UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

SCHOOL OF MEDICINE

Informatics and molecular studies of SNPs, in relation to metabolic and cardiovascular phenotypes

By

Mikkel Bjerregaard Christensen

Thesis for the degree of Doctor of Philosophy

June 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES
SCHOOL OF MEDICINE

Doctor of Philosophy

Informatics and molecular studies of SNPs, in relation to metabolic and cardiovascular phenotypes

by Mikkel Bjerregaard Christensen

The aim of human genetics is to find the variation in the genome, which underlies phenotypic traits and leads to diseases. The degree to which the genomic variation decides the phenotype varies. Monogenetic diseases such as Cystic fibrosis are caused by a rare mutation in a single gene. Alzheimer's disease is caused by variation in several genes with moderate effect. Complex diseases are caused by lifestyle, the environment and genomic variation. The most abundant form of variation in the human genome is single nucleotide polymorphisms *SNPs*. The aim of this thesis was to investigate the genetic background for complex diseases with focus on coronary heart disease and its risk traits. To do this it combines informatics, molecular genetics and statistics.

Due to the nature of complex diseases, many claimed associations between genes and phenotypes in the literature are uncertain. A set of computer programs *Perl scripts* were written to go through Medline abstracts and retrieve information about gene phenotype associations. The computer program identifies gene-phenotype associations with specificity of 83%, precision of 68% and a balance F measure of 75%. The summary of previous studies makes it possible to select candidate genes and found prior hypotheses.

The genes *LTA*, *TNF* (located on Chr 6) and *LGALS2* (located Chr 22) were tested for their association with metabolic syndrome traits and myocardial infarction. Four SNPs in *LTA* (+80/81, +252, T26N) two in *TNF* (-238, -308) and one tagSNP in *LGALS2* (rs7291467) were genotyped in a cohort of 3500 British women aged 60 to 79 (BWHHS). All SNPs were analysed using linear regression. The *LTA* and *TNF* SNPs were also analysed together as haplotypes. Rs7291467 was found to be associated with insulin-glucose profile, the mean difference in fasting insulin per minor allele was -4% (p=0.01 for trend by allele) and the mean difference in fasting glucose per minor allele was -1% (p=0.02 for trend by allele). An *LTA* haplotype *2 (31%) (TGT-Thr) was associated with a decrease of 0.4 BMI units and 0.064 mM triglycerides compared to *1 (37%) (GGC-Asn) (p=0.01). Analysed by itself a haplotype harbouring *TNF* -308 A is associated with lower BMI but analysed together with *LTA* this association is explained by LTA +252 C. No associations were found with coronary heart disease.

Content

Tables and figures

Acknowledgements

I would like to thank Professor Ian Day for the opportunity to conduct a PhD in his group and for supporting me in the development of a bioinformatics strategy. The association study performed in this PhD was only possible because of the work of many others. I would therefore like to thank the following persons for their excellent work and the good times we had together: Tom R Gaunt, Santi Rodriguez, Lesley Hinks, Tricia Briggs, Sylvia Diaper, Matt Kiessling, Nikki Graham, Nicola Ball and Jamil Baban. I would also like to thank the Bristol group[1]: Debbie A Lawlor, Nicholas J Timpson, George Davey Smith and Shah Ebrahim. Last but not least all the Volunteers contributing to the BWHHS study and the British Heart Foundation.

[1] Genetic and Molecular Epidemiology, Department of Social Medicine, Canynge Hall, Whiteladies Road Clifton,Bristol,BS8 2PR

Abbreviation

Abs Abstracts
BC Bayesian Classifier
BMI Body Mass Index
BP Blood Pressure
BWHHS British Women's Heart and Health Study
CAD Coronary Artery Disease
CD Crohn's Disease
cDNA Contact DNA
CHD Coronary Heart Disease
Chr Chromosome
CI Confidence Intervals
DNA Deoxyribonucleic acid
dNTP Deoxynucleotide triphosphate
DOP Degenerated Oligo Primer
ELISA Enzyme-linked immunosorbent assay
FFA Free Fatty Acid
GO Gene Ontology
HDL High Density Lipoprotein
HDL-c High density lipoprotein cholesterol
HLA Human Leukocyte Antigen
HMM Hidden Markov Models
HUGO http://www.gene.ucl.ac.uk/nomenclature
HWE Hardy-Weinberg Equilibrium
Kb Kilo bases
LD Linkage Disequilibrium
LDL Low Density Lipoprotein
MAF Minor Allele Frequency
MeSH Medline Subject Headings
MI Myocardial Infarction
MR Mendelian Randomization
NCBI National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov)
NN Neural Networks
OMIM Online Mendelian Inheritance in Man http://www.ncbi.nlm.nih.gov
OR Odds Ratio
PCR Polymerase Chain Reaction
RA Rheumatoid Arthritis
RCT Randomized Controlled Trials
SD Standard Deviation
SNP Single Nucleotide Polymorphism
T2D/T1D Type 2/1 diabetes
tRNA Transfer ribonucleic acid
WHR Waist Hip Ratio
WTCCC Wellcome Trust Case Control Consortium

Preface

Coronary heart disease is the most important cause of death in British women (1, 2). The environmental risk factors such as smoking, age, hypertension, obesity, homocysteine and raised triglyceride levels are well established (1, 2). Coronary heart disease is believed to have an important genetic component (1-3), but no gene has been established as risk factor for CHD. It has become clear that large-scale association studies are necessary to find genes that predisposes to CHD. This has led to a need for informatics support structures and a change in the scope of association studies from single SNPs to haplotypes. The work presented in this thesis is part of that development. This thesis has two distinct parts.

Chapter 2, 3 and 4 deals with building a bioinformatics framework to support large-scale association studies. Chapter 2 focuses on making a tool for automated prediction of candidate genes and Chapter 3 is about making the necessary database structures. Chapter 4 is on choosing the genes and SNPs for investigation in BWHHS.

Chapter 5 and 6 is a large-scale association study. Chapter 5 in relation to myocardial infarction and the metabolic syndrome and Chapter 6 on the subject of nutritional genetics in collaboration with the Bristol epidemiology group.

Chapter 1: Association Studies

1.1 Introduction

1.1.1 Monogenic Diseases

Since 1980 approximately 1200 diseases have had their underlying genes found (4). Most of them are so-called monogenic diseases where only one gene is responsible for the phenotype (4). If one summarises what we have learned about mutations and diseases in the last ten years three trends become clear (3, 4) The first is the distribution: miss-sense and nonsense mutations are the most frequently seen disease mutations and they account for 60%, deletions account for 30% and only a fraction is accounted for by regulatory (promoter) mutations (4). Secondly we can learn that the more dramatic an amino acid change, the more likely is it that it will lead to a disease (4). A dramatic change is from a small amino acid to a big bulky amino acid or from a positively charged to a negatively charged amino acid. Thirdly the more conserved the region where the mutation takes place, the more severe is the disease (4) (consistent with conserved regions are functionally important). One critical question is how much we can rely on these observations for monogenic diseases when we try to find complex diseases. There are only a few examples where complex disease genes have been found, but for the few established disease genes it seems that the distribution is the same (5).

1.1.2 Complex Diseases

One way to identify genes involved in complex human traits is by a statistical test of association of genetic polymorphisms with disease phenotypes (3). The hypothesis is that common genetic variants explain the predisposition to common complex diseases (3). For cardiovascular disease, this approach started ten years ago, when a paper in Nature claimed an association between an insertion-deletion polymorphism in the angiotensin-converting enzyme (ACE) and myocardial infarction (6). Since then, more than 200 genes have been claimed to be associated with cardiovascular traits or cardiovascular diseases (5), but few have been consistently reproducible (7, 8). The inability to repeat many results is common among association studies (7). Recent reviews and a meta-analysis of genetic association studies have concluded that no more than 20% of positive associations in the literature are true (5, 9). It is also clear that replication in large studies is needed for confirmation or rejection of previous claimed associations.

1.1.3 Coronary heart disease genetics

Coronary heart disease is considered a complex disease. This implies that many genes are suspected to contribute to the phenotype and that a single gene is neither necessary nor sufficient for the phenotype (2, 3). It is also suspected that interaction with the environment and among genes influences the phenotype (2).

As many as 1300 genes have been suggested to be involved in cardiovascular diseases based on their function and pathways (Omim search), and more

than 261 genes have been studied in association studies (pubmed search). In contrast to the large number of studies there are few, if any, well-established associations between genes and cardiovascular diseases (5). This has led to debate. The largest problem is that association studies are not consistently reproducible. Recent reviews have concluded that underpowered studies and failure to exclude chance are the biggest problems(7). They have also inferred that there are true positive associations, and these can be identified in studies using a large sample size. One estimate states that about 20% of the present associations are true (9).

The important implication for the study design is that it must be able to detect a weak signal for each gene and that it would be desirable to genotype many genes. One way to find the true associations is to test all candidate genes systematically but this is not feasible as the number of candidate genes exceeds our capacity. In addition, if many genes are tested, it would need a very strict p value to rule out chance (7). It is therefore necessary to design the experiments in order to either meet the requirement for a strict p value ($5*10^{-5}$) or create a well-founded prior hypothesis (7). One way to form a prior hypothesis is to base it on solid knowledge about the disease and biochemical pathways. Another way is to use the existing association studies to form a new hypothesis. To what extent this new hypothesis will need a less strict p value depends on the number of previous positive associations. It is necessary to estimate this parameter for all candidate genes and it is not possible to evaluate each individual gene. A few genes are settled in the literature. This is true for *APOE*, *ACE* and *MTHFR* associated with the risk factors cholesterol, blood pressure and homocysteine respectively (2). To allow as many genes to be included in the study as possible, it will be desirable to restrict the number of SNPs per gene.

This leads to two questions. What group of SNPs is most likely to cause a disease and how many of them is it necessary to genotype? It is in general accepted that the SNPs that are capable of altering the final gene product must be expected to be the SNPs that cause the disease (in contrast to intron SNPs). The most obvious candidate SNPs are replacement SNPs (non-synonymous), promoter SNPs and SNPs located in splice sites; but SNPs located in the exon splicing enhancer site and other regulatory elements are also interesting. SNPs that change an amino acid codon, but not the amino acid (synonymous) are often regarded proxy markers for non-synonymous SNPs. But because each codon has its own tRNA and not all tRNAs are equally represented in the cell (10), a synonymous SNP could, on the translation state, have the effect of a stop codon or a promoter SNP. For example by making the ribosomes disassemble or slow down (11-13). It is estimated that there are 0.6 replacement SNPs per 1000kb cDNA on average, but with big difference from gene to gene (15 fold) (14-16). This means that it is possible to type all replacements SNPs. How many more SNPs it would be desirable to type depends on what SNPs that are known in other interesting areas and how many SNPs it takes to define the majority of haplotypes. One study in *ACE* defined seven haplotypes that covered 90% of the population by typing ten SNPs (17). The HapMap[2] project was designed to assist researchers in choosing SNPs that tags haplotypes (18).

1.1.4 Population Genetics

This section gives a brief description of the forces that can alter the genetic variation of populations (19, 20). The two main generators of novel variation are mutation and recombination (19). Mutations are created when one of the four DNA

[2] http://www.hapmap.org

bases (Thymine (T), Cytosine (C), Adenine (A), Guanine (G)) is chemical altered into one of the others, most often from T to C or A to G (15, 16). Recombination takes place at meiosis when two homologous chromatids "swaps" adjacent DNA molecules using a DNA break and repair mechanism. The detailed molecular mechanism is not known but it is not random which has important implications for association studies (see next section) (15). Random genetic drift is the increase or decrease in allele frequencies, because segregation of alleles is not exactly 50% (19). Immigration and admixture are of special interest because they can divide a population into groups with different allele frequencies. This can be a problem in population genetic studies, especially case controls studies, if the cases and controls are sampled unequal from the groups (21). If one group has a higher incidence of the phenotype under study, alleles with the highest frequency in this group will appear to be associated with the phenotype, but the association will be false. This scenario is often referred to as stratification (15, 21, 22). Immigration will only create a difference in allele frequency for one generation, but admixture (due to religious, cultural and social reasons) can make the difference in allele frequency permanent (19). Founder effect refers to the special case of migration where a small group of people leaves a population to start a new population (19). Selection of beneficial alleles is what is thought to drive long-term evolution (19). But evidence of selection pressure on modern humans which have resulted in altered genetic variation has only just been investigated (18) and genetic drift is possibly the most important factor in determining allele frequency (16, 19). How the different forces affect variation within or between populations is listed in table 1 (19, 20)

Table 1: Population genetic forces

Force	Variation within populations	Variation between populations
Inbreeding or genetic drift	-	+
Mutation	+	-
Migration	+	-
Selection:		
Directional	-	+/-
Balancing	+	-
Incompatible	-	+

1.1.5 LD, haplotypes and the HapMap

Recombination of the genome is not a random process- it occurs in some regions more often than others. This gives the genome a special pattern of high recombination (called hot-spots) and long stretches of low recombination (called haplotype blocks). This pattern has been established in large sequencing studies (18, 23). This property of the genome can be used to reduce the number of SNPs one needs to genotype in associations studies, which is the main reason for the study of the recombination pattern in the genome(18, 24).

The international HapMap project has completed the genotyping of 4 million SNPs in four populations (see section 1.1.5.2 for details) (25) and the genotype frequencies are available to researchers from the website (25). The genotype frequencies can be used to calculate the correlation between SNPs and for graphical illustration of LD (25, 26); Figure 1 and 2 shows the haplotype block structure around *LGALS2, LTA* and *TNF*. This information helps researchers to select a reduced set of SNPs (tagSNPs

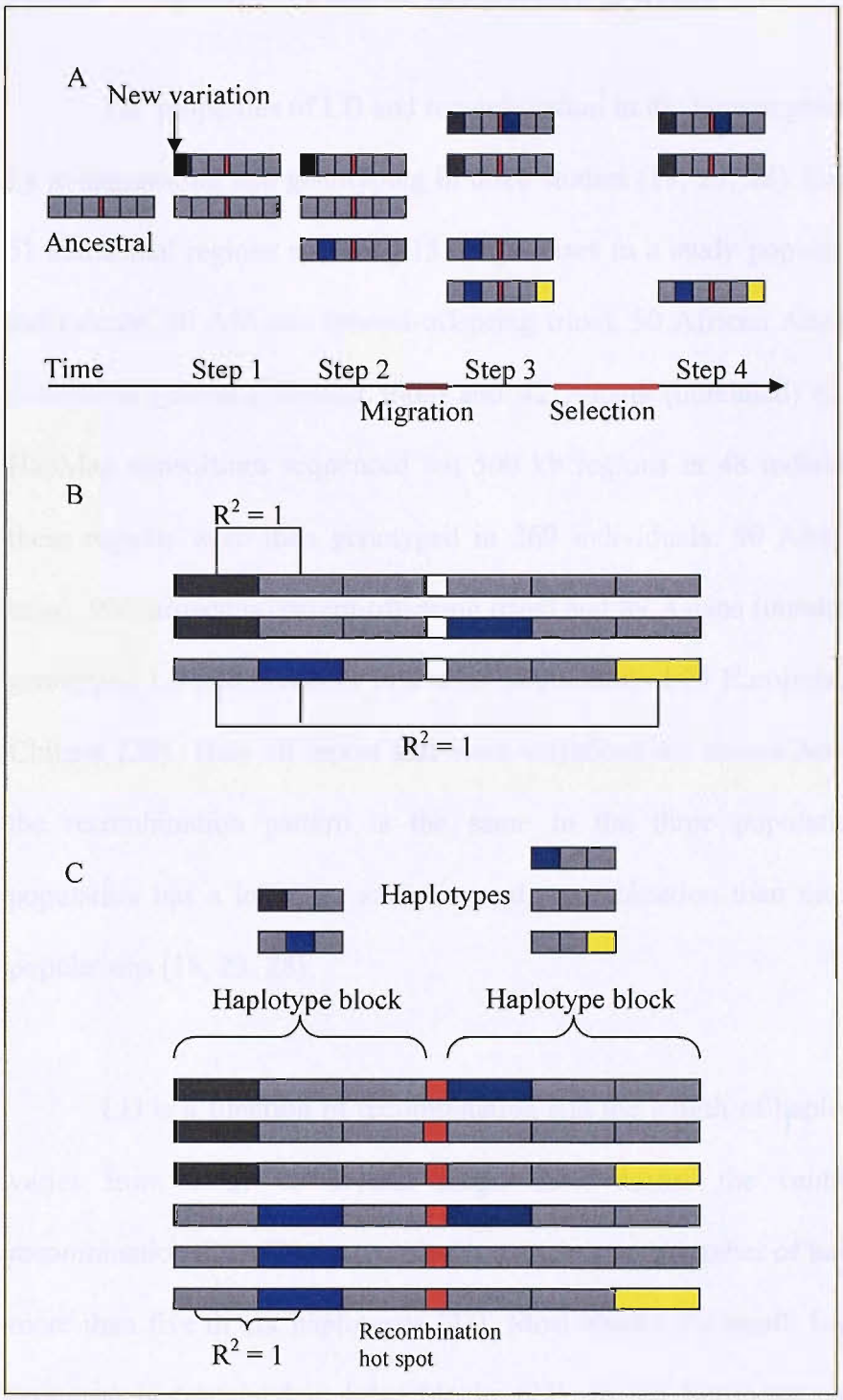
see section 1.1.5.3) for association studies (24), it has also been used to design DNA chips (illumina[3]) for genome-wide association studies (27).

1.1.5.1 Creation and loss of LD

The easiest way to explain how LD and haplotypes are created is to start with a new SNP on an ancestral background (Figure 1a). When a mutation occurs, it creates a new (second) haplotype at the same time (Figure 1a Step 1) (15, 16). If only one base is polymorphic, the concept of LD does not apply (15, 16). A second SNP creates a third haplotype and the two SNPs are in complete LD as long as there is no recombination between them (Figure 1a Step 2) (15, 16). If there were no recombination or other forces, affecting haplotype frequencies this process would create a forever-branching tree of haplotypes. However, genetic drift, migration and selection affect the frequency of haplotypes (Figure 1a Step 3 and 4) (15, 16, 18). Some haplotypes will increase in frequency whilst others will decrease or are lost from the population. This explains how some SNPs can be highly correlated (Figure 1b). Recombination breaks down this correlation, but the number of recombination sites are more dense in some regions of the genome (hot spots) than others (18). This creates a pattern of haplotype blocks, each with a low number of haplotypes, separated by recombination hot spots (figure 1c) (18). Inside a haplotype block SNPs are highly correlated, but there is less correlation between SNPs in different blocks (18). It is important to say that because recombination is not 100% long range correlations between SNPs can still exist (18).

[3] http://www.illumina.com/

Figure 1: Creation and loss of LD



The Figure is adapted from (15, 16, 18, 24)

1.1.5.2 Properties of LD in the human genome

The properties of LD and recombination in the human genome were investigated by re-sequencing and genotyping in three studies (18, 23, 28). Gabriel *et al.* sequenced 51 autosomal regions spanning 13 Mega bases in a study population consisting of 275 individuals: 90 Africans (parent-offspring trios), 50 African Americans (unrelated), 93 Europeans (parent-offspring trios) and 42 Asians (unrelated) (23). The International HapMap consortium sequenced ten 500 kb regions in 48 individuals; SNPs found in these regions were then genotyped in 269 individuals: 90 Africans (parent-offspring trios), 90 Europeans (parent-offspring trios) and 89 Asians (unrelated) (18). Hinds *et al.* genotyped 1.6 million SNPs in a study population of 24 Europeans, 23 Africans and 24 Chinese (28). They all report that most variations are shared between populations and the recombination pattern is the same in the three populations but the African population has a lot more variation and recombination than the European and Asian populations (18, 23, 28).

LD is a function of recombination and the length of haplotype blocks therefore varies from 1 kb to several mega bases across the centromeres, which lack recombination (18). Haplotype blocks contain a low number of haplotypes, normally no more than five or six haplotypes (23). Most blocks are small, less that 5 kb, but most sequence is contained in large blocks (23). In the European population, 50% of all sequence is in blocks larger than 44 Kb (23). This means that one can target 50% of common SNPs by genotyping 50,000 SNPs but needs to genotype 250,000 SNPs to cover 94% of common SNPs (18).

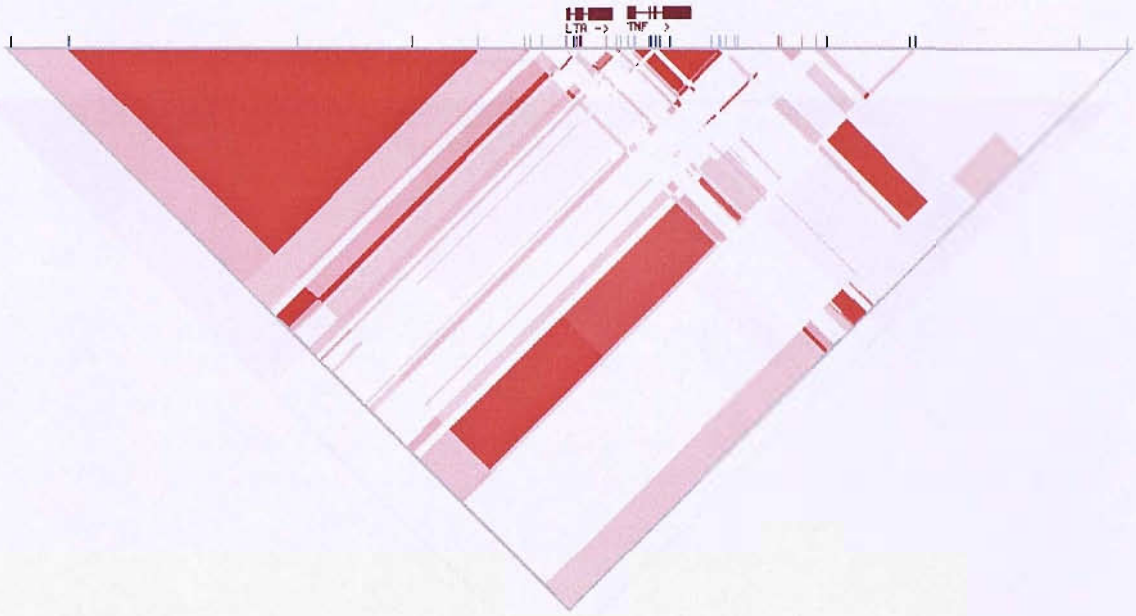
1.1.5.3 Algorithm for tagSNP selection.

To capture the most common variation in the genome, Carlson *et al.* suggests a new algorithm (29). Their first step in this algorithm is to calculate all pair wise r^2 values. R^2 is chosen because it is a direct measure for the information and an r^2 value of 1 would mean complete information whereas an r^2 value of zero would mean no information. The next step in the algorithm is to find the SNP that has a correlation over a certain threshold with most other SNPs. This SNP and all SNPs it is correlated with are removed from the data set and put in a bin and then this step is repeated for the remaining SNPs. In each bin, all r^2 values are recalculated and all SNPs, which have an r2 value over the threshold with all other SNPs in the bin, are tagSNPs for the bin. To test the algorithm one hundred genes were sequenced in a study population of 23 Europeans and 24 Africans (29). The number of tagSNPs it takes to cover a gene depends on how large the gene is and how much recombination there have been. *PON1* and *TRPV5* are the same length and have similar nucleotide diversity, but *PON1* needs 28 SNPs whereas *TRPV5* needs 9 SNPs in the African population (29). The tagSNP approach covered 85% of all common haplotypes in the European population (29). When the authors compared random SNP selection, haplotype SNP selection and tagSNP selection they found that random SNP selection covered 76% of the information from common SNPs, Haplotype SNP selection covered 86% whereas the tagSNP approach covered a 100% (29).

The algorithm described above was used to find tagSNP (30). It was found that 40% of all SNPs were in a bin with 10 or more SNPs and another 22% were in a bin with 5 to 9 SNPs (30). It was also showed that at an r^2 threshold of 0.8- 0.87% of all common SNP would be covered compared with complete resequencing. Based on these

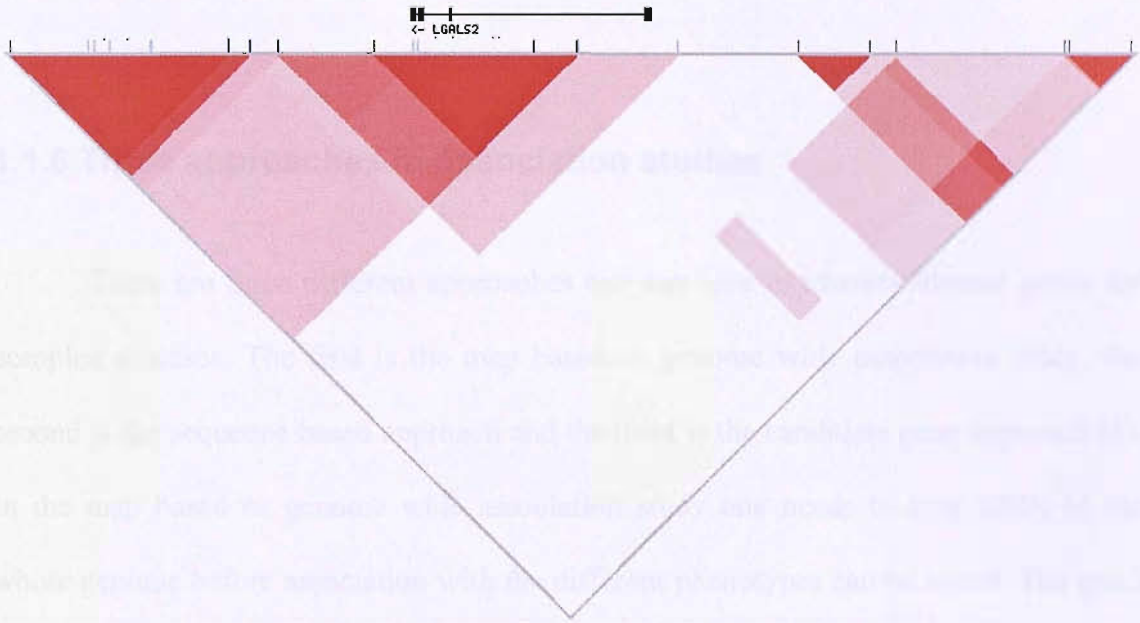
findings one can calculate that 300,000 SNPs is needed to cover the genome in a European association study and 500,000 in an African association study (30).

Figure 2: Haplotype block structure for the genome region around *LTA* and *TNF*



Schematic of the haplotype block structure for the genome region around *LTA* and *TNF*(26). The R^2 value between two SNPs is represented with colours. White indicates an R^2 value of 0. Increasing intensity of red indicate an increasing R^2 value. Red indicates an R^2 value of 1. *LTA* and *TNF* are superimposed on the schematics.

Figure 3: Haplotype block structure for the genome region around *LGALS2*



Schematic of the haplotype block structure for the genome region around *LGALS2* (26). The R^2 value between two SNPs is represented with colours. White indicates an R^2 value of 0. Increasing intensity of red indicate an increasing R^2 value. Red indicates an R^2 value of 1. *LGALS2* is superimposed on the schematics.

The difference between haplotypes and bins is that haplotypes cover a continues stretch of DNA whereas bins have the ability to cover long distance LD. If one whish to study the gene as a functional unit the haplotype approach is most logical, but the bin approach is more power full in statistical terms.

1.1.6 Three approaches to association studies

There are three different approaches one can take to identify disease genes for complex diseases. The first is the map based or genome wide association study, the second is the sequence based approach and the third is the candidate gene approach (4). In the map based or genome wide association study one needs to type SNPs in the whole genome before association with the different phenotypes can be tested. The good thing is that one does not need to assume anything about which genes or SNPs are involved, but one needs to type between 300,000 and 500,000 SNPs and have a large sample size to obtain significant signals(4, 30-32). In the sequence based approach one uses the knowledge that was gained from the study of monogenic diseases and only type the miss sense or nonsense mutation in conserved areas. This will reduce the number of SNPs drastically to between 50,000 and 100,000; this will also reduce the minimum sample size (4). The third approach is the candidate gene approach, where a candidate gene is suggested because it is in a pathway or a system that is known to influence the disease; this is the approach most researchers have followed until now. The third approach can be done either by defining haplotypes or by typing functional SNPs (4). I will briefly discuss the three approaches: The first one should in theory identify the

disease genes if one has the money and capacity to type enough SNPs, technology to type 300,000 SNPs is now available from Illumina[4] and Affymetrix[5]. It is also a problem to obtain phenotypes for a large number of people. The second approach is more economically viable, but its success depends on the assumption that the coding SNPs are the SNPs that cause complex diseases. The third approach relies on our ability to predict candidate genes upon existing knowledge. One could ask whether we have the knowledge that we need to possess and will we ever gain that knowledge about which pathways are important, without finding the candidate genes first.

1.1.7 Wellcome Trust Case Control Consortium (WTCCC)

The Wellcome Trust Case Control Consortium (WTCCC) has performed genome-wide association studies for seven diseases with 2000 cases for each disease (14,000 cases in all) and 3000 shared controls (32). Five hundred thousand SNPs were genotyped in each individual using a DNA chip from Affymetrix[4]. The seven diseases are bipolar disorder, coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), Type 1 diabetes (T1D), type 2 diabetes (T2D) and hypertension (HT) (32). All cases and controls were from the British population (Caucasian) (32). Strong association ($p< 5x10^{-7}$) was found for 24 independent loci as follows, CD 9, T1D 7, RA 3, T2D 3, CAD 1, bipolar disorder 1 and hypertension none (32). In addition, 58 moderate ($p< 5x10^{-6}$) associations were found (32). The one association found for CAD is located on chromosome 9p21.3 with SNPs across a 100 Kb region showing association (32). The strongest association was obtained with rs 1333049 (1.8×10^{-14}) and this association was replicated in a German population (32, 33). This region does not contain any genes

[4] http://www.illumina.com/
[5] www.affymetrix.com/

that have been associated with CAD before and the only three know genes are *CDKN2A*, *CDKN2B*, *MTAB* (32). Six moderate associations were found but again none that has been associated with CAD before. One of the associations was located on 22q12 but *LGALS2,* which has been associated with MI (34), is located on 22q13.1 a million bases down stream of 22q12 and therefore unlikely to represent the same association. *APOE* is one of the few genes, which association with MI has been confirmed by a meta-analysis (35). But a SNP in *APOE* (rs4420638) only showed a weak signal (32). Of the two cardiovascular risk factors: hypertension and T2D, only T2D showed strong associations (32). That no association was found with hypertension possibly because hypertension is more common in the population than other diseases and therefore present in many of the controls (32). This would mask a true association.T2D had four strong associations and nine moderate associations (32). One previous genome-wide association study (in a French population) had identified *TCF7L2* (rs7903146) as a strong genetic risk factor for T2D (27) this association was replicated in the WTCCC study (32). Two others loci identified in the French study *HHEX* and *IDE* had modest (10^{-3}>P>10^{-6}) association and three others *SLC30A8*, *LOC387761* and *EXT2* were not detected at all (27, 32). Two novel strong associations were detected in the WTCCC study one in the *FTO* gene (rs9939609) and another in the region around *CDKAL1(32)*. The association between *FTO* and T2D was replicated in 3757 cases and 5346 controls, but the association was not present after adjustment for BMI (31). The association with BMI was tested in 13 cohorts with 38,759 participants (31). The A allele that was associated with T2D was also associated with an increase in BMI (median 0.36, range 0.34 to 0.46 BMI units per allele) in all cohorts (31). The A allele is also strongly associated with the risk of being overweight (OR = 1.38; 95% CI

= 1.26 to 1.52; P = 4 x 10^{-11}) and obese (OR = 1.67; 95% CI = 1.47 to 1.89; P = 1 x 10^{-14}) (31). This association is present in all age groups (7 to 74 years) (31). Even though the association between rs9939609 and BMI is strong it is most likely not the causal variant. Rs9939609 is in a region with high LD (r^2> 0.5) spanning 47 kb including the first two introns as well as exon 2 (31). However, there is no obvious functional SNP and the function of FTO is not known (31). Besides *FTO,* 8 other genes which showed association with T2D in the WTCCC study, have been replicated in additional 14,586 cases and 17,968 controls (36). These genes are *CDKAL1*, *HHEX*, *CDKN2B*, *IGF2BP2*, *SLC30A8*, *TCF7L2*, and *KCNJ11*(36). What these genes all have in common is that they are in pathways that are connected to pancreatic cell function or development (36).

1.2 Hypothesis and Development

1.2.1 Aim

The aim of this study is to define haplotypes that predispose people to coronary heart disease and associated risk traits. If we know which genes influence a disease, it will help us to understand why the disease starts and how it develops. This can lead to several possible improvements in preventing and treatment of the disease. For example prediction of drug response through genotyping or identification of a subset of patients for a specific treatment.

1.2.2 Development

A number of true positive associations exist in the literature and can be found by systematic testing of known SNPs in large cohorts. To accomplish this, a comprehensive evaluation of the literature is necessary and this can be done by developing an automated method.

Chapter: 2 Development of a Bioinformatics strategy

2.1 Introduction

2.1.2 Text data mining

Text data mining algorithms extract information from free text. Free text is the kind of text obtained from Medline abstracts or any other written text, but it is not extraction from classified field as in Medline subject headings. Many areas use text data mining and it is not unique to science. But the technique used can not be directly applied to biomedical literature and it is therefore necessary to develop tools for biology(37, 38). Text data mining has changed from the recognition of gene and protein names to the recognition of interactions(39, 40). One of the simplest approaches is to use the co-occurrences of two protein or gene names in a text and interpret this as an association between the two(38, 41). The text data mining techniques can in general be divided into two groups unsupervised and supervised (39). Hidden Markov models (HMM) and Bayesian classifier (BC) are examples of unsupervised, whereas template and pattern matching is supervised (39). One can also mix the two methods. The unsupervised method treats the text as a bag of words, then it counts the frequency of each word and an HMM or BC is used to classify the text (39). A recent review recognises that these methods are not well suited to extracting information from text in biology(38).

Recently there has been a challenge cup to make an information extraction system to help the curation of the fly database[6] (38). Based on this, three papers have been published, which all propose different systems for curation of the fly base. One

[6] http://flybase.net

important thing to notice is that all papers have abandoned the unsupervised approach, where the text is treated as a bag of words and have moved into supervised systems. One paper presents a rule based approach(42). It uses pattern matching, natural language processing and sinimatic constraints. Domain experts decide the patterns for extracting. It also uses "part of speech tagger" to locate where the patterns should be found. To find the protein and gene names it uses a given lexical list. One of the other papers use a mixture of supervised selection of keywords and a classifier(43). The keywords are extracted from the abstract and the gene names are selected by domain experts and the distance between gene names and keywords are found. This is shown to a classic classifier to distinguish different types of papers. The last paper uses patterns to capture the association among words that appear in the documents(44). It uses an automated approach where it makes some general patterns and searches through the text. The next step is to sort patterns and the most common patterns that distinguish the different types of papers are given to a classifier.

2.1.3 Other applications for gene selection

The following is a description of other applications that aims to help researcher rank candidate genes for monogenic or complex diseases studies.

2.1.3.1 Summary of Iratxeta *et al.*

In a study with the title "Association of genes to genetically inherited diseases using data mining" (45) Iratxeta and colleagues have tried to associate diseases that have already been linked to a genome region with a gene in that region. The motivation for this is that out of the 4000 diseases that are thought be to genetically

inherit 1000 have been linked to the genome, but 450 have not yet been associated with a specific gene.

The method used is to combine information from Medline, RefSeq and the Gene ontology (G0). Medline records contain disease terms e.g. epilepsy and chemical terms e.g. GABA[7]. RefSeq is a collection of sequences that have been annotated with gene ontology terms. GO terms describe either a process or a function. The first step is to link MeSH disease terms (MeSH C) with MeSH chemical terms (MeSH D). Out of 10 million Medline records, one million have at least one MeSH C and one MeSH D terms. The algorithm uses Medline record to link MeSH C and MeSH D terms that co-occur. RefSeq have a set of papers linked to it that provides the experimental evidence for the annotation of each GO term. The next step is to link MeSH D terms, which are present in this set of papers, to that GO term. The link between MeSH C terms and MeSH D terms are based on the co-occurring in Medline records, and the link to GO terms are based on the presence of the MeSH D terms in the papers with the evidence that was used to annotate a sequence with a specific GO term. This allows for the association between MeSH C terms and GO terms. The authors use fuzzy relations to calculate the strength of the association. A simple way to explain them is to say the more often two terms co-occur the stronger is the association. Each disease now has a set of GO terms connected to it with a score for each GO term as a measure for the strength of the association. The RefSeq sequences (genes) can now be sorted according to the GO terms they have annotated. The RefSeq sequences (genes) with the most relevant GO annotations are then basted against the region where the

[7] Gamma aminobutyric acid

disease has been linked. The sequences (genes) that match that region are then again sorted based on the Gene ontology terms.

The authors have validated their method by using it to re-identify genes for 100 diseases for which genes are already known. The algorithm found the gene for 55 diseases. On average the correct gene was among the 3% best scoring genes. One possible fault of this approach is that much of the information in RefSeq comes from the study of disease genes. To confirm if the system can find new genes the authors used a version of medline\RefSeq from 2000 and tried to identify 27 genes that have only been linked to diseases after that. This time the algorithm found 7 genes outside the region where the diseases have been linked; another 10 diseases had no relevant GO annotation. For the remaining 10 diseases 5 had their gene in the 5% best scoring genes and 5 in the 15%.

The accuracy of the database annotation decides the accuracy of this method. Humans have annotated Medline, but not RefSeq. Even though Medline is annotated correctly, it does not mean that all co-occurrences of MeSH C and MeSH D terms are meaningful. A paper could be concerned with two different subjects. This will create nonsense associations between MeSH C and MeSH D terms. A more fundamental criticism is that the paper makes the underlying assumption that disease and genes is related because of function and not pathways (See section 2.6).

Summary of Freudenberg *et al.*

In a study with the title "A similarity-based method for genome-wide prediction of disease-relevant human genes" (46) J. Freudenberg and P. Propping aim to find and prioritise candidate genes for diseases with no genetic background, based on

their phenotypic similarity with diseases with known genetic background. Their approach is based on the following assumption "The algorithm starts from the assumption that phenotypic similar diseases are caused by similar molecular mechanisms".

Their method is to combine phenotypic data for diseases from the OMIM morbid map with the functional GO annotation for the genes that cause these diseases. The OMIM morbid map has the following information indexed: episodic, aetiology, tissue, onset and inheritance. As an example, colorectal cancer has the following terms in the index: no, neoplastic gastro-intestinal, late adult, autosomal dominant. The algorithm clusters diseases from OMIM with known genetic background together based on the similarity between their phenotypic terms. The algorithm clusters diseases together according to how many index terms they have in common. A threshold is used to decide how similar the diseases have to be to get in the same cluster; this threshold will change the size and number of clusters. The cluster algorithm is only semi-automatic and a human expert is needed both before and after to increase the quality of the clusters. OMIM consists mostly of monogenic diseases that are caused by a single mutation in a single gene. The GO annotation from that gene is considered to belong to the disease. A cluster can then be considered a group of diseases with two interfaces. One interface is all the phenotypic characteristics from OMIM and the other interface is all the GO annotation for the genes that are the genetic background for the diseases in the cluster. These two interfaces can now be used to couple diseases with no known genetic background to genes with GO annotation. The first step in the procedure is to associate a disease with one or more clusters based on phenotypic similarities, the similarity between a disease and a cluster must be above the same threshold that was

used for the clustering. The next step is to score genes with GO annotation to these clusters based on the similarities between the genes GO annotation and the clusters GO annotation. For each disease the system well produce a ranked list of candidate genes.

To test their approach the authors make a so-called leave-one-out test. The test leaves one disease and its gene out and then clusters all other diseases; this is repeated for each disease. The disease that was omitted is now considered a disease with no known genetic background. The system scores all genes with GO annotation, approximate 10000, to the relevant disease clusters. As mentioned above the cluster algorithm produces a different number and size of clusters depending on the threshold. The results will therefore depend on which threshold the algorithm used for similarity. For a so-called medium strength threshold the correct gene is found within the 321 (3%) best scoring genes for a third of the diseases and within the 1600 (15%) best scoring genes for two third of the diseases. To test the underlying assumption, stated in the beginning, the diseases are clustered in two different ways. One cluster is based on how phenotypically similar the diseases are, as in the approach just described. The other is based on how similar the GO annotation is. GO annotation has different levels, top-level annotation is broad and low level is specific. To get a better measure for how similar two genes are the level of their shared GO annotation is taken in to account. If the assumption is true, diseases that cluster together in the phenotypic cluster should also cluster together in the functional cluster. As explained above different thresholds for both the phenotypic and functional clustering will give a different outcome. To test independence between clusters the authors performed a chi-square test for a range of different set of thresholds. The authors performed 25 tests and obtained significant

results for five tests, the most significant test had p = 0.004. The authors do not make it clear if they the have corrected for multiple testing.

Summary of Turner *et al.*

Pocus is a protocol that aims to find candidate genes for diseases that have already been linked to at least two loci on the genome (47). It assumes that the disease genes in the two loci will share functional properties and therefore functional annotation.

It builds upon the effort by GO to annotate already known genes. If a disease has been linked to two or more regions on the genome the protocol will compare GO IDs for annotated genes in the different regions. Then each gene will have a score assigned depending on which and how many IDs it shares with genes in the other regions. Genes that have significantly more IDs in common compared to a simulation with random loci are considered candidate genes. This infers that if gene A in region 1 shares 10 IDs with gene B in region 2 and if 10 IDs are significantly more than two random genes shares in the simulation then gene A and B are candidate genes.

The authors test their protocol on 29 oligogenic diseases with known genes. Regions of 2 to 19 Mb around the genes were used as an artificial link regions; this corresponds to from 10 to 187 genes per region. GO ID in the artificial linkage region was used to identify the candidate genes. The outcome of the protocol is a shortlist of genes from the linkage region. In general the script make a shortlist from which it is 12 to 42 times more likely to find the candidate gene compare to the full list of genes from the linkage region. The authors also preformed a case study with two genes that were first discovered after the protocol was finished and they are not in the version of GO that was used. Therefore the protocol and GO annotation should not be

biased for these genes. The two genes are two autism genes NLG3 and NLG4. Their protocol successfully identified these two genes.

As with the previous papers this study also builds on the assumption that genes that underlie the same disease share some functional properties. As will discussed at a later stage this may be a problem. In addition, the study aims to find genes for complex diseases but because there are no genes yet, it uses oligogenic diseases this may also be a problem. They also assume that there is a candidate in each linkage region, but that is far from true and many linkage regions with no gene could give false positive.

Summary of Hu *et al.*

A paper in Journal of Proteome Research (48) is the most similar to my approach, but with some important differences. The paper aims to assist in the interpretation of proteome and DNA microarray data. It tries to predict all possible gene disease relationships.

Their strategy uses Medline subject headings to define diseases; their program goes thorough 29 diseases and all its sub-categories in an automated way. Their program has a string search algorithm that looks for names, symbols and alias in the abstract to identify genes. They get their names from locus link. They do not take context into account. This means that a gene name or symbol has to be unique not to create a lot of false positives, and if a gene name is not unique they have to delete it. The ranking of the strength between a gene and a phenotype is based on the number of abstracts where the disease and gene co-occur compared with the number of abstract where the disease or gene is alone. They use four different statistical methods to

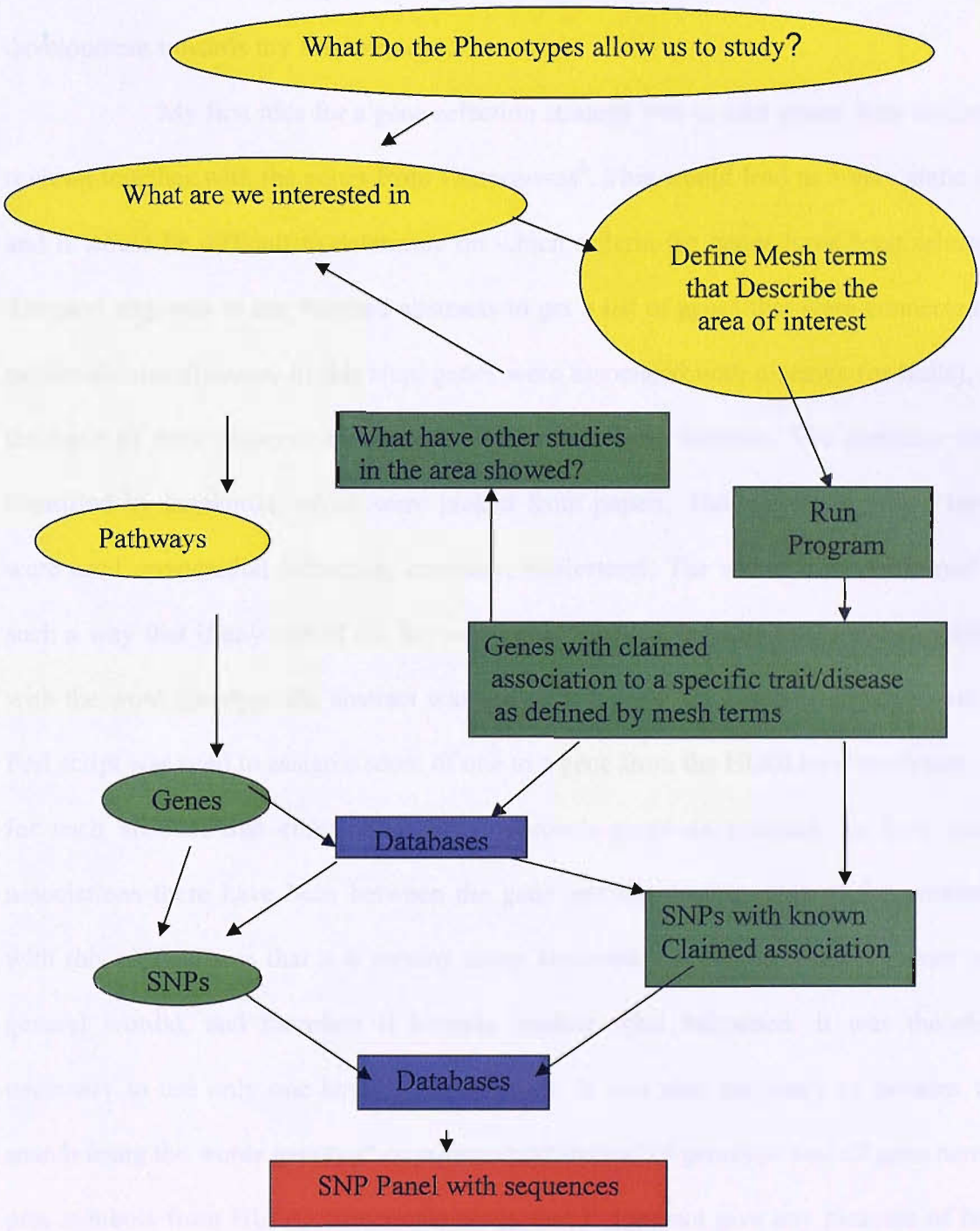
evaluate if the difference is statistically significant. The different relationships the authors have determined to be meaningful are the following. Clinical associations which they estimate are about 19% of the associations found; direct or indirect biochemical evidence which is 33%; histochemistry or differential expression 18.9%; genetic associations, mutant chromosome mapping 5.9%; shares high sequences homology to known disease genes 0.3%. The estimates are made upon a manual search in 600 abstracts, but their program does not classify the associations to any of these categories.

