UN-EXPLORED REGIONS OF THE HUMAN GENOME AND PREDISPOSITION TO CORONARY ARTERY DISEASE

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Paraskevi Christofidou, BSc MSc

Department of Cardiovascular Sciences

University of Leicester

September 2013

ABSTRACT

Un-explored regions of the human genome and predisposition to coronary artery disease - Paraskevi Christofidou

Coronary artery disease (CAD) is a leading cause of morbidity and mortality worldwide. Previous genome-wide association (GWA) studies have identified several common variants underlying the risk of CAD. However, the collective contribution of these variants to CAD risk is modest and explains only a small proportion (~10%) of its overall heritability. There are largely un-explored regions/variants of human genome that may harbor alleles/genes/loci/pathways associated with susceptibility to CAD and account for a portion of its missing heritability. Runs of homozygosity (ROHs), rare alleles and pseudoautosomal regions (PARs) are amongst most overlooked regions/variants by GWA studies.

Genome-wide homozygosity analysis in CARDIoGRAM Consortium revealed statistically significant differences in the overall homozygosity levels between 10,548 CAD patients and 10,273 CAD-free controls. The distribution of consensus regions of overlapping ROHs showed over-representation amongst patients with CAD, suggesting that accumulation of recessive alleles may increase the risk of CAD.

The aggregate association analysis of low-frequency and rare variants represented on the HumanCVD 50K array in >13,000 CAD patients and >14,000 controls from the IBC 50K CAD Consortium validated previously reported association between *LPA* and CAD in populations of European ancestry. This analysis also revealed new associations between *F10, F7* and *TRAF2* genes and CAD in South Asians.

Common intergenic variant in PAR1 was associated with CAD risk in 9,536 women in the meta-analysis of CARDIoGRAM Consortium. New generation RNA-sequencing analysis provided first glimpse into PARs transcriptome in human monocytes and macrophages and uncovered expression of a lincRNA in close proximity to the PAR1 association signal. Sex-stratified comparative gene expression analysis in human monocytes and macrophages revealed statistically significant differences in PAR1 gene expression levels between men and women.

These data revealed novel associations between CAD and ROHs as well as PAR1 and showed that in-depth exploration of regions commonly neglected by previous GWA studies has a potential to provide new insights into genetic architecture of common complex diseases.

ACKNOWLEDGEMENTS

I would like to give my deepest gratitude to Dr. Maciej Tomaszewski, my main supervisor and mentor, who believed in me and was always there for me. He convincingly conveyed a spirit of adventure, excitement and determination in regard to research and he inspired me tremendously to pursue a career in cardiovascular genetics. I would also like to give a big thank you to my co-supervisor Prof. Nilesh Samani. Both of them had been real role models in my life and it was an honour and exceptional training experience working alongside them. Under their mentorship and the continued guidance, I was able to interact with other senior researchers in the field and participate in conferences and most importantly lead three international multi-collaborative projects.

This thesis would have not been possible without the help of so many people in so many ways. I would like to thank you my PhD committee members, namely Prof. Mark Jobling and Prof John Thompson for their support. In addition many special thanks to all of our collaborators (group leaders and supportive staff members from Germany, Canada and US) that participated in the projects.

I would also like to thank the Alumni association of University of Leicester for funding my PhD.

Many special thanks to Dr. Chris Nelson for all the statistical support during my PhD. His valuable guidance and help significantly contributed to the successful completion of this PhD. I would also like to thank Dr. James Eales for all the bioninformatics help.

A big thank you to my colleagues and fellow PhD students, Dr. Radek Debiec and Lisa Bloomer. A very special thank you to Mr Matthew Denniff for the practical and

especially emotional support. Thanks to my friend and co-traveller for the past 4 years Giovanna Nicolaou. Many thanks to all the members of the department for making a nice working environment.

Last but not least, I am greatly indebted to my parents, my brother and sister and my grandparents and who have been a source of encouragement and inspiration to me throughout my life. Thank you for the myriad of ways in which, throughout this whole journey, you have actively supported me in my determination to find and realise my potential, and just believe in myself.

I would like to dedicate this work to my mentor and my family. I hope I make all of you very proud.

CONTENTS

LIST OF TABLES

Chapter 3: Low-frequency/rare variants and predisposition to coronary artery disease

LIST OF FIGURES

Chapter 3: Low-frequency/rare variants and predisposition to coronary artery disease

Chapter 4: Genetic architecture of pseudoautosomal regions and susceptibility to coronary artery disease

ABBREVIATIONS

apoA1	apolipoprotein A1
apoB	apolipoprotein B
apoB-LPs	apolipoprotein B-containing lipoproteins
apoE	apolipoprotein E
ARIC	Atherosclerosis Risk in Communities study
BHF-FHS	British heart foundation family heart study
B	beta coefficient
BMI	body mass index
CABG	coronary artery bypass grafting
CAD	coronary artery disease
CCGB	Cleveland Clinic Gene Bank study
CD/CV	common disease/common variant
CDKN2A	cyclin dependent kinase inhibitor 2A
CDKN2B	cyclin dependent kinase inhibitor 2
CD/RV	common disease/rare variant
Chr	chromosome
CI	confidence interval
CNV	copy number variant
CVD	cardiovascular disease
DBP	diastolic blood pressure
DUKE	DUKE Cathgen study
DZ	dizygotic
e-QTL	expression quantitative trait loci
F	inbreeding coefficient
FDB	familial defective apoB100

FDR	false discovery rate
FH	hypercholesterolemia
FHS	Framingham Heart Study
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GERMIFS	German myocardial infarction family study
GWA	Genome-wide association
HDL	high density lipoprotein
HLA-DRA	major histocompatibility complex, class II, DR alpha
HWE	Hardy Weinberg equilibriium
ICAM-1	intracelluler adhesion molecule1
K1F1B	kinesin family member 1B
KB	kilobase
IBD	identical by descent
IBS	identical by state
IL2RA	interleukin 2 receptor alpha gene
IL7RA	interleukin 7 receptor alpha gene
LCI	lower confidence interval
LD	linkage disequilibrium
LDL	low density lipoprotein
LDL-R	low density lipoprotein receptor
LDLRAP1	low density lipoprotein receptor adaptor protein 1
LFA-1	lymphocyte function-associated antigen 1
lincRNA	long non-coding RNA
LOD	logarithm of the odds
LOLIPOP	London Life Sciences Prospective Population cohort
Lp(a)	lipoprotein(a)
MAF	minor allele frequency

Mb	Mega bases
M-CSF	macrophage colony stimulating factor
MEF2A	myocyte-specific enhancer factor 2A
MI	myocardial infarction
MIGen	Myocardial Infarction Genetics Consortium
MTNR1B	melatonin receptor 1B
MZ	monozygotic
N	number
NHANES	National Health and Nutrition Examination Survey
OHGS	Ottawa Heart Genomic study
OR	odds ratio
PAR	pseudoautosomal region
PAR1	pseudoautosomal region 1
PAR2	pseudoautosomal region 2
PCSK9	proprotein convertase subtilisin/kexin type 9 gene
PROCAM	Prospective cardiovascular munster study
PROCARDIS	Precocious Coronary Artery Disease study
PSGL-1	P-selectin glycoprotein ligand-1
PTCA	percutaneous transluminal coronary angioplasty
QTL	quantitative trait loci
QQ	quantile-quantile plot
ROH	run of homozygosity
RV	rara variant
SD	standard deviation
SE	standard error
SBP	systolic blood pressure
SNP	single nuvleotide polymorphism

T2D	type 2 diabetes
UCI	upper confidence interval
UK	United Kigndom
US	United States
UTR	untranslated region
VCAM-1	vascular cell adhesion molecule1
VLA-4	very late antigen 4
VLDL	very low density lipoprotein
WHO	World Health Organization
WOSCOPS	West of Scotland
WTCCC	Wellcome Trust Case Control Consortium

PUBLICATIONS

Bloomer LD, Nelson CP, Eales J, Denniff M, **Christofidou P**, Debiec R, Moore J, Consortium C, Zukowska-Szczechowska E, Goodall AH, Thompson J, Samani NJ, Charchar FJ, Tomaszewski M. *Male-specific region of the Y chromosome and cardiovascular risk: phylogenetic analysis and gene expression studies.* Arterioscler Thromb Vasc Biol**. 2013 Jul;33(7):1722-7.**

Charchar FJ, Bloomer LDS, Barnes TA, Cowley MJ, Nelson CP, Wang Y, Denniff M, Debiec R, **Christofidou P**, Nankervis S, Dominiczak AF, Bani-Mustafa A, Balmforth AJ, Hall AS, Erdmann J, Cambien F, Deloukas P, Hengstenberg C, Packard C, Schunkert H, Ouwehand WH, Ford I, Goodall AH, Jobling MA, Samani NJ, Tomaszewski M. *Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome.* **Lancet 2012;379:915-22.**

Tomaszewski M, Charchar F, Nelson CP, Barnes T, Denniff M, Debiec R, **Christofidou P**, Rafelt S, Van der Harst P, Wang WYS, Maric C, Zukowska-Szczechowska E, Samani NJ. *Pathway analysis shows association between FGFBP1 and hypertension*. **Journal of the American Society of Nephrology 2011;22:947-55**.

Tomaszewski M, Debiec R, Braund PS, Nelson CP, Hardwick R, **Christofidou P**, Denniff M, Codd V, Rafelt S, Van der Harst P, Waterworth D, Song K, Vollenweider P, Waeber G, Zukowska-Szczechowska E, Burton PR, Mooser V, Charchar F, Thompson JR, Tobin MD, Samani NJ. *Genetic architecture of ambulatory blood pressure in the general population - insights from cardiovascular gene-centric array*. **Hypertension 2010;56:1069-76.**

PRESENTATIONS AND PUBLISHED ABSTRACTS

Christofidou P, Nelson CP, Nikpay M, Qu L, Li M, Debiec R, Braund PS, Bloomer LDS, Denniff M, Roberts R, Schunkert H, Reilly MP, Erdmann **J,** McPherson R, König IR, Thompson JR, Samani NJ, Tomaszewski M. *Signatures of recessive alleles and susceptibility to coronary artery disease – genome-wide homozygosity meta-analysis.*American Heart Association, Los Angeles (03/11/2012-07/11/2012) **Circulation 2012.**

Christofidou P, Debiec R, Nelson CP, Braund PS, Bloomer LDS, Ball SG, Balmforth AJ, Hall AS, Tomaszewski M, Samani NJ. *Rare alleles in genetic predisposition to coronary artery disease: Insights form the novel analysis of gene-centric array*. British Cardiovascular Society, Manchester (13/06/2011-15/06/2011) **Heart 2011;97:42.**

Christofidou P. *Rare alleles and genetic predisposition to coronary artery disease*. Third Annual Progress Meeting, Cardiogenics Consortium, Varenna, Italy (17/05/2010).

AWARDS

International Society of Hypertension New Investigator of the Month – April 2012

American Heart Association – International mentoring travel award 2012

CHAPTER 1

INTRODUCTION TO CORONARY ARTERY DISEASE

1. INTRODUCTION

Cardiovascular disease (CVD), the collective term for all diseases affecting the circulatory system (heart and blood vessels) (http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097) is the leading cause of morbidity and mortality worldwide (http://www.who.int/en/).

CVD claims more than 2,150 lives each day in the USA, averaging one death every 40 seconds (Go *et al.* 2013). It is the main mortality cause in the UK – in 2010 it accounted for approximately 180,000 deaths (http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097). According to World Health Organization (WHO) 17.3 million people around the globe died of CVD in 2008 (30% of deaths worldwide) (http://www.who.int/mediacentre/factsheets/fs317/en/index.html). Of these, 7.3 million deaths were due to coronary artery disease (CAD), the most common terminal clinical manifestation of CVD.

CAD is one of the most important diseases from the public health point of view incurring enormous costs in financial terms. Of all CAD deaths in 2009, 73% occurred before reaching hospital (Go *et al.* 2013). According to National Center Health Statistics mortality data, 281,000 CAD deaths occur out of the hospital or in hospital emergency departments annually (Go *et al.* 2013). The estimated cost of CAD burden in 2009 was $195.2 billion (Go *et al.* 2013). By 2030, 40.5% of the USA population is expected to suffer from some form of CAD disease, with an estimated direct medical cost of $818 billion and an increase of 61% of indirect costs, due to lost productivity resulting from morbidity and premature mortality (Heidenreich *et al*. 2011).

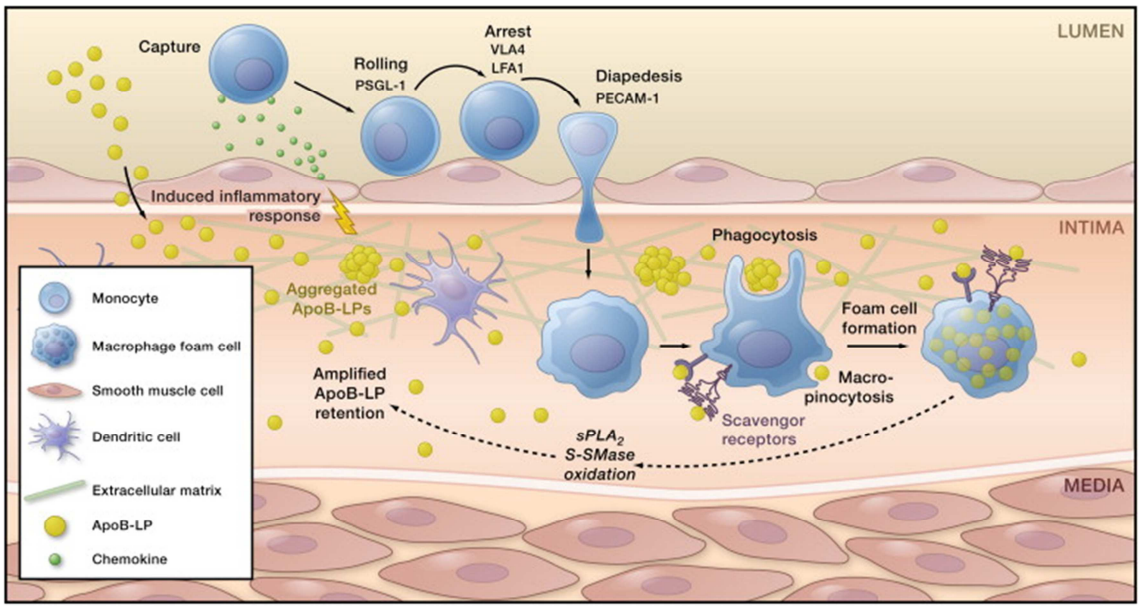
1.1. Biology of CAD - Atherosclerosis

CAD occurs when the walls of the coronary arteries become narrowed by a gradual buildup of atherosclerotic lesions. Atherosclerosis is a silent, chronic, multistep inflammatory process that involves expansion of the arterial intima (normally small area between the endothelium and the underlying smooth muscle cells of the media), with the progressive accumulation of lipids, a variety of cells and extracellular matrix (Moore and Tabas, 2011). This process is asymptomatic for many years. Eventually, some of atherosclerotic lesions may undergo necrotic breakdown and luminal thrombosis, leading to blood vessel occlusion and acute manifestations such as myocardial infarction (Virmani *et al*. 2002).

The key initiating step in atherogenesis is the sub-endothelial accumulation of apolipoprotein B-containing lipoproteins (apoB-LPs) (Williams and Tabas, 1995). ApoB-LPs are composed of a core of neutral lipids, surrounded by a monolayer of phospholipids and proteins (apolipoprotein B). Hepatic apoB-LPs are secreted as very low density lipoproteins (VLDLs) and are converted to low density lipoprotein (LDL) in circulation. Intestinal apoB-LPs are secreted as chylomicrons and are converted by lipolysis into pro-atherogenic particles called remnant lipoproteins (Moore and Tabas, 2011).

The early inflammatory response to apoB-LPs retention triggers activation of overlying endothelial cells, leading to recruitment of circulating monocytes (Figure 1.1) (Glass and Witztum, 2001; Mestas and Ley, 2008). This directional migration process is facilitated by chemokines secreted from activated endothelial cells interacting with related chemokine receptors on monocytes.

Figure 1.1: The process of foam cell formation. Endothelial dysfunction is characterised by increasing stickness of the endothelial cells to circulating monocytes. Recruited monocytes are able to effect passage through the single cell layer of endothelium, between the endothelial cells and into the coronary vessel intima where they differentiate into macrophages. Endothelial dysfunction is also characterised by elevated permeability to lipoproteins. The combination of macrophages and oxidized lipoproteins activate molecular scavenger receptors that recognize and rapidly accumulate oxidized lipoproteins. Macrophages that have taken up great quantities of lipids are called foam cells [Taken from Moore and Tapas, 2011].



Following chemokinesis, monocytes become tethered and roll on endothelial cells overlying retained apoB-LPs through the interaction of monocyte P-selectin glycoprotein ligand-1 (PSGL-1) with endothelial selectins (Mestas and Ley, 2008). Monocytes stick firmly to lesional endothelial cells through interaction of monocyte integrins [VLA-4 (very late antigen4) and LFA-1 (lymphocyte function-associated antigen1)] with endothelial cell ligands [VCAM-1 (vascular cell adhesion molecule1) and ICAM-1 (intracellular adhesion molecule-1)] (Moore and Tabas, 2011). Finally, firm adhesion of monocytes is followed by their entry into the sub-endothelial space (diapedesis) (Kamei and Carman, 2010). There, intralesional monocytes are influenced by macrophage colony stimulating factor (M-CSF) and other factors and differentiate

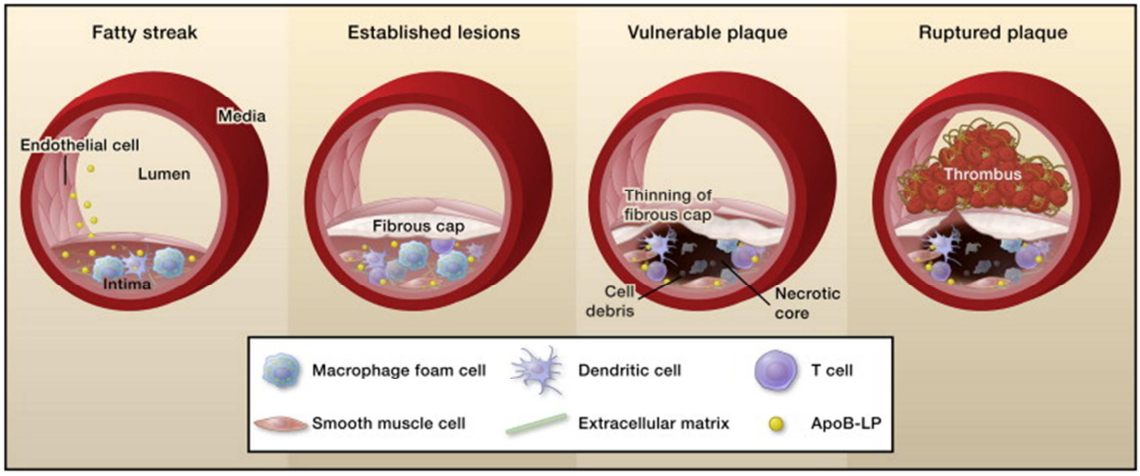
into macrophages or dendritic-like cells (Johnson and Newby, 2009; Paulson *et al.* 2010).

At very initial stages of atherogenesis, phagocytes ingest and process apoB-LPs in macrophages and dendritic-like cells and lead to the formation of foam cells - membrane bound lipid droplets (Moore and Tabas, 2011). These form the bulk of fatty-streaks.

The next stage in the development of atherosclerosis is the formation of intermediate lesions, which are more complex in composition (Figure 1.2). Ongoing inflammation stimulates smooth muscle cells to migrate to the area of injury. Contractile smooth muscle cells undergo phenotypic changes, becoming non-contractile and then fibrous. Neither the fatty-streaks nor the intermediate lesions are immediately harmful. However, this constant and chronic process becomes a maladaptive, nonresolving inflammatory response that expands the sub-endothelial layer and generates atherosclerotic lesions.

The following stage of atherogenesis is the formation of arterial plaques (Figure 1.2). These are well-defined lesions containing a necrotic lipid core, collagen, elastic fibres and proteoglycans and covered with cap of fibrous tissue, composed of smooth muscle cells and connective tissue. As the plaque grows, the fibrous cap gets thinner and the artery lumen narrower. A thinning fibrous cap decreases lesion stability making these plaques susceptible to rupture and the formation of a thrombus. This ultimately results in the manifestation of acute thrombotic vascular disease, including MI.

Figure 1.2: The stages of atherosclerosis – The pathologic process of lesion formation from early fatty streaks, plaque formation, rupture and thrombosis. Early fatty streak, characterized by the accumulation of apoB-LPs in the matrix beneath the endothelial cell layer of blood vessels, stimulates the recruitment of macrophage foam cells and dendritic like cells. Atherosclerotic lesion drives the atheromatous process from initial endothelial injury to final plaque disruption [Taken from: Moore and Tabas, 2011].



1.2. Epidemiology of CAD

Data from the Framingham Heart Study (FHS) have shown that more than 50% of all cardiovascular events in men and women <75 years of age are due to CAD (Go *et al*. 2013). The lifetime risk of developing CAD after 40 years of age is 50% for men and 33% for women (Lloyd-Jones *et al*. 1999) as the incidence rate of developing coronary events in females lags behind men by 10 years and by 20 years for MI and sudden death (Go *et al.* 2013).

Shockingly, about every 34 seconds, an American will experience a coronary event and about every minute a patient will die from one, (Lloyd-Jones *et al*. 2009, Go *et al.* 2013). In 2009, 1 in every 6 deaths was caused by CAD with a mortality rate rising up to 386,324 (Table 1.1) (Go *et al*. 2013). From 1999 to 2009, the annual death number due to CAD declined 27.1%; however, CAD remains the major killer of American males and females (http://www.cdc.gov/nchs/fastats/heart.htm; Go *et al*. 2013).

The latest update from the American Heart Association Statistics Committee, reports that in the United States, the total CAD prevalence is 15,400,000 (6.4%) and the overall prevalence for MI is 7,600,000 (2.9%) in adults ≥20 years of age (Go *et al*. 2013) (Table 1.1). Projections show that by 2030 an additional ~8,675,000 people could have CAD, an 18% increase in prevalence from 2013 (Heidenreich *et al*. 2011; Go et al. 2013). A total number of around 635,000 Americans were estimated having a new coronary event and about 280,000 having a recurrent attack this year (Go *et al*. 2013).

Table 1.1: Epidemiology of CAD and MI in United States among individuals ≥20 years of age - data from National Health and Nutrition Examination Survey (NHANEs) 2007-2010.

Population Group	Prevalence of CAD (n)	Prevalence of MI (n)	Mortality of CAD	Mortality of MI
Both sexes	15 400 000 (6.4%)	7 600 000 (2.9%)	386 324	125 464
Males	8 800 000 (7.9%)	5 000 000 (4.2%)	210 069 (54.4%)	68 814 (54.8%)
Females	6 600 000 (5.1%)	2 600 000 (1.7%)	176 255 (45.6%)	56 650 (45.2%)

Mortality estimates are based on data from Centres for Disease and Prevention/National Center for Health Statistics 2009, Data adopted from Heart disease and Stroke statistics - 2013 update.

An MI is diagnosed every 44 seconds in the US (Go *et al*. 2013). It has been estimated that around 126,000 deaths a year in the US are caused by MI (Table 1.1). The estimated annual incidence of MI is 715,000 (525,000 new and 190,000 recurrent attacks) of which 150,000 (~21%) occur silently (Boland *et al*. 2002; Go *et al.* 2013). The average age at first MI is 64.7 years for men and 72.2 years for women (Lloyd-Jones *et al*. 2009). The estimated average number of years of life lost because of an MI is 16.6 (Go *et al*. 2013). Depending on their sex and clinical outcome, individuals surviving the acute stage of an MI have a chance of illness and death 1.5 to 15 times greater than that of the general population (Go *et al.* 2013).

In the UK, about half (45%) of all deaths from CVD are from CAD (http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097). In 2010, around 1 in 5 male deaths and 1 in 10 female deaths were from CAD, a total of around 80,000 deaths (http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097). Data from the 2006 National Health Survey for England estimate the prevalence of CAD in England as 6.5% in men and 4.0% in women (Smolina *et al*. 2012) (Table 1.2).

Table 1.2: Epidemiology of CAD and MI in United Kingdom among individuals of all ages

Population Group	Prevalence of CAD (n)	Prevalence of MI (n)	CAD mortality	MI Mortality
Both sexes	N/A	N/A	80 568	N/A
Males	6.5%	4.1%	46 591	N/A
Females	4.0%	1.7%	33 977	N/A

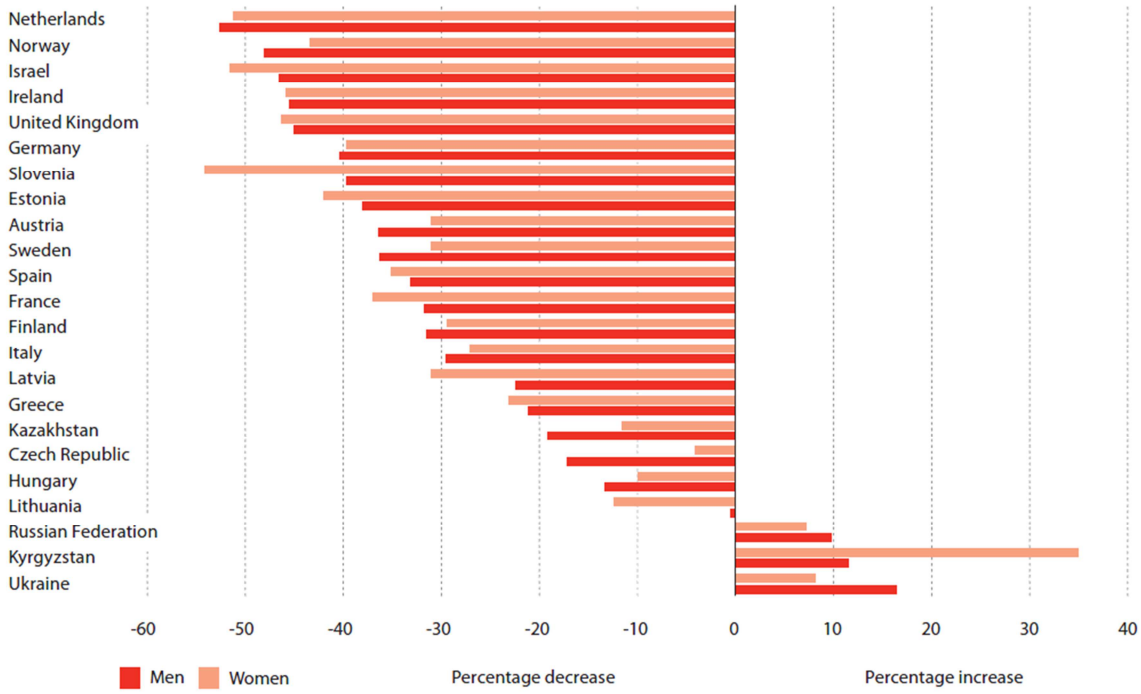
N/A – no estimates are available. Prevalence estimates are based on the Health Survey for England 2006: Cardiovascular disease and risk factors, CAD mortality estimates are based on England and Wales, Office for National Statistics (2010).

Overall, there are just over 1.1 million men and around 850,000 women (≥35 years) with a history of angina, and around 970,000 men and 439,000 women with a history of MI in the UK (http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097).

CAD is the leading death cause in many other countries as well. There are significant differences in CAD disease rates between countries. The highest death rates from CAD are found in Eastern and Northern Europe, the former Soviet Union, USA, Australia and New Zealand. Low CAD mortality rates are observed in Japan and the Mediterranean countries of Europe (Menotti *et al.* 1993). Although there is a paucity of data from most developing countries, CAD accounted for at least one third of CVD deaths in India and more than 50% of deaths in urban areas of China (Okrainec *et al.* 2004; Reddy and Yusuf, 1998).

CAD mortality has declined across most of Europe, with the exception of some Eastern European countries (Figure 1.3) (Grimes, 2012). Overall, Russia and Ukraine experienced an increase in CAD mortality. The CAD death rate in UK has been falling at one of the fastest rates in Europe and decreased by 45% between 1998 and 2008. However, despite this decline, rates are still relatively high compared to other European countries.

Figure 1.3: Percentage change in CAD mortality rates, by sex, in selected countries between 1998 and 2008.



Data are shown as rates per 100,000 individuals. [Taken from Coronary heart disease statistics – A compendium of health statistics 2012 edition - BHF website]

1.3. Risk determination – overview of risk factors

Early identification, monitoring and treatment of high risk individuals are the most critical steps towards reducing CAD burden.

Less than 10% of the population is at high risk for CVD, but the low to intermediate risk group is so large that the majority of cardiovascular events will actually arise from these groups of individuals (Hingorani and Psaty, 2009; Lloyds-Jones *et al*. 2009). As a result, 90% or more of cardiovascular events will occur in individuals with one or more risk factors, just under half of the population (Berger *et al*. 2010). To reduce the greater number of events, lowering risk in the entire population is the most promising strategy. A pharmacological reduction in risk factors within the population accounted for 47-65% of the lowering in death rates during the past 40 years (Bjorck *et al*. 2009; Capewell *et al.* 2009). Unfortunately, the current obesity and diabetes epidemics have reversed these trends (Kones, 2011).

Cardiovascular risk is the product of the effect of several risk factors and as a result a single all-embracing solution is highly improbable. Individual risk factors can cluster and tend to have a multiplicative effect on an individual's total cardiovascular risk (Anderson *et al*. 1991). The goal of primary prevention is to stop the development of disease in an asymptomatic individual via the early identification and treatment of risk factors.

There are numerous risk factors associated with the development of CAD; each of which makes from significant through moderate to small contribution to the ultimate risk of the disease phenotype. Some of these factors such as sex and familiar history are not amenable to change. Others, such as smoking and diet are examples of environmental exposures. Some, such as blood pressure, body weight, diabetes mellitus

and serum cholesterol (Ding and Kullo, 2009), are modifiable and are seen as intermediate processes in the development of CAD originated from a complicated interplay between genetic factors and environmental factors.

A large study of men and women in three prospective cohorts emphasized that ~90% of CAD patients have one of the main risk factors including high blood pressure, high blood cholesterol levels, cigarette smoking and diabetes mellitus (Greenland *et al*. 2003). Similarly, data from the National Health and Nutrition Examination Survey (NHANEs) noted that 45%, or 81 million American adults, have at least one of three conventional risk factors: hypertension, dyslipidemia or diabetes mellitus (Fryar *et al*. 2010).

The INTERHEART study assessed the importance of risk factors for CAD worldwide (52 countries) (Yusuf *et al*. 2004). It showed that optimization of nine main modifiable risk factors (cigarette use, blood lipid disturbances, hypertension, diabetes, obesity, a lack of exercise, low daily fruit and vegetable consumption, alcohol consumption and psychosocial index) could result in a 90% reduction in the risk of an initial MI. The effect of these risk factors follows the same pattern in men and women from different geographic backgrounds and ethnicities.

The prevalence of traditional risk factors is almost as high in those without the disease as in affected individuals (Freedman *et al*. 2002; Khot *et al*. 2003). As a consequence, the predictive models for risk assessment have a lower than desired accuracy in predicting CAD as in any individual patient (Conroy, 2003; Grundy *et al*. 1999). The pathophysiology of CAD stems from development of atherosclerotic lesion. The proliferation of research related to vascular biology in recent years has led to the discovery of a plethora of circulating biomarkers (molecular, heamostatic and

inflammatory) that are implicated in the pathology of atherosclerosis, the common phenotype for all these risk factors (Hackam and Anand, 2003).

1.3.1. Sex

Incidence and prevalence of CAD shows a strong sexual dimorphism. It is well documented that CAD is a male predominated disease with a 2:1 ratio of men to women (Barrett-Connor *et al*. 1997). Indeed, men develop CAD approximately 9 years earlier than women (Yusuf *et al*. 2004). Rates of death attributed to CAD in men are consistently 3 to 4 times higher than those in women across countries with differing background levels of disease. The “male disadvantage” is most obvious when compared pre-menopausal women and age-matched men. This gender gap progressively decreases with aging, in particular after menopause.

1.3.2. Age

Cardiovascular risk increases with age. Age is the strongest indicator for CAD incidence and mortality. Compared to men aged 40 years, 50-year, 60-year and 70-year old men have 5- 15- and more than 40-fold increase in the risk of death from CAD. A similar steep gradient with age is seen for women.

1.3.3. Hypertension

Hypertension is usually defined as systolic blood pressure (SBP) ≥140 mm Hg and/or diastolic blood pressure (DBP) ≥90 mm Hg on repeated office measurements, or use of antihypertensive medications (Chobanian *et al*. 2003). Hypertension is the most common cardiovascular diagnosis in the USA with an estimated 78 million Americans being hypertensive (Go *et al*. 2013). In the 2010 Health Survey for England, 31% of men and 29% of women had hypertension or were taking antihypertensive treatment

(http://www.bhf.org.uk/publications/view-publication.aspx?ps=1002097). An estimated 2 million individuals are diagnosed with hypertension every year (Fields *et al*. 2004) and overall 1 billion people are hypertensive worldwide.

Prevalence of hypertension increases with age, from 7% in individuals aged 18 to 39, to 65% in individuals over age of 59 years (Hajjar and Kotchen, 2003). In the Framingham Heart Study, more than 50% of individuals aged 55 years developed hypertension within the following 10 years (Vasan *et al*. 2002).

Large scale epidemiological studies have established a continuous, consistent, linear relationship between blood pressure and CVD that is independent of other cardiac risk factors (Vasan *et al*. 2002). Hypertension is a major cardiovascular risk factor that directly contributes to CAD (Levy *et al*. 1996). The World Health Report 2002 estimated that over 50% of CAD in developed countries is due to SBP levels > than 115 mm Hg (http://www.who.int/en/). In the INTERHEART study, hypertension accounted for 18% of the population attributable risk of first MI (Yusuf *et al*. 2004). Meta-analysis of prospective data from > 1 million individuals revealed that an increase of 20 mm Hg in SBP or 10 mm Hg in DBP doubled the CAD risk for adults aged 40 to 69 years (Lewington *et al*. 2002).

Randomised clinical trials demonstrate that a blood pressure decrease is beneficial in reducing CAD morbidity and mortality (Staessen *et al*. 2001). Antihypertensive treatment resulted in reduction of MI prevalence from 20% to 25% (Neal *et al*. 2000). A meta-analysis of blood lowering treatment trials of 47,000 subjects with mild to moderate hypertension showed that DBP decrease of 5 to 6 mm Hg reduced MI prevalence by 14% and total CVD by 42% (Collins *et al*. 1990).

1.3.4. Hypercholesterolaemia

Hypercholesterolaemia is considered as a primary atherogenic factor and has a prevalence of 13.8% in the USA (Go *et al*. 2013). There is a continuous, graded and strong relationship between serum cholesterol and CAD death rate (Stamler *et al*. 1986). More than 60% of CAD prevalence in developed countries is attributable to increased levels of cholesterol (>3.8mmol/L) (http://www.who.int/en/). The INTERHEART study demonstrated that 45% of MI in Western Europe and 35% of MI in Central and Eastern Europe are due to abnormal blood lipid profiles, and that individuals with abnormal lipid levels have >3-fold increase in risk of a coronary event compared to individuals with normal lipid levels (Yusuf *et al*. 2004). In the ARIC study, LDL-cholesterol levels greater than 118 mg/dL were associated with an age-adjusted increase in CAD of 42% and 37% in men and women, respectively (Sharrett *et al*. 2001).

The West of Scotland Coronary Prevention Study (WOSCOPS), a large-scale prevention study showed that lipid-lowering treatment reduces CAD events in primary prevention in men with hypercholesteroleamia (Shepherd *et al*. 1995). On average, 1% reduction in LDL cholesterol is associated with 1% reduction in CAD mortality (Grundy *et al*. 2004).

1.3.5. Overweight and obesity

Overweight is defined as BMI of 25 to 29.9 kg/m^2 and obesity as BMI≥ 30 kg/m^2. Obesity constitutes a major public health challenge for developed and developing countries (Finucane *et al*. 2011). The change towards a more affluent Western lifestyle that has taken place during the last 50 years has started a worldwide epidemic increase in the prevalence of obesity. In 2010, 154.7 million (68.2%) of US adults (≥20 years of age) were overweight or obese (Go *et al*. 2013). 12.7 million children aged 2 to 19 years

are obese (Go *et al*. 2013). Over the past 3 decades, the prevalence of obesity in children aged 6 to 11 years has increased from ~4% to >20% (Go *et al*. 2013). Individuals with a BMI>30 kg/m^2 have a 40-fold increased risk of developing diabetes and a 2 to 3 fold increased risk of CAD (Hamm *et al*. 1989) compared to individuals with a normal BMI ≤25 kg/m^2 (Abbott *et al*. 1994). The World Health Report 2002 estimated that > a third of CAD in developed countries is due to overweight (http://www.who.int/en/). The adverse consequences of obesity on CAD risk are possibly mediated through other cardiovascular risk factors such as higher blood pressure, abnormal lipid profile and diabetes (Pearson *et al*. 2002). Reducing weight is accompanied by reduction in the risk of CAD.

1.3.6. Diabetes mellitus

Diabetes mellitus is a common complex disorder with dramatically increasing prevalence worldwide. Over the past two decades, there has been an explosive increase in the number of patients diagnosed with diabetes worldwide. The latest alarming data from the United States, estimate that 19.7 million (8.3%) of the adult population were diabetic, additional 8.2 million (3.5%) had undiagnosed diabetes and 87.3 million (38.2%) were pre-diabetic (Go *et al*. 2013). These numbers are expected to increase epidemically as a consequence of population aging and changes in lifestyle; primarily - obesity. The World Health Organization estimated that the number of diabetic patients will reach 300 million in 2025 (Zimmet *et al.* 2001).

Diabetes is a powerful and independent risk factor for CAD (Beckman *et al*. 2002). Diabetic patients have from 2 to 4 fold increase in CAD risk, independent of other cardiovascular risk factors (Greenland *et al*. 2003; Almdal *et al*. 2004). Some experts even consider diabetes as equivalent to CAD diagnosis (Fadini *et al*. 2009).

1.3.7. Smoking

Tobacco smoke is a major threat to public health; by 2020 smoking is expected to kill 10 million people per year (Go *et al*. 2013). Among Americans aged > 18 years, 34.8 million (21.3% of men and 16.7% of women) were active cigarette smokers (Go *et al*. 2013). The ARIC study demonstrated that smoking is associated with approximately 50% increase in the progression of atherosclerosis based on measurements of carotid intima-media thickness (Howard *et al*. 1998). The relative risk of CAD in these who smoke one pack of cigarettes per day is six-fold higher for men and three-fold higher for women compared with non-smokers (Prescott *et al*. 1998). The INTERHEART study reported that smoking accounted for 36% of population attributable risk for first MI (Yusuf *et al*. 2004). Passive cigarette exposure also increases the CAD risk by about 30% (Barnoya and Glantz, 2005). There is a dose-response relationship between passive cigarette exposure and CAD (Kawachi *et al*. 1997; Coggins, 1998). Quitting smoking is associated with 36% reduced risk of total mortality (Critchley and Capewell, 2003).

1.3.8. Physical inactivity

Physical inactivity is a risk factor for the development of CAD and is associated with a higher all-cause mortality rate (Blair *et al*. 1989). The proportion of youth who report engaging in no regular physical activity is high, and this proportion increases with age (Go *et al*. 2013). Conversely, physical activity is associated with cardiovascular benefits - it increases HDL-cholesterol, lowers LDL-cholesterol, triglycerides, low grade inflammation and blood pressure, improves fasting and postprandial glucose-insulin homeostasis, endothelial function, triggers and maintains weight loss and facilitates smoking cessation (Thompson *et al*. 2003; Bassuk and Manson, 2005; Taylor *et al*. 2007). In patients with established cardiovascular disease, physical activity reduces

angina symptoms, benefits heart failure and lowers mortality after MI (Thompson *et al*. 2003).

1.3.9. Poor diet

Dietary habits are recognized associates of cardiovascular risk. Prospective studies indicate consistent and substantial reduction in cardiovascular risk related to lower unsaturated fat consumption (Mozaffarian *et al*. 2006), consumption of whole grains, legumes and cereal fiber and consumption of fruits and vegetables (Dauchet *et al*. 2006). In one of secondary prevention trials, advice to consume a Mediterranean type diet reduced risk of MI or cardiac death by up to 72% over a 4-year follow-up (De Lorgeril *et al*. 1999).

1.3.10. Fibrinogen

Fibrinogen has been identified as a major independent risk factor for CVD (Kannel *et al.* 1987). This glycoprotein regulates cell adhesion, chemotaxis and proliferation, influences platelet aggregation and blood viscosity, interacts with plasminogen binding and in combination with thrombin, mediates the final step of coagulation and the response to vascular injury (Rabbani and Loscalzo, 1994; Smith *et al*. 1990). There is also evidence suggesting an association among fibrinogen and CAD (Andreotti *et al*. 1999). The ARIC study showed that in over 14,000 middle-aged adults, elevated levels of fibrinogen were associated with 1.5 fold increased risk of developing MI or coronary death over 5 years of follow-up (Folsom *et al*. 1997).

1.3.11. Lipoprotein(a)

Lipoprotein(a) is a recognised proatherogenic factor and contributor to CAD. Asymptomatic individuals with Lp(a) levels in the top tertile of the distribution had 1.7-

fold higher risk of CAD compared to those with Lp(a) in the lower tertile (Danesh *et al*. 2000).

1.3.12. Homocysteine

Homocysteine is a sulfur-containing amino-acid formed as a by-product of the metabolism of the essential amino acid methionine (Mangoni and Jackson, 2002). Epidemiological studies suggested that elevated homocysteine levels were associated with a moderately increased risk of CAD (Boushey *et al.* 1995).

1.3.13. Apolipoproteins

ApoB and ApoA1 are recognized surrogates of non-HDL cholesterol (LDL-C, VLDL-C, IDL-C) and HDL-cholesterol, respectively. Data from INTERHEART project suggest that the ratio of APO-B/APOA-I is in fact the strongest determinant of MI risk in population (55% of MI risk is explained by this ratio), followed by blood pressure and smoking (Yusuf *et al*. 2004).

1.4. Familial history of CAD

Familial history has been characterised as a "free well-proven personalized genomic tool that captures many of the genes and environmental interactions and can serve as the cornerstone for individualized disease prevention" (Guttmacher *et al*. 2004). It has been used as a surrogate of genetic factors both in clinic and research projects.

Familial clustering of CAD has been noted for more than 100 years. For example, in 1910 a family in which three generations were affected with angina was described (Evans *et al*. 2003). In 1966 the incidence risk of first-degree relatives of 121 male and 96 female cases with premature CAD was estimated. Male relatives of male cases had a 5-fold increased incidence, whereas male relatives of female cases had a 7-fold increased incidence of risk (Slack and Evans, 1966). Extensive studies in Finland in the 1970s revealed 3.5 fold-increased risk of CAD in brothers of male CAD cases and a 2-fold increased risk in sisters (Rissanen and Nikkila, 1977; Rissanen, 1979). A study in the early 1980s assessing 19 risk factors associated with premature CAD had shown that the most significant risk contributor was familial history (Nora *et al*. 1980). Since then, consistent evidence from various epidemiological studies indicates that familial history is an important independent cardiovascular risk factor (Kullo and Ding, 2007; Scheuner, 2003). A positive familial history of premature CAD is usually defined as any male first degree relative with proven CAD younger than 55 years or female younger than 65 years (British Cardiac Society *et al*. 2005).

Adoptee studies have also suggested that disease risk can not only explained by shared environmental influences. The landmark Danish study of 960 families with adopted children revealed that the death of a biologic parent before the age of 50 years from a cardiovascular cause was associated with a 4.5-fold increase in mortality for the

offspring, whereas the death of an adoptive parent did not significantly increase the risk (Sorensen *et al*. 1988).

In the Framingham Offspring study, a familial history of CAD was associated with a 2.4 and 2.2 fold increase in risk of CAD in men and women respectively (Genest *et al*. 1992). In the InterHeart study, a familial history of CAD was associated with 1.5 and 1.45 fold increase in risk of CAD after correction of other risk factors in men and women, respectively (Yusuf *et al.* 2004). A family history of MI in the Prospective Cardiovascular Munster (PROCAM) study indicated it was an independent risk factor for CAD (Cooper *et al*. 2005). Relatives and descendants of a patient with premature MI carry a 50-80% increase in relative risk to develop a heart attack as well (Myers *et al.* 1990). Several other epidemiological studies have consistently shown a 2 to 3-fold increase in risk for CAD in first degree relatives when compared to the general population (Arnett *et al*. 2007; Kullo and Ding, 2007; Mayer *et al*. 2007).

Such predictive power of a positive familial history is the hallmark of a genetic component in the etiology of CAD.

1.5. Heritability estimates of CAD

Heritability of a trait refers to the proportion of observed phenotypic differences within a population that is due to genetic differences among the individuals in that population (Vissher *et al.* 2008). More precisely narrow-sense heritability refers to differences among the additive genetic values and the broad-sense heritability refers to genetic differences as differences between genotypic values (Tenesa and Haley, 2013). The narrow-sense heritability reflects the degree to which the genes transmitted from the

parents determine the phenotype of their children and is most useful in predicting disease risk from parental family history (Tenesa and Haley, 2013).

Twin studies represent a useful recourse for estimating the extent to which genetic and environmental variations determine the phenotypic variation of a trait (Boomsma *et al*. 2002). Twins share intrauterine environment and age: monozygotic (MZ) twins share 100% of their genes and dizygotic (DZ) twins share on average, 50% of their segregating genes.

Seminal data from more than 21,000 Swedish twins have revealed that among males, the relative hazard of death from CAD was 8.1 for MZ and 3.8 for DZ when one's twin died of CAD before the age of 55 years. Among females the relative hazard of death was calculated at 15.0 for MZ and 2.6 for DZ twins when one's suffered a fatal coronary event before the age of 65 years (Marenberg *et al*. 1994). These findings clearly highlighted the presence of a genetic component in CAD.

The authors validated their findings in a 36 years follow-up of the same cohort, showing that CAD mortality is also heritable, with values ranging from 38% in females to 57% in males (Zdravkovic *et al*. 2002). A subsequent study of 15,910 Danish twins estimated CAD heritability at 53% both in males and females (Wienkr *et al*. 2001).

Twin studies have yielded evidence for the heritability of many CAD risk factors (Table 1.3). The estimated heritability of these cardiovascular risk factors varies from ~45% to ~90% depending on the phenotype and study characteristics. For example, plasma lipid concentrations appear to have a particularly strong genetic component (Kathiresan *et al*. 2007). Type 2 diabetes (Barroso, 2005) and blood pressure (Havlik *et al*. 1979; Levy *et al.* 2000) also have substantial heritable component.

Table 1.3: Narrow-sense heritability estimates of CAD risk factors

Risk factor	Heritability Estimates	Reference
HDL-cholesterol	69 %	Snieder *et al*. 1997
LDL-cholesterol	69-77 %	Snieder *et al*. 1997
Triglycerides	51-69 %	Snieder *et al*. 1997
Total cholesterol	66-77 %	Snieder *et al*. 1997
Apolipoprotein AI	46/55 %	Snieder *et al*. 1997
Apolipoprotein B	65/59 %	Snieder *et al*. 1997
Lipoprotein(a)	89/89 %	Snieder *et al*. 1997
SBP	53 %	Evans *et al*. 2003
DBP	48 %	Evans *et al*. 2003
Body mass index	83/74 %	Lajunen *et al*. 2009
Smoking (cigarettes/day)	86%	Koopmans *et al*. 1999
Homocysteine	57%	Siva *et al.* 2007
Diabetes	75%	Kaprio *et al*. 1992

Heritability estimates differ among populations due to differences in both genetic and environmental factors.

1.6. Examples of Mendelian disorders associated with CAD

Several examples of monogenic forms of cardiovascular and metabolic disorders associated with increased susceptibility to premature CAD are listed in Table 1.4. These diseases are caused by single rare, highly penetrant mutant alleles with large effect on the phenotype. A majority of these mutations map to genes within pathways of lipid metabolism. Genetically, these are usually non-synonymous, non-sense, frameshift or splice variant mutations that lead to defects of structure and or function of the encoded protein.

The classical example of monogenic form leading to premature CAD is familial hypercholesterolemia (FH), an autosomal dominant disorder caused by mutations in the LDL receptor gene (*LDLR*) (Brown and Goldstein, 1986), proprotein convertase

subtilisin/kexin type 9 gene (*PCSK9*) (Abifadel *et al*. 2003) or apolipoprotein B gene (*ApoB*) (Soria *et al*. 1989). FH manifests as very high plasma concentrations of LDL-C, clustering with familial history of CAD, premature onset of CAD and commonly - cutaneous stigmata of skin lipid deposition (xanthomata and xanthelasmata palpebrarum) (Marks *et al*. 2003).

The most common type of FH is caused by mutations in *LDLR*. Its heterozygotic type affects roughly 1 in 500 individuals globally (Roy *et al*. 2009). More than 1,000 *LDLR* mutations from FH patients have been identified so far and these mutations have been identified in all exons of *LDLR* and lead to defects in functional domains of the encoded protein (Leigh *et al*. 2008). The molecular nature of the genetic defect implicates severity of cardiovascular involvement in FH. Receptor-negative mutations (inability to produce mature LDL-receptors) are associated with an earlier onset and more severe disease phenotype than when mature but abnormal receptors are produced (receptor-defective mutations) (Naoumova *et al*. 2004). FH heterozygotes, with one mutant copy of the *LDLR* gene, typically have plasma cholesterol levels ranging from 300 to 550 mg/dL, whereas FH homozygotes, with both copies mutant, typically have cholesterol levels ranging from 550 to greater than 1000 mg/dL (Jansen *et al*. 2004).

ApoB constitutes a key glycoprotein playing a role in the lipoprotein metabolism. Mutations in *ApoB* gene account for 5% of cases with autosomal dominant monogenic hypercholesterolaemia (Burnett and Hooper, 2008). Several *ApoB* mutations causing familial defective apoB100 (FDB) have been identified (Roy *et al*. 2009). The most common mutation in ApoB substitutes glutamine for argirine at amino acid 3500 (Arg3500Gln), impairing the ability of LDL-C particles to bind to the LDL receptor (Tarugi *et al*. 2007). Normally, the LDLR-binding region of apoB (site B) is available to

interact with LDLR; the interaction between arginine R3500 and tryptophan W3469 being particularly important. In FDB, mutations alter the confirmation on the C-terminal region of apoB, leading to occlusion of site B (Tarugi *et al*. 2007). *ApoB* mutations lead to defect in binding of LDL-C particles to an LDL receptor resulting in impaired plasma clearance and subsequent elevation of circulating concentrations of LDL-C (Burnett and Hooper, 2008).

Mutations in *PCSK9* account for 2% of cases with autosomal dominant monogenic hypercholesteroleamia (Burnett and Hooper, 2008; Soutar and Naoumovaa, 2007). *PCSK9* encodes a serine protease that destroys LDLR receptors in liver and thereby controls the level of LDL in plasma (Abifadel *et al*. 2003). "Gain of function" mutations increase the intra-cellular degradation of the LDLR leading to a reduced number of LDLRs on the surface of hepatocytes (Tarugi *et al*. 2007). This, results in decreased LDL-C internalization and subsequently, increased LDL-C plasma levels.

Autosomal recessive hypercholesterolaemia is caused by homozygous mutations of low density lipoprotein receptor adaptor protein 1 gene (*LDLRAP1*) (Table 1.4) (Garcia *et al*. 2001). Sitosteroleamia is another example of very rare disorder in which plant sterols and cholesterol levels are dramatically increased secondary to mutations in either of the sterol transporters encoded by ATP-Binding Cassette Sub-Family G Member 5 (*ABCG5*) or ATP-Binding Cassette Sub-Family G Member 8 (*ABCG8*) genes (Berge *et al.* 2001). Both lead to premature CAD.

Although rare deleterious mutations such as those discussed above increase enormously the individual risk of developing CAD in mutation carriers, their population effect is low from an epidemiological perspective, (Cambien and Tiret, 2007). However, the unraveling of the genetic component of these single-gene diseases was critical to

understanding of the mechanisms underlying hypercholesterolaemia and its role in CAD (Antonarakis and Beckmann, 2006).

Table 1.4: Examples of monogenic disorders associated with premature CAD

Lipid Disease	Gene	Mode of inheritance	Heterozygous mutant frequency	Homozygous mutant frequency	Molecular etiology
Familial Hypercholesterolaemia type 3	*PCSK9* 1p32	Autosomal dominant	<1:2500	-	Increased LDL-C due to heightened degradation of LDLR in the liver
Familial ligand defective APOB-100	*APOB* 2p23-24	Autosomal dominant	1:1000	$1x4x10^{-6}$	Increased LDL-C due to decreased affinity of APOB to LDLR
Autosomal recessive hypercholesterolaemia	*LDLRAP1* 1p36	Autosomal recessive	-	$<1x5x10^{-6}$	Increased LDL levels due to a defect of intracellular processing of LDL-R
Sitosterolaemia	*ABCG5* or *ABCG8*	Autosomal recessive	-	ABCG5 – very rare ABCG8 - 1:50 000	Increase in phytosterols (sitosterol, campesterol, stigmasterol, avenosterol)

LDLR – low density lipoprotein receptor, **PCSK9** – proprotein convertase subtilisin/kexin type 9 gene, **APOB** – apolipoprotein B100, **LDLRAP1** – low density lipoprotein receptor adaptor protein 1, **ABCG5** - ATP-Binding Cassette Sub-Family G Member 5, **ABCG8** - ATP-Binding Cassette Sub-Family G Member 8

Common complex human diseases such as CAD tend to cluster in families, but they do not exhibit the characteristic Mendelian segregation of monogenic disorders. Unlike familial hypercholesterolaemia and other monogenic disorders, CAD is a product of interaction between many single alleles mapping to different regulatory pathways together interacting with each other as well as environmental factors (such as smoking, diet, physical activity, etc). As a result, many questions about their genetic architecture have been raised including the number of genetic variants acting synergistically, their location in the human genome, their nature, function, characteristics (frequencies, effect sizes) and the model of interactions between them and environmental factors.

1.7. Genetic approaches to studying CAD

Several main genetic strategies have been developed to address some of the questions and investigate the genetic architecture of complex diseases over the last decade.

1.7.1. Candidate gene studies

Initial efforts to elucidate the genetic component of CAD were based on selection of gene(s) that mapped to pathways known to play a role in the disease process. To this end, many family and population-based studies examined candidate genes encoding proteins known to participate in the pathogenesis of atherosclerosis (such as lipoprotein metabolism) to find variants underlying the increased risk of CAD (Kullo and Cooper, 2010; Mayer *et al*. 2007).

More than 5,000 candidate gene-based studies on CAD have been published so far. Only few of them have successfully identified and replicated associations between a candidate gene and CAD (Schunkert *et al*. 2010). These studies implicated *PCSK9* and

Apo E (Kathiresan *et al*. 2008a), *LDLR* (Linsel-Nitschke *et al*. 2008), *Apo B* (Willer *et al.* 2008) and *LPA* (Clarke *et al*. 2009) genes.

However, a majority of candidate studies failed to identify/replicate the association between a candidate gene(s) and CAD. This was related to inherent limitations of the selected strategy. Many of these studies failed primarily due to insufficient power (Charchar *et al*. 2008; Hardy and Singleton, 2009).

One explanation for the lack of success of this strategy to dissect the genetic background of CAD may be related to an incorrect a priori assumption that predisposition to CAD is driven by genes that have an obvious physiological relationship to the common CAD (Charchar *et al*. 2008). This unproductive effort highlighted our insufficient understanding of biology of CAD.

In addition, candidate gene studies usually focus on the exons, introns and immediate flanking regions of the genes of interest. However, these loci represent only a small fraction of the genome, whereas the intergenic regions contain many DNA elements that regulate gene expression.

Also, most of candidate gene studies were undertaken before the era of HapMap and tagging approaches and as a result reflected insufficient coverage for genetic variants in a gene.

1.7.2. Linkage studies

An intensive exploration of the molecular mechanisms behind the inherited predisposition of CAD was also carried out with genome-wide linkage scans, defined as searches for chromosomal regions linked to a phenotype of interest (QTL - Quantitative trait loci). Linkage studies require enrollment of families with individuals affected by

the disease or phenotype of interest over generations (Cambien and Tiret, 2007; Laird and Lange, 2006; Seo and Goldschmidt-Clermont, 2008) and are based on investigation of the co-segregation of genetic markers, panels of microsatellites or large set of SNPs regularly spaced throughout the genome. This strategy examines if specific alleles are co-transmitted with the disease at a higher frequency than expected by chance (Cambien and Tiret, 2007). The evidence of linkage is usually expressed as a "log of the odds" or "LOD" score. This is the $_{10}$logarithm of the odds ratio for the probability of the results of linkage divided by the probability of the data in the absence of linkage. An odds ratio of 1,000 equivalent to a LOD score of 3 is considered strong evidence for linkage a chromosomal region to a trait of interest.

Several genome-wide linkage analyses reported a number of signals suggestive of linkage to CAD/MI in the human genome (Table 1.5).

The largest study, the British Heart Foundation Family Heart Study (BHF-FHS, included 4,175 CAD subjects from 1,933 families recruited throughout the UK (Samani *et al*. 2005). In the genome-wide analysis, not a single linkage peak exceeded LOD score>3 for any of the cardiovascular end points examined. For CAD, the highest LOD score was recorded at 2.70 on chromosome 2. The genome-wide linkage analysis conducted in BHF-FHS illustrates in general a majority of outcomes of this strategy in search of QTLs for CAD. Indeed, few strong signals of linkage to CAD were identified by microsatellite-based genome-wide analysis and very few of them exceeded the conservative threshold of significance suggested (LOD>4.1). In addition, consistency in location of identified QTLs was very poor across different studies (Table 1.5). The identified QTLs usually span millions of base pairs and contain from few to hundreds of candidate genes, making further dissection of their role in CAD very difficult. In fact,

many attempts to identify the drivers of the identified linkage signals were unsuccessful. In general, linkage analysis was extremely useful to identify loci underlying monogenic forms of cardiovascular disease (Wang *et al*. 2003).

For example, a linkage study in a pedigree with clustering of CAD over three generations narrowed a linkage signal to chromosome 15q26 (Wang *et al*. 2003). Of 93 genes within this chromosomal region, *MEF2A* gene was selected as a promising candidate gene based on its profile of expression in embryonic coronary vasculature. A 21-bp deletion in *MEF2A* resulting in the removal of 7 amino acids from the protein product was detected in the affected individuals compared to the unaffected ones (Wang *et al*. 2003). However, a large follow-up study to confirm deleterious mutations in *MEF2A* in sporadic cases of MI did not find any conclusive mutations (Wang *et al*. 2004).

