UNIVERSITÄT ZU LÜBECK

From the Institute for Cardiogenetics
of the University of Lübeck
Director: Prof. Dr. rer. nat. Jeanette Erdmann

Rare variants and coronary artery disease

Seltene Varianten und koronare Herzerkrankung

Dissertation
for Fulfillment of Requirements
for the Doctoral Degree of the University of Lübeck
- from the Department of Natural Sciences -

submitted by
Mariana Kleinecke
from Radebeul

Lübeck, the 18. September 2019

First referee:
Prof. Dr. Jeanette Erdmann

Second referee:
Prof. Dr. Hauke Busch

Date of oral examination:
16th of September 2019

Approved for printing. Lübeck, 17th of September 2019

Your time is limited, so don't waste it living some-
one else's life. Don't be trapped by dogma - which
is living with the results of other people's think-
ing. Don't let the noise of other's opinions drown
out your own inner voice. And most important,
have the courage to follow your heart and intu-
ition. They somehow already know what you truly
want to become. Everything else is secondary.

Steve Jobs (1955 - 2011), US computer engineer
& industrialist

Abstract

Coronary Artery Disease (CAD) is the number one killer in the world. It is a complex disease with many different mechanisms and factors underlaying it, and despite extensive research, many of these mechanisms are still unknown. Although CAD's heritability is estimated at 40-60% , less than 20 percent can be explained with the known variations so far. This work focuses on the role of rare variants in CAD.

Firstly, we had previously shown that a subset of our cohort of 255 myocardial infarction patients actually suffered from familial hypercholesterolemia (FH), a disease that leads to myocardial infarction (MI) via an increase of low-density lipoprotein (LDL) cholesterol. FH is treatable and treatment can, when diagnosed early, significantly reduce the risk for CAD. Hence, we were interested to see if there were other rare variants causing risk factors for CAD in our cohort. We looked at mutations from the Human Gene Mutation Database (HGMD®) and ClinVar known to cause obesity, hypertension, type 2 diabetes, high cholesterol, LDL cholesterol, triglyceride-levels and MI, and found that 17.25% of our cohort carry such a variant.

Additionally, we performed family studies in ten extensive families suffering from CAD to identify potential disease-causing rare variants. Three affected family members were exome sequenced and plausible variants were validated in other family members to analyze the co-segregation of disease with those variants. We were able to identify potential causative variants in five of our ten families, namely chr11:102650461A>T in *MMP10*, chr1:55085648C>T in *REN*, chr1:154574541G>C in *ADAR*, chr1:227203825T>C in *CDC42BPA*, chr16:23569403G>A in *UBFD1* and chr18:45566811C>T in *ZBTB7C*. Future functional studies will have to be done to determine the specific effect of these mutations.

Zusammenfassung

Koronare Herzerkrankung (KHK) ist nach wie vor die weltweit häufigste Todesursache. KHK ist eine komplexe Erkrankung mit vielen beeinflussenden Faktoren und Mechanismen. Viele dieser Mechanismen sind immer noch unbekannt und obwohl die Vererbung von KHK um 40-60% geschätzt wird, können nur weniger als 20 Prozent mit den bisher bekannten Varianten erklärt werden.

In vorheriger Arbeit ist es uns gelungen zu zeigen, dass um 5 Prozent unserer Kohorte von 255 Herzinfarkts (HI) Patienten eigentlich an familiärer Hypercholesterolemie (FH) leiden. FH ist eine Krankheit, die zu erhöhten low-density lipoprotein (LDL) Cholesterin Werten und damit einem erhöhten Risiko für HI führt. Wenn FH frühzeitig erkannt und behandelt wird, kann das Risiko für HI deutlich reduziert werden. In der vorliegenden Arbeit bestimmen wir das Vorkommen seltener Varianten von der Human Gene Mutation Database (HGMD®) und ClinVar, die dafür bekannt sind bestimmte Risikofaktoren für KHK zu verursachen. Wir konnten zeigen das 17.25% unserer Kohorte eine Variante für Fettleibigkeit, Typ 2 Diabetes, Bluthochdruck, hohe Cholesterin-, LDL-Cholesterin, und Triglyceride Werte tragen.

Zusätzliche haben wir zehn Familien studiert, um seltene Varianten mit starkem Effekt auf KHK zu identifizieren, in der Hoffnung einen weiteren Teil der fehlenden Vererbung zu erklären. Dafür wurden die Exome von drei betroffenen Familienmitgliedern pro Familie sequenziert und die so gefundenen plausiblen Varianten wurden in weiteren Familienmitgliedern validiert um die Co-Segregierung der Varianten mit der Krankheit zu untersuchen. Es ist uns gelungen Kandidaten in fünf dieser zehn Familien zu identifizieren. Dazu gehören die Mutationen chr11:102650461A>T in *MMP10*, chr1:55085648C>T in *REN*, chr1:154574541G>C in *ADAR*, chr1:227203825T>C in *CDC42BPA*, chr16:23569403G>A in *UBFD1* und chr18:45566811C>T in *ZBTB7C*. Der genaue Effekt dieser Varianten muss in zukünftigen funktionellen Studien weiter analysiert werden.

Contents

Contents

1. Motivation

Coronary artery disease (CAD) is the most common heart diseases and thus poses a severe public health burden [1]. For more than 15 years, heart disease has been the leading cause fo death globally. According to the World Health Organization (WHO), more than 30 million people died of heart disease in 2015, which accounts for 54% of deaths recorded for that year [2].

In previous work, we demonstrated that 5% of a cohort of 255 unselected German MI patients had familial hypercholesterolemia (FH) [3]. Familial hypercholesterolemia leads to elevated low-density lipoprotein (LDL) cholesterol levels leading to plaque formation at the aterial wall, which in turn promotes MI. However, when diagnosed early, FH is treatable and the risk for MI can be reduced [4]. This leads to the question, if there are any other genetic risk factors accumulated in our dataset. Therefore, we screened the cohort of 255 MI patients for rare variants known to cause MI, diabetes, obesity, hypertension, high cholesterol, and LDL cholesterol levels. Identfiying the underlying mechanism can lead to a tailored treatment, slow disease progression and improve their prognosis.

It is estimated that 40-60% of CAD is inherited [5], however our current knowledge can only explain less than 20% of this heritability, with many underlying mechanisms yet to be discovered [6].

So far, we have identified several common variants associated with CAD. These variants have a minor allele frequence (MAF) over 5% and explain less than 20% of the inherited proportion of the disease. Only a few rare variants have been identified (here denoted by a MAF <5%). Rare variants cluster within families and hence family studies are especially suited to identify these variants. In this study, we exome sequenced three individuals from ten multi-generation families severely affected with MI to identify novel rare variants causing CAD.

Although the variants identified in family studies may be family specific, it helps us to identify new CAD genes and in turn better understand the mechanisms underlying the disease. This could potentially lead to the identification of novel drug targets by targeting the specific genes/pathways identified. Additionally, CAD is often seen as a multifactorial disease and attempts have been made to calculate and predict the risk for the individual patient over the accumulation of risk variants. Identifiying further risk variants might improve those predictions [7].

2. State of the art

2.1. Introduction into Coronary artery disease - a brief history

It is difficult to determine the first occurrence of CAD, but descriptions of heart failure have been recorded as early as the ancient Egyptians [8]. Leonardo da Vinci was one of the first individuals to describe the coronary arteries, however, it was not until 1628 that the circulatory system was actually recognized as such by the physiologist William Harvey [9, 8]. Angina pectoris, the chest pain induced by CAD, was first described by the physician William Herberden over a hundred years later [9, 10]. In the second half of the 19th century the pathologist Ludwig Hektoen observed sclerotic changes in the artery and subsequent thrombosis causing myocardial infarction (MI) [10]. James Herrick, who coined the term 'heart attack', was the first person to diagnose acute MI by electrocardiography in the early 20th century [9, 10]. The history of CAD is strongly linked to the development of the techniques and technology during this period. Electrocardiography, measuring the hearts electrical activity, was one of the big discoveries for cardiology, dating back to the physiologist Augustus Waller in 1887 [11]. The one to build an actual recording device and distinguish the specific readings was Willem Einthoven in 1895 [11]. In the same year, Wilhelm Roentgen discovered electromagnetic radiation (X-rays), which have remained important for the visualization of the heart. The next important breakthrough was the first documented coronary heart catheterization performed by Werner Forssmann on himself in 1929. This technique was only later acknowledged by Cournand and Richards, who repeated and continued Forssmans work, and were the first to use it as a diagnostic tool [11][9]. This led to the development of angiography, a procedure in which the blood vessels are made visible via contrast agents and X-ray imaging. Mason Sones has to be mentioned here, as he improved this technology, resulting in selective coronary angiography, which uses a smaller amount

of contrast agent, applied directly to the relevant vessel [9]. In 1964, the first results of the Framingham study were published. It was the first study to successfully link smoking, high blood pressure and high cholesterol to the development of CAD, defining the term 'risk factors' for the first time. Not only was the Framingham study the first large case control cohort study implemented on CAD, but it was also the first study to incorporate women as part of the tested individuals [11, 10]. Dotter and Grüntzig are the pioneers of reopening blocked or narrowed vessels. Dotter was actually the first to reopen an occluded peripheral artery in 1964, while Grüntzig witnessed the procedure and successfully repeated the procedure in 1977 [9]. Another great development was the implementation called "coronary care units" into hospitals in early 1960. They assembled cases of acute MI together, whilst supplying them with specialized nurses, doctors, and equipment. This lead to a reduction of the acute MI -related death rate in the hospital to 15 % [9, 11, 10]. Coronary angioplasty and stenting then further reduced the death rate to 7%.

2.1.1. Pathophysiology

The major underlying cause for CAD is atherosclerosis, a disease where plaques form at the artery endothelium. A healthy endothelium is a monolayer of cells at the inner artery walls, and it has many functions. Among other things, the endothelium mediates thrombogenesis by expressing hemostatic factors like procoagulant and anticoagulants [12]. The endothelium also plays an important role in platelet adhesion, acts as a barrier to control vascular permeability, and regulates the remodeling of the vessel wall and the vasomotor tone by releasing vasoactive substances. Many of these functions are impaired in CAD, such as coagulation dysfunctions, leukocyte adhesion and migration, and vascular tone alterations [13]. Specifically, when the endothelium is subjected to oxidative, hemodynamic, or biochemical stimuli and inflammatory factors, it changes its permeability, promoting entry and retention of monocytes and LDL cholesterol [10] (see fig. 2.1). The accumulation of lipids leads to the formation of foam cells and the migration of smooth muscles cells into the plaque. The endothelial and smooth muscle cells start to proliferate and release extracellular matrix molecules which form a fibrous cap over the plaque.
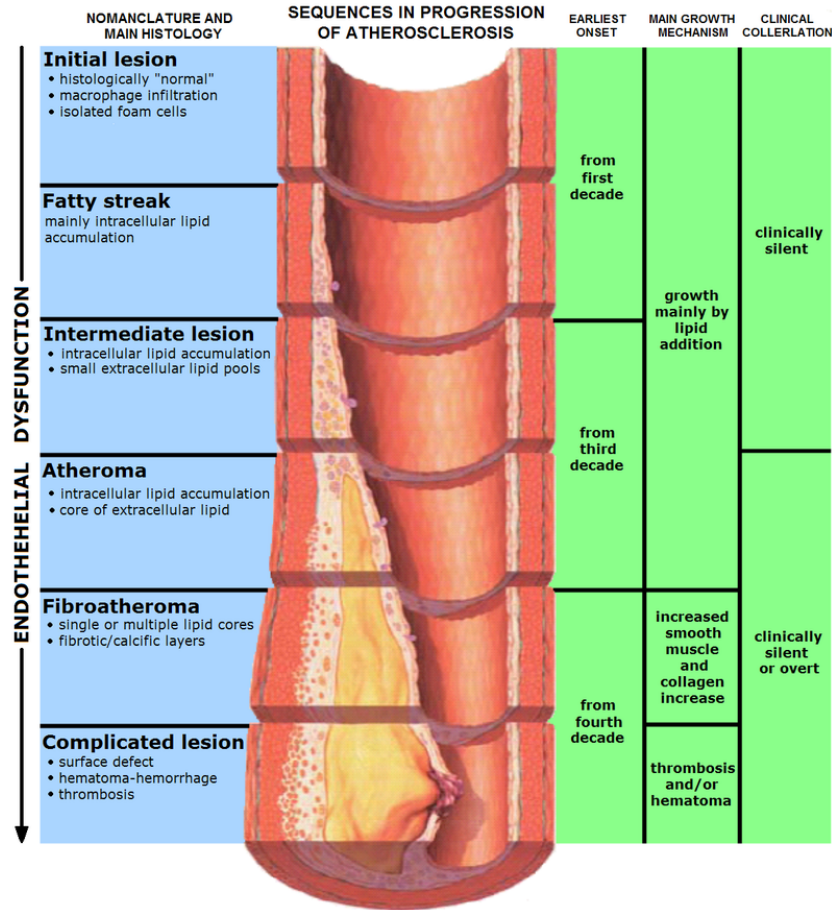
Figure 2.1.: The different stages of atherosclerosis [14].

The progression time of CAD can vary from years to decades, and for most part, it will be clinically silent. Symptoms can vary from individual to individual and most commonly include chest pain, shortness of breath, and sweating. In the event of a heart attack, a crushing pressure and pain in shoulder, arm or other locations can occur [15].

Depending on the development of the plaque, CAD can be divided into several classes. In stable angina, the plaque only limits the blood flow. When thrombosis is provoked by rupturing of the plaque, the blood flow can be temporarily interrupted, resulting in unstable angina. If the thrombus completely and permanently blocks the artery, the oxygen supply to the heart muscle tissue is cut off, resulting in irreparable muscle damage, scarring of the myocardial tissue and heart muscle death. This is commonly known as MI [16].

2.2. Genetics

Genetics is the study of heredity and genes, as well as their variation. It is a complex field not only involving biology, but also technological sciences such as bioinformatics and engineering. Considering that the human genome contains approximately three billion base pairs (bp) and these encode the blueprint for the whole human, errors are very likely and can lead to the development of disease [17].

These errors include insertions/deletions, chromosome abnormalities, mitochondrial variation and single nucleotide polymorphism (SNP) [18].

Insertion and deletions can be either non-frameshift, when exactly or a multiple of three bases are inserted/deleted, or frameshift, when the reading frame is disrupted. In a SNP, one bp has been changed. If this change does not affect the protein sequence (synonymous), the mutation is silent. Despite being silent, this form of variant can have an effect, for example on the regulation of the gene. A missense mutation leads to a change in amino acids (AA) which in turn changes the protein sequence (nonsynonymous). However, these variants might not change the protein function. In some cases the variant creates a new stop codon due to the change (stop-gain), which leads to a truncated or non-functioning protein. A variant can furthermore be gain-of-function (GOF) when it increases the gene and/or protein activity or a loss-of-function (LOF), when the protein/gene function has been decreased or completely lost [19, 20].

Single variation can be sufficient (single-gene inheritance) to cause disease. However, most diseases are caused by multiple variations (multifactorial inheritance), abnormalities with the chromosomes or mitochondrial variations. In single-gene inheritance a mutation in one gene can lead to disease in a dominant, recessive, autosomal or sex-linked fashion. These are also referred to as Mendelian diseases, and are very rare [18]. Multifactorial inheritance is more common, involving not only multiple genes and variations, but also the effect of variations on environmental factors and vice versa. A new study even suggests an 'omnigenic' mode, in which all expressed genes in a relevant tissue contribute to the disease [21].

The diseases with multifactorial inheritance are broadly classified as complex diseases, with CAD being a member of this group.

2.2.1. Genetics of CAD

Coronary artery disease has been estimated to have a heritability of 40-60% [5, 22]. There have been several approaches to determine the causality of CAD, a very successful approach being genome-wide association studies (GWAS). In GWAS, common variants are compared between controls and cases to determine if any of these occur more often in one group. The first GWAS were published eleven years ago and since then 161 loci have been associated to CAD [23, 24]. Less than half of these loci have been connected to lipid levels and other known risk factors, and many of the mechanisms underlying these loci and variants are still unknown [5, 22, 6]. Additionally, one of the limitations of GWAS is that an association is not equal to causality. Most of the GWAS loci feature multiple variants and lay in intronic regions. Without additional data, they cannot be clearly assigned to a specific gene. First of all, they may not lie within a gene to begin with, or they could be assigned to several different genes surrounding the loci. Instead of having a direct effect on the gene, these loci can also affect gene regulation by variating gene expression levels. These expression quantitative trait loci (eQTLs) can be local and affect surrounding genes, but they can also be distant and have an effect on the expression of a gene far away from the original loci [6, 5].
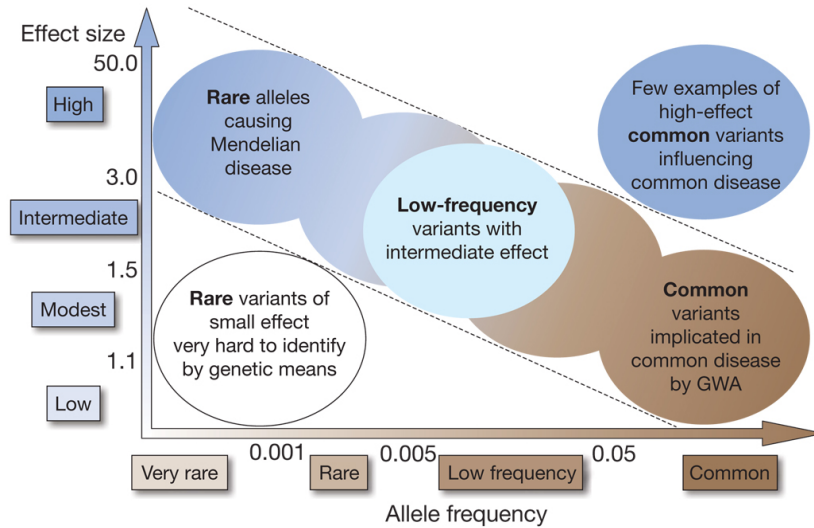

Figure 2.2.: Allele frequency and effect size of variants [25].

The loci identified by GWAS only explain a small fraction (<20%) [6] of CAD's heritability. Genome-wide association studies identify mostly common variants with a small effect size (see fig.: 2.2). With the increase in study sizes, we can identify variants

at lower frequency, and new loci reach genome-wide significance. However, rare variants with a small effect are generally difficult to detect. Only very few rare variants have been identified for CAD, with little contribution to the heritability of the disease [5, 6]. In families rare variants cluster, making family studies a suitable approach for the identification of these variants.

2.2.2. Risk factors

Coronary artery disease is not only modulated by genetic but also by environmental factors. Traditional risk factors used for calculating cardiovascular disease risk include gender, age, race, diabetes, dyslipidemia, hypertension, and smoking [13]. However, there are many more factors involved and these should be included in the risk prediction.

A positive family history is a strong marker for cardiovascular risk. Studies have shown that individuals with a family history of CAD have a 1.5- to 2.5-fold higher risk for CAD compared to those without family history, independent of other risk factors [26].

Age and gender also play an important role in CAD risk. After the age of 45 years, the risk for CAD significantly increases for men. For women, that age lies at 55 years. Gender also seems to make a difference when studying the genetic risk factors as some variants have been shown to have an effect in only one sex. For example, a variant in the *CPS1* gene has been associated with protection from CAD only in women [27].

Diabetes, hypertension (HT), smoking, lack of physical activity, obesity, mental stress and depression all promote cardiovascular disease. Looking at dyslipidemia, high level of triglycerides, total and LDL cholesterol pose a huge risk for the development of atherosclerosis, and with that CAD. For a long time, a low level of high-density lipoprotein (HDL) had been associated with an increased risk of CAD, which was disproved by Mendelian randomization studies [6]. Some of the many other risk factors associated with disease are homocystein levels and sleep apnea. Each of these risk factors impairs function through various complex mechanisms, many of which are still unknown.

It has been shown that reducing risk factors such as high level of LDL cholesterol, strongly decrease risk burden and disease progression. In the case of LDL cholesterol levels, a decrease by 1 unit reduces the risk of cardiac events by 54.5% [26] Therefore,

it is important to identify the underlying cause as well as responsible mechanism to improve the health of any affected individual [28].

In this work we will analyze several of these risk factors, in particular obesity, hypertension, type 2 diabetes and dyslipidemia.

2.2.2.1. Obesity

According to the WHO, 13% of the adult world population and 18% of children and teenagers were obese in 2016, a number that has tripled in the last 30 years [29].

The Body Mass Index (BMI) is used as a measurement of obesity. The BMI is defined as the ratio of weight in kilograms to height in meters squared (kg/m^2), with a BMI of 30 or greater being classified as obese for adults.

In general, obesity occurs due to an imbalance of energy intake and expenditure. The main contributors to the development of obesity are lifestyle factors including eating habits, physical inactivity, and stress. However, research has shown, that genetic and epigenetic factors promote weight gain and obesity with the heritability being estimated at 80% [30, 31, 32].

Over 100 genes have been found to be possibly associated with weight [33], including the melanocortin 4 receptor (*MC4R*) gene, and the fat mass and obesity-associated (*FTO*) gene [30]. The mode of action of these genetic factors are highly variable and include for example, a difference in the metabolic response to food intake [34], the regulation of the distribution of body fat, and the effect of physical activity [35]. Additionally, it is important to consider the genetic make up when treating an obese patient, as some variants can have an effect on the success of the treatment [36, 30]. Most of these variants seem to be working in a polygenic matter, but there have been monogenic mutations reported [31].

Obesity also increases the risk for type 2 diabetes, hypertension and dyslipidemia (amongst other things, high plasma triglycerides). It has been postulated to induce CAD through these comorbidities, with an excess risk of 44% [31]. However, it has been shown that obesity can independently predict the risk for atherosclerosis. For example, middle-aged women with an increased BMI show a 50% higher risk for CAD [37].

2.2.2.2. Hypertension

In 2010, approximately 31% of adults worldwide had an elevated blood pressure due to an increased cardiac output and/or systemic vascular resistance [38].

For adults, HT is defined as a systolic blood pressure over or equal to 140, or a diastolic blood pressure over or equal to 90mmHg [39], and it can be mainly divided into primary (essential) and secondary HT. The most common form is primary hypertension, accounting for 95% of all cases and has no identifiable primary cause [39]. Secondary hypertension is thought to be due to a primary defect, such as tumors or medication.

Hypertension can be caused by environmental factors, such as diet and physical activity, by a combination of common risk variants, by a single variant in a monogenic manner or by a combination of these factors. It is estimated that 40-50% of the variability in blood pressure is inherited [40, 32]. Additionally, the variant or variants can cause HT directly or indirectly by other mechanisms and disorders such as obesity. Over 100 loci have been identified to correlate with blood pressure [41] and include genes such as *CYP11B1* and *SCNN1B* that i.e. lead to an increased absorption of salt and water and with that, hypertension [42].

Most individuals with hypertension do not suffer from any symptoms, but it will greatly increase their risk for coronary heart disease [43].

2.2.2.3. Type 2 diabetes

In the year 2014, 422 million people suffered from type 2 diabetes, highlighting the public health burden of this disease [44, 45].

Individuals with type 2 diabetes have a deficiency or a resistance to insulin, a hormone important for the regulation of blood sugar levels. Therefore, diabetics suffer from high blood sugar which severely damages the body, especially the cardiovascular and the nervous systems [46]. In type 2 diabetes, environmental and behavioral risk factors such as physical inactivity and diet have a strong influence on the disease. Often developing as a result of obesity, type 2 diabetes is the most common form of diabetes (90% of all cases). It has been estimated, that the heritability of diabetes in general varies between 25 - 80% depending on the population [46, 45]. Type 2 diabetes is thought to be acting in a polygenic manner with only a few rare cases of monogenic variants causing the disease. So far 120 loci have been identified involving genes such as *CAPN10* and *PPARG* [45],

which are important in pathways for the secretion and the tissue sensibility of insulin [47, 48].

A diabetic individual has a two-fold increased risk for myocardial infarction [49].

2.2.2.4. Dyslipidemia

Dyslipidemia describes an abnormal amount of lipid, such as cholesterol, phospholipid, and triglyceride in the blood. Too much (hyperlipidemia), or too little (hypolipidemia) of the lipid can each be further subcategorized into primary and secondary dyslipidemia. While primary dyslipidemia is mostly caused by genetic variation, secondary dyslipidemia is caused by a different underlaying (genetic or non-genetic) condition such as obesity [50]. For our work, we are primarily interested in cholesterol and triglycerides.

The amount of total cholesterol in the blood is calculated from the amount of HDL and LDL cholesterol and triglycerides. Total cholesterol >200mg/dL is considered high, with levels >240mg/dL being very high [51]. Cholesterol is an essential component of the human cell membrane, and is required for the biosynthesis of bile acid, steroid hormones and vitamin D [52]. Being lipophilic, cholesterol needs to be transported through the bloodstream by lipoproteins. Based on their density, lipoproteins can be classified as chylomicrons, very-low density (VLDL), LDL and HDL lipoproteins [53, 54, 55].

An increased level of LDL cholesterol is termed hypercholesterolemia, and affects nearly 40% of adults worldwide [56]. The threshold at which an individual is considered to have hypercholesterolemia is at LDL cholesterol levels of ≥190mg/dL. As with other risk factors, there are environmental and genetic factors playing a role in the development of the disease. Familial hypercholesterolemia, the inherited form of hypercholesterolemia, is probably the most common inherited metabolic disease, with a reported prevalence of 1 in 250. It is mainly inherited in a dominant mode, and most mutations were observed in the *LDLR*, *APOB*, *apoB100* and *PSCK9* genes. These variants affect the pathway for cholesterol clearance, by for example, reducing the uptake of LDL cholesterol over the LDL receptor, which then leads to an increased turnover of the LDL cholesterol into foam cells, building a part of the atherosclerotic plaque [57]. Hypercholesterolemia significantly increases the risk for atherosclerosis and hence MI. However, the risk depends on the duration and the amount of increased cholesterol [58, 59, 60].

Triglycerides, the main component of body fat, are lipophilic molecules which are mainly transported via chylomicrons and VLDL lipoproteins [51]. Approximately 10% of men over 30 and women over 60 have elevated triglyceride levels [61], with levels over 200mg/dL being considered high and levels over 500mg/dL being considered 'very high'.

There are several causes for high triglyceride levels. For one, it can be a result of other factors such as obesity and diabetes. On the other hand, there are genetic variants known to cause this condition in diseases such as familial hypertriglyceridemia or familial combined hyperlipidemia. These are inherited both in a monogenic or a polygenic matter, although, monogenic hypertriglyceridemia is very rare and includes genes such as *LPL* and *GPIHBP1*. Variants in these genes can impair either the clearance or transportation pathways of triglycerides [62, 51].

Like hypercholesterolemia, hypertriglyceridemia increases the risk for CAD [50, 63, 64]. Triglyceride levels have been shown to increase the risk for CAD independently from other riskfactors, as individuals with >150mg/dL have a 15-25% increased risk compared to controls [65, 66, 64].

2.2.3. Sequencing strategies

The ability and capability of generating nucleotide sequences has come a long way, from sequencing less than one kilobase in the 1970s, to full genomes and 120 gigabyte of data output in the next generation sequencers [67]. With the improvement of technology, the sequencing price significantly dropped, making it possible to sequence a whole genome for roughly $1000 with the costs still falling [68]. This development opens up many new possibilities not only for research but also for medical purposes. The vision is that a scan of the genome could identify the risk a patient has for the disease and also the correct treatment methods, as for example, some variants can have a significant effect on the drug response [69]. There are different scales of sequencing techniques for different purposes. Whole-genome sequencing (WGS) is the most expensive of these and poses a huge bioinformatic burden. The sheer amount of information produced needs careful handling and filtering to reveal useful information. It can capture everything and is therefore an amazing tool to identify new pathways and disease relations. Panel sequencing contains specific genes of interest which leads to a cost reduction, but not all will be captured and causal variants missed due to the predetermination of genes. This approach is very useful for those screens where the pathways and causal variants

are well known. Whole-exome sequencing (WES) is a good alternative as it contains only the coding sequence which should facilitate the evaluation of the results. However, it will fail to capture regulatory variants as these often lie in the non-coding area.

2.2.3.1. Sanger sequencing

Sanger sequencing is the method of choice for validating specific single variants, as it produces short sequences with a low error rate. Nowadays, all it needs are the deoxyribonucleic acid (DNA) sample, a primer, deoxynucleoside triphosphates (dNTPs), fluorescently-labeled dideoxynucleotide triphosphates (ddNTPs) and the polymerase (fig.2.3). In the first step the DNA sample will be heated to denature the DNA strand into single strands. The temperature will be reduced to allow primer binding. The polymerase will then synthesize a new DNA strand starting from the primer and will incorporate the appropriate dNTPs until it incorporates a ddNTP by chance, which leads to an immediate stop of the elongation. This process will be repeated for several cycles, resulting in DNA strands with different lengths that will cover each position of the requested sequence. These will be separated by length over a capillary gel electrophoresis, where shorter strands travel faster through the gel. The final ddNTPs are specifically labelled with different colors for each nucleotide, which can then be measured by a laser when they are running through the capillary [70].
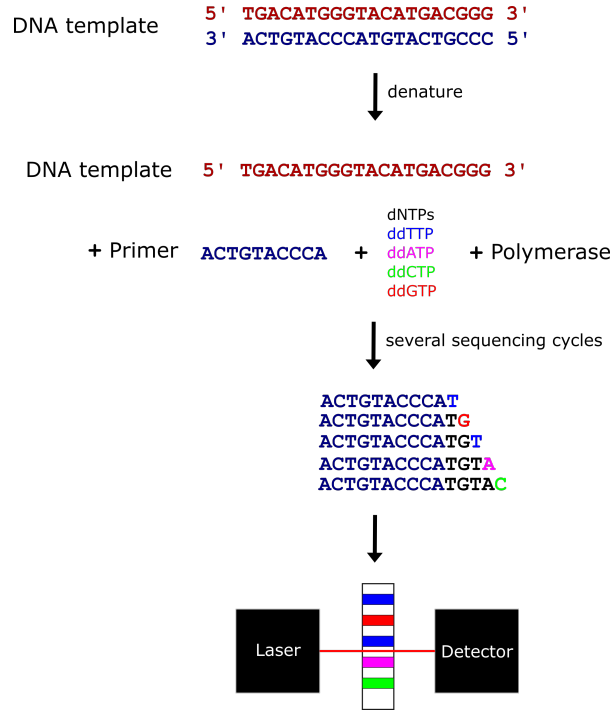
Figure 2.3.: Flow chart of Sanger sequencing process.

2.2.3.2. Exome sequencing

There are several sequencing approaches for WGS and WES. In the following, we will only describe the Illuminas Genome Analyzer [71]. For WES the first step is to enrich the requested sequence, in this case the whole coding region. Then the library is prepared by fragmenting the sample into random pieces and adding adapters to these fragments (see fig. 2.4). These fragments are then amplified, purified and loaded into a flow cell. In these, every fragment will be individually amplified via bridge amplification to build clonal clusters of around one million copies each. After the completion of this step the sequencing can begin. The DNA polymerase is added as well as the four nucleotides. The nucleotides are fluorescently labelled and chemically blocked so no further incorporation happens, similar to Sanger sequencing. After each incorporation, the nucleotides are measured and then the chemical block gets removed to allow new incorporation and the process starts anew. At the end all reads are evaluated and filtered for poor quality [72]. This produces a massive amount of relatively short (around 200bp) long sequences, which will have to be aligned to a reference in further downstream analysis.
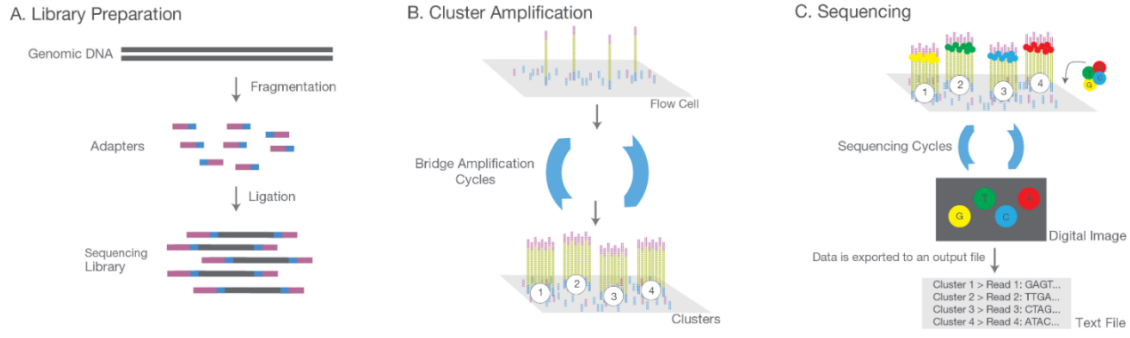
Figure 2.4.: Flow chart of Illumina sequencing process [72].

2.3. Aims of this thesis

Coronary artery disease is the number one cause of death in the world and it is a complex disease, with many complicated gene and environmental factors underlying it. Some of the risk factors for CAD, such as obesity, hypertension, type 2 diabetes, increased level of cholesterol, LDL cholesterol and triglycerides, have rare genetic variants underlying them.

We previously identified the prevalence of familial hypercholesterolemia patients in a cohort of 255 myocardial infarction patients. Thus, we aim to identify the prevalence of other risk factors in this cohort to explore the frequency of treatable risk factors in the dataset. We aim to identify known risk variants in our cohort, as well as risk phenotypes and their occurrence in our data.

Despite all the research on CAD, the mechanisms are still poorly understood, and the identified genetic variants still only explain a small portion of the disease. To identify new variants that are too rare to be linked with CAD in GWAS, we use family studies to explain some of this missing heritability. These newly identified variants might point to new genes possibly involved in the development of atherosclerosis, and with that, potential mechanisms underlying it. This is not only important for understanding the disease, but also to identify potential drug targets and treatment options tailored for each individual patient (personalized medicine).

3. Material and methods

3.1. Bioinformatics

3.1.1. The German MI family study (Germif)

The underlying data consists of 401 German families with an accumulation of CAD. In this study, the inclusion criteria were as follows: the index case had a clinical manifestation of CAD at young age (<60) and there is at least one additional affected sibling with CAD before the age of 70 [73]. Ten of these families were chosen for exome sequencing and analyzed in this work (see section 3.1.2). Additionally, the exome sequencing data of 255 unrelated patients from the Germif study were used for analysis. We conducted our analysis with this data on prealigned and called variant files (VCF - variant call format) [74]. These files lists all variants with, amongst others, genomic coordinates, the quality of the sequencing, the reference and alternative allele. We used hg19/GrCh37 assembly version throughout this work.

3.1.2. Preparatory work

For exome sequencing, we selected the most promising ten of our 401 CAD affected families. For this, we drew pedigrees for all families to allow a clear overview of each family structure. We ranked and filtered all families based on the family size, the amount of affected (minimum three) and unaffected (minimum five) family members and the genetic variability (at least one affected cousin). These steps ensured that we increased the chance of finding the cause of disease in this family as we minimized the number of non-pathogenic variants and maximized the number for improved statistics.

3.1.3. Genome Analysis Toolkit

We established a processing pipeline for the raw exome-sequencing data (chapter.3.2.2.1) using the Genome Analysis Toolkit (GATK) best practices pipeline[75, 76].

The pipeline works as follows (see fig.3.1):
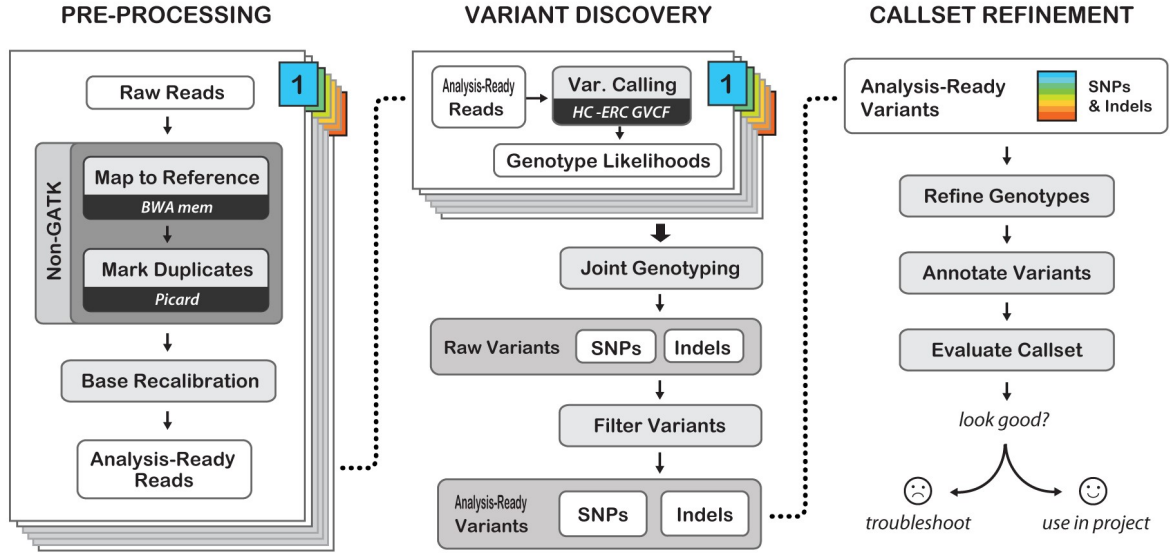
Figure 3.1.: GATK best practices pipeline [77].

We received the raw sequencing reads as FASTQ files, a text-based format that contains the short reads of around 150bp in length and the corresponding quality scores for each called base in the sequence reads. First, we mapped these reads to the human reference genome using the Burrows-Wheeler Aligner (BWA) version 0.7.15 (code snippet 1). We used the newest BWA-mem algorithm as it is supposed to have a better performance, be faster and more accurate [78].

Code snippet 1 Mapping the reads to a reference genome.

```
#mapping
bwa mem -M [split read handling on] -R [RGline] -t [nr of threads]
reference.fasta \
'< zcat $(find -name "*R1_001.fastq.gz" |sort -n)' \
'< zcat $(find -name "*R2_001.fastq.gz" |sort -n)' \
> aln_pe_$1.sam
```

This results in a Sequence Alignment Map (SAM) file containing all the information from the FASTQ files as well as positions of the reads in the reference and mapping quality scores. We then used Picard version 1.119, a set of command line tools, to compress and sort the SAM file into Binary Alignment Map (BAM) format and to mark all duplicates to remove biases from polymerase chain reaction (PCR) amplification. The BAM file is then indexed via samtools [79] (code snippet 2).

Code snippet 2 Compress and sort SAM file into BAM format and remove duplicates.

```
#convert from sam to bam
java -jar SortSam.jar INPUT=aln_pe_$1.sam OUTPUT=aln_pe_$1.bam
SORT_ORDER=coordinate

#mark and remove duplicates
java -jar ~/Programs/picard-tools-1.119/MarkDuplicates.jar
INPUT=aln_pe_$1.bam
OUTPUT=aln_pe_dedupped_$1.bam METRICS_FILE=metrics_$1

#indexing Bam file
samtools index aln_pe_dedupped_$1.bam
```

For every further step GATK nightly build 2016-07-15-ge5173a7 was used [80].

Due to the large amount of reads, some of these might be mapped incorrectly. GATK has the ability to identify regions which are difficult to map such as insertions and deletions and remap these to improve the quality of the mapping (code snippet 3). In the new versions of GATK these steps have become obsolete and are no longer part of the best practices.

Code snippet 3 Realignment of potentially incorrectly mapped reads.

```
# create list of intervals for realignment
java -Xmx8g -jar GenomeAnalysisTK.jar \
-nt [nr of threads] 4 \
-I input.bam \
-R reference.fasta \
-T RealignerTargetCreator \
-ip 100 \
-L exons_intervals_sorted_merged.bed \
-known Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz \
-o realigned_$1.intervals

#realignement
java -XX:+UseParallelGC -XX:ParallelGCThreads=8 -Xmx8g -jar GenomeAnalysisTK.jar \
-I Inputfile.bam \
-R reference.fasta \
-T IndelRealigner \
-targetIntervals realigned_$1.intervals [output from previous step]\
-knownSites Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz] \
-o realigned_$1.bam
```

There are multiple systematic technical errors that happen, such as biases due to the biochemical process and sequencing technique. In the next step GATK uses the base quality scores given to each called base during sequencing to collect statistics and build a model, which it uses to adjust/recalibrate the quality scores and the dataset. For this, it uses databases of known human variants (found at dbSNP [81]) and known variable indel sites (found at Mills and 1000G gold standard [82]) (code snippet 4).

Code snippet 4 Recalibrating reads.

```
#recalibrating reads
java -Xmx8g -jar GenomeAnalysisTK.jar \
-nct 4 \
-T BaseRecalibrator \
-R reference.fasta \
-I  input.bam \
-ip 100 \
-L exons_intervals_sorted_merged.bed \
-knownSites dbsnp_138.hg19.vcf.gz \
-knownSites Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz \
-o recal_data_$1.grp

#recalibration
java -Xmx8g -jar GenomeAnalysisTK.jar \
-nct 4 \
-T PrintReads \
-R reference.fasta \
-BQSR recal_data_$1.grp [output from previous step] \
-I realigned_$1.bam \
-o realigned_$1.recal.bam
```

This completes the preprocessing of the data files and they are then ready for the variant calling. We called variants using the HaplotypeCaller in Genomic Variant Call Format (GVCF) mode. The difference to the usual Variant Call Format (VCF) format is that GVCF records every position in the genome even if there is no variant present. This enables us to differentiate between a variant that has not been called or the individual carrying the reference allele. In the next step all GVCF files were genotyped together by the GenotypeGVCFs command (code snippet 5). The joint analysis improves sensitivity and detection at difficult sites.

Code snippet 5 Variant calling.

```
#variant calling
java -Xmx8g -jar GenomeAnalysisTK.jar \
-nct 4 \
-T HaplotypeCaller \
-R reference.fasta \
-I realigned_$1.recal.bam \
-ip 100 \
-L exons_intervals_sorted_merged.bed \
--emitRefConfidence GVCF \
--variant_index_type LINEAR \
--variant_index_parameter 128000 \
--dbsnp dbsnp_138.hg19.vcf.gz \
-o output_$1.raw.snps.indels.g.vcf


#joint genotyping

java -Xmx8g # -jar GenomeAnalysisTK.jar \
-nt 4 \
-R reference.fasata \
-T GenotypeGVCFs \
--variant file1_output_$1.raw.snps.indels.g.vcf \
.
.
.
--variant file30_output_$1.raw.snps.indels.g.vcf \
-o output_all_LUEB0001_0030.vcf
```

Since the GATK variant calling tools have their main focus on increasing the sensitivity, filtering the variants is of importance to increase the specificity of the called variants. GATK offers a variant quality score recalibration (VQSR), which uses machine learning and databases of known variants (here Hapmap [83], the 1000 Genomes Project (1000G) [84] and the Single Nucleotide Polymorphism Database (dbSNP) [81]) to calculate a score that identifies the likeliness for a variant to be true and assigns a specific score to each variant. This score is then used to filter the variants. The final GVCF file is then ready for annotation (code snippet 6).

Code snippet 6 Recalibrating variants.

```
#variant recalibration
 java -Xmx4g -jar GenomeAnalysisTK.jar
-T VariantRecalibrator \
-nt 4 \
-R reference.fasta \
-input output_all_LUEB0001_0030.vcf \
-resource:hapmap,known=false,training=true,truth=true,
prior=15.0 hapmap_3.3.hg19.sites.vcf \
-resource:omni,known=false,training=true,truth=false,
prior=12.0 1000G_omni2.5.hg19.sites.vcf \
-resource:1000G,known=false,training=true,truth=false,
prior=10.0 1000G_phase1.snps.high_confidence.hg19.vcf \
-resource:dbsnp,known=true,training=false,truth=false,
prior=6.0 dbsnp_138.hg19.vcf \
-an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR \
--maxGaussians 6 --minNumBadVariants 4000 \
-mode SNP \
-recalFile output.recal \
-tranchesFile output.tranches \
-rscriptFile output.plots.R

#apply recal
java -Xmx4g -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-nt 4 \
-R reference.fasta \
-input output_all_LUEB0001_0030.vcf \
--ts_filter_level 99.0 \
-tranchesFile output.tranches [from previous step] \
-recalFile output.recal [from previous step]  \
-mode SNP \
-o output.recalibrated.filtered.vcf.tmp.vcf

java -Xmx4g -jar GenomeAnalysisTK.jar \
-R /media/QNAP/Public/GATKResources/2.8/hg19/ucsc.hg19.fasta \
-T VariantFiltration \
--variant output.recalibrated.filtered.vcf.tmp.vcf \
--filterExpression "vc.isIndel() && QD < 2.0" --filterName "QDfilter" \
--filterExpression "vc.isIndel() && FS > 200.0" --filterName "FSfilter" \
--filterExpression "vc.isIndel() && ReadPosRankSum < -20.0"
--filterName "ReadPosRankSumfilter" \
-o output.recalibrated.filtered.vcf
```

3.1.3.1. ANNOVAR

In this work we used ANNOVAR [85] to annotate variants from exome sequencing data. The software performs three types of annotation. Gene-based annotation defines if the variant is found in an intron, exon, intergenic region, untranslated region (UTR) or a splice site. Additionally, this annotation gives information about possible AA changes. For this annotation, ANNOVAR utilizes data from the NCBI Reference Sequence Database (RefSeq) [86], USCS [87], Ensembl [88], GENCODE [89] or user-defined tables. In the region-based annotation ANNOVAR uses ENCODE data [89] to annotate variants to specific regions such as duplicated regions and/or transcription factor binding sites. It also determines whether or not the variant is located in a conserved region. The third annotation is based on filters such as position and nucleotide exchange. Depending on the requested databases, this annotation will give information about the pathogenicity of a variant via known connection to diseases or prediction scores, as well as the allele frequency.

The databases used by ANNOVAR as well as additional databases used for our analysis will be further described below.

3.1.4. Databases

All databases were accessed between July 2014 and February 2018 (tab.: 3.1).

For the screening of rare reported pathogenic variants in our data set, we mainly utilized the data of four databases: ClinVar [90], HGMD® [91], 1000G [84] and the Exome Aggregation Consortium (ExAC) [92].

ClinVar and HGMD® provide information about the relationship of variants and their associated disease with supporting evidence or corresponding literature.

The variants noted in ClinVar have been submitted by researchers alongside their interpretation and clinical relevance, and it is freely accessible. In comparison, HGMD® takes the variants and their relationship to human disease from publications and the most up to date version can only be accessed when paid for. However, there is a public version freely available for registered users from academic institutions/non-profit organizations.

Both ExAC and 1000G were used throughout this thesis to determine the frequency of variants. The Exome Aggregation Consortium consists of 60,706 unrelated individuals from a variety of disease-specific and population genetic studies. The consortium was founded to combine these various projects and make them accessible for the scientific community [92]. The 1000 Genomes project [84] contains the variants of 2,504 sequenced individuals, including deletions and single nucleotide polymorphisms. Its goal was to find the majority of genetic variants with frequencies over 1% in the healthy population.