To evaluate their approach they have done three tests. First, they went through 600 abstracts manually and used that to estimate a false positive rate of 26%. To estimate a false negative rate they have compared the output from their program to known databases and their program did not predict 9% of the genes in the databases.

2.2 Aim of the program development

The aim of this development is to conduct a survey of the existing biomedical literature that will enable us to make a prioritized list of the gene disease associations that are the more likely to be true.

Figure 4: Flow diagram for the gene and SNP selection strategy.



2.3 Development

2.3.1 Preliminary method

The following is a brief description of the process I went through in the development towards my final strategy.

My first idea for a gene selection strategy was to take genes from different reviews together with the genes from Genecanvas[8]. This would lead to a very static list and it would be difficult to determine on which criteria the genes have been selected. The next step was to use Pubmed abstracts to get a list of genes that were connected to cardiovascular diseases. In this step, genes were associated with diseases (or traits), on the basis of their presence in abstracts concerning these diseases. The abstracts were identified by keywords, which were picked from papers. The following search terms were used: myocardial infarction, coronary, cholesterol. The search was performed in such a way that if any one of the keywords was found in the title or abstracts together with the word genotype the abstract was selected. Circa 2200 abstracts were chosen. A Perl script was used to assign a score of one to a gene from the HUGO nomenclature list for each abstract that contains it. This approach gives an estimate for how many associations there have been between the gene and the disease. One of the problems with this method was that it is mixing many keywords (risk traits, disease names and general words), and therefore it became unclear what happened. It was therefore necessary to use only one keyword at the time. It was also necessary to broaden the search using the words genotyp* or polymorph* instead of genotype and all gene names plus symbols from HUGO. The weakness is that it does not give any measure of how

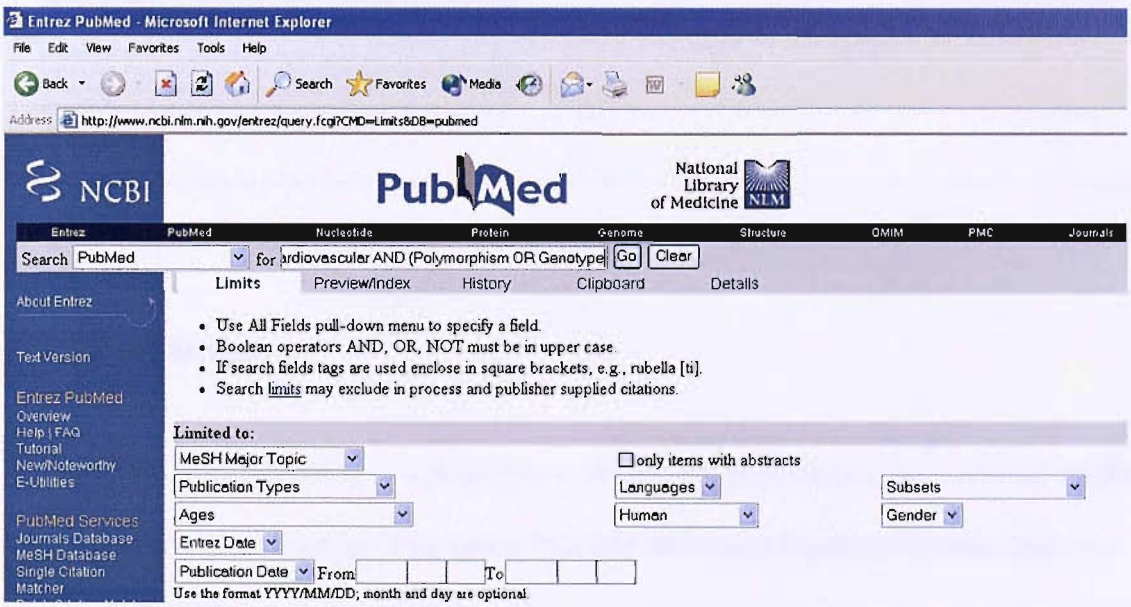
[8] http://genecanvas.idf.inserm.fr

strong the association is. This was solved by using the numbers of papers published per year as an estimate for the strength of the correlation between the gene and the trait.

2.3.2 Final method

The first step in the search strategy makes use of the Medline subject's headings (MeSH)[9]; this is equivalent to the first step in many meta-analyses and comprehensive reviews. I use the MeSH terms genotype and polymorphism to define a subset of abstracts relevant to associations studies. This subset is searched for abstracts that contain the MeSH term for the phenotype of interest and if necessary one or more text words for further specification (NCBI screen shot).

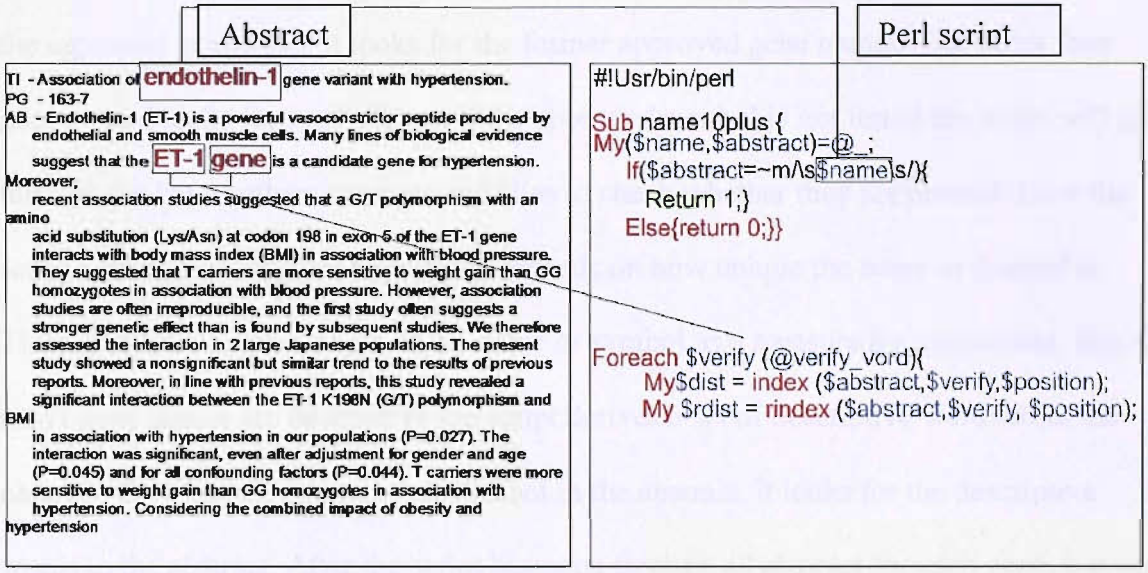
Figure 5: Screen shot from NCBI.



[9] http://www.nlm.nih.gov/mesh/meshhome.html

This set of abstracts is searched through by the Perl script I have made to record gene names (figure 6). The output is a text file with two parts. The first part has one line for each gene-abstract pair with a statement that tells if this pair was approved or not by the program. The Second part has one line per gene each containing information about the number of abstract gene pairs for the years 1992 onwards.

Figure 6: Illustrates how the Perl script looks for gene names in the abstract.



2.3.3 Perl script

The following is a description of how the Perl script I have written works (See figure 7). The script has four parts: Part one processes Medline abstract; part two processes the HUGO nomenclature gene list; part three is searching for gene names in the abstract; part four is an evaluation of the gene abstract connections that the program found. The first part processes Medline records in the following way. The script omits reviews, then saves publication ID, publication date, title and abstract for the rest of the

records. In the next step the script loads the HUGO nomenclature list and processes gene names and symbols. Gene names often contain notes in brackets that have to be removed before the gene name can be used in a string search. Symbols and aliases are together in two cells and have to be split out so a cell contains only one symbol or alias. Once the script has processed a gene name, it uses the obtained information to search through all abstracts. For each gene the script goes through all abstracts in the following way.

First the script searches the approved gene name and if the script does not find the approved gene name it looks for the former approved gene names. The script then goes on and looks for symbols and if the approved symbol is not found the script will go through the list of others symbols and alias to check whether they are present. How the script search, for names and symbols, depends on how unique the name or symbol is. The algorithm uses the length of the name or symbol as a measure for uniqueness. Since many gene names are descriptive the script derives a set of descriptive words from the name. If the script finds a name or symbol in the abstract, it looks for the descriptive words in the abstract. After the script has gone through all abstract for each gene, it goes through all the information that has been stored about each gene found in an abstract. It then passes the information through a series of filters to test if the gene name is real. The information contains:

Where in the abstract the gene name was found first, the context in which the gene name was found, if there was any descriptive words in the abstract that mach the descriptive words that came out of the gene name. Finally, the script looks for how many gene abstract pairs that were approved compared with the total number of abstract the gene name was found in. The thresholds in the filters, which decide if an abstract

should be approved or not, were found by running the script on a test set of abstracts and then examined the result for faults- then adjusting the parameters and running the script again. The process was repeated until the result was satisfactory (source code example 1 in appendix II).

Figure 7: Diagram illustrating working process of the Perl script.

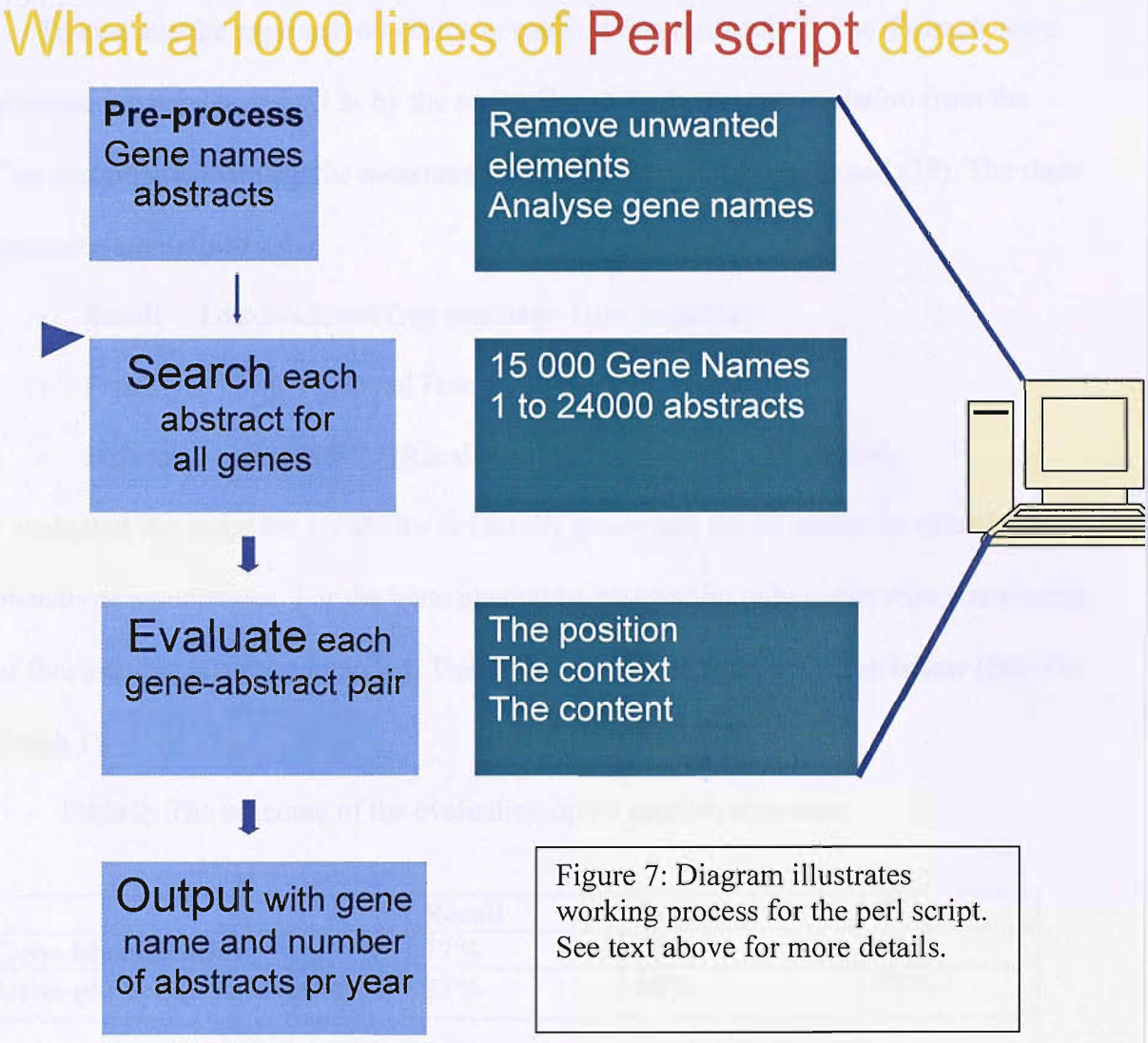


Figure 7: Diagram illustrates working process for the perl script. See text above for more details.

2.4 Evaluation

To evaluate the approach 60 abstracts were chosen randomly[10]. The abstracts were processed manually as well as by the script. Based on the recommendation from the Cup described previously the measures recall, Precision and F were used (39). The three measures are defined as:

- Recall = True Positive/(True positive+ false negative)
- Precision = True Positive/(True positive+ false positive)
- Balance F measure = 2 *(Recall * Precision)/(Recall + Precision)

I evaluated the script for its ability to identify genes and for its ability to identify gene-phenotype associations. For the gene phenotype association only genes with a minimum of five associations were included. The score for the two tasks are given below (See also Graph 1).

Table 2: The outcome of the evaluation on 60 random abstracts

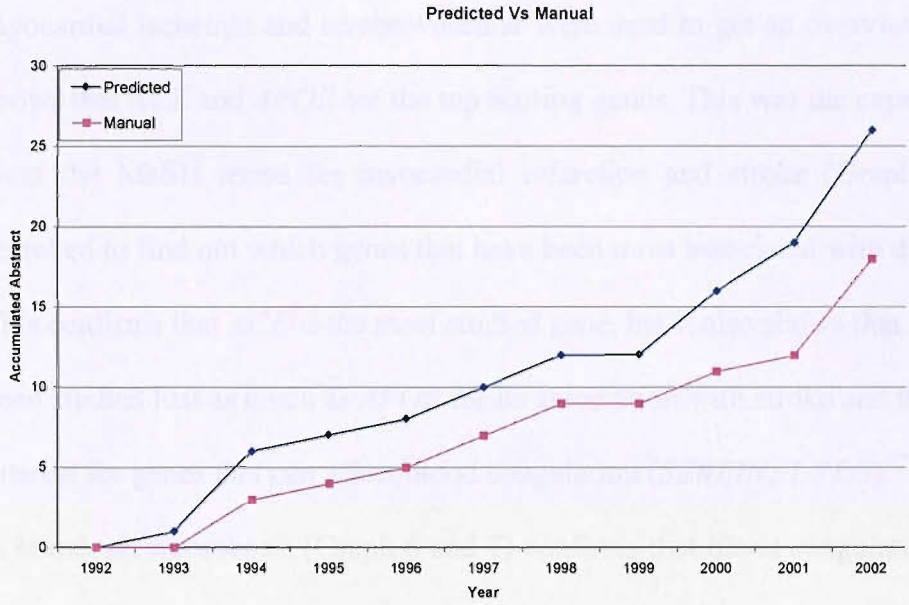
	Recall	**Precision**	**F**
Gene Identification	**77%**	**68%**	**72%**
Gene-phenotype association	**83%**	**68%**	**75%**

The F measure for the Fly base curation systems were between 67% and 84%. Another way to evaluate the approach is to benchmark it against a comprehensive manual review (5). Out of 19 cardiovascular[11] genes identified by Hirschorn the script identified 14. Three of the unidentified genes were *F2*, *F5*, and *F7*. These genes represent a category of gene names that the script has difficulty to find for two reasons. The first reason is that the name most often used is not in the HUGO nomenclature list.

[10] Random numbers were obtained from Random.org
[11] Because the definition of cardiovascular differ some genes had to be removed

The other reason is that these gene names (F2, F5, and F7) are very short and the script is therefore biased against them. Some genes are also overrepresented (see Graph1). This is because even though abstracts are reviewed not only genes of relevance to the disease is mentioned.

Graph 1: Comparison between automated and manual search.



The Graph compares the results of an automated and manual search on the *CETP* gene in Cholesterol. The result is in line with an F measure at 75%

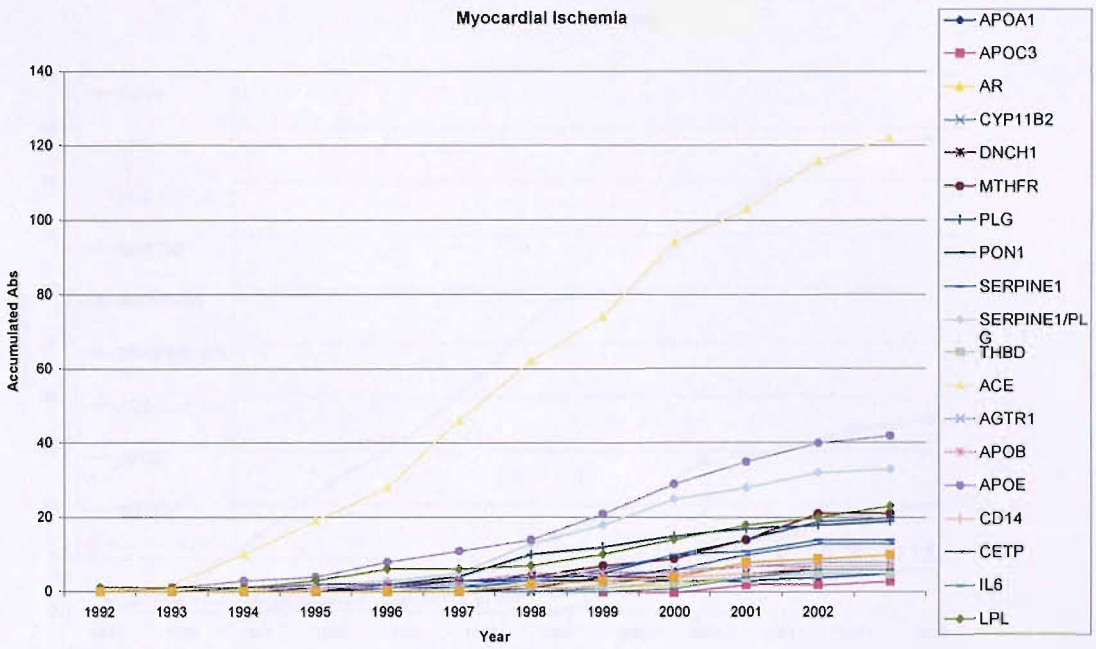
2.5 Results

A subset of MeSH cardiovascular diseases [C14] was searched using the method I have developed. A full list of MeSH terms is given in appendix I. The terms myocardial ischemia and cerebrovascular were used to get an overview. Graph 2 and 3 shows that *ACE* and *APOE* are the top scoring genes. This was the expected outcome. Next the MeSH terms for myocardial infarction and stroke (Graph 4 and 5) were searched to find out which genes that have been most associated with disease endpoints. This confirms that *ACE* is the most studied gene, but it also shows that *MTHFR* has been studied just as much as *APOE* for its association with stroke and there is still an interest for genes that can affect blood coagulation (*SERPINE1/PLG*)[12]
A search on thrombosis (Graph 6 and 7) confirms that blood coagulation genes still are studied and that *MTHFR* is studied in venous thrombosis.

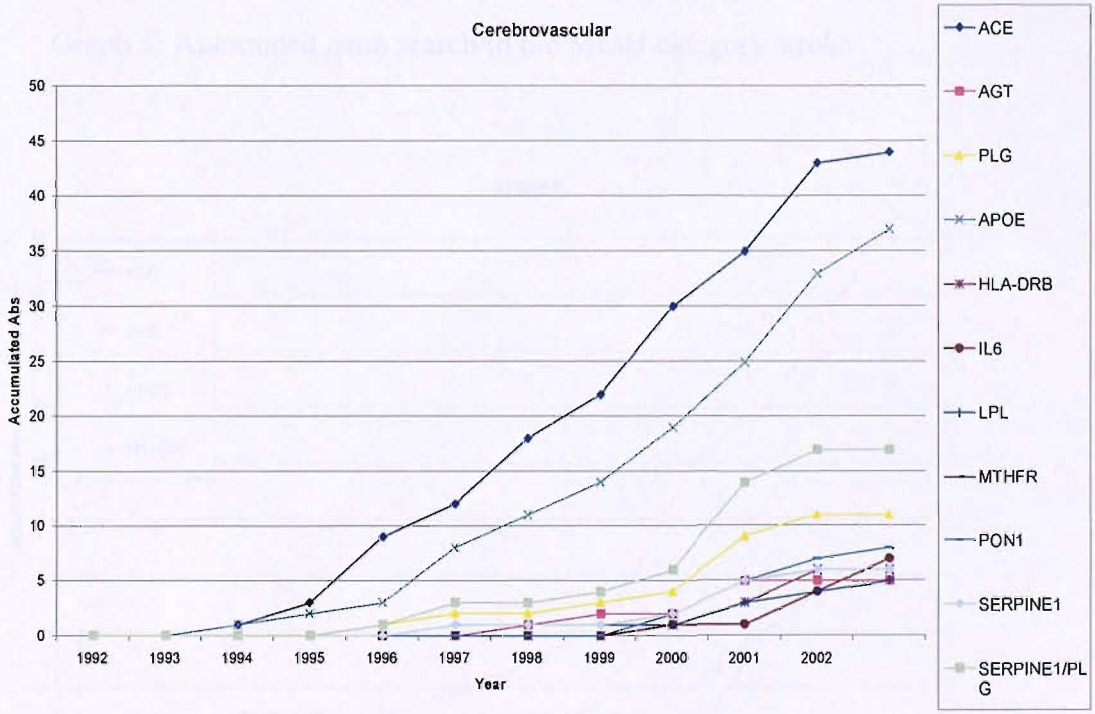
A look at specific risk factors indicates that *ACE*, *MTFHR* and *APOE* are tested for their associations with hypertension, hyperhomocysteinemia and hyperlipidemia respectively (Graph 8, 9 and 10). The search also shows that *CETP* and *LPL* are among the most studied genes in relation to cholesterol (Graph 11). Inflammation has emerged as an important player in atherosclerosis. To make a category that covers inflammation the MeSH term cardiovascular disease and the text word "inflam*" were combined. *TNF* and *IL6* are the most studied genes (Graph 12).

[12] The computer program often mistakes these two gene names for each other.

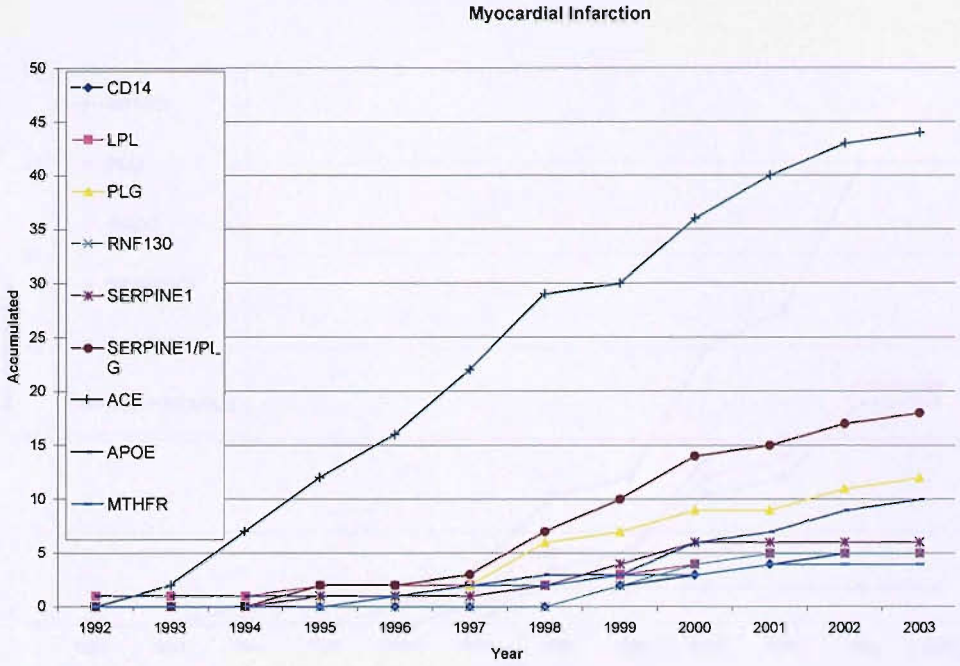
Graph 2: Automated gene search in the MesH category myocardial ischemia



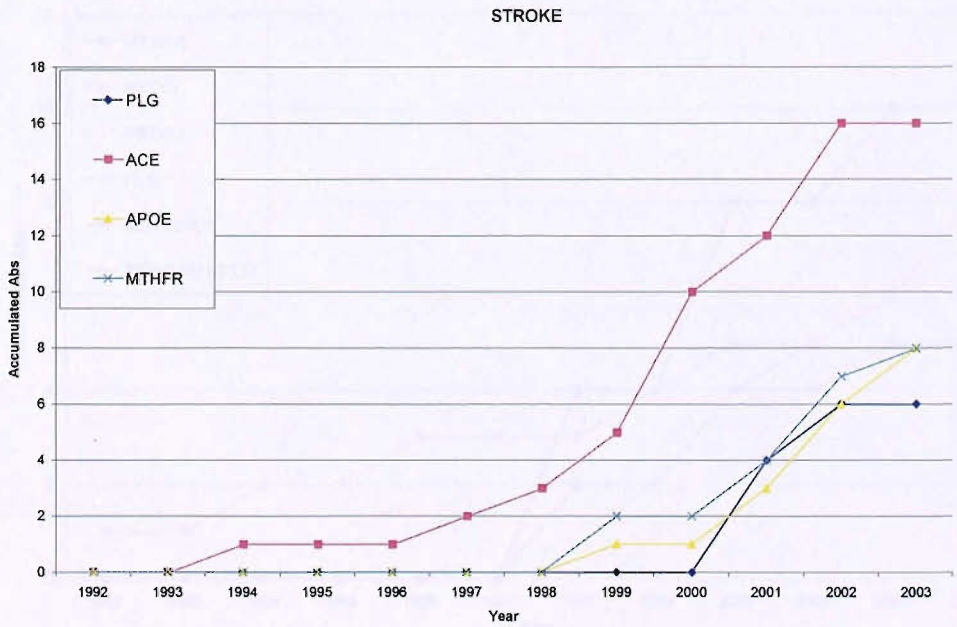
Graph 3: Automated gene search in the MesH category cerebrovascular



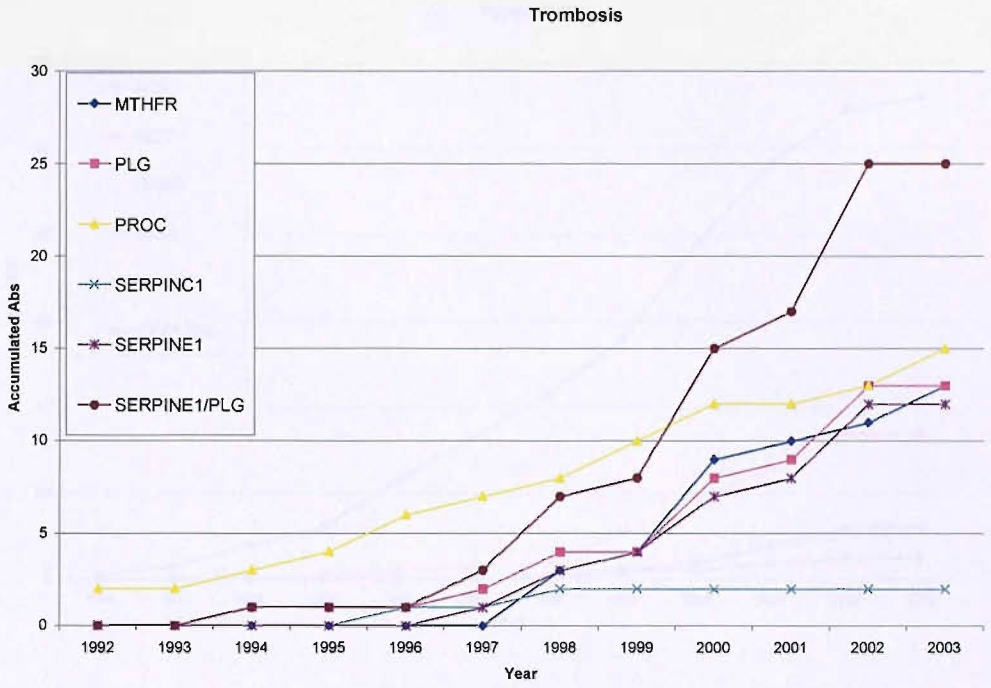
Graph:4 Automated gene search in the MesH category MI



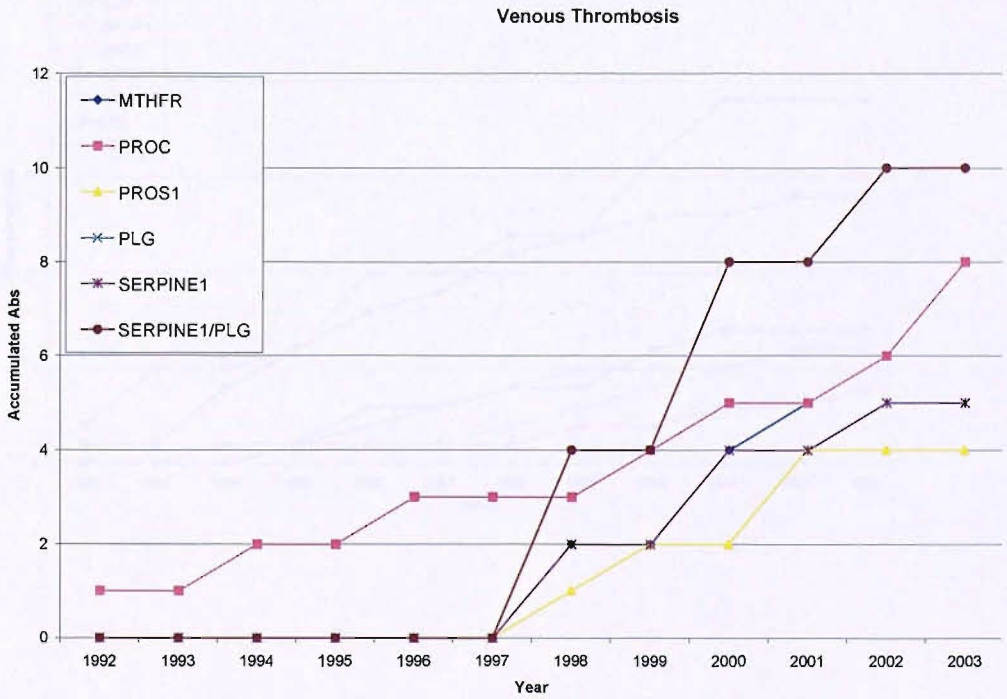
Graph 5: Automated gene search in the MesH category stroke



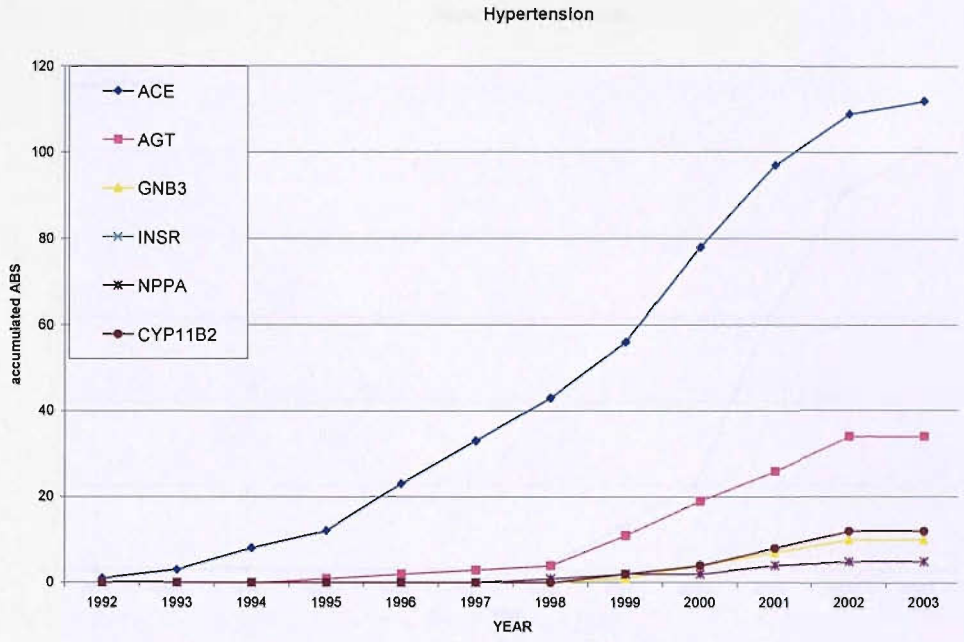
Graph 6: Automated gene search in the MesH category thrombosis



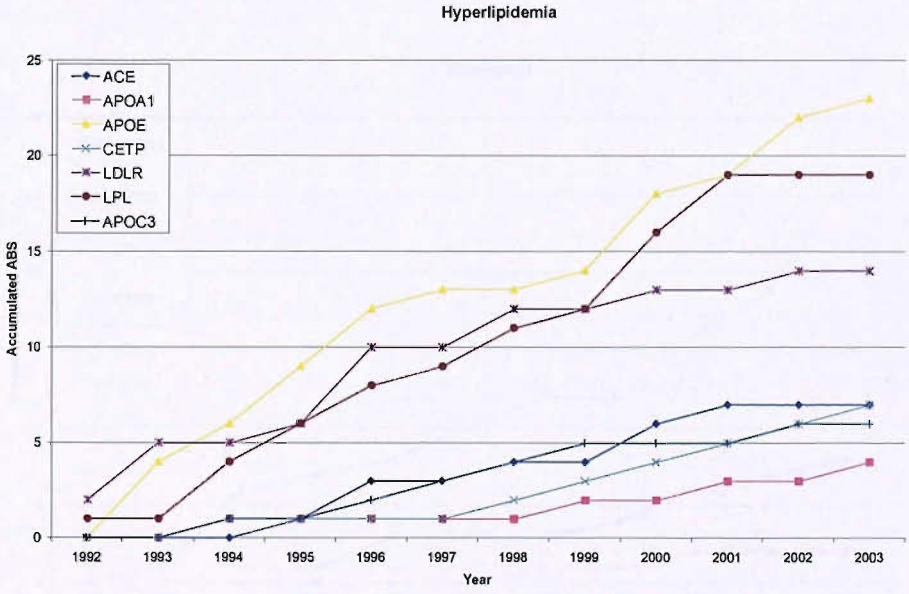
Graph 7: Automated gene search in the MesH category venous thrombosis.



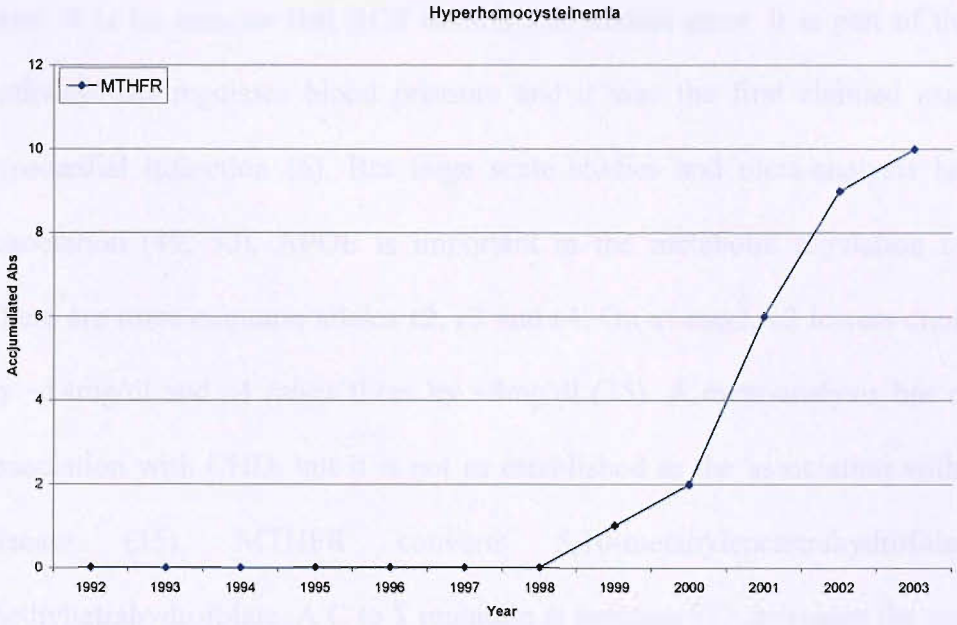
Graph 8: Automated gene search in the MesH category Hypertension.



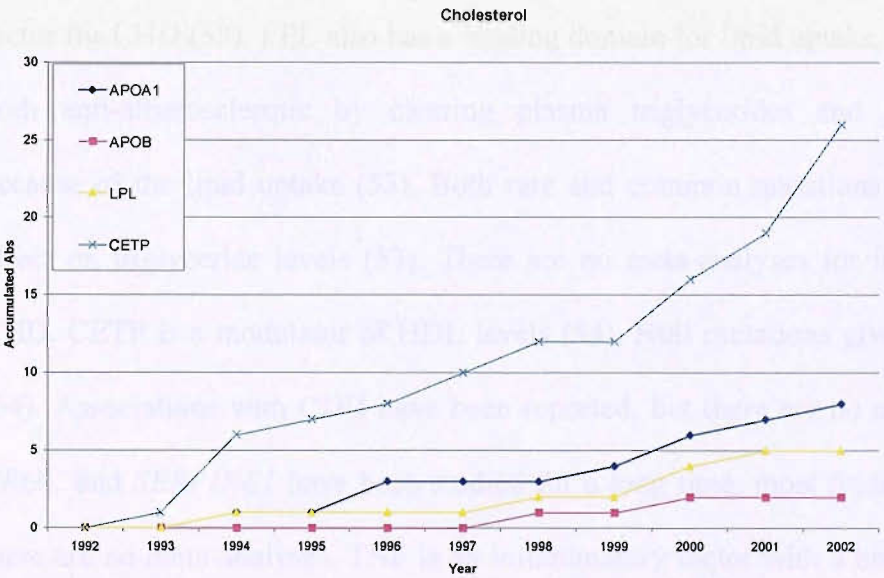
Graph 9: Automated gene search in the MesH category hyperlipidemia.



Graph 10: Automated gene search in the MesH category Hyperhomocysteinemia.



Graph 11: Automated gene search in the MesH category cholesterol.



The above is only a description of the search output. To decide on which genes are good candidate genes it, is necessary to go through the evidence for each gene. It is no surprise that *ACE* is the most studied gene. It is part of the angiotensin pathway that regulates blood pressure and it was the first claimed association with myocardial infarction (6). But large scale studies and meta-analysis have found no association (49, 50). APOE is important in the metabolic regulation of cholesterol. There are three common alleles ε2, ε3 and ε4. On average, ε2 lowers cholesterol levels by ~14mg/dl and ε4 raises them by ~8mg/dl (35). A meta-analysis has confirmed the association with CHD, but it is not as established as the association with Alzheimer's disease (35). MTHFR converts 5,10-metahylenetetrahydrofolate to 5-methyltetrahydrofolate. A C to T mutation at position 677 decreases the enzyme activity which leads to higher levels of homocysteine and lower folate levels (51). Two meta-analyses have confirmed the association between *MTHFR* and CHD (51, 52). LPL converts plasma triglycerides (TG) to free fatty acid (53). Triglycerides are a known risk factor for CHD (53). LPL also has a binding domain for lipid uptake. It can therefore be both anti-atherosclerotic by clearing plasma triglycerides and pro-atherosclerotic because of the lipid uptake (53). Both rare and common mutations exist that have an effect on triglyceride levels (53). There are no meta-analyses for its association with CHD. CETP is a modulator of HDL levels (54). Null mutations give high HDL levels (54). Associations with CDH have been reported, but there are no meta-analyses (54). *PROC* and *SERPINE1* have been studied for a long time, most findings are small, and there are no meta-analyses. TNF is an inflammatory factor with a broad spectrum. One of the first genome-wide association studies has found that *LTA (55)*, another inflammatory factor located near to *TNF*, is associated with CHD. In conclusion *ACE*,

APOE and *MTHFR* have all been extensively investigated and only very large studies would be able to bring new information. There are no convincing results for *PROC* or *SERPINE1*. This leaves *LPL*, *CETP* and *TNF* as the best candidate genes. *LPL* and *CETP* because they effect modulators (TG, HDL) of CHD risk. *TNF/LTA* because of the genome wide study that has identified *LTA* as a candidate for CHD.

2.6 Discussion

2.6.1 Principal findings

I have developed a literature search tool. It estimates the number of Medline publications per year, which reports the association or lack of association between a gene and a phenotype. It uses all available information in the HUGO nomenclature database to predict gene names in Medline abstracts. Medline subject headings are used to choose abstracts for association studies. I have used this tool to conduct a comprehensive search in cardiovascular literature and identified *CETP*, *LPL*, and *TNF/LTA* as the best candidate genes.

2.6.2 Strengths and weaknesses of study

This tool has been evaluated in three different ways, which all estimates an F measure of 75%. This is in line with other systems from dedicated bioinformatics groups. My tool uses all information in the HUGO nomenclature database and it can identify gene names correctly because it evaluates potential gene names based on the context where they were found. As I use a simple algorithm counting Medline abstracts the outcome is relevant biology, in contrast to other systems (see related literature). The number of false negatives for the whole of my method have not be calculated, this is because it would require a manual evaluation of Medline's 10 million abstracts. It is important to note that the false negative rate of the first step of my method is the same as the error rate of the human annotation done by Medline staff, which would be expected to be very low. This tool can only describe what is in the literature. It cannot

compensate for Medline abstracts lacking the necessary information to make a thorough analysis. Statistics such as "(OR: 4.29; CI 95%: 1.6-11.76)" can be extracted from the abstract; mostly from positive papers but their meaning cannot be interpreted precisely without reading the text. Many studies that have mainly negative results will often have vague or conflicting formulations. Abstracts often include a bit of review explaining the background for the study and it is difficult to distinguish this from the authors work.

