BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

BACHELOR'S THESIS

Brno, 2021 Klára Kocmanová

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

GENETIC VARIATIONS IN CORONARY ARTERY DISEASE
GENETICKÉ VARIACE U ONEMOCNĚNÍ KORONÁRNÍCH TEPEN

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — **Klára Kocmanová**
AUTOR PRÁCE

SUPERVISOR — **prof. Ing. Ivo Provazník, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2021

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Bachelor's Thesis

Bachelor's study program **Biomedical Technology and Bioinformatics**

Department of Biomedical Engineering

Student: Klára Kocmanová ***ID:*** 211660

Year of study: 3 ***Academic year:*** 2020/21

TITLE OF THESIS:

Genetic variations in coronary artery disease

INSTRUCTION:

1) Study monogenetic and polygenetic architectures in cardiovascular diseases. Include the results of loci mapping in large-scale genomic projects in the literature research. 2) Study the mapped and described molecular mechanisms of cardiovascular diseases. 3) Get familiar with a structure of the 1000 Genomes Project and Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Consortium Project databases. Design a procedure to read and process data from these databases. 4) Design and perform bioinformatics, statistical and population analysis of data leading to the search for genetic variations in coronary artery disease. 5) Appropriately graphically present and discuss results of the analysis.

RECOMMENDED LITERATURE:

[1] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073, 2010.

[2] CARDIoGRAMplusC4D Consortium. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 47(10): 1121-30, 2015.

Date of project specification: 8.2.2021 ***Deadline for submission:*** 2.6.2021

Supervisor: prof. Ing. Ivo Provazník, Ph.D.

doc. Ing. Jana Kolářová, Ph.D.
Chair of study program board

Abstract

This bachelor's thesis deals with the genetics of coronary artery disease. Studying risk variants helps us identify risk patients and allows us to develop targeted medicine. The first chapter deals with the anatomy, histology, and pathophysiology of CAD. The second chapter focuses on the current knowledge of the genetics of coronary artery disease. The third chapter mentions two genomic projects that have contributed to studying variations in coronary artery disease and the means of reading data from their databases. These projects are the 1000 Genomes Project and CARDIoGRAM plus CD4 consortium. 1000 Genomes-based linkage disequilibrium reference panel was obtained. The selection operator for jointly analyzing multiple variants was applied to the loci discovered in 1000 Genomes-based genome-wide association meta-analysis performed by CARDIoGRAM plus CD4 consortium.

Keywords

Coronary artery disease, genetic variants, 1000 Genomes Project, CARDIoGRAM plus CD4 consortium, allelic heterogeneity

Abstrakt

Tato bakalářská práce se zabývá genetikou onemocnění koronárních tepen. Díky studiu rizikových genetických variací je možné identifikovat ohrožené pacienty a vyvíjet cílenou léčbu. V první kapitole se zmiňuje anatomie, histologie a patofyziologie koronárních tepen. Druhá kapitola je zaměřena na dosavadní znalosti dědičnosti onemocnění koronárních tepen. Ve třetí kapitole jsou zmíněny dva genomické projekty, které přispěly ke studiu variací u onemocnění koronárních tepen, a způsoby čtení dat z jejich databází. Těmito projekty jsou 1000 Genomes Project a CARDIoGRAM plus CD4 consortium. Na základě dat z 1000 Genomes databáze byl získán referenční panel vazebné nerovnováhy. Dále byla provedena asociační analýza na základě dat ze souhrnné statistiky metaanalýzy provedené CARDIoGRAM plus CD4 konsorciem.

Klíčová slova

Onemocnění koronárních tepen, genetické variace, 1000 Genomes Project, CARDIoGRAM plus CD4 consortium

Rozšířený abstrakt

Onemocnění koronárních tepen patří mezi nejčastější příčiny úmrtí ve vyspělých státech. Rozvoj tohoto onemocnění je podmíněn mnoha faktory. Mezi rizikové faktory onemocnění koronárních tepen patří například nedostatek pohybu, dieta s vysokým obsahem lipidů, hypertenze, diabetes mellitus nebo kouření. Bylo prokázáno, že onemocnění koronárních tepen je geneticky podmíněné.

Onemocnění koronárních tepen je způsobeno tvorbou aterosklerotických plátů v koronárních tepnách. Na počátku ateroskleróza dochází k narušení endoteliálních buněk ve stěně cévy. To může být způsobeno mnoha faktory, například působením volných radikálů, toxických látek, zvýšeným tlakem nebo mediátory zánětu. Do stěny porušené cévy začnou pronikat částice LDL cholesterolu. Zde jsou oxidovány. Oxidované LDL částice přitahují monocyty, které difundují do cévní stěny, přeměňují se na makrofágy a fagocytují LDL částice. Dochází ke stimulaci dalších typů leukocytů a k rozvoji zánětu. U makrofágů pohlcujících LDL částice poté dochází k nekróze. Do nekrotického ložiska proliferují buňky hladké svaloviny. Tyto buňky produkují kolagenová vlákna a extracelulární matrix. Kolem nekrotického ložiska se vytvoří fibrózní aterosklerotický plát. Obsah nekrotického ložiska má trombogenní vlastnosti. Pokud dojde k ruptuře plátu, obsah ložiska se dostane do kontaktu s krví a začnou se vytvářet tromby. Následkem vzniku trombózy dochází k okluzi koronární tepny.

Ačkoliv se onemocnění koronárních tepen projevuje převážně v pozdějším věku, aterosklerotické pláty se mohou zakládat už v raných stádiích vývoje člověka. Při pokročilé ateroskleróze může dojít k srdečnímu selhání následkem chronické ischemie myokardu, způsobené ztrátou elasticity a částečnou okluzí koronární tepny. Pacienty s onemocněním koronárních tepen však nejvíce ohrožuje riziko ruptury aterosklerotického plátu a následná trombóza, způsobující akutní koronární syndrom.

Onemocnění koronárních tepen má polygenetický charakter. Vznik tohoto onemocnění je podmíněný přítomností několika variací, z nichž každá má na výsledný fenotyp jen malý vliv. Čím více rizikových alel jedinec zdědí, tím je větší pravděpodobnost, že se u něj onemocnění rozvine. Polygenetické onemocnění lze zkoumat pomocí celogenomové asociační studie. Při celogenomové asociační studii jsou vybráni jedinci do skupiny případů a skupiny kontrol. Frekvence výskytu jednotlivých variací je poté mezi

skupinami porovnávána. Rizikové variace onemocnění koronárních tepen lze očekávat v genech kódujících proteiny, které se účastní aterosklerózy, například proteiny spojené s metabolismem lipidů, buněčnou proliferací a zánětlivými nebo trombogenními procesy. Díky celogenomovým asociačním studiím však byly objeveny také variace, jejichž mechanismus působení při tomto onemocnění nebyl dosud objasněn.

Rozvoj celogenomových asociačních studií byl výrazně usnadněn díky projektům zabývajícím se mapováním genetických variací. Jedním z takovýchto projektů byl 1000 Genomes project. Cílem tohoto projektu bylo osekvenovat alespoň 1000 jedinců a vytvořit referenční databázi genetických variací. Tato databáze slouží jako reference genetických variací a může být použita při imputaci variací v celogenomové asociační studii. Mapováním genetických variací spojených s onemocněním koronárních tepen se zabývá CARDIoGRAM plus CD4 consortium. V rámci tohoto projektu proběhlo již 7 metaanalýz celogenomových asociačních studií zabývajících se onemocněním koronárních tepen.

V rámci této práce jsou analyzována data, která byla stažena z databáze CARDIoGRAM plus CD4 consortium. Jedná se o souhrnnou statistiku metaanalýzy, jejíž výsledky byly publikovány v časopise Nature genetics v roce 2015 pod názvem A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Do této metaanalýzy bylo zahrnuto 48 studií, kterých se účastnilo celkem 60801 případů a 123504 kontrol. Studií se účastnilo 77 % jedinců z evropské populace, 13 % jedinců z populací v jižní Asii a 6 % jedinců z populací ve východní Asii. Zbytek jedinců pocházel z hispánské a africko-americké populace.

Na základě dat z 1000 Genomes databáze byly pro vybrané lokusy sestaveny matice vazebné nerovnováhy pomocí LDlink programu. Pro analýzu byly použity pouze lokusy, ve kterých se nacházela více než jedna signifikantní variace. Pro výpočet vazebné nerovnováhy byly použity populace evropské, asijské, hispánské z LA a africko-americké. Byly staženy matice korelačních koeficientů r^2 vazebné nerovnováhy a jejich vizualizace. Z vizualizací bylo patrné, že v některých lokusech byly variace silně korelované. Tyto lokusy nebyly dále analyzovány. V jiných lokusech se nacházelo několik variací, které s ostatními nekorelovaly. Dá se předpokládat, že v těchto lokusech se nachází více rizikových variací. V jiných lokusech se tvořilo více skupin, ve kterých spolu variace korelovaly. V těchto lokusech by se mohlo nacházet více genů.

Ve vybraných lokusech byla provedena analýza alelické heterogenity. K analýze byl použit sestavený referenční panel vazebné nerovnováhy. Lokusy, ve kterých parametr r^2 vazebné nerovnováhy dosahoval mezi všemi variacemi hodnoty větší než 0.9, a tudíž vykazovaly vysokou míru korelace, nebyly do analýzy alelické heterogenity zahrnuty. V těchto místech byla za rizikovou variaci označena variace s nejnižší p hodnotou. Jedná se o lokus 1 na chromozomu 4, lokus na chromozomu 18 a lokus na chromozomu 22 v aditivním datasetu a lokus na chromozomu 2 v recesivním datasetu. Na ostatní lokusy byla aplikována LASSO analýza pomocí balíčku SOJO v r. Na základě LASSO analýzy bylo identifikováno 371 variací v aditivním datasetu a 55 variací v recesivním datasetu.

Variace byly dále anotovány pomocí prediktoru efektu variací Ensembl. Téměř 50 % variací v aditivním i recesivním datasetu se nacházelo v oblasti intronů. V aditivním datasetu bylo identifikováno 68 genů a 10 pseudogenů. V recesivním datasetu bylo identifikováno 16 genů. 10 genů se nacházelo v obou datasetech. Celkem tedy bylo identifikováno 74 genů spojených s onemocněním koronárních tepen.

Bibliographic citation:

KOCMANOVÁ, Klára. *Genetické variace u onemocnění koronárních tepen* . Brno, 2021. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/134399. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Ivo Provazník.

Declaration

I hereby declare that I have written the Bachelor's Thesis titled "Genetic variations in coronary artery disease" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author, I furthermore declare, that concerning the creation of this Bachelor's Thesis I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation S 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno:

……

……

author's signature

Acknowledgment

I would like to express my very great appreciation to my advisor, Professor Ivo Provazník, for his patient guidance during the planning and development of this thesis.

Contents

INTRODUCTION

Coronary artery disease (CAD) is the leading cause of death worldwide. It is strongly associated with a sedentary lifestyle and a high-fat diet. It is caused by an accumulation of low-density lipoprotein particles in the inner layer of coronary arteries, which results in thickening and loss of elasticity of the vessels. This leads to the insufficient blood supply and ischemia of the myocardium.

Studies have proven CAD clusters in families, thus having heritable components. Multiple genome-wide association studies (GWAS) were performed to discover the genetic nature of the disease. Currently, there are over fifty suspicious loci on chromosomes, which are linked with an increased risk of CAD. However, these explain only a small percentage of CAD heritability. Studying genetic predispositions to CAD helps to better understand the mechanisms and to discover new pathways of CAD. Risk patients can be identified and treated appropriately. It also allows to development of more targeted medicine.

In the first chapter of this thesis anatomy, histology, and pathophysiology of coronary arteries will be described. The second chapter will explain the known principles of CAD heredity. Monogenic and polygenic causes of CAD will be described and the importance of sex and ethnicity will be mentioned. In the third chapter the 1000 Genomes project and CARDIoGRAM plus CD4 consortium and accessing data from 1000 Genomes project database and CARDIoGRAM plus CD4 database will be introduced. Allelic heterogeneity will be investigated on the summary statistics data from CARDIoGRAM plus CD4 Consortium database.

1 CORONARY ARTERY DISEASE

Coronary artery disease (CAD) is a condition in which the coronary arteries become rigid and narrow, causing insufficient blood flow to the myocardium. It has one of the highest morbidity and mortality rates in developed countries, despite the dramatic drop of premature CAD death rates over the past 40 years.

The pathophysiological process of CAD is atherosclerosis. Although previously considered a cholesterol disorder, it is now viewed as an inflammatory process. It is caused by the deposition of low-density lipoprotein (LDL) particles in the subendothelial space of tunica intima and subsequent accumulation of monocytes, leading to inflammation. The whole process results in a formation of an atherosclerotic plaque.

Both genetic and environmental factors can lead to CAD. Familiar studies have proven the clustering of CAD in families. However, certain habits, which are known to lead to the development of CAD, such as diet, smoking, or physical activity, are passed on in families. Therefore, it has to be differentiated whether a condition is caused by a genetic disorder or by the lifestyle children adopt from their parents. Patients suffering from the symptoms of metabolic syndrome, such as obesity, high blood pressure, and diabetes mellitus, also show an increased risk of the development of CAD.

1.1 Anatomy and histology of coronary arteries

Coronary arteries are vessels that nourish the myocardium. Two arteries are protruding from the aortic sinus, arteria coronaria dextra and arteria coronaria sinistra. Arteria coronaria dextra nourishes the right atrium, part of the left atrium, and right ventricle wall. Arteria coronaria sinistra divides into ramus interventricularis anterior and ramus circumflexus. Ramus interventricularis anterior nourishes the front left ventricle wall. Ramus circumflexus nourishes the left atrium and the back-left ventricle wall. (*Figure 1*) [33]

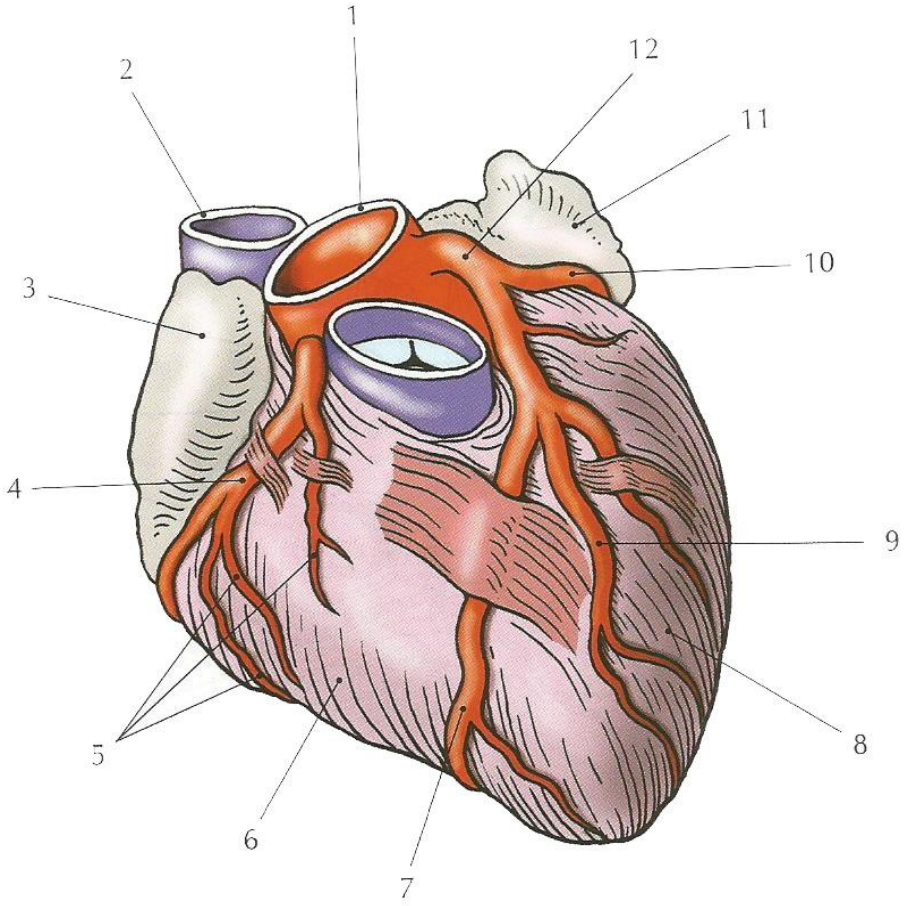


Figure 1: Anatomy of coronary arteries

[33]

1 - Aorta, 2 - v. cava superior, 3 - right atrium, 4 - a. coronaria dx, 5 - aa. protruding from a. coronaria dx., 6 - right ventricle, 7 - ramus interventricularis anterior, 8 - left ventricle, 9 - aa. protruding from ramus interventricularis anterior, 10 - ramus circumflexus, 11 - left atrium, 12 - a. coronaria sin.

The wall of the coronary artery consists of three layers. The inner layer is tunica intima. It is made of a layer of endothelial cells, which are attached to the internal elastic lamina. These cells provide a smooth surface that facilitates blood flow. The middle layer, tunica media, consists of smooth muscle tissue, allowing arteries to contract. It gives the artery its elastic properties. The outer layer, tunica adventitia, is made of fibrous tissue. It supports and protects the vessel. [39]

1.2 Pathophysiology of coronary artery disease

Clinical manifestations of coronary artery disease

Clinical manifestations of CAD depend on the severity and location of the narrowing of the artery. Although they usually appear later in life, the pathological process can start during fetal development. Occlusion in more proximal arteries results in a greater ischemic region. Partial occlusion leads to stable angina pectoris. Heart Failure may also occur due to long-term ischemia caused by reduced perfusion. Total occlusion results in acute coronary syndrome. [29]

Stable angina pectoris

Stable angina pectoris is caused by an insufficient blood supply resulting in reversible myocardial ischemia. It is characterized by a paroxysmal pain located behind the sternum, that might radiate to arms, neck, or jaw, and usually appears during exertion. At rest, the pain subsides after a couple of minutes. Other symptoms, such as sweating, nausea, or breathlessness might be present. Stable angina pectoris usually manifests, when the coronary arteries are severely occluded by atherosclerotic plaques. The myocardium most affected by the hypoxia is in the subendocardial region. [29]

Acute coronary syndrome

Acute coronary syndrome (ACS) manifests with similar symptoms as stable angina pectoris, however, these symptoms do not subside. It is a life-threatening condition, that is caused by a severe occlusion of a coronary artery, which, if not treated, leads to irreversible damage. ACS can be divided into unstable angina pectoris, myocardial infarction without ST elevation (nSTEMI), and myocardial infarction with ST elevation (STEMI). In unstable angina pectoris and nSTEMI, the perfusion is partially restored after occlusion. Patients with unstable angina pectoris and nSTEMI are treated with thrombolytics. In STEMI the perfusion is not restored. On the ECG recording, STEMI is manifested as an elevation of the ST segment on the leads corresponding to the affected region. Patients with STEMI require an emergency angioplasty. [29]

ACS occurs when the atherosclerotic plaque ruptures. This results in the formation of a thrombus. Thrombosis on the surface of the atherosclerotic plaque itself may not cause a complete blockage of the artery, however, a portion of the thrombus may tear off and embolize into the distal artery. In the early stages, the changes in the tissue are reversible, in later stages, irreversible damage occurs due to necrosis of the myocardium. [29]

In two-thirds of patients, ACS is caused by rupture of the plaque, that is not occluding the artery itself. Therefore, these patients do not experience symptoms before the incident. The consequences of ACS depend on the extent of the damaged tissue and the duration of ischemia. For 25% of the patients, ACS is fatal. In the surviving patients, coagulation scar forms at the site of the damaged region. The scar tissue may cause arrhythmias. If the necrotic region ruptures, before the scar is formed, subsequent bleeding into the pericardial cavity causes a cardiac tamponade. [29]

Heart Failure

Heart Failure usually occurs after myocardial infarction. However, it can be caused by severe CAD itself due to long-term ischemia of the myocardium. As a result of the damage of the myocardium, the heart is unable to pump enough blood to meet the metabolic demand. The decreased tissue perfusion is compensated by an increase in blood pressure. This increases the demands on the myocardium, thus causing further damage to the heart. Dilated cardiomyopathy might be present. Symptoms of Heart Failure include fatigue, shortness of breath, swelling, and palpitations. [29]

Molecular mechanisms of atherosclerosis

The process of atherosclerosis begins with disruption in the layer of endothelial cells of the artery. The disruption usually begins in regions with irregular, slow blood flow. Some molecules, such as high concentration of glucose, inflammation mediators, or cigarette smoke toxins, can stick to the vessel walls and negatively affect their condition. This can also be caused by viral or bacterial infection. [2] [12] [15] [24] [37]

LDL cholesterol particles penetrate the subendothelial space, where they are oxidized. The oxidized LDL particles act chemotactically on blood monocytes. In subendothelial space, macrophages differentiate from monocytes. These macrophages

phagocytize LDL cholesterol particles and turn themselves into foam cells. They also produce proinflammatory cytokines, such as tumor necrosis factor and interleukins. Other types of leukocytes accumulate, thus leading to chronic infection of tunica intima.[2] [12] [15] [24] [37]

Smooth muscle cells then migrate from tunica media. Those cells, along with collagen and elastin molecules, form a fibrous cap over the lipid-rich atherosclerotic lesion, resulting in a formation of a fibrous plaque. Calcification can occur in the necrotic lesion. It is caused by intracellular microcalcifications in SMCs. If necrosis occurs in these SMCs, calcium deposits in the extracellular space. The thickness of the fibrous cap determines whether the plaque is stable or vulnerable. Vulnerable plaque is prone to rupture. The rupture of the plaque causes thrombosis. [2] [12] [15] [24] [37]

Oxidative stress

Oxidation plays an important role both in endothelial dysfunction and in LDL modification. LDL cholesterol particles penetrate the subendothelial space, where they are oxidized. The oxidized LDL particles are easier to uptake by macrophages than the unoxidized ones. The oxidation is caused by reactive oxygen species (ROS), that come from both extracorporal and intracorporal sources. [46]

Important intracorporal sources of ROS are oxidative enzymes, such as nicotinamide adenine dinucleotide phosphate oxidase, nitric oxide synthase, xanthine oxidase, myeloperoxidase, and lipoxygenases. Nicotinamide adenine dinucleotide phosphate oxidase is a membrane-bound enzyme, which, under normal circumstances, is inactive. When activated, it catalyzes NADPH oxidation. The by-product of this reaction is a free radical. Xanthine oxidase catalyzes the oxidation of xanthine to uric acid. Under physiological conditions, nitric oxide synthase catalyzes NO synthesis. If the levels of the cofactor tetrahydrobiopterin lower, it can lead to oxygen reduction and production of ROS and peroxynitrite. Myeloperoxidase is a lysosomal enzyme produced by activated phagocytes. It catalyzes the synthesis of hypochlorous acid from hydrogen peroxide. It also oxidizes tyrosine into tyrosyl radical. Both substances are under physiological conditions used to kill pathogens. [46]

Inflammation mediators

The mediators of inflammation play an important role in the development of CAD. Atherosclerosis was previously considered a lipid metabolism disorder. The inflammatory process was thought to be triggered by the deposition of oxidized LDL particles in subendothelial space. However, it has been shown, that inflammatory mediators can also be involved in vascular wall disruption. For example, histamine is thought to loosen junctions between endothelial cells, thus increasing permeability for LDL particles. [14]

The first leukocytes associated with CAD were monocytes. Monocytes are commonly found in arteries even under physiological conditions. They adhere to the fatty streak lesions and turn into macrophages. The macrophages produce proinflammatory cytokines, participate in lipid retention and express pattern recognition receptors (PRRs), which mediate the triggering of an antigen-specific response by the immune system. Oxidized lipoprotein and heat shock protein-specific T lymphocytes are also found in atherosclerotic lesions. They produce antibodies against oxidized LDL particles and inflammatory cytokines and are thought to regulate B lymphocytes' function. Recent studies have shown that B lymphocytes have a role in directing the inflammatory response. [14]

Smooth muscle cells

Smooth muscle cells (SMCs) in the tunica media physiologically produce a range of SMC markers, such as cyclin D and cyclin E[63] and transcription factors E2F1-3. Due to a phenotypic change in SMCs, the levels of SMC markers decrease. The decrease results in increased proliferation, migration to the subendothelial space, and production of extracellular matrix proteins and cytokines. In the forming atherosclerotic plaque apoptosis of SMCs occurs, both in the necrotic lesion and in the plaque. The apoptosis is probably induced by the release of death-inducing markers from macrophages. SMCs, along with collagen and elastin molecules, form a fibrous cap over the lipid-rich atherosclerotic lesion, resulting in a formation of a fibrous plaque. In later stages of atherosclerosis, SMCs are thought to protect the atherosclerotic plaque from rupturing and promote plaque repair. [5]

