UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

School of Medicine

Molecular genetic epidemiological studies of smoking behaviour and growth promoting genes

by

Shuwen Huang

Thesis for the degree of Doctor of Philosophy

April 2007

UNIVERSITY OF SOUTHAMPTON

Abstract

Molecular genetic epidemiological studies of smoking behaviour and growth promoting genes

By Shuwen Huang

This study aimed to advance our understanding of the genetics of two independent complex traits: (i) adolescent smoking behaviour and (ii) growth pathway genes in relation to birthweight and metabolic risk. The strategy employed included:

- Establishment of a DNA bank from a well characterised cohort and development of quality control methods
- Identification of relevant candidate genes and genetic variants
- Development of accurate, cost-efficient, and high throughput genotyping methods; typing variants in relevant populations
- Detailed genotype-phenotype analysis

Smoking is a major cause of death and often initiates in adolescence. Haploinsufficiency for *CYP2A6* in adults is associated with lower cigarette consumption, lower cotinine level and higher rates of giving up. Other genes have also been implicated in smoking behaviour. I genotyped 1,520 subjects from the UK-wide Ten Towns Heart Health Study (TTHHS) for *CYP2A6* alleles **1A, *1B, *2, *4, *5, *9* and **12* plus variants of *DBH*, *MAOA*, *DRD4, TH* and *5HT2A*. I found that individuals predicted to be half normal metabolisers due to haploinsufficiency of *CYP2A6* (carriers of one fully inactive allele, i.e. **2, *4*, or **5*) were more likely to be a current smoker at age 18 years. The current smoker versus ex-smoker odds ratio was 1.95 (95% C.I. 1.03-3.68). Salivary cotinine levels in current smokers were lower in the 'normal' group compared with 'slow' group (p=0.0092), however the significance diminished on correction for multiple testing. I concluded that *CYP2A6* haploinsufficiency may be associated with increased likelihood of continuing smoking in teenagers but effects of the other genes studied were not detectable.

The association between low birthweight and adult disorders has been widely investigated, particularly with regard to environmental factors but there has been less

focus on genetics. There is however evidence that *GH-CSH* influences adult growth hormone (GH) levels, weight at one year of age, metabolic syndrome traits and bone loss. To explore this in more detail I compared microsatellite genotypes at the major growth controlling genes *CSH1.01* and *IGF1* to birthweight and other metabolic traits in the TTHHS cohort. The *CSH1.01* TT genotype in the TTHHS boys showed a higher birthweight compared with those D1D1 or D2D2 genotype (p=0.027; corrected for multiple testing). There was no evidence for association of *IGF1*$(CA)_n$ with birthweight in the TTHHS, but boys of 192/192bp genotype did show a leg length 0.65cm greater than those not homozygous for allele 192bp. My study thus provides evidence for effects of *GH-CSH* on birthweight in males. The complexities of my data relative to previous data are discussed.

Results from previous studies have led to the hypothesis that some *ACE D* allele phenotypic associations may represent proxy marking of causal factors located in the *GH-CSH* gene cluster. I tested this hypothesis for cardiac muscle growth response to exercise and ACE level in a British army heart study cohort. Combined analyses of *ACE I/D* and *GH-CSH Bgl IIB* in relation to exercise induced left ventricular mass (LVM) change and serum ACE activity were investigated, A significant association between *GH-CSH Bgl IIB* and ACE activity was detected but only through its linkage disequilibrium with *ACE I/D*. No association between *GH-CSH Bgl IIB* and LVM change was found. Cardiac muscle growth in response to exercise and ACE levels both appear to be attributable to the *ACE* polymorphism and not to nearby *GH-CSH* polymorphism, although this does not exclude other *ACE* attributed phenotypes being due to *GH-CSH* variation.

Table of contents

Abstract ii

Table of contents iv

List of Tables x

List of Figures xii

Acknowledgement xiii

Abbreviations xiv

List of publications xvi

Chapter 1 Introduction 1

1.1 Introduction to genetic analysis 1

1.1.1 Genetic variations 1

1.1.1.1 Single nucleotide variations 2

1.1.1.2 Variable number of tandem repeats (VNTRs) 4

1.1.1.3 Transposable elements 5

1.1.1.4 Insertion and deletion (Indels) 5

1.1.1.5 Copy number variants 6

1.1.1.6 Other structural variations 8

1.1.2 Genetic diseases 9

1.1.2.1 Single-gene (also called Mendelian or monogenic) disease 9

1.1.2.2 Multifactorial (also called complex or polygenic) disease 10

1.1.2.3 Chromosomal disease 11

1.1.2.4 Mitochondrial disease 11

1.1.3 Genetic analysis of complex diseases 12

1.1.3.1 Linkage analysis 12

1.1.3.2 Genetic association analysis 15

1.1.3.3 Linkage disequilibrium and genetic association study 18

1.2 Genetic analysis of smoking 25

1.2.1 The importance of genetic studies of smoking 25

1.2.1.1 The 'popularity' of cigarette smoking 25

1.2.1.2 The consequences of smoking 26

1.2.1.3 Smoking is addictive 27

1.2.1.4 Smoking is a multifactorial complex behavioural trait 28

1.2.2 Approaches to genetic studies of smoking behaviour 28

1.2.2.1 Twin studies 29

1.2.2.2 Genome-wide linkage studies 31

1.2.2.3. Genome-wide association study 35

1.2.2.4 Candidate gene studies 35

1.2.3 The nicotine metabolism pathway 36

1.2.3.1 Nicotine and its metabolism 36

1.2.3.2 Cytochrome P450 2A6 (CYP2A6) 39

1.2.3.3 *CYP2A6* polymorphism, nicotine metabolism and smoking behaviour 47

1.2.3.4 Other genes influencing metabolism of nicotine (*CYP2D6* and *CYP2B6*) 48

1.2.4 The dopamine reward pathway 50

1.2.4.1 Dopamine rewarding system 50

1.2.4.2 Dopamine synthesis 53

1.2.4.3 Dopamine metabolism 53

1.2.4.4. Dopamine receptors 56

1.2.4.5 Dopamine transporter 59

1.2.5 Serotonin system 60

1.2.6 Cholinergic nicotinic receptor genes 62

1.3 Genetic epidemiological study of growth pathway genes 64

1.3.1 Fetal programming hypothesis and its development 64

1.3.1.1 Thrifty phenotype hypothesis 64

1.3.1.2 Intergenerational effects of fetal programming 66

1.3.1.3 Predictive adaptive response and developmental origins of health and disease 67

1.3.2 Genetic determinants of early growth and adult disease risk 69

1.3.3 Linkage disequilibrium between *GH-CSH* gene cluster and *ACE* gene 74

1.4 Hypotheses, aim and strategies **76**

1.4.1 Hypothesis, aim and strategies of smoking study 76

1.4.2 Hypothesis, aim and strategies of growth pathway genes study 77

1.4.2.*1 CSH1.01* genotype, birthweight and metabolic traits 77

1.4.2.2 *IGF1* gene, birthweight and metabolic traits 78

1.4.2.3 *ACE* I/D, *GH-CSH* markers and ACE activity and heart growth 78

Chapter 2 Materials and methods **80**

2.1 Study cohorts and phenotypes **80**

2.1.1 The Ten Towns Heart Health Study (TTHHS) 80

2.1.2 The Lichfield Army Recruit Growth in Exercise Heart Study 82

2.2 Chemicals and reagents **83**

2.2.1 Chemicals and reagents for establishment of the DNA bank 83

2.2.2 Chemicals and reagents for genotyping 83

2.2.2.1 Chemicals and reagents for PCRs 83

2.2.2.2 Chemicals and reagents for making gels 84

2.2.2.3 Chemicals and reagents for gel staining 84

2.2.3 Other commonly used materials and chemicals 85

2.3 Equipment **86**

2.3.1 Equipment for DNA extraction 86

2.3.2 Equipment for genotyping and aliquoting 86

2.4 Methods **87**

2.4.1 DNA bank establishment and quality control 87

2.4.1.1 Extraction of genomic DNA 87

2.4.1.2 Quantitation of genomic DNA concentration 88

2.4.1.3 Equalization of genomic DNA 88

2.4.1.4 Setting up of a small component of the genomic DNA stocks 89

2.4.1.5 Quality control of DNA bank by gender test 89

2.4.2 Genotyping smoking markers 90

2.4.2.1 Genotyping of *CYP2A6* 91

2.4.2.2 Genotyping of *HUMTH01*-VNTR 96

2.4.2.3 Genotyping of *DBH* G1368A 98

2.4.2.4 Genotyping of *MAOA* T1460C 99

2.4.2.5 Genotyping of *MAOA* VNTR 99

2.4.2.6 Genotyping of *DRD4* VNTR 101

2.4.2.7 Genotyping of *5HT2A* T102C 102

2.4.3 Genotyping growth pathway genes 103

2.4.3.1 Microsatellite marker *CSH1.01* and *IGF1*(CA)n 103

2.4.3.2 Angiotensin converting enzyme (*ACE*) insertion/deletion 105

2.4.3.3 *GH-CSH Bgl llB* 106

2.4.4 Image analysis 108

2.4.4.1 384 ARMS MADGE 108

2.4.4.2 Microsatellite image analysis 108

2.4.5 Dilution and aliquoting 109

2.5 Statistical analysis **110**

2.5.1 Statistical analysis of smoking markers 110

2.5.2 Statistical analysis of growth pathway genes 111

2.5.3 Power calculations 113

Chapter 3 Results **117**

3.1 DNA bank **117**

3.1.1 DNA concentration and yields 117

3.1.2 Sex test applied to the DNA banks 118

3.1.3 Sex comparison between genotype data and phenotype data 118

3.1.4 Sex confirmation combining sex test with *MAO-A* genotypes 121

3.2 Genotype and allele frequencies of smoking markers **122**

3.2.1 Genotype and allele frequencies of *CYP2A6* 122

3.2.1.1 Genotype and allele frequencies by gender 122

3.2.1.2 Genotype and allele frequencies by ethnicity 125

3.2.2 Genotype and allele frequencies of SNPs for *DBH, MAOA,* and *5HT2A* 127

3.2.2.1 Genotype and allele frequencies of *DBH* G1368A 127

3.2.2.2 Genotype and allele frequencies of *MAOA* T1460C 128

3.2.2.3 Genotype and allele frequencies of *5HT2A* T102C 129

3.2.3 Genotype and allele frequencies of VNTRs 130

3.2.3.1 Genotype and allele frequencies of *MAOA* VNTR 131

3.2.3.2 Genotype and allele frequencies of *DRD4* VNTR 133

3.2.3.3 Genotype and allele frequencies of *TH01* VNTR 134

3.2.4 Hardy-Weinberg Equilibrium test 136

3.2.5 Linkage disequilibrium of *MAOA* T1460C and *MAOA* VNTR 138

3.3 Genotype-phenotype analyses in relation to smoking behaviour **139**

3.3.1 Smoking status analyses by genotypes at ages 13–15 and 18 years 140

3.3.2 Smoking traits by genotypes at ages 13–15 and 18 years (current smokers only) 143

3.4 Results for growth pathway genes **146**

3.4.1 *CSH1.01* 146

3.4.1.1 *CSH1.01* allele frequency and distribution 146

3.4.1.2 Phenotypic analysis in relation to *CSH1.01* in the TTHHS 150

3.4.2 *IGF1*(CA)n 152

3.4.2.1 *IGF1*(CA)n allele frequency and distribution 152

3.4.2.2 Phenotypic analysis in relation to *IGF1*(CA)n in the TTHHS 154

3.4.3 *ACE* I/D and *GH-CSH Bgl IIB* 156

3.4.3.1 Allele frequencies and linkage disequilibrium of *ACE* I/D and *GH-CSH BglIIB* 156

3.4.3.2 Genotype-phenotype associations 157

Chapter 4 Discussion **160**

4.1 DNA bank -- the foundation of genetic epidemiology study **160**

4.2 Assay development and validation **162**

4.2.1 Assay development 162

4.2.2 Assay validations 163

4.3 One of the largest genetic epidemiology studies in adolescent smoking **167**

4.4 Phenotype-genotype association in growth pathway genes **172**

4.4.1 *GH-CSH* gene polymorphism, birthweight, and metabolic traits 172

4.4.2 Association of *IGF-I* CA microsatellite with leg length in adolescence 175

4.4.3 Serum ACE activity and LV growth: combined analysis of the linked growth hormone (*GH-CSH*) and *ACE* gene regions 176

Chapter 5 General discussion and perspectives **179**

5.1 How to interpret significant association? **179**

5.2 Common errors and guidelines for association studies **181**

5.3 High throughput genotyping and whole genome scan technologies **182**

5.3.1 array CGH 183

5.3.2 SNP array 184

5.3.3 Investigating the relationship between genetic variations and complex diseases using current technologies 187

Appendix Methods for making up stock reagents **195**

References **198**

List of Tables

Table 1.1 Twin studies of smoking behaviour in relation to genetic effects 30

Table 1.2 Overview of positive results of linkage studies to smoking behaviour 32

Table 1.3 Current *CYP2A6* allele nomenclature 46

Table 2.1 Primers used for genotyping smoking candidate genes 92

Table 2.2 Protocols for generating *CYP2A6* LPCR1, LPCR2, and LPCR3 95

Table 2.3 Protocols for second PCR reactions of *CYP2A6* 95

Table 2.4 Protocols for PCR reactions of dopaminergic genes 103

Table 2.5 *CYP2A6* allele grouping, coding and predicted characteristics 111

Table 3.1 Gender disagreement between phenotype and quality control genotype data 119

Table 3.2 Genotype of *MAO* T1460C and *MAO*-VNTR for gender disagreement 121

Table 3.3 Genotype and allele frequencies for *CYP2A6*1A, *1B, *2, *4* and **5* by gender 124

Table 3.4 Genotype and allele frequencies for *CYP2A6*12* and **9* by gender 125

Table 3.5 Genotype and allele frequencies for *CYP2A6*1A, *1B, *2, *4* and **5* by ethnicity 126

Table 3.6 Genotype and allele frequencies for *CYP2A6*12* and **9* by ethnicity 127

Table 3.7 Genotype and allele frequencies for *DBH* G1368A by gender 128

Table 3.8 Genotype and allele frequencies for *DBH* G1368A by ethnicity 128

Table 3.9 Genotype and allele frequencies for *MAOA* T1460C by gender and by ethnicity 129

Table 3.10 Genotype and allele frequencies for *5HT2A* T102C by gender 130

Table 3.11 Genotype and allele frequencies for *5HT2A* T102C by ethnicity 130

Table 3.12 Genotype and allele frequencies for *MAOA* VNTR by gender and ethnicity 132

Table 3.13 Comparison for *MAOA* VNTR allele frequencies (combined) between different ethnic groups 132

Table 3.14 Comparison for *DRD4* VNTR allele frequencies (combined) between different ethnic groups 133

Table 3.15 Genotype and allele frequencies for *DRD4* VNTR by gender and ethnicity 134

Table 3.16 Genotype and allele frequencies for *TH01* VNTR by gender and ethnicity 135

Table 3.17 Comparison for *TH01* VNTR allele frequencies (combined) between different ethnic groups 136
Table 3.18 Genotype counts, minor allele frequency (q) and HWE 137
Table 3.19 Linkage disequilibrium between *MAOA* T1460C and VNTR 138
Table 3.20 Linkage disequilibrium between *MAOA* T1460C and VNTR (females only) 139
Table 3.21 Smoking behaviour by genotypes at age 13-15 141
Table 3.22 Smoking behaviour by genotypes at age 18 142
Table 3.23 Smoking behaviour by genotypes at age 18 (Caucasian only) 142
Table 3.24 Smoking behaviour markers in current smokers only in relation to *CYP2A6* genotype means(standard error), adjusted for age, sex, town and ethnicity 145
Table 3.25 *CSH1.01* allele frequencies in TTHHS (in dichotomy) 147
Table 3.26 *CSH1.01* allele frequencies in TTHHS (in SGEL coding) 149
Table 3.27 Characteristics of population (Europeans only) 150
Table 3.28 Phenotypic analyses for *CSH1.01* D1/T genotype 151
Table 3.29 Phenotypic analyses for *CSH1.01* D2/T genotype 152
Table 3.30 *IGF1*(CA)n allele frequencies in TTHHS 154
Table 3.31 Phenotypic analyses for *IGF1*(CA)n genotype (1) 155
Table 3.32 Phenotypic analyses for *IGF1*(CA)n genotype (1) 155
Table 3.33 Serum ACE activity and exercise-induced LVM change of studied genotypes 158
Table 3.34 Associations between haplotypes and PreACE level and LVM response 159
Table 3.35 Associations between haplotypes and PreACE level and LVM response in homozygotes 159
Table 5.1 Genotype relative risk (GRR) models and GRR for Aa and AA 189
Table 5.2 Sample size and power calculation using Genetic Power Calculator 189
Table 5.3 Sample size and power calculation using Genetic Power Calculator and CHROMSCAN software 189

List of Figures

Figure 1.1 Quantititative scheme of nicotine metabolism 38

Figure 1.2 Schematic diagram of the position and structure of *CYP2A6* gene 42

Figure 1.3 Schematic diagram of the *CYP2A6* gene structure and variants 45

Figure 1.4 Schematic diagram of synaptic neurotransmission 50

Figure 1.5 Schematic diagram of nicotine working mechanism 51

Figure 1.6 Schematic diagram of dopamine synthesis and metabolism 52

Figure 2.1 Schematic diagram of two-step PCR method for genotyping *CYP2A6* 91

Figure 2.2 Schematic presentation of the tetra-primer ARMS-PCR method 98

Figure 2.3 *ACE* I/D genotyping primers design and bands size 106

Figure 2.4 Power calculator window for continuous quantitative trait analysis 114

Figure 2.5 Schematic illustration of the 384-well MADGE system 115

Figure 2.6 ARMS-MADGE for genotyping *CYP2A6*1A* and **2* 116

Figure 3.1 768-Well MADGE 120

Figure 3.2 *CYP2A6* genotype designations 123

Figure 3.3 *CSH1.01* allele distribution and designation 147

Figure 3.4 *IGF1*(CA)n allele distribution and designation 153

Figure 4.1 The DNA sequence of the *CYP2A6*12* allele as compared to *CYP2A6*1* in the region of –148bp to exon 2 165

Figure 5.1 Demonstration of SNP array coverage of segmental duplication region 186

Acknowledgement

I extend my sincere gratitude and appreciation to many people who made this thesis possible.

My special thanks go to my supervisors Professor Nick Cross and Professor Ian Day for their constant supervision, inspiration, encouragement and guidance. I also want to thank Professor Pat Jacobs and Dr John Crolla for their strong support and encouragement.

I gratefully acknowledge the support from my colleagues in both Human Genetics Division of Southampton University and Wessex Regional Genetics Laboratory at Salisbury District Hospital.

Professor Derek Cook, Peter Whincup and Julie Gilg from St Georges Hospital, Hugh Montgomery and John Payne from University College London are also thanked for their support through collaborative projects.

I would like to thank my family, my husband Dr Xiao-he Chen and my daughter Yanmei Chen. They have been of great support throughout my PhD study.

Last but not the least, I would like to thank my examiners Professor Anthony Brookes and Dr Clive Holmes for their constructive suggestions and advice for revision of my thesis.

Abbreviations

ACE I/D	Angiotensin converting enzyme insertion/deletion
aCGH	Array comparative genomic hybridisation
ARMS	Amplification Refractory Mutation System
ASP	Affected sib-pair
BMI	Body mass index
CNV	Copy number variant
CNP	Copy number polymorphism
CYP2A6	Cytochrome P450 2A6
CSH	Chorionic somatomammotrophin (placental lactogen)
DASH	Dynamic Allelic-Specific Hybridisation
DBH	Dopamine β-hydroxylase
DNA	Deoxyribonucleic acid
DOP	Degenerate oligo primer amplifications
DRD4	Dopamine receptor 4
GH	Growth hormone
GWAS	Genome-wide association study
HGP	Human genome project
HI	Haploinsufficiency
HTR	Haplotype trend regression
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IGF-1	Insulin-like growth factor-1
IHMP	International HapMap Project
ISV	Intermediate-sized variant
LARGE	Lichfield Army Recruit Growth in Exercise (Heart Study)
LCR	Low-copy repeats
LCV	Large-scale copy number variant
LD	Linkage disequilibrium
LOD	Logarithm of the odds (to the base of 10)

LVM	Left ventricular mass
MADGE	Microplate Array Diagonal Gel Electrophoresis
MAOA	Monoamine oxidase A
PHI	Partially haploinsufficiency
PCR	Polymerase Chain Reaction
SGEL	Southampton Genetic Epidemiology Laboratory
SNPs	Single nucleotide polymorphisms
TDT	Transmission disequilibrium test
TH	Tyrosine hydroxylase
TTHHS	Ten Town School Heart Study
VNTR	Variant number of tandem repeat
5HT2A	Serotonin receptor 2A
2LD	Two-locus linkage disequilibrium

List of publications

1. Santiago Rodriguez, Tom R Gaunt, **Shuwen Huang**, Mohammad R. Abdollahi, Khalid Al-Harbi, Xiao-he Chen, Ian N. M. Day (2005). Genomic Applications of Microplate-Array Diagonal-Gel Electrophoresis (MADGE). BIOforum Europe.
2. **Shuwen Huang**, Derek Cook, Lesley J Hinks, Xiao-he Chen, Shu Ye, Julie Gilg, Martin Jarvis, Peter Whincup and Ian N.M. Day (2005). *CYP2A6*, *MAOA*, *DBH*, *DRD4*, and *5HT2A* genotypes smoking behaviour and cotinine levels in 1,518 UK adolescents. Pharmacogenetics and Genomics 15(12):839-850.
3. Santiago Rodriguez, **Shuwen Huang**, Xiao-he Chen, Tom R. Gaunt, Holly E. Syddall, Julie A. Gilg, George G. Miller, Cyrus C. Cooper, Derek G. Cook, Peter H. Whincup and Ian N.M. Day (2006). A study of *TH01* and *IGF2-INS-TH* haplotypes in relation to smoking initiation in three independent surveys. Pharmacogenetics and Genomics 16(1):15-23.
4. **Shuwen Huang**, Xiao-he Chen, Ian N. M. Day (2006). Combination of 768-well microplate array diagonal gel electrophoresis with duplex PCR of X and Y chromosome markers for quality control of epidemiological DNA banks. Electrophoresis Aug;27(15):3038-41.
5. **Huang, Shuwen**, Chen, Xiao-he, Payne, John R, Pennell, Dudley J, Gohlke, Peter, Smith, Matt J, Day, Ian N.M, Montgomery, Hugh E, Gaunt, Tom R, (2007) Haplotype of growth hormone and angiotensin I-converting enzyme genes, serum angiotensin I-converting enzyme and ventricular growth: pathway inference in pharmacogenetics. Pharmacogenetics & Genomics. 17(4):291-294, April 2007.
6. Mohammad Reza Abdollahi, **Shuwen Huang**, Santiago Rodriguez, Philip Alexander Isles Guthrie, George Davey Smith, Shah Ebrahim, Debbie A Lawlor, Ian N M Day, Tom R. Gaunt, (2007) Homogeneous assay of rs4343, an *ACE* I/D proxy, and an analysis in the British Women's Heart and Health Study (BWHHS). Disease Markers (in press).

Chapter 1 Introduction

1.1 Introduction to genetic analysis

Deciphering the aetiology of human disease has long been the goal of medical research. It is widely recognised that most human diseases result from the interaction of genetic susceptibility factors and modifiable environmental factors. Even so-called single-gene disorders may actually develop from the interaction of both genetic and environmental factors, e.g. phenylketonuria (PKU) only occurs when both the genetic variant (phenylalanine hydroxylase deficiency) and the environmental exposure (dietary phenylalanine) co-incide[1]. Environmental factors work on the human body such that people who are exposed to these factors will have a greater chance of developing a certain disease than those without the exposure. Genetic factors influence predisposition to disease so that when exposed to the same disease risk factors, some people are more likely to develop the disease compared to others. Comprehensive understanding of disease causation relies on thoroughly identifying environmental risk factors, causal genes and variants, as well as gene-gene and gene-environmental interactions. Public health research has made great progress in identifying relevant environmental factors and preventing people from the exposure of these factors, which has greatly decreased the prevalence of many diseases and improved human health. Decoding the genetic variants associated with clinical traits represents an important step in the path towards improved prevention, diagnosis and treatment of diseases [2, 3].

1.1.1 Genetic variations

Genetic information is encoded by DNA (deoxyribonucleic acid) molecules - usually in the form of a double helix, organised as strands running start-to-end and joined by

hydrogen bonds along their lengths. Each strand is a chain of chemical "building blocks", called nucleotides, of which there are four types; two purines: adenine (A) and guanine (G), and two pyrimidines: thymine (T) and cytosine (C). These bases pair specifically such that A is opposite T through two hydrogen bonds and C pairs with G through three hydrogen bonds. A gene is an ordered sequence of nucleotides located in a particular position along the long chain that encodes a specific functional product (i.e., a protein or RNA molecule). An organism's complete set of DNA is called its genome. The human genome sequence contains three billion base pairs of genetic code, strung together into different combinations in long chains within 23 pairs of chromosomes: 22 pairs of autosomal chromosomes and a pair of sex chromosomes. The human genome contains about 20,000 to 25,000 protein-coding genes [4], representing about 3% of the human genome sequence. It was reported that different individuals share about 99.9 percent of their genetic code [5], but recent studies revealed that we are more different to one another than was previously thought due to large scale copy number variation [6]. In addition to these copy number variants (CNV), genetic variation can be present in many forms, including single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs; e.g., mini- and microsatellites), the presence/absence of transposable elements (e.g., *Alu* elements), and structural alterations (e.g., deletions, duplications, and inversions) [7]. The following sections consider these variations in more detail.

1.1.1.1 Single nucleotide variations

Single nucleotide variations are DNA sequence variants that occur when a single nucleotide is changed. For the same position in the DNA sequence (locus), alternative base pairs are observed (alleles), whose combination in one individual is a genotype. One allele (majority allele) may be found more frequently than the other (minor allele)

in a population. When the minor allele is present at a frequency of 1% or more in the population, the locus is categorised as single nucleotide polymorphism or SNP. SNPs have been thought of as the most simple form and most common source of genetic polymorphisms in the human genome. There are estimated to be over 10 million SNPs within the human population, averaging one every 300 nucleotides among the 3 billion nucleotide base pairs in the human genome [8].

SNPs can be grouped into different categories according to their base pair substitution, location and function. There are two types of nucleotide base substitutions resulting in SNPs: (i) transitions - substitutions between purines (A, G) or between pyrimidines (C, T), which constitute two thirds of all SNPs, and (ii) transversions - substitutions between a purine and a pyrimidine. SNPs may fall within coding sequences of genes, noncoding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to redundancy in the genetic code. A SNP in which both forms lead to the same protein sequence is termed synonymous; if different proteins are produced they are termed non-synonymous. There are two types of non-synonymous polymorphisms: a missense mutation changes the protein by causing a change of a codon, while a nonsense mutation results in a premature stop codon. Most of the functionally important SNPs are believed to be located within genes. However, SNPs that are not in protein coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNAs. The association between SNPs and disease is considered in more detail below.

1.1.1.2 Variable number of tandem repeats (VNTRs)

There are three types of tandem repeated sequences categorised according to array size: satellite DNA, minisatellite DNA and microsatellite DNA. VNTRs mainly refer to minisatellites and microsatellites. Satellite DNA is comprised of very large arrays (often within 100kb to several Mb size range) of tandemly repeated DNA. The repeat unit may be a simple sequence of a few nucleotides or a moderately complex sequence. Satellite DNA makes up most of the heterochromatic regions of the genome, and is notably found in the vicinity of the centromeres. Satellite DNA is transcriptionally inactive. The vast majority of minisatellites are also transcriptionally inactive, but unlike satellite DNA, minisatellites are often located at or close to telomeres, have core base sequences of 9 to 80bp that are repeated to give arrays within the range 0.1-20kb. The number of repeats for a given minisatellite may differ between individuals and this feature is the basis of DNA fingerprinting [9]. Microsatellite DNA, also known as short tandem repeat (STR), are small arrays of tandem repeats of a simple sequence (usually less than 10bp), and the whole repetitive region spans less than 150bp. They are interspersed throughout the genome, and account for about 3% of the human DNA complement [10]. Although microsatellite DNA has generally been identified in intergenic DNA or within the introns of genes, a significant percentage is located in coding DNA; especially tri- and hexa-nucleotide repeats, which are more abundant in exons [10, 11]. Microsatellites are commonly used for mapping, linkage analysis and to trace inheritance patterns. In addition some microsatellites that fall within coding or non-coding exons may have pathological consequences if they expand, as is the case for Fragile X syndrome and Huntington's disease [12, 13].

1.1.1.3 Transposable elements

Almost all of the interspersed repetitive noncoding DNA in the human genome is derived from transposable elements (transposons), mobile DNA sequences which can migrate to different regions of the genome [11]. There are four major classes of transposon in the human genome, which can be organised into two groups according to the method of transposition: retrotransposons (copy and insert) and DNA transponsons (cut and paste). Retrotransposons include short interspersed nuclear elements (SINEs) (100-400bp, e.g., *Alu* family), long interspersed nuclear elements (LINEs, e.g., a full length of ~ 6kb Kpnl repeats), and transposable elements with long terminal repeats (LTR transposons). Although thesė sequences are derived from mobile elements, the great majority have lost the ability to move around due to the accrual of multiple mutations [11].

Alu sequences are the most abundant mobile element-derived sequences in the human genome. They are about 280bp, GC rich and located in intergenic regions, introns and also untranslated regions of some genes. There are over 1 million *Alu* sequences interspersed throughout the human genome, and it is estimated that about 10% of the mass of the human genome consists of *Alu* sequences. Although only less than 0.5% are polymorphic [14], *Alu* sequences have been implicated in several inherited human diseases, including various forms of cancer [15, 16]

1.1.1.4 Insertion and deletion (Indels)

The word **indel** is a contraction of insertion and deletion, referring to the two types of genetic mutation that are often considered together because of their similar effect. At the DNA level, an insertion means the addition of one or more base pairs into a genetic sequence. Although insertions are typically short, on a chromosome level an insertion

refers to the addition of a larger sequence into a chromosome that is visible by conventional microsocopy. This can happen due to unequal crossover during meiosis. Deletions are mutations in which a part of a chromosome or a sequence of DNA is missing. Any number of nucleotides can be deleted, from a single base to thousands of bases to larger pieces of chromosomes [17]. As inserted or deleted segments of DNA > 1kb are generally referred to as CNVs, indels typically refer to small insertions/deletions (e.g., < 1kb, and often just a few bases). If indels occur in a coding region, unless the length of an indel is a multiple of 3, they produce a frameshift mutation.

1.1.1.5 Copy number variants

A CNV is a segment of DNA that is 1kb or larger and is present at a variable copy number in comparison with a reference genome, and may include insertions, deletions and duplications [8]. There are several terms in current use that refer CNVs, such as large-scale copy number variants (LCVs), intermediate-sized variants (ISVs), and copy number polymorphisms (CNPs). Insertion or deletion of transposable elements (e.g., ~6kb KpnI repeats) was suggested not to be counted as CNVs to minimise the complexity of CNV analysis [7]. Segmental duplications, also known as low-copy repeats (LCR), has been defined as a duplicated region > 1kb in size with the different copies sharing >90% sequence identity. Segmental duplications can be considered CNVs when the copy numbers vary among individuals [8]. LCVs refer to homologous DNA segments of around 50kb to 5Mb, while ISVs typically range from 5kb to 50kb in size [18].

Copy number variants can be observed throughout the whole genome, including genic regions, intergenic regions, or even partially spaning a gene, but they are nonrandomly

distributed. Previously known large duplications and deletions (e.g. > 3-5Mb) are often localised to repeat-rich regions such as telomeres, centromeres, and heterochromatin [19]. Recent studies found that CNVs are more likely to be found near known segmental duplications [20-23], and a notable feature of segmental duplications is their tendency to cluster within pericentromeric and subtelomeric regions [24, 25]. It is estimated that there is a ~4-10-fold enrichment for segmental duplications at sites of human structural variations>10kb in size [18], and larger CNVs are more frequently associated with segmental duplications than smaller ones [7, 22, 26].

Redon et al. (2006) reported 1447 discrete copy number variable regions (CNVRs), covering 12% (approximately 360Mb) of the human genome, that were detected using HapMap samples (270 people with Asian, African or European ancestry) [6]. Although some structural variants have no obvious phenotypic consequences, others play a fundamental role in influencing human disease susceptibility, either through deletion or duplication of the entire copies of dosage sensitive genes, unmasking recessive mutations due to deletion polymorphism, or by position effect of regulatory elements, which have consequences for gene function [8]. There is a wide spectrum of relationships between copy number changes and gene expressions. At one extreme, different copy numbers of the α-defension genes have no effect on total *DEFA1*and *DEFA3* mRNA levels [27]. At the other extreme, triplication of the *SNCA* gene results in a doubling in the amount of α-synuclein protein in blood, and doubles the level of *SNCA* message in brain tissue, with an almost perfect positive correlation between gene dosage, mRNA, and protein levels [28]. In the middle are some reports suggesting that 26-88% of observed variations in expression levels of human genes can be explained by gene copy number alterations [18, 29, 30]. More recently, Stranger et al. (2007) carried out an association analysis of gene expression in cell lines from HapMap individuals in

relation to both SNPs and CNVs, focusing on the genomic regions around expressed genes. Of the 14,072 genes for which expression levels were measured, the authors estimated that 8.75-17.7% of heritable variation in gene expression is due to copy number variation [31]. Diseases and disease susceptibility associated with CNVs are just beginning to emerge, e.g. individuals who had a higher number of copies *CCL3L1* were less likely to become infected with HIV, or to progress to AIDS once they were infected [32]. Deletions associated with segmental duplications are associated with an expanding number of syndromes, for example those involving 15q24 or 17q21 [33-36]

1.1.1.6 Other structural variations

CNVs are unbalanced rearrangements which involve an increase or decrease of sequence compared to a reference genome. Some structural variations are balanced, such as inversions and translocations. Inversions represent a chromosomal defect in which a segment of the chromosome breaks off and reattaches in the reverse direction. Inversions are of two types: paracentric (do not include the centromere) and pericentric (include the centromere). Inversions usually do not cause any abnormalities in carriers as long as the rearrangement is balanced with no gain or loss of genetic information. However, there is an increased chance for the production of unbalanced chromosome rearrangements in the offspring of carriers. Translocation means a transfer of a chromosomal segment to a new position, especially on a nonhomologous chromosome. There are two main types, reciprocal and Robertsonian. Reciprocal translocations usually result in an exchange of material between nonhomologous chromosomes, which are found in about 1 in 600 newborns [37]. Robertsonian translocations involve two acrocentric chromosomes that fuse near the centromere region with loss of the short arms. The resulting karyotype has only 45 chromosomes since two chromosomes have fused together. Carriers of both translocations are phenotypically normal, but there is a

risk of unbalanced gametes that lead to miscarriages or abnormal offspring. With the advent of detailed molecular analysis it has become apparent that rearrangements which appear balanced cytogenetically are often associated with deletions, and it is these deletions which often have pathogenetic consequences [431].

1.1.2 Genetic diseases

A genetic disease, or disorder, is a disease caused by an abnormality in an individual's DNA. Abnormalities can range from a small mutation in a single gene to the addition or subtraction of an entire chromosome or set of chromosomes (www.medicinenet.com). Almost all human disorders are believed to have a genetic component, although the importance of that component varies.

Genetic diseases are normally categorised into the following types: (1) single-gene diseases, (2) multifactorial complex diseases, (3) chromosomal diseases, and (4) mitochondrial genetic diseases.

1.1.2.1 Single-gene (also called Mendelian or monogenic) disease

This type of disease is caused by changes or mutations that occur in the DNA sequence of one gene, resulting in an enzyme or other protein not being produced or having altered functionality. More than 10,000 human diseases are believed to be caused by defects in single genes, affecting 1% of the population as a whole (Online Mendelian Inheritance in Man; www.ncbi.nlm.nih.gov/Omim/). The change can be relatively trivial in its effects, such as colour blindness, or lethal such as Tay-Sachs (a fatal inherited disease of the central nervous system; babies with Tay-Sachs lack an enzyme called hexosaminidase A (hex A) which is necessary for breaking down certain fatty substances in brain and nerve cells). Other disorders, though harmful to those afflicted

with them, appear to offer some advantage to carriers. For example, carriers of sickle cell anaemia and thalassemia appear to have enhanced resistance to malaria. Some other examples of single-gene diseases are cystic fibrosis, Marfan syndrome, Huntington's disease, and hereditary hemochromatosis.

Single gene diseases can be divided into a number of different categories according to how they are transmitted from generation to generation. Some are described as dominant diseases because only one mutant allele is required, and such diseases tend to crop up in every generation, depending on the penetrance of the mutation. Other diseases are described as recessive because both copies of the gene must be defective in order for the disease to occur. These recessive diseases often skip generations because mutant alleles can be carried without any effect if a normal allele is also present. Many single gene disorders affect both sexes equally. However, where the relevant gene is present on the X-chromosome, the associated disease is usually more common in males (genome.wellcome.ac.uk).

1.1.2.2 Multifactorial (also called complex or polygenic) disease

Many of our most common diseases run in families, but they lack the simple inheritance patterns characteristic of single gene disorders. In complex disorders, pedigrees reveal no Mendelian inheritance patterns, and gene mutations are often neither sufficient nor necessary to explain the disease phenotype [432]. Their more complicated nature makes genetic analysis much more difficult than single-gene or chromosomal disorders. Some of the most common chronic disorders are multifactorial disorders e.g. heart disease, high blood pressure, Alzheimer's disease, arthritis, diabetes, cancer, obesity, and behavioural disorders.

1.1.2.3 Chromosomal disease

Conditions caused by a change in the number (e.g. aneuploidy) or structure of chromosomes (e.g. translocation, inversion, deletion, and duplication) are known as chromosomal disorders. Changes that affect entire chromosomes or segments of chromosomes can cause problems with growth, development, and systemic functions. These changes can affect many genes along the chromosome and alter the proteins coded for by those genes. Some types of major chromosomal abnormalities can be detected by microscopic examination, while others are too small to be detected using microscopy. The uncovering of these sub-microscopic abnormalities relies on new technologies, such as microarray. Surprisingly, a range of chromosomal abnormalities that are visible by conventional microscopy may be inherited without any obvious phenotypic consequences [38].

1.1.2.4 Mitochondrial disease

Mitochondrial diseases comprise those disorders that in one way or another affect the function of the mitochondria and/or are due to changes in mitochondrial DNA. Cells not only have 46 chromosomes with DNA inside the nucleus, but small cellular organelles called mitochondria outside the nucleus also have small amounts of DNA. The mitochondrial genome is only around 16,500 bases, however this is enough to encode several proteins and RNA molecules. Mitochondria serve as a cell's power plant, since they processes glucose to produce energy. Mitochondrial disease can be caused by mutations in the mitochondrial DNA ("pure" mitochondrial DNA disease, fully maternally inherited) or by mutations in the chromosomal DNA that affect the function of mitochondria (such diseases do not have the classic maternal inheritance pattern but

often have the traditional Mendelian inheritance patterns, e.g. autosomal recessive or dominant).

1.1.3 Genetic analysis of complex diseases

The many possible approaches to mapping the genes that underlie common diseases fall broadly into two categories: linkage analysis and association analysis.

1.1.3.1 Linkage analysis

Linkage analysis is a statistical technique used to identify the location on a chromosome of a given gene involved in a trait/disease relative to the known location of chromosomal markers. Linkage is the tendency for genes and other genetic markers to be inherited together because of their location near one another on the same chromosome. The further apart two genes on the same chromosome are, the more likely it is that recombination between them will occur. Recombination is the exchange (crossing over) of DNA between a chromosome pair, usually at meiosis. The recombination fraction (θ) is a measure of genetic distance that represents the proportion of recombinant gametes with respect to two loci. Two loci that show 1% recombination are defined as being 1 centimorgan (cM) apart on a genetic map. As a very rough rule of thumb, 1cM=1Mb, but recombination fractions do not exceed 0.5 however great the physical distance. For loci on different chromosomes or far apart on the same chromosome, recombinant and non-recombinant gametes are equally likely to occur, so θ equals 0.5 and the loci are independent. For loci that are close together on the same chromosome, non-recombinant gametes are more likely to occur, and therefore $\theta < 0.5$, and the loci are linked. When θ is between 0 and 0.5, the smaller the value of θ is, the closer the two loci are. The basic idea of linkage mapping is to use the pattern of

allele segregation in the pedigrees (families) to estimate θ between a marker locus and an unknown trait locus.

There are two basic types of linkage analysis: parametric or model-based approach (classical LOD score analysis), and non-parametric or model-free approach (allele sharing analysis). Parametric linkage analysis requires specification of genetic parameters, such as mode of inheritance (dominant/recessive), penetrance, disease-allele frequency, whether phenocopies may occur (an environmentally induced, nonhereditary variation in an organism, closely resembling a genetically determined trait), and mutation rates (hence the term 'model-based') describing the mode of disease inheritance. The segregation of the disease phenotype due to an unknown gene and random genetic markers at known locations in the genome is followed in informative pedigrees, so one can 'scan' the genome for loci that might influence that disease and localise the disease gene with respect to the known position of the linked genetic marker. Typically, genome scans have been performed using hundreds of polymorphic microsatellite markers that span the genome. The statistical method used to estimate the genetic distance between the hypothetical disease locus and specific markers is the LOD score method which is based on calculation of the likelihood of the observed pedigree data, given the assumed input parameter values. **LOD** stands for logarithm of the odds (to the base of 10). It is calculated using the following formula:

$$\text{LOD score} = z = \log \left[\frac{\text{probability of birth sequence with a given linkage value}}{\text{probability of birth sequence with no linkage}} \right]$$

Or $$\text{LOD}(\theta) = \log_{10}[L(\theta) / L(0.5)]$$

A LOD score of 3 means the odds are a thousand to one in favour of genetic linkage and is conventionally the threshold used in human linkage analysis to declare significant

linkage, although in genome-wide testing with many markers higher critical limits usually apply [39]. Lander et al showed that a LOD score of 3 is equivalent to a genome-wide significance level of about 0.09. A higher threshold of at least 3.3 would be necessary to ensure that the genome-wide type I error rate was in fact 0.05 [40]. In the past two decades, many genes, implicated in simple (Mendelian) diseases, have been identified by using genetic linkage and positional cloning methods. Notably, it has been for rare, severe, high-risk mutations, that linkage mapping has been spectacularly successful, for which the diagnosis is least ambiguous and there is a near one-to-one correspondence between genotype, and phenotype. In addition, linkage analysis usually requires large families with multiple affected members. However when there is a significant degree of inbreeding, smaller families may be analysed by screening for regions of homozygosity that are shared between affected individuals. This is known as autozygosity mapping [41].

For complex diseases, the genetic model of disease transmission is unknown and thus model-free or non-parametric methods are applied to search for underlying genetic abnormalities. Analogous to autozygosity mapping, allele sharing methods are based on the prediction that for a marker locus near a disease-causing gene, pairs of siblings both affected with the disease will share marker alleles more often than expected from the laws of Mendelian inheritance. The most common allele-sharing design is the affected-sib-pair (ASP), in which the observed IBD (identity by descent) sharing for 0, 1, or 2 alleles is compared with that which is expected under no linkage (0.25, 0.5, and 0.25, respectively). Chromosomal regions that are shared between affected siblings more often than expected under random segregation will have a higher probability of containing gene(s) that contribute to susceptibility for a given disease.

1.1.3.2 Genetic association analysis

The application of linkage analysis to complex disorders has had limited success, most likely because of genetic heterogeneity (multiple genetic causes leading to the same disease), environmental phenocopies (purely environmental forms of the disease), incomplete penetrance (subjects inheriting a disease gene but not developing the disease), genotype-by-environment interactions (non-additivity of genetic and environmental influences on disease development), and multilocus effects (more than one gene influences disease development) [42, 43]. It remains largely unknown what sort of genetic variants explain the inherited component of complex traits, but recent evidence suggests that common genetic variants will explain at least some of the inherited variation in susceptibility to common disease (common variant/common disease hypothesis) [44]. Genetic association studies, in which the allele or genotype frequencies are determined in affected individuals and compared with those of controls, has been suggested as an effective approach to detecting the effects of common variants with modest effects [45]. Genetic association studies can be categorised into either family-based, or population based according to the subjects selected; also candidate gene association studies, or genome-wide/whole genome association studies according to the coverage of variation investigated.

Family based association studies use small nuclear families and extended pedigrees containing affected and unaffected individuals. They focus on transmission of marker alleles from parents to offspring and compare transmission of marker alleles to affected and unaffected offspring using standard statistical tests. The basic premise of its basic format, the transmission disequilibrium test (TDT), is that certain alleles at a locus are disproportionately transmitted from parents to affected offspring; therefore the basic sampling unit for the TDT is a nuclear family with at least one affected child (i.e. father,

mother & affected child). Various further extensions have been developed for multiallelic markers [46], multiple siblings in a family [47], missing parental data [48, 49] and quantitative traits [50-53]. The TDT-related tests have the advantage that they are immune to population stratification. However, a few common pitfalls of the TDT are worth mentioning. First, in its original form, the TDT requires parental genotypes, which are not easily accessible for late-onset disorders. Second, the method effectively throws away some genotype information, owing to its reliance on heterozygous parents. This creates a loss of statistical power to detect genuine allelic association. Consequently, although more robust in the presence of population stratification, the family-based methodologies can be more difficult or even impossible to implement, and require significantly more patients and family collections than case–control studies [54].

Population based association analysis is performed by cohort or case-control studies. The former is a study in which a population (i.e. a cohort) is defined according to the presence or absence of factor(s) (e.g. a certain allele, genotype or haplotype) that might influence the probability of a given disease or other outcome occurring. The cohort is then followed to determine if those with the factor(s) are indeed at greater risk of the outcome. Although cohort studies can measure risk at the population level, and therefore remain essential for genetic epidemiologists, they are usually not efficient for the initial stages of gene discovery, as the study samples generated will have far fewer individuals with the disease than without the disease, and the nested case-control samples that emerge will often be small [55]. Consequently, the mainstream of genetic association analysis is undertaken as case-control studies, in which genetic determinants are investigated in individuals with the disease (or other outcome) of interest (cases), compared with a suitable control group of persons without the disease. Case-control genetic association studies are no different conceptually from the case-control methods

that have been developed for use in epidemiological studies. Cases and controls are enrolled from the same source population but are unrelated [56]. Significant differences in allele frequencies between cases and controls are taken as evidence for involvement of an allele in disease susceptibility. Alternatively, haplotype frequencies rather than allele frequencies can be compared in cases and controls [42].

Until recently, techologies were not available to scan the whole genome at a sufficient density for association studies and so most case-control studies have employed a candidate gene approach. Candidate gene analysis involves the selection of a gene or genes on the basis that their known function fits in some way with knowledge of the pathogenesis of the disease. These properties may include the expression pattern of the gene (i.e. a gene involved in a renal disease should be expressed in the kidney), the function of the gene product and the relationship of the gene to other known disease gene [57]. Candidate gene studies focus on genotyping markers (usually SNPs) in a set of candidate genes in samples of unrelated affected cases and unaffected controls. Hattersley and McCarthy summarised a range of sources that can be used to identify candidate genes including biology, pharmacology, animal models, monogenic or syndromic forms of disease, positional information and prior association data [55]. The benefits of candidate gene studies include that they are hypothesis driven and can focus on aspects of disease, they have increased statistical power and positive association results can easily be understood and often tested by functional assays. However, there are also some disadvantages of candidate gene approaches, most obviously their dependence on incomplete knowledge of the molecular mechanisms underlying complex traits makes systematic discovery of all associated genetic variants impossible. Furthermore, different gene-gene interactions and gene-environmental interactions affect the results and lead to the frequent lack of replication in other studies.

Genome-Wide Association Studies (GWAS) are defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or other condition. Progress in the following areas has facilitated the realisation of GWAS: (i) the completion of the Human Genome Project, (ii) the determination of LD patterns on a genome-wide scale through the International HapMap Project, (iii) the usage of minimum sets of tagging SNPs to cover most of the genetic variants, (iv) the development of high-throughput genotyping methods and whole genome scan technologies, (v) the reduction of genotyping costs, and (vi) the development of appropriate statistical programs. Early studies have been fraught with problems, particularly the inability to confirm positive findings in follow up studies and it has become apparent that the sample sizes need to be larger and the criteria for candidate positive associations more strict [58]. Many large GWAS projects are currently underway, e.g. as part of the Wellcome Trust Case Control Consortium (www.wtccc.org.uk/). Some initial studies have already been published and have proved that the concept of GWAS really does work, although as with other case-control studies the composition of the control group is critical. Early successes have included the association between a complement factor H polymorphism and age-related macular degeneration [59-61], and the association of a haplotype at 8q24 with prostate cancer predisposition [62, 63].

Linkage disequilibrium is the key element in genetic association studies; therefore this is discussed in more detail in the following section.

1.1.3.3 Linkage disequilibrium and genetic association study

An association between a genetic variant and a disease is not necessarily because that variant is biologically causal, but may be because it is statistically correlated with one or

more true causal variants. This possibility arises because alleles at different loci are frequently found together, either more or less often than expected based on their frequencies [64]. In population genetics, this non-random pattern is called linkage disequilibrium (LD). In the following section, LD, haplotypes, tag SNPs and the International HapMap Project will be discussed in more detail.

Linkage disequilibrium and haplotype

Individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites, a correlation is known as LD. A particular combination of alleles along a chromosome is termed the haplotype, which is a contraction of "haploid genotype" [3]. LD is created when a new mutation occurs on a chromosome that carries a particular allele at a nearby locus, and is gradually eroded by recombination. The extent of linkage disequilibrium in populations is expected to decrease with both time (t) and recombinational distance (r, or the recombination fraction) between markers. Theoretically, LD decays with time and distance according to the following formula, where D_0 is the extent of disequilibrium at some starting point and D_t, is the extent of disequilibrium t generations later:

$$D_t=(1-r)^tD_0$$

There are several measures of LD which have different strengths, depending on the context. **D** was initially developed by Lewontin (1964) to quantify disequilibrium as the difference between the observed frequency of a two-loci haplotype and the frequency it would be expected to show if the alleles are segregating at random [65]. Consider two adjacent loci-A and B, with two alleles (A, a, and B, b) at each locus with the observed frequency of the AB haplotype represented by P_{AB}. Assuming independent assortment of alleles at the two loci, the expected haplotype frequency is calculated as the product of the allele frequency of each of the two alleles, or $P_A \times P_B$, where P_A is the frequency

of allele A at the first locus and P_B is the frequency of allele B at the second locus. So, one of the simplest measures of disequilibrium is

$D=P_{AB}-P_A \times P_B$

D is known as the linkage disequilibrium coefficient, and is dependent on allele frequencies in the population [66]. Other measures of linkage disequilibrium have also been developed. The two most common measures are the absolute values of **D'** and $\mathbf{r^2}$. They can be calculated by the following formulae:

$$D' = \begin{cases} \dfrac{D}{\min(P_A \times P_b, P_a \times P_B)} & D > 0 \\ \dfrac{D}{\min(P_A \times P_B, P_a \times P_b)} & D < 0 \end{cases}$$

$$r^2 = \frac{D^2}{P_A \times P_a \times P_B \times P_b}$$

The case of D'=1 is known as complete LD (observations at one marker provide complete information on the other marker, making one of the loci redundant for purposes of gene mapping). Values of D'<1 indicate that the ancestral LD has been disrupted. The magnitude of values of D'<1 has no clear interpretation (e.g. it is difficult to say how different 0.3 is from 0.7), but statistically significant values of D' that are near to one provide a useful indication of minimal historical recombination [66]. The measure r^2 is in some ways complementary to D'. r^2 is equal to D^2 divided by the product of the allele frequencies at the two loci. $r^2 = 1$ is known as perfect LD [67]. Hill and Roberson deduced that $E[r^2] = 1/1+4Nc$ where c is the recombination rate in morgans between the two markers and N is the effective population size. This equation illustrates that LD is a function of recombination rate and property of population. To arrive at this equation, Hill and Roberson assumed that the population was an "ideal" large, random-mating population without natural selection and mutation [68]. Early

studies involved the use of microsatellite markers and simulations to predict the extent of LD in the human genome [69-71]. However SNPs have become the main tool for LD research in human populations as they are abundant, have a relatively low mutation rate and are easily typed by methods that are at least partially automated [69, 72-74]. How does LD help to map disease genes? For example, consider two independent mutations (locus A and locus B) on the same chromosome. Suppose locus B is an unknown causative region with allele B1 being a risk factor for breast cancer compared to the other allele B2. We expect B1 to be elevated in breast cancer cases relative to controls. Now imagine that the A and B loci are in linkage disequilibrium, and specifically that the A1 allele is more often found with the B1 allele than the B2 allele. In that case, not only B1 is elevated in cases, but A1 is too because of its LD with B. Thus if we did not know about the B locus, but typed the A locus, it could lead us to the B locus. In this example, there are 4 possible allele combinations (haplotypes) at these two loci: A1B1, A1B2, A2B1, and A2B2. In general for n SNPs, there are 2^n possible haplotypes. The number of SNPs required for a genome-wide association study depends on the pattern of LD, and studies showed that LD pattern varies greatly across the human genome with some regions of high LD interspersed by regions of low LD [3]. These high LD regions are referred to as haplotype blocks in the literature [75]. The size of block varies depending on the regions of the chromosome and the algorithm used for analysis. Between 65% and 85% of the human genome is organised in haplotype blocks [3]. Each block comes in three or four common versions that capture the majority of genetic diversity throughout the entire human population. Blocks may contain a large number of SNPs, but a few SNPs are enough to uniquely identify the haplotypes in a block due to LD. The specific SNPs that identify the haplotypes are called tag SNPs. These tag SNPs are extremely useful for association studies, as genotyping effort could be greatly reduced without much loss of power for association studies [75] [3].

International HapMap Project (IHMP)

In October 2002, the International HapMap Project (IHMP) was launched to create a freely-available public resource to increase the power and efficiency of genetic association studies, and to determine the common patterns of DNA sequence variation in the human genome [76]. The HapMap is a catalog of common genetic variants that occur in human beings. It describes what these variants are, where they occur, and how they are distributed among people within populations and amongst populations in different parts of the world (http://www.hapmap.org). The HapMap DNA samples came from a total of 270 people: Yoruba people in Ibadan, Nigeria (30 both-parent-and-adult-child trios), Japanese from Tokyo (45 unrelated individuals), Han Chinese from Beijing (45 unrelated individuals), and Caucasians (30 trios, residents of Utah with ancestry from Northern and Western Europe). A total of ~1 million SNPs were successfully typed in all 270 samples during PHASE I of the project, and a total of >3,500,000 SNPs have been typed in PHASE II. All data can be accessed at http://www.hapmap.org/.

The IHMP is a natural extension of the Human Genome Project. Whereas the reference sequence constructed by the Human Genome Project is informative about the vast majority of bases that are invariant across individuals, the HapMap focuses on DNA sequence differences among individuals [3]. HapMap has provided data on LD pattern, recombination rate, haplotype pattern and tag SNPs in the whole genome. It has broad applications in selecting relevant tag SNPs [77-79], identifying large common deletions [26, 30], integrating with genomic features [59], meta-analysis of association studies [80], loss of heterozygosity in cancer studies and other applications [81].

Linkage disequilibrium between different types of genetic variation

As discussed above, the IHMP has provided genome-wide data on tag SNPs, haplotype patterns and LD patterns, which has substantially improved the cost-effectiveness of

association studies based on SNPs. However, it has become apparent that other abnormalities, such as small indels and CNVs are also common sources of genetic variation and may be causally associated with disease. To what extent can tag SNPs capture this sort of variation? Bhangale et al (2005) compared 2393 diallelic indels in a set of 330 candidate genes with flanking SNPs,and found strong linkage disequilibrium between the two [82, 83]. Recently, three studies have focused on the identification and characterisation of fine-scale deletions by taking advantage of technology designed to catalogue SNPs in the human population during the HapMap project [26, 30, 82]. The key issue here is whether structural variation events are ancient or recurrent [18, 84]. If a structural variant occurred only once during human evolution on a single ancient haplotype, it may show linkage disequilibrium with nearby genetic markers. However there would be little or no LD with flanking SNPs if the structural variant events occured on multiple haplotype backgrounds [18]. McCarroll et al. (2006) used dense SNP genotype data from the HapMap to discover clustered sites that do not pass the Hardy-Weinberg equilibrium test as a means to characterise and catalogue segregating deletion variants across the human genome. They identified 541 deletion variants ranging from 1-745kb (median 7.0kb) in size, and observed strong linkage disequilibrium between SNPs and 90 validated deletions [30]. Hinds et al. (2006), in an independent study have also reached the conclusion that linkage disequilibrium between SNPs and intermediate-length deletions (70bp to 10kb) is similar to linkage disequilibrium between SNPs by using a more direct approach, in which regions of reduced signal intensity were identified by hybridising long-range PCR products against an oligonucleotide microarray [82]. Both studies suggested that at least in the majority of cases, common deletions and SNPs have a similar evolutionary history and that most deletion polymorphisms can be effectively assayed by neighbouring SNPs. Consequently their association with disease can be evaluated in the course of SNP-

based whole-genome association studies without the need to directly assess the deletion variant itself [30, 82]. However, the third study conducted by Conrad and colleagues using the HapMap data gave a different opinion. By applying customised array comparative genomic hybridisation (CGH) technology to validate predicted deletions, they observed that even within small regions, deletion polymorphisms were often complex with multiple deletion breakpoints. They also found that some deletions that occurred in a single region were often present on different haplotype backgrounds, suggesting that these deletions may have occurred in multiple founders [18, 26]. So it seems likely that tag SNPs have the potential to capture perhaps most genetic variation but certainly not all of it.

Stranger et al. reported that CNVs have a smaller contribution to gene expression phenotypes than SNPs do, but the CNV and SNP contributions are largely independent, so interrogation of the genome for both types of variants may be an effective way to elucidate the causes of complex phenotypes and disease in humans [31]. Thus, although considerable progress has been made, a systematic view of interactions between various kinds of variations in the genome will not be formed until a comprehensive understanding of the architecture of human genome has been achieved. Ultimately this is likely to be resolved by cheap whole genome sequencing.

In summary, a variety of increasingly sophisticated approaches have been developed as a means to understand the genetic basis of human disease. The following sections detail specific progress that has been made in the understanding of the genetics of two independent complex traits that are the subjects of this thesis: smoking and growth pathways.

1.2 Genetic analysis of smoking

1.2.1 The importance of genetic studies of smoking

1.2.1.1 The 'popularity' of cigarette smoking

Despite the well-known adverse health effects of cigarette smoking, there are still about 1.25 billion smokers in the world, representing about one third of the global population aged 15 and over. Between 80,000 and 100,000 children worldwide start smoking every day, and almost half of the world's children are exposed to second-hand smoke. In the United Kingdom, about 10 million adults smoke cigarettes: 25% of men and 23% of women. About 450 children start smoking every day. By the age of 11, one-third of children and by 16 years two-thirds of children have experimented with smoking. Currently, about one fifth of Britain's 15 year-olds – 16% of boys and 25% of girls – are regular smokers (at least one cigarette a week) despite the fact that it is illegal to sell cigarettes to children aged under16 [85]. Smoking is spreading rapidly in developing countries and among women. Currently, 50% of men and 9% of women in developing countries smoke, as compared with 35% of men and 22% of women in developed countries [86]. For example, the prevalence of smoking among males in populous Asian countries is now far higher than in Western countries—45% in India, 53% in Japan, 63% in China, 69% in Indonesia, and 73% in Vietnam[87].

In general, there are several trends in cigarette smoking: 1) the age of smoking initiation is decreasing and tobacco use is rising among the youth in many countries; 2) more girls than boys are acquiring the habit; and 3) smoking is decreasing in most developed countries, yet increasing in most developing nations.

1.2.1.2 The consequences of smoking

Smoking is the single largest preventable cause of premature death and a major cause of morbidity. It is currently responsible for the death of one in ten adults worldwide (about 5 million deaths each year) [88]. Worldwide about 4 million people die annually from tobacco related causes and by the late 2020s the estimated toll will be about 10 million. Smoking-related diseases are responsible for 1 in 10 adult deaths worldwide [89]. Every eight seconds, someone dies from tobacco use. In the UK, about 114,000 smokers die from smoking related causes every year, accounting for one fifth of all UK deaths [85]. On average, cigarette smokers die 10 years younger than non-smokers, according to a 50-year-long study of smoking and death among British doctors [90].

Smoking is associated with a high incidence of various types of cancers, respiratory and cardiovascular diseases, gastrointestinal disorders, as well as many other medical complications [91]. Data from the 50 years of follow up study of male British doctors show that the strongest cause-specific associations are with respiratory cancers and chronic obstructive pulmonary disease; in numeric terms, the greatest health impacts of smoking are on respiratory and cardiovascular diseases. Smokers are also at greater risk of many other non-fatal diseases, including osteoporosis, periodontal disease, impotence, male infertility, and cataracts. Smoking in pregnancy is associated with increased rates of fetal and perinatal death and reduced birth weight for gestational age. Passive smoking after birth is associated with cot death and respiratory disease in childhood and lung cancer, heart disease, and stroke in adults [87].

The economic costs of tobacco use are equally devastating. Together, tobacco and poverty create a vicious circle. In most countries, tobacco use tends to be higher among the poor. Poor families, in turn, spend a larger proportion of their income on tobacco.

Money spent on tobacco cannot be spent on basic human needs such as food, shelter, education and health care. Tobacco can also worsen poverty among users and their families since tobacco users are at much higher risk of falling ill and dying prematurely of cancers, heart attacks, respiratory diseases or other tobacco-related diseases, depriving families of much-needed income and imposing additional costs for health care. Tobacco not only impoverishes those who use it, but also it puts an enormous financial burden on countries. The costs of tobacco use at the national level encompass increased health-care costs, lost productivity due to illness and early death, foreign exchange losses, and environmental damage [88].

1.2.1.3 Smoking is addictive

With the knowledge of the health and financial consequences and risks of smoking, more and more smokers are trying to give up, but only a small number of them are successful. Even the most carefully planned smoking cessation programs rarely have success rates higher than 10% to 20% after 1 year [92]. The reason why so many people fail to stop smoking is that they are addicted.

Addiction to smoking means a person has formed an uncontrollable dependence on cigarettes to the point where stopping smoking would cause severe emotional, mental, or physical reactions. Following the termination of cigarette use, addicted people will experience withdrawal symptoms. Physical symptoms include tingling in the hands and feet, sweating, intestinal disorders, headache, sore throats, coughing, and other signs of colds and respiratory problems as the lungs begin to clear. Psychological symptoms include insomnia, mental confusion, vagueness, irritability, and a build up of tension during periods of craving and depression. Nicotine is the addictive chemical in tobacco,

and is one of the most powerful addictive drugs. It is easily available and more socially accepted than other highly addictive substances.

1.2.1.4 Smoking is a multifactorial complex behavioural trait

Smoking is a complex behavioural trait that is influenced by both genetic and environmental factors, as well as the interaction of both aspects. Environmental influences on the development of tobacco dependence are well documented, and it has been hypothesised that the overall rates of decline in smoking is a response to social pressures. The biological factors have been little studied, and genetic influences may be particularly strong within the remaining population of smokers[93, 94].

Understanding the genetics of smoking presents a major challenge. In addition to being a complex trait with important environmental influences, smoking behaviour includes a number of stages such as initiation, experimentation, regular use, dependence, cessation and relapse [95]. Different genetic factors might work on these different stages. Multiple genes might interact with each other and individually show a mild or moderate effect. Therefore, a study of the genetics of smoking needs careful consideration as to possible approaches.