Three further databases utilized for determining the frequency of the identified variants were the NHLBI Exome Sequencing Project (ESP) [93], dbSNP and control data from Tim Strom at the Institute for Human Genetics at the Helmholtz Center in Munich.

The ESP is a project that aimed to discover novel genes and mechanism responsible for heart, lung and blood disease . It contains the exome data of 6,503 African-American and European-American samples with phenotypes ranging from heart to lung diseases. Phenotype information is however not available in the database [93]. Despite its name, the dbSNP contains collected data of small-scale multi-base deletions or insertions, retroposable element insertions and microsatellite repeat variations besides the SNPs [81]. It is a public archive collecting data from various organisms and projects, and aims to support research into a wide range of biological problems like physical mapping, functional analysis, pharmacogenomics, association studies and evolutionary studies. The control data by Tim Strom includes the variants of 3,239 individuals without heart disease. The allele frequency (AF) of our variants in the control data was determined, to compare these with the AF in the MI data.

As noted in the previous section there were several databases involved in variant annotation with our analysis focusing on the Reference Sequence (RefSeq), USCS and Encyclopedia of DNA Elements (ENCODE) data.

While RefSeq provides a well-annotated set of genomic DNA, transcript and protein sequences [86], the UCSC data consists of the merged data sets of several different projects such as RefSeq and GenBank [94, 95].

The UCSC Genome Browser database offers information on genomic sequence for a variety of organisms, primarily vertebrates. In addition, regulatory information from the ENCODE project is easily accessible. The Encyclopedia of DNA Elements contains information about regulatory elements, controlling gene activity and acting on protein or RNA levels [96].

Several variant scores (see 3.1.4.1) were used to filter our variants. The SIFT, PolyPhen2 and MutationTaster score were taken from the dbSNP database, a database for functional prediction and annotation of human variants. While writing this thesis it contained 83,422,341 variations [97].

To check our data for possible eQTLs, HaploReg v4.1 was used. HaploReg is a tool developed by the Broad Institute that aims to give information about the possible impact of non-coding variants on disease. It contains the eQTL data from, for example, the Genotype-Tissue Expression (GTEx) project [98], GEUVADIS [99] as well as others [100]. In the GTEx project, tissue-specific gene expression and regulation has been studied by taking samples from 53 non-diseased tissue sites across nearly 1000 individuals. The GEUVADIS project contains mRNA and sRNA sequencing data from lymphoblastoid cell lines of 465 individuals.

HaploReg is a tool for exploring annotations of the non-coding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using LD information from the 1000 Genomes Project, linked SNPs and small indels can be visualized along with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE projects, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. HaploReg is designed for researchers developing mechanistic hypotheses of the impact of non-coding variants on clinical phenotypes and normal variation.

For protein modeling, the freely accessible RCSB Protein Data Bank (PDB) was used to obtain structural information for specific genes. The PDB contains structures of biological molecules like proteins and nucleic acids. It was established in 1971 and includes more than 77,000 3D structures [101].

Table 3.1.: Data and databases and their use in this work.

Database	Used to analyze	Reference
ClinVar	known pathogenic variants	[90]
Human Gene Mutation Database®	known pathogenic variants	[91]
Reference Sequence collection	variant annotation	[86]
UCSC Genome Browser Annotation database	gene prediction	[94, 95]
ENCODE	regulatory annotation	[96]
NHLBI Exome Sequencing Project	frequency	[93]
1000 Genomes project	frequency	[84]
Exome Aggregation Consortium	frequency	[92]
Single Nucleotide Polymorphism database	frequency	[81]
Tim Strom database	controls	
dbNSFP database	variantscores	[97]
HaploReg	eQTLs	[100]
RCSB Protein Data Bank	protein structure	[101]

3.1.4.1. Variant scores and phenotype thresholds

The following scores where used to analyze and filter variants:

PhastCons PhastCons is a conservation score that identifies evolutionarily conserved small regions in genomic sequences based on the phylogenetic hidden Markov model [102].

SIFT The Sorting Intolerant From Tolerant algorithm (SIFT) predicts the effect of a variant on the protein function [103]. The approach is based on the degree of conservation of these AA, as it assumes that protein function is correlated to protein evolution. Hence, specific positions should be conserved in the alignment of protein families. Scores under 0.05 are predicted to be deleterious, those greater or equal to 0.05 are predicted to be tolerated.

PolyPhen2 Polymorphism Phenotyping v2 (PolyPhen2) is calculated based on the AA sequence, physical (structure) and evolutionary comparisons [104]. It estimates pairs of the probability to identify a benign variant as damaging regarding the protein structure and function (false positive rates) and the probability to identify a pathogenic variant as such (true positive rates). Variants are subsequently classified based on false positive rate thresholds as benign (over the threshold), possibly damaging (under the higher threshold) or probably damaging (under the lower threshold).

MutationTaster MutationTaster uses the data of different biomedical databases to score the deleteriousness of human genetic variants [105]. For that, scores like evolutionary conservation, splice-site changes, loss of protein features and changes that might affect the amount of mRNA are analyzed by a naive Bayes classifier.

CADD-score Combined Annotation Dependent Depletion uses a wide range of information about the nature of different variants, such as conservation, functional genomic data, transcript information and combines them into one score predicting the effect of a variant [106]. It combines known scores like SIFT, Polyphen and GERP and calculates a metascore for every base pair substitution.

Z- and pLI score These two scores are provided by the ExAC database [92]. The Z-score equals the deviation of observed counts from the expected amount of synonymous or missense variants in a gene. Hence, a positive score means there were less variants observed than expected, which indicates a certain intolerance to variance in that gene. The pLI score calculates the same as the Z-score, however, the pLI score is calculated for LOF variants. A gene with pLI score over 0.9 is considered extremely intolerant for LOF variants.

The data set of German MI families used in this work also contains medical information about each patient such as blood test results, weight, height, CAD related phenotypes, and risk factors including smoking and diabetes.

We used the following (standardized) thresholds to phenotype individuals based on BMI, cholesterol, LDL cholesterol, triglycerides and systole/diastole readings (table:3.2).

Table 3.2.: Phenotype thresholds used in this work.

Phenotype	Threshold	Reference
Obesity	BMI $\geq$ 30	[29]
High Cholesterol	Borderline high: total Cholesterol $\geq$ 200	
	High: total Cholesterol $\geq$ 240	[107]
High LDL-Cholesterol	LDL-Cholesterol $\geq$ 190	[107]
High Triglycerides	High: triglycerides $\geq$ 200	[108]
	Very high: triglycerides $\geq$ 500	
Hypertension	Systole $\geq$ 140 / Diastole $\geq$ 90	[109]

In regards to their LDL-cholesterol thresholds, we had to calculate corrected cholesterol level for some patients as they were receiving statin therapy. All values were calculated using the CURVES study [110]. Table 3.3 shows the percentage of reduction taken from

the study. In case a daily dose value sits between two given doses, the average has been used for calculation. For instance, individual 4349503 has a measured LDL cholesterol of 140 mg/dL. He takes 40 mg Atorvastatin a day, which leads to a reduction in LDL cholesterol of 51%. The measured 140 mg/dL is around 49% of the initial level. Hence he had a previously (corrected) LDL cholesterol of 140mg/dL / 0.49 = 285.7mg/dL.

Table 3.3.: Treatment, daily dose and their corresponding mean percent reduction of LDL cholesterol.

Treatment	daily dose [mg]	mean change of LDL cholesterol [%]
Atorvastatin	10	38
Pravastatin	10	19
Simvastatin	10	28
Atorvastatin	20	46
Pravastatin	20	24
Simvastatin	20	35
Fluvastatin	20	17
Lovastatin	20	29
Atorvastatin	40	51
Pravastatin	40	34
Simvastatin	40	41
Fluvastatin	40	23
Lovastatin	40	31
Atorvastatin	80	54
Lovastatin	80	48

3.1.4.2. Statistical analysis of the families

For each family, an age cut-off was calculated. First we calculated the arithmetic mean of disease with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We then added one standard deviation $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$ to the mean to determine the age cut-off.

We additionally calculated the odds ratio with 95% confidence intervals and the LOD score for each variant in every family. The odds ratios were calculated with $OR = \frac{\frac{a}{b}}{\frac{c}{d}}$ where a stands for the amount of affected individuals with the genetic variation, b for healthy individuals with variation, c are sick individuals without the genetic variance and d are healthy individuals without it. We calculated the confidence intervals from the odds ratio with $e^{(log(OR)\pm[1.96xSE(log(OR))]}$ with $SE(log(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

The LOD score calculations and graph where done with R version 3.5.1. The script used for the calculation can be found in the supplement.

3.1.4.3. Protein modeling and molecular dynamics simulation

To analyze a possible structural effect of a variant we modeled the partial protein with the variant and analyzed the effect by performing a molecular dynamics simulation for both the wild type and the mutant protein. The protein sequence and structure were retrieved from the PDB as previously described. We used homology modeling to model the structure of the mutant protein part using Yet Another Scientific Artificial Reality Application (YASARA) [111]. The model was subsequently subjected to an energy minimization with the YAMBER3 force field [112].

The YASARA molecular simulation program [113] and the YAMBER3 force field were used in this work to perform the molecular dynamics (MD) simulations of the protein in a water-filled cuboid box, considering a minimum distance between protein and box wall of 10Å. Since we modeled only a part of the protein, we added terminal caps to avoid the net charge of normal N-termini (NH3+) and C-termini (COO-). The simulations had a duration of 10ns without constraints, with a constant temperature of 298K and 1 bar pressure. The system was first equilibrated for 250ps and the integration time step was 2.5fs. A cut-off of 7.86Å was used for Van der Waals interactions and the particle mesh Ewald algorithm [114] was used for long-range electrostatic interactions without any cut-off. The coordinates have been saved every 2ps as a trajectory.

3.2. Laboratory

3.2.1. Material

3.2.1.1. Equipment

- Tubes 0.2 & 0.5µl - Greiner (Kremsmünster, Germany)

- Bio-Photometer - Eppendorf (Hamburg)

- Pipettes 0.1-2.5µl, 1-10µl, 2-20µl, 20-200µl, 100-1000µl - Eppendorf (Hamburg, Germany)

- Pipette tips - Eppendorf (Hamburg, Germany)

- Labcycler - Sensoquest (Göttingen, Germany)

- Universal Hood II - BioRad (Hercules/California, USA)

- NucleoSpin® Gel & PCR Clean-up - Macherey Nagel (Düren, Germany)

- Collection tubes 2ml - Macherey Nagel (Düren, Germany)

- Biofuge heraeus pico - Heraeus (Frankfurt am Main, Germany)

- Freezer -20 electronic - Siemens (Berlin, Germany)

- Fridge - Liebherr (Bulle, Switzerland)

- WypAll - Kymberly Clark (Dallas, USA)

- Erlenmeyer flask borosylicate glass - Schott (Mainz, Germany)

- Autoclave - Systec (Wettenberg, Germany)

- Vortex - Micro-Bio-Tec-Brand (Gießen, Germany)

- UV-cuvettes - Micro-Bio-Tec-Brand (Gießen, Germany)

3.2.1.2. Chemicals

- 5'Prime Mix - 5PRIME (Hamburg, Germany)

- MyTaq® polymerase - Bioline Pharmaceutical AG (Baar, Switzerland)

- MyTaq® mix - Bioline Pharmaceutical AG (Baar, Switzerland)

- Agarose - Biozym Scientific GmbH (Hessisch Oldendorf, Germany)

- Tris - Carl Roth (Karlsruhe, Germany)

- Boric acid - Sigma Alderich (St. Louis, USA)

- EDTA - Sigma Alderich (St. Louis, USA)

- Oligonucleotides - Eurofins Genomics (Ebersberg, Germany)

- Ethanol - Sigma Alderich (St. Louis, USA)

- Elution buffer NE - Macherey Nagel (Düren, Germany)

- Washing buffer NT3 (+ 200ml ethanol) - Macherey Nagel (Düren, Germany)

- Binding buffer NTI 1:5 mix with ddH2O - Macherey Nagel (Düren, Germany)

3.2.1.3. Primer

The primers were designed using Primer3 [115] with a product size of 600-700bp and a primer size of up to 20bp. The synthesis has been done by Eurofins Genomics. The following tables show the sequence, the annealing temperature used and PCR-reagents for each primer.

Table 3.4.: Primers used for sequencing validation in family 1 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
neu3	GGGAAAGTCTTCTCTCAGAATG	CAAGCAGGCTTGAATTGTAAC	MyTaq, 59-64
Chr3:51908114	CTGCCCAGTGCTGAACTTTA	TGCATTGTTGCACCTTACAA	MyTaq, 59-64
Chr11:102650461	CACACTTCAGTCTTAAACCAGGTC	TGCCAGGAAAGGAGCTGA	MyTaq, 59-64
Chr19:5705841	AGCTATGATGGTGCCTTTGC	CACCTGCCACACCAAGAAG	MyTaq, 59-64
BSN_49679930	TCAGGTGAAGGAGTGGCTCT	TCTGCCATGCTTTGGACTG	MyTaq, 59-64

Table 3.5.: Primers used for sequencing validation in family 2 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr1:204125895	CAACCTCTGGTTGGTCTTGG	AGAGAGCAAATGGTGGCTGT	MyTaq, 59-64
Chr7:112090748	CTCCCGCGCTAGAGAGAAAC	CCTGTCCCGACACACTCTC	MyTaq, 59-64
Chr7:142829211	CTGAATCTGTAAACTCGGGAAA	GGGAGCTCCTGGGACTAGAG	MyTaq, 59-64
Chr14:57103317	TGGAGGGATTGTTGTTGGTT	CACCCCCTGCTTTGTTACAT	MyTaq, 59-64
Chr15:91475021	GACCTGACTTCCTGCTTTGC	TTTCAGGGGAACTGACCAAC	MyTaq, 59-64
Chr21:35742799	AAAGTGAATAAAATGTACGCAGTCA	TGTACTGGTGGTAGGGGTCA	MyTaq, 59-64
GRM7_7503321	ATTCCTCAGAAAGGGGAAAA	TGCACAAAACTCAATGCACA	MyTaq, 59-64
PRTG_55912391	TGTGGTTTTCAGTGGGGAGT	AAAGGCTTTGGTTCCCTGTT	MyTaq, 59-64
FAM151A	AATGGCTCTTCAGGCAACTG	AGCCCCTATTTCCTCATGCT	MyTaq, 59-64
NCAPG	TCAATGGAATCAGGCTATCAA	CTCAGGCGGTTTCATCTTTG	MyTaq, 59-64
NUP205	ACCCAAGAGGCAGAGGTTGT	CCACCTGACTCCCCAATTTA	MyTaq, 59-64
MGAM	TGGGAATTGTTCATTTTCAGC	CAAGGCTGGTTACCTGGTGT	MyTaq, 59-64
GTF2H1	AGTTCAGACCCCTCGTTTGA	ACCTCCCCAGCATTTAGTCC	MyTaq, 59-64

Table 3.6.: Primers used for sequencing validation in family 3 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
ADAR neu1	CCCCCAGTAGTTTCCTGCTT	GGAGTTGCTGTCTTCAGGTTC	5'PRIME 65
Chr7:25191295	AAGATCATTTGAGCCCAAGG	CCACCTTGGCTTCCCTAAG	MyTaq, 59-64
Chr20:3641868	CCTTTGGGCCACACTCTATG	GAGACCTGAAAGGCCAGGAG	GC I, 64
GRIN2A_9858173	CCAAAGAGGTTCCCTCATCA	TTCACTGGGATCCTGGTAGG	MyTaq, 59-64
ABCA4	GCACAGACCAGATGCAGAAG	AGAAGCAGGAAGGGTTTGGT	MyTaq, 59-64
GSTM5	AGCTGGGGACCTAAGAGACC	AAGGGAGCCTCAGGGAATAA	MyTaq, 59-64
PGLYRP3	TCTGGAAAACACATGCAAGC	CTTCTCTGCCTCTGGCAGTC	MyTaq, 59-64
TMEM79	GGATCTCCCACAGCCATAGA	CCAGGTCAAGTGAAGGGAAA	MyTaq, 59-64
ACTR1B	CTCAGCCAGCTTTGTGTTCA	CCTTTTTGGACACCCACATC	MyTaq, 59-64
AP2A2	TGATTGTTGGGTGACTCAGG	GGAGTCTCCAAGGCATGAAC	MyTaq, 59-64
SLC22A18	CAGGCCAGGACCATATAGGA	GATCTGCAGGTGTGCAGTGT	MyTaq, 59-64
MRC2	CAGAGGAGCCGCTGAACTAC	GCAGCCTGAGTAGAGGATGG	MyTaq, 59-64
ACAA2	TCCTTCATCCACCACAGTACC	TGCGATACCACCTTCTCCAT	MyTaq, 59-64

Table 3.7.: Primers used for sequencing validation in family 4 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr3:10346794	GGAGCTGACTCTTTGGCTGT	CACTTGTGGCTTCTGACCAA	MyTaq, 59-64
neu10:29818731	GAGGGGGTCTCACTTTGTTG	AGGAGGAAAGGGGAAATGAA	GC I, 64
SETD1A Chr16:30991240	ACTATGCCCTGGCCGTCA	AGACAGAGTGGGGCACAGAT	nested: GC II 65
and Neu16	CTGCTGTTGAAGACTCAGAGG	ATGCTTTGGAAGGCAGAAAA	and MyTaq 65
ITPR1_4702696	CCTGTTGCTCTGTCTCATGC	GGCAGAAGCCATGGAATAAA	MyTaq, 64
RADIL	CGTATCCAGAACGTGGGAGT	CATCCCTGACCCTCGAAATA	MyTaq, 59-64
KIAA1147	TCTACGTGAACGTGGCTGAC	CGACCAACACAATGAAGGTG	MyTaq, 59-64
CLCN1	GAGAGTGGGAGGTTCTGAGG	TCCCGGGTTCTAATTACAGG	MyTaq, 65
NECAB1	ATCCAGACGCCGTCAACTAC	TCTACGAAACCCAGCCTGTT	MyTaq, 59-64
PRSS36	GCTCCCAGCTCTCAGTCCTA	GGGATTCAGACCCGGATATT	MyTaq, 59-64
ABCA8	CAGGTAATGTGCCTCCCAAA	TAGATGGGACCTGCCACATT	MyTaq, 59-64

Table 3.8.: Primers used for sequencing validation in family 5 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr4:37445144	TCAGGCTTATGGCTGGCTAC	TCCCTGAAGGTCAGGTTCAC	MyTaq, TD61
Chr7:95001590	TGGAAACAGCTGACAGAGTGA	GAAGAGGAGGAGGCCACA	MyTaq, TD65
Chr7:121651348	TTTCCTAGCTCTACAGACATAACA	AATTCCAGCGTCTCTGAAGC	MyTaq, TD61
Chr9:121929598	TCTTTTGCTGGGCAATC	GCTCCACAAAAGGCTGAGAC	MyTaq, TD61

Table 3.9.: Primers used for sequencing validation in family 6 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr2:61522328	TGCTGCTGTTTGTATATGAATTTT	TCCAGGAATTTTTACTATTTTCCA	MyTaq, TD61
Chr3:184039075	GGAGCTTTTATGGGGAAAGG	TGGGGGTTGATTCTTCTACAG	MyTaq, TD61
Chr9:140352246	CAGGCCCAGATAGGAGTGAG	CCTGTCTATGCCCTGACACA	MyTaq, TD65
Chr9:140356442	CAGGTGTTGAACATGGCAGA	CCTGGCCCTGGAGTCTTT	MyTaq, TD61
neu10:16737156	ATTTCTGGGGCCAAGGTACT	TGTGAGCCACACCACACC	MyTaq, TD65
Chr11:20676319	TCTTCCTAACTGCCTCTCTTGG	GGACCAACCTGACCAGAAAC	MyTaq, TD61
Chr12:64521724	GGTAAGTAGAGCCTGGGAATCA	TGTAATTAAAAGCATCGTACCTG	MyTaq, TD61

Table 3.10.: Primers used for sequencing validation in family 7 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr8:110457620	CCCAACAGCTTGTGGATGTA	CCAGGGCATATGTAAAAAGCA	MyTaq, 59-64
Chr17:49371294	CACCACTGTGCAAGAAATTCA	CCCAAAGTAAGATAATGAAGATGAGAG	MyTaq, 59-64
C1orf185_51613245	TCTTTTTCTGTTTAAGCAGGTTGA	GTTAAGAGACTCAGAAGCTGATTGTA	MyTaq, 59-64
PRTG_55964778	GGAGTGTGGCCAAAATTAGAA	ATGGCTGCCTGATATTTTGC	MyTaq, 59-64
TLCD2_1611329	TCTTCAGGCCTGGGTTATTG	AAGGGGAAGGGATGTGAGTT	MyTaq, 61-64
CDC42BPA	TGAGCAAACAGAAGTAGGACTG	CCTGATTTGTTTGGCTTTTGA	MyTaq, 59-64
TARBP1	AAAGGTGTGCGCCATTACTC	AATCCTCCCAAAATCCAAGG	MyTaq, 59-64
FZD6	TGGTGACTCTGGCAAGTGAT	TGTCTGTCCCTCTCACCTCA	MyTaq, 59-64

Table 3.11.: Primers used for sequencing validation in family 8 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr6:170871046	CAGCCAGCCTAACCTGTTTT	GTGCCACTCCCTCCCTTAAT	MyTaq, 59-64
Chr11:111608216	CCATCAGTGACCAGGATGTG	TGAAGTAAGCCAGCAACAGG	MyTaq, 59-64
Chr18:60646248	TGGCTTCCGAGATTAGCAGT	ACATTGTGGCTCCTTTCCTG	MyTaq, 59-64
Chr19:3733883	GGGAGTGGCTGTTTTTAGCA	GAACCCAGGAGGCAGAGAAT	MyTaq, 59-64
Chr20:44594306	TATGCAGGGGTAGGGAAGTG	TGAGTCTTTGAGGGCTTTCC	MyTaq, 59-64
Chr21:26972181	TTTGTGATTTGGGGTTGACA	GGCCTCTTGAAAACCACTACC	MyTaq, 59-64
neu22:29446901	GACAGATGGGGTGAAATACGA	ACTGCTGTGGACATGACAGG	MyTaq, 64
SYTL2_85438948	TGTCATAGGGCACCTTTGGT	TGCAATATACCTGGGCCTTC	MyTaq, 59-64
METTL21B_58174275	TAAAAGCCAGTGTGGGCTCT	CCAGGGAAATCCATCCTTTT	MyTaq, 59-64
RAB3GAP1	ATCCTGGGACATGGAGAGAA	GCAAGCATGAGTGACAATGG	MyTaq, 59-64
DNAH5	ATGGCCCCAGGAAATAAAAC	CACAGAGCACTGCAAGTAGAGAA	MyTaq, 59-64
FAM159B	ATGGAAGCATGGAGGAAAAA	TGTTCTTCACTGAGATAGGACTGC	MyTaq, 59
MDGA1	ACACACCATTGGACCCATTT	TAAAGAAGGTGCTGGCAGGT	MyTaq, 59
ACO1	CACAGGCTGCTTCTCAATCA	CTCTTGGGCCTTAGCTTCCT	MyTaq, 59-64
DDX58	TTAGAGGGTGAGGCCTGTTT	ACCTGGATTTCTGAGCATGG	MyTaq, 59-64
MAMDC2	GCATGCCCTTTGGATAAAGA	GGCAACCACTGAAATCTGCT	MyTaq, 59-64
TECTA	ATGCCCAGGTTACTGCTTTG	CATCCCAAACCAAGGAAAAA	MyTaq, 59-64
DYNC1H1	CCCAGGCTCAAGTGATTCTT	TACAGCTGGGCTTCCTTGTT	MyTaq, 65
TLE2	CGGGAGGAGGATAGAGAAGG	TCTCTAAGCGCCTCTGGAAG	5' PRIME, 65
SEC14L3	CCAGGAGGTGGTGAAAGTGT	TCCTTCAGGGCCTGTATGTC	MyTaq, 65

Table 3.12.: Primer used for sequencing validation in family 9 and its corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
neu19:11319636	GTGAGTCCCCTCCTCACATC	GCCCAGCCTATTCTCCTTCT	MyTaq, 65

Table 3.13.: Primers used for sequencing validation in family 10 and their corresponding PCR (kind of recipe used, temperature).

Primer name	Forward primer (5'-3')	Reverse primer (3'-5')	PCR information
Chr1:208219387	CATGGAATGGAGACAAGCAA	TTAAGAAGCCTGGGTGCAAG	MyTaq, 59-64
neu19:12155456	CATCCCTTAATAGGCACCTCA	AACTGAAGGCTTTCACACATTTT	MyTaq, 59-64
Chr22:37420611	TGCTTCCCTTCTGACATCCT	CCAGGTTCTCCTTGATGTCC	MyTaq, 59-64
PTPRD_8486248	AAACGTGCAGGTGTCAAAAA	TTCTCTCTGCCAGGACAGGT	MyTaq, 59-64
EGLN1	GGACTGGAAGAAGCACAAGC	GTATCAGGTCGTCCATGCTG	GC I 65
FAM83H	CGGAGCTGCTGGAGAAGTA	GACTCCCCGGAGATGGTAAG	GC I 59-65
DOCK8	AAATGACCGCACCTCTGAAG	CCTTAGGGGCACCAAGAGTC	5' PRIME 65
PDCD1LG2	CTGCACAGGAGGGAATGTTT	CCGAGCAGATGCTTTCATTT	MyTaq,65
AGTPBP1	CTTGAAGAACGTGGGGAGTC	CGTTACACAACCCGTCAGG	GC I 63
COG7	GAGTGATGGTGTGCCAAGTG	CGAACTCCTAGGCTCAAACG	MyTaq,65
UBFD1	GCCAATCTCAGCCAGCTC	GGTGCCTAGTAGGGTGCAAG	GC I 59-65
SRCAP	CAGCCCAACAGTGGTTCTCT	AAACTCCTGGCTGAAGCAAA	MyTaq,65
ZBTB7C	GAGATCCAGTGCATCGTGAA	TAGTCGTTCTCCGCCTTGAT	MyTaq, 59-64
SMAD7	GCAGGCAAACGACTTTTCTC	TCACAGCAACACAGCCTCTT	GC II 59-65

3.2.2. Methods

3.2.2.1. Exome sequencing

Exome sequencing was performed by the Institute of Human Genetics at the Helmholtz Center in Munich. It was performed on a Genome Analyzer IIx system (Illumina) after in-solution enrichment of exonic sequences yielding on average 6.2Gb of sequence per individual. The average read depth was 78 with between 84.5% and 85.6% of the target regions covered at least 20x. The alignment and variant calling of the ten CAD families was done using our pipeline (see 3.1.3) and the data of the 255 individuals has been aligned and called by the Helmholtz Center.

3.2.2.2. Polymerase chain reaction

To amplify the fragments, PCR was used. PCR comprises three main steps: denaturation, annealing and extension [116, 117]. In the denaturation step, the DNA gets heated to 94-98°C. This leads to denaturation of the double stranded DNA to enable primers to bind the strands. The binding of the primer to the DNA takes place in the annealing step. Here, the temperature decreases to an optimal primer annealing temperature (usually between 60-65°C depending on the specific primer pair). Temperatures that are too low will produce unspecific binding and therefore unspecific products. High temperatures increase the risk of no binding, as the thermal energy increases the movement of

the primer. In the third step the system is heated to reach optimal temperature for the Taq-polymerase to synthesize and elongate the new DNA strands .

Gradient-Touchdown-PCR

The gradient PCR uses different annealing temperatures depending on the tube position in the cycler. It was used to identify the optimal annealing temperature for the synthesized primers in the cycle (tab.3.14 step 3) [117]. All identified temperatures can be found in tab.3.4-3.13 . Touchdown PCR [55] is used to increase specificity of the amplificates. Multiple cycles (here four) are used, each reducing the annealing temperature by 2°C. A high temperature in the first annealing step ensures the specificity of the first products. Every further circle can be performed at a lower temperature to further amplify the specific fragments (tab.3.14 step 3,6,9 and 12).

Table 3.14.: Course of the PCR program. The temperature in step 3,6, 9 and 11 is depending on the tube position (gradient PCR) or the chosen annealing temperature (in this work either 59°C, 63°C or 65°C).

Step	Temperature [°C]	Time	Runs	Repeat
1	94	3 min	1	
2	94	30s	3	
3	59-65.5	30s	3	
4	72	45s	3	step 2
5	94	30s	3	
6	57-63.5	30s	3	
7	72	45s	3	step 5
8	94	30s	3	
9	55-61.5	30s	3	
10	72	45s	3	step 8
11	94	30s	35	
12	53-59.5	30s	35	
13	72	45s	35	step 11
14	72	10min	1	
15	16		1	

Three different buffer mixes were used for the samples depending on their primer (see tab.3.4-3.13). Table 3.15 shows the recipes for all samples.

Table 3.15.: Recipes for MyTaq, 5'Prime, GC I and GC II DNA sample preparation [20μl].

MyTaq recipe [20μl]	
4μl	Taq-buffer mix
0.1μl	Taq-Polymerase
1μl	per primer
4μl	DNA (50 ng/μl)
9.9μl	Water
5'Prime recipe [20μl]	
8μl	2.5x buffer mix
1μl	per primer
4μl	DNA (50 ng/μl)
6μl	Water
GC I & II recipe [20μl]	
10μl	2x buffer mix (GC I or GC II)
2μl	deoxynucleotide triphosphates
1μl	per primer
0.8μl	DNA (50 ng/μl)
0.13μl	Taq-Polymerase
5.07μl	Water

3.2.2.3. Agarose gel electrophoresis

Agarose gel electrophoresis was used to check the quality of the products. The DNA wanders through pores built up by the agarose due to an electric field [118]. This separates the various length of the substrates as longer molecules move more slowly through the pores compared to short molecules. If the PCR was successful, we will see only one band. Unspecific binding and therefore unspecific products would lead to several bands. Here, 1% gels were used (see tab.3.16). Each lane contains 8μl 100bp DNA ladder and 10μl each of the PCR product. 4μl SYBR Green were added to each sample and DNA ladder. The electrophoresis ran at 400mA/100V for one hour.

3.2.2.4. Sample preparation for Sanger sequencing

Sanger sequencing is a DNA sequencing method, described in depth in chapter 2.2.3.1. The Sanger sequencing was done by Seqlab- Sequence Laboratories in Göttingen and Eurofins Genomics in Ebersberg.

To prepare the samples for sequencing, the PCR products were purified using Nucle-oSpin® Gel and PCR Clean-up. We added 200μl of the NTI binding buffer to the

Table 3.16.: Recipe for 10x TBS-buffer and 1% agarose gel.

10x TBS-buffer	
121.1g	Tris
51.3g	Boric acid
3.7g	EDTA
fill up to 1l with double distilled water	
1% agarose gel	
3.5g	agarose
150 ml	1x TBS-buffer

sample, mixed thoroughly and put on top of the purification column. The binding buffer maintains the pH of 5-6 that is needed for the PCR product to bind to the silica membrane in the column. The column with the mixture was then centrifuged for three minutes at 10000rpm and all flow through was discarded. Afterwards, the sample was washed twice using 600μl ethanolic Buffer NT3 and centrifuged for three minutes each at 10000 rpm again discarding the flow through, which removes contamination. To make sure all washing buffer and contamination have been removed, the column was centrifuged again for three minutes at 13000 rpm. 17μl Elution Buffer NE was added directly on the membrane and centrifuged at 13000rpm for four minutes to collect the then purified DNA sample. We determined the sample concentration via photometer.

Each sample sent to Eurofins Genomics contained between 90ng and 100ng of DNA in 15μl Elution buffer and 1.5μl of the corresponding Primer with a concentration of 10pmol.

4. Results

4.1. Screening for known rare causal variants

As described in section 2.2.2, CAD has many risk factors influencing the development of the disease. Since some of these are treatable and a reduction of risk factors ultimately reduces your risk of CAD and MI, it is important to thoroughly investigate the underlying cause of disease. For example, we previously showed that roughly 5 percent of our cohort of myocardial infarction patients suffered from FH, a disease that leads to elevated LDL cholesterol levels and subsequently promotes CAD [3]. Timely use of lipid-lowering medication is shown to lower the risk of CAD and our results underline the need for systematic screening. Along the lines of the previous study, we evaluated the potential prevalence of known CAD risk factors. This was done by identifying rare disease-causing variants in the CAD cohort. To validate potential findings, we compared the expected phenotype given the variants with the phenotype seen in the cases. Phenotypes included were high levels of total cholesterol, LDL cholesterol, triglycerides, hypertension, diabetes type 2 and obesity.

We used HGMD® [91] and ClinVar [90] to identify disease-causing variants for the selected traits. This was done by identifying all phenotypic search words that were associated with our phenotypes of interest and flagged as known damaging mutations. We excluded most syndromes due to their complex manifestation and concentrated on diseases that primarily induce at least one of our phenotypes of interest. We then identified all variants reported to cause at least one of the identified traits. Additionally we only included heterozygous variants and therefore excluded variants known to cause a disease with a recessive mode of inheritance. If the mode of inheritance was unknown, the variant was included to avoid false negative results. Subsequently, we investigated the prevalence of said variants in our cohort and compared these to the frequency in the control databases 1000G [84] and ExAC [92].

38

4.1.1. Frequency distribution of rare variants in our cohort versus controls

Using the HGMD® database, we identified 11 unique phenotypes reported to cause to our chosen phenotypes (table 4.1). We found 19 variants flagged as disease-causing for these phenotypes in our cohort of 255 myocardial infarction patients. It is important to note that some of these phenotypes and therefore variants are overlapping. For example, a variant causing a metabolic syndrome is not only associated with obesity but also hypertension and high triglyceride levels. This was the only syndrome that we included in our results due to its high relevance in regards to the investigated traits [119]. Another risk factor that overlaps is hyperlipidemia which causes elevation of at least one of the lipids or lipoproteins in the blood and can therefore lead to high cholesterol, LDL cholesterol and/or triglycerides. In this case, the variants associated with hyperlipidemia have been analyzed individually to assign them to the correct phenotype. For the calculation of the prevalence, only unique variants were counted. Seven variants were correlated to obesity, 10 to hypercholesterolemia, two to hypertension and two to high triglyceride levels. We could not identify any variants causing high cholesterol levels or diabetes according to HGMD® in our data.

We found at least one variant in 28 of our 255 myocardial infarction patients, resulting in a prevalence of 10.98 % of risk factor variants in our data. Three of the 28 individuals also carried a second variant.

Table 4.1.: Phenotypic search words for identifying causal flagged risk factor variants in HGMD®. Here, cholesterol, LDL cholesterol and triglycerides correspond to high levels of these factors.

Risk factor	Database disease labels
Obesity	"Metabolic syndrome" [120] "Obesity", "Obesity autosomal dominant"
Cholesterol	"Apolipoprotein E deficiency" [121] "Increased total cholesterol" "Hyperlipidaemia" [122]
Hypertension	"Metabolic syndrome" [120] "Pulmonary hypertension"
LDL cholesterol	"Apolipoprotein B deficiency" [123] "Hypercholesterolaemia" "Hyperlipidaemia" [122]
Triglycerides	"Apolipoprotein E deficiency" [121] "Hypertriglyceridaemia" "Lipoprotein lipase deficiency" [124] "Metabolic syndrome" [120] "Hyperlipidaemia" [122]

We repeated the analysis for the ClinVar database, however, we also included variants known to cause CAD and/or MI. The analysis resulted in 10 unique phenotypes (table 4.2) and 13 respective variants. One for obesity, two for high total cholesterol, one for diabetes, five for hypercholesterolemia, one for high triglycerides and four for CAD. In total 7.84% of our patients were carrying a ClinVar risk factor variant.

Table 4.2.: Phenotypic search words for identifying causal flagged risk factor variants in ClinVar. Here, cholesterol, LDL cholesterol and triglycerides correspond to high levels of these factors.

Risk factor	Database disease labels
Obesity	"Obesity"
Cholesterol	"Familial_type_3_hyperlipoproteinemia" [125]
Diabetes	"Diabetes_mellitus_type_2|Pineal_hyperplasia_AND _diabetes_mellitus_syndrome"
LDL cholesterol	"Familial_hypercholesterolemia" "Hypercholesterolaemia|Familial_hypercholesterolemia" "Hypercholesterolaemia|not_provided|Familial_hypercholesterolemia"
Triglycerides	"Familial_type_3_hyperlipoproteinemia" [125] "Hyperlipoproteinemia_x2c_type_I" [126]
CAD	"Coronary_artery_disease/myocardial_infarction" "Coronary_artery_disease_x2c_autosomal_dominant_2" "Myocardial_infarction_1"

Interestingly, there was little consistency between HGMD® and ClinVar in our analysis. We only found two variants that overlapped, a variant for obesity in the *MC4R* gene and a variant for high LDL cholesterol in the *LDLR* gene. In total we found less ClinVar variants than HGMD® variants in our data (13 to 19, see tab.4.3). When combining variants from HGMD® and ClinVar, 17.25% of our patients were carriers of at least one risk factor variant.

Table 4.3.: All variants from HGMD® and ClinVar identified in our cohort of 255 MI patients.

Gene	Position and change	HGVS	Reference sequence	Database source	Reported Phenotype
ABCA3	chr16:g.2350115G>T	c.1502C>A(p.(A501E))	NM_001089	HGMD®	HT
APOA4	chr11:g.116692293C>A	c.481G>T(p.(A161S))	NM_000482	HGMD®	HC
APOB	chr2:g.21229160C>T	c.10580G>A(p.(R3527Q))	NM_000384	HGMD®	HC
CCT7	chr2:g.73479931C>T	c.962C>T(p.(S321L))	NM_001009570	ClinVar	MI
GCKR	chr2:g.27730170_27730171insA	c.1135dupA(p.(T379fs))	NM_001486	HGMD®	Tri
INSR	chr19:g.7125518C>T	c.2998G>A(p.(V1000M))	NM_001079817	ClinVar	DB
LDLR	chr19:g.11210962G>A	c.131G>A(p.(W44X))	NM_000527	HGMD®	HC
LDLR	chr19:g.11213464T>C	c.313+2T>C	NM_001195798	HGMD®	HC
LDLR	chr19:g.11217303C>T	c.634C>T(p.(R212W))	NM_001195799	HGMD®	HC
LDLR	chr19:g.11217344T>A	c.675T>A(p.(D225E))	NM_001195799	HGMD®	HC
LDLR	chr19:g.11218078C>A	c.324C>A(p.(C108X))	NM_001195800	HGMD®	HC
LDLR	chr19:g.11224052G>A	c.781G>A(p.V261M))	NM_001195800	HGMD®	HC
LDLR	chr19:g.11224296G>A	c.940G>A(p.(D314N))	NM_001195800	HGMD®	HC
LDLR	chr19:g.11227604G>A	c.1271G>A(p.(G424E))	NM_001195800	HGMD®	HC
LDLR	chr19:g.11217357G>A	c.688G>A(p.(V230I))	NM_001195799	ClinVar	HC
LDLR	chr19:g.11227604G>A	c.1271G>A(p.(G424E))	NM_001195800	ClinVar	HC
LIPC	chr15:g.58855748C>T	c.1214C>T(p.(T405M))	NM_000236	ClinVar	Chol
LPL	chr8:g.19811733G>A	c.644G>A(p.(G215E))	NM_000237	ClinVar	Chol,Tri
LRP6	chr12:g.12334052T>C	c.1298A>G(p.(N433S))	NM_002336	HGMD®,ClinVar	Obesity,HT,Tri,MI
MC4R	chr18:g.58038877G>A	c.706C>T(p.(R236C))	NM_005912	HGMD®	Obesity
MC4R	chr18:g.58039203G>A	c.380C>T(p.(S127L))	NM_005912	HGMD®,ClinVar	Obesity
MCHR1	chr22:g.41077613G>A	c.950G>A(p.(R317Q))	NM_005297	HGMD®	Obesity
MEF2A	chr15:g.100230557A>G	c.578A>G(p.(N193S))	NM_001130928	ClinVar	MI
MEF2A	chr15:g.100230605C>T	c.626C>T(p.(P209L))	NM_001130928	ClinVar	MI
MIR6886;LDLR	chr19:g.11224210G>A	c.1359-1G>A	NM_001195798	ClinVar	HC
PCSK1	chr5:g.95761545C>T	c.375G>A(p.(M125I))	NM_000439	HGMD®	Obesity
PCSK9	chr1:g.55518037G>A	c.610G>A(p.(D204N))	NM_174936	ClinVar	HC
POMC	chr2:g.25384048G>C	c.706C>G(p.(R236G))	NM_000939	HGMD®	Obesity
POMC	chr2:g.25384360G>C	c.394C>G(p.(P132A))	NM_000939	HGMD®	Obesity
STAP1	chr4:g.68436820A>G	c.139A>G(p.(T47A))	NM_012108	ClinVar	HC

To see if the frequency of CAD risk variants was higher our CAD database than in controls, we analyzed ExAC and the 1000 Genome databases using the phenotypes identified from ClinVar. This was done using the same search strategy so the number of variants identified could be compared between these control databases and our cohort. We found three risk variants in 1000G and 86 in ExAC.

Combining the results from HGMD® and ClinVar, we found a higher rate of variants in our cohort compared to the control database for all phenotypes except diabetes (tab.4.4). Comparing HGMD® variants in our 255 patients to the control databases, we found a higher prevalence of variants for obesity, hypercholesterolemia, hypertension and high triglyceride level in our cohort. However, we did not identify any HGMD® variants related to high cholesterol level and diabetes in our cohort. There was a slightly higher frequency of obesity and triglyceride-increasing variants in ExAC compared to those found in ClinVar. There was also a higher frequency for diabetes and hypertension, as we could not identify any hypertension-related ClinVar variants in our data. However, we found a higher frequency of cholesterol-increasing and MI ClinVar variants in our cohort compared to the controls.

Table 4.4.: Percentage distribution of rare pathogenic variants in our cohort of 255 myocardial infarction patients taken from HGMD® (HGMDin255) and ClinVar (ClinVarin255) compared to the ClinVar variants found in the control databases ExAC and the European 1000G frequencies.

Phenotype	HGMDin255 [%]	ClinVarin255 [%]	ExAC [%]	1000G [%]
Obesity	0.0275	0.0039	0.0058	0
Cholesterol	0	0.0235	0	0
Diabetes	0	0.0118	0.0180	0.0050
Hypertension	0.0824	0	0.0046	0
Hypercholeterolemia	0.0118	0.0196	0.0028	0
Triglycerides	0.0078	0.0039	0.0042	0
MI	n.a.	0.0196	0.0030	0.0050

Looking more closely at the specific variants we have identified in our cohort, we could see that their frequency in our dataset is increased compared to the controls (tab.4.5,4.6). The only exception being a variant for obesity (Chr22:41077613G>A), which had a slightly higher frequency in 1000G.

Table 4.5.: Frequency comparison of rare pathogenic variants in our cohort of 255 myocardial infarction patients taken from ClinVar (ClinVarin255) compared to the control databases ExAC and the European 1000G frequencies (1000G_Eur).

Phenotype	Variant	ClinVarIn255	ExAC	1000G_Eur
Obesity	Chr18:58039203G>A	0.00392	0.00016	0
Cholesterol	Chr15:58855748C>T	0.01961	0.00276	0.00500
	Chr8:19811733G>A	0.00392	0.00013	0.00100
Diabetes	Chr19:7125518C>T	0.01176	0.00900	0.00500
,	Chr1:55518037G>A	0.00392	0	0
	Chr19:11217357G>A	0.00392	0.00002	0
Hypercholesterolemia	Chr19:11227604G>A	0.00392	0.00005	0
	Chr19:11224210G>A	0.00392	0	0
	Chr4:68436820A>G	0.00392	0	0
Triglycerides	Chr8:19811733G>A	0.00392	0.00013	0.00100
	Chr15:100230557A>G	0.00784	0.00095	0.00200
MI	Chr15:100230605C>T	0.00392	0.00078	0.00300
	Chr12:12334052T>C	0.00392	0.00002	0
	Chr2:73479931C>T	0.00392	0.00002	0

Table 4.6.: Frequency comparison of rare pathogenic variants in our cohort of 255 myocardial infarction patients taken from ClinVar (ClinVarin255) compared to the control databases ExAC and the European 1000G frequencies (1000G_Eur). .

Phenotype	Variant	HGMDIn255	ExAC	1000G_Eur
Obesity	Chr12:12334052T>C	0.00392	0.00002	0
	Chr18:58038877G>A	0.00392	0.00003	0
	Chr18:58039203G>A	0.00392	0.00016	0
	Chr2:25384048G>C	0.00392	0.00231	0.00300
	Chr2:25384360G>C	0.00392	0.00132	0.00100
	Chr22:41077613G>A	0.00392	0.00383	0.00400
	Chr5:95761545C>T	0.00392	0.00017	0.00100
Hypertension	Chr12:12334052T>C	0.00392	0.00002	0
	Chr16:2350115G>T	0.00784	0.00385	0.00100
Hypercholesterolemia	Chr11:116692293C>A	0.03922	0.00739	0.00990
	Chr19:11210962G>A	0.00392	0	0
	Chr19:11213464T>C	0.00392	0	0
	Chr19:11217303C>T	0.00392	0.00018	0
	Chr19:11217344T>A	0.00392	0.00002	0
	Chr19:11218078C>A	0.00392	0	0
	Chr19:11224052G>A	0.00392	0.00001	0
	Chr19:11224296G>A	0.00392	0.00004	0
	Chr19:11227604G>A	0.00392	0.00005	0
	Chr2:21229160C>T	0.01176	0.00023	0
Triglycerides	Chr12:12334052T>C	0.00392	0.00002	0
	Chr2:27730170dupA	0.00392	0	0

4.1.2. Observed and predicted phenotypes

We aimed to compare the observed and expected phenotype to estimate how well the variants predict the outcome. In addition, we aimed to find overlaps between risk factors.

Predicted HGMD® phenotypes

First, we validated variants flagged as disease causing in HGMD® with the observed phenotypes in our cohort. We then analyzed if the variant carriers were showing any additional risk factor phenotypes.

In regards to obesity, only one out of seven individuals that carry a variant connected to obesity had a high BMI of over 30 (32.1.) (see fig.4.1). This patient also suffered from

diabetes type 2, high triglyceride levels, high total cholesterol levels and hypertension. The other six patients showed BMIs between 23.1 and 29.3, with some representing as overweight but not obese. In total, 43% of individuals (three) carrying the obesity-variant had high total cholesterol values, 57% (four) suffered from hypertension, and one showed elevated LDL cholesterol, elevated triglyceride level, and/or has diabetes.

Roughly three-quarters of the individuals (15) carrying a variant for hypercholesterolemia had LDL cholesterol levels over 190mg/dL (191 - 299md/dL). We had no data for two patients, with four others having LDL cholesterol levels between 121 and 177mg/dL. This range varied between near-optimal and high, as only values under 100mg/dL are considered optimal. Sixty-two percent of the 21 variant carriers (13) also had high total cholesterol values. A third (seven) suffered from hypertension and high triglycerides. Only 9% (two) were obese and had diabetes.

