UNIVERSITAT POMPEU FABRA

Department of Experimental and Health Sciences

PhD Programme in Biomedicine 2012

DOCTORAL THESIS

GENETIC FACTORS ASSOCIATED WITH CORONARY HEART DISEASE AND ANALYSIS OF THEIR PREDICTIVE CAPACITY

Ph.D candidate: Carla Lluís Ganella
This work has been carried out under the supervision of **Roberto Elosua** and **Gavin Lucas** in the Cardiovascular Epidemiology and Genetics Research Group from the Program in Research in Inflammatory and Cardiovascular Disorders at the IMIM.

IMIM Institut de Recerca Hospital del Mar

Parc de Salut MAR Barcelona

Printed in May 2012
with the support of *Fundació IMIM*

Dedico aquest treball als meus
pares i a tu, Giovanni. Gràcies per
estar al meu costat

ACKNOWLEDGMENTS

Durant tots aquests anys que comprenen no només la part del doctorat, sinó també molt abans, he acumulat una gran quantitat de persones que han representat pilars importants per la meva formació, no només acadèmica sinó majoritàriament com a persona. Algunes d'aquestes persones podran llegir aquesta tesi, altres no, però no per això mereixen un menor agraïment, ans el contrari.

Voldria començar per agrair als meus pares tot l'esforç que han fet durant els últims 30 anys per fer de mi qui sóc, i per no deixar que mai em faltés de res, renunciant a moltes de les coses que ells haguessin pogut tenir. Voldria agrair-li en especial a ma mare totes les hores i hores que ha dedicat a estar al meu costat, que m'ha fet molt més que de mare i d'amiga, i sense la que jo no estaria aquí.

Voldria també donar les gràcies al Gavin i al Roberto, per la infinita paciència que han tingut amb mi, i per tot el que m'han ensenyat. Durant els anys que he pogut compartir al seu costat, m'han encomanat la seva passió per la feina ben feta i dia a dia m'han deixat gaudir de la seva forma de ser i de tots els seus coneixements entre altres coses. Gavin, per moltes hores que ens quedéssim treballant, que sàpigues que mai se m'ha fet llarga una sola tarda. Thanks Gavin, gracias Roberto.

A tota la resta de membres que han fet possibles els projectes que hem realitzat, des dels participants en l'estudi REGICOR, fins a les persones que han recollit les dades i mostres, i també als que han dissenyat i dut a terme l'estudi, ja en marxa des d'abans que naixés. En especial, i per proximitat durant tot aquest temps, voldria agrair a tot l'equip d'estadística i maneig de dades de la unitat pel gran treball que realitzen, atenent sempre les demandes de tots els que els necessitem. En especial a l'Isaac i al Joan, per totes les hores que m'han dedicat i per tota la paciència que han tingut. Han estat tots dos uns grans mestres.

També vull deixar un record especial per les meves amigues que m'han acompanyat des que era petita (Eva, Laia, Nuria), i entre les quals encara conservo les meves millors amistats. Amb elles he compartit els millors i els pitjors moments de la meva vida, i res hauria estat possible sense elles. I com la vida sempre ens va deixant bones coses pel camí, no em voldria oblidar de les amistats que he fet al llarg d'aquests anys (Esther, Marta, Irene, Gerard), i que no han estat menys importants. Ah!, i amb tots els que he compartit els últims 4 anys (Mariana, Marc, Mari, Vaquero, Gianni, Giulia, Simone, Tobias, Edu)! Han estat realment uns anys fantàstics!!

PUBLICATIONS ARISING FROM WORK CARRIED OUT IN THIS THESIS

Lluís-Ganella C, Subirana I, Lucas G, Tomás M, Muñoz M, Sentí M, Salas E, Sala J, Ramos R, Ordovas JM, Marrugat J, Elosua R. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis 2012; Article in press.

Lucas G, **Lluís-Ganella C**, Subirana I, Sentí M, Willenborg C, Musameh MD, Schwartz SM, O'Donnell CJ, Melander O, Salomaa V, Elosua R. Post-Genomic Update on a Classical Candidate Gene for Coronary Artery Disease: *ESR1*. Circ Cardiovasc Genet. 2011; 4(6): 647-654.

Lluís-Ganella C, Lucas G, Subirana I, Sentí M, Jimenez-Conde J, Marrugat J, Tomás M, Elosua R. Additive effects of multiple genetic variants on risk of ischemic cardiopathy. Rev Esp Cardiol. 2010; 63(8):925-33.

Lluís-Ganella C, Lucas G, Subirana I, Escurriol V, Tomás M, Sentí M, Sala J, Marrugat J, Elosua R. Qualitative assessment of previous evidence and an updated meta-analysis confirms lack of association between the *ESR1* rs2234693 (*PvuII*) variant and coronary heart disease in men and women. Atherosclerosis 2009; 207(2): 480-486.

ORAL PRESENTATIONS AND AWARDS

Lluís-Ganella C, Subirana I, Lucas G, Tomás M, Muñoz M, Sentí M, Salas E, Sala J, Ramos R, Ordovas JM, Marrugat J, Elosua R. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis 2012; Article in press. *Presented at the Congress of the Spanish Society of Cardiology 2011, Maspalomas, Las Palmas de Gran Canaria, 2011.*

Lluís-Ganella C, Lucas G, Subirana I, Escurriol V, Tomás M, Sentí M, Sala J, Marrugat J, Elosua R. Qualitative assessment of previous evidence and an updated meta-analysis confirms lack of association between the *ESR1* rs2234693 (*PvuII*) variant and coronary heart disease in men and women. Atherosclerosis 2009; 207(2): 480-486. *Presented at the Congress of the Heracles Network: in the category of: "Best Publication from HERACLES 2009 from a Young Investigator", Granada 2010. Received the first prize.*

Lluís-Ganella C, Lucas G, Subirana I, Sentí M, Jimenez-Conde J, Marrugat J, Tomás M, Elosua R. Additive effects of multiple genetic variants on risk of ischemic cardiopathy. Rev Esp Cardiol. 2010; 63(8):925-33. *Presented at the Congress of the Heracles Network, Granada 2010.*

Lluís-Ganella C, Lucas G, Subirana I, Escurriol V, Tomás M, Sentí M, Sala J, Marrugat J, Elosua R. Qualitative assessment of previous evidence and an updated meta-analysis confirms lack of association between the *ESR1* rs2234693 (*PvuII*) variant and coronary heart disease in men and women. Atherosclerosis 2009; 207(2): 480-486. *Presented at the Congress of the Spanish Society of Cardiology: in the category of: "Premios de la Sociedad Española de Cardiología a las mejores Comunicaciones Libres de SEC 2009", Barcelona 2009. Received the consolation prize.*

ABBREVIATIONS

AHA: American Heart Association

BMI: Body Mass Index

CAD: Coronary Artery Disease

CARDIoGRAM: Coronary ARtery DIsease Genome-Wide Replicationa And Meta-analysis

CHD: Coronary Heart Disease

Chr: Chromosome

CI: Confidence Interval

CVD: CardioVascular Disease

CVRF: CardioVascular Risk Factor

DNA: DeoxiriboNucleic Acid

ESR1: Estrogen Receptor Alpha gene

GRS: Genetic Risk Score

GWAS: Genome Wide Association Study

HDL: High Density Lipoprotein

HGP: Human Genome Project

HR: Hazard Ratio

HRT: Hormone Replacement Therapy

HWE: Hardy-Weinberg Equilibrium

IDI: Integrated Discrimination Improvement

IHD: Ischaemic Heart Disease

LD: Linkage Disequilibrium

LDL: Low Density Lipoprotein

MeSH: Medical Subject Heading

MI: Myocardial Infarction

MIGen: Myocardial Infarction Genetics consortium

NHGRI: National Human Genome Research Institute

NRI: Net Reclassification index

OR: Odds Ratio

QS: Quality Score

REGICOR: REgistre GIroni del COR

ROC: Receiver Operating Characteristic

RNA: RiboNucleic Acid

RR: Relative Risk

SD: Standard Deviation

SNP: Single Nucleotide Polymorphism

STREGA: STrengthening the REporting of Genetic Associations

WHI: Woman Health Initiative

WHO: World Health Organization

WTCCC: Wellcome Trust Case-Control Consortium

GENERAL SUMMARY

The main expansion of the discovery of genetic variants associated with complex diseases has occurred during the last decade. This expansion has been accompanied, and in some sense motivated, by the desire to use this information to improve the predictive capacity of many diseases with an unidentified familial component, including coronary heart disease (CHD), with the aim of translating this genetic knowledge into clinical practice. This doctoral thesis is structured in two lines of investigation that address distinct aspects of this issue, first to evaluate the possible role of genetic variation in a candidate gene in modulating CHD risk, and second to evaluate whether genetic information can be used to improve risk assessment tools used in clinical practice.

In the first research line (described in *Part I*), I investigate the contribution of genetic variation in one of the most widely-studied genes in cardiovascular genetics, *ESR1*, which encodes the Oestrogen receptor α protein. I provide a solid meta-analysis of evidence regarding the most widely-studied variant in this gene and we further explore the role of a broad range of common and uncommon variants in this gene in CHD risk. Using these approaches, we find no evidence of association between the genetic variants studied and CHD risk. However, although we can confidently accept that common genetic polymorphisms are not associated with CHD, we cannot discard the possibility that other types of variation in this gene (for instance epigenetic variation) could modify susceptibility to CHD, or that other elements of this pathway are associated with an increased risk of CHD. In this research I have provided a reliable answer to this long running unanswered question in cardiovascular genetics, allowing research to re-focus on other elements of this system or other pathways.

In the second line (described in *Part II*), I explored the possible utility of genetic information obtained from genome-wide association studies (GWAS) in prediction of 10-year risk of CHD events by adding this information to cardiovascular risk functions. I have followed the recommendations proposed by the American Heart Association for evaluating the utility of novel biomarkers in clinical practice, and have demonstrated that although the magnitudes of the effects of these genetic variants on CHD risk are modest, there is a tendency towards improvement in the capacity of the risk functions to predict future CHD events. The translation of genetic information into clinical practice was one of the main motivations for the investment in genome-wide association studies, and my research represents one of the first efforts to explore this possibility.

RESUM GENERAL

L'expansió principal pel que fa al descobriment de variants genètiques associades amb malalties complexes s'ha dut a terme durant la última dècada. Aquesta expansió ha estat acompanyada, i d'alguna forma motivada, pel desig d'usar aquesta informació per millorar la capacitat de predicció d'aquelles malalties on hi és present un cert component familiar però en les que no es coneixien les variants que conferien un major risc de patir la malaltia, entre elles la cardiopatia isquèmica (CI). La present tesis doctoral està estructurada en dues línies d'investigació que avaluen el possible rol d'un gen candidat en la susceptibilitat de la CI i també avalua la millora en la capacitat de predicció d'un esdeveniment coronari de les eines usades habitualment en la pràctica clínica mitjançant la inclusió d'informació genètica.

Més concretament, la primera línea d'investigació es centra en la contribució de la variació genètica en un dels gens més estudiats en relació amb CI: el gen que codifica pel receptor d'estrogens α (*ESR1*). En aquesta línea he proveït un sòlid meta-anàlisis entre la variant més àmpliament estudiada d'aquest gen i risc coronari i també hem explorat el paper de la majoria de les variants comunes descrites en aquest gen i risc de CI. Mitjançant cap dels anàlisis he trobat evidència d'associació entre les variants genètiques en aquest gen i el risc de CI. No obstant això, i encara que podem acceptar que les variants genètiques comunes d'aquest gen no estan associades amb esdeveniments coronaris, no podem descartar que altres tipus de variació en aquest gen (com per exemple variació epigenètica) pugui estar modificant la susceptibilitat a patir un esdeveniment coronari, ni tampoc que altres elements de la mateixa cadena de senyalització estiguin associats amb la malaltia. En aquesta recerca he donat resposta a una qüestió que havia estat plantejada des de feia molts anys en el camp de la genètica de malalties cardiovasculars, permetent als investigadors de centrar tots els seus esforços en investigar altres elements d'aquest sistema o altres rutes metabòliques.

En la segona línea d'investigació, hem explorat el possible paper de les variants genètiques, obtingudes mitjançant estudis d'associació global del genoma (GWAS), en la millora de la capacitat de predicció a 10 anys dels esdeveniments coronaris, mitjançant la seva addició en les funcions de risc cardiovascular clàssiques. Hem seguit les recomanacions proposades per la American Heart Association per l'avaluació en la pràctica clínica de nous biomarcadors, i hem demostrat que, tot i que la magnitud de l'associació d'aquestes variants és modesta, hi ha una tendència cap a la millora de la capacitat de predicció de les funcions de risc. La translació de la informació genètica en la pràctica clínica fou una de les principals motivacions dels estudis d'associació global del genoma, i la meva recerca representa un dels primers intents de posar en pràctica aquesta translació.

TABLE OF CONTENTS

0. PREFACE

The difference in incidence of cardiovascular diseases between males and females is one of the most striking observations in the epidemiology of ischaemic heart disease but still remains largely unexplained. Not even the observed differences between genders in terms of cardiovascular risk factors, lifestyle, environmental exposures, or any other known differences are able to explain this. Historically, the fact that this difference disappears almost completely after menopause has driven researchers to focus on physiological or social characteristics that also change during this period, such as the reproductive hormone system.

Previous studies have extensively evaluated but not conclusively determined the role of genetic variation in one of the most attractive candidate genes for cardiovascular risk, *ESR1* (encoding the Oestrogen Receptor α protein). In this context, my aim was to perform a detailed exploration of the genetic variation in this gene in order to provide a more definitive answer to this question. The results of these studies are presented in the first research line of this doctoral thesis.

An important priority in modern medical research is that the knowledge obtained be rapidly and efficiently applied in clinical practice. Thus, one of the main motivations of my research is to explore how genetic information can be used to improve cardiovascular medicine.

During the last decade there has been an impressive expansion in our understanding of the genetic basis of complex diseases. My research represents one of the first attempts to begin the process of translating this new information into clinical practice, the results of which are presented in the second research line of this doctoral thesis.

In this work, the approaches I have taken and the techniques I have used to investigate the role of genetic variation in cardiovascular risk are among the most advanced in our field, and are applicable to most other complex diseases. This is mainly thanks to the fact that I have been fortunate to conduct my research during one of the most important periods of advance in our ability to study complex diseases.

1. INTRODUCTION: Table of contents

1.1. GENERAL INTRODUCTION

1.1.1. Epidemiological studies

Definition of epidemiology and uses

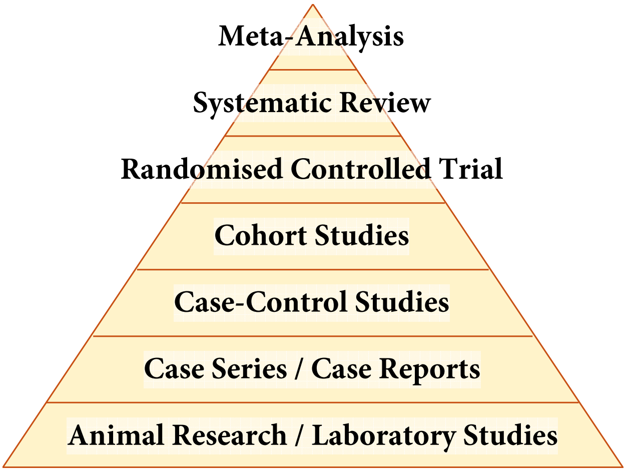
The main focus of epidemiology is the study of the distribution and determinants of disease frequency in populations and the application of this study to control health problems [Aschengrau, 2008]. Factors that are associated with an increased probability of a specific disease are known as risk factors, and can be studied using several types of study designs.

There are two main groups of designs: *interventional or experimental studies*[*] (clinical trials) and *observational studies* (such as cohort or case-control studies) (see *Box 1*). Although interventional studies are the preferred design because they provide the most powerful evidence, the use of observational studies can help to avoid some of the economic and ethical limitations that clinical trials or interventional studies might have, in the case of both human subjects as well as other living organisms.

The different types of study designs can be classified in a hierarchy (see *Box 1*), according to the power of evidence provided by each design. The World Health Organization (WHO) has established a working group to develop a common, sensible and transparent approach to grading the quality of evidence. This grading is based on four domains: study design, study quality, consistency (measuring the internal validity of a study), and directness (external validity of a study, refers to the extent to which the people, interventions, and outcome measures are similar to those of interest. E.g. there may be uncertainty about the directness of the evidence if the people of interest are older or sicker than those in the studies) [Atkins, 2004].

[*] Items highlighted in this font are defined and described in the glossary, page 167

Box 1. A hierarchical representation of the different **types of study designs**, according to the power of the evidence they provide.


Meta-Analysis
Systematic Review
Randomised Controlled Trial
Cohort Studies
Case-Control Studies
Case Series / Case Reports
Animal Research / Laboratory Studies

Advantages and disadvantages of the use of case-control or cohort studies.

To evaluate the influence of an exposure in the incidence of a disease given a certain population, and depending on the study design, some measurements can be obtained, such as relative risk or the odds ratio:

Cohort Studies	Case-Control Studies
Useful for common disorders with short latency periods	Useful for rare diseases or common diseases with long latency periods
Need to be relatively large for the study of rare diseases	Relatively smaller for any frequency of disease. Can be even smaller for rare diseases
Allow study of multiple disease outcomes and multiple exposures	Usually focused on one disease, but allow study of multiple exposures
Considers the effect of time	Limited ability to identify temporal patterns
Relative risk (RR) is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group [Sistrom, 2004]. RR is often used when the binary outcome that is being measured has a relatively low probability. It is thus often suited to clinical trials or cohort studies.	*Odds ratio* (OR) is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group [Sistrom, 2004].

These two measures can be computed as follows:

	Cases	Controls	
Exposed	a	b	a + b
Non-exposed	c	d	c + d
	a + c	b + d	

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \qquad OR = \frac{a \cdot d}{b \cdot c}$$

In both measures, a value of 1 means absence of effect of the factor evaluated, a value <1 means a protective effect, and a value >1 means a harmful effect.
These two measures can be considered almost equivalent, except when the incidence of a disease is very high, and there is a big difference in the incidences of the disease in those individuals exposed and non-exposed to the factor.

Example. Three different scenarios are presented below. In each scenario, the incidence of the disease varies among the exposed and unexposed groups, but the total of individuals (10,000), and the prevalence of smokers (40% of smokers) is constant in all situations.

- **Scenario 1:** low disease incidence:
  - disease incidence among smokers: 4%
  - disease incidence among non smokers: 0.4%

- **Scenario 2:** high disease incidence:
  - disease incidence among smokers: 40%
  - disease incidence among non smokers: 4%

- **Scenario 3:** high disease incidence in both exposed and non-exposed:
  - disease incidence among smokers: 40%
  - disease incidence among non smokers: 35%

Scenario 1

	cancer	no cancer	
Smokers	160	3,840	4,000
Non-smokers	24	5,976	6,000
	184	9,816	10,000

Scenario 2

	cancer	no cancer	
Smokers	1,600	2,400	4,000
Non-smokers	240	5,760	6,000
	1,840	8,160	10,000

Scenario 3

	cancer	no cancer	
Smokers	1,600	2,400	4,000
Non-smokers	2,100	3,900	6,000
	3,700	6,300	10,000

RR→

$$\frac{\frac{160}{4{,}000}}{\frac{24}{6{,}000}} = 10.0 \qquad \frac{\frac{1{,}600}{4{,}000}}{\frac{240}{6{,}000}} = 10.0 \qquad \frac{\frac{1{,}600}{4{,}000}}{\frac{2{,}100}{6{,}000}} = 1.1$$

OR→

$$\frac{160 \cdot 5{,}976}{24 \cdot 3{,}840} = 10.4 \qquad \frac{1{,}600 \cdot 5{,}760}{240 \cdot 2{,}400} = 16.0 \qquad \frac{1{,}600 \cdot 3{,}900}{2{,}100 \cdot 2{,}400} = 1.2$$

Note that the values obtained in the RR and the OR are more similar when the incidence of the disease is low, but also when the difference in the disease incidence in the exposed group and the non-exposed group is small.

Experimental studies: introduction, definition and uses

In biology, when we want to evaluate the “effect” that a certain characteristic or factor has on another condition (for example, a disease), the ideal type of study design is generally an interventional study, in which the investigator designs an experiment to evaluate the changes in disease status of a group exposed to the factor when compared to an unexposed group. The exposure to the factor of interest is usually introduced by the investigator in a random way: some individuals are randomly exposed to the factor and the others are not exposed (randomised clinical trial). The aim of randomisation is to ensure that the two groups are equal in all ways except with regard to their exposure to the factor of interest. If this is true and the disease or condition is more frequent in one of the two groups, we can conclude that the factor is causally associated with the condition. As an example, to evaluate the effects of a pesticide, we plant 200 roses at the same time, and monitor all the conditions. During a two month period, we treat half of the flowers with pesticide and leave the rest untreated. If the untreated roses flower normally, and the treated roses fail to develop flowers, we could conclude that the pesticide prevents flower development in roses. If we try to perform the same type of experiment in other species, for instance humans, we may encounter ethical problems. For example, we could not expose individuals to chemicals with unknown effects in order to identify their side effects. Therefore, other types of study designs are often required for the evaluation of specific relations with certain exposures.

Non-experimental studies: Cohort studies

While there are several types of non-experimental or observational study design [Rothman, 2008] (e.g. cohort studies, proportional mortality studies, case-control, case-cohort, etc.), cohort studies give the closest approximation to the true influence of a risk factor in the populations under study. A cohort study is an analytical study in which a representative sample of individuals from a population is selected, each individual’s level of natural exposure to the factor of interest is measured, and the rate of occurrence of the outcome of interest (outcome *incidence*) during a specific time period (follow-up) is recorded [Rothman, 2008]. A requirement of this type of study is that all subjects must be free of disease at the time of enrolment, because in these longitudinal studies we compare the incidence of the disease of interest in those who are exposed and non-exposed to the factor under study (time to event). A classical example of this type of study is the Framingham Heart Study, which started in 1948 with the follow-up of 5,209 residents of Framingham (Massachusetts, USA). This study has identified most of the currently known risk factors for cardiovascular disease (www.framinghamheartstudy.org).

Non-experimental studies: Case-control studies

Due to the elevated cost and the long period of time required to perform cohort studies, another type of study design that has also been widely used is the case-control design. In this type of study, a group of individuals who have a specific disease (cases) are compared with a group of individuals who are free from that disease (controls) [Rothman, 2008] (see *Box 1*). A classical example of an application of this type of study design is the demonstration by John Snow that persons who drank well water from the Broad Street Pump in London in 1849 had a much higher death rate from cholera than those who did not. The author also showed that the rate of death from cholera was much higher among those who had drunk water that had been polluted by sewage (www.sph.umich.edu/epid/GSS/pub.html).

Deciding when to use cohort and case-control study designs

These distinct study designs can give answers to the same or different questions (see *Table 1*), and each has strengths and limitations (see *Box 1*). A common criticism of case-control studies is the potential for biases related to the individual levels of the exposures and the selection of participants. These biases are related to the fact that data collection is carried out after the disease occurs [White, 1998] and often relies on medical records and patient recall (recall bias). In addition, the cases in case-control studies are the same cases that would normally be included in a cohort study, whereas the sampling of controls from the population that gave rise to the cases is a key issue in this type of study and affords the efficiency gain from a case-control design over a cohort design [Rothman, 2008], especially for those diseases with low incidence, in which very large numbers of individuals would need to be collected in order to have the same number of events with a cohort study compared to a case-control study. However, as researchers' general understanding of the principles and limitations of case-control studies has evolved, their design and acceptance has also improved. Currently, case-control studies are commonly used to study factors associated with disease because cohort studies are very expensive and usually require long-term follow-up. A case-control study can be conceptualised as a more efficient version of a cohort study, while this advantage could be hampered by the potential for greater biases. Therefore, the choice of study design will depend on the type of question to be addressed (see *Table 1*), and on the resources available.

Table 1. Types of study designs that can be used to address specific questions.

Type of question	Recommended study design
Therapeutic	Randomised clinical trial; cohort; case-control; case series
Diagnostic	Prospective; blind comparison to a gold standard
Aetiology (the study of causation)	Randomised clinical trial; cohort; case-control; case series
Prognostic	Cohort; case-control; case series
Prevention	Randomised clinical trial; cohort; case-control; case series
Clinical Exam	Prospective; blind comparison to a gold standard
Cost	Economic analysis

A brief description of the specific studies that have been used in the present doctoral thesis is presented in Section 7.1.

Genetic epidemiology: Genetic factors as potential risk factors

Many different types of factors can be evaluated as potential contributors to disease incidence, including environmental elements (pollution, water contaminants, etc.), sociodemographic characteristics (age, socioeconomic position, etc.), and physical or *phenotypic* characteristics (body weight, blood pressure levels, etc.). Classically, because of their big contribution on disease, the most widely studied factors have been both environmental and physical characteristics. However, there is a clear pattern of inheritance in some diseases since the incidence among siblings of patients with a disease have increased risk with respect to the general population. Therefore, we could also define genetics as a possible risk factor. In this sense, genetic epidemiology deals with the aetiology, distribution, transmission and management of disease among relative individuals, and with heritable factors that contribute to disease risk in populations [Morton, 1982].

As a result of classical epidemiological studies, in which various clinical forms of heart disease were found to occur more frequently in individuals with a family history of the disease (familial aggregation) [Thomas, 1955; White, 1957; Mayer, 2007], it has been known for several decades that cardiovascular diseases (CVD) have an important hereditary component; this has also been observed for many other diseases (e.g. cancer). More recent studies indicate that a family history of ischaemic heart disease in parents [Lloyd-Jones, 2004] or siblings [Marenberg, 1994; Murabito, 2005] is a risk factor for development of the disease, independent of traditional risk factors. This familial aggregation highlights the genetic component of cardiovascular disease risk, although this could also be related to environmental and behavioural factors, which are common and also display familial aggregation [Deutscher, 1966].

1.1.2. Brief introduction to genetics

Origins of the concept of genetics and Mendel’s laws

Beginning with the observations of Darwin and Wallace in the 19th century, the process of natural selection began to gain acceptance as an explanation for the design of living organisms [Barahona, 2009], where individuals were gradually transformed from simpler to more complex life forms [Fontdevila, 2009]. Later in the same century, Mendel initiated the concept of modern genetics through his experiments on the inheritance of morphologic characteristics in the pea plant, on the basis of which he formulated two laws, the “Law of Segregation” and the “Law of Independent Assortment” [Barahona, 2009]. His observations were possible due to the strong *penetrance* of the phenotypic characteristics studied in the pea plant. Mendel's observations were of extreme importance in determining the pattern of inheritance observed in some diseases, and in understanding the way in which genetic information is transmitted between generations.

Discovery of Mendelian inheritance patterns in human disease

In 1902, Garrod observed that the disease *alkaptonuria* followed the same patterns of inheritance that Mendel had described in the pea plant [Garrod, 1902; Dronamraju, 1992]. This meant that the inheritance pattern observed in some plants was not exclusive to that kingdom, and could be extrapolated to some human diseases, and these would later be denoted as *Mendelian diseases* (discussed in Griffiths *et al.* [2000]). To date, the genetic basis of more than 2.000 Mendelian diseases have been established, including familial hypercholesterolemia [Civeira, 2004], familial defective apolipoprotein B-100 [Innerarity, 1987] and Brugada Syndrome [Lehnart, 2007]. However, from the point of view of public health, these diseases generally affect few individuals and do not have a very significant impact on the health of the general population.

Exceptions to Mendel's Law of Independent Assortment

In 1905 Bateson coined the word "genetics" for the study of heredity, and also demonstrated the general validity and importance of Mendelian inheritance (as reported by Harper *et al*. [2005] and Barahona *et al.* [2009]). In the same year, he observed, together with Saunders and Punnett, one of the earliest exceptions to normal Mendelian ratios. In their work with pea plants, these researchers noticed that not all of their crosses yielded results that reflected the principle of independent assortment. Specifically, some phenotypes appeared far more frequently than Mendelian genetics would predict [Bateson, 1909]. Based on these findings, the researchers proposed that certain alleles must somehow be coupled or linked to each another, although they weren't sure how this linkage occurred. The answer to this question came in 1911, when Morgan demonstrated that linked genes must be real physical objects that are located in close proximity on the same chromosome [Morgan, 1911; Lobo, 2008]. Based on the observation that some traits are inherited together, Morgan and colleagues also defined the first genetic map of an organism (*Drosophila melanogaster*), and deduced that these traits must be located in a linear arrangement on the chromosomes.

Moreover, Morgan and his team discovered that some characteristics in the fruit fly were determined by the combined action of two or more genes, which is reminiscent of the type of inheritance observed in complex diseases as reported by Barahona *et al.* [2009]. In contrast to Mendelian diseases, complex diseases, which account for the majority of morbidity and mortality in industrialised countries, are caused by a combination of genetic as well as environmental and behavioural factors, and are the diseases with the greatest impact on general population health. The complex combination of various risk factors makes the identification of the genetic component of complex diseases more difficult than for Mendelian diseases.

DNA as a carrier of genetic information

In 1944, deoxyribonucleic acid (DNA) was identified by Avery and colleagues as the carrier of the genetic information [Avery, 1944], although some doubts remained about the assertion that genetic information was contained in DNA. At that time, the appearance of genetic effects was thought to be under the direct control of proteins. Therefore, the identification and characterisation of the DNA molecule was a crucial objective, not only in order to understand this molecule, but also to determine how and where the genetic information was stored. The idea of one gene-one enzyme had already been suggested as early as 1917, although with limited experimental support as reported by Beadle *et al.* [1941], but this theory became accepted when Watson and Crick described the structure of the double helix in 1953 [Watson, 1953; Arber, 1978]. The importance of this theory was centred on the idea that genes encoded in the DNA were responsible for the generation of proteins, which were, in turn, the elements responsible for known molecular actions. In their discovery article, Watson and Crick proposed what is now accepted as the first correct double-helix model of DNA structure and described a novel feature of this structure, which is the manner in which the two chains are held together [Watson, 1953]. This model focused attention, in particular, on the biological meaning of its physical structure [Khorana, 1968], but one of the most valuable characteristics of this discovery was the fact that DNA is composed of two mirror strands or chains; while genes are encoded on just one strand, the sequence of either strand can be established by determine the sequence of the other one. This is called complementary base pairing.

Cracking the Genome: the development of genotyping techniques

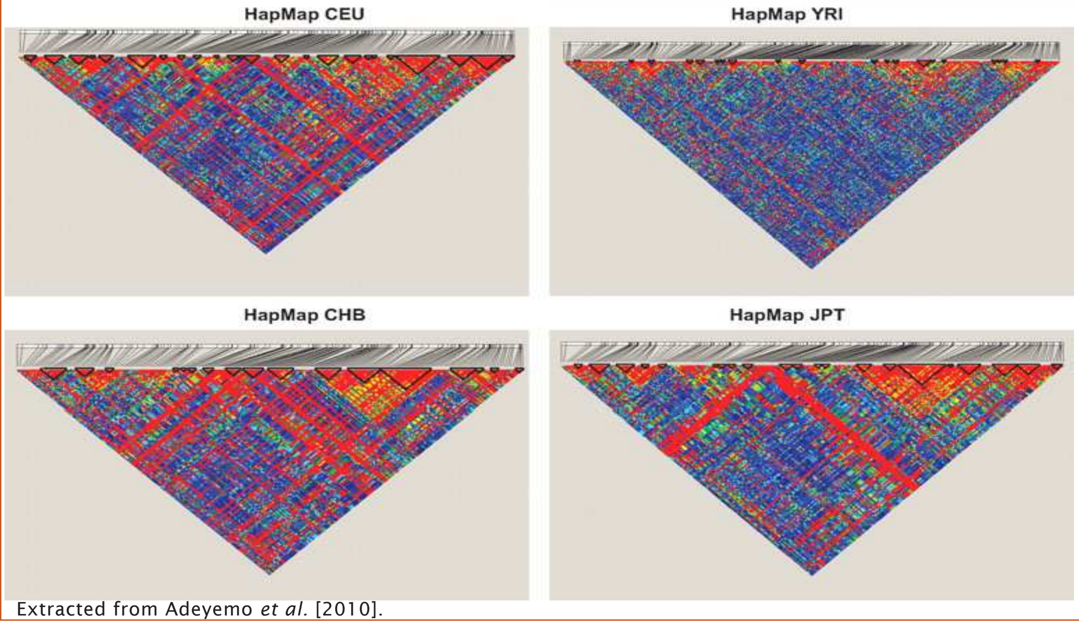
Regardless of the structural characteristics of DNA, the so-called "Cracking of the Genome" could not begin until the discoveries of Ochoa and Kornberg in the 1950s, who identified the enzyme PNPase (Polynucleotide Phosphorylase), and its ability to synthesize RNA in vitro. In his Nobel lecture, Ochoa commented that, *"Since there are good indications that the genetic information stored in DNA is first transmitted to RNA, it is believed that DNA may function as a template for RNA replication"* [Ochoa, 1959]. However, the enzymology of DNA began to develop rapidly with the work of Kornberg and co-workers [Nobelprize.org, 2006], who detailed molecular images of RNA polymerase (the molecule responsible for DNA translation) during various stages of the transcription process. The discovery of this enzyme clarified the manner in which information in DNA is transcribed into RNA, now known as messenger RNA (mRNA), and made possible the development of techniques that are still used for identification of genetic polymorphisms (for example the invention of the polymerase chain reaction (PCR) technique by Mullis in 1986 [Mullis, 1986]). For more than two decades, PCR (and its adaptations) has been the most widely used technique in research into the *genetic architecture* of disease.

Describing the human genome: The Human Genome Project & The HapMap Project

The early 1990s saw the start of the *Human Genome Project* (HGP; 1990-2003), whose main aim was to map and sequence the human genome [Watson, 1989; Pearson, 1991; Roberts, 2001; Venter, 2001] (genomics.energy.gov). In 2003, the goals of the HGP were achieved with the completion of the first human genome sequence (see *Box 2*).

That same year also saw the initiation of another project that has had a crucial role in the field of genetics, *The International HapMap Project* (HapMap; 2002-2009). HapMap was a multi-country effort designed to identify and catalogue human genetic variation [The International HapMap Consortium, 2003], and played an important role in providing better estimates of allele frequencies [Fellay, 2007], identifying additional variants for testing, and defining patterns of correlation between them (*linkage disequilibrium, LD*) [Manolio, 2008]. LD patterns across the genome were found to have a block structure (see *Figure 1*), which is the result of the molecular mechanism of chromosomal recombination throughout the history of our species [The International HapMap Consortium, 2007]. By computing the LD between variants across the genome, the HapMap project established that by genotyping only a small number of *single nucleotide polymorphisms (SNPs)* (called *tag SNPs*), the majority of common genetic variation throughout the genome could be captured. The first phase of the HapMap was completed in 2005, and phases II and III where carried out later [The International HapMap Consortium, 2007].

Figure 1. Linkage Disequilibrium patterns across the *FTO* gene in four HapMap population samples (European: CEU; East Asian: CHB and JPT; and African: YRI).



Extracted from Adeyemo *et al.* [2010].

From genetic variation to the discovery of the genetic basis of disease

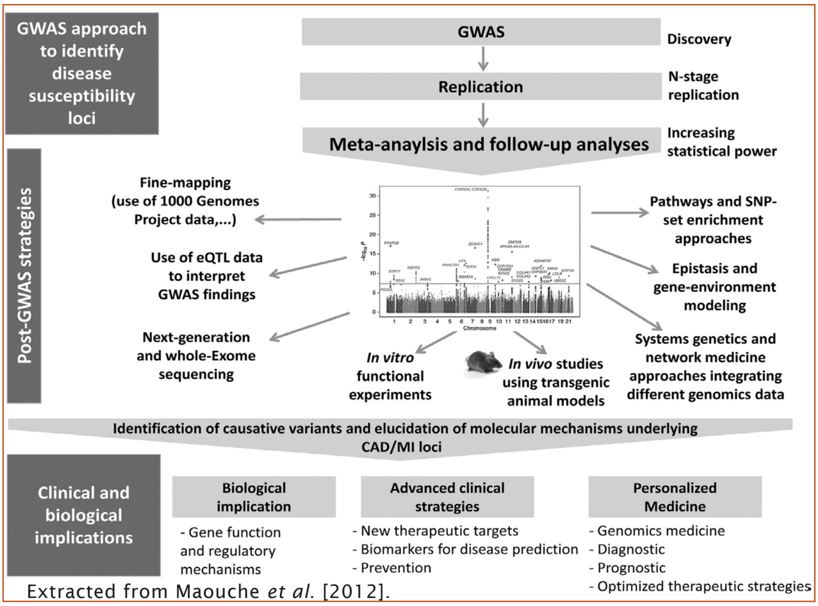
Once the LD patterns across the genome had been clarified, researchers were able to use this information to locate genes involved in clinically important traits. Moreover, thanks to this researchers can now perform a more detailed exploration of specific association with disease of candidate genes, or even search on genome-wide scale for chromosomal regions that may be associated with a disease. These resources have driven disease gene discovery during the first generation of *genome-wide association studies (GWAS)*, in which having data for hundreds of thousands of variants allow to test for association with disease for the vast majority of common variants in the genome (in this context, variants with a minor allele frequency of ≥5% are generally referred to as common) [The International HapMap Consortium, 2007].

Present and future of genotyping techniques

In recent decades, great improvements in the methodologies used to study the influence of genetic variation on diseases have been achieved through the use of information from collaborative projects such as the HGP and HapMap projects in combination with the advanced statistical methods that take advantage of the correlation between common variants in order to impute genotypes at additional variants not directly tested [Hirschhorn, 2011]. These technological developments and developments in high-throughput genotyping technologies have driven an order-of-magnitude expansion of genetic studies on a wide range of diseases in recent years [Visscher, 2009], from studying one or few genetic variants at a time to hundreds of thousands in a single experiment (GWAS) [Manolio, 2008], and more recently using complete sequences of the exome or the entire genome [Singleton, 2011]. Moreover, the rapid drop in cost and increase in scale of DNA sequencing has often been compared to the trend seen in the semiconductor industry in the second half of the twentieth century, which was described by Moore's law (which described that the number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two year) [Muers, 2011]. This is true to the point where the rate-limiting step is not variant identification, but the management and interpretation of the resulting data [Cooper, 2011]. This technical expansion has been accompanied by the creation of multi-centre consortiums and broad data-sharing, which have supported the identification of many new genetic variants associated with complex diseases.

Ways to use this information

In order to define a variant as being associated with a specific disease, it must first be identified, and the association replicated in multiple independent samples. Several such studies can then be meta-analysed in order to increase statistical power to identify additional variants (see upper part of *Figure 2*). Once these genetic variants are identified, the question that arises is how they may affect disease risk. To answer this question, several types of studies can be performed, including functional studies in animal models, integration of different sources of data (e.g. gene expression), and others (see central part of *Figure 2*). Finally, studies to demonstrate the clinical and biological effects of the genetic variants on disease can also to be performed (*Figure 2*, lower part). Above all, one of the most important motivations for the investments made in GWAS technology was the expectation that this new information could be translated to clinical practice; this issue will be explored in more detail in the next section and is also the focus of *Part II* of this doctoral thesis.


GWAS approach to identify disease susceptibility loci
GWAS
Discovery
Replication
N-stage replication
Meta-anaylsis and follow-up analyses
Increasing statistical power
Post-GWAS strategies
Fine-mapping (use of 1000 Genomes Project data,...)
Use of eQTL data to interpret GWAS findings
Next-generation and whole-Exome sequencing
Pathways and SNP-set enrichment approaches
Epistasis and gene-environment modeling
Systems genetics and network medicine approaches integrating different genomics data
In vitro functional experiments
In vivo studies using transgenic animal models
-log10 P
Chromosome
Identification of causative variants and elucidation of molecular mechanisms underlying CAD/MI loci
Clinical and biological implications
Biological implication
- Gene function and regulatory mechanisms
Advanced clinical strategies
- New therapeutic targets
- Biomarkers for disease prediction
- Prevention
Personalized Medicine
- Genomics medicine
- Diagnostic
- Prognostic
- Optimized therapeutic strategies

Extracted from Maouche *et al.* [2012].

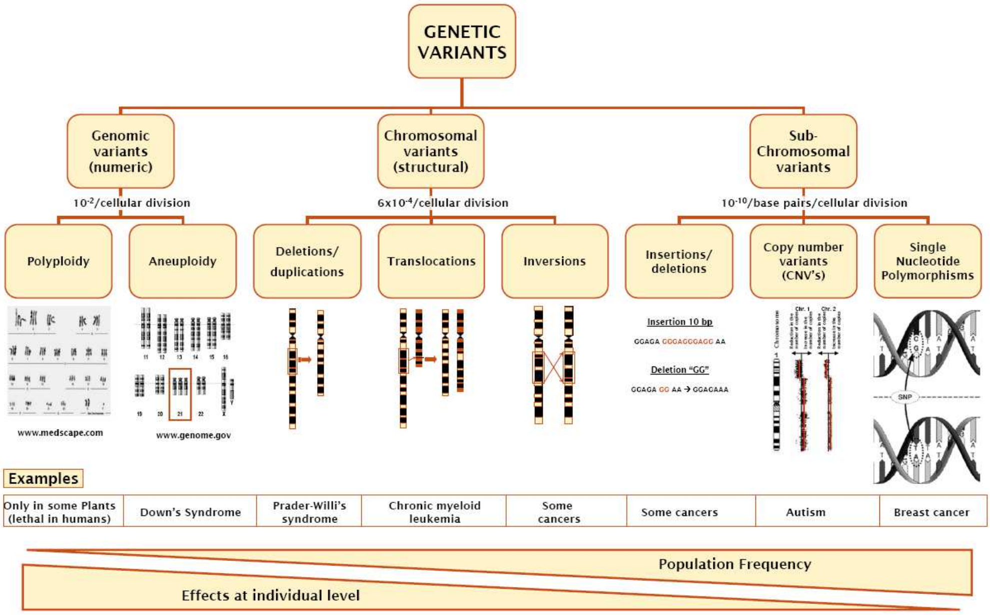
Figure 2. Post-genome-wide association study (GWAS) strategies, and biological and clinical implications of GWAS findings.

Box 2. **Genomes** vary widely in size: the smallest known genome for a free-living organism (a bacterium) contains about 600,000 DNA base pairs, while human and mouse genomes have ~3 billion. Except for mature red blood cells, all human cells contain a complete genome.

Each chromosome (physically separate molecules that range in length from about 50 million to 250 million base pairs) contains many genes, the basic physical and functional units of heredity. Genes comprise only about 2% of the human genome (which is estimated to contain >30,000 genes); the remainder consists of non-coding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

> Note that the numbers presented in this box have been extracted from the summary of the GenBank data provided by NCBI, based on genomic sequence information available on Oct 05, 2011.

There are several types of **genetic variation**, classified according the amount of genetic material that is involved in that genetic variation:



GENETIC VARIANTS

Genomic variants (numeric)

10^{-2}/cellular division

Chromosomal variants (structural)

$6x10^{-4}$/cellular division

Sub-Chromosomal variants

10^{-10}/base pairs/cellular division

Polyploidy

Aneuploidy

Deletions/ duplications

Translocations

Inversions

Insertions/ deletions

Copy number variants (CNV's)

Single Nucleotide Polymorphisms

www.medscape.com

www.genome.gov

Insertion 10 bp

GGAGA GGGAGGGAGG AA

Deletion "GG"

GGAGA GG AA → GGAGAAA

Examples

Only in some Plants (lethal in humans)	Down's Syndrome	Prader-Willi's syndrome	Chronic myeloid leukemia	Some cancers	Some cancers	Autism	Breast cancer

Population Frequency

Effects at individual level

> Note that we have to consider that there are also genome modifications that do not involve a change in the nucleotide sequence (i.e. DNA methylation and histone modification), known as epigenetic variation.

By 2012 the public catalogue of variant sites (dbSNP135) contained approximately 41 million SNPs (www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi).

1.1.3. Translation from genetics to clinics

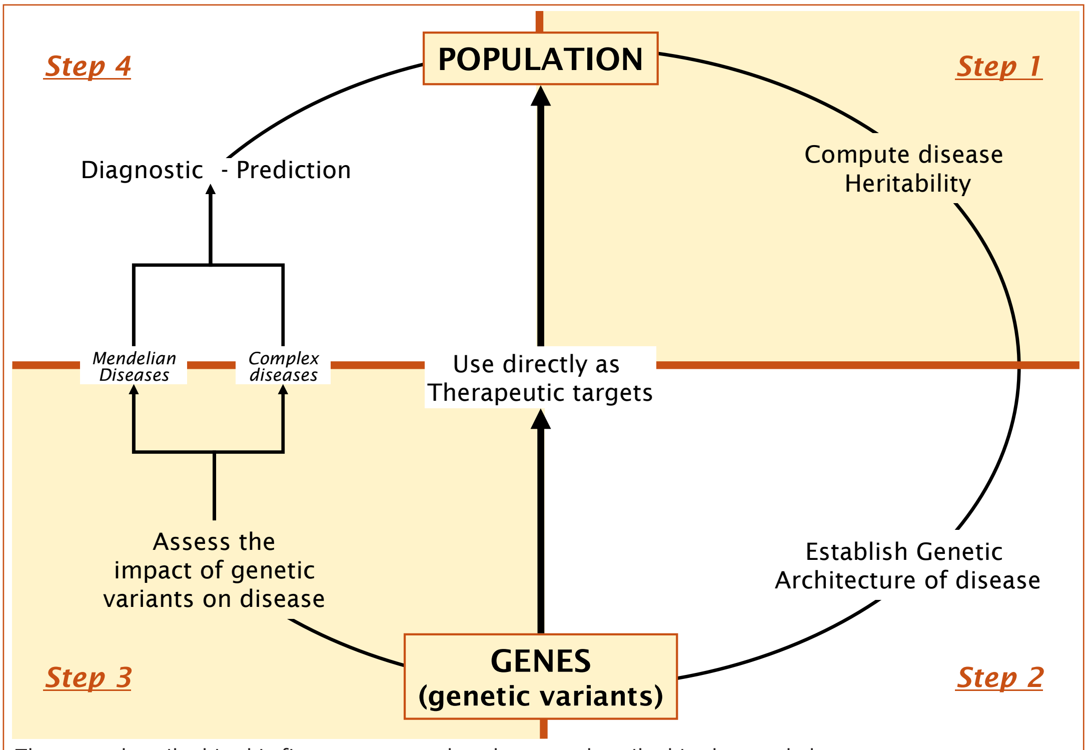
Why and when is it necessary to translate genetics to clinical practice?

One of the primary motivations for investing in large GWAS studies was to translate this information to clinical practice, which can be achieved in two main ways (see *Figure 3*):

i) *Determining new therapeutic targets*. Until now, the selection of targets to be studied in genetics was mainly based on knowledge of physiology. Currently, the hypothesis-free nature of GWAS allows us identify new target regions in the genome, and may help identify new drug targets for the treatment or prevention of disease. A classic example is the gene encoding HMGCoA reductase, variation in which explains only a small fraction of the total variance in cholesterol levels, but is a target for powerful cholesterol-lowering drugs [Lander, 2011; Zuk, 2012].

ii) *Improving diagnosis and prediction of disease*. For some diseases, early detection, often before signs of the disease are visible, is crucial for the survival of the individual, and the identification of individuals that are at high risk of disease is of particular importance. Genetic information can serve as a diagnostic tool (mainly in the case of monogenic or Mendelian diseases such as Tay-Sachs Disease), or it can also be used to improve the predictive capacity of other tools (mainly for complex diseases).

However, before using genetic information in practical applications at the population level, a series of natural steps, shown in *Figure 3* and detailed below, need to be taken. Briefly, these steps go from identifying and quantifying the genetic contribution to disease risk to evaluating the use of this information in clinical practice.


Step 4
POPULATION
Step 1
Diagnostic - Prediction
Compute disease Heritability
Mendelian Diseases
Complex diseases
Use directly as Therapeutic targets
Assess the impact of genetic variants on disease
Establish Genetic Architecture of disease
Step 3
GENES (genetic variants)
Step 2

The steps described in this figure correspond to the steps described in the text below.

Figure 3. Steps from populations to genes, and from genes to populations.

Step 1: Determine whether the disease being studied has a genetic component

To quantify this genetic contribution, measures such as heritability may be used (see *Box 3*). This heritability can be defined as the proportion of phenotypic variation that can be attributed to genetic variation [Visscher, 2008]. In order to compute the heritability values, mainly family-based or twin studies are required. If the disease under evaluation has a genetic component, the next step will be designed to identify which are the specific genetic determinants of that disease.

Step 2: Establish the genetic architecture of the disease

Second, the genetic variants that confer an excess in risk of the disease must be identified (*Part I* of the present doctoral thesis is designed to answer this specific question). The identification of these genetic variants has been widely successful for simple Mendelian diseases as they also usually express a highly penetrant phenotype and a very clear inheritance pattern. In the case of complex diseases, the approaches used for Mendelian diseases have not been very successful to identify genetic factors, because they generally make a smaller contribution to these diseases.

Step 3: Determine the impact of the genetic variability on disease risk

The third step that needs to be achieved is to determine the amount of variance in risk of the disease that can be explained by the genetic variants observed. For example, if the sole presence of a SNP was necessary and sufficient for a disease status, we would say that 100% of the risk variance was explained by that variant. For Mendelian diseases the variance explained by genetic variants is much higher than for complex diseases.

Step 4: Evaluate the utility of the genetic information in the diagnosis, prediction, prevention and treatment of the disease

In the last step, the utility of genetic variation in general clinical practice is evaluated. To this end, specific guidelines have been proposed by the American Heart Association [Hlatky, 2009] for the evaluation of novel cardiovascular risk biomarkers, and these guidelines can be extended to genetic markers, and might also be applied to other diseases (discussed in more detail in *Part II* of the doctoral thesis).

Box 3. The **Phenotypic variance** (V_P) observed within a population consists of a genetic component (V_G) and/or an environmental component (V_E), which can be computed as follows [Falconer, 1996; Lynch, 1998]:

$$V_P = V_G + V_E$$

Note that in order to determine the values for both V_G and V_E, researchers must consider that both genetic and environmental sources of variation are a composite of different components.

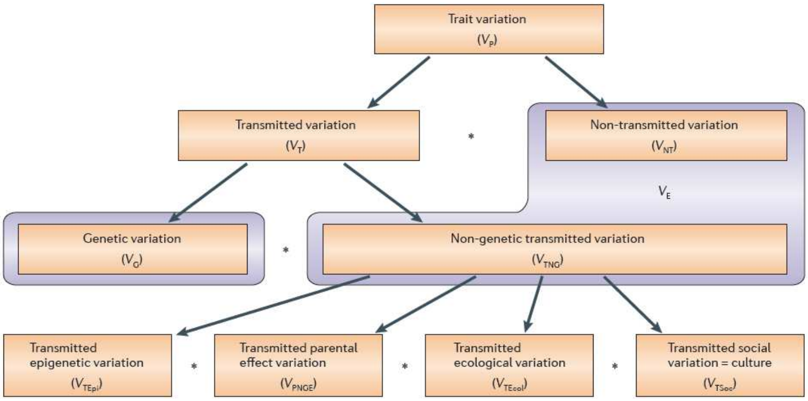
- **Environmental Variation:** Environmental variation can be subdivided into various subcategories; including specific environmental variance (V_{Es}: deviation from the population mean due to environmental conditions experienced by each individual, known as residual variance or error), general environmental variance (V_{Eg}: non-genetic sources of variation between individuals that are experienced by many individuals in a population), and genotype-environment interaction (V_{GxE}: involves the unique or different responses of genetic lines to general environmental variation):

$$V_E = V_{Es} + V_{GxE} + V_{Eg}$$

- **Genetic Variation:** Genetic variation can also be divided into several subcategories, including additive variance (V_A: deviation from the phenotypic mean due to inheritance of a particular allele and this allele's relative effect on phenotype), dominance variance (V_D: involves deviation due to interactions between alternative alleles at a specific locus), and epistatic variance (V_I: involves an interaction between alleles at different loci):

$$V_G = V_A + V_D + V_I$$

Schema of the quantitative contribution of the components of inclusive heritability, from Danchin *et al.* [2011].


Trait variation (V_P)
Transmitted variation (V_T)
*
Non-transmitted variation (V_{NT})
V_E
Genetic variation (V_G)
*
Non-genetic transmitted variation (V_{TNG})
Transmitted epigenetic variation (V_{TEpi})
*
Transmitted parental effect variation (V_{PNGE})
*
Transmitted ecological variation (V_{TEcol})
*
Transmitted social variation = culture (V_{TSoc})

Phenotypes that vary between individuals in a population do so because of both environmental factors and the genes that influence traits, as well as various interactions between genes and environmental factors. Therefore, we can measure the proportion of phenotypic variation in a population that is due to genetic variation between individuals, a measure known as **heritability** [Visscher, 2008].

Under this last definition, there are two types of heritability estimate:

- **Broad-sense heritability:** The proportion of phenotypic variation due to genetic variation, including effects due to dominance and epistasis.

$$H^2 = \frac{V_G}{V_P}$$

- **Narrow-sense heritability:** Captures only that proportion of genetic variation that is due to additive genetic values.