1.2.2 Approaches to genetic studies of smoking behaviour

There are several approaches that can be used for genetic studies into cigarette smoking and other addictive behaviours: 1) heritability studies, such as twin, adoption and family studies, can demonstrate a genetic influence in smoking behaviour; 2) genome-wide linkage analysis can be used to locate areas in the genome that segregate with for a specific phenotype such as nicotine dependence; 3) Candidate-gene association studies focusing primarily on targets identified in linkage studies and on genes selected for their

biological relevance [95]; 4) genome-wide association studies to obtain a global overview [96].

1.2.2.1 Twin studies

Among many other smoking heritablility studies, twin studies have been the most commonly used approach to explore genetic and environmental factors that result in the development of dependence and addiction. If genes significantly influence the risk for addiction, identical (monozygotic) twins will tend to be alike (concordant) in their response to cigarette smoking. On the other hand, fraternal twin pairs who only have 50 percent of their genes in common will be less concordant. By comparing the degree of concordance in monozygotic and dizygotic twin pairs, the extent to which genes influence vulnerability to cigarettes can be estimated.

Fisher (1958) first reported that concordance for smoking was significantly higher in monozygotic than dizygotic twin pairs among the German population [97]. These observations have been replicated many times over the years (Table 1.1). These twin studies have demonstrated that genetic factors not only contribute to the initiation of smoking, but also influence the age of the onset of smoking and the number of cigarettes smoked per day, as well as the persistence of smoking and intensity of smoking. Table 1.1 summarises some of the twin studies on genetic effects of smoking behaviour during the last several decades. As can be seen, the estimates of heritability, i.e. the proportion of the variance that is thought to be genetic, ranges from 0.31-0.86 but in many studies is over 0.5 and thus strongly supports the hypothesis that genetic factors play a significant role with regard to smoking behaviour.

Table 1.1 Twin studies of smoking behaviour in relation to genetic effects

	Authors and date	Sample	Phenotype	Heritability (%)	Reference
1	Kaprio et al., 1984	307 twin pairs	Years smoked	0.62	[98]
			Number of cigarettes	0.54	
			Pack years	0.62	
2	Carmelli et al., 1990	4960 male twin pairs, US	Number of cigarettes	0.53 (unadjusted); 0.35 (adjusted)	[99]
3	Swan et al., 1990	360 male twin pairs, US (52-66y)	Number of cigarettes	0.52	[100]
4	Carmelli et al., 1992	1882 male twin pairs, US	Current cigarette smoking	Yes (moderate)	[101]
		56-66 years old	Smoking persistence	Yes (moderate)	
			Current cigar smoking	Yes (moderate)	
5	Kendler et al., 1993	1566 female twins (avg age 31)	Lifetime smoking	Yes	[102]
6	Boomsma et al., 1994	1600 Dutch twin pairs (13-22 y)	Lifetime smoking	0.31	[103]
7	Edwards et al., 1995	434 female twin pairs, US	smoking initiation	0.77	[104]
		Average age 42.3	smoking maintenance	Yes	
8	Heath & Madden 1995	12898 twin pairs,	Current smoking	Yes	[105]
		Ages 16 - 47	Smoking initiation	Yes	
			Smoking persistence	Yes	
9	Swan et al., 1996	356 male twin pairs, Caucasian (52-66 y)	Number of cigarettes	0.56	[106]
10	Swan et al., 1997	4593 male twin pairs, US	Heavy smoking	0.49	[107]
		Average age 62.1	>30 cigarettes per day		
11	Hettema et al., 1999	1583 twin pairs,	Number of cigarettes - men	0.51	[108]
		Average age 30.3	Number of cigarettes – women	0.49	
12	Kendler et al., 1999	851 female twin pairs,	Smoking initiation	0.85	[109]
		Average age 30.1	Smoking dependence	0.72	
13	Koopmans et al., 1999	1676 Dutch twin pairs	Smoking initiation	0.39	[110]
		Adolescent	Smoking dependence	0.86	
14	Maes et al., 1999	1412 Virginia twin pairs	lifetime smoking	0.84	[111]
		Ages 8-16	current tobacco use	0.82	
15	Kendler et al., 2000	778 twin pairs, Swedish	Regular smoking - men	0.61	[112]
		Mean age 58.6	Regular smoking - women	0.63 (born after 1940)	
16	Madden et al., 2004	Australia 2284 pairs; Finland 10,948 pairs;	lifetime smoking	Women - 0.46	[113]
		Sweden 8651 pairs (18-46)		Men - 0.57	
17	Vink et al., 2004	Dutch 3657 twin pairs	smoking initiation	0.36	[114]
		Average age 24.7	Max number of cigarettes	0.51	
18	Broms et al., 2006	Finnish twin cohort, 2923 monozygotic and 6018 dizygotic twin pairs aged 24 to 88 years	Age at initiation	Men- 0.59 Women - 0.36	[115]
			Amount smoked	Men- 0.54 Women- 0.61	
			Smoking cessation	Men- 0.58 Women- 0.60	

1.2.2.2 Genome-wide linkage studies

Although twin studies have demonstrated that genetic factors play a significant role in determining smoking behaviour, little information is available about the chromosomal locations harbouring susceptibility loci/genes for nicotine dependence [116] [95].

There have been several genomewide linkage scans for smoking and related phenotypes. Straub *et al.* (1999) reported results of a 451-marker genomewide scan on the Christchurch sample of New Zealand (130 families with 343 genotyped individuals) and the Richmond sample of Virginia (91 families with 264 genotyped individuals). Phenotypes were assessed by means of score on the Fagerstrom Tolerance Questionnaire [117]. The highest observed pairwise LOD score in that sample was 2.63, on chromosome 2 [118]. Using the Collaborative Study on the Genetics of Alcoholism data, Bergen *et al.* (1999) performed a sibling-pair linkage analysis of two smoking-related traits, ever/never smokers and pack-years. There was some evidence of linkage of the ever/never smoking trait to regions on chromosome 6, 9, and 14[119]. Similarly, in a family based pedigree study, Duggirala *et al.* (1999) found strong evidence for linkage of smoking behaviour to a genetic location on chromosome 5q and weaker evidence for linkage of smoking behaviour to locations on chromosome 4, 15, and 17[120]. Li *et al.* (2003) reported results on the Framingham Heart Study data, using both GENEHUNTER and S.A.G.E. program, and provided evidence for significant linkage of smoking rate to chromosome 11 and suggestive linkage to chromosomes 4, 7, 9, 14 and 17[121]. Bierut *et al.* (2004) based on the entire COGA sample, identified regions on chromosomes 5, 9, 11, and 21 with increased allele sharing among sib pairs for habitual smoking. Modest evidence of linkage was also found on chromosomes 1, 2, 11, and 15 for a combined phenotype of habitual smoking and alcohol dependence [122]. Vink *et al.* (2004) performed linkage analyses on Dutch twin pairs of 536

dizygotic twins and siblings from 192 families, suggested quantitative trait loci (QTL) on chromosome 6 and 14 for smoking initiation, on chromosome 3 for number of cigarettes smoked. Linkage was also found on chromosome 10 for both smoking initiation and number of cigarettes smoked which may partly explain the overlapping aetiological factors for smoking initiation and number of cigarettes smoked[114]. Gelernter *et al.* (2004) worked on a set of families originally identified because they segregate panic disorders, and supported the significant linkage of chromosome 9 and 11 for increasing risk of cigarette smoking. They also identified evidence supporting linkage to regions of chromosome 14, 16 and X[123]. More recently, several genome scans have been performed with regard to smoking initiation, cigarette consumption, and smoking persistence. Numerous loci have been implicated, but few reached statistical significance and few were replicated in independent studies [124-126]. A major hurdle in genome scans for complex disease is the presence of genetic heterogeneity and multiple interacting genes each with moderate effects. The size of the genetic effect, the density of the markers, the subject size and composition, the narrowness of the phenotype, the statistical approaches used, and the number of genes involved will alter the likelihood of detecting a significant signal [95]. All available linkage results are summarised in Table 1.2.

Table 1.2 Overview of positive results of linkage studies to smoking behaviour

Study	Sample	Number of markers	Phenotype	Chromosome location	Linkage (LOD)/ or p-value[5]	Reference
Straub et al., 1999	Christchurch[1]	451	Nicotine dependence	2	2.63/p=0.004	[118]
				10	1.67/p=0.008	
Bergen et al., 1999	COGA[2]	285	Ever smoked	6	3	[119]
				9	3	
				14	3	
Duggirala et al., 1999	COGA	296	Ever smoked	3	1.71	[120]

			(number of cigarette	4	2.17	
			packs per day for 1 year)	5	3.2	
				6	1.1	
				9	1.14	
				15	1.97	
				17	2.88	
Goode et al., 2003	FHS[3]	401	Max no. of cigarettes /day	2	1.98	[127]
				17	1.52	
				20	1.3	
Li et al., 2003	FHS	401	Smoking rate	11	p=0.000001	[121]
				4	p=0.0017	
				7	p=0.0087	
				9	p=0.00057	
				14	p=0.0011	
				17	p=0.0043	
Saccone et al., 2003	FHS	401	Cigarettes per day	5	1.1	[128]
				9	1.67	
				13	1.28	
				14	1.29	
				22	1.04	
Vink et al., 2004	Dutch twin pairs	379	Smoking initiation	6	3.05	[114]
				14	1.66	
				10	1.92	
			Number of cigarettes smoked	3	1.98	
				10	2.29	
Sullivan et al., 2004	Christchurch	476	Nicotine dependence	2	2.53	[129]
				10	2.02	
Gelernter et al., 2004	PD segregation	416	Habitual smoking	9	3.04	[123]
	families[4]			11	3.43	
	(Panic disorder)			7	2.25	
				14	2.68	
				16	2.63	
				X	2.78	
Bierut et al., 2004	COGA	296	Heavy smoking	5	1.12	[122]
				11	1.64	
			Ever smoked	9	1.51	
				21	1.51	
Wang et al., 2005	FHS	401	Cigarettes per day	1	p=0.0001	[130]
				3	p=0.0302	
				4	p=0.002	
				7	p=0.037	

				8	p=0.0171	
				9	p=0.0419	
				11	p=0.0281	
				16	p=0.0185	
				17	p=0.0251	
				20	p=0.0321	
Vink et al., 2006	Dutch sibling pairs	379	Age of smoking initiation	5	2.09	[124]
				6	1.75	
				7	1.77	
				10	1.84	
				14	2.3	
				22	2.5	
Morley et al., 2006	Australian Twin Registry	1376	Age of smoking initiation	16	p=0.00423	[126]
				20	p=0.00359	
			Cigarette consumption	4	p=0.00982	
				6	p=0.00774	
				11	p=0.00399	
Ehlers et al., 2006	Native Americans	791	Persistant tobacco use (smoked at 10 cigs/day for more than 1 year)	1	1.3	[125]
				4	1.5	
			Regular tobacco usage (smoked daily for at least 1 month)	1	1.3	
				6	2	
				7	1.5	
				8	2	
				13	1.3	
				22	1.3	

1.Christchurch: Christchurch, New Zealand.

2. COGA: Cooperative Study of the Genetics of Alcoholism.

3. FHS: Framinham Heart Study.

4. PD: Panic Disorder.

5. Some publications listed LOD scores, some listed p-values, only Straub et al. paper listed both.

1.2.2.3. Genome-wide association study

In Jan. 2007 a genome-wide association study was published on smoking behaviour that was conducted in parallel with a candidate genes study using the same case-control samples [96, 131]. Bierut et al. perfomed a comprehensive genome wide association study using a two-stage design. In the first stage, genotyping of over 2.4 million SNPs was completed in a portion of case and control pools. In the second stage, 31,960 SNPs selected from the first stage analysis were individually genotyped in 1050 case and 879 controls. A total number of 35 SNPs with p-values less than 10^{-4} were identified, but none of the individual findings were statistically significant after correcting for multiple tests. Nevertheless, several novel smoking related genes, such as Neurexin 1 (*NRXN1*), were suggested through this study, while a known candidate gene, the β3 nicotinic cholinergic receptor, was also identified [96]. This study was combined with a focused study on genes already suspected of playing a role in addiction. Saccone et al. targeted 348 candidate genes and analyzed 3713 SNPs in the same case-control samples. Among the genes genotyped, 52 genes were chosen from all soursecs, while the remaining 296 genes were selected from the companion genome wide association study. After correcting for multiple testing by controlling the false discovery rate, several cholinergic nicotinic receptor genes dominated the top signals. The strongest association was from an SNP represnting *CHRNB3*, the β3 nicotinic receptor subunit gene ($p=9.4 \times 10^{-5}$) [131].

1.2.2.4 Candidate gene studies

Based on the biological relevance, there are a number of plausible candidate genes for nicotine dependence, primarily in two main areas: nicotine metabolic enzymes and relevant neurotransmitter pathways. The first group of candidate genes are those that

regulate nicotine metabolism. Individuals with variants of these genes may differ in their susceptibility to the actions of nicotine at different levels of the drug. Another promising group comprises genes related to the dopaminergic reward system, such as those encoding tyrosine hydroxylase (dopamine synthesis), dopamine transporters (dopamine re-uptake), dopamine receptors (receptor activation), and monoamine oxidases A and B (dopamine metabolism). Additional genes of interest include serotonin transporters and serotonin receptors.

Candidate gene analysis is the approach used in this thesis, and will be discussed in more detail.

1.2.3 The nicotine metabolism pathway

It has been proposed that genetic polymorphisms in nicotine metabolism may be a major determinant of an individual's smoking behaviour [132]. This hypothesis is based on the premise that the smoking behaviour is primarily aimed at maintaining peripheral and central nicotine levels [133, 134]. Hence, nicotine is the compound in tobacco primarily responsible for the development and maintenance of tobacco dependence[135].

1.2.3.1 Nicotine and its metabolism

Nicotine ($C_{10}H_{14}N_2$) is a naturally occurring alkaloid derived exclusively from the leaves of tobacco plant, *nicotiana tabacum.* Nicotine readily diffuses through mucous membranes in the respiratory tract and through the skin into the bloodstream, and almost immediately into the brain where it activates specific types of receptors known as cholinergic receptors. These receptors are found in the brain, autonomic nervous system and neuromuscular junctions [136] and are normally activated by the

neurotransmitter acetylcholine. Many activities, including respiration, maintenance of heart rate, memory, alertness, and muscle movement are influenced by acetylcholine and its receptors. Nicotine can also activate cholinergic receptors because of its similar structure to acetylcholine, but the body does not regulate it. Neurons typically release small amounts of acetylcholine in a regulated manner. Nicotine, however, activates cholinergic neurons leading to the increased release of acetylcholine causing heightened activity in cholinergic pathways throughout the brain. The stimulation of cholinergic neurons promotes the release of the neurotransmitter dopamine in the reward pathways of the brain bringing on feelings of pleasure and reward, and then addiction. Nicotine also causes the release of glutamate, a neurotransmitter involved in learning and memory; this may create a memory loop of good feelings of nicotine use. Increased levels of endorphins, small proteins that are often called the body's natural painkillers, also lead to feelings of euphoria.

First ever puffs of tobacco are typically toxic, and can induce nausea, vomiting, headache, sweating, dizziness, tremor and palpitations in sensitive individuals. However, the nicotine effect on the dopamine pathway seems to represent the key first 'reward' step that induces addiction [137]. Nicotine has a half-life of average about 90 minutes. This leads people to smoke periodically throughout the day to dose themselves with nicotine. Regular nicotine use causes changes in both the number of cholinergic receptors and the sensitivity of these receptors to nicotine and acetylcholine. Some of these changes may be responsible for the development of tolerance to nicotine. Once tolerance has developed, a nicotine user need to use more and more nicotine to reach the same degree of stimulation or relaxation. A withdrawal syndrome that may last a month or more follows cessation of nicotine use.

Nicotine is extensively metabolised primarily in the liver but also to a small extent in the lung and brain. The major pathways of nicotine metabolism are summarised in Figure1.1 by Byrd *et al.* [138].

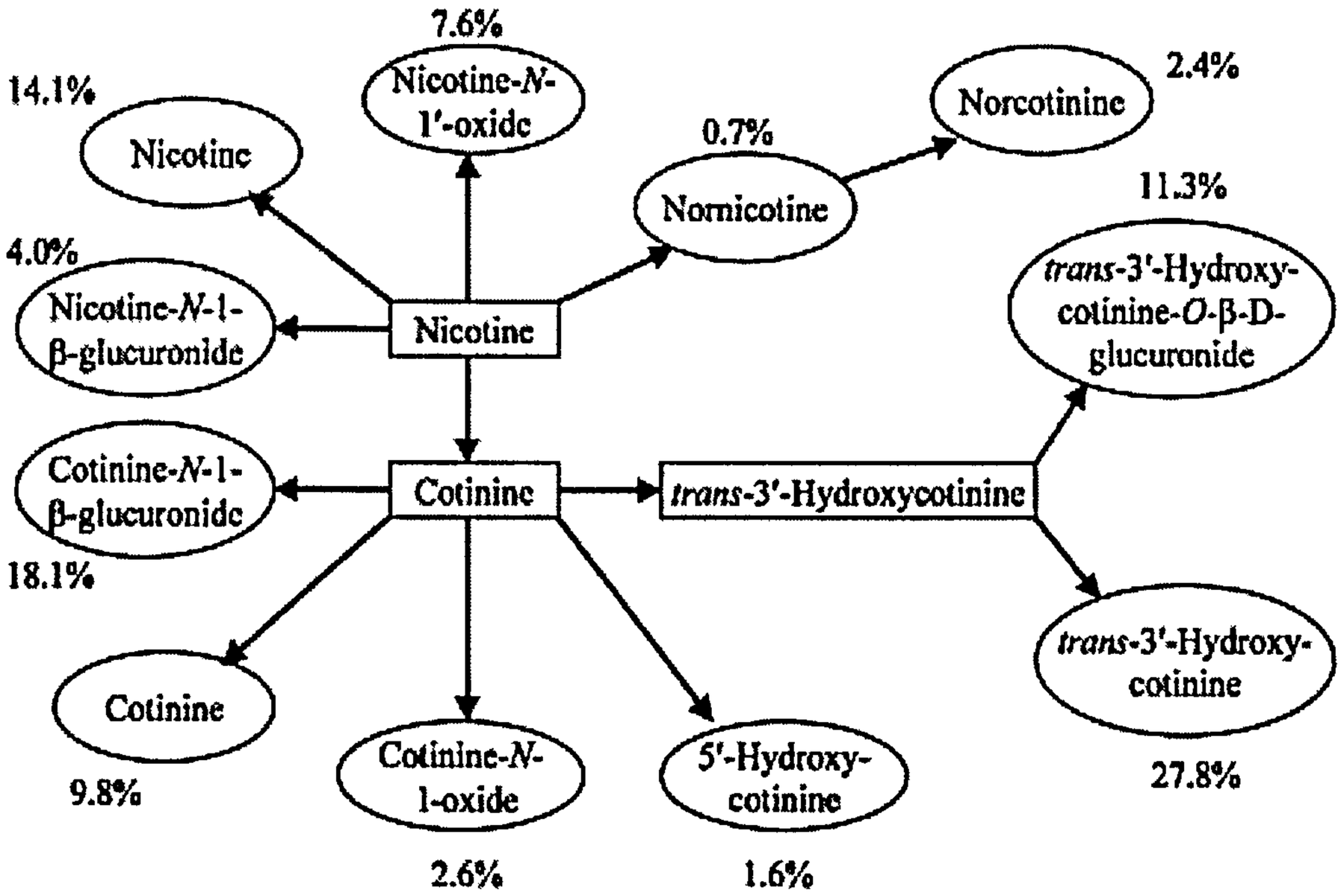


Figure 1.1 Quantitative scheme of nicotine metabolism, based on average excretion of metabolites as percentage of total urinary nicotine. Compounds in square boxes detected in blood, circled compounds indicate major metabolites excreted in urine.

About 70-80% of nicotine is metabolised to cotinine via C-oxidation, and to a lesser extent by N-oxidation to nicotine N-1'-oxide, N-demethylation, and N-glucuronidation. Cotinine is further metabolised by hydroxylation to trans-3'-hydroxycotinine and 5'-hydroxycotinine, N-oxidation to cotinine N-1-oxide, and N-glucuronidation. Trans-3'-hydroxycotinine is further metabolised by O-glucuronidation, and several additional minor metabolites have also been identified which probably account for less than 10% of total nicotine metabolism [132]. Both nicotine and main metabolite cotinine are biologically active compound influencing e.g. catecholamine and eicosanoid systems [139]. Cotinine is less active than nicotine, but it has a much longer half-life than nicotine (14-20 hours vs 90 minutes), and consequently is used as a marker of nicotine

intake [140, 141]. Several cytochrome P450 enzymes, as well as flavin mono-oxygenases, are purported to play a role in nicotine metabolism, but CYP2A6 appears to be the principal enzyme involved in converting nicotine to cotinine, and further oxidation of cotinine to trans-3'-hydroxycotinine.

The *CYP2A6* gene is highly polymorphic. Some variants have been shown to result in altered CYP2A6 enzyme activity. Many studies have demonstrated that there is considerable interindividual and interethnic variations in CYP2A6 activities and in the rate of nicotine metabolism that is due predominantly to genetic variation in the *CYP2A6* gene locus [142-149].

1.2.3.2 Cytochrome P450 2A6 (CYP2A6)

CYP2A6 is a member of the superfamily cytochrome P450, haem-containing isoenzymes located on the membrane of the smooth endoplasmic reticulum, which are important for the metabolism of endogenous substances (steroids, hormones, prostaglandins, lipids, and fatty acids) and in detoxification of exogenous compounds (environmental pollutants and dietary chemicals). The name cytochrome P450 comes from the fact that these enzymes are characterised by a maximum absorption wavelength of light 450nm in the reduced state in the presence of carbon monoxide [150]. All CYP isoenzymes in the same family have at least 40% structural similarity, and those in the same subfamily have at least 55% structural similarity.

CYP2A6 is expressed primarily in liver, but also in nasal epithelium and in lung and other extrahepatic tissues (e.g. kidney, brain) [132]. Studies show that hepatic CYP2A6 activity is not affected by smoking [151, 152], but nicotine metabolism is decreased in smokers compared with non-smokers [153], although the mechanism responsible for the slower metabolism is still unknown.

The *CYP2A6* gene (encoding the CYP2A6 enzyme) spans a region of 6,897 basepairs, contains 9 exons, 8 introns, and has been mapped to the long arm of chromosome 19 (19 q13.2). It is located within a 350-kilobase pair gene cluster together with the *CYP2A7* and *CYP2A13* genes, two *CYP2A7* pseudogenes, as well as genes in the *CYP2B* and *CYP2F* subfamilies [154]. The *CYP2A7* gene is located approximately 25kb upstream of the *CYP2A6* gene, and both genes have a 96.5% similarity in the coding nucleotide sequence [155-157]. (see Figure1.2).Interindividual variability in *CYP2A6* expression can be attributable to genetic polymorphisms in the *CYP2A6* gene; some of them are created by unequal crossover events and/or gene conversions between *CYP2A6* gene and *CYP2A7* gene [158] [159] [160, 161].

The wild type of the *CYP2A6* was termed *CYP2A6*1A*. Two variant alleles (*CYP2A6*2* and *CYP2A6*3*) were described in several studies [154, 157, 162, 163]. The *CYP2A6*2* allele encoded an enzyme with a L160H substitution, which was inactive because it did not incorporate haem [157]. The *CYP2A6*3* allele was initially reported as a hybrid allele created through multiple gene conversions with the inactive *CYP2A7* gene [154] resulting in an allele where exons 3, 6, 8 are of *CYP2A7* origin. This allele was also considered inactive, but many studies failed to detect any true *CYP2A6*3* alleles. Gu *et al.* (2000) used a long PCR-nested method in a bank of 1032 subjects but did not find any cases with *CYP2A6*3* identified by Fernandez-Salguero [164]. *CYP2A6*3* was not detected in many other studies both in Caucasian and Asian populations [146, 147, 165-170]. It was subsequently reported that many individuals had mutations at the primer-binding site and lead to no amplification of one of the alleles using Fernandez-Salguero's genotyping method, leading to false reporting of the *CYP2A6*3* allele [163].

The *CYP2A6* gene deletion variants, known collectively as the *CYP2A6*4* alleles are thought to have resulted from unequal crossover events between *CYP2A7* and *CYP2A6*

[146, 160, 168, 171-173]. Multiple crossover positions have been identified in the highly homologous regions of *CYP2A6* and *CYP2A7* (intron 8 – 3' untranslated region) creating different deletion variants (A-D), and a hypothetical allele containing two *CYP2A6* genes in tandem could also exist [161]. Rao *et al* identified a duplicated form of *CYP2A6*, *CYP2A6*1X2,* corresponding to the *CYP2A6*4* deletion allele [168]. Fukami et al. (2007) reported a novel duplication type of *CYP2A6* gene in African-American population that is created through an unequal crossover event with the *CYP2A7* gene at 5.2-5.6kb downstream from the stop codon [155].

The *CYP2A6*1B* and *CYP2A6*5* alleles were reported by Oscarson and colleagues [167]. The *CYP2A6*1B* allele had a 58bp gene conversion with *CYP2A7* in the 3'-untranslated region [167, 171]. The *CYP2A6*5* allele had a point mutation in exon 9 leading to a single amino acid substitution (G479V), as well as gene conversion in the 3'-untranslated region. It has been reported that *CYP2A6*5* encodes an unstable enzyme [167].

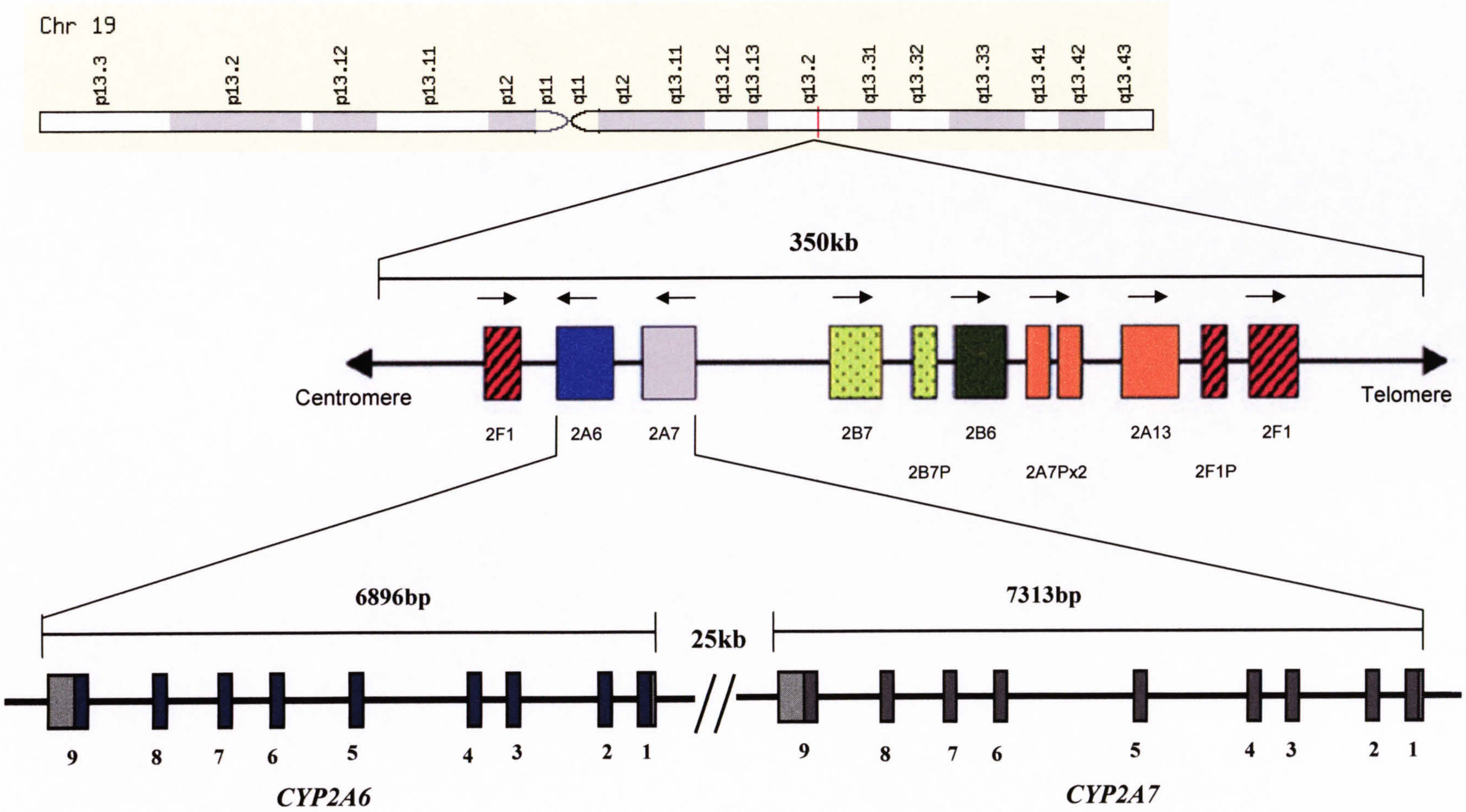


Figure 1. 2 Schematic diagram of the position and structure of *CYP2A6* gene. The 350kb *CYP2A-2B-2F* gene cluster is shown at chromosome 19, q13.2. Arrows indicate the direction of transcription. P denotes a pseudogene. Blue and grey hatched boxes represent exons of *CYP2A6* and *CYP2A7*, respectively. Solid lines between the boxes are introns. White hatched boxes indicate 5' and 3' untranslated regions.

A novel polymorphism of *CYP2A6*, *CYP2A6*6*, was detected in 0.4% of the Japanese population, by using the PCR technique combined with restriction enzyme fragment length polymorphism (PCR-RFLP). The variant was found to have a single nucleotide mutation in exon 3, resulting in an amino acid substitution of arginine (R) 128 by glutamine (Q) [174].

Two novel mutations in exon 9 of *CYP2A6* gene (*CYP2A6*7* and *CYP2A6*8*) were reported by Ariyoshi *et al.* and Xu *et al.* [171, 175]. *CYP2A6*7* had a single nucleotide polymorphism of T1412C resulting in Ile471Thr substitution. *CYP2A6*8* had a single nucleotide polymorphism of G1454T resulting in Arg485Leu substitution. Ariyoshi *et al.* also found that 38.1% of the *CYP2A6*1B* allele possessed the T1412C mutation concomitantly. Combining the *CYP2A6*1B* allele frequency of 41.3% with 38.1% of the C variant allele in Japanese, they calculated the *CYP2A6*7* allele frequency to be 15.7% [171].

A single nucleotide polymorphism, a –48T>G substitution in the TATA box of the 5' flanking region of the *CYP2A6* was identified by Pitarque *et al.* The new allele, named *CYP2A6*9*, has been shown to be common in Caucasian and Oriental populations and affects gene expression. The activity of the TATA-mutated promoter was about 50% of that of the wild type [176]

Xu *et al.*, 2002 described *CYP2A6*10*, the combination of *CYP2A6*7* and *CYP2A6*8* substitutions on the same allele [175]. They reported that the combination allele appeared to decrease activity of the enzyme more than either allelic substitution alone. Daigo *et al.*, found a single nucleotide polymorphism, *CYP2A6*11*, in which thymine at nucleotide 670 was changed to cytosine and caused an amino acid substitution from serine to proline at residue 224 [177], resulting in decreased enzyme activity.

Oscarson and his colleagues reported a new *CYP2A6* allele (*CYP2A6*12*) which carried an unequal crossover between the *CYP2A6* and *CYP2A7* genes in intron 2. This resulted in a hybrid allele where the 5' regulatory region and exons 1-2 were of *CYP2A7* origin and exons 3-9 were of *CYP2A6* origin, resulting in 10 amino acid substitutions compared to the *CYP2A6*1* allele. It caused reduced *CYP2A6* activity in vivo [173].

In addition, Kiyotani *et al.* (2002) reported 21 novel SNPs in Japanese and Caucasians [178]. Recently, more SNPs have been reported [172, 179-185]. Figure 1.3 and Table 1.3 summarise some major variants of *CYP2A6*. See www.imm.ki.se/CYPalleles/ for a complete update of *CYP2A6* alleles.

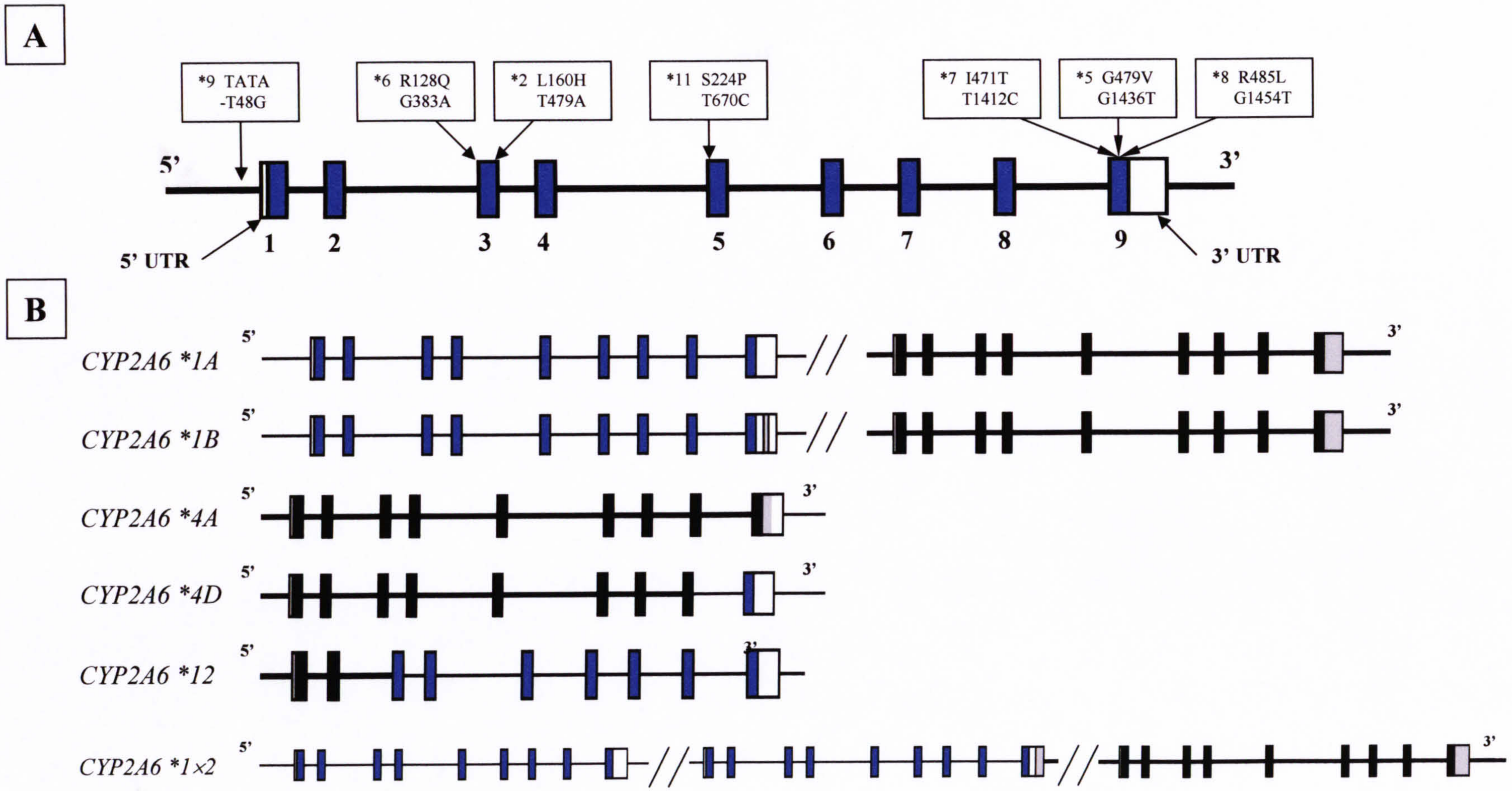


Figure 1. 4 Schematic diagram of the *CYP2A6* gene structure and variants. (A) *CYP2A6* point mutation variants. Exons are represented by blue boxes (▮). 5'UTR and 3'UTR indicate 5' and 3' untranslated region, represented by open boxes (□). Solid lines (___) are introns. *CYP2A6*10* consists of the combination of *CYP2A6*7* and *CYP2A6*8*. The *CYP2A6*3* allele is not shown since it is either extremely rare or is an artifact. (B) Schematic diagram of the variants that have been created through crossover events. Black and blue boxes represent exons of *CYP2A7* and *CYP2A6*, respectively. 5' and 3' untranslated regions are expressed using grey (*CYP2A7*) and open boxes (*CYP2A6*). Solid lines (___) are introns of *CYP2A7*; light lines (___) are introns of *CYP2A6*. See www.imm.kise/CYPalleles/ for a complete update of *CYP2A6* alleles.

Table 1.3 Current *CYP2A6* allele nomenclature

Allele	Nucleotide changes	Effect	Enzyme activity		References
			In vivo	In vitro	
*CYP2A6*1A*	None		Normal	Normal	Yamano et al, 1990
*CYP2A6*1B*	gene conversion in the 3' flanking region				Oscarson et al, 1999b Ariyoshi et al, 2000
*CYP2A6*1C*	-395G>A; 51A>G; gene conversion in the 3' flanking region				Kiyotani et al, 2002
*CYP2A6*1X2*		*CYP2A6* gene duplication			Rao et al, 2000
*CYP2A6*2*	479T>A	L160H	None	None	Yamano et al, 1990 Hadidi et al, 1997 Oscarson et al, 1998
*CYP2A6*3*	*CYP2A6/CYP2A7* hybrid		?	?	Fernandez-Salguero et al, 1995
*CYP2A6*4A*	*CYP2A6* deleted	*CYP2A6* deleted	None		Oscarson et al, 1999a Nunoya et al, 1999a Nunoya et al, 1999b Ariyoshi et al, 2000
*CYP2A6*4B*	*CYP2A6* deleted	*CYP2A6* deleted	None		Nunoya et al, 1998
*CYP2A6*4C*					Identical to the *CYP2A6*4A allele*
*CYP2A6*4D*	*CYP2A6* deleted	*CYP2A6* deleted	None		Oscarson et al, 1999b
*CYP2A6*5*	1436G>T	G479V	None	None	Oscarson et al, 1999b
*CYP2A6*6*	383G>A	R128Q		Decreased	Kitagawa et al, 2001
*CYP2A6*7*	1412T>C; gene conversion in the 3'-flanking region	I471T	(Decreased)	Decreased	Ariyoshi et al, 2001 Xu et al, 2002
*CYP2A6*8*	1454G>T; gene conversion in the 3'-flanking region	R485L	Normal		Ariyoshi et al, 2001 Xu et al, 2002
*CYP2A6*9*	-48T>G	TATA box		Decreased	Pitarque et al, 2001
*CYP2A6*10*	1412T>C; 1454G>T; gene conversion in the 3'-flanking region	I471T; R485L	Decreased		Yoshida et al, 2002 Xu et al, 2002
*CYP2A6*11*	670T>C	S224P	Decreased	Decreased	Daigo et al, 2002
*CYP2A6*12*	exons 1-2 of *CYP2A7* origin; exons 3-9 of *CYP2A6* origin	10 aa substitutions	Decreased	Decreased	Oscarson et al, 2002
*CYP2A6*13*	-48T>G; 13G>A; 51A>G	G5R			Kiyotani et al, 2002
*CYP2A6*14*	86G>A	S29N			Kiyotani et al, 2002
*CYP2A6*15*	-48T>G; 22C>T; 51A>G; 580A>G	K194E			Kiyotani et al, 2002
*CYP2A6*16*	51A>G; 607C>T	R203S			Kiyotani et al, 2002

In the literature, people with at least one enzymatically inactive allele (e.g. *2, *4 and *5) or two alleles with decreased activity (e.g. *9, *12) are defined as 'very slow metabolisers' whereas people with one decreased acitivity allele are defined as 'slow metabolisers'.

1.2.3.3 *CYP2A6* polymorphism, nicotine metabolism and smoking behaviour

Smoking behaviour is the maintenance of a certain level of addictive nicotine in the body, which is controlled by nicotine intake (i.e. smoking) and removal (i.e. rate of nicotine metabolism). Based on the above evidence, many studies have tried to investigate whether any of the identified polymorphisms of *CYP2A6* are functionally associated with differences in nicotine metabolism and smoking behaviour.

In vivo phenotype studies using coumarin as a probe drug showed a pronounced interindividual variability in the expression of *CYP2A6* before the possible influence of *CYP2A6* polymorphism on nicotine metabolism was suggested [186-188]. Oscarson *et al.* (1998) reported a coumarin metabolic index in vivo of three individuals with *CYP2A6*1A/*2* lower than that of wildtype *CYP2A6*1/*1*; two individuals homozygous for the *CYP2A6*2* allele displayed no coumarin 7-hydroxylation activity (Finland cohort of 100 samples) [163]. Using liver samples of 39 Japanese and 43 Caucasians, Inoue *et al.* (2000) revealed that two individual Caucasians who were classified to be heterozygous *CYP2A6*1/*2* were low in catalytic activities towards coumarin 7-hydroxylation and nicotine C-oxidation by liver microsomes, compared with those of wildtype homozygotes [189]. The frequency of the *CYP2A6* deletion is significantly higher in Orientals than in Caucasians. Therefore, the associations between *CYP2A6* deletion and nicotine metabolism were investigated mainly in Asian populations [146, 147, 166, 175, 190, 191]. It was reported that *CYP2A6*4B/*4B* homozygotes were completely deficient in cotinine formation, while subjects with a *CYP2A6*4* allele had lower mean cotinine-nicotine ratios than subjects without a *CYP2A6*4* allele[147, 191]. Subjects with an allele of *CYP2A6*7 or CYP2A6*10* also showed different degrees of reduced enzymatic activity [175, 190].

There is a greater consistency of findings concerning genotypic effects on nicotine metabolism than for studies of smoking behaviour in adults. Pianezza *et al.* first claimed the association between *CYP2A6* genotype and self-reported smoking behaviour. It showed that among tobacco-dependent smokers, the frequency of individuals with impaired nicotine metabolism (carrier of 1 or 2 *CYP2A6*-null alleles) was lower than in the control group (13.2% versus 19.6%, $p<0.04$) [192]. Considering that the *CYP2A6* genotype was determined by the now discredited method of Fernandez-Salguero *et al.* 1995, the association between *CYP2A6* and smoking behaviour reported in this paper should be viewed with caution. Rao *et al.* and Tyndale *et al.* using higher resolution genotyping assays reported that carriers with an allele of *CYP2A6*2* or *CYP2A6*4* had significantly lower self-reported cigarette consumption (13.5 versus 19.5 cigarette per day, $p<0.03$) [91, 168]. A previous study by Gu, *et al.* examined self-reported smoking habit from a cohort of average age 60-70 years and showed that carriers of a *CYP2A6*2* allele were almost twice as likely to give up smoking during their lifetime, and likely to start smoking 3 years later, than were those with wild-type genotype [164]. Schoedel *et al.* (2004) demonstrated that slow nicotine inactivators (*CYP2A6*2*, *CYP2A6*4*, *CYP2A6*9*, *CYP2A6*12*) were less likely to be adult smokers (dependent or non-dependent). Slow inactivators also smoked fewer cigarettes per day and had an earlier age of first smoking (only dependent smokers) [193]. Some studies however did not show any association between self-reported cigarette smoking and *CYP2A6* genotype in either the Caucasian or the Asian population [169, 170, 194-198].

Inconsistent association between polymorphism of *CYP2A6* and smoking behaviour could be attributed to several factors. 1) One of the earlier genotyping methods was shown to give rise to false reporting of the *CYP2A6*3* allele [167]. 2) There were discrepancies in the assessment of smoking behaviour [132]. 3) The frequencies of

CYP2A6 alleles between Caucasians and Orientals were different. 4) It was shown that the frequencies of inactive alleles (e.g. *2, *4, or *5) were low (except *4 in Orientals) [132], therefore, conclusions were drawn from small sample size studies which are inevitably susceptible to statistical uncertainty.

1.2.3.4 Other genes influencing metabolism of nicotine (*CYP2D6* and *CYP2B6*)

CYP2D6 *CYP2D6,* mapped to chromosome 22q13.1[199], is also involved in oxidation of nicotine to cotinine. Variable expression and function of CYP2D6 leads to distinct 'ultra-rapid' (UR), 'extensive' (EM), 'intermediate' (IM), and 'poor metaboliser' (PM) phenotypes. Several studies have investigated this polymorphism in tobacco dependence with conflicting results. Cholerton *et al.* (1996) failed to find a difference in the *CYP2D6* genotype among smokers and nonsmokers [200]. However, the same group concluded that although *CYP2D6* status does not influence whether or not a person becomes a smoker, once an individual starts to smoke, subsequent smoking behaviour may be influenced by their *CYP2D6* genotype [201]. Further support for the involvement of this genotype in smoking comes from report that the prevalence of ultrarapid metabolisers among heavy smokers (defined as >20 cigarettes per day for at least 20 years) was fourfold compared to self-reported nonsmokers and twofold compared to smokers with variable smoking habits [202].

CYP2B6 *CYP2B6* maps to chromosome 19q12-13.2 [203], and is highly polymorphic [204]. It is thought that the low hepatic levels of constitutive CYP2B6 play only a relatively minor role in the metabolism of xenobiotics [205] including that of nicotine [206]. However, its effect on smoking cessation has been reported by Lerman *et al.* (2002)[207]. They examined 426 smokers enrolled in a randomised clinical trial of bupropion for smoking cessation and found that participants with a decreased activity

variant of the *CYP2B6* gene reported greater increases in cravings for cigarettes following the cessation date and were about one and a half times more likely to relapse during the treatment phase.

1.2.4 The dopamine reward pathway

1.2.4.1 Dopamine rewarding system

Addictive drugs stimulate a neural "reward pathway" in the limbic system of the brain, which controls basic emotions and behaviours and provides a physiological basis for the perception of pleasure [208]. The major structure of the reward pathway includes the ventral tegmental area (VTA), the nucleus accumbens and the prefrontal cortex. The VTA is connected to both the nucleus accumbens and the prefrontal cortex via this pathway and it sends information to these structures via its neurons. The neurons of the VTA contain the neurotransmitter dopamine that is released in the nucleus accumbens and in the prefrontal cortex.

This pathway is activated by a rewarding stimulus (positive reinforcement) either natural reinforcement like food, water, etc or by some unnatural reinforcement, like drugs. Nicotine is one of them. Figure 1.4 shows how it functions.

The neurotransmitter dopamine is synthesized in the nerve terminal and packaged in vesicles. As an electrical impulse arrives at the terminal, it triggers vesicles containing dopamine to move toward the

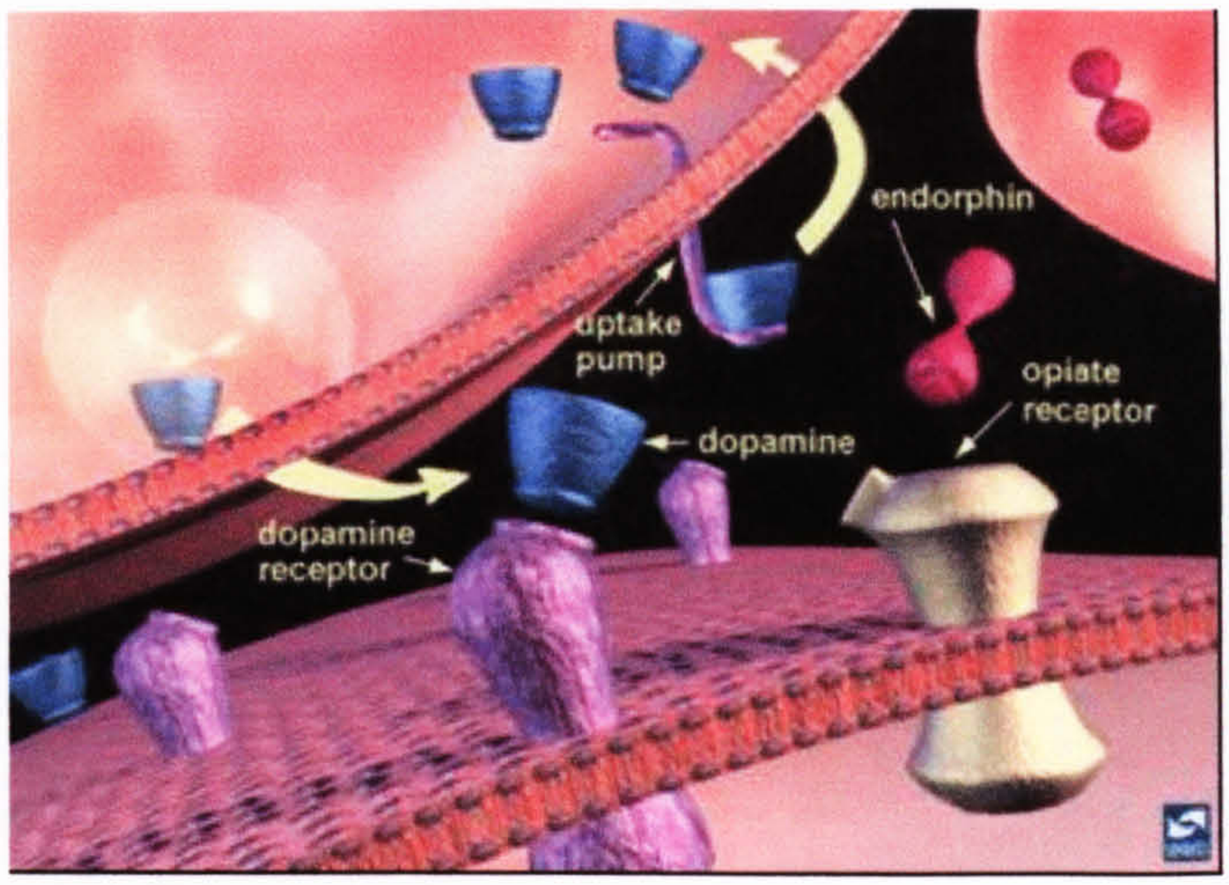


Figure 1.4 Schematic diagram of synaptic neurotransmission

terminal membrane. The vesicles fuse with the terminal membrane to release the dopamine. Once inside the synaptic cleft (the space between the two neurons) dopamine can bind to specific proteins called dopamine receptors (in purple) on the membrane of a neighbouring neuron. After binding, the dopamine comes off the receptor and is removed from the synaptic cleft by uptake pumps (dopamine transporter) that reside on the terminal. This process is important because it ensures that not too much dopamine remains in the synaptic cleft at any one time.

Nicotine binds to specific receptors (Nicotinic acetylcholine receptors) on the presynaptic neuron (see Figure 1.5). The opening of these receptors is usually triggered by the neurotransmitter acetylcholine, but they are also opened by nicotine [209, 210]. When nicotine binds to receptors at the cell body, it excites the neuron so that it fires more action potentials (electrical signals) that move toward the synapse, causing more dopamine release. Dopamine causes a feeling of pleasure, and as a positive reinforcement, leads to more cigarettes smoking (Figure 1.5).

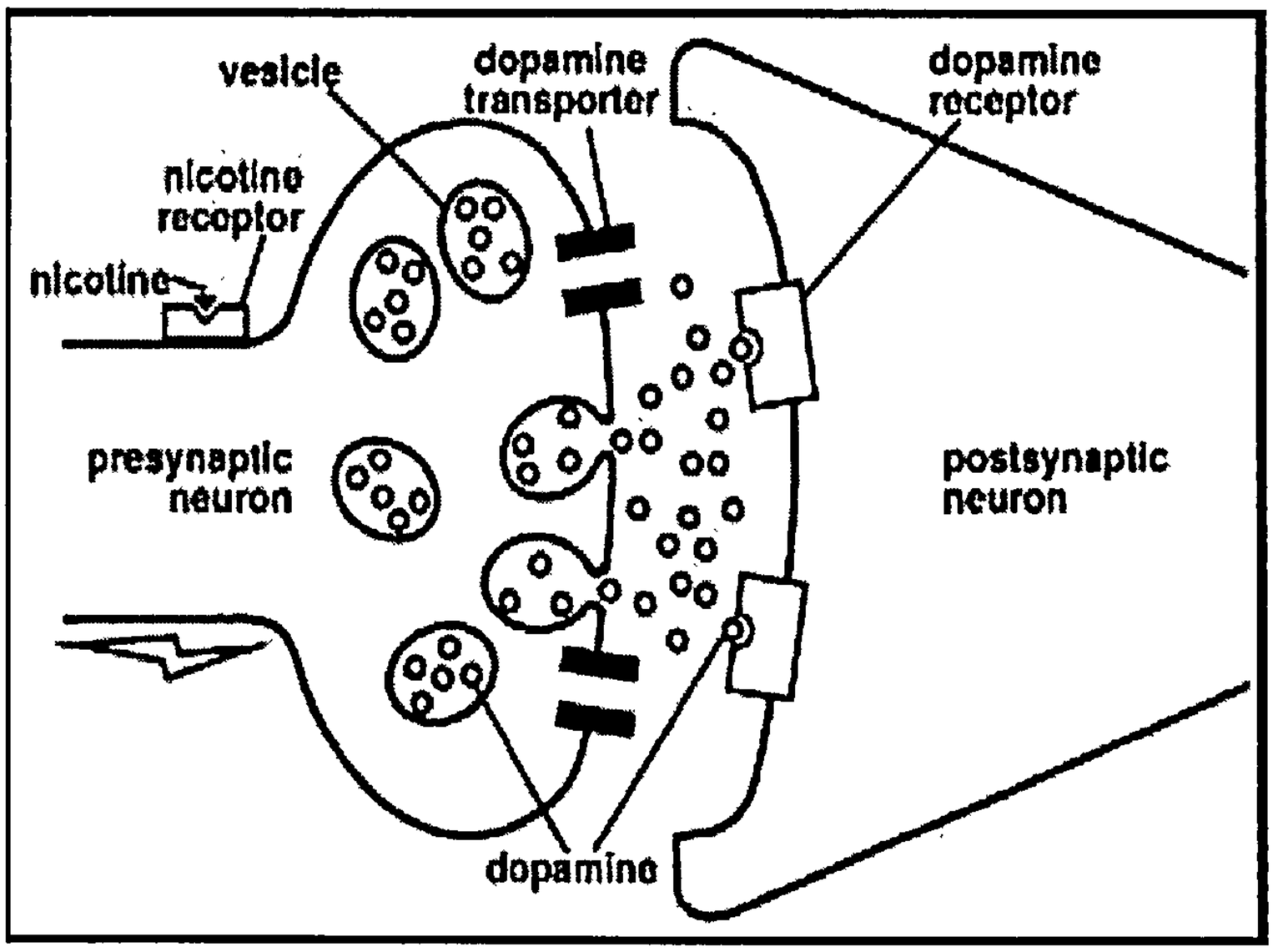


Figure 1.5 Schematic diagram of nicotine working mechanism

As discussed above, the important components of the dopaminergic pathway include 1) synthesis of dopamine in dopaminergic neurons; 2) release of dopamine by presynaptic neurons; 3) receptor activation of postsynaptic neurons; 4) dopamine re-uptake by the presynaptic neuron, and 5) metabolism of released dopamine. Many candidate genes involved in dopaminergic neurotransmission have been characterised and have identified polymorphisms, including those for tyrosine hydroxylase (*TH*), dopamine receptors (*DRD1-DRD5*), dopamine transporter (*SLC6A3*), dopamine ß-hydroxylase (*DBH*), catechol-*o*-methyltransferase (*COMT*), and monoamine oxidase (*MAO*) A and B. Figure 1.6 shows the dopamine synthesis and metabolism system, as well as the enzymes that catalyse each process. Genes encoding these enzymes have been investigated in relation to substance addiction, including cigarette smoking.

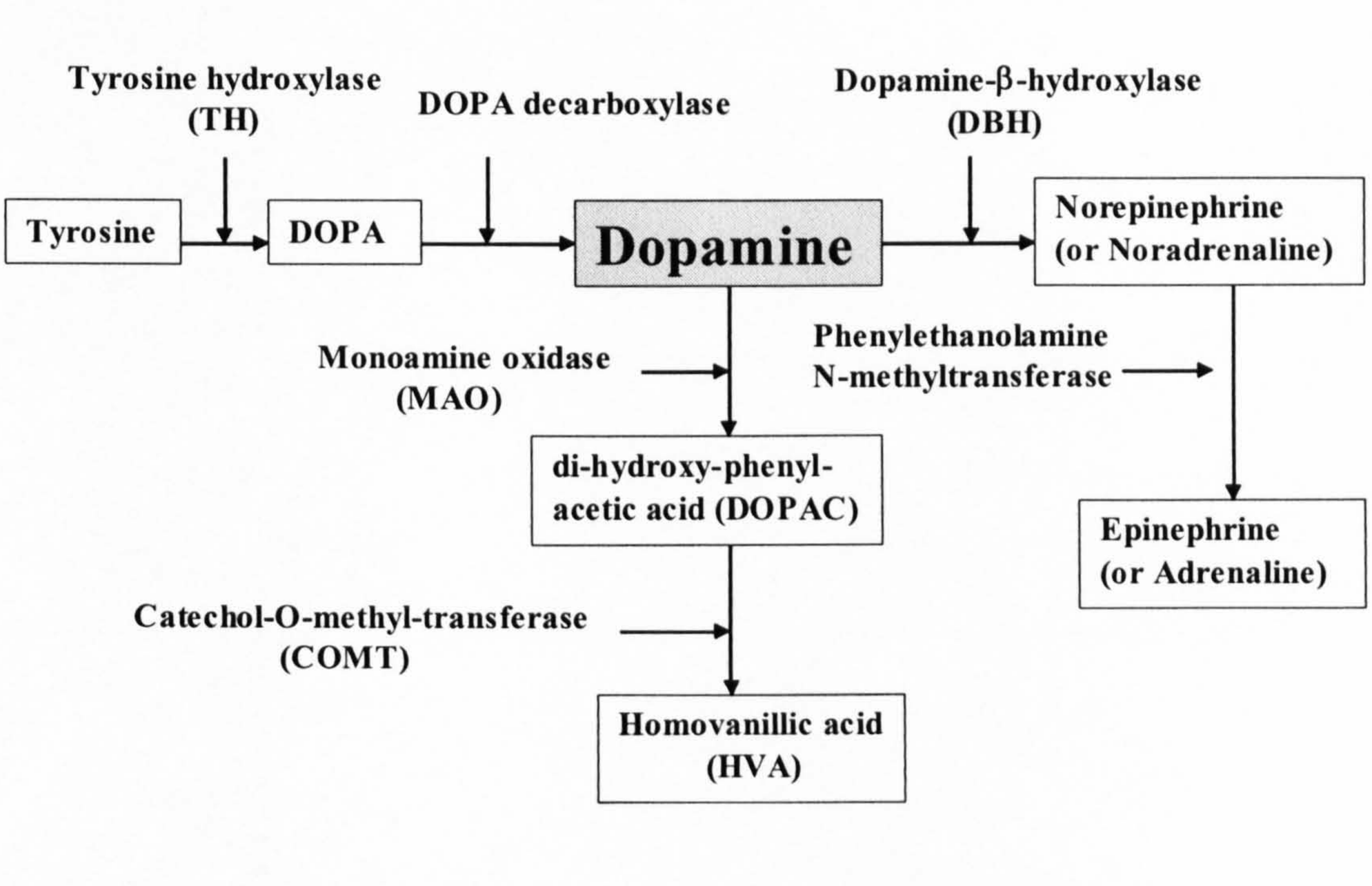


Figure 1.6 Schematic diagram of dopamine synthesis and metabolism

1.2.4.2 Dopamine synthesis

Tyrosine, an amino acid that is abundant in dietary proteins, is the starting point for biosynthesis of dopamine. Tyrosine enters dopaminergic neurons and is converted to dihydroxyphenylalanine (l-DOPA) by tyrosine hydroxylase (TH). TH is a rate-limiting enzyme. The *TH* gene, which encodes the TH enzyme, is located on chromosome 11p15.5, and a region near this gene has been implicated with habitual smoking in genome-wide linkage scans [123]. A tetranucleotide (TCAT) repeat polymorphism (*HUMTH01*) in intron one of the *TH* gene has been identified and used in several different studies. A protective relationship between the K4 (seven repeat) of *TH01* and high-level smoking has been reported [211, 212]. In our laboratory, data from Hertfordshire Cohort Study (HCS) and Northwick Park Heart Study II (NPHSII) show that haplotype *IGF2-INS-TH*5* (9 repeat of *TH01*) has a positive relationship to high-level smoking [213].

1.2.4.3 Dopamine metabolism

As shown in Figure 1.6, dopamine is converted into norepinephrine catalysed by dopamine β-hydroxylase (DBH) and further into epinephrine by phenylethanolamine N-methyltransferase. Dopamine is also broken down by MAO and a further reaction by COMT. All these enzymes have effect on the amount and activity of dopamine. Variations of genes encoding these enzymes have been reported to be related to the activity of dopamine and therefore may affect smoking behaviour.

Dopamine β-hydroxylase (DBH) The *DBH* gene is located on chromosome 9q34. High serum DBH enzyme activity has been associated with alcohol consumption [214]. It has been reported that a silent mutation G1368A in exon 9 was associated with smoking behaviour in a group of 225 smokers. More heavy smokers (> 20 a day) had the *DBH*

1368A allele when compared to light smokers (< 10 a day). (Relative risk 2.3, 95% CI 1.1, 5.0, P = 0.024). The trend for increasing prevalence of the *DBH* A allele in heavy smokers was greater when analysis was restricted to Caucasians (relative risk 3.2, 95% CI 1.3, 8.2, P = 0.004) [215]. Johnstone *et al.* (2004) studied 755 heavy smoker patch users and found that short-term effectiveness of the nicotine patch may be related to *DBH* and *DRD2* genotype. At 1 week, the nicotine patch was more effective for smokers with both *DRD2* CT/TT and *DBH* GA/AA genotypes which had an odds ratio (OR) of 3.6 (2.0-6.5) compared to 1.4 (1.0-2.1) for others (P = 0.01). At 12 weeks, the ORs for these genotypic groups were 3.6 (1.7-7.8) and 1.4 (0.9-2.3), respectively (P = 0.04) [216].

Monoamine oxidase (MAO) Genes encoding MAO A and B are closely linked and are located on the X chromosome. Abnormal MAO enzyme activity levels have been associated with a variety of conditions, including smoking, alcoholism, schizophrenia, and Parkinson's disease [217-220]. Fowler *et al.* reported that the brains of individuals who currently smoke show a 40% reduction in MAO B enzyme relative to never or former smokers [219], and that MAO A enzyme levels are lower in smokers than non-smokers [218]. In a cessation trial, heavy smokers treated for 3 months with moclobemide, a reversible inhibitor of MAO A enzyme, had superior abstinence rates for up to 6 months after giving up compared with heavy smokers treated with a placebo[221].

McKinney *et al.* (2000) reported that a SNP T1460C on exon 14 of *MAOA* gene was associated with cigarette smoking. The heavy smokers were less likely to possess the 1460C allele than were light smokers[215]. It was suggested that the SNP alone was not responsible for variations in phenotype. It is probable that these alleles are in linkage disequilibrium with alleles that cause functional changes in MAO activity. A possible

candidate for this is a VNTR with a 30bp repeat in the promoter region of this gene. The two common alleles are 3 and 4 repeats. Ito *et al.* (2003) reported that the median Fargastrom test score for nicotine dependence (FTND) among male current smokers was significantly higher with than without the 4-repeat allele (5.8 and 4.7, respectively). The adjusted odds ratio (aOR) was 2.72 (95% confidence interval 1.13-6.50) for males with the 4-repeat allele. Among females, the aOR of being current smokers compared to never smokers was 0.49 (0.26-0.93) for individuals with the 4-repeat allele [222].