Genome-wide linkage scans had turned out to be more difficult and partly unsuccessful in identifying CAD pathophysiology.

Table 1.5: Linkage studies of CAD and MI

Chromosomal locus	Population (individual/families)	Max LOD score	Phenotype	Candidate genes	Reference
1p34-p36	American (-/428)	11.7	MI	GJA4	(Wang *et al.* 2004)
2p12-q23.3	British (4175/1933)	2.7	CAD/MI	IL1A, IL1B, PROC	(Samani *et al.* 2005)
2q21.1-q22	Finnish (526/156)	3.2	CAD	-	(Pajukanta *et al*. 2000)
3q13	International (1168/438)	3.5	CAD	-	(Hauser *et al*. 2004)
3q27	Indo-Mauritians (535/99)	2.4	CAD/MI	-	(Francke *et al*. 2001)
10q23	Indo-Mauritians (535/99)	2.1	CAD	-	(Francke *et al.* 2001)
13q12	Icelandic (741/296)	2.5	MI	ALOX5AP	(Helgadottir *et al*. 2004)
14q	German (1406/513)	3.9	MI	-	(Broeckel *et al*. 2002)
15q26	European-Americans (21/1)	4.2	CAD/MI	MEF2A	(Wang *et al.* 2003)
16p13-pter	Indo-Mauritian (535/99)	3.1	CAD	SOCS1, ACSM3	(Francke *et al*. 2001)
17p11.2-q21	Europeans(-/739)	2.9	MI	-	(Farrall *et al.* 2006)
Xq23-q26	Finnish (526/156)	3.5	CAD	AGTR2	(Pajukanta *et al.* 2000)

CAD – coronary artery disease, MI – myocardial infarction, LOD – logarithm of odds of linkage, GJA4 – gap junction protein alpha 4, **IL1A** – interleukin 1 alpha, **IL1B** – interleukin 1 beta, **PROC** – protein C, **ALOX5AP** – arachidonate 5-lipoxygenase-activating protein, **MEF2A** – MADS box transcription enhancer factor 2 polypeptide A, **SOCS1**- suppressor of cytokine signaling 1, **ACSM3** – acyl-Coa synthetase medium-chain family member 3, **AGTR2** – angiotensin II receptor type 2

1.8. The road to genome-wide association studies

The substantial progress in genetic studies of complex diseases has been driven through the Human Genome Project with the sequencing and cataloguing of the 3 billion base pairs and ~21,000 genes of the human genome (Lander *et al.* 2001; McPherson *et al.* 2001; International Human Genome Sequencing Consortium, 2004). Further characterisation of the natural genetic variation in human genome in individuals from four different ethnicity backgrounds and precise mapping of their linkage disequilibrium (LD) structure patterns through the International HapMap project (Frazer *et al.* 2007) was another important step forward to progress in this area of research. Knowledge of LD patterns across human chromosomes permitted to estimate the minimum number of SNPs needed in genotyping to capture most of the variation across the selected locus ("tagging strategy"). This strategy reduced dramatically the logistics and most importantly – the costs of genotyping and was probably a major breakthrough in the genetic revolution of complex disorders (Frazer *et al*. 2007; Hinds *et al*. 2005).

The advent of high throughput genotyping technology – parallel typing of several hundred thousand to over a million SNPs located on arrays ("chips") provided scientists with massive amount of information and had paved the way for proliferation of large-scale genome-wide association studies (GWA studies) (McCarthy and Hirschhorn, 2008; Hardy and Singleton, 2009).

In contrast to linkage studies that rely on biologically related subjects, GWA studies use usually cases and disease-free controls to compare allele frequencies of SNPs across the entire genome (Wellcome Trust Case Control Consortium, 2007). Their design makes them suitable mainly for the discovery of common variants conferring low/moderate risks, in the context of the common disease-common variant hypothesis (Reich and

Lander, 2001). According to the hypothesis common polymorphisms (MAF>5%) have modest to moderate magnitude-effects and it is their collective contribution that is responsible for the ultimate genetic predisposition to common-chronic disease (Manolio *et al*. 2008).

Based on this interrogation of genetic variation on a genome-wide basis, GWA studies are an example of "hypothesis free" approach (Altshuler and Daly, 2007). This agnostic basis of GWA studies offered the opportunity to overcome difficulties and obstacles imposed by the incomplete understanding of disease pathophysiology and gave scientists the opportunity to localise disease-related regions more precisely at an unparalleled scale.

1.9. Successes of GWA studies

Much of the excitement that GWA studies brought to the scientific community was based on the expectation that because these studies are hypothesis-free, and thus independent of the pre-existing bias of traditional biology, a comprehensive description of the genetic causes of complex disease would become feasible.

In the past few years, GWA studies have offered valuable knowledge about the biological pathways underlying complex diseases (Frazer *et al*. 2009) and have provided valuable insights into the complexities of their genetic architecture (Hindorff *et al*. 2009). According to the NIHR GWA studies Catalog, there are ~11,500 genetic variants associated with >300 human phenotypes from most common diseases such as CAD to the most unusual such as restless leg syndrome (www.genome.gov/gwastudies). Many previously "unsuspected" genes and pathways were uncovered (Hirschhorn,

2009). Good examples of such pathways are the autophagy and interleukin-23-related pathways in Crohn's disease (Lettre and Rioux, 2008) and the encoding chromatin proteins and hedgehog signaling pathway in human height (Weedon and Frayling, 2008).

One of the most valuable insights (challenge to be solved) provided by GWA studies is that not all of the common complex diseases are same in their genetic architecture. It appears that heritability of some complex disorders/phenotypes is much more driven by common variants than the others. For instance, common variants have been reported to explain approximately 40% of genetic variation in serum-transferrin levels (Benyamin *et al*. 2009), whereas approximately 50 fairly common loci account for only 6% of human height variation (Visscher, 2008).

Several surprising (and the same time challenging) associations uncovered by GWA studies (such as 9p21.3 locus and CAD) posed unprecedented questions regarding the mechanistic role of genetic variation. Variants in such gene-reduced areas would have never been uncovered by the candidate-gene approach. The vast majority (>80%) of associated variants were present outside coding regions, emphasizing the importance of non-coding regions in the human genome (Hindorff *et al*. 2009).

1.10. Challenges of GWA studies

The major generic feature of a majority of discoveries made by these experiments is the small effect size of the individual identified variants that are of little predictive value. Only 3% of the examined putative risk loci showed odds ratios greater than 3 (Pawitan *et al.* 2009). The usual per-allele odds ratios of 80% of the reported associations were

≤1.5 (Pawitan *et al.* 2009) meaning that the current discoveries explain only a small fraction of the genetic contribution to the disease.

Furthermore, discovered variants usually represent markers, rather than causative alleles. A majority of these statistically associated SNPs are just pointers of promising regions (a LD block), rather than the culprit functional defects (Frazer *et al.* 2009; Visscher and Montgomery, 2009). Thus, a challenge that immediately follows the discovery of association between a SNP and a disease is the search for the affected gene and the causal variant(s) at the chromosomal locus.

Clearly, GWA studies are a starting point of a long journey aiming to elucidate and understand the genetic basis of complex diseases and finally translate this information into clinically useful insights.

1.11. GWA studies and CAD

The search for genes that affect the risk of CAD has been fruitless for many decades until the recent advent of GWA studies. The discovery of common risk alleles in CAD from GWAs studies has been a challenge, but that endeavour has yielded some intriguing new findings.

In a landmark study in 2007, the Wellcome Trust Case Control Consortium (WTCCC) genotyped 17,000 samples for 500,000 SNPs, using a set of 3000 common controls and 2000 case subjects from each of 7 complex diseases, including CAD (Wellcome Trust Case Control Consortium, 2007). Along with two additional independent scientific groups which undertook GWA study for CAD, they successfully reported, in parallel, association between common SNPs on the chromosome arm 9p.21.3 and CAD (Helgadottir *et al.* 2007; McPherson *et al.* 2007; Samani *et al.* 2007). Despite the fact

that each study identified a different SNP in the same 30kb region, the reported variants are highly correlated in European populations (r^2=0.85-1.0) indicating that the same underlying genetic risk factor was identified (Arking and Chakravarti, 2009). Since then, the association with this major locus was confirmed by additional studies (Schunkert *et al*. 2008). This finding constitutes the strongest common genetic effect on the CAD risk known today – the risk allele has a frequency of ~46% in individuals of white European ethnicity.

No prior genetic studies had implicated this region as associated with CAD. The CAD associated SNPs in this region were not linked to any other traditional cardiovascular risk factors (such as smoking or lipids), indicating that the biological mechanism underlying this association signal is operating through a novel pathway.

The excitement that this discovery brought was overshadowed by the complexity of connecting some of the identified genetic associations with a pathological mechanism. No genes are located in close proximity to the GWA studies SNPs as defined by LD, which is around 58kb in Europeans (Musunuru and Kathiresan, 2010). However, three genes called cyclin dependent kinase inhibitor 2A (*CDKN2A*) encoding INK4 protein p161NK4a, cyclin dependent kinase inhibitor 2B (*CDKN2B*) encoding p151NK4b, and *ANRIL* were found more than 100kb away from the SNPs associated with the risk of CAD/MI and they are under investigation.

CDKN2A-2B encode for INK4 proteins which belong to a family of cell cycle suppressors (Samani *et al*. 2007). Alterations in the gene expression of these genes could be postulated to lead to senescence and apoptosis, both of which are processes involved in plaque progression and rupture.

ANRIL is a large antisense non-coding RNA gene. It is expressed in atheromatous human blood vessels (vascular endothelial cells, monocyte-derived macrophages, and coronary smooth muscle cells) (Broadbent *et al.* 2008) but its biological functions have not been elucidated yet.

In the PROCARDIS (Precocious Coronary Artery Disease) study, susceptibility to CAD was encoded by two common haplotypes than span the 53kb region that overlaps with *ANRIL* (Broadbent *et al.* 2008). A recent paper using a mouse model has presented data suggesting that *ANRIL* expression is not the most likely mechanism and identified a cis-acting element that influenced expression of *CDKN2A-2B* and thus cell apoptosis (Visel *et al*. 2010).

The 9p21 chromosomal locus is clearly a disease "hot-spot", as it has been also shown to be associated with risk of heart failure (Yamagishi *et al.* 2009), type-2 diabetes mellitus (Saxena *et al*. 2007), abdominal aortic aneurysms (Helgadottir *et al*. 2008), and stroke (Yamagishi *et al*. 2009).

Since then, additional similar studies employed this genome-wide approach and successfully expanded the list with promising genes by confirming associations at various novel loci (Table 1.6).

The first wave of several GWA studies on CAD (Helgadottir *et al*. 2007; McPherson *et al.* 2007; Samani *et al*. 2007; Erdmann *et al.* 2009; Tregouet *et al*. 2009; Kathiresan *et al*. 2009) identified overall 12 risk loci (Table 1.6). In order to achieve higher resolution and an unbiased view on the entire genome, sample sizes of thousands of individuals were studied by international consortia and shared information for subsequent meta-analyses (Preuss *et al*. 2010; Coronary artery disease C4D Genetics consortium, 2011; Schunkert *et al.* 2011, Deloukas *et al*. 2013). Such cooperative effort ultimately led in

successful validation of previously reported loci and identification of additional, novel genetic variants with an increased susceptibility to CAD.

The examples of successful collaborations were the CARDIoGRAM Study (Schunkert *et al.* 2011), the Coronary Artery Disease (C4D) Genetics Consortium (Coronary artery disease C4D Genetics consortium, 2011) the IBC 50K CAD Consortium (The IBC 50K CAD Consortium, 2011) and the Myocardial Infarction Genetics (MIGen) Consortium (Kathiresan *et al.* 2009). These consortia have performed meta-analyses combining the association signals from multiple GWA studies, thus maximizing the power.

The recently published CARDIoGRAMplusC4D study (Deloukas *et al*. 2013) is the largest GWA study assessing the impact of common variants on CAD risk to date. The analysis examined 63,746 CAD cases and 130,681 controls and brought the total number of confirmed CAD susceptibility loci in Europeans and South Asians to 47.

Additional study in a Chinese population mapped a novel genetic variant at 6p21 that increases their risk of CAD but has no effect in the Caucasian population (Wang *et al*. 2011).

In a short span of 7 years, 47 loci were linked to susceptibility to CAD indicating the importance of genetic predisposition for CAD.

Table 1.6: Loci associated with CAD and MI – discoveries of GWA studies

Chromosome	Published lead SNP or proxy	Risk allele frequency	OR	P-value	Candidate genes	References
1	rs602633 (tagging rs599839)	0.77	1.12	$1.47x10^{-25}$	SORT1	Samani *et al.* 2007 Schunkert *et al.* 2011
1	rs11206510	0.84	1.06	$1.79x10^{-5}$	PCSK9	Kathiresan *et al.* 2009
1	rs17114036	0.91	1.11	$5.80x10^{-12}$	PPAP2B	Schunkert *et al.* 2011
1	rs17464857	0.87	1.05	$6.06x10^{-5}$	MIA3	Samani *et al.* 2007 Schunkert *et al.* 2011
1	rs4845625	0.47	1.04	$3.64x10^{-10}$	IL6R	Deloukas *et al.* 2013
2	rs6725887	0.11	1.12	$1.16x10^{-15}$	WDR12	Schunkert *et al.* 2011 Kathiresan *et al.* 2009
2	rs515135	0.83	1.08	$2.56x10^{-10}$	APOB	Deloukas *et al.* 2013
2	rs2252641	0.46	1.04	$5.30x10^{-8}$	ZEB2-ACO74093.1	Deloukas *et al.* 2013
2	rs1561198	0.45	1.05	$1.22x10^{-10}$	VAMP5-VAMP8-GGCX	Deloukas *et al.* 2013
2	rs6544713	0.30	1.06	$2.12x10^{-9}$	ABCG5-ABCG8	IBC 50K CAD, 2011 Deloukas *et al.* 2013
3	rs9818870	0.14	1.07	$2.62x10^{-9}$	MRAS	Erdmann *et al.* 2009 Schunkert *et al.* 2011
4	rs7692387	0.81	1.06	$2.65x10^{-11}$	GUCY1A3	Deloukas *et al.* 2013
4	rs1878406	0.15	1.06	$2.54x10^{-8}$	EDNRA	Deloukas *et al.* 2013
5	rs273909	0.14	1.09	$9.62x10^{-10}$	SLC22A4-SLC22A5	Deloukas *et al.* 2013

Chromosome	Published lead SNP or proxy	Risk allele frequency	OR	P-value	Candidate genes	References
6	rs12190287	0.59	1.07	$4.94x10^{-13}$	TCF21	Schunkert *et al*. 2011
6	rs2048327	0.35	1.06	$6.86x10^{-11}$	SLC22A3-LPAL2-LPA	Tregouet *et al.* 2009 Schunkert *et al*. 2011
6	rs12205331 (tagging rs17609940)	0.81	1.04	$4.18x10^{-5}$	ANKS1A	Schunkert *et al*. 2011
6	rs9369640 (tagging rs12526453)	0.65	1.09	$7.53x10^{-22}$	PHACTR1	Kathiresan *et al.* 2009 Schunkert *et al.* 2011
6	rs10947789	0.76	1.06	$9.81x10^{-9}$	KCNK5	Deloukas *et al.* 2013
6	rs4252120	0.73	1.06	$4.88x10^{-10}$	PLG	Deloukas *et al.* 2013
7	rs11556924	0.65	1.09	$6.74x10^{-17}$	ZC3HC1	Schunkert *et al.* 2011
7	rs12539895	0.19	1.08	$5.33x10^{-4}$	7q22	Deloukas *et al*. 2013
7	rs2023938	0.10	1.07	$4.94x10^{-8}$	HDAC9	Deloukas *et al*. 2013
8	rs264	0.86	1.05	$2.88x10^{-9}$	LPL	Deloukas *et al.* 2013
8	rs2954029	0.55	1.04	$4.75x10^{-9}$	TRIB1	IBC 50K CAD, 2011 Deloukas *et al*. 2013
9	rs1333049	0.47	1.23	$1.39x10^{-52}$	CDKN2BAS1	Samani *et al*. 2007 McPherson *et al*. 2007 Schunkert *et al.* 2011
9	rs579459	0.21	1.07	$2.66x10^{-8}$	ABO	Reilly *et al*. 2011 Schunkert *et al*. 2011
10	rs124134009	0.89	1.10	$6.26x10^{-8}$	CYP17A1-CNNM2-NT5C2	Schunkert *et al*. 2011
10	rs2505083	0.42	1.06	$1.35x10^{-11}$	KIAA1462	Erdmann *et al.* 2009 C4D Consortium, 2011

Chromosome	Published lead SNP or proxy	Risk allele frequency	OR	P-value	Candidate genes	References
10	rs501120	0.83	1.07	$1.79x10^{-9}$	CXCL12	Samani *et al*. 2007 Schunkert *et al*. 2011
10	rs2246833 (tagging rs1412444)	0.38	1.06	$9.49x10^{-6}$	LIPA	C4D Consortium, 2011
11	rs974819	0.290	1.07	$3.55x10^{-11}$	PDGFD	McPherson *et al.* 2007
11	rs9326246	0.10	1.09	$1.51x10^{-7}$	ZNF259-APOA5-APOA1	Schunkert *et al.* 2011
12	rs3184504	0.40	1.07	$5.44x10^{-11}$	SH2B3	Gudbjartsson *et al*. 2009 Schunkert *et al.* 2011
13	rs4773144	0.42	1.07	$1.43x10^{-11}$	COL4A1-COL4A2	Schunkert *et al.* 2011
13	rs9319428	0.32	1.05	$7.32x10^{-11}$	FLT1	Deloukas *et al*. 2013
14	rs2895811	0.43	1.06	$4.08x10^{-10}$	HHIPL1	Schunkert *et al*. 2011
15	rs7173743	0.58	1.07	$6.74x10^{-13}$	ADAMTS7	Reilly *et al*. 2011 C4D Consortium, 2011 Schunkert *et al.* 2011
15	rs17514846	0.44	1.05	$9.33x10^{-11}$	FURIN-FES	Deloukas *et al.* 2013
17	rs12936587	0.59	1.06	$1.24x10^{-9}$	RAI1-PEMT-RASD1	Schunkert *et al*. 2011
17	rs15563 (tagging rs46522)	0.52	1.04	$9.37x10^{-6}$	UBE2Z	Schunkert *et al.* 2011
17	rs2281727 (tagging rs216172)	0.36	1.05	$7.38x10^{-9}$	SMG6	Schunkert *et al.* 2011
19	rs1122608	0.76	1.10	$6.33x10^{-14}$	LDLR	Kathiresan *et al.* 2009 Schunkert *et al*. 2011
19	rs2075650	0.14	1.11	$5.86x10^{-11}$	ApoE-ApoC1	IBC 50K CAD, 2011

Chromosome	Published lead SNP or proxy	Risk allele frequency	OR	P-value	Candidate genes	References
21	rs9982601	0.13	1.13	$7.67x10^{-17}$	Gene desert (KCNE2)	Kathiresan *et al*. 2009

SNP **–** single nucleotide polymorphism, OR – odds ratio, P-value – level of statistical significance**, SORT1** – sortilin 1, **PCSK9** – proprotein convertase subtilisin/kexin type 9, **PPAP2B** – phosphatidic acid phosphatase type 2B, **MIA3** – melanoma inhibitory activity family member 3, **IL6R** – interleukin 6 receptor, **WDR12** – WD repeat domain 12, **APOB** – apolipoprotein B, **ZEB2** – zinc finger E-box binding homeobox 2, **VAMP5** – vesicle-associated membrane protein 5, **VAMP8** – vesicle-associated membrane protein 8, **GGCX** – gamma-glutamyl carboxylase, **ABCG5** – ATP-binding cassette, sub-family G member 5, **ABCG8** – ATP-binding cassette, sub-family G member 8, **MRAS** – muscle RAS oncogene homolog, **GUCY1A3** – guanylate cyclase 1 soluble alpha 3, **EDNRA** – endothelin receptor type A, **SLC22A4** – solute carrier family 22 (organic cation/ergothioneine transporter) member 4, **SLC22A5** – solute carrier family 22 (organic cation/carnitine transporter) member 5, **TCF21** – transcription factor 21, **SLC22A3** – solute carrier family 22 (extraneuronal monoamine transporter) member 3, **LPAL2** – lipoprotein Lp(a)-like 2 pseudogene, **LPA** – lipoprotein Lp(a), **ANKS1A** – ankyrin repeat and sterile alpha motif domain containing 1A, **PHACTR1** – phosphatase and actin regulator 1, **KCNK5** – potassium channel subfamily K member 5, **PLG –** plasminogen, **ZC3HC1** – zinc finger C3HC-type containing 1, **HDAC9** – histone deacetylase 9, **LPL** – lipoprotein lipase, **TRIB1** – tribbles homolog 1, **CDKN2BAS1** – cyclin-dependent kinase inhibitor 2A antisense RNA1, **ABO** – ABO blood group, **CYP17A1** – cytochrome P450 family 17 subfamily A polypeptide 1, **CNNM2** – cyclin M2, **NT5C2** – 5' nucleotidase cytosolic II, **CXCL12** – chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1), **LIPA** – lipase A lysosomal acid cholesterol esterase, **PDGFD** – platelet derived growth factor D, **ZNF259** – zinc finger protein 259, **APOA5** – apolipoprotein A-V, **APOA1** – apolipoprotein A-I, **SH2B3** – SH2B adaptor protein 3, **COL4A1** – collagen type IV alpha 1, **COL4A2** – collagen type IV alpha 2, **FLT1** – fms-related tyrosine kinase 1, **HHIPL1** – HHIPL1 like1, **ADAMTS7** – ADAM metallopeptidase with thrombospondin type 1 motif7, **FURIN** – furin (paired basic amino acid cleaving enzyme), **FES** – feline sarcoma oncogene, **RAI1** – retinoic acid induced 1, **PEMT** – phosphatidylethanolamine N-methyltransferase, **RASD1** – RAS dexamethasone-induced 1, **UBE2Z** – ubiquitin-conjugating enzyme E2Z, **SMG6** – smg-6 homolog, nonsense mediated mRNA decay factor, **LDLR** – low density lipoprotein receptor, **ApoE** – apolipoprotein E, **ApoC1** – apolipoprotein C-I, **KCNE2** – potassium voltage-gated channel Isk-related family member. 2

1.11. Characteristics of identified SNPs and genes in CAD

The findings from GWA studies revealed that genetic variants associated with CAD were not limited to a small number of genomic loci, but instead map to many different regions throughout the genome. Nearly all currently identified risk alleles for CAD are common (MAF>5%). For instance, an individual of European descent has ~50% probability of carrying one risk allele and ~20% probability of carrying both risk alleles of rs4977574 on the chromosome 9p21.3 locus (Roberts, 2008; Schunkert *et al*. 2008). As a result, only ~30% of Europeans are free of this genetic risk factor for MI. Taking into consideration the large number and high frequency of so far identified risk alleles, it is believed that every subject in the population carries multiple genetic variants that increase susceptibility to CAD.

The effect sizes for most of the identified SNPs are modest as each risk allele increases the probability of CAD by only 10-20%. One known exception from this rule is a fairly low frequency variant (approximately 3% in population of white European ethnicity) in *LPA* gene on chromosome 6q25 - it increases CAD risk by 51% (Clarke *et al*. 2009).

The SNPs associated with CAD in GWA studies are not necessarily the causal ones. They rather mark a region on a chromosome where the driving variant(s) is/are located. The studies aiming to identify the causal, biologically active SNPs were primarily based on gene expression studies (through e-QTL like analysis) (Coronary artery disease C4D Genetics consortium, 2011; Wild *et al.* 2011). Several studies using mRNA profiling in tissues /cells of relevance to cardiovascular regulation identified associations between DNA variants within the regions implicated by GWA studies and expression of closely located genes in a quantitative fashion (cis-acting variants) (Kessler *et al.* 2013). For example, one study showed that sortilin 1 (*SORT1*), cadherin, EGF LAG seven-pass G-

type receptor 2 (*CELSR2*) and proline/serine-rich coiled-coil 1 *(PSRC1)* genes at 1p13 displayed decreased expression in carriers of the risk allele due to a disruption of a transcription factor binding site (Musunuru *et al*. 2010). The gene affected most in its expression was *SORT1* encoding for sortilin. The connection of sortilin and LDL metabolism was proved by the finding that liver-specific overexpression of the gene in mice lowered LDL serum levels (Musunuru *et al.* 2010).

Several SNPs associated with CAD in GWA studies are located in regions without known protein-coding genes. The majority of genes underneath the CAD association signals have not been previously implicated in the pathogenesis of CAD.

Only few of the CAD associated genetic variants are related to one of the traditional cardiovascular risk factors. For 17 (36%) of the 47 reported SNPs, the adjacent gene(s) have been implicated in dyslipidemia (*PCSK9, SORT1, ABCG5/8, LPA, TRIB1, ABO, APOA1-C3-A4-A5, LDLR, APOE, APOB, ANKS1A, LPL*) (Kathiresan *et al.* 2008; Teslovich *et al.* 2010, Clarke *et al.* 2009; Deloukas *et al.* 2013) or hypertension (*CYP17A1/ CNNM2/NT5C2, SH2B3, GUCY1A3, FURIN-FES, ZCEHC1*) (Deloukas *et al.* 2013; Newton-Cheh *et al*. 2009; Levy *et al.* 2009; Ehret *et al*. 2011), suggesting a mechanistic pathway for the detected associations with CAD (Kessler *et al.* 2013). For the majority of remaining 30 SNPs the underlying biological mechanisms remain elusive.

Finally, some of the chromosomal SNPs/loci associated with CAD risk are linked to other more distant/unrelated diseases/phenotypes. For example, rs4977574 was associated with hematological parameters (Soranzo *et al.* 2009), and abdominal and intracranial aneurysms (Helgadottir *et al*. 2008). The underlying mechanisms of this phenomenon, known as pleiotropy, are not clear.

Cumulatively, the associated risk variants explain approximately 10.6% of the additive genetic variance of CAD (Coronary artery disease C4D Genetics consortium, 2011; Deloukas *et al*. 2013).

The clinical implications of this research are less apparent and the pathophysiological mechanisms of these discoveries may need further dissection for appropriate counseling of CAD/MI patients and their relatives (Kessler *et al*. 2013).

1.12. Missing heritability of complex diseases

With rare exceptions, the variance explained even by the replicated SNPs is small (usually <1% for each allele) leaving unexplained more than 90% of the heritable component of a disease/phenotype. This raises the question about the nature of the remaining genetic factors contributing to disease or what has been termed the "missing heritability" (Eichler *et al.* 2010; Maher, 2008; Manolio *et al*. 2009).

A number of explanations have been proposed to account for this phenomenon including: (1) additional common variants of small effect, (2) low-frequency/rare variants, (3) structural variation such as copy number variants, (3) gene-gene and gene-environment interactions, (4) epigenetic modifications such as methylation (Maher, 2008; Manolio *et al*. 2009, Prins *et al.* 2012). Other parameters also influence the genetic component of a complex disease such as its phenotypic complexity and genetic heterogeneity (Kullo and Ding, 2007).

1.12.1. Common variants

It is believed that part of the missing heritability is likely to encompass many additional common variants of small to very small genetic effect - early GWA studies were underpowered to detect them (Yang *et al.* 2010). Increasing the sample size of the

studies is an important first step to reduce false negatives. This idea is also supported by the observation that meta-analyses of published GWA studies are discovering a substantial number of new susceptibility loci (Newton-Cheh *et al*. 2009; Sotoodehnia *et al.* 2010; Teslovich *et al*. 2010). Indeed, an extension of a GWA study for plasma lipids from ~20,000 to >100,000 individuals led to the identification of 95 loci (of which 59 were novel) that, in aggregate, explain 10-12% of the total variance (representing 25-30% of the genetic variance) of lipids (Teslovich *et al.* 2010). Despite that the associated SNPs had small effect sizes, some of the new loci contained genes of biological and clinical importance.

However, the number of risk alleles increases at an exponential rate with decreasing relative risk (Prins *et al*. 2012). It is believed that these larger studies will suffer at some point from a plateau phenomenon in which either no additional common variants will be found or any common variants that will be identified will have too small effect to be of biological interest.

1.13.2. Rare variants

Evidence supports the contribution of both common and rare variants to disease risk. It is becoming increasingly clear that low-frequency/rare independent variants, with MAF less than 5%, could probably account for a large fraction of the heritability unexplained by common polymorphisms (Pritchard, 2001; Iyengar and Elston, 2007). This section is extensively explained in Chapter 3.

1.13.3. Structural variation

Analysis of GWA data has been mainly focused on SNPs, but there are other types of genetic variation widespread throughout the human genome. Structural variation,

including copy number variants (CNVs) (such as insertions and deletions) (Sebat *et al*. 2004) and copy neutral variation (such as inversions and translocations) may explain some of the missing heritability of complex diseases (Manolio *et al.* 2009). Variation due to CNVs arises from a combination of rare and common alleles. CNVs may contribute to risk of common diseases i.e. obesity (Bochukova *et al.* 2010). Earlier studies have linked CAD risk with low kringle IV type 2 copy number repeats at the *SLC2A-LPAL2-LPA* locus (Kraft *et al.* 1996) and high number of CA repeats at the *NOS3* locus (Stangl *et al*. 2000). However, GWA studies have been unsuccessful in detecting CNV effects on CAD. Despite the good coverage of CNVs on commercial genotyping platforms, efforts that had been undertaken with aim to examine their effects on CAD and MI showed no significant association (Kathiresan *et al*. 2009; Craddock *et al.* 2010). As a result, CNVs seem unlikely to account for a substantial proportion of CAD missing heritability (Craddock *et al*. 2010).

1.13.4. Gene-gene and genotype-genotype and gene-environment interactions

Many of genes identified by GWA studies cluster together in pathways, co-expression networks and protein-protein interaction networks (Lage *et al.* 2007). Formal examination of gene-gene interactions or genotype-genotype interactions (epistasis) could potentially explain some of missing heritability (Cordell, 2009; Zuk *et al*. 2012).

Up to date the total heritability estimate, is based on the assumption that there are no interactions among the alleles and the effects are additive. Initial efforts to examine gene-gene interactions have been completed with limited success. The detection of epistatic effects statistically remains challenging as it is seriously disadvantaged by statistical corrections due to the large number of multiple tests. For example, to fully

investigate pairwise interactions in a GWA study of 500,000 SNPs over 100 billion tests would have to be performed (Hosking *et al*. 2011).

1.13.5. Epigenetics

Epigenetic effects refer to all heritable gene expression alterations and may be caused through methylation of CpG islands or through methylation, acetylation, phosphorylation, or other modifications of histone proteins (Baccarelli *et al*. 2010; Prins *et al.* 2012). Epigenetic mechanisms are essential and reversible regulators of gene transcription in complex organisms and a driving force of development, evolutionary adaptation and complex diseases (Feinberg *et al.* 2010).

Differential global DNA methylation levels have been shown in global leukocytes in CAD patients when compared to controls. However, there are some inconsistencies regarding the direction of the effect due to the limited resolution of the global methylation measures (Baccarelli *et al*. 2010; Castro *et al*. 2003; Sharma et al. 2008).

It is possible that epigenetic modifications contribute to inter-individual variability in expression of complex phenotypes and in part account for the missing heritability (Furrow *et al*. 2011).

1.12.6. Un-explored regions of the human genome

There are several regions across the human genome that have been neglected in GWA studies. Examples of these un-explored regions are the X and Y chromosomes, the pseudoautosomal regions (PARs) located at the tips of XY chromosomes, and runs of homozygosity (ROHs) that are abundant in autosomes. All of these offer additional avenues to discover genetic risk loci and explaining a portion of the missing heritability for complex phenotypes such as CAD.

Due to the sexual dimorphism of CAD and most of its risk factors (Charhar *et al.* 2003) some attention has been already given on the role of Y chromosome with blood pressure (Charchar *et al.* 2002) and cholesterol levels (Charchar *et al*. 2004). An exploration of the role of the Y chromosome in CAD that was completed by our research group reported a 50% increase in CAD risk in a subgroup of men carrying haplogroup I compared to all the others (Charchar *et al.* 2012; Bloomer *et al*. 2013). The contribution of X chromosome and PARs has not been unraveled yet.

The role of ROHs has been examined in several complex phenotypes, and analysis of genome-wide homozygosity showed a promise towards dissecting the recessive allelic architecture of complex phenotype (Lencz *et al.* 2007; Nalls *et al*. 2009; Keller *et al*. 2012).

1.13. Hypothesis

Un-explored portions/variants of human genome may harbor alleles/genes/loci associated with susceptibility to CAD.

CHAPTER 2

RUNS OF HOMOZYGOSITY AND PREDISPOSITION TO CORONARY ARTERY DISEASE

2.1. Introduction

2.1.1. Historical perspective on recessive inheritance

In 1902, Archibald Garrod reported that certain rare human phenotypes/disorders, such as albinism and alkaptonuria, were more frequent among offspring of consanguineous unions (Garrod, 2002). He attributed that to the fact that relatives have increased probability of possessing two copies of an ancestral allele that is mutated and thus may unmask its pathogenic effect recessively (Khoury *et al*. 1987).

Generally, inbred individuals tend to have higher rates of congenital disorders and lower survival rates and fertility. This phenomenon is called "inbreeding depression" (Charlesworth and Willis, 2009; Keller *et al*. 2011). The magnitude of this effect is linked to the strength of directional selection on the trait (DeRose and Roff, 1999). Fitness traits such as survival, reproduction and disease resistance are more affected by inbreeding than traits under weaker directional selection (DeRose and Roff, 1999). Several studies also suggested inbreeding effects on human complex traits such as cancer (Lebel and Gallagher, 1989), hypertension (Rudan *et al.* 2003b) and osteoporosis (Rudan *et al.* 2004).

Two main theories have been put forward to explain inbreeding depression (Charlesworth and Charlesworth, 1999; Charlesworth and Willis, 2009):

The "partial-dominance" hypothesis concentrates on the role of homozygosity of rare recessive/partially recessive deleterious mutations (Charlesworth and Willis, 2009). A number of deleterious mutations constantly increase in inbred populations and purifying selection rapidly eliminates a majority of additive and dominant ones. However, a segregating pool of deleterious recessive and partially recessive mutations, called

"mutational load" is retained because selection against recessive mutations is inefficient as they have not reached high enough frequencies to start appearing in homozygotes.

The "overdominance" hypothesis postulates that inbreeding depression is caused by a reduction in heterozygosity of common alleles maintained at equilibrium at genomic regions under a heterozygote advantage, so called "segregation load" (Charlesworth and Willis, 2009).

Inbreeding depression is a manifestation of increased levels of homozygosity. However, the role of homozygosity is a subject of interest not only for Mendelian genetics but also complex genetics and structural and functional genomics. Regions of homozygosity in the human genome represent "re-union" of human pieces from common ancestors in their descendants and create a unique opportunity to better understand the consequences of recessive inheritance. This may be particularly relevant to complex polygenic traits as the variants segregating under additive mode of inheritance failed to explain the major component of heritability in the recent GWA studies (Manolio *et al*. 2009). Indeed, one possible explanation for the remainder of missing heritability in complex diseases are highly penetrant recessive alleles missed in GWA studies. Homozygosity mapping is a strategy with the potential to uncover such rare variants hidden within long stretches of homozygous SNPs (McQuillan *et al*. 2008, Ku *et al*. 2010).

2.1.2. What is homozygosity and what it represents?

Homozygosity (autozygosity) arises from identical alleles present on both homologous chromosomes. Runs of homozygosity (ROHs) define long segments of uninterrupted sequences of homozygous SNPs on both homologous chromosomes.

Two identical segments of DNA can reflect a common origin and be identical by descent (IBD) (McQuillan *et al*. 2008; Browning and Browning, 2012) or can be introduced to the genetic pool of a population independently and therefore be identical by state (IBS). The distinction between IBD and IBS is important; IBS ROHs are less likely to contain rare, recessive deleterious mutations in their homozygous form.

To date, a range of different terminologies have been used to describe ROHs, such as "extended tracts of homozygosity" (Gibson *et al*. 2006), "long contiguous stretches of homozygosity" (Li *et al*. 2006), "runs of homozygosity" (Nothnagel *et al*. 2010; McQuillan *et al*. 2008), and "autozygosity regions" (Nalls *et al*. 2009b).

2.1.3. Mechanisms that generate long homozygous segments

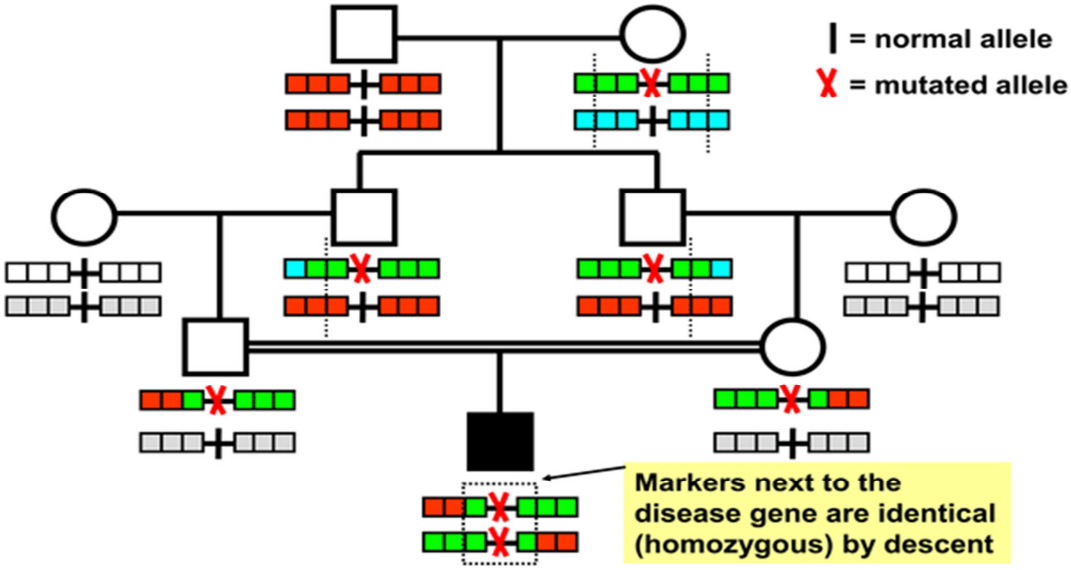
There are three recognised mechanisms leading to ROHs in the human genome:

A. Parental consanguinity

Consanguinity is estimated to be practised by 10% of the world's population (Bittles and Black, 2010). The most common form of consanguineous unions involves 3rd degree relatives (first cousins), which is common in the Middle East, North and Sub-Saharan Africa, Indian Subcontinent and Brazil (Khlat and Khoury, 1991).

This is the most common and well-established mechanism leading to very long ROHs (usually several Mb) - the offspring inherit chromosomal segments that are IBD from each parent (Figure 2.1.1) (McQuillan *et al*. 2008; Pemberton *et al*. 2012). Published data have shown that the number of ROHs of several Mb increases markedly in the offspring of consanguineous marriages (Li *et al*. 2006; Woods *et al.* 2006) with up to 6% of homozygosity anticipated in the genome of first cousin marriages (Broman and Weber, 1999). The number and size of ROHs in offspring of consanguineous unions depends on the degree of parental relatedness (Sund *et al*. 2013).

Figure 2.1.1: Schematic presentation of homozygosity by descent. Filled black symbol defines an individual with an autosomal recessive disease whose parents are consanguineous. A mutation has segregated to the child from both the father's and mother's line, rendering it homozygous for the mutation. [Taken from: Hildebrandt *et al*. 2009]



B. Cytogenetic abnormalities

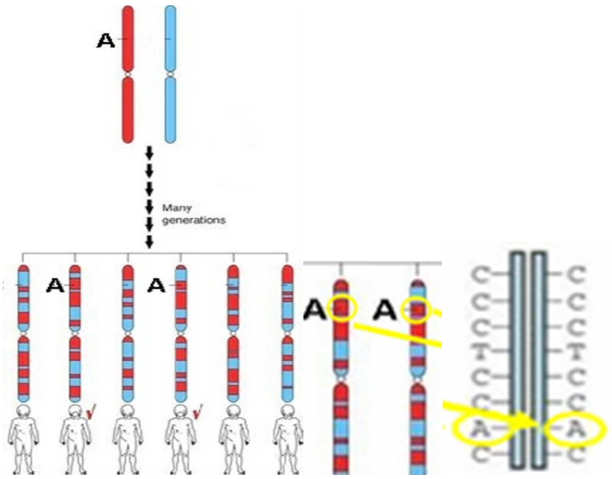
Uniparental isodisomy can also result in homozygosity. The offspring inherit two identical copies of a homologous chromosomal segment from only one parent and as a result no heterozygosity would be observed in that particular homologous chromosomal segment (Ting *et al.* 2007). However, the likelihood that a considerable fraction of

ROHs will be accounted for by uniparental isodisomies is low. Indeed, these are rare genetic abnormalities leading to rare genomic disorders (Liehr, 2010). The most common examples of these disorders are Prader-Willi syndrome (maternally inherited genetic defect in 25% of patients – maps to chromosome 15) (Gurrieri and Accadia, 2009), Angelman syndrome (paternally inherited genetic defect in 2-3% of patients - maps to chromosome 15) (Van Buggenhout and Fryns, 2009), Silver-Russell syndrome (5% of patients inherit a genetic defect on chromosome 7 from their mothers) (Abu-Amero *et al*. 2008). Typically, the type of uniparental disomies in these syndromes is either whole chromosome or segmental isodisomy or a combination of segmental heterodisomy and isodisomy caused by meiotic recombination events. The segments are typically very large, exceeding well over 10Mb (Bruce *et al*. 2005; Altug-Teber *et al.* 2005).

C. *Long haplotypes shared on both chromosomes of an individual*

The presence of common extended haplotypes that happened to be shared on both homologous chromosomes is a recognised mechanism of ROHs in the genomes of outbred populations (Ku *et al.* 2011). Data demonstrating the presence of ROHs in regions with extensive LD and low recombination rates also support this hypothesis (Gibson *et al*. 2006; Curtis *et al*. 2008). Numerous long and frequent ROHs were found to be indicative of positive selection pressure (Enciso-Mora *et al*. 2010; Hosking *et al*. 2010).

Figure 2.1.2: Long haplotypes and ROHs. An ancient chromosome (in red) carries a mutated allele A. The red chromosome pairs with the blue chromosome and over the course of many generations and recombination events this leads to various haplotypes. There are individuals carrying the same haplotypes. Two individuals marked with a yellow circle carry a segment of the red chromosome carrying the haplotype with the risk allele A. As a result, their offspring may inherit this haplotype on both parental chromosomes (with the mutated allele A). The chromosomal regions surrounding the homozygous mutation will not have been broken by crossing overs and thus SNP markers present in these segments will also be homozygous by descent. These runs may potentially harbour functional relatively rare variants that could exert their pathological effects in the homozygous recessive state. [Taken from: http://hapmap.ncbi.nlm.nih.gov/originhaplotype.html.en]



2.1.4. Human population structure and genome homozygosity

Population history and cultural factors can affect levels of homozygosity in individual genomes. There are populations in which a historical bottleneck (small population size) or geographic isolation (island) resulted in elevated levels of background relatedness (Li *et al*. 2006; Jakkula *et al.* 2008; McQuillan *et al.* 2008; Gross *et al.* 2011; Humphreys *et al.* 2011). For example, in the Republic of Croatia, there are 15 Adriatic Sea islands with just >1,000 inhabitants. Some of the villages on these islands have been genetically isolated for centuries from other villages and the outside world and as a result they show increased levels of homozygosity (Rudan *et al*. 2006).

Familial traditions that promote consanguineous marriage or endogamy can lead in high inbreeding levels and therefore increased levels of homozygosity in human genome,

even when the overall size is large (Li *et al.* 2006; Woods *et al.* 2006; Hunter-Zinck *et al.* 2010). Close-kin marriage continues to be preferential in many major populations such as sub-Saharan Africa, and in populous Asian countries including Saudi Arabia Pakistan and India (Bittles, 2002; Bittles, 2008; Bittles and Black, 2010). Even today 10.4% of the 6.7 billion global population are related as second cousins or closer (Bittles and Black, 2010).

In those populations evidence has been reported for several effects, including an increased risk of monogenic disorders (Bittles, 2003; Khlat and Khoury, 1991; Modell and Darr, 2002), an increased risk of complex diseases and disease traits such as blood pressure (Rudan *et al*. 2003; Campbell *et al*. 2007) and LDL-cholesterol (Campbell *et al*. 2007).

Prior to the 19th century, most of the human populations lived and reproduced in small societies with limited mate choice. Over the last 200 years, dramatic changes occurred in the demographic structure of the world's population. Some of the processes involved, on both regional and global level, were increase in the population size, outbreeding, gene flow and admixture (Rudan *et al*. 2006).

The world's total population size expanded from 1 to 6.2 billion, and the percentage of the global population living in cities has increased from 2% to ~50% (Bittles, 2008; Campbell *et al.* 2009). This unprecedented increase in the population size happened as a result of several measures (vaccination, antibiotics, treatment of infections, improved nutrition) to reduce childhood mortality. These actions led to reduced selection in childhood and kept the human population reasonably constant for centuries. Mutations have also increased genetic diversity through the generation and preservation of large numbers of new rare genetic variants (Wright *et al.* 2003; Reich and Lander, 2001).

Therefore, the population size increase is characterised as a possible cause of increased genetic diversity of contemporary human populations in comparison with those that lived two centuries ago, and the effect is expected to be largely due to rare and recently introduced genetic variation.

The composition of gene pools is also different nowadays due to migration. People from small, usually isolated and genetically uniform communities have moved to larger cities and this has led to increased levels of marriage among individuals from different geographical, ethnic, religious and social backgrounds (Bittles, 2008; Darvasi and Shifman, 2005).

This process of "isolate break-up" has led to the admixture of many genetically differentiated populations and in conjunction with migration and rapid urbanization, have all contributed to gene flow and generated more genetically diverse breeding pools. Collectively, the changes have led to a decrease in the level of population substructure and LD in European populations (Helgason *et al*. 2005; Vitart *et al*. 2005).

The key role of homozygosity in many human diseases has fuelled a continued interest in studying the causes and patterns of homozygosity (Pemberton *et al*. 2012).

2.1.5. Genome-wide mapping of ROHs in the human genome

Studies that have investigated inbreeding effects on complex disease using pedigree data suggest that inbreeding is a risk factor (Rudan *et al*. 2003, Campbell *et al.* 2007). However, close inbreeding cannot be a major contributor to late-onset disease risk in modern populations given its rarity. Nevertheless, inbreeding is a matter of degree, when distant relatives are considered; and to some extent "all humans are related"

(Alkuraya, 2010) and are inbred to some degree as there is a limited number of founders to whom contemporary human populations trace their ancestry. It is likely that the parents of the vast majority of people alive today share a common ancestor within ~15 generations (Keller *et al*. 2011). On the other hand, the limited migration until recently, and the creation of many bottlenecks by famines, warfare and epidemics, have reduced the mating pool, such that a more recent common ancestry can be expected for many of current human populations (Weber, 2006). The Finnish population is a prime example of a bottleneck where the population of around 5 million can trace their ancestry to a small number of founders in the not too distant past (Norio, 2003).

Although such distant inbreeding would be very difficult to detect from pedigrees, it can leave signals/traces in the genome that are detectable using genome-wide microsatellite or SNP data. Indeed, this prediction was confirmed first in the Centre d'Etude du Polymorphisme Human panel by microsatellites, and since the advent of SNP chip-based genotyping, a number of other studies clearly demonstrated the presence of relatively frequent ROHs in outbred populations where consanguinity is outlawed (Broman and Weber, 1999; Li *et al*. 2006; McQuillan *et al*. 2008). Observational studies have revealed that ROHs longer than 1Mb are more common in outbred individuals than previously thought (Broman and Weber, 1999; Lencz *et al.* 2007a; Simon-Sanchez *et al*. 2007; Gibson *et al*. 2006; Curtis *et al*. 2008; Li *et al*. 2006).

It was not previously expected that the genomes of outbred populations contain ROHs as long as several megabases until the first few reports in 2006 (Gibson *et al*. 2006). These ROHs are unlikely to be explained on the basis of uniparental isodisomy or deletions (Curtis, 2007; Simon-Sanchez *et al*. 2007).

2.1.6. Characteristics of ROHs

I. ROHs size classification

The ROHs size classification is determined by LD patterns, population history and consanguinity levels.

A. Short ROHs measuring tens of kb probably reflect homozygosity for ancient haplotypes that contribute local LD patterns (from ~100 generations back).
B. Intermediate length ROHs measuring hundreds of kb to several Mb might result from background relatedness owing to limited population size.
C. Long ROHs measuring tens of Mb are the signature of recent parental relatedness (from less than 10 generations back).

Previous studies showed that short ROHs are common covering up to 1/3 of the human genome (Frazer *et al*. 2007).

At the other end of the spectrum, very long ROHs reflect a different and more recent phenomenon. They have been observed in as many as 28-90% of individuals from populations with higher levels of background relatedness; (Li *et al*. 2006; Jakkula *et al*. 2008; McQuillan *et al.* 2008; Gross *et al*. 2011, Broman and Weber, 1999; Kirin *et al*. 2010; Roy-Gagnon *et al.* 2011) surprisingly, they have also been observed in 2-26% of individuals in apparently outbred populations (Li *et al.* 2006; McQuillan *et al*. 2008; Gross *et al*. 2011; Frazer *et al.* 2007; Gibson *et al.* 2006; Lencz *et al*. 2007a; Curtis *et al*. 2008; Kirin *et al.* 2010; Nothnagel *et al*. 2010; O'Dushlaine *et al.* 2010; Roy-Gagnon *et al.* 2011; Teo *et al*. 2011; Teo *et al.* 2012; Simon-Sanchez *et al.* 2007; Auton *et al.* 2009). Long genomic segments containing ROH are common in many populations including Han Chinese, indigenous Taiwanese, Caucasians and African Americans (Broman and Weber, 1999; Gibson *et al*. 2006; Li *et al.* 2006; Simon-Sanchez *et al*.

2007, Curtis *et a*l. 2008). It is worth noting that some of these ROHs potentially result from lack of recombination that allows unusually long ancestral segments to be maintained in the general population (Pemberton *et al.* 2012).

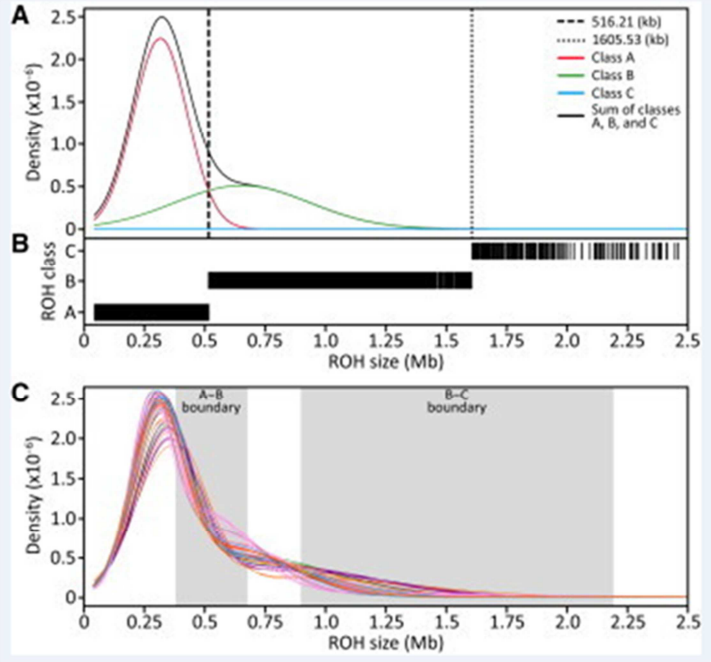
In between, ROHs of intermediate sizes, also occur frequently (Jakkula *et al*. 2008; McQuillan *et al*. 2008; Gross *et al.* 2011; Gibson *et al*. 2006; Lencz *et al*. 2007a; Curtis *et al*. 2008; Nalls *et al*. 2009; Kirin *et al*. 2010; Nothnagel *et al*. 2010; O'Dushlaine *et al.* 2010; Roy-Gagnon *et al*. 2011; Teo *et al.* 2011; Teo *et al.* 2012), probably as a result of recent (but unknown kinship among the parents of sampled individuals) parental relatedness. Also, they might be autozygous segments of much older pedigrees that have occurred by chance inheritance through both parents of extended haplotypes that are at a high frequency in the general population, possibly because they convey some selective advantage (Lencz *et al.* 2007a).

The Phase II HapMap study estimates that ROHs measuring in excess of around 100kb constitute 13-14% of the genome in Europeans (Frazer *et al*. 2007).

II. Classification of ROHs based on location

Across the genome there are regions where ROHs are very frequent (ROHs hotspots) or infrequent (ROHs coldspots) (Pemberton *et al.* 2012). ROHs hotspots have reduced genetic diversity and thus increased homozygosity compared to the rest of the genome, whereas ROHs coldspots show the opposite. The existence of ROH hotspots and ROH coldspots can be explained partly by the frequency of recombination events across the genome or variation across the genome in the effects of demographic processes influencing genetic diversity. However, ROHs hotspots could also represent regions that harbour targets for positive selection and have reduced genetic diversity and an increase in homozygosity around selected loci (Pemberton *et al*. 2012).

Figure 2.1.3: Size-based classification of ROHs. (A) An example of Gaussian kernel density estimation of the ROH size distribution, the boundary between A and B classes is marked by the vertical dashed line and the boundary between B and C classes by the vertical dotted line. (B) ROHs are divided into three classes; Only ROHs less than 2.5Mb in length are shown (C) Gaussian kernel density estimates of the ROH size distribution in each of 64 populations; each line represents a different population. The boundaries between A and B classes and between B and C classes are shown by the shaded boxes. [Taken from: Pemberton *et al.* 2012]



The non-uniform distribution of ROH across the human genome could reflect local genomic properties to maintain homozygosity around certain regions (Pemberton *et al*. 2012). Because recombination events reduce LD and the probability of having two copies of the same haplotype, local recombination rate is expected to be negatively correlated with ROH frequency (Pemberton *et al*. 2012). As recombination acts over many generations, its effect might be expected to be greater on class A and B ROHs than on class C ROHs. Indeed, classes A and B of ROHs probably result from population-level LD patterns on longer time scales, whereas class C ROH probably result from recent inbreeding and thus might have had fewer opportunities for recombination events to systematically occur in high recombination regions. Conversely, as recombination disrupts longer haplotypes more frequently than shorter

haplotypes, one might instead expect the influence of recombination rate to be greater on classes of B and C of ROHs than on class A of ROHs.

Several studies have suggested that ROHs cluster in regions of the genome where recombination rates are low (Simon-Sachez *et al.* 2007; Gibson *et al*. 2006; Curtis *et al*. 2008; Li *et al*. 2006; McQuillan *et al*. 2008).

2.1.7. Homozygosity mapping

Homozygosity mapping, also called autozygosity mapping has served as the most successful disease gene discovery strategy in the recent history of human genetics (Altshuler *et al.* 2008). This method was first proposed by Smith (Smith, 1953) and later developed by Lander and Botstein (Lander and Botstein, 1987). Its principle is the detection of recessive mutations through mapping them to homozygous regions in patients born from consanguineous matings. The greater the number of affected family members who have a shared homozygous region and the longer the length of the region, the more likely it is to harbour the causal disease mutation (Woods *et al.* 2006). Mueller and Bishop, suggested that use of a single multi-affected pedigree is the most efficient autozygosity-based strategy to map a disease region as it reduces the genetic and phenotypic heterogeneity (Mueller and Bishop, 1993).

2.1.7.1. Drawbacks of homozygosity mapping

Historically, the method has faced several major challenges: finding a consanguineous family with the disease in question, observing enough recombination events within the family to narrow down the autozygous region surrounding the mutation, and prioritising candidate genes within that region for further sequencing (Alkuraya, 2010).

Firstly, it is possible for an extended consanguineous family to harbour mutations in two or more different genes giving rise to the same phenotype, particularly when the phenotype is heterogeneous (Benayoun *et al*. 2009; Lezirovitz *et al*. 2008; Frishberg *et al.* 2007; Ducroq *et al*. 2006; Laurier *et al.* 2006; Miano *et al.* 2000; Pannain *et al*. 1999). Secondly, apparently shared autozygous blocks may in fact be IBS. This is particularly difficult when dealing with smaller intervals because the probability of sharing two haplotypes by chance only, is inversely correlated to their lengths (Alkuraya, 2010). Furthermore, the number of shared autozygous blocks between the different members of a given family is a function of the random crossing over events and their frequency. Although their randomness may not be predicted, the number of crossing over events correlates with the number of meiotic events separating the proband from the shared parental ancestor (Genim *et al*. 1998).

On the other hand, higher degrees of autozygosity that would be expected based simply on the level of parental relatedness is frequently observed because extensive background inbreeding is often not reflected in small pedigrees. As a consequence, more blocks of autozygosity will be shared between the affected members by chance only without being truly linked to the disease locus. This may complicate the analysis and reduces the significance of the mapping signal (Carr *et al.* 2006; Leutenegger *et al.* 2006; Woods *et al*. 2006).

2.1.7.2. Genetic markers used to detect homozygosity

Homozygosity mapping became practical with the discovery of highly polymorphic microsatellite repeat markers spread throughout the genome (Alkuraya, 2010). However, it was the advent of SNP chips that enabled high throughput genotyping and successful completion based on autozygosity mapping. Typically, provided the

appropriate family pedigree is available; such projects can now be finished in several weeks (Alkuraya, 2010). Although homozygosity mapping has been focused primarily on inbred populations, recent advances provided a basis for homozygosity mapping in affected individuals who are not inbred (Gibbs and Singleton, 2006; Hildebrandt *et al*. 2009; Collin *et al*. 2011; Hagiwara *et al*. 2011; Schuurs-Hoeijmakers *et al*. 2011). Deleterious recessive variants in such individuals might reside in smaller ROHs than in inbred populations (Pemberton *et al*. 2012).

Whilst each SNP has far less power to detect a homozygous chromosomal segment than a microsatellite marker, it is both their number and their ability to detect a heterozygous region, and hence exclude linkage, that suggested their potential use in autozygosity mapping (Carr *et al*. 2009).

For example, a locus with only one variant that has a heterozygosity score of just 0.5 has a 50% chance of being homozygous by chance. On the other hand, two variants of which each has a heterozygosity score of 0.25 can be expected to be both homozygous by chance with a comparable probability of (1-0.25)x(1-0.25)=56%. Thus, even markers with low heterozygosity scores can be informative when present in high density (Alkuraya *et al*. 2010). It is this simple concept that gives SNPs, despite their low individual heterozygosity (when compared to microsatellites) but high frequency in the human genome, higher overall power as markers in homozygosity mapping (Polasek *et al*. 2010; Carr *et al*. 2006; Evans and Cardon, 2004).