Two of the three individuals (75%) with a variant for hypertension showed the expected phenotype with a blood pressure of 150/80mmHg and 140/90mmHg, which is classified as a mild hypertension. The third individual was very close with 136/83mmHg. Two of three (67%) were also positive for obesity and high triglycerides. One had diabetes and high total cholesterol level.

We identified two individuals with a variant reported to cause triglyceride level and both individuals had levels over 200md/dL (228 and 285 mg/dL). While they did not reach the threshold of >500mg/dL, these levels are still considered clinically high. One of two was obese, had diabetes, and/or high total and LDL cholesterol, and both had hypertension.
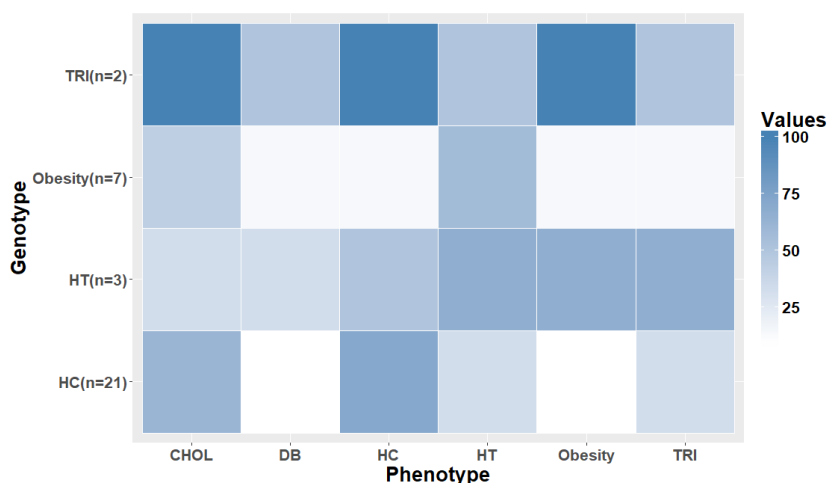
Figure 4.1.: Comparison of expected and observed phenotype based on HGMD® variants. The shading corresponds to the overlap in percent. Chol stand for high total cholesterol levels, DB for diabetes type 2, HT for hypertension, HC for hypercholesterolemia and TRI for high triglyceride level.

Predicted ClinVar phenotypes: We repeated the same analysis with the ClinVar variants and found only one individual carrying a variant for obesity. We could not confirm this phenotype as the patient's BMI was 23. The patient did however have hypertension and high total cholesterol levels (see fig.4.2).

Of the six individuals carrying variants that increase the total cholesterol levels, two (33%) had high levels, with both of them exceeding 240mg/dL. The other four all showed values over 190mg/dL (193-196mg/dL) which is very close to the set threshold. Additionally they had already been treated for hypercholesterolemia with statin therapy. Three (50%) were obese, two (33%) had hypertension and high triglyceride level and one (17%) had diabetes and hypercholesterolemia.

All three patients with a diabetes variant were negative for diabetes, but all of them had high total cholesterol and LDL cholesterol values. One had hypertension and high triglycerides.

Five individuals carried a variant for hypercholesterolemia and all of them were positive for the disease. One of five (20%) was diabetic, three (60%) of them had hypertension, high levels of triglycerides and high levels of total cholesterol.

There was only one individual identified with a variant for high triglyceride levels, however said patient has a very low level of 53mg/dL. He showed no other risk phenotypes except for a high total cholesterol level.

The five individuals with a variant causing CAD and/or myocardial infarction also suffered from CAD-promoting risk factors. One had high triglycerides, two (40%) were obese, had diabetes and/or hypercholesterolemia. Sixty percent (three) were positive for hypertension and 80% (four) had high total cholesterol level.
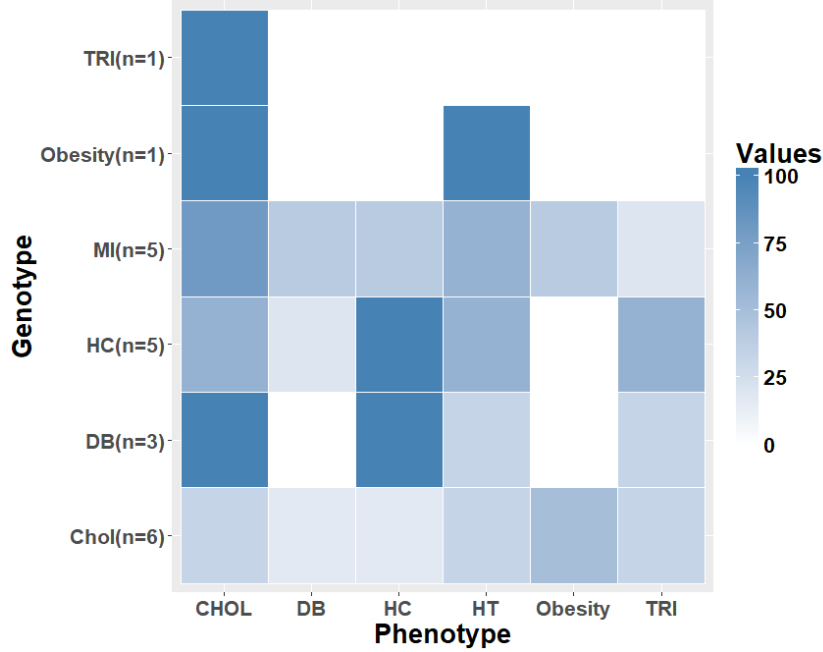
Figure 4.2.: Comparison of expected and observed phenotype based on ClinVar variants. The shading corresponds to the overlap in percent. CHOL stands for high total cholesterol levels, DB for diabetes type 2, HT for hypertension, HC for hypercholesterolemia and TRI for high triglyceride level.

Phenotype overlap in our data

To investigate if the phenotypes correlate with each other, we gathered all available phenotypic data for the complete cohort of 255 myocardial infarction patients and identified all individuals suffering from one of our investigated phenotypes. We then picked one phenotype and looked at the occurrence of other risk factors in these specific individuals as shown in figure 4.3.

Considering the whole cohort 23.5% of our 255 patients are obese, 43.5% have hypertension and 12.5% are diabetic. Hypercholesterolemia was present in 40.8% of the 255 patients, with 35.3% suffering from high cholesterol, and 41.2% from high triglyceride levels.

4. Results

Looking at the obese patients only (60 of 255), 62% had high triglyceride levels, 55% suffered from hypertension, 40% had high cholesterol levels. Thirty-eight percent showed hypercholesterolemia and 22% were diabetic.

Half of the patients with hypertension (111 of 255) had high triglyceride levels, 44% had increased levels of LDL cholesterol, 40% had high total cholesterol levels, 29% were obese and 14% were diabetic.

Of the diabetic individuals (32 of 255), 63% had high triglyceride levels, 47% had hypertension and 41% were obese. There are 44% with hypercholesterolemia and 40% with high total cholesterol values.

The biggest proportion of the individuals with high total cholesterol (90 of 255) had hypercholesterolemia (72%). Over half of them also had high triglyceride levels. Forty-nine percent had hypertension, 27% were obese and 13% diabetic.

Around half of the individuals with high triglyceride levels (105 of 255) had hypertension, hypercholesterolemia and high total cholesterol levels. A third of these were obese and 19% diabetic.

Of the patients with hypercholesterolemia (104 of 255), 63% also had high total cholesterol levels and 51% had high triglyceride levels. Nearly half of them also suffered from hypertension, 22% were obese and 13% diabetic.
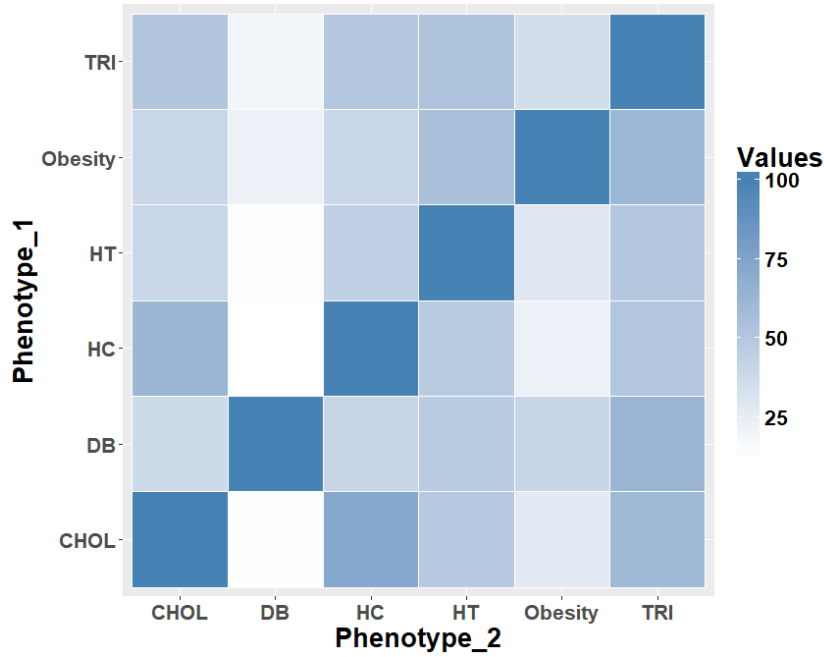
Figure 4.3.: This figure shows the percentage of individuals with a specific risk factor (y-axis) also suffering from other risk factors (y-axis). HT stands for hypertension, CHOL for high total cholesterol levels, DB for diabetes, HC for hypercholesterolemia and TRI for high triglyceride levels.

4.2. Family studies to identify rare causal variants

4.2.1. Prologue - Family matters

Looking at families, heredity can barely be more obvious. Simply put, tall parents produce tall offspring and likewise the size and shape of the nose is passed on to the next generations. Logically, not only visible traits are inherited, the risk of a disease is too. We see families severely affected with coronary artery disease or diabetes. But, brown eyed parents can get blue eyed offspring and likewise, the risk of a disease can change. Hence, in a family, we can identify the genetic variants responsible for a trait if we look for variants that are inherited along with the trait. Family studies are the most powerful and cost-efficient approach when it comes to rare causal variants, as they cluster in families. In addition, a unfavorable combination of variants or a burden or risk variants can be identified. Hence, alongside case control studies, family studies are an

important brick in the puzzle of heredity. Family matters. Here we analyze ten families with prevalent MI to investigate the cause of disease.

4.2.2. General results

After selecting the ten most promising families out of 401, we checked the quality of the available DNA to make sure it was suitable for exome sequencing. The sequencing required the DNA to be intact and in a sufficient concentration. The best three individuals in regards to their DNA quality as well as their position in the pedigree were chosen and exome sequenced. After we received the sequencing files, we set up a pipeline to process the data and call the variants (see material and methods 3.1.3).

Before filtering, we found 321,272 variants (see fig.4.4). Variant prioritization is very important when trying to identify causal variants, hence we applied several filters. Because we expected CAD to be caused by a rare variant with strong effect in these families, we filtered for a frequency under 5% in the general population. Variants in segmental duplications are often false positive, as reads in these regions are difficult to align. Filtering synonymous and intronic variants left us with variants with a probable altered protein function.

After filtering for exonic and splicing variants, we had 73,416 variants left. Filtering for rare variants reduced the amount further to 22,745. We then removed variants in segmental duplications, which decreased it to 18,224. Lastly, we only included nonsynonymous variants which decreased the amount to 11,714.
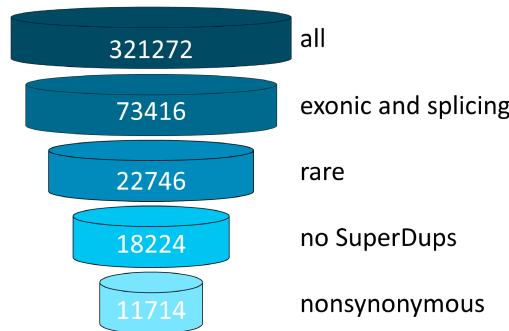
Figure 4.4.: This figure shows the amount of variants after the different filter steps.

We then wanted to see if there were any enriched pathways when looking at the filtered gene list of all our families, and compared these to the enrichment in known CAD genes

from previous GWAS studies. The list of CAD genes as well as a table of all significant pathways, their p and q values, and the overlapping genes can be found in the supplement (see list A.1, tab.A.2, tab.A.3).

With a threshold of q<0.05, we found eight enriched pathways for the CAD genes. All of them were related to lipid levels, namely lipoprotein metabolism, statin pathway, lipid digestion, mobilization, and transport, binding and uptake of ligands by scavenger receptors, chylomicron-mediated lipid transport, LDL-mediated lipid transport and the composition of lipid particles (see fig.4.5). Some genes included in these pathways were *LDLR*, *PCSK9* and *APOE*.
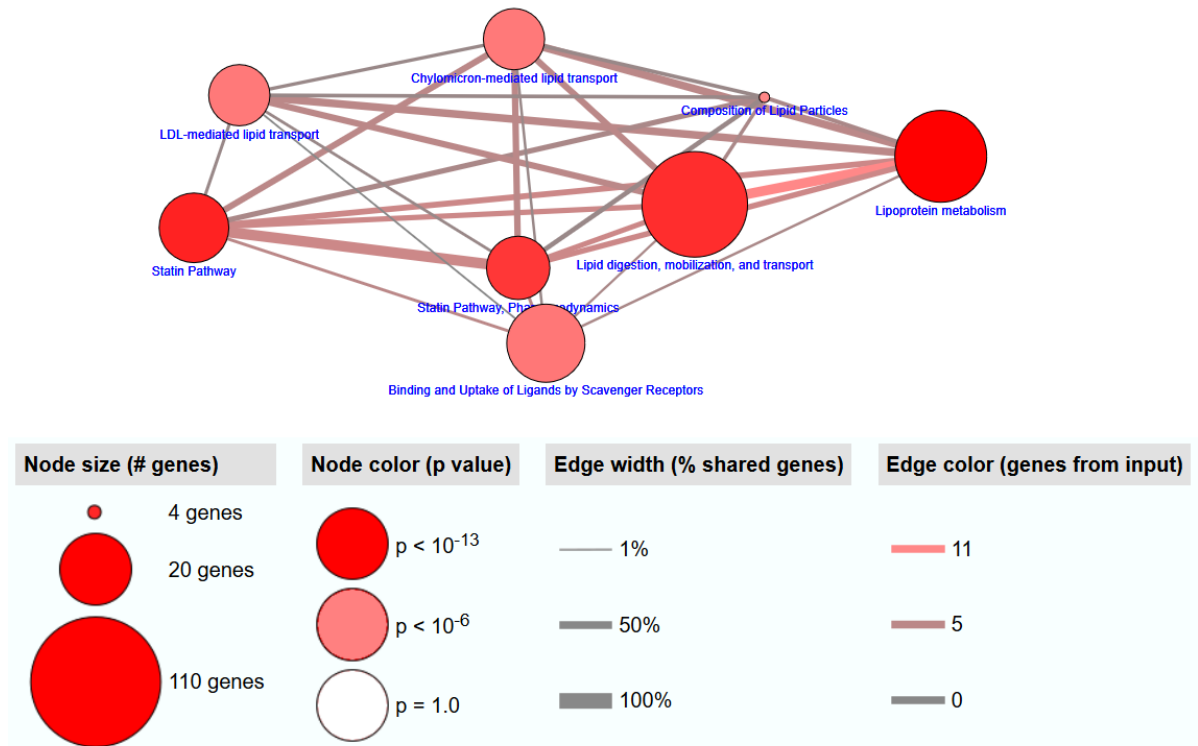
Figure 4.5.: Pathway enrichment in the CAD gene list.

We found no overlap between the enriched pathways from the CAD gene list and our gene list retrieved from the filtered exome-data of our families. The significant pathways in our families included extracellular matrix (ECM) receptor interaction, ECM organization, collagen chain trimerisation, collagen formation, collagen biosynthesis and modifying enzymes, beta1 integrin cell surface interactions, assembly of collagen fibrils and other

multimeric structures, laminin interactions, integrin and ABC transporters (see fig.4.6). Most of these pathways play an important role in the ECM. The genes partaking in these pathways were for example *MMP10* and *DOCK8*.

+
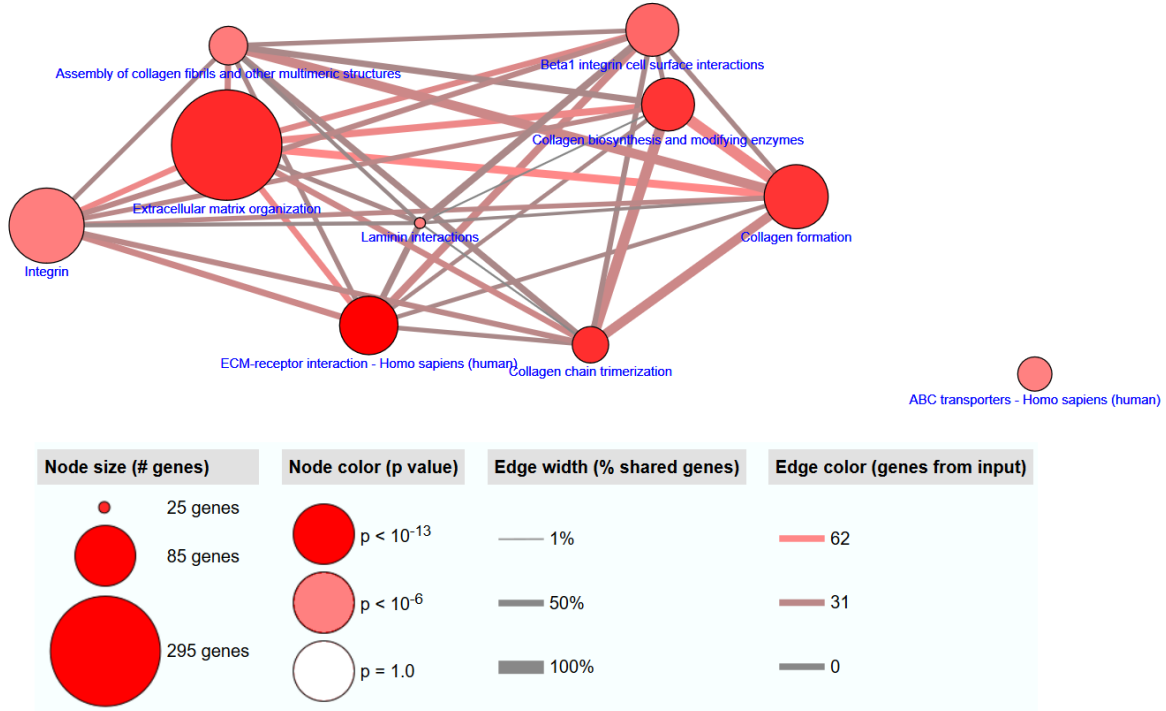
Figure 4.6.: Pathway enrichment in the genes retrieved from the filtered exome-sequencing data.

All variants were filtered for those that were shared by all three affected family members. The most promising variants according to prediction scores were validated in other available family members via Sanger Sequencing. After evaluating the results of the first validation, we decided to sequence more variants/all variants and included all family members available for sequencing. This was the case, when the results remained inconclusive or the originally chosen variants did not co-segregate with disease in the family. There are two instances we had to consider and be aware of when choosing the individuals for analysis. For once, some of the healthy family members were fairly young at the point of data collection and CAD is a late-onset disease. Hence, including them as healthy might be incorrect. Second, if the family member developed MI very late in life, it would be possible that his/her disease is unrelated to the genetic CAD occurring

in the family and might be a case of age-related/spontaneous CAD. To account for these cases we tried to correct for age and calculated all statistics twice, once for all family members and once for those that were within the defined age cut-off. The age cut-off was calculated as the mean age of disease in the family plus one standard deviation (see material and methods on page 28).

4.2.3. Family 1 - 8409

In family 1 we had data available for 24 individuals stretching over three generations (see fig.4.7). After applying our first filters, 486 variants were found in family 8409 alone. We only considered variants that were shared by all three affected and exome sequenced individuals, which left us with nine variants (tab.4.7), eight of which were nonsynonymous variants and one stop-gain.

We then validated the five most promising variants in eight family members (tab.4.8). Promising was defined as three or more prediction scores (SIFT, PolyPhen2, DANN & CADD) showing a possible deleteriousness. The variants that fulfilled these conditions were chr3:44684302C>A in *ZNF197*, chr19:5705841C>T in *LONP1* and chr11:102650461A>T in *MMP10*. The chr3:51908114G>A in *IQCF5* has been included as it is predicted to lead to a stop-gain and might therefore have a strong effect. All variants but the *MMP10* variant were found in three of four affected individuals, and in three of four healthy individuals (see tab.4.8).

Table 4.7.: Variants shared by all three exome sequenced individuals in family 1. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	phastCons	SIFT	Polyphen2	DANN	CADD
chr2	5833511	G	A	exonic	*SOX11*	nonsynonymous SNV	y	T	P	0.999	27.4
chr3	44684302	C	A	exonic	*ZNF197*	nonsynonymous SNV	y	D	D	0.991	15.16
chr3	49679930	C	T	exonic	*BSN*	nonsynonymous SNV	y	T	B	0.767	9.47
chr3	50129547	C	T	exonic	*RBM5*	nonsynonymous SNV	y	0	B	0.997	17.3
chr3	51908114	G	A	exonic	*IQCF5*	stop-gain	y	T	0	0.997	18.01
chr3	142274770	T	C	exonic	*ATR*	nonsynonymous SNV	y	T	B	0.997	14.35
chr11	94533087	C	T	exonic	*AMOTL1*	nonsynonymous SNV	y	0	B	0.982	13.47
chr11	102650461	A	T	exonic	*MMP10*	nonsynonymous SNV	y	D	D	0.984	18.19
chr19	5705841	C	T	exonic	*LONP1*	nonsynonymous SNV	y	D	D	0.998	18.16
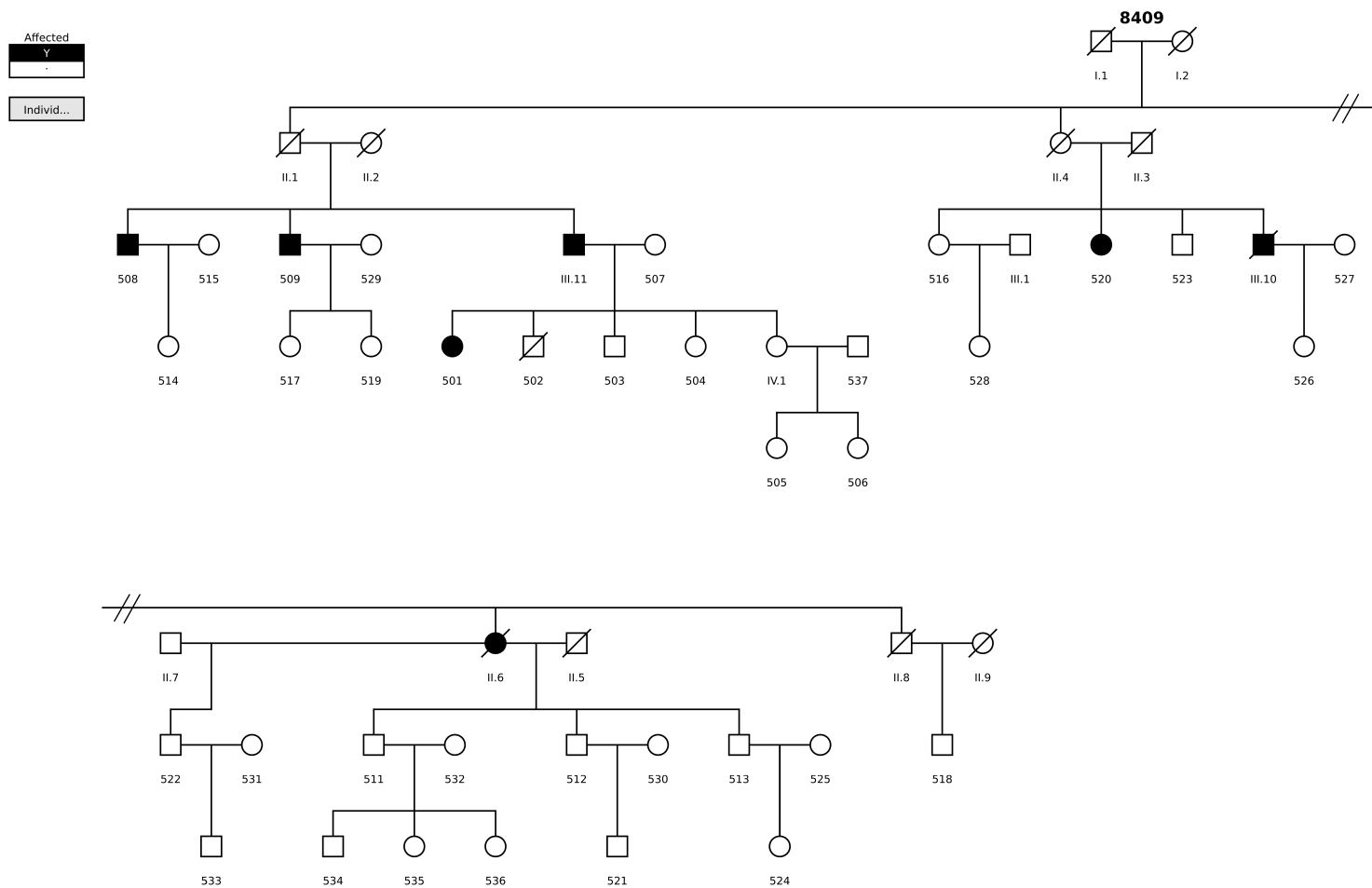
Figure 4.7.: This figure shows the pedigree of family 8409. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.8.: Results of the first validation round. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	*ZNF197*	*IQCF5*	*LONP1*	*MMP10*
8409501*	44	57	2	2	2	2
8409508*	56	71	2	2	2	2
8409509	56	68	1	1	1	2
8409520*	54	66	2	2	2	2
8409511	-	75	1	1	2	1
8409512	-	68	2	2	2	1
8409516	-	70	2	2	1	1
8409523	-	67	2	2	2	1
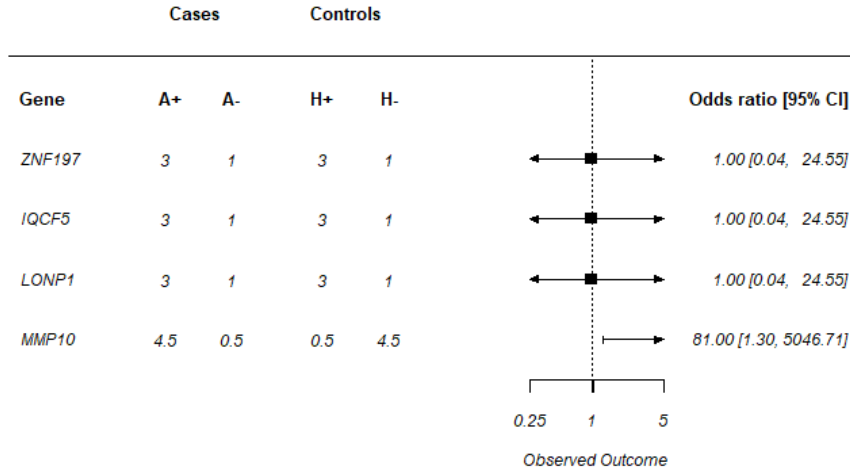
Figure 4.8.: Forest plot of genes in family 1. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

MMP10 was the only variant with perfect co-segregation as well as a significant odds ratio of 81 (see fig.4.8), hence we decided to validate this variant in the rest of the available family members. A complete table of the validation results of all family members can be found below (table 4.9).

Table 4.9.: Results of the complete MMP10 validation round. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

IndividualID	age@CAD	LAA	MMP10
8409501*	44	57	2
8409502	-	44	2
8409503	-	53	2
8409504	-	43	2
8409505	-	38	1
8409508*	56	71	2
8409509	56	68	2
8409511	-	75	1
8409512	-	68	1
8409513	-	70	1
8409514	-	42	1
8409516	-	70	1
8409517	-	42	2
8409519	-	41	2
8409520*	54	66	2
8409521	-	36	1
8409523	-	67	1
8409524	-	47	1
8409526	-	49	2
8409528	-	49	1
8409533	-	37	1
8409534	-	48	1
8409535	-	40	1
8409536	-	36	1

The mean age of disease was at 52.5 years for this family with a standard deviation of 4.97. This resulted in an age cut-off of 57 years. We calculated a significant age-unadjusted odds ratio of 22.85 and a significant age-adjusted odds ratio of 135 (see table 4.10).

Table 4.10.: Calculation of the odds ratio and its confidence interval for the *MMP10* variant. A stands for affected, H for healthy and the +/- indicate whether the variant is present (+) or not (-). AC is short for the age-adjusted calculations

Variant	A+	A-	H+	H-	odds ratio	95% Confidence Interval	
MMP10	4.5	0.5	6.5	16.5	22.85	1.07	487.02
MMP10 AC	4.5	0.5	0.5	7.5	135	2.26	8069.03

The highest unadjusted logarithm of the odds (LOD) score is 1.73 at an r of 0.23, the adjusted one is very close to significance with 2.97 at an r of 0.01 (see fig 4.9).
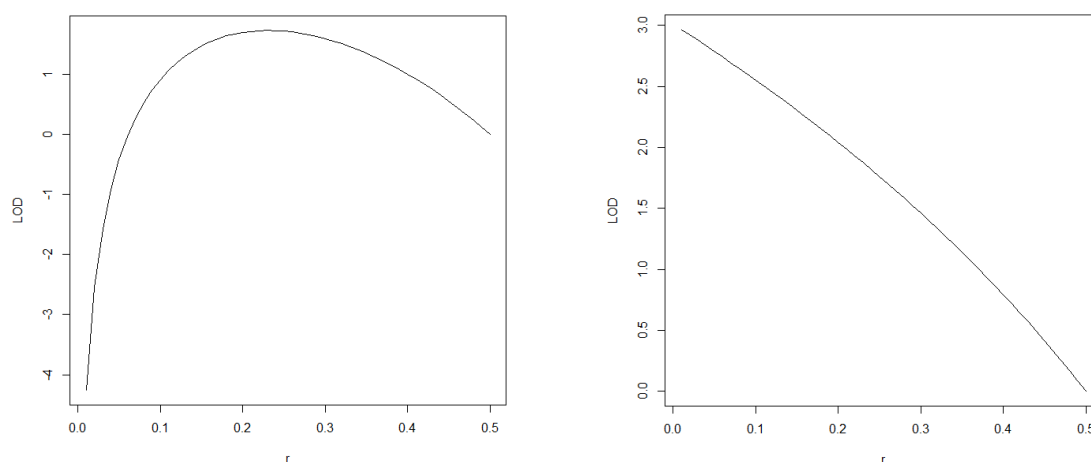
Figure 4.9.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *MMP10* variant.

To consider the influence of environmental factors, we collected the phenotypic data for obesity, hypercholesterolemia, statin treatment, diabetes, hypertension and smoking for each validated family member (see tab.4.11). Only one family member is diabetic, a quarter have hypercholesterolemia, nearly half of them suffer from hypertension and 64% smoke.

Table 4.11.: Phenotypic data for all family members in family. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
8409501	48	2	28.65	44	138	Fluvastatin	40	1	2	2	1
8409502	43	1	24.81		169.7			2	1	2	1
8409503	45	1	27.78		198			1	1	2	1
8409504	36	2	18.34		130.3			2	2	2	1
8409505	28	2	27.76		185			2	1	1	1
8409508	62	1	28.73	56		Pravastatin	20	0	2	2	1
8409509	59	1	31.02	56		Simvastatin	20	0	1	2	1
8409511	67	1	28.41		136.3			2	1	2	1
8409512	60	1	24.42		139.39			2	2	2	1
8409513	63	1	31.56		130			2	1	2	2
8409514	32	2	19.94		135			2	0	2	2
8409516	62	2	25.15		116	Atorvastatin	20	1	1	2	2
8409517	32	2	21.8		111			2	2	2	1
8409519	33	2	18.93		87			2	2	2	1
8409520	55	2	27.94	54	96	Atorvastatin	20	1	0	2	2
8409521	29	1	27.47		123			2	2	2	1
8409523	59	1	26		146	Lovastatin	20	1	1	2	2
8409524	39	2	24.34		122			2	2	2	1
8409526	39	2	21.26		94			2	1	2	1
8409528	39	2	22.77		104			2	1	2	2
8409533	27	1	26.83					0	1	2	2
8409534	38	1	24.86		87			2	2	2	1
8409535	33	2	21.22		98			2	2	2	2
8409536	29	2	24.16		88			2	1	2	2
8409537	52	1	23.39		229			1	2	2	1

The variant leads to an AA change from a Tyrosine to an Asparagine at position 41. While Tyrosine has a polar side group with a phenol group, Asparagine is smaller and has no charged side chain. The variant is located in the pro-peptid of the MMP10 protein, which would in principle be cleaved to activate the protein. Hence we were interested to see if the variant had an effect on the protein structure. We modeled the pro-peptid (AA 18-98) with and without the variant and then ran a molecular dynamics (MD) simulation to investigate the stability of the hydrogen bonds (see material and methods page 29).

Table 4.12.: Number of residues in favored, allowed and outlier regions according to the Ramachandran plot. WT stands for wild type and Y41N for the mutant

model	residues in favored region	residues in alllowed region	residues in outlier region
MMP10_WT	62 (91.2%)	6 (8.8%)	0 (0%)
MMP10_Y41N	66 (97.1%)	2 (2.9%)	0 (0%)

We modeled both, the wild type (WT) and the mutant pro-peptid and evaluated the quality of the modeled pro-peptides. The z-scores of -4.7 and -4.54 are well within the range of native conformations and the residues energy are largely negative. Additionally, we used Ramachandran plots (fig.4.10 and tab.4.12) to analyze their quality. The torsion angle of all residues (phi and psi) are within allowed regions. Hence we assume the models of the pro-peptide to be of high quality.
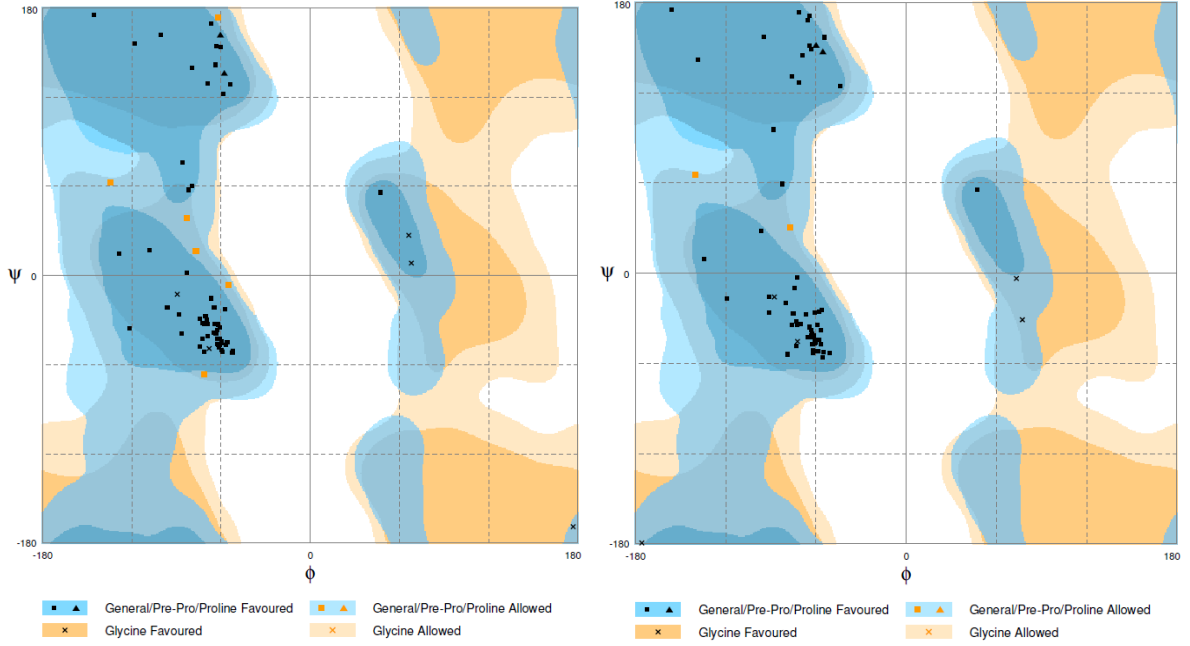
Figure 4.10.: Ramachandran plot of MMP10 wild type (left) and mutant (right) pro-peptid.

We analyzed the root-mean-square deviation of the C alpha (RMSDCα) to evaluate the stability of the MD-simulation. We assume a model as stable when the RMSDCα fluctuates around a value with less then 0.5Ådifference, which is the case for both models at around 2ns (see fig.4.11). Both models converged against a RMSDCα of 1.75Å. For all further analysis of the MD simulation we then solely considered the stable simulation period.

The root mean square fluctuation (RMSF) shows the fluctuation of the modeled residues over the specified time frame. It can be seen in fig.4.12 that there were no significant fluctuation differences between wild type and mutant pro-peptid. Some differences can be explained by their positioning close to or in the loop region.
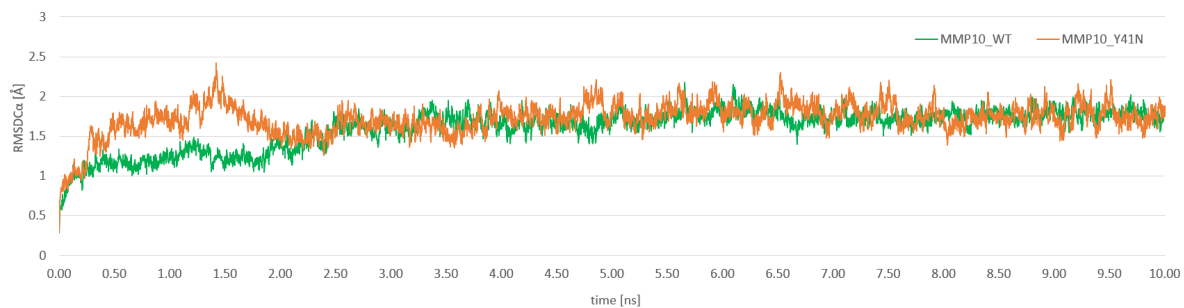
Figure 4.11.: RMSDCα of WT and mutant (Y41N) model across the MD-simulation. The WT model is shown in green and the mutant in orange.
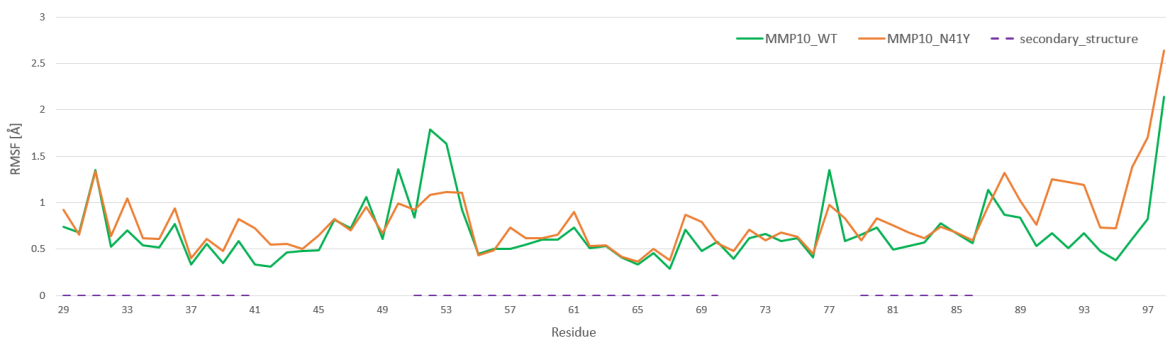
Figure 4.12.: This figure shows the RMSF values per residue of the WT and the mutant MMP10 propeptid.

We then analyzed the formation of secondary structures such as alpha-helices in the stable simulation period and we could see very little differences between wild type and mutant. All residues formed the same secondary structure (alpha-helix or coil) in mutant and wild type, with only small deviations in and close to loop regions with the difference reaching average of 1.89%.

However, when we looked at the stability of the hydrogen bonds we saw a difference between the wild type and the mutant. We took all hydrogen bonds into account that were formed for 50% of the simulation time (see tab.4.13). In the wild type pro-peptide, four hydrogen bonds were formed for more than 90% of the time, in the mutant only two of those reached the 50% and none of them were present for 90% of the simulation time. We observed a bond between 37TYR and 41TYR or 41ASN for both wild type or mutant. The same was true for a bond between 95ASP and 41TYR or 41 ASN. However, this hydrogen bond was only present in 68% of the time in the mutant compared to 96% in the wild type. Additionally the bond between 96VAL and 41ASN only formed 38% of the time, despite coming from the same oxygen as the previous hydrogen bond. The bond between 91 CYS and residue 41 was not formed at all, which could be explained by the missing polar side chain compared to the wild type. This clear difference in distance from the residue 91 to residue 41 between wild type and mutant is shown in figure 4.13.

Table 4.13.: Hydrogen bonds (h-bond) that were observed in 50% of the simulation time of the wild type and the mutant MMP10 pro-peptide.

Atom 1		Atom 2 WT/mutant		h-bond WT [%]	h-bond Y41N [%]
95ASP	N	41TYR/41ASN	O	95.53	67.77
91CYS	O	41TYR	OH	92.63	0
37TYR	O	41TYR/41ASN	N	99.87	83.13
96VAL	N	41TYR/41ASN	O	97.7	38.47

Figure 4.13.: Energy-minimized MD models of the WT MMP10 pro-peptide on the left and the mutant on the right. The residue marked blue (left) and red (right) are the tyrosine (left) and asparagine (right). Residues in orange bond to our residue of interest over hydrogen bond depicted as yellow lines. The hydrogen bonds depicted in this figure were the average occurring h-bonds during the MD.

Finally we decided to see if we could find other variants in the *MMP10* gene in our cohort of 255 MI patients and if these variants co-segregate with disease in their respective families.

We found five additional MMP10 variants in our cohort. A chr11:102643604G>T variant, leading to a stop-gain, three nonsynonymous variants (chr11:102643636C>A, chr11:102647090T>A and chr11:102647396A>G), and a splice site mutation at chr11:102649482T>C.

The stop-gain was identified in five families : 8960, 8525, 4390, 9110 and 8215. We had no DNA available for family 4390. The variant did not co-segregate with disease in the other families. Taking all families together we found that six out of ten CAD-affected individuals carried the variant. Additionally, there was one healthy individual that also carried the variant.

Variant chr11:102643636C>A was found in families 7080 and 7434. Only two out of six affected family members carried the variant, none of the five healthy individuals carried this variant.

Chr11:102647090T>A was found in family 7538. This pedigree consisted of two brothers and one child. Both brothers were affected by CAD but only one carried the variant. The healthy child did not carry the variant.

The chr11:102647396A>G variant was found in families 7257 and 4337. It co-segregated perfectly in both families, as all family members with CAD carry the variant (4) and the healthy individual does not.

The splice site mutation was the most common variant and was found in eight families (9219, 4150, 7285, 7312, 6867, 4500, 6035 and 8887). Twelve out of 18 affected family members carried the variant. However, also three of the nine healthy individuals carried it.

4.2.4. Family 2 - 6849

Two generations with a total of 20 individuals were available for study in family 2 (see the pedigree in fig.4.14). We found 445 exonic, nonsynonymous variants in this family, 13 of which where shared by all three affected family members (tab.4.14). Twelve were nonsynonymous mutations and one is predicted to lead to a splice change.

We validated all the variants in this family as the first validation of the most promising variants remained inconclusive. The variants chr1:55085648C>T in *REN*, chr4:17845015C>G in *NCAPG* and chr15:91475021C>A in *HDDC3* looked most promising and were also validated in another three individuals (see table 4.15). We found that seven of the 11 patients affected with CAD carried the *REN* variant, and all nine healthy individuals had the wild type. All other variants had between one (*GTF2H1* and *HDDC3*) to three healthy individuals with the variant, and between two (*PIP* and *MGAM*) to seven affected family members without it.

