Discovering Common Genetic Variants for Hypertension Using an Extreme Case-control Strategy

Claire E. Hastie, M.Sc.

This being a thesis submitted for the degree of Doctor of Philosophy (Ph.D.) in the Faculty of Medicine, University of Glasgow
October 2010

BHF Glasgow Cardiovascular Research Centre
Institute of Cardiovascular and Medical Sciences
College of Medical, Veterinary, and Life Sciences
University of Glasgow

Declaration

I declare that this thesis has been written entirely by myself and is a record of research performed by myself with the exception of discovery cohort genotyping (Dr Wai K. Lee, Dr Anna Maria Di Blasio, Stewart Laing, and Dr Davide Gentilini), genotyping and association analysis of replication cohorts (undertaken by investigators from each cohort, respectively), and analysis of data from the BRIGHT (Dr Sandosh Padmanabhan), GRECO and HERCULES clinical functional cohorts (undertaken by investigators from each cohort, respectively). This work has not been submitted previously for a higher degree and was carried out under the supervision of Professor Anna F. Dominiczak and Professor Jill P. Pell.

Claire E. Hastie

Acknowledgements

Firstly, I would like to thank my supervisors, Professor Anna Dominiczak and Professor Jill Pell, for their support and guidance. They are both wonderful role models and have provided me with many exciting opportunities over the last four years.

Dr Sandosh Padmanabhan, although advisor in name, has always been much more than that. I would not have been able to complete this project without extensive training and advice from him.

I also wish to thank Dr Wai Kwong Lee for his patience in helping me to understand the genotyping process.

Cristina Menni has been a great source of practical and emotional help through the most difficult stages of this project.

Special thanks to all of my colleagues who selflessly gave up their time to view cluster plots; Adyani Md Redzuan, Aiste Monkeviciute, Angela Bradshaw, Annika Delles, Alan Parker, Caline Koh-Tan, Carolyn Haggerty, Chiara Taurino, David Carty, Elisabeth Beattie, Emily Ord, Fernando Martinez Garcia, Jenny Greig, Jennifer McLachlan, Jim Mcculloch, John McClure, Kirsten Douglas, Laura Graham, Laura Denby, Laura Paul, Lorraine Work, Ruth Mackenzie, Samantha Alvarez-Madrazo, Sandra MacDonald, Teresa, Ulf Neisius, Wendy Crawford, Weiling Sarah Li.

Finally, I cannot express enough my gratitude towards my parents, brother, and husband for their constant encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations, Acronyms & Symbols

A1	major allele
ACADSB	acyl-CoA dehydrogenase, short/branched chain
ACSM1	acyl-CoA synthetase medium-chain family member 1
ACSM2A	acyl-CoA synthetase medium-chain family member 2A
ACSM2B	acyl-CoA synthetase medium-chain family member 2B
ACSM3	acyl-CoA synthetase medium-chain family member 3
ACSM5	acyl-CoA synthetase medium-chain family member 5
ADAMTSL1	ADAMTS-like 1
AIDS	acquired immunodeficiency syndrome
AKAP11	A kinase (PRKA) anchor protein 11
ALDH2	aldehyde dehydrogenase 2
APOE4	apolipoprotein E(*4)
ASCOT	Anglo-Scandinavian Cardiac Outcomes Trial
ASP	affected sibling pair
ASW	African ancestry individuals in the southwestern USA
ATP2B1	ATPase, Ca++ transporting, plasma membrane 1
ATP6V1A	ATPase, H+ transporting, lysosomal 70kDa, V1 subunit A
ATP8B1	ATPase, aminophospholipid transporter, class I, type 8B, member 1
AUC	area under the curve
B58C	British 1958 Birth Cohort
BANK1	B-cell scaffold protein with ankyrin repeats 1
BFSP2	beaded filament structural protein 2, phakinin
BMI	body mass index
BP	blood pressure
BRIGHT	British Genetics of Hypertension study
BSA	body surface area
C10orf88	chromosome 10 open reading frame 88
C13orf30	chromosome 13 open reading frame 30
C16orf72	chromosome 16 open reading frame 72
C19orf61	chromosome 19 open reading frame 61
C3orf36	chromosome 3 open reading frame 36
C5orf42	chromosome 5 open reading frame 42
C9orf150	chromosome 9 open reading frame 150
CAD	coronary artery disease
CARKD	carbohydrate kinase domain containing
CARS2	cysteinyl-tRNA synthetase 2, mitochondrial (putative)
CASR	calcium-sensing receptor
CD163L1	CD163 molecule-like 1
CDCV	common disease-common variant
CDV3	CDV3 homolog (mouse)
CEP57	centrosomal protein 57kDa
CEU	Caucasians in Utah, USA
CHARGE	Cohorts for Heart and Aging Research in Genome Epidemiology Consortium
CHB	Han Chinese in Beijing, China
CHD	coronary heart disease
CHD	metropolitan Chinese in Colorado, USA
CHD2	chromodomain helicase DNA binding protein 2
CHR	chromosome

CI	confidence interval
CKD	chronic kidney disease
CLNK	cytokine-dependent hematopoietic cell linker
CMTM6	CKLF-like MARVEL transmembrane domain containing 6
CMTM7	CKLF-like MARVEL transmembrane domain containing 7
CMTM8	CKLF-like MARVEL transmembrane domain containing 8
CNOT10	CCR4-NOT transcription complex, subunit 10
CNV	copy number variant
COL4A1	collagen, type IV, alpha 1
COL4A2	collagen, type IV, alpha 2
CoLaus	Cohorte Lausannoise
COX-2	cyclooxygenase-2
CV	cardiovascular
CVD	cardiovascular disease
CYP17A1	cytochrome P450, family 17, subfamily A, polypeptide 1
D'	unit of measurement of linkage disequilibrium
DALY	disability adjusted life year
dbGaP	database of Genotypes and Phenotypes
DBP	diastolic blood pressure
DGI	Diabetes Genetics Initiative
DGKH	diacylglycerol kinase, eta
DNA	deoxyribonucleic acid
DPP4	dipeptidyl-peptidase 4
DTC	direct-to-consumer
DYNC1LI1	dynein, cytoplasmic 1, light intermediate chain 1
DZ	dizygotic
EBAG9	estrogen receptor binding site associated, antigen, 9
ECV	extracellular fluid volume
eGFR	estimated glomerular filtration rate
ENY2	enhancer of yellow 2 homolog (Drosophila)
EPSTI1	epithelial stromal interaction 1 (breast)
eQTL	expression quantitative trait locus
ERI2	ERI1 exoribonuclease family member 2
ERPF	effective renal plasma flow
eSNP	expression single nucleotide polymorphism
ETHE1	ethylmalonic encephalopathy 1
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)
ETV6	ets variant 6
EUROSPAN	European Special Populations Research Network
FABP3P2	fatty acid binding protein 3, pseudogene 2
FAM154A	family with sequence similarity 154, member A
FAM24A	family with sequence similarity 24, member A
FAM76B	family with sequence similarity 76, member B
FAP	fibroblast activation protein, alpha"
FECH	ferrochelatase
FENa	fractional excretion of sodium
Fenland Study	Fenland Study
FGF5	fibroblast growth factor 5
FHS	Framingham Heart Study
FJHN	familial juvenile hyperuricaemic nephropathy
FLJ20581	acyl-CoA synthetase medium-chain family member 5
FPRP	false positive report probability
FTO	fat mass and obesity associated

GC	genomic control
GCA	grancalcin, EF-hand calcium binding protein
GCG	glucagon
GFR	glomerular filtration rate
GIH	Gujarati Indians in Texas, USA
Global BPgen	Global Blood Pressure Genetics Consortium
GOLSYN	syntabulin (syntaxin-interacting)
GP2	glycoprotein 2 (zymogen granule membrane)
GPD1L	glycerol-3-phosphate dehydrogenase 1-like
GPR139	G protein-coupled receptor 139
GPRC5B	G protein-coupled receptor, family C, group 5, member B
GRECO	Groningen Renal Hemodynamic Cohort Study Group
GRIN2A	glutamate receptor, ionotropic, N-methyl D-aspartate 2A
GS:SFHS	Generation Scotland:Scottish Family Health Study
GTEx	Genotype-Tissue Expression
GWAS	genome-wide association study
h^2	heritability
HAUS6	HAUS augmin-like complex, subunit 6
HERCULES	Hypertension Evaluation by Remler and CalciUria LEvel Study
HIV	human immunodeficiency virus
HLA	human leukocyte antigen system
HS	high sodium
HS3ST1	heparan sulfate (glucosamine) 3-O-sulfotransferase 1
HTN	hypertension
IBD	identical by descent
IBS	identical by state
ICBP-GWAS	International Consortium for Blood Pressure-Genome-Wide Association Study
IFIH1	interferon induced with helicase C domain 1
IHD	ischaemic heart disease
IKZF5	IKAROS family zinc finger 5
IQCK	IQ motif containing K
IRGC	immunity-related GTPase family, cinema
IRGM	immunity-related GTPase family, M
IRGQ	immunity-related GTPase family, Q
IRS2	insulin receptor substrate 2
ITPA	inosine triphosphatase
JPT	Japanese in Tokyo, Japan
kb	kilobase
KCNH7	potassium voltage-gated channel, subfamily H (eag-related), member 7
KCNJ1	potassium inwardly-rectifying channel, subfamily J, member 1
KCNN4	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4
KIAA0564	KIAA0564
KIAA2018	KIAA2018
KORA	Kooperative Gesundheitsforschung in der Region Augsburg
LD	linkage disequilibrium
LDL	low-density lipoprotein
LEKR1	leucine, glutamate and lysine rich 1
LN	lupus nephritis
LOC123876	acyl-CoA synthetase medium-chain family member 2A
LOC399815	chromosome 10 open reading frame 88 pseudogene
LS	low sodium

LWK	Luhya in Webuye, Kenya
LYPD3	LY6/PLAUR domain containing 3
LYPD5	LY6/PLAUR domain containing 5
MAF	minor allele frequency
Mb	megabase
MC4R	melanocortin 4 receptor
MCKD2	medullary cystic kidney disease 2
MDC	Malmö Diet and Cancer Study
MDRD	Modification of Diet in Renal Disease
MDS	multidimensional scaling
MEX	Mexican origin individuals in California, USA
MI	myocardial infarction
MIGen	Myocardial Infarction Genetics Consortium
MIM	Mendelian Inheritance in Man
MIR122	microRNA 122
MIR572	microRNA 572
MKK	Maasai in Kinyawa, Kenya
MONICA	World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project
MPP	Malmö Preventive Project
mRNA	messenger ribonucleic acid
MTMR2	myotubularin related protein 2
MZ	monozygotic
NAA50	N-alpha-acetyltransferase 50, NatE catalytic subunit
NARS	asparaginyl-tRNA synthetase
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like
NESDA	Netherlands Study of Depression and Anxiety
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
NINDS	National Institute of Neurological Disorders and Stroke
NIPBL	Nipped-B homolog (Drosophila)
NORDIL	Nordic Diltiazem study
NPPA	natriuretic peptide precursor A
NR3C2	nuclear receptor subfamily 3, group C, member 2
NUDCD1	NudC domain containing 1
NUP155	nucleoporin 155kDa
OMIM	Online Mendelian Inheritance in Man
ONECUT2	one cut homeobox 2
OR	odds ratio
OSBPL10	oxysterol binding protein-like 10
PAMELA	Pressioni Arteriose Monitorate e Loro Associazioni
PC	principal component
PCA	principal components analysis
PDILT	protein disulfide isomerase-like, testis expressed
PHLDB3	pleckstrin homology-like domain, family B, member 3
PKHD1L1	polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1
PLCB1	phospholipase C, beta 1 (phosphoinositide-specific)
PLIN2	perilipin 2
POLS	polymerase (DNA-directed) sigma
PPP3CA	protein phosphatase 3, catalytic subunit, alpha isozyme
PRA	plasma renin activity

PREVEND	Prevention of REnal and Vascular ENd stage Disease
PSMD14	proteasome (prosome, macropain) 26S subunit, non-ATPase, 14
PSTK	phosphoseryl-tRNA kinase
QC	quality control
QTL	quantitative trait locus
r^2	unit of measurement of linkage disequilibrium
RAB20	RAB20, member RAS oncogene family
RAB6B	RAB6B, member RAS oncogene family
RGMA	RGM domain family, member A
RNA	ribonucleic acid
RRAGA	Ras-related GTP binding A
RYK	RYK receptor-like tyrosine kinase
S100Z	S100 calcium binding protein Z
SA	structure assessment
SBP	systolic blood pressure
SCNN1B	sodium channel, nonvoltage-gated 1, beta
SCNN1G	sodium channel, nonvoltage-gated 1, gamma
SH2B3	SH2B adaptor protein 3
SH3TC2	SH3 domain and tetratricopeptide repeats 2
SHIP	Study of Health in Pomerania
SHISA9	shisa homolog 0 (Xenopus laevis)
SLC12A1	solute carrier family 12 (sodium/potassium/chloride transporters), member 1
SLC12A3	solute carrier family 12 (sodium/chloride transporters), member 3
SLC1A3	solute carrier family 1 (glial high affinity glutamate transporter), member 3
SLC4A10	solute carrier family 4, sodium bicarbonate transporter, member 10
SLCO2A1	solute carrier organic anion transporter family, member 2A1
SMOC2	SPARC related modular calcium binding 2
SNP	single nucleotide polymorphism
SRPRB	signal recognition particle receptor, B subunit
SRRM5	serine/arginine repetitive matrix 5
SUPT3H	suppressor of Ty 3 homolog (S. cerevisiae)
SWE	Swedish (discovery) sample
TAL	thick ascending limb of the loop of Henle
TBR1	T-box, brain, 1
TDT	transmission disequilibrium test
TF	transferrin
TG	targeted genotyping
THBS2	thrombospondin 2
THUMPD1	THUMP domain containing 1
TMEM108	transmembrane protein 108
TMTC2	transmembrane and tetratricopeptide repeat containing 2
TNFSF11	tumour necrosis factor (ligand) superfamily, member 11
TOPBP1	topoisomerase (DNA) II binding protein 1
TRHR	thyrotropin-releasing hormone receptor
TRIM71	tripartite motif-containing 71
TSHZ2	teashirt zinc finger homeobox 2
TSI	Tuscans in Italy
TSPAN8	tetraspanin 8
TwinsUK	TwinsUK
TYRP1	tyrosinase-related protein 1
UMOD	uromodulin

UMOD	uromodulin
UTI	urinary tract infection
WDR27	WD repeat domain 27
WNK1	WNK lysine deficient protein kinase 1
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan, Nigeria
ZFP112	zinc finger protein 112 homolog (mouse)
ZNF155	zinc finger protein 155
ZNF221	zinc finger protein 221
ZNF222	zinc finger protein 222
ZNF223	zinc finger protein 223
ZNF224	zinc finger protein 224
ZNF225	zinc finger protein 225
ZNF226	zinc finger protein 226
ZNF227	zinc finger protein 227
ZNF229	zinc finger protein 229
ZNF230	zinc finger protein 230
ZNF233	zinc finger protein 233
ZNF234	zinc finger protein 234
ZNF235	zinc finger protein 235
ZNF283	zinc finger protein 283
ZNF285A	zinc finger protein 285A
ZNF404	zinc finger protein 404
ZNF428	zinc finger protein 428
ZNF45	zinc finger protein 45
ZNF536	zinc finger protein 536
ZNF860	zinc finger protein 860
λs	sibling recurrent risk

Publications and Oral Presentations

Publications

Padmanabhan S, Melander O, Johnson T, Di Blasio AM, Lee WK, Gentilini D, **Hastie CE**, Menni C, Monti MC, Delles C, Laing S, Corso B, Navis G, Kwakernaak A, van der Harst P, Bochud M, Maillard M, Burnier M, Hedner T, Kjeldsen S, Wahlstrand B, Sjögren M, Fava C, Montagnana M, Danese E, Torffvit O, Hedblad B, Snieder H, Connell JMC, Brown M, Samani NJ, Farrall M, Cesana G, Mancia G, Signorini S, Grassi G, Eyheramendy S, Wichmann HE, Laan M, Strachan DP, Sever P, Shields DC, Stanton A, Vollenweider P, Teumer A, Völzke H, Rettig R, Newton-Cheh C, Arora P, Zhang F, Soranzo N, Spector TD, Lucas G, Kathiresan S, Siscovick DS, Luan J, Loos RJF, Wareham NJ, Penninx BW, Nolte IM, McBride M, Miller WH, Nicklin SA, Baker AH, Graham D, McDonald RA, Pell JP, Sattar N, Welsh P, Munroe P, Caulfield MJ, Zanchetti A, Dominiczak AF. Genome-wide association study of blood pressure extremes identifies variant near *UMOD* associated with hypertension, *PLoS Genetics*, 2010; 6(10): e1001177.

Padmanabhan S, **Hastie C**, Prabhakaran D, Dominiczak AF. Genomic approaches to coronary artery disease, *Indian Journal of Medical Research*, 2010; 132: 567-78.

Paul L, **Hastie CE**, Li WS, Harrow C, Muir S, Connell JM, Dominiczak AF, McInnes GT, Padmanabhan S. Resting heart rate pattern during follow-up and mortality in hypertensive patients, *Hypertension*, 2010; 55(2): 567-74.

Hastie CE, Padmanabhan P, Dominiczak AF. Genome-Wide Association Studies of Hypertension: Light at the End of the Tunnel, *International Journal of Hypertension*, 2010.

Hastie CE, Padmanabhan S, Slack R, Pell AC, Oldroyd KG, Flapan AD, Jennings KP, Irving J, Eteiba H, Dominiczak AF, Pell JP. Obesity paradox in a cohort of 4880 consecutive patients undergoing percutaneous coronary intervention, *European Heart Journal*, 2010; 31(2): 222-6.

Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, Gaunt TR, Pallas J, Lovering R, Li K, Casas JP, Sofat R, Kumari M, Rodriguez S, Johnson T, Newhouse SJ, Dominiczak A, Samani NJ, Caulfield M, Sever P, Stanton A, Shields DC, Padmanabhan S, Melander O, **Hastie C**, Delles C, Ebrahim S, Marmot MG, Smith GD, Lawlor DA,

Munroe PB, Day IN, Kivimaki M, Whittaker J, Humphries SE, Hingorani AD, ASCOT investigators, NORDIL investigators, BRIGHT Consortium. Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip, *American Journal of Human Genetics*, 2009; 85(5): 628-42.

Padmanabhan S, **Hastie C**, Sainsbury CA, McBride MW, Connell JM, Dominiczak AF. The cat, the fly and the beetle – why genetics need a semantic education, *International Journal of Semantic Computing*, 2009; 3(1): 77-90.

Padmanabhan S, Melander O, **Hastie C**, Menni C, Delles C, Connell JM, Dominiczak AF. Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols, *Journal of Hypertension*, 2008; 26(7): 1275-81.

Hastie CE, Haw S, Pell JP. Impact of smoking cessation and lifetime exposure on C-reactive protein. *Nicotine & Tobacco Research*, 2008; 10(4): 637-42.

Oral Presentations

Hastie CE, Padmanabhan S, Melander O, Johnson T, Di Blasio AM, Munroe PB, Caulfield MJ, Zanchetti A, Dominiczak, AF. Genome wide association study of blood pressure extremes identifies variant in uromodulin gene associated with hypertension, British Hypertension Society Annual Scientific Meeting, Cambridge, UK, 2010.

Hastie C, Smith GCS, Mackay D, Greig K, Pell JP. Relationship between preterm delivery and subsequent C-reactive protein and ischaemic heart disease risk, British Hypertension Society Annual Scientific Meeting, Cambridge, UK, 2009.

Hastie C, Padmanabhan S, Slack R, Isles C, Pell J on behalf of the Scottish Coronary Revascularisation Register Steering Committee. A study of the "obesity paradox" across the spectrum of cardiovascular risk, British Hypertension Society Annual Scientific Meeting, Cambridge, UK, 2008.

Hastie C, Padmanabhan S, Slack R, Isles C, Dominiczak AF, Pell J on behalf of the Scottish Coronary Revascularisation Register Steering Committee. A study of the "obesity paradox" across the spectrum of cardiovascular risk, Joint Scientific Meeting of the

European Society of Hypertension and the International Society of Hypertension, Berlin, Germany, 2008.

Hastie C, Smith GCS, Pell JP. Relationship between preterm delivery and subsequent C-reactive protein, Scottish Society for Experimental Medicine, Glasgow, UK, 2007.

Hastie CE, Craig P, Haw S, Pell JP. Impact of smoking cessation on C-reactive protein, and the role of life-time and passive smoking, World Congress of Cardiology, Barcelona, Spain, 2006.

Summary

Hypertension is a common, highly heritable trait of complex aetiology. Multiple environmental and lifestyle factors contribute to blood pressure variation. Hence the study of hypertension causality is not straightforward. Genetic linkage studies have implicated a number of loci involved in blood pressure regulation and the development of hypertension. Candidate gene association studies, however, have not reported any reproducible associations. Early genome-wide association studies (GWAS) showed remarkable success in identifying validated common variants associated with common diseases such as coronary artery disease and type 1 diabetes. However, the first GWAS of hypertension showed little success. This was largely because of a lack of statistical power and insufficient genomic coverage. Furthermore, it is widely believed that the failure of one GWAS of hypertension was partly due to misclassification of controls that were not phenotyped for blood pressure. Subsequently, two large international consortia-run GWAS of blood pressure as a quantitative trait produced tangible results.

The current study is a GWAS of hypertension using an extreme case-control design. It employed intensive phenotyping and extreme case-control definitions to select a sample of individuals from a restricted geographical area of relative homogeneity. The aim was to reduce misclassification bias and increase the likelihood of detecting any genetic effects. Cases were sampled from the Nordic Diltiazem study, and defined as individuals younger than 60 years with at least two consecutive measurements of systolic blood pressure (SBP) $\geq$ 160 mmHg or diastolic blood pressure (DBP) $\geq$ 100 mmHg. Controls were sampled from the prospective Malmö Diet and Cancer Study, and defined as individuals aged at least 50 years with SBP $\leq$ 120 mmHg and DBP $\leq$ 80 mmHg with no evidence of cardiovascular disease during ten years of follow-up. The groups represent, respectively, the upper 1.7% and lower 9.2% of the Swedish blood pressure distribution. Comparison of groups from the extreme tails of distribution increased statistical power by inflating observed effect sizes. With genome-wide SNP coverage we were able to adjust for population stratification using principal components analysis.

Following quality control exclusions, a final set of 521,220 single nucleotide polymorphisms was available for analysis in 1,621 cases and 1,699 controls. Seventeen SNPs were associated with hypertension at a $P < 1 \times 10^{-5}$ threshold of significance, of which three attained genome-wide significance, defined as $P < 5 \times 10^{-7}$.

The top hit, rs13333226, underwent a two stage validation process in a total of 14 independent cohorts. The combined odds ratio for the discovery cohort and all replication cohorts meta-analysed was 0.87 (95% CI 0.84 – 0.91, $P = 3.67 \times 10^{-11}$) with the minor G allele associated with a lower risk of hypertension. In total 21,466 cases and 18,240 controls were included. After adjustment for age, age^2, sex, and BMI, and when the discovery cohort was excluded from analysis, the association remained significant. Estimated glomerular filtration rate (eGFR), a measure of kidney function, was available in seven of the cohorts. When the analysis was repeated with adjustment for eGFR the effect was marginally strengthened. rs13333226 is located in close proximity, at -1617 base pairs, to the uromodulin (*UMOD*) transcription start site. *UMOD* encodes uromodulin, also known as the Tamm-Horsfall protein. Uromodulin is produced predominantly in the thick ascending limb of the loop of Henle and is the most abundant protein in urine. Its function is unclear; however, variants in *UMOD* have been associated with chronic kidney disease.

Clinical functional studies were conducted in three separate populations. The minor G allele of rs13333226 (associated with a lower risk of hypertension) was associated with lower urinary uromodulin excretion. Furthermore, in one sample following a low salt diet urinary uromodulin excretion was significantly lower in the presence of the G allele, whereas after a high salt diet genotype was no longer associated with urinary uromodulin. If this were verified, this would entail a gene-environment interaction. Our combined results suggest that *UMOD* may have a role in regulating blood pressure, possibly through an effect on sodium homeostasis.

There is ample evidence of a strong, graded relationship between blood pressure and subsequent renal disease. Hence the current finding is biologically plausible. Information on kidney disease was not available for the discovery samples so this could not be explored. However, the association between rs13333226 and hypertension was not substantively altered by adjustment for eGFR in the seven validation cohorts in which it was recorded, suggesting that it is independent of renal function.

In conclusion, we have performed a GWAS of hypertension using an extreme case-control design. The most significant hit was validated in a meta-analysis of the discovery sample and 14 additional cohorts. Moreover, functional studies showed a relationship between genotype and urinary protein excretion. Overall, we demonstrate that with careful

methodological planning and phenotyping it is possible to generate replicable hypertension GWAS results in a relatively small sample size.

1 Introduction

1.1 Cardiovascular disease

Cardiovascular disease (CVD) is a common complex disease of major public health importance with high prevalence throughout the world. This was highlighted by the Global Burden of Disease Study which analysed data from 47 countries between 1950 and 1990 to estimate the leading causes of mortality and disability worldwide [1, 2]. In addition, it predicted the impact of the same causes in 2020 using these estimates and regression equations by region, based on projected changes in key socioeconomic parameters (gross domestic product per person, average number of years of education, smoking intensity) and time. Ischaemic heart disease and cerebrovascular disease were identified as the first and second most common causes of death, respectively, and were predicted to remain so in 2020. Their proportional impact on disability adjusted life years (DALYs) was projected to increase from being the fifth and sixth most common causes, to the first and fourth. [The DALY is a measure of overall burden of disease, commonly used in public health research, which combines the impact of premature death and disability [3].] The reason for this increase is that most of the world is developing economically, and in the process the prevalence of many cardiovascular (CV) risk factors, such as older age, obesity, smoking, alcohol, decreased physical activity, is increasing. The change in pattern of disease as countries develop, from predominantly communicable diseases to chronic noncommunicable degenerative diseases, is termed epidemiologic transition (Figure 1.1)[4]. The 2003 World Health Report found CVD to be the leading cause of mortality in developing countries (Figure 1.2)[5], which translates to more than 10 million deaths. Moreover these deaths occur at a relatively younger age compared with developed countries. A further point made in the report is that CVD accounts for as many deaths in young and middle-aged adults globally as HIV/AIDS. Within the United Kingdom coronary heart disease (CHD) mortality varies geographically with higher rates in Northern England and Scotland (Figure 1.3)[6]. High blood pressure (i.e. hypertension) is the leading risk factor for mortality globally [7]. This is through the effect it has on various cardiovascular diseases.

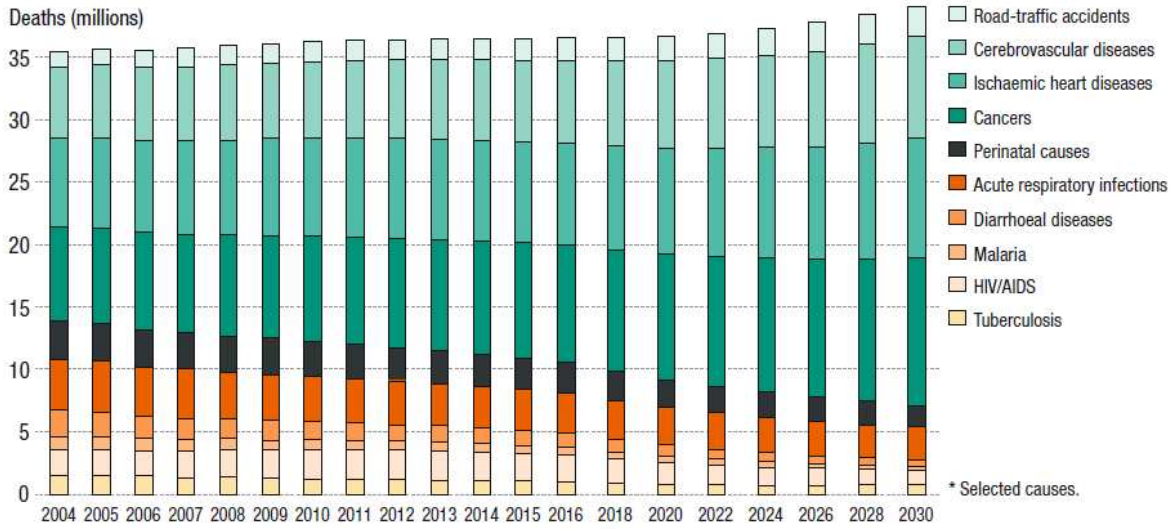


Figure 1.1 The shift towards noncommunicable diseases and accidents as causes of death. (reproduced from [4])

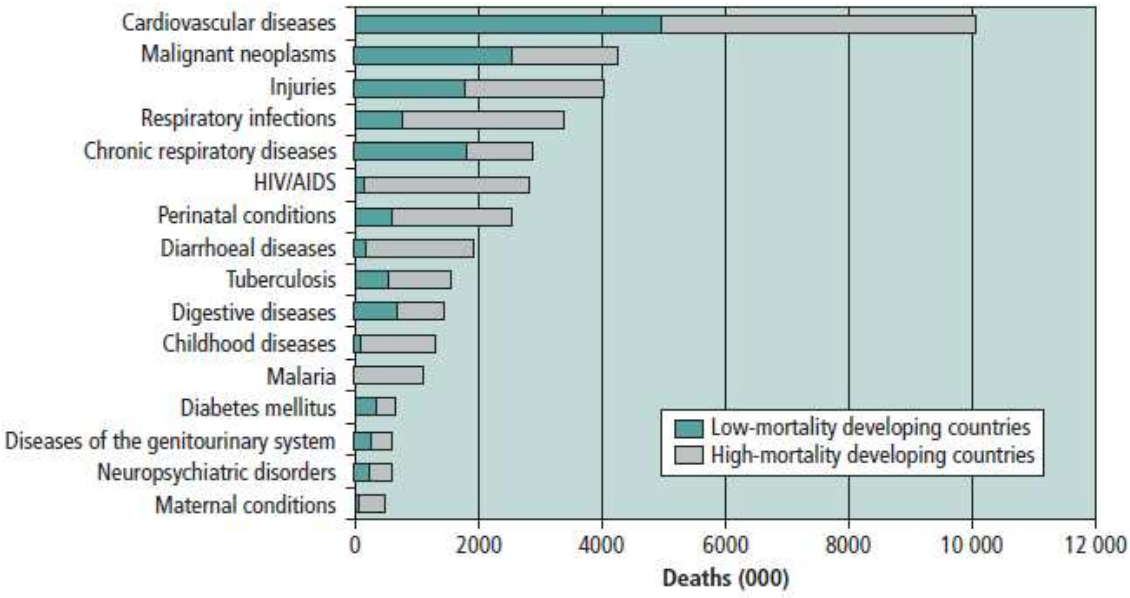


Figure 1.2 Deaths attributable to 16 leading causes in developing countries, 2001.
(reproduced from [5])

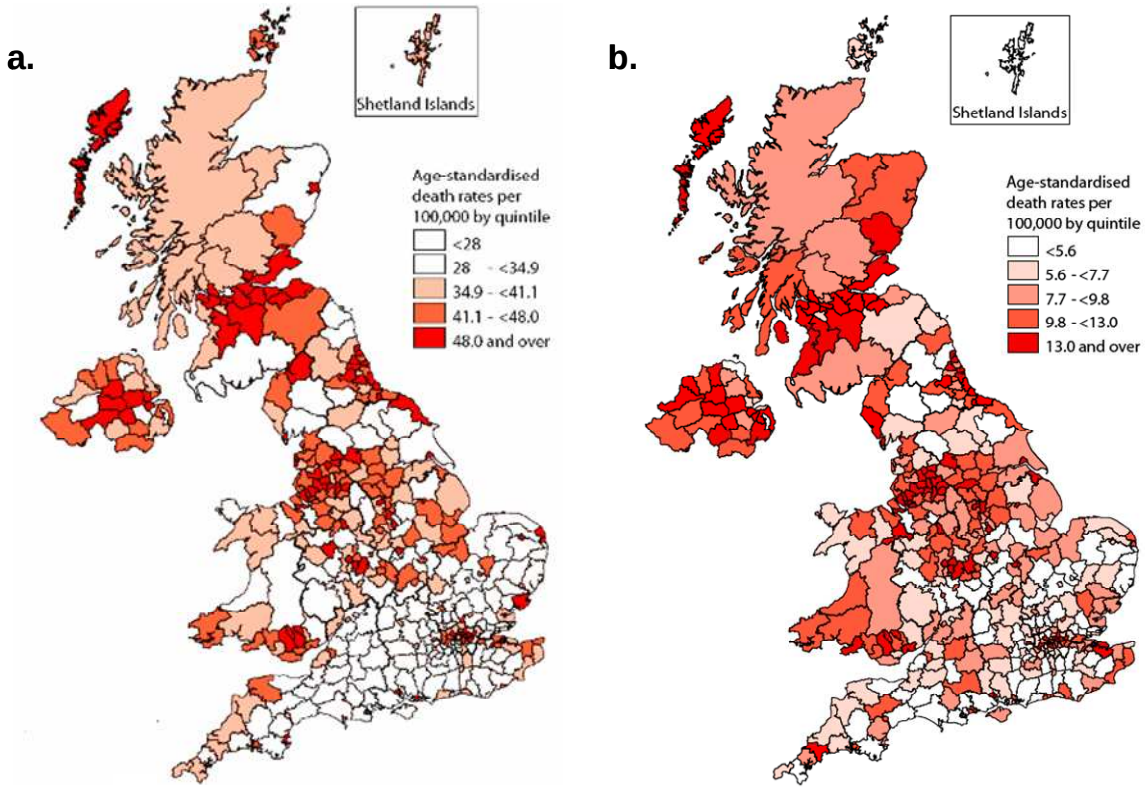


Figure 1.3 Age-standardised death rates per 100,000 population from CHD. For a. men and b. women, under 65 by local authority, 2004/2006, United Kingdom (reproduced from [6])

1.2 Human essential hypertension

Blood pressure refers to the pressure exerted by circulating blood on the walls of blood vessels, and is chiefly determined by cardiac output and total peripheral resistance. It is a quantitative trait that is highly variable both between and within individuals [8]. Blood pressure is typically measured non-invasively by sphygmomanometer via a cuff around the upper arm. It varies throughout the day in a circadian rhythm. Consequently, ambulatory pressure measured over 24 hours is a more reliable method of measurement than single (or a few averaged) measurements taken in the clinic or the home. Furthermore, measurements made in the clinic may be higher than average due to white-coat syndrome, the term used to describe the phenomenon of elevated blood pressure due to anxiety induced by the clinical setting. In the population blood pressure follows a slightly positively skewed distribution [9]. The blood pressure of an individual is expressed in terms of their maximum (termed systolic) and minimum (termed diastolic) pressures per heartbeat. These values are reported in millimetres of mercury (mmHg).

Hypertension is defined as a clinically significant increase in blood pressure, often considered to be diastolic blood pressure (DBP) ≥90mmHg or systolic blood pressure (SBP) ≥140mmHg. This cut-off is required for patient diagnosis and treatment; however it is somewhat arbitrary as there is evidence of a continuous relationship between blood pressure and CVD risk. For example, the Prospective Studies Collaboration comprised 61 prospective observational studies of blood pressure and mortality [10]. Age-specific survival analysis was conducted on a total of 958,074 participants. The subgroups of age at event were 40-49, 50-59, 60-69, 70-79, and 80-89 (deaths out with the range 40-89 years were ignored). Further analysis stratified by sex was also performed. Within each age group blood pressure was found to be strongly and directly related to vascular and overall mortality. There was no evidence of a threshold for risk, down to at least 115/75 mmHg (i.e. far lower than the cut-off typically employed for hypertension diagnosis). In other words, the mortality risk posed by blood pressure was continuous throughout most of its normal range. The findings were similar for males and females.

There has long been debate regarding the better way to classify and define essential hypertension, i.e. as a quantitative versus as a qualitative construct. Discussion of the topic was initiated in the 1940s and 1950s by the competing views expressed by Sir Robert Platt [11, 12] and Sir George Pickering [13], which developed into the “Platt versus Pickering debate”. Platt’s belief was that essential hypertension is a qualitative Mendelian trait that follows a

bimodal distribution, with one peak at the level of normotension and another at hypertension. Because blood pressure is also affected by environmental factors these curves would likely overlap. Conversely, Pickering believed that essential hypertension is a quantitative non-Mendelian trait with a unimodal distribution. According to this paradigm hypertension merely represents the extreme top end of the overall blood pressure distribution, and does not exist as a separate entity. Pickering and colleagues opposed the categorisation of blood pressure values into normal and abnormal as artificial. At the time of the debate Platt's viewpoint was favoured. However, over time Pickering's has come to dominate.

Hypertension is more common in men than women (at least until menopause age), people of African ancestry than European ancestry [14], and prevalence increases with age [15]. Kearney and colleagues estimated the global burden of hypertension by searching the published literature from 1980 to 2002 [16]. They concluded that 26.4% of the global adult population had hypertension in 2000, and projected that this would increase to 29.2% by 2025. It was estimated that the total number of adults with hypertension would increase by over 60% during this time period from 972 million to 1.56 billion. The rise will be far greater in economically developing than developed countries.

The impact of the high prevalence of hypertension is substantial because of its large effects on mortality and morbidity. A comprehensive review undertaken by expert working groups of 26 selected risk factors identified high blood pressure as the leading global risk factor for mortality and third largest cause of DALYs [7]. Figure 1.4 is adapted from the review findings and shows the contribution of this and other risk factors to mortality by region. Out of the ten highest contributors, five are major modifiable cardiovascular risk factors (high blood pressure, tobacco, high cholesterol, high BMI, physical inactivity). Underweight and unsafe sex result in a high proportion of mortality in developing regions with high mortality, but have very little impact in other regions. By contrast, hypertension causes a relatively high proportion of mortality in all regions, regardless of stage of development.

1.2.1 Causation of hypertension

90-95% of hypertension cases are described as essential or primary hypertension, meaning that no medical cause is known for the elevation in blood pressure. In the remaining cases

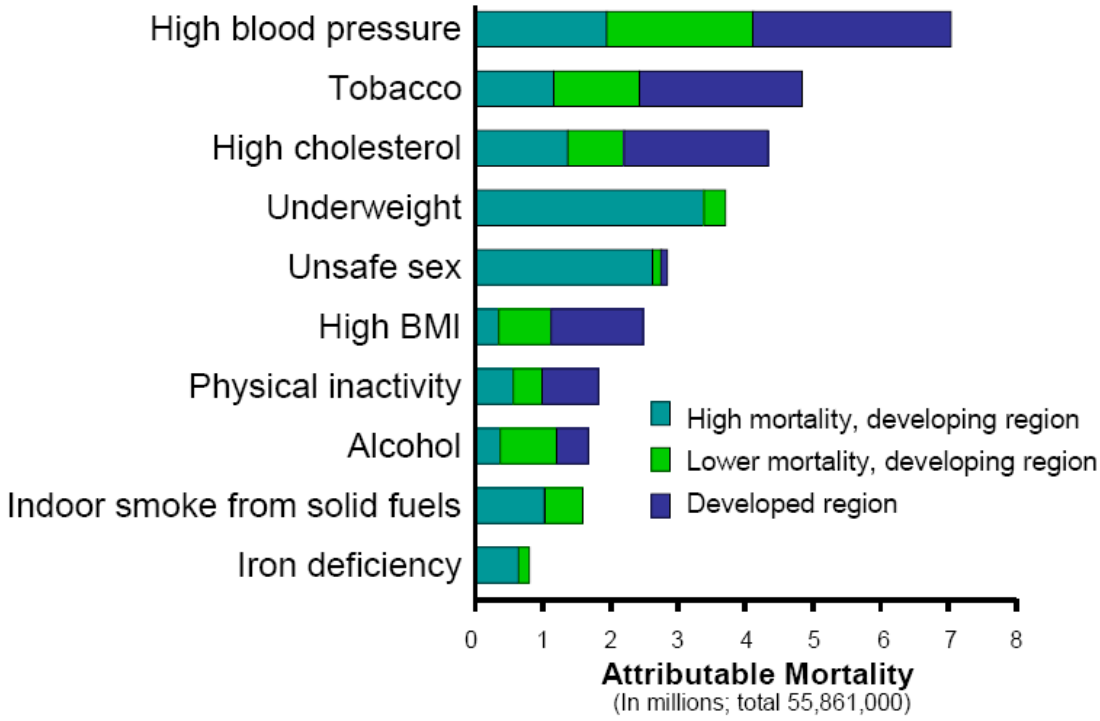


Figure 1.4 Number of deaths attributable to major risk factors worldwide. (adapted from [7]).

it is secondary to renal disease, endocrine disorders or other causes, or is monogenic (due to variation in a single gene). Essential hypertension is a complex heterogeneous disorder and it is thought that many factors contribute to it. Two that have been identified and studied a great deal are salt intake and obesity. Others include insulin resistance, high alcohol intake, lack of exercise, stress, low calcium intake and low potassium intake [17, 18]. These risk factors as well as genes, gene-gene interactions and gene-environment interactions, contribute to the complexity of hypertension and make it inherently difficult to study [19].

Some of the variables influencing hypertension risk, such as level of exercise and dietary calcium and potassium, are solely environmental and can be significantly improved through lifestyle modification. Others are themselves heterogeneous variables affected by both genes and the environment. For example, alcohol intake is clearly mainly determined by consumption. But individuals inheriting a variant of aldehyde dehydrogenase 2 (*ALDH2*), common in the Japanese, experience a more extreme negative response to alcohol and hence consume less on average [20]. The relationship between salt intake and blood pressure is mediated by a person's salt sensitivity which is partly genetically determined [21]. Furthermore all of the genes hitherto implicated in monogenic forms of hypertension and hypotension regulate renal salt re-absorption [22]. There is now substantial evidence, including many genome-wide association studies (GWAS), of a genetic component to obesity [23]. However the huge increase, especially in the developed world, in obesity prevalence is on the whole due to the widespread adoption of sedentary lifestyles.

1.3 Genetic factors in hypertension

1.3.1 Evidence of genetic determinants of blood pressure regulation and hypertension

There are multiple strands of evidence showing that genetic factors contribute to blood pressure and hypertension. Firstly, the normal distribution of blood pressure in the general population indicates the presence of multiple environmental and genetic factors and thus a polygenic aetiology. Secondly, rare monogenic forms of hypertension associated with major defects in renal salt handling prove that gene mutations can cause hypertension, and there is a hypothesis that minor variations in these genes may contribute to essential

hypertension. Finally, from a population perspective, there is considerable evidence from twins and family aggregation studies indicating the presence of a heritable component.

It is estimated that around 30% of variation in blood pressure is due to genetic factors [24]. Hypertension is about twice as common in individuals who have one or two hypertensive parents, and blood pressure is more closely correlated in identical (monozygotic, MZ) than nonidentical (dizygotic, DZ) twins [9, 25]. The Victorian Family Heart Study recruited a sample of adult families, comprising both parents and at least one offspring, that was enriched with families containing twins. Figure 1.5 presents plots of correlation coefficients and their standard errors for SBP, DBP and pulse rate, for different types of family member pairs. There is a trend towards decreasing correlation with decreasing genetic similarity. As expected MZ twins have the highest correlation coefficients for each parameter, and spouse-spouse pairs the lowest. In the Montreal Adoption Study investigators compared blood pressure correlation between biological sibling pairs and adoptive sibling pairs (as well as parent-child correlations). SBP correlation coefficients were 0.38 and 0.16 for biological and adopted siblings respectively, and DBP coefficients 0.53 compared with 0.29 respectively [26].

Two measures that are commonly used to assess the genetic component of a trait are heritability (h2) which is the fraction of variation in disease susceptibility due to genetic factors, and sibling recurrent risk (λs) which is the degree of elevated risk of disease for a sibling of an affected individual compared with a member of the general population. The heritabilities of clinic systolic blood pressure and clinic diastolic blood pressure are around 15-40% [27, 28], whereas for ambulatory night-time systolic and diastolic blood pressure the heritabilities are 69% and 51%[27]. It is pertinent to point out that though the heritability estimates are considerable, this does not equate to magnitude of genetic effect. This is because the denominator in the estimate of heritability comprises measurement error and variances attributable to genes, shared environment, unshared environment and unmeasured determinants. This is illustrated by the example above where minimising measurement errors by using ambulatory night-time values inflates the heritability estimates. Heritability is also a property of the population studied and low heritability estimates would suggest that genetic mapping would be difficult for that phenotype. The sibling recurrent risk of hypertension is around 1.2-1.5[29] indicating a phenotype with modest genetic effect. The complicated interplay between genetic and environmental factors that influence intermediary phenotypes in the development of hypertension is shown in Figure 1.6.

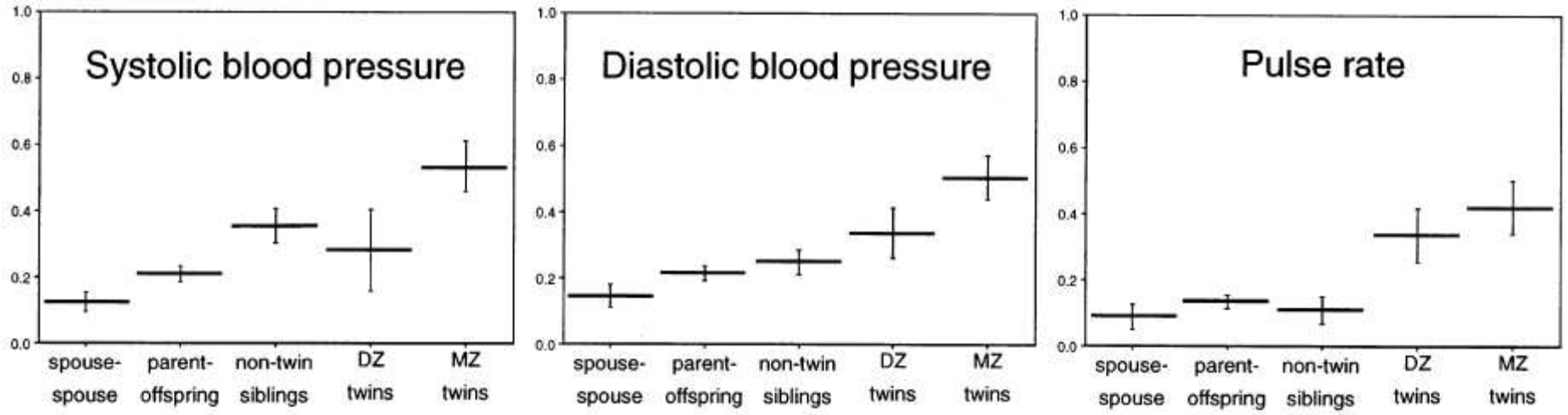


Figure 1.5 Correlation coefficients and their standard errors for pairs of family members. Pairs are spouse-spouse, parent-offspring, non-twin siblings, dizygotic (DZ) twins, and monozygotic (MZ) twins; for systolic blood pressure (left), diastolic blood pressure (middle), and pulse rate (right) (reproduced from [30]).

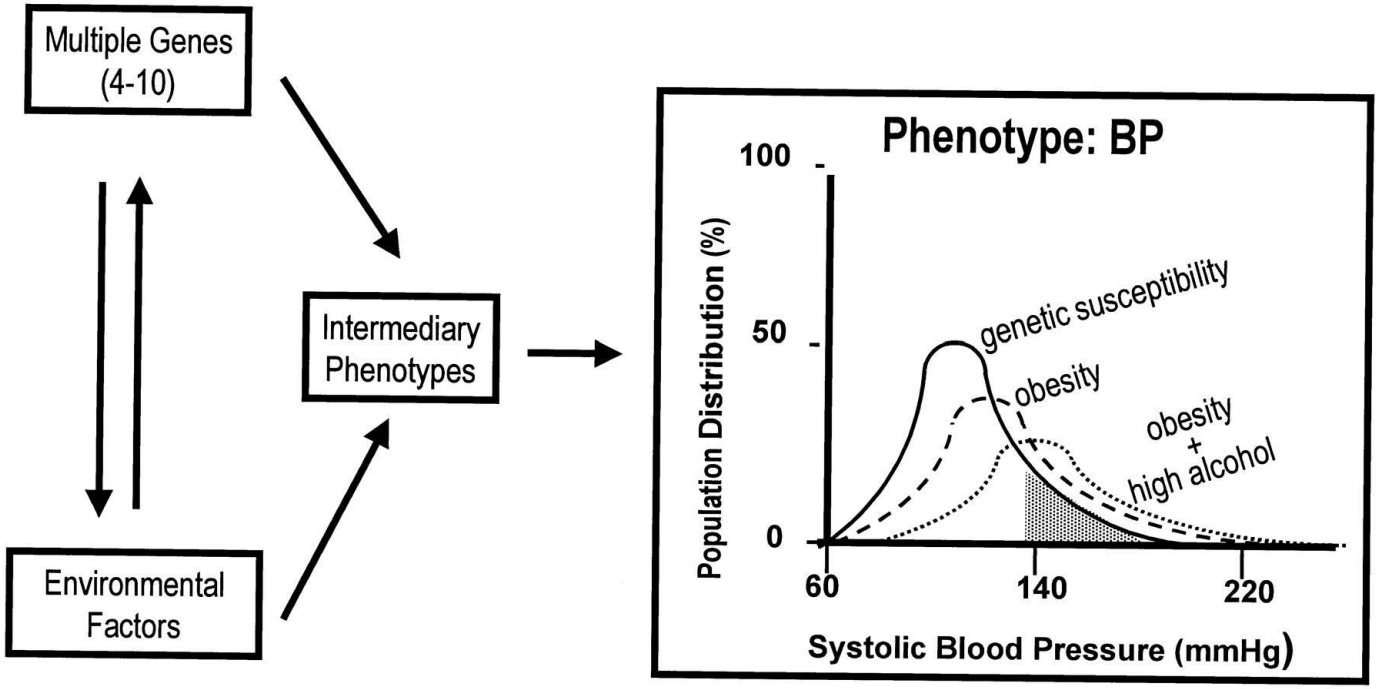


Figure 1.6 Interaction among genetic and environmental factors in the development of hypertension. Left side of figure shows how environmental factors and multiple genes responsible for high blood pressure interact and affect intermediary phenotypes. In the graph on right side of figure, unbroken line indicates theoretical blood pressure (BP) of the population that is not affect by factors that increase BP; the shaded area indicates SBP in the hypertensive range. Broken and dotted lines indicate populations in which 1 (obesity) or 2 (obesity plus high alcohol intake) factors which increase BP have been added. In these two populations the distribution curves are shifted to the right and the number of hypertensive individuals is significantly increased (reproduced from [9]).

1.3.2 Monogenic forms of hypertension

Seven monogenic (i.e. due to a single gene mutation), or Mendelian, forms of hypertension have been identified [31]. These forms of the disease are characterised by their severity and early onset, but their rarity means that they account for less than 1% of human hypertension. They can be categorised by initial mechanism into three distinct groups, all of which ultimately lead to a common mechanism of increased distal tubular re-absorption of sodium and chloride, volume expansion and hypertension. The first group includes Liddle syndrome, Gordon's syndrome, and activating mineralocorticoid receptor mutation in hypertension exacerbated by pregnancy; the mechanism involves mutations in sodium and chloride transporters that lead to their hyperactivity, or mutations of mineralocorticoid receptors that mimic mineralocorticoid excess. The second group includes congenital adrenal hyperplasia and apparent mineralocorticoid excess; the mechanism involves deficiencies of enzymes that regulate adrenal steroid synthesis and activity, thus increasing volume of precursors with mineralocorticoid activity. The third group includes familial hyperaldosteronism types I and II; the mechanism is excessive aldosterone synthesis that avoids regulatory mechanisms, and results in volume-dependent hypertension that suppresses renin release.

Mutations in 8 genes have been discovered that cause Mendelian/monogenic forms of hypertension [22]. In each instance the mutated gene products act in the same physiologic pathway in the kidney. The effect is to increase net renal salt re-absorption. Genes identified that cause Mendelian hypotension also act in this pathway but have the opposite effect of decreasing renal salt re-absorption (Figure 1.7). Raised renal salt re-absorption leads to increased intravascular volume and increased volume delivery to the heart, which in turn raises cardiac output hence blood pressure. Studies of candidate genes within the same pathway have had some success in identifying common genetic variants associated with essential hypertension [32-35].

Tobin et al conducted a study of the effect, in the general population, of common variation in all causal genes for monogenic hypertension and hypotension [36]. The sample studied comprised 2019 individuals from 520 nuclear families, unselected for blood pressure. Primary analyses were of mean 24 hour SBP and mean 24 hour DBP. Secondary analyses were of other blood pressure phenotypes and biochemical measurements. The key findings were for variants in the *KCNJ1* (potassium inwardly-rectifying channel, subfamily J,

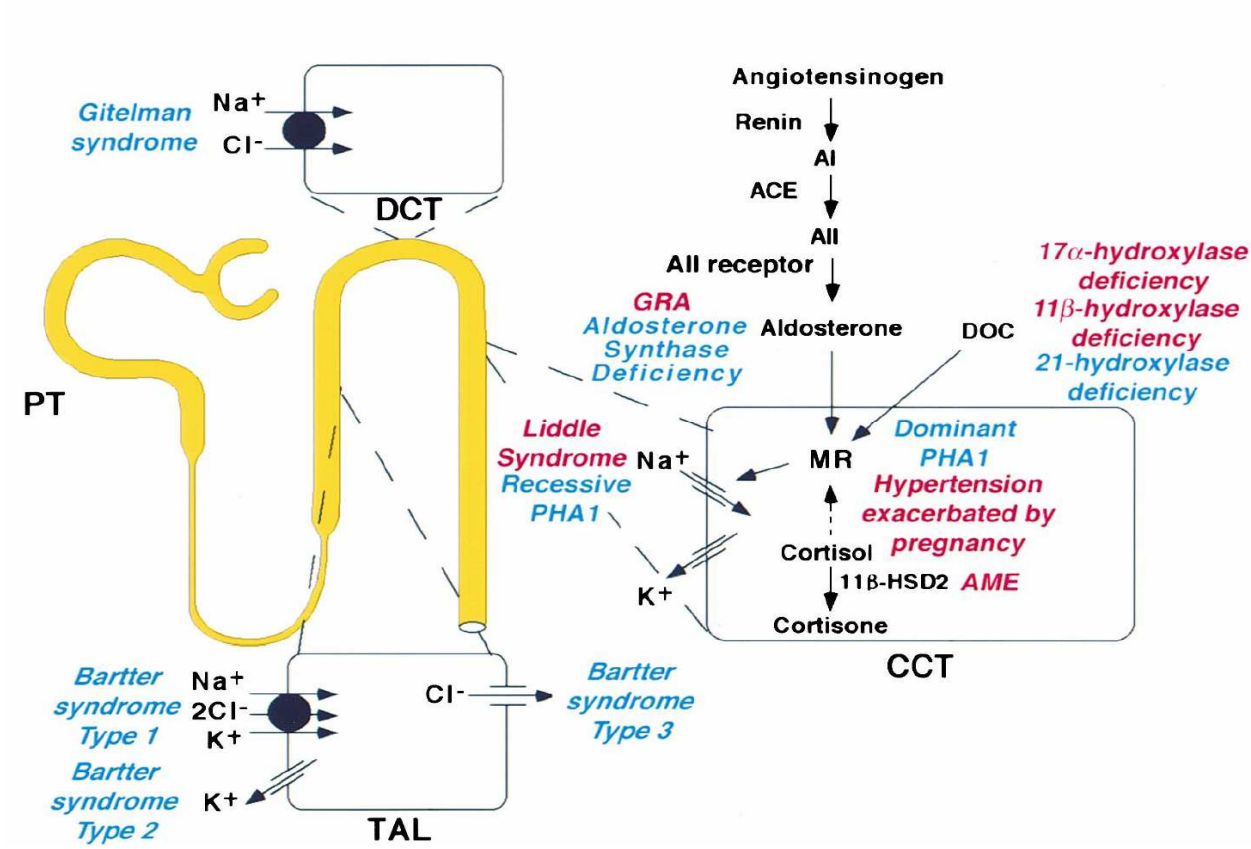


Figure 1.7 Mutations altering blood pressure in humans. A diagram of a nephron, the filtering unit of the kidney, is shown. The molecular pathways mediating NaCl re-absorption in individual renal cells in the thick ascending limb of the loop of Henle (TAL), distal convoluted tubule (DCT), and the cortical collecting tubule (CCT) are indicated, along with the pathway of the renin-angiotensin system, the major regulator of renal salt re-absorption. Inherited diseases affecting these pathways are indicated, with hypertensive disorders in red and hypotensive disorders in blue. AI = angiotensin I. ACE = angiotensin converting enzyme. AII = angiotensin II. MR = mineralocorticoid receptor. GRA = glucocorticoid-remediable aldosteronism. PHA1 = pseudohypoaldosteronism, type-1. AME = apparent mineralocorticoid excess. 11 βHSD2 = 11β-hydroxysteroid dehydrogenase-2. DOC = deoxycorticosterone. PT = proximal tubule (adapted from [37]).

member 1) gene: minor allele of the top hit for mean 24 hour SBP associated with a -1.58 mmHg change (95% CI -2.47 to -0.69, $p = 0.00048$); minor allele of the top hit for mean 24 hour DBP associated with a -0.95 mmHg change (95% CI -1.52 to -0.39, $p = 0.00095$). There were also nominally significant associations in *NR3C2* (nuclear receptor subfamily 3, group C, member 2), *CASR* (calcium-sensing receptor), *SCNN1B* (sodium channel, nonvoltage-gated 1, beta), and *SCNN1G* (sodium channel, nonvoltage-gated 1, gamma). These findings suggest that minor variants in the genes causing monogenic forms of hypertension may underlie common essential hypertension and variations in blood pressure more generally.

1.3.3 Linkage studies

Linkage studies map genetic loci in related individuals. The main advantages of family studies are that they allow the measurement of sex specific effects and heritability estimates, and avoid the problem of population stratification. Linkage analysis identifies large genomic regions that contain disease predisposing genes. If two loci are transmitted together from parent to offspring more often than expected under independent inheritance they are considered to be linked. Analysis can be parametric or non-parametric. Parametric linkage tracks the co-segregation of a marker and putative disease locus in large pedigrees, and then estimates recombination probabilities. If two loci are close together on the same chromosome they segregate together more often, and are less likely to be separated by a recombination event at meiosis. With decreasing relatedness of individuals the power to detect small effects increases, however the likelihood of recombination also increases meaning areas of linkage are shorter hence more markers are required. Parametric linkage makes many assumptions about the disease process, hence the term parametric, and is therefore effective for Mendelian disorders. The assumptions are required because of missing data, unknown phase, degree of penetrance, and sex-specific recombination.

Recently the focus of genetic studies has been on common (more than one case per 1,000 individuals [38]) complex diseases. For these diseases, hypertension amongst them, parametric linkage analysis is not possible because there is not a simple disease model and mode of inheritance. Instead non-parametric, i.e. assumption free, linkage analysis is used. This examines the proportion of haplotypes that are shared identical by descent (IBD) between affected relatives (common complex diseases have lower penetrance so unaffected family members provide much less information for linkage than affecteds). It is expected

that in the region of a disease-susceptibility gene there will be an excess of IBD sharing between affected relatives. The most straightforward way to perform non-parametric linkage is in affected sibling pairs (ASPs). Under the null hypothesis of no linkage, at a single locus the probability that no alleles are shared IBD by an ASP is 0.25, the probability that one is shared is 0.50, and the probability that two are shared is 0.25. These expected frequencies are compared with those observed, to assess whether significantly more alleles are shared IBD than expected by chance. Analysis can also be performed on relative pairs other than siblings. There has been some success in identifying susceptibility genes that modestly increase risk of hypertension by non-parametric linkage, for example the MRC British Genetics of Hypertension (BRIGHT) study [39].

The field of common complex disease genetics has in recent years moved from linkage to association study design because association analysis has far greater power to detect variants of modest effect and of lower frequency. The gene mutations responsible for monogenic Mendelian forms of hypertension are highly penetrant, and under very strong selection which keeps them at low frequencies with high levels of allelic heterogeneity. Thus these are highly amenable to linkage analysis. By contrast, susceptibility variants involved in essential hypertension are likely to have low or medium penetrance, and are probably not subject to such strong selection resulting in lower allelic heterogeneity and greater prevalence of the trait. Thus linkage analysis as expected has not really provided robust validated loci for hypertension. The minor allele frequency (MAF) of a polymorphic locus is defined as the frequency of the less/least common allele and consequently varies between 0 and 0.5. It is used to define the commonality of single nucleotide polymorphisms (SNPs) for analysis. When individuals are compared a SNP refers to a single base difference in the DNA sequence. SNPs are the most common variation in the genome [40]. At an MAF of 1% (i.e. 0.01) it is estimated that SNPs occur once every 290 base pairs (bps), amounting to around 11 million in total [41]. At an MAF of 5% they are expected to occur once every 450 bps. To detect loci conferring a genotypic relative risk of 1.5 (MAF=0.1) by linkage analysis requires an estimated 67,816 ASPs, whereas detection is possible through association with just 2218 singletons [42]. Moreover it is easier to recruit participants from the general population (than families), and there are fewer sampling restrictions in some disease categories such as late onset.

1.3.4 Association studies

Association studies are typically performed in unrelated population samples (although it is possible to conduct them on related individuals). The same allele(s) is associated with the trait across the population, whereas linkage can associate the trait with different alleles in different families. For qualitative traits, association analysis measures statistical association between a disease (phenotype) and genetic marker (genotype) directly by comparing allele frequencies of cases and controls. The aim is to establish whether a particular allele occurs in cases (compared with controls) more frequently than expected by chance. Quantitative traits, e.g. blood pressure, cholesterol, glucose, are assessed for association using linear regression. Association operates over shorter genomic distances than linkage and requires far more markers. In the past it was primarily used to fine map loci identified by linkage. There are two forms of association; direct and indirect. Direct association studies measure polymorphisms (usually SNPs) which are putatively causal. They are more powerful than indirect association studies, however the identification of candidate polymorphisms is not easy. It is probable that many causal variants for common complex diseases will be non-coding, instead influencing gene regulation, expression and splicing. There is not currently sufficient information on the causality of common diseases, including hypertension, for such variants to be identified and assessed for association.