With the increasing availability of data from high-density SNP-based genome scans, it has become feasible to identify and map ROHs in the human DNA at a large scale (McQuillan *et al*. 2008). One of the examples of their use in autozygosity mapping is to monitor the impact of outbreeding on individual and community levels. In the USA

population, ROHs size and frequency dropped over the 20th century (Nalls *et al*. 2009). Indeed, 14% decrease in ROHs frequency, a 24% decrease in the genome length covered by ROHs and a 30% reduction in the inbreeding coefficient (F) was detected through analysis of SNP-based data from genome-wide association scan (Nalls *et al.* 2009).

2.1.8. Homozygosity mapping and monogenic human disorders

The homozygosity mapping approach has successfully been used to map genes associated with recessive Mendelian diseases in hundreds of studies (Botstein and Risch, 2003). Nearly 200 studies published between 1995 and 2003 used homozygosity mapping in consanguineous families (typically first line cousins) to identify rare recessive mutations underlying rare disorders. For example, the Fanconi anaemias are caused by loss of any of 5 genes; 3 of which were mapped by homozygosity analysis (Gschwend *et al*. 1996; Saar *et al.* 1998; Waisfisz *et al.* 1999). However, recessive variants appear also in offspring of parents whose genetic ancestry is more distant or even in those with no evidence of common parental ancestry (Nalls *et al*. 2009).

2.1.9. Regions of homozygosity and their impact on complex diseases

Although the role of ROHs in unmasking recessively acting mutations is well established in Mendelian genetics, much less is known about their contribution to more complex disorders such as cardiovascular disease. Identification of ROHs at a genome-wide scale provides a measure of autozygosity extent and may ultimately expose regions harbouring recessively acting mutations.

Several studies in consanguineous or small isolated populations with above than average levels of parental relatedness have found evidence for a genome-wide effect of homozygosity on CAD (Shami *et al*. 1991; Puzyrev *et al.* 1992), cancer (Shami *et al*. 1991), blood pressure (Rudan *et al*. 2003; Campbell *et al*. 2007) and LDL-cholesterol (Campbell *et al*. 2007). These findings may suggest that the variants associated with increased risk of common complex disease are more likely to be rare than common in populations (Freimer *et al.* 2004), distributed abundantly rather than sparsely across the genome (Wright *et al*. 2003), and being recessive than dominant (Reich and Lander, 2001).

It is believed that homozygosity is linked to both the common disease/common variant (CD/CV) and common disease/rare variant (CD/RV) hypotheses (Schork et al. 2009). With the additive model of inheritance, many risk SNPs identified by GWAs have been shown to exert their pathologic effects more strongly when present in two copies. Based on this it is sensible to investigate if weak recessive alleles can increase the predisposition to complex diseases in the homozygous state (Bittles and Black, 2010; Schork *et al*. 2009).

The relationship between homozygosity and complex diseases has not been extensively examined, and it is only recently that it started being tested. The first study applying the homozygosity association approach at the genome-wide scale for complex diseases only appeared in 2007 (Lencz *et al*. 2007a). A genome-wide homozygosity analysis was applied in 178 schizophrenic cases and 144 healthy controls and ROHs were found significantly more frequent in cases compared to controls. Nine of those ROHs were over-represented in cases. Four of these risk ROHs overlapped or neighboured genes associated with schizophrenia (Lencz *et al*. 2007a). Another very recent study used

ROHs to estimate the proportion of the autosomal genome that is covered in homozygous tracts in >9,000 schizophrenic cases and >12,000 controls (Keller *et al*. 2012). Interestingly stated that for every 1% increase in genome-wide homozygosity there is a ~17% increase in schizophrenia risk (Keller *et al*. 2012).

The success of ROHs approach has been demonstrated in several studies. An association study involving 837 Alzheimer's disease cases and 550 controls identified one ROH on chromosome 8 containing three biologically relevant candidate genes (Nalls *et al.* 2009). A homozygosity study on height, a complex quantitative and highly heritable trait, revealed a ROH on chromosome 12 that had a strong association with adult height variation in both discovery and replication stages (Yang *et al*. 2010). Individuals with this ROH were 3.5cm taller than individuals without it (Yang et al. 2010). Moreover, a genome-wide homozygosity analysis in ~1,500 early onset Parkinson disease cases and ~7,000 controls showed an increased homozygosity level in Parkinson's disease (Simon-Sanchez *et al.* 2012). The biggest difference was observed in ROHs of 9Mb and above (4.4% cases vs 1.4% controls). A locus was identified on chromosome 19p13.3 as over-represented in cases compared to controls but sequencing analysis within the locus failed to identify a novel mutation.

Other examples include genome-wide homozygosity analysis of lymphoblastic leukemia (Hosking *et al.* 2010), bipolar disorder (Vine et al. 2009), rheumatoid arthritis (Yang et al. 2012a), colorectal cancer (Bacolod *et al.* 2008), breast and prostate cancer (Enciso-Mora *et al.* 2010). CAD has never been examined in ROH-based analyses.

2.1.10. Hypothesis

CAD patients and controls will show different patterns of homozygosity in their genome.

2.1.11. Objectives

- to perform genome-wide homozygosity analysis of the overall genetic architecture of ROHs in the general population
- to perform comprehensive analysis of association between homozygosity measures and CAD in individuals of white European ancestry
- to identify individual consensus regions of overlapping ROHs in relation to CAD

2.2. Materials and Methods

2.2.1. Characteristics of study cohorts

A total of 20,821 individuals recruited from 9 populations of white European ancestry included in the CARDIoGRAM consortium was used in this project (Schunkert *et al*. 2011). Full details for each cohort used are given below.

- ***Wellcome Trust Case-Control Consortium Study (WTCCC)***

WTCCC cases: This British population includes 1,988 patients with a validated history of either myocardial infarction (MI) or coronary revascularization (coronary artery bypass surgery or percutaneous coronary angioplasty) before the age of 66 years. Recruitment was carried out on a national basis and lasted 5 years (April 1998 – November 2003). A major proportion of biologically unrelated patients with CAD in this study come from families with two or more affected siblings conducted in UK - British Heart Foundation Family Heart Study (BHF-FHS) (Samani *et al*. 2005). Some individuals in this resource come from GRACE Study (Alfakih K *et al.* 2007) that recruited patients with CAD and familial history of premature CAD (in parents or sibling) but in whom an affected sibling was not available for recruitment.

In the WTCCC study, subjects from both BHF-FHS and GRACE cohorts were used together totaling 2,000 unrelated cases affected by premature CAD. 1,518 of those subjects were derived from BHF-FHS (Samani *et al*. 2005) and 470 subjects were from families included in GRACE study (Alfakih K *et al.* 2007).

WTCCC controls: A population of 3,004 control subjects derived from two independent publicly accessible sources: the British 1958 Birth Cohort (58BC) and a sample of blood donors from UK Blood Service (UKBS).

The 58BC, also known as the National Child Development Study (Power and Elliott, 2006), includes all births in England, Wales and Scotland during one week in 1958. From an original sample of over 17,000 births, survivors were followed up at ages 7, 11, 16, 23, 33 and 42 years. In a biomedical examination at 44-45 years, 9,377 cohort members were visited at home providing 7,692 blood samples. DNA samples extracted from 1,500 cell lines of subjects with self-reported white ethnicity and representative of gender and each geographical region were selected for use as controls.

The UKBS provided 1,500 controls from a sample of healthy blood donors recruited as a part of the WTCCC project (Samani *et al*. 2007). WTCCC in collaboration with UKBS set up a UK national repository of anonymised samples of DNA and viable mononuclear cells form 3,622 consenting blood donors (age range: 18-69 years, majority of them between 40-59 years). A set of 1,564 samples was selected from the total number of samples based on sex and geographical region to reproduce the distribution of the samples of the 1958 Birth Cohort for use as common controls in the WTCCC study. The subjects were about equally divided into males and females and both control groups were geographically widely distributed across the UK. Except from gender information and 10 year age-band, additional phenotypic information was not available on the control cohorts.

- ***German Myocardial Infarction Family Study I (GerMIFSI)***

GerMIFSI cases: This population consists of 875 patients with a validated history of MI before the age of 60 years. The recruitment was carried out after screening >200,000 patients charts in 17 cardiac in-hospital rehabilitation centres. The majority (>70%) of patients were recruited in the surrounding area of Augsburg and the Southern part of Germany between 1997 and 2002. All patients were of white German origin. If at least

one additional first-degree family member (preferentially a sibling) had suffered from MI or had severe coronary artery disease (percutaneous transluminal coronary angioplasty [PTCA] or coronary artery bypass grafting surgery [CABG]), the family (index patient, available parents and all siblings) was approached and asked to join the study (Broeckel *et al*. 2002; Fischer *et al* .2005). All events in index patients and family members were validated through inspection of hospital charts.

GerMIFSI controls: All 1,644 CAD-free controls had German descent and were collected for the MONICA/KORA Augsburg population study in the years 1994/1995 and a follow up of this project in the years 2004/2005 that was undertaken as part of the German National Genome Research Network (NGFN) (Wichmann *et al.* 2005). This population represents a gender- and age- stratified random sample of all German residents of the Augsburg area and consists of individuals 25 to 74 years of age, with about 300 subjects for each 10-year age band. These individuals were studied by physical examination, blood testing, and a standardized interview including medical history, physical activity, medication and personal habits.

- ***German Myocardial Infarction Family Study II (GerMIFSII)***

GerMIFSII cases: This population consists of 1,222 patients with a validated history of MI before the age of 60 years for both men and women (Erdmann *et al*. 2009). A positive family history for CAD was documented in 726 (59.4%) of patients. Patients were identified following their admission for acute treatment of MI or in cardiac rehabilitation clinics.

GerMIFSII controls: A total of 820 CAD-free controls were derived from the MONICA/KORA Augsburg survey S4 (Holle *et al*. 2005). A sample of 478 controls was taken from PopGen blood donor sample 2 (PopGen-DSP) (Krawczak *et al*. 2006).

- ***German Myocardial Infarction Family Study III (GerMIFSIII - KORA)***

GerMIFSIII cases: A total of 1,157 subjects with available DNA were derived from patients with non-fatal MI in the KORA registry (Erdmann *et al*. 2011). Hospitalised survivors of MI aged 26-74 years are routinely entered into this registry (Lowel *et al*. 2005). The diagnosis of MI was made with the use of algorithm of the World Health Organization's Multinational Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA) project (Peters *et al*. 2004).

GerMIFSIII controls: A total of 996 and 752 CAD-free controls were derived from the population-based Augsburg KORA S4/F4 study (Kolz *et al*. 2009) and PopGen respectively (Krawczak *et al*. 2006).

- ***Ottawa Heart Genomics Study (OHGS)***

The OHGS is a hospital-based study of CAD conducted at the Ottawa Heart Institute in Ottawa, Canada. All patients who undergo CABG, coronary artery angiography, or receive treatment for MI are invited to participate in the study (McPherson *et al*. 2007; Davies *et a*l. 2010). Three independent samples (OHGS-A, OHGS-B and OHGS-C) were ascertained consecutively for this study.

OHGS cases: Patients with documented CAD (stenosis in a major epicardial vessel of at least 50%; PCI; CABG or MI before the age of 55 years in men and 65 years in women) were recruited for OHGS. Patients with history of diabetes mellitus of severe dyslipidemia were excluded.

OHGS controls: Healthy elderly controls (men aged >65 years and women aged >70 years) were recruited via an extensive newspaper and television advertising campaign in the Ottawa community. Controls were carefully interviewed by a physician or nurse to ascertain that they were free of symptoms of possible ischemic arterial disease and had

no past history of cardiovascular symptoms, a positive stress test, coronary angiography demonstrating stenosis (>50%) in any artery or clinical cardiovascular events. Individuals with the same ethnic background as the cases were included in this study. The mean age of CAD-free control was 74 years. Controls for all three samples were collected consecutively as described for cases.

- ***Cleveland Clinic Gene Bank (CCGB)***

All cases in CCGB study were recruited using the same criteria as in OHGS (Davies *et al.* 2010). Patients were included if they had at least one of the following before the age of 55 years (men) or 65 years (women): angiographically documented stenosis in a major epicardial artery of at least 50%; history of MI, a percutaneous intervention (PCI), or CABG. Diabetes was an exclusion criterion.

- ***The Duke Cathgen Study (DUKE)***

Subjects were recruited from the cardiac catherisation laboratory at Duke, Canada. Cases had at least one epicardial coronary vessel with ≥50% stenosis documented before age of 55 years (men) or 65 years (women). Controls (≥50 years of age) were recruited from amongst patients who underwent coronary angiography and had no more than 30% stenosis in maximally one epicardial coronary artery. Any subject with diabetes mellitus, severe pulmonary hypertension or congenital heart disease was excluded from DUKE Study. These data have not been published yet.

- ***PennCATH***

PennCATH is a university of Pennsylvania Medical Center based coronary angiographic study. Briefly, between July 1998 and March 2003, PennCATH recruited a consecutive cohort of patients undergoing cardiac catheterization and coronary angiography (Grant *et al.* 2006; Kathiresan *et al.* 2009). Cases were required to have at

least one stenosis of ≥50% in at least one of major epicardial arteries before the age of 60 for males and 65 for women.

In contrast, CAD-free controls (aged >40 years /men/ and 45 years /women/) had no stenosis exceeding 10% in any major epicardial artery.

All participants gave written consent for participation in genetic studies, and the protocol of each study was approved by the corresponding local research ethics committee or institutional review board.

2.2.2. Genotyping and imputation

All cohorts had available genome-wide genotype information from previously conducted GWA studies (Samani *et al*. 2007; Schunkert *et al*. 2011). Apart from directly genotyped SNPs (from ~250,000 to ~900,000 dependent on the population and the genotyping platform used), imputed genotypes (based on HapMapII CEU build 36 as a reference) were available in each population (Table 2.2.1).

The genome-wide analysis was based on approximately 2.5 million SNPs. Imputation was performed using specialized imputation algorithms (either IMPUTE (Howie *et al.* 2009) or MACH (Li *et al*. 2010)), separately in each cohort. The imputation software predicts untyped genotypes at the genome-wide level based on a set of known haplotypes. The accuracy of the imputation depends on SNP density and the similarity among the LD patterns between the study sample and the reference HapMap population (Marcini *et al.* 2007). Only genotypes with a posterior probability of ≥90% were included in this analysis.

2.2.3. Quality control filters

Prior to the association analysis, both the SNP and population sample data were subjected to stringent quality control and cleaning procedures. SNPs were removed from further analysis if their MAF<1% and/or their genotype distribution deviated from Hardy-Weinberg equilibrium (HWE) ($p<0.001$ on the HWE test) and/or the genotype missingness rate was ≥5%.

Individual samples were excluded from further analysis if they failed any of the following quality control filters: (SNP genotype missingness rate >5%, low heterozygosity, external discordance, non-European ancestry, duplicate, cryptic relatedness) were excluded from the analysis. All quality control filters were applied individually at the cohort level.

Table 2.2.1: Genotyping and imputation information for each study

Study acronym		**CCGB**	**DUKE**	**GerMIFSI**	**GerMIFSII**	**GerMIFSIII**	**OHGS-A**	**OHGS-C**	**PennCATH**	**WTCCC**
Genotyping	Platform	Affymetrix 6.0	Affymetrix Axiom	Affymetrix NSP and STY	Affymetrix 6.0	Affymetrix Genome-Wide Human SNP Array 6.0	Affymetrix Mapping 500K	Affymetrix Axiom	Affymetrix 6.0	Affymetrix Mapping 500K Array Set
	Calling algorithm	Birdseed	AxiomGT1	Birdseed	BRLMM	Birdseed	BRLMM	AxiomGT1	Birdseed	CHIAMO
	Genotyped SNPs	NA	NA	262,338(NSP)/ 238,378(STY)	909,622	503,590(5.0)/ 904,954(6.0)	325,040	NA	869,223	477,459
	Imputed SNPs	-	-	2,543,887	2,543,887	2,536,369	2,469,454	-	2,749,197	2,614,446

2.2.4. ROHs idenfitification

A number of ROHs definitions and methods of their detection have been proposed (Auton *et al.* 2009, Curtis *et al.* 2008, Gibson *et al.* 2006, MacLeod *et al.* 2009). In this project ROHs were identified via the "Runs of Homozygosity" program implemented in PLINK version 1.07 (Purcell et al. 2007).

The PLINK ROH tool moves a sliding window of a defined number of SNPs and size across the entire human genome to detect runs of homozygous genotypes. Based on set thresholds, it is determined at each position whether the window meets the required level of homozygosity. For each SNP, the proportion of homozygous windows that overlaps the SNP position is calculated and used to determine whether ROH meets the minimum criteria (numbers of SNPs or size) (Purcell *et al*. 2007). Sliding window size, number of SNPs and the region length were taken into consideration to define what constitutes a ROH.

A sliding window of 50 SNPs in 5,000 kb length region was used to scan the genome (Purcell *et al*. 2007). To prevent underestimating the number and size of ROHs, 1 heterozygote and 2 missing calls in each window were permitted (Nalls *et al*. 2009; Hosking *et al*. 2010). These precautions were taken to allow for a possible minor genotyping error within a stretch of truly homozygous SNPs or other sources of artificial heterozygosity, such as paralogous sequences. A SNP was counted as a part of a ROH if >5% windows spanning it were homozygous. These parameters were selected to minimise the probability of a window being called homozygous by chance, while ensuring that SNPs on the edge of a true ROH will be counted as a part of that ROH (Purcell *et al.* 2007).

A threshold was set as the minimum length (kb) needed for a tract to qualify as homozygous. The existence of LD blocks in DNA means that relatively short ROHs (those spanning from tens to hundreds of kb) are very prevalent across the genome (Abecasis *et al.* 2005; Wall and Pritchard, 2003; Abecasis *et al.* 2001; Reich *et al.* 2001). In order to exclude these very common short tracts of homozygous SNPs, the minimum length for a ROH was set at 1Mb. The selection of this threshold was used by several other studies (Lencz *et al.* 2007a; Nalls *et al*. 2009). Studies have identified only very few long stretches of LD, measuring up to several hundred kb in length could result in longer ROHs in outbred populations.

The longer the homozygous region is (~1Mb), the lower the probability that the SNPs tagging that region are homozygous by simply chance. This probability is the function of both the numbers of SNPs located on the segment and their degree of informativeness, which is expressed in terms of heterozygosity level in the population (Carr *et al.* 2009). To ensure that the analysis captures only regions that are entirely homozygous between the first and the last SNP a threshold for the minimum number of SNPs constituting a ROH was selected. In line with the previous studies on homozygosity of complex disorders, the minimum number of homozygous SNPs to qualify as a ROH in this project was set at 100 (Lencz *et al*. 2007a).

This robust size and SNP density thresholds for inclusion into ROHs allows for the algorithmic exclusion of copy number variants, centromeric and SNP-poor regions (Nalls *et al.* 2009).

Two additional parameters were added to ensure that estimates were not artificially inflated by apparently homozygous tracts in sparsely covered genomic regions. First, required minimum SNP density was defined to 50 so that 1 SNP had to be present per at

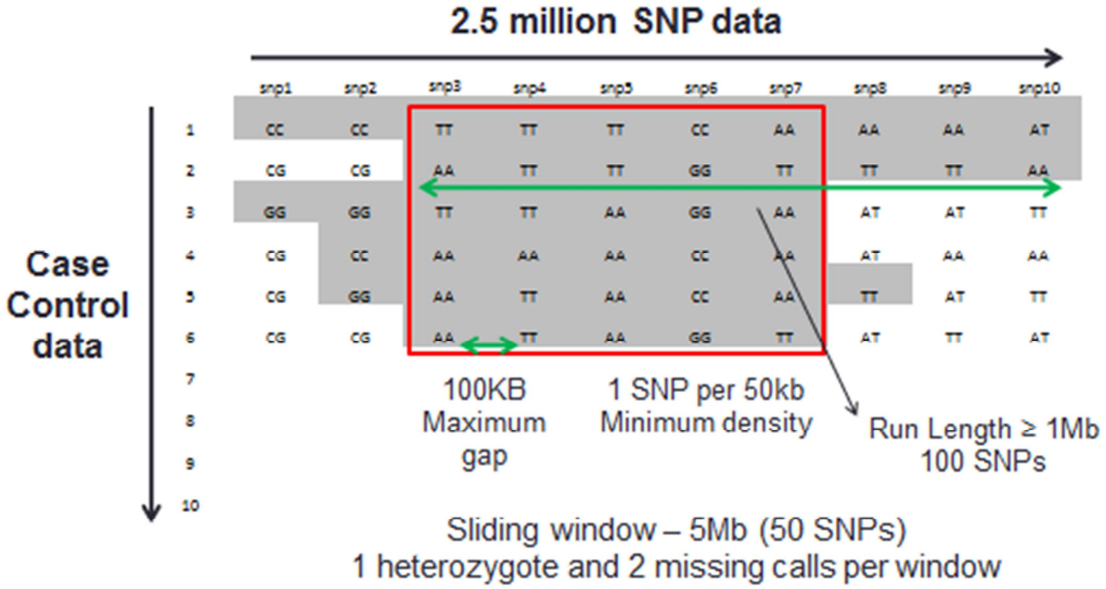
least 50kb of DNA and the maximum gap between two consecutive homozygous SNPs was set at 100kb. These thresholds were also used before in other studies (Nalls *et al*. 2009; Hosking *et al*. 2010; Vine *et al*. 2009; Yang *et al*. 2010).

In summary, the following parameters were used in identification of ROHs in all cohorts (Figure 2.2.1):

- Sliding window of 5000kb with 50 SNPs
- Minimum length of ROHs - 1Mb
- Minimum SNP number – 100
- Maximum gap between two consecutive homozygous SNPs – 100kb
- Minimum SNP density – 50kb
- Allowance for 1 heterozygote and 2 missing calls

Identification of ROHs based on this set of criteria was conducted in each population separately.

Figure 2.2.1: Identification of ROHs in PLINK software. A snapshot of the sliding window approach moving across the chromosome to identify ROHs in consecutive SNP genotype data (horizontally) of individuals (in vertical order) is shown. In grey are ROHs across different individuals. The big green double arrow indicates a at least 1Mb ROH in length, the small green double arrow refers to the maximum gap between two consecutive SNPs and the red window represents the consensus overlapping region across individuals.



2.2.5. Definition of overlapping chromosomal regions with homozygous SNPs

The genomic region spanned by all the ROHs in a certain pool (in at least 5 individuals) was used to define an overlapping ROH. Overlapping ROHs were separated into groups and the number of cases and controls carrying them were identified using the homozyg-group option in PLINK.

Figure 2.2.2: Schematic diagram illustrating an overlapping region. Blue lines represent several individual ROHs ≥1Mb and the red rectangle illustrates the overlapping region across all the individual ROHs in the group [Taken from Ku *et al.* 2011]

2.2.6. Calculation of homozygosity measures

The following measures of homozygosity were calculated in each study (a) the average (range) number of ROHs, (b) the total and average length of ROHs (Nalls *et al.* 2009) (c) proportion of autosomal genome in ROHs above a specific length threshold (FROH) (McQuillan *et al.* 2012). Subjects without any ROHs on a chromosome were also counted.

The total number of ROHs was defined as the sum of all ROHs per individual. The average ROH number was calculated by dividing the total number of runs by the total number of subjects. The average total ROH length is the sum of the length of each individual ROH per participant and was calculated by dividing the total ROH length by the number of individuals having ROHs [for example chr1: 4864 (total number of individuals in WTCCC study) minus 73 (subjects without any ROHs) = 4791; 22765.10 (sum of length of each individual ROH per subject on chr1) (22765.10/4791=4751.64kb is the average total ROH length per individual on chr1)]. The average run length was calculated by dividing the total genomic length of the ROHs by the total number of ROHs per participant [for example chr1: 22765.10 (sum of length of each individual ROH per participant); 17,110 (total number of ROHs on chr1) (22765.10/17,110=1.3Mb)].

FROH was defined as the percentage of the typed autosomal genome in ROHs greater than or equal to 1Mb in length. FROH is optimal for inferring the degree of genome-wide autozygosity and for detecting inbreeding effects (McQuillan *et al.* 2012). However, given the small variation in genome-wide FROH in unrelated individuals, large sample sizes are necessary to detect inbreeding depression for likely effect sizes in outbred samples (Keller *et al.* 2011).

To calculate the proportion of the autosomal genome covered in ROHs, percentage of homozygosity was calculated by summing ROHs>1Mb across the covered autosomal genome and dividing by the total autosomal base pairs represented on the microarray platform. As a result the summed length of identified ROHs was then divided by a factor of 2772.7 and subsequently converted to a percent by multiplying the dividend by 100. A factor of 2772.7 is the number of megabases covered by SNPs included in the array used in the genome-wide dataset. This estimate was calculated by summing the distance between the first and the last available SNP of each chromosomal arm for each of the 22 autosomes.

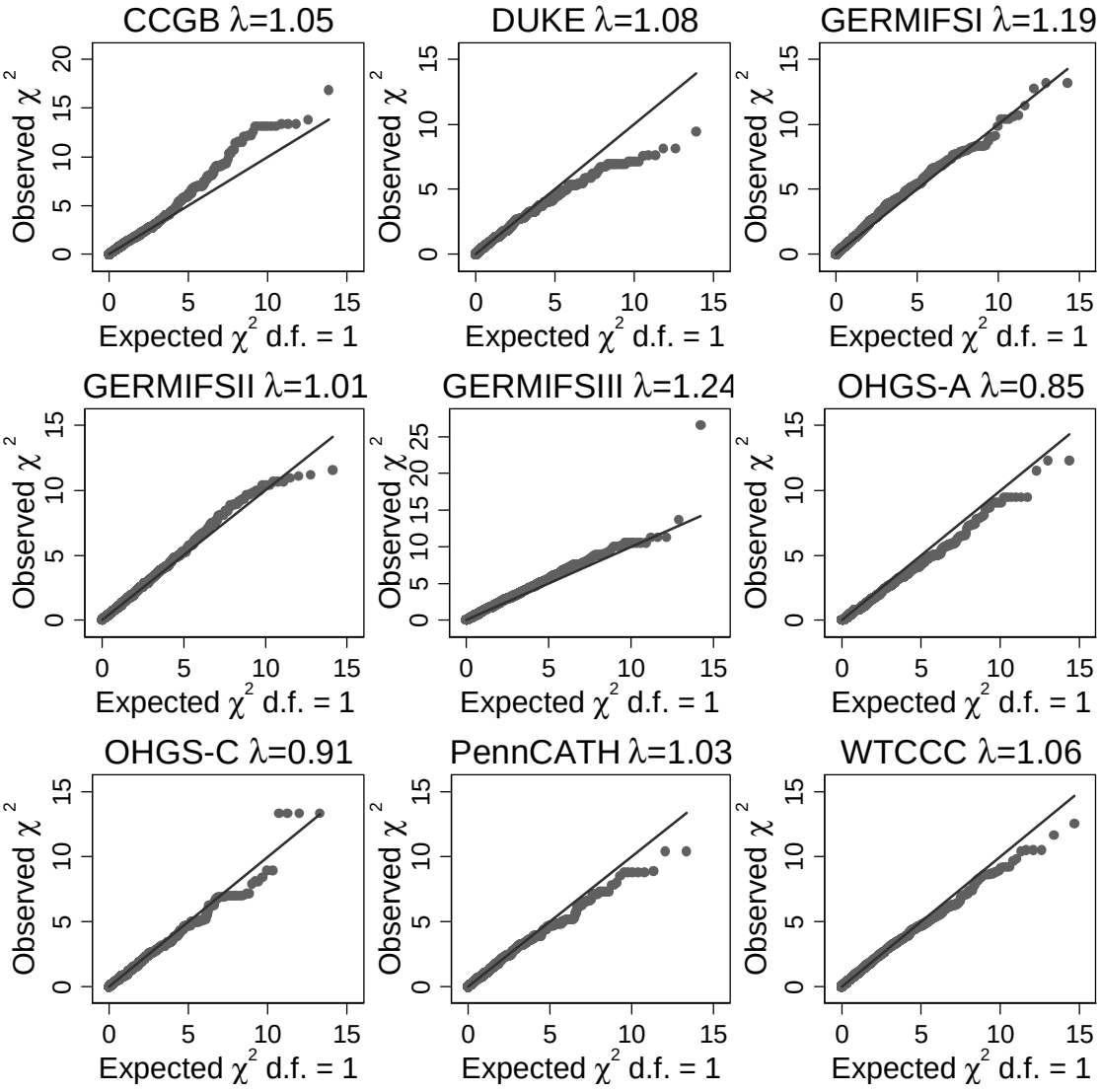
A genome-wide meta-analysis of homozygosity measures in all 9 populations from CARDIoGRAM Consortium was conducted. Differences in the homozygosity measures across populations were examined with ANOVA test. The distributions of homozygosity measures were examined using Kernel density estimates. Correlation between total ROH length and ROH number was examined using Pearson's coefficient (r). All the statistical analysis was undertaken in STATA v12.

Association between measures of homozygosity and CAD was conducted using linear regression adjusted for cohort differences. The beta coefficient/odds ratio and the p-values reflect the magnitude of the effect and statistical significance for the subjects from all cohorts.

2.2.7. Quality controls before the association analysis between overlapping consensus regions and CAD

QQ plots were drawn to compare the distribution of observed chi^2-values to the expected distribution under the null hypothesis of no association to ensure the good quality of the data (Figure 2.2.3). Genomic inflation lambda (λ) value was also estimated for each study. In GERMIFS -I and -III studies λ were slightly inflated in contrast to OHGS –A and –C slightly deflated.

Figure 2.2.3: Quantile-quantile plots of chi^2-values for all 15,441 identified overlapping consensus regions in at least two studies from CARDIoGRAM Consortium



2.2.8. Association analysis between overlapping consensus regions and CAD

Meta-analysis of all individual study ROH associations was conducted using inverse variance weighting in STATA v12. No covariates were taken into account. Heterogeneity was calculated using I^2 statistics. Bonferroni correction was used to correct for multiple testing.

Binomial test was used to examine if there was a deviation from the expected distribution (50%/50%) between the CAD protective (OR<1) and CAD detrimental (OR>1) consensus regions of overlapping ROHs in CARDIoGRAM Consortium. OR>1 reflected a consensus region of overlapping ROHs that was over-represented in CAD cases and a OR<1 reflected a consensus region of overlapping ROHs that was over-represented in controls.

A Cochran Armitage trend test was used to check the direction of the observed difference across CAD detrimental and CAD protective consensus regions of overlapping ROHs.

2.3. Results

2.3.1. Characteristics of study cohorts

A total of 20,821 individuals of European ancestry from 9 populations of CARDIoGRAM Consortium were included. The key characteristics of the cohorts used in the homozygosity analysis are summarized in Table 2.3.1. There was a fair numerical balance between patients with CAD and CAD-free controls included in the analysis. Consistent with the recognised sexual dimorphism in CAD prevalence, a majority of patients with CAD were male. In populations recruited in Canada, controls were older than patients with CAD, in an effort to ensure that controls were old enough without developing CAD.

Table 2.3.1: Characteristics of CARDIoGRAM populations

Study	Number of subjects	Cases/Controls	%MI	%female (cases/controls)	Cases age (years)	Controls age (years)
CCGB	1996	1628/368	60.3	24.4/46.2	48.6±7.3	73.82±5.2
DUKE	1848	1200/648	48.1	30.6/58.0	56.6±9.7	63.3±8.7
GERMIFSI	2488	884/1604	100	49.4/50.8	50.2±7.8	62.6±10.0
GERMIFSII	2509	1222/1287	100	33.1/48.3	51.4±7.5	51.2±11.9
GERMIFSIII	2905	1157/1748	100	20.1/48.9	58.6±8.7	55.9±10.7
OHGS-A	1955	947/1008	64.3	21.9/45.5	48.1±7.0	74.9±4.9
OHGS-C	1161	843/318	44.3	5.9/64.8	56.1±6.9	76.0±6.3
PennCATH	1084	732/352	NA	NA	NA	NA
WTCCC	4864	1926/2938	71.5	20.7/50.0	49.8±7.7	-*
In total	20,821	10,548/10,273	-	-	-	-

Data are counts and percentages or means and standard deviations, MI - myocardial infarction; n - sample size. *WTCCC controls included an equal number of subjects from the 1958 Birth Cohort and from the National Blood Service donors. The latter were recruited in equal 10 years age bands from 11 to 70 years of age.

2.3.2. Analysis of the overall genetic architecture of ROHs

A meta-analysis of all 20,821 subjects from CARDIoGRAM was first undertaken to explore the global genetic architecture of ROHs in human DNA. The summary of homozygosity measures from this analysis is shown in Table 2.3.2.

Table 2.3.2: Genome-wide measures of homozygosity in CARDIoGRAM Consortium

No	Homozygosity measure	Values per genome
1	Average number of ROHs	32.14±8.72
2	Number of ROHs – range	2-260
3	Average total length of ROHs (Mb)	44.05±14.99
4	FROH (%)	1.59±0.54
5	Average ROH length (kb)	1370.68±488.98
6	Length of ROHs – range (Mb)	1.00-29.40

ROHs – runs of homozygosity, FROH – proportion of autosomal genome in ROHs, Data are means and standard deviations or absolute values

This analysis revealed that each individual of white European ancestry has on average 32.14±8.72 ROHs in their DNA. These stretches of homozygous SNPs have an average length of 1370.68±488.98 kb and they cover on average a total length of 44.05±14.99 Mb (1.59% of the human genome). The number and length of ROHs per individual ranged from 2-260 and 1-29.40 Mb, (respectively) in the overall sample.

Homozygosity measures were also examined in each population, separately (Table 2.3.3). A total of 669,313 ROHs ranging in size from 1Mb to 29.40 Mb and containing >100 consecutive homozygous SNPs, were identified across 20,821 genomes from 9 studies (Figure 2.3.1).

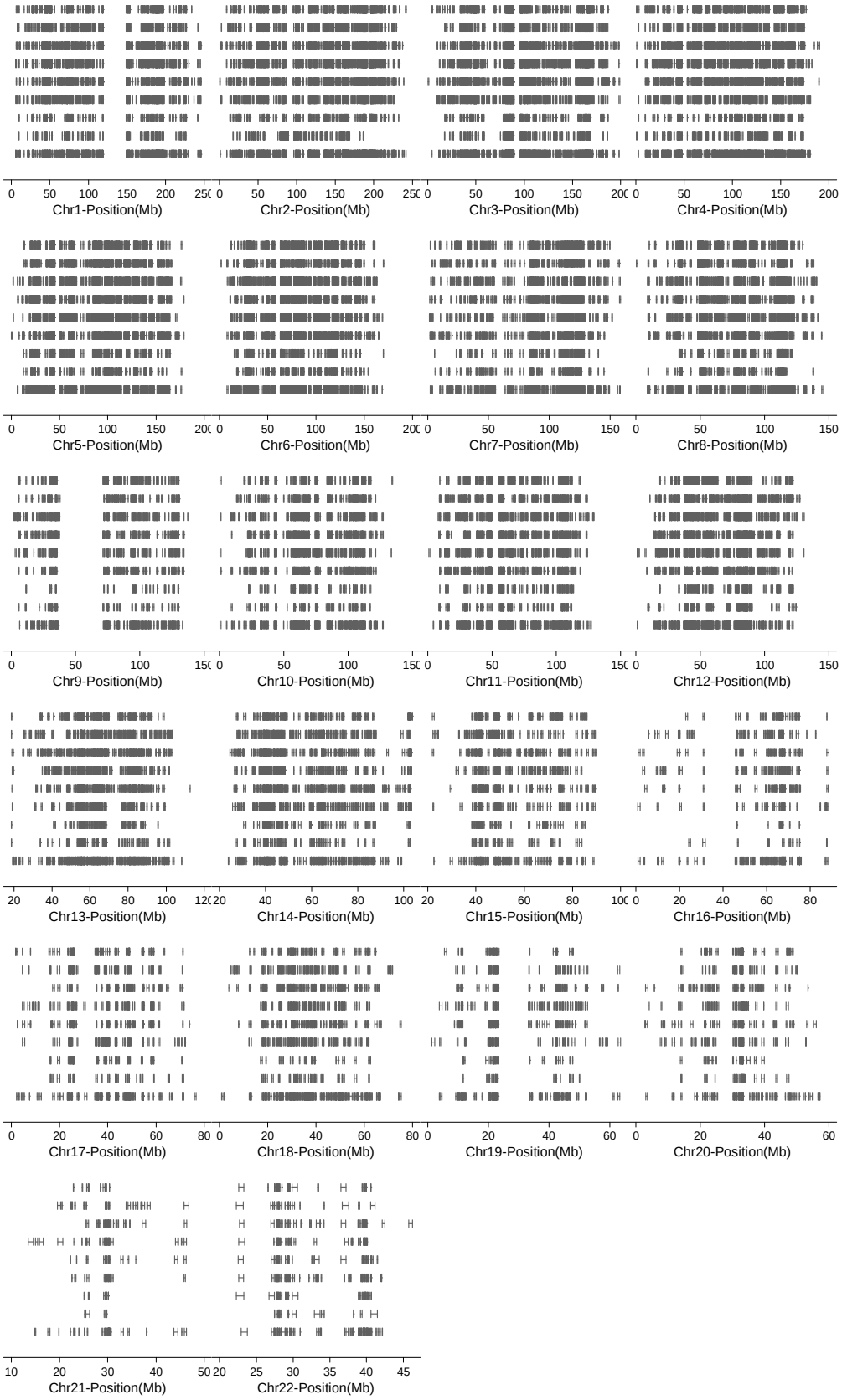
The WTCCC has the largest proportion of the genome covered by ROHs (FROH =1.85%) and PennCATH the smallest (FROH=1.23%). The longest identified ROH was identified in PennCATH (29.40Mb).

Table 2.3.3: Summary of homozygosity measures in 9 populations from CARDIoGRAM Consortium.

Study	Number of subjects	Average run number (n)	ROHs range	Average total run length (kb)	Average run length (kb)	Longest run length (Mb)	% of genome in ROHs (FROH)
CCGB	1996	29.50±6.35	2-83	40199.32±10616.25	1362.62±445.57	10.91	1.45±0.38
DUKE	1869	25.51±10.13	8-260	35335.86±19390.20	1385.15±490.67	10.00	1.27±0.70
GERMIFSI	2486	36.35±7.40	11-145	49924.54±15203.16	1373.59±509.82	18.39	1.80±0.55
GERMIFSII	2506	28.48±6.09	12-81	39399.01±11876.63	1383.16±535.41	15.33	1.42±0.43
GERMIFSIII	2900	31.79±8.12	4-180	43538.19±14700.81	1369.58±485.01	13.61	1.57±0.53
OHGS-A	1955	35.43±8.99	6-158	48904.37±17289.22	1380.22±480.23	11.23	1.76±0.62
OHGS-C	1161	25.22±6.35	9-72	34430.34±11169.43	1365.41±465.48	10.17	1.24±0.40
PennCATH	1084	24.32±5.74	7-47	34093.18±10505.12	1401.59±693.27	29.40	1.23±0.38
WTCCC	4864	37.78±6.09	19-88	51250.29±10376.44	1356.58±443.29	24.14	1.85±0.37

Data are means and standard deviations or absolute values and percentages.

Figure 2.3.1: Genome-wide distribution of ROHs. Each row represents a population. Populations are ordered from top to bottom by alphabetical order (CCGB, DUKE, GERMIFSI, GERMIFSI, GERMIFSII, GERMIFSIII, OHGS-A, OHGS-C, PennCATH, WTCCC). The bars represent ROHs.



The average number of ROHs in the autosomal genome is shown in Figure 2.3.2. Overall, there are significant differences in average number of ROHs amongst nine populations ($P<1x10^{-300}$). Individuals from WTCCC cohort have the largest average ROH number (37.78) and those from PennCATH the smallest (24.32) (Table 2.3.3). Approximatley, 80% (16,620) of individuals have 21-40 ROHs in their autosomal genome. Of these, a majority (43.8% - 9,120) belong to 31-40 ROHs category. A fraction of individuals (36.0% - 7500) have on average 21-30 ROHs. Less common are individuals with 41-50 ROHs (12% - 2489). Only 1.5% of subjects have more than 50 ROHs and just 0.06% (13) of analysed individuals have more than 100 ROHs in their genome. One individual from DUKE study has 260 ROHs in the DNA.

The average ROH length per participant in each study is shown in Figure 2.3.2. A similar general pattern of distribution is observed across all populations ($P=2.5x10^{-27}$). The average ROH length varies between 1362.62kb in CCGB and 1401.59kb in PennCATH.

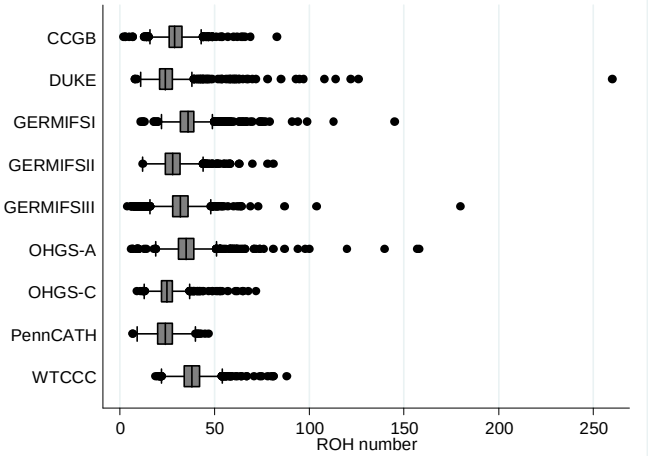
The average total ROH length per participant in each study is presented in Figure 2.3.2. There are statistically significant differences between the populations ($P<1x10^{-300}$). The total ROH length varied between 34.09Mb in PennCATH and 51.25Mb in WTCCC.

Outliers for all measures of homozygosity exist almost in every population. They appear on both sides of homozygosity measures distribution. For example, an individual in CCGB had only 2 ROHs covering just 2.31Mb of the autosomal genome whilst in a different individual from DUKE 260 ROHs covering 476.57Mb were identified (Figure 2.3.2).

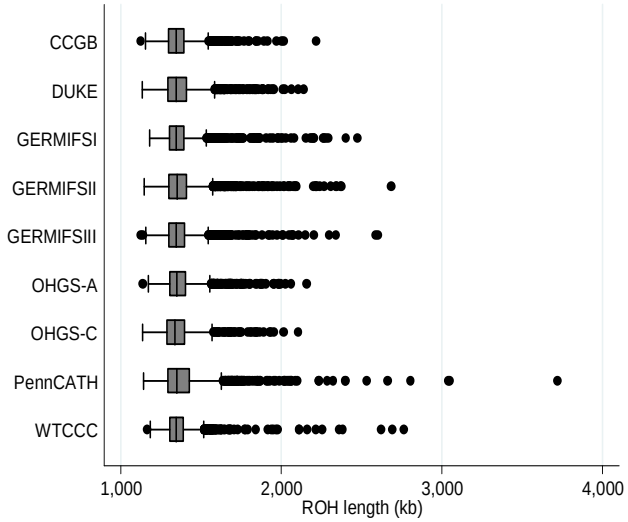
To visualise the shape and range of homozygosity in each population in the context of data collected from other cohorts, the distribution of each homozygosity measure was plotted using Kernell density (Figure 2.3.3). The distribution of each homozygosity measure is fairly symmetrical in each cohort with only one peak observed. There is trend to a positive skewness but the sample size is large enough to assume normality.

Figure 2.3.2: Measures of homozygosity in each cohort from CARDIoGRAM Consortium. A - average number of ROHs, B - average length of ROHs, C - average total length of ROHs in the autosomal genome. The boxes are median values, the upper edge of the box - the 75th percentile, the lower edge - the 25th percentile of ROH distribution. Outliers are represented by black dots.

A.



B.



C.

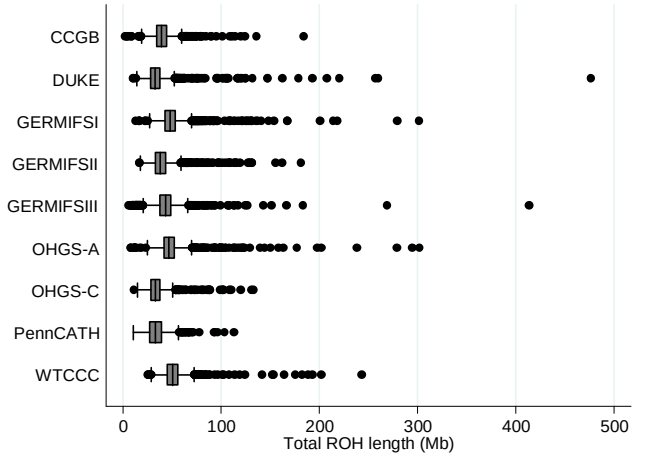
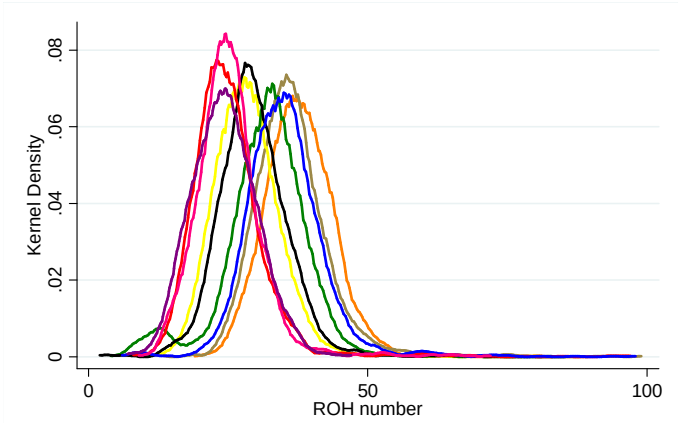
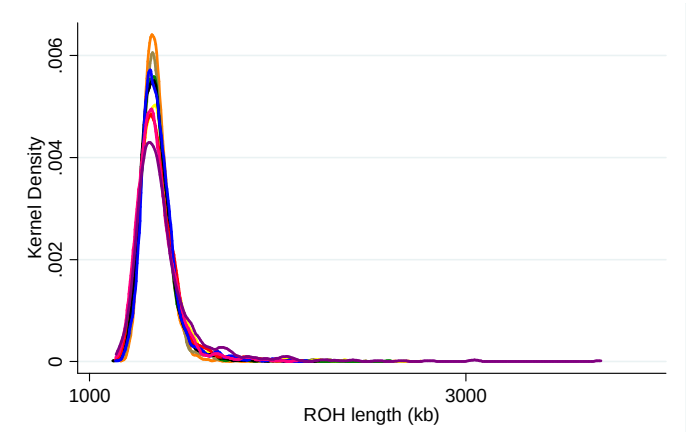


Figure 2.3.3: Kernel density estimates of the homozygosity measures distributions in each population. A - average number of ROHs, B - average length of ROHs, C - average total length of ROHs in the autosomal genome. Each coloured line represents the data distribution from a different population; CCGB - black, DUKE - red, GERMIFSI - brown, GERMIFSII - yellow, GERMIFSIII - green, OHGS-A - blue, OHGS-C - pink, PennCATH - purple, WTCCC - orange. In Panel A only individuals with less than 100 ROHs are shown.

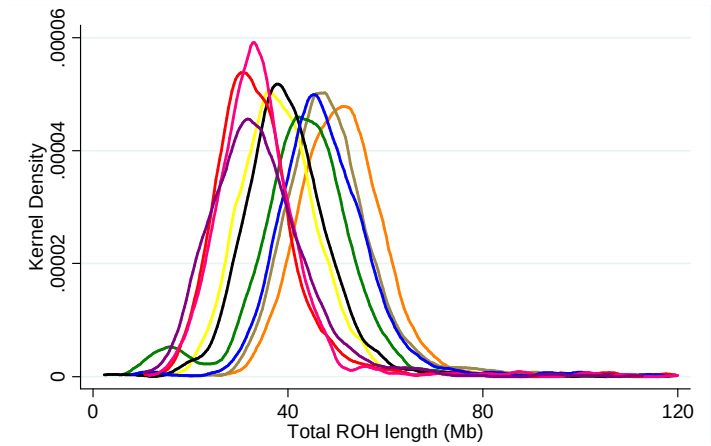
A.



B.



C.



The results of chromosome-specific homozygosity measures analysis are shown in Table 2.3.4.

Table 2.3.4: Chromosome-specific measures of homozygosity in CARDIoGRAM Consortium

Chr	Average number of ROHs	Number of ROHs – range	Average total length of ROHs (kb)	Average ROH length (kb)	Subjects without any ROHs (%)
1	2.90±1.70	0-30	3878.97±2661.08	1337.22±415.04	956 (4.6)
2	3.05±1.78	0-21	4095.44±2730.44	1342.90±457.26	911 (4.4)
3	2.85±1.75	0-33	3948.04±2815.81	1384.03±513.12	1243 (6.0)
4	2.33±1.52	0-27	3159.04±2502.63	1354.29±471.64	1558 (7.5)
5	2.45±1.61	0-27	3424.43±2683.98	1396.06±536.78	1688 (8.1)
6	1.89±1.48	0-30	2646.13±2525.11	1401.99±596.57	3223 (15.5)
7	1.70±1.33	0-19	2308.23±2119.13	1357.09±491.95	3677 (17.7)
8	2.95±1.72	0-28	3990.22±2560.16	1353.39±408.10	1074 (5.2)
9	0.60±0.88	0-19	770.97±1387.99	1283.40±463.77	11648 (55.9)
10	1.64±1.20	0-19	2284.78±1936.97	1392.04±456.28	3138 (15.1)
11	1.48±1.30	0-29	2308.14±2345.26	1563.76±740.05	4939 (23.7)
12	1.85±1.43	0-28	2531.41±2335.44	1370.04±459.67	3247 (15.6)
13	0.68±0.93	0-23	899.17±1667.21	1327.29±568.45	10702 (51.4)
14	1.19±0.97	0-18	1585.64±1541.10	1332.95±378.61	4621 (22.2)
15	1.13±0.97	0-15	1591.92±1540.12	1413.86±471.86	5733 (27.5)
16	1.00±0.86	0-14	1449.76±1370.11	1446.96±401.04	5907 (28.4)
17	0.71±0.84	0-10	896.37±1187.85	1267.56±379.13	10074 (48.4)
18	0.32±0.66	0-20	416.36±1106.69	1309.32±508.11	15320 (73.6)
19	0.41±0.65	0-12	539.77±952.68	1306.80±381.88	13677 (65.7)
20	0.57±0.73	0-12	750.85±1102.83	1324.52±450.12	11373 (54.6)
21	0.11±0.35	0-8	136.38±526.78	1293.52±478.79	18788 (90.2)
22	0.34±0.58	0-11	440.04±832.68	1284.85±361.73	14600 (70.1)

Data are means and standard deviations or counts and percentages, Chr - chromosome

ROHs are widely distributed across the entire genome and are present on each human autosome. The number of ROHs is a function of chromosomal length. Indeed, larger chromosomes tend to have on average more ROHs. Average number of ROHs and average total length of ROHs depend on chromosome size. For example each individual has on average 3.05 ROHs covering ~4.1Mb of DNA on chromosome 2 (the largest chromosome) compared to just 0.11 ROHs covering only 0.1Mb of DNA on chromosome 21 (the shortest chromosome). There are individuals that do not have any

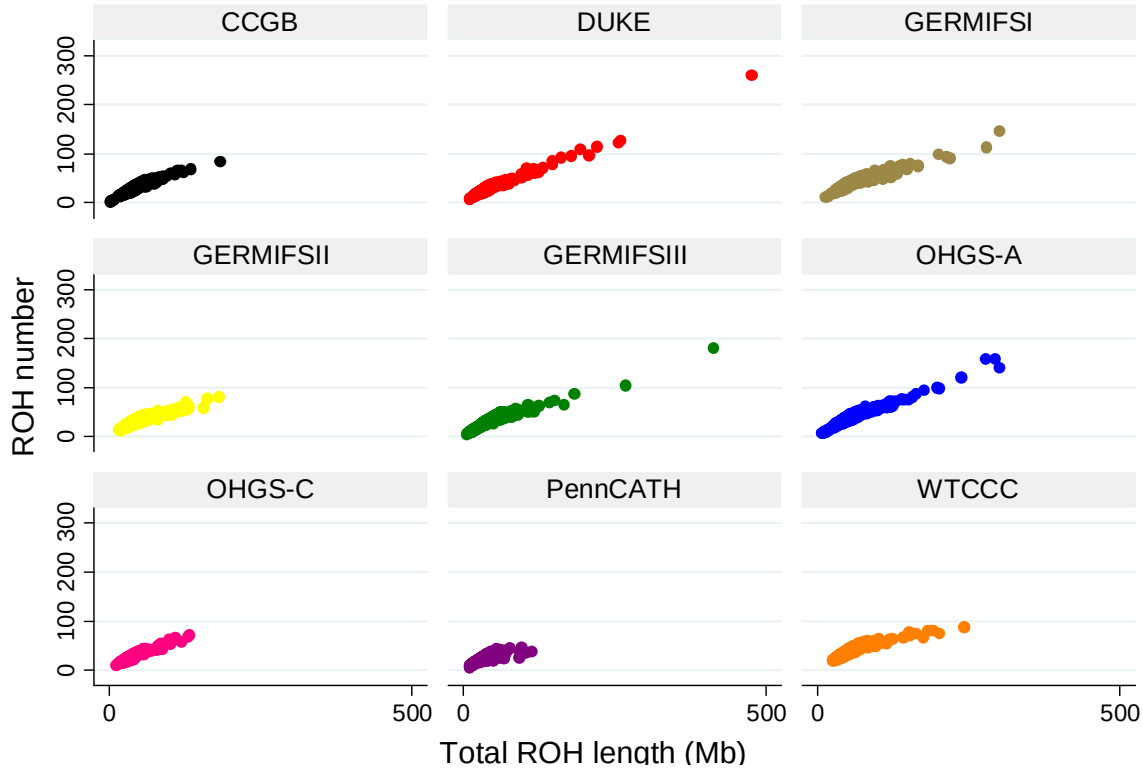
ROHs in some of their chromosomes. Again, the number of these individuals was related to the size of the chromosome. For example, only 4.4% subjects do not contain any ROHs on chromosome 2 compared to 90.2% who have no ROHs on chromosome 21.

In contrast, average ROH length does not depend on chromosomal length. The average length of ROHs on chromosome 11, chromosome 15 and chromosome 16 (shorter chromosomes) is greater than that on chromosome 1 (the second longest chromosome). The shortest average ROH length (1267.56kb) was observed on chromosome 17 and the longest (1563.76kb) on chromosome 11.

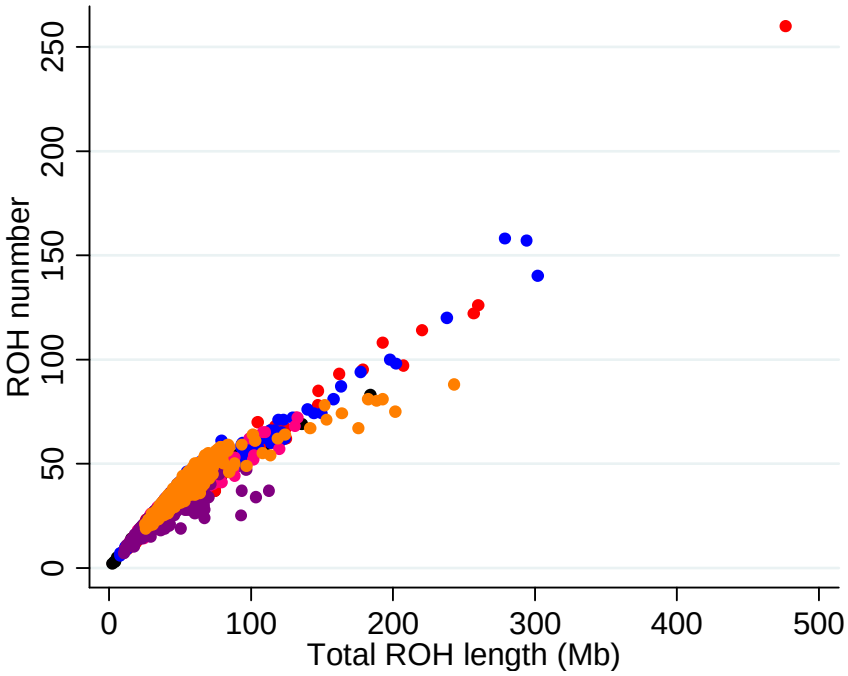
The analysis of association between ROH number and total ROH length showed that they were in strong linear correlation in each population analysed separately (r=0.86 PennCATH to r=0.98 DUKE) and in the combined analysis (r=0.94) (Figure 2.3.4). Indeed, the higher number of ROHs in the autosomal genome, the longer area of DNA is covered by segments of homozygous SNPs. Only few individuals have many long ROHs and cover several hundred of Mb in their autosomal genome.

Figure 2.3.4: Correlation between total ROH length and ROH number in CARDIoGRAM Consortium. A - The number of ROHs at least 1Mb or longer in length plotted against the total length of ROHs, per individual, for each population separately, B - Combined linear correlation of population specific data from panel A.

A.



B.



2.3.3. Analysis of association between CAD and homozygosity measures

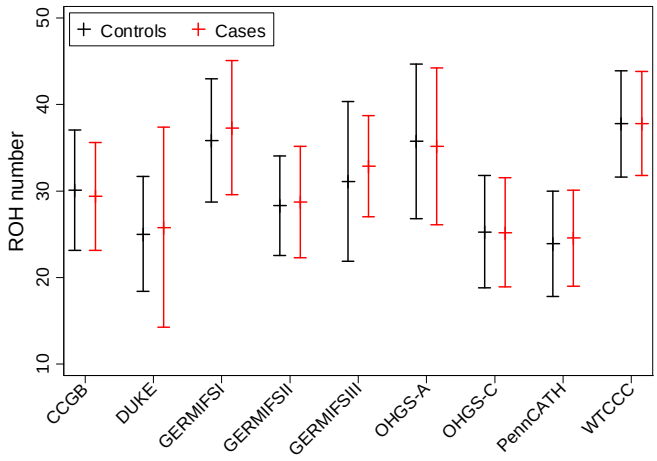
The characteristics of global architecture of homozygosity were followed by comparative analysis of average ROH number, average ROH length and average total ROH length between 10,548 CAD cases and 10,273 healthy controls.

The mean values for average ROH number, average ROH length and average total ROH length for cases and controls in each examined population are shown in Figure 2.3.5 and Table 2.3.5. As a general trend, the mean values (together with 95% confidence intervals) were comparable between cases and controls from the same populations. Of 3 measures of homozygosity, average ROH length showed least variation between populations. As expected from the previous data on global genetic architecture of homozygosity, both average ROH number, and average total ROH length showed significant differences between populations within both cases and controls stratum.