The risk of thrombosis

If the atherosclerotic lesion grows inwards, the artery lumen is narrowed. However, in most cases, the lesion grows outwards, thus not creating stenosis. (*Figure 2*). Although there is no narrowing of the vessel, the non-stenotic plaque also poses a risk to a patient. If the fibrous cap of the vulnerable plaque is ruptured, the lipid core is exposed to the artery lumen. The lipid core has thrombogenic properties. If it comes to contact with blood, a thrombus can be formed. On stable plaques, a thrombus may also be formed, due to the irregularity in the arterial lumen. [1] [24]

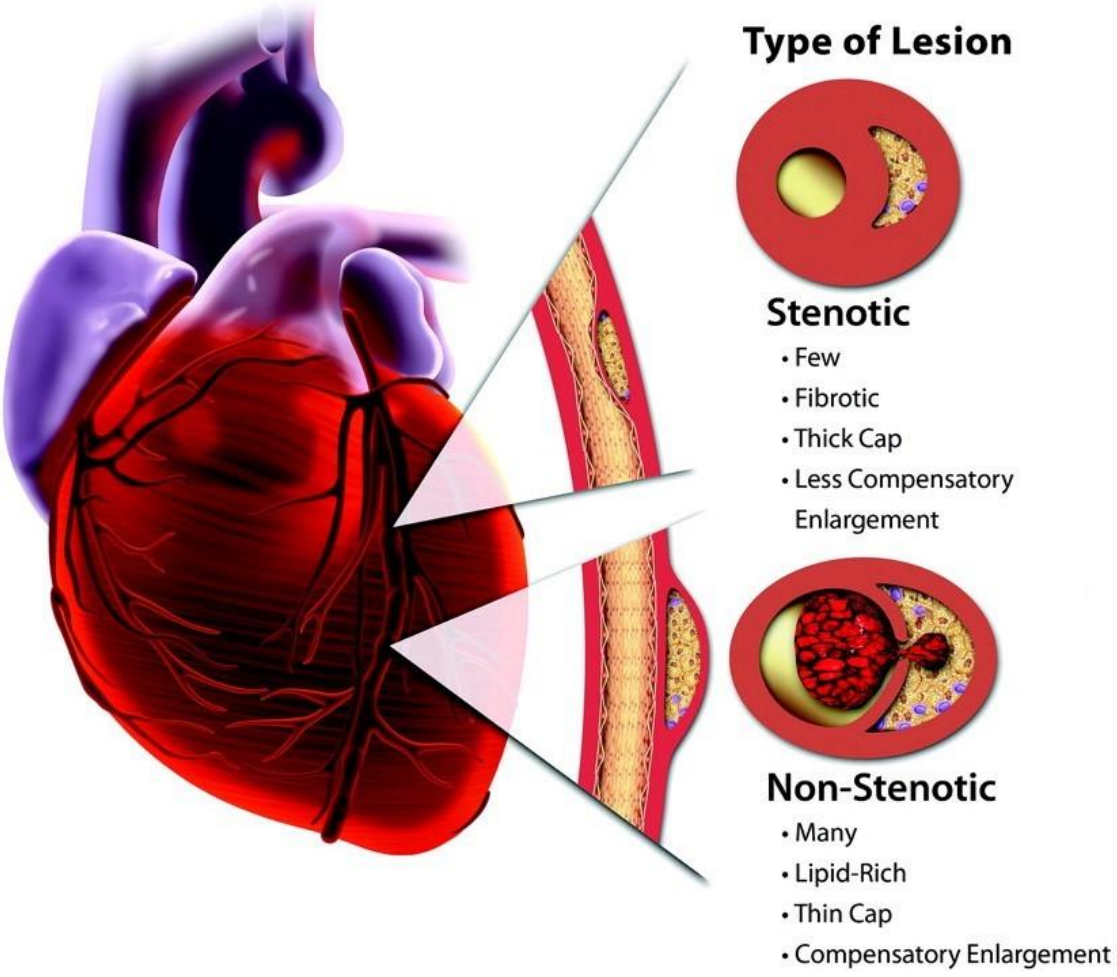


Figure 2: Simplified schema of a diversity of lesions in human coronary atherosclerosis

[24]

The formation of a thrombus is determined by multiple factors. Factors connected with the ruptured plaque are called solid-state determinants. These factors include tissue factor and collagen. The tissue factor produced by smooth muscle cells and macrophages triggers a coagulation cascade. Platelets are activated by the contact with collagen present in fibrous plaque. The thrombogenic factors in plasma are called fluid phase determinants. Based on fluid phase factors high-risk patients can be defined, who are prone to thrombosis. The two major fluid phase determinants are fibrinogen and plasminogen activator inhibitor-1 (PAI-1), which is an inhibitor of fibrinolysis mediators. Both solid-state determinants and fluid phase determinants can be altered by inflammatory mediators. [24]

Coronary artery disease risk factors

Obesity

Obesity is characterized by an excessive accumulation of fat tissue in visceral and subcutaneous regions. Poor eating habits, characterized by an increased lipid intake, result in increased levels of plasma LDL cholesterol particles and lipoprotein clearance dysfunction. The adipose tissue produces various adipokines, which are cell-signaling proteins. Some adipokines, such as adiponectin, have anti-inflammatory properties, some, such as resistin and leptin, are pro-inflammatory. The excessive amount of adipose tissue causes an imbalance between proinflammatory and anti-inflammatory cytokines, which results in vascular inflammation and endothelial cell dysfunction. Higher levels of resistin were measured in obese patients. Resistin stimulates the synthesis of proinflammatory cytokines. The adipokine leptin is responsible for food intake control. Levels of blood leptin are increased with overfeeding. Due to the similarity in the structure of leptin and proinflammatory cytokines, higher levels of leptin may stimulate T-cell proliferation. [27]

Hypertension

System blood pressure is regulated by antidiuretic hormone and the renin-angiotensin-aldosterone system (RAAS). Blood pressure can be lowered by the production of atrial natriuretic peptide (ANP) by the heart. At the local level pressure is regulated by endothelial cells in vessel walls. Endothelial cells produce nitric oxide through the l-arginine nitric oxide system. The production of nitric oxide contributes to vessel relaxation, thus decreasing blood pressure. [19]

Hypertension is a major factor in CAD, stroke, and renal insufficiency. The main effect on CAD development is mechanical stress. The mechanical stress aggravates atherosclerosis and facilitates plaque rupturing. Increased vascular permeability and biosynthesis of collagen, elastin, and mucopolysaccharides have been reported in patients with hypertension, along with increased DNA synthesis, which facilitates proliferation. Increased numbers of SMCs are present in thickened intima and media of hypertensive arteries. Increased tension simulates vessel injury, which results in the production of tissue and plasma mediators of vascular injury, as well as platelet aggregation. [19]

Diabetes mellitus

Around 80 % of diabetic patients die from complications linked with atherosclerosis. High glucose concentrations cause alterations of the vessel tissue. One of the atherogenic mechanisms linked with hypoglycemia is the nonenzymatic glycosylation of proteins and lipoproteins in arterial walls. The process of glycosylation in proteins causes conformation changes and affects their function. In LDL particles glycosylation leads to alterations in LDL clearance and increases susceptibility to oxidation. [4]

The glycosylation is followed by a series of reactions, which result in an irreversible formation of stable advanced glycosylation end products (AGEs). AGEs promote atherosclerosis by non-receptor-dependent and receptor-mediated mechanisms. Receptor-mediated mechanisms are linked with a specific receptor of AGE determinants (RAGE). This receptor can be found in macrophages, endothelial cells, and smooth muscle cells. The expression of RAGE on these cells is usually low, however, under certain circumstances, upregulation may occur. The reaction of AGEs with RAGE results in increased oxidative stress and decreased barrier function of endothelial cells. It also triggers chemotaxis in monocytes. [4]

The glucose itself may increase oxidative stress by multiple pathways. Metal-catalyzed autooxidation may occur, which results in superoxide anion and hydrogen peroxide production. Hyperglycaemia may also reduce natural antioxidant mechanisms. Decreased plasma levels of vitamin C were measured in patients with diabetes mellitus. Molecules of glucose in the intracellular space may promote the formation of oxygen radicals from the products of cellular respiration on mitochondria. Hyperglycaemia in the intracellular space also results in the activation of the protein kinase C (PKC) system. The PKC system is involved in the transcription of growth factors. This promotes smooth muscle cell proliferation and extracellular matrix production. [4]

Smoking

Cigarette smoking is linked with many negative effects, such as lung cancer, renal failure, stroke, and CAD. Cigarette smoke has two phases, a tar phase, and a gas phase. The average cigarette smoke inhaled by a smoker consists of 8 % of tar and 95 % of gaseous components. Both the tar phase and a gas phase are a source of free radicals.

Free radicals from the gas phase have a shorter life span than the ones from the tar phase. [3] [32]

Cigarette smoke promotes inflammation, thrombosis, and LDL oxidation. It is related to increased levels of inflammation mediators, such as C-reactive protein, interleukins, or tumor necrosis factor. The particles in cigarette smoke interfere with nitric oxide synthesis, thus decreasing vasodilatory function. Besides LDL oxidation, smoking is also related to an increase in serum LDL concentration. The molecular mechanisms behind this process are still unclear. Recently, polymorphisms in endothelial NO synthase intron 4 gene and cytochrome P450 (CYP1A1) gene were linked to smoke exposure-related atherosclerosis. The cytochrome P450 is an enzyme, which catalyzes reactions involved in lipid metabolism. [3] [32]

Cigarette smoke also plays an important role in the formation of thrombi. Platelets from smokers exhibit increased adhesion and aggregation. The levels of circulating fibrinogen, along with tissue factor, correlate with the number of cigarettes smoked. On the contrary, the concentrations of fibrinolysis mediators are decreased. Due to the reduction of oxygen concentration in the inhaled air, smokers also have a higher number of erythrocytes, which results in increased blood viscosity. [3] [32]

Cigarette smoke contains nicotine molecules. Nicotine can bind to nicotinic cholinergic receptors. These receptors mediate cell signaling through acetylcholine. Overstimulation of cholinergic receptors leads to sympathetic nervous system activation, stress reaction, and an increase in blood pressure. Nicotine activation of cholinergic receptors also promotes the proliferation and migration of smooth muscle cells. In addition, nicotine-induced catecholamine stimulates platelet activation.[3] [23] [32]

2 GENETICS OF CORONARY ARTERY DISEASE

2.1 Genetic variants

Genetic variation is the difference in the genome between individuals in one population. Genetic variants are specific regions of the nucleotide sequence that vary amongst individuals. If the variant has a lesser frequency than 1% in a population it is considered a mutation, otherwise, it is called polymorphism. The variants in one region are called alleles. Humans have two alleles of one gene, one inherited from a mother and one from a father. Alleles can be dominant or recessive. Individuals who have two dominant or two recessive alleles are homozygous, individuals who have one dominant and one recessive allele are heterozygous. Three types of variants are distinguished - single nucleotide polymorphisms, indels, and structural variants.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are the most common type of variants. SNPs occur when one nucleotide in the chain is substituted by another type of nucleotide. Genetic code is degenerated, meaning that one amino acid is coded by multiple triplets. If the substitution of one nucleotide results in a triplet coding the same amino acid, the same protein is coded by both variants. If not, the resulting protein has a slightly different structure. This may or may not result in changes in the phenotype, depending on whether the amino acid has similar properties that the original.

Indels

Insertions or deletions of one or more nucleotides in the nucleotide chain are called indels. Indels cause a change in the reading frame, so unless the number of inserted or deleted base pairs is multiple of three the rest of the protein-coding sequence is affected. The resulting protein is almost always non-functional. Indels are not very frequent in protein-coding regions, however, they are common in non-coding highly repetitive regions.

Structural variants

Structural variants are larger changes in the structure of a chromosome. They include inversions, translocations, and copy number variants (CNVs). Inversion occurs when a chromosome is broken and the region between brakes rotates 180 degrees before it is repaired. Unless the brake is within an essential gene, the phenotype is not affected. Translocations occur when two chromosomes are broken at the same time and while repairing, they change parts. Larger amounts of such protein can be produced if the affected gene ends up near a more active promoter. CNVs are insertions or deletions in repetitive regions. They usually don't cause protein inactivation, however, the dosage of a protein coded by this region may be changed. [16] [17]

2.2 Monogenic causes of coronary artery disease

Monogenic variants are in genes that code a protein essential in some physiological process. Change in one gene usually affects the phenotype and highly increases the risk of the disease. Although CAD is mostly a polygenic disease, there are three known monogenic causes – familial hypercholesterolemia, triglyceride-rich lipoproteins, and lipoprotein(a).

Familial hypercholesterolemia

Familial hypercholesterolemia (FH) is an autosomal dominant disease caused by a deletion in a gene coding LDL receptor. Two other mutations also cause FH. One of them is the APOB gene, coding Apolipoprotein B. Apolipoprotein B is a part of a lipoprotein particle that binds to LDL receptors. The other is the proprotein convertase subtilisin/kexin type 9 (PCSK9) gene. PCSK9 is an enzyme responsible for the reaction between APOB and LDL receptors. Mutation in one of these genes disables the uptake and processing of LDL cholesterol by the liver. Patients with FH have high blood LDL. Xanthomas might also be present as a result of LDL cholesterol deposition in the skin and tendons. [18] [43]

Triglyceride-rich lipoproteins

The level of circulating triglycerides (TGs) is also associated with the risk of CAD. The level of TGs in the blood is regulated by enzyme lipoprotein lipase (LPL). It hydrolyses triacylglycerols into two fatty acids and monoacylglycerol. Chylomicrons and very-low-density lipoprotein (VLDL) particles are rich in TGs. LPL bonds to chylomicrons and VLDL particles through apolipoprotein A (APOA). Variants in LPL or APOA genes prevent the clearance of triglyceride-rich lipoprotein particles from the bloodstream. [11] [38]

Lipoprotein(a)

Lipoprotein(a) (Lp(a)) is a particle with a structure of LDL particle with apolipoprotein B100, to which apolipoprotein(a) is bonded via a disulfide bridge. Lp(a) is synthesized in the liver. A higher concentration of Lp(a) in plasma increases the risk of CAD. For its structural homology with plasminogen, it has prothrombotic properties and is interfering with the fibrinolytic cascade. It is also more prone to oxidation than LDL particles. Blood concentration of Lp(a) is determined by a gene encoding the apo(a) protein, the LPA, located on the long arm of chromosome 6. The concentration is not altered by environmental factors and it does not change during life. [30] [36]

2.3 Polygenic causes of coronary artery disease

The development of polygenic disease is determined by the presence of multiple risk variants in a group of genes. This group of genes usually codes one trait. The resulting phenotype is determined by the number of risk variants in this group and can be influenced by environmental factors. CAD shows a non-Mendelian pattern of inheritance. Therefore, each variant has a low phenotypic penetrance. [13] [22] [25] [31]

One way of investigating polygenic disease is a genome-wide association study (GWAS). GWAS is a powerful tool in discovering common variants with polygenic traits. When performing GWAS, whole genomes of individuals are sequenced. Genomes of cases are compared with genomes of controls. Variants are examined for reaching genome-wide significant p-value, which is usually set to $5*10^{-8}$. [13] [22] [25] [31]

Many polygenic determinants have been associated with the development of CAD. These concerns genes that are linked with lipid metabolism, proinflammatory, and

thrombogenic factors, vascular remodeling, and high blood pressure. In addition to that, in recent GWAS variants have been identified, whose pathway in CAD development is still unknown. In the table below (

Table 1), the top 20 loci associated with CAD are listed. The first column contains information about the gene loci, the second column contains the gene names and in the third column, possible CAD mechanisms affected by the gene are listed. [22] [25] [31]

Table 1: Top twenty loci associated with CAD [31]

rsID	**Gene**	**Mechanism**
rs11206510	PCSK9	Lipid metabolism
rs17114036	PPAP2B	Cell-cell interactions
rs4845625	IL6R	Inflammation
rs67258870	MIA3	Collagen secretion
rs515135	APOB	Lipid metabolism
rs2252641	ZEB2-ACO74093	Transcription regulation
rs1561198	VAMP5-VAMP8-GGCX	Intracellular processes
rs599839	SORT1	Lipid metabolism
rs9818870	MRAS	Cell growth
rs7692387	GUCY1A3	NO signalling
rs1878406	EDNRA	Vasoconsrtiction
rs12190287	TCF21	Transcription regulation
rs12526453	PHACTR1	Phosphatase 1 activity
rs10947789	KCKN5	Potassium channel
rs4252120	PLG	Fibrinolysis
rs4252120	NOS3	NO synthesis
rs2023938	HDAC9	Represses adipogenesis
rs264	LPL	Lipid metabolism
rs2954029	TRIB1	Cellular proliferation
rs4420638	APOE	Lipid metabolism

2.4 XY linked variation

Sex is determined by the 23rd pair of chromosomes. Male individuals have one X and one Y chromosome, females have two X chromosomes. Due to nondisjunction one sperm can carry two X or two Y gonosomes, whereas the other sperm has none. This results in some individuals having X0, XXX, or XYY karyotype. On chromosome X there are 813 protein-coding genes. On chromosome Y there are 143, and only 46 genes are male-specific. To balance the level of proteins one of the X chromosomes in females is inactivated. If there is a mutation in one of the X chromosomes, the affected chromosome is more likely to inactivate. For their unique properties, special analytical methods are required when studying sex-specific chromosomes. Therefore, they have been excluded from most of the GWAS. [29] [48]

Male individuals have a higher chance of developing CAD than females. A higher level of circulating estrogen decreases the risk of CAD by angiogenesis and vasodilatation and reducing oxidative stress. The Y chromosome contains genes linked with macrophage activation, thus increasing the risk of inflammation. However, in 2016 an X chromosome genome-wide meta-analysis was performed. Over 43,000 CAD cases and 58,000 controls were interrogated. According to this study, no significant CAD loci have been found on the X chromosome. [21] [26] [48]

2.5 Ethnicity linked variation

Variation differs among populations. Being the oldest, the African population shows more diversion than younger European and Asian populations. The differences in CAD prevalence between ethnicities are hugely influenced by lifestyle and the differences in access to healthcare. The influence of ethnicity on CAD development is not well mapped. Most of the GWAS are performed on specific populations, and in multi-racial studies, ethnicities are usually self-reported and ancestry is not considered. The results from GWAS performed on the European population were replicated on the Asian population, but failed to replicate on the African population. Some studies suggest that the Han Chinese population has a higher risk of CAD development than Europeans. Four CAD loci have been associated exclusively with the Han Chinese population. [31] [48]

3 GENOMIC PROJECTS

3.1 The 1000 Genomes project

The 1000 Genomes project is an international collaboration between scientists aiming to map human genome variation. It launched in 2008. It aimed to sequence at least 1,000 individuals and create a database of human genome variants, which could be used in further research. It was divided into stages - pilot and the main project divided into three stages. The results from the final stage were published in 2015. [44] [50]

The pilot

The purpose of the pilot was to determine the best strategy for detecting variants of various types and frequencies. The pilot consists of three projects, the two trios project, the low-coverage project, and the exon project. The two trios project used high-coverage sequencing of two families, parents, and one daughter. It helped to determine the number of de novo mutations, which was approximately 10^{-8} per bp per generation. The low-coverage project used low-coverage sequencing of 179 unrelated individuals from different populations. The exon project used a high coverage of 8,140 exons from 906 randomly selected genes in 697 individuals. The low-coverage was efficient for discovering common variants, the high-coverage exon sequencing for rare variants. Data were aligned to the GRCh36 reference sequence. [44] [50]

First phase

In phase 1, low-coverage and exome analysis was performed on data from 1,092 individuals from 14 populations. 38 million SNPs, 14 million short indels, and 14,000 large indels were identified. Data were aligned to GRCh37 reference sequence. [45] [50]

Second phase

In the second phase, another 1,700 individuals were sequenced and low coverage and exome analysis was performed. The current methods were improved and new methods were implemented. Data were aligned to GRCh37 reference sequence. [50]

Third phase

The third phase of the 1000 Genomes project was focused on discovering structural variants. 2,504 individuals from 26 populations were scanned. Both whole genome sequencing and targeted sequencing were used on all individuals. It was estimated that individuals have a median of 18.4 Mbp of structural variants. Data were aligned to GRCh38 reference sequence. [42] [50]

3.2 1000 Genomes project database

The data from the 1000 Genomes project are publicly available on two sites. The first is 1000 Genomes FTP site (see Index pro /vol1/ftp/ (ebi.ac.uk)), the other is NCBI FTP site (see Index pro /1000genomes/ftp/ (nih.gov)). Both sites have a similar structure. The CHANGELOG file contains information about site modifications. The README file contains additional information about data. The aligned data are stored in BAM format, the results in VCF format. [6]

VFC format is a special format for storing variants designed by the 1000 Genomes project consortium. It has a header part and a data part. The header begins with characters '##'. It contains meta-information about the data in the file. The data part contents columns representing chromosome (CHROM), the position of the variant on the chromosome (POS), variant identifier (ID), reference allele (REF), non-reference alleles (ALT), quality score (QUAL), site filtering information (FILTER) and additional information (INFO). [8]

Accessing 1000 Genomes project data

The data from phase 3 of the 1000 Genomes project can be accessed and downloaded from the NCBI data browser (*Figure 3*) (see Chr2: 1-243.2M - 1000 Genomes Browser (nih.gov)). There are six links on this page. Reset all allows resetting all settings. Share this page generates a temporary link to the page with current settings. FAQ contains currently asked questions. The link Help contains browser documentation. Version X.X contains release notes. YouTube is a link to a video tutorial on YouTube. [51]

The page contains several widgets. Ideogram view allows users to select the displayed chromosome. Chromosome overview helps to navigate the selected

chromosome. Exon navigator is located under the chromosome overview widget. It shows genes located on the selected sequence. Sequence viewer provides graphic representation on variation annotations. The search widget allows users to search for specific locations, genes, or phenotypes. The subject selection widget allows users to add alignment files. History shows recent gene searches. In the genotypes table, allele frequencies are displayed within populations. Your data allows users to upload custom tracks. Through the download widget, data can be downloaded. [51]

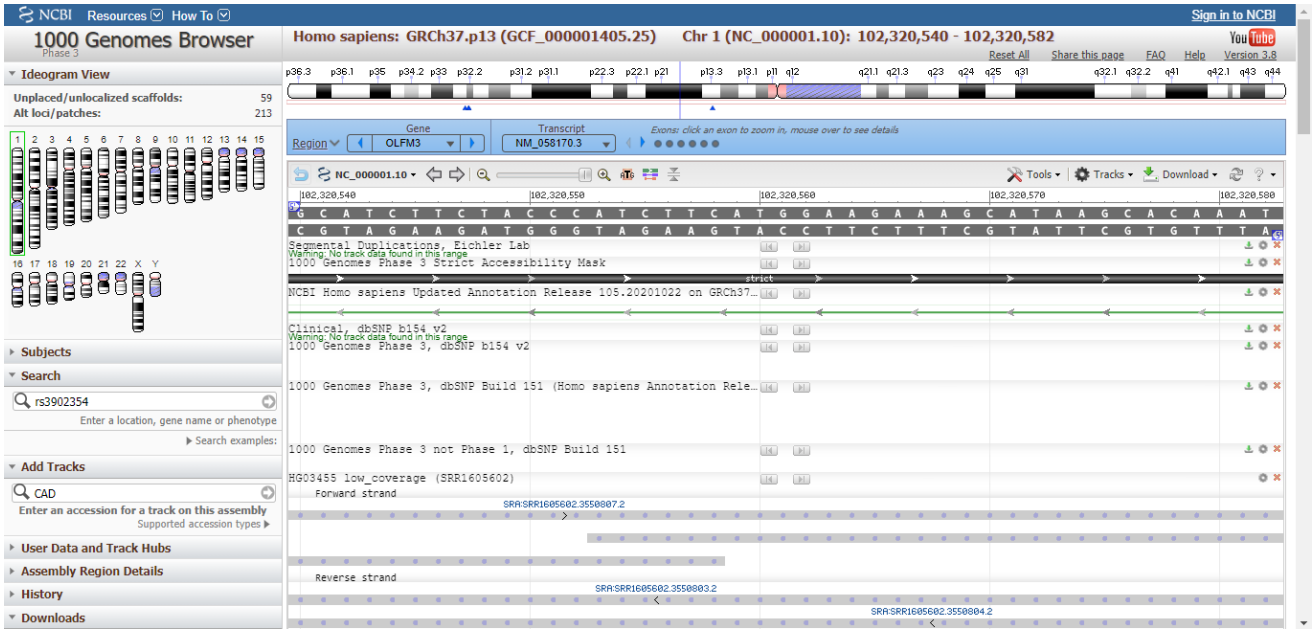


Figure 3: 1000 Genomes Browser

3.3 CARDIoGRAM plus CD4 consortium

Coronary artery disease genome-wide replication and meta-analysis (CARDIoGRAM) plus CD4 consortium is an international collaboration aiming to collect data from multiple GWAS to identify significant CAD loci. The results from the first meta-analysis were published in 2011. To this day seven analyses were performed. [9]

CARDIoGRAM GWAS was a meta-analysis of 22 GWAS. 22,233 cases and 64,762 controls from the European population were analyzed and 13 new CAD loci were identified. [9] [40]

CD4 GWAS was an analysis of 15,420 cases and 15,062 controls from European and South Asian populations. 11 common variants were associated with CAD.