***Catechol-O-methyltransferase* (*COMT*)** COMT is considered a key regulator of dopaminergic neurotransmission[223, 224]. A common biallelic functional polymorphism occurs within the *COMT* gene results in high and low activity gene products and a three- to four fold variation of enzyme activity. This is due to a G to A transition in the fourth exon of the gene, resulting in a valine to methionine substitution at codon 158 of the membrane-bound forms of *COMT* (corresponding to codon 108 of the soluble or cytoplasmic form)[225]. Homozygosity for the allele encoding methionine is associated with low enzyme activity and thermolability. Homozygosity for the allele encoding valine is associated with high activity and thermostability and heterozygous individuals have an intermediate level of COMT activity and heat stability[225]. Vandenbergh *et al.* (1997) reported that individuals with the high-activity COMT variant might have greater genetic vulnerability to drug abuse[226]. The association between this polymorphism and smoking behaviour was first investigated in a British study and no association was found in relation to cigarette consumption[215]. David *et al.* (2002) investigated the relationship between this polymorphism and smoking initiation, smoking persistence and smoking cessation among 266 current smokers, 270 ex-smokers and 265 lifetime non-smokers (never smokers), but yet again no association was found [227].

1.2.4.4. Dopamine receptors

Dopamine exerts its effects in the brain by acting upon a family of receptors. These receptors can be divided into two broad categories, D-1 type and D-2 type, which are defined primarily by their pharmacology although each group also has significant sequence homology. The D1 class of dopamine receptors contains the D1 and D5 receptors; the D2 class includes the D2, D3, and D4 receptors. Polymorphisms in several dopamine receptor genes have been detected and studied in cigarette smoking research.

Dopamine D_1 receptor gene (*DRD1*)

The *DRD1* gene is located on chromosome 5q35.2. Duggirala *et al.* (1999) tested for linkage between smoking behaviour (as defined by the number of cigarette packs per day for 1 year) and chromosomal locations and found strong evidence (LOD= 3.2) for linkage of smoking behaviour to a genetic location on chromosome 5q (D5S1354)[120]. Comings *et al.* (1997) investigated the relationship between smoking behaviour and *DRD1* polymorphisms, which used the *DdeI* polymorphism (−48A/G) located in the 5′ UTR, and found a significant increase in the frequency of homozygosity for either the *DRD1* 1- or 2-allele in smokers relative to non-smoking controls. Among smokers, homozygosity of *DRD1* was also associated with the number of cigarettes smoked per day. The authors also suggested an interaction of genetic variants at the *DRD1* and *DRD2* genes [228]

Dopamine D_2 receptor gene (*DRD2*)

The *DRD2* gene is mapped to chromosome 11q23.2. Various polymorphisms of *DRD2* have been identified and investigated for potential association with various

neuropsychiatric disorders [229]. One of the most thoroughly studied polymorphic sites is a TaqI A polymorphism located in the 3'-flanking region of the gene. Munafo *et al.*, (2004) summarised 41 studies of candidate smoking genes including 13 studies reporting data on the dopamine receptor D2 (*DRD2*) Taq1A polymorphism and smoking status [230]. Noble *et al.* first reported a significantly higher prevalence of the *DRD2* TaqI A1 allele among the current smokers (p=0.024) and ex-smokers (p=0.044) compared to non-smokers in Caucasians [231]. These findings were replicated in a study of non-Hispanic whites, in whom 48.7% of smokers carried the A1 allele, compared to only 25.9% of non-smokers having the A1 allele [232]. The study also showed a significant, inverse relationship between the prevalence of the D2A1 allele and the age of smoking onset, and the maximum duration of time the smokers had been able to refrain from smoking. Li *et al.*, (2004) performed a meta-analysis based on 12 published case-control studies and made it clear that there is a significantly higher prevalence of the TaqI A1 allele in the 2182 smokers than in the 2117 non-smokers ($p<0.0001$; pooled OR=1.47; 95% CI=1.31–1.66) [233].

Dopamine D_3 receptor gene (*DRD3*)

The *DRD3* gene maps to chromosome 3q13.3, and has rarely been studied in relation to smoking behaviour. Recently, Czermak *et al.*, (2004) reported a significant 30% reduction of *DRD3* mRNA expression in PBLs (peripheral blood lymphocytes) in smokers but not former smokers in comparison with controls. *DRD3* mRNA expression in PBLs in current smokers was negatively correlated with daily number of cigarettes consumed. The authors suggested a selective inhibiting effect of smoking on *DRD3* mRNA expression and proposed a vicious-circle explanation for the motivation for continued smoking [234].

Dopamine D_4 receptor gene (*DRD4*)

The *DRD4* gene is located at chromosome 11p15.5. The third exon of *DRD4* contains an unusually polymorphic 16-amino acid repeat, and physiologic differences in ligand binding have been observed between the short (2-5 repeats; most commonly 4) and long (6-10 repeats; most commonly 7) forms of this repeat. There have been a number of reports that this polymorphism is associated with variation in novelty seeking [235-237] or in substance abuse and addictive behaviours [238-241].

Smokers showed higher novelty-seeking rates than did the general population [242, 243], therefore it is likely that the L allele is associated with smoking. The association between smoking behaviour and the L allele has been reported in an African-American population. In African-Americans, the L allele carrier predicted increased smoking, a shorter time before the first cigarette in the morning and an earlier age of smoking initiation [244]. Shields *et al.* also found that, after smoking cessation counselling, none of the subjects with an L allele were abstinent at 2 months compared with 35% of the subjects who were homozygous for the S allele [244]. Hutchison *et al.*, (2002) suggested that participants in the L group demonstrated significantly greater craving, more arousal, less positive effects, and more attention to the smoking cues than did the participants in the S group [239]. Lerman *et al.*, 1998 also found a significant interaction between the *DRD4* exon III VNTR genotypes and depression for stimulation smoking and negative-effect reduction smoking. Specifically, smoking increased significantly in depressed smokers homozygous for the S allele but not in smokers heterozygous or homozygous for the L allele [245]. However, Luciano *et al.*, 2004 reported that *DRD4* variation did not have an effect on the use of alcohol and the problems that arose from it, on smoking, or on novelty seeking behaviour[246]. Its association with smoking remains to be examined.

Dopamine D_5 receptor gene (*DRD5*)

The *DRD5* gene is on chromosome 4p16.1. Six polymorphisms have been identified in this gene, and functional analyses have shown reduced levels of functional protein [247]. An association between a microsatellite marker within *DRD5* and substance abuse has been found [248]. Sullivan *et al.*, 2001 investigated the association of four *DRD5* polymorphisms compared to smoking initiation and progression to nicotine dependence in a population-based sample of over 900 subjects. For smoking initiation, there was no significant association with the four *DRD5* markers they studied, however maximum likelihood analyses suggested the presence of a haplotype protective against smoking initiation. For progression to nicotine dependence, there were no strongly significant associations with the four *DRD5* markers or for the estimated haplotypes [249].

1.2.4.5 Dopamine transporter

The dopamine transporter protein is encoded by the *SLC6A3* gene, which is located on chromosome 5p15.3. Variations in the dopamine transporter gene have been reported to mediate concentrations of and responses to synaptic dopamine. A variable number tandem repeat polymorphism in this gene has been described [250]. The 9-repeat allele (*SLC6A3–9*) has been associated with dopamine excess disorders [251], and the10-repeat allele (*SLC6A3–10*) has been linked to conditions with insufficient dopamine [252, 253]. Lerman *et al.* (1999) reported an association between smoking behaviour and the *SLC6A3-9* allele in their case-control study of non-smokers and current smokers. They found that individuals with the *SLC6A3-9* allele were less likely to be smokers compared to individuals without *SLC6A3-9*, and that smokers with *SLC6A3-9* started to smoke later and were able to abstain for longer periods than were other

smokers [254]. Sabol *et al.* (1999) replicated this study but added more information by including former, current, and non-smokers as phenotypes. They also included a study of the association between personality traits and *SLC6A3-9* and/or smoking behaviour. They found that *SLC6A3-9* allele carriers were approximately 1.5 times more likely to have given up smoking than those without allele 9. *SLC6A3-9* was also associated with relatively low novelty-seeking scores, which was the most significant personality correlate of smoking cessation. However, Sabol *et al.* did not observe the association between *SLC6A3-9* and smoking initiation [92]. Jorm et al. (2000) replicated the study in a non-volunteer community sample of 861 Caucasians, but no associations were found with either smoking initiation or smoking cessation [255] . Vandenbergh *et al.* (2002) failed to replicate the findings by both Lerman *et al.* (1999) and Sabol *et al.* (1999) and support the absence of effects found by Jorm *et al.* (2000) [256].

Lerman *et al.* (1999) also found a significant gene-gene interaction between *DRD2* and *SLC6A3*, indicating that individuals with *SLC6A3-9* genotypes were significantly less likely to be smokers, especially if they also had *DRD2-A2* genotypes [254]. Erblich *et al.* (2004) investigated 88 African-American smokers and 108 healthy adult smokers and found that smokers carrying either the *DRD2* TaqI A1 RFLP or the *SLC6A3* 9-repeat VNTR polymorphisms had stronger cue-induced cravings than noncarriers [257, 258].

1.2.5 Serotonin system

Serotonin modulates dopamine release and has been implicated in nicotine reinforcement [259]. Serotonin levels are increased by nicotine administration and decreased during withdrawal [260, 261]. In addition, serotonin system play an important role in depression and anxiety, two traits linked with smoking behaviour [262, 263].

Polymorphisms in the serotonin system, such as serotonin metabolism genes [264, 265], serotonin transporter genes [266], and serotonin receptor genes [267] have been linked to smoking behaviour.

The serotonin metabolism gene-tryptophan hydroxylase (*TPH*), located on chromosome 11p15.1, codes for a primary enzyme in the biosynthesis of serotonin. Two polymorphisms in intron 7 (A218C and A779C) have been identified and found to be in complete LD [268]. Lerman *et al.* (2001) studied 249 Caucasian smokers and 202 non-smokers and found that the presence of the 779A-allele was associated with an earlier age of smoking initiation, but found no association with smoking status [264]. Consistent with the findings of Lerman *et al.*, Sullivan *et al.* found a highly significant association between smoking initiation and the two intron 7 markers of *TPH* [265]. Mizuno *et al.* (2004) reported an early starting age of smoking among the *TPH* 218A/A genotyped Japanese population [269].

The serotonin transporter gene (*SLC6A4*) is located on chromosome 17q11.2. A polymorphism in the 5' promoter region of this gene has been reported that yields a short (S) and a long (L) variant [270]. The S variant has been associated with reduced serotonin expression and uptake [271]. The few studies that examined the role of the *SLC6A3* gene in smoking have yielded contradictory results. Lerman *et al.* (1998) found no significant difference in the distribution of *SLC6A4* genotypes among smokers and nonsmokers in either Caucasians or African-Americans [263]. Ishikawa *et al.* suggested that individuals with the S/S genotype were less inclined to smoke and/or could more easily stop smoking than others, supporting a role of the serotoninergic system in smoking behaviour [266].

Serotonergic receptors are divided into seven groups (*5-HT1–7*). The type 2 receptors are categorised into three sub-types (A, B, and C). The *5-HT2A* receptor gene is located on chromosome 13q14.2. A silent T102C polymorphism has been described in this receptor [272] however the C allele is associated with a lower expression level [273] and diminished synthesis of *5-HT2A* receptors.

Many studies have shown that *5HT2A* receptors play a role in schizophrenia and alcohol dependence, two disorders that are strongly associated with smoking behaviour. A research group from Brazil investigated whether T102C is associated with the development and maintenance of the smoking habit. They found differences in the distribution of the genotypes when the current smokers were compared with the never and former smokers, suggesting that T102C is associated with maintenance, but not with initiation of smoking [267].

1.2.6 Cholinergic nicotinic receptor genes

Nicotine functions by binding to nicotinic acetylcholine receptors (nAChRs), which, in turn, modulate the release of dopamine in the mesolimbic system [274]. To date, nine α (α_2 - α_{10}) and three β (β_2 - β_4) subunit genes have been identified [275]. Animal studies have shown that receptors with α_4 and β_2 subunits comprise about 90% of the high-affinity nicotine binding sites in the brian [276]. A few research groups tried to investigate the relationship between nicotine receptor genes and smoking behaviour and came up with conflicting results. Silverman, *et al.* (2000), tested 4 SNPs of *CHRNB2* in Caucasians, and no association was found in relation to smoking initiation and nicotine dependence [277]. Lueders, *et al.* (2002) reported that there was no association between 6 SNPs of *CHRNB2* (individual marker or haplotypes) and smoking behaviour [278]. Feng *et al.* (2004) investigated 6 SNPs of *CHRNA4* and 4 SNPs of *CHRNB2* in 621 men

from 206 families with multiple nicotine-addicted siblings and found an association of two synonymous SNPs in exon 5 (rs1044396 and rs1044397) with a lower risk of nicotine dependence. They also reported that a common (22.5%) *CHRNA4* haplotype, GCTATA, was significantly associated with nicotine addiction. But no association was found in *CHRNB2* [279]. These findings were not confirmed in either male or female subjects of other ethnicities [280, 281]. Li et al. analysed six SNPs (including rs1044396), only one (rs2236196), and its haplotype, was associated with protective effects against nicotine dependence [280]. Saccone et al. targeted 348 candidate genes and analysed 3713 SNPs in 1050 cases and 879 controls. They found that the strongest association was between a SNP (rs6474413) at *CHRNB3*, the β_3 nicotinic receptor subunit gene ($p=9.4 \times 10^{-5}$). They also reported a non-synonymous SNP in the α_5 nicotinic receptor subunit gene *CHRNA5*, which exhibited evidence of a recessive mode of inheritance, resulting in individuals having a 2-fold increase in risk of developing nicotine dependence once exposed to cigarette smoking [131].

1.3 The genetic epidemiology of growth pathway genes

It has been recognised that low birth weight is associated with late onset diseases such as cardiovascular diseases and metabolic syndrome, which includes insulin resistance, hypertension, abnormalities of blood clotting, low HDL and high LDL cholesterol levels, and high triglyceride levels [282-289]. Increasing evidence suggests that either low birthweight or accelerated postnatal weight gain or a combination of the two, may predispose the above diseases [290, 291]. The mechanism of this established association is still uncertain and controversial. For the past few decades, researches have mainly focused on the role of the intra-uterine environment. It has been proposed that undernutrition *in utero* results in a permanent re-programming of fetal metabolism [282]. Heritable effects on low birthweight and late onset diseases were noticed as early as 1962 [292], but research in this area has only recently made significant progress [293-306]. In the following section, I will briefly overview fetal programming and development, followed by an overview of current genetic understanding of growth pathway genes.

1.3.1 Fetal programming hypothesis and its development

The fetal origins of adult disease hypothesis states that environmental factors, particularly nutrition, act in early life to program the risks of cardiovascular and metabolic disease in adult life [282, 283, 288]. The association between low birthweight and coronary heart disease or type 2 diabetes has been replicated in studies from around the world [284]. In order to explain the mechanism of this association, various hypotheses have been proposed and developed. These include the 'thrifty phenotype' hypothesis [289], 'intergenerational effects of fetal programming' [286], the 'predictive

adaptive response' hypothesis [287];Bateson, 2004 258 /id}, and the 'developmental origin' hypothesis [284].

1.3.1.1 The thrifty phenotype hypothesis

The thrifty phenotype hypothesis states that when the fetal environment is poor, there is an adaptive response which optimises the growth of key body organs to the detriment of others and leads to an altered postnatal metabolism, which is designed to enhance postnatal survival under conditions of intermittent or poor nutrition. It was proposed that poor development of pancreatic ß-cell mass and function (including islet of Langerhans vasculature and possibly innervation) were key elements linking poor early nutrition to later type 2 diabetes. It was also suggested that fetal malnutrition led to insulin resistance. Fetal nutrition thereby sets in train mechanisms of fetal nutritional thrift, which have a differential impact on the growth of different organs, with selective protection of brain growth. Altered growth permanently changes the structure and function of the body. A poor functional capacity for insulin secretion would not be detrimental to individuals who continued to be poorly nourished and remained thin and, therefore, insulin-sensitive. Glucose intolerance would be triggered by a positive calorie balance as a result of increased food intake and decreased energy expenditure leading to obesity. The combination of malnutrition during fetal life and infancy followed by overnutrition in childhood and adult life characterises populations undergoing the transition from chronic malnutrition to adequate nutrition. The hypothesis also proposed that the emergence of pathological changes following undernutrition in early life was critically dependent upon the interaction with other factors, notably obesity, ageing and physical inactivity [289].

1.3.1.2 Intergenerational effects of fetal programming

Intergenerational influence was defined as 'those factors, conditions, exposures and environments experienced by one generation that relate to the health, growth and development of the next generation [307]. Potential explanations for intergenerational effects include genetic factors, persistence of adverse conditions across generations, and a programming effect e.g. adverse *in utero* experiences may permanently affect maternal growth and development, altering her metabolism in such a way as to provide an adverse environment for her fetus [286]. Drake et al. reviewed intergenerational effects on birthweight and adult diseases, as well as the intergenerational effects in animal models of programming, and suggested that exposure of the mother to an adverse environment results in a developmental signal to her offspring, resulting in physiological changes or 'programming'. These changes (with or without continuing environmental stress) lead to a developmental signal for the next (F2) generation, and thus a cycle of intergenerational programming is established [286]. The authors also conducted an experiment to explore intergenerational effects in the dexamethasone-programmed rat, a model in which fetal exposure to excess glucocorticoid results in low birth weight with subsequent adult hyperinsulinemia and hyperglycemia underpinned by increased activity of the key hepatic gluconeogenic enzyme, phosphoenolpyruvate carboxykinase (PEPCK). They found that the male offspring of female rats that had been exposed prenatally to dexamethasone, but were not manipulated in their own pregnancy, also had reduced birth weight, glucose intolerance, and elevated hepatic PEPCK activity. These effects resolved in a third generation. Similar intergenerational programming was observed in offspring of male rats exposed prenatally to dexamethasone mated with control females. The persistence of such programming effects through several generations, transmitted by either maternal or paternal lines,

indicates the potential importance of epigenetic factors (modification of the genome over and above alterations in nucleotide sequence) in the intergenerational inheritance of the "programming phenotype" and provides a basis for the inherited association between low birth weight and cardiovascular risk factors [308]. The authors also suggested that these effects may masquerade as a 'genetic' explanation for the association between birth weight and inherited risk of cardiovascular disease [286] [308].

1.3.1.3 Predictive adaptive response and developmental origins of health and disease

Predictive adaptive responses, or PARs, are defined as a form of developmental plasticity that evolved as adaptive responses to environmental cues acting early in the life cycle, but where the advantage of the induced phenotype is primarily manifest in a later phase of the life cycle. The cue affects the processes of developmental plasticity and thus induces changes in the developmental trajectory of form and function such that the organism presets its physiology in expectation of that physiology matching its future environment. The nature of the PARs is determined by the predicted and actual postnatal environment. The fetus sets a range of homeostatic settings appropriate for postnatal life according to the information it receives *in utero*. If the actual postnatal environment matches the prenatal prediction, then the PARs are appropriate and disease risk is low; if they do not match, then the disease risk is increased[287] [309]. The concept of PARs led to the development of the DOHaD (Developmental origins of health and disease) paradigm [310]. They suggested that the developing organism senses its metabolic milieu and adjusts its physiological homeostatic setpoints according to the environment it forecasts will exist after birth. If the signals in the developmental phase suggest limited nutrient availability, then the organism will adjust its

developmental trajectory such that the mature individual has a metabolic homeostasis better adapted for survival in a sparse environment. By pre-adapting to a predicted future environment, they may gain a survival advantage. These can also have deleterious consequences for the risk of later disease. This is particularly true if the developmental prediction has been inaccurate, e.g. if the adult environment is nutritionally rich and sedentary, premature death and disability may thus be the long-term price of adaptations to the developmental environment [311]. The authors suggested that small birth size is not an obligatory part of the pathway leading to increased disease risk, given the evidence of the graded relation between disease risk and birth size that occurs across the normal birthweight range [312]. They also suggested that the underlying mechanisms are thought to involve epigenetic modifications in nonimprinted genes induced by aspects of the developmental environment, which modify gene expression without altering DNA sequences [313].

David Barker recently applied 'developmental plasticity' and 'life-history theory' to explain the relationship between early life conditions and later life diseases. 'Developmental plasticity' is the phenomenon by which one genotype can give rise to a range of different physiological or morphological states in response to different environmental conditions during development. 'Life-history theory' suggests that increased allocation of energy to the development of one trait, such as brain growth, necessarily reduces its allocation to one or more other traits, such as tissue repair processes. Small babies, who have had a lesser allocation of energy, must incur higher costs, including disease in later life through at least three kinds of processes: fewer cells in key organs, such as kidneys; 'thrifty' way of handling food (hormones and metabolism); and more vulnerable to adverse environmental influences in later life [284].

1.3.2 Genetic determinants of early growth and adult disease risk

Although observations of low birthweight and adult disease have been predominantly reported to be attributable to environmental factors, especially undernutrition conditions during pregnancy, heritable effects could also be important contributory factors. As early as 1962, Neel proposed the 'thrifty genotype' hypothesis, which suggested that there may be common genetic polymorphisms which conferred some survival advantage during earlier times of undernutrition, and through the process of selection are now over-represented in certain populations who have adapted to conditions of poor or intermittent nutrient supply [292]. While the original 'thrifty genotype hypothesis' and subsequent debate has been centred on how genetic variations might enhance survival during adult life [303], Ong and Dunger have modified the original hypothesis. They argued that 'thrifty fetal genotypes' which act in the fetus to promote adaptive responses to fetal stress or undernuitrition could have a much greater selection bias than adulthood survival genotypes due to higher fetal and infancy mortality rates compared to those during adulthood [300]. They suggested that thrifty fetal genotypes may enhance these adaptations and improve perinatal survival but predispose to adulthood disease [301]. They also suggested that fetal growth is influenced by both fetal genes and maternal-uterine-placental factors according to the results from the Avon Longitudinal Study of Pregnancy and Childhood (ALSPAC). They pointed out that important maternal-placental factors include parity, smoking and weight gain, as well as genetic factors in the mother or fetal placenta [302].

Based on the observation of reduced birthweight in infants who inherited rare glucokinase gene mutations also responsible for maturity-onset diabetes of the young (MODY), Hattersley and Tooke proposed the 'fetal insulin hypothesis': genetically determined insulin resistance could result in low insulin-mediated fetal growth *in utero*

as well as insulin resistance in childhood and adulthood. Low birthweight, measures of insulin resistance in life, and ultimately glucose intolerance, diabetes, and hypertension, would all be phenotypes of the same insulin-resistant genotype. The crucial components of this proposal are that: (i) genetic susceptibility is important for both birth weight and the adult phenotype of disease like diabetes; and (ii) genes that reduce insulin secretion or increase insulin resistance will predispose to small babies as well as to diabetes [314]. Central to this fetal insulin hypothesis is the concept that insulin-mediated fetal growth will be affected by fetal genetic factors that regulate either fetal insulin secretion or the sensitivity of fetal tissues to the effects of insulin [296].

Lindsay et al. (2000) supported the genetic link between lower birth weight and later adult disease by investigating the role of paternal inheritance in the association of low birth weight and diabetes. They reported that fathers of lower birthweight babies are more likely to develop diabetes later in life, i.e., fathers of offspring in the lowest quintile of birth weight, who were not diabetic at the time of birth of their children, had a 1.8-fold increased risk of developing diabetes later in life (95% CI 1.2-2.7; p=0.004) [298].

Based on the observations that type 2 diabetes is associated with a small body size at birth [315] and an association between the peroxisome proliferator-activated receptor (PPAR)-γ2 gene and type 2 diabetes [316], Eriksson et al. assessed whether the effects of the Pro12Ala polymorphism of the PPAR-γ2 gene on insulin sensitivity and insulin concentrations in adult life are modified by size at birth. They found that the effects of the Pro12 and Ala12 polymorphisms of the PPAR-γ2 gene in elderly people depended on their body size at birth. The well-known association between small body size at birth and insulin resistance was seen only in individuals with the high-risk Pro12 allele. In those who had low birth weight, the Pro12 polymorphism was associated with increased

insulin resistance ($P < 0.002$) and elevated insulin concentrations ($P < 0.003$) [317]. They also reported that the Ala12 allele was associated with increased serum total, low-density lipoprotein (LDL), and non-high-density lipoprotein (non-HDL) cholesterol concentrations but only among those who had birth weights below 3000 g [318]. The authors suggested that these interactions between the effects of the PPARγ-2 gene on adult traits and the effects of birth weight may be interpreted as examples of gene-environmental interactions, which underlie plasticity during development.

The plasma cell glycoprotein 1 (PC-1) gene was also investigated by the same research group, who found that those subjects carring the 121Q allele had a significantly higher prevalence of type 2 diabetes and hypertension combined, but only in the presence of a small body size at birth [319, 320].

It was reported that insulin-like growth factor-I (IGF-I) and insulin are essential fetal growth factors [297]. Vaessen et al. reported that a shortage of IGF-I has been implicated in the pathogenesis of type 2 diabetes and cardiovascular disease in later life, and the absence of the most common allele (192bp) in the promoter region of the *IGF-I* gene (CA microsatellite) results in lower IGF-I concentration [304], and lower birthweight [305]. They supported the hypothesis that association between low birthweight, diabetes and cardiovascular disease, could be a result of genetic variation affecting both fetal growth and susceptibility to late-onset disease. However, other investigators could not replicate these results and suggested type I statistical errors in the original study. Day et al. did not find any association between the CA microsatellite and lower birthweight after reassessing 627 men and 426 women, representing the Hertfordshire, UK, cohorts in which the birthweight associations with later onset diseases were first identified [293]. Frayling et al. reported an opposite IGF-I level in relation to 192bp allele of *IGF-I* gene whilst investigating the involvement of the CA

microsatellite in type 2 diabetes using a case-control cohort of 640 U.K. individuals aged 25 [295]. An inverse correlationship was observed between the number of CA repeats in the *IGF-I* gene and circulating IGF-I levels in two association studies in breast cancer [299, 306].

Growth hormone (GH) genes are potential candidates for regulating the growth trajectory after birth since abnormalities of GH/IGF axis are observed in a number of the chronic adult diseases associated with low birth weight, including cardiovascular disease, diabetes and osteoporosis [294]. The *GH* gene cluster is located on chromosome 17 and contains 5 related genes, encoding adult GH (*GH1*) and placental GH (*GH2*), as well as placental lactogens (*CSH1*, *CSH2* and *CSH5*), although *CSH5* is an almost inactive pseudogene. The genes are in the order *GH1*, *CSH5*, *CSH1*, *GH2*, and *CSH2* and span 66.5 kb. Strong linkage disequilibrium occurs across this gene cluster such that in two or more of the genes, inherited together, may reduce growth in early life while predisposing to disease later in life [294]. The gene cluster is an excellent candidate region for predisposition to restricted early growth and later insulin resistance [321]. The placentally-expressed CSH1, CSH2, and GH2 genes have key roles in the regulation of fetal glucose supply and growth [322, 323]. GH1, through transcriptional regulation of the gene for insulin-like growth factor-I (IGF1) and related genes, has a critical role in the regulation of postnatal growth [324]. A complex microsatellite in the promoter region of *CSH1*, previously called "GH microsatellite" and used as a 17q23 linkage marker in hypertension studies [325-327] has been designated as *CSH1.01* and investigated as an association marker in early growth and late life metabolic syndrome [294]. In that study, 909 Caucasian subjects aged 60~65years from Hertfordshire, UK were investigated. Men with *TT* genotype were significantly lighter at 1 year of age and were shorter as adults compared with *D2D2*

genotype. They also had higher blood pressure, fasting insulin levels (*TT* 66% higher than *D2D2*) and elevated triglyceride and glucose concentrations during a glucose tolerance test, but these associations were only found in males. Sexual dimorphism of GH diurnal profile and tissue response is well known [328], and family based linkage studies of hypertension have found effects attributable to this genomic region but only in males [325, 327].

The human *ACE* (angiotensin converting enzyme) gene is also located on human chromosome 17 and has been extensively studied for genotypic associations with a wide range of disease traits, such as coronary artery disease and myocardial infarction [329-332], hypertension [327], left ventricular hypertrophy and dysfunction [333, 334] and metabolic traits [335]. The *ACE* gene has an insertion/deletion polymorphism, which is due to the presence or absence of a 287bp fragment inside intron 16. It has been identified that the absence (deletion, *D* allele) rather than the presence (insertion, *I* allele) of the 287bp fragment is associated with high ACE activity in circulation and tissues and high cardiac growth [335-338]. Cambien et al. (1998) investigated a group of young adults born at term whose length or weight at birth were in the lowest 3% of the sex and gestational age adjusted distribution (SGA, $n = 172$) compared to a group of control individuals born with an appropriate birth weight for gestational age (AGA, $n = 207$). They reported that the I allele was related to greater insulin secretion only in those born small but not in those born appropriate for gestational age [339]. Kajantie et al. further examined interactions between the *ACE* gene I/D polymorphism and size at birth on glucose and insulin metabolism in a group of elderly men and women around 70 years of age. They reported that the *ACE* I allele is associated with shorter duration of gestation and higher birth weight. The association between the presence of the *ACE* I

allele and increased indices of adult insulin secretion is confined to subjects with low birth weight [340].

1.3.3 Linkage disequilibrium between *GH-CSH* gene cluster and *ACE* gene

The *ACE I/D* gene has been extensively studied for genotypic associations with a wide range of disease traits, as discussed in the previous section. The significant relationship between *D* allele dose and ACE concentration, with the highest ACE levels found for the *D/D* genotype, lowest for the *I/I* and intermediate for the *I/D*, was first reported by Rigat et al and then repeatedly confirmed by other studies, for circulating [338, 341], cellular [342] and tissue ACE [343]. The relationship of *D/D* genotype associated with elevated ACE levels in ventricular tissue has helped the investigation of association between local renin-angiotensin systems within LV tissue and left ventricular hypertrophy which is recognised as a major independent risk factor for morbidity and mortality from cardiovascular causes [344, 345]. It was hypothesised that if cardiac renin-angiotensin systems are important regulators of myocardial growth, then individuals with higher cardiac ACE levels may be expected to exhibit a greater response to a hypertrophic stimulus such as physical training [346, 347]. This hypothesis has been confirmed by Montgomery et al. with a positive association between *ACE D/D* genotype and exercise induced left ventricular growth in 420 young males recruited to the Army Training Regiment Bassingbourn, UK [345] and other studies [335-337, 348].

However, because the *ACE I/D* polymorphism is intronic and does not have a known functional effect, the mechanism of ACE overexpression in subjects with *D/D* genotype is unclear. Systematic investigation of SNPs across *ACE* suggested that several sites (at

least one in the promoter and at least two in the distal half of the gene) are jointly associated with ACE concentration [349]. It is also possible that this relationship is the result of tight linkage to another locus involved in the regulation of *ACE* gene expression [350].

As discussed previously, the *GH* (growth hormone) gene cluster is also located on chromosome 17 and contains 5 related genes, which are in the order *GH1*, *CSH5*, *CSH1*, *GH2*, and *CSH2*. The *ACE* gene is about 370kb downstream of the *CSH2* gene. A complex microsatellite in the promoter region of *CSH1 (CSH1.01)* has been investigated as an association marker in early growth and late life metabolic syndrome [294]. In addition, moderate LD ($|D'| = 0.24$) between *CSH1.01* and the *ACE I/D* polymorphism was also observed. A suggestion that some of the many effects ascribed to the *ACE* genotype, such as myocardial infarction, metabolic traits, birth size and insulin resistance, may reflect a causality originating in the *GH-CSH* cluster has been proposed.

1.4 Hypotheses, aim and strategies

This study aims to advance our understanding of the genetics of two independent complex traits: 1) adolescent smoking behaviour; and 2) growth pathway genes in relation to birthweight and metabolic risk

1.4.1 Hypothesis, aim and strategies of smoking study

As summarised in detail above, smoking is the biggest risk factor for coronary disease and lifelong habits often initiate in adolescence. Inactivating mutations in *CYP2A6,* which encodes the main enzyme that inactivates nicotine to cotinine, affect both cotinine levels and smoking behaviour in adults. Diversity of genes relevant to dopaminergic and related pathways, which are important in addiction, has also been reported to be associated with smoking habits in adults. Our hypothesis is that genetic variations in *CYP2A6* and dopaminergic genes affect smoking behaviour at inception in adolescence, which would be of considerable interest, given that addiction tends to be long lasting once established. It is also possible that individual therapy, personalised by genotype, could be of future utility.

The aim of the smoking study was to examine whether smoking behaviour in teenagers was related to genotypes of *CYP2A6, TH01, DBH, MAOA, DRD4* and *5HT2A* genes, using the Ten Towns Heart Health Study (TTHHS) a longitudinal study of 1520 adolescents which assessed smoking behaviour at 13-15 and 18 years.

The strengths of this study include: 1) analysis of an adolescent population, a critical period in the formation of the lifelong habit of smoking. When this project started in 2001, there was no reported behavioural molecular genetic study of smoking inception in adolescence; 2) large sample size. There are 1520 adolescents in the TTHHS, which

provides a larger sample size than previously used genetic smoking studies; 3) more candidate genes and variations in relation to adolescent smoking behaviour were to be investigated, including both nicotine metabolism pathway gene and dopaminergic rewarding genes.

The strategy for this study included: 1) a DNA bank set up and quality control measurements; 2) identification of relevant candidate genes and variants for cigarette smoking; 3) development of accurate, cost-efficient, and high throughput genotyping methods; 4) genotype-phenotype analyses for the association of genotypes of candidate genes and smoking behaviour in adolescents.

1.4.2 Hypothesis, aim and strategies of growth pathway genes study

The associations between low birth weight and adult disorders, such as cardiovascular diseases and metabolic syndrome, are attributed to both environmental factors, especially maternal undernutrition during pregnancy, and genetic factors such as GH/IGF markers, *ACE* gene I/D, as well as gene-gene and gene-environmental interactions. Environmental factors have been widely investigated, while genetic factors and their interaction with environmental elements still need to be strengthened. My study is mainly focused on investigating associations between candidate growth pathway genes, birthweight, and late onset disease traits.

1.4.2.1 *CSH1.01* genotype, birthweight and metabolic traits

Previous studies in an elderly (60-75y) cohort from Hertfordshire, UK has shown that a common genetic marker *CSH1.01* in the growth hormone – chorionic somatomammotrophin (placental lactogen) (*GH-CSH*) gene cluster associates both with traits of the metabolic syndrome and with weight at one year (only in males), but not

with birthweight [294]. We propose that this could represent a survival or ascertainment bias, particularly in a cohort born into the prevailing environmental and healthcare conditions of 1930-1940s Britain. We aimed to examine the same genetic marker against birthweight and adolescent anthropometric and metabolic traits in a cohort born more recently (between 1982-86 in England and Wales, with the specific hypotheses based on the Hertfordshire study [210, 294], ie. that any association might only be expected in males, that the 'D2' coding of marker *CSH1.01* might show a greater magnitude of effect than the 'D1' coding, and that heterozygotes might emerge showing intermediate values for the 'D2' coding.

1.4.2.2 *IGF1* gene, birthweight and metabolic traits

A previous report studying older Dutch subjects (mean age 66y) identified that for a polymorphism near the promoter region of the *IGF1* gene, the commonest allele (192bp) was associated with higher IGF-I levels, greater height and lower risk of type 2 diabetes and myocardial infarction [304] and higher birthweight [305]. A UK 'replication' study found an opposite association for IGF-I levels, but no associations for the other traits [295]. A previous study in our laboratory did not find the association between *IGF1* microsatellite and low birthweight [293]. We aimed to examine this same genotype in relation to birthweight and other anthropometric and metabolic traits in the TTHHS.

1.4.2.3 *ACE I/D, GH-CSH* markers and ACE activity and heart growth

Previous studies demonstrated an association between *ACE I/D* genotypes and ACE activity, LVM growth response to exercise, and a wide variety of cardiovascular diseases and metabolic traits. Markers on *GH-CSH* gene cluster, such as microsatellite *CSH1.01* and a SNP *GH-CSH Bgl IIB* also showed correlation with metabolic and

cardiovascular traits. The intronic position and lack of known functional effect of *ACE I/D*, the informative *GH-CSH* markers, as well as their linkage disequilibrium led to the hypothesis that some *ACE D* allele phenotypic associations may represent proxy marking of causal factors located in the *GH-CSH* gene cluster. In order to examine this hypothesis, we aimed to investigate *ACE I/D* genotype and *GH-CSH Bgl IIB* genotype, which is in strong linkage with *CSH1.01* (D' = 0.87), in a group of Caucasian males recruited to the Army Training Regiment, Lichfield, UK using cardiovascular data such as ACE activity and LV mass.

Chapter 2 Materials and methods

2.1 Study cohorts and phenotypes

2.1.1 The Ten Towns Heart Health Study (TTHHS)

TTHHS is a mixed longitudinal study based on children who originally attended a stratified random sample of 100 primary schools in towns across England and Wales [351, 352]. The third phase of this study, carried out in 1998–2000, included 69 secondary schools attended by at least 20 pupils who participated in the earlier phases of the study. Pupils who had previously taken part were recruited (average 45 pupils per school), together with a random sample of new participants (average 13 pupils per school). All assessments were made by a trained field team (four observers) who visited each town in turn. Pupils completed a questionnaire, had physical measurements and provided a saliva sample for cotinine assay [353]. Older pupils provided a venous blood sample.

All children providing a blood sample in 1998–2000 were followed up by post on their 18th birthday. A short questionnaire enquired about smoking habits at age 18 and a saliva sample was requested, to be returned by post, for cotinine assay. Non-respondents were followed up by post and telephone where possible.

Approval was obtained from all relevant Local Research Ethics Committees. Invitations were sent jointly to parents and pupils; written informed consent was obtained from all participants and their parents. The TTHHS was supported by a project grant from The Wellcome Trust (reference 051187/Z/97/A) and the genetic studies and 18-year follow-up by a grant from the Medical Research Council (G9900686).

Smoking status at 13–15

Smoking status at 13–15 was based on a shortened version of the questionnaire used for the national survey of smoking among secondary school children [354]. Current smoking at ages 13–15 was defined as a positive response to 'Do you smoke at all nowadays?', or an acknowledgement that they 'smoke ≥ 1 cigarette per week' or that 'they do sometimes smoke cigarettes'. Finally anyone with a cotinine ≥ 15 ng/ml [355] was classified as a smoker irrespective of their questionnaire responses. A never smoker was someone who 'had never tried smoking a cigarette, not even a puff or two' and whose cotinine, if available, was <15 ng/ml.

Smoking status at 18

Current smokers at age 18 were defined as those who admitted smoking (any amount) currently (n=392) plus a further 42 who had a cotinine level above 15 ng/ml. In practice, virtually all participants either admitted smoking at least one cigarette per day or had a raised cotinine level (≥15 ng/ml). Ex-smokers were those who admitted they had smoked cigarettes, but who said they did not smoke at all nowadays and had a cotinine, if available, <15 ng/ml (n=373). Never smokers were those who denied ever having smoked a cigarette and who had a cotinine, if available, <15 ng/ml (n=448).

Additional information on amount smoked and on markers of tobacco addiction was requested at both ages. At age 13–15 children were asked 'if they usually smoke more than six cigarettes per week'. At age 18 they were asked, 'About how many cigarettes do you usually smoke per day?' Questions on addiction (both ages) included: 'How old were you when you first tried smoking a cigarette, even if it was only a puff or two?'; 'How easy would you find it to go without smoking for as long as a week?' 1=very difficult, 2=fairly difficult, 3=fairly easy, 4=very easy; 'How soon after waking do you

usually smoke your first cigarette of the day?' 1=first 5 min, 2=5–14 min, 3=15–29 min, 4=30 min– 1 h, 5=1–2 h, 6=2 hours or more; 'If you have smoked in the last 2 days, how many hours ago was the last one?'; 'Have you ever tried to give up smoking?' yes or no.

For the growth study, measurements were all taken at age 13-15 for TTHHS. They are height and weight, serum glucose and insulin levels. In addition, leg length was determined by subtracting sitting height (measured using a standard chair) from standing height. Body mass index (BMI) was calculated by a person's weight in kilograms divided by height in meters squared (BMI=kg/m^2). Birthweight measurements were available from several sources and where available were taken in order of preference from (1) birth record data; (2) birth registers; (3) from parental recall (questionnaire data) [356].

2.1.2 The Lichfield Army Recruit Growth in Exercise Heart Study (LARGE Heart Study)

The LARGE Heart Study, a component of the 'Lichfield Study', was designed to elucidate the genetic and environmental factors which modulate exercise-induced left ventricular growth responses in normal healthy male Caucasians. Participants were Caucasian males recruited to the Army Training Regiment between July 2002 and April 2004. All participants successfully completed an Army medical examination before undergoing a 12-week period of intensive physical training. All exercises were standardized and supervised to ensure full compliance. Anthropometric and lifestyle data were collected, including participant's height and weight, history of smoking, alcohol consumption, and physical activity. Blood pressure, serum ACE level (activity), and LVM were measured at the start (pre) and end (post) of the training period. Venous

blood samples were centrifuged immediately at 10 000 r.p.m. for 10 min and the supernatant stored at – 80°C; subsequently ACE activity was assayed by a modified fluorometric method using carbobenzoxy-phenyl-alanyl-histidyl-leucine (Z-Phe-His-Leu) as substrate [357]. Cardiovascular magnetic resonance was performed to measure LVM using standard techniques [358].

2.2 Chemicals and reagents

2.2.1 Chemicals and reagents for establishment of the DNA bank

Potassium hydrogen carbonate, Fisher Scienfific UK Ltd, UK

Ammoniun chloride, Fisher Scienfific UK Ltd, UK

Ethylenediaminetetraacetic acid disodium salt, Fisher Scienfific UK Ltd, UK

Tris(hydroxymethyl)-methylamine, Fisher Scienfific UK Ltd, UK

Sodium chloride, Fisher Scienfific UK Ltd, UK

Sodium dodecyl sulphate, Fisher Scienfific UK Ltd, UK

Protease, from Streptomyces griseus, SIGMA-ALDRICH CO. USA

Absolute ethanol, VMR International Ltd, UK.

PicoGreen®, Molecular Probes, Inc., Oregon, USA

2.2.2 Chemicals and reagents for genotyping

2.2.2.1 Chemicals and reagents for PCRs

10 × Long PCR buffer: 500mM Tris-HCl, pH 8.9, 140mM Ammonium Sulphate

10 × PCR buffer (minus Magnesium), Gibco BRL, Life Technologies, UK

10 × PCR buffer (containes 15mM $MgCl_2$), Qiagen Ltd. UK

5 × Q-solution, Qiagen Ltd. UK

Dimethyl Sulphoxide (DMSO), DBH Fisher Scienfific UK Ltd, UK.

Deoxynucleotide Triphosphates (dNTPs), 100mM, pH 7.5, Promega, USA

Deoxyinosine 5'Triphosphate (dITP), 100mM, pH7.5, Bioline, USA

Magnesium Chloride, 50mM, Gibco BRL, Life Technologies, UK

Taq DNA polymerase, Gibco BRL, Life Technologies, UK

HotstarTaq® DNA Polymerase, Qiagen Ltd. UK

Betaine, SIGMA, Chemical Co. USA

Pwo DNA polymerase, Roche Diagnostics, Lewes, UK

Primers, 100pmol/µl, MWG-Biotech, Ebersberg, Germany (www.mwgdna.com).

2.2.2.2 Chemicals and reagents for making gels

Agarose, Gibco, BRL, Life Technologies, UK

Acrylamide-bis-acrylamide stock solution, Severn Biotech Ltd. UK.

Duracryl matrix,Genomic Solutions, Ann Arbor, MI, USA

Ammonium persulphate, Fisher Scientific International Co. UK. Ltd

TEMED (N, N, N', N'-Tetramethylethylendediamine), Sigma Chemical Co., USA

10 × TBE, consisting following chemicals:

Trisbase, Fisher Scienfific UK Ltd, UK

Na EDTA, Fisher Scienfific UK Ltd, UK

Orthoboric Acid, Fisher Chemicals UK Ltd, UK.

Sticky silane, consisting following chemicals:

Glacial acetic acid, BDH Laboratory Suppliers, UK

γ-methylacryloxypropyltrimetoxy silane, Sigma Chemical Co., USA

Ethanol, BDH Laboratory Suppliers, UK

2.2.2.3 Chemicals and reagents for gel staining

Ethidium Bromide, Sigma Chemical Co., USA

Vistra Green®, Amersham Pharmacia Biotech, UK Ltd, UK

2.2.3 Other commonly used materials and chemicals

Gilson Pipetman® Pipette, (2µl, 20µl, 200µl, 1000µl), Anachem, UK.

Disposable tips, Life Science International Ltd., Basingstoke, Hampshire, UK.

Multiple channel Finnpipette (8 channel), Labsystem, Helsinki, Finland.

96-pin passive replicator, MADGE Bio. Ltd, Nottingham, UK.

MADGE former (96-well, 384-well, 768well), MADGE Bio. Ltd, Nottingham, UK.

Glass plates, (110mm × 170mm), local suppliers.

96-well plates, Thermo-Fast® Abgene house, UK Cat. No.AB-0800.

384-well plates, Robbins Scientific Corporation, CA. Cat. No. 1047-00-0

Microseal™ `A` Film, Cat. No. MSA-5001. MJ Research Inc, USA (www.mjr.com).

Rubber mat (OmniSeal TD Mats), Hybaid Limited, Middlesex, UK

Plate sealer, B.I.S. Ltd., Kirkham, Lancastershire, UK.

Microcentrifuge tubes, 0.5ml and 1.5ml, Alpha Laboratories Ltd. Hampshire, UK.

Deionised formamide, SIGMA Chemical Co., USA

Xylene cyanole FF, SIGMA Chemical Co., USA

100bp DNA ladder, Promega, MADISON, USA Cat No. G2101

1kb DNA ladder, Promega, MADISON, USA Cat No. G571A12864704.

Sodium hypochlorite solution, Cat. No. 53, 200-2, Sigma-Aldrich CO. Ltd. UK

Chill-Out™ wax, Genetic Research Instrumentation, UK

2.3 Equipment

2.3.1 Equipment for DNA extraction

SORVALL® RT7 Centrifuge, Kendro Labrotory Products. Newtown, Ct 06470 USA

Rotatest shaker, Model R100, Genetic Research Instrumentation Ltd, England

Incubator Shaker, New Brunswick Scientific Co. INC, EDISON, N.J., USA

Microcentrifuge, International Equipment Company, centra® MP4 Model No. 230, USA

Fluorimager® 595, Molecular Dynamics, Inc., California, USA.

2.3.2 Equipment for genotyping and aliquoting

Centrifuge, International Equipment Company, centra®, MP4 Model No. 23690, USA

PCR Thermal Cycler (DNA Engine Tetrad, PTC-225), MJ Research Inc., USA.

Fluorimager® 595, Molecular Dynamics, Inc., California, USA.

Submerge gel electrophoresis tank, Sigma-Aldrich, UK

Dry gel electrophoresis tank, Medical Physics and Bioengineering, Southampton

General Hospital

Thermal-stable MADGE tank, Medical Physics and Bioengineering, Southampton

General Hospital.

Power pack, 200V/A, 50-60 Hz, BIO-RAD Laboratories Inc., CA, USA.

QUADRA96® SV MODEL 345 (aliquoting robot), Cat No. 196-345, TOMTEC, USA.

ABI Prism™ 377 DNA sequencer, Perkin Elmer, Applied Biosystems International,

UK.

Idaho Technology 384-well Odyssey, Salt Lake City, Utah, USA.

LightTyper software, Roche Diagnostics ltd.

2.4 Methods

2.4.1 DNA bank establishment and quality control

Genomic DNA was extracted and then stored in Tris-EDTA pH 7.5 (TE) buffer as primary stock. The samples were diluted to 10ng/µl with water into 96 deep well arrays to give a working stock. 50µl of this stock was subaliquoted into a 96 well plate to be used for subsequent genotyping procedures. A duplex PCR assay was designed and performed for DNA quality control.

2.4.1.1 Extraction of genomic DNA

DNA extractions were processed by the salting out method [359]. In this method, the red cells were removed by washing three times with erythrocyte lysis buffer (ELB). The presence of 0.15M ammonium salt disrupted the red cells leaving the white cells intact. The white cell pellet was lysed and digested with 10% sodium dodecyl sulphate (SDS) and the enzyme protease in nuclei lysis buffer (NLB) and incubated overnight at 37°C. The next day saturated salt was added to the solution to deproteinise by dehydration and precipitation. The precipitate was removed by centrifugation. The supernatant was transferred to another tube and the DNA was precipitated out by adding twice the volume of cold absolute ethanol. The DNA pellet was removed, washed with 70% ethanol then dissolved in Tris EDTA buffer (TE buffer) and stored at 4°C. Methods for making up stock reagents are shown in Appendix.

2.4.1.2 Quantitation of genomic DNA concentration

A fluorescent dye method was used to quantitate genomic DNA concentrations (adapted from Fluorescence Imaging Application Guide, Molecular Dynamics, http://www.link.med.ed.ac.uk/ELEGI/Resources.htm).

PicoGreen® is a DNA specific fluorescent stain allowing accurate quantitation in the presence of RNA and protein. The stock standard solution, DNA concentration 360ng/µl, was diluted 1 in 10 with sterile water to give a working stock solution of 36ng/µl. Serial 1 in 2 dilutions were performed to give a standard curve of 36, 18, 9 and 4.5ng/µl. Genomic DNA was diluted 1 in 20 to fall within the range of the DNA concentration of the calibration curve. Stock PicoGreen® solution was diluted 1 in 200 with TE buffer. 50µl of TE buffer was pipetted into each well of a flat-bottomed 96-well plate. 2µl of mixed diluted DNA sample was added (the last 10 wells of the plate were used for each standard in duplicate and 2 for blanks). 50µl diluted PicoGreen reagent was added to each well, mixed and incubated for 15 minutes in the dark. The plate was scanned on the Fluorimager using excitation wavelength of 488nm, 530nm emission wavelength bandpass filter. The quantitative digital images were analysed using Image QUANT software. The data were processed and the DNA final concentration was calculated using Microsoft Excel®.

2.4.1.3 Equalization of genomic DNA

The genomic DNA samples were divided into two aliquots. One set was kept at -80°C at St Georges Hospital, London and the other set was stored at -80°C in the Human Genetics Division, Duthie Building, Southampton General Hospital.

In order to get good quality PCR products, the DNA was equalised to 10ng/μl in 250μl of sterile water in 96 well arrays. The samples were ranked in order of concentration from the highest to the lowest for convenience. These working stocks were stored at -20°C.

2.4.1.4 Setting up of a small component of the genomic DNA stocks

Two sets of 50μl subaliquots of the working stock were prepared in 96 well plates with re-sealable lids for subsequent genotyping procedures. They were all stored at -20°C.

2.4.1.5 Quality control of DNA bank by gender test

One of the quality control approaches to the DNA bank was designed and performed by combining a duplex PCR assay of X and Y chromosome markers with 768 well microplate array diagonal gel electrophoresis (MADGE).

Aliquots of 2 μl genomic DNA of 384-well polypropylene plates were carried out using robot Quadra Tomtec and dried for subsequent genotyping.

Primer design Two primer sets (primers MAO-F and MAO-R and primers M253F and M253R) were used in the duplex PCR assay. Primers MAO-F (5'-GGA GGC TAG AAG AAG GGA GAT GTG A-3') and MAO-R (5'-GGG GCA GGG CCT TGA ATC TGT AGA AAC T-3'), originally used as an outer pair of primers for genotyping a SNP (T1460C) at the MAO-A locus of X chromosome, acting as a control band and as a secondary template for an allele specific primer operation with one primer of the pair, generate a 211bp PCR product for both males and females. Primers M253F (5'-GGG AGG TTA GAA TTT GAG GCT ATT TTA G-3') and M253R (5'-CAG GAG CAC CTC TAT GCA GTT TAT G-3') were used to produce a 320bp Y-chromosome specific PCR product from gene DDX3Y, which can only be observed in males.

PCR conditions PCRs were performed in a total volume of 10µl containing 20ng template genomic DNA, 0.015µl Taq polymerase, 1x PCR buffer, 2mM deoxynucleotide triphosphates, 2.0mM $MgCl_2$, and 4pM each primer per reaction. The reaction was carried out on high performance Peltier Thermal Cycler. Initial denaturation was performed at 95°C for 3 minutes, followed by 35 cycles each consisting of denaturation at 94°C for 45 seconds, annealing at 56°C for 40 seconds, and extension at 72°C for 45 seconds, followed by a final extension at 72°C for 10 minutes.

Examination of PCR product The duplex PCR products were examined using a palmtop sized 768 well microplate compatible electrophoresis system. 5 µl of formamide dye was added to the 10 µl PCR product and 2µl of this mix was loaded onto the 768 MADGE gel, which was prestained with ethidium bromide for 15-20 minutes, by 96-pin passive replicator. PCR products were electrophoresed using 5% polyacrylamide gel at 150V for 5 minutes and scanned using a Fluorinager 595®. Gel images were viewed and analysed combining Phoretix ID Advanced computer software for MADGE analysis and Microsoft Excel®. Protocol for 768-well gel setting up, loading process and image analyses was exactly as for 96-well, 192-well and 384-well gel protocols [http://tto.trends.com/tto/index_of_TTO.shtml].

2.4.2 Genotyping smoking markers

Markers genotyped for smoking study in TTHHS cohort included *CYP2A6 (*1A, *1B, *2, *4, *5, *9* and **12*); dopamine synthesis marker *HUMTH01* VNTR; dopamine metabolism markers *DBH* G1368A, *MAOA* T1460C and *MAOA* VNTR; dopamine receptor gene *DRD4* VNTR and serotonin receptor gene *5HT2A* T102C.

2.4.2.1 Genotyping of *CYP2A6*

In order to simplify the specific assay of different alleles, a two step assay was constructed for each allele. All primers were listed in Table 2.1 and schematics of the genotyping assays devised were shown in Figure 2.1. Allele-specific PCR combined with MADGE (384 ARMS) was used to genotype *CYP2A6*1A, CYP2A6*2, CYP2A6*1B, CYP2A6*5, CYP2A6*4, *9* and **12*.

A

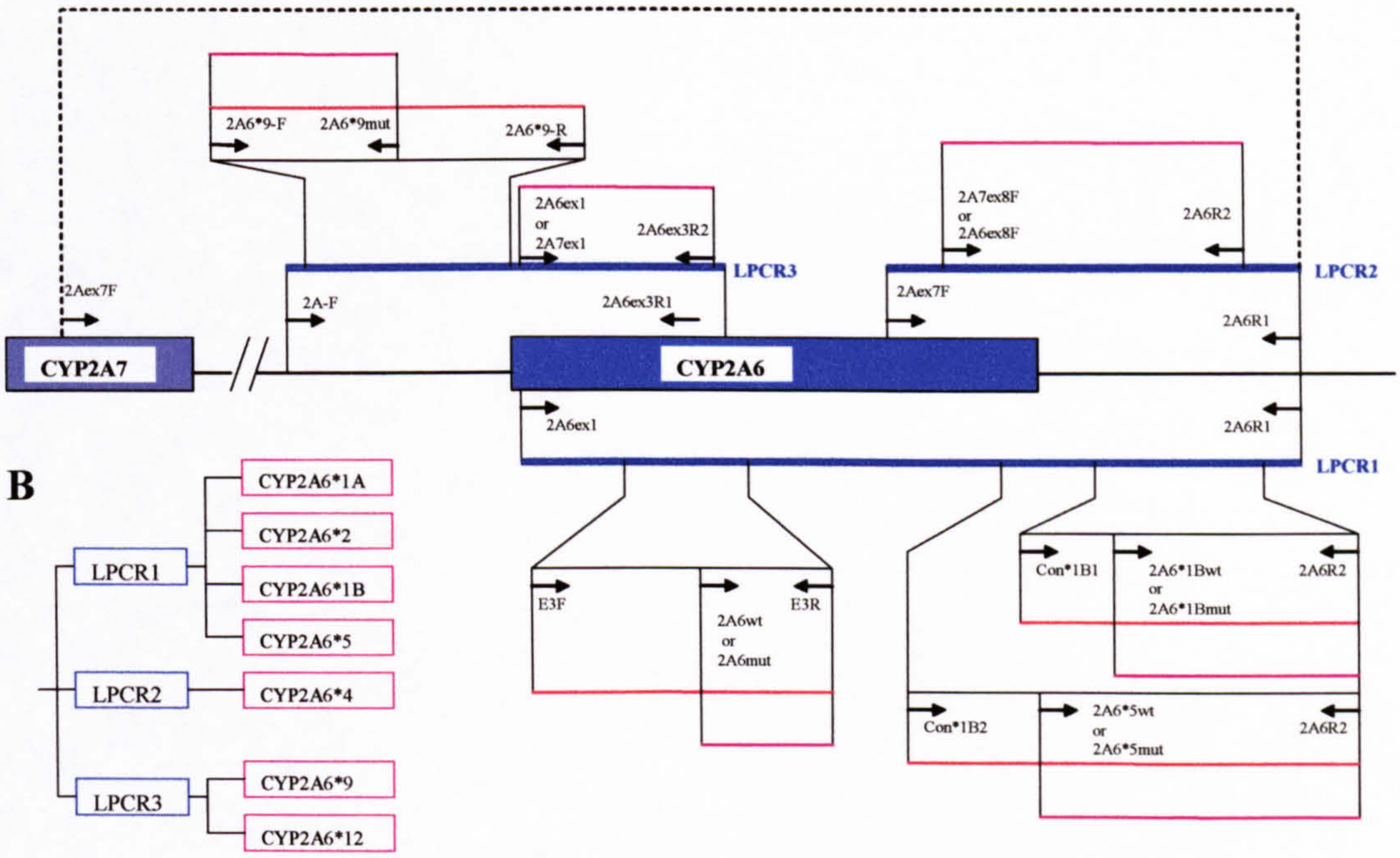


Figure 2.1 Schematic diagram of two-step PCR method for genotyping *CYP2A6*. **A**, the three heavy blue lines represent three long PCR fragments (LPCR1, LPCR2 and LPCR3). '……' represents that if the deletion was present, the LPCR2 product was generated from exon 7 of *CYP2A7* to the 3' UTR of *CYP2A6*. Two primers used for each LPCR were shown on both ends of the blue lines. From each LPCR fragment, different short PCR reactions were performed mainly using ARMS. The red lines represents the control band, and the pink line srepresented allele specific band. **B**, summarised different alleles were generated from different LPCR fragments.

Table 2.1 Primers used for genotyping smoking candidate genes

Primers	Sequence	Location	References	Marker
2A6ex1	5'-GCT GAA CAC AGA GCA GAT GTA CA-3'	Exon 1	Oscarson et al., 1998	*CYP2A6* LPCR1
2A6R1	5'-GCA CTT ATG TTT TGT GAG ACA TCA GAG ACA A-3'	3' flanking region	Oscarson et al., 1999a	*CYP2A6* LPCR1, 2
E3F	5'-GCG TGG TAT TCA GCA ACG GG-3'	Exon 3	Fernandez-Salguero et al., 1995	*CYP2A6*1A, *2*
E3R	5'-TCG TCC TGG GTG TTT TCC TTC-3'	Intron 3	Fernandez-Salguero et al., 1995	*CYP2A6*1A, *2*
2A6wt	5'-CTC ATC GAC GCC CT-3'	Exon 3	Oscarson et al., 1998	*CYP2A6*1A, *2*
2A6mut	5'-CTC ATC GAC GCC CA-3'	Exon 3	Oscarson et al., 1998	*CYP2A6*1A, *2*
con*1B1	5'-GGC CTG GCC AGA ATG GAG-3'	Exon 9	Self design	*CYP2A6*1B*
con*1B2	5'-TGC AGA ACT TCC GCC TCA-3'	Exon 9	Self design	*CYP2A6*1B,*5*
2A6*1Bwt	5'-ACT GGG GGC AGG ATG GC-3'	3'-UTR	Oscarson et al., 1999b	*CYP2A6*1B*
2A6*1Bmut	5'-AAT GGG GGG AAG ATG CG-3'	3'-UTR	Oscarson et al., 1999b	*CYP2A6*1B*
2A6R2	5'-AAA ATG GGC ATG AAC GCC C-3'	3' flanking region	Oscarson et al., 1999a	*CYP2A6*1B,*5*, 4*
2A6*5wt	5'-CCC CAA ACA CGT GGG-3'	Exon 9	Oscarson et al., 1999b	*CYP2A6*5*
2A6*5mut	5'-CCC CAA ACA CGT GGT-3'	Exon 9	Oscarson et al., 1999b	*CYP2A6*5*
2Aex7F	5'-GGC CAA CAT GCC CTA CAT G-3'	Exon 7	Oscarson et al., 1999a	*CYP2A6* LPCR2
2A6ex8F	5'-CAC TTC CTG AAT GAG-3'	Exon 8	Oscarson et al., 1999a	*CYP2A6*4*
2A7ex8F	5'-CAT TTC CTG GAT GAC-3'	Exon 8	Oscarson et al., 1999a	*CYP2A6*4*
2A-F	5'-GCA CCC CTC CTG AGG TAC CAC-3'	5' flanking region	Oscarson et al., 2002	*CYP2A6* LPCR3
2A6ex3R1	5'-GTC CCC TGC TCA CCG CCA-3'	Exon 3	Oscarson et al., 2002	*CYP2A6* LPCR3
CYP2A6*9-F	5'-TCC CTT GCT GCC TGT GTC CT-3'	5' flanking region	Self design	*CYP2A6*9*
CYP2A6*9-R	5'-GCG AGT TGT ACA TCT GCT CTG-3'	Exon 1	Self design	*CYP2A6*9*
CYP2A6*9mut	5'-GCT GGG GTG GTT TGC CTC TC-3'	5'flanking region	Self design	*CYP2A6*9*
2A6ex3R2	5'-CGC TCC CCG TTG CTG AAT A-3'	Exon 3	Oscarson et al., 2002	*CYP2A6*12*
2A6ex1	5'-AAC ACA GAG CAG ATG TAC A-3'	Exon 1	Oscarson et al., 2002	*CYP2A6*12*
2A7ex1	5'-AAC ACA GAG CAC ATA TGT G-3'	Exon 1	Oscarson et al., 2002	*CYP2A6*12*
DBH-F	5'-GAG CTG AGA ACG ACT GGC TTA TTG A-3'	Exon 9	Self design	*DBH* G1368A
DBH-R	5'-GCT GAT GGT CAC ATT GGC TTT TCC T-3'	Exon9	Self design	*DBH* G1368A
DBH-wt	5'-TGT GGC CAG CTC CCG GTC TAC T-3'	Exon 9	Self design	*DBH* G1368A
DBH-mut	5'-CAC CTC CTG CAC GTA CAA CCC G-3'	Exon 9	Self design	*DBH* G1368A
MAO-F	5'-GGA GGC TAG AAG AAG GGA GAT GTG A-3'	Intron 13	Self design	*MAOA* T1460C
MAO-R	5'-GCA GGG CCT TGA ATC TGT AGA AAC T-3'	Intron 15	Self design	*MAOA* T1460C
MAO-Wt	5'-GGG GAA GGT GAC CGA GAA ATA T-3'	Exon 14	Self design	*MAOA* T1460C
MAO-Mut	5'-GGG GAA GGT GAC CGA GAA ATA C-3'	Exon 14	Self design	*MAOA* T1460C
MAO-VNTR-F	5' -ACA GCC TGA CCG TGG AGA AG -3'	5'flanking region	Sabol et al., 1998	*MAOA* VNTR
MAO-VNTR-R	5' -GAA CGG ACG CTC CAT TCG GA -3'	5'flanking region	Sabol et al., 1998	*MAOA* VNTR
232bp-F	5'-ATG ACA AAG GGT GGA GGG ATG A-3'	-	Self design	*MAOA* VNTR sizing
232bp-R	5'-AGA GGG CGA TGT CTT GAT GGA C-3'	-	Self design	*MAOA* VNTR sizing
454bp-F	-GTG CCT GGT GTT TGA GTC CTG TT-3'	-	Self design	*MAOA* VNTR sizing
454bp-R	5'-TGG TGA ATG AAT GAA TGA ATG AAT G-3'	-	Self design	*MAOA* VNTR sizing
DRD4-VNTR-F	5'-GCG ACT ACG TGG TCT ACT CG-3'	Exon 3	Lechter et al., 1993	*DRD4* VNTR
DRD4-VNTR-R	5'-AGG ACC CTC ATG GCC TTG-3'	Exon 3	Lechter et al., 1993	*DRD4* VNTR
pGEM264bp-F	5'-GGTGAGCAAAAACAGGAAGG-3'	-	Self design	*DRD4* VNTR sizing
pGEM264bp-R	-GCCTCGTGATACGCCTATTT-3'	-	Self design	*DRD4* VNTR sizing
F822	5'-CAA CCA AGA CTA CAA GTA CCG CGT CAG TGA-3'	-	Self design	*DRD4* VNTR sizing
R822	5'-AAC CAG CGG AAG AGG TCA AGG G-3'	-	Self design	*DRD4* VNTR sizing
5HT2a-F	5'-TCA CTA CGG CTG TCA GTA AAG CAG A-3'	Exon 1	Self design	*5HT2A* T102C
5HT2a-R	5'-CAT CAA GGT GAA TGG TGA GCA GAA AC-3'	Exon 1	Self design	*5HT2A* T102C
5HT2a-rt	5'-CAG GCT CTA CAG TAA TGA CTT CAA CTC T-3'	Exon 1	Self design	*5HT2A* T102C
5HT2a-rc	5'-GGC TCT ACA GTA ATG ACT TCA ACT CC-3'	Exon 1	Self design	*5HT2A* T102C

Genotyping of *CYP2A6*1A, CYP2A6*1B, CYP2A6*2* and *CYP2A6*5*

The first step for *CYP2A6*1A, CYP2A6*1B, CYP2A6*2 and CYP2A6*5* was an 8kb long PCR (LPCR1- blue line) spanning from exon 1 of *CYP2A6* to the region 3' to the gene, using primers designated respectively 2A6ex1 and 2A6R1. The reaction mixture (20µl) contained 20ng genomic DNA, 0.4µM of each primer, 250µM dNTP, 2mM MgCl2, 1.0unit of Gibco Taq polymerase, 2µl Long PCR Buffer, 1.3M Betaine (PCR Enhancer), and 0.008 unit of Pwo (3' editing polymerase to correct mistakes). The reaction was carried out on Tetrad Thermal Cycler. Initial denaturation was performed at 94˚C for 2 minutes, followed by 35 cycles each consisting of denaturation at 94˚C for 20 seconds, annealing at 60˚C for 30 seconds, and extension at 68˚C for 8 minutes, followed by a final extension at 68˚C for 20 minutes. The long PCR products were examined on a 0.7% agarose 96-MADGE gel stained with ethidium bromide in a submerge tank (filled with about 400ml 1×TBE buffer and 40ul ethidium bromide) for 20 minutes, and scanned using Fluorimager® 595. The long PCR products were then diluted 100 fold and a 2µl aliquot was taken as a template for genotyping *CYP2A6*1A, *2, *1B* and *5.