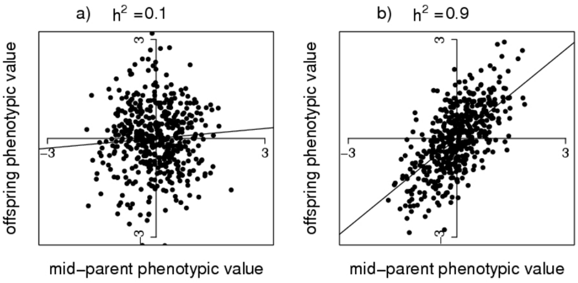
$$h^2 = \frac{V_A}{V_P}$$

Note that since heritability is a proportion, its numerical value will range from 0 (genetic variation does not contribute to individual phenotypic differences in any way) to 1 (genetic variation is responsible for all individual variation).

Example. Traditionally, heritability was estimated from simple, often balanced, designs, such as the correlation between offspring and parental phenotypes, the correlation between full and half siblings, and the difference in correlation between monozygotic (MZ) and dizygotic (DZ) twin pairs. In artificial selection experiments, heritability can also be estimated from the ratio of the observed selection response (R) to the observed selection differential (S). This relationship is summarized in the "breeder's equation", R = h^2S, where:

- R: is the “Response to Selection”, which is the difference between the mean of the parents before selection and the mean of the offspring.
- S: is the “Selection Differential”, which is the difference between the mean of the population and the mean of the individuals that reproduce.

Therefore, it is possible to estimate *Narrow sense heritability* simply from the regression of offspring phenotypic values on the average of parental phenotypic values. The following example [Wray, 2008] concerns traits with high (0.9) and low (0.1) heritability.



Note that to compute the proportions of the environmental and genetic contributions to disease risk, it is usually assumed that the resemblance between monozygotic and dizygotic twins due to shared environment is the same.

Note also that the Breeder’s equation is mainly used in selective breeding of plants and animals.

There are also more sophisticated ways of estimating heritability [Lee, 2011], although a specific description of the methods available is beyond the scope of this thesis.

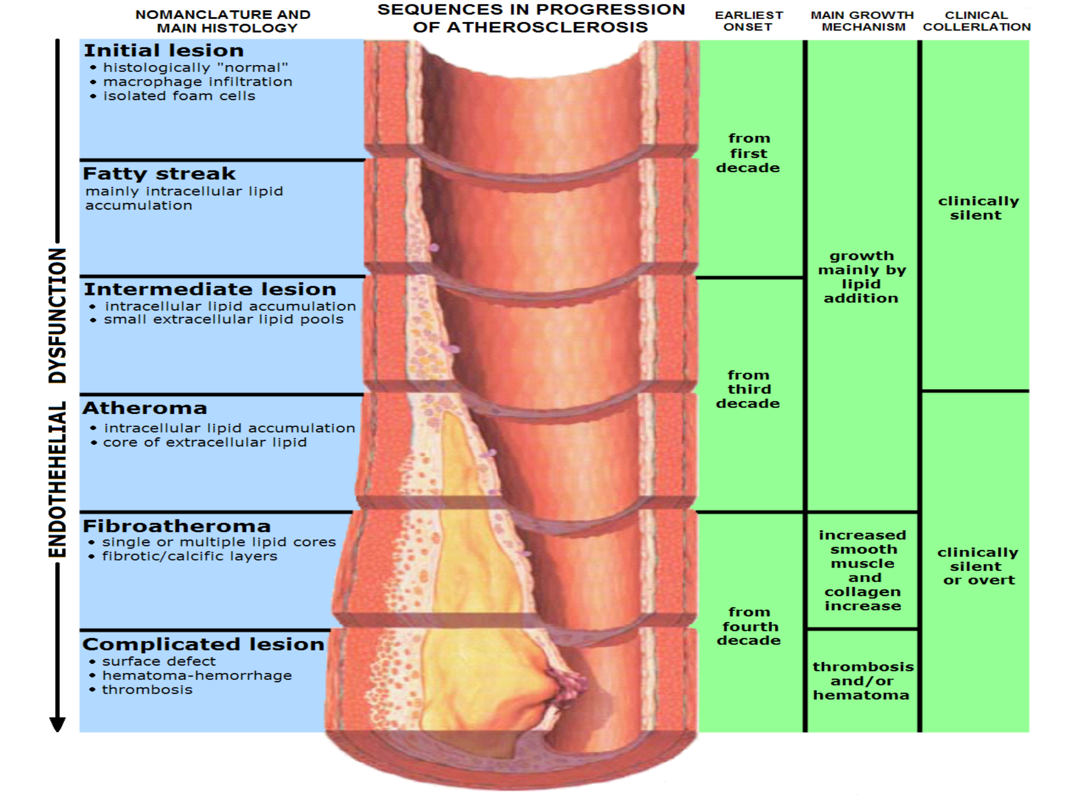
1.1.4. Coronary heart disease: an example of complex disease

Definition and stages of atherosclerosis

Atherosclerosis can be defined as a chronic, complex inflammatory disease that causes a narrowing of the small blood vessels that supply oxygen to the cells, due to the formation of atheroma plaques consisting of deposits of cholesterol and other lipids, which ultimately cause a chronic inflammatory response in the artery walls) [Ross, 1999]. The American Heart Association [Stary, 1995] identifies six stages of atherosclerosis progression (see *Figure 4*):

- *Type I (initial):* isolated macrophage foam cells.
- *Type II (fatty streak):* primarily intracellular lipid accumulation.
- *Type III (intermediate):* small extracellular lipid pools.
- *Type IV (atheroma):* extracellular lipid core.
- *Type V (fibroatheroma):* Lipid core and fibrotic layer; formation of prominent new fibrous connective tissue.
- *Type VI (complicated):* complicated surface, with haemorrhage or thrombus formation.

Figure 4. Graphical representation of the stages of progression of arteriosclerosis over time.


NOMANCLATURE AND MAIN HISTOLOGY
SEQUENCES IN PROGRESSION OF ATHEROSCLEROSIS
EARLIEST ONSET
MAIN GROWTH MECHANISM
CLINICAL COLLERLATION
ENDOTHEHELIAL DYSFUNCTION
Initial lesion
• histologically "normal"
• macrophage infiltration
• isolated foam cells
Fatty streak
mainly intracellular lipid accumulation
Intermediate lesion
• intracellular lipid accumulation
• small extracellular lipid pools
Atheroma
• intracellular lipid accumulation
• core of extracellular lipid
Fibroatheroma
• single or multiple lipid cores
• fibrotic/calcific layers
Complicated lesion
• surface defect
• hematoma-hemorrhage
• thrombosis
from first decade
from third decade
from fourth decade
growth mainly by lipid addition
increased smooth muscle and collagen increase
thrombosis and/or hematoma
clinically silent
clinically silent or overt

Released under the GNU Free Documentation License: commons.wikimedia.org/wiki/File:Endo_dysfunction_Athero.PNG.

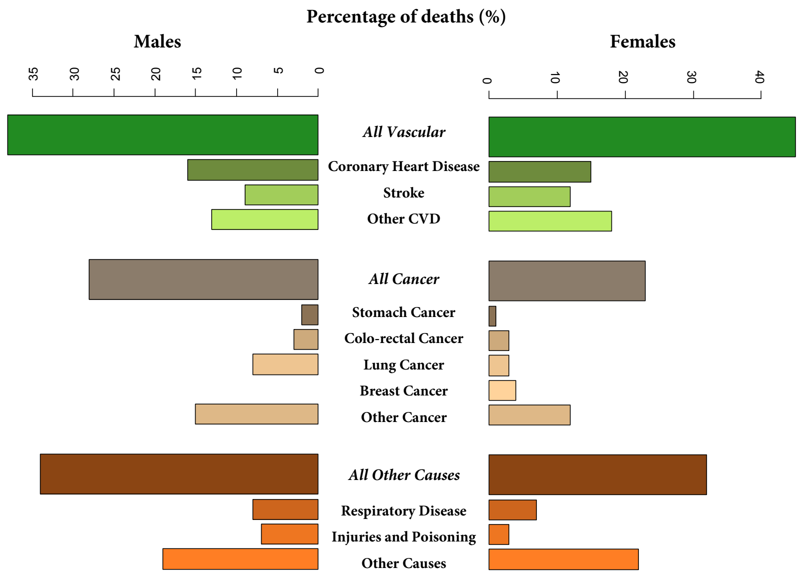
Atherosclerotic plaques can also be separated into two broad categories: stable and unstable (also called vulnerable) [Ross, 1999]. Vulnerable plaques are rich in foam cells, lipids and inflammatory cells, and have a thin fibrous cap. These types of plaques are prone to rupture, which causes an acute thrombus that may occlude the arterial lumen, triggering an acute cardiovascular event. Stable plaques are rich in extracellular matrix and smooth muscle cells, making them more difficult to break, with the result that they are usually asymptomatic.

Definition and manifestations of coronary heart disease

Coronary heart disease (CHD) is one of the main manifestations of atherosclerosis. It is a complex disease characterised by various clinical presentations, a complex aetiopathogenesis, and a strong environmental component (diet, smoking habit, physical activity). The two main clinical manifestations of CHD are acute coronary syndrome and stable angina. Two main types of acute coronary syndrome have been characterised: *i)* myocardial infarction (MI), which results from the interruption of blood supply to a part of the heart, causing heart cells to die (in order to determine an MI event it is required that evidence of myocardial necrosis exists by laboratory tests) [Thygesen, 2007]; and *ii)* unstable angina, which is a strong indicator of an impending MI, is caused by a reduction of coronary blood flow due to transient platelet aggregation on apparently normal endothelium, coronary artery spasms (temporary, sudden narrowing of one of the coronary arteries) or coronary thrombosis (formation of a clot in one of the arteries that conduct blood to the heart muscle) [Lenfant, 2010]. In stable angina, the blood flow and oxygen supply to the myocardium is compromised, causing oppressive chest discomfort/pain that occurs mainly when performing some physical activity and usually disappears with rest.

Global burden of disease

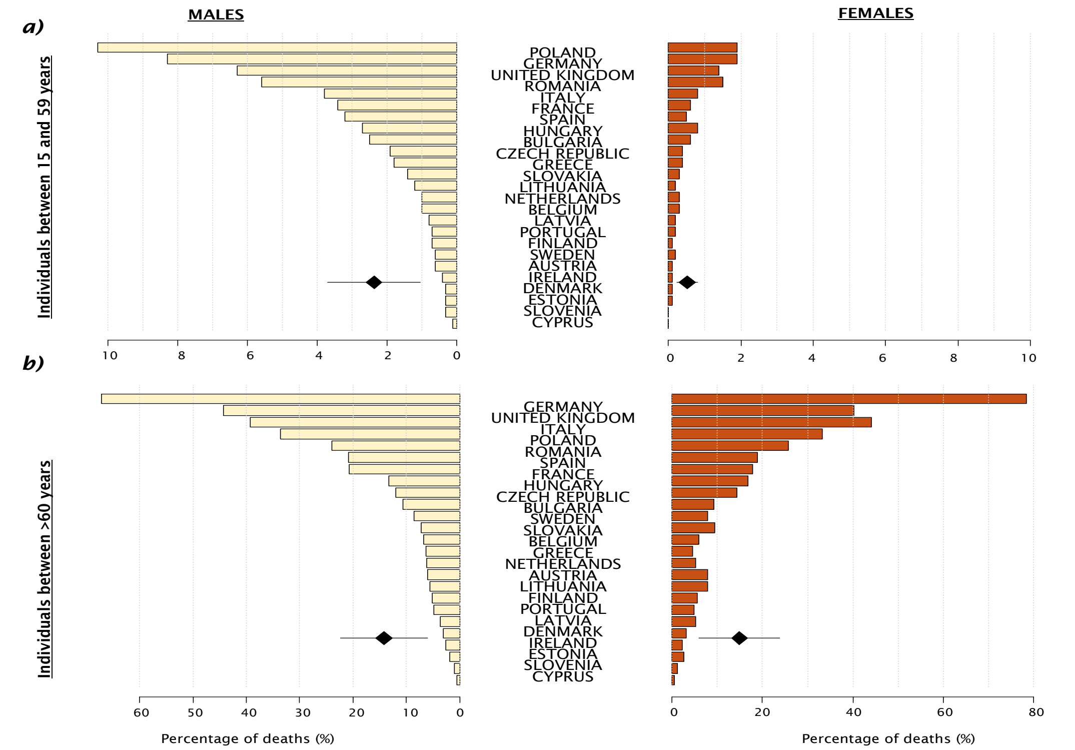
CHD accounts for nearly 20% of deaths worldwide [European Heart Network, 2008; World Health Organization, 2009] (see *Figure 5*) and it mainly occurs from the fifth or sixth decade in men, and from the sixth or seventh decade in women (see *Figure 6*). Moreover, the incidence of CHD is expected to increase in the coming decades due to an increase in the prevalence of cardiovascular risk factors (CVRFs; described in more detail below).



CVD: Cardiovascular disease.

Figure 5. Causes of death in Europeans (data extracted from the European Heart Network Report [2008]).

Figure 6. Deaths estimates due to ischaemic heart disease in 2008 stratified by country from the European Union (data extracted from the European Heart Network Report [2008]).



a) Death estimates for individuals between 15 and 59 years, and separated by sex.
b) Death estimates for individuals with 60 or more years, and separated by sex.

Factors that increase risk of coronary heart disease

The first study performed to identify the causes or factors that modulate risk of CHD to be completely performed and that obtained the most remarkable results was the Framingham Heart Study (www.framinghamheartstudy.org) [Dawber, 1951]. The main objective of that project was to search for factors that influence the development of disease by performing a cohort study comparing individuals who had had a cardiovascular event to the group who did not suffer from any event at the end of the follow-up [Dawber, 1951]. After just 4 years of follow-up, the authors observed that certain attributes were strongly related to risk of developing CHD, such as elevated lipid levels and elevated blood pressure. Later, they also identified smoking, excess of body weight, lack of physical activity, low vital capacity (the maximum amount of air a person can expel from the lungs after a maximum inspiration), gout, and diabetes as relevant risk factors. Moreover, they also concluded that when more than one of these risk factors were present, there was a marked increase in susceptibility to CHD [Dawber, 1966]. These findings had important consequences for our understanding of the physiological aspects of the disease, as well as for primary prevention. Currently, the cardiovascular risk factors (CVRFs) that are considered to cause an increase in risk can be classified as modifiable or non-modifiable (see *Table 2*). Also, the importance of genetic factors in risk prediction has long been appreciated, and is exemplified in a simple form by the value of family history in increasing CVD risk [Jostins, 2011].

Table 2. Classification of cardiovascular risk factors.

Non-modifiable	**Modifiable**
Age	Smoking
Sex	Hypertension
Genetics	Hypercholesterolemia
Family history	Type II Diabetes
	Body Mass Index
	Physical activity

1.1.5. Current state of the field of complex disease genetics

Problems in genetic studies: Lack of robust findings

Using the candidate gene approach, a large number of genetic variants have been studied during the last 20 years, but robust, replicable evidence has been found for only a limited number, while the majority showed contradictory results between studies [Ioannidis, 2001]. This lack of robust findings for genetic studies in complex diseases may be due to different reasons. One of the possible explanations has been described as the “winner’s curse” effect, where the magnitude of the associations observed for the genetic variants is higher in early studies than in subsequent replication studies. This effect may represent either *i)* a spurious finding that is not validated by subsequent research, *ii)* an exaggerated finding that eventually finds its appropriate measure, or *iii)* an effect that is stronger in some subpopulations than in others [Ioannidis, 2001].

Some recommendations have been proposed to increase the reliability of the reported evidence, particularly the STREGA guidelines (STrengthening the REporting of Genetic Association studies) [Little, 2009], which suggest a series of steps that might be followed in order to enhance the transparency of the reporting of genetic association studies. Briefly, these guidelines are an extension of the STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) [von Elm, 2007] statement that include recommendations on items that are specifically relevant to genetic studies, including laboratory methods for genotyping and allele-calling, genotyping accuracy, *haplotype* modelling, population stratification, relatedness between subjects, and statistical adjustments to allow for multiple hypothesis testing [Hudson, 2009]. STREGA also recommends that the stage involving identification of genetic variants be followed by a subsequent stage of validation of those findings in independent populations in order to verify or improve the estimate of the true effect [Hlatky, 2009] and minimise the publication of false positive results.

Problems in genetic studies: Correct selection of genes/loci

Another of the reasons of such lack of robust results in genetic association studies was the selection of appropriate genes and genetic variants to test for association with diseases. Even more, the correct coverage of the genes selected had been shown to be a poorly considered methodological point for many studies [Drago, 2007]. Moreover, the selection of those loci had to be performed on the basis of pathophysiological knowledge, biasing in some cases the possible loci that presented a real association with disease. During the last decade, the development and application of GWAS study designs, which facilitate genotyping of hundreds of thousands of SNPs in thousands of individuals, have largely circumvented the need to select specific candidate genes or SNPs for study.

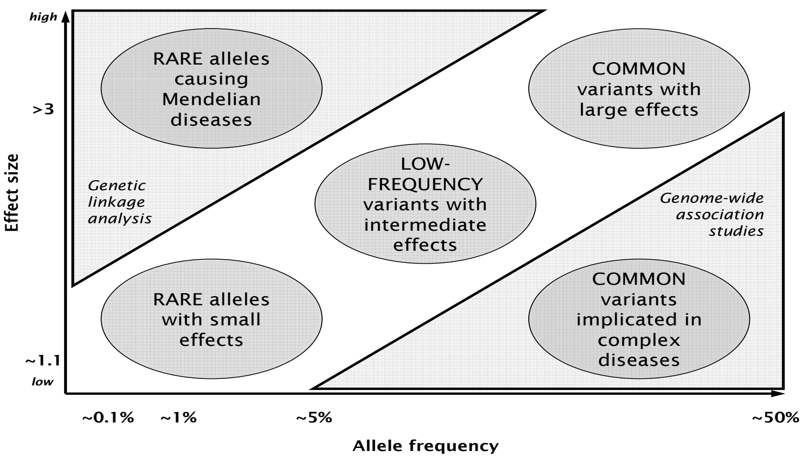
How much information have we gained using GWAS? Is it enough?

Advanced new methods for studying genetic variation including high-throughput genotyping [Ding, 2009], GWAS [Manolio, 2008], genotype imputation [Howie, 2009], second generation sequencing [Wheeler, 2008], in combination with some projects that describe natural human genetic variation (e.g. HapMap [2007] or 1000 Genomes Project [Durbin, 2010]) allow us to explore the effect of genetic variation on phenotype more thoroughly. Moreover, the effort that scientists are making to collaborate and to standardise their procedures and reporting of the articles is increasing the quality and reliability of the works published. Since 2006-2007 a large number of GWAS have been performed for complex diseases revealing loci robustly associated with these diseases [Visscher, 2009; Jostins, 2011].

Although the number of loci identified is increasing, these variants still explain only a minor proportion of the heritability of complex diseases (e.g. ~10% for variants associated with CHD [The CARDIoGRAM Consortium, 2011]), leading to the question of where the ‘‘missing heritability’’ lies. To explain this observation, various hypotheses have emerged that focus on two main lines of thought: the first one questions the initial heritability estimates and the second considers genetic or non-genetic factors that could explain this missing heritability [Manolio, 2009]: *i)* a much larger numbers of variants with smaller effects yet to be found; *ii)* rarer variants (possibly with larger effects) that are poorly captured by current genotyping arrays, which focus on variants with a population frequency ≥0.05; *iii)* structural variants, some of which may be poorly captured by current arrays; *iv)* gene-gene interactions; *v)* shared environment among relatives; *vi)* true causal may have stronger effects than those estimated for the observed variant as a result of imperfect LD between them; *vii)* heritable effects that are not represented by primary genomic sequence.

Characteristics of the variants obtained with GWAS

The variants identified by GWAS for complex diseases generally have minor allele frequencies ranging from 5-50%, and also tend to have moderate to weak effects on the phenotype [Park, 2010]. In contrast, variants identified by linkage analysis to be associated with Mendelian diseases have large effect sizes and rarer allele frequencies (see *Figure 7*). The reason why these variants could be detected by linkage analysis is that they usually cause severe disruptions or truncations of the encoded protein that are a strong predictor of disease onset, thereby resulting in clearly recognisable familial inheritance patterns. This fact made linkage analysis highly successful in the mapping and identification of the genetic causes of Mendelian diseases [Hirschhorn, 2011]. Common variants with large effect sizes are expected to be removed from the population by the process of (purifying) natural selection [Ng, 2006] if they express their effects before reproductive age, and therefore, are largely absent from the populations.


high
>3
~1.1
low
Effect size
RARE alleles causing Mendelian diseases
COMMON variants with large effects
Genetic linkage analysis
LOW-FREQUENCY variants with intermediate effects
Genome-wide association studies
RARE alleles with small effects
COMMON variants implicated in complex diseases
~0.1%
~1%
~5%
~50%
Allele frequency

Adapted from Manolio *et al.* [2009] and McCarthy *et al.* [2008].

Figure 7. Effect sizes observed and expected for genetic variants associated with complex or Mendelian (or monogenic) diseases.

Is there more?

Results obtained from GWAS do not discard that other previously studied candidate genes, selected on the basis of pathophysiological knowledge, could still play important roles in the development of the diseases. Studies performed in the past could still bring us valuable information. For example, maybe some common variants could not be captured correctly by the classical genotyping techniques and therefore the association would have remained undetected, or some rare variants could have been not detected by GWAS because of their effect size. Evidence from GWAS to support the possibility of some rarer variants having greater effects than the common variants that tagged the rarer variant on a disease can be found for example in an article by Nejentsev *et al.* [2009], where they were able to detect greater effects in risk of type I diabetes with 4 rare variants (variants with ~1% frequency) in a gene, than with a common variant associated with the disease in the same gene. Therefore, some other type of variants (i.e. rare variants, epigenetic variants, etc.) could be the cause of this so called missing heritability. In the following article, published in the Spanish Journal of Cardiology, a review about the state of the field in 2009 of genetics of ischaemic heart was disease was presented.

1.1.6. Revision Article: Elosua, R; Lluís, C; Lucas, G.

Ressearch into the genetic component of heart disease: from linkage studies to genome-wide genotyping. *Rev Esp Cardiol Supl. 2009; 9:24B-38B.*

1.1.7. Update on cardiovascular genetics research in the GWAS era

In the case of CHD, the majority of loci now known to be robustly associated with disease risk have been discovered since the beginning of the GWAS era. Most of these loci have been replicated in meta-analyses of those GWAS (meta-GWAS), involving more than 200,000 individuals in total [The CARDIoGRAM Consortium, 2011; Coronary Artery Disease (C4D) Genetics Consortium, 2011] (see *Table 3*). Briefly, the CARDIoGRAM consortium [2011] performed a meta-analysis of 14 genome-wide association studies and 56,682 additional replication samples (26 studies), representing 143,677 individuals of European descent in total. Simultaneously, the Coronary Artery Disease (C4D) Genetics Consortium [2011] performed a similar meta-GWAS including 71,075 individuals from four large GWAS carried out in individuals of European and South Asian descent. The majority of GWAS focussed on CHD and MI that have been performed up to 2011 have been included in one of these two meta-GWAS consortia. Thus, the results provided by these two studies (detailed in *Table 3*) consolidated the evidence reported by the contributing GWAS studies, and capture most or all of the best evidence currently available regarding the genetic component of CHD risk.

						CARDIoGRAM		C4D GC	
	Band	**SNP**	**Gene(s) in region**	***n***	**RAF (RA)**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
Previously known loci	1p32.3	rs11206510	*PCSK9*	102,352	0.82 (T)	1.08 (1.05–1.11)	9.10×10^{-08}	1.05 (1.00-1.11)	7.00×10^{-02}
	1p13.3	rs599839	*SORT1*	83,873	0.78 (A)	1.11 (1.08–1.15)	2.89×10^{-10}	1.14 (1.09-1.19)	6.05×10^{-10}
	1q41	rs17465637	*MIA3*	25,197	0.74 (C)	1.14 (1.09–1.20)	1.36×10^{-08}	1.09 (1.02-1.16)	1.13×10^{-02}
	2q33.1	rs6725887	*WDR12*	77,954	0.15 (C)	1.14 (1.09–1.19)	1.12×10^{-09}	1.11 (1.05-1.19)	6.19×10^{-04}
	3q22.3	rs2306374	*MRAS*	77,843	0.18 (C)	1.12 (1.07–1.16)	3.34×10^{-08}	1.08 (1.02-1.13)	4.34×10^{-03}
	6p24.1	rs12526453	*PHACTR1*	83,050	0.67 (C)	1.10 (1.06–1.13)	1.15×10^{-09}	1.11 (1.07-1.15)	5.82×10^{-08}
	6q25.3	rs3798220	*LPA*	32,584	0.02 (C)	1.51 (1.33–1.70)	3.00×10^{-11}	-	-
	9p21.3	rs4977574	*CDKN2A,CDKN2B*	84,256	0.46 (G)	1.29 (1.23–1.36)	1.35×10^{-22}	1.20 (1.16-1.25)	1.62×10^{-25}
	10q11.21	rs1746048	*CXCL12*	136,416	0.87 (C)	1.09 (1.07–1.13)	2.93×10^{-10}	1.06 (1.01-1.10)	8.52×10^{-03}
	12q24.12	rs3184504	*SH2B3*	67,746	0.44 (T)	1.07 (1.04–1.10)	6.35×10^{-06}	1.05 (1.00-1.11)	6.61×10^{-02}
	19p13.2	rs1122608	*LDLR*	49,693	0.77 (G)	1.14 (1.09–1.18)	9.73×10^{-10}	1.09 (1.02-1.15)	5.88×10^{-03}
	21q22.11	rs9982601	*MRPS6*	46,230	0.15 (T)	1.18 (1.12–1.24)	4.22×10^{-10}	1.09 (1.03-1.15)	3.13×10^{-03}
New loci	1p32.2	rs17114036	*PPAP2B*	133,226	0.91 (A)	1.17 (1.13–1.22)	3.81×10^{-19}	-	-
	6p21.31	rs17609940	*ANKS1A*	137,412	0.75 (G)	1.07 (1.05–1.10)	1.36×10^{-08}	-	-
	6q23.2	rs12190287	*TCF21*	130,888	0.62 (C)	1.08 (1.06–1.10)	1.07×10^{-12}	-	-
	7q32.2	rs11556924	*ZC3HC1*	134,200	0.62 (C)	1.09 (1.07–1.12)	9.18×10^{-18}	-	-
	9q34.2	rs579459	*ABO*	123,978	0.21 (C)	1.10 (1.07–1.13)	4.08×10^{-14}	-	-
	10q24.32	rs12413409	*CYP17A1, CNNM2, NT5C2*	129,741	0.89 (G)	1.12 (1.08–1.16)	1.03×10^{-09}	-	-
	11q23.3	rs964184	*ZNF259, APOA5-A4-C3-A1*	135,492	0.13 (G)	1.13 (1.10–1.16)	1.02×10^{-17}	-	-
	13q34	rs4773144	*COL4A1, COL4A2*	114,731	0.44 (G)	1.07 (1.05–1.09)	3.84×10^{-09}	-	-
	14q32.2	rs2895811	*HHIPL1*	114,238	0.43 (C)	1.07 (1.05–1.10)	1.14×10^{-10}	-	-
	15q25.1	rs3825807	*ADAMTS7*[*]	129,652	0.57 (A)	1.08 (1.06–1.10)	1.07×10^{-12}	1.07 (1.05-1.10)	3.71×10^{-09}
	17p13.3	rs216172	*SMG6, SRR*	111,538	0.37 (C)	1.07 (1.05–1.09)	1.15×10^{-09}	-	-
	17p11.2	rs12936587	*RASD1, SMCR3, PEMT*	129,600	0.56 (G)	1.07 (1.05–1.09)	4.45×10^{-10}	-	-
	17q21.32	rs46522	*UBE2Z, GIP, ATP5G1, SNF8*	137,633	0.53 (T)	1.06 (1.04–1.08)	1.81×10^{-08}	-	-
	10q23.31	rs1412444	*LIPA*	66,220	0.42 (T)	-	-	1.09 (1.07-1.12)	2.76×10^{-13}
	11q22.3	rs974819	*PDGFD*	64,881	0.32 (T)	-	-	1.07 (1.04-1.09)	2.41×10^{-09}
	7q22.3	rs10953541	*7q22*	65,906	0.80 (C)	-	-	1.08 (1.05-1.11)	3.12×10^{-08}
	10q11.23	rs2505083	*KIAA1462*	68,102	0.38 (C)	-	-	1.07 (1.04-1.09)	3.87×10^{-08}

Minor allele frequency average: 0.28. Band: Chromosomal band; OR (95%CI): Odds ratio (95% Confidence interval); n=number of individuals with available data; RAF (RA): Risk allele Frequency (Risk Allele); SNP: Single Nucleotide Polymorphism.

[*] In the C4D GC the variant reported for this gene was rs4380028 (C allele).

1.2. GLOBAL HYPOTHESES

i) Some previously studied candidate genes for coronary heart disease (CHD) harbour genetic variants that are associated with disease risk, but have not yet been discovered because of the low genetic coverage or the small sample size of previous studies.

ii) The information provided by genetic variants that modulate increased risk of CHD, could be used to improve estimates of coronary risk above those provided by classical cardiovascular risk factors alone.

1.3. GLOBAL OBJECTIVES

i) To explore the role of variation in a previously studied candidate gene for CHD risk through an evaluation of previous evidence and the use new post-genomic resources and tools.

ii) To evaluate whether genetic variants identified by GWAS to be associated with CHD risk improve coronary risk estimation when added to classical risk functions.

Table 3. Summary of results from the CARDIoGRAM [2011] and C4D [2011] consortia, representing current best evidence for the genetic component of CHD risk.

2. PART I: Table of contents

Evaluation of the role of genetic variation in ESR1 *on risk of CHD*

PART I

2

2.0. ABSTRACT

ESR1 has been one of the most widely studied candidate genes for CHD, driven by the observation of marked differences in CHD risk between sexes and the implication of the reproductive hormone system as possible explanation. In this part of the doctoral thesis, I perform a qualitative and quantitative evaluation of all reported evidence regarding a widely studied putative association between the rs2234693 variant in the first intron of the *ESR1* gene and CHD. I update and extend two previous meta-analyses of association studies in >32,000 individuals. I also provide evidence to suggest that the quality of studies is a key determinant of the results obtained in the genetic association analyses they report. I also exploit powerful new post-genomic methods and resources to perform more thorough evaluation of the role of genetic variation in the *ESR1* gene in risk of CHD, in order to resolve this long-running unanswered question in cardiovascular genetics. For common variation in a genomic region centred on *ESR1*, I present association results from a large meta-analysis of GWAS of MI and CHD. I also perform an *in-silico* fine mapping analysis of additional common and uncommon genetic variation in this region, and explore possible gender differences. None of the genetic variants tested for association with CHD present a statistically significant association, suggesting that primary genetic variation in this gene is not the cause of differences in CHD risk.

2.1. INTRODUCTION

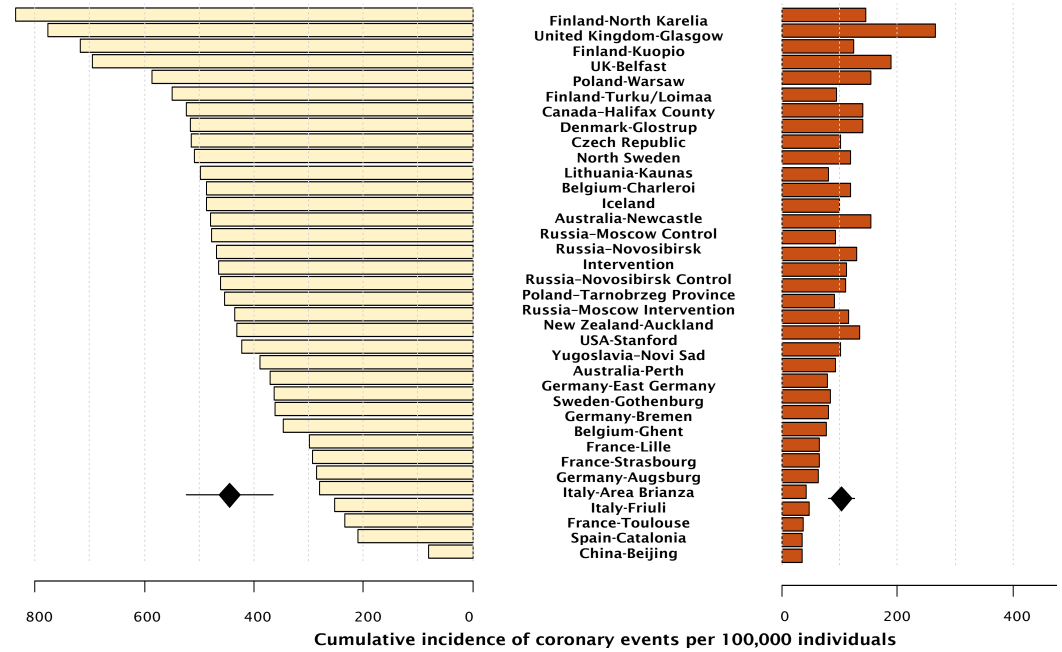
2.1.1. Gender differences in CHD risk

Epidemiologic evidence of gender differences in CHD risk

After age, gender is the most important risk factor for CHD events, with women aged 35 to 64 years having two to four times lower MI incidence than age-matched men (see *Figure 8*) [Tunstall-Pedoe, 1999]. However, the differences in CHD incidence between genders diminish gradually, with the rate among women increasing in later decades to the point where it approaches that among men (see *Figure 9*) [Evangelista, 2009; Healthy, 2009]. While cumulative incidence rates for men and women tend to converge, they never actually cross [Marrugat, 2004]. The mechanism that underlies this difference in rates of incidence between males and females is not understood.

Differences in cardiovascular risk factor profile between sexes

It has been suggested that differences in the prevalence of CVRFs could drive the differences observed for CHD incidence between sexes [Barrett-Connor, 1997] (see *Table*



The average incidence and 95% confidence intervals of the populations shown are indicated by a black diamond and a horizontal line, respectively.

Figure 8. Cumulative incidence of coronary events per 100,000 individuals aged 35-64yrs in the WHO MONICA study (1985-94) [Tunstall-Pedoe, 1999].

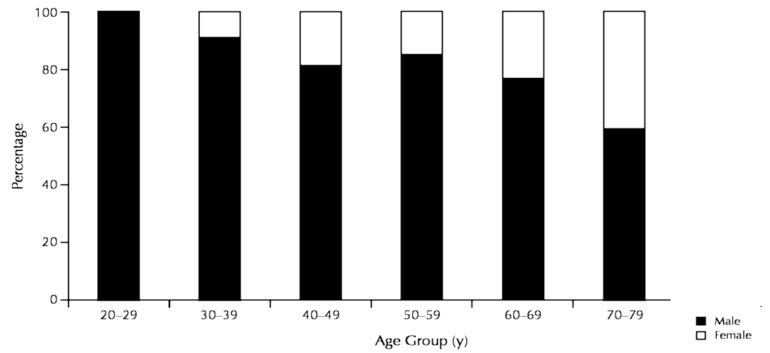


Figure 9. Sex distribution of mortality due to CHD, stratified by age group [Evangelista, 2009; Healthy, 2009].

4). Although men have generally less favourable CVRF profiles than women [Fowkes, 1994; Grau, 2011], the most important CVRFs do not appear to explain the gender gap entirely [Wingard, 1983; Isles, 1992].

Table 4. Prevalence of cardiovascular risk factors in males and females in Spain (2000-2010; pooled analysis with individual data from 11 population-based studies: The DARIOS study). Based on epidemiological data from Grau *et al.* [2011].

	Males	Females	P-value
N	13,425	15,462	
Age	53.81 [42,64]	53.37 [42,64]	0.9558
Diabetes (%)	16 [14,18]	11 [9,13]	0.0005
Hypertension (%)	47 [42,51]	39 [34,43]	0.0137
SBP, mmHg	131 [128,133]	122 [121,123]	6.1x10^{-7}
DBP, mmHg	79 [78,80]	75[74,77]	1.4x10^{-5}
Overweight (BMI [25,30)) (%)	51 [49,52]	36 [34,48]	4.0x10^{-5}
Obesity (BMI ≥30) (%)	29 [26,32]	29 [25,34]	1
Smoking (%)	33 [32,35]	21 [18,24]	2.3x10^{-12}
LDL colesterol (mg/dL)	140 [137,144]	138 [134,141]	0.4284
HDL cholesterol (mg/dL)	49 [48,50]	58 [56,59]	2.6x10^{-21}
Triglycerides (mg/dL)	142 [135,149]	108 [102,114]	4.9x10^{-13}

DBP: diastolic blood pressure; SBP: systolic blood pressure. Values are presented in the units shown [95% Confidence interval].

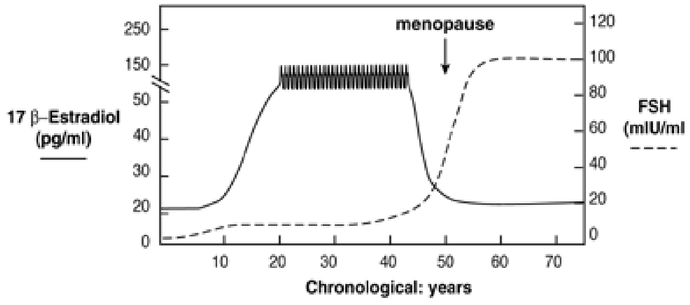
Reproductive steroid hormones as a candidate system to explain gender differences in CHD risk

The exact mechanism through which women are protected from CHD is still largely unknown, but data from epidemiological and observational studies and the fact that CHD risk in women after menopause approaches that of males suggests that elements of the sex steroid hormone system could be involved in the variation in CHD risk [Grodstein, 1997; Mendelsohn, 2005]. The idea that the endogenous oestrogen system is cardio-protective has been reinforced by the observation that women who experience early menopause also show increased CHD risk [Barrett-Connor, 1997]. This hypothesis was also initially supported by the results of observational studies showing lower CHD risk among postmenopausal women undergoing hormone replacement therapy (HRT) [Grodstein, 1997; Barrett-Connor, 1998; Varas-Lorenzo, 2000]. Moreover, differences in hormonal levels are associated with changes in several CVRFs, such as increasing LDL cholesterol levels, decreasing HDL cholesterol levels [Kuller, 1994], thereby increasing the incidence of CHD events in postmenopausal women. Nonetheless, as mentioned above, the changes observed in CVRFs do not entirely explain the differences observed in CHD incidence between genders.

Lifetime variation in sex steroid hormone levels in women

Levels of oestrogen start to increase during puberty in girls as a result of low-amplitude nocturnal pulses of gonadotropin that raise serum oestradiol concentrations [Molina, 2004]. During menstrual cycles, oestradiol production varies cyclically. During the perimenopausal period, depletion of ovarian follicles leads to a steady decline in ovarian oestradiol production. During menopause itself, levels of oestrogen and other female hormones decline such that the hormone profile of postmenopausal women is characterised by lower levels of oestrogens and oestrone as the predominant oestrogen (see *Figure 10*) [Gruber, 2002; Sigelman, 2009]. Serum androgen levels decline steeply in the early reproductive years and do not vary as a consequence of natural menopause [Davison, 2005], therefore making it less likely that this hormone is involved in modulating CHD risk after menopause. For this reason, most research on the potential role of reproductive hormones on CHD risk has focused on oestrogen-related and not androgen-related molecules.

Figure 10. Schematic representation of 17β-oestradiol levels variation through a woman's life [Butts, 2009].



FSH: Basal follicle-stimulating hormone

Hormone replacement therapy and the effects of exogenous oestrogens on health: before and after the WHI trial

Hormone replacement therapy (HRT) is a medical treatment for the effects of surgical menopause, and peri- and, to a lesser extent, postmenopausal symptoms. HRT contains hormone supplements such as oestrogens, typically combined with progestagens (which serve as precursors for all other steroids and are also present in different phases of the menstrual cycle, see *Figure 11*).

Oestrogens have been used since the late 1930s to slow and prevent aging, to stop hot flashes, to avoid pregnancy or miscarriage and physical changes that women experience during menopause. The Food and Drug Administration of the United States (U.S. FDA) initially approved hormone treatment for hot flashes and other problems associated with menopause, but not for disease prevention [Boston Women's Health Book Collective, 2007]. Because exogenous ovarian steroid hormones have multiple target tissues, such as bone, endometrium, the vascular system, and breast tissue [Grady, 1992; Women's Health Initiative, 1998], during the 1960s and subsequent decades drug companies promoted and doctors prescribed hormones to women to prevent and treat an increasingly broad range of ailments associated with aging, from wrinkles to Alzheimer's disease, depression, and heart disease. In the late 1980s and 1990s several observational studies suggested that hormone treatment might improve women's quality of life and even protect them against CHD. A description of the effects of HRT use on cardiovascular traits is shown in *Table 5* [Taylor, 2011]. However, evidence from clinical trials on the use of exogenous oestrogens and their effects on women's health was limited, until publication of the results of the Women's Health Initiative in 2002 [Rossouw, 2002]. Initial clinical trials of HRT showed unexpected negative results [Hulley, 1998; Rossouw, 2002], including the finding that HRT use did not reduce the rate of CHD among women as had been observed in observational studies, but was actually associated with greater risk, not only of CHD, but also of other diseases [Rossouw, 2002; Taylor, 2011]. Currently, new clinical trials are underway to explore, among other questions, the early and late effects of exogenous 17β-oestradiol administration on the progression of subclinical atherosclerosis and cognitive decline in healthy postmenopausal women (ClinicalTrials.gov Identifier: NCT00114517). According to the Clinical Trials database, more than 900 trials regarding the effects of the oestradiol molecule on different diseases and conditions are currently underway.

Disease	Study (or study type)	Treatment	Effect
CHD	WHI: randomised trial [Rossouw, 2002]	E+P	↑ CHD risk *vs.* placebo*
	WHI: randomised trial [Rossouw, 2002]	E alone	~ CHD risk *vs.* placebo*
	WHI: nested case control [Bray, 2008]	E+P & E alone	↑ CHD risk *vs.* placebo when LDL/HDL baseline ratio >2,5mg/dl. Otherwise: ↓ CHD risk. It was suggested that baseline cardiovascular risk may modulate CHD outcome among women on hormone therapy
	Meta-analysis of 19 randomised trials [Salpeter, 2009]	E+P & E alone	↓ CHD risk *vs.* placebo in younger postmenopausal women. The analysis also demonstrated a cardiovascular benefit when MHT was initiated early*
Lipid profile	Association studies	OE vs. TDE	Oral regimens provide superior benefits on lipids, showing greater reductions in total cholesterol and LDL and increases in HDL than do transdermal preparations. However, clotting factors and triglycerides are raised to a greater degree by the use of oral preparations.

CHD: Coronary heart disease; E+P: conjugated equine oestrogens (CEE) + medroxyprogesterone acetate (MPA); E alone: conjugated equine oestrogens; MHT: Menopause hormone therapy; OE: oral oestrogen treatment; TDE: transdermal therapy; WHI: Women's Health Initiative.

* Prompted the proposal of the timing hypothesis: the results varied depending on the time of initiation of the therapy (protective when the therapy was begun less than a decade after menopause)

Table 5. Effects of the use of oestrogens on coronary heart disease and lipid profile, summarised from Taylor *et al.* [2011].

Possible explanations for the negative findings in HRT: should this therapy still be used?

Inconsistencies between the results of observational and experimental studies may be explained by *i)* the use of oral oestrogen treatment versus transdermal therapy (the administration route hypothesis [Canonico, 2007]); *ii)* the type and combination of oestrogens used, such as natural versus synthetic oestrogens, or oestrogen only versus oestrogen plus progestagens; or *iii)* the time from menopause to the initiation of HRT therapy (the timing hypothesis [Dubey, 2005; Salpeter, 2006; Rossouw, 2007; Harman, 2011]).

It is noteworthy that the use of HRT has known effects on other phenotypes, such as risk of cancer or stroke, and cognitive function, evidence that should be taken into account when prescribing HRT. The available evidence indicates that hormone therapy in younger postmenopausal women increases risk of breast cancer and pulmonary embolism, and reduces risk of cardiovascular events, colon cancer, and hip fracture [Grodstein, 2000; Beral, 2002; Nelson, 2002; Beral, 2003; Salpeter, 2006; Manson, 2007; Rossouw, 2007]. It has been postulated that the cardiovascular benefit is a result of a small absolute increase in stroke risk but a greater reduction in risk of coronary events [Grodstein, 2000; Bath, 2005; Rossouw, 2007]. Guidelines published after the first WHI report concluded that hormone therapy represented greater harm than benefits in women of all ages and should be used only for short durations in women with severe menopausal symptoms [U.S. Preventive Services Task Force, 2002; North American Menopause Society, 2003; Wathen, 2004]. However, in the publication of age-specific data from the WHI [Rossouw, 2007], it was concluded that the initiation of hormone therapy in younger women may in fact reduce cardiovascular morbidity and mortality, but no mention was made of an overall reduction in mortality [Pines, 2007; North American Menopause Society, 2007]. Therefore, HRT can not be considered a general therapy for the target symptoms, but rather requires the consideration of various factors.

2.1.2. Oestrogen physiology

Sources and types of oestrogens

Cholesterol is the precursor of the five major classes of steroid hormones, progestagens, glucocorticoids, mineralocorticoids, androgens, and oestrogens (see *Figure 11*). Oestrogens are made from androgens [Berg, 2002] in the ovaries and are rapidly delivered throughout the body [Goodsell, 2003]. Some of the molecules that form part of the oestrogen system are: oestrone (E1), which is derived from androstenedione; oestradiol (E2 or 17β-oestradiol), formed from testosterone; and oestriol, which is a less active metabolite derived from oestrone and oestradiol.

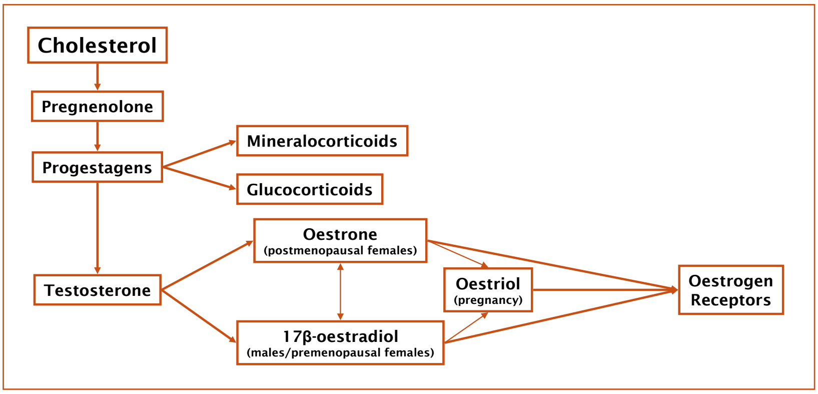


Figure 11. Overview of the steroid hormone biosynthesis pathway [Berg, 2002].

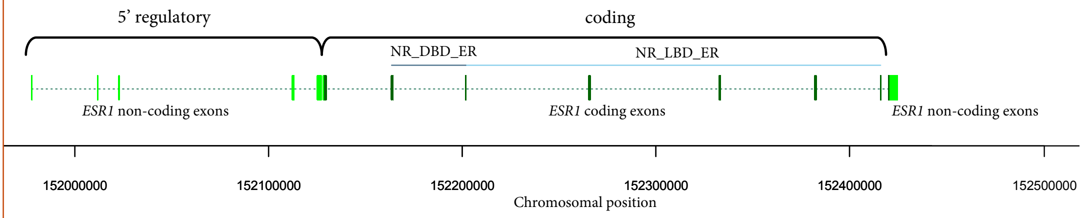
Oestrogen receptors: structure, transcripts and localisation

Oestrogen receptors are members of a large family of proteins that act as receptors for a wide range of hydrophobic molecules, including steroid hormones, thyroid hormones, and retinoids. They are ligand-activated transcription factors composed of several domains important for hormone binding, DNA binding, and activation of transcription (OMIM: 133430). While several oestrogen receptors have been described [Murphy, 2011], ERα, which is encoded by the *ESR1* gene (chromosome 6q25.1) is an important signalling gateway within this system and is expressed in multiple cardiovascular tissues in both males and females [Mendelsohn, 2005]. ERα is mainly expressed in reproductive tissues (breast, uterus, ovaries), cardiovascular tissues, liver and the central nervous system, whereas ERβ is expressed in tissues such as bone, lungs, endothelium, urogenital tract, the central nervous system, ovaries and prostate [Kuiper, 1997; Krege, 1998; Couse, 1999; Couse, 1999; Nilsson, 2001; Anderson, 2002; Palmieri, 2002]. Alternative splicing results in several transcript variants, which differ in their 5' UTRs and use different promoters (see *Figure 12*; extracted from: www.ncbi.nlm.nih.gov/gene/2099). Despite in vitro demonstrations of a possible role for some of the *ESR1* isoforms in hormonal sensitivity, the clinical significance of this evidence is uncertain [Balleine, 1999], and may also offer an explanation for discordant results between genetic association studies.

Molecular actions of oestrogens

It has been known since 1962 that oestrogen receptor(s) in oestrogen's target tissues capture circulating steroids and initiate the cascade of biochemical events associated with oestrogen action in that particular tissue [Jensen,

Figure 12. Structure of the Oestrogen Receptor 1 gene (*ESR1*).



NR_DBD_ER (dark blue line): DNA-binding domain of oestrogen receptors (ER); NR_LBD_ER (light blue line): Ligand binding domain of oestrogen receptor. Chromosomal position according to GRCh37.p5. Coding and non-coding exons are represented in dark and light green, respectively.

1962]. There are two known types of action associated with oestrogens: *i)* the classical pathway (slow response) [McDevitt, 2008] and *ii)* a fast cytoplasmic response [Boulware, 2005; Revankar, 2005]. The main receptors associated with the classical pathway are ERα and ERβ, and the main mediator of the fast cytoplasmic response is the G-protein coupled receptor (*GPER*, described in more detail below). For both mechanisms, the specific actions of oestrogens are mainly determined by the structure of the hormone, the isoform of the oestrogen receptor involved, the characteristics of the target gene promoter (in the case of the nuclear response), and the balance of co-activators and co-repressors that modulate the final response [Gruber, 2002].

It has been suggested that in mice, when there is presence of injured vessels, ERα but not ERβ mediates the beneficial effect of 17β-oestradiol on re-endothelialisation [Brouchet, 2001], that it inhibits smooth muscle cell proliferation and matrix deposition following vascular injury [Pare, 2002], it alters endothelial nitric oxide (NO) production [Pendaries, 2002], and attenuates atherosclerotic plaque progression [Hodgin, 2002; Egan, 2004]. Both ERα and ERβ are considered to exert their main effects via their role as transcription factors in mammals [White, 1987; Toran-Allerand, 2004].

2.1.3. Previous research into the *ESR1* gene and other oestrogen receptors

ESR1 discovery

ESR1 was first cloned in 1985 by Walter *et al.* [1985]. Using *in situ* hybridization, Gosden *et al.* [1986] localised the gene to 6q24-q27, by means of a cDNA probe containing the coding sequence for the oestrogen receptor. In 1988 Ponglikitmongkol *et al.* [1988] showed that the human *ESR1* gene is more than 140 kb long and contains 8 exons, and that the positions of its introns is remarkably similar to those of one of the chicken thyroid hormone receptor genes. In 2001, Koš *et al.* reviewed the organisation of the *ESR1* gene, describing the promoters used in the generation of *ESR1* transcripts in humans and other species. The possible role of multiple promoters in the differential expression of *ESR1* in tissues and during development was also discussed.

Other oestrogen receptors

For a long time, ERα was thought to be the only receptor for oestrogens, but in 1996 Mosselman *et al.* [1996] identified and characterised human ERβ, which has an overlapping but non-identical tissue distribution with ERα, and which is encoded by a gene (*ESR2*) that is homologous to the previously identified *ESR1*. The DNA-binding domain of ERβ is 96% conserved with respect to ERα, and the ligand-binding domain shows 58% conservation. In 1997, Enmark *et al.* [1997] mapped the *ESR2* gene to 14q22-q24 and identified its 8 exons. Between 1996 and 1997 the G protein-coupled oestrogen receptor (*GPER*) was also identified [Owman, 1996; Carmeci, 1997].

First signs of association between *ESR1* and CHD

In 1994 Smith *et al.* [1994] described a 28-year-old man with oestrogen resistance, characterised by continued linear growth into adulthood despite otherwise normal pubertal development and bone mineral density of the lumbar spine below the mean for age-matched normal men. Single-strand conformation polymorphism analysis followed by direct sequencing revealed a homozygous C-T transition at codon 157 in exon 2 of the *ESR1* gene, resulting in the introduction of a termination codon and truncation of the protein. From the subject's physiological characteristics, the authors concluded that oestrogen is important for bone maturation and mineralization in men as well as in women. Moreover, a second study of the same subject revealed evidence of early atherosclerosis in the left anterior descending coronary artery, leading to the hypothesis that the absence of functional oestrogen receptors may be a novel risk factor for coronary artery disease in men [Sudhir, 1997].

