB, ATP7B, BAG3, BMPR1A, BRCA1, BRCA2, BTBD, CACNA1S, CALM1, CALM2, CALM3, CASQ2, COL3A1, DES, DSC2, DSG2, DSP, ENG, FBN1, FLNC, GAA, GLA, HFE, HNF1A, KCNH2, KCNQ1, LDLR, LMNA, MAX, MEN1, MLH1, MSH2, MSH6, MUTYH, MYBPC3, MYH11, MYH7, MYL2, MYL3, NF2, OTC, PALB2, PCSK9, PKP2, PMS2, PRKAG2, PTEN, RB1, RBM20, RET, RPE65, RYR1, RYR2, SCN5A, SDHAF2, SDHB, SDHC, SDHD, SMAD3, SMAD4, STK11, TGFB1, TGFB2, TMEM127, TMEM43, TNNC1, TNNT3, TNNT2, TP53, TPM1, TRDN, TSC1, TSC2, TTN, TTR, VHL, WT1

3. List of libraries used.

Table1 package, Pandas, Numpy, Datetime, Seaborn, Matplotlib, scipy.stats, fisher exact, and Networkx.

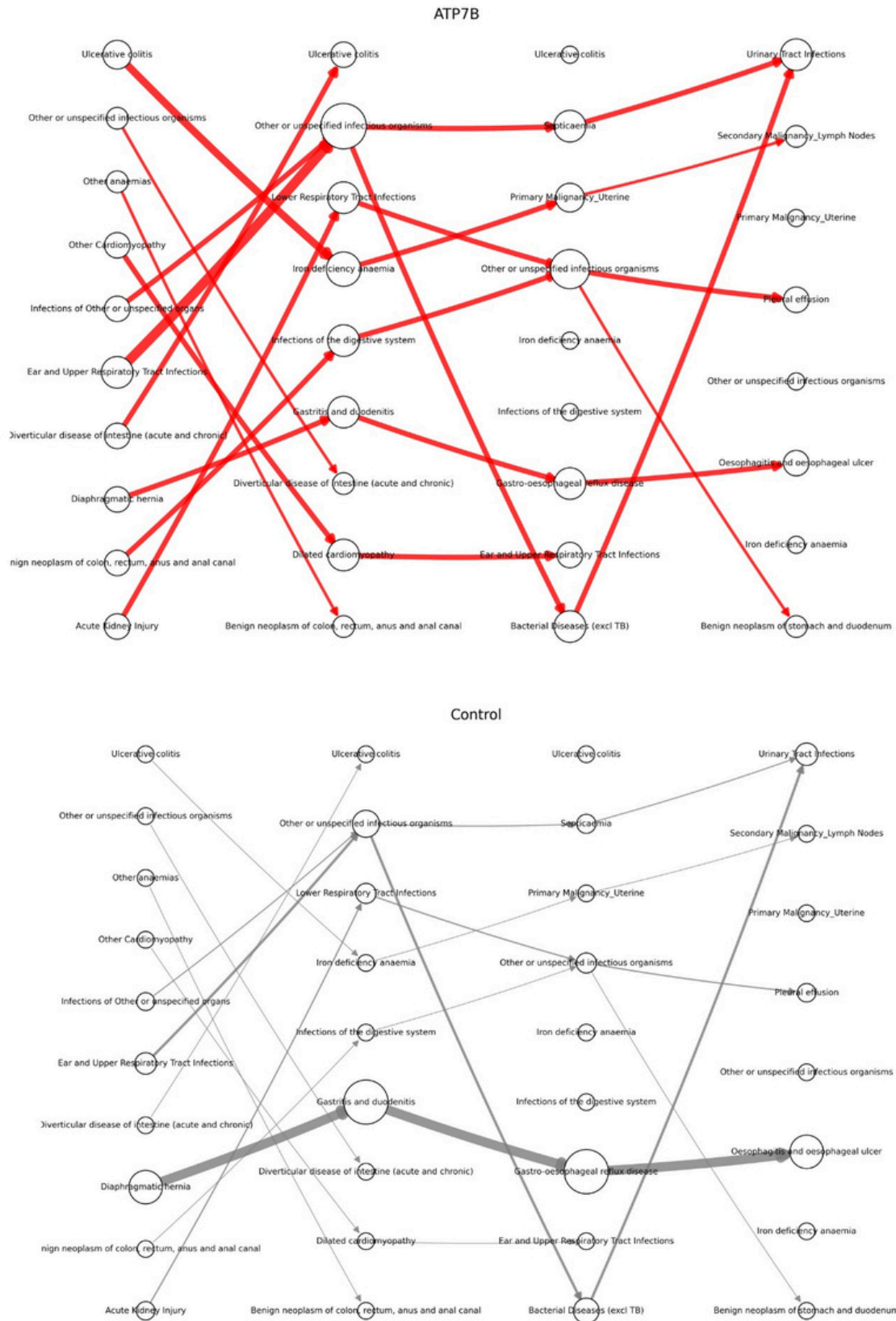
4. Trajectory plots for all the genes except *RYR1*