Table 4.14.: Variants shared by all three exome sequenced individuals in family 2. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	PhastCons	SIFT	Polyphen2	DANN	CADD
chr1	55085648	C	T	exonic	FAM151A	nonsynonymous SNV	y	T	B	0.927	11.71
chr1	204125895	A	G	exonic	REN	nonsynonymous SNV	y	D	P	0.998	21.5
chr3	7503321	G	A	exonic	GRM7	nonsynonymous SNV	y	T	B	0.999	20.5
chr4	17845015	C	G	exonic	NCAPG	nonsynonymous SNV	y	D	B	0.903	14.18
chr7	112090748	C	T	exonic	IFRD1	nonsynonymous SNV	y	D	P	0.999	32
chr7	135323351	G	T	exonic	NUP205	nonsynonymous SNV	y	D	B	0.989	14.27
chr7	141732665	C	T	exonic	MGAM	nonsynonymous SNV	y	T	B	0.991	12.73
chr7	142829211	T	G	exonic	PIP	nonsynonymous SNV	y	D	D	0.983	15.09
chr11	18387342	A	G	exonic	GTF2H1	nonsynonymous SNV	y	T	B	0.985	11.62
chr14	57103317	G	A	splicing	TMEM260	NA	y	0	0	0.995	17.8
chr15	55912391	G	A	exonic	PRTG	nonsynonymous SNV	y	D	B	0.973	11.53
chr15	91475021	C	A	exonic	HDDC3	nonsynonymous SNV	y	D	P	0.997	32
chr21	35742799	A	G	exonic	KCNE2	nonsynonymous SNV	y	D	D	0.995	16.04

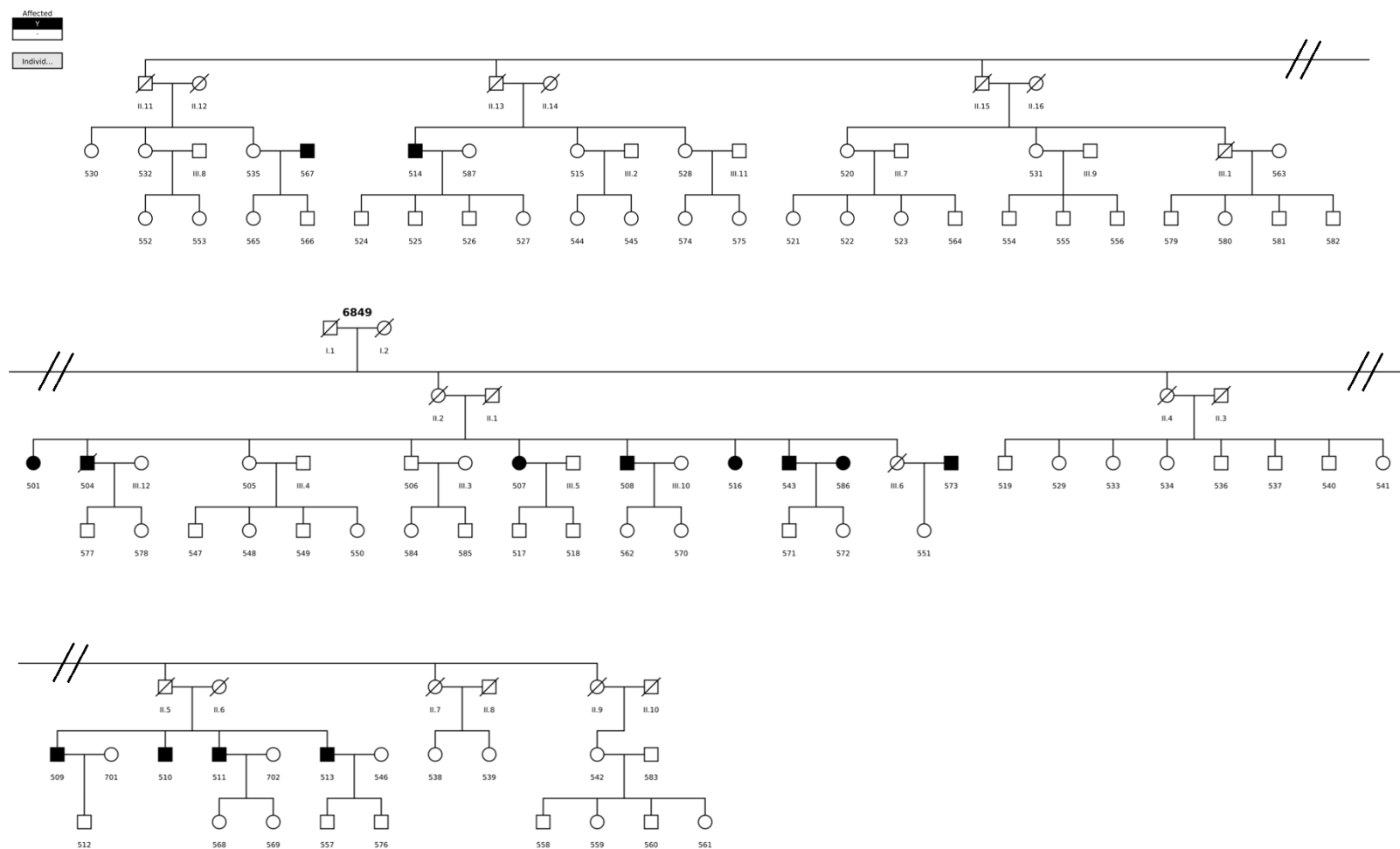**Figure 4.14.:** This figure shows the pedigree of family 2. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.15.: Results of the variant validation in family 2. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	REN	HDDC3	NCAPG	IFRD1	NUP205	MGAM	PIP	GTF2H1	TMEM260	PRTG	KCNE2	FAM151A	GRM7
6849501*	67	79	2	2	2	2	2	2	2	2	2	2	2	2	2
6849504	58	68	1	1	2	2	2	2	2	2	2	2	2	2	2
6849505	-	73	1	1	1	1	1	1	1	1	2	2	1	2	2
6849506	-	64	1	1	2	1	1	1	1	1	1	1	2	1	1
6849507	58	66	2	1	2	1	2	2	2	1	2	2	2	1	1
6849508	74	71	1	1	2	1	1	1	1	2	1	2	1	2	2
6849509*	37	71	2	2	2	2	2	2	2	2	2	2	2	2	2
6849510	52	72	1	2	1	1	1	2	2	2	2	2	2	1	2
6849511	36	75	2	1	1	2	2	2	2	2	2	1	1	2	2
6849513	53	64	1	1	1	1	2	2	2	1	1	1	2	1	2
6849514*	?	60	2	2	2	2	2	2	2	2	2	2	2	2	2
6849515	-	71	1	1	1	2	2	2	2	2	1	1	2	1	2
6849516	66	82	2	1	2	2	2	2	2	1	2	2	2	2	1
6849520	-	78	1	1	1	1	1	1	1	1	1	1	1	1	1
6849528	-	71	1	1	2	2	2	2	2	1	2	1	2	1	1
6849531	-	75	1	1	1	1	2	1	1	1	1	2	1	2	1
6849532	-	67	1	1	1										
6849533	-	69	1	1	1										
6849542	-	68	1	2	1										
6849543	72	75	2	1	1	1	1	1	1	2	1	1	1	2	1

The mean age of disease in this family was 58 years with a standard deviation of 12. This leads to an age cut-off of 70 years. Statistically the variant in *REN* had the highest significant odds ratio of 31.67 age-unadjusted, followed by a variant in *GTD2H1* with 13.33. When adjusted for age, none of the variants reached significance, which could be due to the small study size (n=13). The *REN* variant had the second highest odds ratio of 17.29, surpassed by variants in *MGAM* and *PIP* with 23.8. All variants had an odds ratio >1 implicating a damaging effect (see fig.4.15 and fig.4.16)

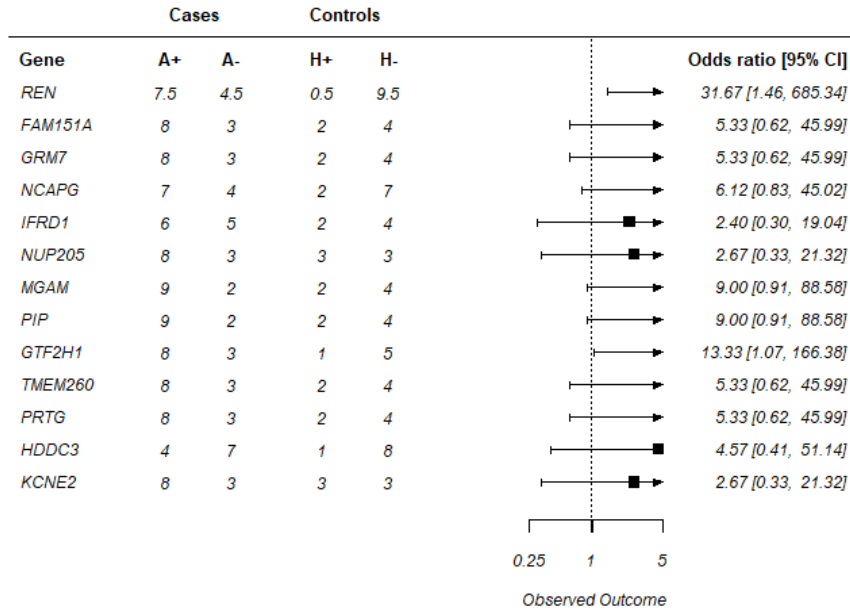|	Cases		Controls			
Gene	**A+**	**A-**	**H+**	**H-**		**Odds ratio [95% CI]**
REN	7.5	4.5	0.5	9.5		31.67 [1.46, 685.34]
FAM151A	8	3	2	4		5.33 [0.62, 45.99]
GRM7	8	3	2	4		5.33 [0.62, 45.99]
NCAPG	7	4	2	7		6.12 [0.83, 45.02]
IFRD1	6	5	2	4		2.40 [0.30, 19.04]
NUP205	8	3	3	3		2.67 [0.33, 21.32]
MGAM	9	2	2	4		9.00 [0.91, 88.58]
PIP	9	2	2	4		9.00 [0.91, 88.58]
GTF2H1	8	3	1	5		13.33 [1.07, 166.38]
TMEM260	8	3	2	4		5.33 [0.62, 45.99]
PRTG	8	3	2	4		5.33 [0.62, 45.99]
HDDC3	4	7	1	8		4.57 [0.41, 51.14]
KCNE2	8	3	3	3		2.67 [0.33, 21.32]

0.25   1   5

Observed Outcome

Figure 4.15.: Forest plot of genes in family 2. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
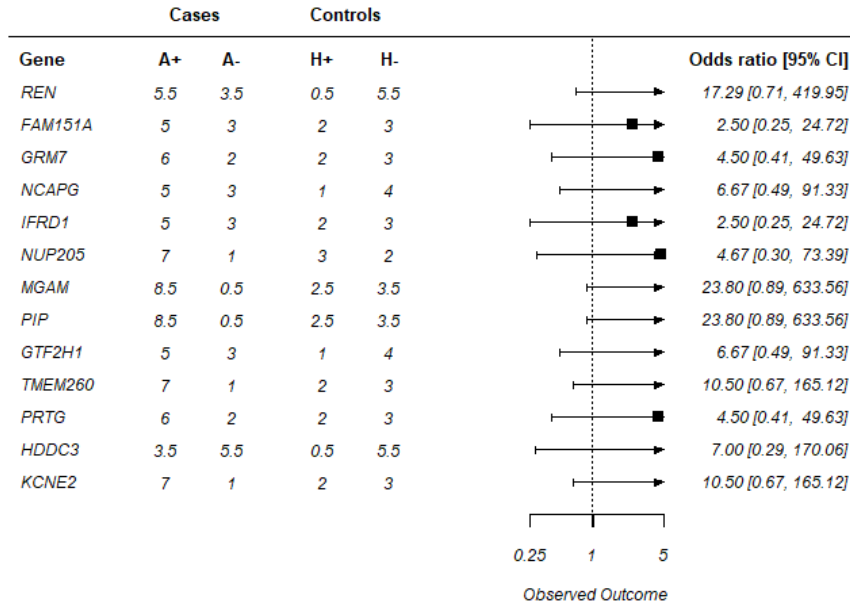
Figure 4.16.: Age-adjusted forest plot of genes in family 2. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

Since the variants in *REN* and *GTD2H1* reached significant odds ratios, we also calculated their LOD scores. None of them reached significance. The highest unadjusted LOD score for *REN* is 1.67 at an r of 0.2 and 0.86 at 0.23 for the adjusted one (see fig 4.17). *GTF2H1* has the highest LOD score at 1.09 with an r of 0.24 when unadjusted and 0.43 with an r of 0.31 when adjusted for age (see fig 4.18).
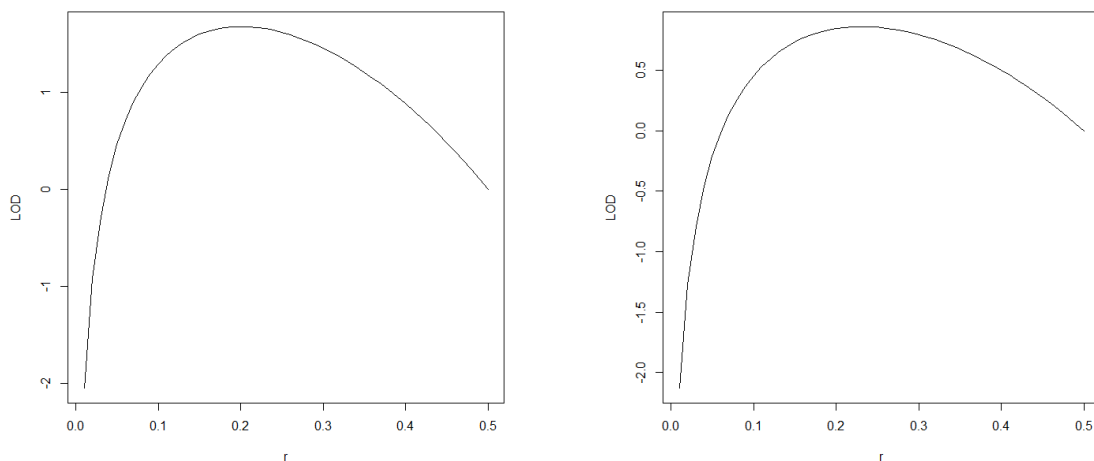
Figure 4.17.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *REN* variant.
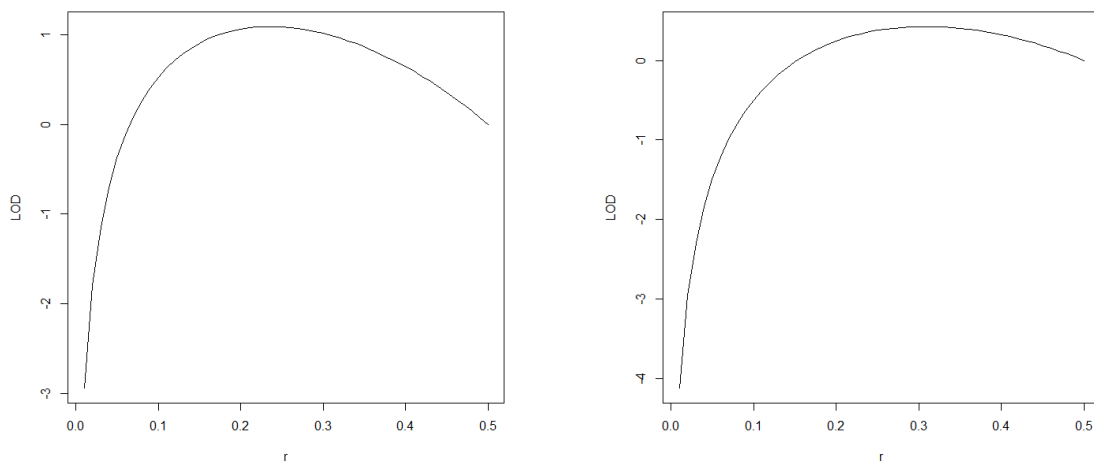
Figure 4.18.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *GTF2H1* variant.

To consider other risk factors, we also analyzed the BMI, hypercholesterolemia, smoking status, diabetes and hypertension in this family (see tab.4.16). Half of the validated family members were obese, 40% suffer from hypercholesterolemia, 65% had hypertension, 25% were diabetic and 35% were smokers. It is important to note that six of the seven individuals with the *REN* variant suffered from hypertension. *REN* codes for

renin which an important player in the renin angiotensin pathway and is involved in blood pressure regulation.

Table 4.16.: Phenotypic data for all family members in family 2. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
6849501	69	2	31.22	67	146	Atorvastatin	40	1	1	2	2
6849504	67	1	20.66	58	69	Simvastatin	10	1	2	1	1
6849505	66	2	30.82		149			2	1	1	2
6849506	64	1	35.43		170			2	1	1	2
6849507	58	2	30.48	58	241			1	1	2	2
6849508	70	1	29.75	74	169			2	1	2	2
6849509	61	1	24.58	37	136	Atorvastatin	10	1	1	2	2
6849510	62	1	29.41	52	146	Atorvastatin	20	1	2	2	1
6849511	65	1	28.06	36	145			2	2	2	1
6849513	63	1	29.41	53	207			1	0	2	1
6849514	59	1	29.41		162			2	1	2	1
6849515	63	2	32.27		155			2	1	2	2
6849516	75	2	22.31	66	131	Atorvastatin	80	1	1	2	1
6849520	70	2	32.03		108			2	2	2	2
6849528	61	2	36.2		119			2	1	1	2
6849531	67	2	33.31		85			2	0	2	2
6849532	60	2	27.18		186			2	0	2	2
6849533	59	2	33.27		171			2	1	2	2
6849542	60	2	31.89		136			2	1	1	2
6849543	73	1	28.34	72	93	Fluvastatin	20	1	1	2	1

4.2.5. Family 3 - 8882 and 8973

We had data available to analyze 37 individuals spanning two generations in family 3 (see pedigree in fig.4.19). We found 446 variants in total. Thirteen nonsynonymous variants were shared between all affected individuals (see table 4.17).

After the first validation of the variants with at least three deleterious scores, no conclusion could be drawn and we validated the variants not fulfilling our initial requirements as well. All variants were validated in 13 individuals (see tab.4.18).

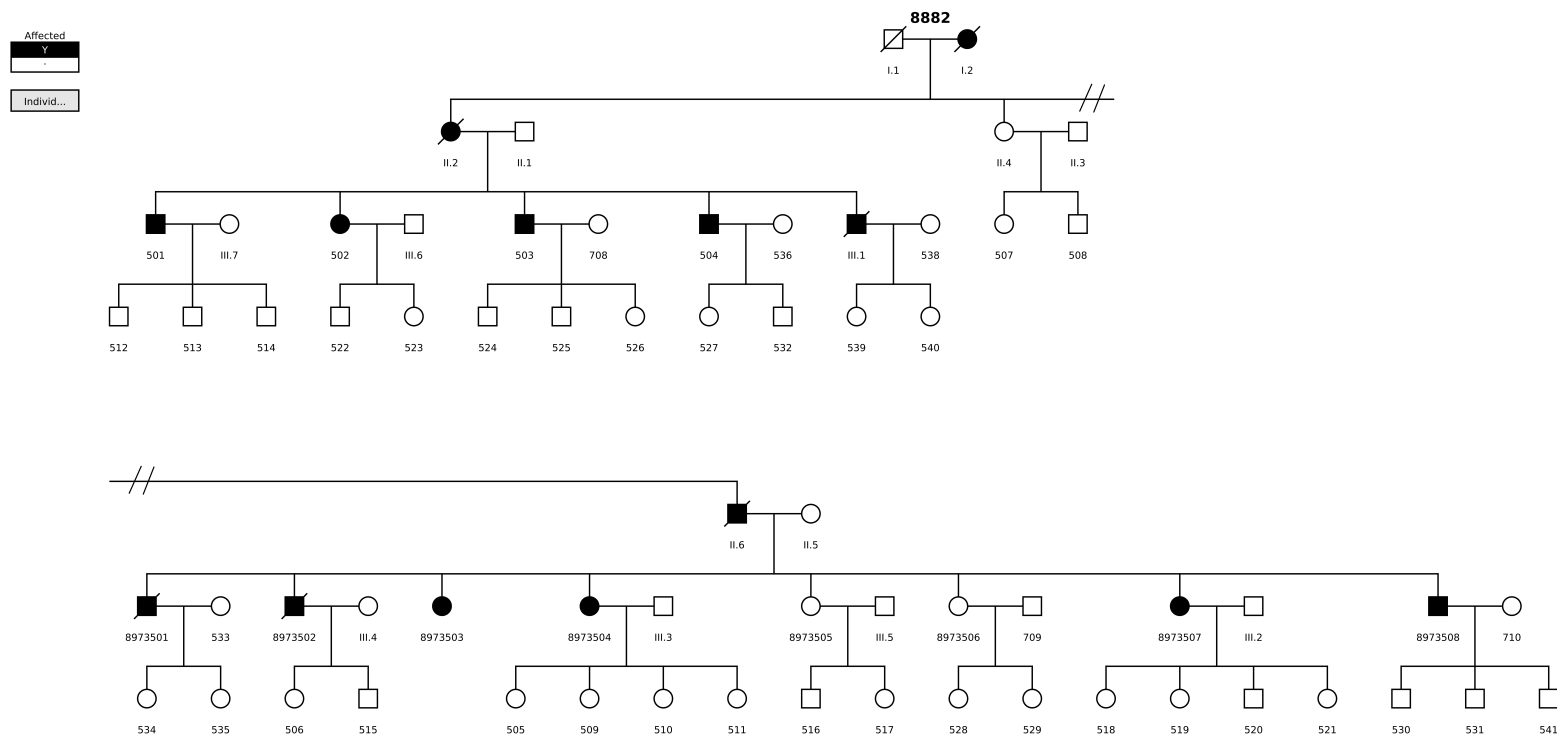**Figure 4.19.:** This figure shows the pedigree of family 3. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.17.: Variants shared by all three exome sequenced individuals in family 3. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	PhastCons	SIFT	Polyphen2	MutationTaster	CADD
chr1	94487490	A	G	exonic	ABCA4	nonsynonymous SNV	y	D	B	D	14.45
chr1	110260040	A	G	exonic	GSTM5	nonsynonymous SNV	y	T	B	N	12.54
chr1	153277491	A	G	exonic	PGLYRP3	nonsynonymous SNV	y	T	B	N	2.91
chr1	154574541	G	C	exonic	ADAR	nonsynonymous SNV	y	D	D	D	17.3
chr1	156255416	C	G	exonic	TMEM79	nonsynonymous SNV	y	T	D	N	20.3
chr2	98273651	G	A	exonic	ACTR1B	nonsynonymous SNV	y	D	B	D	22.1
chr7	25191295	A	C	exonic	C7orf31	nonsynonymous SNV	y	D	D	D	17.91
chr11	972248	G	A	exonic	AP2A2	nonsynonymous SNV	y	T	P	D	10.19
chr11	2939269	G	A	exonic	SLC22A18	nonsynonymous SNV	y	T	B	N	7.95
chr16	9858173	G	T	exonic	GRIN2A	nonsynonymous SNV	y	0	D	D	13.37
chr17	60767649	G	A	exonic	MRC2	nonsynonymous SNV	y	T	B	N	4.78
chr18	47317912	G	A	exonic	ACAA2	nonsynonymous SNV	y	T	B	D	9.315
chr20	3641868	A	T	exonic	GFRA4	nonsynonymous SNV	y	D	D	D	18.29

Table 4.18.: Results of the variant validation in family 3. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position, two is a heterozygous variant genotype and three indicates a homozygous variant. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	GSTM5	PGLYRP3	ADAR	TMEM79	ABCA4	C7orf31	SLC22A18	AP2A2	GRIN2A	MRC2	ACAA2	GFRA4	ACTR1B
8882501	58	82	2	2	2	2	2	2	2	2	2	2	1	2	1
8882502*	62	75	2	2	2	2	2	2	2	2	2	2	2	2	2
8882503*	49	77	2	2	2	2	2	2	2	2	3	2	2	2	2
8882504	68	80	1	1	1	1	1	1	2	2	2	1	1	1	1
8882537	-	85	1	1	1	1	1	1	1	1	1	1	1	1	1
8973501*	67	69	2	2	2	2	2	2	2	2	2	2	2	2	2
8973502	41	67	1	1	1	1	1	2	1	1	2	2	2	1	1
8973503	72	83	2	2	2	2	2	2	1	1	2	1	2	2	1
8973504	76	80	1	1	1	1	1	1	1	1	1	1	2	1	2
8973505	-	80	1	1	1	1	1	1	1	1	1	1	1	2	2
8973506	-	79	1	1	1	1	1	1	1	1	2	1	2	2	1
8973507	72	73	2	2	2	2	2	2	1	1	2	2	1	1	1
8973508	67	71	2	2	2	2	2	1	2	2	1	2	1	2	1

In this family the mean age of disease is 63 years with a standard deviation of 10. This leads to an age cut-off of 73 years.
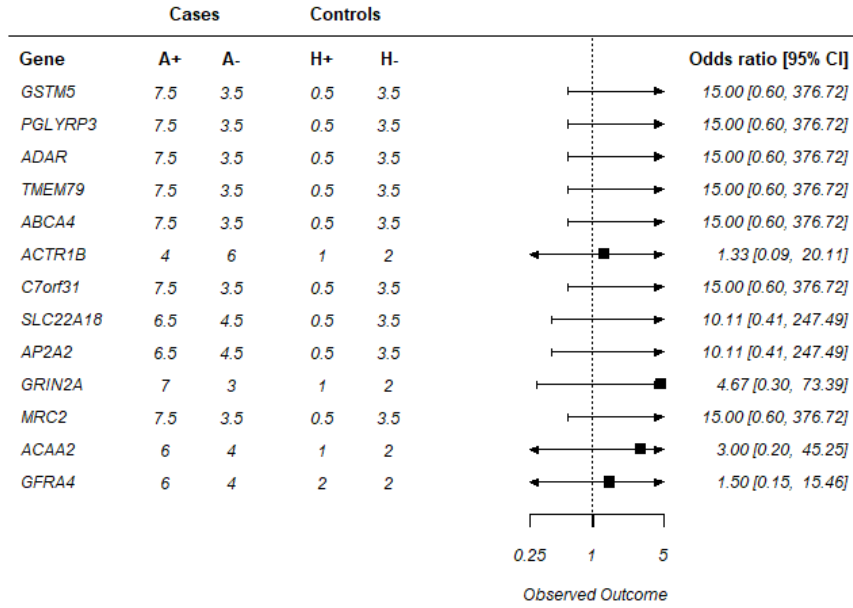

Figure 4.20.: Forest plot of genes in family 3. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
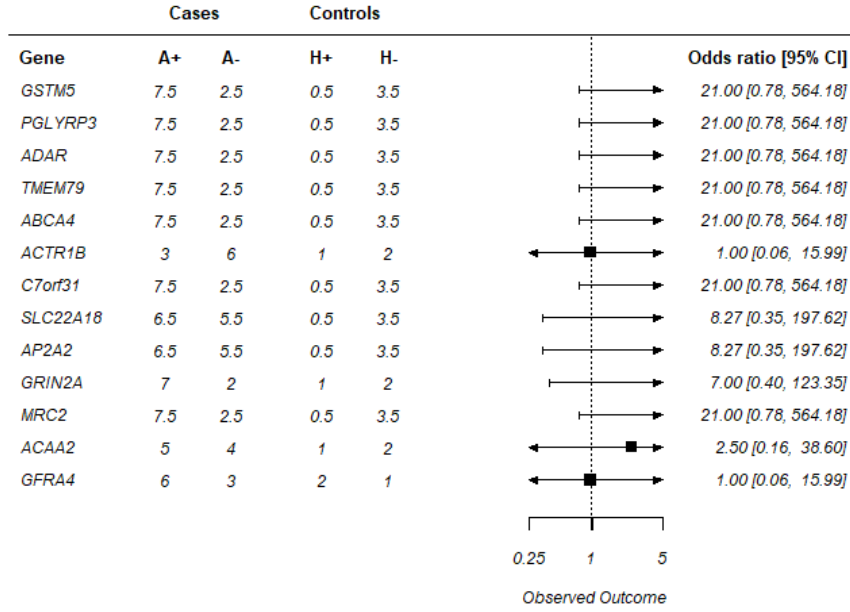
Figure 4.21.: Age-adjusted forest plot of genes in family 3. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

All odds ratios for the validated family members are greater than one, but none of them were significant (see fig.4.20 and fig.4.21) However, a paper has been published linking a variant in $ADAR$ to CAD, so we decided to validate the $ADAR$ variant in the rest of the available family members [127]. See below for a complete table of the $ADAR$ validation (tab.4.19). In total seven of the eleven individuals affected by CAD carried the variant, and five of 28 healthy family members carried it.

Table 4.19.: Results of the complete *ADAR* validation round. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

IndividualID	age@CAD	LAA	ADAR
8882501	58	82	2
8882502	62	75	2
8882503	49	77	2
8882504	68	80	1
8882505	-	50	1
8882506	-	41	1
8882509	-	58	1
8882511	-	57	1
8882512	-	61	2
8882513	-	57	1
8882515	-	46	1
8882516	-	52	1
8882517	-	49	1
8882518	-	52	1
8882519	-	48	2
8882520	-	48	1
8882521	-	44	1
8882523	-	46	1
8882526	-	42	1
8882527	-	47	1
8882528	-	44	1
8882530	-	47	2
8882531	-	36	2
8882532	-	39	1
8882534	-	44	1
8882535	-	46	2
8882537	-	85	1
8882540	-	45	1
8882541	-	41	1
8973501	67	69	2
8973502	41	67	1
8973503	72	83	2
8973504	76	80	1
8973505	-	80	1
8973506	-	79	1
8973507	72	73	2
8973508	67	71	2

Including the newly validated family members, we calculated a significant age-unadjusted odds ratio of 8.05 and 15 for non-adjusted samples, which is not significant (see table 4.22).

Table 4.20.: Calculation of the odds ratio and its confidence interval for the $ADAR$ variant. A stands for affected, H for healthy and the $+$/- indicate whether the variant is present ($+$) or not (-). AC is short for the age-adjusted calculations

Variant	A+	A-	H+	H-	odds ratio	95% Confidence Interval	
ADAR	7	4	5	23	8.05	1.69	38.44
ADAR AC	7.5	3.5	0.5	3.5	15	0.60	376.72

If not adjusted for age, the highest LOD score for $ADAR$ is 2.59 at an r of 0.23. For the age-adjusted values the highest is 0.86 at 0.23 (see fig 4.22).
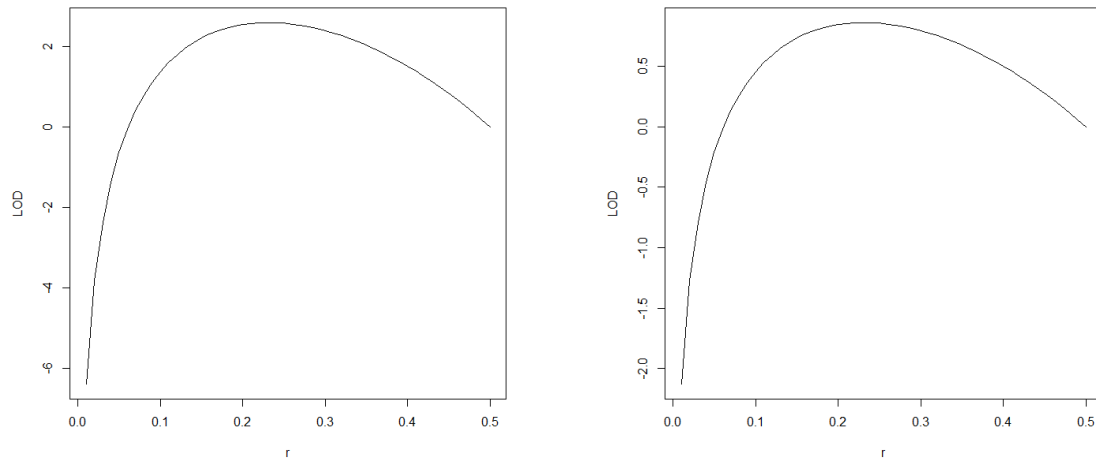

Figure 4.22.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) $ADAR$ variant.

When looking at other potential risk factors, we can see that with the exception of smoking (73% were smokers), only one of the validated individuals was obese, 14% had hypercholesterolemia and 27% suffered from hypertension (see tab.4.21).

Table 4.21.: Phenotypic data for all family members in family 3. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
8882501	83	1	25.99	58	162			2	1	2	1
8882502	77	2	27.99	62	123	Atorvastatin	20	1	1	2	1
8882503	73	1	27.68	49 0	119	Simvastatin	20	1	1	2	1
8882504	72	1	24.62	68	179			2	2	2	1
8882505	45	2	19.72		147			2	2	2	1
8882506	33	2	22.49		124			2	2	2	1
8882509	50	2	19.78		136			2	2	2	1
8882511	47	2	24.61		148			2	2	2	1
8882512	51	1	26.12		185			2	2	2	1
8882513	47	1	25.08		181			2	2	2	1
8882515	38	1	32.89		134			2	2	2	1
8882516	45	1	23.74		187	Atorvastatin	5	1	2	2	1
8882517	41	2	23.03		88			2	2	2	1
8882518	45	2	25.28		112			2	0	2	2
8882519	41	2	21.26		147			2	2	2	1
8882520	38	1	22.49		146			2	2	2	1
8882521	36	2	22.96		138			2	2	2	1
8882523	39	2	17.93		154			2	2	2	2
8882526	34	2	20.55		124			2	2	2	1
8882527	40	2	20.20		118			2	2	2	1
8882528	37	2	20.52		107			2	2	2	1
8882530	39	1	25.06		109			2	2	2	2
8882531	36	1	26.23		169			2	1	2	1
8882532	31	1	21.05		107			2	2	2	1
8882534	36	2	25.71		111			2	2	2	1
8882535	39	2	24.61		96			2	2	2	1
8882537	75	1	27.68		176			2	1	2	1
8882540	38	2	20.76		86			2	2	2	2
8882541	31	1	21.97		74			2	0	2	2
8973501	71	1	28.72	67	177			2	2	2	1
8973502	69	1	27.4	41	189			2	2	2	1
8973503	74	2	25.95	72	162	Pravastatin	10	1	1	2	2
8973504	76	2	24.01		145			2	1	2	2
8973505	73	2	25.95		86			2	1	2	2
8973506	72	2	23.44		140	Pravastatin	10	1	1	2	2
8973507	65	2	23.74		130			2	2	2	2
8973508	67	1	26.47		166			2	1	2	1

Next, we looked if we could find further variants in the *ADAR* gene in our cohort of 255 MI patients.

We found three more nonsynonymous variants, one G to A transition at position 154557515 (family 6750), one A to C transition at position 154574056 (family 6544) and a C to T transition at 154574631. Additionally, we found the original chr1:154574541G>C variant in another family.

The chr1:154557515G>A did not co-segregate with disease in this family. Of the four affected individuals, only two carried the variant. One of the three healthy individuals also carried it.

Family 6544 showed no co-segregation between the chr1:154574056A>C variant and disease. Two of three CAD affected individuals carried the variants and so did one of the two healthy family members.

The third variant (chr1:154574631C>T) co-segregated with disease, however there were only two indivduals available. Both were affected and carry the *ADAR* variant.

The second family (4519) with the original variant consisted of three individuals, two of which were affected and carried the variant and one is healthy and does not. If these individuals are added to family 3, both the odds ratio of 10.8 and the LOD score of 3.16 at an r of 0.21 reach significance (see tab.4.22 and fig.4.23)

Table 4.22.: Calculation of the odds ratio and its confidence interval for the *ADAR* variant including both families. A stands for affected, H for healthy and the +/- indicate whether the variant is present (+) or not (-).

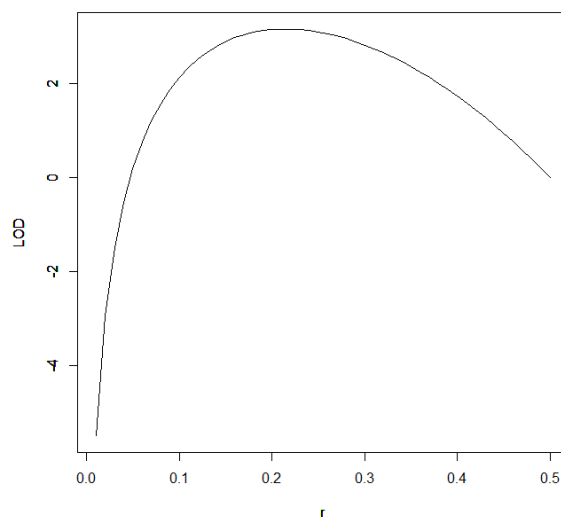| Variant	A+	A-	H+	H-	odds ratio	95% Confidence Interval	
ADAR	9	4	5	25	10.8	2.36	49.47

Figure 4.23.: Plot of the LOD score of the *ADAR* variant including both families.

81

4.2.6. Family 4 - 6556

We had blood available for two generations in family 4 (see pedigree at fig.4.24), where we found 473 variants. Ten of these were shared by all three exome sequenced family members, with eight nonsynonymous variations, one frameshift deletion and one non-frameshift insertion (see tab.4.23).

In this family we validated all variants in all twelve available family members (see tab.4.24). None of the variants co-segregates with disease. Some variants appear in all CAD affected individuals (*KIAA1147*, *CLCN1*, *SETD1A* and *PRSS36*) but are also found in 2-3 healthy individuals. When adjusted for age they co-segregate, but the number of individuals included (six) is too low to make reasonable assumptions.

Table 4.23.: Variants shared by all three exome sequenced individuals in family 4. The variants were annotated to the hg19/GrCh37 build.

| Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr3	4702696	C	-	exonic	ITPR1	frameshift deletion	NA	NA	NA	NA
chr3	10346794	G	A	exonic	SEC13	nonsynonymous SNV	0	D	D	29
chr7	4856904	T	C	exonic	RADIL	nonsynonymous SNV	T	D	D	18.25
chr7	141364841	C	A	exonic	KIAA1147	nonsynonymous SNV	T	B	D	18.19
chr7	143039510	G	C	exonic	CLCN1	nonsynonymous SNV	T	B	D	16.44
chr8	91804181	C	A	exonic	NECAB1	nonsynonymous SNV	T	B	D	13.06
chr10	29818731	G	A	exonic	SVIL	nonsynonymous SNV	0	D	D	15.7
chr16	30991240	-	CAG	exonic	SETD1A	nonframeshift insertion	NA	NA	NA	NA
chr16	31160532	G	T	exonic	PRSS36	nonsynonymous SNV	T	B	N	10.28
chr17	66879986	C	T	exonic	ABCA8	nonsynonymous SNV	T	D	D	20.4
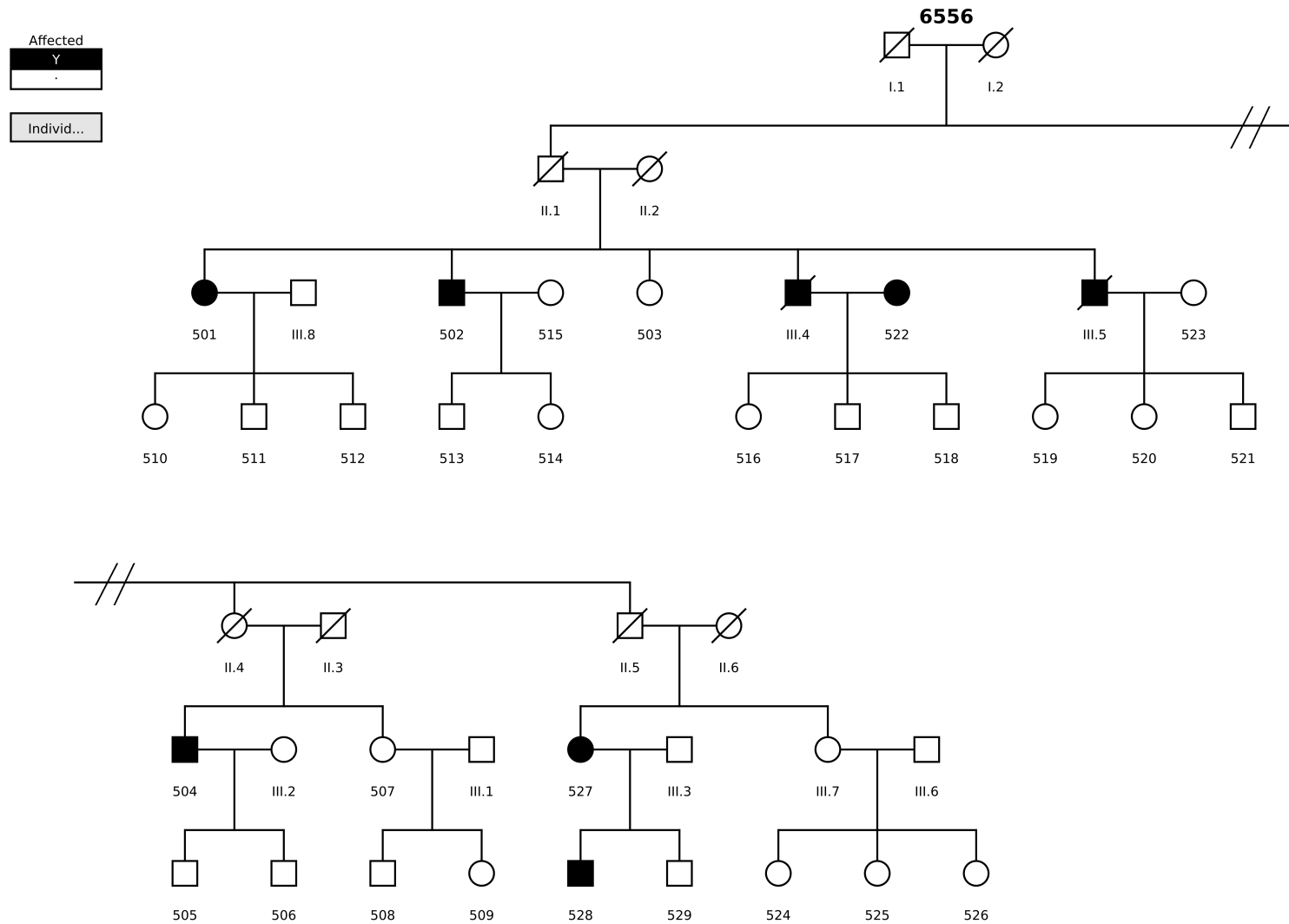
Figure 4.24.: This figure shows the pedigree of family 4. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.24.: Results of the variant validation in family 4. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	ITPR1	SEC13	RADIL	KIAA1147	CLCN1	NECAB1	SVIL	SETD1A	PRSS36	ABCA8
6556501*	61	73	2	2	2	2	2	2	2	2	2	2
6556502	66	71	1	1	1	3	2	1	1	2	2	2
6556503	-	68	1	1	1	2	2	2	2	1	1	1
6556504*	62	69	2	2	2	2	2	2	2	2	2	2
6556507	-	69	2	2	2	2	2	1	1	1	1	1
6556510	-	41	2	2	1	1	1	2	1	1	1	2
6556511	-	45	1	1	1	2	2	2	1	2	2	2
6556512	-	36	2	1	2	1	1	1	1	2	2	2
6556515	-	79	1	1	1	1	1	1	1	1	1	1
6556527*	71	78	2	2	2	2	2	2	2	2	2	2
6556528	52	54	1	2	1	2	2	2	1	2	2	1
6556530	-	72	1	1	1	1	1	1	1	1	1	1
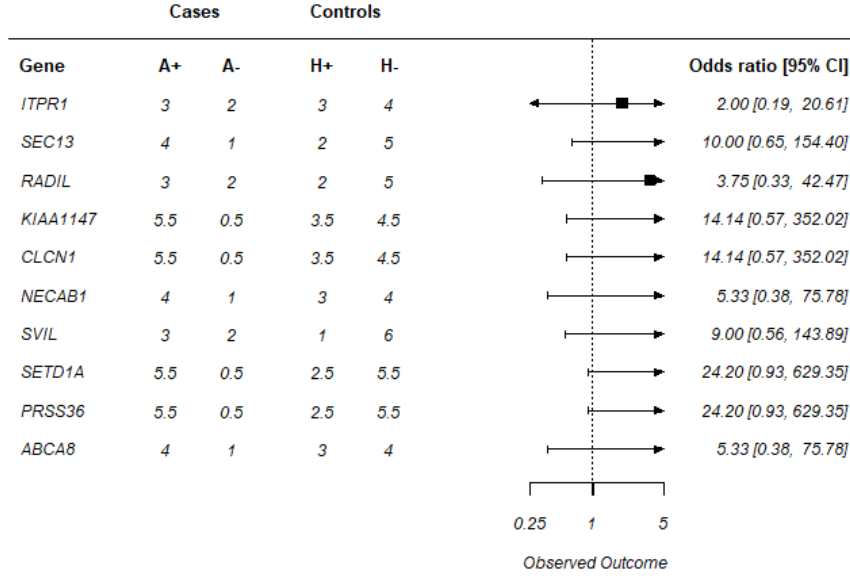
Figure 4.25.: Forest plot of genes in family 4. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
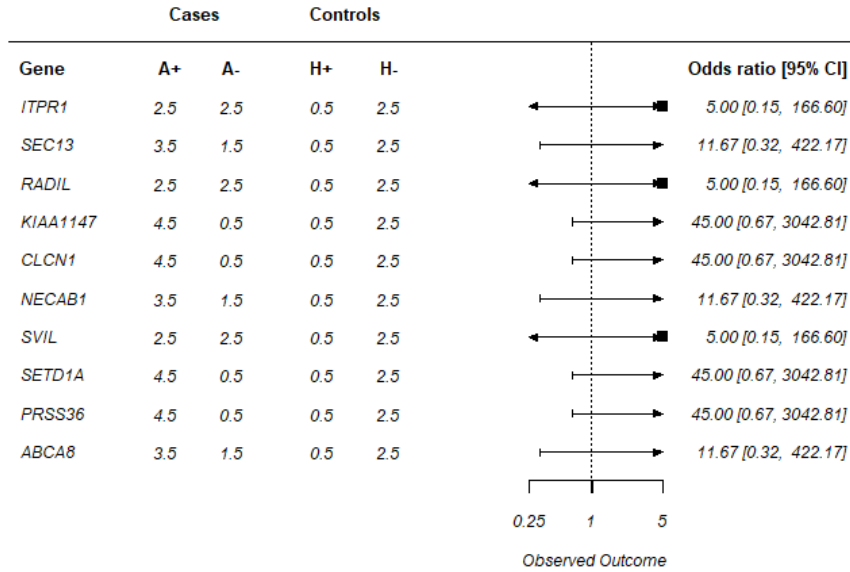

Figure 4.26.: Age-adjusted forest plot of genes in family 4. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

Family members were diagnosed with disease at an average of 63 years with a standard deviation of 6 years. Hence, the age cut-off was set to 69 for this family. All odds ratios were greater than one, with some even reaching 24.2 (*SETD1A* and *PRSS36*) for the unadjusted values and 45 for the age-adjusted ones (*SETD1A* and *PRSS36*). However, none of them were significant (see fig.4.25 and fig.4.26).

Around 15% of the validated family members were obese. Forty-one percent have hypercholesterolemia and/or hypertension, one was diabetic and half of them smoked (see tab.4.25)

Table 4.25.: Phenotypic data for all family members in family 4. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
6556501	63	2	28.73	61	122	Cerivastatin	0.09	1	1	2	1
6556502	71	1	30.76	66	147	Atorvastatin	5	1	1	2	1
6556503	60	2	25.35		154			2	1	2	2
6556504	69	1	23.99	62	116	Cerivastatin	0.30	1	1	2	2
6556507	63	2	28.65		161	0	0	1	2	1	2
6556510	34	2	20.9		97			2	2	2	1
6556511	35	1	21.13		141			2	2	2	1
6556512	29	1	23.8		116			2	2	2	1
6556515	69	2	21.26		127			2	2	2	2
6556527	71	2	26.73	72	139	0	0	1	2	2	2
6556528	47	1	28.40			Atorvastatin	10	0	1	2	1
6556530	71	2	31		132			2	0	2	2

4.2.7. Family 5 - 6642

In family five we had two generations available for analysis (see fig.4.27). We found a total of 447 variants for the the three exome sequenced individuals with four variants shared between them (see tab.4.26).

The median age for disease was at 56 with a standard deviation of 7. We calculated the cut-off at 63 years. The variants were validated in 39 family members (see tab.4.27).