Studies of genetic variation in the *ESR1* gene

Initial genetic association studies (in the 1980s and 1990s) for most phenotypes focussed on the evaluation of one or few variants within a single gene, often those that could be assayed using a popular technique based on restriction enzymes, such as *EcoRI*, *BamHI* or *PvuII*. Variants discovered in this way were frequently evaluated for association in various diseases. This is the case for an *ESR1* SNP described in 1987 by Castagnoli *et al.* [1987] (dbSNP id rs2234693, often referred to by the name of a restriction enzyme widely used to genotype it, *PvuII*), and then studied in relation to cancer [Hill, 1989; Kjaergaard, 2007], bone mineral density and osteoporosis [Kobayashi, 1996; Ioannidis, 2002], Alzheimer's disease [Brandi, 1999], multiple sclerosis [Niino, 2000] and stroke [Shearman, 2005; Kjaergaard, 2007] among others. In relation to CHD, *ESR1* has been the subject of several candidate gene association studies over the past 15 years, primarily focussed on rs2234693, and with generally inconsistent results [Shearman, 2006; Kjaergaard, 2007; Lluís-Ganella, 2009]. Even by 2007 two meta-analyses of association studies including data on several thousands of individuals and centred on the role of rs2234693 in CHD risk had been published, but reported conflicting conclusions [Shearman, 2006; Kjaergaard, 2007]. However, from the ~7.200 variants described for the *ESR1* gene, less than 10 variants have been explored in relation to CHD and their role in CHD risk remains to be clarified.

By design, GWAS must use stringent criteria to determine which results represent statistically significant effects, in order to reduce the number of false positive results declared. However, some variants with moderate risk effects may not reach this stringency level. Therefore, under the hypothesis that some truly associated genetic variants have not shown statistically significant evidence of association with CHD at the genome-wide level, I sought to perform an in depth evaluation of the role of genetic variation in the *ESR1* gene in relation to CHD risk.

2.2. HYPOTHESES

i) The most widely studied variant in *ESR1* (rs2234693) is not associated with CHD risk. The inconsistency between studies is explained by aspects related to their quality.

ii) Some genetic variants located in the *ESR1* gene can modulate risk of CHD, but these variants have not been captured correctly in previous studies.

iii) The effects of putatively associated variants on CHD risk differ between men and women.

2.3. OBJECTIVES

i) To use recently published guidelines on the reporting and interpretation of genetic association studies to evaluate the quality of ours and other published studies as a possible explanation for the discordance between their reported results.

ii) To expand previously published meta-analyses of association studies focussed on the role of the *ESR1* rs2234693 variant in CHD risk, including new data from our population and results from other recently published association studies.

iii) To evaluate the effects on CHD risk of a broad range of both common and uncommon variation in a genomic region centred on *ESR1*.

iv) To search for differences between males and females in the effects of *ESR1* variation on CHD risk.

2.4. Article 1: Qualitative assessment of previous evidence and updated meta-analysis confirms lack of association between the *ESR1* rs2234693 (*PvuII*) variant and coronary heart disease in men and women. *Atherosclerosis 2009; 207(2): 480-486.*

2.5. Article 2:

Post-genomic update on a classial candidate gene for coronary artery disease: *ESR1*. *Circulation: Cardiovascular Genetics 2011; 4(6): 647-654.*

2.6. DISCUSSION

General overview

In *Part I* of the doctoral thesis, I have *i)* provided an explanation for the inconsistency between the results of studies that assessed the association between the *ESR1* rs2234693 variant and CHD risk [Lluís-Ganella, 2009], and *ii)* I have also provided well-powered evidence to contest the hypothesis that common and uncommon genetic variants in the primary sequence of the *ESR1* gene are associated with CHD [Lucas, 2011]. In these two studies I have improved the quantity and quality of evidence regarding this question by increasing both the number of individuals (~5.3 times more individuals: from 16,706 to ~87,000) and the number of genetic variants (~460 times more genetic variants) analysed, with respect to previous studies. Therefore, I can now make a fairly conclusive statement regarding the role of primary variation in this gene in modulating CHD risk.

Implications of the use of a quality assessment framework

In Lluís-Ganella *et al.* [2009], we performed a qualitative and quantitative evaluation of evidence regarding the roles of the rs2234693 variant in *ESR1* in modulating CHD risk. Two meta-analyses focussed on this question were published shortly before our paper [Shearman, 2006; Kjaergaard, 2007], with contradictory conclusions: Shearman *et al.* [2006] reported an association with MI of OR ~1.4 (P<0.0001) in more than 7,000 males (in 5 cohorts from 4 countries), whereas Kjærgaard *et al.* [2007] reported a lack of association between this variant and CHD using both a cohort study design (9244 participants followed up for 23-25 years) and a case–control study design (2495 CHD cases vs. 4447 controls). In addition to expanding the sample size of the meta-analysis to include >32,000 individuals, the use of a quality assessment framework to assess the quality of the previous studies and identify heterogeneity in their results was key to providing a solid justification for our negative conclusions. As reported by Haynes *et al.* [2009], the implementation of quality control checklists dramatically improved the quality and results of studies and, in our case, the use of these tools helped to explain the heterogeneity of the results observed in the studies that were included. Although these guidelines were not developed to evaluate study quality a posteriori, they generally reflect the quality of the work presented. Other measures, such as the phenotype definition, study design and sample size, were also evaluated in our study, but none of these factors alone was capable of explaining the between-study heterogeneity as well as the quality score metric we designed and implemented.

In addition, we must also consider the publication bias of negative studies, where studies with positive results are generally easier to publish than studies with negative results simply because the negative result can not be attributed with certainty to lack of a real effect or to a lack of statistical power, and this can clearly generate a bias in favour a positive report of association with disease.

Curtailing the reporting of spurious results through replication

In addition to recommendations on the conduct and reporting of genetic association studies in order to improve their reliability, evidence of replication of the reported results in at least one independent population is now highly valued in the research community. Specifically, this recommendation aims to address the widespread reporting of spurious results and/or different effects in specific populations. For example, in a review published by Hirschhorn *et al.* [2002] they showed that from 166 putative associations evaluated three or more times for complex diseases, only 6 were consistently replicated. The majority of the studies included in the meta-analysis described in *Article 1* [Lluís-Ganella, 2009] were not supported by evidence from replication cohorts; the only study that performed a replication found a negative result for association. By using data from multiple studies, the analyses reported in *Article 2* [Lucas, 2011] represent a mutual replication of our results across several populations, and we believe that these results therefore represent a reliable statement on the true role in CHD risk of genetic variation in *ESR1*.

For complex diseases in general, the efforts being made to replicate the results of genetic studies are contributing to the improved reliability of the results reported in the genetic epidemiology literature, and a smaller number of false positive results are currently being published.

No evidence of association between risk of CHD events and common & rare variants in *ESR1*

In Lucas *et al.* [2011], we found no evidence of association between genetic variants in the *ESR1* gene and CHD events, despite the fact that up to ~87,000 individuals were analysed. The effect sizes of associations between common variants and complex diseases are not expected to exceed those already discovered by GWAS (e.g. not bigger than those found between chromosome 9 variants and CHD risk), because these studies had high power to detect associations moderate effect sizes. In contrast, the expected effect sizes for rare variants range from weak (~1) to bigger effect sizes (>2), but because of the design of the current studies rare variants with moderate to large effects could still be missed. Using data for ~85,000 individuals from the CARDIoGRAM consortium, our analysis had high power to discover weak associations between common variants and CHD, but lower power to find very subtle effects associated with rare variants. Only rare variants are expected to have large effect on risk of complex diseases because selective pressure acts mainly against variants with strong effects (which are more likely to express themselves before the end of reproductive life), preventing them from increasing in frequency.

No evidence of association with CHD events: regulatory regions

As far as we are aware, our study is the first to examine genetic variation in the regulatory regions of this gene in detail. By extending our association analysis to the 5' regulatory region, we cover all the possible transcripts of this gene. However, if the effect of a genetic variant was in a promoter used in only few transcripts, and the transcripts were expressed at different stages of life, we would not have enough statistical power to declare a significant association. Considering this limitation, our study suggests that there are no variants conferring a different CHD risk in the regulatory regions of *ESR1*.

No evidence of association with CHD events: can other elements of the sex steroid hormone system be involved in CHD risk?

Our data suggest that common primary genetic variation in the *ESR1* gene, one of the main mediators of oestrogen response, does not explain the observed gender differences in CHD risk. As discussed in the introduction of Part I of this doctoral thesis, the sex steroid hormone system is a complex network of many molecules, and therefore, if this system is responsible for the observed gender differences in CHD incidence, these differences could be driven by any of a wide range of elements. Moreover, some molecules in this system can interact with ERα, and other receptors from the intracellular receptor superfamily can act as receptors for the oestrogen molecule. Domain-swap experiments suggest that many of the hormone-binding, transcription-activating, and DNA-binding domains in these receptors can function as interchangeable modules [Alberts, 2008; Bonduriansky, 2009], and therefore elements not identified as part of the sex steroid hormone system could also be implicated in the effects of oestrogen on cell signalling, and could also compensate for minor disruptions of in the main receptors.

Could primary genetic variation cause gender differences in CHD incidence?

Mendel's First Law (equal segregation of traits in males and females) reflects the fact that autosomal loci are in linkage equilibrium with the sex determining locus on the Y-chromosome. This leads to the conclusion that observed differences in CHD risk between genders cannot be directly due to primary autosomal genetic variation. Therefore, the observed differences in CHD risk between sexes could be explained by primary genetic variation, via two mechanisms: *i)* primary genetic variation on the sex chromosomes, or *ii)* interactions between autosomal variants and some other factor that differs between males and females. A recent study that analysed the association between different lineages of the Y chromosome and CHD [Charchar, 2012] showed that men who inherit haplogroup I (one of the most common types of Y chromosome in Europe) from their fathers have a roughly 50% higher risk of CHD than men with from other haplogroups, independently of known classical cardiovascular risk factors. The authors also showed that this effect on risk of CHD is most likely mediated through immune response.

Could other types of variation cause gender differences in CHD incidence?

Epigenetic variation, differences in protein expression and/or function, lifestyle, environmental variation or any other factor that differs between males and females (e.g. emotional factors), could contribute to the observed differences in CHD incidence between males and females [Barrett-Connor, 1997]. Such sex differences could cause molecular changes that may persist through subsequent cell divisions [Chong, 2004] for the remainder of the cell's life and may also last for multiple generations. It is possible that these molecular changes could be generated differently in each sex, or in genes that are differentially expressed between genders, therefore explaining CHD incidence differences.

Further research

The following areas of further research are a priority to determine the possible differences in cardiovascular disease incidence between sexes:

i) Although much less effort has been invested in exploring the rest of the elements of the reproductive hormone system, it is essential to give conclusive evidence to all of the elements of this system. In addition to the strategies we have performed to explore the role of ESR1 in CHD, also other types of variation, such as epigenetic variation or protein expression levels, have to be explored on all the elements of this system.

ii) Recent evidence suggests that the human Y chromosome is associated with risk of coronary artery disease in men despite the small number of genes it harbours. Functional experiments, sequencing and tests for interaction between genetic variants on this chromosome and variants in other regions of the genome could help to understand the pathways that are relevant for CHD in males, which could guide efforts to better understand the gender differences.

3. PART II: Table of contents

Improvement of cardiovascular risk assessment using genetic information

PART II

3

3.0. ABSTRACT

In *Part II* of this doctoral thesis, I evaluated the role of a genetic risk score, composed of genetic variants robustly shown to be associated with CHD independently of CVRFs, in modulating CHD risk. I also evaluated whether the capacity of the Framingham risk function to predict 10-year CHD risk was improved by the addition of the genetic risk score. To achieve this goal, I have followed the first three stages of the American Heart Association's "criteria for the evaluation of novel markers of cardiovascular risk" [Hlatky, 2009], which include *i)* an initial demonstration of association between the marker, in this case a genetic score, and increased risk of CHD (proof of concept), *ii)* validation of this relationship in prospective cohort studies, and *iii)* assessment of the incremental value of using this genetic risk score in combination with the classical cardiovascular screening tools to improve the estimation of 10-year CHD risk.

Our results provide an indication of the potential utility of genetic information in improving the efficiency of classical cardiovascular risk functions. They also suggest that these markers may be most informative in individuals with intermediate CHD risk, precisely the group in which most CHD events occur, and where there is greatest need for improved stratification of CHD risk.

3.1. INTRODUCTION

3.1.1. Cardiovascular risk factors, epidemiology and risk functions

Cardiovascular prevention strategies

One of the main goals of preventive medicine is to reduce the incidence of disease in a specific group of people [Evans, 1997], but the best strategy for doing this is not clear [Emberson, 2004]. In the case of CVD, this goal is of special importance for the following reasons: *i)* CVD is the major cause of premature death in Europe; *ii)* the process of atherosclerosis that underlies CVD develops insidiously over many years and is usually advanced by the time symptoms occur; *iii)* death from CVD often occurs suddenly and before medical action can be taken; *iv)* the majority of CVD cases are strongly related to lifestyle and to modifiable physiological and biochemical factors; *v)* modification of a person's risk factor profile has been shown to reduce CVD morbidity and mortality, particularly in high risk patients [Graham, 2007].

Three complementary strategies can be used for CVD prevention: *i)* population intervention, which is mainly based on the promotion of healthy lifestyles, improving eating habits and anti-smoking legislation; *ii)* screening methods, which are used to identify and intervene in high risk individuals; and *iii)* secondary prevention, which aim to diminish the total cardiovascular risk of patients with established cardiovascular organ damage or disease (not addressed in this doctoral thesis).

Population interventions

The main goal of this type of intervention is to improve population-wide risk factor profiles (see *Table 6*), and is mostly achieved by developing public health policies and community interventions. While difficult to implement at the population level, some such interventions, such as public smoking bans [Haw, 2006] or salt reduction initiatives (e.g. www.food.gov.uk/multimedia/pdfs/saltreductioninitiatives.pdf), aim to reduce exposure to some risk factor, with immediate consequences for health at the population level. Other strategies, such as those whose aim is to improve diet, encourage physical activity or reduce excess alcohol consumption, produce more subtle effects that may take some time to emerge [Craig, 2012]. In the case of CVD, this attempt to modify population-wide risk factor profiles responds to evidence suggesting that most CVD events are preventable [Stamler, 1999; Rosengren, 2001] and their risk factors are modifiable [Pearson, 2002]. For example, data from the Nurses Health Study [Stampfer, 2000] suggest that maintaining a desirable body weight, eating a healthy diet, exercising regularly, not smoking, and consuming a moderate amount of alcohol could produce an 84% reduction in CHD risk in women.

Preventive intervention	Primary Goal
Smoking	No smoking
BP control	<140/90 mm Hg
Dietary intake	Healthy food choices
Blood lipid management	LDL-C <115 mg/dL; total cholesterol <190mg/dL; blood glucose <110mg/dL
Physical activity	30 min of moderate physical activity a day
Weight management	body mass index <25 kg/m^2 and avoidance of central obesity

BP indicates blood pressure; LDL-C, low-density lipoprotein cholesterol.

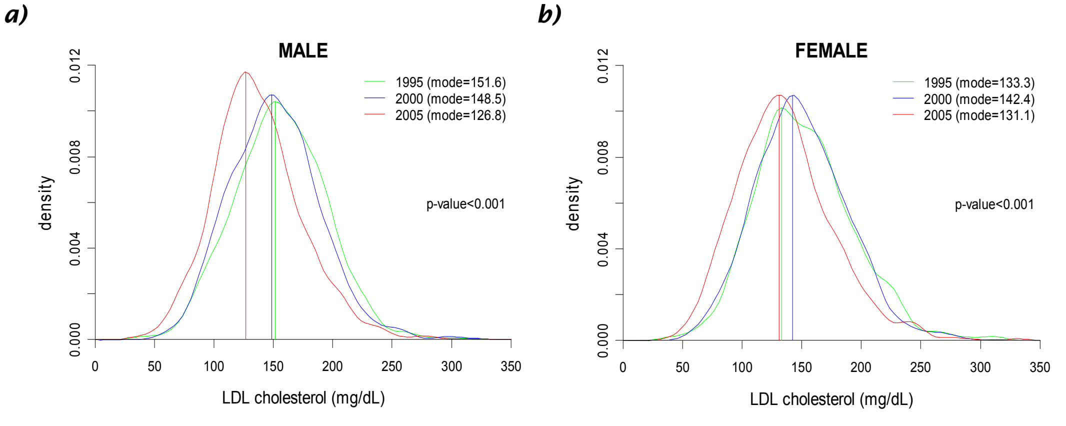
Table 6. Goals for primary prevention of cardiovascular disease [Graham, 2007].

Therefore, the goal of this type of strategy would ideally be, for example, to cause a shift in the distribution of a cardiovascular risk factor (CVRF) in the general population towards a better profile, such as that reported by Grau *et al.* [2007] for LDL-cholesterol levels in the REGICOR cohorts in Girona, Catalonia, Spain (see *Figure 13*).

Screening methods

Since the first clinical symptom of CHD is often catastrophic (MI or sudden death), there is considerable interest in improving diagnosis in asymptomatic individuals. Although mass screening would be the ideal way to detect early stages of CVD, there is no evidence that this strategy would be a cost-effective way to prevent disease [Graham, 2007], so screening must be limited to a subset of the population. To do so, two main types of screening are performed: *i)* opportunistic screening or *ii)* high risk screening. In opportunistic screening, evaluation of CVRFs and estimation

Figure 13. Distribution of LDL cholesterol levels in the REGICOR study, 1995-2000-2005 (from Grau *et al.* [2007].



a) males; ***b)*** females

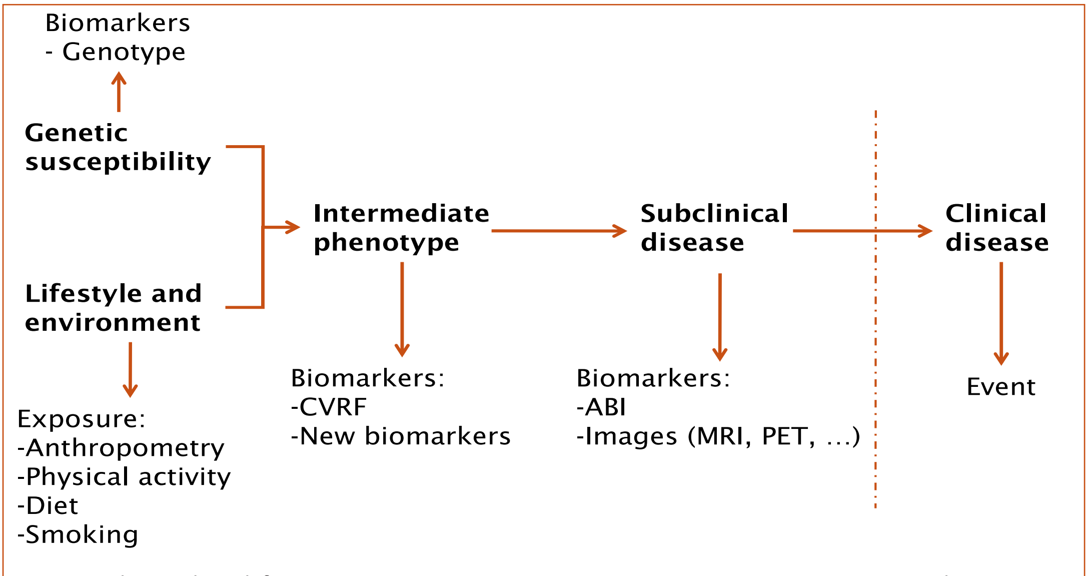
of CVD risk is carried out in all individuals that come into contact with the health system for any cause (including work-related medical examinations). High risk screening is limited to individuals with an elevated probability of suffering a CVD event because they have a family history of early-onset CHD events or familial hypercholesterolemia, or because they present other risk factors, such as renal dysfunction. A summary of the screening strategies that can be implemented at various stages of life or of disease progression is shown in *Figure 14*.

In both types of screening, risk functions are the most commonly used method for evaluating individual risk of having a CHD/CVD event, usually computed for a 10-year time period [Wilson, 1998; Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults, 2001; Graham, 2007]. Risk functions are mathematical equations that estimate the probability of developing CHD/CVD using information about CVRFs that are strongly and independently related to CHD and can be evaluated in simple office procedures and laboratory results [Anderson, 1991].

History of risk functions

An historical summary of the development of the various cardiovascular risk functions currently in use is presented

Figure 14. Screening strategies for atherosclerosis.



CVRF: cardiovascular risk factor; MRI: Magnetic resonance imaging; PET: positron emission tomography.

in *Figure 15*. The first function was developed by the Framingham Heart Study [Truett, 1967] and adopted by the American Heart Association in 1973 in the form of a handbook [American Heart Association, 1973] containing (6-year) CHD risk tables based on the Framingham function [Gordon, 1971]. New CHD risk functions were published by the Framingham investigators in 1991 [Anderson, 1991] and 1998 [Wilson, 1998]. The Wilson function is widely used in clinical practice and has been successfully adapted to and calibrated for different populations [D'Agostino, 2001; Marrugat, 2003a; Liu, 2004]. The Framingham investigators have also developed new functions to estimate 10-year [D'Agostino, 2008] and lifetime [Lloyd-Jones, 2006] global cardiovascular risk, as well as risk of specific cardiovascular events, such as cerebrovascular disease [Seshadri, 2006], atrial fibrillation [Lloyd-Jones, 2004a], peripheral artery disease [Murabito, 1997], and others.

Figure 15. Historical summary of the development of cardiovascular risk functions and their adaptations, highlighting the predictor variables on which they are based.

In parallel, a number of other risk functions have been developed and are in use in different clinical settings:

- The SCORE function [Conroy, 2003], which measures 10-year risk of fatal CVD, is recommended by the European Society of Cardiology and other European Scientific Societies and has been calibrated for use in Spain [Sans, 2007].
- The Reynolds functions measure cardiovascular risk separately in women [Ridker, 2007] and men [Ridker, 2008].
- The PROCAM function uses a scoring system to calculate coronary risk in men [Assmann, 2002].
- The QRISK funcions have recently been developed in the UK to estimate 10-year [Hippisley-Cox, 2007] and lifetime [Hippisley-Cox, 2010] risk of cardiovascular disease.

The Framingham risk function and its adaptations

The Framingham risk function [Wilson, 1998] and its adaptations [D'Agostino, 2001; Marrugat, 2003a; Liu, 2004] estimate individual 10-year risk of presenting a MI or coronary death within the following 10 years, and are based on the incidence of CHD in the population (1-S), individual CVRF profile, the population means the CVRFs ($\overline{CVRF}_p$),

		Age	*Gender*	*Smoking*	*Diabetes*	*SBP*	*DBP*	*HDL-cholesterol*	*LDL-cholesterol*	*Total cholesterol*	*Obesity/BMI*	*Exercice*	*Stress*	*Left ventricular hypertrophy*	*HTN treatment*	*Family history of premature CAD*	*Townsend deprivation score**	*High-sensitivity C-reactive protein*	*Hemoglobin A_{1c}*	*Triglicerides*
2009	Pencina MJ, *et al.* Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation.	■	■	■	■	■	□	■	□	■	■	□	□	□	□	□	□	□	□	□
2008	D'Agostino RB Sr, *et al.* General Cardiovascular Risk Profile for Use in Primary Care: the Framingham Heart Study. Circulation.	■	■	■	■	■	□	■	□	■	□	□	□	□	□	□	□	□	□	□
2007	Ridker PM, *et al.* Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA.	■	■	■	■	■	□	■	□	■	□	□	□	□	□	■	□	■	■	□
	Hippisley-Cox J, *et al.* Derivation and Validation of QRISK, a New Cardiovascular Disease Risk Score for the United Kingdom: Prospective Open Cohort Study. BMJ	■	■	■	□	■	□	■	□	■	■	□	□	□	■	■	■	□	□	□
	Marrugat J, *et al.* Validity of an Adaptation of the Framingham Cardiovascular Risk Function: the VERIFICA Study. J Epidemiol Comm Health.	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□
	Sans S, *et al.* Calibrating the SCORE cardiovascular risk chart for use in Spain. Rev Esp Cardiol.	■	■	■	□	■	□	□	□	■	□	□	□	□	□	□	□	□	□	□
2006	Lloyd-Jones DM, *et al.* Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation.	■	■	■	■	■	□	■	□	■	■	□	□	□	□	□	□	□	□	□
2004	Liu J, *et al.* Predictive Value for the Chinese Population of the Framingham CHD Risk Assessment Tool Compared With the Chinese Multi-Provincial Cohort Study. JAMA.	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□
2003	Marrugat J, *et al.* An Adaptation of the Framingham Coronary Heart Disease Risk Function to European Mediterranean Areas. J Epidemiol Comm Health.	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□
	Conroy RM, *et al.* Estimation of Ten-Year Risk of Fatal Cardiovascular Disease in Europe: the SCORE Project. Eur Heart J.	■	■	■	□	■	□	■	□	■	□	□	□	□	□	□	□	□	□	□
2002	Assmann G, *et al.* Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation.	■	■	■	■	■	□	■	■	□	□	□	□	□	□	■	□	□	□	■
2001	D'Agostino RB Sr, *et al.* Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. JAMA.	■	■	■	■	■	■	■	□	■	□	□	□	□	□	□	□	□	□	□
1998	Wilson PW, *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation.	■	■	■	■	■	■	■	■	■	□	□	□	□	□	□	□	□	□	□
1991	Anderson KM, *et al.* Cardiovascular Disease Risk Profiles. Am Heart J.	■	■	■	■	■	■	■	□	■	□	□	□	■	□	□	□	□	□	□
1973	American Heart Association. Coronary Risk Handbook: Estimating Risk of Coronary Heart Disease in Daily Practice.	■	■	■	■	■	□	□	□	■	■	■	■	■	□	□	□	□	□	□

* The Townsend deprivation index is a simple, census-based index of material deprivation, which is calculated from a combination of four census variables: *i)* percentage of households without a car; *ii)* percentage of overcrowded households; *iii)* percentage of households not owner-occupied; and *iv)* percentage of persons unemployed. The specific variables used vary slightly between censuses [www.geog.soton.ac.uk].

and the magnitude of the effect of each risk factor on CHD risk (β_{CVRF}) (see *Equation 1*).

Equation 1. The Framingham coronary heart disease risk function [D'Agostino, 2001].

$$\text{prob}(\text{event}\,|\,\text{CVRF}_i) = 1 - \widehat{S}^{\exp\left[\sum_{p=1}^{P}\beta_{CVRF_p}\cdot CVRF_{p,i} - \sum_{p=1}^{P}\beta_{CVRF_p}\cdot\overline{CVRF_P}\right]}$$

Where:

- prob(event|CVRF$_i$): Individual probability of having a coronary event, given individual CVRF profile;
- $\widehat{S}$: survival value for the population average.
- exp: exponential value (or anti-logarithm function).

- β_{CVRFp}: logarithm of the hazard ratios of each cardiovascular risk factor (see *Table 7*).
- *CVRF*$_{p,i}$: value of each cardiovascular risk factor for individual i.
- $\overline{CVRF}_p$: population mean of each cardiovascular risk factor.

Table 7. Logarithm of the hazard ratios or β coefficients of CVRF used in the risk functions in this thesis.

Variables		β for Males	β for Females	Variables		β for Males	β for Females
Age (years)		0.0483	0.3377	Total cholesterol (mg/dl)	<160	-0.6595	-0.2614
Age^2 (years)		0.0000	-0.0027		160 - <200	0.0000	0.0000
Diabetes		0.4284	0.5963		200 - <240	0.1769	0.2077
Smoking		0.5234	0.2925		240 - <280	0.5054	0.2439
HDL cholesterol (mg/dl)	<35	0.4974	0.8431		>=280	0.6571	0.5351
	≥35; <45	0.2431	0.3780	Blood pressure (mmHg) (SBP; DBP)	Optimal (<120; <80)	-0.0023	-0.5336
	≥45; <50	0.0000	0.1979		Normal (120-130; 80-85)	0.0000	0.0000
	≥50; <60	-0.0511	0.0000		High (130-140; 85-90)	0.2832	-0.0677
	≥60	-0.4866	-0.4295		HTN grade I (140-160; 90-100)	0.5217	0.2629
					HTN grade II (>160; >100)	0.6186	0.4657

The adaptation and calibration of the risk function in different populations is based on substituting the Framingham CHD survival rate and risk factor prevalence for their values in the target population [D'Agostino, 2001]. This adaptation has been carried out and validated in different populations [D'Agostino, 2001; Liu, 2004], including Spain (the REGICOR adaptation to the Framingham risk function) [Marrugat, 2003a].

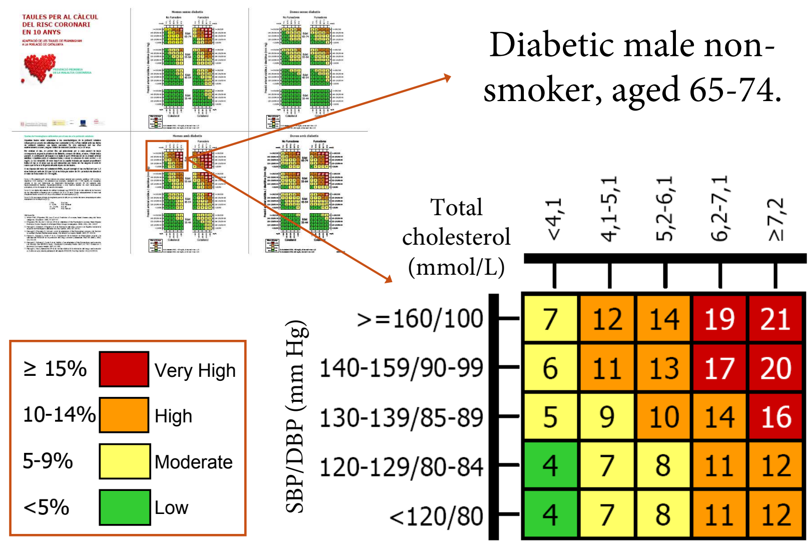
The application of this tool to clinical practice is simplified by the use of tables (see *Figure 16*) [Marrugat, 2003b; Marrugat, 2011], such that, for example, a 70-year-old diabetic male who has never smoked and whose blood pressure is 135/87 mmHg and total cholesterol is 6 mmol/L would be classified as having a high risk of suffering a CHD event within the next 10 years, with an estimated probability of ~10%.

3.1.2. Improving the precision of risk functions by adding new risk markers

Risk classification

Risk functions provide an estimate of individual probability of having a cardiovascular event, and this value can also be used to classify individuals into different risk categories.

Figure 16.
Implementation of the REGICOR adaptation to the Framingham risk function using coronary heart disease risk tables.


TAULES PER AL CÀLCUL DEL RISC CORONARI EN 10 ANYS

Diabetic male non-smoker, aged 65-74.

SBP/DBP (mm Hg) \ Total cholesterol (mmol/L)	<4,1	4,1-5,1	5,2-6,1	6,2-7,1	≥7,2
>=160/100	7	12	14	19	21
140-159/90-99	6	11	13	17	20
130-139/85-89	5	9	10	14	16
120-129/80-84	4	7	8	11	12
<120/80	4	7	8	11	12

≥ 15%	🟥	Very High
10-14%	🟧	High
5-9%	🟨	Moderate
<5%	🟩	Low

DBP: diastolic blood pressure; SBP: Systolic blood pressure.

These categories are used to define the intensity of cardiovascular risk intervention measures, which may range from lifestyle recommendations to drug prescription and periodic follow-up. The cut-points used to define risk categories vary depending on the population. In Spain, the cut points used to define categories of low, intermediate-low, intermediate-high and very high CHD risk are 5%, 10% and 15%, whereas in USA these cut points are 10%, 15% and 20%, respectively.

Population distribution of CHD incidence and risk, and the limitations of risk functions

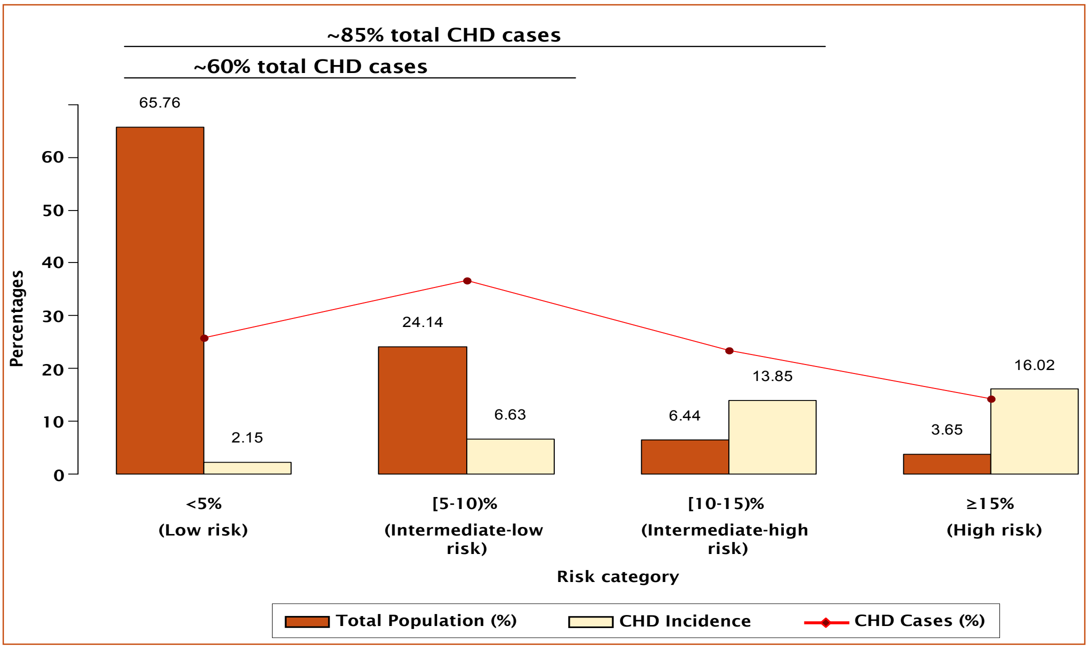
The distribution of CHD risk as estimated by risk functions fits quite well with that of the observed incidence of CHD at the population level (calibration), in that a higher percentage of individuals who are estimated to have high risk go on to have an event than those who are estimated to have lower risk. However, one of the main problems of risk functions is their low sensitivity, which is a function of how the population is distributed between risk categories (*Figure 17*). As an example, >80% of the REGICOR population is classified as having low or intermediate-low risk, but these groups account for 49% of the CHD events observed in the population (*Figure 17*).

Therefore, risk functions quite accurately predict the number of events that will occur in each of the risk categories, but the majority of the CHD events ultimately occur in individuals who are not estimated to have a high enough risk to be subjected to intensive treatment, Therefore, efforts to improve the correct classification of these individuals into higher risk categories is a priority for research and public health, and one of the possible strategies for achieving this improved sensitivity is the inclusion of new biomarkers in classical risk functions.

Evaluating the value of new risk markers

Recently, the American Heart Association (AHA) has published guidelines for the step-wise evaluation of new risk biomarkers and their subsequent application in clinical practice [Hlatky, 2009] (summarised in *Table 8*). The steps suggested by these guidelines require the use of different study designs (e.g. case-controls, cohorts, randomised trials), and evaluation is carried out using metrics that are specific to each stage. An important part of this process is that which evaluates the incremental value of the biomarker of interest in the predictive capacity of the risk function (*Table 8, Step 3*). To this end, various aspects of the function's performance can be evaluated [Steyerberg, 2011]: calibration (of how the expected risk adjusts to

Figure 17. Ten-year incidence of CHD in the REGICOR cohort, and distribution of the population into risk categories defined by the REGICOR adaptation to the Framingham risk function (n~3,800) [Marrugat, 2011].



the observed incidence; Hosmer-Lemeshow, Brier Score), discrimination (between events and non-events; c-statistic, c-statistic improvement) and reclassification (among higher or lower risk categories; Net Reclassification Improvement, NRI and Integrated Discrimination Improvement, IDI) (see *Box 4*).

Steps	Design	Statistical Metric
1. Proof of concept: Do novel marker levels differ between subjects with and without the outcome?	Case-control	OR
2. Prospective validation: Does the novel marker predict the development of future outcomes in a prospective cohort or nested case-cohort/ case-cohort study?	Case-cohort Cohort	RR
3. Incremental value: Does the novel marker add predictive information to established, standard risk markers?	Case-cohort Cohort	Discrimination Calibration Reclassification
4. Clinical utility: Does the novel risk marker change predicted risk sufficiently to change recommended therapy?	Case-cohort Cohort	Net Benefit
5. Clinical outcomes: Does use of the novel risk marker improve clinical outcomes, especially when tested in a randomised clinical trial?	Clinical trial	RR
6. Cost-effectiveness: Does use of the marker improve clinical outcomes sufficiently to justify the additional costs of testing and treatment?	Cost-effectiveness analyses	Cost per QALY

OR: Odds Ratio; RR: Relative Risk; QALY: Quality-adjusted life year (which is a measure of disease burden, including both the quality and the quantity of life lived).

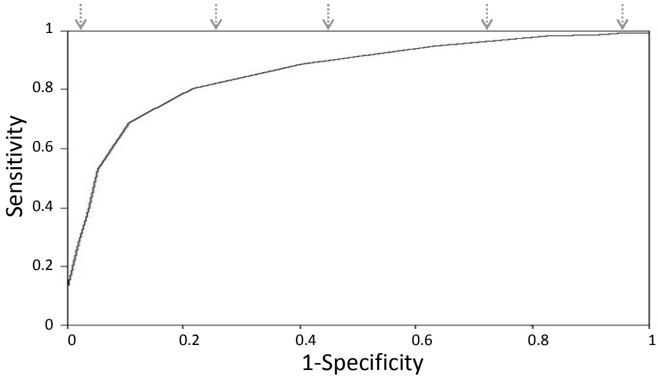
Table 8. Summary of the American Heart Association guidelines for evaluating the utility of new biomarkers [Hlatky, 2009], and the study designs and statistical metrics used in each step.

Box 4. Metrics used for the evaluation of novel biomarkers. Extracted from Steyerberg *et al.* [2011].

Evaluated Aspect	Measure	Characteristics
Evaluation of predictions performed		
Discrimination	AUC or *c*-statistic	AUC or the *c*-statistic is a rank-ordered statistic, which is interpreted as the probability of correct classification of a pair of patients with and without the outcome
Calibration	Intercept and slope of a calibration model	The intercept (*a*|*b*=*1*) reflects the level of calibration in general, or the difference between average prediction and average outcome The slope (*b*) reflects the average effect of predictors on the outcome
Evaluation of classifications		
Classification	Youden index	Sum of sensitivity and specificity-1, which represents the maximum vertical distance between the ROC curve and the diagonal line, which represents randomness [Schisterman, 2005]
Clinical usefulness	NB and DCA	Net fraction of true positives gained by making decisions based on predictions at a single threshold (NB) or over a range of thresholds (DCA)
Evaluation of incremental value by a marker		
Increase in discrimination	Delta AUC	Increase in discrimination is usually a modest number
Reclassification	NRI	Net fraction of reclassifications in the right direction obtained by making decisions based on predictions that take marker data into account, compared to decisions without the marker
Clinical usefulness	Difference in NB and DCA; Weighted NRI	Net fraction of true positives gained by making decisions based on predictions that take marker data into account, compared to decisions without marker data at a single threshold (NB) or over a range of thresholds (DCA); this technique weights the results according to the consequences of the decisions taken (NB and weighted NRI).

AUC, area under the ROC curve; DCA, decision curve analysis; NB, net benefit; NRI, net reclassification index; ROC, receiver operating characteristic.

AUC or c-statistic: This measure is a numerical value representing the area under the Receiver Operator Curve (ROC), which is a plot of the **sensitivity** (computed as: TP / (TP + FN)) on the *y-axis* for each of a series of values of **1-specificity** (computed as: FP / (FP + TN) on the *x-axis*. In the case of CHD, the AUC of ~0.8 is obtained using the REGICOR risk function with classical cardiovascular risk factors only [Marrugat, 2011]. Some authors are concerned with the c-statistic as the main discrimination metric when the goal in clinical practice is mainly to stratify individuals into risk categories in order to decide the intensity of preventive measures to apply, as is the case for cardiovascular prevention [Cook, 2007].



Sensitivity:

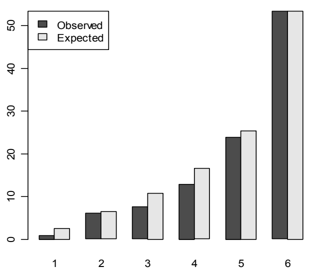
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Specificity:

$$\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$$

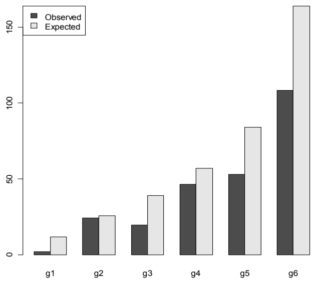
Calibration: While Steyerberg *et al.* [2011] report the intercept and slope of their recalibration model as a means to assess the calibration of the risk function, we used a version of the Hosmer-Lemeshow test [D'Agostino, 2003]. This test assesses whether or not the **observed event rates match expected event rates** in subgroups of the population. The following figure shows an example of the calibration of two models, one of which is well calibrated (model 1), and the other poorly calibrated (model 2).

Model 1



Chi-square = 3.00 (df = 4), p-value = 0.557

Model 2



Chi-square = 60.38 (df = 4), p-value <0.001

Net Reclassification Improvement: (NRI) [Chambless, 2010; Pencina, 2011] This metric is used **to compare two risk functions** (e.g. functions with and without the factor being evaluated) in terms of how well they classify individuals into different risk categories. All individuals are classified using each functions (see figure), and then the NRI is computed by subtracting the number of individuals that are better classified (cases reclassified into higher risk categories and non-cases reclassified into lower risk categories; black arrows) minus the number of individuals that are more inappropriately classified (cases reclassified into lower risk categories and non-cases reclassified into higher risk categories; brown arrows).

RISK FUNCTION +
ADDITIONAL FACTOR

RISK FUNCTION

	Low risk	Intermediate-low risk	Intermediate-high risk	High risk	
CASES					
Low risk	15 →	0	0	0	15
Intermediate-low risk	0	← 33 →	0	0	33
Intermediate-high risk	0	0	← 32 →	0	32
High risk	0	0	0	← 20	20
	15	33	32	20	
NON CASES					
Low risk	1700 →	0	0	0	1700
Intermediate-low risk	0	← 400 →	0	0	400
Intermediate-high risk	0	0	← 100 →	0	100
High risk	0	0	0	← 60	60
	1700	400	100	60	

Integrated Discrimination Improvement: (IDI) [Chambless, 2010] As for the NRI, the IDI is used to compare changes in risk when using two functions in the same individuals. The IDI considers the change in the estimated prediction probabilities **as a continuous variable**, and can be seen as continuous version of NRI with probability differences used instead of categories.

3.1.3. Genetic variants as novel biomarker of cardiovascular risk

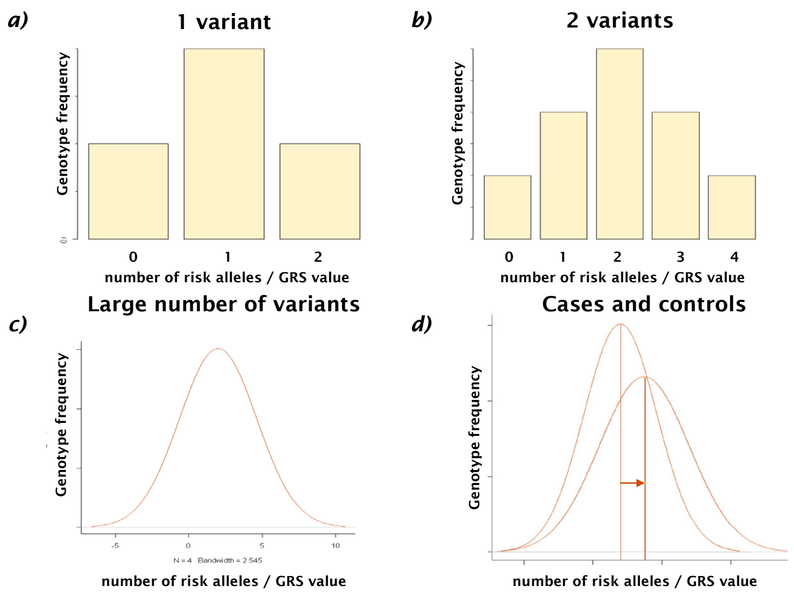
Use of genetic markers

A multitude of biomarkers are currently being evaluated for their capacity to improve the predictive capacity of the cardiovascular risk functions, including genetic variants [Wang, 2011]. The most important disadvantage of using genetic variants information for risk assessment is the modest effects of individual variants on CHD, which range from an OR of 1.06 to 1.51 for known variants. To address this problem, the use of genetic risk scores (GRS) to capture the additive effects of multiple variants has been proposed, thereby summarising the information for all CHD-related genetic variation carried by an individual in a single value.

Computation, behaviour and inclusion of genetic risk scores in the risk functions

GRSs are usually expressed as the number of alleles known to increase disease risk that are carried by an individual. For example, for risk alleles that are independent both within (*Hardy-Weinberg equilibrium*) and between (linkage equilibrium) variants, a GRS composed of a single SNP with MAF=0.5 would have a distribution of 0.25, 0.50 and 0.25 in individuals with 0, 1 and 2 risk alleles; a GRS composed of 2 SNPs, each with MAF=0.5 would have a distribution of 0.0625 0.25 0.375 0.25 and 0.0625 for individuals with 0, 1, 2, 3 and 4 risk alleles, respectively (*Figure 18a*). In a large, randomly selected sample of individuals, this distribution begins to approach a normal distribution as we increase the number of independent genetic variants from which the score is composed (see *Figure 18b-c*). This holds true in any finite population, and for any number of genetic variants that comprise the GRS and with any distribution of risk allele frequencies. However, depending on the allele frequencies of the genetic variants in the score, the shape of the distribution varies. As the frequencies of the genetic variants increase, the distribution of the GRS in the population tends towards the right (carrying more risk alleles), and the opposite happens as the alleles tend to have rarer allele frequencies (*Figure 18d*).

Figure 18. Different genetic risk score distributions according to different number of genetic variants included. Adapted form Plomin *et al.* [2009].



Furthermore, when a case-control design is used to compare the distributions of the allelic load of the individuals, if the variants are truly associated with the disease and are independent of each other the distribution observed in the cases group is going to be presented shifted (to the right) in relation to the control group, representing the excess in risk conferred by the genetic variants.

While a GRS can be expressed as the integer number of risk alleles carried by an individual, it can also be weighted by the magnitudes of the effects of the individual variants on disease risk, thereby accounting for the differences in risk attributable to variants with stronger or more subtle effects on risk (see *Equation 2*).

Equation 2. Formula used for the generation of a genetic risk score, including weighting of individual variants where necessary.

$$GRS = \sum_{i=1}^{n} \beta_i \cdot SNP_i$$

Where:

- β_i: effect size reported for variant i;
- SNP_i={0,1,2}: the number of copies of SNP i; where a score contains SNPs that have been imputed, the estimate genotype of SNPi can be expressed as the dosage of the risk allele, taking a value within the range [0,2].

The value of the GRS computed for each individual can be included as a variable in the risk function in the same manner as all other risk variables (see *Equation 3*), where the deviation of each individual's score (*Equation 2*) from the population mean of the score is multiplied by the per-unit effect size of the score.

$$\text{risk} = 1 - \widehat{S}^{\exp\left[\sum_{p=1}^{P}\left(\beta_{CVRF_p} \cdot CVRF_{p,i}\right) - \sum_{p=1}^{P}\left(\beta_{CVRF_p} \cdot \overline{CVRF_P}\right) + \beta_{GRS} \cdot \left(GRS_i - \overline{GRS}\right)\right]}$$

Equation 3. Extension of the REGICOR adaptation to the Framingham risk function to include a genetic risk score.

Where:

- risk: individual probability of suffering a coronary event for a given CVRF profile and a given set of risk variant genotypes.
- $CVRF_{p,i}$: value of individual i for cardiovascular risk factor *p*.
- β_{CVRFp}: risk effect (logarithm of the hazard ratio) for cardiovascular risk factor *p*.
- $\overline{CVRF}_p$: population mean of cardiovascular risk factor *p*.
- GRS_i: individual value of the genetic risk score.
- βGRS: log-hazard-ratios of the genetic risk score.
- $\overline{\text{GRS}}$: population average value of the genetic risk score.

Advantages and limitations of using genetic variants to improve risk functions

The greatest advantage of the introduction of genetic information when compared to other biomarkers is that the values remain unchanged throughout life. Therefore, the information provided by a genetic test could be more representative of the lifetime exposure to the specific risk factor with which it is associated than a single laboratory measurement, which is maybe susceptible to greater measurement error and intra-individual variation. Another important advantage is that genotyping even hundreds of polymorphism is likely to be much cheaper and more replicable than for some CVRF measurements (for example HDL-cholesterol).

Among the limitations of using genetic information as a marker of risk is the fact that the heritability currently explained by the genetic variants that are known to be associated with CHD is lower than 10%, and the fact that this type of biomarker could not be used to monitor changes during life or responses to treatments or interventions.

3.2. HYPOTHESES

i) A GRS composed of variants associated with CHD independently of CVRFs presents a distinct distribution in individuals with and without disease.

ii) That GRS is a predictor of future CHD/CVD events.

iii) The addition of this GRS in the classical cardiovascular function is able to improve the category in which individuals are classified.

3.3. OBJECTIVES

i) To assess, using a case-control study design, the magnitude of the association between CHD risk and a multi-locus genetic risk score composed of variants that are individually associated with CHD risk independently of CVRFs.

ii) To determine, using a population-based cohort design, the per-unit effect on risk of incident CVD and CHD of a multi-locus genetic risk score composed of variants that are individually associated with CHD risk independently of CVRFs.

iii) To assess whether the inclusion of this genetic risk score in the classical cardiovascular risk function improve its capacity to predict CVD and CHD events in populations with low and high CVD mortality.

3.4. Article 3:
Additive effects of multiple genetic variants on the risk of coronary artery disease. *Rev Esp Cardiol. 2010; 63(8):925-33.*

3.5. Article 4:
Assessment of the value of a genetic risk score in improving the estimation of coronary risk.

Atherosclerosis. 2012; Article in press.

G Model
ATH-12515; No. of Pages 8

ARTICLE IN PRESS

Atherosclerosis xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Atherosclerosis

journal homepage: www.elsevier.com/locate/atherosclerosis

Assessment of the value of a genetic risk score in improving the estimation of coronary risk

Carla Lluis-Ganella [a,1], Isaac Subirana [b,a,1], Gavin Lucas [a], Marta Tomás [a], Daniel Muñoz [a], Mariano Sentí [a,c], Eduardo Salas [d], Joan Sala [e], Rafel Ramos [f,g], Jose M. Ordovas [h,i], Jaume Marrugat [a], Roberto Elosua [a,b,*]

[a] Cardiovascular Epidemiology and Genetics, IMIM, Barcelona, Spain
[b] CIBER Epidemiology and Public Health (CIBERESP), Barcelona, Spain
[c] Pompeu Fabra University, Barcelona, Spain
[d] Gendiag.exe, Barcelona, Spain
[e] Cardiology Department, Hospital Universitari Josep Trueta, Girona, Spain
[f] Primary Care Research Institute (IDIAP-Jordi Gol), Girona, Spain
[g] Medical Science Department, Medical School, Universitat de Girona, Spain
[h] Nutrition and Genomics Laboratory, Jean Mayer US Department of Agriculture Human Nutrition Research Center on Aging, Tufts University School of Medicine, Boston, MA, United States
[i] The Department of Epidemiology and Population Genetics, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain

ARTICLE INFO

Article history:
Received 9 February 2012
Received in revised form 9 March 2012
Accepted 17 March 2012
Available online xxx

Keywords:
Coronary artery disease
Cardiovascular disease
Genetic variants
Genetic risk
Polymorphisms
Risk assessment

ABSTRACT

Background: The American Heart Association has established criteria for the evaluation of novel markers of cardiovascular risk. In accordance with these criteria, we assessed the association between a multi-locus genetic risk score (GRS) and incident coronary heart disease (CHD), and evaluated whether this GRS improves the predictive capacity of the Framingham risk function.

Methods and results: Using eight genetic variants associated with CHD but not with classical cardiovascular risk factors (CVRFs), we generated a multi-locus GRS, and found it to be linearly associated with CHD in two population based cohorts: The REGICOR Study ($n = 2351$) and The Framingham Heart Study ($n = 3537$) (meta-analyzed HR [95%CI]: ∼1.13 [1.01–1.27], per unit). Inclusion of the GRS in the Framingham risk function improved its discriminative capacity in the Framingham sample (*c*-statistic: 72.81 vs.72.37, $p = 0.042$) but not in the REGICOR sample. According to both the net reclassification improvement (NRI) index and the integrated discrimination index (IDI), the GRS improved re-classification among individuals with intermediate coronary risk (meta-analysis NRI [95%CI]: 17.44 [8.04; 26.83]), but not overall.

Conclusions: A multi-locus GRS based on genetic variants unrelated to CVRFs was associated with a linear increase in risk of CHD events in two distinct populations. This GRS improves risk reclassification particularly in the population at intermediate coronary risk. These results indicate the potential value of the inclusion of genetic information in classical functions for risk assessment in the intermediate risk population group.