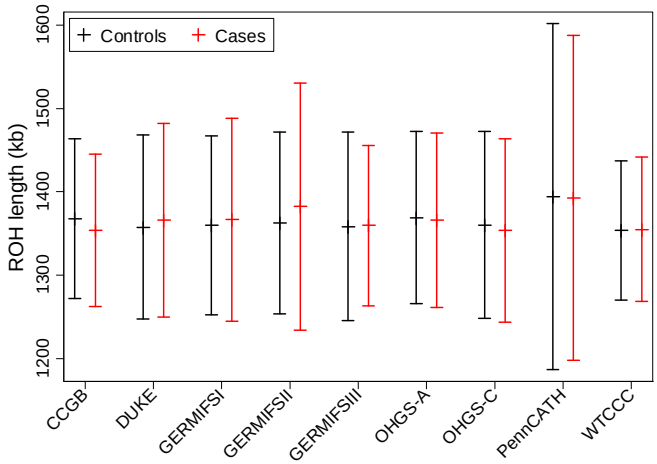
After adjustment for study origin, there were significant differences in overall homozygosity between patients with CAD and CAD-free controls (Table 2.3.6). On average, patients had 0.5 ROH more than controls (β=0.5, 95% CI: 0.30-0.71, $P=2.06 \times 10^{-6}$). The logistic regression revealed that each ROH in the autosomal genome increased CAD risk by approximately 1% (OR=1.01, 95% CI: 1.006-1.014, $P=2.57 \times 10^{6}$). The average total length of ROHs in autosomal genome of patients with CAD was approximately 816 kb longer compared to controls ($P=3.6 \times 10^{-5}$) (Table 2.3.6).

Figure 2.3.5: Comparison of homozygosity measures in CAD cases and controls. A - average number of ROHs, B - average length of ROHs, C - average total length of ROHs in the autosomal genome. Data are means with 95% confidence intervals.

A.



B.



C.

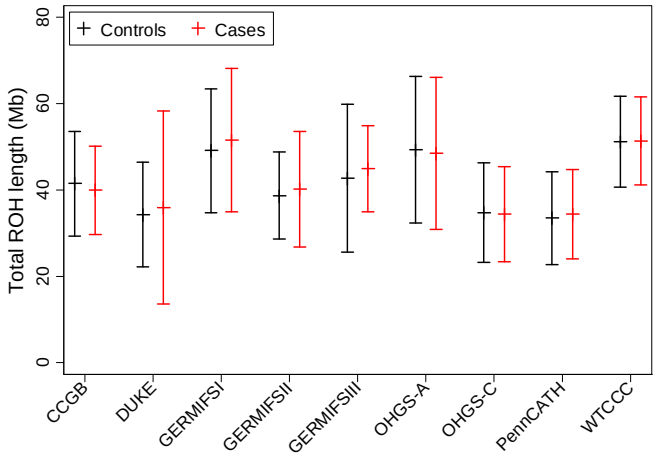


Table 2.3.5: Differences in homozygosity measures between CAD cases and CAD-free controls in 9 populations from CARDIoGRAM Consortium

Study	A. Average number of ROHs			B. Average length of ROHs			C. Average total length of ROHs		
	Controls	Cases	P-value	Controls	Cases	P-value	Controls	Cases	P-value
CCGB	30.08±6.92	29.37±6.21	0.05	1367.71±95.69	1354.05±91.57	0.01	41420.13±12101.32	39923.37±10234.78	0.01
DUKE	25.02±6.68	25.78±11.56	0.12	1357.74±110.10	1366.00±115.95	0.14	34290.94±12130.53	35896.99±22320.03	0.09
GERMIFSI	35.81±7.12	37.31±7.79	$1.24x10^{-6}$	1359.89±107.31	1366.68±121.20	0.15	49051.46±14271.41	51512.31±16657.01	$1.0x10^{-4}$
GERMIFSII	28.26±5.76	28.72±6.41	0.06	1362.52±108.88	1382.54±4.22	$1.0x10^{-4}$	38689.01±10157.20	40145.04±13412.38	$2.1x10^{-3}$
GERMIFSIII	31.08±9.25	32.87±5.83	$5.03x10^{-9}$	1358.58±112.68	1359.86±96.13	0.75	42678.64±17058.93	44842.44±9990.16	$1.0x10^{-4}$
OHGS-A	35.72±8.92	35.13±9.06	0.15	1368.89±103.11	1366.43±104.26	0.60	49309.82±16917.36	48472.79±17675.18	0.28
OHGS-C	25.27±6.49	25.20±6.30	0.87	1360.20±111.96	1353.93±109.88	0.39	34671.03±11550.06	34339.55±11028.07	0.65
PennCATH	23.87±6.11	24.54±5.54	0.07	1394.42±207.10	1392.76±194.21	0.90	33463.15±10822.82	34396.15±10342.72	0.17
WTCCC	37.77±6.15	37.80±5.99	0.86	1353.73±83.42	1354.92±86.51	0.63	51222.46±10502.53	51292.75±10183.65	0.82

Data are means and standard deviations, ROHs – runs of homozygosity, P-value – level of statistical significance for a difference between cases and controls

Table 2.3.6: Differences in homozygosity measures between CAD cases and controls from CARDIoGRAM Consortium

Measure	β-coefficient/OR	LCI	UCI	P-value
Average ROH number	0.50	0.30	0.71	2.06×10^{-6}
Average ROH length (kb)	3.17	-0.01	6.36	0.05
Average total length of ROHs (kb)	816.49	429.04	1203.95	3.6×10^{-5}
FROH*	1.20	1.09	1.32	1.15×10^{-4}

Data are either β-coefficients or odds ratios (OR) with respective confidence intervals and level of statistical significance from regressing homozygosity measures on case-control status after adjustment for cohort. 95% LCI – lower confidence interval, 95% UCI – upper confidence interval, FROH – proportion of autosomal genome in ROHs, * – analysis was conducted in 20,503 individuals (318 excluded because of FROH>3%).

Examination of FROH distribution revealed that it was highly right-skewed with maximal value of 17.19%. Therefore, 318 individuals with FROH>3% were excluded to facilitate regression analysis. Logistic regression analysis revealed that each 1% increase in FROH translated into 20% increase in odds of CAD risk (OR=1.20, 95% CI: 1.09-1.32, $P=1.15 \times 10^{-4}$). The sensitivity analysis conducted with inclusion of these 318 showed no effect on the case-control difference.

Due to large age differences across CAD cases and controls in the examined populations (Table 2.3.1) and a possible age effect, the comparative analysis of homozygosity measures was repeated with adjustement of age as a covariate. Of, the 20,821 individuals 99 didn't have age data and were excluded from the analysis. After adjustment for age, there were even stronger significant differences in all homozygosity measures across CAD patients and CAD-free controls (Table 2.3.7).

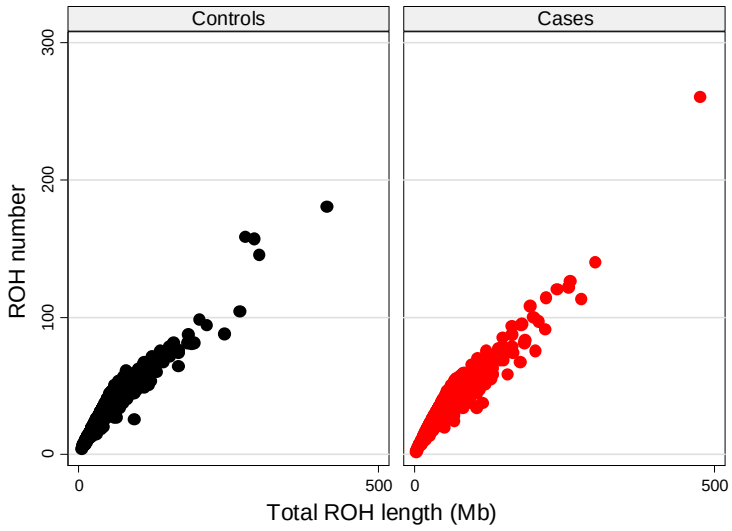
Table 2.3.7: Differences in homozygosity measures between CAD cases and controls from CARDIoGRAM Consortium - age adjustment

Measure	β-coefficient/OR	LCI	UCI	P-value
Average ROH number	0.72	0.50	0.93	9.32×10^{-11}
Average ROH length (kb)	5.16	1.84	8.47	0.002
Average total length of ROHs (kb)	1162.04	758.39	1565.68	1.70×10^{-8}
FROH*	1.30	1.18	1.43	8.38×10^{-8}

Data are either β-coefficients or odds ratios (OR) with respective confidence intervals and level of statistical significance from regressing homozygosity measures on case-control status after adjustment for age and cohort. 95% LCI – lower confidence interval, 95% UCI – upper confidence interval, FROH – proportion of autosomal genome in ROHs, * – analysis was conducted in 20,347 individuals (375 excluded because of FROH>3%).

The analysis of relationship between ROH number and total ROH length conducted separately in cases and controls showed the same linear correlation in both groups (r=0.94) (Figure 2.3.6).

Figure 2.3.6: Association between ROH number and the total length of ROHs in CAD cases and CAD-free controls from CARDIoGRAM Consortium.



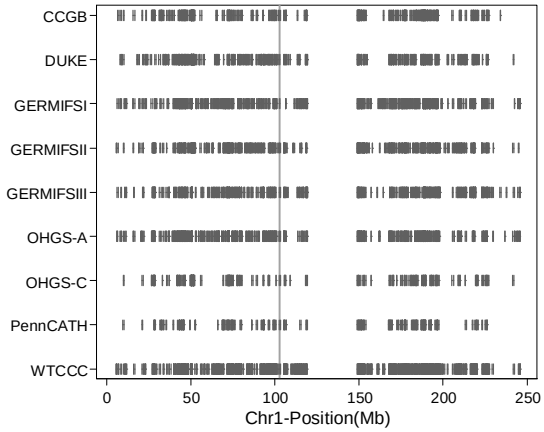
2.3.4. Analysis of association between CAD and overlapping chromosomal regions with homozygous SNPs

Analysis of overlapping chromosomal homozygous regions was undertaken in order to identify the ones that were not “private” (exclusive to one population) but present across the CARDIoGRAM cohorts. Identification of CAD-related regions of consecutive homozygous SNPs shared between populations was supposed to also narrow down the number of potential candidate genes and SNPs for further studies including in-depth sequencing.

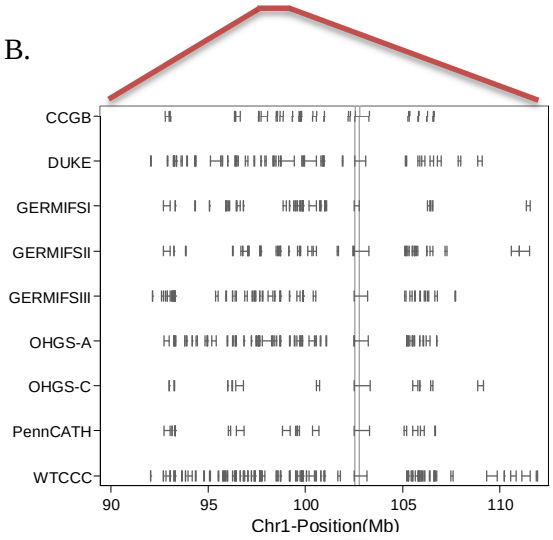
The analysis screened the autosomal genome for ROHs that overlapped in at least two of the examined populations. A total of 15,441 consensus regions of consecutive homozygous SNPs shared by at least 2 populations were identified in 22 autosomes. An example of such a homozygous region identified through analysis of overlapping ROHs on chromosome 1 is illustrated on Figure 2.3.7.

Figure 2.3.7: An example of consensus sequence of consecutive homozygous SNPs on chromosome 1 shared by 9 populations in analysis of overlapping ROHs. A - ROHs on chromosome 1, B - overlapping ROH mapping to approximately 102-104 MB on chromosome 1; double vertical lines indicate the window of consensus sequence of consecutive homozygous SNPs shared by 9 populations C – consensus region of consecutive homozygous SNPs marked by 2 vertical lines in the context of 5Mb sequence on chromosome 1. X axis - physical distance of the chromosome, Y axis lists each of 9 populations studied.

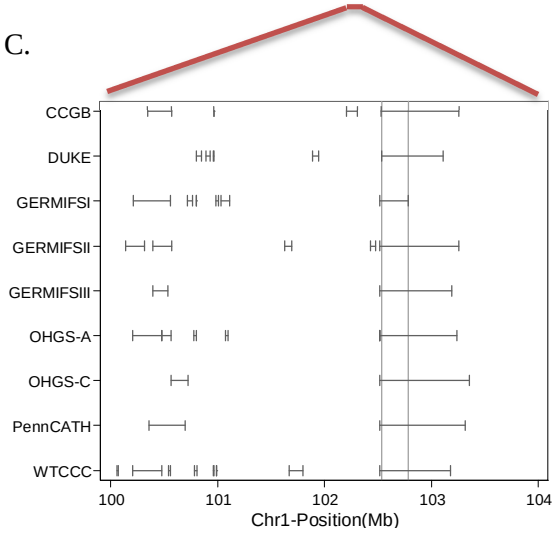
A.



B.



C.



To further characterise these segments of consecutive homozygous SNPs, they were classified into 4 groups (based on their SNP enrichment) and into 3 groups (based on their overall frequency in the meta-analysed sample) (Table 2.3.8). Approximately 2/3 (64.04%) of consensus regions were present in only 2-3 of the examined cohorts and only 8.27% mapped to overlapping ROHs in 6-9 cohorts. A majority (64.45%) of consensus regions were short – they contained 2-9 homozygous SNPs. A small percentage of the identified consensus regions (2.8%) contained more than 100 homozygous SNPs. Table 2.3.8 lists the total number of identified chromosomal regions with homozygous SNPs across 22 autosomes and indicates in how many populations they exist.

Table 2.3.8: Characteristics of consensus regions of consecutive homozygous SNPs identified through analysis of overlapping ROHs in CARDIoGRAM Consortium

No of SNPs in a region	**No of studies where consensus regions were identified**			**Total**
	2-3	4-5	6-9	
2-9	6403 (64.34%)	2821 (28.35%)	748 (7.32%)	9952 (64.45%)
10-49	2687 (63.85%)	1128 (26.81%)	393 (9.34%)	4208 (27.25%)
50-99	515 (60.73%)	232 (27.36%)	101 (11.91%)	848 (5.49%)
100+	284 (65.59%)	94 (21.71%)	55 (12.70%)	433 (2.8%)
Total	**9,889 (64.04%)**	**4,275 (27.69%)**	**1,277 (8.27%)**	**15,441**

Data are counts and percentages; percentages in column 2 and 3 are counted horizontally and vertically, respectively. SNP – single nucleotide polymorphism

Only 41 (0.27%) of consensus regions were shared across all 9 studies [in comparison with 5624 (36.42%) shared just 2 populations]. The shortest consensus region on chromosome 2 was just 2bp in length, contained 2 SNPs and was identified through overlapping ROHs from 4 studies. The longest consensus region on chromosome 7 was 920,570bp in length, had 118 consecutive homozygous SNPs and was shared by 4 populations. The region spanning the largest number (753) of SNPs was mapped to chromosome 6 in 4 studies.

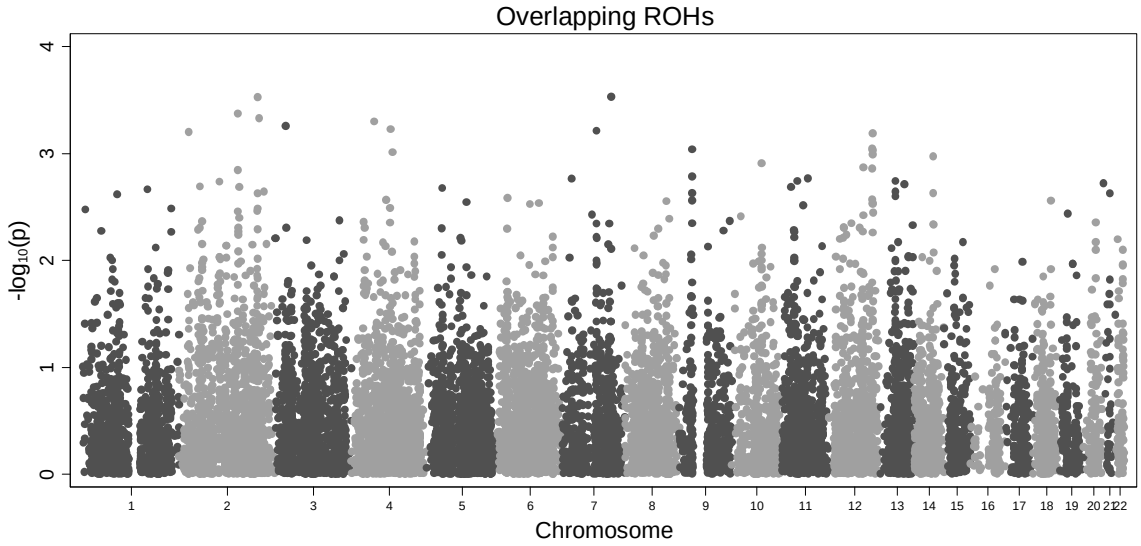
Table 2.3.9: Chromosome-stratified list of regions with consecutive homozygous SNPs identified through analysis of overlapping ROHs in CARDIoGRAM Consortium.

Chromosome	No of overlapping regions (%)	No of populations sharing consensus region		
		2-3	4-5	6-9
1	1094 (7.09)	725	278	91
2	1571 (10.17)	1038	433	100
3	1384 (8.96)	869	407	108
4	1406 (9.11)	938	366	102
5	1551 (10.04)	860	501	190
6	1402 (9.08)	854	412	136
7	818 (5.30)	513	226	79
8	983 (6.37)	643	267	73
9	422 (2.73)	288	102	32
10	572 (3.70)	373	148	51
11	793 (5.14)	506	236	51
12	887 (5.74)	583	235	69
13	635 (4.11)	410	176	49
14	501 (3.24)	320	143	38
15	306 (1.98)	208	72	26
16	165 (1.07)	121	33	11
17	178 (1.15)	113	58	7
18	329 (2.13)	220	86	23
19	167 (1.08)	120	40	7
20	145 (0.94)	100	27	18
21	44 (0.28)	31	5	8
22	88 (0.57)	56	24	8
Total	**15,441**	**9,889**	**4,275**	**1,277**

Data are counts and percentages; percentages in column 1 are counted horizontally

None of 15,441 consensus homozygous regions was associated with CAD after correction for multiple testing (Figure 2.3.8). Indeed, the corrected P-value (based on Bonferroni correction) was calculated at 3.2×10^{-6} and the most significant signal in this genome-wide analysis was calculated at 2.9×10^{-4}.

Figure 2.3.8: Association between CAD and consensus chromosomal regions of consecutive SNPs – genome-wide signal-intensity plot. The Y axis shows the logarithmic level of statistical significance for association of each individual region with CAD, X axis - 22 chromosomes in numerical order.



The top 20 most significant regions from this analysis are listed in Table 2.3.9. Their magnitude of association with CAD (odds ratio) varied between 0.28 and 3.39 whilst the nominal level of statistical significance for association with CAD was from 2.9×10^{-4} to 1.4×10^{-3}.

Table 2.3.10: Top 20 consensus regions of consecutive homozygous SNPs from analysis of association with CAD in CARDIoGRAM Consortium

Reg	Chr	BP1	BP2	Length (kb)	Studies	SNPs	Cases	Controls	OR (95% CI)	PValue	Genes
1	7	122,852,984	122,863,568	10.58	3	8	36 (1.00%)	45 (2.25%)	0.42 (0.25, 0.86)	$2.9x10^{-4}$	IQUB
2	2	197,194,416	197,195,872	1.45	3	2	150 (5.21%)	136 (3.71%)	1.57 (1.22, 2.01)	$3.0x10^{-4}$	HECW2
3	2	147,560,112	147,586,272	26.17	2	23	112 (3.64%)	107 (1.90%)	1.62 (1.23, 2.12)	$4.2x10^{-4}$	-
4	2	201,345,488	201,352,912	7.40	2	9	13 (0.53%)	13 (1.90%)	0.28 (0.13, 0.60)	$4.6x10^{-4}$	BC047410, AOX2, BZW1
5	4	58,734,992	58,766,208	31.21	3	11	34 (0.95%)	12 (0.33%)	3.39 (1.63, 7.04)	$5.0x10^{-4}$	-
6	3	34,972,504	34,990,024	17.52	6	8	52 (0.99%)	24 (0.43%)	2.21 (1.39, 3.51)	$5.5x10^{-4}$	-
7	4	98,416,504	98,417,128	6.20	2	3	174 (8.55%)	204 (6.09%)	1.44 (1.17, 1.78)	$5.9x10^{-4}$	-
8	7	86,648,720	86,661,552	12.84	4	9	132 (3.49%)	99 (2.33%)	1.60 (1.22, 2.10)	$6.1x10^{-4}$	DMFT1, C7orf23
9	2	25,035,624	25,047,502	11.88	2	6	27 (1.29%)	75 (3.32%)	0.47 (0.30, 0.73)	$6.3x10^{-4}$	ADCY3, DNAJC27
10	12	110,131,840	110,133,560	1.73	5	3	1062 (19.98%)	1027 (18.32%)	1.19 (1.08, 1.31)	$6.4x10^{-4}$	CUX2
11	12	109,640,328	109,642,544	22.14	3	3	362 (11.25%)	306 (9.13%)	1.32 (1.12, 1.55)	$8.9x10^{-4}$	TCTN1, HVCN1, PPP1CC
12	9	34,495,480	34,501,456	59.74	3	7	85 (2.74%)	53 (1.57%)	1.80 (1.27, 2.57)	$9.1x10^{-4}$	DNAI1,ENHO, CNTFR
13	12	110,772,472	110,785,160	12.69	8	4	3219 (34.26%)	2673 (31.35%)	1.12 (1.05, 1.20)	$9.3x10^{-4}$	ALDH2, PNAS1, MAPKAPK5
14	4	103,494,864	103,502,480	76.20	2	2	64 (3.15%)	59 (1.76%)	1.83 (1.27, 2.63)	$9.7\ x10^{-4}$	SLC39A8
15	12	110,744,448	110,767,928	23.48	8	6	3219 (34.26%)	2672 (31.34%)	1.12 (1.05, 1.19)	$1.0\ x10^{-3}$	ALDH2, PNAS1, MAPKAPK5
16	14	73,352,952	73,355,664	27.05	2	2	12 (0.42%)	46 (1.17%)	0.36 (0.19, 0.68)	$1.1\ x10^{-3}$	PGR2, ZADH1, PZADH2, C14orf43
17	10	75,318,256	75,318,296	0.04	5	2	2055 (28.77%)	1877 (26.85%)	1.14 (1.05, 1.23)	$1.2\ x10^{-3}$	CAMK2G, PLAU, C10orf55
18	12	87,174,992	87,188,312	13.32	7	9	1373 (20.96%)	1541 (16.66%)	1.14 (1.05, 1.24)	$1.3\ x10^{-3}$	TMTC3
19	12	109,729,664	109,731,024	13.60	2	2	270 (11.37%)	266 (8.77%)	1.34 (1.12, 1.60)	$1.4\ x10^{-3}$	PPP1CC, CCD63

Reg	Chr	BP1	BP2	Length (kb)	Studies	SNPs	Cases	Controls	OR (95% CI)	PValue	Genes
20	2	147,676,864	147,731,328	54.46	4	14	216 (3.82%)	172 (2.84%)	1.40 (1.14, 1.72)	1.4×10^{-3}	-

Chr – chromosome, nSNPs – number of SNPs in overlapping region, OR – odds ratio. P-value – level of statistical significance for difference in frequency between cases and controls. **IQUB** – IQ motif and ubiquitin domain containing, **HECW2** – HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2, **AOX2** – aldehyde oxidase 2 pseudogene, **BZW1** – basic leucine zipper and W2 domains 1, **ADCY3 –** adenylate cyclase 3, **DNAJC27** – DnaJ (Hsp40) homolog subfamily C member 27, **CUX2** – cut-like homeobox 2, **TCTN1** – tectonic family member 1, **HVCN1** – hydrogen voltage-gated channel 1, **PPP1CC** – protein phosphatase 1 catalytic subunit gamma isozyme, **DNAI1** – dynein axonemal intermediate chain 1, **ENHO** – energy homeostasis associated, **CNTFR** – ciliary neurotrophic factor receptor, **ALDH2** – aldehyde dehydrogenase 2 family (mitochondrial), **MAPKAPK5** – mitogen-activated protein kinase-activated protein kinase 5, **SLC39A8** – solute carrier family 39 (zinc transporter), member 8, **PGR2** – G protein-coupled receptor 142, **ZADH1** – zinc binding alcohol dehydrogenase domain containing 1, **CAMK2G** – calcium/calmodulin-dependent protein kinase II gamma, **PLAU** – plasminogen activator, urokinase, **TMTC3** – transmembrane and tetratricopeptide repeat containing 3.

To see if there is a global enrichment for CAD-related consensus regions of consecutive homozygous SNPS in the human genome, all 15,441 regions identified through analysis of overlapping ROHs were classified into either potentially CAD protective or potentially detrimental based on their crude OR from analysis of association with CAD (OR>1: detrimental, OR<1: protective). Under a null hypothesis of no association between consensus regions with CAD, the frequency of each category should be close to 50%/50%. There was a clear deviation from the expected distribution of the consensus regions of consecutive homozygous SNPs in the CARDIoGRAM Consortium (Table 2.3.10). Indeed, there was an overall enrichment in potentially CAD-detrimental consensus regions in the CARDIoGRAM Consortium – approximately 10% excess of regions favouring increased risk of CAD was noted over those that tend to protect against CAD ($P=1.34x10^{-36}$) (Table 2.3.10). These data confirmed that consensus regions of consecutive homozygous SNPs on ROHs overlapping between populations are generally more common in patients with CAD than in CAD-free controls. The same enrichment for regions potentially detrimental to CAD was apparent in each category after stratification based on number of SNPs in each consensus segment – the excess of regions with OR>1.0 for CAD was close to 10% (Table 2.3.10).

The consensus regions of consecutive homozygous SNPs were then binned in 5 categories based on their nominal level of statistical significance for association with CAD (Table 2.3.11). This analysis revealed an enrichment for CAD-detrimental regions within the bin of the most significant results and a linear trend for a decrease of these regions towards the least significant bin ($P=8.25x10^{-20}$).

Table 2.3.11: Frequency of CAD protective and detrimental consensus regions of consecutive homozygous SNPs in CARDIoGRAM Consortium – analysis stratified on number of SNPs

	Regions	Expected	Observed	P-value
Overall	CAD detrimental	50% - 7720	55% - 8502	$1.34x10^{-36}$
	CAD protective	50% - 7720	45% - 6939	
Group 0 2-9 SNPs				
	CAD detrimental	50% - 4976	55% - 5462	$1.01x10^{-22}$
	CAD protective	50% - 4976	45% - 4490	
Group 1 10-49 SNPs				
	CAD detrimental	50% - 2104	55% - 2307	$2.08x10^{-10}$
	CAD protective	50% - 2104	45% - 1901	
Group 2 50-99 SNPs				
	CAD detrimental	50% - 424	56% - 479	$8.9x10^{-5}$
	CAD protective	50% - 424	44% - 369	
Group 3 100+ SNPs				
	CAD detrimental	50% - 216	59% - 254	$1.8x10^{-4}$
	CAD protective	50% - 216	41% - 179	

Regions were classified as detrimental and protective if their odds ratio (OR) for CAD was >1.0 and <1.0 (respectively). Data are counts and percentages, P-value – level of statistical significance from a test of binomial test

Table 2.3.12: Frequency of CAD protective and detrimental consensus regions of consecutive homozygous SNPs in CARDIoGRAM Consortium – analysis stratified on bins of nominal statistical significance for association with CAD

P-value bin	CAD detrimental regions	CAD protective regions	P-value*
0.8-1.0	1509 (17%)	1483 (21%)	
0.6-0.8	1581 (19%)	1442 (21%)	
0.4-0.6	1710 (20%)	1368 (20%)	$8.25x10^{-20}$
0.2.0.4	1665 (20%)	1406 (20%)	
0.0-0.2	2037 (24%)	1240 (18%)	
15,441	**8,502**	**6,939**	**-**

Regions were classified as detrimental and protective if their odds ratio (OR) for CAD was >1.0 and <1.0 (respectively). P-value bin – category based on the level of statistical significance, P-value* - level of statistical significance from test for trend

2.4. Discussion

Homozygosity mapping is a strategy with the potential to uncover such recessive variants hidden within long stretches of homozygous SNPs (Keller *et al*. 2012). This analysis has been useful in the identification of disease susceptibility genes in both monogenic and complex diseases (Miyazawa *et al*. 2007; Jiang *et al*. 2009). The effects of inbreeding and recessive variants on the risk of complex diseases have been previously well established (Rudan *et al*. 2003a, Rudan *et al*. 2003b, Rudan *et al*. 2006; Campbell *et al*. 2007).

Indeed, a strong linear relationship between the inbreeding coefficient and blood pressure was found and several hundred recessive loci were predicted as contributing to blood pressure variation. Recessive or partially recessive variants account for 10-15% of the total variation in blood pressure (Rudan *et al*. 2003a). Inbreeding was a significant predictor for a number of late-onset complex diseases such as CAD, stroke, cancer and asthma (Rudan *et al*. 2003b). These studies have strongly supported the hypothesis that the genetics of complex phenotypes include a component of recessively acting variants.

Nowadays, high-density genome-wide SNP arrays represent a powerful tool for discovering regions of extended homozygosity in the human genome (Lencz *et al*. 2007a; McQuillan *et al*. 2008).

ROH analyses conducted to date have primarily focused on questions relevant to both basic population genetics theory and disease risk. Population genetics studies have analysed the distribution, prevalence and location of ROHs across various sub-populations to infer population substructure, history and natural selection (Gibson *et al*.

2006; Kirin *et al*. 2010; Li *et al*. 2006; McQuillan *et al*. 2008; Nothnagel *et al*. 2010; Sabeti *et al*. 2007; Voight *et al.* 2006). Phenotypic studies have used both family-based and population-based samples to identify specific associated risk ROHs as well as differences in overall ROH burden (Lencz *et al*. 2007a; Keller *et al*. 2012; Howrigan *et al.* 2011).

As the offspring of inbred populations may have lower mean health and fitness because of the homozygous expression of detrimental recessive alleles (Rudan *et al.* 2003a, Rudan *et al*. 2003b, Rudan *et al*. 2006), similar effects could operate with the more numerous partially recessive variants influencing complex diseases in outbred populations.

The importance of ROHs on CAD or any other cardiovascular phenotype is completely unexplored. The data collected here come from the first genome-wide homozygosity analysis for CAD. The study made use of both directly genotyped SNPs and imputed genotypes in the homozygosity analysis to increase the coverage for the genome and extract the maximum information. The extremely dense SNP typing in the populations of CARDIoGRAM Consortium provided a unique opportunity to examine the distribution, number, size, and location of these homozygous tracts. Homozygosity measures such as ROH number and ROH length were used to investigate the genetic architecture of ROHs.

ROH number and ROH length constitute two important descriptive characteristics of homozygosity. ROH number is used to count ROHs and ROH length to size ROHs across the genome. They are two complimentary measures and provide a description of the overall genetic architecture of ROHs.

This study revealed that on average every individual has ~32 ROHs in their autosomal genome, similar to what was previously published (Simon-Sanchez *et al*. 2012). Comparative case-control analysis showed that CAD patients have ~ an excess of 0.5 ROH when compared to controls. When it comes to the extreme ends of the distribution curve it becomes obvious that only very few CAD patients carry <10 ROHs covering <10Mb or >100 ROHs covering >50Mb.

FROH is a measure of inbreeding effects and correlates most highly with the homozygous mutational load, the putative causal mechanism underlying inbreeding depression (McQuillan *et al*. 2012). This homozygosity measure is important because it has low prediction error variance, especially when SNP density is high (Keller *et al*. 2012). However, given the small variation in genome-wide FROH across unrelated individuals, large sample sizes are necessary to detect inbreeding depression for likely effect sizes (Keller *et al*. 2012).

Previous studies investigating the effects of FROH on human complex traits have sample sizes<3,000 and have failed to find significant inbreeding effects (Nalls *et al*. 2009; Spain *et al*. 2009; Vine *et al*. 2009; Enciso-Mora *et al*. 2010; Hosking *et al*. 2010) most likely because they were underpowered. Furthermore, small studies (<1,000) that did find significant inbreeding depression effects using FROH (Lencz *et al*. 2007a) may greatly overestimated the size of the effects.

The sample study of this study (~21,000), the largest examined one so far, provided a well-powered tool for finding signatures of autozygosity on CAD risk. A 20% increase in CAD risk for every 1% increase in the proportion of autosomal autozygosity suggesting the collective role of multiple recessive variants on disease risk.

This study showed a good agreement with previous studies regarding all homozygosity measures (Simon-Sanchez *et al*. 2012). However, one should keep in mind that the rate of detected ROHs in homozygosity mapping tends to vary along the genome due to differences in informativeness and differences in haplotype genealogies. The relative performance of homozygosity mapping depends on population demographic history and the strength of selection against causal variants (Browning and Thompson, 2012). Comparison between different studies may be also challenging because ROH identification criteria vary, especially the minimum length consisting a ROH (Ku *et al*. 2011).

Previous research showed the existence of a trend for decreasing autozygosity with younger chronological age in the North American population of European ancestry (Nalls *et al.* 2009). CAD cases in the analysis were generally younger than controls. As a result, someone would expect controls to present longer, more frequent ROHs. However, CAD patients showed excess homozygosity when compared to controls.

Potential confounding factors such as age and sex were not adjusted in the analysis. However, the sensitivity analysis conducted in WTCCC showed no differences in homozygosity measures between males and females and across different age categories.

Analysis of individual consensus regions of overlapping ROHs showed no statistically significant association with CAD risk. A region with high frequency in affected individuals and very low frequency in unaffected individuals could provide strong evidence for the presence of disease-associated gene/genes. None of the identified consensus regions were exclusive to CAD cases. Several of these regions mapped to gene deserts contain only a single or very few genes. None consensus region has been implicated in previous CAD GWA studies. This is not surprising given that GWA

studies have analysed data under additive model of inheritance whilst ROH analysis operated under recessive mode of inheritance.

Most of the top consensus regions are uncommon (<5%) or indeed very rare (<1%). A few consensus homozygous sequences (i.e. present on chromosome 12) have high population frequencies (>25%) and appear to coincide with previously identified ROH islands. ROH islands show high frequency of ROH in European populations (Nothnagel *et al*. 2010). These regions are unlikely to be disease specific as they just reflect the natural architecture of our DNA.

Although there was no evidence for association between individual consensus regions and CAD (possibly due to their low frequency and hence the power issues), further analysis was undertaken to examine if collectively they may associated with CAD.

Evaluation of the distribution of these consensus regions of homozygous SNPs showed a 10% excess of the consensus regions favouring increased risk of CAD over those that tend to protect against it, a result inconsistent with chance. This supported the hypothesis and suggested that patients with CAD may have accumulated more recessive alleles than controls. To some extent, this enrichment of consensus regions of homozygous SNPs may be the potential driver for the excess homozygosity observed in CAD cases compared to controls in the analysis of genetic architecture of homozygosity measures. This finding is important because it provides evidence for an excess of ROHs as a potential contributor to CAD and therefore supports a theory on the role of recessive component in the genetic architecture of CAD. Additional work in needed to unravel the exact synergistic role of multiple recessive variants, homozygosity levels and their association to CAD. DNA-sequencing studies will offer in-depth coverage for these regions and will advantage genome-wide homozygosity analysis.

Strengths and limitations of genome-wide homozygosity analysis

1. Optimum-length threshold

A central limitation to current studies analysing ROHs is the lack of consensus on criteria for ROHs identification (Ku *et al*. 2011). In outbred population, ROHs reflect distant consanguinity and originate from common ancestors (Charlesworth and Willis, 2009); however the length of ROHs decreases on average over time due to recombination. The expected length of a ROH follows an exponential distribution with mean equal 1/2g Morgans, where g is the number of generations since the common ancestor (Howrigan *et al*. 2011). As a result, the expected length of a ROH caused by sib-sib inbreeding (g=2) is calculated at 1/4 Morgan or 25cM, while the expected length of a ROH originating from a common ancestor 50 generations in the past is 1/100 or 1cM (Houwen *et al*. 1994). Because the shortening of ROHs across generations is gradual, any choice of length threshold to define ROHs is ultimately arbitrary. Consequently, the discrepancy between definitions of ROHs across studies makes their comparisons difficult (Howrigan *et al*. 2011). Unfortunately, to date there has been no systematic investigation of ROHs detection across different statistical methods and power.

This project detected ROHs using a genotype-counting approach that relies on a fixed size, allowing for occasional missing or heterozygous genotypes to account for possible genotyping errors. In contrast, an approach incorporating population-specific allele frequency estimates to determine autozygosity status of a window could enable more rigorous assessments of the possibility of genotyping errors (more sensitive detection of ROHs).

2. Overlapping deletions

Hemizygous deletions may affect the estimates of ROHs (Nothnagel *et al*. 2010). This study was not able to check if ROHs overlap with regions of hemizygous deletions. Previous studies showed that ROHs are true homozygous tracts and not deletions or other chromosomal abnormalities (Brown and Weber, 1999; Frazer *et al*. 2007; Simon-Sanchez *et al*. 2007; Li *et al*. 2006).

3. LD-pruned dataset

LD can act as a potential confounder in comparative ROH analyses of different populations because the local level of LD determines the effective number of SNPs used for ROH definition (Nalls *et al*. 2009). However, it is unclear how much specificity and sensitivity this brings to the homozygosity analysis.

Summary of findings

Genome-wide homozygosity analysis in CARDIoGRAM Consortium revealed statistically significant differences in the overall homozygosity levels between CAD patients and CAD-free controls. Also, a 20% increase in CAD risk was suggested for every 1% increase in the proportion of autosomal homozygosity. The distribution of consensus regions of overlapping ROHs showed over-representation amongst patients with CAD, suggesting that accumulation of recessive alleles may increase the risk of CAD.

CHAPTER 3

LOW-FREQUENCY/RARE VARIANTS AND PREDISPOSITION TO CORONARY ARTERY DISEASE

3.1. Introduction

GWA analysis is a powerful method of genomic screening for common variants underlying complex diseases (Manolio *et al.* 2008). Several hundred such variants have been identified so far through GWA studies (http://www.genome.gov/gwastudies/). However, the collective contribution of these variants to overall heritability of complex disorders is very modest and explains only its small proportion (Frazer *et al*. 2009). One potential contributor to the remainder of "missing heritability" is the effect of multiple low-frequency/rare alleles (Kryukov *et al*. 2007; Bodmer and Bonilla, 2008; Manolio *et al*. 2009; Schork *et al.* 2009; Eichler *et al*. 2010). Such variants are likely to contribute to phenotypic expression in conjunction with, or over and above, common variants (Bansal *et al.* 2010). The extent to which these variants may actually contribute to disease predisposition is of great interest and represents one of the major unanswered questions in complex disease genetics.

3.1.1. Low-frequency/rare variants

Low-frequency alleles are usually defined as those with frequency between 1% and 5% whilst rare variants have frequency <1%. The MAF distribution of SNPs from the International Hap-Map project shows that >40% of SNPs have MAF<5% (Gorlov *et al.* 2008).

3.1.2. Rare variants and susceptibility to complex diseases

3.1.2.1. Rare familial disorders and rare variants

The role of both low-frequency and rare alleles as the sole genetic determinants of Mendelian forms of disorders is well-known (Gibson, 2012). Heterozygous familial hypercholesteroleamia and monogenic forms of hypertension are the excellent illustrations of how a single mutant allele travelling from one generation to another may account for almost entire phenotypic spectrum in the clinical manifestation of rare or very rare disease (rare variant – rare disease) (Goldstein and Brown, 1979). Other examples include, the breast cancer 1 (*BRCA1*) and breast cancer 2 (*BRCA2*) susceptibility mutations (Easton *et al.* 2007), and variants that are responsible for maturity onset diabetes of the young (MODY) (Weedon and Frayling, 2007). The growing body of evidence appears to suggest that rare variants may also play a role in special cases of complex diseases that have familial analogues.

3.1.2.2. Evolutionary theory and the rare allele model

The strongest argument for the rare allele model derives from evolutionary theory which favours the existence of numerous rare polymorphisms rather than common variants in the development of common late-onset diseases (Pritchard, 2001; Gorlov *et al.* 2008; Gibson, 2012). It is believed that disease-promoting variants deleterious to fitness should be rare since they are under negative selection (Barton and Turelli, 1989; Pritchard and Cox, 2002). Rare disease-related variants reflect the balance between mutation process generating them and purifying selection process manipulating their occurrence and preventing them from drifting to a higher frequency in the population (Reich and Lander, 2001). Mutation rates are sufficiently large to promote new disease variants but on the other hand purifying selection cannot eliminate all deleterious

variants especially the ones that affect late-onset diseases. As a result they have the opportunity to rise to allele frequencies of 1% or even more, particularly if their effect is recessive (Gibson, 2012).

3.1.2.3. Empirical population genetic data and rare variants

It has been shown that the distribution of MAFs is strongly skewed towards an excess of low-frequency/rare variants. Whole-exome sequence data indicated that non-synonymous substitutions are significantly over-represented towards low-frequencies suggesting the operation of purifying selection (Cargill *et al.* 1999; Kryukov *et al.* 2007; Zhu *et al.* 2011).

3.1.2.4. Synthetic associations

It has also been argued that some of the identified common variant association signals may actually reflect rare variants (Dickson *et al.* 2010). The term "synthetic association" is used to describe disease association of a common variant that is actually driven to its LD relationship with several disease-promoting rare variants located on the same haplotype block (Dickson *et al*. 2010; Gibson, 2012). It is believed that a common variant may highlight the presence of two-three rare variants that each substantially increases disease risk in just 1-2% of the cases (Gibson, 2012).

3.1.3. Strategies to identify rare variants

Rare susceptibility variants are a considerable analytical challenge, because established disease association methods are tailored to common susceptibility variants and are unlikely to be powerful enough for rare variants (Zeggini *et al.* 2005).

The statistical power to detect susceptibility alleles is positively correlated with their frequency and penetrance. Detection of rare alleles with high penetrance is possible through GWA studies as the detection of common alleles with modest penetrance but the problem is how well these low-frequency variants are captured with the GWA arrays that are designed to tag common SNPs or how well are called (Bodmer and Bonilla, 2008; McCarthy and Hirschhorn, 2008).

Current approaches to investigate the effect of rare variants on complex phenotypes tend to use direct whole-genome re-sequencing (Metzker, 2010). Sequencing of candidate genes, the whole exome or the entire human genome is the optimal way to identify rare variants (Li and Leal, 2009). The most commonly used whole genome sequencing available platforms generate millions of short sequence reads that are then aligned to a reference genome through read mapping. Variant calling algorithms are subsequently employed to identify candidate sites at which one or more samples differ from the reference sequence and to call genotypes across samples (Panoutsopoulou *et al.* 2013).

In addition, the development of the 1000 Genomes project, a large international effort, which sequenced 1000 genomes of individuals from 10 different ethnic backgrounds (Siva, 2008; The 1000 Genomes Consortium et al. 2010), brought to light new information on human variation. A detailed catalogue of variants enabled rapid progress in association studies (Li and Leal, 2008). As a result now it is also possible to use GWA studies as a template to impute low-frequency and rare variants based on sequenced reference panel such as the 1000 Genomes project. A recent GWA analysis of imputed rare variants across seven common complex diseases, identified genome-wide significant evidence of rare variant association in PR domain containing 10 gene

(PRDM10) with CAD and multiple genes in the major histocompatibility complex (MHC) with type 1 diabetes (Magi *et al.* 2012).

3.1.4. Rare variant association analysis

Association mapping of rare variants has focused on the aggregation of the effects of all rare variants within a genomic region because it is neither powerful nor numerically stable to analyse each variant individually (Magi *et al*. 2012).

A plethora of such novel locus-specific experimental strategies and statistical models have been developed to detect binary or quantitative trait associations. These strategies are classified into a few main groups (Panotsopoulou *et al.* 2013):

- Collapsing approaches based on summary statistics such as the Cohort Allelic Sum Test (CAST) (Morgenthaler and Thilly, 2007), Combined Multivariate and collapsing test (CMC) (Li and Leal, 2008), Weighted Sum Statistic test (WSS) (Madsen and Browning, 2009) and Variable-Threshold approach (VT) (Price *et al.* 2010).
- Methods based on similarities among individual sequences such as Kernel Based Association Test (KBAT) (Mukhopadhyay *et al.* 2010) and Sequence Kernel Association Test (SKAT) (Wu *et al.* 2011).
- Regression models that use collapsed sets of variants and other factors as predictors such as collapsing test using proportion of rare variants (GRANVIL) (Morris and Zeggini, 2010), Adaptive Sum test (Han and Pan, 2010), LASSO and Ridge regression (Zhou *et al.* 2010; Asimit and Zeggini, 2010).

Collapsing approaches aggregate information across multiple variants within a genomic locus into a single unit, which is then examined for trait association with an accumulation of rare minor alleles. Alternative approaches such as KBAT (Mukhopadhyay *et al*. 2010) and SKAT are (Wu *et al.* 2011) multivariate tests combining single-variant test statistics. Given that the allelic architecture of complex traits is largely unknown; these tests make no assumptions about the probability or direction of each variant effect and are therefore more flexible (Panoutsopoulou *et al*. 2013).

3.1.5. Examples of rare variants contributing to complex traits

The early evidence for the role of rare variants in human disease comes from in-depth sequencing and re-sequencing studies in the area of oncology (Schork *et al.* 2009). One of the first examples were variants in *BRCA1* and *BRCA2* genes underlying susceptibility to breast and ovarian cancer (Stratton and Radman, 2008). These disease causing variants confer a 10- to 20- fold relative breast cancer risk. This translates into a 30 to 60% increase in risk in carriers of a mutant variant by the age of 60, compared to just 3% in the general population (Stratton and Radman, 2008). Approximately 1 in 1,000 individuals (MAF<0.01%) are heterozygous mutation carriers of each gene, and there are numerous different mutations each of which is very rare (Stratton and Radman, 2008).

Analysis of cardiovascular quantitative traits and in particular circulating lipid levels also revealed rare variants acting collectively on the phenotype. Screening for variants in genes implicated in Mendelian forms of low HDL-cholesterol levels revealed an aggregation of rare alleles in individuals with low HDL-cholesterol compared to those

with high HDL-cholesterol levels (Cohen *et al.* 2004). Re-sequencing of angiopoietin-like 4 gene (*ANGPTL4*) uncovered both rare and common variants that reduce triglycerides and increase HDL-cholesterol (Romeo *et al.* 2007).

A ground-breaking study by Lifton's group revealed associations between rare independent mutations in genes responsible for renal salt handling and blood pressure and risk of hypertension (Ji *et al*. 2008). Screening of individuals of the FHS offspring cohort for variation in three genes known for their role in rare Mendelian forms of hypertension/hypotension [solute carrier family 12 (sodium/chloride Transporter) member3 (*SLC12A3*), solute carrier family 12 (sodium/potassium/chloride Transporter) member1 (*SLC12A1*) and potassium inwardly-rectifying channel, subfamily J, member1 (*KCNJ1*)] identified rare heterozygous mutations. These mutations lead to significant decrease in blood pressure and protection from development of hypertension. The mean long-term SBP among mutation carriers was 6.3mm Hg lower than the mean of the cohort and showed a 59% reduction of developing hypertension by age 60 compared to non-carriers.

A large-scale, gene-centric study conducted by our group on the genetic architecture of ambulatory blood pressure in the general population revealed a significant over-representation of low-frequency/rare variants (MAF<0.05) among variants showing at least nominal association with mean 24-hour blood pressure (Figure 3.1.1) (Tomaszewski *et al*. 2010). This difference became even more striking when applying a lower threshold for definition of low-frequency variants (MAF<2%) - SNPs with MAF<2% were almost two times more common among variants associated with mean 24-hour blood pressure than variants of higher frequency. This observation supported

the hypothesis that low-frequency/rare alleles may play a role in genetic predisposition to blood pressure elevation (Ji *et al*. 2008; Tomaszewski *et al.* 2010).

Figure 3.1.1: Distribution of SNPs associated nominally (P<0.05) with mean 24-hour BP according to their characteristic and frequency. [Taken from Tomaszewski *et al*. 2010]

Descriptor	Illumina gene-centric array	SNPs associated with mean 24-hour SBP at p<0.05	P-value	SNPs associated with mean 24-hour DBP at p<0.05	P-value
All SNPs	33,577	1,782	-	1,842	-
Exonic SNPs	2,298 (6.8%)	126 (7.1%)	0.7004	138 (7.5%)	0.2770
Non-synonymous SNPs	1,634 (4.9%)	84 (4.7%)	0.8211	94 (5.1%)	0.6175
All SNPs with MAF<0.05	4,355 (13.0%)	290 (16.3%)	8.7×10^{-5}	283 (15.4%)	0.0036
Exonic SNPs with MAF<0.05*	516 (11.8%)	32 (11.0%)	0.7778	34 (12.0%)	0.9244
Non-synonymous SNPs with MAF<0.05*	406 (9.3%)	24 (8.3%)	0.6021	24 (8.5%)	0.7507
All SNPs with MAF<0.02	1,188 (3.5%)	104 (5.8%)	3.0×10^{-6}	112 (6.1%)	2.0×10^{-7}
Exonic SNPs with MAF<0.02†	162 (13.6%)	14 (13.5%)	1.0	20 (17.9%)	0.2530
Non-synonymous SNPs with MAF<0.02†	128 (10.8%)	10 (9.6%)	0.8685	14 (12.5%)	0.5292

SNP – single nucleotide polymorphism, MAF – minor allele frequency

*percentage calculated in relation to total number of SNPs with MAF<0.05

†percentage calculated in relation to total number of SNPs with MAF<0.02

Re-sequencing of four triglyceride-modulating candidate genes – apolipoprotein A-V (*APOA5*), glucokinase (hexokinase 4) regulator (*GCKR*), lipoprotein lipase (*LPL*) and apolipoprotein B (*APOB*) revealed a significant burden of 154 rare missense or nonsense variants in 438 individuals with hypertriglyceridemia, compared to 53 variants in 327 controls corresponding to a carrier frequency of 28.1% of affected individuals and 15.3% of controls (Johansen *et al*. 2010). This study showed accumulation of rare variants in genes identified through GWA and that these contribute to the heritability of complex traits among individuals at the extreme of a lipid phenotype.

More recently, re-sequencing in a region previously implicated in GWA studies uncovered four rare new variants (each MAF of approximately 1%) within interferon induced with helicase C domain 1 gene (*IFIH1)* associated with type 1 diabetes

independently of each other (Nejentsev *et al.* 2009). Each of these rare variants showed a protective effect on the risk of type 1 diabetes (OR = 0.51-0.74).

Sequencing of melatonin receptors 1B gene (*MTNR1B*) associated with type 2 diabetes identified several rare variants (MAF<1%) that through impairing receptor function contribute to risk of type 2 diabetes (Bonnefold *et al*. 2012).

Rare and low-frequency variants with individual large effect sizes have also been reported in other complex diseases. Five rare (MAF <1%) variants in nucleotide-binding oligomerization domain containing 2 gene (*NOD2)* were associated with Crohn's disease; they appear to act independently from each other as well as from the previously implicated low frequency causal variants (Hugot *et al.* 2001; Ogura *et al.* 2001; Rivas *et al*. 2011). For example, a rare missense variant in myosin, heavy chain 6, cardiac muscle, alpha gene *(MYH6*) was associated with ~12-fold increase in risk of sick sinus syndrome (Holm *et al.* 2011). Whole genome sequencing efforts of affected trios has led to the identification of several *de novo* mutations implicated in the aetiology of autism (O'Roak *et al.* 2011; Sanders *et al*. 2012; Neale *et al*. 2012; O'Roak *et al*. 2012), schizophrenia (Girard *et al*. 2011; Xu *et al*. 2011) and intellectual disability (Vissers *et al.* 2010).

3.1.6. Low-frequency/rare variants and CAD

Several low-frequency/rare variants have been associated with CAD so far. The most well-known example is rs3798220 (MAF~2% - OR=1.92) at the *SLC2A-LPAL2-LPA* locus that was first came to light by haplotype association analysis (Tregouet *et al*. 2009) and subsequently was replicated by a custom-made 50K gene array (Clarke *et al*.

2009; The IBC 50K CAD Consortium, 2011). Moreover, application of haplotype association analysis to the WTCCC GWA data identified rare variants at a known locus cyclin-dependent kinase inhibitor 2B (*CDKN2B)* and three new genes for CAD – eukaryotic translation initiation factor 4H (*EIF4H*), hemochromatosis type 2 (*HFE2*) and zinc finger and BTB domain containing 43 (*ZBTB43*) (Zhu *et al.* 2010) (Table 3.1.1).

Table 3.1.1: Risk haplotypes and their corresponding frequencies in cases and controls for WTCCC CAD data

Gene	Chromosome	Start SNP	End SNP	Freq in cases	Freq in controls	P-value
CDKN2B	9	rs3217986	rs10965245	0.0119	0.0060	1.27×10^{-3}
				0.0156	0.0077	1.55×10^{-4}
				0.0457	0.0373	2.40×10^{-2}
				0.0114	0.0075	2.45×10^{-2}
EIF4H	7	rs150880	rs17146094	0.0119	0.0005	1.13×10^{-15}
HFE2	1	rs12091564	rs10218795	0.0065	0.0005	6.54×10^{-8}
ZBTB43	9	rs10987465	rs7038622	0.0068	0.0009	5.15×10^{-7}
				0.0039	0.0005	1.86×10^{-4}
				0.0039	0.0003	4.90×10^{-5}

SNP – single nucleotide polymorphism, Freq- frequency, P-value – level of statistical significance on Fisher's exact test

Despite these initial findings, our knowledge of the impact of low-frequency/rare variants on CAD remains limited and further investigation and elucidation of their role and susceptibility/protective properties is needed in order to offer a better individual prediction of CAD disease risk.

3.1.7. Hypothesis

Patients with CAD show over-representation of low-frequency/rare alleles at some loci compared to CAD-free controls.

3.1.8. Objectives

- To evaluate the burden of low-frequency/rare variants in individuals with CAD compared with controls using data from the IBC 50K CAD chip.

- To identify biologically strong CAD candidate genes for further examination in relation to low-frequency/rare variants

3.2. Materials and Methods

3.2.1. Characteristics of study cohorts

The primary analysis was performed in >18,000 individuals from four European populations, all members of the IBC 50K CAD Consortium (The IBC 50K CAD Consortium, 2011) known as BLOODOMICS (Bezzina *et al*. 2010; Winkelmann *et al.* 2001), the British Heart Foundation Family Heart Study (BHF-FHS) (Samani *et al*. 2005; Wellcome Trust Case Control Consortium, 2007), the PennCATH study (Lehrke *et al*. 2007; Kathiresan *et al.* 2009) and the Precorious Coronary Artery Disease (PROCARDIS) study (Clarke *et al*. 2009).

- ***BLOODOMICS***

The Dutch component of the BLOODOMICS collaboration includes patients with CAD drawn from the Academic Medical Centre Amsterdam Premature Atherosclerosis Study (AMC-PAS) and the AGNES study (Bezzina *et al.* 2010). The CAD-free controls were recruited from the Sanquin Common Controls (SANQUIN-CC) study.

AMC-PAS/Sanquin: Patients with symptomatic CAD (defined as MI, coronary revascularization, or evidence of at least 70% stenosis in a major epicardial artery) before the age of 51 years were recruited as part of a prospective cohort study (AMC-PAS) (Bezzina *et al*. 2010). Blood donors from the north-west region of the Netherlands were established as controls for this study. Participating donors were recruited at routine Sanquin Blood Bank donation sessions (SANQUIN-CC). More than 95% of the controls are from the same region as the cases of the AMC-PAS cohort (Bezzina *et al.* 2010).

AGNES: The AGNES case-control set consisted of individuals with a first acute ST-elevation myocardial infarction (Bezzina et al. 2010), hence the whole set was

considered as cases in this project. AGNES cases had ECG-registered ventricular fibrillation occurring before reperfusion therapy for an acute and first ST-elevation myocardial infarction. AGNES controls were individuals with a first acute ST-elevation myocardial infarction but without ventricular fibrillation. All individuals were recruited at seven heart centres in the Netherlands between 2001–2008 (Bezzina *et al.* 2010). Individuals with an actual non-ST-elevation MI, prior MI, congenital heart defects, known structural heart disease, severe comorbidity, electrolyte disturbances, trauma at presentation, recent surgery, previous coronary artery bypass graft or use of class I and III antiarrhythmic drugs were excluded (Bezzina *et al.* 2010). Individuals who developed ventricular fibrillation during or after percutaneous coronary intervention were not eligible. Furthermore, because early reperfusion limits the opportunity of developing ventricular fibrillation, potential control subjects undergoing percutaneous coronary intervention within 2h after onset of myocardial ischemia symptoms were not included (Bezzina *et al.* 2010). This time interval was based on the observation that >90% of cases developed ventricular fibrillation within 2h after onset of symptoms (Bezzina *et al.* 2010).

The German component of the BLOODOMICS collaboration provided cases and controls from the LUdwigshafen RIsk and Cardiovascular Health (LURIC) study, supplemented by additional controls from the Mannheim study (Bugert *et al.* 2003).

LURIC: LURIC is a prospective study of cardiovascular death in individuals of German ancestry resident in southwest Germany. Each subject who underwent elective coronary angiography and left ventriculography between June 1997 and January 2000 (Winkelmann *et al*. 2001) was included in the study. CAD in this project was defined by troponin-confirmed MI or presence of visible luminal narrowing of ≥50% in at least one

coronary vessel. Individuals with ≥20% but <50% stenosis were excluded from the analyses. Individuals with stenosis <20% were regarded as controls (Winkelmann *et al.* 2001).

Mannheim study: Additional controls consisted of 1,187 healthy, unrelated blood donors 18–68 years of age (Bugert *et al.* 2003). They were recruited in 2004 and 2005 by the Institute of Transfusion Medicine and Immunology (Mannheim, Germany) and share the ethnic background with the LURIC patients. According to the German guidelines for blood donation, all blood donors were examined by standard questionnaires. All blood donors consented to the use of their samples for research studies.

- ***British Heart Foundation Family Heart Study (BHF-FHS)***

Individuals of European ancestry with validated history of either MI or coronary revascularisation (coronary artery bypass surgery or percutaneous coronary angioplasty) before their 66th birthday (Samani *et al.* 2005) were included in this study. Recruitment was carried out on a national basis in the UK through (a) responses to a continued UK-wide media campaign (b) responses to posters placed within hospitals and GP surgeries through the UK and (c) in a pilot-phase contacting patients listed on computer based CAD databases in the two lead centres (Leeds and Leicester) (Samani *et al.* 2005). The recruitment phase lasted five years (April 1998 to November 2003).

Controls were European Caucasian healthy blood donors between 30-70 years of age recruited all over the UK through the UK National Blood Service as part of the Wellcome Trust Case Control Consortium study (Wellcome Trust Case Control Consortium, 2007). Apart from age and sex, limited information was available on the controls.

