

Marit Kveberg Skinderhaug

A Polygenic Risk Score Analysis for Cardiovascular Diseases using a PheWAS Network

Masteroppgave i Industriell Kjemi og Bioteknologi

Veileder: Eivind Almaas

Medveileder: Martina Hall

Juni 2021

Marit Kveberg Skinderhaug

A Polygenic Risk Score Analysis for Cardiovascular Diseases using a PheWAS Network

Masteroppgave i Industriell Kjemi og Bioteknologi
Veileder: Eivind Almaas
Medveileder: Martina Hall
Juni 2021

Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for bioteknologi og matvitenskap

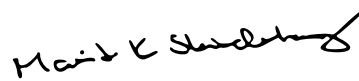


NTNU

Kunnskap for en bedre verden

I hereby declare that the work done in this thesis is independent and in accordance with the exam regulations of the Norwegian University of Science and Technology.

Trondheim, June 7th, 2021

A handwritten signature in black ink, appearing to read "Marit K Skinderhaug", written over a horizontal line.

Marit Kveberg Skinderhaug

Preface

This master thesis has been carried out at the Department of Biotechnology and Food Science in The Norwegian University of Science and Technology in the spring of 2021. It has been supervised by Professor Eivind Almaas and PhD candidate Martina Hall.

I would like to thank Professor Eivind Almaas for all encouraging counsel and help during this time. I thank PhD candidate Martina Hall for always having the time and patience to answer my countless questions and for guiding me along the way when I needed it. I would also like to send a special thanks to the CVDPgX lab at the K. G. Jebsen Center for Genetic Epidemiology, for allowing me access to the HUNT Cloud.

I would like to thank my family for all support during these past five years, and lastly, my friends and fellow students for making the last five years truly unforgettable and joyful.

Abstract

The scientific fields of genome-wide association studies and network theory have experienced considerable development over the past years, and have generated and contributed to important discoveries and new technologies. In this master thesis, the use of network theory and genome-wide association studies have been combined to perform a polygenic risk score analysis, and to construct a network using the PheWeb dataset from the UK Biobank. This network, the gene-phenotype-phenotype network, was constructed in order to compare it with the human disease network. The human disease network was presented in an article published in 2007, and shows diseases connected through mutations in common genes. This network demonstrated that a large number of diseases have a common genetic origin. The same type of network was therefore constructed for this thesis, but here, phenotypes are connected when associated with single nucleotide polymorphisms (SNPs) in common genes. The comparison showed that the two networks have few diseases and connections in common, however, there are certain similarities in the clustering pattern of cancers.

The polygenic risk scores for participants of The Trøndelag Health Study were calculated for the cardiovascular diseases angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension. The hypothesis was that using a larger number of SNPs in the calculations, and using the SNP-phenotype network to determine which SNPs to include, would improve the prediction accuracy of the polygenic risk scores. The SNP-phenotype network shows connections between SNPs and phenotypes, and was also constructed using the PheWeb dataset. The results for angina pectoris and essential hypertension showed an improvement in prediction accuracy when a larger number of SNPs were included in the calculations, and were thus most highly correlated with the initial hypothesis. For these diseases, the estimated odds ratio of developing the disease with a score in the top percentiles of the distribution, were higher when a larger number of SNPs was included in the calculations. However, the majority of the estimations were unreliable due to high p-values. Even though there was an improvement when a larger number of SNPs was included for some of the diseases, the overall disease prediction accuracy of the polygenic risk score was lower than expected.

Sammendrag

Feltene genome-wide association studies og nettverksteori har hatt stor utvikling over de siste årene, og har bidratt til viktige oppdagelser og teknologier. I denne masteroppgaven har bruk av nettverksteori og genome-wide association studies blitt kombinert for å gjennomføre en polygenic risk score analyse, i tillegg til å lage et nettverk ved bruk av PheWeb-datasettet fra UK Biobank. Dette nettverket, gen-fenotype-fenotype nettverket, ble konstruert for å sammenligne det med the human disease network. The human disease network ble presentert i en artikkel publisert i 2007, og viser sykdommer som er forbundet hvis de er assosiert med mutasjoner i felles gener. Dette nettverket demonstrerte at et stort antall sykdommer har et felles genetisk opphav. Et slikt nettverk ble derfor konstruert for denne masteroppgaven også, men her er fenotyper forbundet hvis de er assosiert med enkelt nukleotidpolymorfi (SNPs) i de samme genene. Sammenligningen viste at de to nettverkene har få sykdommer og forbindelser til felles, men at det er visse likheter i grupperingsmønsteret til noen krefttyper.

Polygenic risk scores for deltagere av Helseundersøkelsen i Trøndelag ble regnet ut for hjerte- og karsykdommene angina pectoris, hjerteinfarkt, koronar aterosklerose og essensiell hypertensjon. Hypotesen var at å bruke et større antall SNPs i kalkulasjonene, og å bruke SNP-fenotype nettverket til å avgjøre hvilke SNPs som skulle inkluderes, ville forbedre hvor nøyaktig polygenic risk scores kunne predikere sykdom. SNP-fenotype nettverket viser forbindelser mellom SNPs og fenotyper, og ble også konstruert ved bruk av PheWeb-datasettet. Resultatene for angina pectoris og essensiell hypertensjon viste en forbedring i sykdomsprediksjon når flere SNP-er ble brukt i kalkulasjonene, og korrelerte derfor mest med de forventede resultatene. Estimert odds ratio for å få disse sykdommene med en PRS i de øverste persentilene av PRS distribusjonen, ble også høyere når et større antall SNP-er ble inkludert. Men, majoriteten av estimerte odds ratios var upålitelige på grunn av høye p-verdier. I denne analysen varierer evnen polygenic risk scores har til å forutse tilfeller av de fire ulike sykdommene, men generelt er nøyaktigheten av sykdomsprediksjonen lavere enn forventet.

Contents

Preface	ii
Abstract	iii
Sammendrag	v
Table of Contents	ix
List of Tables	xi
List of Figures	xiv
List of Abbreviations	xv
1 Introduction	1
2 Theory	3
2.1 Genetic Material	3
2.2 Single Nucleotide Polymorphisms (SNPs)	4
2.3 Genome-Wide Association Studies (GWAS)	5
2.3.1 History of GWAS	6
2.3.2 Statistical Model	6
2.3.3 Optimisation and Challenges	10
2.3.4 Quality Control of Genetic Data	11
2.3.5 Quality Control of Results	14
2.3.6 Example: Schizophrenia	16
2.3.7 Conclusion	16
2.4 Networks	16
2.4.1 Properties	17
2.4.2 Types of Networks	19
2.4.3 Clustering Algorithm	20
2.5 The Human Disease Network	21
2.5.1 The Diseasome	21
2.5.2 Functional Modules	24

2.5.3	The Role of Cellular Networks in Human Diseases	25
2.6	Polygenic Risk Score (PRS)	26
2.6.1	PRS Calculation	26
2.6.2	Requirements and Considerations	26
2.6.3	Quality Control	27
2.6.4	PRS Analysis for Coronary Artery Disease	28
2.7	The HUNT Study	29
2.8	Diseases	31
2.8.1	Angina Pectoris	31
2.8.2	Myocardial Infarction	31
2.8.3	Coronary Atherosclerosis	32
2.8.4	Essential Hypertension	32
3	Methods	33
3.1	The UK Biobank Dataset	33
3.2	The SNP-Phenotype- and Phenotype-Phenotype Network	34
3.3	The Gene-Phenotype Network	35
3.4	HUNT	35
3.5	Polygenic Risk Score Analysis	37
3.5.1	Polygenic Risk Score Calculations	37
3.5.2	Visualization of Results	40
4	Results and Analysis	41
4.1	SNP-Phenotype Network	41
4.2	Gene-Phenotype and Phenotype-Phenotype Network	42
4.3	Comparison of Networks	45
4.4	Polygenic Risk Score Analysis	46
4.4.1	Prevalence- and Case-Control Plots	46
4.4.2	Cumulative Disease Risk Plots	50
4.4.3	Statistical Analysis	53
5	Discussion	57
5.1	Comparison of Networks	57
5.2	Polygenic Risk Score Analysis	58
5.2.1	Prevalence Plots	59
5.2.2	Case-Control Plots	60
5.2.3	Cumulative Disease Risk Plots	60
5.2.4	Statistical Analysis	62
5.2.5	Comparison of Results	64
5.2.6	Sources of Error	65
6	Conclusion and Outlook	67
6.1	Conclusion	67
6.2	Outlook	69
	References	77
	A Appendix	I

A.1 SNP-Phenotype Network	I
A.2 The Human Disease Network	III
A.3 Data Availability	V

List of Tables

2.3.1 Possible hypothesis testing errors.	9
2.3.2 Possible mating types in a randomly mating population.	12
3.1.1 Characteristics of the PheWeb dataset from the UK Biobank and the phenotype dataset from the Lee Lab.	34
3.4.1 The properties of the registers on the HUNT Cloud used in this thesis.	36
3.4.2 The number of cases and controls for each disease.	36
3.5.1 The number of SNPs used in the calculations.	37
4.4.1 The odds ratios for developing angina pectoris with polygenic risk scores in the highest percentiles of the distribution.	54
4.4.2 The odds ratios for developing myocardial infarction with polygenic risk scores in the highest percentiles of the distribution.	55
4.4.3 The odds ratios for developing coronary atherosclerosis with polygenic risk scores in the highest percentiles of the distribution.	55
4.4.4 The odds ratios for developing essential hypertension with polygenic risk scores in the highest percentiles of the distribution.	56

List of Figures

2.1.1 The structure of double-stranded DNA.	4
2.2.1 The general structure of a gene.	5
2.3.1 Illustration of how genetic variants can differ between case- and control individuals ^[1]	7
2.3.2 An example of quantile-quantile plots.	15
2.3.3 An example of a Manhattan plot.	15
2.4.1 An illustration of an undirected and directed network with their corresponding adjacency matrix.	19
2.4.2 An illustration of the Poisson distribution compared to a power law distribution.	20
2.5.1 The disease network.	22
2.5.2 The human disease network and the disease gene network.	23
3.5.1 An illustration of how first degree SNPs and LD blocks were chosen for the PRS calculations	38
3.5.2 An illustration of how first- and second degree SNPs and LD blocks were chosen for the PRS calculations.	39
4.1.1 A fraction of the SNP-phenotype network, displaying the diseases for which the PRS calculations were performed.	42
4.2.1 The gene-phenotype-phenotype network (GPPN) where phenotypes are connected through SNPs in common genes.	44
4.4.1 The prevalence plots for angina pectoris.	47
4.4.2 The case-control plots for angina pectoris.	47
4.4.3 The prevalence plots for myocardial infarction.	48
4.4.4 The case-control plots for myocardial infarction.	48
4.4.5 The prevalence plots for coronary atherosclerosis.	49
4.4.6 The case-control plots for coronary atherosclerosis.	49
4.4.7 The prevalence plots for essential hypertension.	50
4.4.8 The case-control plots for essential hypertension.	50
4.4.9 Cumulative disease risks over a lifetime for angina pectoris.	51
4.4.10 Cumulative disease risks over a lifetime for myocardial infarction.	52
4.4.11 Cumulative disease risks over a lifetime for coronary atherosclerosis.	52
4.4.12 Cumulative disease risks over a lifetime for essential hypertension.	53

A.1.1 The complete SNP-phenotype network.	II
A.2.1 The complete human disease network.	IV

List of Abbreviations

AP Angina Pectoris

CA Coronary Atherosclerosis

CDR Cumulative Disease Risk

CI Confidence Interval

COD Cause of Death

CVD Cardiovascular Disease

dbSNP SNP Database Number

DGN Disease Gene Network

DNA Deoxynucleic Acid

EH Essential Hypertension

EHR Electronic Health Record

GO Gene Ontology

GPN Gene-Phenotype Network

GPPN Gene Phenotype-Phenotype Network

GWAS Genome-Wide Association Study

HDN Human Disease Network

HGP Human Genome Project

HNT Helse Nord-Trøndelag

HUNT The Trøndelag Health Study

HUNT1 The HUNT1 Survey (1984-1986)

HUNT2 The HUNT2 Survey (1995-1997)

HUNT3 The HUNT3 Survey (2006-2008)

HUNT4 The HUNT4 Survey (2018-2019)

HWL Hardy-Weinberg Law

IBD Identity By Descent

ICD International Classification of Diseases

ICD10 The International Classification of Diseases, Tenth Revision

ICD9 The International Classification of Diseases, Ninth Revision

KUHR Norway Control and Payment of Health Reimbursement

LASSO Least Absolute Shrinkage and Selection Operator

LD Linkage Disequilibrium

MAF Minor Allele Frequency

MI Myocardial Infarction

MoBa Norwegian Mother, Father and Child Cohort Study

OMIM Online Mendelian Inheritance in Man

OR Odds Ratio

PCC Pearson Correlation Coefficient

PGC Psychiatric Genomic Consortium

PheWAS Phenome-Wide Association Study

PPN Phenotype-Phenotype Network

PRS Polygenic Risk Score

Q-Q plot Quantile-Quantile Plot

QC Quality Control

RNA Ribonucleic Acid

SAIGE Scalable and Accurate Implementation of GEneralized mixed model

SNP Single Nucleotide Polymorphism

SPN SNP-Phenotype Network

UTR Untranslated Region

Introduction

After the discovery of the DNA structure in the 1953, further advancements in comprehending the genome and its coding regions were made [23]. This led to the formulation of the central dogma of biology, which states that genetic information is transferred from DNA to RNA, and from RNA to proteins [2]. Mendelian genetics, established in the middle of the 1800s, had made it clear that when diseases appeared in anticipated patterns within families, this was caused by mutations in a single gene [24]. The first of these genetic mutations to be detected was the mutation that causes Huntington's disease [25]. This discovery was made in 1983, and since then, the causal variants of a series of Mendelian disorders have been discovered, among them cystic fibrosis and sickle cell anemia [26].

However, some of the most common diseases in today's society, such as cardiovascular diseases, Alzheimer's disease, diabetes type 2 and cancer, are complex traits [2]. This means that they are caused by the additive effect of numerous genetic variants, in addition to being influenced by environmental risk factors [2]. During the past two decades, the use of genome-wide association studies (GWAS) has led to the discovery of genetic variants, or single nucleotide polymorphisms (SNPs), associated with an increased risk for a series of common, complex disorders [7]. The first GWAS was published in 2002, and presented the finding of a specific chromosome position, a chromosomal locus, associated with myocardial infarction [8]. Since then, GWAS have been performed for diseases such as coronary artery disease, diabetes type 2 and schizophrenia [9][10][11].

The development of GWAS is partly attributable to the technological advancements over the past decades [12]. This includes the development of algorithms to perform and improve the effectivity of GWAS, but also an increase in the amount of publicly available GWAS summary statistics and resources that link electronic health records with genotype data [12]. The latter is highly advantageous for acquiring a more thorough understanding of the link between genotypes and traits [12]. An example of such a resource is the The Trøndelag Health Study (HUNT) databank, which contains information from questionnaires, clinical studies and blood analyses, in addition to being linked to various Norwegian registers [13].

GWAS can be used to detect genetic variants that are more frequently found in the genome of diseased individuals, and which chromosomal loci these are associated with. However, each single variant contributes only a small amount to the overall disease risk [14][7]. Even though GWAS

have contributed to the discovery of a series of candidate chromosomal loci, there are still a significant amount of disease heritability that cannot be explained by these variants^[15]. GWAS are usually performed by using known, common genetic variants to scan the genome. However, there most likely exist numerous common and rare variants, which have both smaller and larger effects on the occurrence of various diseases, but which have not yet been discovered^[15].

Even though GWAS have not caused as much progress in the field of disease prediction as initially believed, the resulting discoveries of genetic variants are useful. When a statistically significant association between a genetic variant and disease has been found, and the chromosomal locus of this variant has been located, other methods are required to investigate this finding further. The calculation of polygenic risk scores (PRSs) is one such method that is used for risk prediction^[16]. PRS calculations require the knowledge of which variants are associated with the disease and their effect size^[16]. The effect size is a measure of how much that genetic variant contributes to the risk of acquiring the disease, and is obtained from GWAS summary statistics^[16].

Another field that has had a significant development over the past decades, is network theory. Networks are found within all systems, from cells to societies, and at the beginning of the 21st century it was discovered that even though networks are found across all fields, they are based on the same fundamental laws and principles^[17]. For the specialization project in TBT4500, the SNP-phenotype network (SPN) was constructed based on the PheWeb dataset from the UK Biobank, where SNPs and phenotypes are connected if an association has been found between them. For the first part of this thesis, a network connecting phenotypes through SNPs in common genes was constructed. This was done such that a comparison could be made between this and the human disease network (HDN), to detect any similarities in connections or clustering patterns^[18]. For the second part, the use of networks was combined with PRSs to predict disease risk. The SPN was utilized to determine which SNPs to include in the PRS calculations. Using a network approach to perform a PRS analysis is not known to have been done previously.

The PRS analysis was performed for the cardiovascular diseases angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension. The PRSs are calculated using GWAS summary statistics from the UK Biobank and information regarding HUNT participants obtained from the HUNT databank. The HUNT participants included in the PRS analysis for a particular disease were chosen based on the presence of certain SNPs in their genome. The SNPs to be included were determined using two different procedures. First, only SNPs associated with the disease were included in the calculations. These are the SNPs directly connected to the disease in the SPN. In the second procedure, both SNPs directly linked to the disease, and the SNPs linked to its neighbouring diseases in the SPN, were included. **The main hypothesis for this thesis is that using a larger number of SNPs in the PRS calculations, and using the SNP-phenotype network to determine which SNPs to include, increases the disease prediction accuracy of the PRS.**

Chapter 2

Theory

This chapter contains the background theory for this master thesis. The topics covered are genetic material, single nucleotide polymorphisms (SNPs), genome-wide association studies (GWAS), network theory, a summary of the human disease network article, polygenic risk scores (PRS), the HUNT Study and lastly, a section regarding the diseases considered in the PRS analysis. Section 2.1, 2.2, 2.3, 2.4 and 2.5 are taken from the specialization project in TBT4500 delivered the previous semester, and these sections are clearly marked.

2.1 Genetic Material

Section 2.1 is taken from the specialization project in TBT4500^[19]. The genetic material of all living organisms is deoxyribonucleic acid (DNA)^[20]. DNA consists of a double helix composed of two anti-parallel nucleotide strands, as illustrated in Figure 2.1.1. Each strand has a backbone consisting of alternating deoxyribose- and phosphate units. In addition to this, there are four different nitrogen bases; adenine (A), thymine (T), cytosine (C) and guanine (G)^[20]. Each nucleotide consists of a nitrogen base, a deoxyribose molecule and a phosphate group. The two strands are complementary, and thus the nucleotides on the opposite strands of the double helix can base-pair with each other through hydrogen bonds. Adenine forms two hydrogen bonds with thymine, and guanine forms three hydrogen bonds with cytosine. All genetic material of the organism is organized into 23 chromosomes, where number 23 is one of the sex chromosomes, X or Y^[20].

Genes are sequences of DNA with lengths from a couple of hundred up to more than two million base pairs^[21]. All human beings share 99 % of their DNA, while 1 % varies^[22]. This 1 % is what causes the genetic diversity among individuals. Genetic variability can be observed at specific DNA loci. Diploid cells contain two sets of homologous chromosomes, where one is inherited from each parent^[4]. The homologous chromosomes consist of the same genes, but there are some genetic variations. Different variants of a particular gene are called alleles. If the alleles on both homologous chromosomes are equal, that particular allele is homozygous, while if the alleles differ, the allele is heterozygous^[4].

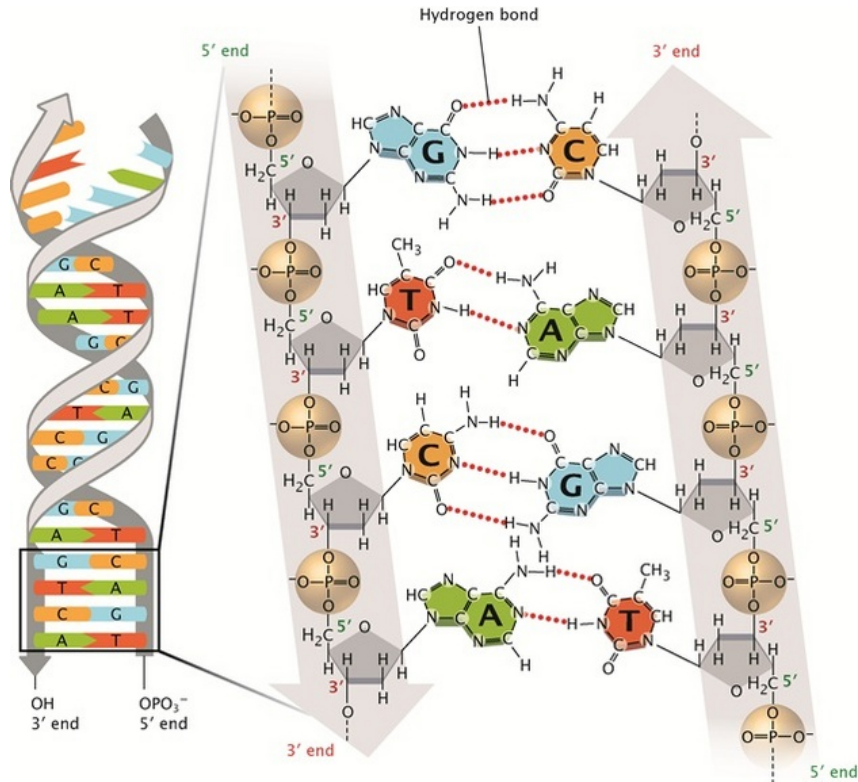


Figure 2.1.1: Illustration of the structure of double-stranded DNA. The nucleotide bases on each strand are complementary, and can base pair with each other through hydrogen bonds. Each nucleotide consists of a deoxyribose-, nitrogen- and a phosphate group. Figure from Nature Education [3].

2.2 Single Nucleotide Polymorphisms (SNPs)

Section 2.2 is taken from the specialization project in TBT4500 [19]. Single Nucleotide Polymorphisms (SNPs) are the most frequent genetic variation found within the genome of human beings [23]. They arise from the exchange of one single nucleotide base within the DNA sequence. An example would be the exchange of G for A in the top base pair in Figure 2.1.1. Diallelic SNPs are mostly the case in humans, even though four different alleles are theoretically possible when a nucleotide base is exchanged [20]. It is estimated that SNPs are found in 1 out of every 1,000 nucleotides, which means that they are quite frequent throughout the human genome [23]. For the polymorphism to be defined as a SNP, its minor allele frequency (MAF) must be at least 1 % throughout the population [20].

SNPs can be located in exons, introns, promoters and also 5'- and 3' untranslated regions (UTRs) [24]. The organization of these elements within a gene is illustrated in Figure 2.2.1. Exons are the DNA-sequences within the gene which encode proteins [4]. SNPs located here can cause repression of transcription and thereby translation of certain proteins [24]. The functional mechanisms of these proteins within the cell determines the consequences of the SNP, and in certain cases it may be a contributor to cancer or other serious diseases.

Introns are sequences within a gene that do not encode proteins [4]. SNPs located in these regions can cause different splicing variants of the gene [24]. Splicing is the post-transcriptional mechanism of extracting the intron-sequences from the gene and attaching the exons together to acquire a protein-encoding sequence [4]. Promoters are sequences where RNA polymerase

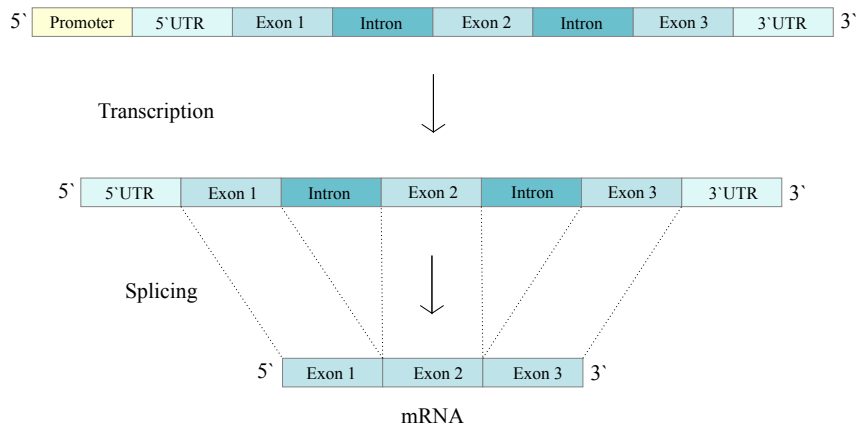


Figure 2.2.1: Illustration of the general structure of a gene, with the organization of the promoter, 5′ and 3′ untranslated regions (UTR), introns and exons. It also shows what occurs during splicing. The figure is inspired by Snustad and Simmons [4].

binds to initiate transcription. SNPs within this part of the sequence can cause deviations in the binding of RNA-polymerase and certain transcription factors, and it can also cause differences in the pattern of DNA methylation and histone modifications [24]. This can again cause changes in gene expression. 5′ and 3′-UTR are untranslated regions, which means that they do not encode proteins [4]. However, SNPs in these regions can still have an effect on the cell’s gene expression [24].

Some SNPs may occur in numerous individuals, while others are more rare [23]. Silent SNPs located within coding regions are called synonymous SNPs [25]. They do not change the amino acid inserted during translation, and therefore have no effect. However, some SNPs cause a change in the amino acid sequence. These are called non-synonymous SNPs and can have more serious consequences for the transcription and thus translation of the DNA [25]. The risk is that the protein encoded by the DNA cannot be produced or might be dysfunctional. The consequence of this depends on the function of the protein and also the amino acid inserted. An amino acid with the same properties as the one it replaced, will cause less dramatic consequences than a completely different one [25].

The consequences of SNPs can be observed through individuals’ tolerance and response to toxins or drugs, and also in the risk of developing certain diseases [23]. The localization of SNPs have been used to investigate the inheritance of genetic diseases within families, and advances have been made to investigate their involvement in more complex diseases [23]. These approaches include genome-wide association studies.

2.3 Genome-Wide Association Studies (GWAS)

Section 2.3 is taken from the specialization project in TBT4500 [19]. Genome-wide association studies (GWAS) evaluate the relationship between gene variant frequency and susceptibility to certain diseases or traits [26]. The DNA of several individuals is scanned to detect genetic markers that are known to cause genetic variation in a population. The target is to find associations between genotype frequencies and traits, which can be used to detect genetic susceptibility to certain diseases [26]. GWAS has strongly enhanced the understanding of the allelic architecture behind complex traits. Among the diseases where associations between genetic variants and

predisposition to disease have been found are inflammatory bowel disease, diabetes type 1 and 2, breast cancer and prostate cancer [26]. The ultimate goal is to locate the genetic variants at each locus that contribute to an individual's predisposition to a disease or trait.

2.3.1 History of GWAS

The initial, main purpose of GWAS was to better understand the genetics and biology behind diseases [27]. There was hope that a more detailed understanding would help improve the treatment of these diseases, or potentially prevent them from arising in the first place. The problem was, and still is, to understand the mechanisms through which the genetic variants work. Even though a significant association has been found between a genetic variant at a specific chromosome locus and a trait, the molecular mechanism behind this is usually unknown [27]. Over the last years, new molecular technologies and analytical methods have helped fill in information regarding these mechanisms. GWAS has also enhanced knowledge regarding the contribution of both genes and environment to disease risk [27].

GWAS has been performed for complex and common diseases, and behavioural, social, and quantitative traits that contribute to disease risk [27]. For the phenotypes studied thus far, it seems that several genetic variants at different chromosomal loci contribute to the genetic variation found within a population [27]. This means that the effect a single genetic variant has on genetic variation is quite low. Therefore, using a larger sample size has made it possible to locate more associations. This has been shown with GWAS regarding different types of cancer. Using GWAS, 45 susceptibility loci associated with lung cancer have been found [28], while 170 loci have been detected for breast cancer [29]. The increasing findings of associations are in part due to the use of larger sample sizes. In addition to this, there has been technological improvements with denser genotyping assays and also an increase in publicly available genetic information [27]. These are all factors that have contributed to the increasing amount of associations found since the first GWAS study was published almost 20 years ago [7].

As mentioned above, a large contributor to the advancement of GWAS during the past years is the increase in genetic data shared with the public [27]. The summary statistics of these datasets, such as p-values and effect sizes, are then made available. An example of such a contributor is the GWAS Catalog, which provides accessible and searchable datasets containing SNP-trait associations [30]. It was founded by the National Human Genome Research Institute (NHGRI) in 2008 because of the rising amount of published GWAS. The GWAS catalog is constructed by curators who evaluate all valid GWAS studies published to detect associations between genetic variants and traits. As of 2020, the catalog contains 4,741 published studies and 212,730 associations [30].

2.3.2 Statistical Model

GWAS is usually performed using case-control studies [26]. This involves choosing case individuals who have a higher susceptibility to the trait in question, and perform hypothesis testing to see whether they have a higher number of susceptibility alleles [26]. The control individuals are not susceptible to the trait, and are tested for a lower number of susceptibility alleles. Figure 2.3.1 illustrates the association study design, and how genetic variants can differ between cases and controls.

$$\text{logit}(p(X)) = \log\left(\frac{P(Y_i = 1|X)}{1 - P(Y_i = 1|X)}\right) = \beta_0 + \beta_1 x$$

is obtained, which is a linear function of the covariate, x [31]. This is defined as the *log odds* or *logit*. A single unit change in X will change the log odds by β_1 . However, the change in $p(X)$ will not be equivalent to β_1 , because the relationship between X and $p(X)$ is not linear [31]. The change in $p(X)$ depends on the value of X . Either way, a positive value of β_1 will provide a higher log odds when x is increased, while a negative β_1 will lower the log odds with increasing values of x .