2.6.3 Related literature

The three first papers reviewed in the introduction (45-47) have one thing in common; they all assume that the (biochemical) function of a protein and the phenotype of a disease, which the protein might cause if its gene is mutated, is related. If this assumption is invalid the foundation for their systems is greatly weakened. Their arguments for the function-phenotype relation are based on the following observation. There is a connection between the time of occurrence for a disease and the category of proteins that cause the disease; most diseases before birth are caused by transcription factors while after birth, enzymes cause most of the diseases(56). Freudenberg have also made a statistical test to prove the assumption. My argument against the function-phenotype relation is that two proteins that are similar in primary structure and biochemical function, can work on different substrates in different pathways. As an example, the *ADH* genes are all very similar but ADH5 works on a different substrate[13] than ADH3 (OMIM). If a mutation makes them dysfunctional, the resulting diseases will most likely have different phenotypes. It is also true that different genes cause the

[13] AHD5 S-hydroxymethylglutathione: ADH3 Ethanol.

same disease; *GNAT2* (Transduction alpha-subunits in G-protein), *CNGA2* and *CNGB3* (both cation channels) all cause achromatopsia (one symptom is total colour-blindness) (OMIM). As mentioned in the introduction (section 1.1.6), most researchers review pathways that are relevant to the disease to find new candidate genes not on function. Besides the two autism genes, there are no concrete examples in the papers to underpin their assumption. I think the authors have not established that the function-phenotype relation is general enough to make their tools useful in the mapping of monogenic disease genes, even less for complex diseases. The last paper (Hu *et al.*) is a literature search tool that detects all gene disease connections, but because gene disease associations are only 6% of all connections, the output is crude. It also excludes many gene names because they are similar to common abbreviations. This means that *LTA* is not used because it is similar to the abbreviation for lipoteichoic acid (LTA).

2.6.4 Important differences in results

Do meta-analyses support the idea that most published means best candidate gene? If one examines the three best scoring genes *ACE*, *APOE* and *MTHFR* the answer is conflicting both no and yes. *ACE* is by far the most published gene and the association is most likely not true but it is also an exception. It emerged first and therefore got more attention than otherwise. The next most published genes are *APOE* and *MTHFR* and their association with CHD looks to be true. Below top three the result becomes crude and no gene have had its association proven but even though an association is not true in itself, it can carry true information (this is because of LD). This could be the case for *TNF* that is in LD with *LTA*. Most studies have been on single

SNPs until now, but the nature of disease mutation could be much more complex. The many negative results can therefore not be used to exclude a gene.

2.7 Conclusion

The number of publications per year is valuable information for one to decide which the best candidate genes are. The utility of this tool is illustrated in the second part of this thesis, because without it the association between *TNF/LTA* and metabolic systems traits would not have been included in the a priori hypothesis.

Chapter 3: SNP retrieval

3.1 Introduction

The objective was to create a database that contains genes that are thought to be involved in coronary heart disease. For each of these genes the database contains SNPs and their position. The database also contains information about structural elements related to the gene (Base line annotation). The position will be used to get sequence from the chromosomes. (See flow diagram on page 20 for how this is implemented in the overall search strategy).

3.1.2 Data resources

The main SNP databases are dbSNP, HGVbase, and the genecanvas. DbSNP is an NBCI database that contains SNPs from all the other major projects and also has SNPs that have been submitted. DbSNP has at present about 10 million SNPs; of this 4 millions are validated and annotated. Of the annotated SNPs sixty-five thousand are non-synonymous and 46 thousand are synonymous (2004). (http://www.ncbi.nlm.nih.gov/SNPs/index.html). HGVbase (Human Genome variation Database) is maintained by the Karolinska Institute in collaboration with EBI and EMBL. All SNPs are subject to minimum validation before released. There are 1.7 million entries, of which thousands are experimentally validated. HGVbase and dbSNP have agreed to undertake bi-directional data exchange of core data. Annotation data is generated by different methods and is not exchanged. About 40% of dbSNP are

included in HGVbase (http://hgvbase.cgb.ki.se). HGVbase have previously collected data from other databases and papers but HGVbase is no longer maintained.

The canvas (http://genecanvas.idf.insern.fr) the French Institute of Health and Medical Research is a cardiovascular specific database. It contains several genes that are thought to be connected with common cardiovascular disorders (30 studies, 119 genes, 597 SNPs). The other databases are TSC (1.4M SNPs, http://SNPs.cshl.org, Nature: vol. 409, pp 928-933); HGBASE (60K SNPs, http://www.hgbase.interactive.de) and Human-SNPs-database (3K SNPs).

Even though there are 10 million SNPs in the public domain, there are only around 4 million validated SNPs. The problem is that many SNPs identified by sequencing are false positives. The SNP consortium, which until now has delivered 1.8M SNPs, only has 2-5 fold sequence depth (2000 SNPs have been validated), while the HapMap project have determined that eight fold sequence depth is necessary to get reliable results.

3.1.3 Structure of databases

The dbSNP flat file has a section for each reference SNP (unique SNPs) and each section have a number of lines with a defined content. If the SNP is located in a gene or in a distance of 2KB of a gene, the gene name is in the LOG line. Accession numbers to all gene bank files that contain sequence (locus sequence, gene sequence and golden path sequence) where the SNPs is present and the position of the SNPs in

each sequence is given in a number of SEQ lines. The LOG and SEQ lines will be used to locate SNPs in and around the genes. The Canvas also contains SNPs, haplotype frequencies and primer details (ASO and PCR). The data are not provided in a flat file but can only be downloaded as web pages. The SNPs from Canvas need to be mapped on the golden path chromosome sequence. The golden path sequence is available from http://www.genome.ucsc.edu/goldenPath/22dec2001/chromosomes/.

3.2 Methods

3.2.1 Perl script for SNPs retrieval

I have made two series of Perl script for SNPs retrieval- one for dbSNPs and another for genecanvas. The dbSNP retrieval system consists of two scripts. One that searches dbSNP for gene names and saves SNP reports that contain a chosen name. The other script extracts selected information. The data that has been extracted from dbSNP are: refSNP number, number of submits, chromosome position, alleles and annotation. Not all SNPs had the information about chromosome position; this is because the orientation of the contig where they are located is not known. This can be solved by getting flanking sequence and use the in silico mapping script (see canvas) to find the position. The Canvas retrieval system consists of three scripts. The first script extracts selected information from the manually downloaded web pages. Another script was made for checking the information. The SNPs were mapped on to the chromosomes by a third script (Source code example 2 in appendix II). The ASO and flanking primer sequences were used to map the SNPs. The script needs all three oligos to map a SNP. Out of approximate 500 SNPs there were ASO and primer information for 386 and 290 were successfully mapped. I expect that the 100 SNPs that were not mapped have

sequence differences between ASO/primer sequences and the chromosome sequence. This could be improved by allowing one primer not to match.

3.2.2 Annotation

The canvas and dbSNP contains some annotation but to get a more precise exon/intron structure and to get information about possible alternative splicing RefSeq was downloaded. A script was made to combine the annotation. The annotation map contains information about exon/intron structure, coding sequence, alleles and replacements SNPs. Sequence analysis makes it possible to obtain information about promoter region (pol II transcription start using neural networks NN[14]), intron splice site and ESE (exon splicing enhancer). The NN makes it possible to test the effect of the SNPs.

3.2 Results

I have built a flexible data retrieval structure, which can be used to obtain selected information from SNP databases. The selection can include all available information or combinations of information. For example, one can search for all coding SNPs with a minimum minor allele frequency. What makes this tool powerful is that it can integrate information from many SNPs; this gives a new set of possibilities. For example one can choose SNPs that do not have any adjacent SNPs, which is important for genotyping. It also possible to choose genes with a special SNP pattern.

[14]http://www.cbs.dtu.dk/

3.4 Discussion

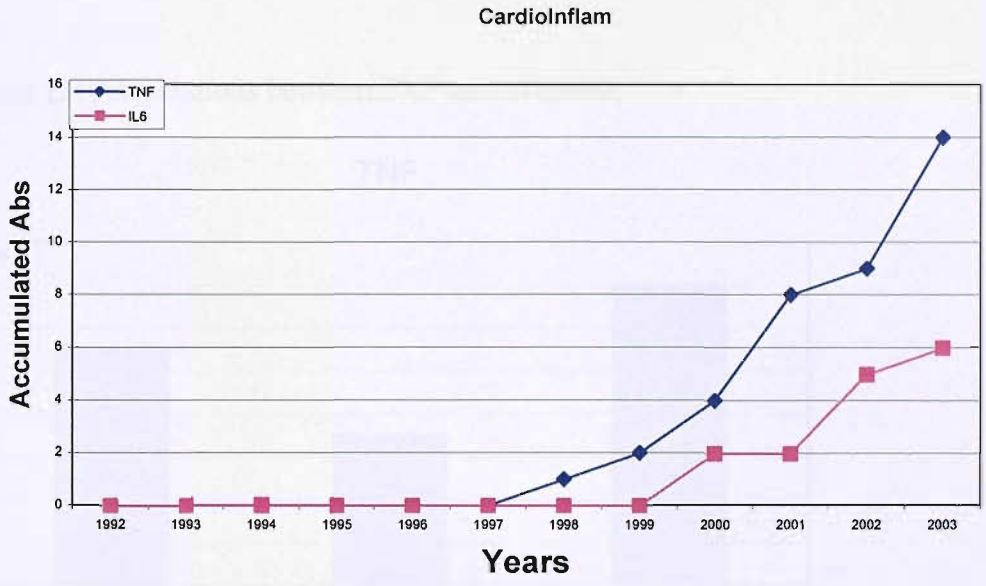
These tools enable large association studies that would otherwise not have been possible and they can be expanded to create new information. Some of the simple search functions have been made redundant by search tools (Biomart[15]) from large bioinformatic centres. Moreover, the ability to combine and integrate information is still only possible with local tools. Other groups have made similar tools but they have not been published. This is most likely because such tools need constant updates and because making a user interface takes too long time. My tool would be difficult to publish for the same reasons.

[15] http://www.biomart.org/

Chapter 4: Selection of genes and SNPs for investigation in BWHHS

The analysis of associations between genes and CHD or its risk factors, performed in chapter 2, covers the period 1992-2003 and showed that investigators were focusing on genes relating to the two established risk factors hypertension and cholesterol. However, the focus was changing in 2002 and there was a strong emphasis on inflammation in atherosclerosis (57, 58). A search in the constructed cardiovascular inflammation category showed that *TNF* and *IL6* were the most studied genes (Graph 12). In addition, the first genome-wide association study also identified *LTA* as a strong candidate gene for MI. *LTA* and *TNF* are located close to each other on Chr 6 (Graph 9). There are SNPs with functional claims in both *LTA* and *TNF* (Graph 14) and this added to the impression that *LTA* and *TNF* were attractive candidate genes. *LGALS2* was found to be associated with MI and its product galectin-2 showed to interacting with LTA; *LGALS2* was therefore included in the study.

Graph 12: Automated gene search in the constructed MesH category cardioInflam.

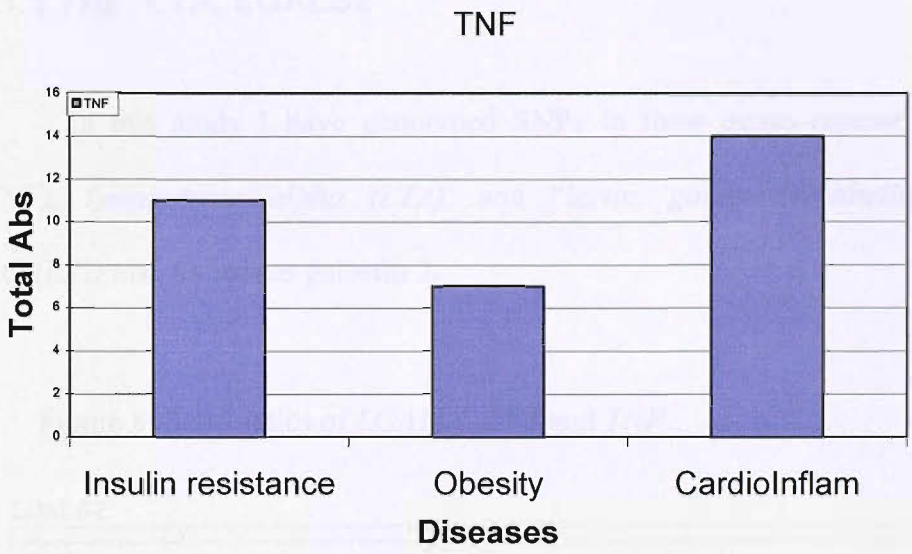


The SNP selection was determined by the availability of functional SNPs and LD. There are six known functional SNPs in *LTA* and *TNF* (Figure 14). Sequencing data available at the time suggested that it would take 15 SNPs to cover the region with a haplotype tagging approach (59). It was therefore decided to genotype functional SNPs but two of the functional SNPs also defines the three major haplotypes in *LTA* (Table 10). There is extensive LD in *LGALS2* and it was therefore only necessary to genotype one SNP to cover the gene (Figure 2).

To find all cardiovascular phenotypes *LTA* and *TNF* have been associated with, the table in appendix II was used as a reference. This identified an association between *TNF* and the phenotypes insulin resistance and obesity (Graph 13). Both component of

the metabolic syndrome, recognised as an emerging risk factor for CHD (60). LTA has been associated with insulin resistance (61); the paper showing this association was not in appendix II as only gene disease associations with at least five papers were included.

Graph 13: Associations between *TNF* and diseases.



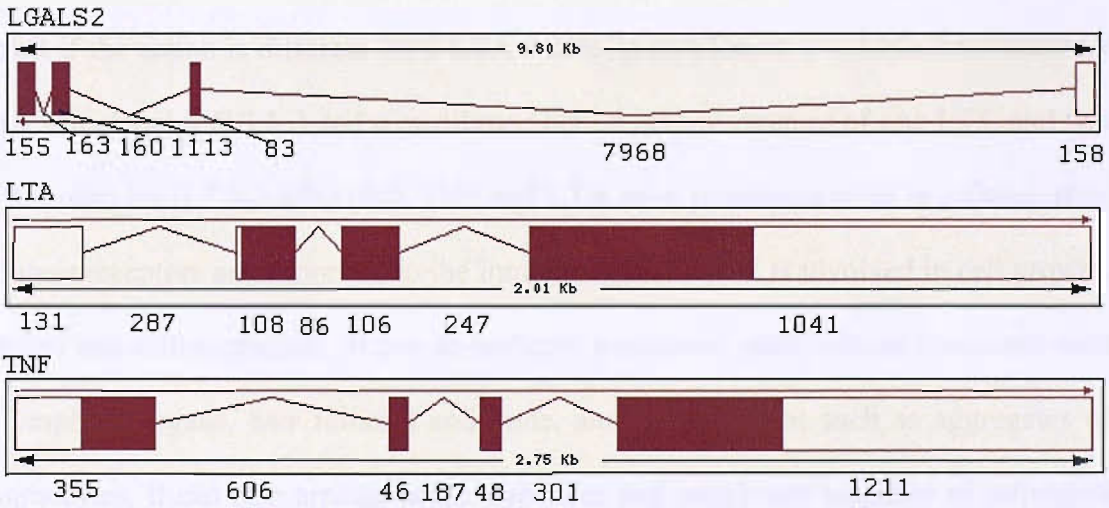
Chapter 5: *LTA, TNF, LGALS2* genotypes; metabolic and cardiovascular phenotypes.

5.1 Introduction

5.1.1 TNF, LTA, LGALS2

In this study I have genotyped SNPs in three genes *tumour necrosis factor (TNF), lymphotoxin alpha (LTA),* and *"lectin, galactoside-binding, soluble, 2" (LGALS2)* also known as galectin 2.

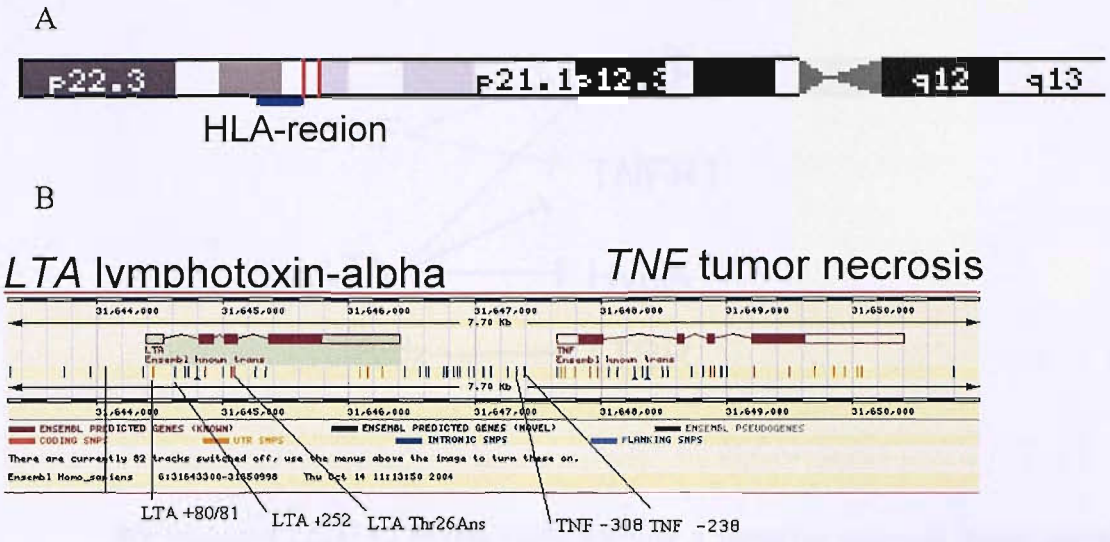
Figure 8: Schematics of *LGALS2*, *LTA* and *TNF.*



Schematics of *LGALS2*, *LTA* and *TNF.* Boxes represent exons and lines represent introns. Coloured areas of boxes are translated. Double arrows denote length of gene and single arrow denotes transcription direction. Each exon and intron has been annotated with length in bp.

TNF and *LTA* are located close together on chromosome 6 downstream of the HLA region (figure 9). TNF is by far the most studied with around 64,000 papers; there are only around 3000 LTA papers. TNF is the cytokine that is most widespread with homologies found in fish whilst LTA is only found in mammals (62). This indicates that TNF was present early in evolutionary history and explains how TNF can be important in many systems. TNF-like ligands have between 25 and 30% amino acid similarity and the all share the same threefold symmetric structure (63). Lymphocytes and macrophages are the main producers of TNF and LTA (62). Adipose tissue also produces TNF but not LTA; TNF as an adipokine will be covered in a separate section. TNF is produced as a precursor protein that is membrane bound. TACE is responsible for cleaving of the membrane bound TNF, and TACE may therefore be part of TNF regulation (63). TNF can act both as a membrane bound and soluble protein- it is not known if the action is different (63). LTA exists in two forms, a soluble homotetramer of LTA molecules (LTA_3) and a membrane bound hetero tetramer of one LTA and two LTB molecules (LTA_1LTB_2) (64). TNF and LTA have prominent roles in inflammation and their receptors are important to the immune system. TNF is involved in cell growth, survival and differentiation. It can co-ordinate permanent multicellular structures such as lymphoid organs, hair follicles and bone, and impermanent such as aggregates of lymphocytes. It can also arrange acute responses and coordinate response to pathogens (65).

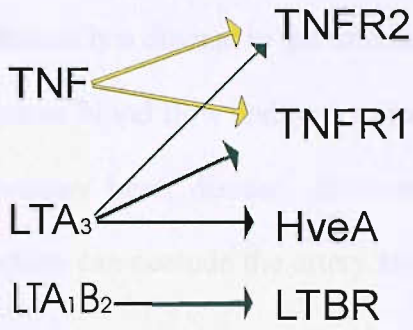
Figure 9: Annotation of the LTA and TN F region



A, Location of *LTA* and *TNF* (red box) down stream of the HLA-region (blue line) on chromosome 6. B, Ensembl output showing a 7.70 kb region of chromosome 6, which includes *LTA* and *TNF*. Vertical lines under the genes represent SNPs. The SNPs genotyped in this study have been annotated with names.

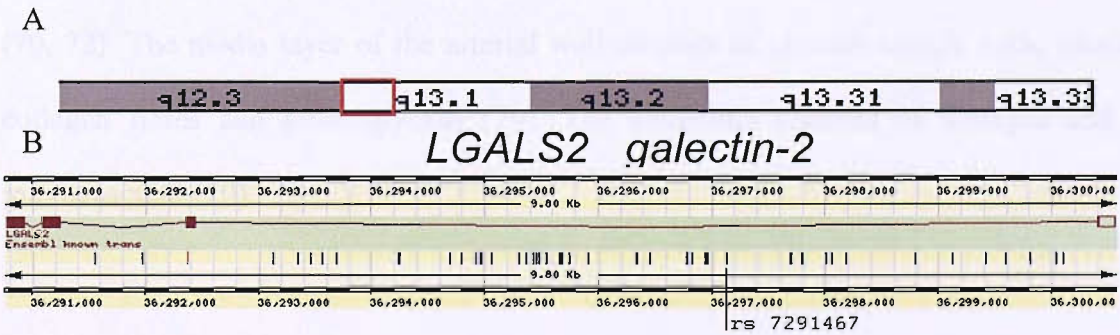
TNF and LTA bind to two receptors TNFR1 and TNFR2, which are found on nearly all cell surfaces (63). LTA can also bind to two extra receptors LTBR and HveA (Figure 10) (63). TNF receptors have two kinds of domains; DD domains, also known as death domains and TRAF binding motifs. The DD domain leads to cell death (apoptosis)(63). The TRAF domain has many roles; TRAF2 is for example negative feedback. TRAF works through the NFκB pathway but the precise way of activating NFκB is not known (66, 67).

Figure 10: TNF and LTA and their respective receptors



We included *LGALS2* in this study because a Japanese research group identified the gene product galectin 2 as a binding partner for LTA (34). Galectins are carbohydrate-binding proteins and there are 14 galectin proteins. Not much is known about LGALS2 in particular but galectins are involved in: regulation of inflammation, cell adhesion, cell growth and cell death (68) (69). *LGALS2* is located at chromosome 22 100 kb upstream of *LGALS1*.

Figure 11: Annotation of the *LGALS2* region.



A, Location of *LGALS2* (red box) on chromosome 22. B, Ensembl output showing a 9.80 kb region of chromosome 22, which include LGALS2. Vertical lines under the gene represent SNPs. The SNP genotyped in this study has been annotated with rs number.

5.1.2 Atherosclerosis

Atherosclerosis is a disease in the arterial wall, where the arterial lumen is narrowed. This can obstruct blood flow and lead to lacking oxygen supply to the tissue and possible cause coronary heart disease. Atherosclerosis can also predispose the vessel to thrombosis, which can occlude the artery and lead to myocardial infarction or stroke (70).

5.1.2.1 The arterial wall

The arterial wall consists of four layers; the endothelium, the intima, the media and the adventitia (70). The endothelium is a single layer of cells that forms a permeable barrier between the lumen and the underlying layers. It has many roles such as being non-thrombogenic and regulation of vascular tone (70, 71). The intima consists of smooth muscles cells and elastic fibres. The thickness of the intima in a healthy vessel is not well defined; sometimes the endothelium is directly on the internal elastic lamina (70, 72). The media layer of the arterial wall consists of smooth muscle cells, elastin, collagen fibres and proteoglycans (70). The adventitia consists of collagen and is microvascular (70).

5.1.2.2 The beginning of Atherosclerosis

The first sign of atherosclerosis is intimal thickening by lipid deposition. The first influx of lipid to the intima through the endothelium is probably due to general permeability and not specific transport (70, 73). This general permeability could be due to endothelium turn over among other things. The endothelium expresses vascular adhesion molecules that attach white blood cells (57). The white blood cells are then transported through the junction between the endothelium into the intima. When the monocytes come into the intima they are transformed into macrophages and they become lipid loaded (foam cells) (57, 73). These macrophages are thought to play an important role in the inflammation that goes on in atherosclerosis(74). In the early lesion, lipid is mostly found inside macrophage and smooth muscle cells.

5.1.2.3 The progression of Atherosclerosis

As atherosclerosis progresses it develops a yellowish fatty streak. The fatty streak consists of intracellular lipid, where the lipid is in smooth muscles cells and foam cells. As the disease progresses further, extra cellular lipid pools are seen. In advanced lesions there are cholesterol clefts with crystallised cholesterol, calcification and a cellular core. As the advanced lesion goes into its final stage a fibrous cap is seen (70).

5.1.2.4 Lipoprotein influx and efflux in the Atherosclerotic intima

Deposition of lipids in the arterial wall is one of the early signs of atherosclerosis. Influx of lipid starts in infancy and is present in the whole population at 10 years old. The first coronary lesion is noted from around 15 years old. The influx of lipid concentrates in regions where the intima has been thicken by connective tissue. There is strong evidence that the lipid influx comes from low density lipoprotein (LDL). The first build-up of lipid is extra cellular and cholesterol rich (75).

Experiments in hypercholesterolemic animals show pre-lesion transport of plasma lipid into the arterial wall. This confirms that lipid influx is the first sign of atherosclerosis (70, 75). Several lines of evidence support that the influx of lipid is passive transport, due to osmosis. Transvascular water can transport lipid and influx increases with evaluated transmural pressure. The size and serum concentration of the lipoprotein effects influx into the arterial wall. Dividing cells have a higher influx rate and increased permeability. This also explains why high cholesterol and hypertension are risk factors. It has also been shown that the concentration near the lumen wall depends on local flow. This could explain why there are some areas that are more prone to atherosclerosis than others (70, 75). Lipids accumulate less than one would expect based on the influx rate and this indicates that removal, exit and degradation in the arterial wall must occur (75). If LDL and lipid can be passively transported into the arterial wall it should also be able to be transported the other way and there should be no large build-up (75). Several mechanisms can hold LDL in the arterial wall. One is aggregation of LDL that of course changes the size and therefore hinders efflux to the lumen; another is that LDL binds to proteoglycans and enzymes such as LPL (75). The internal elastic

lamina also plays a role in the build-up of LDL by preventing removal of LDL to the outer media. This is specific for LDL and has nothing to do with size because HDL is removed (75). Because LDL is retained in the arterial wall one of the first hypotheses for atherosclerosis was response to retention.

5.1.2.5 Inflammation

Studies, mainly in animals, have made a strong case for the involvement of inflammations in atherosclerosis(74). The hypothesis states that inflammation is behind the beginning, progression and complication of atherosclerosis. What is known so far is that VCAM-1 binds monocytes similar to them found in lesions. In addition, MCP-1 is responsible for the migration of monocytes through the endothelium junction by binding to the CCR2 receptor. The evidence for VCAM-1, MCP-1 and CCR2 being involved in the pathology of atherosclerosis comes from experiments in genetic altered mice (74), where knock out of these genes lower lipid levels. It is the endothelium that secrets VCAM-1 and it is thought that it is inflammatory proteins such as NF-kB, IL-1B, TNF or LTA that induce the expression of VCAM-1 (57). These proteins have been found in human lesions. Inflammation also contributes to the transformation of monocytes to macrophages and the uptake of lipid to become foam cells. M-CSF has been identified as an important factor and M-CSF knock out mice have less lesion development (57). The classic view of plaque growths is that it is due to smooth muscle cell multiplication. But some studies suggest that the plaque grows in burst (57). The mechanism behind this is that small superficial erosion or disruption of the plaque leads to thrombosis that is incorporated into the plaque. This means that the plaque grows suddenly. Matrix metalloproteinases (MMPs) plays an important part in the degradation

of the fibrous cap and it is thought that TNF and possible other inflammatory factors can induce MMPs and thereby inflammation becomes a part of thrombosis (74). There are several theories for what it is that triggers the inflammation in atherosclerosis (76). One of the theories is the so-called oxidation theory where modified lipoproteins and oxidised phospholipids that triggers the inflammation but the theory lacks experimental support (77). It has also been proposed that hypertension through angiotension 2 can induce the expression of inflammatory factors such as IL-6. Diabetes (through AGE) and obesity (fat cells express TNF or IL-6) have also been suggested as triggers for inflammation (58, 78). Another theory is that inflammation is a natural occurring process that will start unless it is inhibited. One theory is the atheroprotective theory where endothelium that is under shear-stress will express protective genes that will inhibit inflammation. Areas that are prone to atherosclerosis normally have disturbed blood flow and will therefore not express the protective genes and inflammation starts (79).

5.1.3 Description of the metabolic system

The metabolic system consists of five main organs: liver, muscle, adipose tissue, brain and kidney (80). The liver, muscles and adipose tissue are the organs that are responsible for glucose homeostasis (80). Two major hormones insulin and glycagon regulates the energy metabolism. Insulin conducts the metabolic system when blood glucose levels are high, glycagon when blood glucose levels are low (80). Insulin is synthesised by the beta cells of the pancreas. Insulin promotes the uptake of glucose by the liver, muscles and adipose tissue and it makes the liver, muscles and adipose tissue

store energy (80). The liver absorbs glucose from the intestine and converts it into glucose-6-phosphate and then further into glycogen, it also produces triglycerides (80). The muscles store glucose and glycogen as fuel and the adipose tissue store energy in the form of triglyceride. Triglycerides are transported from the liver to the fat cells as very low density lipoprotein; the triglycerides are released and processed to free fatty acids by lipoprotein lipase (LPL) before uptake (80). When blood levels of glucose are low the liver can produce glucose from glycogen, glycol and lactate, and the adipose tissue releases fatty acid. The muscles can use the fatty acid instead of glucose as fuel (80).

5.1.4 The adipose tissue.

The adipose tissue secretes many hormones to the bloodstream: Leptin, TNF, IL-6, tissue factor, angiotensiogen, adipsin, ASP, adipophilin and INOS (81). Leptin (also known as the Ob the gene) is an endocrine hormone that signals to the brain the status of the fat store and regulates appetite and energy expenditure (81). The fat cells secret leptin in proportion to fat stores and nutrition flux into muscles and adipose tissue (81). Leptin secretion is increased with nutrition intake (81). Insulin also stimulates secretion of leptin from the adipose tissue (81). Angiotensiogen is a precursor of angiotensin II a regulator of blood pressure. INOS is another adipokine secreted by the adipose tissue that has been linked to blood pressure (81). It is possible leptin also has an affect on blood pressure (81). The correlation between the metabolic traits and high blood pressure could be explained by the secretion of adipokines, by the adipose tissue, to support the growing number of fat cells with blood supplies (81, 82).

5.1.5 TNF role in the adipose cell and insulin resistance

TNF has been shown to be directly responsible for insulin resistance (83). TNF is over expressed in adipose tissue from obese mice and rats (84). Blocking of TNF leads to increased insulin sensitivity in rats and long term exposure to TNF leads to less insulin sensitivity (83, 84). The lack of insulin sensitivity is restricted to glucose uptake and TNF do not affect the release of glucose from the liver (84). Mice knock out of TNF or the TNF receptors leads to less insulin plasma concentration on a high fat diet, but not on a standard diet, compared with wild type. Mice lacking TNF or its receptors also have higher insulin sensitivity than wild type mice. There was no difference in body weight between knock out or wild type on the same diet (85). Plasma concentration of triglycerides and FFA is also raised after TNF treatment (83). It has also been shown that weight loss in human leads to subsequent lower expression of TNF. It has long been known that free fatty acids can reduce glucose uptake and induce insulin resistance (82). AP2 is an adipose specific fatty acid binding protein. It transports fatty acids to the membrane of the organelles, which are responsible for triglyceride synthesis. Mice that lack aP2 have a phenotype that is similar to TNF knock outs (86).

It has been proposed that "the development of the metabolic syndrome is mediated by fat tissue mass"(87). However, experiments in mice have demonstrated that knock out of TNF, TNF receptors or aP2 protects from insulin resistance and chemically induced obesity does not lead to insulin resistance (85, 86). It is therefore possible that it is fatty acid metabolism in the adipose tissue that turns on TNF, which then induce insulin resistance. Body weight, weight gain and loss are associated with

insulin resistance and TNF expression but they are also linked to FFA metabolism. It can therefore be hypothesised that fat mass does not cause insulin resistance but is a marker for FFA metabolism. This view can also explain some epidemiology observations (see section about the metabolic syndrome).

TNF act in three ways, it modifies the insulin receptor, it induces changes in cellular gene expression and it promotes cell proliferation. The action is mostly locally in the adipose tissue. TNF inhibits insulin signalling by modifying insulin receptor substrate 1 so it interferes with insulin receptor tyrosine kinase activity (88). DNA microarray studies have shown that TNF down regulates *CEBPα* and *PPARγ*, *LPL* and *GLUT4*. It also up regulates a number of genes among them *TGFβ* and *VCAM1*(83, 89). CEBPα and PPARγ are transcription factors that regulate the expression of a broad range of genes (89). LPL is a membrane bound protein that hydrolyzes triglycerides in to free fatty acids before they are transported into the adipocyte (89). GLUT4 is a glucose transport protein that is responsible for the uptake of glucose in the adipocyte (89). TNF also stimulates cell proliferation possible by up regulation TGFβ (90). These complex relationships are illustrated in figure 11.

TNF and LTA signals to the NF-κB system through their receptors: TNFR1, TNFR2, and LTBR. The NF-κB system consists of five members and seven inhibitor proteins, together they make a flexible set of transcription factors that can be precisely regulated (67). The five members are: P65 (RalA), RelB, RelC, NF-κB1 (p50/p105 precursor) and NF-κB2 (p52/p100 precursor). They all have a common 300 amino acids motif towards the N terminal called RHD (67). RHD is responsible for the formation of homo or hetero dimmers; it is also involved in binding of inhibitor protein. There are seven inhibitor proteins (IκB): IκBa; IκBb; BCL-3; IκBe; IκBy and precursor's p105

and p100 (67). NF-κB is located in the cytosol as long as an inhibitor protein is bound to the nuclear translocation signal. When the IκB is degraded the NF-κB is transported into the nucleus (67). There are two NF-κB pathways- a classic and an alternative. The alternative pathway activates p100, which is special because it does not have a nucleus translocation domain. P100 can therefore repress signalling through the NF- κB pathway (67). The most important issue in this context is probably the LTA can bind to the LTB receptor, which can activate the alternative pathway and NFκB2 (66, 91). This nuclear factor can silence the gene expression that was turned on by the other TNF receptors, which signals to the classic pathway.

Figure 12: TNF and insulin resistance.

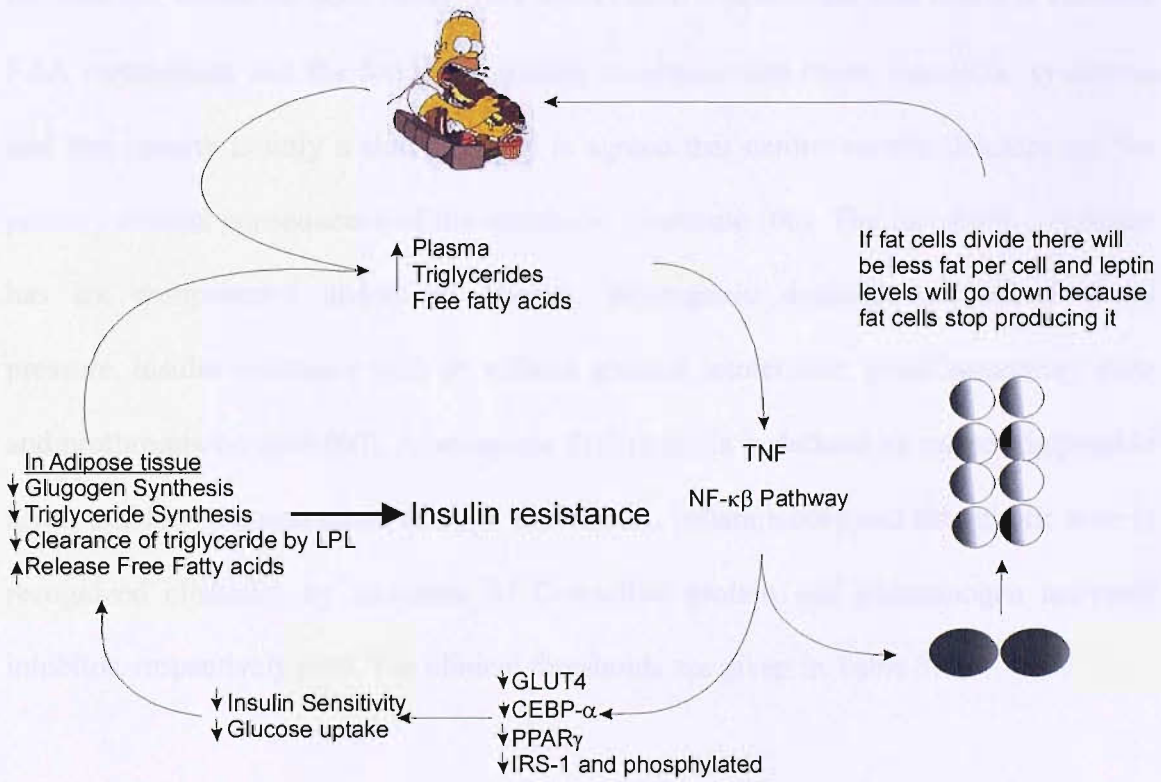


Illustration of how TNF influences gene regulation and metabolic processes in the adipose tissue, which leads to insulin resistance.

5.1.6 The metabolic syndrome.

The metabolic syndrome was first recognised as the clustering of metabolic diseases. The metabolic syndrome has three potential aetiological categories: obesity, insulin resistance and independent factors (60). There is still a debate on which of the factors are the most important. One argument is that obesity is most important and that the metabolic syndrome is a set of complications caused by obesity, others think insulin resistance is more important. Because obesity and insulin resistance is seen hand-in-hand it is difficult to see the contribution each of them make (60). Insulin resistance can be observed within all BMI strata. This observation supports the idea that it is elevated FAA metabolism and the following insulin resistance that cause metabolic syndrome and that obesity is only a side effect. It is agreed that cardiovascular diseases are the primary clinical consequence of the metabolic syndrome (60). The metabolic syndrome has six components: abdominal obesity, atherogenic dyslipidemia, raised blood pressure, insulin resistance with or without glucose intolerance, proinflammatory state and prothrombotic state (60). Atherogenic dyslipidemia is defined by raised triglyceride levels and low concentrations of HDL cholesterol. Inflammatory and thrombotic state is recognized clinically by elevation of C-reactive protein and plasminogen activator inhibitor, respectively (60). The clinical thresholds are given in Table 3.

Figure 13: Illustration of the main components of the metabolic syndrome.

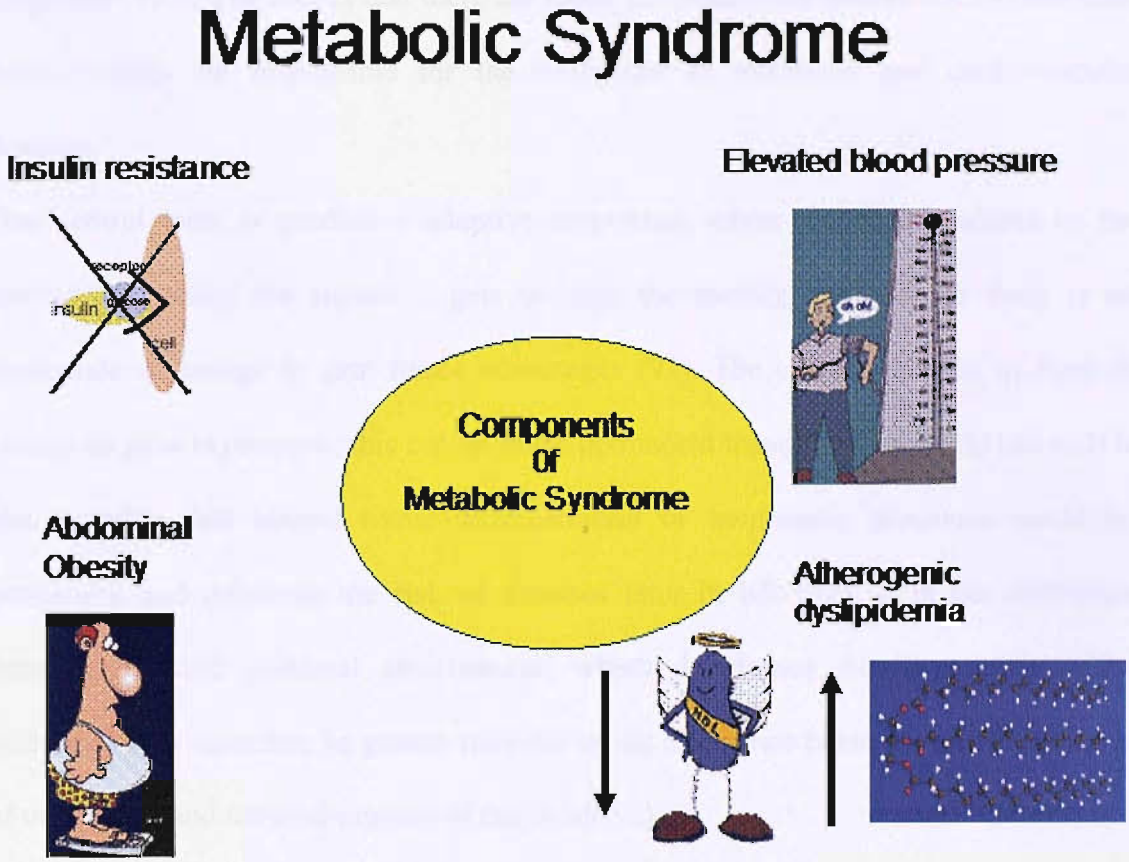


Table 3: Metabolic syndrome clinical thresholds

Abdominal obesity	waist circumference M >102cm F > 88cm	Waist: hip ratio M > 0.9 F >0.85 BMI > 30
Atherogenic dyslipidemia	Triglycerides >= 150mg/dl HDL cholesterol M < 40mg/dl F <50 mg/dl	HDL cholesterol M < 35mg/dl F <39 mg/dl
Raised blood pressure	Systolic >= 130/diastolic >= 85 mmHG	Systolic >= 140/diastolic >= 90 mmHG
Insulin resistance	Fasting glucose >= 110mg/dl	
Proinflammaory state		
Prothombotic state		

There are different ways one can explain why some people are more predisposed to the metabolic syndrome than others. This thesis is focused on genetics but there is a non genetic explanation which is the so-called "developmental origins of the metabolic syndrome" (92). The idea is that there are foetal environmental factors such as nutrition levels, which are responsible for the difference in metabolic and cardiovascular diseases.

One central term is predictive adaptive responses, where the foetus adapts to the environment using the signals it gets through the mother even though there is no immediate advantage to gain future advantages (92). The changes can be in from of change in gene expression; this can be made permanent through DNA methylation. It is also possible that altered tissue differentiation or homostatic processes could be permanent and influence the risk of diseases later in life (92). It is the difference between pre and postnatal environment, which determines the disease risk. The difference may therefore be greater than the actual difference between the environment of the mother and the environment of the child (92).

5.1.7 LTA associations

Four studies have tested the association between the two *LTA* SNPs +252 and T26N, which are in complete LD, and myocardial infarction: three Japanese case-control studies and one European trio family study (55, 93-95). All studies used similar criteria for myocardial infarction diagnostics. Two out of the following three criteria

had to be present: I central chest pain; II electrocardiographic changes; III elevated levels of creatine kinase.

The first study to identify *LTA* as a risk gene for myocardial infarction used a step wise approach (55). They genotyped 65671 SNPs, which were randomly chosen from approximately 14000 genes (exons 12000, introns 44000, flanking regions 2500), in 94 MI individuals. The frequencies of the SNPs were compared with the frequencies in a sample 658 individuals from the general population. SNPs with a P value less than 0.01 were genotyped in 656 MI individuals and one SNP in the *LTA* region showed positive association with myocardial infarction (55). 132 kilo bases of the *LTA* region were sequenced; this identified 187 SNPs of which 120 were genotyped in 94 MI individuals and 94 controls. One haplotype block consisting of 26 SNPs was defined. These 26 SNPs were genotyped in 1133 myocardial infarction individuals and 1006 controls. Five SNPs and the haplotype they define were associated with myocardial infarction (55). The strongest association was with *LTA* T26N in a recessive genetic model. The chi-square value was 20.1 with a p value of 7.3×10^{-6} and the odds ratio was 1.79 (95% CI: 1.38-2.31) (55). The SNPs associated with MI were not associated with the risk factors hypertension, hyperlipidemia or diabetes. Stratification was investigated in the 94 individuals that had 65000 SNPs genotyped and no evidence was found. Functional studies with a vector construct (-307 to +268 of the *LTA* gene), showed that the haplotype bearing the +252 C allele[16] had 1.5 fold higher expression than the T allele (55). This haplotype may also be a binding site for an unknown nuclear binding protein. The T26N Asn allele induces VCAM1 mRNA expression more than the Thr allele in smooth muscle cells (55).

[16] Most literature reports this SNP as A to G, which corresponds to the anti-sense strand. In this thesis all SNPs are reported on the sense strand.

A second Japanese study tried to replicate the association between *LTA* and myocardial infarction (93). They used a large and well described study sample consisting of 1493 men and 398 women with MI and a control population of 993 men and 805 women. The phenotypes include BMI, smoking, hypertension, diabetes mellitus, hypercholesterolemia, hyperuricemia. They performed logistic regression analysis to adjust for the risk factors. Men and women were analysed separately. They could not confirm the association between *LTA* +252 or T26N and myocardial infarction no matter which genetic model they used. They did however find an association with diabetes in men (93). A European trio family study has tested the association between *LTA* T26N and myocardial infarction (94). The criterion for inclusion was: both parent or one parent and at least one sibling were available. They had 303 informative transmissions. They observed an excess transmission of the T26N Asn allele to offspring with MI. They estimated an odds ratio of 1.7, which is similar to that of the first Japanese study (94).