Most GWAS rely on indirect association, on a large scale, to detect causal variants [43]. Indirect association measures the association between a phenotype and a marker polymorphism (or 'tag' SNP), which is correlated with the true causal allele due to linkage disequilibrium (LD). This is illustrated in Figure 1.8. Linkage disequilibrium is defined as "The statistical association, within gametes in a population, of the alleles at two loci" [44] (on the same chromosome). It is assumed that typically a causal variant will not have been typed in a given study. The amount of LD between two loci is summarised by the metric r^2 which varies between 0 and 1 and is inversely proportional to sample size, so with increasing LD a lower sample size is required. To cover unobserved loci well an r^2 value of ≥ 0.8 with typed loci is considered sufficient [45]. A related measure of LD is *D'*, which provides additional information about recombination breakpoints. In general SNPs in LD are more likely to be inherited together because they are physically close to each other on the genome. But this is not necessarily the case; studies have shown that levels of local LD vary, with some adjacent SNPs being independent despite their proximity and others of $\geq$ 100kb apart being in useful LD [46]. Patterns of LD are affected by many factors such as

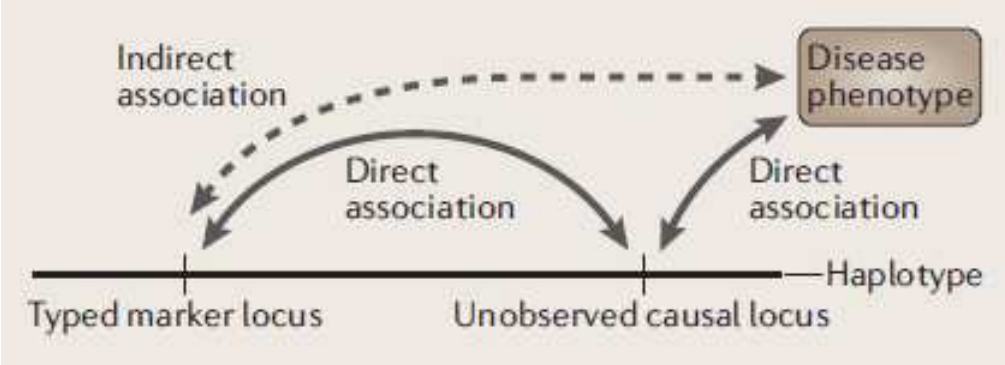


Figure 1.8 Pictorial representation of indirect association. Because the causal locus is unobserved the two direct associations cannot be observed. However if LD between the typed marker locus and causal locus is high then it may be possible to detect the indirect association between the marker locus and disease phenotype (reproduced from [44]).

population growth, population structure, admixture, natural selection, genetic drift, rate of recombination and mutation, and gene conversion [43].

1.3.5 Candidate gene studies

Traditionally association studies tested hypotheses based on candidate genes, for which there was prior evidence (of known physiological pathways that affect the phenotype in question) that a genetic variant influenced disease risk (Figure 1.7). Findings from candidate gene studies suggest that numerous polymorphisms act together (along with environmental variables) to produce a cardiovascular phenotype.

No candidate gene study has yet demonstrated a reproducible association with hypertension [47]. There are several potential reasons for this which highlight the drawbacks of candidate gene studies:

- the wrong genes may have been selected
- the causative genes may be upstream or downstream from the genes studied
- discovery of genetic variants in novel pathways is not possible as candidate gene studies rely on a priori information regarding disease mechanisms

In addition to the above there are the possibilities of population stratification, phenotypic and locus heterogeneity, and insufficient sample size: problems common to candidate gene studies and GWAS. Finally the SNPs studied might not provide complete coverage of the variants within the genes. Candidate gene studies do have the advantage over GWAS, however, that markers can be typed more densely. Thus the probability of detecting any true causal effect is improved as well as the probability that negative findings are true negatives.

1.4 Genome-wide association studies

In recent years there has been a great increase in the number of GWAS. They are hypothesis generating with the aim of identifying variants not previously implicated in the disease process, and no assumptions are made regarding the location or function of the causal variant. Typically, tag SNPs are used at suitable intervals along the entire genome,

in order to economically cover it without the genotyping of every SNP. The dense genotyping chips that are now available contain hundreds of thousands of SNPs so offer increasingly better coverage of the human genome (whether within or outside genes). Adequate coverage requires ≥ 300,000 SNPs with more needed for African samples due to greater genetic diversity in those populations [44] and less LD [48, 49]. One limitation is that tagging can only effectively capture variants that are common (see common disease common variant hypothesis below). Furthermore there is some concern that a set of tag SNPs that were selected in one population may not perform well in another, particularly in reference to the generalisability of the International HapMap Project samples. But there is evidence of good tag SNP transference across populations [50, 51]. This is especially true for different populations within the same continent; and the greatest disparities are between African and non-African samples.

An alternative genotyping method (to using tag SNPs) which similarly reduces the overall burden is to use direct association and genotype all potentially functional SNPs [43]. In this case the assumption is that certain variants are more likely to be associated with complex traits than others. They are chosen based on information recorded in publicly available SNP databases. There are not usually a priori hypotheses available about specific variants in relation to common complex diseases, hence the widespread use of indirect association [52].

Jorgenson and Witte have argued for a gene-centric approach to GWAS [53]. The reasons they outline are: genic variants are more likely to be functionally important than non-genic; and variants in many genes are in lower LD than those outside genes so may be difficult to capture through indirect association. By focusing solely on genes and not the whole genome there is potential to increase coverage of genes and decrease the genotyping burden. The genic approach has greater power to detect variants within genes but suffers from a loss of power for non-genic variants. Despite this Jorgenson and Witte demonstrated empirically using HapMap data that it is more efficient in detecting causal variants than the indirect whole-genome approach when related to genotyping burden. Their suggestion of the best overall GWAS approach is to combine indirect genotyping data with gene-based SNPs in high priority regions, or alternatively to use a more stringent LD threshold in genic regions to ‘over-capture’ genic SNPs.

A limitation of GWAS is that they are very expensive, especially with the large sample sizes that are required for small effects. But technological advancements are rapidly reducing the cost of genotyping. In a bid to further reduce costs, some researchers have adopted a two-stage GWAS study approach. In stage 1 a proportion of the samples are genotyped on all markers, and then in stage 2 a proportion of these markers are genotyped in the remaining samples [54]. Another approach to make studies more economical is the use of common controls for several groups of different disease cases. This was recently demonstrated by the Wellcome Trust Case Control Consortium (WTCCC) [55]. In 2008 Donnelly summarised all GWAS recorded at that time in the National Human Genome Research Institute (NHGRI) catalogue in the United States [56], which amounted to more than 300 replicated associations for more than 70 common diseases and quantitative traits [57]. As of September 2009 there were more than 500 published associations of genome-wide significance, defined as $P \leq 5 \times 10^{-8}$ (Figure 1.9).

1.4.1 Subject ascertainment/ phenotyping

Technological advances have recently improved the quality of genotyping so that increasingly the chance of identifying genetic effects depends upon phenotyping quality. This is especially true for hypertension as it exists on a continuum of blood pressure which is affected by factors such as antihypertensive medication, and time and method of measurement. Furthermore, as mentioned above its definition is somewhat arbitrary and has changed over time. Because of this some investigators have studied blood pressure as a quantitative trait [58, 59]. It has become apparent that controls as well as cases must be intensively phenotyped, to avoid the inclusion of controls with undiagnosed hypertension [47, 55]. Ideally phenotyping and genotyping protocols should be the same for cases and controls.

1.4.2 Phenotypic enrichment for genetic effects

In order to increase the chance of detecting any genetic effect some studies employ phenotypic enrichment. This involves selecting a subsample of cases with severe disease. For example, in their study of loci for essential hypertension, the BRIGHT study sampled individuals in the top 5% of the UK blood pressure distribution [39]. Furthermore participants were not diabetic and not obese, decreasing environmental risk. Low environmental risk increases the chances of a disease case having substantial genetic causation. For example, the study of obstructive lung disease in non-smokers minimises

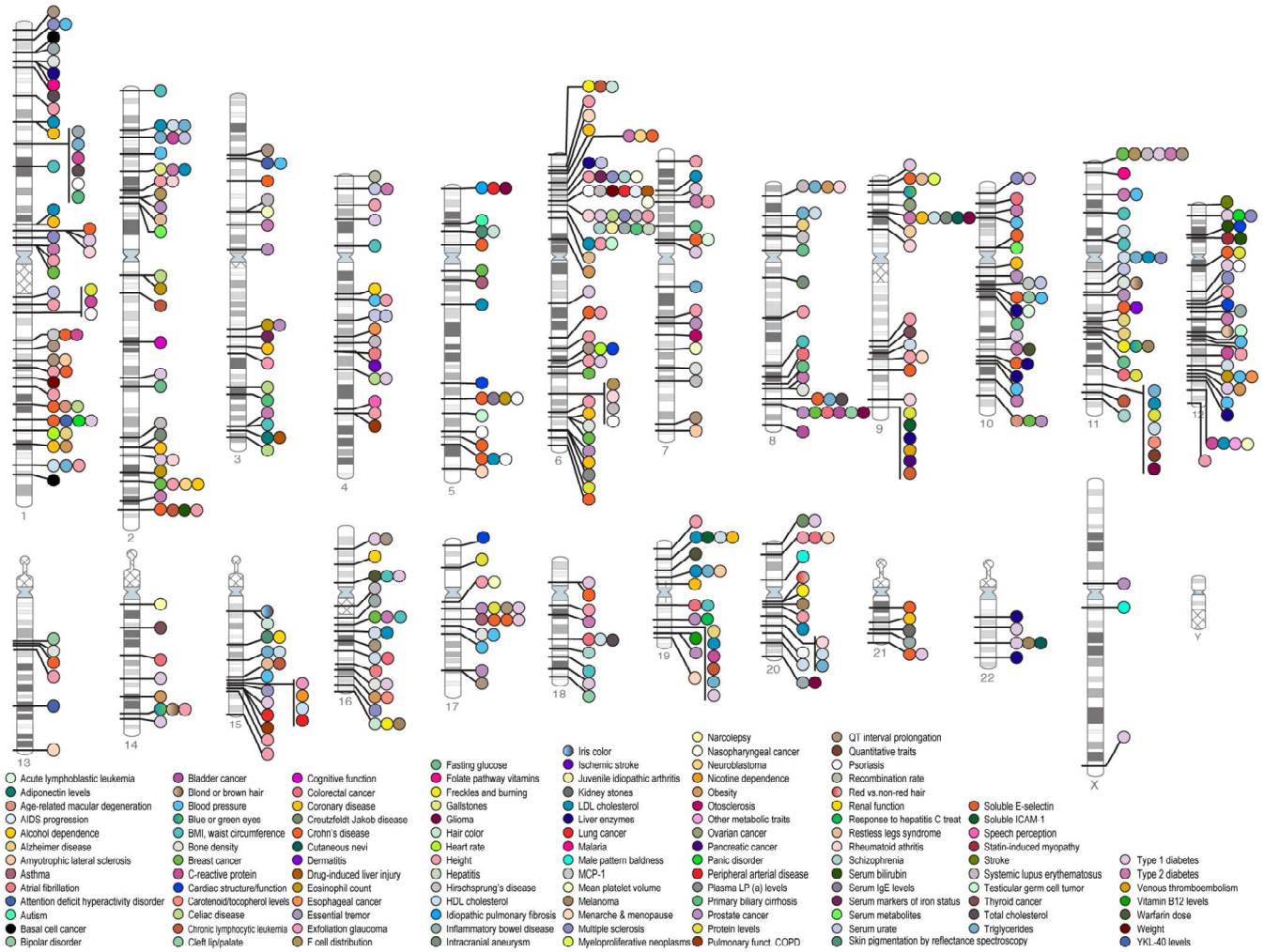


Figure 1.9 Published genome-wide associations up until September 2009. 536 published genome-wide significant associations at $P \leq 5 \times 10^{-8}$ (reproduced from [56])

environmental risk, and increases the likelihood of detecting genetic effects, because smoking is a major contributing factor. The study of cases with premature disease and a strong family history also increases the chance of finding genetic effects [60, 61].

1.4.3 Population stratification

Stratification acts as a confounder and can result in artefactual evidence of association. It occurs when there are two or more strata in a population, and both the risk of disease and the frequency of marker alleles differ between strata. It therefore may appear that the risk of disease is related to the marker alleles when in fact it is not. A similar concept is admixture, which refers to "the mixture of two or more genetically distinct populations" [45]. The International HapMap Project, described below, has demonstrated clear genetic differences between geographically separated populations [62].

There are two examples of confounding by population stratification that are frequently cited in the literature [63]. The first is a significant association between diabetes and an HLA (human leukocyte antigen system) haplotype amongst individuals on a Pima Indian reservation in the United States [64]. This observation was found to be due to ethnic admixture of white European and Pima Indian ancestry, in that prevalence of diabetes and the susceptibility haplotype were both much higher in those of Pima Indian ancestry. Once analysis was restricted to only those of purely Native American ancestry the association disappeared. The second example is studies that observed a significant association between alcoholism and the dopamine D2 receptor [65, 66]. It has been established that this is confounded by varying allele frequencies and alcoholism prevalence by ethnic group. Other than these there are few examples of type 1 error introduced by population stratification, and it is thought that in the past the issue was overstated. In a case control study associating the N-acetyltransferase slow acetylation genotype with male bladder cancer and female breast cancer in American Caucasians with ancestries from eight European countries, Wacholder et al demonstrated empirically that when ethnicity was ignored the resulting bias was only $\leq 1\%$ [67]. Moreover in a theoretical study that assessed a range of cancer rates and genotype frequencies they found that the risk ratio was biased by less than 10% in U.S. studies except in extreme circumstances. That is when there are large differences in genetic variance and disease prevalence between ethnic groups, and the information provided on ethnicity is insufficient for adjustment. The effect of stratification on analysis increases with increasing sample size because even modest levels of underlying population structure are amplified [68, 69]. This has particular relevance to GWAS as they are

employing larger and larger samples. A final important point is that confounding by population stratification tends to actually decrease (counter intuitively) with increasing number of ethnic groups [67]. This is because the direction of bias may differ between groups so that the overall combined effect is diluted. Whether population stratification is a real concern or not, to avoid any possibility of bias it is now commonplace in studies of unrelated cases and controls to employ stratification detection and correction methods.

1.4.4 Solutions to population stratification

Aside from matching cases and controls for genetic background and relevant environmental factors as well as possible and straightforward adjustment for ethnic group, a number of possible solutions to population stratification have been proposed. One is genomic control (GC) which employs detection and correction methods, e.g. by using a bank of randomly selected markers (preferably >100 [44]) that are unrelated to the question of interest to assess artefactual association [68, 70-72]. GC calculates an inflation factor λ by dividing the median of the genome-wide chi-square distribution (for data being assessed) by the median of the ideal, i.e. without population stratification, chi-square distribution [73]. If there is not significant evidence of population stratification then λ will equal 1 or be close to 1. Each association test chi-square statistic is divided by λ which has the global effect, if stratification is present, of adjusting p-values to be less significant across the dataset. The GC method of correction is crude in that all genotype-phenotype associations are adjusted by a uniform amount. This does not take into account the variability of marker allele frequency differences across ancestral populations [74]; therefore GC may lead to under adjustment at markers that have strong differentiation across ancestral populations and over adjustment at those that are not differentiated.

A similar approach is structure assessment (SA), which too uses unlinked genetic markers for detection but then attempts to match homogeneous subgroups of the sample for association analysis within these subpopulations [75-78]. It is assumed that any significant association observed within a subpopulation cannot be due to population structure; there is an issue, however, about how many subpopulations to apply since they are a theoretical concept [44].

Explicit detection/correction methods by principal components analysis (PCA) have also been employed [74, 79]. The aim of PCA is to reduce the dimensionality of multivariate data. It transforms variables into axes that account for decreasing proportions of variance in the

data and that are uncorrelated with each other. For example, the program EIGENSTRAT [74] uses PCA to infer continuous axes of variation which describe as much data variability as possible in a few dimensions. These axes are termed the top eigenvectors of a covariance matrix between samples, and can be entered as covariates into logistic (for qualitative traits) or linear (for quantitative traits) regression analysis. This enables the identification of extreme outlying individuals for exclusion, and for the remaining individuals genotypes and phenotypes are adjusted along each axis by amounts attributable to ancestry. Finally association analysis is conducted on the adjusted genotypes and phenotypes. EIGENSTRAT has the advantage that each marker is considered separately, hence the correction for ancestry effects is more precise than the uniform correction provided by GC.

Due to the large number of markers genotyped in GWAS it is possible to detect low levels of stratification. A caveat to all of these detection methods is that with a large enough sample size even small biases will be statistically significant, and may lead to overcorrection. Another approach is to study genetic isolates: areas that have a limited number of founders, short evolutionary history and limited mixing, making stratification less likely. In addition environmental conditions tend to be more homogeneous within isolates. For example the European Special Populations Research Network (EUROSPAN) is a collaboration linking five genetic isolate populations in Europe; Orkney, Croatian islands, part of the Netherlands, the Saami people and the Tyroleans [80]. Other isolates include Inuit people, the Amish and the Hutterites. A recent study of the Sorbs, a Slavic population isolate in Germany, identified a locus for hypertension on chromosome 1p36 [81]. Limitations of genetic isolate populations are their excess homozygosity due to inbreeding and their long regions of LD, which will make fine mapping and identification of causative variants next to impossible.

A commonly used method of avoiding population stratification in association studies is the transmission disequilibrium test (TDT) [82] which uses family based controls. The TDT is a method of assessing linkage in the presence of association. The basic study design is trios of an offspring proband and both parents. The genotype that is not transmitted to the proband (case) becomes the matched pseudo-control. Because cases and controls have the same genetic background there is no confounding by stratification. The TDT evaluates the frequency with which an allele associated with disease or its alternate is transmitted to the proband. Alleles have a 50% chance of being transmitted to offspring under Mendelian inheritance. So if an allele is transmitted to affecteds more than 50% of the time this

would indicate that it is associated with disease risk. This approach also allows the assessment of parent-of-origin effects, i.e. whether allele effects differ depending on whether they are inherited from the mother or father, which is not possible in case-control studies [83]. Moreover the inclusion of parental phenotypes in the analysis improves power. There are, however, some drawbacks to the TDT which limit its use. Parents must be heterozygous at a SNP for allele transmission to be evaluated, so ≥ 50% of data collected for each parent cannot be analysed [63]. Three people are required to gain information that would otherwise be provided by two (a case and a control), therefore efficiency is only two-thirds. Finally, as already mentioned above family recruitment can be very difficult for late onset disorders and there is the risk of an age-at-onset bias towards younger patients [84]. The inclusion of unaffected sibs in the TDT is possible, and should provide information on negative transmission of alleles. But for complex diseases of low penetrance their addition is not useful in practice because the amount of noise introduced outweighs any gain, thus reducing power [85].

1.4.5 Common Disease Common Variant Hypothesis

The chance of detecting genetic variants that influence common disease depends on the underlying genetic architecture. That is to say, the number of susceptibility alleles, whether they are common or rare, their effect size, and whether their action is neutral or deleterious. Allelic spectra vary greatly between disease genes. Common alleles and those of large effect are of course easier to detect, as are deleterious alleles. Thus far the success of GWAS has been dependent upon the validity of the common disease-common variant (CDCV) hypothesis, in that studies have had insufficient power to detect rare variants. The CDCV hypothesis predicts that the causative genes for common diseases have relatively simple allelic spectra, i.e. one or a few predisposing alleles of relatively high frequency. As yet there is insufficient empirical evidence to determine the validity of the CDCV hypothesis, and arguments for and against have put forward. These are crucial to research using SNP mapping to predict common disease risk which assumes that the theory is by and large accurate (linkage studies of families or ASPs, by contrast, are robust to allelic heterogeneity). For GWAS it has been suggested that, as a rough guide, SNPs should meet a threshold of MAF ≥ 1% [84] or MAF ≥ 5% [44] to be considered common.

Reich and Lander have outlined a model for predicting disease allele diversity that supports the CDCV hypothesis [86]. They argue that in human founder populations an allele's equilibrium frequency was determined by its effect on reproductive fitness.

Therefore, at least for late onset disorders such as hypertension and type 2 diabetes that have no discernable effect on fitness, disease-predisposing alleles could achieve high frequency. This is coupled with the estimation that the human population expanded rapidly from a small founder pool where neutral alleles had low diversity [87, 88]. During population growth many new alleles are generated, but Reich and Lander argue that it takes a long time for these to dilute out common alleles from the founder population, perhaps more than a million years. Taking these observations together, because common disease-predisposing alleles were likely not affected by natural selection they may have had high frequencies in the past and consequently remain at high frequency now. Hence they will be detectable by genome wide association. Reich and Lander state that it is in principle possible that the opposite is true, i.e. that some common disease risk is attributable to a large number of loci with rare disease-predisposing alleles. But the relative risks observed in family members (a more rapid than linear decline in risk with increasing distance of relationship) support the conclusion that the majority of risk is due to a modest number of loci with common disease-predisposing alleles.

A key part of the argument against the CDCV hypothesis rests on the fact that the risk of common disease depends on the interaction of many genes and environmental factors. In particular late-onset disorders of high prevalence in modern western society have been heavily influenced by changes in lifestyle factors such as diet and physical activity, and not by common disease-predisposing alleles. The risk conferred by any one factor, whether genetic or environmental, is weak and most cases are not predominantly determined by genetic variance. Pritchard devised a method to model the likely allelic spectra underlying common disease [89], and reaches very different conclusions to Reich and Lander. He argues that the frequency of disease-predisposing alleles is random, and results from the joint effects of selection, mutation, and random genetic drift. Furthermore he posits that, though it is possible that susceptibility alleles for common disease may be selectively neutral, it is also plausible that they are under weak selection in early life. Neutral alleles by their very nature tend to have disappeared or else become almost fixed in the population, so do not contribute much to the genetic variance associated with disease. Pritchard concludes that, with allelic heterogeneity underlying disease being high, the power of current association analysis is reduced and new statistical methods are needed.

Additional evidence against the CDCV hypothesis is taken from observations of the allelic diversity of late-onset disorders with Mendelian inheritance patterns [90]. Causal genes for these should also be selectively neutral, due to their late onset. However empirical

evidence shows that despite this, and contrary to the CDCV hypothesis prediction of low allelic diversity, these disorders can be highly diverse. This suggests adverse selection and low allele frequencies in founder populations [90]. For example >735 rare alleles of the Low-Density Lipoprotein (LDL) receptor have been found that increase the risk of premature coronary artery disease through familial hypercholesterolaemia. No common LDL receptor alleles have been found to influence disease risk. In contrast there are examples, such as the association between apolipoprotein E(*4) (*APOE4*) and early onset Alzheimer disease, that support the CDCV hypothesis. Variability of allelic spectra cannot, it appears, be predicted on the basis of a disorder's prevalence or time of onset. Moreover the existing evidence suggests that alleles of both high and low frequency play a part in common disease [38, 91-95].

Wang et al have argued against making a distinction between rare and common disease-predisposing alleles [45]. Instead they propose that the allelic spectrum of disease associated variants be considered in the context of all variants in the human genome. In this framework the neutral model is that the allelic spectrum of disease variants and that of all variants are the same. Most susceptibility variants would be rare (MAF < 0.01), however SNPs with MAF > 0.01 would still account for more than 90% of genetic differences between individuals and therefore make a significant contribution to phenotypic variation [41]. This lies somewhere between the two opposing views regarding the CDCV hypothesis discussed above. Common diseases will vary in their allelic spectra depending on the evolutionary forces exerted upon them; nevertheless it is estimated that each will likely have hundreds of common and rare variants contributing to their familial clustering [45].

The completion of Phase I of the International HapMap Project in 2005 represented a great step forward in the GWAS field [96]. The resulting HapMap (i.e. haplotype map) resource is a public database of common variation (defined as MAF $\geq$ 0.05) in over a million SNPs in the human genome. This translates to a SNP density of at least one every 5 kb. These SNPs were completely genotyped in a sample of 269 individuals from four populations: the Yoruba in Ibadan, Nigeria (YRI); Caucasians in Utah, USA (CEU); Han Chinese in Beijing, China (CHB); and Japanese in Tokyo, Japan (JPT). Because of the international sampling method used, information on genetic variation is available within and between populations. In the characterisation of patterns of LD across the genome, HapMap facilitates the design of GWAS in the choosing of tag SNPs, improving study efficiency. Knowledge of LD patterns is crucial for study design; low levels can make it easy to miss a variant even when a large number of SNPs are genotyped, and conversely high levels make

detection more likely but can impede identification of the causal variant (as many SNPs in a region may be in LD together). HapMap can also aid data analysis through haplotype imputation of untyped variants. Phase II has since been completed which genotyped a further 2.1 million SNPs (approximately one per kb) in the same sample of people [97]. Additional samples are now being collected from the original four populations, as well as seven further populations: Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); Tuscans in Italy (TSI); Gujarati Indians in Texas, USA (GIH); metropolitan Chinese in Colorado, USA (CHD); Mexican origin individuals in California, USA (MEX); and African ancestry individuals in the southwestern USA (ASW).

1.4.6 Significance thresholds for GWAS

By convention statistical significance using frequentist methods is determined using the *P* value threshold of 0.05, with values below this considered significant (i.e. there is evidence to reject the null hypothesis of no effect). This is not appropriate for GWAS because the large number of tests performed increases the chance of type I error. An alternative threshold, proposed by Risch and Merikangas and now widely adopted, is $P < 5 \times 10^{-8}$, which corresponds to an equivalent false positive rate of 5% for 1,000,000 independent tests of association [42]. This is calculated using the simple Bonferroni correction for multiple testing, which calculates a new significance threshold by dividing 0.05 by the number of tests performed. In practice this is conservative as it does not take levels of LD into account; the use of tag SNPs means the genome can be covered sufficiently with around half this number of SNPs (i.e. ~500,000). This threshold is preferable to the traditional 0.05 but it has been argued that *P* value alone is not adequate for assessing significance. In addition to the possibility of false positives within a study, the issue of multiple testing can be viewed in the context of replication across studies. If several groups publish the same nominally significant association and these are combined then the finding may be given undue weight [44], unless negative results are also made public thus avoiding publication bias.

The Bayesian school of statistical thought states that the prior probability of an association being true and the power of the study must also be taken into account. This is a complicated process in that probability must be calculated for each variant, and due to the large size of the genome and number of possible disease models the probability of any particular variant being causally associated with the phenotype in question is low. A major advantage of the Bayesian approach is that, unlike the frequentist method, conducting

additional post-hoc analysis does not lead to a more stringent significance threshold because the a priori probability is unaffected. Wacholder et al employed a variation on Bayesian methods to calculate their false positive report probability (FPRP), the probability of no true association between a genetic association and disease given a statistically significant finding [98]. The FPRP takes into account the observed *P* value, the prior probability that the association is real, and the statistical power of the test. It is compared against a criterion for assessing whether a finding is important. If the prior probability is high then the FPRP can be markedly reduced by increasing sample size. However if prior probability is low the benefit of added power is minor. Wacholder et al argue that evaluating significance by *P* value alone, even if an association has a very low *P* value, is incorrect because if prior probability it low then FPRP may be high. Conversely, setting a very low *P* threshold for statistical significance can be unnecessarily conservative if prior probability is high.

1.4.7 Statistical power

Statistical power to detect a phenotype-genotype association is dependent upon the magnitude of effect, the frequency of causal allele(s), and the sample size. Moreover for indirect association it is not only the disease predisposing allele frequency that matters, but also the marker allele (or tag SNP) frequency, and the power to detect an association is greatest when these frequencies match [52]. The extent of LD also influences the likelihood of observing an association. However if the effect size is large this is less important with power being high even at low to moderate LD. Effect size for case control studies is measured as an odds ratio (OR), which estimates the odds of an individual in a given exposure group (i.e. with a certain genetic variant) being a case versus being a control. If the OR is significantly greater than 1 then the variant confers susceptibility to the disease, if it is significantly less than 1 then its effect is protective. Unfortunately large effects are usually rare. For common complex disease the causal variants that are likely to be observed are typically of low or moderate effect (OR < 2.0), so a reasonable level of LD is necessary as well as the disease allele being common and close to the marker allele frequency. These conditions translate to a feasible sample size of several thousand of cases and controls. Increasing sample size should lower the p-value of true positive results, enabling their detection, and raise the p-value of false positive results [99].

It is possible for variants of small effect to have large clinical implications by their interaction with other genetic variants (epistasis) or environmental factors. Analysis of

gene-gene and gene-environment interactions can be performed within usual GWAS of independent SNP effects, but greater sample sizes are required to achieve sufficient power. Moreover the problem of multiple testing is amplified, for example if a study tests 300,000 SNPs there are potentially ~100 billion pairwise tests of epistasis [44]. The identification of gene-environment interactions is especially important because the avoidance of an environmental exposure can prevent the detrimental modification of a gene, thus avoiding the final disease product (since the exposure and genetic variant must be present to produce disease) [100]. Measurement of environmental factors is prone to recall bias in case-control studies. Therefore a prospective cohort study design is preferable for assessment of gene-environment interactions. Data are currently being collected in large scale population cohorts which will make this possible, e.g. UK Biobank [101], Generation Scotland [102].

Manolio and colleagues examined power to detect environmental, genetic, gene-environment interaction, and gene-gene interaction effects for various ORs, disease incidence rates and sample sizes [100]. Estimated minimum detectable ORs were far greater for interaction effects than simple effects (Figure 1.10). This was expected because fewer participants in a sample will be exposed to both a genetic and an environmental risk factor, or two genetic risk factors, than a single genetic or environmental factor. For hypertension, assuming an incidence rate of 5%, a sample size of 200,000 is estimated to be more than sufficient to detect marginal effects and interactions with ORs of ~1.5, the magnitude predicted to be important in complex diseases based on prior evidence. However diseases that fall into the rare incidence category of 0.01% (e.g. Parkinson disease, schizophrenia) will require extremely large samples for there to be any chance of detecting interactions. Even with an as yet unfeasible sample of 1,000,000 follow-up would need to be far longer than 5 years to accrue enough incident cases of rare diseases.

For common complex diseases most published genetic effects have to date been modest (OR ~1.1 – 1.5) [91, 103]. An exception, which no doubt increased expectations of similar findings, is the association between *APOE4* and late-onset Alzheimer disease [104] for which the allelic OR is 3.3 [52]. As yet only a small percentage of the human genome has been subject to well-designed association study so it is unknown whether the published effect sizes are representative of the genome overall [45]. The effect sizes observed are expected due to the multifactorial nature of the diseases concerned, and individually translate to only a small increase in population absolute risk. Multiple common risk variants of small effect have been combined theoretically, however, to construct risk scores of greater practical significance. Studies of the distribution of genetic effect sizes in other species such as

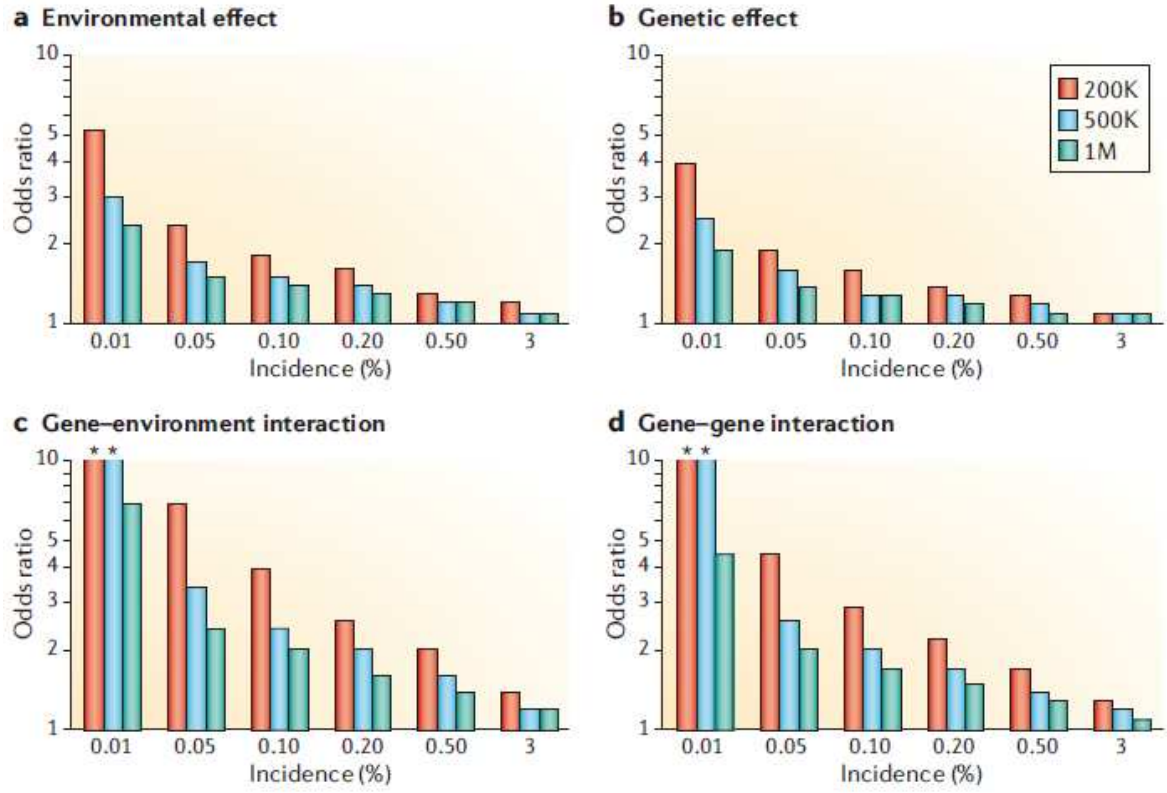


Figure 1.10 Sample-size requirements in prospective cohort studies. The estimated minimum detectable odds ratios after 5 years of follow-up for various cohort sizes and disease incidences are shown, assuming: 10% allele frequency for a dominant risk allele, 10% environmental exposure frequency, no prevalent cases in the cohort at the start of the study, 3% annual loss to follow-up, 80% power, and a type 1 error rate of 0.0001. Minimum odds ratios are shown for: a. an environmental exposure effect; b. a genetic effect (for a dominant variant); c. a gene-environment interaction, assuming genetic and environmental marginal effects of 1.5; d. a gene-gene interaction, assuming genetic and environmental marginal effects of 1.5. Asterisks indicate minimum detectable odds ratios in excess of 10. The colour key refers to sample size (reproduced from [100]).

rodents, *Drosophila melanogaster*, crops and livestock suggest that there will be few genetic loci of large effect and many loci of small effect [105-111]. This view is now widely accepted in the field of common disease genetics [112].

1.4.8 Replication

Many published association findings have failed to be replicated. This is partly due to the so-called "winner's curse", which is a bias whereby genetic effect size estimates are overestimated in initial discovery studies of disease-predisposing variants. Lohmueller and colleagues conducted a meta-analysis of 301 published studies covering 25 associations between common variants and common diseases [91]. They found that in 23 of the 25 initial studies there was inflation of effect size consistent with winner's curse. Similar biases were observed in an earlier meta-analysis of case-control association studies [113], and by linkage analysis of a quantitative trait locus (QTL) where data consisted of a genome wide simulation along with analytical results [114]. The degree of bias can be reduced or eliminated by increasing sample size, and several correction methods have been proposed (such as [115, 116]).

As with all fields of scientific literature there will exist a certain level of publication bias that leads to positive studies being reported more often than negative studies, with the consequence that a positive finding may occur by chance in an area where others have tried and failed to find anything of significance. However Lohmueller et al found that publication bias does not explain the level of replication observed in the literature [91], coming to the conclusion that there are many real associations reported as well as false positives. Technical bias can occur if cases and controls are not genotyped and analysed together in the same way. It is also thought that poor choice of controls and population stratification affect data quality and therefore play a part in replication failure.

Replication of GWAS findings is as important as replication of candidate gene associations, if not more so. The large number of SNPs studied in GWAS and resulting volume of statistical tests performed increases the likelihood of observing type I errors, i.e. false positives. In 2007 an NCI (National Cancer Institute) – NHGRI Working Group on Replication in Association Studies published an excellent summary of their recommendations on the reporting of initial association studies and criteria for replication [117]. These criteria include: replication studies should be of sufficient sample size; a similar population should be studied and ideally the same phenotype; similar

magnitude of effect and statistical significance in the same direction should be demonstrated; initially significance should be obtained using the same genetic model as the discovery study; a strong rationale should be provided for attempting replication of the chosen SNPs. There are certain situations in which there are insufficient participant numbers for replication, such as rare diseases or environmental exposures. These concerns do not affect most association studies of common diseases, though, so usually replication is advised.

Newton-Cheh and Hirschhorn summarised the three main reasons for failure to reproduce an initial significant finding when replication is attempted [99]:

- initial association is a false positive therefore correctly not replicated

- initial association is true but follow-up study underpowered to detect it

- association is true in one population but not another due to genetic or environmental heterogeneity

Meta-analyses that have examined replication failure indicate that in the majority of instances false positives are to blame [91, 113].

A strategy that increases power over that of individual studies and may be more cost effective than replication is meta-analysis of genome wide datasets. This collaborative way of working is increasingly common as investigators attempt to detect loci with smaller effects. One caveat is that genomic coverage does depend on the genotyping platform used, so retrospective meta-analysis where studies have used different platforms is less successful. Another is that the combination of results from multiple centres may increase genetic and environmental heterogeneity; consequently the gain in power may not be as great as one might expect [52]. Evangelou et al examined different meta-analytic strategies empirically with specific application to Parkinson disease [118]. They used three genome wide datasets, two of which were stages of the same study (referred to as Mayo tier 1 and Mayo tier 2) and the other a different study altogether (referred to as NINDS due to sponsorship by the National Institute of Neurological Disorders and Stroke). Three strategies for combining the datasets were considered: enhancement of replication data, where Mayo 1 was considered to be stage 1 analysis and Mayo 2 and NINDS combined to form an independent stage 2; enhancement of first-stage data, where Mayo 1 and NINDS combined as stage 1 and significant SNPs

were examined independently for replication in Mayo 2; and joint meta-analysis of all three datasets. Of the three strategies, the third of joint analysis proved to have the greatest power to detect associations between SNPs and Parkinson disease. However a major drawback was that there were only 527 SNPs available for study that were common to all datasets. This was despite ~200,000 SNPs passing quality control in the Mayo study and ~400,000 in NINDS, and highlights the difference in genotyping platform coverage. Between-study heterogeneity can also affect the results of meta-analyses and should be taken into account as much as possible [119].

1.4.9 GWAS of hypertension and blood pressure

To date there have been few GWAS of hypertension and/or blood pressure. Those studies that have been published have demonstrated very little success in identifying genetic variants that are associated with either hypertension, SBP or DBP. Many have not observed any SNPs that reached genome wide significance [120-124]. One possible reason for this failure in some studies is that hypertension and/or blood pressure were not the primary trait of interest [120], or not of a priori interest when the cohort was recruited [123] therefore phenotyping may not have been necessarily thorough. Another possibility is that there were not enough SNPs studied to provide sufficient coverage of the whole genome, in some studies fewer than 100,000 passing quality control measures [121, 122, 124].

Table 1.1 summarises the most significant hits from GWAS of hypertension and/or blood pressure published since 2007 (other than WTCCC, Global BPgen, and CHARGE), along with meta-analysed discovery and replication results if replication was attempted. One study conducted by Sabatti et al (2009) is omitted because the authors did not publish any results for blood pressure, instead reporting that analysis of blood pressure did not produce any genome-wide significant results. However it was not the primary trait of interest; the authors also studied triglycerides, HDL, LDL, CRP, glucose, insulin and BMI.

Levy et al (2007) and Wang et al (2009) failed to find any associations of genome-wide significance but had limited genomic coverage, studying just 70,897 and 79,447 SNPs respectively. Furthermore the discovery sample employed by Wang et al was small at 542 participants. Org et al (2009) and Cho et al (2009) also did not find any associations of genome-wide significance.

Table 1.1 Summary of recent GWAS of hypertension and/or blood pressure

			Discovery sample			Discovery & replication meta-analysis		
	Publication date	**Phenotype**	**N**	**OR/beta**	**Lowest *P*-value**	**N**	**OR/beta**	***P*-value**
Levy et al [121]	Sep-07	DBP	1233	-	3.31×10^{-6}	-	-	-
		SBP	1260	-	1.69×10^{-6}	-	-	-
*Wang et al [124]	Jan-09	SBP	542	-	7.6×10^{-5}	7125	1.9	1.6×10^{-7}
Org et al [125]	Mar-09	hypertension	364/596	0.49	2.34×10^{-6}	3808/4334	0.78	1.39×10^{-6}
Cho et al [126]	May-09	SBP	8842	-1.309	9.1×10^{-7}	16703	-1.064	1.3×10^{-7}
		DBP	8842	-0.882	1.2×10^{-6}	16703	-0.63	3.0×10^{-6}
Adeyemo et al [127]	Jul-09	SBP	1017	-	4.72×10^{-8}	-	-	-
		DBP	1017	-	0.448	1997	-	0.162
		hypertension	509/508	0.58	5.10×10^{-7}	875/1122	-	0.009

*** Observed a SNP with a lower p-value but did not report on it as situated in gene desert.**
- Value not reported (or in the case of meta-analysis replication not attempted).
OR = odds ratio
Effect sizes for associations with hypertension are presented as odds ratios, and for associations with continuous blood pressure as beta coefficients.

The vast majority of GWAS thus far have been conducted on samples of Caucasian individuals of European ancestry. One of the few studies to examine African Americans was that of Adeyemo et al (2009). Their initial findings were promising, however replication was either not attempted (in the case of SBP) or the replication findings were in the opposite direction of effect (DBP and hypertension).

Three recent studies conducted by large consortia, the WTCCC [55], the Global Blood Pressure Genetics Consortium (Global BPgen) [58], and the Cohorts for Heart and Aging Research in Genome Epidemiology Consortium (CHARGE) [59], were more wide-ranging in their scope and implications. They are described in detail below.

1.4.9.1 Wellcome Trust Case Control Consortium (WTCCC)

The WTCCC, made up of over 50 British research groups, conducted a GWA study of 2,000 cases each for 7 complex diseases of major public health importance; bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. These were compared with 3,000 shared common controls that came from two sources: 1,500 from the 1958 British Birth Cohort and 1,500 blood donors that were recruited for the project. The reason for the use of shared common controls was to reduce the huge cost of GWA, as described above. The 2,000 hypertension cases were unrelated participants from the BRIGHT study [128]. The primary aim of the study was to gain insights into the genetic contributions to each of the diseases, while at the same time discovering differences in allelic architecture between them. Furthermore, because GWAS is a comparatively new area of research, the study aimed to address methodological issues of relevance to all GWAS.

The majority of participants were self-reported white Europeans. All samples were genotyped with the GeneChip 500K Mapping Array Set, produced by Affymetrix, which comprises 500,568 SNPs. The average power of the study for SNPs with minor allele frequencies (MAFs) above 5% was estimated to be 43% for alleles with a relative risk of 1.3, increasing to 80% for a relative risk of 1.5, where the threshold for genome wide significance is $P < 5 \times 10^{-7}$. The investigators developed a new algorithm, CHIAMO, which was applied to simultaneously call the genotypes from all individuals. Of the total sample 809 individuals were excluded due to contamination, false identity, non-Caucasian ancestry, or relatedness, leaving 16,179 study participants. 469,577 SNPs (93.8%) passed quality control filters with an average call rate of 99.63%. Of those, 392,575 had study-

wide MAFs > 1%. All of the SNPs that passed quality control filters were used in the association analyses. Those that showed strong association underwent visual cluster plot inspection, identifying 638 SNPs with poor clustering that were removed.

The control groups were compared to assess bias in their ascertainment and processing (which differed). Few significant differences were found, indicating that there would be little bias introduced by the use of either of them as a control group for any of the case groups. Additionally it meant that they could be combined into a single control group of 3,000 individuals.

Samples were seeded with those from the HapMap panels and then examined for non-European ancestry using multidimensional scaling. This led to the exclusion of 153 individuals, of which many were people of South Asian origin in the diabetes case groups. Next the samples were split into 12 geographical regions of Great Britain defined by postcode, which were compared for allele frequency differences as evidence of population heterogeneity. Table 1.2 shows SNPs in thirteen genomic regions that were highly differentiated by geographical region. This would have been a cause for concern, but no associations were found within the regions so they were not of interest. Otherwise there was only a small effect of population structure found; therefore the association analyses were not adjusted.

The investigators used both frequentist and Bayesian statistical methods. Trend tests, genotype tests, and sex-differentiated tests were performed between each disease group and the pooled controls. Imputation analysis was also performed using HapMap reference samples to impute 2,193,483 SNPs not covered by the Affymetrix 500K chip.

The association analysis began with an investigation of 15 variants for which there was strong prior evidence of association with at least one of the diseases. These significant associations were replicated in the WTCCC sample for 13 of the variants, with effect sizes similar to those observed in previous studies (Table 1.3). Over the entire genome there were 21 SNPs identified with *P* values lower than the genome wide significance threshold of 5×10^{-7} (Figure 1.11). Of these 10 were known associations. Unfortunately, of the 7 diseases of interest, hypertension could be described as the loser in that it was not associated with any SNPs at $P < 5 \times 10^{-7}$. Moreover there was no evidence for any of the variants previously associated with hypertension (at least partly due to some not being well tagged by the Affymetrix chip, e.g. promoter of the *WNK1* (WNK lysine deficient protein

Table 1.2 WTCCC results: SNPs highly differentiated by graphical region [55]

Chromosome	Genes	Region (Mb)	SNP	Position	*P* value
2q21	*LCT*	135.16-136.82	rs1042712	136,379,576	5.54×10^{-13}
4p14	*TLR1, TLR6, TLR10*	38.51-38.74	rs7696175	386,43,552	1.51×10^{-12}
4p28		137.97-138.01	rs1460133	137,999,953	4.43×10^{-08}
6p25	*IRF4*	0.32-0.42	rs9378805	362,727	5.39×10^{-13}
6p21	*HLA*	31.10-31.55	rs3873375	31,359,339	1.07×10^{-11}
9p24	*DMRT1*	0.86-0.88	rs11790408	866,418	4.96×10^{-07}
11p15	*NAV2*	19.55-19.70	rs12295525	19,661,808	7.44×10^{-08}
11q13	*NADSYN1, DHCR7*	70.78-70.93	rs12797951	70,820,914	3.01×10^{-08}
12p13	*DYRK4, AKAP3, NDUFA9, RAD51AP1, GALNT8*	4.37-4.82	rs10774241	45,537,27	2.73×10^{-08}
14q12	*HECTD1, AP4S1, STRN3*	30.41-31.03	rs17449560	30,598,823	1.46×10^{-07}
19q13	*GIPR, SNRPD2, QPCTL, SIX5, DMPK, DMWD, RSHL1, SYMPK, FOXA3*	50.84-51.09	rs3760843	50,980,546	4.19×10^{-07}
20q12		38.30-38.77	rs2143877	38,526,309	1.12×10^{-09}
×p22		2.06-2.08	rs6644913	2,061,160	1.23×10^{-07}

Properties of SNPs that show large allele frequency differences between samples of individuals from 12 regions across Great Britain. Regions showing differentiated SNPs are given with details of the SNP with the smallest *P* value in each region for differentiation on the 11-d.f. test of differences in SNP allele frequencies between geographical regions, within the 9 collections. Positions are in NCBI build 35 coordinates. LCT = lactase. TLR1 = toll-like receptor 1. TLR6 = toll-like receptor 6. TLR10 = toll-like receptor 10. IRF4 = interferon regulatory factor 4. HLA = major histocompatibility complex. DMRT1 = doublesex and mab-3 related transcription factor 1. NAV2 = neuron navigator 2. NADSYN1 = NAD synthetase 1. DHCR7 = 7-dehydrocholesterol reductase. DYRK4 = dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 4. AKAP3 = A kinase (PRKA) anchor protein 3. NDUFA9 = NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9, 39kDa. RAD51AP1 = RAD51 associated protein 1. GALNT8 = UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 8 (GalNAc-T8). HECTD1 = HECT domain containing 1. AP4S1 = adaptor-related protein complex 4, sigma 1 subunit. STRN3 = striatin, calmodulin binding protein 3. GIPR = gastric inhibitory polypeptide receptor. SNRPD2 = small nuclear ribonucleoprotein D2 polypeptide 16.5kDa. QPCTL = glutaminyl-peptide cyclotransferase-like. SIX5 = SIX homeobox 5. DMPK = dystrophia myotonica-protein kinase. DMWD = dystrophia myotonica, WD repeat containing. RSHL1 = radial spokehead-like 1. SYMPK = symplekin. FOXA3 = forkhead box A3.

Table 1.3 WTCCC results: Evidence for signals of association at previously robustly replicated loci [55]

Collection	Genes	Chromosome	Reported SNP	WTCCC SNP	HapMap r^2	Trend *P* value	Genotypic *P* value
CAD	*APOE*	19q13	*	rs4420638	-	1.7×10^{-01}	1.7×10^{-01}
CD	*NOD2*	16q12	rs2066844	rs17221417	0.23	9.4×10^{-12}	4.0×10^{-11}
CD	*IL23R*	1p31	rs11209026	rs11805303	0.01	6.5×10^{-13}	5.9×10^{-12}
RA	*HLA-DRB1*	6p21	*	rs615672	-	2.6×10^{-27}	7.5×10^{-27}
RA	*PTPN22*	1p13	rs2476601	rs6679677	0.75	4.9×10^{-26}	5.6×10^{-25}
T1D	*HLA-DRB1*	6p21	*	rs9270986	-	4.0×10^{-116}	2.3×10^{-122}
T1D	*INS*	11p15	rs689	†	-	-	-
T1D	*CTLA4*	2q33	rs3087243	rs3087243	1	2.5×10^{-05}	1.8×10^{-05}
T1D	*PTPN22*	1p13	rs2476601	rs6679677	0.75	1.2×10^{-26}	5.4×10^{-26}
T1D	*IL2RA*	10p15	rs706778	rs2104286	0.25	8.0×10^{-06}	4.3×10^{-05}
T1D	*IFIH1*	2q24	rs1990760	rs3788964	0.26	1.9×10^{-03}	7.6×10^{-03}
T2D	*PPARG*	3p25	rs1801282	rs1801282	1	1.3×10^{-03}	5.4×10^{-03}
T2D	*KCNJ11*	11p15	rs5219	rs5215	0.9	1.3×10^{-03}	5.6×10^{-03}
T2D	*TCF7L2*	10q25	rs7903146	rs4506565	0.92	5.7×10^{-13}	5.1×10^{-12}

Where information on the strength of association at a particular SNP had been previously published and replicated the *P* value of both the trend and genotype test at the same SNP are tabulated, or the best tag SNP (defined to be the SNP with the highest r^2 with the reported SNP, calculated in the CEU sample of the HapMap project). *Previous reports relate to haplotypes rather than single SNPs. †Not well tagged by SNPs that pass the quality control. APOE = apolipoprotein E. NOD2 = nucleotide-binding oligomerization domain containing 2. IL23R = interleukin 23 receptor. HLA-DRB1 = major histocompatibility complex, class II, DR beta 1. PTPN22 = protein tyrosine phosphatase, non-receptor type 22. INS = insulin. CTLA4 = cytotoxic T-lymphocyte-associated protein 4. IL2RA = interleukin 2 receptor, alpha. IFIH1 = interferon induced with helicase C domain 1. PPARG = interferon induced with helicase C domain 1. KCNJ11 = potassium inwardly-rectifying channel, subfamily J, member 11. TCF7L2 = transcription factor 7-like 2 (T-cell specific, HMG-box).

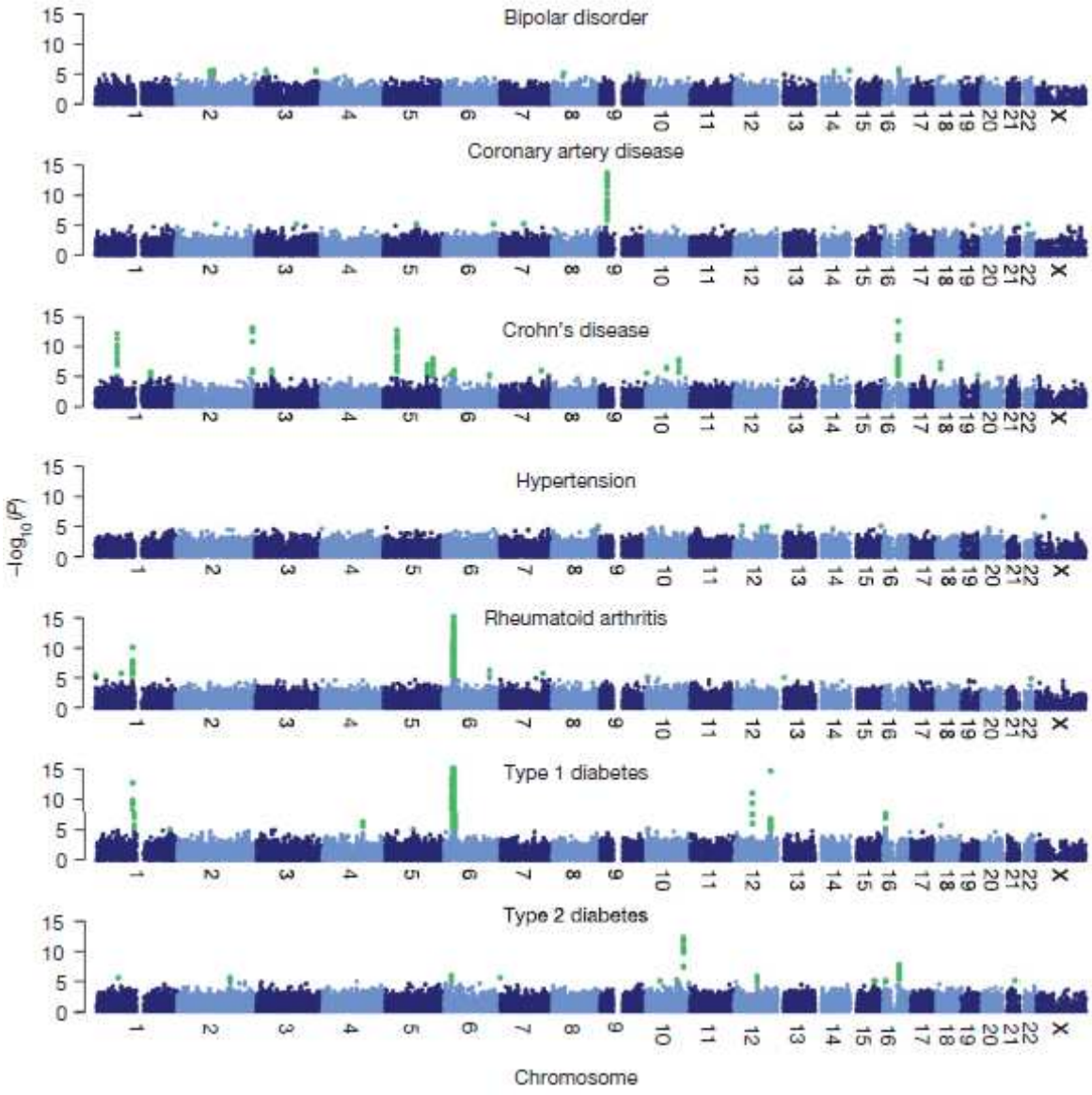


Figure 1.11 WTCCC results. For each of seven diseases –log10 of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values <1 x 10-5 highlighted in green. All panels are truncated at –log10(P value) = 15, although some markers (for example, in the major histocompatibility complex in type 1 diabetes and rheumatoid arthritis) exceed this significance threshold (reproduced from [55]).

kinase 1) gene). There were, however, a similar number and distribution of marginal results (with p-values between 10^{-4} and 10^{-7}) to the other case groups. It was speculated that the lack of a genome-wide significant result may have been due to poorly tagged variants or that hypertension may have few common risk alleles with larger effect sizes. Moreover, misclassification bias may have reduced the power of detecting effects. The common controls were not specifically phenotyped for blood pressure. Due to the high prevalence of hypertension and its existence on the continuum of blood pressure some of the controls may have been misclassified cases. The WTCCC estimated that the misclassification of 5% of controls (i.e. if 5% of controls were in fact undiagnosed cases) would translate to a loss of power equivalent to a 10% reduction in sample size. This is because of the dilution of any observable genetic difference, caused by the blurring of the distinction between cases and controls. Considering the expense of genome-wide association analysis and the anticipated relatively small effect sizes any reduction in power poses a serious problem. Moreover individuals with blood pressure in the mid-range of normotension (that is not considered to pose a risk clinically) may still be at increased risk in relation to individuals with low blood pressure.

Partly due to the failure to find any significant association with hypertension, there have been several recent reviews and commentaries published about the genetics of cardiovascular disease and how to proceed from here [47, 129-131]. One suggestion is to take a pathway approach that takes into account gene-gene and gene-environment interactions, although currently single studies are not powered for this [131].

The WTCCC study makes several recommendations for GWAS in general:

- Extensive data quality control checking is essential including visual inspection of cluster plots for SNPs of interest

- After excluding participants with non-European ancestry, and out with the loci listed in Table 1.2, the effects of population structure on case-control association are not sufficient to warrant concern of confounding by stratification (at least in Britain).

- The validity of common control group usage demonstrates that concern about matching of cases and controls (in genetic association studies) for socio-demographic variables has been exaggerated (again at least in Britain).

- The modest effect sizes observed are in line with the now widely held view that for common diseases there will be few large genetic effects ("low hanging fruit"), some modest effects, and many small (probably undetectable in the sample sizes of most studies) effects.

- Because estimates of variant-disease association effect size in initial discovery studies are often inflated ("winner's curse"), crucial replication attempts will need to employ study samples of even greater size.

1.4.9.2 Global Blood Pressure Genetics (Global BPgen) Consortium

The Global BPgen consortium, a collaboration of 17 GWAS from Europe and the USA (including the WTCCC 1958 British Birth Cohort control group), examined genetic associations with SBP and DBP [58]. Some of the component studies were population-based and some case-control. Participants were of European ancestry and, after exclusion of those who had been ascertained to their original study group on the basis of case status for hypertension, type 1 or 2 diabetes, or coronary artery disease (CAD), totalled 34,433. Investigators dealt with the confounding effect of antihypertensive medication by adding 15 mm Hg to recorded SBP and 10 mm Hg to DBP for those who were on such treatment. The genotyping platform used varied by study and included: Affymetrix 500K; Illumina 550K; AffymetrixSNP 5.0; Illumina HumanHap 300; Affymetrix 6.0; Illumina 370; Illumina 1M; Affymetrix 500; AffymetrixSNP 6.0; and Illumina 317K. Studies also varied in their participant ascertainment and method of blood pressure measurement.

In Stage 1, for each study ~2.5 million autosomal SNPs in the HapMap CEU sample (Caucasians in Utah, USA) were imputed and tested for association with SBP and DBP separately. At a significance threshold of $P < 5 \times 10^{-5}$ there were 11 independent signals for SBP and 15 for DBP, two of which achieved $P < 5 \times 10^{-8}$ (considered genome wide significance for this analysis).

Stage 2a of the study followed-up 12 SNPs from Stage 1. These were genotyped in 13 cohorts of European ancestry (N≤71,225), one of them the BRIGHT cohort, and one cohort of Indian Asian ancestry (N≤12,889). Finally, stage 2b was *in-silico* analysis of ten independent signals each for DBP and SBP in the CHARGE cohort (described in section 1.4.3 below). Meta-analysis of stages 1, 2a and 2b association results identified

associations of genome wide significance (for either DBP or SBP) at eight loci, listed in Table 1.4.

The results for the eight genome wide significant associations were compared for DBP and SBP for stage 1 (validation having only been attempted at stages 2a and 2b for the phenotype with the lower stage 1 *P* value). All eight showed some level of association with both phenotypes in the same direction of effect (Table 1.4). Investigators also assessed whether these loci were associated with hypertension, defined as SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or self-reportedly on antihypertensive medication. Normotension was defined as not taking antihypertensives, SBP ≤ 120 mm Hg and DBP ≤ 85 mm Hg. Stage 1 genome wide analysis was not conducted for hypertension due to lack of power, so instead the significant loci were examined in planned secondary analysis (N range = 57,410 – 99,802). In the secondary samples all eight alleles showed association with hypertension in the same direction of effect as continuous blood pressure (Table 1.5). However, in the stage 1 sample alone four of the SNPs had p-values in the range $0.01 < P \leq 0.10$.

In the Indian Asian sample two of the 12 stage 2a SNPs were replicated with $P < 0.01$. These were rs16998073 on chromosome 4q21 ($P = 5 \times 10^{-4}$) and rs11191548 on chromosome 10q24 ($P = 0.008$). It should be noted that all of the reported associations translate to a very small change in blood pressure, approximately 1 mm Hg per allele SBP or 0.5 mm Hg per allele DBP. However, the effects of multiple variants can be combined to produce a meaningful change in population cardiovascular risk.

1.4.9.3 Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium

The CHARGE consortium is made up of six population-based studies of individuals of European descent from Europe and the USA, N = 29,136 [59]. It conducted GWA studies of SBP, DBP and hypertension (defined as SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg or antihypertensive treatment). Participants taking antihypertensive medication had 10 mmHg added to their recorded SBP and 5 mmHg to their DBP. The genotyping platform used varied by study and included: Illumina 550K; Affymetrix 6.0; Illumina 370CNV; and Affymetrix 500K & MIPS 50K combined.

The association results for SBP (Table 1.6), DBP (Table 1.7) and hypertension (Table 1.8) are presented for SNPs of genome-wide significance, considered $P < 4 \times 10^{-7}$, along with

Table 1.4 Global BPgen results. Relationship of SNPs at 8 genome-wide significant loci to both blood pressure traits [58]

SNP ID	Chromosome	*N* (effective)	Trait	Beta mm Hg	s.e.	*P*
rs17367504	1	34,158	SBP	-0.79	0.18	1×10^{-05}
			DBP	**-0.50**	**0.12**	**3×10^{-05}**
rs11191548	10	33,123	SBP	1.17	0.22	3×10^{-07}
			DBP	**0.56**	**0.15**	**2×10^{-04}**
rs12946454	17	32,120	SBP	0.68	0.15	4×10^{-06}
			DBP	**0.34**	**0.09**	**6×10^{-04}**
rs16998073	4	26,106	DBP	0.65	0.11	7×10^{-09}
			SBP	**0.74**	**0.17**	**1×10^{-05}**
rs1530440	10	32,718	DBP	-0.51	0.11	3×10^{-06}
			SBP	**-0.43**	**0.16**	**7×10^{-03}**
rs653178	12	30,853	DBP	-0.46	0.09	1×10^{-07}
			SBP	**-0.47**	**0.13**	**3×10^{-04}**
rs1378942	15	34,126	DBP	0.48	0.09	6×10^{-08}
			SBP	**0.62**	**0.13**	**2×10^{-06}**
rs16948048	17	34,052	DBP	0.40	0.09	5×10^{-06}
			SBP	**0.41**	**0.13**	**2×10^{-03}**

For each of eight SNPs, the upper row shows association statistics for the blood pressure trait used for the analysis in which they were selected (SBP or DBP). The lower row (in boldface) shows the equivalent association statistics for the alternate blood pressure trait. Results are shown for the 34,433 individuals in the stage 1 Global BPgen GWAS samples.

