Investigation of the mechanisms mediating genetic susceptibility to cardiovascular disease on chromosomes 9p21 and 2q24

Michael Sheridan Cunnington

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Medicine

Newcastle University
Faculty of Medical Sciences
Institute of Human Genetics
April 2010

Abstract

Recent genome-wide studies have identified novel loci associated with cardiovascular diseases, but the mechanisms mediating these associations are unknown. Investigation of intermediate phenotypes can identify the pathways involved and potential targets for therapeutic intervention. This study investigated the relationship with intermediate phenotypes at risk loci on chromosome 9p21 and 2q24.

Chromosome 9p21 polymorphisms are associated with coronary artery disease and congenital intracranial aneurysms. In the present study risk variants were not associated with traditional risk factors, inflammatory mediators, carotid artery intima-media thickness, echocardiographic measures of cardiac structure and function, or congenital heart defects. There was no evidence of copy number variation using MLPA.

To identify genes involved in mediating disease susceptibility this study examined the association of chromosome 9p21 variants with peripheral blood expression in healthy subjects of three neighbouring genes: two cyclin-dependent kinase inhibitors, *CDKN2A* and *CDKN2B*, and a non-coding RNA of unknown function, *ANRIL*. Novel methodology combining allelic expression data from multiple transcribed markers was more powerful than total expression analysis for mapping *cis*-acting effects. Multiple loci were independently associated with expression of each gene, suggesting that several sites may modulate disease susceptibility. Disease-associated variants were all associated with allelic expression of *ANRIL*, while association with the other two genes was only detectable for some risk variants. Variants had an inverse effect on *ANRIL* and *CDKN2B* expression, supporting a role of antisense transcription in *CDKN2B* regulation. This study suggests that modulation of *ANRIL* expression mediates susceptibility to several important human diseases.

Chromosome 2q24 polymorphisms were associated with hypertension in a study involving an Amish population; *in vitro* experiments suggested that influences on *STK39* expression might mediate these effects. In the present study allelic expression analysis confirmed that reported SNPs were associated with *STK39* expression *in vivo*, but were not associated with blood pressure in a large British cohort.

Acknowledgements

I would like to thank the many people who have contributed to this research.

Thanks must firstly go to my supervisors Professor Bernard Keavney and Dr Mauro Santibanez Koref for all their help, enthusiasm and expertise, without which this research would not have been possible.

I am grateful to the British Cardiovascular Society/Swire, the Newcastle Healthcare Charity, and the British Heart Foundation for funding the project. Thanks also to my collaborators, especially Professor Bongani Mayosi who made it possible for me to travel to Cape Town to collect the samples from the South African cohort, and Professor Sir John Burn, who provided the samples from the northeast Caucasian cohort. The assistance of the volunteers who contributed samples to the collections used in this project is greatly appreciated.

I received guidance and training from many people at the Institute of Human Genetics and am very grateful to everyone who gave up their time to help. In particular, Dr Darroch Hall for general laboratory and Sequenom training, Dr Valerie Wilson for training in allelic expression analysis, and Dr Peter Avery for statistical assistance. Dr Mauro Santibanez Koref designed and wrote the custom programmes that were used for the simulations and allelic expression analyses, Dr Darroch Hall assisted with genotyping of the congenital heart disease samples, and Mr Chris Kay assisted with the *STK39* genotyping.

Thanks also to the BHF lab group for their friendship, and finally, to my wife Jo and children Ollie and Lexie whose love and patience I could not do without.

The work in this thesis, except where otherwise specified, is entirely my own, and I have not submitted it previously for a degree in this or any other institution.

Dr Michael Cunnington
British Cardiovascular Society/Swire Research Fellow

Publications

Cunnington MS, Mayosi BM, Hall DH, Avery PJ, Farrall M, Vickers MA, Watkins H, Keavney B. Novel genetic variants linked to coronary artery disease by genome-wide association are not associated with carotid artery intima-media thickness or intermediate risk phenotypes. *Atherosclerosis*. 2009;203(1):41-44.

Cunnington MS, Kay C, Avery PJ, Mayosi BM, Santibanez Koref M, Keavney B. *STK39* polymorphisms and blood pressure: an association study in British Caucasians and assessment of *cis*-acting influences on gene expression. *BMC Med Genetics*. 2009;10:135.

Cunnington MS, Keavney BD. Genetics of coronary heart disease. In *Evidence Based Cardiology*. 3rd edition. Yusuf S, Cairns JA, Camm AJ, Fallen EL, Gersh BJ (eds). 2010; BMJ publishing Ltd, Chichester.

Cunnington MS, Santibanez Koref M, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 SNPs associated with multiple disease phenotypes correlate with *ANRIL* expression. *PLoS Genetics*. 2010;6(4): e1000899

Santibanez Koref M, Wilson V, Cartwright N, **Cunnington MS**, Mathers J, Bishop T, Curtis A, Dunlop MG, Burn J. *MLH1* differential allelic expression in mutation carriers and controls. *Annals of Human Genetics*. 2010;74:479-488.

Cunnington MS, Keavney BD. Genetic mechanisms mediating atherosclerosis susceptibility at the chromosome 9p21 locus. *Current Atherosclerosis Reports* – in press.

Publications in preparation

Cunnington MS, Mayosi BM, Burn J, Keavney B, Santibanez Koref M. Comparing total and allelic expression for mapping *cis*-acting polymorphisms. Manuscript in preparation.

Cunnington MS, Topf A, Hussain R, Keavney B. Chromosome 9p21 polymorphisms associated with coronary artery disease and intracranial aneurysm are not associated with congenital heart disease. Manuscript in preparation.

National and international presentations

British Cardiovascular Society Annual Conference 2010. Cunnington MS, Santibanez Koref M, et al. Modulation of *ANRIL* Expression may mediate the Association between Chromosome 9p21 variants and Coronary Atherosclerosis Risk (**Young research worker's prize runner-up**).

European Atherosclerosis Society Congress 2010. Cunnington MS, Santibanez Koref M, et al. Modulation of *ANRIL* expression may mediate the association of chromosome 9p21 variants with coronary artery disease and stroke (poster discussion).

American Heart Association Annual Scientific Sessions 2009. Cunnington MS, Santibanez Koref M, et al. Modulation of *ANRIL* Expression is a Possible Mechanism Mediating the Association between Chromosome 9p21 Polymorphisms and Coronary Atherosclerosis Risk (oral presentation).

British Cardiovascular Society Annual Conference 2009. Cunnington MS, Mayosi BM, et al. Genetic polymorphisms linked to coronary artery disease are not associated with carotid artery intima-media thickness, left ventricular size, or intermediate risk phenotypes (moderated poster).

British Atherosclerosis Society Spring Meeting 2008. Cunnington MS, Mayosi BM, et al. Genetic variants linked to coronary artery disease are not associated with carotid intima-media thickness or intermediate risk markers (poster).

Table of contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

List of tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Appendix 1

List of abbreviations

µ	micro
A	adenine
AEI	allelic expression imbalance
aeQTL	allelic expression quantitative trait locus
AER	allelic expression ratio
ANRIL	antisense noncoding RNA in the INK4 locus
ARF	alternative reading frame transcript of CDKN2A
B2M	beta-2-microglobulin
BLAST	basic local alignment search tool
BMI	body mass index
bp	base pairs
BP	blood pressure
BSA	body surface area
^{o}C	degrees Celsius
C	cytosine
CAD	coronary artery disease
CDCV	common disease common variant
CDKN2A	cyclin-dependent kinase 2A
CDKN2B	cyclin-dependent kinase 2B
CHANGE study	Congenital Hearts: A National Gene Environment study
cDNA	complementary deoxyribonucleic acid
CDRV	common disease rare variant
CEPH	Centre d'Etude du Polymorphisme Humain Caucasian cohort
CEU	Caucasian (CEPH) HapMap cohort
CI	confidence interval
CIMT	carotid artery intima-media thickness
CNV	copy number variation
CRP	C-reactive protein
Ct	cycle threshold for real-time PCR analysis
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
EDTA	ethylenediaminetetraacetic acid
EST	expressed sequence tag
eQTL	expression quantitative trait locus
FCH study	Freeman Congenital Heart study
FH	familial hypercholesterolaemia
FLAP	5-lipoxygenase activating protein
FWER	family-wise error rate
g	gram
g	gravitational force
G	guanine
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gDNA	genomic DNA
GWA	genome-wide association
HDL	high density lipoprotein
HEK293	human embryonic kidney 293 cell line
HeLa	immortalised cell line derived from cervical cancer cells
HPRT1	hypoxanthine phosphoribosyltransferase 1
HUVEC	human umbilical vein endothelial cells
hME	homogeneous mass extend
HWE	Hardy-Weinberg equilibrium
HTO cohort	hypertension origins family study cohort

IL-6	interleukin-6
kb	kilobase
L	litre
LD	linkage disequilibrium
LDL	low-density lipoprotein
LDLR	low-density-lipoprotein receptor
LOD score	logarithm of the odds score
LPO	left primer oligo
LTA	lymphotoxin-α
LV	left ventricle
LVIDd	left ventricular internal diameter diastole
LVIDs	left ventricular internal diameter systole
M	moles/molar
MAF	minor allele frequency
MI	myocardial Infarction
min	minute
MALDI-TOF	matrix-assisted laser desorption/ionisation time-of-flight
MLH1	mutL homolog 1
MLPA	multiplex ligation-dependent probe amplification
MTAP	methylthioadenosine phosphorylase
NCBI	National Center for Biotechnology Information
NE Caucasian	northeast Caucasian cohort
nm	nanometre
OR	odds ratio
PCR	polymerase chain reaction
PCSK9	proprotein convertase subtilisin/kexin type 9
RACE	rapid amplification of cDNA ends
ROC	receiver operating characteristic
RNA	ribonucleic acid
rpm	revolutions per minute
RPO	right primer oligo
RT	reverse transcription
SA cohort	South African cohort
SAP	shrimp alkaline phosphatase
sec	second
SNP	single nucleotide polymorphism
SPAK	Ste20-related proline-alanine-rich kinase
STK39	serine threonine kinase 39
T	thymidine
TAE	Tris-acetate-EDTA electrophoresis buffer
TE	Tris-EDTA buffer
TOF	tetralogy of Fallot
TNF-α	tumour necrosis factor-α
UK	United Kingdom
USA	United States of America
UTR	untranslated region
V	volts
VSD	ventricular septal defect
WGA	whole genome amplification
WHR	waist-hip ratio
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan Nigeria HapMap cohort

Chapter 1

Introduction

1 Introduction

1.1 Genetics of coronary artery disease

Coronary artery disease (CAD) is the leading cause of death and morbidity worldwide and is rapidly increasing in incidence in nations undergoing industrial development[1, 2]. Understanding the genetic basis of CAD will increase our knowledge of the pathophysiological processes involved, with the ultimate clinical goals of improving risk-stratification of individuals and offering novel targets for therapeutic and preventative interventions[3].

1.1.1 Evidence for genetic susceptibility to CAD

CAD is a complex phenotype which arises from the interaction of genetic and environmental risk factors including smoking, hyperlipidemia, hypertension, obesity, and diabetes[4]. However, evidence from a range of epidemiological studies has consistently identified a family history of CAD as an independent risk factor for disease. The 1994 report from the longitudinal Swedish Twin Study included 21,004 individuals, of whom 2,810 had fatal CAD[5]. Among male twin pairs in which the first twin had died of CAD before the age of 55, the relative risk of fatal CAD in the second twin was 8.1 (95% CI 2.7 – 24.5) for monozygotic (i.e. genetically identical) twins and 3.8 (95% CI 1.4 – 10.5) for dizygotic twins. Among female twin pairs in which the first twin had died of CAD before the age of 65, the relative risk of fatal CAD in the second twin was 15.0 (95% CI 7.1 – 31.9) for monozygotic twins and 2.6 (95% CI 1.0 – 7.1) for dizygotic twins. These results indicate a significant genetic contribution to the risk of CAD death. A 2002 analysis of the same cohort, which included 4,007 CAD deaths and a follow-up time of up to 36 years, used more sophisticated statistical modelling approaches to conclude that the heritability of fatal CAD events was 57% for males and 38% for females, and that the genetic risk of CAD death persisted even into old age[6]. A similar analysis among 7,955 Danish twin pairs including 2,476 CAD deaths was broadly concordant with this data, reporting a heritability of fatal CAD events of 53% for males and 58% for females[7].

Although genetic factors are clearly associated with CAD risk, twin and family genetic studies selecting for populations with premature-onset CAD tend to over-

estimate the risk applicable to a more general population with atherosclerosis. Studies in the offspring of the original Framingham Heart Study participants have shown that a family history of cardiovascular disease remains a significant independent predictor of disease after multivariable adjustment for other measured risk factors[8, 9]. In these studies, the relative risk associated with parental or sibling cardiovascular disease was between 1.45 and 2.0. A similar estimate was provided by the INTERHEART genetic case-control study[4]. This included 15,152 cases with MI and 14,820 controls from 52 countries in populations derived from every inhabited continent, and identified nine modifiable risk factors (smoking, dyslipidemia, hypertension, diabetes, abdominal obesity, psychosocial factors, daily consumption of fruit and vegetables, regular alcohol consumption, and regular physical activity) that were all significantly related to the risk of myocardial infarction (MI). A family history of CAD also showed an independent association with MI, conferring a relative risk of 1.45 (95% CI 1.31 – 1.60) after adjustment for the nine other risk factors described above. The population attributable risk associated with a family history of CAD was 9.8%, but when family history was added to the information from the other risk factors, the overall population attributable risk rose from 90.4% to only 91.4%, suggesting that most of the associated risk burden can be accounted for through other risk factors. Many of these risk factors, such as hypertension and hypercholesterolaemia, are themselves under a similarly moderate degree of genetic control and investigations have been unable to find evidence for single 'major genes' with appreciable population frequency exerting individual large effects[10].

Some studies have examined the heritability of quantitative phenotypes related to atherosclerotic disease, such as carotid artery intima-media thickness (CIMT) and arterial calcification[11-13]. These phenotypes have also been shown to have substantial heritabilities, in the order of 30-60%, which provides further evidence that genetic factors are important in conferring susceptibility to atherosclerosis and its consequences.

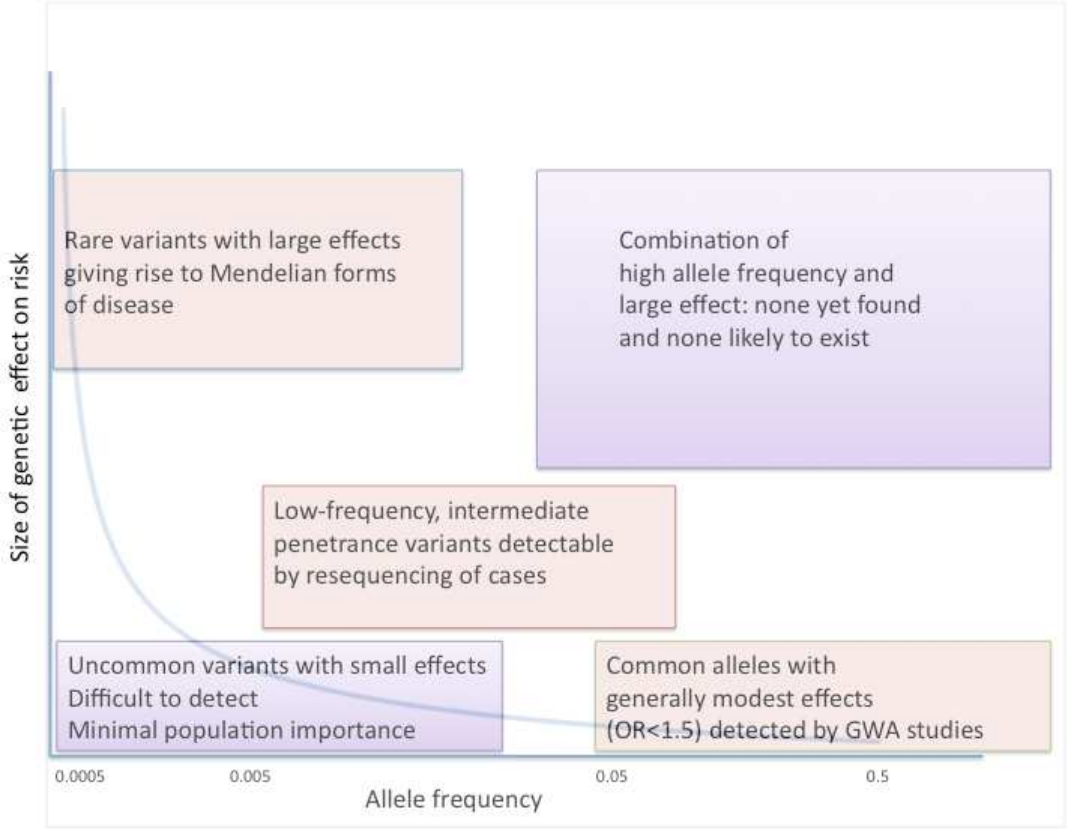
1.1.2 Genetic architecture of CAD susceptibility

Although CAD susceptibility does appear to have a substantial genetic component, identifying the genetic elements involved in causation is complicated by its multifactorial nature, involving a complex interplay between multiple genetic loci and

multiple environmental modifiers. Many genes have been implicated in conferring CAD susceptibility[14], but even with the newly-discovered loci from recent genome-wide association (GWA) studies that are discussed in detail below, there remains a substantial 'missing heritability'– that is, the fraction of the heritability of CAD that is accounted for by confirmed loci to date is much lower than the total calculated heritability of the condition[3, 15]. Furthermore, the exact architecture of that component, in terms of the number of genetic effects and the size of each, remains uncertain at present. Population genetic theory predicts that common alleles are likely to have small biological effects, while large effects are confined to rare alleles (as illustrated in Figure 1.1). As discussed below, a small number of Mendelian (single gene) disorders with high penetrance have been recognised for CAD. Although such mutations confer very markedly increased risks for affected individuals, they are rare and their individual contribution to the population burden of disease is modest. There has been considerable debate about the nature of the genetic effects conferring most of the susceptibility to CAD and other complex diseases at the population level. The 'common disease common variant' (CDCV) hypothesis argues that common polymorphisms each with relatively low penetrance are the major contributors to genetic susceptibility, whereas the 'common disease rare variant' (CDRV) hypothesis argues that many individually rare variants, each of relatively high penetrance, are the major contributors[16]. Common variants tend to be evolutionarily ancient, but it has been argued that since CAD is a late onset disease and is unlikely to have influenced reproductive fitness, CAD risk variants may not have been subject to negative selection pressure and may therefore persist with relatively high population frequencies[16].

Figure 1.1. Effect size and allele frequency in populations.

Both population genetic theory and empirical evidence suggest that the frequency of an allele will vary in the opposite direction to the size of its biological effect (blue curve). Three general classes of detectable variants (red boxes) can be distinguished; different approaches are needed to detect members of the different classes. Figure reproduced from Cunnington and Keavney[17].



Recent findings of multiple common variants associated with CAD and other complex diseases in GWA studies lend support the CDCV hypothesis, with common risk alleles associated with 20-30% increases in CAD risk[18-25]. Perhaps the best support for the CDRV hypothesis with respect to CAD comes from studies investigating mutations in the *PCSK9* gene. *PCSK9* regulates the availability of LDL receptors on the cell surface; activating mutations reduce the number of receptors and give rise to autosomal dominant hypercholesterolaemia[26], whilst nonsense mutations present in 2.6% of African Americans increase the number of receptors, resulting in a 28% reduction in plasma LDL and 88% reduction in CAD risk[27]. The effects on risk associated with the low frequency intermediate penetrance variants in *PCSK9* are much larger than those associated with the common variants identified in GWA studies, but the variants are much less prevalent in the population.

Far from being an academic debate, the degree to which CAD susceptibility is determined by common low penetrance variants (as proposed by the CDCV hypothesis) and low frequency intermediate penetrance variants (as proposed by the CDRV hypothesis) is important since different experimental strategies are required for the identification of each. Common variants may be detected in association studies involving genotyping of polymorphisms throughout the genome in very large numbers of cases and controls, exploiting the fact that associations can be detected for polymorphisms that are in linkage disequilibrium (LD) with the functionally-important variants[28]. However, because of their low frequency and individually small contribution to overall genetic susceptibility, rare variants may not be detectable by association studies based on the use of linked polymorphic markers, even in very large studies. These rare variants are usually discovered by extensive resequencing of carefully selected candidate genes in relatively large numbers of carefully selected cases[28], although other methods for screening for such effects have been proposed[29].

It is likely that multiple common and rare variants of varying effects together account for the overall heritability of CAD that has been observed at population levels, meaning that an integrated approach will be required to identify the majority of the variants and pathways involved in conferring disease susceptibility. Recent studies have largely focused on the identification of common variants, not least because this was technologically feasible with the development of ultra high-throughput genotyping platforms at a time when next generation sequencing technologies were not widely available[30, 31]. Traditional Sanger sequencing was too slow and costly to permit large-scale screening of very large numbers of genes[32], but recent technological innovations that have dramatically increased the speed and cost effectiveness of genome sequencing will enable systematic genome-wide evaluation of rare variants in the near future[33].

The discussion below will focus on the approaches used to try and identify genetic loci conferring susceptibility to CAD to date, and highlight the major findings from such studies. Before doing this, however, a number of insights from population genetic studies which laid the foundations for the framework which has ultimately resulted in the recently published GWA studies, and which are important for the subsequent work presented in this thesis, will be reviewed.

1.1.3 Concepts and tools in population genetics

1.1.3.1 Single nucleotide polymorphisms and the human haplotype map

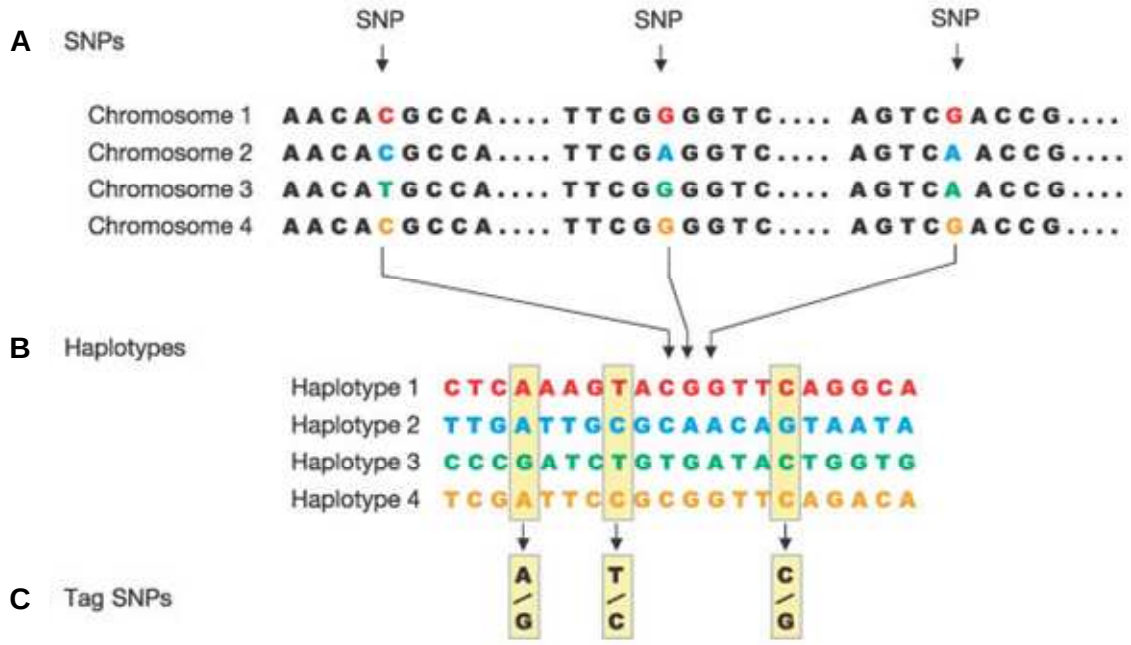
In 1990 the ambitious Human Genome Project set out to sequence the entire human genome, and a draft of the complete sequence was reported in 2001[34, 35]. The large-scale sequencing studies conducted to achieve this established that much of the sequence variation among individuals was accounted for by common variants such as single nucleotide polymorphisms (SNPs)[34, 36, 37], although the frequency and importance of copy number variants (CNVs) had not been fully appreciated at that time[38, 39]. SNPs are due to a single base difference between chromosomes at a particular site, as illustrated in Figure 1.2. There are estimated to be at least 10 million SNPs in the human genome with minor allele frequencies of at least 1% – that is, on average, one SNP roughly every 300 bases[37]. Proponents of the CDCV hypothesis argued that common variants with modest effects were likely to contribute to many common diseases, and models demonstrated that the statistical power to detect disease associations was greater for population based association testing than family based linkage studies[40]. It was therefore argued that common genetic variants in the human genome should be identified to enable their use as markers for association testing[40, 41].

The genome sequence characterised by the Human Genome Project provided the reference framework necessary for the identification and accurate cataloguing of human polymorphisms. A systematic discovery programme for SNPs was implemented by The SNP Consortium, a public-private partnership, which was formed in 1999. The consortium released 1.4 million SNPs into the public domain by 2001, and there are presently more than 7 million SNPs in public databases[42-44]. This constitutes a fairly complete picture of human common SNP variation. However, simple knowledge of the SNPs did not make the task of assessing variation across the genome possible, since no platform exists that could type 7 million SNPs in each of a large number of cases and controls in a cost-effective manner. It was therefore necessary to additionally determine the relationships between the SNPs that had been discovered, in order to establish which SNPs gave redundant or substantially overlapping information about genetic variation in different regions of the genome, and establish a minimal set of SNPs that could be feasibly typed in case-control

studies while retaining relatively full information. Such data was provided by the collaborative International Haplotype Map (HapMap) Project, formed in 2002, which aimed to obtain detailed genomewide information on patterns of common genetic variation in multiple world populations, that could be used for association studies[36, 45, 46].

Figure 1.2. SNPs, haplotypes and tag SNPs.

A, SNPs. Shown is a short stretch of DNA from four versions of the same chromosome region in different people. Most of the DNA sequence is identical in these chromosomes, but three bases are shown where variation occurs. Each SNP has two possible alleles; the first SNP in panel A has the alleles C and T. B, Haplotypes. A haplotype is made up of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6,000 bases of DNA. Only the variable bases are shown, including the three SNPs that are shown in panel A. For this region, most of the chromosomes in a population survey turn out to have haplotypes 1–4. C, Tag SNPs. Genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely. For instance, if a particular chromosome has the pattern A–T–C at these three tag SNPs, this pattern matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population. Figure reproduced from The International HapMap Project[36].



HapMap characterised 270 individuals from four geographically diverse populations: 30 mother-father-adult child trios of the Yoruba people in Ibadan Nigeria (YRI); 30 trios of northern and western European ancestry living in Utah USA (CEU); 45 unrelated Han Chinese individuals in Beijing China (CHB); 45 unrelated Japanese individuals in Tokyo Japan (JPT) – with CHB and JPT populations pooled to form an

Asian analysis panel[36, 46]. This allowed genetic diversity between these world-wide populations to be captured and compared in detail. Phase I of the HapMap project published in 2005 genotyped approximately 1.3 million SNPs[45], and Phase II which was published in October 2007 characterised over 3.1 million SNPs[46]. An online database provides free access to this data[47]. The HapMap data has been instrumental in the development of methods for the design and analysis of GWA studies. The fundamental concepts of LD and haplotype structure which underpin the HapMap project and the design of GWA studies are reviewed below.

1.1.3.2 Linkage disequilibrium and haplotype diversity

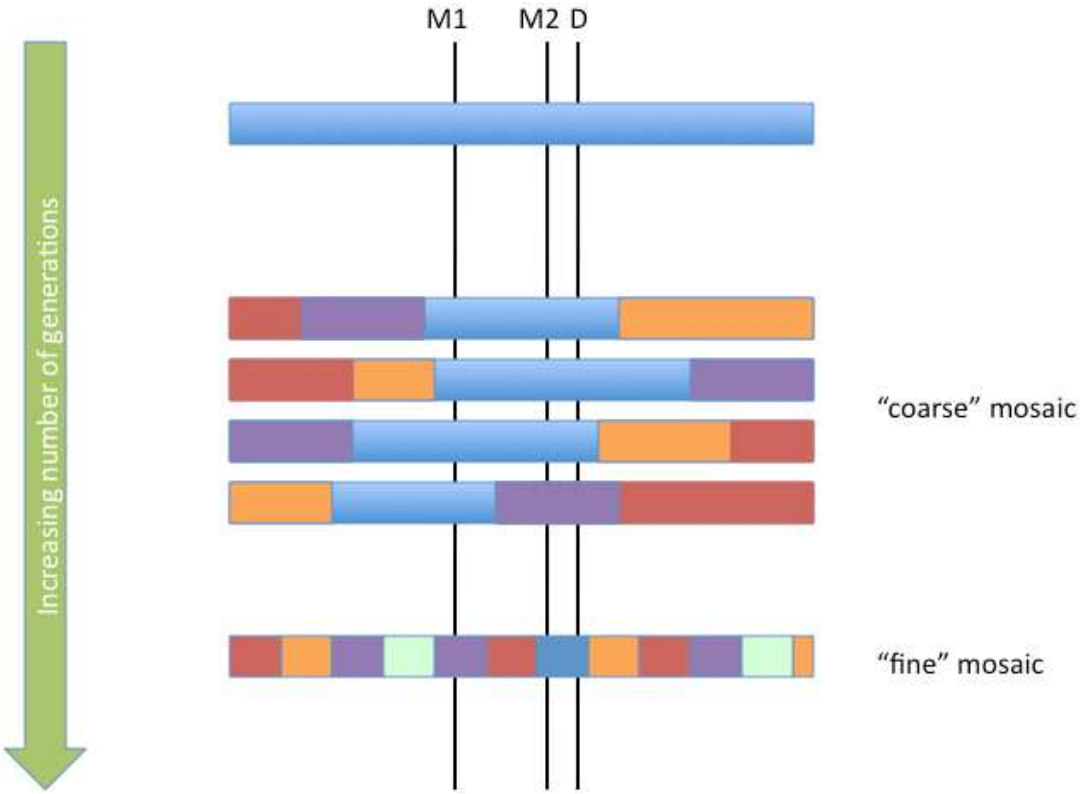
Most SNPs arose from single historical mutation events, and are therefore associated with nearby variants that were present on the ancestral chromosome on which the mutation occurred[36]. These variants which are inherited together on a single chromosome are known as a haplotype, as illustrated in Figure 1.2. The co-inheritance of these alleles produces statistical associations between these alleles in a population, which is known as linkage disequilibrium (LD). SNPs which are more strongly associated with each other are referred to as having higher LD. For the purposes of association mapping, LD is most usefully measured by the correlation coefficient r^2 which varies between 0 in the case of two SNPs whose genotypes have no predictive value for each other, and 1 in the case of two SNPs that are perfect proxies for each other (and therefore one of which is redundant for the purposes of association studies).

The association between such alleles can be disrupted by new mutations or recombination events occurring during meiosis, which in general is more likely to occur the greater the distance is between the two alleles (as illustrated in Figure 1.3). On average, therefore, LD between SNPs tends to decrease the further apart they are. However, recombination events do not occur with equal frequency throughout the genome, but tend to occur repeatedly around certain recombination 'hotspots'[48, 49]. The overall result is that the human genome is structured in 'LD blocks', each of which contains alleles in high LD with one another, with adjacent blocks separated by recombination hotspots resulting in lower LD between different blocks. The situation is more complex than this simple model implies since population history influences

the number, size and frequency of such blocks in human populations. However, the practical value of LD blocks is that genotyping only certain SNPs in a region is sufficient to predict information about the other SNPs nearby, thus the common variation can be captured by genotyping a limited number of 'tag' SNPs[50, 51], as illustrated in Figure 1.2.

Figure 1.3. Changes in haplotype structure and LD between alleles over generations.

Consider a new disease susceptibility allele D arising on an ancestral chromosome (blue bar). In that generation, all the markers in the region including markers M1 and M2 will be correlated with D. Genetic recombination will break down the region of association over time, introducing new haplotypes into the region. If the mosaic is 'coarse' (which is the case with the human genome) fewer markers are needed to tag each piece ('haplotype block') of the mosaic: thus, M1 and M2 are both still good markers for D when the mosaic is coarse, but only M2 is a good marker after further time has elapsed and the mosaic is finer. However, the situation is complicated by population history, since founder populations and population bottlenecks reduce haplotype diversity. Populations of African ancestry therefore have a finer mosaic than Caucasian populations. Figure reproduced from Cunnington and Keavney[17].



These associations between SNPs mean that an 'indirect' approach can be used for performing association studies, since tag SNP markers can detect association between a locus and disease irrespective of whether the tag SNP itself has a functional effect[36, 52]. The length of haplotype blocks varies, but the mean size is estimated to be approximately 22kb in populations of European or Asian ancestry, and 11kb in populations of African ancestry (defining a haplotype block as a region over which <5% of comparisons between SNP pairs showed evidence of historical recombination)[53]. Based on these data, it has been estimated that fully powered haplotype association studies could require up to 300,000 to 1,000,000 tag SNPs (in non-African and African samples, respectively), although these are upper limits since LD between blocks permits fewer markers to be used without loss of power[53].

The HapMap project has characterised in great detail the location and associations between SNPs, providing novel data about the haplotype block structure in different human populations. This enables tag SNPs to be selected for genotyping studies, aids in phasing of genotype data, and provides the potential to infer missing genotypes, which may improve the design and analysis of GWA studies[46]. Some authors have criticised the HapMap approach of using indirect assessment of SNPs markers for GWA testing, arguing that assumptions may not hold in practice, such that under some circumstances this approach could fail to detect associations with disease even with infinite sample sizes[54]. However, the associations and novel insights provided by even the first round of GWA studies are regarded by many as a testament to the success of this approach[25].

Such studies have been made possible by the conceptual insights and collaborative frameworks outlined above, but also by the technological advances that have enabled such high-throughput genotyping to be reliably performed.

1.1.3.3 Genotyping technologies

The theoretical advantages of GWA studies for detecting common variants influencing disease susceptibility had been recognised for some time[40], but the limiting factor to achieving them in reality was the availability of high-throughput, accurate, cost-effective genotyping platforms. Many methods were investigated for

suitability, but until recently none had proven capable of the massive parallel throughput required[31]. Developments in high-density oligonucleotide microarrays succeeded in achieving these goals and have demonstrated success in identifying risk loci in the first wave of genome-wide association analyses in complex diseases including CAD. The Affymetrix GeneChip Human Mapping 500K Array simultaneously types approximately 500,000 SNPs and was used in some of the first reported GWA studies for CAD[18, 21]. The Illumina human Hap300 array, which types 318,000 SNPs, was also used in early GWA studies for CAD[19]. In contrast to the Affymetrix GeneChip 500K, the choice of SNPs was selected using an LD-based selection strategy, in an attempt to capture the maximum amount of variation across the genome with fewer SNPs. The details of these technologies are beyond the scope of this chapter, but have been recently reviewed in detail[31]. Newer products from these major manufacturers type even greater numbers of SNPs and tag up to 90% of genome-wide SNPs in non-African populations and also allow assessment of other forms of common genetic variation such as copy number variation (CNV)[55], which is discussed in more detail in 1.3.3.3.

1.1.4 Approaches to studying genetic susceptibility to CAD

A number of approaches have been used to assess the relationship between genetic factors and CAD susceptibility. Many studies look for direct evidence of association between genetic markers and presence of the CAD phenotype itself. The methodologies used for this are described below. However, studies may also investigate the association of genetic markers with intermediate CAD phenotypes. The main disadvantage of such an approach is that intermediate phenotypes are usually not perfect proxies for disease, and must be selected with great care to avoid creating spurious associations or missing true associations. However, the advantages of this approach are that intermediate phenotypes are often easier to ascertain, and they may give important insights into the mechanisms and pathways involved in conferring susceptibility to disease. Pathways identified in this way may provide targets for therapeutic modulation aimed at reducing disease risk.

1.1.4.1 Mendelian disorders

A number of Mendelian disorders associated with premature CAD have been recognised, as summarised in Table 1.1. Such disorders explain only a very small proportion of CAD cases, but knowledge of them is important for a number of reasons. First, it allows the development of effective strategies for the detection, treatment and prevention of disease in affected individuals and families, which is particularly important as disease in these individuals often presents at a young age and follows an aggressive clinical course. Second, certain genes causing Mendelian forms of CAD have also been implicated in the risk of non-Mendelian CAD in the general population (highlighted in Table 1.1). Third, novel insights into the pathophysiological mechanisms of disease gained through the study of Mendelian disorders may allow the identification of new biomarkers or targets for therapeutic intervention, as illustrated in the examples below.

Familial hypercholesterolemia (FH) is the commonest Mendelian disorder conferring CAD risk. It is an autosomal dominant condition caused by mutations in the low-density lipoprotein receptor (*LDLR*) gene, which disrupt the ability of the receptor to bind LDL at the cell membrane. Cellular uptake of LDL cholesterol is impaired, resulting in increased circulating levels of LDL cholesterol and stimulation of intracellular cholesterol synthesis[56-59]. FH heterozygotes have elevated circulating LDL cholesterol levels and develop tendon xanthomata, corneal arcus and premature coronary artery disease often before the age of 50 years. Affected homozygotes are more severely affected, often with LDL cholesterol above 8.5mmol/L, cutaneous xanthomata, and coronary disease which may present below the age of 30 years. Untreated, the risk of CAD in affected heterozygotes by age 60 years is approximately 50-85% for men and 30-55% for women[60, 61]. Over 1000 mutations in the *LDLR* gene have been reported to be associated with the syndrome. Point mutations account for 91% of mutations and 9% are major rearrangements[62]. The heterozygote frequency of familial forms of hypercholesterolemia is around 1 in 500 in Western populations.

Table 1.1. Mendelian disorders involving CAD.

Disorder	Inheritance	Affected genes	Consequences of gene mutation	Prevalence & clinical features
Familial hypercholesterolemia (OMIM# 143890)	Autosomal dominant	*LDLR**	*LDLR* encodes the cell membrane LDL receptor. More than 1000 mutations have been reported in the LDLR gene, which disrupt the ability of the LDL receptor to bind and endocytose LDL.[62] This results in accumulation of LDL cholesterol in the circulation.	Heterozygote frequency 1/500. Heterozygotes have elevated circulating LDL levels, tendon xanthomata, corneal arcus, and premature CAD. Homozygotes more severely affected with higher LDL levels, cutaneous xanthomata, and CAD may present below the age of 30 years.
Familial defective Apo B (OMIM# 144010)	Autosomal dominant	*APOB**	*APOB* encodes apolipoprotein B-100, which is the protein component of LDL. Apo B-100 is the major ligand for the LDL receptor and mediates binding of LDL to the receptor. 10 mutations reported in APOB, which reduce the binding affinity of Apo B-100 to the LDL receptor, resulting in impaired cellular uptake and processing of LDL.	Heterozygote frequency 1/1000. Elevated circulating LDL with phenotype very similar to FH, but slightly milder in some studies. Risk of premature CAD is lower than with FH, but remains substantially elevated.
Autosomal dominant hypercholesterolemia 3 (OMIM# 603776)	Autosomal dominant	*PCSK9**	*PCSK9* encodes proprotein convertase subtilisin/kexin type 9. The exact function of this glycoprotein is not well understood, but several gain-of-function mutations seem to be associated with a reduction in the number of LDL receptors at the cell surface.	Studied in several pedigrees. Phenotype similar to FH, but prevalence and variations in phenotype not yet well defined.
Autosomal recessive hypercholesterolemia (OMIM# 603813)	Autosomal recessive	*ARH*	*ARH* encodes LDL receptor adaptor protein 1 (LDLRRAP1). This interacts with the cytoplasmic domain of the LDL receptor and components of the clathrin endocytic machinery. Mutations lead to defective endocytosis of the LDL receptor.	Rare. Raised circulating LDL levels with phenotype similar to FH.
Apolipoprotein A-I (Apo A-I) deficiency (OMIM# 107680)	Autosomal dominant	*APOA1**	*APOA1* encodes Apo A-I, which is the major protein constituent of HDL. It mediates the interaction of HDL with cell surface receptors and other pathways. Mutations lead to very low circulating HDL levels.	Rare. Heterozygotes have low circulating HDL levels and some variants are associated with premature CAD risk[63, 64] whilst others are not.[65] Homozygotes with certain mutations may have xanthomata or corneal opacities and early CAD.
Tangier disease (OMIM# 205400)	Autosomal recessive	*ABCA1**	*ABCA1* encodes ATP-binding cassette A1. This acts as a cholesterol efflux pump, involved in cellular cholesterol homeostasis and HDL formation. Mutations lead to lipid accumulation within cells, including macrophages, which produces characteristic clinical and histological features.	Rare. Affected homozygotes usually have pathognomonic enlarged orange tonsils. HDL levels very low, with increased CAD risk. Other features include: hepatosplenomegaly, lymphadenopathy, thrombocytopenia, anaemia,

				GI disturbance, neuropathy, corneal opacities. Heterozygotes have low HDL and increased CAD risk without the other clinical features.[66]
Homocystinuria (OMIM# 236200)	Autosomal recessive	Cystathionine β-synthase	Cystathionine β-synthase is an enzyme in the methionine metabolism pathway. It catalyses conversion of homocysteine and serine to cystathionine. Multiple loss of function mutations have been reported, producing high circulating homocysteine levels. The mechanisms leading to CAD are incompletely defined.	Worldwide 1/300000, in Ireland 1/65000.[67] Affected homozygotes have 'Marfanoid' skeletal abnormalities, developmental delay/mental retardation, ectopia lentis, osteoporosis, and thromboembolism. Untreated, 50% risk of thromboembolic vascular events (including CAD) by age 30 years.[68]
Autosomal dominant coronary artery disease 2 (OMIM# 610947)	Autosomal dominant	*LRP6*	*LRP6* encodes LDL receptor related protein 6, which is related to the LDLR gene and acts as a co-receptor in the Wnt signalling pathway. Missense mutation associated with reduced Wnt signalling, raised LDL, and premature CAD.	Single Iranian pedigree. Elevated LDL, premature CAD and metabolic syndrome described in homozygotes and heterozygotes.[69]
Autosomal dominant coronary artery disease 1 (OMIM# 608320)	Autosomal dominant	Uncertain ?*MEF2A*	Genome-wide linkage analysis in a family with an autosomal dominant pattern of premature CAD identified chromosome 15q26 as a possible susceptibility locus.[70] Sequencing revealed a 21bp deletion in the *MEF2A* gene at this locus in affected individuals in that family, but analysis in other populations demonstrated that the 21bp *MEF2A* deletion did not co-segregate with CAD, suggesting it is not causative.[71] Other *MEF2A* mutations do not appear to be a common cause of CAD in Caucasians. The gene and mechanism remain to be elucidated.	Reported in isolated pedigrees. Premature CAD and MI.
Sitosterolemia (OMIM# 210250)	Autosomal recessive	*ABCG5* *ABCG8*	*ABCG5/8* encode ATP-binding cassette G5/G8. These limit intestinal absorption, and promote biliary excretion, of non-cholesterol sterols. Inactivating mutations lead to high levels of plasma sterols.	Rare. Affected homozygotes have high plasma sterol levels, xanthomata, and premature CAD.

OMIM# = NCBI Online Mendelian Inheritance in Man identifier. * genes for which variants have been shown to be contribute to lipid/cardiovascular phenotypes in the general population. FH = familial hypercholesterolemia.

The term familial hypercholesterolemia is usually used specifically for the syndrome associated with mutations in the *LDLR* gene, but similar phenotypes are associated with abnormalities of the *APOB* and *PCSK9* genes. Detailed investigation of patients with 'FH' phenotypes demonstrated that 79.1% were due to *LDLR* mutations, with 5.5% due to *APOB* mutations, and 1.5% due to mutations in *PCSK9*[72]. Involvement of other genes may account for some of the remaining proportion. The genes identified through studies of FH and other rarer Mendelian disorders may also have common variants that are associated with disease in the general population, and thus these rare conditions can identify candidate genes for further studies. For example, commoner variants in both *LDLR* and *PCSK9* have been shown in recent studies to contribute to population lipid levels and CAD susceptibility. Furthermore, insights from investigations of FH were important in the development of statins, now used widely to reduce CAD risk at the population level.

Apo A-I is the major protein constituent of HDL. Mutations in the *APOA1* gene cause Apo A-I deficiency, which is an autosomal dominant disorder. Numerous mutations in the *APOA1* gene have been reported, some of which lead to very low circulating HDL levels and increased risk of atherosclerosis[63, 64, 73]. However, a variant of Apo A-I (named Apo A-I Milano) identified in an Italian pedigree was associated with reduced CAD risk in heterozygous carriers, despite reduced HDL levels and raised triglycerides[65, 74, 75]. The Apo A-I Milano protein differs from native Apo A-I by the substitution of cysteine for arginine at position 173, which changes its properties and allows the formation of disulphide-linked homodimers and heterodimers with apoA-II[76]. Recombinant Apo A-I Milano complexed with a naturally occurring phospholipid has been manufactured to mimic the properties of nascent HDL, and this has been shown to reduce the lipid and macrophage content of atherosclerotic plaques in animal models of atherosclerosis[77-80]. Furthermore, a randomised double-blinded controlled trial in human subjects with acute coronary syndromes showed a significant reduction in coronary atheroma volume measured using intravascular ultrasound[81]. These investigations suggest that recombinant Apo A-I Milano, or drugs mimicking the effect of the mutation on lipoprotein trafficking, could be useful for the stabilization, and possibly even regression, of atheromatous plaque in the wider population with CAD.

1.1.4.2 Candidate gene association studies

CAD has been the subject of large numbers of 'candidate gene' association studies – there are several hundred such published studies currently in the Genetic Association Database, an archive of association studies in human complex diseases[14]. These studies investigate variation in genes that are already thought to be involved in the pathogenesis of the disease (based on experimental or theoretical grounds) by comparing allele or haplotype frequencies between cases and control groups. A major limitation of this approach has been that relatively few gene-polymorphism associations have been consistently replicated[82]. Many potential explanations for the inconsistent results between different studies have been proposed, which highlight the problems inherent in studying the genetics of a complex disease in which multiple genetic and non-genetic factors operate. These include differences in genetic and environmental factors between populations studied, including genetic heterogeneity (whereby causative variants differ between the populations), differing patterns of linkage disequilibrium (whereby the causal variant is not captured by the same marker in a different population), or phenotypic heterogeneity (whereby the complex 'phenotype' defined as CAD may differ between populations and be attributable to different risk factors)[82]. Methodological factors pertaining to study design have also played a significant role, as many studies lack adequate statistical power to detect associations (false negative results); whereas in others, the presence of confounding factors between cases and control groups, or excessive subgroup analyses, increases the likelihood of false positive associations[10]. However, perhaps the most important factors accounting for the apparently contradictory results of candidate gene studies are the criteria widely used to define and interpret 'statistical significance' and the presence of 'publication bias' favouring the reporting of positive associations[83]. Using a 5% threshold to define 'significance' a positive association would be expected for 1 in 20 tests due to chance alone. Therefore in view of the multiple tests performed for different candidate genes in hundreds of different studies, a great number of false positive associations are expected. However, the *a priori* chance of any particular gene of the 30,000 in the human genome being associated with a particular disease is likely to be small, even allowing for the fact that plausible candidates are selected for testing. This needs to be considered in interpreting the results of such studies, and reinforces the importance of replicating results in

adequately powered studies. The replication problems in the field were highlighted in a 2007 paper by Morgan *et al* which typed 85 variants in 70 genes previously claimed to be associated with the risk of acute coronary syndrome, finding just one borderline significant (P=0.03) association that was most likely due to chance[84].

Detailed discussion of the hundreds of association studies and candidate genes that have been investigated is not possible here. Generally, however, the relative risks attributable to variants reported in association studies have been modest, in keeping with the CDCV hypothesis. As individual studies were often too small to reliably confirm or refute associations, meta-analyses have been used to provide an assessment of composite effect, although the heterogeneity between studies must be considered when interpreting such data. The results of several large-scale meta-analyses are summarised in Table 1.2. Considered as a whole, candidate gene association studies have identified multiple genes which may be connected with CAD, but replication and elucidation of causative mechanisms remain to be established in most cases.

Increasing availability of high-throughput genotyping has allowed larger populations and larger numbers of polymorphisms to be screened, both within the same gene and in different genes. This offers the potential for such studies to overcome many of the shortcomings that have limited them in the past, particularly with respect to inadequate sample size and statistical power. However, an inherent limitation of the candidate gene approach is that it is restricted to the investigation of pathways within the sphere of prior, and incomplete, knowledge of the pathophysiology of CAD. Alternative approaches with the potential to identify novel associations and mechanisms are therefore extremely attractive, and one such approach that has been employed for CAD is the genome-wide linkage study.

1.1.4.3 Family based linkage studies

Linkage analysis involves the use of genetic and phenotypic data collected from families. Polymorphic markers (commonly microsatellites or SNPs) at known locations throughout the genome are genotyped and statistical evidence of co-segregation of these markers with the CAD phenotype is examined in family

Table 1.2. Candidate gene polymorphisms associated with CAD in large-scale meta-analyses (>5000 CAD cases).

Gene	Variant (rs number) and risk allele	Number of studies included	Number of CAD cases / controls	Relative risk (95% CI)	References
Cholesteryl ester transfer protein (*CETP*)	TaqIB (rs708272) –A allele	38	19035 / 32368	0.95 (0.92-0.99) per allele	[101]
	I405V (rs5882) –G allele	18	10313 / 32244	0.94 (0.89-1.00) per allele	[101]
	-629C>A (rs1800775) –A allele	17	11599 / 23185	0.95 (0.91-1.00) per allele	[101]
Apolipoprotein E (*APOE*)	ε2 isoform carrier	17	21331 / 47467	0.80 (0.70-0.90)	[102]
	ε4 isoform carrier	17	21331 / 47467	1.06 (0.99-1.13)	[102]
	ε4 / ε4 homozygote	17	21331 / 47467	1.22 (1.08-1.38)	[102]
Angiotensin type 1 receptor (*AGTR1*)	+1166A/C (rs5186) –C allele	27	10180 / 17129	1.13 (1.04-1.23) per allele	[103]
Factor V	G1691A (rs6025) –A [Factor V Leiden]	60	15704 / 26686	1.17 (1.08-1.28) per allele	[104]
Prothrombin (factor II)	G20210A - A allele	40	11625 / 14462	1.31 (1.12-1.52) per allele	[104]
Plasminogen activator inhibitor (*PAI-1*)	[-675]4G/5G – 4G allele	37	11763 / 13905	1.06 (1.02-1.10) per allele	[104]
Paraoxonase (*PON1*)	Q192R (rs662) -R allele	43	10106 / 11786	1.12 (1.07-1.16) per allele	[105]
Endothelial nitric oxide synthase (*eNOS*)	Glu298Asp –Asp/Asp homozygote	14	6036 / 6106	1.31 (1.13-1.51)	[106]
	Intron-4a –a/a homozygote	16	6212 / 6737	1.34 (1.03-1.75)	[106]
Apolipoprotein B (*APOB*)	SpIns/Del –DD homozygote	22	6007 / 5609	1.19 (1.05-1.35)	[107]
Methylene tetrahydrofolate reductase (*MTHFR*)	677C>T (rs1801133) –TT homozygote	40	11162 / 12758	1.16 (1.05-1.28)	[108]

pedigrees using bio-statistical algorithms[82]. Evidence of cosegregation (suggested by a logarithm of the odds, or LOD score, greater than 3)[40] suggests that the marker is near to a disease susceptibility locus. Fine mapping of the region can then be performed by genotyping more markers in the region of interest and repeating the analysis, or gene maps can be consulted to suggest nearby candidates for association testing.

Genome-wide linkage studies have the advantage that analyses are undertaken without the need for any *a priori* assumptions, and can therefore be regarded as hypothesis generating and capable of identifying entirely novel genetic associations (in contrast to the candidate gene approach). However, the positional resolution and power to detect a given effect is lower than for association studies[40], and extended pedigrees of sufficient size can be difficult to obtain.

Several linkage studies have been performed in large collections of families with CAD[85-94]. Linkage associations have been reported for chromosomes 1, 2, 3, 13, 14, 16, 17 and X, but there was limited replication of any particular region and very few novel susceptibility genes have actually been identified using this method. In the most well known study a genome-wide linkage scan of 713 individuals from 296 families identified a susceptibility locus for MI at chromosome 13q12-13, where the *ALOX5AP* gene encoding 5-lipoxygenase activating protein (FLAP) was identified[91]. Association between SNPs in this gene and MI were demonstrated in a case-control study, and involvement of the locus was then replicated in an independent cohort. The genetic findings have not, however, been consistently replicated in other studies, some of which are substantially larger than the hypothesis-generating study[84, 95-99]. Summary of the evidence to date suggests that the effect of the haplotypes studied, if it exists, is of a much smaller size than originally estimated. Based on the initial genetic findings, a clinical trial was conducted to examine the effect of FLAP inhibition on levels of biomarkers associated with MI risk in 191 patients who had already suffered an MI and carried at-risk variants in FLAP[100]. The FLAP inhibitor led to suppression of plasma levels of leukotriene B4, myeloperoxidase, and C-reactive protein without any adverse events, suggesting that larger-scale trials powered to detect differences in CAD endpoints would be of interest.

1.1.4.4 Genome-wide association studies

Large scale GWA studies have some advantages over linkage studies and candidate gene studies for the investigation of common variants conferring susceptibility to complex diseases such as CAD. Like genome-wide linkage studies they investigate association in a hypothesis-generating manner, without the need for *a priori* assumptions, and can therefore identify novel associations. However, they have greater statistical power to detect common variants than linkage studies[40] and do not require extended pedigree collections, which can be difficult to ascertain. GWA studies still require careful attention to study design to ensure that past failings are not repeated – rigorous phenotyping, appropriate matching of cases and controls, accurate genotyping, adequate sample size, appropriate analysis techniques and significance thresholds, and robust replication of findings are all of paramount importance[109].

The large datasets that are generated from GWA studies have also proved challenging with respect to bioinformatic and statistical analysis. The large number of association tests performed provides the potential for false positive results, and very stringent thresholds have therefore been adopted to define genome-wide significance (typically $5x10^{-8}$)[40, 109]. Analysis also requires strategies to confirm data quality, deal with missing and imputed data, and consider multi-marker effects[109]. Another limitation with sample sizes reported in the first generation of such studies is that the power to detect associations is low for alleles with lower minor allele frequency (MAF) or weaker effects. Once an association between common variants and phenotype has been found, establishing the mechanism of the association may not be straightforward. A recent theoretical paper suggested that some association signals from GWA studies may be attributable to effects of multiple uncommon causal variants distributed over a relatively wide region, in which case the causal variants and pathways responsible for the association may be difficult to identify[110].

The first novel association with CAD using a 'genome-wide type approach' was published in 2002 by a Japanese group[111], although the fact that only gene-based SNPs were typed and the relatively small number of included SNPs mean that this study is generally not considered to offer truly 'genome-wide' coverage. Almost 93,000 gene-based SNPs were typed in 1,133 cases with MI and 1,006 controls,

which identified a susceptibility locus on chromosome 6p21. LD mapping and analysis of haplotype structure showed significant association with a five-SNP haplotype in the lymphotoxin-α (*LTA*) gene, which encodes a member of the TNF ligand family. This finding was subsequently replicated in two additional cohorts[112, 113]. The risk genotype was associated with increased expression of cell-adhesion molecules in coronary artery smooth muscle cells, which may facilitate neutrophil recruitment in a pro-inflammatory mechanism leading to plaque rupture[3]. *LTA* knockout mouse studies supported a role of the gene in atherogenesis[114]. However, the promising early results have not been successfully replicated in most other large studies. Clarke *et al* tested the SNPs defining the risk haplotype at *LTA* in 6,928 cases of MI and 2712 controls from the ISIS genetic study[115]. In this large study and a meta-analysis of other published data, they found no evidence of association between any *LTA* SNP and MI risk. Those results ruled out the effect size obtained in the original study with a high degree of confidence and suggested at most a marginal association with disease risk. *LTA* was not identified with genome-wide significance in the association studies described below, which were sufficiently large to detect effects of the magnitude suggested by the initial *LTA* studies.

It was not until 2007, with the widespread availability of microarray based genotyping, that the first wave of a flood of GWA studies looking at CAD and a range of other complex diseases were published. The main findings of those studies of relevance to phenotypes related to CAD are outlined below.

1.1.5 GWA studies of CAD-related phenotypes

1.1.5.1 CAD

Four GWA studies of CAD reported in 2007 demonstrated conclusive evidence for association between common SNPs in the same ~100kb region on chromosome 9p21 and CAD risk[18-21]. The results of these studies are summarised in Table 1.3.

Table 1.3. CAD susceptibility loci replicated in multiple GWA studies.

Study	Phenotype	Total patients / controls	SNPs on array	Locus / SNP / risk allele	Allele Frequency		Relative risk (95% CI) heterozygote / homozygote	P value
					Controls	Cases		
WTCCC[18] and Samani[21]	CAD / MI	2,863 / 4,648	500,000	9p21 / rs1333049 / C	0.47	0.55	1.36 (1.27-1.46) per copy of risk allele	$2.9x10^{-19}$
				6q25.1 / rs6922269 / A	0.25	0.29	1.23 (1.15-1.33) per copy of risk allele	$2.9x10^{-8}$
				2q36.1 / rs2943634 / C	0.34	0.30	1.21 (1.13-1.30) per copy of risk allele	$1.6x10^{-7}$
Helgadottir[19]	MI / CAD	4,587 / 12,767	305,000	9p21 / rs2383207 / G	0.492	0.548	1.25 (1.18-1.31) per copy of risk allele	$2.0x10^{-16}$
				9p21 / rs10757278 / G	0.453	0.517	1.28 (1.22-1.35) per copy of risk allele	$1.2x10^{-20}$
McPherson[20]	CAD	4,306 / 20,119	100,000	9p21 / rs10757274 / G	0.487*	0.525*	1.18 (1.02-1.37) / 1.29 (1.09-1.52)*	$4x10^{-3}$*
				9p21 / rs2383206 / G	0.505*	0.541*	1.26 (1.09-1.46) / 1.26 (1.07-1.48)*	$7x10^{-4}$*

*Data shown from ARIC cohort only.

The Wellcome Trust Case-Control Consortium (WTCCC) study examined around 400,000 SNPs in 1,988 British Caucasians with a validated history of MI or coronary revascularisation before the age of 66 years, and 3,004 controls[18]. Associations were seen for SNPs across >100kb in the 9p21 region, with the strongest association demonstrated for rs1333049 ($P=1.8x10^{-14}$). For this SNP the heterozygote odds ratio was 1.47 (95% CI 1.27-1.70) and the homozygote odds ratio was 1.90 (95% CI 1.61-2.24). 'Moderate' associations, defined as SNPs with a P-value greater than $5x10^{-7}$ but less than $1x10^{-5}$, were reported for six other loci. Replication of the chromosome 9 locus was achieved in a subsequent paper by Samani *et al* which added data from a cohort comprising 875 German Caucasians with MI before the age of 60 years and at least one family member with premature CAD, and 1,644 controls, which had been genotyped using the same chip[21]. As shown in Table 1.3, two loci on chromosome 6 and chromosome 2 were also replicated in the German cohort. The combined analysis identified four additional loci potentially associated with CAD, one of which (chromosome 1p13.3) has subsequently been shown to be strongly associated with plasma LDL-cholesterol[116]. An analysis of 55 candidate genes previously reported to show association with CAD only confirmed association for two SNPs tagging a variant in the lipoprotein lipase gene in these cohorts.

Helgadottir *et al* performed a similar study in 1,607 Icelandic patients with MI (before age 70 years in males and 75 years in females) and 6,728 controls, typed for around 300,000 genomewide SNPs using the Illumina Hap300 chip[19]. The strongest association was found for three correlated SNPs in the same chromosome 9p21 region that was identified in the WTCCC study, each with P values of approximately $1x10^{-6}$. The association was replicated in four additional case-control cohorts. Combining data from all groups, allele G of the SNP rs10757278 showed the strongest association with MI, with an odds ratio of 1.28 (95% CI 1.22-1.35, $P=1.2x10^{-20}$) per allele.

McPherson *et al* typed 100,000 SNPs in 322 cases of premature CAD and 312 controls, and replicated positive associations in a further 1,658 cases and 9,380 controls[20]. Two SNPs located within 20kb of each other in the same region of chromosome 9p21 were significantly associated with CAD. These two SNPs were

validated in three additional independent cohorts (with varying inclusion parameters for defining CAD).

The 9p21 region had not been previously implicated in CAD susceptibility studies and its identification by genome-wide association highlights the ability of such studies to discover novel risk loci. Furthermore, these studies showed a remarkably consistent association with CAD for the chromosome 9p21 risk region in multiple independent populations using varying inclusion criteria, in marked contrast to the replication problems that had been observed in previous studies of CAD genetics. Further studies investigating this region are discussed in detail below (section 1.2).

The only other loci replicated in separate studies were the chromosome 6q25.1 and chromosome and 2q36.3 loci reported by Samani *et al*. Associated SNPs on chromosome 6q25.1 are located in introns within the gene for methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1-like protein (MTHFD1L). This encodes the mitochondrial isozyme of C1-tetrahydrofolate synthase which is involved in the synthesis of purines and regeneration of methionine from homocysteine. It has therefore been suggested that MTHFD1L activity may influence plasma homocysteine levels, which are a risk factor for CAD, although preliminary analysis of the lead SNP at this locus and homocysteine levels in 1,070 individuals did not show such an association[21]. The chromosome 2q36.3 locus contains one pseudogene and the mechanism of any association at this locus remains unclear.

Subsequent GWA studies have identified multiple additional loci for CAD susceptibility including the *MRAS* gene on chromosome 3q22.3[22], the *SLC22A3-LPAL2-LPA* gene cluster on chromosome 6q26[23], and the *SH2B3* gene on chromosome 12q24[24]. However, individual GWA studies have been adequately powered to detect only high frequency alleles or large effects and it is likely that further loci will be identified through larger studies or combined meta-analyses[109, 117].

1.1.5.2 Plasma lipids

Recent GWA studies investigating plasma lipoprotein concentrations as quantitative traits have produced convincing associations at a number of loci, many of which have been replicated in separate analyses[118-121]. These have confirmed multiple loci previously implicated in lipid metabolism and identified a number of novel genetic associations with HDL cholesterol, LDL cholesterol and triglyceride levels.

Newly identified associations with LDL cholesterol include variants close to the *CELSR2-PSRC1-SORT1* locus, which was identified in three independent studies[118, 120, 121]. No mechanistic connection is obvious for the two genes closest to the association signal (*CELSR2* and *PSRC1*), but variants may influence the nearby *SORT1* gene which is involved in lipoprotein lipase metabolism. An allele on chromosome 19p13.11 showed association with both increased LDL and triglycerides in multiple studies[118, 121]. This variant is close to *CILP2* and *PBX4* and in strong LD with a neuronal proteoglycan *NCAN*, but with no obvious mechanism for its association with plasma lipids.

Novel associations with HDL cholesterol include a locus on chromosome 12q24.11 near to the neighbouring *MVK* and *MMAB* genes, which encode enzymes involved in cholesterol biosynthesis and degradation pathways respectively[121]. A novel association for both HDL and triglycerides was found for a locus on chromosome 1q42.13 within the first intron of *GALNT2*, a widely expressed glycosylotransferase with no known role in lipid metabolism[118, 121].

Replicated novel associations with triglycerides were found for variants near to *MLX1PL* and *ANGPTL3*, both of which are known to be regulators of lipid metabolism. The probable mechanism of a further association with variants near to the *TRIB1* gene on chromosome 8q24.13 remains obscure[118, 119, 121].

The first round of GWA studies have also provided confirmation of variants at loci previously implicated in lipid metabolism, including apolipoproteins (*APOE*, *APOB* and *APOA5*), cholesterol and cholesterol ester transporters (*ABCA1* and *CETP*), lipoprotein receptors (*LDLR*), and lipases (*LPL*, *LIPC* and *LIPG*)[121].

Willer *et al* examined the association of alleles linked to lipid metabolism with CAD in a combined cohort of the WTCCC sample and an expanded panel of British individuals[121]. All of the alleles associated with LDL metabolism were more common in CAD cases than controls, with modest odds ratios of 1.04 to 1.29 per allele; the association was significant for eight of the 11 SNPs. The CAD risk variant rs1333049 was not associated with lipid levels in the same cohort however. These findings suggest that the newly identified loci are potential targets for novel therapies affecting lipoprotein levels and CAD risk.

1.1.5.3 Type II diabetes mellitus

The first round of GWA studies provided similarly exciting insights into the genetic associations of type II diabetes, which is a major risk factor for cardiovascular disease. In the case of type II diabetes, recent studies have confirmed loci in genes for which there was robust prior evidence of involvement (such as *PPARG*, *KCNJ11* and *TCF7L2*); substantiated evidence of a link for other previously implicated genes (such as *IGF2BP2*, *HHEX-IDE* and *SLC30A8*); and identified novel loci and potential candidate genes previously unsuspected of involvement[18, 122-127]. The latter group includes a locus on chromosome 6p22 in intron 5 of the *CDKAL1* gene, the association with which has been robustly replicated in multiple studies (odds ratio for heterozygotes 1.12 to 1.20)[18, 123-126]. The function of this gene is unknown, but it shares protein domain homology with kinases involved in beta-cell metabolism. Variants in the *FTO* gene which is associated with adiposity were also associated with type II diabetes[18, 124].

Interestingly, the recent GWA studies have also identified SNPs on chromosome 9p21 associated with an odds ratio of 1.20 for diabetes risk in some populations[124-126], although this was not reported in other groups[127, 128]. Diabetes is a major risk factor for CAD, associated with features such as vascular endothelial dysfunction and hyperlipidemia that are also important in atherogenesis, raising the prospect of a shared pathophysiological mechanism for the two diseases. However, the most strongly associated SNP for type II diabetes in the region (rs10811661) lies beyond the LD block that is strongly associated with CAD risk. A large study simultaneously comparing the diabetes variant rs10811661 and CAD SNPs with CAD and diabetes

phenotypes showed that they were independent of each other, suggesting that the underlying pathophysiological mechanism may not be the same for the two diseases[129]. This was confirmed by Helgadottir *et al* who in multiple large cohorts examined the association between the CAD risk variant rs10757278 and diabetes risk variant rs10811661 and a number of vascular phenotypes including abdominal aortic aneurysm, intracranial aneurysm, peripheral arterial disease, large artery atherosclerotic or cardiogenic stroke, CAD, and type II diabetes[130]. The diabetes variant did not show association with any of the arterial disease phenotypes.

A second SNP, rs564398 which is not in LD with rs10811661, was associated with diabetes in a meta-analysis of 14,586 cases. This lies within the CAD associated region described in the WTCCC study, and shows moderate correlation ($0.28<r^2<0.42$) with SNPs in the core CAD haplotype[129]. rs564398 showed association with CAD in the study by Broadbent *et al.* ($P=4x10^{-8}$), but once a marker of the CAD risk haplotype was included in the statistical model, this became non-significant ($P=0.08$)[129]. There is therefore no evidence for an effect of rs564398 on CAD risk independent of its association with the risk haplotype at the present time. However, insufficient numbers of people with and without CAD that are concordant and discordant for diabetes have been studied so far to enable a possible effect of rs564398 on both diabetes and CAD to be entirely ruled out.

1.1.5.4 Hypertension

In contrast to the phenotypes described above, the first wave of GWA studies investigating BP phenotypes failed to identify associations at the level of genome-wide significance, despite documented heritability of this trait and similar sample sizes to studies which have detected loci for other diseases[18, 126, 131-134]. Potential explanations cited for the lack of association were that hypertension may have fewer common risk alleles of sufficient effect size, variants may be poorly tagged by the typed SNPs, or the potential for misclassification of hypertensive individuals within the control group[18].

A GWA study published in January 2009 by Wang *et al* analysed 79,447 SNPs in the American Old Order Amish, a closed population descended from a small number of

common founders who emigrated from Switzerland in the early 1700s who have a relatively homogeneous lifestyle[134]. Their initial genome-wide screen of 542 subjects from the Amish Family Diabetes Study[135], in which families were ascertained through a proband with type 2 diabetes, identified a cluster of SNPs in the *STK39* (serine threonine kinase 39) locus on chromosome 2q24.3 that were associated with BP ($P=8.9x10^{-6}$ to $9.1x10^{-5}$) [134]. Although the P-value did not reach the conventionally accepted threshold for genome-wide significance[136], the inclusion of additional data from an Amish and four non-Amish Caucasian cohorts led to a significant result ($P=1.6x10^{-7}$) in the combined dataset of 7,125 individuals. The replication cohorts included a further population of 2,842 individuals from a case-control study of diabetes and three smaller groups not selected for diabetes.

The *STK39* gene encodes the SPAK (Ste20-related proline-alanine-rich kinase) protein which interacts with ion cotransporters involved in salt transport and osmotic cell volume regulation, including the thiazide-sensitive and loop diuretic-sensitive cotransporters involved in renal salt excretion[137, 138]. The *STK39* gene contains 18 exons spanning approximately 300kb on chromosome 2q24.3; SNPs associated with BP by GWA studies were located within introns 1-8, with no coding or splice variants identified by sequencing[134]. Using transfection experiments with luciferase reporter constructs Wang *et al* demonstrated that alleles of one SNP (rs35929607), located in a conserved region of intron 2 altered expression of the reporter constructs *in vitro*. This SNP was in LD with the SNPs identified by GWA studies, suggesting a possible functional mechanism for the observed association.

Despite these results, other large GWA studies and meta-analyses involving up to 71,225 Caucasian individuals have failed to identify association between *STK39* SNPs and BP that satisfies genome-wide significance thresholds[18, 139, 140]. However, such data do not exclude an effect of this locus on BP, since the statistical threshold for replicating an association differs from the threshold for genome-wide 'discovery'. Differences in the actual SNPs genotyped, patterns of LD, and phenotypic variation between populations might contribute to the lack of genome-wide significance in these studies, compared to the GWAS by Wang *et al*. One recent GWA study looked specifically for evidence of correlation between *STK39* SNPs and BP in 1,017 African American subjects and provided some supportive evidence for an association[141].

SNPs in the *STK39* region did not reach genome-wide significance, but the number of significantly associated SNPs at a nominal P threshold of P<0.05 was somewhat higher than expected by chance alone (9/136 for systolic BP and 33/136 for diastolic BP, compared to 7/136 expected by chance). The association of *STK39* SNPs with BP remains to be confirmed at present and further studies are required to investigate this.

Despite the disappointing results of the early hypertension GWAs, subsequent GWA studies and very large meta-analyses (the Global BPgen and CHARGE consortia) have subsequently identify multiple loci convincingly associated with BP phenotypes[139-142]. In common with the findings for other phenotypes discussed above, these studies confirmed previously implicated genes such as *CYP17A1*, a gene associated with a rare Mendelian form of hypertension, as well as identifying novel loci, such as *ULK3* and *ULK4* for which the mechanisms of the association with BP are completely unknown. The small effect on BP of the individual common variants identified, which were in the order of 1mmHg systolic and 0.5mmHg diastolic per copy of the risk allele, underscores the need for large collaborative studies and may account for the lack of association seen in the initial smaller GWA studies.

1.2 The chromosome 9p21 susceptibility locus

1.2.1 Replication and further characterisation

The association of the chromosome 9p21 locus with CAD has been replicated in multiple additional Caucasian case-control series and meta-analyses[129, 143-146]. Association has also been reported for populations of other ethnicities including Japanese[147, 148], Korean[148], US Hispanic[149], Asian Indians[150], and Chinese[151-153] populations. Fine mapping and detailed characterisation of the locus was performed in a large case-control study of 4,251 CAD cases and 4,443 controls from four European populations by Broadbent *et al*[129]. From an analysis of 62 SNPs in the region they identified a region of strong LD containing 14 SNPs that form four common haplotypes (cumulative frequency 89%), with the association of the linked SNPs at this locus a consequence of a perfect 'yin-yang' haplotype spanning 53kb.

The CAD association has shown little evidence of gene-environment interactions, with consistent findings in subgroups analysed by age, gender, smoking status, hypertension, or diabetes[129]. Similarly, there has been little difference in the strength of association between populations ascertained for MI versus those with CAD but no previous MI, although the trend towards a stronger association with the non-MI phenotype suggests that the primary effect of the 9p21 locus is not on plaque rupture[19, 21, 129]. This is in keeping with the findings from a number of studies suggesting that the risk allele is associated with the extent or progression of atherosclerosis[154, 155].

The mechanisms through which the 9p21 locus influences CAD susceptibility are not known. Approaches to identifying pathways involved in causation include first, consideration of the relationship with other cardiovascular intermediate phenotypes which may provide mechanistic clues; and second, consideration of effects on expression of nearby candidate genes.

1.2.2 Relationship to other phenotypes

1.2.2.1 Intermediate phenotypes for CAD

McPherson *et al* examined the relationship between risk alleles and traditional atherosclerosis risk factors in 2,872 cases with CAD and 18,107 controls. They found no association of the risk alleles with age, gender, body mass index, BP, smoking, physical activity, fatty diet, or plasma levels of glucose, triglycerides, LDL cholesterol, HDL cholesterol, CRP, intercellular adhesion molecule-1, and vascular cell adhesion molecule-1[20]. This finding was subsequently extended by further studies that, in addition to replication of the negative association with the previously listed parameters, also found no association between the chromosome 9p21 risk locus and cardiovascular risk phenotypes including plasma levels of lipoprotein(a), albumin, fibrinogen, uric acid, and homocysteine[129, 156]. The lack of association with traditional risk factors and intermediate phenotypes for atherosclerosis suggests that this locus may act via a novel mechanism. This is discussed in greater detail in Chapter 6.

1.2.2.2 Other vascular phenotypes

The relationship between risk SNPs for CAD and diabetes has already been discussed in detail. Helgadottir *et al* also examined the association between the CAD risk variant rs10757278 and a number of vascular phenotypes including abdominal aortic aneurysm, intracranial aneurysm, peripheral arterial disease, and large artery atherosclerotic or cardiogenic stroke[130]. The CAD risk variant was significantly associated with abdominal aortic aneurysm (OR 1.31, $P=1.2x10^{-12}$) and intracranial berry aneurysm (OR 1.29, $P=2.5x10^{-6}$), but after excluding potentially confounding cases of known CAD, it was not significantly associated with peripheral arterial disease, ischaemic stroke, or diabetes. Other early studies also reported no association of 9p21 CAD variants with stroke risk[157, 158], but subsequent studies have now shown convincing association of these SNPs with ischaemic stroke[116, 159-162], and other markers of systemic atherosclerosis[163].

1.2.2.3 Non-vascular phenotypes

Interestingly, SNPs close to the CAD risk region on chromosome 9p21 have also been identified in recent GWA studies as being associated with glioma[164, 165], naevi/melanoma[166, 167], and basal cell carcinoma[168]. As discussed below, the region contains the recognised tumour suppressor genes *CDKN2A* and *CDKN2B*, providing a plausible mechanistic association for the involvement of these variants and genes in cancer aetiology. In view of the known involvement of *CDKN2A* and *CDKN2B* in rare familial cancers, common variants in these genes had been previously investigated in 'candidate genes' studies and were reported to be associated with a range of different cancer phenotypes. These studies, which are summarised in Table 1.4, shared many of the limitations common to other candidate gene studies as previously discussed. The sample sizes were often small, most of the positive findings were unreplicated, and significance thresholds were often unadjusted for multiple testing with over-interpretation of borderline results. The issue of reporting bias favouring positive associations also needs to be considered in interpreting the results of these studies. The location of SNPs that have been associated with disease phenotypes in the 9p21 region are illustrated in Figure 1.4.

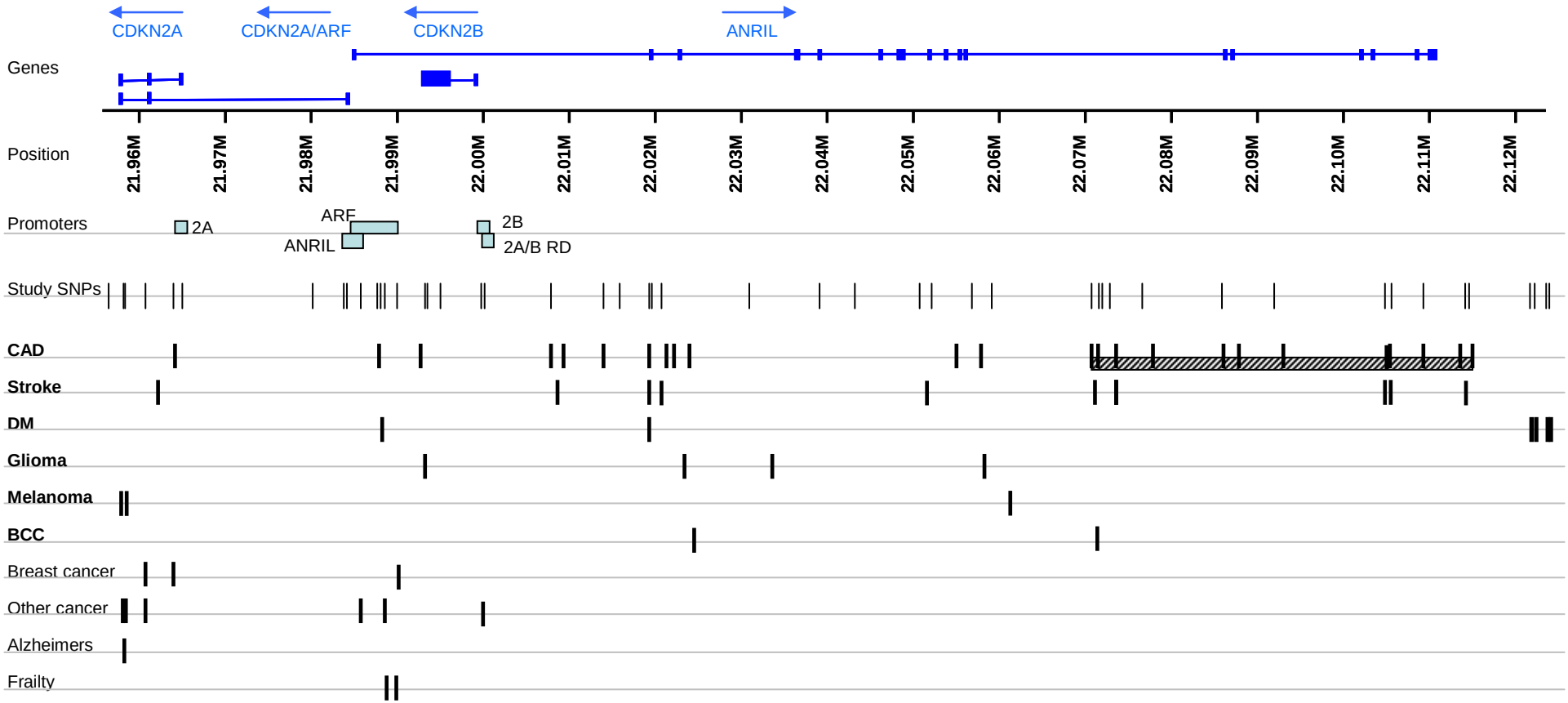
Table 1.4. Candidate gene polymorphisms previously associated with disease in the *CDKN2A/ARF/CDKN2B* region.

Study	Disease	Polymorphism	Population	Evidence of association
Kumar *et al* 2001[169]	Malignant melanoma	rs3088440	2 SNPs in the 5'UTR of CDKN2A typed in 229 patients with melanoma and 235 controls.	T allele of rs3088440 associated with melanoma (OR=1.7, 95%CI 1.1-2.7, P=0.01).
Straume *et al* 2002[170]	Survival in malignant melanoma	rs3088440	2 SNPs in the 5'UTR of CDKN2A typed in 185 malignant melanoma cases.	No association with methylation status, p16 expression, or other clinicopathological variables. rs11515-C associated with improved survival (P=0.03). Not adjusted for multiple testing.
Sakano *et al* 2002[171]	Clinical course of bladder cancer	rs11515 rs3088440	2 SNPs in the 5'UTR of CDKN2A typed in 309 patients with bladder cancer and 235 controls.	No association with bladder cancer. rs11515 associated with tumour progression in 219 patients with superficial bladder cancer (P=0.04) and tumour-specific survival (P=0.02). Not adjusted for multiple testing.
Debniak *et al* 2005[172]	Malignant melanoma	rs3731249	3 CDKN2A SNPs typed in 471 cases and 1210 controls.	Risk (G) allele of rs3731249 associated with disease OR 2.5 (P=0.0003).
Debniak *et al* 2005[173] and 2007[174]	Breast cancer	rs3731249	Single SNP typed in 4,209 cases and 3,000 controls.	Risk (G) allele present in 8/66 cases under 30 years of age, OR 3.8 (P=0.0002) and 168/3318 cases under 50 years, OR 1.5 (P=0.002).
Chen *et al* 2007[175]	Age of onset of pancreatic cancer	rs11515 rs3088440	2 SNPs in CDKN2A and 1 in Aurora-A typed in 148 patients with pancreatic cancer.	Aurora-A and rs3088440-T had a synergistic effect on age-associated risk for early diagnosis, with median age at diagnosis 12.6 years earlier (OR=3.9, 95%CI 1.9-7.8, P=0.0002).
Gayther *et al* 2007[176]	Invasive epithelial ovarian cancer	rs3731257	88 SNPs in cell cycle genes typed in 1,500 cases and 2,500 controls, then 5 most significant SNPs typed in additional 2,000 cases and 3,200 controls.	Combined analysis showed rare allele for rs3731257 associated with disease, OR for homozygotes 0.79 (95%CI 0.65-0.95), P=0.008. Non-significant (P=0.37) after adjustment for multiple testing.
Healy *et al* 2007[177]	Acute lymphoblastic leukaemia	rs3731249 rs2069416	10 SNPs in 4 cell cycle genes typed in 240 patients with ALL and 277 controls.	Risk alleles at rs3731249 (P=0.008) and rs2069416 (P=0.02) associated with ALL.

Kang *et al* 2008[178]	p14ARF promoter methylation in colorectal cancer	rs3218012 rs2518723	21 SNPs close to p14ARF CpG island typed in 188 cases of colorectal cancer and 300 controls.	Risk haplotype associated with p14ARF promoter methylation (OR=8.3, 95%CI 2.4-28.4, P=0.0007). Not adjusted for multiple testing.
Melzer *et al* 2007[179]	Reduced physical functioning in elderly	rs2811712 rs3218005	25 CDKN2A/B SNPs typed in 3 elderly (aged 65-80 years) cohorts (total n=3372).	In screening cohort G alleles of rs2811712 and rs3218005 associated with better physical function. rs2811712 analysed in replication cohorts. Prevalence of severely limited physical function 15% in common homozygotes, 7% in rare homozygotes (OR=1.48, 95%CI 1.17-1.88, P=0.001)
Driver *et al* 2008[180]	Breast cancer	rs3731239 rs3218005	240 SNPs in cell cycle genes typed in 2270 cases and 2,280 controls. SNPs with P<0.1 then typed in further 2,200 cases and 2,280 controls.	rs37371239 OR (CC/TT) = 0.90 (95%CI 0.79-1.03, P=0.013). rs3218005 OR (GG/AA) = 1.55 (95%CI 1.02-2.37, P=0.013). P-values unadjusted for multiple testing.
Yan *et al* 2008[181]	Histologic subtypes of epithelial ovarian cancer	rs11515 rs3088440	2 SNPs analysed in 205 patients with ovarian cancer and 268 controls.	No association with cancer. Subgroup analysis rs3088440 associated with histologic subtype of tumour (P=0.02 unadjusted for multiple testing).
Zuchner *et al* 2008[182]	Late onset Alzheimer disease	rs11515 rs3731246	80 9p21 SNPs typed in 674 families with late onset Alzheimer disease.	rs11515 and rs3731246 significant for family-based association after adjustment for multiple testing.

Figure 1.4. SNPs associated with disease in the chromosome 9p21.3 region.

Genes are illustrated in blue at the top, with arrows representing the direction of transcription. SNPs typed in this study and SNPs associated with various diseases are represented by black bars. Diseases in bold are those with association data from genomewide association studies. The hatched box represents the core risk haplotype for CAD defined by Broadbent *et al* [129]. Promoter regions for each gene are shown as pale blue boxes. DM = diabetes mellitus type II; BCC = basal cell carcinoma.



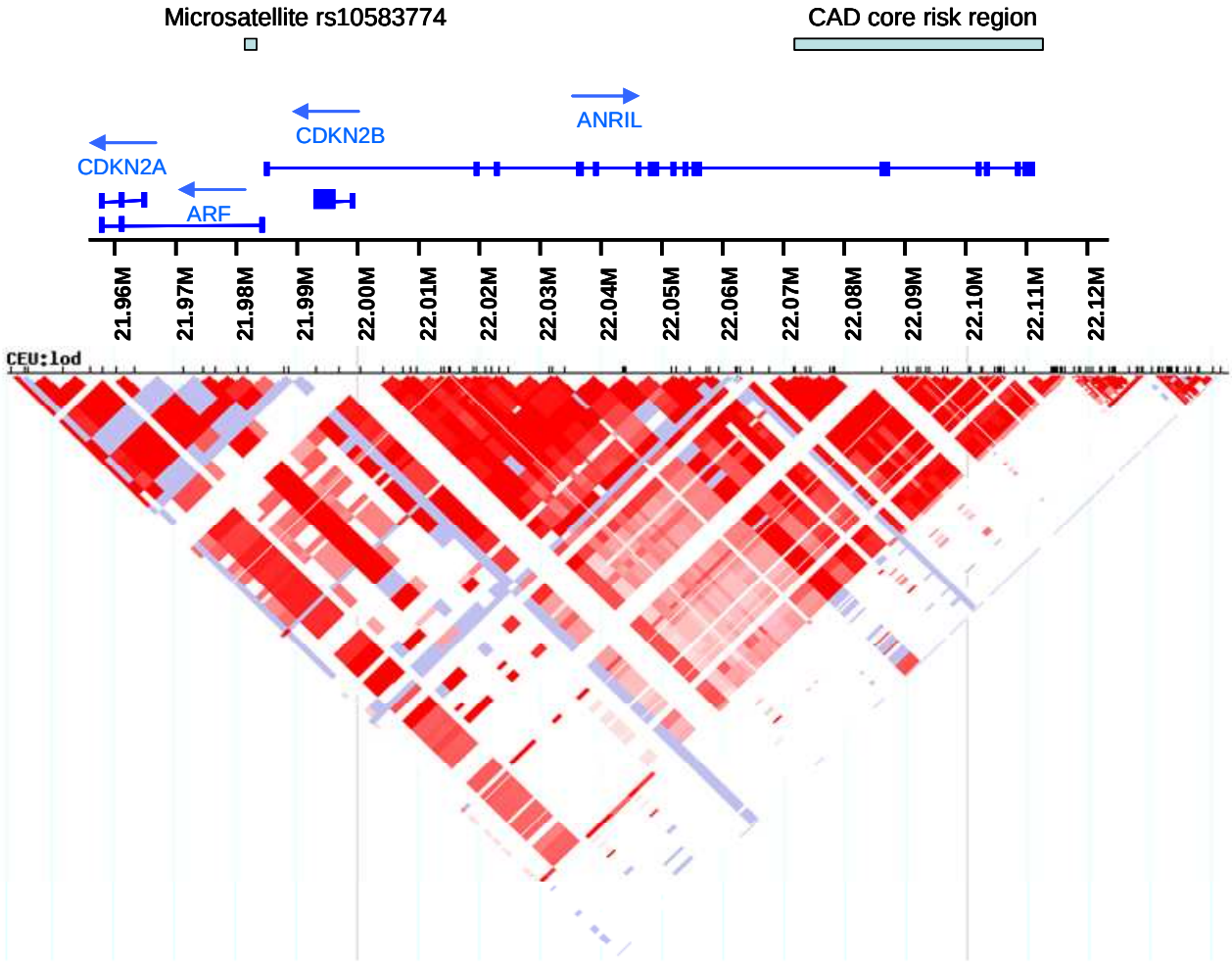
1.2.3 Candidate susceptibility genes in the 9p21 region

1.2.3.1 *CDKN2A* and *CDKN2B*

As shown in Figure 1.5, the chromosome 9p21 variants associated with CAD are located in a block of high LD that contains the cyclin-dependent kinase inhibitor genes *CDKN2A* and *CDKN2B*. These genes are well recognised as tumour-suppressor genes and are involved in regulation of cell cycle, ageing, senescence and apoptosis[183, 184].

Figure 1.5. LD in the chromosome 9p21 region between CAD risk variants, nearby genes, and microsatellite rs10583774.