A third Japanese study, including 400 men, confirmed the observed association between *LTA* and MI (95).

LTA has also been associated with traits of the metabolic syndrome (61). In a Japanese cohort of 200 men *LTA* +252 C was associated with lower HOMA score; the HOMA score for the genotypes was CC 1.28, CT 1.54, TT 1.57 (61). A recent European study found a recessive association between *LTA* +252 C and an increase of 1 in WHR in a cohort of 6000, but no association with any other of the metabolic syndrome traits (96).

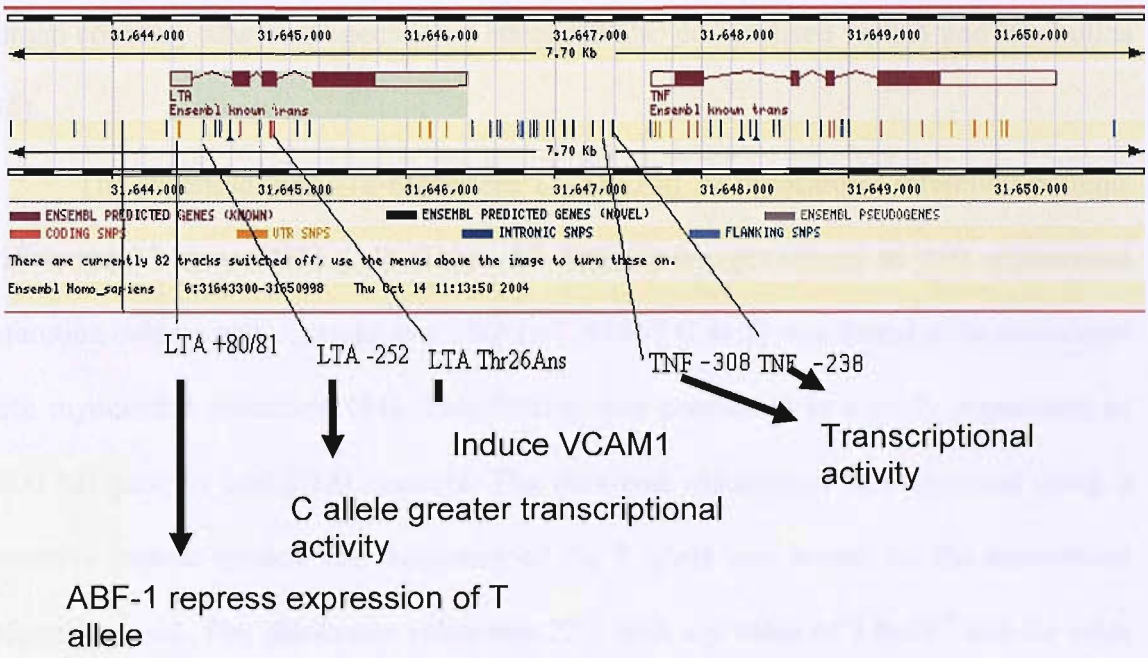
5.1.8 TNF associations and expression levels

TNF is responsible for insulin resistance and the two polymorphisms in the *TNF* promoter *TNF* -308 and *TNF* -238 have been associated with TNF expression levels (82, 97). This has led to intensive studies of these two SNPs and their association with not only insulin resistance and myocardial infarction but also many other diseases mostly related to the immune system (97). The literature is very conflicting and the *TNF* -308 G allele has been associated with both a decrease and an increase in insulin resistance for example. A systematic review that looked at many of the *TNF* associations found that most if not all of the claimed associations were not true (97). This has been backed up by studies that found no difference in expression between *TNF* alleles (98).

The most widespread method to find out if two alleles have different gene expression has been reporter gene constructs. In these constructs a DNA segment spanning the polymorphism is cloned in front of a reporter gene, the construct is then transfected into cell lines. The biological relevance of this method is limited by the absence of natural chromatin structure and regulation. A new method use the property of transcription that the amount of DNA polymerase associated with a segment of DNA is proportional to the transcriptional activity (98). This new method was used to investigate the transcription pattern of the *TNF* -308 A and G alleles. There was no difference in transcription levels from these two alleles (98). In a study of *LTA* haplotypes (99), A (*LTA* +80 G, *LTA* +252 C, *LTA* T26N Asn), B (*LTA* +80 T, *LTA* +252 T, *LTA* T26N Thr) and C (*LTA* +80 G, *LTA* +252 T, *LTA* T26N Thr), it was found that there was more transcription from haplotype A than B. A:B ratio 1.31 (95% CI 1.19- 1.44) (99). Cell lines with diplotype BC and BA produce less LTA than cell lines with diplotype AC. This lead to the conclusion that there is less transcription from

haplotype B than A or C, and that *LTA* +80 is possibly functional (99). It was found that a nuclear extract binds to the *LTA* +80 T allele but not the G allele. The sequence around *LTA* +80 looks like an E-box and a consensus E-box can compete for the nuclear extract that binds the T allele. ABF-1 was identified as a likely binding candidate and it was shown that the T allele has less expression than the G allele in the present of ABF-1 (99).

Figure 14: Annotation of *LTA* and *TNF* SNPs.



Ensembl output showing a 7.70 kb region of chromosome 6, which include *LTA* and *TNF*. Vertical lines under the genes represent SNPs. The SNPs genotyped in this study have been annotated with names and functional claims.

5.1.9 LGALS2

To find proteins that interact with LTA a Japanese research group performed a two-hybrid experiment in *E.coli* (34). They identified galectin-2 (LGALS2) as a binding protein for LTA. The binding was confirmed in an in vitro assay. They further validated the functional and biological interaction by obtaining the following evidence. LTA and galectin-2 co-immuno precipitate, they co-localised in cytoplasm and they co-stain in human coronary atheroma specimen. Galectin-2 also co-localised with A and B tubulins (34).

The Japanese group re-sequenced *LGALS2* in 32 myocardial infarction patients and found 17 new SNPs (34). These 17 SNPs were genotyped in 600 myocardial infarction subject and controls; one SNP (rs7291467 C to T) was found to be associated with myocardial infarction (34). This finding was confirmed in a study population of 2300 MI patients and 2000 controls. The strongest association was obtained using a recessive genetic model. The frequency of the T allele was lowest for the myocardial infarction cases. The chi-square value was 22.1 with a p value of 2.6×10^{-6} and the odds ratio was 1.57 (95% CI: 1.30-1.90) (34). It was then investigated if the SNP was functional. In a reporter gene construct, the T allele had 50% less transcription than the C allele in HELA and HepG2 cells (34). In an experiment where *LGALS2* expression was reduced by a siRNA the LTA concentration was also reduced even though *LTA* expression was the same. This led the authors to speculate that galectin-2 is involved in LTA secretion (34).

5.2 Prior hypothesis

There is a strong case for the association between the two *LTA* SNPs T26N and +252, and an increased risk for myocardial infarction. This association has been observed in three studies, two Japanese and one European, and there is a plausible molecular explanation for the mechanism behind the association. The association between *LGALS2* rs7291467 and a decreased risk for myocardial infarction has only been observed in one study, but the two associations are linked together so it is most likely that the both are either true or false.

1. We expect to find an association between *LTA* T26N N and *LTA* +252 C, and an increased risk for myocardial infarction. We also expect to find an association between *LGALS2* rs7291467 T and a decreased risk for myocardial infarction.

A single study has observed an association between *LTA* +252 and a decrease in insulin resistance. There have not been any reports of associations between *LGALS2* and insulin resistance but because galectin-2 is thought to affect LTA levels any association found with *LTA* could also be obtained with *LGALS2*. The metabolic syndrome traits are tightly linked and there may be a common biological mechanism behind them all. It is therefore reasonable to test for an association with all the metabolic syndrome traits. Based on the one study published, one could expect a decrease in the metabolic syndrome traits but this can be countered by the fact that the metabolic syndrome is a risk factor for myocardial infarction. Any hypothesis must therefore be two tailed.

2. We expect to find an association between *LTA* +252 C and haplotypes defined by *LTA* +252 C (*2 figure 16), and a decrease or increase in the metabolic syndrome traits. We also expect to find an association between *LGALS2* rs7291467 T and a decreased or increase in the metabolic syndrome traits.

The other *LTA* SNPs as well as the two *TNF* SNPs do not have strong evidence to found an a priori hypothesis. There is compelling evidence for that *LTA* +80/+81 is functional and can affect levels of LTA but there is no epidemiological data to suggest how these SNPs affect phenotypes. For *TNF* the data are so conflicting that no conclusion can be made. The analysis of *LTA* +80/+81 and the *TNF* SNPs must therefore be descriptive.

3. We will make a descriptive haplotype analysis to find out if *LTA* +80/+81 can affect any of the metabolic syndrome traits. We also wish to find out if there is any association for the *TNF* SNPs and if there is how it relates to *LTA*.

5.3 Subjects and Methods

5.3.1 Participants

Full details about the selection of participants are given in (100) and (101). A total of 4286 British women aged 60 to 79 years participated and 3,817 (89%) had successful blood samples and 3,606 (83%) had DNA available for genotyping. For some women it were either impossible to give or they refused to donate blood. DNA had not been extracted for a small number of samples, for reasons that did not bias the sample(102).

5.3.2 Measurements

Standing height was measured without shoes to the nearest millimetre. Weight was measured in light clothing without shoes to the nearest 0.1kg. Waist circumference was taken as midpoint between the lower rib and the iliac crest and hip circumference was taken as the largest circumference below the waist. Both two were measured to the nearest millimetre. Blood samples were taken after a 12-h fast. Glucose and insulin were measured on fasting venous plasma samples. Insulin was measured with a specific ELISA assay which does not cross-react with pro-insulin. Homa Score was calculated as the product of glucose and insulin concentration divided with 22.5. High density lipoprotein cholesterol (HDL-C) and triglycerides were measured on frozen serum. Blood pressure was taken seated using the right arm. MI was defined as self-report of a doctor diagnosis of angina or myocardial infarction and/or evidence in medical record, review at baseline, of either of these diagnoses (100, 103).

5.3.3 DNA preparation

DNA was extracted by the salting out procedure (104) from K-EDTA whole blood or red and white cell residues, which had been stored at -80°C for 1 to 2 years. Quantitation was by picoGreen assay and DNA concentrations were equalised by dilutions with water. Long term stock DNA aliquots were laid down and working 96-well plates of DNA dilutions to 10 ng/µl prepared. Degenerate oligo primer amplifications ('DOP-DNA') were made from dilution plates in order to conserve stock DNA and 384-well PCRs were performed from DOP-DNA representing 0.1 ng of original genomic DNA. The DOP protocol was a modified version of the method used by Cheung and Nelson (105) designed to minimise loss of representation of %GC-rich genomic regions.

5.3.4 Genotyping

Asymmetric PCR was performed on 2 µl of dried DOP amplified template in 384-well white PCR plates (Abgene®, Epsom, Surrey, UK) on a MJ Research PTC-225 DNA Engine Tetrad® (Genetic Research Instrumentation Ltd., Braintree, Essex, UK).
A detailed description of the LightTyper (Roche Diagnostics GmbH, Germany, Cat. No. 03357414001) Methodology and software is given in (106) The genotyping was performed using an ASO with a 5' florescent molecule and 3' phosphate. A second probe is designed so it anneals two nucleotides upstream of the ASO, this probe have a quencher (Dabcyl) attached 3'. Reaction conditions are in Table 4.

5.3.5 Primer and probe design

The LightTyper has a lower limit of 35°C for reliable detection of genotypes. Therefore the ASO should have an annealing temperature above 50°C, this would ensure that the mismatch ASO would melt of above 35°C. The quencher probe must have an annealing temperature 10°C higher than the ASO to ensure that it does not melt off before the ASO. Primer3 was used for the design of PCR primers. The annealing temperature was set to be 5°C higher than the dactyls probe. This is importance for the primer that anneals to the same string as the dabcyl probe. Primers and probes are in table 5.

5.3.6 Protocol for genotyping

1) Spin 384 plate 2000rpm 30 sec to ensure that the dry DNA pellet is in the bottom

2) Load 5µl PCR mix to each well (2500µl PCR mix is needed pr 384 plate)

3) Spin 348 plate 4000rpm 1 min, to ensure that PCR mix is in the bottom of the well and that there is no air bubbles

4) Seal PCR plate with Microseal

5) Perform PCR

6) Spin 2000 30 sec, to ensure that the PCR mix is in the bottom, if PCR mix is left at the edge of the well it could lead to contamination

7) Load 5µl mineral oil

8) Spin 4000 1min, to separate oil and PCR mix and ensure there is no air bubbles.

9) Scan plate

Table 4: Summary of PCR conditions.

PCR Mix	TNF -308/238	LTA T26N	LTA +252	LTA+80	LGALS2
Taq buffer	1x	1x	1x	1x	1x
dNTP's	0.2mM	0.2mM	0.2mM	0.2mM	0.2mM
$MgCL_2$	1,5mM	1,5mM	1,5mM	1,5mM	1,5mM
Right Primer	0.5uM	0.5uM	0.1uM	0.5uM	0.5uM
Left Primer	0.1uM	0.1uM	0.5uM	0.1uM	0.1uM
Flu Probe	0.1uM	0.2uM	0.025uM	0.1uM	0.1uM
Dabsyl Probe	0.2uM	0.2uM	0.2uM	0.4uM	0.2uM
TaqU/uL	0.01	0.01	0.01	0.05	0.01
PCR Reaction					
94c	2min	2min	2min	2min	2min
94c	20sec	20sec	20sec	20sec	20sec
Anneal 30sec	60	56	62	66	68
72C	1Min	1Min	1Min	30sec	1Min
Cycles	49	49	49	49	49
72C	2Min	2Min	2Min	2Min	2Min

Taq buffer: 10mM Thris-HCL, 10Mm KCL, pH 8.3

Table 5: Primer and Probes.

Name	Left Primer	Right primer
LTA+80	AGGCCCAGGCAGGCCGGGGAT	GCTGCCACTGCCGCTTCCTCTA
LTA+252	CGTGCTTTGGACTACCGCCCCGCAGTG	CCCCCGACCCCCGAGAGAGAGATCG
LTA T26N	GGGCTGCTGCTGGTTCTGCTG	GGGTGGATGCTTGGGTTCCTGAG
TNF -308	CCTCACACTCCCCATCCTC	CCTGCATCCTGTCTGGAAGT
LGALS2	AGCCTGGGCGACAGAGCGAGAC	TGTTGCAGGGCTCGGGGTGT
TNF-238	AAA GTT GGG GAC ACA CAA GC	ATC AGT CAG TGG CCC AGA AG

Name	Flu Probe	Dabcyl Probe
LTA+80	CACTGCTGGGCGGTA	GAGGCCCAGGCAGAGGGCAG
LTA +252	CAGAGAGGAACCATGGCAG	CAGAGGGAGACAGAGAGAGACAGGAAGGG
LTA T26N	CACAGCAACCTCAAACCT	GTCAGCACCCCAAGATGCATCTTG
TNF -308	CCGTCCCCATGCCCC	TGTGTGTAGGACCCTGGAGGCTGAA
LGALS2	CACACACACGTCTAACACCA	ACAGACACTCACAGACGTGTGCCCTG
TNF-238	CCT GCT CCG ATT CCG	ACC CCT CAC ACT CCC CAT CC

Table 6: Summary of the genotyping for SNPs used in the haplotype analyses.

Genotype code	LTA+80	LTA+81	LTA+252	LTAT26N	TNF308	TNF238	
11	1281	2977	1348	1397	2161	3049	Homozygote major allele
12	1518	298	1456	1442	947	325	Heterozygote
22	483	7	452	466	144	11	Homozygote minor allele
Total (11+12+22)	3282/91.4%	3282/91.4%	3256/90.6%	3305/92.0%	3252/90.5%	3385/94.2%	Genotype efficiency
-1	6	19	5	2	8	13	11 in Blank
-2	5	1	6	5	7	1	12 in Blank
-3	9	0	3	4	0	1	22 in Blank
-6	71	71	11	3	0	6	unknown in blank
-7	121	121	83	62	32	60	assay failure
-9	189	189	253	225	308	147	Unknown
-10	157	157	223	234	233	227	Blank
HWE							
X^2	0.93	0.025	3.413	9	9.3	0.55	
P	0.33	0.873	0.064	0.003	0.002	0.456	

5.3.7 Statistics

Genetic model

The SNPs analysed here are thought to affect transcription and therefore levels of LGALS2, LTA and TNF themselves or other proteins such as VCAM-1. The effect is moderate and both alleles are therefore expected to contribute to the phenotype. An additive genetic model is therefore most appropriate when analysing quantitative traits. We therefore decided to analyse our data using an additive genetic model.

For categorical traits such as myocardial infarction there could be a threshold for the phenotype to manifest, which would only be reached in homozygote individuals. A recessive genetic model could therefore be appropriate when analysing categorical traits.

Power calculation;

Power calculation was performed using the program PS[17] (107, 108). For all calculations, the significant level (ά) was set to 0.05 and the number of individuals was 3200. Standard Deviation of the phenotypes can be found in table 7 and the Standard Deviation of the regression error was calculated using SPSS 14.

SPSS 12 was used to make box whisker plot for all traits. Two individuals had insulin readings of 1000, which were classified as outliers. As insulin levels of 1000μunits/l are non-physiological, they must be instrument errors and these two individuals were excluded from all analyses.

[17] (http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize)

Insulin, glucose and triglycerides were positively skewed and they did not fit to a normal distribution. The natural logarithm of the values was used in all analyses. The results were back transformed.

SPSS 12 basic tables function was used to calculate mean, Standard Deviation and Standard Error of Mean (SEM).

SPSS 12 linear regression (An F test, 1 degree of freedom) was used to calculate regression coefficients, 95% confidence intervals and p values.

The haplotype analysis was preformed with HTR[18] according to the instruction give in the manual.

Statistical analyses were performed following the advice of Dr Debbie A Lawlor or Santiago Rodriguez. The analysis of MI data was performed by Dr Debbie A Lawlor.

[18] http://statgen.ncsu.edu/zaykin/htr.htm

5.4 Results

5.4.1 LGALS2

Of the 3,272 women who were genotyped 1,210 (37.0%) were homozygous for the major allele (TT), 1527 (46.7%) were heterozygous (TC) and 535 (16.4%) were homozygous for the minor allele (CC). The genotype frequencies were in Hardy-Weinberg equilibrium (p = 0.15)

LGALS2 rs7291467 was associated with levels of: glucose, insulin and HOMA score, such that the lowest level or score was observed for the C allele. The decrease per allele (additive genetic model) was 0.011 (95% CI, -0.02,-0.00) p = 0.02 for glucose, 0.040 (95% CI, -0,07,-0.00) p = 0.017 for insulin and 0.036 (95% CI, -0.07, -0.00) p= 0.035 for HOMA score. After adjustment for age, BMI and triglycerides the association with glucose and insulin was unchanged. As it can be seen form table 7 insulin do not satisfy the condition for an additive model, but clearly follows a recessive genetic model. It is harder to judge which genetic model glucose and HOMA score follow. In a recessive genetic model the association with insulin improved 0.066 (95% CI, -0.1, -0.02) p= 0.001 while the association with HOMA score was unchanged and glucose was no longer statistically significant (Table 7). There was no association with myocardial infarction (Table 9).

Table 7: Genotypes of *LGALS2* rs7291467 and their association with the metabolic syndrome traits

	Mean (SD) [SEM]			TT vs TC + CC		TT vs TC vs CC		TT + TC vs CC	
	TT N=1209 37%	TC N=1526 46.7%	CC N=535 16.3%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	69 (5.5) [0.2]	68.8 (5.4) [0.1]	68.8 (5.5) [0.2]	-0.24 (-0.62,0.15)	0.2	-0.13 (-0.40,0.14)	0.3	-0.07 (-0.57,0.44)	0,8
BMI (kg/m^2)	27.6 (5.0) [0.1]	27.6 (5.1) [0.1]	27.5 (4.9) [0.2]	-0.2 (-0.38,0.34)	0.9	-0.03 (-0.28,0.21)	0.8	-0.08 (-0.54,0.39)	0.7
Waist: hip ratio (x100)	82.1 (6.7) [0.0]	82 (6.8) [0.0]	81.8 (7.0) [0.0]	-0.02 (-0.01,-0.00)	0.4	-0.18 (-0.52,0.15)	0.3	-0.00 (-0.01,0.00)	0.4
Systolic BP (mmHG)	147.7 (24.8) [0.7]	148 (25.4) [0.7]	146.7 (25.9) [1.1]	-0.04 (-1.84,1.76)	1	-0.33 (-1.57,0.91)	0.6	-1.14 (-3.45,1.20)	0.3
Diastolic BP (mmHG)	79.3 (11.4) [0.3]	79.7 (11.9) [0.3]	79.7 (11.8) [0.5]	0.42 (-0.41,1.26)	0.3	0.27 (-0.31,0.84)	0.8	0.2 (-0.88,1.29)	0.7
HDL-C (mmol/l)	1.7 (0.5) [0.0]	1.6 (0.4) [0.0]	1.7 (0.5) [0.0]	0.00 (-0.03,0.03)	1	0.01 (-0.02,0.03)	0.6	0.02 (-0.02,0.06)	0.4
Triglycerides (mmol/l)	1.67 (1.60) [0.01]	1.70 (1.55) [0.01]	1.67 (1.62) [0.02]	0.01 (-0.02,0.4)	0.6	0.00 (-0.02,0.02)	0.9	-0.01 (-0.06,0.03)	0.5
Glucose (mmol/l)	5.99 (1.22) [0.01]	5.93 (1.20) [0.00]	5.81 (1.21) [0.01]	-0.01 (-0.03,0.00)	0.1	-0.011 (-0.02,-0.0)	0.02	-0.02 (-0.04,0.00)	0.03
Insulin (µunit/l)	7.32 (1.97) [0.02]	6.82 (1.92) [0.02]	6.82 (1.95) [0.03]	-0.07 (-0.11,-0.02)	0.007	-0.040 (-0.07,-0.00)	0.017	-0.03 (-0.09,-0.03)	0.3
HOMA score	1.73 (1.90) [0.02]	1.65 (1.86) [0.02]	1.62 (1.86) [0.03]	-0.05 (-0.10,-0.00)	0.036	-0.036 (-0.07,-0.00)	0.035	-0.04 (-0.1,0.02)	0.2

Genotypes of *LGALS2* rs7291467 [T/C MAF = 40%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of Mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate the regression coefficient, confidence interval and the associated p value.

Table 9: Genotypes of *LGALS2* rs7291467 and their association with MI

	LGALS2	
	No	MI
11	990 (82%)	220 (18%)
12	1300 (85%)	225 (15%)
22	446 (83%)	89 (17%)
X^2	2	
P	0.16	

5.4.2 LTA and TNF

Six SNPs were genotyped in the *LTA-TNF* region on chromosome 6. Four *LTA* SNPs: +80, +81, +252, T26N and two *TNF* SNPs: -308, -238. Genotype efficiency was between 90 and 94%. *LTA* +252 T and T26N Thr were always observed together; out of 3018 individuals who had genotypes for both SNPs only 18 individuals were not in phase. Genotype frequencies for all SNPs, except *LTA* T26N and *TNF* -308, were in Hardy-Weinberg equilibrium. A detailed summary for all SNP are given in table 6.

The *LTA* +252 C allele was associated with lower BMI and diastolic blood pressure. There were a mean decrease of, 0.44 BMI units (95% CI -0.69,-0.18; p = 0.001) and 0.70 mmHG (95% CI -1.30, -0.11; p = 0.02), per C allele over all three genotypes. Other SNPs showed borderline association, but after correction for multiple testing they

were reduced to weak trends. No SNP was associated with myocardial infarction, using an additive or recessive genetic model. Data for all SNPs are given in table 10 to 15.

Table 10: Genotypes of *LTA* +252 and their association with the metabolic syndrome traits.

	Mean (SD) [SEM]			TT vs TC + CC		TT vs TC vs CC		TT + TC vs CC	
	TT N=1337 42%	TC N=1437 45%	CC N=442 13%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	68.9 (5.5) [0.1]	68.8 (5.5) [0.1]	68.6 (5.3) [0.3]	-0.2 (-0.58,0.18)	0.3	-0.16 (-0.43,0.12)	0.2	-0.24 (-0.79,-0.31)	0.4
BMI (kg/m^2)	28 (5.3) [0.1]	27.5 (4.8) [0.1]	27.2 (4.9) [0.2]	-0.57 (-0.92,-0.22)	0.001	-0.44 (-0.69,-0.18)	0.001	-0.57 (-1.0,-0.06)	0.03
Waist: hip ratio (x100)	82.1 (6.8) [0.0]	81.9 (6.7) [0.0]	81.4 (7) [0.0]	0.00 (0.00,0.00)	0.12	-0.34 (-0.69,0.00)	0.05	-0.00 (-0.01,00)	0.09
Systolic BP (mmHG)	148.4 (25.3) [0.7]	146.8 (24.9) [0.7]	147.4 (26.2) [1.2]	-1.47 (-3.24,0.31)	0.10	-0.77 (-2.03,0.50)	0.2	-0.14 (-2.68,2.40)	0.9
Diastolic BP (mmHG)	79.9 (12) [0.3]	79.5 (11.4) [0.3]	78.3 (11.8) [0.6]	-0.68 (-1.50,0.15)	0.1	-0.7 (-1.3,-0.11)	0.02	-1.4 (-2.6,-0.24)	0.02
HDL-C (mmol/l)	1.7 (0.4) [0.0]	1.7 (0.5) [0.0]	1.7 (0.4) [0.0]	0.01 (-0.02,0.04)	0.4	0.01 (-0.02,0.03)	0.6	0.00 (-0.05,0.0)	0.9
Triglycerides (mmol/l)	1.72 (1.56) [0.01]	1.65 (1.60) [0.01]	1.67 (1.58) [0.02]	-0.041 (-0.07,-0.00)	0.01	-0.023 (-0.46,0.0)	0.05	-0.00 (-0.5,0.04)	0.8
Glucose (mmol/l)	5.93 (1.20) [0.01]	5.93 (1.2) [0.01]	5.99 (1.22) [0.01]	-0.00 (-0.01,0.01)	1	0.00 (-0.01,0,01)	0.5	0.01 (-0.00,0.03)	0.2
Insulin (μunit/l)	7.03 (1.93) [0.02]	7.10 (1.9) [0.02]	6.96 (2.08) [0.03]	0.00 (-0.05,0.05)	0.9	0.01 (-0.03,0.04)	0.9	-0.01 (-0.07,0.5)	0.8
HOMA score	1.70 (1.88) [0.02]	1.68 (1.84) [0.02]	1.60 (1.86) [0.03]	-0.02 (-0.07,02)	0.3	-0.03 (-0.06,0.01)	0.1	-0.05 (-0.12,0.01)	0.12

Genotypes of *LTA* +252 [T/C MAF 36%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate the regression coefficient, confidence interval, and the associated p value.

Table 11: Genotypes of *LTA* +80 and their association with the metabolic syndrome traits.

	Mean (SD) [SEM]			GG vs GT + TT		GG vs GT vs TT		GG + GT vs TT	
	GG N=1259 39%	GT N=1502 46%	TT N=482 15%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	68.7 (5.4) [0.2]	68.9 (5.5) [0.1]	68.8 (5.5) [0.3]	0.1 (-0.28,-0.49)	0.59	0.04 (-0.23,0.32)	0.7	-0.02 (-0.55,0.50)	0.9
BMI (kg/m^2)	27.5 (4.9) [0.1]	27.4 (4.8) [0.1]	28.1 (5.5) [0.3]	0.13 (-0.22,0.48)	0.46	0.22 (-0.03,0.47)	0.08	0.61 (0.12,1.09)	0.013
Waist: hip ratio (x100)	81.9 (6.7) [0.0]	81.7 (6.7) [0.0]	81.9 (7.3) [0.0]	-0.00 (-0.00,0.00)	0.58	-0.03 (-0.37,0.30)	0.86	0.00 (-0.00,0.00)	0.7
Systolic BP (mmHG)	146.7 (25.6) [0.7]	147.9 (25) [0.6]	147.6 (24.4) [1.1]	1.07 (-0.70,2.85)	0.2	0.58 (-0.67,1.83)	0.36	0.22 (-2.22,2.66)	0.9
Diastolic BP (mmHG)	78.7 (11.7) [0.3]	79.9 (11.8) [0.3]	79.5 (11.8) [0.5]	1.1 (0.28,1.94)	0.009	0.23 (-0.2,0.5)	0.053	0.11 (-1.02,1.26)	0.8
HDL-C (mmol/l)	1.7 (0.5) [0.0]	1.7 (0.4) [0.0]	1.7 (0.4) [0.0]	0.02 (-0.02,0.05)	0.3	0.00 (-0.02,0.03)	0.6	0.00 (-0.05,0.04)	0.77
Triglycerides (mmol/l)	1.68 (1.58) [0.01]	1.65 (1.58) [0.01]	1.72 (1.52) [0.02]	-0.00 (-0.04,0.02)	0.6	0.00 (-0.03,0.02)	0,8	0.03 (-0.02,0.07)	0.2
Glucose (mmol/l)	5.93 (1.2) [0.01]	5.93 (1.18) [0.00]	5.93 (1.2) [0.01]	-0.00 (-0.02,0.01)	0.8	0.00 (-0,01,0.00)	0.8	0.02 (-0.05,0.08)	0.6
Insulin (µunit/l)	6.89 (1.97) [0.02]	6.89 (1.90) [0.02]	7.03 (1.97) [0.03]	0.00 (-0.04,0.05)	0.9	0.00 (-0.03,0.04)	0.7	0.00 (-0.01,0.02)	0.5
HOMA score	1.65 (1.90) [0.02]	1.65 (1.82) [0.02]	1.68 (1.84) [0.03]	0.01 (-0.04,0.06)	0.7	0.01 (-0.03,0.04)	0.62	0.01 (-0.05,0.08)	0.62

Genotypes of *LTA* +80 [G/T MAF 38%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate the regression coefficient, confidence interval and the associated p value.

Table 12: Genotypes of *LTA* T26N and their association with the metabolic syndrome traits

	Mean (SD) [SEM]			GG vs GT + TT		GG vs GT vs TT		GG + GT vs TT	
	CC N=1386 42%	CA N=1423 44%	AA N=458 14%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	68.9 (5.5) [0.1]	68.9 (5.5) [0.1]	68.7 (5.3) [0.2]	-0.07 (-0.44,0.31)	0.72	-0.07 (-0.34,-0.20)	0.6	-0.15 (-0.69,0.39)	0.6
BMI (kg/m^2)	27.8 (5.3) [0.1]	27.5 (4.8) [0.1]	27.2 (5.0) [0.2]	-0.4 (-0.77,-0.06)	0.02	-0.30 (-0.55,-0.06)	0.013	-0.45 (-0.95,0.05)	0.08
Waist: hip ratio (x100)	82 (6.9) [0.0]	82 (6.7) [0.0]	81.5 (7.0) [0.0]	-0.00 (-0.00,0.00)	0.44	-0.22 (-0.56,0.12)	0.2	-0.01 (-0.01,0.00)	0.1
Systolic BP (mmHG)	148.1 (25.2) [0.7]	146.8 (24.8) [0.7]	147.3 (25.8) [1.2]	-1.15 (-2.89,0.60)	0.2	-0.59 (-1.83,0.65)	0.35	-0.08 (-2.6,2.4)	0.9
Diastolic BP (mmHG)	79.7 (12) [0.3]	79.4 (11.4) [0.3]	78.3 (11.8) [0.6]	-0.53 (-1.35,0.28)	0.2	-0.61 (-1.19,-0.03)	0.043	-1.3 (-2.47,-0.15)	0.03
HDL-C (mmol/l)	1.7 (0.4) [0.0]	1.7 (0.4) [0.0]	1.7 (0.5) [0.0]	0.00 (-0.03,0.03)	0.9	0.00 (-0.02,0.03)	0.67	0.02 (-0.03,0.06)	0.5
Triglycerides (mmol/l)	1.72 (1.57) [0.01]	1.65 (1.58) [0.01]	1.67 (1.58) [0.02]	-0.04 (-0.07,-0.00)	0.027	-0.02 (-0.04,-0.03)	0.08	0.00 (-0.05,0.04)	0.7
Glucose (mmol/l)	5.93 (1.20) [0.00]	5.93 (1.22) [0.01]	5.99 (1.22) [0.01]	0.00 (0.00,0.02)	0.5	-0.01 (-0.01,0.02)	0.3	-0.02 (-0.08,0.05)	0.6
Insulin (µunit/l)	6.96 (1.93) [0.02]	7.1 (1.92) [0.02]	6.89 (2.08) [0.03]	0.02 (-0.03,0.06)	0.49	-0.01 (-0.02,0.05)	0.8	0.00 (-0.01,0.03)	0.4
HOMA score	1.68 (1.88) [0.02]	1.70 (1.84) [0.02]	1.58 (1.88) [0.03]	-0.01 (-0.06,0.03)	0.6	-0.02 (-0.05,0.01)	0.2	-0.06 (-0.13,0.00)	0.08

Genotypes of *LTA* T26N [C/A MAF 36%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate the regression coefficient, confidence interval and the associated p value.

Table 13: Genotypes of *TNF* -308 and their association with the metabolic syndrome traits.

	Mean (SD) [SEM]			GG vs GA + AA		GG vs GA vs AA		GG + GA vs AA	
	GG N=2141 67%	GA N=933 29%	AA N=138 4%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	69 (5.5) [0.1]	68.5 (5.4) [0.2]	69 (5.4) [0.5]	-0.4 (-0.80,0.00)	0.05	-0.25 (-0.58,0.09)	0.1	0.2 (-0.74,1.13)	0.7
BMI (kg/m^2)	27.7 (5.1) [0.1]	27.6 (4.9) [0.2]	27.1 (4.3) [0.4]	-0.2 (-0.59,0.15)	0.25	-0.21 (-0.52,0.10)	0.2	-0.6 (-1.41,0.31)	0.21
Waist: hip ratio (x100)	82 (6.8) [0.0]	81.8 (6.8) [0.0]	81.7 (6.6) [0.0]	0.00 (0.00,0.00)	0.5	-0.15 (-0.57,0.27)	0.5	-0.00 (-0.01,0.01)	0.7
Systolic BP (mmHG)	147.9 (25) [0.5]	146.9 (25.3) [0.8]	147.1 (28.2) [2.4]	-1 (-2.8,0.86)	0.3	-0.74 (-2.29,0.80)	0.3	-0.5 (-4.78,3.82)	0.828
Diastolic BP (mmHG)	79.5 (11.8) [0.3]	79.4 (11.4) [0.4]	78.4 (12.6) [1.1]	-0.2 (-1.1,0.6)	0.6	-0.32 (-1.04,0.04)	0.4	-1.13 (-3.13,0.87)	0.27
HDL-C (mmol/l)	1.7 (0.4) [0.0]	1.6 (0.5) [0.0]	1.7 (0.5) [0.0]	-0.01 (-0.04,0.2)	0.7	0.00 (-0.03,0.02)	0.8	0.21 (-0.06,0.10)	0.6
Triglycerides (mmol/l)	1.68 (1.58) [0.01]	1.68 (1.58) [0.02]	1.67 (1.58) [0.04]	-0.00 (-0.04,0.03)	0.7	-0.01 (0.03,0.02)	0.7	-0.00 (-0.09,0.07)	0.83
Glucose (mmol/l)	5.93 (1.20) [0.00]	5.93 (1.22) [0.01]	6.05 (1.19) [0.01]	0.01 (-0.00,-0.02)	0.2	-0.01 (0.00,0.02)	0.1	0.02 (-0.01,0.05)	0.2
Insulin (µunit/l)	6.96 (1.93) [0.01]	7.17 (1.95) [0.02]	6.96 (1.20) [0.05]	0.03 (-0.02,0.08)	0.21	0.03 (-0.01,0.07)	0.3	-0.00 (-0.11,0.11)	1
HOMA score	1.67 (1.86) [0.01]	1.70 (1.86) [0.02]	1.7 (1.8) [0.05]	0.03 (-0.02,0.08)	0.3	0.02 (-0.02,0.06)	0.3	0.04 (-0.08,0.15)	0.57

Genotypes of *TNF* -308 [G/A MAF 19%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate the regression coefficient, confidence interval the associated p value.

Table 14: Genotypes of *TNF* -238 and their association with the metabolic syndrome traits.

	Mean (SD) [SEM]			AA vs AG + GG		AA vs AG vs GG		AA + AG vs GG	
	AA N=3012 90%	AG N=321 9%	GG N=11 1%	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)	Regression coefficient (95% CI)	p-value Trend (1df)
Age (years)	68.9 (5.5) [0.1]	68.7 (5.3) [0.3]	67.5 (4.9) [1.5]	-0.2 (-0.8,0.41)	0.52	-0.23 (-0.82,0.36)	0.4	-1.3 (-4.54,1.95)	0.43
BMI (kg/m^2)	27.6 (5.0) [0.1]	27.5 (5.2) [0.3]	25.8 (5.7) [1.7]	-0.13 (-0.7,0.44)	0.65	-0.18 (-0.73,0.36)	0.5	-1.8 (-4.77,1.17)	0.23
Waist: hip ratio (x100)	81.9 (6.8) [0.0]	82.2 (6.5) [0.0]	83.3 (11.1) [0.0]	0.00 (-0.01,0.01)	0.4	0.36 (-0.37,1.09)	0.3	0.014 (-0.03,0.06)	0.5
Systolic BP (mmHG)	147.6 (25) [0.0]	147 (26.7) [1.5]	156.7 (27.9) [8.4]	-0.3 (-3.13,2.6)	0.9	0.05 (-2.67,2.77)	1	9.17 (-5.73,24.08)	0.2
Diastolic BP (mmHG)	79.5 (11.6) [0.2]	79.1 (12.6) [0.7]	87.8 (19.4) [5.8]	-0.05 (-1.4,1.3)	0.9	0.24 (-1.03,1.51)	0.7	8.4 (1.44,15.34)	0.02
HDL-C (mmol/l)	1.7 (0.5) [0.0]	1.7 (0.5) [0.0]	1.7 (0.4) [0.1]	0.014 (-0.04,0.07)	0.6	0.01 (-0.03,0.06)	0.6	0.07 (-0.2,033)	0.612
Triglycerides (mmol/l)	1.68 (1.58) [0.01]	1.69 (1.55) [0.02]	1.67 (1.63) [0.15]	0.00 (-0.04,0.06)	0.77	0.01 (-0.04,0.06)	0.8	-0.011 (-0.3,0.26)	0.9
Glucose (mmol/l)	5.93 (1.21) [0.00]	5.87 (1.16) [0.01]	6.69 (1.57) [0.13]	-0.00 (-0.03,0.01)	0.45	0.00 (-0.02,0.02)	0.7	0.12 (0.01,0.23)	0.03
Insulin (μunit/l)	6.96 (1.93) [0.01]	6.82 (1.91) [0.04]	7.10 (2.69) [0.30]	-0.02 (-0.09,0.06)	0.6	0.02 (0.09,0.05)	0.7	0.022 (-0.37,0.41)	0.9
HOMA score	1.67 (1.84) [0.01]	1.67 (1.93) [0.04]	1.58 (2.44) [0.28]	-0.00 (-0.08,0.07)	0.9	0.01 (0.07,0.06)	0.9	-0.06 (-0.44,0.33)	0.77

Genotypes of *TNF* -238 [A/G MAF 5%] and their association with the metabolic syndrome traits. Mean, Standard Deviation (SD) and Standard Error of mean (SEM) were calculated using SPSS basic table functions. Linear regression was used to calculate per allele mean difference and the associated p value.

Table 15: Genotypes of *LTA* and *TNF* and their association with MI

	LTA +80		LTA +81		LTA +252		LTA T26N		TNF -308		TNF -238	
	No	MI	No	MI	No	MI	No	MI	No	MI	No	MI
11	1048 (83%)	213 (17%)	2425 (82%)	518 (18%)	1096 (82%)	241 (18%)	1136 (82%)	250 (18%)	1750 (82%)	391 (18%)	2529 (84%)	485 (16%)
12	1231 (82%)	271 (18%)	254 (86%)	41 (14%)	1174 (82%)	263 (18%)	1152 (81%)	271 (19%)	775 (83%)	160 (17%)	269 (84%)	52 (16%)
22	406 (76%)	76 (16%)	6 (86%)	1 (14%)	375 (84%)	63 (16%)	388 (85%)	70 (15%)	120 (87%)	18 (13%)	10 (91%)	1 (9%)
X^2	0.03		2.7		0.8		0.63		2.3		0.03	
P	0.86		0.1		0.4		0.42		0.1		0.86	

5.4.3 Haplotype analyses

Haplotype analyses were performed on four data sets for each trait. The full data set containing all 6 SNPs. The *LTA* and *TNF* SNPs were also analysed separately. Two reduced sets of *LTA* SNPs were also analysed for them self.

Table 16: Haplotypes.

NAME	LTA +80	LTA +81	LTA +252	LTA T26N	TNF -308	TNF -238	Freq
*1	2 (T)	1 (G)	1 (T)	1 (Thr)	1 (G)	1 (A)	0.383
*2	1 (G)	1 (G)	1 (T)	1 (Thr)	1 (G)	1 (A)	0.207
*3	1 (G)	1 (G)	2 (C)	2(Asn)	1 (G)	1 (A)	0.171
*4	1 (G)	1 (G)	2 (C)	2(Asn)	2 (A)	1 (A)	0.138
*5	1 (G)	1 (G)	1 (T)	1 (Thr)	1 (G)	2 (G)	0.048
*6	1 (G)	2 (A)	2 (C)	2(Asn)	2 (A)	1 (A)	0.046

Sum of Freq: 0.997

Haplotypes. Rows (2-7) represent haplotypes *1 to *6. Columns (2-7) represent individual SNPs. Frequency for each haplotype is given column 8. 1 denote major allele, 2 denotes minor allele. A, T, C, G indicates nucleotides. Thr and Asn indicate amino acids. Haplotypes were computed using Phase 2 with the following settings 1000 100 1000.

There are three main haplotypes for *TNF*, which cover 99.6% of the 3067 individuals for which genotypes were available. There is a slight variation in the number for each trait because data was not obtained for all individuals. The *TNF* *1 (-308 G, -238 A) haplotype is associated with higher BMI then *2 (-308A, -238A) and *3 (-308G, -

238G). The mean difference between *1 and *2 or *1 and *3 is 0.29 and 0.4 BMI units, respectively. The P value obtained for *1 being different from all other haplotypes is p=0.043 (Figure 15). The interpretation of this is that * 2 and *3 is associated with lower BMI. There was no difference between the three haplotypes for the other traits. Full details for all traits are given in table 17 to 25

.

Figure 15: Haplotype analysis of *TNF* SNPs.

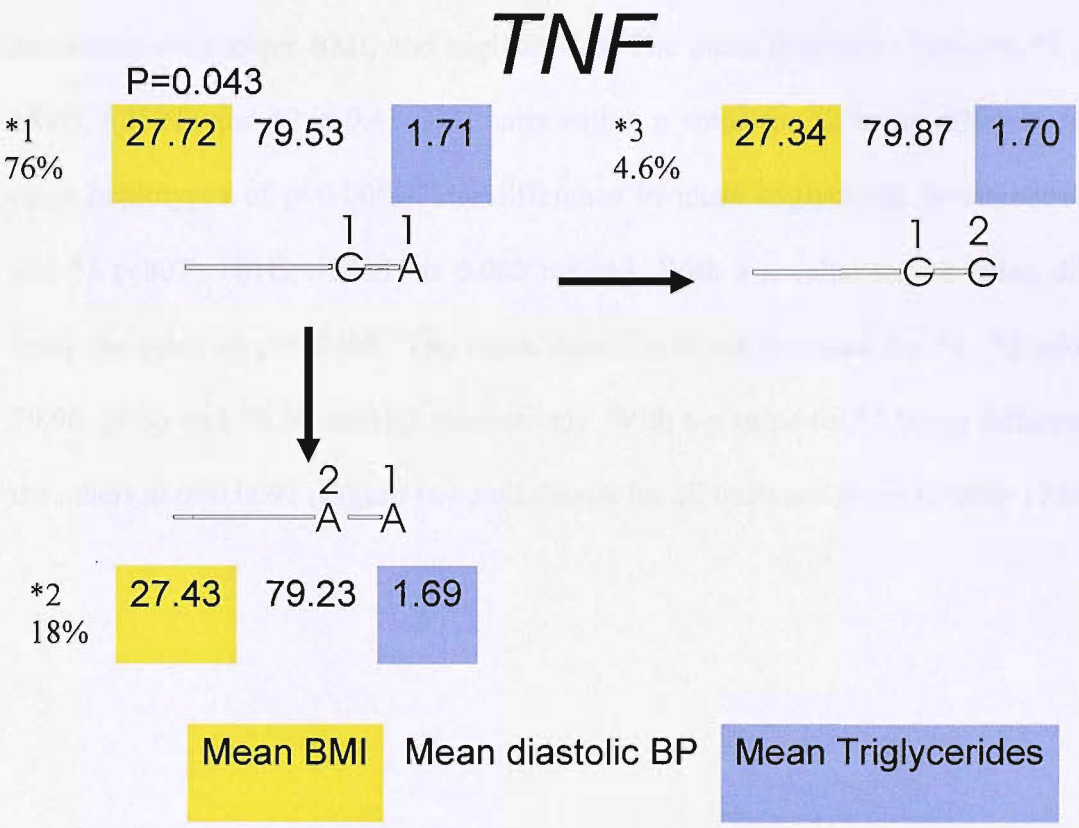


Diagram of *TNF* haplotypes showing their mutual relations. Means for the traits BMI, Blood pressure and triglycerides are given for each haplotype. If the mean of a trait for one haplotype is statistically significant from the other haplotypes it is indicated with a p value (Over all P value was non-significant $p > 0.05$ for all analyses see appendix III). The results were obtained with HTR.