Table 1.5 Global BPgen results. Association of eight SBP- and DBP- associated loci with hypertension [59]

SNP ID	Chromosome	Continuous Trait	Continuous BP effect	HTN OR	HTN 95% CI	HTN *P*	*N*
rs17367504	1	SBP	↓	0.89	0.86-0.93	2×10^{-09}	62,803
rs11191548	10	SBP	↑	1.16	1.11-1.21	3×10^{-13}	99,153
rs12946454	17	SBP	↑	1.07	1.04-1.11	2×10^{-05}	57,410
rs16998073	4	DBP	↑	1.10	1.07-1.13	7×10^{-10}	73,756
rs1530440	10	DBP	↓	0.95	0.91-0.98	2×10^{-03}	83,156
rs653178	12	DBP	↓	0.93	0.91-0.96	8×10^{-07}	60,030
rs1378942	15	DBP	↑	1.10	1.07-1.12	2×10^{-14}	99,802
rs16948048	17	DBP	↑	1.06	1.03-1.09	1×10^{-04}	62,411

Shown are the meta-analysis results for the top SNP from each genome-wide significant SBP or DBP locus from a logistic regression analysis of the odds of hypertension compared to normotension.
HTN = hypertension. OR = odds ratio.

Table 1.6 CHARGE results. Association of 13 loci significantly associated genome-wide with SBP, and corresponding results for DBP and hypertension [59]

		Meta-analysis, SBP			Meta-analysis, DBP			Meta-analysis, hypertension		
SNP ID	**Chromosome**	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***
rs2681492	12	-1.26	0.19	3.0×10^{-11}	-0.62	0.11	4.6×10^{-08}	-0.14	0.03	8.4×10^{-08}
rs2681472	12	-1.29	0.19	3.5×10^{-11}	-0.64	0.11	3.7×10^{-08}	-0.16	0.03	1.7×10^{-08}
rs11105354	12	-1.30	0.20	3.7×10^{-11}	-0.63	0.11	5.8×10^{-08}	-0.16	0.03	1.8×10^{-08}
rs11105364	12	-1.30	0.20	4.8×10^{-11}	-0.63	0.12	1.2×10^{-07}	-0.16	0.03	2.1×10^{-08}
rs17249754	12	-1.30	0.20	5.2×10^{-11}	-0.63	0.12	1.0×10^{-07}	-0.16	0.03	2.2×10^{-08}
rs11105368	12	-1.30	0.20	5.3×10^{-11}	-0.63	0.12	1.3×10^{-07}	-0.16	0.03	2.2×10^{-08}
rs12579302	12	-1.29	0.20	6.2×10^{-11}	-0.62	0.12	1.3×10^{-07}	-0.16	0.03	2.2×10^{-08}
rs12230074	12	-1.31	0.20	9.1×10^{-11}	-0.62	0.12	3.4×10^{-07}	-0.17	0.03	2.9×10^{-08}
rs11105378	12	-1.31	0.20	9.1×10^{-11}	-0.62	0.12	3.1×10^{-07}	-0.17	0.03	2.8×10^{-08}
rs4842666	12	-1.20	0.21	6.5×10^{-09}	-0.62	0.12	4.5×10^{-07}	-0.15	0.03	3.4×10^{-07}
rs8096897	18	-12.87	2.33	3.2×10^{-08}	-4.07	1.33	2.9×10^{-03}	-0.73	0.35	0.04
rs11105328	12	-1.11	0.20	4.2×10^{-08}	-0.61	0.12	5.1×10^{-07}	-0.15	0.03	7.1×10^{-07}
rs880315	1	0.89	0.17	2.1×10^{-07}	0.30	0.10	2.9×10^{-03}	0.09	0.02	6.2×10^{-05}

Beta is the effect size on blood pressure in mmHg, per allele based on the additive genetic model.

Table 1.7 CHARGE results. Association of 20 loci significantly associated genome-wide with DBP, and corresponding results for SBP and hypertension [59]

		Meta-analysis, DBP			Meta-analysis, SBP			Meta-analysis, hypertension		
SNP ID	**Chromosome**	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***
rs3184504	12	0.50	0.09	1.7×10^{-08}	0.75	0.15	5.7×10^{-07}	0.07	0.02	7.4×10^{-04}
rs653178	12	0.50	0.09	2.0×10^{-08}	0.74	0.15	8.5×10^{-07}	0.07	0.02	7.7×10^{-04}
rs2681472	12	-0.64	0.12	3.7×10^{-08}	-1.29	0.19	3.5×10^{-11}	-0.16	0.03	1.7×10^{-08}
rs4766578	12	0.49	0.09	4.2×10^{-08}	0.73	0.15	1.2×10^{-06}	0.06	0.02	1.9×10^{-03}
rs10774625	12	0.49	0.09	4.2×10^{-08}	0.73	0.15	1.1×10^{-06}	0.06	0.02	1.8×10^{-03}
rs2681492	12	-0.62	0.11	4.6×10^{-08}	-1.26	0.18	3.0×10^{-11}	-0.14	0.03	8.4×10^{-08}
rs11105354	12	-0.63	0.12	5.8×10^{-08}	-1.30	0.19	3.7×10^{-11}	-0.16	0.03	1.8×10^{-08}
rs17630235	12	0.50	0.09	1.0×10^{-07}	0.69	0.15	1.1×10^{-05}	0.06	0.02	4.3×10^{-03}
rs17249754	12	-0.63	0.12	1.0×10^{-07}	-1.30	0.19	5.2×10^{-11}	-0.16	0.03	2.2×10^{-08}
rs11066188	12	0.50	0.09	1.1×10^{-07}	0.68	0.15	1.3×10^{-05}	0.06	0.02	4.2×10^{-03}
rs11105364	12	-0.63	0.12	1.2×10^{-07}	-1.30	0.19	4.8×10^{-11}	-0.16	0.03	2.1×10^{-08}
rs11105368	12	-0.63	0.12	1.2×10^{-07}	-1.30	0.19	5.3×10^{-11}	-0.16	0.03	2.2×10^{-08}
rs12579302	12	-0.62	0.12	1.2×10^{-07}	-1.29	0.19	6.2×10^{-11}	-0.16	0.03	2.2×10^{-08}
rs2384550	12	-0.48	0.09	1.3×10^{-07}	-0.71	0.15	4.3×10^{-06}	-0.08	0.02	5.6×10^{-05}
rs1991391	12	-0.48	0.09	1.4×10^{-07}	-0.71	0.15	3.8×10^{-06}	-0.09	0.02	5.6×10^{-05}
rs6489992	12	-0.48	0.09	2.0×10^{-07}	-0.71	0.15	4.7×10^{-06}	-0.08	0.02	1.9×10^{-04}
rs11065987	12	0.48	0.09	2.2×10^{-07}	0.70	0.15	9.4×10^{-06}	0.06	0.02	4.1×10^{-03}
rs11024074	11	0.50	0.10	2.8×10^{-07}	0.79	0.16	1.6×10^{-06}	0.09	0.02	5.2×10^{-05}
rs11105378	12	-0.62	0.12	3.1×10^{-07}	-1.31	0.20	9.1×10^{-11}	-0.17	0.03	2.8×10^{-08}
rs12230074	12	-0.62	0.12	3.4×10^{-07}	-1.31	0.20	9.1×10^{-11}	-0.17	0.03	2.9×10^{-08}

Beta is the effect size on blood pressure in mmHg, per allele based on the additive genetic model.

Table 1.8 CHARGE results. Association of 10 loci significantly associated genome-wide with hypertension, and corresponding results for SBP and DBP [59]

		Meta-analysis, SBP			Meta-analysis, DBP			Meta-analysis, hypertension		
SNP ID	**Chromosome**	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***	**Beta**	**s.e.**	***P***
rs2681472	12	-0.16	0.03	1.7×10^{-08}	-1.29	0.19	3.5×10^{-11}	-0.64	0.11	3.7×10^{-08}
rs11105354	12	-0.16	0.03	1.8×10^{-08}	-1.30	0.19	3.7×10^{-11}	-0.63	0.11	5.8×10^{-08}
rs11105364	12	-0.16	0.03	2.1×10^{-08}	-1.30	0.19	4.8×10^{-11}	-0.63	0.12	1.2×10^{-07}
rs17249754	12	-0.16	0.03	2.2×10^{-08}	-1.30	0.19	5.2×10^{-11}	-0.63	0.12	1.0×10^{-07}
rs11105368	12	-0.16	0.03	2.2×10^{-08}	-1.30	0.19	5.3×10^{-11}	-0.63	0.12	1.2×10^{-07}
rs12579302	12	-0.16	0.03	2.2×10^{-08}	-1.29	0.19	6.2×10^{-11}	-0.62	0.12	1.2×10^{-07}
rs11105378	12	-0.17	0.03	2.8×10^{-08}	-1.31	0.20	9.1×10^{-11}	-0.62	0.12	3.1×10^{-07}
rs12230074	12	-0.17	0.03	2.8×10^{-08}	-1.31	0.20	9.1×10^{-11}	-0.62	0.12	3.4×10^{-07}
rs2681492	12	-0.14	0.03	8.4×10^{-08}	-1.26	0.18	3.0×10^{-11}	-0.62	0.11	4.6×10^{-08}
rs4842666	12	-0.15	0.03	3.4×10^{-07}	-1.20	0.20	6.5×10^{-09}	-0.62	0.12	4.5×10^{-07}

Beta is the effect size on blood pressure in mmHg, per allele based on the additive genetic model.

the results for the other two phenotypes. Taking this significance threshold there were 13 significant SNP associations for SBP, 20 for DBP, and ten for hypertension. There is quite a bit of overlap between phenotypes with many of the top hits attaining significance in the same direction of effect for more than one phenotype. The top ten loci for SBP, DBP and hypertension (30 in total) in the CHARGE cohort were checked for significance in the Global BPgen results (described above). One SNP for SBP, four for DBP and one for hypertension were assessed for independent replication in Global BPgen. Five SNPs out of this six attained P <0.008, the threshold for external replication in Global BPgen. When the results for the same 30 SNPs in both studies were analysed together by meta-analysis, there were four associations of genome wide significance ($P < 5 \times 10^{-8}$) for SBP, six for DBP, and one for hypertension. Again effect sizes were very small, approximately 1 mm Hg change in SBP per allele or 0.5 mm Hg change in DBP per allele.

1.4.9.4 Potential candidate genes for blood pressure regulation identified by Global BPgen and/or CHARGE

Several of the loci associated with SBP, DBP, or hypertension in the Global BPgen study and/or the CHARGE study are potential candidate genes for blood pressure regulation. For example, *CYP17A1* (cytochrome P450, family 17, subfamily A, polypeptide 1) on chromosome 10q24, implicated in both studies, has been associated with a rare Mendelian form of hypertension [132]. Furthermore, the protein encoded is involved in the biosynthesis of mineralocorticoids and glucocorticoids that affect sodium handling. In a region of chromosome 1p36, identified in Global BPgen, lies *NPPA* (natriuretic peptide precursor A) which has previously been associated with blood pressure and hypertension [133]. Another possibility is *ATP2B1* (ATPase, Ca++ transporting, plasma membrane 1) on chromosome 12q21, variants in which were associated with all three traits in CHARGE. Elevated mRNA levels of the encoded protein, PMCA1, have been found in aortic smooth muscle cells of spontaneously hypertensive rats compared with controls [134]. Global BPgen identified *FGF5* (fibroblast growth factor 5) on chromosome 4q21 as a possible candidate gene because it has been associated with angiogenesis in the heart [135]. A missense SNP located in exon 3 of *SH2B3* (SH2B adaptor protein 3) on chromosome 12q24, and implicated in both studies, has been previously associated with type 1 diabetes [136, 137], celiac disease [137, 138], and MI [139].

1.5 Aims

The specific aims of this study are:

1) to identify common polymorphisms associated with hypertension using a genome wide association approach on the extremes of the blood pressure distribution;

2) to validate the top hit(s) in independent cohorts using a similar strategy and perform a meta-analysis of the combined samples;

3) to elucidate the possible functional underpinnings of the validated hit(s).

2 Materials and methods

2.1 Description of samples

Cases and controls were both sampled from existing cohorts in Norway and Sweden: cases from the Nordic Diltiazem study (NORDIL) [140, 141]; and controls from the Malmö Diet and Cancer Study (MDC) [142-144].

2.1.1 Nordic Diltiazem study (NORDIL)

The NORDIL study was a prospective randomised controlled trial of diltiazem, a calcium antagonist, versus conventional (at time of recruitment) antihypertensive treatment [141]. Conventional treatment was mainly considered to be diuretics and/or beta-blockers, although participants could also be prescribed other classes of drug. Recruitment began in September 1992 and took place in 1032 health centres in Norway and Sweden. The primary endpoints considered were CV mortality defined as fatal acute myocardial infarction (MI), fatal acute stroke, sudden death and other fatal CVD; and CV morbidity defined as MI and stroke. Secondary endpoints were total mortality, development or deterioration of ischaemic heart disease (IHD), congestive heart failure, atrial fibrillation, transient ischaemic attacks, diabetes mellitus and renal insufficiency. Participants were hypertensive patients aged between 50 and 69 years at recruitment, with an untreated DBP of at least 100 mmHg during a run-in period without antihypertensive treatment. Previously treated and untreated patients were included. The final sample size was 10,881, of whom 5410 participants were randomised to diltiazem and 5471 to diuretics/beta-blockers. In both groups a little over 51% of individuals were female and the mean age was 60. At baseline mean SBP was ~173 mmHg and mean DBP ~106 mmHg. During the mean follow-up period of 4.5 years this reduced to 154.9/88.6 mmHg in the diltiazem group and 151.7/88.7 mmHg in the diuretic/beta-blocker group. Survival analysis showed that the only endpoint for which there was a significant difference between groups was all stroke, with a lower risk of events in the diltiazem group (RR 0.80, 95% CI 0.65-0.99, p=0.04). For all other endpoints there was no significant difference in risk observed between treatment groups. Examination of the 12 most frequently reported adverse effects found that those in the diltiazem group were more likely to experience headaches but less likely to experience fatigue, dyspnoea or impotence. In the current study participants were selected from the 5,280 Swedish NORDIL patients.

2.1.2 Malmö Diet and Cancer Study (MDC)

The MDC was established as a resource to examine the relationship between diet and the subsequent development of cancer [145]. It is a population based sample of 28,098 individuals aged 40-70 years living in the Swedish city of Malmö (total population 235,000). Recruitment occurred between 1991 and 1996. Blood pressure was measured twice in the supine position at recruitment, and the mean of the measurements recorded. As well as detailed dietary data, the study collected additional information on medical history, medication, anthropometry, and covariates such as alcohol consumption, smoking, physical activity, weight and socioeconomic category. For cardiovascular endpoints (defined as fatal and non-fatal coronary events and stroke) participants have been followed-up for 10.5 years through routine data linkage. There were 860 prevalent cardiovascular events at baseline and 2,100 incident events during follow-up. The age range of the MDC and NORDIL studies is similar and recruitment occurred during the same period of time.

2.1.3 Use of hypercontrols

As mentioned in Chapter 1, one of the reasons proposed for the failure of the WTCCC and other studies to find a genome-wide significant result for hypertension may have been misclassification bias. This is particularly likely for hypertension in comparison with other common diseases, because of its high prevalence and the continuum of risk conferred by elevated blood pressure. Many published studies have not phenotyped controls as carefully as they have cases and the use of retrospectively collected common controls amplifies this issue. In the WTCCC it was estimated that the misclassification of 5% of controls (i.e. if 5% of controls were in fact undiagnosed cases) would translate to a loss of power equivalent to a 10% reduction in sample size [55]. This is because of the dilution of any observable genetic difference, caused by the blurring of the distinction between cases and controls. Considering the expense of genome-wide association analysis and the anticipated relatively small effect sizes any reduction in power poses a serious problem. Moreover individuals with blood pressure in the mid-range of normotension that is not considered to pose a risk clinically may still be at increased risk in relation to individuals with low blood pressure. For this reason and to increase the likelihood of detecting genetic effects we advocate the novel approach of using hypercontrols [47]. Hence the current study compares cases and controls at the extreme high and low ends, respectively, of the blood pressure distribution.

2.1.4 Case inclusion criteria

Cases, selected from the NORDIL study sample, were defined as individuals younger than 60 years with at least two consecutive measurements of SBP ≥ 160 mmHg or DBP ≥ 100 mmHg. The blood pressure readings were taken while participants were off treatment, following a wash-out period of one week. According to these criteria 2,000 cases were identified, representing the top 1.7% of the Swedish population blood pressure distribution (Figure 2.1).

2.1.5 Control inclusion and exclusion criteria

Controls, collected from the MDC sample, were defined as individuals aged at least 50 years with SBP ≤ 120 mmHg and DBP ≤ 80 mmHg who were not prescribed any blood pressure lowering medication. The exclusion criterion was any evidence of cardiovascular disease, defined as no prevalent CAD or stroke and no incident CAD or stroke in the last ten years. 2585 individuals meeting these criteria were identified in the MDC sample, representing the lower 9.2% of the Swedish population blood pressure distribution (Figure 2.1). This does not reflect very extreme low blood pressure (unlike the cases who have very extreme high blood pressure) because in the adult population the blood pressure distribution curve is skewed to the right. Therefore, using the same percentage cut-off at both tails would provide fewer controls than cases. Instead we have selected the bottom 9.2% in order to sample at least 2000 controls (compared with top 1.7% for cases) and have enhanced the definition by utilising ten year follow-up information to exclude prevalent disease.

2.2 Software

Phenotypic data were summarised and analysed using SPSS 15.0 [146] and Stata 10 [147]. Genetic data were analysed using the program described below.

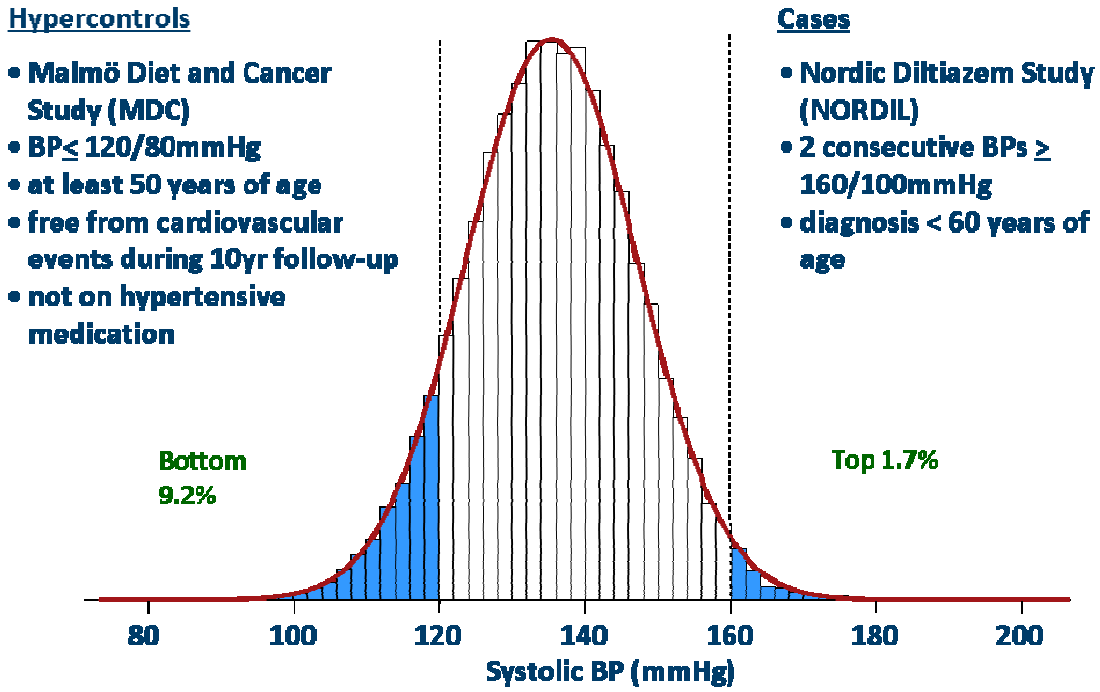


Figure 2.1 Blood pressure distribution in the current study sample.

2.2.1 DNA extraction and genotyping

High quality DNA was extracted in Glasgow for all participants on an Autopure LS DNA extraction system from Qiagen. The Autopure LS workstation allows fully automated high throughput purification of genomic DNA. A NanoDrop at OD260/280 is used to assess quantification and quality of DNA. Samples were genotyped using Illumina Infinium DNA Analysis BeadChips [148]. Due to cost and availability considerations the current analysis used three BeadChip types consecutively; HumanHap550, HumanHap550-Duo, and Human610-Quad. The HumanHap550 chip genotypes a single sample for more than 555,000 SNPs; this amounts to 90% of all Phase I and II CEU (Caucasian) HapMap loci with MAF $\geq$ 0.05. The same SNPs are covered by the HumanHap550-Duo chip but there is the added advantage that two samples can be genotyped concurrently. Furthermore it provides deep coverage of over 2,900 copy number variant (CNV) regions. The Human610-Quad chip also covers > 555,000 SNPs and an additional ~60,000 CNV specific markers, and as indicated by its name can genotype four samples simultaneously. Both BeadChip versions are reported by Illumina to have high validity and reliability with average call rate, reproducibility and HapMap concordance all > 99%. Half of the samples were genotyped at the British Heart Foundation Glasgow Cardiovascular Research Centre, and half at the Istituto Auxologico Italiano, Milan, Italy.

2.2.2 PLINK

PLINK [149, 150] is an open-source tool set for whole genome association and population-based linkage analyses. PLINK enables the user to manage large genome-wide datasets and perform standard summary statistics and association analyses. It offers tests of confounding due to both population stratification and non-random genotyping failure, and a method to assay rare variation with the use of common SNP panels. Furthermore gene-gene and gene-environment interactions can be assessed as well as copy number variation. There is not a fixed limit to the number of samples or SNPs, or indeed overall size of data file that PLINK can cope with, but computer processing capability limits what is possible. Analysis can be run separately for each chromosome if the machine in use has insufficient memory for the whole dataset. In addition analyses can be run on parallel processors to reduce computational time.

In order to run analysis PLINK requires data in the format of two input files: PED and MAP. PED files contain six mandatory columns: Family ID; Individual ID; Paternal ID;

Maternal ID; Sex; and Phenotype. Phenotype can be a single affection status or single quantitative trait. Genotypes are stored in column 7 and onwards, and markers must be biallelic. Flags can be used within analysis code to specify that family ID, parents' identities, sex or phenotype are unknown. The simplest file format possible is an individual ID followed by genotypic data. In MAP files each row represents a single marker and four columns are required: Chromosome; rs number or SNP identifier; Genetic distance; and Base-pair position. It is possible to use a flag to indicate that genetic distance is excluded as for many analyses it is not necessary. PLINK files do not contain column headings.

In the current study we have used binary PED (BED) files. These are smaller than PED files because pedigree/phenotype information is stored in a separate file (FAM), hence analysis is accelerated. BED files contain the chromosome name, the start and end positions of the feature (in this case a SNP), and binary genotype information, and are accompanied by a FAM file and a BIM file. FAM files contain the first six columns of PED files (i.e. they are PED files without genotypes). BIM files are extended MAP files that contain the same four columns with the addition of two further columns containing allele names.

For reasons of computational power the analyses reported here were performed on a remote server via the open-source Telnet/SSH client PuTTY (http://www.chiark.greenend.org.uk/~sgtatham/putty/) and files were managed remotely using WinSCP (http://winscp.net/eng/index.php), an open-source SFTP, FTP and SCP client for Windows. A useful aspect of PLINK is the ability to view genome-wide output in Haploview [151] and in tables and figures created in R [152]. Both of these facilities have been employed in the current study.

2.2.3 Haploview

Haploview [151] is a software package that provides tools for haplotype analysis. Like PLINK it can perform single marker association analysis and quality control analysis of markers. Moreover it generates LD information, haplotype blocks and population haplotype frequencies. As mentioned above there is the ability to import PLINK association and quality-control results into Haploview in order to create LD plots and Manhattan plots amongst other things. Several pairwise measures of LD are calculated and the user has a choice of LD block definitions to base plots on. Data can be sorted and filtered based on parameters such as association p-value, GC adjusted p-value,

missingness, MAF, Hardy-Weinberg frequency etc. If there is information on affection status in the input file Haploview can calculate the χ^2 statistic for case/control data or the TDT statistic for trio data.

2.2.4 EIGENSTRAT

As described in Chapter 1, PCA has been employed as an explicit method of detecting and correcting population stratification in GWAS. EIGENSTRAT [74] is a program which does this in three steps:

1. PCA is applied to genotype data to infer continuous axes of genetic variation

2. genotypes and phenotypes are continuously adjusted by amounts attributable to ancestry along each axis, via computing residuals of linear regressions

3. association statistics are computed using ancestry-adjusted genotypes and phenotypes

The axes of variation aim to describe the maximum amount of data variability possible in a small number of dimensions. These dimensions are principal components (PCs), termed "eigenvectors", and are rated in EIGENSTRAT output based on the amount of data variability explained, enabling the user to select the "top" eigenvectors to adjust association analysis for. If ancestry differences exist between samples then the axes may be interpreted geographically (e.g. north to south, east to west). It should be noted that EIGENSTRAT requires genome-wide data for population stratification to be assessed.

The developers of EIGENSTRAT have run simulations to determine how much data are required to accurately infer population structure and then correct for stratification. More data are needed for correction than for detection alone. Correction was found to be insensitive to number of samples, being effective in sample sizes as small as 100. When sample size was fixed at 1000 and $F_{ST} = 0.005$ full correction of stratification at highly differentiated SNPs required 20,000 SNPs. This rose to 100,000 SNPs when $F_{ST} = 0.001$ (the smaller value indicating that allele frequencies within the populations being compared are more similar). Hence the magnitude of currently reported datasets, of thousands of participants and hundreds of thousands of SNPs, is sufficient for EIGENSTRAT use. F_{ST} is defined as "the correlation between gametes chosen randomly from within the same

subpopulation relative to the entire population"[153]. It is used as a measure of genetic differentiation.

An additional benefit of using EIGENSTRAT is that it can detect problems with experimental design, for example if the laboratory treatment of cases and controls is not fully matched. Any resultant bias is potentially a far greater concern than population stratification. In an example outlined by its authors, Price and colleagues, the two most significant PCs produced by EIGENSTRAT described ancestry effects with the third detecting subtle differences in laboratory treatment among samples [74].

The current study samples are all Swedish Caucasian, so it is likely that any population structure present is low level and caused by mixed European ancestry. Price and colleagues studied six European population samples and an American Ashkenazi Jewish sample using EIGENSTRAT [154]. The most significant axes clustered individuals into three groups, approximately representing northwest European, southeast European, and Ashkenazi Jewish ancestry. The Swedish sample clustered in the northwest European group along with the UK and Poland. In a similar study of samples from 16 European countries Nelis and colleagues also observed a northwest to southeast gradient [155]. Figure 2.2 is a scatterplot of the first two axes of variation in 5,847 European samples in an analysis conducted by Heath et al [156]. Their data most clearly demonstrates a correlation between geographic origin and genetic origin, with Swedish participants clustering close to Norwegians and Germans. Furthermore, when compared with the STRUCTURE method of detecting population structure, EIGENSTRAT was superior in assigning origins to unknown samples [156]. The HapMap CEU samples were most similar to participants from the UK, followed by Germany, Belgium, Norway and Sweden. There was no evidence of non-European origin in any of the CEU samples.

2.3 Statistical analysis

2.3.1 Power calculations

Sample size was calculated using PBAT (http://www.goldenhelix.com/SNP_Variation/Manual/svs7/pbat_power_calculations.html), for 500,000 SNPs with 80% power for various odds ratios, assuming an equal number of

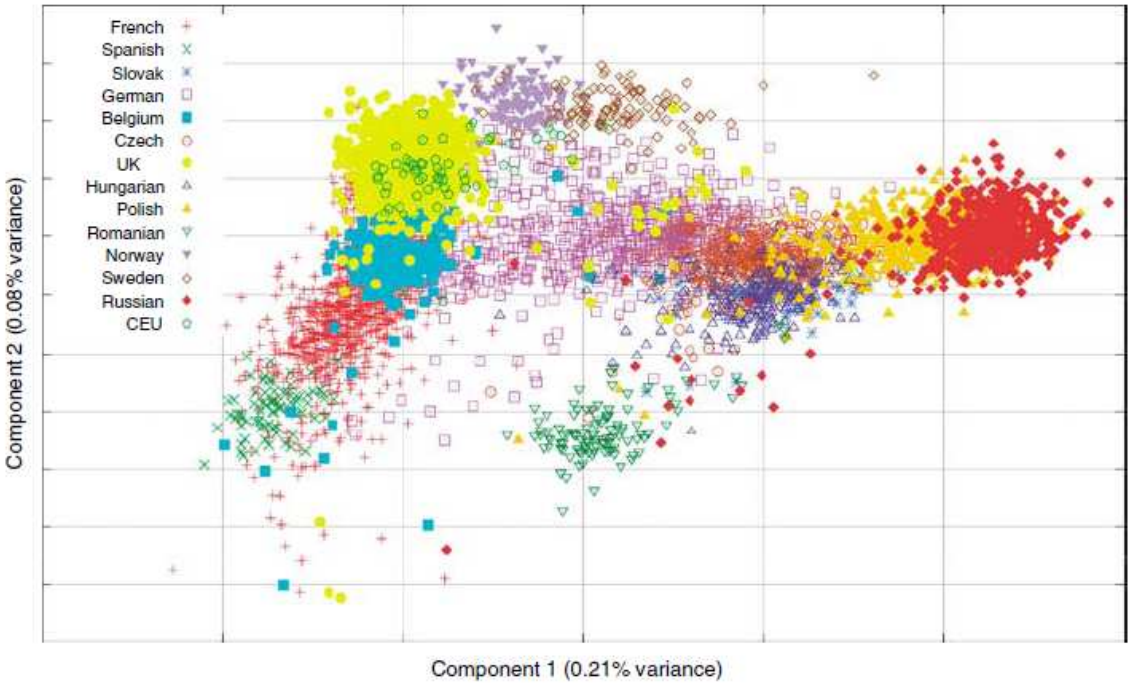


Figure 2.2 The top two axes of variation of a dataset of diverse European samples. Demonstrating both an East-West and a North-South gradient (reproduced from [156]).

cases and controls (Figure 2.3 [47]). Sample size estimates were modelled using Monte-Carlo simulations, with 1,000,000 simulations computed. Based on the blood pressure distribution in the general population and the commonly used hypertension cut-off of DBP ≥90mmHg or SBP ≥140mmHg, the expected OR is less than 1.3. At an MAF of 0.01 this translates to a sample size of at least 6,000 cases and as many controls. However high-fidelity phenotyping of extreme cases and controls will artificially inflate the OR to an estimated 1.5, meaning only ~2,000 cases and ~2,000 controls are required for 80% power to detect any effects. Furthermore the recruitment of both cases and controls from the same country, Sweden, should minimise the confounding effect of population structure and the resultant loss of power.

2.3.2 Summary of phenotypic data

Phenotypic data were analysed using SPSS 15.0. Information was available on age, BMI, SBP and DBP. Cases and controls were compared with the two sample t-test.

2.3.3 Reformatting of Illumina output files

Illumina report files contain nine rows of descriptive information followed by four column headings; sample ID, SNP name, allele 1, and allele 2. Beneath these headings the data are formatted as a row for each individual for each SNP. Report files were converted to LGEN (long-format) PLINK files by the removal of the first ten rows (descriptive information and column headings) and the addition of a fifth column for family ID. In this study there is no family ID since the samples are unrelated individuals. Therefore to meet PLINK file requirements the individual ID column was duplicated. This process was carried out separately for the Illumina report files generated by the three BeadChip types; HumanHap550, HumanHap550-Duo, and Human610-Quad. To differentiate between data genotyped by the different chip types we termed them “Singles”, “Duos”, and “Quads”, respectively. The three resulting LGEN files were merged longitudinally. Finally this LGEN file for the total sample of all cases and controls was converted to BED format with the use of a FAM file listing individual IDs and a MAP file listing marker information (for those common to all three chip types).

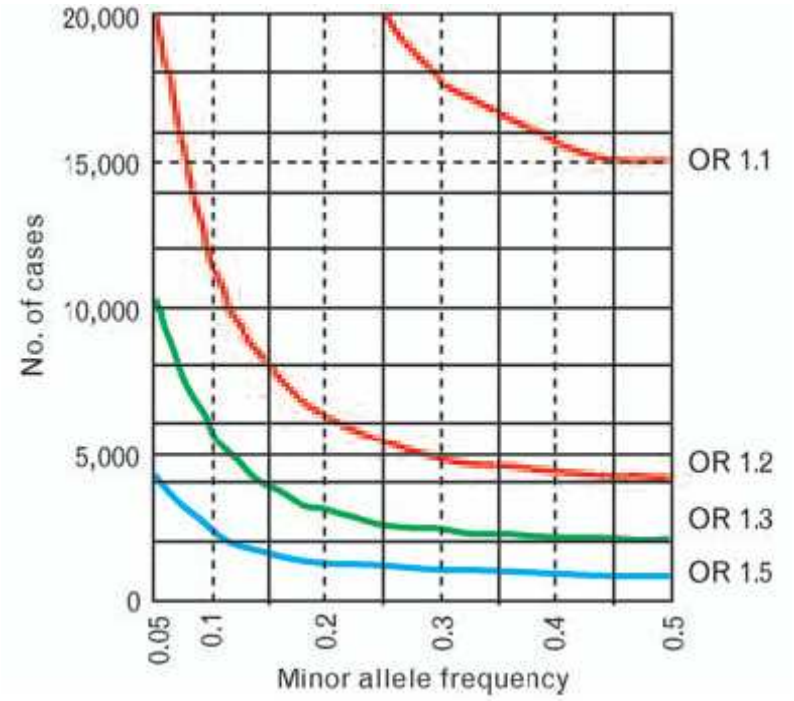


Figure 2.3 Sample size for a case-control genome-wide association study. Using 500,000 SNPs with 80% power for various odds ratios assuming an equal number of cases and controls; prevalence = 30%; P = 5 × 10-7 (reproduced from [47]).

2.3.4 Quality control

Rates of missingness were assessed in PLINK. This produces two output files; IMISS which gives missing rates per individual, and LMISS which gives rates per marker. Columns include the number missing, number genotyped, and missing as a proportion. Hardy-Weinberg test statistics for each SNP were computed in PLINK. The resultant output file contains genotype counts for both homozygotes and the heterozygote, observed heterozygosity, expected heterozygosity, and Hardy-Weinberg p-value.

The allele frequency command was run in PLINK which generates an output file with minor allele frequencies for each SNP.

2.3.5 Assessment of population stratification

The degree of population stratification present in the sample was assessed in various ways. Initially PLINK methods were used that are based on the average proportion of alleles shared identical by state (IBS) between any two individuals genome-wide. As opposed to identical by descent (IBD) alleles which are identical copies of the same ancestral allele (i.e. inherited from the same parent in the case of family studies), those that are IBS have the same DNA sequence but are not derived from a known common ancestor. In the case of monozygotic twins 100% of the genome is shared IBD. On average a parent shares 50% of the genome IBD with each offspring.

The first method employed in PLINK uses complete-linkage hierarchical clustering that clusters individuals into homogeneous subsets. The process begins by considering each individual as a separate cluster. The two closest clusters, i.e. those with the highest proportion of the genome shared IBS, are then repeatedly merged. This process continues until all individuals belong to one cluster, or else merging stops according to prespecified constraints.

The second produces multidimensional scaling (MDS) plots. These provide a visual representation of any substructure rather than clustering participants into groups (above). The dimensions of the representation can be included as covariates in association analysis to adjust for population stratification.

The final method identifies participants who are outliers. The IBS distance between each individual and its nearest neighbour is calculated. The distribution of all distances is then standardised and inspected for outliers, defined as individuals whose nearest neighbour is far less near than the average nearest neighbour.

EIGENSTRAT was also used to evaluate stratification. Beforehand the current samples of cases and controls were merged with a sample of 988 individuals from the International HapMap Project. This enabled comparison with densely genotyped population groups. The HapMap sample comprised: 112 CEU; 84 CHB; 86 JPT; 113 YRI; 49 ASW; 85 CHD; 88 GIH; 90 LWK; 50 MEX; 143 MKK; and 88 TSI. There were 300,000 SNPs common to the Illumina array and HapMap data, which were used to calculate eigenvectors.

2.3.6 Association

The standard case-control allelic association test was run in PLINK with the following limits: maximum proportion of SNPs missing per individual 0.05; minimum minor allele frequency 0.01; minimum Hardy-Weinberg disequilibrium frequency p-value 1×10^{-7}; and maximum proportion of participants missing per SNP 0.05. The test compares allele frequencies between cases and controls. Only SNPs that were common to all three Illumina BeadChips were examined for association with case-control status, and those whose genotypes were poorly clustered (determined by eye; see section 2.3.7) were excluded. Participants who were found to deviate from the group when population stratification was assessed were excluded. Genome-wide significance was defined as $P < 5 \times 10^{-7}$.

A further association test was run in PLINK that adjusts for multiple testing. Amongst other output variables this produces a genomic control corrected p-value for each marker.

Finally logistic regression analysis was performed in PLINK with adjustment for eigenvectors that on average significantly explained data variability.

2.3.7 Examination of cluster plots

Based on initial association analysis results SNPs were sorted by p-value for statistical significance from lowest to highest (i.e. most to least significant). A cluster plot for each SNP was generated in Microsoft Excel and the most significant reviewed by eye. In total

more than 110,000 plots were viewed and if necessary reclustered by hand using Excel macros specifically created for that purpose. Figure 2.4 shows examples of good and poor plots.

2.3.8 Meta-analysis

Four of the validation cohorts were case-control samples selected using the same blood pressure thresholds as the discovery cohort: the Malmö Preventive Project (MPP); additional participants from the MDC study; the combined World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project (MONICA) and Pressioni Arteriose Monitorate e Loro Associazioni (PAMELA) study; and the Netherlands Study of Depression and Anxiety (NESDA). Ten further validation cohorts were obtained through collaboration with the Global BPgen consortium [58]. From these validation cohorts individuals were selected as cases if less than 60 years of age with SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg or current treatment with antihypertensive or blood pressure lowering medication, or if $\geq$ 60 years treatment commenced before age 60.

Individuals were selected as controls if at least 50 years of age with SBP $\leq$ 120 mmHg and DBP $\leq$ 80 mmHg and not treated with any blood pressure lowering medication. Individuals younger than 50 years were included as controls if they met a lower blood pressure threshold of SBP $\leq$ 115 mmHg and DBP $\leq$ 80 mmHg (and free from blood pressure lowering medication).

The top hit was assessed in a two stage validation process. A flowchart of the cohorts included and their sample sizes is shown in Figure 2.5. In stage 1 genotyping was performed in a combined World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project (MONICA) and Pressioni Arteriose Monitorate e Loro Associazioni (PAMELA) sample, a Malmö Preventive Project sample, and in a larger additional Malmö Diet and Cancer Study sample. In total stage 1 validation analysed 9,827 cases and 8,694 controls. Stage 2 analysis was conducted on a total sample of 10,018 cases and 7,847 controls from eleven further cohorts; combined British Genetics of Hypertension Study (BRIGHT)/ Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT), Prevention of Renal and Vascular End Stage Disease Study (PREVEND), Cohorte Lausannoise (CoLaus), Kooperative Gesundheitsforschung in der Region Augsburg (KORA), Study of Health in Pomerania (SHIP), British 1958 Birth Cohort (B58C), TwinsUK, Myocardial Infarction Genetics Consortium (MIGen), Diabetes Genetics

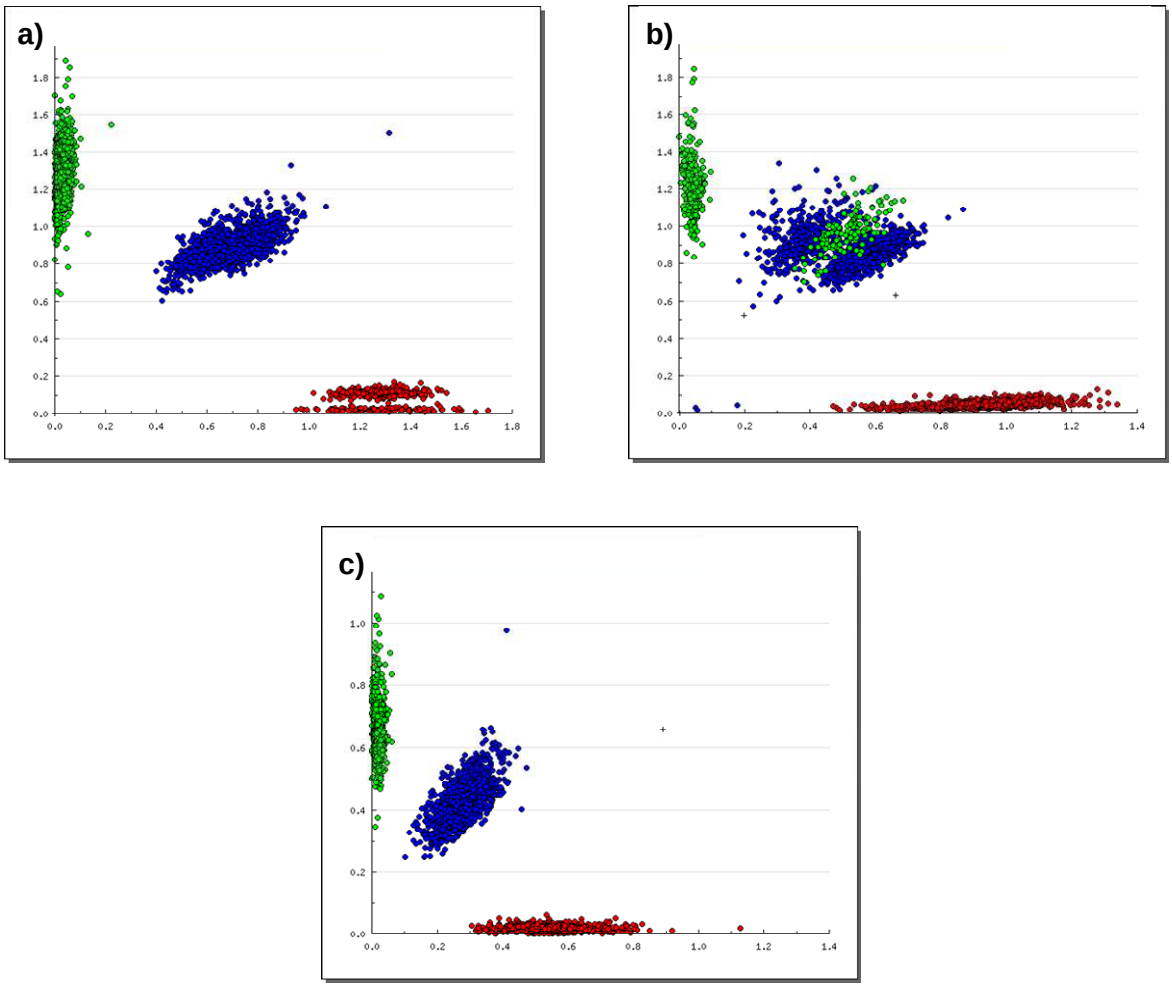
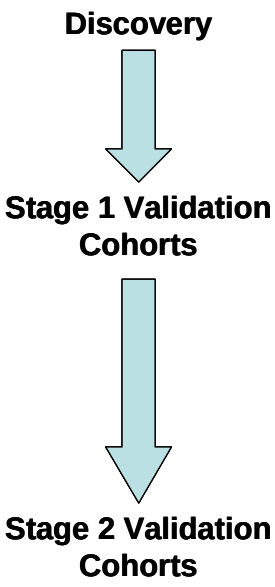


Figure 2.4 Examples of genotype cluster plots. Each plot represents an individual SNP and each data point a person, where green and red circles have been identified by Illumina as the two homozygotes and blue the heterozygote. a) good plot, included in formal analysis ; b) poor plot, removed from formal analysis; c) included in formal analysis following removal of the uppermost heterozygote individual who has not been clustered in any group. The top right individual, represented as a black cross, has already been excluded by Illumina.



Study	Origin	Cases	Controls	N of SNPs
Discovery	Swedish	1621	1699	521,220

Study	Origin	Cases	Controls	N of SNPs	Comments
MDC	Swedish	6977	6891	1	Top SNP (rs13333226)
MPP	Swedish	1956	1057	1	Top SNP (rs13333226)
MONICA/ PAMELA	Italian	894	746	87	Top 87 SNPs selected by P value
Total		**9827**	**8694**		

Study	Origin	Cases	Controls	N of SNPs
BRIGHT/ ASCOT	British/ Irish	3069	1787	1 SNP rs13333226
PREVEND	Dutch	2411	1613	
CoLaus	Swiss	1300	1375	
KORA	German	457	300	
SHIP	German	656	240	
B58C	British	514	529	
TwinsUK	British	245	845	
MIGen	European	316	278	
DGI	Swedish/ Finnish	277	161	
Fenland	British	264	510	
NESDA	Dutch	509	209	
Total		**10018**	**7847**	

Figure 2.5 Flow chart of discovery, validation stage 1, and validation stage 2 analyses. MDC = Malmö Diet and Cancer Study. MPP = Malmö Preventive Project. MONICA = World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project. PAMELA = Pressioni Arteriose Monitorate e Loro Associazioni. BRIGHT = British Genetics of Hypertension Study. ASCOT = Anglo-Scandinavian Cardiac Outcomes Trial. PREVEND = Prevention of Renal and Vascular End Stage Disease Study. CoLaus = Cohorte Lausannoise. KORA = Kooperative Gesundheitsforschung in der Region Augsburg. SHIP = Study of Health in Pomerania. B58C = British 1958 Birth Cohort. MIGen = Myocardial Infarction Genetics Consortium. DGI = Diabetes Genetics Initiative. NESDA = Netherlands Study of Depression and Anxiety.

Initiative (DGI), the Fenland Study, and Netherlands Study of Depression and Anxiety (NESDA). Finally, an overall meta-analysis was performed of the discovery sample and 14 validation cohorts, using an inverse-variance weighted fixed-effects model in Stata version 10 [147]. The beta coefficient estimate and standard error for each study were entered into the analysis, which produced a global (i.e. average) OR and associated 95% confidence interval. This is presented in graphical form as a forest plot of the individual and global ORs and 95% confidence intervals. Analysis was performed unadjusted, adjusted for age, age^2, sex, and BMI, and with further adjustment for estimated glomerular filtration rate (eGFR) in seven cohorts in which this was available. eGFR was calculated using the Modification of Diet in Renal Disease (MDRD) formula [157].

Heterogeneity of studies within the meta-analysis was assessed to determine whether a fixed-effects model, which assumes the true effect size is the same in each dataset, is appropriate [158]. Precision of effect size estimation and statistical power are both decreased by the presence of study heterogeneity. If within-study variability (ubiquitous sampling error/chance) alone is present then a global effect size can be calculated using a fixed-effects model. However, if there is evidence of further heterogeneity in the form of between-study variability then a random-effects model, which assumes the effect sizes in individual datasets vary around an overall average, must be used to take this into account. Alternatively, a search can be conducted within the fixed-effects model for covariates that may be introducing heterogeneity. This heterogeneity, over and above sampling error, is termed true heterogeneity. Its extent is summarised by the Q statistic and I^2 statistic. To calculate the Q statistic, the squared deviations of each individual study effect estimate from the global estimate are summed. The contribution of each study is weighted by its inverse variance. The null hypothesis is homogeneity and under it the Q statistic follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of studies. A P-value threshold of 0.10 is applied to determine significance. The I^2 statistic is calculated by dividing the difference between the Q statistic and its degrees of freedom by the Q statistic and multiplying by 100, and varies between 0 and 100%. I^2 is interpreted as the percentage of total variability that is due to between-studies variability, or true heterogeneity.

Funnel plots were created, also using Stata version 10, and assessed for evidence of bias resulting from overestimation of effect size in smaller studies of poor methodological quality. In general funnel plots are scatter plots in which treatment effects (horizontal axis) are plotted against a measure of study precision or sample size (vertical axis). Each data

point represents an individual study. Various measures can be used as the vertical and horizontal axes. In this case the vertical axis is the standard error of the effect estimate, with larger studies located towards the top of the graph. The horizontal axis is the exponential of the beta coefficient (i.e. the OR) on the log scale. This is as per the recommendations made by Sterne and Egger, who conducted a study evaluating the efficacy of several axes variables [159]. If there is no evidence of bias the plot is approximately symmetrical and looks like an inverted funnel, with the smaller studies at the bottom spread more widely because they have less precision than larger studies. Examples of hypothetical plots in the presence and absence of bias are shown in Figure 2.6 [160].

2.3.9 Annotation of top hits

A meta-analysis of the discovery cohort and combined MONICA/PAMELA dataset was performed for the 87 SNPs with association $P \leq 5.6 \times 10^{-4}$, again using an inverse-variance weighted fixed-effects model. Included in the analysis were 2,515 cases and 2,445 controls. Decisions on the importance of any genotype-phenotype association should be based on more than *P* value and effect size. Thus the top hits from this combined analysis were investigated to determine whether any SNPs were located in or near genes of potential biological significance, their functional relevance if within a gene, and which phenotypes they had been previously associated with, if any. To obtain this information the association results were uploaded to WGAViewer [161], a whole genome association annotation software package. PLINK association files can be uploaded to it directly without reformatting. WGAViewer enables polymorphisms to be considered in their wider genomic context and their relation via LD to other genotyped and ungenotyped SNPs, as well as providing hyperlinks to online genetic databases. Its comprehensive annotation facility performs LD tests using HapMap data by default (alternatively, pre-calculated LD datasets can be uploaded from PLINK or Haploview and used as LD sources) to assess whether there are functional proxies for hits in the study under consideration. A choice of the four original HapMap populations (CEU, CHB, JPT, and YRI) is offered; in the current study SNPs were annotated with the CEU population. Furthermore, the current study employed all default annotation parameters, namely; 500kb area around SNP searched for closest gene, 200bp area around SNP searched for closest exon, LD window span 200kb, and minimum r^2 threshold of 0.8 for LD.

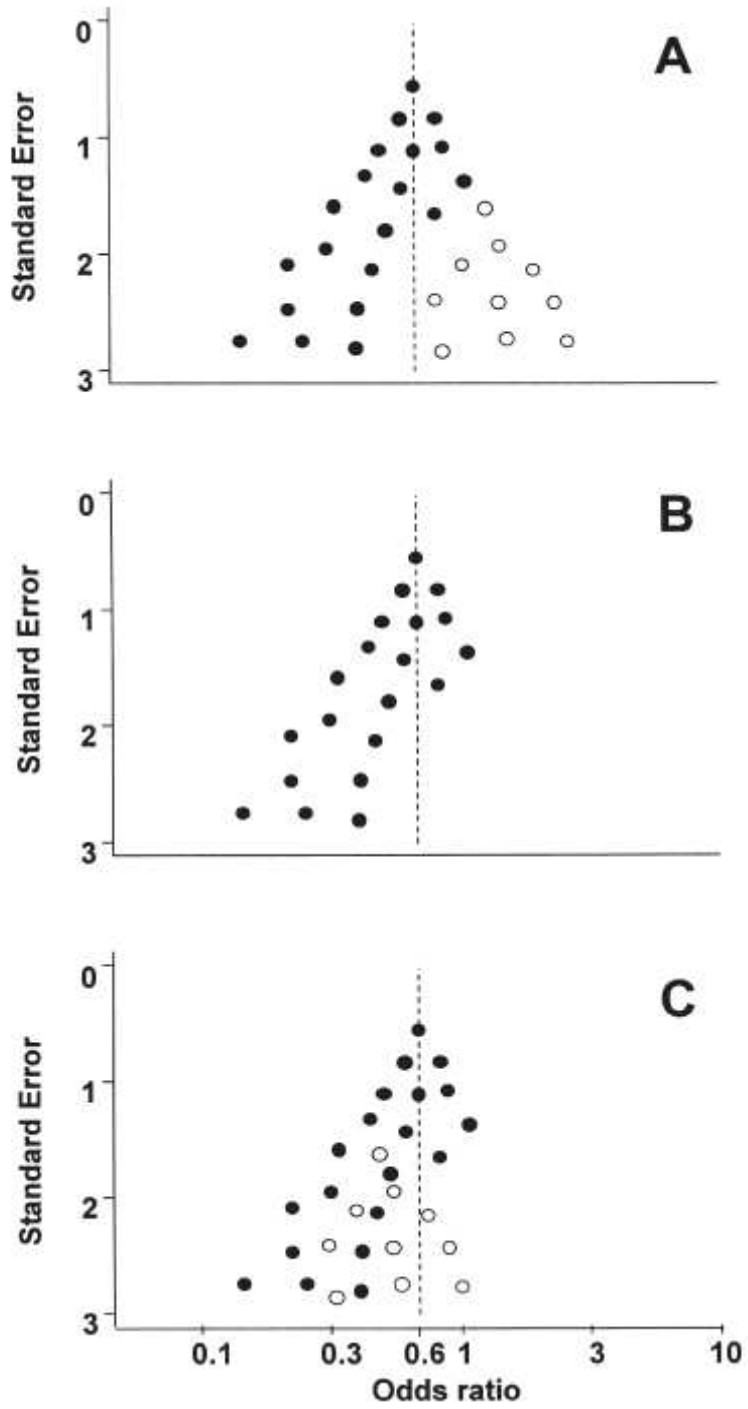


Figure 2.6 Hypothetical funnel plots. A) Symmetrical plot in the absence of bias (open circles indicate smaller studies showing no statistically significant effects); B) asymmetrical plot in the presence of publication bias (smaller studies showing no statistically significant effects are missing); C) asymmetrical plot in the presence of bias due to low methodological quality of smaller studies (open circles indicate small studies of inadequate quality whose results are biased toward larger effects) (reproduced from [160]).

Following annotation, WGAViewer hyperlinks specific to the gene closest to each SNP were followed to Online Mendelian Inheritance in Man (OMIM) [162, 163], Entrez Gene [164, 165], and GeneCards [166, 167]. These three databases were accessed, rather than only one, to avoid missing data on a potentially interesting finding. All are freely available to the public for unrestricted use. The amount of information available varies depending on how much is known about a gene. For example those that have been associated with disease have typically been extensively studied hence much is known about them, whereas for others there is no detail beyond genomic location. The potential information supplied by each database is outlined below.

OMIM and Entrez Gene both belong to the Entrez suite of databases [168], produced by the National Center for Biotechnology Information (NCBI) in the United States. OMIM began life in 1966 as Mendelian Inheritance in Man (MIM), a printed reference guide to genetic disorders and genes. There were 12 print editions of MIM, the last published in 1998 [169]. Since then it has been distributed electronically as OMIM and now contains around 20,000 entries (database accessed 6th of May 2010). The database contains information on: gene name and symbol; alternative names and symbols; cloning; gene structure, mapping and function; molecular genetics; associations with phenotypes; animal models; allelic variants; and a reference list with links to relevant abstracts and sometimes full articles (if available) in PubMed.

Entrez Gene is a gene-specific database that aims to provide tracked, unique identifiers for genes of multiple genomes and associated information. Genes are reported for several model organisms with species-specific identifiers. Information provided by Entrez Gene includes: gene name, symbol and synonyms; gene type; organism; genomic regions, transcripts and products; genomic context; expression; a list of references regarding gene discovery, mapping, and function with links to PubMed; pathways; markers; associated phenotypes with links to citations; homologs; and proteins. It contains entries for 42,506 genes in humans, and more than 6 million genes in total (database accessed 6th of May 2010).

The GeneCards database integrates information about human genes, proteins and diseases, extracted from over 80 databases including both OMIM and Entrez Gene. Specifically, a GeneCard provides information on: official gene name; synonyms; orthologs in homologous species and species with no ortholog; chromosomal location of the gene and its homologues; protein(s) encoded by the gene along with their function, expression,

association with disease; diseases that have been associated with the gene; new diagnoses and treatments that have been developed from knowledge of the gene; and links to other sites providing more information [170]. GeneCards are available for more than 70,000 genes (database accessed 5th of May 2010). The database authors estimate that the prevalence of false negatives, i.e. missing data from a source that does in fact contain gene information, is in range 0-10% depending on the source [171]. This should not be of concern as multiple sources contribute to GeneCards, moreover the current study also searched OMIM and Entrez Gene.

Further to the above, each gene was entered into the HuGE Navigator (http://hugenavigator.net/) [172, 173] Genopedia search engine to establish whether any pertinent information had been missed. The HuGE Navigator is a continuously updated online knowledge base of genetic associations and genome epidemiology. The information contained within is taken from published, population-based epidemiologic studies of human genes, extracted and curated from PubMed. The Navigator is managed and maintained by the Human Genome Epidemiology Network (HuGENet: http://www.cdc.gov/genomics/hugenet/default.htm), a voluntary international collaboration. It provides several search tools and informatics utilities as well as two encyclopaedias: Phenopedia which is searchable by phenotype; and Genopedia which is searchable by gene [174].

As a final check for any information that may have been missed, each gene was entered into the NCBI dbGaP database of genotypes and phenotypes (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) [175]. dbGaP archives the results of studies that have investigated associations between genotype and phenotype. The types of study it contains include GWAS, medical sequencing, molecular diagnostic assays, and associations between genotypes and non-clinical traits. Data is summarised by study, with PubMed links to related articles. There is a facility to apply for access to data, for the exploration of new research hypotheses.

Regional association plots of the top hits were created using LocusZoom Version 1.1 (http://csg.sph.umich.edu/locuszoom/), a tool that plots local association results along with information about the locus including the location and orientation of genes, local estimates of recombination rates, and levels of LD. In the current study all plots were created at the same time using the batch mode facility. Each individual plot was specified by the SNP of interest, acting as the key marker for the region. All markers within an area flanking

500kb each side of the index SNP were included. Plots were generated based on Human Genome Build 18 (hg18). Pairwise LD coefficients, measured as r^2, between each SNP and the index SNP were calculated by LocusZoom using the HapMap Phase II CEU population as a reference. Data points are coloured accordingly and an explanatory key provided. Recombination rates were also estimated from the HapMap Phase II CEU samples.

2.3.10 Clinical functional studies

Functional associations of the top SNP were studied in samples from three cohorts: the BRIGHT study [128]; the Hypertension Evaluation by Remler and CalciUria LEvel Study (HERCULES) [176]; and a Groningen Renal Hemodynamic Cohort Study Group (GRECO) study [177, 178]. The BRIGHT sample comprised 256 hypertensive participants who had completed 24-hour urine collection with urinary sodium, potassium, creatinine and microalbuminuria recorded. The HERCULES sample was 100 middle aged general population participants with 24-hour ambulatory blood pressure measurement and 24-hour urine collection phenotyped for variables including urinary sodium, creatinine clearance, endogenous lithium clearance, potassium and uric acid excretion, and microalbuminuria. Finally, the GRECO sample comprised 64 healthy young males from a crossover protocol consisting of two 7-day periods, one a high sodium diet (HS; 200 mmol Na^+/day), and the other a low sodium diet (LS; 50 mmol Na^+/day). 24-hour urine collection was used to assess dietary compliance and the achievement of a stable sodium balance.

In the current analysis the primary measurement of interest was urinary uromodulin. This was measured in duplicate in 24-hour urine samples using a commercially available enzyme-linked immunosorbent assay (ELISA) produced by MD Biosciences in Zurich, Switzerland (http://www.mdbiosciences.com/), and applied as recommended by the manufacturer. The range of the assay is 9.375 – 150 ng/ml and its sensitivity is <5.50 ng/ml. The inter-assay coefficient of variation was 11.9%. For the BRIGHT and GRECO samples urinary uromodulin was measured in Glasgow, whereas the HERCULES samples were processed in Lausanne by HERCULES investigators. Measured uromodulin was corrected for urinary creatinine and then statistical analysis performed. The BRIGHT data were analysed in Glasgow, and the GRECO and HERCULES data analysed by their respective investigators. Multiple regression was used to assess association between genotype and uromodulin, as well as other functional parameters such as creatinine clearance, eGFR, and fractional excretion of sodium (FENa). In the BRIGHT samples

eGFR was calculated using the MDRD equation, and in the HERCULES samples the abbreviated MDRD [179], whereas in the GRECO study glomerular filtration rate (GFR) was measured directly. Traits with non-normal distributions were tested using the non-parametric Kruskal Wallis test. In the GRECO sample only five individuals had the GG genotype, therefore AG and GG were combined for analysis that compared them with AA.

3 Genome-wide association study in extremes of blood pressure distribution

The first part of the current study is a genome-wide association study of hypertension as defined by extreme case and control criteria. This chapter describes the initial quality control measures undertaken and resultant exclusion of participants and SNPs that did not pass predetermined thresholds. Following this the results of the population stratification detection and correction method employed are outlined, and the final sample population characteristics are presented. Lastly the results of the formal GWAS analysis are presented and discussed.

3.1 Sample quality control

3.1.1 Specification of gender

Where available the observed genotypes of SNPs on chromosomes X and Y were examined to confirm the gender assignments contained in the phenotype file. Individuals were inspected who were coded as "male" but had a significant amount of heterozygous X genotypes (≥1%), or who were coded as "female" but had a high frequency of homozygous X genotypes (≥80%) or Y genotype readings. If it was not possible to correct any discrepancy, the individual was excluded from formal analysis. In this step five individuals were removed.

3.1.2 Cryptic relatedness

This step was taken to check for unexpected relatedness between study participants. Sharing of genetic information was estimated using identity by state (IBS) values calculated by PLINK, for every possible pairwise comparison of participants. All pairs of DNA samples with IBS ≥ 0.80 were individually inspected, and the sample with the lower call rate in each pair was excluded from further analyses. In total, 68 subjects were removed in this step.

3.1.3 Skewed missingness

A test of missingness by case/control status was performed, to determine whether the missing genotypes were skewed and hence may give rise to spurious association p values.

There was no evidence of bias for the most associated SNPs that were carried forward for validation.