Table 4.26.: Variants shared by all three exome sequenced individuals in family 5. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr4	37445144	A	G	exonic	NWD2	nonsynonymous SNV	T	B	D	13.22
chr7	95001590	T	C	exonic	PON3	nonsynonymous SNV	D	B	D	12.32
chr7	121651348	C	G	exonic	PTPRZ1	nonsynonymous SNV	D	D	D	12.33
chr9	121929598	C	T	exonic	BRINP1	nonsynonymous SNV	T	B	D	4.41
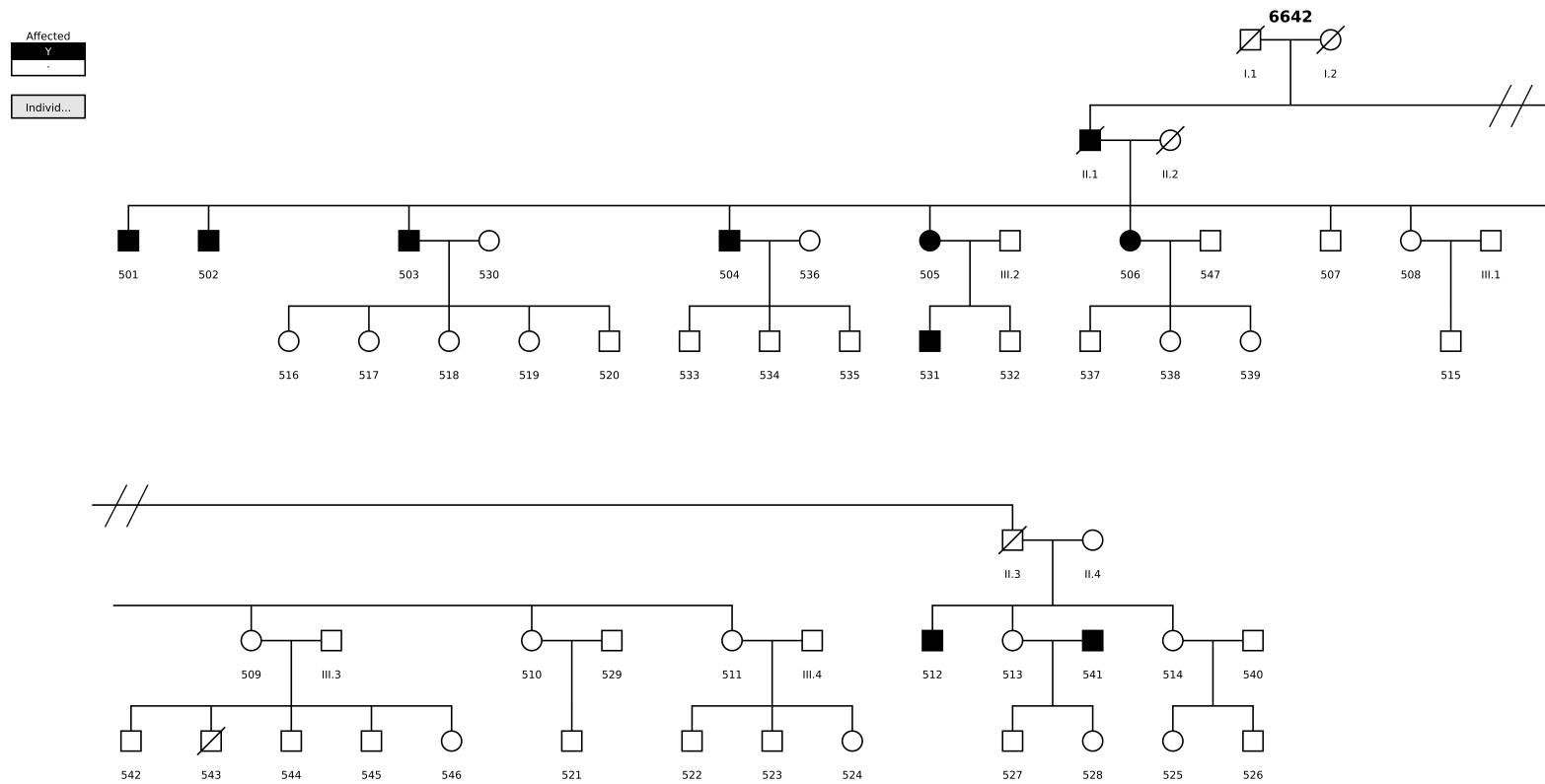
Figure 4.27.: This figure shows the pedigree of family 6642. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

For all variants, the odds ratios are above 1 when age unadjusted. Two variants, chr7:121651348C>G in *PTPRZ1* and chr9:121929598C>T in *BRINP1* reached significance with odds ratios of 8.6 and 11.1 (see fig.4.28 and fig.4.29). With the age cut-off, no variants reached significance.
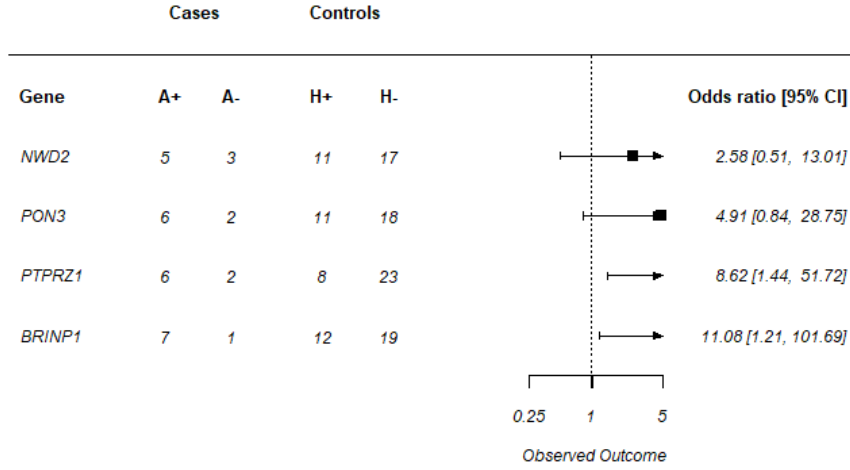
Figure 4.28.: Forest plot of genes in family 5. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
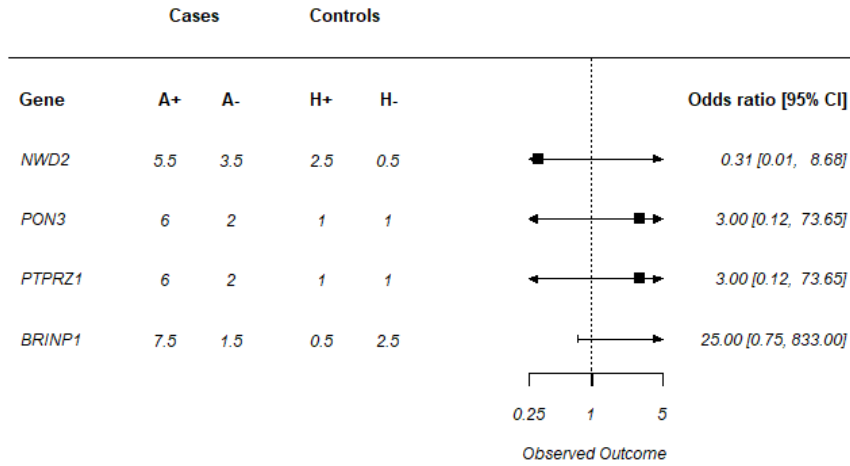
Figure 4.29.: Age-adjusted forest plot of genes in family 5. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

Table 4.27.: Results of the variant validation in family 5. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	NWD2	PON3	PTPRZ1	BRINP1
6642501*	57	68	2	2	2	2
6642502	49	65	2	1	2	1
6642503	63	76	1	1	1	2
6642504	57	70	1	2	2	2
6642505	56	74	1	2	2	2
6642506	59	72	2	2	1	2
6642507	-	66			1	2
6642508	-	58	1	1	1	2
6642509	-	60	1	2	2	2
6642510	-	67	2	2	2	1
6642511	-	59	2	1	1	2
6642512*	?	74	2	2	2	2
6642513	-	79	2	2	1	1
6642514	-	73	2	1	2	1
6642515	-	37	1	1	1	1
6642516	-	46	1	1	1	1
6642517	-	48	1	1	1	2
6642518	-	43	1	1	1	2
6642519	-	41	1	1	1	2
6642520	-	41	1	1	1	2
6642521	-	43	2	1	1	1
6642522	-	41	1		1	2
6642524	-	38	2	1	1	2
6642525	-	49	1	1	1	1
6642526	-	45	1	1	2	1
6642527	-	43	2	1	1	1
6642528	-	54	2	2	1	1
6642531*	41	54	2	2	2	2
6642532	-	53	2	1	1	1
6642533	-	43	1	1	1	1
6642534	-	46		1	1	1
6642535	-	42		2	1	1
6642537	-	43	2	1	1	1
6642538	-	45	2	2	1	1
6642539	-	36	1	2	1	1
6642542	-	48	1	2	2	2
6642544	-	38	1	2	2	1
6642545	-	42	1	2	2	1
6642546	-	39	1	2	2	2

We calculated the LOD scores for the *PTPRZ1* and the *BRINP1* variant, as these had reached significance when not adjusted for age. The highest LOD score for the *PTPRZ1* variant is 2.1 at an r of 0.26 when not adjusted for age. If adjusted for age the highest LOD score is 0.36 at 0.3 (see fig 4.30).
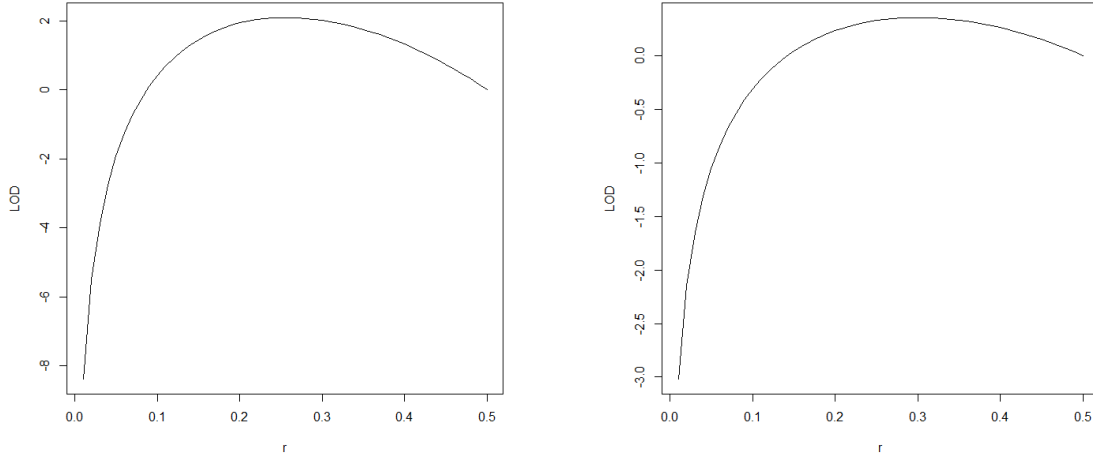

Figure 4.30.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *PTPRZ1* variant.

If not adjusted for age, the highest LOD score for *BRINP1* is 0.96 at an r of 0.33. For the age-adjusted values the highest is 1.6 at 0.1 (see fig 4.31).
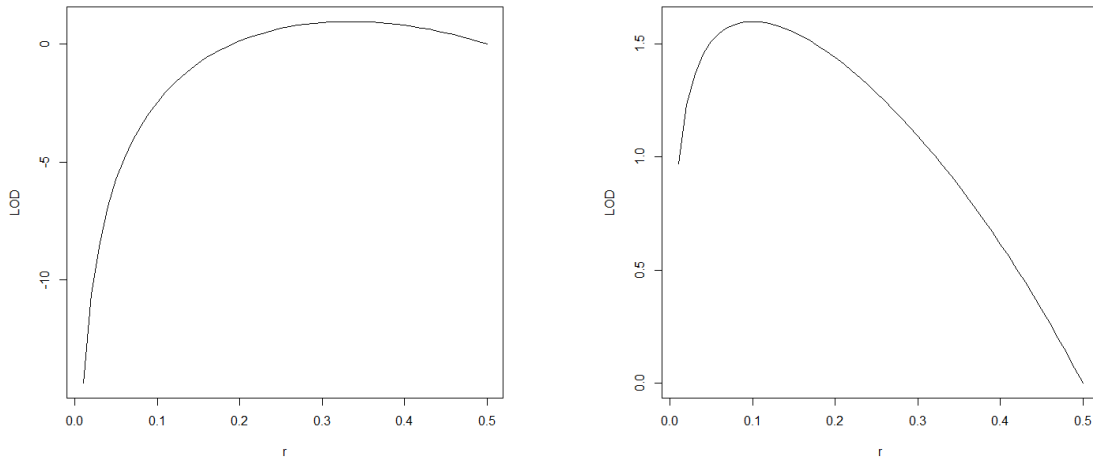
Figure 4.31.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *BRINP1* variant.

Looking at other risk factors in this family, we found that 28% of the validated family members were obese, 44% had hypercholesterolemia, 44% suffered from hypertension and 56% smoked (see tab.4.28).

Table 4.28.: Phenotypic data for all family members in family 5. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
6642501	61	1	30.12	57	167			2	2	2	1
6642502	55	1	29.04	49	211	Simvastatin	10	1	1	2	2
6642503	66	1	29.41	63	191			1	1	2	1
6642504	60	1	33.91	57	197	Cerivastatin	0.10	1	1	2	1
6642505	64	2	29.64	56	169	Atorvastatin	10	1	1	2	2
6642506	62	2	29.38	59	199			1	1	2	2
6642507	56	1	27.06		238			1	2	2	1
6642508	51	2	28.67		196			1	0	2	1
6642509	59	2	30.85		149	Cerivastatin	0.20	1	0	2	1
6642510	57	2	25		157			2	2	2	2
6642511	52	2	31.2		180	Cerivastatin	0.20	1	2	2	1
6642512	64	1	24.49		153	0	0	1	1	2	2
6642513	71	2	27.64		194			1	1	2	2
6642514	66	2	31.60		156			2	1	2	2
6642515	30	1	29.39		142			2	1	2	2
6642516	39	2	31.89		108			2	2	2	1
6642517	38	2	24.16		177			2	2	2	2
6642518	36	2	27.28		209			1	2	2	2
6642519	33	2	30.11		199			1	1	2	1
6642520	31	1	23.15		195			1	2	2	2
6642521	36	1	33.80		164			2	1	2	2
6642522	34	1	23.92		188			2	2	2	1
6642524	31	2	21.22		132			2	2	2	1
6642525	41	2	23.34		123			2	2	2	2
6642526	34	1	23.67					0	2	2	1
6642527	42	1	28.41		163			2	0	2	1
6642528	44	2	23.44		99			2	2	2	2
6642531	44	1	31.95	41	121	Simvastatin	10	1	2	2	1
6642532	43	1	29.35		189	Cerivastatin	0.20	1	2	2	2
6642533	33	1	25.26		90			2	0	2	2
6642534	34	1	27.47		108			2	2	2	1
6642535	32	1	25.25		160			2	2	2	1
6642537	33	1	23.55		101			2	2	2	2
6642538	37	2	30.12		179			2	2	2	1
6642539	28	2	38.10		168			2	2	2	1
6642542	40	1	24.28		187			2	2	2	1
6642544	37	1	27.78		197			1	1	2	1
6642545	36	1	29.03		181			2	1	2	1
6642546	31	2	27.1		152			2	1	2	1

4.2.8. Family 6 - 8942

Two generations were available for analysis in family 6 (see fig.4.32). After filtering, there were 467 variants found, of which six were shared. All six were nonsynonymous mutations (see tab.4.29).

We validated these six variants in 18 available family members and none of them showed perfect co-segregation with disease (see tab.4.30).

The mean age of disease was at 60 years with a standard deviation of 11, leading to an age cut-off of 71 years. The chr11:20676319G>A variant in *SLC6A5* shows a significant unadjusted odds ratio of 12, which increased to 17.5 when adjusted for age (see fig.4.33 and fig.4.34). All other variants did not reach significance. Surprisingly, the variant in *USP34* showed odd ratios below one, possibly showing a protective effect. The variant in *SRGAP1* had odds ratios around 1, indicating that this variant had no effect on disease in this family.
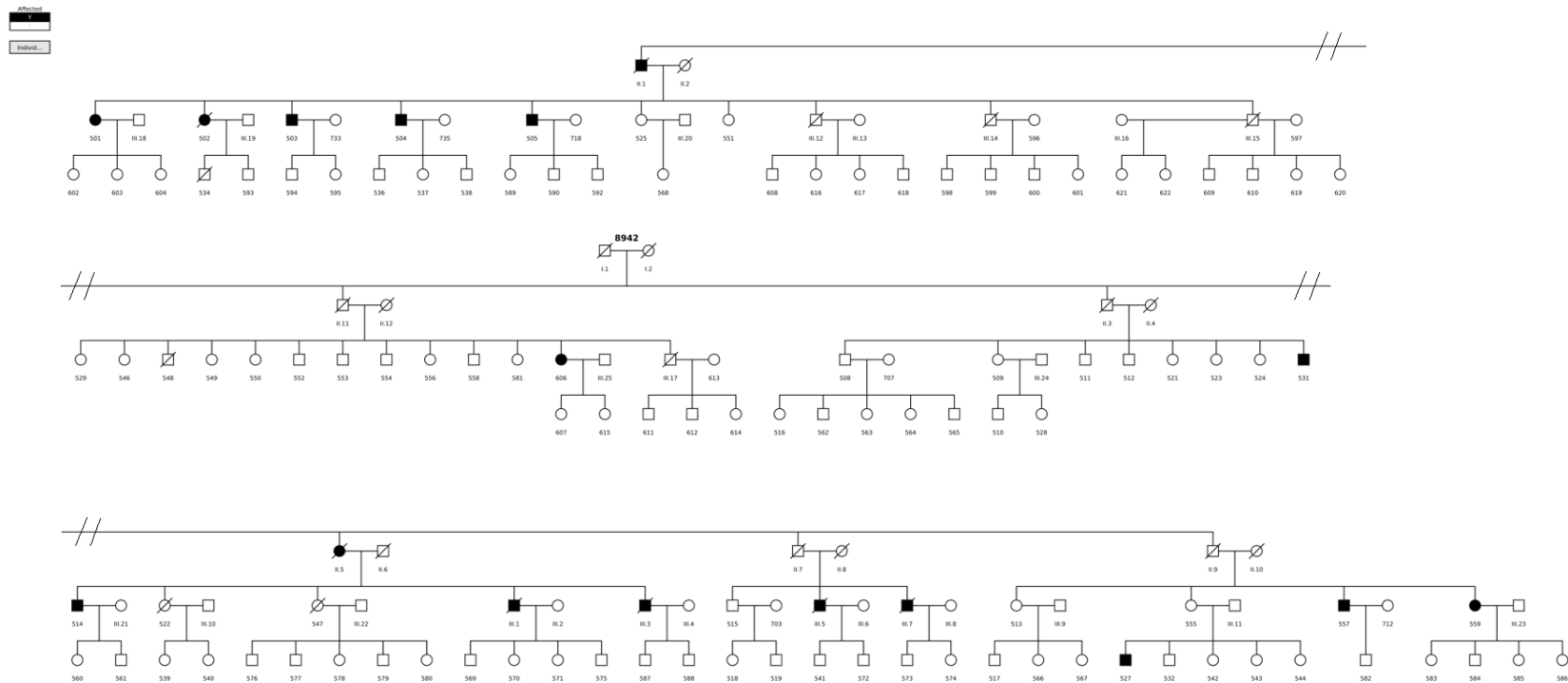
Figure 4.32.: This figure shows the pedigree of family 8942. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.29.: Variants shared by all three exome sequenced individuals in family 6. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr2	61522328	T	C	exonic	USP34	nonsynonymous SNV	T	B	D	12
chr3	184039075	C	T	exonic	EIF4G1	nonsynonymous SNV	D	D	D	16
chr9	140352246	C	T	exonic	NSMF	nonsynonymous SNV	0	B	D	12.16
chr9	140356442	C	T	exonic	PNPLA7	nonsynonymous SNV	T	D	D	24.3
chr11	20676319	G	A	exonic	SLC6A5	nonsynonymous SNV	T	B	D	20.8
chr12	64521724	A	G	exonic	SRGAP1	nonsynonymous SNV	0	B	D	3.34

Table 4.30.: Results of the variant validation in family 6. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	USP34	EIF4G1	NSMF	PNPLA7	SLC6A5	SRGAP1
8942501	51	75	1	2	2	2	2	2
8942502	49	65	1	2	2	2	2	2
8942503	52	67	1	2	1	1	2	1
8942504*	51	74	2	2	2	2	2	2
8942505	59	78	2	2	1	1	2	1
8942508	-	83	2	1	1	1	1	2
8942509	75	78	1	1	1	1	2	1
8942511	-	76	1	2	1	1	2	2
8942512	-	69	2	2	2	2	2	2
8942514*	70	83	2	2	2	2	2	2
8942515	-	80	1	2	1	1	1	1
8942522	-	81	2	2	1	1	1	2
8942527	35	50	1	1	1	1	1	1
8942531*	53	70	2	2	2	2	2	2
8942532	-	42	1	1	1	1	1	1
8942547	-	74	2	2	1	1	1	1
8942555	-	76	1	1	1	1	1	2
8942559	72	75	1	1	1	1	1	2
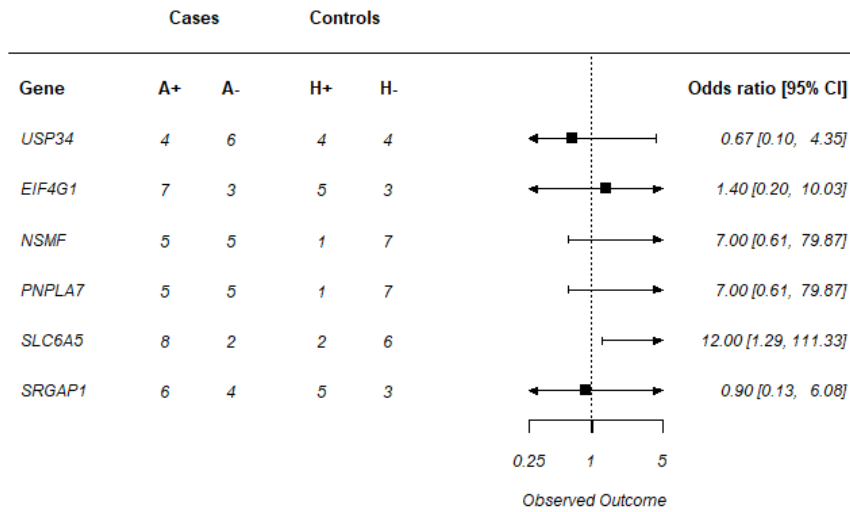
Figure 4.33.: Forest plot of genes in family 6. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
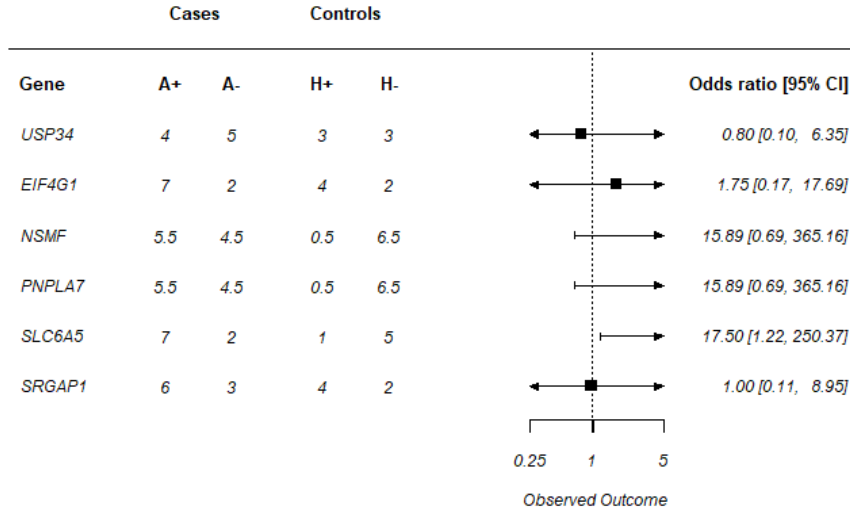
Figure 4.34.: Age-adjusted forest plot of genes in family 6. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

The LOD score did not reach three for the *SLC6A5* variant. When not adjusted for age the highest score was as 1.28 at an r of 0.22. If adjusted for age the highest LOD score was 1.26 at 0.2 (see fig 4.35).
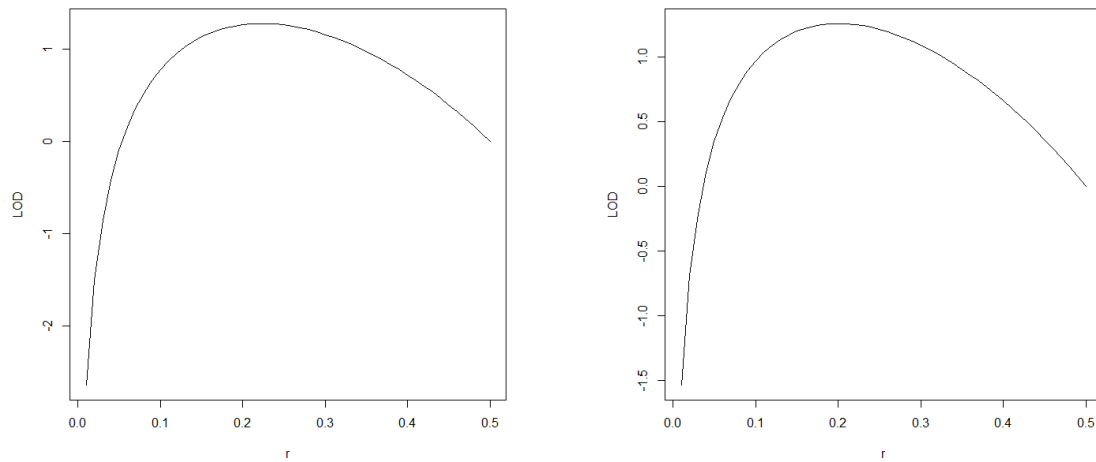
Figure 4.35.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) *SLC6A5* variant.

Only one of the validated family members was obese, but half of them suffered from hypercholesterolemia and over half of them smoked. About 78 % had hypertension and 17 % suffered from diabetes.

Table 4.31.: Phenotypic data for all family members in family 6. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m^2]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
8942501	66	2	23.62	51	218	0	0	1	1	2	2
8942502	66	2	22.27	49	180	0	0	1	1	1	2
8942503	58	1	22.4	52	166			2	1	2	2
8942504	75	1	22.86	51	159			2	1	2	1
8942505	69	1	26.2	59	125.8	Simvastatin	20	1	1	2	1
8942508	75	1	24.91		140	Pravastatin	10	1	1	2	1
8942509	71	2	30.82	75	127			2	1	2	2
8942511	69	1	26.45		211			1	1	2	1
8942512	62	1	24.86		167			2	1	2	1
8942514	73	1	28.345	70	179			2	1	2	1
8942515	73	1	28.73		89			2	1	2	1
8942522	81	2	27.48		199	Cerivastatin		1	2	2	2
8942527	40	1	27.94	35	184	Simvastatin	80	1	2	2	1
8942531	59	1	27.72	53	145			2	2	2	1
8942532	34	1	28.29		111			2	2	2	1
8942547	74	2	29.76		172			2	1	1	2
8942555	68	2	28.37		212			1	1	2	2
8942559	75	2	24.77	72	191			1	1	1	2

4.2.9. Family 7 - 8948

We had data to analyze two generations of family seven (see fig.4.36), with 452 variants after filtering. Eight were shared by the three exome sequenced family members and all eight are nonsynonymous variants (see tab.4.32).

We validated these in eleven family members with no perfect co-segregation for any of the variants (see tab.4.33). For the variants in *C1orf185*, *CDC42BPA*, *TARBP1*, *PRTG*, and *TLCD2*, there were no affected individuals without variants. However, there were between 1 and 5 healthy individuals who carried the mutation (*CDC42BPA* and *TARBP1*).
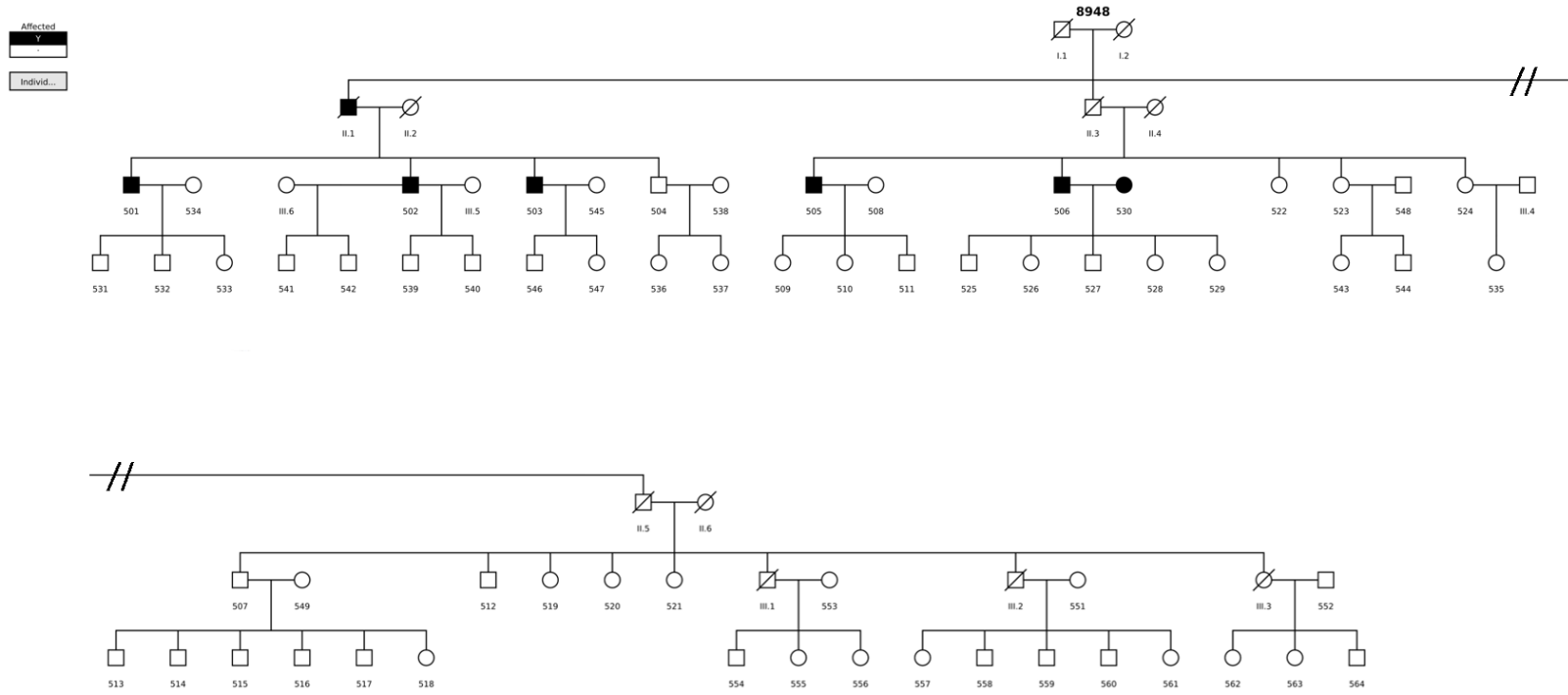
Figure 4.36.: This figure shows the pedigree of family 8948. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.32.: Variants shared by all three exome sequenced individuals in family 7. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr1	51613245	C	A	exonic	C1orf185	nonsynonymous SNV	D	P	N	12.21
chr1	227203825	T	C	exonic	CDC42BPA	nonsynonymous SNV	T	D	D	14.68
chr1	234595070	T	C	exonic	TARBP1	nonsynonymous SNV	D	B	D	11.32
chr8	104331011	A	G	exonic	FZD6	nonsynonymous SNV	D	B	D	14.46
chr8	110457620	G	A	exonic	PKHD1L1	nonsynonymous SNV	D	D	D	25
chr15	55964778	C	G	exonic	PRTG	nonsynonymous SNV	T	B	D	15.36
chr17	1611329	C	T	exonic	TLCD2	nonsynonymous SNV	D	0	D	11.77
chr17	49371294	A	G	exonic	UTP18	nonsynonymous SNV	D	D	D	20.8

Table 4.33.: Results of the variant validation in family 7. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	C1orf185	CDC42BPA	TARBP1	FZD6	PKHD1L1	PRTG	TLCD2	UTP18
8948501*	51	61	2	2	2	2	2	2	2	2
8948502*	47	67	2	2	2	2	2	2	2	2
8948503	63	73	2	2	2	1	1	2	2	2
8948504	-	70	2	1	1	1	1	1	2	2
8948505*	67	67	2	2	2	2	2	2	2	2
8948506	63	74	2	2	2	2	2	2	2	1
8948507	-	78	1	1	1	1	1	2	1	1
8948512	-	71	2	1	1	1	2	2	2	2
8948522	-	60	1	2	2	1	2	2	2	2
8948523	-	68	2	1	1	2	2	2	2	1
8948524	-	65	2	1	1	1	1	2	2	2

With a standard deviation of 7 and a mean age of disease at 58, the age cut-off was set to 65 in this family. All odds ratios were greater than one (see fig.4.37 and fig.4.38) , with the odds ratios being significant for the chr1:227203825T>C variant in *CDC42BPA* and the chr1:234595070T>C variant in *TARBP1*. Both have the same inheritance pattern and have therefore the same odds ratio and the same LOD score. The unadjusted odds ratio is reaching 40.3 and 81 when adjusted for age.
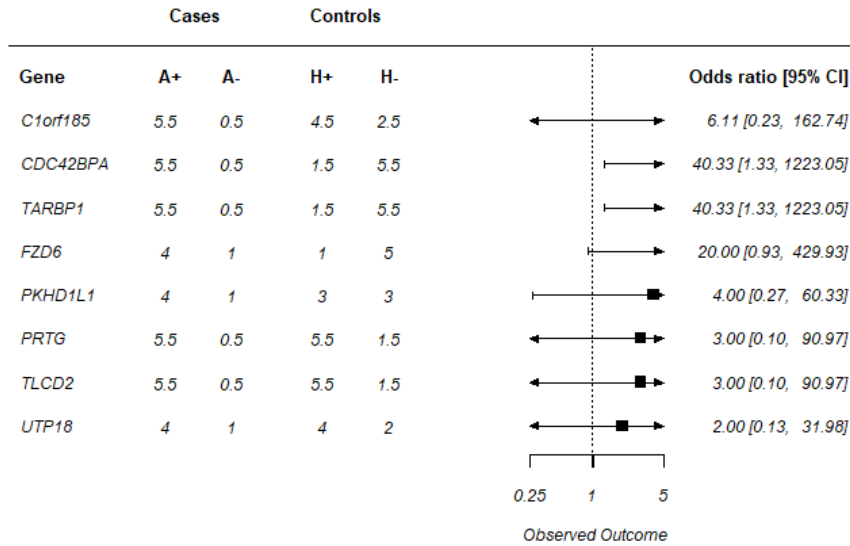
Figure 4.37.: Forest plot of genes in family 7. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.
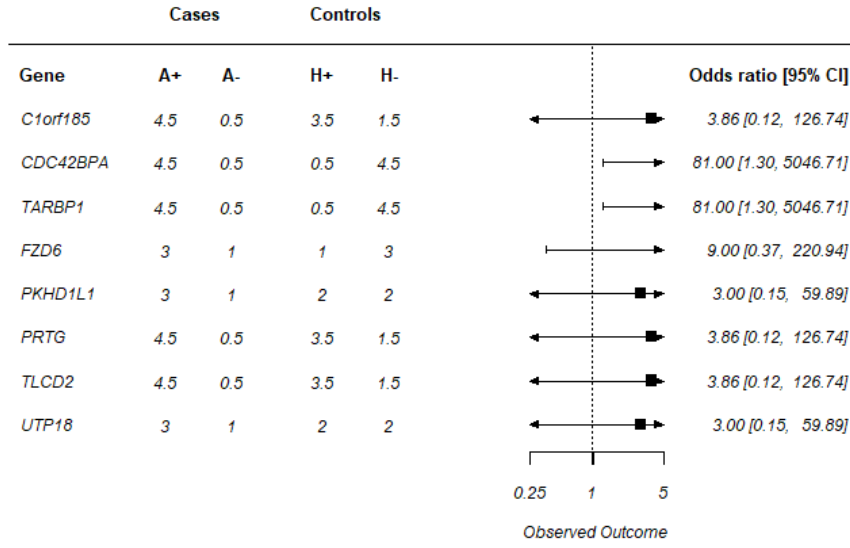
Figure 4.38.: Age-adjusted forest plot of genes in family 7. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

The highest LOD score was 1.86 at 0.09, when not adjusted for age. After adjustment the LOD score was increased to 2.37 with an r of 0.01 (see fig 4.39).
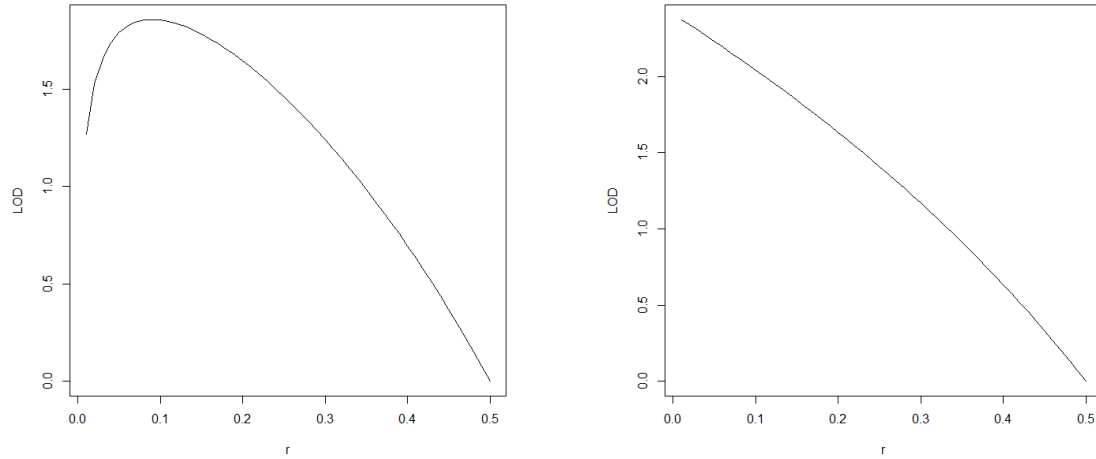
Figure 4.39.: Plot of the LOD score of the unadjusted (left) and age-adjusted (right) of the *CDC42BPA* and *TARBP1* variant.

Sixty-three percent of the validated family members smoked and 36 % had hypertension. Two (18%) were obese and one (9%) had hypercholesterolemia or diabetes.

Table 4.34.: Phenotypic data for all family members in family 7. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
8948501	63	1	28.68	51	157			2	2	1	1
8948502	68	1	33.9	47	130			2	1	2	1
8948503	64	1	27.68	63	159.1	Atorvastatin	20	1	1	2	2
8948504	66	1	29.07		177			2	2	2	1
8948505	67	1	26.96	67	95			2	2	2	1
8948506	64	1	29.98	63	150			2	2	2	1
8948507	78	1	31.02		153			2	1	2	1
8948512	65	1	29.98		137			2	2	2	1
8948522	52	2	25.26		130			2	2	2	2
8948523	68	2	24.61		158			2	2	2	2
8948524	58	2	29.38		159			2	1	2	2

4.2.10. Family 8 - 6815

Family 8 was the smallest of our ten families with two generations for analysis (see fig.4.40). Possibly due to size and with that the relative genetic similarity, we found the highest amount of variants in this family (508), with twenty variants being shared by the exome sequenced individuals (see tab.4.35). Seventeen variants were nonsynonymous mutations, one was a splice site change, one a non frameshift insertion and one a non frameshift deletion.

We validated the 20 variants in seven available family members. None of the variants co-segregated with disease (see tab.4.36). Half of the variants had one affected individual that carried the mutation and only one variant (*METTL21B*) was not found in unaffected relatives. The other variants were carried by one-to-three healthy individuals with the mutation.
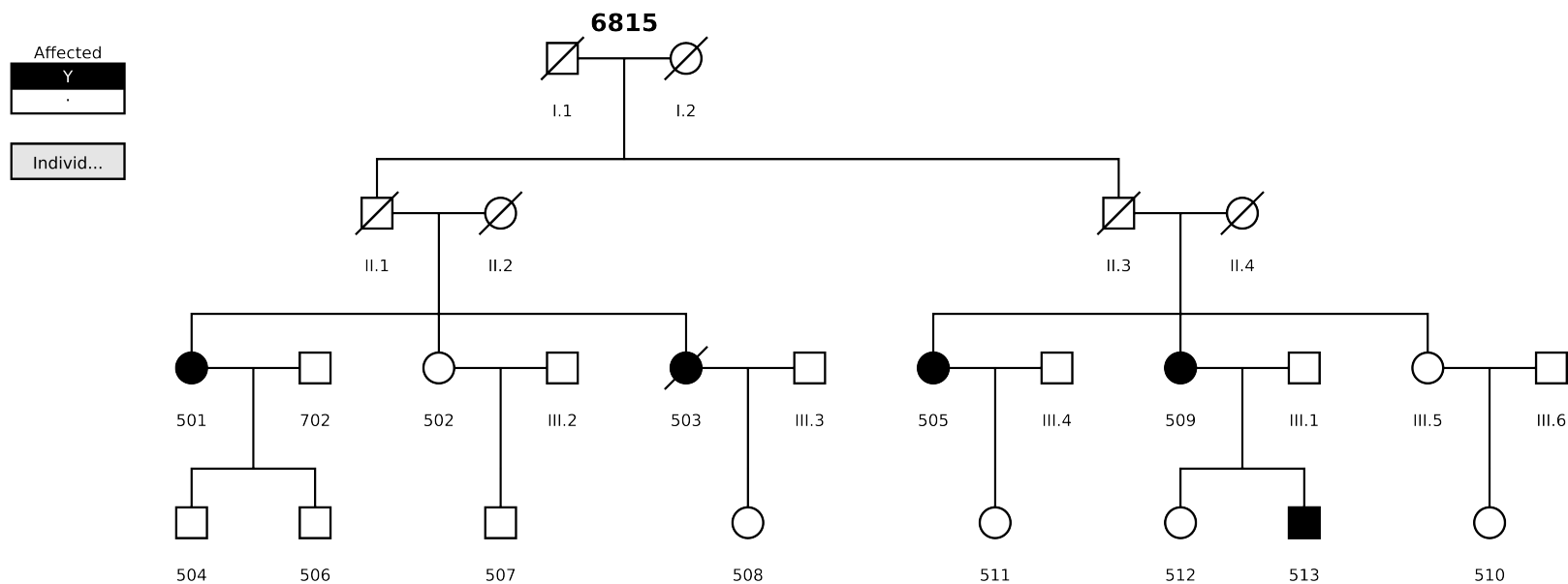
Figure 4.40.: This figure shows the pedigree of family 6815. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.35.: Variants shared by all three exome sequenced individuals in family 8. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr2	135887597	C	T	exonic	RAB3GAP1	nonsynonymous SNV	T	B	D	13.59
chr5	13891219	A	T	exonic	DNAH5	nonsynonymous SNV	T	B	D	12.99
chr5	64013795	T	A	exonic	FAM159B	nonsynonymous SNV	T	0	N	15.75
chr6	37605158	C	T	exonic	MDGA1	nonsynonymous SNV	D	B	N	17.24
chr6	170871046	-	CAG	exonic	TBP	nonframeshift insertion	NA	NA	NA	NA
chr9	32405589	G	T	exonic	ACO1	nonsynonymous SNV	D	B	D	12.2
chr9	32481339	C	T	exonic	DDX58	nonsynonymous SNV	T	D	D	24.3
chr9	72755178	G	A	exonic	MAMDC2	nonsynonymous SNV	T	P	D	21.9
chr11	85438948	G	A	exonic	SYTL2	nonsynonymous SNV	0	P	N	16.72
chr11	111608216	T	A	exonic	PPP2R1B	nonsynonymous SNV	D	P	D	10.82
chr11	121037422	T	C	exonic	TECTA	nonsynonymous SNV	T	B	D	12.52
chr12	58174275	T	C	exonic	METTL21B	nonsynonymous SNV	0	P	D	15.11
chr14	102481630	A	C	exonic	DYNC1H1	nonsynonymous SNV	T	B	D	11.84
chr18	60646248	CAG	-	exonic	PHLPP1	nonframeshift deletion	NA	NA	NA	NA
chr19	3013678	C	T	exonic	TLE2	nonsynonymous SNV	T	B	D	15.56
chr19	3733883	A	T	exonic	TJP3	nonsynonymous SNV	D	D	N	14.18
chr20	44594306	G	A	exonic	ZNF335	nonsynonymous SNV	D	D	D	16.46
chr21	26972181	G	A	splicing	MRPL39	NA	NA	NA	NA	NA
chr22	29446901	C	T	exonic	ZNRF3	nonsynonymous SNV	D	P	D	19.15
chr22	30857329	T	A	exonic	SEC14L3	nonsynonymous SNV	T	B	D	24.2

Table 4.36.: Results of the variant validation in family 8. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	6815501*	6815503*	6815505	6815509*	6815502	6815504	6815506
CAD age	56	57	60	73	-	-	-
LAA	70	67	66	75	65	41	47
RAB3GAP1	2	2	2	2	2	2	1
DNAH5	2	2	2	2	2	2	2
FAM159B	2	2	2	2	2	2	1
MDGA1	2	2	1	2	2	1	2
TBP	2	2	2	2	2	2	2
ACO1	2	2	1	2	2	2	2
DDX58	2	2	1	2	2	2	2
MAMDC2	2	2	1	2	2	2	2
PPP2R1B	2	2	2	2	2	1	1
SYTL2	2	2	2	2	2	2	2
TECTA	2	2	1	2	2	1	1
METTL21B	2	2	1	2	1	1	1
DYNC1H1	2	2	1	2	2	2	2
PHLPP1	2	2	1	2	2	2	1
TLE2	2	2	2	2	1	2	2
TJP3	2	2	2	2	1	2	2
ZNF335	2	2	1	2	2	2	1
MRPL39	2	2	2	2	1	2	2
ZNRF3	2	2	2	2	2	1	1
SEC14L3	2	2	1	2	2	1	1

The mean age of disease was at 57 with a standard deviation of 7. Due to the low number of individuals in this family, significance could not be achieved for the odds ratios or the LOD score. Additionally, we could not apply any age adjustment, as this would have only further reduced the individuals.

There were four variants with an odds ratio below one and six variants with an odds ratio close to one (see fig:4.41). The highest odds ratio was 16 for the variant in *METTL21B*, followed by 15 for the variants in *ZNRF3* and *PPP2R1B*.
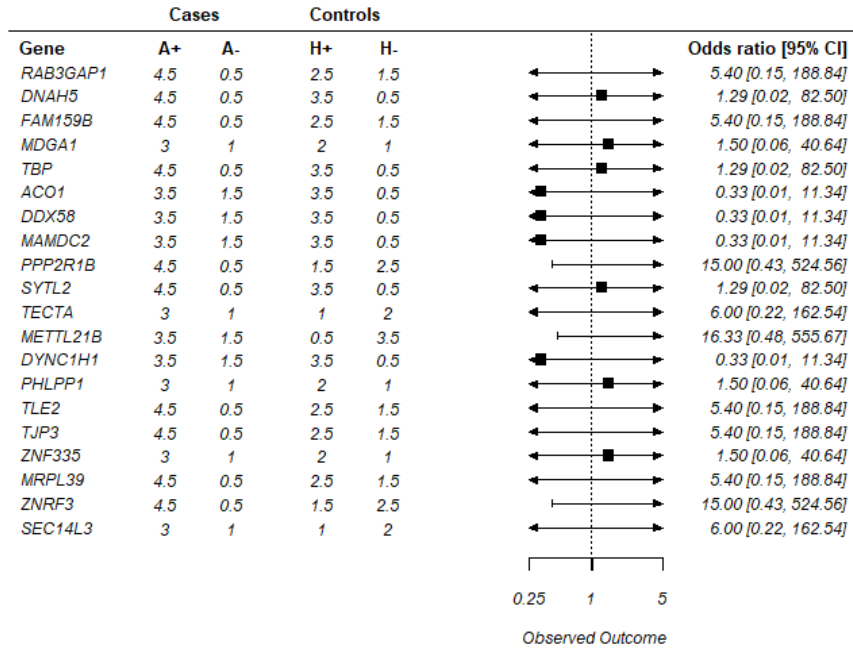
Figure 4.41.: Forest plots of genes in family 8. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

Only one of the validated family members was obese or diabetic and two had hypercholesterolemia (see tab.4.37). Over half of them smoked and 42% suffered from hypertension.