1. Introduction

The main goal of primary cardiovascular prevention is to reduce the incidence of clinical events [1]. Generally, two strategies are used: (i) population-wide interventions based on the promotion of healthy lifestyles and public health policies; and (ii) targeting of high risk individuals, in whom intensive strategies are implemented to control cardiovascular risk factors. In clinical practice, cardiovascular risk functions are used to identify the high risk individuals by estimating the probability of presenting a coronary (CHD) event, usually in the subsequent 10 years [2]. Although these screening methods are well established and widely used, the majority of the CHD events occur in individuals who are classified as having low or intermediate risk [3]. Therefore, the improvement of risk estimation, especially in the intermediate risk group, is a priority for research. In this regard, the identification of new biomarkers, particularly those that provide information complementary to that already provided by classical cardiovascular risk factors (CVRFs)

Abbreviations: AHA, American Heart Association; CHD, coronary heart disease; CVD, Cardiovascular diseases; CVRFs, cardiovascular risk factors; DNA, deoxyribonucleic Acid; GRS, genetic risk score; GWAS, genome-wide association studies; IDI, integrated discrimination improvement; MI, myocardial infarction; NRI, Net reclassification improvement; REGICOR, Registre Gironí del Cor.

[*] Corresponding author at: Cardiovascular Epidemiology and Genetics Research Group, IMIM, Doctor Aiguader 88, 08003 Barcelona, Spain. Tel.: +34 933 160 800.
E-mail address: relosua@imim.es (R. Elosua).

[1] These authors contributed equally to this work.

0021-9150/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

[4], has been the subject of intense research in recent years. To that end, the American Heart Association (AHA) has proposed several essential steps [5] for assessing the potential value of such novel biomarkers in estimating risk: (i) initial demonstration of association between marker and event risk (proof of concept), (ii) validation of this relationship in prospective cohort studies, (iii) assessment of the improvement of the predictive capacity of the risk function due to the addition of the marker, (iv) assessment of effects on patient management and outcomes, and, (v) cost-effectiveness of population-wide implementation.

Genome-wide association studies (GWAS) have led to the identification of a series of genetic variants that are robustly associated with CHD risk [6], although their individual effects on risk are generally quite small. Since these effects have also been observed to be generally additive, overall genetic risk load, formulated as a multi-locus genetic risk score (GRS), has been proposed [7,8] as a potentially informative biomarker for improving the estimation of coronary risk [1,9]. We have recently reported the results of a large case-control study aimed at addressing the first step of the AHA recommendations, in which we observed a robust association between CHD risk and a GRS composed of variants associated with CHD, but not with classical CVRFs [10].

Following on from our previous work, the aims of the current study were to address steps 2 and 3 of the AHA recommendations for the same GRS. First, we assessed the association between the multi-locus GRS and incident CHD events in two prospective cohort studies with low and high CHD mortality (AHA, step 2). Second, we assessed whether the inclusion of this GRS improves the predictive capacity of the Framingham risk function (AHA, step 3). In addition, we evaluated the hypothesis that the improvement in predictive capacity provided by the GRS is greater among individuals with intermediate risk.

2. Methods

An extended description of the methods used is given in the Supplementary methods. Supplementary materials section (*Sx.x*), table (*S.Tx*), figure (*S.Fx*) and analysis (*S.Ax*) numbers are indicated in parentheses throughout the manuscript.

2.1. Design

Two prospective population-based cohorts were analyzed in this study. (i) The REGICOR (Registre Gironí del Cor) cohort originally included 4778 individuals from two population-based cross-sectional studies conducted in the province of Girona, in north-eastern Spain, in 1995 and 2000 [11]. This population has low CHD mortality [12]. (ii) The Framingham Heart Study originally included 5209 men and women recruited in 1948 [13] and 5124 offspring of the original participants and their spouses recruited in 1971 [14], from whom DNA was collected during the 1980s and 1990s [15]. This population has relatively high CHD mortality. We obtained access to phenotype and genotype data for the Framingham sample under the Framingham Share initiative via the Database of Genotypes and Phenotypes (dbGaP, ncbi.nlm.nih.gov/dbgap; Project number 1534). To maximize the number of participants included in the analysis for whom genetic data was available, we set exams 15 and 5 as the baseline visits for the Original Cohort (2632 individuals, 1977–1979) and the Offspring Cohort (3799 individuals, 1991–1995), respectively (*S.F1*).

For both cohorts we selected participants aged 35–74 years at the time of the exams, who were free of cardiovascular disease (CVD) at that time, and for whom DNA and complete follow-up information was available.

2.2. Selection of genetic variants, genotyping and multi-locus risk score generation

We selected 8 genetic variants associated with CHD but not with CVRFs (blood pressure, total cholesterol, low density lipoprotein (LDL) cholesterol, high density lipoprotein (HDL) cholesterol, triglycerides, diabetes, smoking) and generated a multi-locus GRS as previously described [10]. Briefly, the genetic variants were selected from the catalog of GWA studies of the National Human Genome Research Institute (NHGRI GWAS catalog [6], reviewed in August 2010) using the following criteria: (a) the genetic variants were associated with CHD ($p \leq 1 \times 10^{-6}$); (b) when two or more genetic variants were in linkage disequilibrium ($r^2 > 0.3$) only one was selected; (c) we excluded SNPs that were previously reported, either in the literature or the NHGRI GWAS catalog, to be associated with one or more CVRFs (see more detail of this process in *S1.1* and *S.F2*). The variants selected were: rs17465637 in *MIA3*; rs6725887 in *WDR12*; rs9818870 in *MRAS*; rs12526453 in *PHACTR1*; rs1333049 near *CDKN2A/2B*; rs1746048 near *CXCL12*; rs9982601 near *SCL5A3*. We also included the rs10455872 variant in *LPA*, which has recently been shown to be strongly associated with CHD risk independently of CVRFs [16].

REGICOR samples were genotyped by Centro Nacional de Investigación Oncológica (CNIO, Madrid, Spain) using the Cardio inCode chip (Ferrer inCode, Barcelona, Spain), which is based on Veracode (Illumina, San Diego, USA) and KASPar (KBioscience, Hoddesdon, United Kingdom) technologies. Genotype data for the Framingham participants was obtained via dbGaP for genotyped (Affymetrix 500 K and 50 K chips) and imputed variants (HapMap CEU release 22, b36) (*S1.3*). Quality control criteria were applied both to individuals and selected SNPs (*S1.4*).

A multi-locus GRS was computed for each individual as the sum of the number of risk alleles across all 8 variants [10], after weighting each one by its estimated effect size in the CARDIoGRAM study (*S1.2*) [17].

2.3. Follow-up and phenotype definition

All REGICOR participants were periodically contacted to ascertain whether they had presented any CHD event up until the end of 2009, and events were reviewed using hospital or primary care records. Fatal events were identified from regional and national mortality registers. After reviewing all medical records and physician notes, suspected CHD events were classified in committee according to standardized criteria [18].

Among Framingham participants, a record was made of all CHD events that occurred during follow-up until the end of 2007. Suspected CHD events were reviewed by a panel of Framingham physician investigators after reviewing all available medical records and physician notes using standardized criteria [19].

CHD events included myocardial infarction (MI), angina, coronary revascularization and death due to CHD (*S2*).

2.4. Estimation of ten-year cardiovascular risk

Coronary risk was estimated using the standard 10-year Framingham risk function [19] and the REGICOR function, which is an adaptation of the former that has been validated and calibrated for the Spanish population (*S3 and S4*) [9]. Both functions included age, sex, systolic and diastolic blood pressure, total cholesterol level, HDL cholesterol level, smoking status, diabetes status and the GRS, where appropriate. Risk was computed using the following formula (also see *S4*),

$$Risk = 1 - S_{\bar{X}}^{\exp\left(\sum_{j=1}^{p} \beta_j^F \cdot (F_j - \bar{F}_j) + \beta^{GRS} \cdot (GRS - \overline{GRS})\right)},$$

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

Table 1
Description of the phenotypic characteristics of the individuals included in the analysis from the REGICOR and from the Framingham Heart Study cohorts.

	REGICOR				Framingham			
	All	None	CHD	*p*-Value	All	None	CHD	*p*-Value
N	2351	2190	107	–	3537	2863	429	–
Age (years)[a]	53.9 (11.2)	53.3 (11.1)	61.4 (9.40)	<0.001	56.0 (9.3)	54.8 (9.2)	60.5 (7.8)	<0.001
Gender (male)[b]	1123 (47.8)	1016 (46.4)	74 (69.2)	<0.001	1540 (43.5)	1190 (41.6)	250 (58.3)	<0.001
SBP (mmHg)[a]	132 (20.8)	131 (20.5)	147 (18.0)	<0.001	127 (18.3)	125 (17.9)	134 (17.4)	<0.001
DBP (mmHg)[a]	79.5 (10.4)	79.3 (10.3)	82.6 (10.7)	0.004	75.0 (9.8)	74.6 (9.8)	77.7 (9.6)	<0.001
Hypertension[b]	938 (40.1)	822 (37.7)	78 (72.9)	<0.001	1121 (31.7)	802 (28.0)	214 (50.0)	<0.001
Smoking[b]	511 (22.0)	476 (22.0)	27 (25.5)	0.469	713 (20.2)	531 (18.5)	111 (25.9)	0.002
Total cholesterol (mg/dL)[a]	225 (42.4)	224 (42.0)	233 (46.6)	0.103	210 (38.6)	207 (37.4)	224 (41.0)	<0.001
LDL cholesterol (mg/dL)[a]	152 (37.9)	151 (37.7)	159 (39.6)	0.125	126 (34.0)	124 (33.3)	133 (35.7)	0.001
HDL cholesterol (mg/dL)[a]	51.7 (13.3)	52.1 (13.2)	44.8 (12.4)	<0.001	51 (15.2)	52 (15.3)	46 (13.1)	<0.001
Triglycerides (mg/dL)[c]	92 (70–127)	91 (69–125)	123 (90–170)	<0.001	116 (83–172)	112 (80–164)	158 (104–217)	<0.001
Cholesterol treatment[b]	157 (6.7)	136 (6.2)	16 (15.0)	0.003	166 (4.7)	118 (4.1)	28 (6.5)	0.055
Diabetes[b]	316 (13.8)	280 (13.1)	29 (27.6)	<0.001	226 (6.4)	138 (4.8)	60 (14.0)	<0.001
Diabetes treatment[b]	96 (4.11)	74 (3.4)	18 (16.8)	<0.001	90 (2.5)	48 (1.7)	31 (7.2)	<0.001
Body mass index (kg/m^2)[a]	27.4 (4.47)	27.3 (4.46)	28.9 (4.47)	0.001	27.1 (4.8)	27.0 (4.8)	27.9 (4.4)	<0.001
Obesity (BMI≥30 kg/m^2)[b]	596 (25.6)	540 (24.9)	38 (35.8)	0.046	780 (22.1)	604 (21.2)	117 (27.3)	0.006
Family history of CHD[b]	272 (11.7)	301 (11.5)	19 (17.9)	0.150	551 (24.8)	478 (24.3)	55 (32.5)	0.016

CHD, individuals who presented a coronary event during the follow-up; SBP, systolic blood pressure; DBP, diastolic blood pressure; LDL, low density lipoprotein; HDL, high density lipoprotein; BMI, body mass index.
"*None*": all individuals except those who presented any cardiovascular event (CHD, stroke or peripheral arterial disease).

[a] Mean (standard deviation).
[b] *n* (proportion, %).
[c] Median (25th and 75th percentiles).

where $(1-S)$ is the probability of presenting a CHD event in the next 10 years based on the incidence of CHD in the population, (F_j) is the individual's exposure to the various risk factors considered, including the genetic risk factor (*GRS*), $(\bar{F}_j, \overline{GRS})$ is the population mean of those risk factors, and (β) is the effect size of each risk factor.

3. Statistical analysis

We used standard parametric and non-parametric methods to compare the characteristics of different groups of individuals (*S4*). We tested for association between incidence of coronary events and individual genetic variants and the GRS using Cox proportional hazards models, with adjustment for CVRFs (see formula above). We accounted for family relatedness in the Framingham cohort by adjusting for the first five genetic principal components [20]. Each cohort was analyzed separately, and the estimates were pooled using an inverse-variance weighted meta-analysis under a random effects model [21].

We used three different statistics to assess the potential value of including the GRS in risk prediction:

(a) the goodness-of-fit of the models was evaluated using a version of the Hosmer–Lemeshow test [22];
(b) the discriminative capacity of the model was evaluated using the concordance index (*c*-statistic) [23];
(c) reclassification improvement was calculated using the net reclassification improvement (NRI) index [24] and the integrated discrimination improvement (IDI) index [25].

For the assessment of reclassification improvement, we defined four risk categories (low, intermediate-low, intermediate-high and high) with cut-off points defined in each cohort, according to current guidelines in each country (REGICOR: [0–5)%, [5–10)%, [10–15)%, ≥15%; Framingham: [0–10)%, [10–15)%, [15–20)%, ≥20%, respectively). Analyses that focused on individuals with intermediate risk included individuals from both the intermediate-low and intermediate-high groups. We calculated the expected number of events at 10-years in each risk category and in each cohort using Kaplan–Meier estimates [26]. A bootstrapping method was used to construct confidence intervals for IDI and NRI in order to account for uncertainty in the Kaplan–Meier estimates, as suggested by Steyerberg et al. [26].

For each SNP and for the GRS we computed our study's power to detect associations in each cohort and in the meta-analysis (*S5*).

All analyses were performed using the R statistical package (version 2.11) [27].

4. Results

4.1. Sample selection and sample characteristics

The process of selection of individuals to include in our analysis is described in *S.F1*. From the REGICOR sample we included 2351 individuals, including 107 CHD events, with a mean follow-up of 9.75 years. From the Framingham sample we included 3537 individuals, including 429 events, with a mean follow-up of 13.32 years. In the REGICOR sample, we observed no significant difference in the estimated 10-year coronary risk between individuals who were included in the analysis compared to those who were not included (*S.T1*). In the Framingham sample, many individuals were excluded from our study due to the non-availability of genetic data, with the result that those who were included presented a better cardiovascular risk profile (*S.T1*) and a lower incidence of CHD events than those who were not included (suggesting a survival bias related to DNA availability; *S.F3*).

The characteristics of the participants from each cohort that were included in our analyses, stratified by presence of CHD events are shown in Table 1. The observed effects of each cardiovascular risk factor on risk of having a CHD event were concordant with those expected and are presented in *S.T2*.

4.2. Validation of the association between the GRS and risk of CHD

The results of the genotyping quality control process are shown in *S.T3*, and those of the test for association between the genetic variants included in the GRS and incidence of CHD events is shown in *S.T3* (also see *S.T4* for power computations). Only the rs1333049 variant in *CDKN2A/2B* was nominally associated with CHD events in the meta-analysis of both studies.

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

Table 2
Description of the characteristics of the participants across quintiles of the genetic risk score in both cohorts.

Variables	Quintiles of genetic score						
	Q1	Q2	Q3	Q4	Q5	*p*-Value	*p*-Trend
REGICOR							
N	524	416	473	471	467		
Age (years)[a]	54.1 (11.1)	52.9 (11.0)	54.6 (11.4)	54.2 (11.0)	53.6 (11.3)	0.170	0.998
Gender (men)[b]	243 (46.4)	205 (49.3)	217 (45.9)	234 (49.7)	224 (48.0)	0.705	0.581
Total cholesterol (mg/dL)[a]	221 (42.8)	225 (41.8)	227 (42.5)	228 (42.0)	225 (42.8)	0.072	0.049
HDL cholesterol (mg/dL)[a]	51.1 (12.9)	52.4 (13.5)	52.5 (13.4)	51.0 (13.0)	51.5 (13.4)	0.304	0.866
SBP (mmHg)[a]	132.0 (22.0)	131.0 (20.4)	132.0 (20.4)	134.0 (21.5)	132.0 (19.5)	0.278	0.749
DBP (mmHg)[a]	78.9 (10.2)	79.5 (10.8)	79.0 (10.2)	80.2 (10.6)	79.8 (10.0)	0.257	0.099
Diabetes[b]	62 (12.1)	71 (17.5)	66 (14.3)	61 (13.3)	56 (12.3)	0.137	0.590
Smoking[b]	107 (20.7)	87 (21.0)	98 (20.8)	107 (23.1)	112 (24.3)	0.577	0.128
Family history of CHD[b]	46 (8.88)	51 (12.4)	55 (11.6)	63 (13.5)	57 (12.4)	0.207	0.064
Estimated 10-year coronary risk[c]	3.2 (1.7–6.4)	3.2 (1.6–5.6)	3.4 (1.69–6.5)	3.5 (1.8–6.6)	3.1 (1.8–5.9)	0.196	0.607
Incidence of coronary events[d]	5.08	3.44	3.97	5.98	7.06	0.038	0.015
FRAMINGHAM							
N	743	712	681	711	690		
Age (years)[a]	56.6 (9.10)	56.1 (9.12)	55.6 (9.58)	56.1 (9.12)	55.6 (9.41)	0.172	0.060
Gender (men)[b]	351 (47.2)	321 (45.1)	305 (44.8)	299 (42.1)	264 (38.3)	0.008	<0.001
Total cholesterol (mg/dL)[a]	208 (37.1)	209 (37.6)	213 (39.0)	211 (39.3)	210 (39.8)	0.151	0.242
HDL cholesterol (mg/dL)[a]	50.5 (14.7)	50.2 (14.9)	51.1 (15.2)	52.0 (15.8)	51.3 (15.2)	0.151	0.048
SBP (mmHg)[a]	127 (18.4)	126 (17.0)	127 (18.8)	126 (18.2)	127 (18.9)	0.938	0.647
DBP (mmHg)[a]	75.2 (10.2)	75.1 (9.54)	74.8 (9.81)	75.0 (9.65)	74.7 (9.73)	0.872	0.329
Diabetes[b]	47 (6.33)	59 (8.29)	32 (4.70)	39 (5.49)	49 (7.10)	0.059	0.658
Smoking[b]	132 (17.8)	146 (20.5)	146 (21.4)	140 (19.7)	149 (21.6)	0.358	0.144
Family history of CHD[b]	113 (24.6)	112 (24.7)	105 (24.7)	109 (24.8)	112 (25.3)	0.999	0.763
Estimated 10-year coronary risk[c]	8.3 (4.7–14.4)	8.0 (4.8–13.9)	8.5 (4.4–14.7)	7.7 (4.1–13.7)	7.7 (4.2–13.9)	0.261	0.229
Incidence of coronary events[d]	5.39	6.60	7.62	7.50	8.42	0.361	0.054

HDL, high density lipoprotein; SBP, systolic blood pressure; DBP, diastolic blood pressure; CHD, coronary heart disease.
[a] Mean (standard deviation).
[b] *n* (proportion, %).
[c] Estimation of 10-year coronary risk based on the classical risk function without the GRS, mean (95% confidence interval).
[d] Number of cases/100 individuals in 10 years.

Clinical characteristics of the participants within each quintile of the GRS are shown in Table 2. The GRS was not directly associated with any of the classical CVRFs in either cohort, with the exception of gender in Framingham (which we believe to be an artefact of the survival bias among individuals for whom DNA was available). The proportion of participants with a positive family history of CHD did not change between quintiles of the GRS. We observed a general increase in the incidence of coronary events from the bottom to the top quintile of the GRS in both cohorts (Table 2).

For the GRS, we estimated that our study had 80% power to detect a HR of 1.17, 1.09 and 1.18 per unit increase in REGICOR, Framingham, and the meta-analysis, respectively (*S.T4*). Both the models with and without the GRS were well calibrated in the REGICOR sample, but not in the Framingham sample, where we observed fewer events than expected, likely due to the survival bias mentioned above (*S.F4*).

The GRS was linearly associated with incidence of CHD in both cohorts ($p = 0.001$ in REGICOR and $p = 0.016$ in Framingham; Table 3), and in the meta-analysis (HR = 1.13; 95% CI: 1.01–1.27) (Table 3). This association remained statistically significant after further adjustment for family history of CHD (HR = 1.17; 95% CI: 1.09–1.26). Participants in the top quintile of the GRS had 1.44 times greater risk of CHD, compared to those in the bottom quintile (*p*-value for linear trend 0.002) (Table 3). In both cohorts the distribution of the GRS was slightly shifted to the right in individuals who had had an event, compared to those who had not (Fig. 1).

4.3. Improvement in predictive capacity: discrimination and reclassification

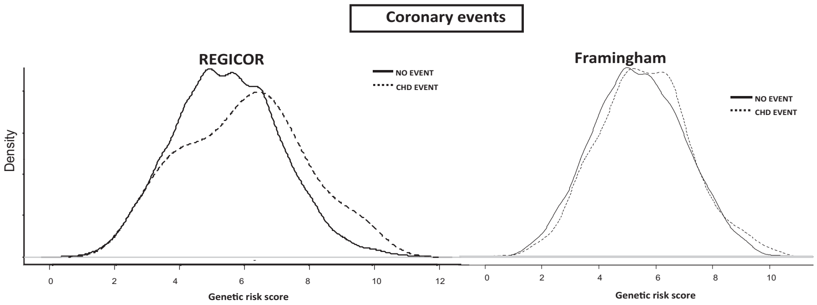
The addition of the GRS to the basic risk function improved its capacity to predict CHD in the Framingham cohort (*c*-statistic, 72.81 vs. 72.37, *p*-value = 0.042) but not in the REGICOR cohort (78.35 vs. 78.33, *p*-value = 0.806).

Table 3
Multivariate adjusted association between the genetic risk score and risk of coronary events as a continuous variable and between quintiles.

Genetic risk score	REGICOR		Framingham		Meta-analysis	
	HR [95%CI][a]	*p*-Value	HR [95%CI][a]	*p*-Value	HR [95%CI][a]	*P*-Value
Continuous	1.21 [1.09–1.36]	0.001	1.07 [1.01–1.14]	0.016	1.13 [1.01–1.27]	0.038
Quintiles	*p*-Trend	0.010	*p*-Trend	0.032	*p*-Trend	0.002
Q1	1	–	1	–	1	–
Q2	0.76 [0.37–1.53]	0.437	1.06 [0.78–1.45]	0.711	1.00 [0.76–1.34]	0.973
Q3	0.84 [0.45–1.58]	0.586	1.22 [0.90–1.66]	0.206	1.12 [0.83–1.52]	0.448
Q4	1.19 [0.67–2.12]	0.555	1.33 [0.99–1.80]	0.060	1.30 [1.00–1.69]	0.053
Q5	1.86 [1.08–3.20]	0.025	1.29 [0.95–1.75]	0.104	1.44 [1.04–2.01]	0.030

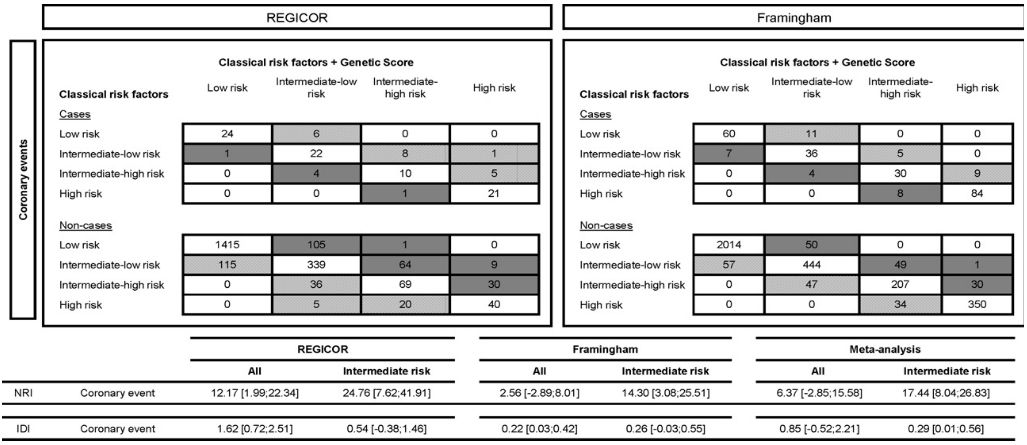
All models were adjusted for the sum of the products of the coefficient for each classical risk factor estimated in the Framingham original and calibrated risk functions and the difference between the participant's value and the population mean of that risk factor (see main text for formula). To account for family structure in the Framingham cohort we also adjusted for the first five genetic principal components.
[a] HR [95%CI]: Hazard ratio [95% confidence interval].

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024



	REGICOR NO EVENT	REGICOR CHD EVENT		Framingham NO EVENT	Framingham CHD EVENT	
	N=2,190	N=107	P-value	N=2,863	N=429	P-value
Weighted genetic score (SD)	5.46 (1.67)	5.99 (1.89)	0.005	5.38 (1.62)	5.59 (1.63)	0.028
Coronary risk (95%CI)	3.21 (1.65-5.89)	7.85 (4.64-12.4)	<0.001	7.07 (3.65-12.4)	14.2 (8.77-21.7)	<0.001
Coronary risk + genetic score (95%CI)	3.23 (1.59-5.92)	7.98 (5.15-13.3)	<0.001	6.82 (3.67-12.2)	14.9 (9.02-22.9)	<0.001

Fig. 1. Distribution of genetic risk score in REGICOR and Framingham participants according to the incidence of coronary events during the follow-up. The genetic risk score is represented in the ordinal axis (*X* axis) and is computed as a cumulative sum of all the risk alleles that a person carries, weighted by the effect of each SNP, and theoretically ranging from 0 to 16 copies. "*No event*": All individuals except those who presented any cardiovascular event (CHD, stroke or peripheral arterial disease).



Coronary events

REGICOR

Classical risk factors	Classical risk factors + Genetic Score: Low risk	Intermediate-low risk	Intermediate-high risk	High risk
Cases				
Low risk	24	6	0	0
Intermediate-low risk	1	22	8	1
Intermediate-high risk	0	4	10	5
High risk	0	0	1	21
Non-cases				
Low risk	1415	105	1	0
Intermediate-low risk	115	339	64	9
Intermediate-high risk	0	36	69	30
High risk	0	5	20	40

Framingham

Classical risk factors	Classical risk factors + Genetic Score: Low risk	Intermediate-low risk	Intermediate-high risk	High risk
Cases				
Low risk	60	11	0	0
Intermediate-low risk	7	36	5	0
Intermediate-high risk	0	4	30	9
High risk	0	0	8	84
Non-cases				
Low risk	2014	50	0	0
Intermediate-low risk	57	444	49	1
Intermediate-high risk	0	47	207	30
High risk	0	0	34	350

		REGICOR		Framingham		Meta-analysis	
		All	Intermediate risk	All	Intermediate risk	All	Intermediate risk
NRI	Coronary event	12.17 [1.99;22.34]	24.76 [7.62;41.91]	2.56 [-2.89;8.01]	14.30 [3.08;25.51]	6.37 [-2.85;15.58]	17.44 [8.04;26.83]
IDI	Coronary event	1.62 [0.72;2.51]	0.54 [-0.38;1.46]	0.22 [0.03;0.42]	0.26 [-0.03;0.55]	0.85 [-0.52;2.21]	0.29 [0.01;0.56]

Fig. 2. Reclassification of individuals based on the 10-year predicted risk of coronary heart disease with and without the genetic risk score. Risk categories were defined using national recommendations. In REGICOR the cut-off points were: low [0–5)%, intermediate-low [5–10)%, intermediate-high [10–15)%, and high ≥15% risk; in Framingham the cut-off points were: low [0–10)%, intermediate-low [10–15)%, intermediate-high [15–20)% and high ≥20% risk. Light gray cells represent an improvement in reclassification and dark gray cells represent the opposite.

We observed a general tendency for both measures of reclassification improvement, the NRI and IDI, to increase after addition of the GRS to the basic risk function, although this improvement was not statistically significant for either measure in the meta-analysis of the two cohorts. However, reclassification improvement was more marked in the group with intermediate risk, and was statistically significant for both measures (NRI: 17.44, 95%CI 8.04;26.83; IDI: 0.29, 95%CI 0.01;0.56). Reclassification data and NRI and IDI for each cohort are shown in Fig. 2.

Results for a GRS constructed from 4 SNPs that had consistent directions of effect in both cohorts, and for a GRS without the *CDKN2A/B* variant were similar and are described in *S.A2* and *S.A3*.

5. Discussion

In accordance with the AHA statement regarding assessment of the value of novel risk biomarkers [5], we have validated the association between a multi-locus GRS and incidence of CHD events in two prospective cohort studies, and have shown that this GRS improves the capacity of the Framingham risk function to predict CHD events, primarily among individuals with intermediate risk.

5.1. Validation of the association between the GRS and risk of CHD

In this study, we selected a series of genetic variants that have been found to be robustly associated with CHD risk in multiple large

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

independent samples and populations, but have not been reported to be associated with CVRFs. Unsurprisingly, most of these variants were not nominally associated with CHD incidence in either of the cohorts in this study, mainly due to their sample size and the weak risk effects of the variants. However, the relevance of these variants for CHD risk is beyond doubt and has been validated in different meta-analyses [17].

A GRS constructed using these variants was linearly associated with incidence of CHD events in two cohorts with distinct background levels of 10-year coronary risk. The effect size of the GRS was modest (~13% increase in risk of CHD per unit), and was also independent of familial history of CHD [4]. This effect size is smaller than that reported in the initial discovery case-control studies [10], which is likely due to these studies' tendency to overestimate the effect sizes of real associations. In fact, the effect size of our GRS could even be slightly underestimated because of the fact that the individuals included in the Framingham analysis have a more favorable cardiovascular risk profile than those who were excluded due to non-availability of DNA samples, thereby introducing a survival bias.

A recent study of the Framingham Heart Study investigators using a GRS comprising 13 SNPs associated with CHD reports the same results that we have obtained in this analysis, although the group of SNPs is slightly different and the events of interest include only myocardial infarction and coronary death [28]. We also observed a similar difference in risk between the top and bottom quintiles of the score (HR = 1.44) to that reported by Ripatti et al. [8] (meta-analyzed HR = 1.66) for a similar GRS comprising 13 SNPs associated with CHD, but not explicitly independent of CVRFs. However, this association has not been confirmed by other authors [29]. A number of differences between the Women's Health Study (WHS) and the rest of studies may explain the observed discordant results, but probably the most important is related to the different sampling strategy used in the WHS which included young women with relatively low baseline risk for CHD whereas the rest of studies are community- or population-based including men and women that may have a higher baseline CHD risk.

5.2. Improvement in predictive capacity: discrimination and reclassification

As has been observed for several other biomarkers [30], we observed no marked improvement in the discriminative capacity of the risk function, as measured by the *c*-statistic, which highlights the challenge of risk prediction for complex traits [31]. However, some authors have expressed concerned about the use of the *c*-statistic as the main predictive metric, when the main goal in clinical practice is to better estimate an individual's risk category, leading to more effective preventive treatment decisions [32]. To address this problem metrics such as IDI and NRI have been proposed that assess a risk function's ability to re-classify individuals who go on to have a coronary event and those who do not into higher and lower risk categories, respectively [24].

In this study, we observed a general tendency for reclassification to improve after addition of the GRS to the basic risk function (Fig. 2), although, as has been observed in previous studies [8,33], the numbers of cases correctly reclassified into higher risk categories was a modest fraction of the total number of cases, and also some individuals were also incorrectly reclassified. This reclassification improvement was not statistically significant overall.

5.3. Improved reclassification in individuals with intermediate coronary risk using the GRS

From a clinical perspective, the low sensitivity of risk functions is exemplified by the fact that a significant proportion of CHD events occur in individuals with intermediate coronary risk [3,34], so improving risk estimation in this group could have a significant impact on the total burden of CHD, and on the effectiveness of population-wide treatment strategies. We observed that the GRS significantly improved the reclassification of individuals with intermediate risk, above the level of improvement observed overall. Similarly, Ripatti et al. have recently reported a higher NRI in individuals with intermediate CHD risk (9.7%) than that observed in the population as a whole [8], although the improvement was less marked than for the intermediate risk group in our study (17.44%). Improvements in risk reclassification have also been observed in other studies through the inclusion of single genetic variants or a GRS in cardiovascular risk functions [8,28,33,35,36], with greater improvement in the intermediate risk group, where this has been assessed [8,33].

5.4. Strengths and limitations of the study

We highlight the following strengths in our study. First, we included two cohorts, which allowed us to evaluate the robustness of the effect size of the GRS, and to verify this effect size in populations with distinct basal cardiovascular risks. Second, the variants included in our score are likely to represent loci that are truly relevant for CHD risk. The fact that most of these variants individually were not significantly associated with CHD incidence, but that the GRS was significantly associated and also generally improved risk reclassification highlights the potential gain in information afforded by using the GRS. Third, these variants are largely independent of CVRFs, which is considered as an optimal strategy [4]. Consequently, we found that the GRS constructed from these variants was also independent of the CVRFs, and of the 10-year risk estimation based only on those CVRFs [4]. This indicates that this GRS provides complementary information to that already provided by the classical risk function. Moreover, the GRS was also found to be independent of family history of CHD [4].

Finally, and in accordance with European guidelines highlighting the importance of assessing overall cardiovascular risk [1], we have also extended our analysis to a broader definition of CVD events, including coronary events, stroke and peripheral artery disease, and observed largely consistent results to those for coronary events only (*S.A1*).

The main limitation of this study is the fact that the size of the individual cohorts and the number of events observed is limited. This is especially true in the REGICOR sample because of the low incidence of disease in this population. Moreover, a number of additional markers that fulfill our selection criteria have been reported since we performed our initial SNP selection in August 2010 (rs12936587, rs2505083, rs17114036 and rs11556924, reported in refs [17,37]). However, adding these 4 SNPs to the 8-SNP GRS and repeating the analyses in the Framingham cohort (genotype data for these SNPs were not available in REGICOR), we obtained similar results in terms of the strength of the per-unit and per-quintile risk effects, and similar improvements in reclassification (*S.A4*). These results are also consistent with those of a recent study [28], which indicated that the addition of 16 recently discovered SNPs to a 13-SNP GRS did not provide a significant improvement in discrimination between individuals with and without CVD events. Also, the findings in this study may be applicable only to European Caucasians or their descendants. Finally, due to the survival bias mentioned above, we have probably underestimated the true per unit effect size of the GRS on risk of CHD in the Framingham study.

6. Conclusions

A multi-locus GRS based on genetic variants unrelated to CVRFs was associated with a linear increase in risk of CHD events in two

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

distinct populations. This GRS improves risk reclassification particularly in the population at intermediate coronary risk. These results indicate the potential value of the inclusion of genetic information in classical functions for risk assessment in the intermediate risk population group.

Acknowledgments

Funding sources: This work was supported by a grant from ACC1Ó (RD08-1-0024), the European Funds for Development (ERDF-FEDER), the Spanish Ministry of Science and Innovation through the Carlos III Health Institute [CIBER Epidemiología y Salud Pública, Red HERACLES RD06/0009, PI061254, PI09/90506], Fundació Marató TV3 (081810 and 080431), and by the Catalan Research and Technology Innovation Interdepartmental Commission [SGR 1195]. GL was funded by the Juan de la Cierva Program, Ministerio de Educación (JCI-2009-04684). MT was funded the Comissionat per a Universitat I Recerca del Departament d'Innovació, Universitats I Empresa de la Generalitat de Catalunya (2007-BP-B1-0068). JMO was supported by the National Heart, Lung, and Blood Institute grant HL-54776, National Institute of Diabetes and Digestive and Kidney Diseases, Grant Number DK075030 and by contracts 53-K06-5-10 and 58-1950-9-001 from the US Department of Agriculture Research. We also thank the Spanish National DNA Bank for providing DNA samples (Project 2008/0012).

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278.The corresponding author had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Disclosures: The authors from the Epidemiology and Cardiovascular Genetics Group, IMIM, declare that they are collaborating with the companies Gendiag.exe and Ferrer InCode in the assessment of genetic markers as a tool to improve the predictive capacity of cardiovascular risk functions. This collaboration was formalized by an agreement between these companies and IMIM, without personal benefit for the signing authors. Access to the Framingham data through the Framingham Share initiative in the Database of Genotypes and Phenotypes (Project number 1534) was available only for some of the authors (CL, IS, GL, MT, RE). ES is employee of Gendiag.exe.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.atherosclerosis.2012.03.024.

References

[1] Graham I, Atar D, Borch-Johnsen K, et al. European guidelines on cardiovascular disease prevention in clinical practice: full text. Fourth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). European Journal of Cardiovascular Prevention and Rehabilitation 2007;14(Suppl. 2): S1–113.
[2] Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–47.
[3] Marrugat J, Vila J, Baena-Diez JM, et al. Relative validity of the 10-year cardiovascular risk estimate in a population cohort of the REGICOR study. Revista Española de Cardiologia 2011;64:385–94.
[4] Thanassoulis G, Vasan RS. Genetic cardiovascular risk prediction: will we get there. Circulation 2010;122:2323–34.
[5] Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119:2408–16.
[6] Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A catalog of published genome-wide association studies; 2009, http://www.genome.gov/26525384.
[7] Janssens AC, Aulchenko YS, Elefante S, et al. Predictive testing for complex diseases using multiple genes: fact or fiction. Genetics in Medicine 2006;8:395–400.
[8] Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet 2010;376:1393–400.
[9] Marrugat J, D'Agostino R, Sullivan L, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. Journal of Epidemiology and Community Health 2003;57: 634–8.
[10] Lluis-Ganella C, Lucas G, Subirana I, et al. Additive effect of multiple genetic variants on the risk of coronary artery disease. Revista Española de Cardiologia 2010;63:925–33.
[11] Grau M, Subirana I, Elosua R, et al. Trends in cardiovascular risk factor prevalence (1995–2000–2005) in northeastern Spain. European Journal of Cardiovascular Prevention and Rehabilitation 2007;14: 653–9.
[12] Masia R, Pena A, Marrugat J, et al. High prevalence of cardiovascular risk factors in Gerona, Spain, a province with low myocardial infarction incidence. REGICOR Investigators. Journal of Epidemiology and Community Health 1998;52:707–15.
[13] Dawber TR, Kannel WB. The Framingham Study. An epidemiological approach to coronary heart disease. Circulation 1966;34:553–5.
[14] Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. American Journal of Epidemiology 1979;110: 281–90.
[15] Larson MG, Atwood LD, Benjamin EJ, et al. Framingham Heart Study 100 K project: genome-wide associations for cardiovascular disease outcomes. BMC Medical Genetics 2007;8(Suppl. 1):S5.
[16] Shiffman D, Louie JZ, Rowland CM, et al. Single variants can explain the association between coronary heart disease and haplotypes in the apolipoprotein(a) locus. Atherosclerosis 2010;212:193–6.
[17] The CARDIoGRAM Consortium. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature Genetics 2011;43:333–8.
[18] Grau M, Subirana I, Elosua R, et al. Why should population attributable fractions be periodically recalculated. An example from cardiovascular risk estimation in southern Europe. Preventive Medicine 2010;51: 78–84.
[19] Cupples LA, D'Agostino RB, Kiely D. The Framingham Heart Study, Section 35. An epidemiological investigation of cardiovascular disease. Survival following cardiovascular events: 30 year follow-up. Bethesda, MD: National Heart, Lung and Blood Institute; 1988.
[20] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 2007;81:559–75.
[21] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:177–88.
[22] D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Handbook of statistics advances in survival analysis. Elsevier; 2003. pp. 1–25.
[23] Newson R. Confidence intervals for rank statistics: Somers' D and extensions. Stata Journal 2006;6:309–34.
[24] Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine 2011;30:11–21.
[25] Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: Extension to survival analysis. Statistics in Medicine 2010;30:22–38.
[26] Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Annals of Internal Medicine 2010;152:195–6.
[27] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010, http://www.R-project.org/.
[28] Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium – The Framingham Heart Study. Circulation: Cardiovascular Genetics 2012 [Epub ahead of print].
[29] Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA 2010;303: 631–7.
[30] Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine 2006;355:2631–9.
[31] Ware JH. The limitations of risk factors as prognostic tools. New England Journal of Medicine 2006;355:2615–7.
[32] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

[33] Brautbar A, Ballantyne CM, Lawson K, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study. Circulation: Cardiovascular Genetics 2009;2:279–85.

[34] Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. JAMA 2004;291: 210–5.

[35] Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. New England Journal of Medicine 2008;358:1240–9.

[36] Talmud PJ, Cooper JA, Palmen J, et al. Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. Clinical Chemistry 2008;54:467–74.

[37] Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nature Genetics 2011;43: 339–44.

Please cite this article in press as: Lluis-Ganella C, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis (2012), http://dx.doi.org/10.1016/j.atherosclerosis.2012.03.024

3.6. DISCUSSION

General overview

In a case-control association study [Lluís-Ganella, 2010] and two prospective cohorts studies [Lluís-Ganella, 2011], I have established and validated the association between a multi-locus GRS composed of genetic variants associated with CHD, but not with CVRFs, according to current evidence. Moreover, I have shown that this GRS improves the capacity of the Framingham risk function to predict CHD events, primarily among individuals with intermediate risk. This represents the accomplishment of the three first stages recommended by the AHA [Hlatky, 2009] for the evaluation of novel biomarkers for risk prediction.

Selecting genetic variants independent of cardiovascular risk factors

In our analyses we have selected genetic variants that are robustly associated with CHD but not with classical cardiovascular risk factors. This strategy is based on two main arguments:

i) The variants selected are expected to introduce additional information that is complementary to that provided by the classical cardiovascular risk factors already included in the risk functions. While the mechanisms through which these variants modulate CHD risk are unknown, an understanding of the specific functional mechanism underlying the association between a variant and disease is not necessary in order to use it as a risk marker.

ii) The use of independent variants is expected to reduce multicollinearity among variables of the model, as would be the case where variables representing genetic variants were correlated with those representing CVRFs. Although multicollinearity does not create bias, it results in large standard errors [Dohoo, 1997;Mendenhall, 2011].

We took steps to ensure lack of association between the variants in our GRS and classical cardiovascular risk factors both by reviewing the literature and testing for association in our data sets.

Using the true causal variant

As humans migrated out of Africa, they carried part but not all of the genetic variation that existed in the ancestral population. As a result, the haplotypes seen outside Africa tend to be subsets of the haplotypes observed in African populations [The International HapMap Consortium, 2007], and are relatively conserved in different populations (particularly within ethnicities). The fact that the variants used to construct the GRS in this thesis have been so robustly replicated (several in various ethnicities) indicates that the LD between them and the true functional variant/s is very high. However, we can't be sure that our results will be applicable outwith the Caucasian population. In the case of variants that are characterised by highly conserved local LD patterns in the populations studied, it's possible that the observed common variants simply capture the signal from rarer but stronger functional variants. In this case, the effect sizes of these rarer causal variants would be even higher, such that they may have even higher sensitivity and specificity for disease prediction. However, for the purposes of risk estimation, it is not necessary either to genotype the causal variant nor to understand the mechanism underlying the association in order to use this information provided by the genetic variant for risk assessment. An illustrative example of this fact is the use of sex/gender as one of the most powerful independent predictors of cardiovascular risk, despite the fact that the mechanism underlying this relationship is not completely understood; in this case, sex is simply used as a marker of risk.

Could genetic variation be more powerful for risk estimation than other biomarkers?

The genotyping of genetic variants is less susceptible to error than other laboratory/physician techniques for measuring other cardiovascular risk factors, minimising intra-individual variability in repeated measurements. Moreover, genotypes represent a constant lifetime exposure, in contrast to laboratory measurements, such as lipid levels or blood pressure, which capture only the current exposure. However, there is no evidence to support the hypothesis that genetic information provides a better estimate of risk than the current laboratory tests for CVRFs or other biomarkers. For *Articles*

3 and 4, we also evaluated whether the inclusion of SNPS associated with CVRFs on the basis of GWAS results [Hindorff, 2009] would be better able to predict a cardiovascular event at 10 years instead of using the measured value of the CVRF itself (data not shown). These analyses showed that none of the scores composed of SNPs associated with CVRFs provided a better estimation of risk than the CVRF itself, which may have at least two possible explanations: *i)* the genetic variants discovered do not yet explain enough of the variability of the risk factors; *ii)* the environment is a stronger determinant for the variability of the CVRFs than genetic variability, and therefore genetic variability will not completely capture the variance of the CVRF.

Missing heritability and rare variants

There is increasing awareness of the fact that in addition to primary genetic variation, other types of heritable information, such as epigenetic variation, can be transmitted between generations [Bird, 2007; Bonduriansky, 2009], and therefore that previous heritability estimates may be incorrect [Maher, 2008; Danchin, 2011; Zuk, 2012]. If this is true, the results of GWAS studies may explain even more of the true variance of complex traits that is currently recognized. Further, more causal variants are expected to exist, and to account for a significant part of the missing heritability [Zuk, 2012]. As a result of the technical characteristics of GWAS focused on complex diseases, we might expect that genetic variants yet to be discovered using this approach will be rarer and have stronger effects on risk.

It is not yet clear what might be the potential advantages and disadvantages of using rare variants with stronger effects compared to using common variants with weaker effects in cardiovascular risk assessment, and this question warrants further research. However, in mathematical terms, one of the possible outcomes on the per-allele risk effects of the GRS could be the lost of linearity between the different score categories, as rarer variants with stronger effects would tend to be clustered among individuals who fall at the upper end of the score's range (i.e. rare risk alleles are more likely to be observed among individuals with more risk alleles in general). In *Articles 3* and *4* of this thesis, we observed a relatively linear increase in effect between categories of the score because the genetic variants included in the GRS are mostly common variants and the effect sizes are relatively similar.

Effect of the genetic risk score

One of the strengths of the GRS in terms of its utility for risk assessment is that in many respects it behaves in a similar way to CVRFs in terms of the effect size (β_{GRS}). For example, the per-allele increase in risk was not significantly different between the populations analysed, even though those populations have different basal cardiovascular risk [Marrugat, 2007; D'Agostino, 2008]. This is extremely convenient because fitting a risk function requires at least ~7-8 events for each variable included in the model [Quentin, 2004]. Therefore, if the effect sizes differed between populations, they would need to be re-calculated for each population in which the function was to be implemented, requiring large cohort studies (>130 events for each sex in the case of the Framingham risk function). Furthermore, the use of a GRS in the risk function instead of individual genetic variants allows the inclusion of genetic information for an unlimited number of risk variants, without affecting its mathematical properties.

Target population of genetic risk assessment

As discussed in *Article 4*, the population subgroup that may benefit most from the inclusion of genetic information cardiovascular risk functions is the intermediate risk group. Individuals with intermediate risk are an important target group for two main reasons. First, although this group accounts for ~60% of cardiovascular events at 10 years, the clinical interventions used are less aggressive than in the high risk group, and a distinct set of drugs and other treatments are used [World Health Organization, 2007]. Moreover, clinicians are often unclear about the type of treatment that needs to be prescribed to these individuals. Since this group accounts for ~30% of the total population, more aggressive measures to intervene in all members of this group would probably not be cost-effective, and may create an unacceptable burden of pharmacological side effects [Gerber, 2011]. Second, current genetic evidence still explains only a fraction of the total variance of CHD risk, and therefore it would be unethical to intervene less aggressively in high risk individuals who are re-classified into lower risk groups on the basis of genetic information alone. On the other hand, this would be more acceptable in the intermediate risk group because a change in intervention strategy would be less likely to lead to under-treatment.

Further research

The following areas of further research are a priority:

i) It is widely accepted that GWAS have identified only a small fraction of the genetic variation that explains risk of complex diseases. It is essential that our field continues to search for new risk variants, and to further explore known loci and their mechanisms of action. Studies that evaluate the improvement of risk functions need to be continuously updated as new genetic information comes to light.

ii) The utility of genetic information in the clinical setting has not yet been addressed for most complex diseases. This can be approached in the following ways:

- Evaluate whether the use of genetic information provides a sufficient change in predicted risk to change recommended therapy.
- Evaluation of whether the use of genetic information improves clinical outcomes.
- Assessment of the cost-effectiveness of implementing this technology in the population.
- Assessment of the effect of this information on adherence to drug treatments and healthy lifestyle patterns.

iii) The development of strategies to improve the education and training of health professionals and society in the utility of this type of biomarker is essential for its success in improving health and in minimising misinterpretation of genetic risks (determinism and false security).

4. CONCLUSIONS

i) The rs2234693 variant in the *ESR1* gene is not associated with CHD risk. The inconsistency found between the results of previous studies that have addressed this question can be partly explained by aspects related to the quality of the study [Lluís-Ganella, 2009].

ii) The results of our well-powered study of variation throughout *ESR1* does not support the hypothesis that CHD risk is modulated by either common or uncommon variants in the coding, noncoding, or flanking regions of the gene, either in the general population or in men and women separately [Lucas, 2011].

iii) Our results suggest that a genetic risk score, based on the additive effects of the risk alleles at several genetic loci that are associated with CHD risk in a manner that is independent of CVRFs, is associated with an increased risk of CHD [Lluís-Ganella, 2010].

iv) This GRS has a similar linear effect on risk of CHD events in two populations with distinct basal cardiovascular risk [Lluís-Ganella, 2010; Lluís-Ganella, 2012].

v) The addition of this GRS to classical cardiovascular risk functions improves their capacity to predict CHD/CVD events, compared to the basic risk function, particularly among individuals with intermediate coronary risk [Lluís-Ganella, 2012].

5. BIBLIOGRAPHY

Adeyemo A, Chen G, Zhou J, Shriner D, Doumatey A, Huang H, Rotimi C. 2010. FTO genetic variation and association with obesity in West Africans and African Americans. Diabetes 59:1549-1554.

Alberts B, Johnson AD, Lewis J, Raff M, Roberts K, Walter P. 2008. Molecular biology of the cell. Garland Sience.

American Heart Association. 1973. Coronary risk handbook: estimating risk of coronary heart disease in daily practice.

Anderson E. 2002. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. Breast Cancer Res 4:197-201.

Anderson KM, Wilson PW, Odell PM, Kannel WB. 1991. An updated coronary risk profile. A statement for health professionals. Circulation 83:356-362.

Arber W. 1978. Nobel Lecture: Promotion and limitation of genetic exchange. Nobel Lectures, Physiology or Medicine.

Aschengrau A, Seage GR. 2008. Essentials of Epidemiology in Public Health: Overview of Epidemiologic Study Designs. Jones and Bartlett Publishers.

Assmann G, Cullen P, Schulte H. 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation 105:310-315.

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D and others. 2004. Grading quality of evidence and strength of recommendations. BMJ 328:1490.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Introduction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J Exp Med 79:137-158.

Balleine RL, Hunt SM, Clarke CL. 1999. Coexpression of alternatively spliced estrogen and progesterone receptor transcripts in human breast cancer. J Clin Endocrinol Metab 84:1370-1377.

Barahona A, Ayala F. 2009. El siglo de los genes. Patrones de explicación en genética. Alianza Editorial.

Barrett-Connor E. 1997. Sex differences in coronary heart disease. Why are women so superior? The 1995 Ancel Keys Lecture. Circulation 95:252-264.

Barrett-Connor E, Grady D. 1998. Hormone replacement therapy, heart disease, and other considerations. Annu Rev Public Health 19:55-72.

Bateson W. 1909. Mendel’s principles of heredity. Cambridge [Eng]

University Press

Bath PM, Gray LJ. 2005. Association between hormone replacement therapy and subsequent stroke: a meta-analysis. BMJ 330:342.

Beadle GW, Tatum EL. 1941. Genetic Control of Biochemical Reactions in Neurospora. Proc Natl Acad Sci U S A 27:499-506.

Beral V. 2003. Breast cancer and hormone-replacement therapy in the Million Women Study. Lancet 362:419-427.

Beral V, Banks E, Reeves G. 2002. Evidence from randomised trials on the long-term effects of hormone replacement therapy. Lancet 360:942-944.

Berg JM, Tymoczko JL, Stryer L. 2002. Biochemistry: Important Derivatives of Cholesterol Include Bile Salts and Steroid Hormones. W.H.Freeman and Company.

Bird A. 2007. Perceptions of epigenetics. Nature 447:396-398.

Bonduriansky R, Day T. 2009. Nongenetic Inheritance and Its Evolutionary Implications. Annu Rev Ecol Evol Syst 40:103-125.

Boston Women's Health Book Collective. 2007. Our Bodies, Ourselves: Menopause. Touchstone.

Boulware MI, Mermelstein PG. 2005. The influence of estradiol on nervous system function. Drug News Perspect 18:631-637.

Brandi ML, Becherini L, Gennari L, Racchi M, Bianchetti A, Nacmias B, Sorbi S, Mecocci P, Senin U, Govoni S. 1999. Association of the estrogen receptor alpha gene polymorphisms with sporadic Alzheimer's disease. Biochem Biophys Res Commun 265:335-338.

Bray PF, Larson JC, LaCroix AZ, Manson J, Limacher MC, Rossouw JE, Lasser NL, Lawson WE, Stefanick ML, Langer RD and others. 2008. Usefulness of baseline lipids and C-reactive protein in women receiving menopausal hormone therapy as predictors of treatment-related coronary events. Am J Cardiol 101:1599-1605.

Brouchet L, Krust A, Dupont S, Chambon P, Bayard F, Arnal JF. 2001. Estradiol accelerates reendothelialization in mouse carotid artery through estrogen receptor-alpha but not estrogen receptor-beta. Circulation 103:423-428.