- ***PennCATH***

The PennCATH cohort is a University of Pennsylvania Medical Centre based coronary angiographic study (Lehrke *et al.* 2007; Kathiresan *et al.* 2009). Information on this cohort is extensively provided in Chapter 2.

- ***Precocious Coronary Artery Disease (PROCARDIS)***

Ascertainment criteria for PROCARDIS probands were MI or symptomatic acute coronary syndrome (ACS), before the age of 66 years (Clarke *et al.* 2009). Diagnosis of MI required documentation of two or more of: (a) typical ischaemic chest pain, pulmonary oedema, syncope or shock; (b) development of pathological Q-waves and/or appearance or disappearance of localised ST-elevation followed by T-wave inversion in two or more standard electrocardiograph leads; (c) increase in concentration of serum enzymes consistent with MI (eg creatine kinase more than twice the upper limit of normal) (Clarke *et al.* 2009). Diagnosis of ACS required documentation of hospitalisation for one of the following indications: (a) unstable angina diagnosed by typical ischemic chest pain at rest associated with reversible ST-depression in two or more standard electrocardiograph leads; (b) thrombolysis for suspected MI (as indicated by localised ST-elevation in two or more standard electrocardiograph leads) even without later development of T-wave inversion, Q-waves, or a significant enzyme rise; or (c) emergency revascularisation (i.e. during same admission) following presentation with typical ischemic chest pain at rest (Clarke *et al.* 2009). Parents and up to four unaffected siblings per family were recruited wherever possible to augment the recovery of linkage phase information. This is a multicentre study and subjects were recruited in four countries: Sweden, UK, Germany and Italy. 99.5% of the study participants reported white European ancestry.

- ***The London Life Sciences Prospective Population Cohort (LOLIPOP)***

LOLIPOP is an ongoing population based cohort study that recruited 30,000 Indian Asian and European white men and women, aged 35-75 years, from the lists of 58 GP surgeries in West London, United Kingdom (Chambers *et al.* 2008, The IBC 50K CAD Consortium, 2011). Response rates averaged 62%; there are no major differences between responders and non-responders with respect to age, sex, co-morbidity and available risk factors (Chambers *et al.* 2008). DNA was available for ≈5000 Indian Asian participants. Indian Asians were selected if all four grandparents originated from the Indian subcontinent.

- ***Pakistan Risk of Myocardial Infarction Study (PROMIS)***

PROMIS is an ongoing case-control study of MI in six centres in urban Pakistan (Saleheen *et al.* 2009; The IBC 50K CAD Consortium, 2011). Participants in the recent study were recruited between 2005 and 2008. MI cases had symptoms within 24 hours of hospital presentation, typical electrocardiographic changes, and a positive troponin-I test.

Controls were individuals without a history of cardiovascular disease. They were frequency-matched to cases by sex and age (in 5 years bands) and concurrently identified in the same hospitals as index cases because they were either: (a) visitors of patients attending the outpatient department; (b) patients attending the outpatient department for routine non-cardiac complaints, or (c) non-blood related visitors of index MI cases. People with recent illnesses or infections were not eligible. Information was recorded on personal and paternal ethnicity, spoken language dietary intake, lifestyle factors and other characteristics.

3.2.2. Illumina HumanCVD BeadChip

HumanCVD BeadChip array (Illumina), also known as the "ITMAT-Broad-CARe" (IBC) 50K array, harbours ~50,000 SNPs to efficiently capture genetic diversity across ~2,100 candidate genes and pathways related to cardiovascular, inflammatory and metabolic phenotypes such as CAD, type 2 diabetes, lipids and hypertension (Keating *et al.* 2008). Variants in genes associated with sleep, lung and blood diseases were also included. Genetic variation within the majority of these regions is captured in a density equal to or greater than that afforded by GWAs (Keating *et al.* 2008). The IBC array has content derived from the International HapMap Consortium and re-sequencing data from the SeattleSNPs and National Institute of Environmental Health Sciences (NIEHS) SNPs consortia, with a focus upon inclusion of lower-frequency variants and variants with a higher likelihood of functionality (Lanktree *et al.* 2011).

Genes were prioritised; "high-priority genes" were densely tagged (all SNPs with MAF>2% tagged at $r^2>0.8$), "intermediate priority genes" were moderately well covered (all SNPs with MAF>5% tagged at $r^2>0.5$) and "low-priority genes" had only non-synonymous SNPs and known functional variants with MAF>1% (Keating *et al.* 2008).

A "cosmopolitan tagging" approach was used to select SNPs providing high coverage of selected genes in 4 HapMap populations (CEPH Caucasians, Han Chinese, Japanese, Yorubans). Approximately 17,000 SNPs included on the IBC array have a MAF<5% in individuals of European ancestry. For the majority of regions, SNPs were designed to be inclusive of the intronic, exonic, and flanking un-translated regions (UTRs), as well as to provide coverage of the proximal promoter regions designed for the higher-priority

loci. Of SNPs included on the IBC array, ~65% are intronic, 9.9% are exonic and 7.7% are non-synonymous (Lanktree *et al.* 2011).

3.2.3. Quality controls

Quality controls were performed in each cohort independently prior to association analysis. Samples were excluded where individual call rates were<90%. SNPs were removed for call rates of <90% or for the Hardy-Weinberg equilibrium cut-off $P \leq 0.0001$.

3.2.4. Rare variant analysis

The single-point analysis of rare variants is usually largely under-powered, because rare alleles are observed in very few subjects. In order to maximize the statistical power over single marker analysis, all low-frequency/rare alleles (MAF<3%) within defined regions (genes) were combined into a single "super locus". Contingency tables of the absence or presence (at least 1) of low frequency SNP variants in cases and controls for each region were constructed. Differences in the proportion of cases and controls carrying rare "super loci" were tested in a logistic regression model, adjusted for the effects of age and sex covariates in STATA v12. Genes were defined based on the co-ordinates of known genes and included 5,000 base pairs flanking either side of each gene's transcriptional start and stop site to include SNPs affecting regulatory elements.

3.2.5. Statistical methods

The association analysis consisted of two main stages:

The primary analysis was conducted separately in each participating study. Within each cohort “super loci” associations with CAD were analysed by logistic regression model with adjustment for age and sex and any other study-specific covariates where appropriate. The β coefficient reflects the magnitude of the accumulation of low-frequency/rare variants for the subjects in each cohort. Sensitivity analyses were undertaken (MAF 1-5%) in an effort to investigate whether MAF definition had an impact on the observed associations.

Meta-analysis of all individual study associations was conducted using Fisher’s method (based on P-values) separately in each ethnic group (Europeans and South Asians) using STATA v12. Bonferroni correction was used to correct for multiple testing. Cross-ethnicity replication of the statistically significant findings was then undertaken.

Single-variant analyses (logistic regression tests in PLINK) were also conducted for the LPA locus low-frequency/rare variants in each European study.

Conditional analysis was undertaken for LPA locus in order to investigate if the association signal resulted from the collective contribution of all low-frequency/rare variants in the locus or it was just driven by one leading SNP.

3.2.6. Power calculation

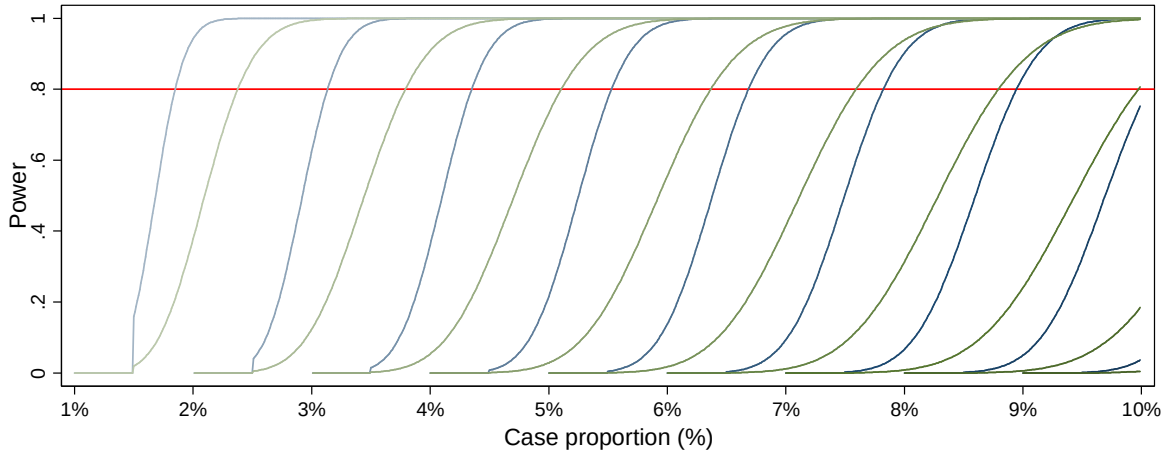
Using a Bonferroni adjusted alpha value (0.05/number of tests) to achieve a power of ≥80%, the detectable change in proportion of subjects with at least one rare allele (MAF≤3%) in the case group is listed in the Table 3.2.1 for a given proportion of subjects with at least one rare allele in the control group. For example in Europeans where 3% of controls have at least one rare allele, a significant difference with 80% power will be detected, if ≥4.36% (3%+1.36%) of the case group had at least one rare allele.

Table 3.2.1: Statistical power calculations for a fixed proportion of individuals in the control group in Europeans and South Asians

	Control%				
Detectable Delta	1	2	3	4	5
Europeans	0.85%	1.14%	1.36%	1.53%	1.69%
South Asians	1.38%	1.80%	2.11%	2.37%	2.60%

Detectable delta – Difference between CAD cases and controls that need to be observed for ≥80% power

Figure 3.2.1: Power curve estimates in Europeans and South Asians. The control proportion is fixed whereas the detectable case proportion with 80% power is indicated by the red line; Europeans – blue, South Asians – green



The power calculations were conducted in STATA v12.

3.3. Results

3.3.1. Characteristics of study cohorts

Following quality control procedures, meta-analyses of four European studies of 9,139 cases and 9,974 controls, and two South Asian studies of 4,448 cases and 4,313 controls, all part of the IBC 50K CAD Consortium (The IBC 50K CAD Consortium 2011), were performed. The characteristics of the European and South Asian individuals included in the analysis are shown in Table 3.3.1 and Table 3.3.2 respectively.

Table 3.3.1: Characteristics of European populations from IBC 50K CAD Consortium included in the analysis.

Study	Number of subjects	Cases/Controls	%male	Cases - age (years)
BHF-FHS	4621	2158/2463	63.3	40.8±7.7
BLOODOMICS Dutch	2684	1462/1222	72.6	48.8±12.0
BLOODOMICS German	3842	1910/1932	63.3	59.2±10.9
PennCATH	1516	489/1027	66.0	54.2±8.8
PROCARDIS	6450	3120/3330	59.2	61.0±8.7
In total	19,113	9,139/9,974	-	-

Data are counts and percentages or means and standard deviations

Table 3.3.2: Characteristics of South Asian populations from IBC 50K CAD Consortium included in the analysis.

Study	Number of subjects	Cases/Controls	%male	Cases - age (years)
LOLIPOP	5000	2592/2408	83.7	NA
PROMIS	3761	1856/1905	82.5	53.3±10.7
In total	8,761	4,448/4,313	-	-

Data are counts and percentages or means and standard deviations

There was a good numerical balance between patients with CAD and CAD-free controls included in the analysis. As expected, a majority of CAD patients were males, this was particularly apparent in the two South Asian populations.

After exclusion of common variants (here defined as those MAF>3%), the number of SNPs taken forward for low-frequency/rare variant focused analysis in each study ranged from 11,688-12,910. Application of the collapsing statistical method yielded a total of 1,908 regions (genes/super loci), of which 1,362 contained at least two low-frequency/rare variants. On average, 5 low-frequency/rare variants were within each locus in each study.

Of 1,362 identified loci with low-frequency/rare variants, 94 (4.9%) showed nominal association with CAD (P<0.05) in the meta-analysis of individuals of white European ancestry. The magnitude of identified associations between the loci and CAD was not dependent on the total number of low-frequency/rare variants in the locus. For example in European studies, nuclear receptor subfamily 4, group A, member 1 (*NR4A1*) gene with just 4 low-frequency/rare variants was associated with CAD at P=0.0006, compared to methionine sulfoxide reductase A (*MSRA*) gene that contained 24 to 28 low-frequency/rare variants and showed a P=0.0011 (Table 3.3.3).

3.3.2. Analysis of association between low-frequency/rare variants and CAD in European populations from IBC 50K CAD Consortium

Lipoprotein(a) *(LPA)* gene on chromosome 6 showed the most significant association with CAD in the meta-analysis of individuals of white European ancestry ($P=1.26x10^{-9}$). It was the only gene that retained its statistical association after Bonferroni correction for multiple testing (calculated at $P<2.62x10^{-5}$) (Table 3.3.3).

Table 3.3.3: Analysis of association between low-frequency/rare SNPs and CAD in Europeans – top association signals from the meta-analysis

Candidate Gene	BHF-FHS P-value	No of SNPs in locus	BLOODOMICS P-value	No of SNPs in locus	PennCATH P-value	No of SNPs in locus	PROCARDIS P-value	No of SNPs in locus	Meta-analysis P-value
LPA	0.0004	6	0.0005	6	0.0060	5	$5.22x10^{-6}$	5	$1.26x10^{-9}$
NR4A1	0.6815	4	0.0008	4	0.5946	4	0.0115	4	0.0006
TNFRSF11A	0.5504	5	$1.05x10^{-5}$	3	0.9668	3	0.9763	4	0.0010
MSRA	0.9131	25	0.0140	26	0.0061	28	0.1989	24	0.0011
RARB	0.9363	17	0.0008	9	0.0171	13	0.6076	12	0.0012
VEGFC	0.9354	9	$6.73x10^{-5}$	9	0.3570	10	0.8967	9	0.0015
PPARGC1B	0.0030	12	0.0448	12	0.0048	14	0.8451	11	0.0019
LCT	0.3950	5	$8.98x10^{-6}$	3	0.2648	5	0.5962	3	0.0021
TAP1	0.0072	4	0.4862	4	0.0100	3	0.9631	5	0.0021
TNFRSF10A	0.2495	5	0.0051	4	0.1094	5	0.2688	3	0.0021

LPA – lipoprotein(a) gene, *NR4A1* – nuclear receptor subfamily 4 group A member 1 gene, TNFRSF11A – tumor necrosis factor receptor superfamily member 11a NFKB activator gene, *MSRA* – methionine sulfoxide reductase A gene, *RARB* – retinoic acid receptor beta gene, *VEGFC* – vascular endothelial growth factor C gene, *PPARGC1B* – peroxisome proliferator-activated receptor gamma co-activator 1 beta gene, *LCT* – lactase gene, *TAP1* – transporter 1 ATP-binding cassette sub-family B (MDR/TAP) gene, *TNFRSF10A* – tumor necrosis factor receptor superfamily member 10a gene, P-value – level of statistical significance

In three studies (BHF-FHS, BLOODOMICS and PROCARDIS) there was a consistent over-representation of rare alleles of *LPA* in CAD cases compared to controls (Table 3.3.4). In PennCATH, the association was in the opposite direction. Overall, carriers of *LPA* low-frequency/rare alleles were more common in individuals with CAD than amongst CAD-free controls (5.9% versus 3.7%, respectively). The power calculation showed that there is reasonable power to detect differences between the two groups.

Table 3.3.4: Association between *LPA* gene and CAD in Europeans – analysis based on low frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
BHF-FHS	0.50±0.34	0.2208	0.7783	0.0004	154/1965 (7.3%)	102/2338 (4.2%)
BLOODOMICS	0.58±0.17	0.2562	0.9058	0.0005	127/2896 (4.2%)	59/2445 (2.4%)
PennCATH	-0.93±0.34	-1.6014	-0.2673	0.0060	11/478 (2.2%)	58/969 (5.6%)
PROCARDIS	0.62±0.14	0.3514	0.8821	$5.22x10^{-6}$	197/2930 (6.3%)	116/3214 (3.5%)
Overall	-	-	-	$1.29x10^{-9}$	489/8269 (5.9%)	335/8966 (3.7%)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low frequency/rare variants, "-" – individuals with no low frequency/rare variants.

Of six *LPA* variants in the locus, five were intronic and very rare (MAF<0.001) and one exonic and low frequency variant (MAF=0.026) (Table 3.3.5).

Table 3.3.5: Low frequency/rare SNPs in *LPA* gene in BHF-FHS

SNP	Position	Major/minor allele	MAF	Functional class
rs6922557	160878574	C/G	0.0004	Intronic
rs3798220	160881127	C/T	0.0264	Exonic (non-synonymous)
rs9347412	160886103	G/A	0.0001	Intronic
rs6922216	160929111	G/A	0.0007	Intronic
rs7755463	160932260	T/C	0.0014	Intronic
rs6455697	160988674	C/A	0.0008	Intronic

SNP – single nucleotide polymorphism

Cross-analysis of this locus in South Asians did not show evidence for its association with CAD (P=0.67) (Table 3.3.6).

Table 3.3.6: Association between LPA gene and CAD is South Asians – analysis based on low frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
LOLIPOP	0.06±0.12	-0.1674	0.2845	0.6116	178/2375 (7.0%)	151/2206 (6.4%)
PROMIS	-0.16±0.16	-0.4709	0.1415	0.2918	79/1777 (4.3%)	99/1806 (5.2%)
Overall	-	-	-	0.67	257/4152 (6.2%)	250/4012 (6.2%)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low frequency/rare variants, "-" – individuals with no low frequency/rare variants.

Further analysis of each single rare variant in *LPA* confirmed that only rs3798220 was associated with CAD, individually. No other single rare variant in the region was statistically significantly associated with CAD (Table 3.3.7).

Table 3.3.7: *LPA* individual SNP-based analysis of association with CAD

SNP	BHF-FHS P-value	BLOODMICS Dutch P-value	BLOODOMICS German P-value	PennCATH P-value	PROCARDIS P-value
rs6922557	0.3521	NA	NA	0.9993	NA
rs3798220	$1.31x10^{-7}$	0.0868	0.0015	0.0502	$1.07x10^{-6}$
rs9347412	0.3521	0.3605	0.1596	NA	NA
rs6922216	0.5202	0.6719	0.5706	0.9985	0.2453
rs7755463	0.2578	0.8038	0.5150	0.0879	0.4062
rs6455697	0.3438	NA	NA	0.2341	0.9786

SNP – single nucleotide polymorphism, P-value – level of statistical significance

The conditional analysis indicated that the association signal identified between *LPA* and CAD was mainly driven by rs3798220.

3.3.3. Analysis of association between low-frequency/rare variants and CAD in South Asian populations from IBC 50K CAD Consortium

A meta-analysis of two South Asian populations with a total of 4,448 CAD cases and 4,313 CAD-free controls was also undertaken. A total of 1,880 genes/loci; each with at least 2 low-frequency/rare variants were identified in this analysis. Of those, 86 (4.6%) showed at least nominal association with CAD (P<0.05). Coagulation factor VII – serum prothrombin conversion accelerator (*F7*) and coagulation factor X (*F10*) genes on chromosome 13 and TNF receptor-associated factor 2 (*TRAF2*) gene on chromosome 9 retained their statistical significance of association with CAD in the meta-analysis after multiple testing correction (Table 3.3.8). The identified associations were driven by signals from LOLIPOP Study; there were no significant differences in the collective accumulation of rare variants among cases and controls in PROMIS study (Table 3.3.9, Table 3.3.10, and Table 3.3.11).

Table 3.3.8: Analysis of association between low-frequency/rare SNPs and CAD in South Asians – top association signals from the meta-analysis

Candidate Gene	LOLIPOP P-value	No of SNPs in locus	PROMIS P-value	No of SNPs in locus	Meta-analysis P-value
F10	$7.88x10^{-12}$	13	0.1287	15	$1.52x10^{-11}$
F7	$6.57x10^{-12}$	15	0.5808	14	$5.47x10^{-11}$
TRAF2	$1.09X10^{-8}$	24	0.9618	25	$1.09x10^{-7}$
MYBPC2	0.6140	9	$9.29x10^{-5}$	9	0.0003
IL8RA	0.0261	2	0.0160	5	0.0021
APOE	0.0490	12	0.0086	12	0.0021
XRCC1	0.6888	5	0.0013	4	0.0043
DRD2	0.2047	5	0.0049	5	0.0047
ADRBK1	0.0074	6	0.1611	7	0.0054
LIPA	0.1426	3	0.0007	4	0.0058

F10 – coagulation factor X gene, F7 – coagulation factor VII (serum prothrombin conversion accelerator) gene, TRAF2 – TNF receptor-associated factor 2 gene, *MYBPC2* – myosin binding protein C fast type gene, *IL8RA* – interleukin 8 receptor A gene, *APOE* – apolipoprotein E gene, *XRCC1* – Xray repair complementing defective repair in Chinese hamster cells 1 gene, *DRD2* – dopamine receptor D2 gene, *ADRBK1* – adrenergic, beta, receptor kinase 1 gene, *LIPA* – lipase A lysosomal acid cholesterol esterase gene, P-value – level of statistical significance.

Table 3.3.9: Association between *F10* gene and CAD in South Asians – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
LOLIPOP	0.78±0.11	0.55	1.00	$7.88x10^{-12}$	271/2552 (10.6)	122/2357 (5.2)
PROMIS	0.19±0.13	-0.06	0.44	0.1287	144/1856 (7.8)	129/1905 (6.8)
Overall	-	-	-	$1.52x10^{-11}$	415/4408 (9.4)	251/4262 (5.9)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Table 3.3.10: Association between F7 gene and CAD in South Asians – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
LOLIPOP	0.75±0.11	0.54	0.97	$6.57x10^{-12}$	286/2553 (11.2)	133/2357 (5.6)
PROMIS	0.07±0.13	-0.18	0.32	0.5808	139/1856 (7.5%)	140/1905 (7.3%)
Overall	-	-	-	$5.47x10^{-11}$	425/4409 (9.6)	273/4262 (6.4)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Table 3.3.11: Association between TRAF2 gene and CAD in South Asians – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
LOLIPOP	0.48±0.83	0.31	0.64	$1.09x10^{-8}$	446/2553 (17.5)	275/2357 (11.7)
PROMIS	0.01±0.11	-0.22	0.23	0.96	167/1856 (9.0)	177.1905 (9.3)
Overall	-	-	-	$1.09x10^{-7}$	613/4409 (13.9)	452/4262 (10.6)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Cross-replication of these genes in the European population did not provide evidence for association (*F10*: P=0.18, *F7*: P=0.11, and *TRAF2*: P=0.5) (Tables 3.12, Table 3.13 and Table 3.14).

Table 3.3.12: Association between *F10* gene and CAD in Europeans – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
BHF-FHS	-0.22±0.23	-0.68	0.23	0.33	40/2120 (1.9)	53/2440 (2.2)
BLOODOMICS	0.09±0.20	-0.29	0.48	0.64	66/3023 (2.2)	52/2503 (2.1)
PennCATH	0.21±0.34	-0.45	0.88	0.53	14/489 (2.9)	28/1027 (2.7)
PROCARDIS	0.43±0.20	0.04	0.82	0.30	76/3127 (2.4)	58/3331 (1.7)
Overall	-	-	-	0.18	196/8759 (2.2)	191/9301 (2.1)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Table 3.3.13: Association between *F7* gene and CAD in Europeans – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
BHF-FHS	0.14±0.16	-0.18	0.46	0.39	100/2120 (4.7)	95/2439 (3.9)
BLOODOMICS	0.72±0.13	-0.18	0.32	0.57	160/3022 (5.3)	129/2503 (5.2)
PennCATH	0.30±0.28	-0.26	0.85	0.30	22/489 (4.5)	36/1027 (3.5)
PROCARDIS	0.24±0.15	-0.06	0.53	0.12	135/3127 (4.3)	120/3331 (3.6)
Overall	-	-	-	0.11	417/8758 (4.8)	380/9300 (4.1)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Table 3.3.14: Association between *TRAF2* gene and CAD in Europeans – analysis based on low-frequency/rare variants

Study population	B±SE	95% LCI	95% UCI	P-value	Cases +/- (%)	Controls +/- (%)
BHF-FHS	-0.01±0.08	-0.17	0.15	0.92	440/2120 (20.8)	507/2440 (20.8)
BLOODOMICS	0.01±0.07	-0.13	0.16	0.87	528/3023 (17.5)	426/2504 (17.0)
PennCATH	-0.06±0.14	-0.34	0.22	0.69	94/489 (19.2)	214/1027 (20.8)
PROCARDIS	-0.11±0.07	-0.25	0.04	0.14	569/3127 (18.2)	652/3331 (19.6)
Overall	-	-	-	0.52	1631/8759 (18.6)	1799/9302 (19.3)

P value – level of statistical significance, B – beta coefficient, SE – standard error, LCI – lower confidence interval, UCI – upper confidence interval, "+" – number (%) of carriers of low-frequency/rare variants, "-" – individuals with no low-frequency/rare variants.

Further sensitivity analyses using different minor allele frequency thresholds were undertaken to examine whether selected low frequency variant definition had an impact on the observed associations. Using a low threshold in the analysis (MAF<1%) led to elimination of all four significant loci associations reported above (Table 3.3.15, Table 3.3.16, Table 3.3.17, Table 3.3.18).

Table 3.3.15: Sensitivity analysis between *LPA* gene and different minor allele frequency thresholds in Europeans

Study	MAF-1%		MAF-2%		MAF-3%		MAF-4%		MAF-5%	
	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	Rare SNPs
BHF-FHS	0.1354	5	0.1354	5	0.0004	6	0.0122	7	0.0122	7
BLOODOMICS	0.6662	5	0.0004	6	0.0005	6	0.1730	7	0.1730	7
PennCATH	0.1374	4	0.0060	5	0.0060	5	0.1435	6	0.1435	6
PROCARDIS	0.4421	4	0.4421	4	$5.22x10^{-6}$	5	0.0233	6	0.0233	6
Meta-analysis P-value	0.2979	-	0.0508	-	$1.26x10^{-9}$	-	0.0043	-	0.0043	-

MAF – minor allele frequency, P-value – level of statistical significance

Table 3.3.16: Sensitivity analysis between *F10* gene and different minor allele frequency thresholds in South Asians

Study	MAF-1%		MAF-2%		MAF-3%		MAF-4%		MAF-5%	
	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus
LOLIPOP	0.3213	12	0.3213	12	$7.88x10^{-12}$	13	$2.89x10^{-12}$	14	$2.89x10^{-12}$	14
PROMIS	0.2164	14	0.2164	14	0.1287	15	0.0307	16	0.0307	16
Meta-analysis P-value	0.6173	-	0.6173	-	$1.52x10^{-11}$	-	$7.24x10^{-6}$	-	$7.24x10^{-6}$	-

MAF – minor allele frequency, P-value – level of statistical significance

Table 3.3.17: Sensitivity analysis between *F7* gene and different minor allele frequency thresholds in South Asians

Study	MAF-1%		MAF-2%		MAF-3%		MAF-4%		MAF-5%	
	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus
LOLIPOP	0.6095	14	0.6095	14	$6.57x10^{-12}$	15	$1.43x10^{-5}$	16	$1.43x10^{-5}$	16
PROMIS	0.8117	13	0.8117	13	0.5808	14	0.7640	15	0.7640	15
Meta-analysis P-value	0.7643	-	0.7643	-	$5.47x10^{-11}$	-	$7.53x10^{-5}$	-	$7.53x10^{-5}$	-

MAF – minor allele frequency, P-value – level of statistical significance

Table 3.3.18: Sensitivity analysis between *TRAF2* gene and different minor allele frequency thresholds in South Asians

Study	MAF-1%		MAF-2%		MAF-3%		MAF-4%		MAF-5%	
	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus	P-value	No of SNPs in locus
LOLIPOP	0.0543	20	0.1996	23	$1.09x10^{-8}$	24	$6.26x10^{-5}$	25	$6.26x10^{-5}$	25
PROMIS	0.6692	22	0.9618	25	0.9618	25	0.7433	26	0.7433	26
Meta-analysis P-value	0.1812	-	0.4110	-	$1.09x10^{-7}$	-	0.0005	-	0.0005	-

MAF – minor allele frequency, P-value – level of statistical significance

3.4. DISCUSSION

Analysis of associations between low-frequency/rare variants and complex disease is a challenging task. GWA studies concentrated on the investigation of common variation therefore discarding a significant proportion of data collected - a common quality control criterion was the exclusion of SNPs with low MAF (Barrett and Cardon, 2006). Because of the small number of observations for any given rare allele the power to detect its association with a phenotype is a major limiting factor in genetic analysis. However, the investigation of this previously discarded information may illuminate the potential of low-frequency/rare variants as a complementary approach to the primary GWA studies of common variants.

Analysis of low-frequency/rare variants using conventional genome-wide SNP arrays is limited by a number of factors. Firstly, such arrays contain only a small proportion of rare variants that actually exist; this is due to the array design being motivated by the emphasis of common susceptibility variants (Evans *et al.* 2008). Secondly, the genotype quality of rare variants typed on GWA platforms tends to be low, mainly driven by poor automated clustering and genotype calling.

The gold standard strategy for investigating low-frequency/rare variants is the use of sequencing data that capture a much higher proportion of “rare” genetic variation (Morris and Zeggini, 2010). However, despite the increasing availability of high-throughput sequencing technologies nowadays, these approaches were limited when this project was conducted. The best alternative was the utilisation of existing data that were generally excluded in GWA studies.

Though not providing genome-wide coverage, the HumanCVD BeadChip targets many low-frequency/rare SNPs (Keating *et al.* 2008) and as a result had an advantage over

standard GWA platforms to detect association with low-frequency/rare alleles (MAF<5%). This array was one of the first disease-specific custom arrays with highly focused rare variant content for CAD to be used on a large scale. An exploratory analysis of SNPs represented on the IBC 50K array showed that low-frequency/rare variants make up a large proportion of all polymorphisms and this was what motivated this analysis. However, it should be noted that the primary purpose of this analysis was signal detection rather than accurate effect estimation.

Previously, the IBC 50K CAD Consortium examined the association between low frequency variants (MAF=1-5%) and CAD. Although previously reported associations of lower frequency variants in *LPA* and *PCSK9* with CAD risk were confirmed, no other strongly associated variants in the 1-5% range or enrichment of low frequency variants amongst SNPs that showed nominal association with CAD were identified (The IBC 50K CAD consortium, 2011).

To overcome the power issues associated with testing rare variants individually, sets of low-frequency/rare variants were collapsed into a single group and tested collectively as "aggregate genotype" for frequency differences between cases and controls. It was hypothesised that patients with CAD may exhibit over-representation of low frequency/rare alleles compared to controls. The collapsing method examined ~1,900 genes and yielded some interesting findings.

In the meta-analysis of European populations, one locus – *LPA* – showed over-accumulation of low-frequency/rare variants in CAD cases compared to controls, a previously established CAD gene. Lipoprotein(a) [Lp(a)], is a lipoprotein particle that consists of an apolipoprotein(a) [apo(a)] molecule covalently linked by a disulfide bond to the apolipoprotein B-100 (apoB-100) component of LDL-like particle (Li *et al.*

2011). The *LPA* gene, which encodes apo(a), has evolved from a duplication of plasminogen (*PLG*) gene (McLean *et al.* 1987). The apo(a) has a protease-like domain and multiple kringle domains. Lipoprotein(a) levels in human plasma undoubtedly have the strictest genetic control of all lipoproteins; more than 90% of the variance of Lp(a) concentrations is explained by genetics (Snieder *et al.* 1997; Snieder *et al.* 1999). The major genetic determinant of Lp(a) levels resides in the *LPA* gene itself. Elevated levels of serum Lp(a) lead to premature CAD (Rhoads *et al*. 1986; Li *et al.* 2011). So far, multiple SNPs in the LPA gene on the long arm of chromosome 6 were associated with Lp(a) levels (Ober *et al.* 2009). Amongst them, low-frequency SNP in the LPA region - rs3798220 (MAF~3%) was associated with CAD in Caucasian in multiple studies (Chasman *et al*. 2009; Clarke *et al.* 2009; Luke *et al.* 2007; Schunkert *et al.* 2011; Shiffman *et al.* 2008; Shiffman *et al.* 2010). This polymorphism results in an amino acid residue substitution (an isoleucine to methionine) at position 4399 of apo(a) (Li *et al*. 2011). The effect of this variant on the risk of CAD is correlated with the effects on the Lp(a) lipoprotein level (Clarke et al. 2009). This linear dose-response relationship of rs3798220 and both the Lp(a) lipoprotein level and the risk of CAD supports a causal role of an elevated plasma level of Lp(a) lipoprotein in the risk of CAD (Clarke *et al*. 2009).

In the analysis conducted in European studies, the association signal in *LPA* was mainly driven by the rs3798220 mainly due to its low-frequency (MAF~2%) compared to other variants with very rare (MAF<0.001) ones. The *LPA* did not show association in South Asians; this may be due to its low MAF in this population (MAF<0.01). It is true that interesting and informative variants often segregate in a population-specific manner. For example, a nonsense variant in *PCSK9* that significantly affects LDL-cholesterol levels

reducing CAD risk has frequency of 0.8% in African-ancestry individuals but is almost absent in European-ancestry samples (Cohen *et al.* 2006*a*).

Similar pattern of association was observed in the Dallas Heart Study data where an association between hepatic fibrinogen/angiopoietin-related protein gene (*ANGPTL4*) and triglycerides was mainly driven by a low-frequency variant within the European population (MAF~3%) (Liu and Leal, 2012).

In the meta-analysis of South Asian studies, 3 genomic loci - *F10, F7* and *TRAF2* - showed over-accumulation of low-frequency/rare variants in CAD cases compared to CAD-free controls. The protein encoded by *TRAF2* gene is a member of the TNF receptor associated factor TRAF protein family (Rothe *et al*. 1994). TRAF proteins bind to, and mediate the signal transduction from members of the TNF receptor superfamily. This protein directly interacts with TNF receptors, and forms a heterodimeric complex with its partner – encoded by its sister gene – TRAF1 (Rothe *et al.* 1994). The TRAF2 protein is needed for TNF-alpha-mediated activation of MAPK8/JNK and NF-kappaB pathways (Blackwell et al. 2009). The protein complex formed by TRAF2 and TRAF1 interacts with the inhibitor-of-apoptosis proteins (IAPs), and functions as a mediator of the anti-apoptotic signals from TNF receptors (Vince *et al.* 2009). The detected association signal is a novel finding; there is no prior published data on the role of TRAF2 in genetic predisposition of CAD.

F10 gene encodes the vitamin K-dependent coagulation factor X of the blood coagulation cascade. Factor X undergoes multiple processing before its preproprotein is converted to a mature two-chain form activated by factor IXa, or by factor VIIa (Davie *et al.* 1991). The activated Xa acts as a converter of prothrombin to thrombin in the

presence of factor Va, Ca^{+2}, and phospholipid during blood clotting (Kamata *et al.* 1998).

F7 gene encodes coagulation factor VII which is a vitamin K-dependent factor essential for hemostasis (Mackman *et al.* 2007). This factor circulates in the blood in a zymogen form, and is converted to an active form by factor IXa, factor Xa, factor XIIa, or thrombin by minor proteolysis (Chen, 2013). In the presence of factor III and calcium ions, the activated VII then further activates the coagulation cascade by converting factor IX to factor IXa and/or factor X to factor Xa.

Coagulation activation plays a key role in thrombus formation and variation and its factors have been associated with the risk of CAD (Mo *et al*. 2011). For example, activation of the extrinsic coagulation pathway plays a key role in hemostasis, and as a result factor VII contributes to the occurrence of thrombotic events. High factor VII levels might disproportionately enhance the coagulation cascade at the time of plaque rupture, which could explain the apparently differential association with fatal and nonfatal coronary events (Mo *et al*. 2011). A number of reports suggested that increased coagulation factor VII activity is a risk factor for CAD (Meade *et al.* 1986; Noto *et al.* 2002; Cai *et al*. 2000).

Several studies have examined the association between polymorphisms in the *F7* and the risk of CAD but they have been inconclusive. Some studies concluded a protective role for SNP variants and the risk of MI among patients with CAD (Di Castelnuovo *et al.* 2000; Iacoviello *et al.* 1998; Girelli *et al*. 2000) and other failed to confirm such associations (Wang *et al*. 1997; Tamaki *et al*. 1999; Ardissimo *et al*. 1999; Feng *et al*. 2000). As a result the exact biological role of these polymorphisms remains to be

determined. The association between F7 polymorphisms and CAD has been shown to vary in different ethnicities (Mo *et al.* 2011).

In this analysis conducted in Asian populations, the uncovered association signal in the meta-analysis was driven clearly by one of the studies. Indeed, the identified associations were driven by signals from LOLIPOP study. One may argue that these differences in association of low frequency/rare variants with CAD between LOLIPOP and PROMIS could be due, at least in part, by differences in ascertainment/origin of both populations. Whereas, LOLIPOP study was of Indian origin and recruited in the UK, PROMIS study was of Pakistani origin and collected in Pakistan. It should be also noted that LOLIPOP is a larger study and as a result has more power to detect low frequency/rare variants. The differences in genetic architecture of rare variants between both populations cannot be excluded. Indeed, rare alleles have typically arisen recently and tend to have higher variation in different geographic distributions than more common variants that are typically evolutionarily older (Nelson *et al*. 2012). Therefore, further studies in CAD should minimise the potential confounding effects by minimising the sources of genetic heterogeneity.

The statistical method relies on a pre-specified threshold for inclusion of alleles into a set of variants considered as “rare” (Morris and Zegini, 2010). Thus, it is perhaps not surprising that using different arbitrary thresholds influence the association results. Unfortunately, there is little guidance in this area and allele frequency thresholds of 1% or 5% are commonly chosen (Manolio *et al*. 2009). These cut-offs are subjective and dependent on the spectrum of the variant frequency within a locus. For example, if the allele frequencies of the variants are relatively rare, the 1% MAF cut-off threshold may be used. On the other hand, if different allele frequencies are observed, it would be

better to use several MAF cut-off thresholds to classify the variants into multiple groups.

In each of the top signals, association do not seem to be driven by the collective contribution of very rare SNPs. For example, the strong signal is largely due to MAF=2-3% and the inclusion of only rarer SNPs or low frequency SNPs reduces the power and increases the noise.

In addition, different genes may have very different relationships between allele frequency and functional effects. Some genes may harbour functional alleles at higher frequencies, whereas other genes may have only private functional variants. As a result the value of the optimal allele-frequency threshold often varies considerably (Price *et al.* 2010). This analysis showed how important the MAF minimum and maximum limits are and that a variety of frequency cut-offs should be considered when analysing low-frequency and rare variants.

Another limitation of the strategy used here, is that the individual effects of each rare variant collapsed into aggregate genotype on phenotype may differ not only in terms of the magnitude but also – direction (neutral, protective or detrimental for a given disease trait) and as a result, the estimates of the average genetic effects will be affected (Liu *et al.* 2012). On the other hand, the average genetic effect variance that is explained in the aggregate analysis is always no greater than the true locus-specific genetic variance (Liu and Leal, 2012).

One should also acknowledge that the method used here does not calculate exactly how many low-frequency/rare variants each individual has in a gene. It would be interesting to find out if subjects with several rare variants were over-represented among individuals with CAD compared to controls. Any given rare variant in affected individuals is not necessarily sufficient to cause disease but rare variants contribute to

the heterogeneity observed among affected individuals. Studies using high-throughput next generation sequencing are required to determine whether these associations extend to additional CAD-associated genes.

Another limitation is that collapsing method can be seriously impaired by misclassification of collapsing regions (Do *et al.* 2012). Collapsing methods can focus on gene-centric bins, known conserved non-coding regions or functional pathways as the functional unit for analysis. As a result it depends on how many low-frequency/rare variants are found in these regions. Finally, it is likely that the nature of the relationship between rare variants and a phenotype varies from gene to gene, even in different parts of a gene.

Recently several factors combined together made the direct investigation of rare variants possible. First, the size of GWA studies and meta-samples has increased, approaching cohort sizes of 100,000 through large-scale international collaborations, strengthens the power. Second, the ascertainment of many rare variants through the 1000 Genomes Project (The 1000 Genomes Project Consortium *et al.* 2010) has enabled imputation of millions of rare and low-frequency variants and led to the development of a new generation of low-cost genotyping platforms that interrogate rare variants directly. Third, the decline in the cost of sequencing technologies has enabled large-scale sequencing studies to be performed, which, in principle, allow the detection of all variants in a sample. The emerging wealth of re-sequencing data yet to be generated for CAD will shed light on the true contribution of rare variants to CAD disease risk. Finally, the recent development and improvement of statistical tests for association studies for rare variants (Ionita-Laza *et al.* 2011; Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Mukhopadhyay *et al.* 2010; Neale *et al.*

2012) provides power to detect genes or pathways harbouring multiple rare variants for which individually there would be low power to detect association.

CHAPTER 4

GENETIC ARCHITECTURE OF PSEUDOAUTOSOMAL REGIONS AND SUSCEPTIBILITY TO CORONARY ARTERY DISEASE

4.1. Introduction

4.1.1. The X and Y chromosomes

The human X and Y sex chromosomes are morphologically and genetically different (Flaquer *et al.* 2009). The X chromosome is large, generally more euchromatic (six-times longer) and has many more genes than the Y chromosome (Ross *et al.* 2005). Both sex chromosomes originate from an ancestral pair of autosomes (Charlesworth *et al.* 2005), which during mammalian evolution lost homology further to acquiring of the sex determining locus by the Y chromosome (Charlesworth, 1991). Unlike autosomal pairs of chromosome, X and Y do not exchange the genetic information during male meiosis along the majority of their length; the only parts where the recombination occurs are the terminal portions of both chromosomes called pseudoautosomal regions (PARs) (Graves *et al.* 1998).

4.1.2. The human pseudoautosomal regions (PARs)

PARs are located on the tips of both the short (p) and long (q) arms of the X and Y chromosomes (Cooke *et al.* 1985; Rappold, 1993). PARs show X-Y sequence homology, act like autosomes during meiosis and are inherited in an autosomal rather than sex-linked manner. In humans, PARs represent about 2% of the X and ~5% of the Y chromosomal sequence length (Blaschke and Rappold, 2006). PARs are two remarkable regions in the mammalian genome that have evolved more recently than autosomes (Graves *et al.* 1998; Blaschke and Rappold, 2006).

4.1.2.1. Pseudoautosomal region 1 (PAR1)

Mapping to the short-arm region of sex chromosomes, PAR1 has a physical length of ~2.7Mb (Ross *et al.* 2005). It originated from a part of autosomal region transferred to the mammalian X and Y chromosomes 100-150 million years ago (Graves *et al.* 1998). It shows sequence homology to PARs of several species, including great apes and Old World monkeys (Ciccodicola *et al*. 2000; Charchar *et al*. 2003). Compared to the X chromosome, PAR1 exhibits significantly higher GC content and has a higher proportion of minisatellite repeats and other duplicated structures (Ried *et al.* 1998). It also has 4-5 times elevated Alu repeat content when compared to the rest of the X chromosome (Blaschke and Rappold, 2006).

In humans, pairing and cross-over between the X and Y is limited to the PAR1 region (Cooke *et al.* 1985; Simmler *et al.* 1985). Family data are consistent with a single obligatory cross over during each male meiosis in PAR1 (Rouyer *et al.* 1986; Lien *et al*. 2000). However, there is a gradient of sex linkage along PAR1 length. Alleles in close proximity to MSY undergo recombination less frequently and show stronger linkage to the Y chromosome than those mapping to the telomeric portion of PAR1 (Blaschke and Rappold, 2006). Deletion of PAR1 results in failure of X-Y pairing and male sterility (Mohandas *et al*. 1992).

4.1.2.2. Pseudoautosomal region 2 (PAR2)

A second pseudoautosomal region, Xq/Yq PAR, was discovered only in the mid-90s as a part of the project on X-chromosome mapping (Freije and Schlessinger, 1992). Located within the distal portion of the long arm region of both sex chromosomes, PAR2 is short – only ~0.33Mb in length (Ross *et al.* 2005). PAR2 has a much shorter evolutionary history than PAR1 and is human specific (Blaschke and Rappold, 2006).

This peculiar part of the human genome originated from an L1-mediated ectopic recombination event that transferred the subtelomeric region of the X onto Y chromosome after the divergence of human and chimpanzee lineages about 6 million years ago (Kvaløy *et al*. 1994; Ciccodicola *et al.* 2000; Charchar *et al.* 2003). As a result, PAR2 contains sequences that had been earlier recovered from both the X and Y and shows recombination over its entire extent (Freije *et al.* 1992). It should be noted that PAR2 in many aspects is different from PAR1. PAR2 is neither necessary nor sufficient for sex-chromosome segregation in male meiosis (Li and Hamer, 1995). However; the region still exhibits a six-fold higher recombination frequency when compared with the average rate of the remaining of the X chromosome (Li and Hamer, 1995).

4.1.3. Recombination rates across PARs in males and females

The most extreme example of gender differences in recombination rates across the human genome is on the sex chromosomes. In the male germline, recombination between X and Y is almost entirely restricted to PAR1 and a lesser extent - PAR2, whereas in the female germline the two X chromosomes can recombine anywhere along their entire length. A high recombination rate in female meiosis is seen only at the Xp telomere within the first 100kb (Flaquer *et al.* 2009). Within this telomeric region no sex-specific recombination rates are observed, resulting in similar genetic distances in both males and females (Flaquer *et al.* 2009). After 100kb, genetic distances are becoming increasingly sex-specific as they approach the pseudoautosomal boundary. This results in marked sex-specific differences in PAR1 genetic map length, which is 50cM in males and around one tenth of this in females (Rouyer *et al*. 1986; Page *et al.*

1987). PAR1 is thus a male-specific recombination hot domain with a mean crossover frequency 20 times higher than the genome average (Rappold, 1993).

4.1.4. Genes within the PARs

Currently there are 50 genes mapping to PARs. Apart from protein coding genes, PARs contain pseudogenes and non-coding RNA genes (miRNAs and lincRNAs). The characteristics of PAR1 and PAR2 gene are listed in Table 4.1.1 and Table 4.1.2, respectively.

PAR1 contains 16 protein-coding genes. Together with the 3 annotated protein coding genes in PAR2, 19 protein-coding genes lie entirely within these recombining regions of the sex chromosomes. Several genes, including *PLCXD1*, *P2RY8* and *DHRSX* have been identified only recently. Many novel transcripts were recently mapped to the PAR1 region (Ross *et al*. 2005). The biological function and the role in susceptibility to disease is not well described for a majority of these genes.

Genes in close proximity to the pseudoautomal boundary have unique features compared to the remaining of PAR1 genes. An example is a group of genes in *XG* blood regions, close to the pseudoautosomal boundary (Ellis *et al*. 1994a; Ellis *et al* 1994b). The first four exons are located within PAR1 and are subjected to high recombination rate, whereas the nine downstream exons lie on the X-specific portion of the chromosome.

Table 4.1.1: Genes within human PAR1 – general characteristics

	Gene Symbol	Gene name	Start (bp)	End (bp)	Length (bp)	Gene biotype	Function	N of transcripts
1	NCRNA00108	-	170410	172712	2302	Pseudogene	Unkown	1
2	PLCXD1	phosphatidylinositol-specific phospholipase C, X domain containing 1	192989	220023	27034	Protein coding	Lipid metabolism Inflammation	11
3	GTPBP6	GTP binding protein 6 (putative)	220025	230886	10861	Protein coding	Unknown	3
4	LINC00685	long intergenic non-protein coding RNA 685	281725	282586	861	RNA gene	Unknown	1
5	PPP2R3B	protein phosphatase 2, regulatory subunit B	294698	347690	52992	Protein coding	DNA replication	10
6	AL732314.1	-	425316	425416	100	Novel miRNA	-	1
7	FABP5P13	fatty acid binding protein 5 pseudogene 13	484510	484837	327	Pseudogene	Unknown	1
8	KRT18P53	-	505971	506087	116	Pseudogene	- Unknown	1
9	SHOX	short stature homeobox	585079	620146	35067	Protein coding	Transcription	7
10	RP11-309M23.1	-	950956	955100	4144	lincRNA	-	1
11	RPL14P5	-	969238	970836	1598	Pseudogene	-	1
12	CRLF2	cytokine receptor-like factor 2	1314890	1331616	16726	Protein coding	Immunity	4
13	CSF2RA	colony stimulating factor 2 receptor, alpha (granulocyte-macrophage)	1387693	1429274	41581	Protein coding	Immunity	21
14	BX649553.4	-	1410987	1411060	73	Novel miRNA	-	1
15	BX649553.2	-	1412508	1412582	74	Novel miRNA	-	1
16	MIR3690	-	1412811	1412885	74	Known miRNA	-	1
17	BX649553.1	-	1413025	1413099	74	Novel miRNA	-	1
18	RNA5SP498	-	1419149	1419268	119	Known rRNA	-	
19	RN7SL355P	-	1437739	1438052	313	Known miscRNA	-	1
20	IL3RA	interleukin 3 receptor, alpha	1455509	1501578	46069	Protein coding	Immunity	3
21	SLC25A6	solute carrier family 25 (mitochondrial carrier; adenine nucleotide	1505045	1511617	6572	Protein coding	Immunity	3

		translocator), member 6						
	Gene Symbol	**Gene name**	**Start (bp)**	**End (bp)**	**Length (bp)**	**Gene biotype**	**Function**	**N of transcripts**
22	LINC00106	long intergenic non-protein coding RNA 106	1515320	1518295	2975	RNA gene	Unknown	2
23	ASMTL-AS1	*ASMTL* antisense RNA 1	1520662	1532921	12259	RNA gene	Unknown	5
24	ASMTL	acetylserotonin O-methyltransferase-like	1522032	1572655	50623	Protein coding	Unknown	7
25	P2RY8	purinergic receptor P2Y, G-protein coupled, 8	1581465	1656000	74535	Protein coding	Unknown	2
26	AKAP17A	A kinase (PRKA) anchor protein 17A	1710486	1721407	10921	Protein coding	mRNA processing	3
27	ASMT	acetylserotonin O-methyltransferase	1733894	1761974	28080	Protein coding	Serotonine metabolism, melatonin synthesis	4
28	RP13-297E16.3	-	1746639	1755356	8717	Processed transcript	-	1
29	RP13-297E16.4	-	1851477	1874878	23401	Novel lincRNA	-	2
30	RP13-297E16.5	-	1886240	1887669	1429	Novel lincRNA	-	1
31	DHRSX	dehydrogenase/reductase (SDR family) X-linked	2137557	2420846	283289	Protein coding	Unknown	7
32	DHRSX-IT1	*DHRSX*intronic transcript 1 (non-protein coding)	2252336	2254451	2115	RNA gene	Unknown	1
33	ZBED1	zinc finger, BED-type containing 1	2404455	2419008	14553	Protein coding	Transcription	4
34	RP11-325D5.3	-	2405023	2407012	1989	Processed transcript	-	1
35	CD99P1	CD99 molecule pseudogene 1	2527389	2575270	47881	Pseudogene	Unknown	5
36	LINC00102	long intergenic non-protein coding RNA 102	2531029	2533388	2359	RNA gene	Unknown	1
37	CD99	*CD99* molecule	2609220	2659350	50130	Protein coding	Immunity	9
38	XG	*Xg* blood group	2670091	2734539	64448	Protein coding	Immunity	6

miRNA – miCRNA precursors, lincRNA - long intergenic non-coding RNAs, miscRNA- miscellaneous other RNA; Start and end positions are based on build 37. Information on number of gene transcripts and their lengths were found in Ensembl Genome Browser.

Table 4.1.2Genes within human PAR2 – general characteristics

	Gene Symbol	Gene name	Start (bp)	End (bp)	Length (bp)	Gene biotype	Function	N of transcripts
1	SPRY3	Sprouty homolog 3 (drosophila)	154997474	155012121	14647	Protein coding	Unknown	1
2	AMDP1	adenosylmethionine decarboxylase pseudogene 1	155058248	155059239	991	Pseudogene	Unknown	1
3	DPH3P2	DPH3, KTI11 homolog (S. cerevisiae) pseudogene 2	155105299	155105548	249	Pseudogene	Unknown	1
4	VAMP7	vesicle-associated membrane protein 7	155110956	155173433	62477	Protein coding	Transportation	6
5	TCEB1P24	transcription elongation factor B (SIII), polypeptide 1 pseudogene 24	155208657	155208990	333	Pseudogene	Unknown	1
6	TRPC6P	transient receptor potential cation channel, subfamily C, member 6 pseudogene	155215035	155215914	879	Pseudogene	Unknown	1
7	IL9R	interleukin 9 receptor	155227246	155240482	13236	Protein coding	Immunity	6
8	AJ271736.10	-				Processed transcript	-	1
9	WASIR1	WASH and IL9R antisense RNA 1	155244288	155246502	2214	RNA gene	-	1
10	WASH6P	WAS protein family homolog 6 pseudogene	155250491	155255375	4884	Pseudogene	Unknown	16
11	DDX11L16	DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11 like 16	155255323	155257848	2525	Pseudogene	Unknown	2

Start and end positions are based on build 37. Information on number of gene transcripts and their lengths were found in Ensembl Genome Browser.

4.1.5. PARs and human disease

Except of the short stature homeobox gene (*SHOX*), there are no known phenotypic traits firmly attributed to variations in PAR1genes. *SHOX* encodes a transcription factor of the homeodomain class and is located in PAR1 (Braschke and Rappold, 2006). Mutations of this gene are associated with various growth deficit disorders, including isolated short stature (Binder, 2011), Leri-Weill syndrome (Evers *et al*. 2011) and Langer syndrome (Belin *et al.* 1998). Patients with some loss-of-function mutations (including deletions, missense and nonsense SNPs) of the *SHOX* gene have short stature (without any other apparent abnormalities). Leri-Weill dyschondrosteosis (affected individuals have shortening and deformities of the bones of the forearm) results from mutations in a single copy of the *SHOX* gene (Belin *et al*. 1998; Shears *et al.* 1998). Langer disorder known as Langer mesomelic dwarfism is also caused by *SHOX* mutations (Belin *et al* 1998). *SHOX* mutations account for ~2 to 7% of isolated short stature patients and 50 to 90% of Leri-Weill syndrome patients (Rappold *et al.* 2002; Schiller *et al.* 2000; Binder *et al*. 2003; Blaschke and Rappold, 2006).

Two of PAR2 genes have been hypothesised to play a role in susceptibility to complex human diseases. Vesicle-associated membrane protein 7 gene (*VAMP7*) is located in PAR2 but differs from most other PAR genes in that sense that it undergoes both X and Y inactivation (Helena Mangs and Morris, 2007). Evolutionarily, *VAMP7* is highly conserved. Its protein product is a member of the synaptobrevins – molecules implicated in cellular exocytosis (Filipinni *et al*. 2001). A role of *VAMP7* gene in bipolar affective disorder has been suggested (Saito *et al.* 2000). Another PAR2 gene, interleukin receptor 9 (*IL9R*) belongs to the hematopoietin receptor subfamily and is expressed in both membrane-bound and soluble forms (Renauld *et al.* 1992). A role for

IL9R gene has been suggested in the development of asthma (Holroyd *et al.* 1998; Kauppi *et al*. 2000).

Very recently, two additional studies revealed associations of PARs and susceptibility to psychiatric disorders. A GWA study uncovered an association between schizophrenia and a common SNP near colony stimulating factor 2 receptor, alpha gene (*CSF2RA*) in the PAR1 region (Lencz *et al*. 2007b). A linkage study mapped a new susceptibility locus for bipolar affective disorder to XP22.3/Yp11.3 near acetylserotonin O-methyltransferase-like gene (*ASMTL*) and acetylserotonin O-methyltransferase gene (*ASMT*) in PAR1 (Flaquer *et al.* 2010). Further evidence is needed to support these observations.

4.1.6. PARs and cardiovascular disease

In the past few years, genetic studies of families with congenital heart malformations (including bicuspid aortic valve, aortic coarctation (CoA) and left heart hypoplasia) have shown clustering of mechanistically related left ventricular outflow tract (LVOT) defects (McBride *et al.* 2005; Loffredo *et al*. 2004; Wessels *et al*. 2005; Lewin *et al*. 2004; McBride *et al*. 2008). There is a sexual dimorphism of LVOT defects with a 2:1 male to female ratio. Women with a single X chromosome (Turner syndrome) demonstrate a unique profile of LVOT abnormalities (Bondy, 2012). Female patients with Turner syndrome show cardiovascular phenotypes only when they lack PAR1 region (Sachdev *et al.* 2008). This observation led to the hypothesis that 2 copies of yet unknown genes in PAR1 region are important for normal cardiovascular development (Bondy, 2008) and that haploinsufficiency for genes in the region causes the LVOT

defects characteristic of Turner Syndrome. However, further studies are needed to identify PAR1 genes related to premature CoA and CAD.

A very recent study revealed a different gene expression pattern of five PAR genes (*CSF2RA, DHRSX, PLCXD1, VAMP7, SPRY3*) in males with ischeamic stroke compared to apparently healthy controls. This is so far the most direct evidence for association between PAR genes and human cardiovascular disease (Tian *et al*. 2012).

4.1.7. Hypothesis

Genes of pseudoautosomal regions are transcriptionally active in human monocytes/macrophages and may play a role in CAD.

4.1.8. Objectives

- to perform comprehensive analysis of association between common SNPs within PARs and CAD in individuals of white European ancestry.

- to fully characterise PAR1 and PAR2 transcriptome in human monocytes and macrophages through new generation RNA-sequencing

- to examine if expression of PAR genes in human monocytes and macrophages exhibits sexual dimorphism in available microarray-based resources

.

4.2. Materials and Methods

4.2.1. Characteristics of study cohorts

A total of 23,975 individuals recruited from 12 populations of white European ancestry included in the CARDIoGRAM Consortium was used in this project. Full details of CCGB, DUKE, GerMIFSI, GerMIFSII, OHGS-A, OHGS-B, OHGS-C, PennCATH and WTCCC studies are extensively described in chapter 2. Characteristics of other cohorts used are given below.

- ***German Myocardial Infarction Family Study IV (GerMIFSIV)***

GerMIFSIV cases: This population consists of 2,746 patients with angiographically proven CAD. The recruitment was carried out at the University Hospital Schleswing-Holstein Campus Lubeck between 2005 and 2008 ("Lubeck angiographic registry of patients with structural heart disease"). Patients were not selected for particular risk factors or phenotypes. Information on this population has not been published yet.

GerMIFSIV controls: All controls were considered CAD-free and derived from the "Berlin aging study II" - BASE-II study (Bertram *et al.* 2013).

- ***The Ludwigshafen Risk and Cardiovascular Health study (LURIC)***

LURIC study recruited patients referred for coronary angiography. The project was designed to investigate environmental and genetic risk factors for CVD (Winkelmann *et al.* 2001) including CAD, MI, dyslipidaemia, hypertension, metabolic syndrome and diabetes mellitus. The baseline examination was performed between July 1997 and January 2000 at a single tertiary care centre in southwest Germany (Herzzentrum Ludwigshafen) and included 3,316 study participants. Inclusion criteria for LURIC

were: the availability of a coronary angiogram suggestive of CAD and German ancestry (to reduce genetic heterogeneity). Patients with a history of malignancy within the past five years, any acute illness other than acute coronary syndrome, and any predominant non-cardiac disease were excluded from the study. Angiographic CAD was defined as at least one 50% luminal stenosis within at least one of 15 coronary segments.