Disease Network Graph: APC vs Controls — top 10 sequences



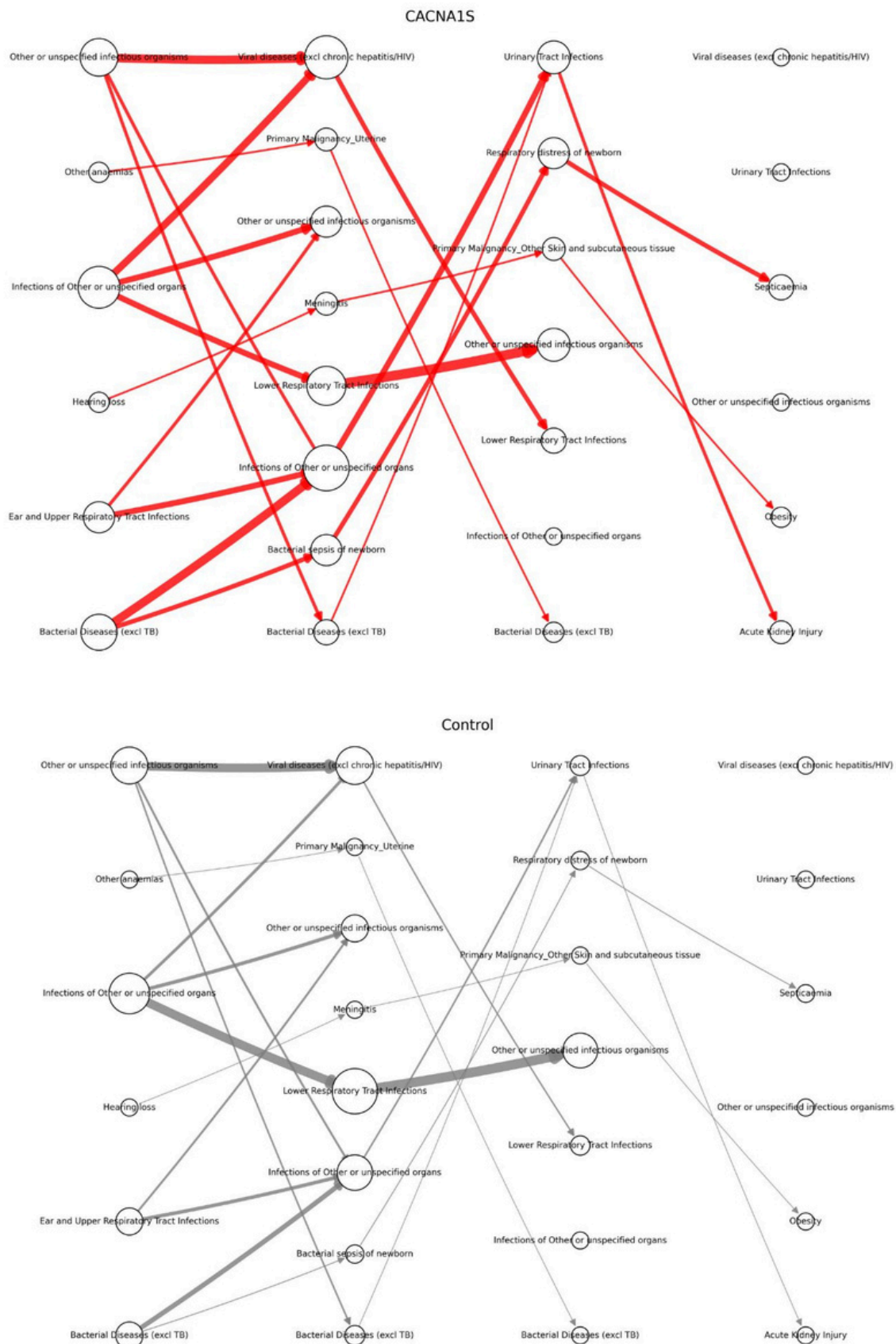
Appendix Figure 4.1: Disease trajectory network plot for carriers of gene *APC*