The *CYP2A6*2* allele exists only on a *CYP2A6*1A* background. A pair of primers E3F and E3R were used in conjunction with one of two allele specific primers 2A6wt or 2A6mut, the latter testing for *2 (codon L160H mutation). In each of these three-oligonucleotide reactions, the allele specific binding site is internal to the E3F/E3R amplicon, giving a constant band of 201bp (E3F/E3R- red line) and if present, mutant or wild type band of 60bp (2A6wt or 2A6mut/E3R- purple line). The reaction mixture (10 µl) contained 1x PCR Buffer, 2mM dNTP, 2mM $MgCl_2$, 0.02µM forward primer E3F and 0.4µM reverse primer E3R, 0.4µM of either 2A6wt (for wild type reaction) or 2A6mut (for mutant reaction), and 0.1 unit Gibco Taq Polymerase. The amplification

was performed by initial denaturation at 94°C for 2 minutes, followed by 19 cycles each involving denaturation at 94°C for 20 seconds, annealing at 48°C for 30 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 2 minutes. The PCR products were examined by electrophoresis in 384-MADGE using 5% polyacrylamide gel that was prestained with ethidium bromide.

Essentially identical conditions were used to test for *CYP2A6* 1B* and *CYP2A6* 5* alleles, except using different primers (Table 2.3).

Genotyping of CYP2A6*4

A different PCR (LPCR2) was used for the first step of the test for the *CYP2A6* 4* allele. This 1.9kb amplification used primer 2A6R1 described above, in conjunction with a generic primer recognising exon 7 of both *CYP2A6* and *CYP2A7* (2Aex7F). This generate a 1.9kb amplicon if *CYP2A6* is intact and does the same if unequal crossover has created a *CYP2A7/2A6* hybrid by deleting a region of the genes such that *CYP2A7* exon 7 is now located where *CYP2A6* exon 7 should be, relative to primer 2A6R1. Two specific secondary reactions, one for *CYP2A6* 4* and one for wild type, were then executed. Both reactions used a primer (2A6R2) located just 5' to 2A6R1. The *CYP2A6* 4* test used a primer 2A7ex8F specific to *CYP2A7* exon 8; the wild type test used a primer 2A6ex8F specific to *CYP2A6* exon 8. Depending on which exon 8, that of *CYP2A6* or that of *CYP2A7* is located where exon 8 of *CYP2A6* should be, relative to primer 2A6R2, a 1.2kb amplicon will, or will not, be formed.

Genotyping of *CYP2A6*12* and **9*

The third long PCR fragment (LPCR3) were generated for genotyping *CYP2A6*12* and **9*. The 2.3kb long PCR was amplified using a forward primer 2A-F with sequence

common to both *CYP2A6* and *CYP2A7* in the 5' flanking region upward about 240bp, and reverse primer 2A6ex3R1 that was *CYP2A6* specific in exon 3. The 2.3kb PCR products were then used as a template for genotyping *12 and *9. PCR conditions for three long PCR amplicons are shown in Table 2.2, for second PCR reactions in Table 2.3.

Table 2.2 Protocols for generating CYP2A6 LPCR1, LPCR2, and LPCR3

Recipe		Program	
Template	20ng genomic DNA	Initial denaturation	94 2mins
Long PCR buffer	2ul 10 x long PCR buffer*	Denaturation	94 20secs
dNTP	0.25mM		60 30secs (LPCR1)
$MgCl_2$	2mM	Annealing	45 30secs (LPCR2)
Forward primer	0.4μM		55 30secs (LPCR3)
Reverse primer	0.4μM	Extension	68 8mins
Betaine	1.3M	Cycles	35
Taq polymerase	1.0 unit	Final extension	68 20mins
1/250 Pwo	0.008 unit		

* 500mM Tris-HCl, PH8.9 and 140mM Ammonium Sulphate

Table 2.3 Protocols for second PCR reactions of CYP2A6

PCR recipe	*2	*1B	*5	*4	*12	*9
Template (2ul 1/100 dilution of)	LPCR1	LPCR1	LPCR1	LPCR2	LPCR3	LPCR3
PCR buffer	1xPCR buffer	1xPCR buffer	1xPCR buffer	1xPCR buffer	1xPCR buffer	1xPCR buffer
dNTP	0.2mM	0.2mM	0.2mM	0.2mM	0.2mM	0.2mM
MgCl2	2.0mM	1.5mM	1.5mM	1.0mM	1.0mM	1.0mM
Outer forward primer	0.02uM	0.8uM	0.08uM			0.6uM
Outer reverse primer	0.4uM	0.4uM	0.4uM	0.25uM	0.25uM	0.4uM
Wild type allele primer	0.4uM	0.1uM	0.4uM	0.25uM	0.25uM	
Mutant allele primer	0.4uM	0.4uM	0.4uM	0.25uM	0.25uM	0.5uM
Taq polymerase	0.1 unit	0.1 unit	0.1 unit	0.5 unit	0.5 unit	0.1 unit
PCR program						
Initial denaturation	94°C 2mins	94°C 2mins	94°C 2mins	95°C 1min	95°C 1min	95°C 3mins
Denaturation	94°C 20secs	94°C 20secs	94°C 20secs	95°C 15secs	95°C 15secs	94°C 30secs
Annealing	48°C 30secs	48°C 30secs	48°C 30secs	52°C 20secs	62°C (wild) 58°C (mut) 20secs	55°C 40secs
Extension	72°C 30secs	72°C 30secs	72°C 30secs	72°C 2mins	72°C 2mins	72°C 30secs
Cycles	20	20 (wild type) 22 (mutant)	19	20 (wild type) 16 (mutant)	16	35
Final extension	72°C 2mins	72°C 2mins	72°C 2mins			

2.4.2.2 Genotyping of *HUMTH01*-VNTR

Genotyping of HUMTH01-VNTR

The *HUMTH01*-VNTR was amplified using forward primer FH269-F (5'-CAG CTG CCC TAG TCA GCA C-3') and reverse primer FH270-R (5'- GCT TCC GAG TGC AGG TCA CA-3') [360]. The 10µl reaction mixture included 20ng of template DNA, 0.04µM of each primer, 2mM dNTP, 1.5mM $MgCl_2$, 0.02 unit of Taq polymerase. The reaction was carried out in 1x PCR buffer by initial denaturation at 95°C for 3 minutes, followed by 35 cycles each involving denaturation at 94°C for 45 seconds, annealing at 56°C for 40 seconds, extension at 72°C for 45 seconds, and a final extension at 72°C for 10 minutes.

Fragment sizing markers

The fragment sizing markers were pairs of PCR products that bracket the allele size ranges of each microsatellite. The *HUMTH01* sizing markers were 201 and 333bp. The template for the 201bp marker was 2µl of a 1 in 1000 dilution of long PCR1 (LPCR1) of *CYP2A6*. The forward primer was 5'-GCG TGG TAT TCA GCA ACG GG-3' (E3F), reverse primer was 5'-TCG TCC TGG GTG TTT TCC TTC-3' (E3R). The 20µl reaction mixture included 2µl of 1 in 1000 dilution of LPCR1, 0.004µM of forward primer E3F, 0.08µM of reverse primer E3R, 4mM dNTP, 2.0mM $MgCl_2$, 0.08 unit of Taq polymerase. The reaction was carried out in 1x PCR buffer by initial denaturation at 94°C for 2 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 48°C for 30 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 5 minutes. The template of 333bp marker was 2µl of 1 in 100 dilution of long PCR3 (LPCR3) of *CYP2A6*. The forward primer was 5'-TCC CTT GCT GCC TGT GTC CT-3' (CYP2A6*9-F), reverse primer was 5'-GCG AGT TGT ACA

TCT GCT CTG-3' (CYP2A6*9-R). The 20µl reaction mixture included 2µl of 1 in 100 dilution of LPCR3, 0.08µM of each primer, 0.2mM dNTP, 2.0mM $MgCl_2$, 0.02 unit of Taq polymerase. The reaction was done in 1x PCR buffer by initial denaturation at 95°C for 3 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 55°C for 40 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 10 minutes.

Electrophoresis

The 96-well 'stretch' MADGE gel was applied for genotyping the *HUMTH01* microsatellite [361] The 'stretch' gel offers longer track length (43.5mm) compared to the standard MADGE gel (26.5mm) and therefore is suitable for the high resolution genotyping of microsatellites. A gel consisting of 5% Duracryl matrix in 1 x TAE buffer was used. Duracryl™ is a mechanically strong and elastic acrylamide-based matrix which contains longer chain polymers than standard acrylamide. To 50mL gel mixture, 150µl each of TEMED and 25% ammonium persulphate were added as polymerization catalysts.

The gel was equilibrated in 1x TAE buffer containing 5mM KCl for 15 minutes before use to help equalize local ionic conditions near the anode and cathode. A loading buffer was made first consisting of formamide dye mixture, two sizing markers 2:1:1 by volume. *HUMTH01* PCR products were diluted 1 in 6 with formamide dye. 2µl diluted PCR products were added to 2µl loading buffer, and 2µl were then loaded onto the gel using a 96-pin passive replicator (loading once). A second glass plate was then slid carefully over the surface and secured in position with two longitudinally placed rubber bands. A silicone tubing gasket was inserted down each side of the gel, leaving the anode and cathode ends unsealed, to eliminate the distortion of bands at the edge of the

gel due to ionic effects between gel and buffer. The gel "sandwich" was then lowered into the MADGE tank which contained 1x TAE buffer pre-equilibrated to a temperature of 40°C. Electrophoresis was at a constant voltage of 150V (10V/cm) and current of 0.5 A for 110 minutes, during which time the temperature of the buffer was thermostatically controlled at 40°C.

2.4.2.3 Genotyping of *DBH* G1368A

A tetra-primer ARMS-PCR method was used to genotype *DBH* G1368A. The principle of tetra-primer method was described by Ye *et al*. 2001[362]. Figure 2.2 is the schematic diagram of the method.

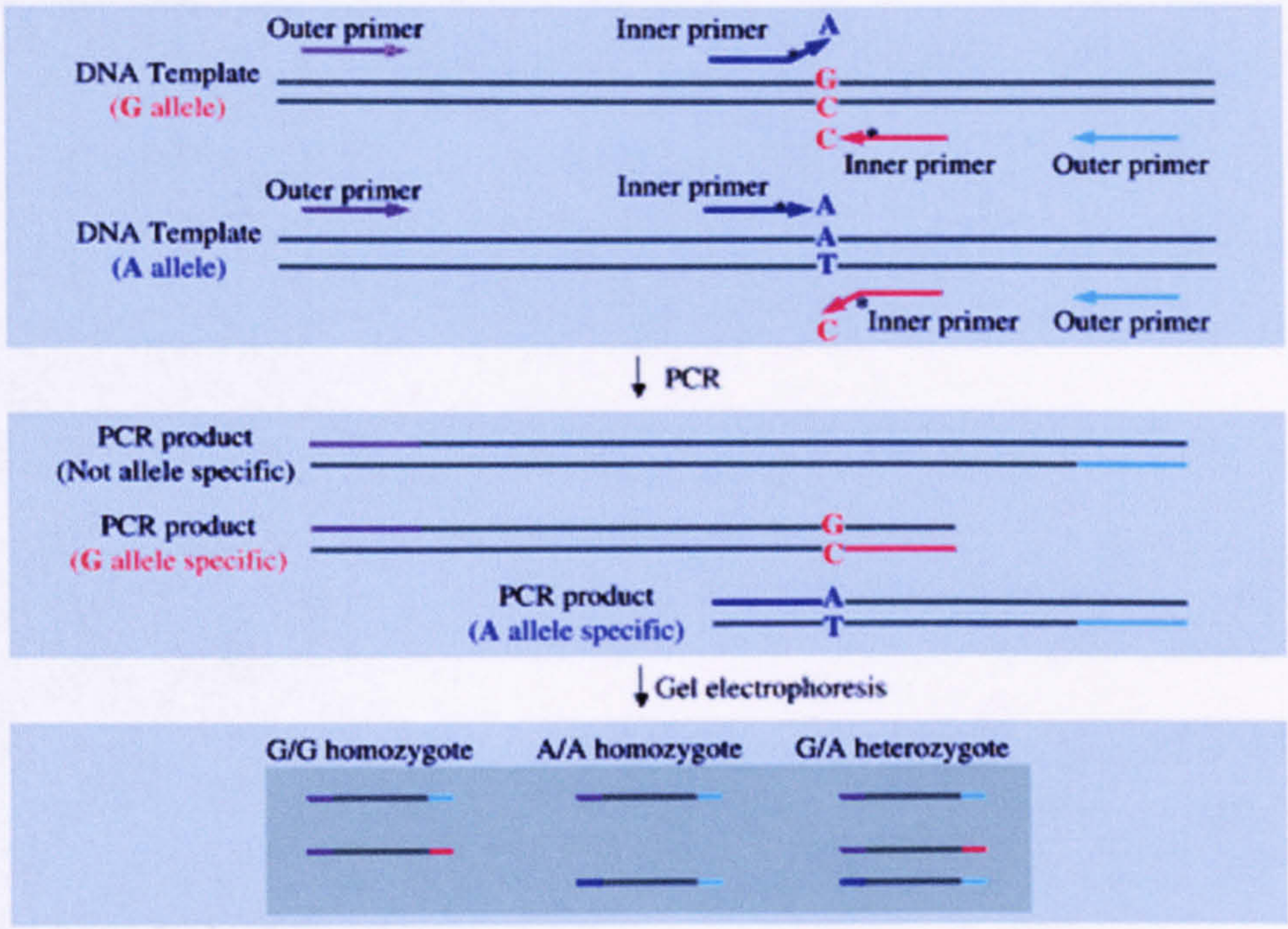


Figure 2.2 Schematic presentation of the tetra-primer ARMS-PCR method. Two allele-specific amplicons were generated using two pairs of primers, one pair (indicated by purple and red arrows, respectively) producing an amplicon representing the G allele and the other pair (indicated by dark blue and light blue arrows, respectively) producing an amplicon representing the A allele. By positioning the two outer primers at different distances from the polymorphic nucleotide, the two allele-specific amplicons differ in length, allowing them to be discriminated by gel electrophoresis [362].

Each PCR reaction was done in a total volume of 10 µl, containing 20 ng of template DNA, 0.002µM of outer forward primer, 0.001µM of outer reverse primer, 0.002µM of inner forward primer, 0.005µM of inner reverse primer, 2mM dNTP, 1.5mM $MgCl_2$, 0.1 unit of Taq polymerase. The reaction was carried out in 1x PCR buffer by initial denaturation at 95°C for 3 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 60°C for 30 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 10 minutes. The PCR products were examined by electrophoresis in 96-MADGE using 5% polyacrylamide gel that was prestained with ethidium bromide, and scanned on Fluorimager® 595.

2.4.2.4 Genotyping of *MAOA T1460C*

An ARMS MADGE assay was used to genotype *MAOA* T1460C. Each PCR was carried out in a total volume of 15 µl, containing 20 ng of template DNA, 0.04µM of each primer, 0.2mM dNTP, 1.5mM $MgCl_2$, 0.1 unit of Taq polymerase. The PCR program was the same with *DBH* G1368A, except that the annealing temperature of 58°C was applied (Table 2.4).

2.4.2.5 Genotyping of *MAOA* VNTR

Genotyping of MAOA VNTR

The *MAOA* VNTR was amplified using forward primer MAO-VNTR-F 5' -ACA GCC TGA CCG TGG AGA AG -3' and reverse primer MAO-VNTR-R 5' -GAA CGG ACG CTC CAT TCG GA -3' as previously described [363]. Each PCR was carried out in a total volume of 15 µl, containing 20 ng of template DNA, 0.04µM of each primer, 0.15mM dNTP, 1.0mM $MgCl_2$, 0.15 unit of Taq polymerase. The PCR program was the

same with *DBH* G1368A, except that the annealing temperature of 68°C was applied (Table 2.4).

Fragment sizing markers

The *MAOA* VNTR sizing markers were 232bp and 454bp. The forward primer of 232bp marker was 232bp-F 5'-ATG ACA AAG GGT GGA GGG ATG A-3', reverse primer 232bp-R 5'-AGA GGG CGA TGT CTT GAT GGA C-3'. The forward primer of 454bp was 454bp-F 5'-GTG CCT GGT GTT TGA GTC CTG TT-3', reverse primer 454bp-R 5'-TGG TGA ATG AAT GAA TGA ATG AAT G-3'. Both PCR mixtures included 20ng genomic DNA, 0.08μM of each primer, 0.2mM dNTP, 2.0mM $MgCl_2$, and 0.02 μl Taq polymerase. The reaction was carried out in 1x PCR buffer by initial denaturation at 95°C for 3 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 62°C for 30 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 10 minutes.

Electrophoresis

Like the *HUMTH01* VNTR, 96-well "stretch" MADGE gel was applied for electrophoresis. However, unlike *HUMTH01*, dry gels without thermal control were used. Formamide dye, 232bp, 454bp and *MAOA* VNTR PCR product 1:1:1:1 by volume, and 4μl were then loaded onto the gel using a 96-pin passive replicator. Electrophoresis was run at the voltage of 120V for 50 minutes.

2.4.2.6 Genotyping of *DRD4* VNTR

Genotyping of DRD4 VNTR

The *DRD4* VNTR was amplified using forward primer DRD4-VNTR-F 5'-GCG ACT ACG TGG TCT ACT CG-3' and reverse primer DRD4-VNTR-R 5'-AGG ACC CTC ATG GCC TTG-3' as previously described [364]. The 10μl reaction mixture included 20ng of template DNA, 0.04μM of each primer, 100μM dATP, dCTP, dTTP, 50μM dGTP, dITP, 1 x buffer (1.5mM $MgCl_2$), 10 x Q solution, 2% DMSO, and 0.2 unit of Hotstar Taq polymerase. The reaction was carried out in an initial 10 minutes denaturation at 95°C, followed by 40 cycles each involving denaturation at 95°C for 1 minute, annealing at 60°C for 40 seconds, extension at 72°C for 2 minutes, and a final extension at 72°C for 10 minutes.

Fragment sizing markers

The *DRD4* VNTR sizing markers were 264bp and 822bp. The template of 264bp marker was 2μl 1/100,000 diluted pGEM3Zf(+). The forward primer was pGEM264bp-F 5'-GGTGAGCAAAAACAGGAAGG-3', reverse primer pGEM264bp-R 5'-GCCTCGTGATACGCCTATTT-3'. The 20μl reaction mixture included 2μl 1/100,000 diluted pGEM3Zf(+), 0.04μM of each primer, 4mM dNTP, 3.0mM $MgCl_2$, 0.2 unit of Taq polymerase (Promega). The reaction was carried out in 1x PCR buffer by initial denaturation at 94°C for 5 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 55°C for 30 seconds, extension at 72°C for 1 minute, and a final extension at 72°C for 5 minutes.

The template of 822bp marker was 20ng genomic DNA. The forward primer was F822 5'-CAA CCA AGA CTA CAA GTA CCG CGT CAG TGA-3', reverse primer R822 5'-

AAC CAG CGG AAG AGG TCA AGG G-3'. The 20μl reaction mixture included 20ng genomic DNA, 0.08μM of each primer, 4mM dNTP, 3.0mM $MgCl_2$, 0.3 unit of Taq polymerase (Gibco). The reaction was carried out in 1x PCR buffer by initial denaturation at 95°C for 10 minutes, followed by 40 cycles each involving denaturation at 95°C for 1 minute, annealing at 70°C for 40 seconds, extension at 72°C for 1 minute, and a final extension at 72°C for 10 minutes.

Electrophoresis of the *DRD4* VNTR was similar to that of the *MAOA* VNTR, except that the formamide dye, 264bp, 822bp and *DRD4* VNTR PCR product were 4:1:1.5:7 by volume.

2.4.2.7 Genotyping of *5HT2A* T102C

The 768 MADGE ARMS were applied for genotyping *5HT2A* T102C. The PCR was carried out in a total volume of 15 μl, containing 20ng of template DNA, 0.06μM of all primers except 0.03μM of reverse primer, 0.2mM dNTP, 1.5mM $MgCl_2$, 0.l unit of Taq polymerase. The PCR program was the same with *DBH* G1368A, except that the annealing temperature of 63°C for T reaction, and 59°C for C reaction were used (Table 2.4).

Table 2.4 Protocols for PCR reactions of dopaminergic genes

PCR recipe	dNTP	MgCl2	Forward primer	Reverse primer	Wild type allele primer	Mutant allele primer	Taq polymerase
MAOA VNTR (Di-primer)	0.15mM	1.0mM	0.4uM	0.4uM			0.1 unit
DRD4 VNTR (Di-primer)	0.01mM dATP, dCTP, dTTP 0.005mM dGTP and dITP	1.5mM	0.4uM	0.4uM			0.15 unit (hotstar taq)
TH01 VNTR (Di-primer)	0.8mM	1.5mM	0.4uM	0.4uM			0.1 unit
MAOA T1460C (Tri-primer ARMS)	0.2mM	1.5mM	0.4uM	0.4uM	0.3uM	0.4uM	0.1 unit
5-HT2A (Tri-Primer ARMS)	0.2mM	1.5mM	0.4uM	0.2uM	0.4uM	0.4uM	0.1unit
DBH G1368A (Tetra-primer ARMS)	0.2mM	1.5mM	0.2uM	0.1uM	0.2uM (inner forward)	0.5uM (inner reverse)	0.1 unit

* Template for each assay was 20ng genomic DNA.
1x PCR buffer was used for each reaction.

Program	Initial denaturation	Denaturation	Annealing	Extension	Cycles	Final extension
MAOA VNTR (Di-primer)	95°C 3mins	94°C 30secs	68°C 30secs	72°C 30secs	35	72°C 10mins
DRD4 VNTR (Di-primer)	95°C 3mins	94°C 30secs	58°C 30secs	72°C 30secs	35	72°C 10mins
TH01 VNTR (Di-primer)	95°C 3mins	94°C 45secs	56°C 40secs	72°C 45secs	35	72°C 10mins
MAOA T1460C (Tri-primer ARMS)	95°C 3mins	94°C 30secs	60°C 30secs	72°C 30secs	35	72°C 10mins
5-HT2A (Tri-Primer ARMS)	95°C 10mins	95°C 1min	60°C 40secs	72°C 2mins	40	72°C 10mins
DBH G1368A (Tetra-primer ARMS)	95°C 3mins	94°C 30secs	62°C (T reaction) 58°C (C reaction) 30secs	72°C 30secs	35	72°C 10mins

2.4.3 Genotyping growth pathway genes

2.4.3.1 Microsatellite marker *CSH1.01* and *IGF1*(CA)n

Microsatellite marker *CSH1.01* (size range, 220-311bp) was coamplified with an *IGF1*(CA)n microsatellite (size range, 175-199bp) using primers 5'-ACT GCA CTC CAG CCT CGG AG-3' and 5'- ACA AAA GTC CTT TCT CCA GAG CA-3' for *CSH1.01* and 5'- GCT AGC CAG CTG GTG TTA TT-3' and 5'- ACC ACT CTG GGA GAA GGG TA-3' for the *IGF1* marker. One primer of each pair was labelled with

the same fluorescent dye. For some subjects we used 6-carboxyfluorescein, for some we used HEX, and for some we used NED. Post PCR pooling of three subjects enabled two-size, three-colour multiplexing for ABI377 size calling.

In order to save stock DNA, degenerate oligo primer amplifications (DOP-DNA) were made from 20ng genomic DNA using partially degenerate sequence: 5'-CCG ACT CGA G(AGCT) (AGCT) (AGCT) (AGCT) (AGCT)(AGCT)AT GTG G-3' The DOP protocol was a modified version of the method used by Cheung and Nelson [365]. The reaction mixture (50µl) contained 5µl 10 x long PCR Buffer, 0.25mM dNTP, 2.5mM $MgCl_2$, 1µM DOP primer, 1.3M Betaine, 0.008 unit PWO, and 2.5 unit Taq polymerase. The amplification was performed by initial denaturation at 94°C for 5 minutes, followed by 8 cycles each involving denaturation at 94°C for 1 minute, annealing at 30°C for 1 minute, extension at 72°C for 3 minute, and then followed by 28 cycles of denaturation at 94°C for 1 minute, annealing at 60°C for 1 minute, extension at 72°C for 3 minute.

A 2µl aliquot DOP-DNA, representing 0.8ng original genomic DNA, was used for genotyping microsatellite markers *CSH1.01* and *IGF1* (CA)n. The reaction mixture (10 µl) contained 1x PCR Buffer, 0.2mM dNTP, 1.5mM $MgCl_2$, 0.5µM IGF1-F and IGF1-R, 0.4µM CSH1.01-F and CSH1.01-R, 0.8 unit Taq polymerase. The amplification was performed by initial denaturation at 95°C for 3 minutes, followed by 35 cycles each involving denaturation at 94°C for 30 seconds, annealing at 61.5°C for 30 seconds, extension at 72°C for 30 seconds, and a final extension at 72°C for 10 minutes. The PCR products were checked by electrophoresis using 5% polyacrylamide gel that was prestained with ethidium bromide.

An ABI Prism™ 377 was used for size calling, which included three major steps: 1) generation of an ABI 377 gel matrix; 2) genotyping using ABI Prism™ 377 DNA Sequencer; 3) post-run preliminary analysis using Genescan® 3.1 and Genotyper® 2.1.

A gel matrix file suitable for the samples was generated by running 6-FAM, HEX, NED and internal-lane size standard GS-400HD ROX size STD (P/N 402985) in individual lanes of the ABI 377. A long (36cm) gel was formed by mixing 35ml Gene-Page Plus gel mixture (5.0% polyacrylamide with 6M urea, Amresco), 35µl TEMED, 350µl 10% APS. The mixture was injected between two crystal quartz ABI 377 plates and left to polymerise for 2 hours. 6µl loading buffer (4µl formamide, 1µl blue dye, 1µl GS-400HD ROX size STD was mixed with 2µl 1/5 diluted PCR samples, denatured at 95°C for 5 minutes and then chilled on ice. 1µl of each denatured samples was loaded into separate lanes of the ABI gel. Virtual filter set D was selected and program "GS Run 36D 2400" was used. Genescan® 3.1 and Genotyper® 2.1 software were used to view and process raw data, which were then exported to Microsoft Excel® for further analysis.

2.4.3.2 *Angiotensin converting enzyme (ACE)* insertion/deletion

A tri-primer PCR method was optimised based on the method previously reported [366] to genotype *ACE* gene insertion/deletion. Each PCR was carried out in a total volume of 10 µl, containing 20ng of template DNA, 0.06µM of primer aceid-1(5'- CAT CCT TTC TCC CAT TTC TC-3') and aceid-3 (5'- ATT TCA GAG CTG GAA TAA AAT T-3'), 0.03µM of aceid-2 (5'- TGG GAT TAC AGG CGT GAT ACA G-3'), 2mM dNTP, 2.0mM $MgCl_2$, 0.15 unit of Taq polymerase. The reaction was carried out in an initial 2 minutes denaturation at 95°C, followed by 36 cycles each involving denaturation at 95°C for 30 seconds, annealing at 53°C for 30 seconds, extension at 72°C for 30

seconds, and a final extension at 72°C for 10 minutes. The PCR products were examined by electrophoresis using 5% polyacrylamide MADGE gel that was prestained with ethidium bromide. Primers aceid-1 and aceid-3 were chosen outside the insertion sequence and primer aceid-2 was chosen in the insertion (see primers sequences and bands sizes in Figure 2.3). If the insertion exists, primer aceid-2 and aceid-3 will form a 65bp band. If there is no insertion, sequences underlined are deleted; primer aceid-1 and aceid-3 will generate an 83bp band.

Part of intron 16 of ACE gene on chromosome 17q23

```
1321 tcaagcacgc ccctcacagg actgctgagg ccctgcaggt gtctgcagca tgtgcccagg
1381 ccggggactc tgtaagccac tgctggagac cactcccatc ctttctccca tttctctaga
1441 cctgctgcct atacagtcac ttttttttct tttttgagac ggagtctcgc tctgtcgccc
1501 aggctggagt gcagtggcgg gatctcggct cactgcaacg tccgcctccc gggttcacgc
1561 cattctcctg cctcagcctc ccaagtagct gggaccacag cgcccgccac tacgcccggc
1621 taatttttg tatttttagt agagacgggg tttcaccgtt ttagccggga tggtctcgat
1681 ctcctgacct cgtgatccgc ccgcctcggc ctcccaaagt gctgggatta caggcgtgat
1741 acagtcactt ttatgtggtt tcgccaattt tattccagct ctgaaattct ctgagctccc
1801 cttacaagca gaggtgagct aagggctgga gctcaagcca ttcaaccccc taccag
```

```
aceid-1 atcctttctcccatttctct
aceid-2 tgggattacaggcgtgatacag
aceid-3 aattttattccagctctgaaat
Deletion -- 83bp band
Insertion-- 65bp band
```

Figure 2.3 *ACE I/D* genotyping primers design and bands sizes. Underlined bases represent the 287bp fragment that is either inserted or deleted. Pink, blue and red base pairs represent three primers. An 83bp band or 65bp band would be detected when the 287bp fragment is deleted or inserted.

2.4.3.3 *GH-CSH Bgl IIB*

The GH-CSH BglII-B sequence was determined by sequencing the restriction-fragment length polymorphism site at nucleotides 51984–9 in GenBank sequence J03071.1 (single nucleotide polymorphism sequence: CCTTCAAG[A/G]TCTATTGGG at

nucleotide 51985). The *GH-CSH Bgl IIB* genotype was determined using fluorescence-labelled oligonuclotide melting from matched or mismatched target, monitored in an Idaho Technology 384-well Odyssey. Detection utilised reduction of opposed G-base quenching of fluorescence during a thermal ramp. Asymmetric PCR was performed on 20ng genomic DNA template in 384-well white PCR plates on a MJ Research PTC-225 DNA Engine Tetrad®. Samples were amplified with primers *GHBglIIB-F* CGCGTGGCATCCTGTGAC at 500nM and *GHBglIIB-RS2* GGAATGGCCTTAGTGAGTAATAGCG at 100nM. A fluorescein isothiocyanate (FITC) labelled probe (with 3' phosphate) *BglIIB-OPr1* 5'-FITC-CAATAGATCTTGAAGGCCC-PHO-3' and a Dabcyl probe *BglIIB-Dab1* 5'-AAGCCTCTTATATTTCAGCGGCCTAGTT-DAB-3' was included at 200nM each in the PCR for the Odyssey melting assay. 5 µl PCR reaction mix also contained: 1x PCR buffer, 0.2mM dNTPs, 2 mM $MgCl_2$ and 0.1 units of Taq DNA polymerase. PCR cycling conditions were: 94°C for 2 minutes then 75 cycles of 94°C for 20 seconds, 68°C for 30 seconds and 72°C for 30 seconds, followed by 72°C for 2 minutes. Samples were overlaid with 5µl Chill-Out™ wax (Genetic Research Instrumentation) to prevent evaporation during analysis. Following PCR, amplification samples were melted from 40°C to 75°C in the 384-well Odyssey. LightTyper software was used to analyse the change in fluorescence during melting, and to group melting profiles into genotype groups. These were then manually verified using in-house software.

2.4.4 Image analysis

2.4.4.1 384 ARMS MADGE

The 384 ARMS MADGE image from Fluorimager® 595 was analysed in 'MADGE mode' of the Phoretix 1D Advanced Image Analysis Software (www.phoretix.com). As shown in Figure 2.5 the 384-MADGE system is composed of four superimposed 96-well arrays. In the system, the second (yellow) and the fourth (purple) sets of wells were located half-way along the tracks of the first (red) and the third (green) sets of samples, so that two 96 well arrays of allele-specific reactions in ARMS can be run in series, e.g. array 1 and array 2 were used for two reactions (wild type and mutant) of one array, array 3 and array 4 for another array. The two allele-specific tracks were treated as one lane when the sample was called. The identification of genotype can be judged by allele band patterns on the ARMS MADGE image (Figure 2.5 and Figure 2.6).

192-well MADGE and 768-well MADGE were also used for genotyping according to the size of the PCR product. The 192-well MADGE was applied when the PCR fragment was relatively long, but with lower efficiency per gel (one array, two reactions). The 768-well MADGE was of higher efficiency per gel (4 arrays, 8 reactions) but with the limitation of certain range of PCR fragments.

2.4.4.2 Microsatellite image analysis

Take *HUMTH01* VNTR for example, after electrophoresis, the glass plate covering the gel was removed and the gel was stained with Vistra Green™ dye. The wet gel was then scanned on a Fluorimager® 595. The *HUMTH01* PCR fragments were sized with reference to two marker bands which span the allele size ranges. Gel tracks were extracted for analysis using Phoretix™ software. Measurements generated by the

Phoretix software were transferred to microsoft Excel® for PCR fragment size computation. The formula for converting the size of bands in centimetre or pixel distance into molecular weight in base pairs of nucleotides is as follow:

$$U = \frac{LMW}{(LMW/HMW)^{[(d'LMW-d'U)/d'LMW]}}$$

The formula represents log-linear relationship of allele migration according to size (bp). Where "U" is the unknown allele size in bp, LMW is the lower molecular weight marker (bp) and HMW the upper marker (bp). d'LMW is the distance between the LMW marker and HMW marker, and d'U the distance between the unknown allele and HMW marker. For *HUMTH01*, LMW is 201, HMW is 333.

Determination of both *MAOA* VNTR and *DRD4* VNTR fragment size were by the same formula but with different sizing marker values substituted.

2.4.5 Dilution and aliquoting

The aliquoting of 96 well-plate genomic DNAs from working stocks was done manually using 8 channel Finnpipette and yellow tips. The aliquoting of 96 well DOP-PCR was also done manually. The 96 well plate of 8kb long PCR to 384 well plate 1 in 100 dilution and aliquoting, genomic stock DNA to 384 well plate aliquoting, and the 384 well plate of 1.9kb LPCR2 and 2.3kb LPCR3 to 384 well plate 1 in 100 dilution and aliquoting were processed by the aliquoting robot QUADRA96® SV MODEL 345 (6-Station Automatic Shuttle), using appropriate QUADRA programs.

2.5 Statistical analysis

2.5.1 Statistical analysis of smoking markers

Microsoft Excel® was used for the analysis of allele and genotype frequency distribution. Hardy-Weinberg equilibrium was examined using χ^2 test. All phenotypic analyses against genotypes were performed by a statistician, Dr Julie Gilg, Department of Community Health Sciences, St. George's Hospital Medical School, London, UK. All phenotypic analyses were carried out in SAS (version 8.1). At each age, and for each genotype, three odds ratios were calculated using the LOGISTIC procedure: (i) current or ex-smoker versus never smoker; (ii) current versus ex-smoker; (iii) current versus never or ex-smoker. All analyses were adjusted for age, sex, town and ethnicity. Which comparison is most appropriate depends on the hypothesis being considered. Thus, for a genotype which it is hypothesized relates to experimentation with smoking, it would be appropriate to use the current+ex-smoker versus never comparison; for a genotype which it is hypothesized influences addiction, such as *CYP2A6*, it is appropriate to focus on current versus ex-smoker since the never group are not at risk. The comparison current versus ex-smoker+never is presented for comparison with other studies since this is what has usually been analysed. Mean levels of addiction markers amongst current smokers by *CYP2A6* genotype were produced adjusted for age, sex, town and ethnicity using the LSMEANS option of the GLM procedure.

For convenience of referring to the genetic subgroups, we adopted a nomenclature as follows: WT1A (*1A/*1A homozygotes), wild-type; WT1B (*1B carriers) with possibly faster cotinine formation than *1A group; HI (haploinsufficiency) (*2, *4 and *5 carriers) with one allele inactive; and PHI (partial haploinsufficiency) (*9 and *12 carriers) with one allele part inactive. Subjects were pooled into these four groups on

the basis of the known functional effects of the alleles on metabolism of nicotine to cotinine in order to conserve power for the analyses of the relatively rare *CYP2A6* alleles.

Table 2.5 *CYP2A6* allele grouping, coding and predicted characteristics

Name of group	Short code for group	Content of group	Predicted characteristic of group
*1 wildtype (*1B carrier)	WT1B	*1B carriers	possibly faster cotinine formation than *1A group
*1 wildtype (*1A/*1A)	WT1A	*1A homozygotes	wild type
partial haploinsufficiency	PHI	*9 and *12 carriers	one allele partially inactive$
haploinsufficiency	HI	*2,*4 and *5 carriers	one allele inactive

$. Rarely, both alleles partially inactive.

In the analyses of the *TH01* gene, according to previous literature, subjects with at least one 9-repeat allele were compared with all other cases, and similarly those possessing at least one 7-repeat allele were compared with all other cases. For the *DRD4* VNTR, those with at least one long allele (6-repeat or longer) were compared with all other cases.

MAOA gene analyses were carried out separately for males and females as this gene is on the X chromosome. Haplotype (diplotype) frequencies of *MAOA* T1460C and *MAOA* VNTR were calculated using Microsoft Excel®.

2.5.2 Statistical analysis of growth pathway genes

Dichotomisation of *CSH1.01* was used to define a large allele class, as shown in Figure 3.3, each of 10 major alleles being one tetranucleotide repeat different in length from the next, as a 'T' group; a small allele class, each of 14 major alleles is a dinucleotide repeat different from the next, as a 'D1' allele group; and a biphasic harmonic of allele frequencies within group D, removal of five higher frequency alleles spaced at four base

intervals leaves a subgroup of D1, as a 'D2' group. Analyses were then performed on genotype groups T/T vs. D1/T vs. D1/D1 and on T/T vs. D2/T vs. D2/D2, the latter after the exclusion of subjects possessing one or more alleles not in T or D2 categories.

Genotype classification of *IGF1*(CA)n adopted the prevailing approach, ie. to group allele lengths which are not 192bp, and to compare 192/192; not 192/not192; and heterozygotes. Since there is strong sexual dimorphism within the growth pathways, males and females were analysed separately for *CSH1.01* and *IGF1*$(CA)_n$.

Microsoft Excel® was used for counts and histogram plots. The BETA program (Collins, www.cedar.genetics.soton.ac.uk) was used to test for Hardy-Weinberg equilibrium.

The 2LD program (Jing hua Zhao (2002) 2LD -- two-locus linkage disequilibrium (LD) calculator,
http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.shtml) was used to calculate pairwise linkage disequilibrium between each marker pair at two loci. Considering two adjacent loci A and B, with two alleles (A, a, and B, b) at each locus, the observed frequency of the haplotype that consists of alleles A and B is represented by f_{AB}. Assuming independent assortment of alleles at the two loci, the expected haplotype frequency is calculated as the product of the allele frequency of each of the two alleles, or $f_A \times f_B$, where f_A is the frequency of allele A at the first locus and f_B is the frequency of allele B at the second locus. The linkage disequilibrium coefficient $D_{AB}= f_{AB} - f_A \times f_B$. The intensity of pairwise disequilibrium was calculated by use of the D′ coefficient, defined as D_{AB}/D_{max}, D_{max} is min $[f_A f_b, f_a f_B]$, when D>0, or min $[f_A f_B, f_a f_b)]$, when D<0. D′ ranges from -1 to +1. The significance of nonrandom associations

was also calculated by the program 2LD according to the formula $\chi_{AB}^2=2ND^2/f_Af_af_Bf_b$, where N is the sample size, or $\chi_{AB}^2=2ND^2/f_A(1- f_A)f_B(1-f_B)$ if they are multiallelic.

Haplotype Trend Regression (HTR) was used to associate phenotypic traits with haplotype frequencies of individuals. HTR estimates the haplotype frequencies by use of the expectation–maximisation (EM) algorithm and then relates the inferred haplotype frequencies to the observed phenotype using a regression model [367]. This approach gives the average trait value for each haplotype with a significance for that-allele against all-others-combined. Assuming additive effects of the haplotypes on phenotypic traits, the HTR approach was used to test for the contribution of individual haplotypes. Analysis using a linear regression model (Stata 8.2) was performed to test the contributions of individual markers.

2.5.3 Power calculations

Power calculations were performed using the 'two sample t test' of Java applets for power and sample size [368](http://www.stat.uiowa.edu/~rlenth/Power). The calculator window is shown below (Figure 2.4). The basic data needed for power calculation include standard deviation (SD) of each genotype group, sample size, desired significance threshold (e.g. 0.05), difference of means between genotype groups.

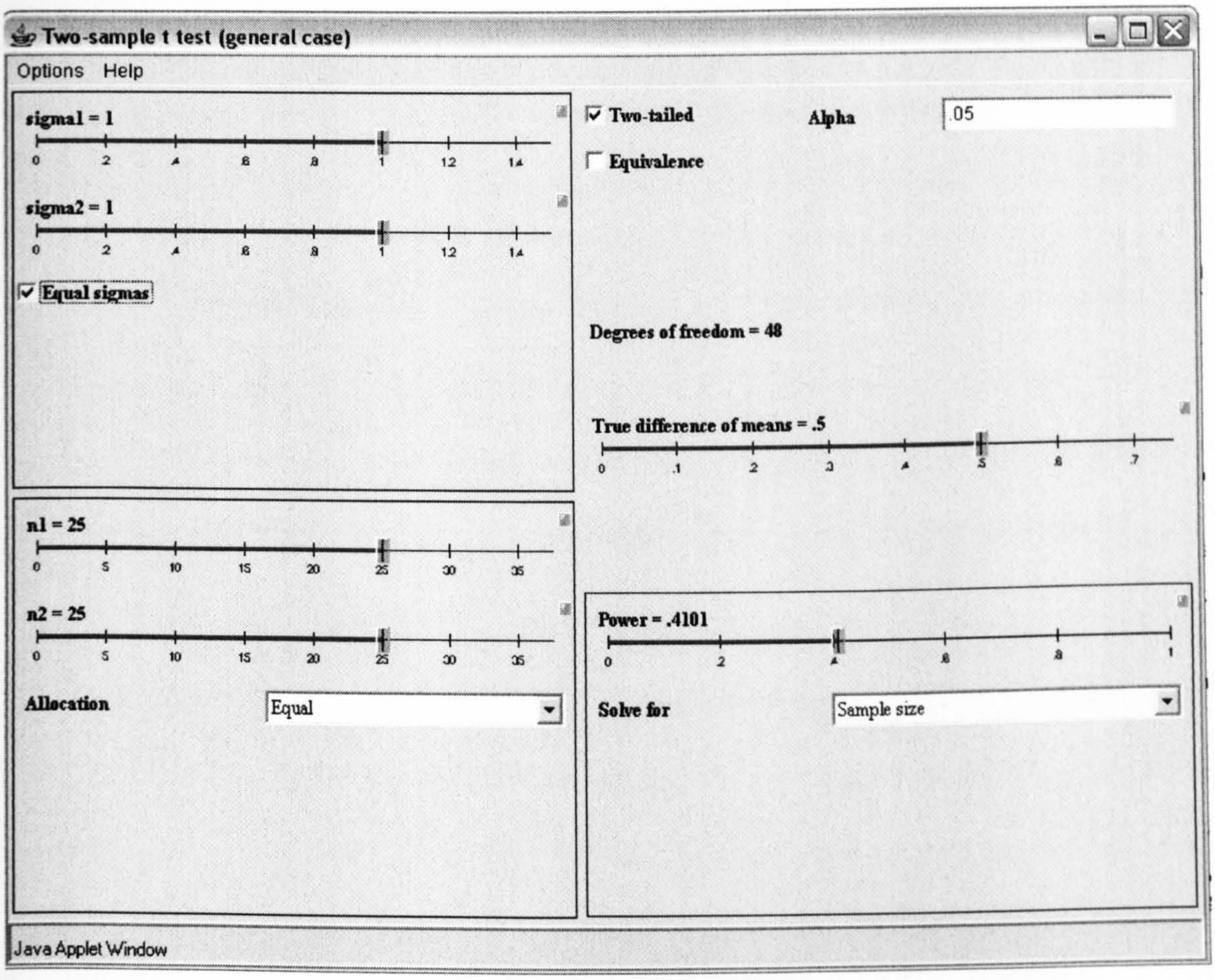


Figure 2.4 Power calculator window for continuous quantitative trait analyses. Sigma 1 and sigma 2 are the standard deviations (SD) of the trait in the two genotype groups. The 'allocation' option could be changed to 'independent' if two differentlly-sized groups are used. Either check or uncheck the two-tailed box, depending on the nature of the hypothesis and set alpha to the desired significance threshold (e.g. 0.05). Then set "solve for" to "diff of means" and type in the true difference of means. The power of the analysis is then given by the program. Another usage is to set the desired power (e.g 0.8), and calculate the sample size needed for each genotype group.

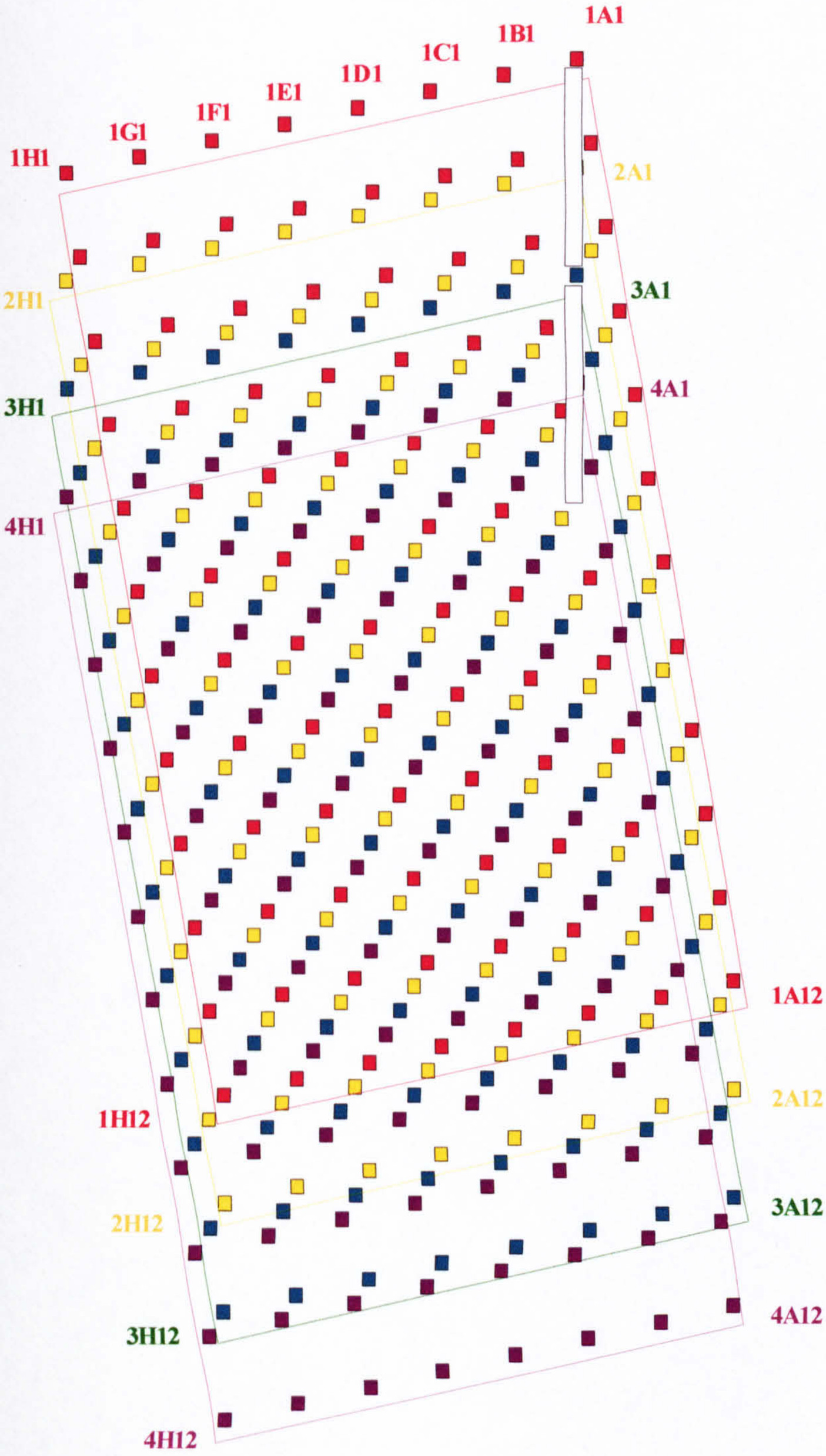
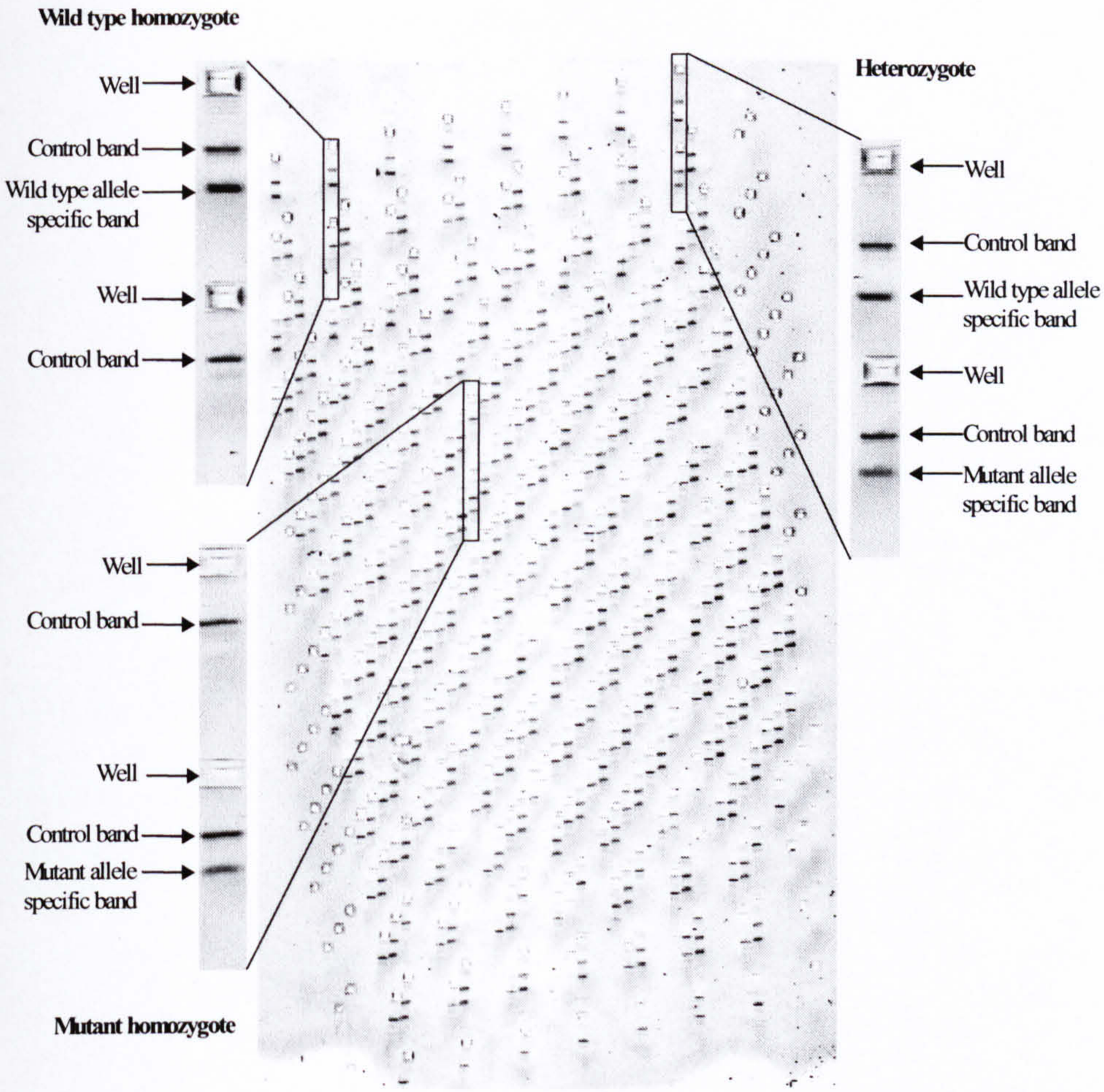


Figure 2.5 Schematic illustration of the 384-well MADGE system. The four colours (red, yellow, green and purple) show that the system is composed of four superimposed 96-well MADGE arrays. The coordinates following this: A1-H12 are those of the original 96-well plate. The numbers 1-4 before A-H represent array 1-array 4. In ARMS, array 1 and array 2 are used for two reactions (wild type and mutant reaction) of one array. Array 3 and array 4 for another array. The two allele-specific tracks were treated as one lane when the sample was called.



B

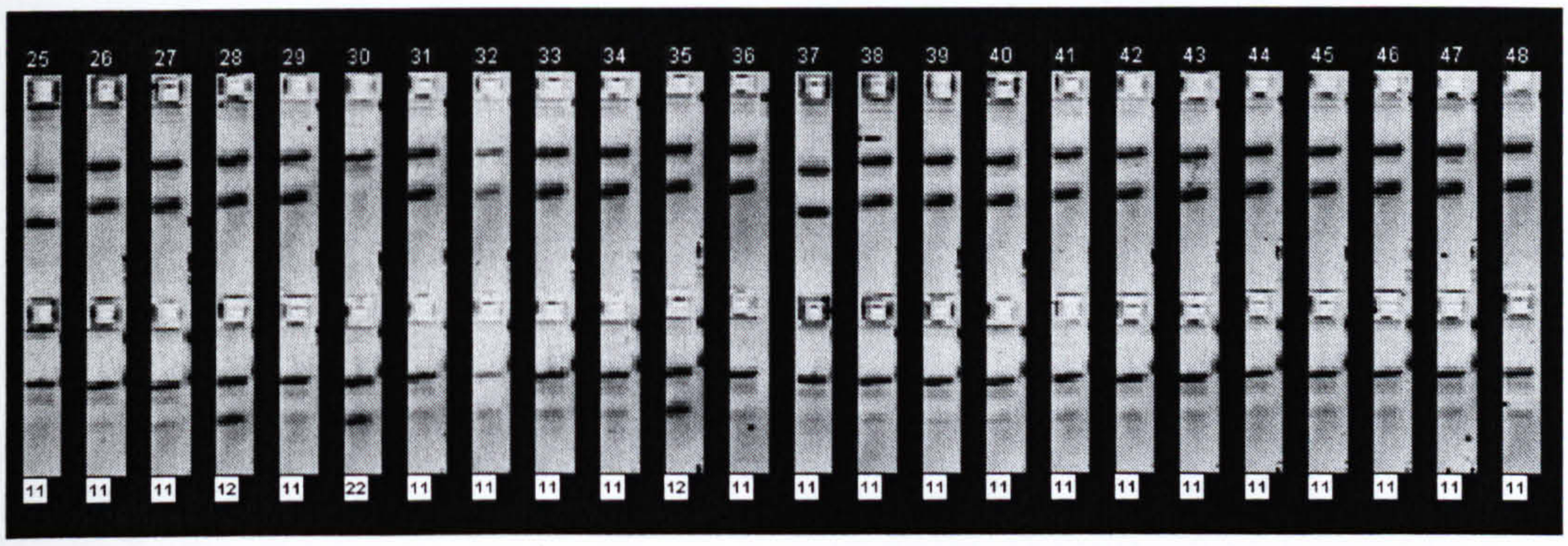


Figure 2.6 ARMS-MADGE for genotyping *CYP2A6*1A* and **2*. A. 384 ARMS-MADGE image. *CYP2A6*2* is a T to A L160H mutation. ARMS PCR was performed using an outer pair of primers (E3F and E3R), giving a 201bp product (control band), and allele specific primers (2A6wt and 2A6mut) as forward primers, together with reverse primer E3R to get a 60bp product. B. The gel image is imported into Phoretix 1D Advanced image-analysis software so that individual lanes can be created, facilitating allele calling. 11 represents wild type homozygote, 12 indicates heterozygote, and 22 is mutant homozygote.

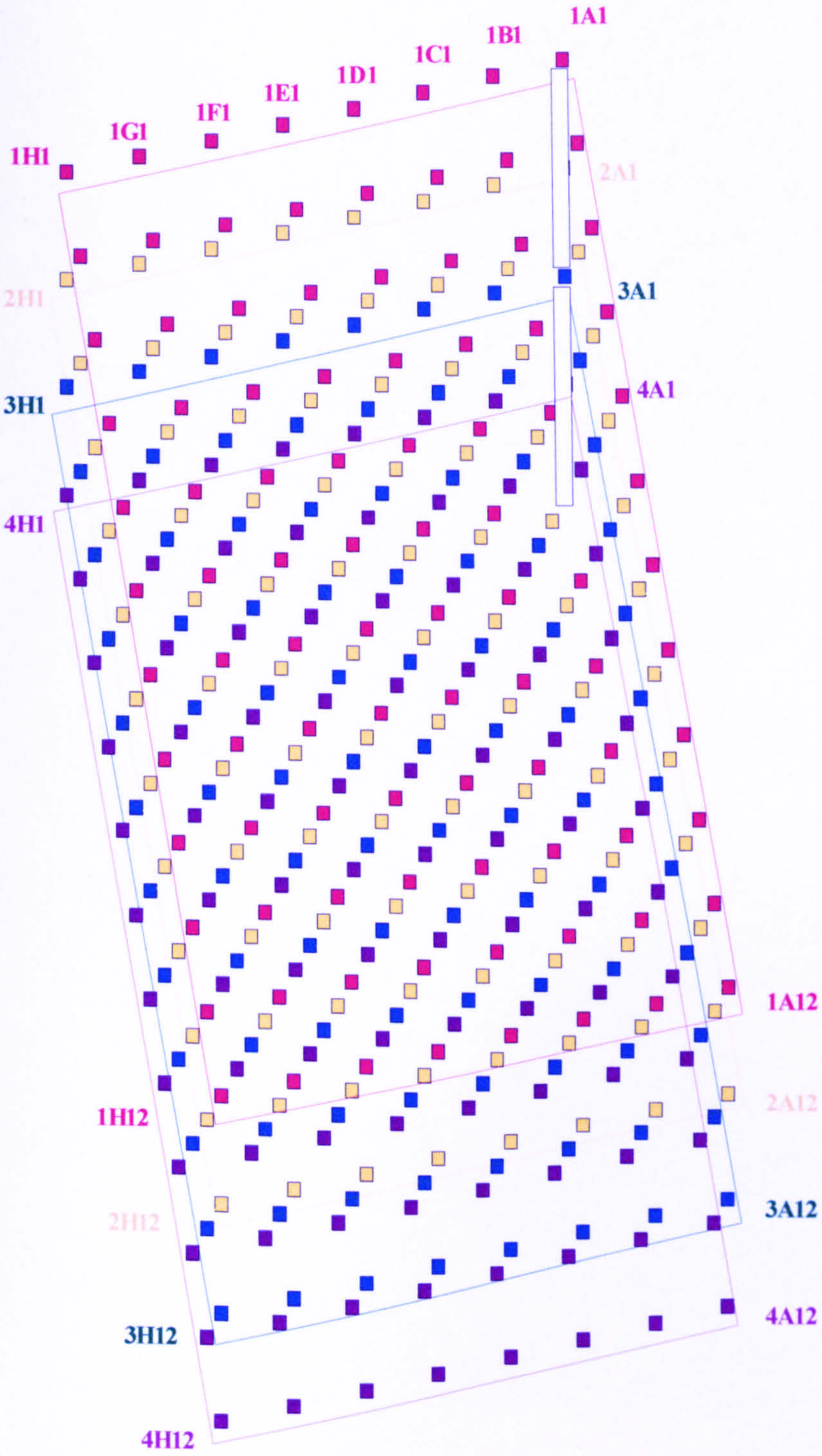
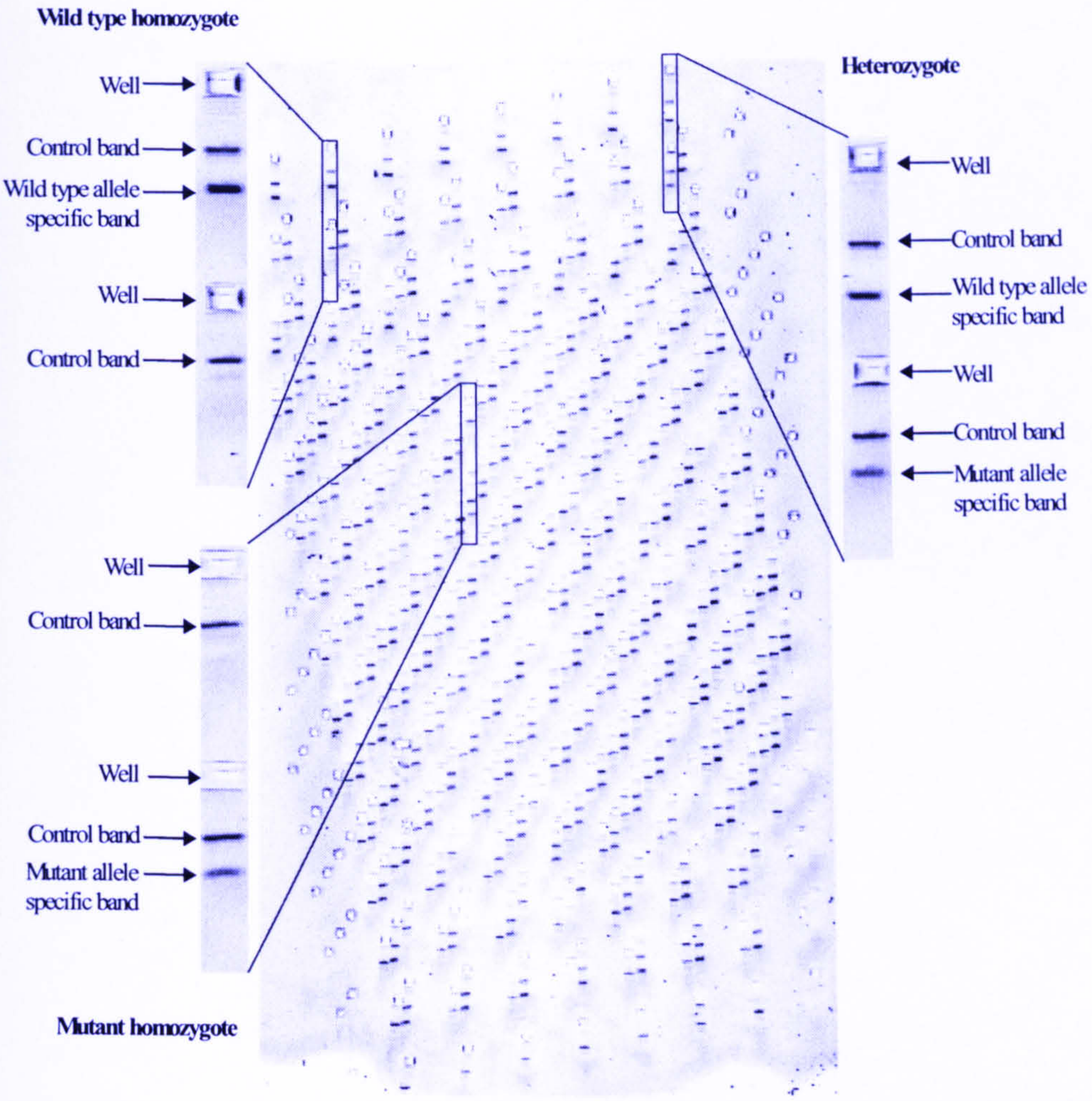


Figure 2.5 Schematic illustration of the 384-well MADGE system. The four colours (red, yellow, green and purple) show that the system is composed of four superimposed 96-well MADGE arrays. The coordinates following this: A1-H12 are those of the original 96-well plate. The numbers 1-4 before A-H represent array 1-array 4. In ARMS, array 1 and array 2 are used for two reactions (wild type and mutant reaction) of one array. Array 3 and array 4 for another array. The two allele-specific tracks were treated as one lane when the sample was called.