The figure shows LD in the HapMap CEU population, adapted from the HapMap website[47]. Black numbers represent the chromosome 9 location (M=megabases). Shading indicates LD ranging from D'=1 in red to D'=0 in white.



The *CDKN2A* gene generates alternative transcript variants that have different first exons spliced to a common second and third exon; CDKN2A (p16, INK4a), and CDKN2A-ARF (p19, ARF)[183, 185]. Although the proteins are encoded in alternative reading frames and do not show amino acid homology, both inhibit cell cycle progression. The CDKN2A protein inactivates the cyclin-dependent kinases CDK4/6 and thereby inhibits cellular proliferation by inducing G1 cell-cycle arrest. The CDKN2A-ARF protein binds to and inactivates the MDM2 protein, resulting in p53 stabilisation and inhibition of proliferation. The CDKN2B protein is structurally similar to CDKN2A and acts through a similar mechanism. Although the three main transcripts have separate promoters and may respond independently in certain contexts[186-188], recent evidence including discovery of a common *cis*-acting regulatory domain and co-ordinated repression by Polycomb-group complexes suggests that the entire locus is co-ordinately regulated[184, 189].

Deletions involving *CDKN2A* and *CDKN2B* have been implicated in multiple cancers[185]. Experimental evidence suggests that these genes also play an important role in ageing, cellular senescence and apoptosis[185]. Increased age is the dominant risk factor for atherosclerosis, and is associated with structural and functional changes in the vascular wall, and these genes have been shown to be widely expressed, including in atherosclerotic tissue[190, 191]. Abnormal proliferation of leukocytes and vascular smooth muscle cells are a hallmark of atherosclerosis[192], but other evidence also suggests that senescence and apoptosis are important in the atherosclerotic process[193, 194]. Furthermore, *CDKN2B* expression may be induced by transforming growth factor beta, which has been implicated in the pathogenesis of atherosclerosis[190, 191]. Taken as a whole, such data provide a plausible mechanism for the involvement of *CDKN2A* and *CDKN2B* in atherosclerosis and CAD, although whether expression would be up or down-regulated is uncertain.

Helgadottir *et al* sequenced 93 early-onset MI patients across exons, exon-intron boundaries and regulatory regions of *CDKN2A* and *CDKN2B*[19], and McPherson *et al* sequenced the coding regions of these genes in 96 affected individuals[20]. These studies revealed no variants likely to account for the observed association with CAD risk.

1.2.3.2 *ANRIL*

The CAD core risk region overlaps with exons 13-19 of *ANRIL*[129], a newly annotated gene that was first reported in April 2007 by investigators examining a deletion in a large melanoma-neural system tumour syndrome family[195]. In this original report, Pasmant *et al* used the dbEST database to identify several human expressed sequence tags (ESTs) of an unidentified transcript in the region of the 403kb chromosome 9p21 deletion that they were investigating. Using real-time PCR of cDNA from normal human testis with primers placed in the identified ESTs followed by sequence analysis of the PCR products, Pasmant *et al* identified two alternatively-spliced transcripts. The first was 3,834 bp long and alignment to reference genomic DNA sequences showed that it contained 19 exons, with exon/intron junctions corresponding to the consensus sequences of donor/acceptor splice sites and a polyadenylation site in the last exon. A second shorter transcript 2,659 bp in length composed of the same first 12 exons and an alternative 3' exon 13 including an additional polyadenylation site was also characterised. These transcripts overlapped and were transcribed in the opposite direction to *CDKN2B* (which is located entirely within the first intron of *ANRIL*, as was illustrated in Figure 1.5 on page 36). These novel transcripts did not possess an open reading frame and appeared to be untranslated, and the gene was designated *ANRIL* for 'antisense noncoding RNA in the *INK4* locus'. This has recently been given the official gene name '*CDKN2B* antisense RNA' (*CDKN2BAS*)[196], but as the name *ANRIL* is used widely in the literature and in common usage, that term will be used throughout this thesis. At the time this project was undertaken, only these two transcripts had been reported. However, very recent work published in November 2009 has suggested that the situation may be more complex with the existence of multiple alternatively-spliced transcripts reported in cell lines derived from different tissues[197]. Recent studies pertaining to alternative transcripts are discussed in detail in Chapter 4 (section 4.6.8, page 182), and a summary of all transcripts reported to date is also presented in Chapter 4 (Figure 4.14 on page 159).

ANRIL produces a large non-coding RNA whose function is unknown. Other large non-coding RNAs such as *Xist* and *HOTAIR* have been reported to be involved in regulation of gene expression through transcriptional and translational control

mechanisms[198, 199]. Similar to *Xist*, *ANRIL* contains a high number of repetitive elements and encompasses binding sites for transcription factors responsible for transcriptional repression, suggesting that it may act through similar mechanisms[200]. Transcriptional repression by *HOTAIR* is mediated by recruitment of Polycomb complexes, which have also been shown to influence expression of *CDKN2A* and *CDKN2B*[201]. It is therefore possible that *ANRIL* influences disease susceptibility through similar mechanisms, and it has been shown to be expressed in tissues relevant to CAD including heart, coronary smooth muscle, vascular endothelial cells, human monocyte-derived macrophages, carotid endarterectomy specimens, and abdominal aortic aneurysm samples[129, 195].

In view of its location, the CAD risk allele may act by altering *ANRIL* expression or function. SNPs in the core risk region for CAD are not located within exons of *ANRIL*, but map to intronic and downstream sequences, which may be involved in the regulation of *ANRIL* expression. Several expressed sequence tags map within the risk locus, but do not contain open reading frames extending more than a few amino acids. McPherson *et al* sequenced the CAD risk region in two homozygotes for the risk allele and one homozygote for the reference allele, identifying 35 sequence variants that were specific to the risk allele, one of which (a CAT repeat CNV) mapped to a splice transcript of *ANRIL*[20]. *CDKN2B* antisense transcription, mapping to the first intron of *ANRIL*, has been shown to be associated with downregulation of *CDKN2B* expression in leukaemia cells and mouse embryonic stem cells, mediated in *cis* and *trans* through heterochromatin formation[202]. This suggests that *ANRIL* expression may be involved in the regulation of *CDKN2B* expression. A number of studies published after this project began that have examined the association of CAD risk SNPs with *ANRIL* expression are discussed in Chapter 4 (section 4.6.7, page 179).

1.2.3.3 *MTAP*

This gene encodes methylthioadenosine phosphorylase (MTAP), an enzyme that plays a role in polyamine metabolism, which appears to be constitutively expressed in human cells. Polyamines are essential for cell growth and normal function, and *MTAP* has been shown to have a role in inhibiting cell growth under certain conditions *in vitro*[203]. It was speculated that variants may influence plasma levels of

homocysteine which have been associated with CAD risk, although a recent study reported no variation in homocysteine levels in relation to genotype of the CAD risk SNP rs10757274[144]. The mechanism through which *MTAP* might influence CAD susceptibility is unclear at present.

1.2.4 Clinical implications

1.2.4.1 Risk prediction

The finding of new loci associated with CAD has attracted great excitement in the medical and scientific communities, but the way in which these findings could translate into tangible health benefits needs to be considered. The use of SNP markers for genetic testing to improve risk stratification of individuals and the promise of 'personalised medicine' has been an often cited prospect. However, statistical significance is not the same as clinical relevance, and the magnitude of the risk conferred by SNPs at the chromosome 9p21 locus is modest for individuals, with relative risks of ranging from 1.18-1.36 for heterozygotes, and 1.26-1.72 for homozygotes. Because the frequency of the risk alleles is high, the population attributable risk associated with the chromosome 9p21 locus has been estimated to be between 10 and 22%, which is significant from a public health perspective, but it should be remembered that the case-control cohorts which have been used in studies to date have been specifically collected to be 'genetically loaded', predominantly including early onset cases and families in whom there is an affected relative-pair. The generalisability of results regarding the relative magnitude of a genetic contribution to risk in such selected cohorts to the entire population at risk of CAD is therefore questionable.

From a practical point of view, the contribution of genotype at associated SNPs to the capacity to predict CAD in a representative population assessed for the 'classical' CAD risk factors is perhaps the most important question. Several studies have investigated this. Talmud *et al* typed the chromosome 9p21 CAD risk SNP rs10757274 in 2,742 men aged 50-64 years old who were free of CAD at baseline, who experienced 270 CAD events over a 15-year follow-up period[144]. The area under the receiver operating characteristic (ROC) curve for the 'classical' risk factors was 0.62 and did not increase significantly when rs10757274 genotype was included. 369

men (13.5%) were reclassified into more accurate risk categories (based on four categories of 10-year CAD risk), which the authors argued might have clinical utility. However, the analysis is biased by the fact that the study did not include family history of CAD in the 'classical' risk model they used. Brautbar *et al* reported similar findings in an analysis in 9,998 Caucasians, 1349 of whom developed incident CAD over 14.6 years of follow-up[204]. Addition of rs10757274 genotype produced a statistically significant, but not clinically meaningful, change in the area under the ROC curve from 0.782 to 0.786, and reclassified 12% of individuals in the intermediate-low and intermediate-high risk categories. Paynter *et al* performed a similar analysis in 22,129 white females aged over 45 (median age 52) years, who experienced 715 cardiovascular events over 10 years of follow-up[205]. Addition of genotype at rs10757274 to a model based on traditional risk factors, CRP and family history of CAD did not improve risk prediction measured using the c-index. The addition of genotype at rs10757274 also failed to improve the Net Reclassification Improvement score.

Simulations performed by Talmud *et al* suggested that if 10 additional SNP genotypes of similar effect to rs10757274 could be identified, the area under the ROC curve could be increased to 0.76, a gain in predictive capacity which the authors judged to be potentially useful[144]. However, it was noted that the proportion of individuals segregating multiple independently associated SNPs in any population would be low, limiting the likely clinical utility of such SNPs as predictors even if discovered. Moreover, the chromosome 9p21 locus is the strongest common genetic risk factor for CAD that has been discovered so far, and it seems unlikely that additional common loci will be discovered that confer a similar level of risk. A subsequent analysis published in February 2010 by Paynter *et al* studied the predictive value of a genetic risk score based on genotypes at 101 SNPs reported to be associated with cardiovascular disease phenotypes with $P<10^{-7}$ (and $r^2<0.5$ between included markers)[206]. In the same cohort as their previous study the genetic risk score did not improve risk prediction or reclassification compared to classical risk factors. After correction for other risk factors, the genetic risk score was not associated with cardiovascular disease risk, whereas self-reported family history remained significantly associated. No studies have shown that testing 9p21 genotypes leads to improvements in measurable health outcomes, and although some private companies

are already offering testing of chromosome 9p21 SNPs for prediction of cardiovascular risk[207], the current data do not support the use of such an approach in routine clinical practice.

1.2.4.2 Novel insights into pathophysiology

More important than the role of such markers for risk prediction is the potential of these genetic variants to provide novel insights into pathways involved in the pathophysiology of disease. This can enhance our fundamental understanding of the biology and ultimately identify biomarkers for risk stratification and novel targets for treatment. Genetic studies that identified the relationship between FLAP and CAD have already led to human trials evaluating the effect of inhibitors of this pathway on clinical parameters[100]. The biological importance of such insights are not necessarily related to the strength of the genotype-phenotype association since the strength of association of a gene with a trait does not predict the potential effect of a drug designed to agonise/antagonise the product of that gene. For example, common variants in the *HMGCR* gene involved in endogenous cholesterol synthesis have only a modest influence on plasma lipid levels[118], yet statins which act by inhibiting the HMGCR enzyme have very significant effects on LDL reduction and hard clinical endpoints such as reduction of death and cardiovascular morbidity[208]. Similarly, common variants of the *KCNJ11* gene, which encodes the sulfonylurea receptor, have an odds ratio of 1.14 for diabetes in GWA studies[125], but this gene has been the target of successful diabetes therapies that are in widespread clinical use[209, 210]. Genes and pathways involved in mediating disease associations may be identified by studying the influence of risk variants on gene expression, as discussed below.

1.3 Assessing effects of risk loci on gene expression

1.3.1 Variation and regulation of gene expression

Phenotypic variation is a consequence of genetic and non-genetic (environmental) factors. DNA sequence variation may act by modifying either the nature of a gene product (its 'quality') or its expression levels (its 'quantity'). Before the results of recent GWA studies were available, most DNA variants that had been convincingly associated with human disease were in coding regions and their effects were mediated

through alterations in protein structure. This was predominantly a reflection of the fact that such variants were easier to identify than those that influence gene expression; variants affecting amino acid sequence often have large phenotypic effects, and since their location is predictable, the coding regions for many genes have been systematically examined[211]. Although these exonic polymorphisms are important causes of Mendelian disease, they are unlikely to account for most of the heritability of complex diseases such as CAD. This has been confirmed by the results of multiple GWA studies reported since 2007, in which the SNPs associated with disease are frequently located in non-coding regions[25, 212], with sequencing and mapping studies failing to reveal exonic variants in nearby genes that could account for the effects[19, 20, 212].

Mutations in non-coding regions causing disease have been well described for a number of disorders, including some complex diseases[213-217], but the regulatory elements and their potential associations with disease remain poorly characterised for most genes[218]. A number of factors complicate the genetic study of gene expression[211, 219]: multiple regulatory elements may influence expression of a gene; the location of regulatory elements may be difficult to predict (they may be within or near to the gene, but may be hundreds of kilobases away or on different chromosomes); the magnitude of phenotypic effect caused by individual variants in these elements may be small; effects may vary between tissues; the influence of regulatory elements on expression may vary depending on other factors (such as environmental or physiological conditions); and expression profiles can be altered by sample handling or immortalisation of cells in the formation of cell lines. Of particular relevance to expression studies involving the chromosome 9p21 locus, expression of *CDKN2A* and *CDKN2B* in murine embryos has been shown to be altered by the process of culturing cells, with further changes as these cultured cells approached senescence[220].

1.3.2 *Cis* versus *trans* effects on expression

Studies of gene expression profiles using RNA microarrays have confirmed the presence of significant inter-individual variation in gene expression levels in human populations, and demonstrated that heritable genetic factors are important determinants of expression levels[221-224]. Genetic influences on expression levels may

act in *cis* or *trans*[218, 225]. *Cis*-acting elements act on genes that are on the same chromosome and affect transcript synthesis or stability in an allele-specific manner. These include regulatory elements such as promoters and enhancers which are usually found close to the genes they regulate, although they may be hundreds of kilobases away[226]. *Trans*-acting factors, such as transcription factors, are not usually located close to the genes whose expression they regulate (often being located on a different chromosome), and they affect the transcripts of both alleles of a gene. These *trans*-acting factors are themselves often regulated by other genetic or environmental influences.

It has been estimated that *cis*-acting variation accounts for 25-35% of genetically determined variation in gene expression between individuals[225]. However, estimates derived from genome-wide studies of total expression which classify *cis* and *trans*-acting effects based on the proximity of the variant to the gene (typically defining *cis* effects as those within 100kb of a gene[227]) are likely to be inaccurate since distance cannot reliably distinguish effects that act in *cis* or *trans*. Furthermore, physiological feedback mechanisms which tightly regulate total expression levels of a gene can prevent the detection of *cis*-acting effects in analyses measuring only total expression levels and underestimate the contribution of effects acting in *cis*[225]. Evidence of common *cis*-acting effects have been detected in around 20-40% of human genes in large studies[228, 229], although these are likely to be underestimates due to the relative insensitivity of the microarray platforms used in these studies to detect small effects.

Identification of *trans*-acting variants influencing gene expression is relatively difficult since *trans*-acting elements may be located on any chromosome and genome-wide approaches to discovery are therefore required. Since the risk variants for CAD identified in recent GWA studies are not found in mature transcripts they do not encode diffusible *trans*-acting factors and are therefore likely to influence expression in *cis*. The different genetic variants that may influence expression in *cis*, and the strategies used to detect and map such effects, are discussed below.

1.3.3 Genetic variation influencing expression in *cis*

1.3.3.1 SNPs

SNPs associated with *cis*-acting effects on gene expression have been widely reported[230]. Most of these are located in upstream promoter regions and exert their effects through alteration of transcription[225], although the involvement of variants in distant enhancers and other regulatory elements that also influence transcription has been increasingly appreciated[212, 231, 232]. Variants have also been shown to have *cis*-acting effects through other mechanisms including alteration of mRNA stability[233], mRNA processing efficiency[234], mRNA splicing[215], or via epigenetic changes such as DNA methylation[235]. The non-coding SNPs associated with CAD in the chromosome 9p21 region may influence disease susceptibility directly through such mechanisms, or the association may be indirect, due to LD with other variants which have functional effects. Studies investigating SNP associations with expression of genes in the chromosome 9p21 region are discussed in detail in Chapter 4.

1.3.3.2 Microsatellites and variable number tandem repeats

In addition to SNPs, other polymorphisms involving DNA segments of various sizes may have *cis*-acting influences on expression. Microsatellites and minisatellites are repetitive sequences of 1-6 bp and 6-100 bp elements respectively that are present throughout the human genome. These elements can influence expression in *cis* by interfering with transcription[236], altering DNA methylation[237], changing chromatin structure[238], or changing the affinity for transcription factor binding[239, 240]. Increasing microsatellite size has been associated with both increased and decreased gene expression in different settings[238, 241]. Changes in expression associated with microsatellites can lead to human diseases such as the fragile X syndrome[242] and Friedreich ataxia[236]. Like CNVs, microsatellite alleles may also be tagged by SNPs, such that functional microsatellite effects could account for SNP associations with disease.

Microsatellites influencing transcription may be located in various positions relative to the genes involved, but have been most frequently reported in promoter regions. Proximity to the 5’ end of transcripts and a high density of CpG islands have been

shown to be predictive of regions with promoter functions[243]. The arrangement of genes in the chromosome 9p21 region is interesting as *CDKN2A-ARF* and *ANRIL* are transcribed in opposite directions, with just 300 base pairs separating the transcription start sites of *ARF* and *ANRIL*. This region overlaps a CpG island, contains a number of conserved elements, and has been shown to have promoter activity for the ARF transcript of *CDKN2A*. However, bidirectional promoters have been described at other loci with similar gene arrangements and it therefore seems possible that this region also has promoter activity with respect to *ANRIL*[244]. A microsatellite is located close to this region, within the first intron of *CDKN2A-ARF* and 5' to *ANRIL*, as shown in Figure 1.6. This microsatellite does not have a specific ID, but is annotated in the SNP database as an insertion/deletion polymorphism, rs10583774. That identifier will be used throughout this thesis. Its proximity to the promoter region and evidence of cross-species sequence conservation suggest that it is subject to selection pressure and may have functionally important influences on expression of these genes. As shown in Figure 1.7, *in silico* analysis using PROMO v3.0.2 software[245-247] showed multiple transcription factor binding sites in the region surrounding the microsatellite, which provides further evidence for involvement of this region in the regulation of gene expression. Transcription factor activity may be variably disrupted by the number of 'TG' repeats, which is known to be polymorphic at this site[42]. As previously shown in Figure 1.5 (page 36), it is in an LD block with CAD risk variants and variation at this microsatellite influencing expression could therefore contribute to CAD susceptibility at this locus.

Figure 1.6. Location and sequence conservation of microsatellite rs10583774.

Panel A shows the location of rs10583774 relative to *CDKN2A-ARF*, *ANRIL* and CpG islands. Panel B is a magnified view showing cross-species sequence homology for the repeat sequence within rs10583774. Figure reproduced from the UCSC Genome Bioinformatics Database[44].

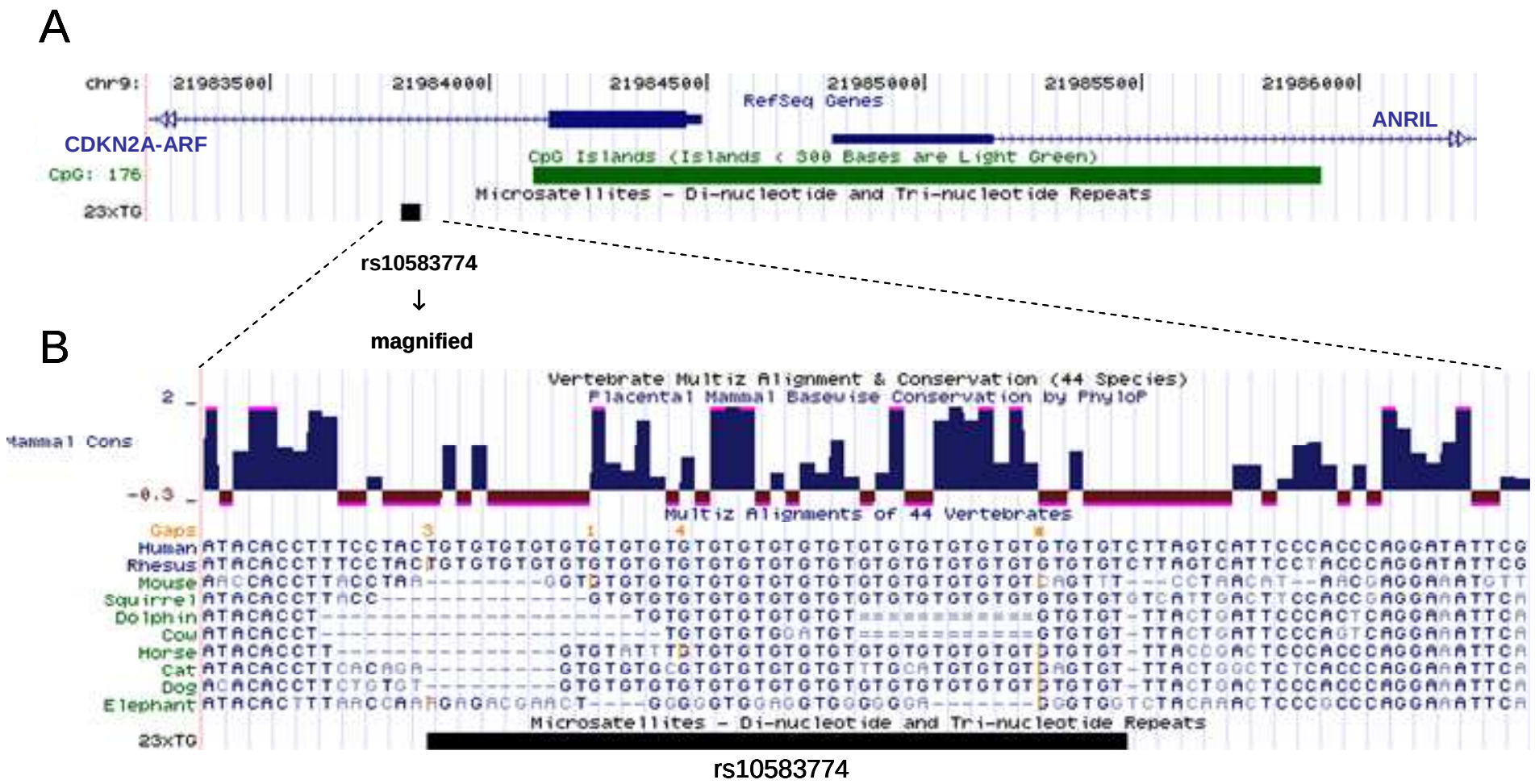
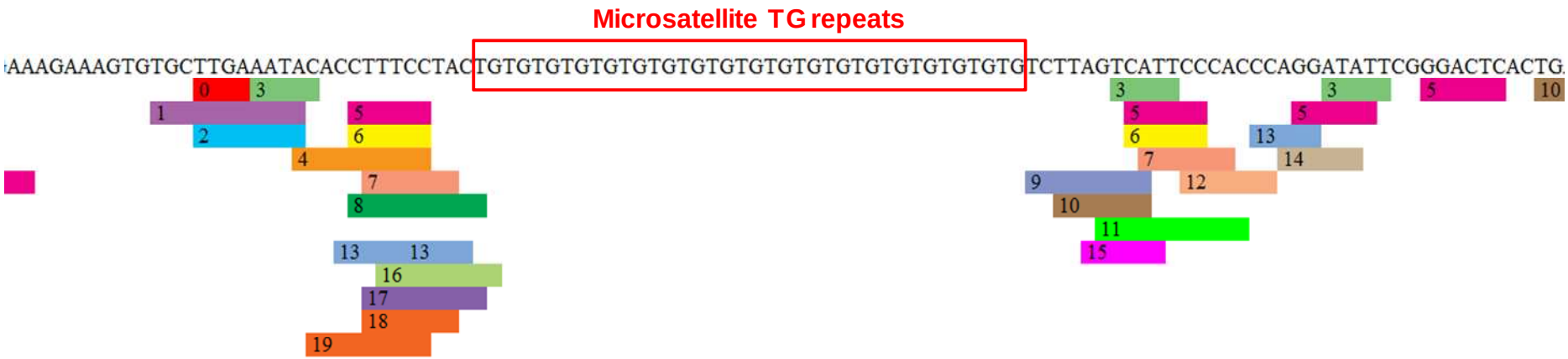


Figure 1.7. Human transcription factor binding sites around microsatellite rs10583774.

Microsatellite repeats and surrounding sequence shown along the top. Coloured bars represent human transcription factor binding sites. The microsatellite is located in a region with a high density of transcription factor binding sites. Figure produced using PROMO v3.0.2 software[245].



Key to human transcription factor binding sites:

1.3.3.3 Insertion/deletion variants and CNVs

Large-scale pathological duplications and deletions have been known to occur in the human genome for many years from cytogenetic observations, but the widespread occurrence of CNVs within the human genome has been only recently appreciated[248]. Although the term CNV may be used to refer specifically to deletions or duplications involving stretches of DNA larger than 1kb[249], smaller insertion/deletion (InDel) variants could also influence expression in *cis*, and in this thesis the term CNV is used to refer to copy number changes of any size. The reported mechanisms through which CNVs may influence expression include altering gene dosage, disrupting coding sequences, or interrupting regulatory elements[230]. Such structural variants have been associated with a range of human diseases, including complex phenotypes such as schizophrenia[250] and autism[251]. Common CNVs and SNPs may be in LD in the human genome[252], such that SNP associations seen in GWA studies could be the result of the SNPs 'tagging' functional CNVs. The possibility of CNVs in the chromosome 9p21 region is discussed in detail in Chapter 5.

1.3.4 Assessment of *cis*-acting effects on expression

1.3.4.1 *In vitro* approaches

Cis-acting regulatory factors have traditionally been investigated using *in vitro* approaches[225]. Transfection assays are often used to monitor the transcriptional activity of a synthetic reporter construct and assess whether a candidate regulatory polymorphism influences gene expression. However, there are several limitations of these techniques. First, appropriate selection of the regulatory elements to be incorporated into the construct is required, yet these are often poorly defined. Empirical targeting of upstream flanking sequences is frequently performed, but such sequences may not capture the complete regulatory elements that are biologically-active, and do not consider the effects of covariation at other loci. Second, *in vitro* transfection studies do not always model the situation *in vivo* accurately. The prevailing environmental conditions and *trans*-acting influences *in vitro* are likely to be very different to those *in vivo*, which may alter the overall effect of polymorphisms in regulatory regions. The problems with extrapolation of *in vitro* data was highlighted by Cirulli *et al*, who demonstrated that *in vitro* assays failed to predict *in*

vivo effects of regulatory polymorphisms in four genes well documented through reporter assays to have promoter polymorphisms influencing expression[253].

1.3.4.2 *In vivo* approaches: total expression versus allelic expression imbalance

The most commonly used approach for investigating *cis*-acting influences on expression *in vivo* has been to compare total expression levels (without differentiating the contributions from each of the two alleles) between individuals with different genotypes at the putative *cis*-acting locus. This approach treats total expression levels as a quantitative trait that is modified by the putative *cis*-acting locus (expression quantitative trait locus; eQTL). The proximity of the eQTL to its target gene is seen as evidence for its *cis*-acting effect. However, this approach is not specific for *cis*-acting effects since total expression levels reflect the net effect of both *cis* and *trans*-acting influences. The sensitivity to detect *cis*-acting effects is therefore reduced in the presence of significant variation in *trans*-acting influences.

An alternative approach that is specific for mapping *cis*-acting influences is to compare the relative expression of each allele within an individual. An unequal amount of transcript arising from each allele in an individual who is heterozygous for a transcribed polymorphism, termed allelic expression imbalance (AEI), indicates the presence of *cis*-acting influences on expression[218]. Correlation between genotype and quantitative allelic expression ratios can be tested to identify SNPs associated with *cis*-acting effects (allelic expression QTL, aeQTL)[254]. While traditional analysis using total expression levels assesses the influence of polymorphisms by comparing expression *between* samples, AEI analysis compares the expression levels of alleles *within* individual samples. This makes it much more robust to *trans*-acting influences that affect both alleles, such as experimental variability and inter-individual variation in other genetic/environmental factors, thereby maximising the sensitivity for detecting *cis*-acting effects. In contrast to the *in vitro* methods discussed above, alleles are studied in their native environment with respect to the haplotype, chromatin, and tissue context.

Whereas eQTL analysis uses information from all members of the population, a limitation of aeQTL analysis is that AEI can only be measured in individuals who are heterozygous for a transcribed polymorphism. A suitable polymorphism must therefore be present at a reasonable minor allele frequency in the population to permit sufficient numbers of informative individuals to be feasibly studied. Sensitive techniques for accurately determining the relative expression of two alleles are also required. AEI assessment may be possible using intronic SNPs in heteronuclear RNA, but this is more challenging and in general has only been shown to be successful for highly expressed genes[228, 255]. AEI can result from epigenetic factors such as DNA methylation.

For assessment of AEI to be a useful tool for the investigation of *cis*-acting genetic regulation in complex disease it must be both common and substantially influenced by genetic polymorphisms. A number of surveys using different methodologies have confirmed that AEI is heritable and can be detected for 20-50% of human genes[225, 228, 256-259]. To date, allelic expression analysis has provided evidence for sites of *cis*-acting regulation for a number of genes involved in a range of human diseases[260-265].

AEI analysis has been used to confirm *cis*-acting loci previously identified as eQTLs[266], and eQTLs have been used to validate loci showing AEI[228], but the relative power of the two approaches to identify *cis*-acting loci has never been systematically compared. Identifying the most efficient methodology for mapping *cis*-acting elements is important since this minimises the sample size that is required to detect variants with significant effects. This is particularly relevant since tissue-specific influences on expression may require the analysis of tissues where large sample collections may be difficult to establish[219]. Additionally, expression may be affected by several polymorphisms simultaneously and increased power may allow the contributions of individual sites to be investigated.

1.3.4.3 Simulations comparing the power of eQTL and aeQTL mapping

Simulations to explore the power of eQTL and aeQTL analysis to detect the effects of polymorphisms affecting expression in *cis* were conducted in collaboration with my co-supervisor Dr Mauro Santibanez-Koref (Institute of Human Genetics, Newcastle

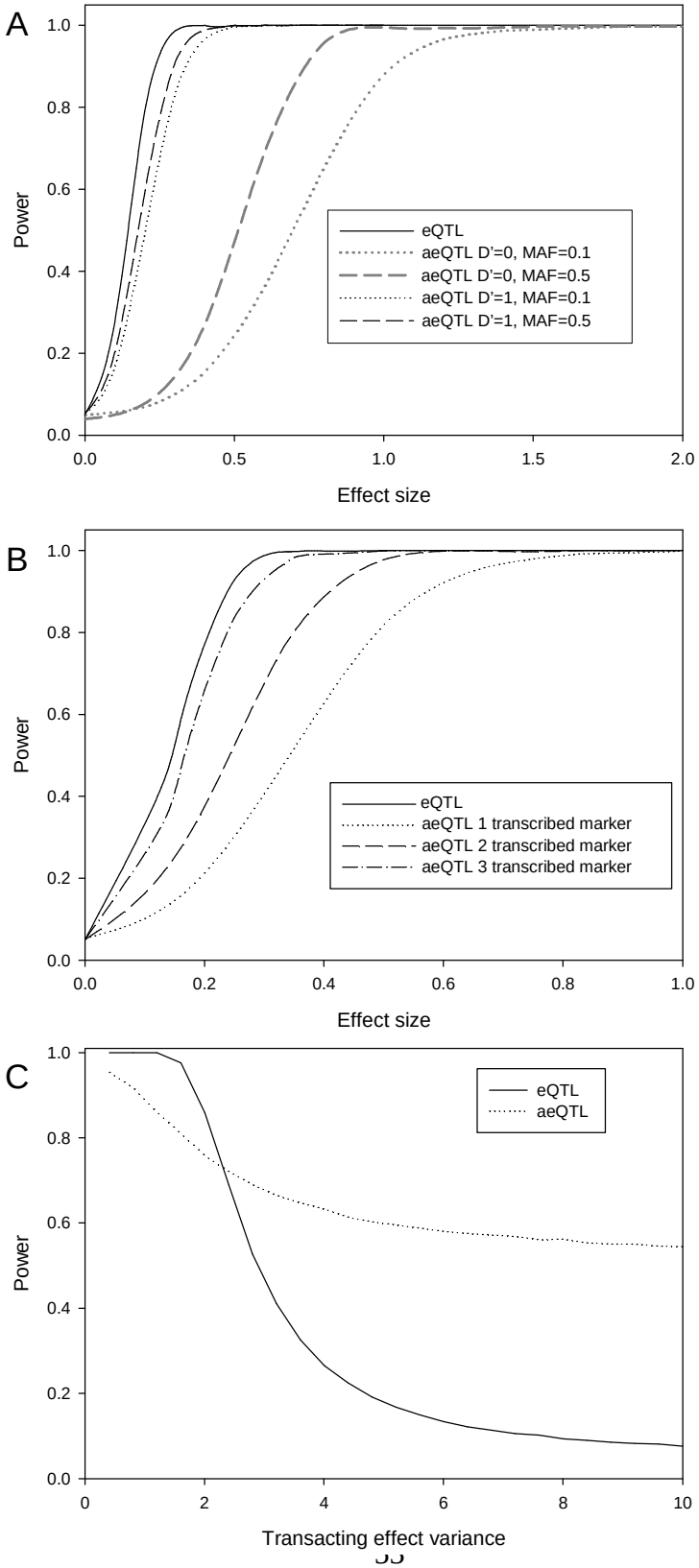
upon Tyne). Simulations were based on one transcribed and one *cis*-acting polymorphism, which were in linkage equilibrium unless otherwise specified. The default MAF was set at 0.1 for the transcribed marker and 0.2 for the *cis*-acting polymorphism. For each parameter configuration 10,000 replicates were analysed. Representative results of these simulations are displayed in Figure 1.8.

Panel A shows the relationship between power and effect size for a single *cis*-acting polymorphism in the absence of other influences on transcription. The power of the eQTL approach is higher than that of aeQTL analysis, since it uses information from all members of the population. The power of aeQTL analysis increases with increasing LD between the transcribed and *cis*-acting polymorphisms, and also as the heterozygosity of the transcribed polymorphism increases. The latter reflects the fact that allelic expression can only be measured in individuals who are heterozygous for a transcribed polymorphism and suggests that the power may be increased by combining information from several transcribed polymorphisms. Such a strategy is illustrated in Panel B, which shows the effect of combining data from multiple transcribed polymorphisms, each with MAF of 0.1. Increasing the number of transcribed polymorphisms increases the number of individuals in which allelic expression can be measured (assuming little LD between the transcribed SNPs) and therefore increases the power of the aeQTL analysis.

The above simulations only considered a single *cis*-acting effect influencing gene expression, but a more realistic scenario would be to also take into account the effects of *trans*-acting variation. As previously discussed, *trans* effects such as response to environmental stimuli, tissue specific factors, and experimental variables act on both alleles and can differ between individuals. Comparing alleles within a sample should minimise the effect of *trans*-acting variation, thereby maximising the power to detect *cis*-acting effects. A second set of simulations was therefore performed assuming that expression of both alleles was additionally affected by a *trans* acting factor, and that *cis* and *trans* effects act in an additive manner. In these simulations, total expression for one individual was defined as the sum of the contributions from both chromosomes (i.e. $e_1 + e_2$) and the ratio e_{1_1} / e_2 was only available for individuals heterozygous for the transcribed allele. As the *trans*-acting variance between individuals increases, the power to detect *cis*-acting effects decreases substantially

Figure 1.8. Comparison of the power of eQTL and aeQTL mapping to detect a *cis* acting polymorphism using simulated data.

Panel A investigates the effect of MAF at the transcribed locus and LD between the transcribed and *cis*-acting locus. Panel B shows the effect of using more than one transcribed marker, each with a minor allele frequency of 0.1. Panel C considers an effect acting in *trans* and explores the effect of increasing its variance across individuals. Effect size is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



faster for eQTL than for aeQTL analysis. This suggests that allelic expression will have greater power for mapping *cis*-acting effects in genes which are substantially influenced by *trans*-acting factors.

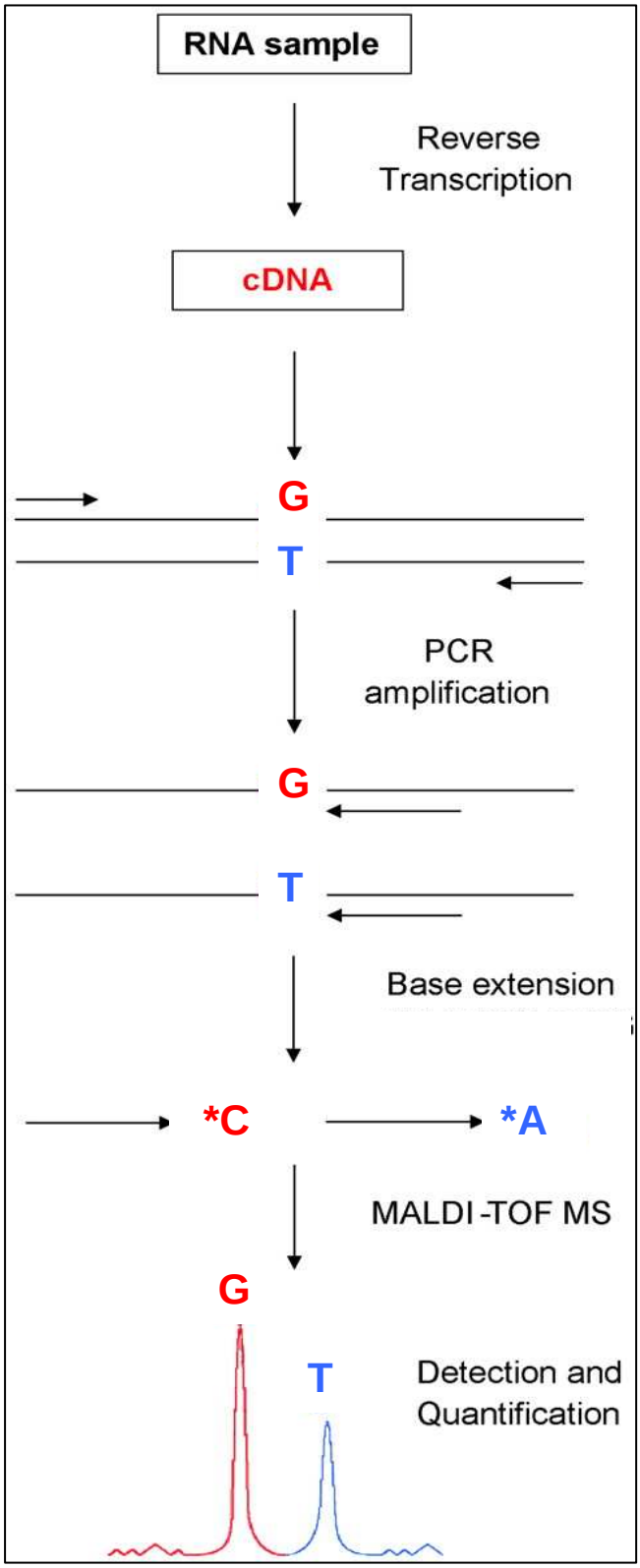
In the chromosome 9p21 region, many of the risk alleles associated with disease have MAF greater than 0.4[19-21] and there is substantial LD between risk variants and transcribed SNPs, suggesting that aeQTL mapping may be a powerful method for mapping *cis*-acting effects in this region even if there is relatively little *trans*-acting variation between individuals. A number of *trans*-acting factors have been previously shown to influence expression at the *CDKN2A*/*CDKN2B* locus[183], suggesting that aeQTL mapping may be more powerful then eQTL mapping, but the ability of both approaches needs to be compared experimentally since the extent of *cis* and *trans*-acting variation, as well as the manner in which such effects interact, is poorly characterised.

1.3.4.4 AEI assessment using the Sequenom platform

AEI has been assessed using a range of different technologies, based on the use of fluorescent dideoxy terminators[256], restriction-fragment length polymorphisms[260], microarray analysis[228, 258, 259], real time PCR[267], polymerase colonies[268], sequencing[269-271], and real competitive PCR with MALDI-TOF mass spectrometry on the Sequenom platform[272, 273]. Assessment using the Sequenom methodology is very sensitive for detection of AEI, allows accurate analysis from even small amounts of template, and offers the potential for assay multiplexing. Coefficients of variation of <10% from three independent reverse transcription reactions, and <3% for four PCR replicates from the same reverse transcription product have been reported[272].

The methodology for the Sequenom technique described by Ding and Cantor is summarised in Figure 1.9[272]. Essentially, total RNA from individuals heterozygous for a transcribed SNP is reverse transcribed to cDNA using random hexamers. PCR is performed using the cDNA template to amplify the region containing the SNP of interest. Because the two alleles are amplified in the same reaction and differ in sequence by only one base, the experimental conditions and reaction kinetics are the

Figure 1.9. Assessment of AEI using Sequenom (iPLEX).
Figure adapted from Ding and Cantor[272].



Total RNA from heterozygote for transcribed SNP is reverse transcribed with random hexamers. This produces cDNA containing each allele.

cDNA is amplified using PCR primers for the SNP of interest.

A single base extension reaction containing mass-modified nucleotides is performed to produce two oligonucleotide products with different molecular weights.

These two products are detected and quantified by MALDI-TOF mass spectrometry. The relative amounts of each allele are used to quantify the amount of allelic expression imbalance using the allelic expression ratio (AER).

same for each allele. Ideally, PCR primers should be designed to be cDNA specific (across exon boundaries) to prevent genomic DNA being amplified in the reaction.

The amplified PCR product is treated with shrimp alkaline phosphatase to remove excess dNTPs. A base extension reaction is performed using a single primer that binds adjacent to the SNP site and reaction mix containing mass-modified dNTPs. The resulting oligonucleotide products for each SNP differ by one base and have different masses. The products are cleaned up with resin to optimise mass spectrometric analysis and then dispensed onto a SpectroCHIP prespotted with a matrix of 3-hydropicolinic acid. The SpectroCHIP is analysed on the Sequenom platform by matrix assisted laser desorption/ionisation time of flight (MALDI-TOF) mass spectrometry. Under high vacuum the matrix is irradiated with an ultraviolet laser that vaporises and ionises the DNA/matrix. The ions are separated by their mass to charge ratio, and the peak area of each allele product is proportional to the relative amount of the allele that was present in the starting RNA. The AEI ratio is calculated by dividing the peak area of allele 1 by the peak area of allele 2.

The same process is repeated using gDNA to act as a control for biases in the detection method which could create an artefactual imbalance in allelic expression. Primers to amplify a portion of genomic DNA similar in size and sequence to the cDNA amplicon are used, with all other steps being identical. The corrected allelic expression ratio is the ratio in cDNA divided by the ratio in gDNA.

Detailed methodological issues relating to AEI assessment and normalisation are considered in Chapter 3.

1.3.4.5 Mapping genetic effects on expression

Mapping eQTLs using total expression may be performed by linear regression, in which expression levels are compared between groups of individuals who have zero, one, or two copies of the putative *cis*-acting allele. Mapping aeQTLs is more complex and requires genotypes to be phased. The principles underlying this process are outlined below.

Once the presence of AEI has been demonstrated at a transcribed polymorphism, providing evidence of *cis*-acting regulation of expression, the goal is to identify the causative polymorphisms responsible for (or at least predictive of) expression differences. In this context, AEI is characterised not only by its presence or absence, but also by which of the alleles is over-expressed and by the extent of the differential expression, quantified using the allelic expression ratio (AER). The principle behind the analysis is to compare the observed AER values with those that would be predicted under different assumptions. The simplest case is when an association between the transcribed marker itself and AER is analysed, and predictions from two models are compared. The first model assumes that allelic expression levels are independent of which allele is present at the transcribed locus, (i.e. there is no association between AER and genotype), in which case individual AER values would not deviate systematically from a 1:1 ratio. The alternative model is that one of the alleles is preferentially overexpressed, in which case a systematic deviation from a 1:1 ratio would be expected. Any test that compares the mean AER to a 1:1 ratio would be suitable to assess association (such as a t-test).

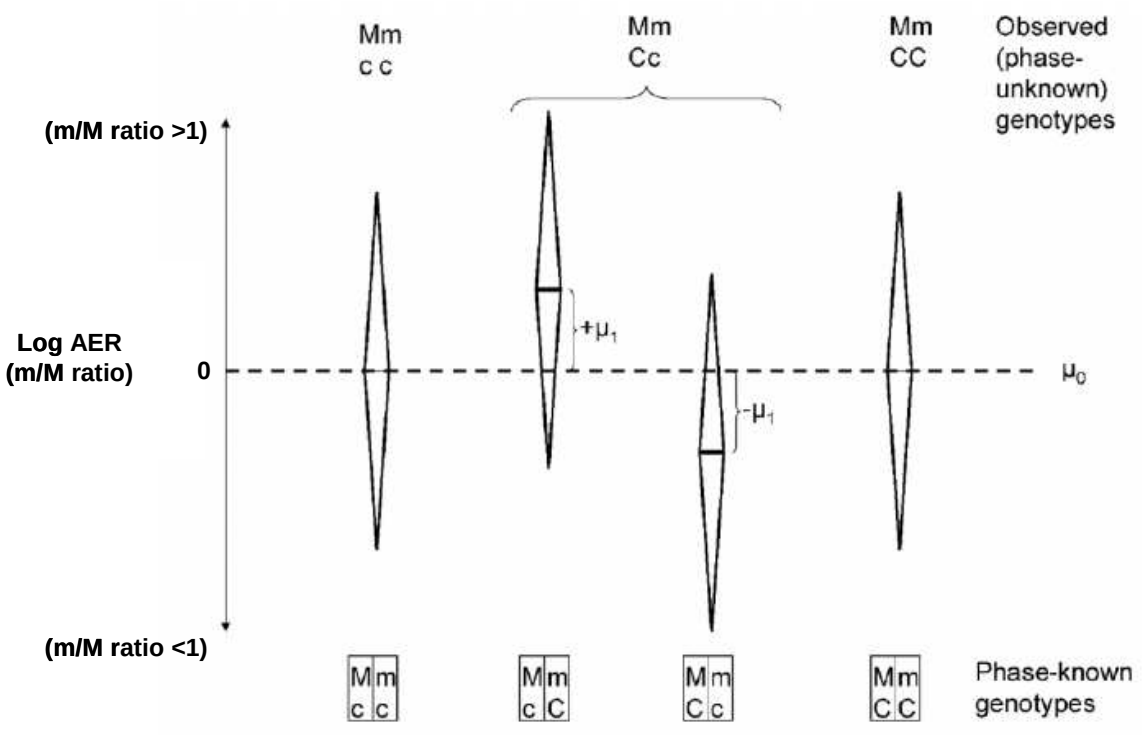
A more complicated situation is when trying to assess the *cis*-acting effect of a polymorphism that is not in the transcript. In this situation two polymorphisms need to be considered; the transcribed one and the potential *cis*-acting one. By experimental design all individuals are heterozygous at the transcribed SNP. An association between genotype at the *cis*-acting site and expression can be detected if there are differences between AER in individuals with different genotypes at the *cis*-acting locus. As above, the first model assumes that AER is independent of which allele is present at the *cis*-acting locus, hence individual AER values would not deviate from a specific ratio (1:1 if the transcribed marker has no effect). The second model assumes that genotype at the candidate *cis*-acting locus influences expression. This concept is illustrated in

Figure 1.10. The alleles at the transcribed marker are designated ‘M’ and ‘m’, and the alleles at the putative *cis*-acting marker as ‘C’ and ‘c’. In this example the transcribed polymorphism has no effect and the ‘C’ allele at the *cis*-acting locus causes overexpression of the transcript from the same chromosome. Alleles ‘MC’ and ‘mC’ will be expressed at the same level as each other, and overexpressed compared to ‘Mc’ and ‘mc’ (which are both expressed at the same lower level). Since only relative transcript levels at the transcribed polymorphism are compared, AEI will only be detected in individuals who are heterozygous at the *cis*-acting site (since genotypes ‘CC’ and ‘cc’ will have the same *cis*-acting effect on each allele of the transcribed polymorphism and therefore the ratio is 1:1 for both of these homozygotes). If the *cis*-acting polymorphism influences expression, the AER in individuals who are heterozygous for the *cis*-acting polymorphism deviates from that seen in homozygotes at the *cis*-acting site, as shown in

Figure 1.10. Since the 'C' allele causes overexpression, the 'MC/mc' phased genotype would cause relative overexpression of the 'M' allele, while the genotype 'Mc/mC' would cause the 'm' allele to be overexpressed. Therefore to determine the effect of the putative *cis*-acting polymorphism requires the phase between the *cis*-acting and transcribed polymorphisms to be estimated (i.e. the probability that individuals who are heterozygous at both sites have the genotypes 'MC/mc' or 'Mc/mC'). Estimating phase and effect can either be done simultaneously or as separate steps. The analysis is performed as a likelihood ratio test comparing the likelihood of the observations occurring assuming no effect from *cis*-acting polymorphism with the probability of the observations occurring assuming an effect of the putative *cis*-acting polymorphism. It assumes that the variance of the observations is independent of the genotype, that the ratios follow a log-normal distribution, and that the mean AER effects observed are representative of the true effects. The latter assumption means that the sample size has to be large enough to accurately assess the effect of a given genotype, which restricts the number of genotypes that can be reliably analysed for a given sample size.

Figure 1.10. Effect of *cis*-acting SNPs on AER at the transcribed marker.

All individuals are heterozygous for the transcribed marker (i.e.'M/m'). The four phase-known genotypes and the corresponding three phase-unknown genotypes are represented on the horizontal axis. The vertical diamonds represent the distribution of log AERs for each genotype, with the horizontal bar representing the mean. *Cis*-acting differences will be seen only in those individuals heterozygous at the *cis*-acting locus, hence the mean log AER is zero (corresponding to an AER of 1:1) in both 'CC' and 'cc' homozygous groups. The *cis*-acting effect is seen as a deviation from the 1:1 ratio in 'Cc' heterozygotes, but the direction and magnitude (μ_1)of this effect can only be estimated once the genotypes are phased (when it can be seen that the 'C' allele at the *cis*-acting SNP causes overexpression) . Figure adapted from Teare *et al*[254].



1.4 Summary and overall project aims

Recent GWA studies have identified a novel, well-replicated association between SNPs at the chromosome 9p21 locus and CAD. The mechanism of the association of SNPs at this locus with CAD is unknown, but elucidating the pathways involved may provide valuable insights into the pathogenesis of CAD and provide novel risk biomarkers and therapeutic targets.

An association of these SNPs with other diseases and intermediate phenotypes known to be important in cardiovascular disease may help to identify pathways involved in causation, and define the influence of these SNPs on other important cardiovascular diseases besides CAD.

The chromosome 9p21 SNPs associated with CAD are not in protein-coding sequence, but may act through *cis*-acting influences on the expression of nearby genes. Candidate genes in the region include *ANRIL* which produces a large non-coding RNA of unknown function, the cell-cycle regulators *CDKN2A* and *CDKN2B*, and *MTAP*, which encodes an enzyme involved in polyamine metabolism. Allelic and total expression can be used to investigate the influence of polymorphisms influencing expression in *cis*. An association between CAD risk SNPs and expression of particular genes would suggest that the SNPs lie in regulatory elements that influence gene expression, and that the genes implicated in this way may play a role in the causation of CAD.

This study aimed to investigate the association of chromosome 9p21 polymorphisms with intermediate phenotypes including traditional CAD risk factors and plasma levels of inflammatory mediators, as well as with other cardiovascular phenotypes including CIMT, LV structure/function, and congenital heart disease. The study also aimed to investigate whether 9p21 SNPs are associated with allelic and total expression of nearby candidate genes, and determine whether other polymorphisms in the region such as microsatellites and CNVs might also play a role.

A recent GWA study has reported an association between SNPs on chromosome 2q24 and BP, but the findings need to be investigated in additional independent cohorts. *In*

vitro functional studies suggested that the effect may be mediated by influences on expression of the nearby *STK39* gene, which encodes a protein that interacts with ion cotransporters involved in salt transport. Association of SNPs with allelic expression of *STK39* could be used to confirm the effects on expression of this gene *in vivo*.

This study aimed to investigate the association between reported risk SNPs for hypertension at the 2q24 locus and BP in a large British Caucasian cohort, and to use allelic expression analysis to determine whether these SNPs influence *STK39* expression *in vivo*.

Chapter 2

Materials and methods

2 Materials and methods

This chapter outlines the general materials and molecular biology methods used in the following chapters. Assay details specific to particular studies are highlighted in the materials and methods sections of the relevant chapters.

2.1 Materials

2.1.1 Participants and samples

2.1.1.1 Northeast (NE) British Caucasian cohort

Anonymised DNA and RNA samples were available for 187 healthy adults. Five samples were excluded from analysis on the basis of inconsistencies between the genotyping results obtained in the DNA and cDNA samples, and five samples subsequently found to be duplicates were also excluded. Of the 177 included samples, 152 were collected through a sub-study of the People of the British Isles study in 2007[274], and 25 were control samples collected for a study investigating the genetics of colon cancer (the CAPP study)[275]. 50% were male and the median age was 63 years (range 25-101, lower quartile 51, upper quartile 69).

The People of the British Isles study recruited healthy adults from different regions to investigate genetic variation within the UK. To be eligible, at least three of their grandparents had to be born within a 30-40 mile radius of one another, and ideally be from a rural population. Of the samples used in this study, 7 were from Cumbria, 1 from Lancashire, and the remainder from northeast England (Northumberland, Tyneside or County Durham).

The CAPP study controls were collected in 2006 as controls for a genetic study investigating colon cancer[275] and comprised four healthy adult staff members at the Institute of Human Genetics, and 21 phenotypically unaffected adults who had undergone screening for cystic fibrosis or haemochromatosis carrier status.

Informed consent was obtained from all participants and the studies were approved by the Newcastle and North Tyneside Local Research Ethics Committee.

2.1.1.2 South African (SA) cohort

DNA and RNA samples were collected from 310 healthy adult volunteers at a blood donor clinic at the University of the Western Cape, Cape Town, South Africa. Samples were anonymised, but limited demographic data including age, gender, and self-reported ethnicity were collected. Informed consent was obtained from all participants and the study was approved by the University of Cape Town Faculty of Health Sciences Research Ethics Committee.

The self-reported ethnicity of the SA cohort was: 200 Cape mixed-ancestry; 67 African black; 19 Indian; 10 white; 4 other/unknown. 42% were male, with median age 20 years (range 17-60, lower quartile 19, upper quartile 23).

2.1.1.3 *MLH1* validation samples

For the purpose of training and optimisation of AEI techniques, initial validation work was carried out to assess AEI in stored RNA samples using *MLH1*, as pilot studies investigating AEI in this gene had been previously performed. This validation work was performed using five heterozygous samples in which AEI had been previously assessed: two healthy volunteers without imbalance and three patients known to have imbalance of *MLH1* expression who were recruited for the CAPP study. Informed consent was obtained from all participants and the studies were approved by the Newcastle and North Tyneside Local Research Ethics Committee.

2.1.1.4 HTO cohort

DNA was available for this cohort comprising of 1425 members of 248 British Caucasian families ascertained through hypertensive probands and phenotyped for a quantitative genetic study of cardiovascular risk factors from 1993-2001[11, 156, 276].

Sample ascertainment

A detailed description of this series and the ascertainment strategy has been published previously[276]. Families were selected through a proband with essential hypertension whose systolic and diastolic BP were in the top 5% of the population distribution (defined as daytime ambulatory BP >140/90mmHg; three clinic BP measurements

>160/95mmHg; or treatment with at least two antihypertensive medications). Secondary hypertension was excluded using the screening protocol applied in the hypertension clinic. Families were required to consist of at least three siblings (including the proband) clinically assessable for BP if DNA from a parent of the sibship was available, or at least four siblings if no parental DNA was available. Qualifying sibships could be in the generation of the proband, or the offspring. There was no requirement for additional members of the family to be hypertensive, but where additional members of the sibship were found to have hypertension (using the same criteria), families were extended and the spouses and offspring of hypertensive members also collected. The majority (64%) of the individuals in the family collection therefore have BP within the conventionally accepted "normal range", and the family collection includes some extended families, though most are nuclear families. The median family size was 5 people, 60% of families comprising between 4 and 6 genotyped and phenotyped members. 71% of families were 2-generation and 29% were 3-generation. 84% of families had an assessable sibship in the generation of the proband, while 16% of families consisted of a proband and their nuclear family (spouse and children over 18 years) only. Informed consent was obtained from all participants and the study was approved by the Central Oxford Research Ethics Committee and Newcastle and North Tyneside Local Research Ethics Committee.

BP measurement

The BP measurement protocol has been previously described[276]. BP was measured using ambulatory monitoring for a period of 24 hours in all subjects willing to undergo monitoring, using the A&D TM2421 monitor. Three readings were taken with the patient in a relaxed seated position at the start of the monitoring period. Simultaneous auscultation was carried out by a trained observer, to confirm satisfactory (within 5mmHg) agreement between the monitor and auscultatory values; if this criterion was not met, the cuff was repositioned until satisfactory agreement was obtained. The three readings which had satisfactory agreement between the monitor and the observer in the final cuff position are referred to as "clinic readings". The monitor was programmed to record blood pressure every half-hour during the daytime and every hour during the night, and a recording was considered of satisfactory technical quality if at least 20 daytime ambulatory data points were

available for analysis. Patients also recorded the time they went to bed and rose in the morning to enable individualised calculation of the "daytime" and "night-time" periods. Mean values for systolic and diastolic blood pressures for the clinic, daytime and night-time periods were analysed for association with genotypes.

Echocardiography and CIMT measurement

Echocardiography and CIMT measurement was performed in a subset of individuals between 1999 and 2001, as previously described[11]. Carotid artery ultrasonography was performed in 953 individuals by two sonographers, with all measurements made by a single observer. The right and left common carotid arteries were scanned using a 7.5 MHz linear array transducer (HP Sonos 5500) and measurements were made from the far wall of the distal 10mm of the common carotid artery. Images were recorded for later offline analysis using computerised edge-detection software with manual editing. Mean and maximal CIMT measurements were performed at end-diastole on each side, and the average reading from these was calculated.

Trans-thoracic echocardiography was performed in 904 individuals according to a standard protocol in which 2D imaging, M-mode imaging and Doppler studies were performed. Left ventricular (LV) wall and cavity measurements were performed by a single observer from the standard M-mode images. LV mass (in grams) was estimated from M-mode images as recommended by the American Society of Echocardiography guidelines[277], using the formula described by Devereux *et al*[278]. Measurements of LV mass were corrected for body surface area as described by Levy *et al* in the Framingham heart study cohort[279]. Fractional shortening was calculated as 100(LVIDs+LVIDd)/LVIDd, and LV ejection fraction was estimated according to the formula described by Teichholz et al[280]. The E/A ratio was calculated using standard methodology from pulse-wave Doppler estimates of mitral valve inflow velocity[281].

Additional phenotyping

A full clinical history was taken, which included the subject's medical history and lifestyle factors including consumption of alcohol and tobacco, and habitual physical exercise. Anthropometric data including height, weight and waist and hip circumferences were measured. Blood samples were also analysed for other

phenotypes relevant to cardiovascular disease, using commercially-available assays. These included: plasma levels of total cholesterol, interleukin-6 (IL-6), tumour necrosis factor α (TNF-α), C-reactive protein (CRP)[282], and leptin[283].

2.1.1.5 Congenital heart disease cohort

The congenital heart disease cohort comprised 888 probands with congenital heart disease collected from two studies, the CHANGE study (Congenital Hearts: A National Gene/Environment study) and the FCH study (Freeman Congenital Heart Disease study)[284].

The multi-centre CHANGE study included Caucasian patients with tetralogy of Fallot (TOF), or the related conditions pulmonary stenosis/VSD and double outlet right ventricle[284]. Blood or saliva samples were obtained from patients at centres in Newcastle, Leeds, Bristol and Liverpool. Patients with recognised causes for TOF were excluded (including known chromosome 22q11 deletion syndrome, other malformation syndromes, known chromosomal abnormalities, developmental delay and learning difficulties, and known maternal exposure to significant teratogens during pregnancy). 444 TOF samples were available for analysis.

The FCH study included Caucasian patients with other forms of congenital heart disease recruited at the Freeman Hospital in Newcastle. Blood or saliva samples and phenotypic information were obtained from all patients. 444 non-TOF congenital heart disease samples were available for analysis.

Informed consent was obtained from all patients or their parents/guardians and the studies were approved by the regional ethics committees.

2.1.1.6 Cumbria control cohort

The Cumbria control cohort comprised 1,089 Caucasian women of childbearing age from the North Cumbria Community Genetics Project (NCCGP)[285] for whom usable DNA was available. Between 1996 and 2003 cord blood samples were collected from infants born consecutively at West Cumberland Hospital, and from 1999 to 2003 maternal blood samples were also collected[286]. DNA extracted from these maternal

samples was used for the analysis. Informed consent was obtained from all individuals and the studies were approved by the regional ethics committee.

2.1.2 Labware

All experiments were performed using standard sterile nuclease-free plasticware from recognised laboratory suppliers. Barrier pipette tips were used for all pre-PCR preparation steps, which were performed in designated laminar-flow hoods.

2.2 Methods

2.2.1 DNA extraction and quantification

2.2.1.1 DNA sample collection and extraction

For the SA cohort, peripheral blood samples for DNA analysis were collected in 5mL EDTA tubes using standard venesection from veins in the antecubital fossa. Samples were stored at -80^{o}C until extraction and were extracted within two months. DNA was extracted from the SA cohort using a standard phenol/chloroform method[287] by technicians at the Department of Molecular Biology and Human Genetics, University of Stellenbosch, Cape Town.

For the NE British Caucasian cohort, peripheral blood samples for DNA analysis had been previously collected using the PAXgene Blood RNA Kit (PreAnalytiX, Belgium). DNA was obtained from 1µL of the eluate from the RNA extraction performed without the DNase step, and amplified using whole genome amplification (as described in section 2.2.1.4).

For the HTO, Cumbria, and *MLH1* samples, DNA from peripheral blood had been previously collected and extracted using standard phenol/chloroform methods[287]. For the Congenital heart disease samples, DNA from either peripheral blood or saliva had been previously collected and extracted using standard methods: phenol/chloroform method[287] for blood samples or the Oragene Self-Collection Kit protocol (DNA Genotek, Canada) for saliva samples.

Stock DNA solutions were diluted to working aliquots of 20ng/µL in 1xTE buffer (Fluka Analytical, Sigma-Aldrich, UK) and stored at -20^{o}C.

2.2.1.2 Quantification by Nanodrop

DNA concentration was quantified by absorbance at 260/280nm using a NanoDrop ND-3000 Spectrophotometer (NanoDrop Technologies, USA) in 1µL of solution. Samples were mixed by vortexing or pipetting prior to measurement and two or three measurement replicates were performed per sample.

2.2.1.3 Quantification by PicoGreen

DNA quantification was also performed using the Quant-iT PicoGreen dsDNA kit (Invitrogen, USA) using the Thermo Fluoroskan Ascent FL (ThermoFisher Scientific, UK) following the manufacturer's standard protocol.

2.2.1.4 Whole genome DNA amplification

For the northeast Caucasian samples, 4µL of 1/10 dilution of the extracted RNA/DNA solution was amplified using GenomePlex Complete Whole Genome Amplification Kit (Sigma-Aldrich, UK) following the manufacturer's standard protocol.

2.2.1.5 Ampure DNA purification

Following WGA, DNA purification was performed using AMPure magnetic beads (Agencourt, Beckman Coulter, UK) following the manufacturer's standard protocol and eluting the DNA in 40µL of TE buffer. Amplified DNA concentration was quantified using optical densitometry, and working aliquots were diluted to a concentration of 20ng/µL in 1xTE buffer. All amplified DNA was stored at -20^{o}C.

2.2.1.6 Electrophoresis and visualisation of DNA

Samples and no-template controls from PCR reactions were analysed using agarose gel electrophoresis prior to Sequenom analysis or other downstream applications to confirm that the PCR reaction had been successful and exclude product in the no-template controls.

2.5% agarose gels were made using 2.5g of SeaKem LE Agarose powder (Cambrex, USA) in 100mL of 1xTAE, containing 1µL of 1000xGel Red per mL of gel. 1.5µL of Orange G loading dye (Sigma-Aldrich, UK) was added to 3µL of PCR product and samples were loaded alongside 5µL of GeneRuler 100bp DNA ladder (Fermentas, Germany) for size determination of the products. Electrophoresis was carried out between electrodes 30cm apart for 30-40 minutes at ~160V in 1xTAE buffer. Gels were visualised using a GeneGenius bioimaging system (Syngene, Synoptics Ltd, UK) and photographed using GeneSnap image acquisition software (Syngene, Synoptics Ltd, UK).

2.2.2 RNA extraction and quantification

2.2.2.1 General RNA procedures

RNA work was performed in designated areas using dedicated equipment where possible. External surfaces of equipment and working surfaces were prepared using RNase Away (Sigma-Aldrich, UK) prior to use, according to standard protocols. Designated RNA-only pipettes and sterile RNase-free labware were used for all RNA work. Sterile nuclease-free water (Sigma-Aldrich, UK) was used for all applications.

2.2.2.2 RNA sample collection

Peripheral blood samples for RNA analysis were collected in a 2.5mL PAXgene Blood RNA tube using standard venesection from veins in the antecubital fossa. Samples were incubated at room temperature for 2-6 hours then stored at 4^{o}C for up to 72 hours before being frozen, initially at -20^{o}C for 24 hours and then at -80^{o}C for long-term storage until extraction. RNA extraction was performed within 3 months of sample collection.

2.2.2.3 RNA extraction

PAXgene Blood RNA extraction and purification was performed according to the manufacturer's standard protocol, with omission of the DNase steps, such that the eluted solution contained both RNA and DNA. This permitted DNA from the sample to be used for genotyping. Nucleic acids were eluted in 80µL of elution buffer BR5 and stored in nuclease-free capped tubes at -80^{o}C.

2.2.2.4 DNase treatment

RNA samples were DNase treated using RQ1 RNase-Free DNase (Promega, USA) prior to reverse transcription following the manufacturer's standard protocol.

2.2.2.5 Whole transcriptome amplification

Whole transcriptome amplification was performed using Quantitect Whole Transcriptome Kit (Qiagen, Germany) following the manufacturer's standard protocol.

2.2.2.6 RNA LabChips

RNA and cDNA fragment size and quality were tested for selected samples using the RNA 6000 Pico LabChip kit (Agilent, USA) for the Agilent 2100 bioanalyzer, following the manufacturer's standard protocol.

2.2.3 Reverse transcription

For AEI measurements, approximately 2 μg of total RNA was reverse transcribed using SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen, USA) or Superscript VILO cDNA Synthesis Kit (Invitrogen, USA) following the manufacturer's standard protocol, and eluted in 20μL.

For teal-time PCR measurements, 500ng of total RNA was reverse transcribed using High capacity RNA-to-cDNA Master Mix (Applied Biosystems, USA) following the manufacturer's standard protocol, and eluted in 20μL. cDNA was stored at -80^{o}C.

2.2.4 Sequenom assays for genotyping and AEI analysis

2.2.4.1 Assay design

Genotyping was carried out using Sequenom iPLEX Gold technology using multiplexed assays. Initial AEI analysis was performed in uniplex using Sequenom hME technology based on the method reported by Ding and Cantor[273]. This methodology was subsequently adapted to utilise iPLEX Gold technology and allow multiplexing, as described below.

Multiplex assays for Sequenom genotyping were designed using Sequenom's RealSNP[288] and MassARRAY Assay Design v3.0 software (Sequenom, USA). This designs primers, checks for compatibility between primers in a multiplex, and performs a BLAST search. PCR primers which performed suboptimally were redesigned into subsequent assay designs, with manual selection of alternative PCR or extension primer sequences.

Primers for AEI analysis were designed manually and using Primer3 v.0.4.0 software[289, 290] and genomic specificity was confirmed using BLASTN against the human reference sequence in the NCBI database[196].

A full list of the assay primers can be found in Appendix 1. A 10 base non-complementary tag was added to PCR primers for Sequenom analysis to increase their mass so that they would not fall within the window of the extension products during MALDI-TOF analysis.

2.2.4.2 Primers

PCR and extension primers were supplied lyophilised (Metabion, Germany) and were resuspended in nuclease-free water (to a concentration of 100µM for PCR primers and 300µM for extension primers), aliquoted, and stored at -20^{o}C.

2.2.4.3 PCR

PCR setup was performed in a designated pre-PCR area in a laminar flow hood using barrier pipette tips and designated pre-PCR pipettes. DNA was added in a separate laminar flow hood with DNA pipettes. PCR reactions were performed using 10µL or 15µL reaction volumes in 96-well plates (ThermoFisher Scientific, UK) sealed with adhesive lids or capped thin-walled PCR tubes (ThermoFisher Scientific, UK) using a thermal cycler (BioRad DNA engine Tetrad 2, BioRad, USA). HotStar Taq polymerase (Qiagen, Germany), 10xPCR buffer (Qiagen), 25mM magnesium chloride (Qiagen, Germany), and 25mM dNTP mix (New England BioLabs, USA) were used for all reactions. Optimisation of annealing temperatures and PCR conditions was performed for AEI assays using 50-70^{o}C temperature gradients. No-template controls and duplicate samples were included for quality control. Agarose gel electrophoresis

of PCR products was performed for all plates to confirm successful PCR in sample wells and exclude PCR product in the no-template controls before proceeding to further analysis.

PCR conditions for multiplex iPLEX genotyping

PCR was performed using 20ng of DNA template in a 10μL reaction volume. A primer mix was prepared containing 5μL of each 100μM PCR primer solution in a total volume of 1000μL (final concentration of each primer 500nM). The 10μL reaction mix for each sample contained 4.7μL nuclease-free water, 1.25μL 10xPCR buffer, 0.65μL 25mM magnesium chloride solution, 0.20μL 25mM dNTP mix, 2μL primer mix, 0.20μL HotStar Taq (5U/μL), and 1μL DNA in 1xTE buffer. PCR conditions were 95^{o}C for 15min, then 35 cycles of (95^{o}C for 20sec, 56^{o}C for 30sec, 72^{o}C for 1min), then 72^{o}C for 1min, then 4^{o}C hold.

PCR conditions for multiplex iPLEX AEI assays

PCR was performed for samples heterozygous at the transcribed SNP using 25ng of cDNA template in a 10μL reaction for rs3217992/rs1063192, and a 15μL reaction for rs3088440/rs10965215 and rs11515/rs564398 (to allow PCR products to be divided into two separate plexes for Sequenom analysis). A primer mix was prepared containing 5μL of each 100μM PCR primer solution in a total volume of 1000μL (final concentration of each primer 500nM). The reaction mix for each sample contained 4.7μL nuclease-free water, 1.25μL 10xPCR buffer, 0.65μL 25mM magnesium chloride solution, 0.20μL 25mM dNTP mix, 2μL primer mix, 0.20μL HotStar Taq (5U/μL), and 1μL cDNA in elution buffer. PCR conditions for rs3217992/rs1063192 were 95^{o}C for 15min, then 45 cycles of (95^{o}C for 20sec, 56^{o}C for 30sec, 72^{o}C for 1min), then 72^{o}C for 1min, then 4^{o}C hold. Conditions were the same for the other assays except for a temperature of 58.6^{o}C in the annealing step.

Estimation of allelic ratios in genomic DNA samples heterozygous for the transcribed SNP was performed using the identical protocol with 25ng of gDNA template in place of the cDNA template.

PCR conditions for uniplex iPLEX AEI assays

PCR was performed for samples heterozygous at the transcribed SNP using 25ng of cDNA template in a 10µL reaction volume. The reaction mix for each sample contained 7.08µL nuclease-free water, 1µL 10xPCR buffer, 0.4µL 25mM magnesium chloride, 0.32µL 25mM dNTP mix, 0.08µL 25µM forward primer, 0.08µL 25µM reverse primer, 0.04µL HotStar Taq (5U/µL), and 1µL cDNA in elution buffer. PCR conditions for rs3217992 and rs1063192 were 95^{o}C for 15min, then 45 cycles of (95^{o}C for 20sec, 56^{o}C for 30sec, 72^{o}C for 1min), then 72^{o}C for 1min, then 4^{o}C hold. Conditions were the same for the other assays except for a temperature of 58.6^{o}C in the annealing step.

Estimation of allelic ratios in genomic DNA samples heterozygous for the transcribed SNP was performed using the identical protocol with 25ng of gDNA template in place of the cDNA template.

PCR conditions for uniplex hME AEI assays

PCR was performed for samples heterozygous at the transcribed SNP using 25ng of cDNA template in a 10µL reaction volume. The reaction mix for each sample contained 3.9µL nuclease-free water, 1µL 10xPCR buffer, 0.8µL 25mM magnesium chloride solution, 0.4µL 20mM dNTP mix, 0.1µL 5µM forward primer, 0.1µL 5µM reverse primer, 0.1µL HotStar Taq (5U/µL), and 3µL cDNA in elution buffer. PCR conditions were 95^{o}C for 15min, then 45 cycles of (94^{o}C for 20sec, 51^{o}C for 30sec, 72^{o}C for 1min), then 72^{o}C for 3min, then 4^{o}C hold.

Estimation of allelic ratios in genomic DNA samples heterozygous for the transcribed SNP was performed using the identical protocol with 25ng of gDNA template in place of the cDNA template.

2.2.4.4 SAP treatment

5µL of PCR product for each sample was transferred to a 384-well plate. A SAP reaction mix was prepared containing 1.53µL water, 017µL 10xSAP buffer (Sequenom), and 0.3µL SAP enzyme (1U/µL) for each sample. 2µL of SAP reaction

mix were added to each well using a Multimek 96 automated pipettor (Beckman Coulter, USA). The plate was sealed with an adhesive lid and centrifuged briefly at 700rpm, and then incubated in a thermocycler (BioRad DNA engine Tetrad 2, BioRad, USA). SAP reaction conditions were 37^{o}C for 40min, followed by 85^{o}C for 5min, followed by 4^{o}C hold.

2.2.4.5 Primer extension reaction

Extension reactions for iPLEX assays

For iPLEX extension reactions containing three or fewer extension primers, a reaction mix was prepared containing 0.20µL 10x iPLEX buffer, 0.20µL iPLEX termination mix, 0.0375µL of each 300µM extension primer solution, 0.041µL iPLEX enzyme, and water to make the volume up to 2µL per sample.

Because larger mass products are detected less efficiently in the MALDI-TOF detection system, a higher concentration of the larger mass extension primers was needed for highly multiplexed reactions (more then four extension primers). In this case, extension primers were divided into four approximately equal groups based on their mass, and the volume of 300µM extension primer solution added for each of the four groups is shown in Table 2.1.

Table 2.1. iPLEX extension reaction mix for multiplexed reactions

Reagent	Volume (µL) per sample
10x iPLEX buffer	0.20
iPLEX termination mix	0.20
Primer group 1 (300µM)	0.01875
Primer group 2 (300µM)	0.0249
Primer group 3 (300µM)	0.0312
Primer group 4 (300µM)	0.0375
iPLEX enzyme	0.041
Water	To make volume up to 2µL per sample

2µL of the reaction mix was transferred to each well of the 384-well plate containing the product of the SAP reaction using a Multimek 96 automated 96-channel pipettor (Beckman Coulter, USA). The plate was sealed with an adhesive lid, briefly

centrifuged at 700rpm, and then incubated in a thermocycler (BioRad DNA engine Tetrad 2, BioRad, USA).

For iPLEX genotyping assays extension reaction conditions were 94^{o}C for 30sec, then 40 cycles of (94^{o}C for 5sec, followed by 5 cycles of (52^{o}C for 5sec, then by 80^{o}C for 5sec)), then 72^{o}C for 3min, then 4^{o}C hold.

For iPLEX AEI assays extension reaction conditions were 94^{o}C for 30sec, then 30 cycles of (94^{o}C for 5sec, followed by 5 cycles of (52^{o}C for 5sec, then by 80^{o}C for 5sec)), then 72^{o}C for 3min, then 4^{o}C hold.

Extension reactions for hME assays

A reaction mix was prepared containing 1.728µL water, 0.20µL ACG termination mix, 0.0545µL of 100µM extension primer, and 0.018µL ThermoSequenase enzyme (32U/µL). 2µl of the reaction mix was transferred to each well of the 384-well plate containing the product of the SAP reaction using a Multimek 96 automated 96-channel pipettor (Beckman Coulter, USA). The plate was sealed with an adhesive lid, briefly centrifuged at 700rpm, and then incubated in a thermocycler (BioRad DNA engine Tetrad 2, BioRad, USA). Reaction conditions were 94^{o}C for 2min, followed by 55 cycles of (94^{o}C for 5sec, then 52^{o}C for 5sec, then by 72^{o}C for 5sec), followed by 4^{o}C hold.

2.2.4.6 Clean Resin step

Reaction products were desalted with Clean Resin (Sequenom) prior to MALDI-TOF analysis. For iPLEX reactions, 6mg of Clean Resin were added to each well of the 384-well reaction plate containing the product of the extension reaction, using a dimple plate. For hME reactions, 3mg of Clean Resin were added to each well of the 384-well reaction plate containing the product of the extension reaction, using a dimple plate. Samples were diluted with16µl of water using a Multimek 96 automated 96-channel pipettor (Beckman Coulter, USA).

2.2.4.7 Quantitative MALDI-TOF analysis

A Sequenom MassARRAY nanodispenser (Samsung, USA) was used to dispense 15nL of reaction product onto a 384-element SpectroCHIP (Sequenom, USA) and MassARRAY Typer version 3.4 software (Sequenom, USA) was used to perform MALDI-TOF analysis, using the appropriate settings for genotyping or AEI assays as specified in the manufacturer's protocol.

2.2.4.8 Determining genotypes

Spectra and automated genotype calls were manually reviewed and automated calls that appeared erroneous or ambiguous were excluded or manually reassigned. This process was performed at the time of plate analysis and was not influenced by results of subsequent data analysis. Individual samples with low genotype call rates (<80%) and SNP assays with poor quality spectra or cluster plots were excluded.

2.2.4.9 Estimation of allelic expression ratios

Spectra were manually reviewed and poor quality spectra were excluded. Allelic expression ratios were estimated as the ratios of the area under the peak representing allele 1 to that representing allele 2. Measurements were performed in four replicates for each sample. Results from amplification of genomic DNA were used as an equimolar reference to normalise the cDNA values.

2.2.5 TaqMan Genotyping

TaqMan genotyping was performed using predesigned TaqMan SNP Genotyping Assays (Applied Biosystems, USA), using the Applied Biosystems 7900HT Fast System. PCR was performed using 20ng of DNA template in a 5μL reaction volume in 384-well optical reaction plates. The reaction mix for each sample contained 2.5μL Universal PCR Master Mix No AmpEraseUNG (Applied Biosystems, USA), 0.125μL 40xAssay Mix (Applied Biosystems, USA), 1.375μL nuclease-free water, 1μL DNA solution in 1x TE. PCR conditions were 95^{o}C for 10min, then 40 cycles of (92^{o}C for 15sec, followed by 60^{o}C for 1min). No-template controls and duplicate samples were included for quality control. Data were analysed using the allelic discrimination algorithms in SDS Software v2.3 (Applied Biosystems, USA) according to the

manufacturer's protocol. Automatic genotype calls were manually reviewed and edited if necessary.

2.2.6 Quantitative real-time PCR

Quantitative real-time PCR reactions were performed using TaqMan gene expression probes and reagents (Applied Biosystems, USA), using the 7900HT Real-Time PCR System (Applied Biosystems, USA). PCR was performed using 25ng of cDNA template in a 15µL reaction volume in 384-well optical reaction plates. Reactions were multiplexed to include a FAM-labelled target gene assay and VIC-labelled control gene assay in the same reaction. The reaction mix for each sample contained 5.5µL nuclease-free water, 7.5µL 2xTaqMan Gene Expression Master Mix (Applied Biosystems, USA), 0.75µL 20x target gene primer/probe mix (Applied Biosystems, USA), 0.75µL 20x control gene primer/probe mix (Applied Biosystems, USA), and 0.50µL cDNA. PCR conditions were 50^{o}C for 2min, then 95^{o}C for 10min, then 40 cycles of (95^{o}C for 15sec, followed by 60^{o}C for 1min). Target genes and all control genes used for normalisation of the sample were analysed on the same plate, and no-template controls were included on each plate. All reactions were performed using four replicates. Specific assay details are given in Chapter 4.

Relative total expression was analysed using the comparative cycle threshold (Ct) method using SDS 2.3 and RQ Manager 1.2 software (Applied Biosystems, USA). Ct values for each target gene were normalised to the mean Ct value of the reference genes, which is more reliable than normalisation to a single endogenous control gene[291]. Normalised Ct values (= mean target gene Ct value – mean reference genes Ct value) were used for all analyses.

2.2.7 Microsatellite analysis

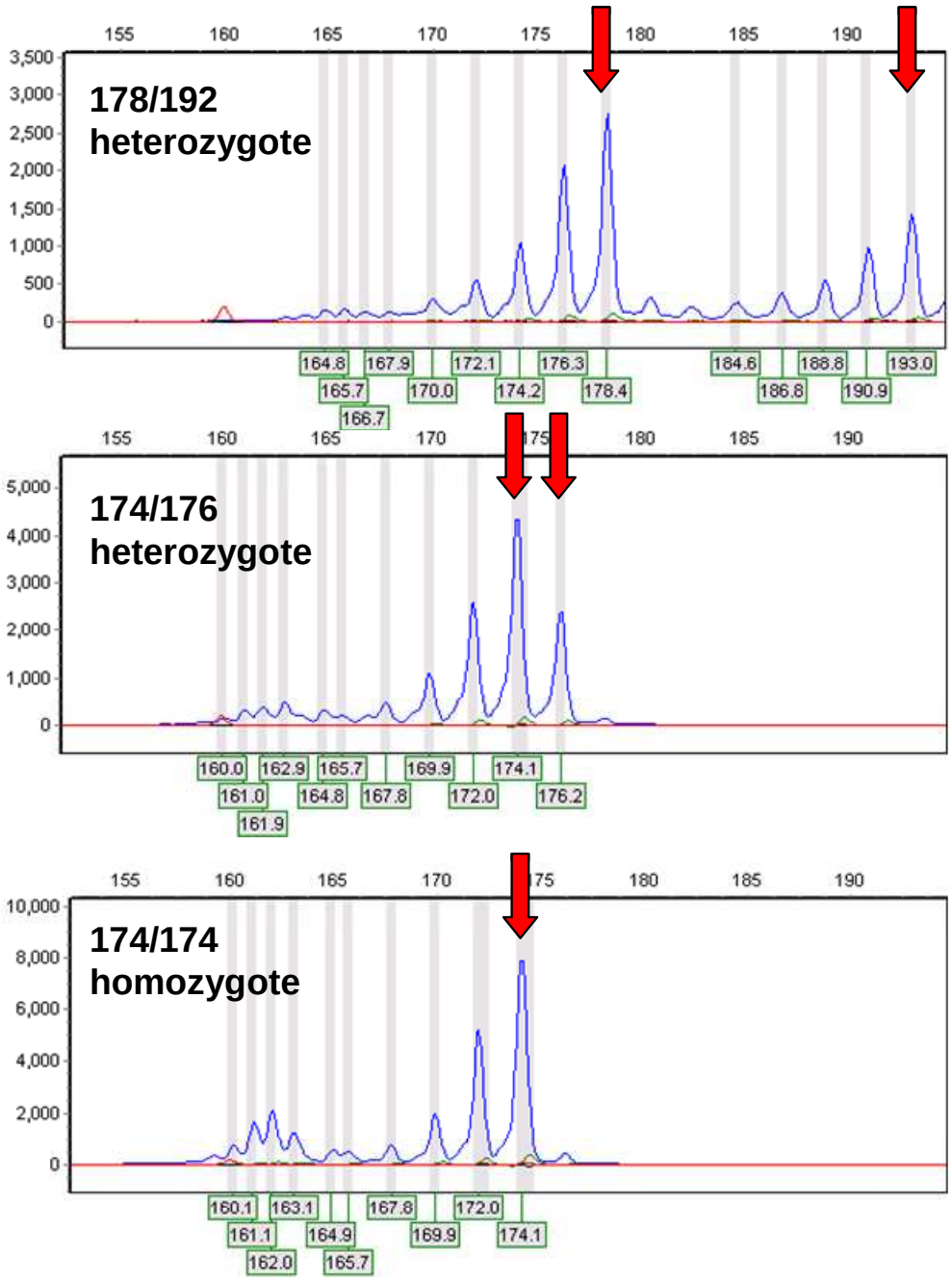
Microsatellites in genomic DNA were amplified by PCR with FAM-labelled primers, and the size of the products was quantified using capillary electrophoresis. PCR was performed using 25ng of cDNA template in a 10µL reaction volume in 96-well reaction plates. Custom PCR primers were used, with the forward primer FAM-labelled (Metabion, Germany). Primer sequences are shown in Appendix 1. The reaction mix for each sample contained 7.08µL nuclease-free water, 1µL 10xPCR

buffer, 0.4µL 25mM magnesium chloride, 0.32µL 25mM dNTP mix, 0.08µL 25µM forward primer solution, 0.08µL 25µM reverse primer solution, 0.04µL HotStar Taq (5U/µL), and 1µL DNA in 1xTE. PCR conditions were 95^{o}C for 15min, then 35 cycles of (95^{o}C for 20sec, 61^{o}C for 30sec, 72^{o}C for 1min), then 72^{o}C for 3min, then 4^{o}C hold. Duplicate and no-template controls were included. A master mix was then prepared containing 8.5µL of Hi-Di formamide (Applied Biosystems, USA) and 0.3µL of Genescan-500 ROX size standard (Applied Biosystems, USA) per sample. 8.8µL of the master mix was added to each well of a new 96-well semi-skirted, straight-edge PCR plate (StarLab, Germany), and 1µL of the PCR product was transferred to each well of this plate. The plate was covered in foil to prevent photo-degradation, before analysis by capillary electrophoresis on a 3130xl Genetic Analyzer (Applied Biosystems, USA). Data were analysed using GeneMarker v1.8 software (SoftGenetics, USA).

Microsatellite genotypes were manually assigned by a single observer (MSC), with selected spectra and genotype calls reviewed independently by another observer (MSK). Examples of selected spectra patterns and their interpretation are shown in Figure 2.1. The spectra estimate the product size of the PCR amplicon, which consisted of 55 bases before, and 79 bases after, the microsatellite TG repeats. The number of TG repeats in the microsatellite was therefore calculated as: N = (amplicon size – 134)/2 for each allele.

Figure 2.1. Examples of microsatellite analysis spectra.

Spectra show the amplicon size along the X-axis and peak height on the Y-axis. The boxed numbers on the X-axis represent the peak positions detected using the GeneMarker autodetection algorithm. The location of true peaks are highlighted with red arrows for each spectrum and the interpretation of each spectrum is shown in the top left of each panel. It can be seen that interpretation is complicated by the presence of 'slippage' peaks, which are seen as peaks of lower magnitude to the left of each true peak.



2.2.8 MLPA

Analysis for copy number variation was performed using custom SALSA MLPA (Multiplex Ligation-dependent Probe Amplification) analysis (MRC Holland). The SALSA MLPA P200-A1 reference probemix (MRC Holland, The Netherlands) was used to provide reference probes and control fragments. Custom MLPA probes were designed manually and using the H-MAPD Human MLPA Probe Design software[293, 294], using the default settings for probe design. From the H-MAPD output, probes were manually reviewed and sorted for additional design criteria not included in that software: left primer oligo (LPO) hybridising sequence including a maximum of two G/C nucleotides in the five nucleotides at the 3' end adjacent to the ligation site, and for LPO and right primer oligo (RPO) hybridisation sequences including a maximum of three G/C nucleotides directly adjacent to the primer recognition sequence). Probe lengths of 92-140 nucleotides at four nucleotide intervals were designed, with the first nucleotide of the LPO hybridisation sequence T for the shortest probes, G for probes of intermediate length, and C for the longest probes. Selected primers were checked for specificity using NCBI BLASTN, and checked against the human reference sequence to exclude the presence of SNPs in the hybridisation sequence using the Ensembl Genome Browser[43]. Sequences of the MLPA probes are shown in Appendix 1. All custom MLPA probes were ordered from Integrated DNA Technologies (Belgium), with probes longer than 60 nucleotides ordered as Ultramers.