Replication on 21,408 cases and 19,185 controls confirmed five new loci associated with CAD. [7] [9]

CARDIoGRAM plus CD4 Metabochip involved analysis of 63,746 cases and 130,681 controls from European and South Asian populations. 15 genome-wide significant loci were identified. [9] [10]

CARDIOGRAM plus CD4 1000 Genomes based GWAS is a meta-analysis of 185,000 CAD cases and controls from multiple genome-wide association studies. 6.7 million common and 2.7 million low-frequency variants were interrogated and 10 novel CAD loci were discovered. [1] [9]

Myocardial Infarction Genetics and CARDIoGRAM Exome is a study of 42,335 cases and 78,240 controls of European ancestry. The aim was to test if variants previously associated with non-CAD diseases are also linked to CAD. Six new loci on genome-wide significance were discovered. [9] [47]

Chromosome-X CAD: A meta-analysis of X-chromosomal variants for CAD was aiming to map variation on the X chromosome. 101,000 cases and controls were interrogated and no genome-wide significant loci were discovered. [9] [26]

A meta-analysis of UK Biobank SOFT CAD GWAS (interim release) with CARDIoGRAMplusC4D 1000 Genomes-based GWAS and the Myocardial Infarction Genetics and CARDIoGRAM Exome aimed to interrogate variants previously discovered with a 5 % false-positive rate. 13 new CAD loci were identified. [9] [34]

3.4 CARDIoGRAM plus CD4 consortium database

The data from CARDIoGRAM and CD4 meta-analyses are publicly available (see CARDIoGRAMplusC4D Consortium). The summary statistics from each project can be downloaded as a zip file, which contains a README file, where additional information is added, and the data proper in text format. [9]

Accessing CARDIoGRAM plus CD4 data

CARDIoGRAM plus CD4 data is accessible on CARDIoGRAM plus CD4 consortium web page in text format. The downloaded file can be opened in any text editor.

The columns are separated by a tab. The data can be opened in a MS Excel worksheet or loaded as a table in R software. [9]

4 DATA ANALYSIS

4.1 Input data description

Data on coronary artery disease / myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from CARDIoGRAM plus CD4 website (see www.CARDIOGRAMPLUSC4D.ORG). It is a summary statistic of a meta-analysis from 48 studies. The data were obtained from 184,305 individuals, from which 60801 were cases and 123,504 were controls. 77% of the individuals were from Europe, 13 % from South Asia, and 6 % from East Asian populations. The rest of the individuals were Hispanic and African Americans. [1]

The data were imputed using the 1000 Genomes Project phase 1 dataset. Variants with low imputation quality were excluded. Variants were filtered for MAF > 0,005. Only variants that were retained in at least 60 % of the individual studies were used in the meta-analysis. Fixed effects model was used in the meta-analysis. Heterogeneity was assessed and variants with high heterogeneity were reanalyzed using a random-effects model. [1]

Three genetic models were tested, the additive model, the recessive model, and the dominance model. In the additive model, the odds ratio is proportional to the number of risk alleles. In the recessive model frequency of the recessive haplotype is compared with the frequency of dominant homozygotes and heterozygotes combined. The results from the dominance model were consistent with the additive results, so only additive and recessive datasets with summary statistics were created. Both datasets are investigated in this thesis. The additive dataset contains 9,455,778 variants. There are 95 insertions, 72 deletions, and 2,046 SNPs, that reach genome-wide significance with a p-value less than $5*10^{-8}$. The recessive dataset contains 6,914,384 variants. There are 4 insertions, 6 deletions, and 126 SNPs under the target p-value. [1]

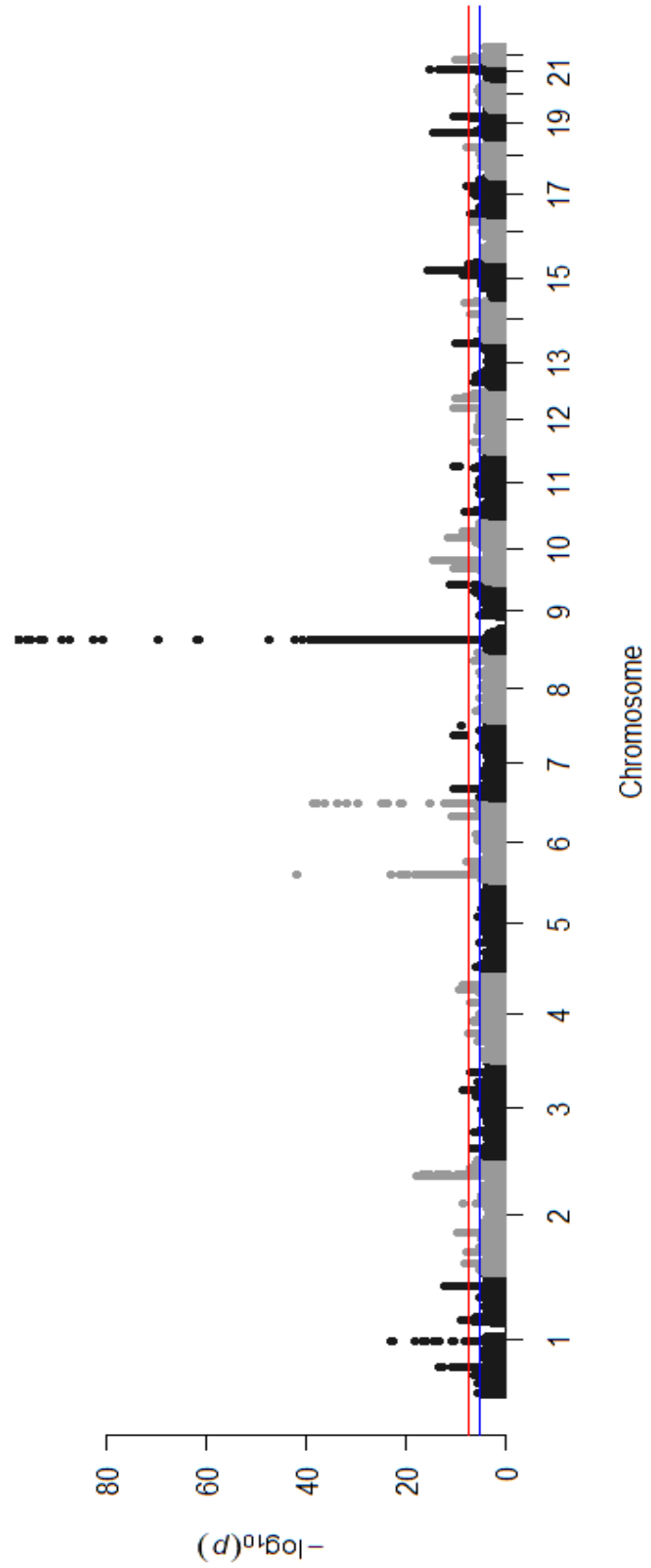
The example of the data from the 1000 Genomes-based GWAS dataset is in the picture below (*Figure 4*). The file contains the following columns. Markername contains the codename of the variant. The column Chr contains the number of the chromosome on which the variation is located. The column Bp_hg19 contains information about the position on the chromosome. In the column Effect_allele, the risk CAD allele is stored, and the column Noneffect_allele contains the reference allele. In Effect_allele_freq the

frequency of the CAD risk allele is stored. The column median_info contains median imputation quality based on all participating studies. The column model contains information, whether fixed or random effects statistical model was used in the meta-analysis. The column beta contains the natural logarithm of the odds ratio. In the column se_dgc, the standard error is stored. P_dgc contains association p-value based on beta & se_dgc. In the column, het_pvalue heterogeneity p-value is stored. The column n_studies contains the number of studies in which the variant was included. [9]

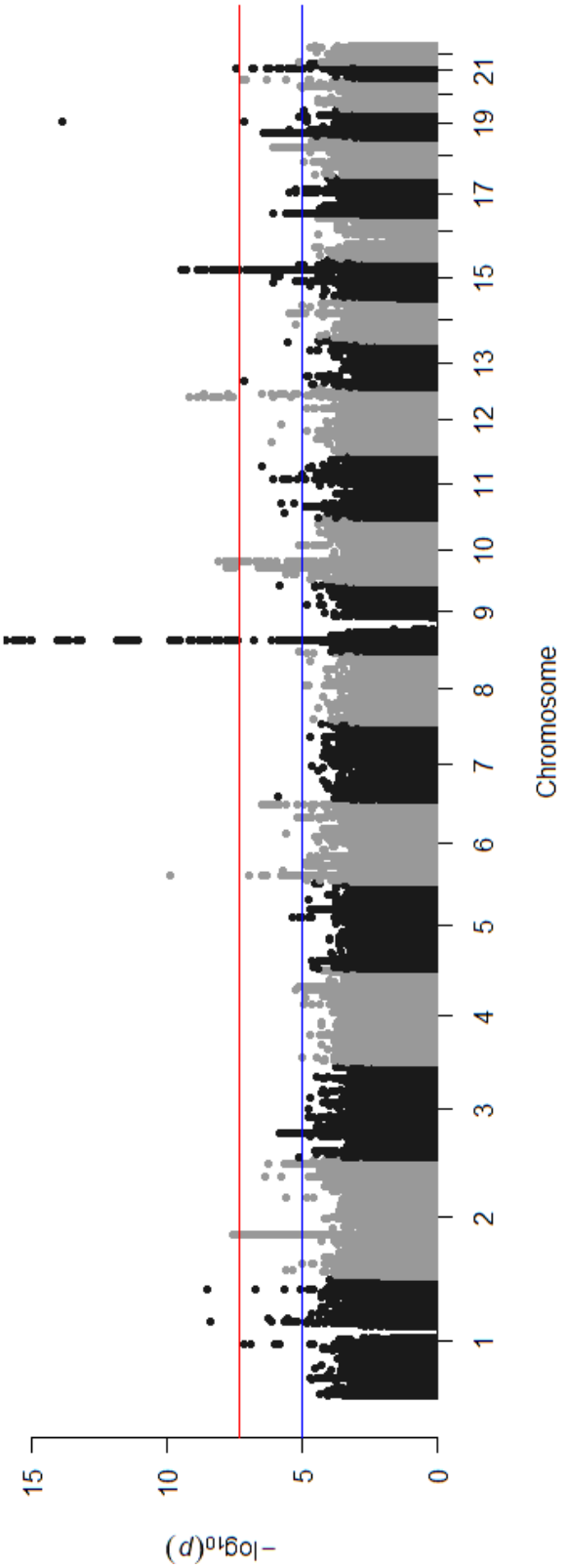
markername	chr	bp_hg19	effect_allele	noneffect_allele	effect_allele_freq	median_info	model	beta	se_dgc	p_dgc	het_pvalue	n_studies
rs143225517	1	751756	C	T	0,158264	0,92	FIXED	0,013006	0,017324	0,452802	0,303481	35
rs3094315	1	752566	A	G	0,763018	1	FIXED	-0,00524	0,015765	0,73946	0,146867	36
rs3131972	1	752721	G	A	0,740969	0,96034	FIXED	-0,00303	0,015638	0,846265	0,340843	36
rs3131971	1	752894	C	T	0,744287	0,793	FIXED	0,00464	0,016238	0,775066	0,035821	37
rs61770173	1	753405	A	C	0,775368	0,91694	FIXED	-0,00629	0,016708	0,706527	0,377485	36
rs2073814	1	753474	G	C	0,716742	0,8848	FIXED	0,000407	0,015746	0,979378	0,472543	35
rs2073813	1	753541	A	G	0,194804	0,91034	FIXED	0,005802	0,016781	0,729529	0,36886	35
rs3131969	1	754182	G	A	0,760434	0,92385	FIXED	-0,00652	0,016557	0,693647	0,321205	34
rs3131968	1	754192	G	A	0,759886	0,920915	FIXED	-0,00679	0,016567	0,681868	0,306346	34
rs3131967	1	754334	C	T	0,747971	0,841665	FIXED	-0,00934	0,016848	0,579209	0,015972	34
rs3115859	1	754503	A	G	0,742041	0,92128	FIXED	-0,00015	0,015827	0,99264	0,39667	36
rs3131966	1	754964	T	C	0,727432	0,82858	FIXED	0,002221	0,016231	0,891161	0,049833	34
rs3131965	1	755775	G	A	0,71467	0,758	FIXED	0,006636	0,016473	0,687073	0,167983	35
rs3115858	1	755890	T	A	0,776798	0,90905	FIXED	-0,00891	0,017013	0,600318	0,167264	36
rs3131962	1	756604	G	A	0,756247	0,848395	FIXED	-0,00486	0,016819	0,772702	0,245825	36
rs3115853	1	757640	A	G	0,749175	0,93417	FIXED	-0,00879	0,01687	0,602541	0,28788	33
rs4951929	1	757734	T	C	0,782655	0,941855	FIXED	-0,01407	0,017018	0,408257	0,267617	36
rs4951862	1	757936	A	C	0,78203	0,93908	FIXED	-0,01472	0,017026	0,387418	0,24359	36
rs3131956	1	758144	G	A	0,767599	0,95246	FIXED	-0,00688	0,016438	0,675676	0,259495	36
rs3131954	1	758626	T	C	0,781067	0,92878	FIXED	-0,01363	0,017155	0,426743	0,308077	36
rs137919983	1	759700	C	T	0,736839	0,873965	FIXED	-0,00235	0,016558	0,886949	0,183303	34
rs114111569	1	759837	A	T	0,784208	0,95476	FIXED	-0,01408	0,017224	0,413531	0,312214	36

Figure 4: A 1000 Genomes based GWAS meta-analysis dataset

Manhattan plots were constructed from the datasets, to assess the genome-wide significant loci. Plot 1 (Plot 1: Manhattan plot from the additive dataset*(Plot 1*) depicts the Manhattan plot for the additive dataset. On the x axis, there are positions of SNPs on chromosomes. On the y axis, there are negative natural logarithms of p values. 5 additive loci on chromosomes 2 and 10, 4 on chromosomes 1 and 6, 3 on chromosome 7 and 4, 2 on chromosome 9, 11, 12, 15, and 19, 1 on chromosomes 3, 13, 17, 18, 21 and 22. Plot 2 (*Plot 2*) depicts the Manhattan plot from the recessive dataset. 1 recessive locus was located on chromosomes 2, 6, 9, 12, 15, 19, and 21. 2 on chromosomes 1 and 10. The subsequent analysis is focused on genome-wide significant variants in these loci. Gonosomes were not sequenced and included in the analysis.



Plot 1: Manhattan plot from the additive dataset



Plot 2: Manhattan plot from the recessive dataset

4.2 Population analysis

Let's have an allele A, which occurs with frequency p_A, and allele B, which occurs with frequency p_B. The frequency of the AB haplotype is then p_{AB}. If the occurrence of allele A and allele B is random and the occurrence of one does not affect the occurrence of the other, the frequency p_{AB} would be equal to $p_A p_B$. If the occurrence of one allele is affected by the occurrence of the other, the two variants are in linkage disequilibrium (LD). The level of LD can be quantified by the coefficient of linkage disequilibrium, which is defined by the equation [1].

$$D = p_{AB} - p_A p_B \quad [1]$$

D is a coefficient of linkage disequilibrium, p_{AB} is the frequency of haplotype AB, p_A is the frequency of allele A and p_B is the frequency of allele B. [41]

If D is equal to 0, variants are non-dependent. In other words, they are in linkage equilibrium. The magnitude of D corresponds to the degree of linkage disequilibrium. If the sign of D is positive, the variants are positively correlated, so the haplotype is more frequent than expected. If the sign of D is negative, the variants are negatively correlated and the haplotype is less frequent than expected. [41]

Another way of quantifying LD is the square root of correlation coefficient r^2, which can be calculated by the equation [2].

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \quad [2]$$

r^2 is the square root of the correlation coefficient, D^2 is the square root of the coefficient of linkage disequilibrium, p_A is the frequency of allele A and p_B is the frequency of allele B. [41]

The LD matrixes were calculated using LDlink. LDlink is an online tool for investigating linkage disequilibrium patterns among populations, that was developed by Mitchell Machiela in collaboration with NCI's Center for Biomedical Informatics and Information Technology. It is available from the LDlink website (see LDlink | An Interactive Web Tool for Exploring Linkage Disequilibrium in Population Groups

(nih.gov)). LD parameters are calculated from the 1000 Genomes database. LDmatrix module was used to calculate and visualize pairwise linkage disequilibrium in loci of interest. As an input, it takes a list of a maximum of 300 variants and reference populations from the drop-down menu. The output is a visualization of the D and r^2 matrix and two datasets with calculated D and r^2. Both datasets and the graph can be downloaded. [28]

LD matrixes were calculated for the identified genome-wide significant loci, which contained more than one significant SNP. Loci, where only one SNP was genome-wide significant variant was discovered, were not included in the analysis. Populations were selected according to the population structure of the input data. East Asian (EAS), European (EUR), South Asian (SAS), Americans of African ancestry (ASW), and Mexican ancestry from Los Angeles (MXL) were included.

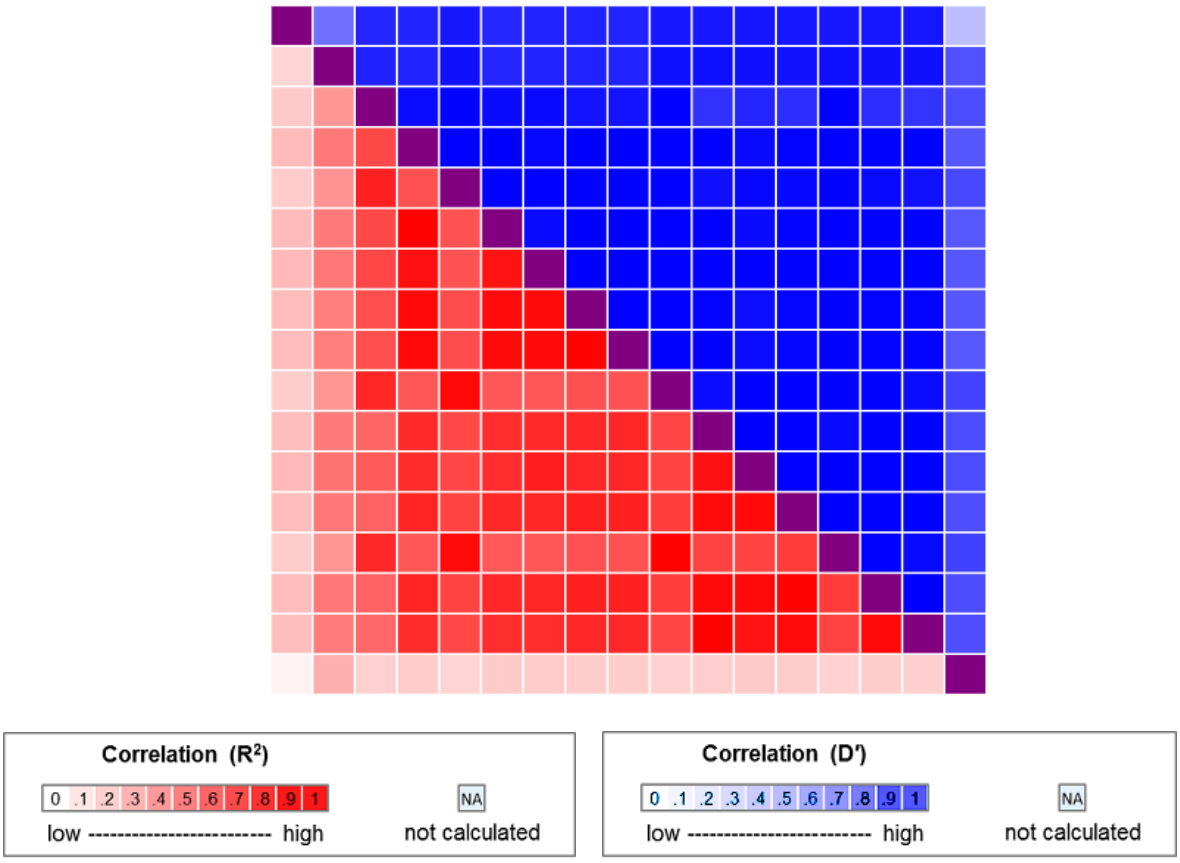


Figure 5: LD matrix chromosome 1 locus 1

A visualization of the LD matrix from locus 1 on chromosome 1 is in picture 5 (

Figure 5). Visualizations of LD matrixes from the rest of the loci are in appendix A. In the bottom left corner of each matrix is the r^2 parameter colored in red. In the top right corner is the D parameter colored in blue. The more saturated colors are used for highly correlated variants. In locus 3 on chromosome 2, a locus on chromosome 3, locus 1 on chromosome 4, locus 1 on chromosome 7, locus 4 on chromosome 10, both loci on chromosome 11, locus 2 on chromosome 12, locus 1 on chromosome 15, and loci on chromosomes 18 and 22, variants are highly correlated, suggesting only one risk allele is present. In other loci white stripes in the pictures are present, which indicates a presence of SNPs in low LD with the leading variant, suggesting allelic heterogeneity. In some loci, such as locus 2 on chromosome 1, groups of variants formed, which are in a strong LD, suggesting more genes might be located in a locus.

4.3 Statistical analysis

When calculating p values for variants association tests are usually performed on a single SNP basis. The variant with the most significant p-value in a locus is then selected as the pathogenic variant under the assumption, that other variants in the region show significance because they are in linkage disequilibrium with the leading SNP. However, if there are multiple casual variants present in a locus, they won't be discovered by this procedure and the overall variation in a said locus might be underestimated. If a variant is negatively correlated with the leading SNP, the effect of this allele is higher than the test on a single SNP basis suggests. In some cases, two variants relatively distant from each other might be in a strong LD and one of the variants is a false positive. The phenomenon, where more than one variant in a locus is responsible for a certain phenotype, is called allelic heterogeneity. The joint analysis aims to investigate allelic heterogeneity and adjust the effect sizes of genome-wide significant SNPs according to the LD patterns. [35] [49]

One of the tools to examine allelic heterogeneity is conditional and joint (COJO) analysis in GCTA software. Another option is to use a selection operator for jointly analyzing multiple variants (SOJO). SOJO was introduced in a study published in 2017. It is based on the least absolute shrinkage and selection operator (LASSO). According to the study, SOJO provides better sensitivity and specificity in predicting risk alleles than COJO. In this study R package for SOJO was developed. [35] [49]

In a single-SNP model, each variant is examined separately. According to a multiple trait model, a list of phenotypes y is affected by a list of genotypes X. Each variant from genotype matrix X has an effect β on phenotype y. The multiple trait model is described by the equation [3].

$$\boldsymbol{y} = \boldsymbol{X}\beta + e \qquad [3]$$

X is n*p genotype matrix, y is n*1 phenotype vector, where n is a number of individuals and p is a number of variants, and β is the effect size vector and e is an error. In a single-SNP model, each variant is examined separately.[35]

LASSO is a regression analysis and prediction model. It is a great tool for the correction of overfitting and feature selection. It minimizes the residual sum of squares

and adds L1 regularization. L1 regularization is applied by multiplying the absolute value of β by a penalization parameter λ. The parameter λ allows setting coefficients β to zero. As the value of λ increases, more variables are removed. Large coefficients are allowed only if they lead to a better fit. The model can be described by the equation [4].

$$\min_{\hat{\beta} \epsilon R^p} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{X}\hat{\beta} \right\|_2^2 + \lambda \left\| \hat{\beta} \right\|_1 \quad [4]$$

Y is a phenotype vector, X is a genotype vector, β is an estimate of effect size and λ is a correction parameter.[35]

When performing joint analysis, individual genotype data are required to obtain LD. However, these data are often not available due to the protection of participants' data. Tools were developed for performing joint analysis on summary-level GWAS meta-analysis data. In that case, a reference LD panel is required. Ideally, the LD is calculated from one of the individual studies used in a meta-analysis. If this is not possible, it can be obtained from Hap Map or 1000 Genomes data. However, the 1000 Genomes-based LD panel is less accurate due to the small sample size. [35][49]

The analysis involved loci, which contained multiple significant variants. Linkage disequilibrium patterns were examined in these loci. If the r^2 value between all variants in a locus was greater than 0.9, variants were deemed to have no allelic heterogeneity and the locus was excluded from further analysis. This concerns locus 1 on chromosome 4, a locus on chromosome 18 and locus on chromosome 22 in the additive dataset, and locus on chromosome 2 in the recessive dataset. In these loci, the variant with the most significant p-value was selected as a risk variant.