3.1.4 Multidimensional scaling plot outliers

In the first stage of analysis, MDS was used to remove 33 individuals who were outliers as they did not cluster with the rest of the samples. However, despite this, there were issues with structure and genotyping quality. Therefore 752 samples were regenotyped (for further explanation see section 3.2), and remaining population structure identified and corrected using EIGENSTRAT.

3.1.5 Genotyping success

92 individuals were removed who had genotyping success less than 95%.

To summarise, through the QC detailed above the following removals were made: 388 participants due to ancestry problems (identified by EIGENSTRAT, below), five with unspecified sex, 68 duplicates or participants with evidence of relatedness, 33 MDS plot outliers, and 92 individuals with genotyping success less than 95%. The remaining final sample size is 3,320, comprising 1,621 cases and 1,699 controls.

3.2 SNP quality control

SNP data were screened within BeadStudio using a two step procedure. First of all, SNPs with a cluster separation value below 0.3 were manually checked to ensure correct calls. Many of these were fixed manually, but some were excluded. The second step evaluated any SNP that had a Het Excess value between -1.0 to -0.1 and 0.1 to 1.0. The Het Excess value indicates the quantity of excess heterozygote calls relative to expectations based on Hardy-Weinberg equilibrium. It varies from -1 (no heterozygotes) to 1 (100% heterozygotes). This is with the exception of SNPs on the X chromosome, which are not assessed because males are not expected to be heterozygous for X chromosome loci.

Following the above procedure the genotyping success rate was 98.4%. When the different chip types were examined separately it became apparent that the single and duo

chips performed sub-optimally, in terms of genotyping call rate, when compared with the quad chip. Therefore 752 samples that had been genotyped on the single or duo chips were repeated using quad. Table 3.1 presents the numbers of cases and controls genotyped on each bead chip type, before and after regenotyping. The concordance rate for duplicate genotyping was 99.99%.

3.2.1 Visual cluster plot inspection

With assistance from colleagues at the British Heart Foundation Glasgow Cardiovascular Research Centre, more than 110,000 cluster plots were individually examined. Of these, I personally examined >30,000. In total 3319 SNPs were removed for poor clustering.

3.2.2 Minor allele frequency

23,562 SNPs were removed prior to formal analysis because they had an MAF <0.01.

3.2.3 Hardy-Weinberg disequilibrium

Deviation from Hardy-Weinberg equilibrium was checked in both cases and controls because technically both were selected from extreme ends of the trait distribution, and thus do not represent normal randomly mating populations. 1,915 SNPs had a Hardy-Weinberg P value $\leq 1 \times 10^{-7}$ in either cases or controls and were therefore removed.

3.2.4 Missingness

12,097 SNPs had a genotype missing rate of >5% in either cases or controls and were therefore removed.

Following SNP exclusions due to low MAF, Hardy-Weinberg disequilibrium, and/or high rate of missingness, a final set of 521,220 SNPs was available for analysis. The overall exclusion rate is comparable with those reported for similar studies [58, 59].

3.2.5 Adjustment for stratification using principal components

Examples of biplots of eigenvectors for the current sample and HapMap samples are shown in Figures 3.1 and 3.2. Ten PCs (eigenvectors) were extracted using

Table 3.1 Numbers of cases and controls genotyped on each bead chip type. Before and after regenotyping.

	Singles		Duos		Quads	
	Cases	Controls	Cases	Controls	Cases	Controls
Original genotyping	726	3	540	5	796	1976
After regenotyping	565	2	321	4	1123	1895

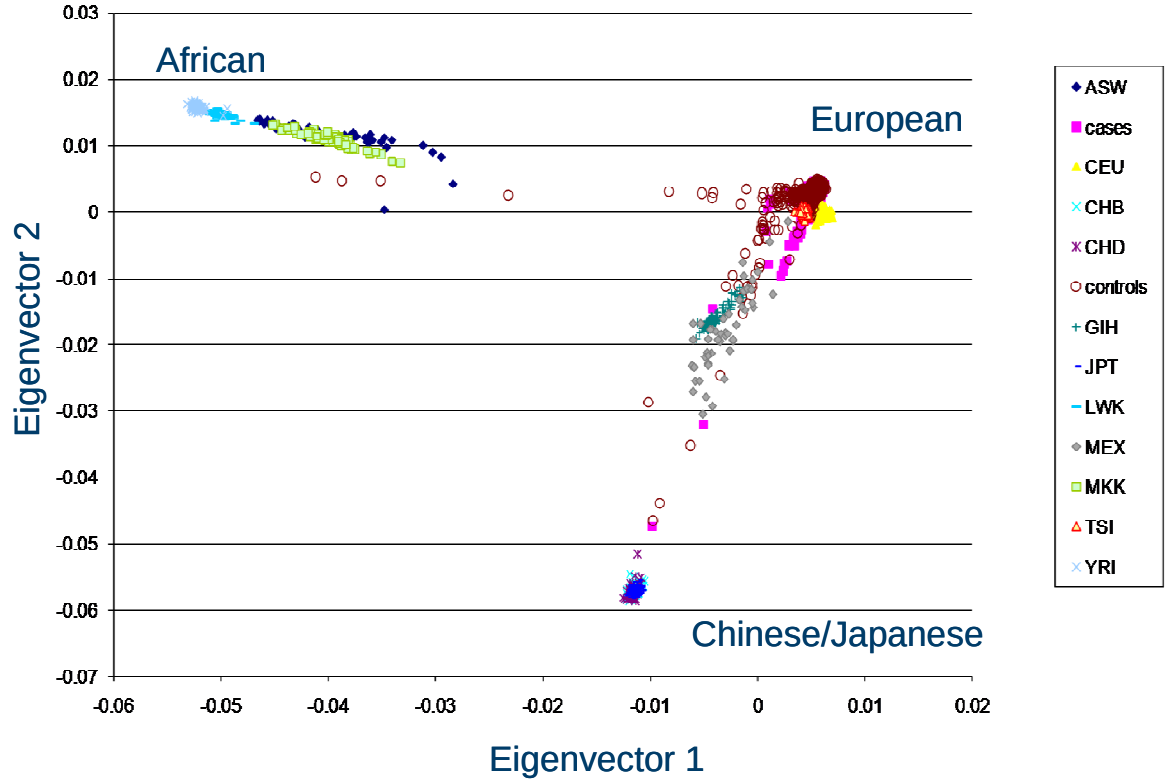


Figure 3.1 Biplot of first two principal components for all HapMap samples and the current study cases and controls. Cases are represented as pink squares, controls as open red circles. Outlying individuals in tails of European distribution (i.e. closer to African and Chinese/Japanese HapMap samples than European) were excluded from association analysis. ASW = African ancestry in Southwest USA. CEU = Utah residents with Northern and Western European ancestry. CHB = Han Chinese in Beijing, China. CHD = Chinese in metropolitan Denver, Colorado. GIH = Gujarati Indians in Houston, Texas. JPT = Japanese in Tokyo, Japan. LWK = Luhya in Webuye, Kenya. MEX = Mexican ancestry in Los Angeles, California. MKK = Maasai in Kinyawa, Kenya. TSI = Toscani in Italy. YRI = Yoruba in Ibadan, Nigeria.

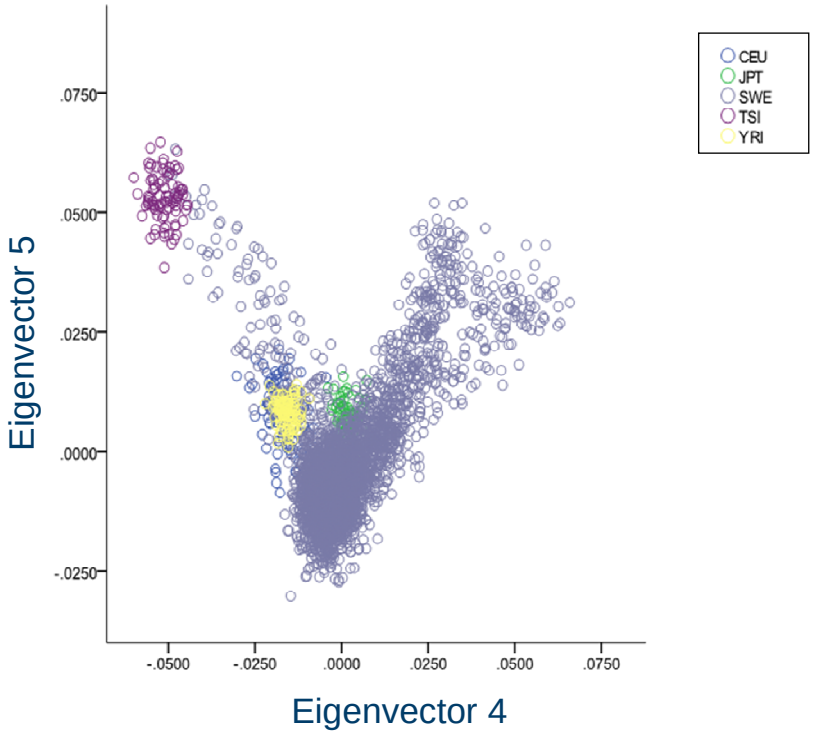


Figure 3.2 Biplot of principal components 4 and 5 for some of HapMap samples and the Swedish sample from the current study. Swedish samples are represented as open lilac circles. CEU = Utah residents with Northern and Western European ancestry. JPT = Japanese in Tokyo, Japan. SWE = current study sample in Sweden. TSI = Toscani in Italy. YRI = Yoruba in Ibadan, Nigeria.

EIGENSTRAT [74], and were then entered into logistic regression analysis sequentially to assess which were most effective in explaining any population structure within the sample. The genomic control inflation factor, λ, was used to determine the impact of their inclusion. Five PCs (3,4,5,6, and 10) accounted for a significant amount of population structure, defined as a significant p-value for the t statistic coefficient, and their combined effect had the maximum impact on λ. Therefore these were included as covariates in all subsequent association analysis. In addition extreme genetic outliers were identified, defined as individuals whose ancestry is at least six standard deviations from the mean on one of the top ten eigenvectors. The process is applied iteratively, and the EIGENSTRAT default setting of 5 iterations was used. In this manner 388 outliers were identified and excluded from the formal analysis.

3.3 Formal analysis

3.3.1 Final sample population characteristics

The population characteristics for the final sample of 3,320 participants are presented in Table 3.2. Data were available for the variables SBP, DBP, age, and BMI. As determined by recruitment criteria cases had higher SBP and DBP. They were also younger and had higher BMI on average.

3.3.2 Association analysis

After removing SNPs that were clustered badly and QC, 521,220 SNPs were available for analysis. Genotype information was compared for 1,621 cases and 1,699 controls using logistic regression, assuming an additive model. The Manhattan plot of GC adjusted – $\log_{10} P$ values for association of hypertension status with markers in all chromosomes is shown in Figure 3.3. This is after adjustment for PCs.

A quantile-quantile plot of (GC adjusted) observed versus expected $-\log_{10} P$ values is shown in Figure 3.4, where $\lambda = 1.07$. This is close to 1 which indicates that, after removal of genetic outliers and adjustment for PCs, there is some residual inflation but no substantial evidence of population stratification. Prior to outlier removal λ was far greater at 2.08, and before PC adjustment it remained high at 1.8; hence these quality control measures were necessary and effective.

Table 3.2 Population characteristics of controls and cases. Summarised as mean (SD). *P*-value is for two sample t-test.

	Controls (n=1699)	Cases (n=1621)	*P*
Age at enrolment, years	57.4 (5.9)	55.4 (7.1)	<0.001
BMI, kg/m^2	24.2 (3.5)	27.1 (7.8)	<0.001
SBP, mmHg	115.8 (6.8)	175.8 (22.5)	<0.001
DBP, mmHg	73.7 (5.7)	104.7 (11.8)	<0.001

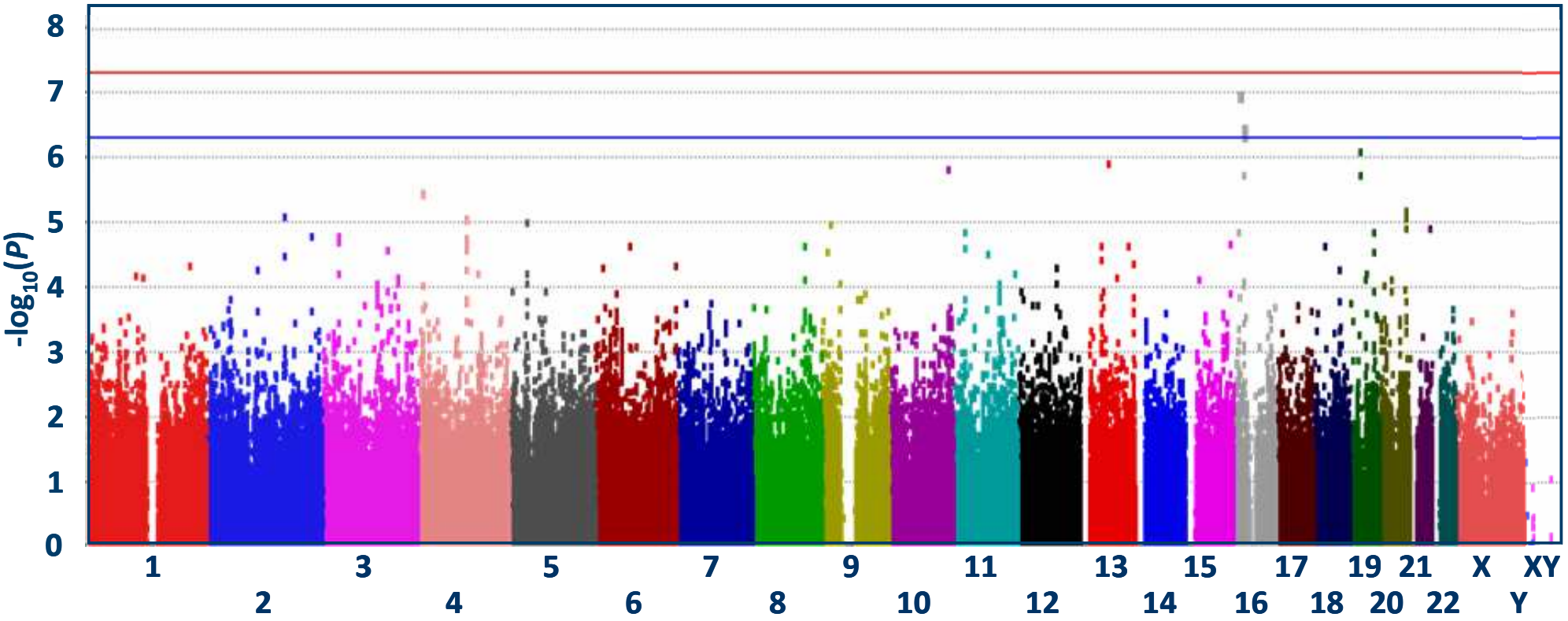


Figure 3.3 Manhattan plot of $-\log_{10}$ transformed *P* values against genomic position for association of hypertension status with markers in all chromosomes.
Red line indicates $P=5\times10^{-8}$ and blue line indicates $P=5\times10^{-7}$.

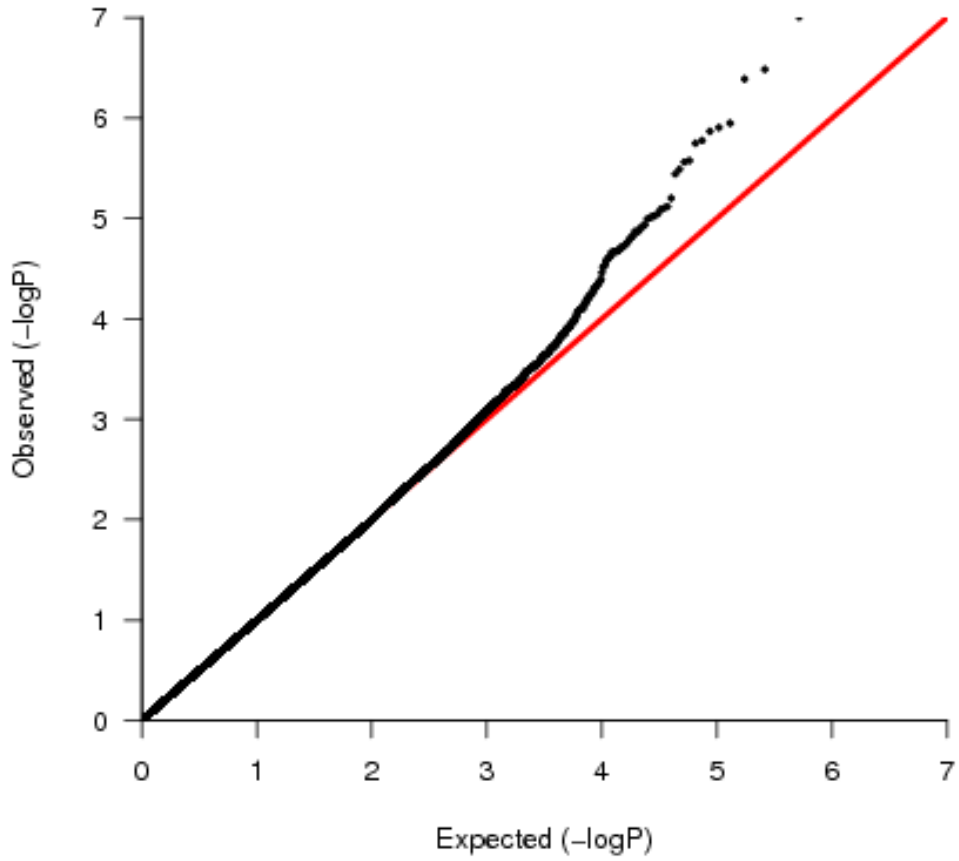


Figure 3.4 Quantile-quantile plot of observed versus expected $-\log_{10}$ *P* values for genome-wide data. Red line represents line of equality, i.e. no association. λ = 1.07.

Cluster plots of the 119 SNPs with a GC adjusted P value $\leq 1 \times 10^{-4}$ were visually inspected, leading to the exclusion of 39 poorly clustered SNPs. Table 3.3 presents the association results for all markers with a $P < 1 \times 10^{-5}$. Of the total 521,220 SNPs entered into analysis, seventeen met this threshold of significance, of which three attained $P < 5 \times 10^{-7}$.

Table 3.3 Association results for SNPs with GC adjusted P < 1 × 10^{-5}

SNP ID	CHR	Position	OR	95% CI	Unadjusted *P*	GC adjusted *P*
rs13333226	16	20273155	0.67	0.58 – 0.78	1.14×10^{-07}	3.28×10^{-07}
rs4293393	16	20272089	0.67	0.58 – 0.78	1.45×10^{-07}	4.09×10^{-07}
rs4932779	19	22573041	1.45	1.26 – 1.67	2.83×10^{-07}	7.63×10^{-07}
rs13378149	13	58050022	0.54	0.43 – 0.69	4.30×10^{-07}	1.13×10^{-06}
rs13353058	10	124740712	1.62	1.34 – 1.96	5.26×10^{-07}	1.36×10^{-06}
rs8111998	19	22533515	0.52	0.41 – 0.68	6.59×10^{-07}	1.68×10^{-06}
rs11647727	16	20263666	0.72	0.63 – 0.82	7.03×10^{-07}	1.78×10^{-06}
rs10009111	4	10580521	0.76	0.68 – 0.85	1.35×10^{-06}	3.26×10^{-06}
rs10011697	4	10580930	0.76	0.68 – 0.85	1.50×10^{-06}	3.61×10^{-06}
rs487331	20	53608771	0.77	0.69 – 0.86	2.73×10^{-06}	6.29×10^{-06}
rs2084543	2	162404660	0.60	0.48 – 0.74	3.36×10^{-06}	7.63×10^{-06}
rs555848	20	53613135	0.77	0.69 – 0.86	3.48×10^{-06}	7.86×10^{-06}
rs7669524	4	102833192	1.36	1.19 – 1.54	3.60×10^{-06}	8.12×10^{-06}
rs13124455	4	102830679	1.36	1.19 – 1.54	4.00×10^{-06}	8.97×10^{-06}
rs292196	5	36949936	0.70	0.60 – 0.81	4.21×10^{-06}	9.41×10^{-06}
rs172384	5	36839982	0.71	0.61 – 0.82	4.27×10^{-06}	9.53×10^{-06}
rs2289006	9	18768319	0.76	0.68 – 0.86	4.41×10^{-06}	9.82×10^{-06}

CHR = chromosome. OR = odds ratio. CI = confidence interval. GC = genomic control.

3.4 Discussion

We have demonstrated a novel extreme phenotyping method to identify common genetic markers associated with hypertension. High fidelity phenotyping, comparing the top and bottom end of the blood pressure distribution, has reduced misclassification of controls and resultant noise [130]. This experimental design increases statistical power to detect effects, thus a smaller sample size is required. We have identified markers significantly associated with hypertension from a discovery sample of just 3,320 individuals, far fewer than other successful GWAS of hypertension e.g. Global BPgen [58] and CHARGE [59]. Furthermore the strategy allows the sampling of participants from a single population, rather than combining multiple cohorts from heterogeneous populations, which reduces confounding by stratification and allows for uniform phenotypic characterisation. This confers practical benefits and reduces costs. The estimated odds ratios are likely to be inflated compared with the true odds ratios for hypertension as typically defined.

There is more than one way to conduct a GWAS of hypertension and/or blood pressure. The study of blood pressure as a quantitative trait, rather than qualitative hypertension (as conventionally defined), may confer more statistical power [180, 181]. This is exemplified by the Global BPgen study, which in Stage 1 analysis identified eight loci associated with SBP and/or DBP at the level of genome-wide significance [58]. However, there was insufficient power to examine associations with hypertension in Stage 1. Instead the eight significant loci were associated with hypertension in planned secondary analysis, which showed that all were indeed associated in the same direction of effect. Furthermore, in the CHARGE study the loci discovered in scans of SBP and DBP were not all discovered in the hypertension scan, whereas all those discovered in the hypertension scan were also associated with continuous blood pressure [59]. Thus, had the group studied hypertension alone, some markers would have been missed. Undoubtedly power would have been yet greater in both studies had they been single centre or had the multiple component studies been prospective with rigorous identical phenotyping.

Plomin et al argue that, as common disorders are affected by multiple genetic variants, they are best represented as quantitative rather than qualitative traits [182]. This is because several markers combined can resemble a risk score with a continuous, normal distribution. Within this paradigm qualitative disorders represent merely the quantitative extremes of a continuous distribution of genetic risk. They point out that for most common disorders, e.g. cancers, arthritis, autism, the relevant quantitative traits are not clear. However in

hypertension research it is fortunate that blood pressure is the obvious quantitative equivalent. It was Fisher who originally showed that complex quantitative traits that are affected by several genes can follow a qualitative Mendelian pattern of inheritance [183]. This resolved the early 1900s conflict between Mendelians who believed that all traits have simple Mendelian inheritance patterns, and biometricians who believed that Mendel's laws do not apply to complex traits.

To date, the SNPs associated with continuous blood pressure have only explained a small amount of its variability; approximately 1 mmHg SBP per allele and 0.5 mmHg DBP per allele [58, 59]. When combined, however, these SNPs can exert cumulative effects of a magnitude found to produce meaningful changes in cardiovascular risk at a population level. For example the Prospective Studies Collaboration observed that in middle age a 2 mmHg reduction in SBP translated to a 7% lower mortality from ischaemic heart disease (IHD) and other vascular causes, and a 10% lower stroke mortality [10]. This applied throughout the normal range of blood pressure down to 115/75 mmHg. Hence even small reductions in blood pressure will have a clinically significant impact on absolute population risk. It seems rational to view the qualitative definition of hypertension and quantitative definition of blood pressure as complementary rather than contradictory. Both have led to whole genome experimental designs with positive results.

Despite the above described successes, there remains a large proportion of unexplained blood pressure heritability. For example, the cumulative effect of the top ten loci identified by the CHARGE consortium explains only 1% of blood pressure variability [59]. This is expected, however, when viewed in the context of findings for other highly heritable complex traits [184]. In a GWAS of adult height, Weedon et al identified 20 associated variants in a discovery sample of 13,655, which were then replicated in an additional 16,482 individuals [185]. Taken together the 20 SNPs explained around 3% of height variation. Similarly, Willer et al further confirmed two previously reported loci (in *FTO* and *MC4R*) and discovered six novel loci associated with BMI in a GWAS with a combined discovery and replication sample of more than 90,000 individuals [186]. In spite of the size of the study and positive findings, when combined the eight SNPs explained just 0.84% of variation in BMI.

In an effort to explain the remaining heritability of height, Yang and colleagues combined the effects of all common SNPs measured genome-wide in a single study [187], instead of testing the significance of individual SNPs. They show that in this manner 45% of height

variation is explained, and argue that the discrepancy between this and the total heritability of 80% is in part due to incomplete LD between causal variants and genotyped SNPs. Therefore, a substantial number of causal SNPs are being missed because they do not meet the stringent threshold for genome-wide significance; most heritability is not missing but is hidden [188]. However, this being said, the work of Yang et al does not suggest a practical use for the additional explanatory variants in terms of risk prediction or treatment. Their method has no way to differentiate individual variants of potential interest from the rest. Rather, their analysis serves as an exemplar study on the meaning of the entire spectrum of genome-wide data.

Meta-analyses of GWAS are becoming ever larger, in a bid to detect rare variants and common variants with smaller effect sizes. The International Consortium for Blood Pressure-Genome-Wide Association Study (ICBP-GWAS) [180] comprises all cohorts from Global BPgen [58] and CHARGE [59] as well as some additional cohorts. The combined sample is more than 70,000 individuals of European ancestry plus smaller numbers of other ethnic groups. This should be a sufficient sample to detect further novel variants of similar effect size to those already identified [189]. It has been argued that earlier GWAS power calculations underestimated adequate sample size in certain circumstances. For example, Burton et al recommend that calculations take into account realistic assessment error rates in exposures and outcomes, plus the impact of unmeasured aetiological determinants [190]. This is relevant to large collaborations where participants were not specifically recruited and phenotyped for the trait being examined, such as Global BPgen and CHARGE where blanket blood pressure adjustments were applied to account for blood pressure reducing treatment. Furthermore, the INTERSALT study estimated the test-retest reliability of clinic SBP as 0.69-0.74 and DBP as just 0.63-0.67, with an average of 14 days between measurements [191]. Home and ambulatory blood pressure measurements are shown to be more reproducible [192], and would be preferable methods in prospective studies. Technological advances mean that genetic exposure variables can be measured almost without error. However the same cannot be said for many environmental exposures and errors in their assessment can reduce study power.

In the current study the chance of detecting true genetic effects was increased through the method of comparing individuals with very low and very high blood pressure. An alternative, complementary strategy is to decrease the influence of environmental factors contributing to overall risk. For example, Spence and colleagues have employed multiple regression modelling to identify participants with excessively high carotid plaque area

once traditional risk factors were taken into account [193]. Such individuals were described as having "unexplained atherosclerosis", and Spence et al proposed the use of this trait to ascertain subjects for the study of genetic loci. In theory individuals with atherosclerosis and low levels of traditional risk factors should have high genetic risk, thus this approach should increase statistical power.

Lanktree et al developed this method in the context of GWAS [194]. Again using plaque area as a quantitative measure of carotid atherosclerosis, they conducted power calculations for different theoretical phenotyping strategies. They first ran a stepwise regression model to predict plaque area with known CVD risk factors as input variables including sex, age, smoking, blood pressure, cholesterol, diabetes, and blood pressure and lipid lowering treatment. Individuals falling into the bottom 5% of regression residuals were defined as having "unexplained atherosclerosis" (i.e. large observed plaque area compared with area predicted by model), and those in the top 5% of residuals were defined as having "unexplained protection" (i.e. small observed plaque area compared with that predicted by model). The power conferred by using these groups as cases and controls in a GWAS was calculated as well as the power conferred by analysing plaque area as a simple quantitative trait. Under comparable effect sizes and allele frequencies the quantitative strategy required a sample four times larger to obtain the same level of power as extreme qualitative selection. The advantage of this novel approach is that it excludes potential participants who are not of interest and would introduce noise into the genetic analysis, namely those whose plaque area is predicted by traditional risk factors. Of course a limitation is the requirement of additional phenotyping which may not be possible in retrospective studies.

Two companies, Affymetrix [195] and Illumina [148], are at the forefront of whole-genome genotyping technology and provide genotyping platforms for SNPs and copy number variants (CNVs). Both companies manufacture standard chips as well as customised chips that cover variants of interest to the researcher. However they differ in SNP selection strategy. Illumina uses genome-wide arrays of tag SNPs derived from HapMap data. By contrast, Affymetrix distributes SNPs randomly across the genome ignoring levels of LD. Consequently their standard chips do not cover the same SNPs, but they do share some in common, allowing comparisons of genotyping agreement and reliability. One study conducted by Suarez and colleagues found that for 94 shared SNPs genotypic concordance was 99.85% between platforms [196]. However Affymetrix had a much greater no-call rate, with 6251 missing genotypes compared with 726 for Illumina.

More recently Kim and colleagues compared, for 757 common SNPs, the genotyping accuracy of the Illumina HumHap550, the Affymetrix 500K, and the Affymetrix custom-made GeneChip Targeted Genotyping (TG) 25K [197]. The TG 25K was used as an independent reference panel to evaluate the accuracy of the other two. A consensus dataset was created, containing genotypes called identically by at least two platforms. The percentage consistency was then calculated for each platform, i.e. the percentage of genotypes that the platform called as the same as the consensus genotype. After excluding SNPs with genotype call rate <80% and deviations from Hardy-Weinberg equilibrium, consistency was 99.08% in TG, 98.64% in Affymetrix, and 99.95% in Illumina (which also had the highest consistency without QC exclusions).

Other studies have analysed the proportion of the genome covered by different SNP panels. Barrett and colleagues compared the Illumina HumanHap300 and the Affymetrix 500K depending on the HapMap population samples studied [198]. Whereas Illumina coverage was better for CEU samples (Illumina, 75%; Affymetrix, 65%), Affymetrix performed better for YRI samples (Illumina, 28%; Affymetrix, 41%). Whole-genome coverage of JPT+CHB samples was similar (Illumina, 63%; Affymetrix, 66%). Conversely, when Wollstein and colleagues examined five SNP chips (Affymetrix 100K and 500K, and Illumina HumanHap100, 300, and 550) for the same HapMap populations using their own measure of coverage, they found little difference between Illumina HumanHap300 and Affymetrix 500K [199]. The greatest coverage in all three samples was provided by Illumina HumanHap550. This conclusion was also reached in a study of the same SNP panels (apart from HumanHap100) conducted by Mägi and colleagues [200]. In addition to the HapMap samples they looked at genome-wide SNP coverage in an Estonian Caucasian population sample. For all four chips the Estonian sample coverage was as good as for the CEU sample, confirming that commercially available chips are suitable for use in non-reference populations. Li and colleagues found that the Illumina HumanHap650Y and Human1M chips provide superior coverage, both local and gene specific, compared with the Affymetrix SNP Array 5.0 and SNP Array 6.0 [201]. This is achieved despite the significantly higher number of SNPs genotyped by Affymetrix, because Illumina employs LD information to a greater degree to increase efficiency. An important point is that in all of these comparison studies the platforms examined are the most recent releases at the time of study, so depending on the timing either company could have a more up to date product.

The current study used the Illumina genotyping platform because it has been shown to have better coverage, especially among Caucasians. This is because the SNP tagging

strategy employed by Illumina provides better coverage of the common variants identified through HapMap, than other genotyping platform strategies. Three BeadChip types were used; HumanHap550, HumanHap550-Duo, and Human610-Quad. This was primarily for two reasons, the first cost saving and the second the increased coverage provided by newer chips. In hindsight this was a mistake. Cases and controls did not arrive at the lab at the same time and for expedience and convenience were genotyped sequentially rather than in parallel. As a result most cases were genotyped using the single and duo chips, and most controls using the quad chips. This breaks a central rule of epidemiological study design: cases and controls must be treated and analysed under the same experimental conditions. During the QC procedure the MDS plots showed evidence of substructure. It transpired that some of this was due to experimental differences, as well as population stratification. Single and duo chips had performed sub-optimally when compared with the quad chip, and for that reason 752 samples that had been genotyped on the single or duo chips were repeated using quad. This additional expense could have perhaps been avoided had the methodology been more carefully thought through from the outset. However the differences between chip types were entirely unexpected because they were different releases of the same product from Illumina.

In summary, the results presented validate the extreme case-control phenotyping method for the study of hypertension as a qualitative trait. The genome-wide association analysis of 521,220 SNPs in 3,320 individuals uncovered three SNPs associated with hypertension at a significance level of $P < 5 \times 10^{-7}$. The top hit, rs13333226 on chromosome 16, is located in close proximity to the uromodulin (*UMOD*) transcription start site. Variants in *UMOD* have been associated with chronic kidney disease (CKD) and estimated glomerular filtration rate (eGFR), a quantitative measure of kidney function [202]. In the current study, the association between hypertension and rs13333226 was followed up in a two stage validation analysis in a total of 14 independent cohorts. The results of this and the combined meta-analyses are presented in the following chapter.

4 Validation and clinical functional studies

This chapter summarises the cohorts included in the initial validation analysis and larger meta-analysis of the top hit, as determined by combined association *P*-value. Unadjusted analyses and analyses adjusted for possible confounding variables have been performed. The association results for each cohort and the combined summary measure are presented, as well as measures to assess heterogeneity between component studies.

Functional associations between the top hit, rs13333226, and urinary uromodulin levels were studied in three samples: a hypertensive cohort; a population cohort with extensive urine phenotypes; and an interventional study of low and high salt intake. The latter study recorded extensive measurements of sodium balance that were also analysed. This chapter summarises the cohorts and presents the results.

4.1 Validation cohorts

Table 4.1 provides summary demographics of the validation cohorts.

4.1.1 Malmö Preventive Project

The Malmö Preventive Project (MPP) is a cardiovascular risk screening programme of the general population in Malmö, Sweden [142]. Blood pressure measurement was the mean of two supine readings using a mercury sphygmomanometer. From the sample 1,956 cases and 1,057 controls were included in the replication analysis.

4.1.2 Malmö Diet and Cancer Study

An additional (i.e. not overlapping with discovery samples) 6,977 case individuals and 6,891 controls were selected from the MDC study for replication analysis [145].

4.1.3 MONICA/PAMELA

The World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease (MONICA) Project is an international collaborative study of 37 populations in 21 countries [203, 204]. Recruitment of individuals aged 25-64 years began in 1981. Blood pressure measurement was the mean of two seated readings using a random-zero

Table 4.1 Summary demographics of the validation cohorts. Data are presented as mean (SD).

Study	Controls					Cases				
	N	Age, years	BMI, kg/m^2	SBP, mmHg	DBP, mmHg	N	Age, years	BMI, kg/m^2	SBP, mmHg	DBP, mmHg
BRIGHT/ ASCOT	1787	58.7 (8.92)	25.2 (3.26)	123.0 (10.47)	76.3 (7.19)	3069	60.0 (9.77)	28.1 (4.22)	165.6 (20.35)	99.1 (11.92)
MPP	1057	65.7 (6.4)	25.3 (3.4)	120.4 (6.8)	72.7 (4.9)	1956	67.4 (6.3)	28.3 (4.1)	169.8 (15.6)	98.5 (7.0)
MDC	6891	54.3 (6.7)	24.3 (3.4)	119.9 (7.7)	75.1 (5.1)	6977	60.8 (7.5)	27.0 (4.3)	165.4 (13.5)	97.2 (6.6)
PREVEND	1613	44.6 (10.5)	24.1 (3.4)	109.1 (6.1)	65.9 (5.9)	2411	47.6 (7.7)	27.9 (4.7)	142.3 (17.2)	80.9 (9.6)
CoLaus	1375	49.1 (9.2)	23.6 (3.6)	108.8 (6.7)	68.8 (6.1)	1300	54.8 (8.8)	28.2 (4.9)	141.9 (16.2)	88.1 (10.8)
KORA	300	46.3 (9.2)	25.1 (3.7)	109.9 (6.2)	70.4 (5.6)	457	51.0 (6.6)	28.7 (4.1)	147.5 (15.5)	91.3 (9.6)
SHIP	240	62.1 (9.0)	26.5 (3.9)	110.7 (7.1)	70.1 (6.4)	656	48.2 (7.8)	29.4 (5.2)	144.5 (15.5)	91.7 (9.5)
B58C	529	44.9 (0.3)	25.6 (4.1)	108.7 (5.0)	68.3 (5.2)	514	45.0 (0.3)	29.4 (5.5)	148.1 (11.9)	92.7 (8.2)
TwinsUK	845	45.7 (11.8)	24.8 (4.6)	117.5 (13.3)	74.8 (8.8)	245	47.2 (12.1)	25.1 (4.7)	139.3 (16.0)	88.3 (11.7)
MIGen	278	45.9 (7.0)	25.0 (4.0)	107.3 (7.1)	69.7 (7.0)	316	48.9 (5.9)	29.0 (5.5)	141.4 (14.0)	89.5 (11.3)
DGI	161	60.1 (7.4)	25.5 (3.2)	113.2 (6.8)	70.4 (6.7)	277	52.7 (5.7)	27.6 (3.7)	145.8 (15.0)	87.6 (8.9)
Fenland	510	44.1 (7.4)	25.4 (4.6)	107.2 (6.8)	66.7 (6.2)	264	48.8 (6.4)	29.5 (4.9)	143.7 (14.1)	88.0 (9.4)
MONICA/ PAMELA	746	56.1(5.2	25.4(3.8)	119.6(8.5)	78.4(7.4)	894	55.8(7.2)	27.6(4.4)	156.6(20.1)	94.3(10.7)
NESDA	209	38.6 (11.8)	22.9 (3.4)	111 (4.2)	69.9 (5.2)	509	46.6 (10.9)	27.1 (5.1)	149.6 (14.8)	88.8 (9.5)

BRIGHT = British Genetics of Hypertension Study. ASCOT = Anglo-Scandinavian Cardiac Outcomes Trial. MPP = Malmö Preventive Project. MDC = Malmö Diet and Cancer Study. PREVEND = Prevention of Renal and Vascular End Stage Disease Study. CoLaus = Cohorte Lausannoise. KORA = Kooperative Gesundheitsforschung in der Region Augsburg. SHIP = Study of Health in Pomerania. B58C = British 1958 Birth Cohort. MIGen = Myocardial Infarction Genetics Consortium. DGI = Diabetes Genetics Initiative. MONICA = World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project. PAMELA = Pressioni Arteriose Monitorate e Loro Associazioni. NESDA = Netherlands Study of Depression and Anxiety.

sphygmomanometer or a mercury sphygmomanometer. Participants from the MONICA study included in the current validation analysis are all Italian.

The PAMELA (Pressioni Arteriose Monitorate e Loro Associazioni) study is a prospective general population study that randomly sampled residents aged 25-64 years in the Italian city of Monza [205]. It compared ambulatory and home blood pressure measurement with clinic blood pressure. In the current analysis phenotype was determined using clinic measurements.

In the current analysis 894 cases and 746 controls were included from MONICA and PAMELA collectively.

4.1.4 Netherlands Study of Depression and Anxiety

The Netherlands Study of Depression and Anxiety (NESDA) is a cohort study of individuals aged 18-65 years [206]. Its aim is to examine the course of depressive and anxiety disorders throughout life. Blood pressure measurement was the mean of two supine readings using the Omron HEM-907XL machine. From the sample 509 participants met our case criteria and 209 our control criteria, and were included in the replication analysis.

4.1.5 BRIGHT/ASCOT

As previously described, the BRIGHT study is a UK hypertension case-control study with the following exclusion criteria; BMI $\geq$35 kg/m^2, diabetes, secondary hypertension or a co-existing illness (http://www.brightstudy.ac.uk)[128]. Blood pressure measurement was the mean of three seated readings using the Omron 705CP machine.

The Anglo-Scandinavian Cardiac Outcomes Trial (ASCOT) is a study of 19,342 hypertensive patients in the UK, Ireland, and Scandinavia (http://www.ascotstudy.org/) [207]. Its initial primary aims were to investigate the effects of different categories of blood pressure lowering medication and statins on non-fatal MI and fatal CHD. All participants had at least three pre-specified risk factors for a CV event, in addition to hypertension, at the time of recruitment. Blood pressure was measured using the Omron HEM-605CP machine at screening and at randomisation. Individuals defined as hypertensive at both visits were included in the study.

In the current analysis BRIGHT and ASCOT cases were combined and compared with BRIGHT controls. 3,069 and 1,787 individuals, respectively, met our case and control criteria and were included in the replication analysis.

4.1.6 Prevention of Renal and Vascular End Stage Disease Study

The Prevention of REnal and Vascular ENd stage Disease (PREVEND) study is a prospective general population study of individuals aged 28-75 years in Groningen, The Netherlands [208, 209]. Its aim is to investigate the natural course of increased levels of urinary albumin excretion and its association with renal and cardiovascular disease. 47% of eligible individuals responded to a questionnaire of which cases were selected with urinary albumin concentration ≥ 10 mg/L and controls with urinary albumin concentration < 10 mg/L. Blood pressure was measured every minute for 10 and 8 minutes, respectively, in the supine position using an automatic DINAMAP XL Model 9300 series monitor, and the mean of the last two measures used in analysis. For the current replication analysis 2,411 cases and 1,613 controls were included who met our inclusion criteria.

4.1.7 Cohorte Lausannoise

Cohorte Lausannoise (CoLaus) is a population based study of the Caucasian population in Lausanne, Switzerland (http://www.colaus.ch/en/cls_home/cls_pro_home.htm) [210], where Caucasian is defined as both parents and grandparents born in any of a list of predefined European countries. Its primary aims were to assess the prevalence and molecular determinants of cardiovascular risk factors and diseases as well as mental health. Participants were randomly selected from the population register of people in Lausanne aged 35-75 years in 2003. Blood pressure was measured three times in the seated position using the Omron HEM-907 machine, and the mean of the last two measures used in analysis. From the sample 1,300 and 1,375 individuals, respectively, met our case and control criteria and were included in the replication analysis.

4.1.8 Kooperative Gesundheitsforschung in der Region Augsburg

Kooperative Gesundheitsforschung in der Region Augsburg (KORA) is a general population cohort in Augsburg, Germany recruited in 1994-1995 (http://epi.helmholtz-muenchen.de/kora-gen/)[211, 212]. Blood pressure measurement was the mean of two seated readings using a random zero sphygmomanometer. Participants for the current replication

analysis were selected from a subset of individuals with BMI <35 kg/m^2 and no diabetes. From the subsample 457 cases and 300 controls met our inclusion criteria.

4.1.9 Study of Health in Pomerania

The Study of Health in Pomerania (SHIP) is a population based sample of individuals aged 20-79 years in north-east Germany, drawn from population registries in the region [213]. Blood pressure was measured three times in the seated position at three minute intervals using the Omron HEM-705CP machine, and the mean of the last two measures used in analysis. From the sample 656 cases and 240 controls were included in the replication analysis.

4.1.10 British 1958 Birth Cohort

The British 1958 Birth Cohort (B58C) is a population based sample of all individuals born in a single week in Britain in 1958, followed from birth to age 44-45 years (http://www.b58cgene.sgul.ac.uk/collection.php). Blood pressure measurement was the mean of three seated recordings using the Omron 705CP machine. From the sample 514 participants met our case criteria and 529 our control criteria, and were included in the replication analysis.

4.1.11 TwinsUK

The TwinsUK study is a sample of healthy female Caucasians aged 18-76 years recruited from the general population through the TwinsUK registry (http://www.twin-research.ac.uk). Blood pressure was measured three times in the seated position using the Omron HEM-907 machine, and the mean of the last two measures used in analysis. Participants for the current study were selected from a subset of individuals, one of each twin pair. From the subsample 245 participants met our case criteria and 845 our control criteria, and were included in the replication analysis.

4.1.12 Myocardial Infarction Genetics Consortium

The Myocardial Infarction Genetics Consortium (MIGen) cohort is a subset of the controls from a study to identify genetic variants associated with early-onset MI. Most of the controls were selected from population based studies and came from the USA, Spain,

Finland, and Sweden. In the majority of studies blood pressure was the mean of two seated recordings using calibrated sphygmomanometers. From the sample 316 and 278 individuals, respectively, met our case and control criteria and were included in the replication analysis.

4.1.13 Diabetes Genetics Initiative

The Diabetes Genetics Initiative (DGI) is a type 2 diabetes (T2D) case-control study of individuals from Sweden and Finland [120]. Blood pressure measurement was the mean of two seated recordings using a mercury sphygmomanometer. For the current study participants were selected from the controls (i.e. T2D free), of which 277 hypertension cases and 161 normotensive controls were included in the replication analysis.

4.1.14 Fenland Study

The Fenland Study is a UK population based cohort study of individuals aged 30-55 years. Its aim is to investigate the influence of genetic and environmental factors on the risk of obesity, insulin sensitivity, hyperglycemia and related metabolic traits. Blood pressure measurement was the mean of three seated recordings at one minute intervals, using an Accutorr automated sphygmomanometer. From the sample 264 cases and 510 controls were included in the current replication analysis.

4.2 Validation analysis

The top hit was rs13333226 with the minor G allele associated with a lower risk of hypertension (OR = 0.67, 95% CI = 0.58 – 0.78, $P = 1.14 \times 10^{-7}$), and was selected for validation in two stages. It is located on chromosome 16 in close proximity, at -1617 base pairs, to the *UMOD* transcription start site (Figure 4.1). The genotype cluster plot for rs1333226 in the discovery sample is shown in Figure 4.2. Table 4.2 presents the association results for each stage of the discovery and validation process as well as the combined results. In stage 1 validation rs13333226 was genotyped in the combined MONICA/PAMELA sample, the MPP sample, and in the larger additional MDC sample. In a combined analysis of 9,827 cases and 8,694 controls, the minor G allele remained associated with a lower risk of hypertension (OR = 0.87, 95% CI = 0.82 – 0.92, $P = 3.6 \times$

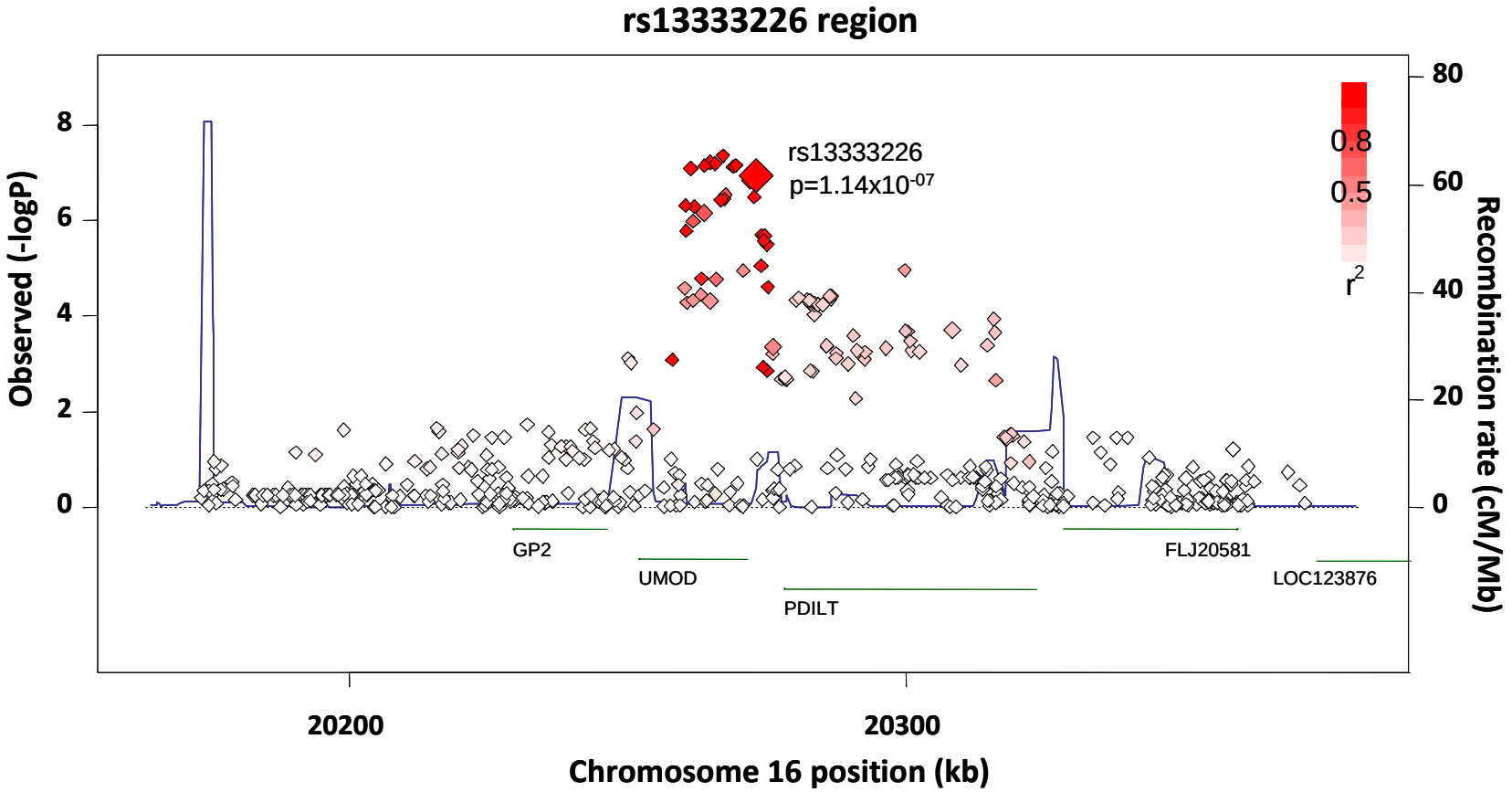


Figure 4.1 Association plot of the genomic region around rs13333226 showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs13333226 and the surrounding SNPs, measured by r 2, is indicated by the key with red meaning high LD. GP2 = glycoprotein 2 (zymogen granule membrane). UMOD = uromodulin. PDILT = protein disulfide isomerase-like, testis expressed. FLJ20581 = acyl-CoA synthetase medium-chain family member 5. LOC123876 = acyl-CoA synthetase medium-chain family member 2A.

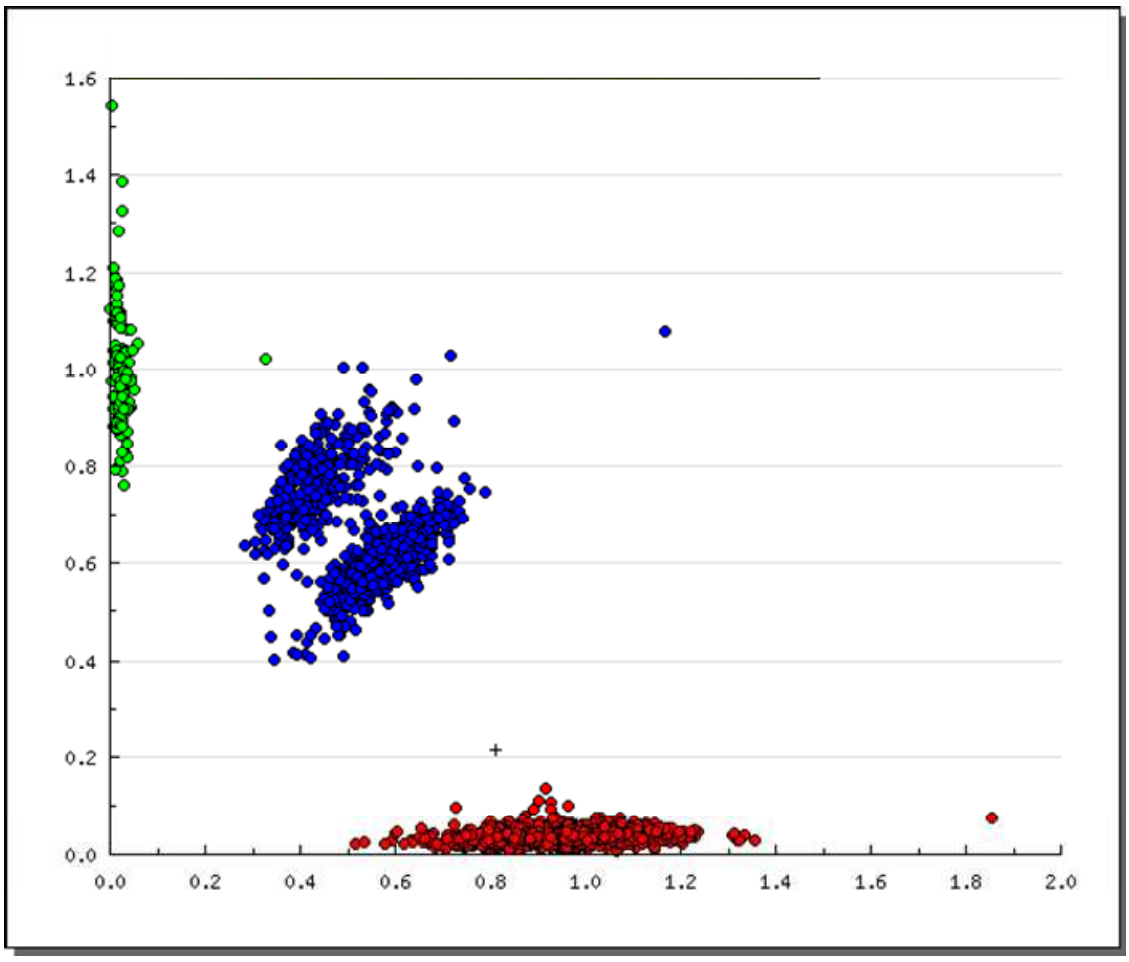


Figure 4.2 Genotype cluster plot for rs1333226 in the discovery sample. Each data point represents a person, where green and red circles have been identified by Illumina as the two homozygotes and blue the heterozygote. The black cross represents an individual that has been excluded by Illumina.

Table 4.2 Results from the meta-analysis of rs13333226 and hypertension in the discovery sample and after validation

Study	origin	cases	controls	maf	Unadjusted Analysis		Adjusted for age, age^2, sex BMI		Q (Unadj/Adj)
					OR [95%-CI]	p	OR [95%-CI]	p	
Swedish BP Extremes (Discovery)	Swedish	1621	1699	0.17	0.67 [0.58-0.78]	$1.14x10^{-07}$	0.6 [0.5-0.73]	$3.3 x10^{-07}$	
Stage 1									
MONICA/ PAMELA	Italian	894	746	0.19	0.91 [0.76-1.08]	0.282	0.87 [0.72-1.05]	0.145	
MPP	Swedish	1956	1057	0.18	0.91 [0.78-1.05]	0.193	0.91 [0.78-1.05]	0.186	
MDC	Swedish	6977	6891	0.18	0.86 [0.80-0.92]	0.001	0.86 [0.80-0.92]	$3.0x10^{-05}$	
Stage 1 Analysis		9827	8694	0.183	0.87 [0.82-0.93]	$6.7x10^{-6}$	0.87 [0.82-0.92]	$3.6x10^{-6}$	0.73/0.81
Stage 1 + Discovery		21275	19087	0.18	0.84 [0.79-0.89]	$4.4x10^{-10}$	0.84 [0.79-0.89]	$2.5x10^{-9}$	0.01/0.01
Stage 2									
BRIGHT/ ASCOT	British/ Irish	3069	1787	0.18	0.94 [0.84-1.04]	0.229	0.9 [0.80-1.02]	0.103	
PREVEND	Dutch	2411	1613	0.18	0.9 [0.80-1.02]	0.091	0.89 [0.77-1.03]	0.113	
CoLaus	Swiss	1300	1375	0.19	0.97 [0.84-1.11]	0.634	0.93 [0.79-1.1]	0.375	

KORA	German	457	300	0.16	0.8 [0.61-1.06]	0.128	0.7 [0.51-0.97]	0.03	
SHIP	German	656	240	0.18	1.07 [0.81-1.41]	0.627	0.74 [0.50-1.1]	0.137	
B58C	British	514	529	0.19	0.82 [0.66-1.02]	0.077	0.77 [0.61-0.97]	0.026	
TwinsUK	British	245	845	0.19	0.88 [0.68-1.14]	0.332	0.84 [0.63-1.12]	0.236	
MIGen	European Ancestry	316	278	0.21	0.68 [0.51-0.9]	0.004	0.61 [0.44-0.84]	0.002	
DGI	Swedish/ Finnish	277	161	0.23	1.11 [0.77-1.62]	0.572	1.15 [0.78-1.68]	0.483	
Fenland	British	264	510	0.19	0.91 [0.69-1.19]	0.478	0.8 [0.58-1.09]	0.158	
NESDA	Dutch	509	209	0.18	0.98 [0.73-1.31]	0.898	0.93 [0.63-1.35]	0.689	
Stage 2 Analysis		10018	7847	0.189	0.91 [0.86-0.96]	0.0019	0.86 [0.81-0.92]	$1.0x10^{-5}$	0.5/0.3
Stage 2 Analysis + Discovery		11639	9546	0.188	0.88 [0.83-0.93]	$1.2x10^{-6}$	0.83 [0.78-0.88]	$5.4x10^{-9}$	0.01/0.02
Combined Analysis - Stage 1 + Stage 2		19845	16541	0.188	0.89 [0.85-0.93]	$8.98x10^{-08}$	0.86 [0.83-0.90]	$1.61x10^{-10}$	0.52/0.51
Combined Analysis - Discovery + Stage 1 + Stage 2		21466	18240	0.187	0.87 [0.84- 0.91]	$3.67x10^{-11}$	0.85 [0.81- 0.89]	$1.5x10^{-13}$	0.02/0.04

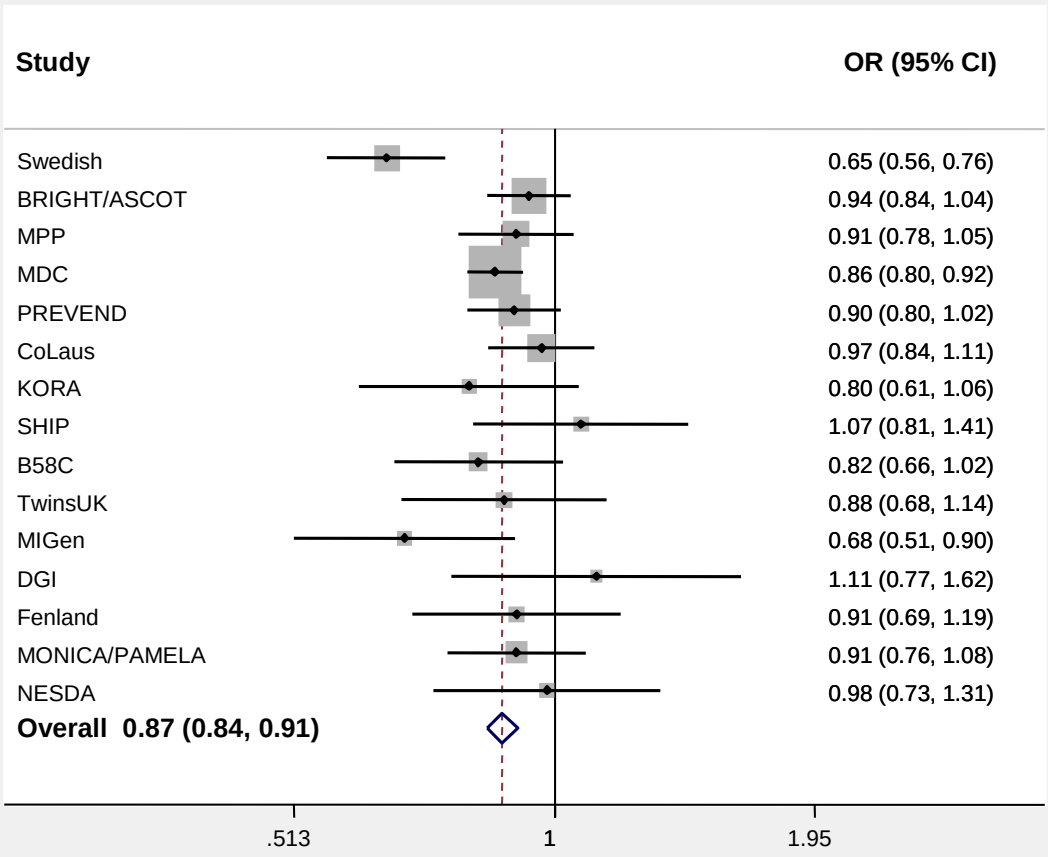
Q(unadj/adj) = P value of the meta-analysis Q test for heterogeneity for the unadjusted and adjusted meta-analysis respectively. OR = odds ratio.CI = confidence interval. MONICA = World Health Organization Monitoring Trends and Determinants in Cardiovascular Disease Project. PAMELA = Pressioni Arteriose Monitorate e Loro Associazioni. MPP = Malmö Preventive Project. MDC = Malmö Diet and Cancer Study. BRIGHT = British Genetics of Hypertension Study. ASCOT = Anglo-Scandinavian Cardiac Outcomes Trial. PREVEND = Prevention of Renal and Vascular End Stage Disease Study. CoLaus = Cohorte Lausannoise. KORA = Kooperative Gesundheitsforschung in der Region Augsburg. SHIP = Study of Health in Pomerania. B58C = British 1958 Birth Cohort. MIGen = Myocardial Infarction Genetics Consortium. DGI = Diabetes Genetics Initiative. NESDA = Netherlands Study of Depression and Anxiety.

10^{-6}) after adjustment for age, age^2, sex, and BMI. These adjustments were made because not all of the replication cohorts met our strict criteria for age cut-offs, and because these covariates are known to associate with blood pressure. Stage 2 analysis was conducted on a total sample of 10,018 cases and 7,847 controls from the eleven remaining cohorts; BRIGHT/ASCOT, PREVEND, CoLaus, KORA, SHIP, B58C, TwinsUK, MIGen, DGI, Fenland, and NESDA. The results were similar with the G allele again associated with reduced risk of hypertension (adjusted OR = 0.86, 95% CI = 0.81 – 0.92, $P = 1.0 \times 10^{-5}$). When stage 1 and stage 2 were combined the effect size was unchanged but strength of the association was greater (OR = 0.86, 95% CI = 0.83 – 0.90, $P = 1.61 \times 10^{-10}$). As assessed by the Q statistic, there was no evidence of heterogeneity between studies in stage 1, stage 2, or the combined stage 1 and stage 2 samples ($P > 0.10$). The meta-analyses of all validation samples and the discovery sample are presented in more detail below.