A logistic model is used to find the effect of a particular SNP on a trait, where the dependent variable (Y) is the presence/absence of the trait and the covariate (x_1) represents whether the SNP has a value of 0, 1 or 2 [32]. This is because the SNP may be present on either none, one or both chromosomes in a homologous pair. In the equation below, a second covariate (x_2) is included, which could represent a factor such as gender. To evaluate the effect of the SNP on a specific trait, the odds ratio (OR) of the SNP, ($OR(\beta_1)$), is found by using the equation

$$OR(\beta_1) = \frac{\text{odds}(x_1 = 1)}{\text{odds}(x_1 = 0)} = \frac{e^{\beta_0 + \beta_1 \cdot 1 + \beta_2 x_2}}{e^{\beta_0 + \beta_1 \cdot 0 + \beta_2 x_2}} = \frac{e^{\beta_0 + \beta_1 + \beta_2 x_2}}{e^{\beta_0 + \beta_2 x_2}} = e^{\beta_1}.$$

If $OR(\beta_1) > 1$, this indicates that the SNP increases the risk of acquiring the trait in question [32]. If $OR(\beta_1) < 1$, this means that the absence of the SNP increases the risk for acquiring the trait. Alternatively, this can also mean that having the SNP decreases the risk for acquiring the trait. A higher value of $OR(\beta_1)$ means that the SNP has a higher effect size [32]. A one unit increase of x_1 ($0 \rightarrow 1$ or $1 \rightarrow 2$) thus indicates a larger odds for having the trait. This model is highly simplified, but the principle is the same as that used in GWAS. In GWAS, a higher number of covariates are considered. Typical examples are gender, age certain population structures and principal components [26].

The statistical test used for the hypothesis testing is the t-test. To test for the effects of the SNP on a specific trait or disease, the hypotheses are

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0,$$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis. The assumption is that the effect of the SNP on a trait, $\hat{\beta}_j$, is normally distributed with an expectation value of β_j and variance σ_j . Since a normal distribution is considered, a z-test could be performed. However, the variance is unknown, and the sample standard deviation must be used as an estimation [33]. Therefore, a t-test is used to test the hypotheses above [33]. Assuming that H_0 is true,

$$t = \frac{\hat{\beta}_j - E(\hat{\beta}_j)}{\sqrt{\hat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}}.$$

The t-statistic above is Student-t distributed with $n - p$ degrees of freedom, where n is the number of samples (in this case, individuals) and p is the number of coefficients in the model [33]. The p-value of the two sided t-test is then defined as

Decision	H_0 is true	H_0 is not true
Keep H_0	Right decision	Type II Error (β)
Reject H_0	Type I Error (α)	Right decision

Table 2.3.1: Possible hypothesis testing errors.

$$P = P(T \geq t_{obs}) + P(T \leq -t_{obs}).$$

Using a significance level of α , H_0 is rejected if the p-value is less than α . This would indicate that the hypothesis of the SNP having a zero effect size is not accurate. The p-value is defined as the probability of obtaining a value as extreme or more than that of the actual sample, when H_0 is true [34]. If the p-value is less than α , which in this case is the genome-wide significance level, H_0 is rejected. This means that there is a significant association between a chromosomal locus and a trait. The GWAS significance threshold is set to $5 \cdot 10^{-8}$, because of the large number of false positives resulting from multiple SNPs being tested at the same time [34]. A GWAS of 1 million SNPs will generate 1 million tests, which again will generate a large number of false positives. The threshold of $5 \cdot 10^{-8}$ is generally accepted for European populations, while for African populations it is set to 10^{-8} , because of a larger genetic diversity [34].

An important consideration to take when performing GWAS, is to ensure that the study has a high enough statistical power [34]. The statistical power is defined as the ability to reject the null hypothesis when it is false. In the case of GWAS, H_0 is the non-existence of an association between a chromosomal locus and a trait. This is a zero effect size result, which means that the gene variant has no effect on the phenotype [34]. The alternative hypothesis, H_1 , is to find a chromosomal locus with a non-zero effect size. The different errors that can occur because of an insufficient statistical power are summarized in Table 2.3.1.

A type II error is a false negative and is denoted $1 - \beta$ [34]. This occurs when H_0 is accepted when not true. False positives are called type I errors and are denoted α [34]. This means that H_0 is rejected when true. These are associations that appear significant, but turn out not to be when the study is replicated. Type I- and II errors can be used to define significance level and statistical power:

$$\text{Significance level : } \alpha = P(\text{Type I Error})$$

$$\text{Statistical Power : } 1 - \beta = 1 - P(\text{Type II Error})$$

The probability for type I errors, α , can be determined by the investigator through changing the threshold for accepting the null hypothesis [34]. This is not the case for β , which is affected by several different factors, including the effect size of the genetic variant and the quality of the data [34]. These are factors outside the investigator's control.

2.3.3 Optimisation and Challenges

GWAS is usually performed using case-control studies [26]. The case individuals should be susceptible to the trait under investigation. These can be difficult to find, especially when dealing with rare traits. Choosing the right control individuals is also quite challenging. The control group should not be susceptible to the trait in question, and it is important that they do not cause confounding results [26]. With a large enough sample size this is usually not a problem, even with some cases of miscategorization. Miscategorization implies the presence of individuals with a hidden diagnosis of the phenotype in the control group [26]. When performing studies on rarer traits, disturbance from the control group is less common. However, frequent traits such as obesity can create challenges when choosing the control group. An alternative here can be to choose individuals with the extreme opposite phenotype, such as extremely underweight individuals, but there is then a risk of acquiring other biases [26].

GWAS has increasingly been performed using cohort-studies over the past years [26]. Cohort-studies involves identifying a cohort of individuals that are disease-free, but have a high exposure to risk factors associated with the disease [35]. The causality of these risk factors can then be more thoroughly analyzed, because the subjects are investigated in a time period where they go from being healthy individuals to acquiring the disease. The disadvantage is that this can require long time periods, and also large sample sizes, which is a problem when analyzing rare diseases [35]. Because the diseases considered often are less common, cohort-studies usually also have low statistical power [26]. However, cohort-studies can give new insights into the joint effects of genes and environment, and can also increase knowledge about continuous traits and pleiotropy. A gene is pleiotropic when it affects several different traits at the same time. In such genes, a single mutation can cause changes to several phenotypic traits [26].

Some of the determinants for the statistical power of a study can be controlled by the investigator, while others cannot [34]. The factors which can be controlled include choice of sample subjects, the method of measurement for genotypes and phenotypes, quality analysis method and the statistical approach used [34]. Outside the investigator's control are allele frequency and effect size of the genetic variant, the genetic complexity of the trait, the stability of the phenotype, and the ancestry of the study population [34].

The genetic complexity of the trait influences the difficulty of finding associations between genetic variants and phenotypes [34]. Complex traits are determined by several genetic variants at different loci, and can also be affected by environmental factors. To find SNPs associated to complex diseases the sample size is important, and for each disease there exists a minimum threshold of size [34]. Up to a certain limit, as the sample size increases, so will the number of associations found. The complexity of each disease varies, and it depends on the number of associated SNPs. This again depends on the actual molecular mechanisms that make the disease occur, which are often unknown [34].

Conversely, Mendelian diseases are caused by a single mutation, and do not have such requirements for sample size [34]. The environment and genetic background do not have as great an impact on Mendelian diseases as with complex diseases, although, susceptibility loci can vary between different families for certain diseases. This means that a disease can be caused by mutations in different genes, which is defined as locus heterogeneity [34]. Since only one mutation is required for developing the disease, this mutation can increase the risk for acquiring a disease from 0 to 1 [34]. The SNPs causing Mendelian diseases are therefore easier to detect than what

is the case for complex diseases.

Detection of SNPs causing Mendelian diseases can be done by performing a linkage analysis on family members^[34]. Linkage analysis can be used to detect chromosome loci associated with a specific trait or disease, through the association of this locus to another one of known location^[36]. These genetic variants are inherited together due to their physical proximity. In this way, additional genetic variants associated to the disease in question can be detected, without being directly genotyped. Linkage analysis requires the genotyping of DNA from the family members to detect the presence of SNPs with known positions. However, statistical power can be reduced because of locus heterogeneity^[34]. Also, phenotypic heterogeneity may have an impact. This implies that different mutations within the same gene produce entirely different phenotypes.

2.3.4 Quality Control of Genetic Data

Before GWAS can be performed, a quality control (QC) of the genetic data is required^[20]. If this is not done adequately, the associations found in the study will be invalid. The following steps need to be executed to find the SNPs that should be excluded from the dataset; Missingness of SNPs and individuals, sex discrepancy, minor allele frequency, Hardy-Weinberg equilibrium, heterozygosity rate, relatedness and population stratification^[37].

SNPs that are missing in a large fraction of the sample subjects should be excluded from the data^[37]. This can be done by first filtering SNPs according to a percentage of missingness. This percentage is often set to 20 %, which means that SNPs missing from more than 20 % of the sample subjects, are excluded from the study^[37]. Further, this percentage is lowered to 2 %, so that a larger number of SNPs are excluded. In this way, all remaining SNPs are sufficiently present in the sample subjects. Also, individuals with a high number of missing SNPs should be removed, and this can be done using the same procedure as discussed above^[37].

The dataset must be checked for sex discrepancies between individuals based on the number of heterozygosity or homozygosity sites on X chromosomes^[37]. Males should have a homozygosity rate above 0.8, while females should have a homozygosity rate below 0.2^[37]. Males have a higher homozygosity rate because they have an X and Y chromosome, while females have two X chromosomes. For males, alleles can therefore only be heterozygous when positioned in the pseudo-autosomal region, which are homologous sequences on the X- and Y chromosome^[38]. If several sample subjects are registered as one sex but have an irregular homozygosity rate, this might indicate discrepancies in the data^[37].

The Minor Allele Frequency (MAF) defines the limit for how rare SNPs can be among the sample subjects before they are excluded from the study^[37]. A larger sample size allows for a lower MAF. This is because a larger sample size implies higher statistical power, and this compensates for the decrease in statistical power due to the low MAF. For large sample sizes of 100,000 subjects, it is usually set to 0.01, while for small populations of 10,000 it is set to 0.05^[37].

The fourth step in the QC is to check for SNPs that deviate from the Hardy-Weinberg Law (HWL)^[37]. This law states that the relation between genotype and allele frequency in a large population is stable across generations^[20]. The HWL applies under the following assumptions: An infinitely large population size, random mating, no migration and no mutations or population

Mating		Offspring Genotype			
Father	Mother	Frequency	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	p_{11}^2	1	0	0
	A_1A_2	$p_{11}p_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0
	A_2A_2	$p_{11}p_{22}$	0	1	0
A_1A_2	A_1A_1	$p_{11}p_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0
	A_1A_2	p_{12}^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	A_2A_2	$p_{12}p_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
A_2A_2	A_1A_1	$p_{11}p_{22}$	0	1	0
	A_1A_2	$p_{12}p_{22}$	0	$\frac{1}{2}$	$\frac{1}{2}$
	A_2A_2	p_{22}^2	0	0	1

Table 2.3.2: Possible mating types in a randomly mating population, regarding a diallelic, autosomal chromosome locus. Each genotype is listed with its frequency, and their offspring's genotype probabilities [20].

stratification [20]. If a particular diallelic, autosomal chromosome locus with alleles A_1 and A_2 is considered, then the genotypes are either A_1A_1 , A_1A_2 or A_2A_2 [20]. The genotype frequencies are

$$P(A_1A_1) = p_{11} \quad P(A_1A_2) = p_{12} \quad P(A_2A_2) = p_{22},$$

such that $p_{11} + p_{12} + p_{22} = 1$. The frequencies for allele A_1 and A_2 are given by

$$P(A_1) = p = p_{11} + \frac{1}{2}p_{12} \quad P(A_2) = q = p_{22} + \frac{1}{2}p_{12},$$

where $q + p = 1$. Under random mating between individuals in the population, nine different mating types are possible [20]. These are shown in Table 2.3.2, along with each genotype frequency and their offspring's different genotype probabilities.

From the values in Table 2.3.2, the frequencies of the offspring's genotype can be calculated [20]. The frequencies of parental genotypes are used, along with the probabilities for the different genotypes of the offspring. The frequencies of the genotypes A_1A_1 , A_1A_2 and A_2A_2 are

$$P(A_1A_1) = 1 \cdot p_{11}^2 + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{4} \cdot p_{12}^2 = p^2$$

$$P(A_1A_2) = \frac{1}{2} \cdot p_{11}p_{12} + 1 \cdot p_{11}p_{22} + \frac{1}{2} \cdot p_{11}p_{12} + \frac{1}{2} \cdot p_{12}^2 + \frac{1}{2} \cdot p_{12}p_{22} + 1 \cdot p_{11}p_{22} + \frac{1}{2} \cdot p_{12}p_{22} = 2pq$$

$$P(A_2A_2) = \frac{1}{4} \cdot p_{12}^2 + \frac{1}{2} \cdot p_{12}p_{22} + \frac{1}{2} \cdot p_{12}p_{22} + 1 \cdot p_{22}^2 = q^2.$$

The allele frequencies of the offspring can then be calculated using the genotype frequencies,

$$P(A_1) = P(A_1A_1) + \frac{1}{2} \cdot P(A_1A_2) = p^2 + \frac{1}{2} \cdot 2pq = p(p + q) = p$$

$$P(A_2) = P(A_2A_2) + \frac{1}{2} \cdot P(A_1A_2) = q^2 + \frac{1}{2} \cdot 2pq = q(p + q) = q.$$

These calculations show that the allele frequency is equal for both the parental generation and the offspring, and that there is a stable relationship between allele and genotype frequency. Deviation from the HWL can be caused by genotyping errors, but may also be because of the invalidity of these assumptions for a particular population [37]. When testing for deviation from the HWE for binary traits, the significance threshold for cases are usually less strict than for controls [37]. This is to ensure that SNPs associated to a certain disease under evolutionary pressure are not excluded.

Another factor that should be considered in the QC, is the heterozygosity rate, which is a measure of genetic diversity [37]. High deviations of the heterozygosity rate from the sample mean might indicate contamination or inbreeding, where contamination would cause a higher heterozygosity rate and inbreeding a lower one [37]. Subjects with heterozygosity rates above or below around 3 standard deviations from the sample mean should be excluded from the study [37].

The last two steps in the QC are relatedness and population stratification, which are a result of hidden substructures within the population [37][26]. Population stratification implies that there are sample individuals with another ancestral and demographic background [26]. Genetic variants that are actually associated to this background can be confused with being associated to the disease in question. The analysis for this step should be performed only on independent, uncorrelated SNPs from autosomal chromosomes [37]. A k -dimensional subset is produced from the data, (k usually equals 10) for each population substructure [37].

Cryptic relatedness involves a latent degree of relatedness between sample individuals, which is discovered through the GWAS data analysis [26]. With sample individuals assumed independent, this can cause confounding results. The analysis for relatedness is also performed for independent SNPs from autosomal chromosomes [37]. A particular threshold of relatedness is determined, and the Identity By Descent (IBD) is found for each pair of sample subjects. Subjects with an IBD above the set threshold are excluded from the study [37].

Data errors can occur for several reasons, but often they arise in the experimental procedure [37]. One important aspect is missing genotype data. This can occur due to experimental causes, but also, some SNPs may be found invalid after the QC is performed [37]. These missing SNPs can be replaced by using imputation methods. Imputation methods are used to fill in missing genotype information for untyped variants, and have successfully increased the statistical power of many GWAS [26].

Imputation takes advantage of linkage disequilibrium (LD) between SNPs [39]. LD means that SNPs are non-randomly inherited together [26]. Imputation can be done for both related and unrelated individuals, the principle is still the same. Sample individuals are genotyped for a large number of genetic variants, often from 100,000 to 1,000,000, and are compared to a reference map of genetic variants containing an even larger number of genetic variants [39]. This map is generated from highly resequenced or densely genotyped individuals, and usually comes from the HapMap International Consortium [39][26]. Sequences of haplotypes found in both the study samples and reference map can then be identified, and the alleles missing from this sequence in the sample subject can be copied from the reference map [39]. The difference between performing imputation for related and unrelated subjects, is that these shared sequences

of haplotypes will be much shorter for unrelated individuals and thus more difficult to detect [39].

The HapMap International Consortium was developed to detect genetic variants, study their frequency in the population, and to find correlations between them [40]. It is a map of human haplotypes that consists of haplotype blocks. Haplotypes are sets of SNPs that tend to be inherited together, and haplotype blocks consist of a specific pattern of SNPs that are in LD and which therefore are inherited together [41]. When a genetic disease is inherited down through generations, the haplotype sequence is shortened because of subsequent recombination. The shorter the distance between two gene variants, the less likely recombination is to occur [41]. Thus, a certain sequence of the haplotype containing the disease-causing mutation is conserved throughout the population. The genetic variants in this sequence are non-randomly linked with each other and to the disease [41].

The main purpose of the HapMap is to simplify the process of detecting the genetic variants causing certain diseases [40]. Common genetic variants associated with diseases and traits are more easily detected using this approach, but also rarer variants have been found. Using haplotypes have proven to increase the detection of susceptibility genes for certain rare diseases [40]. If the amount of significant associations found with haplotype-based methods is higher than by typing SNPs directly, this can mean two things; it may be that this haplotype is directly causal to the disease, or that the haplotype tags genetic variants with a higher efficiency than single SNPs do [40].

2.3.5 Quality Control of Results

The results from a GWAS are most commonly represented in a quantile-quantile (Q-Q) plot, as shown in Figure 2.3.2 [26]. Negatively ranked logarithmic p-values of the observed associations are plotted against their expected logarithmic p-values under the null distribution. Here, chi squared test statistics are used [26]. Confounding factors such as population stratification and cryptic relatedness can be easily detected in the Q-Q plot. The blue dotted line in Figure 2.3.2 represents the expected p-values under the null distribution. If the data points follow this line, it implies that there is no significant association between the genetic variant and the trait [26]. Cryptic relatedness and population stratification will lead to deviations from the blue dotted line along the entire plot, as can be seen in panels b) and c) [26]. Panel c) also shows some signs of genetic variants with strong associations. Genetic variants with large effect sizes will generate more local deviations within the plot, and these will appear in the high significance range to the right, as seen in d) [26].

Manhattan plots present the associations with highest significance found in GWAS, and their chromosome location [26]. Figure 2.3.3 shows results from the type 2 diabetes component of the Wellcome Trust Case Control Consortium study [26]. As can be observed, the strongest associations were found on chromosome 6, 10 and 16. By adding a horizontal line signifying the genome-wide significance level, the plot can be easily interpreted to detect significant associations.

Assuming that SNPs are independent in GWAS is not correct, because LD between SNPs at different chromosomal loci throughout the chromosome must be considered [42]. When associated SNPs with p-values below α are detected on the Manhattan plot, these SNPs must be adjusted for LD. With LD, the association between SNPs at different loci will differ from what would be expected for independent SNPs [42]. Numerous methods and algorithms have been

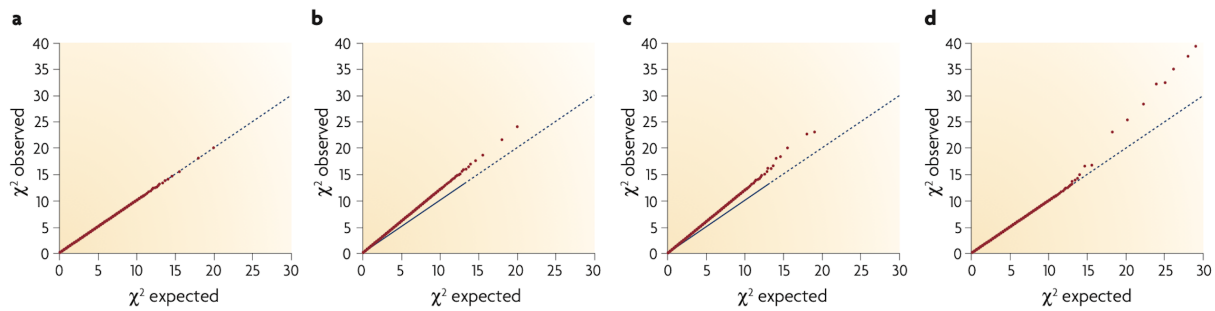


Figure 2.3.2: An example of quantile-quantile (Q-Q) plots, showing the test statistics from a genome-wide association study (GWAS). The y-axis shows negatively ranked logarithmic p-values of the observed association, while the x-axis shows expected logarithmic p-values. The blue dotted line represents the expected p-values under the null distribution. *a*) shows p-values that follow the blue dotted line, while *b*) and *c*) show signs of stratification or relatedness within the population. Panel *c*) displays a greater number of significant associations than what is expected under the null distribution. With permission from Springer Nature [26].

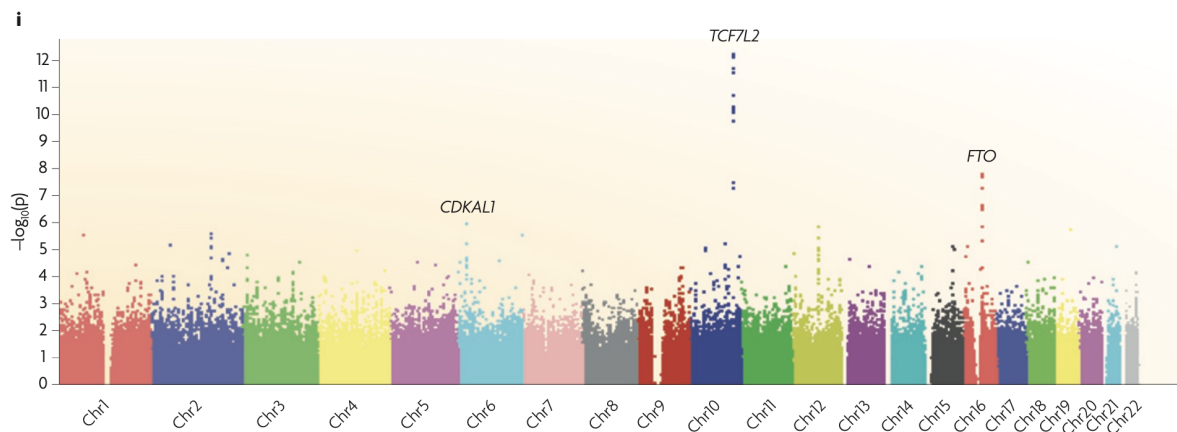


Figure 2.3.3: An example of a Manhattan plot from a Genome-Wide Association study (GWAS). With permission from Springer Nature [26].

developed for the purpose of adjusting for LD in GWAS. Among these algorithms are Proxy-GeneLD, which separates SNPs into LD blocks [43]. The lowest p-value found in each block is adjusted by the number of blocks in the particular gene. In this way, LD between SNPs within the same gene is considered. However, this method does not take into account SNPs that are in the same pathway, but in different genes, which is a disadvantage [43].

Replication is essential to determine whether a detected association is truly significant [26]. This is due to the vulnerability of GWAS to certain errors and biases, that arise because of the large sample sizes and the issue of multiple tests. In addition to this, complex traits and diseases are often caused by genetic variants with low effect sizes, which can also lead to an increased amount of errors [26].

Replication is performed to either confirm or debunk the significance of the association found, and to assess which are the causes of the potential errors of the first study [26]. It is essential that independent samples are used for the replication, and that separate genotyping arrays are utilized [26]. This is to remove any systematic errors coming from technical equipment. The replication should be done with the same allele or haplotype, the same phenotype and should

make use of the same genetic model (dominant, recessive or additive) [26].

2.3.6 Example: Schizophrenia

An example of a disease that GWAS has increased our knowledge of, is Schizophrenia [44]. This is a disease associated with psychosis and social and emotional difficulties. It is believed to cause disturbances in the neurodevelopment, and reduces the life expectancy with 15 - 20 years [44]. Pharmacological treatments exist, but these are usually not sufficient [45]. They target the type 2 dopaminergic receptor, which is involved in a mechanism that was found to be related to schizophrenia 60 years ago. Since then, no new treatments have been developed that differ in their molecular targets [45]. This is why GWAS have been so important for Schizophrenia, and also other psychiatric disorders where the mechanisms causing them have been largely unknown. Schizophrenia is a polygenic and complex disease, and it has been estimated in recent studies that about one third to half of the increased susceptibility to the disease is caused by common genetic variants [45].

The Schizophrenia Working Group in the Psychiatric Genomic Consortium (PGC) published the largest GWAS that has ever been performed concerning Schizophrenia [45]. Here, 108 significant chromosomal loci were detected, of which 83 were not yet reported [45]. They used 36,989 cases and 113,075 controls in their study. Another study published in 2018 located 145 significant chromosomal loci, where 93 of them were also found in the PGC study [44]. Two of the new significant associations found were replicated in other studies. By performing further studies and locating additional genetic variants associated with schizophrenia, the possibility for developing a pharmacological treatment that can target specific genes, increases.

2.3.7 Conclusion

The development of GWAS over the past 20 years has served to strongly enhance the scientific community's knowledge of the genetic architecture behind certain traits and diseases [26]. Numerous chromosomal loci associated with diseases and traits have been identified. However, there is still a long way to go until all genetic variants at every locus is detected. Up until today, only a small fraction of all genetic variation has been analyzed using GWAS [26]. Also, a large part of the molecular mechanisms through which the genetic variants affect traits and diseases are unknown. New developments within technology are required to help fill all these gaps in the field of GWAS, but hopefully this will occur in the near future.

2.4 Networks

Section 2.4 is taken from the specialization project in TBT4500 [19]. The world is built up by complex systems [17]. These complex systems are described by networks which considers each components' interactions with each other. To understand all of these complex systems, we need a more thorough understanding of the networks they consist of. Networks are found everywhere, in nature, science, technology and business. Cellular networks involve genes, proteins and metabolites, and describe the interactions between these factors to maintain the inner workings of an organism [17]. Social networks are a different kind of network which describe the interactions of knowledge between people - friends, family and colleagues. The development of epidemic prediction based on network modelling has highly increased our ability to predict how

contagious diseases spread throughout the population [17]. This has been especially important during the past year, with the spread of Covid-19.

Google is one of the biggest companies of the 21st century and relies on intricate network technology to provide a mapped network of the web [46]. When a search is performed by a user in Google, it will provide all relevant results in less than a second. Billions of web pages are searched and then ranked based on relevance. Relevance is evaluated based on different factors such as search words, reliability of the source, the user's position and their search settings [46]. To do this, many different algorithms are required, and the network technology behind this is probably one of the best in the world.

2.4.1 Properties

Even though networks can be highly different in what they describe, their structure can be quite similar [17]. A network contains nodes (vertices) connected to each other through links (edges). The number of nodes (N) describes the size of the network and these nodes are usually labelled. The number of links between them (L) describes how many of the nodes that are connected to each other. These are usually not labelled. The links can either be directed or undirected, which means they can either only go from one node to another or go both ways [17]. An example of a directed network is the phylogenetic tree, which shows the evolutionary relationship between species. A social network is an undirected network, since a social interaction requires at least two people. There are also networks where both directed and undirected links are present.

Some essential parameters used to describe networks are degree, average degree and degree distribution [17]. Each node, i ($i = 1, 2, \dots, N$), in a network has a degree, k_i , which describes how many nodes i is connected to. When summing up all node degrees throughout an undirected network and dividing by two, the total amount of links, L , is obtained,

$$L = \frac{1}{2} \sum_{i=1}^N k_i.$$

For a directed network, links can either go *from* a particular node or *to* that node [17]. For this reason, either out- or in-degree must be calculated for each node and added together to get the final number of links, L . The calculation for L in a directed network is given by

$$L = \sum_{i=1}^N k_i^{in} = \sum_{i=1}^N k_i^{out}.$$

The average degree of all nodes in an undirected and directed network respectively, is defined as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2L}{N}$$

$$\langle k_{in} \rangle = \frac{1}{N} \sum_{i=0}^N k_i^{in} = \langle k \rangle = \frac{1}{N} \sum_{i=0}^N k_i^{out} = \frac{L}{N}.$$

The degree distribution (p_k) is defined as the probability that a random node in the network has a distribution k , and is described by

$$p_k = \frac{N_k}{N},$$

where N_k is the number of nodes with distribution k [17]. The local clustering coefficient, C_i , of a node i is determined by the neighbouring node's number of links. This is described by

$$C_i = \frac{2L_i}{k_i(k_i - 1)},$$

where k_i is the degree of node i and L_i is the number of links between the k_i neighbouring nodes [17]. The clustering coefficient represents the probability that the neighbouring nodes of i are connected, and has a value between 0 and 1. The average clustering coefficient, $\langle C \rangle$, for the whole network is

$$\langle C \rangle = \frac{1}{N} \sum_{i=0}^N C_i,$$

where $\langle C \rangle$ then represents the probability that a randomly selected node has two neighbouring nodes that are connected to each other [17].

Both the degree (k_i) and the clustering coefficient (C_i) can be used to describe hubs [17]. Hubs are nodes that are central in the network, meaning they are highly connected to other nodes. They will tend to have higher values of both degree and clustering coefficient.

A mathematical representation of all links in a network is done through an adjacency matrix [17]. For a network with N nodes and N links, a corresponding adjacency matrix would have the dimensions $N \times N$. The possible entries in the matrix are

$$A_{ij} = 1 \text{ (If node } j \text{ is linked to node } i)$$

$$A_{ij} = 0 \text{ (If nodes are not connected) [17].}$$

Undirected networks will have a symmetrical matrix since each link is represented in both directions, such that $A_{ij} = A_{ji}$ [17]. This means that the number of non-zero entries in the matrix will be twice the number of links in the network. The degree of the undirected network can be obtained by summing over the rows or columns. For a directed network, summing over the rows or columns provides either in- or out-degree. An illustration of how the adjacency matrix would look for an undirected and directed network is shown in Figure 2.4.1.