The joint analysis was performed using the SOJO package in R. As an input, the SOJO function takes the GWAS summary statistics dataset, which has to contain the column SNP, in which the variant markers are listed, the column A1, which contains the effect allele, the column A2, which contains the reference allele, the column b, which contains the value of beta, the column se, which contains the standard error, and the column N, which contains the number of studies included in the meta-analysis. Another input is the LD matrix. The rows and columns of the matrix are named according to the names of the variants. The third input is a vector of reference alleles of variants present

in the LD matrix. Optional input is the number of required output variants, the default setting is 50. The output of the function SOJO is a vector of lambdas, a matrix of estimated betas through set lambdas, and a chain of selected SNPs.

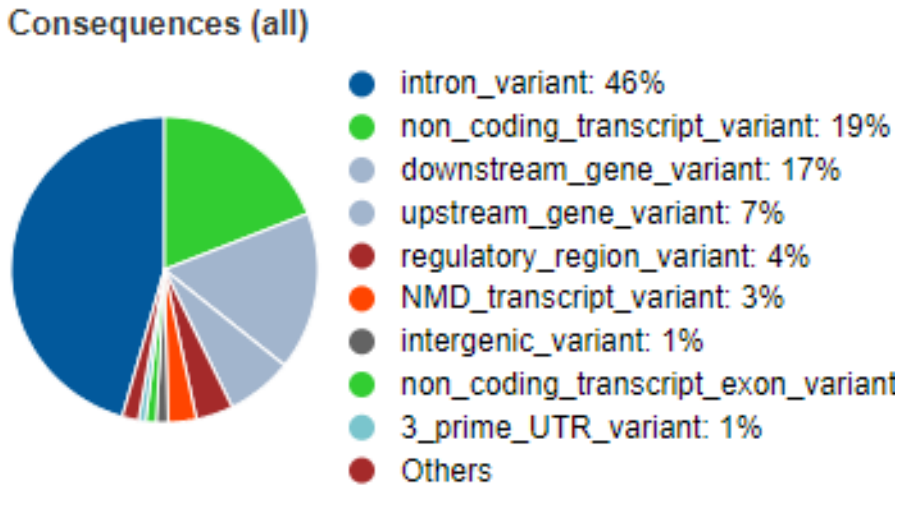
At the end of the analysis 371 variants were identified in the additive dataset. 8 variants were the only genome-wide significant variants in their loci. 4 variants were selected according to the p-value in loci with high LD values. 359 variants were identified in the LASSO analysis. 347 variants were SNPs, 16 were insertions and 8 were deletions. In the recessive dataset, 55 variants were identified. 4 variants were the only genome-wide significant variants in their loci, 1 variant was selected in a locus with high LD values. l54 variants were SNPs and 1 was insertion. All selected variants are listed in tables in Appendix B.

4.4 Bioinformatic analysis

The variants showing genome-wide significance were annotated using Variant Effect Predictor on the Ensembl website (see Variant Effect Predictor - Homo_sapiens - Ensembl genome browser 104). The Ensembl project was launched in 1999 by European Bioinformatics Institute. It is a tool for the management and analysis of genomic data from vertebrate species. Via Ensembl browser, various genomic databases can be remotely accessed. It provides BLAST sequence alignment and alignment visualization Through the Variant Effect Predictor specialist tool effects of variants can be predicted. [20]

The additive dataset

The pie chart below (*Plot 3*) shows percentages of additive variants with their effects. The chart shows the percentages of variants with their consequence from all variants in the dataset. Some variants have more consequences. 46% of the variants were intron variants. 19 % were non-coding transcript variants. 24 % of the variants were located within coding regions of genes. In the additive dataset 68 genes were found. All genes are listed in the table below (*Table* 2). 10 genes were identified both in the additive and in the recessive dataset. 8 pseudogenes, which were also identified, are not listed in the table below. The genes are APOC1P1 on chromosome 19, BCRP1 on chromosome 22, PFN1P11 on chromosome 10, POM121L9P on chromosome 22, PTGES3P4 on chromosome 10, RN7SL126P on chromosome 2, RN7SL192P on chromosome 19, and UBA52P6 on chromosome 9.



Plot 3: Consequences of variants in the additive dataset

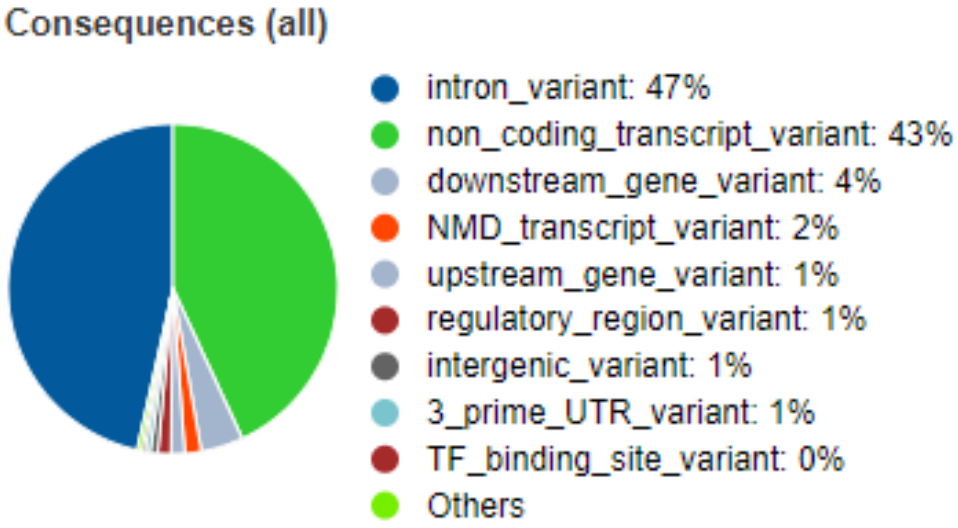
Table 2: Genes discovered in the recessive dataset

gene	leading variant	chr	mechanism
ABO	rs8176643	9	inflammation
ADAMTS7	rs62011004	15	cellular proliferation
AIRN	rs139311851	6	lncRNA
APOC1	rs118147862	19	lipoprotein metabolism
APOE	rs7412	19	lipoprotein metabolism
AS3MT	rs77335224	10	arsenite methyltransferase
ATP2B1	rs7314459	12	calcium homeostasis
ATXN2	rs653178	12	thrombosis
BCAM	rs118147862	19	extracellular matrix
BCAS3	rs7212798	17	cell proliferation
BORCS7	rs3824754	10	lysosomes movement
BORCS7-ASMT	rs150786824	10	nonsense-mediated mRNA decay
BROX	rs4282830	1	unknown
C19orf18	rs12979495	19	unknown
CARM1	rs17850995	19	inflammation
CDKN2B	rs2069418	9	cellular proliferation
CDKN2B-AS1	rs9632885	9	transcription regulation
CELSR2	rs7528419	1	cellular proliferation
COL4A1	rs11617955	13	collagen protein
COL4A2	rs9515203	13	collagen protein
CYP17A1	rs17115100	10	lipid metabolism
CYP46A1	rs10139550	14	lipid metabolism
DNM2	rs2287029	19	myopathy
FAM177B	rs36042506	1	unknown
GGCX	rs6547621	2	inflammation
GUCY1A1	rs72689147	4	NO synthesis
HDAC9	rs7788972	7	lipid metabolism
HHIPL1	rs10139550	14	unknown
IGF2R	rs139311851	6	cellular proliferation
IL6R	rs4845619	1	inflammation
JCAD	rs10826751	10	cell adhesion
LDLR	rs6511721	19	lipid metabolism
LINC00841	rs1870635	10	lncRNA
LINC01312	rs1832181	6	lncRNA
LINC01412	rs17678683	2	lncRNA
LIPA	rs2250644	10	lipid metabolism
LPA	rs55730499	6	lipid metabolism
LPAL2	rs9457927	6	lipid metabolism
MAPKAPK5-AS1	rs11513729	12	lncRNA
MAT2A	rs6547621	2	cell proliferation

MIA3	rs4846770	1	collagen production
MORF4L1	rs62011004	15	transcription regulation
MRAS	rs139016349	3	cellular proliferation
NBEAL1	rs72934591	2	thrombosis
NECTIN2	rs41290120	19	cell adhesion
NOA1	rs17087335	4	NO synthesis
NOS3	rs3918226	7	NO synthesis
PARTICL	rs6705971	2	lncRNA
PHACTR1	rs12202901	6	protein phosphatase activity
PLG	rs4252185	6	thrombosis
PLPP3	rs2017955	1	lipid metabolism
PSRC1	rs599839	1	lipid metabolism
SLC22A2	rs10080815	6	ion transporter
SLC22A3	rs202097157	6	ion transporter
SMAD3	rs56062135	15	cell signalling
SMARCA4	rs73013166	19	transcription regulation
SWAP70	rs12360647	11	inflammation
TARID	rs1832181	6	lncRNA
TCF21	rs6569912	6	transcription regulation
TOMM40	rs429358	19	respiratory chain
U6	rs3824754	10	snRNA
VAMP5	rs55971080	2	SMC proliferation
VAMP8	rs7568458	2	SMC proliferation
WBP1L	rs138009835	10	hypertension
WDR12	rs72932588	2	cellular processes
YIPF2	rs17850995	19	extracellular matrix
ZC3HC1	rs11556924	7	cell proliferation

The recessive dataset

The pie chart below (*Plot 4*) shows percentages of recessive variants with their effects. The chart shows the percentages of variants with said consequence from all variants in the dataset. Some variants have more consequences. 47 % of the variants were intron variants. 43 % were non-coding transcript variants. 5 % of variants were located within coding regions of genes. In the recessive dataset 16 genes were identified. All genes are listed in the table below (*Table 3*). 6 genes were identified exclusively in the recessive dataset. The genes are CDKN2A, DPY19L3DT, NAA25, PTPN11, RPL6, and ZNF507.



Plot 4: Consequences of variants in the recessive dataset

Table 3:Genes discovered in the recessive dataset

Gene	**leading SNP**	**chr**	**mechanism**
ADAMTS7	rs12438645	15	cellular proliferation
CDKN2A	rs2518723	9	cellular proliferation
CDKN2B	rs3217992	9	cellular proliferation
CDKN2B-AS1	rs10757264	9	transcription regulation
DPY19L3DT	rs12976411	19	lncRNA
IL6R	rs10908837	1	inflammation
JCAD	rs12259037	10	cell adhesion
LINC00841	rs12359058	10	lncRNA
MAPKAPK5-AS1	rs11513729	12	lncRNA
MORF4L1	rs7173743	15	transcription regulation
NAA25	rs17696736	12	cell cycle
PTPN11	rs11066301	12	cell proliferation
RPL6	rs11066283	12	inflammation
VAMP5	rs1009	2	SMC proliferation
VAMP8	rs1009	2	SMC proliferation
ZNF507	rs12976411	19	unknown

CONCLUSION

This bachelor's thesis focused on the research of genetic variants in CAD. The pathological process behind CAD is atherosclerosis. Atherosclerosis begins with the disruption of endothelial cells in the tunica intima of coronary arteries. LDL cholesterol particles accumulate in subendothelial space, where they are oxidized. The oxidized LDL particles attract leukocytes, thus causing inflammation. Necrosis occurs in macrophages, which phagocytize LDL particles. SMCs migrate from tunica media to the lesion and produce extracellular matrix and collagen fibers. Fibrous plaque is formed over the lipid lesion. If the plaque ruptures, thrombosis occurs, causing the acute coronary syndrome.

Multiple factors contribute to CAD development. Many CAD risk factors are known, such as a lack of exercise, high lipid diet, and smoking. It has been demonstrated that CAD has a heritable component. CAD is a polygenic disorder. Multiple variants contribute to the risk of CAD development. These concerns variants in genes coding proteins involved in lipid metabolism, inflammatory and thrombogenic processes, and cellular proliferation. Thanks to the development of GWAS, some variants associated with CAD were discovered, whose molecular pathways are still unknown. However, a huge part of CAD heritability is still unexplained.

One way of explaining the missing CAD heritability is allelic heterogeneity. It is a phenomenon, where one locus contains multiple risk variants. These are not discovered in a standard GWAS. Only the most significant SNP is selected as a risk variant in a locus, under the assumption, that genome-wide significant variants in a locus are in LD with the leading variant. Allelic heterogeneity was assessed by the LASSO regression model in 1000 Genomes-based genome-wide meta-analysis summary statistics data downloaded from CARDIoGRAM plus CD4 database. Two datasets were used, additive and recessive. Manhattan plots were constructed to identify genome-wide significant loci. LD matrixes for genome-wide significant loci were constructed using the LDlink online tool. Genome-wide significant loci were investigated using the SOJO function from the SOJO package in r. 347 SNPs, 16 insertions, and 8 deletions were discovered in the additive dataset. In the recessive dataset 154 SNPs and 1 insertion were identified.

Variants were annotated using the variant effect predictor tool on the Ensembl website. Almost 50 % of the variants were located in intronic regions. 24 % of the variants in additive loci were located within coding regions of genes. In the recessive loci, 5 % of the variants were located within coding regions of genes. In recessive loci, 68 genes and 10 pseudogenes were identified. In recessive loci, 16 genes were identified. 10 genes were identified both in the additive and in the recessive loci. Together 72 genes were identified to be associated with CAD.

LITERATURE

[1] A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*. 2015, 47(10), 1121-1130. ISSN 1061-4036. DOI:10.1038/ng.3396

[2] AMBROSE, John A., Manmeet SINGH. Pathophysiology of coronary artery disease leading to acute coronary syndromes. *F1000Prime Reports*. 2015, 7. ISSN 20517599. DOI:10.12703/P7-08

[3] AMBROSE, John A, Rajat S BARUA. The pathophysiology of cigarette smoking and cardiovascular disease. *Journal of the American College of Cardiology*. 2004, **43**(10), 1731-1737. ISSN 07351097. Dostupné z: doi:10.1016/j.jacc.2003.12.047

[4] ARONSON, Doron and Elliot J RAYFIELD. How hyperglycaemia promotes atherosclerosis: molecular mechanisms. *Cardiovascular Diabetology*. **1**(1). ISSN 14752840. Dostupné z: doi:10.1186/1475-2840-1-1

[5] BENNETT, Martin R., Sanjay SINHA a Gary K. OWENS. Vascular Smooth Muscle Cells in Atherosclerosis. *Circulation Research*. 2016, **118**(4), 692-702. ISSN 0009-7330. Dostupné z: doi:10.1161/CIRCRESAHA.115.306361

[6] Clarke, L., Zheng-Bradley, X., Smith, R. et al. The 1000 Genomes Project: data management and community access. Nat Methods 9, 459–462 (2012). DOI: 10.1038/nmeth.1974

[7] Coronary artery disease (CD4) consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature Genetics. 2011, 43(4), 339-344. ISSN 1061-4036. DOI:10.1038/ng.782

[8] DANECEK, P., A. AUTON, G. ABECASIS, et al. The variant call format and VCFtools. Bioinformatics. 2011, 27(15), 2156-2158. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btr330

[9] Data downloads. CARDIoGRAM plus CD4 Consortium [online]. CARDIoGRAM plus CD4 Consortium [cit. 2020-12-30]. Available from: http://www.cardiogramplusc4d.org/data-downloads/

[10] DELOUKAS, Panos, Stavroula KANONI, Christina WILLENBORG, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nature Genetics. 2013, 45(1), 25-33. ISSN 1061-4036. DOI:10.1038/ng.2480

[11] DO, Ron, Cristen J WILLER, Ellen M SCHMIDT, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*. 2013, 45(11), 1345-1352. ISSN 1061-4036. DOI:10.1038/ng.2795

[12] ELOSUA, Roberto, Sergi SAYOLS-BAIXERAS, Carla LLUÍS-GANELLA a Gavin LUCAS. Pathogenesis of coronary artery disease: focus on genetic risk factors and identification of genetic variants. *The Application of Clinical Genetics*. ISSN 1178-704X. DOI:10.2147/TACG.S35301

[13] ERDMANN, Jeanette, Thorsten KESSLER, Loreto MUNOZ VENEGAS a Heribert SCHUNKERT. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovascular Research* [online]. 2018 [cit. 2021-5-27]. ISSN 0008-6363. Dostupné z: doi:10.1093/cvr/cvy084

[14] GALKINA, Elena a Klaus LEY. Immune and Inflammatory Mechanisms of Atherosclerosis. *Annual Review of Immunology*. 2009, **27**(1), 165-197. ISSN 0732-0582. Dostupné z: doi:10.1146/annurev.immunol.021908.132620

[15] GINSBERG, Henry N. New Perspectives on Atherogenesis. *Circulation*. 2002, 106(16), 2137-2142. ISSN 0009-7322. DOI: 10.1161/01.CIR.0000035280.64322.31

[16] GRIFFITHS AJF, Miller JH, Suzuki DT, et al. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Inversions. Available from: https://www.ncbi.nlm.nih.gov/books/NBK22042/

[17] GRIFFITHS AJF, Miller JH, Suzuki DT, et al. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Translocations. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21947/

[18] HILL, J S, M R HAYDEN, J FROHLICH and P H PRITCHARD. Genetic and environmental factors affecting the incidence of coronary artery disease in heterozygous familial hypercholesterolemia. *Arteriosclerosis and Thrombosis: A Journal of Vascular Biology*.1991,11(2),290-297. ISSN 1049-8834. DOI: 10.1161/01.ATV.11.2.290

[19] HOLLANDER, William. Role of hypertension in atherosclerosis and cardiovascular disease. *The American Journal of Cardiology*. 1976, **38**(6), 786-800. ISSN 00029149. Dostupné z: doi:10.1016/0002-9149(76)90357-X

[20] HOWE, Kevin L, Premanand ACHUTHAN, James ALLEN, et al. Ensembl 2021. *Nucleic Acids Research* [online]. 2021, **49**(D1), D884-D891 [cit. 2021-5-26]. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkaa942

[21] IORGA, A., Cunningham, C.M., Moazeni, S. *et al.* The protective role of estrogen and estrogen receptors in cardiovascular disease and the controversial use of estrogen therapy. *Biol Sex Differ* 8, 33 (2017). Available from: https://doi.org/10.1186/s13293-017-0152-8

[22] KHERA, Amit V. and Sekar KATHIRESAN. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature Reviews Genetics*. 2017, 18(6), 331-344. ISSN 1471-0056. DOI: 10.1038/nrg.2016.160

[23] LEE, Jieun a John P. COOKE. The role of nicotine in the pathogenesis of atherosclerosis. *Atherosclerosis*. 2011, **215**(2), 281-283. ISSN 00219150. Dostupné z: doi:10.1016/j.atherosclerosis.2011.01.003

[24] LIBBY, Peter and Pierre THEROUX. Pathophysiology of Coronary Artery Disease. *Circulation*. 2005, 111(25), 3481-3488. ISSN 0009-7322. DOI: 10.1161/CIRCULATIONAHA.105.537878

[25] LIEB, Wolfgang a Ramachandran S. VASAN. Genetics of Coronary Artery Disease. *Circulation*. 2013, 128(10), 1131-1138. ISSN 0009-7322. DOI: 10.1161/CIRCULATIONAHA.113.005350

[26] LOLEY, Christina, Maris ALVER, Themistocles L. ASSIMES, et al. No Association of Coronary Artery Disease with X-Chromosomal Variants in Comprehensive International Meta-Analysis. *Scientific Reports*. 2016, **6**(1). ISSN 2045-2322. DOI: 10.1038/srep35278

[27] LOVREN, Fina, Hwee TEOH a Subodh VERMA. Obesity and Atherosclerosis: Mechanistic Insights. *Canadian Journal of Cardiology*. 2015, **31**(2), 177-183. ISSN 0828282X. Dostupné z: doi:10.1016/j.cjca.2014.11.031

[28] MACHIELA MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015 Nov 1;31(21):3555-7. doi: 10.1093/bioinformatics/btv402

[29] MAČÁK, Jiří, Jana MAČÁKOVÁ a Jana DVOŘÁČKOVÁ. *Patologie*. 2., dopl. vyd. Praha: Grada, 2012. ISBN 978-80-247-3530-6.

[30] MARANHÃO, Raul Cavalcante, Priscila Oliveira CARVALHO, Celia Cassaro STRUNZ and Fulvio PILEGGI. Lipoprotein (a): Structure, Pathophysiology and Clinical Implications. *Arquivos Brasileiros de Cardiologia*. 2010, 31(23), 2844-2853. ISSN 0066-782X. DOI: 10.5935/abc.20140101

[31] MCPHERSON, Ruth and Anne TYBJAERG-HANSEN. Genetics of Coronary Artery Disease. *Circulation Research*. 2016, 118(4), 564-578. ISSN 0009-7330. DOI: 10.1161/CIRCRESAHA.115.306566

[32] MERCADO, C., Jaimes, E.A. Cigarette smoking as a risk factor for atherosclerosis and renal disease: Novel pathogenic insights. *Current Science Inc* **9,** 66–72 (2007). https://doi.org/10.1007/s11906-007-0012-8

[33] NAŇKA, Ondřej, Miloslava ELIŠKOVÁ a Oldřich ELIŠKA. *Přehled anatomie*. 2., dopl. a přeprac. vyd. Praha: Galén, c2009. ISBN 978-80-7262-612-0.