4.3 Unadjusted meta-analysis of rs13333226

The results of the meta-analysis of the crude ORs and 95% CIs for the association between hypertension status and rs13333226, for the discovery sample and all 14 replication cohorts, are presented in Figure 4.3. In total 21,466 cases and 18,240 controls were included. The summary estimate was 0.87 (95% CI 0.84 – 0.91, $P = 3.67 \times 10^{-11}$). The funnel plot was roughly symmetrical, indicating that there was no evidence of bias due to overestimation of effect size in smaller samples. The data point lying outside the 95% confidence interval represents the discovery sample. The Q statistic is 27.34, $P = 0.017$, i.e. there was significant evidence of heterogeneity between studies. Furthermore the I^2 statistic was 48.8%, suggesting a moderate level of heterogeneity. The funnel plot shows the discovery (i.e. Swedish) sample lying outside the odds ratio 95% confidence limits, because it has a greater effect size than the validation cohorts. This suggests that the observed heterogeneity may be due to the extreme case-control study design of the discovery sample, the winner's curse, or a combination of both. Therefore it was excluded and the meta-analysis repeated for the validation cohorts (Figure 4.4). This slightly decreased the effect size to OR = 0.89 (95% CI 0.85 – 0.93, $P = 8.98 \times 10^{-08}$). The Q statistic was 11.46, $P = 0.572$, and the I^2 statistic 0.0%, i.e. no evidence of true heterogeneity, indicating that all of the variability in effect size estimates is due to sampling error. Hence it can be concluded that the initial heterogeneity was due to the inclusion of the discovery cohort. After its omission there remained a significant association between rs13333226 genotype and hypertension status. All cohort ORs now lie within the 95% confidence limits of the funnel plot.

a.



b.

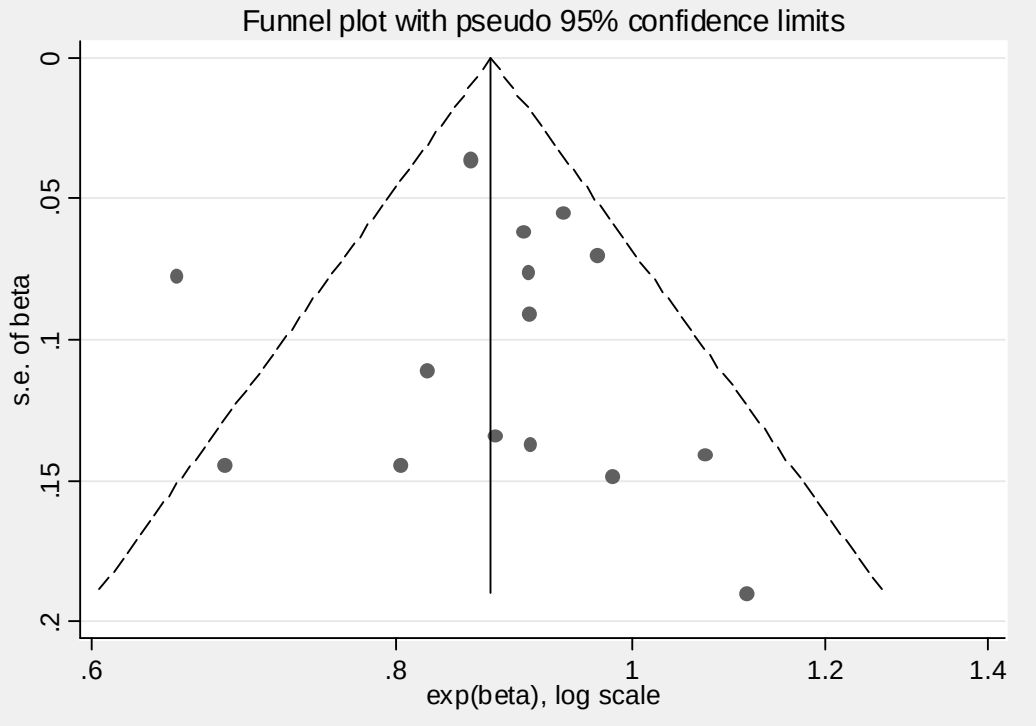
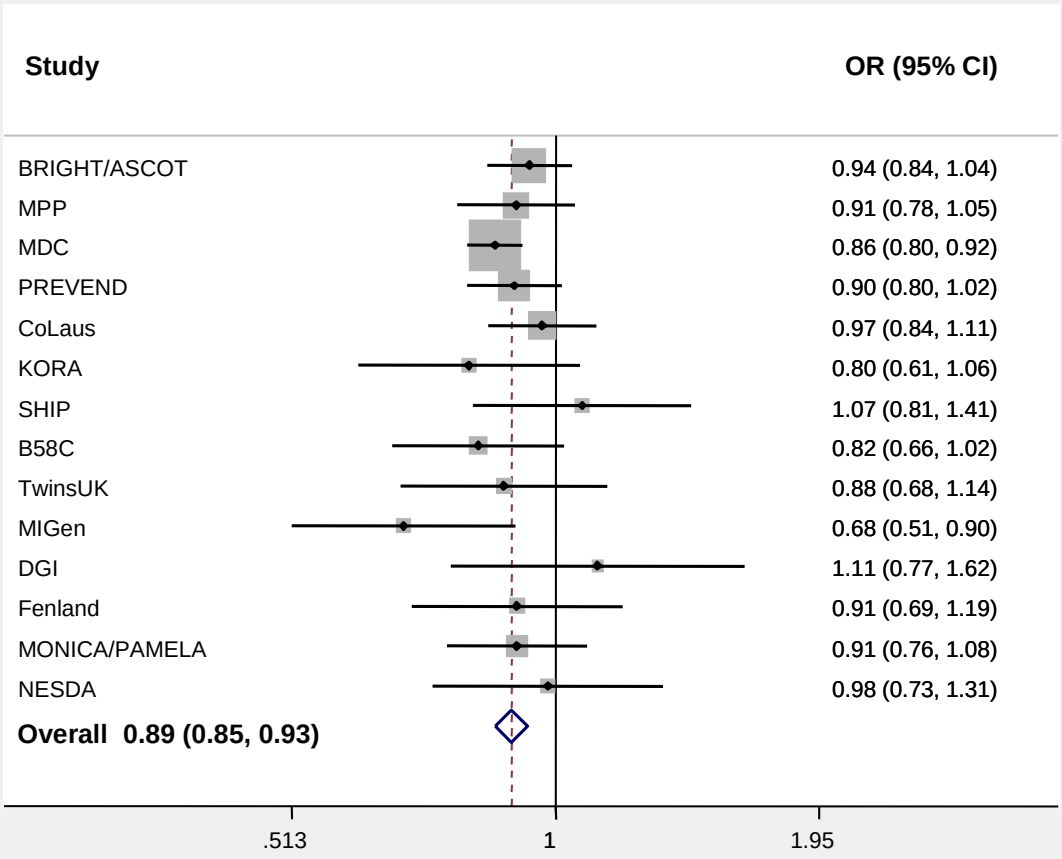


Figure 4.3 Results of meta-analysis of crude odds ratios for the association between rs13333226 and hypertension status in Swedish discovery sample and 14 replication cohorts. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

a.



b.

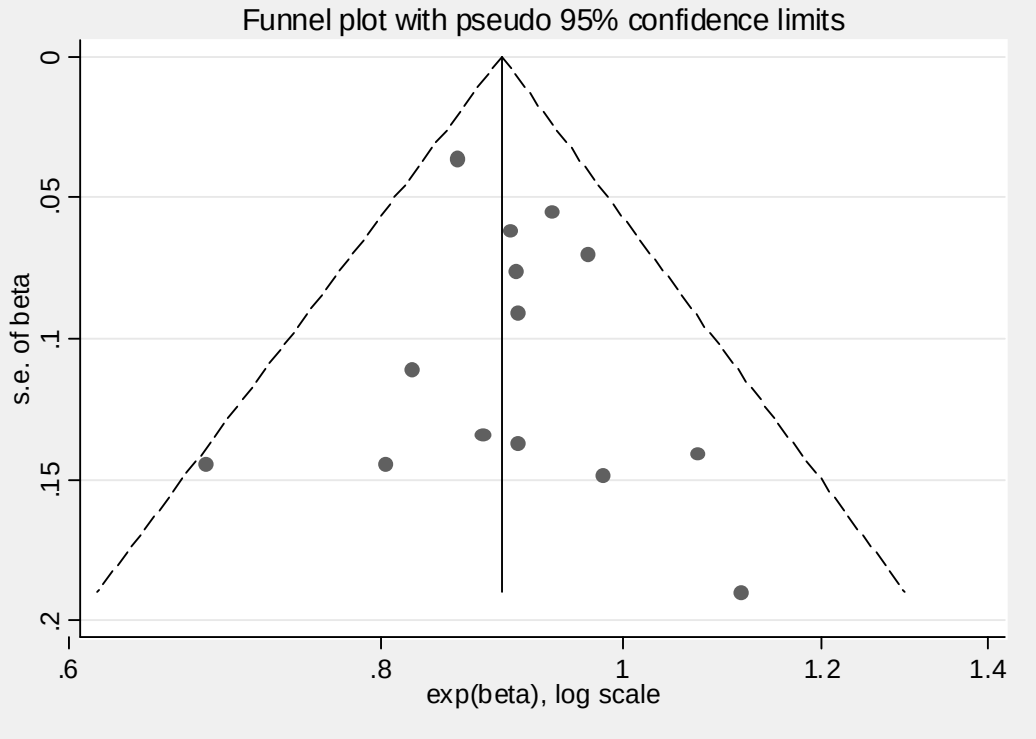


Figure 4.4 Results of meta-analysis of crude odds ratios for the association between rs13333226 and hypertension status in 14 replication cohorts, with exclusion of the Swedish discovery sample. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

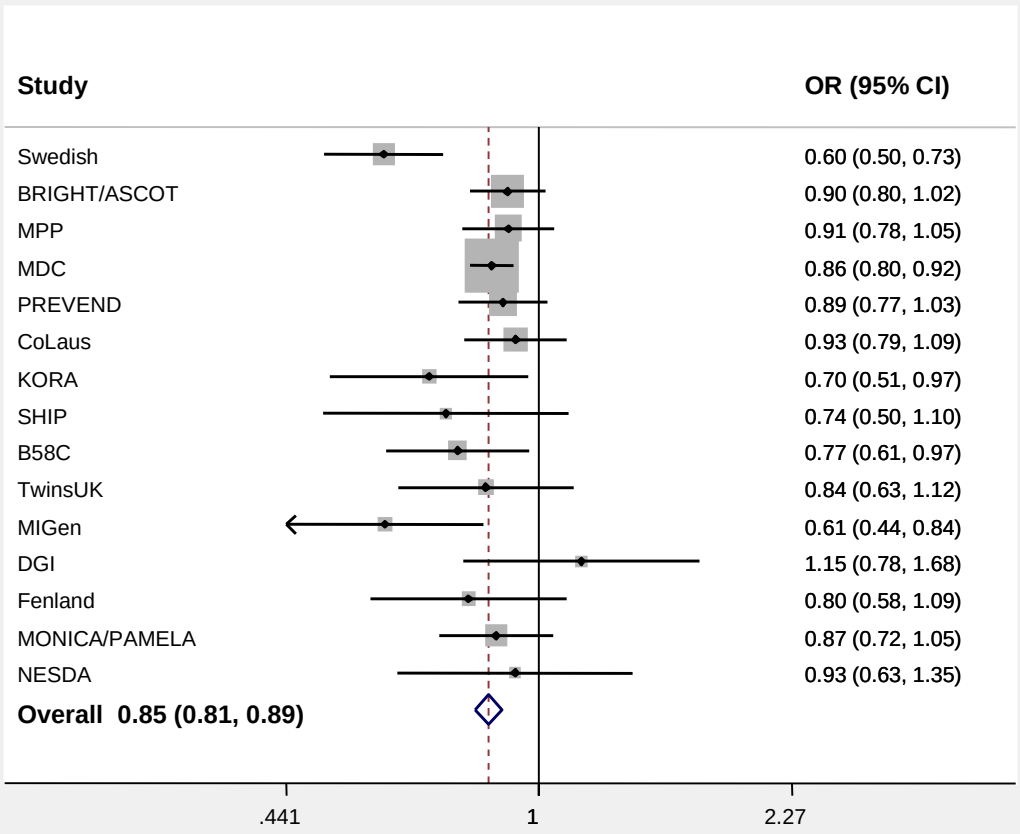
4.4 Analysis of rs13333226 adjusted for age, age^2, sex, and BMI

Adjustment for the potential covariates, age, age^2, sex, and BMI, strengthened the overall meta-analysed association for the discovery sample and replication cohorts (OR = 0.85, 95% CI 0.81 – 0.89, $P = 1.51 \times 10^{-13}$) (Figure 4.5). Again 21,466 cases and 18,240 controls were included. The funnel plot was approximately symmetrical. There was significant heterogeneity present (Q = 24.81, $P = 0.037$; $I^2 = 43.6\%$). The exclusion of the discovery sample again led to a slight decrease in effect size (OR = 0.86, 95% 0.83 – 0.90, $P = 1.61 \times 10^{-10}$) (Figure 4.6), and the disappearance of heterogeneity (Q = 12.17, $P = 0.514$; $I^2 = 0.0\%$).

4.5 Analysis of rs13333226 adjusted for age, age^2, sex, BMI, and eGFR

Values for eGFR were available in only seven cohorts; PREVEND, CoLaus, SHIP, DGI, Fenland, MONICA/PAMELA, and MPP. It was not recorded in the discovery sample. Thus the sample size for this analysis was reduced to 5739 cases and 7427 controls. When these seven ORs from the association analysis adjusted for age, age^2, sex, and BMI were meta-analysed the overall association remained significant (OR = 0.90, 95% CI 0.83 – 0.97 $P = 0.004$; Q = 3.32, $P = 0.767$; $I^2 = 0.0\%$) (Figure 4.7). Additional adjustment for eGFR strengthened the effect marginally (OR = 0.90, 95% CI 0.85 – 0.95, $P < 0.001$; Q = 2.41, $P = 0.879$; $I^2 = 0.0\%$) (Figure 4.8). There was no evidence of heterogeneity between studies in either analysis.

a.



b.

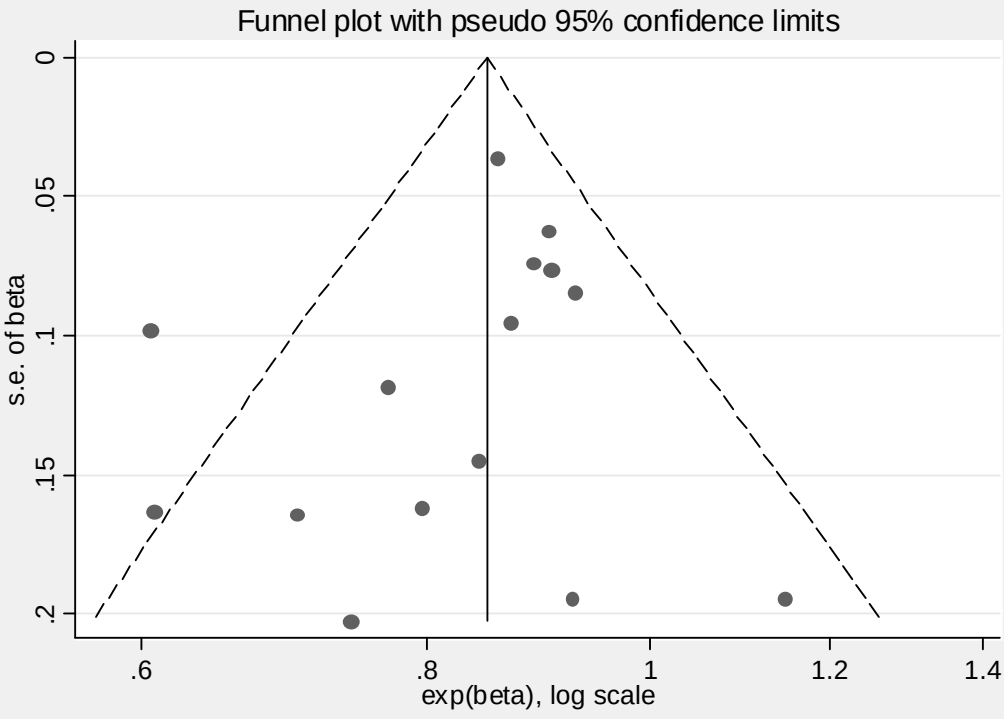
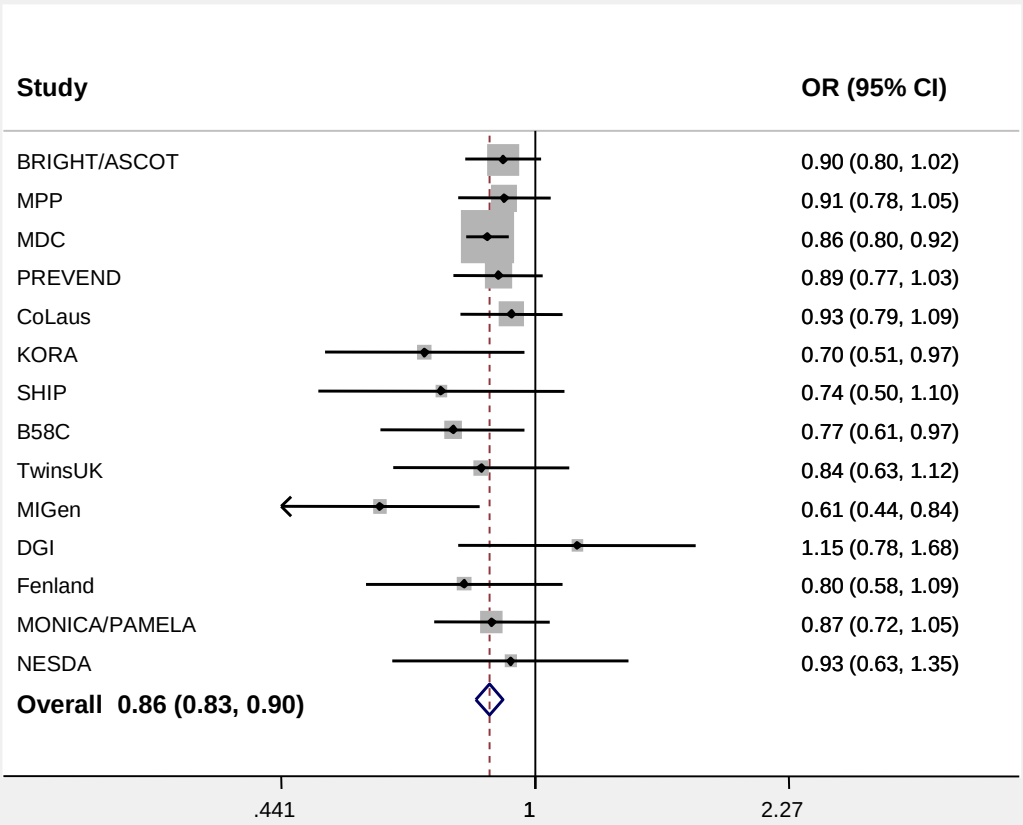


Figure 4.5 Results of meta-analysis of odds ratios for the association between rs13333226 and hypertension status in Swedish discovery sample and 14 replication cohorts, adjusted for age, age^2, sex, and BMI. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

a.



b.

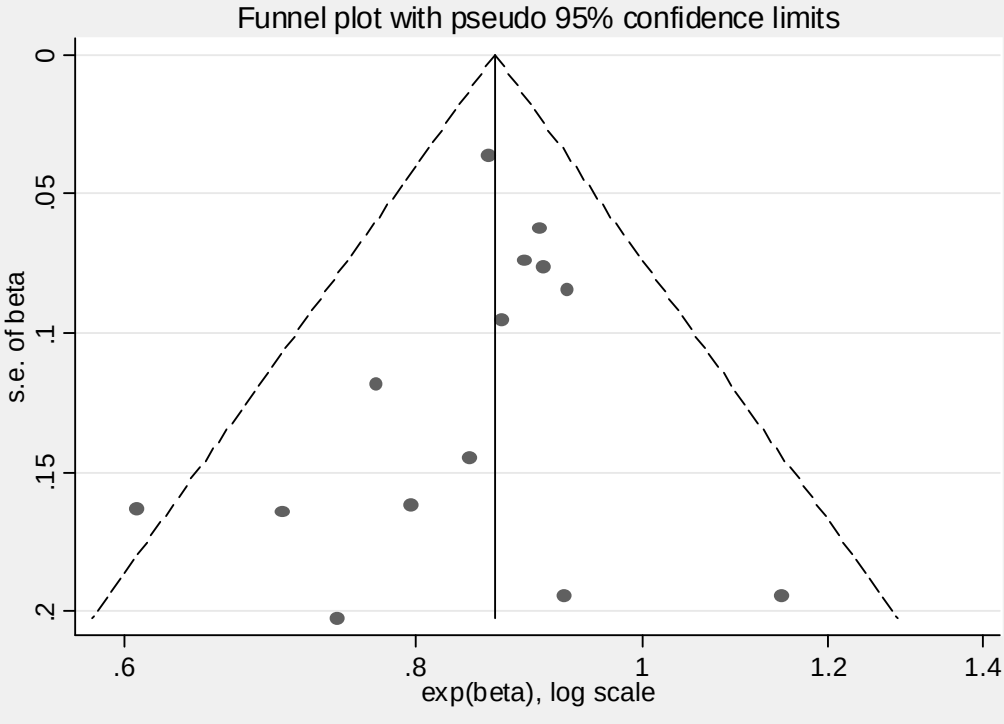
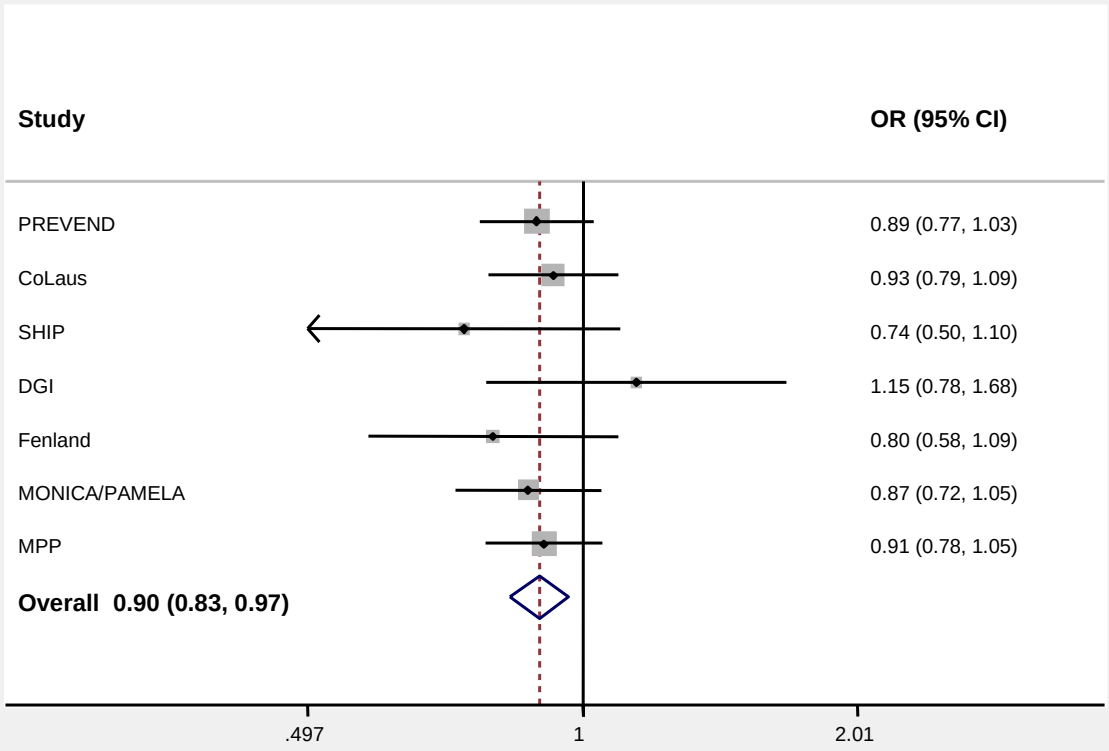


Figure 4.6 Results of meta-analysis of odds ratios for the association between rs13333226 and hypertension status in 14 replication cohorts, with exclusion of the Swedish discovery sample, adjusted for age, age^2, sex, and BMI. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

a.



b.

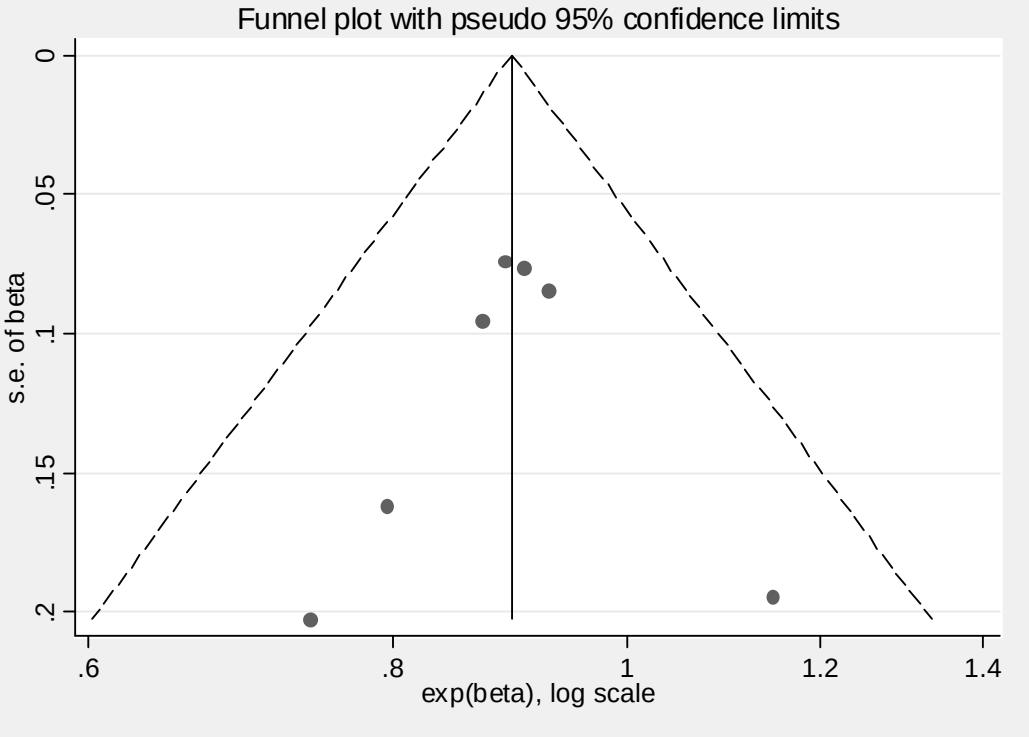
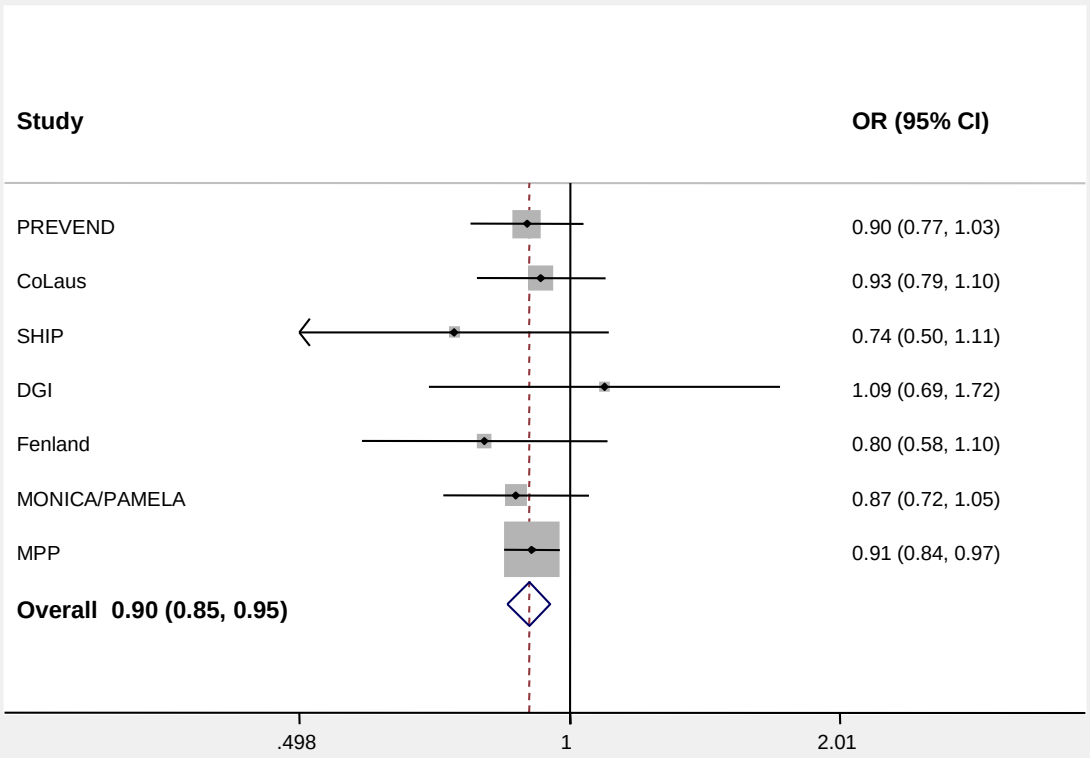


Figure 4.7 Results of meta-analysis of odds ratios for the association between rs13333226 and hypertension status in 7 replication cohorts (those with eGFR available), adjusted for age, age^2, sex, and BMI. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

a.



b.

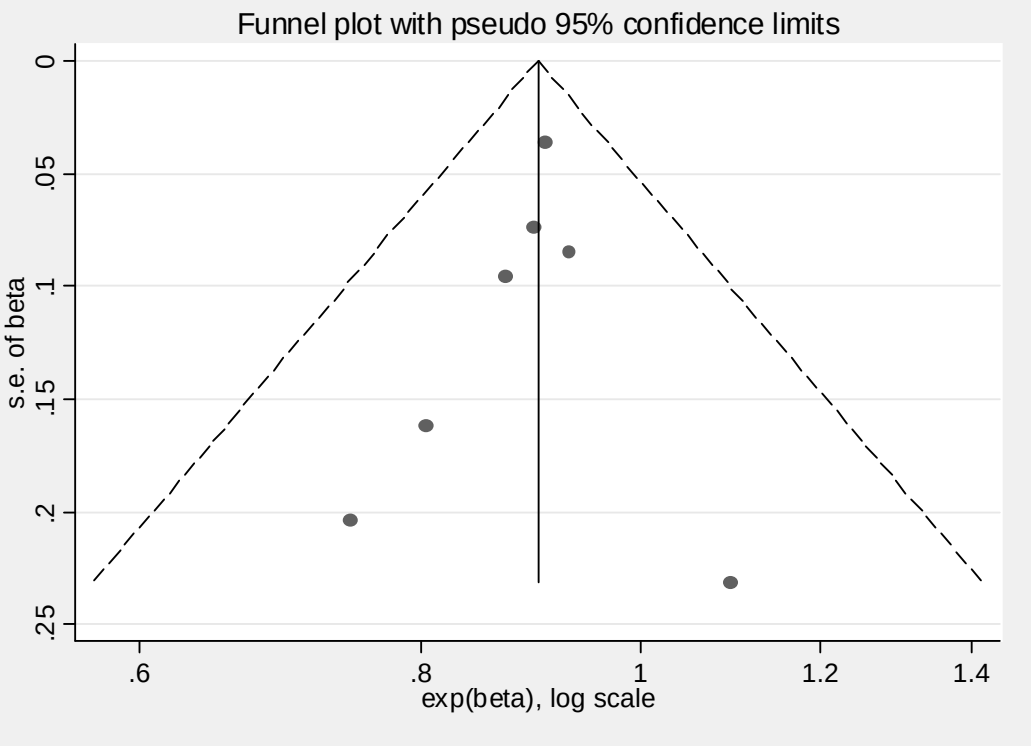


Figure 4.8 Results of meta-analysis of odds ratios for the association between rs13333226 and hypertension status in 7 replication cohorts (those with eGFR available), adjusted for age, age^2, sex, BMI, and eGFR. a) forest plot of odds ratios and 95% confidence intervals for individual studies summary result; b) funnel plot of standard error of coefficient (y axis) against odds ratio (x axis) for individual studies, with 95% confidence interval. Vertical line represents summary odds ratio.

4.6 Clinical functional cohorts

4.6.1 British Genetics of Hypertension study

As previously described, the BRIGHT study is a UK hypertension case-control study with the following exclusion criteria; BMI ≥35 kg/m^2, diabetes, secondary hypertension or a co-existing illness (http://www.brightstudy.ac.uk)[128]. Cases were individuals with a diagnosis of hypertension, defined as >150/100 mmHg, prior to 50 years of age. All cases completed 24-hour urine collection with urinary sodium, potassium, creatinine and microalbuminuria recorded. In the current study urinary uromodulin was measured in 256 hypertensive participants.

4.6.2 Hypertension Evaluation by Remler and CalciUria LEvel Study

Hypertension Evaluation by Remler and CalciUria LEvel Study (HERCULES) is a subsample of 400 randomly selected participants of the CoLaus study [210]. HERCULES (http://www.colaus.ch/en/cls_home/cls_pro_home/cls_pro_studies/cls_pro_studies-hercules.htm) [176] aims to study the prevalence of hypertension using 24-hour ambulatory blood pressure measurement, assess renal function using 24-hour urine collection, and expand understanding of genetic variants associated with hypertension and renal function within CoLaus. Hence as a sample it is ideal for the current study's purposes. Again 24-hour urine was collected and the phenotypes recorded include urinary sodium, creatinine clearance, endogenous lithium clearance, potassium and uric acid excretion, and microalbuminuria. The current study measured urinary uromodulin in 110 HERCULES participants.

4.6.3 Groningen Renal Hemodynamic Cohort Study Group

The Groningen Renal Hemodynamic Cohort Study Group (GRECO) is carrying out a series of studies examining blood pressure, renal function, and extracellular volume following sodium controlled diets [177, 178]. Dietary compliance and the achievement of a stable sodium balance are assessed via 24-hour urine. The current study analysed GRECO

data for 64 healthy males from a crossover protocol consisting of two 7-day periods, one a high sodium diet (HS; 200 mmol Na^+/day), and the other a low sodium diet (LS; 50 mmol Na^+/day).

4.7 Clinical functional results

The primary results of the BRIGHT sample analysis are presented in Table 4.3. The G allele was significantly associated with higher eGFR (P = 0.005), higher creatinine clearance (P = 0.004), and lower FENa (P = 0.032). There was no association between genotype and uromodulin excretion (P = 0.234).

Table 4.4 presents the results of the HERCULES sample analysis. The G allele was significantly associated with lower urinary uromodulin excretion both as a single measurement (P = 0.005) and as averaged over 24 hours (P = 0.006). There was no association between genotype and creatinine clearance (P = 0.866).

The results from the GRECO interventional study are presented in Table 4.5. Urinary uromodulin excretion was significantly lower in the presence of the G allele following the low salt diet (P=0.002). However, following the high salt diet no such difference was observed (P=0.513).

Table 4.3 Univariate association analysis of rs13333226 in 256 hypertensive patients from the BRIGHT study.

	AA (n=141)	**AG (n=93)**	**GG (n=22)**	***P***
Male:Female	0.7	0.8	0.6	0.763
Age (years)	64.7 (8.4)	63.9 (7.8)	59.5 (9.5)	**0.036**
SBP (mmHg)	156 (19.5)	151.5 (18.9)	153.3 (14.5)	0.205
DBP (mmHg)	93.1 (10)	90.9 (10.7)	93.3 (10.3)	0.266
Body mass index (Kg/m^2)	26.8 (4.6)	26.8 (5.4)	27.2 (3.9)	0.927
Body surface area (m^2)	1.8 (0.2)	1.9 (0.2)	1.8 (0.2)	0.494
Sodium (mmol/L)	138.6 (3.1)	138.9 (3.0)	137.8 (2.9)	0.341
Potassium (mmol/L)	4.4 (0.9)	4.2 (0.8)	4.4 (1.0)	0.429
Urea (umol/L)	6.3 (1.6)	5.7 (1.6)	6.0 (1.6)	**0.025**
Creatinine (mmol/L)	92.2 (21.7)	88.4 (18.7)	82.9 (20)	0.096
Urate (mmol/L)	0.3 (0.1)	0.3 (0.1)	0.3 (0.1)	0.726
eGFR ($ml/min/1,73m^2$)	67.6 (16.2)	70.3 (12.3)	79.5 (15.2)	**0.005**
Creatinine Clearance (ml/min)	70.6 (20.3)	76.2 (20)	86.6 (26.6)	**0.004**
Urine Sodium (mmol/24h)	139.1 (61.9)	158.9 (70.6)	142.4 (58.3)	0.073
Urine Potassium (mmol/24h)	66.4 (24.1)	78.8 (54)	69.2 (18.8)	**0.050**
Creatinine excretion (mmol/24h)	10.2 (3.6)	10.8 (4.6)	10.7 (3.1)	0.520
Uromodulin (mg/L)	5.3 (5.3)	5.2 (5.5)	3.2 (3.4)	0.234
Fractional Excretion of Sodium (%)	0.92 (0.37)	0.95 (0.36)	0.73 (0.19)	**0.032**

With the exception of male:female, data are presented as mean (standard deviation).

Table 4.4 Univariate association analysis of rs13333226 in 110 participants from the HERCULES Study.

	AA (n=53)	**AG (n=45)**	**GG (n=12)**	***P***
M / F (n)	29 / 24	17 / 28	7 / 5	0.187
Age (years)	59 (49 - 67)	56 (49 - 66)	59 (49 - 66)	0.845
Body mass index (Kg/m2)	26.1 (23.6 - 29.3)	24.2 (21.8 – 29.0)	24.7 (24.0 – 28.0)	0.130
Body surface area (m2)	1.85 (1.72 - 1.99)	1.75 (1.62 - 1.91)	1.87 (1.77 – 2.0)	**0.028**
24h SBP (mmHg)	115.5 (107.7 - 123.2)	113 (105.8 - 125.6)	118.4 (111.4 - 130.7)	0.562
24h DBP (mmHg)	76.3 (69.7 - 80.4)	76.8 (71.1 - 85.2)	77.7 (71.1 - 87.7)	0.495
Hypertension (%)	37	50	25	0.871
Fasting plasma				
Sodium (mmol/L)	139.1 (137.1 - 140.9)	139.5 (137.8 - 140.9)	138.9 (138.1 - 141.5)	0.869
Potassium (mmol/L)	4.0 (3.9 - 4.3)	4.0 (3.7 - 4.1)	3.7 (3.5 – 4.0)	0.059
Urea (umol/L)	5.2 (4.4 - 6.1)	4.8 (4.4 – 6.0)	4.3 (4.1 - 4.6)	0.090
Creatinine (mmol/L)	81.0 (73.0 – 89.0)	80.5 (73.0 – 88.0)	77.5 (75.0 - 81.5)	0.787
Urate (mmol/L)	316 (287 - 378)	317 (262 - 378)	294 (274 - 317)	0.244
24 h urine				
Uromodulin (mg/L)	30.6 (14.9 - 49.7)	24.7 (14.2 - 42.5)	14 (10.6 - 16.5)	**0.005**
Uromodulin (mg/24h)	53 (25 - 75)	39 (28 - 68)	17 (14 - 33)	**0.006**
Urine volume (mL)	1700 (1200 - 2350)	1600 (1150 - 2050)	1773 (1125 - 2300)	0.780
Creatinine clearance (mL/min)	96.6 (69.8 - 122.3)	98.8 (75.0 - 122.8)	99.1 (79.2 - 128.7)	0.866
Creatinine excretion (mmol/kg/24h)	0.15 (0.10 - 0.19)	0.16 (0.14 - 0.19)	0.15 (0.12 - 0.19)	0.447
Urine Sodium (mmol/24h)	147.35 (96.14 - 187.81)	147.5 (110.13 - 177.18)	103.71 (81.49 - 144.70)	0.322
Urine Potassium (mmol/24h)	61.86 (50.08 - 84.46)	64.16 (54.9 - 74.47)	47.88 (38.06 - 93)	0.662
Fractional Excretion of Sodium	0.013 (0.007 - 0.018)	0.012 (0.009 - 0.017)	0.006 (0.005 - 0.007)	0.130

With the exception of M/F and hypertension, data are presented as median (interquartile range).
Hypertension is defined as 24-hour ambulatory blood pressure >135/85 or on antihypertensive treatment.

Table 4.5 Univariate association analysis of urinary uromodulin in relation to rs13333226 polymorphism and response to high and low salt intake (GRECO Study).

	AA (n=40)	AG and GG (N=24)	p-value
M / F (n)	40 / 0	24 / 0	1.0
Age (years)	26 ± 8	23 ± 6	0.105
Body mass index (Kg/m^2)	23.4 ± 2.7	23.4 ± 2.1	1.0
Body surface area (m^2)	2.05 ± 0.14	2.03 ± 0.15	0.590
SBP LS (mm Hg)	120 ± 10	121 ± 10	0.670
DBP LS (mm Hg)	68 ± 9	70 ± 6	0.453
SBP HS (mm Hg)	123 ± 10	124 ± 10	0.805
DPB HS (mm Hg)	69 ± 8	70 ± 7	0.661
GFR LS ($mL/min/1.73m^2$)	109 ± 13	103 ± 14	0.127
GFR HS ($mL/min/1.73m^2$)	114 ± 14	116 ± 15	0.719
ERPF LS ($mL/min/1.73m^2$)	472 ± 74	449 ± 68	0.209
ERPF HS ($mL/min/1.73m^2$)	502 ± 90	489 ± 68	0.529
ECV LS ($L/1.73m^2$)	16.5 ±1.9	16.3 ± 1.6	0.657
ECV HS ($L/1.73m^2$)	17.2 ± 1.7	18.0 ± 1.9	0.093
FENa LS (%)	0.19 ± 0.18	0.22 ± 0.25	0.342
FENa HS (%)	0.99 ± 0.35	0.82 ± 0.31	**0.001**
PRA LS (nmol/L/h)	6.3 ± 3.7	6.6 ± 3.1	0.723
PRA HS (nmol/L/h)	2.5 ± 1.5	2.0 ± 0.9	0.155
UMOD LS median (IQR) (mg/L)	10.3 (6.9-15.6)	9.0 (6.3-14.2)	**0.002**
UMOD HS median (IQR) (mg/L)	11.9 (7.5-27.9)	12.2 (7.2-21.3)	0.513

BSA = Body Surface Area. LS = Low salt diet. HS = High salt diet. GFR = Glomerular Filtration Rate. ERPF = Effective Renal Plasma Flow. ECV = Extracellular Volume. FENa = Fractional Excretion of Sodium. PRA = Plasma Renin Activity. UMOD = Uromodulin.
With the exception of FENa and UMOD, data are presented as mean (standard deviation).

4.8 Discussion

The initial observation of a significant association between hypertension and rs13333226 genotype was identified as the main result of interest and was successfully replicated in a two stage validation analysis. As anticipated, the meta-analysed effect size was smaller than that of the discovery sample, in which the odds ratio (0.67) was inflated. This is explained in large part by the selection of cases and controls from the extreme ends of the blood pressure distribution in the discovery cohort (see Figure 2.1). Furthermore, the winner's curse may be a contributory factor. The finding remained significant after adjustment for age, age^2, sex, and BMI, and when the discovery cohort was excluded from analysis. Moreover, when the analysis was repeated with adjustment for eGFR, in the seven cohorts for which this was available, the effect was marginally strengthened (compared with initial adjustment of the seven cohort meta-analysis). In three separate populations the minor G allele of rs13333226 (associated with a lower risk of hypertension) was associated with lower urinary uromodulin excretion, although in the BRIGHT sample this did not reach statistical significance. In GRECO participants following a high salt diet, genotype was not associated with urinary uromodulin excretion suggesting a gene-environment interaction. Our combined results suggest that *UMOD* may have a role in regulating blood pressure, possibly through an effect on sodium homeostasis.

The only study to have examined *UMOD* in relation to hypertension was a candidate gene association study conducted by a Japanese group [214]. Iwai et al analysed 161 SNPs of 10 candidate genes and their association with hypertension, defined as SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg or current use of antihypertensive medication. Participants were from the Suita Study, a longitudinal general population study of a random sample of people aged 30-79 years in the Japanese city Suita, which began recruitment in 1989 [215]. Blood pressure measurement was the mean of two seated readings following 10 minutes at rest. Candidate genes were selected based on evidence of physiological function, being involved with blood pressure homeostasis, kidney function, leptin and insulin signalling, antioxidant effects, or familial juvenile stroke. Adjustment was made for age and BMI. In an analysis of 1,509 individuals with hypertension and 2,119 controls the minor allele of rs6497476, located in the 5' region of *UMOD* (-744 bp from *UMOD* transcriptional start point), was associated with lower risk of hypertension with $P = 0.039$. Following Bonferroni correction for multiple testing, however, the association was no longer significant. Moreover, the polymorphism was not associated with blood pressure measured in the recumbent position or with uric acid levels. In the Japanese HapMap population

rs6497476 is in LD with rs13333226 ($r^2 = 0.91$) and shows the same directionality of effect.

A recent GWAS of CKD conducted by Köttgen and colleagues identified rs12917707, located -3653bp upstream from the *UMOD* transcription start site, as the most strongly associated polymorphism ($P = 2.85 \times 10^{-9}$ in discovery sample only; $P = 5 \times 10^{-16}$ across discovery and replication samples combined) [202]. The minor T allele was associated with a 20% reduction in CKD risk. The association was consistent after adjustment for CKD risk factors including SBP, blood pressure lowering medication, and diabetes, and analysis stratified by age, sex, hypertension, and diabetes produced similar odds ratios. In HapMap CEU samples rs12917707 is perfectly correlated with rs13333226 ($r^2 = 1$). In the Köttgen study, rs13333226 was one of seven additional SNPs in or upstream of *UMOD* associated with CKD at the level of genome-wide significance and in high LD ($r^2 > 0.8$) with rs12917707. All eight SNPs were also associated with eGFR.

Specifically, eGFR describes the flow rate of filtered fluid through the kidney, hence higher values indicate better function [179]. The extent of chronic kidney disease, if present, is graded by how much eGFR is reduced [216], with values of less than 60ml per minute per 1.73 m^2 of body-surface area indicative of early stage disease. Our results show that the association between hypertension and rs13333226 is independent of kidney function as defined by eGFR. This is a critical point because rs13333226 is located close to the *UMOD* gene, and mutations of *UMOD* have been associated with chronic renal failure [217, 218]. For example, Hart and colleagues studied large, multigenerational families with familial juvenile hyperuricaemic nephropathy (FJHN) and medullary cystic kidney disease 2 (MCKD2) [219]. Both are autosomal dominant renal diseases and exhibit similar phenotypic characteristics. Primary clinical features vary in occurrence and severity but include juvenile onset of hyperuricaemia, gout, and progressive renal failure. Hypertension is a potential long term complication. Through linkage and haplotype analysis, Hart et al identified four novel *UMOD* mutations segregating with FJHN and MCKD2. This suggests that the protein encoded has a role in renal urate handling and possibly renal development. In total, 37 distinct mutations in *UMOD* have been associated with FJHN and/or MCKD2; there is a clustering in exons 4 and 5[218, 220, 221]. Of the 37 mutations, 33 are single amino acid changes of which 23 modify cysteine and 10 charged residuals.

UMOD encodes uromodulin, also known as the Tamm-Horsfall protein because it was identified and characterised by Tamm and Horsfall [222]. Uromodulin is a glycosylphosphatidylinositol (GPI) anchored glycoprotein. It has long been recognised as the most abundant tubular protein in urine; however its function is unclear. It is expressed predominantly in the thick ascending limb of the loop of Henle (TAL) with negligible expression elsewhere [223, 224]. Studies have shown urinary uromodulin levels to be decreased in older healthy individuals [225] and in patients with reduced renal function [226], although the latter study had a sample size of just 42. The evidence for effects of blood pressure on uromodulin excretion is inconsistent [227, 228]. The TAL is also the site where mutations of tubular transporters have resulted in rare Mendelian high or low blood pressure syndromes [229].

A study conducted by Dahan and colleagues identified further mutations in the *UMOD* gene in families with FJHN [230]. Furthermore, uromodulin excretion was measured in urine and kidney biopsies of patients with *UMOD* mutations. In these individuals urinary uromodulin excretion was decreased, and expression increased in a subset of tubules, in comparison with control participants and patients with renal failure due to other causes, including FJHN patients without a *UMOD* mutation. Similarly, Bleyer et al observed significantly lower urinary uromodulin excretion in individuals with *UMOD* deletion mutations compared with unaffected relatives and spouses [231]. This was independent of sex, age, and glomerular filtration rate.

Studies of uromodulin knockout mice have provided some clues as to its function. The first uromodulin deficient mice were created by Mo et al who deleted the first four exons and a 650bp proximal promoter region of *UMOD* [232], with the effect of rendering it non-functional. Compared with wild-type mice the knockout animals were predisposed to bladder infections, but no major effects on embryonic development or the histology of the kidney were observed. In a similar study undertaken concurrently, Bates et al targeted disruption of exon 2 in a separate knockout model and independently confirmed the increased susceptibility to urinary tract infections (UTI) in uromodulin-null mice [233]. Thus uromodulin is involved in host-defence against *Escherichia coli* (*E.coli*) adhesion to the urothelium, *E.coli* being the cause of 85% of UTI [233]. Mo and colleagues also used their knockout mice to show that uromodulin has a role in preventing the development of calcium oxalate crystal formation (kidney stones) [234]. Conversely, in humans with *UMOD*-associated FJHN or MCKD2 there is not an increase in prevalence of UTI or kidney stones [220]. This is probably because, whereas knockout mice excrete no urinary uromodulin,

patients with *UMOD* mutations retain a small amount of wild-type excretion. It appears that this is enough to protect against UTI or renal stone formation [220].

Lynn and Marshall demonstrated that patients with renal disease have significantly lower levels of uromodulin excretion than individuals with normal renal function (19 mg/24h and 39 mg/24h respectively) [235]. Excretion was yet lower in patients with polycystic kidney disease. This was independent of age, sex, and urine volume. Furthermore, excretion was neither influenced by degree of proteinuria, nor by variations in proteinuria with changes in disease activity or albumin infusions. It was positively correlated with creatinine clearance.

Lupus nephritis (LN) is defined as inflammation of the kidney caused by systemic lupus erythematosus. Tsai et al have shown that uromodulin excretion is lower in patients with active LN and tubulointerstitial inflammation (another complication of systemic lupus erythematosus) compared with those with inactive LN or normal individuals [236].

With the aim of testing the theory that uromodulin has a role in regulating renal salt and water excretion, Bachmann et al ran functional and morphological studies to compare knockout mice (the breed created by Bates and colleagues) and their wild-type siblings [237].The uromodulin-null mice had kidneys that were anatomically normal and there was no change in steady-state electrolyte concentrations. However, there was significant upregulation of major distal transporters, juxtaglomerular immunoreactive cyclooxygenase-2 (COX-2) and renin mRNA expression both were decreased, and creatinine clearance was 63% lower than in wild-type mice. Collectively these observations support the conclusion that uromodulin plays a part in renal function regulation, as does research conducted on rats with hyperthyroidism [238]. Ying et al examined male Sprague-Dawley rats on diets that contained either 0.3%, 1.0%, or 8.0% salt [239]. Higher salt intake led to sustained increases in relative steady-state mRNA and uromodulin levels in the kidney.

The finding of an association between hypertension and a genetic variant in *UMOD* is biologically plausible. An association between malignant hypertension and increased risk of renal disease was first recognised in the 19th century. Following initial observations, a large evidence base was established irrefutably linking the two conditions. However, up until the 1990s any relationship between milder forms of hypertension and renal problems was uncertain. A 1989 review of the clinical and epidemiological evidence by Whelton

and Klag [240] concluded that, although there was ample evidence of a causal relationship between severe hypertension and the occurrence of renal disease, there was insufficient comparable data to make such an assessment for mild to moderate hypertension. Rather they state that at the time there was a suggestive yet inconclusive relationship. Subsequently, they along with others conducted a study of 16 years follow-up of prospective data from the Multiple Risk Factor Intervention Trial [241]. They analysed blood pressure at recruitment and later incidence of end-stage renal disease in 332,544 males, and observed a strong, graded relationship between the two. Compared with normotensive participants (SBP <120 mmHg and DBP <80 mmHg), the relative risk of end-stage renal disease was 3.1 for mild hypertension (SBP 140-159 mmHg and DBP 90-99 mmHg), 6.0 for moderate (SBP 160-179 mmHg and DBP 100-109 mmHg), 11.2 for severe (SBP 180-209 mmHg and DBP 110-119 mmHg), and 22.1 for very severe hypertension (SBP ≥210 mmHg or DBP ≥120 mmHg) (all P<0.001). A limitation of the study was that baseline renal function was not assessed in all participants, therefore confounding by pre-existing disease could not be ruled out. However, a later study by Hsu et al, that was able to screen out baseline kidney disease, demonstrated a similar graded relationship between blood pressure and end-stage renal disease [242]. In a paper recently published in Hypertension risk of chronic kidney disease was shown to be increased even in prehypertension, defined as SBP ≥120 and <140 mmHg or DBP ≥80 and <90 mmHg [243]. Prevalence of chronic kidney disease was 17.3% among prehypertensive participants, compared with 13.4% in those with normotensive blood pressure levels, 22.0% in undiagnosed hypertension, and 27.5% in those diagnosed with hypertension. Hypertension was more strongly related to albuminuria than eGFR. The prevalence of other risk factors for chronic kidney disease was similar across blood pressure categories. It is pertinent to mention that in the current study neither the discovery sample nor the majority of validation samples were phenotyped for renal function.

The meta-analysis was conducted using an inverse-variance fixed-effects model. Random-effects models are generally more conservative, thus require more data to achieve the same statistical power as fixed-effects models. Ioannidis and colleagues examined findings from three GWAS studies of type 2 diabetes where data were meta-analysed using fixed-effects models [119]. They repeated the analysis using random-effects and compared the results of the two model types. At all levels of heterogeneity the average ORs were similar for both models. But when heterogeneity was moderate to high, as estimated by I^2, the associated 95% confidence intervals were wider for random effects estimates and P-values no longer crossed the threshold for genome-wide significance.

Pereira and colleagues compared random- and fixed-effects models, under different experimental circumstances, via simulations of cumulative meta-analyses of GWAS signals [244]. The simulations allowed for genetic model misspecification and true/between-study heterogeneity. They found that random-effects models exhibit a high type I error rate when there are only a few datasets in the analysis. Furthermore, in the presence of heterogeneity in meta-analyses of ≤ 10 datasets of around 1,000 cases and 1,000 controls each, random-effects models did not increase power over a single study and could be less powerful. However, in fixed-effects models when there is true heterogeneity and there is no effect on average, the type I error rate increases considerably as the volume of data increases. Therefore, whereas fixed-effect models are preferable for initial screenings, random-effects models are preferable for subsequent generalisability of findings. For this reason the current study employed a fixed effects model. This decision was supported by the observed lack of heterogeneity.

The extent of study heterogeneity in the meta-analysis was summarised by the Q statistic and I^2 statistic, to determine whether the use of a fixed-effects model was appropriate. The Q statistic does not give a measure of the magnitude of true heterogeneity, only whether it is present. The I^2 statistic, on the other hand, does give a measure of the extent of true heterogeneity [245] and moreover is independent of the number of studies. A study by Huedo-Medina and colleagues that compared the Q statistic and I^2 statistic found them to have comparable Type I error rates and statistical power [158]. There is evidence, however, that when there are only a few studies the Q statistic is greatly underpowered [119]. This is not a concern in the current analysis.

Further functional studies are required to investigate the renal mechanisms by which *UMOD* may influence hypertension and renal sodium handling. The main limitations of the current functional studies are the use of three different populations and single time point renal and blood pressure measurements. To explore genotype-phenotype effects over prolonged periods repeated measurements are essential.

To summarise, rs13333226 on chromosome 16 was followed up for replication meta-analysis in a total of 14 independent cohorts. The combined sample size was 21,466 cases and 18,240 controls. The meta-analysed effect size was diminished relative to the discovery cohort, however, it was in the same direction and statistical significance was strengthened (OR = 0.87, $P = 3.67 \times 10^{-11}$). In analysis adjusted for age, age^2, sex, BMI and eGFR, and when the discovery cohort was excluded, the association remained

significant. Thus it may be concluded that the finding is robust. rs13333226 is located close to the transcription start site of the *UMOD* gene, mutations in which have previously been linked to chronic renal failure. Our findings indicate that *UMOD* is independently associated with hypertension. Following a high salt diet, in healthy males the G-allele was associated with a larger increase in measured GFR and hence filtered sodium load, and a smaller increase in tubular sodium excretion. Together these adaptations achieve sodium balance. Additionally, in G-allele carriers the rise in extracellular fluid volume is greater after increased salt intake. The appendix provides information on the genetic locations and possible biological significance of other SNPs associated with hypertension status at the borderline level of genome-wide significance in the combined discovery and MONICA/PAMELA sample.

5 General discussion

We have demonstrated that using an extreme phenotyping method to ascertain cases and controls can lead to successful identification of genome wide association signals for hypertension. High fidelity phenotyping, comparing the top and bottom end of the blood pressure distribution, reduced misclassification of controls and also inflated the effect sizes allowing for a more efficient design with smaller sample sizes. We have identified and validated a marker significantly associated with hypertension from a discovery sample of just 3,320 individuals, far fewer than other successful GWAS of hypertension e.g. Global BPgen [58] and CHARGE [59]. Furthermore, the strategy allows the sampling of participants from a single population, rather than combining multiple cohorts from heterogeneous populations, which reduces confounding by stratification. This confers practical benefits and reduces costs. The estimated odds ratios are likely to be inflated compared with the true odds ratios for hypertension as typically defined. This is reflected in the smaller effect sizes observed in the validation cohorts that had a somewhat relaxed case/control definition. Importantly, the meta-analysed overall effect size in the current study is comparable to the effect sizes of the robust association signals for blood pressure identified by Global BPgen and CHARGE.

Although the current study validated the association between rs13333226 and hypertension status in cohorts from the Global BPgen study [58], no variants in or near the *UMOD* gene were identified as either blood pressure or hypertension associated in the original Global BPgen or CHARGE analyses [59]. The most likely explanation for this discrepancy is the current study's strict case control criteria for very high and very low blood pressure. In the discovery sample blood pressure cut-offs were SBP $\geq$ 160 mmHg or DBP $\geq$ 100 mmHg for cases, and SBP $\leq$ 120 mmHg and DBP $\leq$ 80 mmHg for controls. This is contrasted with hypertension defined as SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg in Global BPgen and CHARGE, and normotension defined as SBP $\leq$ 120 mmHg and DBP $\leq$ 85 mmHg in Global BPgen (CHARGE normotension criteria not provided). In addition controls in this study had no evidence of CVD in ten years of follow-up, information either not available or not taken into account by the other two studies. Furthermore, the age restrictions of <60 and $\geq$50, respectively, used in the current study reduced the confounding effect of age and increased the power to detect genetic effects. The criteria for participants from Global BPgen samples in the current validation analysis, while somewhat relaxed, were still comparatively stringent in that the original Global BPgen study did not employ age cut-offs. Moreover, in the current discovery sample cases were recruited while off treatment following a washout period, whereas both the Global BPgen and CHARGE studies dealt with the confounding effects of blood pressure lowering medication with the use of blanket

corrections (15 mmHg to recorded SBP and 10 mmHg to recorded DBP in Global BPgen; 10 mmHg to recorded SBP and 5 mmHg to recorded DBP in CHARGE). Once all of the differences in phenotypic definitions between the three studies are taken into account, it is not surprising that their principal findings were different.

The Iwai et al study of *UMOD* as a candidate gene for hypertension offers some evidence of association in a different ethnic group, namely Japanese individuals [214]. While rs6497476, located in the 5' region of *UMOD* (-744 bp from *UMOD* transcriptional start site) was found in initial analysis to be significantly associated with hypertension risk (P = 0.039), significance was lost following Bonferroni multiple testing correction. Sample size, however, was comparatively small at 1,509 cases and 2,119 controls. Therefore, it is possible that lack of significance was due to low statistical power.

The association of rs133332226 and other *UMOD* variants with CKD and eGFR in the GWAS conducted by Köttgen et al [202] supports our finding. Hypertension and renal disease are inextricably linked and each has been shown to increase the risk of the other. In the Köttgen study findings were consistent in models adjusted for and stratified by hypertension. Moreover, the association between rs13333226 genotype and hypertension in the current study remained after adjustment for eGFR, suggesting that it is independent of renal function. However, this observation is limited because eGFR was calculated at a single point in time, as was blood pressure, and we did not have access to measurements over the life course. Further work is required in this area before any firm conclusions can be made.

In the current study only the top hit was followed up in validation analyses. However, there was another SNP in a different region which also attained genome-wide significance; rs13353058 in *IKZF5* on chromosome 10q26. It would be interesting to follow this up in a larger sample as well as any borderline SNPs of potential biological importance. Of the borderline associations the most appealing is probably rs1893469, located close to the *NEDD4L* gene on chromosome 18q21. As discussed in the appendix, the encoded protein is a determinant of sodium reabsorption in the distal nephron [246, 247], and variants in *NEDD4L* have been associated with salt sensitivity [248, 249] and essential hypertension [250-253]. Therefore, further investigation of this locus may provide some insight into the mechanisms linking sodium reabsorption and blood pressure variability. Other possibilities are: *THBS2*, previously associated with risk of thoracic aortic aneurysm in

hypertensive patients [254] and premature MI [255]; and *COL4A1*, associated with arterial stiffness [256] and heredity angiopathy with nephropathy, aneurysms, and muscle cramps [257].

Of course the ultimate goals of GWAS analysis are to improve patient risk prediction, diagnosis, prevention, and finally to develop treatments that improve quality of life and survival. However, current evidence from published GWAS studies indicates that the common variants identified using this method not only explain very little of the population variation of disease traits and heritability, but also have very poor predictive potential on an individual basis [258]. Some studies have combined the risk estimates of several SNPs into a single genotypic risk score that shows association with disease even after adjustment for traditional risk factors. However, odds ratios or similar measures are not sufficient to assess the utility of genetic variants for individual risk prediction [259, 260], and the existing evidence suggests that when more complex measures are employed such genetic information has little or no effect on clinical risk prediction [261, 262]. For accurate prediction using a quantitative risk factor, the population distributions for cases and controls must be sufficiently well separated to allow selection of a cut-off value that discriminates between the groups with adequate sensitivity and specificity. Ware argues that for this to be achieved an odds ratio, cumulative or otherwise, of more than 200 is necessary [263]. This is unrealistic at the moment because the disease-associated variants identified are too few and of too small effect [258]. The reported association between hypertension status and rs13333226 in this study will contribute little to individual risk prediction. Moreover, the case-control definitions applied do not equate to those used in clinical practice. Rather it is hoped that the finding will provide insight into the mechanisms underlying the development of hypertension.