Disease Network Graph: ATP7B vs Controls — top 10 sequences



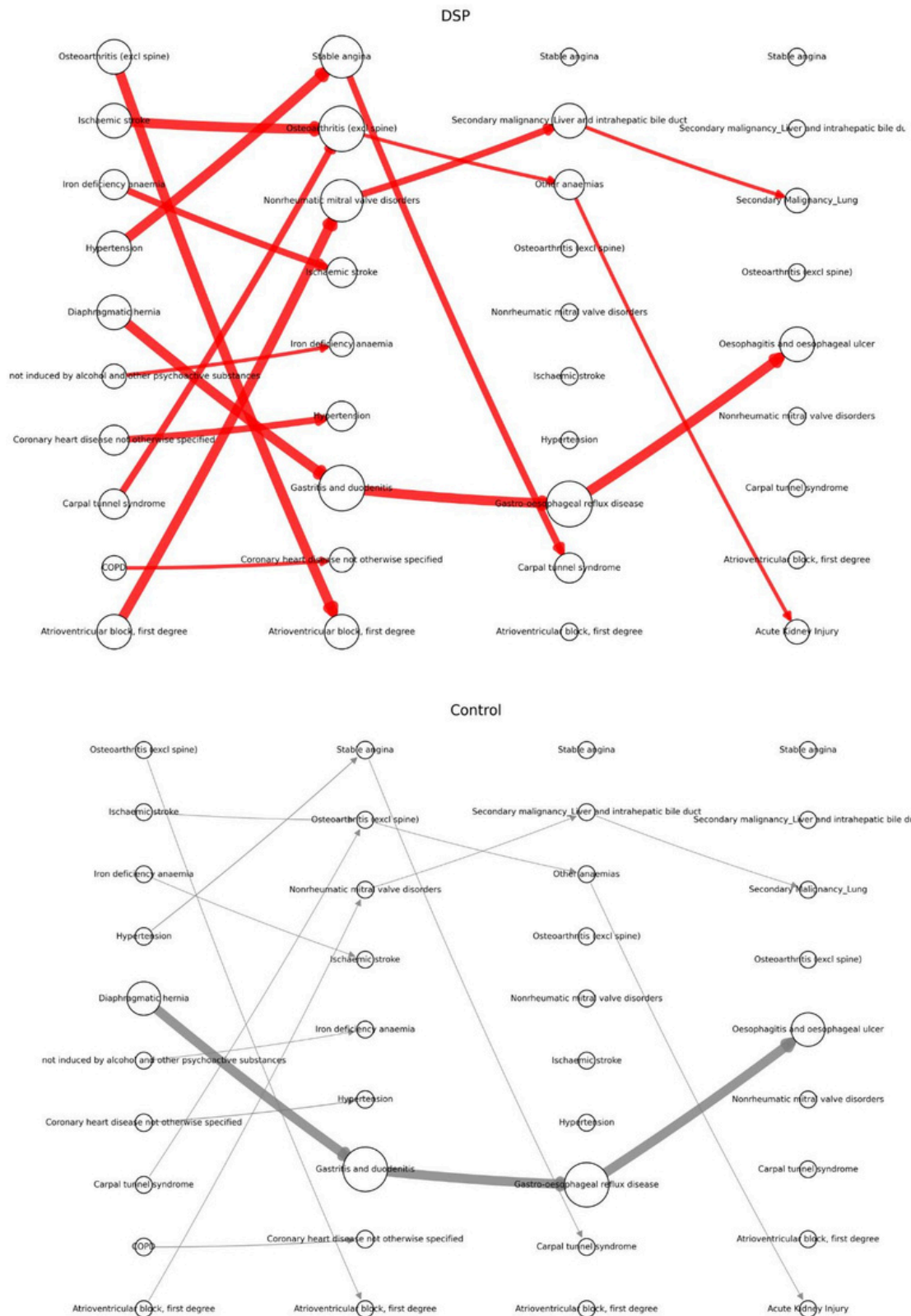
Appendix Figure 4.2: Disease trajectory network plot for carriers of gene *ATP7B*