Four SNPs were genotyped in *LTA,* but one SNP T26N was not in Hardy-Weinberg equilibrium. LTA +252 C and T26N N are always observed on the same haplotype, which makes it possible to exclude T26N. I therefore analysed two data sets, one with all four SNPs and one with three SNPs (+80, +81 and +252). The best result was obtained for the reduced data set, which is therefore summarised here.

There are four haplotypes for *LTA,* which cover 99.8% of the 2964 individuals for which genotype data were available. The *2 (+80G, +81G, +252C) haplotype is associated with lower BMI, and triglycerides. The mean difference between *1 (+80T, +81G, +252T) and *2 is 0.41 BMI units with a p value for *2 being different from all other haplotypes of p=0.0085. The difference in mean triglyceride levels between *2 and *3 (+80T, +81G, +252T) is 0.065 mmol/l. With a p value for *2 being different from the other at p=0.0108. The mean diastolic blood pressure for *1, *2 and *3 is 79.96, 78.93 and 79.26 mmHG, respectively. With a p value for *1 being different from the others at p=0.0097 (Figure 16). Full details for all traits are given in table 17 to 25.

Figure 16: Haplotype analysis of *LTA* SNPs.

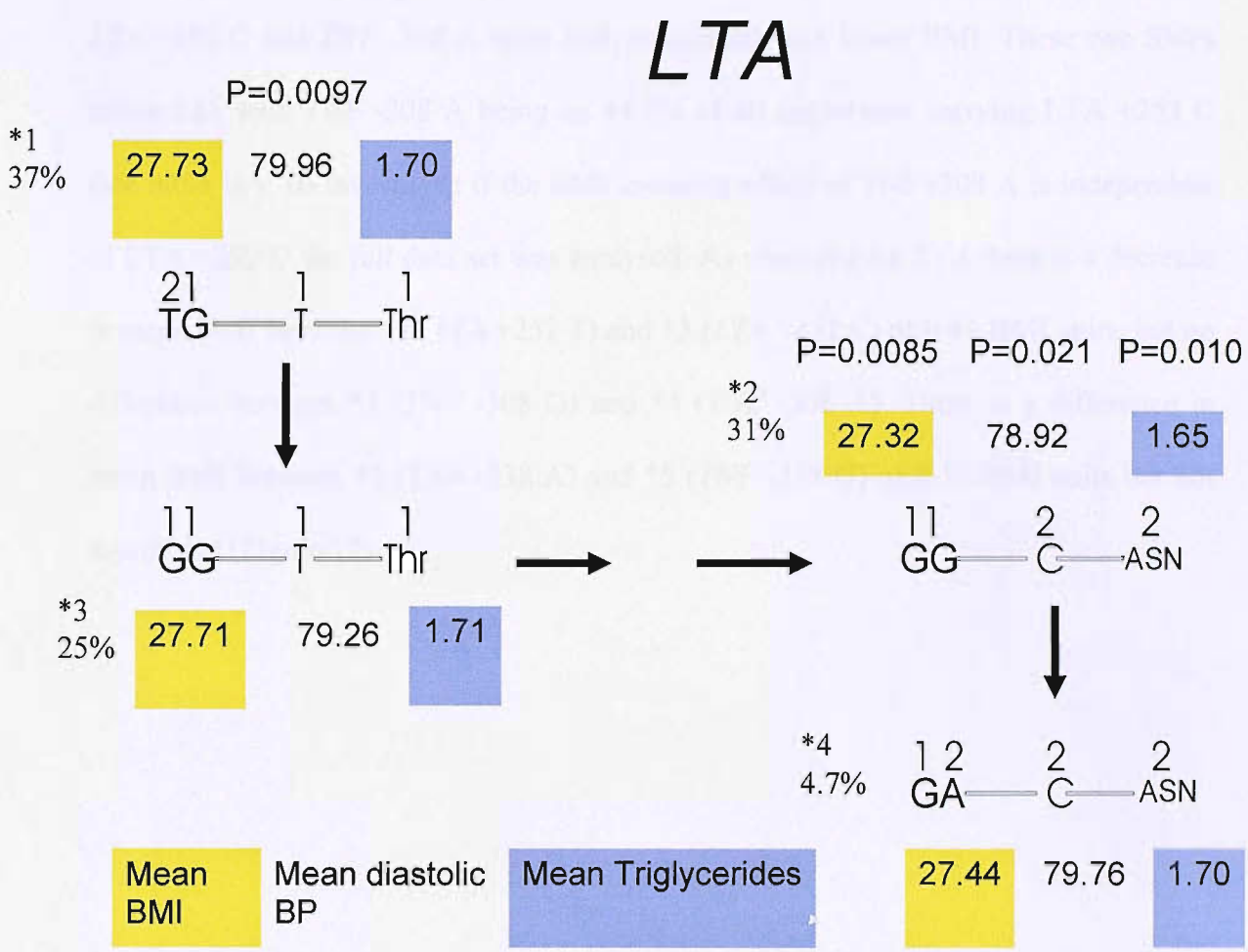


Diagram of *LTA* haplotypes showing their mutual relations. Means for the traits BMI, Blood pressure and triglycerides are given for each haplotype. If the mean of a trait for one haplotype is statistically significant from the other haplotypes it is indicated with a p value (Over all P value was non-significant $p > 0.05$ for all analysis see appendix III). The results were obtained with HTR.

LTA +252 C and *TNF* -308 A were both associated with lower BMI. These two SNPs are in LD, with *TNF* -308 A being on 44.7% of all haplotypes carrying LTA +252 C (see table 16). To investigate if the BMI lowering effect of TNF -308 A is independent of LTA +252 C the full data set was analysed. As observed for *LTA* there is a decrease in mean BMI between *1 (*LTA* +252 T) and *3 (*LTA* +252 C) of 0.41 BMI units, but no difference between *3 (*TNF* -308 G) and *4 (*TNF* -308 A). There is a difference in mean BMI between *2 (*TNF* -238 A) and *5 (*TNF* -238 G) of 0.57 BMI units but not significant (Figure 17).

Figure 17: Haplotype analysis of *LTA* and *TNF* SNPs together.

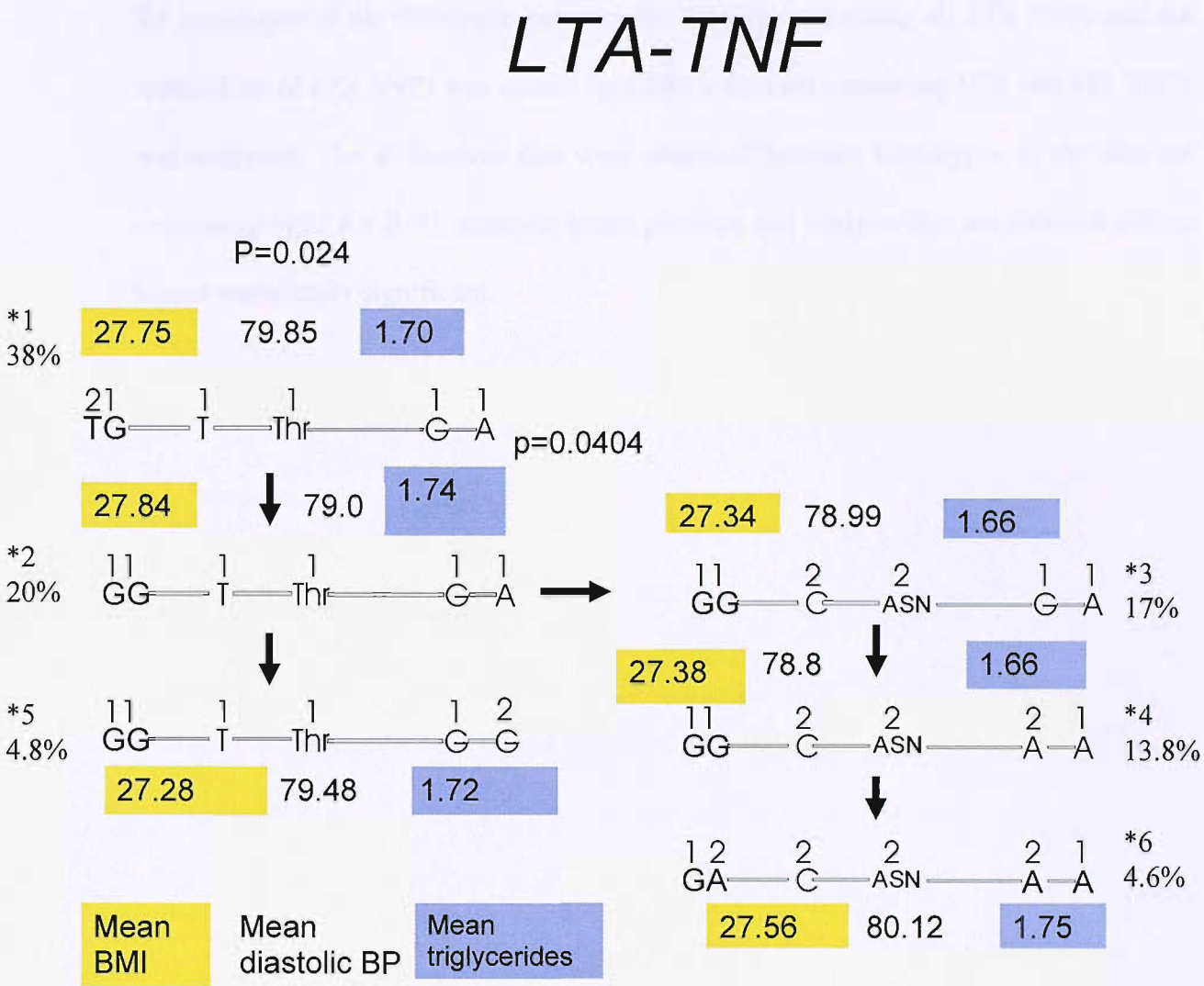


Diagram of *LTA-TNF* haplotypes showing their mutual relations. Means for the traits BMI, blood pressure and triglycerides are given for each haplotype. If the mean of a trait for one haplotype is statistically significant from the other haplotypes it is indicated with a p value (Over all P value was non-significant $p > 0.05$ for all analyses see appendix III). The results were obtained with HTR.

To investigate if the difference between the data set containing all *LTA* SNPs and the reduced set of *LTA* SNPs was caused by T26N a data set containing *LTA* +80 +81 T26N was analysed. The differences that were observed between haplotypes in the data set containing +252 for BMI, diastolic blood pressure and triglycerides are reduced and no longer statistically significant.

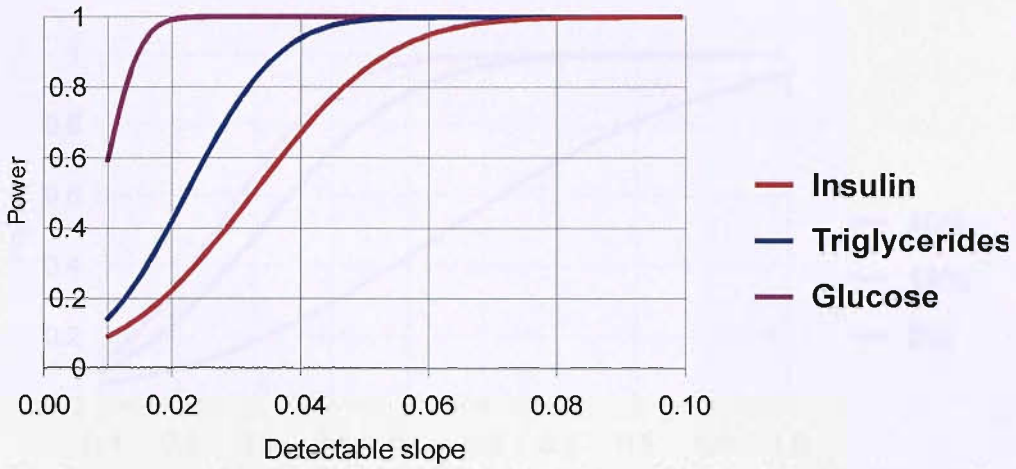
5.4.4 10 years coronary heart disease risk

The 10 years risk for developing CHD has been estimated in large population studies. Based on these estimates people who are homozygote for LTA *3 have 6% less chance for developing CHD over a 10 years period than people homozygote for LTA *1(109-111).

5.4.5 Power calculation.

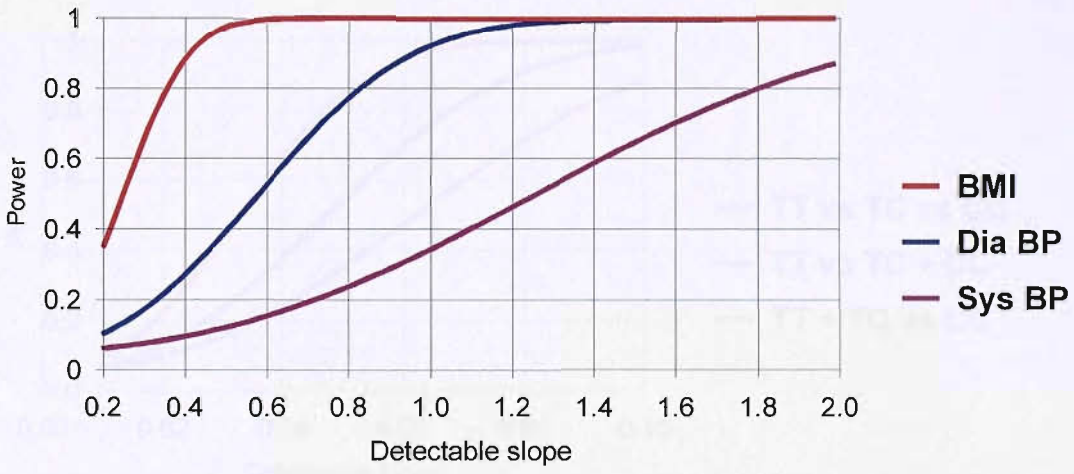
The power to detect a true association is determined by the Standard Deviation (SD) of the phenotype, the minor allele frequency of the SNP (SD of independent variable), the SD of the regression errors (107, 108), and the genetic model. Power calculation was performed for selected phenotypes, allele frequencies and genetic models to investigate if difference in power could explain that association was not observed for all SNPs and traits. The most important observation is that triglycerides is more powered than insulin and lack of power can therefore not explain that no association was observed with triglycerides (Graph 14). Also the power for SNPs with MAF between 19% and 40% is similar and difference in power can therefore not explain that the associations observed for *LGALS2* was not observed for *LTA* or *TNF* except for *TNF* – 238 (Graph 16). The choice of genetic model alters power because the SD of the genotypes changes (Graph 17). The additive model is most powered and should therefore be used unless there is a molecular model that suggests a recessive or dominant model.

Graph 14: Power calculation for; insulin, triglycerides and glucose



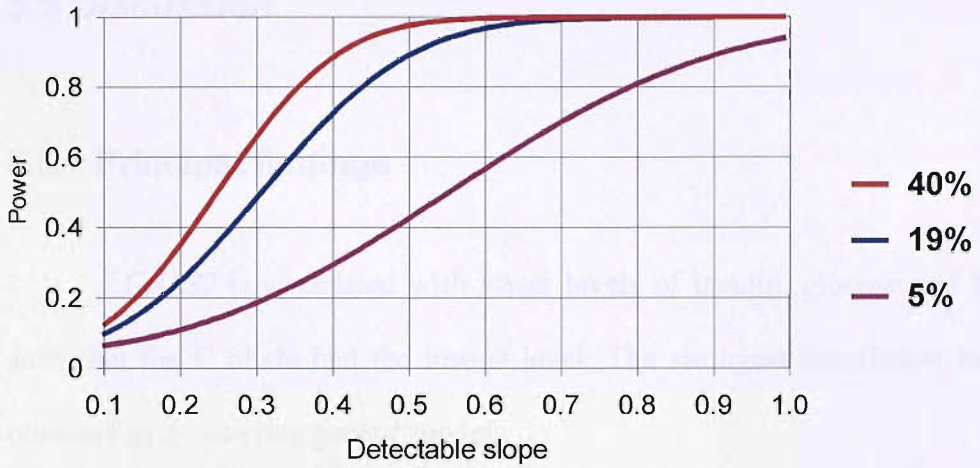
The Power as a function of detectable slope was calculated for insulin, triglycerides and glucose using genotypes of LGALS2 rs7291467 in an additive model as an example.

Graph 15: Power calculation for BMI and Blood pressure.



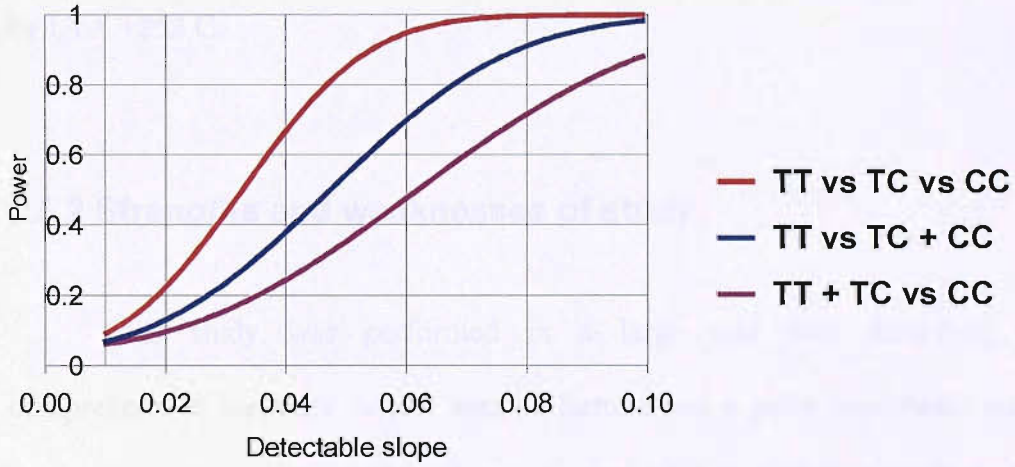
The Power as a function of detectable slope was calculated for BMI and (diastolic and systolic) blood pressure using genotypes of LGALS2 rs7291467 in an additive model as an example.

Graph 16: Power calculation for minor allele frequency.



The Power as a function of detectable slope was calculated for minor allele frequency using BMI in an additive model as an example.

Graph 17: Power calculation for genetic model.



The Power as a function of detectable slope was calculated for BMI in three different genetic models using genotypes of *LGALS2* rs7291467 as an example.

5.5 Discussion

5.5.1 Principal findings

LGALS2 is associated with lower levels of insulin, glucose and HOMA score, such that the C allele had the lowest level. The strongest association for insulin was obtained in a recessive genetic model.

LTA +252 C and a haplotype harbouring + 252 C are associated with lower BMI and triglycerides. Two haplotypes harbouring the *LTA* +80 G allele are associated with lower blood pressure. Analysed by itself a haplotype harbouring *TNF* -308 A is associated with lower BMI, but analysed together with *LTA* this association is explained by LTA +252 C.

5.5.2 Strengths and weaknesses of study

This study was performed in a large and well described cohort. A comprehensive literature review was performed and a prior hypothesis was founded. There is strong experimental evidence for that LTA and TNF through their receptors can influence pathways, which regulates metabolic syndrome traits. The interaction between LGALS2 and LTA is plausible, but still unconfirmed. There is convincing evidence that at least *LTA* +80 and *LTA* T26N are functional (see next section). The size of the genetic effect is as expected for poly genetic traits, and clinical significant. The p values presented in the results section are not corrected for multiple testing. The metabolic syndrome traits are strongly correlated and it is therefore difficult to make a

correction. If the metabolic syndrome is seen as having three etiological categories (obesity related, insulin resistance related and other), as has been proposed [2] a Bonferroni correction of three could be applied. The results obtained for *LGALS2* would still be statistical significant and so would the association between *LTA* +252 and BMI. In the haplotype analysis the association between *LTA* *2 and blood pressure would no longer be statistical significant. None of the results obtained in the descriptive analysis are statistical significant because they not only have to be corrected for the number of traits analysed but also the number of haplotypes.

The metabolic traits are correlated and the expectation could be that if an association is true it would be detected for most of the traits. We only found an association for a subset of the traits. Some traits are strongly correlated; the Pearson correlation for systolic and diastolic blood pressure is 0.699 but we only observed an association for diastolic blood pressure. This can be explained by the size of the measurement, a decrease in systolic blood pressure from 148mmHG to 147mmHG is not statistically significant whereas a decrease in diastolic blood pressure from 79mmHG to 78mmHG is. The Pearson correlation for BMI and WHR is 0.6, but we only saw an association with BMI. Both BMI and WHR are indirect measurement of fat mass, but WHR may not be a good measurement for women because waist and hip circumference expands equally.

Given the functional interaction between LTA and galectin-2 it would be logical if they showed the same associations. In our study *LTA* is associated with BMI, triglycerides and blood pressure, and *LGALS2* is associated with HOMA score, glucose and insulin. *LTA* has previously been associated with HOMA score. The lack of an

accordance between associations is a weakness and do question either our results or the functional relation between LTA and galectin-2. The association between *LTA* and HOMA score was observed in Japanese men and the lack of this association in British women could be explained by the difference in gender and environmental interactions. LTA exists in two forms a soluble homo tetramer of LTA molecules (LTA_3) and a membrane bound hetero tetramer of one LTA and two LTB molecules (LTA_1LTB_2). Galictin-2 would only regulate levels of the soluble form. It is therefore possible that the allelic variation in *LTA* can influence all the metabolic traits depending on environmental interactions, while the allelic variation in *LGALS2* only can influence glucose and insulin. If the action of galectin-2 on LTA_3 levels is stronger than the action of the allelic variation in *LTA* itself, it explains why we only see an association with glucose and insulin for *LGALS2*.

This study does not include all possible variation in *LGALS2* or *LTA*. We have chosen to genotype only SNP rs7291467 in *LGALS2*. This SNP was identified in a gene scan as having the strongest association with myocardial infarction and has been shown to exert functional effects. HapMap data shows nearly perfect LD (all pair wise r^2 close to 1) for the haplotype block containing rs7291467 (Figure 3). Thus, genotyping more common SNPs is this block would provide minimal additional information. For *LTA* and *TNF* we focused on functional SNPs, and the three main haplotypes were covered. Little LD exist within the *LTA* gene and 8 SNPs is needed to cover this relative short gene (Figure 2) (29). Most of the additional haplotypes would though be under 5% and therefore to small to be included in an association study. No association was observed

with myocardial infarction, because of the imprecise classification of MI events and the low number of events, it is not possible to make any conclusions.

5.5.3 Related literature

The background for the study of *LTA* and *LGALS2* is two major papers by a Japanese group (34, 55), which found an association between myocardial infarction and SNPs in these genes. In this section important point about these papers and other relevant literature is discussed.

LTA was identified in a genome wide study; this could suggest that it was the gene with the strongest association. The SNPs were chosen randomly from a database containing SNPs within genes but some genes and regions have been studied more than others and they will therefore be represented with more SNPs in the database. It is therefore no surprise that *LTA* which is in a well studied region is found to be associated but this does not mean that it is the strongest association. The association between *LTA* +252 is strong and not likely to be a type 1 error, if the association is not true it is most likely because of stratification or co-foundering environmental factors. They do test for stratification but only in a small sample, which do not represent the large study sample. There is no correction for known risk factors but they did not found any association between *LTA* and risk factors. The functional claim for +252 on transcription levels is weak and contradicted by Knight (98). The functional claim on T26N is stronger and it is backed by other literature that shows that TNF can induce VCAM-1 through TNFR1, which LTA also binds to.

The Japanese study by Yamada (93) could not confirm the association between *LTA* and myocardial infarction. The cohort used in the Yamada study is of the same size as the one used in the Ozaki study, but has more phenotypes. The most notable difference between the two studies is the frequency of the GG genotype in the two control cohorts. The frequency of the GG genotype in the Ozaki control group is 11.5% whereas it is 17% in the Yamada control group. The expectation would be that the frequency would be the same in the two control groups. As a comparison, the frequency in the third Japanese study is 13.5%, in BWHHS it is 14% and another large European study found a frequency of 13% (96). In the Yamada study a large group of people with hypertension were included in the control group. It is possible that by actively selecting people with hypertension they have selected for a subpopulation with a higher GG frequency and thereby masked the association. The important of the family study in this context is that there is no possibility for stratification (94).

The LGALS2 paper by Ozaki is central to the investigation of genes that predispose people to myocardial infarction. In order to understand this consider the following: First an association between a gene A and a phenotype is detected and then a functional interaction between this gene A and a gene B is established and here after gene B shows a similar association with the same phenotype. The probability to get this sequence of evidence by chance is so small that one can consider it proof of the associations.

The functional interaction between LTA and galectin-2 and the association of *LGALS2* with myocardial infarction, is therefore, if true, a proof of the association between SNPs in the two genes *LTA* and *LGALS2* and myocardial infarction. Several

lines of evidence supports the interaction between LTA and galectin-2 and the use of galectin-1 as a control in determining the effect of galectin-2 expression on LTA levels ensure specificity. There are several reasons for concern. The authors do not make it clear, how many candidate genes the "two hybrid" system identified, how many of these were tested and how many were finally confirmed binding to LTA. They do not tell whether the binding was confirmed before any genotyping were done. If a 100 candidate genes were identified in the two hybrid system, and then one or more SNPs were genotyped in each gene and thereafter genes with positive association had their binding tested, when the claim is less strong. Since galectins have sticky ends, which will bind to many other proteins! It would therefore be of interest to know if galectin-1 also binds to LTA.

LGALS2 rs7291467 was tested in a luciferase assay, but there is no explanation on how an SNP in the middle of a large intron could affect expression (Figure 11). More over the SNP has been cloned into a vector construct, which is designed to test promoter activity (112). It is therefore uncertain if the observed difference in expression levels is biologically relevant. Further to this, our results do not support an additive model but suggest a recessive model. It is difficult to see how a relative small difference in levels of galectin-2 should lead to a recessive model. A recessive model is most often seen with an SNP that have an on/off effect; which is most likely to be a coding SNP.

Prior to the study conducted in this thesis only one paper had been published documenting an association between *LTA* and metabolic system traits (61). The study is small (200 men) but the association is with HOMA score the most direct indicator of insulin resistance and the p value has been Bonferroni corrected. Therefore this study

carries as much weight as the study by Hamid (96), which found an association with WHR but in the other direction.

5.5.4 Important differences in results

It is well established that the traits of the metabolic syndrome are risk factors for myocardial infarction. It is therefore a paradox (see figure 18 and 19) that genotypes that are known to either increase or reduce the risk of myocardial infarction have the opposite effect on traits of the metabolic syndrome. *LTA* +252 C is a risk factor for myocardial infarction but it does also lower BMI and HOMA score. The T allele of the *LGALS2* SNP lowers the risk for myocardial infarction but in our study the T allele has the highest insulin level. The most likely explanation is that the soluble form of LTA (LTA_3) stimulates the expression of VCAM1 in the arterial wall. VCAM1 being responsible for the recruitment of macrophages, an important step in the progression towards myocardial infarction. Both the soluble and the membrane bound form of LTA may interact with the metabolic system as TNF is known to do, and thereby alter metabolic traits. TNF is known to be an apotikine and works locally in apocytes. LTA has not been observed in the apocyte tissue. It is therefore most likely that LTA's action is indirect. A high level of systemic LTA could lead to down regulation of TNF. LTA can activate the LTB receptor which is known to be part of a negative feedback loop (64). It is also known that TNF receptors are cleaved of the membrane to bind circulating TNF; a high level of LTA could therefore mean that more TNF receptors were cleaved and this would lead to less TNF signalling.

5.5.5 Conclusion

LTA and LGALS2 are part of pathways that can affect traits of the metabolic syndrome and allelic variations in these two genes are associated with a decrease in metabolic syndrome traits. This means that some individuals could have a reduced risk for developing cardiovascular diseases because of their genetic makeup, although our study did not have the power to detect an association with myocardial infarction. The observed effect of *TNF* promoter SNPs on BMI can be explained by variation in *LTA*, this is strong evidence against that *TNF* promoter SNPs are functional. Our study provides the first evidence for that *LTA* +80 can affect the metabolic syndrome and supports its presumed functionality.

Figure 18: Illustration of *LGALS2* rs7291467; allele frequency and associations.

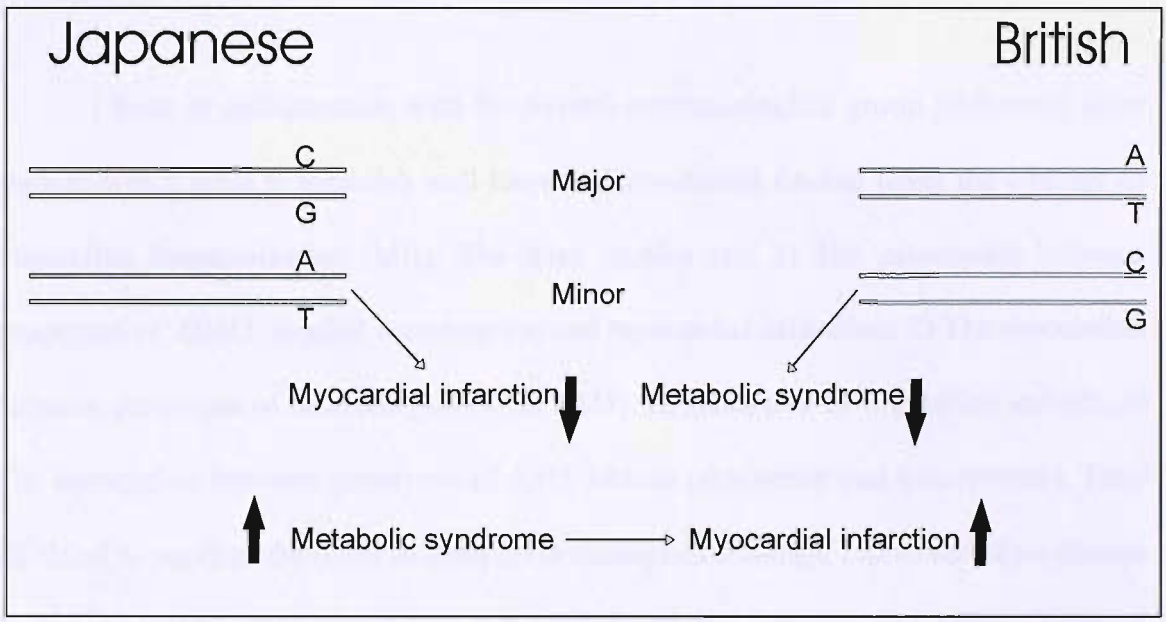


Figure 19: LGALS2, LTA interactions.

Biological Explanation

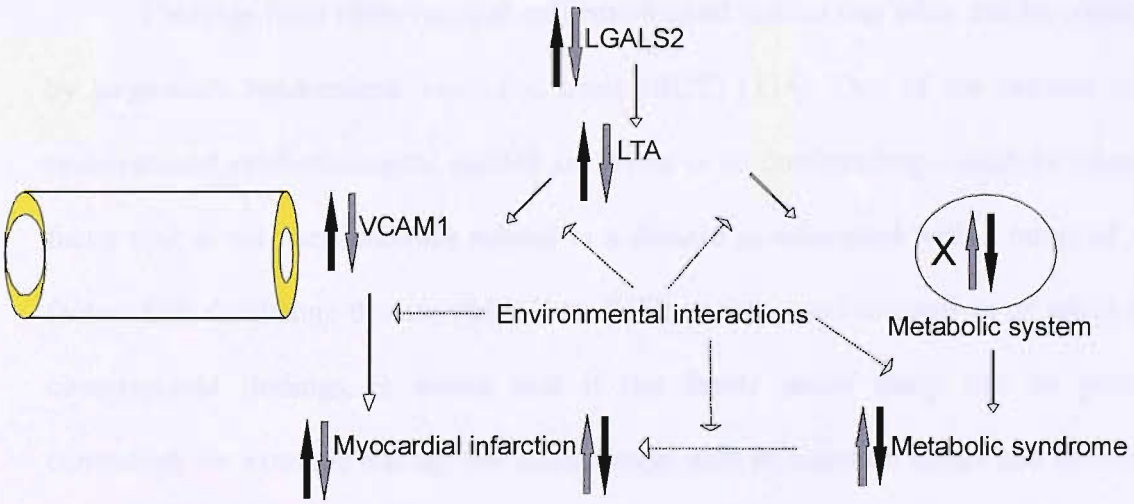


Illustration of interactions between LGALS2, LTA and genes in the arterial cell wall or the metabolic system, which explain how different disease outcomes are possible. Large black or grey arrows denote regulation. Small arrows denote interaction. X denotes action of TNF as illustrated in figure 11.

Chapter 6: Nutritional genetics and Mendelian randomization

6.1 Introduction

I have in collaboration with the Bristol epidemiological group performed three studies, which aims to establish well known observational finding using the concept of Mendelian Randomization (MR). The three studies are: 1) The association between genotypes of *ADH3*, alcohol consumption and myocardial infarction; 2) The association between genotypes of taste receptors (*TAS2R38*), consumption of vegetables and MI; 3) The association between genotypes of *LAC*, lactase persistence and osteoporosis. They all failed to confirm the observational epidemiological findings. I have therefore chosen to include only the first study as an example of MR, the taste receptor study is covered by a recent publication(113).

Findings from observational epidemiological studies can often not be confirmed by large-scale randomized controlled trials (RCT) (114). One of the reasons is that observational epidemiological studies are prone to be confounding- which is when one factor that is not itself causally related to a disease is associated with a range of other factors that do change disease risk (114). RCTs can be used to confirm or reject some observational findings, it works best if the factor under study can be precisely controlled; for example a drug. But other things such as nutrition intake can be difficult to test in RCT (115).

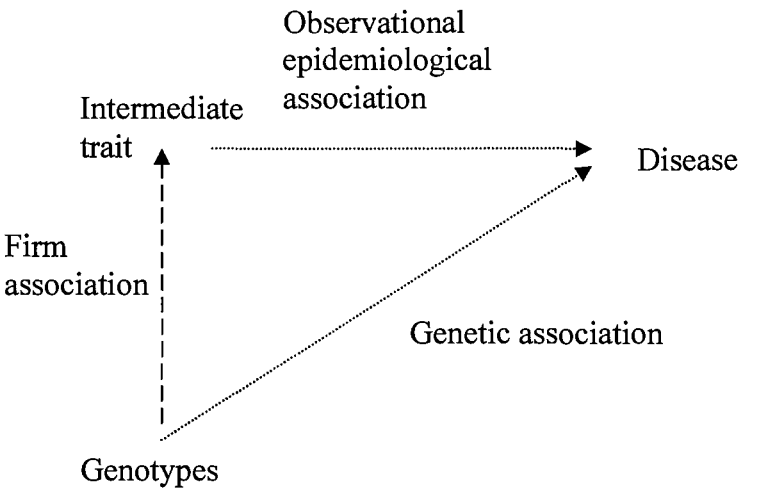
Mendelian randomization is a concept developed to confirm or reject observational findings using genetics. This concept has been described in detail in two

large reviews (114, 116). Figure 20 gives a schematic view of how MR works. An observational epidemiological association between an intermediate trait and a disease can be tested, if a firm association between a gene and the intermediate trait exist, such that the gene can alter the intermediate trait. An association between genotypes of the gene and the disease is a proof of the observational epidemiological finding. A fundamental assumption is that there is no association between genotypes of the gene and social, economic, environmental or behavioural confounders (116).

The observational association between moderate alcohol intake and a decreased risk for myocardial infarction can be tested using Mendelian randomization. Alcohol dehydrogenase type 1C (ADH1C) has two forms, which differ in their metabolic rate. The fast metabolising forms oxidizes alcohol to acetaldehylde 2.5 times faster than the slow form. The difference in metabolic rate is due to two SNPs Arg272Gly and Ile349Val, which are in strong LD. It is expected that the individuals with the slow metabolising haplotype will have a greater beneficial effect from the alcohol and therefore a lower risk of MI (116). Hines *et al* found that individuals homozygote for the slow metabolising genotype, who drank more than 1 drink per day had a lower risk for myocardial infarction relative to individuals homozygous for the fast metabolising genotype who drank fewer than 1 drink per week. They also found an association with HDL levels, such that the slow metabolising genotype had the highest level. The assumption is that the beneficial effect of alcohol is due to its effect on HDL levels (28). Another gene that has been used to test the association between alcohol intake and MI is *ALDH2*, which encodes aldehyde dehydrogenase. ALDH2 converts acetaldehyde, which is toxic, to acetate. A SNP, Gly487Lys, change the metabolic rate. Japanese men, which are homozygous for the slow allele, consume less alcohol, have

lower HDL levels and have higher risk for MI. This supports the observational association between moderate alcohol intake and a decreased risk for myocardial infarction.

Figure 20: Illustration of Mendelian randomization



6.2 Methods

Details for BWHHS are the same as in chapter 5. The Caerphilly study consists of 2513 men, which were recruited between 1979-83 (age 45 to 59). The baseline study is described in (117). The men were examined for myocardial infarction every four to five years over a 20 year follow up period. Details of the assessment of MI are given in(117). DNA extraction and genotyping were done as described in section 5.3.4 (reaction condition in table 17, primer and probes 18). The statistical analysis was performed by collaborators at social medicine in Bristol.

Table 17: PCR conditions.

PCR Mix	Lac 22018	ADH2 272	PTC 49	PTC262
Taq buffer	1x	1x	1x	1x
dNTP's	0.2mM	0.2mM	0.2mM	0.2mM
$MgCL_2$	1,5mM	1,5mM	1,5mM	1,5mM
Right Primer	0.5uM	0.5uM	0.5uM	0.5uM
Left Primer	0.1uM	0.1uM	0.1uM	0.1uM
Flu Probe	0.2uM	0.2uM	0.2uM	0.2uM
Dabsyl Probe	0.2uM	0.2uM	0.2uM	0.2uM
TaqU/uL	0.001	0.001	0.001	0.005
PCR Reaction				
94c	2min	2min	2min	2min
94c	20sec	20sec	20sec	20sec
Anneal 30sec	60	60	62	62
72C	2 Min	2 Min	1 Min	1 Min
Cycles	49	49	100	100
72C	2Min	2Min	2Min	2Min

Taq buffer: 10mM Thris-HCL, 10Mm KCL, pH 8.3

Table 18: Primers and probes

	ADH3 272	LAC22018	PTC 49	PTC 262
Flu probe	TCAAGCCGACCGATGA	caccgcgcccagct	CTCAGTGCCTGCCTCT	GAAGGCAGCACAGGATG
Dabsyl probe	GGGCATGTCACGGATCATACCATGG	tcggcttcccaaagtactgggacaaaggtgtga	GAGACACAGCAGCACACAATCACTGT	GCCACAGAATCAGTAGGGGCACAGAG
LEFT primer	TCCTCCAGGTTGCAGAGGCAGA	gcctcttgagtagctgggaccacaagca	GCCAGAGGTTGGCTTGGTTTGCA	TGCCCAGAGGGACAAGCTGCCATT
RIGHT primer	CCCATTCAGGAAGTGCTAAAGGAAATGA	GGCTGGAGCTTTGATGTTGGCTGA	CCTGGAGTTTGCAGTGGGGTTTCTGACCA	TGGGAAGGCACATGAGGACAATGAAGG

6.3 Results

Genotypes of ADH1C are associated with alcohol consumption when calculated as mean units a week among drinkers, homozygotes for the slow metabolising allele had the highest alcohol intake. The slow metabolising allele associates with high gamma-GT and a status as current smoker but only in women. There was no association with CHD, MI, metabolic syndrome traits or social class (Table 19 and table 1in appendix IV).

Alcohol intake was associated with CHD in women; individuals with the highest alcohol intake had the lowest risk (Table 20 and table 2 in appendix IV). No association was observed with MI either in women or in men (Table 20 and table 2 in appendix IV). Alcohol intake was also associated with metabolic syndrome traits; individuals with the highest alcohol intake had the lowest risk profile. The association was more pronounced in women than in men. Higher alcohol intake was also associated with higher social class in women, but not in men (Table 2 in appendix IV). A trend upper sit to the association observed by Hines was found if alcohol consumption above 3 ½ unit per week was stratified by genotype (Table 3 in appendix IV).

Table 19. *ADH1C* variants and CHD.

	BWHHS (Women aged 60 – 79 years) Percent or mean (95% CI) by *ADH1C* variant N = 3234				Caerphily (Men aged 47 to 67 years) Percent or mean (95% CI) by *ADH1C* variant N = 1313			
Variants	$\gamma^1\gamma^1$ N = 1095	$\gamma^1\gamma^2$ N = 1604	$\gamma^2\gamma^2$ N = 535	P[$]	$\gamma^1\gamma^1$ N = 462	$\gamma^1\gamma^2$ N = 612	$\gamma^2\gamma^2$ N =239	P[$]
Risk of coronary heart disease and myocardial infarction								
CHD(%)	19.1 (16.8,21.5)	21.8 (19.8,23.8)	22.1 (18.7,25.8)	0.10	N/A	N/A	N/A	N/A
MI only(%)	3.1 (2.2,4.3)	3.4 (2.6,4.4)	4.3 (2.9,6.4)	0.24	13.4 (10.3,16.5	11.1 (8.6,13.6)	14.6 (10.1,19.1)	0.91

[$]P-values are for linear trend across groups (1df). This table is an extract from table 1 in appendix IV.

Table 20. Alcohol consumption and CHD.

	BWHHS (Women aged 60 – 79 years) Percent or mean (95% CI) by Current alcohol consumption N = 2716					Caerphily (Men aged 47 to 67 years) Percent or mean (95% CI) by Current alcohol consumption N = 1398				
variants	Life Long abstainers N = 631	Lowest 1/3 N = 1095 Rage 0.5-2 Unit per week	Second 1/3 N = 1604 Rage 3-4 Unit per week	Highest 1/3 N = 535 Rage 5-42 Unit per week	P[$]	Life Long abstainers N = 631	Lowest 1/3 N = 470 Rage 0.2-4 Unit per week	Second 1/3 N = 486 Rage 5-18 Unit per week	Highest 1/3 N = 366 Rage 19-176 Unit per week	P[$]
Risk of coronary heart disease and myocardial infarction										
CHD(%)	24.4 (21.2,27.9)	20.3 (18.0,22.8)	18.5 (14.7,23.0)	16.4 (13.9,19.4)	<0.001	N/A	N/A	N/A	N/A	
MI only(%)	4.1 (2.8,6.0)	4.3 (3.2,5.7)	1.5 (0.6,3.5)	2.6 (0.6,3.5)	0.05	19.7 (10.7,28.8)	14.7 (11.5,17.9)	10.7 (7.9,13.5)	12.6 (9.2,16.0)	0.08

[$]P-values are for linear trend across groups (1df). This table is an extract from table 2 in appendix IV

6.4 Discussion

6.4.1Principal findings

We observed an association between increased alcohol intake and a decrease in risk for CHD, but we could not replicate the previous observed association between genotypes of *ADH1C* and levels of HDL or risk for MI.

6.4.2 Strengths and weaknesses of study

This study was performed in two large cohorts of women and men, with many phenotypes that make it possible to check for co-founders. The molecular mechanism that link ADH1C to alcohol levels is well understood and the SNPs are functional. The precision with which alcohol consumption is reported may be a problem. The difference between fast and slow metabolising is only a factor 2. The imprecision in the reported alcohol consumption could be of the same magnitude because people reports what they think is acceptable. The definition of the cardiovascular phenotypes is also different in the two cohorts. The women's study has loose criteria for CHD and MI while the men's study uses the WHO criteria. This makes it difficult to compare the results from the two cohorts and to compare with other studies.

6.4.3 Related literature

In interpreting the results presented by Hines (28) it is important to note that they use the fast metabolising homozygotes with the lowest alcohol intake as reference group. To see the association between genotypes of *ADH1C* and MI, one needs to compare the relative risk within one strata of alcohol consumption. The confidence intervals for homozygote major allele, heterozygotes and homozygote minor allele all

overlap in each of the three alcohol consumption strata, which could be interpreted as there is no association.

6.4.4 Important differences in results

Two different markers have been used to test the association between alcohol consumption and MI; *ADH1C* Arg272Gly and *ADLH2* Gly487Lys. The important difference between these two markers is that the first do not affect alcohol consumption while the second have a strong effect. This means that the association between *ADLH2* Gly487Lys and MI is not free of co-founding, this could happen if less alcohol intake leads to other lifestyle changes that were associated with MI. Even if there is no obvious co-founders the SNP genotyped could be in LD with SNPs in other genes that have the possibility to co-founder. A test of association with possible co-founders is therefore always necessary.