- ***MedStar Study***

The MedStar study, conducted by the Cardiovascular Research institute of the MedStar Health Research institute, is a Washington Hospital Center based angiographic study on coronary atherosclerosis (Grant *et al*. 2006; Kathiresan *et al*. 2009). MedStar recruited patients undergoing cardiac catheterization at Washington Hospital between August 2004 and March 2007. The main enrolment criterion was clinical indication for cardiac catheterisation. A total of 447 controls with no evidence of CAD and 875 CAD cases with one or more coronary vessels with ≥50% stenosis were included in the study. Controls were aged >45 years. Cases were diagnosed of CAD<55 years (males) and <60 years (females).

- ***The Cardiogenics Transcriptomic Study***

The Cardiogenics Transcriptomic Study - a European collaboration on genetics of CAD (Heinig *et al*. 2010; Rotival *et al.* 2011, Schunkert *et al*. 2011) - provided a unique resource to examine gene expression in human monocytes and macrophages. This initiative recruited 918 participants (459 patients with MI and 459 normal controls) in five centres within the Cardiogenics consortium: Cambridge (UK), Leicester (UK), Lübeck (Germany), Regensburg (Germany), and Paris (France). Healthy individuals were recruited in Cambridge. All participants were of white European origin.

After data quality control, 1,533 RNA samples (849 from monocytes and 684 from macrophages) were available for statistical analysis. Gene expression profiling was performed using the Illumina Human Ref-8 v3 beadchip array containing 24,516 probes corresponding to 18,311 distinct genes and 21,793 Ref Seq annotated transcripts.

All participants in each study included gave written informed consent in accordance with guidelines of local ethical committees. Study protocols and procedure were reviewed and approved by the appropriate regulatory authorities in each country for each study.

4.2.2. DNA analysis - genotyping and imputation

Affymetrix gene arrays offer only a very poor coverage of the PARs. For example, genome-wide human SNP array 5.0 contains 155 SNPs in PAR1 and no SNPs in PAR2, the genome-wide human SNP array 6.0 has 391 PAR1 SNPs and 32 SNPs in PAR2 and the mapping 500K array has just 262 PAR1 SNPs and no PAR2 SNPs.

All cohorts included in this project used one of the above Affymetrix genotyping platforms in previous GWA analysis. As a result, the array-based genotyping provided genotype information for PAR1 in all studies. Between 133 and 404 genotyped PAR1 SNPs (NCBI build 36) were available in each study after extraction of information from array-based databases (Table 4.2.1). However, four studies only had available genotyped PAR2 information (Table 4.2.2). Between 12 and 69 genotyped PAR2 SNPs were available for further investigation. In order to increase the power and resolution of the genetic analysis of PARs, imputation of the untyped PAR1 and PAR2 SNPs was

conducted. Detailed information about individual genotyping platforms, the imputation method and software used in each study is provided in Table 4.2.3.

Imputation is a method used to predict unobserved genotypes in SNP association studies based on observed data (Marchini *et al.* 2007). In other words, it fills in missing genotypes by extrapolating LD patterns from a reference panel (i.e. HapMap project, 1000 Genomes project) to individuals included in the study (Howie *et al*. 2009).

4.2.2.1 Pre-imputation filtering

Study genotypes and individuals

Before imputation was executed, filtering of the genotyped data was completed to remove variants and individuals with low quality of genotyping; these could decrease the accuracy of the association analysis. SNPs were excluded if they violated one of the following criteria: missingness rate >5%, HWE $p \leq 0.0001$, MAF<1%, poor cluster plots on visual inspection. The quality control filters related to samples varied between cohorts. In general, individuals were excluded if they had a poor genotype call rate, non-European ancestry (based on principal component or multi-dimensional scaling plots used before), high heterozygosity levels or any other cohort-specific reason. These filters were used in previous studies of these populations (Samani *et al.* 2007; Kathiresan *et al*. 2009; Schunkert *et al*. 2011).

SNP positions

The reference panel used for imputation was the 1000 Genomes Project Phase I Integrated release version 3 (March 2012) (The 1000 Genomes Project Consortium, 2010). This panel contains ~39 million SNPs, insertion – deletions and structural variants and is considered as a powerful tool for genotype imputation studies.

The imputation software determined which variants were shared across the reference panel and study data and in order to obtain high-quality imputation SNP positions in the examined datasets were mapped to the same coordinate system (assembly) as the reference panel. The 1000G reference panel uses NCBI build 37 and consequently the positions of SNPs in this analysis were converted to NCBI build 37 using the liftover programme from UCSC Genome browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver?hgsid=341413863).

Strand alignment

Study genotypes were aligned to the same strand orientation as the reference panel (default '+' stramd).

4.2.2.2 Post-imputation quality checks

To ensure that appropriate quality controls were applied, QQ plots were produced for each study across males and females and verified visually. Poorly imputed SNPs were removed based on the imputation quality score prior to meta-analysis. Imputation quality score takes values between 0 and 1, where values near 1 indicate that a SNP has been imputed with high certainty. There is no universal cut-off value for post-imputation SNP filtering. In this project we used previously suggested cut-off thresholds of 0.3 (if imputation was completed using MACH) and 0.4 (if it was done using IMPUTE2) (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html).

Table 4.2.1: Genotyped and imputed SNPs used in analysis of association between PAR1 and CAD in CARDIoGRAM Consortium

Study	Genotyped SNPs after QCs	Genotyped and imputed SNPs that passed post imputation QCs	
		Females	Males
CCGB	239	4054	4104
DUKE	357	5067	5012
GerMIFSI	133	2355	2516
GerMIFSII	262	3824	3814
GerMIFSIV	264	3884	3896
LURIC	340	9421	9452
MedSTAR	277	3515	2412
OHGS-A	139	2327	2329
OHGS-B	234	4173	4025
OHGS-C	404	5361	5397
PennCATH	267	3487	2306
WTCCC	164	2718	2726
Meta-analysis	**133-404**	**3083**	**3442**

SNP - single nucleotide polymorphism, QC – quality control filters

Table 4.2.2: Genotyped and imputed SNPs used in analysis of association between PAR2 and CAD in CARDIoGRAM Consortium

Study	Genotyped SNPs after QCs	Genotyped and imputed SNPs that passed post imputation QCs	
		Females	Males
CCGB	12	140	141
DUKE	68	763	738
OHGS-B	11	321	311
OHGS-C	69	723	707
Meta-analysis	**12-69**	**281**	**317**

SNP - single nucleotide polymorphism, QC – quality control filters

QQ plots were drawn to compare the distribution of observed p-values to the expected distribution under the null hypothesis of no association to ensure the good quality of the data (Figure 4.2.1). Observations above the theoretical line indicated that more P-values were significant whereas observations below the theoretical line indicated that fewer P-values were significant than expected by chance. Males in GerMIFSI study were excluded from the meta-analysis since the QQ plot was flat.

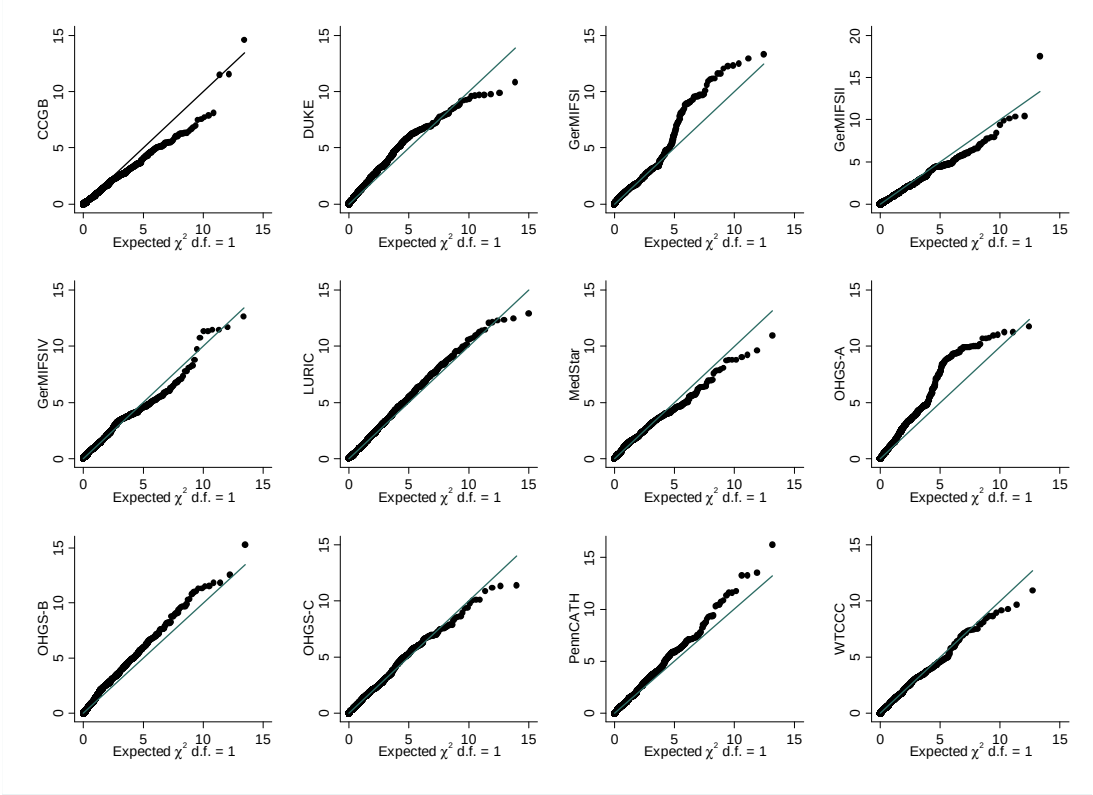
Table 4.2.3: Genotyping, imputation and analysis criteria for each study

Study acronym		CCGB	DUKE	GerMIFSI	GerMIFSII	GerMIFSIV	LURIC
Genotyping	Platform	Affymetrix 6.0	Affymetrix Axiom	Affymetrix NSP and STY	Affymetrix 6.0	Affymetrix 6.0	Affymetrix 6.0
	Calling algorithm	Birdseed	AxiomGT1	Birdseed	BRLMM	BRLMM	Birdseed
Imputation	NCBI build	37	37	37	37	37	37
	Software	IMPUTE2	IMPUTE2	IMPUTE2	IMPUTE2	IMPUTE2	IMPUTE2

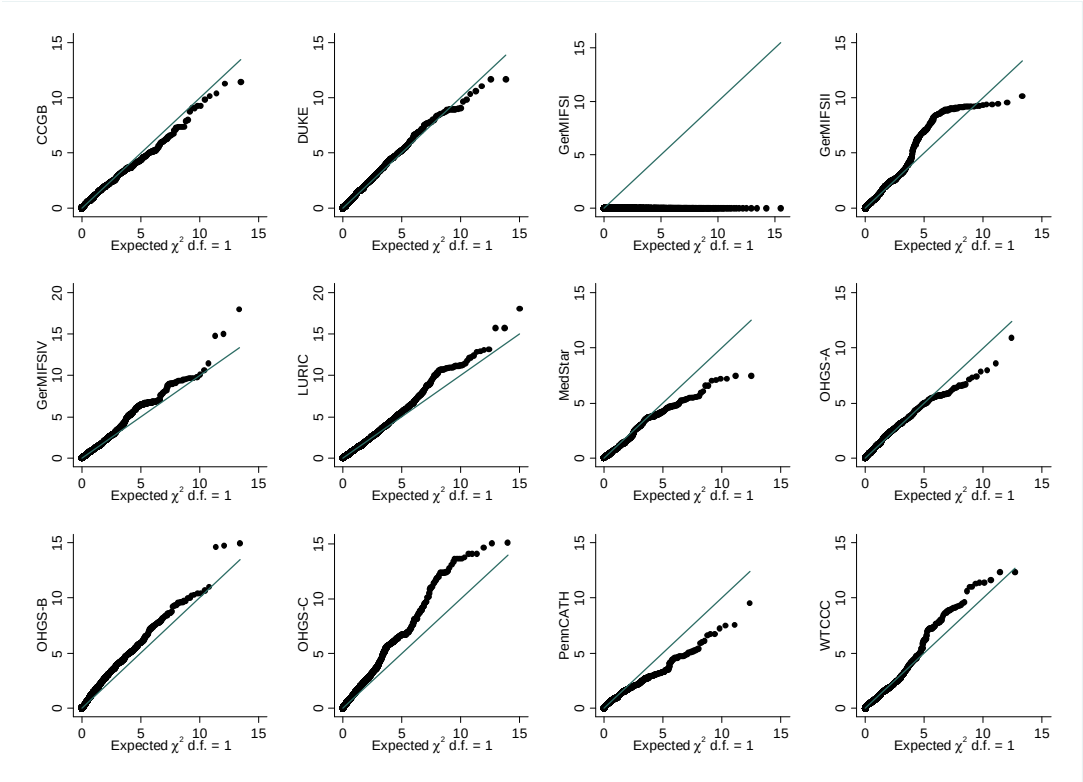
Study acronym		MedStar	OHGS-A	OHGS-B	OHGS-C	PennCATH	WTCCC
Genotyping	Platform	Affymetric 6.0	Affymetrix 500K	Affymetrix 6.0	Affymetrix Axiom	Affymetrix 6.0	Affymetrix 500K Array set
	Calling algorithm	Birdseed	BRLMM	Birdseed	Axiom GT1	Birdseed	CHIAMO
Imputation	NCBI build	37	37	37	37	37	37
	Software	MACH	IMPUTE2	IMPUTE2	IMPUTE2	MACH	IMPUTE2

Figure 4.2.1: Quantile-quantile plots of P-values for all SNPs that passed quality control filters and were used in the meta-analysis - (A) Females and (B) Males

A.



B.



4.2.3. Statistical methods

The association analysis consisted of two stages:

Primary sex-stratified analyses were performed in each study separately. Within each cohort SNP associations with CAD were analysed by logistic regression assuming additive model of inheritance with adjustment for study-specific covariates where appropriate. Results from each study were included in the meta-analysis only if the SNP imputation quality score was>0.4 and if the MAF was>1%. Only SNPs that were available in>50% of the total sample size over all studies were analysed, resulting in a total number of 3,083 SNPs in the female meta-analysis and 3,442 SNPs in the male meta-analysis.

Meta-analysis of all individual study associations was conducted using a fixed-effects inverse variance weighting model in STATA v12. Studies were weighted inversely proportional to the variance of the effect size. As a measure for between studies heterogeneity, I^2 was calculated (Higgins *et al*. 2003).

When there was no indication for heterogeneity for a SNP (I^2<40%), the fixed-effect model was maintained. When heterogeneity was present (I^2>40%), random-effects model was adopted and reported which incorporated between-study variation in the weighting.

To correct for multiple testing, false discovery rate (FDR) (*q*-values) was calculated based on all association tests conducted in sex-stratified meta-analyses using the *R*-based tool *Q*-value (Storey and Tibshirani, 2003). The FDR estimated the expected number of false-positives among all positive results.

Each of the loci containing significantly associated SNPs was visualised using regional association plots generated by LocusZoom (Pruim *et al*. 2010).

4.2.4. PARs – RNA-based analyses

Information on PAR1 and PAR2 genes expression in monocytes and macrophages was extracted *in silico* from microarray-based experiment conducted in Cardiogenics Study reported previously (Schunkert *et al*. 2011). In brief, monocytes were isolated from whole blood by positive selection with CD14 magnetic beads using an AutoMACS system (Miltenyi Biotech, BergischGladbach, Germany) (Heinig *et al.* 2010; Charchar *et al.* 2012). Flow cytometry was used to confirm cell purity and in all samples more than 90% of cells were CD14-positive monocytes (Heinig *et al* 2010; Charchar *et al.* 2012). Macrophages were obtained from culturing of monocytes for 7 days in macrophage-SFM medium (Gibco/Invitrogen, Grand Island, USA) with 50 ng/mL recombinant human M-CSF (R&D Systems, Minneapolis, USA) (Schunkert *et al.* 2011; Charchar et al. 2012). Then RNA was extracted from both cell types as it has been described elsewhere (Schunkert *et al*. 2011, Charchar *et al*. 2012). Preparatory procedures of monocyte, macrophage, and RNA isolation were undertaken separately in each institution using standardised protocols. Further microarray gene-expression profiling of all samples was done in one institution (Paris, France). Every sample was run on the Illumina Human Ref-8 arrays (Illumina, San Diego, USA) containing 24,516 probes (Schunkert *et al*. 2011, Charchar *et al.* 2012). The mRNA was amplified and labelled with the Illumina Total Prep RNA Amplification Kit (Ambion, Austin, USA) (Schunkert *et al*. 2011). Following hybridisation, array images were scanned with an IlluminaBeadArray Reader and probe intensities were extracted with the gene

expression module of Illumina Bead Studio software (Heinig *et al.* 2010).Variance stabilisation transformation was applied to the raw intensities and quantile normalisation was done in the R statistical environment with the Lumi and Beadarray packages (version 1.8.3) (Schunkert *et al.* 2011).

The data on monocyte transcriptome profiling from Cardiogenics were available for 541 men and 308 women. Analysis of macrophage transcriptome was conducted in 449 men and 235 women. The comparative (men versus women) analysis was conducted at the probe level by using linear regression adjusted for age, centre and disease status using STATA v12 software. After quality control filters, information on 9 PAR genes was available to examine sex differences in expression of PAR genes in monocytes and macrophages. Bonferroni correction was applied to count for multiple testing.

4.2.5. PARs – new generation RNA sequencing

To fully characterise PAR1 and PAR2 transcriptome in monocytes and macrophages new generation RNA sequencing (RNA-seq) was conducted in a sample from 38-year old apparently healthy man of white European ancestry. Blood sample was taken after antecubital venepuncture. Monocytes isolation and conversion to macrophages was conducted using the protocol introduced in Cardiogenics (Schunkert *et al*. 2011; Charchar *et al*. 2012). RNA was extracted from both monocytes and macrophages usingTRIzol, followed by clean-up with RNeasy columns (Qiagen, Venlo, Netherlands) and DNase-based treatment (Heinig *et al*. 2010, Charchar *et al*. 2012).

RNA was sequenced on the Illumina HiSeq-2000 sequences using 100 bp paired end reads. The sequencing generated 2Gb of sequence data. Transcripts for genes in PAR

regions were quantified in reads per kilobase per million reads (Mortazavi *et al*. 2008). FPKM is a normalised measure of transcriptional activity and reflects the molar concentration of the transcript within a sample. A minimum FPKM of 0.125 was used as a threshold in order to exclude expression signals at very low levels that may be artefacts from multi-mapping reads spreading a small amount of supposed expression to inactive transcripts.

4.3. Results

4.3.1. Characteristics of study cohorts

A total of 16,226 men and 9,536 women from 12 populations of white European ancestry in CARDIoGRAM Consortium were included in the sex-stratified meta-analyses. The characteristics of the examined populations are listed in table 4.3.1.

Table 4.3.1: Characteristics of populations in CARDIoGRAM Consortium

Study	Total n of subjects	Cases	Males	Females	Age Females	Age Males
CCGB	1993	1626 (81.6)	1427 (71.6)	566 (28.4)	60.48 (12.32)	50.33 (10.64)
DUKE	1809	1177 (65.1)	1088 (60.1)	726 (39.9)	62.10 (9.19)	56.81 (9.87)
GerMIFSI	657	215 (32.7)	139 (21.2)	518 (79.8)	59.56 (10.53)	56.94 (11.41)
GerMIFSII	2520	1222 (48.5)	1650 (65.5)	870 (34.5)	50.06 (13.94)	49.85 (13.03)
GerMIFSIV	2328	1181 (50.7)	1204 (51.7)	1124 (48.3)	58.66 (14.49)	56.10 (12.90)
LURIC	2949	2048 (69.4)	2074 (70.3)	875 (29.7)	-	-
MedStar	1322	875 (66.2)	857 (64.8)	465 (35.2)	-	-
OHGS-A	1955	947 (48.4)	1289 (65.9)	666 (34.1)	68.40 (12.37)	58.56 (14.66)
OHGS-B	2811	1293 (46.0)	1660 (59.1)	1151 (40.9)	69.43 (11.39)	59.22 (14.70)
OHGS-C	1155	839 (72.6)	900 (77.9)	255 (22.1)	70.88 (12.85)	58.99 (9.11)
PennCATH	1401	933 (66.6)	937 (66.9)	464 (33.1)	-	-
WTCCC	4864	1926 (39.6)	2973 (61.1)	1891 (38.9)	-	-
In total	**25,762**	**14,282 (55.4)**	**16,228 (63.0)**	**9,536 (37.0)**	-	-

Data are counts and percentages or means and standard deviations, n – number of individuals in the analysis, (-) - information not available

All individuals were recruited in four countries. CCGB, DUKE and OHGS studies were recruited in Canada, GerMIFSI, GerMIFSII, GerMIFSIV and LURIC in Germany, MedStar and PennCATH in the USA and WTCCC in UK. 63% of participants were males. Overall, approximately 55% were CAD cases and 45% healthy controls. The mean age of participants ranged from 49.85 to 59.22 for males and 50.06 to 70.88 for females.

4.3.2. Analysis of association between PAR1 and CAD in males

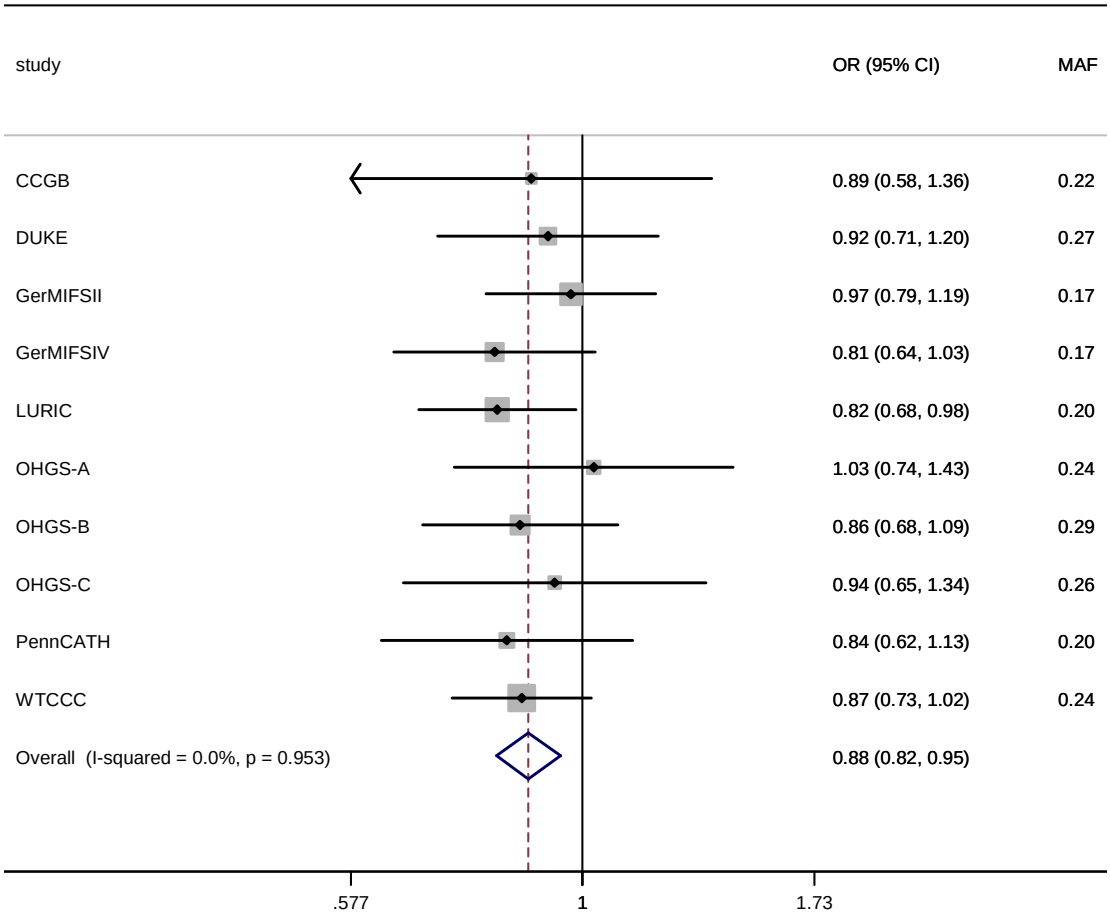
A total of 3,442 SNPs survived post-imputation quality filters and were used in the male-specific meta-analysis. Of those 277 (8.0%) SNPs showed nominal association with CAD ($P<0.05$) (data not shown). Imputed SNP rs141738136 ($P=8.75x10^{-4}$) showed the most significant association with CAD (Figure 4.3.1). This SNP has been merged into rs123468 and is located in the intronic region of *XG* gene, the most centromeric protein coding gene in PAR1 (very close to the pseudoautosomal – MSY boundary). After correction for multiple testing no PAR1 SNPs retained their association with CAD (Table 4.3.2). The top 20 association signals map to three different loci in PAR1 (Figure 4.3.2).

Table 4.3.2: Analysis of association between PAR1 and CAD in males in CARDIoGRAM Consortium – top association signals

	SNP	Position	Minor allele	MAF	N of studies	OR (95% CI)	P value	FDR Qvalue
1	rs141738136	2673098	T	0.23	10 (13107)	0.88 (0.82, 0.95)	$8.75x10^{-4}$	0.32
2	rs183905870	700841	C	0.15	11 (13962)	0.88 (0.82, 0.96)	0.0017	0.32
3	rs141933565	2667544	A	0.23	11 (13962)	0.89 (0.83, 0.96)	0.0017	0.32
4	rs146326145	2672215	G	0.27	11 (13962)	0.90 (0.84, 0.96)	0.0018	0.32
5	rs147056284	680909	G	0.15	8 (11419)	1.18 (1.06, 1.30)	0.0018	0.32
6	rs150156698	2366612	G	0.49	9 (12183)	1.10 (1.04, 1.18)	0.0020	0.32
7	rs150344862	1534293	G	0.25	8 (11327)	0.88 (0.81, 0.96)	0.0023	0.32
8	rs138561386	1533596	G	0.27	9 (12183)	0.89 (0.82, 0.96)	0.0023	0.32
9	rs147969925	2420103	A	0.07	7 (8399)	0.76 (0.64, 0.91)	0.0024	0.32
10	rs180982472	698630	A	0.14	11 (13962)	0.88 (0.81, 0.96)	0.0024	0.32
11	rs148564958	1532533	A	0.27	8 (11327)	0.89 (0.82, 0.96)	0.0025	0.32
12	rs138671348	2366617	T	0.49	9 (12183)	1.10 (1.03, 1.17)	0.0025	0.32
13	rs141006490	2326782	C	0.43	9 (11762)	1.09 (1.01, 1.17)	0.0026	0.32
14	rs150905436	1533771	T	0.28	9 (12183)	1.12 (1.04, 1.20)	0.0027	0.32
15	rs139494123	1455800	T	0.50	7 (9747)	0.91 (0.85, 0.97)	0.0029	0.32
16	rs146697737	1533165	T	0.26	9 (12183)	0.89 (0.82, 0.96)	0.0029	0.32
17	rs146844905	707404	C	0.13	10 (13107)	0.86 (0.78, 0.95)	0.0031	0.32
18	rs146688868	2366547	C	0.45	10 (13038)	1.10 (1.03, 1.17)	0.0032	0.32
19	rs188321807	1534591	A	0.35	7 (10329)	1.13 (1.04, 1.22)	0.0033	0.32
20	rs28665237	2328102	C	0.28	7 (8524)	0.87 (0.80, 0.96)	0.0033	0.32

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – 95% confidence interval, P value – level of statistical significance from inverse variance fixed model effects (or random model effects if I^2 >40%) meta-analysis, FDR Qvalue – corrected level of statistical significance after applying FDR, SNP positions match chrX and are based on build 37.

Figure 4.3.1: Association between rs141738136 and CAD in males – forest plot. Results from each study for rs141738136 are shown as grey squares and a horizontal line represents the effect estimate (OR) together with its confidence intervals. The size of the square corresponds to the sample-size – the weight that the study contributes to the meta-analysis. The combined-effect estimate and its confidence interval are illustrated as the diamond and the broken vertical line, OR – odds ratio, CI – confidence intervals, MAF – minor allele frequency.



In all but one study (OHGS-A), the direction of association between rs141738136 and CAD was consistent with minor allele decreasing the risk of CAD by 3% to19 % and on average - by 12%. There was no heterogeneity between studies (I^2 P=0.95). The frequency of the minor allele in each study is given on the right side of the plot. The MAF of rs141738136 varied across the examined studies (0.17 to 0.29).

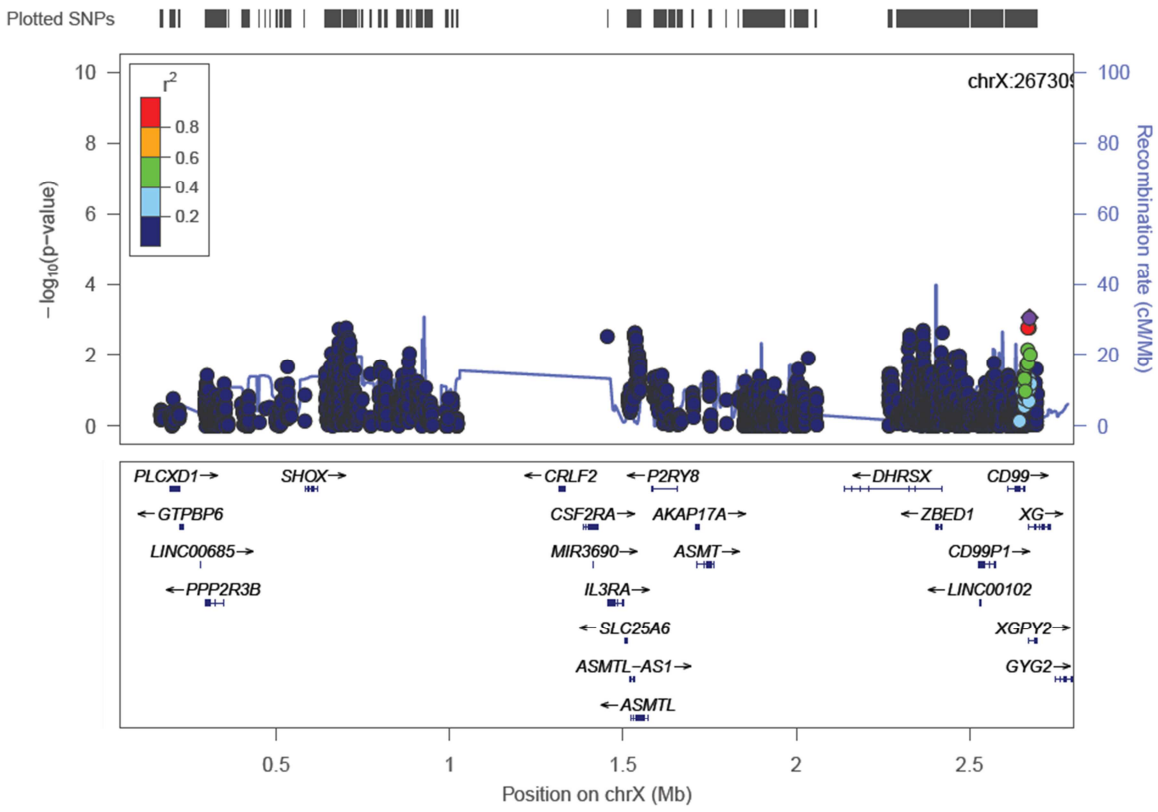
There was no association between this SNP and CAD in women (P=0.97) (Table 4.3.3).

Table 4.3.3: Association between top male SNP (rs141738136) and CAD in women in CARDIoGRAM Consortium

SNP	Position	Minor allele	MAF	N of studies	OR (95% CI)	P value
rs141738136	2673098	T	0.19	12 (8501)	1.00 (0.91, 1.10)	0.97

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – confidence interval, P value – level of statistical significance from inverse variance fixed model effects (or random model effects if I^2 P>40%) meta-analysis.

Figure 4.3.2 Regional association plot – association between PAR1 and CAD in males. Associations of individual PAR1 SNPs with CAD based on meta-analysis from all studies are plotted as –log10 *P* values (Y-axis) against chromosomal bp position (X-axis). The most significant SNP (rs141738136) is shown as a purple diamond. Its LD relationship with the other SNPs is shown by different colours (red is r^2>0.8 and blue is r^2<0.2). The bottom panel of the plot shows the name and location of known genes. Positions of exons are displayed, and the transcribed strand is shown with an arrow to facilitate the visual comparison of association results relative to coding regions. Recombination hotspots are presented as blue peaks.



As illustrated in Figure 4.3.2, there were two gaps in PAR1 where no SNP coverage was provided by the SNP array. The lead SNP rs141738136 is located in an intron of XG blood group gene (*XG*).

4.3.3. Analysis of association between PAR1 and CAD in females

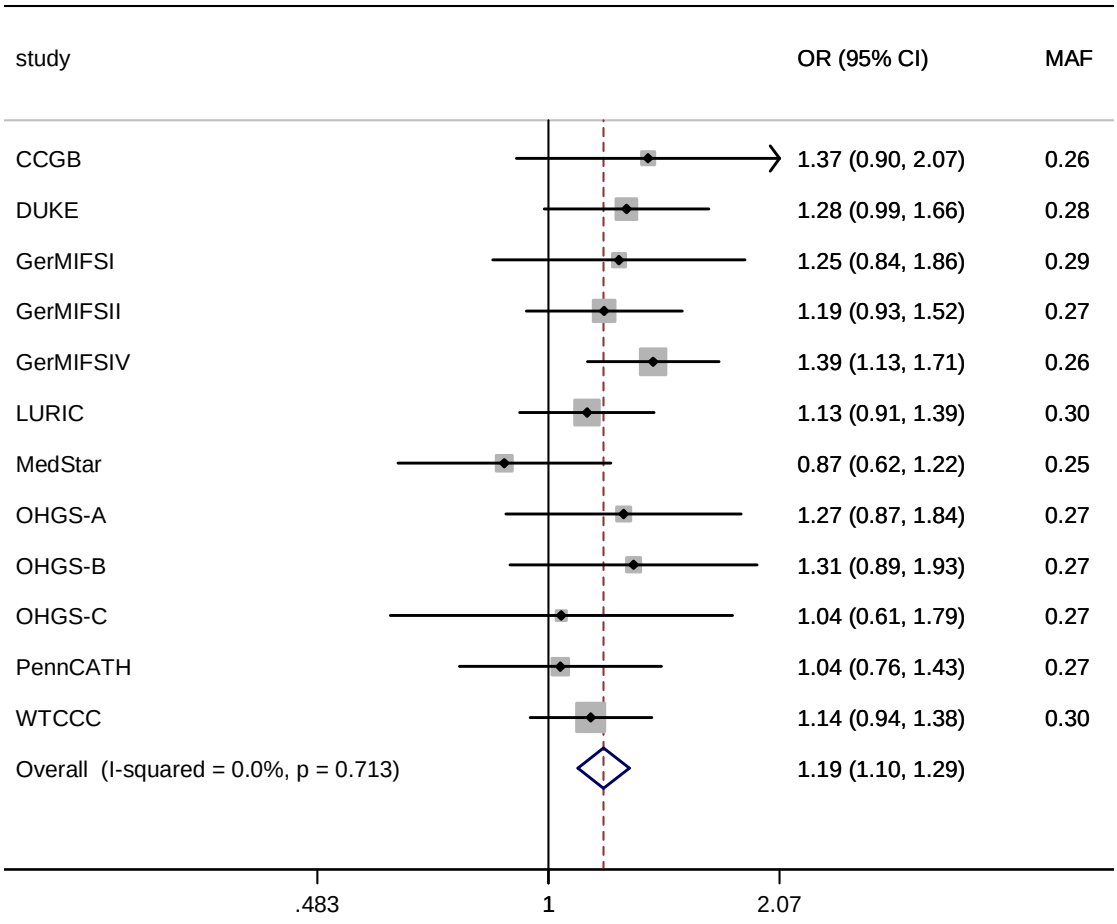
A total of 3,083 SNPs survived post-imputation filters and were used in the female-specific meta-analysis. Of those 115 (3.7%) SNPs showed nominal association with CAD ($P<0.05$) (data not shown). After correction for multiple testing, 20 SNPs retained their association with PAR1 and CAD (Table 4.3.4). All of them were common SNPs (MAF varied between 0.18 and 0.46). The effect size of these SNPs in CAD risk is modest to moderate (OR from 1.13 to 1.25). Imputed SNP rs144253516 ($P=2.44x10^{-5}$) showed the most significant association (Figure 4.3.3) with CAD. Each minor allele (T) copy increases the risk of CAD by approximately 19% [OR=1.19 (1.10-1.29)]. This SNP has been merged into rs113921272 and is located in an intragenic locus (Figure 4.3.4).

Table 4.3.4: Analysis of association between PAR1 and CAD in females in CARDIoGRAM Consortium – top association signals

	SNP	Position	Minor allele	MAF	N of studies	OR (95% CI)	P value	FDR Qvalue
1	rs144253516	946524	T	0.27	12 (8501)	1.19 (1.10, 1.29)	$2.44x10^{-5}$	0.02
2	rs188358977	942659	A	0.24	12 (8531)	1.21 (1.11, 1,21)	$2.56x10^{-5}$	0.02
3	rs145495800	942973	G	0.23	12 (8501)	1.21 (1.11, 1.21)	$2.67x10^{-5}$	0.02
4	rs142933381	944463	T	0.41	12 (8501)	0.85 (0.78, 0.92)	$3.29x10^{-5}$	0.02
5	rs145362525	944713	A	0.27	12 (8501)	1.18 (1.09, 1.27)	$3.51x10^{-5}$	0.02
6	rs146991652	947544	T	0.27	12 (8501)	1.19 (1.10, 1.29)	$3.68x10^{-5}$	0.02
7	rs138100619	947948	T	0.27	12 (8501)	1.19 (1.09, 1.29)	$4.45x10^{-5}$	0.02
8	*rs5946608	944168	T	0.45	12 (9019)	1.16 (1.08, 1.25)	$4.86x10^{-5}$	0.02
9	rs148241722	944436	T	0.40	12 (8501)	0.85 (0.79, 0.82)	$5.54x10^{-5}$	0.02
10	rs151295303	943652	T	0.40	12 (8501)	0.85 (0.79, 0.82)	$7.05x10^{-5}$	0.02
11	rs141390273	945100	A	0.39	12 (8502)	0.86 (0.80, 0.93)	$8.78x10^{-5}$	0.02
12	rs143775210	948002	T	0.28	12 (8501)	1.18 (1.09, 1.29)	$9.06x10^{-5}$	0.02
13	rs145889839	944098	C	0.40	12 (8501)	0.86 (0.80, 0.93)	$9.20x10^{-5}$	0.02
14	rs147465651	942891	C	0.37	12 (8501)	0.84 (0.77, 0.92)	$9.39x10^{-5}$	0.02
15	rs141739675	942367	T	0.18	9 (6924)	1.25 (1.11, 1.41)	$1.92x10^{-4}$	0.04
16	rs150852359	946631	G	0.43	12 (8501)	0.86 (0.79, 0.83)	$2.16x10^{-4}$	0.04
17	rs138614173	945269	T	0.41	12 (8502)	0.87 (0.81, 0.94)	$2.23x10^{-4}$	0.04
18	rs144066964	944385	C	0.42	12 (8501)	0.86 (0.80, 0.93)	$2.31x10^{-4}$	0.04
19	rs147393599	945374	C	0.46	11 (8123)	0.93 (0.86, 1.00)	$2.41x10^{-4}$	0.04
20	rs149476563	944371	C	0.42	12 (8501)	0.86 (0.80, 0.94)	$2.59x10^{-4}$	0.04

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – 95% confidence interval, P value – level of statistical significance from inverse variance fixed model effects (or random model effects if I^2 >40%) meta-analysis, FDR Qvalue – corrected level of statistical significance after applying FDR, SNP positions are based on build 37, *rs5946608 is a directly genotyped SNP.

Figure 4.3.3: Association between rs144253516 and CAD in females – forest plot. Results from each study for rs144253516 are shown as grey squares and a horizontal line represents the effect estimate (OR) together with its confidence intervals. The size of the square corresponds to the sample-size - the weight that the study contributes to the meta-analysis. The combined-effect estimate and its confidence interval are illustrated as the diamond and broken vertical line. OR – odds ratio, CI – confidence intervals, MAF – minor allele frequency.



In all but one study (MedStar) the direction of association between rs144253516 and CAD was consistent with minor allele increasing the risk of CAD by 4% to 39% and overall on average by 19% (Figure 4.3.3). The frequency of the minor allele (0.25 to 0.30) showed a good constancy across studies. There was no heterogeneity between studies included in the meta-analysis (I^2 P=0.71).

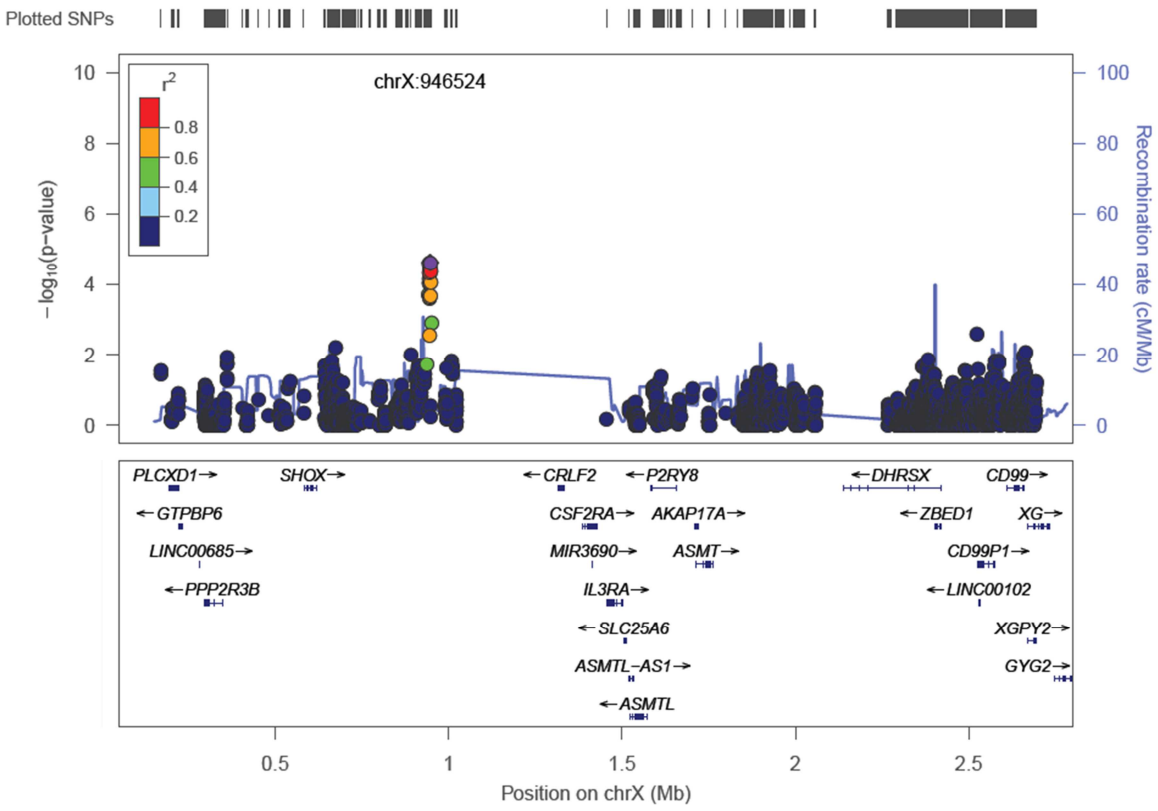
Further exploration of the association between this SNP and CAD in men revealed no significant findings (P=0.40) (Table 4.3.5).

Table 4.3.5: Association between top female SNP (rs144253516) and CAD in men in CARDIoGRAM Consortium

SNP	Position	Minor allele	MAF	N of studies (n)	OR (95% CI)	P value
rs144253516	946524	T	0.28	11 (13962)	1.03 (0.97, 1.10)	0.40

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – confidence interval, P value – level of statistical significance from inverse variance fixed model effects meta-analysis.

Figure 4.3.4: Regional association plot – Association between PAR1 and CAD in females. Associations of individual PAR1 SNPs with CAD based on meta-analysis from all 12 studies are plotted as –log10 *P* values (Y-axis) against chromosomal bp position (X-axis). The most significant SNP (rs144253516) is shown as a purple diamond. Its LD relationship with the other SNPs is shown by different colours (red is r^2>0.8 and blue is r^2<0.2). The bottom panel of the plot shows the name and location of known genes. Positions of exons are displayed, and the transcribed strand is shown with an arrow to facilitate the visual comparison of association results relative to coding regions. Recombination hotspots are presented as blue peaks.



20 SNPs with strongest associations with CAD are in moderate to high LD (r^2>0.6) and map to an intergenic region. The lead SNP rs144253516 is located ~330kb away from the 3'untranslated region of *SHOX* gene and ~370kb and ~440kb away from the 3' untranslated regions of *CRLF2* and 5' *CSF2RA* genes, respectively.

Further exploration of the intergenic region of PAR1for other regulatory genomic features (Figure 4.3.5) in UCSC public database revealed presence of a long intergenic non-coding RNA (lincRNA) - RP11-309M23, only ~4kb away from the CAD-associated region. The length of the RP11-309M23 is 4254bp and its transcript is 588bp in length. It contains several transposable elements. For example, there is an enrichment of retroviral elements, ERV1 and ERVL-MalR families in RP11-309M23. Characteristically, a long terminal repeat (LTR) that belongs to ERV1 LTR family overlaps with exon1 of the lincRNA. There is no information on biological function of this lincRNA. RP11-309M23 lies ~437kb upstream of *CSF2RA* gene, a cytokine which is biologically an interesting candidate gene.

The top imputed PAR1 SNP - rs144253516 maps to a repeat masked region, a SINE AluY element whilst the most significant genotyped polymorphism - rs5946608 is located in an LTR ERV1 sequence (Figure 4.3.6).

Figure 4.3.5: An in-depth view of the female-specific CAD associated locus in PAR1 (in blue). The genomic coordinates 942,367…948,002 (±5000bp) are shown on top of the graph. A detailed annotation of the interspersed repeating elements that are present in the explored region is provided in the bottom panel. The level of colour shading reflects the amount of base mismatch, base deletion, and base insertion associated with a repeat element. The higher the combined number of these, the lighter the shading. SINE – short interspersed nuclear element (which include Alu), LINE – long interspersed nuclear element, LTR – long terminal repeat elements, top imputed SNP rs144253516 and genotyped SNP rs5946608 (circled in red) and lincRNA RP11-309M23 (in green) are highlighted in the top panel [Produced in UCSC database].

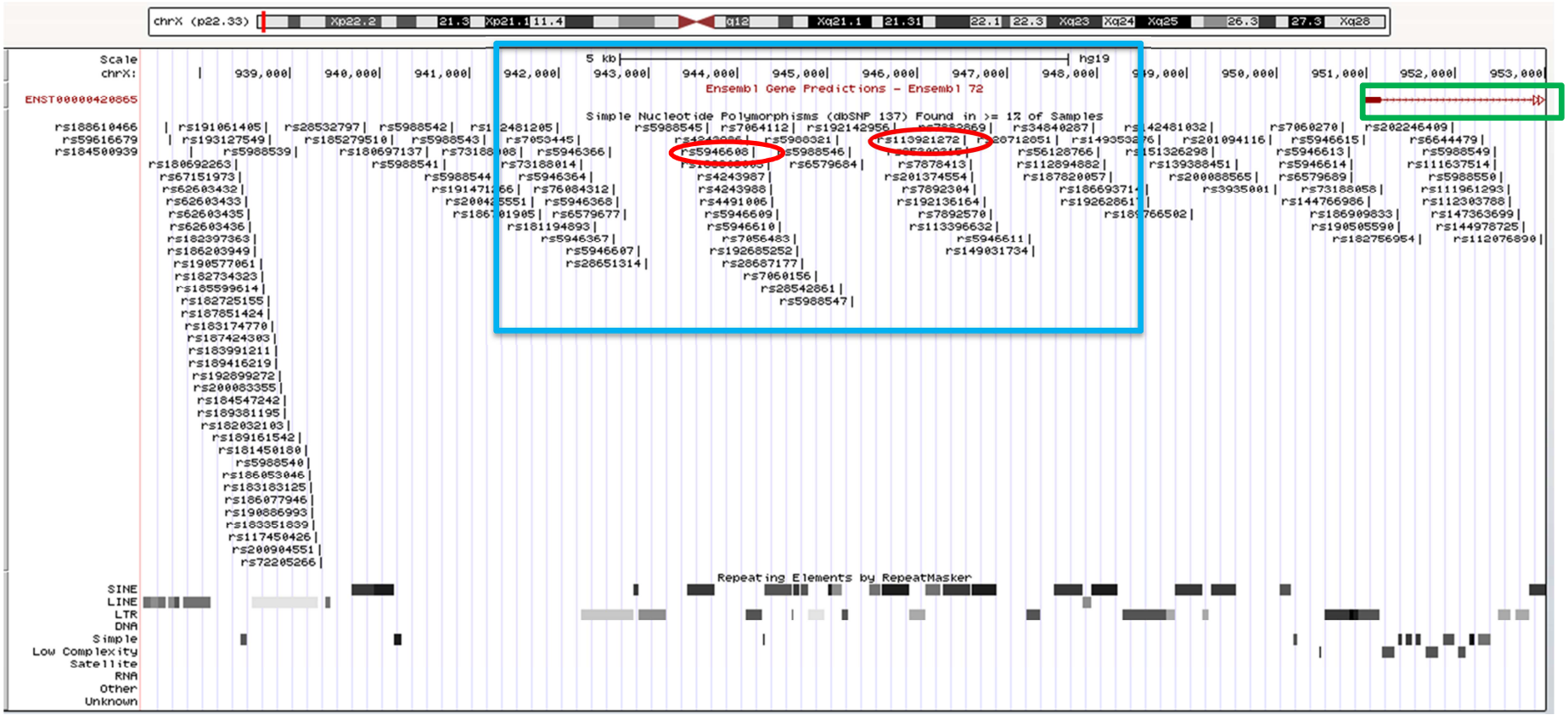
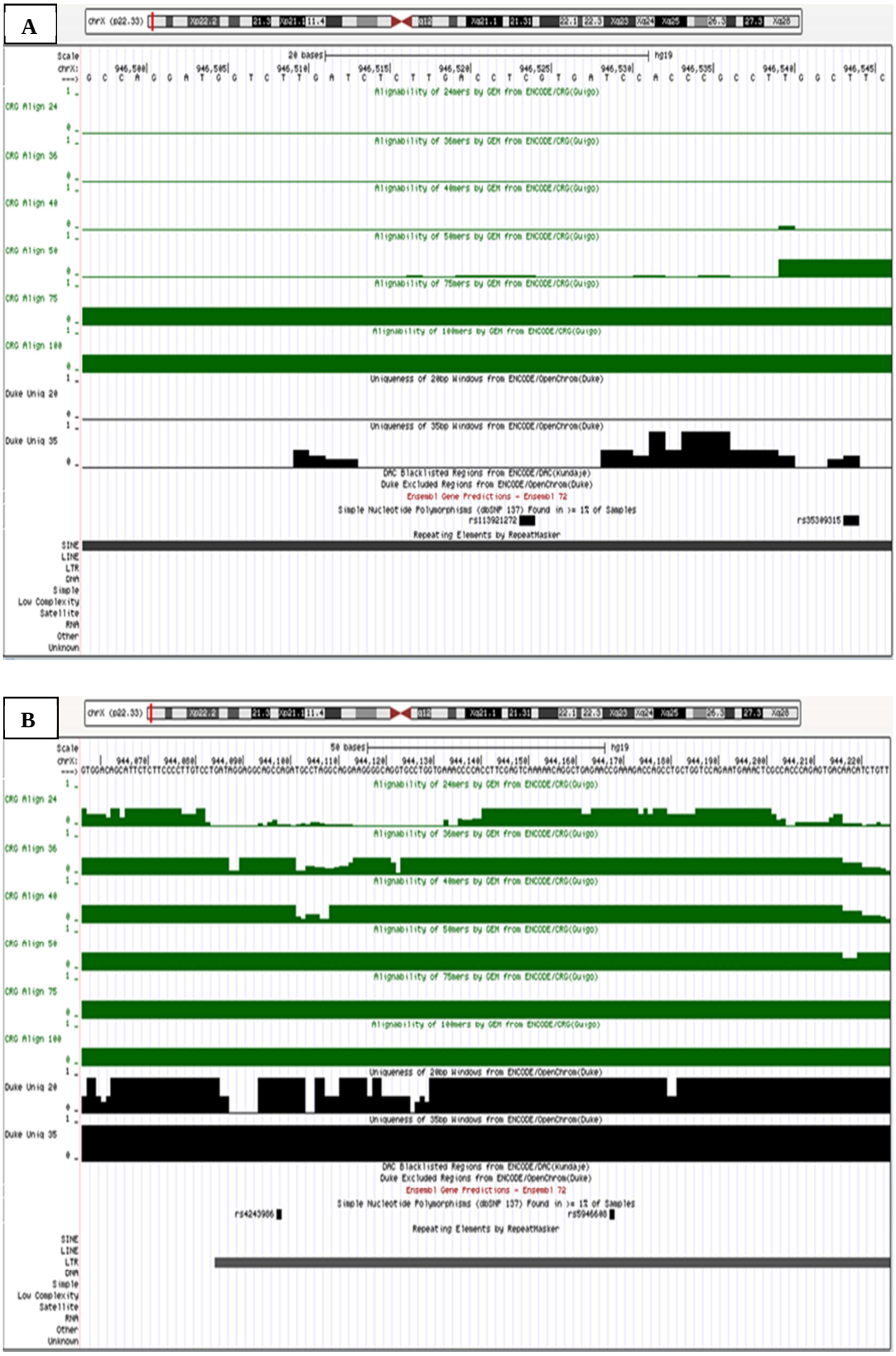


Figure 4.3.6: **UCSC-based view of the genomic location of top PAR1 SNPs** - imputed rs144253516 (A) and directly genotyped rs5946608 (B) SNPs. The green tracks refer to the level of alignability of sequence surrounding the SNP, black tracks refer to the uniqueness of sequence that surround the SNP (signal intensity range: 0-1; whereby 0 – weak and 1 – strong); Repetitive sequences (SINE, LINE LTR) are listed on the bottom. [Obtained from UCSC database].



4.3.4. Analysis of association between PAR2 and CAD

SNP data on PAR2 were available only in Canadian studies, CCGB, DUKE, OHGS-B and OHGS-C. Therefore, the analysis was restricted to 5,075 men and 2,698 women from these studies.

A total of 317 SNPs in PAR2 passed the post-imputation quality filters and were used in the male-specific meta-analysis. The results of the most significant findings from this analysis are shown in Table 4.3.6. Of 317 analysed SNPs, only 2 (rs143917348, rs149183634) showed a nominal level of association with CAD.

Table 4.3.6: Analysis of association between PAR2 and CAD in males – top association signals in 4 Canadian studies from CARDIoGRAM Consortium.

	SNP	Position	Minor allele	MAF	N of studies (n)	OR (95% CI)	P value
1	rs143917348	155037001	G	0.34	3 (2727)	0.84 (0.71, 0.99)	0.0353
2	rs149183634	155229826	T	0.30	3 (2727)	1.18 (1.00, 1.40)	0.0484
3	rs142003474	155207586	G	0.32	3 (2727)	1.17 (1.00, 1.37)	0.0511
4	rs139264026	155037027	G	0.35	3 (2727)	0.85 (0.72, 1.00)	0.0513
5	rs147560965	155207391	G	0.30	3 (2727)	1.17 (1.00, 1.38)	0.0518

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – confidence interval, P value – level of statistical significance from inverse variance fixed model effects meta-analysis, SNP positions are based on build 37.

A total of 281 SNPs in PAR2 survived the post-imputation quality filters and were used in the female meta-analysis. Of those, 12 SNPs showed association with CAD at the nominal level of statistical significance ($P<0.05$) (Table 4.3.7). However, neither of the nominal association signals detected in PAR2 remained significant after correction for multiple testing.

Table 4.3.7: Analysis of association between PAR2 and CAD in females – top association signals in Canadian studies in CARDIoGRAM Consortium.

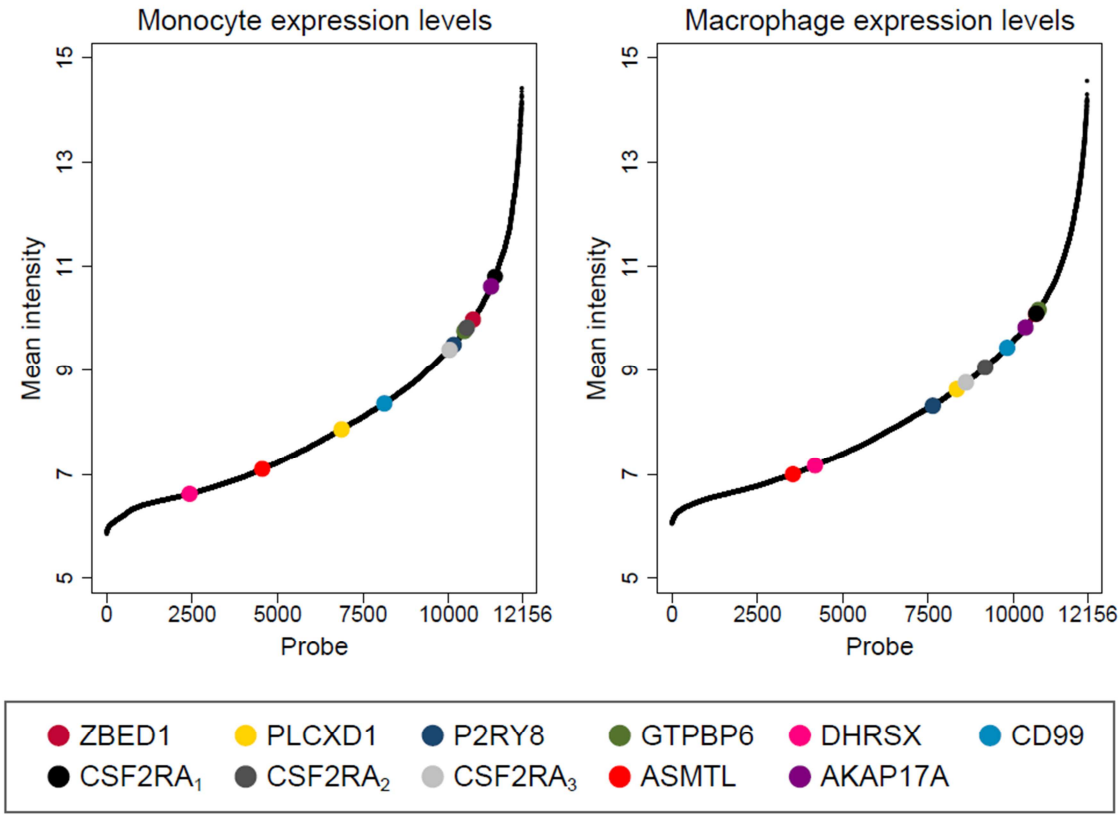
	SNP	Position	Minor allele	MAF	NS (n)	OR (95% CI)	P value
1	rs192915026	154943982	A	0.28	3 (1564)	1.30 (1.05, 1.61)	0.0148
2	rs150396739	154949953	G	0.23	3 (1564)	1.30 (1.05, 1.61)	0.0161
3	rs149164276	155017860	C	0.18	3 (1564)	0.77 (0.61, 0.95)	0.0164
4	rs139362836	154941571	A	0.23	3 (1564)	1.30 (1.05, 1.62)	0.0187
5	rs149869683	154947398	T	0.23	3 (1564)	1.29 (1.04, 1.61)	0.0193

SNP – single nucleotide polymorphism, MAF – minor allele frequency, N – number of studies, n – number of individuals in the analysis, OR – odds ratio, CI – confidence interval, P value – level of statistical significance from inverse variance fixed model effects (or random model effects if I^2 P>40%) meta-analysis, SNP positions are based on build 37 and match chrX.