Disease Network Graph: CACNA1S vs Controls — top 10 sequences



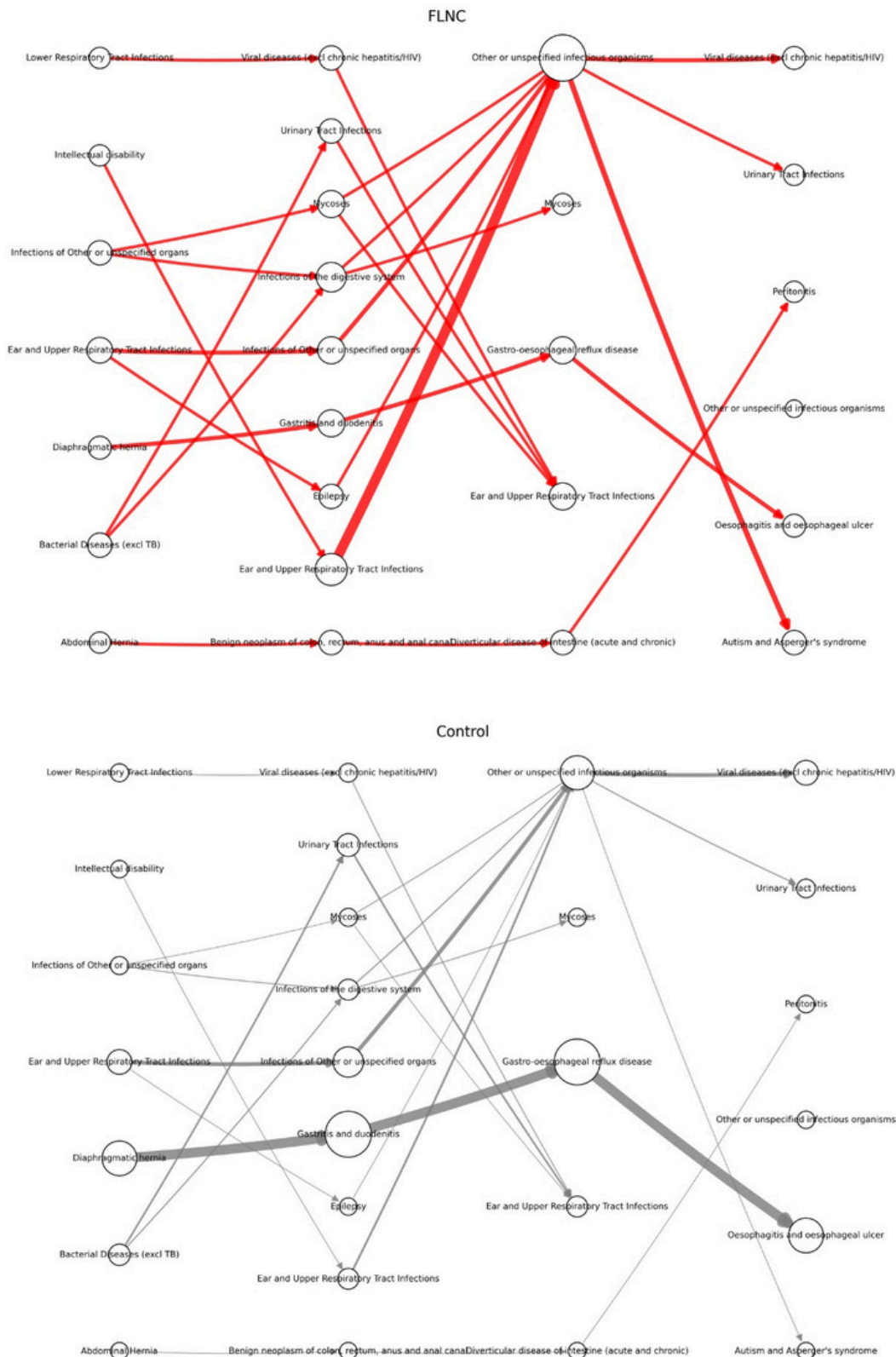
Appendix Figure 4.3: Disease trajectory network plot for carriers of gene CACNA1S