6.4.5 Conclusion

We could not confirm the association between moderate alcohol intake and a decreased risk for myocardial infarction using Mendelian randomization.

7 Summary and future experiments

This study has been centred on the common variant common disease hypothesis, involving both human genetic association studies and informatics.

The work in this thesis supports the idea that the adipose tissue is central to the metabolic syndrome and coronary heart disease. The observed association between the *LGALS2-LTA* pathway and the metabolic syndrome is in agreement with molecular experiments suggesting TNF as a cause of insulin resistance because the TNF receptor family is the most likely mediator between LTA and the adipose tissue.

LGALS2 and *LTA* are in the periphery of the pathway network, which connect them to the phenotype, and the association is therefore prone to interference by other genes and environmental factors. The association could therefore be difficult to replicate unless it was done in a cohort with the exact same properties as the BWHHS. A better verification of the results would therefore be to test the hypothesis "If these associations are true, then there must be genes in downstream pathways that will show a similar association; provided that there are functional variations in these genes". Genes involved in the regulation, transport and biosynthesis of fatty acids and triglycerides are good candidates. Obvious candidates are *LPL* and *aP2;* other candidates could be genes from the NF-κB pathway for example *MyD88*, *IRAK 1*, *TIRAP*, which all have several SNPs.

One view of arteriolosclerosis is as a mechanical disease pushed forward by high levels of cholesterol and blood pressure and it therefore make sense to look at biochemical processes inside the adipose tissue instead of the arterial plaque. Understanding the role of the adipose tissue will enable us to prevent the risk factors

that leads to CHD, by change in early lifestyle based upon our genetic make up. This would be better than treatment for CHD.

Is genome wide association studies the solution to the common variant common disease hypothesis? The drive behind high through put genotyping is the expectation that genome wide association studies will uncover most genes underlying common diseases. Phenotype collection have not had the same attention, and it is not clear if it will be possible to collect high quality and detailed phenotypes for the large sample sizes needed for genome wide association studies. My expectation is that genome wide association studies will be most successful in finding strong but unexpected associations.

Another question that has not yet been answered is how to communicate the many associations found in large studies, positive as well as negative. NCBI have already made a database to collect information from large scale association studies(118), but it is not certain if researchers will report their data. Researchers are most likely to maximize publications and citations. Even though an entry in a database can be claimed as a publication, it will not generate any citations because other researchers are unlikely to site individual papers when they extract information from the database. It is therefore most likely that data will be kept private until they can be published or used in grant applications, or published on private website such as canvas, which are more likely to be cited.

An alternative to whole genome studies would be small samples with high quality phenotypes, stratified to minimize environmental interference. For example to study the genetics of the metabolic syndrome traits one should sample a cohort of women, which a similar level of physical activity (aerobic class three times a week) and

diet. This strategy is most useful for intermediate traits and not disease endpoints. In my view we need to understand the genetic background for all intermediate traits before we can find the genetic background for disease endpoints.

Appendix I

Index of abstracts

Index of Abstracts per Gene

Appendix I

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	SUM
2	**Hypertension**													
3	ACE	1	2	5	4	11	10	10	13	22	19	12	3	112
4	AGT	0	0	0	1	1	1	1	7	8	7	8	0	34
5	GNB3	0	0	0	0	0	0	0	1	3	3	3	0	10
6	INSR	0	0	0	0	0	0	1	1	0	2	1	0	5
7	NPPA	0	0	0	0	0	0	1	1	0	2	1	0	5
8	CYP11B2	0	0	0	0	0	0	0	2	2	4	4	0	12
9	**Embolism**													
10	MTHFR	0	0	0	0	0	0	1	0	1	1	1	1	5
11	Proc	1	0	0	0	0	0	1	0	1	0	1	1	5
12	**Insulin resistance**													
13	INSR	0	0	2	1	0	0	0	1	0	0	3	1	8
14	TNF	0	0	0	0	0	0	0	1	4	1	4	1	11
15	ACE	0	0	0	1	0	0	0	1	0	3	1	0	6
16	**Hyperlpidemia**													
17	ACE	0	0	0	1	2	0	1	0	2	1	0	0	7
18	APOA1	0	0	1	0	0	0	0	1	0	1	0	1	4
19	APOE	0	4	2	3	3	1	0	1	4	1	3	1	23
20	CETP	0	0	1	0	0	0	1	1	1	1	1	1	7
21	LDLR	2	3	0	1	4	0	2	0	1	0	1	0	14
22	LPL	1	0	3	2	2	1	2	1	4	3	0	0	19
23	APOC3	0	0	1	0	1	1	1	1	0	0	1	0	6
24	**HDL-C**													
25	APOA1	0	0	1	0	2	0	0	1	1	0	1	0	6
26	CETP	0	0	3	0	1	2	2	1	2	2	6	2	21
27	**Homocysteine**													
28	MTHFR	0	0	0	0	1	1	4	9	2	9	9	7	42
29	**Hyperhomocysteinemia**													
30	MTHFR	0	0	0	0	0	0	0	1	1	4	3	1	10

Index of Abstracts per Gene

Appendix I

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	SUM
31	**Cholesterol**													
32	APOA1	0	0	1	0	2	0	0	1	2	1	1	0	8
33	APOB	0	0	0	0	0	0	1	0	1	0	0	0	2
34	LPL	0	0	1	0	0	0	1	0	2	1	0	1	6
35	CETP	0	1	5	1	1	2	2	0	4	3	7	3	29
36	**Obesity**													
37	LPL	0	0	0	2	0	3	0	0	2	2	0	0	9
38	UCP1	0	0	1	0	1	0	1	0	4	1	1	0	9
39	INSR	0	0	0	0	1	0	1	0	1	0	1	2	6
40	LEP	0	0	0	0	0	0	3	2	3	5	2	0	15
41	LEPR	0	0	0	0	0	0	1	0	1	5	1	0	8
42	PPARP	0	0	0	0	0	0	1	0	0	2	3	1	7
43	TNF	0	0	0	0	0	0	1	1	1	2	1	1	7
44	**Cerebrovascular**													
45	ACE	0	0	1	2	6	3	6	4	8	5	8	1	44
46	AGT	0	0	0	0	0	0	1	1	0	3	0	0	5
47	PLG	0	0	0	0	1	1	0	1	1	5	2	0	11
48	APOE	0	0	1	1	1	5	3	3	5	6	8	4	37
49	HLA-DRB	0	0	0	0	0	0	0	0	1	2	3	0	6
50	IL6	0	0	0	0	0	0	0	0	1	0	3	3	7
51	LPL	0	0	0	0	0	1	0	0	0	2	1	1	5
52	MTHFR	0	0	0	0	0	1	0	0	0	2	1	1	5
53	PON1	0	0	0	0	0	0	0	0	2	3	2	1	8
54	SERPINE1	0	0	0	0	0	1	0	0	1	3	1	0	6
55	SERPINE1/PLG	0	0	0	0	1	2	0	1	2	8	3	0	17
56	**CoronaryArteriosclerosis**													
57	MTHFR	0	0	0	0	0	0	0	1	1	2	3	0	7
58	ACE	0	0	0	1	0	1	3	0	1	2	2	4	14
59	LPL	0	0	0	0	0	0	1	0	0	1	0	3	5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	SUM
60	**CoronaryThrombosis**													
61	none													
62	**LDL-C**													
63	AOPE	0	1	1	0	0	0	0	0	1	1	1	0	5
64	**Myocardial Infarction**													
65	CD14	0	0	0	0	0	0	0	2	1	1	1	0	5
66	LPL	1	0	0	1	0	0	0	1	1	1	0	0	5
67	PLG	0	0	0	1	0	1	4	1	2	0	2	1	12
68	RNF130	0	0	0	0	0	0	0	2	2	1	0	0	5
69	SERPINE1	0	0	0	1	0	0	1	2	2	0	0	0	6
70	SERPINE1/PLG	0	0	0	2	0	1	4	3	4	1	2	1	18
71	ACE	0	2	5	5	4	6	7	1	6	4	3	1	44
72	APOE	1	0	0	0	0	1	1	0	3	1	2	1	10
73	MTHFR	0	0	0	0	1	1	0	1	0	1	0	0	4
74	**Myocardial Ischemia**													
75	APOA1	0	0	0	0	2	1	0	0	0	0	1	1	5
76	APOC3	0	0	0	0	0	0	0	0	0	2	0	1	3
77	AR	0	0	0	0	0	0	0	1	1	2	2	0	6
78	CYP11B2	0	0	0	0	0	0	0	1	2	1	1	0	5
79	DNCH1	0	0	0	0	0	0	1	3	0	1	1	0	6
80	MTHFR	0	0	0	0	1	2	1	3	2	5	7	0	21
81	PLG	0	0	0	1	1	2	6	2	3	2	1	1	19
82	PON1	0	0	0	0	1	0	4	0	5	4	5	1	20
83	SERPINE1	0	0	0	1	0	0	2	3	4	1	3	0	14
84	SERPINE1/PLG	0	0	0	2	1	2	8	5	7	3	4	1	33
85	THBD	0	0	0	0	0	1	0	0	2	1	1	0	5
86	ACE	0	2	8	9	9	18	16	12	20	9	13	6	122
87	AGTR1	0	0	1	0	0	1	1	2	1	1	1	0	8
88	APOB	0	1	0	0	1	1	2	0	0	2	0	0	7
89	APOE	1	0	2	1	4	3	3	7	8	6	5	2	42
90	CD14	0	0	0	0	0	0	0	2	2	1	2	0	7
91	CETP	0	0	1	1	0	1	1	0	2	4	3	0	13
92	IL6	0	0	0	0	0	0	0	0	1	3	2	0	6
93	LPL	1	0	0	2	3	0	1	3	4	4	2	3	23
94	TNF	0	0	0	0	0	0	2	1	1	4	1	1	10

Index of Abstracts per Gene

Appendix I

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	SUM
95	**Stroke**													
96	PLG	0	0	0	0	0	0	0	0	0	4	2	0	6
97	ACE	0	0	1	0	0	1	1	2	5	2	4	0	16
98	APOE	0	0	0	0	0	0	0	1	0	2	3	2	8
99	MTHFR	0	0	0	0	0	0	0	2	0	2	3	1	8
100	**Trombosis**													
101	MTHFR	0	0	0	0	0	0	3	1	5	1	1	2	13
102	PLG	0	0	1	0	0	1	2	0	4	1	4	0	13
103	PROC	2	0	1	1	2	1	1	2	2	0	1	2	15
104	SERPINC1	0	0	0	0	1	0	1	0	0	0	0	0	2
105	SERPINE1	0	0	0	0	0	1	2	1	3	1	4	0	12
106	SERPINE1/PLG	0	0	1	0	0	2	4	1	7	2	8	0	25
107	**VenousThrombosis**													
108	MTHFR	0	0	0	0	0	0	2	0	2	1	1	2	8
109	PROC	1	0	1	0	1	0	0	1	1	0	1	2	8
110	PROS1	0	0	0	0	0	0	1	1	0	2	0	0	4
111	PLG	0	0	0	0	0	0	2	0	2	0	1	0	5
112	SERPINE1	0	0	0	0	0	0	2	0	2	0	1	0	5
113	SERPINE1/PLG	0	0	0	0	0	0	4	0	4	0	2	0	10
114	Manual													
115	CETP													
116	NPPA pos													
117	NPPA neg													
118	**CardioInflam**													
119	TNF	0	0	0	0	0	0	1	1	2	4	1	5	14
120	IL6	0	0	0	0	0	0	0	0	2	0	3	1	6

Appendix II

Source code

```
#!/usr/bin/perl
#This programme find genename in pubmed ABS
#use strict,
%data=();
$hashkey = '';
open(GENENAMEFILE,$ARGV[0]) or die "can't open GENENAMEFILE";
chomp (@genenames = <GENENAMEFILE>);
close GENENAMEFILE;
open(ABSTRACTFILE,$ARGV[1]) or die "can't open ABSTRACTFILE";
do "termpvalue.dat" or die "can't read probabilitys";
while(defined ($line = <ABSTRACTFILE>)){
    if($line=~ m/^PMID/){
        $absline = '';}
    if($line=~ m/^PT\s+-\s+Review/){
        next;}
    if($line=~ m/^PMID-\s+([0-9]+)/){
        $absline .= "$1~@~";}
    if($line=~ m/^DP.+([0-9]{4})/){
        $pubdate = $1;}
    if(!($line=~m/^MH\s/) and $in_MH ){
        $in_MH = 0;
        }
    if(($line=~ m/^MH\s/) and !$in_MH){
        $in_MH = 1;
        $absline .= "~@~";}
    if($in_MH){
        chomp $line;
        if($line=~ m/^MH\s+-\s+([\w\s\,]+\/)/){
        $absline .= ":$1";}}
    if(!($line=~ m/^\s/)){
        $in_TI = 0;
        $in_AB = 0;}
    if($line=~ m/^TI/){
        $in_TI = 1;}
    if($line=~ m/^AB/){
        $in_AB = 1}
    if($in_TI or $in_AB){
        chomp $line;
        $absline .= $line;}
    if($line=~ m/^SO/){
        $absline .= "~@~$pubdate";
        push (@abstracts,$absline);}}
close ABSTRACTFILE;
```

```
foreach $genename (@genenames){
    ($hgnc_id,$status,$app_long_name,$app_short_symbol,$previous_short_symbol,$enzyme_id,$location,$aliases,$mgi,$pmid1,$pmid2,$seq,$previ
    if(!$location  or $location eq 'reserved'){
            next;}
    #warn"£££££££££££££££new gene  $app_short_symbol £££££££££££££££££££££££££\n";
    (@long_appname,@short_appsymbol,@short_prisymbol,@short_alias,@long_priname) =();
    #warn "approven:$app_short_symbol priname:$previous_short_symbol alias:$aliases prilong:$previous_long_name\n";
    if($app_long_name){
    @long_appname = &name_handel($app_long_name);}
    if($app_short_symbol){
    @short_appsymbol = &symbol_handel($app_short_symbol);}
    if($previous_short_symbol){
    @short_prisymbol = &symbol_handel($previous_short_symbol);}
    if($aliases){
    @short_alias = &symbol_handel($aliases);}
    if($previous_long_name){
    @long_priname = &name_handel($previous_long_name);}
    if($app_short_symbol){
    $proteinname = &trans_pro($app_short_symbol);}
    #warn "proteinename:$proteinname\n";
    foreach $pre_abstract (@abstracts){
        ($pmid,$abstract,$mh,$year) = split /~@~/,$pre_abstract;
        #warn "################new abstract#### $pmid #######$mh##################################\n";
        ($longname_found,$partofname_found,$symbol,$aliassymbol,$proteine) = 0;
        @markers = ();
        foreach $longappname (@long_appname){
            ($name,$length) = split /:/, $longappname;
            #warn "now looking for app_long_name:$name\n";
            if($length > 10){
                $longname_found = &name10plus($name,$abstract);}
            else{   $longname_found = &symbol2or3($name,$abstract);
                #warn "longname_found $longname_found";
                }}
        if(!$longname_found){

            foreach $longpriname (@long_priname){
                ($name,$length) = split /:/, $longpriname;
                #warn "now looking for previous_long_name:$name\n";
                if($length > 10){
                    $longname_found = &name10plus($name,$abstract);}
                else{   $longname_found = &symbol2or3($name,$abstract);}}}


```

```
87
88          foreach $shortappsymbol (@short_appsymbol){
89              ($name,$length) = split /:/,$shortappsymbol;
90              #warn "now looking for app_short_symbol:$name\n";
91              if($length >4){
92                  $symbol = &symbol4plus($name,$abstract);}
93              else{ $symbol = &symbol2or3($name,$abstract);}
94              }
95
96          if(!$symbol){
97
98              foreach $shortprisymbol (@short_prisymbol){
99                  ($name,$length) = split /:/,$shortprisymbol;
100                 #warn "now looking for previous_short_symbol:$name\n";
101                 if($length >4){
102                     $symbol = &symbol2or3($name,$abstract);}
103                 else{ $symbol = &symbol2or3($name,$abstract);}}}
104         if(!$symbol){
105
106             foreach $shortalias (@short_alias){
107                 ($name,$length) = split /:/,$shortalias;
108                 #warn "now looking for aliases:$name\n";
109                 if($length >4){
110                     $aliassymbol = &symbol2or3($name,$abstract);}
111                 else{ $aliassymbol = &symbol2or3($name,$abstract);}
112                 if($aliassymbol){
113                     last;}}}
114
115         if($symbol or $aliassymbol){
116             foreach $longappname (@long_appname){
117                 ($name,$length) = split /:/, $longappname;
118                 $partofname_found = &find_part($name,$abstract);
119                 if($partofname_found){
120                     last;}}}
121
122         $proteine = &symbol4plus ($proteinname,$abstract);
123         #if($abstract=~/\t([0-9]{4})$/){
124             #$year = $1;
125             #warn "year$year\n";
126             # }
127         #if($abstract=~ m/^([0-9]+)\t/){
128                 #$pmid = $1;}
129         @markers = &find_marker($abstract);
```

```
        if($longname_found or $symbol or $aliassymbol or $proteine){
            #warn "found aliassymbol3:$aliassymbol\n";
            &hash($app_short_symbol,$pmid,$mh,$abstract,$year,$longname_found,$symbol,$aliassymbol,$proteine,$partofname_found,@markers);}
    }}



foreach $gene (keys %data){
    @filter_records=@updated_mainmaker=@singel_recordmaker= ();
    $approven=$prefilter_records=$afterfilter_records= 0;
    foreach $medline_id (keys %{$data{$gene}}){
        $mesh_terms = $data{$gene}{$medline_id}{"MesH"};
        ($namestatus,$symbolstatus,$alliasstatus,$partstatus,$proteinestatus,$put_in_out_hash)=0;
        if($data{$gene}{$medline_id}{"name"}){
            $namestatus = &mikkel($data{$gene}{$medline_id}{"name"});}
        if($data{$gene}{$medline_id}{"symbol"}){
            $symbolstatus= &mikkel($data{$gene}{$medline_id}{"symbol"});}
        if($data{$gene}{$medline_id}{"allias"}){
            $alliasstatus= &mikkel($data{$gene}{$medline_id}{"allias"});}
        if($data{$gene}{$medline_id}{"partname"}){
            $partstatus= $data{$gene}{$medline_id}{"partname"};}
        if($data{$gene}{$medline_id}{"proteine"}){
            $proteinestatus= &mikkel($data{$gene}{$medline_id}{"proteine"})}
        #print "status:$gene $medline_id n:$namestatus s:$symbolstatus a:$alliasstatus p:$partstatus\n";
        $prefilter_records +=1;
        warn "pre: $gene $prefilter_records";
        if($namestatus == 1 or $symbolstatus == 1){
            $put_in_out_hash = 1;}
        elsif($namestatus == 2){
            $put_in_out_hash = 1;}
        elsif(($symbolstatus == 2 or $alliasstatus == 2) && $partstatus >= 1){
            $put_in_out_hash = 1;}
        elsif(($symbolstatus == 3 or $alliasstatus == 3) && $partstatus >= 3){
            $put_in_out_hash = 1;}
        elsif(($symbolstatus == 4 or $alliasstatus == 4) && $partstatus >= 4){
            $put_in_out_hash =1;}
        if(!($put_in_out_hash)){
            print "status_notapproven:$gene $medline_id $mesh_terms\n";}
        if($put_in_out_hash){
            print "status_approven:$gene $medline_id $mesh_terms\n";
            $significant = &probality($data{$gene}{$medline_id}{"abstract"});
```

```
            foreach $singel_recordmarker (@{$data{$gene}{$medline_id} ->{"maker"} }){
                if($singel_recordmarker){
                push (@singel_recordmaker,$singel_recordmarker);}}
            $thisyear = $data{$gene}{$medline_id}{"year"};
            $afterfilter_records++;
            warn "after: $gene $afterfilter_records";
            push(@filter_records,"$thisyear:$significant");}}
    $approven = $afterfilter_records/$prefilter_records;
    warn "$afterfilter_records/$prefilter_records $approven";
    @updated_mainmaker = &totalmaker(@singel_recordmaker);
    if(@filter_records and $approven > 0.5){
        $output_hash{"$gene"}={'1' => [@updated_mainmaker]};
        foreach $recordyear (@filter_records){
            ($refyear,$sigcounter) = split /:/,$recordyear;
            if(exists $output_hash{"$gene"}{"$refyear"}){
                ($yearcounter,$oldsigcount) = split /:/,$output_hash{$gene}{$refyear};
                $yearcounter += 1;
                $oldsigcount += $sigcounter;
                $output_hash{$gene}{$refyear}=("$yearcounter:$oldsigcount");}
            else{$output_hash{$gene}{$refyear} =("1:$sigcounter") ;}}}}
foreach $gene_symbol (sort keys %output_hash){

    print "\n$gene_symbol:";
    foreach $oneyear (sort {$a <=> $b} keys %{$output_hash{$gene_symbol}}){
        if($oneyear=~ m/[0-9]{4}/){
            print "$oneyear:$output_hash{$gene_symbol}{$oneyear}|";}
        elsif($oneyear=~ m/1/){
            @genemaker = @{$output_hash{$gene_symbol} -> {"$oneyear"}};
            print "[";
            foreach $gm (@genemaker){
                print "$gm|";}
            print "]";}}}
###############SUBROUTINES############################
sub trans_pro {
    my($name) = @_;
    $name =~tr/A-Z/a-z/;
    substr($name,-1) =~tr/a-z/A-Z/;
    return $name;}
sub name_handel {
    my($genetitle) = @_;
    $genetitle =~ s/\([\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\)//g;
    $genetitle =~ s/\([\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\)//g;
    $genetitle =~ s/\([\w\s,]*?\([\w\s,]*?\)[\w\s,]*?\)//g;
```

```
    $genetitle =~ s/\(.*?\)//g;
    $genetitle =~ s/\)//g;
    $genetitle =~ s£\++£g;
    my$length = length $genetitle;
    return "$genetitle:$length";}
sub symbol_handel {
    my($symbol)= @_;
    my @symbols = split /,/,$symbol;
    my@finish;
    foreach $short_symbol (@symbols){
        my $length = length $short_symbol;
        push (@finish,"$short_symbol:$length");}
        return @finish;}
sub find_part {
    my ($long_name,$abstract) = @_;
    my@part_comma = split /,/,$long_name;
    my@part_space = split /\s+/,$long_name;
    my$yes_partfound;
    foreach $part (@part_comma){

        if($abstract=~ m£$part£){
            $yes_partfound++;}}
    foreach $part (@part_space){

        if($abstract=~ m£$part£){
            $yes_partfound++;}}
    return $yes_partfound;}


sub name10plus {
my ($name,$abstract)= @_;
    if($abstract =~ m/\s$name\s/i){
        return 1;}
    else{return 0;}}
sub symbol4plus {
    my ($symbol,$abstract)= @_;
    if($abstract =~ m/\b$symbol\b/){
        return 1;}
        else{return 0;}}
sub symbol2or3 {
    my ($symbol,$abstract)= @_;
    #warn "$abstract\n";
    my$minlength = 100000;
```

```
259     my$length = 100000;
260     my@position =();
261     while($abstract=~ m/\b$symbol\b/g){
262
263         my$pos = pos $abstract;
264         push(@position,$pos);}
265     if(@position eq ()){
266         return 0}
267     my @verify_word = (" gene "," genes "," gene."," gene,"," genes."," genes,"," enzyme ");
268     my@dist=();
269     my@rdist=();
270     my@sorteddist=();
271     my@sorted_rdist=();
272     foreach $position (@position){
273         my($flength,$rlength) = 100000;
274         my@dist=();
275         foreach $verify (@verify_word){
276             my$dist = index ($abstract,$verify,$position);
277             my$rdist = rindex ($abstract,"$verify",$position);
278             if($dist != -1){
279             #warn "dist $verify $position $dist\n";
280                 $flength = $dist - $position;
281                 push (@dist,$flength);}
282             if($rdist != -1){
283             #warn "rdist $verify $position $rdist\n";
284                 $rlength = $position - $rdist;
285                 push (@dist,$rlength);}}
286         @sorteddist = sort {$a <=> $b} @dist;
287         $length = shift @sorteddist;
288         #warn "length $length\n";
289         if($length <$minlength ){
290             $minlength = $length;}}
291     if($position[0]){
292     return("$position[0]:$minlength");}
293     else{return 0}}
294
295
296
297 sub long_name {
298     my($long_name,$abstract) = @_;
299     my@partname = split/,/,$long_name;
300     foreach $part (@partname){
301         if($abstract =~ m/$part/){
```

```
302             my$found++;}}
303     return $found;}
304 sub hash {
305     my($id_symbol,$pmid,$mh,$abstract,$year,$name,$symbol,$allias,$proteine,$partname,@markers) = @_;
306     $data{$id_symbol}{$pmid}={'abstract' => $abstract,'MesH'=> $mh,'year' => $year,'name' => $name,'symbol' => $symbol, 'proteine' => $pro
307
308 sub find_marker {
309     my($abstract)= @_;
310     my@markers;
311     while ($abstract=~ m/(\b[A-Za-z]{1,3}[0-9]+[A-Za-z]{1,3}\b)/g){
312         my$marker = $1;
313         push(@markers,$marker);}
314     while ($abstract=~ m/(\s\+[0-9]+\s)/g){
315         $marker = $1;
316         push(@markers,$marker);}
317     while ($abstract=~ m/([0-9]+[A-Za-z]{1,3}\/[A-Za-z]{1,3}\b)/g){
318         $marker = $1;
319         push(@markers,$marker);}
320     while ($abstract=~ m/(ins\/del|I\/D)/g){
321         $marker = $1;
322         push(@markers,$marker);}
323     @markers = sort @markers;
324
325     my@singel_mark;
326     my$markone = shift @markers;
327     foreach $marktwo (@markers){
328         if($markone ne $marktwo){
329             push (@singel_mark, "$markone");
330             $markone = $marktwo;}}
331
332     return @singel_mark;}
333 sub totalmaker {
334     my @markers = @_;
335     @markers = sort @markers;
336     my$markcount=1;
337     my@singel_mark;
338     my$markone = shift @markers;
339     foreach $marktwo (@markers){
340         if($markone ne $marktwo){
341              push (@singel_mark, "$markcount:$markone");
342             $markcount=1;
343             $markone = $marktwo;}
344         elsif($markone eq $marktwo){
```

```
            $markcount++;}}


    my@sorted_updatedcounts = sort {
        @a_fields = split /:/, $a;
        @b_fields = split /:/, $b;
        $a_fields[0] <=> $b_fields[0]}@singel_mark;
    return @sorted_updatedcounts;}

sub probality {
    my @abstracts =@_;
    my $prob = 1;
    my (%abs_with_word,%word_per_abs,%terms);
    foreach $abstract (@abstracts){
        @newwords=();
        while($abstract=~ m/\b(\w+)\b/g){
            $newword = $1;
            push (@newwords,$newword);
            if(exists $word_per_abs{"$newword"}){
                $word_per_abs{"$newword"} = ($word_per_abs{"$newword"} + 1);}
            else{$word_per_abs{"$newword"} = 1;}}
        @newwords =sort@newwords;
        $old_word = '';
        foreach $word (@newwords){
            if ($word ne $old_word){
                if(exists $abs_with_word{"$word"}){
                    $abs_with_word{"$word"}=($abs_with_word{"$word"}+1);}
                else{$abs_with_word{"$word"} = 1;}
                $old_word = $word;}}
        if($abstract=~ m/([\(\s\b]?p\s?[\<\=]\s?0?\.0[0-5][0-9]*[\)\s\b]?)/i){
        $terms{"sigp"}= 1;}
        if($abstract=~ m/([\(\s\b]?p\s?[\>\=]\s?0?\.[1-9][0-9]*[\)\s\b]?)/i){
        $terms{"nonsigp"}= 1;}
        elsif($abstract=~ m/([\(\s\b]?p\s?[\>\=]\s?0?\.0[6-9][0-9]*[\)\s\b]?)/i){
        $terms{"nonsigp"}= 1;}}
    while(($key1,$value1) = each %terms){
        while(($key2,$value2) = each %probhash){
            if($key1 eq $key2){
                $prob = $prob *($value1 * $value2);}}}
    if($prob > 1){
        return 1;}
    else{return 0;}
}
```

```
sub mikkel {
    my($value)=@_;
    my($fristlocation,$minlength) = split /:/, $value;
    my$status = 0;
    if($fristlocation == 1){
        $status = 1;}
    elsif($fristlocation < 3000 && $minlength < 20){
        $status = 2;}
    elsif($fristlocation < 1000 && $minlength < 40){
        $status = 3;}
    elsif($fristlocation < 400 && $minlength < 80){
        $status = 4;}
    return $status;}
```

```
#!usr/bin/perl
#This programme map ASO probs on chromosome sequence
#use strict;
@id_chr_probes = '';
$entries = 0;
$index = 0;
#$outline = '';
$old_dna = '';
$pre_id = '';
$old_chr = '';
$f_useold = 0;
open(INPUT_FILE, "$ARGV[0]");
chomp (@id_chr_probes = <INPUT_FILE>);
$entries = @id_chr_probes;
for ($index = 0; $index < $entries ; $index++){
    $id=$allele=$chr=$w_aso=$m_aso=$p1=$p2= '';
    $w_asocr=$m_asocr=$p1cr=$p2cr= '';
    $aso_index=$snp_gp_pos=$prime_status=$prime_statuscr= 0;
    $p1pos=$p1poscr=$w_apos=$w_aposcr=$m_apos=$m_aposcr=$p2pos=$p2poscr= 0;
    $snp_gp_pos=$prime_status = 0;
    ($id,$allele,$chr,$w_aso,$m_aso,$p1,$p2) = split /\t/, $id_chr_probes[$index];
    if($old_chr eq $chr){
        $f_useold = 1;}
    else{$old_chr = $chr;
        $old_dna = '';}
     #($p1,$p2)= &prim_teck($p1,$p2);
    ($aso_index) = &aso_teck($w_aso, $m_aso);
    ($w_asocr,$m_asocr,$p1cr,$p2cr) = &conrev($w_aso,$m_aso,$p1,$p2);
    ($p1pos,$p1poscr,$p2pos,$p2poscr) = &find_prim($chr,$p1,$p2,$p1cr,$p2cr);
    ($ppos,$pposcr) = &select_prim($p1pos,$p1poscr,$p2pos,$p2poscr);
    ($w_apos,$w_aposcr,$m_apos,$m_aposcr) = &find_aso($chr,$w_aso,$m_aso,$w_asocr,$m_asocr,$ppos,$pposcr);
#print ("$id:$p1pos-,$p1poscr-,$w_apos-,$w_aposcr-,$m_apos-,$m_aposcr-,$p2pos-,$p2poscr");
#print "\n";
    ($snp_gp_pos) = &select_aso($w_apos,$w_aposcr,$m_apos,$m_aposcr,$ppos,$pposcr,$aso_index);
    if(0){
    if($id eq $pre_id){
        print "$snp_gp_pos\n";}
    elsif($id ne $pre_id){
        print "//\n$id\n$snp_gp_pos\n";
        $pre_id = $id;}}
    warn "$id\t$snp_gp_pos\t$allele";
    print "$id\t$snp_gp_pos\t$allele\n";}
#print "//";
```

```
44
45  ###########SUBROUTINES#############################
46  sub prim_teck {
47      my ($p1,$p2) = @_;
48      my ($p1new,$p2new) = '';
49      $p1new = substr($p1,-14,14);
50      $p2new = substr($p2,-14,14);
51      return ($p1new,$p2new);}
52  sub aso_teck {
53      my ($bi,$b2) ='';
54      my  $aso_index = 0;
55      my  $p2cr = '';
56      my  $flag = 0,
57      my ($w_aso,$m_aso) = @_;
58      my $index = 1;
59      while($index <= length($w_aso)){
60          $b1 = substr($m_aso, -$index,1);
61          $b2 = substr($w_aso, -$index,1);
62          if($b1 eq $b2){
63              $index++;}
64          elsif($b1 ne $b2 && $flag == 0){
65              $aso_index = $index;
66              $flag = 1;
67              $index++;}
68          else{ return 0;}}
69      return $aso_index;}
70  sub conrev {
71      my($w_asocr,$m_asocr,$p1cr,$p2cr) = '';
72      my @temp;
73      my $index = 0;
74      my@prob = @_;
75      foreach $prob (@prob){
76          $prob =~ tr/ATCGatcg/TAGCtagc/;
77          my $i = 1;
78          while ($i <= length ($prob)){
79              $temp[$index] .= substr($prob,-$i,1);
80              $i++;}
81           $index++;}
82      ($w_asocr,$m_asocr,$p1cr,$p2cr) = ($temp[0],$temp[1],$temp[2],$temp[3]);
83      return ($w_asocr,$m_asocr,$p1cr,$p2cr);}
84  sub find_prim{
85      my ($p1pos,$p1poscr,$p2pos,$p2poscr) = 0;
86      my $dna = '';
```

```
    my $to_use_dna = '';
    my $index = 0;
    my $line = '';
    my $prob = '';
    my ($chr,$p1,$p2,$p1cr,$p2cr) = @_;
    my @prob = ($p1,$p1cr,$p2,$p2cr);
    if(!$f_useold){
    $old_dna = '';
    open(CHR,"/home/mbc3/data_store/chr/chr_hg16_nbci34/chr$chr.fa") or die " cant open file $chr ";
    $idline = <CHR>; # rm id line
    while(defined($line = <CHR> )){
        chomp $line;
        $old_dna .= $line;}
    close CHR;}
    else{$f_useold = 0;}
    foreach $prob (@prob){
        pos $old_dna = 0;
        if($old_dna =~ m/$prob/ig){
            $temp[$index] = pos $old_dna;}
        else{$temp[$index]  = 0;}
        $index++;}
    ($p1pos,$p1poscr,$p2pos,$p2poscr) = ($temp[0],$temp[1],$temp[2],$temp[3]);
    return ($p1pos,$p1poscr,$p2pos,$p2poscr);}
sub find_aso {
    my ($w_apos,$w_aposcr,$m_apos,$m_aposcr) = 0;
    my $dna = '';
    my $index = 0;
    my $line = '';
    my $prob = '';
    my $offset = 0;
    my ($chr,$w_aso,$m_aso,$w_asocr,$m_asocr,$ppos,$pposcr) = @_;
    my $local_dna = '';
    my $length;
    if($ppos < $pposcr){
        $length = $pposcr - $ppos;
        $offset = $ppos; }
    else{$length =$ppos - $pposcr;
        $offset = $pposcr;}
     $local_dna = substr($old_dna,$offset,$length);
    my @prob = ($w_aso,$w_asocr,$m_aso,$m_asocr);
    foreach $prob (@prob){
        if($local_dna =~ m/$prob/ig){
            $temp[$index] = pos $local_dna;
```

```
            $temp[$index] += $offset + 1;}
        else{$temp[$index]  = 0;}
        $index++;}
    ($w_apos,$w_aposcr,$m_apos,$m_aposcr) = ($temp[0],$temp[1],$temp[2],$temp[3]);
    return ($w_apos,$w_aposcr,$m_apos,$m_aposcr);}
sub select_prim {
    my ($p1pos,$p1poscr,$p2pos,$p2poscr) = @_;
    my $ppos = '';
    my $pposcr = '';
    if($p1pos && !$p1poscr && !$p2pos && $p2poscr){
            $pposcr = $p2poscr;
            $ppos = $p1pos;}
    elsif(!$p1pos && $p1poscr && $p2pos && !$p2poscr){
            $pposcr = $p1poscr;
            $ppos = $p2pos;}
    elsif( $p1pos && !$p1poscr && !$p2pos && $p2poscr){
            $pposcr = $p2poscr;
            $ppos = $p1pos;}
    elsif(!$p1pos && $p1poscr && $p2pos && !$p2poscr){
            $pposcr = $p1poscr;
            $ppos = $p2pos;}
    else{ return "faild1";}
    return ($ppos,$pposcr)}
sub select_aso{
    my ($w_apos,$w_aposcr,$m_apos,$m_aposcr,$ppos,$pposcr,$aso_index) = @_;
    my $snp_pos = 0;
    my $asofinalpos = 0;
    if(($ppos < $w_aposcr && $w_aposcr < $pposcr) || ($pposcr < $w_aposcr && $w_aposcr < $ppos)){
            $asofinalpos = $w_aposcr;}
    elsif(($ppos < $w_apos && $w_apos < $pposcr) || ($pposcr < $w_apos && $w_apos < $ppos)){
            $asofinalpos = $w_apos;}
    elsif(($ppos < $m_apos && $m_apos < $pposcr) || ($pposcr < $m_apos && $m_apos < $ppos)){
            $asofinalpos = $m_apos;}
    elsif(($ppos < $m_aposcr && $m_aposcr < $pposcr) || ($pposcr < $m_aposcr && $m_aposcr < $ppos)){
            $asofinalpos = $m_aposcr;}
    else { return "faild2";}

    #find snp pos
    if($aso_index){
        $snp_pos = $asofinalpos - $aso_index;}
    else{$snp_pos = $asofinalpos;}
    return ($snp_pos);}

```

Appendix III

Haplotype analyses

Obesity

Table 21 a-e: BMI haplotype analysis

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2482)	Mean BMI	p-value
2_1_1_1_1_1	0.383715	27.7516	0.1287
1_1_1_1_1_1	0.206189	27.8463	0.0973
1_1_2_2_1_1	0.170962	27.3371	0.0731
1_1_2_2_2_1	0.137773	27.379	0.1812
1_1_1_1_1_2	0.0479669	27.2783	0.2874
1_2_2_2_2_1	0.046089	27.5604	0.8655

Over all p value = 0.52

LTA +80 +81 +252 T26N	Frequencies (N=2763)	Mean BMI	p-value
2_1_1_1	0.382498	27.7126	0.116
1_1_1_1	0.255392	27.6926	0.3199
1_1_2_2	0.309691	27.3473	0.0245
1_2_2_2	0.046869	27.4326	0.637

Over all p value = 0.37

TNF -308 -238	Frequencies (N=3037)	Mean BMI	p-value
1_1	0.760815	27.7185	0.043
1_2	0.0485362	27.3436	0.2973
2_1	0.18716	27.425	0.1054
2_2	0.0034888	27.5195	0.8756

Over all p value = 0.23

LTA +80 +81 +252	Frequencies (N=2935)	Mean BMI	P-value
2_1_1	0.380177	27.7355	0.0704
1_1_1	0.2545	27.7068	0.2713
1_1_2	0.315177	27.3247	0.0085
1_2_2	0.0476235	27.4393	0.6232

Over all p value = 0.14

LTA +80 +81 T26N	Frequencies (N=2981)	Mean BMI	P-value
2_1_1	0.382321	27.631	0.1723
1_1_1	0.256056	27.5869	0.5376
1_1_2	0.311539	27.3413	0.0654
1_2_2	0.047635	27.4342	0.7652

Over all p value = 0.38

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of BMI was performed with HTR.

Table 22 a-d: WH ratio haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2474)	Mean whratio	p-value
2_1_1_1_1_1	0.384751	0.817918	0.6982
1_1_1_1_1_1	0.207062	0.821012	0.1675
1_1_2_2_1_1	0.169891	0.817283	0.5906
1_1_2_2_2_1	0.137214	0.814683	0.123
1_1_1_1_1_2	0.0477217	0.819348	0.8178
1_2_2_2_2_1	0.0460357	0.820851	0.5676

Over all p value = 0.63

LTA +80 +81 +252 T26N	Frequencies (N=2752)	Mean Whratio	p-value
2_1_1_1	0.383481	0.817572	0.5384
1_1_1_1	0.255869	0.821218	0.0642
1_1_2_2	0.308383	0.816114	0.1255
1_2_2_2	0.0466929	0.820858	0.5272

Over all p value = 0.67

TNF -308 -238	Frequencies (N=3029)	Mean WHratio	P-value
1_1	0.761475	0.819289	0.4883
1_2	0.0483634	0.820604	0.6616
2_1	0.186693	0.81695	0.2798
2_2	0.00346884	0.830077	0.3055

Over all p value = 0.47

LTA +80 +81 +252	Frequencies (N=2924)	Mean whratio	P-value
2_1_1	0.381092	0.81797	0.8151
1_1_1	0.254944	0.82111	0.0628
1_1_2	0.313968	0.815803	0.0697
1_2_2	0.0474589	0.820481	0.567

Over all p value = 0.34

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of waist hip ratio was performed with HTR.

Insulin resistance

Table 23 a-d: Insulin haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2493)	Mean Insulin	p-value
2_1_1_1_1_1	0.383424	1.9416	0.5353
1_1_1_1_1_1	0.206686	1.9407	0.6544
1_1_2_2_1_1	0.171008	1.95462	0.7821
1_1_2_2_2_1	0.137767	1.96422	0.5231
1_1_1_1_1_2	0.0477535	1.92333	0.548
1_2_2_2_2_1	0.0460845	2.0063	0.1895

Over all p value = 0.15

LTA +80 +81 +252 T26N	Frequencies (N=2776)	Mean Insulin	p-value
2_1_1_1	0.382146	1.9431	0.6167
1_1_1_1	0.255999	1.93874	0.5195
1_1_2_2	0.309501	1.95485	0.6671
1_2_2_2	0.0468296	2.01605	0.0982

Over all p value = 0.69

TNF -308 -238	Frequencies (N=3048)	Mean Insulin	P-value
1_1	0.761191	1.94383	0.5306
1_2	0.0483566	1.91808	0.4531
2_1	0.186972	1.96574	0.3002
2_2	0.00348062	1.97963	0.7496

Over all p value = 0.66

LTA +80 +81 +252	Frequencies (N=2948)	Mean Insulin	P-value
2_1_1	0.379857	1.94377	0.8406
1_1_1	0.254904	1.93559	0.492
1_1_2	0.315143	1.94787	0.8934
1_2_2	0.0475837	2.00753	0.1177

Over all p value = 0.46

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of Insulin levels (natural log was used and the result have to been back transformed) was performed with HTR.

Table 24 a-c: Glucose haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2466)	Mean rfglucose	p-value
2_1_1_1_1_1	0.384995	1.78027	0.8602
1_1_1_1_1_1	0.20529	1.77537	0.2871
1_1_2_2_1_1	0.170858	1.77951	0.8144
1_1_2_2_2_1	0.137857	1.79118	0.1168
1_1_1_1_1_2	0.0474626	1.76385	0.1467
1_2_2_2_2_1	0.0461843	1.79805	0.1487

Over all p value = 0.09

TNF -308 -238	Frequencies (N=3018)	Mean rfglucose	p-value
1 1	0.761196	1.77844	0.1977
1 2	0.0479494	1.77454	0.5906
2 1	0.18728	1.78864	0.0967
2 2	0.00357474	1.78212	0.9448

Over all p value = 0.40

LTA +80 +81 +252	Frequencies (N=2915)	Mean rfglucose	P-value
2_1_1	0.381249	1.78041	0.8836
1_1_1	0.253667	1.77557	0.207
1_1_2	0.314772	1.7833	0.5023
1_2_2	0.0477797	1.79719	0.1259

Over all p value = 0.38

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of glucose levels (natural log was used and the result have to been back transformed) was performed with HTR.

Table 25 a-c Homascore haplotype analysis

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2171)	Mean homascore	p-value
2_1_1_1_1_1	0.389884	0.512664	0.7536
1_1_1_1_1_1	0.205989	0.539779	0.2059
1_1_2_2_1_1	0.166883	0.485374	0.126
1_1_2_2_2_1	0.135988	0.519159	0.9004
1_1_1_1_1_2	0.0479415	0.511671	0.9137
1_2_2_2_2_1	0.0457792	0.547869	0.4628

Over all p value = 0.71

LTA +80 +81 +252 T26N	Frequencies (N=2416)	Mean HOMAscore	p-value
2_1_1_1	0.389418	0.512271	0.8391
1_1_1_1	0.256067	0.531612	0.2727
1_1_2_2	0.302356	0.49846	0.2422
1_2_2_2	0.0465641	0.552501	0.3392

Over all p value = 0.46

TNF -308 -238	Frequencies (N=2654)	Mean Homascore	p-value
1 1	0.763035	0.510785	0.5798
1 2	0.0483821	0.488902	0.5094
2 1	0.184592	0.528119	0.423
2 2	0.00399169	0.637727	0.2203

Over all p value = 0.47

LTA +80 +81 +252	Frequencies (N=2568)	Mean Homascore	P-value
2_1_1	0.386153	0.512561	0.8991
1_1_1	0.254926	0.524858	0.364
1_1_2	0.309326	0.493781	0.1772
1_2_2	0.0472316	0.542408	0.4145

Over all p value = 0.32

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of HOMA score (natural log was used and the result have to been back transformed) was performed with HTR.