Weighted networks represent cases where the amount of interaction between two nodes matter [17]. This can be the case in social networks, where the strength of social ties between individuals depends on communication, intimacy or the duration of the relationship [47]. In the neural network of the brain, the interaction between two nodes can be weighted based on the number of synapses or gap junctions between them [17]. Each entry in the adjacency matrix is then represented by the weight (w_{ij}) of each interaction between i and j .

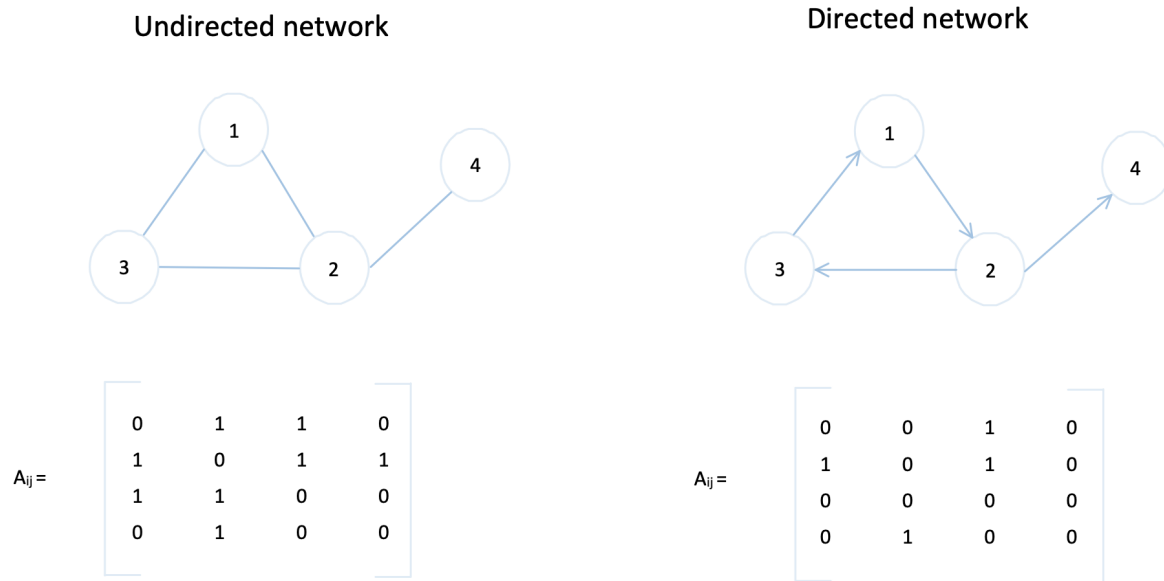


Figure 2.4.1: An illustration of an undirected and directed network with their corresponding adjacency matrix, with inspiration from [Barabási \[17\]](#).

2.4.2 Types of Networks

Bipartite networks consist of two distinct sets of nodes, where each link connects two nodes from each set [\[17\]](#). From this bipartite network, sub-networks can be extracted. If two nodes (A and B) from set 1 are connected to the same node in set 2, a link can be drawn between A and B. In this way, distinct networks from each set can be constructed based on their interaction with the other set of the bipartite network [\[17\]](#). An example of a bipartite network is the "diseasome" described in Section [2.5](#). In that case, disease genes are connected to human diseases if a causal link has been found between them [\[18\]](#). From this bipartite network, the human disease network (HDN) and disease gene network (DGN) are constructed. Genetic diseases are then connected if associated with the same disease gene, and disease genes are connected if associated with the same disease.

The human disease network is an example of a scale-free network [\[17\]](#). The nodes in these networks follow a power-law distribution, instead of a Poisson distribution as is the case with random networks. The Poisson- and power law distribution is defined as

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

$$p_k \sim k^{-\gamma},$$

respectively, where γ represents the degree exponent [\[17\]](#). The two distributions are compared in Figure [2.4.2](#). The linear plot is shown in figure *a*) and the log-log plot is shown in *b*), while figure *c*) and *d*) show a random- and scale-free network, respectively. It can be observed from the log-log plot in figure *b*) that at low levels of degree (k), the power law has a higher degree distribution (p_k) than the Poisson distribution. This implies that there is a higher number of

scarcely connected nodes in the scale-free network [17]. At the average degree ($\langle k \rangle$), the Poisson distribution has a higher p_k than the power law, which means that the random network has a higher number of nodes with degree around $\langle k \rangle$. At higher levels of k , the power law again has a p_k above the Poisson distribution, which means there is a larger number of highly connected nodes, hubs, in the scale-free network [17].

In the construction of scale-free networks, the initial number of randomly connected nodes is m_0 [17]. A number of m links are then constructed, where $m < m_0$. The new node i , has a higher probability of being linked to an existing node with a higher degree. In this way, hubs are created.

2.4.3 Clustering Algorithm

The Louvain method is a hierarchical clustering algorithm used to locate communities within networks [48]. The quality of the distribution of a node to a specific community is measured using the modularity score, Q , which is a quality function given by

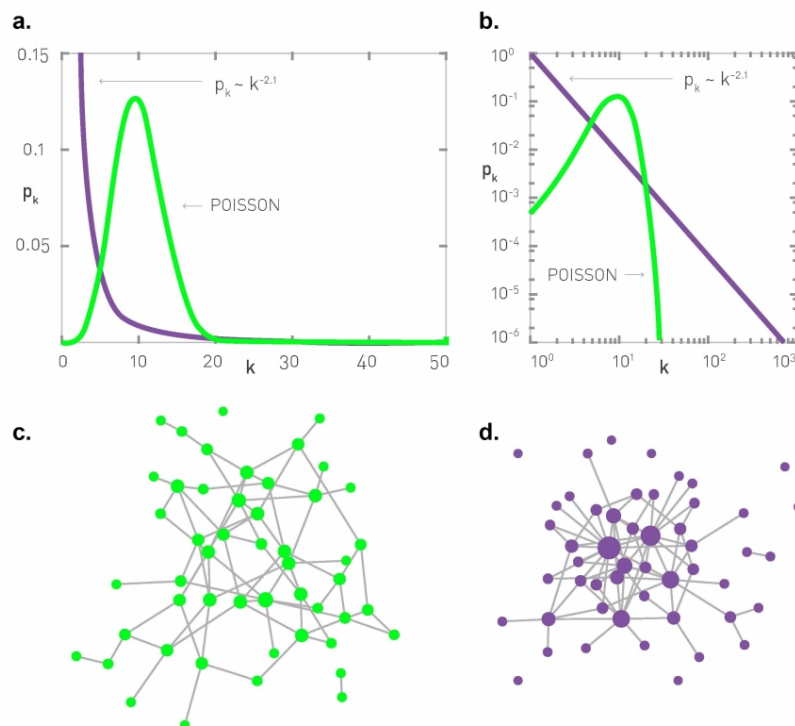


Figure 2.4.2: An illustration of the Poisson distribution compared to the power law distribution, both on a linear plot (a) and a log-log plot (b). In b), when the power-law has a higher degree distribution (p_k) than the Poisson distribution at a particular degree (k), there is a larger number of nodes with that degree in a scale-free network. Conversely, if the Poisson distribution has a higher degree distribution than the power-law, there is a larger number of nodes with that particular degree in a random network. The Poisson distribution is coloured in green, while the power law distribution is purple. A random- and scale-free network are shown in c) and d), respectively [17] (CC BY-NC 3.0).

$$Q = \frac{1}{2m} \sum_c \left(e_c - \gamma \frac{K_c^2}{2m} \right), \quad (2.4.1)$$

where e_c is the number of edges in community c , and $\frac{K_c^2}{2m}$ is the expected number of edges [49]. The variable K_c represents the sum of all degrees of each node in community c , while m is the total number of edges in the network. The resolution parameter is given by γ ($\gamma > 0$). A larger value of γ means that there is a larger number of communities, while a lower value means fewer communities [49]. The modularity score demonstrates how closely connected nodes are within the community, compared to how they would be connected in a random network [49]. The Louvain method attempts to maximize the modularity score for each community, which means that it tries to maximize the difference between the actual and expected number of edges [49].

2.5 The Human Disease Network

Section 2.5 is taken from the specialization project in TBT4500 [19]. This section considers the paper titled The human disease network, which was published in 2007 by Goh et al. [18]. Unless specified otherwise, this article is the source of all material covered in this section. The main purpose of the human disease network is to investigate whether genetic diseases and their corresponding genes are connected at a cellular and organismal level, instead of looking solely on specific diseases and their associated genes. Locus heterogeneity, which means that mutations in several different genes cause the same phenotype, indicates that there is a higher level of organization. Also, several mutations in the same gene can cause different phenotypes, which proves that the cell has a highly connected interior. To demonstrate this connectivity, the "diseasome" was created, where genetic diseases were connected with disease-causing genes.

2.5.1 The Diseasome

The diseasome shown in Figure 2.5.1 is a bipartite graph, thus it consists of two distinct sets of nodes. These sets are in this case human genetic diseases and disease-related genes. A gene is connected to a disease within the network if a mutation in that gene is causal to the disease. From this, two separate networks were generated, the human disease network (HDN) and the disease gene network (DGN).

The dataset used in the creation of this network was taken from Online Mendelian Inheritance in Man (OMIM) [50]. OMIM is a large compendium of human genetic diseases and disease-related genes which was generated by NCBI. It contains information about all known Mendelian diseases, and includes over 15,000 genes [50]. Over the past years, OMIM has developed and has also added complex diseases and their corresponding susceptibility genes.

The HDN shown in Figure 2.5.2 connects genetic diseases through common genes. The diseases are thus nodes, and they are connected if a mutation in the same gene can be causal to both diseases. The HDN contains both large and small clusters, which indicates that both individual diseases and disease classes are connected. Most diseases are connected to a few diseases, while a small number is highly connected. Goh et al. found that 68 % of all diseases were connected to at least one other, while 40 % were incorporated into one large cluster. This means that there

is a small number of highly connected nodes, hubs, within the network. These hubs are often cancer-related, because cancers are often caused by a common tumor suppressor gene.

The genetic diseases within the network tend to organize in clusters according to their disease class. Cancers and neurological diseases form the largest clusters, while metabolic, skeletal and multiple disorders form the smallest ones. This can be explained by locus heterogeneity, which means that several mutations at different chromosomal loci can cause the same phenotype. Cancers and neurological diseases have a high degree of locus heterogeneity, while this is not the case for metabolic, skeletal and multiple disorders. Thus, certain types of cancer can be caused by several different mutations. This probably means that cancers have a high degree of connectedness with other diseases as well.

In the HDN shown in Figure 2.5.2, each node is coloured according to its disease class. All disease classes are listed to the right. The size of the node indicates how many genes are associated with the disease, and the thickness of the linkage between nodes indicates how many genes the two diseases share. Also, the linkages between diseases of different classes are coloured in grey, while linkages between diseases of the same class are coloured with the same tone of colour as the nodes. In the case of the DGN, two nodes are connected if they are associated with the same disease. The size of the nodes represents how many diseases that particular gene is associated with, and the linkage is grey if the two genes are associated with diseases of a different class.

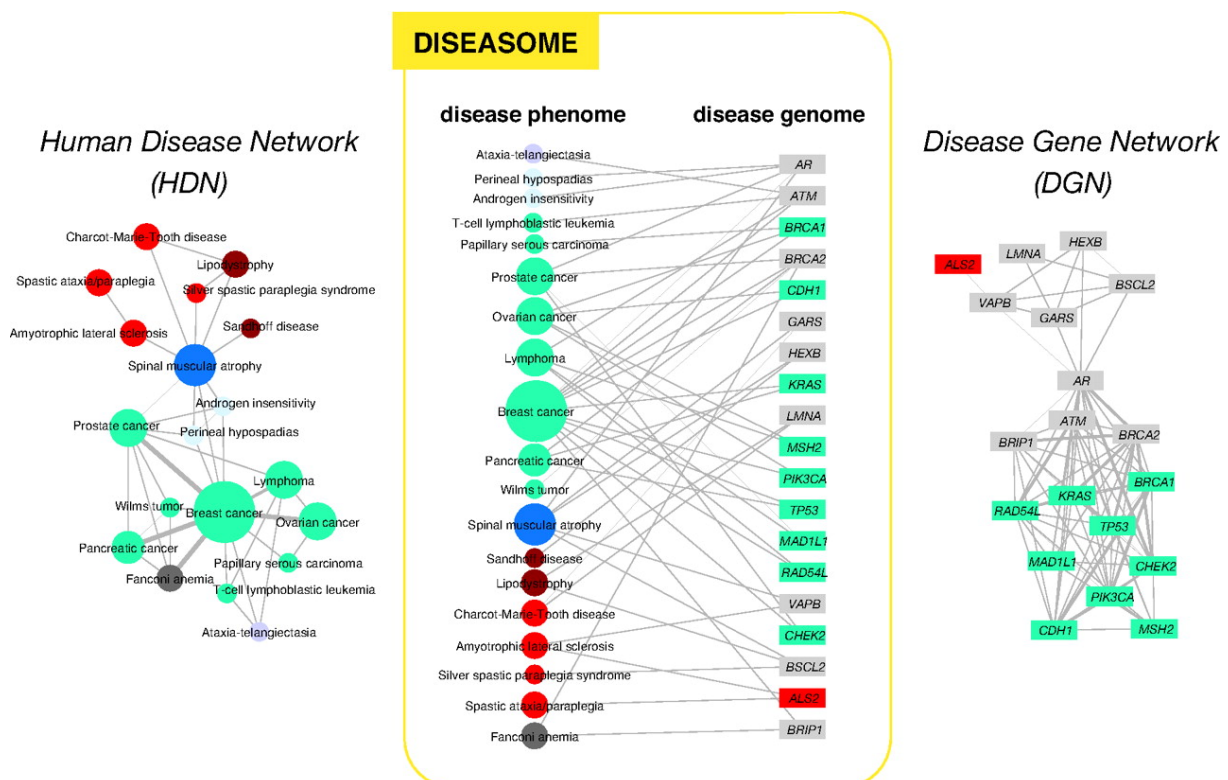


Figure 2.5.1: The diseasome network, where a gene is linked to a disease if an association has been found between them. From the diseasome, the human disease network and disease gene network are constructed [18]. With permission from PNAS, Copyright (2007) National Academy of Sciences, U.S.A.

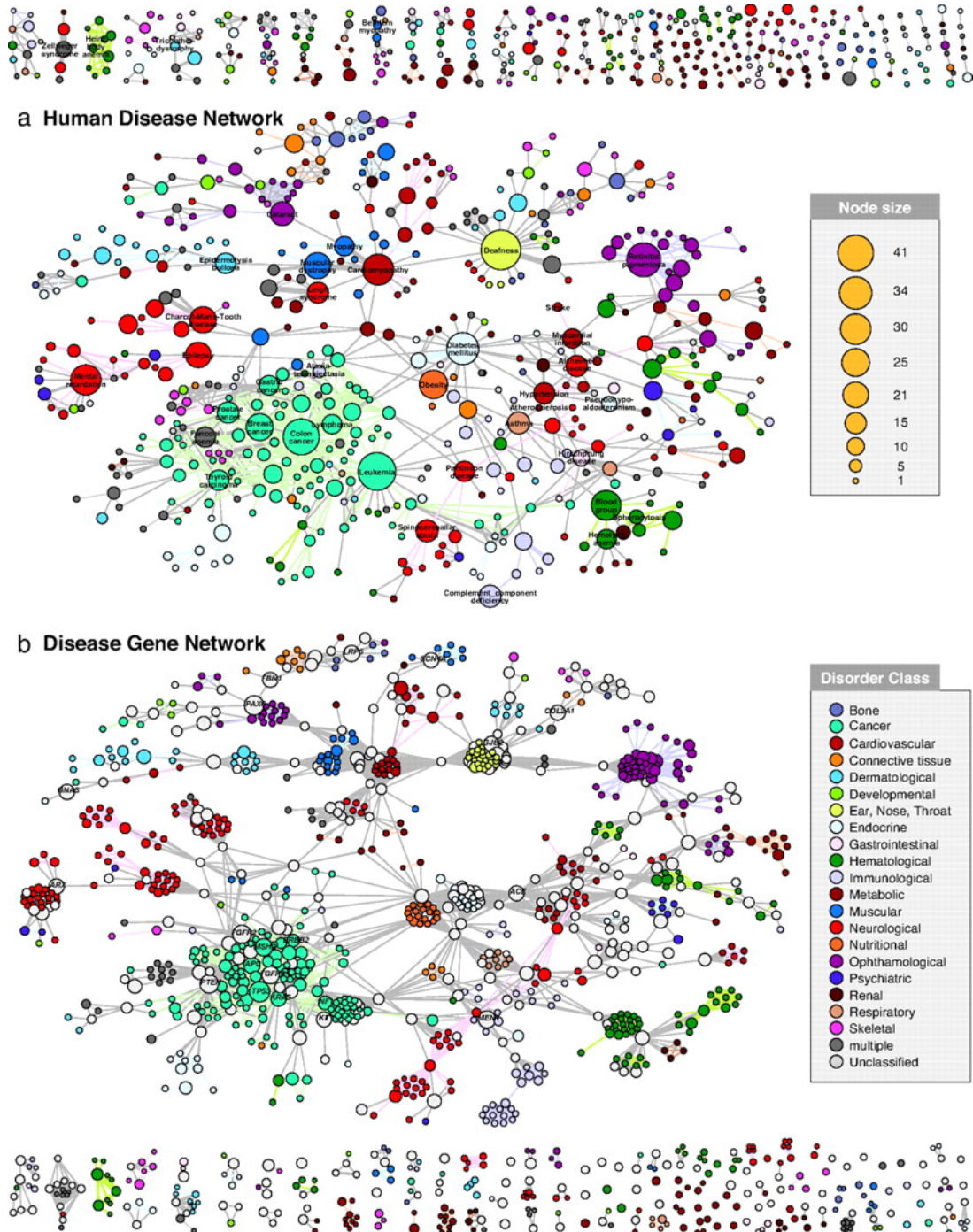


Figure 2.5.2: The human disease network (HDN) and the disease gene network (DGN) [18]. With permission from PNAS, Copyright (2007) National Academy of Sciences, U.S.A.

2.5.2 Functional Modules

Four different hypotheses were tested for the human disease network regarding functional modules; i) that proteins encoded by the disease genes of connected diseases also interact within the same molecular complex, cellular pathway or functional module, ii) that genes associated to the same disease encode proteins with the same cellular and functional features as annotated by the Gene Ontology (GO), iii) that interacting proteins within the same functional module are found in the same tissue, and iv) that disease genes within the same functional module also should show high correlation of gene expression.

Hypothesis i) states that when diseases are connected, their associated genes encode proteins that interact with each other rather than other proteins, in the same molecular complex, cellular pathway or functional module. This hypothesis was tested by comparing the DGN with a protein-protein interaction network to find whether the same interactions were detected in both networks. The assumption was that proteins encoded by disease genes associated with a certain cluster of diseases would more likely interact with each other than with other proteins. [Goh et al.](#) found ten times as many of the same interactions in the two networks, as would be expected to occur randomly. The hypothesis was therefore confirmed.

Hypothesis ii) suggests that genes associated with the same disease, encode proteins which has the same cellular and functional features annotated by the GO. The GO is a computational model which annotates biological processes, cellular pathways and molecular functions to different genes of various organisms, from bacteria to humans [\[51\]](#). This was performed by comparing the GO annotations for each gene associated with a specific disease, for all three categories; biological process, cellular pathway and molecular function. The hypothesis was confirmed by the finding of significant GO similarity between genes within each category, when compared to random expectations.

Hypothesis iii) states that interacting proteins encoded by disease genes within the same functional module are expressed in the same tissue. This was investigated by measuring the maximum proportion of those genes associated with the same disease expressed in a particular tissue, compared to the total number of genes associated with that disease. This is defined as the tissue homogeneity coefficient. [Goh et al.](#) found that 68 % of diseases showed almost absolute tissue homogeneity, compared to 51 % which would be expected to occur by chance.

The final hypothesis to be tested, iv), is whether the expression of genes within the same functional module is highly correlated. This can be tested by measuring the Pearson Correlation Coefficient (PCC), which shows both the extent and direction of correlation. When measured for pairs of genes associated to the same disease, the PCCs were found to be positive and of a higher value than what was expected by chance.

All four hypotheses regarding the diseaseome were thus confirmed, proteins encoded by genes associated to the same diseases tend to; interact more with each other than other proteins, have the same functional and cellular features as annotated by the GO, be expressed in the same tissues and have high co-expression levels. Since genes associated to the same disease encode proteins that work within the same functional module, the breakdown of this functional module can be caused by different factors. A mutation in a single gene can cause a malfunction in a protein, which again causes the breakdown of the functional module. In this way, several mutations in multiple genes can cause the same phenotype. This is what is called polygenic phenotypes.

2.5.3 The Role of Cellular Networks in Human Diseases

Another set of hypotheses were tested by [Goh et al.](#) regarding the role of disease genes in cellular networks. The first hypothesis tests whether human disease genes have a tendency to encode hubs. It was found that this is highly dependent on the type of disease gene in question. Disease genes can be categorized as essential or non-essential genes. Essential disease genes are necessary for growth and development, and mutations in these can be fatal for the organism. The initial analysis concerning this problem found that there is a weak correlation between disease genes and hubs, but when the hypothesis was tested separately for essential and non-essential genes, the result was another. Essential genes displayed a high correlation with hubs. For non-essential genes, no such correlation was found. The conclusion was that the initial weak correlation between disease genes and hubs found, was solely due to essential genes. This indicates that essential disease genes are more connected to the rest of the cell than non-essential genes are.

To test if the essential disease genes were more connected than non-essential genes, the expression pattern of essential disease genes and other important genes within the cell was measured. This was done by measuring the gene co-expression coefficient (PCC) between both essential- and non-essential genes and all other genes within the cell. It was found that genes showing high co-expression coefficients with other genes were more likely to be essential. Conversely, non-essential disease genes were associated with genes that showed anti-correlation or no correlation to other genes.

The next hypothesis to be tested was whether housekeeping genes were more likely to encode disease genes. Housekeeping genes are expressed in all cells within the organism, and are essential for maintaining basic cellular functions. [Goh et al.](#) found that 9.9 % of all housekeeping genes are disease genes, while for non-housekeeping genes this percentage is 13.5 %. Also, 59.8 % of all housekeeping genes were found to be essential, while 40.5 % of non-housekeeping genes were. This means that the majority of housekeeping genes are essential. In addition to this, [Goh et al.](#) tested to what degree essential and non-essential disease genes were expressed in certain tissues. The more tissues in which a gene was expressed, the more likely it was to be essential. The opposite is true for non-essential genes, which were expressed in a lower number of tissues.

All these results indicate that non-essential disease genes are quite peripheral in the cellular network - they are less likely to encode hubs, show low co-expression with other genes within the cell and tend to be expressed in few tissues. The opposite is true for essential disease genes, which have a high probability of encoding hubs, show a high level of co-expression with other genes and tend to be expressed in several tissues. Essential genes are also highly represented among housekeeping genes. This means that essential disease genes have a much more central role in the cellular network than non-essential disease genes.

In the OMIM database containing human genetic diseases and their corresponding genes, 1,267 essential- and 1,370 non-essential disease genes were found. The reason that the majority of disease genes are non-essential can be explained from an evolutionary point of view. Mutations in more central genes within the cellular network are more likely to be fatal for the organism, either before or early after birth. Highly central genes within the cellular network have a higher probability of being essential for the organism, and will therefore be important for its development and growth. Severe mutations in these genes will therefore have a tendency not to be

passed on to further generations. However, disease-causing mutations in more peripheral genes within the cellular network do not have such dramatic consequences, and this therefore allows the organism to develop into its reproductive age. These mutations are thus brought on to further generations.

2.6 Polygenic Risk Score (PRS)

The polygenic risk score (PRS) is a measure of an individual's liability for acquiring a particular phenotype [52]. PRS thus provides a measure for the heritability of a complex trait on the individual level. PRSs can be utilized for personalized medicine or disease diagnosis, and therefore have other usages than GWAS. In GWAS, even if an association between SNP and phenotype is found, the SNP usually only explains a fraction of the phenotypic variance [53]. This is because complex traits are determined by the effects of several different SNPs, compared to Mendelian traits, which are caused by a single SNP.

2.6.1 PRS Calculation

The calculation of PRS for a particular phenotype is performed by summing an individual's number of risk alleles, and multiplying each allele by its effect size [52]. The effect size is obtained from GWAS summary statistics data, and is a measure of the effect the genetic variant has on the phenotype in question [34]. It is thus the proportion of phenotypic variance caused by a particular allele. The formula for PRS is given by

$$PRS_i = \sum_{m=1}^{\#SNPs} \beta_m M_{mi},$$

where β_m is the effect size of a genetic variant m , and M_{mi} is the number of risk alleles of an individual i for a genetic variant m [54]. Since most GWAS performed concerns biallelic SNPs with a MAF larger than 0.1, the majority of SNPs included in a PRS analysis are of this type [52].

2.6.2 Requirements and Considerations

To perform a PRS analysis, both base- and target data are required [52]. Base data consist of GWAS summary statistics, and contain p- and β -values for associations found between SNPs at particular chromosomal loci and phenotypes. The target data should contain genotype- and phenotype information regarding the individuals for which the PRSs are calculated [52]. The base- and target datasets should be independent of each other [52].

Since the effect sizes used in the PRS analysis derive from a GWAS performed on a particular population, the effect sizes may not be accurate when transferred to the population of individuals that make up the target data [52]. In addition to this, linkage disequilibrium (LD) between SNPs, the non-random association between SNPs at different loci, complicates the PRS analysis [52]. It is therefore necessary to perform certain adjustments to the data before performing the analysis. If there were no errors in the effect sizes when transferred to the target data population, the PRS would be equal to the SNP heritability, h_{SNP}^2 , which is defined as the phenotypic variance

caused by a specific set of SNPs, when their effects are additive [55]. However, because of these deviations, the PRS will only approach h_{SNP}^2 as the target population increases.

There are two main strategies for adjusting effect sizes for the target population [52]. The first method is based on the shrinkage of effect sizes for all SNPs, using statistical shrinkage methods. The amount of shrinkage to be applied to the effect sizes may be difficult to predict, since it depends on the distribution of actual SNP effect sizes, which is not known. Because of this uncertainty, PRS should be calculated for all possible parameters, such that the appropriate shrinkage parameters can be found by optimization. LASSO is an example of such a method, where all smaller effect sizes are shrunk to zero [52]. The second strategy for adjusting effect sizes involves using a p-value threshold. In this way, the effect sizes of SNPs with p-values above the threshold, are adjusted to zero, and are thus not considered in the PRS calculation [52]. Since the optimal threshold is not known beforehand, the PRSs are calculated across a range of p-value thresholds. By performing association testing between the PRS and phenotype afterwards, the optimal p-value threshold can be found [52]. Generally, as many SNPs as possible should be included in the analysis, also those with a p-value above the genome-wide significance threshold, since this greatly increases the predictive power of the study [52].

There are also two main approaches for considering LD between SNPs in a PRS analysis; clumping of SNPs and inclusion of all SNPs while taking account of LD [52]. Clumping of SNPs indicates choosing, among SNPs in LD at a particular chromosomal locus, the SNP with the lowest p-value. This can be done based on either R^2 -value, where a cutoff-value is usually set from 0.1 - 0.5, or physical distance between SNPs, where the upper threshold is set to 500 kb [53]. The other associated SNPs at that locus are then excluded from the PRS analysis. In this way, the remaining SNPs will all have independent effects on the phenotype in question, and these effects can thus be summarized [52]. Clumping is usually combined with p-value thresholding as a way of adjusting for effect sizes, which is called the C + T method [52]. Including all SNPs while taking account of LD is often combined with statistical shrinkage methods such as LASSO, which was mentioned above.

Since the C + T method involves choosing the SNPs with the lowest p-value at each p-value threshold, there is a risk of overfitting the PRS analysis to the target data, and thereby making the wrong conclusions [52]. The optimal strategy is to perform an optimization using the target data, and then test the optimized parameters on an independent dataset. When performing a PRS analysis, there are only a base- and target dataset, and therefore no third independent dataset available for out-of-sample prediction [52]. An alternative is therefore to divide the target data into smaller subsets, which can be used as validation datasets. The best possible predictions are obtained by performing the validation on a series of differently divided subsets of the target data [52].

2.6.3 Quality Control

Before using base- and target data for any PRS calculations, a quality control (QC) must be performed [52]. For the base data, h_{SNP}^2 must be above 0.05. In addition to this, it is essential to learn which allele is the effect allele, meaning which of the two alleles on each chromosome actually influence the phenotype [52]. For the target data, it is important to have a large enough sample size to avoid coming to false conclusions. This is also true for the base data. With a smaller sample size, a less thorough QC can be performed, incorrect adjustments for population

substructure and LD might be made and the association testing between PRSs and phenotypes may be under-powered [52].

There are several steps that should be performed for both datasets [52]. Care should be taken when transferring files containing base- and target data, such that no errors occur. The genomic positions considered in both datasets should be based on the same sequence of overlapping DNA sequences, which are assigned in believed chromosomal order [52,56]. Since the GWAS results used as base data usually are taken from an online source, it is important to ensure that QC has been performed according to the guidelines described in section 2.3.4. If the strand on which the effect allele is positioned, is unknown, and the allele is paired with a complementary base on the other strand, the SNP is ambiguous [52]. This is because when these are found in both base- and target set, it is uncertain whether it is the same allele. These should therefore be removed from both datasets. Also, duplicated SNPs must be removed [52].

Further steps in the QC involves the removal of individuals where there is a mismatch between the sex that was reported and the sex signified by the sex chromosomes [52]. This can be checked by measuring the X chromosome homozygosity rate, where the individual is a woman if the rate is below 0.2, and a man if the rate is above 0.8 [52]. If there are sample overlaps between the base and target data, these should be removed, preferably from the base data. This is because the resulting inflation of significance between PRS and phenotype increases proportionately with the fraction of samples in the target set that overlaps with the base set [52]. Lastly, closely related individuals in the target set should be removed, since this also causes an inflation of the association between PRSs and phenotype [52].