Disease Network Graph: DSP vs Controls — top 10 sequences



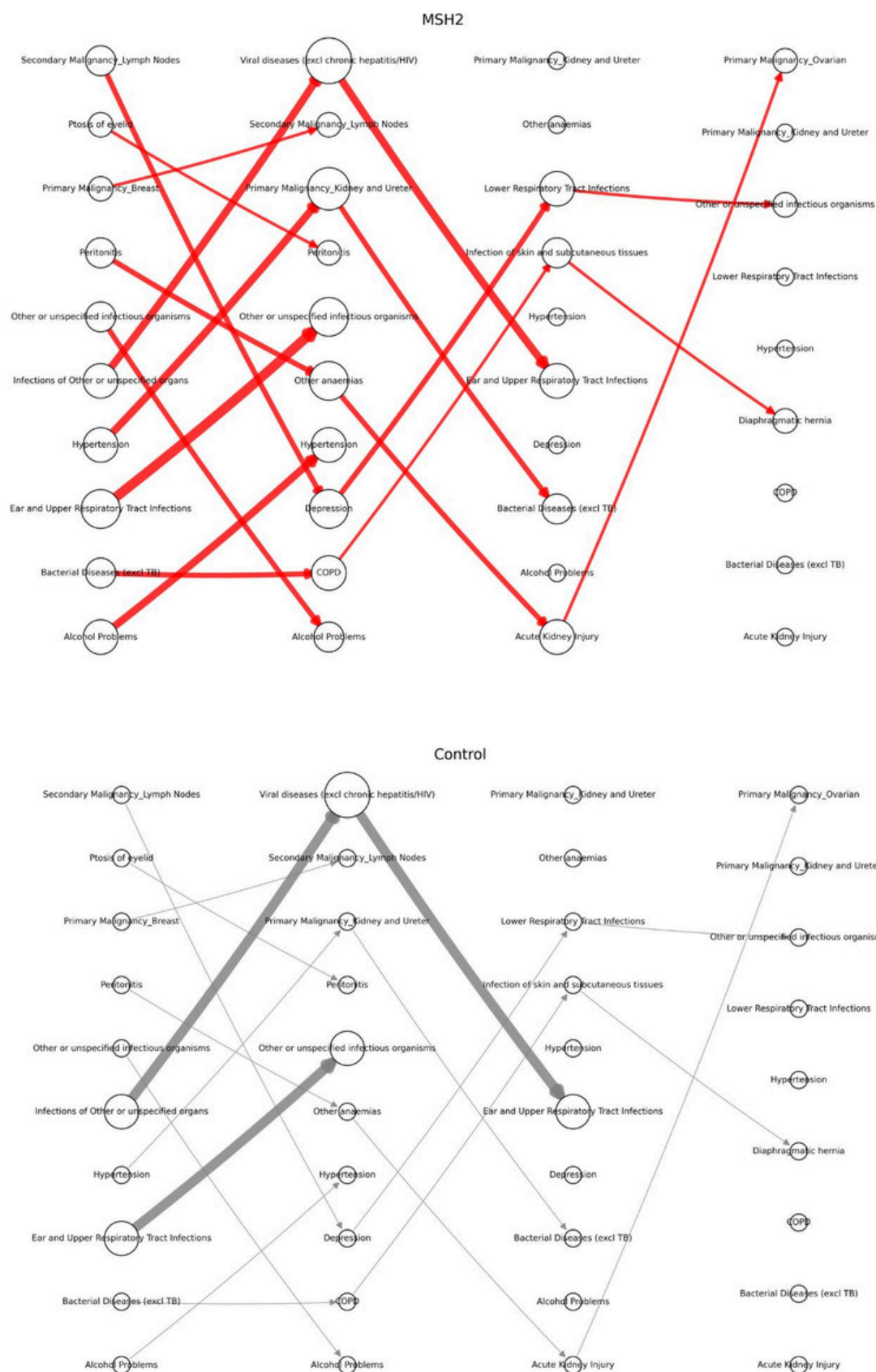
Appendix Figure 4.4: Disease trajectory network plot for carriers of gene *DSP*