Elevated blood pressure

Table 26 a-e Diastolic BP haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2493)	Mean Diastolic	p-value
2_1_1_1_1_1	0.383429	79.8534	0.0241
1_1_1_1_1_1	0.205877	79.0295	0.3018
1_1_2_2_1_1	0.171408	78.9915	0.2999
1_1_2_2_2_1	0.137972	78.8136	0.1833
1_1_1_1_1_2	0.0479575	79.4827	0.8887
1_2_2_2_2_1	0.0460843	80.1213	0.324

Over all p value = 0.38

LTA +80 +81 +252 T26N	Frequencies (N=2776)	Mean diastolic	p-value
2_1_1_1	0.382148	79.9103	0.0129
1_1_1_1	0.255456	79.2571	0.6011
1_1_2_2	0.310042	78.8288	0.0151
1_2_2_2	0.0468296	79.8635	0.5208

Over all p value = 0.28

THF -308 -238	Frequencies (N=3048)	Mean diastolic	p-value
1 1	0.761016	79.5288	0.5628
1 2	0.0485315	79.87	0.5435
2 1	0.186983	79.2333	0.4485
2 2	0.00346982	76.3843	0.1029

Over all p value = 0.32

LTA 80 81 252	Frequencies (N=2948)	Mean Diastolic	P-value
2_1_1	0.379689	79.9627	0.0097
1_1_1	0.254732	79.2629	0.4997
1_1_2	0.315485	78.9291	0.0218
1_2_2	0.0475819	79.7912	0.6152

Over all p value = 0.33

LTA 80 81 T26N	Frequencies (N=2993)	Mean Diastolic	P-value
2_1_1	0.382127	79.7402	0.0445
1_1_1	0.256029	79.292	0.8391
1_1_2	0.311796	78.8248	0.024
1_2_2	0.0476111	79.993	0.3453

Over all p value = 0.36

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of Diastolic blood pressure (natural log was used and the result have to been back transformed) was performed with HTR.

Table 27a-d Systolic BP haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2493)	Mean systolic	p-value
2_1_1_1_1_1	0.383429	147.482	0.3882
1_1_1_1_1_1	0.205877	146.895	0.7755
1_1_2_2_1_1	0.171408	146.93	0.8401
1_1_2_2_2_1	0.137972	146.247	0.3404
1_1_1_1_1_2	0.0479575	147.584	0.7556
1_2_2_2_2_1	0.0460843	148.068	0.5514

Over all p value = 0.84

LTA80 81 252 26	Frequencies (N=2776)	Mean systolic	p-value
1_1_1_1	0.255456	147.153	0.9996
1_1_2_2	0.310042	146.612	0.2893
1_2_2_2	0.0468296	147.6	0.771
2_1_1_1	0.382148	147.608	0.2927

Over all p value = 0.70

TNF -308 -238	Frequencies (N=3048)	Mean systolic	p-value
1 1	0.761016	147.71	0.5175
1 2	0.0485315	147.884	0.8392
2 1	0.186983	147.081	0.4626
2 2	0.00346982	144.36	0.4193

Over all p value = 0.77

LTA +80 +81 +252	Frequencies (N=2948)	Mean Systolic	P-value
2_1_1	0.379689	147.774	0.2855
1_1_1	0.254732	147.171	0.7988
1_1_2	0.315485	146.836	0.3244
1_2_2	0.0475819	148.208	0.5385

Over all p value = 0.12

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of Systolic blood pressure was performed with HTR.

Atherogenic dyslipidemia

Table 28 a-d HDL haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2484)	Mean HDL Cholesterol	p-value
2_1_1_1_1_1	0.384619	1.67525	0.1882
1_1_1_1_1_1	0.205414	1.64241	0.0846
1_1_2_2_1_1	0.171026	1.66684	0.8718
1_1_2_2_2_1	0.138067	1.66374	0.9586
1_1_1_1_1_2	0.0475245	1.6832	0.5197
1_2_2_2_2_1	0.0460505	1.64854	0.5872

Over all p value = 0.92

LTA +80 +81+252 T26N	Frequencies (N=2764)	Mean HDL cholesterol	p-value
2_1_1_1	0.38345	1.67168	0.3199
1_1_1_1	0.254752	1.65422	0.3502
1_1_2_2	0.3094	1.66386	0.9828
2_1_1_1	0.38345	1.67168	0.3199

Over all p value = 0.99

TNF -308 -238	Frequencies (N=2764)	Mean HDL Cholesterol	p-value
1_1	0.761187	1.65751	0.9007
1_2	0.0479994	1.67075	0.6145
2_1	0.187282	1.65725	0.9562
2_2	0.00353167	1.60796	0.4955

Over all p value = 0.83

LTA +80 +81 +252	Frequencies (N=2934)	Mean HDL cholesterol	P-value
2_1_1	0.381163	1.67119	0.4989
1_1_1	0.25373	1.65584	0.3215
1_1_2	0.314948	1.66967	0.6694
1_2_2	0.0476395	1.65639	0.7183

Over all p value = 0.92

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of HDL-Cholesterol was performed with HTR.

Table 29 a-e Triglycerides haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2434)	Mean triglycerides	p-value
2_1_1_1_1_1	0.384305	0.530696	0.9069
1_1_1_1_1_1	0.204083	0.556472	0.0404
1_1_2_2_1_1	0.172681	0.506049	0.0888
1_1_2_2_2_1	0.13763	0.504389	0.1211
1_1_1_1_1_2	0.0476894	0.540191	0.7236
1_2_2_2_2_1	0.0461726	0.560927	0.2862

Over all p value = 0.46

LTA +80 +81 +252 T26N	Frequencies (N=2708)	Mean Triglycerides	p-value
1_1_1_1	0.252998	0.545932	0.0402
1_1_2_2	0.310447	0.502396	0.0208
1_2_2_2	0.0470823	0.543751	0.4833
2_1_1_1	0.383815	0.52511	0.9084

Over all p value = 0.11

TNF -308 -238	Frequencies (N=2980)	Mean triglycerides	p-value
1_1	0.761217	0.535659	0.457
1_2	0.0483466	0.528815	0.8686
2_1	0.186937	0.524566	0.487
2_2	0.00349908	0.518224	0.8386

Over all p value = 0.90

LTA +80 +81 +252	Frequencies (N=2874)	Mean Triglycerides	P-value
2_1_1	0.38147	0.528923	0.587
1_1_1	0.252236	0.541228	0.1154
1_1_2	0.315788	0.502699	0.0108
1_2_2	0.0479371	0.554051	0.2753

Over all p value = 0.12

LTA +80 +81 T26N	Frequencies (N=2919)	Mean Triglycerides	P-value
2_1_1	0.383939	0.522117	0.9829
1_1_1	0.25378	0.542255	0.0519
1_1_2	0.311996	0.503568	0.0376
1_2_2	0.0477903	0.540499	0.4829

Over all p value = 0.13

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of Triglycerides (natural log was used and the result have to been back transformed) was performed with HTR.

Age

Table 30 a-d: Age haplotype analysis.

LTA +80 +81 +252 T26N TNF -308 -238	Frequencies (N=2508)	Mean age	p-value
2_1_1_1_1_1	0.383526	68.6947	0.8898
1_1_1_1_1_1	0.206644	68.9015	0.1987
1_1_2_2_1_1	0.170981	68.7669	0.724
1_1_2_2_2_1	0.137743	68.4923	0.257
1_1_1_1_1_2	0.0476687	68.4349	0.4354
1_2_2_2_2_1	0.0462083	68.7383	0.9326

Over all p value = 0.72

TNF -308 -238	Frequencies (N=3067)	Mean age	p-value
1 1	0.76119	68.8407	0.2236
1 2	0.0482325	68.7594	0.9161
2 1	0.18713	68.6305	0.2763
2 2	0.0034467	67.4491	0.1332

Over all p value = 0.35

LTA +80 +81 +252	Frequencies (N=2964)	Mean Age	P-value
2_1_1	0.379832	68.7587	0.8911
1_1_1	0.255045	68.8499	0.3943
1_1_2	0.314962	68.6481	0.3608
1_2_2	0.0476635	68.812	0.838

Over all p value = 0.53

LTA +80 +81 +252 T26N	Frequencies (N=2791)	Mean age	p-value
1_1_1_1	0.255875	68.8347	0.5344
1_1_2_2	0.309451	68.6976	0.6103
1_2_2_2	0.0469362	68.8206	0.8448
2_1_1_1	0.382244	68.7488	0.9411

Over all p value = 0.53

Rows represent haplotypes. 1 denote major allele, 2 denote minor allele. Frequency for each haplotype is given column 2. Analysis of age was performed with HTR.

Appendix IV

ADHIC Paper

Alcohol dehydrogenase type 1C (*ADH1C*) variants, alcohol consumption traits, HDL cholesterol and risk of coronary heart disease in women and men: British Women's Heart & Health Study and Caerphilly cohorts.

Shah Ebrahim[1,2], Debbie A Lawlor[2], Yoav Ben Shlomo[2], Nicholas J Timpson[2], Roger Harbord[2], Mikkel Christensen[3], Jamil Baban[3], Matt Kiessling[3], Ian N M Day[2,3], Tom R Gaunt[3], George Davey Smith[2]

[1] Department of Epidemiology & Population Health, London School of Hygiene & Tropical Medicine, UK

[2] Department of Social Medicine, University of Bristol, UK

[3] Human Genetics Division, School of Medicine, University of Southampton. UK

Acknowledgements

The British Women's Heart & Health Study is co-directed by Professor Shah Ebrahim, Professor Peter Whincup, Dr Goya Wannamethee together with Dr Debbie A Lawlor. We thank Nicola Ball for running the ADH genotyping assay in the Caerphilly cohort. We thank Carol Bedford, Alison Emerton, Nicola Frecknall, Karen Jones, Rita Patel, Mark Taylor and Katherine Wornell for collecting and entering data, all of the general practitioners and their staff who have supported data collection, and the women who have participated in the study.

The British Women's Heart and Health Study is funded by the Department of Health and British Heart Foundation. DAL is funded by a UK Department of Health Career Scientist Award. NT is funded by a UK Medical Research Council studentship. TRG is a British Heart Foundation Intermediate Fellow. The Caerphilly study was conducted by the former MRC Epidemiology Unit (South Wales) and funded by the Medical Research Council of the United Kingdom. The archive is now maintained by the Department of Social Medicine, University of Bristol. The Caerphilly DNA bank was established with funding from the MRC (G9824960). We thank the staff at the NHS Central Register, Southport for death notification. The Department of Social Medicine of the University of Bristol is the lead centre of the MRC Health Services Research Collaboration. The views expressed in this publication are those of the authors and not necessarily those of any of the funding bodies. The funding bodies have had no influence over the scientific work or its publication.

Word count: text 3885, abstract 230

Abstract

Background. Genetic variants involved in alcohol metabolism provide a means of testing the effects of alcohol intake on coronary heart disease (CHD) and its risk factors. A previous study had identified an interaction between the γ2γ2 slow oxidizer variant of the alcohol dehydrogenase gene (*ADH1C*) and alcohol consumption on HDL cholesterol and coronary heart disease risk. We undertook replication studies in two large population cohorts of women and men.

Design. Prospective general population cohort studies

Participants. 3234 women and 1313 men with relevant genotypic and phenotypic data

Methods. Participants were genotyped for *ADH1C* variant rs1693482 using a competitive allele specific PCR SNP genotyping system. Alcohol intake was assessed by interview and self-completed questionnaires. Standard clinical laboratory biochemical assays were used, and primary care, hospital records and death certificates were used to ascertain CHD events.

Results. No association was found between the single nucleotide polymorphism (SNP) marking γ1 and γ2 alleles at the *ADH1C* locus and HDL cholesterol, blood pressure or CHD risk, although there was an association with alcohol consumption. There was no evidence of interactions between *ADH1C* variants and alcohol intake on HDL cholesterol, blood pressure or CHD risk. Further stratification of women according to hormone replacement treatment did not uncover any evidence of interaction.

Conclusion. Previously reported findings of associations and interactions between *ADH1C*, alcohol and HDL cholesterol and CHD did not replicate in two large population cohorts.

Introduction

Moderate alcohol consumption is associated with reduced coronary heart disease (CHD) risk in large observational epidemiological studies, although debate remains over the interpretation of U- or J-shaped relationships sometimes observed, and whether the benefits are directly attributable to alcohol or are confounded by other lifestyle and social factors.[1 2 3 4 5] Classification of drinking habits may be inaccurate, with heavy drinkers reporting more socially acceptable levels of intake. Separation of ex-drinkers from life-long abstainers is important as the former may have given up because of ill-health related to alcohol use, leading to spurious associations with outcomes in observational studies due to reverse causality. In small short-term experimental studies alcohol has direct biological effects that would reduce the risk of CHD – increasing levels of protective high density lipoprotein (HDL) cholesterol, [6,7] but may also increase blood pressure.[8,9]

Alcohol dehydrogenase (ADH) oxidizes alcohol to acetaldehyde, which is in turn oxidized by aldehyde dehydrogenase (ALDH) to acetate.[10 11] One of the *ADH* genes, *ADH1C (*previously known as *ADH3)*, has two polymorphic forms which produce two different polypeptide enzyme subunits; *ADH1C*1* produces γ1 and *ADH1C*2* produces γ2.[12] The γ1 allele differs from the γ2 allele by two amino acids at positions 271 (non-synonymous arginine to glycine change, rs1693482) and at 349 (non-synonymous isoleucine to valine change, rs698). [11] According to linkage disequilibrium patterns observed between these two loci, variation at the marker rs1693482 is sufficient to score individuals for carriage of the ADH1C alleles.[13] Allele frequencies in European origin populations are roughly 60% γ1 and 40% γ2 [14,15] with differences in the maximal velocity of alcohol oxidation: γ1γ1 gives a 2.5-fold higher rate than γ2γ2.[11] Slow oxidizers would

be predicted to have a lower risk of CHD, higher HDL cholesterol and higher blood pressure which was shown in a case-control study (374 cases and 770 controls) nested within the Physicians Health Study, risk ratios for CHD comparing the homozygous fast oxidizers (γ1γ1) with the heterozygote γ1γ2 and with the homozygous slow oxidisers γ2γ2 were 0.90 (95% CI 0.69, 1.17) and 0.72 (0.50, 1.05) respectively.[16 17] Interactions were demonstrated between *ADH1C* variants and alcohol intake on CHD risk, with a much reduced relative risk in homozygous slow oxidizer (γ2γ2) regular drinkers (0.14, 95% CI 0.04, 0.45) compared with homozygous fast oxidizers (γ1γ1) drinking less than 1 drink a week. Interactions were also demonstrated for HDL-cholesterol with slow oxidizers and drinking regularly having higher levels, but these findings were confined to an augmented study group comprising men and post-menopausal women not taking hormone replacement treatment.[16, 18]

Using data from two large epidemiological studies, one of women and the other of men, we tried to replicate these findings, hypothesizing that slow oxidizers, identified by SNP rs1693482 marking the γ2γ2 polymorphism of the *ADH1C* gene, would have higher HDL-cholesterol, and a reduced risk of CHD, and that interactions with alcohol intake would be found. We also attempted to use Mendelian randomization to estimate the unconfounded and unbiased effects of alcohol intake on HDL-cholesterol as previously proposed.[17 19]

Methods

British Women's Heart and Health Study

Subjects were participants of the British Women's Heart and Health Study (BWHHS). Between 1999 and 2001 4,286 women (60% of those invited) aged 60 to 79 years, who were randomly selected from 23 British towns were interviewed, examined, completed medical questionnaires and had detailed reviews of their medical records. These women have been followed-up over a median of 4.7 years by flagging with the NHS central register for mortality data and review of their medical records every two years. Blood samples were taken after a minimum of a 6-hour fast. Plasma glucose was measured by a glucose oxidase Trinder method[20] using a Flacor 600 automated analyser. Serum insulin was measured using an ELISA assay which does not cross react with proinsulin.[21] Insulin resistance was estimated according to the homeostasis model assessment (HOMA) as the product of fasting glucose (mmol/l) and insulin (μU/ml) divided by the constant 22.5.[22] Levels of gamma-glutamyl transferase (gamma-GT) in serum were determined using an automated analyzer (Technicon Sequential Multiple Analyzer; Technicon Instruments Corporation, Tarrytown, New York). There was no evidence of a diurnal variation in GGT with time of blood sampling being unrelated to GGT ($p = 0.9$). High density lipoprotein cholesterol and triglyceride levels were measured using a Hitachi 747 automated analyser and reagents supplied by Roche Diagnostics. Fibrinogen was assayed in stored citrated plasma by the Clauss assay in an MDA-180 automated coagulometer (Organon Teknika), and CRP by a high-sensitivity immunonephelometric assay on a ProSpec protein analyser (Dade-Behring).

Standard procedures were used to assess blood pressure, height (standing and seated), weight, waist and hip circumference as previously described.[23] Information on occupational social class, smoking and physical activity were obtained from either the research nurse baseline interview or self-complete questionnaire as previously reported.[24] We used a combined measure of prevalent (self-report of a doctor diagnosis of angina or myocardial infarction and/or evidence in medical record review at baseline of either of these diagnoses) and incident (death from CHD (ICD10 codes I20-I25, I51.6) or evidence of angina or myocardial infarction in the medical record review that occurred during the follow-up period up to 31st December 2004) CHD events.

Caerphilly Study

Full details of the Caerphilly study are available elsewhere.[25] Briefly, in Caerphilly, a former mining town in South Wales, electoral rolls for the defined area were used to identify a random sample of men, each of whom was contacted and asked his age. Those aged 45-59 years were invited to take part in the study. Of 2,818 men identified 2,513 (89%) participated in the baseline survey – Phase I - in 1979-83. At recruitment a clinic assessment comprising a standard medical history, weight, height, blood pressure, venepuncture for biochemical and hormonal assays, and a 12-lead ECG was performed. Insulin resistance was estimated according to the homeostatis model assessment (HOMA) as reported above.[22] In the Caerphilly Study, only myocardial infarction was ascertained over the ensuing 20 years at follow up examinations held every four to five years. Men were asked about chest pain, and doctor diagnosis of heart attack. ECG-defined ischaemia at baseline and at subsequent clinic visits was determined using Minnesota codes 1-1-1 to 1-2-5 or 1-2-7. Questions about hospital admissions for severe chest pain,

together with hospital activity analysis notifications of admissions coded 410-414 (9th revision of the International Classification of Diseases), were used as the basis for a detailed search of hospital notes to identify events that satisfied the WHO criteria for acute myocardial infarction. General practitioner records were inspected for events which had not led to hospital admission. Deaths from coronary heart disease up to 1st July 2001 comprised all those coded as 410-414 (ICD 9) and were obtained from the National Health Service central registry.

Assessment of alcohol consumption.

Analyses were conducted on current drinking status for both women and men with aggregation of reports of consumption in a typical week in terms of number of drinks taken, defined colloquially (i.e. a half of beer, a single measure of spirits, a glass of wine) and approximating 1 unit, containing about 10g of alcohol. The distribution of units drunk was very skewed so the data were logged and then grouped into thirds of the distribution and a separate category of life-long abstainers. In women, 518 ex-drinkers on health grounds were excluded from analyses to avoid possible reverse causality effects. Among the few male ex-drinkers, reasons were not reported but given their relatively younger ages, ill-health was unlikely.

Genotyping

DNA was extracted from K-EDTA whole blood samples by salting out procedure.[26] Quantitation was by picoGreen assay and DNA concentrations were equalised by dilutions with water. Long term stock DNA aliquots were laid down and working 96-well plates of DNA dilutions to 10ng/μl prepared. Degenerate oligo primer amplifications

('DOP-DNA') were made from dilution plates in order to conserve stock DNA and 384-well PCRs were performed from DOP-DNA representing 0.1ng of original genomic DNA. The DOP protocol was a modified version of the method used by Cheung and Nelson[27] designed to minimise loss of representation of %GC-rich genomic regions. The ADH1C genotype was determined using fluorescence-labelled oligonuclotide melting from matched or mismatched target, monitored in an Idaho Technology (Salt Lake City, Utah, USA) 384-well Odyssey. Detection depended on dequenching of the fluorescein-based (FITC) probe upon melting during a thermal ramp, quenching being made by a dabcyl group present on an olgionucleotide with a higher melting temperature, located adjacent with the SNP probe on the target strand. Asymmetric PCR was performed on 2µl of dried DOP amplified template in 384-well white PCR plates (Abgene®, Epson, Surrey, UK) on a MJ Research PTC-225 DNA Engine Tetrad® (Genetic Research Instrumentation Ltd., Braintree, Essex, UK). Samples were amplified with primers 5'-TCCTCCAGGTTGCAGAGGCAGA-3' at 100nM and 5'-CCCATTCAGGAAGTGCTAAAGGAAATGA-3' at 500nM. The FITC probe (with 3' phosphate) 5'-F-TCAAGC[C/T]GACCGATGA-PHOS-3' (single base polymorphism rs1693482 underlined) and the dabcyl quencher 5'-GGGCATGTCACGGATCATACCATGG-3' were each included at 200nM in the PCR for the Odyssey melting assay. 5µl PCR reaction mix also contained: 1x PCR buffer (Promega, Southampton, UK), 200µM dNTPs (Promega), 1.5mM $MgCl_2$ (Promega) and 0.01U/µl of Taq DNA polymerase (Promega). PCR cycling conditions were: 94°C for 2 minutes then 49 cycles of 94°C for 20 seconds, 60°C for 30 seconds and 72°C for 120 seconds, followed by 72°C for 2 minutes. Samples were overlaid with 5µl Chill-Out™ wax (Genetic Research Instrumentation) to prevent evaporation during analysis.

Following PCR amplification samples were melted from 45°C to 75°C in the 384-well Odyssey. LightTyper software (Roche Diagnostics Ltd.) was used to analyse the change in fluorescence during melting, and to group melting profiles into genotype groups. These were then manually verified using in-house software. The characteristics of this software and calling (a constant peak derived from an homologous ADH gene occurs in all profiles) have been described previously.[28] All genotyping was done by the Southampton Human Genetics Laboratory.

Statistical methods

Hardy-Weinberg equilibrium was tested on a contingency table of observed-versus-predicted genotypic frequencies using an exact test.[29] HDL-C, blood pressure, CHD and distributions of other risk factors were presented by genotype and categories of alcohol consumption. HOMA-R scores and serum triglyceride levels were positively skewed and therefore were presented as geometric means in descriptive statistics and logged values were used in regression models. Our prior hypothesis was that the relationship between the *ADH1C* variant and alcohol intake, HDL cholesterol and CHD risk would be defined by a per allele association reflecting the speed of oxidation as found in previous studies,[16,18] so groupings of $\gamma 1\gamma 1$, $\gamma 1\gamma 2$, and $\gamma 2\gamma 2$ were used in analyses. Linear and logistic regression models were used to compute statistical tests for the differences between categories and for linear trends across categories. Likelihood ratio tests were computed to test for evidence of statistical interactions.

We used instrumental variable methods [30] to estimate the unconfounded and unbiased effects of alcohol intake on HDL-cholesterol derived from *ADH1C* – HDL-cholesterol

association using a Mendelian randomization design, comparing estimated with observed associations.[31] This gives a point estimate identical to the ratio of the coefficient for the regression of HDL-cholesterol on *ADH1C* to that of alcohol intake on *ADH1C*, with confidence intervals that account for the uncertainty in both associations. This approach is analogous to intention to treat analysis in randomised trials, where the genotype – HDL-cholesterol association represents an "intention to treat" analysis and the alcohol – HDL-cholesterol association represents the "treatment received" analysis.[30] The latter analysis is not a by-randomisation comparison and therefore prone to bias, whereas the intention to treat analysis is by randomization and free of potential confounding. The instrumental variable analyses used the built-in Stata ivreg2[32] command and we examined F-statistics from the first-stage regressions to evaluate the strength of the instruments. Values greater than ten are often taken to indicate sufficient strength to ensure the validity of instrumental variable method.[33] All analyses were conducted in Stata 9.

Results

Of the total 4286 women, 441 had insufficient blood taken for adequate assays to be made and 37 refused consent for the use of stored blood. Of the remaining 3808 DNA samples, 3234 were available for analysis representing those with data on alcohol consumption and described as "white" by the research nurse. In the Caerphilly study, data of 1313 white men were available for analysis. In both samples, the distribution of rs1693482 was in Hardy Weinberg equilibrium (women, p = 0.205, men, p=0.155).

There were no differences in the distribution of risk factors among those included in the main analyses and those excluded because of lack of genotype data, with the exception of

systolic blood pressure in women, where it was slightly higher in those included than excluded (147.7 mm Hg vs. 144.8 mm Hg, p = 0.01). The proportion of life-long alcohol abstainers was higher (23% vs. 5%) among women than men, although the proportions of regular drinkers were similar (19% vs. 21%). The geometric mean units per week among drinkers were substantially higher in the men than women (7.5 vs. 2.7 drinks). The distribution of cardiovascular risk factors by *ADH1C* variants is shown in Table 1. In women and men, weekly consumption of alcohol in slow oxidizers was higher ($p<0.001$ women, $p=0.12$ men), but no association with life-long alcohol abstention was seen. Gamma-GT levels greater or equal to 80 IU/l were weakly associated with *ADH1C* in women. No associations were found between ADH1C variants and HDL-cholesterol, triglycerides or blood pressure in men or women. There were no associations with HOMA, BMI, waist hip ratio, height, leg length, social class measured in childhood and adulthood. Surprisingly, in women, but not men, slow oxidizers were more likely to smoke than fast oxidizers ($\gamma2\gamma2$ 13.6% vs. $\gamma1\gamma2$ 11.9% vs. $\gamma1\gamma1$ 9.2%, $p=0.005$), independently of alcohol consumption. The risk of CHD showed weak evidence of a trend for *higher* CHD risk in slow oxidizers in women, contrary to the direction hypothesized, and point estimates for myocardial infarction also showed a similar trend (see Figure 1). The magnitude of the association between genotype and CHD in women was not altered by adjustment for smoking status. In men, there was no pattern of CHD risk with *ADH1C* genotype.

Associations with some risk factors showed different patterns in men and women (see Table 2). As expected, HDL-cholesterol and triglycerides were associated with alcohol consumption in both sexes. In women, but not men, systolic blood pressure and

prevalence of hypertension were highest among abstainers. In women, but not men, there was some evidence of patterning of lower alcohol consumption by adult and childhood manual social class. Strong negative associations were found between risk of CHD and alcohol intake, with life-long abstainers and the lowest drinking group being at highest risk (See figure 1). Among men and women who drank regularly (at least 3 ½ drinks a week for women and 7 drinks a week for men) there was no evidence of any association between *ADH1C* variants and HDL cholesterol or systolic blood pressure. In women, there was weak evidence for an association with CHD but this was in the opposite direction to that predicted (see Table 3).

There was no evidence of any interactions between *ADH1C* and alcohol consumption on HDL-cholesterol in women (see Figure 2). In men weak evidence of an interaction ($p=0.04$) but none of the main effects were significant, suggesting that the apparent interaction was a chance finding. No evidence of interaction was found between *ADH1C* and alcohol consumption on systolic blood pressure or coronary heart disease risk in women or men. Further exploration of the effects of hormone replacement therapy in women failed to uncover any evidence of an interaction between *ADH1C* and alcohol consumption on HDL-cholesterol among women who had never or ever used hormone replacement, but in never-users a *higher* risk of CHD in slow oxidizers compared to fast oxidizers was found ($p=0.02$) (see Table 4).

In linear regression analysis there was a positive association between units of alcohol drunk (among women who were drinkers) and HDL-cholesterol: for each unit per week of alcohol HDL-cholesterol increased by 0.019 (95%CI: 0.015, 0.023) mmol/l, $p <0.001$. When we repeated this analysis using *ADH1C* categories as instruments for alcohol

consumption there was no association between it and HDL-cholesterol: change per weekly unit of alcohol increase –0.004 (95%CI: -0.041, 0.034), p = 0.85, However, because *ADH1C* explains only a small proportion of the variation in alcohol consumption in the population the confidence intervals for this estimate are wide and include the effect estimate from the ordinary linear regression: difference between observed and instrumental variables estimates, p=0.22). The first-F statistic for the instrumental variable regression was 11.1, suggesting sufficient strength for use in the analysis.

Discussion

We have demonstrated that *ADH1C* variants that are related to alcohol intake, are not associated with either cardiovascular risk factors (particularly HDL-cholesterol) or CHD in two large independent cohort studies. We were unable to find strong evidence of interaction effects between alcohol consumption and *ADH1C* variants on risk factors or coronary heart disease risk. We did find the expected associations between reported alcohol consumption and HDL-cholesterol and coronary heart disease. The association found between smoking habits and *ADH1C* genotype in women was not expected and was probably due to chance given the large number of comparisons made.

Recent findings of an interaction between genotype and hormone replacement treatment on HDL cholesterol levels in women [18] were not replicated in our study. We found evidence of an *increased* risk of CHD in women with γ2γ2 genotype not taking hormone replacement treatment relative to γ1γ1, the opposite to what would be expected. The previously reported interaction between moderate alcohol consumption and *ADH1C* genotype on HDL cholesterol level (a 10% higher HDL cholesterol in γ2γ2 than γ1γ1 groups) was confined to a combined sub-group of older men and postmenopausal women

but was not found in the whole sample or in four component samples (postmenopausal, not taking hormone replacement therapy; postmenopausal, taking hormone replacement therapy; premenopausal; and older men).[18] As with any post hoc sub-group analysis caution is required in interpretation as false positive findings are common.

As the earlier findings [16 18] did not replicate in our studies, it is important to consider the limitations of our study. While we have considerably more data than were available to previous investigators and had sufficient power to detect the differences they had observed, our study is still relatively small with only 842 CHD cases, and for clinical outcomes we had insufficient numbers of myocardial infarctions in women to make direct comparisons with men. However, Hines and colleagues' case control study findings of interaction between *ADH1C* variants and alcohol were based on only 5 cases to 37 controls with γ2γ2 variant relative to 50 cases and 78 controls with γ1γ1 variant.[16] Strengths are that both our samples were randomly selected from general populations, events were ascertained in a systematic way without knowledge of genotype or exposure data, and both men and women were included. It is possible that our inconsistent findings represent a common pattern in genetic association studies of initial positive findings followed by negative findings in larger studies.[34 35 36] It is also possible that differences in classification of alcohol consumption and overall levels of reported intake may have resulted in inconsistent findings. However, our women and men appeared to report similar levels of alcohol intake to those in previous published studies[16, 18]

Other investigators have also been unable to fully replicate these earlier findings as is shown in Table 5 which summarises these studies. In an Australian study of 901 twins, a

strong effect of reported alcohol intake on HDL cholesterol was found but no interactions with *ADH1C* variants and alcohol were demonstrated on HDL cholesterol or apolipoprotiens.[37] While both the Framingham Offspring Study[13] and Northwick Park study[38] analyses demonstrated a reduced risk of coronary heart disease with genotype, neither found evidence of any main effect of genotype on HDL-cholesterol nor any evidence of an alcohol-genotype interaction.. Interestingly, no effect of the *ADH1C* variant on risk of ischaemic stroke was found using data from the Physicians Health Study, which was interpreted as demonstrating the difference in pathological mechanisms of ischaemia between stroke and coronary heart disease.[39] However, there is evidence of a protective effect of HDL cholesterol in ischaemic stroke.[40] While not directly relevant to the ADH1C association, a Japanese study of 826 men and 1295 women found that the ALDH2 variant, that is associated with alcohol intake, was not related to HDL cholesterol in those drinking on average more than half a drink a day.[41]

The hypothesis that *ADH1C* variants interact with alcohol consumption, leading to differences in HDL cholesterol and CHD risk rests on evidence that the *ADH1C* variants alter alcohol metabolism, with the γ2γ2 variant clearing alcohol more slowly, increasing the HDL cholesterol lowering effects of alcohol and thereby reducing the risk of CHD. Although differences in the rate of alcohol oxidation have been reported,[14] these findings have not been replicated in other studies.[42 43 44] *In vitro* evidence suggests that a haplotype derived from three SNP markers and a 66bp in/del in the regulator region upstream of the *ADH1C* gene increases *ADH1C* transcription activity by two-fold in transient transfection assays in H4IIE-C3 cells.[45] Further *in vivo* work examining the role

of this regulatory haplotype may be helpful in understanding the association of *ADH1C* with CHD.

Other enzymes have a role in alcohol metabolism, in particular the aldehyde dehyrogenases, and may be more important determinants of the cardioprotective effects of alcohol than *ADH1C* variants. For example, variants in the *ALDH2* gene are commonly found in Asian but not European or North American populations and they lead to marked differences in the efficiency of alcohol metabolism. Half of Japanese people are heterozygotes or homozygotes for a null variant of *ALDH2* and peak blood acetaldehyde concentrations post alcohol challenge are 18 times and 5 times higher among homozygous null variant and heterozygous individuals compared with homozygous wild type individuals.[46] This renders the consumption of alcohol unpleasant through inducing facial flushing, palpitations, drowsiness and other symptoms, resulting in very considerable differences in alcohol consumption according to genotype.[47] Men either homozygous or heterozygous for null *ALDH2* (and therefore less likely to drink) were at twice the risk of myocardial infarction, and this effect was greatly attenuated by adjustment for HDL cholesterol, indicating the likely mechanism by which alcohol reduces the risk of CHD.[47] Genetic variants that have powerful effects on alcohol habits would be useful for establishing unconfounded and unbiased effects of drinking on disease outcomes using a Mendelian randomization design, although very large sample sizes would be required to demonstrate null associations between alcohol and outcomes.[48] Our own instrumental variable estimate of the alcohol effect on HDL-cholesterol derived from the *ADH1C* variant and alcohol associations were consistent with the observed associations using reported alcohol intake but the confidence intervals were wide,

reflecting the limited precision of estimates based on even a moderate sample size and a small *ADH1C* association with alcohol intake.

In summary, our study did not replicate earlier findings of a role for the *ADH1C* variant in determining HDL cholesterol levels or CHD risk. These earlier findings probably represent chance findings associated with small sample sizes, reflecting a general problem in genetic epidemiology.[35] It is clear that there are genetic influences on the propensity to drink alcohol, and identifying genetic variants that have a more profound effect on alcohol consumption and its metabolic clearance would enable the apparent cardioprotective role to be explored in Mendelian randomization studies. Very large sample sizes are required to provide sufficient power to estimate genetic variant – disease associations and interactions with any precision.[49] Future studies will need to be much larger than those to date, and further work on the functional effects of specific polymorphisms and their regulation, and the ways in which inheritance of different variants interact with each other is needed.

Table 1: Prevalence and means (95% CI) of cardiovascular risk factors and coronary heart disease by *ADH1C* genotype: British Women's Heart & Health Study and Caerphilly Study

	BWHHS (Women aged 60-79 years) Percent or mean (95% CI) by *ADH1C* variants N = 3234				Caerphilly (Men aged 47 to 67 years) Percent or mean (95% CI) by *ADHIC* variants N=1313			
	γ1γ1 N = 1095	γ1γ2 N = 1604	γ2γ2 N = 535	p[$]	γ1γ1 N = 462	γ1γ2 N = 612	γ2γ2 N = 239	p[$]
Age (years)	68.8 (68.5, 69.1)	69.0 (68.7, 69.3)	68.8 (68.3, 69.2)	0.86	56.4 (56.0, 56.9)	56.7 (56.4, 57.1)	57.0 (56.4, 57.5)	0.13
Alcohol consumption markers								
Life-long tee-total (%)	23.6 (21.1, 26.3)	24.7 (22.6, 27.0)	22.1 (18.7, 26.0)	0.72	6.1 (3.7, 8.4)	6.4 (4.4, 8.5)	5.6 (2.5, 8.6)	0.86
Regular daily / most days drinker (%)	20.6 (18.2, 23.2)	16.5 (14.7, 18.5)	22.5 (19.1, 26.4)	0.94	18.9 (15.3, 22.5)	22.0 (18.7, 25.3)	23.0 (17.7, 28.4)	0.16
Mean units per week among drinkers*	2.77 (2.55, 2.99)	2.47 (2.32, 2.64)	3.26 (2.88, 3.67)	<0.03	6.7 (5.8, 7.7)	7.8 (6.9, 8.8)	8.4 (6.9, 10.3)	0.04
Gamma-GT mean units/L*	22.15 (21.36, 22.97)	23.14 (22.42, 23.88)	23.40 (22.14, 24.73)	0.06	25.7 (24.3, 27.2)	24.9 (23.8, 26.1)	27.6 (25.6, 29.7)	0.29
High Gamma-GT (>=80 units/L) %	3.9 (2.9, 5.2)	5.3 (4.2, 6.4)	6.5 (4.6, 8.9)	0.02	9.7 (7.0, 12.4)	10.0 (7.6, 12.3)	11.3 (7.3, 15.3)	0.56
Risk of coronary heart disease and myocardial infarction								
CHD (%)	19.1 (16.8, 21.5)	21.8 (19.8, 23.8)	22.1 (18.7, 25.8)	0.10	N/A	N/A	N/A	N/A
MI only (%)	3.1 (2.2, 4.3)	3.4 (2.6, 4.4)	4.3 (2.9, 6.4)	0.24	13.4 (10.3, 16.5)	11.1 (8.6, 13.6)	14.6 (10.1, 19.1)	0.91
Alcohol related coronary risk factors								
HDLc (mmol / l)	1.67 (1.64, 1.70)	1.65 (1.63, 1.67)	1.64 (1.60, 1.68)	0.19	1.03 (1.00, 1.05)	1.02 (1.00, 1.04)	1.04 (1.01, 1.07)	0.71
Triglycerides (mmol/l)*	1.68 (1.64, 1.73)	1.68 (1.65, 1.72)	1.67 (1.61, 1.74)	0.34	1.71 (1.63, 1.79)	1.68 (1.61, 1.75)	1.64 (1.54, 1.74)	0.29
Systolic blood pressure (mmHg)	148.2 (146.7, 149.7)	146.9 (145.7, 148.2)	148.6 (146.5, 150.7)	0.81	144.7 (142.6, 146.7)	146.0 (144.2, 147.9)	146.6 (143.6, 149.6)	0.25
Diastolic blood pressure (mmHg)	79.5 (78.8, 80.2)	79.5 (78.9, 80.0)	79.9 (78.8, 80.9)	0.73	84.2 (83.1, 85.2)	84.4 (83.4, 85.3)	85.1 (83.6, 86.5)	0.39
Hypertension (BP >=160/90 or on blood pressure medication)	52.7 (49.7, 55.7)	51.7 (49.3, 54.2)	52.7 (48.5, 56.9)	0.90	30.5 (26.3, 34.7)	35.8 (32.0, 39.6)	33.1 (27.1, 39.0)	0.30

Other coronary risk factors								
BMI (m/kg^2)	27.7 (27.4, 28.0)	27.6 (27.3, 27.8)	27.7 (27.3, 28.1)	0.93	26.8 (26.4, 27.1)	26.5 (26.2, 26.8)	26.4 (26.0, 26.8)	0.16
Waist hip ratio	0.820 (0.816, 0.824)	0.818 (0.814, 0.821)	0.822 (0.816, 0.828)	0.97	N/A	N/A	N/A	N/A
Standing height (mm)	1589 (1585, 1593)	1588 (1585, 1591)	1587 (1582, 1592)	0.50	1713 (1707, 1720)	1711 (1705, 1717)	1713 (1704, 1722)	0.86
Leg length (mm)	758 (755, 760)	757 (755, 759)	757 (754, 760)	0.78	802 (797, 806)	801 (797, 805)	800 (794, 805)	0.60
Current smoker	9.2 (7.6, 11.1)	11.9 (10.3, 13.5)	13.6 (10.9, 16.8)	0.005	41.2 (36.7, 45.7)	41.9 (37.9, 45.8)	39.5 (33.2, 45.8)	0.75
Adult head of household manual social class (%)	61.4 (58.4, 64.2)	61.7 (59.3, 64.0)	61.7 (57.5, 65.7)	0.89	87.4 (84.0, 90.8)	84.4 (81.0, 87.8)	86.8 (81.9, 91.7)	0.64
Child manual social class (%)	80.2 (77.7, 82.5)	81.0 (79.1, 82.9)	79.1 (75.4, 82.3)	0.77	65.9 (61.6, 70.3)	65.2 (61.4, 69.0)	63.7 (57.5, 69.9)	0.57

[$] p-values are for linear trend across groups (1df)
CHD = incident and prevalent myocardial infarction / angina/ angioplasty / CABG
MI = incident and prevalent WHO verified myocardial infarction only

Table 2: Prevalence and means (95% CI) of cardiovascular risk factors and coronary heart disease by alcohol intake: British Women's Heart & Health Study and Caerphilly Study

	BWHHS (Women aged 60-79 years) Percent or mean (95% CI) by Current alcohol consumption N = 2716					Caerphilly (Men aged 47 to 67 years) Percent or mean (95% CI) by Current alcohol consumption				
	Life long abstainers N = 631	Lowest 1/3 N = 1056 range 0.5-2 units per week	Second 1/3 N = 335 range 3-4 units per week	Highest 1/3 N = 694 range 5-42 units per week	p[$]	None N=76	Lowest 1/3 N=470 range 0.2 to 4.6 units per week	Second 1/3 N=486 range from 5 to 18.2 units per week	Highest 1/3 N=366 range from 18.3 to 176 units per week	p[$]
Age (years)	69.8 (69.4, 70.2)	69.0 (68.7, 69.3)	68.2 (67.7, 68.8)	68.3 (67.9, 68.7)	<0.001	57.9 (57.0, 58.8)	57.2 (56.8, 57.6)	56.7 (55.3, 56.1)	55.7	<0.001
Alcohol consumption markers										
Gamma-GT mean units/L*	22.19 (21.04, 23.25)	22.46 (21.62, 23.33)	22.27 (20.88, 23.75)	23.97 (22.86, 25.13)	0.11	22.78 (20.16, 25.75)	21.45 (20.49, 22.45)	24.87 (23.69, 26.10)	34.23 (32.08, 36.52)	<0.001
High Gamma-GT (>=80 units/L) %	5.0 (3.5, 7.0)	4.8 (3.7, 6.3)	3.6 (2.1, 6.2)	5.5 (4.1, 7.5)	0.04	11.8 (4.5, 19.2)	5.5 (3.5, 7.6)	6.2 (4.0, 8.3)	13.7 (10.1, 17.2)	0.003
Risk of coronary heart disease and myocardial infarction										
CHD (%)	24.4 (21.2, 27.9)	20.3 (18.0, 22.8)	18.5 (14.7, 23.0)	16.4 (13.9, 19.4)	<0.001	N/A	N/A	N/A	N/A	N/A
MI only (%)	4.1 (2.8, 6.0)	4.3 (3.2, 5.7)	1.5 (0.6, 3.5)	2.6 (1.6, 4.1)	0.05	19.7 (10.7, 28.8)	14.7 (11.5, 17.9)	10.7 (7.9, 13.5)	12.6 (9.2, 16.0)	0.08
Alcohol related coronary risk factors										
HDLc (mmol / l)	1.56 (1.53, 1.60)	1.63 (1.60, 1.66)	1.72 (1.67, 1.77)	1.83 (1.80, 1.87)	<0.001	0.98 (0.93, 1.04)	0.97 (0.95, 0.99)	1.04 (1.01, 1.06)	1.10 (1.07, 1.13)	<0.001
Triglycerides (mmol/l)*	1.78 (1.72, 1.85)	1.70 (1.65, 1.75)	1.58 (1.51, 1.66)	1.52 (1.47, 1.57)	<0.001	1.53 (1.39, 1.69)	1.61 (1.54, 1.69)	1.65 (1.58, 1.73)	1.88 (1.77, 1.99)	<0.001
Systolic blood pressure (mmHg)	150.0 (148.0, 152.0)	146.8 (145.2, 148.3)	145.3 (142.6, 148.1)	146.9 (145.0, 148.7)	0.01	145.1 (139.3, 150.8)	143.9 (141.9, 146.0)	144.9 (143.0, 146.9)	148.9 (146.6, 151.3)	0.005
Diastolic blood pressure (mmHg)	79.6 (78.7, 80.6)	79.3 (78.6, 80.1)	78.4 (77.1, 79.7)	80.3 (79.4, 81.2)	0.41	85.4 (82.5, 88.4)	84.2 (83.1, 85.3)	84.3 (83.2, 85.3)	85.0 (83.9, 86.0)	0.65
Hypertension (BP >=160/90 or on blood pressure medication)	56.9 (53.04, 60.7)	51.7 (48.6, 54.7)	46.5 (41.2, 51.9)	47.4 (43.7, 51.1)	<0.001	32.9 (22.3, 43.5)	30.4 (26.3, 34.6)	31.5 (27.3, 35.6)	35.2 (30.3, 40.2)	0.27