MLPA was performed using the MLPA DNA detection/quantification protocol (MRC Holland, www.mlpa.com), using 100ng of DNA template in 5μL of 1xTE. The volume of the PCR reaction step was reduced to 25μL. The exact protocol used is shown in Appendix 1. Duplicate samples and no-template controls were included. Amplification products were analysed using capillary electrophoresis using a 3130xl Genetic Analyzer (Applied Biosystems). Data were analysed using GeneMarker v1.8 software (SoftGenetics), using custom panel templates for the probesets used. The SALSA MLPA P200-A1 reference probeset was used for internal normalisation of each reaction. This contains reference probes in regions not expected to demonstrate CNV, and also contains control fragments including the Q-fragments that indicates if insufficient amounts of DNA have been used, and the D-fragment that indicates if the sample was not completely denatured. The P200 probes are summarised in Table 2.2.

An example of the spectrum obtained from MLPA analysis using the chromosome 9p21 custom assay is shown in Figure 2.2.

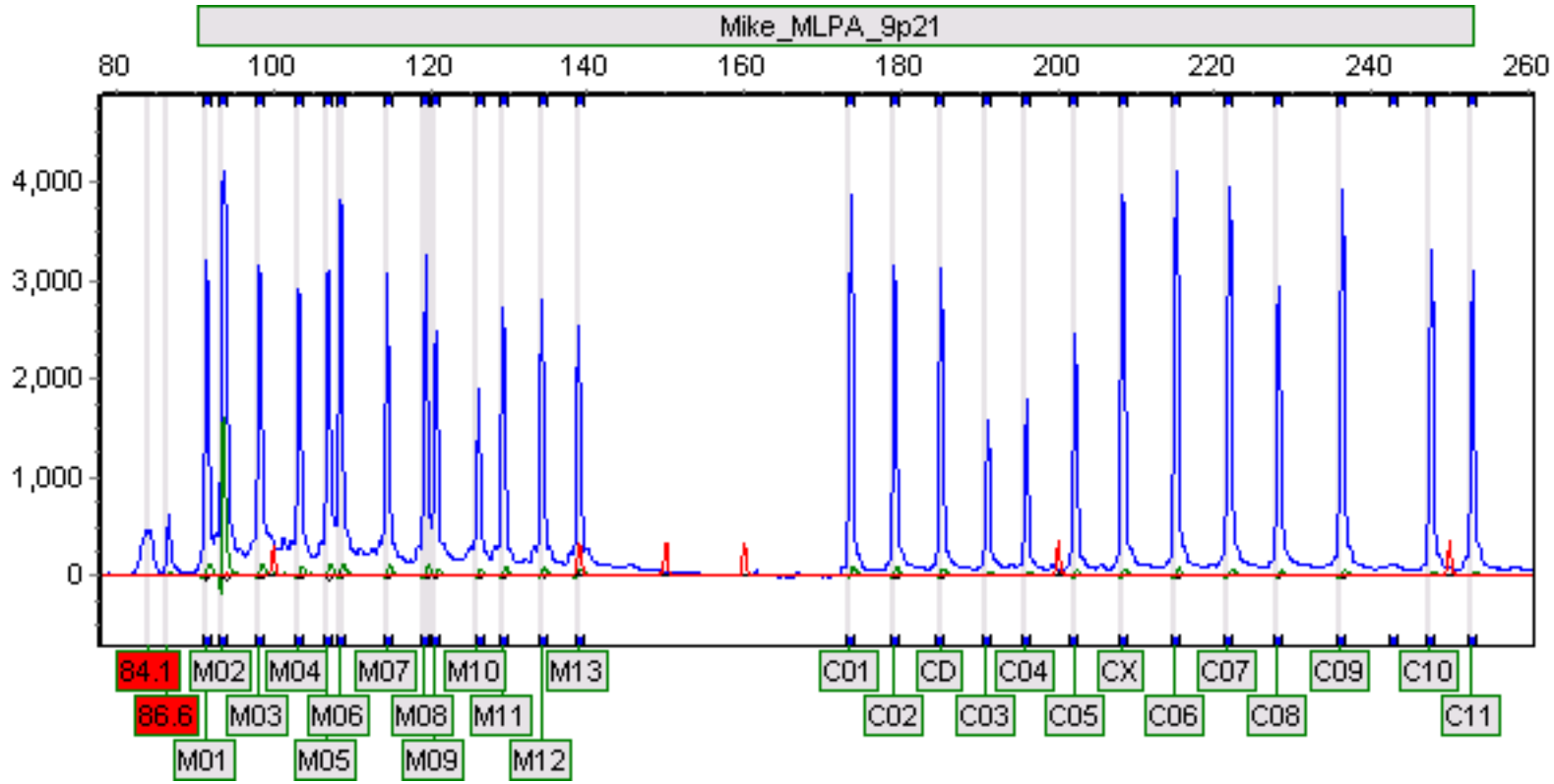
MLPA data for each sample were quality checked with respect to the Q-fragments, D-fragments, and dispersion of the control probes, and samples with poor quality results were excluded from analysis or repeated.

Table 2.2. Probes in the SALSA MLPA P200-A1 reference probemix.

Length (bases)	SALSA MLPA probe	Chromosomal position
64-70-76-82	Q-fragments: DNA quantity; only visible with <100ng sample DNA	NA
172	Reference probe C01	7q31
178	Reference probe C02	14q22
184	D-fragment: low signal indicates incomplete denaturation.	14q32
190	Reference probe C03	20p13
196	Reference probe C04	13q12
202	Reference probe C05	20p12
208	Chromosome X probe CX	Xq26
214	Reference probe C06	10p13
220	Reference probe C07	12q24
226	Reference probe C08	4q25
232	Reference probe C09	18q11
238	Chromosome Y probe CY	Yq11
244	Reference probe C10	5p15
250	Reference probe C11	17p11

Figure 2.2. Example of MLPA spectra for the custom chromosome 9p21 probeset obtained using GeneMarker software.

Spectrum shows the custom probes (M01-M13) and the P200 control probes (C01-C11, CX, CY, and CD, as shown in Table 2.2). The X-axis represents amplicon size (in nucleotides) and the Y-axis represents peak height. The 82nt Q-fragments are small, indicating an adequate amount of DNA template, and the D-fragment peak (CD) is large, indicating adequate sample denaturation.



2.2.9 Statistical analyses

2.2.9.1 General analyses

Normality testing and data transformation

Quantitative data were tested for normality using the Anderson-Darling test using Minitab v15. Normally distributed data are presented as mean and standard deviation, with data that deviate from a normal distribution presented as median, range, and lower/upper quartiles. Quantitative traits that deviated from a normal distribution were transformed (most often using a log transformation) to approximately normalise the distribution prior to correcting for covariates or performing association analysis.

Adjustment for covariates and correlation analyses

Effects of covariates for quantitative phenotypes were adjusted for using linear models in Minitab v15 or R software. Covariate-adjusted data were used for all association analyses unless otherwise stated. All correlations were performed using Pearson's correlation test in SigmaPlot v11.0 unless otherwise stated. Spearman's rank correlation test was used when outliers or deviation from normality were likely to influence the analysis.

Testing genotyping for Mendelian inheritance, D-statistics, and correspondence to Hardy-Weinberg proportions

Family-based genotyping data was checked for Mendelian inheritance errors and correspondence to HWE proportions using PEDSTATS[295]. Data from families with Mendelian errors were reviewed and re-genotyped if necessary, and where Mendelian errors persisted, families were excluded from analyses. HWE checks were performed in the total population and in the unrelated founder members only to account for deviations that may be due to family structure. Pedigree and data files were prepared as described on the PEDSTATS website. Genotyping data from unrelated individuals was checked for correspondence to HWE proportions using standard formulae in Excel or Haploview[296-298].

The extent of deviation from HWE proportions was estimated for each SNP by calculation of the deviation statistic (D-statistic) in Excel according to the

formula $D = H - 2p(1-p)$, where H is the observed heterozygosity and p the minor allele frequency. Using this formula, negative D-values correspond to an excess of homozygosity, and positive values to an excess of heterozygosity.

Significance thresholds for hypothesis testing and correction for multiple-testing

Statistical significance was defined using an uncorrected significance threshold of 0.05 unless otherwise stated. Correction for multiple testing was performed using a Bonferroni correction[299], in which the corrected significance threshold (α') for n tests is calculated as, $\alpha' = \alpha/n$.

Linkage disequilibrium, tag SNPs, and LD plots

LD plots and calculation of LD between SNPs were performed using Haploview v4.0[297]. The Tagger function in Haploview was used to select a minimal set of SNPs that define a haplotype, parameters for which are specified in the relevant sections. LD for pairs of SNPs are presented as r^2 and D' values, following standard convention[297, 299].

2.2.9.2 Genotype-phenotype association testing

Association testing using MERLIN

Family-based association analysis for quantitative traits was performed using MERLIN v1.1.2 software[300], which uses sparse inheritance trees for pedigree analysis. This performs rapid haplotyping, genotype error detection and affected pair linkage analyses and can handle more markers than other pedigree analysis packages. The association test implemented in MERLIN includes an integrated genotype inference feature, which can improve power when some genotypes are missing. Pedigree, data and map files were prepared as described on the MERLIN website[301].

Association testing using UNPHASED

Case-control association analysis for discrete traits was performed using UNPHASED v3.10[302, 303]. This software implements maximum-likelihood inference on genotype effects while allowing for missing data such as missing genotypes. Association testing for binary traits is performed using logistic regression, which can be done

using allele or genotype models. Pedigree, data and map files were prepared as for MERLIN.

Estimation of maximum SNP contribution to overall phenotypic variance

For association analyses of quantitative traits, an approximate estimate of the maximum plausible genetic effect (G) of a given SNP on a particular trait based on the upper 95% confidence interval was calculated using the formula: $G = (A + (2*SE_A))^2 * ((SD_{genotype})^2/(SD_{adj\ trait})^2)$, where A=absolute regression coefficient, SE=standard error, SD=standard deviation.

2.2.9.3 AEI analysis

AERs were analysed using an extension of the approach published previously[254]. This has now been published[304], but the principles are summarised below. After normalisation to an equimolar control, AERs were log-transformed. Outliers from each individual sample's technical replicates were excluded using Grubb's test. In order to assess the effect of a potentially *cis*-acting polymorphism on allelic expression ascertained at a transcribed polymorphism, two things need to be estimated: first, the phase between the alleles at both polymorphic sites; and second, the effect of the potential *cis*-acting marker on expression. The simplest method is to first estimate the phase and then, given that phase, assess the effect of the polymorphism. This approach was used in the chromosome 9p21 analysis, where genotypes for each population were phased using the hap procedure from the R-package gap (as deposited in the CRAN archive)[305]. Once the phase has been estimated it is possible to assess how likely the observed allelic expression ratio is given the phase relationships between transcribed and expressed polymorphisms under different assumptions on the effect. The effect size that best explains the observed ratios is then selected, based on a likelihood maximisation procedure. To establish significance, the effect estimate is compared with the likelihood obtained assuming no effect of the putative *cis*-acting polymorphism. These likelihoods can be compared using a likelihood ratio test, the result of which is summarised as the significance P-value. These analyses were performed using R and MatLab software. An alternative approach to AEI analysis is to determine phase and effect simultaneously, as described by Teare *et al*[254]. This approach was used in the *STK39*

analysis in Chapter 7 using the statistical package R, but was computationally too challenging given the number of SNPs genotyped (and corresponding number of possible haplotypes) at the 9p21 locus. For all analyses the effect size is defined as the logarithm of the ratio of (expression from the chromosome carrying allele 2 at the *cis*-acting locus / expression from the chromosome carrying allele 1 at the *cis*-acting locus), i.e. the log of the change of expression that allele 2 produces compared to allele 1. By convention this was performed defining allele 2 as the rare allele.

2.2.9.4 Quantitative real-time PCR analysis

The association between total expression, as measured by real time PCR, and each of the SNPs was assessed using linear regression of the log transformed normalised expression values on the genotype assuming no dominance or interactions between the effects of different SNPs. Analyses were performed in R and SigmaPlot v11.0. The effect size is defined as the slope of the linear regression anlaysis.

Chapter 3

Preliminary methodological studies for allelic expression analysis

3 Preliminary methodological studies for allelic expression analysis

3.1 Introduction

With the recent identification of multiple non-transcribed SNPs associated with disease phenotypes in GWA studies, interest has increased in techniques for investigating *cis*-acting effects on expression. Previous AEI analyses have only considered the effects measured at individual transcribed SNPs, but simulations suggest that combining information from multiple transcribed SNPs in the same gene may improve the power of this approach for mapping *cis*-acting effects. The chromosome 9p21 region has been recently associated with a range of diseases by GWA studies[21, 126, 130, 164, 167, 168], and nearby genes in this region including *CDKN2A*, *CDKN2B*, *ANRIL*, and *MTAP* each have multiple exonic SNPs that could be used to investigate such an approach[42].

Published studies have varied considerably in the strategies used to normalise AER values. Normalisation to an equimolar control is required to correct for any assay bias that could potentially give artefactual results, for example, with the Sequenom platform higher mass products tend to be detected less efficiently than lower mass products during MALDI-TOF mass spectrometry, and with the hME technology for which the technique was originally described[272], alleles with multi-base extension tend to have smaller peaks than alleles with single-base extension. Appropriate normalisation is important since such systematic assay biases could create artefactual associations for variants in high LD with the transcribed variant used to assess AEI. The most common strategy for normalising AEI data has been to correct for the allelic ratio obtained in genomic DNA from heterozygous individuals, based on the assumption that the alleles are present in a 1:1 ratio in genomic DNA. However, reported approaches have included normalisation to the mean genomic ratio of all heterozygous individuals[306, 307], normalisation to the mean of a subset of individuals[260], normalisation to a single reference sample[308, 309], and normalisation of each individual's cDNA to the genomic DNA from the same individual[310]. An alternative method is to normalise using the AER obtained from an experimentally

prepared equimolar standard. This is done by mixing DNA in a 1:1 ratio from two individuals who are homozygous for alternative alleles at the transcribed polymorphism and using this mix as the template for AER assessment. The extent to which different normalisation strategies influence AEI analysis has not been studied. The most commonly used method has been to normalise to pooled genomic DNA ratios, which should reduce the variability of the normalisation factor, but could be problematic in the presence of CNVs or variants in the amplicon sequence that influence the cDNA AER of individual samples.

At the outset of this study, application notes and published literature for AEI assessment using Sequenom were only available for hME analysis and not the newer iPLEX technology. hME analysis involves a multi-base extension process for detection of the higher mass allele, which tends to reduce the area of the higher mass peak relative to that of the lower mass allele which involves only single base extension for detection. This occurs because some copies of the higher mass allele only extend by a single base during the extension reaction, as evidenced by the presence of a 'pausing peak' in the spectrum (an example of which is shown in Figure 3.3 on page 100). This peak may interfere with measurement of the lower mass allele peak area, and results in reduced peak area of the higher mass allele and skewing of AER. However, the newer iPLEX technology, which involves single base extension for each allele, would be expected to reduce such biases and may offer improved accuracy and reduce reliance on accurate normalisation. A further advantage of the iPLEX technology is that reactions can be multiplexed more easily, which is advantageous where samples are limited. However, the use of iPLEX and multiplexing for AEI analysis needs to be experimentally validated.

Experimental factors such as RNA storage and degradation may influence analyses involving total gene expression, since total expression levels are normalised to expression levels of different reference genes, but it has been shown that different mRNAs may degrade at different rates[311]. AEI analysis would be expected be more resistant to such influences since alleles which differ in sequence by only a single nucleotide are compared within the same sample, but the extent to which such factors influence AEI assessment needs to be confirmed experimentally. Assumptions involved in the AEI analysis such as linearity, normality, and constant variances of

AEI measurements also require experimental validation. Furthermore, whether RNA/cDNA amplification alters AEI measurements is not known.

3.2 Aims

3.2.1 AEI studies using *MLH1*

Preliminary AEI studies were performed using SNP rs1799977 in the *MLH1* gene, for which pilot studies of techniques for AEI assessment had been previously performed by other members of the group. The specific aims of these experiments were:

- To determine the reproducibility of the reverse transcription step.
- To investigate whether AEI could be reliably ascertained using stored RNA samples from the NE Caucasian cohort.
- To compare AEI assessed using hME and iPLEX methodologies.
- To determine whether amplified cDNA could be used for AEI analysis.

3.2.2 AEI studies in the chromosome 9p21 region

Methodological studies were also carried out for AEI assessment of *CDKN2A*, *CDKN2B*, *ANRIL*, and *MTAP* expression at the chromosome 9p21 locus. The specific aims of these experiments were:

- To determine whether expression of these genes could be reliably detected in peripheral blood and whether they displayed sufficient inter-individual variability in AEI to allow mapping of *cis*-acting effects.
- To determine the amount of cDNA needed for AEI analyses of these genes.
- To assess whether AEI assays could be reliably multiplexed using the iPLEX methodology.
- To demonstrate linearity, normality, and constant variance of AEI estimates.
- To compare the influence of different normalisation methods on mapping of *cis*-acting effects using AEI.
- To investigate novel methodology using multiple transcribed SNPs per gene.

3.3 Materials and methods

3.3.1 Participants and samples

Expression studies at the chromosome 9p21 locus included two populations of healthy adult volunteers: the 310 individuals of the SA cohort and the 177 individuals from the NE Caucasian cohort. Preliminary feasibility and optimisation work involving *MLH1*expression were performed using samples from affected cases and unaffected controls of the CAPP study. Details of these populations and the nucleic acid preparation were presented in Chapter 2.

3.3.2 Selection of transcribed SNPs for allelic expression analysis

Using the NCBI Entrez Gene database[196], transcribed transversion SNPs with expected heterozygosity >0.2 in Caucasian populations were selected as suitable candidates for assessment of allelic expression. Variants lacking population frequency data were excluded. Insertion/deletion polymorphisms were also excluded since these would alter the size of the PCR product produced for each allele, which could bias the AEI estimate. The NCBI[196] and Ensembl[43] databases were also used to establish the location of transcribed SNPs with respect to reported transcript variants of each gene.

Transcribed polymorphisms in *ANRIL*, which was not annotated in the databases at the time of the design, were identified by comparing the reported mRNA sequence[195] with NCBI dbSNP using the BLAST tool. The expected heterozygosity of *ANRIL* exonic SNPs identified in this way was then checked in dbSNP.

Transcribed SNPs selected using these criteria were: rs3088440 and rs11515 in exon 3 of *CDKN2A*; rs3217992 and rs1063192 in exon 2 of *CDKN2B*; rs10965215 and rs564398 in exon 2 of *ANRIL*. The two *CDKN2A* SNPs are also present in *ARF*, allowing the assessment of *cis*-acting influences on both of these transcripts. Another SNP rs10738605 in exon 6 of *ANRIL* also satisfied these criteria but was excluded from initial AEI experiments because of extensive skewing of the allelic ratio in genomic DNA.

3.3.3 Selection of mapping SNPs

3.3.3.1 SNPs associated with disease

SNPs previously reported to be associated with disease phenotypes were selected for mapping effects on expression[124, 125, 160, 169-173, 175-182, 312-314]. GWA studies for CAD showed association for multiple SNPs which were in strong LD in the chromosome 9p21 region. To reduce redundancy of genotyping but ensure inclusion of the most important SNPs, the 'lead' SNPs showing the strongest association with disease, subsequent refinement SNPs, and SNPs reported as tagging the risk haplotype were selected from GWA studies. All SNPs reported to have significant associations with disease phenotypes in candidate gene studies (as previously summarised in Table 1.4 on page 33) were included.

3.3.3.2 SNPs altering transcription factor binding sites within regulatory regions

SNPs within previously reported regulatory elements such as the *CDKN2A, ARF* and *CDKN2B* promoters[186-189] or a putative *ANRIL* promoter region (which was arbitrarily defined as 1kb up and downstream of the transcription start site for the purposes of this study) were also selected. SNPs in these regions were included if they were reported more than once in NCBI dbSNP, had expected heterozygosity >5%, and the alternative SNP alleles were predicted to alter human transcription factor binding sites using PROMO v.3.0.2 software[246, 247].

3.3.3.3 Tag SNPs to capture common variation in the region

Additional tag SNPs required to capture common variation in the core region of interest (Chr9:21958155-22115505) based on HapMap CEU data were also selected using HaploView 4.0 Tagger software using the following parameters: minimum minor allele frequency 0.01, pairwise tagging, r^2 threshold >0.8. Transcribed or functional SNPs already selected for genotyping were 'force included' as tag SNPs in Tagger to reduce redundancy of genotyping. Tagger output SNPs were manually checked for potential problems with Sequenom genotyping (e.g. rs6475608 had multiple deletions/repeats which would be problematic for Sequenom SNP analysis). Tagging was then repeated with SNPs likely to be problematic 'force excluded' from

selection, to obtain a final panel of typable SNPs which adequately tagged the region. The SNPs selected for genotyping and reasons for including them are shown in the genotyping results summary table in the next chapter (Table 4.2 on page 130).

3.3.4 Genotyping

Multiplex SNP genotyping was performed using Sequenom methodology as described in Chapter 2. The 56 SNPs were genotyped in five separate reactions (W1-W5).

3.3.5 AEI assays

PCR primers for the selected transcribed SNPs were designed manually and using Primer3 (v.0.4.0) software[289]. For AEI analysis PCR primers were designed to anneal across exon boundaries, thus being specific for cDNA and not binding to genomic DNA. Primers for genomic DNA normalisation were designed to produce products as similar as possible in size and sequence to those produced by the cDNA specific primers, to minimise differences in reaction kinetics. *CDKN2A* primers span exons 3-4 and include both transcribed SNPs (rs3088440 and rs11515) in the same amplicon. *ANRIL* primers span exons 1-2 and include both transcribed SNPs (rs10965215 and rs564398) in the same amplicon. For *CDKN2B*, the distance of transcribed SNPs from the exon boundary meant that amplicons would be more than 1kb in size, and separate primer pairs for transcribed SNPs rs1063192 and rs3217992 were therefore designed entirely within exon 2. These *CDKN2B* primers were therefore not cDNA specific, and analysis was performed using cDNA from DNase treated RNA. Measurements were performed in four replicates using 50ng of cDNA template.

3.3.6 Genomic DNA normalisation assays

Genomic DNA normalisation reactions for *CDKN2B* used the same PCR primers as used for cDNA, but for *CDKN2A* and *ANRIL* (where primers were cDNA-specific) separate assays designed to be as close as possible in size and location to the cDNA primers were used. Genomic normalisation samples were checked both in uniplex reactions and multiplex combinations when these were used for the AEI analysis.

The appropriateness of genomic normalisation ratios and linearity of the AER response were checked by mixing PCR products in varying ratios from individuals

homozygous for the minor and major alleles at each SNP, and using these as template for the allelic expression assays. To do this, PCR was performed from cDNA samples homozygous for each transcribed SNP allele, using standard uniplex AEI protocols but with only 40 cycles. The concentration of the PCR product from each homozygous sample was quantified using PicoGreen, performed in three replicates. For each pair of alleles at each SNP, the product with the highest concentration was diluted with 1xTE to obtain the same concentration as the less concentrated product. Further cycles of PicoGreen measurement and dilution were repeated until the measured concentration of each sample was the same. Volumes of product from each allele pair were then mixed in known ratios (8:1, 4:1, 1:1, 1:4, 1:8), and for each mixture serial dilutions were performed with 1xTE until the molarity with respect to the DNA amplicons roughly equated to the molarity of genomic DNA in a 20ng/μL solution. The resulting allele mixes were then used as template for PCR and AEI analysis following the standard protocol.

3.3.7 Statistical analyses

Allelic expression analysis was performed using the pre-phasing methodology described in Chapter 2. A novel approach of combining allelic ratios from the two transcribed markers in each gene was used to increase the number of informative heterozygotes. This included AERs measured at both transcribed SNPs in the analysis of each gene. For individuals who were heterozygous for both transcribed SNPs, information from both was included allowing for different variances at each transcribed marker. The best fitting parameters were determined using likelihood maximisation algorithms.

3.4 Results

3.4.1 Preliminary feasibility and optimisation work using *MLH1*

3.4.1.1 Reproducibility of reverse transcription and effect of sample storage

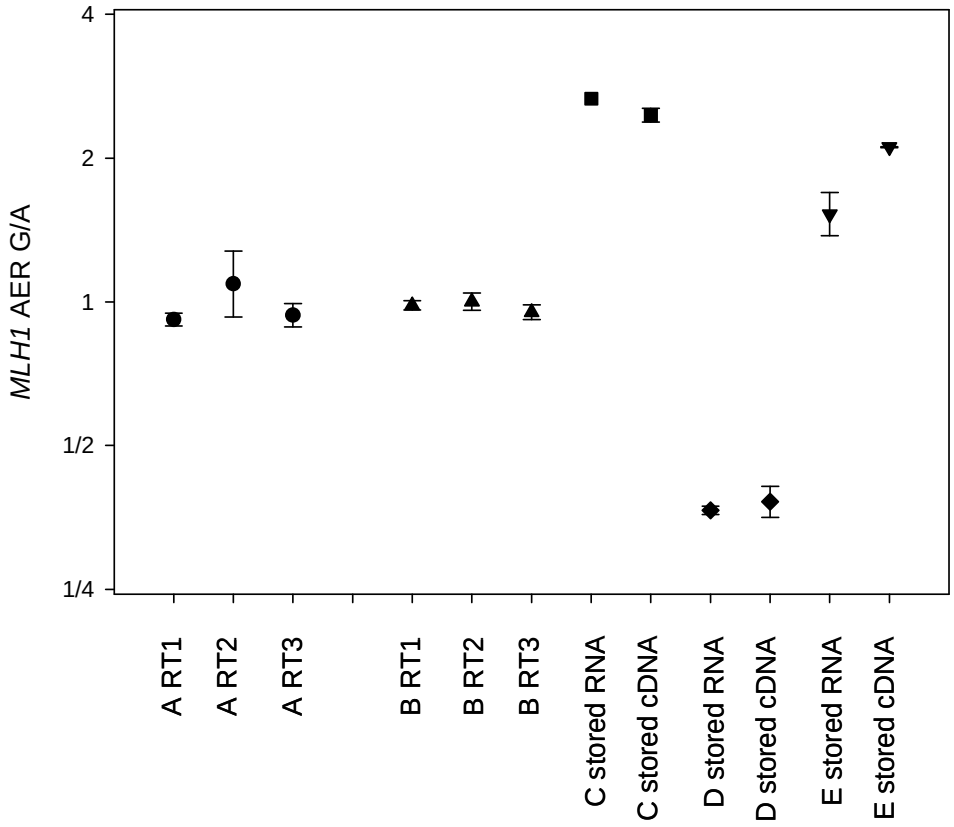
To test the effect of reverse transcription (RT) reactions on AER, the results of three separate RT reactions performed simultaneously from the same RNA solution were compared for two normal individuals. AER for each sample was assessed using four

replicates. As shown in Figure 3.1, there was no substantial difference in AER estimates from separate RT reactions in the same individual.

To examine whether allelic expression analyses might be influenced by degradation of stored RNA, paired samples were used to compare AER results between RT from stored RNA (samples stored at -80^{o}C for 1-2 years) and RT performed at the time of sample extraction and subsequently stored as cDNA (at -20^{o}C for 1-2 years). These analyses were performed using samples from three patients known to have imbalance in *MLH1* expression. As shown in Figure 3.1, the magnitude of AEI was similar for analyses of each patient, confirming that the RT step and sample storage do not appear to substantially affect AEI analysis.

Figure 3.1. Effect of reverse transcription reactions on AER.

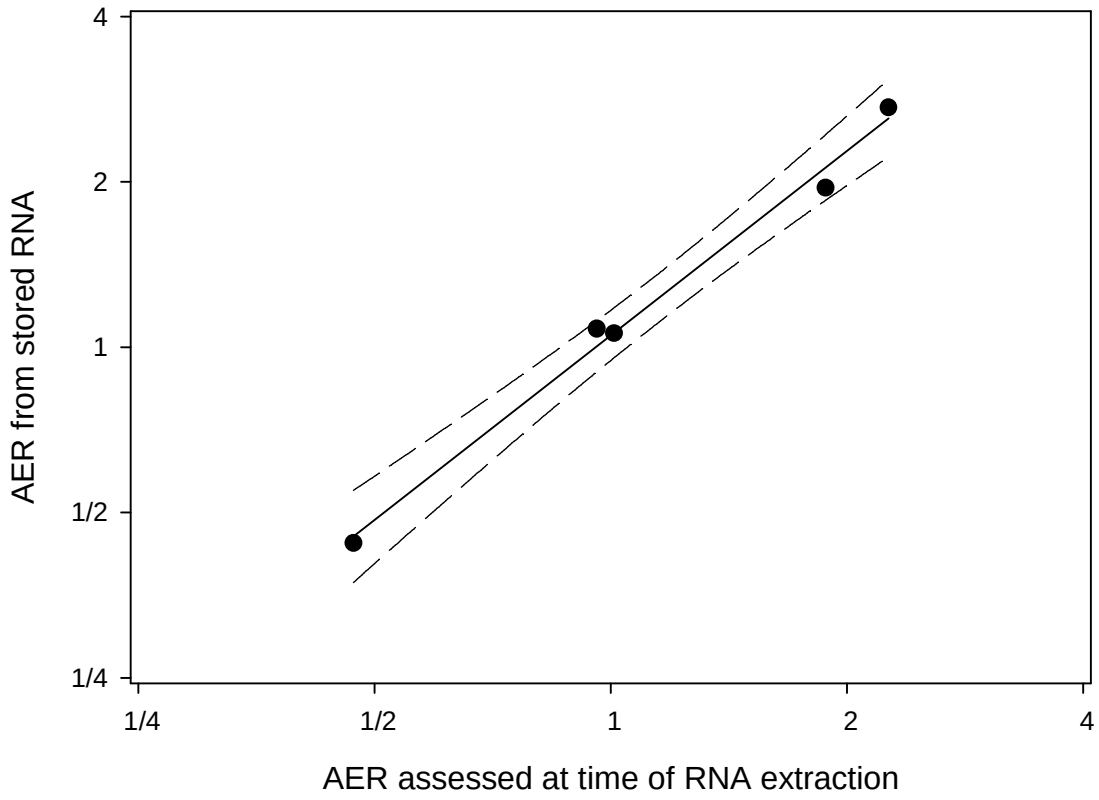
MLH1 AER is shown on the Y-axis. X-axis shows three simultaneous RTs (RT1-RT3) from two normal individuals (A and B) who were heterozygous at the transcribed SNP, and RTs from stored RNA and stored cDNA for three patients with known allelic imbalance (C, D, and E). Mean AER for each individual is represented by a different symbol (with standard error bars shown).



Although AER measurements from stored RNA and stored cDNA were comparable, it was important to confirm that these AER estimates were accurate and not both the result of artefacts due to sample storage (which might occur, for example, if one allele was preferentially degraded to the same extent in both the stored RNA and stored cDNA samples). AER estimates from stored RNA were therefore compared with AER values obtained from the same samples at the time of their original extraction, for five individuals with varying degrees of allelic imbalance. As shown in Figure 3.2, AER values obtained from stored RNA samples were similar to values obtained using 'fresh' RNA in experiments performed by a different operator using the same assay 1-2 years previously (r= 0.99, P=0.001).

Figure 3.2. Comparison of AER from stored RNA with AER obtained at the time of RNA extraction for *MLH1*.

Correlation of AERs obtained at the time of RNA extraction (X-axis) and from stored RNA (Y-axis). Regression line is shown as a solid line with dotted 95% CI.



It was not surprising that the linear RT step introduced little variability into AER estimates, and the data suggest that replicates at the RT stage are not required for AER assessment. Allelic expression analysis compares DNA sequences within individuals that differ by just a single base, and therefore should be relatively robust to the effects of RNA/cDNA degradation which would be expected to affect each allele equally and not alter the ratio between alleles. This is in contrast to total expression analyses where expression is compared between different genes within a sample (for normalisation) and between samples from different individuals, both of which may be subject to differential degradation. These data were particularly important since analyses in the NE Caucasian cohort would involve the use of stored RNA and stored cDNA, whereas analyses in the SA cohort would be performed using newly collected RNA.

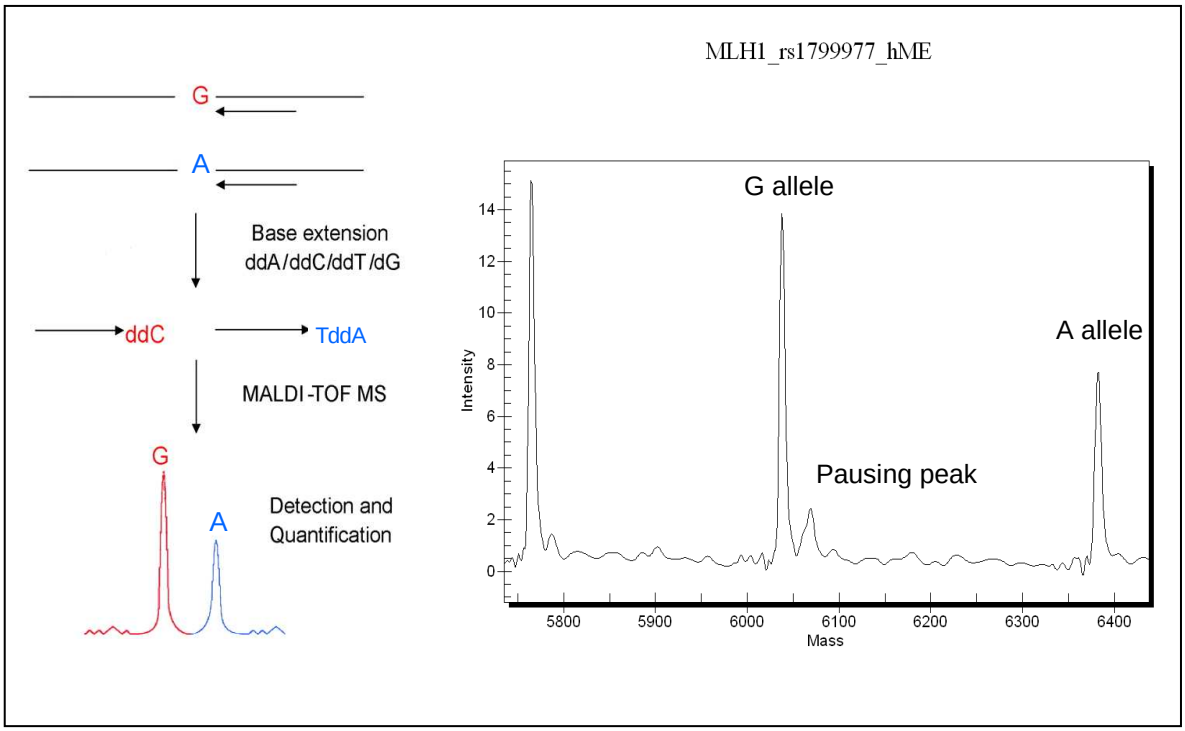
3.4.1.2 Comparison of AEI using hME and iPLEX

Five samples representing a mixture of balanced and imbalanced allelic expression were analysed using each methodology. PCR was performed using standard protocols in a 15µL reaction volume. For each reaction, 5µL of the PCR product was used for hME analysis, and 5µL for iPLEX analysis. This eliminated variation due to the PCR step and allowed the detection techniques to be directly compared. The process was repeated for both hME and iPLEX analysis using genomic DNA from the samples, to allow appropriate correction of each assay to be performed.

Examples of spectra from the same genomic DNA PCR product analysed with hME and iPLEX are shown in Figure 3.3, which clearly shows the pausing peak and difference in ratio between the peak areas for hME. The overall G/A ratio for genomic samples using hME was 1.56, compared to 0.85 for iPLEX on the same samples. This illustrates the importance of appropriate assay correction, as when the allelic expression ratios were corrected for genomic bias the correlation between AER for the two techniques was very high ($r=0.99$, $P<0.0005$), as shown in Figure 3.4. Since iPLEX offers advantages in terms of accuracy and multiplexing, all subsequent analyses were conducted using iPLEX.

Figure 3.3. Comparison of spectra from the same genomic PCR product analysed with hME and iPLEX showing difference in peak ratios.

hME



iPLEX

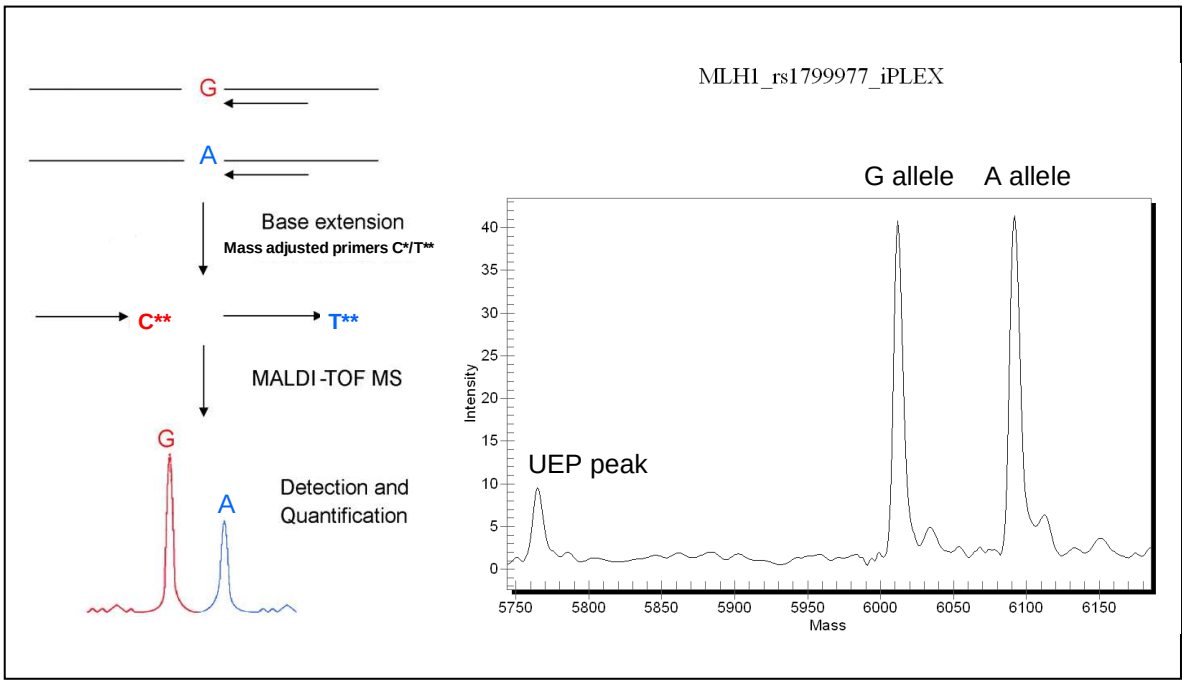
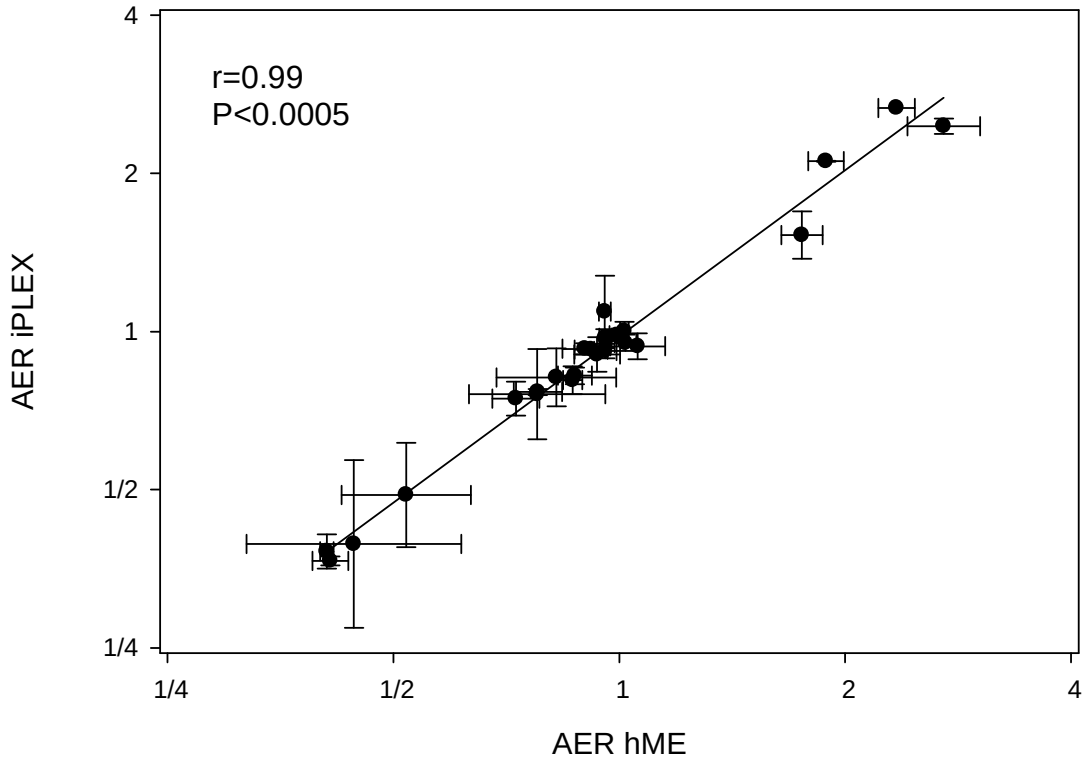


Figure 3.4. Comparison of AER using iPLEX and hME.

Each point represents an individual AER measurement, with standard errors shown. The linear regression line is shown as a solid line.



3.4.1.3 Effect of cDNA amplification on AER

Amplification of RNA/cDNA was investigated as a method of potentially increasing the amount of cDNA available for analysis. This would be advantageous both for the present samples extracted from blood, and for future work using different tissues (such as atheromatous plaque) where working from small amounts of extracted tissue would be important.

Quantitect Whole Transcriptome Kit (Qiagen) provides an integrated system for reverse transcription of RNA using random and oligo-dT primers, ligation of cDNA, and whole transcriptome amplification in an isothermal reaction. This is reported to produce uniform amplification across sequences with negligible sequence bias from as little as 10ng of starting RNA[315]. The transcript profiles of five genes were reported to be similar using amplified and unamplified cDNA according to the

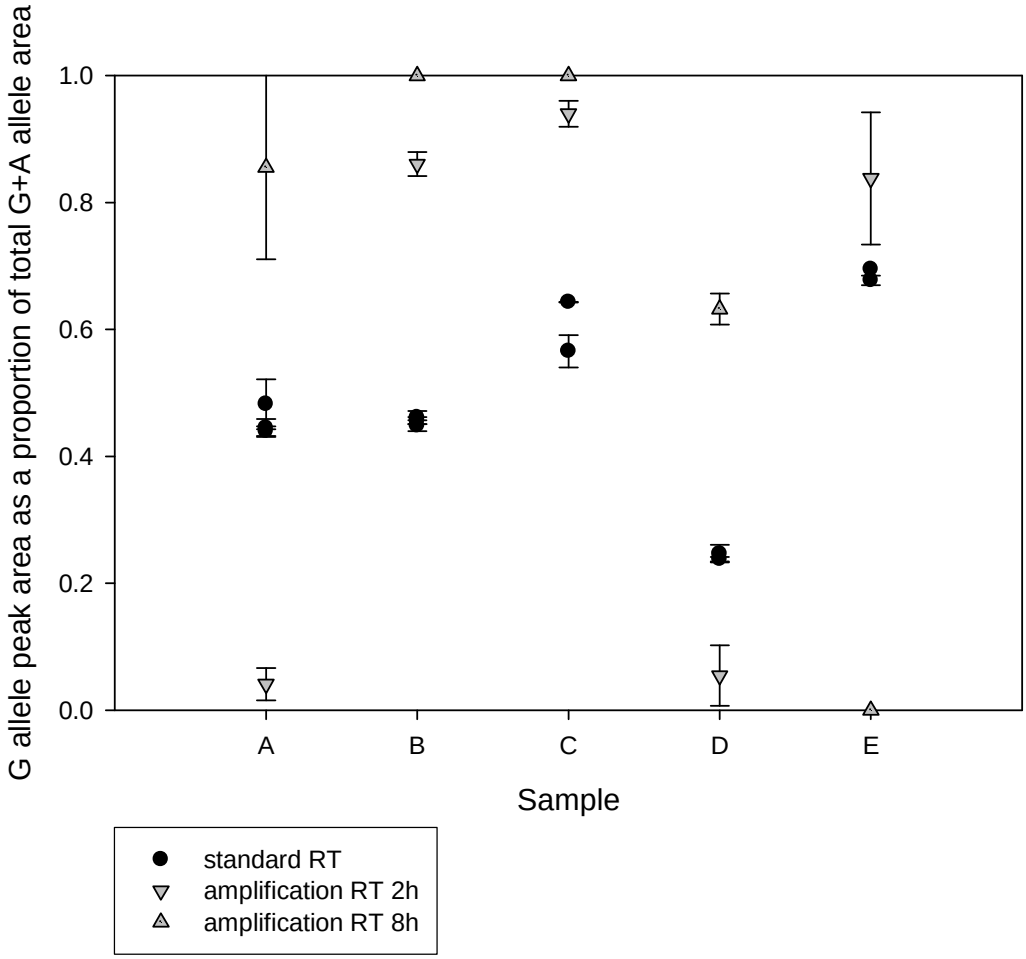
promotional information from Qiagen, but examination of their data did reveal differences in the absolute expression levels between amplified and unamplified expression profiles. Because this is an amplification process, even a small percentage difference in amplification efficiency between the two alleles of a SNP could significantly skew the AEI ratio and the effect of whole transcriptome amplification on AEI was therefore tested experimentally.

The effect of amplification on AER was tested using five samples which had a range of allelic imbalance in *MLH1*. For each RNA sample, RT was performed twice using the standard RT protocol, once with Quantitect 2 hour amplification, and once with Quantitect 8 hour amplification. Amplification RT was carried out using 20-40ng of starting RNA, quantified using optical densitometry. This amount of starting RNA was chosen because Quantitect amplification was reported to be reliable for amounts of starting RNA between 10-300ng, and 10-50ng was the average yield of total RNA that had been previously achieved in our group from laser micro-dissection of atheromatous plaque macrophages, which were potentially being considered for future AEI investigations. AEI was assessed using the standard four replicates from each RT reaction. The mean proportional peak area of the G allele (i.e. area of G allele peak divided by the combined area of the G and A allele peaks) for each reaction is shown in Figure 3.5. This showed that the amplification process substantially distorted the relative allelic peak areas, with one amplified allele tending to become greatly over-represented and the other under-represented in each amplification reaction. In some cases this was so extreme that only a single allele was detectable in the AEI analysis (which is the reason that the proportional area of the G allele rather than the AEI ratio is represented in Figure 3.5, since ratios with a denominator of zero cannot be shown). Interestingly, it was not always the same allele that was over-represented following amplification.

Based on these data, the AER is significantly altered by amplification using the Quantitect kit from 10-40ng of starting RNA. Unamplified RNA/cDNA was therefore used for all subsequent experiments.

Figure 3.5. Effect of Quantitect cDNA amplification on relative peak area of *MLH1* alleles.

The proportional area of the G allele is shown on the Y-axis for five different samples (A-E) shown on the X-axis, for standard RT (black circles), 2h amplification RT (inverted grey triangle ∇), and 8h amplification RT (grey triangle Δ). Points represent mean values and standard errors.



3.4.2 Methodological considerations and optimisation for chromosome 9 assays

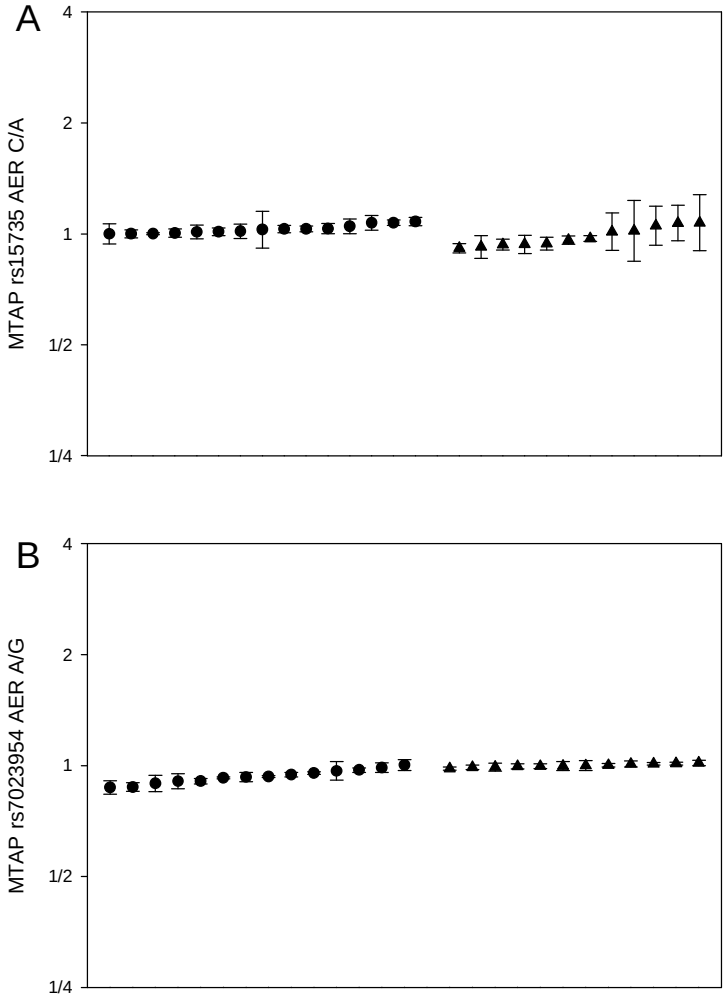
3.4.2.1 Confirmation of gene expression in blood and presence of inter-individual variation in AER

The first step was to confirm that the genes of interest were expressed in peripheral blood. Using cDNA specific primers, agarose gel electrophoresis revealed faint bands for *ANRIL*, *CDKN2A* and *CDKN2B* (after 45 cycles of PCR) suggesting relatively low level expression compared to *MTAP* where very strong bands were seen.

Preliminary investigation of AER in a small number of samples was performed to determine whether there was variation in AER between samples to allow expression mapping using AER. Substantial variation was seen in allelic expression of all markers in *CDKN2A*, *CDKN2B*, and *ANRIL*, which was subsequently confirmed in larger data sets (data shown in Chapter 4, pages 133 and 134). However, *MTAP* showed little inter-individual variation in AER at either transcribed marker, as shown in Figure 3.6, suggesting that it would not be likely to be informative for mapping the effects of common variation acting in *cis* in a study of this size. In light of this, no further analysis of *MTAP* was performed.

Figure 3.6. Lack of inter-individual variation in AER for *MTAP*.

Y-axes show AER for *MTAP* transcribed SNPs rs15735 (A) and rs7023954 (B). Individual cDNA samples are represented as circles, and genomic DNA samples as triangles. Points represent mean AER and standard error bars. AER in cDNA showed little variation between individuals and did not substantially differ from the allelic ratio in gDNA, suggesting little evidence of *cis*-acting effects on expression.

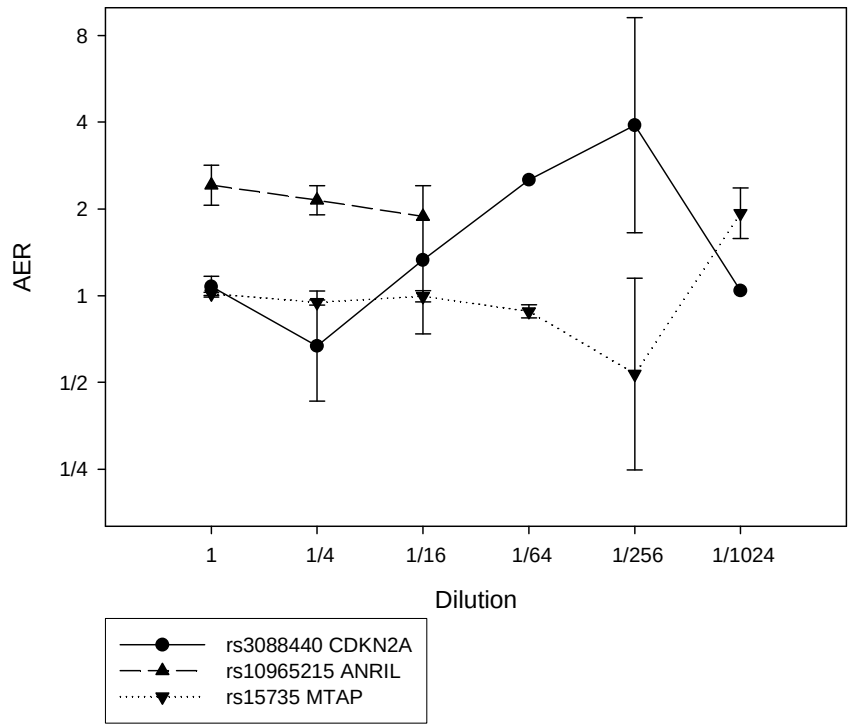


3.4.2.2 cDNA dilution series

A dilution series was repeated for the genes of interest on chromosome 9, since differing expression levels of the genes in blood could affect the amount of cDNA that was needed for AEI experiments. A dilution series was performed for each gene using 1µL of cDNA solution per reaction diluted with water from a starting concentration of 50ng/µL to determine whether lower amounts of cDNA could be used for AEI measurements. Results of these dilutions on AER estimates are shown in Figure 3.7. *MTAP*, which is most highly expressed in blood, had consistent AEI ratios with narrow confidence intervals down to 1/64 dilution. However, *CDKN2A* and *ANRIL* which are expressed at lower levels in blood were not reliably reproduced below 1/4 dilutions. For *ANRIL*, which had the lowest expression level in blood, PCR product was not reliably detected in all replicates below 1/16 dilution. Based on the dilution series, 1µL of cDNA solution (equivalent to 50ng of RNA) was used for AER assessment of all samples. For genomic DNA normalisation, 1µL of 50ng/µL gDNA solution was used.

Figure 3.7. AER at varying cDNA dilutions.

Estimate of AER (Y-axis) at different cDNA dilutions (X-axis) for three different assays (represented by different symbols as shown in the legend). Points show the mean AER estimate, with standard error bars shown. Absent points or error bars indicate the presence of failed replicates, which occurred in the more dilute samples.

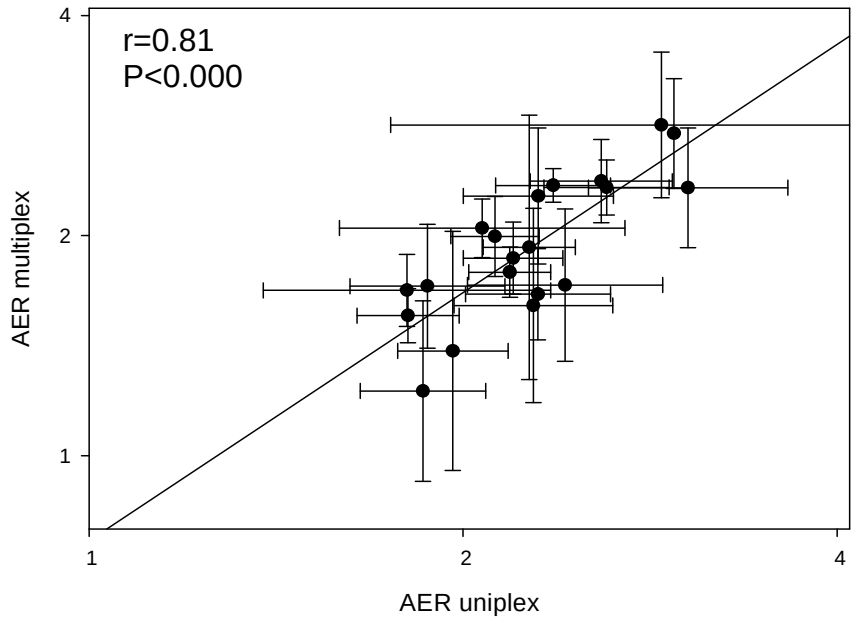


3.4.2.3 Effect of assay multiplexing on AER

The possibility of multiplexing AEI assays was investigated since this might allow the expression of multiple SNPs to be tested in the same sample, thereby reducing the amount of cDNA needed for analyses. In this experiment, PCR was performed for 20 samples using both the *CDKN2A* and *ANRIL* PCR primers in the same reaction. Each reaction was carried out in four replicates. However, during RealSNP assay design for genotyping purposes the expressed polymorphisms for *CDKN2A* and *ANRIL* had been designed into different assays (rs3088440 and rs10965215 in W1; rs11515 and rs564398 in W2) and assays for these SNPs were not compatible for combined analysis because some products were too similar in mass for separate MALDI-TOF peaks to be accurately distinguished. For these SNPs, the PCR reaction volume was increased to 15µL and the product was divided between two separate extension reactions. The correlation of multiplexed AER with uniplexed AER for *ANRIL* SNPs within individuals is shown in Figure 3.8. These values were strongly correlated ($r=0.81$, $P<0.0005$) and these multiplex combinations were therefore used in subsequent AEI experiments.

Figure 3.8. Comparison of AER for ANRIL SNPs assessed in uniplex and multiplex.

Points represent mean AER for each individual with standard error bars and regression line. Pearson correlation coefficient is shown.



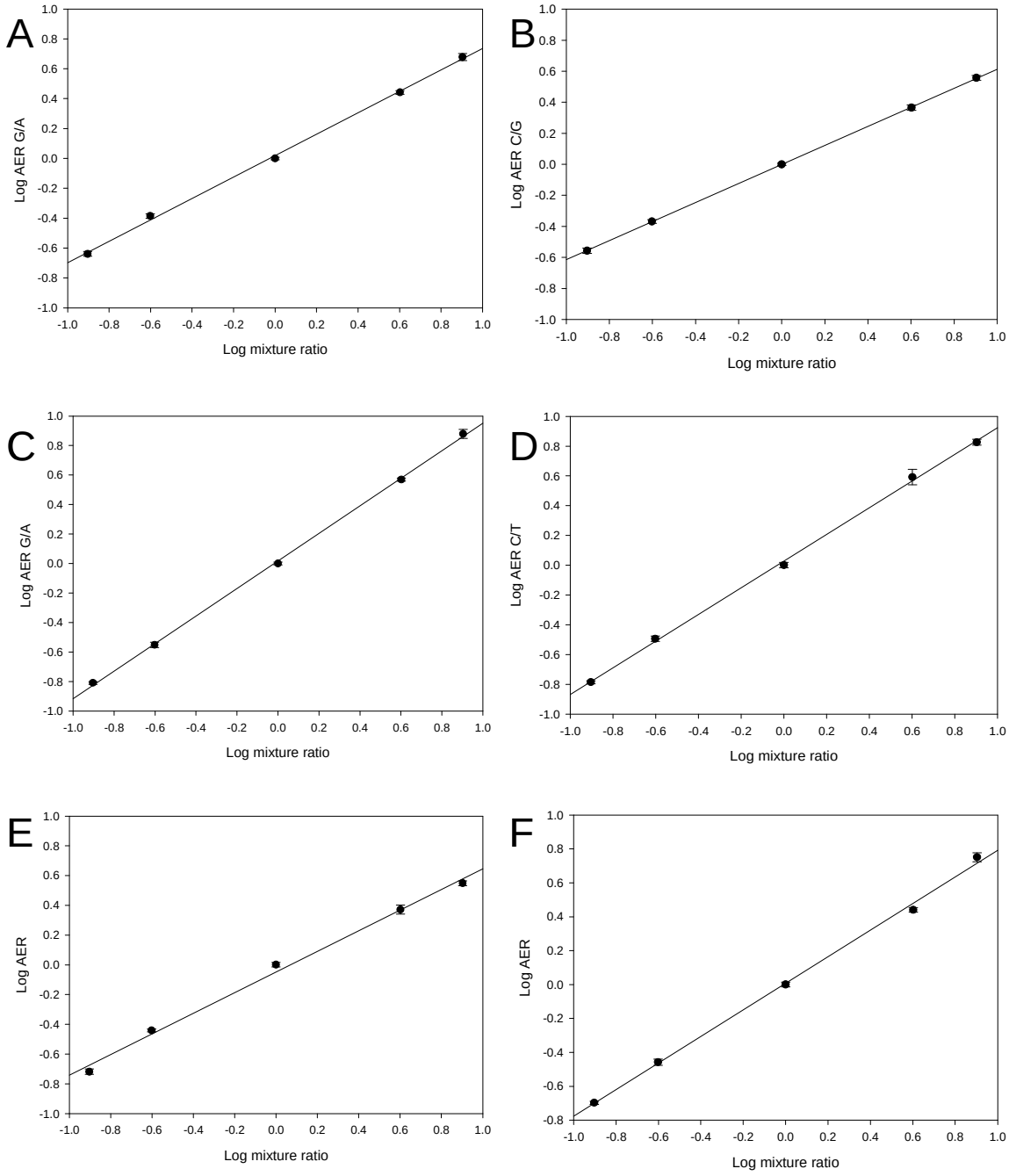
3.4.2.4 Linearity and normality of AER estimates

The appropriateness of genomic normalisation ratios and linearity of the AER response were checked by mixing PCR products from individuals homozygous for the minor and major alleles of each transcribed SNP in varying ratios (8:1, 4:1, 1:1, 1:4, 1:8) and using these as template for the allelic expression assays. As shown in Figure 3.9, these experiments confirmed that allelic expression showed a linear response, which is essential for the analysis procedure, and that normalisation ratios obtained using allelic expression assays on a 1:1 mixture of alleles for each SNP correspond to normalisation ratios obtained from genomic DNA (see also Table 3.1, page 113). The slope of the line was similar for the transcribed SNPs in a particular gene, such that AER values could be combined with no adjustment, as described in the section below. Despite the linearity of these measurements, the slope of the lines was not exactly 1 in all cases. This could represent a bias in the efficiency with which each allele was amplified, or a bias in the dilution process of this experiment. An experimental dilution bias would not influence the results of aeQTL mapping experiments. If there was an effect due to differences in amplification efficiency, this would not affect the significance estimates for aeQTL mapping, but could influence the estimated magnitude of the *cis*-acting effect. The measured effect would tend to underestimate the true effect in each case, and so, if anything the effect sizes reported in this study are conservative estimates of the magnitude of the true *cis*-acting effects.

To assess the distribution of AER measurements and check that the variance of AER assessment was similar in samples with different degrees of allelic imbalance, AER assessment was performed using 20 replicates in two samples with different degrees of AER at the *ANRIL* transcribed SNP rs10965215. Log-transformed AER values did not deviate significantly from a normal distribution ($P>0.05$ for the Anderson-Darling test) and there was no significant difference in the variance of these samples using the equal variance test ($P>0.05$).

Figure 3.9. Linear relationship between measured and expected allelic expression ratios for alleles mixed in known ratios.

Plots show the expected (X-axis) and measured (Y-axis) AER for alleles mixed in known ratios (8:1, 4:1, 1:1, 1:4, 1:8) at transcribed SNPs, along with the linear regression line. (A) *CDKN2A* rs3088440. (B) *CDKN2A* rs11515. (C) *CDKN2B* rs3217992. (D) *CDKN2B* rs1063192. (E) *ANRIL* rs10965215. (F) *ANRIL* rs564398.



3.4.3 Combining multiple transcribed markers for AEI analysis

AEI was assessed for each gene using two transcribed SNPs located in the same exon. AERs measured at the two transcribed SNPs in each gene were highly correlated (*CDKN2A* r=0.68 $P=1.7x10^{-3}$; *CDKN2B* r=0.80 $P=1.7x10^{-12}$; *ANRIL* r=0.90 $P=1.0x10^{-26}$; all genes combined r=0.96 $P=3x10^{-61}$), as shown in Figure 3.10. This was expected since the two transcribed SNPs selected to assess AER in each gene are located in the same exon and the same transcripts. Furthermore, aeQTL mapping results were highly correlated for the two transcribed SNPs in each gene, as shown in Figure 3.11, suggesting that each transcribed SNP identified similar *cis*-acting effects. This was important since in the absence of complete LD between the SNPs, individual transcribed SNPs could differentially represent the effects of different haplotypes which may lead to differences in the *cis*-acting effects detected. Such effects, if present, could potentially be lost in a combined analysis. The outlying points on Figure 3.11 were mostly effect estimates that were not statistically significant and therefore represented point estimates which had higher degrees of uncertainty rather than evidence of true differences in *cis*-acting effects detected at the different transcribed markers.

In view of the correlations between AER and mapping results at the two transcribed SNPs in each gene, the AERs from both transcribed markers were used for aeQTL analysis. As expected based on the above data, the results of aeQTL mapping using two transcribed SNPs per gene were highly correlated with the results obtained using information from single transcribed SNPs, as shown in Figure 3.12. This validates the methodology used for the transcribed SNP analysis.

All subsequent analyses use this methodology unless otherwise stated. The influence of this approach on the number of informative heterozygotes at which allelic expression could be assessed for each gene and the power to detect significant effects, is presented in the following chapter (Table 4.4, page 138).

Figure 3.10. Correlation between AER in individuals heterozygous for both transcribed markers in a gene.

The X- and Y-axes show the allelic expression ratio (AER) at the two transcribed SNPs in each gene. Each point represents an individual who is heterozygous for both transcribed SNPs in that gene. Circles represent *CDKN2A*, squares *CDKN2B*, and triangles *ANRIL*.

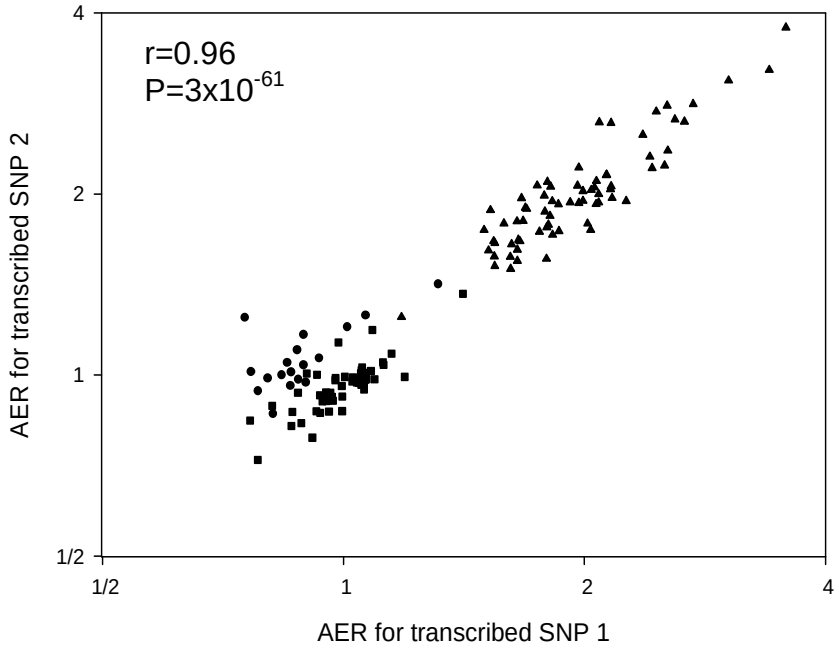


Figure 3.11. Comparison of aeQTL mapping results for the two transcribed SNPs in each gene for *CDKN2A*, *CDKN2B* and *ANRIL*.

Points represent the *cis*-acting effect for each of the 56 mapping SNPs in each of the three genes estimated at transcribed SNP one (X-axis) and transcribed SNP two (Y-axis) for each gene. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).

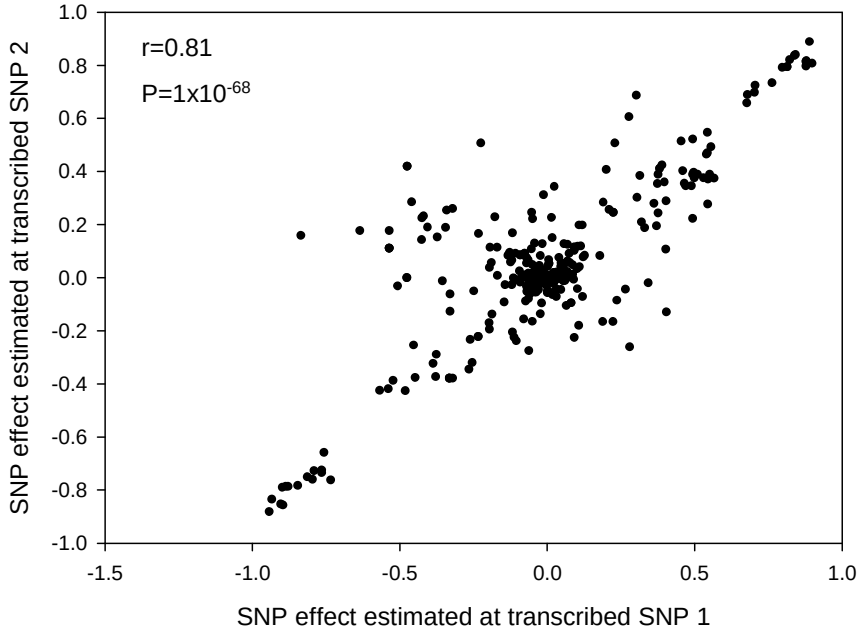
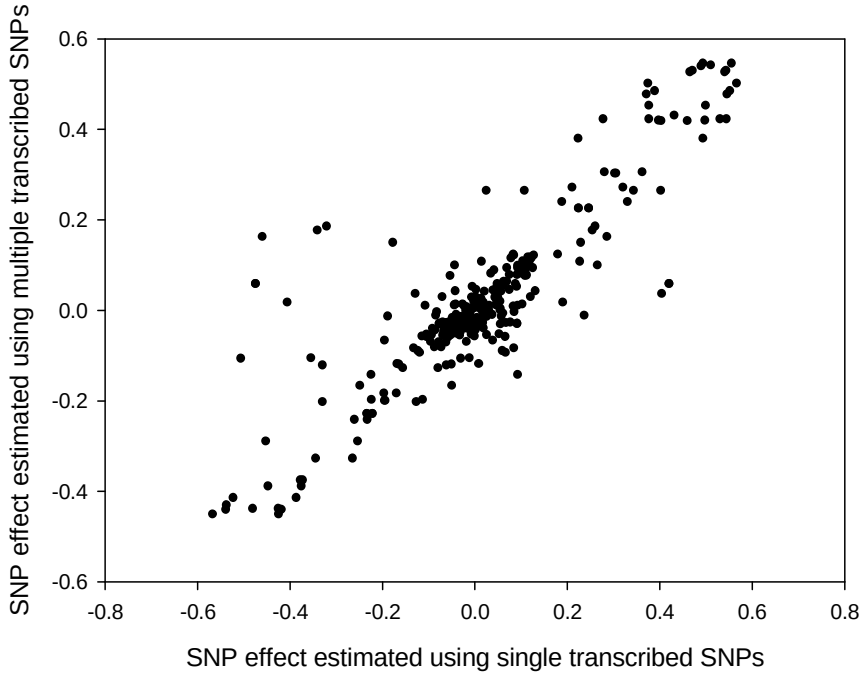


Figure 3.12. Comparison of aeQTL mapping results using single versus multiple transcribed SNPs per gene for *CDKN2A*, *CDKN2B* and *ANRIL*.

Points represent the *cis*-acting effect for each of the 56 mapping SNPs in each of the three genes estimated using a single transcribed SNP per gene (X-axis) and two transcribed SNPs per gene (Y-axis) in the SA population. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



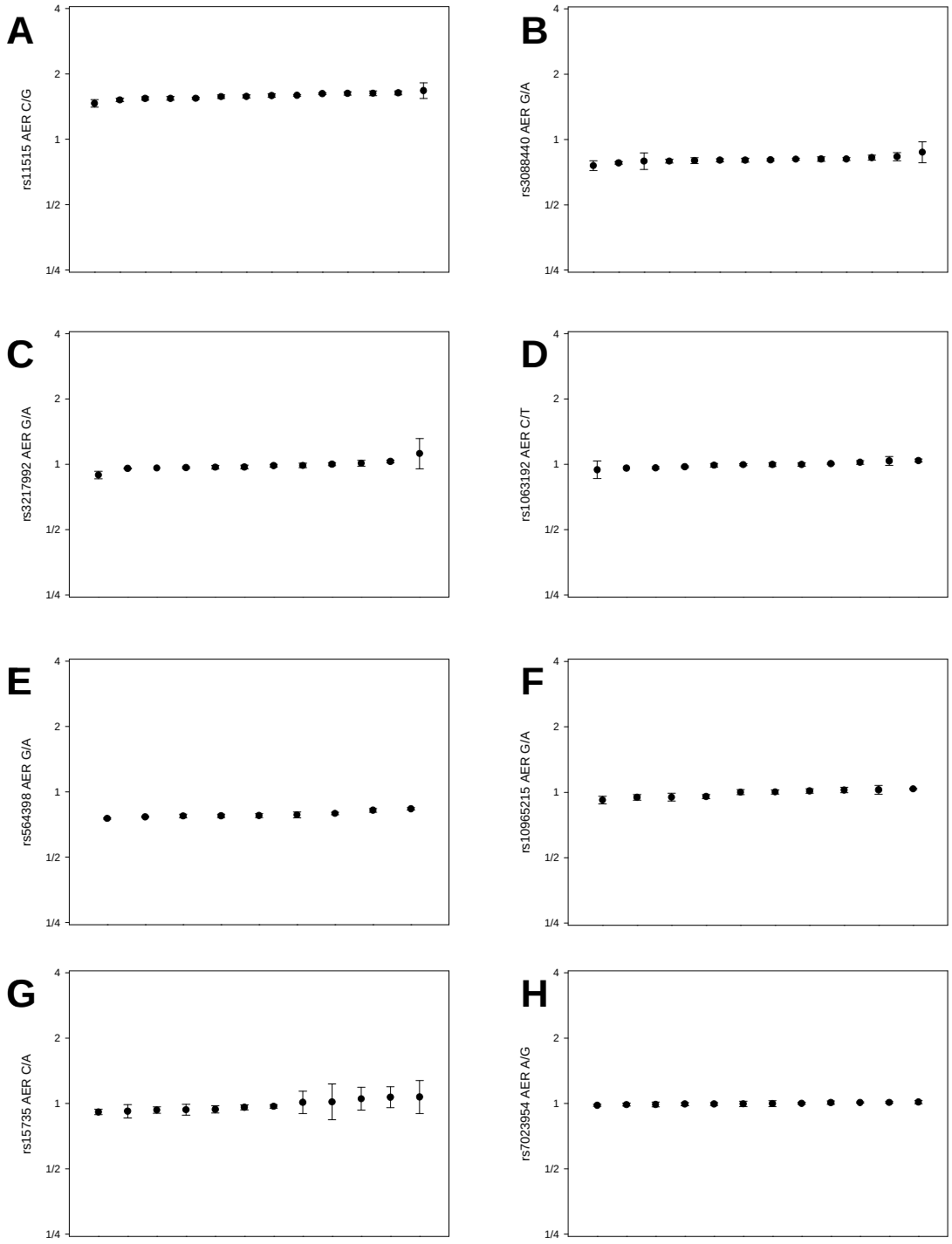
3.4.4 Investigation of normalisation methods

3.4.4.1 Normalisation factor quantification

Initially, a subset of Caucasian samples was used to estimate the gDNA normalisation factor for each transcribed SNP. As shown in Figure 3.13, the gDNA ratios for each assay were highly reproducible for each individual (as indicated by the small standard errors) and there was little inter-individual variability, particularly in comparison with the variability observed between individuals in cDNA ratios (see Chapter 4 pages 133 and 134 for comparison with cDNA ratios). However, for some assays the allelic ratios measured in gDNA ratios did deviate from a 1:1 ratio, confirming that allelic ratios in cDNA require correction for assay bias even using iPLEX single-base extension technology, which has less inherent assay bias than the previous hME technology.

Figure 3.13. Allelic ratios for genomic control DNA.

Y-axes show AER in the Caucasian gDNA normalisation samples for the following transcribed SNPs: (A) *CDKN2A* rs11515; (B) *CDKN2A* rs3088440; (C) *CDKN2B* rs3217992; (D) *CDKN2B* rs1063192; (E) *ANRIL* rs564398; (F) *ANRIL* rs10965215, (G) *MTAP* rs15735, (H) *MTAP* rs7023954. Each point represents an individual, with standard error bars shown.



For *CDKN2A* and *ANRIL* AEI assays, the cDNA primers were designed to be exon-spanning to make them cDNA-specific, meaning that different primers were required for the gDNA normalisation assays. Although these were selected to produce similar sized PCR products, the amplicon sequences partially differed from those in the cDNA assays. Since the normalisation process was primarily intended to correct for biases in the allele detection methodology (at the level of the extension reaction and MALDI-TOF), it was anticipated that sequences differences in the PCR amplicons would make little difference to the allelic ratios, especially since the AER compares ratios between two SNP alleles in the same sample (which differ by only 1bp and are likely to have similar PCR kinetics). To investigate this, the allelic ratios obtained in gDNA were compared with those obtained using experimental equimolar mixes (performed as previously described for the linearity assessment experiments), which are an exact sequence match for the cDNA assay. The normalisation ratios obtained using these two techniques were similar, as shown in Table 3.1.

Table 3.1. Allelic ratios in genomic DNA.

Gene	SNP	Allele 1 / allele 2	Method of assessment	Number of samples	Mean allelic ratio (A1/A2)	Standard error
ANRIL	rs10965215	G/A	gDNA 1	10	0.99	0.013
			gDNA 2	19	0.97	0.011
			Allelic mix	1	1.02	
	rs564398	G/A	gDNA 1	9	0.79	0.009
			gDNA 2	11	0.81	0.004
			Allelic mix	1	0.83	
CDKN2A	rs3088440	G/A	gDNA 1	14	0.81	0.008
			Allelic mix	1	0.82	
	rs11515	C/G	gDNA 1	14	1.58	0.015
			Allelic mix	1	1.36	
CDKN2B	rs1063192	C/T	gDNA 1	12	0.96	0.008
	rs3217992	G/A	gDNA 1	12	0.97	0.018
MTAP	rs15735	C/A	gDNA 1	12	0.97	0.020
	rs7023954	A/G	gDNA 1	12	0.99	0.003

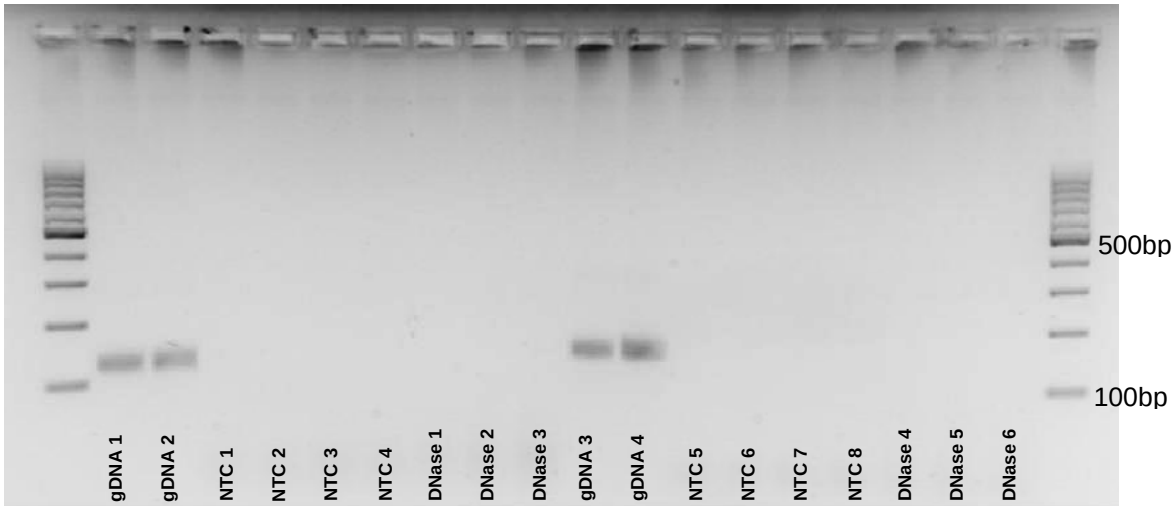
*primers perfect control for cDNA sequence. P1 = primer pair 1, P2 = primer pair 2 (producing genomic amplicons of differing sequence).

As presented in the following chapter, AEI analysis for *ANRIL* showed substantial effects for the transcribed SNPs, which can occur with normalisation artefacts. An additional gDNA primer pair for *ANRIL* that produced a product of similar size but different sequence showed no significant difference in the allelic ratio in gDNA ($P>0.05$), as shown in Table 3.1, confirming the appropriateness of the gDNA normalisation strategy. Biases attributable to sequence differences in the gDNA normalisation assay were likely to be less than the potential biases that could be introduced using an experimental 1:1 equimolar mix, and the gDNA normalisation factor was therefore used for normalisation in all subsequent analyses.

For *CDKN2B* the same primers were used to measure the AER in cDNA and gDNA. This had the advantage that AERs in genomic DNA were perfect experimental controls for normalisation of the cDNA, but the disadvantage that gDNA remaining in the RNA solution or gDNA contamination could potentially influence the results. To exclude effects from DNA remaining in the RNA solution after DNase treatment, control reactions were performed using the DNase treated RNA solution as template for the PCR reaction. There was no amplification in controls, demonstrating that DNase treatment was adequate (as illustrated in Figure 3.14).

Figure 3.14. Agarose gel electrophoresis showing no amplification of DNA from RNA solution after DNase treatment.

Gel shows the products of 45-cycle multiplex *CDKN2B* AEI PCR, run alongside a DNA ladder with 100bp markers. Amplification is seen in genomic DNA controls (gDNA) but not in the no-template controls (NTC) or DNase treated RNA samples without reverse transcription (DNase). This confirmed that DNase treatment adequately removed genomic DNA and that this would not bias the AEI assay.



3.4.4.2 Comparison of allelic ratios in genomic DNA in the Caucasian and SA cohorts

Genomic ratios were subsequently analysed in all individuals from the SA cohort to exclude any effects of genomic copy number variants or rare sequence variants influencing primer binding sites, and to allow a comparison of different normalisation strategies of normalising each individual's cDNA ratio to the ratio obtained in their own genomic DNA, versus normalising all cDNA samples to a mean genomic ratio. As shown in Table 3.2, mean genomic allelic ratio in the SA cohort showed small magnitude but statistically significant differences from the original estimates in the Caucasian population for some assays. However, these experiments were performed several months after the original experiments analysing the cDNA samples and estimating the normalisation ratios in the Caucasian cohort. In the intervening period the Sequenom MALDI-TOF analyser had been physically moved to a different building and been serviced/repaired (including recalibration), and improvements had been made to the SpectroChips, advertised as improving peak resolution and in particular reducing salt adduct peaks, which could potentially influence assessment of AER. Allelic ratios in gDNA were therefore repeated using the same samples originally used for estimation of genomic ratio in the Caucasian cohort. As shown in Table 3.2, the repeated measurements differed from the original estimates, but were similar to the estimates obtained in the SA cohort. It therefore seemed likely that the difference between the original Caucasian ratios and the SA ratios was due to assay shift due to the changes in the Sequenom hardware and SpectroChips. When the SA genomic samples were normalised to the repeated Caucasian genomic results, AER did not differ between the populations, as shown in Table 3.2. The mean gDNA ratio was therefore used for normalisation of all samples. Although the absolute changes resulting from these differences in conditions were small, these experiments highlight the principle that samples that are to be compared should be analysed in exactly the same way, and that genomic normalisation should be performed at the same time as the samples. However, recalibration of the assay using a panel of genomic controls can be used to correct for assay bias if unavoidable changes in conditions do occur.

Table 3.2. Comparison of genomic ratios obtained after Sequenom modifications.

Gene	SNP	Allele 1 / allele 2	Measurement	Mean gDNA ratio	P-value for comparison between original Caucasian AER and adjusted SA AER**
ANRIL	rs10965215	G/A	Original Caucasian	0.97	0.70
			SA	0.96	
			Repeat Caucasian	0.96	
	rs564398	G/A	Original Caucasian	0.81	0.10
			SA	0.87*	
			Repeat Caucasian	0.85	
CDKN2A	rs3088440	G/A	Original Caucasian	0.88	0.37
			SA	0.95*	
			Repeat Caucasian	0.96	
	rs11515	C/G	Original Caucasian	1.26	0.42
			SA	1.29	
			Repeat Caucasian	1.27	
CDKN2B	rs1063192	C/T	Original Caucasian	0.96	0.03
			SA	1.03*	
			Repeat Caucasian	1.07	
	rs3217992	G/A	Original Caucasian	0.98	0.79
			SA	0.98	
			Repeat Caucasian	0.98	

* P<0.05 for comparison between AER in original Caucasian measurement and SA measurement.
** Adjusted SA AER measurement obtained by normalising the SA samples to the mean repeat Caucasian AER (dividing SA AERs by the repeat Caucasian mean AER). Groups compared using 2-sample t-test or Mann-Whitney test if not normally distributed.

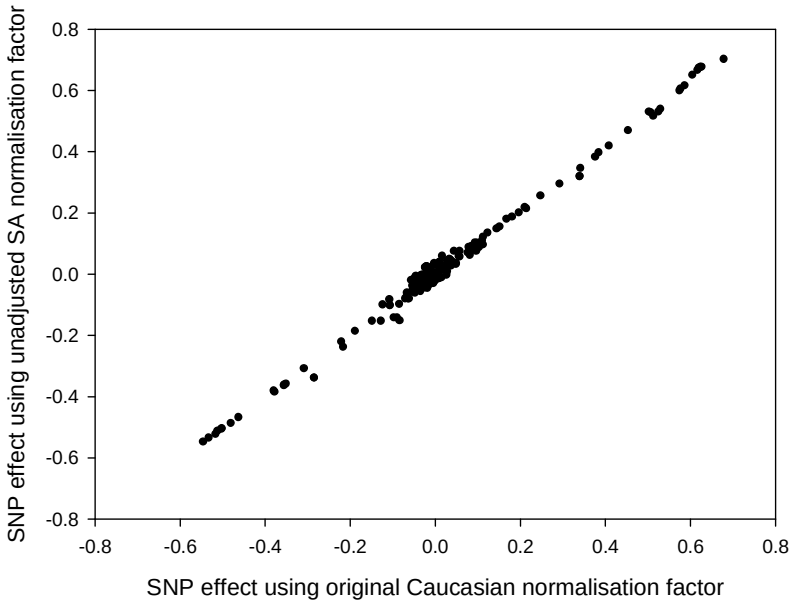
3.4.4.3 Influence of normalisation factor on mapping of *cis*-acting effects

An analysis was performed to assess what effect variations in normalisation factor estimates of the size that had been observed had on mapping of *cis*-acting effects. Results of SNP mapping were compared for data normalised using the original Caucasian gDNA ratios, and data normalised using the uncorrected SA gDNA ratios (i.e. without correction for assay shift), as discussed above and presented in Table 3.2.

As shown in Figure 3.15, mapping results were highly correlated using the different normalisation factors (r=0.997, $P=2x10^{-179}$). As expected, estimates of SNP effects were most influenced for the transcribed SNPs themselves, but little difference was seen for other SNPs, despite the strength of LD in the region, suggesting that the methodology is robust to the effects of small variations in normalisation ratios. The methodology of combining AERs from multiple transcribed SNPs per gene will tend to reduce errors associated with normalisation problems.

Figure 3.15. Effect of different normalisation ratios on mapping of *cis*-acting effects for *CDKN2A, CDKN2B* and *ANRIL*.

Points represent SNP effects for the 56 mapping SNPs in each of the three genes. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).