Disease Network Graph: FLNC vs Controls — top 10 sequences



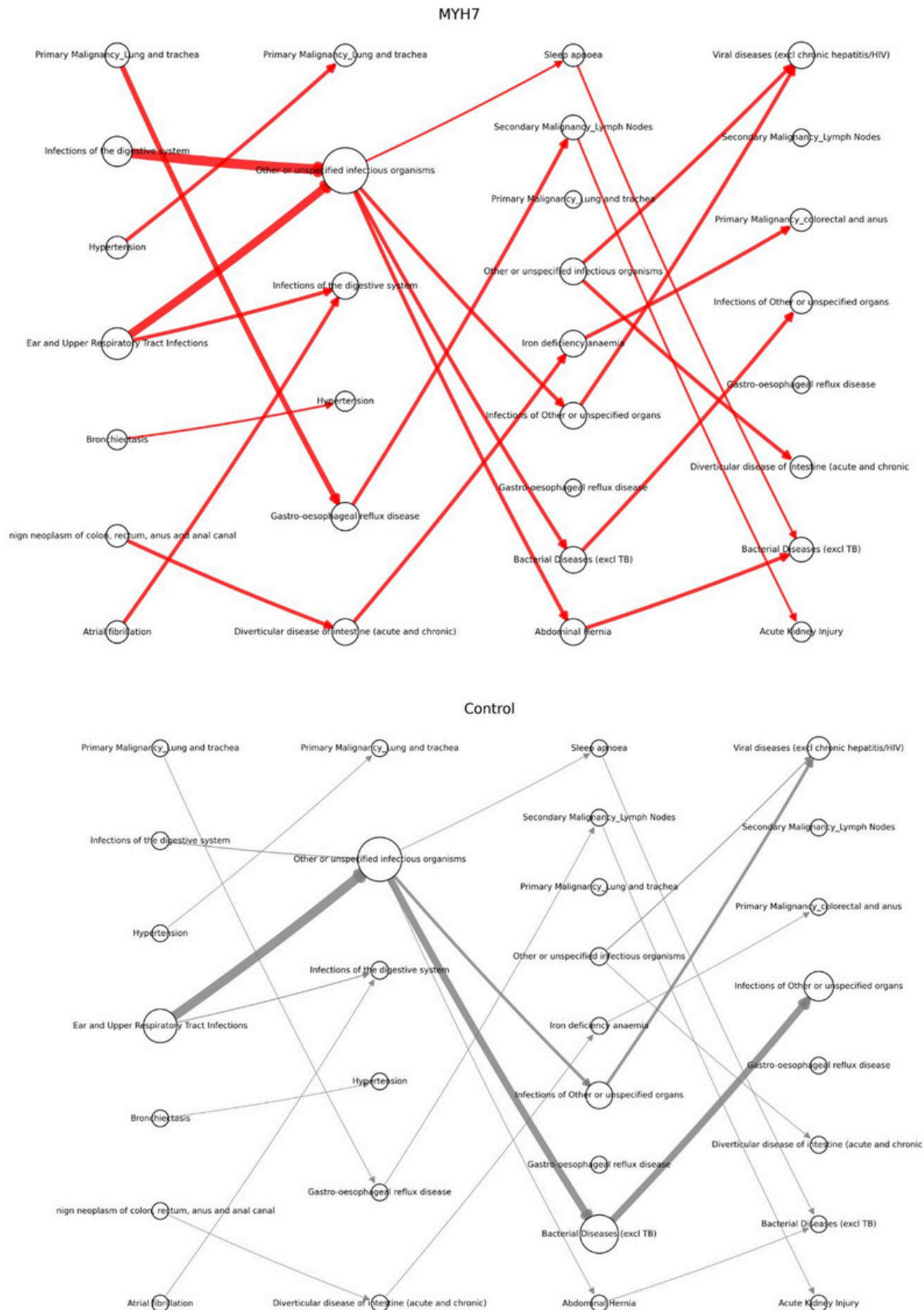
Appendix Figure 4.5: Disease trajectory network plot for carriers of gene FLNC