2.6.4 PRS Analysis for Coronary Artery Disease

In an article from 2016 by [Khera et al.], a PRS analysis for coronary artery disease, also called coronary atherosclerosis (CA), was performed, where both genetic predisposition to the disease and lifestyle were considered [54]. At the time this article was published, 50 risk loci for CA had been discovered, and it is therefore only these SNPs that are considered in the PRS analysis. As of 2020, however, the number of discovered risk loci for CA has increased to 163 [57].

The PRS analysis was in this case performed for three cohorts from different studies, in addition to a cohort from a cross-sectional study [54]. The PRS for CA was calculated by first finding the number of each risk allele present in a single individual, represented by M_{mi} in formula 2.6.1. This number can be either 0, 1 or 2, since the allele can be found on either none-, one- or both of the chromosomes. The number of alleles was multiplied with the effect size of the risk allele in question, represented by β_m in formula 2.6.1. The effect size was given as the natural logarithm of the odds ratio (OR) provided in the base data for the cohorts [54]. The resulting values were summarized for all risk alleles in each individual, to acquire the final PRS. Using formula 2.6.1, along with numbers obtained from [Khera et al.], an example of the calculation for SNP $rs599839$ is given by

$$PRS_i = \beta_{rs599839} M_{rs599839,i} = \ln(1.11) \cdot 1 = 0.104,$$

for individual i . This calculation is performed for each SNP considered a risk allele, and the resulting values are summarized to find the total PRS. The results presented in this article indicate that genetic risk and lifestyle independently contribute to the total risk of acquiring CA [54].

[Khera et al.](#) concludes that PRS indeed can be used to predict cases of CA. However, according to an article from 2020 by [Lewis and Vassos](#), the reliability of the PRS seems to depend on four factors; (i) the knowledge we do have about the genetic risk of each individual, (ii) the insufficient knowledge due to missing genetic data, (iii) the risk of being in possession of inaccurate information due to incorrect assumptions used to approximate the effect sizes, and finally, (iv) the usage of the PRS analysis [\[16\]](#). If the PRSs are used to advice a population about a change in lifestyle, this demands less information than if the PRS is to be used as a justification for distributing personalized medicine [\[16\]](#).

In addition to these four factors, the ethnicity of the target population is determining for the reliability of the PRS [\[58\]](#). GWAS summary statistics are required for acquiring the effect sizes of risk alleles, and the majority of these studies have been performed on populations of European origin. This means that the results of the GWAS may be less transferable to other non-European populations due to differing genetic variant frequency and LD patterns [\[58\]](#). This has serious consequences for the clinical use of PRS, since it can be assumed that PRS predicts disease more accurately for individuals of European descent than other parts of the population. It is believed that PRSs may be used for diagnosis of diseases and for personalized medicine, but currently the advantages of PRS are limited to only a fraction of the population [\[58\]](#).

[Khera et al.](#) found that the individuals with highest genetic risk, thus a higher PRS, had a 91 % greater probability of acquiring CA, compared to the individuals with lowest genetic risk [\[54\]](#). In addition to genetic risk, lifestyle has a great impact on the probability of acquiring CA. Generally, a healthy, favourable lifestyle was found to decrease the risk for CA, independent of genetic risk. For the individuals with highest PRS, a healthy lifestyle decreases the risk of acquiring CA with 46 % [\[54\]](#).

2.7 The HUNT Study

The HUNT (The Trøndelag Health Study) study is made up of four health surveys; HUNT1 in 1984–86 with 77,212 adult participants, HUNT2 in 1995–97 with 62,237 adult participants, HUNT3 in 2006–08 with 50,807 adult participants and HUNT4 in 2017–19 with 161,488 adult participants [\[59\]](#). For the first three studies, all information was obtained from individuals of Nord-Trøndelag county in Norway, aged 20 and above. In HUNT4, individuals from both Nord- and Sør-Trøndelag were included, and the number of participants therefore increased substantially. The data were collected using various questionnaires, clinical tests such as measures of weight, height and blood pressure, and by performing interviews. In total 240,000 individuals have participated [\[13\]](#). The study has been performed at regular 11-year intervals, and certain individuals have been invited to a follow-up study, which was the case for the lung-, diabetes-, osteoporosis- and chronic pain study [\[59\]](#).

The initial purpose of the HUNT study was mainly to evaluate arterial hypertension, diabetes, chest X-ray screening for tuberculosis and quality of life throughout the population [\[59\]](#). Over the years, the objective of HUNT has developed to include health issues related to lifestyle, the occurrences of diseases, and associations between genotype and disease phenotype [\[59\]](#). From HUNT1 to HUNT2, the disorders cardiovascular disease, diabetes, obstructive lung disease, osteoporosis, headache, mental health, chronic musculoskeletal pain and urinary incontinence were included in the study, as they were viewed as the most central health issues among the Norwegian population. In the HUNT2 study, blood samples were taken in addition to the other clin-

ical tests^[59]. The Young-HUNT study was also initiated along with HUNT2, which involved participants at the age of 13-19. Young-HUNT1 was performed in 1995-97, and Young-HUNT2 in 1999-2000^[60].

HUNT3, in addition to the disorders added for HUNT2, also included individuals' attachments to religions and their culture considerations^[59]. HUNT4 continued with much of the same investigations as the previous studies, but also had further developments. Among other things, HUNT4 focused on the health of elderly people above the age of 70, including factors such as hearing impairment, dental health and occurrence of dementia. Other focus areas were body composition of fat, and also medicine targeted to individuals based on genetics^[61]. The Young-HUNT studies were repeated along with the HUNT3- and HUNT4 studies^[60].

At the time of the HUNT3 study, in 2006-08, the HUNT biobank was established^[62]. This biobank contains biological samples from the HUNT participants, which are stored, analyzed and distributed out to various projects. There has been taken 150,000 samples from about 100,000 unique individuals^[63]. The goal is to genotype the DNA of all HUNT participants^[62]. The amount of research data from the HUNT study has strongly increased since its beginning, and in 2013 the HUNT cloud was therefore created^[62]. This is a storage service that contains large amounts of sensitive information regarding HUNT participants, of which only researchers with authorization can access. This allows for the utilization of HUNT data across different institutions and country borders^[62].

Since the beginning of the study, HUNT data have been used in numerous projects and been involved in a total of 1,821 published articles^[13]. There are currently around 300 national and international research projects using HUNT data as of 2021^[13]. The MindMap project is an example of such an international collaboration, where HUNT data are utilized to learn how the mental health of elderly people is negatively affected by living in cities, and what can be done to improve this^[13]. An example of a national project making use of HUNT data is the HUNT-MI project on pharmacogenetics, from which the data used in this thesis have been obtained^[64].

The project HUNT-MI focuses on medicine currently used for the treatment of cardiovascular diseases today, which includes lipid lowering drugs, oral anticoagulants and platelet aggregation inhibitors^[64]. Pharmacogenetics is a research field that analyzes patient response to medication based on genetic differences. There are four main focus areas; to identify any potential genetic variants that can explain lacking effects or side effects in individuals using these medical treatments, to evaluate the genetic variation frequency of known pharmacogenetic genes in the population, analyze individuals who have a reduced gene functionality in previous medical targets in failed phase II/III studies, and to validate any potential results by international cooperation^[64].

This project makes use of the HUNT databank, which contains information collected from the participants of the HUNT study through questionnaires, interviews and measurements^[13]. By also utilizing other data sources, the information from the HUNT databank is correlated with variables such as age, sex, weight and demography, and also life style- and risk factors^[64]. The Norwegian prescription database is used to associate patients' use of medication with their genetic composition, and analyze any potential side effects or complications they may have due to these medications^[64]. In addition to the data sources mentioned above, hospital data from Helse Nord-Trøndelag (HNT) and the Norway Control and Payment of Health Reimbursement (KUHR) database are utilized. HNT contains ICD-codes and other information from HUNT

participants registered with cardiovascular diseases and other associated disorders^[64]. KUHR also contains ICD-codes, but includes cases registered outside hospitals that may be of less severity, such as visits at general practitioners or physiotherapists^[64]. Additional data sources used are the Norwegian cause of death registry (COD), the UK Biobank and the Norwegian Mother, Father and Child Cohort Study (MoBa)^[64].

2.8 Diseases

For this thesis, a PRS analysis is performed for four diseases: Angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension. These are cardiovascular diseases (CVDs), and are all found in individuals of the HUNT population. They are highly connected to each other, and are caused by many of the same risk factors.

2.8.1 Angina Pectoris

Angina pectoris (AP) is a condition involving chest pain or discomfort, which occurs because the heart muscle receives inadequate levels of blood and oxygen^[65]. This can be caused by activities which cause the heart to require an increased blood flow, such as physical activity, emotional strain, excessive heat or cold, large meals, high amounts of alcohol and smoking cigarettes. It can also be caused by coronary atherosclerosis, which is the narrowing or blockage of blood vessels bringing blood to the heart^[65]. Its symptoms can usually be relieved by a period of rest. However, AP increases the risk of acquiring myocardial infarction, since it means that some part of the heart receives less blood than it requires^[65]. AP can be prevented by maintaining a healthy lifestyle, which includes sufficient amounts of exercise, a healthy diet, no smoking and an ability to manage stress. Treatment can be provided by taking nitroglycerin, which widens the blood vessels to ensure an increased blood flow to the heart muscle^[65].

2.8.2 Myocardial Infarction

Myocardial infarction (MI), also termed heart attack, is caused by a reduced blood flow to some parts of the heart^[66]. The reduced blood flow is caused by the formation of plaques within the blood vessels. These plaques consist of calcium, cholesterol and other fatty substances, and when they break, blood clots form^[66]. These blood clots are the main cause of the heart attack. A blood clot hinders blood and oxygen from reaching the heart, and heart tissue will therefore start to die. After about 30 minutes, this will cause the heart muscle to undergo irreversible damage^[66].

An elevated risk for suffering from MI might be caused by either genetic- or acquired risk factors^[66]. Genetic risk factors include hypertension (elevated blood pressure), family history of cardiovascular disease, high levels of low density lipoprotein (LDL) cholesterol and triglycerides, diabetes type 1, old age and having gone through menopause^[66]. Acquired risk factors include elevated blood pressure, high levels of LDL cholesterol and triglycerides, cigarette smoking, high stress levels, excessive alcohol consumption, diabetes type 2, an inactive lifestyle and obesity^[66].

2.8.3 Coronary Atherosclerosis

Coronary Atherosclerosis (CA), also termed coronary heart disease, involves a narrowing or blockage of the coronary arteries, which provide the heart with blood^[67]. The narrowing or blockage of the blood vessel is caused by inflammation and the formation of plaques. These plaques are built up by calcium, cholesterol and other fatty substances, and gradually expand and harden over time, a process called atherosclerosis^[67]. As mentioned previously, CA is a risk factor for MI, because of the reduced levels of blood, oxygen and other nutrients received by the heart due to these plaques. Risk factors for CA include high levels of LDL cholesterol and triglycerides, smoking, hypertension, an inactive lifestyle, obesity, diabetes, a diet high in saturated fat and a family history of heart disease^[67]. Thus, MI and CA share many of the same risk factors, which is as expected since CA itself is a risk factor for MI.

2.8.4 Essential Hypertension

Essential hypertension (EH) indicates an elevated level of blood pressure. This is strongly correlated with an increased risk for certain cardiovascular diseases, such as MI and heart failure^[68]. The blood pressure increases when the heart pumps more blood and the arteries narrow. Essential, or primary, hypertension means that there is no clear cause of the high blood pressure. Secondary hypertension, however, occurs due to an underlying condition, such as severe sleep apnea or kidney disease^[68]. Risk factors for EH include old age, obesity, an inactive lifestyle, family history of high blood pressure, diets high in salt and low in potassium, excessive amounts of alcohol and tobacco and high stress levels. Also, it seems that individuals of African heritage have a higher risk of developing high blood pressure and the severe complications that may follow^[68]. Thus, EH shares many of its risk factors with MI. As was the case with CA, EH is also a potential risk factor for MI, and this is therefore as expected.

Methods

This chapter concerns the materials and procedures used to obtain the results presented in this thesis. Section 3.1 is taken from the specialization project in TBT4500 [19]. Section 3.2 is based on the specialization project in TBT4500 as well, but this and the remaining sections of the chapter were produced solely for this thesis. A new network was generated, the gene-phenotype network (GPN), where phenotypes and genes are connected if the phenotype is associated with a SNP in that gene. From this network, the gene-phenotype-phenotype network (GPPN) was made, where phenotypes are connected through SNPs in common genes. Section 3.4 and 3.5 concern the main topic of this thesis, the use of HUNT data to perform a polygenic risk score analysis for the diseases angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension.

3.1 The UK Biobank Dataset

Section 3.1 is taken from the specialization project in TBT4500 [19]. The PheWeb dataset obtained from the UK Biobank contains genome-wide associations of International Classification of Diseases (ICD) billing codes. The ICD billing codes have been obtained from the Electronic Health Record (EHR), and concerns white, British citizens in the UK Biobank [69]. The ICD is the standard classification system of medical conditions [70]. A couple of modifications were applied to the PheWeb-dataset. The phenotype "eating disorder" had two different phenocodes, 305.0 and 305.2, and the association with phenocode 305.2 was therefore removed. In addition to this, 1,173 rows did not contain a SNP database (dbSNP) number, thus "chr/pos_ref_alt" were inserted into these cells instead. The PheWeb dataset only covers binary traits and diseases, which means that an individual either has the phenotype in question, or not [71]. The characteristics of the PheWeb dataset is displayed in Table 3.1.1.

In a Phenome-Wide Association Study (PheWAS), a GWAS with millions of variants is analyzed against thousands of phenotypes [72]. For binary traits, the number of individuals with the trait, cases, are often quite low compared to the number of individuals without the trait, controls. The case-control ratio can be low, 1:10, or extremely low, 1:100 [72]. This causes higher rates of type I errors when using linear mixed models and logistic mixed models, because these are not made to handle such unbalanced case-control ratios. Confounding factors such as population

Parameter	PheWeb	Lee Lab
Entries	21,532	1,403
Columns	16	7
P-value threshold	10^{-6}	10^{-6}
Unique phenocodes	1,397	1,403
Analysis method	SAIGE	SAIGE
Disease Categories	17	17
Unique SNPs	19,183	-

Table 3.1.1: Characteristics of the PheWeb dataset from the UK Biobank and the phenotype dataset from the Lee Lab.

substructure and relatedness also have a negative impact by increasing the number of type I errors [72].

Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) is a method developed to operate with binary phenotypes having unbalanced case-control ratios and relatedness within the population [72]. SAIGE uses a Saddlepoint Approximation (out of scope for this thesis) to calibrate the unbalanced case-control ratio. It is the only mixed association model found which can both handle unbalanced case-control ratios and large sample sizes [72].

Since unbalanced case-control ratios often are present in PheWAS, the UK Biobank dataset used in this project is analyzed using SAIGE [72]. Certain other logistic mixed models, like the generalized mixed model association test, tend to cause an inflation of type I errors under unbalanced case-control studies, especially for rare genetic variants [72]. These models also struggle to adjust p-values under population stratification. In addition to this, it is evident that SAIGE provides unbalanced case-control studies with a higher statistical power than the other models do [72].

An additional dataset from the Lee Lab was utilized to obtain information regarding the phenotypes in the PheWeb dataset. This dataset was constructed by analyzing phenotypes from the PheWeb dataset against 28 million imputed genetic variants [73]. Its characteristics are shown in Table 3.1.1.

3.2 The SNP-Phenotype- and Phenotype-Phenotype Network

The SNP-phenotype (SPN)- and phenotype-phenotype (PPN) networks were constructed for the specialization project in TBT4500 [19]. They were both constructed using the PheWeb dataset described above. In this master thesis, linkage disequilibrium (LD) between SNPs has been added as an additional element to the SPN. The SNPs of the PheWeb dataset were analyzed in R to detect LD, utilizing the function *LDmatrix* in the R package *LDlinkR* [74]. *LDmatrix* calculates the R^2 -score for each pair of SNPs. This was performed separately for each chromosome, since only SNPs located on the same chromosome can be in LD. If the R^2 -score was found to be above 0.8, the two SNPs were assumed to be in LD. The R^2 -score is the correlation coefficient between two variables, where one variable indicates the presence or absence of a particular allele at one chromosomal locus, while the other variable indicates the presence or absence of the allele at another locus [75]. In addition to this, SNPs were not considered to be in LD if they

were further apart than 500,000 bp. The distance between two SNPs was calculated by using the chromosomal position of each SNP.

The function *LDmatrix* only locates pairs of SNPs in LD. However, if SNP 1 and SNP 2 are in LD, and the same is the case for SNP 2 and SNP 3, then SNP 1 and SNP 3 are also probably in LD. Therefore, all SNPs in LD with each other, in addition to those that are in LD with common SNPs, were added to an LD block. SNPs were only added to an LD block if the average R^2 -score of the SNPs included was above 0.8. The resulting 225 LD blocks were then added to the PheWeb dataset by replacing the individual SNP names with the LD block they belong to. In this way, phenotypes are linked through common SNPs, but they are also linked if the associated SNPs are in the same LD block. In this thesis, the SPN was utilized further to choose the diseases to include in the PRS analysis, and to determine which SNPs to use in the PRS calculations for these diseases.

3.3 The Gene-Phenotype Network

For this thesis, the gene-phenotype network (GPN) was constructed, where a phenotype is connected to a gene if the phenotype is associated with a SNP located within that gene. From the GPN, the gene-phenotype-phenotype network (GPPN) was constructed, where phenotypes are linked when associated with SNPs in common genes. These networks were constructed using the PheWeb dataset described above, where LD is taken into consideration. The PheWeb dataset was slightly modified in R before the networks were constructed. The column "nearest genes" was utilized to detect in which genes each SNP was located, or which gene was known to be the closest. This information was then used to determine which phenotypes should be connected to which genes. There were 85 SNPs with several genes listed as their nearest gene, and these were therefore split into separate rows, such that a single row for each individual SNP was obtained. The PheWeb dataset contains associations with p-values below 10^{-6} , but for the construction of these networks, the data were filtered such that only those SNPs with p-values below the genome-wide significance threshold of $5 \cdot 10^{-8}$ were included. This was done to limit the size of the network.

For the GPPN, the clustering algorithm Louvain, described in Section [2.4.3](#), was used to group the phenotypes into 28 different clusters [\[48\]](#). The phenotypes are distributed such that similarities within clusters are larger than between clusters. Both the GPN and GPPN were visualized in Cytoscape 3.8.0. Node tables for disease categories and clusters were imported along with the network. This allows for disease classes to be easily distinguished and phenotypes to be distributed out to their respective clusters.

3.4 HUNT

The data utilized for the polygenic risk score (PRS) analysis were obtained from the HUNT Cloud, through the CVDPgX lab at the K. G. Jebsen Center for Genetic Epidemiology. A PRS analysis requires, as mentioned in section [2.6](#), both a base- and target dataset. The base data usually consist of GWAS summary statistics and the target data contain information regarding the individuals for which the PRSs are calculated for. For this thesis, the PheWeb dataset was used as the base data, while the HUNT data functioned as the target data.

Register	Unique individuals	Period
HNT	89,212	1987 - 2017
KUHR	75,392	2006 - 2016
COD	37,679	1984 - 2015

Table 3.4.1: The properties of the registers on the HUNT Cloud used in this thesis.

For the PRS calculations, information about the HUNT-participants was obtained from Helse Nord-Trøndelag (HNT), the Norway Control and Payment of Health Reimbursement (KUHR) and the Cause of Death (COD) registers. The properties of the registers are displayed in Table [3.4.1](#). From the HNT- and KUHR registers, information regarding the year of birth, sex, date of diagnosis, and ICD9- and ICD10 codes of the participants was used. For both cases and controls, only individuals that were genotyped and had consented that their information may be used were included in the PRS analysis. From the HNT register, only individuals who had the disease in question as their main diagnosis were used as cases. The COD register was used to determine which individuals had died and which were still alive. This was done by comparing the ID numbers obtained from HNT and KUHR with those in the COD register. Table [3.4.2](#) displays the number of cases and controls for each disease.

Disease	Cases	Controls
Angina Pectoris	6,469	62,954
Myocardial Infarction	5,739	63,684
Coronary Atherosclerosis	1,901	67,522
Essential Hypertension	4,090	65,333

Table 3.4.2: The number of cases and controls for each disease in the HUNT population.

3.5 Polygenic Risk Score Analysis

The polygenic risk score (PRS) calculations were performed for four cardiovascular diseases: Angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH). This was done using information from the PheWeb dataset, the SNP-phenotype network and the HUNT Cloud. The PRSs were calculated for HUNT participants.

3.5.1 Polygenic Risk Score Calculations

To calculate the PRSs, formula 2.6.1 was used. For each SNP considered to be associated with the disease, the number of SNPs present in an individual (0, 1 or 2) and the effect size of the SNP are required. The final PRS is obtained by summarizing the values for each SNP. The PRS calculations were performed using two different procedures. In the first one, the PheWeb dataset was used to detect SNPs found to be associated with the four diseases. These associations are known to have p-values below the significance level of 10^{-6} . The SNPs were then used to calculate the PRSs of HUNT participants, for each of the four diseases. In the second procedure, the PRSs were calculated using a larger number of SNPs. This was done because, as mentioned in Section 2.6, including a larger number of SNPs in the PRS calculations increases the prediction accuracy of the PRS [52]. To determine which SNPs to include, a network approach was used. This was done to ensure that the SNPs included in the PRS calculations have an association to the disease, such that the resulting PRSs provide a higher disease prediction accuracy than what would have been the case if SNPs were chosen at random.

In the SPN described in Section 3.2, SNPs and LD blocks are linked to a phenotype if an association has been found between them. The SPN was used to choose which SNPs to include in order to increase the number of SNPs used in the calculations. Both the SNPs directly linked to the diseases (first degree SNPs) and the SNPs linked to their neighbouring diseases (second degree SNPs), were then included. Table 3.5.1 shows the number of SNPs used in the calculations both when only first degree SNPs were used and when second degree SNPs were included. When second degree SNPs are included, there is a 14-fold increase in the number of SNPs for AP. MI has a 18-fold increase, CA a 27-fold increase, and EH a 15-fold increase in the number of SNPs. An illustration of how SNPs were chosen for AP using the two different procedures is shown in Figure 3.5.1 and 3.5.2. In Figure 3.5.1, only first degree SNPs are marked in yellow, while in Figure 3.5.2, both first- and second degree SNPs are marked. As can be observed, the number of SNPs increases substantially when second degree SNPs are included in the calculations.

Disease	First Degree SNPs	First- and Second Degree SNPs
Angina Pectoris	62	880
Myocardial Infarction	57	1,029
Coronary Atherosclerosis	36	987
Essential Hypertension	50	771

Table 3.5.1: The number of single nucleotide polymorphisms (SNPs) used in the two different procedures for calculating polygenic risk scores (PRSs), for the diseases angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH).

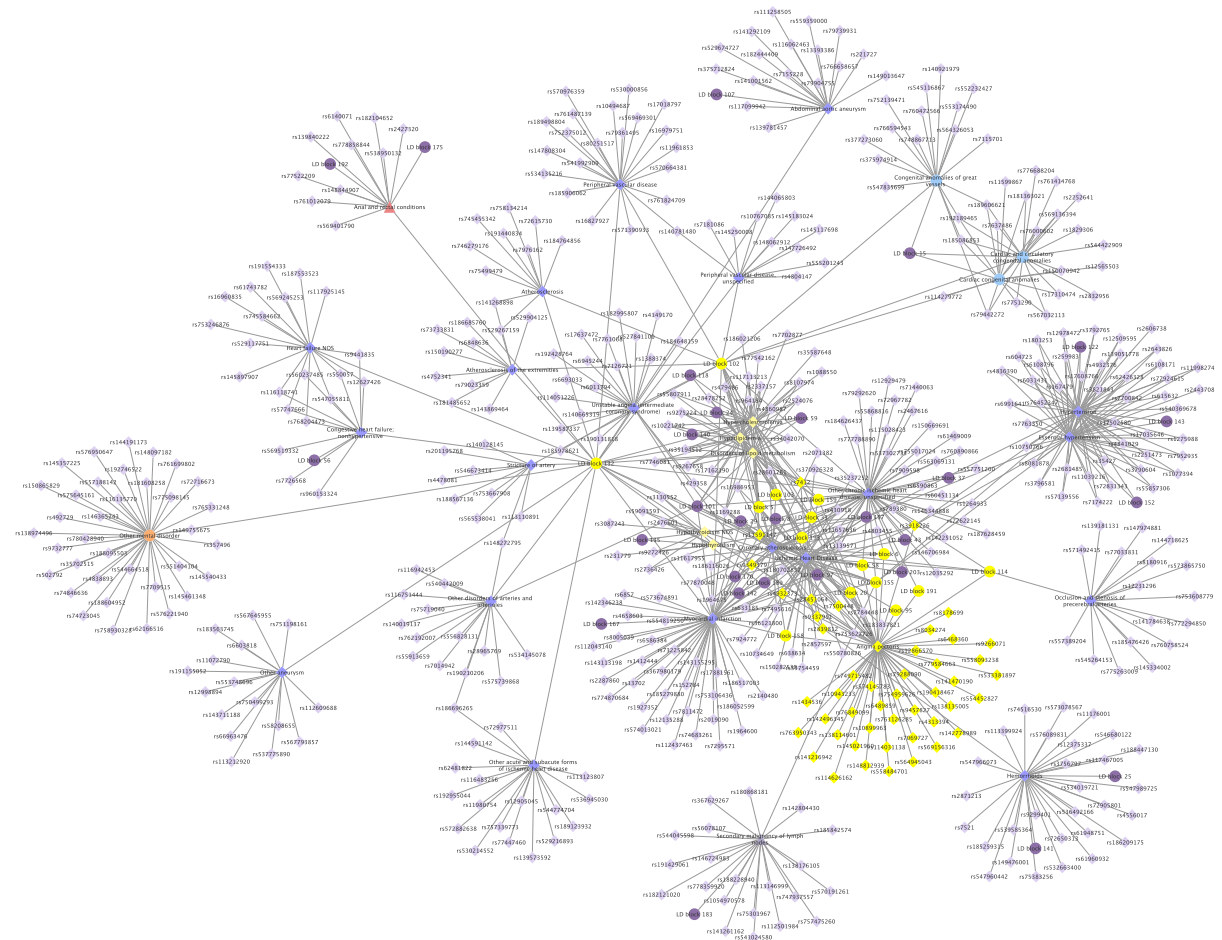


Figure 3.5.1: A network illustrating how first degree single nucleotide polymorphisms (SNPs) and linkage disequilibrium (LD) blocks were chosen for the polygenic risk score (PRS) calculations for angina pectoris (AP). SNPs are the light purple nodes, while LD blocks are the larger, dark purple nodes. First degree SNPs for AP are marked in yellow.

To obtain the correct information from the HUNT Cloud, lists containing the chromosome number- and position of the selected SNPs were run against the HUNT chromosome lists to detect individuals with the correct reference- and alternative allele at the right positions. A list of all SNPs associated with a particular disease was compared with the obtained data from the HUNT Cloud, to ensure that the reference- and alternative allele of the individuals found were the same as for the SNPs associated with the disease in the PheWeb dataset.

Information from HNT and KUHR, registers that were described in Section 2.7, were utilized to identify cases, which are the individuals diagnosed with the four different diseases. The ICD9- and ICD10 codes of AP, MI, CA and EH were run against the two databases to find individuals diagnosed with each of the four diseases. If the individuals found were also present in the list of individuals registered as having the associated SNPs, the diagnosis (0: No disease, 1: Disease) and date of diagnosis were added to the already obtained data. If the same individual was found in both the HNT- and KUHR database, the earliest date of diagnosis was added.

The individuals found were also run against the COD register. Information about whether each individual was dead or alive and the potential year of death were added to the already existing data. This was done for the calculation of age and cumulative disease risk, the calculation of

3.5.2 Visualization of Results

To analyze the PRSs of the four diseases, prevalence- and case-control plots were made. For the prevalence plot, the obtained PRSs were separated into 8.3 % quantiles based on the PRS distribution. The number of individuals within each quantile was plotted against the actual prevalence of the disease in that particular quantile. In this way, it can be observed how well the genetic risk, represented by the PRS, predicts the occurrence of a particular disease in the population. The case-control plots are displayed as box plots, where the PRSs are presented as percentages based on the PRS distribution. These plots show whether there is a difference between cases and controls in where their PRS lies in the PRS distribution.

For each disease, cumulative disease risk (CDR) was calculated for all ages to observe how the probability of developing the diseases changes over a lifetime. This was done solely for the PRSs calculated with both first- and second degree SNPs. A survival analysis involves an evaluation of the time period that goes by until an event occurs [76]. If several events might take place, competing risks should be considered. The function *cuminc*, in the R package *cmprsk* was used to perform the competing risk analysis [77]. In this case, the competing risks were; (i) the patient has the disease in question, (ii) the patient is alive without the disease, (iii) the patient has died without being diagnosed with the disease.

The competing risk analysis is performed because there may be a dependency between the time occurrence of these events [76]. Certain events might not have occurred at the time of data collection, and some events may not occur at all. Therefore, for this analysis, the event of a patient being alive without the disease is set as the censored value. This means that the patients for which this is the case, will not experience the possible events of death or diagnosis of the disease before the time of data collection [76]. PRSs were separated into 20 % quantiles, and the CDR was calculated for each of these groups separately. A CDR was given for each age, from the lowest age registered in the HUNT data, to the highest. The disease risk was given as the probability of developing the disease. When constructing the plots, each line represents a PRS quantile, such that the effect of both age and genetic risk on the CDR can be evaluated.