[34] NELSON, Christopher P, Anuj GOEL, Adam S BUTTERWORTH, et al. Association analyses based on false discovery rate implicate new loci for

coronary artery disease. Nature Genetics [online]. 2017, 49(9), 1385-1391 [cit. 2020-12-28]. ISSN 1061-4036. DOI: 10.1038/ng.3913

[35] NING, Zheng, Youngjo LEE, Peter K. JOSHI, James F. WILSON, Yudi PAWITAN a Xia SHEN. A Selection Operator for Summary Association Statistics Reveals Allelic Heterogeneity of Complex Traits. *The American Journal of Human Genetics*. 2017, **101**(6), 903-912. ISSN 00029297. Dostupné z: doi:10.1016/j.ajhg.2017.09.027

[36] NORDESTGAARD, Børge G., M. John CHAPMAN, Kausik RAY, et al. Lipoprotein(a) as a cardiovascular risk factor: current status. *European Heart Journal*. 2010, 31(23), 2844-2853. ISSN 0195-668X. DOI: 10.1093/eurheartj/ehq386

[37] OHASHI, Ryuji, Hong MU, Qizhi YAO and Changyi CHEN. Cellular and Molecular Mechanisms of Atherosclerosis with Mouse Models. *Trends in Cardiovascular Medicine*. 2004, 14(5), 187-190. ISSN 10501738. DOI: 10.1016/j.tcm.2004.04.002

[38] ROSENSON, Robert S., Michael H. DAVIDSON, Benjamin J. HIRSH, Sekar KATHIRESAN and Daniel GAUDET. Genetics and Causality of Triglyceride-Rich Lipoproteins in Atherosclerotic Cardiovascular Disease. *Journal of the American College of Cardiology*. 2014, 64(23), 2525-2540. ISSN 07351097. DOI: 10.1016/j.jacc.2014.09.042

[39] SAXTON A, CHAUDHRY R, MANNA B. Anatomy, Thorax, Heart Right Coronary Arteries. [Updated 2020 Jul 31]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK537357/

[40] SCHUNKERT, Heribert, Inke R KÖNIG, Sekar KATHIRESAN, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genetics. 2011, 43(4), 333-338. ISSN 1061-4036. DOI: 10.1038/ng.784

[41] SLATKIN, Montgomery. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews*

Genetics [online]. 2008, **9**(6), 477-485 [cit. 2021-5-27]. ISSN 1471-0056. Dostupné z: doi:10.1038/nrg2361

[42] SUDMANT, Peter H., Tobias RAUSCH, Eugene J. GARDNER, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015, 526(7571), 75-81. ISSN 0028-0836. DOI: 10.1038/nature15394

[43] TADA, Hayato, Masa-aki KAWASHIRI, Atsushi NOHARA, Akihiro INAZU, Hiroshi MABUCHI a Masakazu YAMAGISHI. Impact of clinical signs and genetic diagnosis of familial hypercholesterolaemia on the prevalence of coronary artery disease in patients with severe hypercholesterolaemia. *European Heart Journal*. 2017, 38(20), 1573-1579. ISSN 0195-668X. DOI: 10.1093/eurheartj/ehx004

[44] The 1000 Genes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010, 467(7319), 1061-1073. ISSN 0028-0836. DOI: 10.1038/nature09534

[45] The 1000 Genomes Project Consortium., Corresponding Author., McVean, G. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012). DOI: 10.1038/nature11632

[46] VOGIATZI G, Tousoulis D, Stefanadis C. The role of oxidative stress in atherosclerosis. Hellenic J Cardiol. 2009 Sep-Oct;50(5):402-9. PMID: 19767282

[47] WEBB, Thomas R., Jeanette ERDMANN, Kathleen E. STIRRUPS, et al. Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease. Journal of the American College of Cardiology. 2017, 69(7), 823-836. ISSN 07351097. DOI: 10.1016/j.jacc.2016.11.056

[48] WINHAM, Stacey J., Mariza DE ANDRADE a Virginia M. MILLER. Genetics of cardiovascular disease: Importance of sex and ethnicity. *Atherosclerosis*. 2015, 241(1), 219-228. ISSN 0021-9150. DOI: 10.1016/j.atherosclerosis.2015.03.021

[49] YANG, Jian, Teresa FERREIRA, Andrew P MORRIS, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*. 2012, **44**(4), 369-375. ISSN 1061-4036. Dostupné z: doi:10.1038/ng.2213

[50] 1000 Genomes: A Deep Catalog Of Human Genome Variation [online]. EMBL-EBI [cit. 2021-01-01]. Available from: https://www.internationalgenome.org/

[51] 1000 Genomes Browser. NCBI [online]. Bethesda (MD): National Center for Biotechnology Information, U.S. National Library of Medicine [cit. 2020.12.30]. Available from: https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/help/

APPENDIX A – LD MATRIXES

LD matrixes for recessive loci

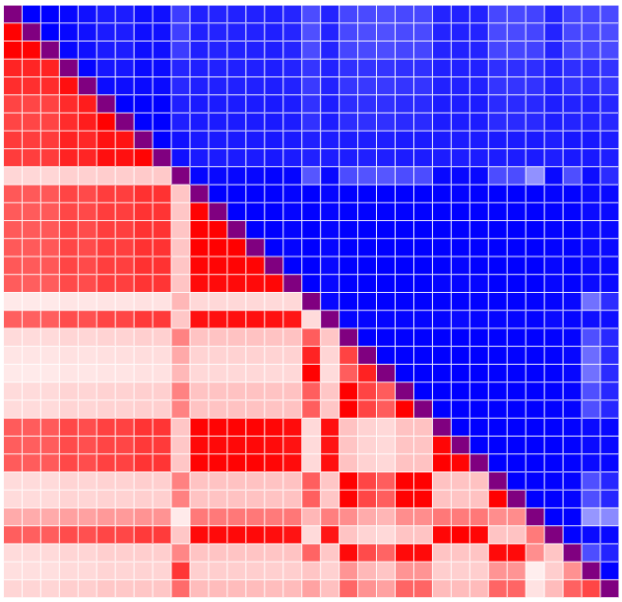


Figure 6: Chromosome 1 locus 2

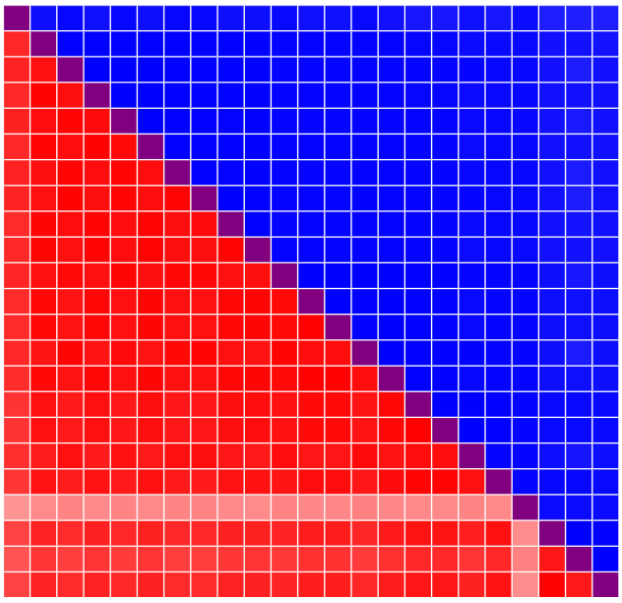
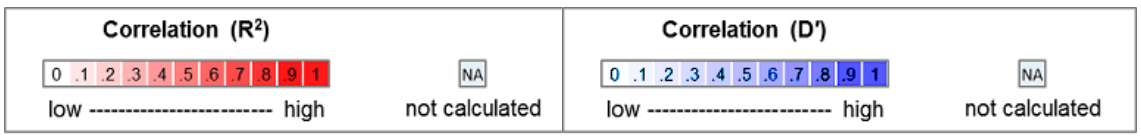


Figure 7: Chromosome 1 locus 3



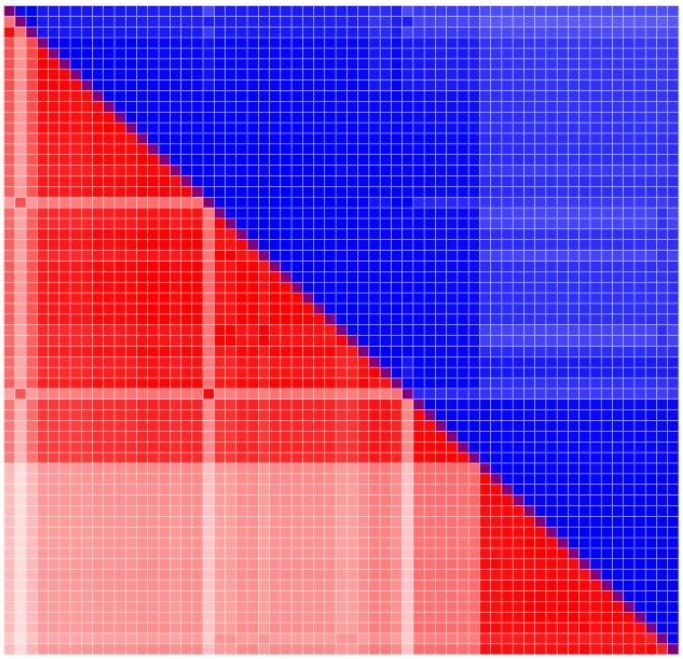


Figure 8: Chromosome 1 locus 4

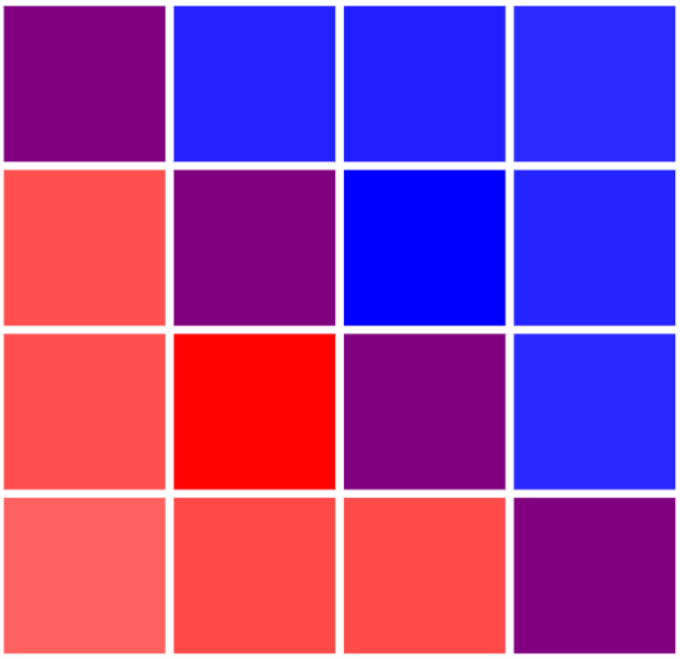
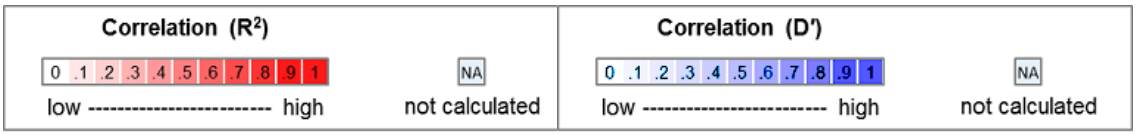


Figure 9: Chromosome 2 locus 1



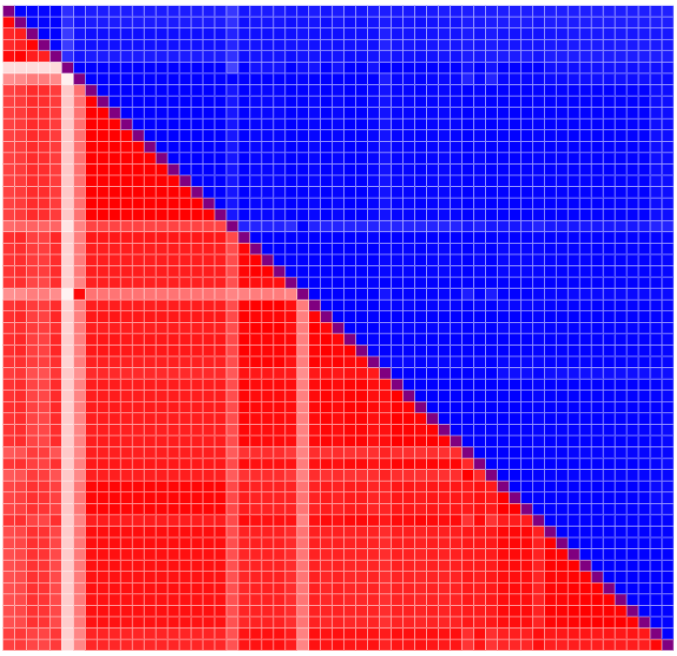


Figure 10: Chromosome 2 locus 2

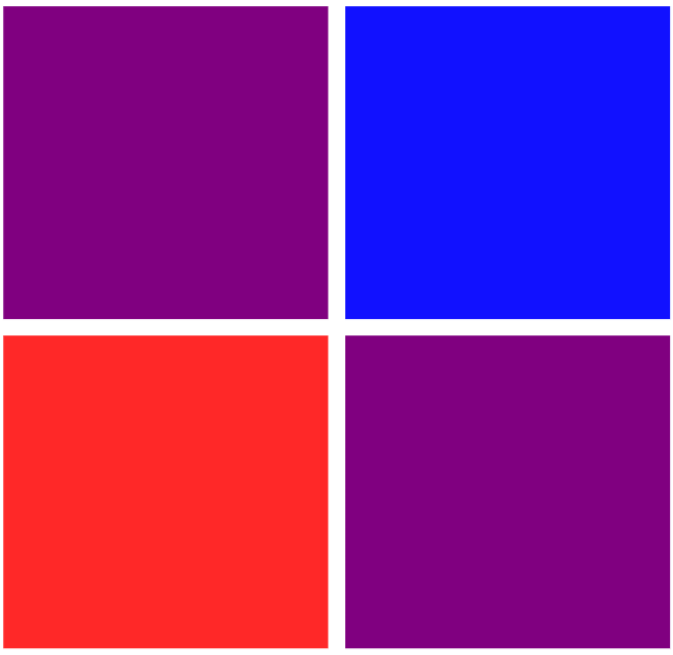
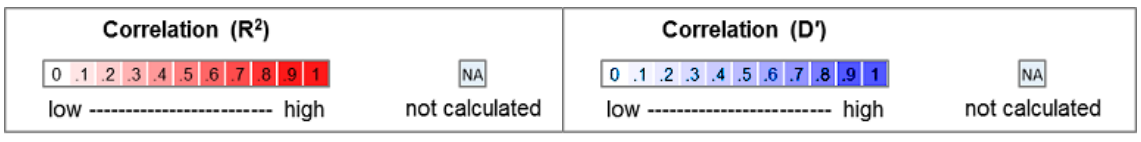


Figure 11: Chromosome 2 locus 3



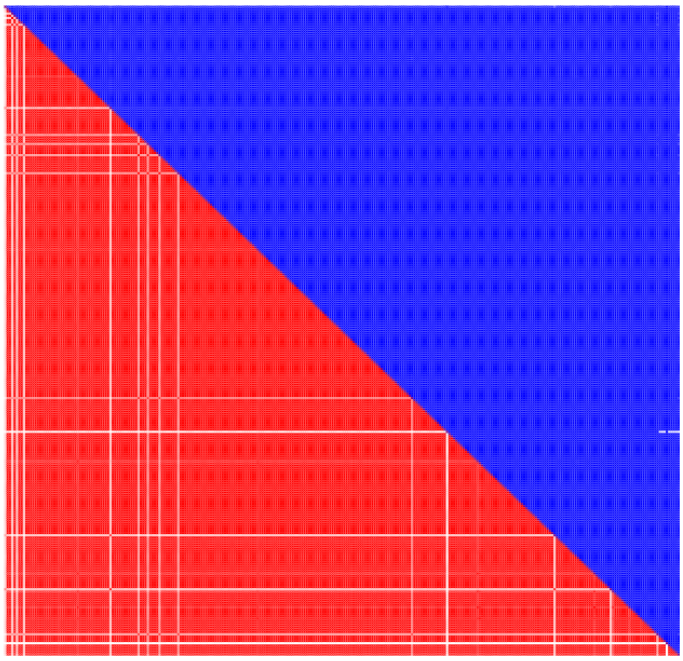


Figure 12: Chromosome 2 locus 4

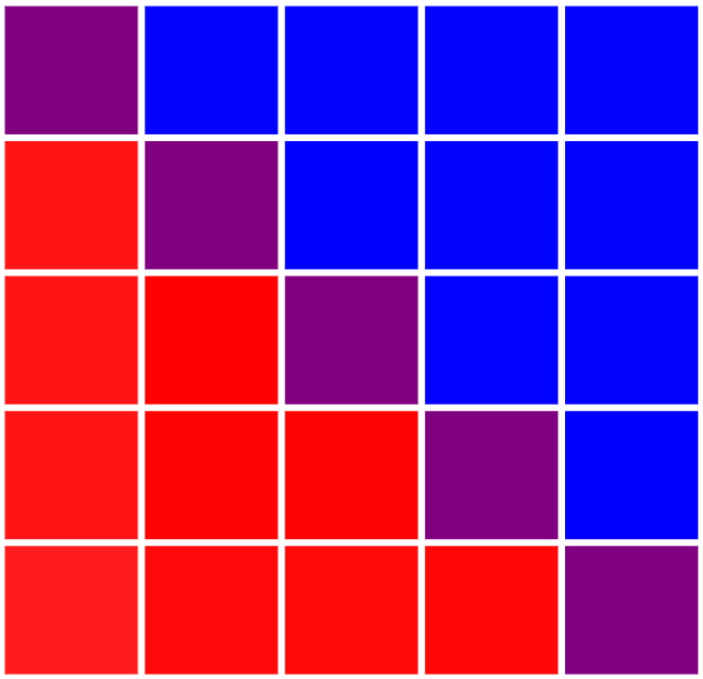
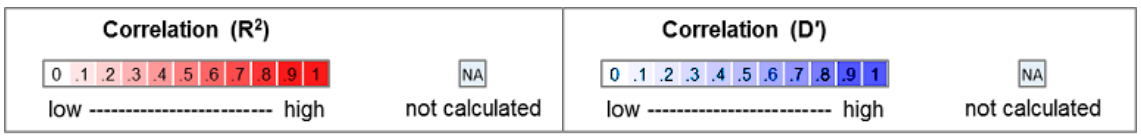


Figure 13: Chromosome 3 locus 1



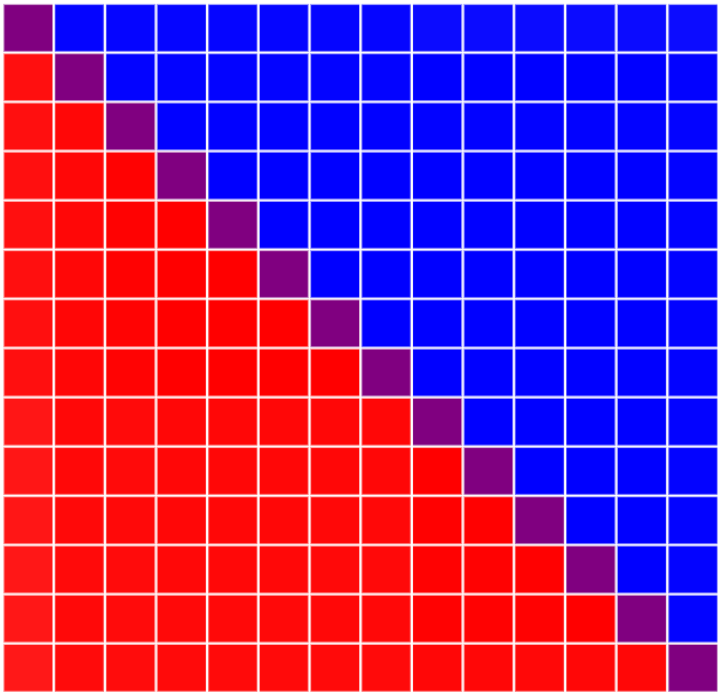


Figure 14: Chromosome 4 locus 1

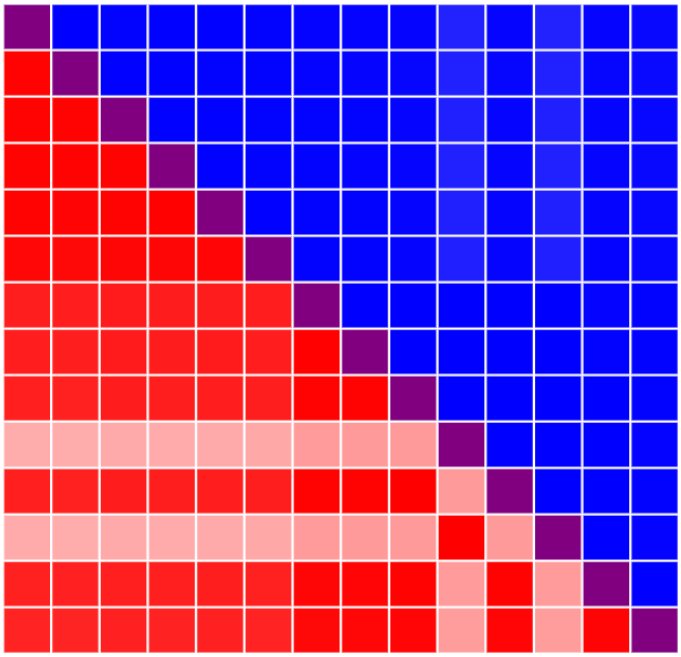
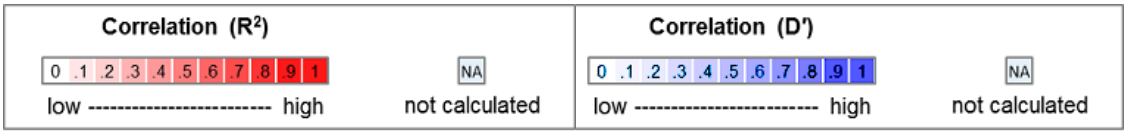


Figure 15: Chromosome 4 locus 2



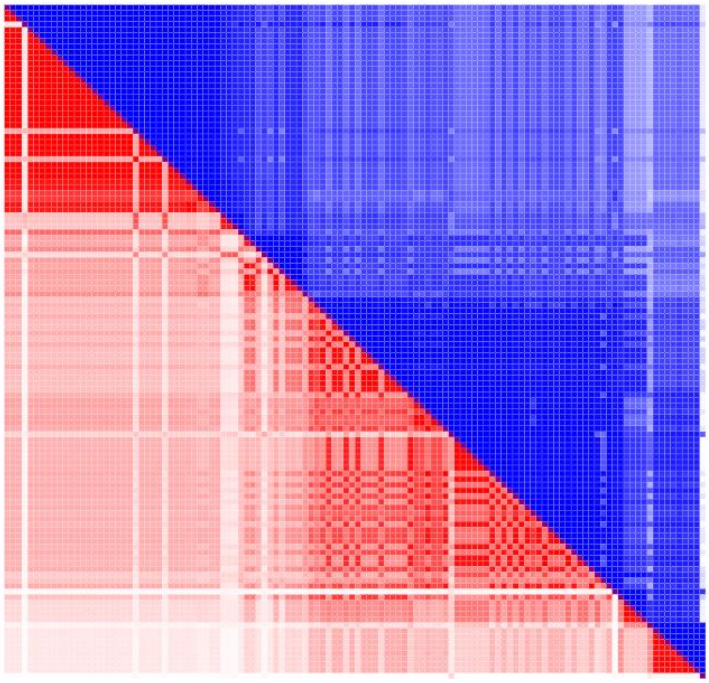


Figure 16: Chromosome 6 locus 1

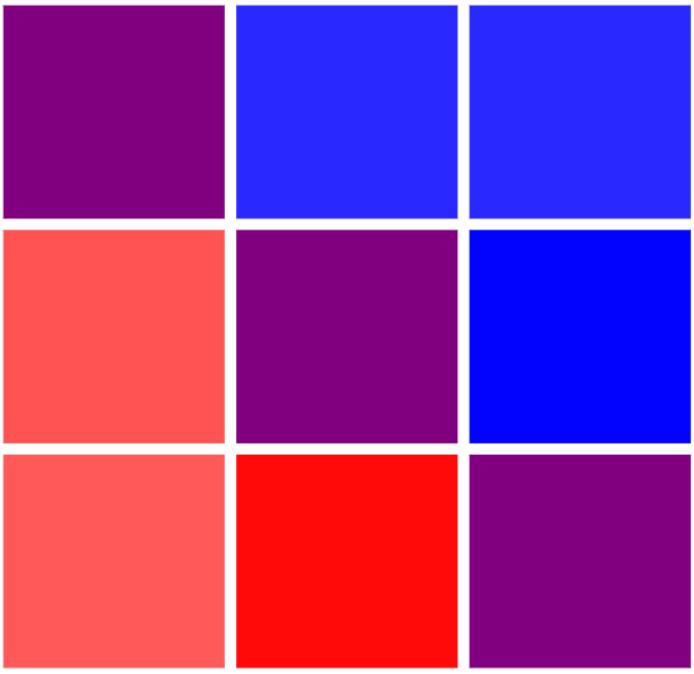
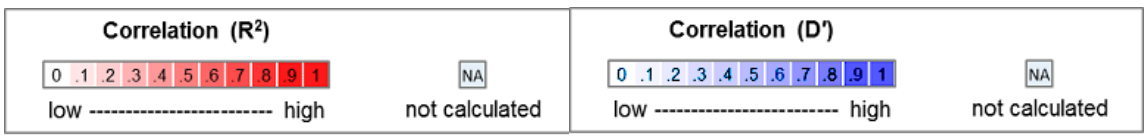