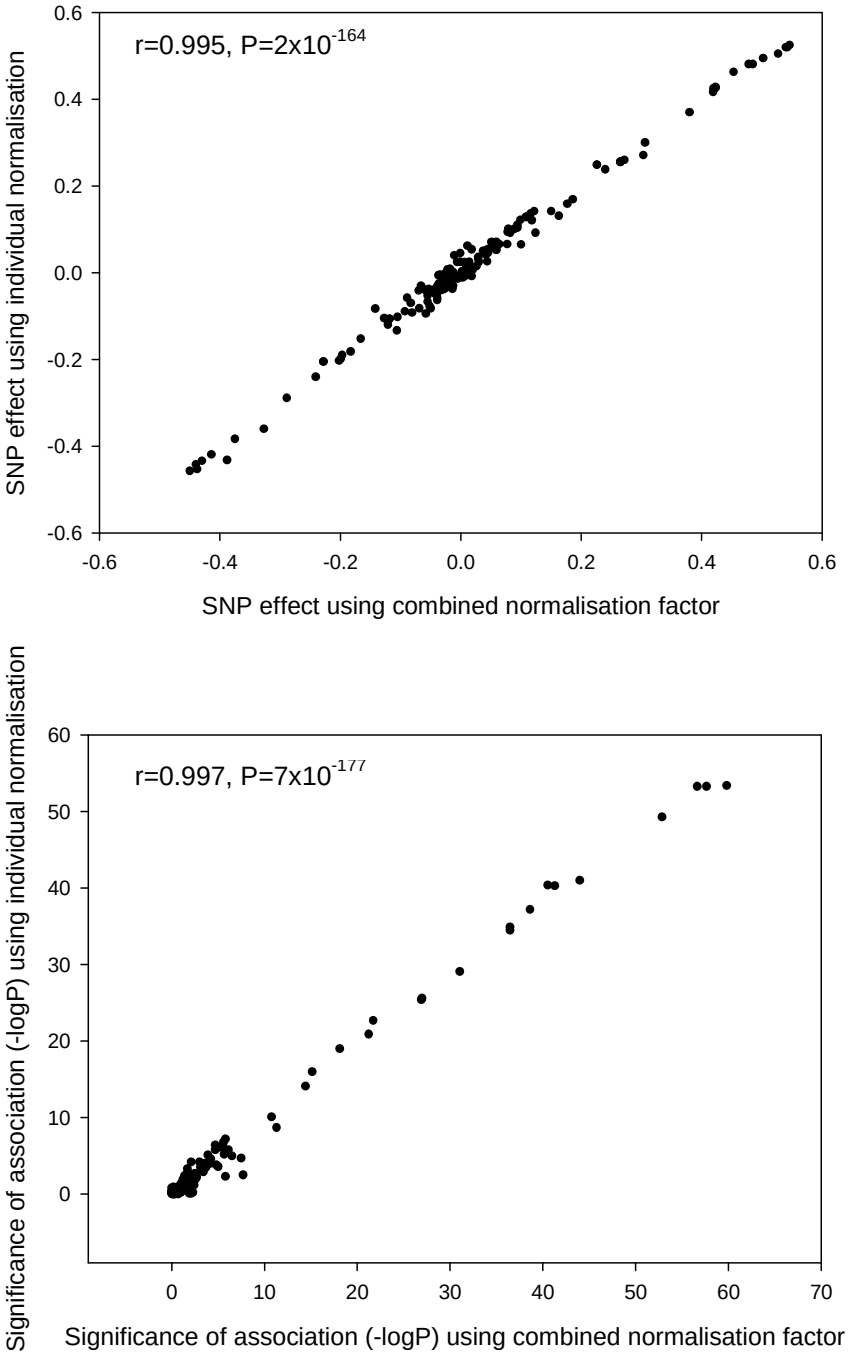


3.4.4.4 Comparison of individual genomic versus pooled genomic normalisation

The results of expression mapping using two different normalisation strategies were compared in the SA cohort: normalising to a mean population normalisation factor versus normalising each individual's cDNA to their own gDNA ratio. There was no significant difference in the results obtained using these two normalisation strategies, as shown in Figure 3.16. The mean gDNA ratio was therefore used for normalisation of all samples.

Figure 3.16. Effect of individual normalisation of allelic expression ratios.

Scatter plots show the estimates of effect size (A) and significance of association (B) for each of the 56 SNPs in the three genes tested. Each plot compares the values obtained using allelic expression ratios normalised to a combined normalisation factor (X-axis) versus individual normalisation of each cDNA ratio to the gDNA ratio from the same individual (Y-axis). Pearson correlation coefficient and the P-value for each association are shown. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



3.5 Discussion

Simulations suggest that combining data from multiple transcribed SNPs would be expected to increase the power of aeQTL analysis, but this study provides important data to support the feasibility of such an approach. The study demonstrates that AER measured at multiple transcribed markers in the same transcript are closely correlated, and moreover, that the *cis*-acting effects estimated using single or multiple transcribed SNPs per gene are also similar. This supports the validity of combining multiple markers. This methodology may allow wider application of the aeQTL approach for genes which have several transcribed SNPs that are individually too rare to otherwise provide sufficient numbers of informative individuals. All four of the genes we investigated had multiple transcribed variants, suggesting that this is not an uncommon situation such that this approach is likely to be of practical value.

MTAP showed little evidence of AEI in the pilot investigations. Although this does not exclude the presence of *cis*-acting effects that could be mapped in a study of this power, it means that any such effects are likely to be of small magnitude. This gene is also the furthest from the CAD risk locus and had the least supportive evidence for a potential role in the pathogenesis of CAD and hence it was not investigated further in this study.

The study demonstrates that AEI can be reliably assessed in RNA that has been stored for longer than a year. Similar AEI results were obtained with hME and iPLEX methodologies and with multiplexed AEI reactions, offering the potential to increase the throughput and decrease the cost of AEI analysis on the Sequenom platform. As previously discussed, a number of different approached have been used for normalisation of AEI measurements[260, 306-309]. The consistency of AEI measurements in gDNA and finding that normalisation using a small number of samples did not substantially alter the results of aeQTL mapping compared to normalising each sample individually to its own gDNA ratio supports the approach of using gDNA as an equimolar reference for normalisation. This suggests that the gDNA ratio may not need to be measured in every individual for these genes, although the situation would need to be investigated at other loci to exclude potential effects of CNVs and unannotated variants in primer-binding sequences that could alter AEI measurements.

A reasonable pragmatic approach might be to investigate the variance of gDNA AEI ratios in a subset of individuals which could then be used for normalisation if there was little variation in gDNA measurements, but with measurement of additional/all gDNA samples if there was significant inter-individual variability. Confirmation of the gDNA ratio in individuals which are outliers for cDNA AEI expression should also be performed. Small differences in the normalisation ratio did not substantially alter the results of AEI mapping, which is likely to be particularly the case when multiple transcribed markers are used. The approach of measuring AEI in experimental equimolar mixes provides a method of confirming the normalisation ratio in cases of doubt. The same machine, reagents, and analysis settings should be used for cDNA and normalisation experiments, preferably performed at a similar time, to reduce the risk of biases due to assay shift.

The results of AEI analysis in the chromosome 9p21 region presented in the following chapter are based on combined transcribed markers, iPLEX methodology, and normalisation to pooled gDNA ratios.

Chapter 4

Influence of chromosome 9p21 polymorphisms on gene expression in *cis*

4 Influence of chromosome 9p21 polymorphisms on gene expression in *cis*

4.1 Abstract

SNPs on chromosome 9p21 are associated with CAD, diabetes and multiple cancers. Most risk SNPs are not present in mature transcripts, suggesting that they influence expression and act in *cis*. This study examined the association between 56 SNPs in this region and peripheral blood expression of the three nearest genes *CDKN2A*, *CDKN2B* and *ANRIL* using total and allelic expression in two populations of healthy volunteers: 177 British Caucasians and 310 mixed-ancestry South Africans. Total expression of the three genes was significantly correlated, suggesting that they are co-regulated. SNP associations mapped by allelic and total expression were similar ($r=0.97$, $P=4.8x10^{-99}$), but the power to detect effects was greater for allelic expression. The proportion of expression variance attributable to *cis*-acting effects was 8% for *CDKN2A*, 5% for *CDKN2B*, and 20% for *ANRIL*. SNP associations were highly correlated in the two populations ($r=0.94$, $P=10^{-72}$). Multiple SNPs were independently associated with expression of each gene after correction for multiple testing, suggesting that several sites may modulate disease susceptibility. Individual SNPs correlated with changes in expression up to 1.4-fold for *CDKN2A*, 1.3-fold for *CDKN2B*, and 2-fold for *ANRIL*. Risk SNPs for CAD, stroke, diabetes, and melanoma were all associated with reduced allelic expression of *ANRIL*, while association with the other two genes was only detectable for some risk SNPs. SNPs had an inverse effect on *ANRIL* and *CDKN2B* expression, supporting a role of antisense transcription in *CDKN2B* regulation. This study suggests that modulation of *ANRIL* expression mediates susceptibility to several important human diseases.

4.2 Introduction

Recent GWA studies and candidate gene studies have shown that chromosome 9p21 variants are associated with a range of conditions (as previously summarised in Figure 1.4 on page 35). Although candidate gene approaches usually focused on transcribed variants, most of the SNPs associated with disease in GWA studies are located in non-coding regions, suggesting that their effects are likely to be mediated by influences on

gene expression. Since most do not appear in mature transcripts, and there are no known or predicted microRNAs mapping to this region[44, 316-318], these variants are unlikely to produce diffusible *trans*-acting factors and are therefore likely to influence expression of nearby genes in *cis*. Genes in the region include the cyclin-dependent kinase inhibitors *CDKN2A* and *CDKN2B*, the products of which are established tumour suppressors that influence physiological processes that may be relevant for CAD such as replicative senescence, apoptosis, and stem-cell self-renewal[184]. The region also contains the non-coding RNA *ANRIL*. The function of *ANRIL* is unknown, but other processed non-coding RNAs are involved in the regulation of gene expression through transcriptional and translational control mechanisms[199].

At the time this study was undertaken, no other studies investigating the relationship between risk variants and expression in the chromosome 9p21 region had been published. Two *ANRIL* transcript variants had been described, a short isoform comprised of exons 1-13, and a longer isoform containing exons 1-12 spliced to exons 14-20[195]. The AEI assay used to investigate *ANRIL* expression in this study used primers located in the first and second exons, as described in the previous chapter. This was primarily selected because transcribed SNPs of suitable allele frequency were located in exon 2. Previous work had also suggested that antisense transcription within the first intron of *ANRIL* downregulates *CDKN2B* expression in *cis*[202], hence the assay spanning *ANRIL* exons 1-2 had the advantage that it would have a plausible mechanistic basis for effects mediated through its expression. However, a number of very recent studies suggest that the situation may be more complicated with respect to *ANRIL* alternative splicing and transcript-specific expression, as presented in the discussion chapter[155, 197, 200]. Throughout this chapter, unless otherwise stated '*ANRIL*' expression will be used to refer to expression measured using the exon 1-2 assay. Preliminary work using assays located in different *ANRIL* exons will be presented separately.

As previously discussed, genetic effects on expression can be investigated using total expression levels (eQTL mapping) or allelic expression levels (aeQTL mapping). Simulations suggest that the power to detect *cis*-acting effects will be greater for the aeQTL approach in the presence of significant *trans*-acting influences. Expression of *CDKN2A/ARF* and *CDKN2B* has been shown to be influenced by factors that are

likely to act in *trans*, including age, chemotherapeutic agents, levels of transcriptional regulators, and DNA damage by ultraviolet or ionizing radiation[183], suggesting that analysis of allelic expression may be superior to total expression for mapping *cis*-acting effects at this locus. However, the power of the two approaches has never been compared experimentally. *Cis*-acting regulatory elements for *CDKN2A* and *CDKN2B* have been identified *in vitro* using reporter assays[186, 187, 189, 319, 320], but the *in vivo* relevance of these elements also needs to be confirmed.

Caucasian populations have strong LD in the chromosome 9p21 region which limits the ability to separate the effects of individual SNPs on expression[129]. Populations of African ancestry in general show less LD than Caucasian populations[321, 322], and this has been exploited to improve the fine-mapping of functional polymorphisms associated with quantitative traits at other loci [266, 323]. Analysis of LD in the HapMap YRI population suggested that such an approach may be useful in the chromosome 9p21 region.

4.3 Aims

The aims of this chapter were:

- To map the variants associated with *CDKN2A*, *CDKN2B* and *ANRIL* expression in the chromosome 9p21 region and determine whether disease-associated variants correlate with expression of these genes.
- To investigate whether effects on gene expression could be mapped with greater resolution using a population of African ancestry.
- To determine whether multiple loci are independently associated with expression at the chromosome 9p21 region.
- To compare the power of total and allelic expression for mapping *cis*-acting effects.
- To investigate whether *cis*-acting effects differ between transcripts for *CDKN2A/ARF* and *ANRIL* isoforms.
- To investigate the influence of microsatellite rs10583774 on expression of *CDKN2A*, *CDKN2B* and *ANRIL* expression.

4.4 Materials and methods

4.4.1 Participants, samples, genotyping, mapping SNPs, and AEI methods

As described in the previous chapter. AEI assays were performed using iPLEX methodology and were normalised to pooled gDNA controls. Microsatellite genotyping was performed as described in Chapter 2.

4.4.2 Relative quantification of total gene expression using real-time PCR

Real-time PCR reactions were performed using TaqMan methodology as described in Chapter 2. Commercially available FAM-labelled TaqMan assays (Applied Biosystems) were used for *CDKN2A* exons 2-3 (Hs00923894_m1) and *ANRIL* exons 1-2 (Hs01390879_m1). A custom FAM-labelled assay was used for exon 2 of *CDKN2B*. Commercially available VIC-labelled TaqMan assays were used for three reference genes shown to be suitable for normalisation of expression in peripheral blood[291, 292]: *B2M* (4326319E), *GAPDH* (4326317E), and *HPRT1* (4326321E). TaqMan assays are validated by the manufacturer to have close to 100% amplification efficiency and assays were selected to quantify the same transcripts as the allelic expression assays. PCR was performed according to the manufacturer's protocol using four replicates, 25ng cDNA template per reaction, and the following multiplex combinations: *CDKN2A*/*B2M*, *CDKN2B*/*GAPDH*, and *ANRIL*/*HPRT1*.

For transcript-specific analyses, a commercially available FAM-labelled assay was used for exon 1β of *ARF*. Custom FAM-labelled assays were also used, with primer/probesets located within *CDKN2A* exon 1α, *ANRIL* exon 13, and *ANRIL* exon 20. The VIC-labelled TaqMan assays for *B2M* and *GAPDH* (described above) were used for reference of these assays, using the following multiplex combinations: *CDKN2A*specific/*B2M*, *ARF*/*GAPDH*, *ANRIL*exon13/*B2M*, and *ANRIL*exon20/*GAPDH*. Normalisation was performed using the reference genes performed on the same plate as the target genes. The methodology was otherwise unchanged.

Relative total expression was analysed using the comparative cycle threshold (Ct) method. Ct values for each target gene were normalised to the average Ct value of the reference genes[291]. Standard errors and variances of measurements for allelic and total expression analyses in the SA population are shown in Table 4.1.

Table 4.1. Comparison of variances between total expression and allelic expression measurements in the SA cohort.

	Total expression			Allelic expression	
	Variance within samples	**Variance between samples**	**Transcribed marker**	**Variance within samples**	**Variance between samples**
CDKN2A	0.057	0.434	rs3088440	0.016	0.111
			rs11515	0.028	0.064
CDKN2B	0.038	0.478	rs1063192	0.043	0.044
			rs3217992	0.018	0.039
ANRIL	0.590	0.697	rs10965215	0.073	0.106
			rs564398	0.063	0.076

4.4.3 Statistical analyses

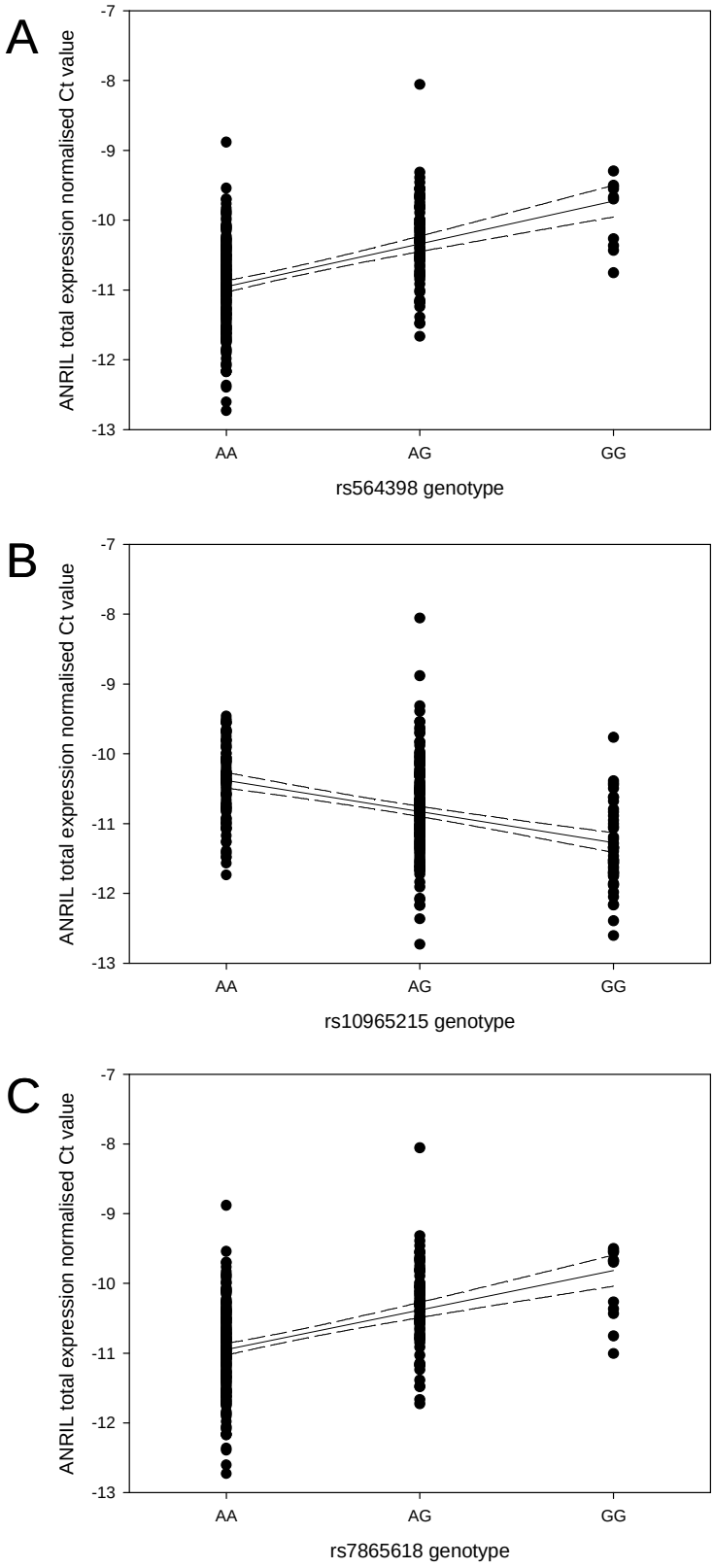
4.4.3.1 aeQTL and eQTL mapping

Allelic expression analysis was performed as described in Chapters 2 and 3[304]. The association between total expression and each of the mapping SNPs was assessed using linear regression of the log transformed normalized expression values on the genotype. Plots illustrating the associations between genotype and total expression for selected SNPs are shown in Figure 4.1. The effect of including age, sex, and ethnicity as covariates, as well as excluding outlying individuals was investigated. Self reported ethnicity was included as a categorical variable (categorised as "Cape mixed-ancestry", "black African", "white", "Indian", and "other").

Since total expression data was only available for the SA cohort, unless otherwise stated analyses directly comparing allelic and total expression data only included allelic expression data from the SA cohort to make the results comparable.

Figure 4.1. Effect of genotype on total expression of *ANRIL* for selected SNPs.

Y-axis shows the normalised total expression value for *ANRIL*. X-axis shows genotype for SNPs with cis-acting effects: (A) rs564398; (B) rs10965215; (C) rs7865618. Linear regression lines are shown as solid lines, with dotted lines indicating the 95% confidence intervals



4.4.3.2 Correction for multiple testing

For both total and allelic expression, multiple testing was taken into account by calculating the family wise error rate using a Bonferroni correction for the 56 SNPs tested. Associations with family wise error rate (FWER) below a threshold of 0.05 (corresponding to a nominal P-value of 8.9×10^{-4}, $-\log_{10}P$ of 3.05, and $-\log_{10}FWER$ of 1.3) were considered significant.

4.4.3.3 Estimation of *cis* and *trans* effects

From the allelic and total expression data it is also possible to estimate the proportion of total expression variance that is due to *cis*-acting effects. This assumes that *cis* and *trans*-acting factors act in an additive manner, do not interact, are independent, and that there is random mating, no segregation distortion (i.e. an equal probability of each allele being passed on to offspring), and the locus is not subject to imprinting. Under these assumptions, expression from each allele (designated as e_1 and e_2) can be decomposed into two components, one that is allele specific resulting from *cis*-acting factors and one that is shared between both alleles and reflects *trans*-acting influences. Total expression is the sum of the contributions from both chromosomes (i.e. $e_1 + e_2$). The ratio e_{1_1} / e_2 and *cis* and *trans* effects on expression are assumed to be the same for maternal and paternal alleles in the population. Based on these assumptions, calculations to estimate the proportion of total expression variance that was due to *cis*-acting effects were derived by Dr Mauro Santibanez Koref as follows:

The variance due to *cis* acting effects, $V(c)$, can therefore be estimated as

$\hat{V}(c) = \frac{1}{2(n-1)} \sum_{i=1}^{n} e_{Ti}^2 \left(\frac{1-r_i}{1+r_i} \right)^2$, where r_i is the allelic expression ratio for individual *i*,

and the proportion of the total variance due to *cis* acting effects can be estimated as

$\frac{2\hat{V}(c)}{\hat{V}(e_T)}$, where $\hat{V}(e_T)$ is the estimated total variance, i.e.

$\hat{V}(e_T) = \frac{1}{n-1} \sum_{i=1}^{n} (e_{Ti} - \bar{e}_T)^2$ with $\bar{e}_T = \frac{1}{n} \sum_{i=1}^{n} e_{Ti}$ and e_{Ti} represent the total expression level for individual *i* as determined by real time PCR.

4.5 Results

4.5.1 Genotyping results

Information on the selected SNPs and genotyping data are summarised in Table 4.2. A higher proportion of SNPs in the NE Caucasian cohort showed significant deviations from HWE at the P<0.05 threshold (12/53, 23%) and the P<0.01 threshold (2/53, 4%) than would be expected by chance alone. Using the same assays in the SA cohort, only 1/56 SNPs (1.7%) deviated from HWE at the P<0.05 threshold. This issue and its further investigation are dealt with in detail in Chapter 5.

4.5.2 Variability of *CDKN2A*, *CDKN2B* and *ANRIL* expression

4.5.2.1 Inter-individual variation in expression

Total expression levels showed substantial inter-individual variation for each of the three genes, up to 13.9-fold for *CDKN2A*, 36.1-fold for *CDKN2B*, and 25.5-fold for *ANRIL*. Plots of the normalised total expression Ct values are shown in Figure 4.2. Allelic expression ratios at individual transcribed markers also showed considerable inter-individual variation, up to 5.6-fold for *CDKN2A*, 2.4-fold for *CDKN2B*, and 6.8-fold for *ANRIL*. Plots of the allelic expression ratios at each transcribed SNP in the SA and Caucasian cohorts are shown in Figure 4.3 and Figure 4.4. Systematic overexpression of one allele was observed at both transcribed SNPs in *ANRIL* (D'=1, r^2=0.14 between these SNPs in the SA cohort), suggesting that *cis*-acting influences on expression were attributable to the transcribed SNPs themselves, or to elements in strong LD with them. The transcribed markers in the other genes did not show consistent overexpression of one allele, but the increased variability of the expression ratios compared to the gDNA samples indicated the presence of *cis*-acting effects attributable to loci that were not in strong LD with the transcribed markers.

Standard errors for *ANRIL* were higher than for the other two genes in both the allelic and total expression assays, which is likely to be due to the fact that peripheral blood expression of *ANRIL* was lower than for *CDKN2A* and *CDKN2B*.

Table 4.2. Summary of included SNPs.

					Caucasian cohort					SA cohort				
SNP ID	Chr position	Gene info	Functional info	Selection reason	% Genotyped	MAF	Heterozygosity	Alleles	HW P-value	% Genotyped	MAF	Heterozygosity	Alleles	HW P-value
rs7023954	21806758	MTAP		MTAP gene	96.0	0.44	0.50	G:A	1.00	94.8	0.34	0.42	G:A	0.33
rs15735	21852271	MTAP		MTAP gene	100.0	0.25	0.44	A:C	0.06	98.7	0.41	0.40	A:C	0.003
rs1134871	21852897	MTAP		MTAP gene	98.9	0.23	0.40	T:A	0.20	100.0	0.43	0.43	A:T	0.04
rs3731257	21956221		Ovarian ca	Phenotypic association	97.7	0.32	0.39	G:A	0.23	99.4	0.21	0.35	G:A	0.41
rs3088440	21958159	CDKN2A transcribed	Melanoma, Pancreatic ca, Ovarian ca, Bladder ca	CDKN2A Transcribed	100.0	0.07	0.13	G:A	1.00	100.0	0.21	0.33	G:A	0.87
rs11515	21958199	CDKN2A transcribed	Alzheimer's, Bladder ca, Pancreatic ca	CDKN2A transcribed	98.9	0.11	0.19	C:G	1.00	100.0	0.14	0.24	C:G	0.64
rs3731249	21960916		Breast ca, melanoma, ALL	Phenotypic association	98.3	0.02	0.04	C:T	1.00	99.7	0.01	0.02	C:T	1.00
rs3731239	21964218		CAD, breast ca	Phenotypic association	95.5	0.35	0.44	T:C	0.73	100.0	0.15	0.26	T:C	1.00
rs3814960	21965017	CDKN2A promoter		Promoter	F					98.7	0.38	0.44	C:T	0.23
rs36228834	21965319	CDKN2A promoter		Promoter	98.3	0.02	0.04	T:A	1.00	99.7	0.01	0.02	T:A	1.00
rs7036656	21980457			Tag	96.6	0.25	0.40	T:C	0.56	100.0	0.23	0.36	T:C	0.87
rs2811711	21983964	ANRIL promoter		Promoter	98.3	0.16	0.28	T:C	1.00	99.4	0.09	0.16	T:C	0.67
rs1801022	21984347	ANRIL promoter		Promoter	97.7	0.00	0.00	C:C	1.00	99.7	0.00	0.00	C:C	1.00
rs2518723	21985882	CDKN2A^{ARF} promoter	Colorectal ca	Phenotypic association	92.1	0.43	0.41	C:T	0.05	99.4	0.45	0.50	C:T	0.97
rs3218022	21987723	CDKN2A^{ARF} promoter		Promoter	97.7	0.00	0.01	T:C	1.00	99.7	0.03	0.06	T:C	1.00
rs3218020	21987872	CDKN2A^{ARF} promoter		Promoter	100.0	0.42	0.42	G:A	0.08	99.7	0.25	0.40	G:A	0.29
rs2811712	21988035	CDKN2A^{ARF} promoter	Frailty, breast ca	Phenotypic association	97.2	0.06	0.12	A:G	1.00	99.7	0.22	0.37	A:G	0.11
rs3218018	21988139	CDKN2A^{ARF} promoter	Diabetes	Phenotypic association	98.3	0.04	0.09	T:G	1.00	99.7	0.08	0.15	T:G	1.00
rs3218012	21988660	CDKN2A^{ARF} promoter	Colorectal ca	Phenotypic association	97.7	0.50	0.45	G:A	0.18	99.7	0.46	0.53	G:A	0.33
rs3218009	21988757	CDKN2A^{ARF} promoter	CAD	Phenotypic association	96.0	0.13	0.22	G:C	1.00	100.0	0.03	0.06	G:C	0.08
rs3218005	21990247		Breast ca	Phenotypic association	96.6	0.04	0.08	T:C	1.00	99.4	0.21	0.36	T:C	0.18
rs3217992	21993223	CDKN2B transcribed	CAD	CDKN2B transcribed	98.3	0.46	0.41	G:A	0.02	100.0	0.24	0.36	G:A	1.00
rs1063192	21993367	CDKN2B transcribed	Glioma	CDKN2B transcribed	98.9	0.42	0.40	T:C	0.02	100.0	0.18	0.28	T:C	0.70
rs3217986	21995330	CDKN2B transcribed		CDKN2B transcribed	96.6	0.08	0.14	A:C	0.55	100.0	0.08	0.14	A:C	0.64
rs2069418	21999698	CDKN2B promoter		Promoter	97.7	0.44	0.43	C:G	0.15	98.7	0.18	0.30	C:G	1.00
rs495490	22000412	RD$^{INK4/ARF}$		Promoter	95.5	0.11	0.19	T:C	0.69	100.0	0.03	0.05	T:C	1.00
rs7044859	22008781		CAD, stroke	Phenotypic association	98.9	0.48	0.44	A:T	0.14	100.0	0.30	0.43	A:T	0.97
rs496892	22014351		CAD, stroke	Phenotypic association	100.0	0.40	0.42	G:A	0.14	100.0	0.30	0.45	G:A	0.35
rs615552	22016077			Tag	95.5	0.41	0.39	A:G	0.02	99.7	0.17	0.29	A:G	1.00
rs10965215	22019445	ANRIL transcribed		ANRIL transcribed	100.0	0.46	0.40	A:G	0.01	100.0	0.42	0.50	G:A	0.71

rs564398	22019547	ANRIL transcribed	Diabetes, CAD, stroke	ANRIL transcribed	97.7	0.38	0.39	A:G	0.02	100.0	0.17	0.28	A:G	0.99
rs7865618	22021005		CAD, stroke	Phenotypic association	98.9	0.41	0.39	A:G	0.02	100.0	0.18	0.27	A:G	0.22
rs17694493	22031998			Tag	96.6	0.09	0.18	C:G	0.50	100.0	0.10	0.18	C:G	1.00
rs10738605	22039130	ANRIL transcribed		ANRIL transcribed	98.9	0.45	0.41	G:C	0.02	99.4	0.45	0.50	G:C	1.00
rs11790231	22043591			Tag	96.6	0.12	0.21	G:A	0.96	100.0	0.06	0.11	G:A	0.64
rs2184061	22051562			Tag	93.2	0.38	0.41	A:C	0.10	99.7	0.40	0.49	A:C	0.91
rs1011970	22052134		Melanoma	Tag	83.6	0.16	0.28	G:T	0.96	97.4	0.28	0.42	G:T	0.45
rs10811650	22057593			Tag	95.5	0.47	0.37	C:G	0.001	100.0	0.29	0.41	C:G	1.00
rs16905599	22059144			Tag	F					91.9	0.19	0.29	G:A	0.51
rs10116277	22071397		CAD, stroke	Phenotypic association	100.0	0.49	0.36	T:G	0.0002	100.0	0.26	0.37	T:G	0.63
rs10965227	22071796			Tag	96.0	0.22	0.35	A:G	1.00	100.0	0.09	0.15	A:G	0.20
rs1547705	22072375			Tag	96.0	0.12	0.19	A:C	0.75	99.7	0.11	0.19	A:C	0.95
rs10965228	22072380			Tag	96.6	0.12	0.21	A:G	0.96	100.0	0.03	0.06	A:G	1.00
rs1333040	22073404		CAD, stroke	Phenotypic association	100.0	0.36	0.37	T:C	0.02	99.4	0.41	0.48	T:C	0.94
rs7857345	22077473			Tag	93.2	0.27	0.39	C:T	1.00	100.0	0.14	0.21	C:T	0.02
rs10757274	22086055		CAD	Phenotypic association	100.0	0.49	0.40	G:A	0.01	100.0	0.38	0.45	A:G	0.50
rs10125231	22092128			Tag	94.9	0.02	0.03	G:A	1.00	100.0	0.01	0.02	G:A	1.00
rs2383206	22105026		CAD, stroke	Phenotypic association	100.0	0.47	0.40	G:A	0.01	100.0	0.48	0.51	G:A	0.81
rs2383207	22105959		CAD, stroke	Phenotypic association	90.4	0.42	0.44	G:A	0.33	100.0	0.23	0.32	G:A	0.15
rs1333045	22109195		CAD	Tag	96.6	0.48	0.42	C:T	0.05	99.7	0.49	0.51	T:C	0.97
rs10757278	22114477		CAD, stroke	Phenotypic association	99.4	0.50	0.42	A:A	0.05	100.0	0.36	0.41	A:G	0.05
rs1333049	22115503		CAD	Phenotypic association	100.0	0.50	0.43	C:G	0.08	100.0	0.38	0.45	G:C	0.41
rs2891169	22121825		Diabetes	Phenotypic association	F					99.4	0.48	0.49	G:A	0.92
rs2383208	22122076		Diabetes	Phenotypic association	93.2	0.18	0.29	A:G	0.84	99.4	0.24	0.35	A:G	0.40
rs10811661	22124094		Diabetes	Phenotypic association	100.0	0.17	0.28	T:C	0.84	99.7	0.11	0.19	T:C	0.61
rs10757283	22124172		Diabetes	Phenotypic association	97.7	0.47	0.42	C:T	0.06	99.7	0.49	0.49	C:T	0.76

F = SNP removed from analysis in this cohort as genotype available for <80% of individuals. CAD = coronary artery disease; MAF = minor allele frequency; HW = Hardy-Weinberg.

Figure 4.2. Total expression values in the SA cohort.

Y-axes show normalised total expression Ct values relative to reference genes for: (A) *CDKN2A*; (B) *CDKN2B*; (C) *ANRIL*. Each point represents an individual, with standard error bars shown.

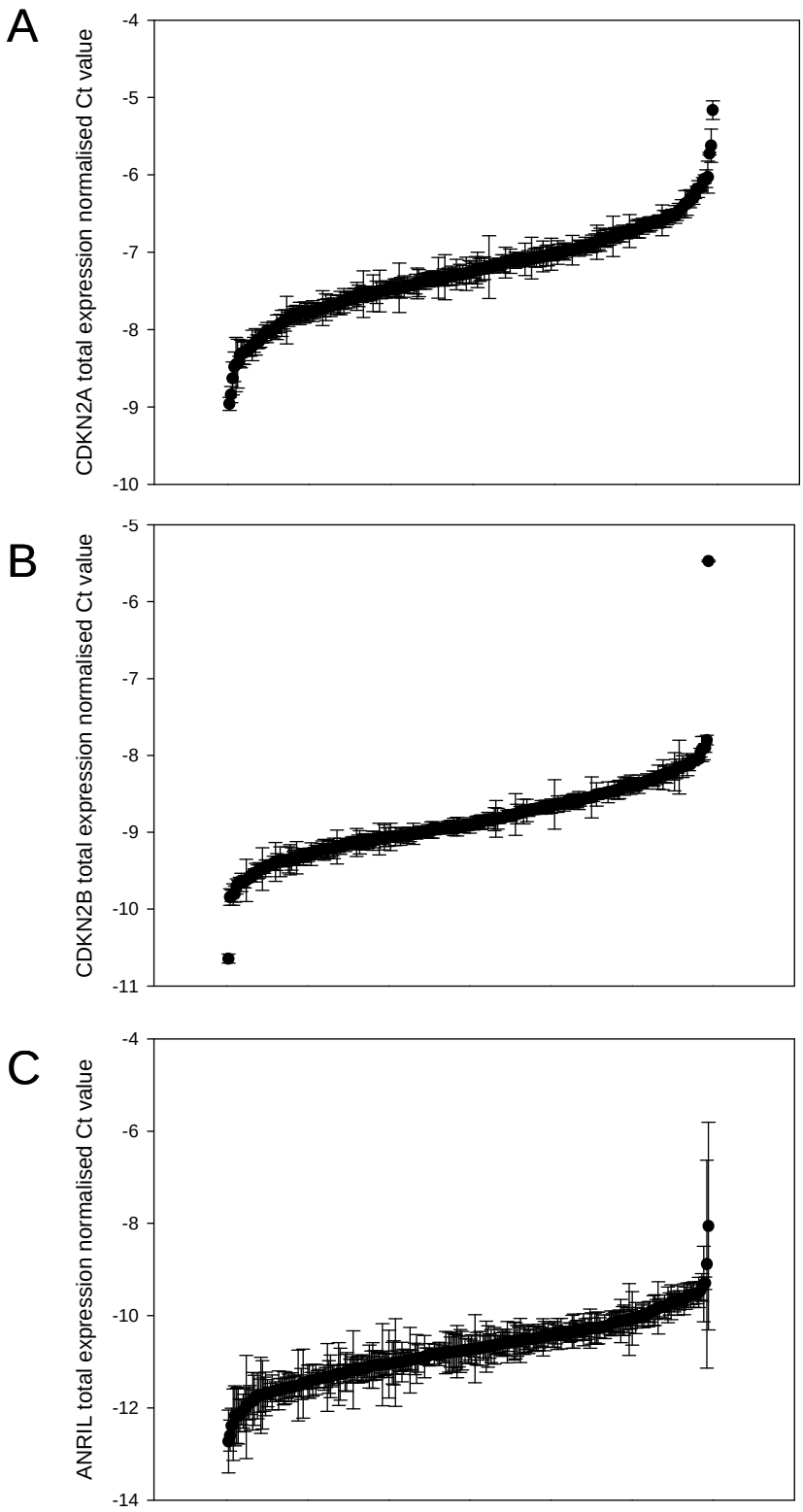


Figure 4.3. Allelic expression ratios at transcribed SNPs in the SA cohort.

Y-axes show AER for the following transcribed SNPs: (A) *CDKN2A* rs11515; (B) *CDKN2A* rs3088440; (C) *CDKN2B* rs3217992; (D) *CDKN2B* rs1063192; (E) *ANRIL* rs564398; (F) *ANRIL* rs10965215. Each point represents an individual, with standard error bars shown. Black circles represent cDNA measurements and blue circles genomic DNA measurements.

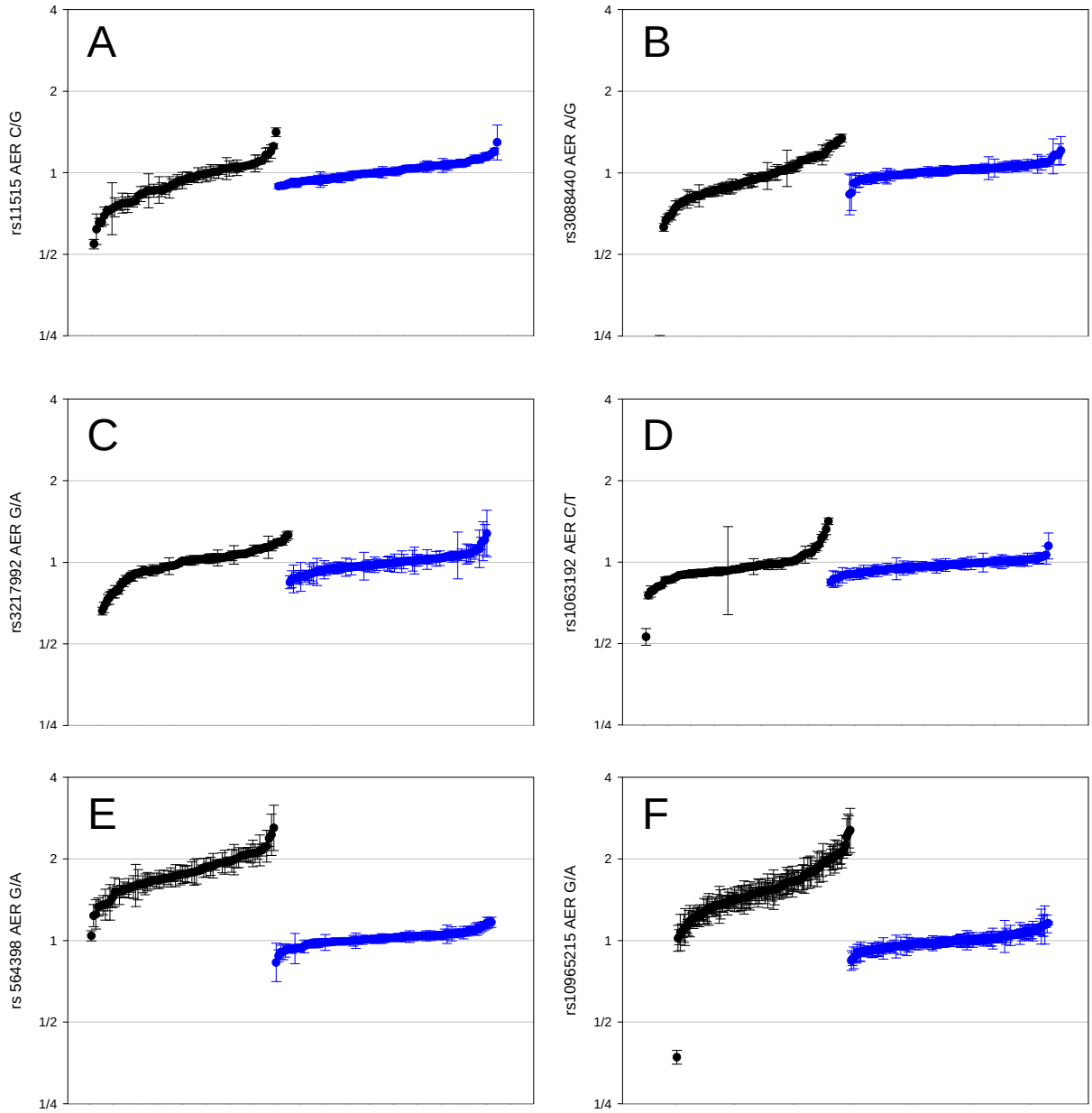
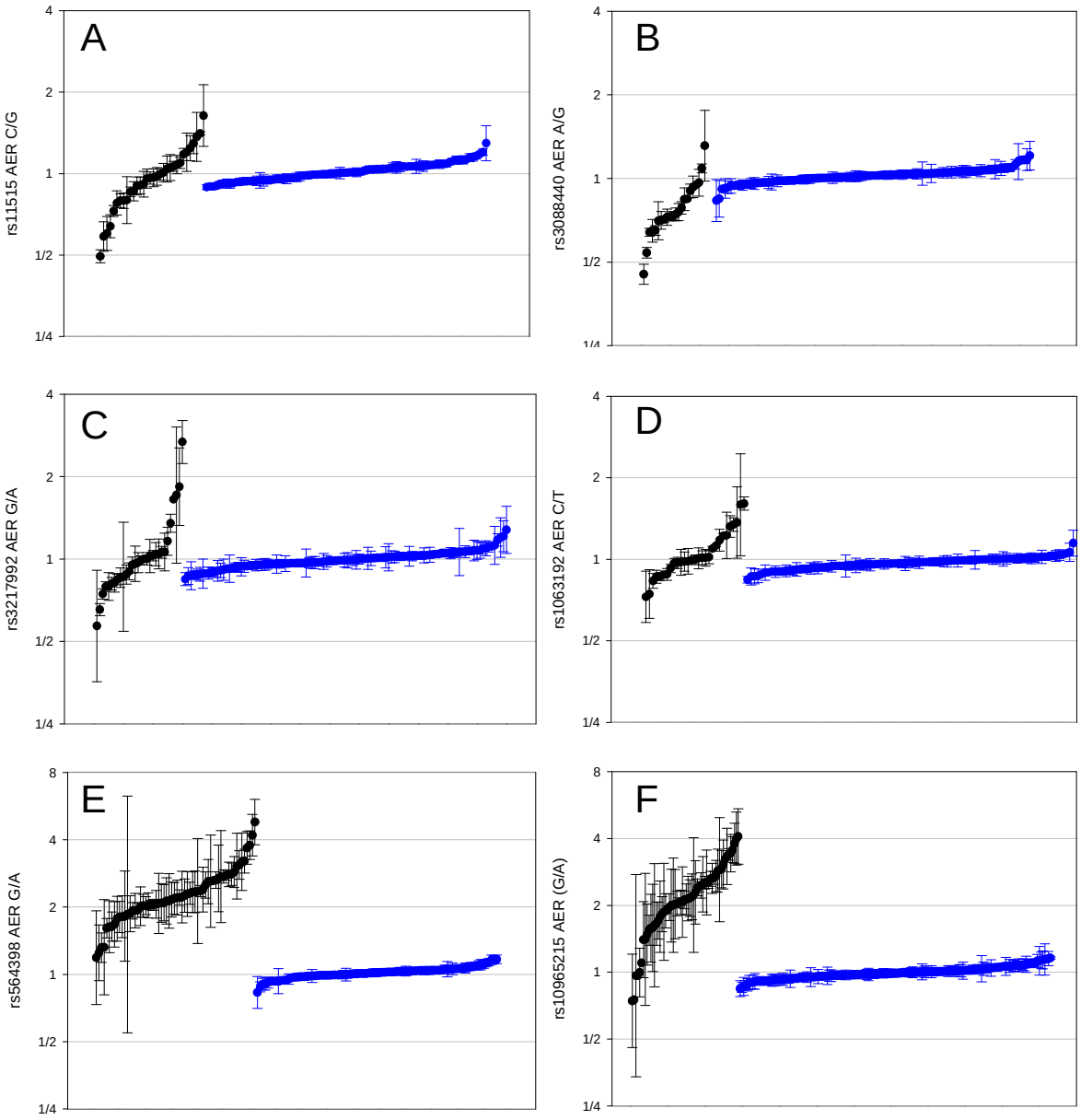


Figure 4.4. Allelic expression ratios at transcribed SNPs in the Caucasian cohort.

Y-axes shows AER for the following transcribed SNPs: (A) *CDKN2A* rs11515; (B) *CDKN2A* rs3088440; (C) *CDKN2B* rs3217992; (D) *CDKN2B* rs1063192; (E) *ANRIL* rs564398; (F) *ANRIL* rs10965215. Each point represents an individual, with standard error bars shown. Black circles represent cDNA measurements and blue circles represent genomic DNA measurements.



4.5.2.2 Proportion of variance attributable to *cis* and *trans* effects

The proportion of the variance in total expression that can be attributed to *cis*-acting effects for each transcribed SNP in the three genes was estimated, as shown in Table 4.3.

Table 4.3. Proportion of variance in total expression attributable to *cis*-acting effects estimated at each transcribed SNP.

Gene	Transcribed SNP	Proportion of variance in total expression attributable to *cis*-acting effects
CDKN2A	rs3088440	8%
	rs11515	4%
CDKN2A	rs3217992	5%
	rs1063192	5%
ANRIL	rs10965215	20%
	rs564398	19%

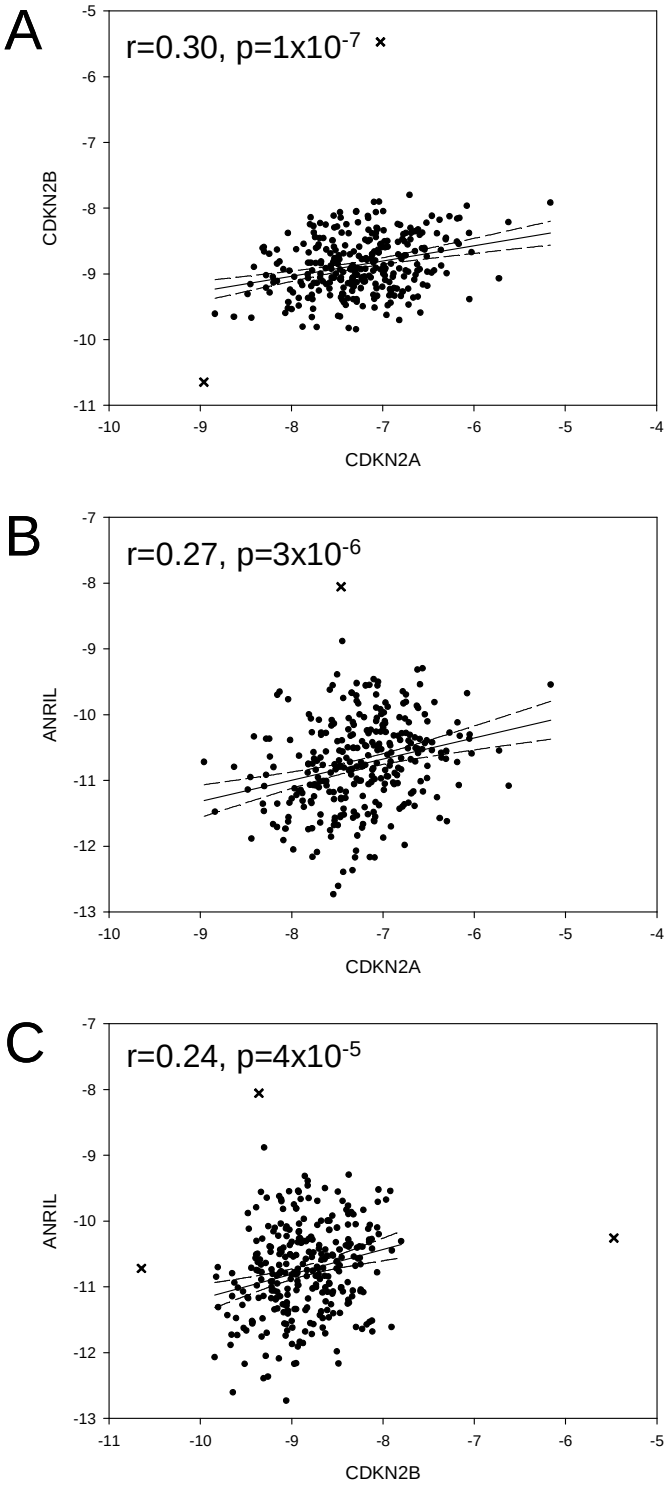
These data demonstrate that most of the variance in expression of these genes observed between different individuals in the population is due to *trans*-acting effects. The proportion attributable to cis-acting effects was higher for *ANRIL* than for *CDKN2A* and *CDKN2B.*

4.5.2.3 Correlation of *CDKN2A*, *CDKN2B* and *ANRIL* expression

Total expression levels of *CDKN2A*, *CDKN2B* and *ANRIL* showed a modest but statistically significant correlation (r=0.24 to 0.30, all $P<4x10^{-5}$), as shown in Figure 4.5. The finding of a very significant correlation between expression of these genes suggests that their expression is co-regulated by common *trans*-acting factors. However, the observation that the degree of correlation was only modest suggests that other influences on expression differ between the genes.

Figure 4.5. Correlations between total expression levels of *CDKN2A*, *CDKN2B* and *ANRIL*.

Scatter plots show correlations between total expression levels for: (A) *CDKN2A* and *CDKN2B*; (B) *CDKN2A* and *ANRIL*; (C) *CDKN2B* and *ANRIL*. Expression values on the X- and Y-axes are shown as delta Ct values for the target gene relative to the three internal control genes. Circles represent individual samples and the crosses represent three outliers excluded from correlation analyses. Linear regression lines are shown as solid lines, with dotted lines indicating the 95% confidence intervals. Pearson correlation coefficient (r) and the P-value for each association are shown in the top left of each plot.



4.5.3 Allelic expression versus total expression for mapping *cis*-acting effects

As previously presented in Chapter 3, AERs measured at the two transcribed SNPs in each gene were highly correlated (r=0.96, $P=3x10^{-61}$). AEI analysis was therefore performed using AERs from both transcribed markers in each gene. This approach increased the number of informative heterozygotes for each gene and the power to detect significant effects compared to the conventional single marker approach, as shown in Table 4.4.

Unlike allelic expression ratios, total expression data may be influenced by covariates that influence expression in *trans*. Total expression values were therefore corrected for covariates (age, sex, and ethnicity) and outlying individuals were excluded. These corrections did not significantly alter the results of the eQTL analysis, as shown in Figure 4.6. All subsequent analyses are presented using the covariate-corrected eQTL data.

Cis-acting effects assessed by eQTL and aeQTL mapping were compared, as shown in Figure 4.7. There was a strong correlation both for the effect size (r=0.87, $P=4.7x10^{-51}$) and significance of association (r=0.97, $P=4.8x10^{-99}$) at each mapping SNP between the two techniques, confirming that they are indeed measuring the same effects as expected. However, the associations were more significant for allelic expression than for total expression analysis, indicating that allelic expression had greater power for detecting *cis*-acting effects. As shown by the simulations presented in Chapter 1, the power of the AEI approach to detect *cis*-acting effects is expected to be greater than that of the total expression approach in the presence of substantial *trans*-acting influences on expression, since allelic expression analysis is more robust to the effects of *trans*-acting variance. The finding of greater power with the aeQTL approach is therefore consistent with the estimates that *cis*-acting effects account for only between 4 and 20% of the overall inter-individual variance in expression of these genes.

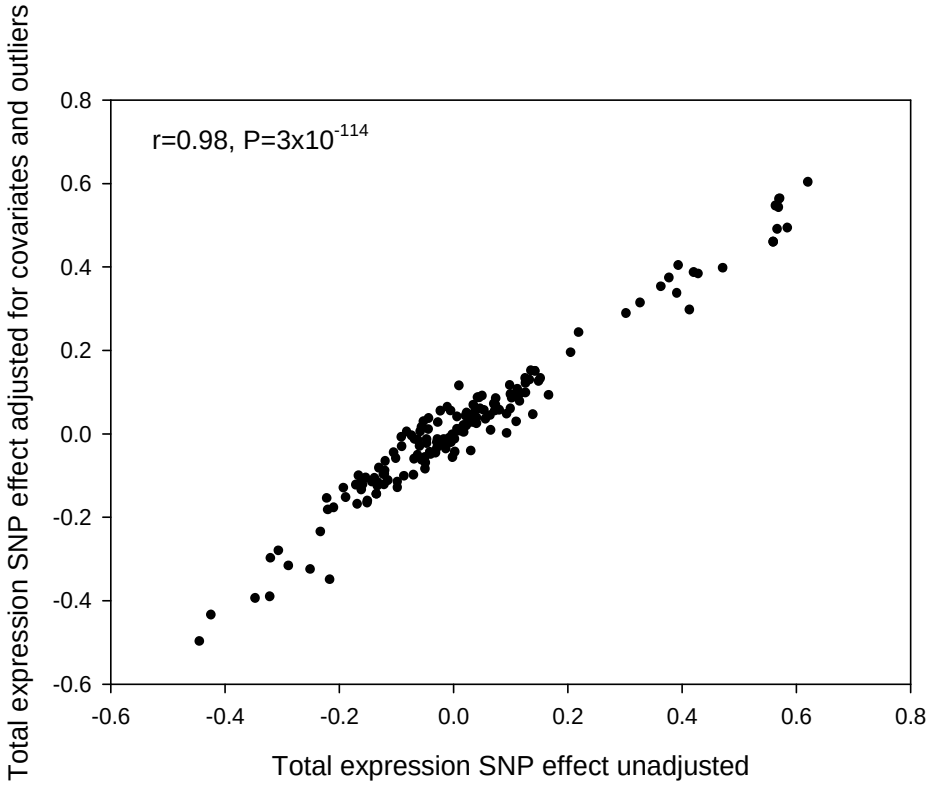
Table 4.4. Increase in number of informative heterozygotes and associated SNPs using two transcribed SNPs per gene.

Transcribed SNP	Number (%) of informative heterozygotes in Caucasian cohort (n=177)	Number (%) of informative heterozygotes in SA cohort (n=310)	Number (%) of mapping SNPs significantly associated with AER at transcribed SNP(s) in Caucasian cohort* (n=53 SNPs)	Number (%) of mapping SNPs significantly associated with AER at transcribed SNP(s) in SA cohort* (n=56 SNPs)
CDKN2A rs3088440	23 (12%)	103 (33%)	5 (9%)	2 (4%)
CDKN2A rs11515	33 (18%)	75 (24%)	0 (0%)	9 (16%)
***CDKN2A* markers combined**	**54 (29%)**	**159 (51%)**	**3 (6%)**	**11 (20%)**
CDKN2B rs3217992	71 (38%)	112 (36%)	0 (0%)	4 (7%)
CDKN2B rs1063192	70 (37%)	87 (28%)	0 (0%)	2 (4%)
***CDKN2B* markers combined**	**90 (48%)**	**164 (53%)**	**5 (9%)**	**5 (9%)**
ANRIL rs10965215	70 (37%)	155 (50%)	25 (47%)	27 (48%)
ANRIL rs564398	67 (36%)	85 (28%)	23 (43%)	22 (39%)
***ANRIL* markers combined**	**80 (43%)**	**187 (61%)**	**30 (57%)**	**31 (55%)**

* Multiple testing was taken into account by calculating the FWER using a Bonferroni correction for the 56 SNPs tested. Associations with FWER using a threshold of 0.05 (that corresponds to a nominal P value of 8.9×10^{-4} or $-\log_{10}P$ of 3.05) were considered significant. SNPs with less than eight informative heterozygotes were excluded.

Figure 4.6. Effect of adjustment for covariates and outliers on total expression mapping.

Scatter plots depict the estimates of effect size (A) and significance of association (B) for each of the 56 SNPs obtained using unadjusted total expression values (X-axis) versus values adjusted for covariates (age, sex, ethnicity) and with outliers removed (Y-axis). Pearson correlation coefficient (r) and the P-value for each association are shown in the top left of each plot. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



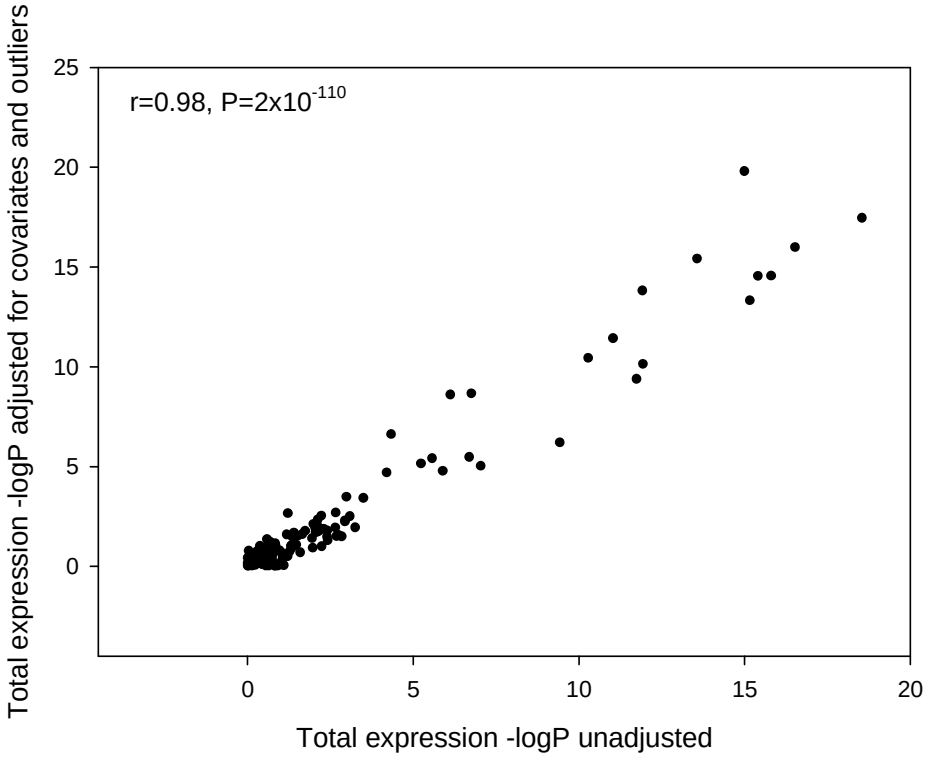
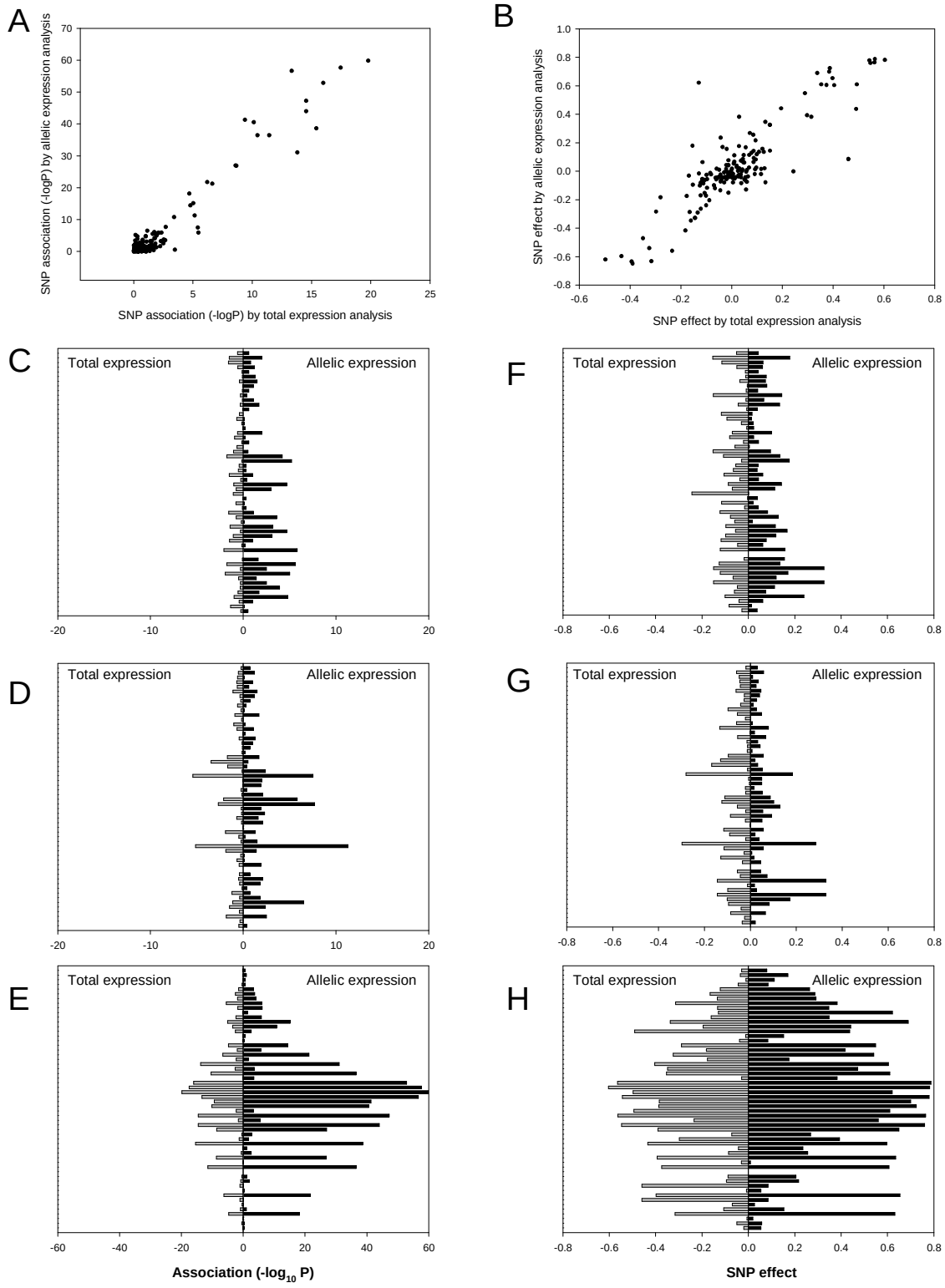


Figure 4.7. Significance of associations and effect size estimates using total and allelic expression.

Scatter plots depict the P-values (A) and estimates of effect size (B) obtained for each SNP for all three genes by eQTL (X-axis) and aeQTL (Y-axis) mapping. Bar charts show the comparison of the significance of association for *CDKN2A* (C), *CDKN2B* (D), and *ANRIL* exons1-2 (E); and the effect size estimates for *CDKN2A* (F), *CDKN2B* (G), and *ANRIL* exons1-2 (H). The Y-axes on the bar charts show the 56 SNPs ordered by chromosome location (most telomeric at the top). Grey bars to the left represent total expression and black bars to the right represent allelic expression. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



4.5.4 Comparison of *cis*-acting effects between populations

As shown in Figure 4.8, LD was lower in the SA cohort than the Caucasian cohort in the region, as expected for a population of African ancestry. The MAF in the SA population was higher for 33 of the 53 SNPs typed in both populations, which increases the proportion of informative heterozygotes for allelic expression analysis.

Results of aeQTL mapping were compared between the SA and British Caucasian samples. These were highly correlated between the two populations, both for the significance of the detected association ($r=0.94$, $P=10^{-72}$) and the estimated magnitude of the effect on expression for each SNP ($r=0.82$, $P=2x10^{-38}$), as shown in Figure 4.9. These data suggest that the *cis*-acting effects on expression for each mapping SNP were similar in the two populations despite the difference sin LD between them.

Figure 4.8. LD in the SA and Caucasian cohorts.

Figures show linkage disequilibrium between the 56 SNPs in each population: (A) D' in Caucasian cohort; (B) D' in SA cohort; (C) r^2 in Caucasian cohort; (D) r^2 in SA cohort. Colouring in (A) and (B) represents D' values: D'=1, LOD<2 (blue); D'=1, LOD>2 (red); D'<1, LOD>2 (shades of pink); D'<1, LOD<2 (white). Shading in (C) and (D) represents r^2 values: r^2=1 (black); 0< r^2<1 (shades of grey); r^2=0 (white).

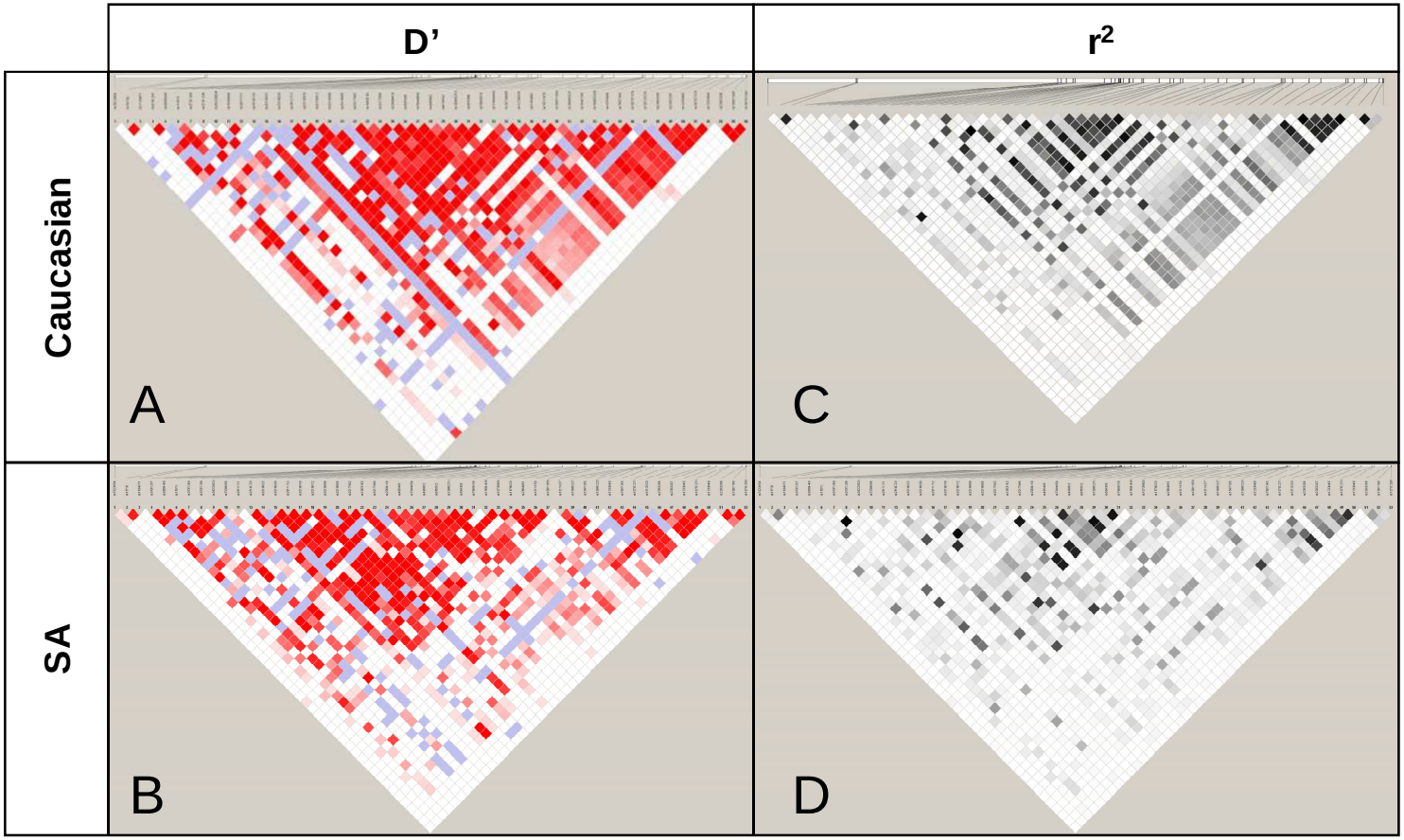
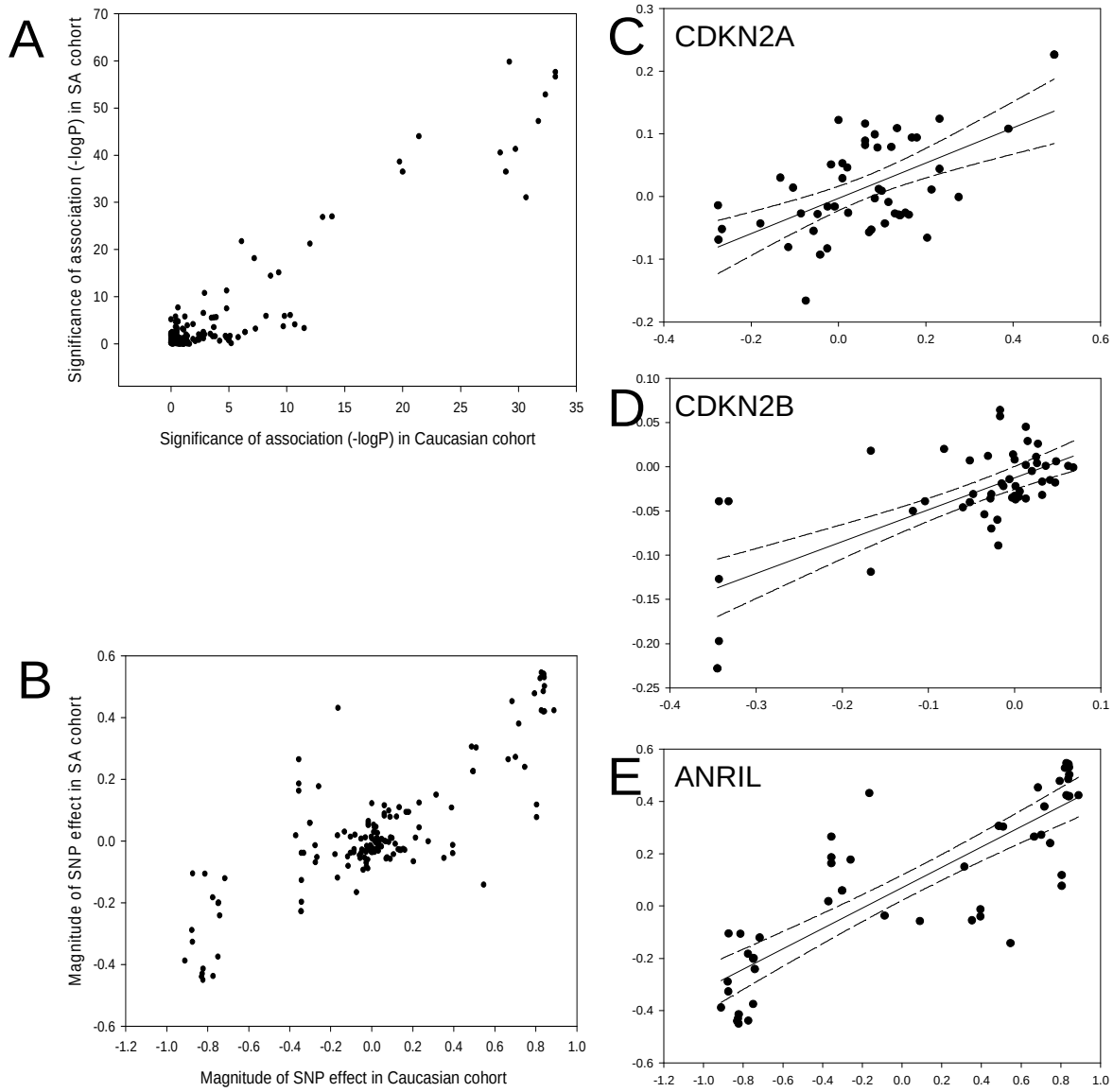


Figure 4.9. SNP effects in the SA and Caucasian cohorts.

Scatter plots show the correlation between aeQTL results obtained in the SA (Y-axis) and Caucasian cohorts (X-axis) for: (A) significance of association with expression (-log P value) for all three genes; (B) effect size at each SNP for all three genes; (C) effect size at each SNP for *CDKN2A* only; (D) effect size at each SNP for *CDKN2B* only; (E) effect size at each SNP for *ANRIL* only. Linear regression line for the association is shown as a solid line with the 95% confidence intervals shown as dotted lines. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



4.5.5 AEI analysis in the combined population

In view of the similarity of the effects in the two cohorts, the data were combined in subsequent analyses, increasing the power to detect *cis*-acting effects of smaller magnitude and enabling adjustment for the effects of individual SNPs. The significance of association and effect estimate for each SNP in the combined cohort are shown in Table 4.5. Associations with a FWER threshold of 0.05 (corresponding to a nominal P-value of 8.9×10^{-4}) were regarded as significant, and nominal P-values are reported in the text. The maximum change in allelic expression associated with any SNP was 1.4-fold for *CDKN2A*, 1.33-fold for *CDKN2B*, and 1.97-fold for *ANRIL*.

The significance of associations for individual SNPs are summarised graphically in Figure 4.10. A greater number of SNPs in the chromosome 9p21 region were significantly associated with expression of *ANRIL* than with *CDKN2A* or *CDKNB*. Furthermore, the significance of the SNP associations was much greater for *ANRIL* than for *CDKN2A* or *CDKN2B*. All risk variants for CAD were highly associated with reduced ANRIL expression, but associations with expression of the other two genes were not consistently seen. Variants associated with diabetes, glioma and melanoma in GWA studies were also all significantly associated with *ANRIL* expression, but not consistently with expression of the other two genes. The effects of SNPs associated with gene regulation *in vitro* and particular diseases are discussed in greater detail in the following sections.

Table 4.5. Effect size and significance of association for all SNPs.

Data shown are for aeQTL mapping in the combined population. Effects are reported as fold changes in expression for individuals who are homozygous for the minor allele relative to individuals who are homozygous for the major allele (calculated from allelic expression data using two transcribed SNPs per gene). Association for each SNP is presented as the $-\log_{10}$ P-value and the $-\log_{10}$ of the FWER using a Bonferroni correction for the 56 SNPs tested. Associations that were significant using a FWER threshold of 0.05 (corresponding to $-\log_{10}$P of 3.05, or $-\log_{10}$FWER of 1.3) were regarded as significant. Significant associations for each gene are shaded grey.

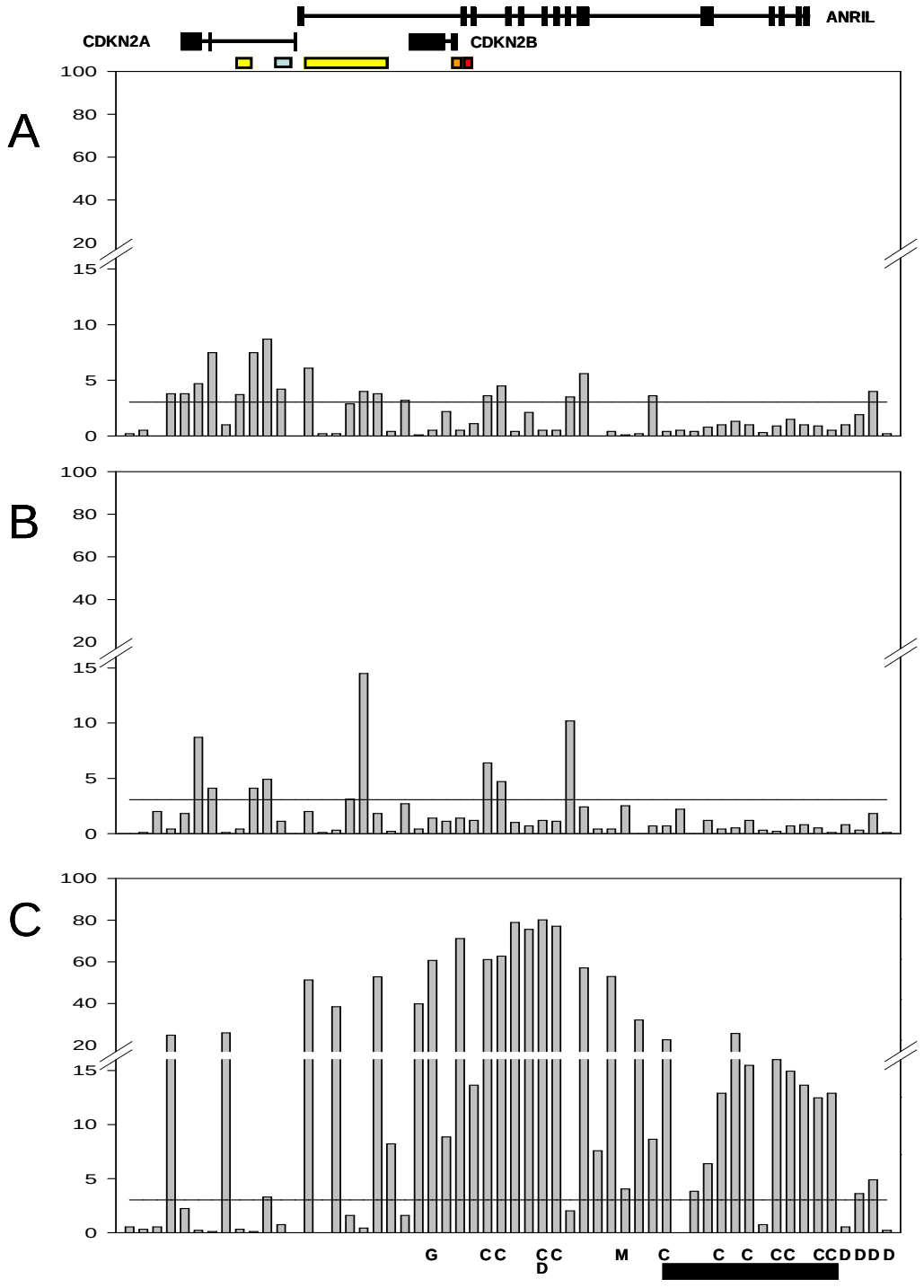
SNP	Promoter	Reported phenotypic associations	Risk allele	Minor allele	*CDKN2A* effect (fold change)	*CDKN2A* -$\log_{10}$P	*CDKN2A* -$\log_{10}$FWER	*CDKN2B* effect (fold change)	*CDKN2B* -$\log_{10}$P	*CDKN2B* -$\log_{10}$FWER	*ANRIL* effect (fold change)	*ANRIL* $-\log_{10}$P	*ANRIL* -$\log_{10}$FWER
rs7023954				A	0.989	0.2	0	1.000	0	0	0.949	0.5	0
rs15735				C	1.026	0.5	0	1.003	0.1	0	1.040	0.3	0
rs1134871				A	0.997	0	0	1.038	2	0.3	1.057	0.5	0
rs3731257		Ovarian ca	G	A	0.883	3.8	2.1	1.011	0.4	0	0.596	24.1	22.4
rs3088440		Melanoma, pancreatic ca, ovarian ca, bladder ca	A	A	0.919	3.8	2.1	1.050	1.8	0.1	0.828	2.1	0.4
rs11515		Alzheimer's, bladder ca, pancreatic ca	C	G	1.084	4.7	3.0	0.880	8.7	7.0	0.968	0.2	0
rs3731249		Breast ca, melanoma, ALL	T	T	1.404	7.5	5.8	0.752	4.1	2.4	0.899	0.1	0
rs3731239		CAD, breast ca	A	C	0.947	1	0	0.997	0.1	0	1.652	25.2	23.5
rs3814960	CDKN2A			T	0.898	3.7	2.0	0.988	0.4	0	1.039	0.3	0
rs36228834	CDKN2A			A	1.404	7.5	5.8	0.752	4.1	2.4	0.899	0.1	0
rs7036656				C	1.119	8.7	7.0	0.931	4.9	3.2	1.217	3.1	1.4
rs2811711	ANRIL			C	1.197	4.2	2.5	0.965	1.1	0	1.099	0.7	0
rs1801022	ANRIL			C	NA	NA	NA	NA	NA	NA	NA	NA	NA
rs2518723	CDKN2A/ARF	Colorectal ca promoter methylation	C	T	1.119	6.1	4.4	0.971	2	0.3	1.669	51.5	49.8
rs3218022	CDKN2A/ARF			C	1.045	0.2	0	0.985	0.1	0	0.995	0	0

rs3218020	CDKN2A/ARF			A	0.989	0.2	0	1.008	0.3	0	0.587	38.2	36.5
rs2811712	CDKN2A/ARF	Frailty, breast ca	A	G	1.079	2.9	1.2	0.938	3.1	1.4	1.155	1.5	0
rs3218018	CDKN2A/ARF	Diabetes	G	G	1.108	4	2.3	0.805	14.5	12.8	1.097	0.4	0
rs3218012	CDKN2A/ARF	Colorectal ca promoter methylation	A	A	0.919	3.8	2.1	1.026	1.8	0.1	0.605	53.1	51.4
rs3218009	CDKN2A/ARF	CAD	G	C	1.059	0.4	0	0.990	0.2	0	1.659	7.7	6.0
rs3218005		Breast ca, frailty	C	C	1.087	3.2	1.5	0.940	2.7	1.0	1.163	1.5	0
rs3217992		CAD	A	A	0.990	0.1	0	1.009	0.4	0	0.579	39.7	38.0
rs1063192		Glioma	C	C	1.036	0.5	0	0.976	1.4	0	1.829	61.3	59.6
rs3217986				C	0.914	2.2	0.5	1.049	1.1	0	0.605	8.3	6.6
rs2069418	CDKN2B			G	1.031	0.5	0	0.976	1.4	0	1.852	72	70.3
rs495490	RD$^{INK4/ARF}$			C	1.131	1.1	0	0.955	1.2	0	1.970	12.8	11.1
rs7044859		CAD, stroke	A	T	1.089	3.6	1.9	0.945	6.4	4.7	1.797	61.7	60.0
rs496892		CAD, stroke	G	A	1.102	4.5	2.8	0.952	4.7	3.0	1.775	63.3	61.6
rs615552				G	1.028	0.4	0	0.980	1	0	1.857	80	78.3
rs10965215				A	0.936	2.1	0.4	1.014	0.7	0	0.599	76.6	74.9
rs564398		Diabetes, CAD, stroke	A	G	1.034	0.5	0	0.978	1.2	0	1.865	81.3	79.6
rs7865618		CAD, stroke	A	G	1.035	0.5	0	0.979	1.1	0	1.870	78.2	76.5
rs17694493				G	1.103	3.5	1.8	0.862	10.2	8.5	1.234	1.9	0.2
rs10738605				C	1.116	5.6	3.9	0.968	2.4	0.7	1.689	57.5	55.8
rs11790231				A	1.006	0	0	0.980	0.4	0	0.618	7.1	5.4
rs2184061				C	1.021	0.4	0	1.010	0.4	0	1.697	53.2	51.5
rs1011970		Melanoma	T	T	0.995	0.1	0	0.953	2.5	0.8	0.802	3.8	2.1
rs10811650				G	0.988	0.2	0	1.000	0	0	0.629	31.7	30.0

rs16905599				A	0.907	3.6	1.9	1.028	0.7	0	0.684	8.1	6.4
rs10116277		CAD, stroke	T	G	0.979	0.4	0	0.984	0.7	0	1.573	21.9	20.2
rs10965227				G	0.966	0.5	0	0.949	2.2	0.5	1.008	0	0
rs1547705				C	1.031	0.4	0	0.999	0	0	0.734	3.6	1.9
rs10965228				G	1.104	0.8	0	0.958	1.2	0	1.504	6	4.3
rs1333040		CAD, stroke	T	C	0.963	1	0	1.012	0.4	0	1.406	12.1	10.4
rs7857345				T	0.935	1.3	0	1.016	0.5	0	1.779	25	23.3
rs10757274		CAD	G	G	1.039	1	0	0.975	1.2	0	0.685	14.5	12.8
rs10125231				A	1.101	0.3	0	0.961	0.3	0	1.280	0.7	0
rs2383206		CAD, stroke	G	A	0.968	0.9	0	1.007	0.2	0	1.456	15	13.3
rs2383207		CAD, stroke	G	A	0.946	1.5	0	0.982	0.7	0	1.468	14	12.3
rs1333045		CAD	C	C	1.037	1	0	0.981	0.8	0	0.700	12.8	11.1
rs10757278		CAD, stroke	G	G	1.039	0.9	0	0.986	0.5	0	0.700	11.7	10.0
rs1333049		CAD	C	C	1.023	0.5	0	0.996	0.1	0	0.704	12.1	10.4
rs2891169		Diabetes	G	A	1.042	1	0	0.977	0.8	0	1.059	0.5	0
rs2383208		Diabetes	G	G	1.081	1.9	0.2	0.989	0.3	0	1.239	3.4	1.7
rs10811661		Diabetes	T	C	1.182	4	2.3	0.957	1.8	0.1	1.339	4.6	2.9
rs10757283		Diabetes	T	T	1.010	0.2	0	1.002	0.1	0	1.024	0.2	0

Figure 4.10. Significance of association with expression for SNPs in the combined population.

The Y-axis represents the –log P value for individual SNPs (shown in chromosomal order along the X-axis) for: *CDKN2A* (A); *CDKN2B* (B); *ANRIL* (C). The horizontal black line on each graph represents the significance threshold after adjustment for multiple testing (FWER=0.05 corresponding to -$\log_{10}$P=3.05). The relative location of genes and promoter elements is represented at the top (*CDKN2A* and *CDKN2A*/ARF promoters yellow; *ANRIL* promoter blue; *CDKN2B* promoter orange; *CDKN2A*/ARF regulatory domain red). Letters along the bottom represent associations from GWA studies (C=CAD, D=diabetes, M=melanoma, G=glioma) and the black bar at the bottom represents the core risk haplotype for CAD defined by Broadbent *et al* [129].



4.5.6 Adjusting for the effects of individual SNPs

As shown in Figure 4.10, multiple SNPs were associated with *cis*-acting influences on expression of *CDKN2A*, *CDKN2B* and *ANRIL*. This could be the result of multiple independent loci influencing expression of each gene, but could also be a reflection of strong LD in the region since associations might be observed for 'non-functional' SNPs (that do not directly influence expression) which are in LD with other 'functional' polymorphisms. Adjusting for the effect of individual SNPs was used to assess whether multiple SNPs were independently correlated with expression of the three genes, as shown in Figure 4.11. For each gene stepwise adjustments were made for the effect of the SNP which showed the most significant association with expression, until independent effects could no longer be detected. Associations remained significant after adjusting for the top SNP for *CDKN2A* and *CDKN2B*, and the top two SNPs for *ANRIL*.

These results indicate that even after adjusting for the effects of the most significant marker, some of the remaining SNPs still showed significant association with *ANRIL* expression. This could be explained by the presence of more than one functional polymorphism affecting expression, but could also reflect the presence of a functional polymorphism that is in LD with both markers. However, examination of the allelic expression patterns provides additional support for the presence of multiple sites affecting expression. For example, Figure 4.12 shows the allelic expression ratios observed at the transcribed SNP rs564398 in *ANRIL*, grouped according to the genotype at rs10965215. These two SNPs are in strong LD (D'=0.98), hence the absence of individuals homozygous for the A allele at rs10965215 that are heterozygous at rs564398. The G allele of the transcribed SNP (rs564398) was overexpressed (G/A AER values greater than 1), however overexpression was stronger ($P=10^{-15}$ using the Mann-Whitney test) for individuals that were also heterozygous at the second polymorphism (rs10965215). This pattern is not consistent with allelic expression being determined by a single biallelic polymorphism acting in *cis* and suggests that there is more than one functional polymorphism or that this polymorphism is multiallelic.

Figure 4.11. Effect of sequential adjustment for most highly associated SNPs.

The Y-axis represents the –log P value for individual SNPs (shown in chromosomal order along the X-axis). (A) Unadjusted. (B) Adjusted for the most highly associated SNP for each gene (*CDKN2A* rs7036656, *CDKN2B* rs3218018, *ANRIL* rs564398). (C) Adjusted for the two most highly associated SNPs for each gene (*CDKN2A* rs7036656 and rs36228834, *CDKN2B* rs3218018 and rs3814960, *ANRIL* rs564398 and rs10965215). Values in the top right corner are the number of significantly associated SNPs after each round of adjustment, following correction for multiple testing. The remaining SNP showing association with *ANRIL* expression was rs495490. The horizontal black line on each graph represents the significance threshold after adjustment for multiple testing (FWER=0.05 corresponding to $-\log_{10}P=3.05$).