B

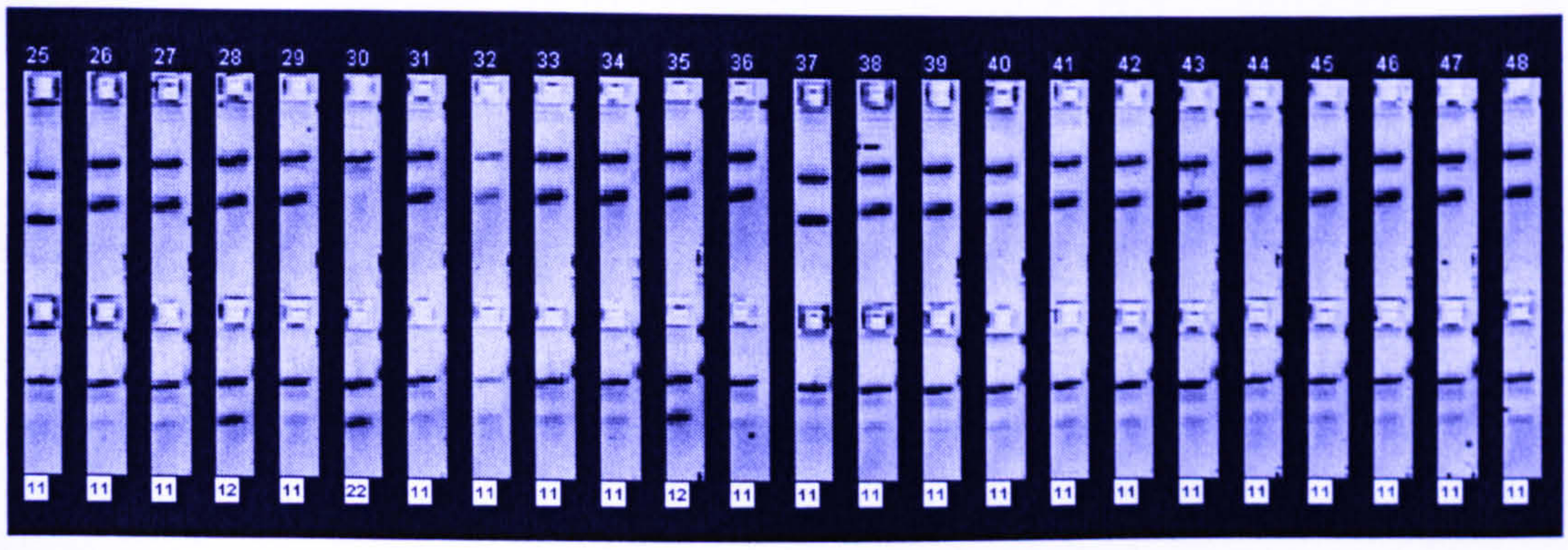


Figure 2.6 ARMS-MADGE for genotyping *CYP2A6*1A* and **2*. A. 384 ARMS-MADGE image. *CYP2A6*2* is a T to A L160H mutation. ARMS PCR was performed using an outer pair of primers (E3F and E3R), giving a 201bp product (control band), and allele specific primers (2A6wt and 2A6mut) as forward primers, together with reverse primer E3R to get a 60bp product. B. The gel image is imported into Phoretix 1D Advanced image-analysis software so that individual lanes can be created, facilitating allele calling. 11 represents wild type homozygote, 12 indicates heterozygote, and 22 is mutant homozygote.

Chapter 3 Results

The TTHHS is a mixed longitudinal study that examines many different elements of the health of young people which are relevant to heart health. Phenotypic data collected in the TTHHS includes smoking status, smoking behaviour, and saliva cotinine level at both ages of 13-15 y and 18y. Growth and metabolic traits were collected when they were 13-15y, including, birthweight, height and weight, serum glucose and insulin levels. The aim of my study was to investigate genetic variation in candidate genes in relation to adolescent smoking behaviour and growth and metabolic traits. To initiate this project, a DNA bank of 1520 adolescents in the TTHHS was set up and a quality control procedure established. A total of 17 markers were investigated. Candidate genes for the smoking study included those belonging to the nicotine metabolism pathway: *CYP2A6 (*1A, *1B, *2, *4, *5, *9* and **12)*, the dopaminergic pathway genes variants *DBH*G1368A, *MAOA* T1460C and VNTR, *TH01*VNTR, *DRD4* VNTR and the serotonine receptor gene *5HT2A* T102C variant. Two microsatellite markers *CSH1.01* and *IGF1*$(CA)_n$ were genotyped in relation to growth and metabolic trait in the TTHHS. In order to investigate the association between a haplotype of growth hormone and *ACE* genes, a separate cohort (a component of the Lichfield study), which was set up by the UCL Centre for Cardivascular Genetics, was analysed for *ACE* I/D and *GH-CSH* BglII.

3.1 DNA bank

3.1.1 DNA concentration and yields

A critical compenent of this study was the development of a high quality DNA bank from the frozen samples that were available. DNA yields are dependent on the volume of blood and the number of white blood cells, sample storage and extraction method.

The normal white cell count ranges from 4 x10^6/ml to 11x10^6/ml of whole blood and each white cell contains 6pg of DNA. The theoretical DNA yield is thus a maximum of 24-66µg per millilitre of blood. For practical reasons, large studies necessitate short or long term storage of samples at -20°C or preferably -80°C. Freezing samples reduces the recovery of DNA.

Of the 1530 samples that were processed from the TTHHS cohort, DNA was successfully recovered from 1520. The DNA concentrations were between 10ng/µl and 1031ng/µl. The DNA was usually dissolved in 500µl TE buffer; therefore, the total yields were between 5µg and 516µg. 90.9% of the yields were over 50µg with most of these being between 50µg and 200µg.

3.1.2 Sex test applied to the DNA banks

As shown in Figure 3.1, there were two possible band patterns from the duplex PCR products, either one band or two bands. The 211-bp amplicon from *MAOA* was from the X chromosome and therefore appears in both males and females, thus effectively acting as a control band for the amplification reaction. The 320-bp amplicon from *DDX3Y* was Y chromosome specific, and would therefore only be observed when the subject was a male. Therefore, in each track of the gel, two bands indicate that the subject is genotypically male, while the absence of the upper 320- bp band indicates a female subject.

3.1.3 Sex comparison between genotype data and phenotype data

The total numbers of males and females were the same between known phenotype and quality control genotype data. All cases were concordant except for 12 subjects (6 males and 6 females) for whom the genders were opposite between the phenotype and

genotype data, For these 12 subjects, every two subjects of each pair were adjacent to each other on the DNA extraction preparation (Prep) code list (Table 3.1).

Table 3.1 Gender disagreement between phenotype and quality control genotype data

	Prep code number	Phenotype gender	Genotype gender
Pair 1	G0511	M	F
	G0512	F	M
Pair 2	G0528	M	F
	G0529	F	M
Pair 3	G0595	F	M
	G0596	M	F
Pair 4	G1045	M	F
	G1046	F	M
Pair 5	G1115	F	M
	G1116	M	F
Pair 6	G1125	F	M
	G1125	M	F
Total number		12	12

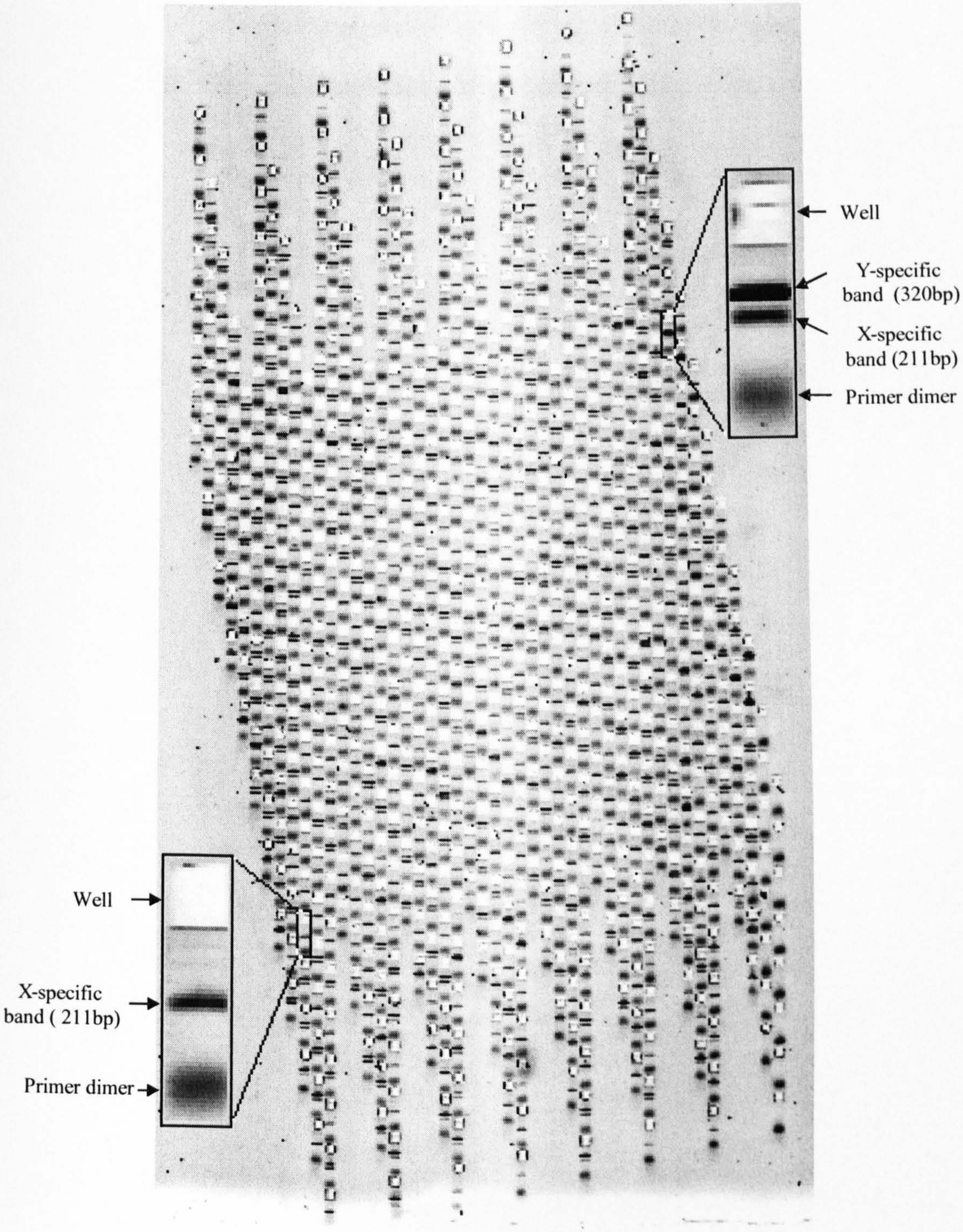


Figure 3.1 768-Well MADGE. Eight 96-well arrays (768 subjects) were genotyped in one gel. In each track of this gel, the two bands of 211bp (from the X chromosome) and 320bp (Y chromosome specific) indicated the subject as a male; while the absence of the upper 320bp band indicates a female subject.

3.1.4 Sex confirmation combining sex test with *MAO-A* genotypes

Two markers in the *MAOA* gene, a SNP (T1460C) in exon 14 and a VNTR in the promoter region, originally designed for the smoking study were used to help to resolve these mixed samples. *MAOA* is on the X chromosome and therefore only females could be heterozygous for markers in this gene. The table below shows the genotypes of *MAOA* markers (T1460C and VNTR) for those discordant gender subjects. Genders judged from these *MAOA* genotypes were consistent with those from quality control assay, ie all three *MAOA* SNP heterozygotes and five VNTR heterozygotes were assigned as females by the genotype assay (Table 3.2).

Table 3.2 Genotype of *MAO* T1460C and *MAO*-VNTR for gender disagreement subjects

Prep code number	Phenotype data	Genotype data	Genotype *MAO* T1460C	Genotype *MAOA* VNTR
G0511	Male	Female	TC	370/340
G0528	Male	Female	TC	370/340
G0596	Male	Female	TT	400/340
G1045	Male	Female	TC	355/340
G1116	Male	Female	CC	370/340
G1126	Male	Female	TT	340/340
G0512	Female	Male	T	340
G0529	Female	Male	C	370
G0595	Female	Male	C	340
G1046	Female	Male	T	340
G1115	Female	Male	T	340
G1125	Female	Male	T	340

The phenotype data were checked and no errors were found. Since all of the mismatches occured in pairs of adjacent samples on the Prep code list the errors probably occurred

during the DNA extraction process. These samples were reinstated in the correct preparation code order.

3.2 Genotype and allele frequencies of smoking markers

Altogether thirteen markers for six genes were genotyped, including *CYP2A6*1A, *1B, *2, *4, *5, *9,* and **12*; *DBH* G1368A; *MAOA* T1460C and VNTR; *TH01* VNTR; *DRD4* VNTR and *5HT2A* T102C using the methods described in Chapter 2. Of the 13 markers tested, between 77.8% (*TH01* VNTR) and 99.6% (*MAOA* T1460C) of cases were successfully genotyped.

3.2.1 Genotype and allele frequencies of *CYP2A6*

CYP2A6 designations are summarised in Figure 3.2. *CYP2A6*12* and *CYP2A6*9* were analysed separately.

3.2.1.1 Genotype and allele frequencies by gender

There were 1520 subjects who consented to be investigated in the TTHHS cohort, among which there were 839 (55.2%) males and 681 (44.8%) females. The genotype and allele frequencies by gender are shown in Table 3.3 and Table 3.4. χ^2 testing (with or without Yates correction) showed that there was no significant difference between male and female subjects both for genotype frequencies and allele frequencies.

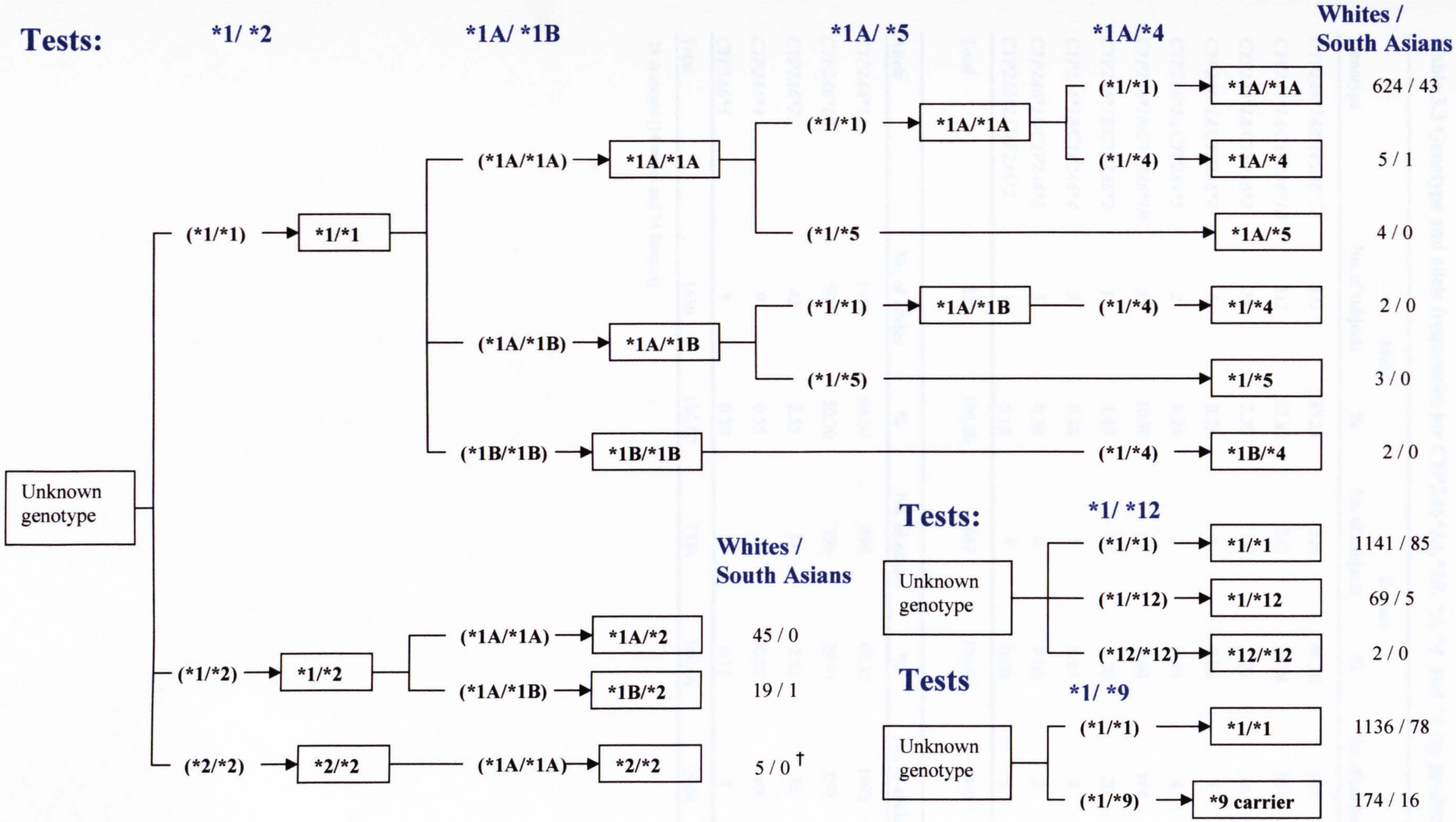


Figure 3.2 *CYP2A6* genotype designations Test outcomes in brackets, genotype designations in boxes. † Among five *2 / *2 apparent homozygotes, three are *2 / *12. Numbers given by each box (a/b) represent a (number of Whites), b (number of South Asians).

Table 3.3 Genotype and allele frequencies for *CYP2A6*1A,*1B, *2, *4* and **5* by gender

	Male		Female		Total	
Genotype	No. of subjects	%	No. of subjects	%	No. of subjects	%
*CYP2A6*1A/CYP2A6*1A*	373	45.21	308	46.18	681	45.64
*CYP2A6*1A/CYP2A6*1B*	312	37.82	262	39.28	574	38.47
*CYP2A6*1A/CYP2A6*2*	28	3.39	18	2.70	46	3.08
*CYP2A6*1A/CYP2A6*4*	6	0.73	0	0.00	6	0.40
*CYP2A6*1A/CYP2A6*5*	2	0.24	2	0.30	4	0.27
*CYP2A6*1B/CYP2A6*1B*	85	10.30	64	9.60	149	9.99
*CYP2A6*1B/CYP2A6*2*	12	1.45	8	1.20	20	1.34
*CYP2A6*1B/CYP2A6*4*	3	0.36	1	0.15	4	0.27
*CYP2A6*1B/CYP2A6*5*	3	0.36	0	0.00	3	0.20
*CYP2A6*2/CYP2A6*2*	1	0.12	4	0.60	5	0.34
Total	825	100.00	667	100.00	1492	100.00

Allele	No. of alleles	%	No. of alleles	%	No, of alleles	%
*CYP2A6*1A*	1094	66.30	898	67.32	1992	66.76
*CYP2A6*1B*	500	30.30	399	29.91	899	30.13
*CYP2A6*2*	42	2.55	34	2.55	76	2.55
*CYP2A6*4*	9	0.55	1	0.07	10	0.34
*CYP2A6*5*	5	0.30	2	0.15	7	0.23
Total	1650	100.00	1334	100.00	2984	100.00

28 dropouts (14 males and 14 females)

Table 3.4 Genotype and allele frequencies for *CYP2A6*12* and *CYP2A6*9* by gender

	Male		Female		Total	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
*CYP2A6*1/*1*	699	94.59	547	93.66	1246	94.18
*CYP2A6*1/*12*	39	5.27	35	5.99	74	5.59
*CYP2A6*12/*12*	1	0.14	2	0.34	3	0.23
Total	739	100.00	584	100.00	1323	100.00
Allele						
[1]*CYP2A6*1*	1437	97.23	1129	96.66	2566	96.98
*CYP2A6*12*	41	2.77	39	3.34	80	3.02
Total	1478	100.00	1168	100.00	2646	100.00
Genotype						
*CYP2A6*1/*1*	673	85.73	559	87.34	1232	86.46
[2]*CYP2A6*9/*X*	112	14.27	81	12.66	193	13.54
Total	785	100.00	640	100.00	1425	100.00

Total number of dropout is 197 for *CYP2A6*12* (100 males and 97 females); 95 for *CYP2A6*9*(54 males and 41 females)

1. *CYP2A6*1* represents alleles which are not *CYP2A6*12 or CYP2A6*9*
2. *X means either *1 or *9

3.2.1.2 Genotype and allele frequencies by ethnicity

Most of the subjects (1401, 92.2%) in the TTHHS cohort were Caucasians. The predominant non-Caucasian group was South Asian, 96 (6.3%), the remaining 23 (1.5%) included African-American, Orientals and others. The genotype and allele frequencies by ethnicity are shown in Table 3.5 and Table 3.6. There was a trend for the frequency of *CYP2A6*2* allele to be higher in Caucasians than in South Asians (2.54% versus 0.52%), but statistically there was no significant difference between these two groups (p=0.0635).

Table 3.5 Genotype and allele frequencies for *CYP2A6*1A,*1B, *2, *4* and **5* by ethnicity

	Caucasian		Asian		Other	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
*CYP2A6*1A/CYP2A6*1A*	624	45.41	43	44.79	14	63.64
*CYP2A6*1A/CYP2A6*1B*	534	38.86	36	37.50	4	18.18
*CYP2A6*1A/CYP2A6*2*	45	3.28	0	0.00	1	4.55
*CYP2A6*1A/CYP2A6*4*	5	0.36	1	1.04	0	0.00
*CYP2A6*1A/CYP2A6*5*	4	0.29	0	0.00	0	0.00
*CYP2A6*1B/CYP2A6*1B*	131	9.53	15	15.63	3	13.64
*CYP2A6*1B/CYP2A6*2*	19	1.38	1	1.04	0	0.00
*CYP2A6*1B/CYP2A6*4*	4	0.29	0	0.00	0	0.00
*CYP2A6*1B/CYP2A6*5*	3	0.22	0	0.00	0	0.00
*CYP2A6*2/CYP2A6*2*	5	0.36	0	0.00	0	0.00
Total	1374	100.00	96	100.00	22	100.00

Allele	**No. of alleles**	**%**	**No. of alleles**	**%**	**No. of alleles**	**%**
*CYP2A6*1A*	1836	66.81	123	64.06	33	75.00
*CYP2A6*1B*	822	29.91	67	34.90	10	22.73
*CYP2A6*2*	74	2.69	1	0.52	1	2.27
*CYP2A6*4*	9	0.33	1	0.52	0	0.00
*CYP2A6*5*	7	0.25	0	0.00	0	0.00
Total	2748	100.00	192	100.00	44	100.00

28 dropouts (27 Caucasians and 1 other)

Table 3.6 Genotype and allele frequencies for *CYP2A6*12* and *CYP2A6*9* by ethnicity

	Caucasian		Asian		Other	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
*CYP2A6*1/*1*	1141	94.06	85	94.44	20	100.00
*CYP2A6*1/*12*	69	5.69	5	5.56	0	0.00
*CYP2A6*12/*12*	3	0.25	0	0.00	0	0.00
Total	1213	100.00	90	100.00	20	100.00
Allele						
[1]*CYP2A6*1*	2351	96.91	175	97.22	36	100.00
*CYP2A6*12*	75	3.09	5	2.78	0	0.00
Total	2426	100.00	180	100.00	36	100.00
Genotype						
*CYP2A6*1/*1*	1136	86.72	78	82.98	18	85.71
[2]*CYP2A6*9/*X*	174	13.28	16	17.02	3	14.29
Total	1310	100.00	94	100.00	21	100.00

Total number of dropout is 197 for *CYP2A6*12* (188 Caucasians, 6 Asians and 3 others);
95 for *CYP2A6*9* (91 Caucasians, 2 Asians and 2 others

1. *CYP2A6*1* represents alleles which are not *CYP2A6*12 or CYP2A6*9*
2. *X means either *1 or *9

3.2.2 Genotype and allele frequencies of SNPs for *DBH, MAOA*, and *5HT2A*

3.2.2.1 Genotype and allele frequencies of *DBH* G1368A

The genotype and allele frequencies of *DBH* G1368A are shown in Table 3.7 and Table 3.8 by gender and ethnicity. There was no significant difference between males and females, and between different ethnic groups.

Table 3.7 Genotype and allele frequencies for *DBH* G1368A by gender

	Male		Female		Total	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
G/G	215	26.16	186	28.40	401	27.15
G/A	419	50.97	323	49.31	742	50.24
A/A	188	22.87	146	22.29	334	22.61
Total	822	100.00	655	100.00	1477	100.00
Allele						
G	849	51.64	695	53.05	1544	52.27
A	795	48.36	615	46.95	1410	47.73
Total	1644	100.00	1310	100.00	2954	100.00

43 dropouts (17 males and 26 females)

Table 3.8 Genotype and allele frequencies for *DBH* G1368A by ethnicity

	Caucasian		Asian		Other	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
G/G	372	27.35	20	21.28	9	39.13
G/A	681	50.07	50	53.19	11	47.83
A/A	307	22.57	24	25.53	3	13.04
Total	1360	100.00	94	100.00	23	100.00
Allele						
G	1425	52.39	90	47.87	29	63.04
A	1295	47.61	98	52.13	17	36.96
Total	2720	100.00	188	100.00	46	100.00

43 dropouts (41 Caucasians, 2 Asians)

3.2.2.2 Genotype and allele frequencies of *MAOA* T1460C

The *MAOA* gene is located on the X chromosome. Males, therefore, have only one allele for this marker. The genotype and allele frequencies are shown in Table 3.9. The χ^2 test showed that for both genotype and allele frequencies for females, there was a significant difference between Caucasians and South Asians ($p<0.001$), as well as Caucasians and others ($p=0.001$), but not between South Asians and others ($p=0.846$). For males, there was significant difference between Caucasians and South Asians

(p<0.001), but not Caucasians and other (p=0.72), nor South Asians and others (p=0.13). There was no significant difference between males and females.

Table 3.9 Genotype and allele frequencies for *MAOA T1460C* by gender and ethnicity

		Caucasian		Asian		Other	
	Genotype	No. of subjects	%	No. of subjects	%	No. of subjects	%
Female	*C/C*	340	53.97	7	17.07	1	14.29
	CT	245	38.89	13	31.71	3	42.86
	TT	45	7.14	21	51.22	3	42.86
	Total	630	100.00	41	100.00	7	100.00
	Allele						
	C	925	73.41	27	32.93	5	35.71
	T	335	26.59	55	67.07	9	64.29
	Total	1260	100.00	82	100.00	14	100.00
Male	*C*	557	72.81	26	47.27	11	68.75
	T	208	27.19	29	52.73	5	31.25
	Total	765	100.00	55	100.00	16	100.00

6 dropouts (all Caucasians, 3 males and 3 females)

3.2.2.3 Genotype and allele frequencies of *5HT2A* T102C

Genotype and allele frequencies by gender and ethnicity are shown in Table 3.10 and 3.11. χ^2 tests showed that there was no significant difference between males and females, as well as between different ethnic groups.

Table 3.10 Genotype and allele frequencies for *5HT2A* T102C by gender

	Male		Female		Total	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
T/T	135	20.00	143	17.15	278	18.42
T/C	305	45.19	396	47.48	701	46.45
C/C	235	34.81	295	35.37	530	35.12
Total	675	100.00	834	100.00	1509	100.00
Allele						
T	575	42.59	682	40.89	1257	41.65
C	775	57.41	986	59.11	1761	58.35
Total	1350	100.00	1668	100.00	3018	100.00

11 dropouts (6 males and 5 females)

Table 3.11 Genotype and allele frequencies for *5HT2A* T102C by ethnicity

	Caucasian		Asian		Other	
Genotype	**No. of subjects**	**%**	**No. of subjects**	**%**	**No. of subjects**	**%**
T/T	254	18.26	20	21.05	4	17.39
T/C	649	46.66	42	44.21	10	43.48
C/C	488	35.08	33	34.74	9	39.13
Total	1391	100.00	95	100.00	23	100.00
Allele						
T	1157	41.59	82	43.16	18	39.13
C	1625	58.41	108	56.84	28	60.87
Total	2782	100.00	190	100.00	46	100.00

11 dropouts (10 Caucasians, 1 from Asian)

3.2.3 Genotype and allele frequencies of VNTRs

Several VNTRs were genotyped, including the *MAOA* 30bp VNTR, the *DRD4* 48bp VNTR and the *TH01* TCAT tetranucleotide.

3.2.3.1 Genotype and allele frequencies of *MAOA* VNTR

The *MAOA* VNTR sequence is internally repetitive in that it consists of 5 repetitions of the core sequence ACC(A/G/C)G(C/T). The PCR product sizes of 310bp, 340bp, 355bp, 370bp and 400bp in this study represented 2, 3, 3.5, 4 and 5 repeats. The genotype and allele frequencies are shown in Table 3.12. As some allele frequencies were very low, especially in non-Caucasian groups, alleles 310, 340, and 355 (2, 3 and 3.5 repeats) were combined into one group (short group), and alleles 370 and 400 (4 and 5 repeats) were combined into another group (long group). Then χ^2 tests were performed to compare allele distributions between different ethnic groups. The results showed that in women there was a significant difference between Caucasians and South Asians ($\chi^2 = 39.27$, $p<0.001$), as well as Caucasians and others ($\chi^2 = 12.21$ $p<0.001$), but not South Asians and others ($\chi^2 = 0.5998$, $p=0.4386$). In men, a significant difference only existed between Caucasians and South Asians ($\chi^2 = 5.7361$, $p=0.0166$), not between Caucasians and others ($\chi^2 = 0.94$, $p= 0.3323$), nor South Asians and others ($\chi^2 =0.0855$, $p= 0.7699$).

There were no significant differences in allele frequency between males and females in Caucasians ($\chi^2 =0.092$, $p=0.7617$) and others ($\chi^2 =3.1309$, $p= 0.0768$). However, there was a significant difference between men and women in the South Asian group ($\chi^2 =$ 4.0917, $p= 0.0431$) (Table 3.13).

Table 3.12 Genotype and allele frequencies for *MAOA VNTR* by gender and ethnicity

		Caucasian		Asian		Other	
	Genotype	No. of subjects	%	No. of subjects	%	No. of subjects	%
Female	310/340	2	0.33	-	-	-	-
	340/340	61	10.10	21	51.22	4	57.14
	340/355	6	0.99	-	-	-	-
	355/355	1	0.17	-	-	-	-
	310/370	1	0.17	-	-	1	14.29
	340/370	249	41.23	14	34.15	2	28.57
	355/370	16	2.65	-	-	-	-
	370/370	257	42.55	6	14.63	-	-
	340/400	4	0.66	-	-	-	-
	370/400	7	1.16	-	-	-	-
	Total	604	100.00	41	100.00	7	100.00
	Allele						
	310	3	0.25	-	-	1	7.14
	340	383	31.71	56	68.29	10	71.43
	355	24	1.99	-	-	-	-
	370	787	65.15	26	31.71	3	21.43
	400	11	0.91	-	-	-	-
	Total	1208	100.00	82	100.00	14	100.00
Male	310	1	0.14	-	-	-	-
	340	238	32.69	26	49.06	6	40.00
	355	13	1.79	1	1.89	1	6.67
	370	467	64.15	24	45.28	7	46.67
	400	9	1.24	2	3.77	1	6.67
	Total	728	100.00	53	100.00	15	100.00

72 dropouts (29 females and 43 males)

1. Allele size 310,340, 355, 370, 400 represent 2, 3, 3.5, 4 and 5 repeats
2. '-' represents no data

Table 3.13 Comparison for *MAOA* VNTR allele frequencies (combined) between different ethnic groups

	Caucasian		South Asian		Other	
	Male	Female	Male	Female	Male	Female
Short (2-3.5)	252	410	27	56	7	11
Long (4-5)	476	798	26	26	8	3
χ^2	5.7361(C-S)	39.27	0.0855(S-O)	0.6	0.94(C-O)	12.21
p	0.0166	<0.001	0.7699	0.4386	0.3323	<0.001

Short: 2, 3 and 3.5 repeats. Long: 4 and 5 repeats.
C-S: comparison between Caucasian and South Asian.
S-O: comparison between South Asian and other.
C-O: comparison between Caucasian and other.

3.2.3.2 Genotype and allele frequencies of *DRD4* VNTR

The variable number of tandem repeats for *DRD4* ranges from 2-11 repeats with the PCR fragment lengths varing between 379-811bp using primers described in Ebstein *et al.* (1996) [236]. The most common alleles in this study were 4 repeats and 7 repeats, and no 10 or 11 repeats were found in the TTHHS cohort. For the same reason as described above, alleles were combined into two groups: short (2-5 repeats) and long (6-9 repeats). χ^2 tests were then performed to compare allele distributions between the different ethnic groups. Table 3.14 showed allele frequencies (combined) were significantly different between Caucasians and South Asians, both for males (χ^2 =13.965, p<0.001) and females (χ^2 =5.593, p=0.018). There was no significant difference between South Asians and others, or between Caucasians and others. Complete genotype and allele frequencies of *DRD4* VNTR are shown in Table 3.15.

Table 3.14 Comparison for *DRD4* VNTR allele frequencies(combined) between different ethnic groups

	Caucasian		South Asian		Other	
	Male	Female	Male	Female	Male	Female
short (2-5)	1124	917	101	73	28	10
long (6-9)	268	221	5	7	4	2
χ^2	13.965(C-S)	5.593	2.442(S-O)	0.741	0.923(C-O)	0.058
p	0.0002	0.018	0.1181	0.3893	0.3367	0.8103

Short: 2, 3, 4 and 5 repeats. Long: 6, 7, 8 and 9 repeats.
C-S: comparison between Caucasian and South Asian.
S-O: comparison between South Asian and Other.
C-O: comparison between Caucasian and Other.

Table 3.15 Genotype and allele frequencies for *DRD4* VNTR by gender and ethnicity

	Caucasian				South Asian				Other			
	Male		Female		Male		Female		Male		Female	
Genotype	No	%	No	%	No	%	No	%	No	%	No	%
2/2	9	1.29	10	1.76	-	-	-	-	3	18.75	-	-
2/3	7	1.01	4	0.70	-	-	-	-	-	-	-	-
2/4	99	14.22	48	8.44	7	13.21	5	12.50	1	6.25	-	-
2/5	-	-	1	0.18	-	-	-	-	-	-	-	-
2/7	18	2.59	14	2.46	-	-	-	-	-	-	1	16.67
2/8	-	-	1	0.18	-	-	-	-	-	-	-	-
3/3	3	0.43	6	1.05	-	-	-	-	-	-	-	-
3/4	33	4.74	33	5.80	-	-	3	7.50	1	6.25	-	-
3/5	-	-	1	0.18	-	-	-	-	-	-	-	-
3/7	9	1.29	11	1.93	-	-	-	-	-	-	-	-
3/8	-	-	1	0.18	-	-	-	-	-	-	-	-
4/4	303	43.53	266	46.75	40	75.47	26	65.00	7	43.75	3	50.00
4/5	2	0.29	6	1.05	1	1.89	-	-	-	-	1	16.67
4/6	2	0.29	3	0.53	-	-	-	-	1	6.25	-	-
4/7	177	25.43	133	23.37	3	5.66	4	10.00	2	12.50	1	16.67
4/8	6	0.86	3	0.53	2	3.77	1	2.50	1	6.25	-	-
5/7	-	-	1	0.18	-	-	-	-	-	-	-	-
6/7	1	0.14	2	0.35	-	-	-	-	-	-	-	-
7/7	26	3.74	22	3.87	-	-	1	2.50	-	-	-	-
7/8	-	-	3	0.53	-	-	-	-	-	-	-	-
7/9	1	0.14	-	-	-	-	-	-	-	-	-	-
Total	696	100.00	569	100.00	53	100.00	40	100.00	16	100.00	6	100.00
Allele												
2	142	10.20	88	7.73	7	6.60	5	6.25	7	21.88	1	8.33
3	55	3.95	62	5.45	-	-	3	3.75	1	3.13	-	-
4	925	66.45	758	66.61	93	87.74	65	81.25	20	62.50	8	66.67
5	2	0.14	9	0.79	1	0.94	-	-	-	-	1	8.33
6	3	0.22	5	0.44	-	-	-	-	1	3.13	-	-
7	258	18.53	208	18.28	3	2.83	6	7.5	2	6.25	2	16.67
8	6	0.43	8	0.70	2	1.89	1	1.25	1	3.13	-	-
9	1	0.07	-	-	-	-	-	-	-	-	-	-
Total	1392	100	1138	100.00	106	100	80	100	32	100.00	12	100

140 dropouts (74 males and 66 females)

1. Numbers 2-9 represent 2-9 repeats

3.2.3.3 Genotype and allele frequencies of *TH01* VNTR

Seven different alleles were reported for this intronic tetranucleotide (TCAT)-repeat polymorphism consisting of 5, 6, 7, 8, 9, 9.3(10), and 11 repeats [211]. The common alleles are 6, 7, 8, 9 and 9.3 repeats. In the TTHHS cohort, no 11-repeat allele was

found, and only one 5-repeat allele was detected. The complete genotype and allele frequencies of *TH01* VNTR by gender and ethnicity are shown in Table 3.16. χ^2 tests were performed to compare combined allele distributions between different ethnic groups. Results showed that there were no combined allele discrepancies between the different ethnic groups (Table 3.17).

Table 3.16 Genotype and allele frequencies for *TH01* VNTR by gender and ethnicity

	Caucasian				South Asian				Other			
	Male		Female		Male		Female		Male		Female	
Genotype	No	%	No	%	No	%	No	%	No	%	No	%
5/6	1	0.15	-	-	-	-	-	-	-	-	-	-
5/8	-	-	1	0.21	-	-	-	-	-	-	-	-
5/9	-	-	4	0.84	-	-	-	-	-	-	-	-
6/6	36	5.26	28	5.85	4	8.33	4	12.50	2	14.29	1	16.67
6/7	50	7.30	38	7.93	4	8.33	5	15.63	2	14.29	2	33.33
6/8	21	3.07	23	4.80	2	4.17	1	3.13	-	-	-	-
6/9	52	7.59	22	4.59	6	12.50	5	15.63	2	14.29	-	-
6/9.3	106	15.47	65	13.57	3	6.25	1	3.13	1	7.14	-	-
7/7	21	3.07	18	3.76	1	2.08	-	-	-	-	-	-
7/8	23	3.36	28	5.85	6	12.50	1	3.13	1	7.14	-	-
7/9	29	4.23	23	4.80	4	8.33	4	12.50	1	7.14	-	-
7/9.3	65	9.49	62	12.94	3	6.25	1	3.13	1	7.14	-	-
8/8	8	1.17	7	1.46	-	-	2	6.25	1	7.14	1	16.67
8/9	21	3.07	24	5.01	1	2.08	2	6.25	-	-	-	-
8/9.3	37	5.40	34	7.10	3	6.25	4	12.50	-	-	1	16.67
9/9	92	13.43	11	2.30	5	10.42	-	-	1	7.14	-	-
9/9.3	51	7.45	34	7.10	4	8.33	2	6.25	1	7.14	1	16.67
9.3/9.3	72	10.51	57	11.90	2	4.17	-	-	1	7.14	-	-
Total	685	100.00	479	100.00	48	100.00	32	100.00	14	100.00	6	100.00
Allele												
5	1	0.07	5	0.52	-	-	-	-	-	-	-	-
6	302	22.04	204	21.29	23	23.96	20	31.25	9	32.14	4	33.33
7	209	15.26	187	19.52	19	19.79	11	17.188	5	17.86	2	16.67
8	118	8.61	124	12.94	12	12.50	12	18.75	3	10.71	3	25.00
9	337	24.60	129	13.47	25	26.04	13	20.313	6	21.43	1	8.33
9.3	403	29.42	309	32.25	17	17.71	8	12.5	5	17.86	2	16.67
Total	1370	100.00	958	100.00	96	100.00	64	100	28	100.00	12	100

337 dropouts (173 males and 164 females)

1. Numbers 5-9.3 represent 5-9.3 repeats

Table 3.17 Comparison for *TH01* VNTR allele frequencies(combined) between different ethnic groups

	Caucasian		South Asian		Other	
	Male	Female	Male	Female	Male	Female
short (5-7)	512	396	42	31	14	6
long (8-9.3)	858	562	54	33	14	6
χ^2	1.152(C-S)	1.244	0.342(S-O)	0.01	1.8643(C-O)	0.367
p	0.2128	0.2647	0.5587	0.9208	0.3367	0.5449

Short: 5, 6 and 7 repeats. Long: 8, 9 and 9.3 repeats
C-S: comparison between Caucasian and South Asian.
S-O. comparison between South Asian and Other
C-O. comparison between Caucasian and Other

3.2.4 Hardy-Weinberg Equilibrium test

It was noted from the above tables that some markers were differentially represented, for example *CYP2A6*2* (L160H) is an allele that was found at higher frequency in Caucasians than in other ethnic groups, and the rarer alleles of both *MAOA* 1460C>T and *MAOA* VNTR alleles 340 and 370 are more common in South Asians. Hardy–Weinberg equilibrium (HWE) was tested on the Caucasian group, South Asian group and overall for each marker by χ^2 analysis. The genotype counts (Table 3.18) for each marker were in HWE. The contribution of the high frequency of *MAOA* homozygotes in 96 South Asians almost led to significance of HWE test in the overall sample (p=0.059) but the Caucasian subset fitted HWE expectations. For *CYP2A6*, an excess of apparent *2 homozygotes (five) was observed in assays defining *1/*2 classification but subsequent analysis revealed that three of these five were in fact heterozygotes for a more recently defined hybrid allele *12 which can be mistaken in *2/*1 heterozygotes as *2 in the 160H assay (*12 was null to the *1 allele assay used in conjunction with the *2 test, since *12 does not amplify in long PCR1 (Fig. 3.2) in the first step of this nested two step assay).

Table 3.18 Genoype counts, minor allele frequency (q) and HWE

A. For overall population sample

	wt/wt	wt/mut	mut/mut	q	χ^2	Signlficance
MAOA 1460C>T †	347	261	69	0.2947	3.5650	0.0590
DBH 1368G>A	401	740	334	0.4773	0.044	0.8338
5HT2A 102C>T	530	699	278	0.4164	3.1395	0.0764
*CYP2A6*1B*	651	597	151	0.3213	0.6499	0.4202
*CYP2A6*2*	1372	68	2	0.0250	1.4194	0.2335
*CYP2A6*5*	1395	7	0	0.0025	0.0087	0.9256
*CYP2A6*4*	1413	10	0	0.0035	0.0177	0.8942
*CYP2A6*12*	1245	73	3	0.0299	0.7269	0.3939

B.For Caucasian subset

	wt/wt	wt/mut	mut/mut	q	χ^2	Significance
MAOA 1460C>T †	339	245	45	0.2663	0.0065	0.9357
DBH 1368G>A	372	679	307	0.4761	0.0072	0.9326
5HT2A 102C>T	488	647	254	0.4158	2.3557	0.1248
*CYP2A6*1B*	594	556	133	0.3203	0.0295	0.8637
*CYP2A6*2*	1258	66	2	0.0264	1.3225	0.2501
*CYP2A6*5*	1279	7	0	0.0027	0.0096	0.922
*CYP2A6*4*	1299	9	0	0.0034	0.0156	0.9006
*CYP2A6*12*	1140	68	3	0.0306	0.8421	0.3588

† 1. *MAOA* 1460C>T data were females only.
2. Caveats of cells with low counts or low expected values, Yates corrections made no difference to conclusions.

C. For South Asian subset

	wt/wt	wt/mut	mut/mut	q	χ^2	Significance
MAOA 1460C>T †	21	13	7	0.3293	3.2641	0.0708
DBH 1368G>A	24	50	20	0.4787	0.4065	0.5238
5HT2A 102C>T	33	40	20	0.4301	0.9295	0.3350
*CYP2A6*1B*	44	37	15	0.3490	2.2110	0.1370
*CYP2A6*2*	94	1	0	0.0053	*	*
*CYP2A6*5*	96	0	0	0.0000	*	*
*CYP2A6*4*	94	1	0	0.0053	*	*
*CYP2A6*12*	85	5	0	0.0278	*	*

† 1. *MAOA* 1460C>T data were females only.
* χ^2 test even with Yates' correction inappropriate.

3.2.5 Linkage disequilibrium between *MAOA* T1460C and *MAOA* VNTR

LD between *MAOA* T1406C and *MAOA* VNTR alleles in females was calculated using 2LD as described in section 2.5.2. Haplotype/diplotype frequencies for T1460C and VNTR in the *MAOA* gene were calculated using Microsoft Excel®. Markers T1460C and the VNTR in *MAOA* were in strong linkage disequilibrium in females. D´= 0.8937, $p<0.001$ with the most common haplotype in males being C-370 (370=allele with 4 repeats) 59.9%; second most common T-340 (340=allele with 3 repeats) 26.2%; third most common C-340 7.76% and fourth most common T-370 2.63%. In females, the most common diplotypes were CC-370/370 (38.8%) and TC-370/340 (30.93%), followed by TT-340/340 (8.73%) and CC-370/340 (8.42%). Haplotype (diplotype) frequencies are tabulated in Table 3.19 and Table 3.20.

Table 3.19 Linkage disequilibrium between *MAOA* T1460C and VNTR

MAOA T1460C	*MAOA*-VNTR	Male		Female	
Allele	Allele (Repeat)	Allele No.	Haplotype Frequency	Estimated allele No.	Estimated Haplotype
T	310 (2)	1	0.0013	4	0.0030
	340 (3)	209	0.2616	356	0.2729
	335 (3.5)	0	0.0000	2	0.0012
	370 (4)	21	0.0263	21	0.0161
	400 (5)	3	0.0038	3	0.0025
C	310 (2)	0	0.0000	0	0.0001
	340 (3)	62	0.0776	93	0.0709
	335 (3.5)	15	0.0188	22	0.0172
	370 (4)	479	0.5995	798	0.6110
	400 (5)	9	0.0113	7	0.0052
Total		799	1.0000	1306	1.0000

3.20 Linkage disequilibrium between *MAOA* T1460C and VNTR (females only)

MAOA T1460C	*MAOA*-VNTR	No. of genotype	genotype frequency
TT	340/310	1	0.0015
	340/340	57	0.0873
	370/340	8	0.0123
	400/340	1	0.0015
TC	340/310	1	0.0015
	340/340	24	0.0368
	355/340	5	0.0077
	370/310	2	0.0031
	370/340	**202**	**0.3093**
	370/355	2	0.0031
	370/370	11	0.0168
	400/340	3	0.0046
	400/370	2	0.0031
CC	340/340	5	0.0077
	355/340	1	0.0015
	355/355	1	0.0015
	370/340	55	0.0842
	370/355	14	0.0214
	370/370	**254**	**0.3890**
	400/370	4	0.0061
	Total	653	1.0000

3.3 Genotype-phenotype analyses in relation to smoking behaviour

As described in the methods section, smoking behaviour was investigated in two stages (13-15y, and 18y). Smoking traits, including the amount smoked, starting age, ease of

giving up, number of attempts to give up, timing of first cigarette of the day and interval since the last cigarette, were analysed by genotypes at both stages.

3.3.1 Analysis of smoking status by genotype at ages 13–15 and 18 years

At age 13–15, smoking data were available for 1511 subjects. There were 403 current smokers, 492 ex-smokers, and 616 never smokers. At age 18, smoking data were available for 1255 subjects. Among them were 434 current smokers, 373 ex-smokers and 448 never smokers. Of the 492 and 373 ex-smokers at age 13–15 and 18 years, cotinine measurement was <15 ng/ml for 480 subjects at 13–15 years and 308 subjects at 18 years, which represents 98% and 83% of ex-smokers at each stage being biochemically validated (cotinine measurements were not available for 12 cases in the 13-15 age group and 65 cases at 18 years). Genotype data were available for most subjects. Odds ratios and their 95% confidence intervals are given for each genotype group at ages 13–15 and 18 years in Table 3.21, for ex and current smokers versus never smokers; for current versus ex-smokers; and for current versus never and ex-smokers. For *CYP2A6*, there was no relation to smoking status at 13–15 years. However, at 18 years the 'slowest inactivators' predicted to be half normal metabolizers (carriers of one fully inactive allele, viz. *2, *4, or *5, i.e. haploinsufficient, HI) were more likely to be a current smoker. Current smoker versus ex-smoker odds ratio was 2.23 (95% C.I. 1.16–4.27), and current smoker versus never and ex-smoker odds ratio was 2.02 (95% C.I. 1.21–3.38). Exclusion of non- Caucasians from these analyses gave similar results (Table 3.23) and these respective odds ratios were 2.05 (1.06, 3.98) and 1.85 (1.09, 3.13). No genotypes except for *CYP2A6* exhibited association with smoking status. For *MAOA,* analysis of the 1460C>T SNP only is presented as it is in strong LD with the VNTR.

Table 3.21 Smoking behaviour by genotypes at age 13-15

Marker	Genotype group	Total Number	Smoking Status at 13-15 corrected for cotinine			OR (95% CI) (adjusted for age, sex, town and ethnicity)		
			Current Number (%)	Ever Number (%)	Never Number (%)	ex + current vs never	current vs ex	current vs never +ex
CYP2A6 [1]	WT1B*	625	174 (27.8)	196 (31.4)	255 (40.8)	1	1	1
	WT1A*	408	105 (25.7)	134 (32.8)	169 (41.4)	0.98 (0.75, 1.26)	0.90 (0.64, 1.27)	0.90 (0.67, 1.20)
	PHI*	251	62 (24.7)	84 (33.5)	105 (41.8)	0.99 (0.73, 1.34)	0.84 (0.56, 1.25)	0.88 (0 62, 1.24)
	HI*	82	24 (29.3)	26 (31.7)	32 (39.0)	1.05 (0.65, 1.70)	0.95 (0.51, 1.77)	1.07 (0.63, 1.81)
DBH [2]	GG	400	114 (28.5)	134 (33.5)	152 (38.0)	1	1	1
	GA	735	186 (25.3)	243 (33.1)	306 (41.6)	0.86 (0 67, 1.11)	0.89 (0.64, 1.23)	0.85 (0.64, 1.12)
	AA	333	90 (27.0)	106 (31.8)	137 (41.1)	0.88 (0.65, 1.19)	0.94 (0.63, 1.39)	0.91 (0 65, 1.27)
DRD4 [3]	short (2-5 repeat)	922	246 (26.7)	312 (33.8)	364 (39.5)	1	1	1
	long (6-9 repeat)	450	113 (25.1)	143 (31.8)	194 (43.1)	0.81 (0.64, 1.03)	0.99 (0.72, 1.35)	0.91 (0.70, 1.20)
5HT2A [4]	TT	274	79 (28 8)	79 (28.8)	116 (42.3)	1	1	1
	TC	697	176 (25.3)	235 (33.7)	286 (41.0)	1.07 (0.80, 1.43)	0.79 (0.54, 1.16)	0.87 (0.63, 1.20)
	CC	529	148 (28.0)	175 (33.1)	206 (38.9)	1.19 (0.88, 1.62)	0.96 (0.64, 1.42)	1.05 (0.75, 1.47)
MAO SNP [5]	Boys C	589	136 (23.1)	195 (33.1)	258 (43.8)	1	1	1
	T	239	59 (24.7)	71 (29.7)	109 (45.6)	1.00 (0.73, 1.37)	1.27 (0 82, 1.97)	1.18 (0.81, 1.71)
	Girls CC	347	112 (32.3)	109 (31.4)	126 (36.3)	1	1	1
	TT or TC	330	96 (29.1)	115 (34.5)	119 (36.1)	1.14 (0.82, 1.59)	0.85 (0.56, 1.28)	0.97 (0.69, 1.37)
TH01 [6]	no 9.0s	856	240 (28.0)	274 (32.0)	342 (40.0)	1	1	1
	at least 1 9.0s	320	74 (23.0)	109 (34.0)	137 (43.0)	0.96 (0.73, 1.25)	0.74 (0.52, 1.07)	0.79 (0.58, 1.08)
	no 7.0s	785	204 (26.0)	243 (31.0)	338 (43.0)	1	1	1
	at least 1 7.0s	391	105 (27.0)	141 (36.0)	145 (37)	1.25 (0.97, 1.61)	0.93 (0 67, 1.29)	1.07 (0.81, 1.42)

[1] "no call" 83; [2] "no call" 43; [3] "no call" 139 ; [4] "no call" 11; [5] "no call" Boys: 3 Girls: 3; [6] "no call" 335.

No calls are included in the models but results are not reported for them.

* WT1B: *1 wildtype (*1B carrier)

WT1A: *1 wild type (*1A/*1A)
PHI: partial haploinsufficency
HI: haploinsufficiency

Table 3.22 Smoking behaviour by genotypes at age 18

Marker	Genotype group	Total Number	Smoking Status at 18 corrected for cotinine			OR (95% CI) (adjusted for age, sex, town and ethnicity)		
			Current Number (%)	Ever Number (%)	Never Number (%)	ex + current vs never	current vs ex	current vs never +ex
CYP2A6 [1]	WT1B*	526	175 (33.3)	168 (31.9)	183 (34.8)	1	1	1
	WT1A*	335	124 (37.0)	85 (25.4)	126 (37.6)	0.86 (0.64, 1.15)	1.44 (1.01, 2.06)	1.16 (0.87, 1.56)
	PHI*	212	66 (31.1)	69 (32.6)	77 (36.3)	0.97 (0.69, 1.38)	0.92 (0.61, 1.37)	0.92 (0.65, 1.31)
	HI*	70	35 (50.0)	15 (21.4)	20 (28.6)	1.34 (0.76, 2.37)	2.23 (1.16, 4.27)	2.02 (1.21, 3.38)
DBH [2]	GG	318	112 (35.2)	103 (32.4)	103 (32.4)	1	1	1
	GA	605	218 (36.0)	172 (28.4)	215 (35.5)	0.89 (0.66, 1.20)	1.15 (0.82, 1.61)	1.03 (0.77, 1.38)
	AA	296	90 (30.4)	91 (30.7)	115 (38.9)	0.78 (0.55, 1.11)	0.86 (0.57, 1.29)	0.79 (0.56, 1.12)
DRD4 [3]	short (2-5 repeat)	769	272 (35.4)	229 (29.8)	268 (34.9)	1	1	1
	long (6-9 repeat)	369	115 (31.2)	112 (30.4)	142 (38.5)	0.81 (0.62, 1.06)	0.87 (0.63, 1.21)	0.83 (0.63, 1.09)
5HT2A [4]	TT	229	87 (38.0)	62 (27.1)	80 (34.9)	1	1	1
	TC	574	186 (32.4)	165 (28.8)	223 (38.9)	0.85 (0.61, 1.19)	0.81 (0.55, 1.21)	0.80 (0.58, 1.11)
	CC	442	160 (36.2)	142 (32.1)	140 (31.7)	1.23 (0.87, 1.76)	0.81 (0.54, 1.22)	0.95 (0.67, 1.33)
MAO SNP [5]	Boys C	456	159 (34.9)	128 (28.1)	169 (37.1)	1	1	1
	T	199	70 (35.2)	54 (27.1)	75 (37.7)	1.06 (0.74, 1.53)	1.10 (0.71, 1.71)	1.11 (0.77, 1.59)
	Girls CC	305	107 (35.1)	102 (33.4)	96 (31.5)	1	1	1
	TT or TC	291	97 (33.3)	87 (29.9)	107 (36.8)	0.94 (0.65, 1.34)	1.04 (0.68, 1.57)	1.00 (0.71, 1.43)
TH01 [6]	no 9.0s	700	252 (36.0)	203 (29.0)	245 (35.0)	1	1	1
	at least 1 9.0s	254	79 (31.1)	84 (33.1)	91 (35.8)	1.01 (0.74, 1.39)	0.76 (0.53, 1.09)	0.83 (0.60, 1.14)
	no 7.0s	641	224 (35.0)	193 (30.0)	224 (35.0)	1	1	1
	at least 1 7.0s	313	103 (32.9)	97 (31.0)	113 (36.1)	0.99 (0.74, 1.32)	0.92 (0.66, 1.30)	0.94 (0.70, 1.26)

[1] "no call" 66; [2] "no call" 36; [3] "no call" 117 ; [4] "no call" 10; [5] "no call" Boys: 2 Girls: 2; [6] "no call" 276.

No calls are included in the models but results are not reported for them.

* WT1B: *1 wildtype (*1B carrier)

WT1A: *1 wild type (*1A/*1A)

PHI: partial haploinsufficency

HI: haploinsufficiency

Table 3.23 Smoking behaviour by genotypes at age 18 [Caucasian only]

Marker	Genotype group	Total Number	Smoking Status at 18 corrected for cotinine			OR (95% CI) (adjusted for age, sex, town and ethnicity)		
			Current Number (%)	Ever Number (%)	Never Number (%)	ex + current vs never	current vs ex	current vs never +ex
CYP2A6 [1]	WT1B*	491	166 (33.8)	161 (32.8)	164 (33.4)	1	1	1
	WT1A*	313	121 (38.7)	78 (24.9)	114 (36.4)	0.89 (0.66, 1.21)	1.57 (1.09, 2.26)	1.26 (0.93, 1.69)
	PHI*	194	62 (32.0)	67 (34.5)	65 (33.5)	1.03 (0.72, 1.48)	0.89 (0.59, 1.35)	0.93 (0.65, 1.33)
	HI*	67	32 (47.8)	15 (22.4)	20 (29.9)	1.25 (0.71, 2.21)	2.05 (1.06, 3.98)	1.85 (1.09, 3.13)

The two most significant odds ratios above were between WT1B and HI of *CYP2A6* at age 18 years (Table 3.22). The significance of these odds ratios can be calculated using the 2-way Contingency Table Analysis software (http://statpages.org/ctab2x2.html), which yields p values of 0.012 and 0.006 for current smoker vs ex-smoker, and current smoker vs ex- + never smoker, respectively. However this does not take into account the fact that multiple testing was performed. Altogether, there are 72 tests (36 tests at each age group), and therefore after Bonferroni correction the p value should be ≤ 0.05/72=0.00069 to keep the type 1 error at a level of 0.05. Thus, after correction for multiple testing none of the associations above can be considered as significant. However it is important to note that this procedure is extremely conservative and may lead to Type II errors (false negatives).

3.3.2 Smoking traits by genotypes at ages 13–15 and 18 years (current smokers only)

Cotinine levels

At age 13–15, salivary cotinine levels amongst current smokers were higher in both the PHI group and the HI group compared to *1B carrier group, though neither difference reached statistical significance (Table 3.24). At age 18, higher cotinine levels were observed among PHI subjects as compared with *1B carriers (164.0 ng/ml vs. 115.0 ng/ml, p=0.0092, uncorrected p value). HI group and *1A homozygotes also showed higher cotinine levels (150.3 ng/ml and 138.2 ng/ml, respectively), but these did not reach statistical significance.

Dose

At age 13–15 only 43% of current smokers smoked more than six cigarettes per week (one cigarette a day). It seems unlikely that many had achieved nicotine dependence at this age. At age 18, nearly all subjects were smoking daily and the average number of cigarettes was approaching 10 per day. However, there was no relationship between genotype and amount smoked.

Timing and other markers of addiction

There were no differences between genotype groups compared to age of starting, ease of giving up, or minutes after waking before first cigarette. However, at age 18 there was some evidence that the PHI and HI subjects were likely to have smoked more recently than the normal group, the difference being significant for the PHI group (Table 3.24). The odds ratios for having tried to stop were higher in the HI group at both 13–15 and 18 years, but the differences approach statistical significance only at 13–15.