Evans and colleagues [264] used the original genome-wide WTCCC data [55] to investigate whether genome-wide data could improve diagnostic accuracy over and above information on loci known to affect risk. The addition of genome-wide data improved discriminative accuracy most for coronary heart disease, type II diabetes, and bipolar disorder. The improvement was of such a small magnitude that the authors considered it unlikely to be of diagnostic or predictive utility; however it suggests that there remain undiscovered variants that are associated with disease. Perhaps surprisingly, when the additional SNPs to be included were determined using a liberal significance threshold any improvement in predictive ability was greater than when more stringent thresholds were employed. This is consistent with there being many unknown loci of small effect. Improvements were far smaller for rheumatoid arthritis and type I diabetes, and for Crohn's disease the inclusion

of whole genome data actually decreased discrimination of case-control status. This may be because more loci associated with these diseases have been discovered; hence genome-wide data has little or no added value. It was not possible to evaluate the use of whole genome data for hypertension case determination in the same manner because at the time no known confirmed loci had been associated with common essential hypertension. Nevertheless, genome-wide data alone provided a median area under the curve (AUC) of 0.61 for hypertension, considerably better than chance defined as AUC = 0.5. As discussed in chapter 1 page 46, the possibility of misclassification bias in the WTCCC samples stands.

The major breakthrough with GWAS has been the identification of robust signals for common diseases and traits and novel pathways of disease. Despite this success, the large amount of unexplained variation that remains for many traits, coupled with high financial cost, means that researchers are looking at other scientific strategies. Several explanations for the missing heritability have been proposed. These include rarer variants in novel pathways that are undetectable through traditional GWAS case-control study design; many more variants of smaller effect size that haven't been discovered; structural variants, such as CNVs, that are poorly captured by existing arrays; insufficient power to detect gene-gene interactions; and shared environment unaccounted for in family studies [265]. Moreover, in GWAS of common variants, imperfect LD between tag SNPs and causal SNPs may have caused underestimation of effect sizes. As noted in a review by Manolio et al, explaining missing heritability is not really an end in itself [265]. Rather it is a step on the path to achieving the ultimate goals of all genetic research into complex disease; improved prevention, diagnosis, and treatment. Figure 5.1, reproduced from the Manolio review, plots the effect size of variants against their allele frequency. It is thought that most of the unexplained heritability for many complex traits is likely to lie in the middle region of the graph, i.e. low-frequency variants with intermediate effect. The primary method in the search for such variants is next-generation deep sequencing.

Our study, using extreme case/control definitions, was able to identify a novel signal in *UMOD* that could potentially identify a novel pathway for blood pressure regulation. We have identified a common variant with an effect-size similar to other GWAS signals, suggesting that there is potential for identifying common variants using different study designs. Our study design has the potential to identify rare and intermediate frequency variants with MAF<5%. However, our current strategy is blind to rare variants and we are now pursuing an exome sequencing (described below) experiment which is underway in

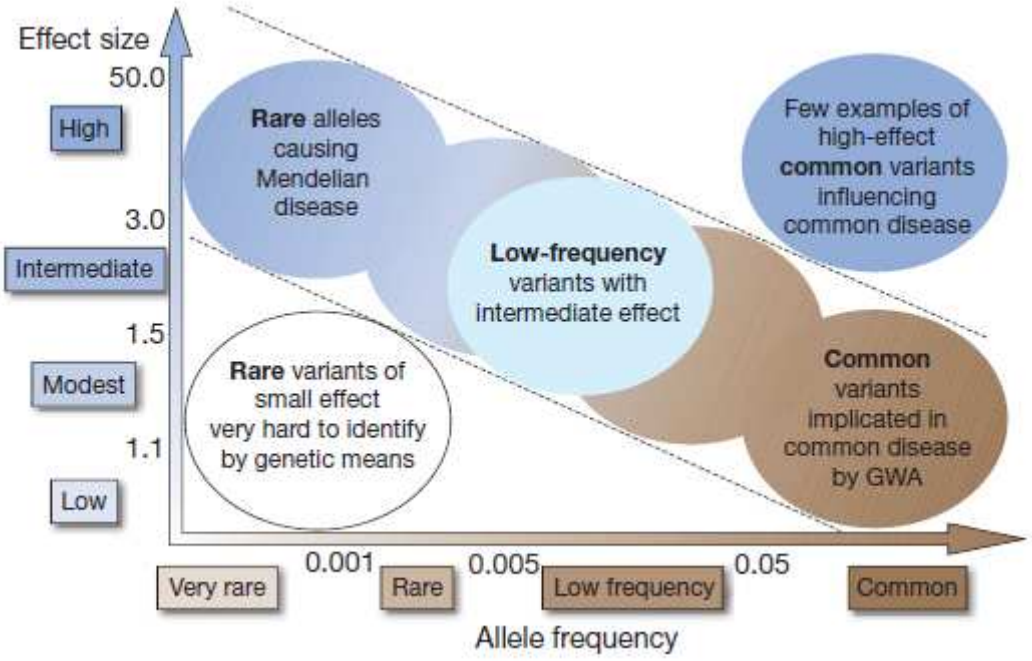


Figure 5.1 Feasibility of identifying genetic variants by strength of genetic effect (odds ratio) and risk allele frequency. Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines (reproduced from [265]).

collaboration with Richard Lifton at Yale to identify rare variants that influence blood pressure.

As noted above, next-generation deep sequencing technology is available which will increase the chances of rare variant detection [266]. Possible sequencing strategies to detect rare variants include focusing on genomic regions where disease has been repeatedly and strongly associated with common variants, sequencing a larger portion of the genome in individuals with extreme phenotypes, and studying populations of recent African ancestry [265]. The latter may be more effective than studying European populations, where most research has been focused so far, because Africans have greater genetic variation. Diversifying GWAS analysis beyond European populations could also contribute greatly to discovering the common genetic determinants of multifactorial disease. But certain methodological issues need to be tackled carefully, including marker ascertainment, tag SNP portability, imputation, replication, and admixture [267]. Some of these issues are of particular concern in sub-Saharan African populations of low LD, due to their dissimilarity from the reference populations employed in the design of analysis tools.

Whole genome resequencing is currently prohibitively expensive, therefore as a practical alternative whole exome arrays were developed. These cover only exomic, i.e. protein coding, regions which account for 1% of the genome yet harbour 85% of mutations with large effects on disease predisposition [268], hence efficiency is increased. A limitation is that they do not measure most structural variation [269]. Choi et al demonstrated the use of whole exome capture to make a clinical diagnosis of congenital chloride diarrhoea in a patient with a rare causal mutation [268]. For complex traits it is likely that whole exome sequencing will be performed in a subset of participants and then followed up with validation sequencing of regions of interest in a larger sample. To maximise the chances of uncovering rare disease-predisposing variants whilst minimising financial cost, sequencing may be best performed in affected individuals within families at the extremes of trait distribution [269]. Following analysis the greatest challenge will be the identification of causal variants amongst the large number of novel variants uncovered, and biological plausibility may again come to the fore [265]. Functional information will help narrow the regions and variants of interest, as will association and linkage evidence for candidate variants. This, in combination with co-segregation analysis of family data, will hopefully pinpoint the causal variants. Statistical power will be greater for recognisable variants, i.e. those that have a clear function, such as ones that delete some or all of a gene [269].

One strategy that may aid the identification of variants with MAF <5% is the study of extreme phenotypes which are likely to be enriched for rare variants; a similar idea to the currently reported study design yet taken to another degree of distance from the normal distribution of trait values in the general population. Two recent articles published back to back in Nature demonstrated the utility of this method. The studies of Bochukova et al [270] and Walters et al [271] began with genome-wide association analysis of CNVs in patients with severe early-onset obesity, with and without developmental delay. Both identified a rare deletion on chromosome 16p11.2 which had been previously associated with autism and mental retardation. The Bochukova study followed the finding up with targeted analysis in 16,053 individuals from eight cohorts. Deletions were present in 0.7% of morbidly obese cases but were absent in healthy non-obese controls. This methodological approach demonstrates that rare variants of large effect identified in extreme cases may also have a role in milder common forms of disease.

Family studies may have a better chance of detecting rare variants, and any association between them and common diseases, than population studies. If a disease predisposing variant is present in a proband individual then it should occur at higher frequency in affected relatives than unaffecteds. Large family studies that are currently collecting data, such as the Generation Scotland: Scottish Family Health Study (GS: SFHS) [102] which aims to recruit a sample of 50,000 individuals, will facilitate such analysis. Two recent studies have employed next-generation sequencing in small family samples with Mendelian disease in order to identify causal variants [272, 273]; it is hoped that similar methods will be applied to common complex diseases once costs have fallen enough to allow sequencing of large samples. Roach and colleagues analysed the whole-genome sequences of two siblings both with two recessive disorders, Miller syndrome and primary ciliary dyskinesia, and their unaffected parents [272]. This allowed the identification of very rare variants and the narrowing of possible candidate genes. Lupski et al sequenced the genome of a single patient from a family with a recessive form of Charcot-Marie-Tooth disease [273]. Once potential functional variants had been identified in the proband these were genotyped in affected family members. The process identified two mutations in *SH3TC2* (SH3 domain and tetratricopeptide repeats 2) segregating independently with separate subclinical phenotypes, demonstrating the utility of whole-genome sequencing in diagnostics. Lupski et al note that, over the 6-month course of their study, “the sequence yield increased by a factor of three, with no appreciable increase in expense”, highlighting the current rapid advancement of sequencing technology. A further benefit of family samples is that they permit the study of parent-of-origin effects, which may be important for common disease

phenotypes [274]. If these are not accounted for they could mask associations hence reducing the proportion of heritability that can be explained [265].

Our plan for whole exome sequencing to uncover rare variants influencing blood pressure combines selection of extremes of trait distribution with family data, the aim being to choose participants that enable optimum study efficiency. We hope that careful selection of individuals from families with very high (extreme hypertensive cases) or low blood pressure (hypercontrols) will maximise power and minimise sample size requirements.

In parallel to the aforementioned search for further blood pressure associated variants, we will follow up the current finding. The actual function of uromodulin is unknown; therefore our plans for future work include functional studies to clarify the mechanisms by which *UMOD* may influence renal sodium handling and the development of hypertension. To this end we hope to obtain well preserved human kidneys for the study of expression levels; this has not been possible to date. Also, we currently await the arrival of knock-out mice from the United States [233, 237] and would like to obtain funding for knock-in mice that are a model of FJHN, bred in Rajesh Thakker's group in Oxford. Finally, the reported reduced urinary uromodulin excretion in the presence of the minor G allele of rs13333226 should be confirmed in larger cohorts.

Intensively genotyped samples are becoming available from the 1000 Genomes Project, an open resource catalogue of human genetic variation [275]. The project is run by an international consortium and aims to describe over 90% of genetic variation down to 1% MAF. To date the genomes of more than 1000 individuals have been sequenced (with planned expansion to 2000 individuals in 2010) and around 11 million novel variants identified. The resource, and other expanded reference panels of genomic variation, will improve GWAS coverage thus enabling the study of low frequency variants [265] and aiding fine mapping of regions of interest.

Analysis of CNVs has been proposed as a possibility for exploring some of the missing heritability of common diseases. CNVs are structural genomic variations that result from duplication or deletion of genomic segments. This approach has had substantial success in rare genomic disorders, with the identification of several causal CNVs [276-278]. Thus far, a few rare CNVs [279-281] and common CNVs [186, 282, 283] have been associated with common diseases. However, a recent WTCCC study has raised some doubt as to the additional information conferred over and above SNP analysis [284]. It analysed 3,432 common (MAF

>5%) CNVs in a sample of 16,000 cases of eight common diseases (those included in the original WTCCC study plus breast cancer) and 3,000 shared controls. CNVs were found to be significantly associated with disease at three loci: *IRGM* (immunity-related GTPase family, M) for Crohn's disease; *HLA* for Crohn's disease, rheumatoid arthritis, and type 1 diabetes; and *TSPAN8* (tetraspanin 8) for type 2 diabetes. But all were well tagged by SNPs and indeed had been previously identified in SNP association studies. Hence the CNV analysis did not add anything to the state of knowledge, and currently it does not look as though the study of CNVs will explain much missing heritability. Nevertheless, compared with SNP assays, CNV assays are at an early stage of development. With improved accuracy, reliability and quality control CNV analysis of common diseases may prove more beneficial in the future. *De novo* CNVs do not contribute to heritability but may explain some of the trait variation currently attributed to the environment [265]. There is also the possibility of studying other structural variations such as inversions or translocations (which are copy neutral), new sequence insertions, microsatellite repeat expansions, and complex rearrangements. There is evidence of effects of all of these in rare Mendelian conditions, but for complex traits little data exists [265].

It could be that the association signals for common variants identified by GWAS are markers for rarer causal variants of large effect [269]. This possibility was recently examined using computer simulations by Dickson et al [285]. They assumed that rare (defined loosely as less common than those routinely studied in GWAS) variants were the only contributors to disease risk, and found that in such a scenario the rare causal variants can create associations of genome-wide significance that are picked up by more common variants megabases (Mb) away. Moreover, the real effect size can be several-fold stronger than that attributed to the common variant. These associations are a special case of indirect association which Dickson et al termed 'synthetic associations'. They propose that the rare variant involved occurs, randomly, more often with one of the alleles of the common variant than the other allele. Currently the proportion of GWAS signals that may be due to synthetic associations is unknown [269], but their likelihood has implications for the interpretation of GWAS results and their follow-up. Thus far, when hits have been located a long way from the nearest gene it has often been concluded that a regulatory variant is the cause. However, since synthetic associations can be due to rare variants that are Mb away, the chance of their existence means that the area examined in follow-up studies should be larger than that typically studied. Thus deep sequencing of the region surrounding a hit should extend beyond the LD block of common variants that it is contained within. Dickson et al recommend sequencing an area at least 4 Mb and ideally

10 Mb around the discovery signal [285]. Although there is little empirical evidence for synthetic associations, there are some examples of rare variants in the same regions as GWAS signals that influence disease risk [185, 286, 287]. The strongest supportive evidence comes from Fellay and colleagues [288], working in the same research centre as Dickson.

As mentioned above, there is some evidence that common minor variants in the causal genes for monogenic forms of hypertension and hypotension may also have a causal role in blood pressure regulation, though this has not been replicated. Moreover, similar associations have been found with rare variants. Ji and colleagues screened participants of the Framingham Heart Study (FHS) for variation in three known candidate genes, *SLC12A3* (solute carrier family 12 (sodium/chloride transporters), member 3), *SLC12A1* (solute carrier family 12 (sodium/potassium/chloride transporters), member 1), and *KCNJI*, in which homozygous loss-of-function mutations are causally associated with monogenic forms of hypotension (namely Bartter's syndrome and Gitelman's syndrome) [289]. The FHS began in 1948 as a prospective general population study with the aim of identifying CVD risk factors. Participants, and later their offspring, grandchildren and spouses, have been followed up since. Risk scores based on its results are in widespread use (http://www.framinghamheartstudy.org/risk/index.html). Resequencing by Ji et al led to the detection of 30 different heterozygous rare (MAF<1%) mutations in 49 individuals. Mutation carriers had mean long-term SBP 6.3 mmHg lower, and DBP 3.4 mmHg lower, than the overall cohort means. Furthermore, the likelihood of developing hypertension by age 60 was 59% lower in mutation carriers compared with noncarriers. Despite the health benefit conferred by these heterozygous mutations, they remain rare due to strong purifying selection caused by the adverse effects of the homozygous states. These along with other findings [94] support the assertion that heterozygosity of rare variants explains some of the population variability of heritable traits. It could be that variants in *UMOD*, such as that identified in the current study, may act to lower blood pressure in a similar manner, albeit to a lesser degree.

The study of gene transcript abundance, which is directly modified by polymorphisms in regulatory elements, may help to elucidate the function of loci underlying complex disease traits [290]. Recent studies have mapped transcript abundance as a quantitative trait, termed expression quantitative trait loci (eQTLs). These are identified through the simultaneous assay of genome-wide association data and global gene expression data. First of all disease-associated variants are ascertained via GWAS in the usual way. Then genome-wide eQTL mapping data are studied for evidence that the same markers are also

associated with quantitative transcript levels (termed eSNPs) of a gene or genes. In eQTL mapping the expression of thousands of genes in a target cell or tissue is measured simultaneously using microarray technology. The data are treated in the same way as other quantitative trait phenotypes, such as blood pressure or lipids, and the same statistical methods used as for QTLs [290]. Many human eQTLs are highly heritable [291-293], and only some of this heritability is accounted for by SNPs. Transcription is also affected by CNVs, deletion-insertion polymorphisms, short tandem repeats, and single amino acid repeats [294]. VarySysDB is a database of information on all published polymorphisms that affect transcription [294].

The simultaneous association of specific markers with both disease and eQTLs confers more power than a simple comparison of gene expression between cases and controls [290]. Some GWAS have already taken the combined approach and have demonstrated its utility in identifying candidate genes [295, 296]. eQTL data have been used as supportive evidence for a candidate gene proposed because of biological plausibility or location, to select one gene from a choice of candidates, and to identify a different gene [297]. The value of eQTLs in the study of the biology underlying complex traits is supported by work recently conducted by Nicolae et al [298]. In lymphoblastoid cell lines from HapMap samples they showed that trait-associated SNPs are significantly more likely to be eQTLs than MAF-matched SNPs chosen from high-throughput GWAS platforms.

The main limitations of eQTL methodology arise from issues in the use of microarrays, including bias from variation in sample preparation and technical variation [299]. However, technology is improving and ultra-high-throughput sequencing systems can overcome many of these problems [300]. There is little known about epigenetic, environmental, and parent-of-origin effects in eQTLs; therefore these areas need investigation. Moreover, future studies should take transcript stability into account as well as how expression levels change at different stages of development [290, 301]. A systems biology approach will help to clarify the function and wider context of single-gene discoveries. To this end, several methods to construct data networks have been proposed [302]. Another limiting factor is unavailability of appropriate tissue samples. But recent achievements in the field of eQTL analysis have prompted the National Institutes of Health (NIH) Genotype-Tissue Expression (GTEx) [303] project, currently a 2-year pilot project to test the feasibility of collecting high-quality RNA and DNA from multiple tissues from ~160 donors. If successful the project will be expanded to around 1000 donors, and will serve as a resource to study human gene expression and regulation.

Perhaps the most challenging area of work in the future will be the study of gene-environment interactions. As already discussed in chapter 1 pages 51-53, this requires extremely large sample sizes [100] and accurate measurement of environmental variables can be difficult. Aside from this, and despite the apparent likelihood of gene-environmental effects in complex disease coupled with some supporting evidence [304], in reality few have been demonstrated [305]. It would be unwise to make assumptions regarding their importance in disease aetiology without further evidence. Gene-gene interaction detection, although still requiring relatively large sample sizes, should be less of a challenge. A pathway analysis approach may improve success rates, and has been demonstrated as effective in linking the SNPs most strongly associated with hypertension in the WTCCC data [306].

Our finding of a variant in the Uromodulin gene with a protective effect, associated with a reduced risk of hypertension, is exciting on several levels. In methodological terms we have effectively demonstrated an alternative phenotyping strategy that enabled the detection of a locus not previously linked to blood pressure or hypertension in large whole-genome studies [58, 59]. The multiple strands of evidence linking *UMOD* to renal disease and kidney function [202, 217-219, 230], which often co-occur and correlate, respectively, with hypertension and blood pressure, lend biological significance to this observation. Furthermore, uromodulin is predominantly expressed in TAL [223, 224] where physiologically crucial mechanisms of sodium handling are located, suggesting that alterations of these mechanisms may underlie the reduced hypertension risk in G allele carriers. In conclusion, we believe that the newly discovered *UMOD* locus for hypertension has the potential to provide unique insights into the mechanisms of high blood pressure, and identify novel drugable targets.

Appendix

This appendix summarises the top 87 hits, as defined by $P \leq 5.6 \times 10^{-4}$ in the discovery sample. These SNPs underwent validation meta-analysis of the discovery sample and combined MONICA/PAMELA sample. Included in the analysis were 2,515 cases and 2,445 controls. A summary of the results of the discovery, validation, and combined analyses is presented in Table A1. In the combined analysis three SNPs, shown in red bold font, crossed a p-value threshold of 5×10^{-7}: rs13333226, $P = 3.86 \times 10^{-7}$; rs4293393, $P = 3.30 \times 10^{-7}$; and rs13353058, $P = 4.78 \times 10^{-7}$. Two of them, rs13333226 and rs4293393, are in high LD with r^2=0.996. Unlike rs13333226, the other 86 SNPs were not taken forward for further replication but may be of future interest. Individual SNPs were annotated using WGAViewer and the results investigated via the OMIM, Entrez Gene, GeneCards, HuGE Navigator Genopedia, and dbGaP online databases. The results are presented by gene, rather than SNP or chromosome, to highlight the potential functional significance of the observed hypertension-related genetic loci. Regional plots were created using LocusZoom, and are presented for unique regions. Several borderline significant SNPs are located in or near genes for which there is no information available at present, therefore they are not included in this appendix.

IKAROS family zinc finger 5 (IKZF5)

The third SNP associated with hypertension at the genome-wide level of significance was rs13353058, located in the three prime untranslated (3' UTR) region of *IKZF5* on chromosome 10q26. Figure A1 is an association plot of the genomic region around rs13353058. The *IKZF5* gene is protein coding and belongs to the Ikaros family of transcription factors, which are expressed in lymphocytes and implicated in control of lymphoid development. A commonly used synonym is *Pegasus*. It is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

Cytokine-dependent hematopoietic cell linker (CLNK)

Three intergenic SNPs on chromosome 4 were closest to the *CLNK* gene, at a distance of between 280 and 290kb. Two, rs10009111 (Figure A2) and rs10011697, were in LD (r^2 = 1.00) but were not in LD with the third rs10516217. *CLNK* is protein coding and a member of the SLP76 family of adaptors. It plays a role in the regulation of immunoreceptor signalling [307]. However, in a mouse model without *CLNK* mast cell, T cell, and NK cell functions were normal [308]. This suggests that its presence is not essential

Table A1 Initial replication analysis of the top 87 SNPs. Results presented are the discovery sample, MONICA/PAMELA replication sample, and combined analysis using inverse-variance weighted fixed-effects meta-analysis.

				DISCOVERY				MONICA/PAMELA				COMBINED ANALYSIS			
CHR	SNP	BP	A1	N	OR	95%CI	P	N	OR	95%CI	P	P-FIXED	OR-FIXED	Q	I^2
1	rs1399291	97349510	T	3315	1.26	1.13-1.41	3.12E-05	1611	1.08	0.94-1.25	2.71E-01	2.55E-05	1.26	0.45	0
1	rs10857978	1.13E+08	T	3319	0.76	0.67-0.86	3.33E-05	1616	0.99	0.83-1.18	8.93E-01	3.49E-05	0.76	0.63	0
2	rs2192615	48975598	G	3319	0.8	0.71-0.89	7.77E-05	1612	0.96	0.84-1.11	5.88E-01	8.09E-05	0.8	0.63	0
2	rs12611661	1.05E+08	C	3314	0.76	0.66-0.86	2.61E-05	1618	1.08	0.92-1.26	3.41E-01	3.33E-05	0.76	0.26	21.95
2	rs9636284	1.62E+08	T	3319	0.63	0.52-0.78	1.48E-05	1615	1.15	0.93-1.43	2.08E-01	2.61E-05	0.64	0.14	54.82
2	rs2084543	1.62E+08	A	3320	0.6	0.48-0.74	3.36E-06	1616	1.18	0.94-1.49	1.54E-01	6.47E-06	0.61	0.12	58.89
2	rs16846179	1.62E+08	G	3320	0.63	0.51-0.78	1.49E-05	1618	1.12	0.9-1.39	2.99E-01	2.14E-05	0.64	0.24	28.52
3	rs9853991	32441801	T	3316	0.71	0.61-0.82	8.98E-06	1615	1.01	0.84-1.22	8.87E-01	9.53E-06	0.71	0.6	0
3	rs3888882	32472578	T	3320	0.73	0.63-0.85	2.92E-05	1614	1.00	0.84-1.2	9.67E-01	2.98E-05	0.73	0.77	0
3	rs12636240	1.15E+08	G	3319	1.26	1.12-1.4	6.15E-05	1615	1.03	0.89-1.19	7.21E-01	6.20E-05	1.26	0.84	0
3	rs9881563	1.15E+08	C	3318	1.26	1.13-1.41	5.16E-05	1613	1.01	0.87-1.17	9.34E-01	5.00E-05	1.26	0.95	0
3	rs9865965	1.15E+08	T	3317	1.26	1.13-1.41	4.28E-05	1615	1.01	0.87-1.17	8.94E-01	4.25E-05	1.26	0.99	0
3	rs13061150	1.15E+08	A	3320	1.25	1.12-1.4	6.88E-05	1615	1.02	0.88-1.18	7.94E-01	6.90E-05	1.25	0.86	0
3	rs9828099	1.15E+08	C	3320	1.25	1.12-1.4	7.92E-05	1615	1.03	0.89-1.19	7.06E-01	7.59E-05	1.25	0.79	0
3	rs3811647	1.35E+08	A	3320	0.77	0.69-0.87	1.21E-05	1613	1.00	0.86-1.17	9.86E-01	1.21E-05	0.77	0.89	0
3	rs6794945	1.35E+08	T	3318	0.78	0.7-0.88	5.88E-05	1614	1.04	0.89-1.21	6.39E-01	6.09E-05	0.78	0.67	0
3	rs7635876	1.58E+08	T	3318	1.39	1.19-1.63	4.18E-05	1615	1.01	0.78-1.31	9.51E-01	4.12E-05	1.39	0.92	0
3	rs1842840	1.58E+08	T	3319	1.27	1.13-1.42	4.20E-05	1617	1.00	0.86-1.15	9.51E-01	4.16E-05	1.27	0.77	0
3	rs11715321	1.58E+08	C	3318	1.27	1.14-1.43	3.35E-05	1618	1.00	0.87-1.16	9.85E-01	3.37E-05	1.27	0.84	0
4	rs10009111	10580521	G	3318	0.76	0.68-0.85	1.35E-06	1616	0.90	0.78-1.03	1.37E-01	1.94E-06	0.77	0.19	42.15

CHR = chromosome. BP = location in base pairs. A1 = major allele. OR = odds ratio. CI = confidence interval. Q = Q statistic. I^2 = I^2 statistic.

Table A1 continued.

				DISCOVERY				MONICA/PAMELA				COMBINED ANALYSIS			
CHR	SNP	BP	A1	N	OR	95%CI	P	N	OR	95%CI	P	P-FIXED	OR-FIXED	Q	I^2
4	rs10011697	10580930	G	3319	0.76	0.68-0.85	1.50E-06	1616	0.90	0.78-1.03	1.37E-01	2.16E-06	0.77	0.19	42.04
4	rs10516217	10582359	A	3319	1.26	1.13-1.4	4.57E-05	1615	0.90	0.78-1.05	1.72E-01	3.78E-05	1.26	0.53	0
4	rs4487344	1.03E+08	G	3320	0.78	0.7-0.87	8.11E-06	1611	1.08	0.94-1.24	2.97E-01	1.23E-05	0.78	0.12	58.12
4	rs13124455	1.03E+08	A	3318	1.36	1.19-1.54	4.00E-06	1613	0.94	0.79-1.11	4.47E-01	3.47E-06	1.36	0.63	0
4	rs7669524	1.03E+08	A	3320	1.36	1.19-1.54	3.60E-06	1613	0.94	0.79-1.11	4.58E-01	3.15E-06	1.36	0.64	0
4	rs768290	1.03E+08	G	3320	1.34	1.18-1.53	7.55E-06	1612	0.93	0.79-1.11	4.43E-01	6.58E-06	1.35	0.64	0
4	rs12505043	1.03E+08	T	3319	0.78	0.68-0.89	1.79E-04	1614	1.05	0.89-1.23	5.76E-01	2.04E-04	0.78	0.4	0
4	rs4482766	1.03E+08	C	3313	1.29	1.14-1.45	2.63E-05	1613	1.02	0.88-1.2	7.74E-01	2.52E-05	1.29	0.69	0
5	rs106415	6845839	A	3320	0.8	0.71-0.89	5.96E-05	1614	0.91	0.8-1.05	1.98E-01	8.62E-05	0.8	0.16	48.95
5	rs172384	36839982	G	3320	0.71	0.61-0.82	4.27E-06	1616	0.80	0.66-0.95	1.21E-02	1.51E-04	0.76	0	92.72
5	rs292196	36949936	T	3320	0.7	0.6-0.81	4.21E-06	1614	0.82	0.68-0.98	2.59E-02	1.01E-04	0.75	0	91.75
5	rs16903459	37067173	G	3317	0.72	0.62-0.84	4.16E-05	1616	0.83	0.69-1	5.40E-02	4.36E-04	0.76	0	89.38
5	rs12658479	37242378	C	3313	0.76	0.66-0.87	8.61E-05	1614	0.89	0.75-1.06	1.82E-01	1.91E-04	0.77	0.06	72.4
5	rs2460498	76177535	A	3320	0.76	0.67-0.87	5.71E-05	1613	0.88	0.73-1.07	2.13E-01	7.87E-05	0.77	0.22	33.48
6	rs10948155	44795935	C	3319	0.79	0.7-0.89	6.47E-05	1616	1.02	0.88-1.19	7.68E-01	7.04E-05	0.79	0.5	0
6	rs633668	1.69E+08	A	3320	1.31	1.16-1.49	2.22E-05	1618	0.96	0.82-1.13	6.55E-01	1.54E-05	1.32	0.41	0
8	rs964307	1.1E+08	G	3320	0.77	0.68-0.86	1.05E-05	1616	0.92	0.79-1.06	2.49E-01	3.27E-05	0.78	0.02	81.5
8	rs9297425	1.1E+08	T	3320	0.77	0.68-0.86	1.05E-05	1611	0.91	0.78-1.06	2.31E-01	3.21E-05	0.78	0.02	81.22
8	rs7015262	1.11E+08	G	3320	0.78	0.7-0.88	3.60E-05	1612	0.93	0.81-1.08	3.68E-01	7.30E-05	0.79	0.06	71.64
9	rs12683218	12411929	G	3318	1.33	1.17-1.51	1.29E-05	1612	0.97	0.83-1.14	7.11E-01	1.15E-05	1.33	0.64	0
9	rs2289006	18768319	T	3320	0.76	0.68-0.86	4.41E-06	1609	1.01	0.87-1.16	9.36E-01	4.88E-06	0.77	0.48	0

CHR = chromosome. BP = location in base pairs. A1 = major allele. OR = odds ratio. CI = confidence interval. Q = Q statistic. I^2 = I^2 statistic.

Table A1 continued.

				DISCOVERY				MONICA/PAMELA				COMBINED ANALYSIS			
CHR	SNP	BP	A1	N	OR	95%CI	P	N	OR	95%CI	P	P-FIXED	OR-FIXED	Q	I^2
9	rs894520	38179527	C	3205	1.29	1.14-1.45	4.35E-05	1610	1.13	0.97-1.31	1.12E-01	2.86E-05	1.29	0.28	15.44
9	rs10867228	80449015	C	3320	1.37	1.17-1.61	7.85E-05	1618	1.07	0.9-1.28	4.52E-01	7.22E-05	1.37	0.69	0
9	rs10868564	89157591	C	3318	1.27	1.13-1.42	6.50E-05	1616	1.03	0.89-1.19	6.61E-01	6.16E-05	1.27	0.72	0
10	**rs13353058**	**1.25E+08**	**G**	**3317**	**1.62**	**1.34-1.96**	**5.26E-07**	**1617**	**1.01**	**0.8-1.26**	**9.55E-01**	**4.78E-07**	**1.62**	**0.76**	**0**
11	rs1255182	95112039	T	3319	0.8	0.72-0.89	6.56E-05	1614	0.89	0.77-1.03	1.07E-01	1.50E-04	0.81	0.04	76.97
11	rs3748256	95161601	G	3319	0.79	0.71-0.89	6.81E-05	1615	0.89	0.78-1.03	1.15E-01	1.36E-04	0.8	0.06	72.15
11	rs1784135	95171396	A	3319	0.79	0.71-0.89	7.90E-05	1609	0.88	0.77-1.02	8.24E-02	1.97E-04	0.81	0.03	78.94
11	rs693364	95261574	C	3320	0.79	0.7-0.88	4.45E-05	1615	0.92	0.8-1.06	2.71E-01	6.71E-05	0.79	0.14	53.52
11	rs10765777	95296033	C	3317	0.79	0.71-0.89	5.01E-05	1615	0.92	0.8-1.06	2.55E-01	7.84E-05	0.8	0.12	57.59
11	rs3808977	95297409	G	3319	0.79	0.71-0.89	6.65E-05	1613	0.92	0.8-1.06	2.34E-01	1.08E-04	0.8	0.11	60.28
11	rs11221390	1.28E+08	T	3320	0.75	0.65-0.86	2.85E-05	1613	0.93	0.77-1.12	4.40E-01	4.41E-05	0.75	0.16	48.23
12	rs10431296	7478801	C	3319	1.64	1.29-2.08	5.67E-05	1618	0.82	0.59-1.14	2.32E-01	3.98E-05	1.64	0.7	0
12	rs7961094	11803634	T	3317	1.34	1.16-1.54	7.69E-05	1614	0.89	0.73-1.07	2.17E-01	5.45E-05	1.34	0.45	0
12	rs6539747	82337800	C	3317	1.3	1.15-1.47	4.22E-05	1612	1.05	0.9-1.23	5.12E-01	4.07E-05	1.3	0.87	0
12	rs7964484	82373606	G	3318	1.3	1.15-1.46	2.36E-05	1615	1.03	0.88-1.19	7.51E-01	2.30E-05	1.3	0.98	0
13	rs9533108	41922710	C	3317	0.78	0.7-0.88	1.72E-05	1615	1.13	0.98-1.31	8.43E-02	4.71E-05	0.79	0.01	83.88
13	rs665657	41987378	T	3316	1.28	1.13-1.45	7.54E-05	1612	1.16	0.99-1.36	6.29E-02	3.33E-05	1.3	0.12	59.36
13	rs990466	48218469	G	3320	0.77	0.68-0.87	5.66E-05	1613	1.21	1.03-1.42	2.11E-02	2.17E-04	0.79	0	87.65
13	rs1164503	75765746	A	3320	1.29	1.14-1.46	3.34E-05	1615	1.09	0.95-1.26	2.31E-01	2.79E-05	1.29	0.5	0
13	rs529041	1.1E+08	A	3320	1.54	1.26-1.88	1.99E-05	1614	0.83	0.63-1.09	1.73E-01	1.58E-05	1.55	0.69	0
13	rs7995158	1.1E+08	A	3318	1.25	1.12-1.4	7.53E-05	1613	1.06	0.92-1.22	4.13E-01	6.36E-05	1.26	0.5	0

CHR = chromosome. BP = location in base pairs. A1 = major allele. OR = odds ratio. CI = confidence interval. Q = Q statistic. I^2 = I^2 statistic.

Table A1 continued.

				DISCOVERY				MONICA/PAMELA				COMBINED ANALYSIS			
CHR	SNP	BP	A1	N	OR	95%CI	P	N	OR	95%CI	P	P-FIXED	OR-FIXED	Q	I^2
15	rs7164857	91693946	T	3320	0.8	0.71-0.89	6.52E-05	1613	1.03	0.9-1.19	6.59E-01	6.53E-05	0.8	0.89	0
15	rs17541566	91697940	G	3318	1.32	1.17-1.49	9.23E-06	1614	0.92	0.79-1.08	3.18E-01	7.49E-06	1.32	0.56	0
16	rs9939858	9585398	T	3320	1.56	1.28-1.89	6.21E-06	1617	0.84	0.66-1.07	1.58E-01	1.17E-06	1.59	0.42	0
16	rs407146	13223156	T	3319	1.26	1.12-1.41	7.06E-05	1614	1.00	0.86-1.16	9.77E-01	7.16E-05	1.26	0.98	0
16	rs11647727	20263666	A	3319	0.72	0.63-0.82	7.03E-07	1613	0.92	0.79-1.07	2.80E-01	2.43E-06	0.73	0.02	80.29
16	rs4506906	20264899	C	3320	0.79	0.7-0.88	4.76E-05	1614	0.90	0.78-1.04	1.39E-01	1.71E-04	0.8	0.01	84.35
16	**rs4293393**	**20272089**	**C**	**3320**	**0.67**	**0.58-0.78**	**1.45E-07**	**1614**	**0.93**	**0.77-1.11**	**4.00E-01**	**3.30E-07**	**0.68**	**0.08**	**67.53**
16	**rs13333226**	**20273155**	**G**	**3319**	**0.67**	**0.58-0.78**	**1.14E-07**	**1615**	**0.91**	**0.76-1.08**	**2.82E-01**	**3.86E-07**	**0.68**	**0.03**	**77.71**
16	rs4496151	20280791	T	3320	0.79	0.7-0.88	4.13E-05	1608	0.91	0.79-1.05	1.84E-01	1.01E-04	0.8	0.03	77.8
18	rs1942526	53769875	G	3318	1.72	1.31-2.25	8.46E-05	1616	0.91	0.72-1.16	4.61E-01	8.20E-05	1.72	0.9	0
18	rs1893469	53776497	G	3320	1.65	1.31-2.08	2.48E-05	1617	0.99	0.8-1.24	9.46E-01	2.38E-05	1.65	0.86	0
19	rs11880417	33022748	G	3320	0.79	0.71-0.89	3.76E-05	1614	1.03	0.89-1.19	6.67E-01	3.77E-05	0.79	0.87	0
19	rs4804925	35515227	G	3319	0.73	0.63-0.84	3.04E-05	1613	1.04	0.88-1.23	6.40E-01	3.09E-05	0.73	0.78	0
19	rs444816	49130206	G	3320	1.22	1.09-1.37	5.66E-04	1613	0.94	0.82-1.08	4.07E-01	4.83E-04	1.22	0.5	0
19	rs381872	49157996	A	3320	1.26	1.13-1.41	5.73E-05	1613	0.94	0.82-1.09	4.21E-01	4.80E-05	1.26	0.49	0
19	rs383133	49158936	C	3316	1.33	1.17-1.51	1.24E-05	1615	0.91	0.78-1.08	2.84E-01	7.93E-06	1.34	0.39	0
20	rs2295179	8626446	G	3320	0.8	0.71-0.9	1.63E-04	1611	0.89	0.76-1.04	1.39E-01	3.44E-04	0.81	0.05	74.18
20	rs8123323	8643581	C	3320	0.78	0.69-0.88	4.50E-05	1614	0.93	0.79-1.09	3.71E-01	5.98E-05	0.78	0.24	28.98
20	rs172038	22309396	G	3317	0.6	0.47-0.77	5.92E-05	1617	0.95	0.74-1.21	6.53E-01	6.02E-05	0.6	0.85	0
20	rs199843	22313546	G	3316	0.59	0.46-0.76	3.80E-05	1617	1.00	0.78-1.28	9.88E-01	4.69E-05	0.6	0.38	0
20	rs6022204	51052745	A	3320	0.52	0.37-0.71	5.23E-05	1614	0.87	0.64-1.17	3.53E-01	8.26E-05	0.53	0.25	23.44

CHR = chromosome. BP = location in base pairs. A1 = major allele. OR = odds ratio. CI = confidence interval. Q = Q statistic. I^2 = I^2 statistic.

Table A1 continued.

				DISCOVERY				MONICA/PAMELA				COMBINED ANALYSIS			
CHR	SNP	BP	A1	N	OR	95%CI	P	N	OR	95%CI	P	P-FIXED	OR-FIXED	Q	I^2
20	rs2244665	53587166	G	3320	1.26	1.12-1.41	8.21E-05	1617	1.04	0.89-1.21	6.15E-01	8.13E-05	1.26	0.78	0
20	rs682132	53605875	G	3317	0.77	0.69-0.86	5.32E-06	1593	1.01	0.87-1.16	9.19E-01	5.34E-06	0.77	0.98	0
20	rs487331	53608771	T	3289	0.77	0.69-0.86	2.73E-06	1612	1.00	0.86-1.15	9.61E-01	2.76E-06	0.77	0.8	0
20	rs555848	53613135	C	3315	0.77	0.69-0.86	3.48E-06	1615	1.04	0.9-1.2	6.40E-01	3.52E-06	0.77	0.79	0

CHR = chromosome. BP = location in base pairs. A1 = major allele. OR = odds ratio. CI = confidence interval. Q = Q statistic. I^2 = I^2 statistic.

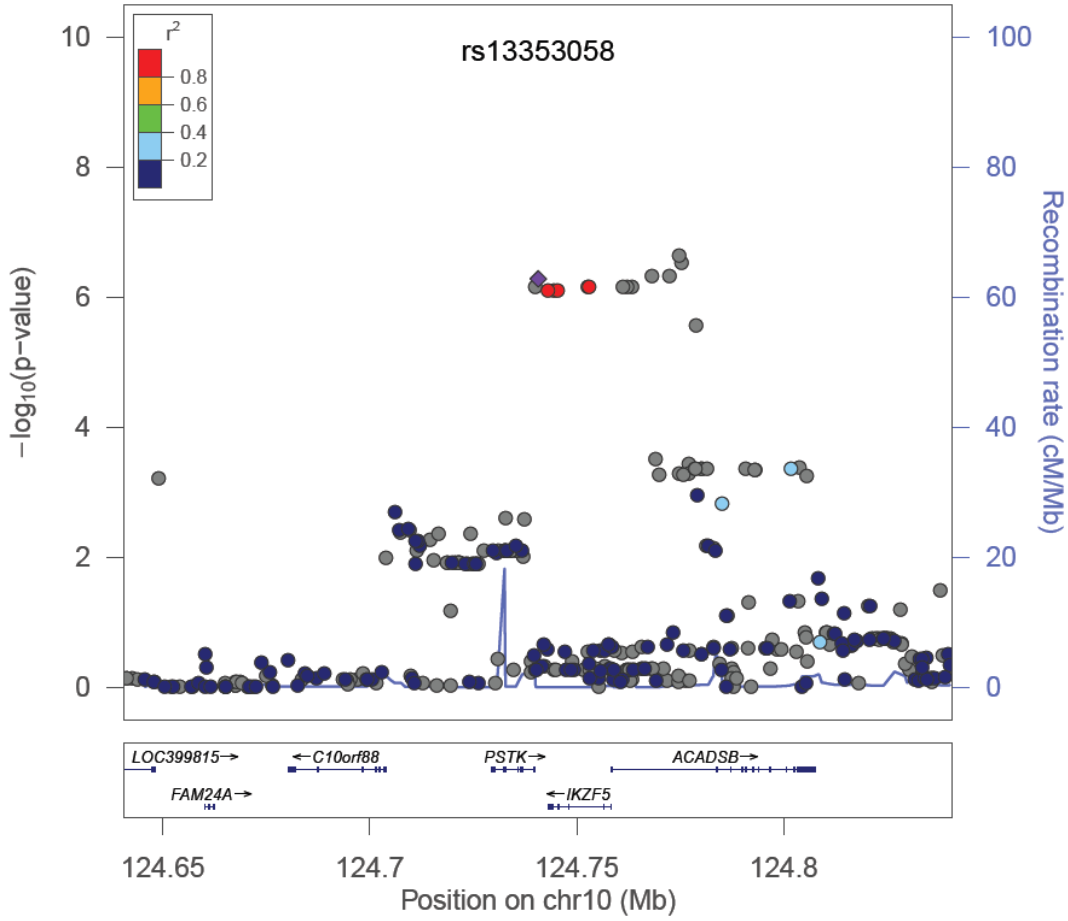


Figure A5.2 Association plot of the genomic region around rs13353058. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs13353058 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown as a purple diamond. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. LOC399815 = chromosome 10 open reading frame 88 pseudogene. C10orf88 = chromosome 10 open reading frame 88. PSTK = phosphoseryl-tRNA kinase. ACADSB = acyl-CoA dehydrogenase, short/branched chain. FAM24A = family with sequence similarity 24, member A. IKZF5 = IKAROS family zinc finger 5.

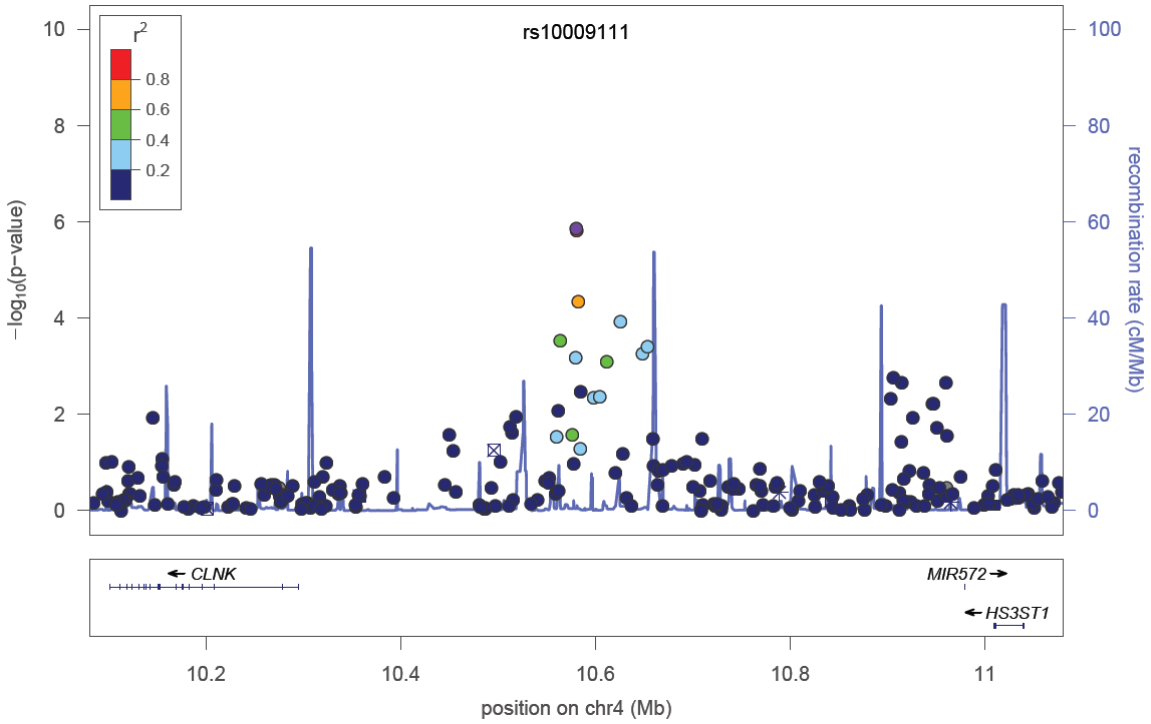


Figure A5.3 Association plot of the genomic region around rs10009111. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs10009111 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. CLNK = cytokine-dependent hematopoietic cell linker. MIR572 = microRNA 572. HS3ST1 = heparan sulfate (glucosamine) 3-O-sulfotransferase 1.

for normal immune function. A commonly used *CLNK* synonym is *MIST*. It is conserved in the chimpanzee, dog, mouse, and rat.

ADAMTS-like 1 (ADAMTSL1)

Located on chromosome 9p22.2-p22.1, rs2289006 is intronic to the *ADAMTSL1* gene, which is a member of the ADAMTS family. Figure A3 is an association plot of the genomic region around rs2289006. *ADAMTSL1* encodes a secreted protein which may have functions in the extracellular matrix. It is conserved in the dog, cow, and mouse.

Solute carrier family 4, sodium bicarbonate transporter, member 10 (SLC4A10)

Three SNPs, rs2084543 (synonymous coding; Figure A4), rs9636284 (intronic; Figure A5), and rs16846179 (intronic; Figure A6), were in LD with each other (r^2 = 0.85-0.88) in the *SLC4A10* gene on chromosome 2q23-q24. *SLC4A10* is an Na(+)-dependent Cl-/HCO3-exchanger; in an *SLC4A10* deficient mouse model regulation of internal pH was compromised and the animals had small brain ventricles and increased seizure threshold [309]. A mutation in *SLC4A10* has been associated with mental retardation and complex partial epilepsy with progressive cognitive decline [310]. It is conserved in the chimpanzee, dog, cow, mouse, rat, and zebrafish.

CKLF-like MARVEL transmembrane domain containing 7 (CMTM7)

The *CMTM7* gene is one of several genes with hypertension-associated variants on chromosome 3 in the current study. Two SNPs, rs9853991 and rs3888882, in LD (r^2 = 0.82) are intronic to *CMTM7* which belongs to the chemokine-like factor gene superfamily. Figure A7 is an association plot of the genomic region around rs17798480, which is also intronic to *CMTM7*. The protein that *CMTM7* encodes is highly expressed in leukocytes; however its function is unknown. The gene is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

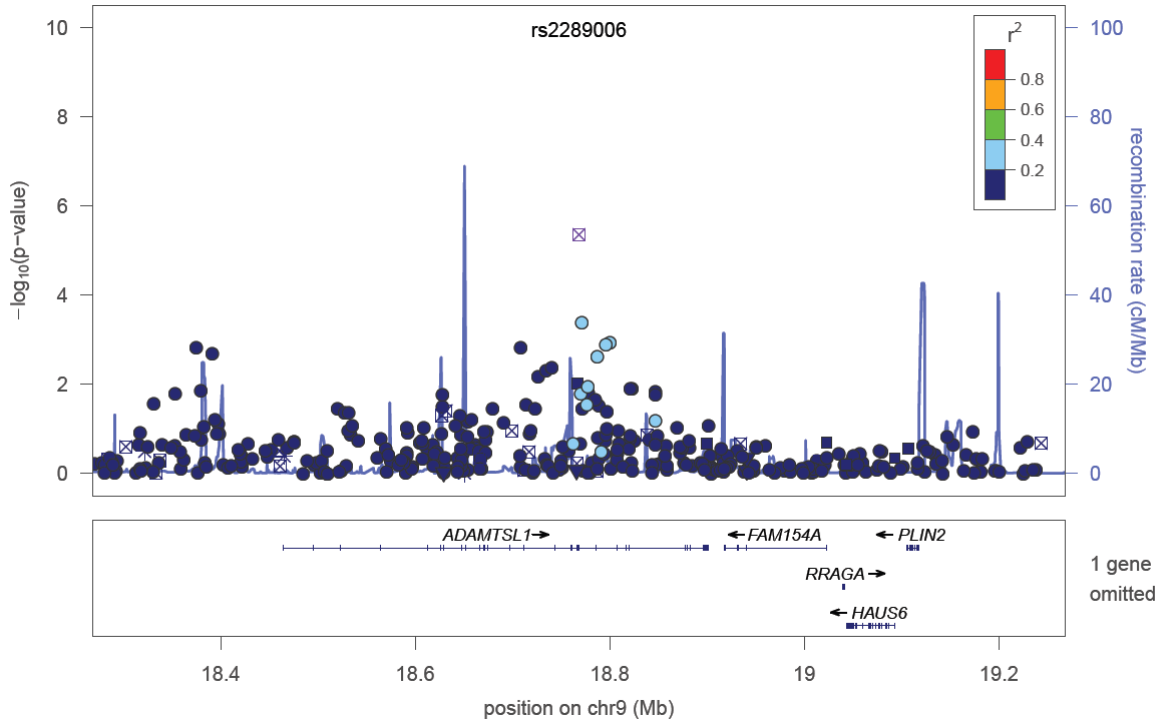


Figure A3 Association plot of the genomic region around rs2289006. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs2289006 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. ADAMTSL1 = ADAMTS-like 1. FAM154A = family with sequence similarity 154, member A. PLIN2 = perilipin 2. RRAGA = Ras-related GTP binding A. HAUS6 = HAUS augmin-like complex, subunit 6.

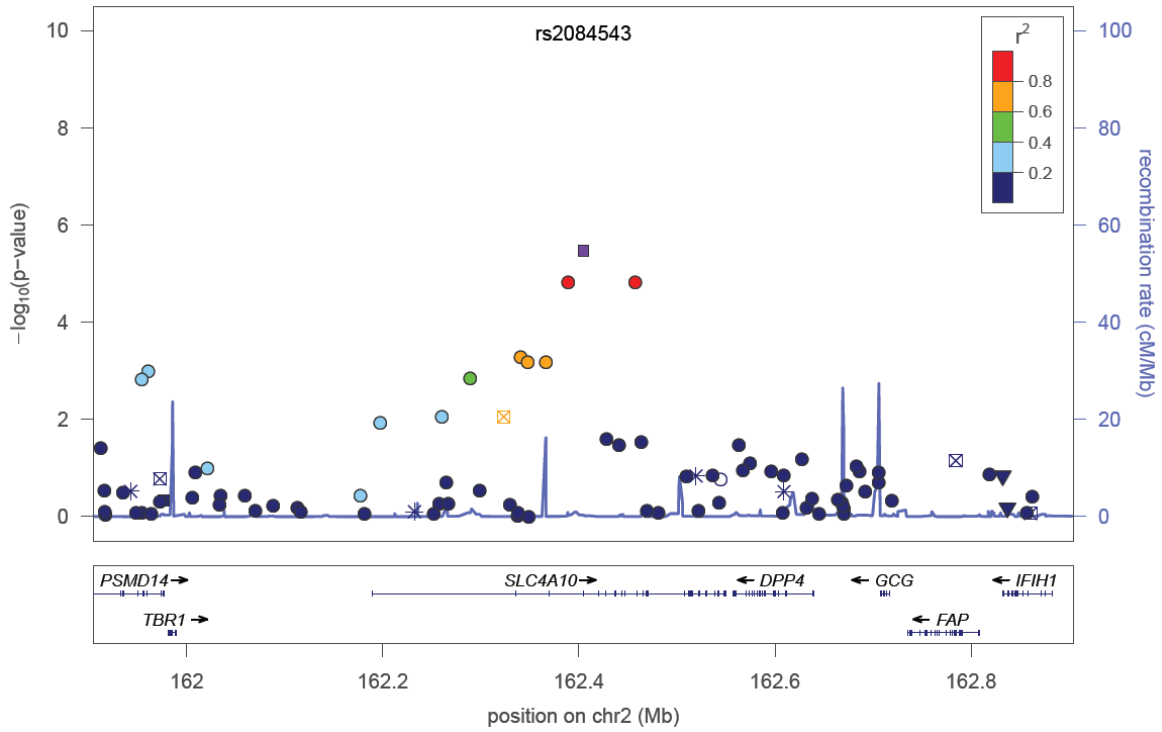


Figure A4 Association plot of the genomic region around rs2084543. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs2084543 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PSMD14 = proteasome (prosome, macropain) 26S subunit, non-ATPase, 14. SLC4A10 = solute carrier family 4, sodium bicarbonate transporter, member 10. DPP4 = dipeptidyl-peptidase 4. GCG = glucagon. IFIH1 = interferon induced with helicase C domain 1. TBR1 = T-box, brain, 1. FAP = fibroblast activation protein, alpha.

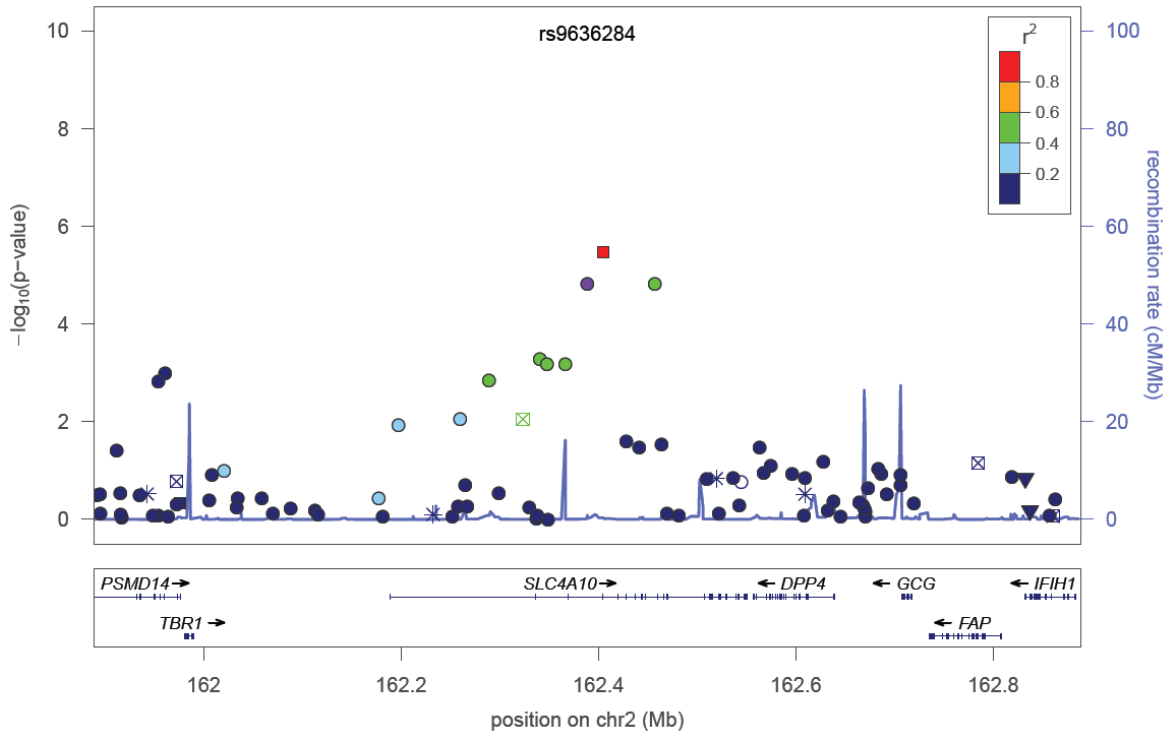


Figure A5 Association plot of the genomic region around rs9636284. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs9636284 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PSMD14 = proteasome (prosome, macropain) 26S subunit, non-ATPase, 14. SLC4A10 = solute carrier family 4, sodium bicarbonate transporter, member 10. DPP4 = dipeptidyl-peptidase 4. GCG = glucagon. IFIH1 = interferon induced with helicase C domain 1. TBR1 = T-box, brain, 1. FAP = fibroblast activation protein, alpha.

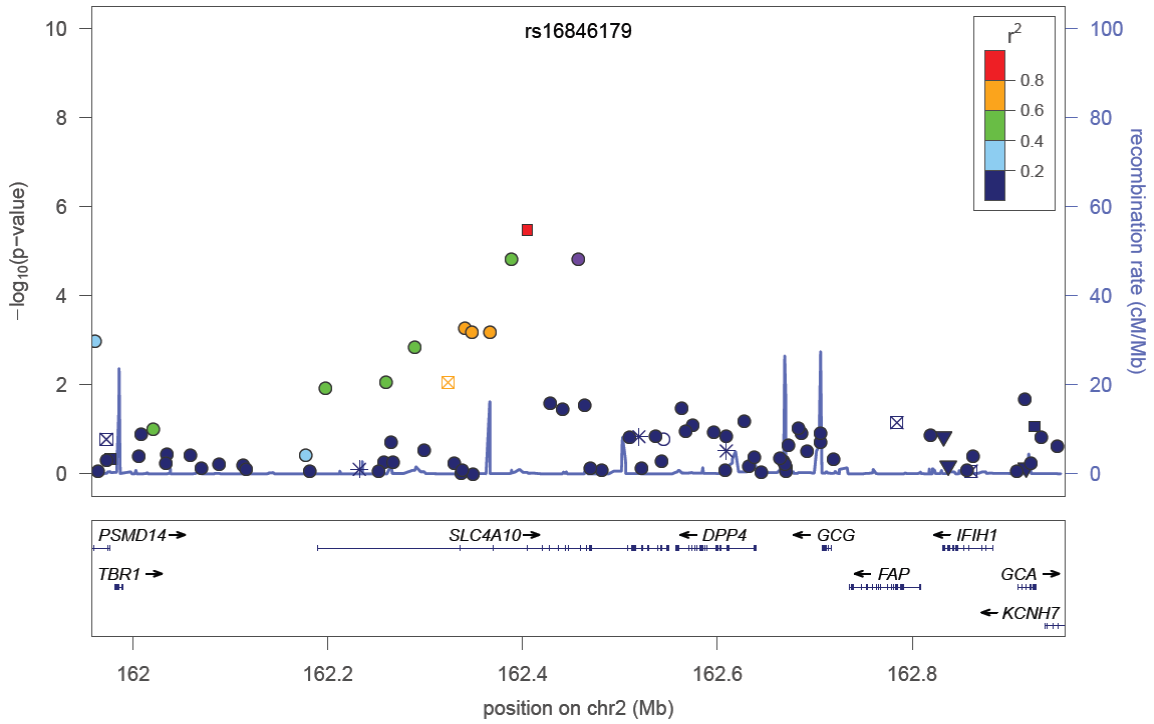


Figure A6 Association plot of the genomic region around rs16846179. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs16846179 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PSMD14 = proteasome (prosome, macropain) 26S subunit, non-ATPase, 14. SLC4A10 = solute carrier family 4, sodium bicarbonate transporter, member 10. DPP4 = dipeptidyl-peptidase 4. GCG = glucagon. IFIH1 = interferon induced with helicase C domain 1. TBR1 = T-box, brain, 1. FAP = fibroblast activation protein, alpha. GCA = grancalcin, EF-hand calcium binding protein. KCNH7 = potassium voltage-gated channel, subfamily H (eag-related), member 7.