Disease Network Graph: MSH2 vs Controls — top 10 sequences



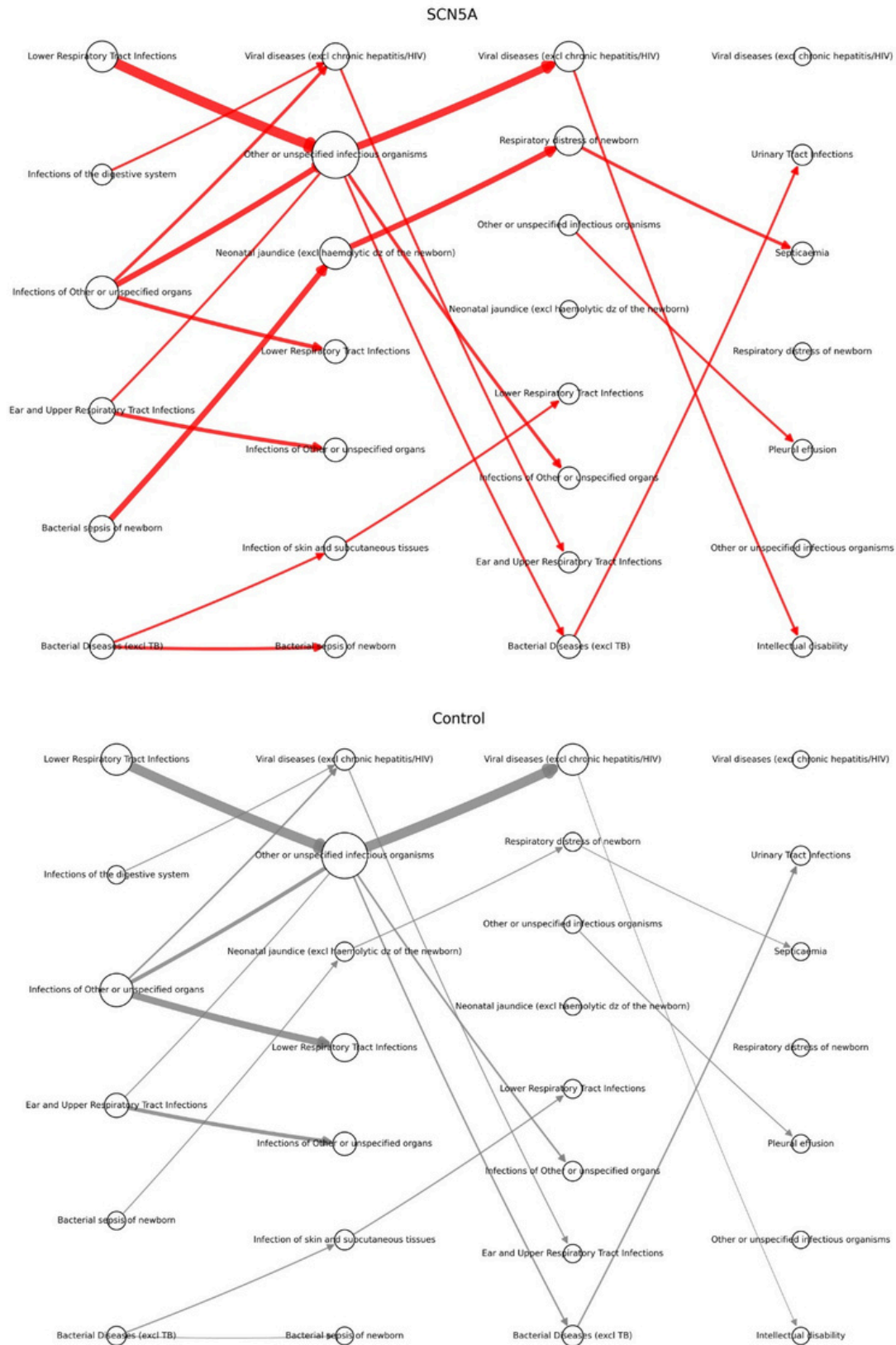
Appendix Figure 4.6: Disease trajectory network plot for carriers of gene *MSH2*