Butts SF, Seifer DB. 2009. 6 office tests to assess ovarian reserve, and what they tell you. Sexuality, Reproduction and Menopause 20. www.obgmanagement.com

Canonico M, Oger E, Plu-Bureau, Conard J, Meyer G, Levesque H, Trillot N, Barrellier MT, Wahl D, Emmerich J and others. 2007. Hormone therapy and venous thromboembolism among postmenopausal women: impact of the route of estrogen administration and progestogens: the ESTHER study. Circulation 115:840-845.

Carmeci C, Thompson DA, Ring HZ, Francke U, Weigel RJ. 1997. Identification of a gene (GPR30) with homology to the G-protein-coupled receptor superfamily associated with estrogen receptor expression in breast cancer. Genomics 45:607-617.

Castagnoli A, Maestri I, Bernardi F, Del Senno L. 1987. PvuII RFLP inside the human estrogen receptor gene. Nucleic Acids Res 15:866-

Chambless LE, Cummiskey CP, Cui G. 2010. Several methods to assess improvement in risk prediction models: Extension to survival analysis. Stat Med 30:22-38.

Charchar FJ, Bloomer LD, Barnes TA, Cowley MJ, Nelson CP, Wang Y, Denniff M, Debiec R, Christofidou P, Nankervis S and others. 2012. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. Lancet 379:915-922.

Chong S, Whitelaw E. 2004. Epigenetic germline inheritance. Curr Opin Genet Dev 14:692-696.

Civeira F. 2004. Guidelines for the diagnosis and management of heterozygous familial hypercholesterolemia. Atherosclerosis 173:55-68.

Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U and others. 2003. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 24:987-1003.

Cook NR. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115:928-935.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628-640.

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. Nat Genet 43:339-344.

Couse JF, Hewitt SC, Bunch DO, Sar M, Walker VR, Davis BJ, Korach KS. 1999a. Postnatal sex reversal of the ovaries in mice lacking estrogen receptors alpha and beta. Science 286:2328-2331.

Couse JF, Korach KS. 1999b. Estrogen receptor null mice: what have we learned and where will they lead us? Endocr Rev 20:358-417.

Craig P, Cooper C, Gunnell D, Haw SJ, Lawson K, Macinttyre S, Ogilvie D, Petticrew M, Reeves B, Sutton M and others. 2012. Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence. Medical Research Council. www.mrc.ac.uk

D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. 2001. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA 286:180-187.

D'Agostino RB, Nam BH. 2003. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. Volume 23:1-25.

D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 117:743-753.

Danchin E, Charmantier A, Champagne FA, Mesoudi A, Pujol B, Blanchet S. 2011. Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. Nat Rev Genet 12:475-486.

Davison SL, Bell R, Donath S, Montalto JG, Davis SR. 2005. Androgen levels in adult females: changes with age, menopause, and oophorectomy. J Clin Endocrinol Metab 90:3847-3853.

Dawber TR, Kannel WB. 1966. The Framingham study. An epidemiological approach to coronary heart disease. Circulation 34:553-555.

Dawber TR, MEADORS GF, MOORE FE, Jr. 1951. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 41:279-281.

Deutscher S, Epstein FH, Kjelsberg MO. 1966. Familial aggregation of factors associated with coronary heart disease. Circulation 33:911-924.

Ding C, Jin S. 2009. High-throughput methods for SNP genotyping. Methods Mol Biol 578:245-254.

Dohoo IR, Ducrot C, Fourichon C, Donald A, Hurnik D. 1997. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. Prev Vet Med 29:221-239.

Drago A, De RD, Serretti A. 2007. Incomplete coverage of candidate genes: a poorly considered bias. Curr Genomics 8:476-483.

Dronamraju K. 1992. Profiles in genetics: Archibald E. Garrod (1857-1936). Am J Hum Genet 51:216-219.

Dubey RK, Imthurn B, Barton M, Jackson EK. 2005. Vascular consequences of menopause and hormone therapy: importance of timing of treatment and type of estrogen. Cardiovasc Res 66:295-306.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Egan KM, Lawson JA, Fries S, Koller B, Rader DJ, Smyth EM, Fitzgerald GA. 2004. COX-2-derived prostacyclin confers atheroprotection on female mice. Science 306:1954-1957.

Elosua R, Lluís-Ganella C, Lucas G. 2009. Research into the genetic component of heart disease: from linkage studies to genome-wide genotyping. Rev Esp Cardiol Supl 9(suplB):24B:38B.

Emberson J, Whincup P, Morris R, Walker M, Ebrahim S. 2004. Evaluating the impact of population and high-risk strategies for the primary prevention of cardiovascular disease. Eur Heart J 25:484-491.

Enmark E, Pelto-Huikko M, Grandien K, Lagercrantz S, Lagercrantz J, Fried G, Nordenskjold M, Gustafsson JA. 1997. Human estrogen receptor beta-gene structure, chromosomal localization, and expression pattern. J Clin Endocrinol Metab 82:4258-4265.

European Heart Network. 2008. European cardiovascular disease statistics. 2008 edition. http://www.ehnheart.org/

Evangelista O, McLaughlin MA. 2009. Review of cardiovascular risk factors in women. Gend Med 6 Suppl 1:17-36.

Evans J. 1997. Epidemiology in practice: disease incidence. J Comm Eye Health 10:60-62.

Expert Panel on Detection EaToHBCiA. 2001. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). JAMA 285:2486-2497.

Falconer, D. S. 1981. Introduction to Quantitative Genetics. Longman.

Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, Cossarizza A and others. 2007. A whole-genome association study of major determinants for host control of HIV-1. Science 317:944-947.

Fontdevila A. 2009. Cent cinquanta anys després de l'origen de les espècies, de Darwin. Treballs de la societat catalana de biologia (filial de l'institut d'estudis catalans).

Fowkes FG, Pell JP, Donnan PT, Housley E, Lowe GD, Riemersma RA, Prescott RJ. 1994. Sex differences in susceptibility to etiologic factors for peripheral atherosclerosis. Importance of plasma fibrinogen and blood viscosity. Arterioscler Thromb 14:862-868.

Garrod AE. 1902. The incidence of alcaptonuria: a study in chemical individuality. Lancet 160:1616-1620. (Croonian Lecture).

Genome.gov. 2011. Genome.gov.

Gerber Y, Melton LJ, III, Weston SA, Roger VL. 2011. Association between myocardial infarction and fractures: an emerging phenomenon. Circulation 124:297-303.

Goodsell D. 2003. Estrogen Receptor. Protein Data Bank 101.

Gordon T, Sorlie P, Kannel WB. 1971. Coronary heart disease, atherothrombotic brain infarction, intermittent claudication - A multivariate analysis of some factors related to their incidence: Framingham Study, 16-year followup, in Kannel WB, Gordon T (eds): The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease, section 27. US Government Printing Office No. 426-1301/1345, Washington, D.C.

Gosden JR, Middleton PG, Rout D. 1986. Localization of the human oestrogen receptor gene to chromosome 6q24----q27 by in situ hybridization. Cytogenet Cell Genet 43:218-220.

Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, Ernster VL, Cummings SR. 1992. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med 117:1016-1037.

Graham I, Atar D, Borch-Johnsen K, Boysen G, Burell G, Cifkova R, Dallongeville J, De Backer G, Ebrahim S, Gjelsvik B and others. 2007. European guidelines on cardiovascular disease prevention in clinical practice: full text. Fourth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). Eur J Cardiovasc Prev Rehabil 14 Suppl 2:S1-113.

Grau M, Elosua R, Cabrera dL, Guembe MJ, Baena-Diez JM, Vega AT, Javier FF, Zorrilla B, Rigo F, Lapetra J and others. 2011. [Cardiovascular risk factors in Spain in the first decade of the 21st Century, a pooled analysis with individual data from 11 population-based studies: the DARIOS study]. Rev Esp Cardiol 64:295-304.

Grau M, Subirana I, Elosua R, Solanas P, Ramos R, Masia R, Cordon F, Sala J, Juvinya D, Cerezo C and others. 2007. Trends in cardiovascular risk factor prevalence (1995-2000-2005) in northeastern Spain. Eur J Cardiovasc Prev Rehabil 14:653-659.

Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. 2000. An Introduction to Genetic Analysis. W.H.Freeman and Company.

Grodstein F, Manson JE, Colditz GA, Willett WC, Speizer FE, Stampfer MJ. 2000. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. Ann Intern Med 133:933-941.

Grodstein F, Stampfer MJ, Colditz GA, Willett WC, Manson JE, Joffe M, Rosner B, Fuchs C, Hankinson SE, Hunter DJ and others. 1997. Postmenopausal hormone therapy and mortality. N Engl J Med 336:1769-1775.

Gruber CJ, Tschugguel W, Schneeberger C, Huber JC. 2002. Production and actions of estrogens. N Engl J Med 346:340-352.

Harman SM, Vittinghoff E, Brinton EA, Budoff MJ, Cedars MI, Lobo RA, Merriam GR, Miller VM, Naftolin F, Pal L and others. 2011. Timing and duration of menopausal hormone treatment may affect cardiovascular outcomes. Am J Med 124:199-205.

Harper PS. 2005. William Bateson, human genetics and medicine. Hum Genet 118:141-151.

Haw SJ, Gruer L, Amos A, Currie C, Fischbacher C, Fong GT, Hastings G, Malam S, Pell J, Scott C and others. 2006. Legislation on smoking in enclosed public places in Scotland: how will we evaluate the impact? J Public Health (Oxf) 28:24-30.

Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP, Herbosa T, Joseph S, Kibatala PL, Lapitan MC and others. 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 360:491-499.

Healthy HK. 2009. Department of health, Government of Hong Kong Special Administrative Region. Coronary heart disease by age group, 2006. www.healthyhk.gov.hk/phisweb/en/healthy_facts/disease_burden/major_causes_death/coronary_heart_disease/.

Hill SM, Fuqua SA, Chamness GC, Greene GL, McGuire WL. 1989. Estrogen receptor expression in human breast cancer associated with an estrogen receptor gene restriction fragment length polymorphism. Cancer Res 49:145-148.

Hindorff LA, Junkins HA, Mehta JP, Manolio TA. 2009. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies.

Hippisley-Cox J, Coupland C, Robson J, Brindle P. 2010. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ 341:c6624.

Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. 2007. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 335:136.

Hirschhorn JN, Gajdos ZK. 2011. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annu Rev Med 62:11-24.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genet Med 4:45-61.

Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, Go AS, Harrell FE, Jr., Hong Y, Howard BV and others. 2009. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 119:2408-2416.

Hodgin JB, Maeda N. 2002. Minireview: estrogen and mouse models of atherosclerosis. Endocrinology 143:4495-4501.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529.

Hudson TJ, Cooper DN. 2009. STREGA: a ‘How-To’ guide for reporting genetic associations. Hum Genet 125:117-118.

Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, Vittinghoff E. 1998. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA 280:605-613.

Human Genome Project. 2011. genomics.energy.gov.

Innerarity TL, Young SG, Poksay KS, Mahley RW, Smith RS, Milne RW, Marcel YL, Weisgraber

KH. 1987. Structural relationship of human apolipoprotein B48 to apolipoprotein B100. J Clin Invest 80:1794-1798.

International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789-796.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. Nat Genet 29:306-9.

Ioannidis JP, Stavrou I, Trikalinos TA, Zois C, Brandi ML, Gennari L, Albagha O, Ralston SH, Tsatsoulis A. 2002. Association of polymorphisms of the estrogen receptor alpha gene with bone mineral density and fracture risk in women: a meta-analysis. J Bone Miner Res 17:2048-2060.

Isles CG, Hole DJ, Hawthorne VM, Lever AF. 1992. Relation between coronary risk and coronary mortality in women of the Renfrew and Paisley survey: comparison with men. Lancet 339:702-706.

Jensen EV. 1962. On the mechanism of estrogen action. Perspect Biol Med 6:47-59.

Jostins L, Barrett JC. 2011. Genetic risk prediction in complex disease. Hum Mol Genet 20:R182-R188.

Kjaergaard AD, Ellervik C, Tybjaerg-Hansen A, Axelsson CK, Gronholdt ML, Grande P, Jensen GB, Nordestgaard BG. 2007. Estrogen receptor alpha polymorphism and risk of cardiovascular disease, cancer, and hip fracture: cross-sectional, cohort, and case-control studies and a meta-analysis. Circulation 115:861-871.

Kobayashi S, Inoue S, Hosoi T, Ouchi Y, Shiraki M, Orimo H. 1996. Association of bone mineral density with polymorphism of the estrogen receptor gene. J Bone Miner Res 11:306-311.

Krege JH, Hodgin JB, Couse JF, Enmark E, Warner M, Mahler JF, Sar M, Korach KS, Gustafsson JA, Smithies O. 1998. Generation and reproductive phenotypes of mice lacking estrogen receptor beta. Proc Natl Acad Sci U S A 95:15677-15682.

Kuiper GG, Carlsson B, Grandien K, Enmark E, Haggblad J, Nilsson S, Gustafsson JA. 1997. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. Endocrinology 138:863-870.

Kuller LH, Meilahn EN, Cauley JA, Gutai JP, Matthews KA. 1994. Epidemiologic studies of menopause: changes in risk factors and disease. Exp Gerontol 29:495-509.

Lander ES. 2011. Initial impact of the sequencing of the human genome. Nature 470:187-197.

Lee SH, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294-305.

Lehnart SE, Ackerman MJ, Benson DW, Jr., Brugada R, Clancy CE, Donahue JK, George AL, Jr., Grant AO, Groft SC, January CT and others. 2007. Inherited arrhythmias: a National Heart, Lung, and Blood Institute and Office of Rare Diseases workshop consensus report about the diagnosis, phenotyping, molecular mechanisms, and therapeutic approaches for primary cardiomyopathies of gene mutations affecting ion channel function. Circulation 116:2325-2345.

Lenfant C. 2010. Chest pain of cardiac and noncardiac origin. Metabolism 59 Suppl 1:S41-S46.

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J and others. 2009. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE Statement. Hum Genet 125:131-151.

Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, Wilson PW, Kannel WB, Zhao D. 2004. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. JAMA 291:2591-2599.

Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PW, Wolf PA, Levy D. 2006. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation 113:791-798.

Lloyd-Jones DM, Wang TJ, Leip EP, Larson MG, Levy D, Vasan RS, D'Agostino RB, Massaro JM, Beiser A, Wolf PA and others. 2004a. Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. Circulation 110:1042-1046.

Lloyd-Jones DM, Wilson PW, Larson MG, Beiser A, Leip EP, D'Agostino RB, Levy D. 2004b. Framingham risk score and prediction of lifetime risk for coronary heart disease. Am J Cardiol 94:20-24.

Lluís-Ganella C, Lucas G, Subirana I, Escurriol V, Tomas M, Senti M, Sala J, Marrugat J, Elosua R. 2009. Qualitative assessment of previous evidence and an updated meta-analysis confirms lack of association between the ESR1 rs2234693 (PvuII) variant and coronary heart disease in men and women. Atherosclerosis 207:480-486.

Lluís-Ganella C, Lucas G, Subirana I, Senti M, Jimenez-Conde J, Marrugat J, Tomas M, Elosua R. 2010. Additive effect of multiple genetic variants on the risk of coronary artery disease. Rev Esp Cardiol 63:925-933.

Lluís-Ganella C, Subirana I, Lucas G, Tomas M, Munoz D, Senti M, Salas E, Sala J, Ramos R, Ordovas JM and others. 2012. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis. Article in press.

Lobo I, Shaw K. 2008. Discovery and Types of Genetic Linkage. Nature Education 1.

Lucas G, Lluís-Ganella C, Subirana I, Senti M, Willenborg C, Musameh MD, Schwartz SM, O'Donnell CJ, Melander O, Salomaa V and others. 2011. Post-Genomic Update on a Classical Candidate Gene for Coronary Artery Disease: ESR1. Circ Cardiovasc Genet

Lynch, M. & Walsh, B. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates.

Maher B. 2008. Personal genomes: The case of the missing heritability. Nature 456:18-21.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590-1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Manson JE, Allison MA, Rossouw JE, Carr JJ, Langer RD, Hsia J, Kuller LH, Cochrane BB, Hunt JR, Ludlam SE and others. 2007. Estrogen therapy and coronary-artery calcification. N Engl J Med 356:2591-2602.

Maouche S, Schunkert H. 2012. Strategies beyond genome-wide association studies for atherosclerosis. Arterioscler Thromb Vasc Biol 32:170-181.

Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. 1994. Genetic susceptibility to death from coronary heart disease in a study of twins. N Engl J Med 330:1041-1046.

Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, Solanas P, Cordon F, Ramos R, Sala J and others. 2003a. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. J Epidemiol Community Health 57:634-638.

Marrugat J, Solanas P, D'Agostino R, Sullivan L, Ordovas J, Cordon F, Ramos R, Sala J, Masia R, Rohlfs I and others. 2003b. Coronary risk estimation in Spain using a calibrated Framingham function. Rev Esp Cardiol 56:253-261.

Marrugat J, Sala J, Manresa JM, Gil M, Elosua R, Perez G, Albert X, Pena A, Masia R. 2004. EPI i GENAcute myocardial infarction population incidence and in-hospital management factors associated to 28-day case-fatality in the 65 year and older. Eur J Epidemiol 19:231-237.

Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, Nam BH, Ramos R, Sala J, Solanas P and others. 2007. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. J Epidemiol Community Health 61:40-47.

Marrugat J, Vila J, Baena-Diez JM, Grau M, Sala J, Ramos R, Subirana I, Fito M, Elosua R. 2011. [Relative validity of the 10-year cardiovascular risk estimate in a population cohort of the REGICOR study]. Rev Esp Cardiol 64:385-394.

Mayer B, Erdmann J, Schunkert H. 2007. Genetics and heritability of coronary artery disease and myocardial infarction. Clin Res Cardiol 96:1-7.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369.

McDevitt MA, Glidewell-Kenney C, Jimenez MA, Ahearn PC, Weiss J, Jameson JL, Levine JE. 2008. New insights into the classical and non-classical actions of estrogen: evidence from estrogen receptor knock-out and knock-in mice. Mol Cell Endocrinol 290:24-30.

Mendelsohn ME, Karas RH. 2005. Molecular and cellular basis of cardiovascular gender differences. Science 308:1583-1587.

Mendenhall W, Sincich T. 2011. A Second Course in Statistics: Regression Analysis.

Molina PE. 2004. Endocrine Physiology: Age related changes in the female reproductive system.

Morgan TH. 1911. Random segregation versus coupling in mendelian inheritance. Science 34:384.

Morton NE. 1982. Outline of Genetic Epidemiology. S Karger Pub

Mosselman S, Polman J, Dijkema R. 1996. ER beta: identification and characterization of a novel human estrogen receptor. FEBS Lett 392:49-53.

Muers M. 2011. Technology: Getting Moore from DNA sequencing. Nat Rev Genet 12:586.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1:263-273.

Murabito JM, D'Agostino RB, Silbershatz H, Wilson WF. 1997. Intermittent claudication. A risk profile from The Framingham Heart Study. Circulation 96:44-49.

Murabito JM, Pencina MJ, Nam BH, D'Agostino RB, Sr., Wang TJ, Lloyd-Jones D, Wilson PW, O'Donnell CJ. 2005. Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. JAMA 294:3117-3123.

Murphy E. 2011. Estrogen signaling and cardiovascular disease. Circ Res 109:687-696.

Myocardial Infarction Genetics Consortium. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat

Genet 41:334-341.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387-389.

Nelson HD, Humphrey LL, Nygren P, Teutsch SM, Allan JD. 2002. Postmenopausal hormone replacement therapy: scientific review. JAMA 288:872-881.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61-80.

Niino M, Kikuchi S, Fukazawa T, Yabe I, Tashiro K. 2000. Estrogen receptor gene polymorphism in Japanese patients with multiple sclerosis. J Neurol Sci 179:70-75.

Nilsson S, Makela S, Treuter E, Tujague M, Thomsen J, Andersson G, Enmark E, Pettersson K, Warner M, Gustafsson JA. 2001. Mechanisms of estrogen action. Physiol Rev 81:1535-1565.

North American Menopause Society. 2003. Amended report from the NAMS Advisory Panel on Postmenopausal Hormone Therapy. Menopause 10:6-12.

North American Menopause Society. 2007. Estrogen and progestogen use in peri- and postmenopausal women: March 2007 position statement of The North American Menopause Society. Menopause 14:168-182.

Ochoa S. 1959. Nobel Lecture: Enzymatic Synthesis of Ribonucleic Acid. Nobel Lectures, Physiology or Medicine.

Owman C, Blay P, Nilsson C, Lolait SJ. 1996. Cloning of human cDNA encoding a novel heptahelix receptor expressed in Burkitt's lymphoma and widely distributed in brain and peripheral tissues. Biochem Biophys Res Commun 228:285-292.

Palmieri C, Cheng GJ, Saji S, Zelada-Hedman M, Warri A, Weihua Z, Van Noorden S, Wahlstrom T, Coombes RC, Warner M and others. 2002. Estrogen receptor beta in breast cancer. Endocr Relat Cancer 9:1-13.

Pare G, Krust A, Karas RH, Dupont S, Aronovitz M, Chambon P, Mendelsohn ME. 2002. Estrogen receptor-alpha mediates the protective effects of estrogen against vascular injury. Circ Res 90:1087-1092.

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570-575.

Pearson ML, Soll D. 1991. The Human Genome Project: a paradigm for information management in the life sciences. FASEB J 5:35-39.

Pearson TA, Blair SN, Daniels SR, Eckel RH, Fair JM, Fortmann SP, Franklin BA, Goldstein LB, Greenland P, Grundy SM and others. 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. American Heart Association Science Advisory and Coordinating Committee. Circulation 106:388-391.

Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 30:11-21.

Pendaries C, Darblade B, Rochaix P, Krust A, Chambon P, Korach KS, Bayard F, Arnal JF. 2002. The AF-1 activation-function of ERalpha may be dispensable to mediate the effect of estradiol on endothelial NO production in mice. Proc Natl Acad Sci U S A 99:2205-2210.

Pines A, Sturdee DW, Birkhauser MH, Schneider HP, Gambacciani M, Panay N. 2007. IMS updated recommendations on postmenopausal hormone therapy. Climacteric 10:181-194.

Plomin R, Haworth CM, Davis OS. 2009. Common disorders are quantitative traits. Nat Rev Genet 10:872-878.

Ponglikitmongkol M, Green S, Chambon P. 1988. Genomic organization of the human oestrogen receptor gene. EMBO J 7:3385-3388.

Quentin LB. 2004. Statistical Rules of Thumb. Journal Of The Royal Statistical Society Series A 167:184-185.

Revankar CM, Cimino DF, Sklar LA, Arterburn JB, Prossnitz ER. 2005. A transmembrane intracellular estrogen receptor mediates rapid cell signaling. Science 307:1625-1630.

Ridker PM, Buring JE, Rifai N, Cook NR. 2007. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA 297:611-619.

Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. 2008. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation 118:2243-2251.

Roberts L. 2001. The human genome. Controversial from the start. Science 291:1182-1188.

Rosengren A, Dotevall A, Eriksson H, Wilhelmsen L. 2001. Optimal risk factors in the population: prognosis, prevalence, and secular trends; data from Goteborg population studies. Eur Heart J 22:136-144.

Ross R. 1999. Atherosclerosis--an inflammatory disease. N Engl J Med 340:115-126.

Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC and others. 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA 288:321-333.

Rossouw JE, Prentice RL, Manson JE, Wu L, Barad D, Barnabei VM, Ko M, LaCroix AZ, Margolis KL, Stefanick ML. 2007. Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. JAMA 297:1465-1477.

Rothman KJ, Greenland S, Lash TL. 2008. Modern Epidemiology: Types of Epidemiologic Studies. third edition:87-99.

Salpeter SR, Cheng J, Thabane L, Buckley NS, Salpeter EE. 2009. Bayesian meta-analysis of hormone therapy and mortality in younger postmenopausal women. Am J Med 122:1016-1022.

Salpeter SR, Walsh JM, Greyber E, Salpeter EE. 2006. Brief report: Coronary heart disease events associated with hormone therapy in younger and older women. A meta-analysis. J Gen Intern Med 21:363-366.

Sans S, Fitzgerald AP, Royo D, Conroy R, Graham I. 2007. Calibrating the SCORE cardiovascular risk chart for use in Spain. Rev Esp Cardiol 60:476-485.

Schisterman EF, Perkins NJ, Liu A, Bondell H. 2005. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 16:73-81.

Seshadri S, Beiser A, Kelly-Hayes M, Kase CS, Au R, Kannel WB, Wolf PA. 2006. The lifetime risk of stroke: estimates from the Framingham Study. Stroke 37:345-350.

Shearman AM, Cooper JA, Kotwinski PJ, Humphries SE, Mendelsohn ME, Housman DE, Miller GJ. 2005. Estrogen receptor alpha gene variation and the risk of stroke. Stroke 36:2281-2282.

Shearman AM, Cooper JA, Kotwinski PJ, Miller GJ, Humphries SE, Ardlie KG, Jordan B, Irenze K, Lunetta KL, Schuit SC and others. 2006. Estrogen receptor alpha gene variation is associated with risk of myocardial infarction in more than seven thousand men from five cohorts. Circ Res 98:590-592.

Sigelman CK, Rider EA. 2009. Health and phisical development: Female menopause. 7th:164-165.

Singleton AB. 2011. Exome sequencing: a transformative technology. Lancet Neurol 10:942-946.

Smith EP, Boyd J, Frank GR, Takahashi H, Cohen RM, Specker B, Williams TC, Lubahn DB, Korach KS. 1994. Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. N Engl J Med 331:1056-1061.

SPH U. 2012. University of Michigan School of Public Health website. http://www.sph.umich edu/epid/GSS/pub.html

Stamler J, Stamler R, Neaton JD, Wentworth D, Daviglus ML, Garside D, Dyer AR, Liu K, Greenland P. 1999. Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy: findings for 5 large cohorts of young adult and middle-aged men and women. JAMA 282:2012-2018.

Stampfer MJ, Hu FB, Manson JE, Rimm EB, Willett WC. 2000. Primary prevention of coronary heart disease in women through diet and lifestyle. N Engl J Med 343:16-22.

Stary HC, Chandler AB, Dinsmore RE, Fuster V, Glagov S, Insull W, Jr., Rosenfeld ME, Schwartz CJ, Wagner WD, Wissler RW. 1995. A definition of advanced types of atherosclerotic lesions and a histological classification of atherosclerosis. A report from the Committee on Vascular Lesions of the Council on Arteriosclerosis, American Heart Association. Circulation 92:1355-1374.

Steyerberg EW, Van Calster B, Pencina MJ. 2011. [Performance measures for prediction models and markers: evaluation of predictions and classifications]. Rev Esp Cardiol 64:788-794.

Sudhir K, Chou TM, Chatterjee K, Smith EP, Williams TC, Kane JP, Malloy MJ, Korach KS, Rubanyi GM. 1997. Premature coronary artery disease associated with a disruptive mutation in the estrogen receptor gene in a man. Circulation 96:3774-3777.

Taylor HS, Manson JE. 2011. Update in hormone therapy use in menopause. J Clin Endocrinol Metab 96:255-264.

the CARDIoGRAM Consortium. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43:333-338.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

“The Nobel Prize in Chemistry 2006 - Advanced Information”.Nobelprize.org. 2012. Nobel Lecture: Nucleic acid synthesis in the study of the genetic code. http://www nobelprize org/nobel_prizes/chemistry/laureates/2006/advanced html

Thomas CB, Cohen BH. 1955. The familial occurrence of hypertension and coronary artery disease, with observations concerning obesity and diabetes. Ann Intern Med 42:90-127.

Thygesen K, Alpert JS, White HD. 2007. Universal definition of myocardial infarction. Eur Heart J 28:2525-2538.

Toran-Allerand CD. 2004. Minireview: A plethora of estrogen receptors in the brain: where will it end? Endocrinology 145:1069-1074.

Truett J, Cornfield J, Kannel W. 1967. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 20:511-524.

Tunstall-Pedoe H, Kuulasmaa K, Mahonen M, Tolonen H, Ruokokoski E, Amouyel P. 1999. Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. Lancet 353:1547-1557.

U.S.Preventive Services Task Force. 2002. Postmenopausal hormone replacement therapy for primary prevention of chronic conditions: recommendations and rationale. Ann Intern Med 137:834-839.

Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, Duque-Oliart A. 2000. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. Circulation 101:2572-2578.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA and others. 2001. The sequence of the human genome. Science 291:1304-1351.

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9:255-266.

Visscher PM, Montgomery GW. 2009. Genome-wide association studies and human disease: from trickle to flood. JAMA 302:2028-2029.

von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 18:800-804.

Walter P, Green S, Greene G, Krust A, Bornert JM, Jeltsch JM, Staub A, Jensen E, Scrace G, Waterfield M and others. 1985. Cloning of the human estrogen receptor cDNA. Proc Natl Acad Sci U S A 82:7889-7893.

Wang TJ. 2011. Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction. Circulation 123:551-565.

Wathen CN, Feig DS, Feightner JW, Abramson BL, Cheung AM. 2004. Hormone replacement therapy for the primary prevention of chronic diseases: recommendation statement from the Canadian Task Force on Preventive Health Care. CMAJ 170:1535-1537.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737-738.

Watson JD, Jordan E. 1989. The Human Genome Program at the National Institutes of Health. Genomics 5:654-656.

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.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT and others. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876.

White E, Hunt JR, Casso D. 1998. Exposure measurement in cohort studies: the challenges of prospective data collection. Epidemiol Rev 20:43-56.

White PD. 1957. Genes, the heart and destiny. N Engl J Med 256:965-969.

White R, Lees JA, Needham M, Ham J, Parker M. 1987. Structural organization and expression of the mouse estrogen receptor. Mol Endocrinol 1:735-744.

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. 1998. Prediction of coronary heart disease using risk factor categories. Circulation 97:1837-1847.

Wingard DL, Suarez L, Barrett-Connor E. 1983. The sex differential in mortality from all causes and ischemic heart disease. Am J Epidemiol 117:165-172.

Women's Health Initiative. 1998. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials 19:61-109.

World Health Organization. 2007. Prevention of cardiovascular disease: guideline for assessment and management of cardiovascular risk. 20-26.

World Health Organization. 2009. GLOBAL HEALTH RISKS. Mortality and burden of disease attributable to selected major risks. 28-29.

Wray NR, Visscher PM. 2008. Estimating trait heritability. Nature Education.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 24:1193-1198.

6. GLOSSARY

Alkaptonuria: Is a disease with several symptoms, of which the most conspicuous is that the urine turns black when exposed to air. In 1898, an English doctor named Archibald Garrod showed that the substance responsible is homogentisic acid, which is excreted in abnormally large amounts into the urine of alkaptonuria patients. In 1902, early in the post-Mendelian era, Garrod suggested, on the basis of pedigree patterns, that alkap-tonuria is inherited as a Mendelian recessive [Griffiths, 2000].

Complex disease: Complex diseases are common disorders that are believed to have many causes i.e. cancer, coronary heart disease, diabetes mellitus, hypertension, bipolar disorder or schizophrenia. (source: medical-dictionary.thefreedictionary.com).

Coronary Heart Disease (CHD): A heart disease due to an abnormality of the arteries that supply blood and oxygen to the heart. (source: medical-dictionary.thefreedictionary.com).

Genetic architecture: The differences observed between the individuals genomes (less than 0.1% of an individual's sequence ~3x10^{9} DNA base pairs [Human Genome Project, 2011]) are those that provide a big part of the phenotypic differences between humans. Although there are many types of genetic variations described, the only ones that are evaluated in this thesis are those named single nucleotide polymorphisms.

Genome wide association study (GWAs): Is an examination of all or most of the genes (the genome) of different individuals of a particular species to see how much the genes vary from individual to individual.

Haplotype: Combination of alleles at adjacent locations on the chromosomes that are transmitted together.

HapMap Project, The International: The International HapMap Project is an organization whose goal is to develop a haplotype map (HapMap) of the human genome, which will describe the common patterns of human genetic variation. The HapMap is expected to be a key resource for researchers to find genetic variants affecting health, disease and responses to drugs and environmental factors. The information produced by the project is made freely available to researchers around the world. The International HapMap Project is a collaboration among researchers at academic centers, non-profit biomedical research groups and private companies in Canada, China, Japan, Nigeria, the United Kingdom, and the United States. It officially started with a meeting on October 27 to 29, 2002, and was expected to take about three years. It comprises two phases; the complete data obtained in Phase I were published on 27 October 2005. The analysis of the Phase II dataset was published in October 2007. The Phase III dataset was released in spring, 2009 (source: medical-dictionary.thefreedictionary.com).

Hardy-Weinberg Equilibrium (HWE): It states that both allele and genotype frequencies in a population remain constant (that is, they are in equilibrium) from generation to generation unless specific disturbing influences are introduced. Those disturbing influences include non-random mating, mutations, selection, limited population size, random genetic drift, gene flow and meiotic drive.

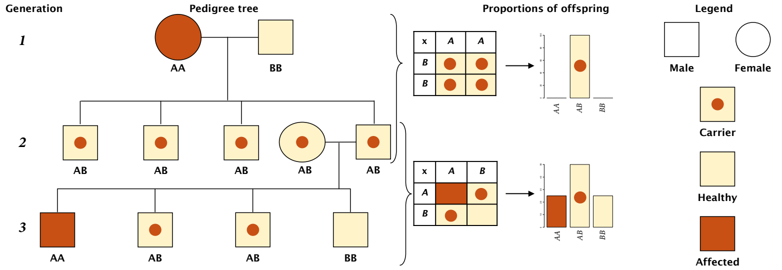
Human Genome Project (HGP): The Human Genome Project is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint. The project began in 1990 and was initially headed by Ari Patrinos, head of the Office of Biological and Environmental Research in the U.S. Department of Energy's Office of Science. Francis Collins directed the National Institutes of Health National Human Genome Research Institute efforts. A working draft of the genome was announced in 2000 and a complete one in 2003, with further, more detailed analysis still being published. A parallel project was conducted outside of government by the Celera Corporation, which was formally launched in 1998. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Japan, France, Germany, and China. The mapping of human genes is an important step in the development of medicines and other aspects of health care (source: medical-dictionary.thefreedictionary.com).

Incidence: Is the probability of developing a particular disease during a given period of time; the numerator is the number of new cases during the specified time period and the denominator is the population at risk during the period (source: medical-dictionary.thefreedictionary.com).

Intervention study: Testing an hypothesized epidemiological cause-effect relationship by intervening in a population and modifying a supposed causal factor and measuring the effect of the change. (source: medical-dictionary.thefreedictionary.com).

Linkage disequilibrium (LD): Is the non-random association of alleles at two or more loci.

Mendelian disease: Diseases in which the phenotypes are largely determined by the action, lack of action, of mutations at individual loci. Example of inheritance of a disease that followed a mendelian inheritance pattern.



In this figure a disease-free individual mates another individual affected of a disease with recessive inheritance (in which the presence of two copies of genetic variation is necessary to cause the disease. If only one copy is present, the disease is not expressed and the individual will only be carrier of the disease). All of the offspring (generation 2) will be carrying the disease without suffering it. If one of the individuals of the second generation mates with another individual carrier of the same disease, 50% of the offspring (generation 3) will be also carrier, 25% will be free of the disease and 25% will suffer the disease.

Observational studies: An observational study draws inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator. This is in contrast with controlled experiments, such as randomized controlled trials, where each subject is randomly assigned to a treated group or a control group before the start of the treatment.

One gene-one enzyme hypothesis: The one gene-one enzyme hypothesis is the idea that genes act through the production of enzymes, with each gene responsible for producing a single enzyme that in turn affects a single step in a metabolic pathway.

Penetrance: Penetrance in genetics is the proportion of individuals carrying a particular variation of a gene (allele or genotype) that also express an associated trait (phenotype).

Phenotype: Is an organism's observable characteristics or traits: such as its morphology, development, biochemical or physiological properties, behaviour, and products of behaviour.

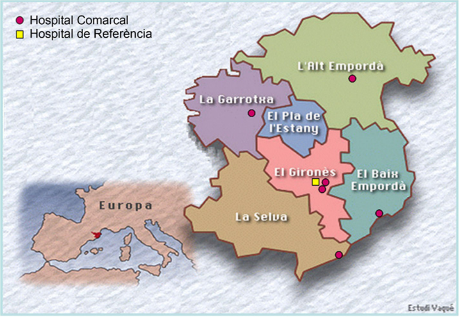
Single nucleotide polymorphisms (SNPs): SNPs are DNA sequence variants occurring when a nucleotide (A, T, C, or G) differs between members of a biological species. For example, two DNA fragments from different individuals, AAGC**C**TA to AAGC**T**TA, contain a difference in a single nucleotide, a SNP. In some cases, these genetic variations can cause diseases by modifying the proteins they code for, by modifying transcription binding sites, or by many other causes (known and unknown causes).

Tag SNPs: Are representative SNPs in a region of the genome with high linkage disequilibrium (the non-random association of alleles at two or more loci). It is possible to identify genetic variation without genotyping every SNP in a chromosomal region. Tag SNPs are useful in whole-genome SNP association studies in which hundreds of thousands of SNPs across the entire genome are genotyped.

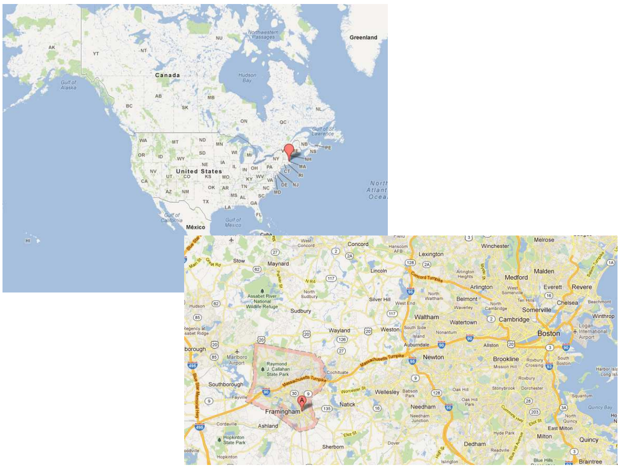
7. APPENDICES

7.1. APPENDIX: Brief description of the cohorts and studies used in the present doctoral thesis

Registre Gironí del Cor (REGICOR) [Grau, 2007]: The REGICOR study is clinical and epidemiology project, both hospital and population based, conducted in the Girona area (Catalunya, Spain) which principal objective is to evaluate the magnitude of ischaemic heart disease and the associated risk factors at population scale, while also monitoring the utilisation of health care resources and the long-term prognosis for this disease and its risk factors. The area of reference is covered by a hospital network that includes Hospital Universitari de Girona Dr. Josep Trueta (reference hospital), Hospital de Figueres, Hospital de Palamós, Hospital Sant Jaume d'Olot, Hospital Santa Caterina de Girona, Hospital Comarcal de la Selva de Blanes, Clínica Girona and Clínica de l'Aliança. The REGICOR study contains data from three cross-sectional studies of cardiovascular risk factors, conducted in 1995 (N=1,748), 2000 (N=3,058) and 2005 (N=6,500). This registry also contains information of all consecutive patients who had undergone a coronary event in the Hospital Universitari de Girona Dr. Josep Trueta from the year 1978.



Framingham Heart Study (FHS) [Dawber, 1951; Dawber, 1966]: The objective of the FHS was to identify the common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. The researchers recruited 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts, and began the first round of extensive physical examinations and lifestyle interviews that they would later analyze for common patterns related to CVD development. Since 1948, the subjects have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests, and in 1971, the Study enrolled a second generation - 5,124 of the original participants' adult children and their spouses - to participate in similar examinations. In 1994, the need to establish a new study reflecting a more diverse community of Framingham was recognized, and the first Omni cohort of the Framingham Heart Study was enrolled. In April 2002 the Study entered a new phase, the enrolment of a third generation of participants, the grandchildren of the Original Cohort. In 2003, a second group of Omni participants was enrolled.

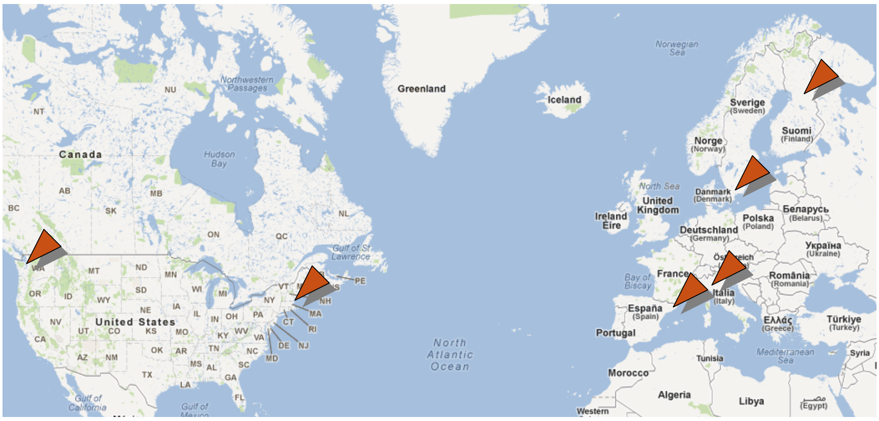


Coronary ARtery DIsease Genome-wide Replication And Meta-analysis (CARDIoGRAM) [The CARDIoGRAM Consortium, 2011]: The CARDIoGRAM consortium was formed with the purpose of identifying novel susceptibility loci for CAD and MI. Briefly, the CARDIoGRAM discovery analysis combined data from 14 published and unpublished primary GWA studies, in individuals of European ancestry, including >22 000 cases with CAD, MI, or both and >60 000 controls.

Wellcome Trust Case-Control Consortium (WTCCC) [Wellcome Trust Case Control Consortium, 2007]: is a group of 50 research groups across the UK which was established in 2005. The WTCCC aims were to exploit progress in understanding of patterns of human genome sequence variation along with advances in high-throughput genotyping technologies, and to explore the utility, design and analyses of genome-wide association (GWA) studies.

Myocardial Infarction Genetics Consortium (MIGen) [Myocardial Infarction Genetics Consortium, 2009]: Is a collection of 2,967 cases of early onset myocardial infarction (in men ≤50 y old or women ≤60 y old) and 3,075 age- and sex-matched controls free of myocardial infarction from six international sites: Boston and Seattle in the United States, as well as Sweden, Finland, Spain and Italy.

.



7.2. APPENDIX: Article 1: Supplementary material

Supplementary Materials & Methods

In order to assess current evidence in favour of an association between genetic variation in *ESR1* and CHD, an extensive literature search was carried out, the resulting articles were then reviewed, association data was extracted and qualitative and quantitative analyses of this evidence were performed. The experimental procedures used are described in detail in the following sections.

1. Literature search

Three different approaches were used to identify articles containing information about the association between genetic variation in *ESR1* and CHD.

1.1 Literature database search

Articles of interest were obtained from the PUBMED (ncbi.nlm.nih.gov/pubmed/) database using a structured (Boolean) search strategy (*Supplementary Table 2*), with search terms falling into three main categories: limits (e.g. date of publication between 1985 and December 2008, humans, etc.); medical search terms; and genetic search terms. The MeSH terms (Medical Subject Heading: National Library of Medicine's controlled vocabulary thesaurus) database was used, which consists of sets of descriptors in a hierarchical structure that permits searching at various levels of specificity.

1.2 Subject review articles

In order to identify articles that are generally considered to be important by experts in the subject area, and that may have been not captured by the Boolean search strategy, data was collected from recent relevant reviews. These review articles were obtained by searching both PubMed and The Cochrane Library (www.update-software.com/Clibplus/Clibplus.asp; Wiley InterScience) for reviews containing the terms "cardiovascular disease" and "estrogen receptor". The bibliographies of these review articles were examined to identify additional articles relevant to the topic of study.

1.3 Retrospective/Prospective search

In order to identify other relevant articles that were not identified by the two previous strategies, a prospective and retrospective search was carried out for each article [*Article N*] that passed all steps of the review process described in *Supplementary Table 3*. For the retrospective search, the references in each *Article N* were reviewed to identify any other relevant articles that were cited by that *Article N* and that were not found by the other search strategies. For the prospective search, relevant articles that subsequently cited *Article N* were identified using the "Cited Reference Search" tool from the ISI Web of Knowledge database (www.isiwebofknowledge.com/).

Articles identified in this way were then subjected to the same three stage review process as articles identified by the other search strategies.

2. Article revision

In order to confirm the relevance of articles to our topic of interest, a common review protocol was applied to each one. Each article was reviewed by two blinded independent reviewers in three stages, as summarised in *Supplementary Table 3*.

3. Data Extraction and Analysis

In order to assess the strength of the evidence reported to date on the association between genetic variation in *ESR1* and risk of CHD, data was extracted from the articles collected above using a common data extraction protocol, and qualitative and quantitative analyses were performed.

3.1 Qualitative analysis

The collection and analysis of data from these articles was carried out according to a) recently published guidelines (Venice Criteria) for classifying the quality and quantity of reported evidence for a particular association [1], and b) a report from the NCI-NHGRI Working Group on Replication in Association Studies which suggests minimum standards for the publication of genetic association studies [2].

3.1.1 Cumulative evidence analysis

Ioannidis *et al.* [1] proposed a classification system for assessing the credibility of cumulative evidence on a reported genetic association, based on three criteria: amount of evidence, replication, and protection from bias. This classification consists of a 3-letter (A, B or C) code to classify the type and quality of the evidence. The first letter relates to the amount of evidence in favour of a genetic association, depending on the cumulative sample size of the least common genetic group of interest. The second letter describes how well a reported association has been replicated (e.g. if it has been consistently replicated or not). The third letter describes the likelihood of an important bias in the association studies reported. With this classification it is possible to evaluate the robustness and credibility of a particular genetic association. We used this approach to evaluate the cumulative evidence of the association studies identified in our systematic literature search.

3.1.2 Individual study analysis

A questionnaire based on guidelines proposed by the NCI-NHGRI Working Group on Replication in Association Studies [2] was used to assess the quality of evidence provided by each article. Briefly, this questionnaire contained questions related to the following issues: study information; data issues; genotyping and quality control procedures; results; replication studies; genotyping data and specifications for deposition in standard databases; and points for reviewers and authors to consider regarding priority for publication. This questionnaire was applied by three independent reviewers (CL, GL, RE) to each of the articles identified above, with the general aim of evaluating how much of the information required by the guidelines was provided in these articles. After each reviewer had filled the questionnaire for each article, discrepancies were resolved in a second stage of revision. For each study/article, each question/condition was evaluated as 1 (Yes) where a given requirement was met and 0 (No) otherwise.

3.2 Quantitative analysis: meta-analysis

An update of a previous meta-analysis [3] of association studies was performed for the most widely studied genetic variant in the *ESR1* gene for which data was reported in the articles mentioned above.

4. Association analysis and meta-analysis of rs9340799 (XbaI) polymorphism in ESR1

4.1 Materials and methods

The rs9340799 variant was genotyped in the same sample as the rs2234693 variant (see *Materials and Methods*). TaqMan primers and probes are shown in *Supplementary Table 1*. Genotype frequencies in cases and controls were compared using a 1df χ^2 test (two separate tests: common homozygotes (GG) versus heterozygotes (AG); common homozygotes (GG) versus rare homozygotes (AA)).

The same protocols for the literature search, article revision, data extraction, association analysis and meta-analysis were used for this variant as for the rs2234693 variant.

4.2 Results

No significant association was observed between rs9340799 and MI in the REGICOR study (AA vs. GG OR 1.13 (95%CI 0.78-1.64; p=0.507); AG vs. GG OR 1.26 (95%CI 0.86-1.85; p=0.208); Table 1). These results remained non-significant after stratifying the analysis by gender (data not shown).

This genetic variant achieved a classification of ACB (weak evidence) under the criteria proposed by Ioannidis *et al.* [1].

Quality Scores for articles that studied the association between this variant and CHD are shown in

Supplementary Table 5.

The meta-analysis of rs9340799 association studies showed no association between this variant and CHD for both tests: AA vs. GG OR 1.03 (95%CI 0.81-1.32, p=0.808; range [0.993-1.082]; χ^2=10.94 (p=0.2049)) and AG vs. GG OR 1.02 (95%CI 0.88-1.17, p=0.829; range [0.978-1.034]; χ^2=7.34 (p=0.5005); *Supplementary Fig 2*). No significant between-study heterogeneity was observed for this variant.

5. Supplementary tables

Supplementary Table 1: Primers and probes used in the genotyping of the genetic variants in *ESR1* gene (rs2234693:*PvuII* and rs9340799:*XbaI*).

SNP	primers		probes	
rs2234693	forward	TTCCCAGAGACCCTGAGTGT	VIC	CTCATCCCAACTCTAG-MGB
	reverse	GCAGGAATATACAATTATTTCAGAACCATTAGAGA	FAM	CTCATCCCAACTCCAG-MGB
rs9340799	forward	TCTGTGTTGTCCATCAGTTCATCTG	VIC	ACAAAGCATAAAACAGCTG-MGB
	reverse	CTCAGGGTCTCTGGGAAACAG	FAM	ACAAAGCATAAAACGGCTG-MGB

Supplementary Table 2: Description of the Boolean search strategy and description of each one of the terms included.

Species: Humans[MeSH] AND
Gene names: estrogen receptor AND
Genetic terms: (“gene” OR “genes” OR “genetic” OR “genetics” OR “exon” OR “exons” OR “intron” OR “introns” OR “polymorphism” OR “polymorphisms” OR “single nucleotide polymorphism” OR “single nucleotide polymorphisms” OR “snp” OR “snps” OR “restriction fragment length polymorphism” OR “restriction fragment length polymorphisms” OR “rflp” OR “rflps” OR “allele” OR “alleles” OR “codon” OR “codons” OR “untranslated region” OR “untranslated regions” OR “microsatellite” OR “microsatellites” OR”mutation” OR “mutations” OR “mutant´” OR “mutants” OR “copy number variant”) AND
Date: 1985:2008[DP] AND
Medical terms: (“ischemic heart disease” [TIAB] OR “coronary heart disease” [TIAB] OR “acute coronary syndrome” [TIAB] OR “STEMI” [TIAB] OR NSTEMI [TIAB] OR “myocardial infarction” [TIAB] OR angina [TIAB] OR “angor pectoris” [TIAB] OR “angiography” [TIAB] OR “angioplasty” [TIAB] OR revascularization [TIAB] OR CAD [TIAB] OR “coronary artery disease” [TIAB])

Supplementary Table 3: Review protocol applied to articles identified by database searches.

Review Stage	Article Sections reviewed	Decisions applied	Next step for discrepancies between reviewers
First	Title	Include, Exclude, Undecided	Just those articles evaluated as “exclude” by both reviewers were removed from the review process. All the rest were submitted to the rest of the revision process.
Second	Title & Abstract	Include, Exclude, Undecided	Just those articles evaluated as “exclude” by both reviewers were removed from the review process. All the rest were submitted to the rest of the revision process.
Third	Full article	Include, Exclude	Discussion to resolve discrepancies

Supplementary Table 4: Quality score results, by study.