Other coronary risk factors										
BMI (m/kg^2)	28.3 (27.8, 28.7)	27.9 (27.6, 28.2)	27.0 (26.5, 27.5)	26.5 (26.2, 26.8)	<0.001	27.4 (26.3, 28.4)	26.4 (26.0, 26.7)	26.5 (26.2, 26.8)	26.8 (26.5, 27.2)	0.60
Waist hip ratio	0.828 (0.822, 0.833)	0.819 (0.815, 0.823)	0.809 (0.803, 0.816)	0.813 (0.807, 0.818)	<0.001	N/A	N/A	N/A	N/A	N/A
Standing height (mm)	1582.1 (1577.3, 1587.0)	1589.5 (1585.7, 1593.3)	1592.7 (1586.4, 1599.0)	1593.5 (1589.0, 1597.9)	<0.001	1704 (1688, 1719)	1713 (1706, 1719)	1710 (1703, 1716)	1717 (1710, 1724)	0.21
Leg length (mm)	753.6 (750.5, 756.8)	757.5 (754.9, 760.0)	760.5 (756.3, 764.6)	760.4 (757.4, 763.4)	0.001	795 (784, 806)	802 (798, 807)	800 (796, 804)	802 (797, 807)	0.67
Current smoker (%)	10.0 (7.9, 12.6)	10.2 (8.5, 12.1)	9.3 (6.6, 12.9)	13.7 (11.3, 16.5)	0.21	26.3 (16.3, 36.3)	41.5 (37.1, 46.0)	42.0 (37.6, 46.4)	44.3 (39.2, 49.4)	0.04
Adult head of household manual social class (%)	72.4 (68.8, 75.8)	63.8 (60.8, 66.6)	55.2 (49.9, 60.5)	46.8 (43.1, 50.6)	<0.001	82.3 (72.7, 91.9)	85.8 (82.2, 89.5)	86.2 (82.7, 89.7)	86.3 (82.3, 90.3)	0.57
Child manual social class (%)	84.2 (81.1, 86.8)	82.2 (79.7, 84.4)	78.5 (73.8, 82.6)	73.3 (69.9, 76.5)	<0.001	68.4 (57.9, 79.0)	62.1 (57.7, 66.5)	71.6 (67.6, 75.6)	62.7 (57.8, 67.7)	0.88

[$] p-values are for linear trend across groups (1df)

CHD = incident and prevalent myocardial infarction / angina/ angioplasty / CABG

MI = incident and prevalent WHO verified myocardial infarction only

Table 3: Associations of *ADH1C* variants, risk factors and coronary heart disease risk in women and men who drink moderate amounts (for BWHHS = those drinking >=3.5 units per week on average, N = 895)

WOMEN					MEN			
ADL1C variants among those drinking at least 3 ½ units a week N = 895					*ADL1C* variants among those drinking at least 7 units a week N = 757			
	γ1γ1	γ1γ2	γ2γ2	P trend	γ1γ1	γ1γ2	γ2γ2	P trend
HDLc (mmol/l)	1.84 (1.79, 1.90)	1.81 (1.76, 1.86)	1.78 (1.71, 1.86)	0.20	1.08 (1.05, 1.12)	1.06 (1.03, 1.09)	1.05 (1.01, 1.09)	0.20
SBP (mmHg)	146.8 (144.1, 149.5)	145.1 (142.5, 147.6)	149.1 (145.5, 152.8)	0.49	147.3 (144.5, 150.1)	147.1 (144.7, 149.4)	147.9 (144.2, 151.7)	0.85
MI	1.6 (0.7, 3.8)	2.5 (1.3, 4.5)	3.4 (1.5, 7.3)	0.22	12.0 (8.0, 16.0)	8.9 (5.9, 11.8)	15.1 (9.2, 20.9)	0.57
CHD (%)	13.1 (9.5, 17.3)	16.6 (13.1, 20.6)	19.8 (14.2, 26.4)	0.05	N/A	N/A	N/A	

CHD = incident and prevalent myocardial infarction / angina/ angioplasty / CABG
MI = incident and prevalent WHO verified myocardial infarction only

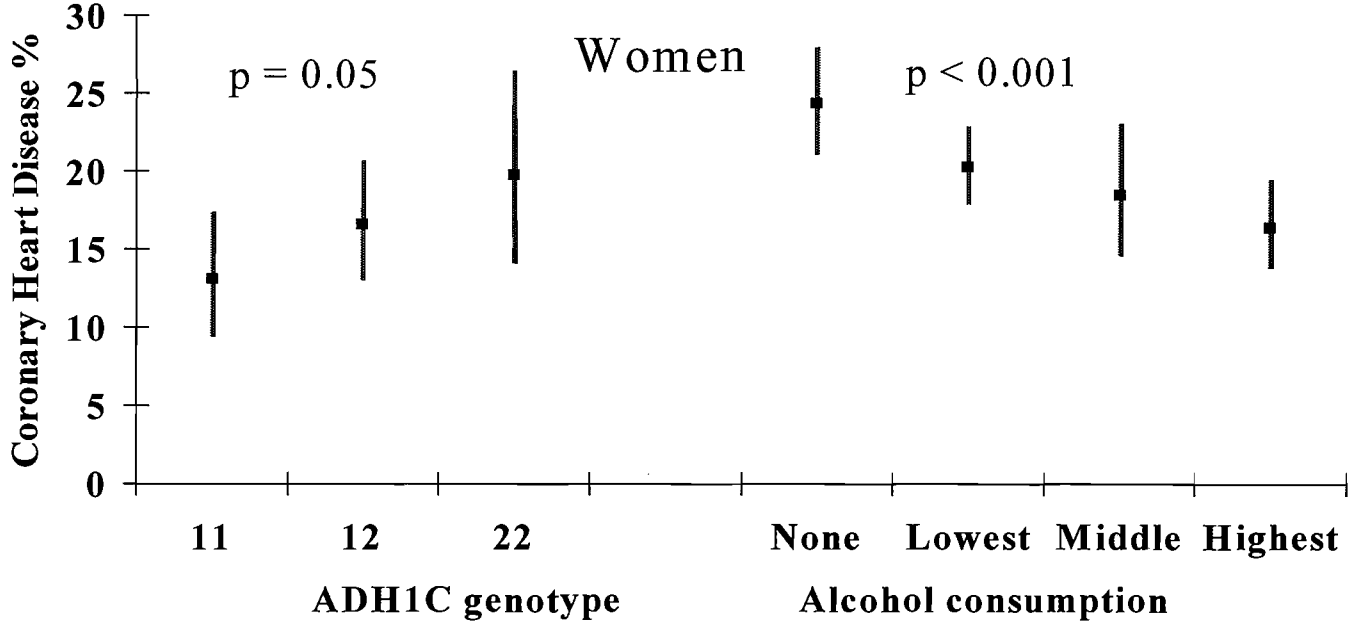
Table 4: Association of genotype with HDL-cholesterol and CHD in women who have ever and never used hormone replacement

	Mean or % (95% CI) of outcome by genotype and hormone replacement therapy			P trend across genotype	P interaction
HDL-c	γ1γ1	γ1γ2	γ2γ2		
Ever used HRT N = 629	1.71 (1.64, 1.77)	1.70 (1.65, 1.75)	1.63 (1.54, 1.72)	0.18	
Never used HRT N = 2605	1.66 (1.63, 1.69)	1.64 (1.61, 1.66)	1.65 (1.60, 1.69)	0.40	0.34
CHD	γ1γ1	γ1γ2	γ2γ2		
Ever used HRT N = 629	20.4 (15.3, 26.3)	17.1 (13.0, 21.8)	14.7 (8.6, 22.7)	0.19	
Never used HRT N = 2605	18.7 (16.2, 21.5)	22.8 (20.6, 25.2)	23.9 (20.0, 28.3)	0.02	0.07

Table 5: Summary findings from studies of ADH1C genotype and associations with coronary heart disease and HDL-cholesterol

Study	Crude odds ratio γ2γ2 vs. γ1γ1 for coronary heart disease	HDL-cholesterol (HDL-C) effect of genotype, evidence of genotype-alcohol interaction on HDL-C, (p for interaction)
Physicians' and Nurses' Health Studies (n=1166), Hines, 2001	0.68 (0.46, 1.01)	Higher HDL-C in men (alcohol-genotype interaction p=0.05) and women (alcohol-genotype interaction p=0.02) who drink and are γ2γ2.
Australian men and women (n=901), Whitfield, 2003	No data	No difference in HDL-C by genotype and no evidence of alcohol-genotype interaction (p=0.69)
Framingham Offspring Study (n=1805), Djoussé, 2005	0.72 (0.42, 1.23)	No difference in HDL-C by genotype, and no evidence of alcohol-genotype interaction (p=0.37)
Northwick Park Study (n=2773), Younis, 2005	0.91 (0.58, 1.42)	No difference in HDL-C by genotype and no evidence of alcohol-genotype interaction
Health Professional Follow Up Study, Nurses' Health Study 2, and Nurses' Health Study (n=1817), Hines, 2005	No data	No difference in HDL-C by genotype in any of the four samples studied. No evidence of alcohol-genotype interaction (men p = 0.26, women p = 0.50). Evidence of alcohol-genotype interaction in pooled analysis of men and women not taking postmenopausal hormones (p=0.02)
British Women's Heart & Health Study (n=3234) and Caerphilly men Study (n=1313), present study	Women: 1.20 (0.92, 1.56) Men 1.11 (0.69, 1.77)	Women: no difference in HDL-C by genotype and no evidence of alcohol-genotype interaction (p=0.73) Men: No difference in HDL-C by genotype but weak evidence of alcohol-genotype interaction with higher HDL-C in abstainers who are γ2γ2. (p=0.04)

Figure 1a and 1b. Risk of coronary heart disease by ACH1C genotype and alcohol consumption: a) women in the British Women's Heart & Health Study and b) men in the Caerphilly study



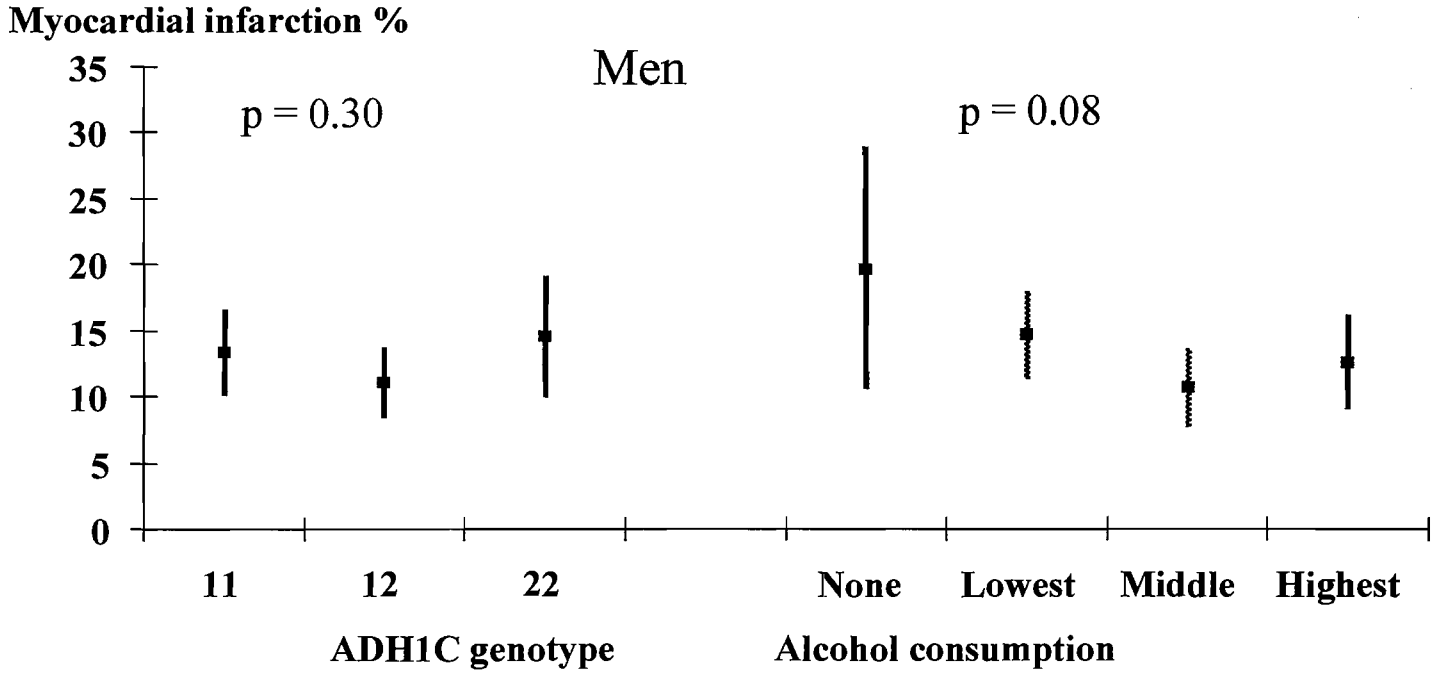
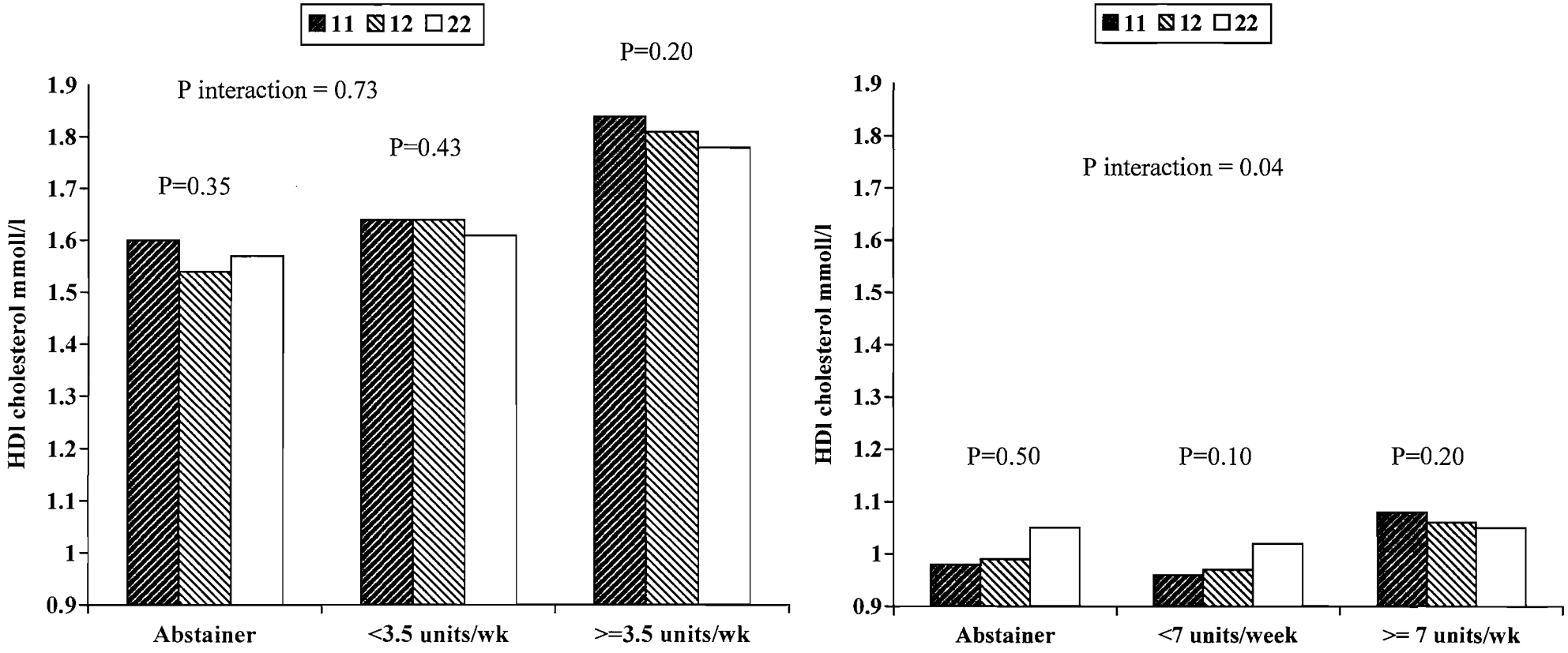

Myocardial infarction %
Men
p = 0.30
p = 0.08
35
30
25
20
15
10
5
0
11
12
22
None
Lowest
Middle
Highest
ADH1C genotype
Alcohol consumption

Figure 2: HDL-cholesterol levels by ADH1C genotype and alcohol consumption in women (left panel) and men (right panel)



[1] Shaper AG. Editorial: alcohol, the heart, and health. Am J Public Health 1993;83:799-801.

[2] Marmot M. Commentary: Reflections on alcohol and coronary heart disease. Int J Epidemiol 2001;30:729-34.

[3] Bovot P, Paccaud F. commentary: Alcohol, coronary heart disease and public health: which evidence-based policy? Int J Epidemiol 2001;30:734-37.

[4] Klatsky AL. Could abstinence from alcohol be hazardous to your health? Int J Epidemiol 2001;30:739-42.

[5] Hart CL, Davey Smith G, Hole DJ, Hawthorne VM. Alcohol consumption and mortality from all causes, coronary heart disease, and stroke: results from a prospective cohort study of Scottish men with 21 years of follow up. BMJ 1999;318:1725-29.

[6] Rimm EB, Williams P, Fosher K, Criqui M, Stampfer MJ. Moderate alcohol intake and lower risk of coronary heart disease: meta-analysis of effects on lipids and haemostatic factors. BMJ 1999;319; 1523-8

[7] Rimm E. Alcohol and coronary heart disease – laying the foundation for future work. Int J Epidemiol 2001;30:738-39.

[8] Puddey I, Beilin L, Vandongen R, Rouse I, Rogers P. Evidence for a direct effect of alcohol consumption on blood pressure in normotensive men. A randomized controlled trial. Hypertension 1985;7:707-713

[9] Puddey I, Beilin L, Vandongen R. Regular alcohol use raises blood pressure in treated hypertensive subjects. Lancet 1987;i:647-51

[10] Eriksson CJP, Fukunaga T, Sarkola T, Chen WJ, Chen CC, Ju JM, Cheng ATA, Yamamoto H, Kohlenberg-Müller K, Kimura M, Murayama M, Matsushita S, Kashima H, Higuchi S, Carr L, Viljoen D, Brooke L, Stewart T, Foroud T, Su J, Li T-K, Whitfield JB. Functional relevance of human ADH polymorphism. *Alcohol Clin Exp Res* 2001;25:157S-163S.

[11] Hines LM. Genetic modification of the effect of alcohol consumption on CHD. Proc Nutrition Soc 2004;63:73-79

[12] Agarwal DP. Genetic polymorphisms of alcohol metabolizing enzymes. *Pathol Biol* 2001;49:703-9.

[13] Djousse L. Levy D. Herbert AG. Wilson PW. D'Agostino RB. Cupples LA. Karamohamed S. Ellison RC. Influence of alcohol dehydrogenase 1C polymorphism on the alcohol-cardiovascular disease association (from the Framingham Offspring Study) *American Journal of Cardiology*.2005; 96:227-32

[14] Bosron WF, Lumeng L, Li TK. Genetic polymorphism of enzymes of alcohol metabolism and susceptibility to alcoholic liver disease. Mol Aspects Med 1988;10:147-158

[15] Whitfield JB, Nightingale BN, Bucholz KK, Madden PAF, Heath AC, Martin NG. ADH Genotypes and alcohol use and dependence in Europeans. *Alcohol Clin Exp Res* 1998;22:1463-1469.

[16] Hines LM, Stampfer MJ, Ma J, Gaziano JM, Ridker PM, Hankinson SE, Sacks F, Rimm EB, Hunter DJ. Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N Engl J Med* 2001;344(8):549-55.

[17] Willett WC. Balancing Life-Style and Genomics Research for Disease Prevention. Science 2002; 296: 695-698

[18] Hines LM, Hunter DJ, Stampfer MJ et al. Alcohol consumption and high-densitry lipoprotein levels: the effect of *ADH1C* genotype, gender and menopausal status. Atherosclerosis 2005;182:293-300

[19] Davey Smith, G. & Ebrahim, S. 'Mendelian Randomisation': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32:**1-22

[20] Trinder P. Determination of blood glucose using 4-amino phenazone as oxygen acceptor. J Clin Pathol 1969; 22(2):246

[21] Andersen L, Dinesen B, Jorgensen PN, Poulsen F, Roder ME. Enzyme immunoassay for intact human insulin in serum or plasma. Clin Chem 1993; 39(4):578-582.

[22] Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985; 28(7):412-419

[23] Lawlor DA, Ebrahim S, Davey Smith G. The metabolic syndrome and coronary heart disease in older women: findings from the British Women's Heart and Health Study. Diabetic Medicine 2004; 21:906-913

[24] Lawlor DA, Bedford C, Taylor M, Ebrahim S. Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. J Epidemiol Community Health 2003; 57(2):134-140.

[25] Anonymous. Caerphilly and Speedwell collaborative heart disease studies. The Caerphilly and Speedwell Collaborative Group. *J.Epidemiol.Community Health* 1984;**38**:259-262

[26] Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988; 16(3):1215

[27] Cheung VG, Nelson SF. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America.* 1996;**93**:14676-14679.

[28] Gaunt TR, Hinks LJ, Christensen MB, Kiessling M, Day INM. Experience applying LightTyper methodology to human SNPs relevant to growth and cardiovascular risk. Day INM (ed) in Genetic Variance Detection: technologies for pharmacogenomics (ed Karl Hecker) DNA Press (in press 09-2005)

[29] Guo,S.W.,Thompson,E.A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles Biometrics 1992;48:361-72

[30] Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:722-729

[31] Davey Smith, G. Lawlor DA, Harbord RM et al. Association of C-reactive protein with blood pressure and hypertension: lifecourse confounding and Mendelian randomisation tests of causality. Atheroscler Thromb Vasc Biol 2005:25:1051-6

[32] Baum CF, Schaffer ME, Stillman S. Instrumental Variables and GMM: Estimation and Testing. The Stata Journal 2003; 3: 1-31

[33] Staiger D, Stock JH. Instrumental Variables Regression with Weak Instruments. *Econometrica: Journal of the Econometric Society* 1997;65:557-586

[34] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genetics 2001;29:306-309

[35] Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet. 2003;361:865-72

[36] Davey Smith G, Ebrahim S, Lewis S, Hansell A, Palmer LJ, Burton P. Genetic epidemiology and public health: hope, hype and future prospects. Lancet (in press)

[37] Whitfield JB. O'Brien ME. Nightingale BN. Zhu G. Heath AC. Martin NG. ADH genotype does not modify the effects of alcohol on high-density lipoprotein. *Alcoholism: Clinical & Experimental Research. 2003;27:509-14*

[38] Younis J, Cooper JA, Miller GJ, Humphries SE, Talmud PJ. Genetic variation in alcohol dehydrogenase 1C and the beneficial effect of alcohol intake on coronary heart disease risk in the Second Northwick Park Heart Study. Atherosclerosis. 2005; 180:225-32

[39] Zee RY, Ricker PM, Cook NR. Prospective evaluation of the alcohol dehydrongenase γ1γ2 gene polymorphism and risk of stroke. Stroke 2004;35:e39-e42

[40] Wannamethee SG, Shaper AG, Ebrahim S. HDL-Cholesterol, total cholesterol, and the risk of stroke in middle-aged British men. *Stroke* 2000;**31**:1882-88

[41] Nakamura Y. Amamoto K. Tamaki S. Okamura T. Tsujita Y. Ueno Y. Kita Y. Kinoshita M. Ueshima H. Genetic variation in aldehyde dehydrogenase 2 and the effect of alcohol consumption on cholesterol levels. *Atherosclerosis. 2002;164:171-7*

[42] Mizoi Y, Yamamoto K, Ueno Y, Fukunaga T, Harada S. Involvement of genetic polymorphism of alcohol and aldehyde dehydrogenases in individual variation of alcohol metabolism. Alcohol Alcoholism 1994;29:707-710

[43] Whitfield JB. ADH and ALDH genotypes in relation to alcohol metabolic rate and sensitivity. Alcohol Alcoholism 1994;2: suppl 59-65

[44] Whifield JB, Zhu G, Duffy DL, Birley AJ, Madden PA, Heath AC, Martin NG. Variation in alcohol pharmacokinetics as a risk factor for alcohol dependence. Alcohol Clin Exp Res 2001;25:1257-1263

[45] Chen HJ, Tian H, Edenberg HJ. Natural haplotypes in the regulatory sequences affect human alcohol dehydrogenase 1C (ADH1C) gene expression. Hum Mutat. 2005; 25:150-5

[46] Enomoto N, Takase S, Yasuhara M, Takada A. Acetaldehyde metabolism in different aldehyde dehydrogenase-2 genotypes. *Alcohol Clin Exp Res* 1991;15:141-4.

[47] Takagi S, Iwai N, Yamauchi R, Kojima S, Yasuno S, Baba T, terashima M, Tsutsumi Y, Suzuki S, Morii I, Hanai S, Ono K, baba S, Tomoike H, Kawamura A, Miyazaki S, Nonogi H, Goto Y. Aldehyde dehydrogenase 2 gene is a risk factor for myocardial infarction in Japanese men. Hypertens Res 2002;25:677-681.

[48] Davey Smith, G. & Ebrahim, S. Mendelian Randomisation: prospects, pitfalls and limitations. Int J Epidemiol 2004;33, 30-42.

[49] Davey Smith G, Harbord R, Ebrahim S. Fibrinogen, C-reactive protein and CHD: does Mendelian randomization suggest the associations are non-causal? QJM 2004;97:163-6

References

1. Knopp, R.H. (2002) Risk factors for coronary artery disease in women. *Am J Cardiol,* **89,** 28E-34E; discussion 34E-35E.
2. Ellsworth, D.L., Sholinsky, P., Jaquish, C., Fabsitz, R.R. and Manolio, T.A. (1999) Coronary heart disease. At the interface of molecular genetics and preventive medicine. *Am J Prev Med,* **16,** 122-33.
3. Daley, G.Q. and Cargill, M. (2001) The heart SNPs a beat: polymorphisms in candidate genes for cardiovascular disease. *Trends Cardiovasc Med,* **11,** 60-6.
4. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* **33 Suppl,** 228-37.
5. Hirschhorn, J.N., Lohmueller, K., Byrne, E. and Hirschhorn, K. (2002) A comprehensive review of genetic association studies. *Genet Med,* **4,** 45-61.
6. Cambien, F., Poirier, O., Lecerf, L., Evans, A., Cambou, J.P., Arveiler, D., Luc, G., Bard, J.M., Bara, L., Ricard, S. *et al.* (1992) Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. *Nature,* **359,** 641-4.
7. Colhoun, H.M., McKeigue, P.M. and Davey Smith, G. (2003) Problems of reporting genetic associations with complex outcomes. *Lancet,* **361,** 865-72.
8. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet,* **33,** 177-82.
9. Lohmueller, K.E., Peace, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-annalysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *nature genetics.*
10. Mathews, C. and Holde, K. (1990) *Biochemistry.* The Benjamin/Cummings Publishing company, Inc.
11. Gu, W., Li, M., Zhao, W.M., Fang, N.X., Bu, S., Frazer, I.H. and Zhao, K.N. (2004) tRNASer(CGA) differentially regulates expression of wild-type and codon-modified papillomavirus L1 genes. *Nucleic Acids Res,* **32,** 4448-61.
12. Nakamura, Y. and Tabata, S. (1997) Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes. *Microb Comp Genomics,* **2,** 299-312.
13. Oba, T., Andachi, Y., Muto, A. and Osawa, S. (1991) Translation in vitro of codon UGA as tryptophan in Mycoplasma capricolum. *Biochimie,* **73,** 1109-12.
14. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet,* **22,** 239-47.
15. Brookes, A.J. (1999) The essence of SNPs. *Gene,* **234,** 177-86.
16. Chakravarti, A. (1999) Population genetics--making sense out of sequence. *Nat Genet,* **21,** 56-60.
17. Keavney, B., McKenzie, C.A., Connell, J.M., Julier, C., Ratcliffe, P.J., Sobel, E., Lathrop, M. and Farrall, M. (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet,* **7,** 1745-51.

18. Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. and Donnelly, P. (2005) A haplotype map of the human genome. *Nature,* **437,** 1299-320.
19. Griffiths, A.J.F., Gelbart, W.M., Miller, J.H. and Lewontin, R.C. (1999) 17. Population and Evolutionary Genetics. In *Modern Genetic Analysis*
W. H. FREEMAN, New York

20. Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C. and Gelbart, W.M. (2000) *An Introduction to Genetic Analysis. Seventh Edition.*
21. Cardon, L.R. and Palmer, L.J. (2003) Population stratification and spurious allelic association. *Lancet,* **361,** 598-604.
22. Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nat Genet,* **36,** 512-7.
23. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science,* **296,** 2225-9.
24. Cardon, L.R. and Abecasis, G.R. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet,* **19,** 135-40.
25. (2007) International HapMap Project. http://www.hapmap.org.
26. (2007) http://www.ensembl.org.
27. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature,* **445,** 881-5.
28. Hines, L.M., Stampfer, M.J., Ma, J., Gaziano, J.M., Ridker, P.M., Hankinson, S.E., Sacks, F., Rimm, E.B. and Hunter, D.J. (2001) Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N Engl J Med,* **344,** 549-55.
29. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet,* **74,** 106-20.
30. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science,* **307,** 1072-9.
31. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W. *et al.* (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science,* **316,** 889-94.
32. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* **447,** 661-78.
33. Samani, N.J., Erdmann, J., Hall, A.S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R.J., Meitinger, T., Braund, P., Wichmann, H.E. *et al.* (2007) Genomewide association analysis of coronary artery disease. *N Engl J Med,* **357,** 443-53.
34. Ozaki, K., Inoue, K., Sato, H., Iida, A., Ohnishi, Y., Sekine, A., Sato, H., Odashiro, K., Nobuyoshi, M., Hori, M. *et al.* (2004) Functional variation in

LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro. *Nature,* **429,** 72-5.
35. Eichner, J.E., Dunn, S.T., Perveen, G., Thompson, D.M., Stewart, K.E. and Stroehla, B.C. (2002) Apolipoprotein E polymorphism and cardiovascular disease: a HuGE review. *Am J Epidemiol,* **155,** 487-95.
36. Zeggini, E. and Weedon, M.N. and Lindgren, C.M. and Frayling, T.M. and Elliott, K.S. and Lango, H. and Timpson, N.J. and Perry, J.R. and Rayner, N.W. and Freathy, R.M. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science,* **316,** 1336-41.
37. Hearst, M.A. (1999) Untangling Text Data mining. *Proceedings of ACL'99.*
38. Yeh, A.S., Hirschman, L. and Morgan, A.A. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics,* **19 Suppl 1,** i331-9.
39. Hirschman, L., Park, J.C., Tsujii, J., Wong, L. and Wu, C.H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics,* **18,** 1553-61.
40. Page, D. and Craven, M. Biological applications of Multi-Relational Data Mining. *SIGKDD Explorations.*
41. Marcotte, E.M., Xenarios, I. and Eisenberg, D. (2001) Mining literature for protein-protein interactions. *Bioinformatics,* **17,** 359-63.
42. Regev, Y., Finkelstein-Landau, M., Feldman, R., Gorodetsky, M., Zheng, X., Levy, S., Charlab, R., Lanwrence, C., Lippert, R.A., zhang, Q. *et al.* (2002) Rule-based Extraction of experimental evidence in the Biomedical Domain- the KDD Cup 2002. *SIGKDD Explorations,* **4,** 90-92.
43. Keerthi, S.S., Ong, C.J., Siah, K.B., Lim, D.B., Chu, W., Shi, M., Edwin, D.S., Menon, R., Shen, L., Lim, J.Y.K. *et al.* (2002) A Machine Learning Approach for the Curation of Biomedical Literature--KDD Cup 2002. *SIGKDD Explorations,* **4,** 93-94.
44. Ghanem, M.M., Guo, Y., Lodhi, H. and Zhang, Y. (2002) Automated Scientific text Classification Using Local Patterns: KDD CUP 2002. *SIGKDD Explorations,* **4,** 95-96.
45. Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat Genet,* **31,** 316-9.
46. Freudenberg, J. and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics,* **18 Suppl 2,** S110-5.
47. Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol,* **4,** R75.
48. Hu, Y., Hines, L.M., Weng, H., Zuo, D., Rivera, M., Richardson, A. and LaBaer, J. (2003) Analysis of genomic and proteomic data using advanced literature mining. *J Proteome Res,* **2,** 405-12.
49. Agerholm-Larsen, B., Nordestgaard, B.G. and Tybjaerg-Hansen, A. (2000) ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites. *Arterioscler Thromb Vasc Biol,* **20,** 484-92.
50. Keavney, B., McKenzie, C., Parish, S., Palmer, A., Clark, S., Youngman, L., Delepine, M., Lathrop, M., Peto, R. and Collins, R. (2000) Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases

and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. *Lancet,* **355,** 434-42.
51. Klerk, M., Verhoef, P., Clarke, R., Blom, H.J., Kok, F.J. and Schouten, E.G. (2002) MTHFR 677C-->T polymorphism and risk of coronary heart disease: a meta-analysis. *Jama,* **288,** 2023-31.
52. Wald, D.S., Law, M. and Morris, J.K. (2002) Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. *Bmj,* **325,** 1202.
53. Talmud, P.J. and Stephens, J.W. (2004) Lipoprotein lipase gene variants and the effect of environmental factors on cardiovascular disease risk. *Diabetes Obes Metab,* **6,** 1-7.
54. Thompson, J.F., Lira, M.E., Durham, L.K., Clark, R.W., Bamberger, M.J. and Milos, P.M. (2003) Polymorphisms in the CETP gene and association with CETP mass and HDL levels. *Atherosclerosis,* **167,** 195-204.
55. Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y. *et al.* (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet,* **32,** 650-4.
56. Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001) Human disease genes. *Nature,* **409,** 853-5.
57. Libby, P. (2002) Inflammation in atherosclerosis. *Nature,* **420,** 868-74.
58. Libby, P., Ridker, P.M. and Maseri, A. (2002) Inflammation and atherosclerosis. *Circulation,* **105,** 1135-43.
59. Posch, P.E., Cruz, I., Bradshaw, D. and Medhekar, B.A. (2003) Novel polymorphisms and the definition of promoter 'alleles' of the tumor necrosis factor and lymphotoxin alpha loci: inclusion in HLA haplotypes. *Genes Immun,* **4,** 547-58.
60. Grundy, S.M., Brewer, H.B., Jr., Cleeman, J.I., Smith, S.C., Jr. and Lenfant, C. (2004) Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Arterioscler Thromb Vasc Biol,* **24,** e13-8.
61. Hayakawa, T., Nagai, Y., Taniguchi, M., Yamashita, H., Takamura, T., Abe, T., Nomura, G. and Kobayashi, K. (2000) Tumor necrosis factor-beta gene NcoI polymorphism decreases insulin resistance in Japanese men. *Metabolism,* **49,** 1506-9.
62. Paul, N.L. and Ruddle, N.H. (1988) Lymphotoxin. *Annu Rev Immunol,* **6,** 407-38.
63. Locksley, R.M., Killeen, N. and Lenardo, M.J. (2001) The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell,* **104,** 487-501.
64. Schneider, K., Potter, K.G. and Ware, C.F. (2004) Lymphotoxin and LIGHT signaling pathways and target genes. *Immunol Rev,* **202,** 49-66.
65. Goetz, F.W., Planas, J.V. and MacKenzie, S. (2004) Tumor necrosis factors. *Developmental & Comparative Immunology,* **28,** 487-497.
66. Dejardin, E., Droin, N.M., Delhase, M., Haas, E., Cao, Y., Makris, C., Li, Z.W., Karin, M., Ware, C.F. and Green, D.R. (2002) The lymphotoxin-beta receptor induces different patterns of gene expression via two NF-kappaB pathways. *Immunity,* **17,** 525-35.
67. Hayden, M.S. and Ghosh, S. (2004) Signaling to NF-kappaB. *Genes Dev,* **18,** 2195-224.

68. Yang, R.Y. and Liu, F.T. (2003) Galectins in cell growth and apoptosis. *Cell Mol Life Sci,* **60,** 267-76.
69. Sturm, A., Lensch, M., Andre, S., Kaltner, H., Wiedenmann, B., Rosewicz, S., Dignass, A.U. and Gabius, H.J. (2004) Human galectin-2: novel inducer of T cell apoptosis with distinct profile of caspase activation. *J Immunol,* **173,** 3825-37.
70. Yuan, X.M., Brunk, U.T. and Hazell, L. (2000) The morphology and natural history of aherosclerosis. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford University press, Oxford.
71. Bannon, p., James, N. and Jessup, W. (2000) The endothelial cell in atherosclerosis. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford university press, Oxford.
72. Campbell, G., Bingley, J., Hayward, I. and Campbell, J. (2000) Smooth muscle cells and the connective tissue matrix of the intima. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford university Press, Oxford.
73. Kritharides, L. and Jessup, W. (2000) Macrophage lipid metabolism and atherosclerosis. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford University press, Oxford.
74. Libby, P. (2002) Inflammation in atherosclerosis. *NATURE,* **420,** 868-874.
75. Kritharides, L. and Redgrave, T.G. (2000) Lipoprotein influx and efflux in the atherosclerotic intima. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford University press, Oxford.
76. Libby, P., Ridker, P.M. and Maseri, A. (2002) Inflammation and Atherosclerosis. *Circulation,* **105,** 1135-1143.
77. Keaney, J.F. and Freedman, J.E. (2000) Oxidative stress and platelet function in atherosclerosis. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford university press, Oxford.
78. Eaton, J. and Dean, R.T. (2000) Diabetes and atherosclerosis. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. Oxford University press, Oxford.
79. Dean, R.T. and Kelly, D.T. (2000) *Atherosclerosis*. Oxford University Press.
80. Berg, J., Tymoczko, J. and Stryer, L. The Integration of Metabolism. In *Biochemistry.* fifth edition ed. http://www.ncbi.nlm.nih.gov/books/.
81. Fruhbeck, G., Gomez-Ambrosi, J., Muruzabal, F.J. and Burrell, M.A. (2001) The adipocyte: a model for integration of endocrine and metabolic signaling in energy metabolism regulation. *Am J Physiol Endocrinol Metab,* **280,** E827-47.
82. Ruan, H. and Lodish, H.F. (2003) Insulin resistance in adipose tissue: direct and indirect effects of tumor necrosis factor-alpha. *Cytokine Growth Factor Rev,* **14,** 447-55.
83. Ruan, H., Miles, P.D., Ladd, C.M., Ross, K., Golub, T.R., Olefsky, J.M. and Lodish, H.F. (2002) Profiling gene transcription in vivo reveals adipose tissue as an immediate target of tumor necrosis factor-alpha: implications for insulin resistance. *Diabetes,* **51,** 3176-88.
84. Hotamisligil, G.S., Shargill, N.S. and Spiegelman, B.M. (1993) Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. *Science,* **259,** 87-91.
85. Uysal, K.T., Wiesbrock, S.M., Marino, M.W. and Hotamisligil, G.S. (1997) Protection from obesity-induced insulin resistance in mice lacking TNF-alpha function. *Nature,* **389,** 610-4.

86. Hotamisligil, G.S., Johnson, R.S., Distel, R.J., Ellis, R., Papaioannou, V.E. and Spiegelman, B.M. (1996) Uncoupling of obesity from insulin resistance through a targeted mutation in aP2, the adipocyte fatty acid binding protein. *Science*, **274,** 1377-9.
87. Sonnenberg, G.E., Krakower, G.R. and Kissebah, A.H. (2004) A novel pathway to the manifestations of metabolic syndrome. *Obes Res,* **12,** 180-6.
88. Hotamisligil, G.S., Peraldi, P., Budavari, A., Ellis, R., White, M.F. and Spiegelman, B.M. (1996) IRS-1-mediated inhibition of insulin receptor tyrosine kinase activity in TNF-alpha- and obesity-induced insulin resistance. *Science*, **271,** 665-8.
89. Ruan, H., Hacohen, N., Golub, T.R., Van Parijs, L. and Lodish, H.F. (2002) Tumor necrosis factor-alpha suppresses adipocyte-specific genes and activates expression of preadipocyte genes in 3T3-L1 adipocytes: nuclear factor-kappaB activation by TNF-alpha is obligatory. *Diabetes,* **51,** 1319-36.
90. Kras, K.M., Hausman, D.B. and Martin, R.J. (2000) Tumor necrosis factor-alpha stimulates cell proliferation in adipose tissue-derived stromal-vascular cell culture: promotion of adipose tissue expansion by paracrine growth factors. *Obes Res,* **8,** 186-93.
91. Cuff, C.A., Sacca, R. and Ruddle, N.H. (1999) Differential induction of adhesion molecule and chemokine expression by LTalpha3 and LTalphabeta in inflammation elucidates potential mechanisms of mesenteric and peripheral lymph node development. *J Immunol,* **162,** 5965-72.
92. Gluckman, P.D. and Hanson, M.A. (2004) The developmental origins of the metabolic syndrome. *Trends Endocrinol Metab,* **15,** 183-7.
93. Yamada, A., Ichihara, S., Murase, Y., Kato, T., Izawa, H., Nagata, K., Murohara, T., Yamada, Y. and Yokota, M. (2004) Lack of association of polymorphisms of the lymphotoxin alpha gene with myocardial infarction in Japanese. *J Mol Med,* **82,** 477-83.
94. (2004) A trio family study showing association of the lymphotoxin-alpha N26 (804A) allele with coronary artery disease. *Eur J Hum Genet,* **12,** 770-4.
95. Iwanaga, Y., Ono, K., Takagi, S., Terashima, M., Tsutsumi, Y., Mannami, T., Yasui, N., Goto, Y., Nonogi, H. and Iwai, N. (2004) Association analysis between polymorphisms of the lymphotoxin-alpha gene and myocardial infarction in a Japanese population. *Atherosclerosis,* **172,** 197-8.
96. Hamid, Y.H., Urhammer, S.A., Glumer, C., Borch-Johnsen, K., Jorgensen, T., Hansen, T. and Pedersen, O. (2005) The common T60N polymorphism of the lymphotoxin-alpha gene is associated with type 2 diabetes and other phenotypes of the metabolic syndrome. *Diabetologia,* **48,** 445-51.
97. Bayley, J.P., Ottenhoff, T.H. and Verweij, C.L. (2004) Is there a future for TNF promoter polymorphisms? *Genes Immun,* **5,** 315-29.
98. Knight, J.C., Keating, B.J., Rockett, K.A. and Kwiatkowski, D.P. (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet,* **33,** 469-75.
99. Knight, J.C., Keating, B.J. and Kwiatkowski, D.P. (2004) Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1. *Nat Genet,* **36,** 394-9.
100. Lawlor, D.A., Bedford, C., Taylor, M. and Ebrahim, S. (2003) Geographical variation in cardiovascular disease, risk factors, and their control in older

women: British Women's Heart and Health Study. *J Epidemiol Community Health,* **57,** 134-40.
101. Lawlor, D.A., Ebrahim, S. and Davey Smith, G. (2002) The association between components of adult height and Type II diabetes and insulin resistance: British Women's Heart and Health Study. *Diabetologia,* **45,** 1097-106.
102. Christensen, M.B., Lawlor, D.A., Gaunt, T.R., Howell, M.W., Davey Smith, G., Ebrahim, S. and Day, I.N. (2006) Genotype of galectin 2 (LGALS2) is associated with insulin-glucose profile in the British Women's Heart and Health Study. *Diabetologia,* **49,** 673-7.
103. Lawlor, D.A., Ebrahim, S. and Davey Smith, G. (2004) The metabolic syndrome and coronary heart disease in older women: findings from the British Women's Heart and Health Study. *Diabet Med,* **21,** 906-13.
104. Miller, S.A., Dykes, D.D. and Polesky, H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res,* **16,** 1215.
105. Cheung, V.G. and Nelson, S.F. (1996) Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc Natl Acad Sci U S A,* **93,** 14676-9.
106. Gaunt, T.R., Hinks, L.J., Christensen, M.B., Kiessling, M. and Day, I.N. (2005) Experience Applying LightTyper™ Methodology to Human SNP's Relevant to Growth and Cardiovascular Risk. In Hecker, K. (ed.), *Genetic Variance Detection: technologies for pharmacogenomics*. DNA Press.
107. Dupont, W.D. and Plummer, W.D., Jr. (1990) Power and sample size calculations. A review and computer program. *Control Clin Trials,* **11,** 116-28.
108. Dupont, W.D. and Plummer, W.D., Jr. (1998) Power and sample size calculations for studies involving linear regression. *Control Clin Trials,* **19,** 589-601.
109. Jousilahti, P., Tuomilehto, J., Vartiainen, E., Pekkanen, J. and Puska, P. (1996) Body weight, cardiovascular risk factors, and coronary mortality. 15-year follow-up of middle-aged men and women in eastern Finland. *Circulation,* **93,** 1372-9.
110. NCBI (2005) Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack. *http://hin.nhlbi.nih.gov/atpiii/calculator.asp*.
111. UK, P. (2005) Primary Cardiovascular Risk Calculator. *http://www.patient.co.uk/showdoc/40000133/*.
112. (2005) http://www.promega.com/tbs/tm033/tm033.pdf.
113. Timpson, N.J., Christensen, M., Lawlor, D.A., Gaunt, T.R., Day, I.N., Ebrahim, S. and Davey Smith, G. (2005) TAS2R38 (phenylthiocarbamide) haplotypes, coronary heart disease traits, and eating behavior in the British Women's Heart and Health Study. *Am J Clin Nutr,* **81,** 1005-11.
114. Smith, G.D. and Ebrahim, S. (2004) Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol,* **33,** 30-42.
115. Patel, A. and Keech, A. (2000) Antioxidant intervention studies in humans. In Dean, R.T. and Kelly, D.T. (eds.), *Atherosclerosis*. oxford university press, Oxford.

116. Davey Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol,* **32,** 1-22.
117. (1984) Caerphilly and Speedwell collaborative heart disease studies. The Caerphilly and Speedwell Collaborative Group. *J Epidemiol Community Health,* **38,** 259-62.
118. NCBI (2006) Whole Genome Association. *http://www.ncbi.nlm.nih.gov/WGA/*.