Table 3.24. Smoking behaviour markers in current smokers only in relation to CYP2A6 genotype means (standard errors), adjusted for age, sex,town and ethnicity

Age			*CYP2A6*				p value			
		N	WT1B	WT1A	PHI	HI	WT1A vs. WT1B	PHI vs. WT1B	HI vs. WT1B	Test for trend
13-15	Cotinine (ng/ml)	355	73.4 (7.2)	70.7 (9.2)	97.3 (12.1)	95.6 (19.1)	0.8158	0.0919	0.2806	0.0957
	Smoke 7 or more a week (%) (OR, CI)	356	1	1.44	1.33	1.86	(0.86, 2.43)	(0.72, 2.46)	(0.73, 4.77)	0.1253
	Starting age (years)	352	11.9 (0.1)	11.9 (0.2)	12.0 (0.2)	11.3 (0.4)	0.9471	0.8096	0.0878	0.3522
	Ease of giving up (score)	327	2.80 (0.09)	2.72 (0.12)	2.69 (0.15)	2.31 (0.23)	0.5734	0.5318	0.0516	0.0880
	Minutes after waking (score)	285	4.73 (0.13)	4.39 (0.17)	4.50 (0.22)	3.52 (0.36)	0.1278	0.3848	0.0018	0.0074
	Hours since last smoked (hours)	186	6.29 (0.78)	6.26 (0.96)	6.07 (1.26)	6.31 (1.73)	0.983	0.8841	0.9917	0.9423
	Tried to give up (%) (OR, CI)	322	1	1.49	1.48	3.05	(0.83,2.69)	(0.72, 3.01)	(1.01, 9.24)	0.0343
18	Cotinine (ng/ml)	319	115.0 (9.9)	138.2 (12.3)	164.0 (15.7)	150.3 (22.7)	0.1440	0.0092	0.1558	0.0118
	Cigarettes per day (number)	324	8.85 (0.45)	8.61 (0.53)	9.99 (0.77)	7.91 (0.99)	0.7362	0.2020	0.3927	0.9992
	Starting age (years)	383	13.50 (0.16)	13.44 (0.19)	13.40 (0.27)	13.03 (0.36)	0.8090	0.7594	0.2351	0.3063
	Ease of giving up (score)	351	2.63 (0.08)	2.73 (0.09)	2.62 (0.13)	2.87 (0.17)	0.3920	0.9434	0.2102	0.3503
	Minutes after waking (score)	339	4.47 (0.13)	4.37 (0.15)	4.10 (0.22)	4.67 (0.27)	0.6203	0.1505	0.5046	0.8072
	Hours since last smoked (hours)	276	8.35 (0.83)	7.15 (1.02)	4.22 (1.32)	5.69 (1.91)	0.3584	0.0092	0.2025	0.0173
	Tried to give up (%) (OR, CI)	345	1	0.82	0.93	1.12	(0.48,1.38)	(0.48,1.82)	(0.49.2.56)	0.9658

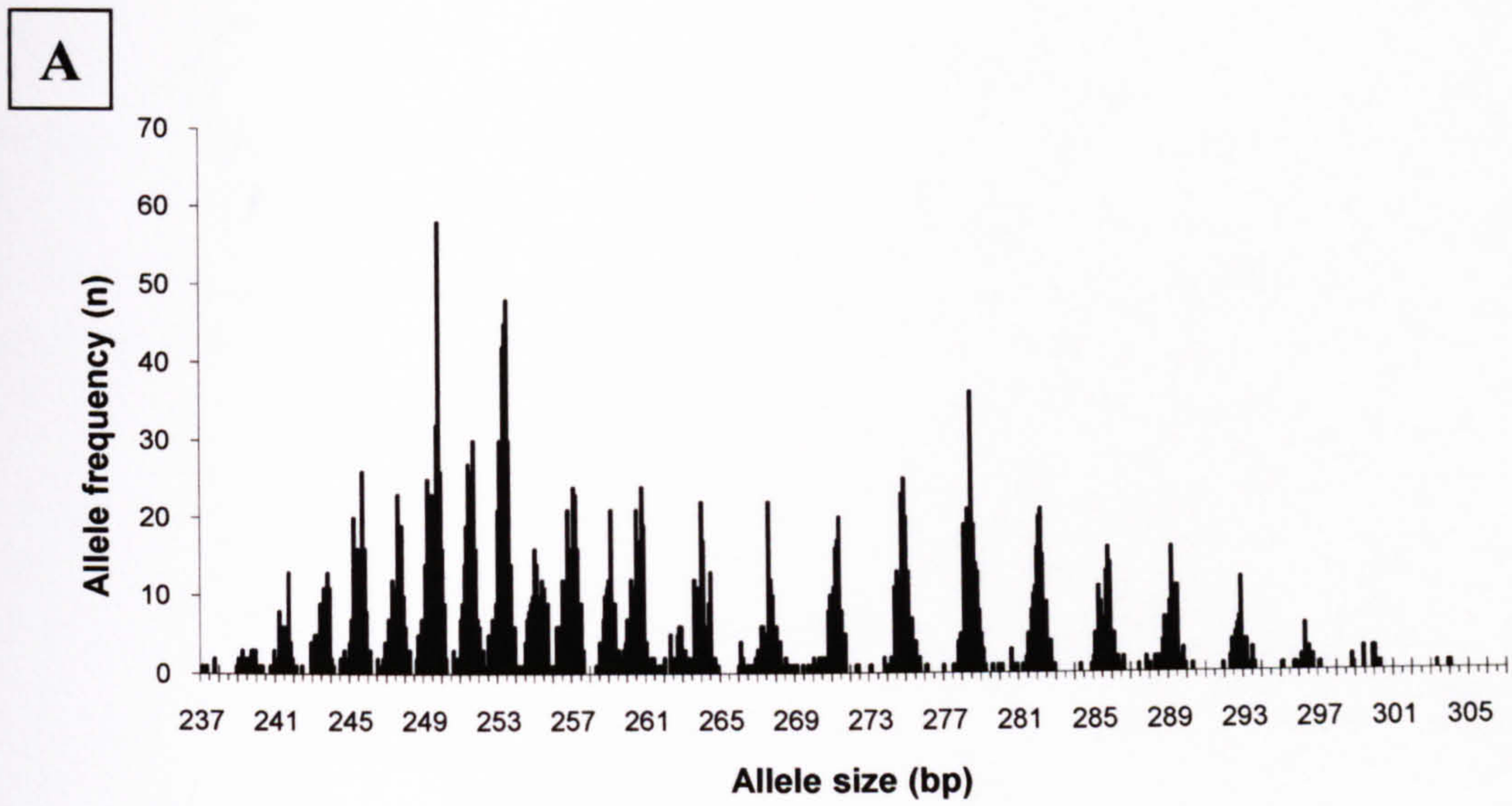
* WT1B: *1 wildtype (*1B carrier)
WT1A: *1 wild type (*1A/*1A)
PHI: partial haploinsufficency
HI: haploinsufficiency

3.4 Results for growth pathway genes

3.4.1 *CSH1.01*

3.4.1.1 *CSH1.01* allele frequency and distribution

The raw data for allele size estimates, from which allele size binning was undertaken, are shown in Figure 3.3A. As described in section 2.5.2, allele sizes were dichotomised for the purposes of statistical analysis. The allele binning size is represented in SGEL code. SGEL represents Southampton Genetic Epidemiology Laboratory, which was set up by Professor Ian Day. The principle of the SGEL code is to take the most frequent allele as 100, with the other alleles sized in relation to this. For example, considering the *CSH1.01* microsatellite, the most frequent allele in the raw data is at 249bp which is designated as 100. The alleles at 247bp, 245bp, 251bp and 253bp in raw data are designated as 98, 96, 102 and 104, respectively as shown in Figure 3.3. Alleles above the allele bin 111 are classified as T (tetranucleotide, shown in red) group as each allele group clusters 4 nucleotides apart, with a mode at bin 126. Alleles below bin 111 are classified as D1 (dinucleotide, shown in yellow and blue) group as each allele group clusters 2 nucleotides apart, with a mode at bin 100. It is also notable that five bins (96, 100, 104, 108, and 111, shown in blue) are more frequent than the other bins 2 nucleotides nearer to each of them. Exclusion of these five bins leaves a subgroup of D1 alleles designated as D2 (shown in yellow). The allele frequencies for TTHHS using this binary classification are shown by sex in Table 3.25.



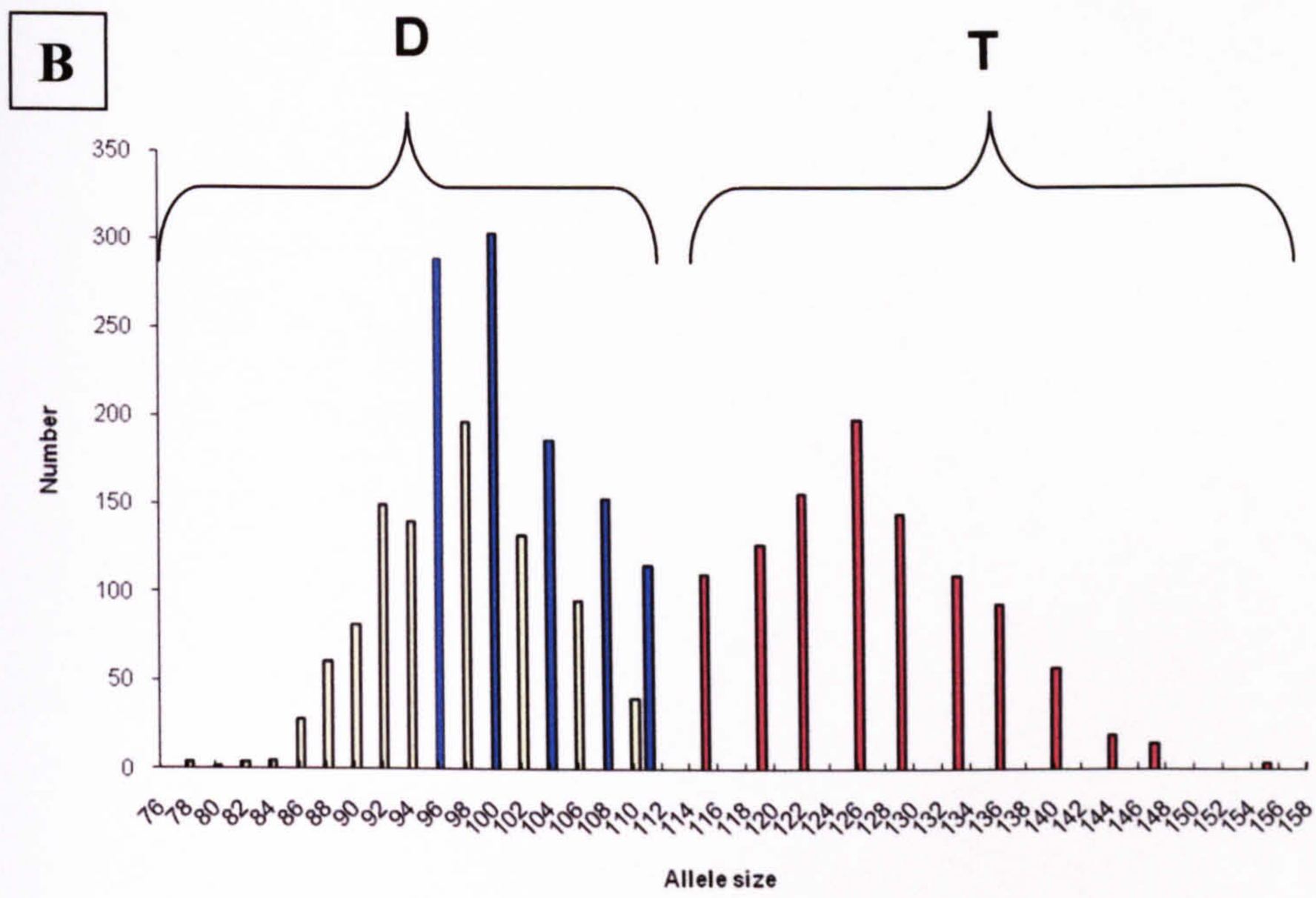


Figure 3.3 *CSH1.01* allele distribution and designation. A represents the raw data of the *CSH1.01* allele distribution. The x-axis shows allele size (bp) and the y-axis shows the numbers of times each allele was observed. B shows the binning allele distribution and *CSH1.01* allele designation. The x-axis shows the allele binning size, which is in SGEL code. Within experimental error, almost all alleles cluster very precisely, consistent with 2- and 40-base repeat length allelic variation. The first group of alleles designated T (tetranucleotide, shown in red), which are spaced at 4-base intervals, starts from allele 114 at the right end. The second group of alleles designated D (dinucleotide, shown in yellow and blue), which are spaced at 2-base intervals, starts from allele 111 at the left end. The D allele (in total, also designated as D1), contains alternating high (shown in blue) and low frequency (shown in yellow) allele bins. Exclusion of the bins showed in blue forms a subgroup of D1, designated D2 (yellow).

Table 3.25 *CSH1.01* allele frequencies in TTHHS (in dichotomy)

Genotype	TTHHS					
	Female			Male		
	Caucasian	South Asian	Other	Caucasian	South Asian	Other
D1D1	266	26	5	330	40	8
D1T	274	12	1	332	11	5
TT	89	3	1	103	4	3
Total	629	41	7	765	55	16
D2D2	67	3	1	80	12	3
D2T	139	6	1	184	5	2
TT	89	3	1	103	4	3
Total	295	12	3	367	21	8

In the TTHHS cohort, a total number of 1520 subjects were examined and genotype data were available for 1513 subjects. The genotype coded from alleles T and D1 was available for 1513 subjects and the coding using the D2 subset of D1 alleles was available for 706 subjects. The allele frequencies between ethnic groups were significantly different (χ^2= 15.834, p<0.001, Caucasians versus South Asians). The full list of allele frequencies at *CSH1.01*, represented in SGEL coding system, is shown in Table 3.26. Again, allele distributions were not significantly different between males and females (p=0.1620), but were significantly different between ethnic groups (χ^2 = 128.60, p<0.0001, Caucasians versus non-Caucasians. In the TTHHS, the majority (92%) of the subjects were Caucasians (1394 out of 1513) and therefore all subsequent genotype/phenotype analyses exclude the 96 South Asians and 23 subjects of other ethnic groups. The Caucasian genotypes were in HWE.

Table 3.26 *CSH1.01* allele frequencies TTHHS cohorts (in SGEL coding)

Allele size	TTHHS Female		TTHHS Male	
	number (n)	frequency (%)	number (n)	frequency (%)
78	3	0.22	1	0.06
80	1	0.07	1	0.06
82	1	0.07	3	0.18
84	2	0.15	3	0.18
86	15	1.11	13	0.78
88	33	2.44	28	1.67
90	44	3.25	38	2.27
92	49	3.62	101	6.04
94	67	4.95	73	4.37
96	132	9.75	157	9.39
98	82	6.06	114	6.82
100	139	10.27	164	9.81
102	57	4.21	75	4.49
104	74	5.47	112	6.70
106	35	2.58	60	3.59
108	71	5.24	82	4.90
110	16	1.18	24	1.44
111	62	4.58	53	3.17
115	47	3.47	63	3.77
119	65	4.80	62	3.71
122	69	5.10	87	5.20
126	91	6.72	107	6.40
129	62	4.58	83	4.96
133	41	3.03	69	4.13
136	46	3.40	48	2.87
140	29	2.14	29	1.73
144	9	0.66	11	0.66
147	8	0.59	7	0.42
151	0	0.00	0	0.00
155	1	0.07	3	0.18
158	3	0.22	1	0.06
Total	1354	100.00	1672	100.00

3.4.1.2 Phenotypic analysis in relation to *CSH1.01* in the TTHHS

Exclusion of 119 subjects from other ethnic groups meant that the analyses were based on a maximum of 765 Caucasian males and 629 females for T/D1 coding and 367 males and 295 females for T/D2 coding. Table 3.27 summarises the characteristics of this study population.

Table 3.27 Characteristics of population (Europeans only)

	Boys			Girls		
	N	Mean	SD	N	Mean	SD
Birthweight (kg)	710	3.41	0.52	596	3.25	0.5
Height (cm)	738	171.0	8.5	608	163.1	6.0
Leg length (cm)	738	86.8	4.8	608	80.7	4.2
BMI (kg/m2)*	736	20.3	N/A	607	21.1	N/A
Glucose (mmol/l)	640	5.12	0.44	543	4.96	0.46
Insulin (mmol/l)*	647	8.25	N/A	554	9.68	N/A

* geometric mean

The phenotypic comparisons for T/D1 coding appear in Table 3.28. There was no significant difference for females and there was only one significant result for males, namely a considerably higher birthweight in T/T males relative to D1/D1 males (0.159kg, 95% confidence interval +0.037kg to 0.281kg, p=0.0287 for trend). D1/T heterozygotes were intermediate but not significantly heavier than D1/D1 males. Girls showed a nonsignificant difference for birthweight, those of T/T genotype averaging 0.059kg heavier than those of D1/D1 genotype. The test for effects being different in girls and boys was not significant (p=0.16). Adjustment for maternal smoking during pregnancy and social class slightly increased the T/T vs. D1/D1 difference in

birthweight to 0.176kg (95% CI 0.045, 0.308) in males and the sex difference for interaction was of borderline significance (p=0.07).

Table 3.28 Phenotypic analyses for *CSH1.01* D1/T genotype

		Boys				Girls			
Variable	Genotype	Number	Difference from D1D1	95% CI	†p	Number	Difference from D1D1	95% CI	†p
Birthweight (kg)	TT	93	0.159	(0.037, 0.281)	0.0287	86	0.059	(-0.065, 0.183)	0.558
	D1T	312	0.019	(-0.064, 0.103)		265	-0.02	(-0.107, 0.067)	
	D1D1	306	0	.		246	0	.	
Height (cm)	TT	99	0.233	(-1.530, 1.997)	0.675	85	0.434	(-1.017, 1.884)	0.8229
	D1T	320	0.33	(-0.886, 1.546)		268	-0.335	(-1.349, 0.679)	
	D1D1	320	0	.		256	0	.	
Leg length (cm)	TT	99	-0.292	(-1.331, 0.747)	0.7047	85	0.16	(-0.862, 1.183)	0.8843
	D1T	320	0.051	(-0.666, 0.768)		268	-0.097	(-0.812, 0.618)	
	D1D1	320	0	.		256	0	.	
BMI (%)*	TT	99	2.20%	(-1.2%, 5.7%)	0.3363	85	-2.10%	(-6.0%, 1.9%)	0.158
	D1T	318	-0.10%	(-2.4%, 2.3%)		267	-2.20%	(-4.9%, 0.6%)	
	D1D1	320	0	.		256	0	.	
Glucose (mmol/l)	TT	90	-0.011	(-0.106, 0.084)	0.8469	76	0.025	(-0.085, 0.134)	0.7001
	D1T	269	-0.001	(-0.068, 0.066)		239	0.004	(-0.073, 0.080)	
	D1D1	272	0	.		228	0	.	
Insulin (%)*	TT	88	7.40%	(-4.0%, 20.2%)	0.2744	76	-8.00%	(-17.5%, 2.6%)	0.128
	D1T	172	1.10%	(-6.6%, 9.3%)		244	-3.20%	(-10.2%, 4.3%)	
	D1D1	278	0	.		234	0	.	

* Based on analyses of log(variable) andexpressed as % difference from D1/D1 group.
† p value are for trend across the three genotype group.
Anylases excluded non-Europeans and those refusing consent to DNA analysis. Insulin and glucose analyses excluded those not fasting.

In the analyses of the coding D2 and T allele groups (Table 3.29), again only one significant association emerged, namely birthweight in males. T/T males were 0.256kg heavier than D2/D2 males (95% confidence interval +0.079kg to 0.433kg); and D2/T male heterozygotes are more clearly intermediate being 0.132kg heavier than D2/D2 males (p=0.0045, trend test across the 3 groups).

Adjustment for maternal smoking during pregnancy and social class had little effect on the D2/D2 vs. T/T difference, which reduced to 0.242g (from 0.256g). The test for sex interaction was p=0.12 before adjustment and p=0.11 after adjustment. Amongst boys, the percentage of birthweight variance explained by the D1/T genotypes was 1% and for D2/T 2%. By contrast, the percentage of birthweight variance explained by maternal smoking during pregnancy was 3%.

Table 3.29 Phenotypic analyses for *CSH1.01* D2/T genotype

		Boys				Girls			
Variable	Genotype	Number	Difference from D2D2	95% CI	†p	Number	Difference from D2D2	95% CI	†p
Birthweight (kg)	TT	93	0.256	(0.079, 0.433)	0.0045	86	0.065	(-0.113, 0.242)	0.4036
	D2T	170	0.132	(-0.029, 0.292)		135	-0.007	(-0.172, 0.158)	
	D2D2	72	0	.		56	0	.	
Height (cm)	TT	99	-0.481	(-2.779, 1.817)	0.6601	85	0.764	(-1.246, 2.775)	0.3584
	D2T	174	-0.047	(-2.135, 2.041)		136	-0.56	(-2.400, 1.281)	
	D2D2	76	0	.		61	0	.	
Leg length (cm)	TT	99	-0.938	(-2.277, 0.400)	0.1489	85	0.631	(-0.819, 2.081)	0.3757
	D2T	174	-0.21	(-1.426, 1.007)		136	0.228	(-1.099, 1.555)	
	D2D2	76	0	.		61	0	.	
BMI (%)*	TT	99	0.50%	(-4.2%, 5.5%)	0.7892	85	-1.60%	(-6.9%, 4.0%)	0.5627
	D2T	172	-0.50%	(-4.8%, 4.0%)		135	-0.60%	(-5.6%, 4.5%)	
	D2D2	76	0	.		61	0	.	
Glucose (mmol/l)	TT	90	0.073	(-0.040, 0.186)	0.2575	76	-0.006	(-0.156, 0.144)	0.946
	D2T	148	0.083	(-0.020, 0.187)		120	-0.006	(-0.145, 0.132)	
	D2D2	64	0	.		54	0	.	
Insulin (%)*	TT	88	13.40%	(-2.5%, 31.8%)	0.1124	76	-13.20%	(-26.4%, 2.5%)	0.1127
	D2T	150	9.70%	(-4.4%, 25.8%)		123	-11.00%	(-23.5%, 3.6%)	
	D2D2	65	0	.		55	0	.	

* Based on analyses of log(variable) andexpressed as % difference from D2/D2 group.

† p value are for trend across the three genotype group.

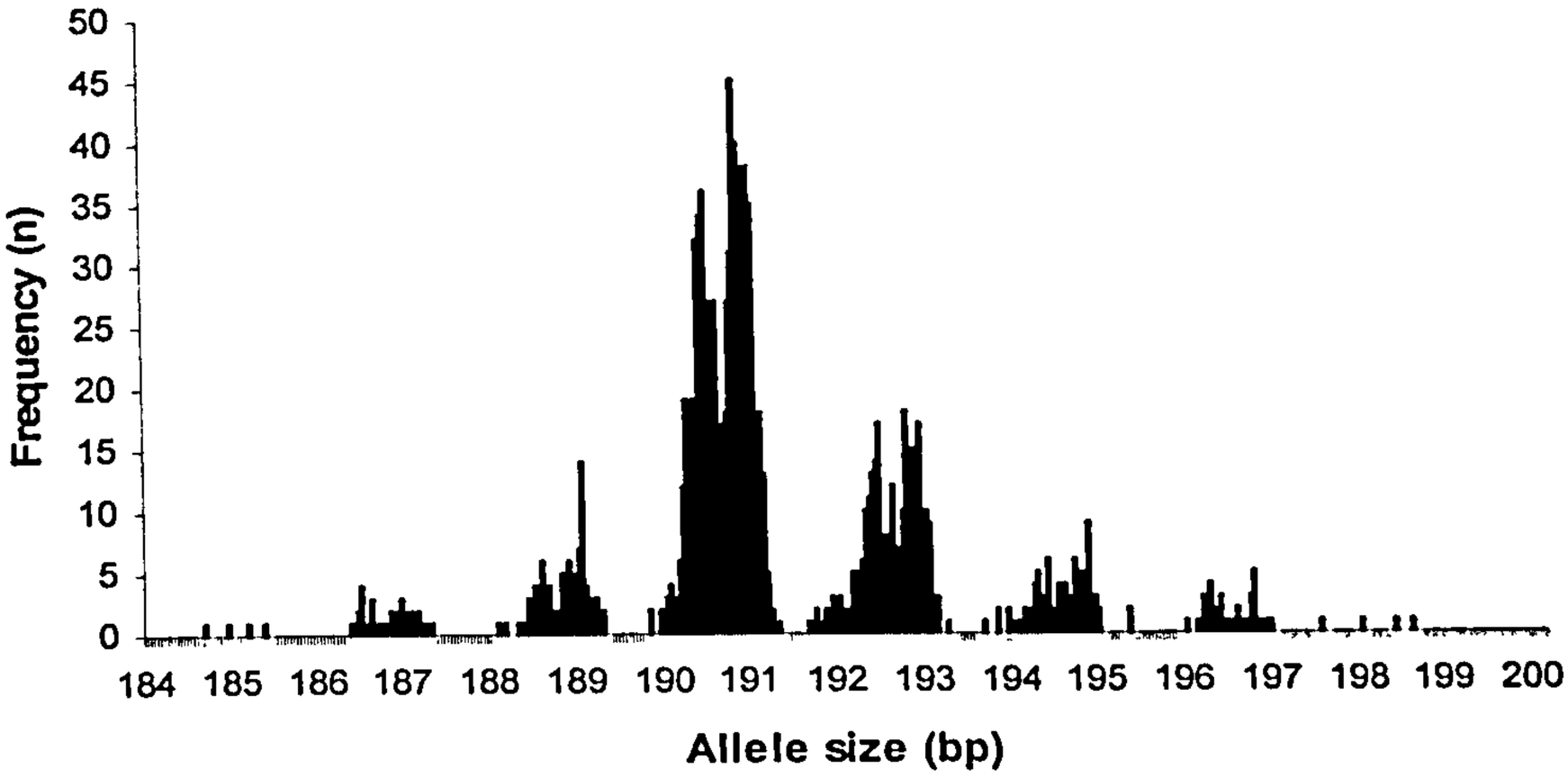
Anylases excluded non-Europeans and those refusing consent to DNA analysis. Insulin and glucose analysesexcluded those not fasting.

3.4.2 *IGF1*$(CA)_n$

3.4.2.1 *IGF1*$(CA)_n$ allele frequency and distribution

The most common allele for *IGF1*$(CA)_n$ was found to be 191bp in this project, which is probably equivalent to the previously reported 192bp allele [304]. The data were then transformed into SGEL coding, with the most common allele 100. Figure 3.4A and 3.4B shows the allele distribution for both raw and binned data.

A



B

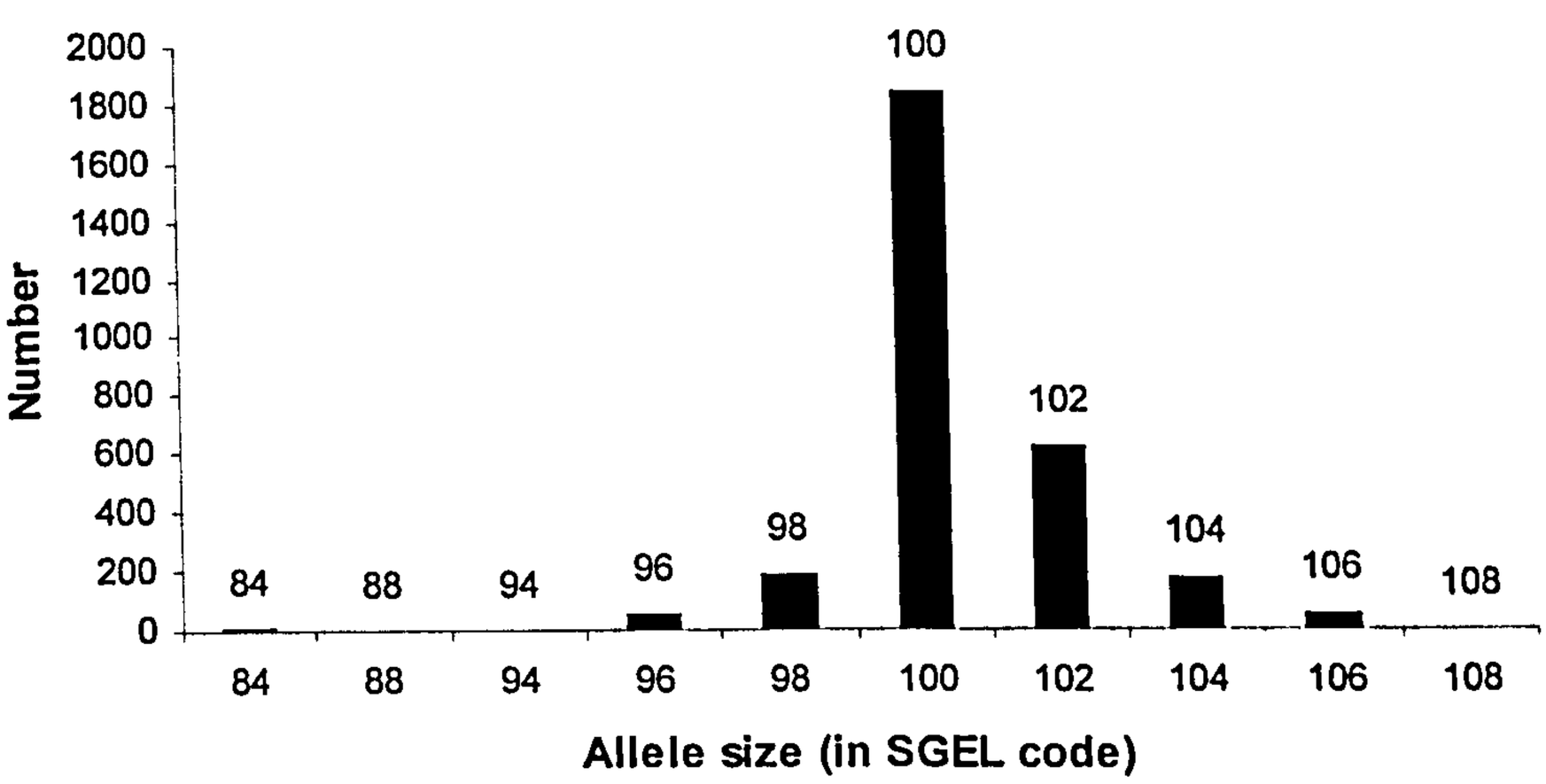


Figure 3.4 *IGF1*$(CA)_n$ allele distribution A. The raw data of *IGF1*$(CA)_n$ alleles. The x-axis shows allele size (bp). The y-axis shows the number of each allele observed. B. The binning allele distribution. The x-axis shows the allele binning size, which is in SGEL coding.

The total number of *IGF1*$(CA)_n$ genotypes was 1489 for the TTHHS cohort and the allele frequencies are shown in Table 3.30.

Table 3.30 *IGF1* (CA)n allele frequencies in TTHHS

Allele size	TTHHS	
	number	frequency
84	13	0.0044
90	1	0.0003
94	4	0.0013
96	56	0.0188
98	186	0.0625
100	1848	0.6206
102	629	0.2112
104	181	0.0608
106	57	0.0191
108	3	0.001
Total	2978	1

The allele distributions were not significantly different between males and females (χ^2= 5.3127 p=0.5044). Allele distributions between ethnic groups were significantly different. The χ^2 was 35.3793 (p<0.0001) when Caucasians were compared with non-Caucasians. The Caucasian genotypes were in HWE.

3.4.2.2 Phenotypic analysis in relation to *IGF1*(CA)n in the TTHHS

Phenotypic analyses were based on grouping all allele lengths which are not 192bp, and to compare 192/192; not 192/not 192; and heterozygotes. Here allele 100 was used to represent allele 192 (Table 3.31). Further grouping and analyses to compare the 100/100 and not(100/100) were shown in Table 3.32.

Table 3.31 Phenotypic analyses for *IGF1* (CA)n genotype (1)

Variable	genotype	Girls			Boys		
		diff. from 100/100	SE	†p.value	diff. from 100/100	SE	†p.value
Birthweight (kg)	Not100/not100	0.022	0.062	0.724	-0.075	0.063	0.228
	100/not100	-0.004	0.045	0.924	-0.062	0.044	0.163
	100/100	0			0		
Height (cm)	Not100/not100	0.358	0.689	0.604	-1.043	0.851	0.221
	100/not100	0.483	0.492	0.327	-1.032	0.593	0.082
	100/100	0			0		
Leg length (cm)	Not100/not100	0.057	0.489	0.907	-0.381	0.505	0.45
	100/not100	0.355	0.349	0.31	-0.733	0.351	0.037
	100/100	0			0		
BMI (%)*	Not100/not100	0.016	0.02	0.408	-0.004	0.017	0.81
	100/not100	-0.007	0.014	0.599	-0.008	0.012	0.484
	100/100	0			0		
Insulin (%)*	Not100/not100	0.088	0.05	0.08	0.014	0.054	0.79
	100/not100	0.033	0.036	0.358	-0.001	0.037	0.986
	100/100	0			0		

*Based on analyses of log (variable) and expressed as % difference from 100/100 group.

† p values are for trend across the three genotype groups.

Table 3.32 Phenotypic analyses for *IGF1* (CA)n genotype (2)

Variable	Genotype	Girls			Boys		
		diff. from 100/100	SE	p value	diff. from 100/100	SE	p value
Birthweight (kg)	Not(100/100)	0.003	0.042	0.952	-0.065	0.042	0.12
	100/100	0			0		
Height (cm)	Not(100/100)	0.452	0.462	0.329	-1.035	0.562	0.066
	100/100	0			0		
Leg length (cm)	Not(100/100)	0.281	0.328	0.392	-0.653	0.333	0.05
	100/100	0			0		
BMI (%)*	Not(100/100)	-0.002	0.013	0.907	-0.007	0.011	0.514
	100/100	0			0		
Insulin (%)*	Not(100/100)	0.047	0.034	0.166	0.003	0.035	0.939
	100/100	0			0		

*Based on analyses of log (variable) and expressed as % difference from 100/100 group.

In the TTHHS cohort, there was no evidence for association of this *IGF1* genotype with birthweight. However, males with the 100/100 genotype did show a leg length 0.65cm greater (p=0.05) than those not homozygous for allele 100. For the 100/100 genotype, there was also a trend in boys toward greater height (p=0.066), of greater magnitude (1.035cm) than the leg length association.

3.4.3 *ACE I/D* and *GH-CSH Bgl IIB*

3.4.3.1 Allele frequencies and linkage disequilibrium of *ACE I/D* and *GH-CSH BglIIB*

This study was part of the Lichfield Army Recruit Growth in Exercise Heart Study and was aimed at examining the hypothesis that some *ACE D* allele phenotypic associations may represent proxy marking of causal factors located in the *GH-CSH* gene cluster. In total, 1330 participants were invited to participate in the study, of which 900 participants volunteered. DNA was extracted from 890 participants. Genotyping data (both *ACE I/D* and *GH-CSH BglIIB*) were obtained for 847 participants, of which there were 605 with pretraining ACE activity data, 138 with both pretraining and posttraining ACE activity data, and 281 with both pretraining and posttraining left ventricular mass (LVM) data.

Among the 847 subjects in the study, the frequencies of the *ACE D* and *I* alleles were 0.52 and 0.48, respectively. The observed frequencies of 0.28, 0.47, and 0.25 for the *D/D, D/I,* and *I/I* genotypes, respectively, agreed with the frequencies predicted by HWE (χ^2 =2.7199; P=0.0991). The alleles of *GH-CSH Bgl IIB* were designated as *Cutter* and *Noncutter.* The frequencies of *GH-CSH Bgl IIB Cutter* and *Noncutter* alleles were 0.61 and 0.39 respectively. The observed frequencies of 0.37, 0.47, and 0.15 for *Cutter/Cutter, Cutter/Noncutter*, and *Noncutter/Noncutter* genotypes, respectively,

agreed with HWE (χ^2=0.0217, P=0.8828). There was linkage disequilibrium between *ACE I/D* and *GH-CSH Bgl IIB*, the D' coefficient =0.2999, P=0.0001.

3.4.3.2 Genotype-phenotype associations

Participant characteristics, such as age, body weight, body height, and BMI, did not differ significantly between *ACE I/D* and *GH-CSH Bgl IIB Cutter/Noncutter* genotypes. However, both *ACE I/D* genotypes and *GH-CSH Bgl IIB Cutter/Noncutter* genotypes were significantly different (P<0.0001) in relation to pre-training ACE activity using one-way ANOVA. The *D/D* genotype of the *ACE* gene showed a highest serum ACE activity (35.98±8.31), the *I/I* genotype the lowest (21.73±5.56), and the *I/D* intermediate (28.34±6.2). *ACE I/D* showed 6-7 units effect per deletion allele. *GH-CSH Bgl IIB Cutter/Noncutter* showed 2 units effect per allele, with the highest serum ACE activity in *Noncutter/Noncutter* individuals (31.23±8.58), lowest in the *Cutter/Cutter* individuals (26.84±7.96), and *Cutter/Noncutter* in the middle (29.94±8.6) (Table 3.33). LVM change before and after training showed significant difference among *D/D*, *I/D* and *I/I* genotypes of *ACE* gene, with the greatest change in *D/D* homozygotes (6.71±10.84), the smallest change in *I/I* homozygotes (2.45 ± 10.83) and *I/D* heterozygotes in the middle (3.06 ± 10.25) (P=0.0253). No significant difference was detected among *GH-CSH Bgl IIB* genotypes in relation to LVM change before and after training (Table 3.33).

Table 3.33 Serum ACE activity and exercise-induced LVM change of studied genotypes

	ACE				*Bgl IIB*			
	D/D	*I/D*	*I/I*	P value	*C/C*	*C/N*	*N/N*	P value
Number	178	283	144		220	292	93	
Pre ACE level	35.98±8.31*	28.34±6.2	21.73±5.56	<0.0001	26.84±7.96	29.94±8.6	31.23±8.58	<0.0001
Number	73	130	78		104	139	38	
LVM change*	6.71±10.84	3.06±10.24	2.45±10.83	0.0253	3.07±11.46	4.24±10.34	4.47±9.74	0.6482

*Values are expressed as mean ± SD;
* *C/C, C/N* and *N/N* represent *Cutter/Cutter, Cutter/Noncutter* and *Noncutter/Noncutter*, respectively
*LVM change=change for left ventricular mass (post-training minus pretraining)

Haplotype analyses of *ACE I/D* and *GH-CSH BglIIB* were performed using the Haplotype Trend Regression (HTR) program. Four haplotypes exist: *D/C*, *D/N*, *I/C*, and *I/N*, in which the first character represents *D* or *I* for *ACE I/D* and the second character *C* or *N* for *GH-CSH BglIIB*. As shown in Table 3.34, very strong association between haplotypes and pre ACE activity was found in this study. The overall p value was less than 0.0001, and each haplotype against all the others combined showed a significant difference ($p<0.0001$). There was no significant difference between haplotypes and LVM change ($p = 0.1066$).

Further analysis within genotype groups showed that a significant association only existed between *ACE I/D* genotypes and pre ACE activity (Table 3.35). Within the *C/C* or *N/N* genotype group, *D* allele carriers showed 6-7 unit higher pre ACE activity compared to *I* allele carriers ($p<0.0001$ for both groups). Within the *D/D* or *I/I* groups, there was, however, no significant difference between *GH-CSH BglIIB C* and *N* allele carriers in relation to pretraining ACE level, indicating that the association observed between *GH-CSH BglIIB* genotypes and pretraining ACE levels described above is a result of the confounding effect of LD, rather than direct causality. None of the haplotypes was significantly associated with exercise-induced change in LVM.

Table 3.34. Associations between haplotypes and Pre ACE level and LVM response

Haplotype	Pre ACE level			LVM change		
	Estimated frequency	Mean	p value	Estimated frequency	Mean	p value
*D/C**	0.2623	32.41	<0.0001	0.2297	4.8528	0.1899
D/N	0.2658	32.77	<0.0001	0.2614	5.1189	0.0655
I/C	0.3427	24.77	<0.0001	0.3877	2.7646	0.0543
I/N	0.1292	25.64	<0.0001	0.1212	2.61	0.2474
Total	1		<0.0001	1		0.1066

** D/C: ACE Deletion /GH-CSH Bgl IIB Cutter*
D/N: ACE Deletion /GH-CSH Bgl IIB Noncutter
I/C: ACE Insertion /GH-CSH Bgl IIB Cutter
I/N: ACE Insertion /GH-CSH Bgl IIB Noncutter

Table 3.35. Associations between haplotypes and Pre ACE level and LVM response in homozygotes

Genotype	Haplotype	Pre ACE level			LVM change		
		Estimated frequency	Mean	p value	Estimated frequency	Mean	p value
Del/Del	[1]*D/C*	0.514	35.54	0.2930	0.5	3.3027	0.7555
	D/N	0.486	36.43		0.5	2.8193	
Ins/Ins	*I/C*	0.7431	21.46	0.1677	0.7821	1.5	0.1751
	I/N	0.2569	22.52		0.2179	3.8974	
Cutter/Cutter	*D/C*	0.4205	31.04	<0.0001	0.351	3.6749	0.1532
	I/C	0.5795	23.8		0.649	1.1866	
Noncutter/ Noncutter	*D/N*	0.6613	33.35	<0.0001	0.671	4.6615	0.9639
	I/N	0.3387	27.07		0.3289	4.5764	

1. D/C: ACE Deletion /GH-CSH Bgl IIB Cutter
D/N: ACE Deletion /GH-CSH Bgl IIB Noncutter
I/C: ACE Insertion /GH-CSH Bgl IIB Cutter
I/N: ACE Insertion /GH-CSH Bgl IIB Noncutter

Chapter 4 Discussion

Candidate gene association analysis has been one of most commonly used approaches in genetic epidemiological studies in the past two decades, principally because it has been the most tractable mode of analysis. Even now, with the whole genome association analysis becoming a reality, candidate gene analysis still plays an important role in annotating the aetiology of common complex diseases and traits. My PhD research project adopted candidate gene approach, focusing on two major health issues, smoking and growth, in large, well characterised cohorts.

4.1 DNA banks -- the foundation of genetic epidemiology studies

To study genetic risk factors for common diseases, researchers have collected DNA specimens in large epidemiological studies. The size of these studies continues to expand and several national 'BioBanks' have been established [369]. A DNA bank, according to the American National Bioethics Advisory Commission, is 'a facility that stores extracted DNA, transformed cell lines, frozen blood or other tissue, or biological materials, for future DNA analysis' [370]. There are major challenges in dealing with a DNA bank, including social, ethical, and legal issues that impact on sample collection, storage and DNA banking practices for biomedical research and in the control of the quality of the DNA banks. A wide variety of information originates from each sample of blood or other tissue. Study power is dependent on the quality of sample analyses in addition to quality of associated information derived from questionnaires, clinical history and clinical examination. In genetic epidemiological studies, the power of association between genotypes and phenotypes is dependent both on the quality of phenotyping and of genotyping and of correctly linking of phenotypes to genotypes. Samples must be tracked through numerous steps between subject or patient and

genotype data validation. Although good laboratory practice is always emphasised in every step, manual handling in blood collection, data inputting, DNA extraction, and genotype data acquisition can lead to random errors. In genetic studies, only one phenotype, sex, has a perfect and binary correlation with a genotype. Sex is usually implicit in a subject's name and is recorded explicitly. In mixed sex studies, it is therefore advantageous for the purposes of quality control to keep sexes mixed during the steps from acquisition to DNA bank, in order to be able to check later for sample swaps. We designed a duplex PCR, combining a *MAOA* amplicon marking the X chromosome and a *DDX3Y* amplicon from the Y chromosome, to test a subject's sex. We applied a simple economical 768-well electrophoresis system, which is microplate compatible, for examination of this duplex PCR. Genotypic sex from this assay can then be compared with phenotypic sex documented from the clinical and epidemiological phase of the study. We undertook such analyses during the course of set up of the DNA bank from the TTHHS cohort.

From the result of the cohort, six pairs of the subjects (0.79%, 12 out of 1520) were discordant in gender between phenotype and quality control genotype data. All of these mismatches occur in pairs of adjacent samples on the Prep code list, strongly suggesting that the errors occurred during the DNA extraction process. In a random allocation model, any male-female sample swap or microplate array location blunder or array orientation or array replication blunder (e.g., 180° rotation) would be detected by genotyping assay, but male-male and female-female swaps would not be detected.

Defining frequencies, f_M and f_F (and $f_M + f_F = 1$), $2K_1N\ f_M\ f_F$ swaps would be detected, and $NK_1(f_M^2 + f_F^2)$ swaps would not be detected, where N is the study size and K_1 the swap rate, and assuming no 'interference' between swaps. If S swaps are detected, then $K_1 = S/2Nf_Mf_F$ and $K_1N = S/2f_Mf_F$ swaps are predicted in total. In the TTHHS, f_M is

0.552, f_F is 0.448. K_1=12/(2*1520*0.552*0.448)=0.016. The predicted total number swapped K_1N=0.016*1520≈24. The number undetected is the total number minus the detected number, which is 24-12=12. The predicted number of male-male swap is therefore 12*0.552=6.6, while the predicted number of female-female swap is 12*0.448=5.4. Overall therefore it can be estimated that 24 errors were introduced during processing (24/1520 = 1.58%) but half of these were identified and corrected by detecting sex mismatches.

This quality control test, in conjunction with the genotypes of *MAOA* markers, is the first experiment designed and applied to the validation of a large mixed gender DNA bank. It could not detect all the errors caused by careless operations such as sample swapping during DNA extraction, or errors from data inputting, but it estimated to halve the error rate.This duplex PCR assay is now one of the routine approaches in our laboratory for the quality control of gender mixed DNA banks.

4.2 Assay development and validation

4.2.1. Assay development

A variety of genotype methods were designed and applied in this project. An integrated set of two-step PCR ARMS was developed combining some designs taken from the literature, together with a new long PCR design combined with formats generic to our laboratory, specifically allele-specific PCR combined with MADGE to genotype *CYP2A6*1A, *1B, *2, *4, *5, *9* and **12*.

Di-primer PCR combined with self-designed sizing markers were used for genotyping VNTRs (*MAOA*, *TH01* and *DRD4*). The *TH01* microsatellite was typed by PCR and MADGE as previously described [361]. The *MAOA* 30bp VNTR and *DRD4* 48bp

VNTR were genotyped using the same method, except that they were examined on dry gels without thermal control. This innovative use of a dry gel system for genotyping VNTRs was both cost-efficient and time-efficient.

Tri-primer and tetra-primer ARMS were designed for SNPs *MAOA* T1460C, *5HT2A* T102C and *DBH* G1368A. The 96, 384 and 768 MADGE gels were chosen to check PCR products according to the size of the PCR fragments.

4.2.2 Assay validations

Assay validations were performed by several approaches: double checking by other colleagues, repetition by a different method (e.g. capillary sequencing for *TH01* assay), HWE testing and comparison of allele frequencies with other studies.

In this project, genotype data were thoroughly double-checked by several colleagues. PCRs were repeated for those samples whose genotypes were not clear and convincing. If the repeat results were still not clear, they were excluded from the phenotype-genotype analyses.

In population genetics, the Hardy–Weinberg law, named after G. H. Hardy and Wilhelm Weinberg, states that, under generalised conditions after one generation of random mating, the genotype frequencies at a single gene locus will become fixed at a particular equilibrium value. It also specifies that those equilibrium frequencies can be represented as a simple function of the allele frequencies at that locus. These conditions (assumptions) include an infinitely large population, that individuals mate at random, that no new mutations appear in the gene pool, there is no migration into or out of population and there are no genotype-dependent differences in the ability to survive and reproduce, ie no selection for specific alleles. If all these assumptions hold, at least

approximately, then the population is in Hardy-Weinberg equibrilium with respect to the marker of interest. In the simplest case of a single locus with two alleles A and a with allele frequencies of p and q, respectively, the Hardy-Weinberg law predicts that the genotypic frequencies for the AA homozygote to be p^2, the Aa heterozygote to be 2pq and the other aa homozygote to be q^2, with $p^2 + 2pq + q^2 = 1$.

The Hardy-Weinberg model enables us to compare a population's actual genetic structure over time with the genetic structure we would expect if the population were in Hardy-Weinberg equilibrium (i.e. not evolving). If genotype frequencies differ from those we would expect under equilibrium, then either one or more of the assumptions above are being violated, or there is something wrong with the genotyping. At the population level, deviation from HWE can be due to inbreeding, population stratification or selection either for the marker itself or for another variant that is in LD with the marker. Genotyping errors may arise from the presence of a common deletion polymorphism, mutations in the PCR-primer binding site, and effects of local sequence context that may affect the performance of specific assays. In addition it has become apparent that many SNPs in the public databases are not simply binary markers as first thought but instead fall in regions of copy number variation, and thus may yield complex and unexpected genotypes [371].

Testing for deviation from HWE can be carried out using χ^2 test, using the formula $\chi^2 = \sum(\text{Observed-Expected})^2/\text{Expected}$. The χ^2 test is easy to compute, but the χ^2 approximation can be poor when there are low genotype counts, and it is better to use either Yates' correction or Fisher Exact test.

HWE was tested for each marker in this project. When the *CYP2A6*2* assay was tested initially, it did not follow HWE, and an excess of apparent **2* homozygotes. A

migration effect was excluded by testing Caucasians only. The existence of a 'null allele' then emerged to explain the deviation. A previous study by Gu *et al*. genotyped the same alleles in 1,032 Caucasians and suggested a system like the ABO blood groups, with alleles *L* (wild type), *H* (*2 mutant) and *O* (null allele) [164]. The *O* allele lacks a long PCR product and would be scored as homozygous, leading to a falsely elevated rate of *2 homozygotes. In the present study, one allele is *2 for the five apparent *2 homozygotes. The other allele could be missing any of the targets for long PCR primers 2A6ex1 or 2A6R1, or short PCR primers E3F or E3R or 2A6wt, leading to no target band and thus being scored homozygous. This assumption is supported by the existence of *CYP2A6*12*, reported by Oscarson *et al*. in 2002 [173]. The *12 allele carries an unequal crossover between the *CYP2A6* and *CYP2A7* genes in intron 2, which results in a hybrid allele where the 5' regulatory region and exons 1-2 are of the *CYP2A7* origin and exons 3-9 are of *CYP2A6* origin, resulting in 10 amino acid substitutions compared to the *CYP2A6*1* allele. There are 5 nucleotide changes, resulting in 4 amino acid substitutions, which are within the forward primer (2A6ex1) sequence for the 8kb long PCR in the present study (Figure 4.1). Therefore, when the 8kb long PCR was amplified, alleles with *12 variation could not amplify any long PCR product. This assumption has been confirmed by genotyping *CYP2A6*12* in TTHHS cohort. Three out of five apparent *2/*2 homozygotes turned out to be *2/*12 heterozygotes.

```
          -148                                                                -76
CYP2A6*1    gtcccaagctaggcaggattcatggtggggcatgtagttgggaggtgaaatgaggtaattatgtaatcagcca
CYP2A6*12   gtcccaagctaggtggcattcatggtggggcgtgtagttgggaggtgaaatagaggtgattatgtaattagcca
          -75                                                                 -1
CYP2A6*1    aagtccatccctctttttcaggcagtataaaggcaaaccaccccagccgtcaccatctatcatcccactaccacc
CYP2A6*12   aagtccatccctctttttcaggcagtataaaggcaaaccaccccacccatcaccatctgtcatcctcactaccacc
          1                                                                   75
CYP2A6*1    ATGCTGGCCTCAGGGATGCTTCTGGTGGCCTTGCTGGTCTGCCTGACTGTGATGGTCTTGATGTCTGTTTGGCAG
CYP2A6*12   ATGCTGGCCTCAGGGCTGCTTCTGGTGGCCTTGCTGGCCTGCCTGACTGTGATGGTCTTGATGTCTGTCTGGCAG
CYP2A6.1    M  L  A  S  G  M  L  L  V  A  L  L  V  C  L  T  V  M  V  L  M  S  V  W  Q
CYP2A6.12   M  L  A  S  G  L  L  L  V  A  L  L  A  C  L  T  V  M  V  L  M  S  V  W  Q
          76                                                                  150
CYP2A6*1    CAGAGGAAGAGCAAGGGGAAGCTGCCTCCGGGACCCACCCCATTGCCCTTCATTGGAAACTACCTGCAGCTGAAC
CYP2A6*12   CAGAGGAAGAGCAGGGGGAAGCTGCCTCCGGGACCCACCCCACTGCCCTTCATTGGAAACTACCTCCAGCTGAAC
CYP2A6.1    Q  R  K  S  K  G  K  L  P  P  G  P  T  P  L  P  F  I  G  N  Y  L  Q  L  N
CYP2A6.12   Q  R  K  S  R  G  K  L  P  P  G  P  T  P  L  P  F  I  G  N  Y  L  Q  L  N
```

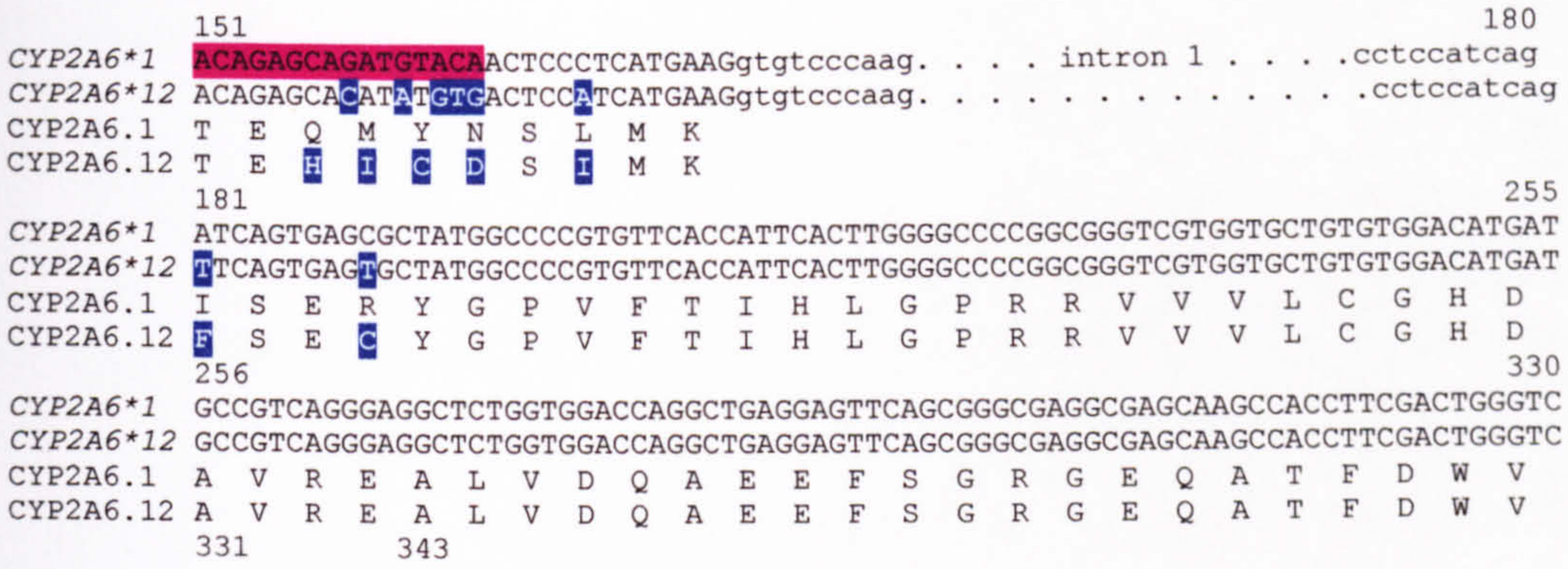


Figure 4.1 The DNA sequence of the *CYP2A6*12* allele as compared to *CYP2A6*1* in the region of –148bp to exon 2. Exon sequences are shown in upper case letters, and the blue boxes represent differences between the two alleles. Purple shaded sequence is forward primer 2A6ex1 for 8kb long PCR[173].

The contribution of the high frequency of *MAOA* 1460T homozygotes in the 96 South Asians almost led to significance of HWE test (p=0.0601) in the overall sample but the Caucasian subset fitted the expectation precisely (p=0.9242) (Table 3.18). The reason for this is unclear but is unlikely to be a genotyping problem otherwise the effect would have been seen across populations. Possibly it could be an artefactual consequence of the relatively small number of South Asians in the cohort, but potentially this could indicate selection in this region (i.e. non-random mating). The allele frequencies of the *MAOA* VNTR in South Asian group in our study is consistant with those in other two studies of the South Asian population [372, 373].

The allele frequencies for each marker were also compared with those of Caucasian populations in other literature studies for the validation of genotyping assays. Comparison results showed that the allele frequencies for each marker were similar to those of reported [167, 215, 321, 372, 374, 375].

4.3 One of the largest genetic epidemiology studies relating to adolescent smoking

It has been observed that smoking behaviour has a marked degree of heritability (over 50% in the USA and UK [376]. Among adult cigarette smokers, specific associations have been claimed for several human genotypes. *CYP2A6* is of particular interest, encoding hepatic cytochrome P4502A6 which catalyses the major step of nicotine inactivation, C-oxidation to cotinine. As described in detail in Chapter 1, two common alleles of *CYP2A6* have been identified, namely *CYP2A6*1A* and *CYP2A6*1B*. *CYP2A6*1B* represents a gene conversion (by the adjacent *CYP2A7* pseudogene [147] in the 3' untranslated region of exon 9. Some reports have suggested that the **1B* allele correlates with a higher rate of nicotine metabolism [147] [191] [377] and smoking behaviour [378] [379]. The presence of inactivating alleles *CYP2A6*2* (which contains an inactivating amino acid substitution) and *CYP2A6*4A* (which contains a major deletion) are associated in adults with significantly lower self-reported cigarette consumption, lower cotinine levels and lower carbon monoxide exhalation [91, 168]– suggesting that inactivation of nicotine metabolism is associated with lower levels of cigarette smoking. Schoedel *et al.* (2004) demonstrated that slow nicotine inactivators (*CYP2A6*2*, *CYP2A6*4*, *CYP2A6*9*, *CYP2A6*12*) were less likely to be adult smokers, smoked fewer cigarettes per day and had an earlier age of first smoking (only dependent smokers) [193]. Gu et al., (2000) reported that carriers of the *CYP2A6*2* allele tended to start smoking later and had an increased likelihood of quitting compared with those subjects with wild type genotype [164]. However, several smaller studies which measured fewer *CYP2A6 alleles*, did not find any association between self-reported cigarette smoking and *CYP2A6* genotype [169, 194-198]. A recent study in adolescents suggested that subjects with 1 or 2 copies of *CYP2A6*2* or **4* tended to smoke less on

average, consistent with the earlier adult studies [380]. However, they appeared to be at substantially increased risk of becoming nicotine dependent. No other studies examining the relationship between *CYP2A6* genotypes and smoking initiation in adolescence have been described.

Other genes implicated in smoking uptake have included *MAOA*, *DBH*, *DRD4* and *5HT2A*. *MAOA* encodes monoamine oxidase A, which is important in neuronal degradation of biogenic amines. *DBH* encodes dopamine beta hydroxylase which converts dopamine to noradrenaline in nerve terminals. McKinney et al [215] found that allelic variations in the *DBH* and *MAOA* genes predict whether a person is a heavy smoker and how many cigarettes they consume. The *DBH* 1368A allele was more frequent in heavy smokers than in light smokers. However, heavy smokers were less likely to possess the *MAOA* 1460C allele than were light smokers. The *MAOA* gene also contains a minisatellite length polymorphism in its promoter region and this appears to have functional effects. Two alleles (3 and 4 repeats) are most common and represent more than 97% of observed alleles. Ito *et al.* (2003) reported that the median Fagerstrom test for nicotine dependence (FTND) score among male current smokers was significantly higher with than without the 4-repeat allele (5.8 and 4.7, respectively)[222]. *DRD4* encodes the dopaminergic receptor 4; a VNTR in exon 3 of this gene has been related to addictive behaviours and substance abuse [236, 238, 239, 241]. Having the L allele (6-10 repeats, commonly 7) predicted increased smoking, a shorter time before the first cigarette in the morning and an earlier age of smoking initiation compared to S allele (2-5 repeats, commonly 4) carriers in an African-American population[244]. The *5HT2A* gene has been associated with emotional disorders and alcohol dependence, two disorders that are strongly associated with smoking behaviour. The 102C>T polymorphism was reported to be associated with

maintenance, but not initiation of the smoking habit in a Brazilian population [267]. Set against these positive functional and genetic epidemiological claims, a number of reports have not found association between genotypes of these genes and smoking behaviour [246] [381].

The smoking study described in this thesis based on the TTHHS is one of the largest genetic epidemiological studies of genotypes relevant to smoking behaviour, both in terms of the number of genes studied and for *CYP2A6*, in the number of inactivating alleles examined. There are very few studies of the behavioural molecular genetics of smoking inception in adolescence, a critical period in the formation of the lifelong habit of smoking, which is in turn the biggest risk factor for CHD and cancer in later life. Markers which have been implicated in addiction in previous studies of adults were not found to associate either with likelihood of trying, or of continuing, smoking in this study, except for *CYP2A6*. *CYP2A6* showed significant associations both with likelihood of remaining a current smoker; and with smoking behavioural traits and cotinine levels among current smokers.

In adult studies, genotype variations in dopaminergic markers (*DBH*, *MAOA*, *TH01*, and *DRD4*) showed significant differences mainly in smoking dependence or regular smokers [211, 212, 215, 222, 244]. In our study, subjects were 13-18 year old teenagers, a group in which smoking dependence is less established. It is possible that associations previously observed in dopamine markers will therefore be less evident in these teenagers. However, there were no trends for any of the markers examined in our study at age 18 when many of the subjects were smoking >10 cigarettes per day.

Although the significance of my findings diminished after correction for multiple testing, several elements deserve discussion. For *CYP2A6*, the HI individuals, predicted

to metabolise nicotine at half the rate of the 'wildtype' individuals, were twice as likely to be current smokers at age 18, compared with the WT1B individuals. Additionally, the partial haploinsufficiency (**9* and **12*) and haploinsufficiency (**2, *4* and **5*) current smokers reported shorter times (respectively 4.22 and 5.69 hours) since their last cigarette compared with the WT1B (8.35 hours) and WT1A (7.15 hours). This also seemed to be reflected in trends in cotinine levels being lower in the 'wildtype' group, again contrasting with published studies of older adults [147, 191, 382]. It should be noted that unlike older adults, the mean consumption was much lower at age 13-15 years (typically 1-6 cigarettes/week) and modest at age 18 years (typically 8-10 cigarettes/day) in smokers. Moreover, interpretation of the cotinine data requires caution given that CYP2A6 is involved in metabolising cotinine to 3-hydroxycotinine as well as nicotine to cotinine [383] leading to suggestions that assessment of CYP2A6 activity might best be based on the ratio of cotinine to 3-hydroxycotinine [384]. Lack of saturation may also explain why cotinine levels are higher for HI and PHI in situations where nicotine dependence has not been achieved, in particular at ages 13-15.

The overall picture of results in predicted slow inactivators (increased likelihood of smoking and related traits) appears paradoxical, since amongst older adults, the 'slowest inactivators' have been associated with lower likelihood of smoking [193], lower consumption [91, 168], and higher likelihood of quitting [164]. We hypothesise that the nicotine pathway may be in a different state in teenagers starting to smoke, compared with older adult addicted smokers. In the early phase (age 13-15 years) of smoking, typically 1-6 cigarettes per week, the pathway of nicotine metabolism will not be saturated and the gaps between cigarettes may lead to zero levels of nicotine (or metabolites). In this situation, a slow inactivator will retain nicotine for longer. Plausibly, the increased temporal exposure to nicotine (and to higher levels) could

enhance likelihood of becoming dependent (an 'occupancy' hypothesis). In contrast, an addicted adult would titrate their timing and dose to match central nervous system satiation by nicotine. A slow inactivator would need a lower dose to achieve a satiety level which is dependent on CNS factors not on CYP2A6. In this model, *CYP2A6* inactivating alleles lead to earlier nicotine dependence but to lower long term consumption. Why the intermediate group of slow metabolisers are not at higher risk requires further study. However an alternative hypothesis is that despite our study being one of the largest to date, the results we found may be an artefact of the population that was studied or a statistical anomaly.

Previous studies reported a higher rate of cotinine formation in *1B carriers compared with *1A wildtype [147, 191, 377]. In our study, we compared *1B carriers with *1A homozygotes (excluding PHI and HI subjects) in relation to cotinine levels and smoking status. We did not find any significant difference in cotinine levels between these two groups, but we did find that at age 18, the odds ratio (OR) of being a current smoker vs. ex-smoker was 1.44 (95%CI 1.01, 2.06) for WT1A as compared with WT1B, which suggests that *1B carrier confers a lower risk of cigarette addiction than does *1A homozygosity, but again this effect diminished after applying Bonferroni correction. Vasconcelos et al., (2005) reported that *CYP2A6*1B* showed a 14-fold lower probability for being tobacco dependent in Brazilians self-classified as white [378]. Our result is consistent with Vasconlelos's, although our signal is much lower. It should be noted that *9 (partial haploinsufficiency) mutation is only found on *1A background [385]. We confirmed this in our data by linkage disequilibrium analysis. Any study which compared *1A with *1B for cotinine formation without taking into account the more recently defined *9 allele [176] could reach an erroneous conclusion because the *9 subgroup of *1A will lower the apparent mean cotinine index for *1A relative to *1B.