A statistical analysis was conducted both for PRSs calculated with only first degree SNPs and first- and second degree SNPs. Logistic regression is advantageous when a binary outcome is to be predicted, and in this case, the binary outcome is to either develop the disease, or not to [31]. The function *glm* in the R package *stats* was therefore used to fit a logistic regression model, with the covariates sex, age, the first 10 principal components to adjust for population stratification and whether the PRS is in the top x % of the distribution or in the reference group [78]. The top 20-, 10-, 5- and 1 % highest PRSs were analyzed against a reference group consisting of the remaining part of the distribution. The tables presented in the results show the odds ratios (ORs) for developing the diseases with a PRS in the top 20-, 10-, 5- or 1 % of the PRS distribution, compared to in the reference group. The confidence interval (CI) and p-value of each OR are also provided to show the significance of the estimate.

Results and Analysis

This chapter contains a description of the results obtained during this master thesis. Section [4.1](#) and [4.2](#) concern the SNP-phenotype network (SPN) and the gene-phenotype-phenotype network (GPPN), respectively, while section [4.3](#) regards a comparison between the GPPN and the human disease network (HDN). Section [4.4](#) concerns the polygenic risk score (PRS) analysis, which was conducted for the diseases angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension. The SPN was used to determine which SNPs to include in the PRS calculations, to discover whether a network approach improves the prediction accuracy of the PRS. The PRSs were calculated for participants of the HUNT study.

4.1 SNP-Phenotype Network

The specialization project in TBT4500 concerned a network analysis of the SNP-phenotype network (SPN), where phenotypes are connected to SNPs if an association has been found between them. The SPN was also utilized in this thesis, however, some alterations were made. Linkage disequilibrium (LD) between SNPs was taken into consideration, and SNPs in high LD were added to the same LD block. SNPs in LD were thus represented by their LD block in the network, instead of their SNP database number. The SPN was used to determine which SNPs and LD blocks to include in the PRS calculations. A part of the SPN is displayed in Figure [4.1.1](#), where the diseases angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH) are marked in yellow. As described in Section [3.5](#), the PRSs for the four diseases were first calculated using SNPs and LD blocks directly linked to the diseases in the SPN. Then, the PRSs were calculated again, using both the SNPs and LD blocks directly linked to the diseases in question, and the SNPs linked to their neighbouring diseases.

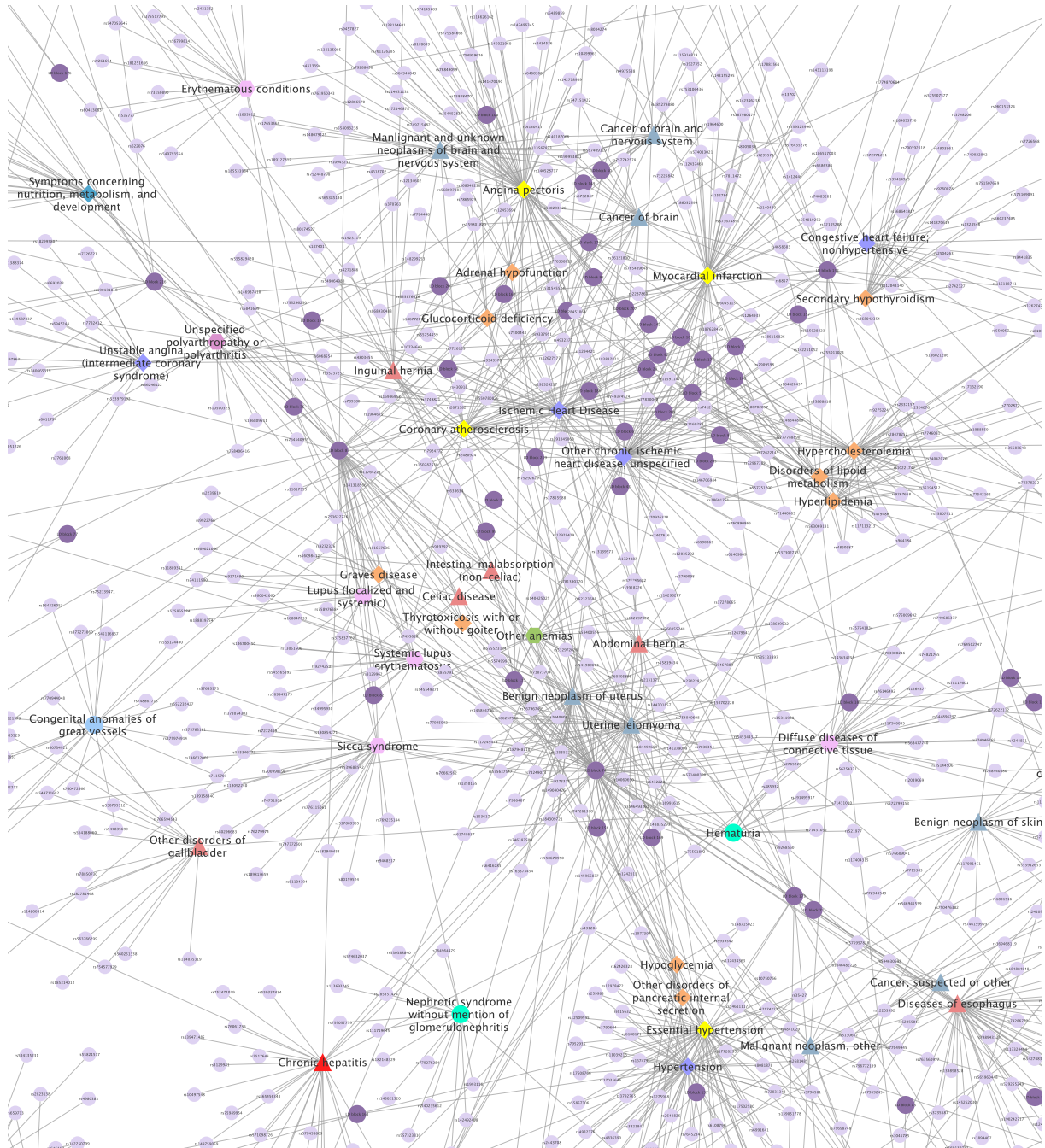


Figure 4.1.1: A part of the SNP-phenotype network which shows the diseases for which the polygenic risk score (PRS) calculations were performed; angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH). These four diseases are marked in yellow. The light purple nodes are single nucleotide polymorphisms (SNPs) and the dark purple nodes are linkage disequilibrium (LD) blocks.

4.2 Gene-Phenotype and Phenotype-Phenotype Network

In the specialization project in TBT4500, a comparison was made between the phenotype-phenotype network (PPN), where phenotypes are linked when associated with common SNPs, and the human disease network (HDN) described in Section 2.5. For this thesis, a network where phenotypes are connected through common genes was constructed for the purpose of

comparing it to the HDN. In the HDN, diseases are connected if associated with a mutation within the same gene.

As the SPN and PPN, the gene-phenotype network (GPN) was built using the PheWeb dataset from the UK Biobank, where LD is taken into consideration. In the GPN, a phenotype is linked to a gene if the phenotype is associated with a SNP in that gene, or an LD block containing a SNP positioned within the gene. From the GPN, the gene-phenotype-phenotype network (GPPN) was created, where phenotypes are connected if associated with a SNP, or an LD block containing a SNP, located within the same gene. The phenotypes in the GPPN were distributed into clusters by using the clustering algorithm Louvain, and the network contains solely associations with p-values below the genome-wide significance threshold of 10^{-8} . The procedure for the construction of the GPN and GPPN is described in section 3.3, and the GPPN is displayed in Figure 4.2.1.

The GPPN consists of 283 nodes and 1,818 edges. Of these 283 nodes, 252 of them have two or more connections. The largest cluster of the network consists of 42 nodes, and is made up of mainly mental disorders, circulatory system- and digestive diseases. While there are 31 nodes with a degree of 1, there is only one node with a degree of 76, which is the highest degree found for any node in the network. There are generally few high degree nodes, hubs, and numerous low degree nodes, which is characteristic of a scale-free network.

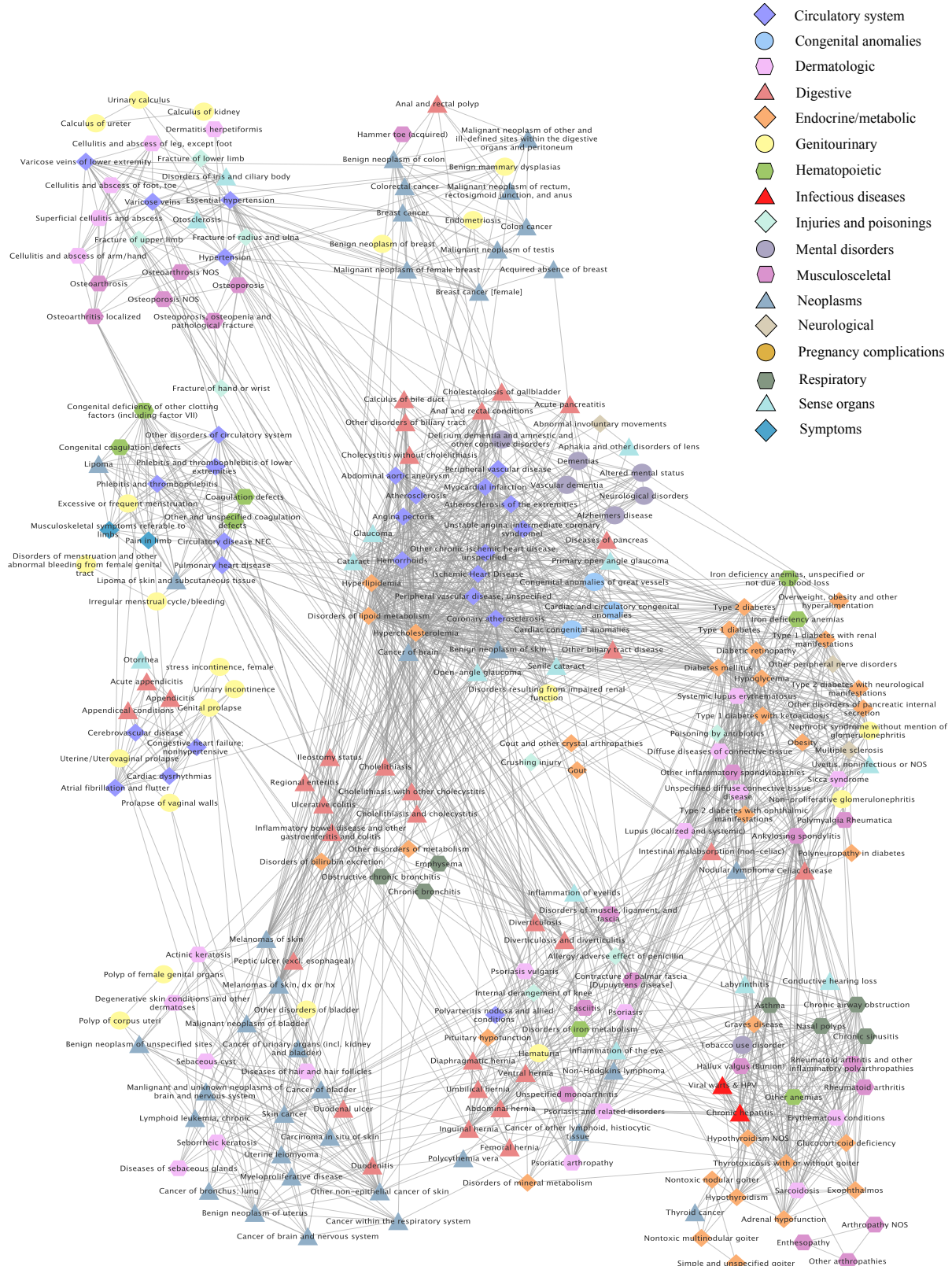


Figure 4.2.1: The gene-phenotype-phenotype network (GPPN), where phenotypes are connected if associated with a single nucleotide polymorphism (SNP) or linkage disequilibrium (LD) block within the same gene. The diseases are classified according to the disease categories displayed in the table.

4.3 Comparison of Networks

The human disease network (HDN) was described in Section 2.5 of the theory. The network is made up of nodes representing diseases, and these are connected to each other if associated with a mutation within the same gene. The HDN is built using data from the OMIM database. The two networks were compared to detect any potential similarities or differences in clustering patterns, disease connections or general structure. However, since the networks are built using different databases, there are disparities between the diseases included in the two networks. Figure A.2.1 in Appendix A.2 shows the HDN with both diseases and genes.

In the HDN, the degree of each node can be determined by its size, since the size of each node represents the number of genes that node is connected to. A large number of the highest degree nodes in the HDN are also present in the GPPN. Among them are the nodes "colon cancer", "breast cancer", "diabetes mellitus", "obesity", "asthma", "Alzheimer's disease" and "myocardial infarction". The nodes "breast cancer" and "colon cancer" are contained within a cluster of the HDN consisting mainly of cancers, while they are allocated between two different clusters in the GPPN. Both of these clusters predominantly contain cancers, however, the diseases and connections within the clusters often differ between the two networks.

The node "leukemia" has the highest degree in the HDN, while this disease is not present in the GPPN. "Lymphoma", another high degree node in the HDN, is present in the form of "nodular lymphoma" and "non-Hodgkins lymphoma" in the GPPN, where these are allocated to two different clusters. Since "non-Hodgkins lymphoma" also is present in the HDN, it can be assumed that the node "lymphoma" indicates Hodgkins lymphoma. Nodular lymphoma, or nodular lymphocyte-predominant Hodgkin lymphoma, is a type of Hodgkins lymphoma [79]. Nodular lymphoma is associated with skin cancer in the GPPN, while lymphoma is associated with breast- and colon cancer in the HDN. These nodes are thus not involved in any common connections.

The link between "hypothyroidism" and "Graves disease" is present in both networks. This is rather surprising, since Graves disease has been found to cause an overactive thyroid gland, which is defined as hyperthyroidism [80]. A link between Graves disease and hyperthyroidism would therefore be expected instead, but this connection does not appear in either network. In the HDN, hypothyroidism is linked to hyperthyroidism through another gene, while hyperthyroidism does not appear at all in the GPPN. The node "diabetes mellitus" is involved in several common connections in the two networks, among them the link between "diabetes mellitus" and "obesity". Studies have shown associations between obesity and both type 1- and type 2 diabetes, but especially type 2 diabetes and obesity seem to be highly connected [81]. The majority of patients with type 2 diabetes are obese, and the insurgence of obesity over the past years is believed to have caused the equivalent increase in diabetes type 2 patients [82]. The node "diabetes mellitus" is also connected to "myocardial infarction" in both networks, which is logical considering that diabetes mellitus has been shown to increase the risk for certain cardiovascular diseases, such as myocardial infarction and coronary atherosclerosis [83].

The HDN appears to contain a greater number of syndromes, which make up a small fraction of the nodes in the GPPN. The diseases are also categorized somewhat differently, which makes comparing the two networks more difficult. In the GPPN, there is a larger number of circulatory system diseases, and many of these are found in the same cluster and are for the most part connected. This is also the case for the HDN, however, there are fewer circulatory system

diseases included here. Neurological diseases are mostly found within the same cluster in the GPPN, while they are more widely distributed throughout the HDN. Digestive diseases are found in numerous clusters of the GPPN, while these are not included at all in the HDN. Thus, the diseases, clustering pattern and structure of the two networks mostly differ.

4.4 Polygenic Risk Score Analysis

The polygenic risk score (PRS) analysis was performed for four diseases: Angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH). The PRSs were calculated for participants of the HUNT study, and the information required for the calculations was obtained from the HUNT data. The SNP-phenotype network (SPN) was utilized to determine which SNPs to include in the calculations.

4.4.1 Prevalence- and Case-Control Plots

Prevalence plots were generated to display the relation between the genetic disease risk of HUNT participants, represented by the PRSs, and the actual prevalence of the disease in the HUNT population. For each of the four diseases, the prevalence plots were first generated for the PRSs calculated with only first degree SNPs, and then for the calculations with both first- and second degree SNPs. As mentioned in Section 2.6, including a larger number of SNPs in the PRS calculations increases the prediction accuracy of the PRS. For this analysis, the SPN was used to determine which SNPs to include, such that the SNPs used in the calculations have stronger associations with the disease compared to what randomly chosen SNPs would have had.

If the PRS actually predicts the occurrence of disease, then individuals with a higher PRS would have a higher genetic risk for developing the disease. Thus, the percentage of individuals diagnosed with the disease within each PRS quantile, should increase with an increasing PRS percentage. The PRS percentages are calculated based on the distribution of PRSs in the population, for each disease. The case-control plots show whether there is a distinct difference between where the PRSs of cases and controls lie in the PRS distribution.

Angina Pectoris

The prevalence- and case-control plots for angina pectoris (AP) are displayed in Figure 4.4.1 and 4.4.2. As can be observed, there was a decreasing tendency of disease prevalence in Figure 4.4.1a, while in Figure 4.4.1b, where both first- and second degree SNPs were considered, there was a tendency of a rise in prevalence with an increasing PRS percentage. However, there were considerable variations in prevalence, and no steady increase was observed. In the case-control plots, cases tended to have somewhat lower PRS percentages than controls in Figure 4.4.2a, while cases had a higher PRS percentage than controls in Figure 4.4.2b. The PRSs calculated with both first- and second degree SNPs therefore seemed to be more predictive of actual disease occurrence.

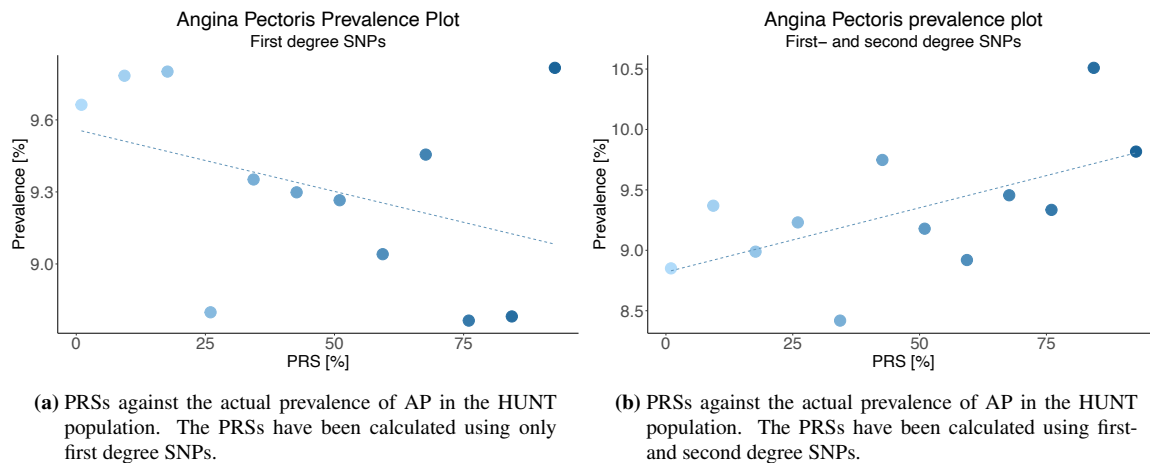


Figure 4.4.1: The prevalence plots for angina pectoris (AP) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The disease prevalence of AP is displayed as the percentage of individuals in that particular PRS quantile with the disease. The PRSs are shown as 8.3 % quantiles based on the PRS distribution for AP in the HUNT population.

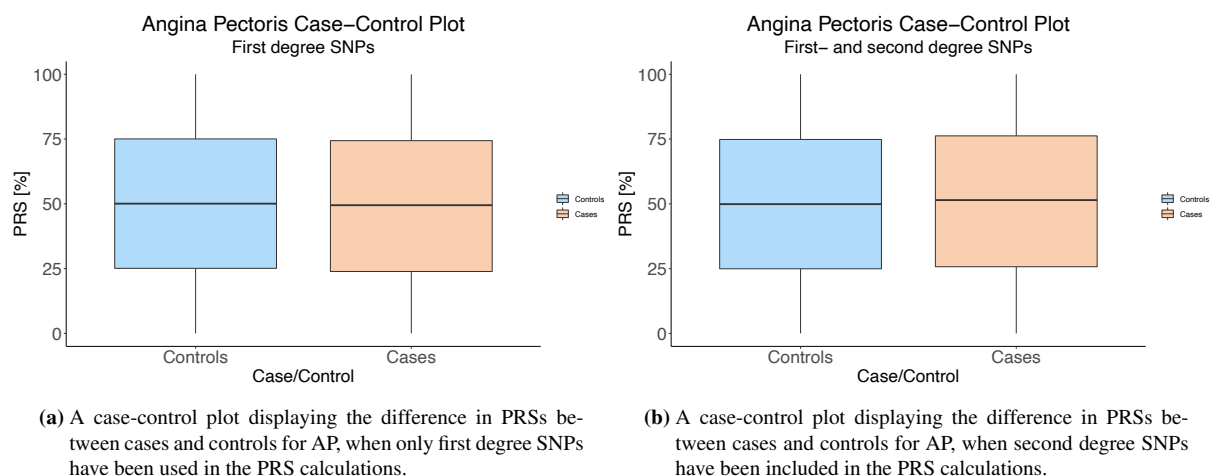


Figure 4.4.2: The case-control plots for angina pectoris (AP) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The PRSs are shown as percentages based on the PRS distribution for AP in the HUNT population.

Myocardial Infarction

The prevalence- and case-control plots for MI are displayed in Figure 4.4.3 and 4.4.4. In both Figure 4.4.3a and 4.4.3b, disease prevalence showed a tendency to increase with an increasing PRS percentage, however, the tendency was somewhat larger in the latter. Also, the PRS percentages were slightly higher when second degree SNPs were included, however, both plots showed a high instability of prevalence. In the case-control plots, cases had a slightly higher PRS percentage than controls, both in Figure 4.4.4a and 4.4.4b. Thus, it seems that the difference in PRS prediction accuracy when using a larger number of SNPs, was smaller for MI than what was the case for AP.

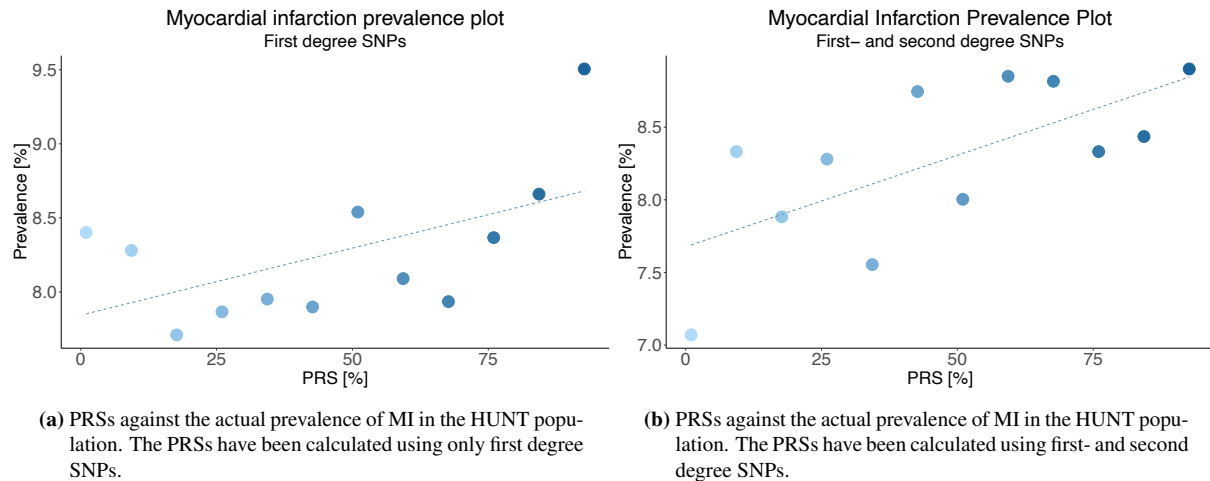


Figure 4.4.3: The prevalence plots for myocardial infarction (MI) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The disease prevalence of MI is displayed as the percentage of individuals in that particular PRS quantile with the disease. The PRSs are displayed as 8.3 % quantiles based on the PRS distribution for MI in the HUNT population.

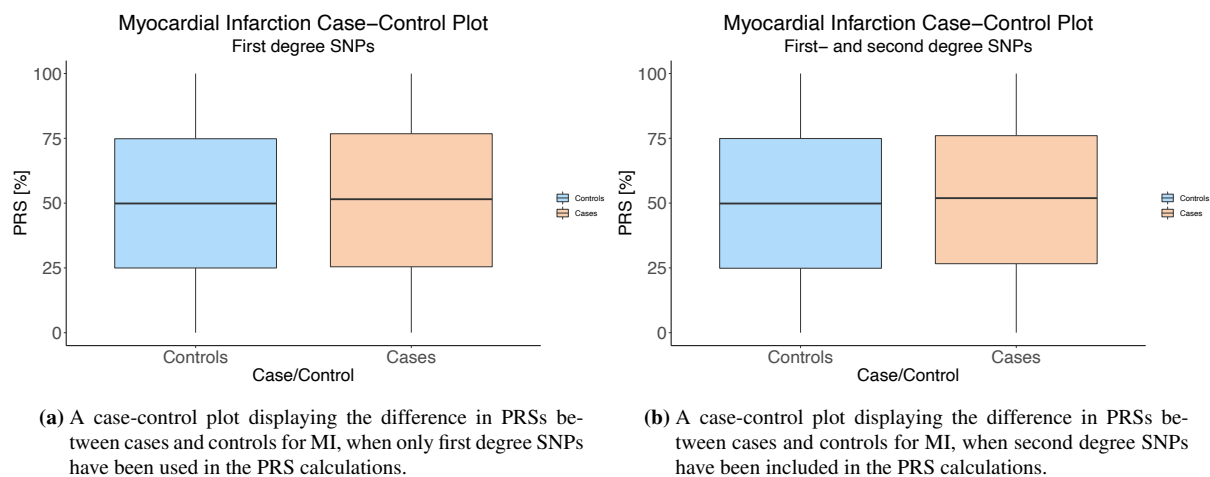


Figure 4.4.4: The case-control plots for myocardial infarction (MI) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The PRSs are shown as percentages based on the PRS distribution for MI in the HUNT population.

Coronary Atherosclerosis

The prevalence- and case-control plots for CA are shown in Figure 4.4.5 and 4.4.6. Neither the plot in Figure 4.4.5a nor 4.4.5b showed a rise in disease prevalence with an increasing PRS percentage. There are quite large variations in both plots, and there seems to be a deterioration of prediction accuracy when including additional SNPs in the PRS calculations, considering that there is a decreasing tendency of disease prevalence. The case-control plot in Figure 4.4.6a, where only first degree SNPs were used in the calculations, showed a higher PRS percentage for cases than controls. In Figure 4.4.6b, where second degree SNPs were included, controls had a slightly higher PRS percentage than cases. Thus, there was no improvement in PRS prediction accuracy when including second degree SNPs, rather, the prediction ability of the PRSs seemed to deteriorate.

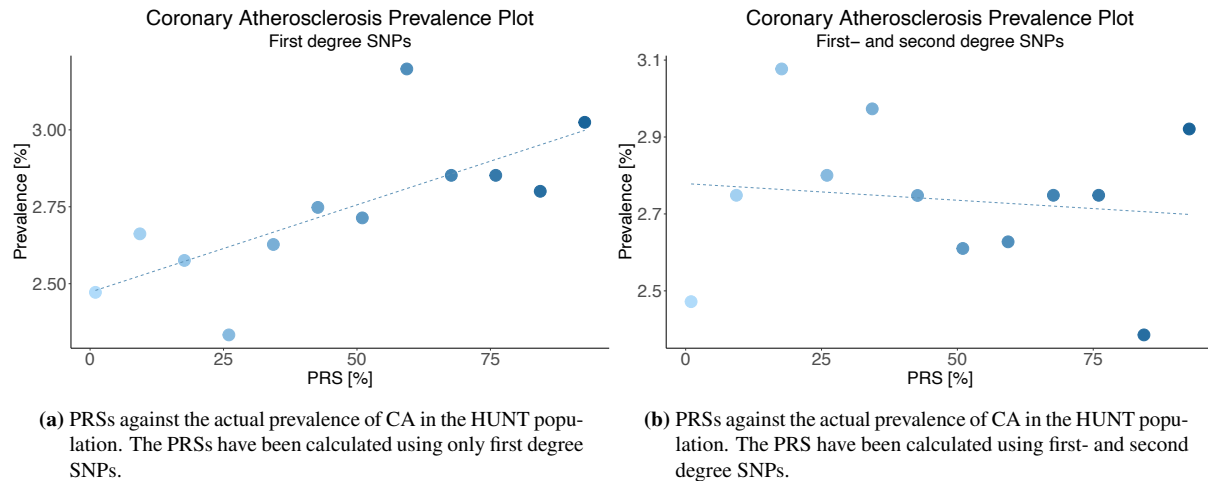


Figure 4.4.5: The prevalence plots for coronary atherosclerosis (CA) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The disease prevalence of CA is displayed as the percentage of individuals in that particular PRS quantile with the disease. The PRSs are displayed as 8.3 % quantiles based on the PRS distribution for CA in the HUNT population.