4.3.5. PAR1 and PAR2 gene expression studies

Sex-stratified comparative analysis of PAR genes expression was conducted in Cardiogenics Study where gene expression at mRNA was measured in both human monocytes and macrophages. Only 9 PAR1 genes (*AKAP17A, ASMTL, CD99, CSF2RA, DHRSX, GTPBP6, PLCXD1, P2RY8* and *ZBED1*) with a total of 11 probes were available on the microarray platform used in transcriptome profiling in this study. Apart from *CSF2RA* gene that had 3 transcripts measured, all other PAR1 genes were represented by single probes. All PAR1 genes represented on the microarray platform were expressed in human monocytes/macrophages (Figure 4.3.7). In context, of all 18,311 genes with measurable expression in monocytes/macrophages, PAR1 genes showed moderate abundance (above inclusion threshold) in both cell types.

Figure 4.3.7: Expression levels of PAR1 genes in monocytes and macrophages.



The findings from sex-stratified comparative gene expression analysis in human monocytes/macrophages are shown in Table 4.3.8. Except of *AKAP17A*, all PAR1 genes were up-regulated in male monocytes and macrophages when compared to females. Of 11 analysed probes, 6 showed an expression difference in monocytes and 8 in macrophages at the nominal level of statistical significance ($P<0.05$). *CD99*, *CSF2RA*, *P2RY8* and *ZBED1* genes showed a statistically significant difference in their expression level between men and women in monocytes and *ASMTL*, *CD99* and *CSF2RA* genes showed a statistically significant difference in their expression level between men and women in macrophages (Corrected $P< 4.5x10^{-3}$).

Table 4.3.8: Sex differences in PAR1 genes expression in human macrophages and monocytes – Cardiogenics cohort

Gene	Monocytes			Macrophages		
	M vs F	Beta ± SE	P value	M vs F	Beta ± SE	P value
AKAP17A	Down	-0.02±0.03	0.55	Up	0.04±0.03	0.22
ASMTL	Up	0.001±0.02	0.87	Up	0.06±0.02	$2.92x10^{-3}$
CD99	Up	0.09±0.02	$8.58x10^{-5}$	Up	0.09±0.03	$2.55x10^{-3}$
$CSF2RA_1$	Up	0.06±0.03	0.02	Up	0.07±0.04	0.03
$CSF2RA_2$	Up	0.09±0.02	$2.02x10^{-7}$	Up	0.16±0.04	$6.08x10^{-5}$
$CSF2RA_3$	Up	0.10±0.02	$1.20x10^{-9}$	Up	0.15±0.04	$5.04x10^{-5}$
DHRSX	Up	0.02±0.01	0.13	Up	0.02±0.02	0.35
GTPBP6	Up	0.05±0.03	0.10	Up	0.03±0.03	0.30
P2RY8	Up	0.13±0.02	$5.29x10^{-12}$	Up	0.14±0.06	0.01
PLCXD1	Up	0.02±0.02	0.29	Up	0.10±0.05	0.04
ZBED1	Up	0.12±0.01	$5.60x10^{-16}$	Up	0.08±0.03	$6.86x10^{-3}$

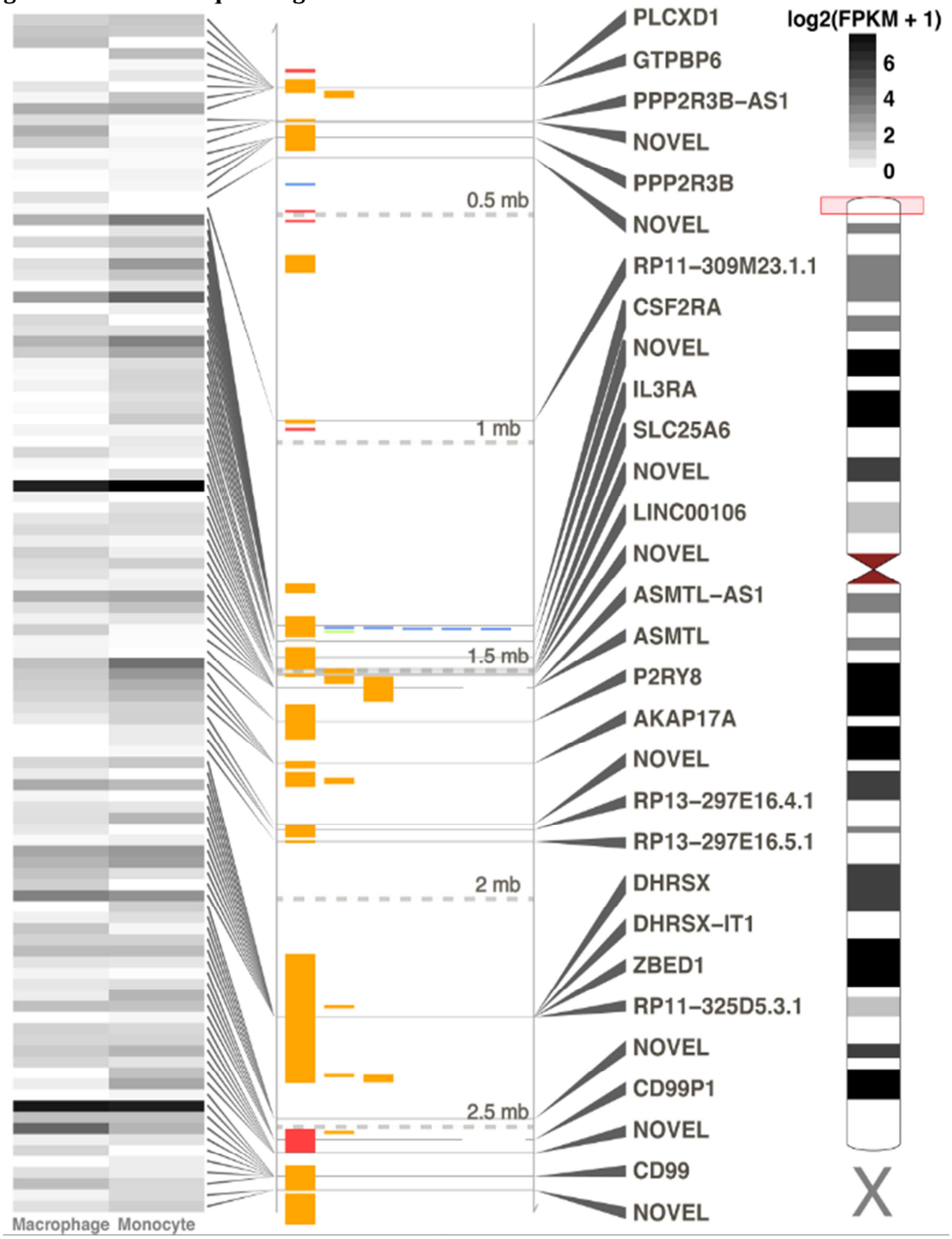
M – males, F – females, Beta – beta coefficient, SE – standard error, P value – level of statistical significance from linear regression analysis.

4.3.6. New generation RNA-sequencing of human monocytes and macrophages

Overall, reads were mapped to 81 PAR1 and 16 PAR2 transcript sequences in human monocytes and macrophages (Table 4.3.9, Table 4.3.12). A majority of PAR1 and PAR2 transcripts showed relatively low expression in monocytes and macrophages when compared to average transcript expression in both types of cells (Figures 4.3.8, Figure 4.3.9). Two PAR genes – *SLC25A6* and *CD99* showed relatively high expression levels (233.83 and 123.07 FPKM, respectively). Nine novel transcripts were identified in PAR1 through RNA-seq (Table 4.3.11).

Of 135 known PAR1 transcripts in 38 genes (including protein coding, pseudogenes, RNA and non-coding RNA genes), 81 showed expression in monocytes (65) or macrophages (61) or both cell types (Table 4.3.10). Four known protein coding genes were not expressed at all in either type of cell (*SHOX*, *CRLF2*, *ASMT* and *XG*).

Figure 4.3.8: Atlas of PAR1 genes expression in human monocytes and macrophages - new generation RNA-sequencing



Left panel – heat maps of each individual PAR1 transcript expressed in either cell type mapping to respective PAR1 genes. Middle panel – alignment of all PAR1 genes based on genomic order (based onEnsembl), in pink – pseudogenes, in orange- protein coding genes. Right panel – symbols of genes that are expressed in at least one of monocytes or macrophages. Far right panel – PAR1 location (in pink) in the context of the cytogenetic banding pattern of the X chromosome. Log2FPKM+1 – units of expression in RNA-sequencing data.

Table 4.3.9: PAR1 transcripts in human monocytes and macrophages – the results of RNA-sequencing

	Gene Symbol	Gene name	Total number of transcripts in Ensembl	Number of transcripts expressed	Transcripts expressed in monocytes	Transcripts expressed in macrophages
1	NCRNA00108	-	1	-	-	-
2	PLCXD1	phosphatidylinositol-specific phospholipase C, X domain containing 1	11	5	4	3
3	GTPBP6	GTP binding protein 6 (putative)	3	4	3	3
4	PPP2R3B-AS1	protein phosphatase 2, regulatory subunit B, antisense RNA 1 (LINC00685)	1	-	-	-
5	PPP2R3B	protein phosphatase 2, regulatory subunit B	10	4	4	1
6	AL732314.1	-	1	2	1	2
7	FABP5P13	fatty acid binding protein 5 pseudogene 13	1	-	-	-
8	KRT18P53	-	1	-	-	-
9	SHOX	short stature homeobox	7	-	-	-
10	RP11-309M23.1	-	1	1	-	1
11	RPL14P5	-	1	-	-	-
12	CRLF2	cytokine receptor-like factor 2	4	-	-	-
13	CSF2RA	colony stimulating factor 2 receptor, alpha (granulocyte-macrophage)	21	11	10	7
14	BX649553.4	-	1	-	-	-
15	BX649553.2	-	1	-	-	-
16	MIR3690	-	1	-	-	-
17	BX649553.1	-	1	-	-	-
18	RNA5SP498	-	-	-	-	-
19	RN7SL355P	-	1	-	-	-
20	IL3RA	interleukin 3 receptor, alpha	3	3	3	2
21	SLC25A6	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6	3	5	3	3

	Gene Symbol	Gene name	Total number of transcripts in Ensembl	Number of transcripts expressed	Transcripts expressed in monocytes	Transcripts expressed in macrophages
22	LINC00106	long intergenic non-protein coding RNA 106	2	2	2	2
23	ASMTL-AS1	*ASMTL* antisense RNA 1	5	2	2	2
24	ASMTL	acetylserotonin O-methyltransferase-like	7	6	4	6
25	P2RY8	purinergic receptor P2Y, G-protein coupled, 8	2	2	1	2
26	AKAP17A	A kinase (PRKA) anchor protein 17A	3	4	4	4
27	ASMT	acetylserotonin O-methyltransferase	4	-	-	-
28	RP13-297E16.3	-	1	-	-	-
29	RP13-297E16.4	-	2	2	2	-
30	RP13-297E16.5	-	1	1	1	-
31	DHRSX	dehydrogenase/reductase (SDR family)	7	7	4	7
32	DHRSX-IT1	*DHRSX*intronic transcript 1 (non-protein coding)	1	1	1	-
33	ZBED1	zinc finger, BED-type containing 1	4	4	3	4
34	RP11-325D5.3	-	1	1	1	1
35	CD99P1	CD99 molecule pseudogene 1	5	5	4	4
36	LINC00102	long intergenic non-protein coding RNA 102	1	-	-	-
37	CD99	*CD99* molecule	9	9	8	7
38	XG	*Xg* blood group	6	-	-	-

Table 4.3.10: Average expression levels of PAR1 gene transcripts in human monocytes and macrophages – the results of RNA-sequencing

Gene/transcript symbol	Transcript length (bp)	Monocytes expression (FPKM)	Macrophages expression (FPKM)	Class code
PLCXD1-001	5287	1.90	1.39	=
PLCXD1-010	587	1.70	1.96	=
PLCXD1-003	284	-	2.38	=
PLCXD1-010	587	2.54	-	=
PLCXD1-004	828	0.15	-	=
GTPBP6-003	2675	0.60	-	=
GTPBP6-201	2677	-	0.65	J
GTPBP6-201	1527	2.06	0.16	=
GTPBP6-001	1186	4.47	3.88	=
LINC00685-001	2378	0.16	0.95	=
LINC00685-001	428	-	3.56	=
PPP2R3B-001	2511	0.14	-	J
PPP2R3B-001	2151	0.13	0.38	=
PPP2R3B-001	458	0.26	-	=
PPP2R3B-009	408	0.17	-	=
RP11-309M23-001	588	-	0.14	=
CSF2RA-006	1816	12.96	3.51	=
CSF2RA-203	2330	0.65	-	J
CSF2RA-203	1575	1.96	1.09	J
CSF2RA-203	2313	0.65	-	J
CSF2RA-001	2291	6.22	0.82	=
CSF2RA-009	802	2.06	0.61	=
CSF2RA-005	1601	0.57	-	=
CSF2RA-012	490	22.26	5.78	=
CSF2RA-011	459	0.40	-	=
CSF2RA-013	391	-	1.03	=
CSF2RA-015	389	0.82	0.76	=
IL3RA-001	1472	0.17	0.24	J
IL3RA-001	1706	0.50	-	=
IL3RA-001	801	0.42	1.13	X
SLC25A6-001	2062	-	0.15	=
SLC25A6-001	2170	0.92	-	J
SLC25A6-001	1583	233.83	117.43	J

Gene/transcript symbol	Transcript length (bp)	Monocytes expression (FPKM)	Macrophages expression (FPKM)	Class code
SLC25A6-001	2173	-	0.40	J
SLC25A6-002	900	0.74	-	=
LINC00106-002	2231	0.88	1.01	=
LINC00106-001	370	0.14	0.41	=
ASMTL-AS1-003	574	1.48	0.95	=
ASMTL-AS1-004	840	0.22	0.70	=
ASMTL-202	2048	0.39	0.25	=
ASMTL-005	2027	4.50	3.73	=
ASMTL-001	2035	2.17	0.85	=
ASMTL-004	573	0.62	0.30	=
ASMTL-202	965	-	1.47	J
ASMTL-002	851	-	0.18	=
P2RY8-002	4272	-	0.22	=
P2RY8-001	4198	16.55	2.37	=
AKAP17A-003	3204	6.72	1.76	=
AKAP17A-001	3232	3.62	1.53	=
AKAP17A-002	2187	2.49	1.24	=
AKAP17A-001	4299	1.37	0.82	J
RP13-297E16.4-001	761	0.47	-	J
RP13-297E16.4-001	445	0.47	-	=
RP13-297E16.5-001	548	0.15	-	=
DHRSX-001	2571	1.89	1.18	=
DHRSX-001	2575	-	0.47	J
DHRSX-004	744	2.80	4.10	=
DHRSX-003	572	-	0.20	=
DHRSX-002	797	0.64	0.76	=
DHRSX-005	539	3.04	0.82	=
DHRSX-006	679	-	0.56	=
DHRSX-IT1-001	651	0.36	-	=
ZBED1-201	4510	6.55	4.65	=
ZBED1-001	4485	5.55	3.20	=
RP11-325D5.3-001	905	0.81	2.19	=
ZBED1-002	2832	-	1.47	=
ZBED1-201	947	7.63	11.24	=
CD99P1-001	1350	-	0.21	J

Gene/transcript symbol	Transcript length (bp)	Monocytes expression (FPKM)	Macrophages expression (FPKM)	Class code
CD99P1-001	943	0.85	0.62	J
CD99P1	984	3.12	0.41	=
CD99P1	561	2.18	2.37	=
CD99P1	3071	0.17	-	=
CD99	1513	0.20	-	J
CD99	1245	123.07	148.94	=
CD99	892	2.37	2.38	=
CD99	918	3.29	20.78	=
CD99	530	0.73	0.31	=
CD99	604	-	1.23	=
CD99	1383	0.41	-	J
CD99	806	0.42	0.78	=
CD99	430	1.15	2.52	=

Data are average expression levels in FPKM (Fragments per kilobase of transcript per million mapped reads), (-) – transcript not expressed, (=) – complete match of intron, (J) – potentially novel isoform (fragment) - at least one splice junction is shared with a reference transcript, (X) - exonic overlap with reference on the opposite strand.

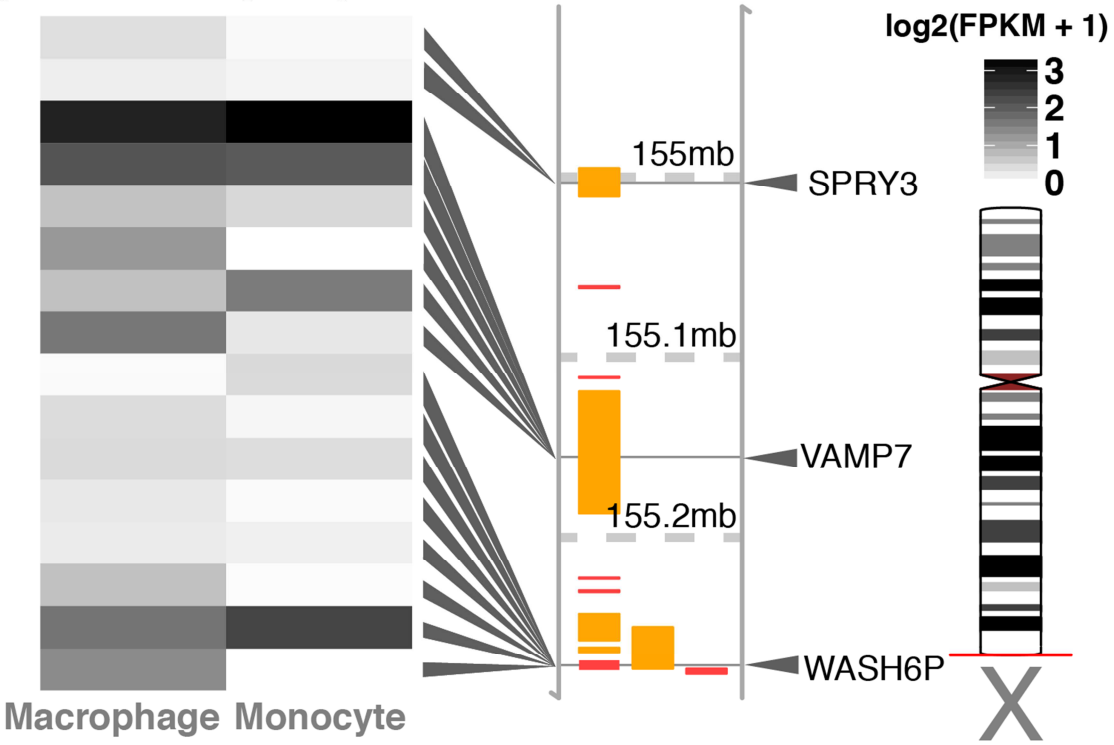
Table 4.3.11: Average expression levels of novel PAR1 gene transcripts in human monocytes and macrophages – the results of RNA-sequencing

Gene/transcript ID	Start (bp)	End (bp)	Transcript length (bp)	Monocytes expression (FPKM)	Macrophages expression (FPKM)
XLOC_056423	284832	285230	398	0.25	1.70
XLOC_056424	366225	366800	575	-	0.77
XLOC_055336	1430497	1430951	454	11.15	3.24
XLOC_055337	1431106	1432068	962	4.43	1.59
XLOC_056425	1433962	1435841	1879	0.94	0.24
XLOC_056426	1436534	1437664	1130	1.25	-
XLOC_056427	1437882	1438764	882	1.23	0.28
XLOC_056428	1438942	1439519	577	0.86	-
XLOC_056429	1439973	1440983	1010	1.04	-
XLOC_056430	1444635	1444968	333	1.77	-
XLOC_056431	1511974	1512667	693	1.06	0.60
XLOC_056432	1519503	1519923	420	0.35	1.44
XLOC_056433	1850514	1850988	474	1.43	0.46
XLOC_055345	2484979	2527216	376	1.55	-
XLOC_055345	2484979	2527216	806	0.83	0.31
XLOC_056434	2484979	2527216	352	0.16	1.95
XLOC_056435	2484979	2527216	382	1.37	1.30
XLOC_056436	2484979	2527216	301	2.63	2.65
XLOC_056437	2484979	2527216	1113	0.59	0.73
XLOC_056438	2578722	2580009	1287	1.05	1.36
XLOC_056439	2580075	2581136	1061	1.28	1.22
XLOC_056440	2581244	2581765	521	3.15	1.68
XLOC_056441	2581887	2582308	421	0.69	1.22
XLOC_056442	2582579	2582912	333	2.30	-
XLOC_056443	2583238	2583478	240	4.45	0.36
XLOC_056444	2662977	2663619	642	1.05	0.24
XLOC_056445	2663764	2664288	524	1.48	0.79

FPKM – Fragments per kilobase of transcript per million mapped reads, unit of expression in RNA sequencing data, (-) – not expressed.

Ensembl identifies 11 genes (including protein coding, pseudogenes, RNA and non-coding RNA genes) with a total of 37 transcripts in PAR2. Of those 16 transcripts showed expression in monocytes (9) or macrophages (15) or both cell types. One known protein coding gene was not expressed at all in either monocytes or macrophages (*IL9R*).

Figure 4.3.9: Atlas of PAR2 genes expression in human monocytes and macrophages - new generation RNA-sequencing



Left panel – heat maps of each individual PAR2 transcript expressed in either cell type mapping to respective PAR2 genes. Middle panel – alignment of all PAR2 genes based on genomic order (based on Ensembl), in pink – pseudogenes, in orange – protein coding genes. Right panel – symbols of genes that are expressed in at least one of monocytes or macrophages. Far right panel – PAR2 location (in pink) in the context of the cytogenetic banding pattern of the X chromosome. Log2FPKM+1 – units of expression in RNA-sequencing data.

Table 4.3.12: PAR2 transcripts in human monocytes and macrophages – the results of RNA-sequencing

	Gene Symbol	Gene name	Total number of transcripts in Ensembl	Number of transcripts expressed	Transcripts expressed – monocytes	Transcripts expressed – macrophages
1	SPRY3	Sprouty homolog 3 (drosophila)	1	2	0	2
2	AMDP1	adenosylmethionine decarboxylase pseudogene 1	1	-	-	-
3	DPH3P2	DPH3, KTI11 homolog (S. cerevisiae) pseudogene 2	1	-	-	-
4	VAMP7	vesicle-associated membrane protein 7	6	6	5	6
5	TCEB1P24	transcription elongation factor B (SIII), polypeptide 1 pseudogene 24	1	-	-	-
6	TRPC6P	transient receptor potential cation channel, subfamily C, member 6 pseudogene	1	-	-	-
7	IL9R	interleukin 9 receptor	6	-	-	-
8	AJ271736.10	-	1	-	-	-
9	WASIR1	WASH and IL9R antisense RNA 1	1	-	-	-
10	WASH6P	WAS protein family homolog 6 pseudogene	16	8	4	7
11	DDX11L16	DEAD/H (Asp-Glu-Ala-Asp/His) box helicase 11 like 16	2	-	-	-

Table 4.3.13: Average expression levels of PAR2 gene transcripts in human monocytes and macrophages – the results of RNA-sequencing

Gene/transcript symbol	Transcript length (bp)	Monocytes expression (FPKM)	Macrophages expression (FPKM)	Class code
SPRY3-201	3302	-	0.29	=
SPRY3-001	9019	-	0.14	=
VAMP7-002	2646	8.68	6.17	=
VAMP7-003	742	3.04	3.27	=
VAMP7-005	696	0.36	0.63	=
VAMP7-004	673	-	1.32	=
VAMP7-001	2496	1.99	0.64	=
VAMP7-006	658	0.19	2.09	=
WASH6P-012	3449	0.33	-	J
WASH6P-012	3069	-	0.31	J
WASH6P-016	4244	0.30	0.34	=
WASH6P-015	3187	-	0.20	=
WASH6P-014	1826	-	0.17	=
WASH6P-008	542	-	0.64	=
WASH6P-010	832	3.98	2.20	=
WASH6P-011	534	-	1.59	=

FPKM - Fragments per kilobase of transcript per million mapped reads, unit of expression in RNA sequencing data, (-) – transcript not expressed (=) – complete match of intron, (J) - potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript.

The lincRNA RP11-309M23.1 in close proximity to rs144253516 that was associated with CAD in females, showed expression in macrophages but not in monocytes. In the Human Body Map 2.0 project (Cabili *et al.* 2011), this lincRNA showed expression in other tissues including brain, breast, kidney, ovary and thyroid (Figure 4.3.10). The lincRNA showed no expression in peripheral white blood in this dataset. The expression data from Human Body Map suggest that this lincRNA might be longer than what Ensembl reports - strong signals were observed downstream of this lincRNA (Figure 4.3.10).

Figure 4.3.10: RP11-309M23.1 lincRNA expression profile across different human tissues in HumanBodyMap 2.0 dataset. RNA sequencing data from the Human Body Map 2.0 dataset (for all different tissue types) are shown. The bottom row shows the Ensembl gene annotations for the RP11-309M23.1lincRNA. Expression abundances were estimated on the gene locus level and are given as raw FPKM.

4.4. Discussion

Both PARs belong to the most un-explored regions of human genome in complex human diseases. The unusual genetic "behaviour" of these regions, their complex biology, small size and very poor coverage in commercial arrays made them unpopular, a genetic "blind-spot" for genetic association studies. They were routinely excluded from GWA studies.

So far only three human diseases have been mapped to PAR1. Isolated short stature, Leri-Weill (Belin *et al*. 1998; Shears *et al*. 1998) and Langer (Belin *et al*. 1998) syndromes are all caused by the functional loss of *SHOX* gene. This gene represents the only known disease locus within the human PAR1 (Blaschke and Rappold, 2006). Two additional studies examined the involvement of PARs in susceptibility to psychiatric disorders. A GWA study in schizophrenia, revealed a strong association signal near *CSF2RA* in the PAR1 region (Lencz *et al*. 2007b) and a linkage study highlighted a new susceptibility locus for bipolar affective disorder very close to *ASMTL* and *ASMT* in PAR1 (Flaquer *et al*. 2010).

The associations between PARs and either risk factors or terminal manifestations of cardiovascular disease have not been explored to date in prior candidate gene studies. Interestingly, several genes in PARs including *IL3RA*, *CSF2RA* appear as strong biological candidates in relation to atherosclerosis. This genetic study is the first robust analysis of association between PARs and predisposition to CAD.

Despite the sequence homology of PARs, gene expression levels in these regions can be different in males and females, which may result in a variety of dose-dependent effects and different functional consequences in both sexes (Carrel and Willard, 2005; Talebizadeh *et al*. 2006; D'Esposito *et al*. 1996). The analysis of an existing genome-

wide microarray expression profiling dataset – Cardiogenics – showed sexual dimorphic differences in expression of ~50% of PAR genes in monocytes and macrophages. Very recently, another study highlighted the sexual gene expression dimorphism of another cardiovascular disease, ischeamic stroke (Stamova *et al.* 2012). There was a sexual dimorphism in expression changes of PAR genes after ischaemic stroke. This was most striking for dehydrogenase/reductase SDR family X-linked (*DHRSX*) and sprouty homolog 3 (*SPRY*3) and a kinase anchor protein 17A (*AKAP17A*) genes. For example at 24 hours after stroke, there was a male-specific up-regulation of *DHRSX* and a female-specific up-regulation of *AKAP17A* (Stamova *et al.* 2012). Although it is not feasible to discriminate whether PAR gene expression can be driven by X or Y chromosome (or both) based on the experimental design of these studies, several PAR genes showed clear differences in expression pattern between both sexes (Stamova *et al*. 2012). These different patterns of PAR expression in women compared with men provided a first glimpse into the sexually dimorphic nature of PAR response to cardiovascular disease.

The most important finding from the genetic meta-analysis of PARs was that rs144253516 showed association with CAD in females. The lead SNP is located in an intergenic region. Four kb away from it, is a long intergenic non-coding RNA gene, RP11-309M23. Long non-coding RNA (lincRNA) is most commonly defined as a non-protein-coding RNA molecule longer than 200 nucleotides (Shi *et al.* 2013). Accumulating evidence suggests that lincRNAs are key regulators in cell differentiation and disease pathways (Huarte *et al.* 2010; Wang *et al*. 2011a; Guttman *et al.* 2011; Prensner *et al.* 2011; Cesana *et al*. 2011; Hu *et al*. 2011; Ng *et al.* 2012; Kretz *et al* 2012; Gupta *et al*. 2010). The transcriptome analysis revealed that RP11-309M23 was expressed in macrophages but not in monocytes. Its expression levels were very low but it is not surprising assuming that its function is regulatory. It is also known that long

non-coding RNAs expression patterns tend to be more tissue-specific than protein-coding genes (Schonrock *et al*. 2012). Its expression in macrophages only, may suggest that differentiation into macrophages, an important initial step in the process of atherosclerosis (Ross, 1993), might be the key stage at which rs144253516 (or any of its proxies) act on the risk of CAD.

This lincRNA contains transposable elements such as long terminal repeat endogenous retroviruses (LTR-ERVs), SINE-Alus and LINE-L1 repetitive sequences. Transposable elements are nucleic acid sequences capable of inserting into genomic DNA (Kelley and Rinn, 2012).They are typically considered "selfish" genomic parasites and occupy 45 to 65% of the human genome (De Koning *et al*. 2011). Whether these transposable elements influence lincRNA sequence is largely unexplored, but various recent studies point to interesting transposable element-associated lincRNA functions (Gong and Maquat, 2011; Cartault *et al*. 2012; Loewer *et al*. 2010). For example a mutated L1 element in a lincRNA is associated with infantile encephalopathy (Cartault *et al*. 2012). In addition, there is a strong enrichment of ERV1 and ERVL-MalR LTR families in this lincRNA. ERVs are remnants of exogenous retrovirus insertions into the germline and contain retroviral protein open reading frames, flanked by transcription-promoting LTRs (Lower *et al*. 1996; Kelley and Rinn, 2012). ERVs usually exhibit position and orientation biases, preferring the 5' end of lincRNA transcripts and sense orientation with the transcript, consequently placing their LTRs in proper position to promote transcription. This suggests that transcription of ERVs may play a role in lincRNA transcriptional regulation (Keller and Rinn, 2012). Characteristically RP11-309M23 lincRNA contains an ERV1-LTR at its 5' end. Alterations in these repetitive elements could affect the functional role of this lincRNA function and can contribute to disease.

However, at this stage, the potential involvement of other genes in close proximity such as *CSF2RA* or *CRLF2* as drivers of the identified association cannot be excluded. Both genes are strong biological candidate genes in relation to CAD. Indeed, because of involvement in inflammatory response, *CRLF2,* a cytokine receptor gene, could have been a promising biological candidate gene; however the RNA-sequencing data showed that it is not expressed in either monocytes or macrophages.

CSF2RA gene is located ~440kb from the lead SNP. The gene encodes the alpha subunit of the heterodimeric receptor for colony stimulating factor 2 (GM-CSF), a cytokine controlling the production, differentiation and function of granulocytes and macrophages (Lencz *et al.* 2007b; Tian *et al.* 2012). The encoded protein is a member of the cytokine family of receptors and contributes to the development and progression of atherosclerosis. This gene has 11 transcripts expressed in monocytes, macrophages or both and shows sexual dimorphism in expression, at least in some of its transcripts. Three *CSF2RA* transcripts were present on the expression platform and two of them showed a significantly different up-regulation in men when compared to women. Up to date, this gene has been associated with schizophrenia (Lencz *et al.* 2007b) and its elevated expression levels were observed in males with stroke compared to healthy controls (Tian *et al*. 2012). *CSF2RA* mRNA level was reported as higher in healthy females than in healthy males in the stomach and lung tissue (Talebizadeh *et al.* 2006). Further studies are required to explain the role of this cytokine in these chronic disease processes and most importantly in relation to the role of PAR1 in CAD.

It is certainly important to identify potential functional effects of rs144253516 in cells of relevance to CAD. To this end, associations between this SNP and both monocyte and macrophage expression of RP11-309M23 lincRNA and *CSF2RA* should be

examined. The rs144253516 is located inside a repeat masked region, a SINE AluY element. This SNP lies in a non-unique DNA area. Examination of both alignability (how often this sequence will align within the whole genome) and uniqueness (how unique is this sequence throughout the reference genome) of the DNA sequence surrounding this SNP, highlighted how difficult it is to design specific primers for this SNP. Indeed, there is no unique base which could be possibly used to anchor rs144253516. The genotyped SNP rs5946608 is in high LD (r^2=0.91) with rs144253516 and is located in a LTR ERV1 repetitive sequence. The rs5946608 is an easier target for genotyping and will be used as a proxy in the future studies.

The association between the lead PAR1 SNP (rs141738136) and CAD in male-stratified analysis did not survive multiple testing correction. This variant mapped to *XG* which encodes the XG blood group antigen, and is located in the boundary of PAR1/MSY. Indeed, three 5' exons reside in the PAR1 and the remaining exons within the MSY. One possibility is that the observed signal is driven by the MSY due to its proximity. Indeed, one of the MSY haplogroups was associated with predisposition to CAD in men (Charchar *et al*. 2012). Alternatively, the signal could be a simple false positive.

It should be acknowledged that similar to other common genetic variants associated with increased risk of cardiovascular disease, rs144253516 on the short arm of the sex chromosomes shows a modest effect size and on its own is unlikely to offer sufficiently high positive predictive value for CAD. Indeed, SNPs that are fairly common in both cases and controls (and lead to a 1.1-1.5 increase in the OR of a disease) are not really specific enough to offer sufficient predictive value (Wald *et al.* 1999; Holmes *et al*. 2011; Charchar *et al*. 2012). The new associations of PAR1 with susceptibility to CAD will require additional independent replication.

Next generation deep sequencing technology has been proven to be a powerful tool for transcriptome analysis of PARs. In this project a detailed atlas of PAR1 and PAR2 genes expression in human monocytes and macrophages was provided. A majority of previous studies exploring the global gene profile change in monocytes and macrophages were based on commercially available microarrays (Schunkert *et al.* 2011; Chachar *et al*. 2012). The coverage of each platform is variable but generally much inferior to RNA-seq. Different microarray platforms only have 30-40% overlap in transcript detection (Barnes *et al.* 2005; Pedotti *et al.* 2008). With the advent of direct ultra-high-throughput sequencing of RNA transcripts, analyses of gene expression are significantly improved. The major advantages of RNA-sequencing are improved detection accuracy (more sensitive), quantification of transcripts with low expression, ability to identify alternative splicing without probe dependency and de *novo* analysis of novel transcripts and long non-coding RNAs (Labaj *et al.* 2011). RNA-sequencing provided the unique opportunity to identify 9 novel transcripts in PAR1 and reveal the expression profiles of miRNAs and lincRNA in human monocytes and macrophages. Further larger scale analyses followed by real-time quantitative PCR validation are necessary to confirm the findings from this analysis and offer more accurate quantitative studies.

There is still a long way to understanding of the exact mechanisms underlying the association between PARs and cardiovascular disease. The collected data provided the first glimpse into their role in CAD. Further larger scale studies will be necessary to replicate the findings from the association analysis. Future studies should focus on DNA-mRNA correlation analysis (eQTL analysis) using appropriately sized samples to clarify which PAR1 transcript(s) is/are the driver of the identified association.

CHAPTER 5

GENERAL DISCUSSION

5. General Discussion

5.1. Lessons from GWA studies on CAD

Genome-wide SNP arrays have been used to identify common risk alleles for CAD. These studies and their meta-analyses have included thousands of patients and healthy individuals and provided the appropriate statistical power to identify fairly common genetic variants associated with CAD. A total of 46 chromosomal regions affecting CAD risk and a further 104 independent variants (r^2<0.2) strongly associated with CAD at a 5% FDR were reported (Deloukas *et al*. 2013; Kessler *et al.* 2013). However, the collective contribution of these variants to CAD is very modest and explains only a small proportion of the overall heritability (up to 10.6%) (Peden and Farrall, 2011; Deloukas *et al*. 2013). To date, several important and novel insights into CAD biology have been provided by GWA studies.

5.2. Un-explored regions of the human genome

Despite their overall success in genetic discovery, GWA studies have overlooked several regions of the human genome. For example, sex chromosomes have been almost routinely excluded from these investigations. In addition, a majority of GWA studies were simply not powered to detect the effect of low-frequency/rare variants on complex disorders. As a result, the contribution of GWA studies to our knowledge on the role of both sex chromosomes and rare alleles in complex polygenic cardiovascular disorders has been limited.

The two sex chromosomes lag behind autosomes in GWA findings despite being represented on all current GWA microarray platforms. Many reasons accounted for this

exclusion. For example, male specific region (95% the Y chromosome length) is haploid in nature and is transmitted from one generation to another along male lineage only (Charchar *et al.* 2012). It does not recombine during meiosis with the X chromosome. The traditional methods of LD-based association mapping applied to autosomal chromosomes cannot be used to examine the variation between MSY and human disease (Charchar *et al.* 2012) and the most suitable approach is the investigation of the Y chromosome phylogenetic tree.

Unlike Y, X chromosome was included in a few GWA studies. However, of a total of >2,800 GWA signals, only 15 (~0.5%) were reported on the X chromosome (Wise *et al*. 2013). Although similar in size to bigger autosomes (155Mb), X exhibits several major differences when compared to the rest of the human chromosomes. For example, a special feature of the X is the process of inactivation as a mechanism of dosage compensation at gene expression level (Chow *et al*. 2005). Due to lack of consensus on how to handle/equalise X chromosomal allele/genotype data across both sexes, X chromosome was usually excluded from GWA studies. Unfortunately, a majority of current GWA platforms are still poorly designed for this region (Wise *et al.* 2013).

Similar to X and Y, PARs were almost completely excluded from previous genetic association analyses, mostly because of their modest size, location within the sex chromosomes and unusual genetic behaviour during meiosis. The coverage for PARs in GWA arrays remains very poor.

These factors together created a significant barrier precluding their inclusion in analyses of complex diseases. In addition, the plethora of findings obtained from the autosomal genome alone, led many researchers to overlook the X, Y and PARs.

ROHs, long segments of uninterrupted consecutive homozygous SNPs and potential signatures for recessive variants were used before in genetic mapping of rare recessive disorders in families. It is now apparent that ROHs are more common in outbred populations than previously thought (Gibson *et al*. 2006, McQuillan *et al*. 2008). Only the last few years brought discoveries from genome-wide analysis of ROHs and complex disorders (Ku *et al.* 2011). For example, ROHs were associated with human height (Yang *et al.* 2010), schizophrenia (Lencz *et al.* 2007a; Keller *et al.* 2012) and Alzheimer's disease (Nalls *et al.* 2009a; Sims *et al.* 2011). It would be fair to acknowledge that the number of these studies is much lower than "traditional" GWA studies and certain diseases, for example CAD, have never been examined in ROH-based analyses.

A majority of low-frequency/rare variants were excluded from GWA studies, usually at the level of quality control checking.

5.3. The sex chromosomes and CAD

The strongest level of evidence for association between sex chromosomes and CAD comes from the recent phylogenetic analysis and gene expression studies on MSY (Charchar *et al.* 2012). Indeed, haplogroup I of the human Y chromosome was associated with 50% risk of CAD in men of European ancestry, possibly through its effect on immunity and inflammation (Charchar *et al.* 2012). The very recent analysis from our group further revealed that CAD predisposing haplogroup I of the Y chromosome was associated with down-regulation of ubiquitously transcribed tetratricopeptide repeat containing Y-linked (*UTY*) and protein kinase, y-linked, pseudogene (*PRKY*) genes in macrophages (Bloomer *et al.* 2013). The association

between haplogroup I of the Y chromosome and CAD was independent of traditional cardiovascular risk factors such as BMI, blood pressure, total cholesterol, HDL-C, triglycerides, LDL-C, and glucose in European men (Bloomer *et al.* 2013). These data put the sex chromosome on the map of genetic predisposition to CAD.

Of all human chromosomes, X contains the largest number of genes related to immunity and inflammation (Bianchil *et al*. 2012), both of which are recognised biological components of atherosclerosis and CAD. X chromosome anomalies are a risk factor of pro-ischaemic phenotype of cardiovascular disease – women with Turner syndrome have two-fold increase in risk of CAD when compared to the general population (Gravholt *et al.* 1998). Men and women differ in expression patterns of X chromosome genes in blood cells after ischaemic stroke (Stamova *et al.* 2012). However, so far there is no published genome-wide based evidence for association with CAD.

The collected body of evidence for potential biological role of the sex chromosomes (mostly MSY) in CAD was a major trigger to explore possibly mostly unexplored part of both X and Y – PARs. These two small parts of the human DNA have proved to be more exciting and informative than their size implies. Indeed, genetic variation in PAR1 was associated with CAD risk in sex-specific manner in the meta-analysis of CARDIoGRAM Consortium. Specifically, common genetic variants were associated with CAD in women but not in men after correction for multiple testing in the meta-analysis. Although the biological foundations of sex specific patterns of association between PAR1 with CAD are not clear, sex-stratified comparative gene expression analysis in human monocytes and macrophages revealed statistically significant differences in PAR1 gene expression levels between men and women. This further emphasises the importance of sex-specific analysis of this region in further studies on

complex disorders. Interestingly, RNA-sequencing of human macrophages uncovered expression of lincRNA in close proximity to the region of PAR1 where an association with CAD was identified. Further studies should verify whether this lincRNA is a mediator of the association between PAR1 and CAD. Results from GWA studies suggest that a significant proportion of genetic variation (>80%) associated with complex diseases falls in non-coding regions of the genome (Hindorff *et al.* 2009; Bernstein *et al.* 2012). Most of these transcripts have little or no protein-coding capacity and may hold the key to understanding the regulatory complexity inherent to advanced biological networks (Amaral *et al.* 2008). Many non-coding RNAs have fundamental indices of functionality such as regulation by tissue-restricted transcription factors, dynamic developmental and cell type-specific expression patterns, localisation to specific subcellular compartments, association with chromatin signatures indicative of active transcription, conservation of promoters, structure and genomic location and association with human disease (Mattick, 2009). All of these renewed enthusiasm for exploring functional implications of these transcripts that ar present are weakly understood.

5.4. ROHs and CAD – genetic signature of recessive variants?

Recessive effects that make a substantial contribution to susceptibility to disease have been ignored by GWA studies that focused on additive mode of inheritance. The presence of ROHs in outbred populations provided evidence for a recessive component to the human genetic architecture. Evaluation of the distribution of ROHs in this project showed their over-representation favouring increased CAD risk. This suggests that accumulation of recessive alleles may increase the risk of CAD.

5.5. Low-frequency/rare variants and CAD

The role of low-frequency/rare variants in relation to CAD has not been examined extensively. There is no doubt that low-frequency/rare variants exist and play a role in CAD biology. For example, low frequency SNP rs3798220 (MAF~3%) in *LPA* gene, is consistently associated with CAD risk (Clarke *et al*. 2009; The IBC 50K CAD Consortium, 2011). Taking advantage of the already available information on low-frequency and rare variants produced by the HumanCVD 50K chip, a search for their association was conducted using an aggregate method where the overall accumulation of low-frequency/rare variants in a locus were compared across CAD patients and controls. The previously reported association in *LPA* was confirmed in populations of European ancestry and new associations in *F10, F7* and *TRAF2* genes were discovered in South Asians. These findings highlight the need of gigantic discovery sample sizes to detect low-frequency/rare variants - ~20,000 individuals from the IBC 50K CAD Consortium were clearly an insufficient resource in terms of power. Finding rare variants is a challenging task, but the advent of next generation sequencing technologies has markedly facilitated discovery of rare variants. A recent study based on whole-genome-based analysis indicated that 7.8% of HDL-cholesterol heritability is attributable to rare variants (MAF<1%) (Morrison *et al.* 2013). A major part of HDL-C heritability (61.8%) was accounted for by common variants. These results suggest that many common variants with individual small effects primarily determine genetic architecture of HDL-C. The data from this study also appear to suggest that rare variants may play a lesser role in HDL-C biology than was initially suggested. Whether these findings apply to other cardiovascular risk factors or indeed CAD remains to be established.

5.6. Common versus rare polymorphisms and risk prediction

In the past, it was predicted that the identification of common genetic variants could eventually lead to stable prediction risk models with significant individual and public health implications (Bowles and Marteau, 1999; Chatterjee *et al*. 2013). The complexity of the genetic architecture of CAD, with multiple variants conferring usually modest increases in its relative risk (Manolio *et al*. 2009) makes the prediction problematic - odds ratio of 1.3-1.7 typical for common variants identified in GWA studies do not offer sufficient specificity in risk classification (Jakobsdottir *et al*. 2009; Manolio, 2009). The precise fraction of risk attributable to single common genetic polymorphism is difficult to determine since all people carry numerous risk alleles. Genetic risk scores aggregate and weight the number of risk alleles carried by each individual - the overall burden of them may have a substantial impact on the predisposition to CAD at the population level (Kessler *et al.* 2013).

Thousands of common susceptibility variants for a wide spectrum of complex traits indicate that they collectively have low predictive value (van Hoek *et al*. 2008; Lango *et al*. 2010; Speliotes *et al*. 2010; Teslovich *et al*. 2010; Jostins and Barrett, 2011; Kraft and Hunter, 2009; Chatterjee *et al*. 2013). Although risk prediction models will continue to improve as total sample sizes increase, the improvement will be slow and modest (Chatterjee *et al.* 2013) even when thousands of them with individually undetectable effect sizes will be encompassed in risk prediction.

5.7. Final conclusions

There is still a long way to elucidating genetic background of CAD. The experiments conducted here revealed novel associations between CAD and ROHs as well as PAR1 and highlighted the difficulties in the analysis of rare alleles in search of novel genes underlying susceptibility to CAD. Most importantly, these data showed that in-depth exploration of regions commonly neglected by previous GWA studies has a potential to provide new insights into genetic architecture of common complex diseases. With the advent of new technologies including whole-genome DNA and RNA sequencing, future studies should focus on further fine mapping and functional characterisation of these regions to bring us closer to full understanding of their roles in CAD.

References

Abbott RD, Behrens GR, Sharp DS, et al. (1994) Body mass index and thromboembolic stroke in nonsmoking men in older middle age. The Honolulu Heart Program. Stroke; a journal of cerebral circulation 25:2370–2376.

Abecasis GR, Ghosh D, Nichols TE (2005) Linkage disequilibrium: ancient history drives the new genetics. Human heredity 59:118–124.

Abecasis GR, Noguchi E, Heinzmann A, et al. (2001) Extent and distribution of linkage disequilibrium in three genomic regions. American journal of human genetics 68:191–197.

Abifadel M, Varret M, Rabès J-P, et al. (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nature genetics 34:154–156.

Abu-Amero S, Monk D, Frost J, et al. (2008) The genetic aetiology of Silver-Russell syndrome. Journal of medical genetics 45:193–199.

Alfakih K, Brown B, Lawrance RA et al. (2007) Effect of a common X-linked angiotensin II type 2-receptor gene polymorphism (-1332 G/A) on the occurrence of premature myocardial infarction and stenotic atherosclerosis requiring revascularization. Atherosclerosis 195:e32-8.

Alkuraya FS (2010) Autozygome decoded. Genetics in medicine : official journal of the American College of Medical Genetics 12:765–771.

Almdal T, Scharling H, Jensen JS, et al. (2004) The Independent Effect of Type 2 Diabetes Mellitus on Ischemic Heart Disease, Stroke, and Death - A population-based study of 13000 men and women with 20 years of follow-up. Archives of internal medicine 164:1422–1426.

Altshuler D, Daly M (2007) Guilt beyond a reasonable doubt. Nature genetics 39:813–815.

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science (New York, NY) 322:881–888.

Altug-Teber O, Dufke A, Poths S, et al. (2005) A rapid microarray based whole genome analysis for detection of uniparental disomy. Human mutation 26:153–159.

Amaral PP, Dinger ME, Mercer TR, et al. (2008) The eukaryotic genome as an RNA machine. Science 319:1787–1789.

Anderson KM, Odell PM, Wilson PW, et al. (1991) Cardiovascular disease risk profiles. American heart journal 121:293–8.

Andreotti F, Burzotta F, Maseri A (1999) Fibrinogen as a marker of inflammation: a clinical view. Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis 10 Suppl 1:S3–4.

Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nature reviews Genetics 7:277–282.

Ardissino D, Mannucci PM, Merlini PA, et al. (1999) Prothrombotic genetic risk factors in young survivors of myocardial infarction. Blood 94:46-51.

Arking DE, Chakravarti A (2009) Understanding cardiovascular disease through the lens of genome-wide association studies. Trends in genetics : TIG 25:387–394.

Arnett DK, Baird AE, Barkley RA, et al. (2007) Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational. Circulation 115:2878–2901.

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annual review of genetics 44:293–308.

Auton A, Bryc K, Boyko AR, et al. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome research 19:795–803.

Baccarelli A, Rienstra M, Benjamin EJ (2010) Cardiovascular epigenetics:basic concepts and results from anaimal and human studies. Circulation cardiovascular genetics 3:567-73.

Bacolod MD, Schemmann GS, Wang S, et al. (2008) The signatures of autozygosity among patients with colorectal cancer. Cancer research 68:2610–2621.

Bansal V, Libiger O, Torkamani A, et al. (2010) Statistical analysis strategies for association studies involving rare variants. Nature reviews Genetics 11:773–785.

Barnes M, Freudenberg J, Thompson S, et al. (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic acids research 33:5914–5923.

Barnoya J, Glantz S (2005) Cardiovascular effects of secondhand smoke: nearly as large as smoking. Circulation 111:2684–2698.

Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nature genetics 38:659–662.

Barrett-Connor EL, Cohn BA, Wingard DL, et al. (1991) Why is diabetes mellitus a stronger risk factor for fatal ischemic heart disease in women than in men? The Rancho Bernardo Study. JAMA : the journal of the American Medical Association 265:627–631.

Barroso I (2005) Genetics of Type 2 diabetes. Diabetic medicine : a journal of the British Diabetic Association 22:517–535.

Barton NH, Turelli M (1989) Evolutionary quantitative genetics: how little do we know? Annual review of genetics 23:337–370.

Bassuk SS, Manson JE (2005) Epidemiological evidence for the role of physical activity in reducing risk of type 2 diabetes and cardiovascular disease. Journal of applied physiology 99:1193–1204.

Beckman JA, Creager MA, Libby P (2002) Diabetes and atherosclerosis: epidemiology, pathophysiology, and management. JAMA : the journal of the American Medical Association 287:2570–81.

Belin V, Cusin V, Viot G, et al. (1998) SHOX mutations in dyschondrosteosis (Leri-Weill syndrome). Nature genetics 19:67–69.

Benayoun L, Spiegel R, Auslender N, et al. (2009) Genetic heterogeneity in two consanguineous families segregating early onset retinal degeneration: the pitfalls of homozygosity mapping. American journal of medical genetics Part A 149A:650–656.

Benyamin B, McRae AF, Zhu G, et al. (2009) Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. American journal of human genetics 84:60–65.

Berge KE, Tian H, Graf GA, et al. (2001) Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. Science 290:1771-5.

Berger JS, Jordan CO, Lloyd-Jones D, et al. (2010) Screening for cardiovascular risk in asymptomatic patients. Journal of the American College of Cardiology 55:1169–1177.

Bernstein BE, Birney E, Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Bertram L, Böckenhoff A, Demuth I, et al. (2013) Cohort Profile: The Berlin Aging Study II (BASE-II). International journal of epidemiology 1–10.

Bezzina CR, Pazoki R, Bardai A, et al. (2010) Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction. Nature genetics 42:688–91.

Bianchi I, Lleo A, Gershwin ME, et al. (2012) The X chromosome and immune associated genes. Journal of autoimmunity 38:J187–92.

Binder G (2011) Short stature due to SHOX deficiency: genotype, phenotype, and therapy. Hormone research in pædiatrics 75:81–89.

Bittles AH (2003) Consanguineous marriage and childhood health. Developmental Medicine & Child Neurology 45:571–576.

Bittles AH (2002) Endogamy, consanguinity and community genetics. Journal of genetics 81:91–98.

Bittles AH (2008) A community genetics perspective on consanguineous marriage. Community genetics 11:324–330.

Bittles AH, Black ML (2010) Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proceedings of the National Academy of Sciences of the United States of America 107 Suppl: 1779–1786.

Björck L, Rosengren A, Bennett K, et al. (2009) Modelling the decreasing coronary heart disease mortality in Sweden between 1986 and 2002. European Heart Journal 30:1046–1056.

Blackwell K, Zhang L, Thomas GS, et al. (2009) TRAF2 phosphorylation modulates tumor necrosis factor alpha-induced gene expression and cell resistance to apoptosis. Molecular and cellular biology 29:303–314.

Blair SN, Kohl HW, Paffenbarger RS, et al. (1989) Physical fitness and all-cause mortality. A prospective study of healthy men and women. JAMA : the journal of the American Medical Association 262:2395–2401.

Blaschke RJ, Rappold G (2006) The pseudoautosomal regions, SHOX and disease. Current opinion in genetics & development 16:233–239.

Bloomer LDS, Nelson CP, Eales J, et al. (2013) Male-specific region of the Y chromosome and cardiovascular risk: phylogenetic analysis and gene expression studies. Arteriosclerosis, thrombosis, and vascular biology 33:1722–7.

Bockukova EG, Huang N, Keogh J, et al. (2010) Large, rare chromosomal deletions associated with severeearly-onset obesity. Nature 463:666-70.

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics 40:695–701.

Boland LL, Folsom AR, Sorlie PD, et al. (2002) Occurrence of unrecognized myocardial infarction in subjects aged 45 to 65 years (the ARIC study). The American journal of cardiology 90:927–931.

Bondy CA (2012) Aortic coarctation and coronary artery disease: the XY factor. Circulation 126:5–7.

Bondy CA (2008) Congenital cardiovascular disease in Turner syndrome. Congenital heart disease 3:2–15.

Bonnefond A, Clément N, Fawcett K, et al. (2012) Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nature genetics 44:297–301.

Boomsma D, Busjahn A, Peltonen L (2002) Classical twin studies and beyond. Nature reviews Genetics 3:872–882.

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics 33 Suppl:228–237.

Boushey CJ, Beresford SA, Omenn GS, et al. (1995) A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. Probable benefits of increasing folic acid intakes. JAMA : the journal of the American Medical Association 274:1049–1057.

Bowles Biesecker B, Marteau TM (1999) The future of genetic counselling: an international perspective. Nature genetics 22:133–137.

British Cardiac Society (2005) JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. Heart. pp v1–52.

Broadbent HM, Peden JF, Lorkowski S, et al. (2008) Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Human molecular genetics 17:806–814.

Broeckel U, Hengstenberg C, Mayer B, et al. (2002) A comprehensive linkage analysis for myocardial infarction and its related risk factors. Nature genetics 30:210–214.

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. The American Journal of Human Genetics 65:1493–1500.

Brown MS, Goldstein JL (1986) A receptor-mediated pathway for cholesterol homeostasis. Science 232:34–47.

Browning SR, Browning BL (2012) Identity by descent between distant relatives: detection and applications. Annual review of genetics 46:617–633.

Browning SR, Thompson EA. (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190:1521-31

Bruce S, Leinonen R, Lindgren CM, et al. (2005) Global analysis of uniparental disomy using high density genotyping arrays. Journal of medical genetics 42:847–851.

Bugert P, Hoffmann MM, Winkelmann BR, et al. (2003) The variable number of tandem repeat polymorphism in the P-selectin glycoprotein ligand-1 gene is not associated with coronary heart disease. Journal of Molecular Medicine 81:495-501.

Burnett JR, Hooper AJ (2008) Common and rare gene variants affecting plasma LDL cholesterol. The Clinical biochemist Reviews / Australian Association of Clinical Biochemists 29:11–26.

Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development 25:1915–1927.

Cai Q, Chen J, Ma H, et al. (2000) Association of coagulation Factor VII with the risk of myocardial infarction in the Chinese. Chinese Medical journal 113:1059-1063.

Cambien F, Tiret L (2007) Genetics of cardiovascular diseases: from single mutations to the whole genome. Circulation 116:1714–1724.

Campbell H, Carothers AD, Rudan I, et al. (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Human molecular genetics 16:233–41.

Campbell H, Manolio T (2007) Commentary: rare alleles, modest genetic effects and the need for collaboration. International journal of epidemiology 36:445–448.

Campbell H, Rudan I, Bittles AH, Wright AF (2009) Human population structure, genome autozygosity and human health. Genome medicine 1:91.

Capewell S, Hayes DK, Ford ES, et al. (2009) Life-years gained among US adults from modern treatments and changes in the prevalence of 6 coronary heart disease risk factors between 1980 and 2000. American Journal of Epidemiology170:229–236.

Cargill M, Altshuler D, Ireland J, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature genetics 22:231–238.

Carr IM, Flintoff KJ, Taylor GR, et al. (2006) Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. Human mutation 27:1041–1046.

Carr IM, Sheridan E, Hayward BE, et al. (2009) IBDfinder and SNPsetter: tools for pedigree-independent identification of autozygous regions in individuals with recessive inherited disease. Human mutation 30:960–967.

Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434:400–404.

Cartault F, Munier P, Benko E, et al. (2012) Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. Proceedings of the National Academy of Sciences of the United States of America 109:4980–4985.

Castro R, River I, Struys EA, et al. (2003) Increased himocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease. Clinical chemisty 49:1292-6.

Cesana M, Cacchiarelli D, Legnini I, et al. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147:358–369.

Chambers JC, Elliott P, Zabaneh D, et al. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nature genetics 40:716–8.