Disease Network Graph: MYH7 vs Controls — top 10 sequences



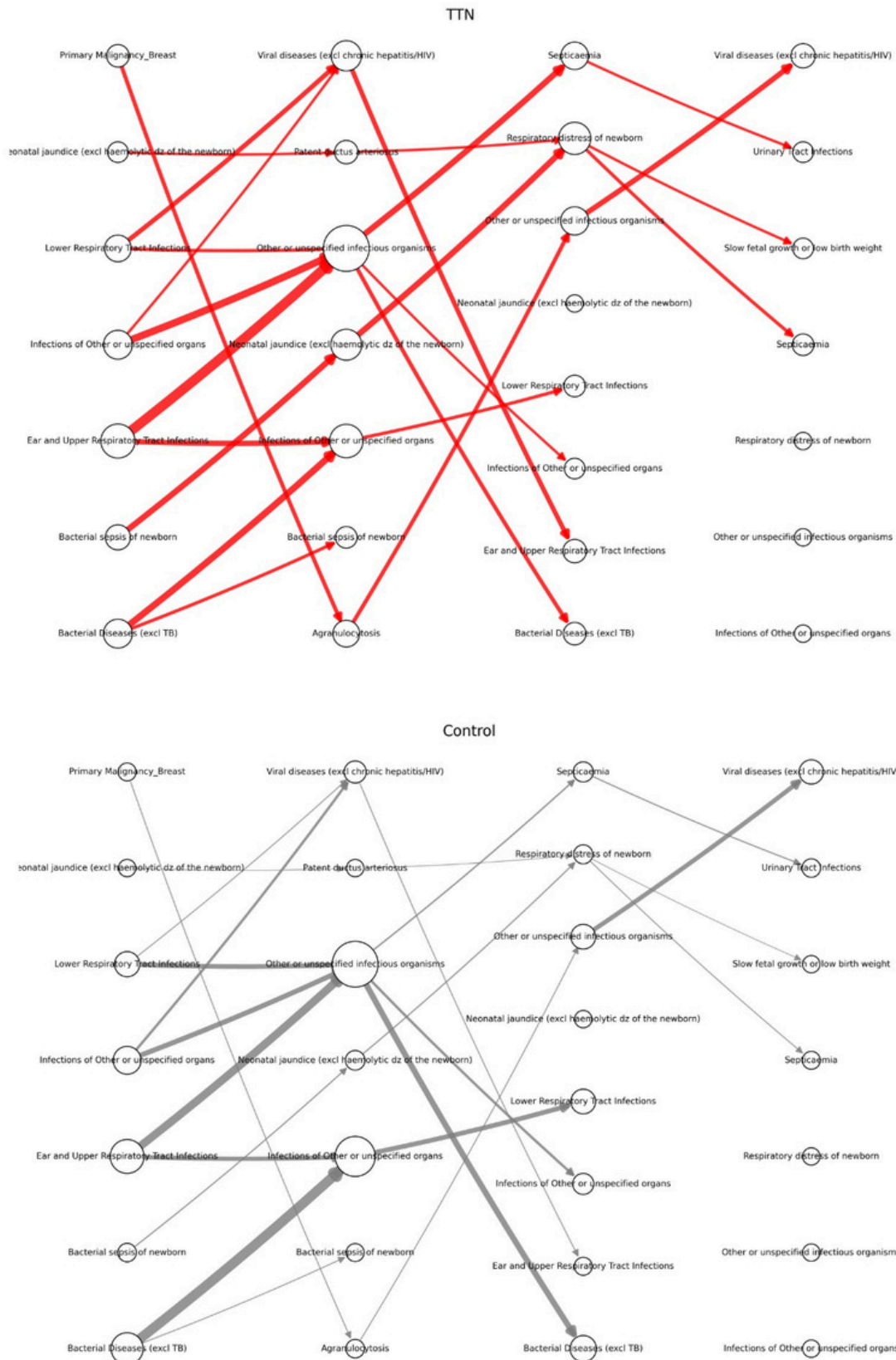
Appendix Figure 4.7: Disease trajectory network plot for carriers of gene MYH7

Disease Network Graph: SCN5A vs Controls — top 10 sequences



Appendix Figure 4.8: Disease trajectory network plot for carriers of gene SCN5A

Disease Network Graph: TTN vs Controls — top 10 sequences



Appendix Figure 4.9: Disease trajectory network plot for carriers of gene *TTN*