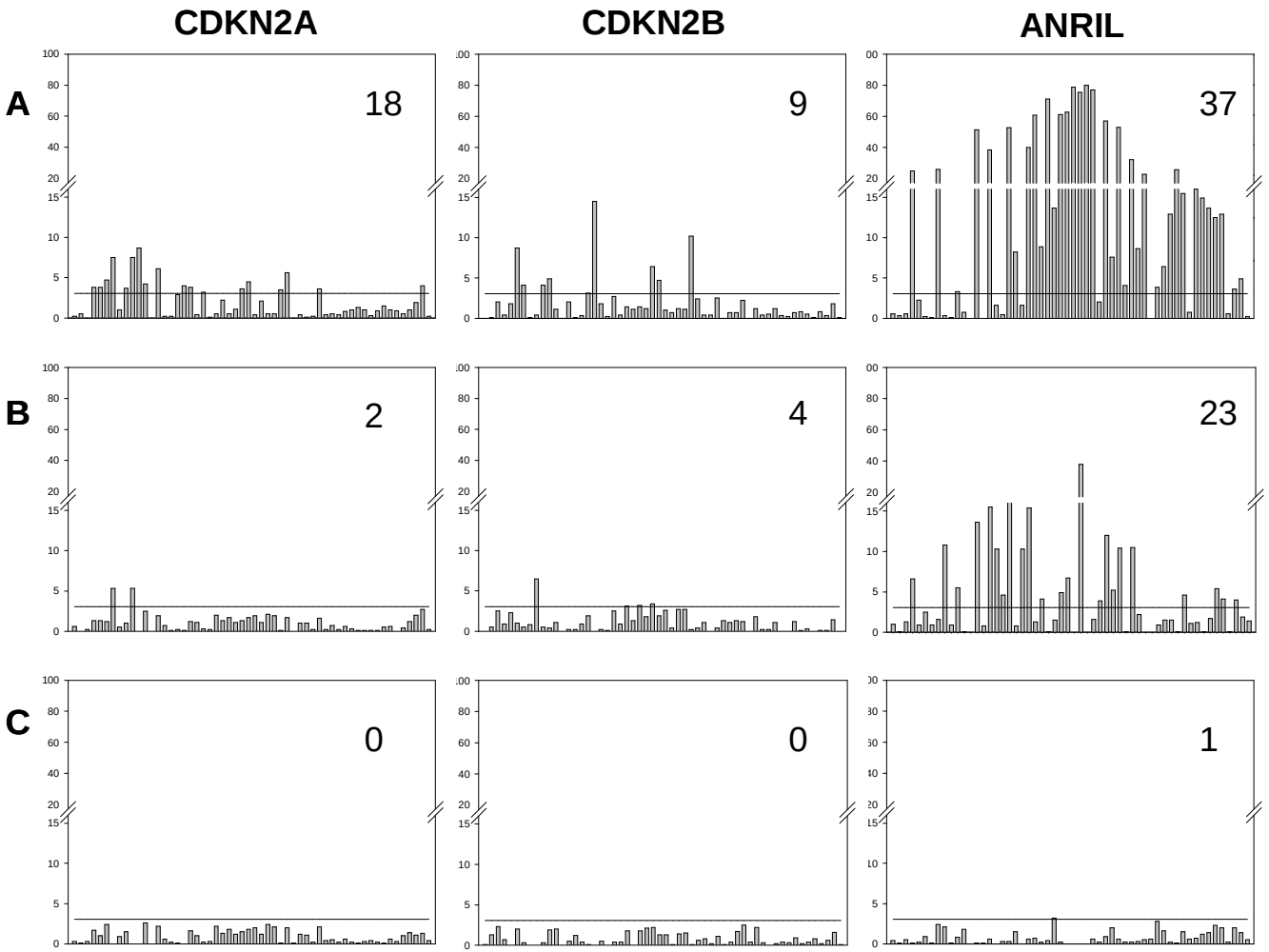
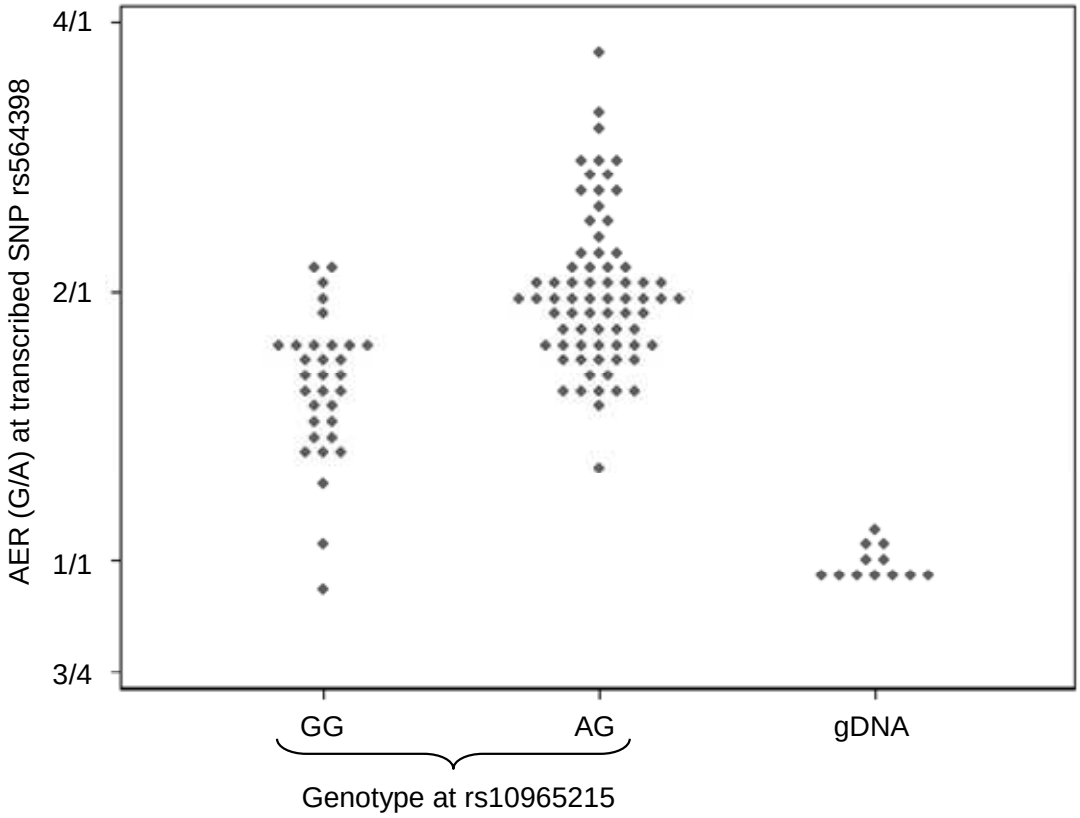


Figure 4.12. Effect of genotype at rs10965215 on allelic expression ratio of transcribed *ANRIL* SNP rs564398.

Diamonds represent the allelic expression ratio for each individual, all of whom are heterozygous for the transcribed SNP rs564398. The first column shows individuals who are homozygous for rs10965215 (mean ratio 1.57), and the second column shows individuals who are heterozygous for rs10965215 (mean ratio 2.00). The third column shows the expression ratio obtained from genomic DNA in individuals who are heterozygous for the transcribed SNP rs564398, where the two alleles are present in a 1:1 ratio (mean ratio 1.00).



The direction of *cis*-acting effects on expression was compared between genes for SNPs showing significant associations with expression of each gene, as shown in Table 4.6**.** SNP effects for *CDKN2A* and *ANRIL* were in the same direction for all 10 SNPs, meaning that alleles associated with overexpression of *CDKN2A* were also associated with overexpression of *ANRIL*. By contrast, for all 8 SNPs that were significantly associated with allelic expression of both *CDKN2A* and *CDKN2B*, the alleles associated with *CDKN2A* overexpression were associated with *CDKN2B* underexpression. Similarly for all 3 SNPs significantly associated with allelic expression of both *CDKN2B* and *ANRIL*, alleles associated with overexpression of *CDKN2B* were associated with *ANRIL* underexpression. The total expression analysis had insufficient power for similar analyses to be performed.

Table 4.6. Correlation of SNP effects between genes by aeQTL mapping.

The table shows the SNP effect between genes for SNPs that show significant association (using FWER threshold of 0.05) with expression of both genes. Gene pairs are shown along the top. SNP effects in the same direction means that a SNP is associated with overexpression or underexpression of both genes, whereas SNP effects in the opposite direction means that a SNP associated with underexpression of one gene is associated with overexpression of the other gene.

	CDKN2A-ANRIL	***CDKN2B-ANRIL***	***CDKN2A-CDKN2B***
SNP effects same direction	10	0	0
SNP effects opposite	0	3	8

4.5.7 *In vivo* effects of putative regulatory elements identified *in vitro*

Whether SNPs within regulatory regions previously identified by *in vitro* reporter assays were associated with *cis*-acting effects on expression *in vivo* was also investigated (full data contained in Table 4.5).

CDKN2A expression was significantly correlated with SNPs in its promoter and the *ARF* transcript promoter[186, 187, 319, 320], and with SNPs close to the regulatory domain ($RD^{INK4/ARF}$) that has been shown to regulate expression of *CDKN2A*, *ARF* and *CDKN2B in vitro*[189].

CDKN2B expression was also significantly correlated with SNPs in the *CDKN2A* and *ARF* promoter regions, suggesting that these elements influence expression of both genes. *CDKN2B* expression was not significantly correlated with the single SNP typed in its promoter (rs2069418) prior to adjustment, but this became significant after adjustment for the most significant SNP in the *ARF* promoter (rs3218018).

ANRIL expression was strongly associated with SNPs in the *CDKN2B* promoter ($P=10^{-72}$), *ARF* promoter (P up to10^{-53}) and $RD^{INK4/ARF}$ domain ($P=10^{-12}$), as well as with SNPs adjacent to the *CDKN2A* promoter (rs3731239, $P=10^{-25}$).

These data confirm *in vivo* the function of the regulatory elements identified by *in vitro* transfection studies, and provide further evidence that shared *cis*-acting elements influence expression of *CDKN2A*, *CDKN2B* and *ANRIL*.

4.5.8 Effects of disease associated SNPs on expression

The study also examined the correlation of allelic expression of *CDKN2A*, *CDKN2B* and *ANRIL* with SNPs reported to confer disease susceptibility (full data contained in Table 4.5).

4.5.8.1 CAD and stroke

SNPs within the core risk haplotype region for CAD[129] were associated with *ANRIL* expression (P up to 10^{-21}), but none were associated with *CDKN2A* or *CDKN2B*

expression. CAD risk alleles were all associated with reduced *ANRIL* expression, up to 1.9-fold, suggesting that expression of *ANRIL*, rather than *CDKN2A* or *CDKN2B*, might mediate atherosclerosis susceptibility. However, other CAD risk variants located telomeric to the core risk haplotype region such as rs7044859 and rs496892 showed substantially larger effects and stronger associations with *ANRIL* expression ($P<10^{-60}$ for each SNP), and were also significantly associated with *CDKN2A* and *CDKN2B* expression ($P<10^{-4}$ for each SNP). The CAD risk alleles at these SNPs correlated with reduced expression of *ANRIL* and *CDKN2A*, but increased *CDKN2B* expression. Associations for these SNPs remained significant after adjusting for the effect of the lead CAD SNPs within the core risk haplotype region (rs10757274, rs2383206, rs10757278 and rs1333049)[129], but SNPs within the core risk haplotype were no longer significantly associated with *ANRIL* expression after adjusting for the effect of SNPs at the distal locus (rs10965215 and rs564398). This suggests that the core CAD risk haplotype does not account for all of the observed association with *ANRIL* expression in peripheral blood.

Based on evolutionary conservation and effects on *ANRIL* transcription, rs1333045 within the core risk haplotype has been previously highlighted as a potential functional variant responsible for conferring susceptibility to CAD at the 9p21 locus[200]. In our analysis rs1333045 was associated with *ANRIL* expression ($P=10^{-12}$), but not with *CDKN2A* or *CDKN2B* expression. Its effects were similar to those of other SNPs in the core risk haplotype for CAD. After adjusting for the effect of rs1333045, 32 SNPs remained significantly associated with *ANRIL* expression, suggesting that the effect attributed to such variants was not due to LD with rs1333045.

4.5.8.2 Diabetes

The lead chromosome 9p21 SNPs associated with diabetes in GWA studies are located in a separate LD block to the CAD risk variants[125, 130], and the phenotypic effects of CAD and diabetes variants have been shown to be independent[129]. Diabetes risk alleles in this region (rs10811661-T and rs2383208-A) were associated with under-expression of *ANRIL*, but were not associated with *CDKN2A* or *CDKN2B* expression in our Caucasian population. However, these SNPs showed no association

with expression of *ANRIL* in the SA population, despite greater power to detect effects in this cohort.

A separate locus for diabetes susceptibility in the chromosome 9p21 region in Caucasians is located within the region associated with CAD risk. The rs564398-T risk allele at this locus is associated with diabetes[124], CAD[129] and stroke[159]. This SNP had the strongest association with *ANRIL* expression of all the SNPs we tested ($P=10^{-81}$), but was not significantly associated with *CDKN2A* or *CDKN2B* expression. The rs564398-T risk allele was associated with *ANRIL* underexpression, and the association remained significant after adjusting for the effect of rs10811661, the lead diabetes SNP. However, the association with rs10811661 was no longer significant after adjustment for rs564398.

4.5.8.3 Cancers and frailty

GWA studies have recently identified chromosome 9p21 SNPs correlated with susceptibility for glioma[164, 165] and malignant melanoma[167]. The glioma risk allele rs1063192-C was highly correlated with increased *ANRIL* expression ($P=10^{-61}$), while the melanoma risk variant rs1011970-T correlated with reduced expression of *ANRIL*. Neither was associated with *CDKN2A or CDKN2B* expression.

Multiple candidate gene association studies have reported associations between SNPs in this region and susceptibility to a variety of diseases. These have mostly involved cancer phenotypes because the cell-cycle regulators *CDKN2A* and *CDKN2B* are recognised to be involved in predisposition to certain cancers. Such association studies have implicated 9p21 SNPs as being potentially involved in the development or therapeutic response to pancreatic[175, 314], breast[173, 174, 180], ovarian[176, 181], and bladder[171] carcinoma, as well as acute lymphoblastic leukaemia[177], and melanoma[169, 170, 172]. All of the SNPs associated with these phenotypes showed a significant correlation with allelic expression of one or more of the genes we examined, as summarised in Table 4.5. A SNP (rs2811712) that was reported to be associated with severely limited physical function in older people[179] was significantly associated with *CDKN2B* expression, but not with *ANRIL* expression.

4.5.9 Preliminary investigation of expression of exons involved in other transcripts

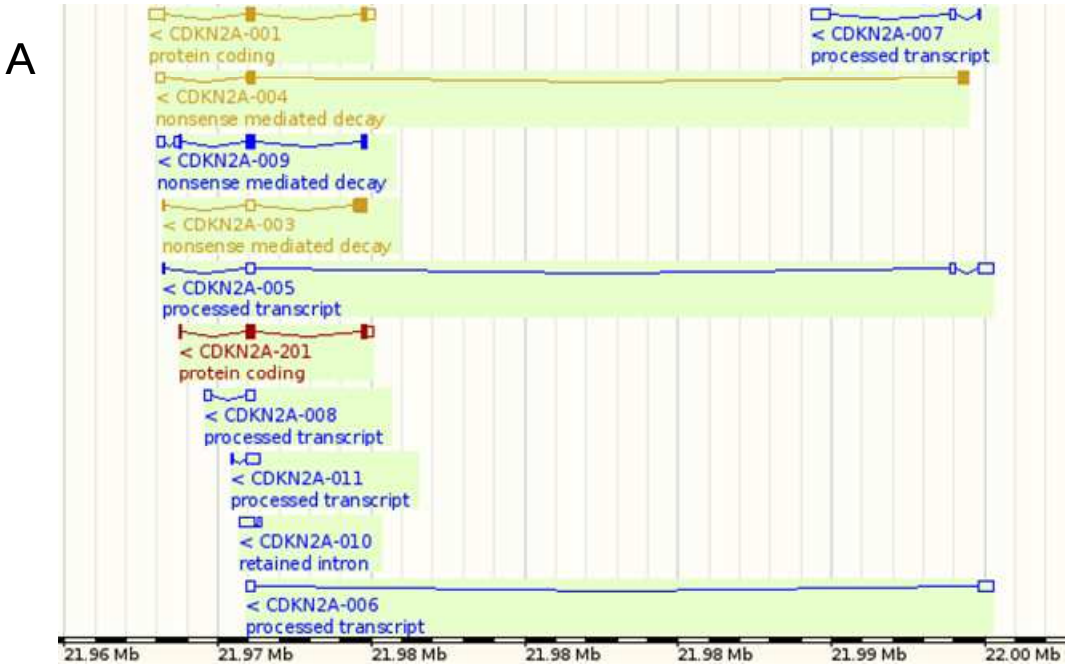
4.5.9.1 Exon-specific AEI analysis

This study explored whether AEI could be used to investigate transcript-specific expression and mapping of *cis*-acting effects for *CDKN2A* and *ANRIL*.

Multiple transcript variants are now annotated for *CDKN2A*[43], two of which are known to produce the functionally-important proteins CDKN2A and CDKN2A-ARF. These transcripts and the location of SNPs with heterozygosity greater than 0.2 that are suitable for AEI assessment are illustrated in Figure 4.13. Several transcript-specific assays were investigated; the primer locations are illustrated in Figure 4.13 and the PCR product sizes are shown in Table 4.7. A limitation of the AEI approach for investigation of expression of particular exons is that it requires a transcribed SNP of reasonable heterozygosity within the amplicon. There were no such SNPs within ARF exon 1β, meaning that transcript-specific amplicons were designed to include a primer in exon 1β and a primer downstream of the transcribed SNPs in exon 3. As shown in Table 4.7, PCR products could not be reliably detected for these long products as initially designed (*CDKN2A*-specific-1 and *ARF*-specific-1), despite using 45 cycles and extensive attempts to optimise the PCR conditions (adjusting temperatures, magnesium concentrations, and adding Q solution). The assays were therefore redesigned to give the shortest transcript-specific products possible (*ARF*-specific-2, *CDKN2A*-specific-2 and *CDKN2A*-specific-3). After optimisation, PCR product was detectable for some samples by agarose gel electrophoresis, but was still not reliably detectable from many samples for the longer amplicons (*ARF*-specific-2 and *CDKN2A*-specific-2). In contrast, the shorter products for the *CDKN2A*-specific-3 and *CDKN2A&ARF* assays were reliably detected for most samples.

Figure 4.13. *CDKN2A* transcripts and transcribed SNPs.

Panel A shows the 11 *CDKN2A* transcript variants annotated in Ensembl 01/03/10[43], with the chromosome 9 reference sequence loction shown on the X-axis. Panel B shows the two protein-coding transcripts. SNPs suitable for AEI analysis are illustrated as black bars with the SNP ID in bold above. Black arrows indicate the location of PCR primer pairs for AEI assays.



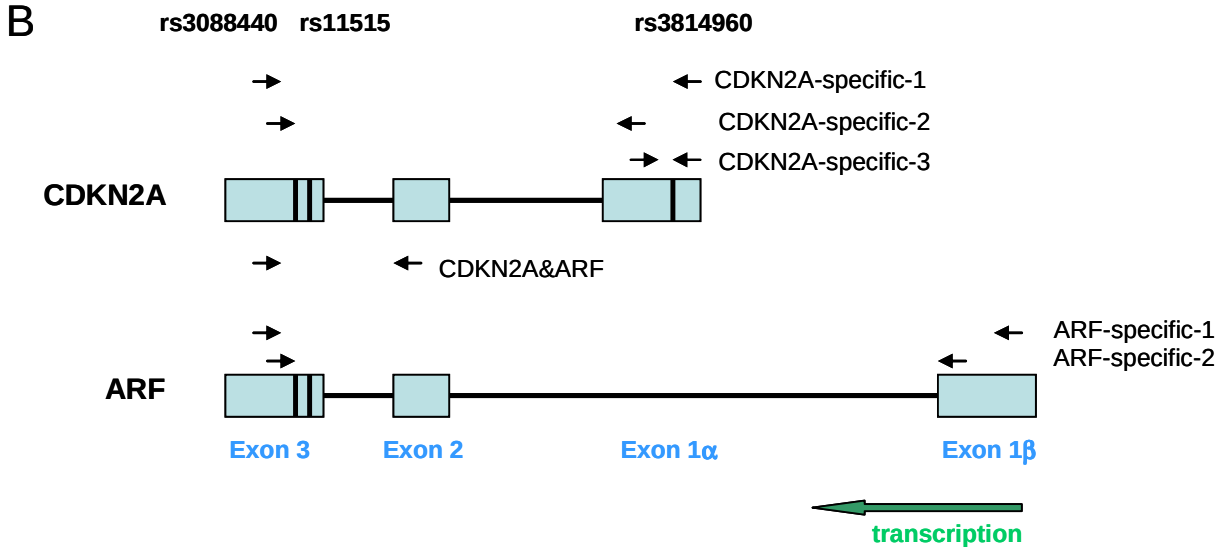


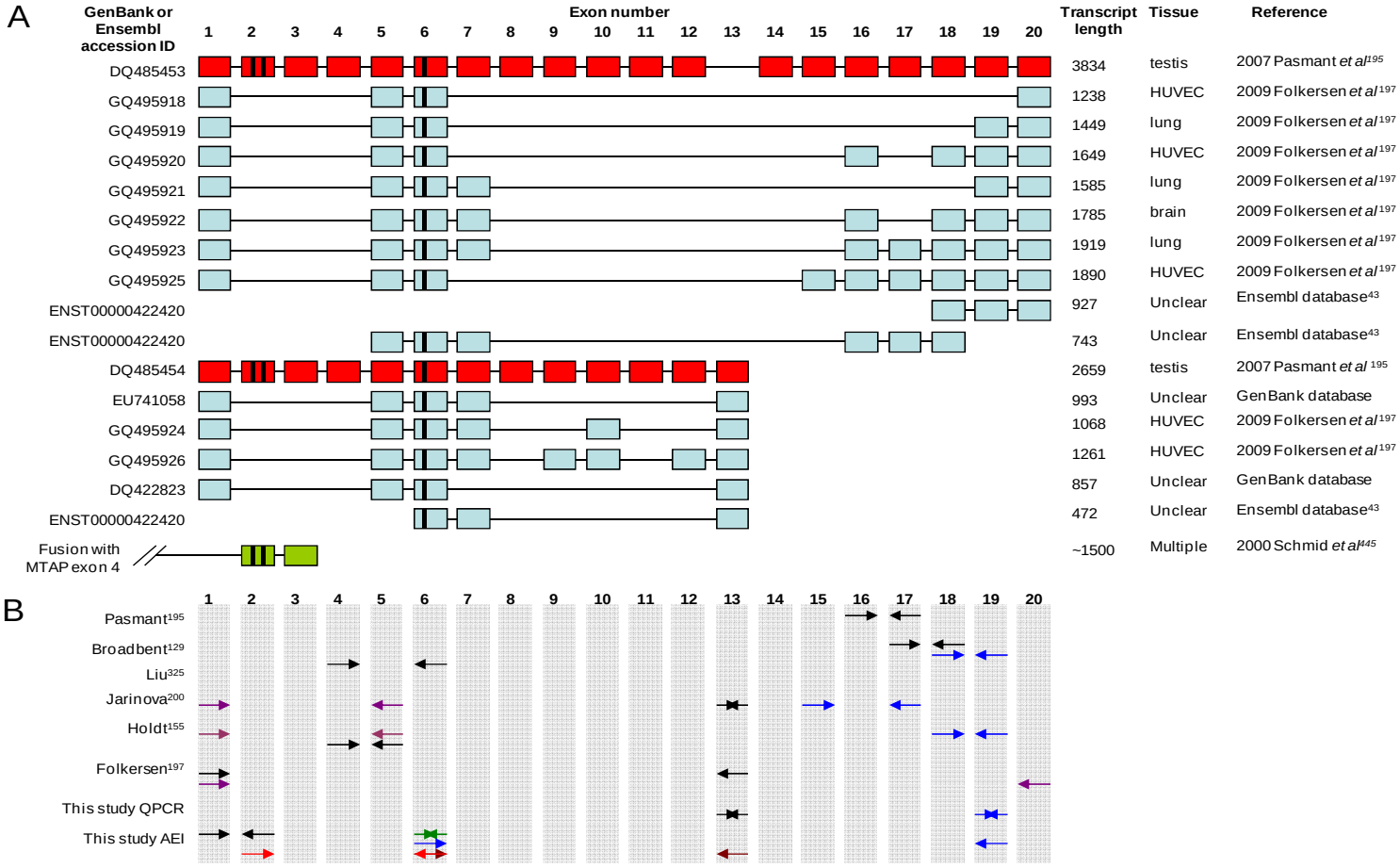
Table 4.7. Transcript-specific AEI assays for *CDKN2A* and *ANRIL*.

Primer pairs	Expected product size	Optimised annealing temperature (°C)	Comment
CDKN2A-specific-1	861	-	PCR product not reliably detected.
CDKN2A-specific-2	479	58.4	PCR product not reliably detected.
CDKN2A-specific-3	73	61.8	Satisfactory PCR on agarose gel.
CDKN2A&ARF	150	58.4	Satisfactory PCR on agarose gel.
ARF-specific-1	635	-	PCR product not reliably detected.
ARF-specific-2	407	66.8	PCR product not reliably detected.
ANRILexons1-2	150	58.4	Satisfactory PCR on agarose gel.
ANRILexons6-19	154, 161, 290, 354, 490, 624, 2058	-	Expected PCR products not detected in 2 tests.
ANRILexons6-13	167, 303, 378, 571	55.5	167 bp and 303 bp products faintly detected in 2 tests.
ANRILexons2-6	314	64.6	314 bp product detected in 2 tests.
MTAP4-ANRIL2	290	-	PCR product not detected in 2 tests.
ANRILexons6-6	49	56.0	Satisfactory PCR on agarose gel.

A recent report suggested that *ANRIL* has more splice variants than the two that were originally described by Pasmant *et al*[195]. The *ANRIL* transcripts annotated in the NCBI and Ensembl databases (as of 01/03/10) are shown in Figure 4.14[43, 196], with the two original isoforms highlighted in red (both of which contain exons 1-2, in which the primers for the AEI assay used in the studies reported above were located). The transcripts reported to date broadly fall into two groups; a short form ending with exon 13, and a long form ending with exon 20[43, 196]. However, this may be related to the methodology used to investigate alternative transcripts by Folkersen *et al*, which only used primers located in exon 13 and exon 20[197].

Figure 4.14. ANRIL transcripts.

Panel A shows reported transcripts with transcribed SNPs indicated as black bars (rs10965215 and rs564398 in exon 2 and rs10738605 in exon 6). The transcripts described in the original report by Pasmant *et al*[195] in human testis are shown in red. Panel B shows the locations of primers used in expression studies.



Some recent reports suggested that short and long transcripts may be differentially expressed[155, 200], and the study therefore aimed to investigate the *cis*-acting effects in assays involving different *ANRIL* exons. Transcribed SNPs with heterozygosity greater than 0.2 which may be suitable for AEI analysis were located in exon 2 and exon 6. AEI assays were designed to investigate expression of the 'long transcript' (with primers in exon 6 and exon 19) and the 'short transcript' (with primers in exon 6 and exon 13), as summarised in Table 4.7 and Figure 4.14. Another assay was designed with primers in exon 2 and exon 6 aiming to assess the correlation between AER measured at rs10965215/rs564398 (in exon 2) and rs10738605 (in exon 6) in the same transcript. An *MTAP-ANRIL* fusion transcript has also been reported[166], and a further assay was designed to determine whether expression of this transcript could be detected in blood and mapped using AEI (using primers in *MTAP* exon 4 and *ANRIL* exon 2).

No evidence was detected for the *MTAP-ANRIL* product on gel electrophoresis. Products of the expected size were also not detected for the 'ANRILexons6-19' assay. Some faint bands of other sizes were seen at low temperatures during temperature gradient optimisations, but these were not consistent between different tests and were likely to represent non-specific primer binding. PCR products at some of the expected sizes were seen for the other assays, as summarised in Table 4.7. These assays require replication in larger numbers of samples using the optimised reaction conditions shown in the Table 4.7. If these assays perform reliably in larger numbers of samples then they may be utilised for aeQTL mapping of different transcripts.

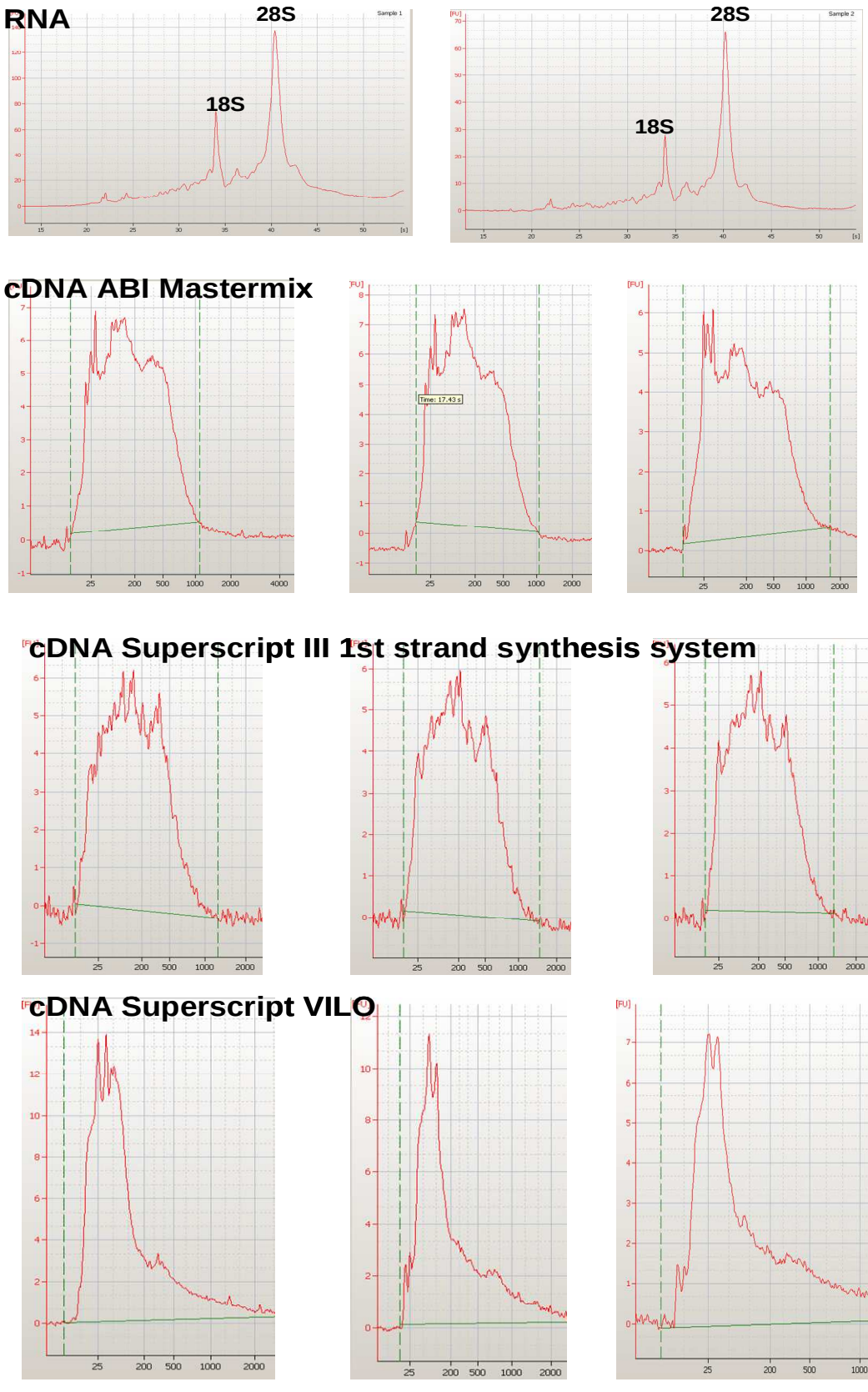
It was noted for both the *CDKN2A* and *ANRIL* assays discussed above that assays involving longer PCR products were not reliably detected in most samples, whereas the shorter amplicons showed less variability between samples. One explanation considered for the failure of assays involving longer products was that RNA/cDNA integrity might be insufficient to allow the identification of complete longer transcripts, especially since the samples had been stored for a considerable time and been through several freeze-thaw cycles. RNA quality and cDNA size were therefore checked for a selection of samples using the RNA 6000 Pico LabChip kit (Agilent) for the Agilent 2100 bioanalyzer.

As shown in Figure 4.15, this confirmed high RNA quality with a 28S:18S ratio greater than two. However, cDNA fragments were predominantly between 25 and 500 bases in length, which is likely to account for the failure of assays with products above that size. On review of the study methods, it was noted that the High capacity RNA-to-cDNA Master Mix (Applied Biosystems) that was used for reverse transcription of the samples in the above analyses is optimised for cDNA targets less than 1000 bases in length. Other RT systems were investigated to see whether longer cDNA could be obtained, including the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen) which is advertised as permitting detection of RNA targets more than 12kb in size[324]. As shown in Figure 4.15, cDNA fragment size was not substantially greater with the different systems, suggesting that analysis of the longer assays was not possible in these samples.

The two assays using previously untested transcribed SNPs which had shown reliable performance on agarose gel electrophoresis were tested for AEI analysis: 'CDKN2A-specific-3' (using transcribed SNP rs3814960) and 'ANRILexons6-6' (using transcribed SNP rs10738605). The 'ANRILexons6-6' assay worked well in both gDNA and cDNA. The 'CDKN2A-specific-3' assay worked well in gDNA and some cDNA samples, but had a higher measurement variability and proportion of failed replicates in cDNA samples than other AEI assays. This may be accounted for by low expression levels of this transcript variant in peripheral blood, suggesting that higher starting amounts of cDNA should be used for analysis using this assay.

Figure 4.15. Assessment of RNA integrity and cDNA size.

Images show the output of obtained with the Agilent 2100 bioanalyzer for RNA samples and cDNA samples reverse transcribed with different RT systems. Y-axis represents fluorescence. X-axis on the RNA figures represents time in seconds, and on the cDNA figures represents cDNA fragment size in nucleotides. The 18S and 28S ribosomal RNA peaks are labelled on the RNA samples.



4.5.9.2 Exon-specific total expression analysis

Exon-specific expression was also investigated using analysis of total expression levels in the SA cohort. This was simpler since there was no requirement for amplicons to include a transcribed polymorphism and therefore large products were not necessary. A suitable cDNA-specific predesigned assay was available for *ARF* (Applied Biosystems assay ID Hs00924091_m1), but not for *CDKN2A* or *ANRIL*. For *CDKN2A*, the assay previously reported by Liu *et al* was tested[325], but replicates did not perform consistently. On investigation of the assay design, the difference between primers and probe annealing temperatures was greater than recommended for such assays and a new custom assay was therefore designed with primers and probes located within exon 1α. For *ANRIL*, custom assays were designed within exon 13 for the 'short' transcript and within exon 20 for the 'long' transcript (as shown in Figure 4.14, page 159). These assays were multiplexed with *B2M* and *GAPDH* reference genes respectively.

Transcript-specific total expression levels of *CDKN2A* and *ARF* were highly correlated with each other ($r=0.63$, $P=7x10^{-61}$) and with expression assessed using the combined *CDKN2A-ARF* assay (using *B2M* and *GAPDH* as reference genes), as shown in Figure 4.16. Furthermore, SNP effects for *CDKN2A* and *ARF* mapped by eQTL analysis were also correlated ($r=0.47$, $P=0.0003$), as shown in Figure 4.17. This suggests that the loci influencing expression in *cis* are similar for these transcripts. Associations for few individual SNP effects achieved statistical significance, reflecting the relatively low power of this analysis compared to the combined AEI analysis.

Figure 4.16. Correlations between total expression levels of *CDKN2A*, *ARF* and *CDKN2A/ARF*.

Axes show expression Ct values normalised to *B2M/GAPDH*. Each point represents an individual. The regression line is shown in solid black, with dotted 95% confidence intervals. Pearson correlation values are shown.

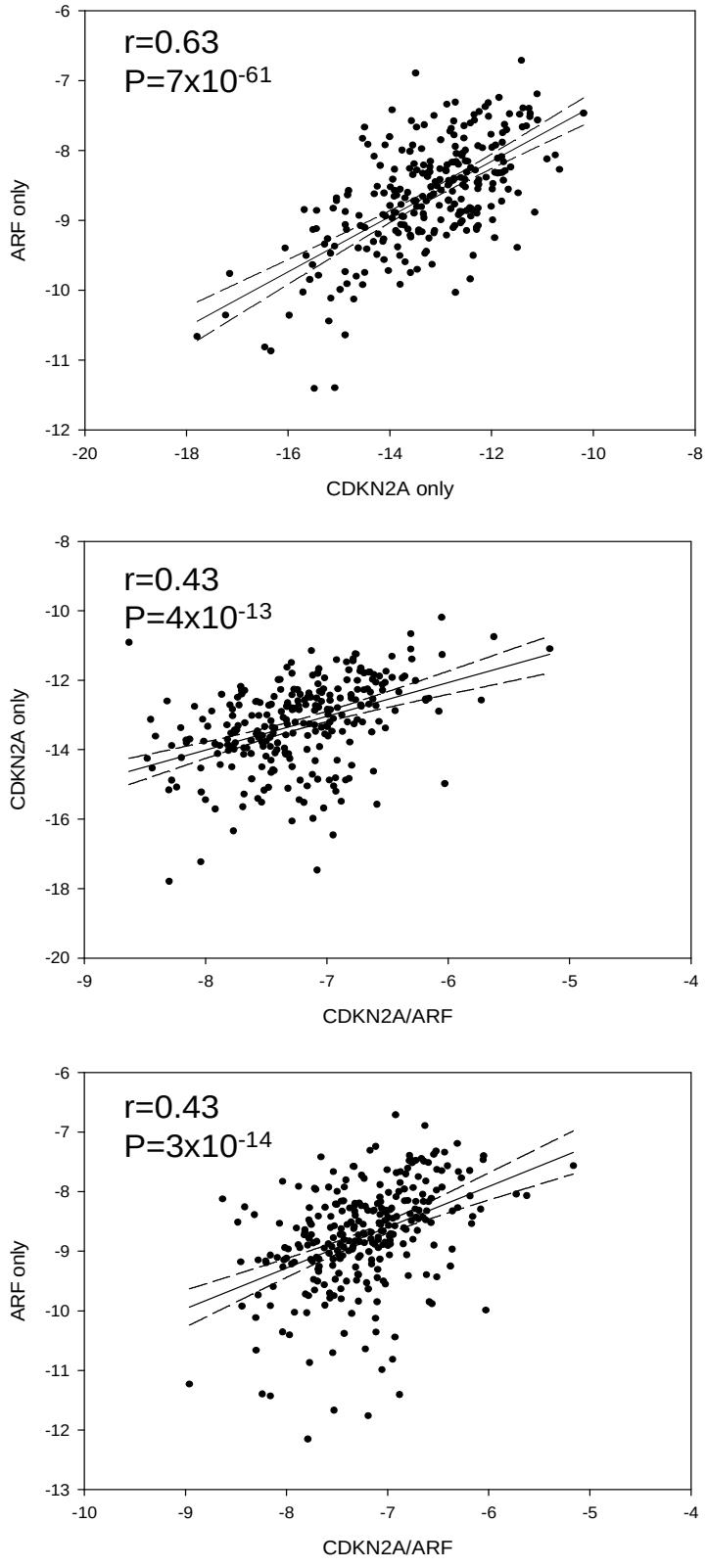
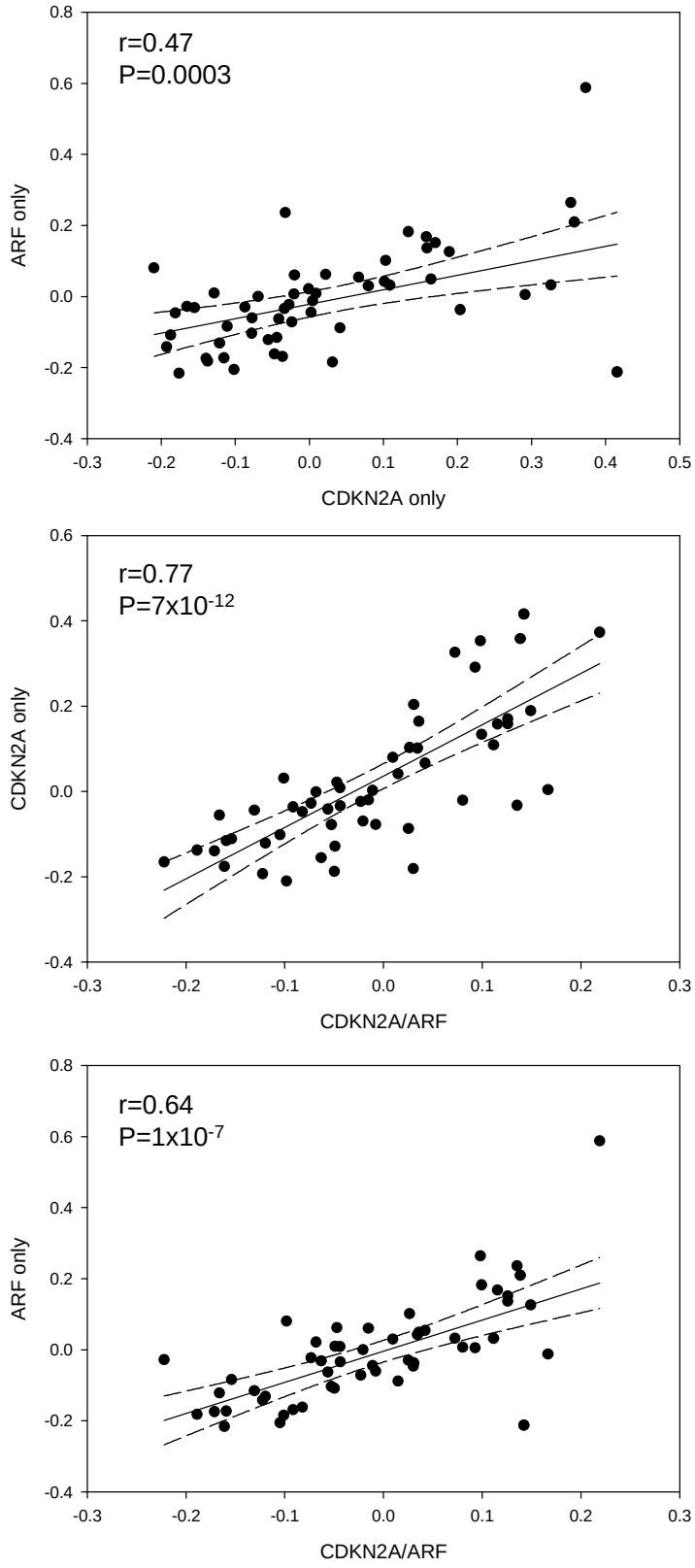


Figure 4.17. SNP effects estimated using eQTL analysis with assays specific for CDKN2A, ARF and CDKN2A/ARF.

Axes show *cis*-acting effect for the 56 mapping SNPs, each of which is represented as a point. The regression line is shown in solid black, with dotted 95% confidence intervals. Pearson correlation values are shown, but results were similar using Spearman's rank correlation. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



Preliminary investigation of transcript-specific expression of *ANRIL* was also performed. As shown in Figure 4.18, transcript-specific total expression of *ANRIL* exon 13 and exon 20 showed a weak but significant correlation. Total expression at exon 13 was weakly correlated with total expression at exons 1-2 as used previously, but there was no significant correlation between expression at exon 20 and exons 1-2. Interestingly, SNP effects for exon 13 and exon 20 mapped by eQTL analysis were positively correlated, but both showed a significant negative correlation with SNP effects measured at exons 1-2, as shown in Figure 4.19. These data suggest that expression of the exon 13 and exon 20 transcripts are co-regulated and influenced by the same *cis*-acting effects. However, the *cis*-acting influences on expression of the exons 1-2 transcript act in the opposite directions, and the overall expression levels of this transcript relate poorly to expression of the other two transcripts studied. Expression levels at exon 13 and exon 20 were higher than at the exons 1-2 when identical reference genes were used for normalisation, which suggests that the exon 13 and exon 20 assays do not reflect the levels of the two originally reported transcript variants (as shown in red in Figure 4.14), both of which contain exons 1-2 meaning that expression measured at this 'summed' assay would be expected to be higher than that of the individual transcript variants. The significance of *cis*-acting effects on expression was substantially less for the eQTL analysis using the exon 13 and exon 20 assays, with associations achieving significance for only a single SNP. This could mean that the transcripts assessed using these assays have less influence of *cis*-acting effects, or it could result from poor precision of the measurements and 'background noise' of the assay. These results therefore require additional investigation with the same and independent assays and must be regarded as preliminary at present.

Figure 4.18. Correlations between total expression levels of different *ANRIL* transcripts.

Axes show expression Ct values normalised to *B2M/GAPDH*. The regression line is shown in solid black, with dotted 95% confidence intervals. Pearson correlation values are shown.

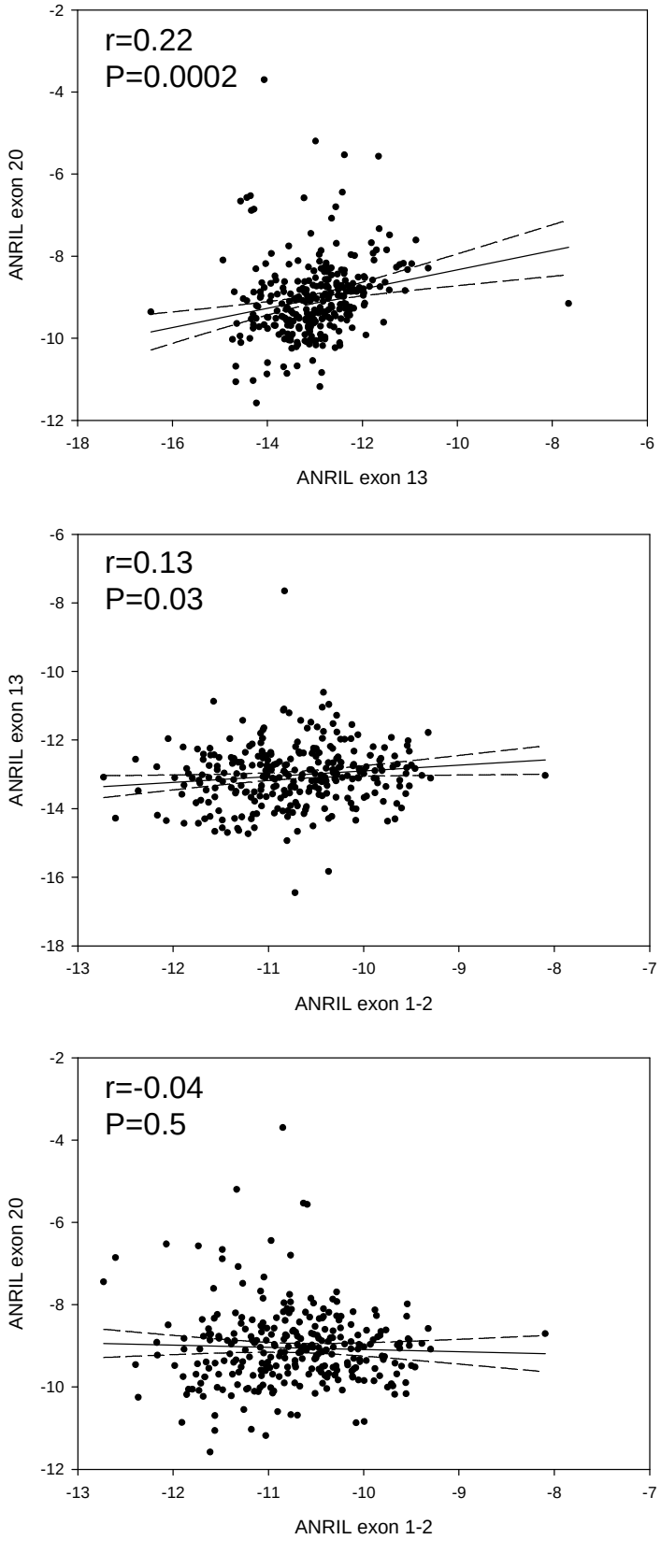
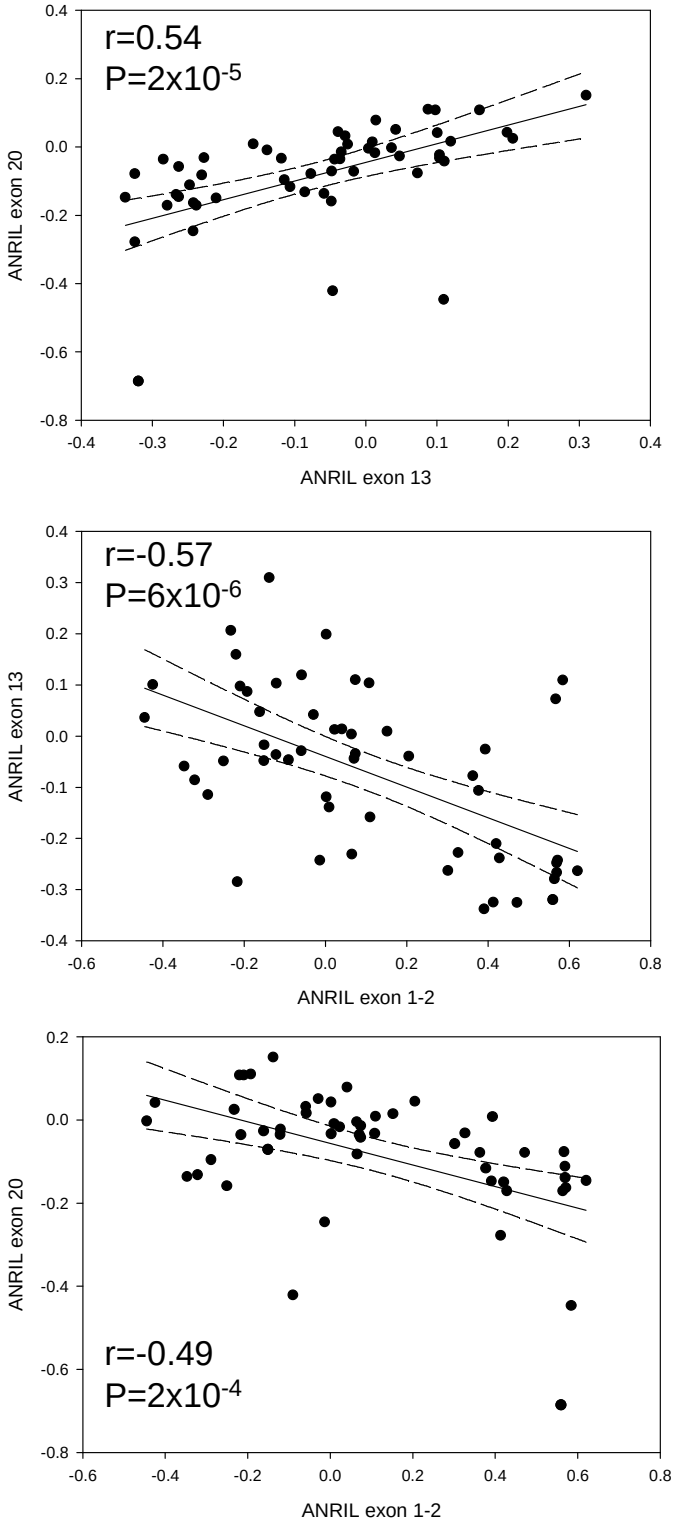


Figure 4.19. SNP effects estimated using eQTL analysis for different *ANRIL* transcripts.

Axes show *cis*-acting effect for the 56 mapping SNPs. The regression line is shown in solid black, with dotted 95% confidence intervals. Pearson correlation values are shown, but results were similar using Spearman's rank correlation. SNP effect is the the log of the change in expression that allele 2 at the *cis*-acting locus produces compared to allele 1 (as defined on p87).



4.5.10 Microsatellite rs10583774 effects on expression

A comparison of microsatellite analyses using WGA DNA and native gDNA showed that spectra were more difficult to interpret in WGA DNA for some samples as the pattern of the 'slippage' peaks was not as predictable as it was in native genomic DNA, as illustrated in Figure 4.20. In unamplified DNA, the true peak was highest, with smaller 'slippage' peaks at two-nucleotide intervals of progressively decreasing size. In amplified DNA, there was less difference in size between the true peak and nearest 'slippage' peak which made spectra more difficult to interpret for individuals who were heterozygous for similarly sized alleles. Analyses were therefore performed using unamplified genomic DNA for all individuals in both cohorts. The distribution of the microsatellite alleles was similar in the SA and British Caucasian cohorts as shown in Figure 4.21.

Figure 4.20. Example of microsatellite spectra from unamplified DNA and WGA DNA from the same sample.

Amplicon size (X-axis) and product yield (Y-axis) are shown for the same sample with and without WGA. Numbers on the X-axis represent the peak size estimates obtained using GeneMarker autodetection software. In unamplified DNA the true peak (giving the greatest size product) had the greatest peak height. However, in WGA DNA the 'slippage' peak height exceeded the height of the true peak, which complicated the genotyping interpretation. In the example below, the native DNA suggests a 178/178 homozygote genotype but the pattern in the WGA sample could be consistent with a 176/178 heterozygote genotype. The explanation and interpretation of the spectra patterns is explained in Chapter 2.

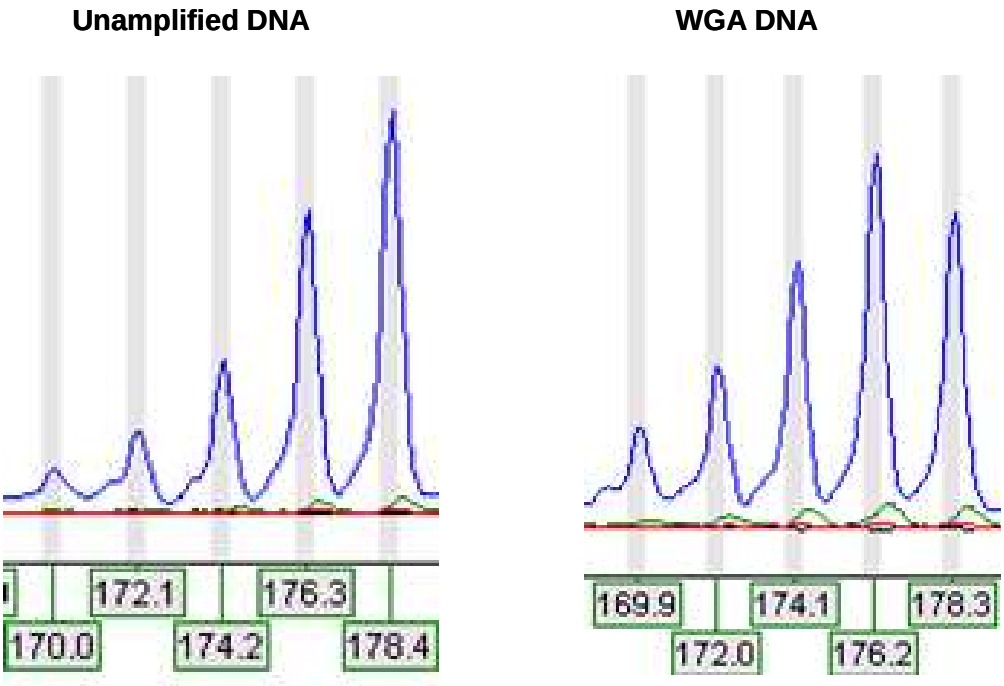
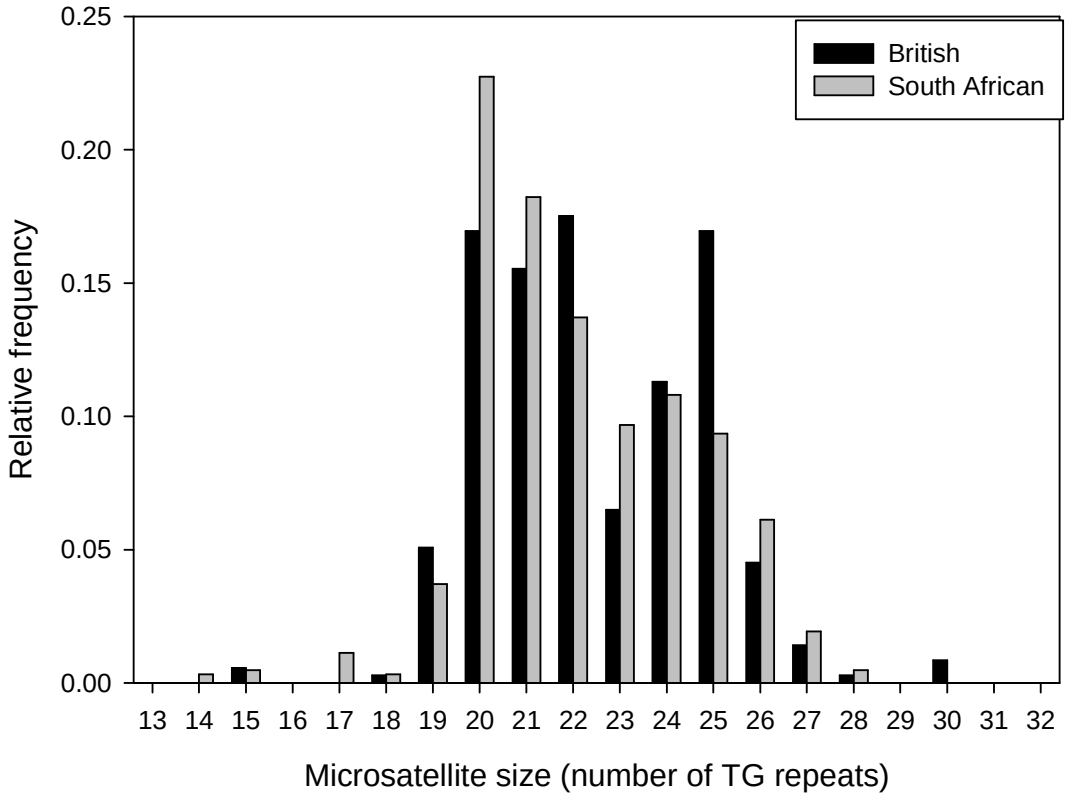


Figure 4.21. Relative frequency of microsatellite rs10583774 alleles in the SA and Caucasian cohorts.



Total microsatellite length (allele 1 + allele 2) showed a weak but significant correlation with total expression of *CDKN2A* and *CDKN2B*, with increasing number of TG repeats associated with reduced expression (as shown in Table 4.8). There was a trend in the same direction for *ANRIL* exons1-2 expression which did not achieve significance. To investigate potential *cis*-acting effects, an analysis including only individuals with at least one copy of the commonest allele (20 TG repeats) was performed. This approach allowed the *cis*-acting effect associated with the other allele in each individual to be estimated, but the power of the analysis is less because a smaller number of individuals are included. As shown in Table 4.8 and

Figure 4.22, there was a borderline significant association with expression of *ANRIL* exons 1-2, with a trend towards a negative correlation for all genes that was consistent

with the results of the previous combined total microsatellite allele length. All of these analyses were performed with exclusion of a single outlying individual with a particularly low number of TG microsatellite repeats (as highlighted in Figure 4.22).

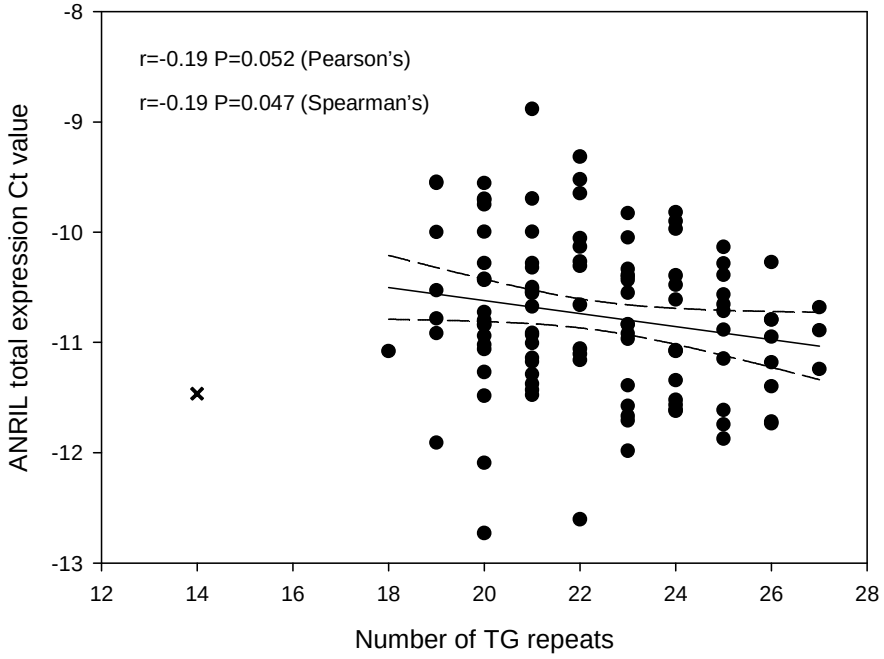
Table 4.8. Correlation of total expression levels with number of microsatellite TG repeats.

	CDKN2A	***CDKN2B***	***ANRIL***
Correlation between TE and total number of TG repeats	-0.12 (P=0.04)	-0.14 (P=0.02)	-0.08 (P=0.14)
Correlation between TE and number of TG repeats at the alternative allele in individuals with 1 copy of the common allele	-0.08 (P=0.38)	-0.05 (P=0.63)	-0.19 (P=0.05)

Shaded cells represent statistically significant correlations. TE = total expression.

Figure 4.22. Effect of number of microsatellite rs10583774 TG repeats at the alternative allele in individuals who have the common (20 TG repeats) allele on total expression of *ANRIL* exons1-2 in the SA cohort.

Each point represents an individual. Regression line and correlation coefficients are shown after exclusion of a single outlier (shown as a cross).



Phasing the microsatellites was difficult because the large number of alleles created a large degree of uncertainty around the phase. Therefore the *cis*-acting influence of microsatellite size was investigated in individuals who had at least one copy of the commonest allele using the approach described above with absolute log-transformed AER values at each transcribed SNP (adjusted for the effect of the transcribed SNP). This analysis was performed for each transcribed AEI marker separately, and could only be performed for individuals who had one copy of the common microsatellite allele *and* who were heterozygous at the transcribed SNP (to permit AEI measurement). The number of informative samples was therefore low and no significant correlation was seen between number of microsatellite repeats and AER at any of the transcribed markers in *CDKN2A, CDKN2B* and *ANRIL* using this approach.

The above analyses assume a linear relationship between microsatellite length and *cis*-acting effect on expression, but previous *in vitro* studies have suggested that the relationship between microsatellite length and expression is not linear in all cases[241]. Using a regression analysis for total expression with multiple alleles as categorical variables (which allows each allele to have an independent effect) showed a significant association between microsatellite genotype and expression of *ANRIL* exons1-2 (P=0.0047), but not *CDKN2A* or *CDKN2B*. This suggests that the genotype at rs10583774 is associated with *ANRIL* exons1-2 expression, but that the relationship may not be linearly proportional to the number of TG repeats at this locus.

4.6 Discussion

This is the most detailed study to date of *cis*-acting influences on expression at the chromosome 9p21 locus. It has shown that multiple sites in the 9p21 region independently influence *CDKN2A, CDKN2B* and *ANRIL* expression. SNPs associated with diseases including CAD, diabetes, and cancers are all highly associated with expression of *ANRIL* exon 1-2, suggesting that modulation of *ANRIL* expression may mediate disease susceptibility. The study also reports novel methodology for allelic expression analysis that allowed data from multiple transcribed polymorphisms to be combined and the effects of particular SNPs to be

adjusted for. The study has demonstrated that this approach has greater power than total expression analysis for mapping *cis*-acting effects.

4.6.1 Relationships between *CDKN2A*, *CDKN2B* and *ANRIL* expression

Total expression levels of *CDKN2A*, *CDKN2B* and *ANRIL* exons 1-2, which reflect the combined influence of *cis* and *trans*-acting factors, were positively correlated. This corroborates other recent data[197, 325, 326], and suggests that expression of these genes is co-regulated. *Trans*-acting influences account for the majority of the observed variance in expression of these genes (80-96%), and the correlation in total expression levels is likely to reflect co-regulation of the genes through *trans*-acting factors. In addition, the allelic expression analysis demonstrated that expression is also influenced by shared *cis*-acting elements in the region. Despite the positive correlation in total expression levels, *cis*-acting effects associated with individual SNP alleles may act in opposite directions; the effect of individual SNPs on *CDKN2B* expression were opposite to effects on expression of *CDKN2A* and *ANRIL* exons 1-2 (which were concordant) in this study. Although *cis*-acting effects account for only a small proportion of the overall variance in expression of these genes, this does not diminish the potential biological significance of the *cis*-acting effects. *ANRIL* overlaps and is transcribed in antisense with respect to *CDKN2B*[195]. It is modestly conserved across species[200] and its function is not known, but recent work has demonstrated that antisense transcription from *CDKN2B* downregulates *CDKN2B* expression in *cis* through heterochromatin formation[202]. This is consistent with our observation of an inverse effect of SNPs on *ANRIL* exons 1-2 and *CDKN2B* expression. By contrast, *CDKN2A* and *ANRIL* exons 1-2 showed positive correlations for both allelic and total expression in our study. *CDKN2A* and *ANRIL* do not overlap, but are transcribed divergently from transcription start sites separated by just 300 base pairs. Although the *ANRIL* promoter is currently not characterised, it may share promoter elements with *CDKN2A* and the resulting co-regulation could account for the positive correlation in expression we observed for these genes, similar to that described at other sites[244]. In this context, inhibition of *CDKN2B* expression by *ANRIL* would enable a level of crosstalk between *CDKN2A* and *CDKN2B* expression, which would be consistent with the inverse *cis*-acting effect of SNPs on *CDKN2A* and *CDKN2B* that was observed. The observation that *cis*-acting genetic effects played a

greater role in expression of *ANRIL* compared to *CDKN2A* and *CDKN2B* (20% compared to less than 8% and 5% respectively) makes it a good candidate for genetic causation mediated through influences on expression.

A recent mouse study demonstrated that deletion of a 70kb non-coding region orthologous to the human CAD core risk region resulted in severely reduced expression of *Cdkn2a* and *Cdkn2b* in heart tissue, indicating the presence of distant-acting regulatory functions in the CAD risk interval[327]. Allele-specific analyses demonstrated that these effects were mediated by *cis*-acting mechanisms. Although sequence conservation of *ANRIL* is not seen in this orthologous region, this may be due to high rates of sequence evolution without loss of functional activity that has been proposed for long non-coding RNAs[328].

4.6.2 aeQTL mapping of *cis*-acting effects

The study compared total expression and allelic expression for the investigation of *cis*-acting influences on expression. For aeQTL mapping a novel adaptation of previously reported methodology[254] was developed to combine multiple transcribed SNPs per gene, which increased the number of informative individuals and the power for detecting *cis*-acting effects. This approach was demonstrated using two transcribed polymorphisms per gene, but the methodology offers the potential for the inclusion of multiple additional transcribed variants. The results obtained by eQTL and aeQTL mapping were similar, consistent with previous work suggesting that the two approaches identify the same *cis*-acting loci[266]. However, the present study demonstrated that aeQTL analysis had substantially greater power than the eQTL approach. Adjusting for *trans*-acting covariates including age, sex and ethnicity in the eQTL analysis did not substantially alter the results. An influence of age on *CDNK2A* has been reported[329], but there was little variability in the age of the SA cohort (90% of whom were between the ages of 18 and 30 years). The fact that allelic expression is a more efficient way to identify *cis*-acting influences on expression has implications for future studies investigating the effects of SNPs on expression at other loci, for example for the hundreds of non-coding SNPs correlated with different diseases by recent GWA studies[25].

Allelic expression quantifies the relative contributions of each allele to the mRNA pool irrespective of the absolute mRNA levels, and therefore provides information about transcriptional effects and polymorphisms within the transcript influencing RNA degradation in *cis*. By contrast, total expression analyses that quantify absolute mRNA levels are also sensitive to post-transcriptional regulatory effects, such as mRNA degradation by microRNAs. In extreme cases tight post-transcriptional regulation could keep total mRNA levels constant irrespective of the contributions of each allele to the total mRNA pool. The fact that the results of eQTL and aeQTL mapping were so similar in this study suggests that the effect of regulation at the post-transcriptional level is limited, although regulation of *CDKN2A* expression by a microRNA has been described[330]. In general, although allelic expression is a robust method for mapping sites influencing expression in *cis*, investigation of total expression and other intermediate phenotypes such as protein levels or protein activity will provide complementary information that contributes to fully understanding the phenotypic effects of *cis*-acting polymorphisms. It would be desirable to determine whether the significant associations with mRNA expression observed for *CDKN2A* and *CDKN2B* are confirmed at the protein level.

4.6.3 Trans-ethnic expression mapping

Although it had been hoped that trans-ethnic fine-mapping could be used to refine the associations with expression, the results of aeQTL mapping were in fact very similar in the SA and Caucasian populations. This replication in a separate cohort strongly supports the validity of the findings and enabled a combined analysis of the two cohorts to be performed. This approach of pooling data from ethnically-divergent populations has been previously shown to increase the power to detect influences on expression that are shared across populations[266, 331]. The principal difference identified between the two populations was for the SNPs associated with type II diabetes. The lead diabetes SNP rs10811661 was correlated with *ANRIL* underexpression in the Caucasian cohort, but not in the SA population, despite greater power to detect effects in that cohort. This may reflect differences in LD between the populations, but suggests that if rs10811661 is the causal variant influencing diabetes susceptibility the effects are unlikely to be mediated through influences on *ANRIL* expression. Alternatively, if modulation of *ANRIL* expression is the mechanism

through which the association of rs10811661 with diabetes is mediated, as suggested by the association in Caucasian populations, rs10811661 itself is unlikely to be the 'functional' variant causing the effect (but is likely to be in LD with the functional variant in the Caucasian but not the SA population). Studies to determine whether this SNP is associated with diabetes in populations of African origin would be of interest.

4.6.4 Consideration of 'causal' variants

The power of the analyses to detect differences in expression enabled the effects of individual SNPs to be adjusted for. Using this approach demonstrated that expression, and therefore probably disease predisposition, was independently influenced by multiple sites and that the observed effects cannot be explained by a single polymorphic site. From this analysis the existence of rare variants with large effects cannot be excluded, but previous resequencing studies in this region did not find rare variants associated with disease phenotypes[19, 20]. It is uncertain whether the individual SNPs for which associations were found are the actual 'causal' variants responsible for the effects on expression, or if the association simply reflects LD between these SNPs and the causative polymorphisms. Although fine mapping studies often purport to identify causal variants, in the context of complex diseases identifying pathways involved in disease predisposition may be more important. This is of particular interest for these genes where variation in expression is mostly due to *trans* effects which may be substantially influenced by non-genetic factors, raising the prospect that it may be amenable to therapeutic modulation. The putative causal variants rs10757278 and rs1333045 previously associated with altered *ANRIL* expression[200, 325] were significantly associated with reduced *ANRIL* exons 1-2 expression *in vivo* in this analysis, but their effects were relatively modest compared to other SNPs in the region and adjustment for the effect of these SNPs accounted for only a small proportion of the effect observed at other SNPs. The maximum changes in expression associated with individual SNPs were substantial, up to 2-fold for *ANRIL* exons 1-2, but effects of much smaller magnitude were also detected; the minimum significant effect was associated with just a 1.05-fold change in expression. Although the associations of SNPs with expression that were observed were statistically highly significant, it is uncertain what impact such effects on expression

have on disease risk. However, even small differences in gene expression due to genetic factors that are present throughout an individual's lifetime could contribute to differences in common late-onset phenotypes such as CAD and diabetes, and the effects may be even greater in tissues related to disease.

Genotype at microsatellite rs10583774 appears to be associated with expression of *CDKN2A*, *CDKN2B* and *ANRIL* exons 1-2. Such an association may be 'indirect', resulting from LD between the microsatellite and other variants which have true functional effects on expression. Alternatively, it is possible that the number of microsatellite TG repeats has a direct functional influence on expression, as has been reported for other microsatellites *in vitro*[238, 241]. The location of this microsatellite within a promoter region likely to influence both *CDKN2A* and *ANRIL* expression supports the notion that it could have direct functional effects, perhaps through affecting transcription factor binding similar to that described for other microsatellites[236]. However, it is interesting that an increasing number of TG repeats was associated with reduced expression of all three genes, including *CDKN2B* which is transcribed in the opposite direction to *ANRIL* (i.e. the microsatellite is located at its 3'-end and not in a promoter region for this gene). These effects could potentially be mediated through changes in chromatin structure which up or downregulate expression of the whole locus, similar to the effects described for other microsatellites[238]. A recent 2010 paper has shown that methylation of the CpG island in this region abolishes binding of the chromatin insulator protein CTCF and reduces expression of *CDKN2A*, *CDKN2B* and *ANRIL*[326]. The association between microsatellite genotype and *ANRIL* expression was stronger using a categorical rather than a linear model. Non-linear relationships between repeat elements and expression have been reported for other microsatellites, which may be due to steric effects such as the proximity and interaction of transcription factors within three-dimensional space of the helical DNA structure[241]. If rs10583774 has direct influences on expression, genotyping this microsatellite may add to the predictive value of SNPs at this locus. The functional role could be investigated using transfection experiments with reporter constructs. It is possible that microsatellites may be the functional elements responsible for SNP associations detected at other loci in GWA studies, and these elements should not be overlooked.

4.6.5 Confirmation of regulatory elements *in vivo*

This study examined *in vivo* expression in primary cells rather than in transformed cell lines. Although cell lines have been extensively used to investigate *cis*-acting influences on expression[227, 331], patterns of expression may be altered in immortalised cells, particularly for genes such as these that are associated with senescence and cell-cycle regulation[220]. Furthermore, widely used cell lines are pauciclonal or monoclonal[255, 332] and since a significant proportion of human genes exhibit random patterns of monoallelic expression within single clones of cell lines[333], *cis*-acting effects in these cells are unlikely to be representative of polyclonal cell populations *in vivo.* Previous studies have delineated the promoters and other elements regulating *CDKN2A/ARF* and *CDKN2B* expression using reporter assays[186, 187, 189, 319, 320]. Such studies are valuable to identify causative polymorphisms, but since they examine the effects on expression outside of the normal haplotype, chromatin and cellular context their findings require confirmation by *in vivo* studies[225, 253]. This analysis confirmed that polymorphisms in upstream regulatory elements identified by *in vitro* assays were significantly associated with *cis*-acting effects on expression *in vivo*, but also demonstrated that other loci located up and downstream were associated with effects on expression of similar or even larger magnitude. These data highlight the complexity and multiplicity of sites influencing expression in the region. The assays used to investigate *CDKN2A* expression also included the *ARF* transcript variant. This gave the potential to detect sites influencing expression of both transcripts, and effects of SNPs in both the *CDKN2A* and *ARF* promoter regions were detected in the study, although differential effects of loci on individual transcripts cannot be distinguished using this approach.

4.6.6 Tissue-specific considerations

All of the SNPs in the region associated with disease in GWA studies were associated with influences on *ANRIL* exons 1-2 expression, suggesting that modulation of *ANRIL* expression may mediate susceptibility to these phenotypes. SNPs in the CAD core risk haplotype region[129] that are most strongly associated with CAD in GWA studies were associated with reduced expression of *ANRIL* exons 1-2, but other SNPs associated with CAD which lie outside of the core risk haplotype region showed independent and stronger associations with underexpression of *ANRIL* exons 1-2.

This may reflect differences in the relative importance of particular sites in the tissues responsible for the association with CAD. Indeed, the patterns of association observed in peripheral blood in healthy individuals may differ from those in primary disease tissues. Similarly, differences in the relative contribution of each SNP to modulation of expression in the tissues crucial for the pathogenesis of the different conditions could explain why particular diseases are associated with different subsets of SNPs that influence *ANRIL* expression. Recent work also suggests that *ANRIL* has multiple transcripts, which may be differentially expressed between tissues[197, 200]. Confirmation of the findings of the present study in tissues relevant to each disease and for different *ANRIL* transcripts would therefore be desirable, although for CAD and other complex diseases the exact cell populations responsible for mediating disease susceptibility are not certain and may be inaccessible. Although tissue specificity of *cis*-acting influences is well documented, variation in *cis*-acting effects is primarily explained by genetic variation, with allele-specific expression for most transcribed SNPs being the same in different tissues of the same individual which express the gene of interest[269]. Analysis of expression in blood is therefore likely to give biologically relevant information despite the fact that this may not be the tissue in which influences on expression actually mediate disease susceptibility.

4.6.7 Other studies of chromosome 9p21 expression

Previous genome-wide expression analyses using microarrays and immortalised cell lines did not identify association of *CDKN2A* and *CDKN2B* expression with markers in this region, although they did not examine *ANRIL* expression[227, 331]. However, several recent studies have specifically examined the relationship between CAD risk variants and total expression levels of transcripts in the chromosome 9p21 region. These studies are described below, and the locations of the primers used relative to the annotated *ANRIL* transcripts are illustrated in Figure 4.14 (page 159).

The first study investigating the association between CAD risk variants and 9p21 gene expression was published by Liu *et al* in April 2009[325]. This study measured total expression levels of *CDKN2A*, *CDKN2A-ARF*, *CDKN2B*, *MTAP* and *ANRIL* in peripheral blood T-cells from 170 healthy subjects using TaqMan assays, and analysed the association with three CAD risk SNPs (rs10757278, rs518394 and rs564398). The risk allele of rs10757278 which is located within the CAD core risk

region defined by Broadbent *et al*[129], was significantly associated with reduced expression of *CDKN2A*, *CDKN2A-ARF*, *CDKN2B* and *ANRIL*. This supports the finding of the present study that CAD risk variants are associated with reduced *ANRIL* expression. However, rs10757278 and other SNPs in the core CAD risk region were not associated with expression of *CDKN2A* or *CDKN2B* in the present study, which conflicts with the findings of Liu *et al*, but is supported by the findings of other larger studies as described below. Liu *et al* found no association with expression for the other two CAD risk variants studied, despite them being in moderate LD with rs10757278 (D'=0.66 with rs10757278 for both). This is surprising since rs564398 was strongly associated with a 1.9-fold reduction in allelic expression of *ANRIL* in the present study, which was larger and more significant than the effect associated with rs10757278. The assay used by Liu *et al* to measure *ANRIL* expression spanned exons 4-5 whereas the assay used in the present study spanned exons 1-2. Recent data suggests that *ANRIL* may have multiple transcripts and it is possible that the different assays used in the studies capture different transcripts that are differentially regulated. This could potentially account for the lack of association between *ANRIL* expression and rs564398 genotype and that was observed by Liu *et al*.

Liu *et al* also found no correlation of risk variants associated with diabetes (rs10811661), frailty (rs2811712), and melanoma (rs11515) with expression of any of the genes studied. However, significant associations with allelic expression were found for these variants in the present study: rs10811661 with *CDKN2A* and *ANRIL*; rs2811712 with *CDKN2B*; rs11515 with *CDKN2A* and *CDKN2B*. The effects and significance of associations for these variants were smaller than those detected between rs10757278 and *ANRIL* expression, and the fact that the present study was able to detect such effects that were not detected by Liu *et al* is likely to reflect the greater power of the present study due to the larger sample size and increased sensitivity of the aeQTL mapping approach.

Jarinova *et al* investigated the functional effects of four conserved sequences within the CAD risk region using *in vitro* reporter gene expression in primary aortic smooth muscle cells[200]. One of the conserved sequences demonstrated enhancer activity, and the CAD risk allele rs1333045-C in this sequence was associated with higher expression compared to the T allele. The study then went on to examine the

association of this putative functional SNP with total expression of *CDKN2A*, *CDKN2B*, *MTAP*, and *ANRIL* in peripheral blood from 120 healthy individuals[200]. In keeping with the results of the present study Jarinova *et al* found no significant association between rs1333045 genotype and expression of *CDKN2A* and *CDKN2B*, but did find a significant association with *ANRIL* expression. Primers located in different exons were used to investigate effects on expression of the 'long' and 'short' *ANRIL* transcripts described by Pasmant *et al*[195]. Primer pairs were located in exons 1 and 5 (transcript EU741058), within exon 13 (transcript DQ485454/EU741058), and in exons 15 and17 (transcript DQ455453). The CAD risk allele was associated with increased expression of the 'short' transcripts (DQ485454/EU741058 and EU741058) but reduced expression of the 'long' transcript (DQ455453), suggesting that these *ANRIL* transcripts are differentially regulated. The *ANRIL* assay spanning exons 1-2 used in the present study is expected to capture both the 'long' and 'short' transcript variants, and the finding that CAD risk variants reduce expression measured at *ANRIL* exons 1-2 is consistent with the findings of Jarinova *et al* if the 'long' isoform is expressed at higher levels in peripheral blood than the 'short' isoform. Exon-specific total expression assays performed in the present study did indeed demonstrate higher expression of the exon 20 assay (specific for the 'long' transcript), compared to the exon 13 assay (specific for the 'short' transcript), although expression measured with these assays was not significantly associated with rs1333045 genotype. Jarinova *et al* also performed whole-genome expression analysis using microarrays in subsets of healthy subjects who were homozygous for the risk and non-risk allele. These analyses suggested that pathways associated with cell proliferation and vascular endothelial growth factor signalling were upregulated in individuals who were homozygous for the risk allele, suggesting that the risk allele may act by modifying expression of cell cycle regulatory genes.

A study by Folkersen *et al* also used investigated the association between a chromosome 9p21 CAD risk variant (rs2891168) and whole genome total expression data from microarrays[197]. Association with microarray expression was analysed in five separate datasets comprising: 57 CEU lymphoblastoid cell lines; 87 CEU and 89 YRI lymphoblastoid cell lines; 117 carotid endarterectomy specimens; 88 mammary artery medial samples; and 89 aorta medial samples. No significant associations with expression in *trans* were found in these datasets, and there was also no evidence of a

cis-acting effect on expression of the neighbouring genes including *CDKN2A*, *CDKN2B* and *ANRIL*. However, the power of the association analysis was limited due to the small numbers studied, low expression levels of *ANRIL*, and relative insensitivity of the microarray platform. These factors are likely to account for the lack of association observed by Folkersen *et al*, which contrasts with the findings of the present study and the other published studies.

A recent study by Holdt *et al* published in January 2010 investigated the association of CAD risk variants with total expression of chromosome 9p21 genes in peripheral blood mononuclear cells from 1,098 patients with varying degrees of CAD assessed by coronary angiography[155]. In keeping with results of the present study CAD variants were not consistently associated with *CDKN2A* or *CDKN2B* expression, but were associated with *ANRIL* expression. The study investigated expression of one 'long' and two 'short' *ANRIL* transcripts using different primer pairs located in *ANRIL* exons 1 and 5 (transcript EU741058), exons 4 and 5 (transcript DQ455454/DQ455453), and exons 18 and 19 (transcript DQ455453). The CAD risk haplotype was associated with increased expression of the 'short' transcript variant EU741058, but had no effect on expression of the other 'short' transcript variant DQ485454. Expression of these transcripts was associated with the severity of atherosclerosis, providing the first direct evidence linking *ANRIL* and atherosclerosis susceptibility. In contrast to the findings of Jarinova *et al*[200], however, the CAD risk haplotype was associated with increased expression of the 'long' transcript DQ455453. Similar findings were obtained in whole blood from 154 individuals free from CAD. The assays used to try and study transcript-specific expression by Jarinova *et al* and Holdt *et al* were located in different exons, and recent data suggests that complexity of *ANRIL* transcripts beyond the three isoforms considered in these studies might account for the differences in the findings, as discussed below.

4.6.8 Consideration of transcript-specific *ANRIL* expression

The report by Folkersen *et al*, which was published in November 2009 after the expression studies discussed above were conducted, suggested that the situation with respect to *ANRIL* splice variants may be more complex than had previously been suspected[197]. Investigation of splice variants was performed using real-time PCR amplification and sequencing in cDNA libraries from cell lines derived from three

different tissues: human umbilical vein endothelial cells (HUVEC), brain, and lung. This study amplified sequence between PCR primers located in exon 1 and exon 20 for each cell type, and for the HUVEC cells additionally amplified sequence between PCR primers located in exon 1 and exon 13. Using these assays, Folkersen *et al* were unable to identify 'full length' transcripts of 3,834 or 2,659 bp as previously described by Pasmant et al[195]. However, a number of shorter transcripts were detected, which appeared to differ between the different cell lines studied (as summarised in Figure 4.14 on page 159). Based on these observations the authors proposed that *ANRIL* undergoes extensive alternative-splicing, and that there is tissue-specific expression of *ANRIL* transcripts. In the report by Folkersen *et al* there is no data presented to show that the cDNA quality and methodology used was adequate for the detection of long transcripts. This is important since failure to detect the 'full length' *ANRIL* transcripts could be the result of inadequate cDNA fragment length or other methodological issues (similar to the problems in detecting long transcripts encountered in the present study) and positive controls for other long transcripts would be useful. Furthermore, other recent work has shown that culture of human cells is commonly associated with culture-induced copy number changes in genomic DNA and changes in gene expression, particularly for genes involved cell-cycle regulation[334]. The implications of the findings from Folkersen *et al's* study with respect to *in vivo* expression remains uncertain at present. Deletions related to cell culture could account for the failure to detect certain blocks of exons that was observed. It seems perhaps unlikely that this would occur in three separate cell lines within the same gene, but particular genes that are prone to culture-induced anomalies have been previously demonstrated[334]. The methodology used by Folkersen *et al* to study transcript variants did not allow the full range of potential transcripts to be surveyed, and the possible range of transcript isoforms in different native tissues is unclear.

Taken together, the results of the present study and other published studies clearly suggest that CAD variants are associated with expression of *ANRIL* transcripts but are not consistently associated with *CDKN2A*, *CDKN2B* or *MTAP* expression. However, additional work is needed to fully characterise the alternative transcripts and loci that influence their expression in different tissues.

4.6.9 Mechanistic considerations

The finding that disease associated SNPs are all associated with *ANRIL* exons 1-2 expression suggests that *ANRIL* plays a role in influencing disease susceptibility. Although little is known about the targets of *ANRIL*, its effects may be mediated through antisense transcription regulation of *CDKN2B* in the tissues critical for the pathogenesis of the different diseases. The observation that the effects of sequence variants acting in *cis* were stronger for *ANRIL* than for *CDKN2B* may reflect selection pressure against variants that have substantial direct effects on the expression of critical genes. *CDKN2A*, *ARF* and *CDKN2B* are cell cycle regulators and are plausible candidates for involvement in the pathogenesis of the diseases for which SNP associations with *ANRIL* were found. Mutations involving these genes are well documented in glioma[335, 336] and melanoma[195, 337, 338]. Overexpression of *CDKN2A* and *CDKN2B* in murine models is associated with pancreatic islet hypoplasia and diabetes[339, 340], and there is also emerging evidence that vascular cell senescence involving these pathways is involved in the pathogenesis of atherosclerosis[193, 194].

4.6.10 Summary

This study shows that multiple independent sites in the chromosome 9p21 region influence *CDKN2A*, *CDKN2B* and *ANRIL* expression. SNPs associated with a number of different diseases in GWA studies are all associated with *ANRIL* expression, indicating that modulation of *ANRIL* expression mediates susceptibility to a variety of conditions.

Chapter 5

Investigation for copy number variation in the chromosome 9p21 region

5 Investigation for copy number variation in the chromosome 9p21 region

5.1 Abstract

Departure from HWE is frequently analysed in genetic studies as a screening check for genotyping errors, although this phenomenon may be due to many other factors. In the NE Caucasian cohort a higher proportion of SNPs in the chromosome 9p21 region deviated from the proportions expected under HWE than expected by chance alone (12/53 at the $P<0.05$ threshold), with genotyping demonstrating an excess of homozygosity. The aim of this chapter was to investigate the possible causes of the observed departure from HWE. Review of the genotyping and repeat genotyping by both Sequenom and TaqMan methodologies in the same cohort showed no evidence of substantial genotyping errors. Analysis of published datasets which had genotyped SNPs in this region revealed that departure from HWE with an excess of homozygosity was also more prevalent for SNPs in these cohorts than would be expected by chance alone; these data could be consistent with the existence of common deletions (leading to null alleles in the typed SNPs) in Caucasian populations. Genotyping the 1425 members of 248 families of the HTO cohort for 17 SNPs that had previously shown HWE departures showed no significant deviations from HWE in that population, but analysis of pedigrees with Mendelian errors revealed patterns which could be consistent with segregation of a null allele in nine families. MLPA analysis with 12 custom probes in the region of interest showed no evidence of CNV in these nine individuals, or 118 unselected individuals from the NE Caucasian and HTO cohorts. In summary, the observed departure from HWE was not attributable to genotyping errors and was not replicated in the HTO cohort; nor was there evidence of deletion polymorphisms in the region using MLPA.