The statistically significantly higher cotinine level in the PHI group but not in the HI group compared with the WT1B group at age 18 (Table 3.24), may reflect sample size and consequent increased statistical power since the PHI group was much larger than the HI group. However, the mean cotinine level in the HI group was similar to the PHI group. In a similar way, some inconsistencies in the literature may also reflect differences in power.

Our data are consistent with the recent report by O'Loughlin and colleagues [380]. In that study, "very slow inactivators" (defined as one or more allele of *2 or *4, or individuals who were homozygous for *9 or *12) were more likely to progress to nicotine dependence during adolescence than either "intermediate inactivator" (defined as individuals heterozygous for *9 or *12) or "normal inactivators" (wildtype). Prospective follow-up of future establishment of smoking habit in TTHHS should have sufficient power [164] to determine the long-term influence of the *CYP2A6* locus on smoking behaviour, and whether low activity of CYP2A6 due to haploinsufficiency may enhance smoking dependence at an early age but reduce it at later ages. This may suggest approaches to primary prevention of smoking based around nicotine metabolism and nicotine-driven behaviour.

4.4 Phenotype-genotype association in growth pathway genes

4.4.1 *GH-CSH* gene polymorphism, birthweight, and metabolic traits

Based on the Hertfordshire study [294], we hypothesized that any association between *CSH.01* and birthweight or adolescent anthropometric and metabolic traits might only be expected in males, that the 'D2' coding of marker *CSH1.01* might show a greater magnitude of effect than the 'D1' coding, and that heterozygotes might emerge showing intermediate values for the 'D2' coding. In the TTHHS girls, there was no association

between *CSH1.01* genotypes and birthweight, nor was there any relationship to metabolic traits. In the TTHHS boys, however, there was significant association between *CSH1.01* alleles and birthweight, but unexpectedly the effect of TT genotype in TTHHS was opposite to that found in the Hertfordshire study, where the TT genotype showed a lower weight at one year of age compared with D1D1 or subgroup of D1D1-D2D2. In the TTHHS teenage boys, the T alleles conferred high weight at birth, T/T genotype displaying approximately 7.5% higher mean birthweight than D2/D2 genotype. Moreover insulin levels were also higher in this group though not significantly and not to the same degree as has been reported in adults [294]. Also consistent with the adult data was that a signal of intermediate magnitude in heterozygotes only emerges when the subdivision of D1 alleles to D2 is carried out. This suggested a quantitative effect per allele of some sequence variant(s) within the *GH-CSH* region, most clearly marked by the D2 allele lineage(s) of *CSH1.01*.

In the Hertfordshire study, there was strong evidence showing that the TT genotype revealed a lower one year weight and later life metabolic traits [294]. Although the TT genotype was not related to lower birthweight, we speculated that it might be due to the survival bias considering the higher fetal and infancy mortality rates, especially in the economic and healthcare conditions of 1930-1940 Britain. We planned to confirm our speculation by examining the same marker in a better nourished cohort and expected to demonstrate a significant association between T alleles and lower birthweight.

However, the results of TT genotype on the boys' birthweight in the TTHHS are not what we expected. The reasons for the discrepancy are not clear. It is known that both low birthweight and later life diseases are multifactorial complex traits. Genetic factors are only part of the aetiologies, each individual marker accounts for only a minor effect of the phenotype and the effect of genetic polymorphisms is influenced by

environmental factors, gene-gene interactions and gene expression. In the TTHHS, factors such as mothers' smoking status, mothers' heart health status and the accuracy of birthweight record might influence or overshadow the effect of TT genotype. It was reported that the most important cause of low birthweight in developed countries is smoking in pregnancy. On average, smokers' babies weigh between 150 and 250g less than non-smokers' babies, and twice as many of the former weigh less than 2500g [386]. However, adjustment for maternal smoking during pregnancy and social class in the present study had little effect on the D2/D2 vs. T/T difference, which reduced slightly to 0.242g (from 0.256g). The accuracy of recording the birthweight is another confounding factor. In the TTHHS cohort, when subjects with birth weight data were from more than one source, data were taken in order of preference from 1) birth record data, 2) birth registers, and 3) parental recall (questionnaire data) [356] The accuracy is questionable if the birthweight was from parental recall. Each of the confounding factors might only show a small influence on the effect of TT genotype on mean birthweight, however, the combination of the known and unknown factors may have a significant effect such that the real specific allele functions becomes obscured.

Recently, Freathy et al. (2006) repeated this study in a Caucasian cohort of 1396 subjects (fathers, mothers and children) from a consecutive birth study for the *CSH1.01* microsatellite and analysed genotypes for association with 1-year weight in boys and fasting insulin in fathers [321]. They did not find an association between the *CSH1.01* genotype with fasting insulin or weight at one year. The discrepancies between the three studies might be a reflection of substantial heterogeneity between the samples as a result of age differences (i.e. Day and Freathy studies), phenotype difference (birthweight vs. one-year weight between this study and Freathy study), or both factors (this study and Day study). We also considered the possibility of statistical artefacts. Power

calculations (using software 'Java applets for power and sample size') showed that there is a 73% power to detect the birthweight difference of 159g between TT and D1D1, and 88% power to detect 256g difference between TT and D2D2 genotypes in this study (different direction from Day study, e.g. TT boys conferred high weight at birth). Freathy et al. reported an 80% power to detect decreases in boys' weight at 1 year of 830 g for TT relative to D2D2 homozygotes (p value < 0.05; same direction as Day study). Thus, although studies of the size employed here are sufficiently powered to detect weight differences with a reasonable probability, larger cohorts would obviously provide more power and are no doubt needed to confirm the relationship between *CSH1.01* genotypes and birthweight and/or one-year weight.

4.4.2 Association of *IGF-I* CA microsatellite with leg length in adolescence

In the TTHHS cohort, there was no evidence for association of the *IGF1* CA microsatellite genotype with birthweight, consistent with two other studies [293, 295]. However, TTHHS males of 192/192bp genotype did show a leg length 0.65cm greater (p=0.05) than those not homozygous for allele 192bp. For 192/192 genotype, there was also a trend toward greater height (p=0.066), of greater magnitude (1.035cm) than the leg length association. In the Dutch study (mean age 66y), there was a 2.7cm height difference between homozygous groups, with a trend toward tallness according to number of 192bp alleles [304]. It is notable that while Frayling et al. did not find significant height association in young adults (mean age 25y), the Z/Z allele group (which corresponds with 192/192) were on average 1.71cm taller than the opposite homozygous group [295]. It appears that *IGF1* genotype exerts a small effect on both leg length and height but very little or no replicable effect on birthweight, BMI, or insulin levels. Genotypic determination of IGF1 expression may be one contributor to

the observed associations between leg length and height in childhood and adult health in later life.

4.4.3 Serum ACE activity and LV growth: combined analysis of the linked growth hormone (*GH-CSH*) and *ACE* gene regions

As part of the Lichfield Army Recruit Growth in Exercise Heart Study, we investigated the linkage disequilibrium between *ACE* and *GH-CSH*, as well as their genotypes and haplotypes in relation to ACE activity and exercise-induced LV growth response. The findings from this study were consistent with the previously reported association between D allele dose and enzymatic levels [338] and, showed no association between a *BglII-B* marker in the *GH-CSH* locus and LVM change. The results also indicated LD between the *ACE* I/D polymorphism and *GH-CSH BglII-B*, D′ =0.3, which accounted for the significant association found between *GH-CSH BglII-B* and pretraining ACE levels.

In the last 15 years, the *ACE I/D* polymorphism has been widely studied in relation to ACE activity and numerous diseases and traits. Conflicting results have been reported, especially regarding a possible association of the *ACE* gene polymorphism with left ventricular (LV) hypertrophy. While no such association was found by Lindpaintner et al. [387], Kupari et al. [388], Wu et al. [389] and Lopez-Contreras et al. [390], others have found an association between *ACE* genotype and LV hypertrophy [333, 334]. Variations in sample selection and size, ethnicity, and analysis method represent possible explanations. In the present study, genotype-phenotype analysis using one-way ANOVA showed that both *ACE I/D* and *GH-CSH Bgl IIB* were associated with ACE activity, while only *ACE I/D* was associated with LVM change (Table 3.33). Haplotype analysis showed that four haplotypes (*D/C, D/N, I/C* and *I/N*) were strongly associated

with ACE activity, with the overall p value <0.0001, but not with LVM change (overall p value = 0.1066), only haplotypes *D/N* and *I/C* showed marginal significance (p=0.0655 and 0.0543 respectively) (Table 3.34). Linear regression analysis also suggested a significant association between LVM change and *ACE I/D* genotype (2.6%, p=0.008). Further haplotype analysis within each homozygote group of *ACE I/D* and *GH-CSH Bgl IIB* showed that the association only existed between *ACE I/D* genotypes and ACE activity (Table 3.35).

It is clear that the positive association between *GH-CSH Bgl IIB* and serum ACE activity was due to its linkage disequilibrium with *ACE I/D*. When each single marker was analysed in relation to serum ACE activity, both *ACE I/D* and *GH-CSH Bgl IIB* were strongly associated (p<0.0001). Haplotype analysis also showed strong association between haplotypes and phenotype, but linear regression analysis indicated that the variation of ACE activity was accounted for mainly by *ACE I/D* (30.34%), *GH-CSH Bgl IIB* only contributed 1.5% of the variation of ACE activity. When the analysis was performed within either *ACE* homozygote groups, the effect of *GH-CSH Bgl IIB* genotype disappeared (Table 3.35). As for the LVM change, single marker analysis showed significant differences for *ACE I/D* genotypes. The differences could not be detected when haplotypes were examined, although linear regression analysis showed a significant effect (2.6%, p=0.008) accounted for by the *ACE I/D*.

It should also be noted the effects of gene-gene and/or gene-environment interactions on either ACE activity or LVM. Diet et al. (2001) reported that no association with left ventricular mass was found for the polymorphisms of *ACE I/D*, angiotensinogen gene M235T and angiotensin II type 1 gene A1166C when studied separately. However, combined analysis showed a synergistic effect of *ACE DD* genotype and angiotensinogen gene *TT* genotype on left ventricular mass in endurance athletes [336].

Kuznetsova et al. (2004) suggested that the relationship between LVMI (LVM index) and the *ACE I/D* differed across populations, and possibly as a consequence of intermediate regulatory mechanisms responsive to varying level of salt intake [391, 392]. Other factors in relation to ACE and LVM include smoking [393] and exercise [335-337, 345].

In conclusion, there is linkage disequilibrium between markers in *ACE* and the *GH-CSH* gene cluster. We have shown here that the association of LVM with *ACE I/D* does not represent a causal effect of *GH-CSH*. Every independent phenotype with which *ACE I/D* has been associated, however, merits investigation as to whether this might represent causal effects due to LD with the *GH-CSH* cluster, otherwise the chain of causality could be misinterpreted. This is highly relevant to potential pharmacogenetic inferences or interventions involving the renin–angiotensin pathway or growth hormone pathway.

Chapter 5 General discussion and perspectives

My study has focused on an investigation of genetic variants in relation to smoking behaviour, growth and metabolic traits in two cohorts, using a candidate gene approach and previously described polymorphisms. A strength of my study, which was initiated in 2001, is that a relatively large number of well characterised subjects were analysed. However human genetic research has developed dramatically in recent years and in the context of new developments it is clear that the analysis I have undertaken is limited in scope. Whilst candidate gene studies still have a role to play, a better understanding of structure and variation in the human genome, better definitions of statistically significant associations, understanding common errors and rapidly developing genetic technologies have enabled whole genome screening of susceptibility variants, genes and regions to become a reality. In this Chapter I discuss these new developments and how the questions addressed in my study may be taken forward.

5.1 How to interpret significant associations?

In genetic association studies, the finding of an apparently significant association between a genetic variant and a disease could be due to the following reasons: 1) random chance; 2) population stratification; 3) a surrogate for the causal locus due to LD; and 4) a true causal association of putative gene/variant. As in any statistical test, false-positive results can arise due to chance, e.g. a conventional significance value of 0.05 would be expected to occur by chance at a frequency of 1 in 20. This means that if a particular variant is compared to a phenotype in 20 separate studies, on average one of those studies will find a false positive association with any variant that is not truly associated with the phenotype. Similarly, on average one false positive at this threshold of significance will arise in any study that compares 20 independent genetic variants

with any phenotype. The conventional way to cater for multiple testing is to apply a Bonferroni correction by dividing the conventional level of significance by the number of tests to give the true level of significance required to call an association. In this example only markers that achieved a p value of 0.05/20 = 0.0025 should be considered as significant.

How stringent should a threshold in a genome wide association study be? Risch and Merikangas suggested that the same Bonferroni correction could be applied if one million SNPs were typed across the genome, ie. a p value of 5×10^{-8} is required to demonstrate a significant association [72], however this remains an area of debate and statistical methodology in this area is still being developed.

Population stratification, also known as population substructure, can result from recent population admixtures or differences in ethnicity between cases and controls [394]. Different approaches have been developed to solve this issue. Family-based association strategies have become increasingly common in genetic studies of complex diseases [395, 396]. Another recently developed method is to genotype a set of polymorphic markers unlinked to any candidate loci for the disease of interest in cases and controls [397]. Absence of association to these unlinked markers provides some assurance that an association with a candidate locus in those cases and controls is not a spurious result [42].

Genotyped markers associated with the disease could be a proxy because of its LD with causal marker. For example, congenital adrenal hyperplasia (CAH) was reported to be associated with HLA-B47 [398], and late-onset adrenal hyperplasia with HLA-B14 [399]. The fact that HLA-B47 is associated with CAH is due to deletion of *CYP21A2* on HLA-B47 DR7 hyplotype; while HLA-B14 associated with late-onset adrenal

hyperplasia is due to an exon 7 missense mutation (V281L) in *CYP21A2* on HLA-B14DR1 hyplotype. In complex disease genetics, replication is especially critical in analysing the true association of a putative gene. However, the role of a putative disease gene variant can only be confirmed after a biologically functional effect has been demonstrated [42, 400]. The use of exonic SNPs and particularly those that change an amino acid are expected to enrich for associations in which the variant is directly causal. However most tagged SNPs are not within coding sequences and thus any associations that are found using them are probably more likely to be associated with an unknown true variant that is in LD.

5.2 Common errors and guidelines for association studies

Many positive findings derived from association studies of complex diseases have failed to be confirmed in subsequent analyses of different patient groups. Cardon and Bell have summarised the common errors encountered in association studies of complex diseases, including small sample sizes, subgroup analysis and multiple testing, random errors, poorly matched control groups, failure to detect LD with adjacent loci, over interpreting results and positive publication bias, and unwarranted 'candidate gene' declaration after identifying associations in arbitrary genetic regions [54]. They also proposed guidelines for conducting association studies, including: 1) replication of any allelic association should be common practice; 2) optimal control population selection should limit the impact of population substructures; 3) in general, prospective studies provide better controls than case–control studies; 4) strategies that use family-based controls and identify linkage and association can be valuable, but are inefficient in their use of available information; 5) the power of a study is influenced by a host of factors, such as the effect size, local patterns of LD, allele frequency differences between marker and trait loci, and genetic heterogeneity. Sample sizes should be chosen to

assume suboptimal conditions, such as weak effects, rare alleles and incomplete LD; 6) multiple testing, either with multiple markers or independent phenotypes, will produce false positives under nominal significance thresholds. All tests considered or conducted should be reported in association results, even if nonsignificant. The multiple-testing effect will be exacerbated by publication bias; 7) selection of controls sets the range of the allele frequencies in a population, particularly those related to ethnic diversity. Consideration of different sampling panels might help refine regions of LD and provide insight into LD patterns and history [54]. These recommendations have been largely accepted by the genetics community and whole genome association studies that are now underway often include several thousand cases/controls and are structured as a nested or cascade study, in which an initial group is used to identify candidate loci which are then tested more rigourously in secondary and sometimes tertiary groups [401-405].

5.3 High throughput genotyping and whole genome scan technologies

Resequencing large amounts of DNA and ultimately whole genomes from cases and controls is the only comprehensive approach to investigate the associations between all kinds of genetic variations (common and rare) and complex diseases. However, the cost for doing this is currently prohibitive. After spending $3 billion to sequence the entire human genome the first time around, researchers are attempting to drive the cost down as low as it can go. A key milestone in the evolution of DNA sequencing is the $1,000 genome. To achieve this goal, the cost per base for an accurate $1,000 genome must approach approximately 40 million raw bases per $1. To motivate its realisation, J. Craig Venter, president of Celera Genomics, announced a $10 million prize for technology leading to the $1,000 genome. Researchers aim to sequence the genome for $100,000 within five years and for $1,000 in ten years time. The current cost is in the order of $1,000,000 [406].

Realising the the $1,000 genome will enable comprehensive association studies to be performed, although the data that will be generated will present a formidable challenge in terms of analysis and validation. In the meantime, high throughput genotyping and whole genome scan technologies have been developed and widely applied to identify causal genes and associated variants. Among various types of high throughput technologies, microarray formats have been developed to analyse gene expression, genotype SNPs and identify copy number changes. Array CGH and SNP arrays are discussed in more detail below.

5.3.1. Array CGH

Array CGH (Array comparative genomic hybridisation) is a type of microarray that has been developed to detect chromosomal copy number changes on a genome-wide and/or high-resolution scale [407]. It has been widely used to identify chromosomal gains and losses in tumours, as well as new microdeletion syndromes associated with mental retardation [407-414] and associations between copy number differences and other complex disorders such as autism and schizophrenia [408-415] [416].

Array CGH builds upon conventional Comparative Genomic Hybridisation (CGH), in which differentially labelled genomic DNA from a 'test' and a 'reference' cell population are co-hybridised to normal metaphase chromosomes, using blocking DNA (Cot-1 DNA) to suppress signals from repetitive sequences. The fluorescence ratio of the test and reference hybridisation signals is determined at different positions along the genome and provides information on the relative copy number of sequences in the test genome compared with a normal diploid genome [417]. The use of metaphase chromosomes limits the resolution of conventional CGH to 10-20Mb [418] but the development of microarray based CGH has greatly increased the resolution. The first

array CGH platforms used large-insert clones as probes, such as BACs (Bacterial Artificial Chromosome), YACs (Yeast Artificial Chromosome) or PACs (P1 derived Artificial Chromosome). These clones are arrayed onto glass slides and used as the hybridisation target rather than metaphase chromosomes. The resolution is determined by the spatial distances between the clones and the size of the cloned DNA fragments [419]. Genome-wide BAC arrays vary in size from 2,400 to over 32,000 unique array elements, with the resolution from 1.4Mb to ~100kb [420-423]. However, because of the large size of the inserted clones (~ 150kb), the BAC arrays can not detect genomic aberrations smaller than ~100kb.

The most promising alternative is the development of oligonucleotide arrays (oligo arrays), which are characterised by single-stranded 25 to 85mer oligonucleotide elements [407]. This platform has the advantage of being rapidly synthesised with very high uniformity and density, and being customised to any region of interest with high resolution. However, array CGH also has a number of limitations independent of the platform used, particularly its inability to detect aberrations that do not result in copy number changes such as balanced translocations or copy neutral loss of heterozygosity. Consequently array CGH will not completely replace conventional cytogenetic analysis as a routine diagnostic tool but it will provide an important means to compare copy number changes with clinical phenotypes.

5.3.2 SNP arrays

SNP arrays have been designed to target and type increasingly large numbers of SNPs across the genome in patient samples. To construct a SNP array, the probes related to a single SNP are organised into groups called probe sets. Each probe in a probe set is a single-stranded 25 to 85mer oligonucleotide element that contains the nucleotide

specific to the SNP, plus in some formats control probes that contain mismatches to provide a measure of hybridisation background. There estimated to be are over 10 million SNPs in the human genome, which means on average there is one SNP per 300bp, although current arrays are only capable of testing a subset of this variation. Although initially designed to produce simple SNP calls (a/a, a/b, b/b), refinements in software and increasingly dense chips have enabled SNP arrays to be used to identify copy number abnormalities as well. Currently there are two principal commercial platforms which can provide dense genomewide SNP coverage: Affymetrix and Illumina.

The Affymetrix system uses test and control oligonucleotides on conventional arrays. The most recently released format, termed SNP array 5.0, is a single microarray featuring approximately 500,000 SNPs as well as 420,000 additional non-polymorphic probes that can measure other genetic differences, such as copy number variation (www.affymetrix.com). Arrays that incorporate one million SNPs will be released soon. Test genomic DNA is digested with a restriction enzyme and ligated to adaptors. A generic primer that recognises the adaptor sequence is used to amplify adaptor-ligated DNA fragments and preferentially amplifies fragments in the 200-1,100 bp size range, thus reducing the genome complexity by about 10 fold. The amplified DNA is then fragmented, labelled, and hybridised to array. After stringent washing signals are read in a scanner and translated into SNP calls and copy number.

The Illumina platform can generate a similar number of genotypes and has focused in particular on producing arrays with tagged SNPs from the HapMap project (www.illumina.com). Test DNA is amplified isothermally, fragmented and hybridised to bead-coupled oligonucleotides. There are over 500,000 bead types couple to oligonucleotides that are specific to each allele per SNP. After hybridisation, allele

specificity is determined by enzymatic single base extension and following flourescent staining. Intensities for each bead are measured using a proprietry reader and converted into genotypes [424].

Both these products are highly suitable for whole-genome association studies with SNPs and new associations from whole genome studies are beginning to emerge [59-63, 403-405]. The degree to which the arrays can cover copy number variation is less clear since their coverage has been limited by the availability and density of informative SNPs, which have typically excluded repeat regions, including significant portions of segmental duplications [18]. As illustrated in Figure 5.1, segmental duplication regions are poorly covered by both the 500K Affymetrix and 300k Illumina arrays. Although the new Affymetrix 5.0 array begins to address some of this deficiency it is clear that new systems need to focus on better representation of CNVs in order to provide comprehensive coverage.

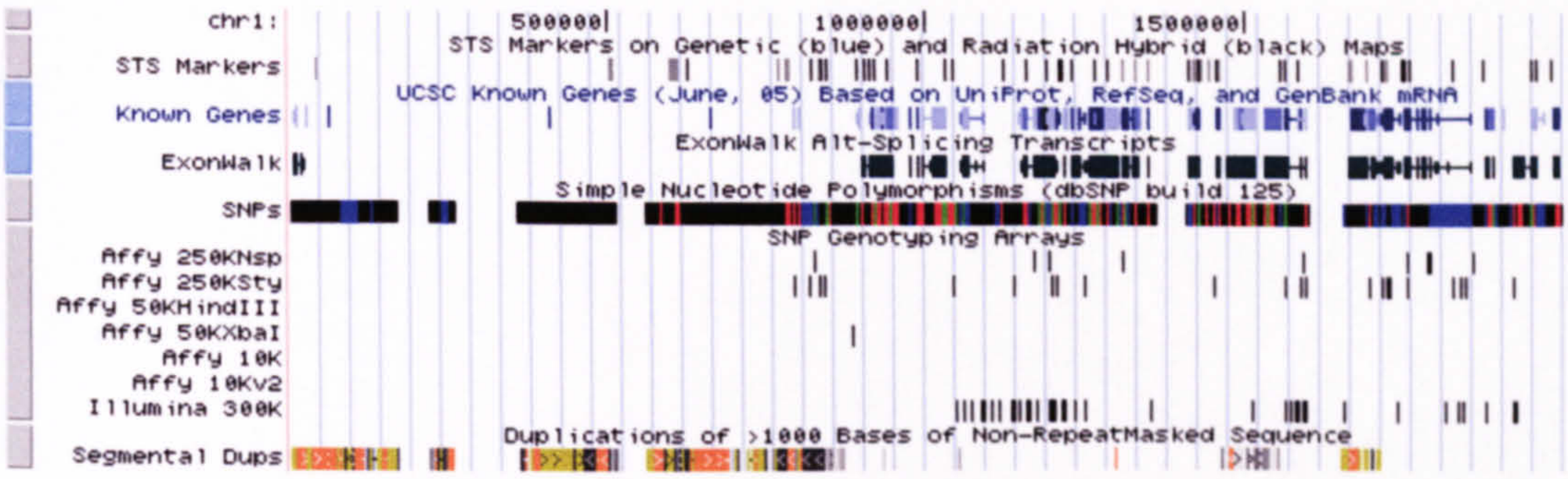


Figure 5.1 Demonstration of SNP array coverage of segmental duplication region.

It has been estimated that segmental duplications (duplicons) comprise at least 5% of the genome and many are yet to be annotated. Although recent intensive mapping by BAC array CGH has most likely identified large duplicons and other repeated

sequences, much less is known about the copy number of sequences below the resolution of this technique. Quantitative analysis of SNPs by Dynamic Allele-Specific Hybridization (DASH) has revealed an appreciable amount of copy number variation at this level, e.g. more than 10% SNPs show copy number variation. [371] [425, 426]. It is clear therefore that there is still a considerable amount of copy number variation below a size of approximately 100kb that is poorly characterised and that the genetic variation that is known is not fully captured by any single technology that is currently available. Nevertheless, the new platforms are able to provide genomewide analysis to a degree that was not possible when I started my project and it seems appropriate therefore to consider how the analysis of genetic factors relating to smoking and growth promoting genes might be studied using current technology.

5.3.4 Investigating the relationship between genetic variations and complex diseases using current technologies.

The main limitations in undertaking a comprehensive genomewide analysis of any complex disease using current technologies are (i) the availability of DNA from a sufficient number of well characterised patients and controls (ii) sufficient money to perform the genotyping and (iii) appropriate bioinformatic support to analyse the data.

Since both smoking and growth differences/metabolic syndrome are very common it would be possible in principal to recruit very large numbers of cases for an association study given adequate resources. Indeed the national biobanks that are currently underway will be ideally suited for this sort of analysis, provided that relevant data is actually collected to accompany the sample. However genotyping large numbers of SNPs in large numbers of samples is still expensive. Assuming the infrastructure for performing Affymetrix or Illumina analysis is available, the current cost to determine

500,000 SNPs is in the order of £500 per sample, and thus the analysis of 2000 cases and controls would cost a million pounds. Assuming that sufficient resources were available then I would plan a 3-stage project, including whole genome scanning, focusing the investigation on a large cohort followed, if possible, by functional or animal experiments to confirm the significance of any association.

Stage 1- whole genome scanning

(1). A sufficient number of cases and controls would be used for susceptibility gene screening according to the disease prevalence and availability of study cohorts. From recently published whole genome association studies, the sample size ranges from about 150 (96 cases and 50 controls) [59] to about 5000 (1924 cases and 2938 controls) [427], and most genome-wide association studies have an average sample size of 2000-3000 (~50% cases and 50% controls) for the first stage investigation [402, 403, 405]. Sample size and power calculation can be estimated using 'Genetic Power Calculator' (http://statgen.iop.kcl.ac.uk/gpc/cc2.html), as shown in Table 5.2 and 5.3. According to Schaid (1999) [428], there are several genotype relative risk (GRR) models, as shown in Table 5.1. Assuming 500,000 SNPs are genotyped, the significance level required would be 0.05/500000=0.0000001 after Bonferroni correction. If an additive model is used with GRR for Aa=1.5 (AA=2.0), disease allele frequency 0.3 and population prevalence 0.05, then 1352 cases and 1352 controls are needed for 90% power at 5% type I error rate, as shown in Table 5.2.

Table 5.1 Genotype relative risk (GRR) models and GRR for Aa and AA

Genotype Relative Risk	Multiplicative	Additive	Recessive	Dominant
Aa	Y	Y	1	Y
AA	Y*Y	2Y-1	Y	Y
Example if Y=1.5				
Aa	1.5	1.5	1	1.5
AA	2.25	2	1.5	1.5

Table 5.2 Sample size and power calculation using Genetic Power Calculator

Case - control for discrete traits

```
High risk allele frequency (A)    : [0.3]      (0 - 1)
Prevalence                        : [0.05]     (0.0001 - 0.9999)
Genotype relative risk Aa         : [1.5]      ( >1 )
Genotype relative risk AA         : [2.0]      ( >1 )

D-prime                           : [1]        (0 - 1)
Marker allele frequency (B)       : [0.3]      (0 - 1)

Number of cases                   : [1000]     (0 - 10000000)
Control : case ratio              : [1]        ( >0 )
                                    ( 1 = equal number of cases
and controls)
User-defined type I error rate    : [0.0000001] (0.00000001 - 0.5)
User-defined power: determine N   : [0.90]     (0 - 1)
(1 - type II error rate)
```

Process | Reset

Case-control parameters

Number of cases	1000
Number of controls	1000
High risk allele frequency (A)	0.3
Prevalence	0.05
Genotypic relative risk Aa	1.5
Genotypic relative risk AA	2

Genotypic risk for aa (baseline)	0.03846

Marker locus B

High risk allele frequency (B)	0.3
Linkage disequilibrium (D')	1
Penetrance at marker genotype bb	0.03846
Penetrance at marker genotype Bb	0.05769
Penetrance at marker genotype BB	0.07692
Genotypic odds ratio Bb	1.531
Genotypic odds ratio BB	2.083

Expected allele frequencies

	Case	Control
B	0.3808	0.2957
b	0.6192	0.7043

Expected genotype frequencies

	Case	Control
BB	0.1385	0.08745
Bb	0.4846	0.4166
bb	0.3769	0.496
H-W test NCP	0.7655	7.528e-06
Power (alpha=0.05)	0.1412	0.05

Case-control statistics
Sample NCP = 32.29

Alpha	**Power**	**N cases for 90% power**
0.1	1	265.2
0.05	0.9999	325.4
0.01	0.9991	460.8
0.001	0.9916	647.3
1e-07	0.6391	**1352**

(2) Another approach to improve power and reduce sample sizes can be obtained by using CHROMSCAN analysis. CHROMSCAN is a statistical based program originally developed by Prof. A. R. Collins at the Genetics Epidemiology and Bioinformatics Group, Human Genetic Division at the University of Southampton for association

mapping of disease genes by linkage disequilibrium (LD). An entire chromosome or a large segment of a chromosome is split into regions of appropriate length e.g. each region is delimitated by a minimum size of 30 SNPs and a minimum length of 10 LD units. There are about ~65000 LD units in the genome [429]; 6000 regions of 10 LD units are generated and therefore 6000 tests for Bonferroni correction. The significance level required is thus = 0.05/6000=0.00000833. Using the same other parameters (Table 5.2), 1020 cases are needed for 90% power at alpha = 0.00000883, as shown in Table 5.3.

Table 5.3 Sample size and power calculation using Genetic Power Calculator and CHROMSCAN software

Case - control for discrete traits

```
High risk allele frequency (A)    : 0.3      (0 - 1)
Prevalence                        : 0.05     (0.0001 - 0.9999)
Genotype relative risk Aa         : 1.5      ( >1 )
Genotype relative risk AA         : 2.0      ( >1 )
D-prime                           : 1        (0 - 1)
Marker allele frequency (B)       : 0.3      (0 - 1)
Number of cases                   : 1000     (0 - 10000000)
Control : case ratio              : 1        ( >0 )
                                    ( 1 = equal number of cases and controls)
User-defined type I error rate    : 0.0000083  (0.00000001 - 0.5)
User-defined power: determine N   : 0.90     (0 - 1)
(1 - type II error rate)
```

Process | Reset

Case-control for discrete traits

Case-control parameters

Number of cases	1000
Number of controls	1000

High risk allele frequency (A)	0.3
Prevalence	0.05
Genotypic relative risk Aa	1.5
Genotypic relative risk AA	2
Genotypic risk for aa (baseline)	0.03846

Marker locus B

High risk allele frequency (B)	0.3
Linkage disequilibrium (D')	1
Penetrance at marker genotype bb	0.03846
Penetrance at marker genotype Bb	0.05769
Penetrance at marker genotype BB	0.07692
Genotypic odds ratio Bb	1.531
Genotypic odds ratio BB	2.083

Expected allele frequencies

	Case	Control
B	0.3808	0.2957
b	0.6192	0.7043

Expected genotype frequencies

	Case	Control
BB	0.1385	0.08745
Bb	0.4846	0.4166
bb	0.3769	0.496
H-W test NCP	0.7655	7.528e-06
Power (alpha=0.05)	0.1412	0.05

Case-control statistics
Sample NCP = 32.29

Alpha	**Power**	**N cases for 90% power**
0.1	1	265.2
0.05	0.9999	325.4
0.01	0.9991	460.8
0.001	0.9916	647.3
8.33e-06	0.8899	**1020**

(3) Criteria for cases should be carefully designed and implemented, including clinical symptoms and laboratory results etc.

(4) Controls should be carefully matched to the cases in terms of age, sex, ethnicity disease stage, and other relevent risk factors specific to the disease.

(5) High-density SNP arrays would be used for the genotype analysis, such as the Affymetrix SNP Array 5.0, or the 550K Illumina beadchip.

(6) Appropriate statistical software for data analysis should be used. Although it was suggested that a Bonferroni correction could be applied if one million SNPs were typed across the genome, with a P value of 5×10^{-8} required to demonstrate a significant association [72], the aim of the primary screen is to enrich for relevant variants rather than unambiguously identify them. A less stringent p value is therefore appropriate, for example, the p value of less than 0.0001 was chosen in the first stage analysis by a recently published genome-wide genotyping study in amyotrophic lateral sclerosis [430]; while $p<0.00001$ was chosen by another study [403].

(7) Prioritise candidate genes based on the primary screen, plus possibly including good candidate genes based on previous functional and genetic studies, animal studies and meta-analysis.The number of prioritised genes is likely to number a few hundred.

(8) Define a minimum set of SNPs to saturate the candidate genes. Even with the capacity of designing as many SNPs as possible, this does not necessarily mean 'the more the better', because: 1) a lot of SNPs are just redundant because of LD, including them will not add any more information; 2) adding more redundant markers can reduce the detection power because of multiple testing.

(9) Design customised array according to the first stage results for second stage investigation.

Stage 2 – running customised array in big clinical or other cohorts

(1) The first repeat cohort could be derived from the same background of case-control trial. Alternatively it might include several small to medium sized cohorts, but criteria for selecting subjects should be standardised. As the number of SNPs being tested is reduced at this stage, larger sample numbers are more affordable and will add power.

(2) Analyse and interpret stage 2 results. Always remember that statistically significant association does not mean that there is definitely a link between the genotyped marker and a disease (trait), it could be due to random chance, or a surrogate of the true marker.

(3) The sample size and the p value required to show significant association can be calculated using the Genetic Power Calculator as described above

Stage 3 – result confirmation using functional experiments

The discovery of causal genes or markers should be confirmed using functional experiments, such as customised gene expression array and animal experiments. Clearly this approach may be possible when studying growth and metabolic differences but is likely to be much more difficult when considering traits such as smoking. However even if no direct test in animals can be established, it may be possible ultimately to link any variant to a relevant biological pathway and thereby generate hypotheses that can be tested in humans or animals.

Appendix Methods for making up stock reagents

Stock reagents for DNA extraction

- **10X Erythrocyte Lysis Buffer (ELB)**

10.1g potassium hydrogen carbonate ($KHCO_3$)

82g ammonium chloride

2ml 0.5M EDTA pH 8.0

Made up to 1 litre (using highly purified water), and stored at room temperature.

- **1X ELB**

50ml 10X ELB

450ml highly purified water

Mixed and autoclaved, stored at 4°C overnight before use.

- **1X Nucleic Lysis Buffer (NLB)**

5ml 1M Tris pH 8.0

40ml 5M Sodium chloride (NaCl)

2ml 0.5M EDTA pH 8.0

Add highly purified water up to 500ml, stored at room temperature.

- **Saturated NaCl Solution (approximately 6M)**

Dissolve 35g of sodium chloride (NaCl) in a total volume of 100ml deionized water (DI water). If the solution has no precipitate, add 2g NaCl and mix. Repeat until no more NaCl will go into solution.

- **10% sodium dodecyl sulphate (SDS)**

Weigh 10g SDS using weighing boat.Take a 100ml beaker, add 80-90ml highly purified water, and put the beaker on the magnetic mixer. Add SDS gradually, mix and dissolve. Add water up to 100ml.

- **Protease**

800mg into 20ml DI water. Mix completely and aliquot (1ml) into 1.5ml microtubes. Stored at -20°C.

- **Tris-EDTA Buffer (TE butter)**

10ml 1.0M Tris pH 7.5

2ml 0.5M EDTA pH 8.0

Made up to 1 litre with DI water. Sterilized by autoclaving.

- **500ml 70% Ethanol**

350ml Absolute Ethanol

150ml DI water

Buffers and solutions for PCR

- **1 litre 10 X TBE buffer**

108g Tris base

55g Orthoboric acid

9.3g Na EDTA

Add DI water up to 1 litre.

- **500ml Sticky silane stock solution**

  2.5ml Glacial acetic acid

  2.5ml λ-methacryloxypropyltrimethoxylilane

  Add ethanol up to 500ml.

- **50ml MADGE loading buffer (blue dye)**

  49ml Deionised formamide (98%)

  1ml Na EDTA (0.5M)

  125mg Xylene cyanl FF

- **25% APS (Ammonium persulphate)**

  2.5g Ammonium persulphate

  Add DI water up to 10ml, and mix.

Reference List

(1) Charles R.Scriver. An Ongoing Debate Over Phenylalanine Hydroxylase Deficiency in Phenylketonuria. The Journal of Clinical Investigation 1998 June 15;101(12), 2613-2614.

(2) Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 2006 Sep;**16**(9):1182-90.

(3) The international HapMap consortium. A haplotype map of the human genome. Nature 2005 Oct 27;**437**(7063):1299-320.

(4) Stein LD. Human genome: end of the beginning. Nature 2004 Oct 21;**431**(7011):915-6.

(5) Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, bu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M. The sequence of the human genome. Science 2001 Feb 16;**291**(5507):1304-51.

(6) Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature 2006 Nov 23;**444**(7118):444-54.

(7) Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C. Copy number variation: new insights in genome diversity. Genome Res 2006 Aug;**16**(8):949-61.

(8) Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006 Feb;**7**(2):85-97.

(9) Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. Nature 1985 Jul 4;**316**(6023):76-9.

(10) Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 2003;**4**(2):R13.

(11) Tom Strachan APR. Human Molecular Genetics. 3 ed. London and New York: Garland Science; 2004.

(12) Richards RI, Shen Y, Holman K, Kozman H, Hyland VJ, Mulley JC, Sutherland GR. Fragile X syndrome: diagnosis using highly polymorphic microsatellite markers. Am J Hum Genet 1991 Jun;**48**(6):1051-7.

(13) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 1993 Mar 26;**72**(6):971-83.

(14) Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL. Alu insertion polymorphisms for the study of human genomic diversity. Genetics 2001 Sep;**159**(1):279-90.

(15) Deininger PL, Batzer MA. Alu repeats and human disease. Mol Genet Metab 1999 Jul;**67**(3):183-93.

(16) Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet 2002 May;**3**(5):370-9.

(17) Ricki Lewis. Human Genetics: Concepts and Applications. 6 ed. McGraw-Hill Science/Engineering/Math; 2004.

(18) Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. Annu Rev Genomics Hum Genet 2006;7:407-42.

(19) Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, Weber JL, Ledbetter DH, Zuffardi O. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 2001 Apr;**68**(4):874-83.

(20) Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. Science 2004 Jul 23;**305**(5683):525-8.

(21) Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet 2004 Sep;**36**(9):949-51.

(22) Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE. Fine-scale structural variation of the human genome. Nat Genet 2005 Jul;**37**(7):727-32.

(23) Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 2005 Jul;**77**(1):78-88.

(24) Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 2005 Sep 1;**437**(7055):94-100.

(25) She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, Bailey JA, Sahinalp C, Rocchi M, Haussler D, Wilson RK, Miller W, Schwartz S, Eichler EE. The structure and evolution of centromeric transition regions within the human genome. Nature 2004 Aug 19;**430**(7002):857-64.

(26) Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 2006 Jan;**38**(1):75-81.

(27) Aldred PM, Hollox EJ, Armour JA. Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet 2005 Jul 15;**14**(14):2045-52.

(28) Miller DW, Hague SM, Clarimon J, Baptista M, Gwinn-Hardy K, Cookson MR, Singleton AB. Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. Neurology 2004 May 25;**62**(10):1835-8.

(29) Hollox EJ, Armour JA, Barber JC. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet 2003 Sep;**73**(3):591-600.

(30) McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM. Common deletion polymorphisms in the human genome. Nat Genet 2006 Jan;**38**(1):86-92.

(31) Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de GA, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007 Feb 9;**315**(5813):848-53.

(32) Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005 Mar 4;**307**(5714):1434-40.

(33) Sharp AJ, Selzer RR, Veltman JA, Gimelli S, Gimelli G, Striano P, Coppola A, Regan R, Price SM, Knoers NV, Eis PS, Brunner HG, Hennekam RC, Knight SJ, de Vries BB, Zuffardi O, Eichler EE. Characterization of a recurrent 15q24 microdeletion syndrome. Hum Mol Genet 2007 Mar 1;**16**(5):567-72.

(34) Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 2006 Sep;**38**(9):1038-42.

(35) Koolen DA, Vissers LE, Pfundt R, de LN, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet 2006 Sep;**38**(9):999-1001.

(36) Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepischi-Santos AC, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de SR, Carter NP. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 2006 Sep;**38**(9):1032-7.

(37) Jacobs PA, Melville M, Ratcliffe S, Keay AJ, Syme J. A cytogenetic survey of 11,680 newborn infants. Ann Hum Genet 1974 May;**37**(4):359-76.

(38) Barber JC. Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. J Med Genet 2005 Aug;**42**(8):609-29.

(39) Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995 Nov;**11**(3):241-7.

(40) Dawn TM, Barrett JH. Genetic linkage studies. Lancet 2005 Sep 17;**366**(9490):1036-44.

(41) Mueller RF, Bishop DT. Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. J Med Genet 1993 Sep;**30**(9):798-9.

(42) Silverman EK, Palmer LJ. Case-control association studies for the genetics of complex respiratory diseases. Am J Respir Cell Mol Biol 2000 Jun;**22**(6):645-8.

(43) Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994 Sep 30;**265**(5181):2037-48.

(44) Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001 Sep;**17**(9):502-10.

(45) Newton-Cheh C, Hirschhorn JN. Genetic association studies of complex traits: design and analysis issues. Mutat Res 2005 Jun 3;**573**(1-2):54-69.

(46) Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Hum Genet 1995 Jul;**59**(Pt 3):323-36.

(47) Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 1998 Feb;**62**(2):450-8.

(48) Curtis D. Use of siblings as controls in case-control association studies. Ann Hum Genet 1997 Jul;**61**(Pt 4):319-33.

(49) Martin ER, Kaplan NL, Weir BS. Tests for linkage and association in nuclear families. Am J Hum Genet 1997 Aug;**61**(2):439-48.

(50) Allison DB. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 1997 Mar;**60**(3):676-90.

(51) Rabinowitz D. A transmission disequilibrium test for quantitative trait loci. Hum Hered 1997 Nov;**47**(6):342-50.

(52) Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000 Jan;**66**(1):279-92.

(53) Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 2000 Jul;**67**(1):146-54.

(54) Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001 Feb;**2**(2):91-9.

(55) Hattersley AT, McCarthy MI. What makes a good genetic association study? Lancet 2005 Oct 8;**366**(9493):1315-23.

(56) Rosand J, Altshuler D. Human genome sequence variation and the search for genes influencing stroke. Stroke 2003 Oct;**34**(10):2512-6.

(57) James J.Dowling and Andrew A.Hack. Introduction to molecular and clinical genetics. Best Practice & Research Clinical Anaesthesiology 2001;**15**(2):155-68.

(58) Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science 2002 Dec 20;**298**(5602):2345-9.

(59) Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science 2005 Apr 15;**308**(5720):385-9.

(60) Edwards AO, Ritter R, III, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. Science 2005 Apr 15;**308**(5720):421-4.

(61) Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA. Complement factor H variant increases the risk of age-related macular degeneration. Science 2005 Apr 15;**308**(5720):419-21.

(62) Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, bers-Akkers MT, Godino-Ivan MJ, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007 Apr 1.

(63) Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF, Jr., Hoover R, Hunter DJ, Chanock SJ, Thomas G. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007 Apr 1.

(64) Goldstein DB, Weale ME. Population genomics: linkage disequilibrium holds the key. Curr Biol 2001 Jul 24;**11**(14):R576-R579.

(65) LEWONTIN RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 1964 Jan;**49**(1):49-67.

(66) Jorde LB. Linkage disequilibrium and the search for complex disease genes. Genome Res 2000 Oct;**10**(10):1435-44.

(67) Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. Nature 2001 May 10;**411**(6834):199-204.

(68) W.G.Hill and Alan Robertson. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics 1968;**38**(6):226-31.

(69) Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999 Jun;**22**(2):139-44.

(70) Peterson AC, Di RA, Lehesjoki AE, de la CA, Slatkin M, Freimer NB. The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet 1995 May;4(5):887-94.

(71) Laan M, Paabo S. Demographic history and linkage disequilibrium in human populations. Nat Genet 1997 Dec;17(4):435-8.

(72) Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996 Sep 13;273(5281):1516-7.

(73) Nordborg M, Tavare S. Linkage disequilibrium: what history has to tell us. Trends Genet 2002 Feb;18(2):83-90.

(74) Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 2002 Jan;18(1):19-24.

(75) Zhang K, Calabrese P, Nordborg M, Sun F. Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 2002 Dec;71(6):1386-94.

(76) The International HapMap Project. Nature 2003 Dec 18;426(6968):789-96.

(77) Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A. An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS Genet 2006 Mar;2(3):e27.

(78) Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF, Kinnunen L, Mohlke KL, Valle TT, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet Epidemiol 2006 Feb;30(2):180-90.

(79) De Bakker PI, Graham RR, Altshuler D, Henderson BE, Haiman CA. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. Pac Symp Biocomput 2006;478-86.

(80) Mutsuddi M, Morris DW, Waggoner SG, Daly MJ, Scolnick EM, Sklar P. Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. Am J Hum Genet 2006 Nov;79(5):903-9.

(81) Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res 2005 Nov;15(11):1566-75.

(82) Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 2006 Jan;38(1):82-5.

(83) Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet 2005 Jan 1;14(1):59-69.

(84) Eichler EE. Widening the spectrum of human genetic variation. Nat Genet 2006 Jan;38(1):9-11.

(85) Action on Smoking and Health. Smoking Statistics. 2007.

(86) World Health Organisation. Neuroscience of Psychoactive Substance Use and Dependence. WHO; 2004.

(87) Edwards R. The problem of tobacco smoking. BMJ 2004 Jan 24;328(7433):217-9.

(88) World Health Organisation. Tobacco Free Initiative (TFI). 2007.

(89) The World Bank Group. Economics of Tobacco Control. 2007.

(90) Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ 2004 Jun 26;328(7455):1519.

(91) Tyndale RF, Sellers EM. Variable CYP2A6-mediated nicotine metabolism alters smoking behavior and risk. Drug Metab Dispos 2001 Apr;29(4 Pt 2):548-52.

(92) Sabol SZ, Nelson ML, Fisher C, Gunzerath L, Brody CL, Hu S, Sirota LA, Marcus SE, Greenberg BD, Lucas FR, Benjamin J, Murphy DL, Hamer DH. A genetic association for cigarette smoking behavior. Health Psychol 1999 Jan;18(1):7-13.

(93) Pomerleau OF. Individual differences in sensitivity to nicotine: implications for genetic research on nicotine dependence. Behav Genet 1995 Mar;25(2):161-77.

(94) Rossing MA. Genetic influences on smoking: candidate genes. Environ Health Perspect 1998 May;106(5):231-8.

(95) Ho MK, Tyndale RF. Overview of the pharmacogenomics of cigarette smoking. Pharmacogenomics J 2007 Jan 16.

(96) Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L, Fugman D, Goate AM, Hinrichs AL, Konvicka K, Martin NG, Montgomery GW, Saccone NL, Saccone SF, Wang JC, Chase GA, Rice JP, Ballinger DG. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 2007 Jan 1;16(1):24-35.

(97) FISHER RA. Cancer and smoking. Nature 1958 Aug 30;182(4635):596.

(98) Kaprio J, Koskenvuo M, Langinvainio H. Finnish twins reared apart. IV: Smoking and drinking habits. A preliminary analysis of the effect of heredity and environment. Acta Genet Med Gemellol (Roma) 1984;33(3):425-33.

(99) Carmelli D, Swan GE, Robinette D, Fabsitz RR. Heritability of substance use in the NAS-NRC Twin Registry. Acta Genet Med Gemellol (Roma) 1990;39(1):91-8.

(100) Swan GE, Carmelli D, Rosenman RH, Fabsitz RR, Christian JC. Smoking and alcohol consumption in adult male twins: genetic heritability and shared environmental influences. J Subst Abuse 1990;2(1):39-50.

(101) Carmelli D, Swan GE, Robinette D, Fabsitz R. Genetic influence on smoking--a study of male twins. N Engl J Med 1992 Sep 17;327(12):829-33.

(102) Kendler KS, Neale MC, MacLean CJ, Heath AC, Eaves LJ, Kessler RC. Smoking and major depression. A causal analysis. Arch Gen Psychiatry 1993 Jan;50(1):36-43.

(103) Boomsma DI, Koopmans JR, Van Doornen LJ, Orlebeke JF. Genetic and social influences on starting to smoke: a study of Dutch adolescent twins and their parents. Addiction 1994 Feb;89(2):219-26.

(104) Edwards KL, Austin MA, Jarvik GP. Evidence for genetic influences on smoking in adult women twins. Clin Genet 1995 May;47(5):236-44.

(105) Heath AC, Madden PA, Slutske WS, Martin NG. Personality and the inheritance of smoking behavior: a genetic perspective. Behav Genet 1995 Mar;25(2):103-17.

(106) Swan GE, Carmelli D, Cardon LR. The consumption of tobacco, alcohol, and coffee in Caucasian male twins: a multivariate genetic analysis. J Subst Abuse 1996;8(1):19-31.

(107) Swan GE, Carmelli D, Cardon LR. Heavy consumption of cigarettes, alcohol and coffee in male twins. J Stud Alcohol 1997 Mar;58(2):182-90.

(108) Hettema JM, Corey LA, Kendler KS. A multivariate genetic analysis of the use of tobacco, alcohol, and caffeine in a population based sample of male and female twins. Drug Alcohol Depend 1999 Nov 1;57(1):69-78.

(109) Kendler KS, Neale MC, Sullivan P, Corey LA, Gardner CO, Prescott CA. A population-based twin study in women of smoking initiation and nicotine dependence. Psychol Med 1999 Mar;29(2):299-308.

(110) Koopmans JR, Slutske WS, Heath AC, Neale MC, Boomsma DI. The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. Behav Genet 1999 Nov;29(6):383-93.

(111) Maes HH, Woodard CE, Murrelle L, Meyer JM, Silberg JL, Hewitt JK, Rutter M, Simonoff E, Pickles A, Carbonneau R, Neale MC, Eaves LJ. Tobacco, alcohol and drug use in eight- to sixteen-year-old twins: the Virginia Twin Study of Adolescent Behavioral Development. J Stud Alcohol 1999 May;60(3):293-305.

(112) Kendler KS, Thornton LM, Pedersen NL. Tobacco consumption in Swedish twins reared apart and reared together. Arch Gen Psychiatry 2000 Sep;57(9):886-92.

(113) Madden PA, Pedersen NL, Kaprio J, Koskenvuo MJ, Martin NG. The epidemiology and genetics of smoking initiation and persistence: crosscultural comparisons of twin study results. Twin Res 2004 Feb;7(1):82-97.

(114) Vink JM, Beem AL, Posthuma D, Neale MC, Willemsen G, Kendler KS, Slagboom PE, Boomsma DI. Linkage analysis of smoking initiation and quantity in Dutch sibling pairs. Pharmacogenomics J 2004;4(5):345-6.

(115) Broms U, Silventoinen K, Madden PA, Heath AC, Kaprio J. Genetic architecture of smoking behavior: a study of Finnish adult twins. Twin Res Hum Genet 2006 Feb;9(1):64-72.

(116) Thomas DC, Haile RW, Duggan D. Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 2005 Sep;77(3):337-45.

(117) Fagerstrom KO. Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. Addict Behav 1978;3(3-4):235-41.

(118) Straub RE, Sullivan PF, Ma Y, Myakishev MV, Harris-Kerr C, Wormley B, Kadambi B, Sadek H, Silverman MA, Webb BT, Neale MC, Bulik CM, Joyce PR, Kendler KS. Susceptibility genes for nicotine dependence: a genome scan and followup in an independent sample suggest that regions on chromosomes 2, 4, 10, 16, 17 and 18 merit further study. Mol Psychiatry 1999 Mar;4(2):129-44.

(119) Bergen AW, Korczak JF, Weissbecker KA, Goldstein AM. A genome-wide search for loci contributing to smoking and alcoholism. Genet Epidemiol 1999;**17 Suppl 1**:S55-S60.

(120) Duggirala R, Almasy L, Blangero J. Smoking behavior is under the influence of a major quantitative trait locus on human chromosome 5q. Genet Epidemiol 1999;**17 Suppl 1**:S139-S144.

(121) Li MD, Ma JZ, Cheng R, Dupont RT, Williams NJ, Crews KM, Payne TJ, Elston RC. A genome-wide scan to identify loci for smoking rate in the Framingham Heart Study population. BMC Genet 2003;**4 Suppl 1**:S103.

(122) Bierut LJ, Rice JP, Goate A, Hinrichs AL, Saccone NL, Foroud T, Edenberg HJ, Cloninger CR, Begleiter H, Conneally PM, Crowe RR, Hesselbrock V, Li TK, Nurnberger JI, Jr., Porjesz B, Schuckit MA, Reich T. A genomic scan for habitual smoking in families of alcoholics: common and specific genetic factors in substance dependence. Am J Med Genet 2004 Jan 1;**124A**(1):19-27.

(123) Gelernter J, Liu X, Hesselbrock V, Page GP, Goddard A, Zhang H. Results of a genomewide linkage scan: support for chromosomes 9 and 11 loci increasing risk for cigarette smoking. Am J Med Genet 2004 Jul 1;**128B**(1):94-101.

(124) Vink JM, Posthuma D, Neale MC, Eline SP, Boomsma DI. Genome-wide linkage scan to identify Loci for age at first cigarette in Dutch sibling pairs. Behav Genet 2006 Jan;**36**(1):100-11.

(125) Ehlers CL, Wilhelmsen KC. Genomic screen for loci associated with tobacco usage in Mission Indians. BMC Med Genet 2006;**7**:9.

(126) Morley KI, Medland SE, Ferreira MA, Lynskey MT, Montgomery GW, Heath AC, Madden PA, Martin NG. A possible smoking susceptibility locus on chromosome

11p12: evidence from sex-limitation linkage analyses in a sample of Australian twin families. Behav Genet 2006 Jan;**36**(1):87-99.

(127) Goode EL, Badzioch MD, Kim H, Gagnon F, Rozek LS, Edwards KL, Jarvik GP. Multiple genome-wide analyses of smoking behavior in the Framingham Heart Study. BMC Genet 2003;**4 Suppl 1**:S102.

(128) Saccone NL, Neuman RJ, Saccone SF, Rice JP. Genetic analysis of maximum cigarette-use phenotypes. BMC Genet 2003;**4 Suppl 1**:S105.

(129) Sullivan PF, Neale BM, van den OE, Miles MF, Neale MC, Bulik CM, Joyce PR, Straub RE, Kendler KS. Candidate genes for nicotine dependence via linkage, epistasis, and bioinformatics. Am J Med Genet 2004 Apr 1;**126B**(1):23-36.

(130) Wang D, Ma JZ, Li MD. Mapping and verification of susceptibility loci for smoking quantity using permutation linkage analysis. Pharmacogenomics J 2005;**5**(3):166-72.

(131) Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, Swan GE, Goate AM, Rutter J, Bertelsen S, Fox L, Fugman D, Martin NG, Montgomery GW, Wang JC, Ballinger DG, Rice JP, Bierut LJ. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 2007 Jan 1;**16**(1):36-49.

(132) Tricker AR. Nicotine metabolism, human drug metabolism polymorphisms, and smoking behaviour. Toxicology 2003 Feb 1;**183**(1-3):151-73.

(133) Benowitz NL, Porchet H, Jacob P, III. Nicotine dependence and tolerance in man: pharmacokinetic and pharmacodynamic investigations. Prog Brain Res 1989;**79**:279-87.

(134) McMorrow MJ, Foxx RM. Nicotine's role in smoking: an analysis of nicotine regulation. Psychol Bull 1983 Mar;**93**(2):302-27.

(135) Henningfield JE, Heishman SJ. The addictive role of nicotine in tobacco use. Psychopharmacology (Berl) 1995 Jan;**117**(1):11-3.

(136) Dani JA, De Biasi M. Cellular mechanisms of nicotine addiction. Pharmacol Biochem Behav 2001 Dec;**70**(4):439-46.

(137) Di Chiara G. Nucleus accumbens shell and core dopamine: differential role in behavior and addiction. Behav Brain Res 2002 Dec 2;**137**(1-2):75-114.

(138) Byrd GD, Robinson JH, Caldwell WS, deBethizy JD. Comparison of measured and FTC-predicted nicotine uptake in smokers. Psychopharmacology (Berl) 1995 Nov;**122**(2):95-103.

(139) Saareks V, Riutta A, Alanko J, Ylitalo P, Parviainen M, Mucha I, Sievi E, Vapaatalo H. Clinical pharmacology of eicosanoids, nicotine induced changes in man. J Physiol Pharmacol 2000 Dec;**51**(4 Pt 1):631-42.

(140) Benowitz NL, Kuyt F, Jacob P, III, Jones RT, Osman AL. Cotinine disposition and effects. Clin Pharmacol Ther 1983 Nov;**34**(5):604-11.

(141) Zevin S, Jacob P, III, Benowitz N. Cotinine effects on nicotine metabolism. Clin Pharmacol Ther 1997 Jun;**61**(6):649-54.

(142) Benowitz NL, Jacob P, III. Metabolism of nicotine to cotinine studied by a dual stable isotope method. Clin Pharmacol Ther 1994 Nov;**56**(5):483-93.

(143) Hadidi H, Irshaid Y, Vagbo CB, Brunsvik A, Cholerton S, Zahlsen K, Idle JR. Variability of coumarin 7- and 3-hydroxylation in a Jordanian population is suggestive of a functional polymorphism in cytochrome P450 CYP2A6. Eur J Clin Pharmacol 1998 Jul;**54**(5):437-41.

(144) Messina ES, Tyndale RF, Sellers EM. A major role for CYP2A6 in nicotine C-oxidation by human liver microsomes. J Pharmacol Exp Ther 1997 Sep;**282**(3):1608-14.

(145) Nakajima M, Yamamoto T, Nunoya K, Yokoi T, Nagashima K, Inoue K, Funae Y, Shimada N, Kamataki T, Kuroiwa Y. Role of human cytochrome P4502A6 in C-oxidation of nicotine. Drug Metab Dispos 1996 Nov;**24**(11):1212-7.

(146) Nakajima M, Yamagishi S, Yamamoto H, Yamamoto T, Kuroiwa Y, Yokoi T. Deficient cotinine formation from nicotine is attributed to the whole deletion of the CYP2A6 gene in humans. Clin Pharmacol Ther 2000 Jan;**67**(1):57-69.

(147) Nakajima M, Kwon JT, Tanaka N, Zenta T, Yamamoto Y, Yamamoto H, Yamazaki H, Yamamoto T, Kuroiwa Y, Yokoi T. Relationship between interindividual differences in nicotine metabolism and CYP2A6 genetic polymorphism in humans. Clin Pharmacol Ther 2001 Jan;**69**(1):72-8.

(148) Tyndale RF, Pianezza ML, Sellers EM. A common genetic defect in nicotine metabolism decreases risk for dependence and lowers cigarette consumption. Nicotine Tob Res 1999;**1 Suppl 2**:S63-S67.

(149) Zabetian CP, Gelernter J, Cubells JF. Functional variants at CYP2A6: new genotyping methods, population genetics, and relevance to studies of tobacco dependence. Am J Med Genet 2000 Oct 9;**96**(5):638-45.

(150) Slaughter RL, Edwards DJ. Recent advances: the cytochrome P450 enzymes. Ann Pharmacother 1995 Jun;**29**(6):619-24.

(151) Nowell S, Sweeney C, Hammons G, Kadlubar FF, Lang NP. CYP2A6 activity determined by caffeine phenotyping: association with colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 2002 Apr;**11**(4):377-83.

(152) Pelkonen O, Pasanen M, Kuha H, Gachalyi B, Kairaluoma M, Sotaniemi EA, Park SS, Friedman FK, Gelboin HV. The effect of cigarette smoking on 7-ethoxyresorufin O-deethylase and other monooxygenase activities in human liver: analyses with monoclonal antibodies. Br J Clin Pharmacol 1986 Aug;**22**(2):125-34.

(153) Schoedel KA, Sellers EM, Palmour R, Tyndale RF. Down-regulation of hepatic nicotine metabolism and a CYP2A6-like enzyme in African green monkeys after long-term nicotine administration. Mol Pharmacol 2003 Jan;**63**(1):96-104.

(154) Fernandez-Salguero P, Hoffman SM, Cholerton S, Mohrenweiser H, Raunio H, Rautio A, Pelkonen O, Huang JD, Evans WE, Idle JR. A genetic polymorphism in coumarin 7-hydroxylation: sequence of the human CYP2A genes and identification of variant CYP2A6 alleles. Am J Hum Genet 1995 Sep;**57**(3):651-60.

(155) Fukami T, Nakajima M, Yamanaka H, Fukushima Y, McLeod HL, Yokoi T. A Novel Duplication Type of CYP2A6 Gene in African-American Population. Drug Metab Dispos 2007 Apr;**35**(4):515-20.

(156) Miles JS, Bickmore W, Brook JD, McLaren AW, Meehan R, Wolf CR. Close linkage of the human cytochrome P450IIA and P450IIB gene subfamilies: implications for the assignment of substrate specificity. Nucleic Acids Res 1989 Apr 25;**17**(8):2907-17.

(157) Yamano S, Tatsuno J, Gonzalez FJ. The CYP2A3 gene product catalyzes coumarin 7-hydroxylation in human liver microsomes. Biochemistry 1990 Feb 6;**29**(5):1322-9.

(158) Pelkonen O, Raunio H, Rautio A, Lang M. Xenobiotic-metabolizing enzymes and cancer risk: correspondence between genotype and phenotype. IARC Sci Publ 1999;(148):77-88.

(159) Nebert DW. Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist? Am J Hum Genet 1997 Feb;**60**(2):265-71.

(160) Nunoya KI, Yokoi T, Kimura K, Kainuma T, Satoh K, Kinoshita M, Kamataki T. A new CYP2A6 gene deletion responsible for the in vivo polymorphic metabolism of (+)-cis-3,5-dimethyl-2-(3-pyridyl)thiazolidin-4-one hydrochloride in humans. J Pharmacol Exp Ther 1999 Apr;**289**(1):437-42.

(161) Oscarson M, McLellan RA, Gullsten H, Yue QY, Lang MA, Bernal ML, Sinues B, Hirvonen A, Raunio H, Pelkonen O, Ingelman-Sundberg M. Characterisation and PCR-based detection of a CYP2A6 gene deletion found at a high frequency in a Chinese population. FEBS Lett 1999 Apr 1;**448**(1):105-10.

(162) Hadidi H, Zahlsen K, Idle JR, Cholerton S. A single amino acid substitution (Leu160His) in cytochrome P450 CYP2A6 causes switching from 7-hydroxylation to 3-hydroxylation of coumarin. Food Chem Toxicol 1997 Sep;**35**(9):903-7.

(163) Oscarson M, Gullsten H, Rautio A, Bernal ML, Sinues B, Dahl ML, Stengard JH, Pelkonen O, Raunio H, Ingelman-Sundberg M. Genotyping of human cytochrome P450 2A6 (CYP2A6), a nicotine C-oxidase. FEBS Lett 1998 Nov 6;**438**(3):201-5.

(164) Gu DF, Hinks LJ, Morton NE, Day IN. The use of long PCR to confirm three common alleles at the CYP2A6 locus and the relationship between genotype and smoking habit. Ann Hum Genet 2000 Sep;**64**(Pt 5):383-90.