Question/Condition		Matsubara 1997 [4]	Alevizaki 2007 [5]	Almeida 2006 [6]	Mansur 2005 [7]	Yilmaz 2007 [8]	Xu 2008 [9]	Koch 2005 [10]	Shearman 2003 [11]	Schuit 2004 [12]	Shearman 2006 [13]	Kjaergaard 2007 [3]	Morgan 2007 [14]	*Current study 2009*
	Author (Publication year) Reference													
Study information														
1	A detailed description of the study design and its implementation	●	●	●	●	●	●	●	●	●		●	●	●
2	The source of cases and controls or cohort members, if based on cohort design	●	●	●		●	●	●	●	●	●	●	●	●
3	Methods for ascertaining and validating affected or unaffected status and reproducibility of classification	●	●	●		●	●	●	●	●	●	●	●	●
4	Participation rates for cases, controls or cohort members		●						●	●		●	●	
5	Presentation of case and control selection in a flow chart		●						●			●		
6	Initial table comparing relevant characteristics of cases and controls	●		●	●	●	●	●		●		●	●	●
7	Success rate for DNA acquisition											●		
Data issues														
8	Statement on availability of results and data													
9	Links to supplemental online resources and database accession numbers										●	●		
Genotyping and quality control procedures														
10	Sample tracking methods, such as barcoding, to ensure accuracy of analysis													
11	Description of genotyping assays and protocols	●	●	●	●	●	●	●	●	●		●		●
12	Description of genotyping calling algorithm													
13	Genotype quality control design for samples		●					●			●	●	●	
14	External control samples from standard accepted sets (such as HapMap)													
15	Internal control samples		●					●			●	●	●	
16	Assay and DNA quality metrics by locus, sample, plate or 'batch'													
17	Assay call rates												●	
18	Average error rates estimated by internal duplicates or external samples		●					●				●	●	
19	Assay reproducibility: concordance for performance of extraction, aliquoting and assay reproducibility							●				●	●	
20	Concordance with published or previously generated genotypes													
21	Mendelian consistency checks if related individuals are present													
22	Detection of inconsistent or cryptic relatedness in study subjects													
23	Evaluation of deviations from Hardy–Weinberg proportions separately in cases and controls	●	●	●	●	●	●	●	●	●	●		●	●
24	Assessment of population heterogeneity, including													
25	Average or median value of chi-square and full distribution													
26	Q–Q plots of chi-square analysis and *P*values													
27	Validation of most critical results on an independent genotyping platform							●				●		
Results														
28	Analysis methods in sufficient detail to reconstruct the analytical approach	●	●	●	●	●	●	●	●	●		●	●	●
29	Description of any pre-analysis weighting scheme for selecting variants for replication													
30	Simple single-locus and multi-marker (haplotype) association analyses		●	●				●		●				
31	Genetic models tested	●	●	●		●	●	●	●	●	●	●	●	●
32	Graphical display of genotype clustering for assays of high interest													
33	Verification of results at highly correlated loci													
34	Discussion of choice of threshold for significance	●			●	●		●	●	●		●	●	●
35	Significance of any known 'positive controls'													
36	Consistency of results before and after application of quality control filters													
Replication studies														
37	Description of replication samples, including source, ascertainment and comparability to initial sample													
38	Discussion of choice of threshold for significance													
39	Summary of replication and analysis attempts by authors											●		
40	Summary of all known replication attempts by others, including non-replications					●	●	●	●			●		●
Genotyping data and specifications for deposition in standard databases														
41	Availability of 'raw' genotype data in the technology and vendor format													
42	Data extraction and processing protocols													
43	Normalization, transformation and data selection procedures and parameters													
43 (total)	Total number of positive answers for each individual study	9	13	9	6	10	9	16	11	11	7	19	14	10

Supplementary Table 5: Data extracted from association studies that reported an association between rs9340799 (*XbaI*) or rs2234693 (*PvuII*) in *ESR1* gene and coronary heart disease.

							cases
	Author	PMID	QS	Outcome	Gender	Study design	Number
rs9340799 (*XbaI*)	Matsubara [4]	9409287	9	coronary heart disease	men & women	case/control	75
	Shearman [11]	14600184	11	myocardial infarction	men & women	cohort	58
	Koch [10]	16203927	16	myocardial infarction	men & women	case/control	3657
	Mansur [7]	16099331	6	coronary heart disease	men & women	case/control	153
	Almeida [6]	16612467	9	coronary heart disease	men & women	case/control	210
	Alevizaki [5]	17389465	13	coronary heart disease	women	case/control	20
	Yilmaz [8]	18294052	10	coronary heart disease	men & women	case/control	168
	Xu [9]	18582450	9	coronary heart disease	men & women	case/control	179
	REGICOR	Unpubl.	10	myocardial infarction	men & women	case/control	423
							Number
rs2234693 (*PvuII*)	Matsubara [4]	9409287	9	coronary heart disease	men & women	case/control	87
	Alevizaki [5]	17389465	13	coronary heart disease	women	case/control	87
	Almeida [6]	16612467	9	coronary heart disease	men & women	case/control	210
	Mansur [7]	16099331	6	coronary heart disease	men & women	case/control	153
	Yilmaz [8]	18294052	10	coronary heart disease	men & women	case/control	168
	Xu [9]	18582450	9	coronary heart disease	men & women	case/control	210
	Koch [10]	16203927	16	myocardial infarction	men & women	case/control	3587
	Shearman FHS [11,13]	16484614	11	myocardial infarction	men	cohort	154
	Shearman Rot [12,13]	16484614	11	myocardial infarction	men	cohort	303
	Shearman NPHS [13]	16484614	7	myocardial infarction	men	cohort	360
	Shearman GCI-USA [13]	16484614	7	myocardial infarction	men	case/control	226
	Shearman GCI-Poland [13]	16484614	7	myocardial infarction	men	case/control	235
	Kjaergaard [3]	17309937	19	myocardial infarction	men & women	cohort	1137
	Kjaergaard [3]	17309937	19	myocardial infarction	men & women	case/control	2495
	Morgan [14]	17426274	14	myocardial infarction	men & women	case/control	805
	REGICOR	Unpubl.	10	myocardial infarction	men & women	case/control	423

QS, Quality Score.

controls	Genotype cases N (%)					Genotype controls N (%)				
Number	AA	AG	GG	A	G	AA	AG	GG	A	G
89	51 (68)	22 (29)	2 (3)	124 (83)	26 (17)	60 (68)	25 (29)	4 (3)	145 (83)	33 (17)
1556	22 (38)	25 (43)	11 (19)	69 (59)	47 (41)	645 (38)	712 (43)	199 (19)	2002 (59)	1110 (41)
1211	1531 (42)	1658 (45)	468 (13)	4720 (65)	2594 (35)	498 (42)	556 (45)	157 (13)	1552 (65)	870 (35)
142	11 (7)	70 (46)	72 (47)	92 (30)	214 (70)	12 (7)	69 (46)	61 (47)	93 (30)	191 (70)
143	24 (11)	89 (42)	97 (46)	137 (33)	283 (67)	8 (11)	68 (42)	67 (46)	84 (33)	202 (67)
70	4 (20)	11 (55)	5 (25)	19 (48)	21 (53)	27 (20)	37 (55)	6 (25)	91 (48)	49 (53)
99	4 (2)	116 (69)	48 (29)	124 (37)	212 (63)	1 (2)	64 (69)	34 (29)	66 (37)	132 (63)
174	36 (20)	48 (27)	95 (53)	120 (34)	238 (66)	33 (20)	43 (27)	98 (53)	109 (34)	239 (66)
1269	175 (41)	202 (48)	46 (11)	552 (65)	294 (35)	544 (41)	563 (48)	162 (11)	1651 (65)	887 (35)
Number	TT	CT	CC	T	C	TT	CT	CC	T	C
94	27 (31)	47 (54)	13 (15)	101 (58)	73 (42)	34 (36)	46 (49)	14 (15)	114 (61)	74 (39)
70	25 (29)	45 (52)	17 (20)	95 (55)	79 (45)	32 (46)	31 (44)	7 (10)	95 (63)	55 (37)
143	72 (34)	96 (46)	42 (20)	240 (57)	180 (43)	54 (38)	72 (50)	17 (12)	180 (63)	106 (37)
142	51 (33)	85 (56)	17 (11)	187 (61)	119 (39)	47 (33)	69 (49)	26 (18)	163 (57)	121 (43)
99	8 (5)	117 (70)	43 (26)	133 (40)	203 (60)	22 (22)	53 (54)	24 (24)	97 (49)	101 (51)
174	92 (44)	88 (42)	30 (14)	272 (65)	148 (35)	82 (47)	78 (45)	14 (8)	242 (70)	106 (30)
1211	1074 (30)	1781 (50)	732 (20)	3929 (55)	3245 (45)	360 (30)	595 (49)	256 (21)	1315 (54)	1107 (46)
721	37 (24)	71 (46)	46 (30)	145 (47)	163 (53)	236 (33)	363 (50)	122 (17)	835 (58)	607 (42)
1869	86 (28)	148 (49)	69 (23)	320 (53)	286 (47)	524 (28)	949 (51)	396 (21)	1997 (53)	1741 (47)
2349	88 (24)	189 (53)	83 (23)	365 (51)	355 (49)	702 (30)	1182 (50)	465 (20)	2586 (55)	2112 (45)
414	67 (30)	105 (46)	54 (24)	239 (53)	213 (47)	125 (30)	206 (50)	83 (20)	456 (55)	372 (45)
441	75 (32)	102 (43)	58 (25)	252 (54)	218 (46)	141 (32)	209 (47)	91 (21)	491 (56)	391 (44)
8044	360 (32)	547 (48)	230 (20)	1267 (56)	1007 (44)	2352 (29)	4023 (50)	1669 (21)	8727 (54)	7361 (46)
4447	740 (30)	1268 (51)	487 (20)	2748 (55)	2242 (45)	1296 (29)	2237 (50)	914 (21)	4829 (54)	4065 (46)
656	239 (30)	421 (52)	145 (18)	899 (56)	711 (44)	187 (29)	326 (50)	143 (22)	700 (53)	612 (47)
1269	117 (28)	231 (55)	75 (18)	465 (55)	381 (45)	383 (30)	636 (50)	250 (20)	1402 (55)	1136 (45)

Supplementary Table 6: Descriptive statistics for REGICOR participants included in the study stratified by sex.

WOMEN		cases (n=105)	controls (n=315)	*p*-value †
Age (year)		67.02 ± 9.01	66.38 ± 8.46	0.511
Hypertension *		57 (54.3)	232 (73.7)	0.676
Diabetes *		23 (21.9)	71 (22.5)	0.174
Dyslipemia *		49 (46.7)	152 (48.3)	0.014
Smoking *	never smoker	59 (56.2)	293 (93.0)	<0.001
	current or ex-smoker <1 year	15 (14.3)	13 (4.1)	
	ex-smoker >1year	0 (0)	3 (1.0)	
BMI (kg/m^2)		27.50 ± 4.93	28.92 ± 4.94	0.055
SBP (mmHg)		111.45 ± 16.87	144.96 ± 21.08	<0.001
DBP (mmHg)		59.42 ± 10.29	80.53 ± 9.84	<0.001
Family history of MI *		14 (13.3)	36 (11.4)	0.014
Total cholesterol (mg/dl)		185.54 ± 39.05	232.31 ± 41.88	<0.001
HDL cholesterol (mg/dl)		53.42 ± 13.68	54.18 ± 13.13	0.712
rs2234693 (*PvuII*)	TT	30 (28.6)	94 (29.8)	0.834
	CT	55 (52.4)	169 (53.7)	
	CC	20 (19.0)	52 (16.5)	
	MAF	0.452	0.434	
rs9340799 (*XbaI*)	AA	43 (41)	132 (42)	0.465
	AG	46 (44)	149 (47)	
	GG	16 (15)	34 (11)	
	MAF	0.371	0.344	

MEN		cases (n=318)	controls (n=954)	*p*-value †
Age (year)		59.29 ± 11.15	58.85 ± 10.65	0.524
Hypertension *		140 (44.0)	561 (58.8)	0.002
Diabetes *		73 (22.9)	199 (20.8)	0.195
Dyslipemia *		169 (53.1)	407 (42.6)	<0.001
Smoking *	never Stoker	49 (15.4)	326 (34.2)	<0.001
	current or ex-smoker <1 year	161 (50.6)	240 (25.2)	
	ex-smoker >1year	81 (25.5)	368 (38.6)	
BMI (kg/m^2)		27.75 ± 4.50	27.93 ± 3.78	0.519
SBP (mmHg)		109.94 ± 17.12	139.65 ± 20.05	<0.001
DBP (mmHg)		61.71 ± 10.75	82.18 ± 9.73	<0.001
Family history of MI *		46 (14.5)	98 (10.3)	0.001
Total cholesterol (mg/dl)		190.77 ± 43.51	224.69 ± 39.94	<0.001
HDL cholesterol (mg/dl)		41.93 ± 9.33	46.26 ± 11.17	<0.001
rs2234693 (*PvuII*)	TT	87 (27.4)	289 (30.3)	0.131
	CT	176 (55.3)	467 (49.0)	
	CC	55 (17.3)	198 (20.8)	
	MAF	0.450	0.452	
rs9340799 (*XbaI*)	AA	132 (42)	412 (43.2)	0.086
	AG	156 (49)	414 (43.4)	
	GG	30 (9)	128 (13.4)	
	MAF	0.340	0.351	

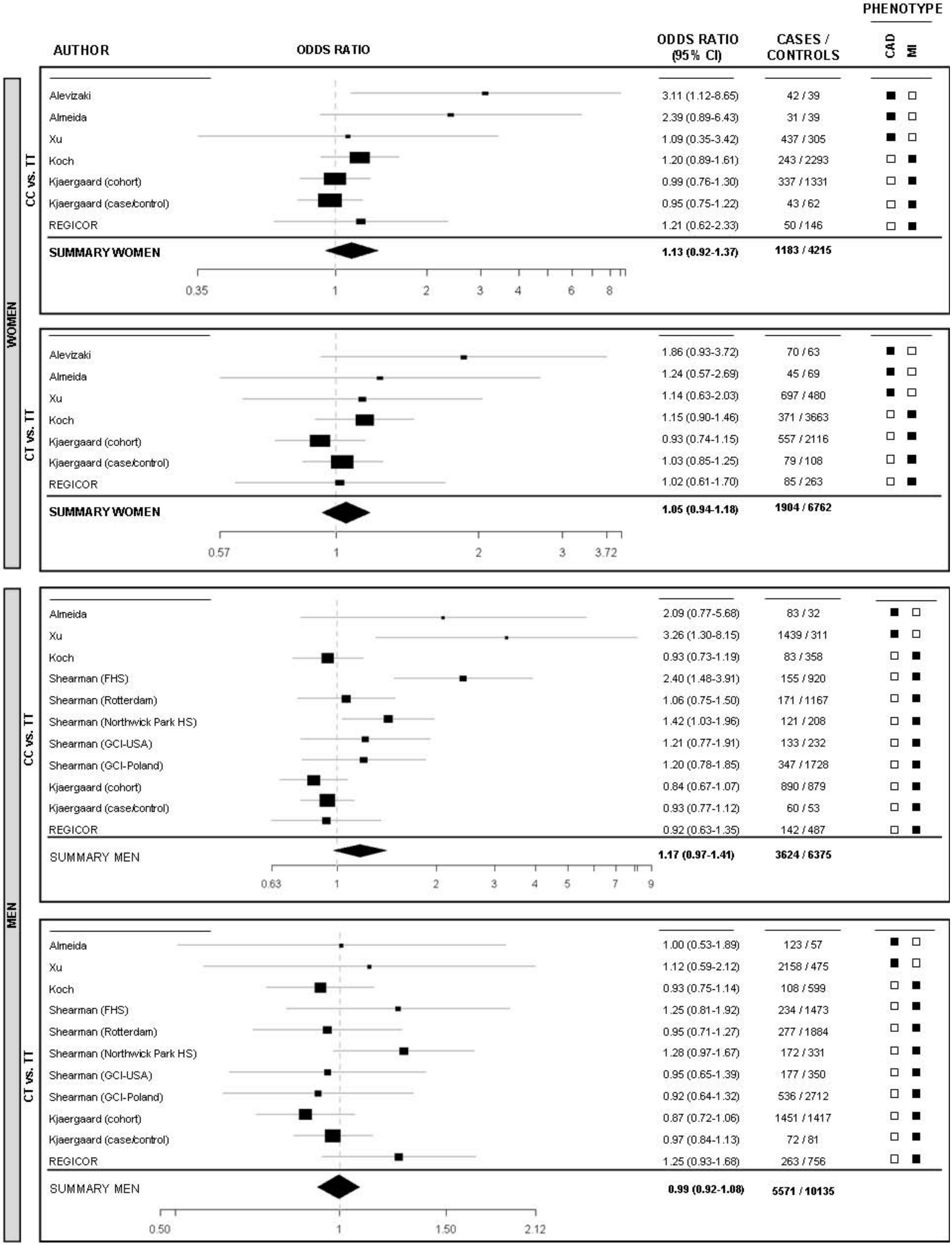
Results are expressed as mean ± SD for normally distributed variables or n (%) for categorical variables. MAF, minor allele frequency; SD, standard deviation. Both polymorphisms were in HWE.

* Self reported history or treatment

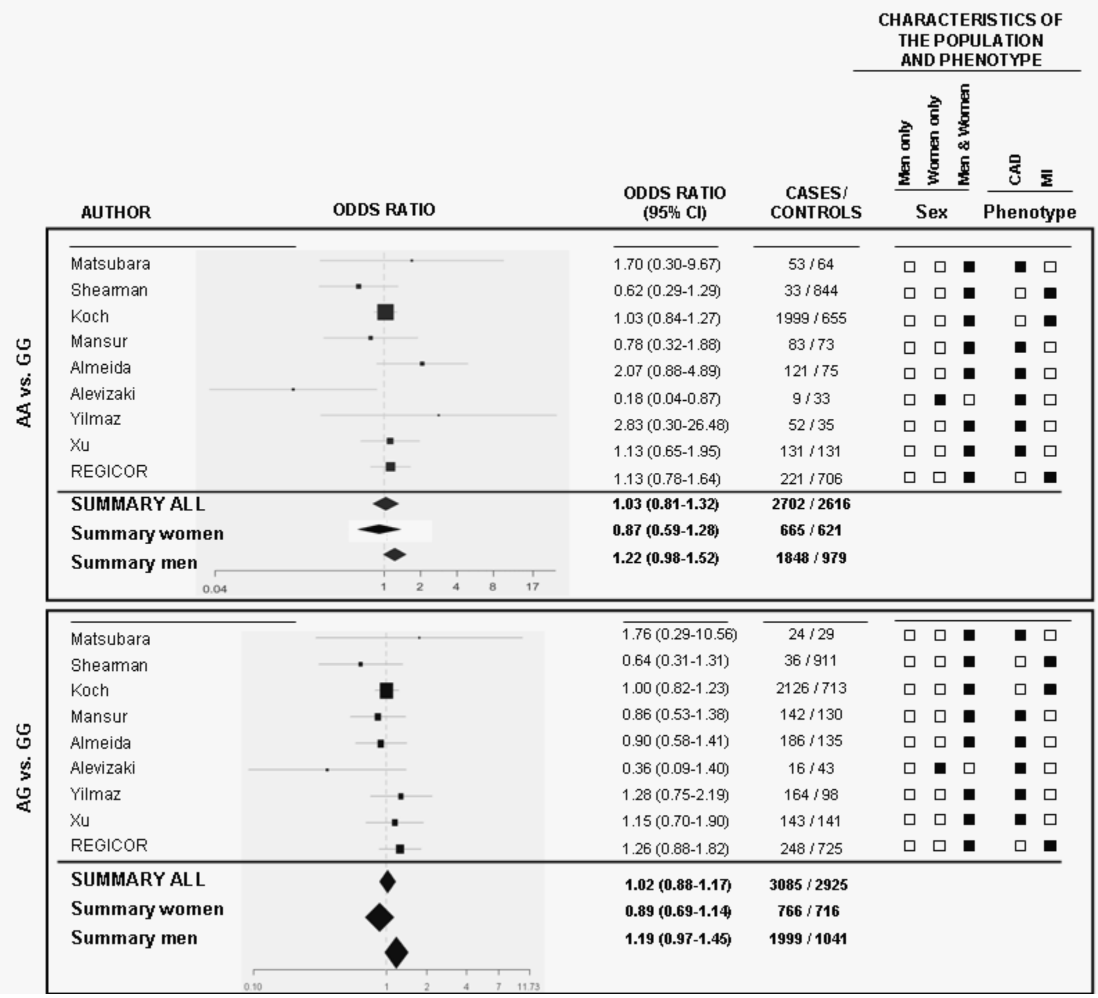
† To test differences between cases and controls, a Pearson χ^2 test was performed for categorical variables and a Student t test for normally distributed variables.

6. Supplementary figures

Supplementary Fig. 1: Meta-analyses of association studies reporting association between rs2234693 and CHD stratified by sex. None of the female populations has been previously included in previous meta-analyses. The different meta-analyses represent the association between just females and just males. Although the studies provided by Mansur *et al.* [7], Matsubara *et al.* [4], Morgan *et al.* [14] and Yilmaz *et al.* [8] had data for both men and women, we could not use it for the stratified analysis because the genotypes are provided for the global sample. The pooled OR is shown as a diamond (•), where the width of the diamond corresponds to the 95%CI of the pooled OR. Data presented in a logarithmic scale.



Supplementary Fig. 2: Meta-analysis of the genetic variant rs9340799 (*XbaI*) in *ESR1* gene in relation to CHD risk.



	AUTHOR	ODDS RATIO (95% CI)	CASES/ CONTROLS	Sex: Men only	Sex: Women only	Sex: Men & Women	Phenotype: CAD	Phenotype: MI
AA vs. GG	Matsubara	1.70 (0.30-9.67)	53 / 64	□	□	■	■	□
	Shearman	0.62 (0.29-1.29)	33 / 844	□	□	■	□	■
	Koch	1.03 (0.84-1.27)	1999 / 655	□	□	■	□	■
	Mansur	0.78 (0.32-1.88)	83 / 73	□	□	■	■	□
	Almeida	2.07 (0.88-4.89)	121 / 75	□	□	■	■	□
	Alevizaki	0.18 (0.04-0.87)	9 / 33	□	■	□	■	□
	Yilmaz	2.83 (0.30-26.48)	52 / 35	□	□	■	■	□
	Xu	1.13 (0.65-1.95)	131 / 131	□	□	■	■	□
	REGICOR	1.13 (0.78-1.64)	221 / 706	□	□	■	□	■
	SUMMARY ALL	**1.03 (0.81-1.32)**	**2702 / 2616**					
	Summary women	**0.87 (0.59-1.28)**	**665 / 621**					
	Summary men	**1.22 (0.98-1.52)**	**1848 / 979**					
AG vs. GG	Matsubara	1.76 (0.29-10.56)	24 / 29	□	□	■	■	□
	Shearman	0.64 (0.31-1.31)	36 / 911	□	□	■	□	■
	Koch	1.00 (0.82-1.23)	2126 / 713	□	□	■	□	■
	Mansur	0.86 (0.53-1.38)	142 / 130	□	□	■	■	□
	Almeida	0.90 (0.58-1.41)	186 / 135	□	□	■	■	□
	Alevizaki	0.36 (0.09-1.40)	16 / 43	□	■	□	■	□
	Yilmaz	1.28 (0.75-2.19)	164 / 98	□	□	■	■	□
	Xu	1.15 (0.70-1.90)	143 / 141	□	□	■	■	□
	REGICOR	1.26 (0.88-1.82)	248 / 725	□	□	■	□	■
	SUMMARY ALL	**1.02 (0.88-1.17)**	**3085 / 2925**					
	Summary women	**0.89 (0.69-1.14)**	**766 / 716**					
	Summary men	**1.19 (0.97-1.45)**	**1999 / 1041**					

The pooled OR is shown as a diamond (◆), where the width of the diamond corresponds to the 95%CI of the pooled OR. Data presented in a logarithmic scale.

References

[1] Ioannidis JP, Boffetta P, Little J, et al. Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol 2007;37(1):120-3.

[2] Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. Nature 2007;447:655-60.

[3] Kjaergaard AD, Ellervik C, Tybjaerg-Hansen A, et al. Estrogen receptor alpha polymorphism and risk of cardiovascular disease, cancer, and hip fracture: cross-sectional, cohort, and case-control studies and a meta-analysis. Circulation 2007;115:861-71.

[4] Matsubara Y, Murata M, Kawano K, et al. Genotype distribution of estrogen receptor polymorphisms in men and postmenopausal women from healthy and coronary populations and its relation to serum lipid levels. Arterioscler Thromb Vasc Biol 1997;17:3006-12.

[5] Alevizaki M, Saltiki K, Cimponeriu A, et al. Severity of cardiovascular disease in postmenopausal women: associations with common estrogen receptor alpha polymorphic variants. Eur J Endocrinol 2007;156:489-96.

[6] Almeida S, Hutz MH. Estrogen receptor 1 gene polymorphisms and coronary artery disease in the Brazilian population. Braz J Med Biol Res 2006;39:447-54.

[7] Mansur AP, Nogueira CC, Strunz CM, Aldrighi JM, Ramires JA. Genetic polymorphisms of estrogen receptors in patients with premature coronary artery disease. Arch Med Res 2005;36:511-7.

[8] Yilmaz A, Menevse S, Erkan AF, et al. The relationship of the ESR1 gene polymorphisms with the presence of coronary artery disease determined by coronary angiography. Genet Test. 2007;11:367-71.

[9] Xu H, Hou X, Wang N, et al. Gender-specific effect of estrogen receptor-1 gene polymorphisms in coronary artery disease and its angiographic severity in Chinese population. Clin Chim Acta 2008;395:130-3.

[10] Koch W, Hoppmann P, Pfeufer A, Mueller JC, Schomig A, Kastrati A. No replication of association between estrogen receptor alpha gene polymorphisms and susceptibility to myocardial infarction in a large sample of patients of European descent. Circulation 2005;112:2138-42.

[11] Shearman AM, Cupples LA, Demissie S, et al. Association between estrogen receptor alpha gene variation and cardiovascular disease. JAMA 2003;290:2263-70.

[12] Schuit SC, Oei HH, Witteman JC, et al. Estrogen receptor alpha gene polymorphisms and risk of myocardial infarction. JAMA 2004;291:2969-77.

[13] Shearman AM, Cooper JA, Kotwinski PJ, et al. Estrogen receptor alpha gene variation is associated with risk of myocardial infarction in more than seven thousand men from five cohorts. Circ Res 2006;98:590-2.

[14] Morgan TM, Krumholz HM, Lifton RP, Spertus JA. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. JAMA 2007;297:1551-61.

7.3. APPENDIX: Article 2: Supplementary material

Post-genomic update on a classical candidate gene for coronary artery disease: *ESR1*

Gavin Lucas MSc, PhD[*1]; Carla Lluís-Ganella, MSc[*1]; Isaac Subirana, MSc[2,1]; Mariano Sentí, MD, PhD[1,3]; Christina Willenborg[4,5]; Muntaser Musameh MD, PhD[6]; CARDIoGRAM Consortium[†]; Stephen M Schwartz MD, PhD[7,8]; Christopher J O'Donnell MD MPH[9,10]; Olle Melander MD, PhD[11]; Veikko Salomaa MD, PhD[12]; Roberto Elosua, MD, PhD[1,2].

* These authors contributed equally to this work

1 Cardiovascular Epidemiology and Genetics, IMIM. Barcelona, Spain.

2 Epidemiology and Public Health Network (CIBERESP), Barcelona, Spain.

3 Pompeu Fabra University. Barcelona, Spain.

4 Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany.

5 Medizinische Klinik II, Universität zu Lübeck, Lübeck, Germany.

6 Department of Cardiovascular Sciences, University of Leicester, United Kingdom.

7 Cardiovascular Health Research Unit, Departments of Medicine and Epidemiology, University of Washington, Seattle, Washington, USA.

8 Department of Epidemiology, University of Washington, Seattle, Washington, USA.

9 National, Heart, Lung, and Blood Institute and Framingham Heart Study, Framingham, Massachusetts, USA.

10 Cardiology Division, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.

11 Department of Clinical Sciences, Hypertension and Cardiovascular Diseases, University Hospital Malmö, Lund University, Malmö, Sweden.

12 National Institute for Health and Welfare, Helsinki, Finland.

† See Supplementary Appendix for a full list of contributors

Supplementary Methods

a. *CARDIoGRAM discovery analysis methods summary*

Genotyping in individual discovery GWA studies was carried out on Affymetrix or Illumina platforms. Approximately 2.3 million imputed genotypes were generated using the MACH, IMPUTE, or BIMBAM imputation algorithms and the HapMap Phase II reference panel{International HapMap Consortium, 2007 364 /id}). Each primary discovery GWAS performed a logistic regression analysis to test for association between genotyped and imputed SNPs and risk of MI/CAD under an additive disease model adjusted for age and sex and taking into account the uncertainty of imputed genotypes. In every study, the variance inflation factor λ was estimated from genotyped SNPs and also used for adjustment. Quality control of these data was performed centrally according to established criteria including a check of consistency of the given alleles across all studies, quality of the imputation, deviation from Hardy-Weinberg equilibrium in the controls, minor allele frequency, and call rate.

In the present study, a meta-analysis was performed separately for every SNP from every CARDIoGRAM study that passed the quality criteria. The default meta-analysis was a fixed effect model with inverse variance weighting and calculation of two homogeneity statistics: Cochran's Q- and I^2 statistic. When there was no indication of heterogeneity for a SNP (P for $Q > 0.01$), the fixed effect model was maintained. When heterogeneity was present (P for $Q < 0.01$), a random effects model (DerSimonian-Laird) was used for that SNP.

b. *Test for interaction between SNP and gender*

To formally test for interaction between each SNP and gender in the CARDIoGRAM and fine mapping analyses (data not shown for the latter), we performed the following steps:

i. Within each CARDIoGRAM study, we computed the beta for the SNP-gender interaction term as the absolute difference between the betas for females and males.

ii. Within each CARDIoGRAM study, we computed the standard error of the SNP-gender interaction term as square root of the sum of the squares of the standard errors of the β from the female and male analysis.

iii. We then used these betas and standard errors to perform fixed or random effects meta-analyses according to the same protocol as that used for the un-stratified analysis.

c. *Power Calculations*

We performed a post-hoc calculation of our analyses' power to detect significant associations. We allowed that power is determined by sample size, the proportion of cases and controls for the case-control studies or the number of events for the cohort study, the effect of a variant on risk (e.g. OR), and the frequency of the minor allele (MAF) of the associated variant, the p-value threshold required to declare statistical significance, LD between correlated and causal variants, genotyping error, the quality of imputation for imputed variants, between-study heterogeneity in the meta-analysis, and possibly other factors. Of these, MAF is the most important non-modifiable determinant of power, and so we estimated power for a series of representative sub-ranges of MAF. Rather than attempting to parameterize all of the other factors, we captured their effects by using the standard error (SE) from the meta-analysis of all three studies, which is inversely correlated with power. In these power calculations, the variant's effect on disease risk was taken as the beta from the meta-analysis of all studies, and thus represented the odds ratios for the case-control studies and hazard ratio for cohort studies, where applicable; ORs and HRs are considered to be comparable because the prevalence of the phenotype in the cohort studies is relatively low. All power computations were based on an alpha value (Type I error rate) equivalent to the threshold required to declare a statistically significant association after adjustment for multiple testing (see main text). Within each analysis we performed the following steps:

i. For each SNP in the analysis, MAF was taken to be the mean MAF across all studies.

ii. SNPs were binned according to the following sub-ranges of MAF: (0,0.01], (0.01,0.02], (0.02,0.03], (0.03,0.04], (0.04,0.05], (0.05,0.06], (0.06,0.07], (0.07,0.08], (0.08,0.09], (0.09,0.1], (0.1,0.15], (0.15,0.2], (0.2,0.3], (0.3,0.4] and (0.4,0.5].

iii. For each sub-range of MAF the mean of the SE of all SNPs within that sub-range was computed, and used to compute and express the power of the analysis in the following two ways.

iv. The minimum effect size (beta) the analysis had high (~80%) or moderate (~50%) power to detect. The definitions of high and moderate power were selected arbitrarily to indicate where our analysis was well powered to detect risk effects (high power), but also to allow for the fact that, if multiple independent but more subtle effects were present, at least some proportion of these could also be detected (e.g. 50%, moderate power).

v. The power of the analysis to detect each of a series of effect sizes (betas, corresponding to the following odds ratios: 1.05, 1.1, 1.2, 1.3, 1.5, 1.7, 2, 2.5 and 3). These data were computed to help indicate the circumstances under which our study was unable to provide conclusive information, e.g. for rarer variants or for more subtle effect sizes.

The results of these power calculations are shown in Supplementary Table 2

Supplementary Note

Preliminary age-stratified analysis to explore possible menopause-related *ESR1* effects among women

After age and gender, menopausal status among women appears to be one of the strongest determinants of CAD risk. We explored the possibility that the effect of genetic variation in *ESR1* on CAD risk may vary among women according to menopausal status. Although this variable was not available for any of the CARDIoGRAM studies or for the three studies included in the fine mapping analysis, we attempted to capture most of its variation using a proxy variable based on age (<50 years or ≥50 years{Palacios, 2010 4274 /id}), and then tested for interaction between this proxy variable and genotype. This analysis was performed only for MIGen owing to the lack of age data for the WTCCC sample, and the low number of events in the Framingham study.

We observed no regionally significant interaction between this proxy variable and genotype for any variant in the region of interest (minimum p-value=0.0012 for rs11968025), although we note the limited sample size of this analysis (number of females aged <50 yrs and ≥50 yrs was 832 (of which 389 were cases) and 582 (of which 274 were cases), respectively).

Supplementary Tables

Supplementary Table 1. Chromosomal locations of coding and non-coding exons in *ESR1*.

Exon Name[*]	Coding[†]	Start[‡]	Stop[‡]	Length (bp)	Position with respect to translation start site[§]	AA length
E2	-	151977808	151977899	91	-151240	
F	-	152011675	152011800	125	-117373	
E1	-	152023011	152023141	130	-106037	
T1	-	152112508	152112595	87	-16540	
T2	-	152112697	152112848	151	-16351	
D	-	152125065	152125160	95	-3983	
C	-	152125748	152126956	1208	-3300	
B	-	152128494	152128645	151	-554	
A	-	152128816	152128978	162	-232	
1	+	152128979[‖]	152129499	521	-70	151
2	+	152163732	152163922	190	34684	64
3	+	152201790	152201906	116	72742	39
4	+	152265308	152265643	335	136260	112
5	+	152332791	152332929	138	203743	46
6	+	152382126	152382259	133	253078	45
7	+	152415520	152415703	183	286472	61
8	+	152419867	152420102	235	290819	77
3' UTR	-	152420103	152424406	4303	291055	

AA: Amino Acid; bp: base pairs. 152177055

[*] Name assigned by Koš *et al.*{Kos, 2001 199 /id} to non-coding exons, or sequentially for coding exons

[†] Non-coding, -; coding, +

[‡] Position in GRCh37.p1 determined using information provided by Koš *et al*. for non-coding exons and the Exon 1 start point, and from the Ensembl exon report for the coding exons (ENSG00000091831; www.ensembl.org) otherwise.

[§] Translation start codon begins at 152129048, 70bp downstream of the beginning of Exon 1

‖ Common splice acceptor site reported by Koš *et al.*

Supplementary Table 2. Power computation – see Supplementary Methods for details**.**

		Minor Allele Frequency range	(0,0.01]	(0.01,0.02]	(0.02,0.03]	(0.03,0.04]	(0.04,0.05]	(0.05,0.06]	(0.06,0.07]	(0.07,0.08]	(0.08,0.09]	(0.09,0.1]	(0.1,0.15]	(0.15,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]
CARDIoGRAM (*n=86,995*)		Number of SNPs	0	4	13	12	3	15	12	11	13	23	60	79	106	87	97
	OR detectable	high power (0.8)	--	1.329	1.277	1.220	1.200	1.283	1.214	1.203	1.150	1.150	1.121	1.103	1.081	1.074	1.071
		moderate power (0.5)	--	1.263	1.223	1.178	1.161	1.227	1.173	1.164	1.122	1.121	1.098	1.084	1.066	1.060	1.058
	Power (*to detect OR*)	1.05	--	0.001	0.0017	0.0034	0.0047	0.0028	0.0055	0.0086	0.015	0.025	0.048	0.07	0.19	0.27	0.33
		1.1	--	0.011	0.023	0.057	0.085	0.047	0.100	0.15	0.28	0.42	0.62	0.77	0.96	0.98	0.98
		1.2	--	0.2	0.37	0.67	0.79	0.39	0.68	0.71	0.97	0.89	0.96	0.99	1	1	1
		1.3	--	0.68	0.87	0.99	1	0.79	0.98	0.98	1	0.98	1	1	1	1	1
		1.5	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		1.7	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2.5	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		3	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CARDIoGRAM Females (*n=30,615*)		Number of SNPs	0	4	1	8	1	10	8	16	10	19	65	71	102	89	96
	OR detectable	high power (0.8)	--	1.577	1.490	1.400	1.412	1.472	1.365	1.292	1.279	1.222	1.201	1.184	1.162	1.13	1.123
		moderate power (0.5)	--	1.454	1.388	1.318	1.328	1.374	1.291	1.235	1.224	1.18	1.162	1.149	1.132	1.106	1.1
	Power (*to detect OR*)	1.05	--	0.00038	0.00048	0.0007	0.00065	0.00068	0.0012	0.0019	0.0019	0.0033	0.0048	0.0064	0.015	0.024	0.03
		1.1	--	0.0021	0.0029	0.0057	0.0050	0.0051	0.014	0.026	0.026	0.055	0.087	0.12	0.28	0.43	0.5
		1.2	--	0.027	0.042	0.098	0.083	0.1	0.23	0.39	0.39	0.66	0.79	0.89	0.91	0.99	1
		1.3	--	0.14	0.22	0.43	0.39	0.35	0.58	0.79	0.84	0.98	0.99	0.99	0.97	1	1
		1.5	--	0.63	0.82	0.96	0.95	0.77	0.93	0.99	1	1	1	1	1	1	1
		1.7	--	0.93	0.99	1	1	0.97	1	1	1	1	1	1	1	1	1
		2	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2.5	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		3	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CARDIoGRAM Males (*n=32,069*)		Number of SNPs	0	1	7	6	3	2	3	14	14	23	68	75	100	94	97
	OR detectable	high power (0.8)	--	1.488	1.382	1.304	1.261	1.226	1.336	1.352	1.344	1.202	1.211	1.152	1.123	1.101	1.096
		moderate power (0.5)	--	1.387	1.305	1.244	1.21	1.182	1.269	1.281	1.275	1.164	1.171	1.123	1.1	1.082	1.078
	Power (*to detect OR*)	1.05	--	0.00048	0.0008	0.0014	0.0021	0.0029	0.0015	0.0016	0.0021	0.0068	0.0099	0.017	0.045	0.072	0.09
		1.1	--	0.003	0.0072	0.016	0.029	0.047	0.021	0.024	0.034	0.13	0.19	0.32	0.62	0.79	0.85
		1.2	--	0.043	0.13	0.28	0.45	0.63	0.29	0.23	0.29	0.76	0.75	0.95	0.96	1	1
		1.3	--	0.22	0.49	0.77	0.91	0.99	0.62	0.56	0.58	0.94	0.86	0.97	0.98	1	1
		1.5	--	0.83	0.97	1	1	1	0.98	0.97	0.96	1	0.98	1	1	1	1
		1.7	--	0.99	1	1	1	1	1	1	1	1	1	1	1	1	1
		2	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		2.5	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		3	--	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Notes:

1. Within each analysis, the number of SNPs whose mean SE was used to compute power is shown for each sub-range of MAF.

2. 'OR detectable' indicates the minimum risk effect detectable (expressed as the exponent of the beta from the meta-analysis) with high or moderate power. 'Power' indicates the study's power to detect the effect sizes (odds ratios) shown.

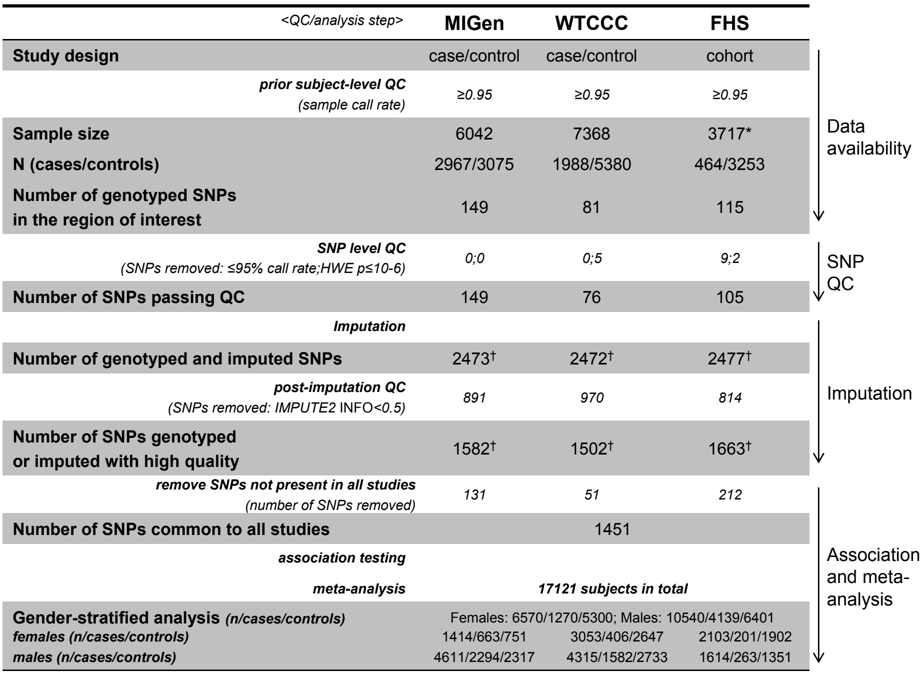
3. Effect sizes detectable or for which power is shown are expressed as the exponent of the absolute beta from the meta-analysis (i.e. the odds ratio computed with the lower risk group set as the reference group). Thus, in the CARDIoGRAM, Females, Males, and Fine Mapping analyses, these are the odds ratios associated for each additional copy of the risk allele; in the Gender*Genotype Interaction analysis these are the odds ratios for difference in risk between sexes.

4. Power does not increase linearly with increase in MAF because these data are based on empirical SE values, which may be affected by other factors (e.g. imputation quality, between-study heterogeneity in the meta-analysis, etc.) for SNPs in some sub-ranges of MAF compared to others.

5. In the computation of power for given effect size, scenarios with high power (≥80%) are shaded dark grey, those with moderate power (≥50% and <80%) are shaded light grey, and those with power lower than 50% are unshaded.

Supplementary Figures

Supplementary Figure 1. Summary of quality control and analysis procedures in the fine mapping analysis.

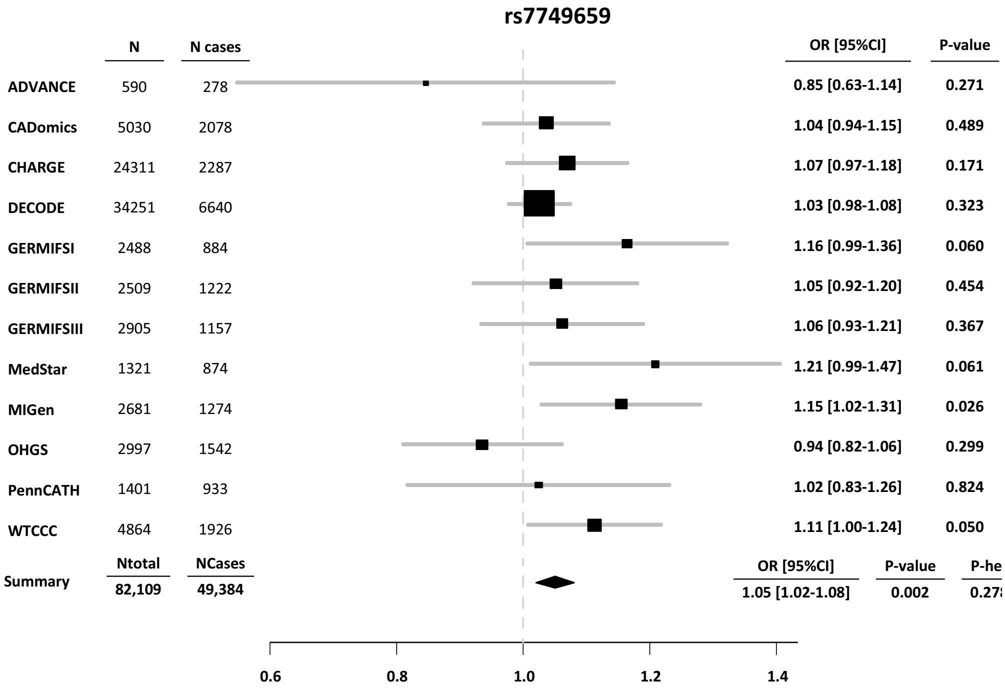


<QC/analysis step>	MIGen	WTCCC	FHS	
Study design	case/control	case/control	cohort	Data availability
prior subject-level QC *(sample call rate)*	*≥0.95*	*≥0.95*	*≥0.95*	
Sample size	6042	7368	3717*	
N (cases/controls)	2967/3075	1988/5380	464/3253	
Number of genotyped SNPs in the region of interest	149	81	115	
SNP level QC *(SNPs removed: ≤95% call rate;HWE p≤10-6)*	*0;0*	*0;5*	*9;2*	SNP QC
Number of SNPs passing QC	149	76	105	
Imputation				Imputation
Number of genotyped and imputed SNPs	2473†	2472†	2477†	
post-imputation QC *(SNPs removed: IMPUTE2 INFO<0.5)*	*891*	*970*	*814*	
Number of SNPs genotyped or imputed with high quality	1582†	1502†	1663†	
remove SNPs not present in all studies *(number of SNPs removed)*	*131*	*51*	*212*	Association and meta-analysis
Number of SNPs common to all studies	1451			
association testing				
meta-analysis	***17121 subjects in total***			
Gender-stratified analysis ***(n/cases/controls)***	Females: 6570/1270/5300; Males: 10540/4139/6401			
females (n/cases/controls)	1414/663/751	3053/406/2647	2103/201/1902	
males (n/cases/controls)	4611/2294/2317	4315/1582/2733	1614/263/1351	

[*] The publicly available dataset for the Framingham study contained genotype and phenotype data for 9,270 individuals. In the current analysis, we included 3,717 individuals from the original and offspring cohorts for whom survival data were available for the follow-up periods beginning at visits 15 and 5 respectively. The sample selection and process used to filter these individuals is described in more detail in Lluís-Ganella *et al.* 2011 (submitted).

[†] For all three studies, all genotyped SNPs were also present in the 1kG reference panel. These values show the total number of SNPs, including genotyped SNPs, after imputation, post-imputation QC, and filtering to include SNPs common to all studies.

Supplementary Figure 2. CARDIoGRAM global meta-analysis results for the top SNP in the region of interest. Total sample size, number of cases, OR and 95% CI are shown for each contributing study, in addition to global sample sizes, OR, 95%CI, and p-values for association and heterogeneity. Note that only 12 of the 14 CARDIoGRAM studies are represented, as data for this variant was not available in the LURIC 1 and LURIC 2 samples.



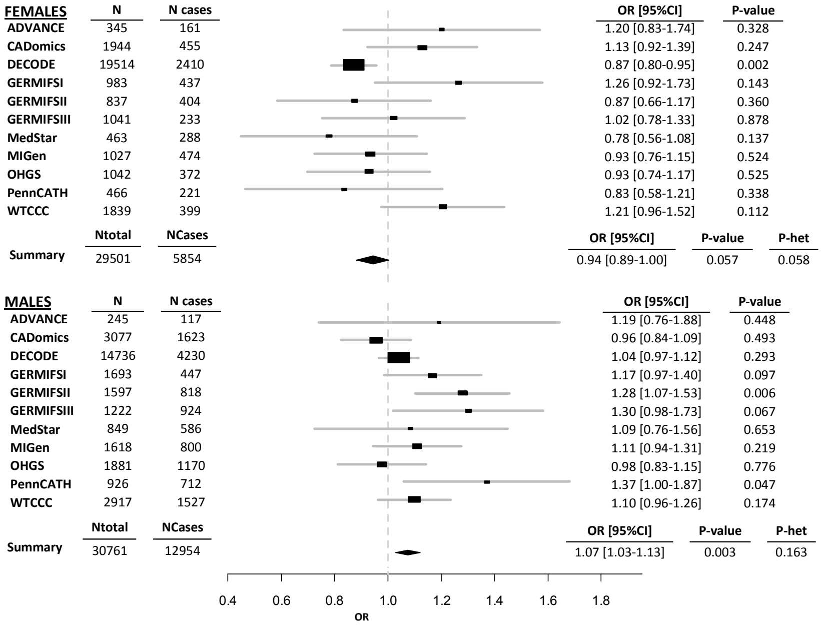
rs7749659

	N	N cases	OR [95%CI]	P-value
ADVANCE	590	278	0.85 [0.63-1.14]	0.271
CADomics	5030	2078	1.04 [0.94-1.15]	0.489
CHARGE	24311	2287	1.07 [0.97-1.18]	0.171
DECODE	34251	6640	1.03 [0.98-1.08]	0.323
GERMIFSI	2488	884	1.16 [0.99-1.36]	0.060
GERMIFSII	2509	1222	1.05 [0.92-1.20]	0.454
GERMIFSIII	2905	1157	1.06 [0.93-1.21]	0.367
MedStar	1321	874	1.21 [0.99-1.47]	0.061
MIGen	2681	1274	1.15 [1.02-1.31]	0.026
OHGS	2997	1542	0.94 [0.82-1.06]	0.299
PennCATH	1401	933	1.02 [0.83-1.26]	0.824
WTCCC	4864	1926	1.11 [1.00-1.24]	0.050

	Ntotal	NCases	OR [95%CI]	P-value	P-he
Summary	82,109	49,384	1.05 [1.02-1.08]	0.002	0.27

Supplementary Figure 3. Meta-analysis results for the CARDIoGRAM SNP with the greatest difference in association between males and females (strongest interaction with gender). Total sample size, number of cases, OR and 95% CI are shown for each contributing study, in addition to global sample sizes, OR, 95%CI, and p-values for association and heterogeneity. Note that only 11 of the 14 CARDIoGRAM studies are represented, as data for this variant was not available in the LURIC 1, LURIC 2 and CHARGE samples.

6-152177055



FEMALES	N	N cases	OR [95%CI]	P-value
ADVANCE	345	161	1.20 [0.83-1.74]	0.328
CADomics	1944	455	1.13 [0.92-1.39]	0.247
DECODE	19514	2410	0.87 [0.80-0.95]	0.002
GERMIFSI	983	437	1.26 [0.92-1.73]	0.143
GERMIFSII	837	404	0.87 [0.66-1.17]	0.360
GERMIFSIII	1041	233	1.02 [0.78-1.33]	0.878
MedStar	463	288	0.78 [0.56-1.08]	0.137
MIGen	1027	474	0.93 [0.76-1.15]	0.524
OHGS	1042	372	0.93 [0.74-1.17]	0.525
PennCATH	466	221	0.83 [0.58-1.21]	0.338
WTCCC	1839	399	1.21 [0.96-1.52]	0.112

	Ntotal	NCases	OR [95%CI]	P-value	P-het
Summary	29501	5854	0.94 [0.89-1.00]	0.057	0.058

MALES	N	N cases	OR [95%CI]	P-value
ADVANCE	245	117	1.19 [0.76-1.88]	0.448
CADomics	3077	1623	0.96 [0.84-1.09]	0.493
DECODE	14736	4230	1.04 [0.97-1.12]	0.293
GERMIFSI	1693	447	1.17 [0.97-1.40]	0.097
GERMIFSII	1597	818	1.28 [1.07-1.53]	0.006
GERMIFSIII	1222	924	1.30 [0.98-1.73]	0.067
MedStar	849	586	1.09 [0.76-1.56]	0.653
MIGen	1618	800	1.11 [0.94-1.31]	0.219
OHGS	1881	1170	0.98 [0.83-1.15]	0.776
PennCATH	926	712	1.37 [1.00-1.87]	0.047
WTCCC	2917	1527	1.10 [0.96-1.26]	0.174

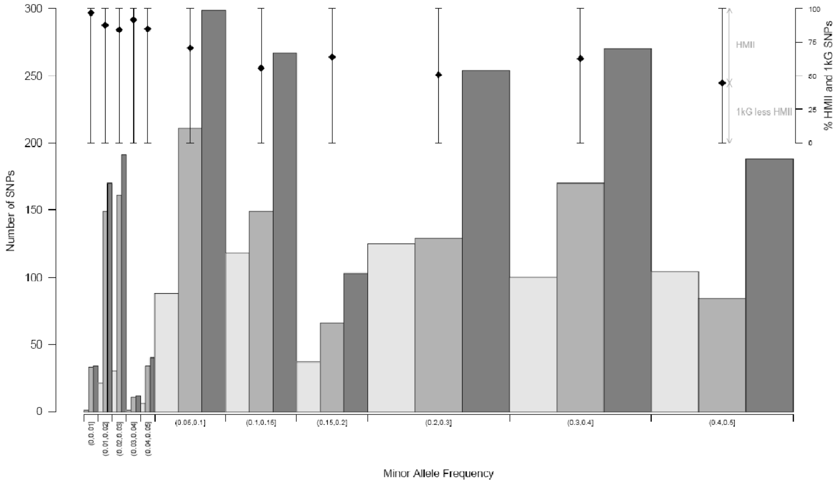
	Ntotal	NCases	OR [95%CI]	P-value	P-het
Summary	30761	12954	1.07 [1.03-1.13]	0.003	0.163

Supplementary Figure 4. Distribution of minor allele frequencies (MAF) for SNPs analyzed in this study.

Data are shown as vertical bars whose widths are proportional to the ranges of MAF indicated on the x-axis, and whose heights correspond to the absolute number (left y-axis) of SNPs whose MAF falls within those ranges (MAF computed as the weighted mean in the MIGen, WTCCC and Framingham samples).

The number of SNPs within the region of interest that were genotyped or imputed in (a) the CARDIoGRAM meta-analysis (corresponding to the high-quality SNPs from the HapMapII reference panel) are indicated as white bars; (b) the number of additional SNPs imputed in the fine mapping analysis in this study are shown as light grey bars; the total the number of SNPs analyzed in the fine mapping analysis in this study (a plus b, corresponding to high-quality SNPs from the 1kG reference panel) are shown as dark grey bars.

Within each sub-range of MAF, the vertical black lines and diamonds at the top of the graph represent the proportions of SNPs analyzed in the fine mapping analysis (1kG panel, see (c) above). The portion of the line above the diamond represents the percentage (right y-axis) of these SNPs that were included in the HapMapII panel, and the portion below the diamond represents those additional SNPs that were imputed in the present study. This graph shows that many additional SNPs with a broad range of MAFs were imputed in this study, but that the greatest gain of information was obtained for rarer SNPs.