Figure 17: Chromosome 6 locus 1



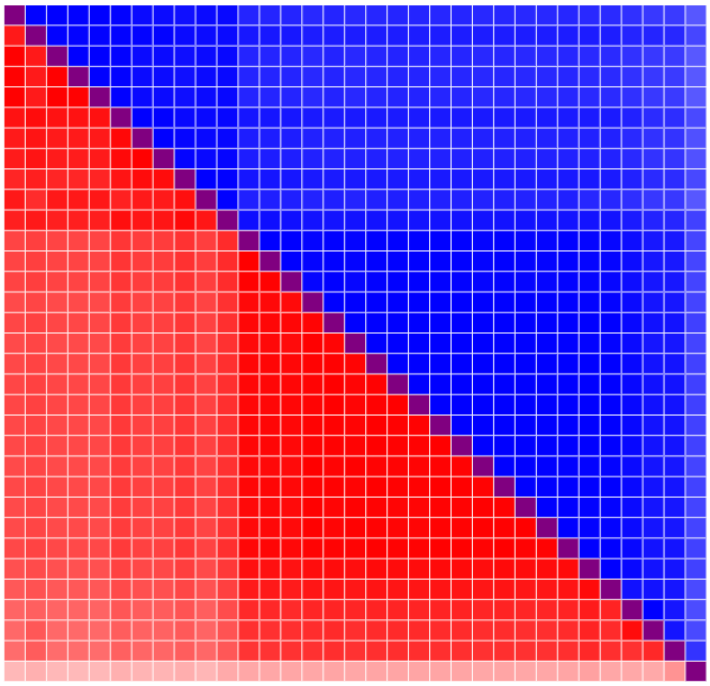


Figure 18: Chromosome 6 locus 3

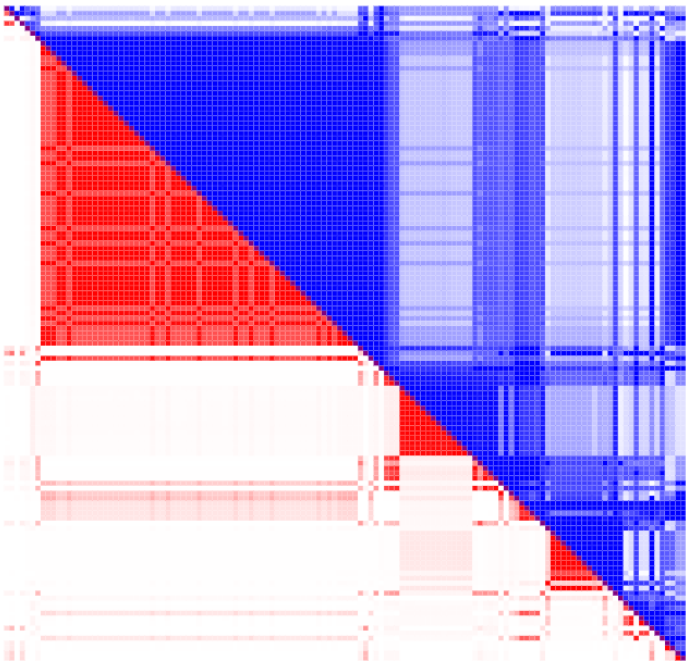
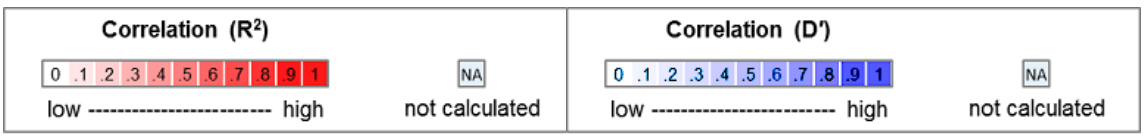


Figure 19: Chromosome 6 locus 4



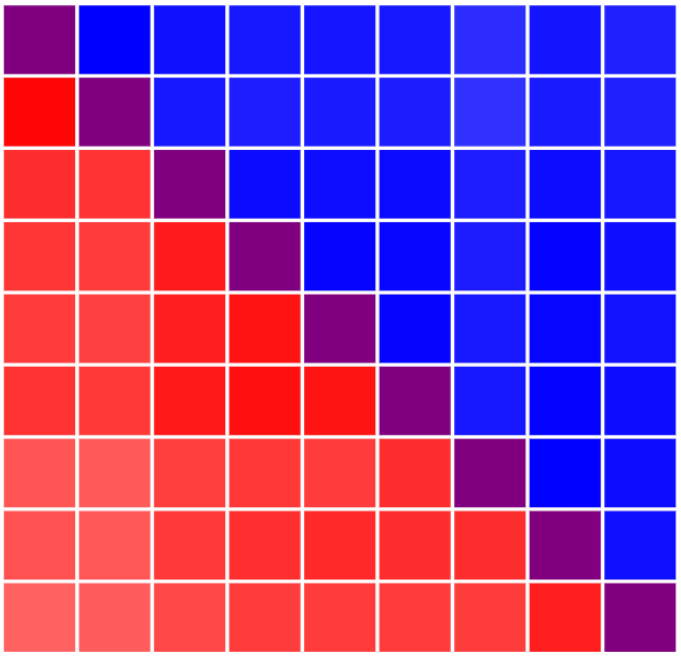


Figure 20: Chromosome 7 locus 1

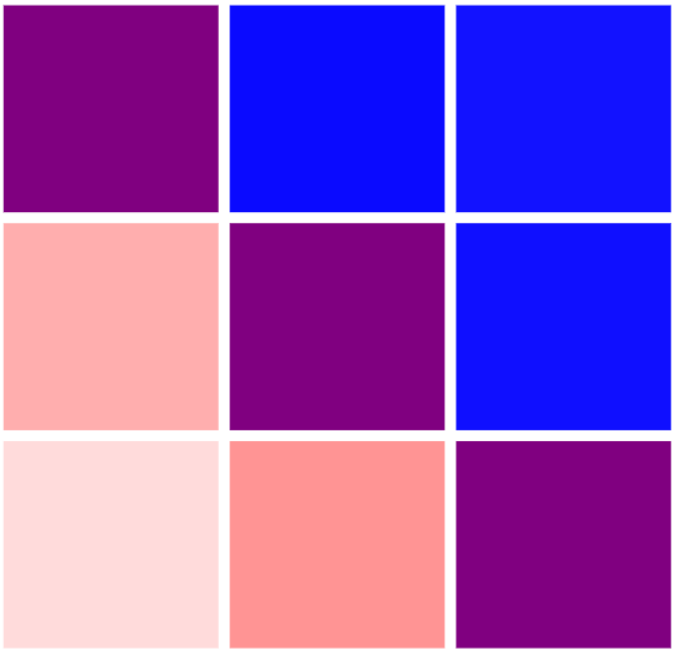
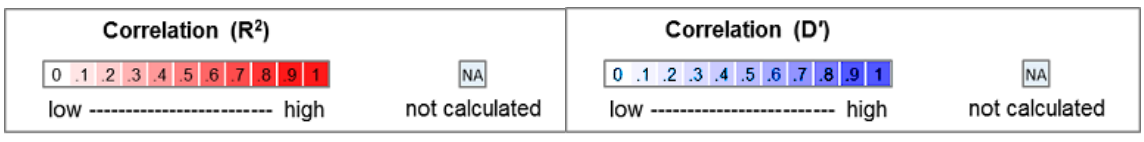


Figure 21: Chromosome 7 locus 2



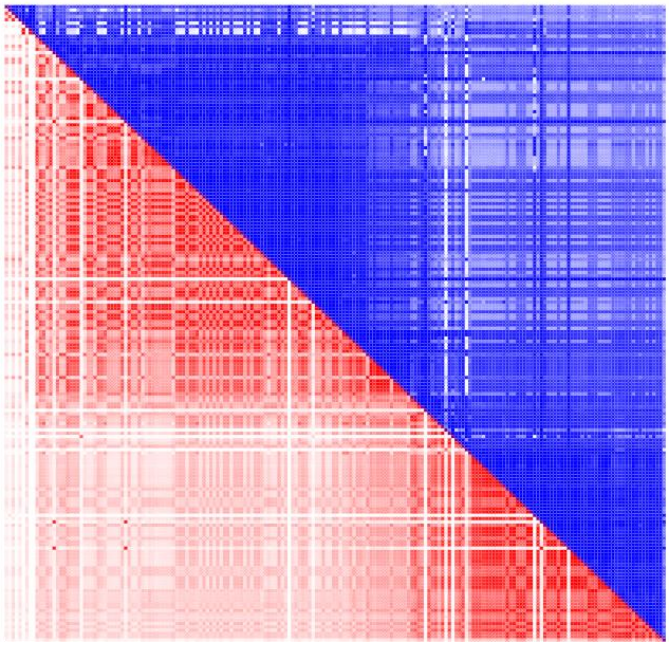


Figure 22: Chromosome 9 locus 1

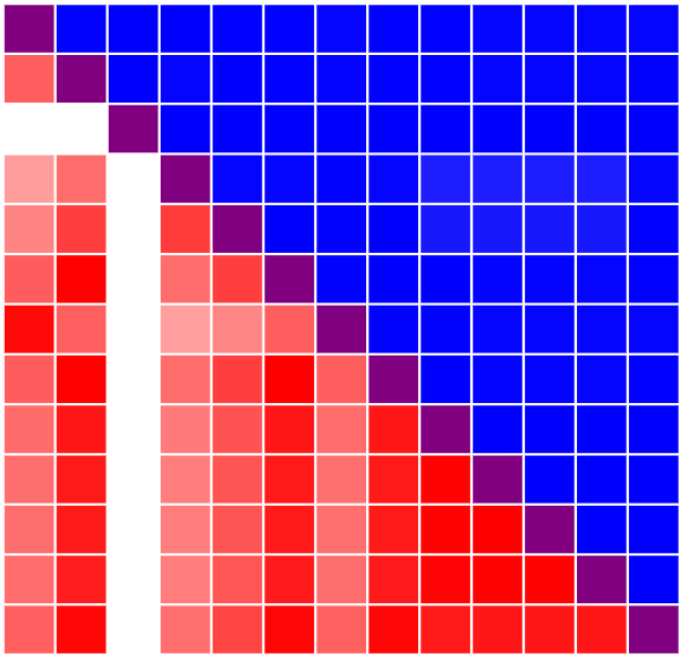
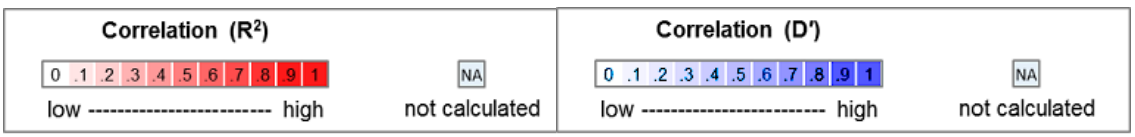


Figure 23: Chromosome 9 locus 2



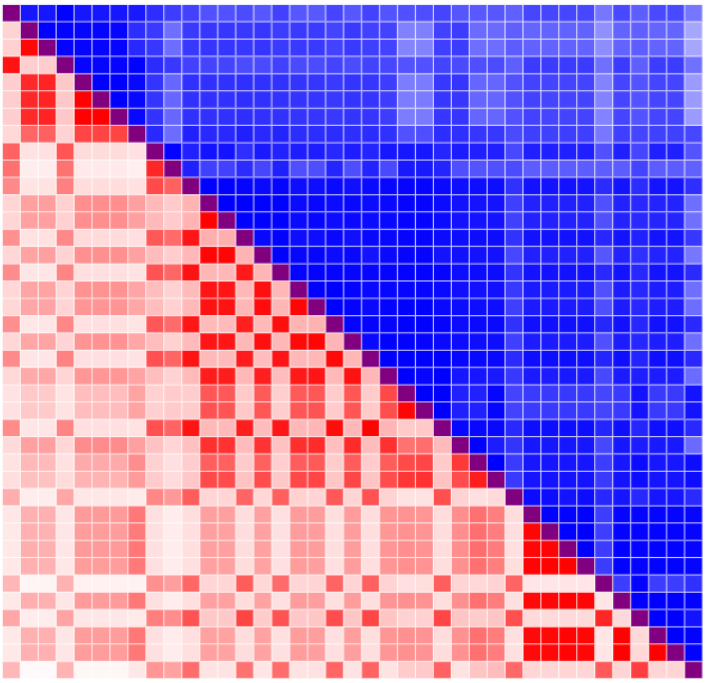


Figure 24: chromosome 10 locus 1

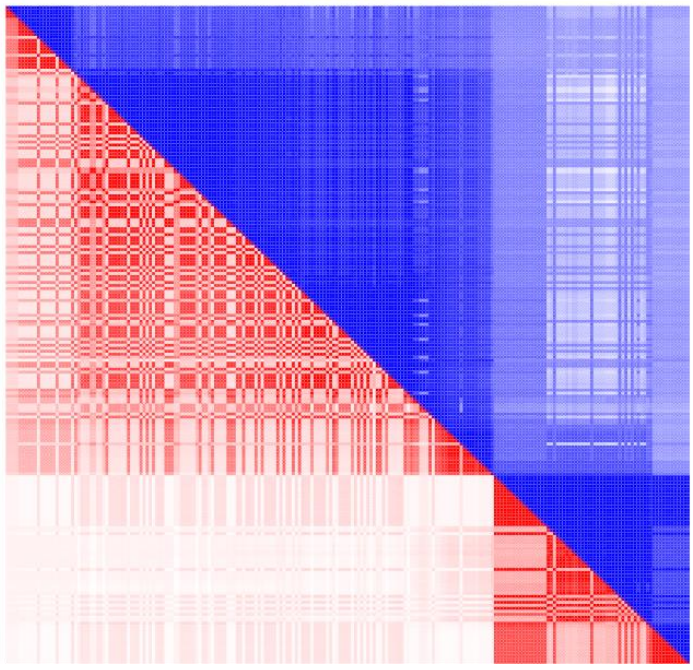
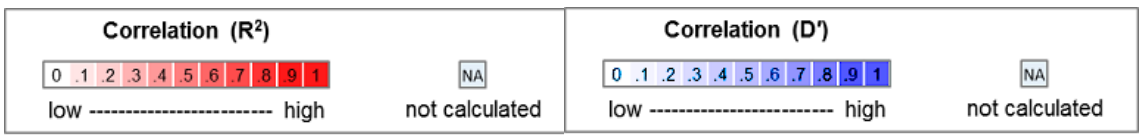


Figure 25: Chromosome 10 locus 2



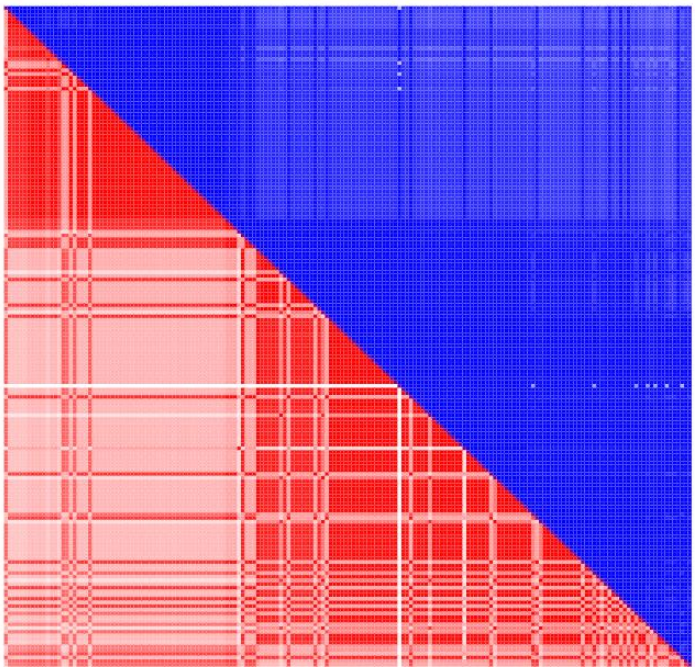


Figure 26: Chromosome 10 locus3

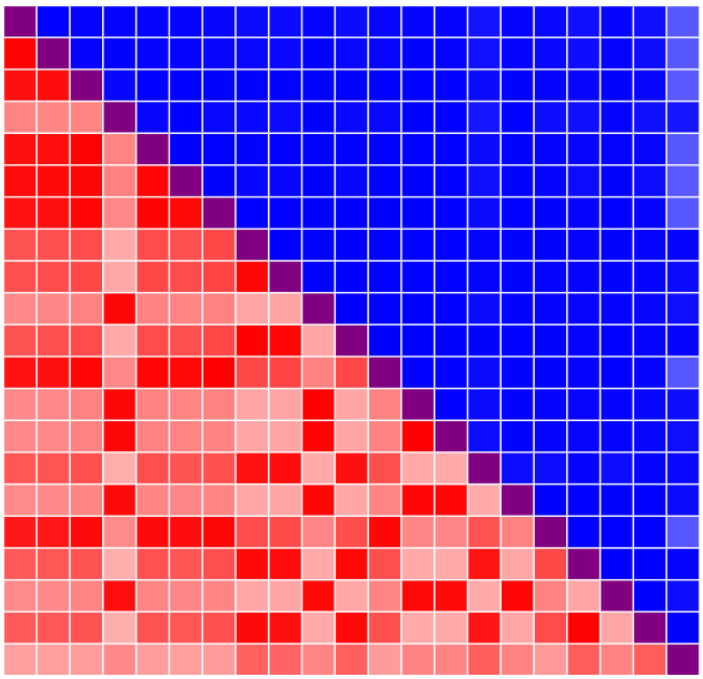
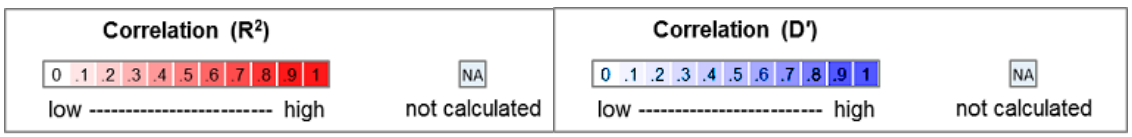


Figure 27: Chromosome 10 locus 4



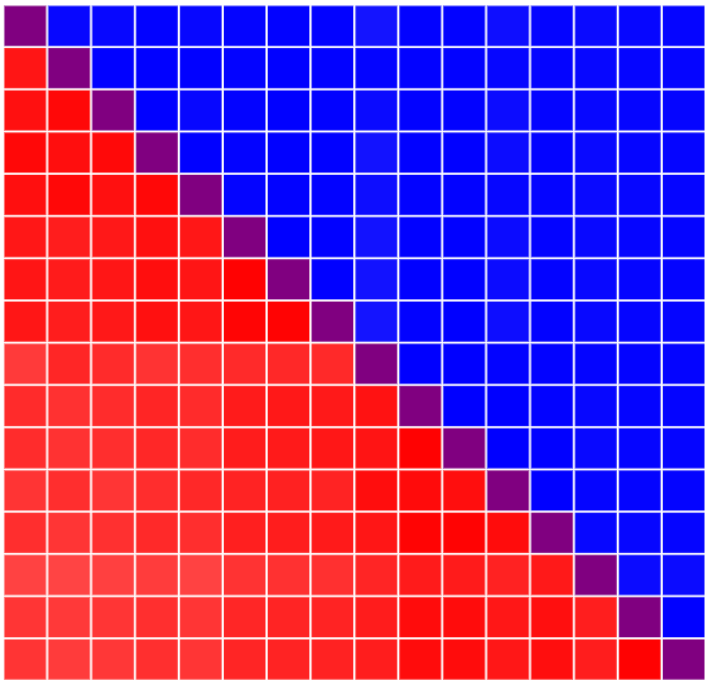


Figure 28: Chromosome 10 locus 5

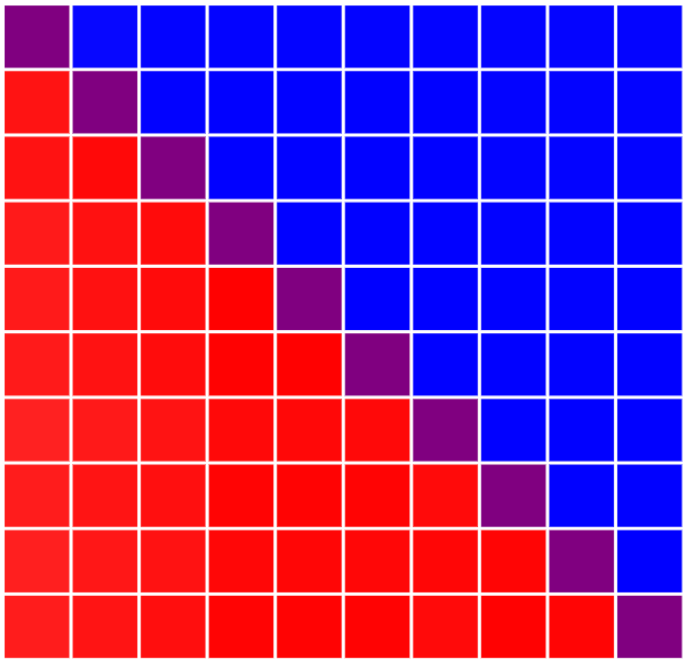
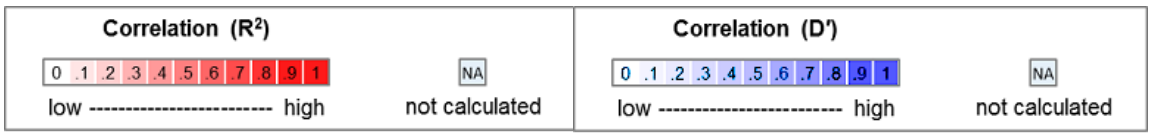


Figure 29: Chromosome 11 locus 1



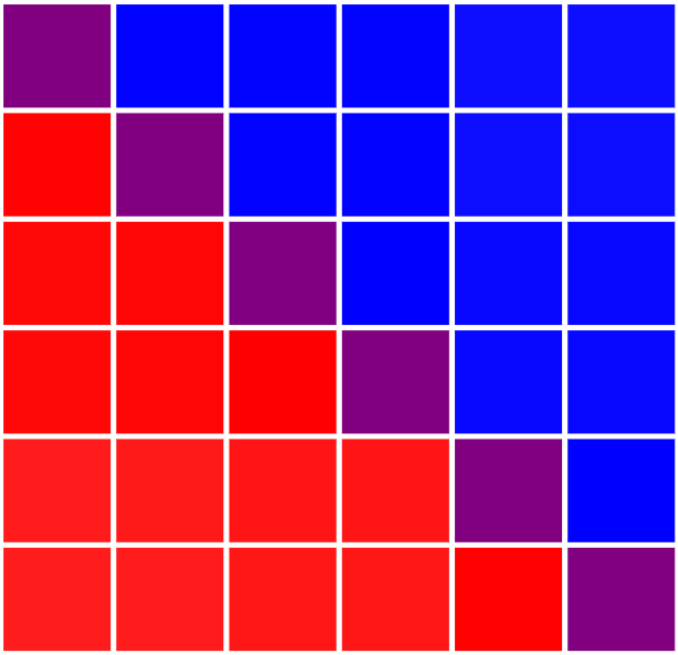


Figure 30: Chromosome 11 locus 2

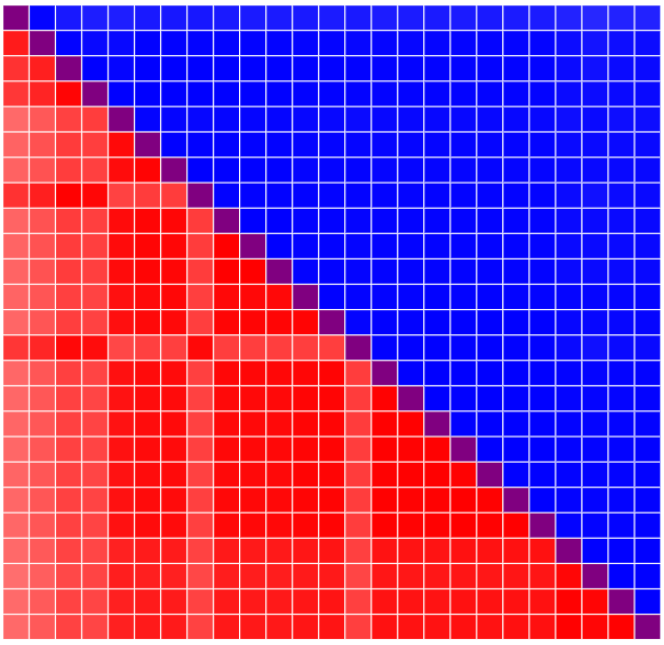
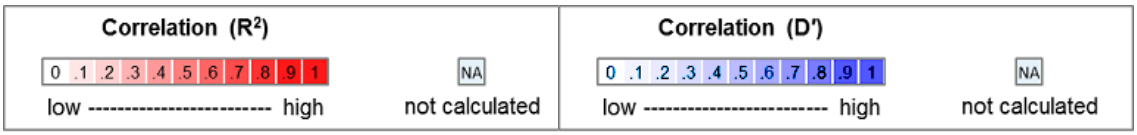