5.2 Introduction

The Hardy-Weinberg law states that allele and genotype frequencies of a large, randomly-mating population remain constant between generations in the absence of migration, mutation, and selection[341, 342]. According to this law, if two alleles, G and g, with frequencies p and $q = 1 - p$, are in equilibrium in a population, then the

proportion of people with genotypes GG, Gg and gg will be p^2, $2pq$ and q^2. Although the assumptions underlying HWE are rarely met in human populations[343], it has been proposed that testing for HWE can be used to detect errors or peculiarities in datasets analysed in genetic association studies[344]. As presented in Chapter 4, a higher proportion of SNPs genotyped in the NE Caucasian cohort showed deviation from HWE than would be expected by chance alone: 12/53 SNPs (23%) at the P<0.05 threshold, and 2/53 SNPs (4%) at the P<0.01 threshold. A higher proportion of SNPs in this region also deviated from HWE in some other published series (presented in section 5.5.1.2 on page 201). This chapter investigates possible causes and implications of the observed departure from HWE in the study population.

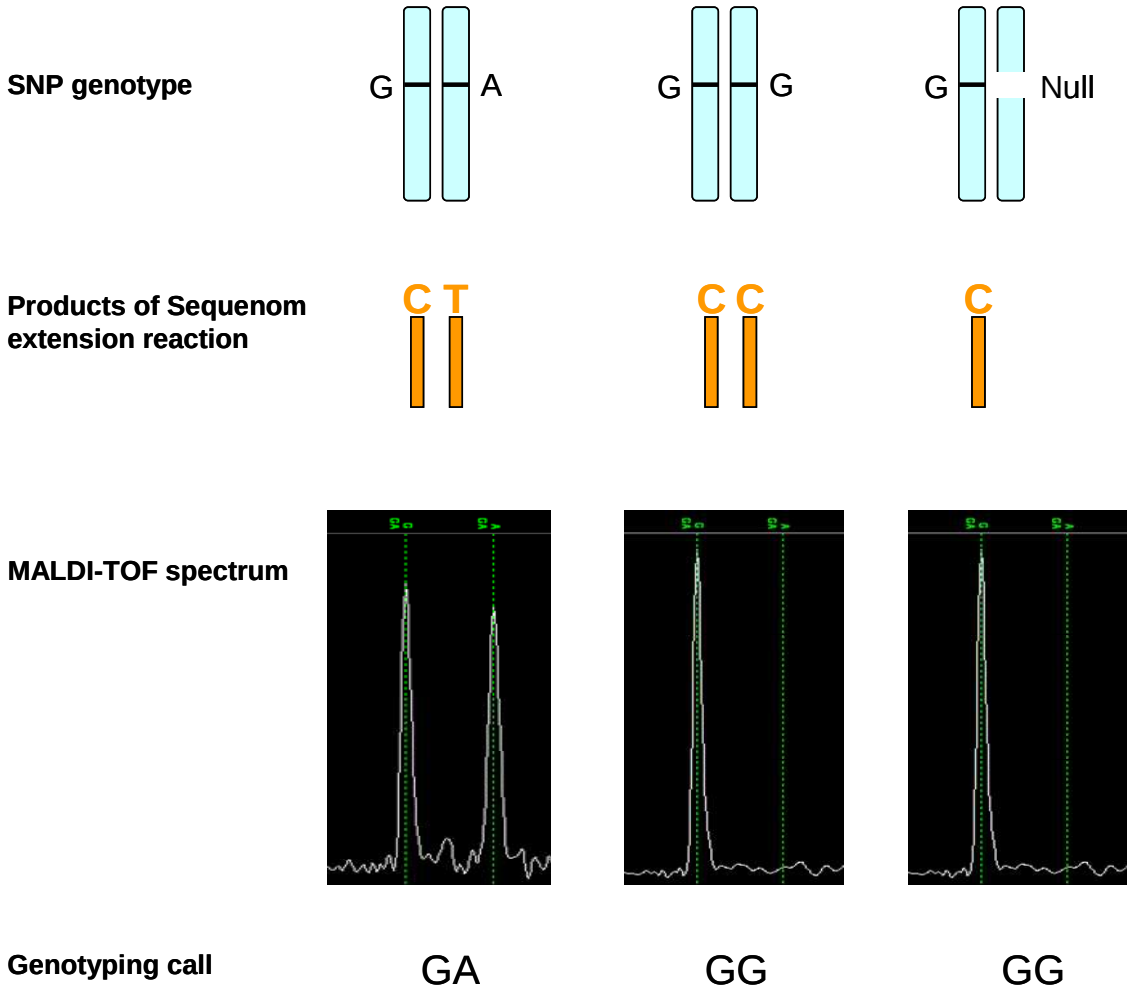
The most common reason for testing for departure from HWE in datasets used for genetic studies is to screen for genotyping errors. However views differ on the usefulness of analysing HWE in this context, as recognised in a recent consensus guideline for good practice in the reporting of genetic association studies[345]. Whilst some authors propose that it should be used as a 'quality control' for genotyping and form an essential part of any genotyping study[344, 346], others have argued that it should not be used as a tool to identify genotyping errors, and that its use is unnecessary and even unhelpful as it may alter the type-I error rate of the association test[347]. Departure from HWE may detect gross systematic genotyping errors, but the power to detect genotyping errors of smaller magnitude has been shown to be low[109, 347]. Consensus is also lacking on the appropriate P-value threshold for HWE-related quality control. Rather than recommending routine HWE testing, the current guidelines simply recommend transparent reporting of whether such testing was done and the thresholds used[348]. The first part of this chapter presents the studies that were performed to investigate genotyping error as a cause of the departure from HWE that was observed in the study population.

Factors that violate any of the assumptions of HWE may lead to deviation from the expected genotype proportions. These include non-random mating, such as inbreeding which tends to increase homozygosity; and population stratification, in which the population is comprised of a mix of sub-populations with different allele frequencies and possibly other confounding factors[349]. The significance of these with respect to the study population is considered in the discussion.

Departure from HWE may also be due to copy number polymorphisms[299]; these are of particular interest since CNVs may modulate gene expression and be involved in complex disease causation[249, 350, 351]. Genome-wide surveys have reported common deletion polymorphisms[352, 353] and segmental duplications[354, 355] in the human genome. With many genotyping techniques, including the Sequenom assay used in this study, SNPs that are hemizygous for a deletion are miscalled as homozygous for the allele that is present, as illustrated in Figure 5.1. Deletion polymorphisms may therefore cause departure from HWE with an apparent over-representation of homozygotes. The disequilibrium coefficient (D-statistic) can be used to determine whether departures from HWE are due to an excess of homozygosity[343, 356]. CNVs have been shown to be associated with complex disease[351], and recent studies have shown that many common deletions are in LD with SNPs, such that they are effectively assayed by proxy in SNP-based association studies[252, 355]. Hence if common deletions in the chromosome 9p21 region were functionally important and influenced disease predisposition, they could potentially account for at least some of the association of SNPs in this region with disease phenotypes. Although previous studies sequencing this region did not identify deletions associated with disease, neither did they exclude them as causes of potentially important effects. McPherson *et al* sequenced the complete 58kb risk interval in only two homozygotes for the risk haplotype[20], and whilst Helgadottir *et al* studied 93 affected individuals[19], they only sequenced exons, exon-intron junctions, and regulatory regions of *CDKN2A* and *CBKN2B* and hence did not exclude CNVs in the *ANRIL* region where the greatest departure from HWE was observed. SNPs and other known variants discovered to date account for only a small proportion of the genetic risk for MI, and CNVs which remain largely unexplored territory may account for some of the unexplained heritability in MI and other complex diseases[357].

Figure 5.1. Illustration of homozygous genotyping miscalling in samples with a hemizygous deletion.

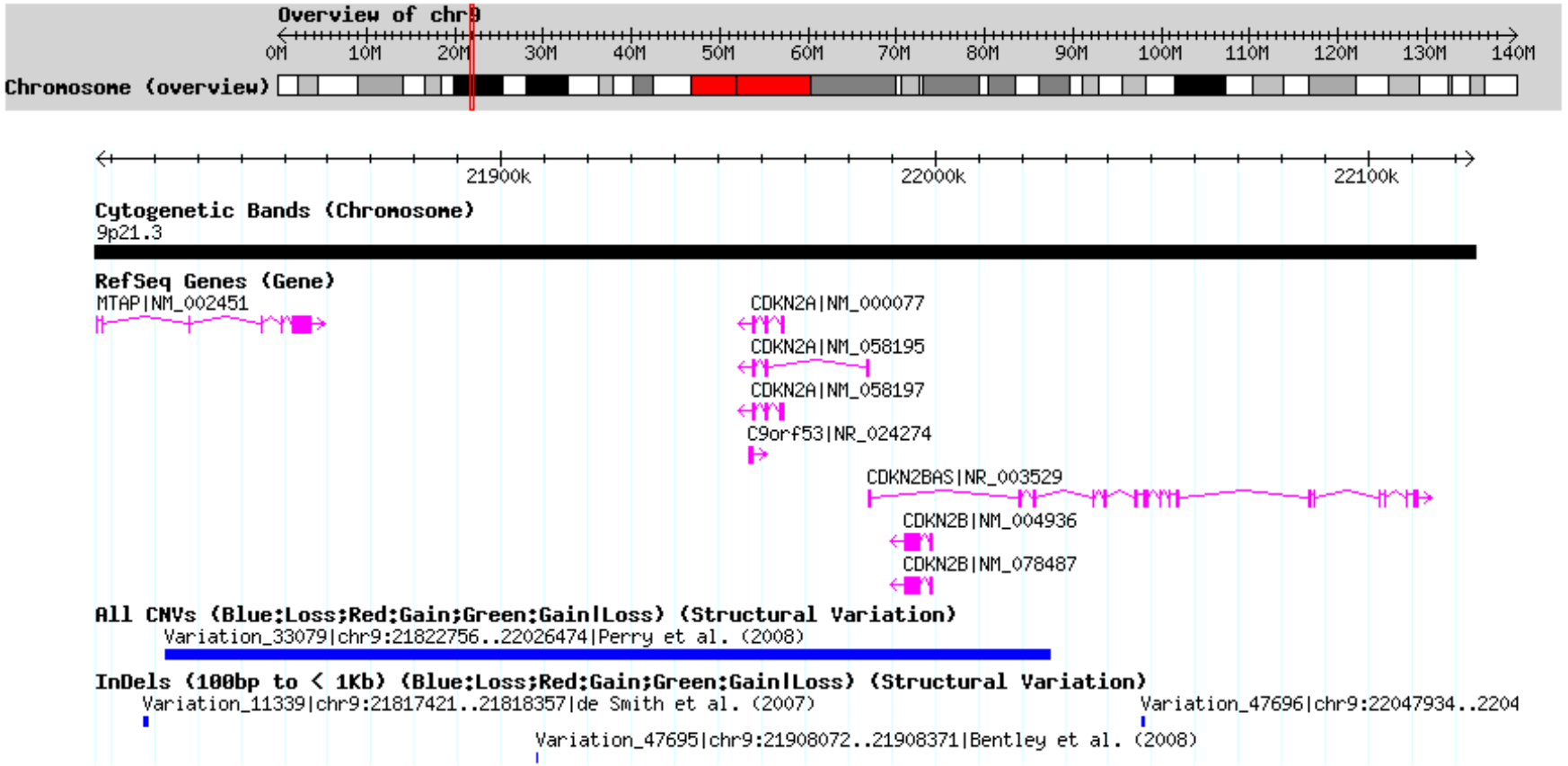
Heterozygous individuals produce two different extension products and two peaks on MALDI-TOF analysis. By contrast, individuals who are homozygous or have a hemizygous deletion (null allele) both produce a single extension product and one peak on MALDI-TOF analysis, resulting in miscalling of samples with a hemizygous deletion as homozygous.



Deletions of the chromosome 9p21 region have been reported in rare pedigrees with familial cancers[195], but whether there are also common CNVs in this region has not been established. The Database of Genomic Variants[358] showed that deletions in this region were detected in a number of small genome-wide CNV surveys, as shown in Figure 5.2. Perry *et al* used CNV arrays in cell lines from 30 HapMap samples from four populations and identified a 200kb deletion in one individual[359]. De Smith *et al* used CNV arrays in peripheral blood from 50 French controls, normalised to a cell line reference, and identified a 1kb deletion in one individual[360]. Bentley *et al* performed Illumina sequencing in a cell line from a single Yoruban individual and identified two small deletions in the region; one of 299bp and one of 147bp[361]. These findings suggest that CNV in this region may exist in apparently healthy individuals, although data derived from cell lines requires corroboration since immortalised cell lines have been shown to exhibit copy number changes that are not present in the primary cells[362]. The detection of CNVs in studies involving small numbers of individuals suggests that they may not be rare, although it must be noted that other genome-wide surveys have not identified CNV in this region[352, 355]. One GWA study has specifically investigated the association of CNVs with MI, using 1,320 CNV probes in 2,967 cases with early-onset MI and 3,075 controls[363]. Although this study was able to identify significant associations for SNPs in nine regions (including chromosome 9p21), no significant associations were found for any of the CNVs tested. However, there was no CNV probe in the region investigated in the present study, with the nearest probes located >1Mb distant each side (chromosome 9: 20,791,650 and 23,353,115); the results therefore do not provide evidence against the possibility of an important CNV associated with CAD in the region showing departure from HWE.

Figure 5.2. CNVs annotated in the Database of Genomic Variants in the chromosome 9p21 region genotyped in the study.

Chromosome 9: 21,806,758 - 22,124,172. Diagram downloaded from Database of Genomic Variants 18/02/2010[358].



1

Key to CNVs shown:

Variation_33079 (Perry et al 2008). CNV array targeting known CNV regions at 1kb resolution in cell lines from 30 control individuals from 4 HapMap populations[359].

Variation_11339 (de Smith et al 2007). Genomewide CNV array survey of 50 French control individuals using peripheral blood referenced to single cell line sample[360].

Variation_47695 & variation_47696 (Bentley *et al* 2008). Whole-genome Illumina sequencing in a cell line of a single Yoruban individual[361].

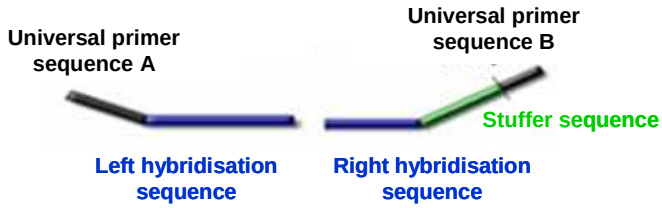
Multiplex Ligation-dependent Probe Amplification (MLPA) is a sensitive method for the detection of CNV[364, 365]. In this technique, which is summarised in Figure 5.3, multiple probesets are designed to hybridise to regions of interest in genomic DNA. Probesets consist of two halves, each of which contains a target-specific hybridising sequence and universal primer sequence (with or without a variably sized non-complementary stuffer sequence). Probesets are designed so that the hydridising sequences of each half bind adjacently on the target DNA and can be joined in a ligation reaction. Each probeset is designed to produce a product of unique size, and only probesets adjacently bound to their target sequence are ligated. This process results in production of a series of continuous probes of different sizes all of which are flanked by universal primer binding sites that can then be amplified by PCR, without amplification of unbound probe halves. The amount of ligated probe and resulting PCR product is proportional to the copy number of the original target sequence. The different sized amplification products are separated using capillary electrophoresis and checked for data quality before proceeding to analysis. The first step of the analysis is intra-sample normalisation, in which the peak area generated by each probe is expressed proportional to a panel of reference probes. The second step is inter-sample normalisation, in which each sample is compared to reference samples of normal copy number, or to the population average (if looking for *de novo* CNVs where reference samples are unavailable). The relative peak heights/areas after normalisation indicate copy number changes, and deletions/duplications can be defined using the MLPA ratio (peak area relative to the area in the reference sample).

Figure 5.3. Summary of the steps involved in MLPA.

Figure adapted from MRC-Holland website[366].

1. Design of MLPA probesets

Multiple probesets each consisting of two halves designed to hybridise adjacently to specific DNA sequences. All probesets contain the same primer-binding sequences A and B. Probesets each of unique length.



2. Denaturisation and hybridisation

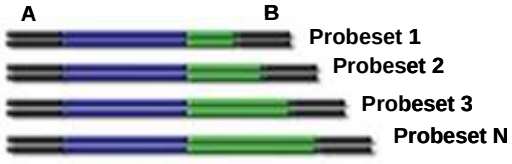
Probes hybridise to target DNA sequence.

3. Ligation of bound probesets

Bound probes ligated. Unbound probes cannot be ligated. Each ligated probeset of unique length.

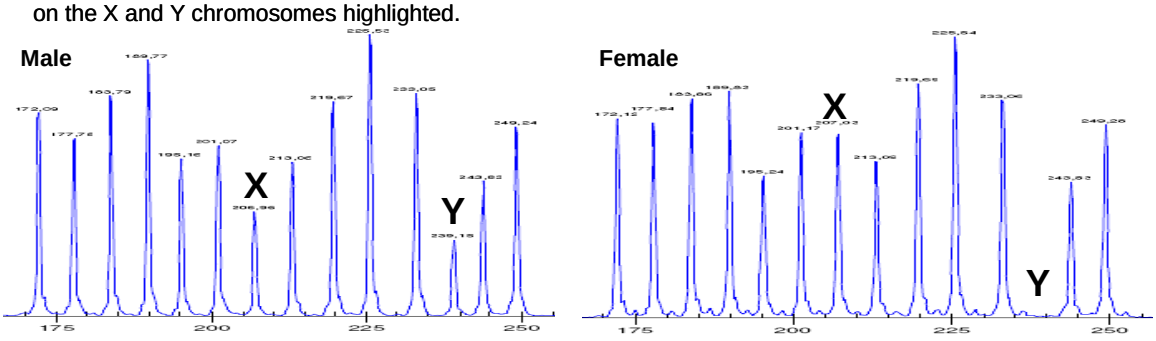
4. PCR with universal primers A and B

Exponential amplification of ligated probes only.

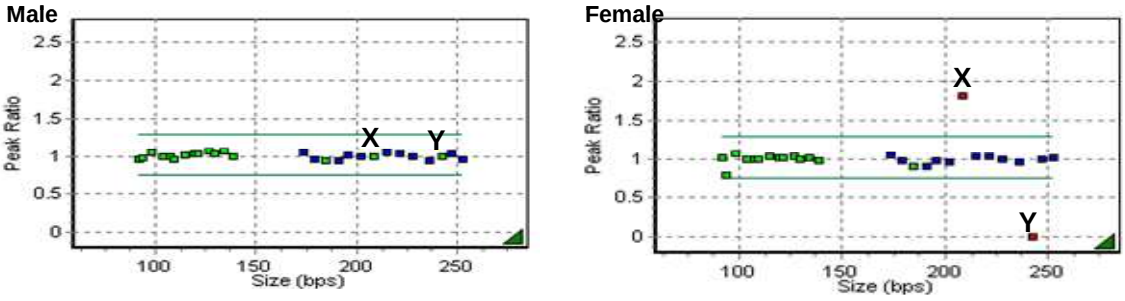


5. Fragment analysis

MLPA spectra forP200 reference probes compared in male and female samples, with peaks from probes on the X and Y chromosomes highlighted.



MLPA ratio plots forP200 reference probes (probes on the right-hand side) compared in male and female samples (normalised to a male sample reference), illustrating the finding of twice the number of copies of the X chromosome probe and no copies of the Y chromosome probe in the female sample.



5.3 Aims

The broad aim of this chapter was to investigate possible causes of the departure from HWE that was observed for SNPs in the chromosome 9p21 region in Caucasian populations.

Specific aims were:

1. To exclude genotyping error as a source of departure from HWE in the NE Caucasian cohort.

2. To investigate the direction and significance of HWE departures for these SNPs in other Caucasian cohorts including:

    a. Previously published datasets

    b. The HTO cohort

3. To look for evidence of inheritance patterns consistent with null allele segregation at these SNPs in families with Mendelian errors of inheritance in the HTO cohort.

4. To investigate for evidence of deletions and CNV in the chromosome 9p21 region in the NE Caucasian and HTO cohorts using custom MLPA analysis of unselected individuals and individuals with Mendelian inheritance errors consistent with null allele segregation.

5.4 Materials and methods

5.4.1 Participants and samples

Genotyping checks and MLPA analysis for CNV were undertaken in the NE Caucasian cohort, in which genotyping results were initially shown to deviate from the proportions expected under HWE. However, the amount of sample remaining for many of these individuals was very limited, with only WGA DNA available for most of the samples. Additional genotyping and MLPA analyses were therefore performed in the HTO Caucasian cohort.

For validation of the ability of the custom MLPA design to detect CNV, positive control DNA samples were used from a small number of individuals in which the number of copies of target genes had been previously quantified. These comprised:

- Five samples from the CHANGE study of families with congenital heart disease that had been characterised for the presence of 22q11 deletion syndrome (DiGeorge syndrome). Three samples had known deletion involving the *CLDN5* gene, and two samples had two copies of this gene (samples and data provided by Dr A Topf, Institute of Human Genetics, Newcastle upon Tyne).

- Six samples that had been characterised for copy number at the *DUSP22* gene. These samples comprised one individual with one copy, one individual with two copies, three individuals with three copies, and one individual with four copies (samples and data provided by Ms R Soemedi and Professor S Pearce, Institute of Human Genetics, Newcastle upon Tyne).

- Five samples provided by the Northern Regional Genetics Service that had been characterised for CNV of the *PMP22* gene that is associated with Hereditary Motor and Sensory Neuropathy (Charcot-Marie-Tooth disease). These samples comprised one individual with one copy, one individual with two copies, and three individuals with three copies (samples and data provided by Dr O O'Brian, Northern Regional Genetics Service, Newcastle upon Tyne).

For each of these sets of control samples, an individual with two copies of the target gene who could be used as a normalisation reference was identified, but the other samples were analysed with blinding to the number of copies expected. The actual number of copies present was revealed after analysis of the samples with custom MLPA.

5.4.2 Genotyping

Genotyping was performed by Sequenom and TaqMan methodologies using the standard protocols outlined in Chapter 2.

5.4.3 HWE and D-statistic analysis

The extent of deviation from HWE proportions was estimated for each SNP by calculation of the deviation statistic (D-statistic) in Excel according to the formula $D = H - 2p(1-p)$, where H is the observed heterozygosity and p the minor allele frequency. Using this formula, negative D-values correspond to an excess of homozygosity, and positive values to an excess of heterozygosity. The significance of the deviation from HWE proportions (which tests the likelihood of D=0 versus D≠0) was estimated by standard formulae using Excel or Haploview for datasets of unrelated individuals[296-298]. For datasets containing related individuals, such as the HTO cohort, PEDSTATS[295] was used to estimate the significance of the deviation in the total dataset and in the unrelated founder members only, to account for deviations caused by the family-structure.

The HWE D-statistics and P-values were calculated for chromosome 9p21 SNPs genotyped in the study populations outlined above, and also in other large populations for which genotyping data in this region were available from a review of the literature and online databases. Details of the other cohorts for which data were available are shown in Table 5.1. Genotype data were obtained from the NCBI dbSNP website[42] for the 'HapMap CEU' and 'AFD/other' cohorts, and from the individual sources for the other cohorts, as referenced in Table 5.1.

Table 5.1. Other cohorts with chromosome 9p21 SNP genotyping data reported.

Cohort	N	Ethnicity	Details
PROCARDIS controls[129]	4443	Caucasian	European ancestry controls self-reportedly free of CAD at age 65.
1958 Birth Cohort[367]	1500	Caucasian (>96%)	Cross-sectional study of all individuals born during one week in March 1958.
UK Blood Service cohort[368]	1500	Caucasian	Healthy UK blood donors.
HapMap CEU cohort[47]	60	Caucasian	Utah residents with Northern and Western European ancestry from the CEPH collection.
AFD/other Caucasian cohorts[42]	24 / variable	Caucasian	Individuals of European American descent.

5.4.4 MLPA

Analysis for CNV using MLPA was performed according to the methodology described in Chapter 2. Population normalisation was used, except for the control experiments investigating copy number in the small number of patients with known deletions/duplications, which used internal control probe normalisation relative to an individual known to have two copies of the target gene. Data were analysed with GeneMarker v1.8 software using MLPA ratio analysis, with deletions defined as ratios <0.75 and duplications as ratios >1.3 unless otherwise stated.

A core region (chromosome 9: 21,990,300-22,117,096) where the highest proportion of SNPs showed departure from HWE was selected for MLPA analysis (based on the SNP 'heatmap' presented in Figure 5.4 on page 202). Using custom MLPA with the P200 reference probeset, a maximum of 13 custom probes can be designed into each assay. Ten probes were spaced evenly throughout the core region of interest to provide a 'screen' for CNV, with a view to performing more detailed mapping of any significant CNVs that were discovered using additional MLPA assays with more closely spaced probes. An additional probe was placed in a different region that showed HWE deviations for a group of SNPs in the HapMap CEU samples only (chromosome 9: 21,854,535-21,921,896). One probe was placed in an adjacent region where SNPs did not deviate from HWE in any cohort (chromosome 9: 21,807,158). The approximate location of the chromosome 9p21 MLPA probes is highlighted in blue on Figure 5.4 (page 202).

To demonstrate the ability to detect CNV using the custom MLPA analysis, control probes were also designed in two different regions known to show common CNV in Caucasian populations (*DEFA3* on chromosome 8p23[369, 370], and *OR4K2* on chromosome 14q11.2[371]), and for three genes for which control samples with previously characterised CNVs were available (*CLDN5* on chromosome 22q11, *PMP22* on chromosome 17p12, and *DUSP22* on chromosome 6p25). The two MLPA assays used are summarised in Table 5.2, and MLPA probe sequences can be found in Appendix 1.

Table 5.2. Location of probes used in the MLPA assays.

MLPA assay 1*			MLPA assay 2 §		
Probe	**Location**	**Length (nt)**	**Probe**	**Location**	**Length (nt)**
9p21_01	9p21 control region 21807158	92	9p21_01	9p21 control region 21807158	92
8p23_*DEFA3*	*DEFA3* CNV 6863598	96	8p23_*DEFA3*	*DEFA3* CNV 6863598	96
9p21_02	9p21 H region 21892247	100	9p21_02	9p21 H region 21892247	100
9p21_03	9p21 core region 22027375	104	14q11_*OR4K2*	*OR4K2* CNV 1344825	104
9p21_04	9p21 core region 22038430	108	9p21_04	9p21 core region 22038430	108
9p21_05	9p21 core region 22048364	112	22q11_ *CLDN5*	*CLDN5* CNV 2667283	112
9p21_06	9p21 core region 22054243	116	9p21_06	9p21 core region 22054243	116
9p21_07	9p21 core region 22064638	120	9p21_07	9p21 core region 22064638	120
9p21_08	9p21 core region 22074944	124	17p12_ *PMP22*	*PMP22* CNV 14743013	124
9p21_09	9p21 core region 22087232	128	9p21_09	9p21 core region 22087232	128
9p21_10	9p21 core region 22094524	132	6p25_ DUSP22	*DUSP22* CNV 260236	132
9p21_11	9p21 core region 22104164	136	9p21_11	9p21 core region 22104164	136
9p21_12	9p21 core region 22114749	140	9p21_12	9p21 core region 22114749	140

* MLPA assay 1 used for 9p21 CNV assessment. § MLPA assay 2 used to check ability to detect CNVs in controls. nt = nucleotides, H region = region deviated in HapMap samples only (as shown on Figure 5.4). Probe locations are given as a general description along with the specific chromosome location of the first nucleotide of the left-hand probe hybridising sequence.

5.5 Results

5.5.1 Deviation from Hardy-Weinberg proportions and genotyping checks

5.5.1.1 Genotyping checks

As presented in Chapter 4, a higher proportion of SNPs in the NE Caucasian cohort showed significant deviations from HWE than would be expected by chance alone: 12/53 SNPs (23%) at the P<0.05 threshold, and 2/53 SNPs (4%) at the P<0.01 threshold. However, using the same assays in the SA cohort, only 1/56 SNPs (1.7%) deviated from HWE at the P<0.05 threshold. Since correspondence of genotypes to the proportions expected under HWE is commonly used as a screen for genotyping errors, a number of steps were taken to investigate this possibility.

Review and repeat analysis of Sequenom genotyping calls

The correspondence of genotypes in the SA cohort to the proportions expected under HWE, which was performed using the same assays and reaction conditions, suggested that the deviations observed in the NE Caucasian cohort were not due to systematic bias in the assays. The results of genotyping in the Caucasian cohort were therefore carefully reanalysed. Sequenom spectra, cluster plots, and genotype call rates were manually reviewed by the operator (MSC) and appeared generally satisfactory, with changes made to few individual sample genotypes. Data processing steps were then repeated to exclude errors during data handling. These steps did not alter the results of the analysis, with the same SNPs still deviating from the proportions expected under HWE. The raw Sequenom data and analysis were then independently reviewed by experienced Sequenom technicians (Dr D Hall and Mr J Eden, Institute of Human Genetics, Newcastle upon Tyne) and judged to be satisfactory.

To check the consistency of genotyping calls and exclude errors due to sample switching, a subgroup of samples was re-genotyped using the same Sequenom assays. 14 SNPs selected to include those that most significantly deviated from HW proportions were re-genotyped (rs3088440, rs10965215, rs10738605, rs3217992, rs1063192, rs10116277, rs10757274, rs10757278, rs1333040, rs1333049, rs2383206, rs496892, rs7044859, rs7865618). Repeat genotyping was performed using WGA DNA from a panel of 64

samples, which during preliminary analysis as a subgroup had shown significant deviation from HWE for some of the SNPs. A total of 896 genotyping comparisons were made (14 SNPs in 64 samples) and were analysed blinded to the results of the original genotyping. A difference in genotype call was found for only one sample at one SNP, which did not significantly alter HWE proportions. This demonstrated that the Sequenom genotyping assays were reproducible, and the same in native DNA (used for original genotyping) and WGA DNA (used for the repeat analysis). These analyses suggested that genotyping errors in the Sequenom assays were unlikely to account for the observed deviations from HWE.

Comparison with TaqMan genotyping results

In the association studies presented in Chapter 6, four of the chromosome 9 SNPs had been previously genotyped in the HTO population using the TaqMan platform: rs7044859, rs496892, rs7865618, rs1333049. Although these SNPs did not deviate significantly from the proportions expected under HWE in either the NE Caucasian samples (Sequenom) or the HTO cohort (TaqMan), genotypes from both methods were available for 90 unrelated individuals of the HTO population, which had been used to optimise the Sequenom assay. Comparison of 360 genotypes for these individuals (4 SNPs in 90 samples) showed no inconsistencies in genotype calls, suggesting that Sequenom genotyping was accurate for those SNPs.

To check the accuracy of genotyping for SNPs which deviated from HWE proportions in the NE Caucasian samples, the SNP showing the most significant deviation (rs10116277) was retyped in the same population using a pre-designed TaqMan assay (C__29991625_20). Genotype calls at this SNP were obtained for 187 samples using Sequenom and 169 samples using TaqMan, as summarised in Table 5.3. Only three samples showed genotyping inconsistencies between the two techniques, with no systematic bias in the genotype calls (SeqGG/TaqGT, SeqGT/TaqTT, SeqGT/TaqGG). The 18 genotypes with missing TaqMan genotypes did not differ from the proportions observed in the rest of the population using Sequenom genotyping (GG 7, GT 6, TT 5).

Table 5.3. Comparison of Sequenom and TaqMan genotyping calls for rs10116277.

Genotypes	TaqMan	Sequenom
GG	52	59
GT	61	68
TT	56	60
Absent	18	0
HWE P-value	0.0003	0.0002

Taken together, the above data suggest that genotyping deviations from HWE proportions for SNPs in the NE Caucasian cohort were not due to genotyping errors.

5.5.1.2 D-statistic analysis and comparison with other cohorts

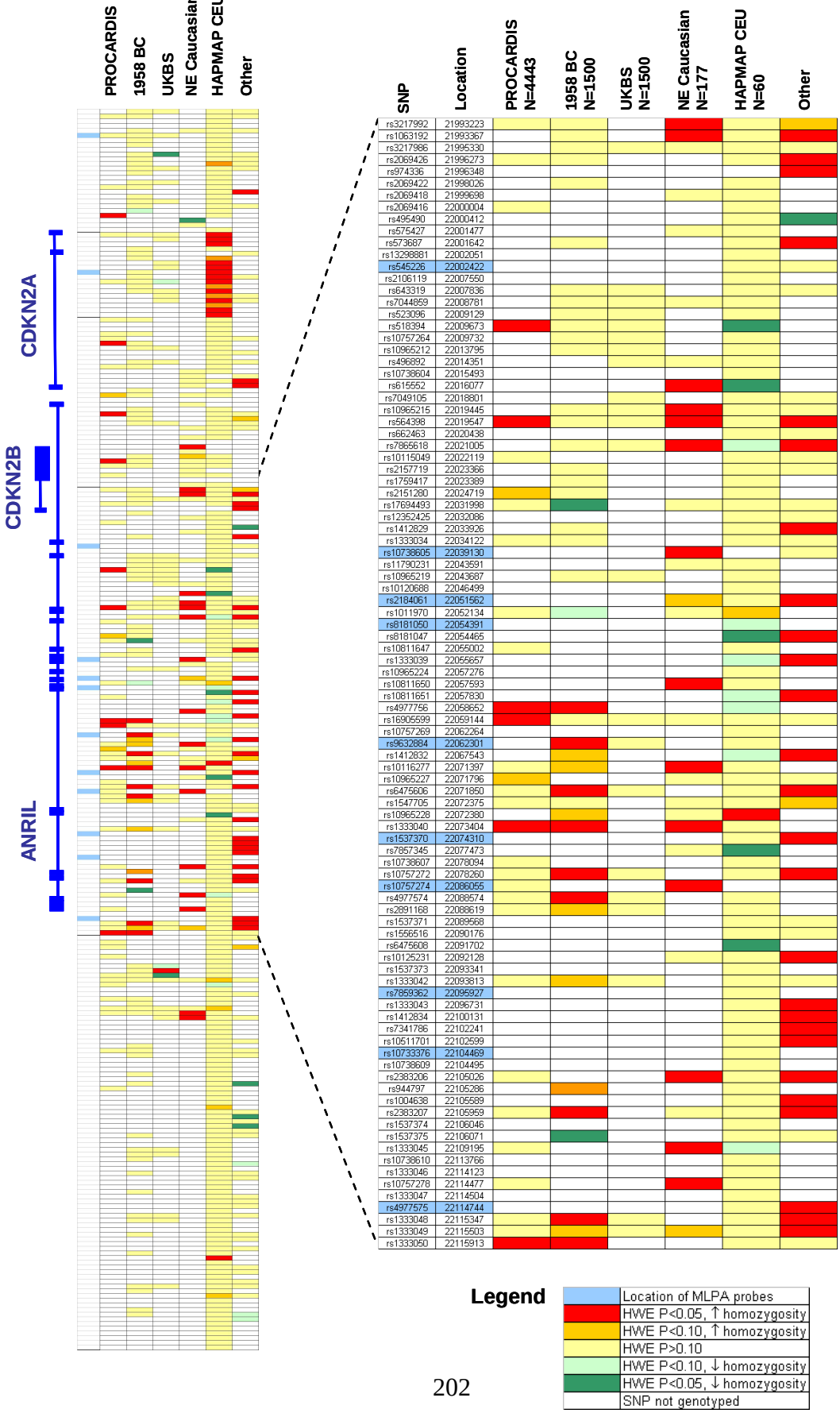
As discussed above, deletions in the region could cause departure from HWE, in which case SNPs showing deviations would be expected to have an excess of homozygotes (for the reasons previously presented in Figure 5.1 on page 189). Analysis of the D-statistics for SNPs showing HWE deviations in this study demonstrated that most of the SNPs showing deviations from proportions expected under HWE had an excess of homozygosity consistent with the presence of common deletions in the region.

The genotyping data from other large cohorts where multiple SNPs in this chromosome 9p21 region had been genotyped were analysed to determine whether a higher than expected proportion of SNPs in these studies also deviated from the proportions expected under HWE assumptions. HWE P-values and D-statistics were calculated for SNPs in these studies to determine whether the pattern suggested an excess of homozygosity consistent with the presence of a null allele.

As shown in Figure 5.4, other cohorts also showed deviations from HWE proportions in the region of interest, which were mostly attributable to an excess of homozygosity. A core region (chromosome 9: 21,990,300-22,117,096) highlighted in Figure 5.4 was identified where SNPs showed deviations in multiple Caucasian cohorts at a higher rate than was observed in surrounding regions. This suggested that the rate of deviation in this region was different to the 'background' deviation rate (for example as observed for SNPs in regions shown at the top and bottom of Figure 5.4).

Figure 5.4. Heatmap of SNP departures from HWE in Caucasian cohorts.

Different cohorts are shown in columns. Each row represents a SNP and colours show the extent and direction of HWE deviation, as shown in the legend. A core region showing greatest HWE departure which was used for MLPA analysis is magnified. The approximate location of MLPA probes is illustrated by highlighting the closest SNP to each probe in blue (all MLPA probes were designed not to overlie SNPs).



It must be noted that due to LD between SNPs these tests were not independent and therefore a deviation from HWE caused by a particular haplotype/mutation may be expected to affect SNPs in LD in a similar fashion, such that the high proportion showing deviation from HWE might be considered a less dramatic finding than if this was seen at multiple unrelated sites. However, this would not account for the observation that the same region showed deviations in multiple independent cohorts. Of particular interest, the core region showing deviation from HWE contained the SNPs associated with CAD in GWA studies, and completely overlapped the *ANRIL* gene, expression of which is associated with variants correlated with disease (see Chapter 4). This raised the interesting possibility that common deletions involving *ANRIL*, leading to HWE deviations in the region of that gene, might be causally associated with disease, and that SNP associations observed in GWA studies may be due to SNPs 'tagging' the CNV. Further investigation to look for evidence of common null alleles in this region in Caucasian cohorts was therefore performed.

Table 5.4. Summary of SNPs in the core region of deviation that differed from expected HWE proportions in each cohort.

Cohort	Number (%) of SNPs in the core region of deviation with P<0.05 for correspondence to HWE	Number (%) of deviating SNPs showing excess homozygosity
PROCARDIS controls[129]	6/31 (19%)	6/6 (100%)
1958 Birth Cohort[367]	11/42 (26%)	9/11 (82%)
UK Blood Service cohort[368]	0/22 (0%)	NA
NE Caucasian cohort	14/31 (45%)	14/14 (100%)
HapMap CEU cohort[42]	6/92 (7%)	1/6 (17%)
Other Caucasian cohorts[42]	27/48 (56%)	26/27 (96%)

5.5.2 Investigation of deviations from HWE and evidence for null alleles in a family cohort

5.5.2.1 HWE analysis

Since other reported Caucasian datasets had also shown evidence of deviations from proportions expected under HWE for SNPs in the chromosome 9p21 region, it was decided to investigate this further in the HTO cohort. This had two main advantages:

- It allowed investigation of whether the observed deviation in HWE proportions was replicated in a larger Caucasian cohort.

- Because the HTO cohort was composed of related family members, it allowed investigation for a possible 'null allele' (resulting from heterozygous allele deletion) by studying allele segregation in families with Mendelian errors in inheritance. This approach could detect deletions that were too rare to account for deviation from HWE proportions and identify samples that could be tested for presence of a null allele using MLPA.

17 SNPs that showed deviations from the proportions expected under HWE in either the NE Caucasian cohort or one of the other cohorts analysed were selected for genotyping in the HTO cohort and redesigned into a single Sequenom assay. As shown in Table 5.5, only a single SNP (rs564398) deviated significantly from the proportions expected under HWE in this cohort including all individuals, and no SNPs showed a significant deviation when only the founders were considered (to account for deviations that could be attributed to the family structure of the population). This did not support the hypothesis that a common null allele in Caucasian populations might explain the deviations observed for these SNPs in the other cohorts.

Table 5.5. SNP genotyping results in the HTO cohort.

SNP	Chr 9p21 position	Alleles	NE Caucasian cohort			HTO cohort			
			% Genotyped	MAF	HW P-value*	% Genotyped	MAF	HW P-Value all	HW P-Value founders
rs1063192	21993367	T:C	98.9	0.42	0.02	99.5	0.47	0.32	0.18
rs615552	22016077	A:G	95.5	0.41	0.02	99.0	0.47	0.13	0.22
rs10965215	22019445	A:G	100.0	0.46	0.01	99.5	0.45	0.38	0.32
rs564398	22019547	A:G	97.7	0.38	0.02	91.9	0.44	0.00003	0.61
rs11790231	22043591	G:A	96.6	0.12	0.96	99.4	0.10	0.37	0.53
rs2184061	22051562	A:C	93.2	0.38	0.10	93.4	0.43	0.08	0.80
rs10811650	22057593	C:G	95.5	0.47	0.001	99.0	0.40	0.95	0.31
rs10116277	22071397	T:G	100.0	0.49	0.0002	99.0	0.44	0.91	0.46
rs1333040	22073404	T:C	100.0	0.36	0.02	99.4	0.43	0.96	0.71
rs7857345	22077473	C:T	93.2	0.27	1.00	99.8	0.30	0.48	0.47
rs10757274	22086055	G:A	100.0	0.49	0.01	96.9	0.45	0.82	0.90
rs10125231	22092128	G:A	94.9	0.02	1.00	93.6	0.03	0.05	1.0
rs2383206	22105026	G:A	100.0	0.47	0.01	98.5	0.46	0.91	0.39
rs2383207	22105959	G:A	90.4	0.42	0.33	99.6	0.46	0.74	0.38
rs1333045	22109195	C:T	96.6	0.48	0.05	98.7	0.48	0.74	0.62
rs10757278	22114477	A:A	99.4	0.50	0.05	97.8	0.44	0.74	1.0
rs1333049	22115503	C:G	100.0	0.50	0.08	99.7	0.44	0.87	1.0

MAF = minor allele frequency; HW = Hardy-Weinberg. * Included SNPs which have HW P-values >0.05 in the NE Caucasian cohort showed significant deviations in other populations analysed.

5.5.2.2 Investigation for Mendelian errors consistent with segregation of a null allele

PEDSTATS software[295] was used to identify 78 Mendelian inheritance errors in 20 families. Genotyping spectra and cluster plots were manually reviewed for parent-offspring trios with Mendelian errors, and samples with low quality spectra and borderline calls likely to represent genotyping miscalls were excluded. The remaining samples were then re-genotyped, leaving 38 Mendelian errors in 16 individuals from 13 families. Pedigrees of these individuals were examined to determine whether the observed patterns were consistent with a null allele, as summarised in Table 5.6.

Table 5.6. Mendelian errors consistent with presence of a null allele.

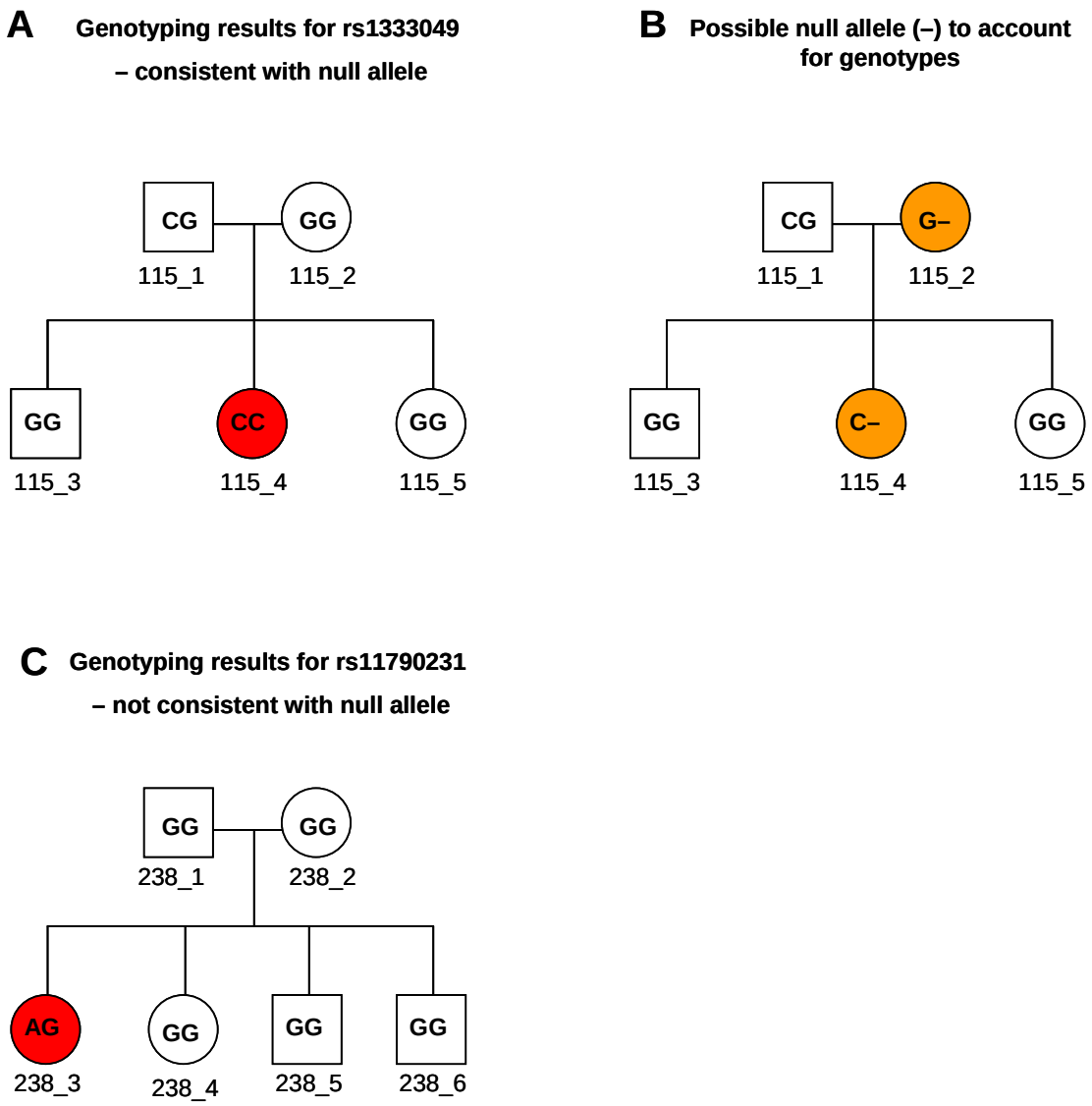
Sample ID (Family_individual)	Number of Mendelian errors in SNPs consistent with null allele	Number of Mendelian errors in SNPs not consistent with null allele
36_9	6	0
104_8	1	0
115_4	1	0
150_6	1	0
155_5	1	0
155_7	1	0
190_3	9	0
118_5	1	1
202_3	3	1
107_4	0	1
118_4	0	1
118_5	0	2
118_6	0	1
124_5	0	2
170_13	0	2
238_3	0	1

For seven individuals the pattern of genotyping in the pedigree was consistent with the presence of a null allele, i.e. one of the parents and the offspring were homozygous for different alleles. Two further individuals had Mendelian errors consistent with a null allele at some SNPs, but other SNPs where the pattern was not consistent with a null allele. In the remaining seven individuals the pattern was inconsistent with a null allele at all SNPs; this could result from sample switching/contamination, genotyping miscalling, or mistaken paternity. Pedigrees illustrating examples of the different patterns observed are shown in

Figure 5.5.

Figure 5.5. Examples of pedigrees with Mendelian errors consistent or inconsistent with a null allele.

Family pedigrees with genotypes and sample identifiers shown below. (A) Mendelian error in individual 115_4 (red) consistent with presence of a possible null allele. (B) Possible null allele in 115_2 and 115_4 (orange) that could explain the genotyping results. (C) Mendelian error in individual 238_3 (red) that could not be attributed to the presence of a null allele.



5.5.3 MLPA results

5.5.3.1 Effect of DNA amplification on MLPA analysis

The NE Caucasian cohort showed the most significant deviations from HWE and it was therefore intended to investigate the possibility of CNV in that population using MLPA. However, DNA for most individuals in that population was only available in limited amounts from the RNA extraction. MLPA ideally requires 50-200ng of high quality DNA since impurities and poor quality DNA can potentially influence MLPA results. Exploratory experiments were performed to determine whether the RNA/DNA solution from the RNA extraction could be used for analysis. MLPA was performed for four samples, using 1μL and 3μL of the RNA/DNA solution as template for each. These samples showed unreliable results, with 'scattering' of the control probes and complete failure of one sample, as shown in Figure 5.6. Separate DNA samples were available for a small number of samples in this cohort, which were analysed using MLPA. As shown in Figure 5.7, the 12 samples which had acceptable MLPA data showed no evidence of CNV in the chromosome 9p21 region.

Figure 5.6. Suboptimal MLPA results from DNA in the RNA extraction eluate for the NE Caucasian samples.

The 13 points on the left-hand side of each plot represent the custom probes for chromosome 9p21 (except for point 2 which is a control in *DEFA3*), and the 14 points on the right-hand side represent the P200 control probes (with control probe 7 on the X chromosome and control probe 12 on the Y chromosome). Points outside of the parallel lines indicate CNVs. Plots show MLPA results in the same four individuals using 1µL and 3µL of RNA/DNA solution, and illustrate poor quality data that could not be used for reliable CNV assessment. For comparison, see Figure 5.7 which shows satisfactory quality plots, with no deviation of the control probes.

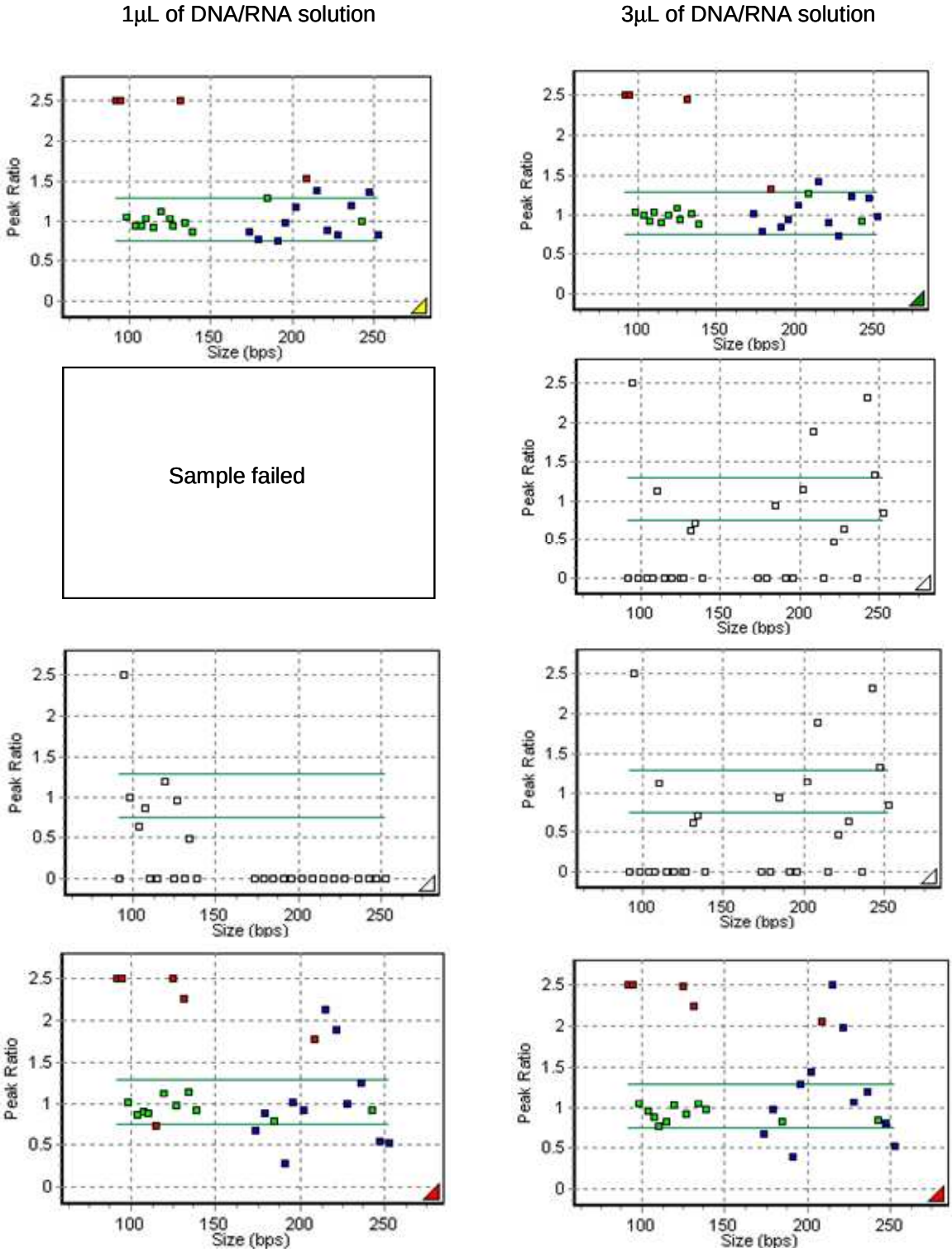
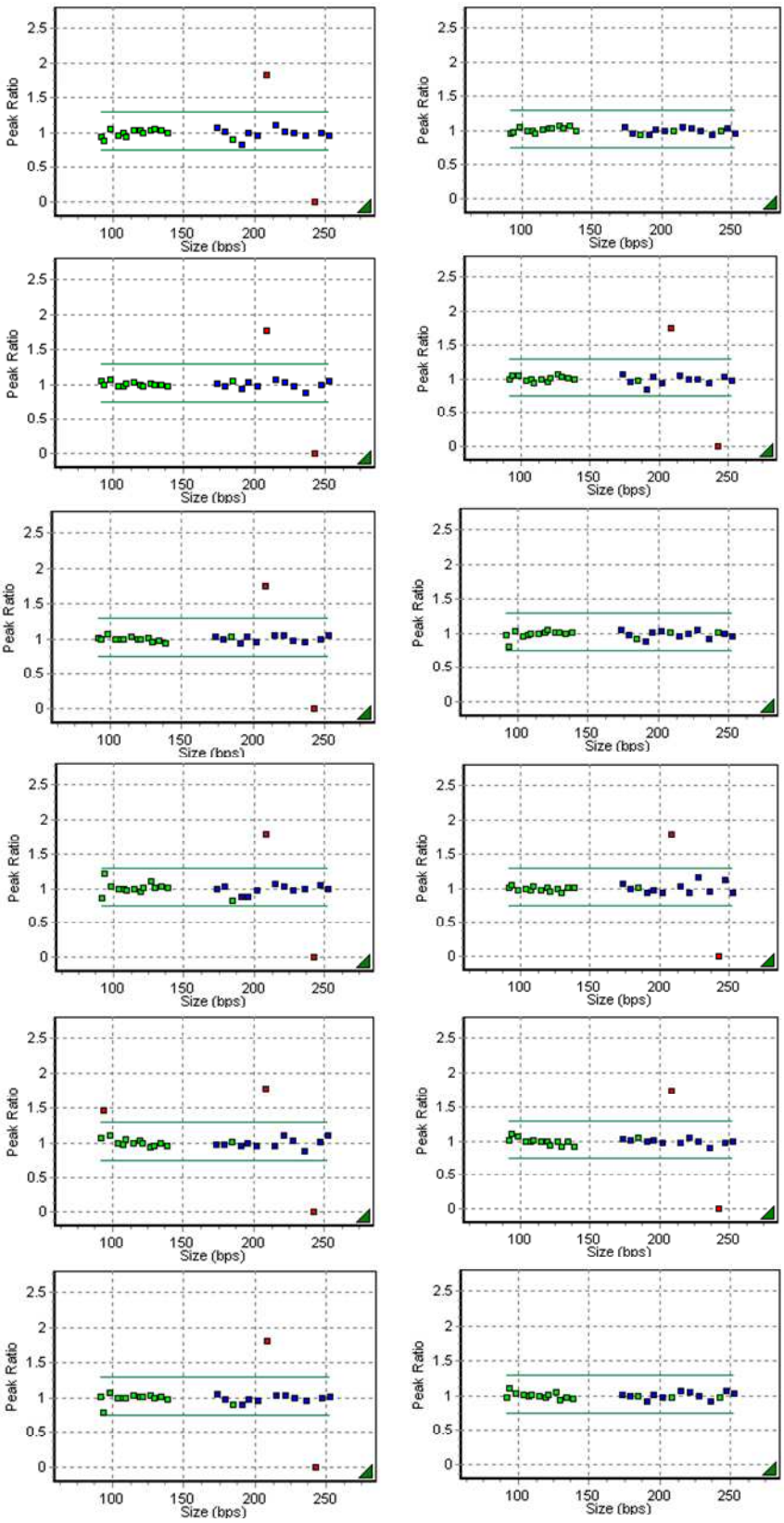


Figure 5.7. MLPA results showing no evidence of CNV in the chromosome 9p21 region in DNA samples from the NE Caucasian cohort.

The 13 points on the left-hand side of each plot represent the custom probes for chromosome 9p21 (except for point 2 which is a control in *DEFA3*), and the 14 points on the right-hand side represent the P200 control probes (with control probe 7 on the X chromosome and control probe 12 on the Y chromosome). Points outside of the parallel lines indicate CNVs. None of these samples showed chromosome 9p21 CNV.



Since amplified DNA was available for all of the NE Caucasian samples, an experiment was performed to investigate whether MLPA detection of CNV could be performed using amplified DNA. As shown in Figure 5.8, MLPA results derived from paired samples of native DNA and WGA DNA from the same individuals performed on the same plate using the same reaction mastermix were substantially different in the 10 samples, indicating that WGA DNA could not be used for reliable assessment of CNV. WGA samples showed apparent copy number variations for multiple probes, including the control probes, which were not present in the native DNA, suggesting that the amplification process amplified some regions of DNA more efficiently than others.

5.5.3.2 Custom MLPA in the chromosome 9p21 region in additional Caucasian samples

In addition to the NE Caucasian samples, MLPA analysis was also performed in 106 unrelated individuals from the HTO cohort to investigate the possibility of low frequency CNV which did not result in deviation from the proportions expected under the assumptions of HWE. Using the standard MLPA ratio thresholds of <0.75 or >1.3, only one sample showed a deviation from two copies (zero copies of probe 9p21_01), and this sample was of borderline quality with some deviation in the control probes, suggesting that it was most likely to be due to artefact. With the deletion threshold of 0.75, a number of samples showed borderline evidence for deletion at the *DEFA3* probe. Increasing the lower threshold to 0.8 resulted in samples that looked suggestive for *DEFA3* deletion becoming significant and led to only one additional chromosome 9p21 probe deviation (one copy of probe 9p21_04). The results of the analysis using the deletion threshold of 0.8 are shown in Table 5.7. Overall, there was no strong evidence for CNV for any of the probes in 118 Caucasian samples tested. This suggested that CNV was not common in this region in Caucasian cohorts, and was unlikely to account for the deviation from HWE proportions that was observed in the genotyping of the NE Caucasian cohort.

Figure 5.8. MLPA results in native DNA versus WGA DNA from the same individuals.

The 13 points on the left-hand side of each plot represent the custom probes for chromosome 9p21 (except for point 2 which is a control in *DEFA3*), and the 14 points on the right-hand side represent the P200 control probes (with control probe 7 on the X chromosome and control probe 12 on the Y chromosome). Points outside of the parallel lines indicate CNVs. Results in gDNA and WGA DNA are shown for the same samples. WGA led to unreliable CNV assessments.

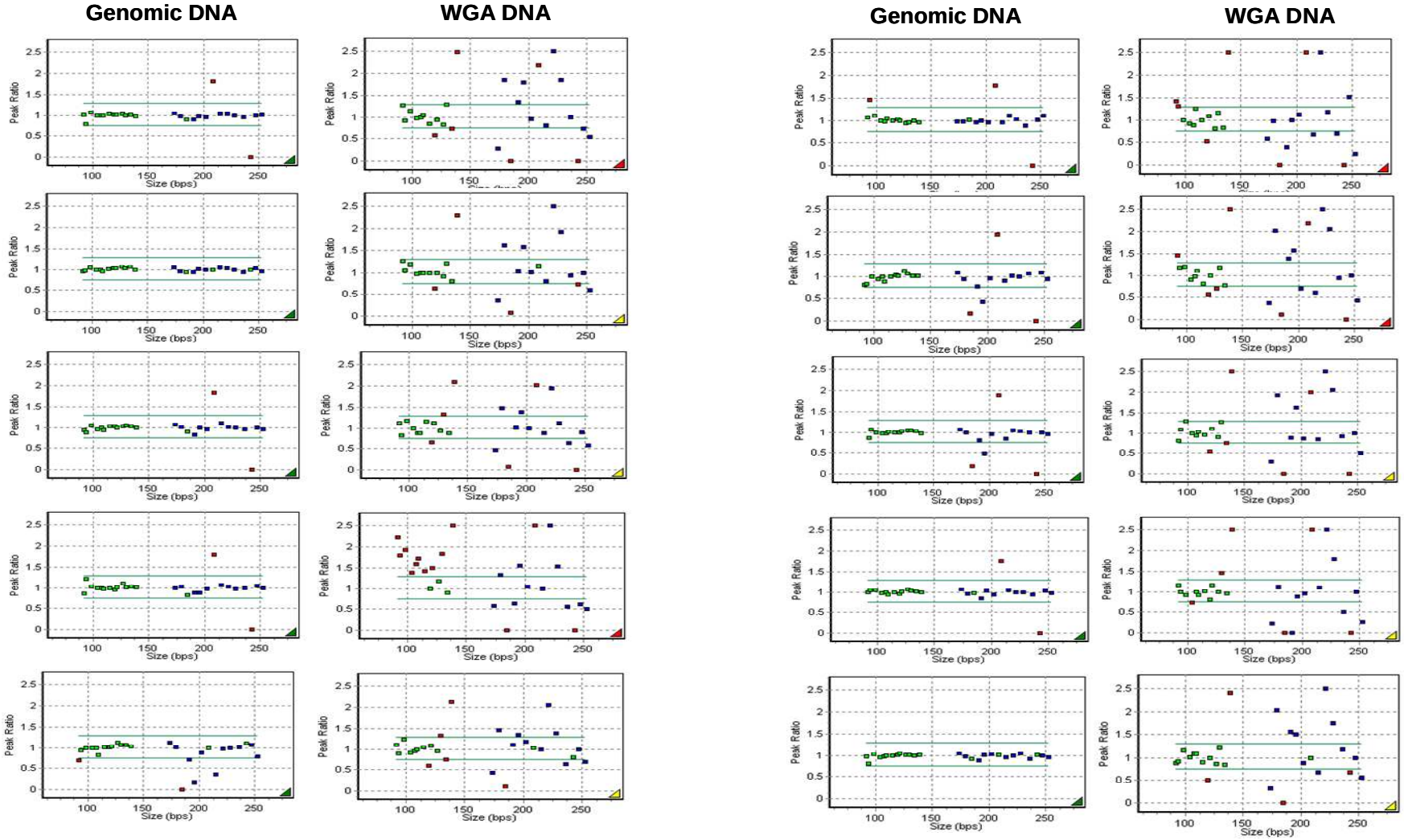


Table 5.7. MLPA results at the chromosome 9p21 locus and two loci known to show CNV.

MLPA probe	Number of copies				
	0	**1**	**2**	**3**	**4**
9p21_01	1**	0	118	0	0
9p21_02	0	0	118	0	0
9p21_03	0	0	75*	0	0
9p21_04	0	1[§]	118	0	0
9p21_05	0	0	75*	0	0
9p21_06	0	0	118	0	0
9p21_07	0	0	118	0	0
9p21_08	0	0	75*	0	0
9p21_19	0	0	118	0	0
9p21_10	0	0	75*	0	0
9p21_11	0	0	118	0	0
9p21_12	0	0	118	0	0
8p23_*DEFA3*	1	33	78	5	1
14q11_*OR4K2*	0	1	23	17	2

Data shown using the MLPA ratio threshold of 0.8 to define deletion. * Fewer samples for probes replaced with alternative probes in assay-2 used to test for known CNV controls in different regions in a subset of 43 samples. ** Deviation in sample of borderline quality, with some deviation in control probes, likely spurious. [§] Two copies using the MLPA threshold of 0.75.

However, since this was a custom assay investigating for the presence of potential CNVs that had not been previously characterised, there were no samples with confirmed deletions that could be used as positive controls to test the sensitivity of the custom MLPA analysis for the detection of true CNVs. To demonstrate that assays designed and analysed using this methodology were capable of detecting CNVs, custom probes were designed for two genes previously reported to display common CNV in human populations (*DEFA3* on chromosome 8p23[369, 370], and *OR4K2* on chromosome 14q11.2[371]), and for three genes for which control samples with previously characterised CNVs were available (*CLDN5* on chromosome 22q11.21, *PMP22* on chromosome 17p12, and *DUSP22* on chromosome 6p25.3).

As shown in Table 5.7, CNVs were detected in the population using the custom *DEFA3* and *OR4K2* probes. The observed *OR4K2* duplication prevalence was 19/43 (44%), which was similar to the prevalence of 37% reported previously[371]. Similarly,

the observed CNV rates for *DEFA3* were broadly comparable to published data[369, 370], as shown in Figure 5.9. The custom assays also accurately quantified the known CNVs from patient samples at the other loci studied, as shown in Table 5.8. The two instances where copy number estimation did not exactly correlate with the known number of copies occurred in reactions with considerable scatter of the control probes, indicating suboptimal quality of the assay.

Figure 5.9. Comparison of CNV detected in *DEFA3* gene with published data[370].

The graph shows the proportion individuals in the population (Y-axis) with different numbers of copies of the gene (X-axis) observed in the current study (black) and a previously published series (grey). Results were similar in the two populations

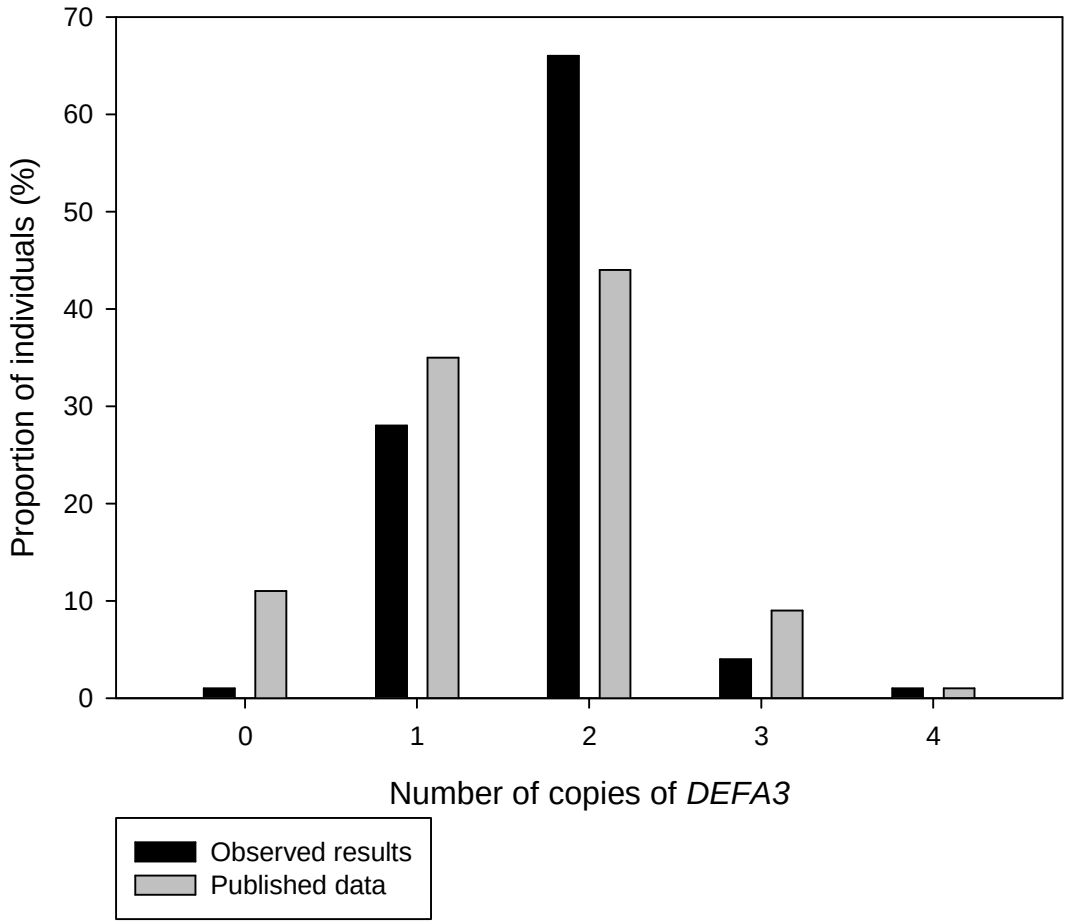


Table 5.8. Copy number variation detected by custom MLPA compared to actual number of copies present for three different loci.

CNV region	Number of copies in sample	Number of copies detected with MLPA
22q11	1	1
	1	1
	1	1
	2	2
	2	2
DUSP22	1	2
	2	2*
	3	3
	3	3
	3	3
	4	4
HMSN	1	1
	2	2*
	3	2
	3	3
	3	3

* Sample used as reference to normalise MLPA assay. Shaded rows highlight samples where the custom MLPA probe did not identify the expected number of copies in the sample (however in both of these cases the control probes were fairly scattered, indicating that the individual reaction or sample may have been suboptimal).

These data indicate that the custom MLPA was able to accurately detect and quantify CNVs. As shown in Table 5.9, in 118 samples screened the probability of not observing a null allele of a frequency that could account for the magnitude of HWE departure observed in the NE Caucasian cohort was small for all SNPs in the MLPA region. The study was therefore adequately powered, suggesting that common CNVs in this region do not account for the observed departure from HWE and are unlikely to account for the associations of common SNPs with phenotypes observed in GWA studies.

Table 5.9. Probability of missing a null allele that could account for the observed magnitude of HWE departure in 118 Caucasian samples for SNPs in the MLPA region.

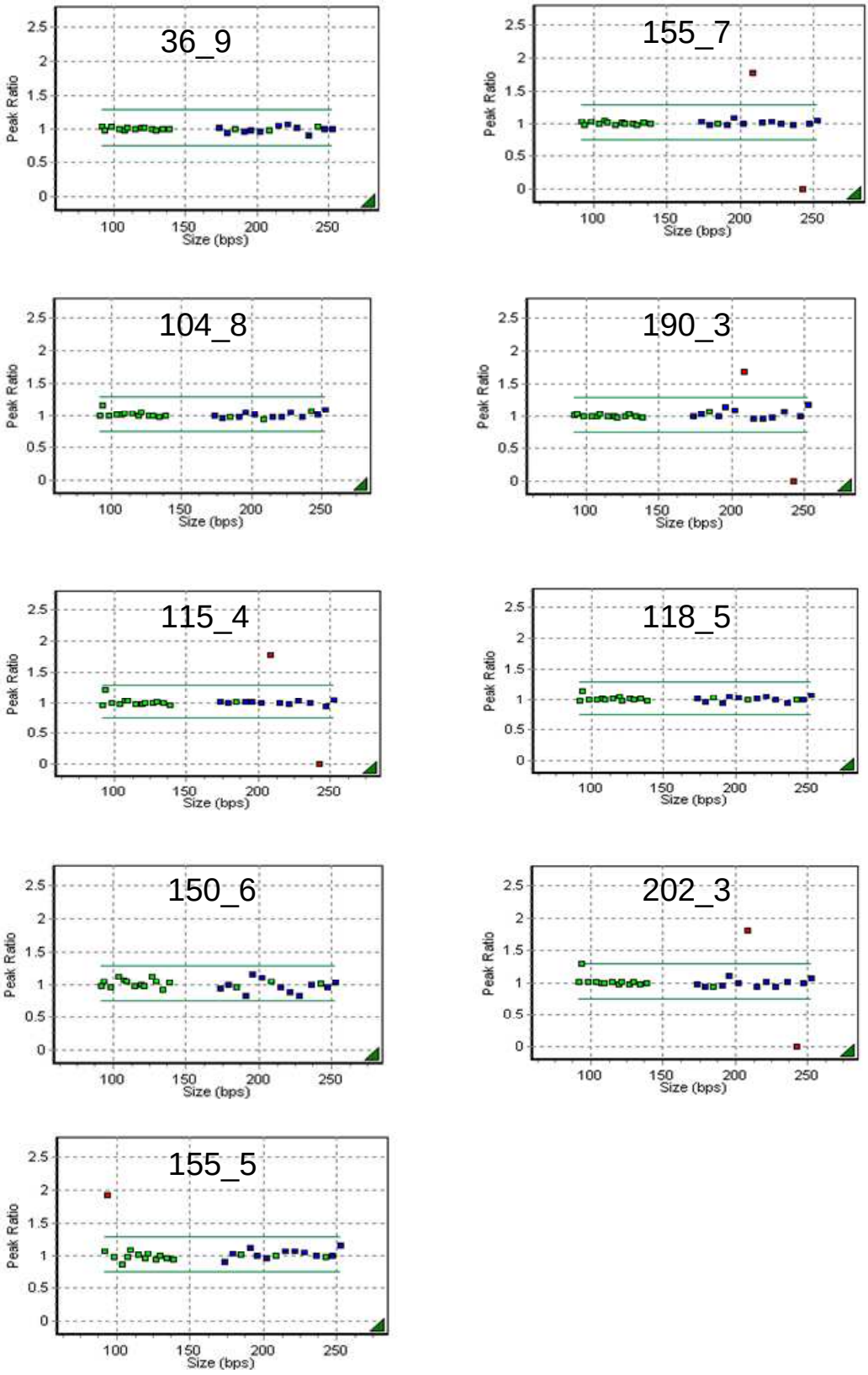
SNP	HWE P-value in NE Caucasian cohort	Frequency of null allele that could account for this magnitude of HWE departure	Probability of not observing a null allele of this frequency in 118 individuals
rs3217992	0.02	0.10	$2.2x10^{-11}$
rs1063192	0.02	0.10	$1.8x10^{-11}$
rs615552	0.01	0.11	$3.0x10^{-12}$
rs10965215	0.01	0.11	$4.5x10^{-13}$
rs564398	0.02	0.10	$1.5x10^{-11}$
rs7865618	0.01	0.10	$1.0x10^{-11}$
rs10738605	0.02	0.10	$2.0x10^{-11}$
rs10811650	0.00	0.14	$1.1x10^{-16}$
rs10116277	0.00	0.16	$1.1x10^{-18}$
rs1333040	0.01	0.10	$6.6x10^{-12}$
rs10757274	0.01	0.12	$1.8x10^{-13}$
rs2383206	0.01	0.11	$1.7x10^{-12}$
rs1333045	0.04	0.09	$7.4x10^{-10}$
rs10757278	0.03	0.09	$5.0x10^{-10}$

5.5.3.3 MLPA analysis in individuals with possible null alleles

MLPA analysis was performed for all nine samples (and the relevant parent) whose patterns of Mendelian errors were consistent with null allele segregation in some or all of the SNPs (as presented in section 5.5.2.2 and Table 5.6). As shown in Figure 5.10, MLPA in the chromosome 9p21 region did not show evidence for deletions or other CNVs in any of these individuals.

Figure 5.10. MLPA results in the nine individuals with Mendelian inheritance errors consistent with null allele segregation within the pedigree.

The 13 points on the left-hand side of each plot represent the custom probes for chromosome 9p21 (except for point 2 which is a control in *DEFA3*), and the 14 points on the right-hand side represent the P200 control probes (with control probe 7 on the X chromosome and control probe 12 on the Y chromosome). Points outside of the parallel lines indicate CNVs. None of the individuals demonstrated chromosome 9p21 CNVs.



5.6 Discussion

SNP genotyping in the NE Caucasian cohort showed departures from HWE with an excess of homozygosity at a higher rate than expected by chance alone. Analysis of HWE in this study was performed as a check for genotyping errors, which is consistent with widespread practice in other genotyping studies[129, 344], although some have suggested that HWE testing should not be performed since real-life populations are never exactly in HWE[343] and it is not a sensitive method for detecting genotyping errors[347]. Extensive genotyping checks, including repeat analysis by different methodologies, excluded genotyping error as the cause of the observed departure from HWE seen in this population. These investigations confirmed that the genotyping performed in this study was highly accurate.

Although the excess of homozygosity observed for SNPs showing departure from HWE could be consistent with the presence of common deletions in the region, CNV was not detected using custom MLPA in 118 Caucasian samples analysed. However, it must be noted that MLPA was not performed for most samples from the NE Caucasian cohort which showed the highest frequency of departure from HWE because insufficient native DNA was available for these individuals, and it was demonstrated that assessment of copy number could not be reliably performed in WGA DNA. No CNV was detected in the small number of samples that were analysed from this population. Although consistent departures from HWE with an excess of homozygosity in this region have also been observed in other large datasets[129], raising the possibility that common deletions may exist in Caucasian populations, we did not find evidence to support this in the HTO cohort. Using the family-based approach to look for Mendelian errors consistent with null allele segregation should provide greater power to detect deletions present at a frequency below that which would cause departure from HWE, but no deletions were seen even in this subset of individuals. Although this approach provides a method to identify individuals with potential null alleles, it cannot identify null alleles that do not result in Mendelian errors of inheritance, such as in offspring where one parent has a null allele but the other parent is homozygous for the undeleted allele (for example: offspring ‘G-’ from parents and ‘GG’ and ‘G-’ would not appear as a Mendelian error), although the screen of 118 Caucasian individuals suggests that such CNVs are

not common in the population. The errors of Mendelian inheritance observed in this study which are not attributable to CNV may be the result of non-paternity, sample confusion/contamination, or genotyping error. Repeat genotyping should have minimised the possibility of genotyping error in the present study.

Because MLPA was being used for the investigation of unknown CNVs, positive controls and reference samples were not available, raising the possibility that the custom MLPA design was insensitive for the detection of CNVs. Although the chromosome 9p21 custom probes could not be directly validated, custom probes designed and analysed using identical methodology for five separate control regions were accurate in detecting CNVs. In the analysis of amplified DNA which contained copy number artefacts, the custom chromosome 9p21 probes detected significant CNVs that were similar to the patterns seen in the control probes, suggesting that they were able to detect CNVs when present. Taken together, these data along with the low variability observed in the 9p21 probes in native DNA, make it unlikely that the MLPA results represent false negative findings. Previous genomewide surveys of CNV have reported four deletions in this region in three separate individuals, one of which was 200kb in size and the other three of which were all less than 1kb. The chromosome 9p21 probes used in the present study were spaced at approximately 10kb intervals throughout the region of interest, and it is therefore possible that CNVs smaller than 10kb would not have been detected. However, deletions sufficient to account for the departure from HWE observed across a region of this size would have been detected with the spacing of our MLPA probes. The sample size of the Caucasian cohort screened was sufficient to confidently exclude deletions of a frequency that could account for the magnitude of the HWE departure observed in the NE Caucasian cohort. However, the analysis does not exclude the presence of rare variants with significant effects on expression or disease phenotypes, such as those that have been previously reported in pedigrees with familial cancers[195]. The possibility that rare CNVs in this region confer susceptibility to CAD or other disease phenotypes in some individuals therefore cannot be excluded.

This study demonstrated that CNV assessment consistently differed in native DNA and WGA from the same individuals, with WGA DNA showing significant deviations for multiple MLPA probes in each sample that are not present in native DNA. Copy

number artefacts associated with DNA amplification have been previously reported for Phi29 polymerase multiple-strand displacement techniques[372-374], which at least in part relate to the GC content of the region[372]. The Genomeplex kit uses degenerate oligonucleotide primer PCR-based amplification, and has been used to detect both whole chromosome copy number changes [375] and copy number changes involving regions as small as 8.3Mb[376], although the latter study did note increased variability associated with amplified compared to unamplified DNA. A subsequent study by the same group, published after the experiments reported in this chapter were performed, showed that analyses using tiling oligo arrays in Genomeplex amplified DNA could not reliably detect CNVs of less than 3Mb due to the variability introduced by amplification[377]. In the present study, the artefacts associated with Genomeplex amplification were non-random, with systematic deviations occurring for particular MLPA probes in different samples. It may be possible to correct for such systematic biases by normalising to other amplified samples, which has been proposed for analysis of amplified data from multiple-strand displacement techniques[373].

The assumptions of HWE are never exactly met in real populations, and violations of any of these assumptions may cause deviation from the proportions expected, particularly population stratification and selection bias[349]. The cohort was composed principally of individuals from the People of the British Isles study, the ascertainment strategy of which involved recruitment of individuals with three generations of family born within a 40-mile radius of one another in rural populations in NE England (from Northumberland, Tyneside, County Durham, and Cumbria). Effectively, this selected individuals whose family had lived for generations the same village/town, geographically distinct from that of the other members of the cohort. This violated the HWE assumption of random-mating and may have produced stratification of the combined population which might account for the higher incidence of HWE departure observed in this population compared to the HTO cohort, although this would not account for the HWE departure seen in other published cohorts at the same locus. However, LD between SNPs in this region means that tests of correspondence to HWE proportions are not independent for each SNP, and deviation at one SNP is likely to be associated with deviation at other SNPs in strong LD with it. This inflates the proportion of analysed SNPs showing deviations, and could account for the clustering of HWE departures seen within the various cohorts.

In summary, the deviations from HWE that were observed were not due to genotyping error, and MLPA analysis showed no evidence of common CNVs in Caucasian cohorts. It is likely that the deviations observed in the NE Caucasian cohort and other published series can in varying degrees be accounted for by the ascertainment strategy, random chance, and LD between markers.

Chapter 6

Association of 9p21 polymorphisms with cardiovascular phenotypes

6 Association of 9p21 polymorphisms with cardiovascular phenotypes

6.1 Abstract

The mechanisms through which novel CAD risk variants identified by GWA studies influence disease is unknown; their relationship to other cardiovascular phenotypes may give mechanistic clues. Carotid artery intima-media thickness (CIMT) is a subclinical marker of atherosclerosis associated with stroke that can be used to investigate the association of these SNPs with extra-coronary arterial manifestations. Chromosome 9p21 CAD risk variants are also associated with non-atherosclerotic intracranial aneurysms, which are themselves associated with congenital cardiovascular abnormalities, suggesting that risk variants may have more generalised influences on vascular and cardiac structure. This study investigated the association of CAD risk variants with intermediate phenotypes, CIMT, and echocardiographic measures of cardiac structure and function in 1425 members of 248 British Caucasian families ascertained through a hypertensive proband. The association of risk variants with congenital heart disease was also investigated in a case-control study of 888 affected individuals and two control cohorts of 1144 and 1500 unaffected individuals. CAD risk SNPs were genotyped using TaqMan and Sequenom. MERLIN software was used for family-based association testing and case-control association analysis was performed by logistic regression using UNPHASED software. No significant association was found between genotype at any SNP and BP, obesity, cholesterol, CRP, interleukin-6, TNF-α, or leptin. Nor were SNPs significantly associated with CIMT, echocardiographic parameters, or congenital heart disease (at 0.05 significance level corrected for multiple testing). In summary, novel CAD variants do not appear to mediate the risk of atherothrombosis through the known risk factors studied and are not associated with CIMT, LV structure/function or congenital heart disease. Further investigations will be required to determine the mechanisms by which these SNPs are associated with CAD.

6.2 Introduction

A number of common SNPs have been associated with CAD in recent GWA studies,[19-21] but, in contrast to associations derived from candidate gene analyses, the mechanisms underlying the associations with CAD are unknown. A relationship of these variants to intermediate cardiovascular phenotypes may give important mechanistic clues to pathways involved in causation.

Previous studies had demonstrated that CAD risk variants at the chromosome 9p21 locus were not associated with traditional risk factors including BP, BMI, cholesterol and glucose[20]; nor with biochemical parameters including lipoprotein(a), fibrinogen, albumin, bilirubin and homocysteine[129]. However, the association with intermediate phenotypes from other pathways known to be important in the pathogenesis of CAD had not been investigated. Neither had the association of phenotypes with CAD risk variants at two other replicated loci from GWA studies (on chromosome 6 and chromosome 2) been investigated.

Atherosclerosis involves an ongoing inflammatory response, and there is increasing evidence for the role of inflammation in mediating all stages of atherosclerosis from plaque initiation and progression to the thrombotic complications[378-380]. The inflammatory cytokine interleukin-6 (IL-6) is a regulator of the acute phase response which stimulates the production of all of the acute phase proteins including CRP and tumour necrosis factor-α (TNF-α)[380]. It has diverse actions that may be involved in atherogenesis including regulation of fibrinogen levels, increasing platelet aggregation, influences on endothelial function, and mediating leukocyte recruitment[380]. Cardiovascular risk factors such as smoking increase plasma IL-6 levels[381, 382], which in turn have been associated with atherosclerosis development and adverse clinical outcomes from atherosclerotic disease[383-386]. Furthermore, genetic variants which increase plasma IL-6 levels have been shown to be associated with atherosclerosis prevalence[387] and adverse clinical outcome after acute coronary syndromes[388], suggesting that IL-6 may be involved in causation rather than simply being a marker of disease. Plasma levels of the inflammatory marker C-reactive protein (CRP) and TNF-α are also associated with atherosclerosis[389-391], although whether they were causally implicated was unknown at the time of this study. The

hormone leptin that is involved in the regulation of body weight has also been shown to be involved in platelet aggregation, vascular calcification, and development of atherosclerosis in murine models[392-394].

Carotid artery intima-media thickness (CIMT) is a subclinical marker of atherosclerosis that is associated with cardiovascular disease risk factors and an independent predictor of atherosclerosis and clinical outcomes including stroke[395, 396]. CIMT progression may be reduced or reversed with interventions that reduce the risk of future cardiovascular events, supporting the concept that CIMT measurement can be used as an early marker for atherosclerosis[395], and it has been validated as a surrogate for cardiovascular disease endpoints in statin trials[397]. Although it does not directly measure carotid atherosclerosis, a close correlation between CIMT and carotid plaques has been demonstrated[398]. Investigating the relationship of CAD risk variants with CIMT may help to establish whether such variants are associated with atherosclerotic manifestations outside of the coronary arteries, and whether CIMT might permit the early effects of genetic variants to be investigated in other cohorts, even before the onset of clinical disease.

CAD risk variants on chromosome 9p21 have also been associated with non-atherosclerotic intracranial aneurysms[130, 399], suggesting that they may have broader influences on cardiovascular structure and integrity that are independent of the association with atherosclerosis. Abnormalities of cardiac structure and function such as LV dysfunction, hypertrophy, and dilatation are important predictors of adverse prognosis in population-based cohorts[400-403], and can be easily assessed non-invasively by transthoracic echocardiography. These phenotypes are heritable[404, 405], suggesting that they have a genetic basis, but the genes responsible for most of the heritability remain unknown.

Interestingly, intracranial aneurysms have also been associated with congenital cardiovascular disease. The association between aortic coarctation, bicuspid aortic valve, and intracranial berry aneurysm has been well reported and reviewed[406-408]. Although it has been suggested that hypertension resulting from coarctation could account for the association with intracranial aneurysms[409], pathological arterial wall changes have been demonstrated in normotensive adults long after coarctation repair

and also distal to the coarctation, suggesting that it may be associated with a more generalised arteriopathy[410, 411]. Moreover, subarachnoid haemorrhage may occur many years after coarctation repair and in normotensive individuals[408, 412]. Case reports and retrospective series have also shown that intracranial aneurysms, as well as intracranial dissections and AV malformations, may be associated with a range of different congenital heart defects including transposition of the great vessels, persistent truncus arteriosus, tricuspid atresia, pulmonary artery stenosis, and pulmonary valve stenosis[413-415]. Furthermore, the combination of intracranial aneurysms and cardiac defects is a component of a number of congenital syndromes, suggesting that these phenotypes may have a common aetiology[416-418]. Neural crest cells are involved embryologically in the development of the heart, aortic arch, and muscular arteries of the head and neck, and ablation of the neural crest has been shown experimentally to produce a range of cardiac defects[419]. It has been proposed that abnormal development of cells arising from the neural crest may be the common pathogenic factor responsible for the association between congenital cardiac defects and intracranial aneurysms[413]. SNPs influencing expression of genes controlling cellular proliferation and senescence (such as *CDKN2A*, *CDKN2B* and *ANRIL*) could be involved in the development of cardiac defects during embryogenesis. In view of the unexpected association between chromosome 9p21 risk variants and intracranial aneurysm[130], investigating whether CAD risk SNPs are associated with congenital heart disease is also of interest.

6.3 Aims

The broad aim of this chapter was to investigate the association of novel CAD risk variants with a range of acquired and congenital cardiovascular phenotypes.

The chapter is divided into two separate parts as outlined below.

1. Investigation of CAD risk variants with acquired quantitative cardiovascular phenotypes in the HTO cohort that has been previously shown to have adequate power to detect small genetic influences on quantitative traits, including CIMT[11, 420].