Table 4.37.: Phenotypic data for all family members in family 8. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
6815501	60	2	25.39	56	120	Simvastatin	10	1	2	2	2
6815502	65	2	29.97		182			2	2	1	1
6815503	67	2	23.42	57	186			2	1	1	1
6815504	33	1	23.15					0	2	2	1
6815505	66	2	30.47	60	118	Simvastatin	10	1	1	2	1
6815506	40	1	23.78					0	0	2	1
6815509	72	2	25.24	73				0	0	2	2

4.2.11. Family 9 - 9214

Three generations were available for analysis in family 9 (see fig.**??**). Filtering left us with 417 variants of which only a nonsynonymous variant in *DOCK6* was shared by all three family members. The chr19:11319636G>A variant was predicted to be benign by SIFT and PolyPhen2, labelled as possibly damaging by MutationTaster and had a CADD score of 11.32.

We validated the variant in eleven family members (see tab.4.38). Three of the affected individuals carried the variant, and two did not. Of the six healthy family members, one was also a variant carrier. Also after adjusting for age with a mean disease age of 54, a standard deviation of 8 and a calculated cut-off of 62, there was still no perfect co-segregation.

Table 4.38.: Results of the variant validation in family 9. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	DOCK6
9214501	-	50	1
9214502	-	54	1
9214503*	53	58	2
9214504	-	55	1
9214506	48	61	1
9214507	-	62	1
9214510	67	75	1
9214512*	42	63	2
9214516*	62	82	2
9214525	-	69	1
9214526	-	83	2
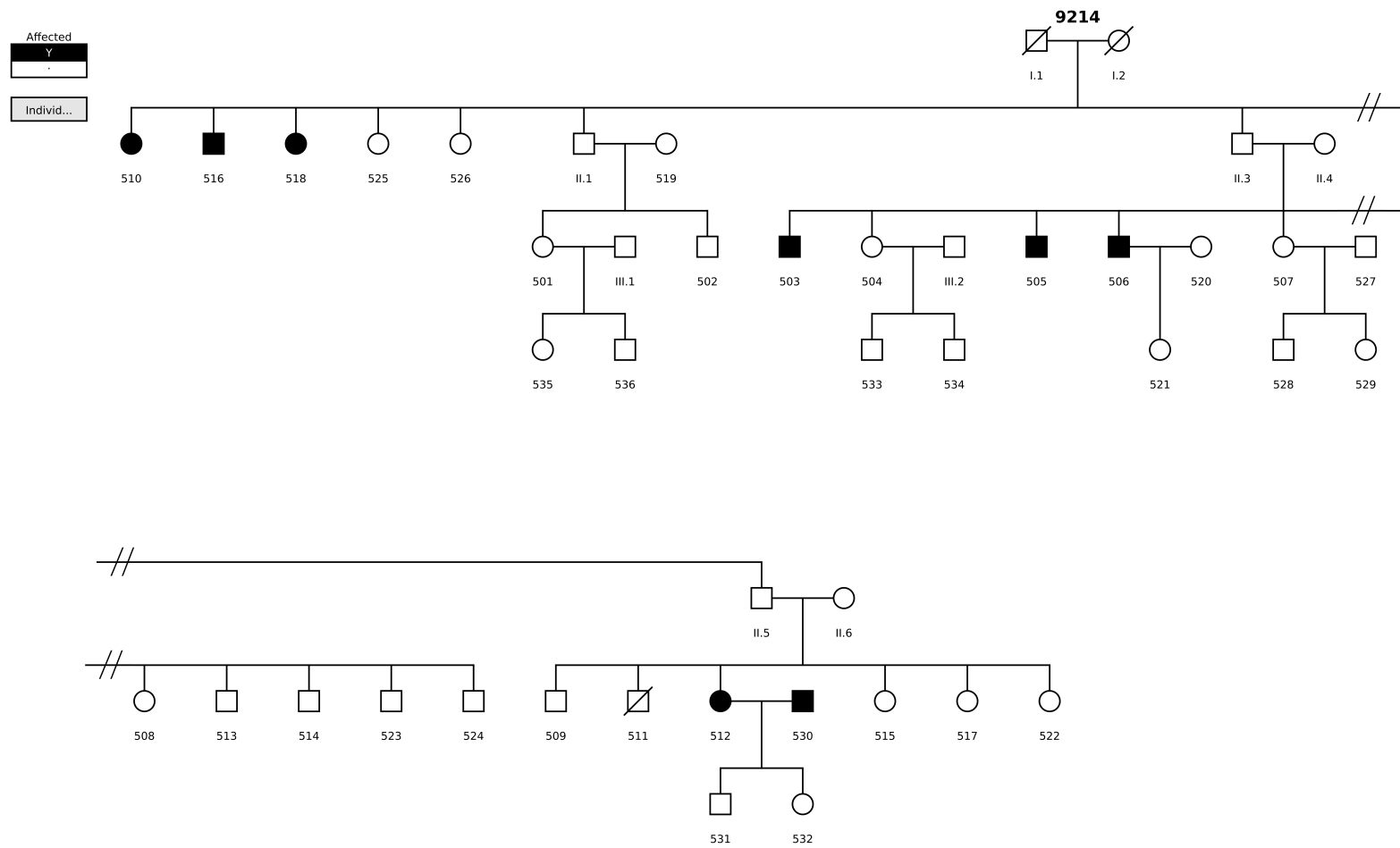
Figure 4.42.: This figure shows the pedigree of family 9214. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Both the adjusted and unadjusted odds ratios were not significant, with the unadjusted on of 7.5 decreasing to three when adjusted for age (see tab.4.39).

Table 4.39.: Calculation of the odds ratio and its confidence interval for the variant in family 9. A stands for affected, H for healthy and RF for risk factor, meaning they carry the variant.

Variant	A+RF	A-RF	H+RF	H-RF	odds ratio	95% Confidence Interval	
DOCK6	3	2	1	5	7.5	0.46	122.70

Ninety percent of the validated individuals smoked. Half of them had hypercholesterolemia, 27% had hypertension and only one (9%) was obese or had diabetes (see tab.4.40).

Table 4.40.: Phenotypic data for all family members in family 9. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
9214501	42	2	22.20		132			2	2	2	1
9214502	45	1	31.24		118			2	1	2	1
9214503	49	1	28.41	53		Atorvastatin	20	0	2	1	1
9214504	46	2	21.87		61	Atorvastatin	10	1	2	2	1
9214506	53	1	26.78	48	124	Simvastatin	20	1	2	2	1
9214507	60	2	23.05		278			1	2	2	1
9214510	72	2	19.81	67	129	Atorvastatin	10	1	2	2	1
9214512	NA	2	25.26	42		Simvastatin	40	0	2	2	2
9214516	73	1	21.20	62	147			2	1	2	1
9214525	67	2	22.86		202			1	0	2	1
9214526	74	2	23.24		198			1	1	3	1

4.2.12. Family 10 - 8662

We analyzed two generations in family 10 (see fig.4.43), with 435 variant found through exome sequencing. When filtering for those that are shared by all three family members, 14 nonsynonymous variants remain (see tab.4.41).
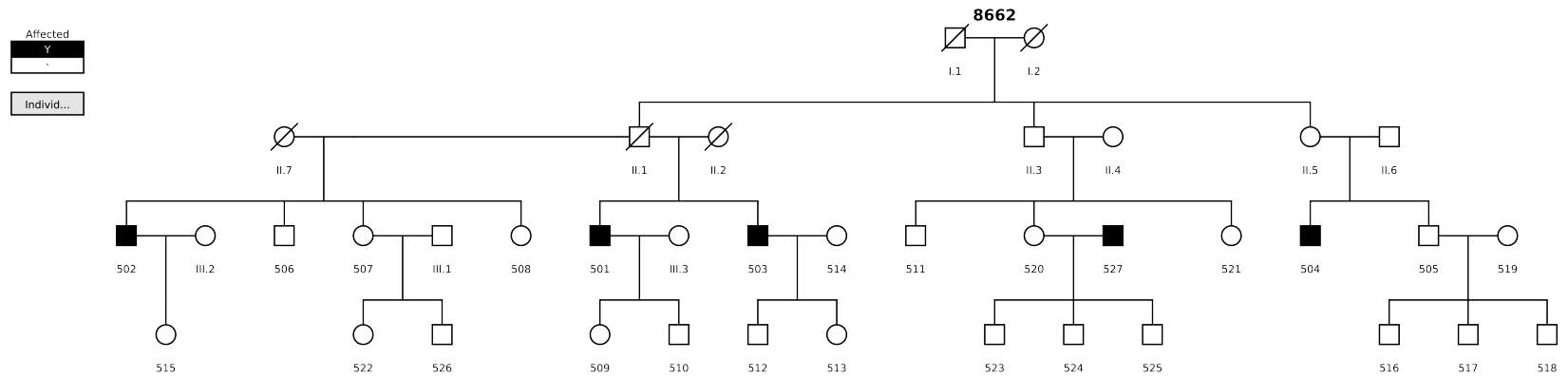
Figure 4.43.: This figure shows the pedigree of family 8662. Females are represented by a circle, males by a square. Open symbols (white) represent unaffected individuals, closed symbols (black) indicates the individual was affected by CAD. An individual with a slash is deceased. The specific individual IDs are shown under each individual. We had no-to-limited information available for individuals with a roman numeral.

Table 4.41.: Variants shared by all three exome sequenced individuals in family 10. The variants were annotated to the hg19/GrCh37 build.

Chr	Start	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	SIFT	Polyphen2	MutationTaster	CADD
chr1	208219387	C	A	exonic	PLXNA2	nonsynonymous SNV	D	D	D	22,9
chr1	231557077	C	G	exonic	EGLN1	nonsynonymous SNV	T	B	D	15,35
chr8	144809017	T	C	exonic	FAM83H	nonsynonymous SNV	T	B	N	7,37
chr9	304628	G	A	exonic	DOCK8	nonsynonymous SNV	T	D	D	15,1
chr9	5534924	C	G	exonic	PDCD1LG2	nonsynonymous SNV	T	P	N	13
chr9	8486248	G	A	exonic	PTPRD	nonsynonymous SNV	0	D	D	23,4
chr9	88356862	C	T	exonic	AGTPBP1	nonsynonymous SNV	0	B	D	13,41
chr16	23444928	G	A	exonic	COG7	nonsynonymous SNV	T	P	D	19,02
chr16	23569403	G	A	exonic	UBFD1	nonsynonymous SNV	T	B	N	3405
chr16	30732558	C	A	exonic	SRCAP	nonsynonymous SNV	T	B	D	11,98
chr18	45566811	C	T	exonic	ZBTB7C	nonsynonymous SNV	T	P	D	16,05
chr18	46476680	C	T	exonic	SMAD7	nonsynonymous SNV	T	B	D	2594
chr19	12155456	G	A	exonic	ZNF878	nonsynonymous SNV	D	P	D	15,41
chr22	37420611	G	A	exonic	MPST	nonsynonymous SNV	D	P	D	19,34

We validated these variants in 16 family members and two variants show perfect co-segregation, the chr16:30732558C>A variant in *SRCAP* and the chr18:45566811C>T in *ZBTB7C*. Both variants were validated in 4 CAD affected family members and were not found in the 12 healthy individuals. Some of the other variants (*FAM83H, AGTPBP1, COG7, UBFD1, SMAD7, ZNF878* and *MPST*) were also found in all affected individuals, but there have been one-to-four healthy individuals with variant (see tab.4.42).

The median age of disease was at 62 years in this family with a standard deviation of 6. Using the calculated age cut-off of 68, only 6 individuals were left for analysis. This small sample size was to underpowered to result in significant odds ratios or LOD scores (even with perfect co-segregation). We calculated significant odds ratios (without the age cut-off) for 5 variants in this family. The chr9:304628G>A in *DOCK8* had the smallest odds ratio with 33, followed by the variants in *COG7* and *UBFD1*, which had the same inheritance pattern and reached an odds ratio of 37.8. An odds ratio of 225 was been calculated for the *SRCAP* and *ZBTB7C* variant, who also have the same inheritance pattern.

Table 4.42.: Results of the variant validation in family 10. LAA stands for last available age and shows the age of the individual at the last data collection point. One means the individual has the wild type at this position and two is a heterozygous variant genotype. A star indicates the three exome sequenced family members.

Individual ID	CAD age	LAA	PLXNA2	EGLN1	FAM83H	DOCK8	PDCD1LG2	PTPRD	AGTPBP1	COG7	UBFD1	SRCAP	ZBTB7C	SMAD7	ZNF878	MPST
8662501	57	62	1	1	2	1	1	1	2	2	2	2	2	2	2	2
8662502*	61	69	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8662503*	58	69	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8662504*	63	74	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8662505	-	69	2	2	2	1	2	2	2	1	1	1	1	1	2	2
8662507	-	84	2	2	1	1	1	1	2	2	2	1	1	1	1	1
8662509	-	47		1	2	1	1		1	1	1	1	1	1		
8662511	-	61	1	1	2	1	1	1	1	1	1	1	1	2	1	1
8662513	-	35		2	1	2	1		1	1	1	1	1	1		
8662516	-	40		1	1	1	2		2	1	1	1	1	1		
8662520	-	68	1	1	1	1	1	1	1	1	1	1	1	2	1	1
8662521	-	65	1	1	2	1	1	1	1	1	1	1	1	2	1	1
8662522	-	53		1	1	1	1		2	2	2	1	1	1		
8662523	-	45		1	1	1	1		1	1	1	1	1	1		
8662524	-	37		1	1	1	1		1	1	1	1	1	1		
8662525	-	34		1	1	1	1		1	1	1	1	1	2		
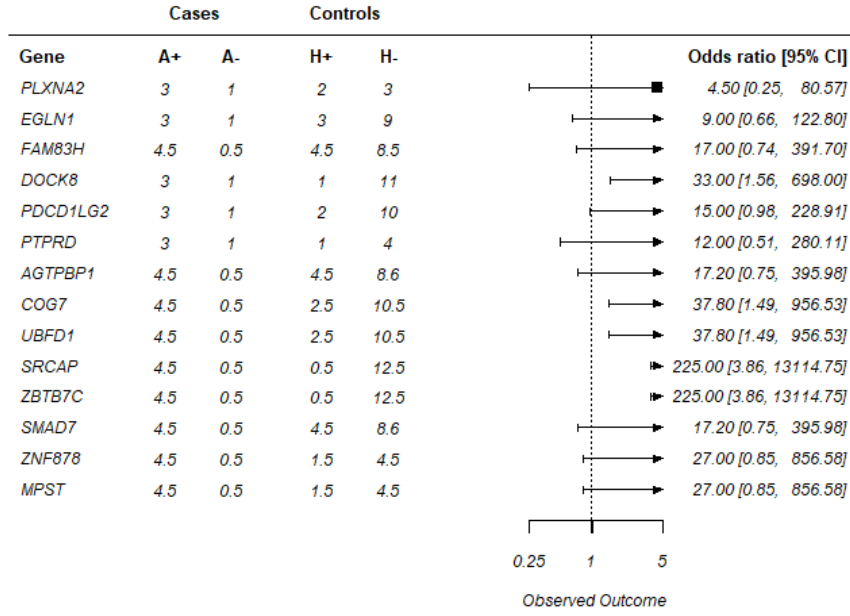
Figure 4.44.: Forest plot of genes in family 10. A+ shows number of cases with the variant in the gene and A- without. H+ shows the number of controls with the variant in the gene and H- without.

Although the inheritance was slightly different between the *DOCK8* variant and the variants in *AGTPBP1* and *COG7*, they all had the same amount of recombinant and non recombinant individuals resulting in the same LOD score. For unadjusted measurement the LOD score was 2.2 at an r of 0.13 (see fig 4.45).
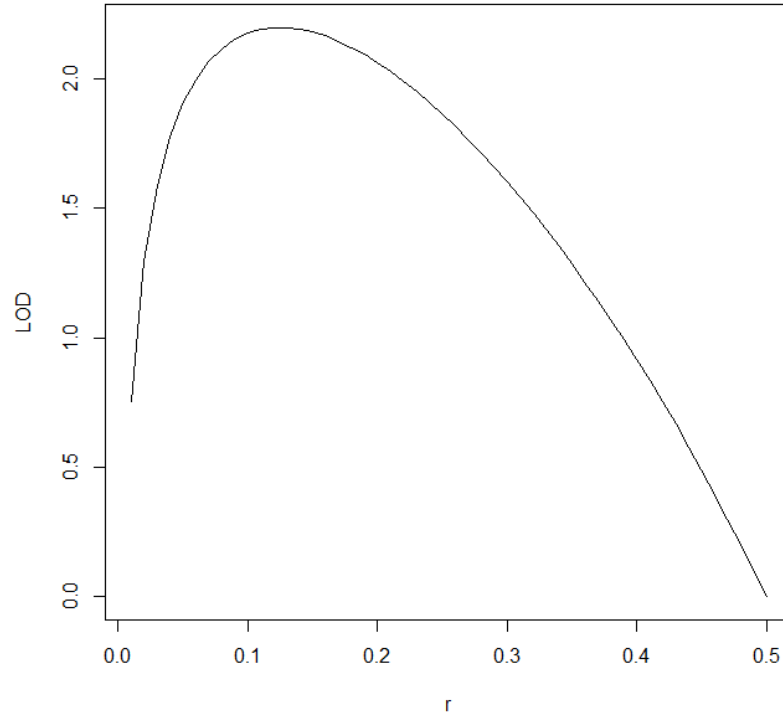
Figure 4.45.: Plot of the LOD score of the *DOCK8* variant and the variants in *AGTPBP1* and *COG7*.

The *SRCAP* and *ZBTB7C* had the highest LOD score, which was significant at 4.75 with an r of 0.01 (see fig 4.46).
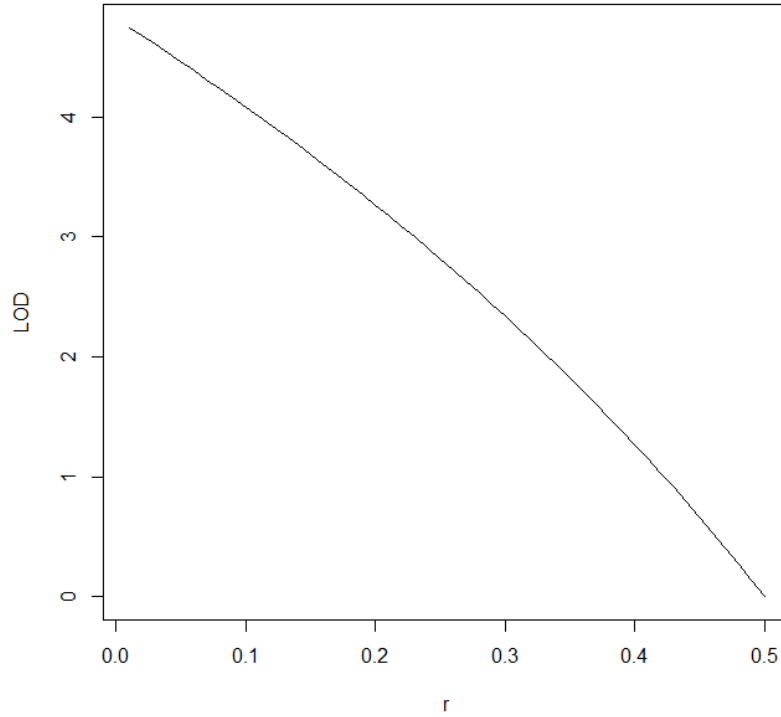
Figure 4.46.: Plot of the LOD score of the *SRCAP* and *ZBTB7C* variant.

Overall this family showed very little comorbidities. None of the validated family members was obese or had diabetes. Two suffered from hypercholesterolemia and four (36%) had hypertension and/or were smoking (see tab.4.43).

Table 4.43.: Phenotypic data for all family members in family 10. Sex: 1 - Male, 2 - Female. Hypercholesterolemia (HC), hypertension (HT), diabetes (DB) and smoking (S) : 1 - yes, 2 - no, 0 - no data

ID	Age at inclusion [years]	Sex	BMI [kg/m²]	age CAD [years]	LDL C [mg/dL]	statin	Daily dose	HC	HT	DB	S
8662501	61	1	23.62	57	137			2	1	2	2
8662502	71	1	23.15	61	166			2	1	2	1
8662503	60	1	25.16	58	155	Simvastatin	10	1	2	2	2
8662504	64	1	22.96	63	114	Simvastatin	10	1	1	2	2
8662506	100	1						0	0	0	0
8662507	76	2	21.48		186			2	1	2	2
8662510	33	1	24.21		110			2	2	2	1
8662512	34	1	27.68					0	0	2	1
8662516	30	1	24.30		136			2	2	2	2
8662525	24	1	19.37		119			2	2	2	1
8662526	100	1						0	0	0	0

4.2.13. Candidate genes

To further evaluate the candidate genes, we firstly looked up the genes in ExAC and evaluated their constraint metrics. These metrics were calculated by the ExAC database as Z scores for synonymous variants and a pLI scores for loss-of-function (LoF) variants and they indicate an intolerance to this sort of variation in a particular gene.

4.2.13.1. Constraint metrics:

A positive Z score stands for a higher intolerance to variations and is a measure for the difference between observed and expected variance frequency in the gene. Positive values indicate a high tolerance and hence variants in this gene less likely to be deleterious. Ten of our candidate genes had a positive Z score (see tab.4.44) and only four were negative (*MMP10*, *PTPRZ1*, *TARBP1* and *DOCK8*). The scores for *CDC42BPA* (1.75), *ZBTB7C* (1.8), *SRCAP* (2.23), *BRINP1* (2.13) and *ADAR* (3.01) had values above one.

The idea of the pLI score is the same as the Z score with the difference being that the pLI score looks at LOF variants only. The closer it is to one, the more intolerant to LOF variants a gene seems to be. All values over or equal to 0.9 are considered extremely intolerant. This is reached by seven variants (*ADAR*, *PTPRZ1* and *ZBTB7C*) with four even reaching one (*SRCAP*, *CDC42BPA*, *BRINP1* and *GTF2H1*). All others with the exception of *REN* (0.01) and *UBFD* (0.84) have been calculated to be zero.

Table 4.44.: Z- and pLI- scores for all candidate genes taken from ExAC.

Gene	Z score	pLI score
MMP10	-2.43	0
REN	0.5	0.01
GTF2H1	0.59	1
ADAR	3.01	0.91
PTPRZ1	-1.11	0.99
BRINP1	2.13	1
SLC6A5	0.3	0
CDC42BPA	1.75	1
TARBP1	-0.66	0
DOCK8	-2.78	0
COG7	0.96	0
UBFD1	2	0.84
SRCAP	2.23	1
ZBTB7C	1.8	0.91

Second, we analyzed the prevalence of filtered variants in our database of 400 controls and 655 exome sequenced cases, to determine whether there was a higher frequency of variants in the candidate genes in the cases compared to the controls. A list of all identified variants can be found in the supplement (tab.A.4).

4.2.13.2. Variant prevalence in cases and controls:

Table 4.45.: Overview of results of variant prevalence in 655 cases and 400 controls.

Gene	Total amount of variants	Amount cases	Amount controls
MMP10	11	19	17
REN	1	1	0
GTF2H1	5	13	8
ADAR	17	15	8
PTPRZ1	43	37	31
BRINP1	15	16	11
SLC6A5	16	34	20
CDC42BPA	23	22	5
TARBP1	40	46	17
DOCK8	62	70	52
COG7	17	14	8
UBFD1	4	3	1
SRCAP	55	71	51
ZBTB7C	14	13	13

We found eleven variants in the *MMP10* gene in our database of 400 controls and 655 exome sequenced cases: eight nonsynonymous, one stop-gain, one frameshift insertion and one splice variant. These variants were found in 19 cases and 17 controls (see tab. 4.45). There was only one variant found in *REN*, a nonsynonymous variant, which was found in one case. Five nonsynonymous variants were identified for *GTF2H1*, with 13 cases and eight controls carrying one. In the *ADAR* gene, we found 17 nonsynonymous variants in 15 cases and eight controls. A lot of the variants were found in *PTPRZ1*. We identified 39 nonsynonymous variants, one splicing and three non-frameshift deletions. In total we found 37 cases and 31 controls carrying a variant in *PTPRZ1*. We identified 15 nonsynonymous variants in 16 cases and 11 controls in *BRINP1*, and 13 nonsynonymous, two stop-gain and one frameshift insertion in 34 cases and 20 controls in *SLC6A5*. Twenty-three variants were found in *CDC42BPA*. Twenty were nonsynonymous, two stop-gain and one stop-loss. In total 22 cases and five controls carried a variant in *CDC42BPA*. Forty variants were found in *TARBP1*, 33 nonsynonymous, one stop-loss, three stop-gain, two splicing and one frameshift deletion. Forty-six cases and 17 controls carried a variant. We found the most variants in *DOCK8*, with 70 cases and 52 controls carrying at least on of the variants. Fifty-eight of these were nonsynonymous, three were splice-site mutations and one was a stop-gain. For *COG7*, one splice and 16

nonsynonymous variations were found in 14 cases and eight controls. We only found three cases and one control that carried one of four nonsynonymous variants in *UBFD1*. There was one frameshift insertion, one non-frameshift deletion and 53 nonsynonymous variants in *SRCAP*, which were identified in 71 cases and 51 controls. The twelve non-synonymous, one non-frameshift insertion and one non-frameshift deletion in *ZBTB7C* were found in 13 cases and 13 controls.

We calculated all odds ratios and performed a Fisher's exact test (see fig.4.47). Only *CDC42BPA*had a nominally significant p-value of 0.04316 and a significant odds ratio of 2.74. Most of the other variants had an odds ratio close to one, some were even smaller than one (*MMP10, PTPRZ1, SRCAP* and *ZBTB7C*). Other than *CDC42BPA*, only *UBFD1, TARBP1* and *REN* had odds ratios over one.
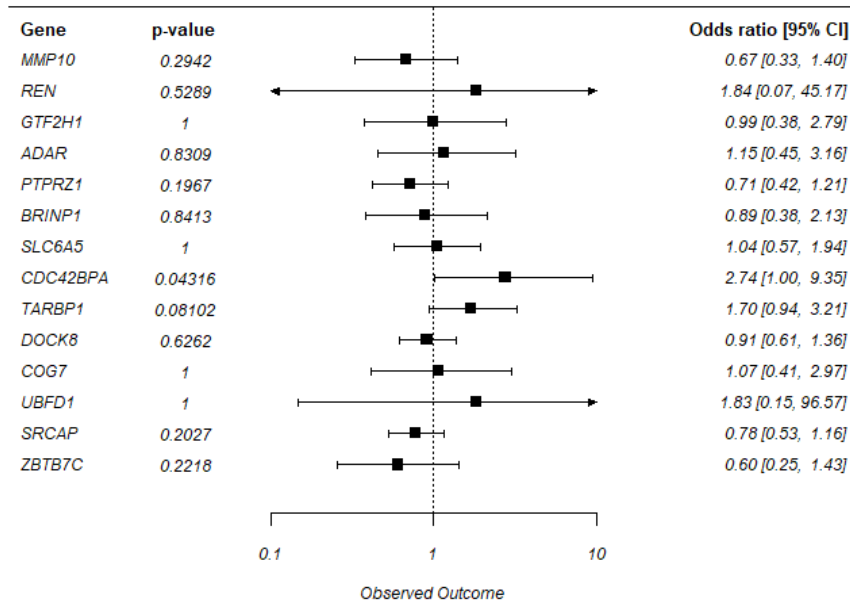
Figure 4.47.: Forest plot of candidate genes including their respective p-values.

5. Discussion

5.1. Screening for known rare causal variants

In the first part of this thesis, we screened for variants reported to cause CAD or CAD related traits. In a previous study, we identified that 5% of CAD patients in our cohort had hypercholesterolemia. As a follow up, in this thesis, we identified a total of 17.25% of these patients carried a variant known to cause CAD or CAD risk factors such as obesity, hypertension, diabetes and dyslipidemia. We only analyzed rare variants and we saw significantly more causal rare variants in our cohort then expected, taking into account that we were looking at variants with a frequency under 5% in the general population.

To compare observed and expected frequencies, we analyzed variants taken from two databases for this. The variants from HGMD$^{\circledR}$ were found in 10.98% of our patients and the variants from ClinVar were found in 7.84% of the patients, with minimal overlap between these two databases. A reason for this could be the different way the data is collected. For one, ClinVar collects their data from submissions made by researchers about variants and their effect on a phenotype, while HGMD$^{\circledR}$ collects the data from publications. We accessed ClinVar in 2017, while our HGMD$^{\circledR}$ access was only valid until 2014. Since then, variants may have been removed and added.

We found more variants in our CAD cohort compared to 1000G and ExAC which we were not surprised to see. However, for the individual databases, we saw more cholesterol variants in the controls than HGMD$^{\circledR}$ and more obesity- and triglyceride-increasing variants compared to ClinVar. It has to be noted that although we treat the individuals in these databases as controls, it is to be expected that some are affected. ExAC is a consortium of many databases and hence includes cases as well as controls. However, ExAC does not provide individual phenotypes. All individuals included in 1000G declared themselves healthy and the minimum inclusion age was over 18. Most of our

phenotypes are fairly complex and often have a long progression time. Depending of the age at inclusion into the control databases, for example the young age for 1000G, the individuals might not have developed symptoms yet.

This made us wonder if the variants actually cause the phenotype they are connected to in our cohort and we saw that this is not the case for all of them. Generally speaking, the variants causing high cholesterol, high LDL cholesterol, triglycerides and hypertension often showed the expected phenotype. Although only two of the six carriers of a ClinVar variant causing high total cholesterol, actually had levels over the set threshold, the other patients were fairly close to this threshold and statin therapy was in place for all of them.

In contrast, of individuals with obesity variants, only one (of eight) obesity patients had a BMI above 30 kg/m^2. Also for carriers of diabetes (type 2) variants, none were diagnosed with the disease. On one hand there could be errors in the databases and the variant does not actually cause the specific phenotype. On the other hand the reason for the missing phenotype could lie in the diseases themselves. Both obesity and type 2 diabetes are complex diseases with a myriad of influencing factors. It is possible that only a combination of these factors (i.e. the genetic variant and environmental factors together) cause the actual disease. It might be that these variants only increase the risk of obesity/diabetes but a healthy lifestyle and diet could counteract the variant. The measurement for obesity is the BMI, which by itself can already be misleading. As it only includes the height and the weight, it does not incorporate factors such as the actual percentage of fat and muscle. Some very musculous individuals can reach a high BMI without being obese. The BMI does not necessarily differentiate between metabolically healthy and unhealthy, as some obese patients can actually be healthy, while some individuals with a normal weight show obese lipid levels [31].

Some of the phenotypes are overlapping. Over half of the obese individuals suffered from high triglyceride levels, hypertension and roughly 40% of them had high total and high LDL cholesterol levels. As obesity has many comorbidities, this was to be expected, especially if you look at phenotypes like the metabolic syndrome, which is known to not only cause obesity but also high triglyceride- and cholesterol- levels and hypertension [119]. Interestingly, obesity is known to increase the risk for type 2 diabetes and only 22% of the obese individuals in this study were diabetic. This could again be a complex interplay of several factors.

As the amount of total cholesterol is calculated from the amount of triglycerides, LDL- and HDL- cholesterol, we expected to see an overlap. Indeed, 72% of the individuals with high total cholesterol also had high LDL cholesterol levels, and over half of them had high levels of triglycerides. Conversely, half of the individuals with high triglyceride levels and over 60% of the hypercholesterolemia patients also had high total cholesterol levels.

Hypertension surprisingly overlapped with triglyceride levels, LDL cholesterol and total cholesterol and only 29% of these individuals were obese. This could be explained with hypertension being an indicator for metabolic syndrome, which in turn would explain the triglyceride levels and the amount of cholesterol [128]. However, metabolic syndrome usually includes obesity, unless the individuals do not show an increased BMI despite being metabolically unhealthy. The diabetic individuals overlapped with all other risk factors. This makes sense as type 2 diabetes is a comorbidity to all of these risk factors and vice versa.

5.2. Family studies

We used exome sequencing to identify the cause of coronary artery disease in ten extended families, severely affected by CAD. We did a gene-set enrichment with the genes found in our families compared to a list of CAD genes obtained from known GWAS loci. The enriched pathways of the known CAD genes were all connected to lipid levels. This is not surprising as the mechanisms behind many GWAS loci had been identified to be part of the lipid metabolism [6]. Surprisingly though, there was no overlap between the significant enriched pathways of the known CAD genes and the genes identified in our families. This could potentially be explained by the approach of GWAS being so fundamentally different to family studies, as GWAS contain common variations, while we were explicitly looking at rare variations. Additionally our data is subject to noise, as not all variants in our list would be causal.

The most prominently affected pathways in our families were mostly connected to the ECM. The ECM plays an important role in structure and function of the heart and the vascular wall, as they contain proteoglycans and fibrous proteins like collagen and elastin. An association of genetic variability in ECM components and CAD is not new, i.e. variants in *COL3A1* have been shown to influence CAD risk [129]. However, there

have been very little associations between genetic variants in the ECM and CAD so far. We were able to identify potential candidate genes for seven of our ten families, with one of these (*MMP10*) falling into the ECM category .

5.2.1. MMP10

A variant in the *MMP10* gene had been found to co-segregate with disease in the first family. There were six healthy individuals that carried the variant, however, when adjusted for age, the variant co-segregated perfectly. The healthy individuals with this variant could possibly be too young (age range from 41 to 53) to have developed the disease. Their age of these individuals ranged between 41 to 53 with the mean age of disease at 52.5 years for this family.

MMP10 codes for the matrix metalloproteinase 10 (also known as stromelysin-2). It is a member of the family of matrix metalloproteinases (MMPs) and is expressed by macrophages and epithelium in various tissues [130, 131]. MMPs do not only play an important role in the breakdown of the extracellular matrix (by cleaving i.e. fibronectin, proteoglycans and collagen), but they are also involved in a myriad of different signaling pathways. These include processing bioactive proteins, such as cell surface receptors and cytokines [132, 133]. MMP10 in particular has been associated with cell adhesion, migration, proliferation, differentiation as well as vascular development and atherogenesis [132, 134].

MMP10 has been linked to pulmonary hypertension, inflammation, carotid plaque stability, increased intima-media thickness and the presence of atherosclerotic plaques [135, 136, 137, 138, 134], making it a plausible candidate for modulating CAD. Hence variants in MMP10 could affect the progression of CAD either via hypertension, plaque stability, or so far unknown mechanisms. We postulate that a variant with an actual affect on CAD would probably be a gain-of-function, as we assume that an increase of proteolytic activity might lead to an unstable plaque. This would also be supported by Zhao Y. et. al., who suggested that it is an over expression of MMP10 that leads to impaired vascular integrity [139]. Additionally, the pLI-score for MMP10 was 0, showing that MMP10 might be very tolerant to LOF variants. This and also our calculated odds ratio of 0.67 when comparing our data of 400 controls and 655 exome sequenced cases indicates that rare variants might generally have a protective effect.

However, MMP10 is expressed as a pre-protein, which gets proteolytically processed to activate the protein. Our variant lies within the pro-peptide and would hence be cleaved off, and should therefore not interfere with the protein activity itself.

The modeling predicted no changes in the protein structure of the MMP10 pro-peptid. We saw a decreased amount of hydrogen bonds in the mutated pro-peptid, potentially leading to a more unstable structure. This could lead to the pro-peptid being cleaved off prematurely or not at all. We might expect a gain-of-function affect by having a prematurely activated protein or we might see a loss of function, due to the protein not being activated. Further functional studies on this specific variant are under way and will hopefully reveal the specific mechanism.

Based on our findings chr11:102650461A>T is likely to be the variant causing coronary artery disease in our extended family. In regards to the other families, we can not make a clear statement about their functionality. For those who do not co-segregate, we do not expect an effect of the variant, and for the one variant that did co-segregate with disease, we can not be sure as the family size was small.

5.2.2. REN and GTF2H1

In family two, we identified two potential candidate genes, *REN* and *GTF2H1*, that could cause the disease in the family. Both showed significant odds ratios, with *REN* having the biggest unadjusted odds ratio of 31.67, and none of them reaching significance in the LOD scores.

GTF2H1 codes for the general transcription factor IIH subunit 1, a subunit of the transcription factor IIH (TFIIH) complex. This complex is involved in RNA transcription initiation and nucleotide excision repair (NER) of damaged DNA [140]. Due to its very general function and very little published knowledge, we can not link this gene to CAD. Additionally, there is a healthy family member with the variant and also several affected family members that did not carry it. The pLI-score of 1 suggests that the gene is highly intolerable to LOF variants, which would make sense as it takes up a very major and general role. Nevertheless, the assumption would be that a LOF variant would lead to a reduction of NER and/or RNA transcription, probably leading to cancer. Indeed variations in subunits of the TFIIH have been connected to xeroderma pigmentosum and other diseases unrelated to CAD [141]. Xeroderma pigmentosum is a disease in which

DNA repair is impaired, leading to extreme intolerance to UV light and the development of skin cancers.

The second candidate gene in the family (*REN*) codes for renin, an important part of the renin-angiotensin system, which regulates the blood pressure, electrolyte balance, inflammation, proliferation and fibrosis [142]. A drop in blood pressure leads to the release of renin from the kidneys, which then converts angiotensinogen from the liver to angiotensin I. The angiotensin-converting enzyme from the lungs acts on antiotensin I to form angiotensin II, which stimulates the release of aldosterone and can act on the vessel itself leading to vasoconstriction. Aldosterone stimulates the reabsorption of salt and water in the kidneys, effectively increasing the bloodpressure [143]. GWAS were not able to establish an association of *REN* variants and hypertension [144], and the positive odds ratio over all cases and controls, as well as the pLI- and z-value are consistent with this. However, there have been limited reports of *REN* variants being associated with hypertension [145], and a substantial portion of patients with hypertension show low or suppressed renin [146]. In our family, six of the seven individuals with the *REN* variant suffer from hypertension and all of them have been affected by CAD.

We believe that the *REN* variant is the most probable candidate in this family, due to the gene's function and the better co-segregation with disease in the family. However, we do not think that this variant is the sole cause of disease in this family and might hence only increase the risk of disease, as we saw several affected family members without the variant.

5.2.3. ADAR

At the beginning, all variants including the *ADAR* variant were non-significant. We then decided to validate the *ADAR* variant due to a newly published paper. In this paper an ADAR deficiency or mutation on the ADAR editase region had been connected to phenotypic modulation of smooth muscle cells (SMC). The phenotypic modulation, when vascular SMCs switch from a contractile phenotype to a proliferative phenotype, plays a critical role in atherosclerosis, as it is said to lead to intima remodeling and the formation of the fibrous cap [127, 147, 148].

The gene codes for the double-stranded RNA-specific adenosine deaminase (DSRAD), an enzyme important for RNA editing by converting adenosine to inosine, regulation of retrotransposons and gene silencing [149]. Other mutations in the *ADAR* gene have

been connected to dyschromatosis symmetrica hereditaria, a dermatosis [150] and many other diseases including atherosclerosis [151].

ADAR has been identified as the main RNA editor in endothelial cells. It can also modulate atherosclerosis via editing of human cathepsins S (CTSS) Alu RNA and CTSS mRNA expression, a cystein protease that has been implicated in the development of atherosclerosis [151, 152]. The overexpression of ADAR leads to increased CTSS mRNA, and both are upregulated in atherosclerotic carotid plaques [152].

This variant does not show a perfect co-segregation and it does not lie in the editase region. We found four individuals that do not carry the variant and have been affected by CAD and five healthy individuals that carry it. These five healthy patients are still very young and are hence excluded when adjusted for age, as they might still develop the disease. Nevertheless, the ADAR variant shows a significant odds ratio of 8 unadjusted and a high but not significant LOD score of 2.59. The z- and pLI-score show extreme intolerance of variants. We postulate that the variant might lead to a gain-of-function and might modulate atherosclerosis over CTSS. All-in-all we think that this variant could contribute to disease in this family but will not be the sole reason for disease.

5.2.4. PTPRZ1 and BRINP1

In family five we identified two possible candidate genes. One variant was found in the *PRPRZ1* gene, which codes for the receptor-type tyrosine-protein phosphatase zeta, a member of the receptor tyrosine phosphatase family. Four products are built from the gene, two transmembrane and two soluble ones [153], which might play a role in myelin formation and differentiation of oligodendrocytes and Schwann cells [154]. It has been associated with multiple sclerosis and cancer [153, 154, 155] and according to RefSeq, the expression is limited to the central nervous system (CNS). Due to these factors a connection to CAD seems unlikely.

BRINP1 is mainly expressed in the brain and codes for the BMP/retinoic acid-inducible neural-specific protein 1. There is not much information available in regards to the gene's exact function, but BRINP1-/- mice exihibt autism spectrum disorder phenotypes [156]. The gene appears in a report about atherosclerotic risk in African Americans [157], however it showed a odds ratio <1, further indicating that this gene might not have an effect on CAD.

Both genes seem to be highly intolerant for LOF mutations (pLI-score of 0.99 and 1) and BRINP1 seems to also be very intolerant to SNPs (z-score of 2.13). In our family, both genes had odds ratios <1 in our case/control data and significant odds ratios in the family when not adjusted for age. We see affected individuals without the variant and healthy individuals that carry it. When adjusted for age, there is no healthy family member with a *BRINP1* variant.

Biologically, both genes do not seem to be connected to CAD as they are expressed in the brain and seem to effect neural diseases. Hence we think it is very unlikely that the variants have an effect in our family but due to their statistics and the fact that, when adjusted for age, there is no healthy family member with the BRINP1 variant, we can not completely rule out that the variant/s have an effect on this particular family.

5.2.5. SLC6A5

A variant in the *SLC6A5* gene was identified in family six, and resulted in significant odds ratio in age unadjusted and adjusted calculations (12 and 17.5). It codes for the sodium- and chloride-dependent glycine transporter 2 (GlyT2), which plays a major role in extracellular glycine clearance. Mutations in the gene are known to cause human startle disease [158, 159], a disease with an exaggerated startle response, hypertonia and neonatal apnea episodes. According to RefSeq, it is found in glycinergic axons. In our family, we do not see significant LOD scores for the variant and we see healthy variant carriers as well as CAD affected non-carriers. Additionally the *SLC6A5* reaches an odds ratio of 1 in our case/control data and seems to be tolerating mutations, hence we do not think that this variant is the cause of disease in our family.

5.2.6. CDC42BPA and TARBP1

The *CDC42BPA* gene codes for the serine/threonine-protein kinase MRCK alpha, an enzyme that is abundant in heart, brain, kidney, pancreas and skeletal muscle [160]. It binds CDC42 and is important for the cytoskeleton reorganization, regulation and cell migration [161, 162, 163]. This could potentially play a role in cardiac disease, as it is well known that alterations in the cardiac cytoskeleton (i.e. mutations in actin, an important cytoskeleton protein), can cause cardiomyopathies [160]. *CDC42BPA* has also been connected to cancer [164].

There was little information available regarding the *TARBP1* gene. It is known to be ubiquitously expressed and codes for the TAR (HIV-1) RNA binding protein 1. The protein has been found to be upregulated in cancer [165, 166] and it can bind TAR RNA an important component of HIV gene expression [167], where it is believed to disengage the RNA polymerase from the RNA.

CDC42BPA was the only candidate gene that reached significant odds ratios in our cases and controls with an odds ratio of 2.74. We also see a high mutation intolerance in the gene, which is not the case for TARBP1. In our family, both variants show the same inheritance pattern and might therefore be inherited together. Both are found on chromosome 1, however there is are more than 7,000,000 bp between them. There are no affected individual carriers without the mutations but there is one healthy individual that carries them. However, this individual is 60 years old and although the mean age of disease is at 58, there are several family members affected by CAD after the age of 63. Despite the small size (8), the variants reached a LOD score of 2.37 when adjusted for age, which is high considering a minimum of 11 members and a perfect co-segregation is needed to reach the significant value of 3.

All things considered, we believe that these variants might be the cause of disease in this family. The *CDC42BPA* variant is more likely to be the causal variant, due to the gene's function and our data compared to *TARBP1*.

5.2.7. DOCK8, COG7, UBFD1, SRCAP and ZBTB7C

There were several genes with statistical significance in family ten.

DOCK8 codes for the dedicator of cytokinesis 8. It is mostly expressed in bone marrow and the immune system, and there is only very little expression in the heart. *DOCK8* plays an important role in cell migration and immune cell survival, as it controls Treg fitness and their function [168, 169]. Mutations in the gene cause hyper-IgE syndrome, an immunodeficiency disorder with symptoms such as recurrent infections, cancer and autoimmunity [170, 171, 169].

The component of oligomeric golgi complex 7 is encoded by the *COG7* gene and is part of 8 subunits of the conserved oligomeric Golgi (COG) complex. This complex in necessary for golgi morphology and localization. It has been shown that COG-/- human fibroblasts show an irregular golgi protein recycling, [172] and deletions in the

gene cause a rare multisystemic congenital disorder of glycosylation, leading to an early death [173, 174].

There is very limited information about *UBFD1*. It is expressed everywhere and it has been identified as one of the proteins binding polyubiquitine involved in the NF-kappaB signaling pathway [175]. The ubiquitous system plays a role in atherosclerosis, i.e. the ubiquitin ligase A20 modulates NFkB, and has been mapped to a loci for atherosclerosis in mice [176].

The *SRCAP* gene codes for the Snf2 Related CREBBP activator protein, a core catalytic component of the multiprotein chromatin-remodeling SRCAP complex. The complex plays a major role in the repair of double strand breaks in homologous recombination [177]. Additionally, it is recruited to promoters, as it regulates translation via insertion of the histone variant H2A.Z [178]. Mutations in SRCAP cause Floating-Harbour syndrome. This syndrome leads to language deficits, short stature and dysmorphic facial features [179, 180, 181].

We also identified a variant in the *ZBTB7C* gene, which codes for the Zinc finger and BTB domain containing 7C. There is limited information available for this gene. For one, it has been related to cancer, where it promotes increased glutamine uptake which in return facilitates the cancer cell proliferation [182]. On the other hand, it has been associated with increased infarct size in mice [183], as well as the regulation of MMP genes (specifically MMP -8, -10, -13, and -16). As previously discussed, MMPs play an important role in degradation of the ECM as well as other biological pathways and have previously been connected to atherosclerosis, making this gene an interesting candidate gene for this family.