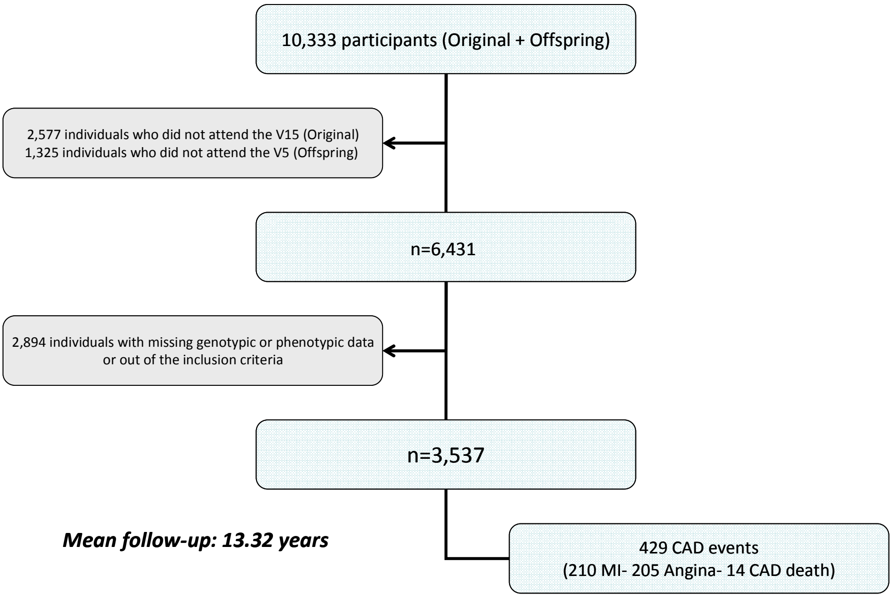
Supplementary References

1. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851-861.

2. Palacios S, Henderson VW, Siseles N, Tan D, Villaseca P. Age of menopause and impact of climacteric symptoms by geographical region. Climacteric. 2010; 13:419-428.

3. Kos M, Reid G, Denger S, Gannon F. Minireview: genomic organization of the human ERalpha gene promoter region. Mol Endocrinol. 2001; 15:2057-2063.

Supplementary Appendix 1. Process of selection of participants from the Framingham study (from Lluís-Ganella *et al.*, submitted).


10,333 participants (Original + Offspring)
2,577 individuals who did not attend the V15 (Original)
1,325 individuals who did not attend the V5 (Offspring)
n=6,431
2,894 individuals with missing genotypic or phenotypic data or out of the inclusion criteria
n=3,537
Mean follow-up: 13.32 years
429 CAD events
(210 MI- 205 Angina- 14 CAD death)

Supplementary Appendix 2. CARDIoGRAM Investigators

Executive Committee: Sekar Kathiresan1,2,3, Muredach P. Reilly4, Nilesh J. Samani5,6, Heribert Schunkert7
Executive Secretary: Jeanette Erdmann7
Steering Committee: Themistocles L. Assimes8, Eric Boerwinkle9, Jeanette Erdmann7, Alistair Hall10, Christian Hengstenberg11, Sekar Kathiresan1,2,3, Inke R.
Konig12, Reijo Laaksonen13, Ruth McPherson14, Muredach P. Reilly4, Nilesh J. Samani5,6, Heribert Schunkert7, John R. Thompson15, Unnur Thorsteinsdottir16,17,
Andreas Ziegler12

ADVANCE: Devin Absher18, Themistocles L. Assimes8, Stephen Fortmann8, Alan Go27, Mark Hlatky8, Carlos Iribarren27, Joshua Knowles8, Richard Myers18,
Thomas Quertermous8, Steven Sidney27, Neil Risch28, Hua Tang29
CADomics: Stefan Blankenberg30, Tanja Zeller30, Arne Schillert12, Philipp Wild30, Andreas Ziegler12, Renate Schnabel30, Christoph Sinning30, Karl Lackner31,
Laurence Tiret32, Viviane Nicaud32, Francois Cambien32, Christoph Bickel30, Hans J. Rupprecht30, Claire Perret32, Carole Proust32, Thomas Munzel30
CHARGE: Maja Barbalic33, Joshua Bis34, Eric Boerwinkle9, Ida Yii-Der Chen35, L. Adrienne Cupples20,21, Abbas Dehghan36, Serkalem Demissie-Banjaw37,21, Aaron
Folsom38, Nicole Glazer39, Vilmundur Gudnason40,41, Tamara Harris42, Susan Heckbert43, Daniel Levy21, Thomas Lumley44, Kristin Marciante45, Alanna
Morrison46, Christopher J. O´Donnell47, Bruce M. Psaty48, Kenneth Rice49, Jerome I. Rotter35, David S. Siscovick50, Nicholas Smith43, Albert Smith40,41, Kent D.
Taylor35, Cornelia van Duijn36, Kelly Volcik46, Jaqueline Whitteman36, Vasan Ramachandran51, Albert Hofman36, Andre Uitterlinden52,36
deCODE: Solveig Gretarsdottir16, Jeffrey R. Gulcher16, Hilma Holm16, Augustine Kong16, Kari Stefansson16,17, Gudmundur Thorgeirsson53,17, Karl Andersen53,17,
Gudmar Thorleifsson16, Unnur Thorsteinsdottir16,17
GERMIFS I and II: Jeanette Erdmann7, Marcus Fischer11, Anika Grosshennig12,7, Christian Hengstenberg11, Inke R. Konig12, Wolfgang Lieb54, Patrick Linsel-
Nitschke7, Michael Preuss12,7, Klaus Stark11, Stefan Schreiber55, H.-Erich Wichmann56,58,59, Andreas Ziegler12, Heribert Schunkert7
GERMIFS III (KORA): Zouhair Aherrahrou7, Petra Bruse7, Angela Doering56, Jeanette Erdmann7, Christian Hengstenberg11, Thomas Illig56, Norman Klopp56, Inke
R. Konig12, Patrick Linsel-Nitschke7, Christina Loley12,7, Anja Medack7, Christina Meisinger56, Thomas Meitinger57,60, Janja Nahrstedt12,7, Annette Peters56,
Michael Preuss12,7, Klaus Stark11, Arnika K. Wagner7, H.-Erich Wichmann56,58,59, Christina Willenborg12,7, Andreas Ziegler12, Heribert Schunkert7
LURIC/AtheroRemo: Bernhard O. Bohm61, Harald Dobnig62, Tanja B. Grammer63, Eran Halperin22, Michael M. Hoffmann64, Marcus Kleber65, Reijo Laaksonen13,
Winfried Marz63,66,67, Andreas Meinitzer66, Bernhard R. Winkelmann68, Stefan Pilz62, Wilfried Renner66, Hubert Scharnagl66, Tatjana Stojakovic66, Andreas
Tomaschitz62, Karl Winkler64
MIGen: Benjamin F. Voight2,3,24, Kiran Musunuru1,2,3, Candace Guiducci3, Noel Burtt3, Stacey B. Gabriel3, David S. Siscovick50, Christopher J. O'Donnell47,
Roberto Elosua69, Leena Peltonen49, Veikko Salomaa70, Stephen M. Schwartz50, Olle Melander26, David Altshuler71,3, Sekar Kathiresan1,2,3
OHGS: Alexandre F. R. Stewart14, Li Chen19, Sonny Dandona14, George A. Wells25, Olga Jarinova14, Ruth McPherson14, Robert Roberts14
PennCATH/MedStar: Muredach P. Reilly4, Mingyao Li23, Liming Qu23, Robert Wilensky4, William Matthai4, Hakon H. Hakonarson72, Joe Devaney73, Mary Susan
Burnett73, Augusto D. Pichard73, Kenneth M. Kent73, Lowell Satler73, Joseph M. Lindsay73, Ron Waksman73, Christopher W. Knouff74, Dawn M. Waterworth74,
Max C. Walker74, Vincent Mooser74, Stephen E. Epstein73, Daniel J. Rader75,4
WTCCC: Nilesh J. Samani5,6, John R. Thompson15, Peter S. Braund5, Christopher P. Nelson5, Benjamin J. Wright76, Anthony J. Balmforth77, Stephen G. Ball78,
Alistair S. Hall10, Wellcome Trust Case Control Consortium

Affiliations

1 Cardiovascular Research Center and Cardiology Division, Massachusetts General Hospital, Boston, MA, USA.
2 Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA.
3 Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.
4 The Cardiovascular Institute, University of Pennsylvania, Philadelphia, PA, USA.
5 Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK.
6 Leicester National Institute for Health Research Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, LE3 9QP, UK.
7 Medizinische Klinik II, Universitat zu Lubeck, Lubeck, Germany.
8 Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.
9 University of Texas Health Science Center, Human Genetics Center and Institute of Molecular Medicine, Houston, TX, USA.
10 Division of Cardiovascular and Neuronal Remodelling, Multidisciplinary Cardiovascular Research Centre, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, UK.
11 Klinik und Poliklinik fur Innere Medizin II, Universitat Regensburg, Regensburg, Germany.
12 Institut fur Medizinische Biometrie und Statistik, Universitat zu Lubeck, Lubeck, Germany.
13 Science Center, Tampere University Hospital, Tampere, Finland.
14 The John & Jennifer Ruddy Canadian Cardiovascular Genetics Centre, University of Ottawa Heart Institute, Ottawa, Canada.
15 Department of Health Sciences, University of Leicester, Leicester, UK.
16 deCODE Genetics, 101 Reykjavik, Iceland.
17 University of Iceland, Faculty of Medicine, 101 Reykjavik, Iceland.
18 Hudson Alpha Institute, Huntsville, Alabama, USA.
19 Cardiovascular Research Methods Centre, University of Ottawa Heart Institute, 40 Ruskin Street, Ottawa, Ontario, Canada, K1Y 4W7.
20 Department of Biostatistics, Boston University School of Public Health, Boston, MA USA.
21 National Heart, Lung and Blood Institute's Framingham Heart Study, Framingham, MA, USA.
22 The Blavatnik School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel, and the International Computer Science Institute, Berkeley, CA, USA.
23 Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA.
24 Department of Medicine, Harvard Medical School, Boston, MA, USA.
25 Research Methods, Univ Ottawa Heart Inst.
26 Department of Clinical Sciences, Hypertension and Cardiovascular Diseases, Scania University Hospital, Lund University, Malmo, Sweden.
27 Division of Research, Kaiser Permanente, Oakland, CA, USA.
28 Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA.
29 Dept Cardiovascular Medicine, Cleveland Clinic.
30 Medizinische Klinik und Poliklinik, Johannes-Gutenberg Universitat Mainz, Universitatsmedizin, Mainz, Germany.
31 Institut fur Klinische Chemie und Laboratoriumsmediizin, Johannes-Gutenberg Universitat Mainz, Universitatsmedizin, Mainz, Germany.
32 INSERM UMRS 937, Pierre and Marie Curie University (UPMC, Paris 6) and Medical School, Paris, France.
33 University of Texas Health Science Center, Human Genetics Center, Houston, TX, USA.
34 Cardiovascular Health Resarch Unit and Department of Medicine, University of Washington, Seattle, WA USA.
35 Cedars-Sinai Medical Center, Medical Genetics Institute, Los Angeles, CA, USA.
36 Erasmus Medical Center, Department of Epidemiology, Rotterdam, The Netherlands.
37 Boston University, School of Public Health, Boston, MA, USA.
38 University of Minnesota School of Public Health, Division of Epidemiology and Community Health, School of Public Health (A.R.F.), Minneapolis, MN, USA.
39 University of Washington, Cardiovascular Health Research Unit and Department of Medicine, Seattle, WA, USA.
40 Icelandic Heart Association, Kopavogur Iceland.
41 University of Iceland, Reykjavik, Iceland.
42 Laboratory of Epidemiology, Demography, and Biometry, Intramural Research Program, National Institute on Aging, National Institutes of Health, Bethesda MD, USA.

43 University of Washington, Department of Epidemiology, Seattle, WA, USA.
44 University of Washington, Department of Biostatistics, Seattle, WA, USA.
45 University of Washington, Department of Internal Medicine, Seattle, WA, USA.
46 University of Texas, School of Public Health, Houston, TX, USA.
47 National Heart, Lung and Blood Institute, Framingham Heart Study, Framingham, MA and Cardiology Division, Massachusetts General Hospital, Boston, MA, USA.
48 Center for Health Studies, Group Health, Departments of Medicine, Epidemiology, and Health Services, Seattle, WA, USA.
49 The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
50 Cardiovascular Health Research Unit, Departments of Medicine and Epidemiology, University of Washington, Seattle.
51 Boston University Medical Center, Boston, MA, USA.
52 Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands.
53 Department of Medicine, Landspitali University Hospital, 101 Reykjavik, Iceland.
54 Boston University School of Medicine, Framingham Heart Study, Framingham, MA, USA.
55 Institut fur Klinische Molekularbiologie, Christian- Albrechts Universitat, Kiel, Germany.
56 Institute of Epidemiology, Helmholtz Zentrum Munchen – German Research Center for Environmental Health, Neuherberg, Germany.
57 Institut fur Humangenetik, Helmholtz Zentrum Munchen, Deutsches Forschungszentrum fur Umwelt und Gesundheit, Neuherberg, Germany.
58 Institute of Medical Information Science, Biometry and Epidemiology, Ludwig-Maximilians-Universitat Munchen, Germany.
59 Klinikum Grosshadern, Munich, Germany.
60 Institut fur Humangenetik, Technische Universitat Munchen, Germany.
61 Division of Endocrinology and Diabetes, Graduate School of Molecular Endocrinology and Diabetes, University of Ulm, Ulm, Germany.
62 Division of Endocrinology, Department of Medicine, Medical University of Graz, Austria.
63 Synlab Center of Laboratory Diagnostics Heidelberg, Heidelberg, Germany.
64 Division of Clinical Chemistry, Department of Medicine, Albert Ludwigs University, Freiburg, Germany.
65 LURIC non profit LLC, Freiburg, Germany.
66 Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University Graz, Austria.
67 Institute of Public Health, Social and Preventive Medicine, Medical Faculty Manneim, University of Heidelberg, Germany.
68 Cardiology Group Frankfurt-Sachsenhausen, Frankfurt, Germany.
69 Cardiovascular Epidemiology and Genetics Group, Institut Municipal d'Investigacio Medica, Barcelona.
Ciber Epidemiologia y Salud Publica (CIBERSP), Spain.
70 Chronic Disease Epidemiology and Prevention Unit, Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland.
71 Department of Molecular Biology and Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, USA.
72 The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
73 Cardiovascular Research Institute, Medstar Health Research Institute, Washington Hospital Center, Washington, DC 20010, USA.
74 Genetics Division and Drug Discovery, GlaxoSmithKline, King of Prussia, Pennsylvania 19406, USA.
75 The Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
76 Department of Cardiovascular Surgery, University of Leicester, Leicester, UK.
77 Division of Cardiovascular and Diabetes Research, Multidisciplinary Cardiovascular Research Centre, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds, LS2 9JT, UK.
78 LIGHT Research Institute, Faculty of Medicine and Health, University of Leeds, Leeds, UK

7.4. APPENDIX: Article 4: Supplementary material

Assessment of the value of a genetic risk score in improving the estimation of coronary risk

Carla Lluís-Ganella[a,*], Isaac Subirana[b,a,*], Gavin Lucas[a], Marta Tomás[a], Daniel Muñoz[a], Mariano Sentí[a,c], Eduardo Salas[d], Joan Sala[e], Rafel Ramos[f,g], Jose Ordovas[h,i], Jaume Marrugat[a], Roberto Elosua[a,b]

[a]Cardiovascular Epidemiology and Genetics, IMIM. Barcelona, Spain.
[b]CIBER Epidemiology and Public Health (CIBERESP). Barcelona, Spain.
[c]Pompeu Fabra University. Barcelona, España.
[d]Gendiag.exe. Barcelona, España.
[e]Cardiology Department, Hospital Universitari Josep Trueta. Girona, Spain.
[f]Primary Care Research Institute (IDIAP-Jordi Gol), Girona, Spain.
[g]Medical Science Department, Medical School, Universitat de Girona, Spain.
[h]Nutrition and Genomics Laboratory. Jean Mayer US Department of Agriculture Human Nutrition Research Center on Aging, Tufts University School of Medicine, Boston, MA.
[i]The Department of Epidemiology and Population Genetics, Centro Nacional Investigación Cardiovasculares (CNIC). Madrid, Spain.

* These authors contributed equally to this work

TABLE OF CONTENTS **Page**

Supplemental Methods section (Sx), table (S.Tx), figure (S.Fx) and analysis (S.Ax) numbers are indicated in parentheses throughout the main manuscript and supplementary material.

SUPPLEMENTARY METHODS

SUPPLEMENTARY TABLES

SUPPLEMENTARY FIGURES

SUPPLEMENTARY ANALYSES

SUPPLEMENTARY REFERENCES *S32*

SUPPLEMENTARY METHODS

S1. Genetic variant selection, genotyping, quality controls and generation of the multi-locus risk score.

S1.1. Genetic variant selection: SNP-selection was carried out as described previously [1]. Briefly, we searched the NHGRI GWAS catalog [2] (August, 2010) for the terms 'Myocardial Infarction/Coronary disease (MI/CAD)' and related phenotypes. This search returned 21 genetic variants. Those variants that reported an association p-value >1x10^{-6} were excluded for the present analysis. In order to minimize redundant information in the genetic risk score (GRS), we computed the linkage equilibrium between variants using data from the HapMap CEU sample, and from those variants that presented high correlation (LD r^2>0.3), one was randomly selected. We evaluated the evidence in the NHGRI GWAS catalogue for each of the 14 remaining variants, and excluded those that had been reported to be associated with classical cardiovascular risk factors (CVRF), such as total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, diabetes, hypertension and smoking. Moreover, we excluded 2 of the remaining SNPs because literature-based evidence strongly suggested an association between those loci and CVRF. From this list we also excluded variants that were not associated with MI/CAD in the CARDIoGRAM study [3]. We added the rs10455872 variant in *LPA* because it has since been reported to be strongly association with MI/CAD [3,4]. See the flow chart of the selection process in *S.F2.*

S1.2. Generation of multi-locus genetic risk score: The GRS was weighted by the estimated effect size reported for each variant in the CARDIoGRAM study [3] using the following formula:

$$GRS = \sum_{i=1}^{8} \beta_i \cdot SNP_i$$

Where:

- β_i is the estimated effect size reported for each variant in the CARDIoGRAM study;

- SNP_i is the number of copies of each individual SNP evaluated (can have values of 0, 1 or 2 for genotyped SNPs and values ranging from 0 to 2 for imputed SNPs)

S1.3. Genotyping and genotyping quality control: REGICOR participants' DNA was obtained from buffy coat using standardized methods [5] (L'ARS services, Barcelona, Spain) and samples were genotyped by Centro Nacional de Investigación Oncológica (CNIO, Madrid, Spain) using the Cardio inCode chip (Ferrer inCode, Barcelona, Spain) based on Veracode (Illumina, San Diego, USA) and KASPar (KBioscience , Hoddesdon, United Kingdom) technologies. The overall percentage of agreement of the chip with reference technology is 99.9% and the analytical sensitivity and specificity are greater than 98.6%. For the Framingham participants, the genotypes for genotyped SNPs were obtained using the Affymetrix 500K and 50K chips, and for additional SNPs by imputation into the HapMapII CEU haplotype panel (build 36, release 22), using MACH version 1.00.15.

S1.4. Quality control: Various quality control measures were applied at both participant and SNP levels to the data from both cohorts: Individuals with low call rates or sex mismatches were excluded before imputation in the Framingham cohort database. Moreover, high levels of missingness ($p<10^{-9}$), highly significant departures from Hardy-Weinberg equilibrium ($p<10^{-6}$), or Mendelian errors

(>100) were used to determine which SNPs to use for the imputation step, and were also applied as quality control criteria for the SNPs selected.

S2. Follow-up and phenotype definition

All REGICOR participants were periodically contacted by telephone or by mail to ascertain whether they had presented any cardiovascular event up until the end of 2009. Fatal events were identified from regional and national mortality registers. All the reported events were reviewed with hospital records or primary care records. An event committee classified the suspected cardiovascular (CVD) events after review of all medical records and physician notes using standardized criteria [6]. This study was approved by the local Ethics Committee and all participants gave written informed consent.

All Framingham participants were analyzed for onset of cardiovascular events during follow-up until the end of 2007. Repeated examinations and clinic visits were carried out approximately every 2 and 4 years, respectively. Suspected cardiovascular events were reviewed and adjudicated by a panel of three Framingham physician investigators after review of all available examination records, hospitalization records and physician notes using standardized criteria [7].

Methodology for laboratory determinations has been described elsewhere [7,8].

Myocardial infarction was defined on the basis of the classical WHO definition by the presence of 2 out of 3 clinical criteria: new diagnostic Q-waves on ECG, prolonged ischemic chest discomfort and elevation of serum biomarkers of myocardial necrosis. Angina was defined by the presence of ischemic chest discomfort with signs of ischemia in the ECG. Coronary artery by-pass grafting or percutaneous coronary interventions were considered as revascularization procedures. CHD death was considered after reviewing the mortality register when the most likely cause of death was CHD and no other cause could be ascribed.

Atherothrombotic stroke was defined as a non-embolic acute-onset focal neurological deficit of vascular origin that persisted for more than 24 hours or an ischemic infarction that was documented at autopsy. Peripheral artery disease was defined by the presence of symptoms of claudication and an objective diagnostic test such as a pathological ankle-brachial index (<0.9) or a pathological arteriography or revascularization procedure.

S3. Ten-year cardiovascular risk estimation

All cardiovascular risk factors required for the risk functions were measured using standard methods [9,10]. Participants were considered to be diabetic if they had been diagnosed with diabetes or treated with oral hypoglycemic agents or insulin or presented a glycemia higher or equal to 126 mg/dL. Those who reported smoking ≥1 cigarette/day in the preceding year were considered smokers. All necessary baseline lipid and blood pressure measurements were collected and used to estimate the risk of each participant.

S4. Statistical analysis

To account for family structure in the Framingham cohort we also adjusted for the first five genetic principal components (computed using PLINK) [11] as covariates in the models [12,13].

All other analyses were performed using R version 2.11 (packages and functions indicated below by <*package*>::<*function*>).

The proportional hazards assumption was tested using *survival::cox.zph*.

The meta-analysis was computed using the *rmeta::meta.DSL* function [14].

We used three different statistics to assess the potential value of including the GRS in risk prediction:

a) to assess the goodness-of-fit of the models we used a version of the Hosmer-Lemeshow test that takes right censoring of the data into account [15];

b) to evaluate the improvement in the discriminative capacity of the model that included the genetic score with respect to a model without the score, we computed the concordance index (c-statistic) using the *Hmisc::rcorr.cens* function [16];

c) to assess the reclassification we calculated the net reclassification improvement (NRI) [17] and integrated discrimination improvement (IDI) [18] in the whole sample and in the subgroup of individuals considered to have intermediate coronary risk according to the classical risk function. To calculate the 10-year expected number of events in each risk category and in each cohort we used the Kaplan-Meier estimates as proposed by Steyerberg and Pencina [15,18]. A bootstrapping method was used to construct confidence intervals for IDI and NRI to take into account the uncertainty of the Kaplan-Meier estimates.

The estimated risk for each individual was computed under the Proportional Hazards assumption (Cox Model)

$$Risk = 1 - S_{\overline{X}}^{\exp \eta},$$

where:

a) $S_{\overline{X}}$ is survival value for the population average. This value depends on gender and has been taken from Framingham equation [19] for the Framingham cohort, and from REGICOR calibrated equation [20] for the REGICOR cohort.

b) $\exp$: exponential value (or anti-logarithm function).

c) η is the linear predictor, i.e, the product of coefficients and factors, and differs for each cohort:

a) For REGICOR $\eta = \sum_{j=1}^{p} \beta_j^F \left(F_j - \overline{F}_j\right) + \beta^G \left(G - \overline{G}\right)$

b) For Framingham $\eta = \sum_{j=1}^{p} \beta_j^F \left(F_j - \overline{F}_j\right) + \beta^G \left(G - \overline{G}\right) + \sum_{k=1}^{5} \beta_k^C \left(C_k - \overline{C}_k\right)$,

where,

- β_j^F : log-hazard-ratios of each of the classical risk factors. These coefficients have not been estimated but taken from the Framingham equation [7].
- F_j : individual value of each classical risk factor.
- $\overline{F}_j$: population average value of each classical risk factor. This value has been taken from Framingham equation [7] for the Framingham cohort, and from REGICOR calibrated equation [20] for the REGICOR cohort.
- β^G : log-hazard-ratios of genetic score, estimated from the data
- G : individual value of genetic score
- $\overline{G}$: average value of genetic score in the sample
- β_k^C : log-hazard-ratios of each of the first five principal components, estimated from the data.
- C_k : individual value of each of the first five principal components.
- $\overline{C}_k$: sample average value of each of the first five principal components.

NOTE: In Framingham cohort, computation of goodness-of-fit (Hosmer-Lemeshow), discrimination (c index), NRI and IDI was performed after adjustment for the first five principal components, in order to allow for the familial nature of the data.

S5. Power calculations

We performed a post-hoc calculation of our analyses' power to detect significant associations. In these power calculations, the variant's effect on disease risk was taken as the beta obtained from each study. All power computations were based on an alpha value (Type I error rate) equivalent to 0.05. Within each analysis we performed the following steps:

i. The minimum effect size (beta) the analysis had high (~80%) or moderate (~50%) power to detect. The definitions of high and moderate power were selected arbitrarily to indicate where our analysis was well powered to detect risk effects (high power), but also to allow for the fact that, if multiple independent but more subtle effects were present, at least some proportion of these could also be detected (e.g. 50%, moderate power).

ii. The power of the analysis to detect each of a series of effect sizes (betas, corresponding to the following hazard ratios: 1.05, 1.09, 1.10, 1.12, 1.14, 1.18, 1.29 and 1.35). These data were computed to help indicate the circumstances under which our study was unable to provide conclusive information, e.g. for rarer variants or for more subtle effect sizes. These hazard ratios were in part selected because are the ones reported in the CARDIoGRAM study for the values we include in this analysis, and therefore we can observe the specific power that we have to achieve each reported HR.

iii. These two computations described were also computed for the GRS and the risk of coronary or cardiovascular disease to evaluate the study power.

The results of these power calculations are shown in *S.T4*.

SUPPLEMENTARY TABLES

S.T1. Clinical characteristics of individuals included in the analysis or not, based on the availability of genetic information.

	Not included	Included	P-value
REGICOR			
Individuals	698	2,351	--
Age (years) *	54.6 (11.0)	53.9 (11.2)	0.128
Gender (male) †	343 (49.1%)	1,123 (47.8%)	0.552
Systolic Blood Pressure (mmHg) *	133 (21.0)	132 (20.8)	0.346
Diastolic Blood Pressure (mmHg) *	79.1 (10.2)	79.5 (10.4)	0.414
Hypertension †	274 (39.5%)	938 (40.1%)	0.843
Smoking †	123 (18.1%)	511 (22.0%)	0.034
Total cholesterol (mg/dL)*	223 (40.7)	225 (42.4)	0.357
LDL cholesterol (mg/dL)*	152 (36.3)	152 (37.9)	0.886
HDL cholesterol (mg/dL)*	50.2 (13.3)	51.7 (13.3)	0.017
Triglycerides (mg/dL)‡	95.0 (69.0-131)	92.0 (70.0-127)	0.523
Cholesterol treatment †	48 (6.91%)	157 (6.71%)	0.926
Diabetic status †	111 (17.2%)	316 (13.8%)	0.036
Diabetes treatment †	35 (5.04%)	96 (4.11%)	0.337
Body mass index (kg/m^2)*	27.6 (4.24)	27.4 (4.47)	0.436
Obesity (BMI≥30 kg/m^2) †	177 (25.8%)	596 (25.6%)	0.962
Estimated 10-y coronary risk §	3.7 (1.9-6.8)	3.3 (1.7-6.2)	0.061
FRAMINGHAM			
Individuals	1,699	3,537	--
Age (years) *	65.8 (12.1)	56.0 (9.26)	<0.001
Gender (male) †	675 (39.7%)	1,540 (43.5%)	0.009
Systolic Blood Pressure (mmHg) *	135 (19.9)	127 (18.3)	<0.001
Diastolic Blood Pressure (mmHg) *	75.3 (10.5)	75.0 (9.79)	0.249
Hypertension †	861 (50.9%)	1,121 (31.7%)	<0.001
Smoking †	449 (26.5%)	713 (20.2%)	<0.001
Total cholesterol (mg/dL)*	222 (43.1)	210 (38.6)	<0.001
LDL cholesterol (mg/dL)*	125 (32.9)	125 (34.1)	0.911
HDL cholesterol (mg/dL)*	50.2 (15.4)	51.0 (15.2)	0.087
Triglycerides (mg/dL)‡	120 (84.0-179)	116 (83.0-172)	0.224
Cholesterol treatment †	55 (3.25%)	166 (4.69%)	0.015
Diabetic status †	164 (10.1%)	226 (6.39%)	<0.001
Diabetes treatment †	72 (4.25%)	90 (2.54%)	0.001
Body mass index (kg/m^2)*	26.7 (4.77)	27.1 (4.78)	0.001
Obesity (BMI≥30 kg/m^2) †	332 (20.2%)	780 (22.1%)	0.126
Estimated 10-y coronary risk §	12.3 (6.9-20.4)	7.79 (4.5-14.1)	<0.001

The 'not included' group includes individuals who were not between 35 and 74 years of age, who had had a previous event, or were missing values for classical risk factors or SNP.

* mean (standard deviation); † n (proportion (%)); ‡ median (25 and 75 percentiles); § mean (95% confidence interval).

S.T2. Effects of classical risk factors on risk of a coronary event.

	HR [95%CI]	**P-value**
REGICOR		
Age (10 years)	2.05 [1.69-2.49]	<0.001
Gender (men)	2.56 [1.69-3.85]	<0.001
Total cholesterol (10 mg/dL)	1.04 [1.00-1.09]	0.092
HDL cholesterol (10 mg/dL)	0.60 [0.50-0.72]	<0.001
Systolic BP (10 mmHg)	1.38 [1.27-1.49]	<0.001
Diastolic BP (10 mmHg)	1.37 [1.15-1.64]	0.001
Diabetes	2.55 [1.66-3.91]	<0.001
Smoker	1.21 [0.78-1.87]	0.392
Family history of CVD*	1.58 [0.96-2.60]	0.068
Estimated 10-y coronary risk†	1.15 [1.12-1.18]	<0.001
FRAMINGHAM		
Age (10 years)	1.60 [1.42-1.81]	<0.001
Gender (men)	2.22 [1.82-2.70]	<0.001
Total cholesterol (10 mg/dL)	1.07 [1.04-1.09]	<0.001
HDL cholesterol (10 mg/dL)	0.74 [0.69-0.80]	<0.001
Systolic BP (10 mmHg)	1.25 [1.19-1.31]	<0.001
Diastolic BP (10 mmHg)	1.33 [1.21-1.47]	<0.001
Diabetes	2.66 [2.02-3.49]	<0.001
Smoker	1.32 [1.07-1.65]	0.011
Family history of CVD‡	1.50 [1.09-2.07]	0.013
Estimated 10-y coronary risk†	1.06 [1.05-1.06]	<0.001

* CVD: Cardiovascular disease.

† Coronary risk was calculated using the original Framingham risk function for the Framingham cohort, and the calibrated function for the REGICOR cohort.

‡ Only in the Offspring sample.

S.T3. SNPs included in the genetic risk score, including genotype quality control.

				Risk	Minor	Weight	REGICOR					Framingham					Meta-analysis	
SNP	Chr	Gene	Position	allele	Allele	(OR)	N total	MAF	p-HWE	HR[95%CI]	p-val	N total	MAF	p-HWE	HR[95%CI]	p-val	HR[95%CI]	p-val
rs17465637	1	*MIA3*	220890152	C	A	1.14	2,351	0.290	0.3409	0.99 [0.74-1.33]	0.482	3,537	0.305	0.9592	0.95 [0.82-1.09]	0.454	0.96 [0.84-1.09]	0.506
rs6725887	2	*WDR12*	203454130	C	C	1.14	2,351	0.144	0.9334	1.10 [0.76-1.60]	0.307	3,537	0.123	0.0572	1.11 [0.91-1.34]	0.299	1.11 [0.93-1.32]	0.242
rs9818870	3	*MRAS*	139604812	T	T	1.12	2,351	0.127	0.0634	1.00 [0.67-1.51]	0.496	3,537	0.142	0.1418	1.15 [0.96-1.37]	0.127	1.12 [0.96-1.32]	0.158
rs12526453	6	*PHACTR1*	13035530	C	G	1.10	2,351	0.353	0.9281	1.19 [0.89-1.59]	0.119	3,537	0.358	0.0098	0.97 [0.84-1.12]	0.656	1.03 [0.86-1.24]	0.739
rs1333049	9	*CDKN2A/2B*	22115503	C	G	1.29	2,351	0.484	0.2006	1.22 [0.93-1.60]	0.077	3,537	0.467	1.0000	1.18 [1.03-1.35]	0.020	1.19 [1.05-1.34]	0.005
rs1746048	10	*CXCL12*	44095830	C	T	1.09	2,351	0.134	0.9291	1.01 [0.68-1.50]	0.475	3,537	0.143	0.0488	0.99 [0.81-1.21]	0.931	0.99 [0.83-1.19]	0.948
rs9982601	21	*SCL5A3*	34520998	T	T	1.18	2,351	0.124	1.0000	1.14 [0.78-1.67]	0.250	3,537	0.147	Imputed	1.15 [0.96-1.39]	0.137	1.15 [0.97-1.36]	0.104
rs10455872	6	*LPA*	160930108	G	G	1.35	2,351	0.078	0.8856	2.26 [1.56-3.29]	<0.001	3,537	0.076	Imputed	1.09 [0.76-1.55]	0.638	1.57 [0.77-3.20]	0.219

Chr: Chromosome; p-HWE: p-value for the Hardy-Weinberg equilibrium; MAF: Minor allele frequency; N total: number of individuals with available genotype (or imputed value) for each variant. P-val: p-value. Weight (OR): odds ratio reported in the CARDIoGRAM study; analyses were weighted by the ln(OR); HR [95%CI]: Hazard ratio [95% confidence interval].

S.T4. Power calculations.

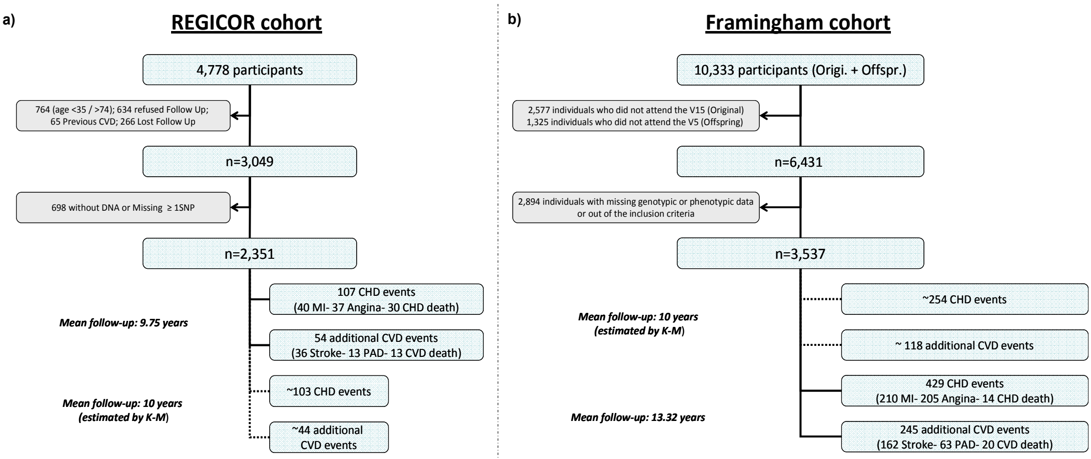
			Minimum HR detectable with high or moderate power		Power to detect a specific HR							
	SNP	se	0.8	0.5	1.05	1.09	1.10	1.12	1.14	1.18	1.29	1.35
REGICOR	rs17465637	0.150	1.52	1.34	0.062	0.089	0.098	0.118	0.141	0.198	0.399	0.519
	rs6725887	0.190	1.70	1.45	0.058	0.074	0.079	0.092	0.106	0.141	0.268	0.352
	rs9818870	0.207	1.79	1.50	0.056	0.070	0.075	0.085	0.097	0.126	0.233	0.305
	rs12526453	0.148	1.51	1.34	0.063	0.090	0.099	0.119	0.143	0.201	0.405	0.527
	rs1333049	0.138	1.47	1.31	0.064	0.095	0.106	0.130	0.157	0.223	0.452	0.582
	rs1746048	0.202	1.76	1.49	0.057	0.071	0.076	0.087	0.100	0.130	0.243	0.318
	rs9982601	0.194	1.72	1.46	0.057	0.073	0.078	0.090	0.104	0.136	0.259	0.339
	rs10455872	0.190	1.70	1.45	0.058	0.074	0.079	0.091	0.106	0.140	0.267	0.351
Framingham	rs17465637	0.073	1.23	1.15	0.103	0.221	0.259	0.345	0.438	0.625	0.939	0.985
	rs6725887	0.099	1.32	1.21	0.078	0.141	0.162	0.209	0.264	0.389	0.732	0.860
	rs9818870	0.091	1.29	1.19	0.084	0.158	0.183	0.239	0.303	0.446	0.801	0.911
	rs12526453	0.073	1.23	1.15	0.102	0.217	0.255	0.339	0.431	0.616	0.934	0.983
	rs1333049	0.069	1.21	1.14	0.109	0.239	0.282	0.375	0.476	0.669	0.958	0.992
	rs1746048	0.102	1.33	1.22	0.076	0.134	0.154	0.198	0.249	0.366	0.701	0.834
	rs9982601	0.094	1.30	1.20	0.081	0.150	0.172	0.225	0.284	0.418	0.769	0.888
	rs10455872	0.182	1.66	1.43	0.058	0.076	0.082	0.096	0.111	0.149	0.288	0.379
Meta-analysis	rs17465637	0.066	1.20	1.14	0.114	0.254	0.300	0.400	0.505	0.702	0.969	0.995
	rs6725887	0.089	1.28	1.19	0.085	0.162	0.187	0.245	0.311	0.457	0.813	0.919
	rs9818870	0.081	1.26	1.17	0.092	0.186	0.217	0.286	0.364	0.531	0.880	0.959
	rs12526453	0.093	1.30	1.20	0.082	0.152	0.175	0.229	0.289	0.426	0.779	0.895
	rs1333049	0.062	1.19	1.13	0.123	0.283	0.335	0.445	0.558	0.758	0.984	0.998
	rs1746048	0.092	1.29	1.20	0.083	0.155	0.179	0.234	0.297	0.437	0.791	0.904
	rs9982601	0.086	1.27	1.18	0.087	0.170	0.198	0.260	0.330	0.484	0.840	0.936
	rs10455872	0.363	2.77	2.04	0.052	0.056	0.058	0.061	0.065	0.074	0.108	0.131

			Minimum HR detectable with high or moderate power		Power to detect a specific HR							
	GRS	se	0.8	0.5	1.05	1.10	1.15	1.20	1.25	1.30	1.40	1.50
REGICOR	Linear	0.056	1.17	1.12	0.139	0.393	0.697	0.898	0.977	0.996	1.000	1.000
	Q2	0.362	2.76	2.03	0.052	0.058	0.067	0.080	0.095	0.112	0.153	0.201
	Q3	0.320	2.45	1.87	0.053	0.060	0.072	0.088	0.107	0.130	0.183	0.244
	Q4	0.294	2.28	1.78	0.053	0.062	0.076	0.095	0.118	0.145	0.209	0.281
	Q5	0.277	2.17	1.72	0.054	0.064	0.080	0.101	0.127	0.157	0.229	0.310
Framingham	Linear	0.031	1.09	1.06	0.352	0.870	0.995	1.000	1.000	1.000	1.000	1.000
	Q2	0.158	1.56	1.36	0.061	0.093	0.143	0.211	0.292	0.382	0.566	0.727
	Q3	0.156	1.55	1.36	0.061	0.094	0.146	0.215	0.298	0.390	0.577	0.738
	Q4	0.153	1.53	1.35	0.062	0.096	0.150	0.223	0.310	0.405	0.597	0.758
	Q5	0.156	1.55	1.36	0.061	0.094	0.146	0.216	0.299	0.391	0.579	0.739
Meta-analysis	Linear	0.058	1.18	1.12	0.133	0.371	0.667	0.877	0.968	0.994	1.000	1.000
	Q2	0.145	1.50	1.33	0.063	0.101	0.162	0.243	0.338	0.442	0.643	0.800
	Q3	0.154	1.54	1.35	0.062	0.095	0.148	0.219	0.304	0.398	0.587	0.748
	Q4	0.134	1.45	1.30	0.065	0.110	0.181	0.275	0.385	0.500	0.710	0.857
	Q5	0.168	1.60	1.39	0.060	0.088	0.132	0.192	0.264	0.345	0.517	0.674

GRS: Genetic risk score; Se: Standard error; 'HR detectable' indicates the minimum risk effect detectable (expressed as the exponent of the beta from the meta-analysis) with high or moderate power. 'Power' indicates the study's power to detect the effects sizes (hazard ratios) shown. In the computation of power for given effect size, scenarios with high power (≥80%) are shaded dark grey, those with moderate power (≥50% and <80%) are shaded light grey, and those with power lower than 50% are unshaded.

SUPPLEMENTARY FIGURES

S.F1. Process of sample inclusion.

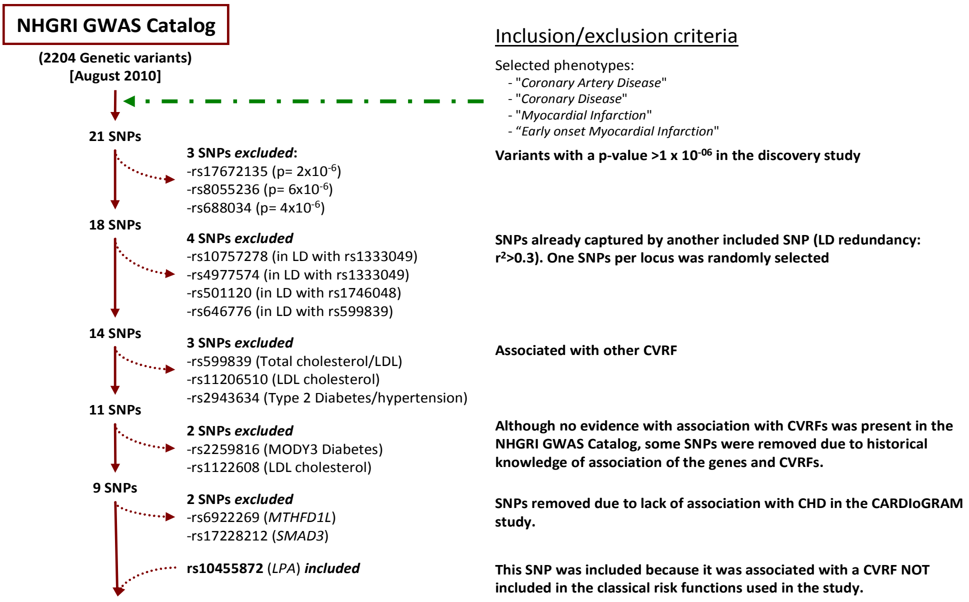

a)
REGICOR cohort
4,778 participants
764 (age <35 / >74); 634 refused Follow Up; 65 Previous CVD; 266 Lost Follow Up
n=3,049
698 without DNA or Missing ≥ 1SNP
n=2,351
107 CHD events (40 MI- 37 Angina- 30 CHD death)
Mean follow-up: 9.75 years
54 additional CVD events (36 Stroke- 13 PAD- 13 CVD death)
~103 CHD events
Mean follow-up: 10 years (estimated by K-M)
~44 additional CVD events
b)
Framingham cohort
10,333 participants (Origi. + Offspr.)
2,577 individuals who did not attend the V15 (Original) 1,325 individuals who did not attend the V5 (Offspring)
n=6,431
2,894 individuals with missing genotypic or phenotypic data or out of the inclusion criteria
n=3,537
~254 CHD events
Mean follow-up: 10 years (estimated by K-M)
~ 118 additional CVD events
429 CHD events (210 MI- 205 Angina- 14 CHD death)
Mean follow-up: 13.32 years
245 additional CVD events (162 Stroke- 63 PAD- 20 CVD death)

CHD: coronary heart disease; CVD: Cardiovascular disease; n: number of individuals; Origi: individuals from the Framingham Original cohort; Offspr: individuals from the Framingham Offspring cohort.

The values for the 10-year follow up in both cohorts have been estimated by Kaplan-Meyer (in REGICOR extending the results from 9.75 years of follow up to 10 years and in Framingham censoring the events from 13.32 to 10 years).

In the REGICOR cohort, the events estimated by Kaplan-Meyer were lower than in the observed sample at a median of 9.75 years because some of the observed events occur at a later stage (>10 years of follow up), and therefore the estimation obtained considers those individuals as event-free. By contrast, some individuals with a follow up <10 years who have not presented an event are considered as event by the estimator. By the same principle, a reduction of ~41% and ~52% of CHD and CVD events from the Framingham cohort can be due to the high number of individuals with unavailability of genetic data (although they were eligible for the present study).

S.F2. Process of SNP selection.



NHGRI GWAS Catalog

(2204 Genetic variants) [August 2010]

Inclusion/exclusion criteria

Selected phenotypes:
- "*Coronary Artery Disease*"
- "*Coronary Disease*"
- "*Myocardial Infarction*"
- "*Early onset Myocardial Infarction*"

21 SNPs

3 SNPs *excluded*:
-rs17672135 (p= 2x10^{-6})
-rs8055236 (p= 6x10^{-6})
-rs688034 (p= 4x10^{-6})

Variants with a p-value >1 x 10^{-06} in the discovery study

18 SNPs

4 SNPs *excluded*
-rs10757278 (in LD with rs1333049)
-rs4977574 (in LD with rs1333049)
-rs501120 (in LD with rs1746048)
-rs646776 (in LD with rs599839)

SNPs already captured by another included SNP (LD redundancy: r^2>0.3). One SNPs per locus was randomly selected

14 SNPs

3 SNPs *excluded*
-rs599839 (Total cholesterol/LDL)
-rs11206510 (LDL cholesterol)
-rs2943634 (Type 2 Diabetes/hypertension)

Associated with other CVRF

11 SNPs

2 SNPs *excluded*
-rs2259816 (MODY3 Diabetes)
-rs1122608 (LDL cholesterol)

Although no evidence with association with CVRFs was present in the NHGRI GWAS Catalog, some SNPs were removed due to historical knowledge of association of the genes and CVRFs.

9 SNPs

2 SNPs *excluded*
-rs6922269 (*MTHFD1L*)
-rs17228212 (*SMAD3*)

SNPs removed due to lack of association with CHD in the CARDIoGRAM study.

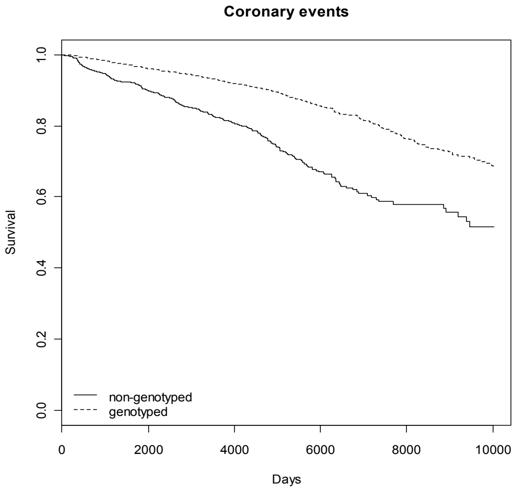
rs10455872 (*LPA*) ***included***

This SNP was included because it was associated with a CVRF NOT included in the classical risk functions used in the study.

8 SNPs selected:

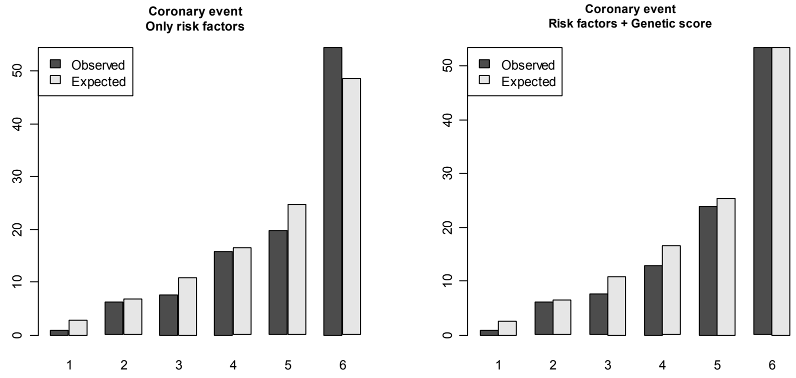
SNP	Chromosome	Gene	Position	Minor Allele
rs17465637	1	*MIA3*	220890152	A
rs6725887	2	*WDR12*	203454130	C
rs9818870	3	*MRAS*	139604812	T
rs12526453	6	*PHACTR1*	13035530	G
rs1333049	9	*CDKN2A/2B*	22115503	G
rs1746048	10	*CXCL12*	44095830	T
rs9982601	21	*SCL5A3*	34520998	T
rs10455872	6	*LPA*	160930108	G

S.F3. Kaplan-Meier curves for those individuals who were included in the analysis or not, based on the availability of phenotypic or genotypic information from the Framingham Heart Study.



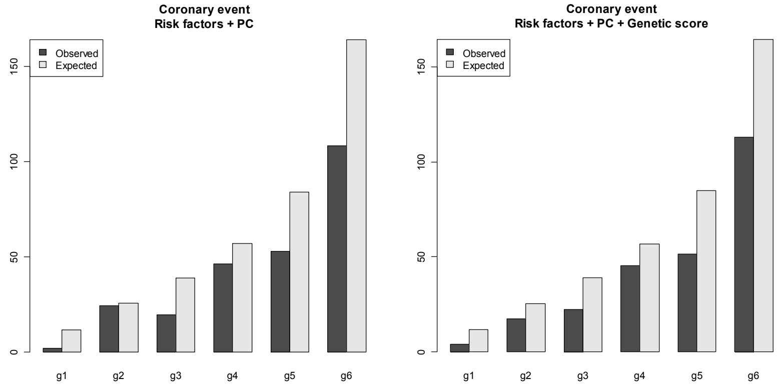
S.F4. Analysis of the goodness-of-fit of the models with and without the genetic risk score, for coronary heart disease events both in REGICOR (a) and Framingham (b) cohorts using the Hosmer-Lemeshow test.

a) REGICOR



REGICOR risk function	Chi-square = 4.20 (df = 4), p-value = 0.383
REGICOR risk function + genetic risk score	Chi-square = 3.00 (df = 4), p-value = 0.557

b) FRAMINGHAM



Framinghan risk function	Chi-square = 60.38 (df = 4), p-value <0.001
Framinghan risk function + genetic risk score	Chi-square = 55.37 (df = 4), p-value <0.001

SUPPLEMENTARY ANALYSES

Supplementary Analysis 1

Predictive capacity of a coronary risk function improved by including a genetic score – extension of main analysis to CVD

1. INTRODUCTION

In 1994 the European Atherosclerosis Society and the European Society of Hypertension published a set of recommendations for CHD prevention [21]. The main reason for separating CHD and total cardiovascular risk (CVD), which are similar but distinct outcomes, was an attempt to simplify the estimation of CVD risk. However, by 2003 the Third Joint Task Force Guidelines proposed a change from CHD to CVD prevention, to reflect the fact that atherosclerosis may affect any part of the vascular tree [22,23], and because some of the clinical manifestations of CVD are thought to share a common etio-pathogenesis with CHD.

Although a population based strategy is critical to reducing the overall incidence of CVD [23], primary prevention in high risk groups is also widely implemented and an improvement of the risk functions for a significant reduction of incidence of the disease is warranted.

The aims of the current analyses were also to address steps 2 and 3 of the AHA recommendations for the same GRS. First, we assessed the association between the multi-locus GRS and incident CVD events in two prospective cohort studies with low and high CVD mortality (AHA, step 2). Second, we assessed whether the inclusion of this GRS improves the predictive capacity of the Framingham risk function (AHA, step 3). In addition, we evaluated the hypothesis that the improvement in predictive capacity provided by the GRS is greater among individuals with intermediate risk.

2. METHODS

Follow-up and phenotype definition

All REGICOR participants were periodically contacted to ascertain whether they had presented any CVD event up until the end of 2009, and events were reviewed using hospital or primary care records. Fatal events were identified from regional and national mortality registers. After reviewing all medical records and physician notes, suspected CVD events were classified in committee according to standardized criteria [6].

Among Framingham participants, a record was made of all CHD events that occurred during follow-up until the end of 2007. Suspected cardiovascular events were reviewed by a panel of Framingham physician investigators after reviewing all available medical records and physician notes using standardized criteria [7].

CVD events included myocardial infarction (MI), angina, coronary revascularization and death due to CHD, plus atherothrombotic stroke and peripheral artery disease.