Charchar FJ, Bloomer LDS, Barnes T a, et al. (2012) Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. Lancet 379:915–922.

Charchar FJ, Svartman M, El-Mogharbel N, et al. (2003) Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome research 13:281–286.

Charchar FJ, Tomaszewski M, Lacka B, et al. (2004) Association of the human Y chromosome with cholesterol levels in the general population. Arteriosclerosis, thrombosis and vascular biology 24:308-12.

Charchar FJ, Tomaszewski M, Padmanabhan S, et al. (2002) The Y chromosome effect on blood pressure in two European populations. Hypertension 39:353-6.

Charchar FJ, Tomaszewski M, Strahorn P, et al. (2003) Y is there a risk to being male? Trends endocrinology metabolism 14:163-8.

Charchar FJ, Zimmerli LU, Tomaszewski M (2008) The pressure of finding human hypertension genes: new tools, old dilemmas. Journal of human hypertension 22:821–828.

Charlesworth B (1991) The evolution of sex chromosomes. Science 251:1030-3.

Charlesworth B, Charlesworth D (1999) The genetic basis of inbreeding depression. Genetical research 74:329–340.

Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118–128.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nature reviews Genetics 10:783–796.

Chasman DI, Shiffman D, Zee RYL, et al. (2009) Polymorphism in the apolipoprotein(a) gene, plasma lipoprotein(a), cardiovascular disease, and low-dose aspirin therapy. Atherosclerosis 203:371–6.

Chatterjee N, Wheeler B, Sampson J, et al. (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature genetics 45:400–5, 405e1–3.

Chen VM (2013) Tissue factor de-encryption, thrombus formation, and thiol-disulfide exchange. Seminars in Thrombosis and Hemostasis 39:40-7.

Chobanian A V, Bakris GL, Black HR, et al. (2003) Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension 42:1206–1252.

Chow JC, Yen Z, Ziesche SM, Brown CJ (2005) Silencing of the mammalian X chromosome. Annual review of genomics and human genetics 6:69–92.

Ciccodicola a, D'Esposito M, Esposito T, et al. (2000) Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Human molecular genetics 9:395–401.

Clarke R, Peden JF, Hopewell JC, et al. (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. The New England journal of medicine 361:2518–2528.

Coggins CR (1998) A prospective study of passive smoking and coronary heart disease. Circulation 97:1870–1873.

Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH (2006a) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. The New England journal of medicine 354:1264–1272.

Cohen JC, Kiss RS, Pertsemlidis A, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872.

Cohen JC, Pertsemlidis A, Fahmi S, et al. (2006b) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences of the United States of America 103:1810–5.

Collins R, Peto R, MacMahon S, et al. (1990) Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. Lancet 335:827-38.

Collin RWJ, van den Born LI, Klevering BJ, et al. (2011) High-resolution homozygosity mapping is a powerful tool to detect novel mutations causative of autosomal recessive RP in the Dutch population. Investigative ophthalmology & visual science 52:2227–2239.

Conroy R (2003) Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European Heart Journal 24:987–1003.

Cooke HJ, Brown WR, Rappold GA (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. Nature 317:687-92.

Cooper J a, Miller GJ, Humphries SE (2005) A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. Atherosclerosis 181:93–100.

Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nature genetics 10:392-404.

Coronary Artery Disease (C4D) Genetics Consortium (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature genetics 43:339-44.

Craddock N, Hurles ME, Cardin N, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464:713-20.

Crawford DC, Carlson CS, Rieder MJ, et al. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. American journal of human genetics 74:610–22.

Critchley JA, Capewell S (2003) Mortality risk reduction associated with smoking cessation in patients with coronary heart disease. JAMA : the journal of the American Medical Association 290:86–97.

Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. BMC genetics 8:67.

Curtis D, Vine E, Knight J (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Annals of human genetics 72:261–278.

D'Esposito M, Ciccodicola A, Gianfrancesco F, et al. (1996) A synaprobrevin-like gene in the Xq28 pseudoautosomal region undergoees X inactivation. Nature genetics 13:227–229.

Danesh J, Collins R, Peto R (2000) Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies. Circulation 102:1082–1085.

Darvasi A, Shifman S (2005) The beauty of admixture. Nature genetics 37:118–9. doi: 10.1038/ng0205-118

Dauchet L, Amouyel P, Hercberg S, Dallongeville J (2006) Fruit and vegetable consumption and risk of coronary heart disease: a meta-analysis of cohort studies. The Journal of nutrition 136:2588–2593.

Davie EW, Fujikawa K, Kisiel W (1991) The coagulation cascade: initiation, maintenance, and regulation. Biochemistry 30:10363-70.

Davies RW, Dandona S, Stewart AFR, et al. (2010) Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. Circulation Cardiovascular genetics 3:468–474.

Deloukas P, Kanoni S, Willenborg C, et al. (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. Nature genetics 45:25–33.

De Koning PJ, Gu W, Castoe T, et al. (2011) Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics 7:e1002384.

De Lorgeril M, Salen P, Martin JL, et al. (1999) Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction: final report of the Lyon Diet Heart Study. Circulation 99:779 –785.

Diet Heart Study. *Circulation*. 1999;99:779 –785.

DeRose MA, Roff DA (1999) A comparison of inbreeding depression in life-history and morphological traits in animals. Evolution 53: 1288–1292.

Di Castelnuovo A, D'Orazio A, Amore C, et al. (2000) The decanucleotide insertion/deletion polymorphism in the promoter region of the coagulation Factor VII gene and the risk of familial myocardial infarction. Thrombosis Research 98:9-17.

Dickson SP, Wang K, Krantz I, et al. (2010) Rare variants create synthetic genome-wide associations. PLoS Biology 8:e1000294.

Ding K, Kullo IJ (2009) Genome-wide association studies for atherosclerotic vascular disease and its risk factors. Circulation Cardiovascular genetics 2:63–72.

Do R, Kathiresan S, Abecasis GR (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Human molecular genetics 21:R1-9.

Ducroq D, Shalev S, Habib A, et al. (2006) Three different ABCA4 mutations in thsame large family with several consanguineous loops affected with autosomal recessive cone-rod dystrophy. European journal of human genetics : EJHG 14:1269–1273.

Easton DF, Pooley KA, Dunning AM et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-1093.

Ehret GB, Munroe PB, Rice KM, et al. (2011) International Consortium for Blood Pressure Genome-Wide Association Studies; genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478:103–9.

Eichler EE, Flint J, Gibson G, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 11:446–450.

Ellis NA, Tippett P, Petty A, et al. (1994a) PBDX is the XG blood group gene. Nature genetics 8:285–290.

Ellis NA, Ye T-Z, Patton S, et al. (1994b) Cloning of PBDX, and MIC2-related gene that spans the pseudoautosomal boundary on chromosome Xp. Nature genetics 6:394–400.

Enciso-Mora V, Hosking FJ, Houlston RS (2010) Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. European journal of human genetics : EJHG 18:909–914.

Erdmann J, Grosshennig A, Braund PS, et al. (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nature genetics 41:280–282.

Erdmann J, Willenborg C, Nahrstaedt J, et al. (2011) Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. European heart journal 32:158–68.

Evans A, Van Baal GCM, McCarron P, et al. (2003) The genetics of coronary heart disease: the contribution of twin studies. Twin research : the official journal of the International Society for Twin Studies 6:432–441.

Evans DM, Barrett JC, Cardon LR (2008) To what extent do scans of non-synonymous SNPs complement denser genome-wide association studies? European journal of human genetics : EJHG 16:718–723.

Evans DM, Cardon LR (2004) Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. American journal of human genetics 75:687–692.

Evers C, Heidemann PH, Dunstheimer D, et al. (2011) Pseudoautosomal inheritance of Léri-Weill syndrome: what does it mean? Clinical genetics 79:489–494.

Fadini GP, de Kreutzenberg SV, Tiengo A, Avogaro A (2009) Why to screen heart disease in diabetes. Atherosclerosis 204:11–15.

Farrall M, Green FR, Peden JF, et al. (2006) Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. PLoS genetics 2:e72.

Feinberg AP, Irizarry RA, Fradin D, et al. (2010) Personalized epigenomic signatures that are stable over time and covary body mass index. Scence translational medicine 2:49ra67.

Feng D, Tofler GH, Larson MG, et al. (2000) Factor VII gene polymorphism, Factor VII levels, and prevalent cardiovascular disease: the Framingham Heart Study. Arteriosclerosis Thrombosis and Vascular Biology 20:593-600.

Fields LE, Burt VL, Cutler J a, et al. (2004) The burden of adult hypertension in the United States 1999 to 2000: a rising tide. Hypertension 44:398–404.

Filippini F, Rossi V, Galli T, et al. (2001) Longins: a new evolutionary conserved VAMP family sharing a novel SNARE domain. Trends in biochemical sciences 26:407–409.

Finucane MM, Stevens G a, Cowan MJ, et al. (2011) National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9·1 million participants. Lancet 377:557–567.

Fischer M, Broeckel U, Holmer S, et al. (2005) Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. Circulation 111:855–862.

Flaquer A, Fischer C, Wienker TF (2009) A new sex-specific genetic map of the human pseudoautosomal regions (PAR1 and PAR2). Human heredity 68:192–200.

Flaquer A, Jamra RA, Etterer K, et al. (2010) A new susceptibility locus for bipolar affective disorder in PAR1 on Xp22.3/Yp11.3. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics 153B:1110–1114.

Folsom AR, Wu KK, Rosamond WD, et al. (1997) Prospective study of hemostatic factors and incidence of coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. Circulation 96:1102–1108.

Francke S, Manraj M, Lacquemant C, et al. (2001) A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. Human molecular genetics 10:2751–2765.

Frazer KA, Ballinger DG, Cox DR, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nature reviews Genetics 10:241–251.

Freedman DS, Khan LK, Serdula MK, et al. (2002) Trends and correlates of class 3 obesity in the United States from 1990 through 2000. JAMA : the journal of the American Medical Association 288:1758–1761.

Freije D, Helms C, Watson MS, et al. (1992) Identification of a second pseudoautosomal region near the Xq and Yq telomeres. Science 258:1784–1787.

Freimer N, Sabatti C (2004) The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nature genetics 36:1045–1051.

Frishberg Y, Ben-Neriah Z, Suvanto M, et al. (2007) Misleading findings of homozygosity mapping resulting from three novel mutations in NPHS1 encoding nephrin in a highly inbred community. Genetics in medicine : official journal of the American College of Medical Genetics 9:180–184.

Fryar CD, Hirsch R, Eberhardt MS, et al. (2010) Hypertension, high serum total cholesterol, and diabetes: Racial and ethnic prevalence differences in US adults, 1999–2006. NCHS data brief, no 36. Hyattsville, MD: National Center for Health Statistics.

Furrow RE, Christiansen FB, Feldman MW (2011) Environment-sensitive epigenetics and the heritability of complex diseases. Genetics 189:1377-87.

Garcia CK, Wilund K, Arca M, et al. (2001) Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. Science 292:1394-1398.

Garrod AE (2002) The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article]. Yale Journal of Biology and Medicine 75:221-31.

Genest JJ, Martin-Munley SS, McNamara JR, et al. (1992) Familial lipoprotein disorders in patients with premature coronary artery disease. Circulation 85:2025–2033.

Génin E, Todorov AA, Clerget-Darpoux F (1998) Optimization of genome search strategies for homozygosity mapping: influence of marker spacing on power and threshold criteria for identification of candidate regions. Annals of human genetics 62:419–429.

Gibbs JR, Singleton A (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS genetics 2:e150.

Gibson G (2012) Rare and common variants: twenty arguments. Nature reviews Genetics 13:135–145.

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Human molecular genetics 15:789–795.

Girard SL, Gauthier J, Noreau A, et al. (2011) Increased exonic de novo mutation rate in individuals with schizophrenia. Nature genetics 43:860–863.

Girelli D, Russo C, Ferraresi P, et al. (2000) Polymorphisms in the Factor VII gene and the risk of myocardial infarction in patients with coronary artery disease. The New England journal of medicine 343:774-780.

Glass CK, Witztum JL (2001) Atherosclerosis: The Road Ahead. Cell 104:503–516.

Go AS, Mozaffarian D, Roger VL, et al. (2013) Heart disease and stroke statistics--2013 update: a report from the American Heart Association. Circulation 127:e6–e245.

Goldstein JL, Brown MS (1979) The LDL receptor locus and the genetics of familial hypercholesterolemia. Annual review of genetics 13:259–89.

Gong C, Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3’ UTRs via Alu elements. Nature 470:284–288.

Gorlov IP, Gorlova OY, Sunyaev SR, et al. (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. American journal of human genetics 82:100–112.

Grant SFA, Thorleifsson G, Reynisdottir I, et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nature genetics 38:320–323.

Graves JA, Wakefield MJ, Toder R (1998) The origin and evolution of the pseudoautosomal regions of human sex chromosomes. Human molecular genetics 7:1991–1996.

Gravholt CH, Juul S, Naeraa RW, Hansen J (1998) Morbidity in Turner syndrome. Journal of clinical epidemiology 51:147–158.

Greenland P, Knoll MD, Stamler J, et al. (2003) Major risk factors as antecedents of fatal and nonfatal coronary heart disease events. JAMA : the journal of the American Medical Association 290:891–897.

Grimes DS (2012) An epidemic of coronary heart disease. QJM 105:509-518.

Gross A, Tönjes A, Kovacs P, et al. (2011) Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. BMC genetics 12:67.

Grundy SM, Cleeman JI, Merz CNB, et al. (2004) Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 110:227–239.

Grundy SM, Pasternak R, Greenland P, et al. (1999) Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology. Circulation 100:1481–1492.

Gschwend M, Levran O, Kruglyak L, et al. (1996) A locus for Fanconi anemia on 16q determined by homozygosity mapping. American journal of human genetics 59:377–384.

Gupta R, Shah N, Wang KC, et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464:1071–1076.

Gurrieri F, Accadia M (2009) Genetic imprinting: the paradigm of Prader-Willi and Angelman syndromes. Endocrine development 14:20–28.

Guttmacher AE, Collins FS, Carmona RH (2004) The family history - More important than ever. The New England journal of medicine 351:2333–2336.

Guttman M, Donaghey J, Carey BW, et al. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300.

Hackam DG, Anand SS (2003) Emerging Risk Factors for Atherosclerotic Vascular Disease. Clinical cardiology 290:932–940.

Hagiwara K, Morino H, Shiihara J, et al. (2011) Homozygosity mapping on homozygosity haplotype analysis to detect recessive disease-causing genes from a small number of unrelated, outbred patients. PloS one 6:e25059.

Hajjar I, Kotchen TA (2003) Trends in prevalence, awareness, treatment, and control of hypertension in the United States, 1988-2000. JAMA: the journal of the American Medical … 290:199–206.

Hamm P, Shekelle RB, Stamler J (1989) Large fluctuations in body weight during young adulthood and twenty-five-year risk of coronary death in men. American journal of epidemiology 129:312–318.

Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Human heredity 70:42–54.

Hardy J, Singleton A (2009) Genomewide association studies and human disease. The New England journal of medicine 360:1759–1768.

Hauser ER, Crossman DC, Granger CB, et al. (2004) A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. American journal of human genetics 75:436–47.

Havlik RJ, Garrison RJ, Feinleib M, et al. (1979) Blood pressure aggregation in families. American journal of epidemiology 110:304–312.

Heidenreich P a, Trogdon JG, Khavjou O a, et al. (2011) Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation 123:933–944.

Heinig M, Petretto E, Wallace C, et al. (2010) A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature 467:460–464.

Helena Mangs a, Morris BJ (2007) The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Current genomics 8:129–136.

Helgadottir A, Thorleifsson G, Magnusson KP, et al. (2008) The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nature genetics 40:217–224.

Helgadottir A, Thorleifsson G, Manolescu A, et al. (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science (New York, NY) 316:1491–1493.

Helgason A, Yngvadóttir B, Hrafnkelsson B, et al. (2005) An Icelandic example of the impact of population structure on association studies. Nature genetics 37:90–95.

Higgins JP, Thompson SG, Deeks JJ, et al (2003) Measuring inconsistency in meta-analyses. British Medical Journal 327:557-60.

Hildebrandt F, Heeringa SF, Rüschendorf F, et al. (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS genetics 5:e1000353.

Hindorff LA, Sethupathy P, Junkins HA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings pf the national academy of sciences 106:9362-7.

Hinds DA, Stuve LL, Nilsen GB, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

Hingorani AD, Psaty BM (2009) Primary Prevention of Cardiovascular Disease - Time to Get More or Less Personal? JAMA : the journal of the American Medical Association 302:2144–2145.

Hirschhorn JN (2009) Genomewide association studies--illuminating biologic pathways. The New England journal of medicine 360:1699–1701.

Holle R, Happich M, Löwel H, Wichmann HE (2005) KORA--a research platform for population based health research. Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)) 67 Suppl 1:S19–25.

Holm H, Gudbjartsson DF, Sulem P, et al. (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nature genetics 43:316–320.

Holmes M V, Harrison S, Talmud PJ, et al. (2011) Utility of genetic determinants of lipids and cardiovascular events in assessing risk. Nature reviews Cardiology 8:207–221.

Holroyd KJ, Martinati LC, Trabetti E, et al. (1998) Asthma and bronchial hyperresponsiveness linked to the XY long arm pseudoautosomal region. Genomics 52:233–235.

Hosking FJ, Papaemmanuil E, Sheridan E, et al. (2010) Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk. Blood 115:4472–4477.

Hosking FJ, Dobbins SE, Houlston RS (2011) Genome-wide association studies for detecting cancer susceptibility. British Medical Bulletin 97:27-46.

Houwen RHJ, Baharloo S, Blankenship K, et al. (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nature genetics 8:380–386.

Howard G, Wagenknecht LE, Burke GL, et al. (1998) Cigarette smoking and progression of atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) Study. JAMA : the journal of the American Medical Association 279:119–124.

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5:e1000529.

Howrigan DP, Simonson M, Keller MC (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC genomics 12:460.

Hu W, Yuan B, Flygare J, Lodish HF (2011) Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. Genes & development 25:2573–2578.

Huarte M, Guttman M, Feldser D, et al. (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419.

Hugot JP, Chamaillard M, Zouali H, et al. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603.

Humphreys K, Grankvist A, Leu M, et al. (2011) The genetic structure of the Swedish population. PloS one 6:e22547.

Hunter-Zinck H, Musharoff S, Salit J, et al. (2010) Population Genetic Structure of the People of Qatar. The American Journal of Human Genetics 87:17–25.

Iacoviello L, Di Castelnuovo A, De Knijff P, et al. (1998) Polymorphisms in the coagulation factor VII gene and risk of myocardial infarction. The New England journal of medicine 338:79-85.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.

Ionita-Laza I, Buxbaum JD, Laird NM, et al. (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genetics 7:e1001289.

Iyengar SK, Elston RC (2007) The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods molecular biology 376:71-84.

Jakkula E, Rehnström K, Varilo T, et al. (2008) The genome-wide patterns of variation expose significant substructure in a founder population. American journal of human genetics 83:787–794.

Jakobsdottir J, Gorin MB, Conley YP, et al. (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS genetics 5:e1000337.

Jansen ACM, van Aalst-Cohen ES, Tanck MW, et al. (2004) The contribution of classical risk factors to cardiovascular disease in familial hypercholesterolaemia: data in 2400 patients. Journal of internal medicine 256:482–490.

Jiang H, Orr A, Guernsey DL, et al (2009) Application of **homozygosity** haplotype analysis to genetic mapping with high-density SNP genotype data. PLoS One 4:e5280.

Ji W, Foo JN, O'Roak BJ, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nature genetics 40:592–599.

Johansen CT, Wang J, Lanktree MB, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nature genetics 42:684–687.

Johnson JL, Newby AC (2009) Macrophage heterogeneity in atherosclerotic plaques. Current opinion in lipidology 20:370–378.

Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Human molecular genetics 20:R182–8.

Kamata K, Kawamoto H, Honma T, et al. (1998) Structural basis for chemical inhibition of human blood coagulation factor Xa. Proceedings of the National Academy of Sciences of the United States of America 95:6630–6635.

Kamei M, Carman C V (2010) New observations on the trafficking and diapedesis of monocytes. Current opinion in hematology 17:43–52.

Kannel WB, Wolf PA, Castelli WP, D'Agostino RB (1987) Fibrinogen and risk of cardiovascular disease. The Framingham Study. JAMA : the journal of the American Medical Association 258:1183–1186.

Kaprio J, Tuomilehto J, Koskenvuo M, et al. (1992) Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35:1060-7.

Kathiresan S, Manning AK, Demissie S, et al. (2007) A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC medical genetics 8 Suppl 1:S17.

Kathiresan S, Melander O, Anevski D, et al. (2008a) Polymorphisms associated with cholesterol and risk of cardiovascular events. The New England journal of medicine 358:1240–9.

Kathiresan S, Melander O, Guiducci C, et al. (2008b) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature genetics 40:189–197.

Kathiresan S, Voight BF, Purcell S, et al. (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature genetics 41:334–341.

Kauppi P, Laitinen T, Ollikainen V, et al. (2000) The IL9R region contribution in asthma is supported by genetic association in an isolated population. European journal of human genetics : EJHG 8:788–792.

Kawachi I, Colditz GA, Speizer FE, et al. (1997) A prospective study of passive smoking and coronary heart disease. Circulation 95:2374-9.

Keating BJ, Tischfield S, Murray SS, et al. (2008) Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PloS one 3:e3583.

Keller MC, Simonson M a, Ripke S, et al. (2012) Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS genetics 8:e1002656.

Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189:237–249.

Kelley D, Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome biology 13:R107.

Kessler T, Erdmann J, Schunkert H (2013) Genetics of coronary artery disease and myocardial infarction--2013. Current cardiology reports 15:368.

Khlat M, Khoury M (1991) Inbreeding and diseases: demographic, genetic, and epidemiologic perspectives. Epidemiologic reviews 13:28–41.

Khot UN, Khot MB, Bajzer CT, et al. (2003) Prevalence of conventional risk factors in patients with coronary heart disease. JAMA : the journal of the American Medical Association 290:898–904.

Khoury MJ, Cohen BH, Chase GA, et al. (1987) An epidemiologic approach to the evaluation of the effect of inbreeding on prereproductive mortality. American journal of epidemiology 125:251–262.

Kirin M, McQuillan R, Franklin CS, et al. (2010) Genomic runs of homozygosity record population history and consanguinity. PloS one 5:e13996.

Knight HM, Maclean A, Irfan M, et al. (2008) Homozygosity mapping in a family presenting with schizophrenia, epilepsy and hearing impairment. European journal of human genetics : EJHG 16:750–758.

Kolz M, Baumert J, Gohlke H, et al. (2009) Association study between variants in the fibrinogen gene cluster, fibrinogen levels and hypertension: results from the MONICA/KORA study. Thrombosis and haemostasis 101:317–324.

Kones R (2011) Primary prevention of coronary heart disease: integration of new data, evolving views, revised goals, and role of rosuvastatin in management. A comprehensive survey. Drug design, development and therapy 5:325–380.

Koopmans JR, Slutske WS, Heath AC, et al. (1999) The genetics of **smoking** initiation and quantity smoked in Dutch adolescent and young adult twins. Behavior Genetics 29:383-93.

Kraft HG, Lingenhel A, Kochl S, et al. (1996) Apolipoprotein(a) kringle IV repeat number predicts risk for coronary artery disease. Arteriosclerosis, thrombosis, and vascular biology 16:713-9.

Kraft P, Hunter DJ (2009) Genetic risk prediction: are we there yet? The New England journal of medicine 360:1701–1703.

Krawczak M, Nikolaus S, von Eberstein H, et al. (2006) PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community genetics 9:55–61.

Kretz M, Webster DE, Flockhart RJ, et al. (2012) Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes & development 26:338–343.

Kryukov G V, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. American journal of human genetics 80:727–739.

Ku CS, Naidoo N, Teo SM, et al. (2011) Regions of homozygosity and their impact on complex diseases and traits. Human genetics 129:1–15.

Kullo IJ, Cooper LT (2010) Early identification of cardiovascular risk using genomics and proteomics. Nature reviews Cardiology 7:309–317.

Kullo IJ, Ding K (2007) Mechanisms of disease: The genetic basis of coronary heart disease. Nature clinical practice Cardiovascular medicine 4:558–569. doi: 10.1038/ncpcardio0982

Kvaløy K, Galvagni F, Brown WR (1994) The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. Human molecular genetics 3:771–778.

Łabaj PP, Leparc GG, Linggi BE, et al (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics 27:383-91.

Lage K, Karlberg EO, Storling ZM, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology 25:309-16.

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nature reviews Genetics 7:385–94.

Lajunen H-R, Kaprio J, Keski-Rahkonen A, et al. (2009) Genetic and environmental effects on body mass index during adolescence: a prospective study among Finnish twins. International Journal of Obesity (London) 33:559–567.

Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570.

Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lango Allen H, Estrada K, Lettre G, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838.

Lanktree MB, Guo Y, Murtaza M, et al. (2011) Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. American journal of Human Genetics 88:6-18.

Laurier V, Stoetzel C, Muller J, et al. (2006) Pitfalls of homozygosity mapping: an extended consanguineous Bardet-Biedl syndrome family with two mutant genes (BBS2, BBS10), three mutations, but no triallelism. European journal of human genetics : EJHG 14:1195–1203.

Lebel RR, Gallagher WB (1989) Wisconsin consanguinity studies. II: Familial adenocarcinomatosis. American journal of medical genetics 33:1–6.

Lehrke M, Millington SC, Lefterova M, et al. (2007) CXCL16 is a marker of inflammation, atherosclerosis, and acute coronary syndromes in humans. Journal of the American College of Cardiology 49:442–449.

Leigh SEA, Foster AH, Whittall RA, et al. (2008) Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database. Annals of human genetics 72:485–498.

Lencz T, Lambert C, DeRosse P, et al. (2007a) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 104:19942–19947.

Lencz T, Morgan T V, Athanasiou M, et al. (2007b) Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Molecular psychiatry 12:572–580.

Lettre G, Rioux JD (2008) Autoimmune diseases: insights from genome-wide association studies. Human molecular genetics 17:R116–21.

Leutenegger A-L, Labalme A, Genin E, et al. (2006) Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. American journal of human genetics 79:62–66.

Levy D, DeStefano a. L, Larson MG, et al. (2000) Evidence for a Gene Influencing Blood Pressure on Chromosome 17 : Genome Scan Linkage Results for Longitudinal Blood Pressure Phenotypes in Subjects From the Framingham Heart Study. Hypertension 36:477–483.

Levy D, Ehret GB, Rice K, et al. (2009) Genome-wide association study of blood pressure and hypertension. Nature genetics 41:677–87.

Levy D, Larson MG, Vasan RS, et al. (1996) The progression from hypertension to congestive heart failure. JAMA : the journal of the American Medical Association 275:1557–62.

Lewin MB, McBride KL, Pignatelli R, et al. (2004) Echocardiographic evaluation of asymptomatic parental and sibling cardiovascular anomalies associated with congenital left ventricular outflow tract lesions. Pediatrics 114:691–696.

Lewington S, Clarke R, Qizilbash N, et al. (2002) Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet 360:1903–1913.

Lezirovitz K, Pardono E, de Mello Auricchio MTB, et al. (2008) Unexpected genetic heterogeneity in a large consanguineous Brazilian pedigree presenting deafness. European journal of human genetics : EJHG 16:89–96.

Li B, Leal SM (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS genetics 5:e1000481.

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics 83:311–321.

Li L, Hamer DH (1995) Recombination and allelic association in the Xq/Yq homology region. Human molecular genetics 4:2013–2016.

Li L, Ho S, Chen C, et al. (2006) Long contiguous stretches of homozygosity in the human genome. Human mutation 27:1115–1121.

Li Y, Luke MM, Shiffman D, Devlin JJ (2011) Genetic variants in the apolipoprotein(a) gene and coronary heart disease. Circulation Cardiovascular genetics 4:565–573.

Li Y, Willer CJ, Ding J, et al. (2010) MaCH: Using sequence and genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genetic epidemiology 34:816–834.

Liehr T (2010) Cytogenetic contribution to uniparental disomy (UPD). Molecular cytogenetics 3:8.

Lien S, Szyda J, Schechinger B, et al. (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. American journal of human genetics 66:557–566.

Linsel-Nitschke P, Götz A, Erdmann J, et al. (2008) Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study. PloS one 3:e2986.

Liu DJ, Leal SM (2012) Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. American journal of human genetics 91:585–596.

Lloyd-Jones D, Adams R, Carnethon M, et al. (2009) Heart disease and stroke statistics--2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 119:480–486.

Lloyd-Jones DM, Larson MG, Beiser A, Levy D (1999) Lifetime risk of developing coronary heart disease. Lancet 353:89–92.

Loewer S, Cabili MN, Guttman M, et al. (2010) Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nature genetics 42:1113–1117.

Loffredo CA, Chokkalingam A, Sill AM, et al. (2004) Prevalence of congenital cardiovascular malformations among relatives of infants with hypoplastic left heart, coarctation of the aorta, and d-transposition of the great arteries. American journal of medical genetics Part A 124A:225–230.

Löwel H, Meisinger C, Heier M, Hörmann A (2005) The population-based acute myocardial infarction (AMI) registry of the MONICA/KORA study region of Augsburg. Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)) 67 Suppl 1:S31–7.

Löwer R, Löwer J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proceedings of the National Academy of Sciences of the United States of America 93:5177–5184.

Luke MM, Kane JP, Liu DM, et al. (2007) A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease. Arteriosclerosis, thrombosis, and vascular biology 27:2030–2036.

Mackman N, Tilley RE, Key NS (2007) Role of the extrinsic pathway of blood coagulation in hemostasis and thrombosis. Arteriosclerosis, thrombosis, and vascular biology 27:1687–1693.

Macleod IM, Larkin DM, Lewin H a, et al. (2013) Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. Molecular biology and evolution 30:2209–2223.

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics 5:e1000384.

Mägi R, Asimit JL, Day-Williams AG, et al. (2012) Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases. Genetic epidemiology 796:785–796.

Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456:18–21.

Mangoni AA, Jackson SHD (2002) Homocysteine and cardiovascular disease: current evidence and future prospects. The American journal of medicine 112:556–565.

Manolio T a, Collins FS, Cox NJ, et al. (2009) Finding the missing heritability of complex diseases. Nature 461:747–753.

Manolio TA (2009) Cohort studies and the genetics of complex disease. Nature genetics 41:5–6.

Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. The Journal of clinical investigation 118:1590–605.

Marchini J, Howie B, Myers S, et al. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics 39:906–913.

Marenberg ME, Risch N, Berkman LF, et al. (1994) Genetic susceptibility to death from coronary heart disease in a study of twins. The New England journal of medicine 330:1041–1046.

Marks D, Thorogood M, Neil HAW, Humphries SE (2003) A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. Atherosclerosis 168:1–14.

Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS genetics 5:e1000459.

Mayer B, Erdmann J, Schunkert H (2007) Genetics and heritability of coronary artery disease and myocardial infarction. Clinical research in cardiology : official journal of the German Cardiac Society 96:1–7.

McBride KL, Pignatelli R, Lewin M, et al. (2005) Inheritance analysis of congenital left ventricular outflow tract obstruction malformations: Segregation, multiplex relative risk, and heritability. American journal of medical genetics Part A 134A:180–186.

McBride KL, Riley MF, Zender GA, et al. (2008) NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling. Human molecular genetics 17:2886–2893.

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Human molecular genetics 17:R156–65.

McLean JW, Tomlinson JE, Kuang WJ, et al. (1987) cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. Nature 330:132–137.

McPherson JD, Marra M, Hillier L, et al. (2001) A physical map of the human genome. Nature 409:934-41.

McPherson R, Pertsemlidis A, Kavaslar N, et al. (2007) A common allele on chromosome 9 associated with coronary heart disease. Science 316:1488–1491.

McQuillan R, Eklund N, Pirastu N, et al. (2012) Evidence of inbreeding depression on human height. PLoS genetics 8:e1002655.

McQuillan R, Leutenegger A, Abdel-rahman R, et al. (2008) Runs of homozygosity in European populations. The American Journal of Human Genetics 83:359–372.

Meade TW, Mellows S, Brozovic M, et al. (1986) Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study. Lancet 2:533-537.

Menotti A, Keys A, Kromhout D, et al. (1993) Inter-cohort differences in coronary heart disease mortality in the 25-year follow-up of the seven countries study. European journal of epidemiology 9:527–536.

Mestas J, Ley K (2008) Monocyte-endothelial cell interactions in the development of atherosclerosis. Trends in cardiovascular medicine 18:228–232.

Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews Genetics 11:31–46.

Miano MG, Jacobson SG, Carothers A, et al. (2000) Pitfalls in Homozygosity Mapping. The American Journal of Human Genetics 67:1348–1351.

Miyazawa H, Kato M, Awata T, et al. (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. American Journal of Human Genetics 80:1090-102.

Mo X, Hao Y, Yang X, et al. (2011) Association between polymorphisms in the coagulation factor VII gene and coronary heart disease risk in different ethnicities: a meta-analysis. BMC medical genetics 12:107.

Modell B, Darr A (2002) Science and society: genetic counselling and customary consanguineous marriage. Nature reviews Genetics 3:225–229.

Mohandas TK, Speed RM, Passage MB, et al. (1992) Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. American journal of human genetics 51:526–533.

Moore KJ, Tabas I (2011) Macrophages in the pathogenesis of atherosclerosis. Cell 145:341–355.

Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutation research 615:28–56.

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology 34:188–193.

Morrison AC, Voorman A, Johnson AD, et al. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nature genetics 45:899–901.

Mortazavi A, Williams BA, McCue K, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5:621–8.

Mozaffarian D, Katan MB, Ascherio A, et al. (2006) Trans fatty acids and cardiovascular disease. The New England journal of medicine 354:1601–1613.

Mueller RF, Bishop DT (1993) Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. Journal of Medical Genetics 30:798 –799.

Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A (2010) Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genetic epidemiology 34:213–221.

Musunuru K, Kathiresan S (2010) Genetics of coronary artery disease. Annual review of genomics and human genetics 11:91–108.

Musunuru K, Strong A, Frank-Kamenetsky M, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466:714-9.

Myers RH, Kiely DK, Cupples LA, et al. (1990) Parental history is an independent risk factor for coronary artery disease: the Framingham Study. American Heart Journal 120:963-9.

Nalls M a, Guerreiro RJ, Simon-Sanchez J, et al. (2009a) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics 10:183–190.

Nalls M a, Simon-Sanchez J, Gibbs JR, et al. (2009b) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS genetics 5:e1000415.

Naoumova RP, Neuwirth C, Lee P, et al. (2004) Autosomal recessive hypercholesterolaemia: long-term follow up and response to treatment. Atherosclerosis 174:165–72.

Neale BM, Kou Y, Liu L, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485:242–245.

Neal B, MacMahon S, Chapman N, et al (2000) Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials. Blood Pressure Lowering Treatment Trialists' Collaboration. Lancet 356:1955-64.

Nejentsev S, Walker N, Riches D, et al. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387–389.

Nelson MR, Wegmann D, Ehm MG, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100-4.

Newton-Cheh C, Johnson T, Gateva V, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nature genetics 41:666–76.

Ng S-Y, Johnson R, Stanton LW (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. The EMBO journal 31:522–33.

Nora JJ, Lortscher RH, Spangler RD, et al. (1980) Genetic--epidemiologic study of early-onset ischemic heart disease. Circulation 61:503–508.

Norio R (2003) Finnish Disease Heritage I: characteristics, causes, background. Human genetics 112:441–456.

Nothnagel M, Lu TT, Kayser M, et al. (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Human molecular genetics 19:2927–2935.

Noto D, Barbagallo CM, Cefalu' AB, et al. (2002) Factor VII activity is an independent predictor of cardiovascular mortality in elderly women of a Sicilian population: results of an 11-year follow-up. Thrombosis and Haemostasis journal 87:206-210.

O'Dushlaine CT, Morris D, Moskvina V, et al. (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. European journal of human genetics : EJHG 18:1248–1254.

Okrainec K, Banerjee DK, Eisenberg MJ (2004) Coronary artery disease in the developing world. American heart journal 148:7-15.

O'Roak BJ, Deriziotis P, Lee C, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nature genetics 43:585–589.

O'Roak BJ, Vives L, Girirajan S, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485:246–50.

Ober C, Nord AS, Thompson EE, et al. (2009) Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. Journal of lipid research 50:798–806.

Ogura Y, Bonen DK, Inohara N, et al. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606.

Page DC, Brown LG, de la Chapelle A (1987) Exchange of terminal portions of X- and Y-chromosomal short arms in human XX males. Nature 328:437–440.

Pajukanta P, Cargill M, Viitanen L, et al. (2000) Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland. American journal of human genetics 67:1481–93.

Pannain S, Weiss RE, Jackson CE, et al. (1999) Two different mutations in the thyroid peroxidase gene of a large inbred Amish kindred: power and limits of homozygosity mapping. The Journal of clinical endocrinology and metabolism 84:1061–1071.

Panoutsopoulou K, Tachmazidou I, Zeggini E (2013) In search of low frequency and rare variants affecting complex traits. Human molecular genetics 1–18.

Paulson KE, Zhu S-N, Chen M, et al. (2010) Resident intimal dendritic cells accumulate lipid and contribute to the initiation of atherosclerosis. Circulation research 106:383–390.

Pawitan Y, Seng KC, Magnusson PKE (2009) How many genetic variants remain to be discovered? PloS one 4:e7969.

Pearson TA, Blair SN, Daniels SR, et al. (2002) AHA Guidelines for Primary Prevention of Cardiovascular Disease and Stroke: 2002 Update: Consensus Panel Guide to Comprehensive Risk Reduction for Adult Patients Without Coronary or Other Atherosclerotic Vascular Diseases. Circulation 106:388–391.

Peden JF, Farrall M (2011) Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. Human molecular genetics 20:R198–205.

Pedotti P, 't Hoen P a C, Vreugdenhil E, et al. (2008) Can subtle changes in gene expression be consistently detected with different microarray platforms? BMC genomics 9:124.

Pemberton TJ, Absher D, Feldman MW, et al. (2012a) Genomic patterns of homozygosity in worldwide human populations. American journal of human genetics 91:275–292.

Peters A, von Klot S, Heier M, et al. (2004) Exposure to traffic and the onset of myocardial infarction. The New England journal of medicine 351:1721–1730.

Polašek O, Hayward C, Bellenguez C, et al. (2010) Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11:139.

Power C, Elliott J (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). International journal of epidemiology 35:34–41

Prensner JR, Iyer MK, Balbin OA, et al. (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature biotechnology 29:742–749.

Prescott E, Hippe M, Schnohr P, et al. (1998) Smoking and risk of myocardial infarction in women and men: longitudinal population study. BMJ (Clinical research ed) 316:1043–1047.

Preuss M, König IR, Thompson JR, et al. (2010) Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. Circulation Cardiovascular genetics 3:475–483.

Price AL, Kryukov G V, de Bakker PIW, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. American journal of human genetics 86:832–8.

Prins BP, Lagou V, Asselbergs FW, et al. (2012) Genetics of coronary artery disease: genome-wide association studies and beyond. Atherosclerosis 225:1–10.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics 69:124–137.

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Human molecular genetics 11:2417–2423.

Pruim RJ, Welch RP, Sanna S, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics (Oxford, England) 26:2336–2337.

Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81:559–575.

Puzyrev VP, Lemza S V, Nazarenko LP, Panphilov VI (1992) Influence of genetic and demographic factors on etiology and pathogenesis of chronic disease in north Siberian aborigines. Arctic medical research 51:136–142.

Rabbani LE, Loscalzo J (1994) Recent observations on the role of hemostatic determinants in the development of the atherothrombotic plaque. Atherosclerosis 105:1–7.

Rappold GA (1993) The pseudoautosomal regions of the human sex chromosomes. Human Genetics 92:315-24.

Rappold G, Fukami M, Niesler B, et al. (2002) Deletions of the homeobox gene SHOX (short stature homeobox) are an important cause of growth failure in children with short stature. The Journal of clinical endocrinology and metabolism 87:1402–1406.

Reddy KS, Yusuf S (1998) Emerging epidemic of cardiovascular disease in developing countries. Circulation 97:596-601

Reich DE, Cargill M, Bolk S, et al. (2001) Linkage disequilibrium in the human genome. Nature 411:199–204.

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends in genetics : TIG 17:502–510.

Reilly MP, Li M, He J, et al. (2011) Identification of *ADAMTS7* as a novel locus for coronary atherosclerosis and association of *ABO* with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. Lancet 377(9763):383-392.

Renauld JC, Druez C, Kermouni a, et al. (1992) Expression cloning of the murine and human interleukin 9 receptor cDNAs. Proceedings of the National Academy of Sciences of the United States of America 89:5690–5694.

Rhoads GG, Dahlen G, Berg K, et al. (1986) Lp(a) lipoprotein as a risk factor for myocardial infarction. JAMA : the journal of the American Medical Association 256:2540–2544.

Ried K, Rao E, Schiebel K, Rappold G a (1998) Gene duplications as a recurrent theme in the evolution of the human pseudoautosomal region 1: isolation of the gene ASMTL. Human molecular genetics 7:1771–1778.

Rissanen AM (1979) Familial aggregation of coronary heart disease in a high incidence area (North Karelia, Finland). British heart journal 42:294–303.

Rissanen AM, Nikkilä EA (1977) Coronary artery disease and its risk factors in families of young men with angina pectoris and in controls. British heart journal 39:875–883.

Rivas M a, Beaudoin M, Gardet A, et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics 43:1066–1073.

Roberts R (2008) Genetics of premature myocardial infarction. Current atherosclerosis reports 10:186–193.

Romeo S, Pennacchio LA, Fu Y, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nature genetics 39:513–516.

Ross MT, Graftham D V, Coffey AJ, et al. (2005) The DNA sequence of the human X chromosome. Nature 434:325–337.

Ross R. (1993) The pathogenesis of atherosclerosis: a perspective for the 1990s. Nature 362: 801–09.

Rothe M, Wong SC, Henzel WJ, et al. (1994) A novel family of putative signal transducers associated with the cytoplasmic domain of the 75 kDa tumor necrosis factor receptor. Cell 78:681-92.

Rotival M, Zeller T, Wild PS, et al. (2011) Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS genetics 7:e1002367.

Rouyer F, Simmler MC, Johnsson C, et al. (1986) A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. Nature 319:291–295.

Roy H, Bhardwaj S, Yla-Herttuala S (2009) Molecular genetics of atherosclerosis. Human genetics 125:467–491.

Roy-Gagnon M-H, Moreau C, Bherer C, et al. (2011) Genomic and genealogical investigation of the French Canadian founder population structure. Human genetics 129:521–31.

Rudan I, Biloglav Z, Vorko-Jović A, et al. (2006) Effects of inbreeding, endogamy, genetic admixture, and outbreeding on human health: a (1001 Dalmatians) study. Croatian medical journal 47:601–610.

Rudan I, Rudan D, Campbell H, et al. (2003a) Inbreeding and risk of late onset complex disease. Journal of Medical genetics 40:925–932.

Rudan I, Skarić-Jurić T, Smolej-Narancić N, et al. (2004) Inbreeding and susceptibility to osteoporosis in Croatian island isolates. Collegium antropologicum 28:585–601.

Rudan I, Smolej-Narancic N, Campbell H, et al. (2003b) Inbreeding and the genetic complexity of human hypertension. Genetics 163:1011–1021.

Saar K, Schindler D, Wegner RD, et al. (1998) Localisation of a Fanconi anaemia gene to chromosome 9p. European journal of human genetics 6:501-8.

Sachdev V, Matura LA, Sidenko S, et al. (2008) Aortic valve disease in Turner syndrome. Journal of the American College of Cardiology 51:1904–1909.

Saito T, Parsia S, Papolos DF, Lachman HM (2000) Analysis of the pseudoautosomal X-linked gene SYBL1in bipolar affective disorder: description of a new candidate allele for psychiatric disorders. American journal of medical genetics 96:317–323.

Saleheen D, Zaidi M, Rasheed A, et al. (2009) The Pakistan Risk of Myocardial Infarction Study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia. European journal of Epidemiology 24:329-338.

Samani NJ, Burton P, Mangino M, et al. (2005) A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. American journal of human genetics 77:1011–1020.

Samani NJ, Erdmann J, Hall AS, et al. (2007) Genomewide association analysis of coronary artery disease. The New England journal of medicine 357:443–453.

Sanders SJ, Murtha MT, Gupta AR, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237–41.

Saxena R, Voight BF, Lyssenko V (2007) Genome-wide association analysis identifies loci for tyoe 2 diabetes and triglyceride levels. Science 316:1331-6.

Scheuner MT (2003) Genetic evaluation for coronary artery disease. Genetics in medicine : official journal of the American College of Medical Genetics 5:269–285.

Schiller S, Spranger S, Schechinger B, et al. (2000) Phenotypic variation and genetic heterogeneity in Léri-Weill syndrome. European journal of human genetics : EJHG 8:54–62.

Schonrock N, Harvey RP, Mattick JS (2012) Long noncoding RNAs in cardiac development and pathophysiology. Circulation research 111:1349–1362.

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Current opinion in genetics & development 19:212–219.

Schunkert H, Erdmann J, Samani NJ (2010) Genetics of myocardial infarction: a progress report. European heart journal 31:918–925.

Schunkert H, Götz A, Braund P, et al. (2008) Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation 117:1675–1684.

Schunkert H, König IR, Kathiresan S, et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics 43:333–338.

Schuurs-Hoeijmakers JHM, Hehir-Kwa JY, Pfundt R, et al. (2011) Homozygosity mapping in outbred families with mental retardation. European journal of human genetics EJHG 19:597–601.

Sebat J, Lakshmi B, Troge J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305;525-8.

Seo D, Goldschmidt-Clermont PJ (2008) Cardiovascular genetic medicine: the genetics of coronary heart disease. Journal of cardiovascular translational research 1:166–70.

Shami SA, Qaisar R, Bittles AH (1991) Consanguinity and adult morbidity in Pakistan. Lancet 338:954.

Sharma P, Kumar J, Garg G, et al. (2008) Detection pf altered global DNA methylation in coronary artery disease patients. DNA cell biology 27:357-65.

Sharrett AR, Ballantyne CM, Coady SA, et al. (2001) Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density subfractions: The Atherosclerosis Risk in Communities (ARIC) Study. Circulation 104:1108-13.

Shears DJ, Vassal HJ, Goodman FR, et al. (1998) Mutation and deletion of the pseudoautosomal gene SHOX cause Leri-Weill dyschondrosteosis. Nature genetics 19:70–73.

Shepherd J, Cobbe SM, Ford I, et al. (1995) Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group. The New England journal of medicine 333:1301-7.

Shi X, Sun M, Liu H, et al. (2013) Long non-coding RNAs: A new frontier in the study of human diseases. Cancer letters.

Shiffman D, Kane JP, Louie JZ, et al. (2008) Analysis of 17,576 potentially functional SNPs in three case-control studies of myocardial infarction. PloS one 3:e2895.

Shiffman D, Louie JZ, Rowland CM, et al. (2010) Single variants can explain the association between coronary heart disease and haplotypes in the apolipoprotein(a) locus. Atherosclerosis 212:193–196.

Simón-Sánchez J, Kilarski LL, Nalls M a, et al. (2012) Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinson's disease. PloS one 7:e28787.

Simon-Sanchez J, Scholz S, Fung H-C, et al. (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Human molecular genetics 16:1–14.

Simmler MC, Rouyer F, Vergnaud G, et al. (1985) Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. Nature 317:692-7.

Sims R, Dwyer S, Harold D, et al. (2011) No evidence that extended tracts of homozygosity are associated with Alzheimer's disease. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics 156B:764–771.

Siva a, De Lange M, Clayton D, et al. (2007) The heritability of plasma homocysteine, and the influence of genetic variation in the homocysteine methylation pathway. QJM : monthly journal of the Association of Physicians 100:495–499.

Siva N (2008) 1000 Genomes project. Nature biotechnology 26:256.

Slack J, Evans KA (1966) The increased risk of death from ischaemic heart disease in first degree relatives of 121 men and 96 women with ischaemic heart disease. Journal of medical genetics 3:239–257.

Smith CAB (1953) Detection of linkage in human genetics. Journal of the Royal Statistical Society: Series B 15:153–192.

Smith EB, Keen G a., Grant a., Stirk C (1990) Fate of fibrinogen in human arterial intima. Arteriosclerosis, Thrombosis, and Vascular Biology 10:263–275.

Smolina K, Wright FL, Rayner M, Goldacre MJ (2012) Long-term survival and recurrence after acute myocardial infarction in England, 2004 to 2010. Circulation Cardiovascular quality and outcomes 5:532–540.

Snieder H, van Doornen LJ, Boomsma DI (1997) The age dependency of gene expression for plasma lipids, lipoproteins, and apolipoproteins. American journal of human genetics 60:638–650.

Snieder H, van Doornen LJ, Boomsma DI (1999) Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. Arteriosclerosis, thrombosis, and vascular biology 19:2826–34.

Soranzo N, Spector TD, Mangino M, et al. (2009) A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nature genetics 41:1182–1190.

Sørensen TI, Nielsen GG, Andersen PK, Teasdale TW (1988) Genetic and environmental influences on premature death in adult adoptees. The New England journal of medicine 318:727–32.

Soria LF, Ludwig EH, Clarke HR, et al. (1989) Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. Proceedings of the National Academy of Sciences of the United States of America 86:587–591.

Sotoodehnia N, Isaacs A, de Bakker PI, et al. (2010) Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. Nature genetics 42:1068-76.

Soutar AK, Naoumova RP (2007) Mechanisms of disease: genetic causes of familial hypercholesterolemia. Nature clinical practice Cardiovascular medicine 4:214–225.

Spain SL, Cazier J-B, Houlston R, et al. (2009) Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. Cancer research 69:7422–7429.

Speliotes EK, Willer CJ, Berndt SI, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics 42:937–948.

Staessen J a, Wang JG, Thijs L (2001) Cardiovascular protection and blood pressure reduction: a meta-analysis. Lancet 358:1305–1315.

Stamova B, Tian Y, Jickling G, et al. (2012) The X-chromosome has a different pattern of gene expression in women compared with men with ischemic stroke. Stroke; a journal of cerebral circulation 43:326–334.

Stangl K, Cascorbi I, Laule M, et al. (2000) High CA repeat numbers in intron 13 of the endothelial nitric oxide synthase gene and increased risk of coronary artery disease. Pharmacogenetics 10:133-40.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100:9440–9445.

Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nature genetics 40:17–22.

Sund KL, Zimmerman SL, Thomas C, et al. (2013) Regions of homozygosity identified by SNP microarray analysis aid in the diagnosis of autosomal recessive disease and incidentally detect parental blood relationships. Genetics in medicine : official journal of the American College of Medical Genetics 15:70–78.

Talebizadeh Z, Simon SD, Butler MG (2006) X chromosome gene expression in human tissues: male and female comparisons. Genomics 88:675–681.

Tamaki S, Iwai N, Nakamura Y, et al. (1999) Variation of the factor VII gene and ischemic heart disease in Japanese subjects. Coronary Artery Disease 10:601-606.

Tarugi P, Averna M, Di Leo E, et al. (2007) Molecular diagnosis of hypobetalipoproteinemia: an ENID review. Atherosclerosis 195:e19–27.

Taylor AH, Ussher MH, Faulkner G (2007) The acute effects of exercise on cigarette cravings, withdrawal symptoms, affect and smoking behaviour: a systematic review. Addiction 102:534–543.

Tenesa A, Haley CS (2013) The heritability of human disease: estimation, uses and abuses. Nature reviews Genetics 14:139–149.

Teo S-M, Ku C-S, Naidoo N, et al. (2011) A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. Journal of human genetics 56:524–533.

Teo S-M, Ku C-S, Salim A, et al. (2012) Regions of homozygosity in three Southeast Asian populations. Journal of human genetics 57:101–8.

Teslovich TM, Musunuru K, Smith AV, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-73.

The IBC 50K CAD Consortium (2011) Large-scale gene-centric analysis identifies novel variants for coronary artery disease. PLoS genetics 7:e1002260.

Thompson PD, Buchner D, Pina IL, et al. (2003) Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease: a statement

from the Council on Clinical Cardiology (Subcommittee on Exercise, Rehabilitation, and Prevention) and the Council on Nutrition, Physical. Circulation 107:3109–3116.

Tian Y, Stamova B, Jickling GC, et al. (2012) Y chromosome gene expression in the blood of male patients with ischemic stroke compared with male controls. Gender medicine 9:68–75.e3.

Ting JC, Roberson ED, Miller ND et al. (2007) Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. Human Mutation 28:1225-35.

Tomaszewski M, Debiec R, Braund PS, et al. (2010) Genetic architecture of ambulatory blood pressure in the general population: insights from cardiovascular gene-centric array. Hypertension 56:1069–1076.

Trégouët D-A, König IR, Erdmann J, et al. (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nature genetics 41:283–285.

Van Buggenhout G, Fryns JP (2009) Angelman syndrome (AS, MIM 105830). European Journal of Human Genetics 17:1367-73.

Van Hoek M, Dehghan A, Witteman JCM, et al. (2008) Predicting Type 2 Diabetes Based on Polymorphisms from Genome-Wide Association Studies. Diabetes 57:3122–3128.

Vasan RS, Beiser A, Seshadri S, et al. (2002) Residual Lifetime Risk for Developing Hypertension in Middle-aged Women and Men - The Framingham Heart Study. JAMA : the journal of the American Medical Association 287:1003–1010.

Vince JE, Pantaki D, Feltham R, et al. (2009) TRAF2 must bind to cellular inhibitors of apoptosis for tumor necrosis factor (tnf) to efficiently activate nf-{kappa}b and to prevent tnf-induced apoptosis. The Journal of biological chemistry 284:35906–35915.

Vine AE, McQuillin A, Bass NJ, et al. (2009) No evidence for excess runs of homozygosity in bipolar disorder. Psychiatric genetics 19:165–170.

Virmani R, Burke AP, Kolodgie FD, Farb A (2002) Vulnerable plaque: the pathology of unstable coronary lesions. Journal of interventional cardiology 15:439–446.

Visel A, Zhu Y, May D, et al. (2010) Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature. 464:409-12.

Visscher PM (2008) Sizing up human height variation. Nature genetics 40:489–490.

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nature reviews Genetics 9:255–266.

Visscher PM, Montgomery GW (2009) Genome-wide association studies and human disease: from trickle to flood. JAMA : the journal of the American Medical Association 302:2028–9.

Vissers LELM, de Ligt J, Gilissen C, et al. (2010) A de novo paradigm for mental retardation. Nature genetics 42:1109–12.

Vitart V, Carothers AD, Hayward C, et al. (2005) Increased level of linkage disequilibrium in rural compared with urban communities: a factor to consider in association-study design. American journal of human genetics 76:763–772.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS biology 4:e72.

Waisfisz Q, Saar K, Morgan N V, et al. (1999) The Fanconi anemia group E gene, FANCE, maps to chromosome 6p. American journal of human genetics 64:1400–1405.

Wald NJ, Hackshaw AK, Frost CD (1999) When can a risk factor be used as a worthwhile screening test? BMJ: British Medical Journal 319:1562–1565.

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nature reviews Genetics 4:587–597.

Wang F, Xu CQ, Cai JP et al. (2011) Genome-wide association identifies asusceptibility locus for coronary artery disease in the Chinese Han population. Nature genetics 43:345-9.

Wang KC, Yang YW, Liu B, et al. (2011a) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472:120–124.

Wang L, Fan C, Topol SE, et al. (2003) Mutation of MEF2A in an inherited disorder with features of coronary artery disease. Science 302:1578–1581.

Wang Q, Rao S, Shen G-Q, et al. (2004) Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis. American journal of human genetics 74:262–271.

Wang XL, Wang J, McCredie RM, et al. (1997) Polymorphisms of factor V, factor VII, and fibrinogen genes: relevance to severity of coronary artery disease. Arteriosclerosis Thrombosis and Vascular Biology 17:246-251.

Weber JL (2006) Clinical applications of Genome Polymorphism Scans. Biology direct 1:16.

Weedon MN, Frayling TM (2008) Reaching new heights: insights into the genetics of human stature. Trends in genetics : TIG 24:595–603.

Weedon MN, Frayling TM (2007) Insights on pathogenesis of type 2 diabetes from MODY genetics. Current diabetes reports 7:131–8.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wichmann H-E, Gieger C, Illig T (2005) KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)) 67 Suppl 1:S26–30.

Wild PS, Zeller T, Schillert A, et al. (2011) A genome-wide association study identifies LIPA as a susceptibility gene for coronary artery disease. Circulation Cardiovascular Genetics 4:403–12.

Willer CJ, Sanna S, Jackson AU, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics 40:161–169.

Williams KJ, Tabas I (1995) The response-to-retention hypothesis of early atherogenesis. Arteriosclerosis, thrombosis, and vascular biology 15:551–561.

Wienkr A, Holm NV, Skytthe A, et al. (2001) The heritability and mortality due to heart diseases: a correlated frailty model applied to Danish twins. Twin research 4:266-74.

Winkelmann BR, März W, Boehm BO, et al. (2001) Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. Pharmacogenomics 2:S1–73.

Wise AL, Gyi L, Manolio T a (2013) eXclusion: toward integrating the X chromosome in genome-wide association analyses. American journal of human genetics 92:643–647.

Woods CG, Cox J, Springell K, et al. (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. American journal of human genetics 78:889–896.

Wright A, Charlesworth B, Rudan I, et al. (2003) A polygenic basis for late-onset disease. Trends in genetics : TIG 19:97–106.

Wu MC, Lee S, Cai T, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89:82-93.

Xu B, Roos JL, Dexheimer P, et al. (2011) Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nature genetics 43:864–8.

Yamagishi K, Folsom AR, Rosamond WD, et al. (2009) A genetic variant on chromosome 9p21 and incident heart failure in the ARIC study. European heart journal 30:1222-8.

Yang H-C, Chang L-C, Liang Y-J, et al. (2012a) A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex. PloS one 7:e34840.

Yang J, Benyamin B, McEvoy BP et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42:565-9.

Yang T-L, Guo Y, Zhang L-S, et al. (2010) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. The Journal of clinical endocrinology and metabolism 95:3777–3782.

Yusuf S, Hawken S, Ounpuu S, et al. (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 364:937–952.

Zdravkovic S, Wienke A, Pedersen NL, et al. (2002) Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. Journal of internal medicine 252:247–254.

Zeggini E, Rayner W, Morris AP, et al. (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nature genetics 37:1320–1322.

Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26:2375–82.

Zhu Q, Ge D, Maia JM, et al. (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American journal of human genetics 88:458–468.

Zhu X, Feng T, Li Y, et al. (2010) Detecting rare variants for complex traits using family and unrelated data. Genetic epidemiology 34:171–187.

Zimmet P, Alberti KG, Shaw J (2001) Global and societal implications of the diabetes epidemic. Nature 414:782–787.

Zuk O, Hechter E, Sunyaev SR, et al. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the national academy of sciences 109:1193-8.