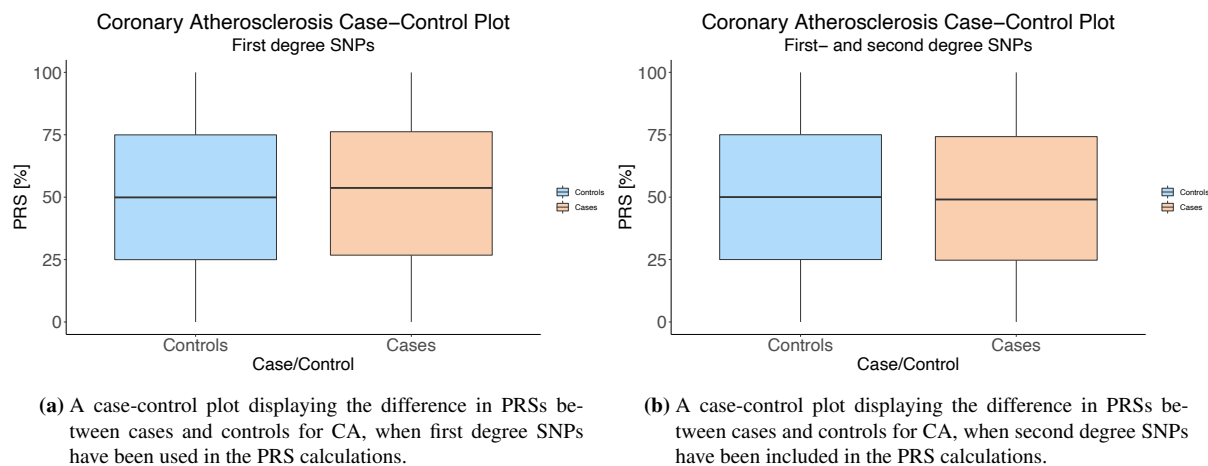


Figure 4.4.6: The case-control plots for coronary atherosclerosis (CA) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The PRSs are shown as percentages based on the PRS distribution for CA in the HUNT population.

Essential Hypertension

The prevalence- and case-control plots for EH are shown in Figure 4.4.7 and 4.4.8. The prevalence plot in Figure 4.4.7a showed a decreasing tendency of prevalence with an increasing PRS percentage. There was an apparent improvement in disease prediction accuracy when including a greater number of SNPs in Figure 4.4.7b, where the disease prevalence tended to rise with an increasing PRS percentage. However, there were quite large variations in both plots. In the case-control plot in Figure 4.4.8b, where both first- and second degree SNPs were included, cases had a higher PRS percentage than controls. This is not the case for Figure 4.4.8a. Thus, there seemed to be an improvement in disease prediction accuracy for EH when including a larger number of SNPs.

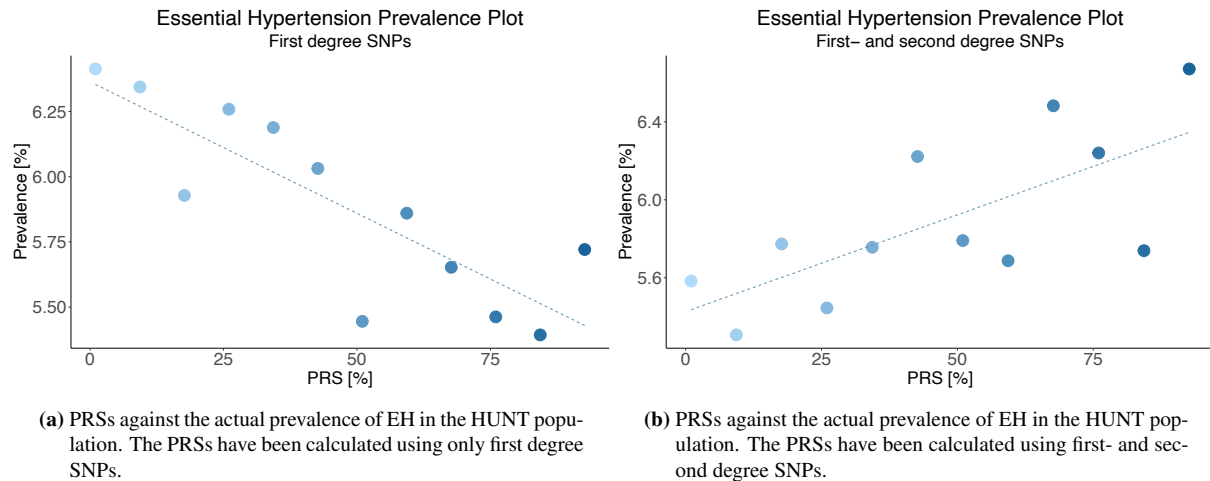


Figure 4.4.7: The prevalence plots for essential hypertension (EH) when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the polygenic risk score (PRS) calculations, and **b)** when both first- and second degree SNPs are included. The disease prevalence of EH is displayed as the percentage of individuals in that particular PRS quantile with the disease. The PRSs are displayed as 8.3 % quantiles based on the PRS distribution for EH in the HUNT population.

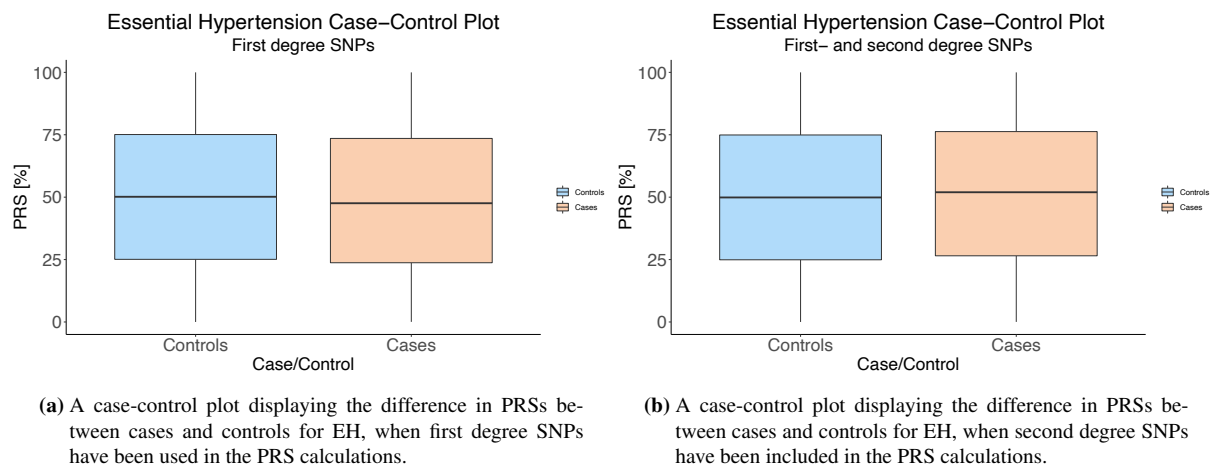


Figure 4.4.8: The case-control plots for essential hypertension when **a)** only first degree single nucleotide polymorphisms (SNPs) are included in the PRS calculations, and **b)** when both first- and second degree SNPs are included. The PRSs are shown as percentages based on the PRS distribution for EH in HUNT population.

4.4.2 Cumulative Disease Risk Plots

Plots relating age and cumulative disease risk (CDR) for different PRS groups were generated to analyze how the probability of developing the diseases changes over a lifetime. In this way, it can be observed how much of the disease risk is caused by age and how much is due to genetic risk. The PRSs were divided into 20 % quantiles based on the distribution of PRSs. The CDR is given as the probability of developing the disease, and the procedure for calculating this risk was described in Section [3.5.2](#). The procedure involved conducting a competing risk analysis for each PRS quantile. This was performed solely for the PRSs calculated with both first- and second degree SNPs, since these PRSs were expected to be the most predictive of actual disease risk.

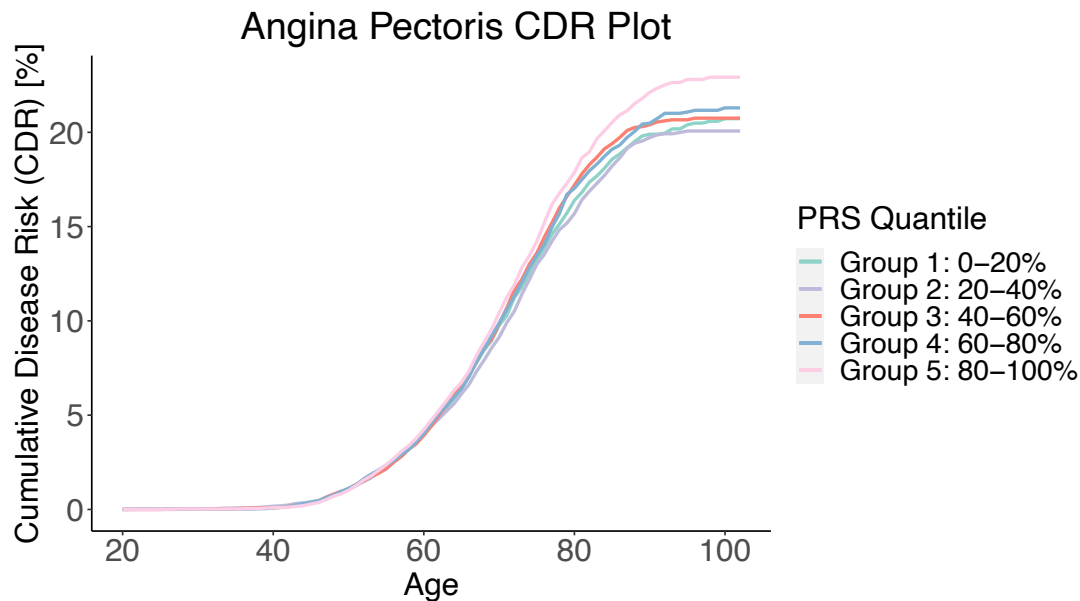


Figure 4.4.9: The plot shows how the cumulative disease risk (CDR) for angina pectoris (AP) changes over a lifetime. The patients are divided into 20 % polygenic risk score (PRS) quantiles, where the PRS is given as a percentage based on the PRS distribution for AP in the HUNT population.

The CDR plot for AP is shown in Figure [4.4.9](#). As can be observed, the CDR for all five PRS groups starts increasing around the age of 40. At the maximum age of 102, group 5 has a 1.14-fold higher probability of developing AP than group 2, which has the lowest CDR. Group 5 reaches a 5 % probability at the age of 61, and a 20 % probability at 83. Group 2 has the lowest CDR from the age of 60 and onwards, and reaches a 5 % probability at the age of 61 and a 20 % probability at 88. The slopes of the various groups differ slightly, especially from the age of 70 and onwards, with the slope of group 5 being the steepest. This implies that for individuals with a PRS among the 80-100 % highest scores, the CDR of developing AP increases more rapidly with aging, compared to individuals with lower PRSs.

The CDR plot for MI is shown in Figure [4.4.10](#). The CDR for group 5, with the 80-100 % highest PRSs, begins to increase slightly before the other groups. However, at the maximum age of 103, group 4 has a 1.14-fold higher CDR than group 1, which has the lowest. Group 4 reaches a CDR of 10 % at the age of 74, while it reaches 20 % at 88. Group 1 reaches a CDR of 10 % at the age of 75, while it reaches 20 % at 91. From the age of 70, there seems to be a larger difference between the slopes of the groups, and this is where the groups with higher PRSs start to deviate from those with lower scores. It can be observed that the slopes of groups 3, 4 and 5 are slightly steeper than the remaining groups. The disease risk for individuals with PRSs in these groups therefore increases more rapidly with age.

The CDR plot for CA is shown in Figure [4.4.11](#). In this case, group 2 has considerably higher CDRs than the other groups from the age of 40 and onwards. At the maximum age of 103, group 2 has 1.12-fold higher CDR than group 1, which has the lowest risk. Group 2 reaches the CDR of 2 % at the age of 65, while it reaches 6 % at 82. Group 1 reaches the probability of 2 % at the age of 68, while it reaches 6 % at 89. The slopes of all groups seem to be quite similar, except for group 2, which is somewhat steeper. This implies that the CDR for developing CA

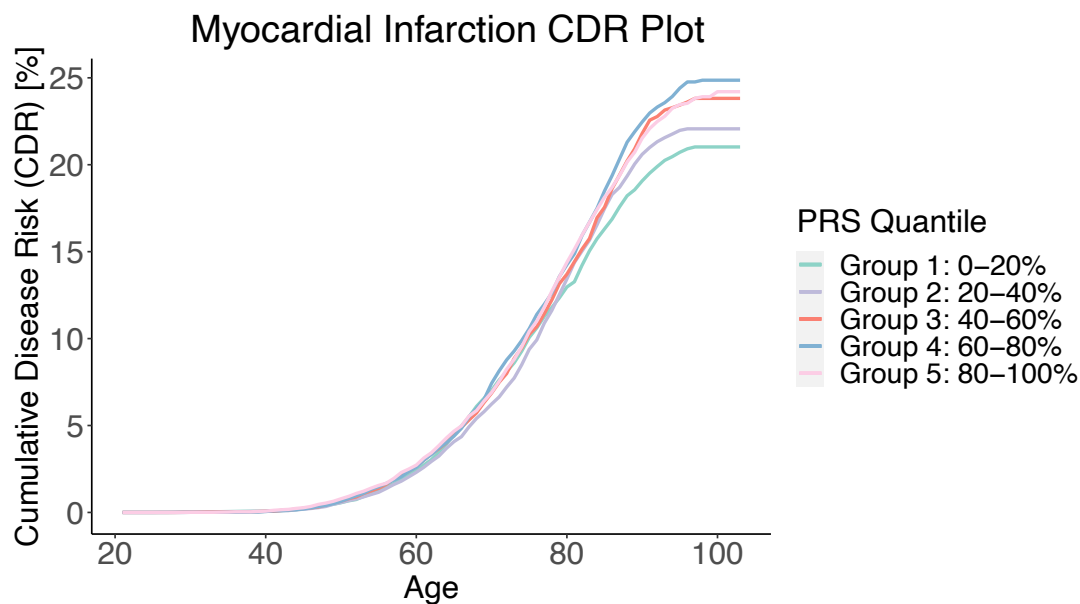


Figure 4.4.10: The plot shows how the cumulative disease risk (CDR) for myocardial infarction (MI) changes over a lifetime. The patients are divided into 20 % polygenic risk score (PRS) quantiles, where the PRS is given as a percentage based on the PRS distribution for MI in the HUNT population.

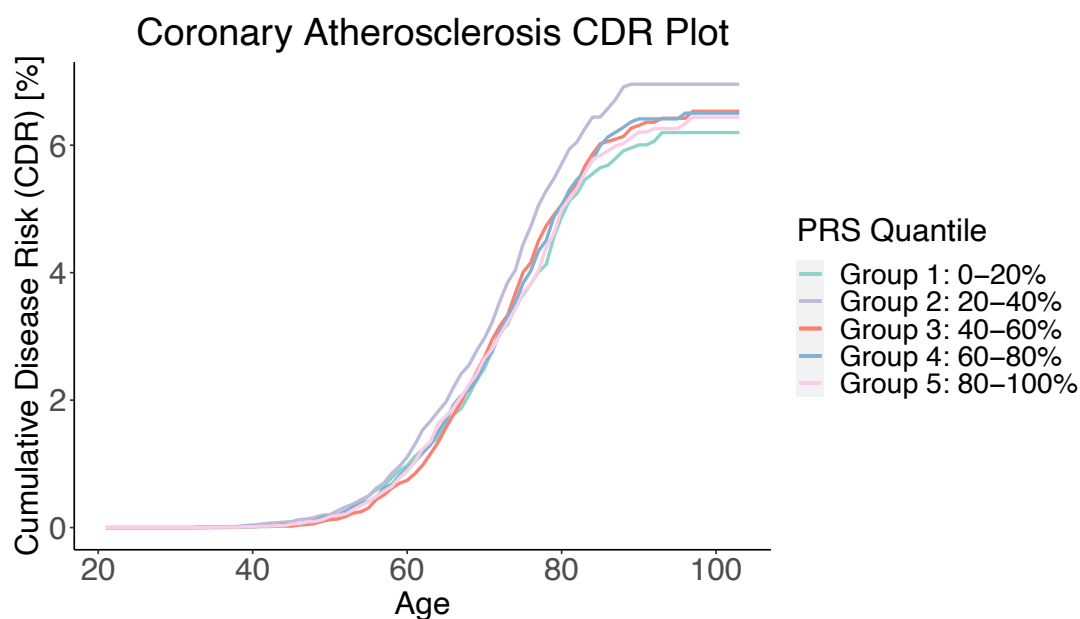


Figure 4.4.11: The plot shows how the cumulative disease risk (CDR) for coronary atherosclerosis (CA) changes over a lifetime. The patients are divided into 20 % polygenic risk score (PRS) quantiles, where the PRS is given as a percentage based on the PRS distribution for CA in the HUNT population.

increases more rapidly for individuals with a PRS among the 20-40 % lowest scores of the population.

The CDR plot for EH is shown in Figure [4.4.12](#). Group 4 and 5 are quite similar at all ages,

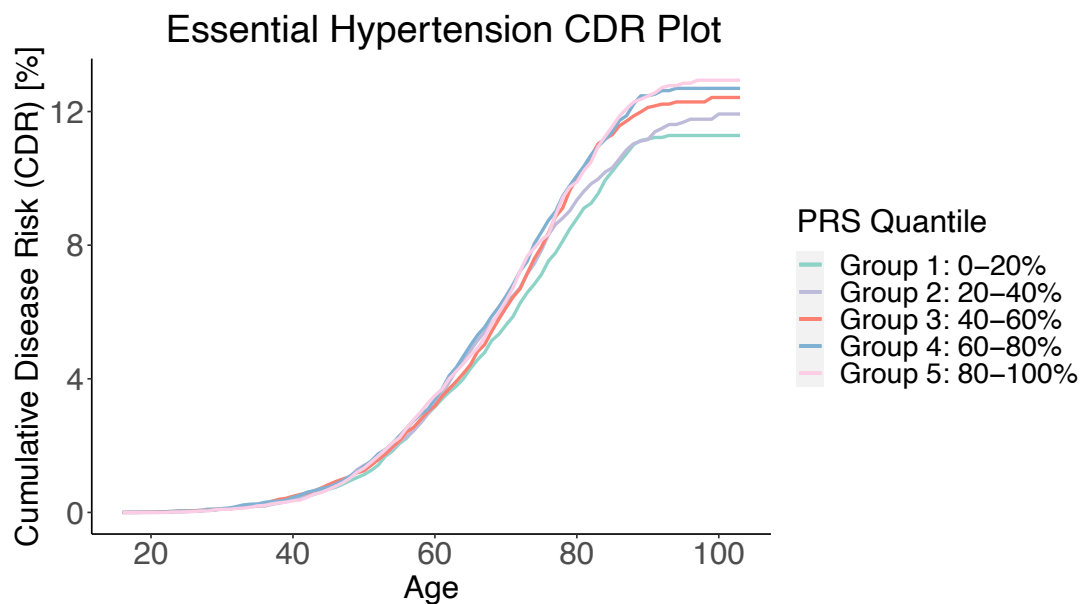


Figure 4.4.12: The plot shows how the cumulative disease risk (CDR) for essential hypertension (EH) changes over a lifetime. The patients are divided into 20 % polygenic risk score (PRS) quantiles, where the PRS is given as a percentage based on the PRS distribution for EH in the HUNT population.

except for at the maximum age of 103, where group 5 has a slightly higher CDR. Group 5 reaches a probability of 5 % at the age of 66, while it reaches a probability of 10 % at 79. Group 1 reaches a probability of 5 % at the age of 66, while it reaches a probability of 10 % at 85. The CDR for all five groups starts increasing quite uniformly at the age of 40, while the slopes starts to differ from the age of 60 and onwards. Group 3, 4 and 5 have the steepest slopes, which implies that the cumulative risk for EH increases more rapidly with age for individuals with higher PRSs. These results are similar to those of AP and MI.

4.4.3 Statistical Analysis

The tables below provide the odds ratios (ORs) that individuals with PRSs among the top x % highest scores of the population acquire the disease, compared with a PRS in the reference group. The reference group consists of those individuals with the $(1 - x)$ % lowest scores. This analysis is performed for the top 20-, 10-, 5- and 1 % highest PRSs in the population, for each of the four diseases; AP, MI, CA and EH. The tables provide the 95 % confidence intervals (CIs) and p-values of each obtained OR, and the significance level is set to 5 %. If a CI contain the null value, which in this case is 1.00, it is non-significant [84]. An OR of 1.00 would mean that individuals with a PRS in the top percentiles have the same odds for developing the disease as individuals with PRSs in the remaining part of the distribution.

Table 4.4.1 shows the ORs that an individual develops AP when their PRS is among the top 20-, 10-, 5- or 1 % highest scores in the population, compared to in the reference group. All ORs are larger than or equal to 1.00, meaning that the odds of acquiring AP increases or stays the same when an individual has a PRS within these groups. It can be observed that the ORs for PRS calculations performed with only first degree SNPs generally are lower than when second

Node degree	Percentage of distribution	Reference group	OR	95 % C.I.	P-value
First degree	Top 20 %	Remaining 80 %	1.00	0.93-1.07	$9.19 \cdot 10^{-1}$
	Top 10 %	Remaining 90 %	1.05	0.96-1.15	$2.81 \cdot 10^{-1}$
	Top 5 %	Remaining 95 %	1.06	0.94-1.20	$3.43 \cdot 10^{-1}$
	Top 1 %	Remaining 99 %	1.16	0.89-1.50	$2.54 \cdot 10^{-1}$
First- and second degree	Top 20 %	Remaining 80 %	1.10	1.03-1.18	$3.14 \cdot 10^{-3}$
	Top 10 %	Remaining 90 %	1.10	1.01-1.20	$2.91 \cdot 10^{-2}$
	Top 5 %	Remaining 95 %	1.07	0.95-1.20	$2.69 \cdot 10^{-1}$
	Top 1 %	Remaining 99 %	1.28	1.00-1.62	$4.58 \cdot 10^{-2}$

Table 4.4.1: The table shows the odds ratios (ORs) of developing angina pectoris (AP) for individuals with a polygenic risk score (PRS) among the top 20-, 10-, 5- and 1 % of the PRS distribution. The OR is given relative to a reference group with a lower PRS, which consists of the remaining part of the population. The confidence interval (CI) and p-value of each obtained OR are also displayed. The upper half of the table contains ORs for AP when only first degree single nucleotide polymorphisms (SNPs) are used in the PRS calculation, while in the lower half, both first- and second degree SNPs are included.

degree SNPs are included. All p-values are non-significant at a 5 % significance level.

For the PRSs for AP calculated using both first- and second degree SNPs, the estimated OR for the top 5 % of the PRS distribution is lower than the OR for the top 10- and 20 %, and its p-value is non-significant at a 5 % significance level. The remaining ORs have significant p-values, however, the OR for the top 1 % is close to being non-significant and its CI starts with the null value. The reliability of the estimated ORs for the top 1- and 5 % of the PRS distribution is therefore questionable.

Table 4.4.2 shows the ORs that an individual develops MI when their PRS is among the top 20-, 10-, 5- or 1 % highest scores in the population, compared to in the reference group. In this case, The ORs for the PRS calculations performed with only first degree SNPs generally are higher than when including second degree SNPs. The only exception is the OR for the top 1 % of the PRS distribution, which is 0.95. An OR lower than 1.00 implies that having a PRS this high decreases the odds for acquiring MI. However, the p-value of this OR is non-significant at a 5 % significance level, which means that a difference between the top 1 % and the remaining part of the PRS distribution cannot be established. The remaining estimations of ORs are significant.

The ORs obtained when second degree SNPs are included, are all larger than 1.00 and increases up to the top 5 % of the PRS distribution. However, the estimated OR for the top 1 % is lower than the remaining ORs. All p-values are non-significant at a 5 % significance level, which means that a difference between the four groups and the remaining part of the distribution cannot be confirmed. All estimated ORs for MI are therefore quite unreliable.

Table 4.4.3 shows the ORs that an individual develops CA when their PRS is among the top 20-, 10-, 5- or 1 % highest scores in the population, compared to in the reference group. As was the case with MI, the ORs are higher for the PRS calculations performed with first degree SNPs, and lower when second degree SNPs are included. All obtained ORs have non-significant p-values at a 5 % significance level.

Node degree	Percentage of distribution	Reference group	OR	95 % C.I.	P-value
First degree	Top 20 %	Remaining 80 %	1.11	1.04-1.19	$2.81 \cdot 10^{-3}$
	Top 10 %	Remaining 90 %	1.18	1.08-1.29	$4.05 \cdot 10^{-4}$
	Top 5 %	Remaining 95 %	1.16	1.02-1.31	$1.96 \cdot 10^{-2}$
	Top 1 %	Remaining 99 %	0.95	0.70-1.25	$7.20 \cdot 10^{-1}$
First- and second degree	Top 20 %	Remaining 80 %	1.05	0.98-1.13	$1.64 \cdot 10^{-1}$
	Top 10 %	Remaining 90 %	1.09	0.99-1.19	$8.21 \cdot 10^{-2}$
	Top 5 %	Remaining 95 %	1.11	0.98-1.26	$9.22 \cdot 10^{-2}$
	Top 1 %	Remaining 99 %	1.06	0.80-1.39	$6.75 \cdot 10^{-1}$

Table 4.4.2: The table shows the odds ratios (ORs) of developing myocardial infarction (MI) for individuals with a polygenic risk score (PRS) among the top 20-, 10-, 5- and 1 % of the PRS distribution. The OR is given relative to a reference group with a lower PRS, which consists of the remaining part of the population. The confidence interval (CI) and p-value of each obtained OR are also displayed. The upper half of the table contains ORs for MI when only first degree single nucleotide polymorphisms (SNPs) are used in the PRS calculation, while in the lower half, both first- and second degree SNPs are included.

Node degree	Percentage of distribution	Reference group	OR	95 % C.I.	P-value
First degree	Top 20 %	Remaining 80 %	1.05	0.94-1.18	$3.90 \cdot 10^{-1}$
	Top 10 %	Remaining 90 %	1.10	0.94-1.27	$2.12 \cdot 10^{-1}$
	Top 5 %	Remaining 95 %	1.11	0.90-1.35	$3.23 \cdot 10^{-1}$
	Top 1 %	Remaining 99 %	1.38	0.90-2.03	$1.17 \cdot 10^{-1}$
First- and second degree	Top 20 %	Remaining 80 %	0.97	0.86-1.09	$6.10 \cdot 10^{-1}$
	Top 10 %	Remaining 90 %	1.03	0.88-1.19	$7.44 \cdot 10^{-1}$
	Top 5 %	Remaining 95 %	1.07	0.87-1.31	$5.06 \cdot 10^{-1}$
	Top 1 %	Remaining 99 %	0.97	0.59-1.50	$9.04 \cdot 10^{-1}$

Table 4.4.3: The table shows the odds ratios (ORs) of developing coronary atherosclerosis (CA) for individuals with a polygenic risk score (PRS) among the top 20-, 10-, 5- and 1 % of the PRS distribution. The OR is given relative to a reference group with a lower PRS, which consists of the remaining part of the population. The confidence interval (CI) and p-value of each obtained OR are also displayed. The upper half of the table contains ORs for CA when only first degree single nucleotide polymorphisms (SNPs) are used in the PRS calculation, while in the lower half, both first- and second degree SNPs are included.

For the PRS calculations using first- and second degree SNPs, both the estimated ORs for the top 1- and 20 % are lower than 1.00, which implies that having a PRS in these groups lowers the odds for developing CA. However, since both these findings are non-significant, the ORs are unreliable. Therefore, a difference between the odds of acquiring CA with a PRS in the top 20-, 10-, 5- or 1 % compared to a PRS in the remaining parts of the distribution, cannot be confirmed.

Table 4.4.4 shows the ORs that an individual acquires EH when their PRS is among the top 20-, 10-, 5- or 1 % highest scores in the population, compared to in the reference group. For the PRS

Node degree	Percentage of distribution	Reference group	OR	95 % C.I.	P-value
First degree	Top 20 %	Remaining 80 %	0.92	0.85-0.99	$3.85 \cdot 10^{-2}$
	Top 10 %	Remaining 90 %	0.92	0.82-1.02	$1.16 \cdot 10^{-1}$
	Top 5 %	Remaining 95 %	0.91	0.78-1.06	$2.34 \cdot 10^{-1}$
	Top 1 %	Remaining 99 %	0.82	0.57-1.14	$2.57 \cdot 10^{-1}$
First- and second degree	Top 20 %	Remaining 80 %	1.06	0.98-1.14	$1.64 \cdot 10^{-1}$
	Top 10 %	Remaining 90 %	1.15	1.03-1.27	$8.11 \cdot 10^{-3}$
	Top 5 %	Remaining 95 %	1.14	0.93-1.31	$5.87 \cdot 10^{-2}$
	Top 1 %	Remaining 99 %	1.33	0.99-1.75	$4.66 \cdot 10^{-2}$

Table 4.4.4: The table shows the odds ratios (ORs) of developing essential hypertension (EH) for individuals with a polygenic risk score (PRS) among the top 20-, 10-, 5- and 1 % of the PRS distribution. The OR is given relative to a reference group with a lower PRS, which consists of the remaining part of the population. The confidence interval (CI) and p-value of each obtained OR are also displayed. The upper half of the table contains ORs for EH when only first degree single nucleotide polymorphisms (SNPs) are used in the PRS calculation, while in the lower half, both first- and second degree SNPs are included.

calculations performed with only first degree SNPs, all ORs are below 1.00, which implies that a PRS within any of these groups would decrease the odds of acquiring EH. The estimated OR for the top 20 % of the PRS distribution is significant at a 5 % significance level, but the p-value is quite high and the CI is quite close to being non-significant. All four estimations of ORs are therefore quite unreliable.

For the PRS calculations performed with both first- and second degree SNPs, all ORs are higher than 1.00. Thus, a PRS within these groups increases the odds of acquiring EH. The ORs are, for the most part, increasing as the top x % of the PRS distribution decreases, and the odds for acquiring EH in the top 1 % is substantially higher than for the top 20 % of the PRS distribution. Both the p-value and CI of the estimated OR for the top 10 % of the PRS distribution is significant at a 5 % significance level, but the CI contains the value 1.03, and is therefore quite close to being non-significant. The p-values of the ORs obtained for the top 5- and 20 % are non-significant, and the p-value for the estimated OR for the top 1 % is close to being non-significant. Their CIs all include the null value, and the ORs are therefore quite unreliable.