Figure 31: Chromosome 12 locus 1



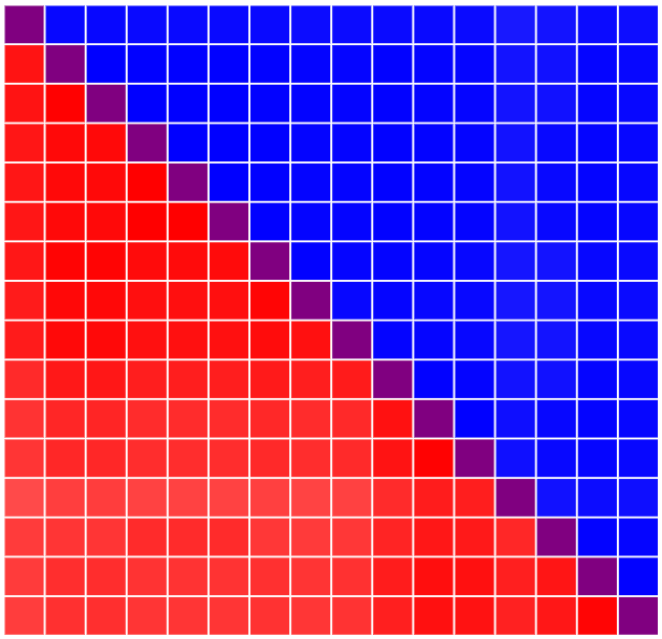


Figure 32: Chromosome 12 locus 2

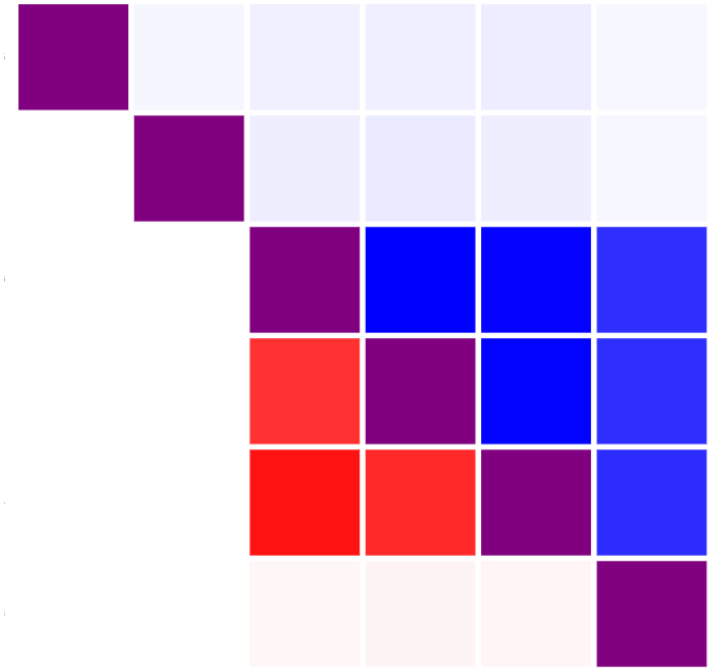
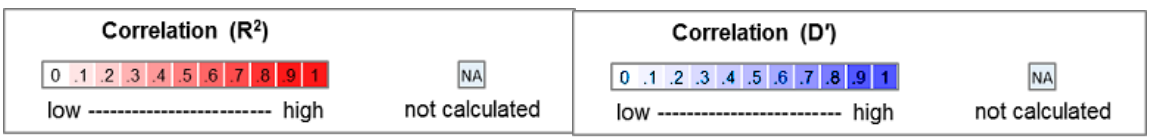


Figure 33: Chromosome 13 locus 1



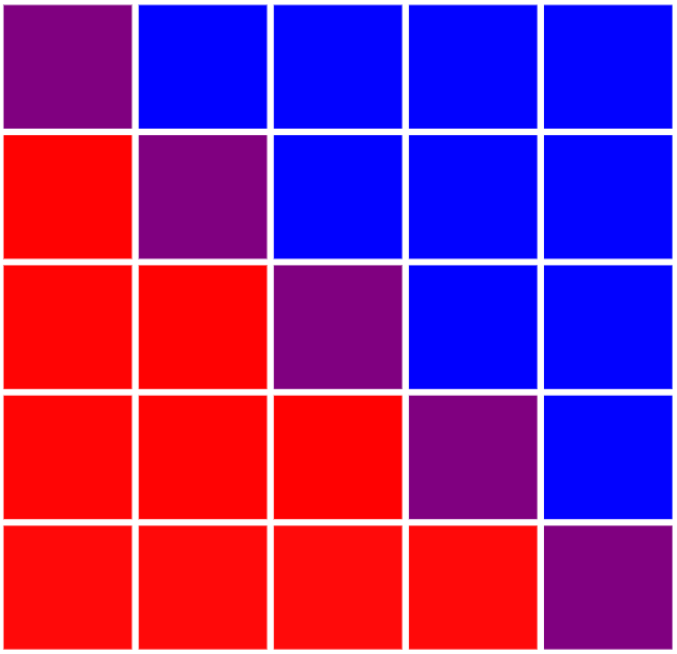


Figure 34: Chromosome 15 locus 1

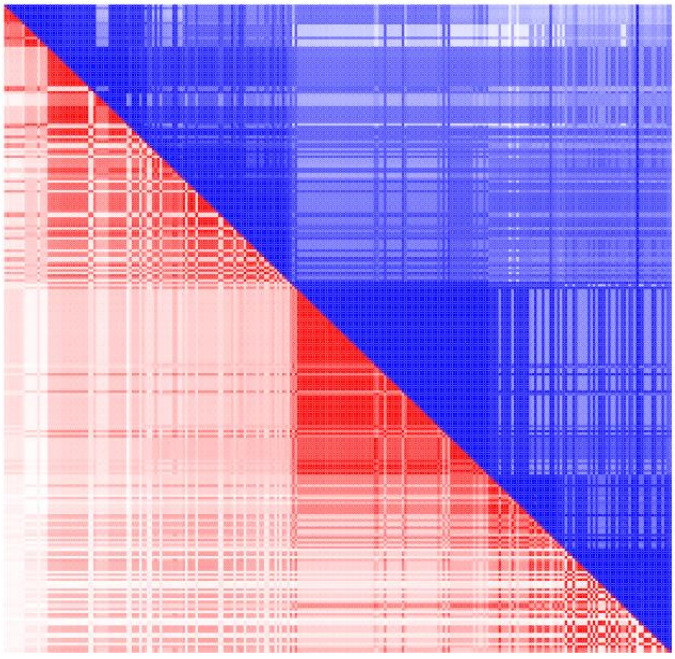
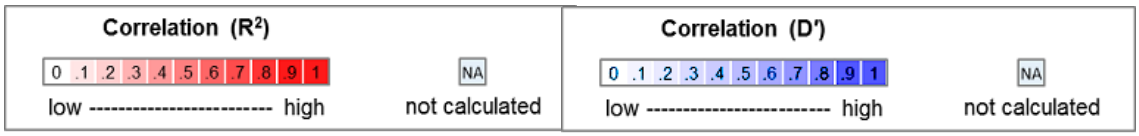


Figure 35: Chromosome 15 locus 2



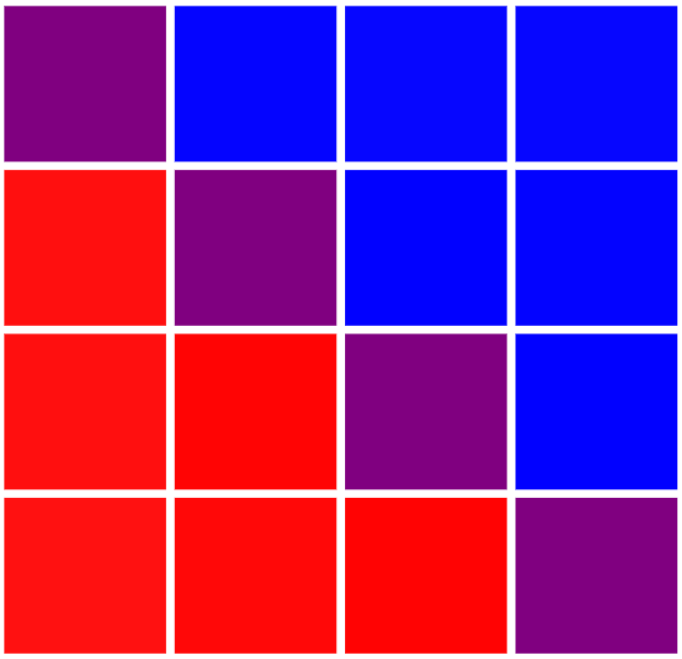


Figure 36: Chromosome 18 locus 1

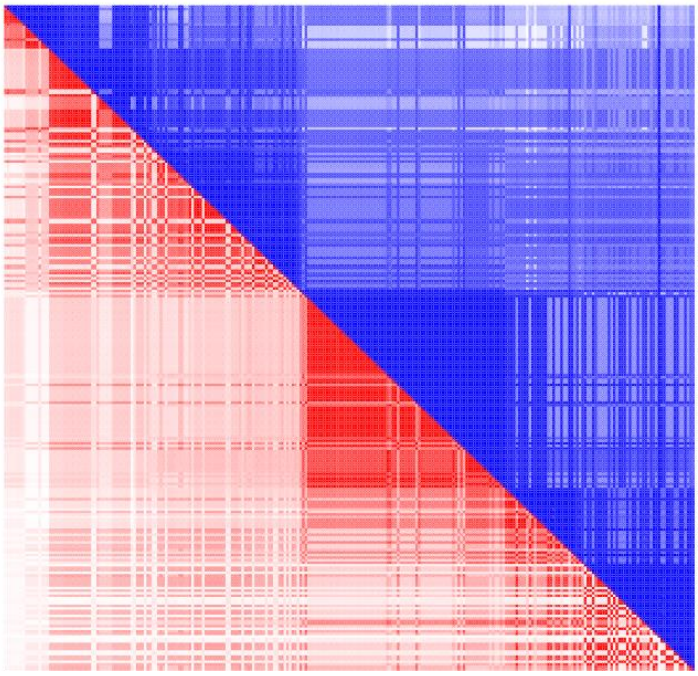
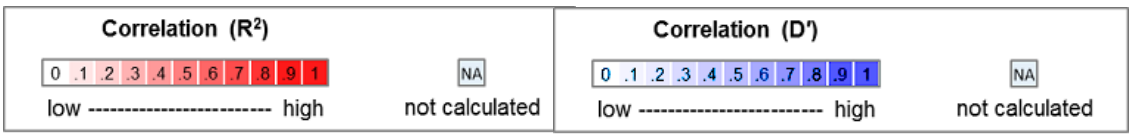


Figure 37: Chromosome 19 locus 1



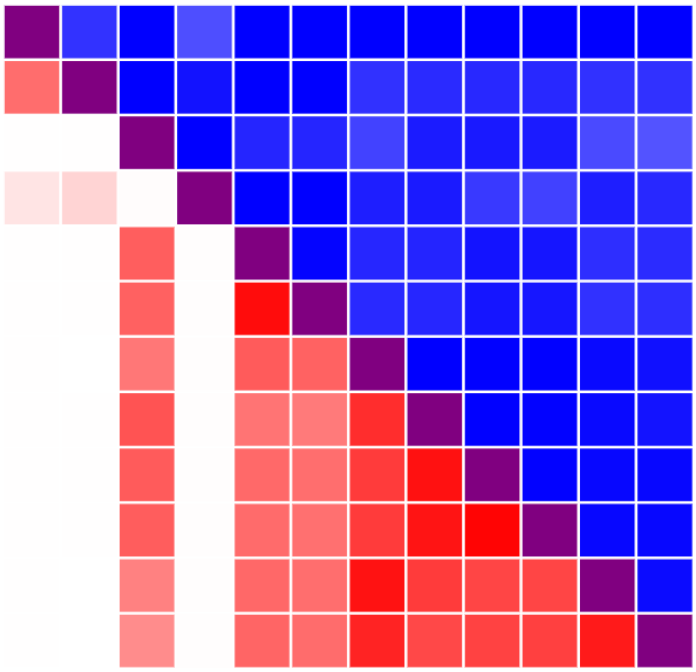


Figure 38: Chromosome 19 locus 2

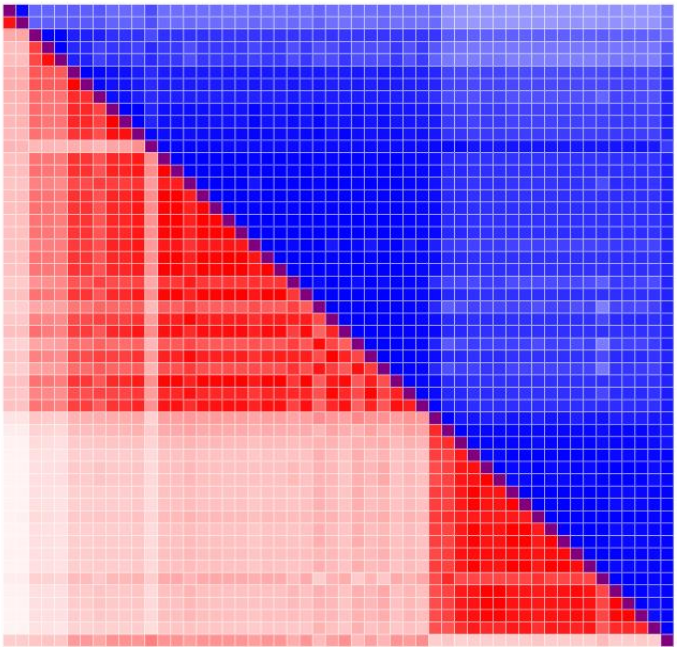


Figure 39: Chromosome 21 locus 1

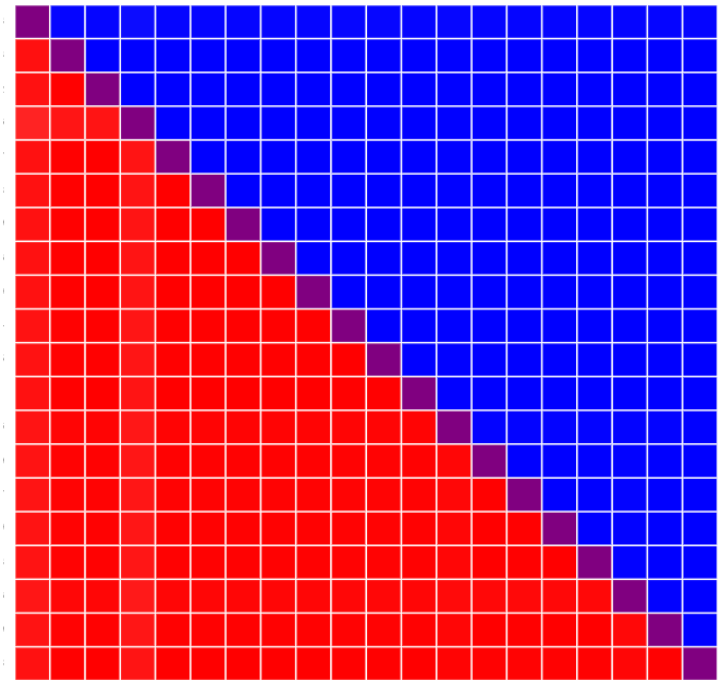
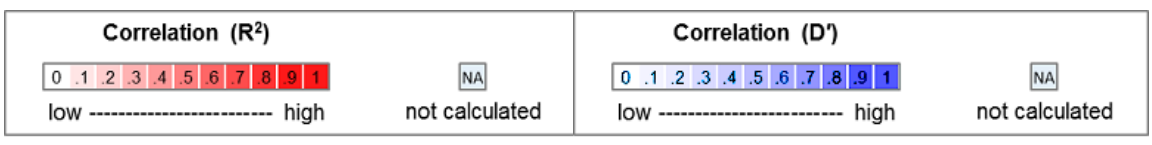


Figure 40: Chromosome 22 locus 1



LD matrixes for recessive loci

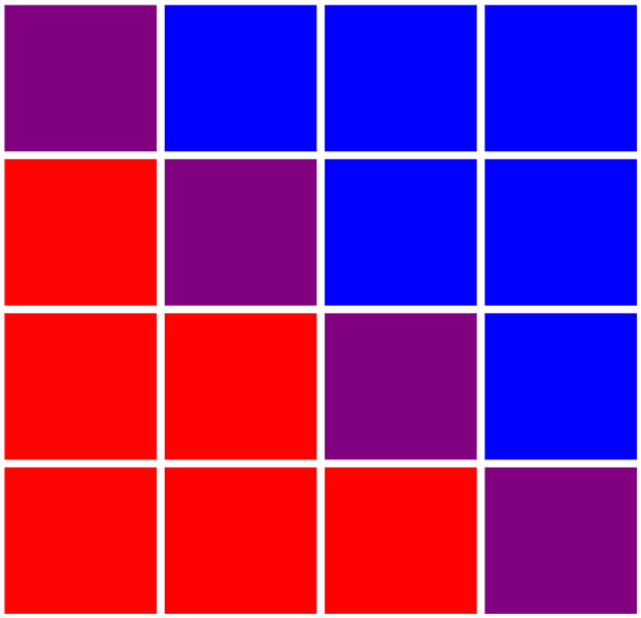


Figure 41: Chromosome 2 locus 1

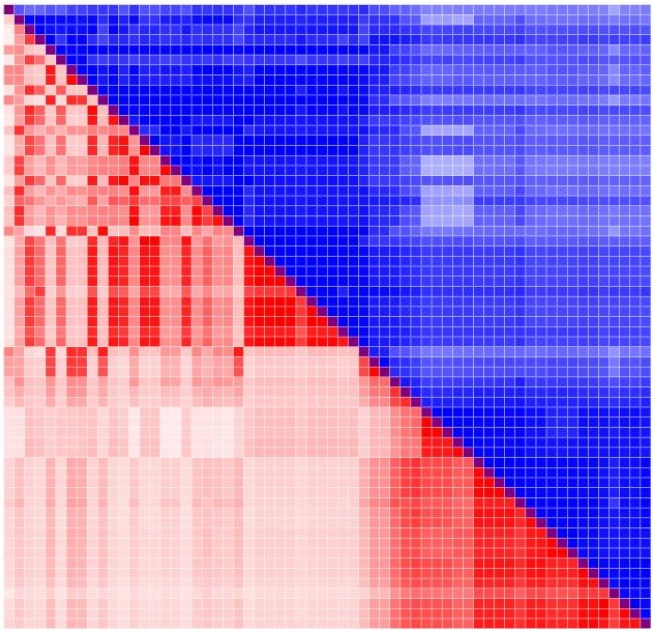
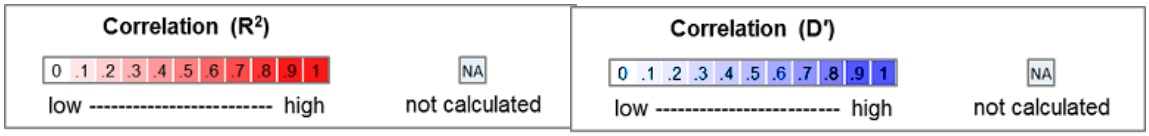


Figure 42: Chromosome 9 locus 1



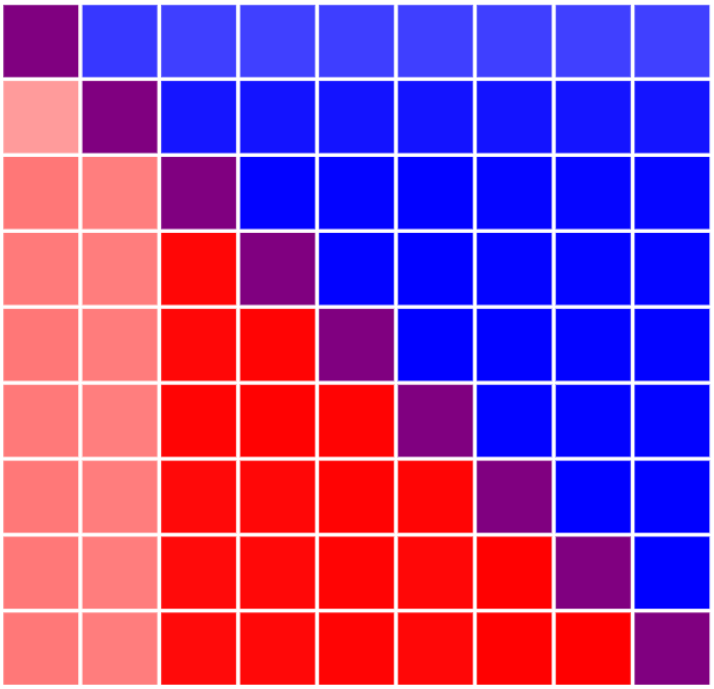


Figure 43: Chromosome 10 locus 1

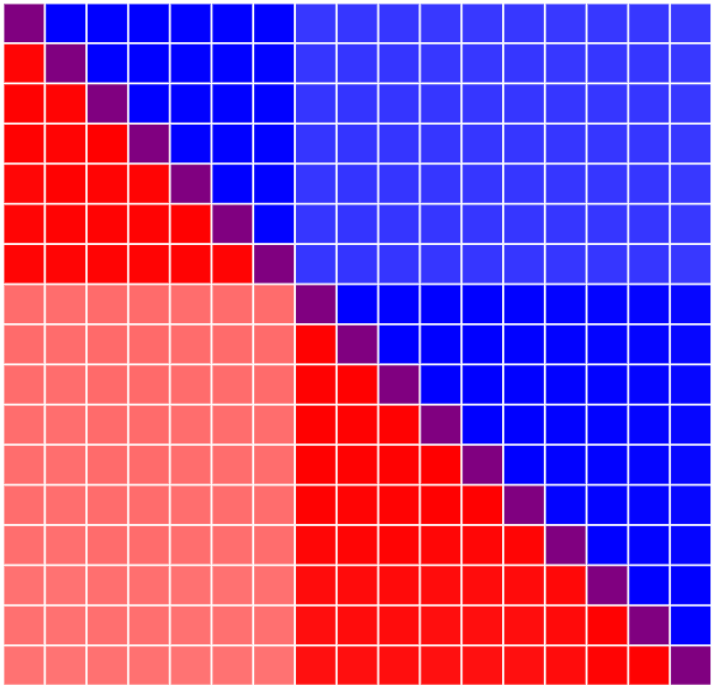
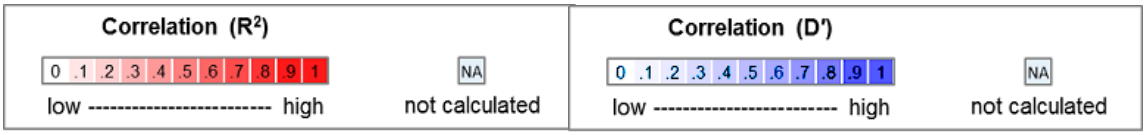


Figure 44: Chromosome 10 locus 2



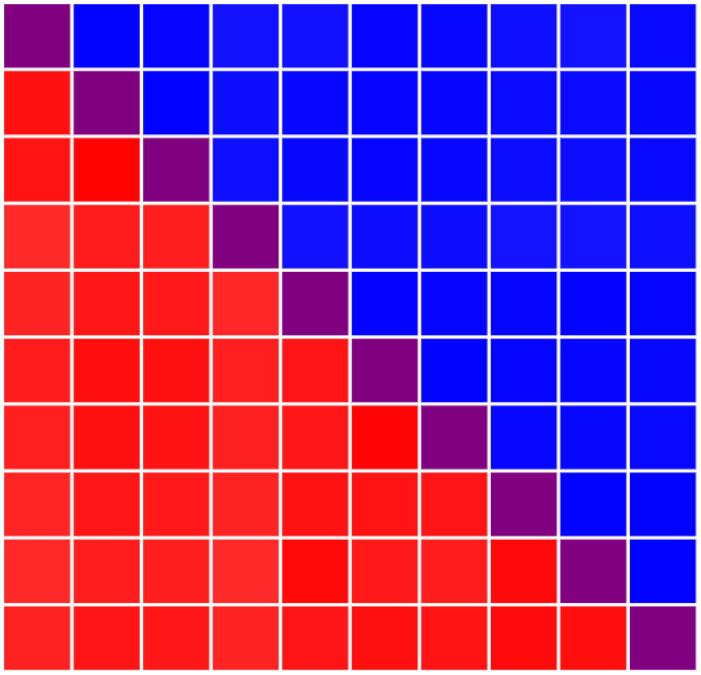


Figure 45: Chromosome 12 locus 1

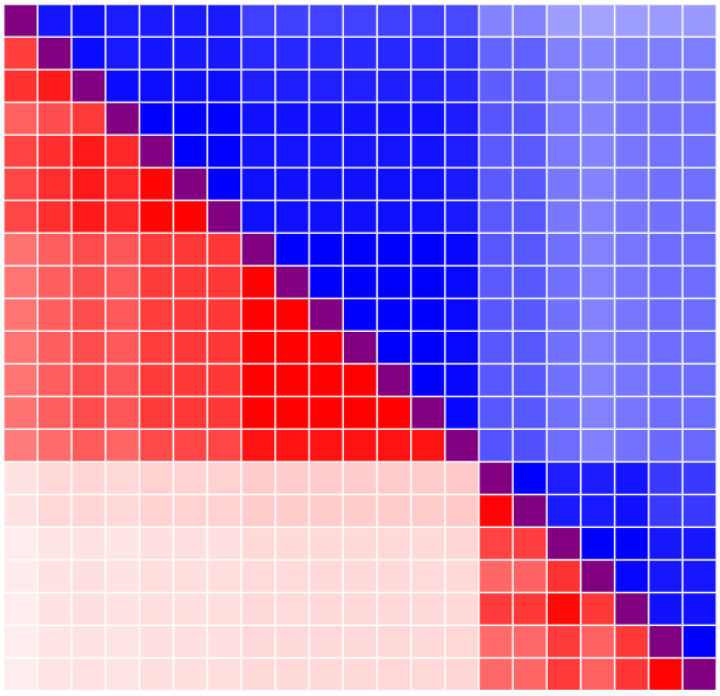
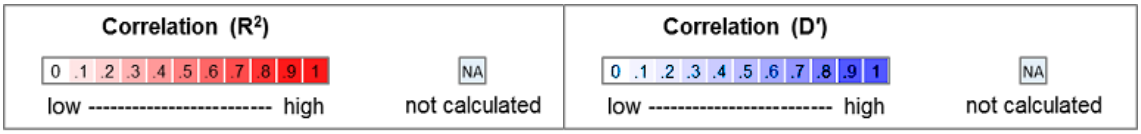