The specific aims were:

- To investigate the association of CAD variants on chromosomes 9p21, 6q25 and 2q36 with intermediate phenotypes not previously examined including plasma IL-6, CRP, TNF-α, and leptin, as well as waist-hip ratio.

- To investigate the association of novel CAD variants on chromosomes 6q25 and 2q36 with traditional risk factors, and confirm the negative associations previously observed with these phenotypes for chromosome 9p21 variants.

- To investigate the association of CAD variants on chromosomes 9p21, 6q25 and 2q36 with CIMT and echocardiographic measures of cardiac structure and function including LV dimensions, LV systolic function and LV diastolic function.

2. Investigation of chromosome 9p21 variants with congenital cardiovascular phenotypes in a case-control study.

The specific aims were:

- To determine whether risk variants for CAD/intracranial aneurysm and SNPs most strongly associated with expression of *CDKN2A/2B/ANRIL* are associated with congenital heart disease.

6.4 Association with CIMT, cardiac function, and intermediate phenotypes for CAD

6.4.1 Materials and methods

6.4.1.1 Participants and samples

Association between SNPs in the chromosome 9p21 region and acquired cardiovascular phenotypes was investigated in the Caucasian HTO population. This cohort comprised 1425 members of 248 British Caucasian families who were ascertained through hypertensive probands and who had been characterised for phenotypes relevant to cardiovascular disease, as described in Chapter 2.

Individuals with an established aetiology for LV impairment or dilatation were excluded from the analysis of echocardiographic parameters, since these would reduce the ability to detect effects attributable to the SNPs under investigation. This resulted in exclusion of 37 individuals with previous MI (based on history, ECG and echo findings) and 38 with valvular heart disease. All individuals with available measurements were included in the analysis of the other phenotypes.

6.4.1.2 Genotyping

SNPs were selected from each of the three regions that had shown replicated association with CAD in more than one GWA study at the time the experiments were carried out (chromosome 9p21.3, chromosome 6q25.1 and chromosome 2q36.3)[21]. Four SNPs selected to tag the two chromosome 9p21 haplotype blocks associated with CAD in Caucasian populations described by Samani et al[21] were genotyped. These comprised rs7044859, rs496892 (formerly known as rs1292136) and rs7865618 in the first LD block, and rs1333049 in the second LD block[18, 21]. The lead SNPs showing the strongest association with CAD at each of the other two replicated CAD loci were also genotyped: rs6922269 on chromosome 6q25.1 and rs2943634 on chromosome 2q36.3[21].

TaqMan genotyping was performed as described in Chapter 2, using made-to-order SNP genotyping assays (Applied Biosystems) with the following assay IDs: C_2618037_10 (rs7044859), C_2618027_10 (rs496892), C_2618016_10 (rs7865618), C_1754666_10 (rs1333049), C_29894051_10 (rs6922269), C_15949769_10 (rs2943634). Genotyping calls were manually reviewed and edited if required.

6.4.1.3 Statistical analysis

Mendelian inheritance and correspondence of genotype frequencies to HWE proportions was checked in both the total population and the unrelated founders using PEDSTATS [295]. Phenotypes were examined for normality and log-transformed if required (log transformed phenotypes highlighted in results Table 6.2, page 231). Phenotypes were adjusted for significant covariates using linear regression, considering age, sex, cardiovascular medications, smoking status

(current/former/never), alcohol consumption (units per week) and exercise habit (frequency per week). For each phenotype, stepwise regression was used initially to identify significant covariates to be included in the linear regression model for that phenotype. The residual values from the linear regression were then used for association testing.

Association between genotypes and adjusted phenotypes was assessed using MERLIN v1.1.2 software[300]. This evaluates the evidence for association under an additive genetic model, which was selected based on the data from previous publications showing that SNP effects were additive[21]. The upper bound of the genetic effect that is plausibly associated with each SNP for each phenotype was estimated using linear regression models. Robust "sandwich" variance estimations, to compensate for family clustering, were fitted to model additive genetic effects using Stata v9.2.

6.4.2 Results

1425 participants were included: mean age 50 (lower quartile 35, upper quartile 60) years, 678 (47.6%) male. General characteristics of the study participants are shown in Table 6.1.

Table 6.1. General characteristics of the included HTO study population.

Characteristic	N	Median (LQ, UQ)
Previous ischaemic heart disease	61 (4.3%)	-
Previous stroke	33 (2.3%)	-
Previous peripheral vascular disease	13 (0.9%)	-
Diabetes	35 (2.5%)	-
Current or former smoker	507 (35.6%)	-
Hypertension	512 (35.9%)	-
Daytime systolic blood pressure (mmHg)	958	131 (121.1, 144.1)
Daytime diastolic blood pressure (mmHg)	958	78.6 (72.0, 88.0)
Plasma total cholesterol (mmol/l)	1289	5.6 (4.8, 6.4)
BMI (kg/m^2)	1402	25.4 (23.1, 28.2)
WHR	1357	0.85 (0.78, 0.91)

N=1425 HTO participants; BMI = body mass index; WHR = waist-hip ratio; LQ = lower quartile; UQ = upper quartile.

854 participants had measurable CIMT, as shown in Table 6.2. Median maximal CIMT values of 0.858mm in men and 0.803mm in women were within the normal population range (0.36-1.07mm).

Genotyping results are shown in Table 6.3. Genotyping data were complete for more than 93% of individuals at all loci and allele frequencies were similar to those reported in the HapMap CEU population[42]. Only one SNP, rs6922269, showed a significant deviation from the proportions expected under HWE at the P=0.05 threshold, but this SNP showed no departure in the unrelated founders and was therefore regarded as acceptable. LD between the typed SNPs on chromosome 9p21 is shown in Figure 6.1.

Table 6.2. Cardiovascular phenotypes in the study population and estimated maximum genetic effect on total phenotypic variance for the typed SNPs.

Variable	N	Median (LQ, UQ)	Maximum genetic effect for typed SNPs †
BMI (kg/m^2)*	1402	25.4 (23.1, 28.2)	0.4-1.7%
WHR	1357	0.85 (0.78, 0.91)	0.6-1.4%
Daytime systolic blood pressure (mmHg)*	958	131 (121.1, 144.1)	0.5-0.7%
Mean CIMT (mm)*	854	0.76 (0.65, 0.91)	0.7-1.5%
Max CIMT (mm)*	856	0.83 (0.71, 1.00)	0.7-1.4%
Plasma total cholesterol (mmol/l)*	1289	5.6 (4.8, 6.4)	0.5-1.4%
Plasma IL-6 (pg/mL)*	1186	0.78 (0.48, 1.38)	0.4-0.7%
Plasma TNF-α (pg/mL)*	1186	0.74 (0.35, 1.66)	0.6-1.3%
Plasma CRP (mg/L)*	1314	1.40 (0.55, 3.2)	0.4-1.7%
Plasma Leptin (ng/μL)*	1319	8.6 (4.6, 15.3)	0.5-1.0%
LVIDd (mm)	829§	49 (46, 53)	0.8-2.0%
LVIDs (mm)	829§	29 (26, 33)	0.6-2.4%
Fractional shortening (%)	823§	39 (35, 45)	0.6-2.0%
Ejection fraction (%)	823§	69 (64, 76)	0.6-2.2%
E/A ratio	563§	1.0 (0.8, 1.3)	0.8-2.0%
LV mass/BSA (g/m^2)	590§	115 (94, 139)	0.8-3.2%

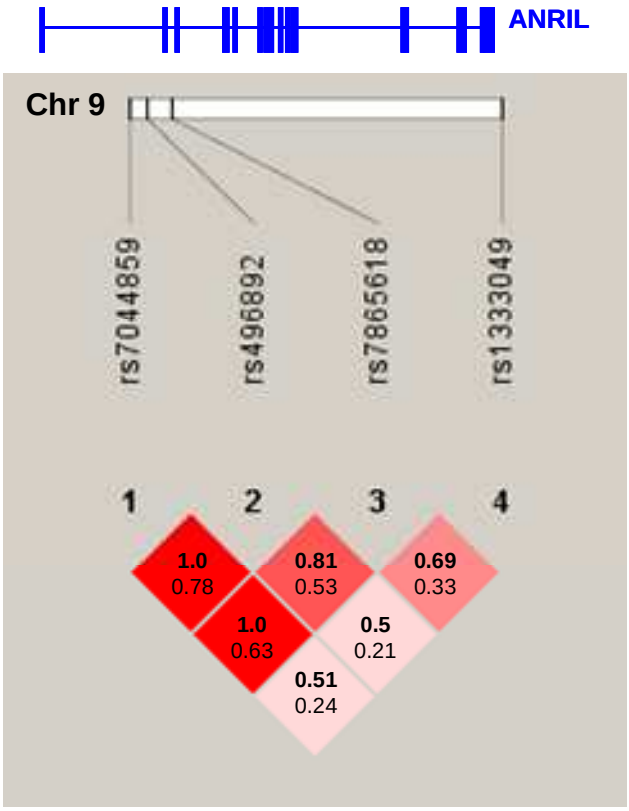
* Variables log-transformed before analysis to approximately normalise the distributions. † Numbers represent the range of maximum plausible genetic contribution to total phenotypic variance of each trait for typed SNPs. § After exclusion of 75 patients with established MI or valvular heart disease. LQ = lower quartile; UQ = upper quartile.

Table 6.3. Genotyping results.

SNP	Chromosome location	Number (%) genotyped	MAF	Alleles	HW P-value all
rs7044859	9p21	1336 (93.7%)	0.44	A:T	0.39
rs496892	9p21	1329 (93.2%)	0.49	C:T	0.85
rs7865618	9p21	1328 (93.1%)	0.45	A:G	0.50
rs1333049	9p21	1380 (94.8%)	0.45	C:G	0.31
rs6922269	6q25	1331 (93.3%)	0.33	A:G	0.01
rs2943634	2q36	1343 (94.0%)	0.35	A:C	0.71

Figure 6.1. LD between typed SNPs on chromosome 9p21.

LD plot for chromosome 9p21 SNPs showing in each diamond the D' (on top) and r^2 values (on bottom) for typed SNP combinations in our population. The relative position of *ANRIL* is shown for reference along the top in blue. Figure adapted from Haploview[297].



There was no evidence of a significant association between genotypes and plasma cholesterol, CRP, IL-6, TNF-α, or leptin concentrations; nor clinical measurements of blood pressure, or waist-hip ratio, using an uncorrected significance threshold of P=0.05. Covariate adjusted log-transformed body mass index (BMI) was associated with rs1333049 genotype with a nominal P-value of 0.03, but association of this magnitude was not unexpected in view of the number of tests performed and was likely to be related to chance. Waist-hip ratio, which is another measure of obesity, was not associated with genotype at rs1333049. The association between BMI and rs1333049 was no longer significant after Bonferroni correction for the three independent chromosomal loci tested, producing a corrected significance threshold of P=0.017.

There was no significant association between genotype and CIMT at the loci studied. The upper bound of the contribution to the total population variance in maximal CIMT was between 0.7% and 1.4% for the typed SNPs, and of a similar order for the other traits tested (Table 6.2, right-hand column). A stepwise backward-elimination multivariate regression procedure failed to find any multilocus models showing association (removal P-value > 0.2) to the CIMT phenotype. This is a powerful approach for examining tightly linked SNPs in association models that avoids the computational complexity and degrees of freedom that accompany a formal haplotype analysis[421]. It is therefore unlikely that chromosome 9p21 haplotypes are associated with CIMT, recognising that the power to detect subtle associations is very low for minor haplotypes with frequencies < 5%.

Echocardiographic measures of LV size, mass, and function also showed no convincing evidence of association with genotype at the tested SNPs. The T allele of rs7044859 on chromosome 9p21 showed nominal association with LVIDs (P=0.018), fractional shortening (P=0.034), and ejection fraction (P=0.024). However, it is important to note that these three measures are not independent since calculation of each depends on the LVIDs, and it is therefore not surprising that they show similar effects. The alternative (A) allele at this SNP is associated with CAD in GWA studies[21]. None of the associations remained significant after Bonferroni correction for the three chromosomal loci tested.

6.5 Association with congenital heart disease

6.5.1 Materials and methods

6.5.1.1 Participants and samples

Association between SNPs in the chromosome 9p21 region and congenital cardiovascular phenotypes was investigated in 888 affected probands from the CHANGE and FCH cohorts, compared to 1089 unaffected individuals from the Cumbria control cohort and 1500 unaffected individuals from the 1958 birth cohort[367]. These populations were described in Chapter 2.

6.5.1.2 Genotyping

Genotyping was performed using Sequenom methodology. Two SNPs that are associated with CAD were selected; the lead SNP rs1333049 from GWA studies[21, 368], and an additional SNP rs1547705 required to fully characterise the core risk haplotype for CAD defined by Broadbent *et al*[129]. Based on the hypothesis that an association between the chromosome 9p21 region and congenital heart disease might be mediated through effects on expression of *CDKN2A*, *CDKN2B*, and *ANRIL*, SNPs highly associated with expression of each of these genes in the NE Caucasian cohort (see Chapter 4) were also selected for genotyping. So that the 1958 birth cohort could be used as controls, only SNPs that had been genotyped in that population were selected for inclusion; this included the SNPs most strongly associated with *CDKN2B* (rs3218018) and *ANRIL* (rs564398) expression, but the SNP most strongly associated with *CDKN2A* expression (rs7036656) was not genotyped in the 1958 birth cohort. An alternative SNP, rs2811708, which was in strong LD with rs7036656 in the HapMap CEU samples (D'=1, r^2=0.92) was therefore selected instead.

6.5.1.3 Statistical analysis

Mendelian inheritance and correspondence of genotype frequencies to HWE proportions was checked using standard formulae in Excel[298]. Association between genotypes and binary phenotypes was assessed by logistic regression using UNPHASED v3.1.3[302, 303] and STATA v10 software.

6.5.2 Results

888 individuals with congenital heart disease were included, 50.9% of whom were male. The principal diagnoses of the affected individuals are summarised in Table 6.4. Classification of congenital heart disease for the investigation of genetic causation is complicated by the great phenotypic heterogeneity and fact that widely-used clinical classifications, such as cyanotic versus non-cyanotic, do not necessarily imply common developmental anomalies. Furthermore, affected individuals often have multiple separate abnormalities. The classification used here was suggested by Professor Robert Anderson, an expert in cardiac anatomy and congenital cardiac anomalies, and groups together pathologies involving common developmental processes (personal communication Professor R Anderson, University College London). Individuals were classified according to the major abnormality present, and those with multiple major abnormalities not separately categorised were termed "complex".

Genotyping results are shown in Table 6.5. Genotyping data were complete for more than 93% of individuals at all loci and allele frequencies in the Cumbria control cohort were similar to those in the 1958 birth cohort. No SNPs showed significant deviation from the proportions expected under HWE.

As shown in Table 6.6, none of the SNPs tested showed a significant association with congenital heart disease using either an allele or a general genotype model in the Cumbria control group or the 1958 birth cohort. Nor was there any association for the tested SNPs when the TOF and non-TOF individuals were analysed separately.

Table 6.4. Cardiac phenotypes in the congenital heart disease cohort.

Predominant phenotype	Number (%)
TOF	450 (50.7%)
Common AV junction (including AVSD)	25 (2.8%)
VSD	29 (3.3%)
Left sided lesions (aortic/mitral valve abnormalities, chamber hypoplasia, coarctation)	86 (9.7%)
TGA and DORV with subpulmonary defect	59 (6.6%)
Abnormal looping (double inlet LV, tricuspid atresia, congenitally corrected TGA)	7 (0.8%)
TOF-like (pulmonary atresia plus VSD, DORV with subaortic VSD)	21 (2.4%)
Common trunk	8 (0.9%)
ASD, PFO	88 (9.9%)
Ebstein's anomaly	9 (1.0%)
PDA	18 (2.0%)
Complex	73 (8.2%)
Other	15 (1.7%)

AV=atrioventricular; VSD=ventricular septal defect; TGA=transposition of the great arteries; DORV=double outlet right ventricle; ASD=atrial septal defect; PFO=patent foramen ovale; PDA=patent ductus arteriosus.

Table 6.5. Genotyping results in the congenital heart disease and control cohorts.

SNP	All congenital heart disease			TOF congenital heart disease			Non-TOF congenital heart disease			Cumbria control cohort			1958 BC
	% genotyped	MAF	HWE P-value	% genotyped	MAF	HWE P-value	% genotyped	MAF	HWE P-value	% genotyped	MAF	HWE P-value	MAF
rs2811708	94.8	0.26	0.75	98.2	0.25	0.17	98.6	0.27	0.28	93.7	0.26	0.25	0.27
rs3218018	97.0	0.09	0.56	99.6	0.09	0.30	99.3	0.09	0.89	99.7	0.10	0.97	0.09
rs564398	95.1	0.43	0.73	99.3	0.44	0.95	99.1	0.43	0.76	96.6	0.43	0.48	0.44
rs1547705	96.9	0.12	0.74	99.6	0.12	0.02	99.3	0.13	0.07	99.7	0.12	0.86	0.12
rs1333049	96.7	0.48	0.93	98.7	0.47	0.90	99.3	0.47	0.09	99.6	0.49	0.41	0.48

Table 6.6. Significance of the association with congenital heart disease for tested SNPs.

SNP	All congenital v 1958 BC P-value (allele / genotype model)	All congenital v Cumbria controls P-value (allele / genotype model)
rs2811708	0.23 / 0.11	0.78 / 0.12
rs3218018	0.33 / 0.71	0.45 / 0.60
rs564398	0.67 / 0.72	0.64 / 0.72
rs1547705	0.56 / 0.40	0.28 / 0.56
rs1333049	0.42 / 0.06	0.25 / 0.19

6.6 Discussion

Chromosome 9p21 SNPs have been associated with CAD and MI, but the mechanisms through which these SNPs influence disease are unknown. This study examined the association of risk SNPs with intermediate phenotypes that may give clues to mechanistic pathways involved in causation, and found no association with any of the clinical and biochemical phenotypes examined, including several inflammatory markers. This suggests that the CAD susceptibility mediated by this locus is not primarily mediated through the pathways and risk factors studied here, and may involve novel mechanisms.

The lack of association observed in this study extends the findings of two previous studies which found no association between chromosome 9 risk variants that are in strong LD (r^2>0.8) with the typed SNPs and traditional CAD risk factors including BP, BMI, and plasma cholesterol and glucose[20]; nor with biochemical parameters including plasma lipoprotein(a), fibrinogen, albumin, bilirubin and homocysteine[129]. The present study replicated and confirmed the lack of association with BP, BMI, and cholesterol, but was the first to demonstrate the lack of association with these inflammatory markers; this is important as inflammation is known to play a key role in the pathogenesis of atherosclerosis[378-380].

This study also investigated the association between chromosome 9p21 risk SNPs and CIMT. This was of interest since these variants had not been previously associated with extra-coronary manifestations of atherosclerosis except for aortic aneurysm, and association with CIMT could permit the early effects of genetic variants to be

investigated in other cohorts prior to the onset of clinical disease. No association was found with CIMT. The lack of association was confirmed in a similar larger study published at around the same time, involving CIMT measurements in 2,277 young Finns (aged 24 to 39 years) and a further 1,295 individuals aged 46 to 76 years[422], which also found no evidence of association with endothelial dysfunction as measured by brachial artery flow-mediated dilatation. These data suggested that chromosome 9p21 variants influence CAD though mechanisms that do not involve changes in carotid artery wall thickness. Interestingly, at the time the present study was reported, only one study examining the association between chromosome 9p21 polymorphisms and stroke had been published, which found no significant association between the CAD risk SNP rs10757278 (for which rs1333049 is a perfect proxy) and ischaemic stroke, once potentially-confounding cases of known CAD were excluded from the analysis[130]. These data and the lack of association with CIMT raised the possibility that chromosome 9p21 variants might confer risk for CAD and MI without influencing stroke risk. Several small studies have reported negative associations with stroke risk[157, 158], but other studies have demonstrated convincing association of these SNPs with ischaemic stroke[116, 159-162], as well as with other atherosclerotic phenotypes such as abdominal aortic aneurysm[130] and stiffness of the abdominal aorta[163]. An analysis of CIMT in 769 Austrian Caucasians found no association with CIMT, but did show a significant association with carotid atherosclerosis assessed ultrasonographically[154]. A subsequent large population-based study of CIMT in more than 14,000 individuals from the Atherosclerosis Risk in Communities (ARIC) study confirmed the lack of association of chromosome 9p21 variants with CIMT, and showed a borderline association with the presence of carotid atheromatous plaques in whites (P=0.047 after adjustment for covariates)[157]. The study that demonstrated an association with aortic stiffness found no association between risk variants and aortic IMT[163].

There are a number of potential explanations that need to be considered to explain the lack of association with IMT for variants that have now been shown to be associated with extra-coronary atherosclerosis and ischaemic stroke. The present study and the others examining IMT were performed in relatively homogeneous Caucasian populations similar to those in which the associations with CAD were demonstrated, making it unlikely that the results are due to bias caused by selection of a different

population. In the present study, the maximum plausible genetic effect for the typed SNPs on CIMT and the other phenotypes studied, calculated based on the observed mean effect and 95% confidence interval, was low in all cases. Therefore, although this study cannot exclude any effect of the typed SNPs on these phenotypes, it demonstrates that such effects could only be of small magnitude and are unlikely to be clinically significant. The consistent lack of association with CIMT in multiple studies and the very large cohorts used in the studies by Samani *et al*[422] and the ARIC study[157] makes it unlikely that the results represent false acceptance of the null hypothesis in the presence of a real effect; the study by Samani *et al* had greater than 99% power at an alpha of 0.01 to detect a 0.2mm difference in CIMT (expected to be associated with a 30-40% increase in stroke risk[395]) between homozygotes for the risk and non-risk alleles. The power of the ARIC study which was substantially larger would be expected to be even greater[157]. In the present study, selection through hypertensive probands could potentially introduce bias if there were an interaction between hypertension and the variants studied, but only 36% of the participants were hypertensive and the variants did not associate with hypertension. On the other hand, using a population enriched for individuals with hypertension, a trait strongly related to cardiovascular risk, resulted in selection of individuals with an increased range of CIMT values and mean CIMT 33% higher than the general population, which should increase the power for the detection of variants with small effects.

CIMT has been commonly used as a surrogate marker for atherosclerosis[397, 423]. It is an independent predictor of atherosclerosis and clinical outcomes[395, 396], but it is important to recognise that despite the close association between CIMT and carotid plaques[398], increased CIMT is not synonymous with carotid atheroma[395]. The presence of carotid plaque, which is more specific for atherosclerosis, showed a significant association with 9p21 variants at the nominal significance threshold in both studies in which this was investigated, in contrast to the negative findings for CIMT in the same studies[154, 157]. The extent to which the pathophysiological mechanisms underlying CAD and stroke overlap with CIMT is uncertain[424]; the lack of association of 9p21 risk variants with CIMT highlights that they are at least partially distinct. Other factors that alter CIMT measurements, such as hypertension, may do so through pathophysiological mechanisms that differ from those involved in atherosclerosis[425].

It appears that the chromosome 9p21 risk variants influence the risk of CAD and stroke through mechanisms that are not manifested by early changes in arterial wall thickness. This suggests that CIMT measures should not be used as a surrogate for disease endpoints in studies of the genetic risk associated with markers at chromosome 9p21.

There was no significant association between chromosome 9p21 CAD variants and echocardiographic assessments of cardiac structure and function in individuals without evidence of MI or valve disease. At the time the study was performed, no similar data from other cohorts had been reported, but two subsequent studies are consistent with the findings of the present study. Farzaneh-Fah *et al* reported no association of chromosome 9p21 SNPs with echo parameters in 593 Caucasians with stable CAD[426], but the limited sample size and use of a population with CAD in that study meant that a significant association between risk variants and cardiac structure and function in the general population could not be excluded[427]. Vasan *et al* performed a GWA study of 2.5 million SNPs with echocardiographic traits in five community-based cohorts totalling 12,612 Caucasians, with subsequent replication in two additional cohorts totalling 4,094 Caucasian individuals[428]. In stage one, 16 loci were associated with five traits, of which five were replicated in stage two. No SNPs in the chromosome 9p21 showed significant associations at either stage using the genomewide significance threshold of $5x10^{-7}$, or a less stringent threshold of 10^{-5}. These data are consistent with the lack of association with chromosome 9p21 variants observed in the present study, although failure to achieve genomewide significance levels does not exclude an effect of these SNPs on such phenotypes.

The association between chromosome 9p21 CAD variants and incident heart failure was assessed in 14,000 individuals of the ARIC study over median follow-up of 17 years[157]. Prevalent cases of CAD were excluded at baseline, and heart failure episodes were ascertained through annual telephone contact and surveillance of hospital and death records. CAD risk alleles were associated with a 30% increase in the risk of heart failure after adjusting for covariates (HR=1.30, 95%CI 1.07-1.58, P=0.008), which remained significant when CAD events were censored at their occurrence (HR=1.28, 95%CI 1.00-1.62, P=0.046). The association of CAD variants with heart failure is not surprising, since heart failure is often a consequence of CAD,

but the significant association after exclusion of both prevalent CAD at baseline and censoring for incident cases of CAD has led to speculation that 9p21 variants might contribute to heart failure independently of clinical CAD events[157]. However, unrecognised or subclinical CAD might still account for the association even after censoring for recognised CAD.

A number of issues should be considered with respect to the present study. Firstly, associations with echo measurements were available for a maximum of 829 individuals, and only 563 for the E/A ratio assessment of diastolic dysfunction. This raises the possibility that the lack of association observed could be due to a lack of statistical power. However, as discussed above for the other phenotypes examined, the maximum plausible genetic effect for the typed SNPs on the echocardiographic parameters was low in all cases. Furthermore, Farzaneh-Far *et al* estimated that they had >80% power to detect a 3% difference in LV ejection fraction in their smaller sample of 593 individuals, although they used a more accurate methodology to quantify LV function in that study. Therefore, although the current study cannot exclude that some of the typed SNPs have effects on cardiac structure and function, any such effects are likely to be small and unlikely to be clinically meaningful. A limitation of the present study is that LV function assessment was derived from single dimension M-mode measurements. This methodology has been widely used and has been shown to be reproducible with low inter-observer and intra-observer variability[429], and provides useful information for clinical studies[277]. However, the geometric assumptions required to estimate three-dimensional volume from two-dimensional measurements may produce inaccuracies, especially in the presence of regional wall motion abnormalities, and its use as a sole measure of LV function is no longer recommended in clinical practice[277]. Similarly, E/A ratio was used as a sole measure of diastolic dysfunction and more recent parameters such as tissue Doppler measurements were not available. The limitations of these estimates may have reduced the power to detect significant associations with cardiac function in the present study.

To date, this is the only study to investigate the association of novel CAD risk variants on chromosome 2q36 and 6q25 with traditional risk factors, intermediate

CAD phenotypes, and cardiac structure and function. No association with these loci was found for any of the phenotypes examined.

This study investigated SNP associations with a number of intermediate phenotypes to identify causative pathways through which genotypes at these SNPs influence risk. No associations were found, suggesting that the effects are mediated through unknown pathways. Further investigations will be required to determine the mechanisms by which these SNPs are associated with CAD.

This is the first study to investigate the association between chromosome 9p21 SNPs and congenital heart disease. No association was found for CAD risk variants or for SNPs that are associated with expression of *CDKN2A/2B/ANRIL*. This suggests that the typed SNPs are not involved in conferring susceptibility to congenital heart disease, although small effects that could not be detected in this sample size cannot be excluded. A number of other factors that could account for the observed lack of association should be considered. The study included a heterogeneous population of probands with diverse congenital heart defects of varying severity. Population stratification caused by differential SNP effects in particular phenotypes could therefore potentially obscure the association with particular subgroups in the combined cohort. Separate analysis of TOF and non-TOF groups did not show any trend towards association for either, but there were insufficient numbers with other phenotypes to permit meaningful separate analyses. However, since intracranial aneurysms and experimental neural crest ablation have been associated with a wide range of cardiovascular phenotypes[419], there were few data to justify restricting the cohort to include only particular phenotypes. Indeed, congenital heart disease is a complex phenotype and identical genetic abnormalities may be associated with a range of cardiac manifestations, such as are seen in the chromosome 22q11 deletion syndrome[430]. Investigation in a larger cohort powered to allow analysis of individual phenotypes could address this issue, but given the rarity of individual phenotypes this would require a multicentre collaborative study.

Selection bias might also reduce the ability to detect an association. All cohorts were Caucasian, but whilst the Cumbrian controls were women of childbearing age, the cases were mostly ascertained during infancy and were of both genders. An

interaction with gender could potentially obscure an association with congenital heart disease using these two cohorts, but the same allele frequencies and lack of association was seen using the mixed-gender 1958 birth cohort controls, making this unlikely.

In addition to the SNPs associated with CAD/intracranial aneurysms, this study also genotyped three SNPs which were strongly associated with *CDKN2A*, *CDKN2B* and *ANRIL* expression in peripheral blood, to investigate the hypothesis that congenital heart disease might be mediated by variants that influence expression of these genes. Although these SNPs were not associated with congenital heart defects, a role of these genes in the pathogenesis of congenital heart disease cannot be excluded since the tissue-specific influences on expression are likely to be substantially different during cardiac development. It therefore remains possible that rare variants or other common variants could mediate disease through effects on expression of these genes during embryogenesis.

Although the aetiology of intracranial aneurysms is widely accepted to be independent of atherosclerosis, intracranial aneurysms and CAD do share some common risk factors in adult populations[431], and some authors have suggested that berry aneurysms are acquired degenerative lesions[409], such that the association with CAD risk variants and intracranial aneurysm could reflect the same pathophysiological process. However, the demographics of the population in which the association with intracranial aneurysms were observed differ substantially from that in the CAD population, with a mean age <50 years and 66% female[130], making it unlikely that they represent the same phenomenon.

In summary, novel CAD variants do not appear to mediate the risk of atherothrombosis through known risk factors and are not associated with CIMT, LV structure/function or congenital heart disease. Further investigations will be required to determine the mechanisms by which these SNPs are associated with CAD.

Chapter 7

Association of *STK39* polymorphisms with blood pressure and *STK39* expression

7 Association of *STK39* polymorphisms with blood pressure and *STK39* expression

7.1 Abstract

Although BP has significant heritability, the genes responsible remain largely unknown. No associations were found in the first wave of GWA studies, but SNPs at the *STK39* locus on chromosome 2q24 were recently associated with hypertension by genome-wide association in an Amish population; *in vitro* data from transient transfection experiments using reporter constructs suggested that altered *STK39* expression might mediate the effect. However, other large studies have not implicated *STK39* in hypertension. To determine whether reported SNPs influenced *STK39* expression *in vivo*, or were associated with BP in a large British cohort, 1425 members of 248 Caucasian families ascertained through a hypertensive proband were genotyped for reported risk variants in *STK39* (rs6749447, rs3754777, rs35929607) using Sequenom technology. MERLIN software was used for family-based association testing. *Cis*-acting influences on expression were assessed *in vivo* using allelic expression ratios in cDNA from peripheral blood cells in 35 South African individuals heterozygous for a transcribed SNP in *STK39* (rs1061471) and quantified by mass spectrometry (Sequenom). No significant association was seen between the SNPs tested and systolic or diastolic BP in clinic or ambulatory measurements (all $P>0.05$). The tested SNPs were all associated with allelic expression differences in peripheral blood cells ($P<0.05$), with the most significant association for the intronic SNP rs6749447 ($P=9.9x10^{-4}$). In individuals who were heterozygous for this SNP, on average the G allele showed 13% overexpression compared to the T allele. In summary, *STK39* expression is modified by polymorphisms acting in *cis* and the typed SNPs are associated with allelic expression of this gene, but there is no evidence for an association with BP in a British Caucasian cohort.

7.2 Introduction

Hypertension is a major risk factor for vascular and renal disease and thereby contributes to substantial worldwide morbidity and mortality[432]. Although BP has been shown to have significant heritability (30-60%)[433], the genes conferring

susceptibility to hypertension remain largely unknown. As presented in Chapter 1, a relatively small study by Wang *et al* suggested that SNPs in the *STK39* gene on chromosome 2q24 may be associated with BP, and that the effects might be mediated by influences on *STK39* expression that were demonstrated *in vitro*[134].

This study aimed to test whether the *STK39* SNPs identified in the report by Wang *et al*[134] were associated with BP in a large family-based Caucasian cohort ascertained through probands with essential hypertension[156, 276]. This cohort has a higher average BP (134 v 128mmHg systolic) and lower proportion of diabetics (2.5 v 22.2%) compared to the Amish population used for the initial genome-wide study by Wang *et al*, but has 24-hour ambulatory BP recordings which provide a more reproducible and precise measure of true BP, with better prediction of cardiovascular events, compared to clinic BP measures[434, 435]. Selecting the same SNPs typed by Wang *et al* and using a cohort enriched for hypertension with ambulatory BP data should increase the power to detect effects associated with these variants in the study population. Transfection studies in cell lines may fail to adequately model *in vivo* effects on expression in complex human tissues[253], but AEI provides a powerful method for evaluating *cis*-acting influences in expression *in vivo*.

7.3 Aims

This chapter is divided into two parts, with the specific aims described below.

1. Investigation of the association of *STK39* SNPs with BP.

The specific aim was to determine whether *STK39* SNPs reported to be associated with BP in the Amish population are associated with BP in a British Caucasian cohort.

2. Investigation of the association of *STK39* SNPs with *STK39* expression in vivo.

The specific aim was to determine whether *STK39* SNPs shown to have *cis*-acting effects on *STK39* expression *in vitro* influence expression of this gene *in vivo* assessed using AEI in the SA cohort.

7.4 Materials and methods

7.4.1 Association of SNPs with blood pressure

7.4.1.1 Participants and samples

Association of *STK39* SNPs and BP was investigated using the Caucasian HTO cohort, which comprised 1425 members of 248 British Caucasian families who were ascertained through hypertensive probands, as described in detail in Chapter 2. All 1372 family members in whom DNA was available were genotyped in the present study. The results of 24-hour ambulatory BP monitoring were available for 1134 participants who had agreed to undergo monitoring. Mean values for systolic and diastolic blood pressures for the clinic, daytime and night-time periods were analysed for association with genotypes.

7.4.1.2 Genotyping

Four SNPs reported to be associated with hypertension by Wang *et al*[134] were selected for genotyping; three at the *STK39* locus on chromosome 2q24.3 (rs6749447, rs3754777 and rs35929607) and one on chromosome 9p21.3 (rs4977950) which gave the most significant signal for genome-wide association in the same study[134]. SNP genotyping was performed using Sequenom methodology, as described in Chapter 2.

7.4.1.3 Statistical analysis

Mendelian inheritance and correspondence of genotype frequencies to HWE proportions was tested using PEDSTATS[295]. The analysis was performed using log transformed BP values to approximately normalise the distribution[276]. Adjustments were then made for significant covariates determined by linear regression, considering age, sex, cardiovascular medications, smoking status (current/former/never), alcohol consumption (units per week) and exercise habit (frequency per week). For those participants taking antihypertensive medication, the effects of the main drug classes (diuretics and β-blockers) were estimated from the data by regression, and the appropriate adjustment made to the on-treatment BP values, as previously described[276]. Association between genotypes and adjusted phenotypes was assessed using MERLIN v1.1.2[300, 436]. In order to estimate the upper bound of the genetic

effect that is plausibly associated with each phenotype/SNP combination, linear regression models were fitted to model additive genetic effects using Minitab v15. .

7.4.2 Association of SNPs with *STK39* allelic expression

7.4.2.1 Identification of transcribed markers for allelic expression analysis

Allelic expression measures the relative amount of transcript arising from each allele in individuals heterozygous for a transcribed polymorphism. NCBI dbSNP was used to identify SNPs in transcribed regions of *STK39* with minor allele frequencies greater than 5%[42]. Nine transcribed variants were identified (rs1061471, rs3769429, rs7425806, rs3769428, rs56330212, rs56697518, rs56031549, rs1802105, rs56048258). Only two of these had previously been validated (minor alleles observed in at least two chromosomes) according to dbSNP[42]: rs1061471 in exon 18 (3' untranslated region) and rs56031549 in exon 11 (missense A/G SNP). Three of the reported variants (rs7425806, rs3769429 and rs3769428) were non-polymorphic in all HapMap populations. SNP rs1061471 was non-polymorphic in the Caucasian HapMap CEU population, but had a minor allele frequency of 11% and informative heterozygote frequency of 24% in the African HapMap YRI population. Population frequency data were unavailable for the other SNPs. Therefore, in order to find transcribed markers suitable for assessing allelic expression, rs1061471 and the five SNPs for which HapMap population frequencies were not available (rs1802105, rs56031549, rs56048258, rs56330212, rs56697518) were genotyped in the SA cohort.

7.4.2.2 Participants, samples and genotyping

In view of the lack of transcribed polymorphisms in Caucasian populations, investigation of SNP effects on expression using AEI was performed in the 310 healthy individuals of the SA cohort. Genotyping was performed using Sequenom methodology as described in Chapter 2.

35 individuals heterozygous for the transcribed variant rs1061471 suitable for allelic expression analysis were identified. The other five SNPs tested had minor allele frequencies <1%, giving an insufficient number of heterozygotes to be used for AEI assessment (Table 7.1).

Table 7.1. Allele frequency of transcribed SNPs in 310 South African individuals.

SNP	Alleles*	Number genotyped	Number of heterozygotes	MAF	HapMap YRI MAF	Hardy-Weinberg p value
rs1061471	G / A	298	35	0.069	0.119	0.15
rs1802105	A / C	298	0	0	NA	-
rs56031549	C / T	298	2	0.003	NA	0.95
rs56048258	A / G	296	5	0.008	NA	0.88
rs56330212	G / A	296	0	0	NA	-
rs56697518	C / T	298	2	0.003	NA	0.95

* Major allele given first; MAF, minor allele frequency; YRI, HapMap YRI African cohort; NA, not available.

7.4.2.3 Measurement of *STK39* allelic expression ratios:

Quantification of the allelic expression ratio using 50ng of cDNA was performed by Sequenom methodology, as described in Chapter 2. Results from amplification of gDNA from 19 individuals (performed in 4 replicates) were used as the equimolar reference to normalise the cDNA values.

7.4.2.4 Statistical analysis:

Mendelian inheritance and correspondence of genotype frequencies to HWE proportions was checked using standard formulae in Excel[298]. Analyses were performed using the logarithm of the normalised allelic expression ratios as published by Teare *et al*[254] and presented in Chapter 2.

7.5 Results

7.5.1 Association with blood pressure

1372 individuals with available DNA were included. 649 (47%) were male and the mean age was 49 (standard deviation 15.5) years. 490 (36%) were hypertensive, 298 (22%) were current smokers, 35 (3%) had diabetes, 55 (4%) had structural heart disease, and 456 (33%) were taking antihypertensive or antianginal medication.

Median alcohol consumption was 3 (interquartile range 0-12) units per week. 43% reported no regular exercise per week, with 21%, 22% and 14% exercising once, twice, and three or more times per week respectively. The BP parameters of participants, heritability (h^2) of each phenotype and proportion of phenotypic variation accounted for by covariates (r^2) are shown in Table 7.2.

Table 7.2. BP characteristics of the Caucasian participants.

Characteristic	*N*	Median (LQ, UQ) in mmHg	Proportion of phenotypic variability (r^2) explained by covariates for log-transformed variable	Heritability (h^2) for adjusted variable
Clinic Systolic BP	1138	134 (121.3, 152.3)	27.2	11.2
Clinic Diastolic BP	1130	82 (73.7, 92)	19.7	12.2
Day Systolic BP	1134	132.5 (122.1, 146)	20.4	13.6
Day Diastolic BP	1133	79.7 (72.5, 90)	17.9	10.7
Night Systolic BP	903	113 (103.8, 126)	11.4	24.3
Night Diastolic BP	902	66 (60.2, 73.6)	13.6	34.6

N=1372 Caucasian participants. BP = blood pressure; LQ = lower quartile; UQ = upper quartile.

Genotyping was complete for more than 99% of individuals at all loci. Allele frequencies for typed SNPs are shown in

Table 7.3; these were similar to the HapMap CEU population and did not deviate significantly from HWE proportions.

There was no significant association between genotype and covariate-adjusted, log transformed BP parameters at the loci studied using an unadjusted P-value threshold of 0.05. Estimates of effect size with 95% confidence intervals are shown in Table 7.4; the upper bound of the contribution of typed SNPs to the total population variance was low for all of the phenotypes tested (ranges: rs3754777 0.4-1.1%; rs35929607 0.5-1.7%; rs6749447 0.4-1.4%; rs4977950 0.4-0.6%).

Table 7.3. Allele frequencies of the SNPs typed in study participants.

SNP	Chromosome location	Alleles*	Number genotyped	HW P-value	Frequency of putative risk allele		
					Caucasian cohort	HapMap CEU	SA cohort
rs3754777	2q24 (*STK39* intron)	G/**A**	1367	0.63	0.17	0.13	0.16
rs35929607	2q24 (*STK39* intron)	A/**G**	1361	0.79	0.19	0.18	0.34
rs6749447	2q24 (*STK39* intron)	T/**G**	1369	0.32	0.29	0.28	0.52
rs4977950	9p21 (not within gene)	G/**C**	1365	0.46	0.15	0.18	Not typed

***** Major allele given first, putative risk allele for hypertension identified by Wang *et al* shown in bold.

Table 7.4. Effect sizes of tested SNPs.

BP phenotype*	*N*	rs3754777	rs35929607	rs6749447	rs4977950
Clinic Systolic	1138	0.2 (-1.8, 2.3)	1 (-1, 2.9)	0.5 (-1.2, 2.2)	0 (-2.1, 2.1)
Clinic Diastolic	1130	-0.3 (-1.6, 1.1)	0.4 (-0.9, 1.7)	0.1 (-1, 1.2)	0.3 (-1.1,1.7)
Day Systolic	1134	-0.1 (-1.9, 1.7)	-0.3 (-1.5, 2)	0.1 (-1.4, 1.6)	0.2 (-1.7, 2)
Day Diastolic	1133	-0.1 (-1.4, 1.1)	0.2 (-1, 1.4)	0.1 (-0.9, 1.2)	0.4 (-0.9, 1.7)
Night Systolic	903	0.4 (-1.4, 2.1)	0.9 (-0.8, 2.6)	0.7 (-0.7, 2.2)	-0.2 (-2.1, 1.7)
Night Diastolic	902	0.7 (-0.5, 1.9)	1.1 (0, 2.3)	0.8 (-0.2, 1.8)	-0.1 (-1.3, 1.2)

* All phenotypes log-transformed to approximately normalise distributions and adjusted for covariates before analysis; Effect sizes are the mmHg difference attributable to 1 copy of the risk (minor) allele, with 95% confidence interval given in brackets. P-value >0.05 for all effects.

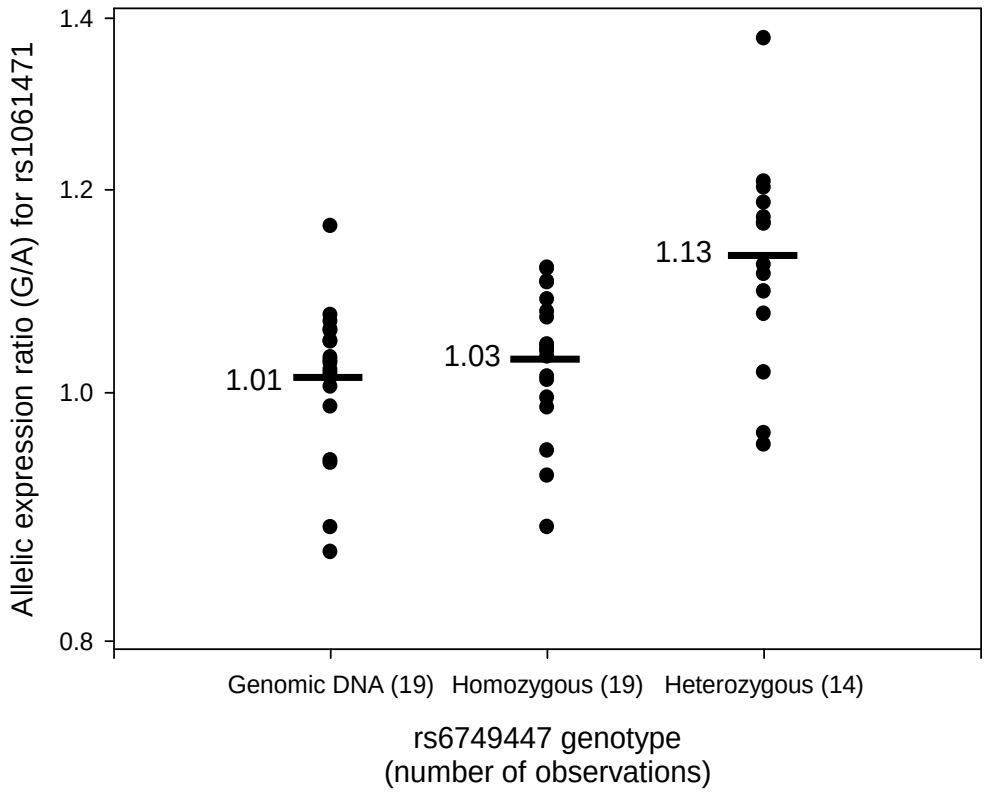
7.5.2 Association with allelic expression

Two of the 35 samples failed quality control criteria for allelic expression measurements at the transcribed SNP rs1061471 because of high standard error between replicates and were excluded from the analysis; all replicates for the remaining 33 samples were included. The experimental variability was low and the average standard error of the log-transformed ratio for the four technical replicates

was the same in gDNA (0.0146) and cDNA (0.0149). The standard deviation of the sample means (represented in Figure 7.1) was 0.030 in genomic DNA and 0.039 in cDNA.

Figure 7.1. Effect of genotype at rs6749447 on allelic expression ratio of the transcribed SNP rs1061471.

Circles represent allelic expression ratio for each individual, with horizontal bars representing the mean values for each group (shown alongside). The first column shows genomic DNA where the alleles are present in a 1:1 ratio, giving a mean ratio of approximately 1. The second column shows individuals who are homozygous for either allele of rs6749447. In this group, the *cis*-acting influence on expression from each allele is the same, giving a mean allelic expression ratio of approximately 1 at the transcribed marker rs1061471 (P>0.05 for the comparison with genomic DNA). The third column shows individuals who are heterozygous for rs6749447. In this group each of the two transcribed alleles is expressed at a different level, causing increased imbalance in allelic expression (P<0.005 for the comparisons with genomic DNA and individuals homozygous for rs6749447)



The tested SNPs were all significantly associated with allelic expression differences in peripheral blood, as shown in Table 7.5. The most significant association with expression was for rs6749447 ($P=9.9x10^{-4}$). This SNP is located in intron 1 of *STK39* and in the SA sample had minor allele frequency 0.48 and D' 0.54 with the transcribed SNP rs1061471 located in the 3' untranslated region of the gene. Of the 33 included individuals who were heterozygous for the transcribed SNP rs1061471, 14 were heterozygous for rs6749447. If a SNP affects expression in *cis* then individuals heterozygous for that SNP should show a greater allelic imbalance at the transcribed SNP than homozygous individuals. Figure 7.1 shows the allelic expression ratios grouped by whether individuals are heterozygous or homozygous for rs6749447, compared to the ratio seen in genomic DNA (where alleles are present in a 1:1 ratio). There was no difference in allelic expression ratios between genomic DNA and individuals homozygous at rs6749447 (P=0.07 using the Mann-Whitney U test). However, individuals heterozygous for rs6749447 have on average a higher allelic expression ratio than the homozygous individuals (P=0.005 using the Mann-Whitney U test), which is consistent with a *cis*-acting effect and also suggests that the overexpressing allele is preferentially in phase with the G allele at the transcribed locus. In individuals who were heterozygous for rs6749447 on average the G allele of rs6749447 showed 13% overexpression compared to the T allele. Once the effect of this SNP was adjusted for, the association for the other SNPs no longer remained significant (as shown in Table 7.5). Differences in allele frequencies and patterns of linkage disequilibrium between the Caucasian and South African populations are summarised in

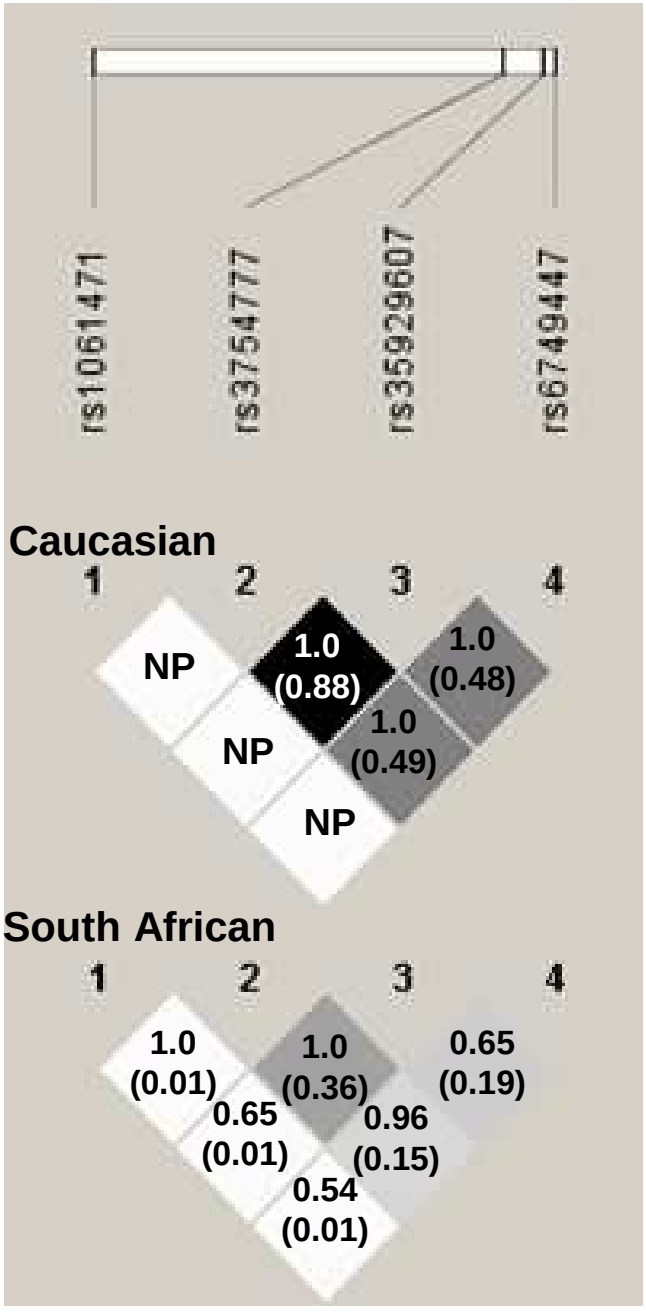
Table 7.3 and Figure 7.2.

Table 7.5. Association of tested SNPs with allelic expression differences.

SNP	P-value	P-value adjusted for rs6749447
rs35929607	0.02	0.39
rs3754777	0.006	0.71
rs6749447	0.001	-

Figure 7.2. LD between typed SNPs at the chromosome 2q24 *STK39* locus in Caucasian and SA participants.

Shading represents r^2 values (r^2=0 white, with higher r^2 values darker shades of grey). Numbers show D' values (and r^2 values in brackets) between SNPs. NP, non-polymorphic in Caucasian population. Figure adapted from Haploview.



7.6 Discussion

The *STK39* SNPs reported to be associated with BP and hypertension by Wang *et al*[134] did not show a significant association with BP in the British Caucasian families. Other published GWA studies and two large recent meta-analyses did not identify associations achieving genome-wide significance levels for *STK39* polymorphisms and BP, despite powerful analyses of combined cohorts involving 34,433 and 71,225 Caucasian individuals which successfully identified multiple other susceptibility loci[139, 140]. Although failure to achieve genome-wide significance levels does not exclude an effect of this locus on BP, any effect of sequence variation at *STK39* on BP is too small to be detected in these data and is perhaps something particular to the Amish population. No association with BP was found for the chromosome 9p21.3 SNP rs4977950 which was the top signal in the GWA studies by Wang *et al* ($P = 9.1x10^{-8}$), and the only SNP to achieve significance at the genome-wide threshold in their study.

Inadequate power is an unlikely explanation for the lack of association in this study as the maximum plausible genetic effect for the typed SNPs on BP phenotypes, which was calculated based on the observed mean effect and 95% confidence intervals, was low in all cases and the data exclude an effect as large as the 3mmHg increase in systolic BP per copy of the risk allele reported by Wang *et al* in the Amish population. Families in the present study were selected through a hypertensive proband and had a higher proportion of hypertensive individuals and wider distribution of BP values; this should increase the power to detect an association with BP phenotypes compared to the cohorts in the study by Wang *et al*, which were not selected for BP. The use of 24-hour ambulatory BP recordings in the present study provides a more reproducible assessment of 'usual' BP and reduces misclassification due to 'white-coat' or masked hypertension that may occur with isolated clinic measurements[434, 435]; this should also reduce the noise and increase the power of the study. The heritability of night-time BP measures was significantly higher than clinic BP measures, which should increase the power for detecting genetic effects. Studies first reporting a novel association often show a more extreme odds ratio than subsequent replication studies[437]; if the effect size estimated by Wang *et al* is an upwardly biased estimate this might explain the negative association in the present

study and the other reported GWA studies. The effect due to *STK39* alleles observed by Wang *et al* was strongest in the Amish population and substantially weaker in the other Caucasian cohorts tested in their study. Data from the present study suggest that the effect of the typed *STK39* SNPs on BP is at most modest in a UK Caucasian population. If the reported association was caused by LD between the typed SNPs and another functional variant, then weaker LD in other populations, compared to the Amish population, could account for the results; this is particularly pertinent since closed founder populations are expected to exhibit extensive LD. Heterogeneity between the populations, including differences in population history, recruitment strategy, and phenotypes (such as the proportion of diabetics and hypertensives) might also contribute to the differences in genetic associations we have observed. The analysis was performed using an additive model because this was the model used to report the most significant overall association by Wang *et al*.

The study has demonstrated that SNPs in the *STK39* gene correlate with *in vivo* allelic expression of this gene in peripheral blood cells. This corroborates and extends the findings from reporter gene constructs in the study by Wang *et al*. Using a luciferase assay in HeLa and HEK293 cell lines they examined the effect on transcription of two SNPs (rs12692877 and rs35929607) that are in conserved elements and which were in complete LD with two of the SNPs associated with BP in their GWAS in the Amish population (rs6749447 and rs3754777 respectively). They demonstrated that the G allele of rs35929607, which was the allele reported to be associated with increased BP, was in isolation associated with a greater than two-fold increase in transcriptional activity compared to the A allele of this SNP, or either allele of rs12692877. Such *in vitro* studies have limitations since results depend on the constructs that are used and expression is considered outside of the normal chromatin and cellular context, which may not reflect true expression in complex tissues *in vivo*[225, 253]. Allelic expression analysis allows *in vivo* assessment of RNA transcripts in their native environment and regulatory context, and by controlling for *trans*-acting influences has high sensitivity to detect *cis*-acting effects. This study found a stronger association between genotype and allelic expression for rs6749447 than for rs35929607, with the G allele of rs6749447 associated with a 13% increase in expression relative to the T allele. However, if the effect of rs6749447 is not accounted for, then the allelic expression analysis found that the G allele of rs35929607 was significantly associated with

increased expression. This is consistent with the *in vitro* data from Wang *et al*, although the magnitude of the relative increase in our study was lower at 7%. The difference in the magnitude of effect we observed compared to the *in vitro* transfection studies is not surprising in view of the fact that *in vivo* expression is likely to result from the interaction of multiple elements that modulate expression, rather than just a single SNP.

AEI can only be assessed in individuals who are heterozygous for a transcribed polymorphism in the gene of interest – there were no suitable transcribed SNPs in *STK39* in Caucasian populations and in the SA cohort the number of suitable heterozygotes was relatively small because of a low minor allele frequency (7%) at the transcribed SNP rs1061471. However, measurement of allelic expression ratios within samples is very sensitive for detection of *cis*-acting influences since *trans*-acting and experimental factors are identical for each allele; this allows significant *cis*-acting effects to be detected in relatively few samples, as demonstrated in this study. Sensitivity of the technique for the detection of significant effects in equivalent sample sizes has been previously demonstrated[260, 266]. Populations of African-descent show on average smaller regions of high LD compared to Caucasian populations[321, 438], and therefore using a South African cohort increases the sensitivity to separate the effects of different SNPs on expression. In the case of *STK39*, using a population of different ethnicity allowed allelic expression to be measured in the absence of transcribed polymorphisms in the Caucasian population. Correlation between *STK39* SNPs and expression of this gene does not necessarily mean that these SNPs will influence BP and the finding of an association with *STK39* expression in this study is therefore not at odds with the lack of association with BP that was observed.

A limitation of the present study is that expression was only tested in white blood cells, rather than in a tissue of potentially greater relevance to BP regulation. However, many *cis*-acting influences on gene expression are expected to be the same in different cell types[439], although tissue-specific differences have been described[440]. In the case of *STK39* this approach is supported by the fact that rs35929607 genotype correlated with expression in immortalised cell lines derived from cervical tumour cells (HeLa) and embryonic kidney cells (HEK293) in the study by Wang *et al*, as well as with expression in blood in the present study. However, the effects may vary

in other tissues. This study analysed the SNPs influencing expression in a SA cohort of mixed ethnicity, but the SNPs associated with expression may vary in different populations due to differing allele frequencies and LD patterns. Although the associations of SNPs with *STK39* expression were highly statistically significant, the biological significance of these findings is less certain, since it is not known what impact such an effect on expression has on disease risk.

Future studies will be necessary to investigate the association of other SNPs at the *STK39* locus with BP, or to determine whether rare mutations at this locus contribute significantly to population blood pressure variation, as has been shown for other genes implicated in hypertension causation[441].

7.7 Conclusions

STK39 expression is modified by polymorphisms acting in *cis*, but there is no evidence that these SNPs affecting *STK39* transcription are associated with BP in a British Caucasian cohort.

Chapter 8

General discussion and future directions

8 General discussion and future directions

CAD and other cardiovascular phenotypes such as hypertension have a significant genetic component, yet the pathways through which genetic factors mediate disease susceptibility remain largely unknown. Studying intermediate phenotypes, including gene expression, may lead to the identification of the genes and pathways involved in the pathogenesis of disease. As well as being of intrinsic scientific interest, this may allow the development of novel biomarkers to aid diagnosis or improve risk stratification, but more importantly, offers the potential for developing novel therapeutic interventions which influence these mechanisms to modify disease risk. An advantage of the recent GWA approach is that associations are identified in a 'hypothesis-free' manner, which offers the potential to provide truly novel insights into biology, as illustrated by discovery of the chromosome 9p21 region which was not previously implicated in the pathogenesis of CAD.

This study examined the association of 9p21 risk variants with a range of intermediate phenotypes to identify pathways implicated in causation. There was no association with traditional cardiovascular risk factors or plasma levels of inflammatory mediators known to be involved in atherogenesis, suggesting that the association at this locus might be mediated through previously unsuspected mechanisms. There was no association with CIMT, suggesting that this cannot be used as an intermediate phenotype in studies investigating modulation of risk associated with this locus. Furthermore, although the association of risk variants with intracranial aneurysms raised the possibility of more widespread cardiovascular effects not linked to atherosclerosis, there was no association with congenital heart disease or echocardiographic measures of cardiac structure and function in adults.

The study also examined the association of risk variants with expression of genes in the 9p21 region that were not previously implicated in the pathogenesis of CAD. The study yielded a number of novel findings. There was a strong relationship between risk variants and expression of *ANRIL* transcripts measured using an assay spanning exons 1-2, with CAD variants associated with an up to two-fold reduction in *ANRIL* expression. Multiple loci were independently associated with expression, suggesting

that several sites rather than a single 'causal' variant may modulate disease susceptibility. Risk variants for diabetes, glioma and melanoma were also associated with *ANRIL* expression. These results suggest that modulation of *ANRIL* expression in the *CDKN2B* antisense region mediates susceptibility to several important human diseases. *ANRIL* is therefore clearly identified as a target for further investigation to determine its function, how it is affected in disease, and whether it can be modified to influence disease susceptibility.

The effects of various disease states on *ANRIL* expression require further investigation, but a very recent 2010 report provided further evidence to support a link between *ANRIL* and CAD, showing that *ANRIL* expression levels in blood and atherosclerotic plaque are associated with the severity of atherosclerosis[155]. The finding that CAD risk variants are associated with *ANRIL* expression in young healthy individuals in the present study suggests that levels of *ANRIL* expression are likely to be a cause, rather than a consequence, of atherosclerosis; the median age of the healthy volunteers in the SA cohort was 20 years making the presence of significant atherosclerosis extremely unlikely in that cohort. Furthermore, the observation that *ANRIL* expression shows considerable variation even in healthy volunteers suggests that therapeutic manipulations aimed at influencing levels of *ANRIL* expression may, at least in principle, be tolerated. A recent 2010 report by Visel *et al* showed that mice with homozygous knockout of a 70kb region orthologous to the CAD core risk region were viable and fertile without obvious morphological or behavioural phenotypes, but did demonstrate severely reduced expression of *CDKN2A* and *CDKN2B*, and increased proliferation of vascular smooth muscle cells as occurs in atherosclerosis[327]. However, these mice also had a reduced life expectancy due to an excess of cancers. A limitation of such models with respect to atherosclerosis is that most mice strains do not develop CAD, and the effect of knocking out this region on atherosclerotic phenotypes is not known. Further studies may investigate this in murine models of atherosclerosis. However, a further problem with respect to studies in mice models is that although the CAD risk region overlaps the 3' end of *ANRIL* in humans, the *ANRIL* sequence is not conserved beyond primates[44, 200], and therefore the relationship of the orthologous knockout region to any functionally similar molecule in mice is unknown. Whether components of the *ANRIL* pathway can be

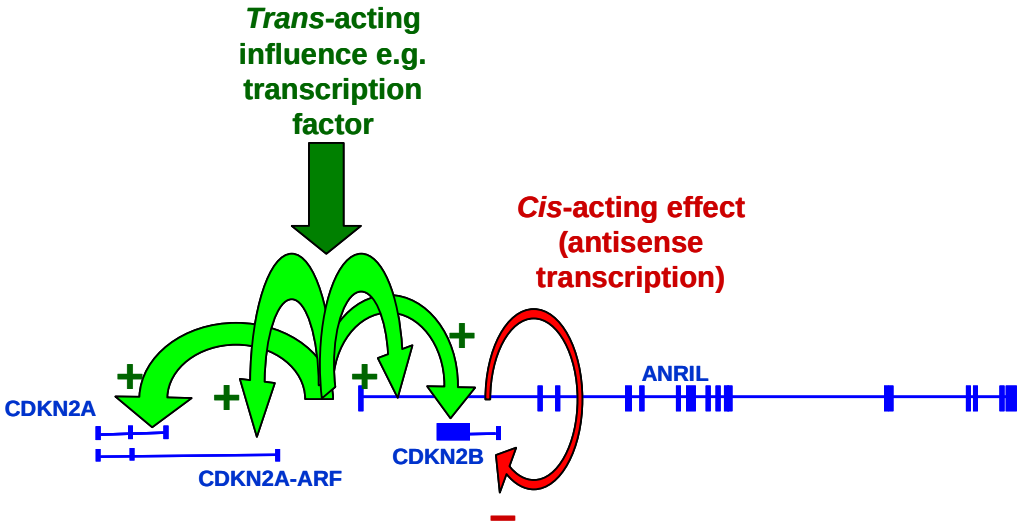
manipulated to influence disease risk without unacceptable sequelae from the perturbation of normal function is as yet unknown.

The study also demonstrated and validated a novel approach for AEI analysis, combining data from multiple transcribed markers per gene. This improved the power to map *cis*-acting effects compared to standard single-marker analysis, and gave considerably higher power than conventional mapping techniques based on analysis of total expression levels. These findings have important implications for future studies investigating *cis*-acting effects, for example for the numerous non-coding variants associated with diseases in GWA studies at other loci. Utilising a more powerful approach means that significant effects can be detected in a smaller sample size, which may be particularly important for investigating tissue-specific effects where large sample collections may be difficult to establish. The combined analysis of allelic and total expression levels gave interesting insights into the potential mechanisms regulating expression at the important 9p21 locus. Total expression levels of *CDKN2A*, *CDKN2B* and *ANRIL* exons1-2 were correlated, as observed in other studies, likely representing co-regulation by shared *trans*-acting elements, perhaps from a common bidirectional promoter between *CDKN2A* and *ANRIL*. However, SNPs had opposite *cis*-acting effects on expression of *ANRIL* exons 1-2 and *CDKN2B*, suggesting that antisense transcription may be involved in *CDKN2B* regulation[202]. A proposed model to unify these observations is shown in Figure 8.1.

The study also used AEI techniques to confirm *in vivo* the association of chromosome 2q24 polymorphisms with *STK39* expression that had been previously identified *in vitro*[134]. However, association analysis showed no association of these variants with BP in a large cohort of British Caucasian families, and excluded an effect of the magnitude previously reported in an Amish population[134]. This highlights the importance of replication of genetic associations in independent populations.

Figure 8.1. Proposed model of factors influencing transcription of genes in the chromosome 9p21 region.

Trans-acting factors such as transcription factors (dark green arrow) influence expression of all genes in a co-ordinated manner (bright green arrows) from shared promoter/enhancer elements. *ANRIL* transcription has an inverse effect on *CDKN2B* expression mediated through antisense transcription. The magnitude of t*rans* effects is greater than *cis* effects (represented by increased arrow thickness) such that total expression levels of *CDKN2B* and *ANRIL* are positively correlated.



Another interesting approach employed in the current studies for both the chromosome 9p21 and chromosome 2q24 regions was the use of populations of African ancestry for association analyses. Characterisation of genomic sequence and genetic variation such as SNPs has primarily been performed using Caucasian populations, and most association studies performed to date have studied populations of European ancestry[25]. However, populations of African ancestry exhibit greater genetic diversity which has a number of potential advantages. First, the average size of haplotype blocks is smaller in African populations than in Caucasian populations[53]. Strong LD limits the ability of association analysis to separate the effects of individual variants within haplotype blocks, and African populations can therefore be used to improve the resolution with which functional elements can be fine-mapped[323]. Although in the present study LD in the core CAD risk region remained strong and did not allow much separation of the effects of individual variants contained within it, interesting differences were detected for other more distant variants. The most notable example of this was the diabetes variant rs10811661[125], which was associated with *ANRIL* expression in the Caucasian but not the SA cohort. The lack of

association with *ANRIL* expression in the SA cohort implies that this SNP is either not the functional variant that acts through effects on *ANRIL* expression, or alternatively, if this SNP is the causative variant, that the association with diabetes is not primarily mediated through effects on *ANRIL* expression. It will be interesting to determine whether this variant is associated with diabetes in African cohorts. A second advantage of African populations is that in addition to greater haplotype diversity, they also display greater nucleotide diversity[321, 438]. Recent sequencing of two complete Khoisan and Bantu genomes from southern Africa revealed that each contained more than a million single base pair changes that were not found in each other or in any of the previously published genomes[442]. Furthermore, analysis of partial genome sequences from three other Khoisan individuals showed that in terms of nucleotide substitutions, these individuals were on average more different from each other than a European and an Asian[442]. These data suggest that even greater genetic diversity remains to be discovered in African populations. Such diversity was exploited in the *STK39* allelic expression analysis presented here, where using the SA population permitted allelic expression to be measured in the absence of transcribed polymorphisms in the Caucasian population. Similarly, the greater number of variants in populations of African ancestry may allow the novel approach of combining AEI data from multiple transcribed SNPs to be used in such populations when multiple transcribed variants are not present in a gene in Caucasians.

Several areas are identified for future study. First, the range of *ANRIL* transcripts and their relationship to chromosome 9p21 variants in different tissues requires further investigation. A systematic survey of the *ANRIL* transcripts present in different primary tissues could be performed using sequencing techniques, such as rapid amplification of cDNA ends (RACE)[287]. The relationship of genetic risk variants to different *ANRIL* transcripts could then be more fully characterised. The effects of *ANRIL* also remain to be fully elucidated. Cell models could be used to evaluate the effect of *ANRIL* knockdown using inhibitory RNAs. These could be designed to selectively target specific *ANRIL* exons, which may allow the effects of different transcripts to be characterised. Effects of *ANRIL* on expression of other unknown targets could be investigated in such models using whole-genome microarray expression data. Studies in animal models are complicated in view of the fact that the *ANRIL* sequence is poorly conserved in other species and at present it is not known

whether orthologous regions, such as that knocked out in mice by Visel *et al*[327], contain large noncoding RNAs that are functionally similar in spite of the sequence differences. The relationship between microsatellite rs10583774 genotype and *ANRIL* expression could be investigated using *in vitro* transfection studies with reporter constructs. Based on the hypothesis that disease susceptibility is mediated by changes in *ANRIL* expression, it would also be interesting to determine whether haplotypes most strongly associated with *ANRIL* expression offer increased power for prediction of disease in association studies.

The relationship between risk variants and expression also needs to be confirmed for the tissues in which the effects are believed to be mediated for the different diseases. A tissue bank of 110 atheromatous carotid endarterectomy specimens has been previously collected by our group between 2004 and 2007. Samples were snap-frozen in liquid nitrogen at the time of the operation and subsequently stored at -80^{o}C. I have performed preliminary optimisation work to establish techniques for extracting RNA from these samples with the aim of performing expression mapping in this tissue which is of potentially greater biological relevance for investigating the effect of chromosome 9p21 risk variants on gene expression related to CAD. Optimised protocols defining the methodology for extracting adequate RNA from these samples are under development. However, the cell types in which chromosome 9p21 variants influence CAD susceptibility through effects on expression are uncertain, and the relevant tissue may not be accessible in the numbers/quantity required to perform expression analyses. In addition to the carotid endarterectomy specimens described above, other sources of arterial tissue that could be utilised include sections of macroscopically-normal radial or internal mammary arteries harvested during coronary artery bypass graft surgery, or samples of diseased aorta from aortic aneurysm repair surgery. Such samples include heterogeneous cell types, and although laser capture microdissection can be used to isolate specific cell populations (such as vascular smooth muscle cells or macrophages) the amount of RNA obtained using this technique is small, and may be insufficient for expression analysis without amplification[443]. Induced pluripotent stem cells could be obtained from individuals with selected genotypes at loci thought to influence expression[444]. These could then be differentiated into various tissue types, such as cardiomyocytes, and the effects of expression studied in cell populations with different genotypes at putative risk loci.

Using an AEI approach for such studies would reduce the influence of experimental and *trans*-acting factors. However, such *in vitro* expression models may not accurately represent the true *in vivo* effects[253].

Identifying the mechanisms through which genetic variants identified in GWA studies mediate disease is an important goal. This work has identified associations between SNPs associated with a number of different diseases and expression of genes at the chromosome 9p21 locus, and has demonstrated a new approach to AEI analysis that may improve the power for mapping variants associated with disease at other loci.

Chapter 9

References

9 References

1 Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. Lancet. 367(9524):1747-57.

2 Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006;3(11):e442.

3 Watkins H, Farrall M. Genetic susceptibility to coronary artery disease: From promise to progress. Nat Rev Genet. 2006;7(3):163-73.

4 Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. Lancet. 2004;364(9438):937-52.

5 Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. N Engl J Med. 1994;330(15):1041-6.

6 Zdravkovic S, Wienke A, Pedersen NL, Marenberg ME, Yashin AI, De Faire U. Heritability of death from coronary heart disease: A 36-year follow-up of 20 966 Swedish twins. J Internal Medicine. 2002;252(3):247-54.

7 Wienke A, Holm NV, Skytthe A, Yashin AI. The heritability of mortality due to heart diseases: A correlated frailty model applied to Danish twins. Twin Research. 2001;4:266-74.

8 Lloyd-Jones DM, Nam B-H, D'Agostino RB, Sr., Levy D, Murabito JM, Wang TJ, et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: A prospective study of parents and offspring. JAMA. 2004;291(18):2204-11.

9 Murabito JM, Pencina MJ, Nam B-H, D'Agostino RB, Sr., Wang TJ, Lloyd-Jones D, et al. Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. JAMA. 2005;294(24):3117-23.

10 Keavney B. Genetic epidemiological studies of coronary heart disease. Int J Epidemiol. 2002;31(4):730-6.

11 Mayosi BM, Avery PJ, Baker M, Gaukrodger N, Imrie H, Green FR, et al. Genotype at the -174g/c polymorphism of the interleukin-6 gene is associated with common carotid artery intimal-medial thickness: Family study and meta-analysis. Stroke. 2005;36(10):2215-9.

12 Peyser PA, Bielak LF, Chu JS, Turner ST, Ellsworth DL, Boerwinkle E, et al. Heritability of coronary artery calcium quantity measured by electron beam computed tomography in asymptomatic adults. Circulation. 2002;106(3):304-8.

13 O'Donnell CJ, Chazaro I, Wilson PWF, Fox C, Hannan MT, Kiel DP, et al. Evidence for heritability of abdominal aortic calcific deposits in the Framingham Heart Study. Circulation. 2002;106(3):337-41.

14 Genetic association database. [cited 2010 8th March]; Available from: http://geneticassociationdb.nih.gov/cgi-bin/index.cgi

15 Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456(7218):18-21.

16 Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs rare allele hypotheses for complex diseases. Current Opinion in Genetics and Development. 2009;19(3):212-9.

17 Cunnington MS, Keavney B. Genetics of coronary heart disease. *Evidence based cardiology*. 3rd ed. Chichester: BMJ publishing Ltd 2010.

18 The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.

19 Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007;316(5830):1491-3.

20 McPherson R. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007;316:1488-91.

21 Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007;357(5):443 - 53.

22 Erdmann J, Groszhennig A, Braund PS, Konig IR, Hengstenberg C, Hall AS, et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet. 2009;41(3):280-2.

23 Tregouet D-A, Konig IR, Erdmann J, Munteanu A, Braund PS, Hall AS, et al. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat Genet. 2009;41(3):283-5.

24 Gudbjartsson DF. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet. 2009;41:342-7.

25 Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Investigation. 2008;118(5):1590-605.

26 Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, Devillers M, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003;34(2):154-6.

27 Cohen JC, Boerwinkle E, Mosley Jr TH, Hobbs HH. Sequence variations in pcsk9, low ldl, and protection against coronary heart disease. N Engl J Med. 2006;354(12):1264-72.

28 Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695-701.

29 Santibanez Koref MF, Wilson V, Cartwright N, Cunnington MS, Mathers J, Bishop T, et al. MLH1 differential allelic expression in mutation carriers and controls. Annals of Human Genetics. 2010:[in press].

30 Tsuchihashi Z, Dracopoli NC. Progress in high throughput SNP genotyping methods. Pharmacogenomics Journal. 2002;2(2):103-10.

31 Maresso K, Broeckel U, Rao DC, Gu CC. Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. *Advances in genetics*: Academic Press 2008:107-39.

32 Hutchison CA, III. DNA sequencing: Bench to bedside and beyond. Nucl Acids Res. 2007;35(18):6227-37.

33 Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 11(1):31-46.

34 Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, et al. New goals for the U.S. Human genome project: 1998-2003. Science. 1998;282(5389):682-9.

35 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860-921.

36 International HapMap Consortium. The International HapMap Project. Nature. 2003;426(6968):789-96.

37 Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet. 2001;27:234-6.

38 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.

39 Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annual Review of Genomics and Human Genetics. 2009;10(1):451-81.

40 Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516-7.

41 Lander ES. The new genomics: Global views of biology. Science. 1996;274(5287):536-9.

42 National Center for Biotechnology Information dbSNP. [cited 2010 24th February]; Available from: http://www.ncbi.nlm.nih.gov/snp

43 Ensembl genome browser. [cited 2009 1st May]; Available from: http://www.ensembl.org/index.html

44 The ucsc genome bioinformatics database. [cited 2010 6th January]; Available from: http://genome.ucsc.edu/

45 International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320.

46 International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.

47 The International HapMap Project. [cited 2009 1st May]; Available from: http://www.hapmap.org

48 Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217-22.

49 Chakravarti A. Nonuniform recombination within the human beta-globin gene cluster. Am J Hum Genet. 1984;36:1239-58.

50 Johnson GCL. Haplotype tagging for the identification of common disease genes. Nature Genet. 2001;29:233-7.

51 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001;29:229-32.

52 Collins FS, Guyer MS, Chakravarti A. Variations on a theme: Cataloging human DNA sequence variation. Science. 1997;278:1580-1.

53 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225-9.

54 Terwilliger JD, Hiekkalinna T. An utter refutation of the `fundamental theorem of the HapMap'. Eur J Hum Genet. 2006;14(4):426-37.

55 Affymetrix website: Affymetrix genome-wide human SNP array 6.0. [cited 2010 27th March]; Available from: http://www.affymetrix.com/browse/products.jsp?productId=131533&navMode=34000&navAction=jump&aId=productsNav#1_1

56 Brown MS, Goldstein JL. Familial hypercholesterolemia: Defective binding of lipoproteins to cultured fibroblasts associated with impaired regulation of 3 hydroxy 3 methylglutaryl coenzyme a reductase activity. PNAS. 1974;71(3):788-92.

57 Brown MS, Dana SE, Goldstein JL. Receptor dependent hydrolysis of cholesteryl esters contained in plasma low density lipoprotein. PNAS. 1975;72(8):2925-9.

58 Goldstein JL, Brown MS. Familial hypercholesterolemia: Identification of a defect in the regulation of 3 hydroxy 3 methylglutaryl coenzyme a reductase activity associated with overproduction of cholesterol. PNAS. 1973;70(10):2804-8.

59 Brown MS, Goldstein JL. A receptor-mediated pathway for cholesterol homeostasis. Science. 1986;232(4746):34-47.

60 Slack J. Risks of ischaemic heart disease in familial hyperlipoproteinaemic states. Lancet. 1969;294(7635):1380-2.

61 Stone NJ, Levy RI, Fredrickson DS, Verter J. Coronary artery disease in 116 kindred with familial type II hyperlipoproteinemia. Circulation. 1974;49(3):476-88.

62 Varret M, Abifadel M, Rabes JP, Boileau C. Genetic heterogeneity of autosomal dominant hypercholesterolemia. Clinical Genetics. 2008;73(1):1-13.

63 Hovingh GK, Brownlie A, Bisoendial RJ, Dube MP, Levels JHM, Petersen W, et al. A novel ApoA-I mutation (l178p) leads to endothelial dysfunction, increased arterial wall thickness, and premature coronary artery disease. JACC 2004;44(7):1429-35.

64 Miller M, Aiello D, Pritchard H, Friel G, Zeller K. Apolipoprotein A-I(zavalla) (leu159-pro): HDL cholesterol deficiency in a kindred associated with premature coronary artery disease. Arterioscler Thromb Vasc Biol. 1998;18(8):1242-7.

65 Gualandri V, Franceschini G, Sirtori CR. Ai(milano) apoprotein identification of the complete kindred and evidence of a dominant genetic transmission. Am J Hum Gen. 1985;37(6):1083-97.

66 Clee SM, Kastelein JJP, Van Dam M, Marcil M, Roomp K, Zwarts KY, et al. Age and residual cholesterol efflux affect HDL cholesterol levels and coronary artery disease in abca1 heterozygotes. J Clin Invest. 2000;106(10):1263-70.

67 Mudd SHL, H. L.; Skovby, F. : . Disorders of transsulfuration. In: Scriver CRB, A. L.; Sly, W. S.; Valle, D., ed. *The metabolic and molecular bases of inherited disease*. New York: McGraw-Hill 1995:1279-327.

68 Mudd SH, Skovby F, Levy HL. The natural history of homocystinura due to cystathionine beta-synthase deficiency. Am J Hum Gen. 1985;37(1):1-31.

69 Mani A, Radhakrishnan J, Wang H, Mani A, Mani M-A, Nelson-Williams C, et al. LRP6 mutation in a family with early coronary disease and metabolic risk factors. Science. 2007;315(5816):1278-82.

70 Wang L, Fan C, Topol SE, Topol EJ, Wang Q. Mutation of mef2a in an inherited disorder with features of coronary artery disease. Science. 2003;302:1578-81.

71 Weng L. Lack of mef2a mutations in coronary artery disease. J Clin Invest. 2005;115:1016-20.

72 Tosi I, Toledo-Leiva P, Neuwirth C, Naoumova RP, Soutar AK. Genetic defects causing familial hypercholesterolaemia: Identification of deletions and duplications in the LDL-receptor gene and summary of all mutations found in patients attending the Hammersmith hospital lipid clinic. Atherosclerosis. 2007;194(1):102-11.

73 von Eckardstein A. Differential diagnosis of familial high density lipoprotein deficiency syndromes. Atherosclerosis. 2006;186(2):231-9.

74 Franceschini G, Sirtori CR, Capurso A. A-i(milano) apoprotein. Decreased high density lipoprotein cholesterol levels with significant lipoprotein modifications and without clinical atherosclerosis in an Italian family. J Clin Invest. 1980;66(5):892-900.

75 Sirtori CR, Calabresi L, Franceschini G, Baldassarre D, Amato M, Johansson J, et al. Cardiovascular status of carriers of the apolipoprotein A-I Milano mutant : The Limone Sul Garda study. Circulation. 2001;103(15):1949-54.

76 Weisgraber KH, Rall SC, Jr., Bersot TP, Mahley RW, Franceschini G, Sirtori CR. Apolipoprotein a-imilano. Detection of normal A-I in affected subjects and evidence for a cysteine for arginine substitution in the variant A- I. J Biol Chem. 1983;258(4):2508-13.

77 Shah PK, Yano J, Reyes O, Chyu K-Y, Kaul S, Bisgaier CL, et al. High-dose recombinant apolipoprotein A-I Milano mobilizes tissue cholesterol and rapidly reduces plaque lipid and macrophage content in apolipoprotein E-deficient mice : Potential implications for acute plaque stabilization. Circulation. 2001;103(25):3047-50.

78 Ameli S, Hultgardh-Nilsson A, Cercek B, Shah PK, Forrester JS, Ageland H, et al. Recombinant apolipoprotein A-I Milano reduces intimal thickening after balloon injury in hypercholesterolemic rabbits. Circulation. 1994;90(4 I):1935-41.

79 Shah PK, Nilsson J, Kaul S, Fishbein MC, Ageland H, Hamsten A, et al. Effects of recombinant apolipoprotein A-I(Milano) on aortic atherosclerosis in apolipoprotein E-deficient mice. Circulation. 1998;97(8):780-5.

80 Eberini I, Gianazza E, Calabresi L, Sirtori CR. ApoA-I Milano from structure to clinical application. Annals of Medicine. 2008;40(SUPPL. 1):48-56.

81 Nissen SE, Tsunoda T, Tuzcu EM, Schoenhagen P, Cooper CJ, Yasin M, et al. Effect of recombinant apoA-I Milano on coronary atherosclerosis in patients with acute coronary syndromes: A randomized controlled trial. JAMA. 2003;290(17):2292-300.

82 Arnett DK, Baird AE, Barkley RA, Basson CT, Boerwinkle E, Ganesh SK, et al. Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: A scientific statement from the American Heart Association council on epidemiology and prevention, the stroke council, and the functional genomics and translational biology interdisciplinary working group. Circulation. 2007;115(22):2878-901.

83 Sterne JAC, Smith GD. Sifting the evidence - what's wrong with significance tests? BMJ. 2001;322(7280):226-31.

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

85 Broeckel U, Hengstenberg C, Mayer B, Holmer S, Martin LJ, Comuzzie AG, et al. A comprehensive linkage analysis for myocardial infarction and its related risk factors. Nat Genet. 2002;30(2):210-4.

86 Chiodini BD, Lewis CM. Meta-analysis of 4 coronary heart disease genome-wide linkage studies confirms a susceptibility locus on chromosome 3q. Arterioscler Thromb Vasc Biol. 2003;23(10):1863-8.

87 Farrall M, Green FR, Peden JF, Olsson PG, Clarke R, Hellenius M-L, et al. Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. PLoS Genetics. 2006;2(5):e72.

88 Francke S, Manraj M, Lacquemant C, Lecoeur C, Lepretre F, Passa P, et al. A genome-wide scan for coronary heart disease suggests in Indo-Mauritians a susceptibility locus on chromosome 16p13 and replicates linkage with the metabolic syndrome on 3q27. Hum Mol Genet. 2001;10(24):2751-65.

89 Harrap SB, Zammit KS, Wong ZYH, Williams FM, Bahlo M, Tonkin AM, et al. Genome-wide linkage analysis of the acute coronary syndrome suggests a locus on chromosome 2. Arterioscler Thromb Vasc Biol. 2002;22(5):874-8.

90 Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJH, Mooser V, et al. A genomewide scan for early-onset coronary artery disease in 438 families: The Genecard study. Am J Hum Gen. 2004;75(3):436-47.

91 Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, Thorsteinsdottir U, et al. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat Genet. 2004;36(3):233-9.

92 Pajukanta P, Cargill M, Viitanen L, Nuotio I, Kareinen A, Perola M, et al. Two loci on chromosomes 2 and X for premature coronary heart disease identified in early- and late-settlement populations of Finland. Am J Hum Gen. 2000;67(6):1481-93.

93 Wang Q, Rao S, Shen G-Q, Li L, Moliterno DJ, Newby LK, et al. Premature myocardial infarction novel susceptibility locus on chromosome 1p34-36 identified by genomewide linkage analysis. Am J Hum Gen. 2004;74(2):262-71.

94 The BHF Family Heart Study Research Group. A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) family heart study. Am J Hum Gen. 2005;77(6):1011-20.

95 Zee RYL, Cheng S, Hegener HH, Erlich HA, Ridker PM. Genetic variants of arachidonate 5-lipoxygenase-activating protein, and risk of incident myocardial infarction and ischemic stroke: A nested case-control approach. Stroke. 2006;37(8):2007-11.

96 Girelli D, Martinelli N, Trabetti E, Olivieri O, Cavallari U, Malerba G, et al. Alox5ap gene variants and risk of coronary artery disease: An angiography-based study. Eur J Hum Genet. 2007;15(9):959-66.

97 Koch W, Hoppmann P, Mueller JC, Scho?mig A, Kastrati A. No association of polymorphisms in the gene encoding 5-lipoxygenase- activating protein and myocardial infarction in a large central European population. Genetics in Medicine. 2007;9(2):123-9.

98 Assimes T, Knowles J, Priest J, Basu A, Volcik K, Southwick A, et al. Common polymorphisms of ALOX5 and ALOX5AP and risk of coronary artery disease. Human Genetics. 2008;123(4):399-408.

99 Linsel-Nitschke P, Götz A, Medack A, König IR, Bruse P, Lieb W, et al. Genetic variation in the arachidonate 5-lipoxygenase-activating protein (ALOX5AP) is associated with myocardial infarction in the German population. Clin Sci (Lond). 2008;115(10):309-15.

100 Hakonarson H, Thorvaldsson S, Helgadottir A, Gudbjartsson D, Zink F, Andresdottir M, et al. Effects of a 5-lipoxygenase-activating protein inhibitor on biomarkers associated with risk of myocardial infarction: A randomized trial. JAMA. 2005;293(18):2245-56.

101 Thompson A, Di Angelantonio E, Sarwar N, Erqou S, Saleheen D, Dullaart RPF, et al. Association of cholesteryl ester transfer protein genotypes with CETP mass and activity, lipid levels, and coronary risk. JAMA. 2008;299(23):2777-88.

102 Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, Ahlbom A, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. JAMA. 2007;298(11):1300-11.

103 Ntzani EE, Rizos EC, Ioannidis JPA. Genetic effects versus bias for candidate polymorphisms in myocardial infarction: Case study and overview of large-scale evidence. Am J Epidemiol. 2007;165(9):973-84.

104 Ye Z, Liu EH, Higgins JP, Keavney BD, Lowe GD, Collins R, et al. Seven haemostatic gene polymorphisms in coronary disease: Meta-analysis of 66,155 cases and 91,307 controls. Lancet. 2006;367(9511):651-8.

105 Wheeler JG, Keavney BD, Watkins H, Collins R, Danesh J. Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: Meta-analysis of 43 studies. Lancet. 2004;363(9410):689-95.

106 Casas JP, Bautista LE, Humphries SE, Hingorani AD. Endothelial nitric oxide synthase genotype and ischemic heart disease: Meta-analysis of 26 studies involving 23028 subjects. Circulation. 2004;109(11):1359-65.

107 Chiodini BD, Barlera S, Franzosi MG, Beceiro VL, Introna M, Tognoni G. Apo B gene polymorphisms and coronary artery disease: A meta-analysis. Atherosclerosis. 2003;167(2):355-66.

108 Klerk M, Verhoef P, Clarke R, Blom HJ, Kok FJ, Schouten EG, et al. MTHFR 677C>T polymorphism and risk of coronary heart disease: A meta-analysis. JAMA. 2002;288(16):2023-31.

109 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356-69.

110 Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 8(1):e1000294.