Looking at all candidate genes, *DOCK8* seems to be the most unlikely causal variant, as it is known to cause hyper-IgE syndrome, neither z- nor pLI value are positive and the odds ratio in our case/control data is <1. Additionally, it had the smallest odds ratio (33), and one healthy variant carrier as well as one affected non-carrier. *COG7* and *UBFD1* have the same inheritance pattern, as do *SRCAP* and *ZBTB7C*. All of these genes had positive z-values, however, only *ZBTB7C* also had a significant pLI-value. None of the genes were significant in our case/control data and only UBFD1 had an odds ratio >1, all others showed values below 1. The *ZBTB7C* and *SRCAP* variants show better statistics and co-segregate perfectly with disease in this family. For *COG7* and *UBFD1* we found two healthy individuals with the variant. Additionally, the LOD

score for *ZBTB7C* and *SRCAP* reached significance (4.75). Nevertheless, the LOD score was still high for the other three genes (2.2).

We believe that the *ZBTB7C* variant might be the most plausible candidate gene due to the co-segregation with disease, the statistical significance and the gene's function. However, the *UBFD1* variant is also a potential candidate, so it can not completely be ruled out that both variants have an effect in this family.

5.3. Conclusion

When looking at the general distribution of risk factor phenotypes in our data, we can clearly see an accumulation of these in our cohort. Roughly 40% had hypertension, high triglyceride level and/or hypercholesterolemia, 35% had high cholesterol, 23% were obese and 12% diabetic. In this work, we were able to determine that 17.25% of our cohort carry a rare variant potentially causing one of these risk factors. The other phenotypes could potentially be explained by environmental factors such as the patients activity level and diet, as we do not have data available for these factors. This highlights a potential for genetic screening for known variants to increase the risk of CAD, as well as to reduce the health burden of these risk factors. If the underlying cause had been identified earlier, MI might have been prevented in these cases.

In this work, we have identified variants that potentially explain the disease in five of our ten families. We were not able to identify candidates in the other five families, with several explanations for this. Firstly, some of the variants we ruled out may have an effect on CAD. CAD is a complex disease and as such has been shown to be modulated by a myriad of factors. CAD is caused by common and rare variants, and the risk of the disease is often increased by a combination of variants, so we can expect to see several variants with little effect accumulating and causing disease [6]. This is something we can not verify in our approach, as our approach was based on a dominant mode of inheritance. Furthermore, we are only looking at exome-data, hence, we are missing intronic and intergenic variants. These variants are more likely to be regulatory, influencing expression and splicing and would facilitate a comparison with GWAS data, as GWAS mostly contain non-coding loci. In the future, genome-wide sequencing might become state of the art, although this makes the analysis of all variants very difficult due to the sheer amount of data that needs to be analyzed. Thirdly, we expected all affected

individuals to carry the same variant. It is possible that one of the exome sequenced individuals actually had a different cause of disease (phenocopy). Due to our filtering, we may have missed potential causal variants, but this step was necessary as the load of information would have been too big to analyze otherwise. If we filtered for two individuals carrying the same variant, we would have been left with over 250 variants. It is also possible that some of the genetic variants only work in a gene-environment interaction, where the variant is only effective when coupled with a certain modifier [5]. However, we do not think that this is very likely in this case, as we expect similar environmental factors in families.

We were hoping to identify new rare variants associated with CAD in our family study approach, and indeed we were able to identify several genes and variants that might lead to disease in our families. The most promising variants were chr11:102650461A>T in *MMP10*, chr1:55085648C>T in *REN*, chr1:154574541G>C in *ADAR*, chr1:227203825T>C in *CDC42BPA*, chr16:23569403G>A in *UBFD1* and chr18:45566811C>T in *ZBTB7C*. In general, these genes may harbor several rare variants that could potentially explain some portion of the missing heritability, however, there is still more to unravel. Nevertheless, functional studies are absolutely necessary to determine an actual effect of these variants on CAD. If functional studies confirm the effect, these genes might be potential drug targets and could help to further explain mechanisms behind the development of CAD.

Bibliography

[1] M. Mack and A. Gopal. Epidemiology, traditional and novel risk factors in coronary artery disease. *Cardiol Clin*, 32(3):323–332, Aug 2014.

[2] World Health Organisation. The top 10 causes of death, January 2017. http://www.who.int/mediacentre/factsheets/fs310/en/.

[3] I. Braenne and M. et al. Kleinecke. Systematic analysis of variants related to familial hypercholesterolemia in families with premature myocardial infarction. *Eur. J. Hum. Genet.*, 24(2):191–197, Feb 2016.

[4] J. et al. Versmissen. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. *BMJ*, 337:a2423, Nov 2008.

[5] R. McPherson and A. Tybjaerg-Hansen. Genetics of Coronary Artery Disease. *Circ. Res.*, 118(4):564–578, Feb 2016.

[6] J. Hartiala, W. S. Schwartzman, J. Gabbay, A. Ghazalpour, B. J. Bennett, and H. Allayee. The Genetic Architecture of Coronary Artery Disease: Current Knowledge and Future Opportunities. *Curr Atheroscler Rep*, 19(2):6, Feb 2017.

[7] H. et al. Tada. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *Eur. Heart J.*, 37(6):561–567, Feb 2016.

[8] R. C. Davis, F. D. Hobbs, and G. Y. Lip. ABC of heart failure. History and epidemiology. *BMJ*, 320(7226):39–42, Jan 2000.

[9] P. R. Lichtlen. History of coronary heart disease. *Zeitschrift für Kardiologie*, Volume 91:pp iv56–iv59, August 2002.

[10] E. G. Nabel and E. Braunwald. A tale of coronary artery disease and myocardial infarction. *N. Engl. J. Med.*, 366(1):54–63, Jan 2012.

[11] J. D. Haller, A. S. Olearchyk, and R. H. Goetz. Cardiology's 10 greatest discoveries. *Tex Heart Inst J*, 29(4):342–344, 2002.

[12] P. L. Gross and W. C. Aird. The endothelium and thrombosis. *Semin. Thromb. Hemost.*, 26(5):463–478, 2000.

[13] Y. Matsuzawa and A. Lerman. Endothelial dysfunction and coronary artery disease: assessment, prognosis, and treatment. *Coron. Artery Dis.*, 25(8):713–724, Dec 2014.

[14] University of Malta THINK Magazine. Let the blood flow, 2014. https://www.um.edu.mt/think/let-the-blood-flow/.

[15] Lung National Heart and Blood Institute. Coronary heart disease, 2018. https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease.

[16] National Heart Foundation of Australia. What is coronary heart disease?, 2018. https://www.heartfoundation.org.au/your-heart/heart-conditions/what-is-coronary-heart-disease.

[17] The International Human Genome Sequencing Consortium. The human genome project completion: Frequently asked questions, October 30, 2010. https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/.

[18] Heidi Chial. Rare genetic disorders: Learning about genetic disease through gene mapping, snps, and microarray data. Nature Education, 2008.

[19] Rolf Knippers. *Molekulare Genetik*. Thieme, 9th edition edition, 2006.

[20] J. T. den Dunnen and S. E. Antonarakis. *Nomenclature for the description of human sequence variations*. pp 121–124. Human Genetics, July 2001.

[21] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, Jun 2017.

[22] D. Girelli, C. Piubelli, N. Martinelli, R. Corrocher, and O. Olivieri. A decade of progress on the genetic basis of coronary artery disease. Practical insights for the internist. *Eur. J. Intern. Med.*, 41:10–17, Jun 2017.

[23] P. van der Harst and N. Verweij. The Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.*, Dec 2017.

[24] R. Roberts. Genetics of coronary artery disease: an update. *Methodist Debakey Cardiovasc J*, 10(1):7–12, 2014.

[25] T. A. et al. Manolio. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.

[26] T. L. Assimes and R. Roberts. Genetics: Implications for Prevention and Management of Coronary Artery Disease. *J. Am. Coll. Cardiol.*, 68(25):2797–2818, Dec 2016.

[27] J. A. et al. Hartiala. Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nat Commun*, 7:10558, Jan 2016.

[28] T. Theodorson. Cardiovascular risk and risk reduction: a review of recent literature. *J Family Community Med*, 2(1):19–26, Jan 1995.

[29] World Health Organisation. Obesity and overweight, October 2017. http://www.who.int/mediacentre/factsheets/fs311/en/.

[30] X. Sun, P. Li, X. Yang, W. Li, X. Qiu, and S. Zhu. From genetics and epigenetics to the future of precision treatment for obesity. *Gastroenterol Rep (Oxf)*, 5(4):266–270, Nov 2017.

[31] J. Upadhyay, O. Farr, N. Perakakis, W. Ghaly, and C. Mantzoros. Obesity as a Disease. *Med. Clin. North Am.*, 102(1):13–33, Jan 2018.

[32] X. Dai, S. Wiernek, J. P. Evans, and M. S. Runge. Genetics of coronary artery disease and myocardial infarction. *World J Cardiol*, 8(1):1–23, Jan 2016.

[33] N. Srivastava, R. Lakhan, and B. Mittal. Pathophysiology and genetics of obesity. *Indian J. Exp. Biol.*, 45(11):929–936, Nov 2007.

[34] T. Steemburgo, M. J. Azevedo, J. L. Gross, F. I. Milagro, J. Campion, and J. A. Martinez. The rs9939609 polymorphism in the FTO gene is associated with fat and fiber intakes in patients with type 2 diabetes. *J Nutrigenet Nutrigenomics*, 6(2):97–106, 2013.

[35] P. et al. Poirier. Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss. *Arterioscler. Thromb. Vasc. Biol.*, 26(5):968–976, May 2006.

[36] M. A. et al. Sarzynski. Associations of markers in 11 obesity candidate genes with maximal weight loss and weight regain in the SOS bariatric surgery cases. *Int J Obes (Lond)*, 35(5):676–683, May 2011.

[37] Robert H. Eckel. Obesity and heart disease. *Circulation*, Volume 96, Issue 9, November 4, 1997.

[38] K. T. et al. Mills. Global Disparities of Hypertension Prevalence and Control: A Systematic Analysis of Population-Based Studies From 90 Countries. *Circulation*, 134(6):441–450, Aug 2016.

[39] Oscar A. Carretero and Suzanne Oparil. Essential hypertension. *Circulation*, January 25, 2000.

[40] A. Aggarwal and D. Rodriguez-Buritica. Monogenic Hypertension in Children: A Review With Emphasis on Genetics. *Adv Chronic Kidney Dis*, 24(6):372–379, Nov 2017.

[41] P. A. Doris. Genetics of hypertension: an assessment of progress in the spontaneously hypertensive rat. *Physiol. Genomics*, 49(11):601–617, Nov 2017.

[42] S. Y. Ahn and C. Gupta. Genetic Programming of Hypertension. *Front Pediatr*, 5:285, 2017.

[43] S. K. et al. Ganesh. Loci influencing blood pressure identified using a cardiovascular gene-centric array. *Hum. Mol. Genet.*, 22(8):1663–1678, Apr 2013.

[44] World Health Organisation. Diabetes, November 2017. http://www.who.int/mediacentre/factsheets/fs312/en/.

[45] M. Szabo, B. Mate, K. Csep, and T. Benedek. Genetic Approaches to the Study of Gene Variants and Their Impact on the Pathophysiology of Type 2 Diabetes. *Biochem. Genet.*, Nov 2017.

[46] T. Dendup, X. Feng, S. Clingan, and T. Astell-Burt. Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *Int J Environ Res Public Health*, 15(1), Jan 2018.

[47] M. Ridderstrale and E. Nilsson. Type 2 diabetes candidate gene CAPN10: first, but not last. *Curr. Hypertens. Rep.*, 10(1):19–24, Feb 2008.

[48] V. et al. Bermudez. PPAR-gamma agonists and their role in type 2 diabetes mellitus management. *Am J Ther*, 17(3):274–283, 2010.

[49] N. et al. Sarwar. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet*, 375(9733):2215–2222, Jun 2010.

[50] P. Anagnostis, S. A. Paschou, D. G. Goulis, V. G. Athyros, and A. Karagiannis. Dietary management of dyslipidaemias. Is there any evidence for cardiovascular benefit? *Maturitas*, 108:45–52, Feb 2018.

[51] Rafael A. Cox and Mario R. García-Palmieri. *Clinical Methods: The History, Physical, and Laboratory Examinations.* Butterworth Publishers, a division of Reed Publishing, 3rd edition edition, 1990.

[52] J. F. de Boer, F. Kuipers, and A. K. Groen. Cholesterol Transport Revisited: A New Turbo Mechanism to Drive Cholesterol Excretion. *Trends Endocrinol. Metab.*, Dec 2017.

[53] M. S. Brown and J. L. Goldstein. Receptor-mediated control of cholesterol metabolism. *Science*, 191(4223):150–154, Jan 1976.

[54] J. L. Goldstein and M. S. Brown. The LDL receptor. *Arterioscler. Thromb. Vasc. Biol.*, 29(4):431–438, Apr 2009.

[55] Trevor Huff; Ishwarlal I. Jialal. *Physiology, Cholesterol.* StatPearls Publishing LLC, November 14, 2017.

[56] World Health Organisation. Global health observatory (gho) data - raised cholesterol, 2018. http://www.who.int/gho/ncd/risk_factors/cholesterol_text/en/.

[57] A. K. Soutar and R. P. Naoumova. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nat Clin Pract Cardiovasc Med*, 4(4):214–225, Apr 2007.

[58] J. C. Defesche, S. S. Gidding, M. Harada-Shiba, R. A. Hegele, R. D. Santos, and A. S. Wierzbicki. Familial hypercholesterolaemia. *Nat Rev Dis Primers*, 3:17093, Dec 2017.

[59] M. et al. Wicinski. PCSK9 signaling pathways and their potential importance in clinical practice. *EPMA J*, 8(4):391–402, Dec 2017.

[60] Michael A. Ibrahim; Ishwarlal I. Jialal. *Hypercholesterolemia*. StatPearls Publishing LLC., 2017.

[61] S J Enna; David B Bylund. *xPharm: The Comprehensive Pharmacology Reference*. Elsevier Science, 2008.

[62] J. S. Dron and R. A. Hegele. Genetics of Triglycerides and the Risk of Atherosclerosis. *Curr Atheroscler Rep*, 19(7):31, Jul 2017.

[63] J. Peng, F. Luo, G. Ruan, R. Peng, and X. Li. Hypertriglyceridemia and atherosclerosis. *Lipids Health Dis*, 16(1):233, Dec 2017.

[64] Michael Miller et. al. Triglycerides and cardiovascular disease. *Circulation*, Volume 123, Issue 20, May 24, 2011.

[65] A. Byrne, S. Makadia, A. Sutherland, and M. Miller. Optimizing Non-Pharmacologic Management of Hypertriglyceridemia. *Arch. Med. Res.*, Dec 2017.

[66] Harsha Karanchi; Kathleen Wyne. *Hypertriglyceridemia*. StatPearls Publishing LLC., 2017.

[67] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, Jan 2016.

[68] National Human Genome Research Institute. The cost of sequencing a human genome, July 06 2016. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/.

[69] M. D. et al. Caldwell. Evaluation of genetic factors for warfarin dose prediction. *Clin Med Res*, 5(1):8–16, Mar 2007.

[70] Amrita Kumari Panda Satpal Singh Bisht. *DNA Sequencing: Methods and Applications*. Number pp 11-23 in Advances in Biotechnology. Springer, 22 October 2014.

[71] Inc. Illumina. Genome analyzer iix system, 2009. https://www.illumina.com/Documents/products/specifications/specification_genome_analyzer

[72] Inc Illumina. An introduction to next-generation sequencing technology, 2017.

[73] M. et al. Fischer. Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. *Circulation*, 111(7):855–862, Feb 2005.

[74] H. Zhang. Overview of Sequence Data Formats. *Methods Mol. Biol.*, 1418:3–17, 2016.

[75] M. A. et al. DePristo. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[76] G. A. et al. Van der Auwera. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 43:1–33, 2013.

[77] Broad Institute, 2017. https://software.broadinstitute.org/gatk/img/BP_workflow_3.6.png.

[78] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[79] H. et al. Li. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[80] A. et al. McKenna. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, Sep 2010.

[81] S. T. et al. Sherry. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, Jan 2001.

[82] R. E. et al. Mills. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, 21(6):830–839, Jun 2011.

[83] No authors listed. The International HapMap Project. *Nature*, 426(6968):789–796, Dec 2003.

[84] A. Auton, L. D. Brooks, R. M. Durbin, and et al. Garrison. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015.

[85] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164, Sep 2010.

[86] K. D. et al. Pruitt. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, 42(Database issue):D756–763, Jan 2014.

[87] M. et al. Haeussler. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, Nov 2018.

[88] D. R. et al. Zerbino. Ensembl 2018. *Nucleic Acids Res.*, 46(D1):D754–D761, Jan 2018.

[89] A. et al. Frankish. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, Oct 2018.

[90] M. J. et al. Landrum. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, 44(D1):D862–868, Jan 2016.

[91] P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. Phillips, and D. N. Cooper. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, 133(1):1–9, Jan 2014.

[92] M. et al. Lek. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, Aug 2016.

[93] NHLBI Exome Sequencing Project (ESP). Exome variant server, December 15 2011. http://evs.gs.washington.edu/EVS/.

[94] M. L. et al. Speir. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, 44(D1):D717–725, Jan 2016.

[95] W. J. et al. Kent. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, Jun 2002.

[96] I. et al. Dunham. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[97] X. Liu, X. Jian, and E. Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, 32(8):894–899, Aug 2011.

[98] J. et al. Lonsdale. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, Jun 2013.

[99] T. et al. Lappalainen. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[100] L. D. Ward and M. Kellis. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, 40(Database issue):D930–934, Jan 2012.

[101] H. M. et al. Berman. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, 58(Pt 6 No 1):899–907, Jun 2002.

[102] J. Felsenstein and G. A. Churchill. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13(1):93–104, Jan 1996.

[103] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, 4(7):1073–1081, 2009.

[104] I. A. et al. Adzhubei. A method and server for predicting damaging missense mutations. *Nat. Methods*, 7(4):248–249, Apr 2010.

[105] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, 11(4):361–362, Apr 2014.

[106] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–315, Mar 2014.

[107] National Institutes of Health, May 2001. https://www.nhlbi.nih.gov/files/docs/guidelines/atgland

[108] Mayo Clinic. Triglycerides: Why do they matter?, September 2018. http://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186.

[109] NHLBI. High blood pressure, June 2014. https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0062996/?

[110] P. Jones, S. Kafonek, I. Laurora, and D. Hunninghake. Comparative dose efficacy study of atorvastatin versus simvastatin, pravastatin, lovastatin, and fluvastatin in patients with hypercholesterolemia (the CURVES study). *Am. J. Cardiol.*, 81(5):582–587, Mar 1998.

[111] E. Krieger and G. Vriend. YASARA View - molecular graphics for all devices - from smartphones to workstations. *Bioinformatics*, 30(20):2981–2982, Oct 2014.

[112] E. Krieger, T. Darden, S. B. Nabuurs, A. Finkelstein, and G. Vriend. Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins*, 57(4):678–683, Dec 2004.

[113] E. Krieger and G. Vriend. New ways to boost molecular dynamics simulations. *J Comput Chem*, 36(13):996–1007, May 2015.

[114] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[115] S. Rozen and H. Skaletsky. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, 132:365–386, 2000.

[116] C. D. Ahrberg, A. Manz, and B. G. Chung. Polymerase chain reaction in microfluidic devices. *Lab Chip*, 16(20):3866–3884, Oct 2016.

[117] C. Mülhardt. *Der Experimentator Molekularbiologie/Genomics.* Springer, 2013.

[118] S. Adkins and M. Burmeister. Visualization of DNA in agarose gels as migrating colored bands: applications for preparative gels and educational demonstrations. *Anal. Biochem.*, 240(1):17–23, Aug 1996.

[119] S. L. Samson and A. J. Garber. Metabolic syndrome. *Endocrinol. Metab. Clin. North Am.*, 43(1):1–23, Mar 2014.

[120] P. L. Huang. A comprehensive definition for metabolic syndrome. *Dis Model Mech*, 2(5-6):231–237, 2009.

[121] E. J. et al. Schaefer. Familial apolipoprotein E deficiency. *J. Clin. Invest.*, 78(5):1206–1219, Nov 1986.

[122] G. R. Thompson. Management of dyslipidaemia. *Heart*, 90(8):949–955, Aug 2004.

[123] P. S. Hansen. Familial defective apolipoprotein B-100. *Dan Med Bull*, 45(4):370–382, Sep 1998.

[124] Robert A Hegele John R Burnett, Amanda J Hooper. *Familial Lipoprotein Lipase Deficiency*. GeneReviews®, 1999 Oct 12 [updated 2017 Jun 22].

[125] R. W. Mahley, Y. Huang, and S. C. Rall. Pathogenesis of type III hyperlipoproteinemia (dysbetalipoproteinemia). Questions, quandaries, and paradoxes. *J. Lipid Res.*, 40(11):1933–1949, Nov 1999.

[126] Seema Jawalekar. The hyperlipoproteinemia - an approach to diagnosis and classification. *Biochemistry & Physiology: Open Access*, March 31, 2012.

[127] J. Fei, X. B. Cui, J. N. Wang, K. Dong, and S. Y. Chen. ADAR1-Mediated RNA Editing, A Novel Mechanism Controlling Phenotypic Modulation of Vascular Smooth Muscle Cells. *Circ. Res.*, 119(3):463–469, Jul 2016.

[128] J. Jeppesen, H. O. Hein, P. Suadicani, and F. Gyntelberg. High triglycerides and low HDL cholesterol and blood pressure and risk of ischemic heart disease. *Hypertension*, 36(2):226–232, Aug 2000.

[129] C. Muckian, A. Fitzgerald, A. O'Neill, A. O'Byrne, D. J. Fitzgerald, and D. C. Shields. Genetic variability in the extracellular matrix as a determinant of cardiovascular risk: association of type III collagen COL3A1 polymorphisms with coronary artery disease. *Blood*, 100(4):1220–1223, Aug 2002.

[130] R. S. et al. McMahan. Stromelysin-2 (MMP10) Moderates Inflammation by Controlling Macrophage Activation. *J. Immunol.*, 197(3):899–909, 08 2016.

[131] M. G. et al. Rohani. MMP-10 Regulates Collagenolytic Activity of Alternatively Activated Resident Macrophages. *J. Invest. Dermatol.*, 135(10):2377–2384, Oct 2015.

[132] P. Schlage, T. Kockmann, F. Sabino, J. N. Kizhakkedathu, and U. Auf dem Keller. Matrix Metalloproteinase 10 Degradomics in Keratinocytes and Epidermal Tissue Identifies Bioactive Substrates With Pleiotropic Functions. *Mol. Cell Proteomics*, 14(12):3234–3246, Dec 2015.

[133] D. Muller, B. Quantin, M. C. Gesnel, R. Millon-Collard, J. Abecassis, and R. Breathnach. The colcollagen gene family in humans consists of at least four members. *Biochemical Journal*, 253:187–192;, Jul 01, 1988.

[134] J. A. et al. Rodriguez. Metalloproteinases and atherothrombosis: MMP-10 mediates vascular remodeling promoted by inflammatory stimuli. *Front. Biosci.*, 13:2916–2921, Jan 2008.

[135] G. Zhang, M. Miyake, A. Lawton, S. Goodison, and C. J. Rosser. Matrix metalloproteinase-10 promotes tumor progression through regulation of angiogenic and apoptotic pathways in cervical tumors. *BMC Cancer*, 14:310, May 2014.

[136] J. Orbe, I. Montero, J. A. Rodriguez, O. Beloqui, C. Roncal, and J. A. Paramo. Independent association of matrix metalloproteinase-10, cardiovascular risk factors and subclinical atherosclerosis. *J. Thromb. Haemost.*, 5(1):91–97, Jan 2007.

[137] J. et al. Avouac. Role of Stromelysin 2 (Matrix Metalloproteinase 10) as a Novel Mediator of Vascular Remodeling Underlying Pulmonary Hypertension Associated With Systemic Sclerosis. *Arthritis Rheumatol*, 69(11):2209–2221, 11 2017.

[138] F. et al. Zhu. [Association between matrix metalloproteinase-10 gene polymorphisms and instability of carotid plaque]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*, 30(6):711–715, Dec 2013.

[139] Y. Zhao, S. Zhou, and C. K. Heng. Celecoxib inhibits serum amyloid a-induced matrix metalloproteinase-10 expression in human endothelial cells. *J. Vasc. Res.*, 46(1):64–72, 2009.

[140] D. Tantin. RNA polymerase II elongation complexes containing the Cockayne syndrome group B protein interact with a molecular complex containing the transcription factor IIH components xeroderma pigmentosum B and p62. *J. Biol. Chem.*, 273(43):27794–27799, Oct 1998.

[141] A. Singh, E. Compe, N. Le May, and J. M. Egly. TFIIH subunit alterations causing xeroderma pigmentosum and trichothiodystrophy specifically disturb several steps during transcription. *Am. J. Hum. Genet.*, 96(2):194–207, Feb 2015.

[142] W. S. D. Tan, W. Liao, S. Zhou, D. Mei, and W. F. Wong. Targeting the renin-angiotensin system as novel therapeutic strategy for pulmonary diseases. *Curr Opin Pharmacol*, 40:9–17, Jun 2018.

[143] A. et al. Mascolo. New and old roles of the peripheral and brain renin-angiotensin-aldosterone system (RAAS): Focus on cardiovascular and neurological diseases. *Int. J. Cardiol.*, 227:734–742, Jan 2017.

[144] L. D. Ji, J. Y. Li, B. B. Yao, X. B. Cai, Q. J. Shen, and J. Xu. Are genetic polymorphisms in the renin-angiotensin-aldosterone system associated with essential hypertension? Evidence from genome-wide association studies. *J Hum Hypertens*, 31(11):695–698, Nov 2017.

[145] X. et al. Zhu. Associations between hypertension and genes in the renin-angiotensin system. *Hypertension*, 41(5):1027–1034, May 2003.

[146] R. Baudrand and A. Vaidya. The Low-Renin Hypertension Phenotype: Genetics and the Role of the Mineralocorticoid Receptor. *Int J Mol Sci*, 19(2), Feb 2018.

[147] Y. N. et al. Zhang. Phenotypic switching of vascular smooth muscle cells in the 'normal region' of aorta from atherosclerosis patients is regulated by miR-145. *J. Cell. Mol. Med.*, 20(6):1049–1061, 06 2016.

[148] L. et al. Perisic Matic. Phenotypic Modulation of Smooth Muscle Cells in Atherosclerosis Is Associated With Downregulation of LMOD1, SYNPO2, PDLIM7, PLN, and SYNM. *Arterioscler. Thromb. Vasc. Biol.*, 36(9):1947–1961, 09 2016.

[149] H. Poulsen, J. Nilsson, C. K. Damgaard, J. Egebjerg, and J. Kjems. CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain. *Mol. Cell. Biol.*, 21(22):7862–7871, Nov 2001.

[150] Z. L. et al. Tang. Eight Novel Mutations of the ADAR1 Gene in Chinese Patients with Dyschromatosis Symmetrica Hereditaria. *Genet Test Mol Biomarkers*, 22(2):104–108, Feb 2018.

[151] C. Song, M. Sakurai, Y. Shiromoto, and K. Nishikura. Functions of the RNA Editing Enzyme ADAR1 and Their Relevance to Human Diseases. *Genes (Basel)*, 7(12), Dec 2016.

[152] K. et al. Stellos. Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. *Nat. Med.*, 22(10):1140–1150, 10 2016.

[153] W. J. Hendriks and R. Pulido. Protein tyrosine phosphatase variants in human hereditary disorders and disease susceptibilities. *Biochim. Biophys. Acta*, 1832(10):1673–1696, Oct 2013.

[154] S. et al. Harroch. A critical role for the protein tyrosine phosphatase receptor type Z in functional recovery from demyelinating lesions. *Nat. Genet.*, 32(3):411–414, Nov 2002.

[155] K. V. et al. Lu. Differential induction of glioblastoma migration and growth by two forms of pleiotrophin. *J. Biol. Chem.*, 280(29):26953–26964, Jul 2005.

[156] S. R. et al. Berkowicz. Brinp1(-/-) mice exhibit autism-like behaviour, altered memory, hyperactivity and increased parvalbumin-positive cortical interneuron density. *Mol Autism*, 7:22, 2016.

[157] A. Shendre, M. R. Irvin, H. Wiener, D. Zhi, N. A. Limdi, E. T. Overton, and S. Shrestha. Local Ancestry and Clinical Cardiovascular Events Among African Americans From the Atherosclerosis Risk in Communities Study. *J Am Heart Assoc*, 6(4), Apr 2017.

[158] M. I. et al. Rees. Mutations in the gene encoding GlyT2 (SLC6A5) define a presynaptic component of human startle disease. *Nat. Genet.*, 38(7):801–806, Jul 2006.

[159] E. et al. Carta. Mutations in the GlyT2 gene (SLC6A5) are a second major cause of startle disease. *J. Biol. Chem.*, 287(34):28975–28985, Aug 2012.

[160] V. Sequeira, L. L. Nijenkamp, J. A. Regan, and J. van der Velden. The physiological role of cardiac cytoskeleton and its alterations in heart failure. *Biochim. Biophys. Acta*, 1838(2):700–722, Feb 2014.

[161] S. Wilkinson, H. F. Paterson, and C. J. Marshall. Cdc42-MRCK and Rho-ROCK signalling cooperate in myosin phosphorylation and cell invasion. *Nat. Cell Biol.*, 7(3):255–261, Mar 2005.

[162] T. Leung, X. Q. Chen, I. Tan, E. Manser, and L. Lim. Myotonic dystrophy kinase-related Cdc42-binding kinase acts as a Cdc42 effector in promoting cytoskeletal reorganization. *Mol. Cell. Biol.*, 18(1):130–140, Jan 1998.

[163] Y. Zhao, P. Loyer, H. Li, V. Valentine, V. Kidd, and A. S. Kraft. Cloning and chromosomal location of a novel member of the myotonic dystrophy family of protein kinases. *J. Biol. Chem.*, 272(15):10013–10020, Apr 1997.

[164] P. Y. et al. He. Inhibition of cell migration and invasion by miRâ€'29aâ€'3p in a colorectal cancer cell line through suppression of CDC42BPA mRNA expression. *Oncol. Rep.*, 38(6):3554–3566, Dec 2017.

[165] J. et al. Ye. Expression of protein TARBP1 in human hepatocellular carcinoma and its prognostic significance. *Int J Clin Exp Pathol*, 8(8):9089–9096, 2015.

[166] M. et al. Sand. Expression levels of the microRNA maturing microprocessor complex component DGCR8 and the RNA-induced silencing complex (RISC) components argonaute-1, argonaute-2, PACT, TARBP1, and TARBP2 in epithelial skin cancer. *Mol. Carcinog.*, 51(11):916–922, Nov 2012.

[167] H. Wu, J. Min, H. Zeng, and A. N. Plotnikov. Crystal structure of the methyl-transferase domain of human TARBP1. *Proteins*, 72(1):519–525, Jul 2008.

[168] A. K. et al. Singh. DOCK8 regulates fitness and function of regulatory T cells through modulation of IL-2 signaling. *JCI Insight*, 2(19), Oct 2017.

[169] C. J. Kearney, K. L. Randall, and J. Oliaro. DOCK8 regulates signal transduction events to control immunity. *Cell. Mol. Immunol.*, 14(5):406–411, May 2017.

[170] S. Wang, W. Mou, Z. Xu, J. Gui, and L. Ma. Autosomal recessive hyper-IgE syndrome in two brothers of a Chinese family with a novel mutation in DOCK8 gene. *J Eur Acad Dermatol Venereol*, Feb 2018.

[171] F. J. et al. Alroqi. DOCK8 Deficiency Presenting as an IPEX-Like Disorder. *J. Clin. Immunol.*, 37(8):811–819, Nov 2017.

[172] R. Steet and S. Kornfeld. COG-7-deficient Human Fibroblasts Exhibit Altered Recycling of Golgi Proteins. *Mol. Biol. Cell*, 17(5):2312–2321, May 2006.

[173] G. Belloni, S. Sechi, M. G. Riparbelli, M. T. Fuller, G. Callaini, and M. G. Giansanti. Mutations in Cog7 affect Golgi structure, meiotic cytokinesis and sperm development during Drosophila spermatogenesis. *J. Cell. Sci.*, 125(Pt 22):5441–5452, Nov 2012.

[174] X. et al. Wu. Mutation of the COG complex subunit gene COG7 causes a lethal congenital disorder. *Nat. Med.*, 10(5):518–523, May 2004.

[175] B. J. Fenner, M. Scannell, and J. H. Prehn. Identification of polyubiquitin binding proteins involved in NF-kappaB signaling using protein arrays. *Biochim. Biophys. Acta*, 1794(7):1010–1016, Jul 2009.

[176] F. Wang, A. Lerman, and J. Herrmann. Dysfunction of the ubiquitin-proteasome system in atherosclerotic cardiovascular disease. *Am J Cardiovasc Dis*, 5(1):83–100, 2015.

[177] S. et al. Dong. The human SRCAP chromatin remodeling complex promotes DNA-end resection. *Curr. Biol.*, 24(18):2097–2110, Sep 2014.

[178] M. M. Wong, L. K. Cox, and J. C. Chrivia. The chromatin remodeling protein, SRCAP, is critical for deposition of the histone variant H2A.Z at promoters. *J. Biol. Chem.*, 282(36):26132–26139, Sep 2007.

[179] D. et al. Milani. Perthes disease: A new finding in Floating-Harbor syndrome. *Am. J. Med. Genet. A*, 176(3):703–706, Mar 2018.

[180] G. Messina, M. T. Atterrato, and P. Dimitri. When chromatin organisation floats astray: the Srcap gene and Floating-Harbor syndrome. *J. Med. Genet.*, 53(12):793–797, 12 2016.

[181] M. Kehrer, A. Beckmann, J. Wyduba, U. Finckh, A. Dufke, U. Gaiser, and A. Tzschach. Floating-Harbor syndrome: SRCAP mutations are not restricted to exon 34. *Clin. Genet.*, 85(5):498–499, May 2014.

[182] M. W. Hur, J. H. Yoon, M. Y. Kim, H. Ko, and B. N. Jeon. Kr-POK (ZBTB7c) regulates cancer cell proliferation through glutamine metabolism. *Biochim. Biophys. Acta*, 1860(8):829–838, 08 2017.

[183] R. et al. Du. Integrative Mouse and Human Studies Implicate ANGPT1 and ZBTB7C as Susceptibility Genes to Ischemic Injury. *Stroke*, 46(12):3514–3522, Dec 2015.

List of Tables

List of Tables

List of Figures

List of Abbreviations

1000G   1000 Genomes Project

AA   amino acid

AF   allele frequency

BAM   Binary Alignment Map

BMI   body mass index

bp   basepairs

BWA   Burrows-Wheeler Aligner

CAD   coronary artery disease

CADD   Combined Annotation Dependent Depletion

COG   conserved oligomeric Golgi

CTSS   cathepsins S

dbSNP   Single Nucleotide Polymorphism Database

ddNTPs   dideoxynucleotide triphosphates

DNA   deoxyribonucleic acid

dNTPs   deoxynucleoside triphosphates

DSRAD   double-stranded RNA-specific adenosine deaminase

ECM   extracellular matrix

ENCODE   Encyclopedia of DNA Elements

eQTLs   expression quantitative trait loci

ESP   Exome Sequencing Project

ExAC   Exome Aggregation Consortium

FH   familial hypercholersterolemia

GATK   Genome Analysis Toolkit

Germif   German MI family

GlyT2   glycine transporter 2

GOF   gain-of-function

GTEx   Genotype-Tissue Expression

GVCF   Genomic Variant Call Format

GWAS   genome-wide association studies

HDL   high-density lipoprotein

HGMD®   Human Gene Mutation Database

HT   hypertension

LDL   low-density lipoprotein

LOD   logarithm of the odds

LOF   loss-of-function

MAF   minor allele frequency

MD   molecular dynamics

MI   myocardial infarction

MMPs   matrix metalloproteinases

NER   nucleotide excision repair

PCR   polymerase chain reaction

PDB   Protein Data Bank

PolyPhen2   Polymorphism Phenotyping v2

RefSeq   NCBI Reference Sequence Database

RefSeq   Reference Sequence

RMSDCα root-mean-square deviation of the C alpha

RMSF   root mean square fluctuation

SAM   Sequence Alignment Map

SDC   sudden cardiac death

SIFT   Sorting Intolerant From Tolerant

SMC   smooth muscle cells

SNP   single nucleotide polymorphism

TFIIH   transcription factor IIH

UTR   untranslated region

VCF   variant call format

VCF   Variant Call Format

VLDL   very-low-density lipoprotein

VQSR   variant quality score recalibration

WES   whole-exome sequencing

WGS   whole-genome sequencing

WHO   World Health Organization

WT   wild type

YASARA   Yet Another Scientific Artificial Reality Application

Code snippet 7 Script used for calculating and plotting the LOD score with R

```
CalcLOD <- function(R,NR,x){

r <- vector(mode="numeric")
LOD <- vector(mode="numeric")
LODs <- 0
rs <- 0.01
while ( x < 0.51 ){

    pL <-((x/2)^R)*((1-x)/2)^NR;

    pNL <- (0.25)^(R+NR);

    LODv <- log10(pL/pNL)
    LOD <- append(LOD,log10(pL/pNL))
    rv <- x
    r <- append(r, x)

    x <- x+0.01;

    if(LODv>LODs){

LODs <- LODv
rs <- rv


    }
}
print(LODs)
print(rs)
plot(r,LOD, type="l")
}
```

Table A.1.: List of known CAD genes, state 26.10.2017

- 9p21
- PHACTR1
- SLC22A3
- LPAL2
- LPA
- PLG
- SORT1
- ADAMTS7
- KCNE2
- FN1
- LDLR
- CXCL12
- ZNF507-LOC400684
- PPAP2B
- ANRIL
- CDKN2A
- CDKN2B

- LRP1
- SCARB1
- MIA3
- LIPA
- C2
- TCF21
- KIAA1462
- ZC3HC1
- ATP2B1
- PDGFD
- APOE
- APOC1
- HDAC9
- COL4A1
- COL4A2
- POM121L9P-ADORA2A

- VAMP5
- VAMP8
- GGCX
- EDNRA
- SH2B3
- ARHGEF26
- LOC646736
- KCNJ13-GIGYF2
- NOS3
- KSR2
- LOX
- IL6R
- ZEB2
- ACO74093.1
- CCDC92
- LMOD1
- SMAD3

A. Supplement

- CYP17A1
- CNNM2
- NT5C2
- GUCY1A3
- CETP
- HHIPL1
- AK097927
- MRAS
- MRVI1-CTR9
- CDH13
- PROCR
- ARHGAP26
- SWAP70
- UMPS-ITGB5
- SERPINH1
- PCSK9
- BCAS3
- WDR12
- RHOA
- KCNK5
- RP11664H17.1

- DHX38
- TGFB1
- BCAR1-CFDP1-TMEM170A
- DDX59/CAMSAP2
- FGD5
- MAD2L1
- PMAIP1-MC4R
- FGF5
- MFGE8-ABHD2
- ATP1B1
- GOSR2
- HNF1A
- REST-NOA1
- PCNXL3
- PECAM1
- PARP12
- RASD1
- PEMT
- RAI1
- ABO

- ZNF259
- APOA1
- APO5A
- FURIN
- FES
- LPL
- SMG6
- NT5C2
- CNNM2
- CYP17A1
- OAZ2
- RBPMS2
- TRIB1
- TNS1a
- ANKS1A

Table A.2.: Full results on pathway enrichment in known CAD genes.

P-value	Q-value	Pathway	Source	Overlapping members from input
4.70E-14	2.06E-11	Lipoprotein metabolism	Reactome	LIPA; APOE; CETP; SCARB1; LPL; LDLR; PCSK9; APOC1; APOA1; LPA; FURIN
2.87E-12	6.29E-10	Statin Pathway	Wikipathways	APOE; CETP; SCARB1; LRP1; LDLR; LPL; APOC1; APOA1
1.10E-1	1.61E-09	Lipid digestion, mobilization, and transport	Reactome	LIPA; PCSK9; CETP; APOE; SCARB1; LDLR; LPL; APOC1; APOA1; LPA; FURIN
3.77E-11	4.13E-09	Statin Pathway, Pharmacodynamics	PharmGKB	LRP1; APOE; SCARB1; LDLR; LPL; APOC1; APOA1
7.92E-08	6.93E-06	Binding and Uptake of Ligands by Scavenger Receptors	Reactome	APOE; SCARB1; LRP1; COL4A2; COL4A1; APOA1
1.12E-07	6.99E-06	Chylomicron-mediated lipid transport	Reactome	LPL; LDLR; APOE; APOA1; FURIN
1.12E-07	6.99E-06	LDL-mediated lipid transport	Reactome	LIPA; LDLR; PCSK9; CETP; LPA
5.65E-07	3.09E-05	Composition of Lipid Particles	Wikipathways	LDLR; LPL; CETP

Table A.3.: Full results on pathway enrichment in the genes retrieved from the filtered exome-sequencing data.

P-value	Q-value	Pathway	Source	Overlapping members from input
2.80E-14	9.92E-11	ECM-receptor interaction - Homo sapiens (human)	KEGG	TNC; ITGA1; ITGA2; HSPG2; ITGB3; CD36; ITGA7; AGRN; ITGA10; ITGA11; LAMB1; LAMB3; LAMB2; COL2A1; LAMB4; LAMA5; LAMC1; ITGA3; ITGA2B; THBS4; TNR; GP1BA; THBS3; ITGA6; THBS1; GP6; RELN; GP5; SV2C; HMMR; LAMA1; LAMA2; LAMA3; ITGB5; ITGB4; ITGB6; CD44; COL9A2; COL9A3; COL9A1; COL4A6; COL4A5; COL4A4; COL4A3; COL4A2; COL4A1; LAMC2; LAMC3; TNN; COL6A3; FN1; SDC1; ITGAV; SPP1; COL1A1; COL6A1; VWF; COL6A2; COL6A5; COL6A6
4.52E-12	8.01E-09	Extracellular matrix organization	Reactome	MUSK; COL11A1; P4HA3; FBLN2; FBLN5; ADAMTS4; ADAMTS1; ADAMTS3; ADAMTS2; COL27A1; CAPN7; CAPN6; CAPN5; P3H3; P3H2; P3H1; CDH1; FGG; FGA; CTSB; MADCAM1; COL19A1; COL15A1; COL7A1; FN1; TRAPPC4; ADAM8; ADAM9; MMP17; MMP10; MMP12; MMP13; ITGA1; ITGA2; ITGA3; ITGB3; ITGA6; NCSTN; ITGA10; ITGA11; ADAMTS18; LAMB1; COL3A1; LAMB3; LAMB2; MATN4; LTBP1; TTR; TNR; THBS1; NID2; NID1; CEACAM6; PCOLCE; SPP1; DST; EFEMP2; COL4A6; CMA1; COL4A4; PLEC; COL4A2; COL4A1; LAMC2; LAMC3; COL16A1; LAMC1; FGF2; ITGAX; SDC1; COL22A1; COL4A5; ITGAV; COL4A3; ITGAE; COL12A1; COLGALT2; ACAN; COL24A1; DMD; COL2A1; A2M; MATN3; TLL1; TLL2; ITGA2B; ASPN; SPARC; ICAM1; COL26A1; CAPN3; CAPN2; LAMA1; LAMA2; LAMA3; FMOD; COL9A2; COL9A3; COL9A1; COL5A2; COL5A3; CAPN13; ITGA7; COL17A1; PHYKPL; CAPN15; NTN4; CAPN12; PLG; COL13A1; HSPG2; COL1A1; COL21A1; HAPLN1; COL10A1; TGFB1; VCAN; AGRN; ADAM17; LOX; LAMA5; PLOD2; PLOD3; COL28A1; MATN1; COL8A1; LTBP4; ITGB5; ITGB4; KLKB1; ITGB6; COL18A1; CD44; FBN2; FBN3; FBN1; TNC; MMP24; TNN; ADAMTS16; LTBP2; COL14A1; COL20A1; DSPP; MMP9; COL6A1; COL6A3; COL6A2; COL6A5; MMP3; COL6A6
8.25E-12	9.75E-09	Collagen chain trimerization	Reactome	COL10A1; COL11A1; COL3A1; COL2A1; COL27A1; COL28A1; COL26A1; COL21A1; COL8A1; COL9A3; COL18A1; COL9A2; COL19A1; COL9A1; COL5A2; COL4A6; COL4A5; COL4A4; COL4A3; COL4A2; COL4A1; COL24A1; COL16A1; COL17A1; COL15A1; COL7A1; COL5A3; COL22A1; COL14A1; COL13A1; COL1A1; COL6A1; COL6A3; COL6A2; COL6A5; COL20A1; COL12A1; COL6A6
1.71E-11	1.25E-08	Collagen formation	Reactome	COL10A1; COLGALT2; COL11A1; MMP13; ITGA6; P4HA3; COL3A1; LAMB3; COL2A1; LOX; TLL1; TLL2; ADAMTS3; ADAMTS2; PLOD2; PLOD3; COL27A1; P3H3; P3H2; P3H1; COL26A1; PCOLCE; DST; COL19A1; COL28A1; COL21A1; COL8A1; LAMA3; ITGB4; COL18A1; CTSB; MMP9; COL9A2; COL9A3; COL9A1; COL5A2; COL4A6; COL4A5; COL4A4; PLEC; COL4A2; COL4A1; LAMC2; COL24A1; COL16A1; COL17A1; COL15A1; COL7A1; COL5A3; COL22A1; COL14A1; COL20A1; COL13A1; COL1A1; COL6A1; COL4A3; COL6A3; COL6A2; COL6A5; MMP3; COL12A1; COL6A6
1.77E-11	1.25E-08	Collagen biosynthesis and modifying enzymes	Reactome	COLGALT2; COL10A1; COL11A1; P4HA3; COL3A1; COL2A1; TLL1; TLL2; ADAMTS3; ADAMTS2; PLOD2; PLOD3; COL27A1; P3H3; P3H2; P3H1; COL26A1; PCOLCE; COL28A1; COL8A1; COL6A1; COL9A3; COL18A1; COL9A2; COL19A1; COL9A1; COL5A2; COL4A6; COL4A5; COL4A4; COL4A3; COL4A2; COL4A1; COL24A1; COL16A1; COL17A1; COL15A1; COL7A1; COL5A3; COL22A1; COL14A1; COL13A1; COL1A1; COL21A1; COL6A3; COL6A2; COL6A5; COL20A1; COL12A1; COL6A6
8.40E-09	4.96E-06	Beta1 integrin cell surface interactions	PID	TGM2; IGSF8; TNC; ITGA1; ITGA2; ITGA3; ITGA6; ITGA7; ITGA10; ITGA11; LAMB1; COL3A1; LAMB3; LAMB2; COL2A1; COL11A1; THBS1; NID1; FGG; FGA; LAMA1; LAMA2; LAMA3; LAMA5; COL18A1; FN1; PLAU; FBN1; COL5A2; COL4A6; COL4A5; COL4A4; COL4A3; COL4A1; LAMC2; LAMC1; COL7A1; CSPG4; ITGAV; NPNT; SPP1; COL1A1; COL6A1; TGFBI; COL6A3; COL6A2

P-value	Q-value	Pathway	Source	Overlapping members from input
9.96E-08	5.05E-05	Assembly of collagen fibrils and other multimeric structures	Reactome	COL10A1; MMP13; ITGA6; LAMB3; LOX; TLL1; TLL2; PCOLCE; CTSB; COL8A1; LAMA3; ITGB4; COL18A1; DST; COL9A2; COL9A3; COL9A1; COL4A6; COL4A5; COL4A4; PLEC; COL4A2; COL4A1; LAMC2; COL15A1; COL7A1; MMP9; COL6A1; COL4A3; COL6A3; COL6A2; COL6A5; MMP3; COL6A6
1.30E-07	5.57E-05	Laminin interactions	Reactome	LAMA1; LAMA2; LAMA3; LAMA5; COL18A1; ITGA1; LAMB1; ITGA3; ITGAV; ITGA7; NID2; NID1; HSPG2; ITGA6; LAMC1; LAMC2; LAMC3; LAMB3; LAMB2; ITGA2 ; ITGB4
1.41E-07	5.57E-05	Integrin	INOH	COL10A1; DOCK8; DOCK9; COL11A1; DOCK2; DOCK3; ITGA2; ITGA3; VCAN; ITGB3; DOCK4; ITGA7; AGRN; MAP2K2; ITGA10; ITGA11; RAPGEF1; COL3A1; LAMB3; COL2A1; COL9A2; ITGA2B; SMC3; TLN2; ITGA6; DOCK7; MAPK3; ITGA1; ACAN; COL19A1; COL8A1; HSPG2; ELN; ITGB4; SHC2; ITGB6; COL18A1; CD44; ITGB5; COL9A3; COL9A1; COL5A2; COL4A6; COL4A5; COL4A4; COL4A3; COL4A2; COL4A1; LAMC2; COL16A1; DOCK5; COL7A1; ITGAX; COL15A1; SOS2; FN1; GPC1; DOCK6; SDC1; ITGAV; DOCK1; TNC; COL1A1; COL6A1; GPC4; COL6A2; ITGAE; COL12A1
2.13E-07	7.56E-05	ABC transporters - Homo sapiens (human)	KEGG	ABCA8; ABCB1; ABCB7; ABCB6; ABCB4; ABCD1; ABCA2; ABCD3; ABCB8; ABCG2; ABCC12; ABCC10; ABCC11; ABCA10; ABCG1; ABCA12; ABCA13; ABCG5; ABCD2; ABCC1; ABCA7; ABCA4; ABCC8; ABCA3; ABCA1; ABCC4; ABCC5; ABCA9; CFTR; ABCC2; ABCC3

A. Supplement

Table A.4.: Filtered variants from candidate genes found in database of 400 controls and 655 exome-sequenced cases.