Discussion

Section 5.1 concerns the comparison of the gene-phenotype-phenotype network (GPPN), where phenotypes are connected through SNPs in common genes, and the human disease network (HDN). Section 5.2 concerns a polygenic risk score (PRS) analysis for participants of the HUNT study, for the four diseases angina pectoris (AP), myocardial infarction (MI), coronary atherosclerosis (CA) and essential hypertension (EH). The SNP-phenotype network (SPN), which was constructed for the specialization project in TBT4500, was used to determine which SNPs or LD blocks to use in the calculation of PRSs.

5.1 Comparison of Networks

In the HDN, phenotypes are connected if associated with a mutation within the same gene [18]. This is also the case for the GPPN, however, here phenotypes are connected if associated to a SNP within the same gene. The GPPN is therefore more specific. The two networks were constructed using different databases, the GPPN was constructed using the PheWeb dataset from the UK Biobank, while the HDN was constructed using the OMIM database [18]. There were therefore expected to be certain differences in the diseases included, however, since the networks were constructed using mainly the same procedure, certain similarities were presumed to be found. For the specialization project in TBT4500, the phenotype-phenotype network (PPN), where phenotypes are connected if associated with common SNPs, was compared to the HDN [19]. For this thesis, it was expected that a network linking phenotypes through common genes instead of SNPs would be more similar to the HDN than what was the case for the PPN.

There are a few common connections in the two networks, such as the link between the nodes "diabetes mellitus" and "obesity". For the most part, these seem to be thoroughly researched and established connections. It is thus reasonable that the networks have these in common. However, most connections differ, which seems to be partly due to the differences in which diseases are included. However, some of the diseases the two networks have in common are involved in different connections, such as the nodes "colon cancer" and "breast cancer". These nodes are found within the same cluster in both networks, but share none of the same connections.

In both networks, cancers tend to cluster together and are highly connected to each other. The reason for this may be that several types of cancer appear to be associated with the same tumor

suppressor genes^[85]. For example, the inactivation of the tumor suppressor gene *p53* has been found to be the cause for numerous common cancer types, such as leukemia, lymphomas, brain cancer and several types of carcinomas^[85]. The tumor suppressor genes might thus be the cause of the interconnectedness between cancers. Even though the clustering pattern of cancers is somewhat similar, the types of cancers and the links between them in the HDN and GPPN mostly differ.

A substantial part of the high degree nodes in the HDN represent diseases that are included in both networks, such as "colon cancer" and "diabetes mellitus". The large number of genes these diseases are associated with, might therefore be part of the reason why they are involved in connections in both the GPPN and HDN. Since the diseases are associated with a higher number of genes, it is more likely that some of these associations are detected in both networks.

It is apparent that the difference in which diseases are included stem from the different datasets used to construct the networks. Also, the HDN connects diseases through mutations in common genes, which means that all kinds of mutations may have been considered. This might cause different genes to be involved than what is the case when considering solely SNPs, which again would generate other connections. It is a possible explanation for why some of the diseases found in both networks are involved in different links. For the construction of the GPPN, the clustering algorithm Louvain was used to distribute diseases into clusters. It is not apparent from the human disease network article by Goh et al. whether a clustering algorithm was used in the construction of the HDN, however, differences in how this procedure was performed may have caused the observed dissimilarities in clustering patterns. The expectation was that there would be a higher degree of similarity between the GPPN and HDN, than what was the case between the PPN and HDN. However, the degree of similarity between the GPPN and HDN was lower than expected.

5.2 Polygenic Risk Score Analysis

The polygenic risk score (PRS) analysis performed for this thesis was based on the use of the SPN, where phenotypes are connected to SNPs or LD blocks if an association has been found between them. First, the PRS calculations were performed by using solely SNPs directly linked to the disease in question. Then, PRSs were calculated for the same individuals again, but this time using the SPN to determine which SNPs to include. Both SNPs directly linked to the disease and SNPs linked to its neighbouring diseases were then included in the calculations.

The main hypothesis was that to include a greater number of SNPs in the calculations, and by using the SPN to determine which SNPs to include, the ability of the PRS to predict the actual occurrence of disease would improve. Prevalence plots were thus made to test this hypothesis, where PRSs were plotted against the actual prevalence of disease in the HUNT population. In addition to these, case-control plots were made, showing the difference in PRSs between cases and controls. To observe how the risk for developing the diseases changes over a lifetime, cumulative disease risks (CDRs) were calculated for all ages. These results were visualized as plots where age was plotted against CDR for five different PRS quantiles. Lastly, a statistical analysis was performed to provide odds ratios (OR) of developing the four different diseases with a PRS in the top 20-, 10-, 5- or 1 % of the PRS distribution.

5.2.1 Prevalence Plots

Prevalence plots were made for each of the four diseases, both for the PRSs calculated with first degree SNPs, and when second second degree SNPs were included. PRSs, given as 8.3 % quantiles based on the distribution of PRSs, were plotted against the actual prevalence of the disease within that particular PRS quantile. In this way, it can be observed how significant the genetic risk, represented by the PRS, is for the the actual occurrence of disease.

As mentioned, the hypothesis was that increasing the number of SNPs included in the PRS calculations, and that choosing these additional SNPs using the SPN, the prediction accuracy of the PRS would improve. Thus, it was expected that when including both first- and second degree SNPs, a tendency of an exponential increase in disease prevalence with increasing PRSs would be observed. This expectation is based on the results presented in the article by [Aeagam and Natarajan](#) published in 2020, where a similar PRS analysis was performed for CA [\[57\]](#). In this article, there is a clear tendency of an exponential increase in disease prevalence with an increasing PRS [\[57\]](#). The PRS analysis was performed on a UK Biobank population and utilized 6.6 million genetic variants in the calculations, and therefore differs somewhat from the analysis in this thesis [\[57\]](#).

The results for AP and EH align somewhat with the initial hypothesis (Figure [4.4.1](#) and [4.4.7](#)). There is a clear improvement in the accuracy of disease prediction when second degree SNPs are included. In Figure [4.4.1a](#), when only first degree SNPs are considered, the prevalence plot for AP shows a negative trend in disease prevalence with an increasing PRS. This is also the case for EH, in Figure [4.4.7a](#). When including second degree SNPs, the disease prevalence tends to rise with an increasing PRS for both diseases. However, for both AP and EH, the prevalence varies considerably and there is no exponential increase, as was expected for these plots.

For MI, there is a tendency of an increase in disease prevalence with an increasing PRS for both prevalence plots (Figure [4.4.3a](#) and [4.4.3b](#)), although the regression line shows a somewhat steeper slope when second degree SNPs are included. The increasing tendency in Figure [4.4.3a](#), where only first degree SNPs are considered, might have occurred by chance. MI is the disease for which the number of SNPs increases the most when second degree SNPs are included. Based on the initial hypothesis for this thesis, such a substantial increase in the number of SNPs should cause a considerable improvement in disease prediction accuracy. In the prevalence plots for CA (Figure [4.4.5a](#) and [4.4.5b](#)), there is a tendency of an increase in disease prevalence when only first degree SNPs are considered. When second degree SNPs are included, there is a larger instability of disease prevalence with increasing PRS and a somewhat decreasing tendency. These results are opposite of what was expected.

The reason for the especially deviant results obtained for CA might be that the HUNT population contains only 1,901 individuals diagnosed with CA, and that the case-control ratio for CA therefore is quite low. By comparison, there were 6,469 individuals diagnosed with AP, 5,739 with MI and 4,090 with EH. A low number of cases compared to controls means that there are fewer individuals within each PRS quantile in the prevalence plots, and thus, outliers will have a higher significance. Even though the prevalence plots for AP, MI and EH all show an increasing tendency of disease prevalence, the ability of the PRS to predict disease is not as precise as expected for these diseases either. There is a high instability of disease prevalence with an increasing PRS, and no exponential increase is observed. Increasing the case-control ratio for all four diseases may therefore be a possible solution for the low prediction accuracy

of the PRSs.

5.2.2 Case-Control Plots

For the case-control plots, cases are expected to have PRSs among the upper percentiles of the distribution, and the difference between cases and controls is expected to be larger for the PRSs calculated with first- and second degree SNPs. This is because an increased genetic risk is expected to be associated with a higher occurrence of disease, and as mentioned in Section 2.6, PRSs calculated using a larger number of SNPs are anticipated to estimate the actual occurrence of disease more accurately [52]. The case-control plots for AP (Figure 4.4.2), MI (Figure 4.4.4) and EH (Figure 4.4.8) show a slightly higher PRS percentage for cases when second degree SNPs are included in the PRS calculations. However, the majority of cases would be expected to have PRSs in the upper percentiles of the PRS distribution, while the opposite is the case for controls. A larger difference between the medians of controls and cases was therefore expected.

For MI, both case-control plots (Figure 4.4.4a and 4.4.4b) show a somewhat higher PRS percentage for cases than controls. This correlates quite well with the prevalence plots for MI, where there was an increasing tendency in prevalence with increasing PRSs, both with first degree SNPs and when second degree SNPs were included. This indicates that the prediction accuracy of the PRS for MI is quite similar, independent of the number of SNPs used in the calculations. However, as was the case with AP and EH, the difference between the PRS of cases and controls was expected to be larger. The case-control plots for CA (Figure 4.4.6a and 4.4.6b) show results quite opposite of what was expected. When only first degree SNPs are used in the calculations, cases have a higher PRS median than controls, while when second degree SNPs are included, cases seem to have a lower PRS. As was the case with the prevalence plots, the deviant results for CA are most likely because of the low number of individuals with CA in the HUNT population. For all four diseases, there is a smaller difference between the PRSs of cases and controls than what was expected, which implies that the ability of the PRS to predict the occurrence of disease is questionable.

5.2.3 Cumulative Disease Risk Plots

The prevalence plots show the relation between the occurrence of disease and the genetic composition of the HUNT participants, represented by the PRSs. However, age is not taken into consideration. To analyze how the risk for developing the four diseases changes over a lifetime, cumulative disease risks (CDR) were calculated for all ages. To do this, a competing risk analysis was performed, as described in Section 3.5.2. The competing risks considered in this analysis are; (i) the patient has the disease in question, (ii) the patient is alive without the disease, (iii) the patient has died without being diagnosed with the disease. The analysis is conducted to evaluate whether there is a dependency between the time of occurrence for these events. The CDR plots presented in the results show age plotted against CDR for each 20 % PRS quantile, where the CDR is given as the probability of developing the disease.

It is expected that the CDR of the groups with highest PRSs will start to increase earlier and increase more rapidly with age, compared to the remaining groups. This is because, the increased disease risk due to an individual's genetic composition, represented by the PRS, will be added to an individual's already existing risk due to age. The probability of developing the disease is expected to increase as an individual grows older, since age is an independent risk factor

for cardiovascular diseases (CVDs)^[86]. As an individual grows older, the time over which the individual has been exposed to potential risk factors increases, such that the overall disease risk becomes higher^[87]. These expectations are based on the results obtained from a similar PRS analysis presented in the article by [Aeagam and Natarajan](#), mentioned previously^[57].

The CDR plot for AP shows that a PRS among the 80-100 % highest scores of the population gives the highest probability of developing AP at almost all ages. At the age of 102, the order of the PRS groups is mostly as expected, except for the order of group 1 and 2, which has been exchanged. The CDR plot for AP is quite similar to the plot for EH. Also here, the PRSs within group 5 provide the highest probabilities of developing EH, at least at the age of 80 and above. For both AP and EH, the age at which a 5 % CDR is obtained, occur around the same age for both the lowest and highest PRS group. However, the lowest and highest PRS groups reach a CDR of 20- and 10 %, for AP and EH, respectively, at different ages. For AP, the age difference is 5 years, while for EH, it is 6 years. These results indicate that with a higher PRS, the CDR increases more rapidly with age, and are therefore as expected. It seems that the majority of the disease risk arise from growing older, but that the additional risk due to genetic composition is more determining for the CDR at a higher age.

The CDR for EH starts to increase around the age of 40, while for AP the increase starts around the age of 50. However, from the age of 60 to 70, the slope of group 5 for AP is steeper than the slope of group 5 for EH. So, it seems that an increased disease risk due to aging starts earlier for EH than for AP. Thus, these results indicate that already at the age of 40 there is an increased risk of developing high blood pressure, which again can lead to other CVDs, such as MI^[68]. Individuals above the age of 40 are advised to check their blood pressure every year, thus, these results are reasonable^[68]. The risk of developing AP due to aging does not start to increase until the age of 50, however, the probability of developing AP increases more rapidly with age after this point, for all PRS groups. The results for both AP and EH are mostly as expected, considering the difference in CDR between the different PRS groups and that the CDR increases more rapidly for the groups with higher PRSs. This aligns well with the prevalence and case-control plots for AP and EH, which both show results that mostly align with the initial hypothesis.

In the CDR plot for MI, group 4 seems to provide higher CDRs than the other groups after the age of 80. Group 5 is relatively similar to group 3 at all ages. As was the case with AP and EH, the CDR seems to increase more rapidly with age for the groups with higher PRSs. The difference between the age at which the groups with lowest and highest PRSs reaches 20 % is 3 years, thus somewhat lower than for AP and EH. For both AP and MI, there is a 1.14-fold difference between the CDR for the lowest and highest PRS groups at the age of 103. However, the CDR seems to be more influenced by genetic risk at a higher age, considering that there is such an even spread between the different PRS groups. The CDR plot for CA shows results that deviate from what was expected. Group 2 has the highest CDRs at all ages, while the other groups are quite similar. It therefore seems that the PRS has a lower effect on the CDR for CA, and that age is the main contributor.

An article published in 2019 by [Inouye et al.](#), presented the results from a corresponding survival analysis performed for CA, where CDRs were calculated over a lifetime for a UK Biobank population^[88]. This population consisted of 22,242 cases and 460,387 controls. Individuals with a PRS among the 0-20 % lowest scores of the population showed a CDR of 10 % at the age of 75, while those with a PRS among the 80-100 % highest scores showed a disease risk

of 10 % at the age of 61 [88]. The maximum CDR was in this case around 30 %, and the risks concerned solely men. These results showed that a higher PRS increased the risk of developing CA at an earlier age, and are therefore similar to the results obtained for AP, MI and EH. The survival analysis by Inouye et al. has been performed on a population with a somewhat higher case-control ratio for CA than in this thesis, which may indicate that the low case-control ratio is part of the reason for the deviant results obtained for CA [88].

The probabilities of developing the four diseases are considerably different. At the age of 102, the probability of acquiring AP with a PRS in group 5 is twice as high as for EH. However, comparing the probabilities between plots may be problematic, since the probabilities are not absolute, but relative to a control group. The CDR for individuals with a PRS in group 5 for AP is 20 % at the age of 83. This does not mean that the absolute risk for developing AP at the age of 83 is 20 %, but rather that the probability is twice as high as at the age of 69, where it is 10 %. The probabilities for CA are considerably lower than what was the case for AP, MI and EH. While MI had a maximum CDR of 25 %, the highest CDR for CA is 7 %. This may indicate that the risk for acquiring CA is less affected by age, but may also be partly because of the low case-control ratio for CA.

There are limitations to the procedure used to generate the CDRs. The CDRs are, as mentioned, calculated using a competing risk analysis where the three competing risks are (i) the patient has the disease in question, (ii) the patient is alive without the disease, (iii) the patient has died without being diagnosed with the disease. The CDR is based on the relation between these three events and their time of occurrence. However, calculating the disease risk with additional risk factors would probably provide a more accurate prediction. Potential risk factors that are common for all CVDs considered in this thesis are diabetes, smoking, alcohol consumption, stress- and activity levels, among others [65][66][67][68]. The HUNT data contain information regarding all these risk factors, through the questionnaires filled out by participants of the study [13].

5.2.4 Statistical Analysis

The odds ratio (OR) tables present the ORs that an individual develops the diseases when their PRS is among the top 20-, 10-, 5- or 1 % highest scores in the population, compared to in the reference group. The null hypothesis is that there is no difference in the number of people that acquire the disease between these distributions and the reference groups. For the PRS calculations performed with both first- and second degree SNPs, the ORs are expected to be higher than for those where only first degree SNPs are used. The reason for this is that, as mentioned previously, including a larger number of SNPs in the calculation of PRSs increases the predictability of the score [52]. In addition to this, the additional SNPs included are chosen based on the use of the SPN, and are therefore known to have some level of association with the disease. This is also believed to strengthen the ability of the PRS to estimate the actual occurrence of disease, which would cause a larger difference between the top 20-, 10-, 5- and 1 % and the remaining part of the PRS distribution.

The 95 % confidence interval (CI) and p-value of the obtained OR are provided in the tables. It is expected that the ORs obtained for the PRSs calculated with first- and second degree SNPs will be more accurate. Thus, higher ORs and lower p-values are expected. This is because, as the prediction accuracy of the PRS improves, the difference between the top x % of the PRS distribution and the reference group will be more easily detected, if there is one [84]. The

larger this difference is, the lower is the p-value [84]. The p-values are expected to decrease as x decreases. This is because the difference between the top x % of the distribution and the reference group should increase as the PRSs in the top x % becomes higher. If a CI contains the null value, which in the case of ORs is 1.00, the interval is non-significant [84]. An OR of 1.00 would mean that there is no difference between the top x % and the remaining part of the PRS distribution. The CIs are expected to increase in width as the top x % of the PRS distribution decreases. This is because a smaller sample size causes a wider CI, and thereby a less precise estimation of the OR [84].

The results for AP in Table 4.4.1 partially support the initial hypothesis that the difference between the top 20-, 10-, 5- and 1 % of the distribution and the reference group becomes more apparent when including a larger number of SNPs. For the PRSs calculated using first degree SNPs, all ORs are non-significant, while when second degree SNPs are included, both the ORs for the top 10- and 20 % of the PRS distribution are significant. The largest OR for developing AP is for the top 1 % of the distribution. This is logical, since the top 1 % of the PRS distribution only contains the highest scores of the population. However, the OR for the top 1 % is close to being non-significant and its CI contains the null value, thus, a difference between the top 1 % and the remaining part of the distribution cannot be established. This does not necessarily mean that there is no difference between these groups, but might be because of the low number of cases, which causes unreliable results when smaller groups are considered.

The ORs for MI displayed in Table 4.4.2 are not quite as expected. Unlike the results for AP, the obtained ORs are generally higher when only first degree SNPs are used in the PRS calculations. For these PRSs, all ORs are significant at a 5 % significance level, except for the OR for the top 1 %. The CIs generally increase in width as the top x % decreases, due to the decrease in sample size. This means that the provided ORs become less precise estimates. The results for MI are not as expected, considering that when including second degree SNPs, the estimated ORs are generally lower and have higher p-values than what is the case with first degree SNPs. This aligns quite well with the obtained results in the prevalence- and case-control plots for MI (Figure 4.4.3 and 4.4.4), which showed a similar PRS prediction accuracy both when first degree SNPs were used, and when second degree SNPs were included.

The ORs for CA in Table 4.4.3 are, as was the case with MI, higher when only first degree SNPs are used in the calculations. However, neither the p-values nor CIs are significant for any ORs, and the estimates are therefore quite unreliable. The results in the OR table for CA deviate from the initial hypothesis, which is as expected, considering that the results for CA deviated from the initial hypothesis in all previously reported results. For both MI and CA, including a larger number of cases compared to controls would most likely improve the PRS prediction accuracy. This would probably increase the number of individuals with PRSs in the top percentiles of the distribution, and thus make the differences between these groups and the remaining fractions of the population, easier to detect.

The ORs for EH in Table 4.4.4 are higher when second degree SNPs are included, as was the case for AP. These ORs also have lower p-values, although, only one of them is significant at a 5 % significance level. The p-value of the estimated OR for the top 1 % of the PRS distribution is, as mentioned, close to being non-significant at a 5 % significance level. When taking into consideration the corresponding CI, which contains the null value, this OR can be assumed to be unreliable. The estimated ORs for EH thus align better with the initial hypothesis than what was the case for MI and CA, but the reliability of the results is lower than what was

expected. Although the p-values are somewhat lower when including second degree SNPs in the calculations, they are still, for the most part, non-significant. Thus, as with the other diseases considered in this thesis, including a larger number of cases compared to controls would most likely increase the predictability of the PRSs and make it easier to distinguish between the top percentiles and the remaining parts of the PRS distribution.

5.2.5 Comparison of Results

The obtained ORs for the four diseases correlate with the prevalence-, case-control- and CDR plots. The obtained results for AP and EH align quite well with the initial hypothesis, while the results for MI and CA are more deviant. However, the PRS prediction accuracy was expected to be higher for all four diseases, and a larger difference was expected to be observed when including second degree SNPs in the calculations. The difference is most apparent for AP and EH, where both the prevalence- and case-control plots for AP and EH show a slight improvement in PRS prediction accuracy when including second degree SNPs. The CDR plots for these diseases present results that align with the initial hypothesis, where the CDR of the groups with highest PRSs increases more rapidly with a higher age. The OR tables for AP and EH do, however, show poorer results than what was expected. Even though there is a slight increase in ORs when including second degree SNPs, a large part of these estimates are non-significant at a 5 % significance level. AP and EH were the diseases with the lowest increase in the number of SNPs used when second degree SNPs were included in the PRS calculations. Despite of this, the results for these diseases were most highly correlated with the initial hypothesis.

The prevalence- and case-control plots, CDRs and ORs for MI and CA also correlate quite well with each other. For MI, the difference in PRS prediction accuracy when second degree SNPs are included in the prevalence- and case-control plots, is smaller than expected. The CDR plot shows somewhat deviating results, and the estimated ORs are both lower and have higher p-values when second degree SNPs are included. For CA, none of the presented results support the initial hypothesis. There is no improvement in PRS prediction accuracy in neither the prevalence- nor case-control plot, and the CDR plot shows a different order of CDRs for the PRS groups than what was expected. Also, none of the estimated ORs for the top 20-, 10-, 5- or 1 % of the distribution were found to be significant at a 5 % significance level.

The results from the PRS analysis performed for this thesis were not entirely as expected. Four quite common CVDs were chosen for the PRS calculations, which were performed on 69,423 participants of the HUNT study. Even though these diseases are common in the general population, quite few individuals in the HUNT population had the diagnoses. As has been discussed, this is a possible explanation for how much some of the results deviated from the initial hypothesis. Especially CA, which was the disease with the fewest cases in the HUNT population, showed curious results. Care should therefore be taken before making use of PRSs for clinical purposes. The methods used for calculating PRSs are still in development, and the values of PRSs differ considerably depending on which method is used, how many SNPs are included, their estimated effect sizes, and which populations the PRS calculations are performed for [87]. The covariates considered, such as age and sex, also vary considerably between different analyses [87]. A more standardized procedure for calculating the PRSs would probably make it easier to compare PRSs between different studies, and could make their utilization for clinical purposes more achievable [87].

Even though there are uncertainties linked to the usage of PRSs, they could be useful in early disease prevention^[87]. Especially for CVDs, common risk factors are not measured early in life. However, by genotyping individuals early and calculating PRSs for a series of diseases, a potentially increased risk for these diseases can be detected. Preventative measures, such as lifestyle choices, or potential therapies and screening procedures might then be made use of^[87]. An example is the use of statins for the treatment of CVDs. Statins are used to lower the levels of low density lipoprotein (LDL) cholesterol, to decrease the risk for disease^[87]. An article published by [Natarajan et al.](#), considers the use of statins for the prevention of CA and whether the PRS has an impact on the effect of statin treatment^[89]. It was found that among individuals with the highest PRSs of the population, statins decreased the risk for suffering a first case of CA with 44 %. For individuals with a low PRS, the risk decreased with 24 %^[89].

5.2.6 Sources of Error

When including a larger number of SNPs in the PRS calculations, the SNP-phenotype network (SPN) is used to choose these SNPs. The hypothesis is that when including a larger number of SNPs, and these SNPs are known to be closely associated with the disease through the network, the disease prediction of the PRS will improve. However, CA had a 27-fold increase in the number of SNPs, which is higher than the remaining diseases. Despite of this, the results for CA correlated the least with the initial hypothesis. MI had the second highest increase, and also showed somewhat deviating results. First degree SNPs are directly associated with the four diseases, while this is not necessarily the case for second degree SNPs. Including a larger number of SNPs may increase the predictability of the PRS, but too many SNPs with a low or erroneous effect size might have a negative effect on this predictability^[87]. There is therefore a trade-off between including a low number of SNPs with precise estimates of effect sizes and a large number of SNPs with less precise estimates^[87]. The PRSs for CA and MI might have been calculated using a greater number of SNPs with low and possibly erroneous effect sizes, which could be part of the reason for the deviating results.

For all four diseases, few of the estimated ORs are significant at a 5 % level. As mentioned previously in the discussion, the low case-control ratio most likely had an effect on the deviating results. According to an article by [Hodge et al.](#) published in 2014, increasing the number of controls compared to cases will only increase the statistical power up to a certain point^[90]. An excessively large number of controls compared to cases is therefore not necessarily advantageous. CA is the disease with the lowest case-control ratio, and has 36 times as many controls as cases. By comparison, AP has the highest case-control ratio, with 10 times as many controls as cases. Thus, the low case-control ratio might have negatively affected the statistical power of the analysis, and caused the somewhat deviating results^[90]. If the PRSs actually can predict an increased risk for developing a disease, a higher case-control ratio would probably make the differences between the top 20-, 10-, 5 and 1 % and the remaining part of the PRS distribution more easily detectable. For the CDR plots, increasing the case-control ratio would most likely improve the prediction accuracy of the cumulative disease risks, and make it easier to detect how the CDR actually changes with age and how much it is affected by genetic risk.

As mentioned in Section [3.5.1](#), when including second degree SNPs, the effect sizes used in the PRS calculations might apply to other phenotypes than the disease the PRSs are calculated for. Since the four CVDs considered in this thesis are highly connected in the SNP-phenotype network (SPN), it means that when including second degree SNPs, the calculations of the PRSs

for the different diseases will have a certain number of SNPs in common. Also, the SNPs used in the calculation of PRSs for a particular disease may not be directly causal to that disease. This might affect the generalizability of the PRSs, which would limit their transferability to other populations^[91]. The reason for this is probably that even though erroneous effect sizes are used in the calculations, a PRS can still be compared to the remaining scores of the distribution. In this way, the PRS can describe the individual's relative disease risk in that population, however, these PRSs cannot be compared to other populations.

A last consideration to be made, is that the GWAS summary statistics used to obtain the effect sizes of SNPs, apply to the UK Biobank population, while the PRS calculations are performed for the HUNT population. There may be genetic differences between the populations, and also, individuals of the UK Biobank population have a lower level of obesity, alcohol consumption and self-reported health conditions than the average population^[92]. In addition to this, they have lower rates of cancer, both in women and men^[92]. Thus, the UK biobank population does not necessarily represent the general population of the United Kingdom, but seems to be somewhat healthier. The usage of effect sizes that have been transferred from the UK Biobank to the HUNT population may therefore be problematic, and is possibly part of the reason for the poor prediction accuracy of the PRSs observed in this thesis.

Conclusion and Outlook

This chapter contains a conclusion of the results obtained during this thesis, followed by the outlook section. This is an evaluation of what could be done further in this thesis, which includes potential developments from the current results that were not done either due to a lack of time or lack of resources.

6.1 Conclusion

For the comparison between the gene-phenotype-phenotype network (GPPN) and the human disease network (HDN), the initial hypothesis was that the GPPN, where phenotypes are linked through SNPs in common genes, would show a high degree of similarity with the HDN, where diseases are linked through mutations in common genes. For the specialization project in TBT4500, the phenotype-phenotype network (PPN), a network where phenotypes are connected when associated with common SNPs, was compared to the HDN. The networks had quite few similarities, and it was therefore presumed that a network connecting phenotypes through genes would show a higher degree of similarity. The GPPN and HDN had certain similarities in the clustering pattern of cancers, which tend to be highly connected to each other and form cancer clusters in both networks. However, the diseases included in the two networks mostly differed, and the few diseases the networks had in common were usually involved in different connections. In addition to this, diseases were classified differently, which made a comparison more difficult. Overall, the two networks had fewer similarities than expected.

For the polygenic risk score (PRS) analysis, the main hypothesis was that using a larger number of SNPs in the PRS calculations would improve the disease prediction accuracy of the PRSs in the HUNT population. In addition to this, it was hypothesized that using the SPN to determine which SNPs to include when increasing the number of SNPs, would further improve this prediction accuracy, compared to if the SNPs were chosen randomly. Out of the four diseases, the results for AP and EH were most highly correlated with the initial hypothesis, even though these diseases had the lowest increase in the number of SNPs. The prevalence plots for AP and EH showed a tendency of increasing disease prevalence with an increasing PRS when second degree SNPs were included, as was expected. However, there was a high instability of prevalence, and the exponential increase that was expected could not be observed. For the case-control plots, the majority of cases was expected to have PRSs in the upper percentiles

of the PRS distribution, while the opposite was expected for controls. The case-control plots showed a slightly higher PRS percentage for cases than controls when second degree SNPs were included, while the opposite was the case when only first degree SNPs were used in the calculations. This aligns with the initial hypothesis, however, a larger difference between cases and controls was expected.

The prevalence plots for MI showed a tendency of increasing disease prevalence with an increasing PRS, both when first degree SNPs were used and when second degree SNPs were included. Both case-control plots showed a slightly higher PRS percentage for cases than controls, however, also here the difference between cases and controls was smaller than expected. The PRSs for MI seemed to show a similar prediction accuracy both when only first degree SNPs were used in the calculations, and when second degree SNPs were included. However, there was a high instability of prevalence, and the prediction accuracy for MI was therefore questionable. The prevalence and case-control plots for CA deviated the most from the initial hypothesis. The disease prevalence had an increasing tendency with increasing PRSs when only first degree SNPs were used in the calculations. When second degree SNPs were included, the disease prediction accuracy seemed to deteriorate, and there was a decreasing tendency. In the case-control plots, cases had a slightly higher PRS percentage than controls when only first degree SNPs were used in the calculations, while the opposite was the case when second degree SNPs were included. These results were therefore not as expected.