111 Ozaki K. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002;32:650-4.

112 Iwanaga Y. Association analysis between polymorphisms of the lymphotoxin-alpha gene and myocardial infarction in a Japanese population. Atherosclerosis. 2004;172:197-8.

113 PROCARDIS Consortium. A trio family study showing association of the lymphotoxin-alpha n26 (804a) allele with coronary artery disease. Eur J Hum Genet. 2004;12:770-4.

114 Schreyer SA, Vick CM, LeBoeuf RC. Loss of lymphotoxin-alpha but not tumor necrosis factor-alpha reduces atherosclerosis in mice. J Biol Chem. 2002;277:12364-8.

115 Clarke R, Xu P, Bennett D, Lewington S, Zondervan K, Parish S, et al. Lymphotoxin-alpha gene and risk of myocardial infarction in 6,928 cases and 2,712 controls in the ISIS case-control study. PLoS Genetics. 2006;2(7):e107.

116 Karvanen J, Silander K, Kee F, Tiret L, Salomaa V, Kuulasmaa K, et al. The impact of newly identified loci on coronary heart disease, stroke and total mortality in the Morgam prospective cohorts. Genetic Epidemiology. 2009;33(3):237-46.

117 Iles MM. What can genome-wide association studies tell us about the genetics of common disease? PLoS Genetics. 2008;4(2):e33.

118 Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet. 2008;40(2):189-97.

119 Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, Warnes GR, et al. Genome-wide scan identifies variation in MlXIPL associated with plasma triglycerides. Nat Genet. 2008;40(2):149-51.

120 Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, et al. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: Serum urate and dyslipidemia. Am J Hum Gen. 2008;82(1):139-49.

121 Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008;40(2):161-9.

122 Frayling TM, McCarthy MI. Genetic studies of diabetes following the advent of the genome-wide association study: Where do we go from here? Diabetologia. 2007;50(11):2229-33.

123 Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, et al. A variant in cdkal1 influences insulin response and risk of type 2 diabetes. Nat Genet. 2007;39(6):770-5.

124 Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007;316(5829):1336-41.

125 Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007;316(5829):1341-5.

126 Diabetes Genetics Initiative of Broad Institute of Harvard and MIT Lund University and Novartis Institutes of BioMedical Research. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007;316:1331-6.

127 Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.

128 Salonen JT, Uimari P, Aalto J-M, Pirskanen M, Kaikkonen J, Todorova B, et al. Type 2 diabetes whole-genome association study in four populations: The Diagen consortium. The Am J Hum Gen. 2007;81(2):338-45.

129 Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, et al. Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet. 2008;17(6):806-14.

130 Helgadottir A, Thorleifsson G, Magnusson KP, Gretarsdottir S, Steinthorsdottir V, Manolescu A, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 2008;40(2):217-24.

131 Levy D, Larson M, Benjamin E, Newton-Cheh C, Wang T, Hwang S-J, et al. Framingham heart study 100k project: Genome-wide associations for blood pressure and arterial stiffness. BMC Medical Genetics. 2007;8(Suppl 1):S3.

132 Kato N. High-density association study and nomination of susceptibility genes for hypertension in the Japanese national project. Hum Mol Genet. 2008;17:617-27.

133 Sabatti C. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet. 2008;41:35-46.

134 Wang Y, O'Connell JR, McArdle PF, Wade JB, Dorff SE, Shah SJ, et al. Whole-genome association study identifies STK39 as a hypertension susceptibility gene. PNAS. 2009;106(1):226-31.

135 Hsueh WC, Mitchell BD, Aburomia R, Pollin T, Sakul H, Ehm MG, et al. Diabetes in the old order amish: Characterization and heritability analysis of the Amish family diabetes study. Diabetes Care. 2000;23(5):595-601.

136 Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology. 2008;32(2):179-85.

137 Delpire E, Gagnon KBE. SPAK and osr1: Ste20 kinases involved in the regulation of ion homoeostasis and volume control in mammalian cells. Biochemical Journal. 2008;409(2):321-31.

138 Moriguchi T, Urushiyama S, Hisamoto N, Iemura SI, Uchida S, Natsume T, et al. WNK1 regulates phosphorylation of cation-chloride-coupled cotransporters via the ste20-related kinases, spak and osr1. J Biol Chem. 2005;280(52):42685-93.

139 Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009;41(6):666-76.

140 Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, et al. Genome-wide association study of blood pressure and hypertension. Nat Genet. 2009;41(6):677-87.

141 Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A, Huang H, et al. A genome-wide association study of hypertension and blood pressure in African Americans. PLoS Genet. 2009;5(7):e1000564.

142 Org E, Eyheramendy S, Juhanson P, Gieger C, Lichtner P, Klopp N, et al. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two european populations. Hum Mol Genet. 2009;18(12):2288-96.

143 Schunkert H, Gotz A, Braund P, McGinnis R, Tregouet DA, Mangino M, et al. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation. 2008;117(13):1675 - 84.

144 Talmud PJ, Cooper JA, Palmen J, Lovering R, Drenos F, Hingorani AD, et al. Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. Clin Chem. 2008;54(3):467-74.

145 Palomaki GE, Melillo S, Bradley LA. Association between 9p21 genomic markers and heart disease: A meta-analysis. JAMA. 2010;303(7):648-56.

146 Coronary Artery Disease Consortium. Large scale association analysis of novel genetic loci for coronary artery disease. Arterioscler Thromb Vasc Biol. 2009;29(5):774-80.

147 Hiura Y, Fukushima Y, Yuno M, Sawamura H, Kokubo Y, Okamura T, et al. Validation of the association of genetic variants on chromosome 9p21 and 1q41 with myocardial infarction in a Japanese population. Circulation Journal. 2008;72(8):1213-7.

148 Hinohara K, Nakajima T, Takahashi M, Hohda S, Sasaoka T, Nakahara K-i, et al. Replication of the association between a chromosome 9p21 polymorphism and coronary artery disease in Japanese and Korean populations. J Hum Genet. 2008;53(4):357-9.

149 Assimes TL, Knowles JW, Basu A, Iribarren C, Southwick A, Tang H, et al. Susceptibility locus for clinical and subclinical coronary artery disease at chromosome 9p21 in the multi-ethnic advance study. Hum Mol Genet. 2008;17(15):2320-8.

150 Maitra A, Dash D, John S, Sannappa PR, Das AP, Shanker J, et al. A common variant in chromosome 9p21 associated with coronary artery disease in Asian Indians. J Genet. 2009;88(1):113-8.

151 Ding H, Xu Y, Wang X, Wang Q, Zhang L, Tu Y, et al. 9p21 is a shared susceptibility locus strongly for coronary artery disease and weakly for ischemic stroke in Chinese Han population. Circ Cardiovasc Genet. 2009;2(4):338-46.

152 Peng WH, Lu L, Zhang Q, Zhang RY, Wang LJ, Yan XX, et al. Chromosome 9p21 polymorphism is associated with myocardial infarction but not with clinical outcome in Han Chinese. Clin Chem Lab Med. 2009;47(8):917-22.

153 Zhou L, Zhang X, He Ma, Cheng L, Chen Y, Hu FB, et al. Associations between single nucleotide polymorphisms on chromosome 9p21 and risk of coronary heart disease in Chinese Han population. Arterioscler Thromb Vasc Biol. 2008;28(11):2085-9.

154 Ye S, Willeit J, Kronenberg F, Xu Q, Kiechl S. Association of genetic variation on chromosome 9p21 with susceptibility and progression of atherosclerosis: A population-based, prospective study. JACC. 2008;52(5):378-84.

155 Holdt LM, Beutner F, Scholz M, Gielen S, GÃ¤bel G, Bergert H, et al. ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. Arterioscler Thromb Vasc Biol. 2010;30(3):620-7.

156 Cunnington MS, Mayosi BM, Hall DH, Avery PJ, Farrall M, Vickers MA, et al. Novel genetic variants linked to coronary artery disease by genome-wide association are not associated with carotid artery intima-media thickness or intermediate risk phenotypes. Atherosclerosis. 2009;203(1):41-4.

157 Yamagishi K, Folsom AR, Rosamond WD, Boerwinkle E, for the AI. A genetic variant on chromosome 9p21 and incident heart failure in the ARIC study. Eur Heart J. 2009;30(10):1222-8.

158 Zee RYL, Ridker PM. Two common gene variants on chromosome 9 and risk of atherothrombosis. Stroke. 2007;38(10):e111-.

159 Gschwendtner A, Bevan S, Cole JW, Plourde A, Matarin M, Ross-Adams H, et al. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. Annals of Neurology. 2009;65(5):531-9.

160 Matarin M, Brown WM, Singleton A, Hardy JA, Meschia JF, for the Ii. Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke. 2008;39(5):1586-9.

161 Hu W-l, Li S-j, Liu D-t, Wang Y, Niu S-q, Yang X-c, et al. Genetic variants on chromosome 9p21 and ischemic stroke in Chinese. Brain Research Bulletin. 2009;79(6):431-5.

162 Wahlstrand B, Orho-Melander M, Delling L, Kjeldsen S, Narkiewicz K, Almgren P, et al. The myocardial infarction associated CDKN2A/CDKN2B locus on chromosome 9p21 is associated with stroke independently of coronary events in patients with hypertension. J Hypertension. 2009;27(4):769-73.

163 Björck HM, Länne T, Alehagen U, Persson K, Rundkvist L, Hamsten A, et al. Association of genetic variation on chromosome 9p21.3 and arterial stiffness. J Intern Med. 2009;265(3):373-81.

164 Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat Genet. 2009;41(8):899-904.

165 Wrensch M, Jenkins RB, Chang JS, Yeh R-F, Xiao Y, Decker PA, et al. Variants in the CDKN2B and RTELL regions are associated with high-grade glioma susceptibility. Nat Genet. 2009;41(8):905-8.

166 Falchi M, Bataille V, Hayward NK, Duffy DL, Bishop JAN, Pastinen T, et al. Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. Nat Genet. 2009;41(8):915-9.

167 Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, et al. Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 2009;41(8):920-5.

168 Stacey SN, Sulem P, Masson G, Gudjonsson SA, Thorleifsson G, Jakobsdottir M, et al. New common variants affecting susceptibility to basal cell carcinoma. Nat Genet. 2009;41(8):909-14.

169 Kumar R, Smeds J, Berggren P, Straume O, Rozell BL, Akslen LA, et al. A single nucleotide polymorphism in the 3'untranslated region of the CDKN2A gene is common in sporadic primary melanomas but mutations in the CDKN2B, CDKN2C, CDK4 and p53 genes are rare. Int J Cancer. 2001;95(6):388-93.

170 Straume O, Smeds J, Kumar R, Hemminki K, Akslen LA. Significant impact of promoter hypermethylation and the 540 C>T polymorphism of CDKN2A in cutaneous melanoma of the vertical growth phase. Am J Pathol. 2002;161(1):229-37.

171 Sakano S, Berggren P, Kumar R, Steineck G, Adolfsson J, Onelov E, et al. Clinical course of bladder neoplasms and single nucleotide polymorphisms in the CDKN2A gene. Int J Cancer. 2003;104(1):98-103.

172 Debniak T, Scott RJ, Huzarski T, Byrski T, Rozmiarek A, Debniak B, et al. CDKN2A common variants and their association with melanoma risk: A population-based study. Cancer Res. 2005;65(3):835-9.

173 Debniak T, Gorski B, Huzarski T, Byrski T, Cybulski C, Mackiewicz A, et al. A common variant of CDKN2A (p16) predisposes to breast cancer. J Med Genet. 2005;42(10):763-5.

174 Dębniak T, Cybulski C, Górski B, Huzarski T, Byrski T, Gronwald J, et al. CDKN2A-positive breast cancers in young women from Poland. Breast Cancer Research and Treatment. 2007;103(3):355-9.

175 Chen J, Li D, Wei C, Sen S, Killary AM, Amos CI, et al. Aurora-a and p16 polymorphisms contribute to an earlier age at diagnosis of pancreatic cancer in Caucasians. Clin Cancer Res. 2007;13(10):3100-4.

176 Gayther SA, Song H, Ramus SJ, Kjaer SK, Whittemore AS, Quaye L, et al. Tagging single nucleotide polymorphisms in cell cycle control genes and susceptibility to invasive epithelial ovarian cancer. Cancer Research. 2007;67(7):3027-35.

177 Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2):683-92.

178 Kang MY, Lee BB, Ji YI, Jung EH, Chun H-K, Song SY, et al. Association of interindividual differences in p14ARF promoter methylation with single nucleotide polymorphism in primary colorectal cancer. Cancer. 2008;112(8):1699-707.

179 Melzer D, Frayling TM, Murray A, Hurst AJ, Harries LW, Song H, et al. A common variant of the p16INK4a genetic region is associated with physical function in older people. Mechanisms of Ageing and Development. 2007;128(5-6):370-7.

180 Driver KE, Song H, Lesueur F, Ahmed S, Barbosa-Morais NL, Tyrer JP, et al. Association of single-nucleotide polymorphisms in the cell cycle genes with breast cancer in the British population. Carcinogenesis. 2008;29(2):333-41.

181 Yan L, Na W, Shan K, Xiao-Wei M, Wei G, Shu-Cheng C. P16(cdkn2) gene polymorphism: Association with histologic subtypes of epithelial ovarian cancer in china. Int J Gynecolog Cancer. 2008;18(1):30-5.

182 Züchner S, Gilbert JR, Martin ER, Leon-Guerrero CR, Xu PT, Browning C, et al. Linkage and association study of late-onset Alzheimer disease families linked to 9p21.3. An Hum Gen. 2008;72(6):725-31.

183 Kim WY, Sharpless NE. The regulation of INK4/ARF in cancer and aging. Cell. 2006;127(2):265-75.

184 Gil J, Peters G. Regulation of the INK4B-ARF-INK4A tumour suppressor locus: All for one or one for all. Nat Rev Mol Cell Biol. 2006;7(9):667-77.

185 Cánepa ET, Scassa ME, Ceruti JM, Marazita MC, Carcagno AL, Sirkin PF, et al. INK4 proteins, a family of mammalian CDK inhibitors with novel biological functions. IUBMB Life. 2007;59(7):419-26.

186 Hara E, Smith R, Parry D, Tahara H, Stone S, Peters G. Regulation of p16CDKN2 expression and its implications for cell immortalization and senescence. Mol Cell Biol. 1996;16(3):859-67.

187 Robertson KD, Jones PA. The human ARF cell cycle regulatory gene promoter is a CpG island which can be silenced by DNA methylation and down-regulated by wild-type p53. Mol Cell Biol. 1998;18(11):6457-73.

188 Li J-M, Nichols MA, Chandrasekharan S, Xiong Y, Wang X-F. Transforming growth factor beta activates the promoter of cyclin-dependent kinase inhibitor p15 (INK4b) through an Spl consensus site. J Biol Chem. 1995;270(45):26750-3.

189 Gonzalez S, Klatt P, Delgado S, Conde E, Lopez-Rios F, Sanchez-Cespedes M, et al. Oncogenic activity of CDC6 through repression of the INK4/ARF locus. Nature. 2006;440(7084):702-6.

190 Hannon GJ, Beach D. Pl5ink4b is a potentia| effector of TGF-beta-induced cell cycle arrest. Nature. 1994;371(6494):257-61.

191 Kalinina N, Agrotis A, Antropova Y, Ilyinskaya O, Smirnov V, Tararak E, et al. Smad expression in human atherosclerotic lesions: Evidence for impaired TGF-beta/SMAD signaling in smooth muscle cells of fibrofatty lesions. Arterioscler Thromb Vasc Biol. 2004;24(8):1391-6.

192 Dzau VJ, Braun-Dullaeus RC, Sedding DG. Vascular proliferation and atherosclerosis: New perspectives and therapeutic strategies. Nat Med. 2002;8(11):1249-56.

193 Andreassi M. DNA damage, vascular senescence and atherosclerosis. J Molec Med. 2008;86(9):1033-43.

194 Minamino T, Komuro I. Vascular cell senescence: Contribution to atherosclerosis. Circ Res. 2007;100(1):15-26.

195 Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D, Bieche I. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: Identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res. 2007;67(8):3963-9.

196 National Center for Biotechnology Information. [cited 2010 24th February]; Available from: http://www.ncbi.nlm.nih.gov/

197 Folkersen L, Kyriakou T, Goel A, Peden J, Malarstig A, Paulsson-Berne G, et al. Relationship between CAD risk genotype in the chromosome 9p21 locus and gene expression. Identification of eight new ANRIL splice variants. PLoS ONE. 2009;4(11):e7677.

198 Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet. 2006;15(suppl_1):R17-29.

199 Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: Insights into functions. Nat Rev Genet. 2009;10(3):155-9.

200 Jarinova O, Stewart AFR, Roberts R, Wells G, Lau P, Naing T, et al. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. Arterioscler Thromb Vasc Biol. 2009;29(10):1671-7.

201 Bracken AP, Kleine-Kohlbrecher D, Dietrich N, Pasini D, Gargiulo G, Beekman C, et al. The polycomb group proteins bind throughout the INK4a-ARF locus and are disassociated in senescent cells. Genes & Development. 2007;21(5):525-30.

202 Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature. 2008;451(7175):202-6.

203 Christopher SA, Diegelman P, Porter CW, Kruger WD. Methylthioadenosine phosphorylase, a gene frequently codeleted with p16CDKN2A/ARF, acts as a tumor suppressor in a breast cancer cell line. Cancer Res. 2002;62(22):6639-44.

204 Brautbar A, Ballantyne CM, Lawson K, Nambi V, Chambless L, Folsom AR, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of

coronary heart disease risk and implications for lipid-modifying therapy in the atherosclerosis risk in communities study. Circ Cardiovasc Genet. 2009;2(3):279-85.

205 Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med. 2009;150(2):65-72.

206 Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, Miletich JP, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010;303(7):631-7.

207 Decode genetics. [cited 2010 3rd April]; Available from: http://www.decode.com/

208 Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: A randomised placebo-controlled trial. Lancet. 2002;360(9326):7-22.

209 Grant SFA, Hakonarson H. Recent development in pharmacogenomics: From candidate genes to genome-wide association studies. Expert Review of Molecular Diagnostics. 2007;7(4):371-93.

210 McCarthy MI. Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. Hum Mol Genet. 2004;13(REV. ISS. 1).

211 Buckland PR. Allele-specific gene expression differences in humans. Hum Mol Genet. 2004;13(suppl_2):R255-60.

212 Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461(7261):199-205.

213 Antonarakis SE, Irkin SH, Cheng TC. Beta -thalassemia in american blacks: Novel mutations in the 'TATA' box and an acceptor splice site. PNAS. 1984;81(4 I):1154-8.

214 Koivisto UM, Palvimo JJ, Janne OA, Kontula K. A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. PNAS. 1994;91(22):10526-30.

215 Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature. 2003;423(6939):506-11.

216 Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, et al. Functional haplotypes of padi4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet. 2003;34(4):395-402.

217 Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, et al. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet. 2000;26(2):163-75.

218 Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. Hum Mol Genet. 2006;15(suppl_1):R9-16.

219 Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009:1174148.

220 Zindy F, Quelle DE, Roussel MF, Sherr CJ. Expression of the p16(INK4A) tumor suppressor versus other INK4 family members during mouse development and aging. Oncogene. 1997;15(2):203-11.

221 Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. Individuality and variation in gene expression patterns in human blood. PNAS. 2003;100(4):1896-901.

222 Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics. 2003;33(3):422-5.

223 Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365-9.

224 Morley M. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430:743-7.

225 Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. Science. 2004;306(5696):647-50.

226 Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific SHH expression and truncation of the mouse limb. Development. 2005;132:797-803.

227 Dixon AL. A genome-wide association study of global gene expression. Nat Genet. 2007;39:1202-7.

228 Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, et al. Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. PLoS Genet. 2008;4(2).

229 Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KCL, Koka V, et al. Targeted screening of cis-regulatory variation in human haplotypes. Genome Research. 2009;19(1):118-27.

230 Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848-53.

231 Lettice LA. A long-range SHH enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003;12:1725-35.

232 Furniss D. A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. Hum Mol Genet. 2008;17:2417-23.

233 Carter AM, Sachchithananthan M, Stasinopoulos S, Maurer F, Medcalf RL. Prothrombin g20210a is a bifunctional gene polymorphism. Thrombosis and Haemostasis. 2002;87(5):846-53.

234 Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, et al. Increased efficiency of mRNA 3' end formation: A new genetic mechanism contributing to hereditary thrombophilia. Nat Genet. 2001;28(4):389-93.

235 Ober C, Aldrich CL, Chervoneva I, Billstrand C, Rahimov F, Gray HL, et al. Variation in the hla-g promoter region influences miscarriage rates. Am J Hum Genet. 2003;72(6):1425-35.

236 Bidichandani SI, Ashizawa T, Patel PI. The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. T Am J Hum Genet. 1998;62(1):111-21.

237 Sutcliffe JS, Nelson DL, Zhang F, Pieretti M, Caskey CT, Saxe D, et al. DNA methylation represses FMR-1 transcription in fragile X syndrome. Hum Mol Genet. 1992;1(6):397-400.

238 Hamada H, Seidman M, Howard BH, Gorman CM. Enhanced gene expression by the poly(dT-dG).Poly(dC-dA) sequence. Mol Cell Biol. 1984;4(12):2622-30.

239 Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet. 2002;30(3):315-20.

240 Kennedy GC, German MS, Rutter WJ. The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. Nat Genet. 1995;9(3):293-8.

241 Albanese V, Biguet NF, Kiefer H, Bayard E, Mallet J, Meloni R. Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. Hum Mol Genet. 2001;10:1785 - 92.

242 Pieretti M, Zhang F, Fu Y-H, Warren ST, Oostra BA, Caskey CT, et al. Absence of expression of the FMR-1 gene in fragile X syndrome. Cell. 1991;66(4):817-22.

243 Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. Nat Genet. 2000;26(1):61-3.

244 Yang MQ, Koehly LM, Elnitski LL. Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. PLoS Comput Biol. 2007;3(4):e72.

245 Promo: A virtual laboratory for the study of transcription factor binding sites in DNA sequences. [cited 2010 15th March]; Available from: http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3

246 Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM. Promo: Detection of known transcription regulatory elements using species-tailored searches. Bioinformatics. 2002;18(2):333-4.

247 Farre D, Roset R, Huerta M, Adsuara JE, Rosello L, Alba MM, et al. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. Nucl Acids Res. 2003;31(13):3651-3.

248 Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: New insights in genome diversity. Genome Res. 2006;16(8):949-61.

249 Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18(R1):R1-8.

250 Stefansson H, Rujescu D, Cichon S, PietilÃ¤inen OPH, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455(7210):232-6.

251 Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316(5823):445-9.

252 Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nature Genetics. 2006;38(1):82-5.

253 Cirulli ET, Goldstein DB. In vitro assays fail to predict in vivo effects of regulatory polymorphisms. Hum Mol Genet. 2007;16(16):1931-9.

254 Teare MD, Heighway J, Santibanez Koref MF. An expectation-maximization algorithm for the analysis of allelic expression imbalance. Am J Hum Genet. 2006;79(3):539-43.

255 Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, et al. A survey of genetic and epigenetic variation affecting human gene expression. Physiological Genomics. 2003;16:184-93.

256 Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. Science. 2002;297(5584):1143-.

257 Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. Cis-acting variation in the expression of a high proportion of genes in human brain. Human Genetics. 2003;113(2):149-53.

258 Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, et al. Allelic variation in gene expression is common in the human genome. Genome Res. 2003;13(8):1855-62.

259 Pant PVK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. Analysis of allelic differential expression in human white blood cells. Genome Res. 2006;16(3):331-9.

260 Heighway J, Bowers NL, Smith S, Betticher DC, Santibanez Koref MF. The use of allelic expression differences to ascertain functional polymorphisms acting in cis: Analysis of mmp1 transcripts in normal lung tissue. Ann Hum Genet. 2005;69(1):127-33.

261 Milani L, Gupta M, Andersen M, Dhar S, Fryknas M, Isaksson A, et al. Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. Nucl Acids Res. 2007;35(5):e34-.

262 Zhang Y, Wang D, Johnson AD, Papp AC, Sadee W. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem. 2005;280(38):32618-24.

263 Li Y, Grupe A, Rowland C, Nowotny P, Kauwe JSK, Smemo S, et al. DAPK1 variants are associated with Alzheimer's disease and allele-specific expression. Hum Mol Genet. 2006;15(17):2560-8.

264 Wang D, Sadee W. Searching for polymorphisms that affect gene expression and mRNA processing: Example ABCB1 (MDR1). AAPS Journal. 2006;8(3).

265 Tao H, Cox DR, Frazer KA. Allele-specific KRT1 expression is a complex trait. PLoS Genet. 2006;2(6):0848-58.

266 Campino S, Forton J, Raj S, Mohr B, Auburn S, Fry A, et al. Validating discovered cis-acting regulatory genetic variants: Application of an allele specific expression approach to HapMap populations. PLoS ONE. 2008;3(12):e4105.

267 Chen X, Weaver J, Bove BA, Vanderveer LA, Weil SC, Miron A, et al. Allelic imbalance in brca1 and brca2 gene expression is associated with an increased breast cancer risk. Hum Mol Genet. 2008;17(9):1336-48.

268 Butz J, Yan H, Mikkilineni V, Edwards J. Detection of allelic variations of human gene expression by polymerase colonies. BMC Genetics. 2004;5(1):3.

269 Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nature Methods. 2009;6(8):613-8.

270 Main B, Bickel R, McIntyre L, Graze R, Calabrese P, Nuzhdin S. Allele-specific expression assays using solexa. BMC Genomics. 2009;10(1):422.

271 Tuupanen S, Niittymaki I, Nousiainen K, Vanharanta S, Mecklin J-P, Nuorva K, et al. Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution. Cancer Res. 2008;68(1):14-7.

272 Ding C, Cantor CR. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. PNAS. 2003;100(6):3059-64.

273 Ding C, Maier E, Roscher AA, Braun A, Cantor CR. Simultaneous quantitative and allele-specific expression analysis with real competitive PCR. BMC Genetics. 2004;5.

274 People of the British isles. [cited 2010 27th March]; Available from: http://www.peopleofthebritishisles.org/

275 Mathers JC, Mickleburgh I, Chapman PC, Bishop DT, Burn J. Can resistant starch and/or aspirin prevent the development of colon neoplasia? The concerted action polyp prevention (capp) 1 study. Proceedings of the Nutrition Society. 2003;62(01):51-7.

276 Palomino-Doza J, Rahman TJ, Avery PJ, Mayosi BM, Farrall M, Watkins H, et al. Ambulatory blood pressure is associated with polymorphic variation in P2X receptor genes. Hypertension. 2008;52(5):980-5.

277 Lang RM, Bierig M, Devereux RB, Flachskampf FA, Foster E, Pellikka PA, et al. Recommendations for chamber quantification: A report from the American society of echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the European association of echocardiography, a branch of the European society of cardiology. J Am Soc Echocardiography. 2005;18(12):1440-63.

278 Devereux RB, Alonso DR, Lutas EM, Gottlieb GJ, Campo E, Sachs I, et al. Echocardiographic assessment of left ventricular hypertrophy: Comparison to necropsy findings. Am J Cardiol. 1986;57(6):450-8.

279 Levy D, Savage DD, Garrison RJ, Anderson KM, Kannel WB, Castelli WP. Echocardiographic criteria for left ventricular hypertrophy: The Framingham Heart Study. Am J Cardiol. 1987;59(9):956-60.

280 Teichholz LE, Kreulen T, Herman MV, Gorlin R. Problems in echocardiographic volume determinations: Echocardiographic-angiographic correlations in the presence or absence of asynergy. Am J Cardiol. 1976;37(1):7-11.

281 Quiñones MA, Otto CM, Stoddard M, Waggoner A, Zoghbi WA. Recommendations for quantification of Doppler echocardiography: A report from the Doppler quantification task force of the nomenclature and standards committee of the American society of echocardiography. J Am Soc Echocardiography. 2002;15(2):167-84.

282 Vickers MA, Green FR, Terry C, Mayosi BM, Julier C, Lathrop M, et al. Genotype at a promoter polymorphism of the interleukin-6 gene is associated with baseline levels of plasma C-reactive protein. Cardiovasc Res. 2002 Mar;53(4):1029-34.

283 Hall DH, Rahman T, Avery PJ, Keavney B. Insig-2 promoter polymorphism and obesity related phenotypes: Association study in 1428 members of 248 families. BMC Med Genet. 2006;7:83.

284 Griffin HR, Hall DH, Topf A, Eden J, Stuart AG, Parsons J, et al. Genetic variation in VEGF does not contribute significantly to the risk of congenital cardiovascular malformation. PLoS ONE. 2009;4(3):e4978.

285 Chase DS, Tawn EJ, Parker L, Jonas P, Parker CO, Burn J. The north Cumbria community genetics project. J Med Genet. 1998;35(5):413-6.

286 Relton CL, Wilding CS, Pearce MS, Laffling AJ, Jonas PA, Lynch SA, et al. Gene-gene interaction in folate-related genes and risk of neural tube defects in a UK population. J Med Genet. 2004;41(4):256-60.

287 Sambrook J, Russell DW. Molecular cloning: A laboratory manual. New York: Cold Spring Harbor 2001.

288 Sequenom RealSNP assay database. [cited 2009 1st May]; Available from: https://www.realsnp.com/default.asp

289 Rozen S, Skaletsky HJ. Primer3 on the www for general users and for biologist programmers. In: Misener S, Krawetz SA, eds. *Bioinformatics methods and protocols: Methods in molecular biology*. Totowa, NJ: Humana Press 2000:365-86.

290 Primer3 v0.4.0 input website. [cited 2010 24th February]; Available from: http://frodo.wi.mit.edu/primer3/

291 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology. 2002;3(7):research0034.1 - research.11.

292 Zhang X, Ding L, Sandford AJ. Selection of reference genes for gene expression studies in human neutrophils by real-time PCR. BMC Molecular Biology. 2005;6.

293 H-mapd: Human MLPA probe design website. [cited 2009 1st May]; Available from: http://genomics01.arcan.stonybrook.edu/mlpa/cgi-bin/mlpa.cgi

294 Zhi J, Hatchwell E. Human MLPA probe design (h-MAPD): A probe design tool for both electrophoresis-based and bead-coupled human multiplex ligation-dependent probe amplification assays. BMC Genomics. 2008;9(1):407.

295 Wigginton JE, Abecasis GR. Pedstats: Descriptive statistics, graphics and quality assessment for gene mapping data. Bioinformatics. 2005;21(16):3445-7.

296 Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76(5):887-93.

297 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263-5.

298 Weir BS. Genetic data analysis 2: Methods for discrete population genetic data. Sunderland MA: Sinauer 1996.

299 Balding DJ. A tutorial on statistical methods for population association studies. 2006 2006/10//print;7(10):781-91.

300 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97-101.

301 University of michigan center for statistical genetics: Merlin. [cited 2010 1st February]; Available from: http://www.sph.umich.edu/csg/abecasis/merlin/index.html

302 Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Human Heredity. 2008;66(2):87-98.

303 Dudbridge F. Unphased: Software for genetic association analysis [cited 2009 1st May]; Available from: http://www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased/

304 Cunnington MS, Santibanez Koref MF, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 snps associated with multiple disease phenotypes correlate with ANRIL expression. PLoS Genet. 2010;6(4):e1000899.

305 Cran archive. [cited 2009 1st May]; Available from: http://cran.r-project.org/

306 Pinsonneault JK, Papp AC, Sade?e W. Allelic mRNA expression of x-linked monoamine oxidase A (MaOA) in human brain: Dissection of epigenetic and genetic factors. Hum Mol Genet. 2006;15(17):2636-49.

307 Ware MD, DeSilva D, Sinilnikova OM, Stoppa-Lyonnet D, Tavtigian SV, Mazoyer S. Does nonsense-mediated mRNA decay explain the ovarian cancer cluster region of the BRCA2 gene? Oncogene. 2005;25(2):323-8.

308 De Luca V, Strauss J, Semeralul M, Huang S, Li PP, Warsh JJ, et al. Analysis of BDNF Val66Met allele-specific mRNA levels in bipolar disorder. Neuroscience Letters. 2008;441(2):229-32.

309 Sun C, Southard C, Witonsky DB, Olopade OI, Rienzo AD. Allelic imbalance (AI) identifies novel tissue-specific cis-regulatory variation for human UGT2B15. Human Mutation. 2009;31(1):99-107.

310 Forton JT, Udalova IA, Campino S, Rockett KA, Hull J, Kwiatkowski DP. Localization of a long-range cis-regulatory element of IL13 by allelic transcript ratio mapping. Genome Res. 2007 January 1, 2007;17(1):82-7.

311 Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. Nature Protocols. 2006;1(3):1559-82.

312 Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008;40(5):638-45.

313 Duesing K, Fatemifar G, Charpentier G, Marre M, Tichet J, Hercberg S, et al. Strong association of common variants in the VDKN2A/CDKN2B region with type 2 diabetes in French Europids. Diabetologia. 2008;51(5):821-6.

314 Chen J, Li D, Killary A, Sen S, Amos C, Evans D, et al. Polymorphisms of p16 , p27 , p73, and mdm2 modulate response and survival of pancreatic cancer patients treated with preoperative chemoradiation. Annals of Surgical Oncology. 2009;16(2):431-9.

315 Qiagen website: Quantitect whole transcriptome kit. [cited 2010 31st March 2010]; Available from: http://www1.qiagen.com/products/RnaStabilizationPurification/WholeTranscriptomeAmplification/QuantiTectWholeTranscriptomeKit.aspx#Tabs=t1

316 Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. Mirbase: Tools for microRNA genomics. Nucl Acids Res. 2008;36(suppl_1):D154-8.

317 Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, et al. The UCSC genome browser database: Update 2010. Nucl Acids Res. 2010;38(suppl_1):D613-9.

318 Mirbase: The microRNA database. [cited 2010 6th January]; Available from: http://www.mirbase.org/

319 Wang X, Feng Y, Pan L, Wang Y, Xu X, Lu J, et al. The proximal GC-rich region of p16INK4a gene promoter plays a role in its transcriptional regulation. Molecular and Cellular Biochemistry. 2007;301(1):259-66.

320 Wang W, Wu J, Zhang Z, Tong T. Characterization of regulatory elements on the promoter region of p16INK4a that contribute to overexpression of p16 in senescent fibroblasts. J Biol Chem. 2001;276(52):48655-61.

321 Campbell MC, Tishkoff SA. African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. Annual Review of Genomics and Human Genetics. 2008;9(1):403-33.

322 Silander K, Tang H, Myles S, Jakkula E, Timpson NJ, Cavalli-Sforza L, et al. Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease. Genome Medicine. 2009;1(5):51.1-.7.

323 McKenzie CA, Abecasis GR, Keavney B, Forrester T, Ratcliffe PJ, Julier C, et al. Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). Hum Mol Genet. 2001;10(10):1077-84.

324 Superscript™ III first-strand synthesis system for rt-pcr manual. [cited 2010 3rd April]; Available from: http://tools.invitrogen.com/content/sfs/manuals/superscriptIIIfirststrand_pps.pdf

325 Liu Y, Sanoff HK, Cho H, Burd CE, Torrice C, Mohlke KL, et al. INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. PLoS ONE. 2009;4(4):e5027.

326 Rodriguez C, Borgel J, Court F, Cathala G, Forné T, Piette J. Ctcf is a DNA methylation-sensitive positive regulator of the INK/ARF locus. Biochemical and Biophysical Research Communications. 2010;392(2):129-34.

327 Visel A, Zhu Y, May D, Afzal V, Gong E, Attanasio C, et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature. 2010;464(7287):409-12.

328 Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. Trends in Genetics. 2006;22(1):1-5.

329 Liu Y, Sanoff HK, Cho H, Burd CE, Torrice C, Ibrahim JG, et al. Expression of p16INK4a in peripheral blood t-cells is a biomarker of human aging. Aging Cell. 2009;8(4):439-48.

330 Lal A, Kim HH, Abdelmohsen K, Kuwano Y, Pullmann R, Jr., Srikantan S, et al. P16INK4a translation suppressed by mir-24. PLoS ONE. 2008;3(3):e1864.

331 Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet. 2007;39(10):1217-24.

332 Plagnol V, Uz E, Wallace C, Stevens H, Clayton D, Ozcelik T, et al. Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. PLoS ONE. 2008;3(8):e2966.

333 Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread monoallelic expression on human autosomes. Science. 2007;318(5853):1136-40.

334 Narva E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, et al. High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nat Biotech. 2010;28(4):371-7.

335 The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.

336 Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321(5897):1807-12.

337 Berwick M, Orlow I, Hummer AJ, Armstrong BK, Kricker A, Marrett LD, et al. The prevalence of CDKN2A germ-line mutations and relative risk for cutaneous malignant melanoma: An international population-based study. Cancer Epidemiology Biomarkers & Prevention. 2006;15(8):1520-5.

338 Goldstein AM, Chan M, Harland M, Hayward NK, Demenais F, Bishop DT, et al. Features associated with germline CDKN2A mutations: A genomel study of melanoma-prone families from three continents. J Medl Genet. 2007;44(2):99-106.

339 Krishnamurthy J, Ramsey MR, Ligon KL, Torrice C, Koh A, Bonner-Weir S, et al. P16INK4a induces an age-dependent decline in islet regenerative potential. Nature. 2006;443(7110):453-7.

340 Moritani M, Yamasaki S, Kagami M, Suzuki T, Yamaoka T, Sano T, et al. Hypoplasia of endocrine and exocrine pancreas in homozygous transgenic TGF-beta1. Molecular and Cellular Endocrinology. 2005;229(1-2):175-84.

341 Sham P. Statistics in human genetics. London: Arnold Publishers 2001.

342 Khoury MJ, Little J, Burke W. Human genome epidemiology: A scientific foundation for using genetic information to improve health and prevent disease. . New York: Oxford University Press 2004.

343 Shoemaker J, Painter I, Weir BS. A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics. 1998;149(4):2079-88.

344 Salanti G, Amountza G, Ntzani EE, Ioannidis JPA. Hardy-Weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. Eur J Hum Genet. 2005;13(7):840-8.

345 Little J. Reporting and review of human genome epidemiology studies. Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. 2004:168-92.

346 Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, et al. Hardy-Weinberg quality control. Annals of Human Genetics. 1999;63(6):535-8.

347 Zou GY, Donner A. The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: A cautionary note. Ann Hum Genet. 2006;70(6):923-33.

348 Little J, Higgins JPT, Ioannidis JPA, Moher D, Gagnon F, Von Elm E, et al. Strengthening the reporting of genetic association studies (STREGA): An extension of the strobe statement. Eur J Epidemiol. 2009;24(1):37-55.

349 Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. How should we use information about hwe in the meta-analyses of genetic association studies? Internat J Epidemiol. 2008;37(1):136-46.

350 Sun YV, Peyser PA, Kardia SLR. A common copy number variation on chromosome 6 association with the gene expression level of endothelin 1 in transformed b lymphocytes from three racial groups. Circ Cardiovasc Genet. 2009;2(5):483-8.

351 Merikangas AK, Corvin AP, Gallagher L. Copy-number variants in neurodevelopmental disorders: Promises and challenges. Trends in Genetics. 2009;25(12):536-44.

352 Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81.

353 McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, et al. Common deletion polymorphisms in the human genome. Nat Genet. 2006;38(1):86-92.

354 Bailey JA, Eichler EE. Primate segmental duplications: Crucibles of evolution, diversity and disease. Nat Rev Genet. 2006;7(7):552-64.

355 McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40(10):1166-74.

356 Hernandez JL, Weir BS. A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics. 1989;45(1):53-70.

357 Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305(5683):525-8.

358 Database of genomic variants: A curated catalogue of structural variation in the human genome. [cited 2010 24th February]; Available from: http://projects.tcag.ca/variation/

359 Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 2008;82(3):685-95.

360 de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, Tsang P, et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: Implications for association studies of complex diseases. Hum Mol Genet. 2007;16(23):2783-94.

361 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

362 Simon-Sanchez J, Scholz S, Fung H-C, Matarin M, Hernandez D, Gibbs JR, et al. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. 2007;16(1):1-14.

363 Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009;41(3):334-41.

364 Loryn NS, Graham RT. MLPA and MAPH: New techniques for detection of gene deletions. Human Mutation. 2004;23(5):413-9.

365 den Dunnen JT, White SJ. MLPA and MAPH: Sensitive detection of deletions and duplications. Current protocols in human genetics. 2006;51:Unit 7.14.

366 MRC-Holland MLPA website. [cited 2010 24th February]; Available from: http://www.mrc-holland.com/WebForms/WebFormMain.aspx

367 Genetic information from the British 1958 birth cohort [cited 2009 1st May]; Available from: http://www.b58cgene.sgul.ac.uk/

368 Wellcome trust case control consortium: Access to WTCCC genotype data. [cited 2009 1st May]; Available from: https://www.wtccc.org.uk/ccc1/access_to_data_samples.shtml

369 Groth M, Szafranski K, Taudien S, Huse K, Mueller O, Rosenstiel P, et al. High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes. Human Mutation. 2008;29(10):1247-54.

370 Aldred PMR, Hollox EJ, Armour JAL. Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet. 2005;14(14):2045-52.

371 Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet. 2008;83(2):228-42.

372 Bredel M, Bredel C, Juric D, Kim Y, Vogel H, Harsh GR, et al. Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. J Mol Diagn. 2005 May 1, 2005;7(2):171-82.

373 Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, Li HI, et al. Impact of whole genome amplification on analysis of copy number variants. Nucl Acids Res. 2008;36(13):e80-.

374 Pinard R, de Winter A, Sarkis G, Gerstein M, Tartaro K, Plant R, et al. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics. 2006;7(1):216.

375 Rusakova V, Nosek L. Amplification of genome-representative DNA from limited sources with genomeplex WGA technology for use in genetic alterations studies. Nature Methods. 1996;3(3).

376 Fiegler H, Geigl JB, Langer S, Rigler D, Porter K, Unger K, et al. High resolution array-CGH analysis of single cells. Nucl Acids Res. 2007;35(3):e15.

377 Geigl JB, Obenauf AC, Waldispuehl-Geigl J, Hoffmann EM, Auer M, Hormann M, et al. Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. Nucl Acids Res. 2009;37(15):e105.

378 Libby P, Ridker PM, Maseri A. Inflammation and atherosclerosis. Circulation. 2002;105(9):1135-43.

379 Ross R. Mechanisms of disease: Atherosclerosis - an inflammatory disease. N Engl J Med. 1999;340(2):115-26.

380 Woods A, Brull DJ, Humphries SE, Montgomery HE. Genetics of inflammation and risk of coronary artery disease: The central role of interleukin-6. Eur Heart J. 2000;21(19):1574-83.

381 De Maat MPM, Pietersma A, Kofflard M, Sluiter W, Kluft C. Association of plasma fibrinogen levels with coronary artery disease, smoking and inflammatory markers. Atherosclerosis. 1996;121(2):185-91.

382 Tappia PS, Troughton KL, Langley-Evans SC, Grimble RF. Cigarette smoking influences cytokine production and antioxidant defences. Clinical Science. 1995;88(4):485-9.

383 Biasucci LM, Vitelli A, Liuzzo G, Altamura S, Caligiuri G, Monaco C, et al. Elevated levels of interleukin-6 in unstable angina. Circulation. 1996;94(5):874-7.

384 Biasucci LM, Liuzzo G, Grillo RL, Caligiuri G, Rebuzzi AG, Buffon A, et al. Elevated levels of C-reactive protein at discharge in patients with unstable angina predict recurrent instability. Circulation. 1999;99(7):855-60.

385 Biasucci LM, Liuzzo G, Fantuzzi G, Caligiuri G, Rebuzzi AG, Ginnetti F, et al. Increasing levels of interleukin (IL)-1Ra and IL-6 during the first 2 days of hospitalization in unstable angina are associated with increased risk of in-hospital coronary events. Circulation. 1999;99(16):2079-84.

386 Harris TB, Ferrucci L, Tracy RP, Corti MC, Wacholder S, Ettinger Jr WH, et al. Associations of elevated interleukin-6 and C-reactive protein levels with mortality in the elderly. Am J Med. 1999;106(5):506-12.

387 Rauramaa R, Vaisanen SB, Luong L-A, Schmidt-Trucksass A, Penttila IM, Bouchard C, et al. Stromelysin-1 and interleukin-6 gene promoter polymorphisms are determinants of asymptomatic carotid artery atherosclerosis. Arterioscler Thromb Vasc Biol. 2000;20(12):2657-62.

388 Antonicelli R, Olivieri F, Bonafe M, Cavallone L, Spazzafumo L, Marchegiani F, et al. The interleukin-6 -174 G>C promoter polymorphism is associated with a higher risk of death after an acute coronary syndrome in male elderly patients. Internat J Cardiol. 2005;103(3):266-71.

389 Lindahl B, Toss H, Siegbahn A, Venge P, Wallentin L. Markers of myocardial damage and inflammation in relation to long-term mortality in unstable coronary artery disease. N Engl J Med. 2000;343(16):1139-47.

390 Haverkate E, Thompson SG, Pyke SD, Gallimore JR, Group MBP. Production of C-reactive protein and risk of coronary events in stable and unstable angina. Lancet. 1997 1997/2/15;349(9050):462-6.

391 Cesari M, Penninx BWJH, Newman AB, Kritchevsky SB, Nicklas BJ, Sutton-Tyrrell K, et al. Inflammatory markers and onset of cardiovascular events: Results from the health ABC study. Circulation. 2003;108(19):2317-22.

392 Konstantinides S, SchÃ¤fer K, Koschnick S, Loskutoff DJ. Leptin-dependent platelet aggregation and arterial thrombosis suggests a mechanism for atherothrombotic disease in obesity. J Clin Investigat. 2001;108(10):1533-40.

393 Schafer K, Halle M, Goeschen C, Dellas C, Pynn M, Loskutoff DJ, et al. Leptin promotes vascular remodeling and neointimal growth in mice. Arterioscler Thromb Vasc Biol. 2004;24(1):112-7.

394 Parhami F, Tintut Y, Ballard A, Fogelman AM, Demer LL. Leptin enhances the calcification of vascular cells : Artery wall as a target of leptin. Circ Res. 2001;88(9):954-60.

395 Stein JH, Korcarz CE, Hurst RT, Lonn E, Kendall CB, Mohler ER, et al. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: A consensus statement from the American society of echocardiography carotid intima-media

thickness task force endorsed by the society for vascular medicine. J Am Soc Echocardiography. 2008;21(2):93-111.

396 Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M. Prediction of clinical cardiovascular events with carotid intima-media thickness: A systematic review and meta-analysis. Circulation. 2007;115(4):459-67.

397 Espeland M, O'Leary D, Terry J, Morgan T, Evans G, Mudra H. Carotid intimal-media thickness as a surrogate for cardiovascular disease events in trials of HMG-CoA reductase inhibitors. Current Controlled Trials in Cardiovascular Medicine. 2005;6(1):3.

398 Ebrahim S, Papacosta O, Whincup P, Wannamethee G, Walker M, Nicolaides AN, et al. Carotid plaque, intima media thickness, cardiovascular risk factors, and prevalent cardiovascular disease in men and women: The British regional heart study. Stroke. 1999 April 1, 1999;30(4):841-50.

399 Bilguvar K, Yasuno K, Niemela M, Ruigrok YM, von und zu Fraunberg M, van Duijn CM, et al. Susceptibility loci for intracranial aneurysm in European and Japanese populations. Nat Genet. 2008;40(12):1472-7.

400 Gardin JM, McClelland R, Kitzman D, Lima JAC, Bommer W, Klopfenstein HS, et al. M-mode echocardiographic predictors of six- to seven-year incidence of coronary heart disease, stroke, congestive heart failure, and mortality in an elderly cohort (the cardiovascular health study). Am J Cardiol. 2001;87(9):1051-7.

401 Levy D, Garrison RJ, Savage DD, Kannel WB, Castelli WP. Prognostic implications of echocardiographically determined left ventricular mass in the Framingham heart study. N Engl J Med. 1990;322(22):1561-6.

402 Vasan RS, Larson MG, Benjamin EJ, Evans JC, Levy D. Left ventricular dilatation and the risk of congestive heart failure in people without myocardial infarction. N Engl J Med. 1997;336(19):1350-5.

403 Wang TJ, Evans JC, Benjamin EJ, Levy D, LeRoy EC, Vasan RS. Natural history of asymptomatic left ventricular systolic dysfunction in the community. Circulation. 2003;108(8):977-82.

404 Bielen E, Fagard R, Amery A. The inheritance of left ventricular structure and function assessed by imaging and doppler echocardiography. American Heart Journal. 1991;121(6 I):1743-9.

405 Palatini P, Krause L, Amerena J, Nesbitt S, Majahalme S, Tikhonoff V, et al. Genetic contribution to the variance in left ventricular mass: The Tecumseh offspring study. Journal of Hypertension. 2001;19(7):1217-22.

406 Matson DD. Intracranial arterial aneurysms in childhood. Journal of Neurosurgery. 1965;23(6):578-83.

407 Patel AN, Richardson AE. Ruptured intracranial aneurysms in the first two decades of life. A study of 58 patients. Journal of Neurosurgery. 1971;35(5):571-6.

408 Mercado R, Lopez S, Cantu C, Sanchez A, Revuelta R, Gomez-Llata S, et al. Intracranial aneurysms associated with unsuspected aortic coarctation. Journal of Neurosurgery. 2002;97(5):1221-5.

409 Stehbens WE. Etiology of intracranial berry aneurysms. Journal of Neurosurgery. 1989;70(6):823-31.

410 Gardiner HM, Celermajer DS, Sorensen KE, Georgakopoulos D, Robinson J, Thomas O, et al. Arterial reactivity is significantly impaired in normotensive young adults after successful repair of aortic coarctation in childhood. Circulation. 1994;89(4):1745-50.

411 Isner JM, Donaldson RF, Fulton D, Bhan I, Payne DD, Cleveland RJ. Cystic medial necrosis in coarctation of the aorta: A potential factor contributing to adverse consequences observed after percutaneous balloon angioplasty of coarctation sites. Circulation. 1987;75(4):689-95.

412 Ostergaard JR, Voldby B. Intracranial arterial aneurysms in children and adolescents. Journal of Neurosurgery. 1983;58(6):832-7.

413 Schievink WI, Mokri B, Piepgras DG, Gittenberger-de Groot AC. Intracranial aneurysms and cervicocephalic arterial dissections associated with congenital heart disease. Neurosurgery. 1996;39(4):685-90.

414 Muzumdar DP, Sateesh M, Goel A. Multiple intracranial aneurysms in a child with congenital cyanotic heart disease. Pediatric Neurosurgery. 2006;42(6):368-73.

415 Uemura S, Takamoto K, Matsukado Y, Ishibashi K. Arteriovenous malformation associated with congenital heart disease, with a remark on accompanying cardiopulmonary dysfunction. Neurological Surgery. 1981;9(3):377-83.

416 de Wit MCY, Kros JM, Halley DJJ, de Coo IFM, Verdijk R, Jacobs BC, et al. Filamin A mutation, a common cause for periventricular heterotopia, aneurysms and cardiac defects. J Neurol Neurosurg Psychiatry. 2009;80(4):426-8.

417 Oza VS, Wang E, Berenstein A, Waner M, Lefton D, Wells J, et al. PHACES association: A neuroradiologic review of 17 patients. American Journal of Neuroradiology. 2008;29(4):807-13.

418 Kamath BM, Spinner NB, Emerick KM, Chudley AE, Booth C, Piccoli DA, et al. Vascular anomalies in Alagille syndrome: A significant cause of morbidity and mortality. Circulation. 2004;109(11):1354-8.

419 Kirby ML, Waldo KL. Role of neural crest in congenital heart disease. Circulation. 1990;82(2):332-40.

420 Imrie H, Freel M, Mayosi BM, Davies E, Fraser R, Ingram M, et al. Association between aldosterone production and variation in the 11beta-hydroxylase (CYB11B1) gene. J Clin Endocrinol Metab. 2006;91(12):5051-6.

421 Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. Am J Hum Genet. 2002;70(1):124-41.

422 Samani NJ, Raitakari OT, Sipila K, Tobin MD, Schunkert H, Juonala M, et al. Coronary artery disease-associated locus on chromosome 9p21 and early markers of atherosclerosis. Arterioscler Thromb Vasc Biol. 2008;28:1679-83.

423 Kastelein JJP, Akdim F, Stroes ESG, Zwinderman AH, Bots ML, Stalenhoef AFH, et al. Simvastatin with or without ezetimibe in familial hypercholesterolemia. N Engl J Med. 2008 April 3, 2008;358(14):1431-43.

424 Sharma K, Blaha MJ, Blumenthal RS, Musunuru K. Clinical and research applications of carotid intima-media thickness. Am J Cardiol. 2009;103(9):1316-20.

425 Zanchetti A. Carotid artery wall alterations as intermediate end points. Clinical and Experimental Hypertension. 1999;21(5-6):595-607.

426 Farzaneh-Far R, Na B, Schiller NB, Whooley MA. Lack of association of chromosome 9p21.3 genotype with cardiovascular structure and function in persons with stable coronary artery disease: The Heart and Soul study. Atherosclerosis. 2009;205(2):492-6.

427 Dandona S, Stewart AFR. Lack of association of chromosome 9p21.3 genotype with cardiovascular function in persons with stable coronary artery disease: The Heart and Soul study. Atherosclerosis. 2009;205(2):367.

428 Vasan RS, Glazer NL, Felix JF, Lieb W, Wild PS, Felix SB, et al. Genetic variants associated with cardiac structure and function: A meta-analysis and replication of genome-wide association data. JAMA. 2009;302(2):168-78.

429 Palmieri V, Dahlöf B, DeQuattro V, Sharpe N, Bella JN, de Simone G, et al. Reliability of echocardiographic assessment of left ventricular structure and function : The preserve study. JACC. 1999;34(5):1625-32.

430 Kobrynski LJ, Sullivan KE. Velocardiofacial syndrome, DiGeorge syndrome: The chromosome 22q11.2 deletion syndromes. Lancet. 2007;370(9596):1443-52.

431 Rinkel GJE. Natural history, epidemiology and screening of unruptured intracranial aneurysms. Journal of Neuroradiology. 2008;35(2):99-103.

432 Lawes CM, Vander Hoorn S, Rodgers A. International society of hypertension. Global burden of blood-pressure-related disease, 2001. Lancet. 2008;371:1513-8.

433 Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, et al. Evidence for a blood pressure gene on chromosome 17: Genome scan results for longitudinal blood pressure phenotypes in subjects from the Framingham heart study. Hypertension. 2000;36:477 - 83.

434 Coats AJS. Benefits of ambulatory blood pressure monitoring in the design of antihypertensive drug trials. Blood Pressure Monitoring. 1996;1(2):157-60.

435 Pickering TG, Shimbo D, Haas D. Ambulatory blood-pressure monitoring. N Engl J Med. 2006;354(22).

436 Chen W-M, Abecasis GR. Family-based association tests for genomewide association scans. Am J Hum Genet. 2007;81(5):913-26.

437 Ioannidis JP. Why most discovered true associations are inflated. Epidemiology (Cambridge, Mass). 2008;19(5):640-8.

438 Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009;324(5930):1035-44.

439 Mahr S, Burmester G-R, Hilke D, Gobel U, Grutzkau A, Haupl T, et al. Cis - and trans -acting gene regulation is associated with osteoarthritis. Am J Hum Genet. 2006;78(5):793-803.

440 Koch O, Kwiatkowski DP, Udalova IA. Context-specific functional effects of IFNGR1 promoter polymorphism. Hum Mol Genet. 2006;15(9):1475-81.

441 Ji W. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet. 2008;40:592-9.

442 Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. Complete Khoisan and Bantu genomes from southern Africa. Nature. 2010;463(7283):943-7.

443 Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, et al. Laser capture microdissection. Science. 1996;274(5289):998-1001.

444 Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131(5):861-72.

445 Schmid M, Sen M, Rosenbach MD, Carrera CJ, Friedman H, Carson DA. A methylthioadenosine phosphorylase (MTAP) fusion transcript identifies a new gene on chromosome 9p21 that is frequently deleted in cancer. Oncogene. 2000;19(50):5747-54

Appendix 1

Assay details

10 Appendix 1: Assay details

10.1 Assay primers

Table 10.1. Sequenom AEI and genotyping assays.

SNP	Assay	PCR primer 1	PCR primer 2	Amplicon size	Extension primer sequence
9p21 genotyping assays for expression mapping					
rs3088440	W1	ACGTTGGATGAGATCATCAGTCACCGAAGG	ACGTTGGATGTCTACGTTAAAAGGCAGGAC	140	CGAAGGTCCTACAGGGC
rs15735	W1	ACGTTGGATGAGTAGAGATTACCCTCCACC	ACGTTGGATGATGACTAGTAAACATGTGGG	136	GAGAAGAGAATGGTGGG
rs1333049	W1	ACGTTGGATGTTGCTTACCTCTGCGAGTGG	ACGTTGGATGGTGTATGACACTTCTTAGGC	140	CTGCGAGTGGCTGCTTTT
rs10965215	W1	ACGTTGGATGAGGAGCTGAGGAATCATCAC	ACGTTGGATGGAGACTTGTCTGACAAATCC	138	TTGGAATCCTTTGAAATGT
rs1333040	W1	ACGTTGGATGCATTCAAGAGAGACAGGAGG	ACGTTGGATGCATTGAGATTTGGGAGCCAC	138	gGGGTCAGAGGTAAGAATG
rs10757278	W1	ACGTTGGATGACTCTGTCTTGATTCTGCAT	ACGTTGGATGTGGAAAGTGACAAAGAGGAC	139	cccTGATTCTGCATCGCTGC
rs10757274	W1	ACGTTGGATGCATGGTGATGGGAGGTACTG	ACGTTGGATGGAATTCCCTACCCCTATCTC	139	gAAATCTAAGCTGAGTGTTG
rs1063192	W1	ACGTTGGATGCATGTTACTTCATTTCTGG	ACGTTGGATGACTGTGGGATCCACAATAAC	121	GGAATCTTTCCTAATGACAAC
rs496892	W1	ACGTTGGATGCCTGTTCCCCACAAAAATGC	ACGTTGGATGACTTTGTTGGAGGACCAAGG	140	ccctGCAGGCACCCCTGAATCC
rs7865618	W1	ACGTTGGATGTTTCACCAGGTGTAGTTAGG	ACGTTGGATGAATGCCCCACAAATACATTG	140	TGTTTGTTTAGCTTCTTAAACC
rs3217992	W1	ACGTTGGATGGCTGATGAAACAGCTAAACC	ACGTTGGATGTGGGTATCAATTACCACCTG	131	TGGCATTGATAAGTTACTATTT
rs10738605	W1	ACGTTGGATGCTAGGGTTCAAGCATCACTG	ACGTTGGATGCTCTGATGGTTTCTTTGGAG	140	ttccAGCATCACTGTTAGGTGTG
rs7023954	W1	ACGTTGGATGTTTCATTGCATGCAGCCATC	ACGTTGGATGTCCTTTAGTAGTAACATCC	134	GGAAGATAAAAAATGTTGATTGC
rs10116277	W1	ACGTTGGATGTGGTAATACCTAGCCTAAAC	ACGTTGGATGCTCTAGAATCGATTCTTGGG	140	ctcCAAGAGTTCACTCATAAATCT
rs2383206	W1	ACGTTGGATGCATGGCCCGATGATTTTCAG	ACGTTGGATGGGTTCAGGATTCAGGCCATC	135	cTTTCCTTAGAAATGTTATTGTAGT
rs7044859	W1	ACGTTGGATGTCCTCCTATTTCAATCTAC	ACGTTGGATGGGGCTCAATATTTTGAGTGG	138	aagAATGTGAACATGAAATCTTGAA
rs2383207	W2	ACGTTGGATGGAACTCAAAGATACTTAGCC	ACGTTGGATGTTGGGCAGCTCTTTTCATAC	150	CCTGTTCGGATCCCTTC
rs11515	W2	ACGTTGGATGCAGACATCCCCGATTGAAAG	ACGTTGGATGATGAAAACTACGAAAGCGGG	150	AGGCTCTGAGAAACCTC
rs564398	W2	ACGTTGGATGTGATGATTCCTCAGCTCCTC	ACGTTGGATGGATGGTTTCCCAAACAGCAC	149	GCCCCCATGACTTTCTTTG
rs1134871	W2	ACGTTGGATGGGGAGGGATACTGGGATAAT	ACGTTGGATGGAACTGGATCTTTCTAAATC	147	TCACATTTTTCTACAGTGAAT
rs11790231	W3	ACGTTGGATGCTCTCTCTTCCCTTTTCCTC	ACGTTGGATGGGGAAGAAAGAAAGTGGTAG	98	CCCTTTTCCTCCAACTC
rs17694493	W3	ACGTTGGATGTTTGGATGGTGAGAGCAGTC	ACGTTGGATGGAGTCTGTCAACTCATGTAA	99	AGAGCAGTCACACTGTT
rs615552	W3	ACGTTGGATGAGACTCTACCCAGTGATGAG	ACGTTGGATGCAAAAATGTGGCCACACTGC	96	AGTGATGAGGAACGGGA
rs16905599	W3	ACGTTGGATGGGCCCAGGAATGAAATTATC	ACGTTGGATGGTGTGAAAATAATGCGTGGG	82	TATCGCTCTGTATGACCTA
rs3217986	W3	ACGTTGGATGCGTGATCCAGGGATTTTAGC	ACGTTGGATGAACTCCACCAGATAGCAGAG	92	CATCTGTCGTCGCTTGCACA
rs10965227	W3	ACGTTGGATGGGTGCCATGTTCTTCCTTTC	ACGTTGGATGTCAGTTTCCCCATCTGGCAC	99	CATGTTCTTCCTTTCTACATT
rs10965228	W3	ACGTTGGATGGTCACTGGGCCTCTATATCA	ACGTTGGATGAGTGGCTCACTATTCACATC	100	GGAGGAAATAACATTTTCTGA
rs10125231	W3	ACGTTGGATGGCATCTCAGTTACCATGAGC	ACGTTGGATGAAAATGCAGGGAATTTAAG	99	aTCAGTTACCATGAGCCCAATA
rs3218009	W3	ACGTTGGATGCCTAAGGGATAAAGCTAGTG	ACGTTGGATGCTTACTGAATATCCCTTGTC	95	AAGGGATAAAGCTAGTGTCATAG
rs2184061	W3	ACGTTGGATGGCATAATCAAAATCAAGCCTG	ACGTTGGATGGCACTACTACTTACCTAATG	99	cagCAAAATCAAGCCTGTTAGATA
rs10811650	W3	ACGTTGGATGTCCTAGGGAATGTGTTACAG	ACGTTGGATGTGTTACCCACACCTTTGATG	87	ggaAGATACAGTGAAGAGTACAAA
rs495490	W3	ACGTTGGATGCAGCACTCCGAACCAAAATG	ACGTTGGATGGGTGAGGAACTGCATAAGTG	98	ACCAAAATGATTCTGTAAACAAAAA
rs7857345	W3	ACGTTGGATGAGTTAGGATGCTTTCAGGTA	ACGTTGGATGCCATCAGGAAATAATCCTTAG	100	cctGATGCTTTCAGGTACTTGTAATA
rs3731239	W3	ACGTTGGATGATCAGTGGAAATCTGTGGTG	ACGTTGGATGGCCGCTTCTGAATAATTTCG	94	GTGTATGTTGGAATAAATATCGAATA
rs7036656	W4	ACGTTGGATGTCCAAAGCCTTTGACACACG	ACGTTGGATGCTACCCATCTAAGGGTAGAG	95	TTTGACACACGGTAACA
rs1547705	W4	ACGTTGGATGGTGCTAGTGGCTCACTATTC	ACGTTGGATGGGCCTCTATATCATGATCAAC	98	GGCTCACTATTCACATCA
rs1333045	W4	ACGTTGGATGCATGTCATCTTAATGAATGC	ACGTTGGATGTTATTGTGAGGCCCAAGTGC	96	ATGCTTACTAGATGCCAG
rs1011970	W4	ACGTTGGATGAGAACTGATAGGGAGCCAGC	ACGTTGGATGGAAGATACAGGTGGAACTGG	99	AGAAATATCATGGTCCTCTG
rs2811712	W4	ACGTTGGATGCTAGGGATAAGGGTAATTGC	ACGTTGGATGGAACAGAGAGCTTACTATAT	98	GGAATGCCATCTGAATAAAAG
rs1801022	W5	ACGTTGGATGAACCTGCGCACCATGTTCTC	ACGTTGGATGTGCTCACCTCTGGTGCCAAA	100	TGTTCTCGCCGCCTCCA
rs3731249	W5	ACGTTGGATGCTCTCAGGGTACAAATTCTC	ACGTTGGATGGCAGTAACCATGCCCGCATA	99	ACCTGAGGGACCTTCCG
rs2811711	W5	ACGTTGGATGCTTAGACATGTGGTTCTCGC	ACGTTGGATGTCAGTTTCCCACGATTGAGG	95	cGTGGTTCTCGCAGTACC
rs3814960	W5	ACGTTGGATGTAATAGCACCTCCTCCGAGC	ACGTTGGATGCTCAAATCCTCTGGAGGGAC	99	CCTCCGAGCACTCGCTCAC
rs3218018	W5	ACGTTGGATGATTCCATCTTGGTTTCCCCC	ACGTTGGATGATTCCCCTGCATACTGCAAC	91	CGCTGCCCAATACCTGTTC
rs36228834	W5	ACGTTGGATGTCCCTATGACACCAAACACC	ACGTTGGATGGTTTCCTTCCTCCGCGATAC	88	CCGATTCAATTTGGCAGTT
rs2518723	W5	ACGTTGGATGACACACCCACCCACTCAGG	ACGTTGGATGACCCGGAGCCTGCGGTCT	96	gcGCCCGCGAGGTTTAGGAC
rs2069418	W5	ACGTTGGATGCCCTTGTGACCGAGAGAAAG	ACGTTGGATGCAGAAAGACACATCCAAGAG	94	AGTCATTCAAAATAACTCCGT
rs3218005	W5	ACGTTGGATGGCCAGGATTGAGAAACAATT	ACGTTGGATGTGGTGTTTGCCTCAGTAGTC	100	ggAGGGGTTTTAAACTTGTTG
rs2383208	W5	ACGTTGGATGTCTTGAAACTAGTAGATGC	ACGTTGGATGGCAACTAAAACTGTGACAGG	80	cCTAGTAGATGCTCAATTCATG
rs3218022	W5	ACGTTGGATGCCCATATTATTGAGGATAATC	ACGTTGGATGGCTGTGGAGGTATTTTTTAG	99	TTATTGAGGATAATCTGCTTTT
rs2891169	W5	ACGTTGGATGGGCCAAGGCAACAGTTTCTT	ACGTTGGATGACTCTTTGTTATGAATGTG	83	AGTTTCTTAATGAGAGAGTAGA
rs10757283	W5	ACGTTGGATGCAGGATGGGGAAGGTTTTTG	ACGTTGGATGCCCGACTGGAAAACCTTATA	99	ggTGGGGAAGGTTTTTGACTTTA
rs10811661	W5	ACGTTGGATGATAAGCGTTCTTGCCCTGTC	ACGTTGGATGAGATCAGGAGGGTAATAGAC	99	AGCTCACCTCCAGCTTTAGTTTTC
rs3731257	W5	ACGTTGGATGCTTGGTTCAGCACTCACTTG	ACGTTGGATGAGCTGCCAGAATTGACTTCG	99	agTGGTTGCTGTAAAACTTTCTTA
rs3218012	W5	ACGTTGGATGGTTTTTTTTAACAGGGGTA	ACGTTGGATGTTTTTCTCCCCTATCCCTGC	98	TTTAACAGGGGTAAATAATTTTCT
rs3218008	W5	ACGTTGGATGCAAAACCCCAAAGTCATAAG	ACGTTGGATGCTTTTGACATATATGATC	97	AGATAACATAAGCAAGAATAAAAAA
rs3218020	W5	ACGTTGGATGAACTGTCACATTATCCCGTC	ACGTTGGATGGATCCTCAGAAAACTGAGGG	80	cCACATTATCCCGTCATGCCTGAGGG
rs2069416	W5	ACGTTGGATGCCTGGCACACATAAGACAAA	ACGTTGGATGGATGTTTCATTATTACTGC	98	CACATAAGACAAAAAAAAATTATACT
9p21 genotyping assays for investigation of HWE departure (HTO cohort)					
rs10757278	W6	ACGTTGGATGCAAGTCAGGGTGTGGTCATT	ACGTTGGATGGAGAGAGAAGGAGAAAC	93	ATTCTGCATCGCTGC
rs2383206	W6	ACGTTGGATGCTTACTATCCTGGTTGCCCC	ACGTTGGATGGGTTCAGGATTCAGGCCATC	100	AGGCCATCTTGCAAA
rs615552	W6	ACGTTGGATGCAAAAATGTGGCCACACTGC	ACGTTGGATGAGACTCTACCCAGTGATGAG	96	TGATGAGGAACGGGA
rs1333049	W6	ACGTTGGATGCACTACCCTACTGTCATTCC	ACGTTGGATGATATCTTGCTTACCTCTGCG	100	GCGAGTGGCTGCTTTT
rs11790231	W6	ACGTTGGATGGGGAAGAAAGAAAGTGGTAG	ACGTTGGATGCTCTCTCTTCCCTTTTCCTC	98	CCCTTTTCCTCCAACTC
rs564398	W6	ACGTTGGATGTCAGTGGCACATACCACACC	ACGTTGGATGAGCTCCTCTCATCTGATCTC	96	CCCCATGACTTTCTTTG
rs10125231	W6	ACGTTGGATGAAAATGCAGGGAATTTAAG	ACGTTGGATGGCATCTCAGTTACCATGAGC	99	GTTACCATGAGCCCAATA
rs1333040	W6	ACGTTGGATGCTCATATCTGTACCTTCCTC	ACGTTGGATGCATTCAAGAGAGACAGGAGG	92	GGGTCAGAGGTAAGAATG
rs2383207	W6	ACGTTGGATGCAAGACATAACTAATAGCTG	ACGTTGGATGGGGACCATTTTTTACTCCTG	99	ggCCTGTTCGGATCCCTTC

rs10116277	W6	ACGTTGGATGCCAAATCAGACAAGAGTTCAC	ACGTTGGATGCGATTCTTGGGGAGGTTATT	99	cGTGATGGACAGAGTGTAG
rs10965215	W6	ACGTTGGATGGGATGTTTTGCAGGACTATT	ACGTTGGATGGGAATCATCACAGCATGGAC	100	aTTGGAATCCTTTGAAATGT
rs1063192	W6	ACGTTGGATGAACTTGTGGAATCTTTCCT	ACGTTGGATGACTGTGGGATCCACAATAAC	95	TTCTTTAGTTTCCCTTAATATCA
rs7857345	W6	ACGTTGGATGAGTTAGGATGCTTTCAGGTA	ACGTTGGATGCCATCAGGAAATAATCCTTAG	100	ccAGATGCTTTTAGTTGTGTTTT
rs10757274	W6	ACGTTGGATGCGTGGGTCAAATCTAAGCTG	ACGTTGGATGGAATTCCCTACCCCTATCTC	97	CTATCTAGTGAATTTCAATTATGTC
rs10811650	W6	ACGTTGGATGTGTTACCCACACCTTTGATG	ACGTTGGATGTCCTAGGGAATGTGTTACAG	87	ggaaAGATACAGTGAAGAGTACAAA
rs2184061	W6	ACGTTGGATGGCATAATCAAAATCAAGCCTG	ACGTTGGATGGCACTACTACTTACCTAATG	99	tcAGTTGCTATTAATGGATGAATTTT
rs1333045	W6	ACGTTGGATGCATGTCATCTTAATGAATGC	ACGTTGGATGTTATTGTGAGGCCCAAGTGC	96	cctTGCAATAATATATAGTACACTGTG
Congenital heart disease study genotyping assays					
rs564398	W7	ACGTTGGATGAGCTCCTCTCATCTGATCTC	ACGTTGGATGTCAGTGGCACATACCACACC	96	ACCACACCCTAACTACC
rs3218018	W7	ACGTTGGATGATTCCCCTGCATACTGCAAC	ACGTTGGATGATTCCATCTTGGTTTCCCCC	91	CTGCCCAATACCTGTTC
rs1333049	W7	ACGTTGGATGCACTACCCTACTGTCATTCC	ACGTTGGATGATATCTTGCTTACCTCTGCG	100	TGCGAGTGGCTGCTTTT
rs2811708	W7	ACGTTGGATGGAAAAGTGGATAGTTTTGAC	ACGTTGGATGCTAAGCCAACATCATTTCCC	98	TTCTACATGTTCTTCTCCC
rs1547705	W7	ACGTTGGATGGGCCTCTATATCATGATCAAC	ACGTTGGATGGTGCTAGTGGCTCACTATTC	98	GTGGCTCACTATTCACATCA
STK39 study genotyping assays (HTO cohort)					
rs3754777	W8	ACGTTGGATGGCCTGAACAAAAATGAGGAC	ACGTTGGATGATCTCGCCTGTTTCACCCAC	96	AGGCGTCTCTGGGTCTTTTACT
rs35929607	W8	ACGTTGGATGCACACTCATGGAATTAAAGG	ACGTTGGATGTCAGAGGGCTCACATTTTGG	90	ATGGAATTAAAGGATTATTAGGATACC
rs6749447	W8	ACGTTGGATGTGGAGTCTGCTAGTACTAGA	ACGTTGGATGCAGTTAGGTCACCTCCTTTC	100	GAGTCTGCTAGTACTAGATTAGGA
rs4977950	W8	ACGTTGGATGCTGAGAACAGTCTTCAACTTG	ACGTTGGATGCACTAGACTGGTTGATTGAA	100	CCTGAAGGTTTTTTTTTATATCACTA
STK39 study genotyping assays (SA cohort)					
rs1061471	W9	ACGTTGGATGTTTAGAAGTTACAAATACTC	ACGTTGGATGGCTTCTTGCAGTTAATCTCG	93	GAAGTTACAAATACTCCAAAGA
rs1802105	W9	ACGTTGGATGGAAGGCTAATGGCACTTACC	ACGTTGGATGAGAGTACCTGATTGAGAAGC	100	CTTTGGGCTATGTCTGGTG
rs56031549	W9	ACGTTGGATGGATGAGAAGAGCGAAGAAGG	ACGTTGGATGACACAGATTAGCTCATCGTC	100	GCGAAGAAGGGAAAGCA
rs56048258	W9	ACGTTGGATGATTCAAGCCATGAGTCAGTG	ACGTTGGATGGCCAAAGTTCATCTTTGACC	99	ATGAGTCAGTGCAGCCA
rs56330212	W9	ACGTTGGATGGGCACAAGAAGAAGCTTCTC	ACGTTGGATGAACACTGTTTGCTGTTTTC	96	TCTGTAGTCTTCATTAGCATT
rs56697518	W9	ACGTTGGATGACACAGATTAGCTCATCGTC	ACGTTGGATGGATGAGAAGAGCGAAGAAGG	100	CGTCAACGATITCATTTAC
AEI cDNA assays					
rs11515/rs3088440	CDKN2A exons2-3 cDNA	GGAAGGTCCCTCAGACATCC	ATTTTCTAAATGAAAACTACGAAAGC	150	AGGCTCTGAGAAACCTC and CGAAGGTCCTACAGGGC
rs1063192	CDKN2B1063192 cDNA	ACGTTGGATGCATGTTACTTCATTTCTGG	ACGTTGGATGACTGTGGGATCCACAATAAC	121	GGAATCTTTCCTAATGACAAC
rs3217992	CDKN2B3217992 cDNA	ACGTTGGATGGCTGATGAAACAGCTAAACC	ACGTTGGATGTGGGTATCAATTACCACCTG	131	TGGCATTGATAAGTTACTATTT
rs10965215/rs564398	ANRIL exons1-2 cDNA	CGCCGGACTAGGACTATTTG	GCACATACCACACCCTAACTACC	150	TTGGAATCCTTTGAAATGT and GCCCCCATGACTTTCTTTG
rs15735	MTAP15735 cDNA	CCTGAAGAATATGGCCCAGT	AGGGGGAGAAGAGAATGGTG	326	GAGAAGAGAATGGTGGG
rs7023954	MTAP7023954 cDNA	GCCGTGAAGATTGGAATAATTG	GACCTTTGAAGGCATGATGG	192	GGAAGATAAAAAATGTTGATTGC
rs1799977	MLH1 cDNA	ACGTTGGATGAGCAAGGAGAGACAGTAGC	ACGTTGGATGGTGATGGAATGATAAACCAAG	229	TCTTTGGAAATGCTGTTAG
rs1061471	STK39 cDNA	ACGTTGGATGTTTAGAAGTTACAAATACTC	ACGTTGGATGGCTTCTTGCAGTTAATCTCG	93	GAAGTTACAAATACTCCAAAGA
AEI gDNA assays					
rs11515/rs3088440	CDKN2A exons2-3 gDNA	ACGTTGGATGAGACATCCCCGATTGAAAG	ACGTTGGATGATTTTCTAAATGAAAACTACGAAAGC	158	AGGCTCTGAGAAACCTC and CGAAGGTCCTACAGGGC
rs1063192	CDKN2B1063192 gDNA	ACGTTGGATGCATGTTACTTCATTTCTGG	ACGTTGGATGACTGTGGGATCCACAATAAC	121	GGAATCTTTCCTAATGACAAC
rs3217992	CDKN2B3217992 gDNA	ACGTTGGATGGCTGATGAAACAGCTAAACC	ACGTTGGATGTGGGTATCAATTACCACCTG	131	TGGCATTGATAAGTTACTATTT
rs10965215/rs564398	ANRIL exons1-2 gDNA	ACGTTGGATGTCTGAATCTAATCACATAGAGACTTG	ACGTTGGATGGCACATACCACACCCTAACTACC	219	TTGGAATCCTTTGAAATGT and GCCCCCATGACTTTCTTTG
rs15735	MTAP15735 gDNA	TGGCCCAGTTTTCTGTTTTATTAC	AGGGGGAGAAGAGAATGGTG	315	GAGAAGAGAATGGTGGG
rs7023954	MTAP7023954 gDNA	ACGTTGGATGTTTCATTGCATGCAGCCATC	ACGTTGGATGTCCTTTAGTAGTAACATCC	134	GGAAGATAAAAAATGTTGATTGC
rs1799977	MLH1 gDNA	acgttggatgGTTTCAGTCTCAGCCATGAGAC	acgttggatgAAGCCTGTGTATTTGACTAAAG	174	TCTTTGGAAATGCTGTTAG
rs1061471	STK39 gDNA	ACGTTGGATGTTTAGAAGTTACAAATACTC	ACGTTGGATGGCTTCTTGCAGTTAATCTCG	93	GAAGTTACAAATACTCCAAAGA

Table 10.2. Taqman custom gene expression assays.

Assay	PCR primer 1	PCR primer 2	Probe sequence
CDKN2A_exon1alpha	GGCGGCTGCGGAGA	CGCCGGCTCCATGCT	CTGCCTGCTCTCCC
CDKN2B	GTGGGATGTCCTACTATCACTGAAC	GTAGTGAAGTATCAGTTGTTCCAATGATATAATGA	CAACCCTGGAAATTAA
ANRIL_exon13	CCTACGAAGCTGGGTGATAAAAG	GGATCCTGGTATAGAATGAAGCTATCTG	ACATTGGACAAAAAC
ANRIL_exon20	TCCCCATGTTACATATGAAGAGACAGA	GGTTAGACTGCTTAAGTTCAAATCCCA	CACCGCTGGTAATTG

Table 10.3. Primers for microsatellite and transcript-specific AEI assays.

Primer	Sequence
Microsatellite study	
Microsatellite_10583774_F	AGCAATAATTCTCCCCAAGG*
Microsatellite_10583774_R	GAAACTGAGGCGAACAGAGC
CDKN2A/ARF transcript-specific 1	
CDKN2A_exon1alphaA_F	CTCCGAGCACTCGCTCAC
ARF_exon1betaA_F	GGTTTTCGTGGTTCACATCC
CDKN2A-ARF_exon3A_R	TTTACGGTAGTGGGGGAAGG
CDKN2A/ARF transcript-specific 2	
CDKN2A_exon1alphaB_F	CAACGCACCGAATAGTTACG
CDKN2A_exon3B_R	TTTTCTAAATGAAAACTACGAAAGC
ARF_exon1betaB_F	GCCGCTTCCTAGAAGACCAG
ARF_exon3B_R	CCTGTAGGACCTTCGGTGAC
CDKN2A transcript-specific 3	
CDKN2A_exon1alphaA_F	CTCCGAGCACTCGCTCAC
CDKN2A_exon1alphaA_R	CTGTCCCTCAAATCCTCTGG
ANRIL transcript-specific	
ANRIL_exon1_F	CTCTGACGCGACATCTGG
ANRIL_exon2_F	GTCCATGCTGTGATGATTCC
ANRIL_exon6_F	GGTTCAAGCATCACTGTTAGG
ANRIL_exon2_R	TCAGTGGCACATACCACACC
ANRIL_exon6_R	GCAGTACTGACTCGGGAAAGG
ANRIL_exon13_R	TTTATCACCCAGCTTCGTAGG
ANRIL_exon19_R	CCACAATGTTCAACTTGCTGTC
ANRIL_exon20_R	CACAGCTTTGATCTATATGCTTGG
MTAP_exon4_F	TGCCTTCAAAGGTCAACTACC

Table 10.4. MLPA probes.

Probe	Total robe length (bases)	Left primer oligo sequence	Right primer oligo sequence
9p21_01	92	GGGTTCCCTAAGGGTTGGATGGGTTACTTGTGTCTGTGCTGTTG	TTTCCTTAAGGAGGCTGGTGCTGCTTCTAGATTGGATCTTGCTGGCAC
9p21_02	100	GGGTTCCCTAAGGGTTGGATTTCCACCTGCTAAGCTAGTGTCCCTTCA	AACCCTTGTTTCCTGTAGCTCATGAGTCATCTAGATTGGATCTTGCTGGCAC
9p21_03	104	GGGTTCCCTAAGGGTTGGATCAGTGCTGGGTGACTCTCTCCTGGCATGTT	CTGGCAAGACTCTACCTGGTTTGCGTAATCTTCTAGATTGGATCTTGCTGGCAC
9p21_04	108	GGGTTCCCTAAGGGTTGGATGAATTGCATGGAAGAGGCTGTTCATGGGGTTA	GTCTGGCAGATAAGAGGTCTCCACTGGATTTTTTCTAGATTGGATCTTGCTGGCAC
9p21_05	112	GGGTTCCCTAAGGGTTGGACAATGAACGCCTTCACTGATATCCAAAGCATGAAG	GACACACCAGGGAAAAACATAGACCTAACACAGGATCTAGATTGGATCTTGCTGGCAC
9p21_06	116	GGGTTCCCTAAGGGTTGGAGTATACACACTGGTATAGTCAGATGGCATGAGTACCA	AGGTGGAAGTGTGGTGGTCCTAAAGTGGCATTAAGGATCTAGATTGGATCTTGCTGGCAC
9p21_07	120	GGGTTCCCTAAGGGTTGGAGTAGAGTGGGCTGGGGAGCGACTTTCTTGGGACTATTAT	GCTTTTCTTGCACTTTACTAGGTTCCTGTTCTTCCCCACTCTAGATTGGATCTTGCTGGCAC
9p21_08	124	GGGTTCCCTAAGGGTTGGAGACAGATATCTAAGATAACACGAACACCCAGACGACAACTC	AGGCAACCCCAGCACAGTGTTTTGGTATGAGAACAGGAGAATCTAGATTGGATCTTGCTGGCAC
9p21_09	128	GGGTTCCCTAAGGGTTGGACTAACTAGTCTGTTTCCCTACCCAGGTGGAGAACTTCAGTAGA	GGAAGTGGCAGGAATTTGGGAATGAGGAGCACAGTGATTAAACTCTAGATTGGATCTTGCTGGCAC
9p21_10	132	GGGTTCCCTAAGGGTTGGACATGCAGCCAGCCACTAAGACTACTTCCCACAGGACTGGAAATCT	CTACCACCGTACCTCTCCTACATGCTGGGTATAAGAAGTTAAAGGTCTAGATTGGATCTTGCTGGCAC
9p21_11	136	GGGTTCCCTAAGGGTTGGACTCCCACTGTCACAGCCACTTCTGTCATGACCTAGTTGTTCTCAGAA	AGTTCCCTCTTATACCTGGTCAAACTTCTCTTACAATCACCTGTGACTCTAGATTGGATCTTGCTGGCAC
9p21_12	140	GGGTTCCCTAAGGGTTGGACAGCTAAGTTAGAATGTGATACCATTGTTTACACAAGTGTGGCCTGCCT	TCAAGATAGGGTGAGGTGTTTTATGACCACAGGCTTTATGAGTTATAGCTCTAGATTGGATCTTGCTGGCAC
8p23_DEFA3	96	GGGTTCCCTAAGGGTTGGATATTGCAGAATACCAGCGTGCATTGC	AGGAGAACGTCGCTATGGAACCTGCATCTCTAGATTGGATCTTGCTGGCAC
14q11_OR4K2	104	GGGTTCCCTAAGGGTTGGATTCTGTCATTAGTCCCCAGGTGTGTGTTGCT	CTCGTGGTGGCTTCCTGGATTATGGGAGTTATCTAGATTGGATCTTGCTGGCAC
22q11_CLDN5	112	GGGTTCCCTAAGGGTTGGAGCTTTTCTGCAGACTCAGGCTCTTCTCACACACAA	TTAATGAGATCCTGCCATTCTTCCCCTGGGAAAGCTCTAGATTGGATCTTGCTGGCAC
17p11_PMP22	124	GGGTTCCCTAAGGGTTGGAGCTGACTGTAGGTTTGGGAGTTGTCTGTCTCCTTGCTCTAG	ATTACAGCTCTGTGTGTTGTGTGGGGTCTCCATGGTTTGCCTCTAGATTGGATCTTGCTGGCAC
6p25_DUSP22	132	GGGTTCCCTAAGGGTTGGACACTGACCAGCCTCTCCTTACAACTGAAGACTTTTGCAGGACACT	CTTCCTAGATGTTGCCTTCATGTTGAGTCCCGGTGTCTGAGCATTTCTAGATTGGATCTTGCTGGCAC

10.2 MLPA protocol

Hybridisation mix preparation

Hybridisation mix prepared containing 0.5µL of synthetic probemix, 1µL of P200 probmix and 1.5µL MLPA buffer per sample. Mixed thoroughly by pipetting.

DNA denaturation and hybridisation of MLPA probes

5µL of DNA (50-200ng) added wells of 96-well plate. Heated on thermocycler at 98^{o}C for 5 min, followed by 25^{o}C hold. Whilst still on block, film lid removed and 3µL hybridisation mix added to each well (using a multichannel pipette from a loading plate). Mixed thoroughly by pipetting. Sealed with new lid and heated at 95^{o}C for 1 min, followed by 60^{o}C for 16 (12-24) hours.

Ligase-65 mix preparation (within 1 hour of use, stored on ice)

Ligase-65 mix prepared containing 3µL Ligase-65 buffer A, 3µL Ligase-65 buffer B, 25µL water, 1µL Ligase-65 per sample. Mixed thoroughly by pipetting.

Ligation reaction

Whilst plate still on block, temperature reduced to 54^{o}C. 32µL Ligase-65 mix added to each well using multichannel pipette and mixed thoroughly by pipetting. Then heated at 54^{o}C for 15 min, followed by 98^{o}C for 5 min, followed by 4^{o}C hold.

Polymerase mix prepared

Polymerase mix prepared containing 1µL SALSA PCR primers, 1µL SALSA enzyme dilution buffer, 15.75µL water, 0.25µL SALSA polymerase, 2µL SALSA PCR buffer per sample.

PCR reaction

20µL of polymerase mix added to wells of a new 96-well PCR plate. 5µL of MLPA ligation reaction product added to each well. Thermocycler heated to 95^{o}C and plate put onto block once reached 95^{o}C. PCR performed using 35 cycles of (95^{o}C for 30 sec, followed by 60^{o}C for 30 sec, followed by 72^{o}C for 1 min), then 72^{o}C for 20 min, then 4^{o}C hold. Product stored in dark at 4^{o}C until analysis.

Analysis

A master mix was then prepared containing 8.5µL of Hi-Di formamide (Applied Biosystems, USA) and 0.3µL of Genescan-500 ROX size standard (Applied Biosystems, USA) per sample. 8.8µL of the master mix was added to each well of a new 96-well semi-skirted, straight-edge PCR plate (StarLab, Germany), and 1µL of the PCR product was transferred to each well of this plate. The plate was covered in foil to prevent photo-degradation, before analysis by capillary electrophoresis on a 3130xl Genetic Analyzer (Applied Biosystems, USA). Data were analysed using GeneMarker v1.8 software (SoftGenetics, USA).

Appendix 2

Published manuscripts

11 Appendix 2: Published manuscripts