Figure 46: Chromosome 15 locus 1



APPENDIX B – SIGNIFICANT VARIANTS AFTER LASSO ANALYSIS

Table 4: Selected variants from the additive dataset

SNP	chr	A1	A2	bp	b
rs11206510	1	T	C	55496039	0,07452
rs72664304	1	A	G	56948290	0,01200
rs17114036	1	A	G	56962821	0,02714
rs72664318	1	A	G	56963627	0,01245
rs9970807	1	C	T	56965664	0,04962
rs17114046	1	A	G	56966350	0,01221
rs72664324	1	G	A	56972353	-0,03377
rs2404716	1	G	A	56979076	0,03005
rs55869368	1	C	T	56986633	-0,06146
rs72664335	1	G	T	56986978	0,00116
rs4634932	1	T	C	56996191	0,05062
rs953856	1	A	T	56997139	0,00214
rs10888977	1	C	T	56997893	0,03262
rs2017955	1	C	T	56998925	-5,92516
rs72664353	1	C	T	57001898	-0,00772
rs11206838	1	T	C	57012791	-0,00345
rs72664358	1	C	T	57015317	0,02306
rs56170783	1	A	C	57016131	0,05359
rs7551402	1	G	A	57016633	-0,00056
rs55751848	1	C	G	57018257	0,00982
rs72664362	1	A	C	57018260	0,01175
rs6657811	1	A	T	109807283	0,04859
rs6657811	1	A	T	109807283	-0,01829
rs4970834	1	C	T	109814880	0,02982
rs611917	1	A	G	109815252	0,00554
rs7528419	1	A	G	109817192	0,22841
rs11102967	1	T	C	109817245	-0,03918
rs629301	1	T	G	109818306	-0,02398
rs646776	1	T	C	109818530	-0,11258
rs3902354	1	A	C	109819296	0,04938
rs583104	1	T	G	109821307	-0,01232
rs602633	1	G	T	109821511	0,01720
rs4970836	1	A	G	109821797	-0,05371
rs4970837	1	T	G	109822008	-0,01434
rs1277930	1	A	G	109822143	0,00380
rs599839	1	A	G	109822166	0,05043
chr1:109826760	1	D	I	109826760	0,04714
rs11265611	1	A	G	154395125	-0,00047
rs6689306	1	G	A	154395946	-0,25094

SNP	chr	A1	A2	bp	b
rs12118721	1	T	C	154397416	2,03065
rs12118770	1	C	T	154397589	0,55980
rs12117832	1	G	A	154397610	-7,64391
rs10908838	1	G	T	154397984	-7,13790
rs4845618	1	T	G	154400015	-0,83065
rs6687726	1	G	A	154400320	-8,09391
rs6427658	1	C	T	154400799	1,34640
rs6694817	1	C	T	154401972	0,75504
rs7549250	1	T	C	154404336	7,80587
rs4845619	1	G	T	154405058	-8,70544
rs59632925	1	G	T	154406540	1,53815
rs6667434	1	G	A	154409100	3,30680
rs4553185	1	T	C	154410955	-2,98746
rs11265612	1	G	A	154417044	0,01894
rs11265613	1	T	C	154418415	0,08040
rs6686750	1	G	A	154419843	0,84903
rs4845625	1	C	T	154422067	-0,17869
rs6694258	1	A	C	154428505	-0,77856
rs4618978	1	C	G	222779187	-0,09254
rs34679168	1	C	T	222781932	0,12489
rs1909196	1	T	C	222786032	-0,21744
rs17163301	1	C	T	222794091	0,07015
rs4846770	1	G	C	222795569	0,08492
rs4846384	1	G	C	222797614	0,01222
rs17163313	1	G	T	222799625	0,03926
rs35626308	1	A	T	222816356	0,03127
rs61824282	1	C	T	222821211	-0,22529
rs17465637	1	C	A	222823529	0,03203
rs35158675	1	A	G	222829550	0,12877
rs4282830	1	T	G	222911960	-0,00104
rs36042506	1	D	I	222915810	-0,03690
rs61825259	1	C	T	222936769	-0,01649
rs145936486	1	A	C	222937591	0,06055
rs16986953	2	G	A	19942473	-0,08516
rs548145	2	C	T	21291312	0,00052
rs1712246	2	A	G	21383514	0,00162
chr2:44074126	2	I	D	44074126	0,06206
rs12618603	2	D	I	85741126	0,00833
rs201955338	2	I	D	85753465	-0,05230
rs6547620	2	C	T	85755928	-0,30595
rs70953953	2	D	I	85759591	-0,09143
rs6731005	2	T	C	85760395	1,86667

SNP	chr	A1	A2	bp	b
rs6705839	2	A	T	85761279	-0,80223
rs6705971	2	A	C	85761417	-4,49306
rs10176176	2	T	A	85762048	0,14355
rs3755014	2	T	C	85764006	0,53637
rs1446668	2	T	G	85764960	2,44511
rs2028900	2	C	T	85767735	-1,78123
rs7605975	2	T	C	85772548	2,79251
rs6547621	2	A	G	85774676	4,50553
rs141959654	2	D	I	85776499	-3,71881
rs7568458	2	T	A	85788175	-1,73720
rs10172544	2	C	A	85788270	1,15026
rs10187424	2	T	C	85794297	-0,80424
rs1009	2	A	G	85808737	0,09357
rs1561198	2	C	T	85809989	-1,17290
rs55971080	2	C	T	85812746	1,29377
rs17678683	2	T	G	145286559	-0,00808
rs72932588	2	C	G	203883193	-0,06752
rs115654617	2	C	A	203893999	-0,01897
rs72934535	2	T	C	203968973	-0,00631
rs72934591	2	G	A	204019762	-0,08871
rs139016349	3	D	I	138099161	-0,07444
rs17087335	4	G	T	57838583	-0,06076
rs4593108	4	C	G	148281001	0,07083
rs7688323	4	A	G	156616621	0,00349
rs3796592	4	T	C	156631345	0,00045
rs3796586	4	C	T	156638142	0,00233
rs72689147	4	G	T	156639888	0,00605
rs12202891	6	C	T	12768218	-0,01788
rs5874382	6	G	A	12788283	0,00874
rs9381401	6	A	C	12801967	-0,02173
rs9381462	6	G	A	12873775	0,03179
rs5874392	6	I	D	12874041	-0,04358
rs9472790	6	T	A	12879101	0,01579
rs1537340	6	G	A	12881855	0,00431
rs13208248	6	C	A	12883524	0,01037
rs4714955	6	C	T	12903435	0,00281
rs9349379	6	A	G	12903957	-0,15063
rs7454157	6	G	A	12909874	0,06763
rs4711863	6	G	C	12915417	0,00238
rs111395058	6	A	T	12921714	0,02432
rs1014342	6	C	T	12923157	0,02019
rs6458545	6	G	C	12934302	0,00634

SNP	chr	A1	A2	bp	b
rs13209002	6	C	T	12938341	-0,03961
rs4715000	6	A	G	12944118	0,01136
rs6929107	6	G	T	12949494	0,00289
rs36049381	6	G	A	12953384	0,00720
rs10948377	6	A	G	12987237	0,06586
rs56336142	6	T	C	39134099	0,01145
rs10456468	6	G	T	39151425	0,00639
rs4895388	6	G	A	134150561	0,00555
rs12193973	6	C	T	134151297	0,01745
rs1966248	6	A	T	134159622	0,01077
rs12202017	6	A	G	134173151	0,01869
rs1832181	6	C	G	134186165	-0,03711
rs12192720	6	G	A	134195719	0,01881
rs12194592	6	A	G	134204247	0,03618
rs6569912	6	C	T	134206802	-0,00724
rs2327429	6	T	C	134209837	0,00146
rs162185	6	T	C	134226147	0,03446
rs139311851	6	I	D	160427471	-0,15119
rs9457761	6	C	T	160626650	-0,12287
rs9457861	6	C	T	160626650	-0,06623
rs624249	6	C	A	160679400	0,05354
rs10080815	6	T	G	160687412	-0,09997
rs79390162	6	C	T	160719592	-0,04793
rs9295128	6	G	T	160751531	-0,20018
rs202097157	6	D	I	160776695	-0,42882
rs1510226	6	T	C	160816409	-0,06223
rs2292334	6	G	A	160858188	-0,00835
rs11267696	6	I	D	160887298	-0,04045
rs9457927	6	A	G	160910282	-0,10804
rs141273416	6	I	D	160963010	-0,00265
rs55730499	6	C	T	161005610	-0,12676
rs10455872	6	A	G	161010118	-0,07607
rs140570886	6	T	C	161013013	-0,13495
rs34448740	6	I	D	161028725	0,00024
rs13202636	6	T	C	161029728	0,03603
rs6905073	6	G	T	161067892	0,04666
rs4708876	6	G	C	161068320	-0,04015
rs56393506	6	C	T	161089307	-0,10086
rs12174589	6	A	T	161099044	0,03130
rs9457995	6	G	A	161102643	0,03270
rs2315065	6	C	A	161108144	-0,04874
rs6455698	6	T	G	161108803	0,01422

SNP	chr	A1	A2	bp	b
rs186696265	6	C	T	161111700	-0,32769
rs783144	6	C	T	161119239	0,01242
rs4252185	6	T	C	161123451	-0,08474
rs74625856	6	A	T	161313619	-0,07211
rs143843429	6	A	G	161383079	-0,02675
rs7788833	7	T	C	19034191	0,00421
rs7788972	7	T	A	19034280	-0,03174
rs2074633	7	T	C	19035920	-0,01044
rs7798197	7	A	G	19037661	0,00251
rs2023936	7	C	G	19039067	-0,01040
rs28688791	7	T	C	19039605	0,00861
rs2107595	7	G	A	19049388	-0,03188
rs57301765	7	G	A	19052733	-0,01542
rs11556924	7	C	T	129663496	0,01031
rs56179563	7	G	A	129685597	0,00380
rs3918226	7	C	T	150690176	-0,13332
rs2069418	9	C	G	22009698	0,06218
rs10115049	9	G	A	22032119	0,07690
rs2151280	9	A	G	22034719	0,00579
rs7028268	9	G	A	22048414	-0,01140
rs9632885	9	A	G	22072638	0,44713
rs10757270	9	A	G	22072719	-0,15987
rs1831733	9	T	C	22076071	-0,12273
rs10757271	9	G	A	22076795	0,11771
rs1537370	9	T	C	22084310	0,01034
rs1970112	9	T	C	22085598	-0,09321
rs9644859	9	G	A	22090521	-0,05666
rs7866503	9	G	T	22091924	-0,03371
rs2891168	9	A	G	22098619	-0,04160
rs7859727	9	T	C	22102165	0,18665
rs10757275	9	G	A	22106225	-0,00259
rs1537375	9	C	T	22116071	0,00720
rs10217586	9	T	A	22121349	0,44697
rs1333046	9	T	A	22124123	-0,19950
rs10757277	9	A	G	22124450	-0,00608
rs1333048	9	A	C	22125347	-0,04443
rs8176685	9	I	D	136138765	0,08977
rs2519093	9	C	T	136141870	-0,09287
rs28850884	9	I	D	136145424	-0,06607
rs9411378	9	C	A	136145425	-0,03043
rs550057	9	C	T	136146597	0,00182
rs507666	9	G	A	136149399	0,04112

SNP	chr	A1	A2	bp	b
rs8176643	9	I	D	136149709	-0,11265
rs600038	9	T	C	136151806	0,00836
rs495828	9	G	T	136154867	-0,03739
rs635634	9	C	T	136155000	0,03802
rs1774240	10	C	T	30291963	-0,00836
rs11007851	10	C	T	30292597	0,03466
rs1774241	10	C	G	30294542	-0,03009
rs4749520	10	G	A	30296358	0,09874
rs10826749	10	T	C	30298522	0,00076
rs12259037	10	G	A	30300079	-0,02921
rs5784169	10	I	D	30310072	-0,11726
rs2014144	10	C	T	30310120	-0,11850
rs11007864	10	C	T	30313117	0,04077
rs3739998	10	C	G	30316072	0,00986
rs7920686	10	G	A	30317826	-0,09146
rs10826751	10	T	C	30319861	-0,48428
rs2066333	10	T	C	30322878	0,80977
rs2066334	10	A	G	30323189	-0,13327
rs11007869	10	C	T	30323410	0,03293
rs10826753	10	T	C	30326776	-0,28984
rs943440	10	G	A	30327316	0,12498
rs943438	10	T	C	30327735	0,21793
rs4749524	10	G	C	30332618	-0,00749
rs2505084	10	G	A	30335520	-0,08145
rs1870635	10	T	C	44480694	0,11877
rs1870634	10	G	T	44480811	0,09035
rs988739	10	A	G	44481673	0,33975
rs4145117	10	G	A	44483521	-0,00037
rs4948807	10	A	T	44490428	-0,01402
rs10793513	10	A	G	44494546	0,00946
rs10793515	10	C	T	44496985	-0,50812
rs11351103	10	I	D	44707570	0,01166
rs146583321	10	I	D	44738619	0,03404
rs617542	10	G	A	44743636	0,00538
rs200695234	10	I	D	44750411	0,04226
rs2437935	10	A	G	44752268	0,01051
rs471451	10	G	C	44754785	0,01672
rs627135	10	T	G	44765000	0,00800
rs11238961	10	C	G	44765095	0,02101
rs9703833	10	C	T	44765933	0,00216
rs1746050	10	C	A	44777188	0,01894
rs1412445	10	C	T	91002804	-0,00629

SNP	chr	A1	A2	bp	b
rs1412444	10	C	T	91002927	-0,03087
rs1332329	10	A	C	91003419	-0,04676
rs2246949	10	T	C	91004668	0,07181
rs2246941	10	C	A	91004916	-0,05716
rs2246833	10	C	T	91005854	-0,31371
rs2246828	10	G	A	91005952	-0,00629
rs1051338	10	T	G	91007360	-0,08348
rs2250781	10	A	C	91007470	0,09574
rs2250645	10	C	T	91008873	0,09000
rs2250644	10	C	T	91008879	0,34510
rs2250398	10	G	A	91010157	-0,03642
rs2243548	10	T	C	91010313	0,01792
rs2243547	10	A	C	91010479	0,04895
rs10612161	10	D	I	91011255	-0,06166
rs1332328	10	C	T	91011458	-0,03133
rs1332327	10	C	T	91011681	0,01006
rs1332326	10	G	T	91011692	0,08675
rs74863780	10	I	D	91012222	-0,00781
rs2266008	10	A	G	91014061	-0,13095
rs17115100	10	G	T	104591393	0,04757
rs1004467	10	A	G	104594507	0,02238
rs138009835	10	G	A	104598995	-0,05239
rs11191416	10	T	G	104604916	0,03248
rs150786824	10	G	A	104612941	0,04135
rs3824754	10	C	T	104614350	-0,04781
rs11191425	10	C	T	104625970	-0,01357
rs7098825	10	T	C	104628234	0,01617
rs17878846	10	A	T	104630412	0,00795
rs77335224	10	C	T	104636276	0,03914
rs12360647	11	G	A	9751089	-0,84216
rs10840293	11	A	G	9751196	0,37104
rs10840294	11	T	G	9751246	0,38065
rs9943599	11	T	C	9752741	0,01128
rs360157	11	T	C	9754221	0,08659
rs974819	11	C	T	103660567	0,00558
rs2019090	11	T	A	103668962	-2,27502
rs2128739	11	C	A	103673277	-0,03918
rs1384705	11	T	C	103696851	-0,01003
rs7314459	12	T	A	89956341	-0,03910
rs11105337	12	A	T	89993507	-0,00992
rs2681472	12	A	G	90008959	-0,02957
rs57481061	12	C	G	90019178	0,01040

SNP	chr	A1	A2	bp	b
rs11105376	12	G	A	90088768	-0,02235
rs35350651	12	I	D	111907431	-0,02943
rs653178	12	T	C	112007756	-0,00366
rs11065979	12	C	T	112059557	-0,02928
rs11065987	12	A	G	112072424	0,01218
rs11513729	12	C	T	112273499	-0,02213
rs11617955	13	T	A	110818102	0,02836
rs4773141	13	C	G	110954353	-0,02642
rs11838776	13	G	A	111040681	-0,01920
rs9515201	13	C	A	111040798	-0,01013
rs9515203	13	T	C	111049623	0,02726
rs10139550	14	C	G	100145710	-0,05538
rs72743461	15	C	A	67441750	0,01507
rs17293632	15	C	T	67442596	0,04049
rs56375023	15	G	A	67448363	-0,02610
rs56062135	15	C	T	67455630	0,04114
rs7177201	15	C	T	79065380	-0,00335
rs35934157	15	A	G	79067385	-0,06455
rs12438645	15	G	A	79068203	0,01852
rs62012629	15	C	A	79070351	-0,00534
rs12907764	15	G	A	79070438	-0,10136
rs138914938	15	G	C	79073454	0,08234
rs35474770	15	A	G	79074000	0,00557
rs4887109	15	C	T	79074294	0,07516
rs2869862	15	A	G	79076885	-0,00611
rs3743058	15	C	T	79080798	-0,00390
rs28505515	15	C	G	79114477	-0,02549
rs7182103	15	T	G	79123946	0,00776
rs4468572	15	C	T	79124475	0,01557
rs62011004	15	T	C	79126536	0,10191
rs8024048	15	G	A	79127907	-0,02393
rs6495338	15	C	T	79128030	-0,00472
rs9920324	15	T	A	79129142	-0,02154
rs8030937	15	G	T	79129587	-0,05544
rs11072810	15	T	C	79132206	-0,02262
rs7173743	15	T	C	79141784	0,03003
rs8042271	15	A	G	89574218	-0,09671
rs7212798	17	T	C	59013488	-0,07996
rs663129	18	G	A	57838401	-0,05816
rs2287029	19	C	T	10916684	0,01677
rs12979495	19	G	A	10964632	0,00502
rs4804142	19	G	A	11004230	0,00277

SNP	chr	A1	A2	bp	b
rs12971616	19	C	T	11014937	0,00368
rs17850995	19	T	A	11034039	0,05271
rs12611283	19	T	A	11055299	0,01071
rs12610374	19	C	T	11080921	0,00946
rs73013166	19	T	C	11126160	0,03668
rs881227	19	C	T	11154452	0,01181
rs8103309	19	T	C	11174935	0,02080
rs73013202	19	C	G	11179709	0,00467
rs73015007	19	G	A	11183837	0,03139
rs56289821	19	G	A	11188247	0,01341
rs73015016	19	G	A	11191300	0,01885
rs10402112	19	T	A	11191677	0,00201
rs61194703	19	A	T	11192193	0,00398
rs17248720	19	C	T	11198187	0,02427
rs57217136	19	T	C	11201124	0,03466
rs6511721	19	A	G	11206575	-0,05257
rs17242395	19	G	A	11206969	0,03668
rs118147862	19	G	A	45319631	0,08800
rs41290120	19	G	A	45382675	0,04430
rs429358	19	T	C	45411941	-0,01733
rs7412	19	C	T	45412079	0,07298
rs7256200	19	G	T	45415935	-0,00965
rs12721046	19	G	A	45421254	-0,00563
rs56131196	19	G	A	45422846	-0,00098
rs4420638	19	A	G	45422946	-0,03494
rs111789331	19	T	A	45427125	-0,00145
rs66626994	19	G	A	45428234	-0,00028
rs9979847	21	C	A	35591400	-0,01034
rs28451064	21	G	A	35593827	-0,04103
rs9976596	21	T	C	35596842	-0,00945
rs8131284	21	T	C	35607496	-0,00831
rs146711790	21	A	G	35612761	-0,03334
rs762158	21	C	G	35653670	-0,00719
rs743339	21	T	C	35655217	-0,00871
rs7277800	21	G	A	35655734	-0,00185
rs9982672	21	A	G	35657043	-0,00670
rs79158929	21	G	A	35682958	-0,00832
rs180803	22	T	G	24658858	-0,18092

Table 5: Selected variants from the recessive dataset

SNP	chr	A1	A2	bp	b
rs10908837	1	A	G	154397933	-0,08948
rs4618978	1	C	G	222779187	0,08886
rs1009	2	G	A	85808737	0,09466
rs7041637	9	A	C	21961866	0,01999
rs2518723	9	T	C	21995882	-0,12377
rs3217992	9	T	C	22003223	0,09423
rs3217991	9	I	D	22003298	0,03545
rs523096	9	G	A	22019129	-0,09524
rs10757264	9	A	G	22019732	3,74756
rs490005	9	A	G	22020493	-0,33971
rs567453	9	C	G	22021737	-0,01463
rs581876	9	T	C	22022376	-0,02230
rs504318	9	T	A	22024023	-0,04506
rs496892	9	T	C	22024351	-0,13219
rs10738604	9	A	G	22025493	-0,16317
rs62560774	9	A	C	22028406	-0,02275
rs1412830	9	T	C	22043612	-0,00622
rs1412829	9	G	A	22043926	-0,00148
rs10811647	9	G	C	22065002	-0,84884
rs10811650	9	G	A	22067593	0,62322
rs141014318	9	G	A	22092924	0,02518
rs1537372	9	T	G	22103183	0,05276
rs10811656	9	T	C	22124472	0,06554
rs12259037	10	A	G	30300079	-0,05999
rs943438	10	C	T	30327735	-0,05894
rs4749523	10	G	A	30332445	0,01971
rs61841115	10	T	C	30333025	-0,03485
rs7071112	10	A	G	30334804	-0,03673
rs10899955	10	C	T	44462959	0,19042
rs10899956	10	T	C	44466480	-0,10219
rs12359058	10	G	T	44471343	-0,14430
rs2085797	10	C	T	44471736	0,07768
rs10899957	10	T	C	44472469	0,02939
rs9633745	10	T	C	44487298	-0,04533
rs4948591	10	T	G	44490414	0,09173
rs1873756	10	A	G	44500350	-0,03484
rs10899965	10	A	G	44501946	0,02071
rs11238808	10	T	G	44505391	0,02265
rs11065979	12	T	C	112059557	0,06382
rs11513729	12	T	C	112273499	0,00378

SNP	chr	A1	A2	bp	b
rs17696736	12	G	A	112486818	0,00183
rs11066283	12	G	A	112840766	0,05782
rs11066301	12	G	A	112871372	0,00455
rs12438645	15	G	A	79068203	0,07051
rs12440925	15	A	G	79075754	0,00422
rs2869862	15	G	A	79076885	0,02091
rs7183257	15	A	G	79105381	0,05748
rs28505515	15	G	C	79114477	0,02682
rs62011004	15	T	C	79126536	0,09066
rs9920324	15	A	T	79129142	0,03025
rs8030937	15	T	G	79129587	0,18247
rs4344704	15	A	T	79141703	-0,07197
rs7173743	15	C	T	79141784	-0,30761
rs12976411	19	T	A	32882020	-0,40120
rs762158	21	G	C	35653670	0,15949