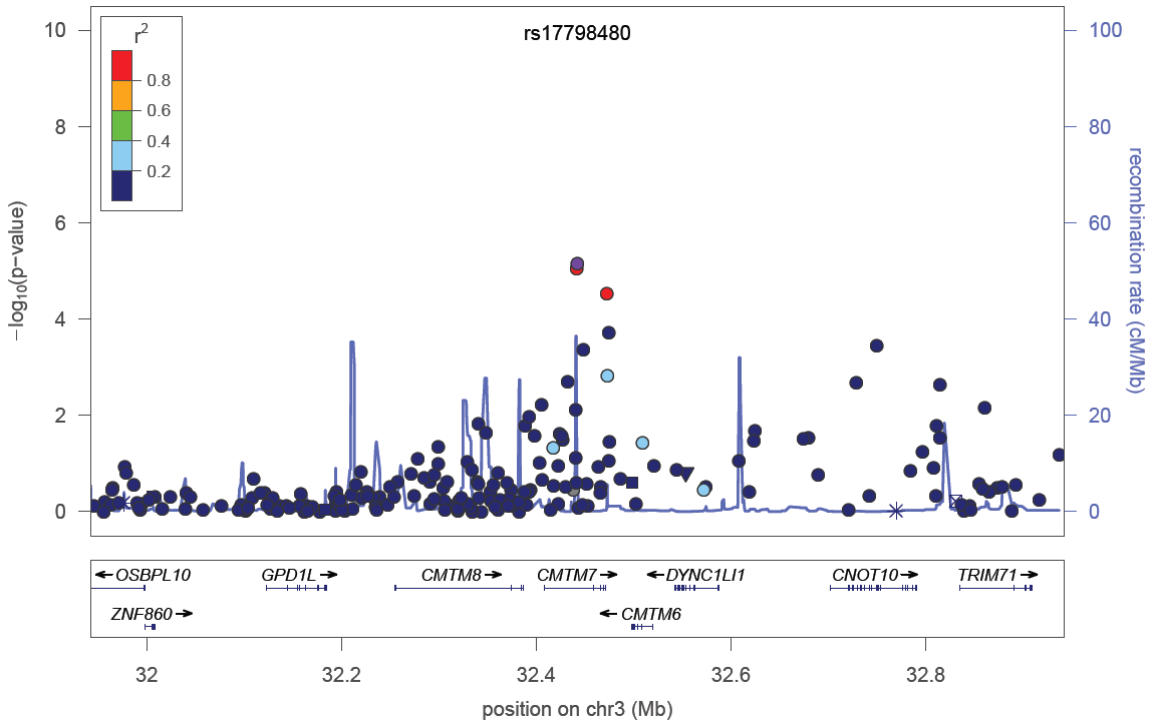


Figure A7 Association plot of the genomic region around rs17798480. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs17798480 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. OSBPL10 = oxysterol binding protein-like 10. GPD1L = glycerol-3-phosphate dehydrogenase 1-like. CMTM8 = CKLF-like MARVEL transmembrane domain containing 8. CMTM7 = CKLF-like MARVEL transmembrane domain containing 7. DYNC1LI1 = dynein, cytoplasmic 1, light intermediate chain 1. CNOT10 = CCR4-NOT transcription complex, subunit 10. TRIM71 = tripartite motif-containing 71. ZNF860 = zinc finger protein 860. CMTM6 = CKLF-like MARVEL transmembrane domain containing 6.

KIAA2018

rs12636240 is intronic to the *KIAA2018* gene on chromosome 3q13.2, but is in LD with SNPs in other genes: rs9865965 ($r^2 = 0.95$) and rs9881563 ($r^2 = 1.00$) which are intronic to *NAA50*; and rs9828099 ($r^2 = 0.95$) and rs13061150 ($r^2 = 0.95$) which are intronic to *ATP6V1A*. Furthermore rs12636240 is in LD with non-synonymous coding SNPs in the region, untyped in the current study. The *KIAA2018* gene is protein coding but little further information is available. It is conserved in the chimpanzee, dog, cow, mouse, rat, and chicken.

N-alpha-acetyltransferase 50, NatE catalytic subunit (NAA50)

Two SNPs, rs9881563 and rs9865965, are intronic to the *NAA50* gene on chromosome 3q13.2 and are in LD with each other ($r^2 = 0.95$). They are also in LD with SNPs in *KIAA2018* and *ATP6V1A* and with non-synonymous coding SNPs in the region, untyped in the current study. The *NAA50* gene encodes a protein whose probable function is as a catalytic component of the ARD1A-NARG1 complex. A commonly used synonym is *NAT13*. It is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, C.elegans, A.thaliana, and rice.

ATPase, H+ transporting, lysosomal 70kDa, V1 subunit A (ATP6V1A)

Two SNPs, rs13061150 and rs9828099, are intronic to the *ATP6V1A* gene on chromosome 3q13.2 and are in LD with each other ($r^2 = 1.00$). They are also in LD with SNPs in *KIAA2018* and *NAA50* and with non-synonymous coding SNPs in the region, untyped in the current study. The *ATP6V1A* gene encodes a component of vacuolar ATPase (V-ATPase), a multisubunit enzyme that mediates acidification of eukaryotic intracellular organelles. The protein encoded by *ATP6V1A* is expressed in all tissues, and has been found to be differentially expressed in both the Wernicke's Area [311] and the dorsolateral prefrontal cortex [312] in patients with schizophrenia. It is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, C.elegans, S.pombe, S.cerevisiae, K.lactis, M.grisea, N.crassa, A.thaliana, rice, and P.falciparum.

Transferrin (TF)

The *TF* gene encodes transferrin, a glycoprotein that transports iron from the intestine, reticuloendothelial system, and liver parenchymal cells to all proliferating cells. Located on chromosome 3, rs3811647 is intronic to *TF* and along with two other variants in *TF* and the *HFE* C282Y mutation explains 40% of the genetic variation in serum transferrin [313]. Figure A8 is an association plot of the genomic region around rs3811647. Variants in *TF* are associated with atransferrinemia, a rare autosomal recessive disorder characterised by iron loading and microcytic anaemia [314]. The gene is conserved in the dog, cow, mouse, rat, chicken, and zebrafish. rs3811647 is in LD with rs6794945 ($r^2 = 0.86$) which is intronic to the *SRPRB* gene.

Signal recognition particle receptor, B subunit (SRPRB)

rs6794945 is located on chromosome 3q22.1 in the *SRPRB* gene, whose encoded protein may be a component of the signal recognition particle receptor and moreover may play a role in the development of colon cancer [315]. A commonly used synonym of *SRPRB* is *APMCF1*, and it is located close to the *TF* gene (Figure A8). There is also evidence that it participates in cell cycle regulation [316]. The gene is conserved in the dog, mouse, rat, zebrafish, fruit fly, mosquito, C.elegans, A.thaliana, rice, and P.falciparum. rs6794945 is in LD with rs3811647 ($r^2 = 0.86$) which is intronic to the *TF* gene.

Leucine, glutamate and lysine rich 1 (LEKR1)

Three SNPs associated with hypertension in the current study, rs7635876, rs1842840, and rs11715321, are intronic to the *LEKR1* gene on chromosome 3q25. rs1842840 and rs11715321 are in LD (r2 = 0.97). A variant near *LEKR1* was associated with birth weight in a recent meta-analysis of six GWAS [317]. The *LEKR1* gene is conserved in the chimpanzee, dog, cow, rat, and chicken.

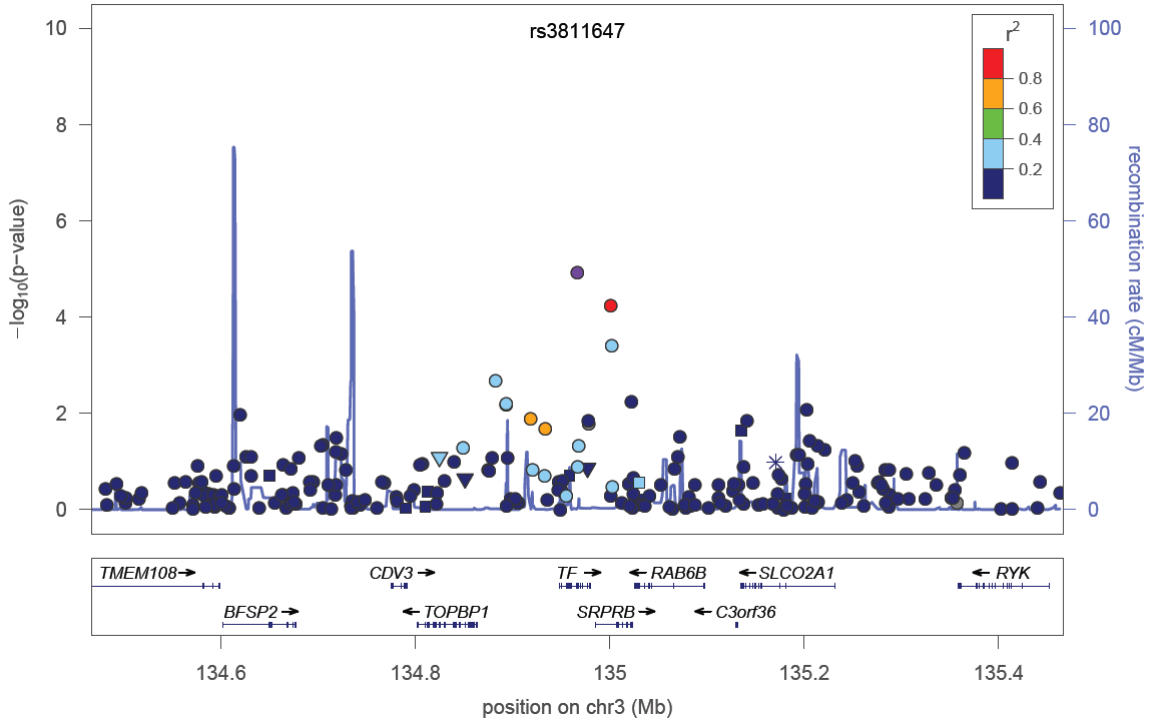


Figure A8 Association plot of the genomic region around rs3811647. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs3811647 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. TMEM108 = transmembrane protein 108. CDV3 = CDV3 homolog (mouse). TF = transferrin. RAB6B = RAB6B, member RAS oncogene family. SLCO2A1 = solute carrier organic anion transporter family, member 2A1. RYK = RYK receptor-like tyrosine kinase. BFSP2 = beaded filament structural protein 2, phakinin. TOPBP1 = topoisomerase (DNA) II binding protein 1. SRPRB = signal recognition particle receptor, B subunit. C3orf36 = chromosome 3 open reading frame 36.

B-cell scaffold protein with ankyrin repeats 1 (BANK1)

rs4482766 is located 18.9kb from the *BANK1* gene on chromosome 4q24. rs2866431 (Figure A9), rs4487344 (Figure A10) and rs13124455 (Figure A11) are also near to *BANK1*. The protein encoded by *BANK1* functions in B-cell receptor-induced calcium mobilisation from intracellular stores. In addition it can promote Lyn-mediated tyrosine phosphorylation of inositol 1,4,5-trisphosphate receptors. Functional polymorphisms in *BANK1* have been associated with systemic lupus erythematosus [318], rheumatoid arthritis [319], and systemic sclerosis [320]. The gene is conserved in the chimpanzee, dog, cow, mouse, and rat.

Polymerase (DNA-directed) sigma (POLS)

rs106415 is located 35.7kb from the *POLS* gene on chromosome 5p15. A common synonym for *POLS* is *PAPD7*. The protein it encodes is a DNA polymerase that is probably involved in DNA repair, and may also be required for sister chromatid adhesion. The *POLS* gene is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

Nipped-B homolog (Drosophila) (NIPBL)

Two SNPs in the current study, rs292196 and rs16903459, were in LD ($r^2 = 0.82$) and located in the *NIPBL* gene on chromosome 5p13.2. *NIPBL* encodes the homolog of the Drosophila melanogaster Nipped-B gene product and fungal sister chromatid cohesion 2 homolog –type sister chromatid cohesion proteins. Figure A12 is an association plot of the genomic region around rs172384 which is also located close to *NIPBL*. The Drosophila protein is involved in developmental regulation and it is homologous to a family of chromosomal adherins (involved in sister chromatid cohesion, chromosome condensation and DNA repair). Mutations in *NIPBL* cause Cornelia de Lange syndrome [321, 322], a heterogeneous developmental disorder characterised by facial dysmorphism, delayed growth, cognitive retardation and other malformations. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, A.thaliana, and rice.

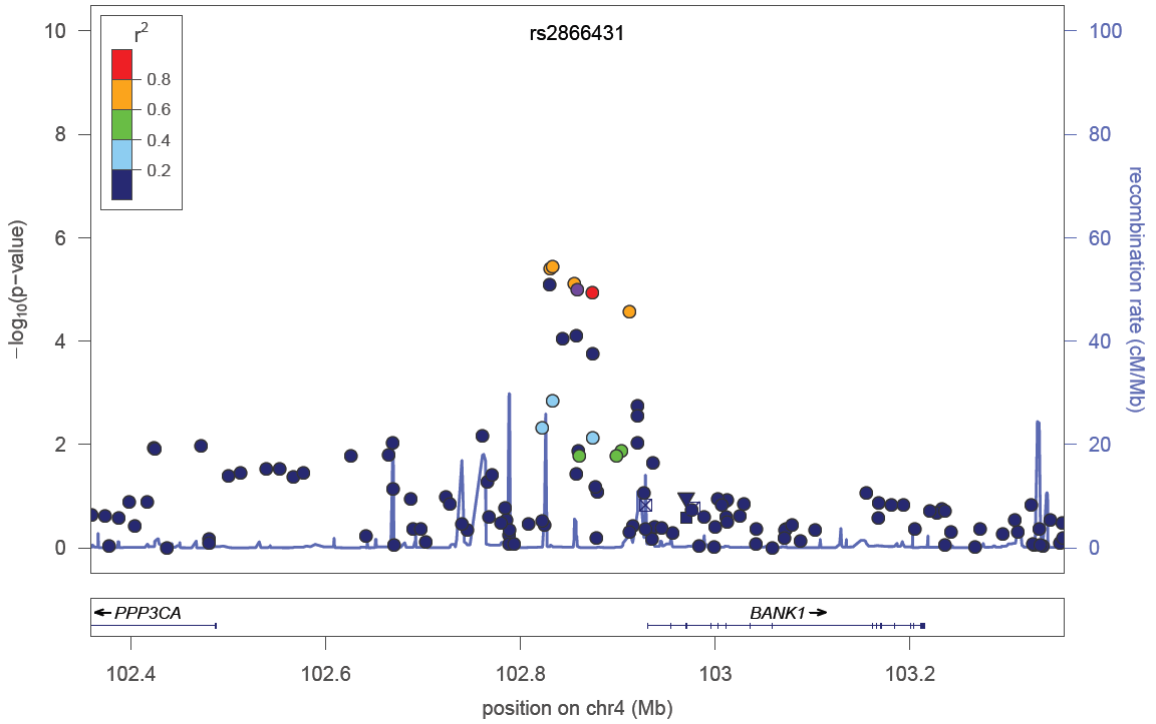


Figure A9 Association plot of the genomic region around rs2866431. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs2866431 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PPP3CA = protein phosphatase 3, catalytic subunit, alpha isozyme. BANK1 = B-cell scaffold protein with ankyrin repeats 1.

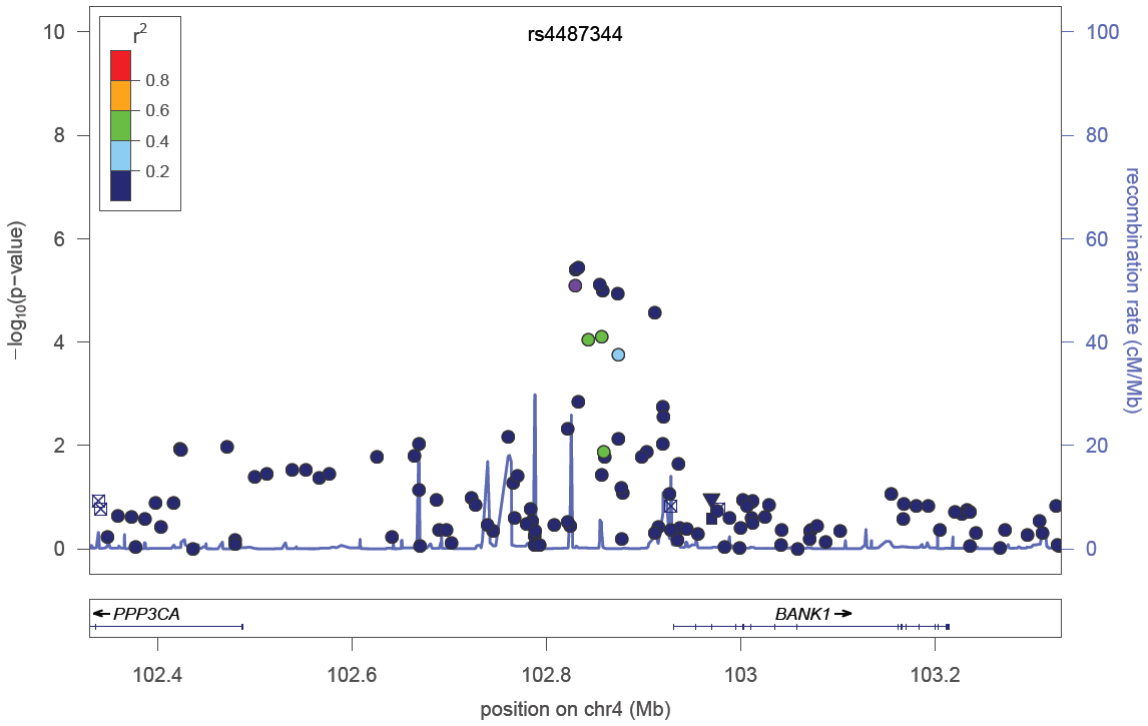


Figure A10 Association plot of the genomic region around rs4487344. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs4487344 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PPP3CA = protein phosphatase 3, catalytic subunit, alpha isozyme. BANK1 = B-cell scaffold protein with ankyrin repeats 1.

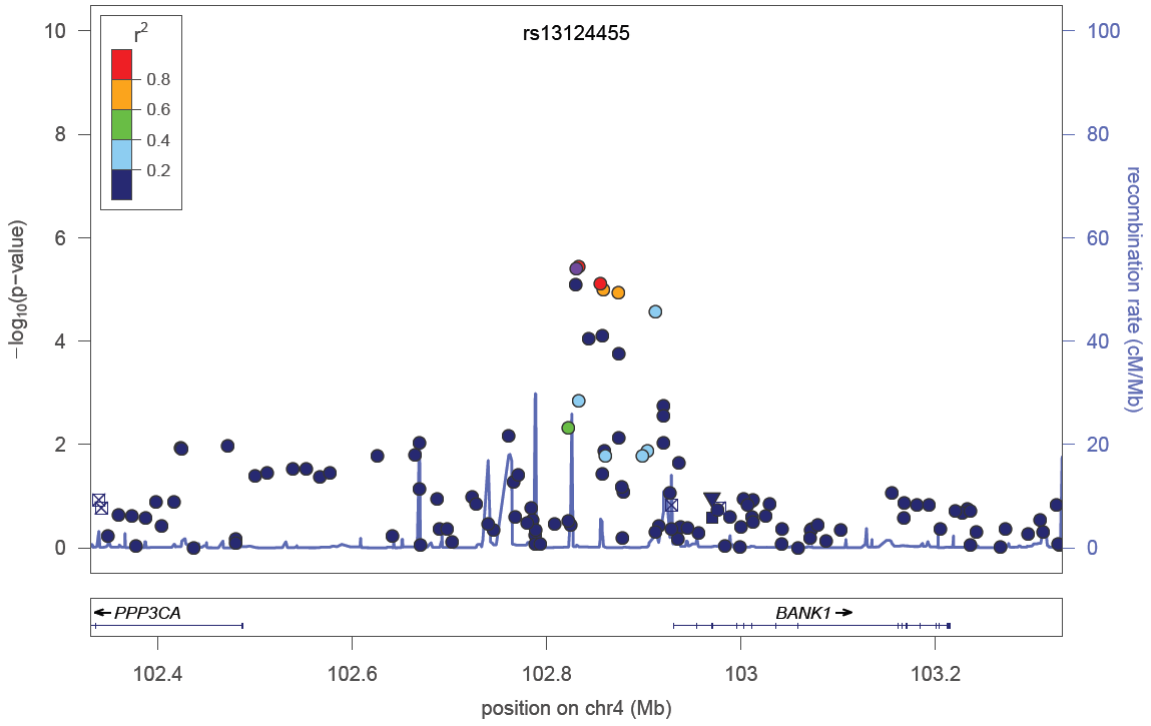


Figure A11 Association plot of the genomic region around rs13124455. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs13124455 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. PPP3CA = protein phosphatase 3, catalytic subunit, alpha isozyme. BANK1 = B-cell scaffold protein with ankyrin repeats 1.

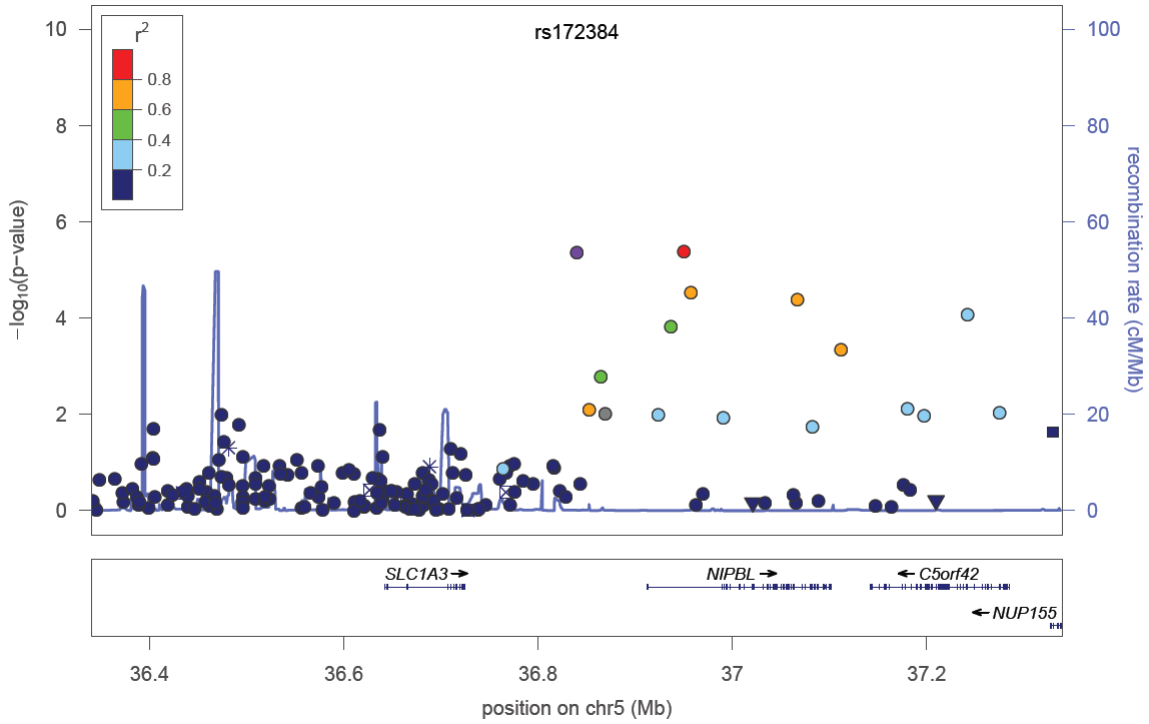


Figure A12 Association plot of the genomic region around rs172384. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs172384 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. SLC1A3 = solute carrier family 1 (glial high affinity glutamate transporter), member 3. NIPBL = Nipped-B homolog (Drosophila). C5orf42 = chromosome 5 open reading frame 42. NUP155 = nucleoporin 155kDa.

S100 calcium binding protein Z (S100Z)

rs2460498 is located 4.0kb upstream of the *S100Z* gene on chromosome 5q13.3. *S100Z* encodes a member of the S100 protein family which is involved in calcium binding. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Supressor of Ty 3 homolog (S.cerevisiae) (SUPT3H)

rs10948155 is located 89.0kb from the *SUPT3H* gene on chromosome 6p21. Polymorphisms in and near *SUPT3H* have been associated with height [323, 324] and adult attention-deficit/hyperactivity disorder [325]. The protein it encodes is likely a transcriptional activator. The *SUPT3H* gene is conserved in the chimpanzee, cow, mouse, chicken, and zebrafish.

Thrombospondin 2 (THBS2)

The gene that rs633668 is closest to, at 4kb downstream, is *RP3-495K2.2* for which there is no information available. However, rs633668 is approximately 142kb upstream from the *THBS2* gene on chromosome 6q27, for which there is further information. A commonly used synonym is *TSP2*. Figure A13 is an association plot of the genomic region around rs633668. The protein encoded by *THBS2* belongs to the thrombospondin family. Studies in mice have shown that it functions as a potent inhibitor of tumour growth and angiogenesis [326] and that it may modulate the cell surface properties of mesenchymal cells and be involved in cell adhesion and migration [327]. Expression in humans has been correlated with microvessel counts in salivary gland carcinomas [328]. A variant in *THBS2* has been associated in Japanese samples with lumbar-disc herniation, a cause of lower back pain and unilateral leg pain [329]. Of greater relevance to the current study, other variants have been associated with risk of thoracic aortic aneurysm in hypertensive patients [254] and premature MI [255]. However, a meta-analysis of all the evidence linking *THBS2* polymorphisms with MI found no association [330]. The *THBS2* gene is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

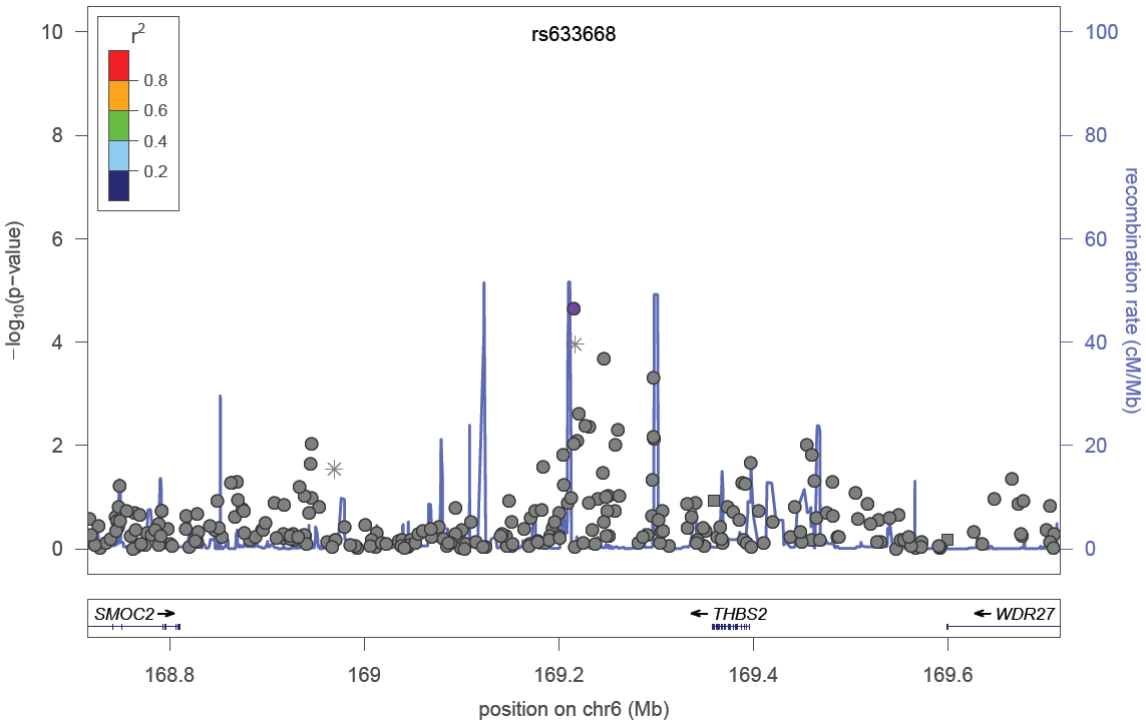


Figure A13 Association plot of the genomic region around rs633668. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs633668 and the surrounding SNPs could not be determined, hence all surrounding SNPs are grey. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. SMOC2 = SPARC related modular calcium binding 2. THBS2 = thrombospondin 2. WDR27 = WD repeat domain 27.

Polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1 (PKHD1L1)

Three SNPs, rs964307 (non-synonymous coding), rs9297425 (intronic; Figure A14), and rs7015262 (intronic) are in LD with each other (r^2 = 0.94-1.00) and located in the *PKHD1L1* gene on chromosome 8q23. *PKHD1L1* may have a role in cellular immunity and mutations in its homolog, *PKHD1*, result in autosomal-recessive polycystic kidney disease [331]. The *PKHD1L1* gene is conserved in the chimpanzee, dog, cow, mouse, and zebrafish.

Tyrosinase-related protein 1 (TYRP1)

rs12683218 is located in an intergenic region 121kb from the *AL162422.1* gene for which there is no information available. At a distance of approximately 268kb from rs12683218 is the *TYRP1* gene on chromosome 9p23. The encoded protein is a melanosomal enzyme that has a role in the melanin biosynthetic pathway. Figure A15 is an association plot of the genomic region around rs12683218. Defects in *TYRP1* cause the autosomal recessive disorder Rufous oculocutaneous albinism [332], in which the biosynthesis of melanin pigment is reduced. A SNP in *TYRP1*, rs1408799, has been associated with blue versus green eye colour [333] and melanoma risk [334, 335]. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Family with sequence similarity 76, member B (FAM76B)

rs1255182 is located 29.7kb from the *FAM76B* gene on chromosome 11q21. rs3748256 is intronic to *FAM76B* and is in LD with SNPs in other genes: rs693364 (r^2 = 0.96), rs10765777 (r^2 = 0.93), and rs3808977 (r^2 = 0.93) which are intronic to *MTMR2*; and rs1784135 (r^2 = 1.00) which is intronic to *CEP57*; as well as other SNPs untyped in the current study. The *FAM76B* gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, mosquito, and C.elegans.

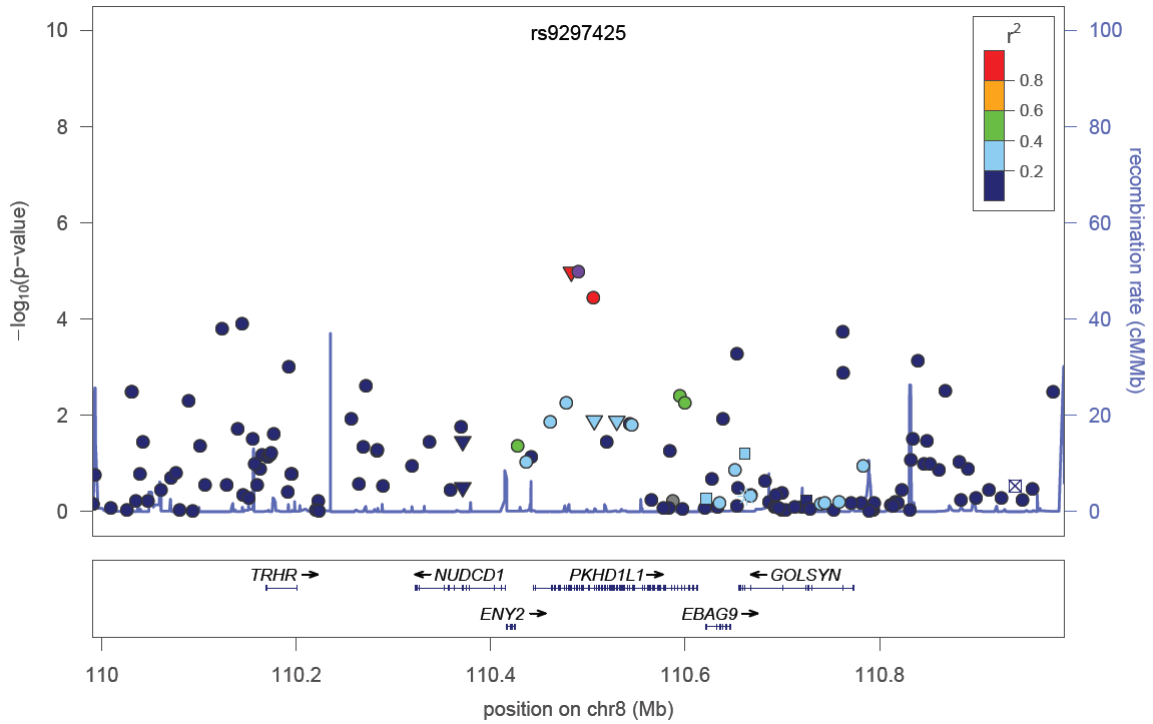


Figure A14 Association plot of the genomic region around rs9297425. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs9297425 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. TRHR = thyrotropin-releasing hormone receptor. NUDCD1 = NudC domain containing 1. PKHD1L1 = polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1. GOLSYN = syntabulin (syntaxin-interacting). ENY2 = enhancer of yellow 2 homolog (Drosophila). EBAG9 = estrogen receptor binding site associated, antigen, 9.

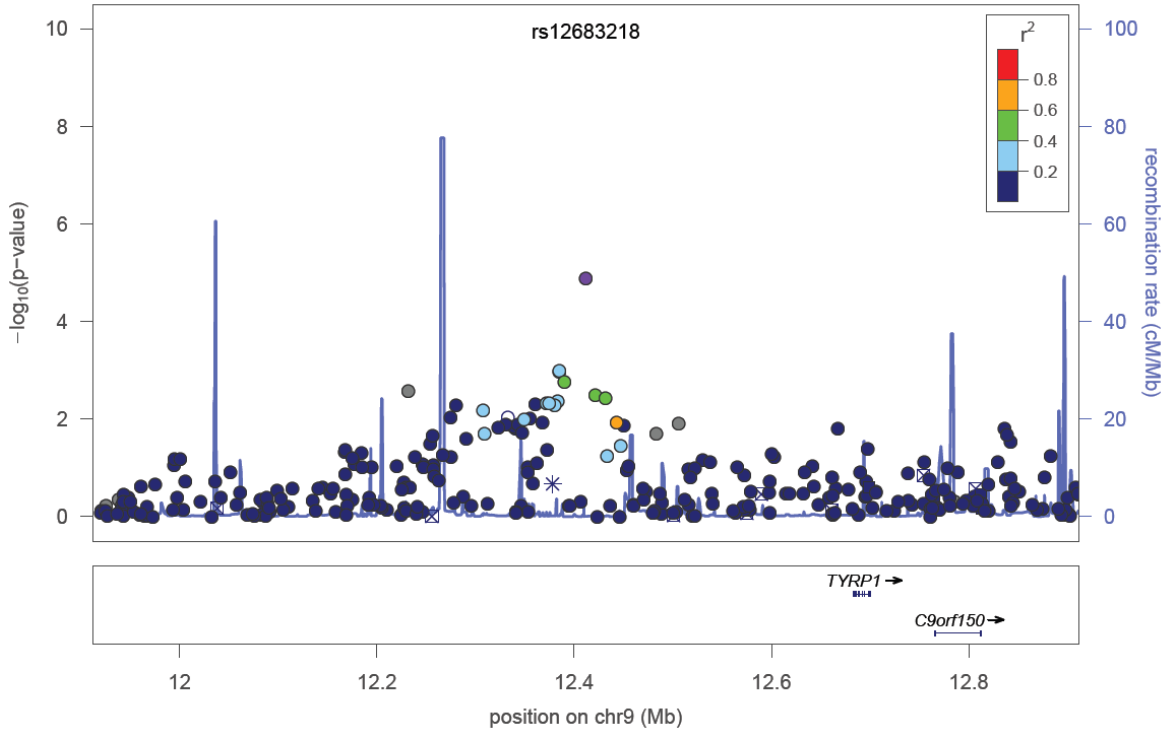


Figure A15 Association plot of the genomic region around rs12683218. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs12683218 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. TYRP1 = tyrosinase-related protein 1. C9orf150 = chromosome 9 open reading frame 150.

Centrosomal protein 57kDa (CEP57)

rs1784135 is intronic to *CEP57* on chromosome 11q21, and is in LD with rs693364 (r^2 = 0.96), rs10765777 (r^2 = 0.93), and rs3808977 (r^2 = 0.93) (all intronic to *MTMR2*), and rs3748256 (r^2 = 1.00) (intronic to *FAM76B*), and other untyped SNPs. *CEP57* encodes the protein translokin which binds basic fibroblast growth factor and mediates its nuclear translocation and mitogenic activity. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Myotubularin related protein 2 (MTMR2)

Three hypertension associated polymorphisms in the current study, rs693364, rs10765777, and rs3808977, are in LD with each other (r^2 = 0.96 – 1.00) and located in the *MTMR2* gene on chromosome 11q22. They are in LD with rs3748256 in *FAM76B* (r^2 = 0.93 – 0.96) and rs1784135 in *CEP57* (r^2 = 0.93 – 0.96), and other untyped SNPs. *MTMR2* is a member of the myotubularin family, and the protein it encodes has phosphatase activity towards lipids with a phosphoinositol headgroup. Mutations in *MTMR2* cause Charcot-Marie-Tooth disease type 4B [336], an autosomal recessive demyelinating peripheral neuropathy that can also result from mutations in *MTMR13*. *MTMR2* is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, A.thaliana, and rice.

V-ets erythroblastosis virus E26 oncogene homolog 1 (avian) (ETS1)

rs11221390 is located 19.7kb from the *ETS1* gene on chromosome 11q23-q24. *ETS1* encodes the protein ETS1, one of the ETS transcription factors that regulate several genes and are involved in cell senescence and death, stem cell development, and tumorigenesis. Recent GWAS have discovered variants in *ETS1* associated, at the level of genome-wide significance, with systemic lupus erythematosus [337, 338] and celiac disease [339]. It is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

CD163 molecule-like 1 (CD163L1)

rs10431296 is located in the *CD163L1* gene on chromosome 12p13.3. The protein encoded by *CD163L1* is a member of the scavenger receptor cysteine-rich (SRCR) superfamily. Members of SRCR are mainly found in immune system related cells and are defined by a 100-110 amino acid SRCR domain, which possibly mediates protein-protein interaction and ligand binding.

Ets variant 6 (ETV6)

rs7961094 is located in the *ETV6* gene on chromosome 12p13. *ETV6* encodes an ETS family transcription factor which is involved in protein-protein interactions and DNA binding. Chromosomal rearrangements in *ETV6* and fusions between it and other genes have been linked to malignant eosinophil proliferation [340], congenital fibrosarcoma [341], and acute myeloid leukaemia [342]. It is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Transmembrane and tetratricopeptide repeat containing 2 (TMTC2)

Two SNPs, rs6539747 and rs7964484, are in LD ($r^2 = 0.85$) and located in an intergenic region around 268-304kb from the *RP11-87P13.1* gene. The closest gene for which there is information available is the *TMTC2* gene, approximately 400kb from rs7964484 on chromosome 12q21.31. Figure A16 is an association plot of the genomic region around rs7964484. It is conserved in the dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, and C.elegans.

Fatty acid binding protein 3, pseudogene 2 (FABP3P2)

rs9533108 is located 303.5kb from the pseudogene *FABP3P2* on chromosome 13q13-q14 (Figure A17). Recent GWAS have observed genome-wide significant associations between decreased bone mineral density and variants in *TNFSF11* [343, 344] and *AKAP11* [345] on chromosome 13q14.11.

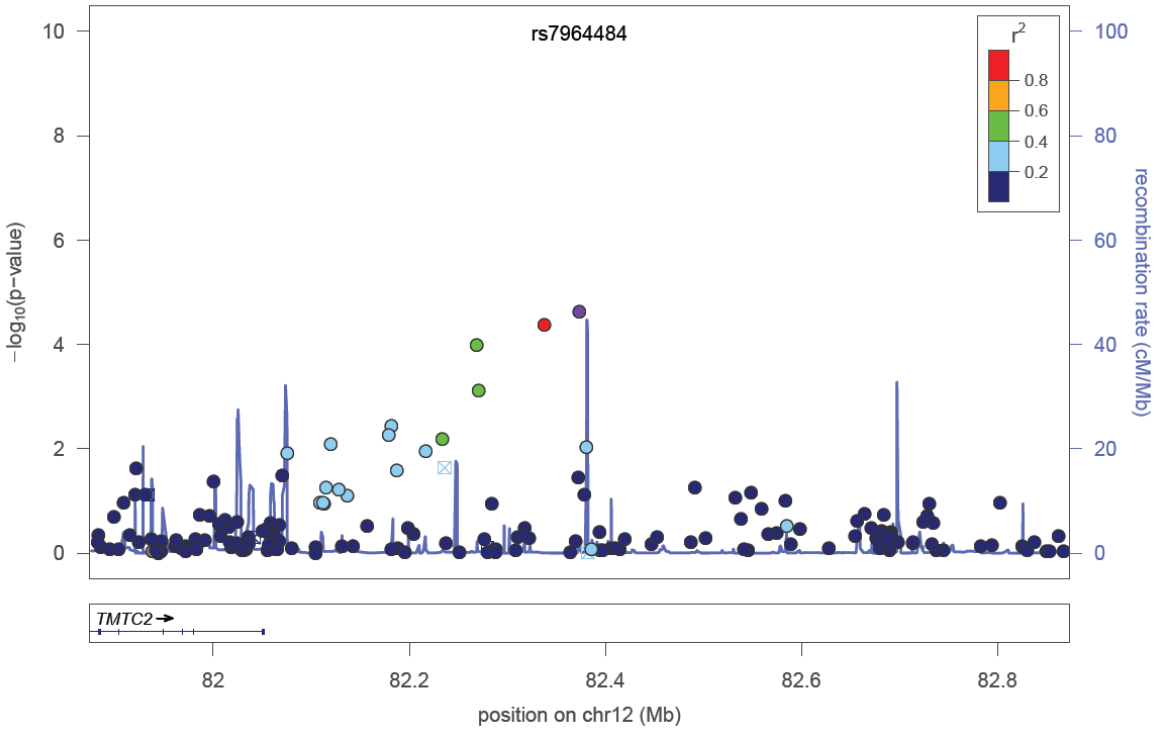


Figure A16 Association plot of the genomic region around rs7964484. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs7964484 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. TMTC2 = transmembrane and tetratricopeptide repeat containing 2.

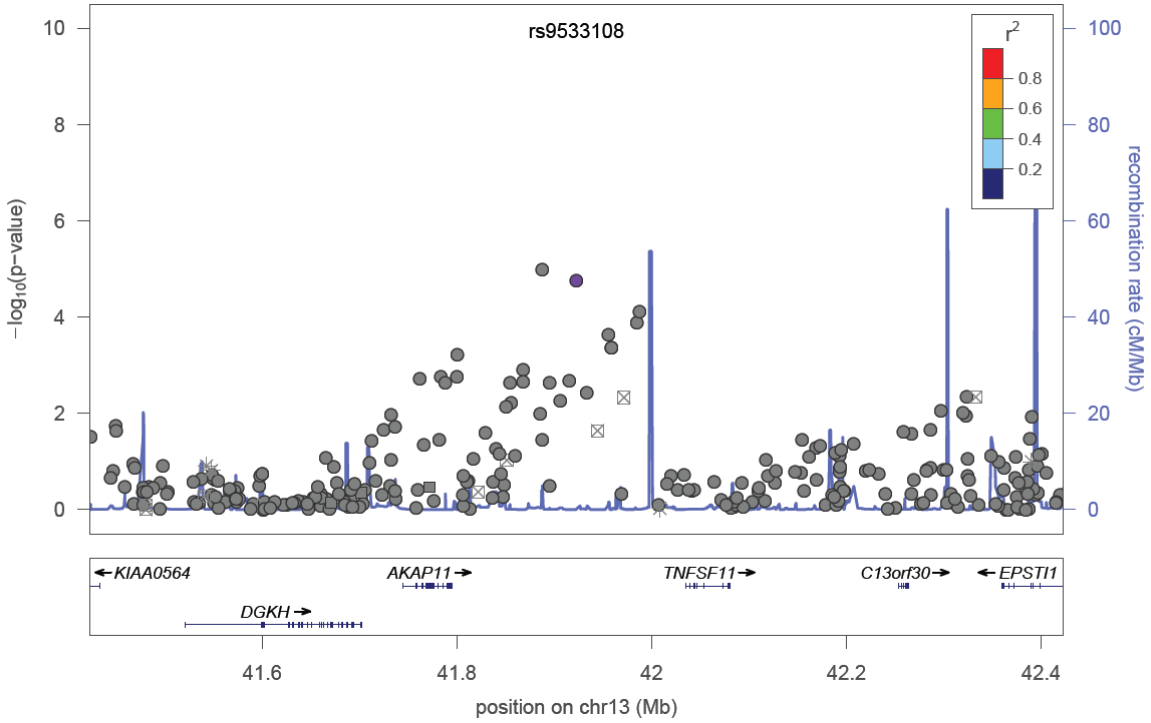


Figure A17 Association plot of the genomic region around rs9533108. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs9533108 and the surrounding SNPs could not be determined, hence all surrounding SNPs are grey. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. KIAA0564 = KIAA0564. AKAP11 = A kinase (PRKA) anchor protein 11. TNFSF11 = tumor necrosis factor (ligand) superfamily, member 11. C13orf30 = chromosome 13 open reading frame 30. EPSTI1 = epithelial stromal interaction 1 (breast). DGKH = diacylglycerol kinase, eta.

Tumour necrosis factor (ligand) superfamily, member 11 (TNFSF11)

rs665657 is located 47.5kb from the *TNFSF11* gene on chromosome 13q14.11. A commonly used synonym for *TNFSF11* is *RANKL*. The genomic region around *TNFSF11* is shown in the association plots for rs9533108 (Figure A17) and rs9594746 (Figure A18). The encoded protein is a member of the tumour necrosis factor (TNF) cytokine family and is a key factor for osteoclast differentiation and activation. Variants in *TNFSF11* have been linked to decreased bone mineral density [343, 344] and osteopetrosis [346], a rare disease characterised by abnormally dense bone that can occur as a severe autosomal recessive form or a benign autosomal dominant form.

Collagen, type IV, alpha 1 (COL4A1)

rs529041 is located in the *COL4A1* gene on chromosome 13q34. Figure A19 is an association plot of the genomic region around rs529041. The protein encoded by *COL4A1* is the major type IV alpha collagen chain of basement membranes. Mutations in *COL4A1* have been implicated in porencephaly, a rare neurological disease characterised by degenerative cavities in the brain [347, 348]; brain small vessel disease with haemorrhage [349] and with Axenfeld-Rieger anomaly [350]; heredity angiopathy with nephropathy, aneurysms, and muscle cramps [257]; and arterial stiffness [256]. *COL4A1* is conserved in the dog, cow, mouse, rat, chicken, and zebrafish.

RGM domain family, member A (RGMA)

rs17541566 is located 67kb from the *AC091078.2* gene, for which no information is available. The nearest gene that has further information is the *RGMA*, approximately 360kb from rs17541566 on chromosome 15q26.1 (Figure A20). The protein it encodes is a glycosylphosphatidylinositol-anchored glycoprotein that functions as an axon guidance protein in the developing and adult central nervous system. Furthermore it may act as a tumour suppressor in Hodgkin's lymphoma [351] and colon cancer [352]. The *RGMA* gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

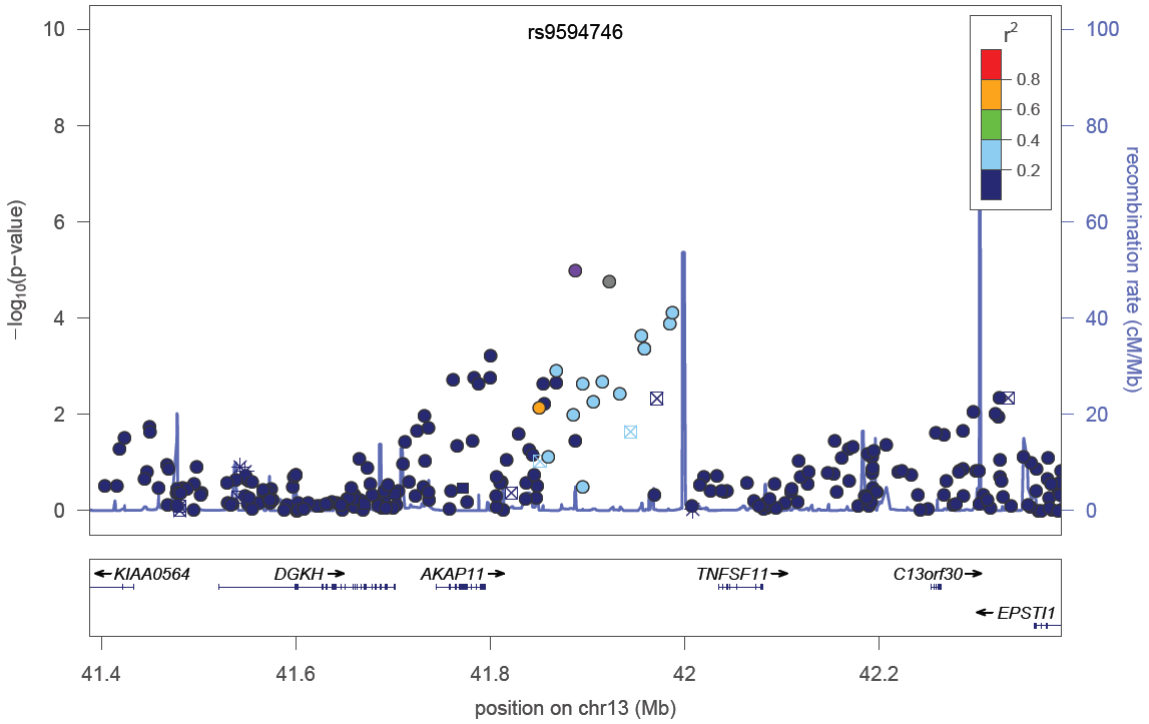


Figure A18 Association plot of the genomic region around rs9594746. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs9594746 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. KIAA0564 = KIAA0564. AKAP11 = A kinase (PRKA) anchor protein 11. TNFSF11 = tumor necrosis factor (ligand) superfamily, member 11. C13orf30 = chromosome 13 open reading frame 30. EPSTI1 = epithelial stromal interaction 1 (breast). DGKH = diacylglycerol kinase, eta.

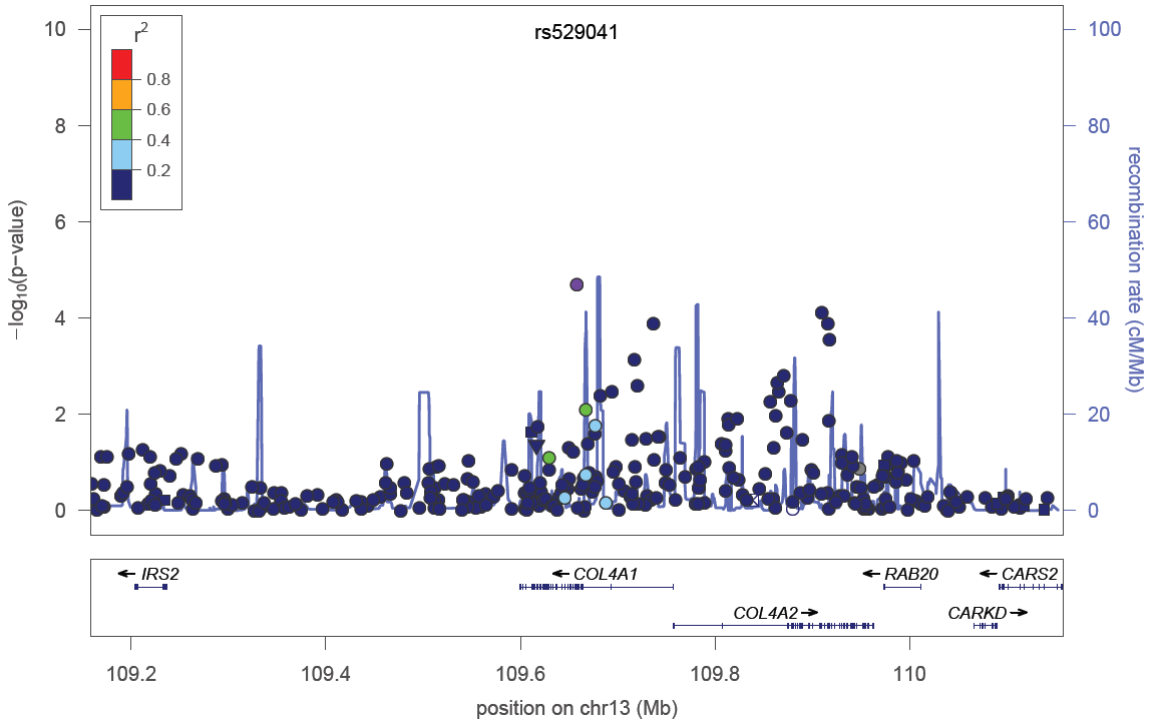


Figure A19 Association plot of the genomic region around rs529041. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs529041 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. IRS2 = insulin receptor substrate 2. COL4A1 = collagen, type IV, alpha 1. RAB20 = RAB20, member RAS oncogene family. CARS2 = cysteinyl-tRNA synthetase 2, mitochondrial (putative). COL4A2 = collagen, type IV, alpha 2. CARKD = carbohydrate kinase domain containing.

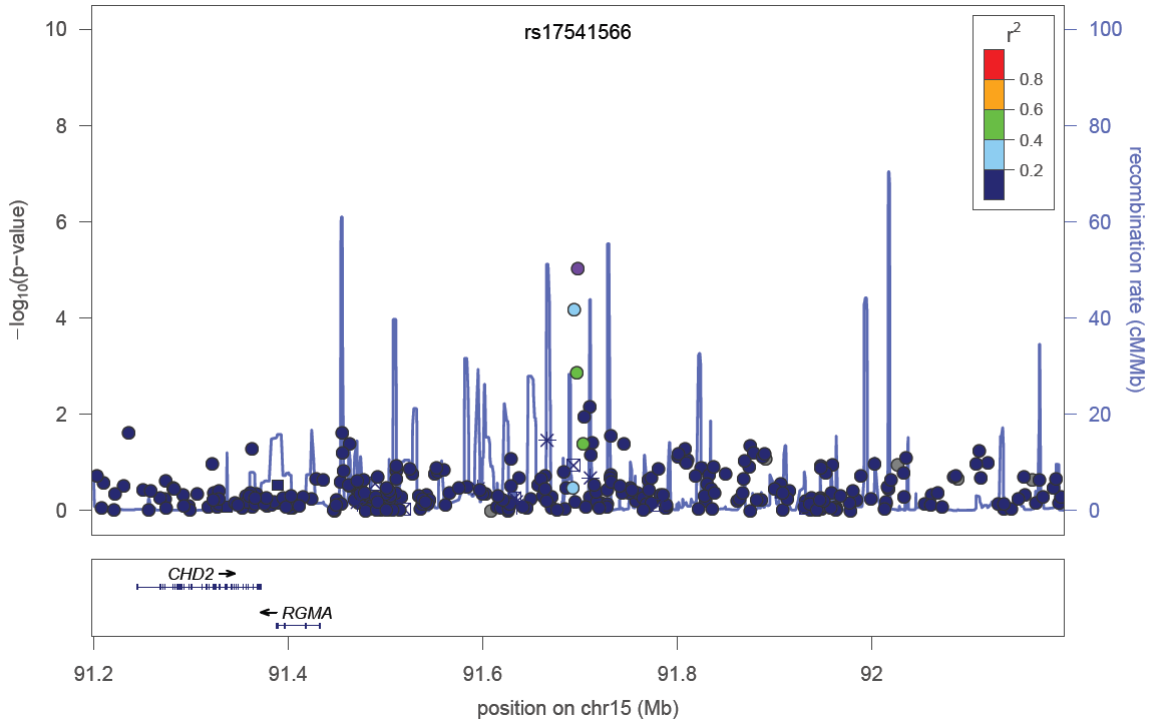


Figure A20 Association plot of the genomic region around rs17541566. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs17541566 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. CHD2 = chromodomain helicase DNA binding protein 2. RGMA = RGM domain family, member A.

Glutamate receptor, ionotropic, N-methyl D-aspartate 2A (GRIN2A)

rs9939858 is located 16kb from the *AC007221.1* gene, for which no information is available. The nearest gene that has further information is the *GRIN2A*, approximately 169kb from rs9939858 on chromosome 16p13.2 (Figure A21). A commonly used synonym is *NR2A*. The encoded protein is an N-methyl-D-aspartate (NMDA) receptor, one of a class of ionotropic glutamate-gated ion channels that have a critical role in excitatory synaptic transmission and plasticity in the central nervous system. Variants in *GRIN2A* have been linked to some aspects of memory and learning [353, 354], attention deficit hyperactivity disorder [355], schizophrenia [356, 357], depression [358, 359] and age of onset in Huntington disease [360-362]. The *GRIN2A* gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Shisa homolog 0 (Xenopus laevis) (SHISA9)

rs407146 is located in the *SHISA9* gene on chromosome 16p13.12. A commonly used synonym is *CKAMP44*. The protein encoded modulates short-term plasticity at specific excitatory synapses [363]. A variant in the intergenic region of 16p13.12 has been associated with schizophrenia by genome-wide association [364]. *SHISA9* is conserved in the chimpanzee, mouse, chicken, and zebrafish.

Protein disulfide isomerase-like, testis expressed (PDILT)

rs4496151 is located in the *PDILT* gene on chromosome 16p12.3, i.e. in the same region as *UMOD*. The genomic region is shown in Figure A22, centred on rs13333226. The protein encoded by *PDILT* is expressed solely in the testis and is thought to perform a specialised chaperone function involved in spermatogenesis [365]. *PDILT* is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

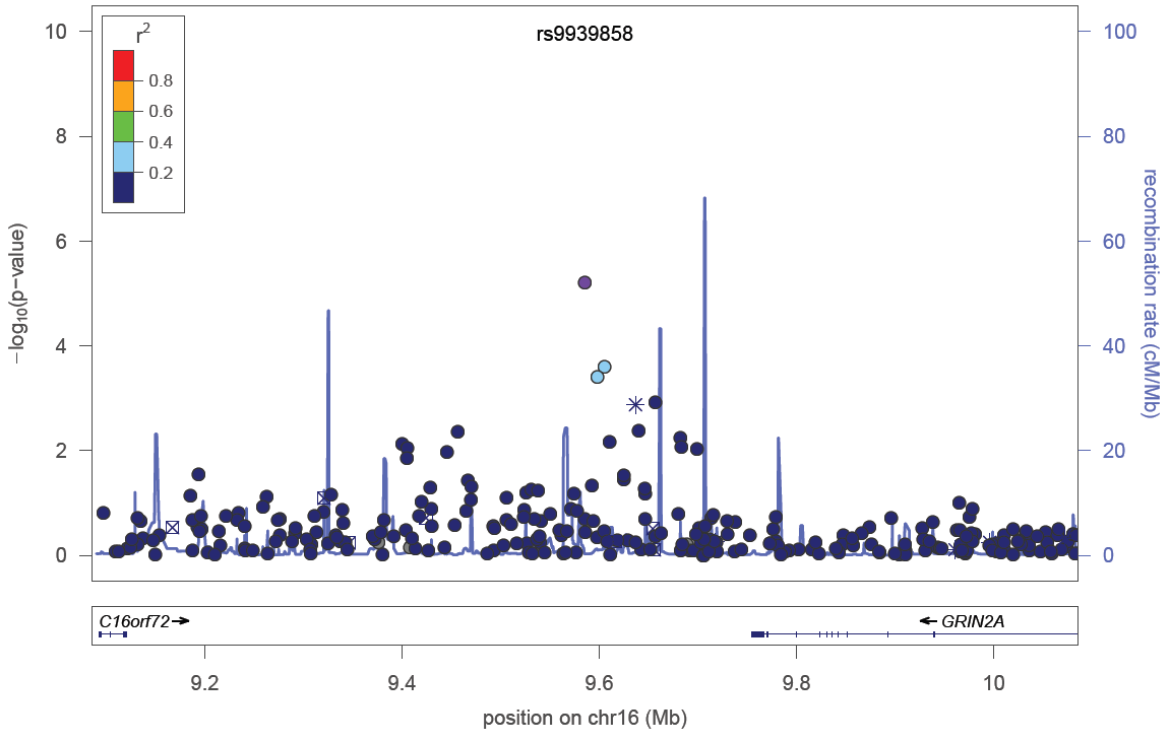


Figure A21 Association plot of the genomic region around rs9939858. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs9939858 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. C16orf72 = chromosome 16 open reading frame 72. GRIN2A = glutamate receptor, ionotropic, N-methyl D-aspartate 2A.

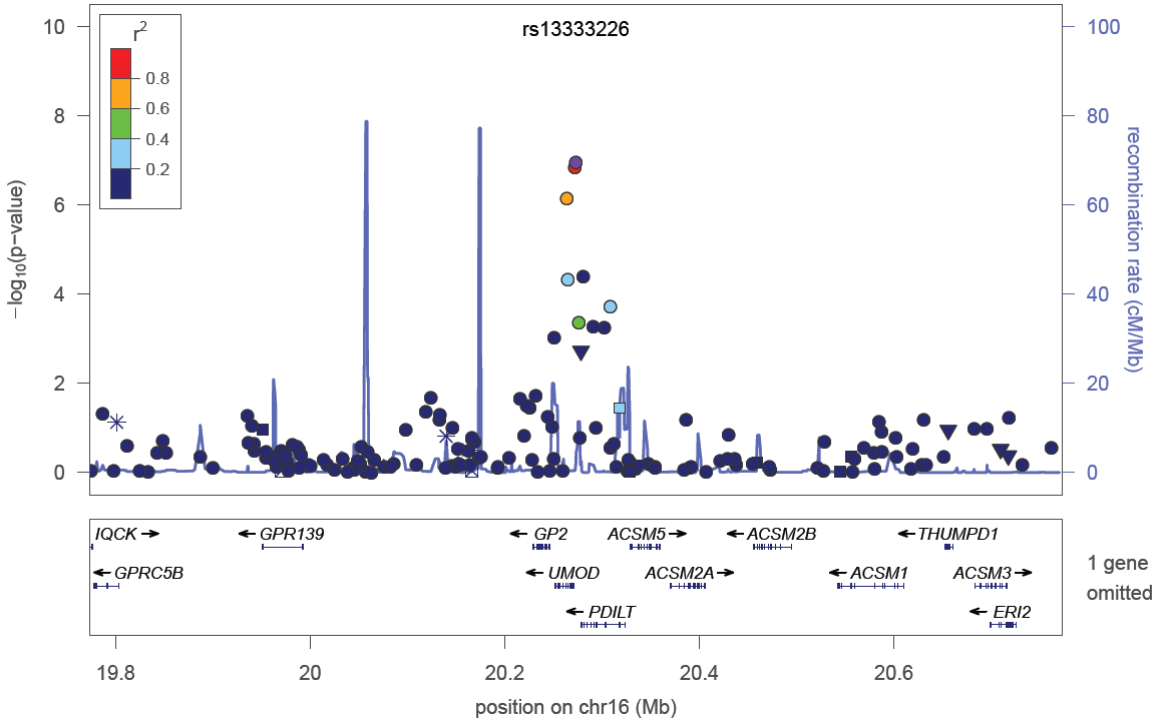


Figure A22 Association plot of the genomic region around rs13333226. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs13333226 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. IQCK = IQ motif containing K. GPR139 = G protein-coupled receptor 139. GP2 = glycoprotein 2 (zymogen granule membrane). ACSM5 = acyl-CoA synthetase medium-chain family member 5. ACSM2B = acyl-CoA synthetase medium-chain family member 2B. THUMPD1 = THUMP domain containing 1. GPRC5B = G protein-coupled receptor, family C, group 5, member B. UMOD = uromodulin. ACSM2A = acyl-CoA synthetase medium-chain family member 2A. ACSM1 = acyl-CoA synthetase medium-chain family member 1. ACSM3 = acyl-CoA synthetase medium-chain family member 3. PDILT = protein disulfide isomerase-like, testis expressed. ERI2 = ERI1 exoribonuclease family member 2.