(165) Chen GF, Tang YM, Green B, Lin DX, Guengerich FP, Daly AK, Caporaso NE, Kadlubar FF. Low frequency of CYP2A6 gene polymorphism as revealed by a one-step polymerase chain reaction method. Pharmacogenetics 1999 Jun;9(3):327-32.

(166) Kitagawa K, Kunugita N, Katoh T, Yang M, Kawamoto T. The significance of the homozygous CYP2A6 deletion on nicotine metabolism: a new genotyping method of CYP2A6 using a single PCR-RFLP. Biochem Biophys Res Commun 1999 Aug 19;**262**(1):146-51.

(167) Oscarson M, McLellan RA, Gullsten H, Agundez JA, Benitez J, Rautio A, Raunio H, Pelkonen O, Ingelman-Sundberg M. Identification and characterisation of novel polymorphisms in the CYP2A locus: implications for nicotine metabolism. FEBS Lett 1999 Oct 29;**460**(2):321-7.

(168) Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, Sellers EM, Tyndale RF. Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol Pharmacol 2000 Oct;**58**(4):747-55.

(169) Schulz TG, Ruhnau P, Hallier E. Lack of correlation between CYP2A6 genotype and smoking habits. Adv Exp Med Biol 2001;**500**:213-5.

(170) Yang M, Kunugita N, Kitagawa K, Kang SH, Coles B, Kadlubar FF, Katoh T, Matsuno K, Kawamoto T. Individual differences in urinary cotinine levels in Japanese smokers: relation to genetic polymorphism of drug-metabolizing enzymes. Cancer Epidemiol Biomarkers Prev 2001 Jun;**10**(6):589-93.

(171) Ariyoshi N, Takahashi Y, Miyamoto M, Umetsu Y, Daigo S, Tateishi T, Kobayashi S, Mizorogi Y, Loriot MA, Stucker I, Beaune P, Kinoshita M, Kamataki T. Structural characterization of a new variant of the CYP2A6 gene (CYP2A6*1B) apparently diagnosed as heterozygotes of CYP2A6*1A and CYP2A6*4C. Pharmacogenetics 2000 Nov;**10**(8):687-93.

(172) Nakajima M, Yoshida R, Fukami T, McLeod HL, Yokoi T. Novel human CYP2A6 alleles confound gene deletion analysis. FEBS Lett 2004 Jul 2;**569**(1-3):75-81.

(173) Oscarson M, McLellan RA, Asp V, Ledesma M, Ruiz ML, Sinues B, Rautio A, Ingelman-Sundberg M. Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. Hum Mutat 2002 Oct;**20**(4):275-83.

(174) Kitagawa K, Kunugita N, Kitagawa M, Kawamoto T. CYP2A6*6, a novel polymorphism in cytochrome p450 2A6, has a single amino acid substitution (R128Q) that inactivates enzymatic activity. J Biol Chem 2001 May 25;**276**(21):17830-5.

(175) Xu C, Rao YS, Xu B, Hoffmann E, Jones J, Sellers EM, Tyndale RF. An in vivo pilot study characterizing the new CYP2A6*7, *8, and *10 alleles. Biochem Biophys Res Commun 2002 Jan 11;**290**(1):318-24.

(176) Pitarque M, von Richter O, Oke B, Berkkan H, Oscarson M, Ingelman-Sundberg M. Identification of a single nucleotide polymorphism in the TATA box of the CYP2A6

gene: impairment of its promoter activity. Biochem Biophys Res Commun 2001 Jun 8;**284**(2):455-60.

(177) Daigo S, Takahashi Y, Fujieda M, Ariyoshi N, Yamazaki H, Koizumi W, Tanabe S, Saigenji K, Nagayama S, Ikeda K, Nishioka Y, Kamataki T. A novel mutant allele of the CYP2A6 gene (CYP2A6*11) found in a cancer patient who showed poor metabolic phenotype towards tegafur. Pharmacogenetics 2002 Jun;**12**(4):299-306.

(178) Kazuma K, Masaki F, Hiroshi Y, Tsutomu S, F.Peter G, Andrew P, Kazuko N, Takashi I, and Tetsuya K. Twenty One Novel Single Nucleotide Polymorphisms (SNPs) of the CYP2A6 Gene in Japanese and Caucasians. Drug Metab Pharmacokin 2002;**17**(5):482-7.

(179) Fukami T, Nakajima M, Higashi E, Yamanaka H, McLeod HL, Yokoi T. A novel CYP2A6*20 allele found in African-American population produces a truncated protein lacking enzymatic activity. Biochem Pharmacol 2005 Sep 1;**70**(5):801-8.

(180) Fukami T, Nakajima M, Higashi E, Yamanaka H, Sakai H, McLeod HL, Yokoi T. Characterization of novel CYP2A6 polymorphic alleles (CYP2A6*18 and CYP2A6*19) that affect enzymatic activity. Drug Metab Dispos 2005 Aug;**33**(8):1202-10.

(181) Haberl M, Anwald B, Klein K, Weil R, Fuss C, Gepdiremen A, Zanger UM, Meyer UA, Wojnowski L. Three haplotypes associated with CYP2A6 phenotypes in Caucasians. Pharmacogenet Genomics 2005 Sep;**15**(9):609-24.

(182) Nakajima M, Fukami T, Yamanaka H, Higashi E, Sakai H, Yoshida R, Kwon JT, McLeod HL, Yokoi T. Comprehensive evaluation of variability in nicotine metabolism and CYP2A6 polymorphic alleles in four ethnic populations. Clin Pharmacol Ther 2006 Sep;**80**(3):282-97.

(183) Pitarque M, von Richter O, Rodriguez-Antona C, Wang J, Oscarson M, Ingelman-Sundberg M. A nicotine C-oxidase gene (CYP2A6) polymorphism important for promoter activity. Hum Mutat 2004 Mar;**23**(3):258-66.

(184) Solus JF, Arietta BJ, Harris JR, Sexton DP, Steward JQ, McMunn C, Ihrie P, Mehall JM, Edwards TL, Dawson EP. Genetic variation in eleven phase I drug metabolism genes in an ethnically diverse population. Pharmacogenomics 2004 Oct;**5**(7):895-931.

(185) von Richter O, Pitarque M, Rodriguez-Antona C, Testa A, Mantovani R, Oscarson M, Ingelman-Sundberg M. Polymorphic NF-Y dependent regulation of human nicotine C-oxidase (CYP2A6). Pharmacogenetics 2004 Jun;**14**(6):369-79.

(186) Cholerton S, Daly AK, Idle JR. The role of individual human cytochromes P450 in drug metabolism and clinical response. Trends Pharmacol Sci 1992 Dec;**13**(12):434-9.

(187) Iscan M, Rostami H, Iscan M, Guray T, Pelkonen O, Rautio A. Interindividual variability of coumarin 7-hydroxylation in a Turkish population. Eur J Clin Pharmacol 1994;**47**(4):315-8.

(188) Rautio A, Kraul H, Kojo A, Salmela E, Pelkonen O. Interindividual variability of coumarin 7-hydroxylation in healthy volunteers. Pharmacogenetics 1992 Oct;2(5):227-33.

(189) Inoue K, Yamazaki H, Shimada T. CYP2A6 genetic polymorphisms and liver microsomal coumarin and nicotine oxidation activities in Japanese and Caucasians. Arch Toxicol 2000 Jan;73(10-11):532-9.

(190) Ariyoshi N, Sawamura Y, Kamataki T. A novel single nucleotide polymorphism altering stability and activity of CYP2a6. Biochem Biophys Res Commun 2001 Mar 2;281(3):810-4.

(191) Kwon JT, Nakajima M, Chai S, Yom YK, Kim HK, Yamazaki H, Sohn DR, Yamamoto T, Kuroiwa Y, Yokoi T. Nicotine metabolism and CYP2A6 allele frequencies in Koreans. Pharmacogenetics 2001 Jun;11(4):317-23.

(192) Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking. Nature 1998 Jun 25;393(6687):750.

(193) Schoedel KA, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF. Ethnic variation in CYP2A6 and association of genetically slow nicotine metabolism and smoking in adult Caucasians. Pharmacogenetics 2004 Sep;14(9):615-26.

(194) London SJ, Idle JR, Daly AK, Coetzee GA. Genetic variation of CYP2A6, smoking, and risk of cancer. Lancet 1999 Mar 13;353(9156):898-9.

(195) Loriot MA, Rebuissou S, Oscarson M, Cenee S, Miyamoto M, Ariyoshi N, Kamataki T, Hemon D, Beaune P, Stucker I. Genetic polymorphisms of cytochrome P450 2A6 in a case-control study on lung cancer in a French population. Pharmacogenetics 2001 Feb;11(1):39-44.

(196) Tan W, Chen GF, Xing DY, Song CY, Kadlubar FF, Lin DX. Frequency of CYP2A6 gene deletion and its relation to risk of lung and esophageal cancer in the Chinese population. Int J Cancer 2001 Mar 20;95(2):96-101.

(197) Tiihonen J, Pesonen U, Kauhanen J, Koulu M, Hallikainen T, Leskinen L, Salonen JT. CYP2A6 genotype and smoking. Mol Psychiatry 2000 Jul;5(4):347-8.

(198) Zhang X, Amemo K, Ameno S, Iwahashi K, Kinoshita H, Kubota T, Mostofa J, Ijiri I. Lack of association between smoking and CYP2A6 gene polymorphisms in A Japanese population. Nihon Arukoru Yakubutsu Igakkai Zasshi 2001 Oct;36(5):486-90.

(199) Gough AC, Smith CA, Howell SM, Wolf CR, Bryant SP, Spurr NK. Localization of the CYP2D gene locus to human chromosome 22q13.1 by polymerase chain reaction, in situ hybridization, and linkage analysis. Genomics 1993 Feb;15(2):430-2.

(200) Cholerton S, Boustead C, Taber H, Arpanahi A, Idle JR. CYP2D6 genotypes in cigarette smokers and non-tobacco users. Pharmacogenetics 1996 Jun;6(3):261-3.

(201) Boustead C, Taber H, Idle JR, Cholerton S. CYP2D6 genotype and smoking behaviour in cigarette smokers. Pharmacogenetics 1997 Oct;7(5):411-4.

(202) Saarikoski ST, Sata F, Husgafvel-Pursiainen K, Rautalahti M, Haukka J, Impivaara O, Jarvisalo J, Vainio H, Hirvonen A. CYP2D6 ultrarapid metabolizer genotype as a potential modifier of smoking behaviour. Pharmacogenetics 2000 Feb;10(1):5-10.

(203) Hoffman SM, Fernandez-Salguero P, Gonzalez FJ, Mohrenweiser HW. Organization and evolution of the cytochrome P450 CYP2A-2B-2F subfamily gene cluster on human chromosome 19. J Mol Evol 1995 Dec;41(6):894-900.

(204) Lang T, Klein K, Fischer J, Nussler AK, Neuhaus P, Hofmann U, Eichelbaum M, Schwab M, Zanger UM. Extensive genetic polymorphism in the human CYP2B6 gene with impact on expression and function in human liver. Pharmacogenetics 2001 Jul;11(5):399-415.

(205) Ekins S, Wrighton SA. The role of CYP2B6 in human xenobiotic metabolism. Drug Metab Rev 1999 Aug;31(3):719-54.

(206) Yamazaki H, Inoue K, Hashimoto M, Shimada T. Roles of CYP2A6 and CYP2B6 in nicotine C-oxidation by human liver microsomes. Arch Toxicol 1999 Mar;73(2):65-70.

(207) Lerman C, Shields PG, Wileyto EP, Audrain J, Pinto A, Hawk L, Krishnan S, Niaura R, Epstein L. Pharmacogenetic investigation of smoking cessation treatment. Pharmacogenetics 2002 Nov;12(8):627-34.

(208) Landry DW. Immunotherapy for cocaine addiction. Sci Am 1997 Feb;276(2):42-5.

(209) Siegel A, Roeling TA, Gregg TR, Kruk MR. Neuropharmacology of brain-stimulation-evoked aggression. Neurosci Biobehav Rev 1999;23(3):359-89.

(210) Itier V, Bertrand D. Neuronal nicotinic receptors: from protein structure to function. FEBS Lett 2001 Aug 31;504(3):118-25.

(211) Anney RJ, Olsson CA, Lotfi-Miri M, Patton GC, Williamson R. Nicotine dependence in a prospective population-based study of adolescents: the protective role of a functional tyrosine hydroxylase polymorphism. Pharmacogenetics 2004 Feb;14(2):73-81.

(212) Olsson C, Anney R, Forrest S, Patton G, Coffey C, Cameron T, Hassett A, Williamson R. Association between dependent smoking and a polymorphism in the tyrosine hydroxylase gene in a prospective population-based study of adolescent health. Behav Genet 2004 Jan;34(1):85-91.

(213) Rodriguez S, Huang S, Chen XH, Gaunt TR, Syddall HE, Gilg JA, Miller GJ, Cooper CC, Cook DG, Whincup PH, Day IN. A study of TH01 and IGF2-INS-TH haplotypes in relation to smoking initiation in three independent surveys. Pharmacogenet Genomics 2006 Jan;16(1):15-23.

(214) La Grange L, Jones TD, Erb L, Reyes E. Alcohol consumption: biochemical and personality correlates in a college student population. Addict Behav 1995 Jan;20(1):93-103.

(215) McKinney EF, Walton RT, Yudkin P, Fuller A, Haldar NA, Mant D, Murphy M, Welsh KI, Marshall SE. Association between polymorphisms in dopamine metabolic enzymes and tobacco consumption in smokers. Pharmacogenetics 2000 Aug;10(6):483-91.

(216) Johnstone EC, Yudkin PL, Hey K, Roberts SJ, Welch SJ, Murphy MF, Griffiths SE, Walton RT. Genetic variation in dopaminergic pathways and short-term effectiveness of the nicotine patch. Pharmacogenetics 2004 Feb;14(2):83-90.

(217) Coron B, Campion D, Thibaut F, Dollfus S, Preterre P, Langlois S, Vasse T, Moreau V, Martin C, Charbonnier F, Laurent C, Mallet J, Petit M, Frebourg T. Association study between schizophrenia and monoamine oxidase A and B DNA polymorphisms. Psychiatry Res 1996 Jun 1;62(3):221-6.

(218) Fowler JS, Volkow ND, Wang GJ, Pappas N, Logan J, Shea C, Alexoff D, MacGregor RR, Schlyer DJ, Zezulkova I, Wolf AP. Brain monoamine oxidase A inhibition in cigarette smokers. Proc Natl Acad Sci U S A 1996 Nov 26;93(24):14065-9.

(219) Fowler JS, Volkow ND, Wang GJ, Pappas N, Logan J, MacGregor R, Alexoff D, Shea C, Schlyer D, Wolf AP, Warner D, Zezulkova I, Cilento R. Inhibition of monoamine oxidase B in the brains of smokers. Nature 1996 Feb 22;379(6567):733-6.

(220) Sobell JL, Lind TJ, Hebrink DD, Heston LL, Sommer SS. Screening the monoamine oxidase B gene in 100 male patients with schizophrenia: a cluster of polymorphisms in African-Americans but lack of functionally significant sequence changes. Am J Med Genet 1997 Feb 21;74(1):44-9.

(221) Berlin I, Said S, Spreux-Varoquaux O, Launay JM, Olivares R, Millet V, Lecrubier Y, Puech AJ. A reversible monoamine oxidase A inhibitor (moclobemide) facilitates smoking cessation and abstinence in heavy, dependent smokers. Clin Pharmacol Ther 1995 Oct;58(4):444-52.

(222) Ito H, Hamajima N, Matsuo K, Okuma K, Sato S, Ueda R, Tajima K. Monoamine oxidase polymorphisms and smoking behaviour in Japanese. Pharmacogenetics 2003 Feb;13(2):73-9.

(223) Elsworth JD, Roth RH. Dopamine synthesis, uptake, metabolism, and receptors: relevance to gene therapy of Parkinson's disease. Exp Neurol 1997 Mar;144(1):4-9.

(224) Karayiorgou M, Altemus M, Galke BL, Goldman D, Murphy DL, Ott J, Gogos JA. Genotype determining low catechol-O-methyltransferase activity as a risk factor for obsessive-compulsive disorder. Proc Natl Acad Sci U S A 1997 Apr 29;94(9):4572-5.

(225) Lachman HM, Papolos DF, Saito T, Yu YM, Szumlanski CL, Weinshilboum RM. Human catechol-O-methyltransferase pharmacogenetics: description of a functional

polymorphism and its potential application to neuropsychiatric disorders. Pharmacogenetics 1996 Jun;6(3):243-50.

(226) Vandenbergh DJ, Rodriguez LA, Miller IT, Uhl GR, Lachman HM. High-activity catechol-O-methyltransferase allele is more prevalent in polysubstance abusers. Am J Med Genet 1997 Jul 25;74(4):439-42.

(227) David SP, Johnstone E, Griffiths SE, Murphy M, Yudkin P, Mant D, Walton R. No association between functional catechol O-methyl transferase 1947A>G polymorphism and smoking initiation, persistent smoking or smoking cessation. Pharmacogenetics 2002 Apr;12(3):265-8.

(228) Comings DE, Gade R, Wu S, Chiu C, Dietz G, Muhleman D, Saucier G, Ferry L, Rosenthal RJ, Lesieur HR, Rugle LJ, MacMurray P. Studies of the potential role of the dopamine D1 receptor gene in addictive behaviors. Mol Psychiatry 1997 Jan;2(1):44-56.

(229) Noble EP. D2 dopamine receptor gene in psychiatric and neurologic disorders and its phenotypes. Am J Med Genet 2003 Jan 1;116B(1):103-25.

(230) Munafo M, Clark T, Johnstone E, Murphy M, Walton R. The genetic basis for smoking behavior: a systematic review and meta-analysis. Nicotine Tob Res 2004 Aug;6(4):583-97.

(231) Noble EP, St Jeor ST, Ritchie T, Syndulko K, St Jeor SC, Fitch RJ, Brunner RL, Sparkes RS. D2 dopamine receptor gene and cigarette smoking: a reward gene? Med Hypotheses 1994 Apr;42(4):257-60.

(232) Comings DE, Ferry L, Bradshaw-Robinson S, Burchette R, Chiu C, Muhleman D. The dopamine D2 receptor (DRD2) gene: a genetic risk factor in smoking. Pharmacogenetics 1996 Feb;6(1):73-9.

(233) Li MD, Ma JZ, Beuten J. Progress in searching for susceptibility loci and genes for smoking-related behaviour. Clin Genet 2004 Nov;66(5):382-92.

(234) Czermak C, Lehofer M, Wagner EM, Prietl B, Gorkiewicz G, Lemonis L, Rohrhofer A, Legl T, Schauenstein K, Liebmann PM. Reduced dopamine D3 receptor expression in blood lymphocytes of smokers is negatively correlated with daily number of smoked cigarettes: a peripheral correlate of dopaminergic alterations in smokers. Nicotine Tob Res 2004 Feb;6(1):49-54.

(235) Benjamin J, Li L, Patterson C, Greenberg BD, Murphy DL, Hamer DH. Population and familial association between the D4 dopamine receptor gene and measures of Novelty Seeking. Nat Genet 1996 Jan;12(1):81-4.

(236) Ebstein RP, Novick O, Umansky R, Priel B, Osher Y, Blaine D, Bennett ER, Nemanov L, Katz M, Belmaker RH. Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of Novelty Seeking. Nat Genet 1996 Jan;12(1):78-80.

(237) Keltikangas-Jarvinen L, Elovainio M, Kivimaki M, Lichtermann D, Ekelund J, Peltonen L. Association between the type 4 dopamine receptor gene polymorphism and novelty seeking. Psychosom Med 2003 May;**65**(3):471-6.

(238) Hutchison KE, McGeary J, Smolen A, Bryan A, Swift RM. The DRD4 VNTR polymorphism moderates craving after alcohol consumption. Health Psychol 2002 Mar;**21**(2):139-46.

(239) Hutchison KE, LaChance H, Niaura R, Bryan A, Smolen A. The DRD4 VNTR polymorphism influences reactivity to smoking cues. J Abnorm Psychol 2002 Feb;**111**(1):134-43.

(240) Li T, Xu K, Deng H, Cai G, Liu J, Liu X, Wang R, Xiang X, Zhao J, Murray RM, Sham PC, Collier DA. Association analysis of the dopamine D4 gene exon III VNTR and heroin abuse in Chinese subjects. Mol Psychiatry 1997 Sep;**2**(5):413-6.

(241) Vandenbergh DJ, Rodriguez LA, Hivert E, Schiller JH, Villareal G, Pugh EW, Lachman H, Uhl GR. Long forms of the dopamine receptor (DRD4) gene VNTR are more prevalent in substance abusers: no interaction with functional alleles of the catechol-o-methyltransferase (COMT) gene. Am J Med Genet 2000 Oct 9;**96**(5):678-83.

(242) Pomerleau CS, Pomerleau OF, Flessland KA, Basson SM. Relationship of Tridimensional Personality Questionnaire scores and smoking variables in female and male smokers. J Subst Abuse 1992;**4**(2):143-54.

(243) Pomerleau OF, Fagerstrom KO, Marks JL, Tate JC, Pomerleau CS. Development and validation of a self-rating scale for positive- and negative-reinforcement smoking: The Michigan Nicotine Reinforcement Questionnaire. Nicotine Tob Res 2003 Oct;**5**(5):711-8.

(244) Shields PG, Lerman C, Audrain J, Bowman ED, Main D, Boyd NR, Caporaso NE. Dopamine D4 receptors and the risk of cigarette smoking in African-Americans and Caucasians. Cancer Epidemiol Biomarkers Prev 1998 Jun;**7**(6):453-8.

(245) Lerman C, Caporaso N, Main D, Audrain J, Boyd NR, Bowman ED, Shields PG. Depression and self-medication with nicotine: the modifying influence of the dopamine D4 receptor gene. Health Psychol 1998 Jan;**17**(1):56-62.

(246) Luciano M, Zhu G, Kirk KM, Whitfield JB, Butler R, Heath AC, Madden PA, Martin NG. Effects of dopamine receptor D4 variation on alcohol and tobacco use and on novelty seeking: multivariate linkage and association analysis. Am J Med Genet 2004 Jan 1;**124B**(1):113-23.

(247) Cravchik A, Gejman PV. Functional analysis of the human D5 dopamine receptor missense and nonsense variants: differences in dopamine binding affinities. Pharmacogenetics 1999 Apr;**9**(2):199-206.

(248) Vanyukov MM, Moss HB, Gioio AE, Hughes HB, Kaplan BB, Tarter RE. An association between a microsatellite polymorphism at the DRD5 gene and the liability to substance abuse: pilot study. Behav Genet 1998 Mar;**28**(2):75-82.

(249) Sullivan PF, Neale MC, Silverman MA, Harris-Kerr C, Myakishev MV, Wormley B, Webb BT, Ma Y, Kendler KS, Straub RE. An association study of DRD5 with smoking initiation and progression to nicotine dependence. Am J Med Genet 2001 Apr 8;**105**(3):259-65.

(250) Vandenbergh DJ, Persico AM, Hawkins AL, Griffin CA, Li X, Jabs EW, Uhl GR. Human dopamine transporter gene (DAT1) maps to chromosome 5p15.3 and displays a VNTR. Genomics 1992 Dec;**14**(4):1104-6.

(251) Gelernter J, Kranzler HR, Satel SL, Rao PA. Genetic association between dopamine transporter protein alleles and cocaine-induced paranoia. Neuropsychopharmacology 1994 Nov;**11**(3):195-200.

(252) Comings DE, Wu S, Chiu C, Ring RH, Gade R, Ahn C, MacMurray JP, Dietz G, Muhleman D. Polygenic inheritance of Tourette syndrome, stuttering, attention deficit hyperactivity, conduct, and oppositional defiant disorder: the additive and subtractive effect of the three dopaminergic genes--DRD2, D beta H, and DAT1. Am J Med Genet 1996 May 31;**67**(3):264-88.

(253) Cook EH, Jr., Stein MA, Krasowski MD, Cox NJ, Olkon DM, Kieffer JE, Leventhal BL. Association of attention-deficit disorder and the dopamine transporter gene. Am J Hum Genet 1995 Apr;**56**(4):993-8.

(254) Lerman C, Caporaso NE, Audrain J, Main D, Bowman ED, Lockshin B, Boyd NR, Shields PG. Evidence suggesting the role of specific genetic factors in cigarette smoking. Health Psychol 1999 Jan;**18**(1):14-20.

(255) Jorm AF, Henderson AS, Jacomb PA, Christensen H, Korten AE, Rodgers B, Tan X, Easteal S. Association of smoking and personality with a polymorphism of the dopamine transporter gene: results from a community survey. Am J Med Genet 2000 Jun 12;96(3):331-4.

(256) Vandenbergh DJ, Bennett CJ, Grant MD, Strasser AA, O'Connor R, Stauffer RL, Vogler GP, Kozlowski LT. Smoking status and the human dopamine transporter variable number of tandem repeats (VNTR) polymorphism: failure to replicate and finding that never-smokers may be different. Nicotine Tob Res 2002 Aug;4(3):333-40.

(257) Erblich J, Lerman C, Self DW, Diaz GA, Bovbjerg DH. Effects of dopamine D2 receptor (DRD2) and transporter (SLC6A3) polymorphisms on smoking cue-induced cigarette craving among African-American smokers. Mol Psychiatry 2004 Sep 21.

(258) Erblich J, Lerman C, Self DW, Diaz GA, Bovbjerg DH. Stress-induced cigarette craving: effects of the DRD2 TaqI RFLP and SLC6A3 VNTR polymorphisms. Pharmacogenomics J 2004;4(2):102-9.

(259) Porras G, Di M, V, Fracasso C, Lucas G, De DP, Caccia S, Esposito E, Spampinato U. 5-HT2A and 5-HT2C/2B receptor subtypes modulate dopamine release induced in vivo by amphetamine and morphine in both the rat nucleus accumbens and striatum. Neuropsychopharmacology 2002 Mar;**26**(3):311-24.

(260) Ribeiro EB, Bettiker RL, Bogdanov M, Wurtman RJ. Effects of systemic nicotine on serotonin release in rat brain. Brain Res 1993 Sep 10;**621**(2):311-8.

(261) Mihailescu S, Palomero-Rivero M, Meade-Huerta P, Maza-Flores A, Drucker-Colin R. Effects of nicotine and mecamylamine on rat dorsal raphe neurons. Eur J Pharmacol 1998 Oct 30;**360**(1):31-6.

(262) Gilbert DG, Gilbert BO. Personality, psychopathology, and nicotine response as mediators of the genetics of smoking. Behav Genet 1995 Mar;**25**(2):133-47.

(263) Lerman C, Shields PG, Audrain J, Main D, Cobb B, Boyd NR, Caporaso N. The role of the serotonin transporter gene in cigarette smoking. Cancer Epidemiol Biomarkers Prev 1998 Mar;7(3):253-5.

(264) Lerman C, Caporaso NE, Bush A, Zheng YL, Audrain J, Main D, Shields PG. Tryptophan hydroxylase gene variant and smoking behavior. Am J Med Genet 2001 Aug 8;**105**(6):518-20.

(265) Sullivan PF, Jiang Y, Neale MC, Kendler KS, Straub RE. Association of the tryptophan hydroxylase gene with smoking initiation but not progression to nicotine dependence. Am J Med Genet 2001 Jul 8;**105**(5):479-84.

(266) Ishikawa H, Ohtsuki T, Ishiguro H, Yamakawa-Kobayashi K, Endo K, Lin YL, Yanagi H, Tsuchiya S, Kawata K, Hamaguchi H, Arinami T. Association between serotonin transporter gene polymorphism and smoking among Japanese males. Cancer Epidemiol Biomarkers Prev 1999 Sep;**8**(9):831-3.

(267) do Prado-Lima PA, Chatkin JM, Taufer M, Oliveira G, Silveira E, Neto CA, Haggstram F, Bodanese LC, da Cruz IB. Polymorphism of 5HT2A serotonin receptor gene is implicated in smoking addiction. Am J Med Genet 2004 Jul 1;**128B**(1):90-3.

(268) Nielsen DA, Jenkins GL, Stefanisko KM, Jefferson KK, Goldman D. Sequence, splice site and population frequency distribution analyses of the polymorphic human tryptophan hydroxylase intron 7. Brain Res Mol Brain Res 1997 Apr;**45**(1):145-8.

(269) Mizuno S, Ito H, Hamajima N, Tamakoshi A, Hirose K, Tajima K. Association between smoking habits and tryptophan hydroxylase gene C218A polymorphism among the Japanese population. J Epidemiol 2004 May;**14**(3):94-9.

(270) Lesch KP, Balling U, Gross J, Strauss K, Wolozin BL, Murphy DL, Riederer P. Organization of the human serotonin transporter gene. J Neural Transm Gen Sect 1994;**95**(2):157-62.

(271) Heils A, Teufel A, Petri S, Stober G, Riederer P, Bengel D, Lesch KP. Allelic variation of human serotonin transporter gene expression. J Neurochem 1996 Jun;**66**(6):2621-4.

(272) Peroutka SJ. Serotonin receptor variants in disease: new therapeutic opportunities? Ann N Y Acad Sci 1998 Dec 15;**861**:16-25.

(273) Polesskaya OO, Sokolov BP. Differential expression of the "C" and "T" alleles of the 5-HT2A receptor gene in the temporal cortex of normal individuals and schizophrenics. J Neurosci Res 2002 Mar 15;67(6):812-22.

(274) Pidoplichko VI, DeBiasi M, Williams JT, Dani JA. Nicotine activates and desensitizes midbrain dopamine neurons. Nature 1997 Nov 27;390(6658):401-4.

(275) Mansvelder HD, McGehee DS. Cellular and synaptic mechanisms of nicotine addiction. J Neurobiol 2002 Dec;53(4):606-17.

(276) Flores CM, Rogers SW, Pabreza LA, Wolfe BB, Kellar KJ. A subtype of nicotinic cholinergic receptor in rat brain is composed of alpha 4 and beta 2 subunits and is up-regulated by chronic nicotine treatment. Mol Pharmacol 1992 Jan;41(1):31-7.

(277) Silverman MA, Neale MC, Sullivan PF, Harris-Kerr C, Wormley B, Sadek H, Ma Y, Kendler KS, Straub RE. Haplotypes of four novel single nucleotide polymorphisms in the nicotinic acetylcholine receptor beta2-subunit (CHRNB2) gene show no association with smoking initiation or nicotine dependence. Am J Med Genet 2000 Oct 9;96(5):646-53.

(278) Lueders KK, Hu S, McHugh L, Myakishev MV, Sirota LA, Hamer DH. Genetic and functional analysis of single nucleotide polymorphisms in the beta2-neuronal nicotinic acetylcholine receptor gene (CHRNB2). Nicotine Tob Res 2002 Feb;4(1):115-25.

(279) Feng Y, Niu T, Xing H, Xu X, Chen C, Peng S, Wang L, Laird N, Xu X. A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. Am J Hum Genet 2004 Jul;75(1):112-21.

(280) Li MD, Beuten J, Ma JZ, Payne TJ, Lou XY, Garcia V, Duenes AS, Crews KM, Elston RC. Ethnic- and gender-specific association of the nicotinic acetylcholine receptor alpha4 subunit gene (CHRNA4) with nicotine dependence. Hum Mol Genet 2005 May 1;14(9):1211-9.

(281) Greenbaum L, Kanyas K, Karni O, Merbl Y, Olender T, Horowitz A, Yakir A, Lancet D, Ben-Asher E, Lerer B. Why do young women smoke? I. Direct and interactive effects of environment, psychological characteristics and nicotinic cholinergic receptor genes. Mol Psychiatry 2006 Mar;11(3):312-22, 223.

(282) Barker DJ. The fetal and infant origins of adult disease. BMJ 1990 Nov 17;301(6761):1111.

(283) Barker DJ, Martyn CN. The maternal and fetal origins of cardiovascular disease 7. J Epidemiol Community Health 1992 Feb;46(1):8-11.

(284) Barker DJ. The developmental origins of well-being 4. Philos Trans R Soc Lond B Biol Sci 2004 Sep 29;359(1449):1359-66.

(285) Bateson P, Barker D, Clutton-Brock T, Deb D, D'Udine B, Foley RA, Gluckman P, Godfrey K, Kirkwood T, Lahr MM, McNamara J, Metcalfe NB, Monaghan P,

Spencer HG, Sultan SE. Developmental plasticity and human health. Nature 2004 Jul 22;430(6998):419-21.

(286) Drake AJ, Walker BR. The intergenerational effects of fetal programming: non-genomic mechanisms for the inheritance of low birth weight and cardiovascular risk 3. J Endocrinol 2004 Jan;180(1):1-16.

(287) Gluckman PD, Hanson MA. Developmental origins of disease paradigm: a mechanistic and evolutionary perspective
2. Pediatr Res 2004 Sep;56(3):311-7.

(288) Hales CN, Barker DJ, Clark PM, Cox LJ, Fall C, Osmond C, Winter PD. Fetal and infant growth and impaired glucose tolerance at age 64
3. BMJ 1991 Oct 26;303(6809):1019-22.

(289) Hales CN, Barker DJ. The thrifty phenotype hypothesis
1. Br Med Bull 2001;60:5-20.

(290) Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. N Engl J Med 2005 Oct 27;353(17):1802-9.

(291) Lebenthal E, Bier DM. Novel concepts in the developmental origins of adult health and disease. J Nutr 2007 Apr;137(4):1073-5.

(292) NEEL JV. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?
1. Am J Hum Genet 1962 Dec;14:353-62.

(293) Day IN, King TH, Chen XH, Voropanov AM, Ye S, Syddall HE, Sayer AA, Cooper C, Barker DJ, Phillips DI. Insulin-like growth factor-I genotype and birthweight
2. Lancet 2002 Sep 21;360(9337):945-6.

(294) Day IN, Chen XH, Gaunt TR, King TH, Voropanov A, Ye S, Rodriguez S, Syddall HE, Sayer AA, Dennison EM, Tabassum F, Barker DJ, Cooper C, Phillips DI. Late life metabolic syndrome, early growth, and common polymorphism in the growth hormone and placental lactogen gene cluster
1. J Clin Endocrinol Metab 2004 Nov;89(11):5569-76.

(295) Frayling TM, Hattersley AT, McCarthy A, Holly J, Mitchell SM, Gloyn AL, Owen K, Davies D, Smith GD, Ben Shlomo Y. A putative functional polymorphism in the IGF-I gene: association studies with type 2 diabetes, adult height, glucose tolerance, and fetal growth in U.K. populations
2. Diabetes 2002 Jul;51(7):2313-6.

(296) Hattersley AT, Tooke JE. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease
1. Lancet 1999 May 22;353(9166):1789-92.

(297) Hill DJ, Petrik J, Arany E. Growth factors and the regulation of fetal growth
2. Diabetes Care 1998 Aug;21 Suppl 2:B60-B69.

(298) Lindsay RS, Dabelea D, Roumain J, Hanson RL, Bennett PH, Knowler WC. Type 2 diabetes and low birth weight: the role of paternal inheritance in the association of low birth weight and diabetes
1. Diabetes 2000 Mar;49(3):445-9.

(299) Missmer SA, Haiman CA, Hunter DJ, Willett WC, Colditz GA, Speizer FE, Pollak MN, Hankinson SE. A sequence repeat in the insulin-like growth factor-1 gene and risk of breast cancer
1. Int J Cancer 2002 Jul 20;100(3):332-6.

(300) Ong KK, Dunger DB. Thrifty genotypes and phenotypes in the pathogenesis of type 2 diabetes mellitus
1. J Pediatr Endocrinol Metab 2000;**13 Suppl 6**:1419-24.

(301) Ong KK, Dunger DB. Perinatal growth failure: the road to obesity, insulin resistance and cardiovascular disease in adults
1. Best Pract Res Clin Endocrinol Metab 2002 Jun;**16**(2):191-207.

(302) Ong KK, Dunger DB. Birth weight, infant growth and insulin resistance
1. Eur J Endocrinol 2004 Nov;**151 Suppl 3**:U131-U139.

(303) Reaven GM. Hypothesis: muscle insulin resistance is the ("not-so") thrifty genotype 14. Diabetologia 1998 Apr;**41**(4):482-4.

(304) Vaessen N, Heutink P, Janssen JA, Witteman JC, Testers L, Hofman A, Lamberts SW, Oostra BA, Pols HA, van Duijn CM. A polymorphism in the gene for IGF-I: functional properties and risk for type 2 diabetes and myocardial infarction
1. Diabetes 2001 Mar;**50**(3):637-42.

(305) Vaessen N, Janssen JA, Heutink P, Hofman A, Lamberts SW, Oostra BA, Pols HA, van Duijn CM. Association between genetic variation in the gene for insulin-like growth factor-I and low birthweight
1. Lancet 2002 Mar 23;**359**(9311):1036-7.

(306) Yu H, Li BD, Smith M, Shi R, Berkel HJ, Kato I. Polymorphic CA repeats in the IGF-I gene and breast cancer
1. Breast Cancer Res Treat 2001 Nov;**70**(2):117-22.

(307) Emanuel I. Maternal health during childhood and later reproductive performance. Ann N Y Acad Sci 1986;**477**:27-39.

(308) Drake AJ, Walker BR, Seckl JR. Intergenerational consequences of fetal programming by in utero exposure to glucocorticoids in rats. Am J Physiol Regul Integr Comp Physiol 2005 Jan;**288**(1):R34-R38.

(309) Gluckman PD, Hanson MA, Spencer HG. Predictive adaptive responses and human evolution. Trends Ecol Evol 2005 Oct;**20**(10):527-33.

(310) Gluckman PD, Hanson MA, Morton SM, Pinal CS. Life-long echoes--a critical analysis of the developmental origins of adult disease model. Biol Neonate 2005;**87**(2):127-39.

(311) Gluckman PD, Hanson MA, Beedle AS. Early life events and their consequences for later disease: a life history and evolutionary perspective. Am J Hum Biol 2007 Jan;**19**(1):1-19.

(312) Barker DJ. In utero programming of chronic disease. Clin Sci (Lond) 1998 Aug;**95**(2):115-28.

(313) Godfrey KM, Lillycrop KA, Burdge GC, Gluckman PD, Hanson MA. Epigenetic Mechanisms and the Mismatch Concept of the Developmental Origins of Health and Disease. Pediatr Res 2007 Mar 15.

(314) Frayling TM, Hattersley AT. The role of genetic susceptibility in the association of low birth weight with type 2 diabetes. Br Med Bull 2001;**60**:89-101.

(315) Rich-Edwards JW, Colditz GA, Stampfer MJ, Willett WC, Gillman MW, Hennekens CH, Speizer FE, Manson JE. Birthweight and the risk for type 2 diabetes mellitus in adult women. Ann Intern Med 1999 Feb 16;**130**(4 Pt 1):278-84.

(316) Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 2000 Sep;**26**(1):76-80.

(317) Eriksson JG, Lindi V, Uusitupa M, Forsen TJ, Laakso M, Osmond C, Barker DJ. The effects of the Pro12Ala polymorphism of the peroxisome proliferator-activated receptor-gamma2 gene on insulin sensitivity and insulin metabolism interact with size at birth. Diabetes 2002 Jul;**51**(7):2321-4.

(318) Eriksson J, Lindi V, Uusitupa M, Forsen T, Laakso M, Osmond C, Barker D. The effects of the Pro12Ala polymorphism of the PPARgamma-2 gene on lipid metabolism interact with body size at birth. Clin Genet 2003 Oct;**64**(4):366-70.

(319) Kubaszek A, Markkanen A, Eriksson JG, Forsen T, Osmond C, Barker DJ, Laakso M. The association of the K121Q polymorphism of the plasma cell glycoprotein-1 gene with type 2 diabetes and hypertension depends on size at birth. J Clin Endocrinol Metab 2004 May;**89**(5):2044-7.

(320) Eriksson JG. Gene polymorphisms, size at birth, and the development of hypertension and type 2 diabetes. J Nutr 2007 Apr;**137**(4):1063-5.

(321) Freathy RM, Mitchell SM, Knight B, Shields B, Weedon MN, Hattersley AT, Frayling TM. A study of association between common variation in the growth hormone-chorionic somatomammotropin hormone gene cluster and adult fasting insulin in a UK Caucasian population. J Negat Results Biomed 2006;**5**:18.

(322) Handwerger S, Freemark M. The roles of placental growth hormone and placental lactogen in the regulation of human fetal growth and development. J Pediatr Endocrinol Metab 2000 Apr;**13**(4):343-56.

(323) Lacroix MC, Bolifraud P, Durieux D, Pauloin A, Vidaud M, Kann G. Placental growth hormone and lactogen production by perifused ovine placental explants:

regulation by growth hormone-releasing hormone and glucose. Biol Reprod 2002 Mar;66(3):555-61.

(324) Rosenfeld RG. Insulin-like growth factors and the basis of growth. N Engl J Med 2003 Dec 4;349(23):2184-6.

(325) Fornage M, Amos CI, Kardia S, Sing CF, Turner ST, Boerwinkle E. Variation in the region of the angiotensin-converting enzyme gene influences interindividual differences in blood pressure levels in young white males. Circulation 1998 May 12;97(18):1773-9.

(326) Jeunemaitre X, Lifton RP, Hunt SC, Williams RR, Lalouel JM. Absence of linkage between the angiotensin converting enzyme locus and human essential hypertension. Nat Genet 1992 Apr;1(1):72-5.

(327) O'Donnell CJ, Lindpaintner K, Larson MG, Rao VS, Ordovas JM, Schaefer EJ, Myers RH, Levy D. Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study. Circulation 1998 May 12;97(18):1766-72.

(328) Jaffe CA, Ocampo-Lim B, Guo W, Krueger K, Sugahara I, DeMott-Friberg R, Bermann M, Barkan AL. Regulatory mechanisms of growth hormone secretion are sexually dimorphic. J Clin Invest 1998 Jul 1;102(1):153-64.

(329) Cambien F, Poirier O, Lecerf L, Evans A, Cambou JP, Arveiler D, Luc G, Bard JM, Bara L, Ricard S, . Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. Nature 1992 Oct 15;359(6396):641-4.

(330) Cambien F, Costerousse O, Tiret L, Poirier O, Lecerf L, Gonzales MF, Evans A, Arveiler D, Cambou JP, Luc G, . Plasma level and gene polymorphism of angiotensin-converting enzyme in relation to myocardial infarction. Circulation 1994 Aug;90(2):669-76.

(331) Keavney B, McKenzie C, Parish S, Palmer A, Clark S, Youngman L, Delepine M, Lathrop M, Peto R, Collins R. Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. Lancet 2000 Feb 5;355(9202):434-42.

(332) Nakai K, Itoh C, Miura Y, Hotta K, Musha T, Itoh T, Miyakawa T, Iwasaki R, Hiramori K. Deletion polymorphism of the angiotensin I-converting enzyme gene is associated with serum ACE concentration and increased risk for CAD in the Japanese. Circulation 1994 Nov;90(5):2199-202.

(333) Iwai N, Ohmichi N, Nakamura Y, Kinoshita M. DD genotype of the angiotensin-converting enzyme gene is a risk factor for left ventricular hypertrophy. Circulation 1994 Dec;90(6):2622-8.

(334) Schunkert H, Hense HW, Holmer SR, Stender M, Perz S, Keil U, Lorell BH, Riegger GA. Association between a deletion polymorphism of the angiotensin-converting-

enzyme gene and left ventricular hypertrophy. N Engl J Med 1994 Jun 9;330(23):1634-8.

(335) Woods DR, Humphries SE, Montgomery HE. The ACE I/D polymorphism and human physical performance. Trends Endocrinol Metab 2000 Dec;11(10):416-20.

(336) Diet F, Graf C, Mahnke N, Wassmer G, Predel HG, Palma-Hohmann I, Rost R, Bohm M. ACE and angiotensinogen gene genotypes and left ventricular mass in athletes. Eur J Clin Invest 2001 Oct;31(10):836-42.

(337) Myerson SG, Montgomery HE, Whittingham M, Jubb M, World MJ, Humphries SE, Pennell DJ. Left ventricular hypertrophy with exercise and ACE gene insertion/deletion polymorphism: a randomized controlled trial with losartan. Circulation 2001 Jan 16;103(2):226-30.

(338) Rigat B, Hubert C, Alhenc-Gelas F, Cambien F, Corvol P, Soubrier F. An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. J Clin Invest 1990 Oct;86(4):1343-6.

(339) Cambien F, Leger J, Mallet C, Levy-Marchal C, Collin D, Czernichow P. Angiotensin I-converting enzyme gene polymorphism modulates the consequences of in utero growth retardation on plasma insulin in young adults. Diabetes 1998 Mar;47(3):470-5.

(340) Kajantie E, Rautanen A, Kere J, Andersson S, Yliharsila H, Osmond C, Barker DJ, Forsen T, Eriksson J. The effects of the ACE gene insertion/deletion polymorphism on glucose tolerance and insulin secretion in elderly people are modified by birth weight. J Clin Endocrinol Metab 2004 Nov;89(11):5738-41.

(341) Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F. Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I-converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 1992 Jul;51(1):197-205.

(342) Costerousse O, Allegrini J, Lopez M, Alhenc-Gelas F. Angiotensin I-converting enzyme in human circulating mononuclear cells: genetic polymorphism of expression in T-lymphocytes. Biochem J 1993 Feb 15;290 (Pt 1):33-40.

(343) Danser AH, Schalekamp MA, Bax WA, van den Brink AM, Saxena PR, Riegger GA, Schunkert H. Angiotensin-converting enzyme in the human heart. Effect of the deletion/insertion polymorphism. Circulation 1995 Sep 15;92(6):1387-8.

(344) 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 May 31;322(22):1561-6.

(345) Montgomery HE, Clarkson P, Dollery CM, Prasad K, Losi MA, Hemingway H, Statters D, Jubb M, Girvain M, Varnava A, World M, Deanfield J, Talmud P, McEwan JR, McKenna WJ, Humphries S. Association of angiotensin-converting

enzyme gene I/D polymorphism with change in left ventricular mass in response to physical training. Circulation 1997 Aug 5;**96**(3):741-7.

(346) DeMaria AN, Neumann A, Lee G, Fowler W, Mason DT. Alterations in ventricular mass and performance induced by exercise training in man evaluated by echocardiography. Circulation 1978 Feb;**57**(2):237-44.

(347) Kanakis C, Hickson RC. Left ventricular responses to a program of lower-limb strength training. Chest 1980 Oct;**78**(4):618-21.

(348) Montgomery H, Brull D, Humphries SE. Analysis of gene-environment interactions by "stressing-the-genotype" studies: the angiotensin converting enzyme and exercise-induced left ventricular hypertrophy as an example. Ital Heart J 2002 Jan;**3**(1):10-4.

(349) Cox R, Bouzekri N, Martin S, Southam L, Hugill A, Golamaully M, Cooper R, Adeyemo A, Soubrier F, Ward R, Lathrop GM, Matsuda F, Farrall M. Angiotensin-1-converting enzyme (ACE) plasma concentration is influenced by multiple ACE-linked quantitative trait nucleotides. Hum Mol Genet 2002 Nov 1;**11**(23):2969-77.

(350) Montgomery H, Clarkson P, Barnard M, Bell J, Brynes A, Dollery C, Hajnal J, Hemingway H, Mercer D, Jarman P, Marshall R, Prasad K, Rayson M, Saeed N, Talmud P, Thomas L, Jubb M, World M, Humphries S. Angiotensin-converting-enzyme gene insertion/deletion polymorphism and response to physical training. Lancet 1999 Feb 13;**353**(9152):541-5.

(351) Whincup PH, Gilg JA, Papacosta O, Seymour C, Miller GJ, Alberti KG, Cook DG. Early evidence of ethnic differences in cardiovascular risk: cross sectional comparison of British South Asian and white children. BMJ 2002 Mar 16;**324**(7338):635.

(352) Whincup PH, Cook DG, Adshead F, Taylor S, Papacosta O, Walker M, Wilson V. Cardiovascular risk factors in British children from towns with widely differing adult cardiovascular mortality. BMJ 1996 Jul 13;**313**(7049):79-84.

(353) Feyerabend C, Russell MA. A rapid gas-liquid chromatographic method for the determination of cotinine and nicotine in biological fluids. J Pharm Pharmacol 1990 Jun;**42**(6):450-2.

(354) Jarvis L. *Smoking among Secondary School Children in 1996: England.* London: The Stationery Office, London; 1997.

(355) McNeill AD, Jarvis MJ, West R, Russell MA, Bryant A. Saliva cotinine as an indicator of cigarette smoking in adolescents. Br J Addict 1987 Dec;**82**(12):1355-60.

(356) Owen CG, Whincup PH, Odoki K, Gilg JA, Cook DG. Birth weight and blood cholesterol level: a study in adolescents and systematic review. Pediatrics 2003 May;**111**(5 Pt 1):1081-9.

(357) Unger T, Schull B, Rascher W, Lang RE, Ganten D. Selective activation of the converting enzyme inhibitor MK 421 and comparison of its active diacid form with

captopril in different tissues of the rat. Biochem Pharmacol 1982 Oct 1;31(19):3063-70.

(358) Maceira AM, Prasad SK, Khan M, Pennell DJ. Normalized left ventricular systolic and diastolic function by steady state free precession cardiovascular magnetic resonance. J Cardiovasc Magn Reson 2006;8(3):417-26.

(359) Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988 Feb 11;16(3):1215.

(360) Polymeropoulos MH, Xiao H, Rath DS, Merril CR. Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH). Nucleic Acids Res 1991 Jul 11;19(13):3753.

(361) Chen XH, O'Dell SD, Day IN. Microplate array diagonal gel electrophoresis for cohort studies of microsatellite loci. Biotechniques 2002 May;32(5):1080-2, 1084, 1086.

(362) Ye S, Dhillon S, Ke X, Collins AR, Day IN. An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic Acids Res 2001 Sep 1;29(17):E88.

(363) Jonsson EG, Norton N, Forslund K, Mattila-Evenden M, Rylander G, Asberg M, Owen MJ, Sedvall GC. Association between a promoter variant in the monoamine oxidase A gene and schizophrenia. Schizophr Res 2003 May 1;61(1):31-7.

(364) Lichter JB, Barr CL, Kennedy JL, Van Tol HH, Kidd KK, Livak KJ. A hypervariable segment in the human dopamine receptor D4 (DRD4) gene. Hum Mol Genet 1993 Jun;2(6):767-73.

(365) Cheung VG, Nelson SF. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc Natl Acad Sci U S A 1996 Dec 10;93(25):14676-9.

(366) Evans AE, Poirier O, Kee F, Lecerf L, McCrum E, Falconer T, Crane J, O'Rourke DF, Cambien F. Polymorphisms of the angiotensin-converting-enzyme gene in subjects who die from coronary heart disease. Q J Med 1994 Apr;87(4):211-4.

(367) Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002;53(2):79-91.

(368) Lenth RV. Statistical power calculations. J Anim Sci 2007 Mar;85(13 Suppl):E24-E29.

(369) Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics 2005 Sep;6(6):639-46.

(370) Godard B, Schmidtke J, Cassiman JJ, Ayme S. Data storage and DNA banking for biomedical research: informed consent, confidentiality, quality issues, ownership,

return of benefits. A professional perspective. Eur J Hum Genet 2003 Dec;**11 Suppl 2**:S88-122.

(371) Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 2004 Aug;**36**(8):861-6.

(372) Sabol SZ, Hu S, Hamer D. A functional polymorphism in the monoamine oxidase A gene promoter. Hum Genet 1998 Sep;**103**(3):273-9.

(373) Das M, Bhowmik AD, Sinha S, Chattopadhyay A, Chaudhuri K, Singh M, Mukhopadhyay K. MAOA promoter polymorphism and attention deficit hyperactivity disorder (ADHD) in indian children. Am J Med Genet B Neuropsychiatr Genet 2006 Sep 5;**141**(6):637-42.

(374) Marshall SE, Bird TG, Hart K, Welsh KI. Unified approach to the analysis of genetic variation in serotonergic pathways. Am J Med Genet 1999 Dec 15;**88**(6):621-7.

(375) Todd RD, Neuman RJ, Lobos EA, Jong YJ, Reich W, Heath AC. Lack of association of dopamine D4 receptor gene polymorphisms with ADHD subtypes in a population sample of twins. Am J Med Genet 2001 Jul 8;**105**(5):432-8.

(376) Hall W, Madden P, Lynskey M. The genetics of tobacco use: methods, findings and policy implications. Tob Control 2002 Jun;**11**(2):119-24.

(377) Ujjin P, Satarug S, Vanavanitkun Y, Daigo S, Ariyoshi N, Yamazaki H, Reilly PE, Moore MR, Kamataki T. Variation in coumarin 7-hydroxylase activity associated with genetic polymorphism of cytochrome P450 2A6 and the body status of iron stores in adult Thai males and females. Pharmacogenetics 2002 Apr;**12**(3):241-9.

(378) Vasconcelos GM, Struchiner CJ, Suarez-Kurtz G. CYP2A6 genetic polymorphisms and correlation with smoking status in Brazilians. Pharmacogenomics J 2005;**5**(1):42-8.

(379) Gambier N, Batt AM, Marie B, Pfister M, Siest G, Visvikis-Siest S. Association of CYP2A6*1B genetic variant with the amount of smoking in French adults from the Stanislas cohort. Pharmacogenomics J 2005;**5**(4):271-5.

(380) O'Loughlin J, Paradis G, Kim W, DiFranza J, Meshefedjian G, Millan-Davey E, Wong S, Hanley J, Tyndale RF. Genetically decreased CYP2A6 and the risk of tobacco dependence: a prospective study of novice smokers. Tob Control 2004 Dec;**13**(4):422-8.

(381) Terayama H, Itoh M, Fukunishi I, Iwahashi K. The serotonin-2A receptor polymorphism and smoking behavior in Japan. Psychiatr Genet 2004 Dec;**14**(4):195-7.

(382) Xu C, Goodz S, Sellers EM, Tyndale RF. CYP2A6 genetic variation and potential consequences. Adv Drug Deliv Rev 2002 Nov 18;**54**(10):1245-56.

(383) Yamanaka H, Nakajima M, Nishimura K, Yoshida R, Fukami T, Katoh M, Yokoi T. Metabolic profile of nicotine in subjects whose CYP2A6 gene is deleted. Eur J Pharm Sci 2004 Aug;22(5):419-25.

(384) Benowitz NL, Pomerleau OF, Pomerleau CS, Jacob P, III. Nicotine metabolite ratio as a predictor of cigarette consumption. Nicotine Tob Res 2003 Oct;5(5):621-4.

(385) Yoshida R, Nakajima M, Nishimura K, Tokudome S, Kwon JT, Yokoi T. Effects of polymorphism in promoter region of human CYP2A6 gene (CYP2A6*9) on expression level of messenger ribonucleic acid and enzymatic activity in vivo and in vitro. Clin Pharmacol Ther 2003 Jul;74(1):69-76.

(386) Fielding JE. Smoking and pregnancy. N Engl J Med 1978 Feb 9;298(6):337-9.

(387) Lindpaintner K, Lee M, Larson MG, Rao VS, Pfeffer MA, Ordovas JM, Schaefer EJ, Wilson AF, Wilson PW, Vasan RS, Myers RH, Levy D. Absence of association or genetic linkage between the angiotensin-converting-enzyme gene and left ventricular mass. N Engl J Med 1996 Apr 18;334(16):1023-8.

(388) Kupari M, Perola M, Koskinen P, Virolainen J, Karhunen PJ. Left ventricular size, mass, and function in relation to angiotensin-converting enzyme gene polymorphism in humans. Am J Physiol 1994 Sep;267(3 Pt 2):H1107-H1111.

(389) Wu S, Hong J, Li H, Feng S, Qiao D, Gao M, Shi J, Nie H. No correlation of polymorphism of angiotensin-converting enzyme genes with left ventricular hypertrophy in essential hypertension. Hypertens Res 2000 May;23(3):261-4.

(390) Lopez-Contreras J, Blanco-Vaca F, Borras X, Carreras F, Pons-Llado G, Gallego F, Sole MJ, Cirera S, Benet MT, Negredo E, Roca-Cusachs A. Usefulness of the I/D angiotensin-converting enzyme genotype for detecting the risk of left ventricular hypertrophy in pharmacologically treated hypertensive men. J Hum Hypertens 2000 May;14(5):327-31.

(391) Kuznetsova T, Staessen JA, Thijs L, Kunath C, Olszanecka A, Ryabikov A, Tikhonoff V, Stolarz K, Bianchi G, Casiglia E, Fagard R, Brand-Herrmann SM, Kawecka-Jaszcz K, Malyutina S, Nikitin Y, Brand E. Left ventricular mass in relation to genetic variation in angiotensin II receptors, renin system genes, and sodium excretion. Circulation 2004 Oct 26;110(17):2644-50.

(392) Kuznetsova T, Staessen JA, Stolarz K, Ryabikov A, Tikhonoff V, Olszanecka A, Bianchi G, Brand E, Casiglia E, Dominiczak A, Fagard R, Malyutina S, Nikitin Y, Kawecka-Jaszcz K. Relationship between left ventricular mass and the ACE D/I polymorphism varies according to sodium intake. J Hypertens 2004 Feb;22(2):287-95.

(393) Maheshwari VD, Pillai A. Influence of smoking and hypertension on left ventricular mass. J Assoc Physicians India 2000 Apr;48(4):397-9.

(394) Silverman EK, Palmer LJ. Case-control association studies for the genetics of complex respiratory diseases. Am J Respir Cell Mol Biol 2000 Jun;22(6):645-8.

(395) Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993 Mar;52(3):506-16.

(396) Schaid DJ. Transmission disequilibrium, family controls, and great expectations. Am J Hum Genet 1998 Oct;63(4):935-41.

(397) Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999 Jul;65(1):220-8.

(398) Dupont B, Oberfield SE, Smithwick EM, Lee TD, Levine LS. Close genetic linkage between HLA and congenital adrenal hyperplasia (21-hydroxylase deficiency). Lancet 1977 Dec 24;2(8052-8053):1309-12.

(399) Pollack MS, Levine LS, O'Neill GJ, Pang S, Lorenzen F, Kohn B, Rondanini GF, Chiumello G, New MI, Dupont B. HLA linkage and B14, DR1, BfS haplotype association with the genes for late onset and cryptic 21-hydroxylase deficiency. Am J Hum Genet 1981 Jul;33(4):540-50.

(400) Fields S. The future is function. Nat Genet 1997 Apr;15(4):325-7.

(401) Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, bers-Akkers MT, Godino-Ivan MJ, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007 Apr 1.

(402) Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF, Jr., Hoover R, Hunter DJ, Chanock SJ, Thomas G. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007 Apr 1.

(403) Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007 Apr 15.

(404) Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De L, V, Briggs J, Gunther S, Prescott NJ, Onnie CM, Hasler R, Sipos B, Folsch UR, Lengauer T, Platzer M, Mathew CG, Krawczak M, Schreiber S. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007 Feb;39(2):207-11.

(405) Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007 Feb 22;445(7130):881-5.

(406) Marshall E. Genetics. Sequencers of a famous genome confront privacy issues. Science 2007 Mar 30;315(5820):1780.

(407) Ylstra B, van d, I, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 2006;34(2):445-50.

(408) Rajcan-Separovic E, Harvard C, Liu X, McGillivray B, Hall JG, Qiao Y, Hurlburt J, Hildebrand J, Mickelson EC, Holden JJ, Lewis ME. Clinical and molecular cytogenetic characterisation of a newly recognised microdeletion syndrome involving 2p15-16.1. J Med Genet 2007 Apr;44(4):269-76.

(409) Paris PL, Andaya A, Fridlyand J, Jain AN, Weinberg V, Kowbel D, Brebner JH, Simko J, Watson JE, Volik S, Albertson DG, Pinkel D, Alers JC, van der Kwast TH, Vissers KJ, Schroder FH, Wildhagen MF, Febbo PG, Chinnaiyan AM, Pienta KJ, Carroll PR, Rubin MA, Collins C, van DH. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. Hum Mol Genet 2004 Jul 1;13(13):1303-13.

(410) Greshock J, Naylor TL, Margolin A, Diskin S, Cleaver SH, Futreal PA, deJong PJ, Zhao S, Liebman M, Weber BL. 1-Mb resolution array-based comparative genomic hybridization using a BAC clone set optimized for cancer gene analysis. Genome Res 2004 Jan;14(1):179-87.

(411) Schoumans J, Anderlid BM, Blennow E, Teh BT, Nordenskjold M. The performance of CGH array for the detection of cryptic constitutional chromosome imbalances. J Med Genet 2004 Mar;41(3):198-202.

(412) Weiss MM, Kuipers EJ, Postma C, Snijders AM, Siccama I, Pinkel D, Westerga J, Meuwissen SG, Albertson DG, Meijer GA. Genomic profiling of gastric cancer predicts lymph node status and survival. Oncogene 2003 Mar 27;22(12):1872-9.

(413) Weiss MM, Kuipers EJ, Meuwissen SG, van Diest PJ, Meijer GA. Comparative genomic hybridisation as a supportive tool in diagnostic pathology. J Clin Pathol 2003 Jul;56(7):522-7.

(414) Trask BJ. Human cytogenetics: 46 chromosomes, 46 years and counting. Nat Rev Genet 2002 Oct;3(10):769-78.

(415) Wilson GM, Flibotte S, Chopra V, Melnyk BL, Honer WG, Holt RA. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. Hum Mol Genet 2006 Mar 1;15(5):743-9.

(416) Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks

J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong Association of De Novo Copy Number Mutations with Autism. Science 2007 Mar 15.

(417) Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet 2003 Oct 15;**12 Spec No 2**:R145-R152.

(418) Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992 Oct 30;**258**(5083):818-21.

(419) Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet 2004 Dec;**66**(6):488-95.

(420) Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 2001 Nov;**29**(3):263-4.

(421) Vissers LE, de Vries BB, Osoegawa K, Janssen IM, Feuth T, Choy CO, Straatman H, van d, V, Huys EH, van RA, Smeets D, van Ravenswaaij-Arts CM, Knoers NV, van dB, I, de Jong PJ, Brunner HG, van Kessel AG, Schoenmakers EF, Veltman JA. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. Am J Hum Genet 2003 Dec;73(6):1261-70.

(422) Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL. A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 2004 Mar;**36**(3):299-303.

(423) Vissers LE, Veltman JA, van Kessel AG, Brunner HG. Identification of disease genes by whole genome CGH arrays. Hum Mol Genet 2005 Oct 15;**14 Spec No. 2**:R215-R223.

(424) Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. Nat Methods 2006 Jan;3(1):31-3.

(425) Howell WM, Jobs M, Gyllensten U, Brookes AJ. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. Nat Biotechnol 1999 Jan;**17**(1):87-8.

(426) Jobs M, Howell WM, Stromqvist L, Mayr T, Brookes AJ. DASH-2: flexible, low-cost, and high-throughput SNP genotyping by dynamic allele-specific hybridization on membrane arrays. Genome Res 2003 May;**13**(5):916-24.

(427) Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-

Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, vey-Smith G, Hattersley AT, McCarthy MI. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science 2007 Apr 12.

(428) Schaid DJ. Likelihoods and TDT for the case-parents design. Genet Epidemiol 1999;16(3):250-60.

(429) Newton Morton, Nikolas Maniatis, Weihua Zhang, Sarah Ennis, Andrew Collins. Genome Scanning by Composite Likelihood. 80, 19-28. 2007.

(430) Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, Crews C, Bruijn L, Rothstein J, Mora G, Restagno G, Chio A, Singleton A, Hardy J, Traynor BJ. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 2007 Apr;6(4):322-8.

(431) Gribble SM, Prigmore E, Burford DC, Porter KM, Ng BL, Douglas EJ, Fiegler H, Carr P, Kalaitzopoulos D, Clegg S, Sandstrom R, Temple IK, Youings SA, Thomas NS, Dennis NR, Jacobs PA, Crolla JA, Carter NP. The complex nature of constitutional de novo apparently balanced translocations in patients presenting with abnormal phenotypes. J Med Genet. 2005 Jan;42(1):8-16.

(432) Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. Science. 2001 Feb 16;291(5507):1224-9.