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr11	102641531	102641531	T	A	exonic	MMP10	NM_002425:exon10:c.A1424T:p.H475L
chr11	102643604	102643604	G	T	exonic	MMP10	NM_002425:exon8:c.C1200A:p.Y400X
chr11	102643636	102643636	C	A	exonic	MMP10	NM_002425:exon8:c.G1168T:p.A390S
chr11	102643657	102643657	T	C	exonic	MMP10	NM_002425:exon8:c.A1147G:p.T383A
chr11	102647090	102647090	T	A	exonic	MMP10	NM_002425:exon6:c.A853T:p.M285L
chr11	102647156	102647156	C	G	splicing	MMP10	NM_002425:exon6:c.788-1G>C
chr11	102647396	102647396	A	G	exonic	MMP10	NM_002425:exon5:c.T734C:p.L245P
chr11	102647452	102647452	A	C	exonic	MMP10	NM_002425:exon5:c.T678G:p.F226L
chr11	102649402	102649402	C	T	exonic	MMP10	NM_002425:exon4:c.G575A:p.G192E
chr11	102649457	102649457	C	T	exonic	MMP10	NM_002425:exon4:c.G520A:p.D174N
chr11	102650052	102650052	-	T	exonic	MMP10	NM_002425:exon3:c.387_388insA:p.D130fs
chr1	204125410	204125410	C	T	exonic	REN	NM_000537:exon8:c.G856A:p.G286S
chr11	18357457	18357457	A	G	exonic	GTF2H1	NM_005316:exon3:c.A311G:p.K104R
chr11	18359721	18359721	G	T	exonic	GTF2H1	NM_005316:exon4:c.G413T:p.S138I
chr11	18362901	18362901	G	A	exonic	GTF2H1	NM_005316:exon6:c.G701A:p.R234Q
chr11	18380090	18380090	T	C	exonic	GTF2H1	NM_005316:exon13:c.T1370C:p.I457T
chr11	18387342	18387342	A	G	exonic	GTF2H1	NM_005316:exon15:c.A1573G:p.I525V
chr1	154557452	154557452	A	C	exonic	ADAR	NM_001025107:exon15:c.T2626G:p.F876V
chr1	154557515	154557515	G	A	exonic	ADAR	NM_001025107:exon15:c.C2563T:p.R855W
chr1	154558285	154558285	C	A	exonic	ADAR	NM_001025107:exon13:c.G2374T:p.D792Y
chr1	154562715	154562715	C	T	exonic	ADAR	NM_001025107:exon7:c.G1556A:p.S519N
chr1	154562842	154562842	C	T	exonic	ADAR	NM_001025107:exon7:c.G1429A:p.V477I
chr1	154570367	154570367	T	C	exonic	ADAR	NM_001025107:exon4:c.A986G:p.H329R
chr1	154570905	154570905	G	C	exonic	ADAR	NM_001025107:exon3:c.C873G:p.H291Q
chr1	154574037	154574037	C	T	exonic	ADAR	NM_001025107:exon2:c.G196A:p.E66K
chr1	154574056	154574056	A	C	exonic	ADAR	NM_001025107:exon2:c.T177G:p.H59Q
chr1	154574135	154574135	C	T	exonic	ADAR	NM_001025107:exon2:c.G98A:p.R33Q
chr1	154574346	154574346	C	T	exonic	ADAR	NM_001111:exon2:c.G772A:p.G258R
chr1	154574457	154574457	G	A	exonic	ADAR	NM_001111:exon2:c.C661T:p.P221S
chr1	154574526	154574526	C	T	exonic	ADAR	NM_001111:exon2:c.G592A:p.A198T
chr1	154574541	154574541	G	C	exonic	ADAR	NM_001111:exon2:c.C577G:p.P193A
chr1	154574631	154574631	C	T	exonic	ADAR	NM_001111:exon2:c.G487A:p.G163R
chr1	154575014	154575014	G	T	exonic	ADAR	NM_001111:exon2:c.C104A:p.S35Y
chr1	154575087	154575087	C	T	exonic	ADAR	NM_001111:exon2:c.C31A:p.G11R
chr7	121513582	121513582	G	T	exonic	PTPRZ1	NM_001206838:exon1:c.G29T:p.C10F
chr7	121513612	121513615	GTGA	-	splicing	PTPRZ1	NA
chr7	121608085	121608085	G	A	exonic	PTPRZ1	NM_001206838:exon3:c.G205A:p.D69N
chr7	121608109	121608109	C	A	exonic	PTPRZ1	NM_001206838:exon3:c.C229A:p.L77I
chr7	121623815	121623815	C	T	exonic	PTPRZ1	NM_001206838:exon7:c.C716T:p.P239L
chr7	121636450	121636450	C	G	exonic	PTPRZ1	NM_001206838:exon9:c.C943G:p.P315A
chr7	121636589	121636589	A	G	exonic	PTPRZ1	NM_001206838:exon9:c.A1082G:p.H361R
chr7	121650440	121650440	G	T	exonic	PTPRZ1	NM_001206838:exon12:c.G1340T:p.R447I
chr7	121650505	121650505	C	T	exonic	PTPRZ1	NM_001206838:exon12:c.C1405T:p.R469C
chr7	121650508	121650508	A	T	exonic	PTPRZ1	NM_001206838:exon12:c.A1408T:p.I470L
chr7	121650917	121650917	G	A	exonic	PTPRZ1	NM_001206838:exon12:c.G1817A:p.G606E
chr7	121651306	121651306	G	A	exonic	PTPRZ1	NM_001206838:exon12:c.G2206A:p.D736N
chr7	121651337	121651337	C	T	exonic	PTPRZ1	NM_001206838:exon12:c.C2237T:p.S746L
chr7	121651351	121651351	C	G	exonic	PTPRZ1	NM_001206838:exon12:c.C2251G:p.P751A
chr7	121651352	121651352	C	T	exonic	PTPRZ1	NM_001206838:exon12:c.C2252T:p.P751L
chr7	121651672	121651672	G	A	exonic	PTPRZ1	NM_002851:exon12:c.G2572A:p.A858T
chr7	121651751	121651751	A	G	exonic	PTPRZ1	NM_002851:exon12:c.A2651G:p.H884R
chr7	121651902	121651902	G	C	exonic	PTPRZ1	NM_002851:exon12:c.G2802C:p.E934D
chr7	121651952	121651952	A	C	exonic	PTPRZ1	NM_002851:exon12:c.A2852C:p.H951P

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr7	121652072	121652072	C	T	exonic	PTPRZ1	NM_002851:exon12:c.C2972T:p.S991F
chr7	121652919	121652919	G	T	exonic	PTPRZ1	NM_002851:exon12:c.G3819T:p.L1273F
chr7	121653226	121653226	A	G	exonic	PTPRZ1	NM_002851:exon12:c.A4126G:p.I1376V
chr7	121653363	121653380	TGATGATGATGGTGATGA	-	exonic	PTPRZ1	NM_002851:exon12:c.4263_4280del:p.1421_1427del
chr7	121653390	121653392	TGA	-	exonic	PTPRZ1	NM_002851:exon12:c.4290_4292del:p.1430_1431del
chr7	121653419	121653419	A	G	exonic	PTPRZ1	NM_002851:exon12:c.A4319G:p.H1440R
chr7	121653466	121653466	A	T	exonic	PTPRZ1	NM_002851:exon12:c.A4366T:p.M1456L
chr7	121653526	121653526	T	C	exonic	PTPRZ1	NM_002851:exon12:c.T4426C:p.S1476P
chr7	121653532	121653532	T	A	exonic	PTPRZ1	NM_002851:exon12:c.T4432A:p.S1478T
chr7	121653539	121653539	A	C	exonic	PTPRZ1	NM_002851:exon12:c.A4439C:p.N1480T
chr7	121653646	121653660	AATGATGGAAAAGAG	-	exonic	PTPRZ1	NM_002851:exon12:c.4546_4560del:p.1516_1520del
chr7	121653910	121653910	G	A	exonic	PTPRZ1	NM_002851:exon12:c.G4810A:p.E1604K
chr7	121659252	121659252	C	A	exonic	PTPRZ1	NM_001206838:exon13:c.C2338A:p.P780T
chr7	121678929	121678929	G	A	exonic	PTPRZ1	NM_001206838:exon18:c.G2887A:p.V963M
chr7	121680926	121680926	G	C	exonic	PTPRZ1	NM_001206839:exon20:c.G3093C:p.Q1031H
chr7	121682694	121682694	T	C	exonic	PTPRZ1	NM_001206839:exon21:c.T3233C:p.V1078A
chr7	121682702	121682702	A	G	exonic	PTPRZ1	NM_001206839:exon21:c.A3241G:p.S1081G
chr7	121684575	121684575	C	G	exonic	PTPRZ1	NM_001206839:exon22:c.C3436G:p.L1146V
chr7	121684599	121684599	A	G	exonic	PTPRZ1	NM_001206839:exon22:c.A3460G:p.T1154A
chr7	121691992	121691992	T	C	exonic	PTPRZ1	NM_001206839:exon24:c.T3650C:p.M1217T
chr7	121693982	121693982	G	A	exonic	PTPRZ1	NM_001206839:exon25:c.G3670A:p.E1224K
chr7	121695073	121695073	G	C	exonic	PTPRZ1	NM_001206839:exon26:c.G3859C:p.E1287Q
chr7	121695103	121695103	A	G	exonic	PTPRZ1	NM_001206839:exon26:c.A3889G:p.K1297E
chr7	121699928	121699928	G	A	exonic	PTPRZ1	NM_001206839:exon28:c.G4192A:p.V1398I
chr9	121929598	121929598	C	T	exonic	BRINP1	NM_014618:exon8:c.G2050A:p.G684S
chr9	121929812	121929812	G	T	exonic	BRINP1	NM_014618:exon8:c.C1836A:p.F612L
chr9	121929828	121929828	C	T	exonic	BRINP1	NM_014618:exon8:c.G1820A:p.R607Q
chr9	121930161	121930161	T	C	exonic	BRINP1	NM_014618:exon8:c.A1487G:p.D496G
chr9	121930416	121930416	C	T	exonic	BRINP1	NM_014618:exon8:c.G1232A:p.R411Q
chr9	121930417	121930417	G	A	exonic	BRINP1	NM_014618:exon8:c.C1231T:p.R411W
chr9	121971061	121971061	G	C	exonic	BRINP1	NM_014618:exon7:c.C1081G:p.R361G
chr9	121971096	121971096	G	A	exonic	BRINP1	NM_014618:exon7:c.C1046T:p.T349M
chr9	121971195	121971195	C	T	exonic	BRINP1	NM_014618:exon7:c.G947A:p.R316H
chr9	122004329	122004329	A	C	exonic	BRINP1	NM_014618:exon4:c.T575G:p.I192S
chr9	122011288	122011288	G	T	exonic	BRINP1	NM_014618:exon3:c.C359A:p.T120N
chr9	122011292	122011292	C	T	exonic	BRINP1	NM_014618:exon3:c.G355A:p.D119N
chr9	122011417	122011417	C	T	exonic	BRINP1	NM_014618:exon3:c.G230A:p.R77H
chr9	122075434	122075434	G	T	exonic	BRINP1	NM_014618:exon2:c.C200A:p.T67N
chr9	122075609	122075609	G	A	exonic	BRINP1	NM_014618:exon2:c.C25T:p.L9F
chr11	20622761	20622761	C	A	exonic	SLC6A5	NM_004211:exon2:c.C90A:p.C30X
chr11	20622796	20622796	-	G	exonic	SLC6A5	NM_004211:exon2:c.126dupG:p.A42fs
chr11	20622808	20622808	C	G	exonic	SLC6A5	NM_004211:exon2:c.C137G:p.P46R
chr11	20623018	20623018	A	G	exonic	SLC6A5	NM_004211:exon2:c.A347G:p.N116S
chr11	20623027	20623027	A	G	exonic	SLC6A5	NM_004211:exon2:c.A356G:p.H119R
chr11	20623159	20623159	C	T	exonic	SLC6A5	NM_004211:exon2:c.C488T:p.T163M
chr11	20625862	20625862	C	T	exonic	SLC6A5	NM_004211:exon3:c.C571T:p.R191X
chr11	20629163	20629163	C	T	exonic	SLC6A5	NM_004211:exon5:c.C950T:p.T317M
chr11	20636285	20636285	C	T	exonic	SLC6A5	NM_004211:exon6:c.C1046T:p.T349I
chr11	20639339	20639339	G	T	exonic	SLC6A5	NM_004211:exon7:c.G1169T:p.G390V
chr11	20657910	20657910	C	T	exonic	SLC6A5	NM_004211:exon11:c.C1682T:p.P561L
chr11	20658776	20658776	G	T	exonic	SLC6A5	NM_004211:exon12:c.G1796T:p.R599L
chr11	20673878	20673878	A	G	exonic	SLC6A5	NM_004211:exon15:c.A2114G:p.Y705C
chr11	20676316	20676316	C	T	exonic	SLC6A5	NM_004211:exon16:c.C2296T:p.R766C
chr11	20676319	20676319	G	A	exonic	SLC6A5	NM_004211:exon16:c.G2299A:p.G767R
chr11	20676326	20676326	G	A	exonic	SLC6A5	NM_004211:exon16:c.G2306A:p.R769H
chr1	227181970	227181970	C	A	exonic	CDC42BPA	NM_014826:exon35:c.G4916T:p.X1639L
chr1	227182055	227182055	TG	T	exonic	CDC42BPA	NM_014826:exon35:c.C4831A:p.L1611I
chr1	227182087	227182087	G	A	exonic	CDC42BPA	NM_014826:exon35:c.C4799T:p.P1600L
chr1	227192742	227192742	C	T	exonic	CDC42BPA	NM_014826:exon33:c.G4580A:p.R1527H
chr1	227204744	227204744	G	T	exonic	CDC42BPA	NM_014826:exon31:c.C4275A:p.D1425E
chr1	227211032	227211032	T	G	exonic	CDC42BPA	NM_014826:exon30:c.A4199C:p.N1400T
chr1	227218145	227218145	C	T	exonic	CDC42BPA	NM_014826:exon27:c.G3454A:p.E1152K
chr1	227219020	227219020	T	C	exonic	CDC42BPA	NM_014826:exon26:c.A3406G:p.T1136A
chr1	227219076	227219076	C	T	exonic	CDC42BPA	NM_014826:exon26:c.G3350A:p.R1117H
chr1	227223274	227223274	G	T	exonic	CDC42BPA	NM_014826:exon23:c.C2886A:p.N962K
chr1	227227874	227227874	A	C	exonic	CDC42BPA	NM_014826:exon22:c.T2809G:p.S937A
chr1	227259986	227259986	G	T	exonic	CDC42BPA	NM_014826:exon19:c.C2507A:p.S836X
chr1	227279603	227279603	G	A	exonic	CDC42BPA	NM_014826:exon15:c.C2096T:p.T699M
chr1	227300509	227300509	G	A	exonic	CDC42BPA	NM_003607:exon13:c.C1753T:p.R585W
chr1	227327390	227327390	T	G	exonic	CDC42BPA	NM_003607:exon10:c.A1277C:p.D426A
chr1	227333351	227333351	G	A	exonic	CDC42BPA	NM_003607:exon8:c.C982T:p.R328X
chr1	227348245	227348245	G	A	exonic	CDC42BPA	NM_003607:exon6:c.C692T:p.T231M
chr1	227348324	227348324	C	A	exonic	CDC42BPA	NM_003607:exon6:c.G613T:p.D205Y
chr1	227381616	227381616	T	C	exonic	CDC42BPA	NM_003607:exon5:c.A470G:p.Y157C
chr1	227387293	227387293	T	C	exonic	CDC42BPA	NM_003607:exon4:c.A415G:p.T139A
chr1	227400895	227400895	G	A	exonic	CDC42BPA	NM_003607:exon3:c.C296T:p.A99V
chr1	227504784	227504784	T	C	exonic	CDC42BPA	NM_003607:exon1:c.A100G:p.I34V
chr1	227504828	227504828	G	C	exonic	CDC42BPA	NM_003607:exon1:c.C56G:p.A19G

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr1	234527325	234527325	A	G	exonic	TARBP1	NM_005646:exon30:c.T4864C:p.X1622R
chr1	234529191	234529191	C	T	exonic	TARBP1	NM_005646:exon28:c.G4477A:p.V1493I
chr1	234529561	234529561	A	T	exonic	TARBP1	NM_005646:exon27:c.T4266A:p.D1422E
chr1	234536986	234536986	G	A	exonic	TARBP1	NM_005646:exon25:c.C4012T:p.R1338C
chr1	234541656	234541656	C	T	exonic	TARBP1	NM_005646:exon24:c.G3982A:p.G1328R
chr1	234541708	234541708	A	C	exonic	TARBP1	NM_005646:exon24:c.T3930G:p.D1310E
chr1	234546190	234546190	C	T	splicing	TARBP1	NM_005646:exon24:c.3792+1G>A
chr1	234561498	234561498	T	C	exonic	TARBP1	NM_005646:exon20:c.A3365G:p.K1122R
chr1	234564964	234564964	G	A	exonic	TARBP1	NM_005646:exon17:c.C2978T:p.A993V
chr1	234564977	234564977	C	T	exonic	TARBP1	NM_005646:exon17:c.G2965A:p.A989T
chr1	234565854	234565854	G	A	exonic	TARBP1	NM_005646:exon15:c.C2588T:p.A863V
chr1	234565857	234565857	T	G	exonic	TARBP1	NM_005646:exon15:c.A2585C:p.Y862S
chr1	234565893	234565893	G	A	exonic	TARBP1	NM_005646:exon15:c.C2549T:p.T850M
chr1	234573025	234573025	T	C	exonic	TARBP1	NM_005646:exon13:c.A2228G:p.N743S
chr1	234582548	234582548	C	T	splicing	TARBP1	NM_005646:exon13:c.2134+1G>A
chr1	234582615	234582615	T	C	exonic	TARBP1	NM_005646:exon12:c.A2068G:p.R690G
chr1	234586237	234586237	T	C	exonic	TARBP1	NM_005646:exon10:c.A1798G:p.T600A
chr1	234593436	234593436	G	A	exonic	TARBP1	NM_005646:exon9:c.C1699T:p.R567X
chr1	234595070	234595070	T	C	exonic	TARBP1	NM_005646:exon8:c.A1538G:p.D513G
chr1	234596122	234596122	T	C	exonic	TARBP1	NM_005646:exon7:c.A1420G:p.I474V
chr1	234601430	234601430	C	T	exonic	TARBP1	NM_005646:exon5:c.G1273A:p.A425T
chr1	234603391	234603391	A	G	exonic	TARBP1	NM_005646:exon4:c.T1105C:p.W369R
chr1	234606978	234606978	G	T	exonic	TARBP1	NM_005646:exon3:c.C1055A:p.P352Q
chr1	234613950	234613950	G	-	exonic	TARBP1	NM_005646:exon1:c.900delC:p.A300fs
chr1	234613952	234613952	C	A	exonic	TARBP1	NM_005646:exon1:c.G898T:p.A300S
chr1	234614018	234614018	C	G	exonic	TARBP1	NM_005646:exon1:c.G832C:p.A278P
chr1	234614123	234614123	C	A	exonic	TARBP1	NM_005646:exon1:c.G727T:p.E243X
chr1	234614126	234614126	G	A	exonic	TARBP1	NM_005646:exon1:c.C724T:p.P242S
chr1	234614128	234614128	A	C	exonic	TARBP1	NM_005646:exon1:c.T722G:p.L241W
chr1	234614138	234614138	C	A	exonic	TARBP1	NM_005646:exon1:c.G712T:p.E238X
chr1	234614140	234614140	G	A	exonic	TARBP1	NM_005646:exon1:c.C710T:p.A237V
chr1	234614155	234614155	A	G	exonic	TARBP1	NM_005646:exon1:c.T695C:p.V232A
chr1	234614239	234614239	C	A	exonic	TARBP1	NM_005646:exon1:c.G611T:p.G204V
chr1	234614257	234614257	A	C	exonic	TARBP1	NM_005646:exon1:c.T593G:p.V198G
chr1	234614273	234614273	C	G	exonic	TARBP1	NM_005646:exon1:c.G577C:p.G193R
chr1	234614366	234614366	G	T	exonic	TARBP1	NM_005646:exon1:c.C484A:p.R162S
chr1	234614416	234614416	G	A	exonic	TARBP1	NM_005646:exon1:c.C434T:p.A145V
chr1	234614462	234614462	C	G	exonic	TARBP1	NM_005646:exon1:c.G388C:p.A130P
chr1	234614692	234614692	C	T	exonic	TARBP1	NM_005646:exon1:c.G158A:p.G53D
chr1	234614810	234614810	G	A	exonic	TARBP1	NM_005646:exon1:c.C40T:p.R14W
chr9	214991	214991	G	A	exonic	C9orf66,DOCK8	NM_152569:exon1:c.C406T:p.R136W
chr9	271626	271626	G	T	splicing	DOCK8	NM_203447:exon2:c.54-1G>T
chr9	286572	286572	G	C	exonic	DOCK8	NM_001190458:exon2:c.G64C:p.D22H
chr9	289557	289557	G	A	exonic	DOCK8	NM_001190458:exon3:c.G176A:p.R59H
chr9	304628	304628	G	A	exonic	DOCK8	NM_001190458:exon4:c.G248A:p.R83Q
chr9	311966	311966	C	G	exonic	DOCK8	NM_001190458:exon5:c.C337G:p.H113D
chr9	311975	311975	G	A	exonic	DOCK8	NM_001190458:exon5:c.G346A:p.V116M
chr9	312005	312005	G	A	exonic	DOCK8	NM_001190458:exon5:c.G376A:p.V126I
chr9	312030	312030	G	A	exonic	DOCK8	NM_001190458:exon5:c.G401A:p.R134H
chr9	312088	312088	C	A	exonic	DOCK8	NM_001190458:exon5:c.C459A:p.D153E
chr9	312104	312104	G	A	exonic	DOCK8	NM_001190458:exon5:c.G475A:p.E159K
chr9	328022	328022	A	T	exonic	DOCK8	NM_001190458:exon8:c.A691T:p.I231F
chr9	328079	328079	G	A	exonic	DOCK8	NM_001190458:exon8:c.G748A:p.A250T
chr9	332443	332443	C	T	exonic	DOCK8	NM_001190458:exon9:c.C886T:p.P296S
chr9	332450	332450	C	T	exonic	DOCK8	NM_001190458:exon9:c.C893T:p.T298M
chr9	334292	334292	G	A	exonic	DOCK8	NM_001190458:exon10:c.G989A:p.R330Q
chr9	339033	339033	T	G	exonic	DOCK8	NM_001190458:exon12:c.T1246G:p.L416V
chr9	340230	340230	G	A	exonic	DOCK8	NM_001190458:exon13:c.G1384A:p.V462M
chr9	340265	340265	C	G	exonic	DOCK8	NM_001190458:exon13:c.C1419G:p.H473Q
chr9	368021	368021	C	A	exonic	DOCK8	NM_001190458:exon14:c.C1479A:p.N493K
chr9	372194	372194	A	T	exonic	DOCK8	NM_001190458:exon17:c.A1813T:p.I605F

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr9	377046	377046	G	A	exonic	DOCK8	NM_001190458:exon19:c.G2071A:p.V691M
chr9	379893	379893	G	A	exonic	DOCK8	NM_001190458:exon20:c.G2359A:p.V787I
chr9	379923	379923	G	A	exonic	DOCK8	NM_001190458:exon20:c.G2389A:p.V797M
chr9	379924	379924	T	C	exonic	DOCK8	NM_001190458:exon20:c.T2390C:p.V797A
chr9	382656	382656	G	A	exonic	DOCK8	NM_001190458:exon21:c.G2545A:p.E849K
chr9	382666	382666	A	G	exonic	DOCK8	NM_001190458:exon21:c.A2555G:p.K852R
chr9	386413	386413	C	G	exonic	DOCK8	NM_001193536:exon22:c.C2657G:p.P886R
chr9	396837	396837	G	A	exonic	DOCK8	NM_001190458:exon23:c.G2723A:p.R908Q
chr9	396872	396872	A	G	exonic	DOCK8	NM_001190458:exon23:c.A2758G:p.I920V
chr9	396927	396927	C	T	exonic	DOCK8	NM_001190458:exon23:c.C2813T:p.P938L
chr9	399233	399233	A	G	exonic	DOCK8	NM_001190458:exon24:c.A2908G:p.N970D
chr9	399245	399245	C	A	exonic	DOCK8	NM_001190458:exon24:c.C2920A:p.H974N
chr9	404995	404995	G	C	exonic	DOCK8	NM_001190458:exon25:c.G3012C:p.E1004D
chr9	405044	405044	A	G	exonic	DOCK8	NM_001190458:exon25:c.A3061G:p.T1021A
chr9	406999	406999	C	T	exonic	DOCK8	NM_001190458:exon26:c.C3160T:p.R1054C
chr9	414816	414816	A	G	exonic	DOCK8	NM_001190458:exon27:c.A3265G:p.I1089V
chr9	414863	414863	A	T	exonic	DOCK8	NM_001190458:exon27:c.A3312T:p.K1104N
chr9	414869	414869	G	C	exonic	DOCK8	NM_001190458:exon27:c.G3318C:p.E1106D
chr9	418080	418080	G	A	exonic	DOCK8	NM_001190458:exon28:c.G3413A:p.R1138H
chr9	418125	418125	G	T	exonic	DOCK8	NM_001190458:exon28:c.G3458T:p.G1153V
chr9	418210	418210	A	G	splicing	DOCK8	NM_001190458:exon28:c.3540+3A>G
chr9	420428	420428	G	A	exonic	DOCK8	NM_001190458:exon29:c.G3568A:p.A1190T
chr9	420470	420470	A	G	exonic	DOCK8	NM_001190458:exon29:c.A3610G:p.M1204V
chr9	420579	420579	A	G	exonic	DOCK8	NM_001190458:exon29:c.A3719G:p.Y1240C
chr9	426926	426926	A	G	exonic	DOCK8	NM_001190458:exon32:c.A3983G:p.N1328S
chr9	426984	426984	G	A	splicing	DOCK8	NM_001190458:exon32:c.4038+3G>A
chr9	428479	428479	C	T	exonic	DOCK8	NM_001190458:exon33:c.C4156T:p.R1386C
chr9	429732	429732	G	C	exonic	DOCK8	NM_001190458:exon34:c.G4204C:p.E1402Q
chr9	432202	432202	C	G	exonic	DOCK8	NM_001190458:exon35:c.C4363G:p.L1455V
chr9	434791	434791	A	G	exonic	DOCK8	NM_001190458:exon37:c.A4595G:p.K1532R
chr9	441328	441328	A	C	exonic	DOCK8	NM_001190458:exon39:c.A4966C:p.I1656L
chr9	443434	443434	A	G	exonic	DOCK8	NM_001190458:exon41:c.A5198G:p.Y1733C
chr9	443478	443478	T	C	exonic	DOCK8	NM_001190458:exon41:c.T5242C:p.S1748P
chr9	443481	443481	A	C	exonic	DOCK8	NM_001190458:exon41:c.A5245C:p.T1749P
chr9	446401	446401	A	G	exonic	DOCK8	NM_001190458:exon42:c.A5312G:p.Y1771C
chr9	446460	446460	C	T	exonic	DOCK8	NM_001190458:exon42:c.C5371T:p.R1791W
chr9	446500	446500	G	A	exonic	DOCK8	NM_001190458:exon42:c.G5411A:p.R1804Q
chr9	449794	449794	C	T	exonic	DOCK8	NM_001190458:exon43:c.C5528T:p.T1843I
chr9	463555	463555	C	G	exonic	DOCK8	NM_001190458:exon45:c.C5807G:p.T1936R
chr9	463579	463579	A	G	exonic	DOCK8	NM_001190458:exon45:c.A5831G:p.Q1944R
chr9	463653	463653	T	A	exonic	DOCK8	NM_001190458:exon45:c.T5905A:p.Y1969N
chr9	463655	463655	C	A	exonic	DOCK8	NM_001190458:exon45:c.C5907A:p.Y1969X
chr16	23400288	23400288	G	A	exonic	COG7	NM_153603:exon17:c.C2266T:p.R756C
chr16	23409402	23409402	C	T	exonic	COG7	NM_153603:exon14:c.G1852A:p.A618T
chr16	23415100	23415100	C	T	exonic	COG7	NM_153603:exon13:c.G1718A:p.R573Q
chr16	23417560	23417560	T	C	exonic	COG7	NM_153603:exon12:c.A1499G:p.Y500C
chr16	23417586	23417586	G	A	splicing	COG7	NM_153603:exon12:c.1476-3C>T
chr16	23428402	23428402	T	C	exonic	COG7	NM_153603:exon9:c.A1178G:p.H393R
chr16	23436124	23436124	C	T	exonic	COG7	NM_153603:exon7:c.G955A:p.D319N
chr16	23445997	23445997	C	T	exonic	COG7	NM_153603:exon5:c.G647A:p.R216Q
chr16	23453882	23453882	A	G	exonic	COG7	NM_153603:exon4:c.T520C:p.C174R
chr16	23453947	23453947	G	A	exonic	COG7	NM_153603:exon4:c.C455T:p.A152V
chr16	23456398	23456398	C	T	exonic	COG7	NM_153603:exon3:c.G406A:p.A136T
chr16	23456418	23456418	T	C	exonic	COG7	NM_153603:exon3:c.A386G:p.D129G

A. Supplement

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr16	23456437	23456437	C	T	exonic	COG7	NM_153603:exon3:c.G367A:p.E123K
chr16	23457211	23457211	A	C	exonic	COG7	NM_153603:exon2:c.T241G:p.S81A
chr16	23464206	23464206	G	A	exonic	COG7	NM_153603:exon1:c.C110T:p.A37V
chr16	23464242	23464242	G	A	exonic	COG7	NM_153603:exon1:c.C74T:p.S25F
chr16	23464278	23464278	A	T	exonic	COG7	NM_153603:exon1:c.T38A:p.V13E
chr16	23569457	23569457	T	G	exonic	UBFD1	NM_019116:exon2:c.T212G:p.V71G
chr16	23570830	23570830	G	A	exonic	UBFD1	NM_019116:exon3:c.G397A:p.V133I
chr16	23573958	23573958	A	G	exonic	UBFD1	NM_019116:exon5:c.A643G:p.T215A
chr16	23581849	23581849	G	A	exonic	UBFD1	NM_019116:exon7:c.G868A:p.V290I
chr16	30715431	30715431	C	T	exonic	SRCAP	NM_006662:exon4:c.C101T:p.S34L
chr16	30720835	30720835	T	G	exonic	SRCAP	NM_006662:exon7:c.T635G:p.V212G
chr16	30720838	30720838	T	G	exonic	SRCAP	NM_006662:exon7:c.T638G:p.V213G
chr16	30723050	30723050	A	G	exonic	SRCAP	NM_006662:exon11:c.A1477G:p.S493G
chr16	30723156	30723156	A	T	exonic	SRCAP	NM_006662:exon12:c.A1493T:p.D498V
chr16	30723222	30723222	G	A	exonic	SRCAP	NM_006662:exon12:c.G1559A:p.S520N
chr16	30723720	30723720	C	G	exonic	SRCAP	NM_006662:exon13:c.C1953G:p.I651M
chr16	30727520	30727520	C	G	exonic	SRCAP	NM_006662:exon17:c.C2627G:p.T876S
chr16	30727792	30727792	C	T	exonic	SRCAP	NM_006662:exon18:c.C2809T:p.P937S
chr16	30731486	30731486	A	G	exonic	SRCAP	NM_006662:exon19:c.A2821G:p.I941V
chr16	30731492	30731492	A	T	exonic	SRCAP	NM_006662:exon19:c.A2827T:p.M943L
chr16	30732218	30732218	C	T	exonic	SRCAP	NM_006662:exon20:c.C3172T:p.P1058S
chr16	30732510	30732510	T	C	exonic	SRCAP	NM_006662:exon21:c.T3254C:p.V1085A
chr16	30732548	30732548	C	T	exonic	SRCAP	NM_006662:exon21:c.C3292T:p.R1098W
chr16	30732558	30732558	C	A	exonic	SRCAP	NM_006662:exon21:c.C3302A:p.T1101K
chr16	30732626	30732626	T	C	exonic	SRCAP	NM_006662:exon21:c.T3370C:p.S1124P
chr16	30732644	30732644	C	T	exonic	SRCAP	NM_006662:exon21:c.C3388T:p.P1130S
chr16	30732651	30732651	C	T	exonic	SRCAP	NM_006662:exon21:c.C3395T:p.T1132I
chr16	30732773	30732773	C	G	exonic	SRCAP	NM_006662:exon21:c.C3517G:p.P1173A
chr16	30733599	30733599	C	T	exonic	SRCAP	NM_006662:exon22:c.C3698T:p.P1233L
chr16	30734045	30734045	A	G	exonic	SRCAP	NM_006662:exon23:c.A3868G:p.S1290G
chr16	30734343	30734343	C	T	exonic	SRCAP	NM_006662:exon24:c.C3952T:p.R1318W
chr16	30735100	30735100	C	T	exonic	SRCAP	NM_006662:exon25:c.C4355T:p.S1452L
chr16	30735322	30735322	T	G	exonic	SRCAP	NM_006662:exon25:c.T4577G:p.L1526W
chr16	30735333	30735333	A	G	exonic	SRCAP	NM_006662:exon25:c.A4588G:p.T1530A
chr16	30735424	30735424	C	T	exonic	SRCAP	NM_006662:exon25:c.C4679T:p.P1560L
chr16	30735699	30735699	A	G	exonic	SRCAP	NM_006662:exon25:c.A4954G:p.T1652A
chr16	30735951	30735951	C	G	exonic	SRCAP	NM_006662:exon25:c.C5206G:p.P1736A
chr16	30735993	30735993	C	T	exonic	SRCAP	NM_006662:exon25:c.C5248T:p.P1750S
chr16	30736045	30736045	C	T	exonic	SRCAP	NM_006662:exon25:c.C5300T:p.T1767M
chr16	30736149	30736149	G	A	exonic	SRCAP	NM_006662:exon25:c.G5404A:p.A1802T
chr16	30736343	30736345	GTT	-	exonic	SRCAP	NM_006662:exon25:c.5598_5600del:p.1866_1867del
chr16	30736370	30736370	-	C	exonic	SRCAP	NM_006662:exon25:c.5626dupC:p.Q1875fs
chr16	30740333	30740333	A	G	exonic	SRCAP	NM_006662:exon26:c.A5705G:p.E1902G
chr16	30745234	30745234	G	A	exonic	SRCAP	NM_006662:exon30:c.G6514A:p.V2172M
chr16	30748932	30748932	C	G	exonic	SRCAP	NM_006662:exon34:c.C7571G:p.S2524C
chr16	30749224	30749224	G	C	exonic	SRCAP	NM_006662:exon34:c.G7863C:p.E2621D
chr16	30749490	30749490	C	T	exonic	SRCAP	NM_006662:exon34:c.C8129T:p.S2710L
chr16	30749654	30749654	C	T	exonic	SRCAP	NM_006662:exon34:c.C8293T:p.R2765W
chr16	30749748	30749748	T	C	exonic	SRCAP	NM_006662:exon34:c.T8387C:p.M2796T
chr16	30749820	30749820	C	T	exonic	SRCAP	NM_006662:exon34:c.C8459T:p.P2820L
chr16	30749843	30749843	C	T	exonic	SRCAP	NM_006662:exon34:c.C8482T:p.R2828C
chr16	30749861	30749861	G	A	exonic	SRCAP	NM_006662:exon34:c.G8500A:p.G2834R
chr16	30749892	30749892	G	C	exonic	SRCAP	NM_006662:exon34:c.G8531C:p.G2844A
chr16	30749965	30749965	G	T	exonic	SRCAP	NM_006662:exon34:c.G8604T:p.R2868S
chr16	30750003	30750003	C	T	exonic	SRCAP	NM_006662:exon34:c.C8642T:p.P2881L
chr16	30750116	30750116	C	T	exonic	SRCAP	NM_006662:exon34:c.C8755T:p.L2919F
chr16	30750414	30750414	G	A	exonic	SRCAP	NM_006662:exon34:c.G9053A:p.R3018Q
chr16	30750498	30750498	G	A	exonic	SRCAP	NM_006662:exon34:c.G9137A:p.G3046D
chr16	30750681	30750681	C	A	exonic	SRCAP	NM_006662:exon34:c.C9320A:p.T3107K
chr16	30750729	30750729	G	A	exonic	SRCAP	NM_006662:exon34:c.G9368A:p.R3123H
chr16	30750788	30750788	G	T	exonic	SRCAP	NM_006662:exon34:c.G9427T:p.A3143S

A. Supplement

Chr	Start	End	Ref	Alt	Function	Gene	RefSeq Transcript ID & CDS change
chr16	30750848	30750848	C	G	exonic	SRCAP	NM_006662:exon34:c.C9487G:p.L3163V
chr16	30751010	30751010	G	A	exonic	SRCAP	NM_006662:exon34:c.G9649A:p.A3217T
chr16	30751031	30751031	A	G	exonic	SRCAP	NM_006662:exon34:c.A9670G:p.R3224G
chr18	45555810	45555810	G	A	exonic	ZBTB7C	NM_001039360:exon3:c.C1681T:p.R561C
chr18	45555900	45555900	C	T	exonic	ZBTB7C	NM_001039360:exon3:c.G1591A:p.A531T
chr18	45556010	45556010	C	T	exonic	ZBTB7C	NM_001039360:exon3:c.G1481A:p.G494E
chr18	45566421	45566421	A	C	exonic	ZBTB7C	NM_001039360:exon2:c.T1058G:p.V353G
chr18	45566448	45566448	C	T	exonic	ZBTB7C	NM_001039360:exon2:c.G1031A:p.G344E
chr18	45566619	45566619	C	T	exonic	ZBTB7C	NM_001039360:exon2:c.G860A:p.R287Q
chr18	45566811	45566811	C	T	exonic	ZBTB7C	NM_001039360:exon2:c.G668A:p.R223Q
chr18	45566927	45566927	G	C	exonic	ZBTB7C	NM_001039360:exon2:c.C552G:p.S184R
chr18	45566996	45566996	C	A	exonic	ZBTB7C	NM_001039360:exon2:c.G483T:p.E161D
chr18	45567002	45567002	-	TCC	exonic	ZBTB7C	NM_001039360:exon2:c.476_477insGGA:p.E159delinsEE
chr18	45567027	45567029	TCA	-	exonic	ZBTB7C	NM_001039360:exon2:c.450_452del:p.150_151del
chr18	45567084	45567084	T	C	exonic	ZBTB7C	NM_001039360:exon2:c.A395G:p.D132G
chr18	45567342	45567342	C	T	exonic	ZBTB7C	NM_001039360:exon2:c.G137A:p.R46Q
chr18	45567465	45567465	A	G	exonic	ZBTB7C	NM_001039360:exon2:c.T14C:p.I5T

B. Acknowledgment

First of all I would like to thank my supervisor Prof. Dr. Erdmann and the Institute of Cardiogenetics for giving me this opportunity and supporting me throughout my journey. Another massive thanks to Dr. Ingrid Braenne, who despite residing on the other side of the world, still found time to help me and guide me during my thesis, you are the best mentor anyone could ask for. Thank you to Bene another part of our amazing team. You made work fun and were always there to help.

I would also like to thank my family for believing in me and supporting me, my husband for enduring me during all this time and my little girls, who patiently waited to come into this world until I fulfilled my promise to myself and sent off my first draft.

To my friends: Thank you for sticking with me, helping me through mood swings and desperation. Especially : Liana, who took the pleasure in reading my whole work. Tegan, who also helped me fix up this thesis. Lana, who offered to read it and thankfully never had to. Mattes, the man who saved my citation list and spent hours with me trying to make latex work. And last but not least Sina, without who I would have never been able to hand it in. You are the best.

Curriculum Vitae

Personal Particulars

Name	Mariana Kleinecke
Date of Birth	October 28, 1988
Nationality	German
Academic Degree	Master of Science
Address	Wendische Straße 18
	23558 Lübeck (Germany)
E-Mail	mariana-barnes@outlook.com
Phone	+61 415044423

Education

2011 - 2014   Master's Program in Molecular Life Science, University of Lübeck, Germany

Master thesis   "Prevalence of familial hypercholesterolemia variants in the exome data of myocardial infarction patients" at the Institute for Integrative and Experimental Genomics of the University of Lübeck, directed by Prof. Dr. Jeanette Erdmann

2008 - 2011   Bachelor's program in Molecular Life Science, University of Lübeck, Germany

Bachelor thesis   "Analysis of the cell-matrix interaction of different melanoma cell lines" at the department for cell biologie and virology of the University of Lübeck, directed by Prof. Dr. Jürgen Brinckmann

1999 - 2008   Academic high school Glinde (Germany)
Advanced courses   Mathematics and Biology

Skills

Language Skills   German (native speaker)
English (fluent)
French (basic)

Software Application   Advanced Microsoft Office skills
Windows & Unix
Shell scripting (mainly AWK)
DNA-Sequencing pipelines (BWA, Picard, GATK)
Basic programming skills in JAVA, HTML, CSS, SQL

Interpersonal Skills   team minded, motivated, determined, independent, adapting, positive attitude

Working Experience

November 2016 – current
High Performance Computing Officer and Researcher at Menzies School of Health Research
- o Software and account manager for the High Performance Computer
- o Short-read and long-read analysis of bacterial genomes in particular *Burkholderia Pseudomallei*

July 2014 – current
PhD student at the Institute for Integrative and Experimental Genomics of the University of Lübeck
- o DNA-sequencing pipeline
- o Co-segregational studies

November-December 2012
Internship at school of Medicine and Pharmacology of the University of Western Australia
- o Cell culture
- o Matrix-degradation assays

August-October 2012
Internship at the institute for human genetics of the University of Hamburg
- o Detection and evaluation of detected genetic variants
- o Functional analyses in cell culture

March - August 2012
Student assistant of the department of medicine II, university hospital Schleswig-Holstein
- o Collecting data for a scientific study about risk factors for myocardial infarction

October 2010 – March 2011
Voluntary member of the student council for molecular life science and computer science
- o Represent students to improve the university- student relationship

August 2010
Internship at the department for Biochemistry of the University of Lübeck in the field of crystallography
- o Protein purification
- o Crystal growth

2009-2014
Voluntary member of P++, a section of culture, environment and sport of the students' union executive committee
- o Planning and organizing recreational activities for students

Publications

Giffard PM, Andersson P, Wilson J, Buckley C, Lilliebridge R, Harris TM, **Kleinecke M**, O'Grady KF, Huston WM, Lambert SB, Whiley DM, Holt DC. CtGEM typing: Discrimination of Chlamydia trachomatis ocular and urogenital strains and major evolutionary lineages by high resolution melting analysis of two amplified DNA fragments.
PLoS One. 2018 Apr 10;13(4):e0195454. doi: 10.1371/journal.pone.0195454. eCollection 2018

Brænne I, Willenborg C, Tragante V, Kessler T, Zeng L, Reiz B, **Kleinecke M**, von Ameln S, Willer CJ, Laakso M, Wild PS, Zeller T, Wallentin L, Franks PW, Salomaa V, Dehghan A, Meitinger T, Samani NJ, Asselbergs FW, Erdmann J, Schunkert H. A genomic exploration identifies mechanisms that may explain adverse cardiovascular effects of COX-2 inhibitors
Sci Rep. 2017 Aug 31;7(1):10252. doi: 10.1038/s41598-017-10928-4

Brænne I*, **Kleinecke M***, Reiz B, Graf E, Strom T, Wieland T, Fischer M, Kessler T, Hengstenberg C, Meitinger T, Erdmann J, Schunkert H. Systematic analysis of variants related to familial hypercholesterolemia in families with premature myocardial infarction.
Eur J Hum Genet. 2016 Feb;24(2):191-7. doi: 10.1038/ejhg.2015.100. Epub 2015 Jun 3.
*equal contribution

.
Brænne I, Reiz B, Medack A, **Kleinecke M**, Fischer M, Tuna S, Hengstenberg C, Deloukas P, Erdmann J, Schunkert H; Cardiogenics consortium. Whole-exome sequencing in an extended family with myocardial infarction unmasks familial hypercholesterolemia
BMC Cardiovasc Disord. 2014 Aug 26;14:108. doi: 10.1186/1471-2261-14-108.