Neural precursor cell expressed, developmentally down-regulated 4-like (NEDD4L)

The gene that rs1893469 is closest to, at a distance of 61kb, is *AC090324.1* for which there is no information available. However, rs1893469 is approximately 87kb from the *NEDD4L* gene on chromosome 18q21 (Figure A23). A commonly used synonym is *KIAA0439*. The encoded protein is a regulator of the epithelial sodium channel [246, 247] and is therefore a determinant of sodium reabsorption in the distal nephron. Of particular relevance to the current study, variants in *NEDD4L* have been associated with salt sensitivity [248, 249] and essential hypertension [250-253]. The *NEDD4L* gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, mosquito, S.pombe, S.cerevisiae, K.lactis, E.gossypii, M.grisea, and N.crassa.

Zinc finger protein 536 (ZNF536)

rs4804925 is located 39.9kb from the *ZNF536* gene on chromosome 19q12. The protein encoded by *ZNF536* is most abundant in the brain and regulates neuron differentiation [366]. In a GWAS of Framingham Heart Study data, a variant in *ZNF536* was associated with C-reactive protein [367]. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, and zebrafish.

Zinc finger proteins 230, 225, 283, 221 (ZNF230, ZNF225, ZNF283, ZNF221)

Two SNPs, rs381872 and rs383133, are intronic to zinc finger protein genes on chromosome 19q13. Figure A24 is an association plot of the genomic region around rs383133. The proteins encoded may be involved in transcriptional regulation, but there is little further information available about them or the distinctions between them. A GWAS of childhood acute lymphoblastic leukemia observed an association with a polymorphism in *ZNF230* at the level of genome-wide significance [368]. The *ZNF230* gene may also be associated with azoospermia, a condition in which there is an absence of sperm in semen [369]. The *ZNF283* gene is conserved in the chimpanzee, dog, and cow.

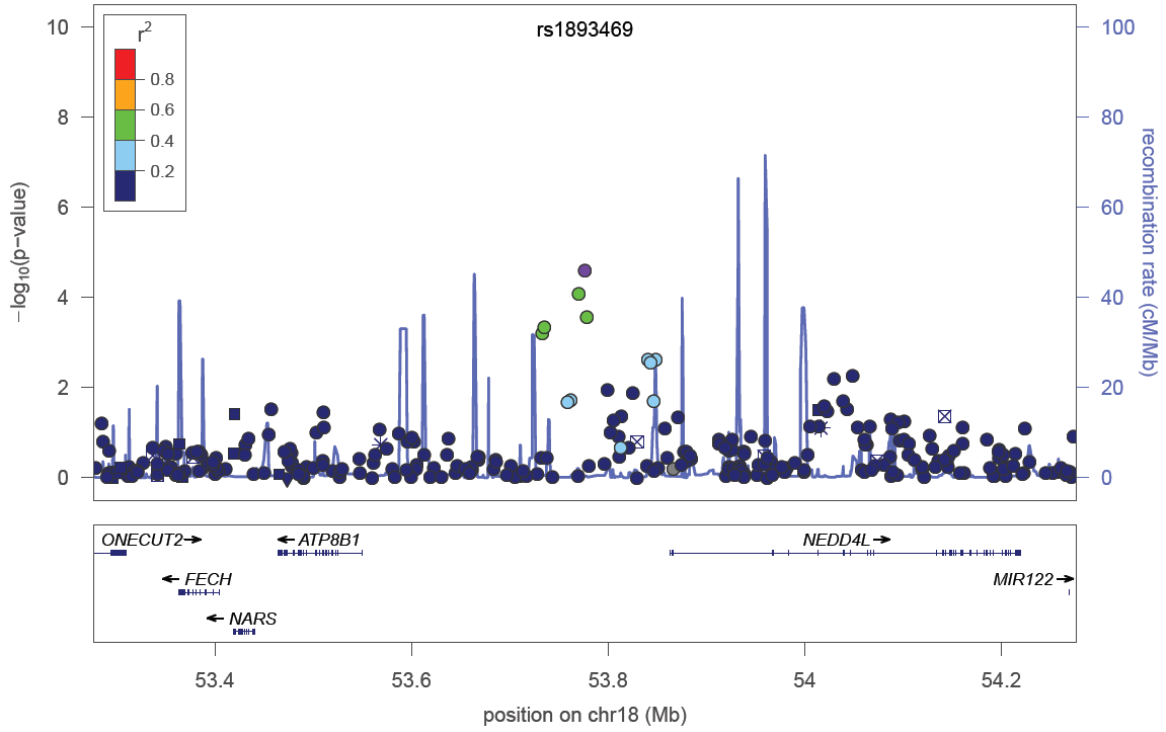


Figure A23 Association plot of the genomic region around rs1893469. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs1893469 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. ONECUT2 = one cut homeobox 2. ATP8B1 = ATPase, aminophospholipid transporter, class I, type 8B, member 1. NEDD4L = neural precursor cell expressed, developmentally down-regulated 4-like. FECH = ferrochelatase. MIR122 = microRNA 122. NARS = asparaginyl-tRNA synthetase.

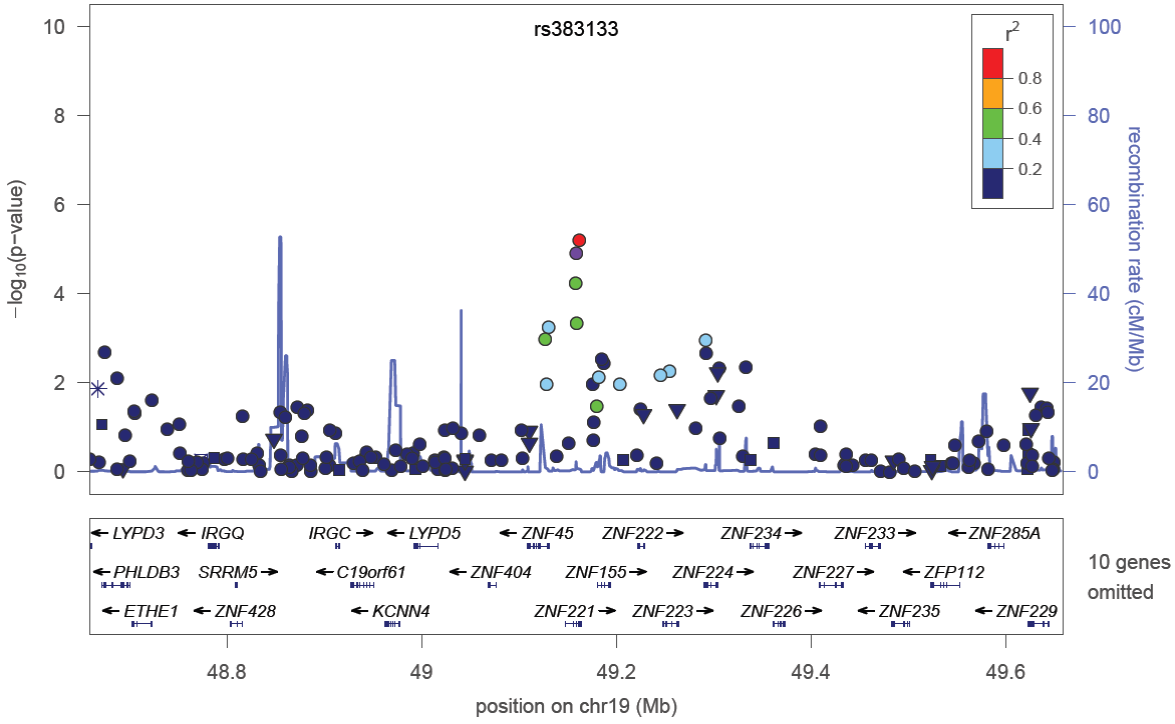


Figure A24 Association plot of the genomic region around rs383133. Showing both typed and imputed SNPs. Observed (-logP) is the –log10 transformed P values for association with hypertension status in the discovery sample. Recombination rate, represented by the blue line, is estimated from HapMap CEU samples. The level of LD between rs383133 and the surrounding SNPs, measured by r^2, is indicated by the key with red meaning high LD. The index SNP is shown in purple. Key to symbols for functional annotation: triangle = framestop or splice, inverted triangle = nonsynonymous, square = synonymous or untranslated, star = conserved transcription factor binding site, square with diagonal lines = region is highly conserved in placental mammals, circle = none-of-the-above. LYPD3 = LY6/PLAUR domain containing 3. IRGQ = immunity-related GTPase family, Q. IRGC = immunity-related GTPase family, cinema. LYPD5 = LY6/PLAUR domain containing 5. ZNF45 = zinc finger protein 45. ZNF222 = zinc finger protein 222. ZNF234 = zinc finger protein 234. ZNF233 = zinc finger protein 233. ZNF285A = zinc finger protein 285A. PHLDB3 = pleckstrin homology-like domain, family B, member 3. SRRM5 = serine/arginine repetitive matrix 5. C19orf61 = chromosome 19 open reading frame 61. ZNF404 = zinc finger protein 404. ZNF155 = zinc finger protein 155. ZNF224 = zinc finger protein 224. ZNF227 = zinc finger protein 227. ZFP112 = zinc finger protein 112 homolog (mouse). ETHE1 = ethylmalonic encephalopathy 1. ZNF428 = zinc finger protein 428. KCNN4 = potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4. ZNF221 = zinc finger protein 221. ZNF223 = zinc finger protein 223. ZNF226 = zinc finger protein 226. ZNF235 = zinc finger protein 235. ZNF229 = zinc finger protein 229.

Phospholipase C, beta 1 (phosphoinositide-specific) (PLCB1)

rs8123323 is located in the *PLCB1* gene on chromosome 20p12. The protein encoded by *PLCB1* catalyses the generation of inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG) from phosphatidylinositol 4,5-bisphosphate (IP2), necessary for the intracellular transduction of many extracellular signals. Variants in *PLCB1* have been associated with performance in cognitive tests by genome-wide association [370, 371]. The gene is conserved in the chimpanzee, dog, cow, mouse, rat, chicken, fruit fly, and mosquito.

Teashirt zinc finger homeobox 2 (TSHZ2)

rs6022204 is located in the *TSHZ2* gene on chromosome 20q13.2. The protein encoded by *TSHZ2* is a transcriptional regulator involved in developmental processes. A GWAS of erythrocyte phenotypes conducted by the CHARGE consortium observed an association between haemoglobin and a polymorphism in *TSHZ2* that reached genome-wide significance [372]. The gene is conserved in the chimpanzee, dog, cow, mouse, chicken, and zebrafish.

Reference List

(1) Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet* 1997;349(9063):1436-42.

(2) Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. *Lancet* 1997;349(9064):1498-504.

(3) Murray CJ. Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bulletin of the World Health Organization* 1994;72(3):429-45.

(4) World Health Organization. The World Health Report 2008: Primary Health Care (Now More Than Ever). 2008.

(5) World Health Organization. The World Health Report 2003: Shaping the Future. 2003.

(6) British Heart Foundation. British Heart Foundation Statistics Website. 2009. 10-12-2009.

(7) Ezzati M, Lopez AD, Rodgers A, Vander HS, Murray CJ, Comparative Risk Assessment Collaborating Group. Selected major risk factors and global and regional burden of disease. *Lancet* 2002;360(9343):1347-60.

(8) Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H. A polygenic basis for late-onset disease. *Trends in Genetics* 2003;19(2):97-106.

(9) Carretero OAM, Oparil SM. Essential Hypertension: Part I: Definition and Etiology. *Circulation* 2000;101(3):329-35.

(10) Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;360(9349):1903-13.

(11) Platt R. Heredity in hypertension. *Quarterly Journal of Medicine* 1947;16(63):111-21.

(12) Platt R. The nature of essential hypertension. *Lancet* 1959;2(7091):55-7.

(13) Oldham PD, Pickering G, Roberts JA, Sowry GS. The nature of essential hypertension. *Lancet* 1960;1(7134):1085-93.

(14) Hertz RP, Unger AN, Cornell JA, Saunders E. Racial disparities in hypertension prevalence, awareness, and management. *Archives of Internal Medicine* 2005;165(18):2098-104.

(15) Joint National Committee on Prevention DEaToHBP. The sixth report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Archives of Internal Medicine* 1997;157(21):2413-46.

(16) Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: analysis of worldwide data. *Lancet* 2005;365(9455):217-23.

(17) The INTERSALT Co-operative Research Group. Sodium, potassium, body mass, alcohol and blood pressure: the INTERSALT Study. *Journal of Hypertension* 1988;Supplement.6(4):S584-S586.

(18) Sever PS, Poulter NR. A hypothesis for the pathogenesis of essential hypertension: the initiating factors. *Journal of Hypertension* 1989;Supplement.7(1):S9-S12.

(19) Dominiczak AF, Brain N, Charchar F, McBride M, Hanlon N, Lee WK. Genetics of hypertension: lessons learnt from mendelian and polygenic syndromes. *Clinical & Experimental Hypertension (New York)* 2004;26(7-8):611-20.

(20) Chen L, Davey SG, Harbord RM, Lewis SJ. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Medicine / Public Library of Science* 2008;5(3):e52.

(21) Beeks E, Kessels AG, Kroon AA, van der Klauw MM, de Leeuw PW. Genetic predisposition to salt-sensitivity: a systematic review. *Journal of Hypertension* 2004;22(7):1243-9.

(22) Lifton RP, Gharavi AG, Geller DS. Molecular mechanisms of human hypertension. *Cell* 2001;104(4):545-56.

(23) Walley AJ, Asher JE, Froguel P. The genetic contribution to non-syndromic human obesity. *Nature Reviews Genetics* 2009;10(7):431-42.

(24) Beevers G, Lip GY, O'Brien E. ABC of hypertension: The pathophysiology of hypertension. *BMJ* 2001;322(7291):912-6.

(25) Luft FC. Twins in cardiovascular genetic research. *Hypertension* 2001;37(2 Part 2):350-6.

(26) Mongeau JG, Biron P, Sing CF. The influence of genetics and household environment upon the variability of normal blood pressure: the Montreal Adoption Survey. *Clinical & Experimental Hypertension - Part A, Theory & Practice* 1986;8(4-5):653-60.

(27) Kotchen TA, Kotchen JM, Grim CE et al. Genetic determinants of hypertension: identification of candidate phenotypes. *Hypertension* 2000;36(1):7-13.

(28) Levy D, DeStefano AL, Larson MG et al. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension* 2000;36(4):477-83.

(29) Tunstall-Pedoe H. The Dundee coronary risk-disk for management of change in risk factors. *BMJ* 1991;303(6805):744-7.

(30) Harrap SB, Stebbing M, Hopper JL, Hoang HN, Giles GG. Familial patterns of covariation for cardiovascular risk factors in adults: The Victorian Family Heart Study. *American Journal of Epidemiology* 2000;152(8):704-15.

(31) Garovic VD, Hilliard AA, Turner ST. Monogenic forms of low-renin hypertension. *Nature Clinical Practice Nephrology* 2006;2(11):624-30.

(32) Jeunemaitre X, Soubrier F, Kotelevtsev YV et al. Molecular basis of human hypertension: role of angiotensinogen. *Cell* 1992;71(1):169-80.

(33) Staessen JA, Kuznetsova T, Wang JG, Emelianov D, Vlietinck R, Fagard R. M235T angiotensinogen gene polymorphism and cardiovascular renal risk. *Journal of Hypertension* 1999;17(1):9-17.

(34) Corvol P, Persu A, Gimenez-Roqueplo AP, Jeunemaitre X. Seven lessons from two candidate genes in human essential hypertension: angiotensinogen and epithelial sodium channel. *Hypertension* 1999;33(6):1324-31.

(35) Newhouse SJ, Wallace C, Dobson R et al. Haplotypes of the WNK1 gene associate with blood pressure variation in a severely hypertensive population from the British Genetics of Hypertension study. *Human Molecular Genetics* 2005;14(13):1805-14.

(36) Tobin MD, Tomaszewski M, Braund PS et al. Common Variants in Genes Underlying Monogenic Hypertension and Hypotension and Blood Pressure in the General Population. *Hypertension* 2008;51(6):1658-64.

(37) Lifton RP, Gharavi AG, Geller DS. Molecular mechanisms of human hypertension. *Cell* 2001;104(4):545-56.

(38) Todd JA. Human genetics. Tackling common disease. *Nature* 2001;411(6837):537.

(39) Caulfield M, Munroe P, Pembroke J et al. Genome-wide mapping of human loci for essential hypertension. *Lancet* 2003;361(9375):2118-23.

(40) Wang DG, Fan JB, Siao CJ et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280(5366):1077-82.

(41) Kruglyak L, Nickerson DA. Variation is the spice of life. *Nature Genetics* 2001;27(3):234-6.

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

(43) Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005;366(9492):1223-34.

(44) Balding DJ. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 2006;7(10):781-91.

(45) Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* 2005;6(2):109-18.

(46) Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* 2002;18(1):19-24.

(47) Padmanabhan S, Melander O, Hastie C et al. Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols. *Journal of Hypertension* 2008;26(7):1275-81.

(48) Gabriel SB, Schaffner SF, Nguyen H et al. The structure of haplotype blocks in the human genome. *Science* 2002;296(5576):2225-9.

(49) Crawford DC, Carlson CS, Rieder MJ et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *American Journal of Human Genetics* 2004;74(4):610-22.

(50) Huang W, He Y, Wang H et al. Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proceedings of the National Academy of Sciences of the United States of America* 2006;103(5):1418-21.

(51) Gonzalez-Neira A, Ke X, Lao O et al. The portability of tagSNPs across populations: a worldwide survey. *Genome Research* 2006;16(3):323-30.

(52) Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* 2004;5(2):89-100.

(53) Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nature Reviews Genetics* 2006;7(11):885-91.

(54) Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genetic Epidemiology* 2007;31(7):776-88.

(55) Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661-78.

(56) Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [27/01/2010]. 2010.

(57) Donnelly P. Progress and challenges in genome-wide association studies in humans. *Nature* 2008;456(7223):728-31.

(58) Newton-Cheh C, Johnson T, Gateva V et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics* 2009;41(6):666-76.

(59) Levy D, Ehret GB, Rice K et al. Genome-wide association study of blood pressure and hypertension. *Nature Genetics* 2009;41(6):677-87.

(60) Meyre D, Delplanque J, Chevre JC et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics* 2009;41(2):157-9.

(61) Myocardial Infarction Genetics Consortium, Kathiresan S, Voight BF et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics* 2009;41(3):334-41.

(62) Lahn BT, Ebenstein L. Let's celebrate human genetic diversity. *Nature* 2009;461(7265):726-8.

(63) Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;361(9357):598-604.

(64) Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics* 1988;43(4):520-6.

(65) Gelernter J, Goldman D, Risch N. The A1 allele at the D2 dopamine receptor gene and alcoholism. A reappraisal. *JAMA* 1993;269(13):1673-7.

(66) Pato CN, Macciardi F, Pato MT, Verga M, Kennedy JL. Review of the putative association of dopamine D2 receptor and alcoholism: a meta-analysis. *American Journal of Medical Genetics* 1993;48(2):78-82.

(67) Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *Journal of the National Cancer Institute* 2000;92(14):1151-8.

(68) Freedman ML, Reich D, Penney KL et al. Assessing the impact of population stratification on genetic association studies. *Nature Genetics* 2004;36(4):388-93.

(69) Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nature Genetics* 2004;36(5):512-7.

(70) Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 1999;65(1):220-8.

(71) Hoggart CJ, Parra EJ, Shriver MD et al. Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics* 2003;72(6):1492-504.

(72) Hao K, Li C, Rosenow C, Wong WH. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. *European Journal of Human Genetics* 2004;12(12):1001-6.

(73) Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55(4):997-1004.

(74) Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 2006;38(8):904-9.

(75) Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155(2):945-59.

(76) Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *American Journal of Human Genetics* 2000;67(1):170-81.

(77) Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics* 2001;68(2):466-77.

(78) Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theoretical Population Biology* 2001;60(3):227-37.

(79) Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology* 2008;32(3):215-26.

(80) Vitart V, Rudan I, Hayward C et al. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nature Genetics* 2008;40(4):437-42.

(81) Hoffmann K, Planitz C, Ruschendorf F et al. A novel locus for arterial hypertension on chromosome 1p36 maps to a metabolic syndrome trait cluster in the Sorbs, a Slavic population isolate in Germany. *Journal of Hypertension* 2009;27(5):983-90.

(82) Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 1993;52(3):506-16.

(83) Frayling TM, Walker M, McCarthy MI et al. Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* 1999;48(12):2475-9.

(84) Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 2005;6(2):95-108.

(85) Schaid DJ. Transmission disequilibrium, family controls, and great expectations. *American Journal of Human Genetics* 1998;63(4):935-41.

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

(87) Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology & Evolution* 1999;16(12):1791-8.

(88) Reich DE, Goldstein DB. Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95(14):8119-23.

(89) Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* 2001;69(1):124-37.

(90) Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. *Genome Biology* 2001;2(8):COMMENT2007.

(91) Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* 2003;33(2):177-82.

(92) Hugot JP, Chamaillard M, Zouali H et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;411(6837):599-603.

(93) Ogura Y, Bonen DK, Inohara N et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;411(6837):603-6.

(94) Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004;305(5685):869-72.

(95) Corder EH, Saunders AM, Strittmatter WJ et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993;261(5123):921-3.

(96) International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437(7063):1299-320.

(97) International HapMap Consortium, Frazer KA, Ballinger DG et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449(7164):851-61.

(98) Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute* 2004;96(6):434-42.

(99) Newton-Cheh C, Hirschhorn JN. Genetic association studies of complex traits: design and analysis issues. *Mutation Research* 2005;573(1-2):54-69.

(100) Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nature Reviews Genetics* 2006;7(10):812-20.

(101) UK Biobank. 2009.

(102) Generation Scotland. 2009.

(103) Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003;361(9357):567-71.

(104) Roses AD. A model for susceptibility polymorphisms for complex diseases: apolipoprotein E and Alzheimer disease. *Neurogenetics* 1997;1(1):3-11.

(105) Paterson AH, Damon S, Hewitt JD et al. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics* 1991;127(1):181-97.

(106) Mackay TF, Lyman RF, Jackson MS. Effects of P element insertions on quantitative traits in Drosophila melanogaster. *Genetics* 1992;130(2):315-32.

(107) Risch N, Ghosh S, Todd JA. Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 1993;53(3):702-14.

(108) Vyse TJ, Todd JA. Genetic analysis of autoimmune disease. *Cell* 1996;85(3):311-8.

(109) Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* 2001;33(3):209-29.

(110) Barton NH, Keightley PD. Understanding quantitative genetic variation. *Nature Reviews Genetics* 2002;3(1):11-21.

(111) Nezer C, Collette C, Moreau L et al. Haplotype sharing refines the location of an imprinted quantitative trait locus with major effect on muscle mass to a 250-kb

chromosome segment containing the porcine IGF2 gene. *Genetics* 2003;165(1):277-85.

(112) Farrall M. Quantitative genetic variation: a post-modern view. *Human Molecular Genetics* 2004;13:Spec-7.

(113) Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nature Genetics* 2001;29(3):306-9.

(114) Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* 2001;69(6):1357-69.

(115) Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics* 2007;80(4):605-15.

(116) Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology* 2009;33(5):453-62.

(117) NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T et al. Replicating genotype-phenotype associations. *Nature* 2007;447(7145):655-60.

(118) Evangelou E, Maraganore DM, Ioannidis JP. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE* 2007;2(2):e196.

(119) Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* 2007;2(9):e841.

(120) Saxena R, Voight BF, Lyssenko V et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316(5829):1331-6.

(121) Levy D, Larson MG, Benjamin EJ et al. Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Medical Genetics* 2007;8(Suppl 1):S3.

(122) Kato N, Miyata T, Tabara Y et al. High-density association study and nomination of susceptibility genes for hypertension in the Japanese National Project. *Human Molecular Genetics* 2008;17(4):617-27.

(123) Sabatti C, Service SK, Hartikainen AL et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* 2009;41(1):35-46.

(124) Wang Y, O'Connell JR, McArdle PF et al. Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106(1):226-31.

(125) Org E, Eyheramendy S, Juhanson P et al. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Human Molecular Genetics* 2009;18(12):2288-96.

(126) Cho YS, Go MJ, Kim YJ et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics* 2009;41(5):527-34.

(127) Adeyemo A, Gerry N, Chen G et al. A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genetics* 2009;5(7):e1000564.

(128) Caulfield M, Munroe P, Pembroke J et al. Genome-wide mapping of human loci for essential hypertension. *Lancet* 2003;361(9375):2118-23.

(129) Bowcock AM. Genomics: guilt by association. *Nature* 2007;447(7145):645-6.

(130) McCarthy MI, Abecasis GR, Cardon LR et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 2008;9(5):356-69.

(131) Delles C, McBride MW, Padmanabhan S, Dominiczak AF. The genetics of cardiovascular disease. *Trends in Endocrinology & Metabolism* 2008;9:309-16.

(132) Costa-Santos M, Kater CE, Auchus RJ. Two prevalent CYP17 mutations and genotype-phenotype correlations in 24 Brazilian patients with 17-hydroxylase deficiency. *J Clin Endocrinol Metab* 2004;89(1):49-60.

(133) Newton-Cheh C, Larson MG, Vasan RS et al. Association of common variants in NPPA and NPPB with circulating natriuretic peptides and blood pressure. *Nat Genet* 2009;41(3):348-53.

(134) Monteith GR, Kable EP, Kuo TH, Roufogalis BD. Elevated plasma membrane and sarcoplasmic reticulum Ca2+ pump mRNA levels in cultured aortic smooth muscle cells from spontaneously hypertensive rats. *Biochem Biophys Res Commun* 1997;230(2):344-6.

(135) Vatner SF. FGF induces hypertrophy and angiogenesis in hibernating myocardium. *Circ Res* 2005;96(7):705-7.

(136) Todd JA, Walker NM, Cooper JD et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007;39(7):857-64.

(137) Smyth DJ, Plagnol V, Walker NM et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N Engl J Med* 2008;359(26):2767-77.

(138) Hunt KA, Zhernakova A, Turner G et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008;40(4):395-402.

(139) Gudbjartsson DF, Bjornsdottir US, Halapi E et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat Genet* 2009;41(3):342-7.

(140) NORDIL Study Group. The Nordic Diltiazem Study (NORDIL). A prospective intervention trial of calcium antagonist therapy in hypertension. *Blood Pressure* 1993;2(4):312-21.

(141) Hansson L, Hedner T, Lund-Johansen P et al. Randomised trial of effects of calcium antagonists compared with diuretics and beta-blockers on cardiovascular morbidity and mortality in hypertension: the Nordic Diltiazem (NORDIL) study. *Lancet* 2000;356(9227):359-65.

(142) Berglund G, Eriksson KF, Israelsson B et al. Cardiovascular risk groups and mortality in an urban swedish male population: the Malmo Preventive Project. *Journal of Internal Medicine* 1996;239(6):489-97.

(143) Berglund G, Nilsson P, Eriksson KF et al. Long-term outcome of the Malmo preventive project: mortality and cardiovascular morbidity. *Journal of Internal Medicine* 2000;247(1):19-29.

(144) Manjer J, Kaaks R, Riboli E, Berglund G. Risk of breast cancer in relation to anthropometry, blood pressure, blood lipids and glucose metabolism: a prospective study within the Malmo Preventive Project. *European Journal of Cancer Prevention* 2001;10(1):33-42.

(145) Berglund G, Elmstahl S, Janzon L, Larsson SA. The Malmo Diet and Cancer Study. Design and feasibility. *Journal of Internal Medicine* 1993;233(1):45-51.

(146) SPSS for Windows, Release 15.0.0 [computer program]. Chicago: SPSS Inc.; 2006.

(147) Stata Statistical Software: Release 10 [computer program]. College Station, TX: StataCorp LP; 2007.

(148) Illumina Inc. 2010.

(149) Purcell S, Neale B, Todd-Brown K et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 2007;81(3):559-75.

(150) PLINK version 1.05 [computer program]. Harvard, MA: 2008.

(151) Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21(2):263-5.

(152) R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2009.

(153) Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews Genetics* 2009;10(9):639-50.

(154) Price AL, Butler J, Patterson N et al. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genetics* 2008;4(1):e236.

(155) Nelis M, Esko T, Magi R et al. Genetic structure of Europeans: a view from the North-East. *PLoS ONE [Electronic Resource]* 2009;4(5):e5472.

(156) Heath SC, Gut IG, Brennan P et al. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008;16(12):1413-29.

(157) Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med* 1999;130(6):461-70.

(158) Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods* 2006;11(2):193-206.

(159) Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54(10):1046-55.

(160) Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;53(11):1119-29.

(161) WGAViewer: Package of Whole Genome Association Annotation. 1.10 edn [computer program]. Durham, NC: 2007.

(162) Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002;30(1):52-5.

(163) Online Mendelian Inheritance in Man, OMIM (TM). 2010. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

(164) Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;35(Database issue):D26-D31.

(165) Entrez Gene. 2010.

(166) Safran M, Solomon I, Shmueli O et al. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* 2002;18(11):1542-3.

(167) GeneCards Version 3. 2010.

(168) Entrez. 2010.

(169) McKusick VA. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. 12th ed. Baltimore, MD: Johns Hopkins University Press; 1998.

(170) Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* 1997;13(4):163.

(171) Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 1998;14(8):656-64.

(172) Yu W, Yesupriya A, Wulf A, Qu J, Khoury MJ, Gwinn M. An open source infrastructure for managing knowledge and finding potential collaborators in a domain-specific subset of PubMed, with an example from human genome epidemiology. *BMC Bioinformatics* 2007;8:436.

(173) Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet* 2008;40(2):124-5.

(174) Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* 2010;26(1):145-6.

(175) Mailman MD, Feolo M, Jin Y et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;39(10):1181-6.

(176) Bochud M, Bovet P, Vollenweider P et al. Association between white-coat effect and blunted dipping of nocturnal blood pressure. *Am J Hypertens* 2009;22(10):1054-61.

(177) Visser FW, Muntinga JH, Dierckx RA, Navis G. Feasibility and impact of the measurement of extracellular fluid volume simultaneous with GFR by 125I-iothalamate. *Clin J Am Soc Nephrol* 2008;3(5):1308-15.

(178) Visser FW, Boonstra AH, Titia LA, Boomsma F, Navis G. Renal response to angiotensin II is blunted in sodium-sensitive normotensive men. *Am J Hypertens* 2008;21(3):323-8.

(179) Levey AS, Coresh J, Balk E et al. National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Ann Intern Med* 2003;139(2):137-47.

(180) Munroe PB, Johnson T, Caulfield MJ. The genetic architecture of blood pressure variation. *Current Cardiovascular Risk Reports* 2009;3:418-25.

(181) Tenesa A, Visscher PM, Carothers AD, Knott SA. Mapping quantitative trait loci using linkage disequilibrium: marker- versus trait-based methods. *Behav Genet* 2005;35(2):219-28.

(182) Plomin R, Haworth CM, Davis OS. Common disorders are quantitative traits. *Nat Rev Genet* 2009;10(12):872-8.

(183) Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 1918;52:399-433.

(184) Zhang K, Weder AB, Eskin E, O'Connor DT. Genome-wide case/control studies in hypertension: only the 'tip of the iceberg'. *J Hypertens* 2010;28(6):1115-23.

(185) Weedon MN, Lango H, Lindgren CM et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 2008;40(5):575-83.

(186) Willer CJ, Speliotes EK, Loos RJ et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 2009;41(1):25-34.

(187) Yang J, Benyamin B, McEvoy BP et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;42(7):565-9.

(188) Gibson G. Hints of hidden heritability in GWAS. *Nat Genet* 2010;42(7):558-60.

(189) Park JH, Wacholder S, Gail MH et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 2010;42(7):570-5.

(190) Burton PR, Hansell AL, Fortier I et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;38(1):263-73.

(191) Dyer AR, Shipley M, Elliott P. Urinary electrolyte excretion in 24 hours and blood pressure in the INTERSALT Study. I. Estimates of reliability. The INTERSALT Cooperative Research Group. *Am J Epidemiol* 1994;139(9):927-39.

(192) Stergiou GS, Baibas NM, Gantzarou AP et al. Reproducibility of home, ambulatory, and clinic blood pressure: implications for the design of trials for the assessment of antihypertensive drug efficacy. *Am J Hypertens* 2002;15(2 Pt 1):101-4.

(193) Spence JD, Barnett PA, Bulman DE, Hegele RA. An approach to ascertain probands with a non-traditional risk factor for carotid atherosclerosis. *Atherosclerosis* 1999;144(2):429-34.

(194) Lanktree MB, Hegele RA, Schork NJ, Spence JD. Extremes of Unexplained Variation as a Phenotype: An Efficient Approach for Genome-Wide Association Studies of Cardiovascular Disease. *Circulation: Cardiovascular Genetics* 2010;3(2):215-21.

(195) Affymetrix Inc. 2010.

(196) Suarez BK, Taylor C, Bertelsen S et al. An analysis of identical single-nucleotide polymorphisms genotyped by two different platforms. *BMC Genetics* 2005;6(Suppl 1):S152.

(197) Kim KK, Won HH, Cho SS et al. Comparison of identical single nucleotide polymorphisms genotyped by the GeneChip Targeted Genotyping 25K, Affymetrix 500K and Illumina 550K platforms. *Genomics* 2009;94(2):89-93.

(198) Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. *Nature Genetics* 2006;38(6):659-62.

(199) Wollstein A, Herrmann A, Wittig M et al. Efficacy assessment of SNP sets for genome-wide disease association studies. *Nucleic Acids Research* 2007;35(17):e113.

(200) Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M. Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics* 2007;8:159.

(201) Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *European Journal of Human Genetics* 2008;16(5):635-43.

(202) Kottgen A, Glazer NL, Dehghan A et al. Multiple loci associated with indices of renal function and chronic kidney disease. *Nature Genetics* 2009;41(6):712-7.

(203) The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. WHO

MONICA Project Principal Investigators. *Journal of Clinical Epidemiology* 1988;41(2):105-14.

(204) Tunstall-Pedoe H, Kuulasmaa K, Mahonen M, Tolonen H, Ruokokoski E, Amouyel P. Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. *Lancet* 1999;353(9164):1547-57.

(205) Mancia G, Sega R, Bravi C et al. Ambulatory blood pressure normality: results from the PAMELA study. *J Hypertens* 1995;13(12 Pt 1):1377-90.

(206) Licht CM, de Geus EJ, Seldenrijk A et al. Depression is associated with decreased blood pressure, but antidepressant use increases the risk for hypertension. *Hypertension* 2009;53(4):631-8.

(207) Sever PS, Dahlof B, Poulter NR et al. Rationale, design, methods and baseline demography of participants of the Anglo-Scandinavian Cardiac Outcomes Trial. ASCOT investigators. *J Hypertens* 2001;19(6):1139-47.

(208) Pinto-Sietsma SJ, Janssen WM, Hillege HL, Navis G, De ZD, de Jong PE. Urinary albumin excretion is associated with renal functional abnormalities in a nondiabetic population. *J Am Soc Nephrol* 2000;11(10):1882-8.

(209) Hillege HL, Fidler V, Diercks GF et al. Urinary albumin excretion predicts cardiovascular and noncardiovascular mortality in general population. *Circulation* 2002;106(14):1777-82.

(210) Firmann M, Mayor V, Vidal PM et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc Disord* 2008;8:6.

(211) Wichmann HE, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* 2005;67 Suppl 1:S26-S30.

(212) Heid IM, Vollmert C, Hinney A et al. Association of the 103I MC4R allele with decreased body mass in 7937 participants of two population based surveys. *J Med Genet* 2005;42(4):e21.

(213) John U, Greiner B, Hensel E et al. Study of Health In Pomerania (SHIP): a health examination survey in an east German region: objectives and design. *Soz Praventivmed* 2001;46(3):186-94.

(214) Iwai N, Kajimoto K, Kokubo Y, Tomoike H. Extensive genetic analysis of 10 candidate genes for hypertension in Japanese. *Hypertension* 2006;48(5):901-7.

(215) Mannami T, Konishi M, Baba S, Nishi N, Terao A. Prevalence of asymptomatic carotid atherosclerotic lesions detected by high-resolution ultrasonography and its relation to cardiovascular risk factors in the general population of a Japanese city: the Suita study. *Stroke* 1997;28(3):518-25.

(216) Stevens LA, Coresh J, Greene T, Levey AS. Assessing kidney function--measured and estimated glomerular filtration rate. *N Engl J Med* 2006;354(23):2473-83.

(217) Turner JJ, Stacey JM, Harding B et al. UROMODULIN mutations cause familial juvenile hyperuricemic nephropathy. *J Clin Endocrinol Metab* 2003;88(3):1398-401.

(218) Rampoldi L, Caridi G, Santon D et al. Allelism of MCKD, FJHN and GCKD caused by impairment of uromodulin export dynamics. *Hum Mol Genet* 2003;12(24):3369-84.

(219) Hart TC, Gorry MC, Hart PS et al. Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. *J Med Genet* 2002;39(12):882-92.

(220) Devuyst O, Dahan K, Pirson Y. Tamm-Horsfall protein or uromodulin: new ideas about an old molecule. *Nephrol Dial Transplant* 2005;20(7):1290-4.

(221) Vylet'al P, Kublova M, Kalbacova M et al. Alterations of uromodulin biology: a common denominator of the genetically heterogeneous FJHN/MCKD syndrome. *Kidney Int* 2006;70(6):1155-69.

(222) Tamm I, Horsfall FL, Jr. Characterization and separation of an inhibitor of viral hemagglutination present in urine. *Proc Soc Exp Biol Med* 1950;74(1):106-8.

(223) Bachmann S, Metzger R, Bunnemann B. Tamm-Horsfall protein-mRNA synthesis is localized to the thick ascending limb of Henle's loop in rat kidney. *Histochemistry* 1990;94(5):517-23.

(224) Malagolini N, Cavallone D, Serafini-Cessi F. Intracellular transport, cell-surface exposure and release of recombinant Tamm-Horsfall glycoprotein. *Kidney Int* 1997;52(5):1340-50.

(225) Zurbig P, Decramer S, Dakna M et al. The human urinary proteome reveals high similarity between kidney aging and chronic kidney disease. *Proteomics* 2009;9(8):2108-17.

(226) Torffvit O, Jorgensen PE, Kamper AL et al. Urinary excretion of Tamm-Horsfall protein and epidermal growth factor in chronic nephropathy. *Nephron* 1998;79(2):167-72.

(227) Dulawa J, Kokot F, Kokot M, Pander H. Urinary excretion of Tamm-Horsfall protein in normotensive and hypertensive elderly patients. *J Hum Hypertens* 1998;12(9):635-7.

(228) Torffvit O, Agardh CD, Thulin T. A study of Tamm-Horsfall protein excretion in hypertensive patients and type 1 diabetic patients. *Scand J Urol Nephrol* 1999;33(3):187-91.

(229) Lifton RP. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lectures* 2004;100:71-101.

(230) Dahan K, Devuyst O, Smaers M et al. A cluster of mutations in the UMOD gene causes familial juvenile hyperuricemic nephropathy with abnormal expression of uromodulin. *J Am Soc Nephrol* 2003;14(11):2883-93.

(231) Bleyer AJ, Hart TC, Shihabi Z, Robins V, Hoyer JR. Mutations in the uromodulin gene decrease urinary excretion of Tamm-Horsfall protein. *Kidney Int* 2004;66(3):974-7.

(232) Mo L, Zhu XH, Huang HY, Shapiro E, Hasty DL, Wu XR. Ablation of the Tamm-Horsfall protein gene increases susceptibility of mice to bladder colonization by type 1-fimbriated Escherichia coli. *Am J Physiol Renal Physiol* 2004;286(4):F795-F802.

(233) Bates JM, Raffi HM, Prasadan K et al. Tamm-Horsfall protein knockout mice are more prone to urinary tract infection: rapid communication. *Kidney Int* 2004;65(3):791-7.

(234) Mo L, Huang HY, Zhu XH, Shapiro E, Hasty DL, Wu XR. Tamm-Horsfall protein is a critical renal defense factor protecting against calcium oxalate crystal formation. *Kidney Int* 2004;66(3):1159-66.

(235) Lynn KL, Marshall RD. Excretion of Tamm-Horsfall glycoprotein in renal disease. *Clin Nephrol* 1984;22(5):253-7.

(236) Tsai CY, Wu TH, Yu CL, Lu JY, Tsai YY. Increased excretions of beta2-microglobulin, IL-6, and IL-8 and decreased excretion of Tamm-Horsfall glycoprotein in urine of patients with active lupus nephritis. *Nephron* 2000;85(3):207-14.

(237) Bachmann S, Mutig K, Bates J et al. Renal effects of Tamm-Horsfall protein (uromodulin) deficiency in mice. *Am J Physiol Renal Physiol* 2005;288(3):F559-F567.

(238) Schmitt R, Kahl T, Mutig K, Bachmann S. Selectively reduced expression of thick ascending limb Tamm-Horsfall protein in hypothyroid kidneys. *Histochem Cell Biol* 2004;121(4):319-27.

(239) Ying WZ, Sanders PW. Dietary salt regulates expression of Tamm-Horsfall glycoprotein in rats. *Kidney Int* 1998;54(4):1150-6.

(240) Whelton PK, Klag MJ. Hypertension as a risk factor for renal disease. Review of clinical and epidemiological evidence. *Hypertension* 1989;13(5 Suppl):I19-I27.

(241) Klag MJ, Whelton PK, Randall BL et al. Blood pressure and end-stage renal disease in men. *N Engl J Med* 1996;334(1):13-8.

(242) Hsu CY, McCulloch CE, Darbinian J, Go AS, Iribarren C. Elevated blood pressure and risk of end-stage renal disease in subjects without baseline kidney disease. *Arch Intern Med* 2005;165(8):923-8.

(243) Crews DC, Plantinga LC, Miller ER, III et al. Prevalence of chronic kidney disease in persons with undiagnosed or prehypertension in the United States. *Hypertension* 2010;55(5):1102-9.

(244) Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP. Discovery properties of genome-wide association signals from cumulatively combined data sets. *American Journal of Epidemiology* 2009;170(10):1197-206.

(245) Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327(7414):557-60.

(246) Harvey KF, Dinudom A, Cook DI, Kumar S. The Nedd4-like protein KIAA0439 is a potential regulator of the epithelial sodium channel. *J Biol Chem* 2001;276(11):8597-601.

(247) Raikwar NS, Thomas CP. Nedd4-2 isoforms ubiquitinate individual epithelial sodium channel subunits and reduce surface expression and function of the epithelial sodium channel. *Am J Physiol Renal Physiol* 2008;294(5):F1157-F1165.

(248) Dahlberg J, Nilsson LO, von WF, Melander O. Polymorphism in NEDD4L is associated with increased salt sensitivity, reduced levels of P-renin and increased levels of Nt-proANP. *PLoS ONE* 2007;2(5):e432.

(249) Manunta P, Lavery G, Lanzani C et al. Physiological interaction between alpha-adducin and WNK1-NEDD4L pathways on sodium-related blood pressure regulation. *Hypertension* 2008;52(2):366-72.

(250) Russo CJ, Melista E, Cui J et al. Association of NEDD4L ubiquitin ligase with essential hypertension. *Hypertension* 2005;46(3):488-91.

(251) Wen H, Lin R, Jiao Y et al. Two polymorphisms in NEDD4L gene and essential hypertension in Chinese Hans - a population-based case-control study. *Clin Exp Hypertens* 2008;30(2):87-94.

(252) Luo F, Wang Y, Wang X, Sun K, Zhou X, Hui R. A functional variant of NEDD4L is associated with hypertension, antihypertensive response, and orthostatic hypotension. *Hypertension* 2009;54(4):796-801.

(253) Li N, Wang H, Yang J et al. Genetic variation of NEDD4L is associated with essential hypertension in female Kazakh general population: a case-control study. *BMC Med Genet* 2009;10:130.

(254) Kato K, Oguri M, Kato N et al. Assessment of genetic risk factors for thoracic aortic aneurysm in hypertensive patients. *Am J Hypertens* 2008;21(9):1023-7.

(255) Boekholdt SM, Trip MD, Peters RJ et al. Thrombospondin-2 polymorphism is associated with a reduced risk of premature myocardial infarction. *Arterioscler Thromb Vasc Biol* 2002;22(12):e24-e27.

(256) Tarasov KV, Sanna S, Scuteri A et al. COL4A1 is associated with arterial stiffness by genome-wide association scan. *Circ Cardiovasc Genet* 2009;2(2):151-8.

(257) Plaisier E, Gribouval O, Alamowitch S et al. COL4A1 mutations and hereditary angiopathy, nephropathy, aneurysms, and muscle cramps. *N Engl J Med* 2007;357(26):2687-95.

(258) Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. *N Engl J Med* 2010;363(2):166-76.

(259) Kraft P, Wacholder S, Cornelis MC et al. Beyond odds ratios--communicating disease risk based on genetic profiles. *Nat Rev Genet* 2009;10(4):264-9.

(260) Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;150(11):795-802.

(261) Zheng SL, Sun J, Wiklund F et al. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 2008;358(9):910-9.

(262) Kathiresan S, Melander O, Anevski D et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* 2008;358(12):1240-9.

(263) Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med* 2006;355(25):2615-7.

(264) Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 2009;18(18):3525-31.

(265) Manolio TA, Collins FS, Cox NJ et al. Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-53.

(266) Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics* 2009;10(5):318-29.

(267) Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010;11(5):356-66.

(268) Choi M, Scholl UI, Ji W et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009;106(45):19096-101.

(269) Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;11(6):415-25.

(270) Bochukova EG, Huang N, Keogh J et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010;463(7281):666-70.

(271) Walters RG, Jacquemont S, Valsesia A et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* 2010;463(7281):671-5.

(272) Roach JC, Glusman G, Smit AF et al. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* 2010.

(273) Lupski JR, Reid JG, Gonzaga-Jauregui C et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010;362(13):1181-91.

(274) Kong A, Steinthorsdottir V, Masson G et al. Parental origin of sequence variants associated with complex diseases. *Nature* 2009;462(7275):868-74.

(275) Durbin R, Altshuler D, Brooks LD, Felsenfeld A, McEwen J. 1000 Genomes Project: A Deep Catalog of Human Genetic Variation. Available at: www.1000genomes.org. Accessed [27/01/2010]. 2010.

(276) Brunetti-Pierri N, Berg JS, Scaglia F et al. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 2008;40(12):1466-71.

(277) Franco LM, de RT, Graham BH et al. A syndrome of short stature, microcephaly and speech delay is associated with duplications reciprocal to the common Sotos syndrome deletion. *Eur J Hum Genet* 2010;18(2):258-61.

(278) Merla G, Brunetti-Pierri N, Micale L, Fusco C. Copy number variants at Williams-Beuren syndrome 7q11.23 region. *Hum Genet* 2010;128(1):3-26.

(279) Sebat J, Lakshmi B, Malhotra D et al. Strong association of de novo copy number mutations with autism. *Science* 2007;316(5823):445-9.

(280) Stefansson H, Rujescu D, Cichon S et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;455(7210):232-6.

(281) The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008;455(7210):237-41.

(282) McCarroll SA, Huett A, Kuballa P et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* 2008;40(9):1107-12.

(283) Diskin SJ, Hou C, Glessner JT et al. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 2009;459(7249):987-91.

(284) Craddock N, Hurles ME, Cardin N et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010;464(7289):713-20.

(285) Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8(1):e1000294.

(286) Sanna S, Jackson AU, Nagaraja R et al. Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 2008;40(2):198-203.

(287) Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009;324(5925):387-9.

(288) Fellay J, Thompson AJ, Ge D et al. ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature* 2010;464(7287):405-8.

(289) Ji W, Foo JN, O'Roak BJ et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics* 2008;40(5):592-9.

(290) Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009;10(3):184-94.

(291) Schadt EE, Monks SA, Drake TA et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422(6929):297-302.

(292) Dixon AL, Liang L, Moffatt MF et al. A genome-wide association study of global gene expression. *Nat Genet* 2007;39(10):1202-7.

(293) Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nature Reviews Genetics* 2008;9(4):255-66.

(294) Shimada MK, Matsumoto R, Hayakawa Y et al. VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res* 2009;37(Database issue):D810-D815.

(295) Moffatt MF, Kabesch M, Liang L et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007;448(7152):470-3.

(296) Barrett JC, Hansoul S, Nicolae DL et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008;40(8):955-62.

(297) Schadt EE, Molony C, Chudin E et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008;6(5):e107.

(298) Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;6(4):e1000888.

(299) Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. *Nat Genet* 2007;39(7):807-8.

(300) Wold B, Myers RM. Sequence census methods for functional genomics. *Nat Methods* 2008;5(1):19-21.

(301) Reis BY, Butte AS, Kohane IS. Extracting knowledge from dynamics in gene expression. *J Biomed Inform* 2001;34(1):15-27.

(302) Schadt EE, Lum PY. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *J Lipid Res* 2006;47(12):2601-13.

(303) Genotype-Tissue Expression project. 2010.

(304) Schneider GM, Jacobs DW, Gevirtz RN, O'Connor DT. Cardiovascular haemodynamic response to repeated mental stress in normotensive subjects at genetic risk of hypertension: evidence of enhanced reactivity, blunted adaptation, and delayed recovery. *J Hum Hypertens* 2003;17(12):829-40.

(305) Hardy J, Singleton A. Genomewide association studies and human disease. *New England Journal of Medicine* 2009;360(17):1759-68.

(306) Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008;92(5):265-72.

(307) Cao MY, Davidson D, Yu J, Latour S, Veillette A. Clnk, a novel SLP-76-related adaptor molecule expressed in cytokine-stimulated hemopoietic cells. *J Exp Med* 1999;190(10):1527-34.

(308) Utting O, Sedgmen BJ, Watts TH et al. Immune functions in mice lacking Clnk, an SLP-76-related adaptor expressed in a subset of immune cells. *Mol Cell Biol* 2004;24(13):6067-75.

(309) Jacobs S, Ruusuvuori E, Sipila ST et al. Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. *Proc Natl Acad Sci U S A* 2008;105(1):311-6.

(310) Gurnett CA, Veile R, Zempel J, Blackburn L, Lovett M, Bowcock A. Disruption of sodium bicarbonate transporter SLC4A10 in a patient with complex partial epilepsy and mental retardation. *Arch Neurol* 2008;65(4):550-3.

(311) Martins-de-Souza D, Gattaz WF, Schmitt A et al. Proteome analysis of schizophrenia patients Wernicke's area reveals an energy metabolism dysregulation. *BMC Psychiatry* 2009;9:17.

(312) Martins-de-Souza D, Gattaz WF, Schmitt A et al. Proteomic analysis of dorsolateral prefrontal cortex indicates the involvement of cytoskeleton, oligodendrocyte, energy metabolism and new potential markers in schizophrenia. *J Psychiatr Res* 2009;43(11):978-86.

(313) Benyamin B, McRae AF, Zhu G et al. Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels. *Am J Hum Genet* 2009;84(1):60-5.

(314) Beutler E, Gelbart T, Lee P, Trevino R, Fernandez MA, Fairbanks VF. Molecular characterization of a case of atransferrinemia. *Blood* 2000;96(13):4071-4.

(315) Yan W, Wang WL, Zhu F, Chen SQ, Li QL, Wang L. Isolation of a novel member of small G protein superfamily and its expression in colon cancer. *World J Gastroenterol* 2003;9(8):1719-24.

(316) Li Q, Yan W, Cheng S et al. Introduction of G1 phase arrest in Human Hepatocellular carcinoma cells (HHCC) by APMCF1 gene transfection through the down-regulation of TIMP3 and up-regulation of the CDK inhibitors p21. *Mol Biol Rep* 2006;33(4):257-63.

(317) Freathy RM, Mook-Kanamori DO, Sovio U et al. Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight. *Nat Genet* 2010;42(5):430-5.

(318) Kozyrev SV, Abelson AK, Wojcik J et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet* 2008;40(2):211-6.

(319) Orozco G, Abelson AK, Gonzalez-Gay MA et al. Study of functional variants of the BANK1 gene in rheumatoid arthritis. *Arthritis Rheum* 2009;60(2):372-9.

(320) Dieude P, Wipff J, Guedj M et al. BANK1 is a genetic risk factor for diffuse cutaneous systemic sclerosis and has additive effects with IRF5 and STAT4. *Arthritis Rheum* 2009;60(11):3447-54.

(321) Krantz ID, McCallum J, DeScipio C et al. Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. *Nat Genet* 2004;36(6):631-5.

(322) Tonkin ET, Wang TJ, Lisgo S, Bamshad MJ, Strachan T. NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat Genet* 2004;36(6):636-41.

(323) Gudbjartsson DF, Walters GB, Thorleifsson G et al. Many sequence variants affecting diversity of adult human height. *Nat Genet* 2008;40(5):609-15.

(324) Kim JJ, Lee HI, Park T et al. Identification of 15 loci influencing height in a Korean population. *J Hum Genet* 2010;55(1):27-31.

(325) Lesch KP, Timmesfeld N, Renner TJ et al. Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. *J Neural Transm* 2008;115(11):1573-85.

(326) Streit M, Riccardi L, Velasco P et al. Thrombospondin-2: a potent endogenous inhibitor of tumor growth and angiogenesis. *Proc Natl Acad Sci U S A* 1999;96(26):14888-93.

(327) Kyriakides TR, Zhu YH, Smith LT et al. Mice that lack thrombospondin 2 display connective tissue abnormalities that are associated with disordered collagen fibrillogenesis, an increased vascular density, and a bleeding diathesis. *J Cell Biol* 1998;140(2):419-30.

(328) Kishi M, Nakamura M, Nishimine M et al. Loss of heterozygosity on chromosome 6q correlates with decreased thrombospondin-2 expression in human salivary gland carcinomas. *Cancer Sci* 2003;94(6):530-5.

(329) Hirose Y, Chiba K, Karasugi T et al. A functional polymorphism in THBS2 that affects alternative splicing and MMP binding is associated with lumbar-disc herniation. *Am J Hum Genet* 2008;82(5):1122-9.

(330) Koch W, Hoppmann P, de WA, Schomig A, Kastrati A. Polymorphisms in thrombospondin genes and myocardial infarction: a case-control study and a meta-analysis of available evidence. *Hum Mol Genet* 2008;17(8):1120-6.

(331) Parati G, Di RM, Ulian L et al. Clinical relevance blood pressure variability. *Journal of Hypertension - Supplement* 1998;16(3):S25-S33.

(332) Manga P, Kromberg JG, Box NF, Sturm RA, Jenkins T, Ramsay M. Rufous oculocutaneous albinism in southern African Blacks is caused by mutations in the TYRP1 gene. *Am J Hum Genet* 1997;61(5):1095-101.

(333) Sulem P, Gudbjartsson DF, Stacey SN et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet* 2008;40(7):835-7.

(334) Gudbjartsson DF, Sulem P, Stacey SN et al. ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nat Genet* 2008;40(7):886-91.

(335) Nan H, Kraft P, Hunter DJ, Han J. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. *Int J Cancer* 2009;125(4):909-17.

(336) Bolino A, Muglia M, Conforti FL et al. Charcot-Marie-Tooth type 4B is caused by mutations in the gene encoding myotubularin-related protein-2. *Nat Genet* 2000;25(1):17-9.

(337) Han JW, Zheng HF, Cui Y et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* 2009;41(11):1234-7.

(338) Yang W, Shen N, Ye DQ et al. Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet* 2010;6(2):e1000841.

(339) Dubois PC, Trynka G, Franke L et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;42(4):295-302.

(340) Keene P, Mendelow B, Pinto MR et al. Abnormalities of chromosome 12p13 and malignant proliferation of eosinophils: a nonrandom association. *Br J Haematol* 1987;67(1):25-31.

(341) Knezevich SR, McFadden DE, Tao W, Lim JF, Sorensen PH. A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma. *Nat Genet* 1998;18(2):184-7.

(342) Barjesteh van Waalwijk van Doorn-Khosrovani, Spensberger D, de KY, Tang M, Lowenberg B, Delwel R. Somatic heterozygous mutations in ETV6 (TEL) and frequent absence of ETV6 protein in acute myeloid leukemia. *Oncogene* 2005;24(25):4129-37.

(343) Styrkarsdottir U, Halldorsson BV, Gretarsdottir S et al. Multiple genetic loci for bone mineral density and fractures. *N Engl J Med* 2008;358(22):2355-65.

(344) Styrkarsdottir U, Halldorsson BV, Gretarsdottir S et al. New sequence variants associated with bone mineral density. *Nat Genet* 2009;41(1):15-7.

(345) Rivadeneira F, Styrkarsdottir U, Estrada K et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* 2009;41(11):1199-206.

(346) Sobacchi C, Frattini A, Guerrini MM et al. Osteoclast-poor human osteopetrosis due to mutations in the gene encoding RANKL. *Nat Genet* 2007;39(8):960-2.

(347) Gould DB, Phalan FC, Breedveld GJ et al. Mutations in Col4a1 cause perinatal cerebral hemorrhage and porencephaly. *Science* 2005;308(5725):1167-71.

(348) Breedveld G, de Coo IF, Lequin MH et al. Novel mutations in three families confirm a major role of COL4A1 in hereditary porencephaly. *J Med Genet* 2006;43(6):490-5.

(349) Gould DB, Phalan FC, van Mil SE et al. Role of COL4A1 in small-vessel disease and hemorrhagic stroke. *N Engl J Med* 2006;354(14):1489-96.

(350) Sibon I, Coupry I, Menegon P et al. COL4A1 mutation in Axenfeld-Rieger anomaly with leukoencephalopathy and stroke. *Ann Neurol* 2007;62(2):177-84.

(351) Feys T, Poppe B, De PK et al. A detailed inventory of DNA copy number alterations in four commonly used Hodgkin's lymphoma cell lines. *Haematologica* 2007;92(7):913-20.

(352) Li VS, Yuen ST, Chan TL et al. Frequent inactivation of axon guidance molecule RGMA in human colon cancer through genetic and epigenetic mechanisms. *Gastroenterology* 2009;137(1):176-87.

(353) de Quervain DJ, Papassotiropoulos A. Identification of a genetic cluster influencing memory performance and hippocampal activity in humans. *Proc Natl Acad Sci U S A* 2006;103(11):4270-4.

(354) Bi H, Sze CI. N-methyl-D-aspartate receptor subunit NR2A and NR2B messenger RNA levels are altered in the hippocampus and entorhinal cortex in Alzheimer's disease. *J Neurol Sci* 2002;200(1-2):11-8.

(355) Turic D, Langley K, Mills S et al. Follow-up of genetic linkage findings on chromosome 16p13: evidence of association of N-methyl-D aspartate glutamate receptor 2A gene polymorphism with ADHD. *Mol Psychiatry* 2004;9(2):169-73.

(356) Itokawa M, Yamada K, Yoshitsugu K et al. A microsatellite repeat in the promoter of the N-methyl-D-aspartate receptor 2A subunit (GRIN2A) gene suppresses transcriptional activity and correlates with chronic outcome in schizophrenia. *Pharmacogenetics* 2003;13(5):271-8.

(357) Iwayama-Shigeno Y, Yamada K, Itokawa M et al. Extended analyses support the association of a functional (GT)n polymorphism in the GRIN2A promoter with Japanese schizophrenia. *Neurosci Lett* 2005;378(2):102-5.

(358) Karolewicz B, Szebeni K, Gilmore T, Maciag D, Stockmeier CA, Ordway GA. Elevated levels of NR2A and PSD-95 in the lateral amygdala in depression. *Int J Neuropsychopharmacol* 2009;12(2):143-53.

(359) Taniguchi S, Nakazawa T, Tanimura A et al. Involvement of NMDAR2A tyrosine phosphorylation in depression-related behaviour. *EMBO J* 2009;28(23):3717-29.

(360) Arning L, Kraus PH, Valentin S, Saft C, Andrich J, Epplen JT. NR2A and NR2B receptor gene variations modify age at onset in Huntington disease. *Neurogenetics* 2005;6(1):25-8.

(361) Andresen JM, Gayan J, Cherny SS et al. Replication of twelve association studies for Huntington's disease residual age of onset in large Venezuelan kindreds. *J Med Genet* 2007;44(1):44-50.

(362) Arning L, Saft C, Wieczorek S, Andrich J, Kraus PH, Epplen JT. NR2A and NR2B receptor gene variations modify age at onset in Huntington disease in a sex-specific manner. *Hum Genet* 2007;122(2):175-82.

(363) von EJ, Mack V, Sprengel R et al. CKAMP44: a brain-specific protein attenuating short-term synaptic plasticity in the dentate gyrus. *Science* 2010;327(5972):1518-22.

(364) O'Donovan MC, Craddock N, Norton N et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 2008;40(9):1053-5.

(365) van LM, Hartigan N, Hatch J, Benham AM. PDILT, a divergent testis-specific protein disulfide isomerase with a non-classical SXXC motif that engages in disulfide-dependent interactions in the endoplasmic reticulum. *J Biol Chem* 2005;280(2):1376-83.

(366) Qin Z, Ren F, Xu X et al. ZNF536, a novel zinc finger protein specifically expressed in the brain, negatively regulates neuron differentiation by repressing retinoic acid-induced gene transcription. *Mol Cell Biol* 2009;29(13):3633-43.

(367) Benjamin EJ, Dupuis J, Larson MG et al. Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med Genet* 2007;8 Suppl 1:S11.

(368) Trevino LR, Yang W, French D et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* 2009;41(9):1001-5.

(369) Dong JT, Zhang SZ, Ma YX et al. Screening for ZNF230 gene mutation and analysis of its correlation with azoospermia. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 2005;22(3):258-60.

(370) Need AC, Attix DK, McEvoy JM et al. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum Mol Genet* 2009;18(23):4650-61.

(371) Cirulli ET, Kasperaviciute D, Attix DK et al. Common genetic variation and performance on standardized cognitive tests. *Eur J Hum Genet* 2010.

(372) Ganesh SK, Zakai NA, van Rooij FJ et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat Genet* 2009;41(11):1191-8.