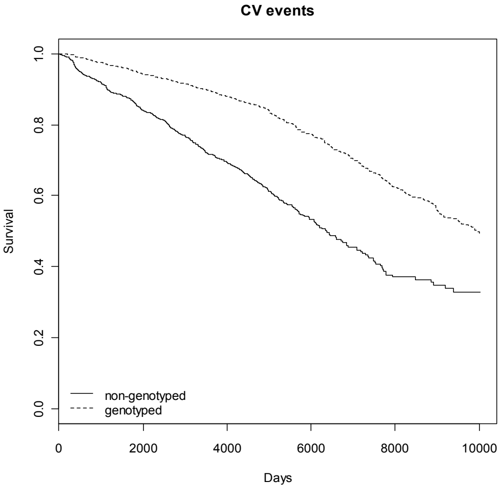
3. RESULTS

Sample selection and sample characteristics

The number of participants included was 2,351 from the REGICOR cohort and 3,537 from the Framingham cohort, and the number of observed CVD events was 161 in a mean follow-up of 9.75 years, and 674 in a mean follow-up of 13.32 years, respectively (*S.F2*).

As observed for CHD, in the Framingham sample, there was a difference in survival rates between individuals who had DNA sample available and those who did not and those included presented a better cardiovascular risk profile (*S.T1*) and a lower incidence of CVD events than those not included (*S.A1.Figure 1*)

S.A1.Figure 1. Kaplan-Meier curves for those individuals who were included in the analysis or not, based on the availability of phenotypic or genotypic information from the Framingham Heart Study.



The characteristics of the participants included in the present analyses stratified by cohort, and by the presence/absence of CVD events are shown in *S.A1.Table 1*. The effect of each cardiovascular risk factor on risk of CVD (hazard ratio) is presented in *S.A1.Table 2*.

S.A1.Table 1. Description of the phenotypic characteristics of the individuals included in the analysis from the REGICOR and from the Framingham Heart Study cohorts.

	All	**None**	**CVD**	**p-value**
REGICOR				
N	2,351	2,19	161	-
Age (years)[a]	53.9 (11.2)	53.3 (11.1)	61.5 (9.52)	<0.001
Gender (male)[b]	1123 (47.8)	1,016 (46.4)	72 (66.5)	<0.001
SBP (mmHg)[a]	132 (20.8)	131 (20.5)	147 (20.1)	<0.001
DBP (mmHg)[a]	79.5 (10.4)	79.3 (10.3)	82.4 (11.5)	0.001
Hypertension[b]	938 (40.1)	822 (37.7)	116 (72.0)	<0.001
Smoking[b]	511 (22.0)	476 (22.0)	35 (21.9)	0.947
Total cholesterol (mg/dL)[a]	225 (42.4)	224 (42.0)	235 (47.3)	0.011
LDL cholesterol (mg/dL)[a]	152 (37.9)	151 (37.7)	161 (40.6)	0.011
HDL cholesterol (mg/dL)[a]	51.7 (13.3)	52.1 (13.2)	46.4 (12.4)	<0.001
Triglycerides (mg/dL)[c]	92 (70-127)	91 (69-125)	116 (82-164)	<0.001
Cholesterol treatment[b]	157 (6.7)	136 (6.2)	21 (13.2)	0.001
Diabetes[b]	316 (13.8)	280 (13.1)	36 (22.9)	0.001
Diabetes treatment[b]	96 (4.11)	74 (3.4)	22 (13.7)	<0.001
Body mass index (kg/m^2)[a]	27.4 (4.47)	27.3 (4.46)	28.8 (4.28)	<0.001
Obesity (BMI≥30 kg/m^2)[b]	596 (25.6)	540 (24.9)	56 (35.2)	0.005
Family history of CHD[b]	272 (11.7)	301 (11.5)	29 (18.1)	0.012
Framingham				
N	3,537	2,863	674	-
Age (years)[a]	56.0 (9.3)	54.8 (9.2)	61.2 (7.4)	<0.001
Gender (male)[b]	1,540 (43.5)	1,190 (41.6)	350 (51.9)	<0.001
SBP (mmHg)[a]	127 (18.3)	125 (17.9)	134 (18.0)	<0.001
DBP (mmHg)[a]	75.0 (9.8)	74.6 (9.8)	76.6 (9.7)	<0.001
Hypertension[b]	1121 (31.7)	802 (28.0)	319 (47.5)	<0.001
Smoking[b]	713 (20.2)	531 (18.5)	182 (27.0)	<0.001
Total cholesterol (mg/dL)[a]	210 (38.6)	207 (37.4)	226 (39.3)	<0.001
LDL cholesterol (mg/dL)[a]	126 (34.0)	124 (33.3)	135 (37.3)	<0.001
HDL cholesterol (mg/dL)[a]	51 (15.2)	52 (15.3)	47 (14.1)	<0.001
Triglycerides (mg/dL)[c]	116 (83-172)	112 (80-164)	157 (107-217)	<0.001
Cholesterol treatment[b]	166 (4.7)	118 (4.1)	48 (7.1)	0.001
Diabetes[b]	226 (6.4)	138 (4.8)	88 (13.1)	<0.001
Diabetes treatment[b]	90 (2.5)	48 (1.7)	42 (6.2)	<0.001
Body mass index (kg/m^2)[a]	27.1 (4.8)	27.0 (4.8)	27.8 (4.5)	<0.001
Obesity (BMI≥30 kg/m^2)[b]	780 (22.1)	604 (21.2)	176 (26.2)	0.005
Family history of CHD[b]	551 (24.8)	478 (24.3)	73 (29.2)	0.089

CVD: individuals who presented a cardiovascular event (includes those with a coronary event); SBP: systolic blood pressure; DBP: diastolic blood pressure; LDL: low density lipoprotein; HDL: high density lipoprotein; BMI: body mass index; CI: confidence interval.

[a] mean (standard deviation); [b] n (proportion, %); [c] median (25 and 75 percentiles); [d] mean (95% confidence interval).

S.A1.Table 2. Effects (hazard ratio) of classical risk factors on risk of cardiovascular events.

	HR [95%CI]	P-value
REGICOR		
Age (10 years)	2.11 [1.79-2.47]	<0.001
Gender (men)	2.27 [1.64-3.23]	<0.001
Total cholesterol (10 mg/dL)	1.05 [1.01-1.09]	0.033
HDL cholesterol (10 mg/dL)	0.69 [0.60-0.79]	<0.001
Systolic BP (10 mmHg)	1.37 [1.29-1.46]	<0.001
Diastolic BP (10 mmHg)	1.37 [1.18-1.58]	<0.001
Diabetes	2.02 [1.39-2.93]	<0.001
Smoker	0.99 [0.68-1.44]	0.957
Family history of CVD[a]	1.59 [1.06-2.37]	0.024
Estimated 10-y CVD risk[b]	1.14 [1.12-1.16]	<0.001
FRAMINGHAM		
Age (10 years)	1.78 [1.61-1.96]	<0.001
Gender (men)	1.75 [1.52-2.04]	<0.001
Total cholesterol (10 mg/dL)	1.07 [1.05-1.09]	<0.001
HDL cholesterol (10 mg/dL)	0.79 [0.75-0.84]	<0.001
Systolic BP (10 mmHg)	1.24 [1.19-1.28]	<0.001
Diastolic BP (10 mmHg)	1.19 [1.10-1.29]	<0.001
Diabetes	2.53 [2.02-3.16]	<0.001
Smoker	1.42 [1.20-1.68]	<0.001
Family history of CVD[c]	1.29 [0.98-1.69]	0.067
Estimated 10-y CVD risk[b]	1.06 [1.05-1.06]	<0.001

[a] CVD: Cardiovascular disease. [b] Coronary risk was calculated using the original Framingham risk function for the Framingham cohort, and the calibrated function for the REGICOR cohort; [c] Only in the Offspring sample.

Validation of the association between the GRS and risk of CVD

The results of the test for association between the genetic variants included in the GRS and incidence of CVD events is shown in *S.A1.Table 3*. The variants nominally associated with CVD events were rs1333049 in *CDKN2A/2B* and rs10455872 in *LPA*. The minimum hazard ratio (HR) we were able to detect with 80% power for each individual variant ranged from 1.36 to 1.64, in REGICOR, from 1.17 to 1.48 in Framingham, and from 1.15 to 1.74 in the meta-analysis (*S.A1.Table 4*).

S.A1.Table 3. Characteristics of the genetic variants included in the multi-locus genetic risk score, magnitude of the association for coronary events in both cohorts and meta-analyses results of the observed effect sizes.

SNP	REGICOR		FRAMINGHAM		Meta-analysis	
	HR[95%CI]	p-value	HR[95%CI]	p-value	HR[95%CI]	p-value
rs17465637	1.03 [0.80-1.31]	0.420	0.99 [0.88-1.11]	0.825	1.00 [0.90-1.11]	0.957
rs6725887	1.30 [0.98-1.74]	0.037	1.07 [0.92-1.25]	0.402	1.13 [0.95-1.35]	0.158
rs9818870	0.99 [0.71-1.39]	0.478	1.13 [0.98-1.30]	0.097	1.11 [0.97-1.26]	0.124
rs12526453	1.02 [0.82-1.29]	0.418	0.95 [0.85-1.07]	0.394	0.96 [0.87-1.07]	0.483
rs1333049	1.12 [0.90-1.39]	0.161	1.23 [1.10-1.37]	<0.001	1.21 [1.09-1.33]	<0.001
rs1746048	1.30 [0.92-1.84]	0.070	0.93 [0.80-1.09]	0.375	1.06 [0.77-1.46]	0.725
rs9982601	1.06 [0.77-1.46]	0.357	1.15 [0.98-1.33]	0.083	1.13 [0.99-1.30]	0.076
rs10455872	1.85 [1.33-2.57]	<0.001	1.25 [0.95-1.64]	0.113	1.50 [1.02-2.21]	0.037

MAF: Minor allele frequency obtained from CEU samples from HapMap; Weight (OR): weight assigned to each genetic variant; HR [95%CI]: Hazard ratio [95% confidence interval].

S.A1.Table 4. Power calculations for cardiovascular disease.

Individual SNPs

			Minimum HR detectable with high or moderate power		Power to detect a specific HR							
	SNP	se	0.8	0.5	1.05	1.09	1.10	1.12	1.14	1.18	1.29	1.35
REGICOR	rs17465637	0.126	1.42	1.28	0.067	0.105	0.118	0.147	0.181	0.260	0.526	0.665
	rs6725887	0.146	1.51	1.33	0.063	0.091	0.100	0.121	0.146	0.204	0.413	0.536
	rs9818870	0.171	1.62	1.40	0.059	0.079	0.086	0.101	0.119	0.162	0.318	0.417
	rs12526453	0.116	1.38	1.25	0.071	0.116	0.131	0.165	0.205	0.299	0.596	0.738
	rs1333049	0.111	1.36	1.24	0.072	0.122	0.138	0.176	0.219	0.320	0.632	0.772
	rs1746048	0.177	1.64	1.41	0.059	0.078	0.084	0.098	0.115	0.155	0.302	0.396
	rs9982601	0.163	1.58	1.38	0.060	0.083	0.090	0.107	0.126	0.174	0.345	0.452
	rs10455872	0.168	1.60	1.39	0.060	0.081	0.088	0.104	0.122	0.166	0.329	0.431
Framingham	rs17465637	0.059	1.18	1.12	0.131	0.307	0.363	0.481	0.600	0.798	0.990	0.999
	rs6725887	0.078	1.24	1.17	0.096	0.197	0.230	0.305	0.388	0.562	0.903	0.970
	rs9818870	0.072	1.22	1.15	0.104	0.223	0.262	0.349	0.444	0.632	0.942	0.986
	rs12526453	0.059	1.18	1.12	0.132	0.312	0.368	0.488	0.607	0.805	0.991	0.999
	rs1333049	0.056	1.17	1.12	0.140	0.337	0.398	0.526	0.648	0.840	0.995	1.000
	rs1746048	0.079	1.25	1.17	0.095	0.194	0.227	0.301	0.382	0.555	0.897	0.967
	rs9982601	0.078	1.24	1.17	0.096	0.198	0.231	0.307	0.391	0.565	0.905	0.971
	rs10455872	0.139	1.48	1.31	0.064	0.095	0.105	0.129	0.156	0.221	0.448	0.577
Meta-analysis	rs17465637	0.054	1.16	1.11	0.149	0.364	0.429	0.563	0.688	0.872	0.997	1.000
	rs6725887	0.090	1.29	1.19	0.085	0.161	0.186	0.244	0.309	0.455	0.811	0.917
	rs9818870	0.067	1.21	1.14	0.113	0.252	0.298	0.397	0.502	0.699	0.968	0.994
	rs12526453	0.053	1.16	1.11	0.152	0.372	0.439	0.574	0.699	0.880	0.998	1.000
	rs1333049	0.051	1.15	1.11	0.161	0.397	0.467	0.607	0.733	0.903	0.999	1.000
	rs1746048	0.163	1.58	1.38	0.060	0.083	0.090	0.107	0.126	0.174	0.345	0.452
	rs9982601	0.069	1.22	1.15	0.108	0.236	0.279	0.371	0.470	0.663	0.956	0.991
	rs10455872	0.197	1.74	1.47	0.057	0.072	0.077	0.089	0.102	0.134	0.252	0.331

GRS

			Minimum HR detectable with high or moderate power		Power to detect a specific HR							
	GRS	se	0.8	0.5	1.05	1.10	1.15	1.20	1.25	1.30	1.40	1.50
REGICOR	Linear	0.046	1.14	1.09	0.185	0.543	0.858	0.977	0.998	1.000	1.000	1.000
	Q2	0.272	2.14	1.70	0.054	0.064	0.081	0.103	0.130	0.161	0.235	0.319
	Q3	0.261	2.08	1.67	0.054	0.065	0.083	0.107	0.137	0.171	0.252	0.342
	Q4	0.244	1.98	1.61	0.055	0.068	0.089	0.116	0.150	0.190	0.282	0.384
	Q5	0.237	1.94	1.59	0.055	0.069	0.091	0.120	0.156	0.197	0.294	0.400
Framingham	Linear	0.023	1.07	1.05	0.549	0.983	1.000	1.000	1.000	1.000	1.000	1.000
	Q2	0.127	1.43	1.28	0.067	0.117	0.196	0.300	0.419	0.542	0.754	0.891
	Q3	0.124	1.42	1.28	0.068	0.120	0.203	0.311	0.435	0.560	0.773	0.904
	Q4	0.123	1.41	1.27	0.068	0.121	0.205	0.315	0.439	0.566	0.778	0.907
	Q5	0.122	1.40	1.27	0.069	0.123	0.210	0.323	0.451	0.579	0.791	0.916
Meta-analysis	Linear	0.028	1.08	1.06	0.424	0.932	0.999	1.000	1.000	1.000	1.000	1.000
	Q2	0.114	1.37	1.25	0.071	0.134	0.234	0.361	0.502	0.637	0.842	0.946
	Q3	0.113	1.37	1.25	0.072	0.134	0.234	0.363	0.504	0.639	0.844	0.947
	Q4	0.110	1.36	1.24	0.073	0.139	0.246	0.381	0.527	0.665	0.864	0.958
	Q5	0.108	1.35	1.24	0.074	0.143	0.253	0.392	0.541	0.679	0.875	0.963

Se: Standard error; 'HR detectable' indicates the minimum risk effect detectable (expressed as the exponent of the beta from the meta-analysis) with high or moderate power. 'Power' indicates the study's power to detect the effects sizes (hazard ratios) shown. In the computation of power for given effect size, scenarios with high power (≥80%) are shaded dark grey, those with moderate power (≥50% and <80%) are shaded light grey, and those with power lower than 50% are unshaded.

The characteristics of the participants within each quintile of the GRS are shown in *S.A1.Table 5*. The GRS was not associated with classical CVRFs but was associated with gender in Framingham.

S.A1.Table 5. Description of the characteristics of the participants across quintiles of the genetic risk score in both cohorts.

	Quintiles of genetic score						
Variables	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**p-value**	**p-trend**
REGICOR							
N	524	416	473	471	467		
Age (years)[a]	54.1 (11.1)	52.9 (11.0)	54.6 (11.4)	54.2 (11.0)	53.6 (11.3)	0.170	0.998
Gender (men)[b]	243 (46.4)	205 (49.3)	217 (45.9)	234 (49.7)	224 (48.0)	0.705	0.581
Total cholesterol (mg/dL)[a]	221 (42.8)	225 (41.8)	227 (42.5)	228 (42.0)	225 (42.8)	0.072	0.049
HDL cholesterol (mg/dL)[a]	51.1 (12.9)	52.4 (13.5)	52.5 (13.4)	51.0 (13.0)	51.5 (13.4)	0.304	0.866
SBP (mmHg)[a]	132.0 (22.0)	131.0 (20.4)	132.0 (20.4)	134.0 (21.5)	132.0 (19.5)	0.278	0.749
DBP (mmHg)[a]	78.9 (10.2)	79.5 (10.8)	79.0 (10.2)	80.2 (10.6)	79.8 (10.0)	0.257	0.099
Diabetes[b]	62 (12.1)	71 (17.5)	66 (14.3)	61 (13.3)	56 (12.3)	0.137	0.590
Smoking[b]	107 (20.7)	87 (21.0)	98 (20.8)	107 (23.1)	112 (24.3)	0.577	0.128
Family history of CHD[b]	46 (8.88)	51 (12.4)	55 (11.6)	63 (13.5)	57 (12.4)	0.207	0.064
Incidence of CVD events[c]	6.46	6.10	5.72	8.42	8.35	0.200	0.028
FRAMINGHAM							
N	743	712	681	711	690		
Age (years)[a]	56.6 (9.10)	56.1 (9.12)	55.6 (9.58)	56.1 (9.12)	55.6 (9.41)	0.172	0.060
Gender (men)[b]	351 (47.2)	321 (45.1)	305 (44.8)	299 (42.1)	264 (38.3)	0.008	<0.001
Total cholesterol (mg/dL)[a]	208 (37.1)	209 (37.6)	213 (39.0)	211 (39.3)	210 (39.8)	0.151	0.242
HDL cholesterol (mg/dL)[a]	50.5 (14.7)	50.2 (14.9)	51.1 (15.2)	52.0 (15.8)	51.3 (15.2)	0.151	0.048
SBP (mmHg)[a]	127 (18.4)	126 (17.0)	127 (18.8)	126 (18.2)	127 (18.9)	0.938	0.647
DBP (mmHg)[a]	75.2 (10.2)	75.1 (9.54)	74.8 (9.81)	75.0 (9.65)	74.7 (9.73)	0.872	0.329
Diabetes[b]	47 (6.33)	59 (8.29)	32 (4.70)	39 (5.49)	49 (7.10)	0.059	0.658
Smoking[b]	132 (17.8)	146 (20.5)	146 (21.4)	140 (19.7)	149 (21.6)	0.358	0.144
Family history of CHD[b]	113 (24.6)	112 (24.7)	105 (24.7)	109 (24.8)	112 (25.3)	0.999	0.763
Incidence of CVD events[c]	8.36	8.99	11.5	10.7	12.8	0.013	0.001

HDL: high density lipoprotein; SBP: systolic blood pressure; DBP: diastolic blood pressure; CHD: coronary heart disease; CVD: cardiovascular disease.

[a] mean (standard deviation); [b] n (proportion, %); [c] number of cases/100 individuals in 10 years.

For the GRS, we estimated that our study had 80% power to detect a HR of 1.14, 1.07 and 1.08 per unit increase in REGICOR, Framingham, and the meta-analysis, respectively (*S.A1.Table 4*). The GRS was linearly associated with incidence of CHD in both cohorts (p=0.002 in REGICOR and p<0.001 in Framingham; *S.A1.Table 6*), and in the meta-analysis, with a ~11% increase in risk of having a CVD event per unit of the GRS (p<0.001; *S.A1.Table 6*). This association remained statistically significant after further adjustment for family history of CHD (HR=1.15; 95% CI: 1.08-1.22). Participants in the top quintile of the GRS had 1.54 times greater risk of CHD, compared to those in the bottom quintile (p-value for linear trend <0.001) (*S.A1.Table 6*)). In both cohorts the distribution of the GRS was slightly shifted to the right in individuals who had had an event, compared to those who had not (*S.A1.Figure 2*).

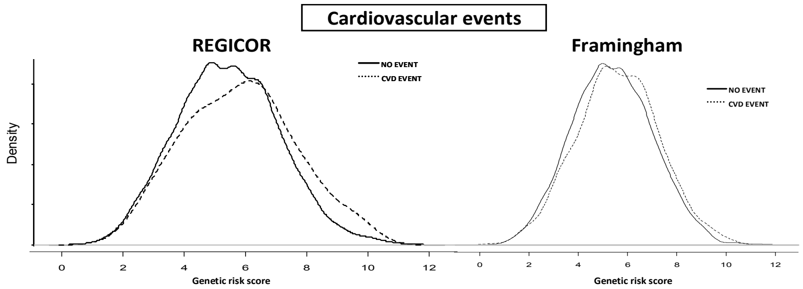
S.A1.Table 6. Multivariate adjusted association between risk of cardiovascular events and the genetic risk score, or quintiles thereof, in both cohorts and meta-analyses results of the observed effect sizes.

Genetic risk score	REGICOR		Framingham		Meta-analysis	
	HR [95%CI][a]	P-value	HR [95%CI][a]	P-value	HR [95%CI][a]	P-value
Linear	1.16 [1.06-1.27]	0.002	1.09 [1.04-1.14]	<0.001	1.11 [1.05-1.17]	<0.001
Quintiles	P-trend	0.018	P-trend	<0.001	P-trend	<0.001
Q1	1	---	1	---	1	---
Q2	1.09 [0.64-1.86]	0.749	1.01 [0.79-1.30]	0.916	1.02 [0.82-1.28]	0.838
Q3	1.00 [0.60-1.67]	0.993	1.20 [0.94-1.53]	0.143	1.16 [0.93-1.45]	0.185
Q4	1.32 [0.82-2.13]	0.255	1.25 [0.98-1.59]	0.075	1.26 [1.02-1.57]	0.033
Q5	1.72 [1.08-2.74]	0.023	1.50 [1.18-1.90]	0.001	1.54 [1.25-1.91]	<0.001

All models were adjusted for the sum of the products of the coefficient for each classical risk factor estimated in the Framingham original and calibrated risk functions and the difference between the participant's value and the population mean of that risk factor (see main text for formula). To account for family structure in the Framingham cohort we also adjusted for the first five genetic principal components.

[a] HR [95%CI]: Hazard ratio [95% confidence interval].

S.A1.Figure 2. Density distribution of genetic risk score in REGICOR and Framingham participants according to the incidence of cardiovascular events during the follow-up. The GRS is represented on the x-axis and is computed as a cumulative sum of all the risk alleles that a person carries, weighted by the effect of each SNP.



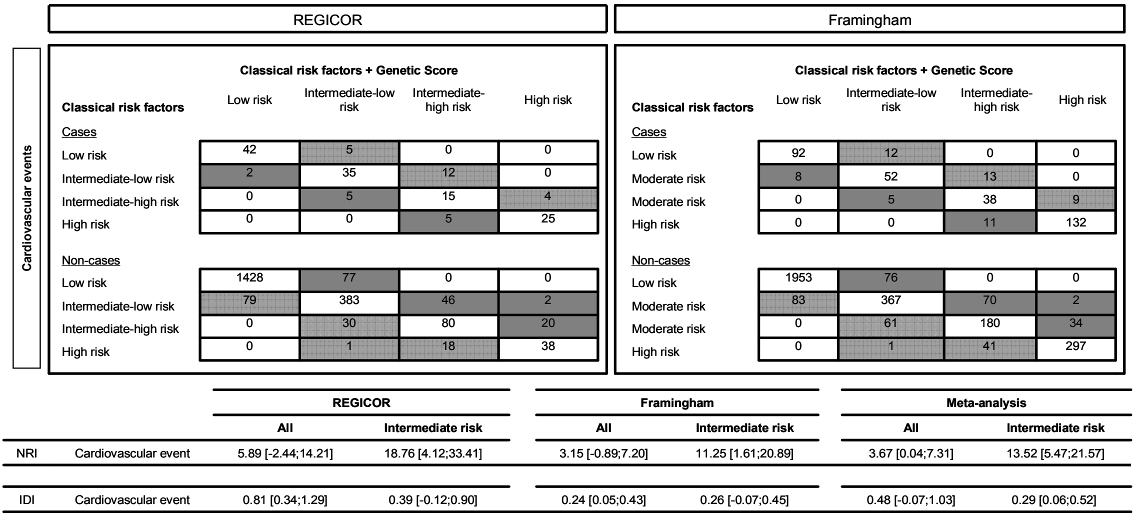
	REGICOR NO EVENT	REGICOR CVD EVENT		Framingham NO EVENT	Framingham CVD EVENT	
	N=2,190	N=161	P-value	N=2,863	N=674	P-value
Weighted genetic score (95%CI)	5.46 [1.67]	5.84 [1.81]	0.010	5.38 [1.62]	5.60 [1.65]	0.002
Coronary risk (95%CI)	3.21 (1.65-5.89)	7.33 (4.40-12.1)	<0.001	7.07 (3.65-12.4)	13.9 (8.08-21.5)	<0.001
Coronary risk + genetic score (95%CI)	3.23 (1.59-5.92)	7.59 (4.50-12.1)	<0.001	6.82 (3.67-12.2)	14.7 (8.39-23.7)	<0.001

Improvement in predictive capacity: discrimination and reclassification

The addition of the GRS to the basic risk function improved its capacity to predict CVD in the Framingham cohort (c-statistic, 73.18 vs. 72.65, p-value=0.005) but not in the REGICOR cohort (76.09 vs. 76.10, p-value=0.621).

We observed a general tendency for both measures of reclassification improvement, the NRI and IDI, to increase after addition of the GRS to the basic risk function, although this improvement was not statistically significant for IDI index in the meta-analysis of the two cohorts. Overall, the NRI index in the meta-analysis was 3.67, 95%CI 0.04-7.31. However, reclassification improvement was more marked in the group with intermediate risk, and was statistically significant for both measures (NRI: 13.52, 95%CI 5.47-21.57; IDI: 0.29, 95%CI 0.06-0.52). Raw reclassification data and NRI and IDI for each cohort are shown in *S.A1.Figure3.*

S.A1.Figure 3. Reclassification of individuals based on the predicted 10-year risk of cardiovascular heart disease with and without the genetic risk score. Four risk categories (low, intermediate-low, intermediate-high and high), with cut-off points defined in each cohort, were defined according to current guidelines in each country (REGICOR: [0-5)%, [5-10)%, [10-15)%, ≥15%; Framingham: [0-10)%, [10-15)%, [15-20)%, ≥20%, respectively). Light grey cells represent an improvement in reclassification and dark grey cells represent the opposite.



Cardiovascular events

REGICOR

Classical risk factors	Classical risk factors + Genetic Score: Low risk	Intermediate-low risk	Intermediate-high risk	High risk
Cases				
Low risk	42	5	0	0
Intermediate-low risk	2	35	12	0
Intermediate-high risk	0	5	15	4
High risk	0	0	5	25
Non-cases				
Low risk	1428	77	0	0
Intermediate-low risk	79	383	46	2
Intermediate-high risk	0	30	80	20
High risk	0	1	18	38

Framingham

Classical risk factors	Classical risk factors + Genetic Score: Low risk	Intermediate-low risk	Intermediate-high risk	High risk
Cases				
Low risk	92	12	0	0
Moderate risk	8	52	13	0
Moderate risk	0	5	38	9
High risk	0	0	11	132
Non-cases				
Low risk	1953	76	0	0
Moderate risk	83	367	70	2
Moderate risk	0	61	180	34
High risk	0	1	41	297

		REGICOR		Framingham		Meta-analysis	
		All	Intermediate risk	All	Intermediate risk	All	Intermediate risk
NRI	Cardiovascular event	5.89 [-2.44;14.21]	18.76 [4.12;33.41]	3.15 [-0.89;7.20]	11.25 [1.61;20.89]	3.67 [0.04;7.31]	13.52 [5.47;21.57]
IDI	Cardiovascular event	0.81 [0.34;1.29]	0.39 [-0.12;0.90]	0.24 [0.05;0.43]	0.26 [-0.07;0.45]	0.48 [-0.07;1.03]	0.29 [0.06;0.52]

4. DISCUSSION

As for CHD events (main manuscript) and in accordance with the AHA statement regarding assessment of the value of novel risk biomarkers s [24], we have validated the association between a multi-locus GRS and incidence of CVD events in two prospective cohort studies, and have shown that this GRS improves the capacity of the Framingham risk function to predict CVD events. In addition, we have also observed greater improvement in risk reclassification among individuals with intermediate risk.

Prospective validation of the association between a novel multi-locus genetic risk score and CHD events

As in the case of CHD, the GRS is linearly and directly associated with the incidence of CVD events in two cohorts with different basal 10-year coronary risks with a ~11% increased risk per unit of the GRS. The association GRS results were similar in both populations and independent of familial history of CHD. As observed for CHD events, this result is mainly driven by the effect size in the Framingham cohort and we believe that the effect size per unit of the GRS could be slightly underestimated.

The 1.54-times increased risk observed for CVD is also very similar to the 1.44-times risk increase in CHD between the extreme quintiles of the GRS.

Incremental value of the genetic risk score for CHD risk prediction

The inclusion of the GRS improved the classification of the individuals in the different risk categories, especially in those individuals with intermediate risk.

The discriminative capacity of the classical risk function was improved by inclusion of the GRS in the Framingham cohort but not the REGICOR.

Risk estimation including information for the GRS in risk functions in individuals with intermediate risk

We observed that the GRS improved the classification of individuals mainly in the intermediate risk group. The results of the NRI for CVD events observed in our study was 13.52%.

Supplementary Analysis 2
Four SNP analysis.

1. METHODS

We sought to evaluate the reclassification of individuals based on the 10-year predicted risk of coronary heart disease, with and without the genetic risk score (GRS), using a GRS composed of the 4 SNPs (rs6725887 [*WDR12*], rs9982601 [*SCL5A3*], rs1333049 [*CDKN2A/2B*], rs10455872 [*LPA*]) that presented consistent effects in the direction of the association in the two cohorts and in the meta-analysis (see *table 2* in the main article).

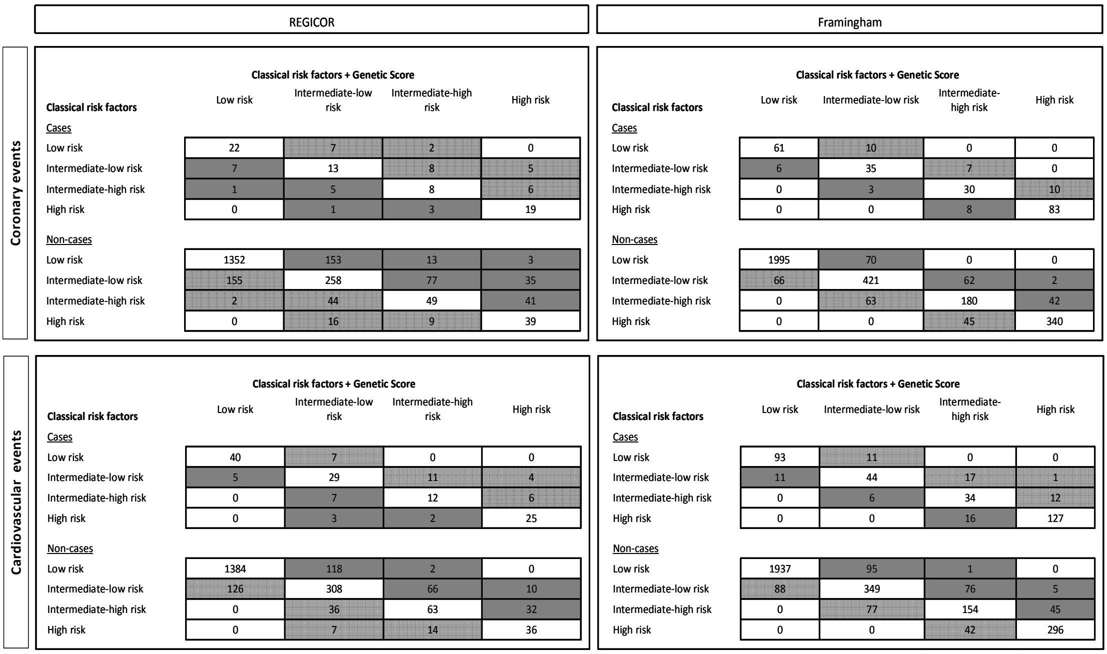
2. RESULTS

S.A2.Table 1. Comparison of the Net Reclassification Index (NRI) results for the analyses using the 4-SNP and 8-SNP scores for the entire sample and separately for the intermediate risk group.

	NRI results obtained using 4-SNP GRS		NRI results obtained using 8-SNP GRS	
	Cardiovascular event	Coronary event	Cardiovascular event	Coronary event
All events				
REGICOR	5.35 [-3.57;14.27]	5.54 [-7.78;18.86]	5.89 [-2.44;14.21]	12.17 [1.99;22.34]
Framingham	2.28 [-2.54;7.11]	3.75 [-1.45;8.95]	3.15 [-0.89;7.20]	11.25 [1.61;20.89]
Meta-analysis	2.97 [-1.27;7.22]	3.99 [-0.86;8.83]	3.67 [0.04;7.31]	13.52 [5.47;21.57]
Intermediate risk				
REGICOR	21.36 [5.05;39.91]	17.71 [-4.49;39.91]	18.76 [4.12;33.41]	24.76 [7.62;41.91]
Framingham	15.10 [4.72;25.47]	18.04 [6.23;29.85]	2.56 [-2.89;8.01]	14.30 [3.08;25.51]
Meta-analysis	16.77 [7.76;25.78]	17.97 [7.54;28.39]	6.37 [-2.85;15.58]	17.44 [8.04;26.83]

Columns 3 and 4 show the NRI results for the 8-SNP GRS from *Figure 2* in the main manuscript. Cell shaded in yellow indicate the results for the score that provided the greatest improvement in reclassification.

S.A2.Figure 1. Reclassification of individuals based on the predicted 10-year risk of coronary heart disease with and without the genetic risk score. Four risk categories (low, intermediate-low, intermediate-high and high), with cut-off points defined in each cohort, were defined according to current guidelines in each country (REGICOR: [0-5)%, [5-10)%, [10-15)%, ≥15%; Framingham: [0-10)%, [10-15)%, [15-20)%, ≥20%, respectively). Light grey cells represent an improvement in reclassification and dark grey cells represent the opposite.



REGICOR — Coronary events

Classical risk factors	Low risk	Intermediate-low risk	Intermediate-high risk	High risk
	Classical risk factors + Genetic Score			
Cases				
Low risk	22	7	2	0
Intermediate-low risk	7	13	8	5
Intermediate-high risk	1	5	8	6
High risk	0	1	3	19
Non-cases				
Low risk	1352	153	13	3
Intermediate-low risk	155	258	77	35
Intermediate-high risk	2	44	49	41
High risk	0	16	9	39

Framingham — Coronary events

Classical risk factors	Low risk	Intermediate-low risk	Intermediate-high risk	High risk
	Classical risk factors + Genetic Score			
Cases				
Low risk	61	10	0	0
Intermediate-low risk	6	35	7	0
Intermediate-high risk	0	3	30	10
High risk	0	0	8	83
Non-cases				
Low risk	1995	70	0	0
Intermediate-low risk	66	421	62	2
Intermediate-high risk	0	63	180	42
High risk	0	0	45	340

REGICOR — Cardiovascular events

Classical risk factors	Low risk	Intermediate-low risk	Intermediate-high risk	High risk
	Classical risk factors + Genetic Score			
Cases				
Low risk	40	7	0	0
Intermediate-low risk	5	29	11	4
Intermediate-high risk	0	7	12	6
High risk	0	3	2	25
Non-cases				
Low risk	1384	118	2	0
Intermediate-low risk	126	308	66	10
Intermediate-high risk	0	36	63	32
High risk	0	7	14	36

Framingham — Cardiovascular events

Classical risk factors	Low risk	Intermediate-low risk	Intermediate-high risk	High risk
	Classical risk factors + Genetic Score			
Cases				
Low risk	93	11	0	0
Intermediate-low risk	11	44	17	1
Intermediate-high risk	0	6	34	12
High risk	0	0	16	127
Non-cases				
Low risk	1937	95	1	0
Intermediate-low risk	88	349	76	5
Intermediate-high risk	0	77	154	45
High risk	0	0	42	296

3. DISCUSSION

The results obtained for the NRI using only the 4 SNPs that presented the same direction of effect both in the REGICOR and Framingham studies, showed that although the SNPs were selected on the basis on the results they have in both cohorts, we still gain more information from the full set of SNPs independent from CVRFs.

Supplementary Analysis 3

Predictive capacity analysis without CDKN2A-2B variant

1. INTRODUCTION

Genetic variants in the chromosomal region 9p21.3, specifically between the genes *CDKN2A* and *CDKN2B*, have been identified by GWAS studies as being associated with several complex diseases, including *Abdominal aortic aneurysm, Breast cancer, Coronary heart disease, Glioma, Intracranial aneurysm, Melanoma, Myocardial infarction* and *Type 2 diabetes* (NHGRI GWAS catalog, accessed in 17th November 2011). Although some variants in this region are known to be associated with T2D, we included in our GRS a variant from chromosomal region 9p21 that is known to be associated with MI/CHD risk independently of T2D risk [25].

In the present analysis we evaluated the sensitivity of our analysis to the inclusion of this variant, not only to avoid the possibility of including a variant that could have some undetected association with T2D, but also because this variant has the largest effect on risk (OR=1.29, according to the CARDIoGRAM study). Our aim was to evaluate if the results in the main analyses are mainly driven variant.

2. RESULTS

S.A3.Table 1. Description of the characteristics of the participants across genetic risk score quintiles in both cohorts.

	Quintiles of genetic score						
Variables	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**p-value**	**p-trend**
REGICOR							
N	511	439	502	438	461		
Age (years)[a]	54.7 (11.2)	52.5 (11.1)	53.6 (11.2)	53.5 (11.2)	55.1 (11.1)	0.005	0.343
Gender (men)[b]	247 (48.3)	207 (47.2)	231 (46.0)	204 (46.6)	234 (50.8)	0.617	0.577
TC (mg/dL)[a]	223 (41.8)	224 (40.6)	226 (43.4)	227 (44.9)	226 (41.6)	0.608	0.135
HDLc (mg/dL)[a]	50.8 (12.6)	52.9 (13.4)	52.5 (13.8)	51.1 (13.2)	51.2 (13.2)	0.058	0.695
SBP (mmHg)[a]	133 (21.9)	132 (21.4)	130 (20.2)	132 (20.3)	134 (20.0)	0.139	0.753
DBP (mmHg)[a]	79.3 (10.5)	80.0 (10.5)	78.9 (10.4)	79.0 (10.2)	80.3 (10.2)	0.151	0.444
Diabetes[b]	73 (14.7)	61 (14.3)	61 (12.3)	67 (15.8)	54 (11.9)	0.404	0.400
Smoking[b]	102 (20.2)	98 (22.4)	106 (21.4)	93 (21.6)	112 (24.4)	0.621	0.202
CHD Family hist [b]	55 (10.8)	39 (9.01)	53 (10.7)	68 (15.7)	57 (12.5)	0.028	0.038
Estimated 10-y coronary risk[c]	3.6 (1.9-6.6)	3.1 (1.4-5.5)	3.1 (1.7-5.9)	3.2 (1.6-6.5)	3.6 (1.9-6.3)	0.015	0.299
Incidence of CVD events[d]	6.23	5.98	5.94	6.82	10.3	0.004	0.004
Incidence of coronary events[d]	4.43	3.93	3.84	4.95	7.95	0.004	0.002

FRAMINGHAM							
N	743	712	681	711	690		
Age (years)[a]	56.3 (9.18)	56.4 (9.12)	55.6 (9.44)	56.0 (9.32)	55.7 (9.27)	0.389	0.145
Gender (men)[b]	371 (50.2)	299 (42.2)	316 (46.2)	282 (41.0)	272 (37.9)	<0.001	<0.001
TC (mg/dL)[a]	209 (37.5)	211 (37.7)	209 (38.5)	209 (38.6)	213 (40.4)	0.158	0.233
HDLc (mg/dL)[a]	50.4 (14.5)	51.0 (14.8)	50.9 (15.7)	51.1 (15.4)	51.8 (15.4)	0.532	0.103
SBP (mmHg)[a]	126 (17.4)	127 (18.3)	127 (19.2)	126 (17.9)	127 (18.6)	0.785	0.941
DBP (mmHg)[a]	75.0 (9.61)	75.3 (9.70)	75.5 (10.3)	74.6 (9.82)	74.4 (9.55)	0.230	0.131
Diabetes[b]	48 (6.50)	53 (7.49)	40 (5.85)	39 (5.67)	46 (6.41)	0.668	0.499
Smoking[b]	135 (18.3)	140 (19.8)	138 (20.2)	135 (19.6)	165 (23.0)	0.250	0.048
CHD Family hist [b]	113 (24.6)	112 (24.7)	105 (24.7)	109 (24.8)	112 (25.3)	0.999	0.763
Estimated 10-y coronary risk[c]	8.6 (4.7-14.5)	8.1 (4.6-14.1)	8.1 (4.4-14.3)	7.5 (4.5-13.3)	7.8 (4.1-14.1)	0.342	0.041
Incidence of CVD events[d]	10.40	11.10	10.70	8.06	12.50	0.200	0.369
Incidence of coronary events[d]	7.20	7.38	7.34	5.43	8.72	0.210	0.672

HDLc: high density lipoprotein cholesterol; SBP: systolic blood pressure; DBP: diastolic blood pressure; CHD: coronary heart disease; CVD: cardiovascular disease; TC: Total cholesterol; CHD Family hist: CHD Family history.

[a] mean (standard deviation); [b] n (proportion, %); [c] mean (95% confidence interval); [d] number of cases/100 individuals in 10 years.

S.A3.Table 2. Multivariate adjusted association of the genetic risk score with cardiovascular and coronary events as a linear variable and across quintiles in both cohorts and meta-analyses results of the observed effect sizes.

	Genetic	REGICOR		Framingham		Meta-analysis	
	risk score	**HR [95%CI][a]**	**P-value**	**HR [95%CI][a]**	**P-value**	**HR [95%CI][a]**	**P-value**
Cardiovascular events	**Linear**	1.21 [1.08-1.35]	0.001	1.05 [0.99-1.12]	0.099	1.12 [0.97-1.28]	0.113
	Quintiles	P-trend	0.0050	P-trend	0.452	P-trend	0.235
	Q1	1	---	1	---	1	---
	Q2	1.02 [0.60-1.73]	0.944	0.92 [0.73-1.17]	0.515	0.94 [0.75-1.16]	0.546
	Q3	0.86 [0.50-1.45]	0.566	1.03 [0.81-1.31]	0.801	1.00 [0.80-1.24]	0.993
	Q4	1.19 [0.73-1.94]	0.487	0.87 [0.68-1.12]	0.278	0.95 [0.72-1.24]	0.685
	Q5	1.87 [1.19-2.91]	0.006	1.13 [0.89-1.42]	0.316	1.40 [0.86-2.28]	0.177
Coronary events	**Linear**	1.26 [1.10-1.43]	0.001	1.05 [0.97-1.13]	0.247	1.14 [0.95-1.36]	0.147
	Quintiles	P-trend	0.0024	P-trend	0.781	P-trend	0.318
	Q1	1	---	1	---	1	---
	Q2	0.88 [0.44-1.77]	0.718	0.98 [0.73-1.31]	0.874	0.96 [0.74-1.26]	0.792
	Q3	0.90 [0.47-1.74]	0.760	1.00 [0.74;1.35]	0.995	0.98 [0.75-1.29]	0.895
	Q4	1.36 [0.75-2.48]	0.311	0.80 [0.59-1.11]	0.179	0.98 [0.59-1.62]	0.935
	Q5	2.10 [1.21-3.64]	0.008	1.13 [0.85-1.51]	0.412	1.47 [0.81-2.68]	0.208

All models were adjusted for the sum of the products of the coefficient for each classical risk factor estimated in the Framingham original and calibrated risk functions and the difference between the participant's value and the population mean of that risk factor (see main text for formula). To account for family structure in the Framingham cohort we also adjusted for the first five genetic principal components. [a] HR [95%CI]: Hazard ratio [95% confidence interval].

Cell shaded in yellow indicate the results for the score that provided a more significant association between the GRS and risk of CVD or CHD events.

S.A3.Table 3. Comparison of the Net Reclassification Index (NRI) results for the 7-SNP score (GRS of the main analysis without the variant of Chromosome 9: *CDKN2A-2B*) and 8-SNP score analyses, for the entire sample and separately for the intermediate risk group.

	NRI results obtained with 7 SNPs GRS		NRI results obtained with 8 SNPs GRS	
	Cardiovascular event	Coronary event	Cardiovascular event	Coronary event
All individuals				
REGICOR	6.76 [-1.60;15.11]	11.02 [-0.78;22.82]	5.89 [-2.44;14.21]	12.17 [1.99;22.34]
Framingham	3.15 [-1.02;7.32]	2.56 [-2.89;8.01]	3.15 [-0.89;7.20]	11.25 [1.61;20.89]
Meta-analysis	3.87 [0.14;7.60]	5.10 [-2.50;12.71]	3.67 [0.04;7.31]	13.52 [5.47;21.57]
Intermediate risk				
REGICOR	21.80 [6.82;36.79]	21.91 [2.25;41.56]	18.76 [4.12;33.41]	24.76 [7.62;41.91]
Framingham	11.25 [1.60;20.90]	14.30 [3.82;24.77]	2.56 [-2.89;8.01]	14.30 [3.08;25.51]
Meta-analysis	14.90 [5.07;27.74]	15.98 [6.74;25.23]	6.37 [-2.85;15.58]	17.44 [8.04;26.83]

The two columns presented for NRI results obtained with a GRS composed of 8 SNPs are the ones presented in the main document.

Cell shaded in yellow indicate the results for the score that provided the greatest improvement in reclassification.

3. DISCUSSION

The results shown in *S.A3.Table 2* and *S.A3.Table 3* suggest that, although the results do not change markedly after excluding the variant on 9p21, it is mainly in the Framingham Heart study that this variant evaluated has a greater effect on the GRS, and in some cases it can drive the meta-analyses to a significant result. This is consistent with the effect sizes observed for the individual SNPs in each cohort, because this variant presents a HR lower than the average in the REGICOR study, and the opposite scenario for both the Framingham and meta-analysis (see *table 2* in the main article).

Supplementary Analysis 4

Predictive capacity analysis with a 12-SNP based GRS in the Framingham cohort

1. METHODS

We sought to evaluate the reclassification of individuals based on the 10-year predicted risk of coronary heart disease, with and without the genetic risk score (GRS), using a GRS composed of the 12 SNPs (rs17465637 [*MIA3*]; rs6725887 [*WDR12*]; rs9818870 [*MRAS*]; rs12526453 [*PHACTR1*]; rs1333049 [*CDKN2A/2B*]; rs1746048 [*CXCL12*]; rs9982601 [*SCL5A3*]; rs10455872 [*LPA*];) representing the addition of 4 additional SNPs obtained from refs [3,26].

2. RESULTS

S.A4.Table 1. Multivariate adjusted association between the genetic risk score and risk of coronary events as a continuous variable and between quintiles.

	Coronary event		Cardiovascular event	
Genetic risk score	HR (95% CI)	p-value	HR (95% CI)	p-value
Continuous	1.06 (1.01-1.11)	0.013	1.08 (1.04-1.12)	<0.001
Quintiles	p-trend	0.017	p-trend	<0.001
Q1	1	--	1	--
Q2	1.08 (0.80-1.46)	0.628	1.08 (0.84-1.39)	0.531
Q3	1.05 (0.78-1.43)	0.737	1.17 (0.91-1.50)	0.221
Q4	1.28 (0.95-1.71)	0.104	1.33 (1.05-1.70)	0.020
Q5	1.36 (1.02-1.81)	0.039	1.52 (1.20-1.93)	0.001

S.A4.Table 2. Reclassification of individuals based on the 10-year predicted risk of coronary heart disease with and without the genetic risk score. Risk categories were defined using national recommendations. Cut-off points: low [0-10)%, intermediate-low [10-15)%, intermediate-high [15-20)% and high =20% risk.

		ALL	Intermediate risk
NRI	Coronary event	0.91 [-4.38;6.21]	7.80 [-1.76;17.36]
	Cardiovascular event	1.30 [-3.16;5.76]	10.55 [0.40;20.70]
IDI	Coronary event	0.22 [0.04; 0.41]	0.22 [-0.06; 0.49]
	Cardiovascular event	0.27 [0.09; 0.46]	0.25 [-0.03; 0.54]

SUPPLEMENTARY REFERENCES

[1] Lluis-Ganella C, Lucas G, Subirana I, *et al*. Additive Effect of Multiple Genetic Variants on the Risk of Coronary Artery Disease. Rev Esp Cardiol 2010;8:925-933.

[2] Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. 2009.

[3] The CARDIoGRAM Consortium. Large-Scale Association Analysis Identifies 13 New Susceptibility Loci for Coronary Artery Disease. Nat Genet 2011;4:333-338.

[4] Shiffman D, Louie JZ, Rowland CM, Malloy MJ, Kane JP, Devlin JJ. Single Variants Can Explain the Association Between Coronary Heart Disease and Haplotypes in the Apolipoprotein(a) Locus. Atherosclerosis 2010;1:193-196.

[5] Miller SA, Dykes DD, Polesky HF. A Simple Salting Out Procedure for Extracting DNA From Human Nucleated Cells. Nucleic Acids Res 1988;3:1215.

[6] Grau M, Subirana I, Elosua R, *et al*. Why Should Population Attributable Fractions Be Periodically Recalculated? An Example From Cardiovascular Risk Estimation in Southern Europe. Prev Med 2010;1:78-84.

[7] Cupples LA, D'Agostino RB, Kiely D. The Framingham Heart Study, Section 35. An Epidemiological Investigation of Cardiovascular Disease. Survival Following Cardiovascular Events: 30 Year Follow-Up. 1988.

[8] Grau M, Elosua R, Cabrera DL, *et al*. Cardiovascular Risk Factors in Spain in the First Decade of the 21st Century, a Pooled Analysis With Individual Data From 11 Population-Based Studies: the DARIOS Study. Rev Esp Cardiol 2011;4:295-304.

[9] Grau M, Subirana I, Elosua R, *et al*. Trends in Cardiovascular Risk Factor Prevalence (1995-2000-2005) in Northeastern Spain. Eur J Cardiovasc Prev Rehabil 2007;5:653-659.

[10] Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An Investigation of Coronary Heart Disease in Families. The Framingham Offspring Study. Am J Epidemiol 1979;3:281-290.

[11] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007;3:559-575.

[12] Jombart T, Devillard S, Balloux F. Discriminant Analysis of Principal Components: a New Method for the Analysis of Genetically Structured Populations. BMC Genet 2010;11:94.

[13] Ma J, Amos CI. Theoretical Formulation of Principal Components Analysis to Detect and Correct for Population Stratification. PLoS One 2010;9:e12510.

[14] Schwarzer G. Meta: Meta-Analysis. R Package Version 0.8-2. 2007.

[15] D'Agostino RB, Nam BH. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. 2003;1-25.

[16] Newson R. Confidence Intervals for Rank Statistics: Somers' D and Extensions. Stata Journal 2006;309-334.

[17] Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of Net Reclassification Improvement Calculations to Measure Usefulness of New Biomarkers. Stat Med 2011;1:11-21.

[18] Steyerberg EW, Pencina MJ. Reclassification Calculations for Persons With Incomplete Follow-Up. Ann Intern Med 2010;3:195-196.

[19] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation 1998;18:1837-1847.

[20] Marrugat J, Solanas P, D'Agostino R, *et al*. Coronary Risk Estimation in Spain Using a Calibrated Framingham Function. Rev Esp Cardiol 2003;03:225-227.

[21] Pyorala K, De Backer G, Graham I, Poole-Wilson P, Wood D. Prevention of Coronary Heart Disease in Clinical Practice: Recommendations of the Task Force of the European Society of Cardiology, European Atherosclerosis Society and European Society of Hypertension. Atherosclerosis 1994;2:121-161.

[22] De Backer G, Ambrosioni E, Borch-Johnsen K, *et al.* European Guidelines on Cardiovascular Disease Prevention in Clinical Practice. Third Joint Task Force of European and Other Societies on Cardiovascular Disease Prevention in Clinical Practice. Eur Heart J 2003;17:1601-1610.

[23] Graham I, Atar D, Borch-Johnsen K, *et al*. European Guidelines on Cardiovascular Disease Prevention in Clinical Practice: Full Text. Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by Representatives of Nine Societies and by Invited Experts). Eur J Cardiovasc Prev Rehabil 2007;S1-113.

[24] Hlatky MA, Greenland P, Arnett DK, *et al*. Criteria for Evaluation of Novel Markers of Cardiovascular Risk: a Scientific Statement From the American Heart Association. Circulation 2009;17:2408-2416.

[25] Broadbent HM, Peden JF, Lorkowski S, *et al*. Susceptibility to Coronary Artery Disease and Diabetes Is Encoded by Distinct, Tightly Linked SNPs in the ANRIL Locus on Chromosome 9p. Hum Mol Genet 2008;6:806-814.

[26] Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet. 2011;43:339-344.