The cumulative disease risk (CDR) plots show how the probability of developing the disease changes over a lifetime with different levels of PRSs. The CDR plots for the four diseases correlated quite well with the results obtained in the prevalence- and case-control plots. For AP and EH, the CDR increased more rapidly with age for groups with higher PRS, and there was a larger difference between groups at a higher age, as expected. Thus, it seems that the majority of the disease risk originates from growing older, but that the additional risk due to genetic composition, represented by the PRS, is more influential at an older age. In the CDR plot for MI, there is also a larger difference between groups at a higher age, however, the group with the 60-80 % highest PRSs showed a higher CDR than the group with the 80-100 % highest PRSs, from the age of 80 and onwards. The CDR plot for CA deviated the most from the initial hypothesis. The group with the 20-40 % lowest PRSs had higher CDRs than the remaining groups from the age of 40 and onwards. The other groups showed quite similar CDRs at all ages. These results indicate that the CDR of developing CA for individuals with a PRS in these groups, was not considerably affected by the PRS.

A statistical analysis was performed to calculate the ORs for developing the diseases with a PRS in the top 20-, 10-, 5- or 1 % of the PRS distribution, compared to a PRS in the reference groups. The obtained ORs were quite deviant from what was expected. The results for AP and EH were most highly correlated with the initial hypothesis, with higher estimations of ORs and lower p-values when second degree SNPs were included in the calculations. However, only a few estimations were statistically significant at a 5 % level. For MI and CA, the estimated ORs were lower and had higher p-values when second degree SNPs were included, and these results were therefore opposite of what was expected. The poor results obtained in the statistical analysis were most likely due to the low case-control ratio of the four diseases in the HUNT population, which made it difficult to compare small groups of the population.

Using a network approach to perform a PRS analysis is not known to have been done previously. By using the SNP-phenotype network to choose which SNPs should be added to the PRS

calculations, it was ensured that these SNPs had some type of connection to the disease. The hypothesis was that using this procedure would lead to a greater disease prediction ability of the PRS, compared to if the SNPs had been chosen randomly. However, the prediction ability of the PRS for the four cardiovascular diseases in the HUNT population was less accurate than expected. Even though an apparent improvement was observed for AP and EH when second degree SNPs were included, the results were not as predictive of the actual occurrence of disease, or as reliable, as hypothesized.

6.2 Outlook

For the PRS analysis, the effect of the PRS in combination with age was evaluated by analyzing CDRs. However, another possibility would have been to analyze the differences in disease risk between males and females. It has for a long time been the impression that males are at higher risk for developing cardiovascular diseases (CVDs)^[93]. However, the appearance of a low occurrence rate of CVDs in women has probably been due to the low prevalence in younger women, however, the risk of CVDs increases more rapidly with age for women over the age of 50^[93]. The majority of research on CVDs have been performed on men, and it has been assumed that the results from these studies apply to women as well. This is, however, not the case, and the differences in risk depends on genetic and other biological factors, as well as social factors^[93]. By evaluating PRSs and CDRs separately for women and men in the HUNT population, observations could be made regarding differences in both genetic risk and age of onset for CVDs. As mentioned in the discussion, common risk factors for the four disease could be taken into consideration as well, such as diabetes, smoking, alcohol consumption, stress- and activity levels^{[65][66][67][68]}. An evaluation of environmental risk factors in addition to genetic risk would provide a more precise estimate for the overall disease risk.

Further, the PRS analysis could be performed again for the same four diseases, but when including second degree SNPs in the calculations, these SNPs could be chosen randomly from the SPN. The PRSs calculated in this thesis showed a lower ability to predict disease than expected, but a comparison with a PRS analysis performed using randomly chosen SNPs would have allowed an observation of how large the effect of using a network approach is, and how much the choice of SNPs matters for the disease prediction of the PRS.

The PRS analysis could also have been performed for other diseases. For this thesis, the focus was cardiovascular diseases. Angina pectoris, myocardial infarction, coronary atherosclerosis and essential hypertension were all connected to each other in the SPN, and were therefore associated with a high number of common SNPs. The SNPs included in the calculations might have had especially low or erroneous effect sizes, and would as a consequence generate imprecise PRSs. Performing the PRS analysis for another set of diseases, which have a lower number of SNPs in common, might have caused a different outcome. When the cardiovascular diseases for the PRS analysis were chosen, this was done partly based on their high prevalence in the HUNT population. However, if other disease categories had been considered as well, diseases with a higher prevalence might have been found, which would have generated a higher case-control ratio when performing the PRS analysis. This would be advantageous, considering that the low case-control ratio for the diseases considered in this thesis is believed to be part of the reason for the low prediction ability of the PRSs.

Bibliography

- [1] EMBL-EBI. *What are genome wide association studies (GWAS)?* <https://www.ebi.ac.uk/training-beta/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/> [downloaded december 2020], 2020.
- [2] Michael D. Gallagher and Alice S. Chen-Plotkin. The post-gwas era: From association to function. *American Journal of Human Genetics*, **102**:717–730, 2018.
- [3] Leslie A. Pray. Discovery of dna structure and function: Watson and crick. *Nature Education*, **1**, 2008.
- [4] D. Peter Snustad and Michael J. Simmons. *Principles of Genetics*. John Wiley and Sons, pp. 786, 6th edition, 2012.
- [5] Heidi Chial. Huntington’s disease: The discovery of the huntingtin gene. *Nature Education*, **1**, 2008.
- [6] Heidi Chial. Mendelian genetics: Patterns of inheritance and single-gene disorders. *Nature Education*, **1**, 2008.
- [7] Shiro Ikegawa. A short history of the genome-wide association study: Where we were and where we are going. *Genomics and Informatics*, **10**:220–225, 2012.
- [8] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, and Toshihiro Tanaka. Functional snps in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, **32**:650–654, November 2002.
- [9] Pim van der Harstcorresponding and Niek Verweij. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation Research*, **122**:433–443, 2018.
- [10] Zhiqiang Li, Jianhua Chen, Hao Yu, Lin He, Yifeng Xu, Dai Zhang, Qizhong Yi, Changgui Li, Xingwang Li, Jiawei Shen, Zhijian Song, Weidong Ji, Meng Wang, Juan Zhou, Boyu Chen, Yahui Liu, Jiqiang Wang, Peng Wang, Ping Yang, Qingzhong Wang, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics*, **49**:1576–1583, 2017.

- [11] Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E. Kemper, Zhili Zheng, Loic Yengo, Luke R. Lloyd-Jones, Julia Sidorenko, Yeda Wu, eQTLGen Consortium, Allan F. McRae, Peter M. Visscher, Jian Zeng, and Jian Yang. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, **9**:1–14, 2018.
- [12] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, **20**:467–484, 2019.
- [13] HUNT Research Center. *The HUNT Study - a longitudinal population health study in Norway*. <https://www.ntnu.edu/hunt> [downloaded 27 November 2020], 2020.
- [14] National Human Genome Research Institute. *Genome-Wide Association Studies Fact Sheet*. <https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet>, [downloaded May 2021], 2020.
- [15] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature; Author Manuscript*, **461**:747–753, 2009.
- [16] Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, **12**:1–11, 2020.
- [17] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.
- [18] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, **104**:8685–8690, 2007.
- [19] Marit K. Skinderhaug. Phewas networks. pages 1–56, 2020.
- [20] Andreas Ziegler and Inke R. König. *A Statistical Approach to Genetic Epidemiology*. WILEY-VCH Verlag GmbH & Co. KGaA, Germany, pp. 497, 2nd edition, 2010.
- [21] MedlinePlus. *What are genes?* <https://medlineplus.gov/genetics/understanding/basics/gene/> [downloaded August 2020], 2020.
- [22] MedlinePlus. *What is DNA?* <https://ghr.nlm.nih.gov/primer/basics/dna>, [downloaded August 17, 2020], 2020.
- [23] MedlinePlus. *What are single nucleotide polymorphisms (SNPs)?* <https://ghr.nlm.nih.gov/primer/genomicresearch/snp> [downloaded August 2020], 2020.
- [24] Na Deng, Heng Zhou, Hua Fan, and Yuan Yuan. Single nucleotide polymorphisms and cancer susceptibility. *Impact journals*, **8**:110635–110649, 2017.

- [25] Barkur S. Shastry. Snps in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics*, **52**:871–880, 2017.
- [26] Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews*, **9**:356–369, 2008.
- [27] Peter M. Visscher, Naomi R. Wray, Pamela Sklar Qian Zhang, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, **101**:5–22, 2017.
- [28] Yohan Bossé and Christopher I. Amos. A decade of gwas results in lung cancer. *Cancer Epidemiology, Biomarkers & Prevention*, **27**:363–379, 2019.
- [29] Haoyu Zhang, Thomas U. Ahearn, Julie Lecarpentier, Daniel Barnes, Jonathan Beesley, Guanghao Qi, Xia Jiang, Tracy A. O’Mara, Ni Zhao, Manjeet K. Bolla, Alison M. Dunning, Joe Dennis, Qin Wang, Zumuruda Abu Ful, Kristiina Aittomäki, Irene L. Andrulis, Hoda Anton-Culver, Volker Arndt, Kristan J. Aronson, Banu K. Arun, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics*, **52**:572–581, 2020.
- [30] GWAS Catalog. *About the GWAS Catalog*. <https://www.ebi.ac.uk/gwas/docs/about> [downloaded October 2020], 2020.
- [31] Gareth James, Daniela Witten, and Trevor Hastie Robert Tibshirani. *An Introduction to Statistical Learning*. Springer-Verlag New York Inc, pp. 426, 2017.
- [32] Germán Rodríguez. *Lecture notes in Generalized Linear Models*. <https://www.math.ntnu.no/emner/TMA4315/2013h/lecture-notes.pdf> [downloaded december 2020], 2013.
- [33] Damir Kalpić, Nikica Hlupić, and Miodrag Lovrić. *Lovric M. (eds) International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg, 2011. Student’s t-Tests.
- [34] Pak C. Sham and Shaun M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews*, **15**:335–346, 2014.
- [35] Jae W. Song, , and Kevin C. Chung. Observational studies: Cohort and case-control studies. *National Institute of Health*, **126**:2234–2242, 2010.
- [36] Andrea R. Waksmunski, Leighanne R. Main, and Jonathan L. Haines. *Genetics and Genomics of Eye Disease*. Elsevier inc., pp. 383, 2019. Chapter 2.
- [37] Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Wiley Online Library*, **27**:1–10, 2018.
- [38] Carl A. Anderson, Fredrik H Pettersson, Geraldine M. Clarke, Lon R Cardon, Andrew P. Morris, and Krina T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, **5**:1564–1573, 2010.
- [39] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, **10**:387–406, 2009.

- [40] The International HapMap Consortium. The international hapmap project. *Nature*, **426**: 789–796, 2003.
- [41] Ho-Youl Jung, Yun-Ju Park, Young-Jin Kim, Jung-Sun Park, Kuchan Kimm, and InSong Koh. New methods for imputation of missing genotype using linkage disequilibrium and haplotype information. *Information Sciences*, **177**:804–814, 2006.
- [42] Huiying Zhao, Dale R. Nyholt, Yuanhao Yang, Jihua Wang, and Yuedong Yang. Improving the detection of pathways in genome-wide association studies by combined effects of snps from linkage disequilibrium blocks. *Scientific Reports*, **7**:1–8, 2017.
- [43] Mun-Gwan Hong, Yudi Pawitan, Patrik K. E. Magnusson, and Jonathan A. Prince. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Human Genetics*, **126**:289–301, 2009.
- [44] Antonio F. Pardiñas, Peter Holmans, Andrew J. Pocklington, Valentina Escott-Price, Stephan Ripke, Noa Carrera, Sophie E. Legge, Sophie Bishop, Darren Cameron, Marian L. Hamshere, Jun Han, Leon Hubbard, Amy Lynham, Kiran Mantripragada, Elliott Rees, and James H. MacCabe. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, **50**: 381–389, 2018.
- [45] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**:421–427, 2014.
- [46] Google. *How Google Search Works*. https://www.google.com/intl/en_uk/search/howsearchworks/, 2020.
- [47] Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, **78**: 1360–1381, 1973.
- [48] Mark Needham and Amy E. Hodler. *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O’Reilly Media, 2019. Chapter 6.
- [49] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**:1–12, 2019.
- [50] Victor A. McKusick. Mendelian inheritance in man and its online version, omim. *American Journal of Human Genetics*, **80**:588–604, 2007.
- [51] GO Consortium. *About the GO*. <http://geneontology.org/docs/introduction-to-go-resource/> [downloaded October 2020], 2020.
- [52] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F. O’Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, **15**:2759–2772, 2020.
- [53] Morgan E. Levine, Peter Langfelder, and Steve Horvath. *A Weighted SNP Correlation Network Method for Estimating Polygenic Risk Scores*. Humana Press, 2017. Volume 1613, pp. 277-290.
- [54] Amit V. Khera, Connor A. Emdin, D. Phil, Isabel Drake, Pradeep Natarajan, Alexander G. Bick, Nancy R. Cook, Daniel I. Chasman, Usman Baber, Roxana Mehran, Daniel J. Rader,

- Valentin Fuster, Eric Boerwinkle, Olle Melander, Marju Orho-Melander, Paul M Ridker, and Sekar Kathiresan. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *The New England Journal of Medicine*, **375**:2349–2358, 2016.
- [55] Kangcheng Hou, Kathryn S. Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics*, **51**:1244–1251, 2019.
- [56] NCBI. *SNP FAQ Archive*. <https://www.ncbi.nlm.nih.gov/books/NBK44377/> [downloaded March 2021], 2006.
- [57] Krishna G. Aeagam and Pradeep Natarajan. Polygenic scores to assess atherosclerotic cardiovascular disease risk. *Circulation Research*, **126**:1159–1177, 2020.
- [58] L. Duncan, H. Shen, B. Gelaye, J. Meijssen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, **10**:1–9, 2019.
- [59] S. Krokstad, A. Langhammer, K. Hveem, T.L. Holmen, K. Midthjell, T.R. Stene, G. Bratberg, J. Heggland, and J. Holmen. Cohort profile: The hunt study, norway. *International Journal of Epidemiology*, **42**:968–977, 2012.
- [60] HUNT Research Centre. *The Young-HUNT Study*. <https://www.ntnu.edu/hunt/young-hunt>, [downloaded March 2021], 2021.
- [61] Vegard Knutsen. *Hva er nytt i HUNT4 – den fjerde store Helseundersøkelsen i Nord-Trøndelag?* <https://www.ntnu.no/blogger/hunt4/2017/09/08/hva-er-nytt-i-hunt4-den-fjerde-store-helseundersokelsen-i-nord-tronde> [downloaded March 2021], 2017.
- [62] HUNT Forskningscenter. *Helseundersøkelsen i Trøndelag (HUNT), HUNT forskningscenter*. <https://forskningsprosjekter.ihelse.net/prosjekt/46053000>, [downloaded March 2021], 2020.
- [63] NTNU HUNT Forskningscenter. *HUNT Biobank*, 2021.
- [64] HUNT-MI. *HUNT-MI: Studiedel på kardiovaskulær farmakogenetikk*, 2021.
- [65] John Hopkins Medicine. *Angina Pectoris*. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/angina-pectoris> [downloaded May 2021], 2021.
- [66] John Hopkins Medicine. *Heart Attack*. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/heart-attack> [downloaded May 2021], 2021.
- [67] John Hopkins Medicine. *Coronary Heart Disease*. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronary-heart-disease> [downloaded May 2021], 2021.
- [68] Mayo Clinic. *High blood pressure (hypertension)*. <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410> [downloaded May 2021], 2021.

- [69] *About PheWeb*. <http://pheweb.sph.umich.edu/SAIGE-UKB/about> [downloaded September 18, 2020], 2020.
- [70] World Health Organization. Icd purpose and uses. <https://www.who.int/classifications/icd/en/> [downloaded October 2020], 2020.
- [71] Cynthia J. Coffman, R. W. Doerge, Katy L. Simonsen, Krista M. Nichols, Christine K. Duarte, Russell D. Wolfinger, and Lauren M. McIntyre. Model selection in binary trait locus mapping. *Genetics*, **170**:1281–1297, 2005.
- [72] Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, **50**:1335–1341, 2018.
- [73] Lee Lab. Lee lab, resources. <https://www.leelabsg.org/resources> [downloaded October 2020], 2020.
- [74] Timothy A. Myers, Stephen J. Chanock, and Mitchell J. Machiela. Ldlinkr: An r package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Frontiers in Genetics*, **11**, 2020.
- [75] Jenna M. VanLiere and Noah A. Rosenberg. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology*, **74**:130–137, 2008.
- [76] Emily C. Zabor. Competing risks. https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html#Part_3:_Competing_Risks [downloaded May 2021], 2019.
- [77] Bob Gray. cmprsk: Subdistribution analysis of competing risks. <https://CRAN.R-project.org/package=cmprsk> [downloaded May 2021], 2020. R package version 2.2-10.
- [78] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.
- [79] Lymphoma Action. *Nodular lymphocyte-predominant Hodgkin lymphoma*. <https://lymphoma-action.org.uk/types-lymphoma-hodgkin-lymphoma/nodular-lymphocyte-predominant-hodgkin-lymphoma> [downloaded March 2021], 2021.
- [80] MedlinePlus. *Graves Disease*. <https://medlineplus.gov/ency/article/000358.htm> [downloaded March 2021], 2021.
- [81] Abdullah S. Al-Goblan, Mohammed A. Al-Alfi, and Muhammad Z. Khan. Mechanism linking diabetes mellitus and obesity. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, **7**:587–591, 2014.
- [82] Robert H. Eckel, Steven E. Kahn, Ele Ferrannini, Allison B. Goldfine, David M. Nathan, Michael W. Schwartz, Robert J. Smith, and Steven R. Smith. Obesity and type 2 dia-

- betes: What can be unified and what needs to be individualized? *The Journal of Clinical Endocrinology and Metabolism*, **96**:1654–1663, 2011.
- [83] Benjamin M Leon and Thomas M Maddox. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. *World Journal of Diabetes*, **6**:1246–1258, 2015.
- [84] Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Confidence interval or p-value? *Deutsches Ärzteblatt International*, **106**:335–339, 2009.
- [85] Cooper G. M. *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates, 2nd edition, 2000. Tumor Suppressor Genes.
- [86] Jennifer L. Rodgers, Jarrod Jones, Samuel I. Bolleddu, Sahit Vanthenapalli, Lydia E. Rodgers, Kinjal Shah, Krishna Karia, and Siva K. Panguluri. Cardiovascular risks associated with gender and aging. *Journal of Cardiovascular Development and Disease*, **6**: 1–19, 2019.
- [87] Samuel A. Lambert, Gad Abraham, and Michael Inouye. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, **28**:133–142, 2019.
- [88] Michael Inouye, Gad Abraham, Christopher P. Nelson, Angela M. Wood, Michael J. Sweeting, Frank Dudbridge, Florence Y. Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, Shu Ye, Thomas R. Webb, Martin K. Rutter, Ioanna Tzoulaki, Riyaz S. Patel, Ruth J.F. Loos, Bernard Keavney, Harry Hemingway, John Thompson, Hugh Watkins, Panos Deloukas, Emanuele Di Angelantonio, Adam S. Butterworth, John Danesh, and Nilesh J. Samani. Genomic risk prediction of coronary artery disease in 480,000 adults. *Journal of the American College of Cardiology*, **72**:1883–1893, 2018.
- [89] Pradeep Natarajan, Robin Young, Nathan O. Stitzel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F. Reilly, Adam Butterworth, Daniel J. Rader, Ian Ford, Naveed Sattar, and Sekar Kathiresan. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation, AHA Journals*, **135**: 2091–2101, 2017.
- [90] Susan E. Hodge, Ryan L. Subaran, Myrna M. Weissman, and Abby J. Fyer. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *American Journal of Psychiatry*, **169**:785–789, 2012.
- [91] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, **19**:581–590, 2018.
- [92] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American Journal of Epidemiology*, **186**:1026–1034, 2017.
- [93] Anne Maria Möller-Leimkühler. Gender differences in cardiovascular disease and comorbid depression. *American Journal of Epidemiology*, **9**:71–83, 2007.

Appendix

A.1 SNP-Phenotype Network

The SNP-phenotype network (SPN) was constructed based on the PheWeb dataset from the UK Biobank [\[69\]](#). In this network, phenotypes are connected to SNPs if an association has been found between them. Linkage disequilibrium (LD) between SNPs is taken into consideration, such that SNPs in LD are found within the same LD block. Figure [A.1.1](#) displays the entire SPN, where solely associations with p-values below 10^{-6} are included.

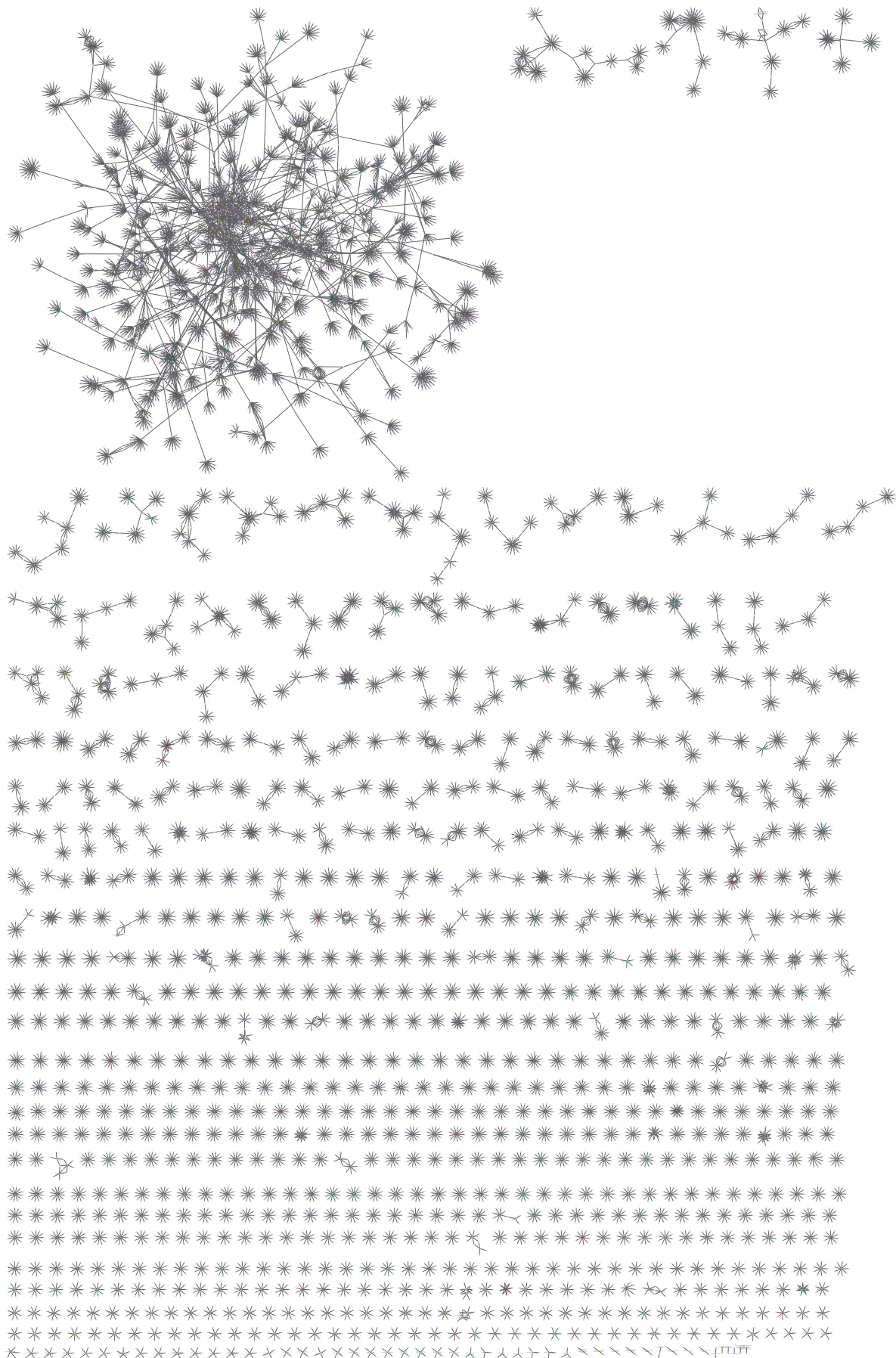


Figure A.1.1: The complete SNP-phenotype network (SPN), where SNPs are linked to phenotypes if an association has been found between them. Linkage disequilibrium (LD) between SNPs is taken into consideration, such that all SNPs in LD with each other are placed within the same LD block.

A.2 The Human Disease Network

The Human Disease Network (HDN) was presented in an article published in 2007 by Goh et al., and was described in Section 2.5. The network was constructed based on a dataset from the OMIM database [18]. The entire network is displayed in Figure A.2.1, where genetic diseases are connected to genes if an association has been found between them [18]. This network is separated into two distinct networks, the human disease network (HDN), where diseases are connected through mutations in common genes, and the disease gene network (DGN), where genes are connected when associated with common diseases [18]. The HDN is equivalent to the gene-phenotype-phenotype network constructed for this thesis.

The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) Proc Natl Acad Sci USA 104:8685-8690

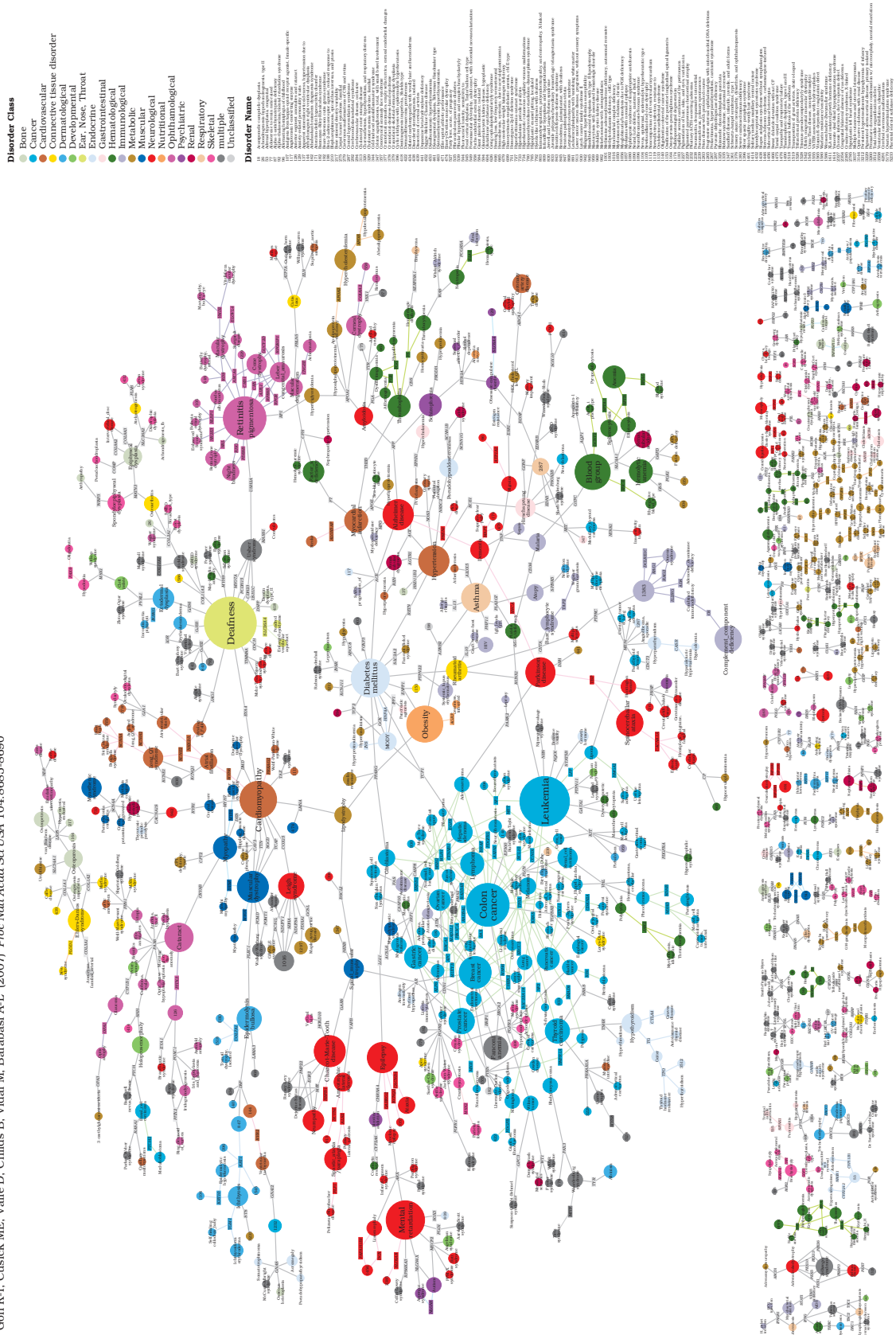


Figure A.2.1: The complete Human Disease Network from the Human Disease Network article by Goh et al. The disease categories included are shown to the right. With permission from PNAS, Copyright (2007) National Academy of Sciences, U.S.A.

A.3 Data Availability

The data used for the findings in this master thesis are from the UK Biobank and are found openly available at **PheWeb** (http://pheweb.sph.umich.edu/SAIGE-UKB/top_hits). This dataset was downloaded August 28th, 2020. The data containing phenotype information are found at **Lee Lab** (<https://www.leelabsg.org/resources>), and was downloaded September 16th, 2020.

